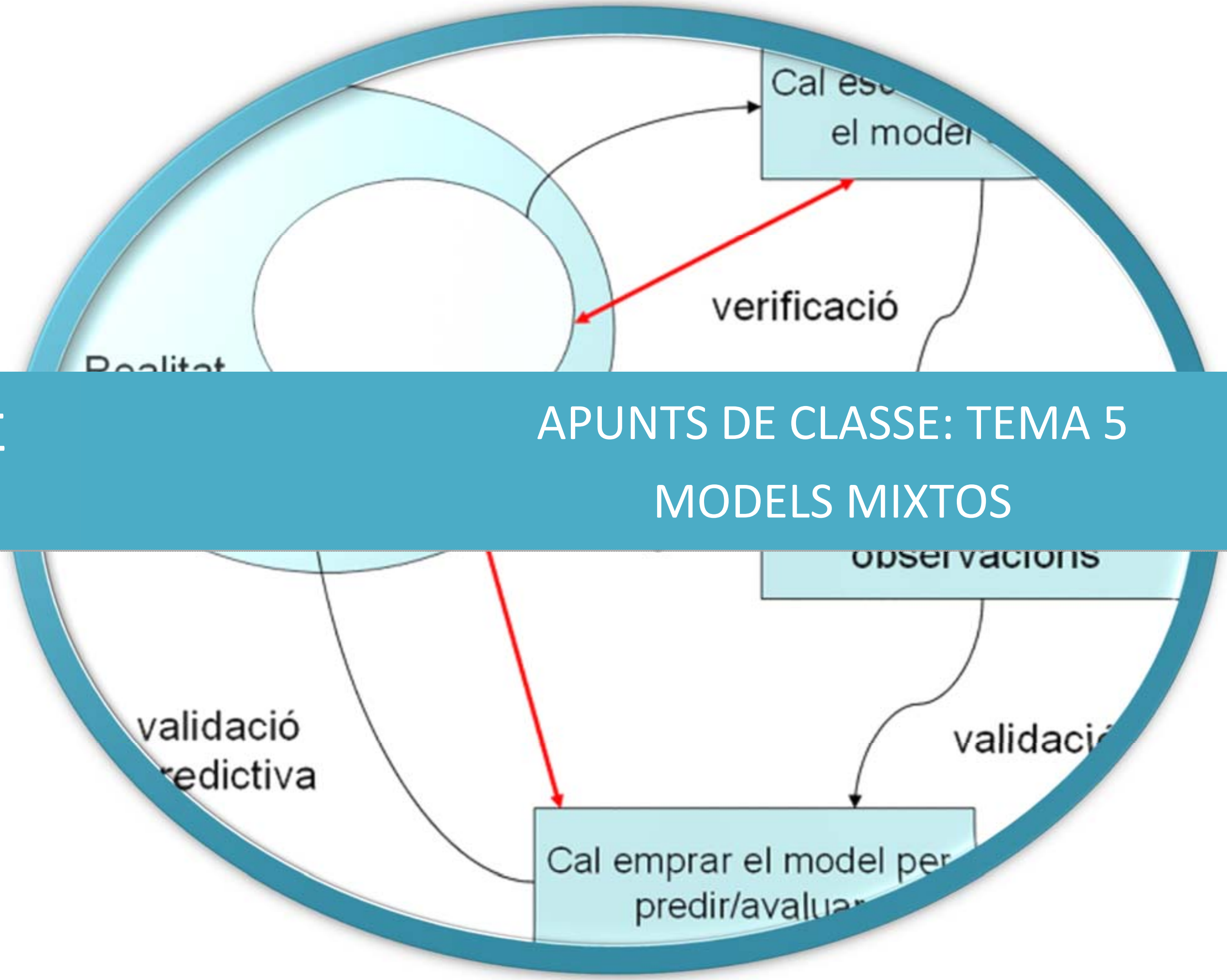
UPC
Departament d'Estadística
i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Curs 2013-2014



MODEL LINEAL GENERALITZAT

APUNTS DE CLASSE: TEMA 5
MODELS MIXTOS

Grau d'Estadística |UB - UPC | Lídia Montero i Josep Anton Sánchez

TABLA DE CONTENIDOS

EXTENSIONES DEL MODELO LINEAL

Objetivo:

Extender la metodología de los Modelos Lineales Generalizados para tratar datos no independientes: medidas repetidas y datos longitudinales.

Modelo Lineal:

Distribución condicional de la respuesta (parte aleatoria):

$$Y_i|X_i \sim N(\mu_i, \sigma)$$

$$E(Y_i|X_i) = \mu_i$$

$$V(Y_i|X_i) = \sigma^2$$

Relación del parámetro Esperanza Condicionada con covariables (parte sistemática)

$$\mu_i = X_i\beta$$

Premisas:

- Linealidad: $E(Y_i|X_i) = \mu_i = X_i\beta$
- Homocedasticidad (var. cte.): $V(Y_i|X_i) = \sigma^2$
- Normalidad: $Y_i|X_i \sim Normal$
- Independència: $Y_i\big|X_i \perp Y_j\big|X_j, i \neq j$

EXTENSIONES DEL MODELO LINEAL

<u>1ª Extensión:</u>

Relajamos las condiciones de la distribución condicional de la respuesta

Modelo Lineal Generalizado:

Distribución condicional de la respuesta (parte aleatoria):

$$Y_i|X_i \sim F(.;\mu_i,\phi)$$

$$E(Y_i|X_i) = \mu_i$$

$$V(Y_i|X_i) = \phi v(\mu_i)$$

Relación del parámetro Esperanza Condicionada con covariables (parte sistemática)

$$g(\mu_i) = X_i\beta$$

Premisas:

- Linealidad en la escala del predictor: $g(E(Y_i|X_i)) = g(\mu_i) = X_i\beta$
- Varianza determinada por la distribución: $V(Y_i|X_i) = \phi v(\mu_i)$
- Distribución condicional: $Y_i|X_i \sim F$
- Independència: $Y_i\big|X_i \perp Y_j\big|X_j, i \neq j$

EXTENSIONES DEL MODELO LINEAL

Modelo Lineal Generalizado con Respuesta Binaria:

Distribución condicional de la respuesta (parte aleatoria): Binomial

$$Y_i|X_i \sim Bern(\pi_i), \phi = 1$$

$$E(Y_i|X_i) = \pi_i$$

$$V(Y_i|X_i) = \phi\pi_i(1-\pi_i)$$

Relación del parámetro Esperanza Condicionada con covariables (parte sistemática)

$$g(\pi_i) = X_i\beta$$

$g(.)$ logit, probit, clog-log,…

Premisas:

- Linealidad en la escala del predictor: $g(E(Y_i|X_i)) = g(\mu_i) = X_i\beta$
- Varianza determinada por la distribución: $V(Y_i|X_i) = \pi_i(1-\pi_i)$ (si $\phi \neq 1$, sobre(o infra)-dispersión)
- Distribución condicional: $Y_i|X_i \sim Bern$
- Independència: $Y_i\big|X_i \perp Y_j\big|X_j, i \neq j$

EXTENSIONES DEL MODELO LINEAL

Modelo Lineal Generalizado con Recuentos:

Distribución condicional de la respuesta (parte aleatoria): Poisson

$$Y_i|X_i \sim Pois(\lambda_i), \phi = 1$$

$$E(Y_i|X_i) = \lambda_i$$

$$V(Y_i|X_i) = \phi\lambda_i$$

Relación del parámetro Esperanza Condicionada con covariables (parte sistemática)

$$g(\lambda_i) = X_i\beta$$

$$g(.)\ \text{log}$$

Premisas:

- Linealidad en la escala del predictor: $g(E(Y_i|X_i)) = g(\lambda_i) = X_i\beta$
- Varianza determinada por la distribución: $V(Y_i|X_i) = \lambda_i$ (si $\phi \neq 1$, sobre(o infra)-dispersión)
- Distribución condicional: $Y_i|X_i \sim Pois$
- Independència: $Y_i\big|X_i \perp Y_j\big|X_j, i \neq j$

EXTENSIONES DEL MODELO LINEAL

Modelo Lineal Generalizado con Valores Positivos (p.ej. tiempos):

Distribución condicional de la respuesta (parte aleatoria): Gamma

$$Y_i|X_i \sim \Gamma(\alpha, \beta_i)$$

$$E(Y_i|X_i) = \mu_i = \alpha\beta_i$$

$$V(Y_i|X_i) = \phi v(\mu_i) = \alpha\beta_i^2, \phi = \frac{1}{\alpha}, v(\mu_i) = \mu_i^2$$

Relación del parámetro Esperanza Condicionada con covariables (parte sistemática)

$$g(\mu_i) = X_i\beta$$

$g(.)$ inversa (μ^{-1})

Premisas:

- Linealidad en la escala del predictor: $g(E(Y_i|X_i)) = g(\mu_i) = X_i\beta$
- Varianza determinada por la distribución: $V(Y_i|X_i) = \frac{\mu_i}{\alpha}$ $\left(CV(Y_i|X_i) = \frac{\sqrt{V(Y_i|X_i)}}{E(Y_i|X_i)} = \frac{1}{\sqrt{\alpha}}\right.$ constante)
- Distribución condicional: $Y_i|X_i \sim \Gamma$
- Independència: $Y_i \big| X_i \perp Y_j \big| X_j, i \neq j$

DATOS AGRUPADOS

En todos los casos, la premisa de independencia es necesaria. La muestra ha de ser aleatoria y simple: cada observación se ha obtenido de unidades experimentales independientes y cada unidad experimental contribuye a la muestra con una sola observación.

