

Elaborado por: Andrés García Medina

Fuente: Wood, S. N. (2017). Generalized additive models: an introduction with R. CRC press.

## Modelos Lineales Mixtos

En general, los modelos lineales mixtos amplían el modelo lineal:

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, I\sigma^2)$$

Para agregar efectos aleatorios a través del vector  $b$

$$y = X\beta + Zb + \epsilon, \quad b \sim N(0, \Sigma_\theta), \quad \epsilon \sim N(0, \Sigma_\phi)$$

Con parametros desconocidos en  $(\theta, \phi)$ , mientras que  $Z$  es un modelo matricial para los efectos aleatorios.

Usualmente  $\Sigma_\phi = I\sigma^2$

Ademas, se asume que  $b$  y  $\epsilon$  son independientes.

La idea es permitir una estructura de modelo lineal para el componente aleatorio de los datos de respuesta,  $y$ , que sea tan rica como la estructura de modelo lineal utilizada para modelar el componente sistemático.

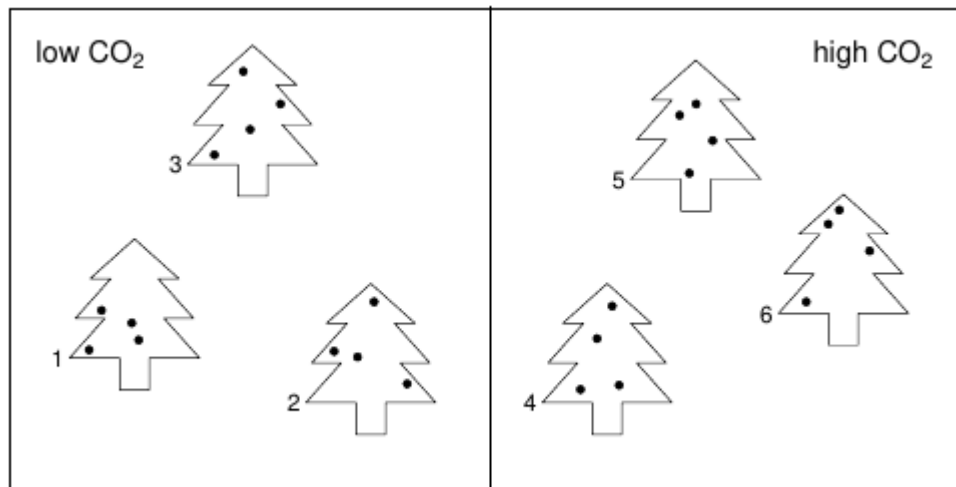
## Modelos mixtos para datos equilibrados.

### Ejemplo guía 1: Motivación

- Las hojas de las plantas tienen pequeños agujeros, llamados "estomas", a través de los cuales toman aire, pero también pierden agua.
- La mayoría de las plantas no tropicales realizan la fotosíntesis de tal manera que, en los días soleados, están limitadas por la cantidad de dióxido de carbono que pueden obtener a través de estos estomas.
- El "problema" de una planta es que si sus estomas son demasiado pequeños no podrá obtener suficiente dióxido de carbono, y si son demasiado grandes perderá demasiada agua en los días soleados.
- Dada la importancia que esto tiene para este tipo de plantas, parece probable que el tamaño de los estomas dependa de la concentración de dióxido de carbono en la atmósfera.

- Esto puede tener implicaciones en el cambio climático si el aumento de la cantidad de CO<sub>2</sub> en la atmósfera hace que las plantas liberen menos agua: el vapor de agua es el gas de efecto invernadero más importante.