Datos Agrupados

En multitud de situaciones es muy común que exista algún tipo de agrupación en los datos:

- Se desea medir el efecto de cuatro fertilizantes (A,B,C y D) en la producción de un cereal. Se realiza un diseño por bloques aleatorizados seleccionando 10 parcelas que se subdividen en cuatro subparcelas a las que se aplican cada uno de los tratamientos

- Se desea comparar la productividad de tres máquinas (A,B y C). Puesto que los operarios tienen diferentes habilidades, se seleccionan 5 operarios y se les hace utilizar las tres máquinas

- Se quiere analizar el efecto de dos metodologías docentes (A y B) en el rendimiento escolar en matemáticas de los estudiantes de secundaria. Se seleccionan 20 escuelas (10 de tipo A y 10 de tipo B) , dentro de cada escuela 3 clases y dentro de cada clase 5 estudiantes

- Se desea comparar la evolución de la altura según el género en la adolescencia. Se seleccionan 10 niños y 10 niñas y se mide su altura semestralmente durante 8 años

DATOS AGRUPADOS

En los casos mencionados anteriormente, la respuesta se podría considerar gaussiana y los tratamientos a analizar son de tipo factor (fertilizantes, máquinas, metodología y género) pero podrían existir covariantes de tipo numérico y otros factores que podrían afectar a la respuesta.

¿Es posible utilizar el Modelo Lineal para estimar el modelo y hacer inferencia u obtener predicciones?

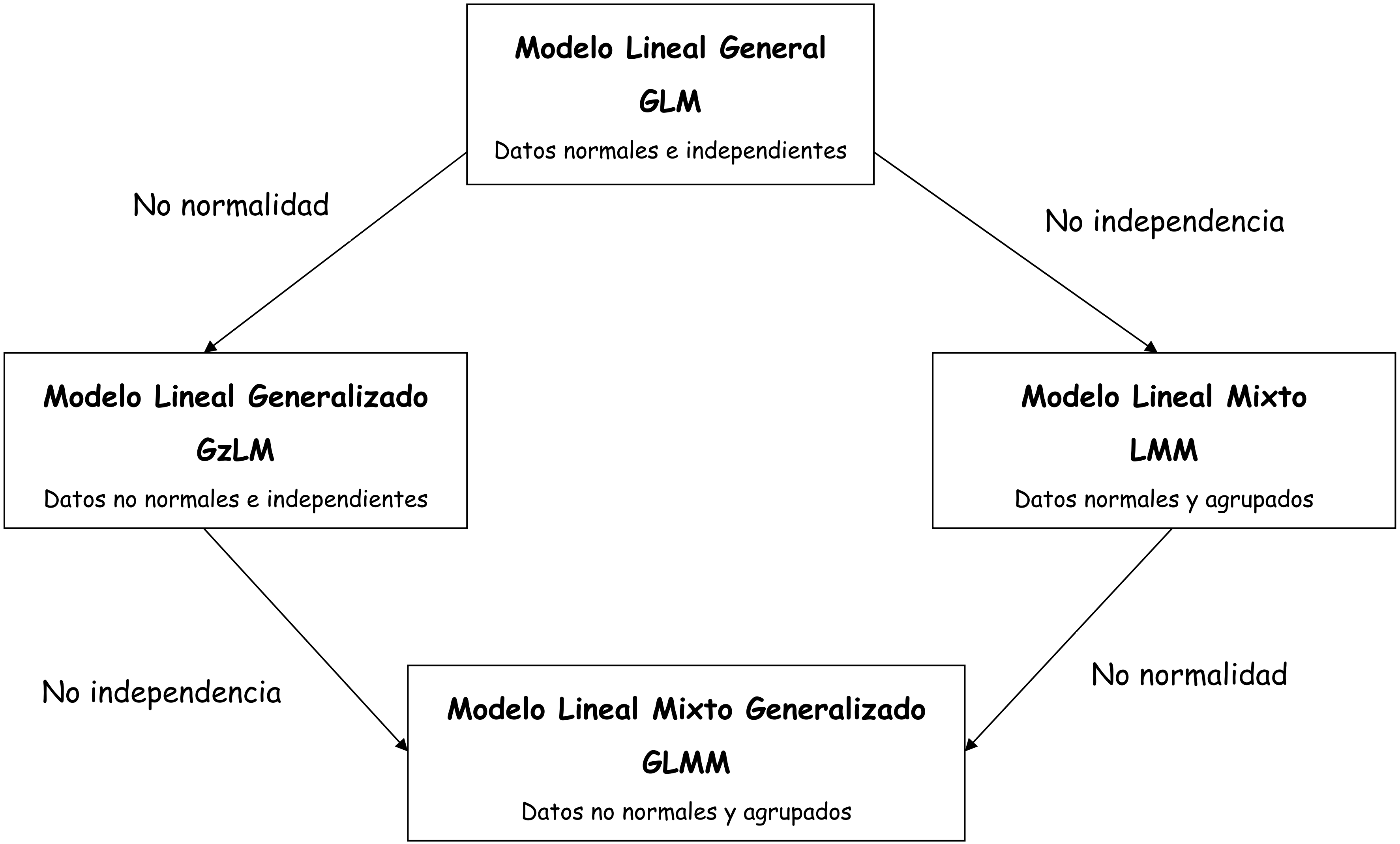
Los datos <u>no son independientes:</u>

- Las 4 observaciones obtenidas en la misma parcela (bloque) están relacionadas
- Las productividades obtenidas por las 3 máquinas cuando el operario es el mismo están relacionadas
- Las 5 notas obtenidas por alumnos de la misma clase en el mismo colegio están relacionadas y las 15 notas de alumnos del mismo colegio, también
- Las 16 medidas realizadas en cada individuo están relacionadas

En todos los casos, además de la respuesta y del tratamiento a testar, existe al menos una variable de tipo categórico que induce una partición de la muestra, haciendo que los datos estén agrupados (parcela, operario, clase/colegio, individuo) y no se puedan considerar independientes. Además, esta variable tiene carácter aleatorio (sus niveles corresponden a una selección aleatoria dentro de la población)

Es necesario extender el Modelo Lineal para relajar la condición de independencia. Hay diferentes formas de hacerlo, pero una de las más potentes son los <u>Modelos Lineales Mixtos</u>

MODELOS LINEALES MIXTOS



MODELOS LINEALES MIXTOS

Diferentes situaciones habituales que dan lugar a datos agrupados:

- Diseños de bloques aleatorizados (parcelas): Los diferentes tratamientos de interés se aplican en cada bloque.
- Diseños de medidas repetidas (operarios): Cada unidad experimental contribuye con más de una observación, pero no existe estructura temporal ni espacial
- Diseños jerárquicos o multinivel (colegios/clases): Hay diferentes capas de agrupación
- Diseños de datos logitudinales (individuos): Existe una covariable temporal/espacial que supone una estructura particular de dependencia entre observaciones

<u>Diseño con k grupos, no necesariamente balanceado (diferente número de observaciones por grupo)</u>

Vector de observaciones del individuo i-ésimo de dimensión ($1 \times n_i$):

$$Y_i = \left(y_{i1}, \dots, y_{in_i}\right)' \quad i = 1..k$$

Matriz de diseño del individuo i-ésimo de dimensión ($p \times n_i$):

$$X_i = \begin{pmatrix} x_{i1}^{(1)} & \cdots & x_{i1}^{(p)} \\ \vdots & \ddots & \vdots \\ x_{in_i}^{(1)} & \cdots & x_{in_i}^{(p)} \end{pmatrix} \quad i = 1..k$$

MODELOS LINEALES MIXTOS

Si realizamos un modelo poblacional con todas las observaciones sin tener en cuenta la estructura de agrupación ➔ modelo con pocos parámetros pero no válido por no ser datos independientes

$$Y_i = X_i\beta + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2 I)$$

p+1 parámetros: vector β $(1xp)$ y escalar σ desviación estándar residual

Si realizamos un modelo individual con los datos de cada grupo ➔ modelo válido (dentro de cada grupo las observaciones son independientes) pero con muchos parámetros y solo útiles para esas unidades experimentales que han sido seleccionadas aleatoriamente.

$$Y_i = X_i\beta_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2 I)$$

kp+1 parámetros: vectores β_i $(1xp)$ y escalar σ desviación estándar residual

MODELOS LINEALES MIXTOS

Solución: Asumimos un modelo poblacional subyacente válido, que servirá para hacer inferencia, y el modelo para cada individuo posee coeficientes que son "desviaciones" aleatorias de los coeficientes del modelo poblacional.

Los parámetros asociados al modelo poblacional son los efectos fijos. Las "desviaciones" para obtener el modelo individual se denominan efectos aleatorios. La presencia de efectos fijos y aleatorios en el modelo da lugar a los Modelos Mixtos. Hay diferentes terminologías para el mismo modelo, según el área de aplicación, pero el modelo es el mismo:

Modelos jerárquicos/Multinivel:

$$\begin{cases} Y_i = X_i\beta_i + \varepsilon_i & \varepsilon_i \sim N\left(0, \sigma^2 I\right) \\ \beta_i = \beta + b_i & b_i \sim N(0, D) \end{cases}$$

Modelos de coeficientes aleatorios:

$$Y_i = X_i\beta_i + \varepsilon_i, \qquad \beta_i \sim N(\beta, D), \quad \varepsilon_i \sim N(0, \sigma^2 I)$$

Modelos con efectos aleatorios

$$Y_i = X_i(\beta + b_i) + \varepsilon_i, \qquad b_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \sigma^2 I)$$

Modelos Mixtos:

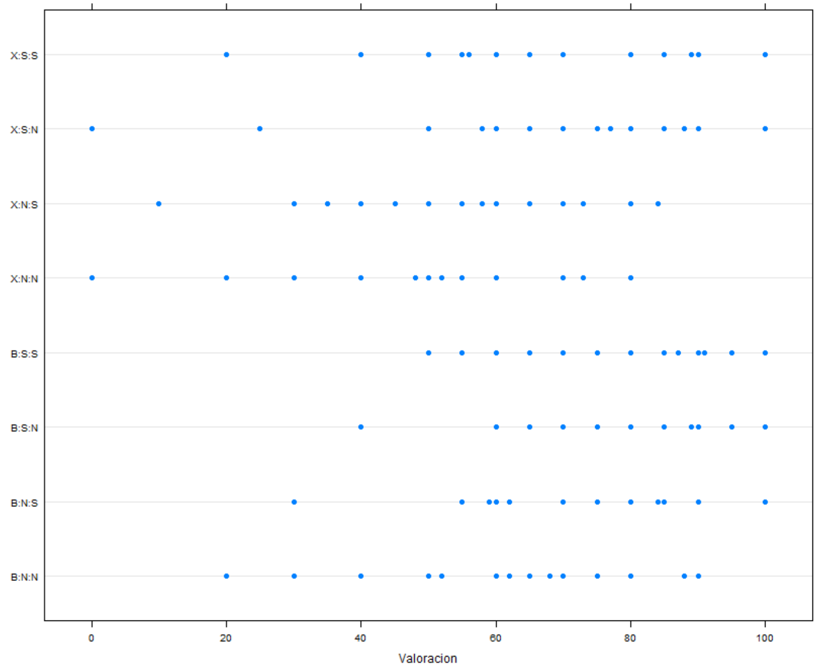
$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i, \qquad b_i \sim N(0, D), \quad \varepsilon_i \sim N(0, \sigma^2 I)$$

EJEMPLO 1: MODELO DE BLOQUES ALEATORIZADOS

Ejemplo: galletas (Valoración sensorial de un tipo de galletas ("filipinos") según marca, tipo de chocolate y enmascaramiento)

Diseño factorial 2^3 con 31 réplicas por condición experimental. Se recogen datos de la valoración sensorial de una galleta realizada por un individuo, en una escala de 0 a 100. Hay dos tipos de chocolate (Blanco:B y Negro:X), las galletas pueden ser de Marca o no, y la degustación se ha realizado con o sin enmascaramiento. En total hay 8 posibles condiciones experimentales, y puesto que existen réplicas, es posible analizar interacciones.

```
>Anova(lm(Valoracion~Tipo*Marca*Enmascaramiento,galletas))
Anova Table (Type II tests)

Response: Valoracion
                           Sum Sq  Df F value    Pr(>F)    
Tipo                        16551   1 62.2081 1.088e-13 ***
Marca                        7701   1 28.9457 1.767e-07 ***
Enmascaramiento               158   1  0.5942   0.44158    
Tipo:Marca                    883   1  3.3194   0.06971 .  
Tipo:Enmascaramiento            2   1  0.0061   0.93800    
Marca:Enmascaramiento         632   1  2.3766   0.12448    
Tipo:Marca:Enmascaramiento    285   1  1.0723   0.30146    
Residuals                   63855 240                      
```



EJEMPLO 1: MODELO DE BLOQUES ALEATORIZADOS

El ajuste del modelo, manteniendo términos con significación inferior a 0.15 es:

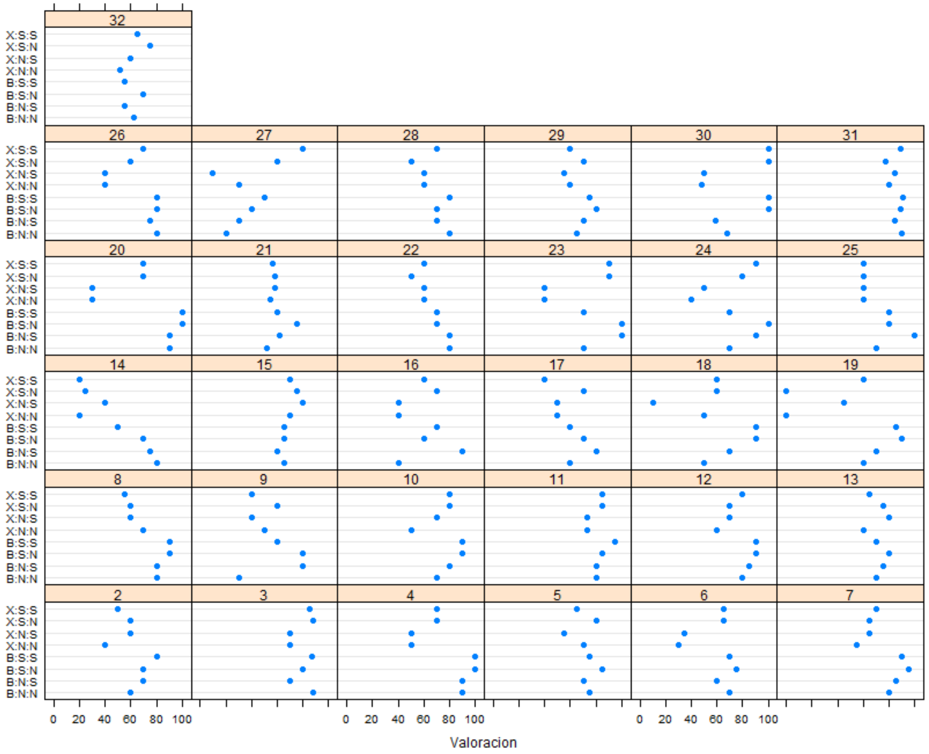
```
lm(formula = Valoracion ~ Tipo + Marca + Enmascaramiento + Tipo:Marca +
    Marca:Enmascaramiento, data = galletas)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               69.540      2.532  27.462  < 2e-16 ***
TipoX                    -20.113      2.924  -6.879 5.13e-11 ***
MarcaS                    10.565      3.581   2.950  0.00349 **
EnmascaramientoS           4.790      2.924   1.638  0.10267
TipoX:MarcaS               7.548      4.135   1.825  0.06917 .
MarcaS:EnmascaramientoS   -6.387      4.135  -1.545  0.12376
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.28 on 242 degrees of freedom
Multiple R-squared: 0.2879,   Adjusted R-squared: 0.2731
F-statistic: 19.56 on 5 and 242 DF,  p-value: 2.345e-16
```

Ahora bien, si los datos son independientes el número de individuos que ha participado en el estudio ha de ser el mismo que el tamaño de muestra (n=248!). En realidad, el diseño se realizó con los 31 alumnos de una clase que probaron y valoraron cada uno de ellos las 8 posibles condiciones experimentales. Por lo tanto, los datos NO son independientes y el procedimiento realizado NO es correcto!