Considere un experimento en el que se cultivan plántulas de árboles bajo 2 niveles de concentración de dióxido de carbono, con 3 árboles asignados a cada tratamiento, y suponga que después de 6 meses de crecimiento, el área estomática se mide en 4 ubicaciones aleatorias de cada planta:



## El enfoque equivocado: un modelo lineal de efectos fijos

- Un modelo de estos datos debe incluir un factor (de 2 niveles) para el tratamiento de CO<sub>2</sub>, pero también un factor (de 6 niveles) para cada árbol individual, ya que tenemos múltiples mediciones en cada árbol y debemos esperar cierta variabilidad en el área de estomas de un árbol a otro.
- Entonces un modelo lineal adecuado es

$$y_i = \alpha_j + \beta_k + \epsilon_i, \quad (1)$$

Si la observación  $i$  es para el nivel  $j$  de CO<sub>2</sub> y el árbol  $k$

donde  $y_i$  es la  $i$ -ésima medición del área estomática,  $\alpha_j$  es el área estomática media de la población en el nivel de CO<sub>2</sub>  $j$ ,  $\beta_k$  es la desviación del árbol  $k$  de esa media y  $\epsilon_i$  son  $N(0, \sigma^2)$  variables aleatorias independientes.

Ahora bien, si este es un modelo de efectos fijos, tenemos dos problemas:

- Los  $\alpha_j$  y los  $\beta_k$  están completamente confundidos. Los árboles están "anidados" dentro del tratamiento, con 3 árboles en un tratamiento y 3 en el otro: cualquier número que quieras podría sumarse a  $\alpha_1$  y restarse simultáneamente de  $\beta_1$ ,  $\beta_2$  y  $\beta_3$ , sin cambiar en absoluto las predicciones del modelo, y lo mismo ocurre con  $\alpha_2$  y los  $\beta_k$  's restantes.

- Realmente queremos aprender sobre los árboles en general, pero esto no es posible con un modelo en el que hay un efecto fijo para cada árbol en particular: a menos que los efectos de los árboles resulten insignificantes, no podemos usar el modelo para predecir qué le sucede a un árbol distinto de uno de los seis del experimento.

## Datos

```
In [63]: library(gamair)
         data(stomata)
```

```
In [64]: stomata
```

A data.frame: 24 × 3

	area	CO2	tree
	<dbl>	<fct>	<fct>
1	1.6055739	1	1
2	1.6300711	1	1
3	1.5391189	1	1
4	1.7187315	1	1
5	1.3896163	1	2
6	1.5858805	1	2
7	1.4697276	1	2
8	1.9493473	1	2
9	1.5397020	1	3
10	1.2436558	1	3
11	0.8752505	1	3
12	0.9932352	1	3
13	3.1149370	2	4
14	2.7402102	2	4
15	2.4825228	2	4
16	2.8192831	2	4
17	2.8924475	2	5
18	2.8622759	2	5
19	2.8410755	2	5
20	3.0183753	2	5
21	2.6576575	2	6
22	2.0839150	2	6
23	2.2310707	2	6
24	2.3464027	2	6

Primero compare modelos con y sin el factor de árbol ( $\beta_k$ ):

```
In [65]: m0 <- lm(area ~ CO2, stomata)
          m1 <- lm(area ~ CO2 + tree, stomata)
          anova(m0,m1)
```

A anova: 2 × 6						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	22	2.1348274	NA	NA	NA	NA
2	18	0.8603993	4	1.274428	6.665424	0.001787974

Claramente, existe evidencia sólida de diferencias entre árboles, lo que significa que con este modelo no podemos decir si el CO2 tuvo un efecto o no.

Para volver a enfatizar este punto, esto es lo que sucede si intentamos probar el efecto del CO2:

```
In [66]: m2 <- lm(area ~ tree, stomata)
```

```
In [67]: anova(m2,m1)
```

A anova: 2 × 6						
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	18	0.8603993	NA	NA	NA	NA
2	18	0.8603993	0	-1.110223e-16	NA	NA

La confusión del CO2 y los tres factores significa que los modelos que se comparan aquí son en realidad el mismo modelo: como resultado, dan la misma suma residual de cuadrados y tienen los mismos grados de libertad residuales; "compararlos" no nos dice nada sobre el efecto del CO2.