EJEMPLO 1: MODELO DE BLOQUES ALEATORIZADOS

En realidad, el diseño es un diseño 2^3 completo con bloques aleatorizados a 31 niveles sin réplicas por condición experimental. Para resolver el diseño en base a modelos mixtos, hemos de suponer que existe un efecto aleatorio no observable (b_i) que justifica la correlación entre las observaciones recogidas en cada individuo.

```
> Anova(lme(Valoracion~Tipo*Marca*Enmascaramiento,galletas,random=~1|ID))

Analysis of Deviance Table (Type II tests)

Response: Valoracion
                            Chisq Df Pr(>Chisq)
Tipo                      93.6437  1  < 2.2e-16 ***
Marca                     43.5728  1  4.085e-11 ***
Enmascaramiento            0.8944  1    0.34429
Tipo:Marca                 4.9968  1    0.02539 *
Tipo:Enmascaramiento       0.0091  1    0.92390
Marca:Enmascaramiento      3.5776  1    0.05856 .
Tipo:Marca:Enmascaramiento 1.6142  1    0.20390
```



EJEMPLO 1: MODELO DE BLOQUES ALEATORIZADOS

La estimación puntual de la parte fija coincide con el análisis anterior, pero la significación de los factores es muy diferente, ya que entes se ignoró que existen componentes de la varianza (debido a los individuos y la residual)

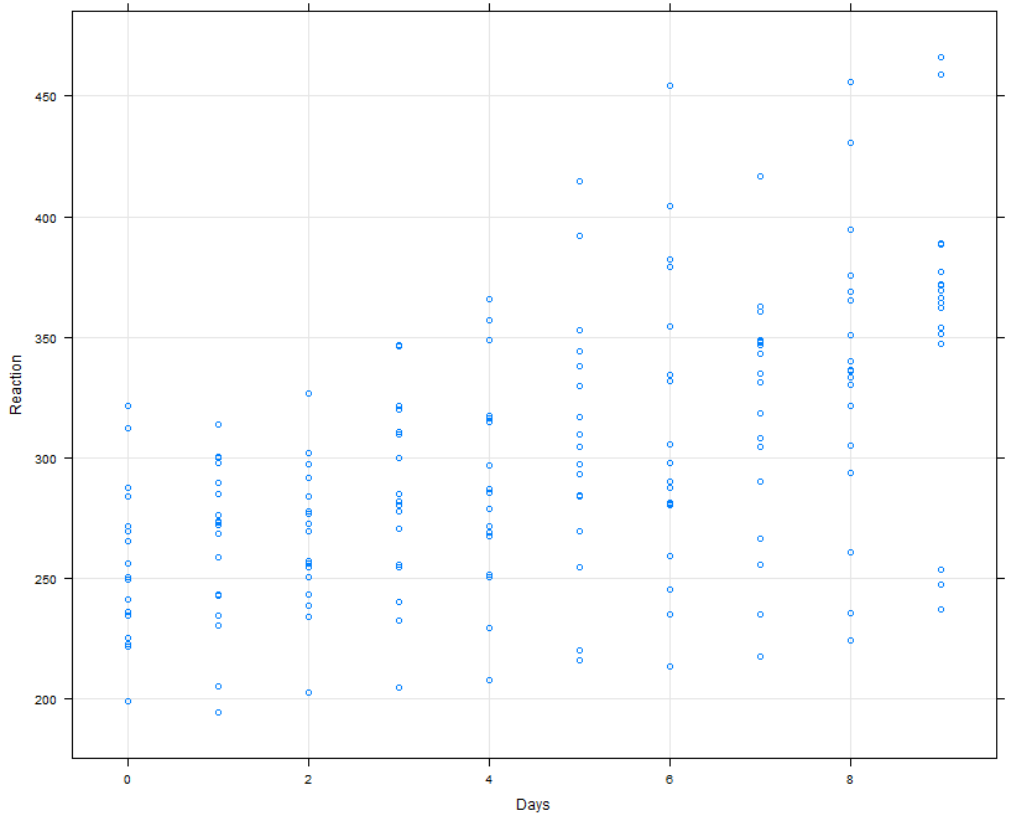
```
> summary(mod<-lme(Valoracion~Tipo*Marca+Marca*Enmascaramiento,galletas,random=~1|ID))
Linear mixed-effects model fit by REML
 Data: galletas
       AIC      BIC    logLik
  2026.584 2054.495 -1005.292

Random effects:
 Formula: ~1 | ID
        (Intercept) Residual
StdDev:    9.452726 13.28276

Fixed effects: Valoracion ~ Tipo * Marca + Marca * Enmascaramiento
                            Value Std.Error  DF   t-value p-value
(Intercept)              69.54032  2.674116 212 26.004976  0.0000
TipoX                   -20.11290  2.385654 212 -8.430772  0.0000
MarcaS                   10.56452  2.921817 212  3.615735  0.0004
EnmascaramientoS          4.79032  2.385654 212  2.007970  0.0459
TipoX:MarcaS              7.54839  3.373824 212  2.237339  0.0263
MarcaS:EnmascaramientoS  -6.38710  3.373824 212 -1.893133  0.0597
```

EJEMPLO 2: MODELO DE PENDIENTES ALEATORIAS

Ejemplo: sleepstudy (Tiempo diario de reacción promedio para sujetos en un estudio de privación del sueño)

Se seleccionan 18 individuos. En el día 0 los sujetos tenían su cantidad normal de sueño. Desde aquella noche se limita a 3 horas de sueño por noche. Las observaciones representan el tiempo medio de reacción en una serie de ensayos cada día a cada sujeto.

Reaction	Days	Subject
249.5600	0	308
258.7047	1	308
250.8006	2	308
321.4398	3	308
356.8519	4	308
414.6901	5	308
382.2038	6	308
290.1486	7	308
430.5853	8	308
466.3535	9	308
222.7339	0	309
205.2658	1	309
202.9778	2	309
204.7070	3	309
207.7161	4	309
215.9618	5	309
213.6303	6	309
217.7272	7	309
224.2957	8	309
:	:	:

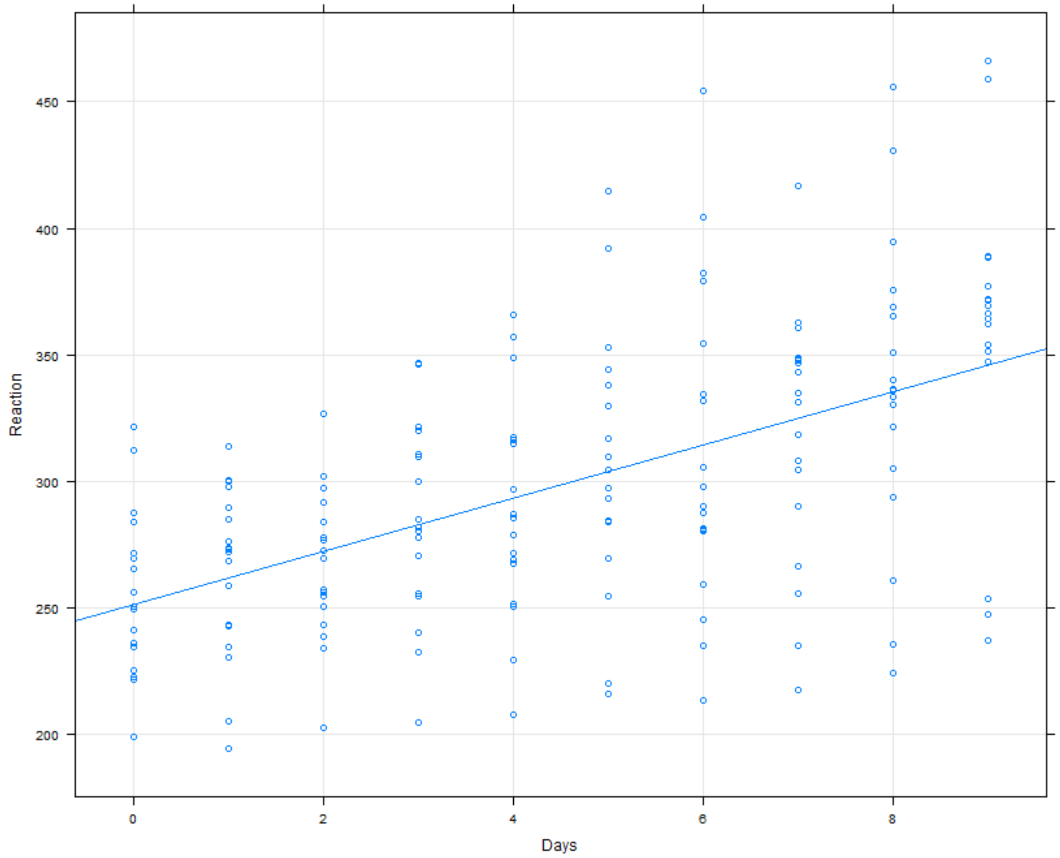


EJEMPLO 2: MODELO DE PENDIENTES ALEATORIAS

Modelo lineal "poblacional": Si ignoramos la estructura de correlación para los datos de un mismo individuo, podemos aplicar el modelo lineal general (OLS) a todo el conjunto:

$$R_{ij} = \beta_0 + \beta_1 D_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2 I)$$

```
> (mod.pobl<-lm(Reaction~Days,sleepstudy))

Coefficients:
(Intercept)         Days
     251.41        10.47
```

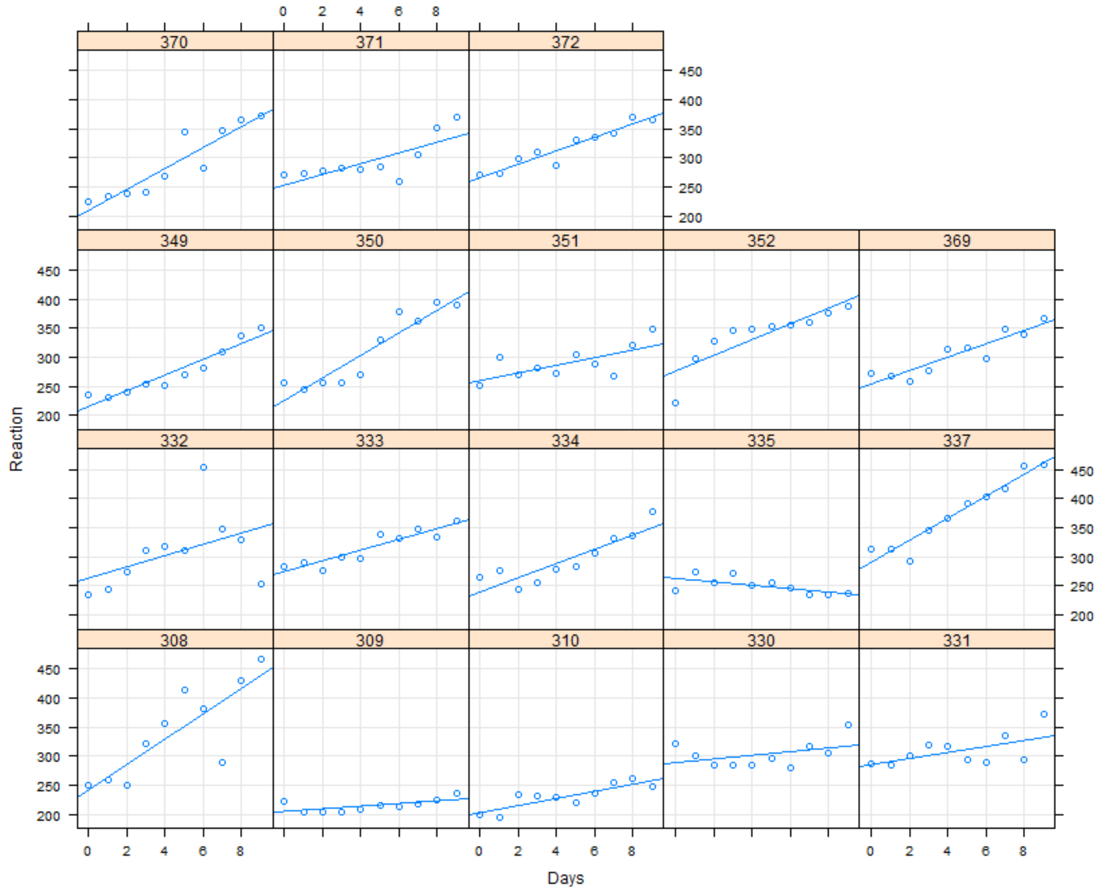


EJEMPLO 2: MODELO DE PENDIENTES ALEATORIAS

Modelo lineal "individual": Hacemos un modelo para cada individuo

$$R_{ij} = \beta_{0i} + \beta_{1i} D_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2 I)$$

```
> (mod.indiv=lmList(Reaction~Days|Subject,sleepstudy))

Coefficients:
    (Intercept)      Days
308    244.1927 21.764702
309    205.0549  2.261785
310    203.4842  6.114899
330    289.6851  3.008073
331    285.7390  5.266019
332    264.2516  9.566768
333    275.0191  9.142045
334    240.1629 12.253141
335    263.0347 -2.881034
337    290.1041 19.025974
349    215.1118 13.493933
350    225.8346 19.504017
351    261.1470  6.433498
352    276.3721 13.566549
369    254.9681 11.348109
370    210.4491 18.056151
371    253.6360  9.188445
372    267.0448 11.298073
```



EJEMPLO 2: MODELO DE PENDIENTES ALEATORIAS

Modelo Lineal Mixto: La parte fija es el modelo poblacional y el factor aleatorio sujeto tiene un efecto aleatorio en el Intercept y otro en la pendiente

$$R_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})D_{ij} + \varepsilon_{ij} \qquad (b_{0i}, b_{1i})' \sim N(0, D) \qquad \varepsilon_{ij} \sim N(0, \sigma^2 I)$$

```
> (mod.mix=lmer(Reaction~Days+(Days|Subject),sleepstudy))

Linear mixed model fit by REML

Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy

  AIC  BIC logLik deviance REMLdev
 1756 1775 -871.8     1752    1744

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.092  24.7405
          Days         35.072   5.9221  0.066
 Residual             654.941  25.5918

Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77
```

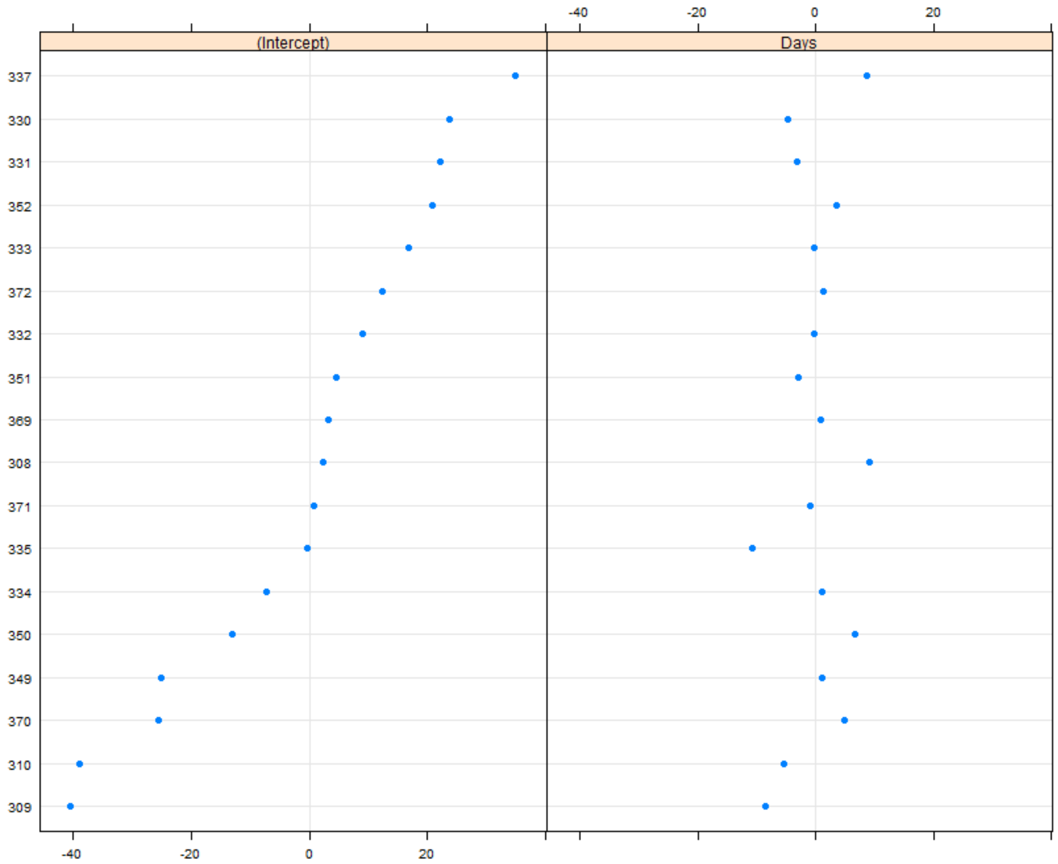
EJEMPLO 2: MODELO DE PENDIENTES ALEATORIAS

Efectos Fijos del modelo mixto (β):

```
> fixef(mod.mix)
(Intercept)        Days
  251.40510    10.46729
```

Efectos Aleatorios del modelo mixto (b):

```
> ranef(mod.mix)
$Subject
    (Intercept)        Days
308   2.2571870   9.1992523
309 -40.3984020  -8.6211497
310 -38.9605268  -5.4502094
330  23.6919454  -4.8137089
331  22.2613341  -3.0692765
332   9.0398323  -0.2718974
333  16.8409575  -0.2231088
334  -7.2329589   1.0743770
335  -0.3320174 -10.7524155
337  34.8900139   8.6295604
349 -25.2110300   1.1726615
350 -13.0713764   6.6139522
351   4.5784421  -3.0151846
352  20.8636459   3.5367394
369   3.2754116   0.8723379
370 -25.6143737   4.8217830
371   0.8072184  -0.9881528
372  12.3146970   1.2844399
```



EJEMPLO 2: MODELO DE PENDIENTES ALEATORIAS

Modelo Lineal General

```
> summary(mod.pobl<-lm(Reaction~Days,sleepstudy))

Residuals:
     Min       1Q   Median       3Q      Max
-110.848  -27.483    1.546   26.142  139.953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  251.405      6.610  38.033  < 2e-16 ***
Days          10.467      1.238   8.454 9.89e-15 ***
---
Residual standard error: 47.71 on 178 degrees of freedom
Multiple R-squared: 0.2865,   Adjusted R-squared: 0.2825
F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.894e-15
```

Incorrecto!!! Datos no independientes!

Modelo Lineal Mixto

```
> (mod.mix=lmer(Reaction~Days+(Days|Subject),sleepstudy))

Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy
  AIC  BIC logLik deviance REMLdev
 1756 1775 -871.8     1752    1744

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.092  24.7405
          Days         35.072   5.9221  0.066
 Residual             654.941  25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84

Days          10.467      1.546    6.77
```

Correcto!!! Efectos aleatorios en datos agrupados

ESTIMACIÓN E INFERENCIA EN MODELOS LINEALES MIXTOS

Verosimilitud:

$$L(\beta,\theta,\sigma|y_1,\dots y_n)=\prod_{i=1}^{n} f_{\beta,\theta,\sigma}(y_i)=\prod_{i=1}^{n}\left(\int_{b_i} f_{\beta,\theta,\sigma}(y_i,b_i)db_i\right)$$
$$=\prod_{i=1}^{n}\left(\int_{b_i} f_{\beta,\sigma}(y_i|b_i)f_\theta(b_i)db_i\right)$$

Puesto que los efectos aleatorios son normales y su matriz de varianzas y covarianzas D está parametrizada con los parámetros θ:

$b_i \sim N\big(0, D(\theta)\big)$

$$f_\theta(b_i)=\frac{1}{\sqrt{2\pi|D(\theta)|}}\exp(-\frac{1}{2}b_i' D(\theta)^{-1}b_i)$$

Y la distribución de y_i conodido el valor del efecto aleatorio b_i también es normal:

$y_i|b_i \sim N(X_i\beta+Z_ib_i,\sigma^2 I)$

$$f_{\beta,\sigma}(y_i|b_i)=\frac{1}{\sqrt{2\pi|\sigma^2 I|}}\exp\left[-\frac{1}{2\sigma^2}\big(y_i-(X_i\beta+Z_ib_i)\big)'\big(y_i-(X_i\beta+Z_ib_i)\big)\right]$$

ESTIMACIÓN E INFERENCIA EN MODELOS LINEALES MIXTOS

El cálculo de la verosimilitud incluye la resolución de una integral (ya que los efectos aleatorios no se observan, es necesario obtener la verosimilitud marginal, integrando la distribución conjunta respecto a los efectos aleatorios).

Estimadores de máxima verosimilitud:

$$argmax_{\beta,\theta,\sigma} L(\beta,\theta,\sigma|y_1,\ldots y_n) = argmax_{\beta,\theta,\sigma} \prod_{i=1}^{n} \left(\int_{b_i} f_{\beta,\sigma}(y_i|b_i) f_\theta(b_i) db_i \right)$$

Los estimadores de las componentes de la varianza están sesgados (el valor esperado del estimador infra-estima los parámetros. Por ello, se trabaja con una variante de la verosimilitud, denominada Verosimilitud Restringida (Restricted Maximum Likelihood, REML)

La parte fija del modelo funciona igual que en los modelos lineales:

Inferencia para β basada en el test de Wald y razón de verosimilitudes:

$$\hat{\beta} \approx N(\beta, I(\beta)^{-1})$$

$$-2\,(\log L_0 - \log L_1) \approx \chi^2_p$$

Para las componentes de la varianza se utiliza el test de razón de verosimilitudes, pero la estimación ha de ser ML y no REML

MODELOS LINEALES MIXTOS GENERALIZADOS

Igual que pasamos del Modelo Lineal al Modelo Lineal Generalizado cambiando la distribución de la respuesta, el Modelo Lineal Mixto Generalizado es un Modelo Lineal Mixto con una distribución diferente para la respuesta.

Datos agrupados:

Vector de observaciones del individuo i-ésimo de dimensión ($1 x n_i$):

$$Y_i = \left(y_{i1}, \dots, y_{in_i}\right)' \quad i = 1..k$$

Matriz de diseño de la parte fija del individuo i-ésimo de dimensión ($p x n_i$):

$$X_i = \begin{pmatrix} x_{i1}^{(1)} & \cdots & x_{i1}^{(p)} \\ \vdots & \ddots & \vdots \\ x_{in_i}^{(1)} & \cdots & x_{in_i}^{(p)} \end{pmatrix} \quad i = 1..k$$

Matriz de diseño de la parte aleatoria del individuo i-ésimo de dimensión ($p x n_i$):

$$Z_i = \begin{pmatrix} z_{i1}^{(1)} & \cdots & z_{i1}^{(q)} \\ \vdots & \ddots & \vdots \\ z_{in_i}^{(1)} & \cdots & z_{in_i}^{(q)} \end{pmatrix} \quad i = 1..k$$

MODELOS LINEALES MIXTOS GENERALIZADOS

Expresión del Modelo Lineal Mixto Generalizado (GLMM):

Distribución condicional de la respuesta (parte aleatoria):

$$Y_i | X_i, Z_i \sim F(.\,; \mu_i, \phi)$$
$$E(Y_i | X_i, Z_i) = \mu_i$$
$$V(Y_i | X_i, Z_i) = \phi v(\mu_i)$$

Relación del parámetro Esperanza Condicionada con covariables (parte sistemática):

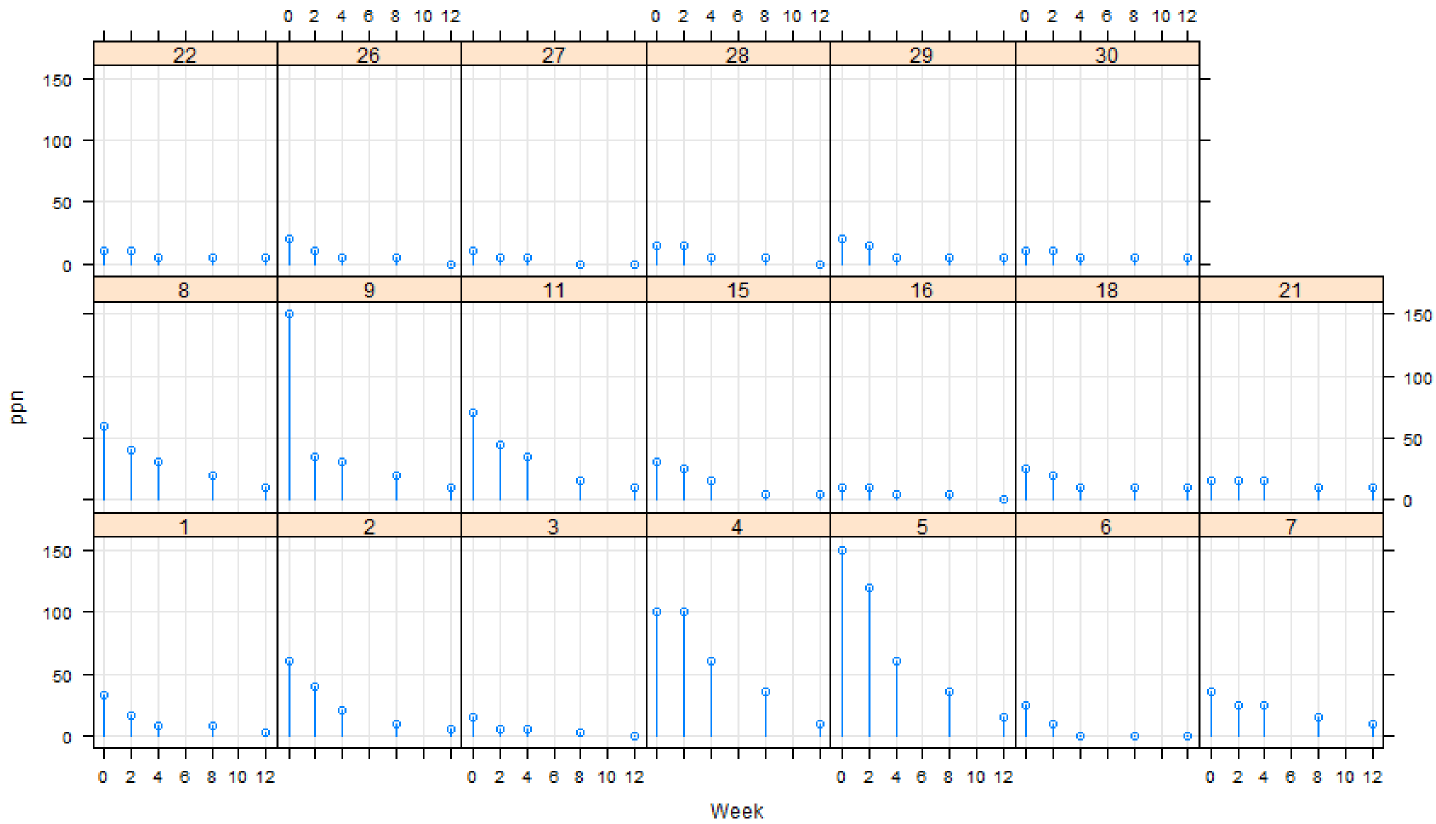
$$g(\mu_i) = X_i \beta + Z_i b_i$$

En los modelos lineales mixtos generalizados, es habitual considerar modelos simples. Si hay datos agrupados se suele trabajar inicialmente con predictores lineales con un único efecto aleatorio afectando al término independiente:

$$Z_i = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad i = 1..k$$

EJEMPLO 3: MODELO LINEAL MIXTO POISSON

En este ejemplo, participan 30 individuos con problemas de lesiones dérmicas en el rostro (pápulas, pústulas y nódulos). A todos se les aplica un tratamiento para corregir estas lesiones. En las semanas 2, 4, 8 y 12 se cuenta el número de lesiones que presentan, para evaluar la eficacia del tratamiento.



EJEMPLO 3: MODELO LINEAL MIXTO POISSON

Para cuantificar la efectividad del tratamiento a lo largo del tiempo, podemos calcular el modelo lineal de Poisson poblacional y sus correspondientes modelos individuales:

y_{ij} Número de lesiones del individuo i-ésimo en la visita j-ésima

Distribución de la respuesta: $Y_{ij} \sim Pois(\lambda_{ij})$

Parte sistemática: $log(\lambda_{ij}) = (\beta_0 + b_i) + \beta_1 Week_{ij}$

Parte aleatoria: $\mathrm{b_i} \sim N(0, \sigma_b)$

```
> (mod=glmer(ppn~Week+(1|Patient),skin,family=poisson))
Generalized linear mixed model fit by the Laplace approximation
Formula: ppn ~ Week + (1 | Patient)
   Data: skin
   AIC   BIC logLik deviance
 301.7 309.6 -147.9    295.7
Random effects:
 Groups  Name        Variance Std.Dev.
 Patient (Intercept) 0.69306  0.8325
Number of obs: 100, groups: Patient, 20

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.353428   0.189389   17.71   <2e-16 ***
Week        -0.175595   0.006853  -25.62   <2e-16 ***
```

EJEMPLO 3: MODELO LINEAL MIXTO POISSON

Evolución de acuerdo al modelo poblacional:

```
> fixef(mod)
(Intercept)        Week
  3.3534285  -0.1755953
```

$$log(\lambda_{ij}) = 3.35 - 0.17Week_{ij}$$

$$\lambda_{ij} = 28.5e^{-0.17Week_{ij}}$$

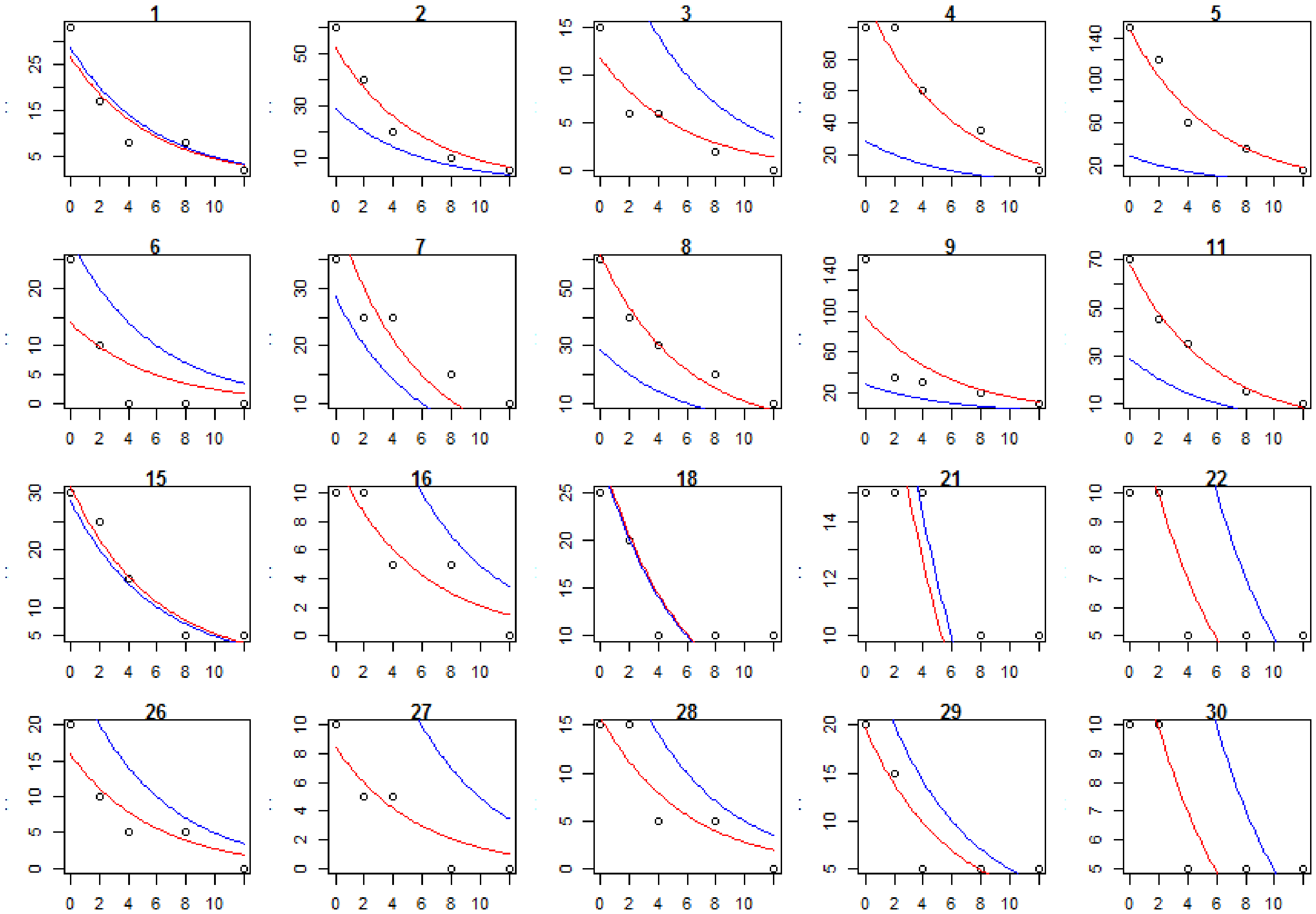
Efectos aleatorios:

```
> ranef(mod)

$Patient
   (Intercept)
1  -0.07478617
2   0.60293028
3  -0.88547025
4   1.41770286
5   1.63805253
6  -0.71161688
7   0.39934878
8   0.77230527
9   1.19829653
11  0.86177572
15  0.08461976
16 -0.85441014
18  0.02120066
21 -0.11885728
22 -0.71161688
26 -0.58607927
27 -1.21605294
28 -0.58607927
29 -0.37314591
30 -0.71161688
```

Modelo individual para el paciente 4:

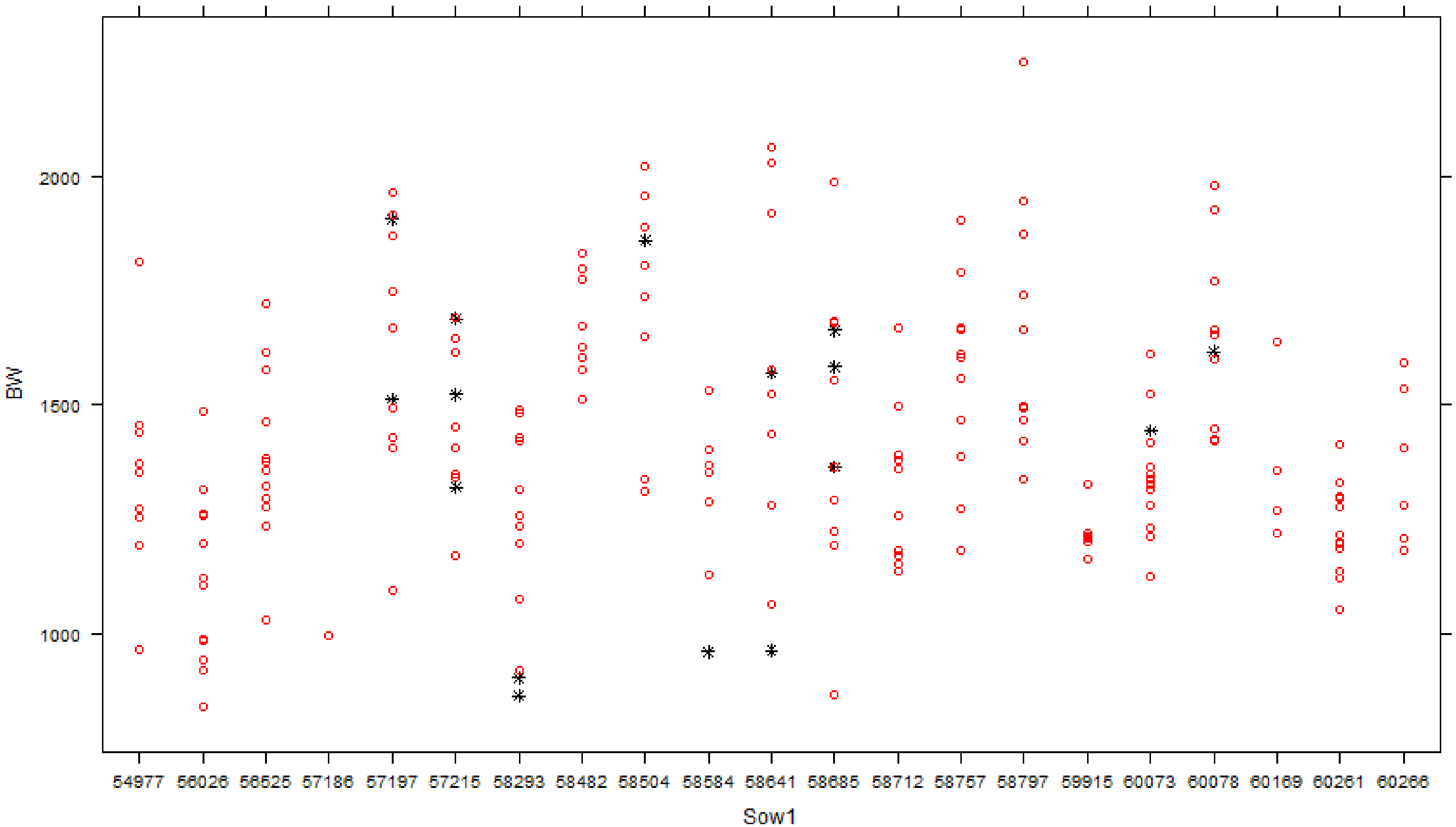
$$log(\lambda_{4j}) = (3.35 + 1.417) - 0.17Week_{4j} = 4.767 - 0.17Week_{ij}$$

EJEMPLO 3: MODELO LINEAL MIXTO POISSON



EJEMPLO 4: MODELO LINEAL MIXTO BINOMIAL

Se desea analizar los factores que determinan la supervivencia de lechones. Para ello se seleccionan 21 madres y para cada cría se registra el sexo, el peso al nacer y su temperatura. Además se desea caracterizar las madres en base a la capacidad de supervivencia de su progenie.



EJEMPLO 4: MODELO LINEAL MIXTO BINOMIAL

Para cuantificar la efectividad del tratamiento a lo largo del tiempo, podemos calcular el modelo lineal de Poisson poblacional y sus correspondientes modelos individuales:

y_{ij} variable binaria (0=no sobrevive, 1=sí sobrevive)

Distribución de la respuesta: $Y_{ij} \sim Bern(\pi_{ij})$

Parte sistemática: $logit(\pi_{ij}) = (\beta_0 + b_i) + \beta_1 Sex_{ij} + \beta_2 T0_{ij} + \beta_3 BW_{ij}$

Parte aleatoria: $\mathrm{b_i} \sim N(0, \sigma_b)$

```
> (mod<-glmer(Survival~Sex+T0+BW+(1|Sow1),lechones,family=binomial))
Generalized linear mixed model fit by the Laplace approximation
Formula: Survival ~ Sex + T0 + BW + (1 | Sow1)
   Data: lechones
   AIC   BIC logLik deviance
 120.2 136.7 -55.08    110.2
Random effects:
 Groups Name        Variance Std.Dev.
 Sow1   (Intercept) 0.27861  0.52783
Number of obs: 201, groups: Sow1, 21
Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.4943374 19.2407750   0.078    0.938
Sex         -0.4673204  0.5521378  -0.846    0.397
T0           0.0240623  0.5198960   0.046    0.963
BW           0.0002858  0.0011247   0.254    0.799
```

EJEMPLO 4: MODELO LINEAL MIXTO BINOMIAL

No parece ser significativo ninguna característica recogida. Puede ser debido a la baja incidencia de defunciones que hace que el modelo sea altamente desbalanceado.

Caracterización de las madres (efectos aleatorios):

```
> ranef(mod)
$Sow1
      (Intercept)
54977  0.14320876
56026  0.22072625
56525  0.22497243
57186  0.02623120
57197 -0.30095563
57215 -0.44747281
58293 -0.25839649
58482  0.12360346
58504 -0.10307085
58584 -0.12154502
58641 -0.29435760
58685 -0.46304800
58712  0.16713390
58757  0.18167053
58797  0.14941353
59915  0.10733159
60073 -0.01759419
60078 -0.06345183
60169  0.06757838
60261  0.22677368
60266  0.10222279
```