Tratar los árboles individuales, no como individuos completamente únicos sino como una muestra aleatoria de la población de árboles objetivo, nos permitirá estimar el efecto del CO2 y generalizar más allá de los 6 árboles del experimento

## El enfoque correcto: un modelo de efectos mixtos

- La clave para establecer si el CO2 tiene un efecto es reconocer que el factor CO2 y los factores de los árboles son de naturaleza diferente.
- Los efectos del CO2 son características fijas de toda la población de árboles que estamos tratando de conocer.
- Por el contrario, el efecto árbol variará aleatoriamente de un árbol a otro en la población.
- En esta circunstancia, tiene sentido modelar la distribución de los efectos de los árboles entre la población de árboles y suponer que los efectos de los árboles

particulares que ocurren en el experimento son sólo observaciones independientes de esta distribución.

- Es decir, el efecto del CO2 se modelará como un efecto fijo, pero el efecto del árbol se modelará como un efecto aleatorio:

- $y_i = \alpha_j + b_k + \epsilon_i$  (2)

- Si la observación  $i$  es para el nivel  $j$  del CO2, y el árbol  $k$ , donde ahora  $b_k \sim N(0, \sigma_b^2)$ ,  $\epsilon_i \sim N(0, \sigma^2)$ , y las  $b_k$  y  $\epsilon_i$  son v.a. mutuamente independientes.

- Ahora las pruebas de efectos de árbol pueden proceder exactamente como lo hicieron en el caso de efectos fijos, comparando los ajustes de mínimos cuadrados de los modelos con y sin efectos de árbol.

- Pero este modelo de efectos mixtos también nos permite probar los efectos del CO2, exista o no evidencia de un efecto árbol.

- Todo lo que se requiere es promediar los datos en cada nivel del efecto aleatorio, es decir, en cada árbol.

- Para datos balanceados, como los que tenemos aquí, la característica clave de un modelo de efectos mixtos es que este "promediado" de un efecto aleatorio implica automáticamente un modelo de efectos mixtos simplificado para los datos agregados: el efecto aleatorio se absorbe en el residuo independiente del término de error.

- Es fácil ver que el modelo para el área estomática promedio por árbol debe ser

- $\bar{y}_k = \alpha_j + e_k$ , (3)

- Si el árbol  $k$  está en el nivel  $j$  de CO2, y donde  $e_k \sim N(0, \sigma_b^2 + \sigma^2/4)$ .

- Hint:  $Var(1/4(\epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4))$

## Usando R

Ahora es sencillo probar el efecto del CO2 en R, primero agregue los datos de cada árbol:

```
In [68]: st <- aggregate(data.matrix(stomata), by=list(tree=stomata$tree), mean)
st$C02 <- as.factor(st$C02)
st
```

A data.frame: 6 × 4

tree	area	CO2	tree
<fct>	<dbl>	<fct>	<dbl>
1	1.623374	1	1
2	1.598643	1	2
3	1.162961	1	3
4	2.789238	2	4
5	2.903544	2	5
6	2.329761	2	6

y luego ajustar el modelo implícito en la agregación:

```
In [69]: m3 <- lm(area ~ CO2, st)
```

```
In [70]: anova(m3)
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
CO2	1	2.205314	2.20531395	27.68695	0.006246682
Residuals	4	0.318607	0.07965176	NA	NA

Aquí hay pruebas sólidas de un efecto del CO2, y ahora procederíamos a examinar la estimación de este efecto fijo

```
In [71]: summary(m3)
```

Call:

```
lm(formula = area ~ CO2, data = st)
```

Residuals:

1	2	3	4	5	6
0.1617	0.1370	-0.2987	0.1151	0.2294	-0.3444

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.4617	0.1629	8.970	0.000855 ***
CO2	1.2125	0.2304	5.262	0.006247 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2822 on 4 degrees of freedom

Multiple R-squared: 0.8738, Adjusted R-squared: 0.8422

F-statistic: 27.69 on 1 and 4 DF, p-value: 0.006247

**Generalmente con un modelo mixto las varianzas de los efectos aleatorios son de más interés que los efectos mismos, por lo que en este**

## ejemplo se debe estimar $\sigma_b^2$

Sea  $RSS_i$  la suma residual de cuadrados del modelo  $i$ .

De la teoría de modelos lineales tenemos que:

$$\hat{\sigma}^2 = RSS_2/18 \text{ (modelos sin promediar)}$$

$$\sigma_b^2 + \widehat{\sigma^2}/4 = RSS_3/4 \text{ (modelos promediando)}$$

Ambos estimadores son insesgados.

Por tanto, un estimador insesgado para  $\sigma_b^2$  es:

$$\hat{\sigma}_b^2 = \sigma_b^2 + \widehat{\sigma^2}/4 - \hat{\sigma}^2/4 = RSS_3/4 - RSS_2/72$$

```
In [72]: summary(m3)$sigma^2 - summary(m2)$sigma^2/4
```

```
0.0677017670742175
```

## Ejemplo guia 2: Modelo con un factor

- Consideremos ahora un ejemplo industrial práctico.
- Una prueba de ingeniería para la tensión longitudinal en rieles implica medir el tiempo que tardan ciertas ondas ultrasónicas en viajar a lo largo del riel.
- Para que sea una prueba útil, los ingenieros necesitan conocer el tiempo de viaje promedio de los rieles y la variabilidad esperada entre rieles, así como la variabilidad en el proceso de medición.
- El marco de datos Rail disponible con el paquete R `nlme` proporciona 3 mediciones del tiempo de viaje para cada uno de los 6 rieles elegidos al azar.
- Esto proporciona una aplicación obvia para el modelo (3).
- Primero examine los datos.

```
In [73]: library(nlme)
```

```
In [74]: data(Rail)
```

```
In [75]: Rail
```



A nffGroupedData:

18 × 2

	Rail	travel
	<ord>	<dbl>
1	1	55
2	1	53
3	1	54
4	2	26
5	2	37
6	2	32
7	3	78
8	3	91
9	3	85
10	4	92
11	4	100
12	4	96
13	5	49
14	5	51
15	5	50
16	6	80
17	6	85
18	6	83

Ahora ajuste el modelo (3) como un modelo de efectos fijos y utilice este modelo para probar

- $H_0 : \sigma_b^2 = 0$ ,

es decir, para probar la evidencia de diferencias entre rieles.

```
In [76]: m1 <- lm(travel ~ Rail,Rail)
summary(m1)
```

```
Call:
lm(formula = travel ~ Rail, data = Rail)

Residuals:
    Min       1Q   Median       3Q      Max
-6.6667 -1.0000  0.1667  1.0000  6.3333

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.5000     0.9477   70.169 < 2e-16 ***
Rail.L       54.3032     2.3214   23.392 2.22e-11 ***
Rail.Q      -4.6917     2.3214   -2.021 0.066161 .
Rail.C      -2.6584     2.3214   -1.145 0.274458
Rail^4      -0.5669     2.3214   -0.244 0.811181
Rail^5      11.1919     2.3214    4.821 0.000418 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.021 on 12 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9711
F-statistic: 115.2 on 5 and 12 DF,  p-value: 1.033e-09
```

```
In [77]: anova(m1)
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
<b>Rail</b>	5	9310.5	1862.10000	115.1814	1.032673e-09
<b>Residuals</b>	12	194.0	16.16667	NA	NA

Por lo tanto, hay pruebas sólidas para rechazar la hipótesis nula y aceptar como reales las diferencias entre ferrocarriles.

Como vimos teóricamente, hasta ahora el análisis no difiere del de un modelo de efectos fijos, pero estimar  $\sigma_b^2$  implica promediar en cada nivel del efecto aleatorio y ajustar el modelo (3):

$$\bar{y}_i = \alpha + e_i,$$

a los promedios resultantes.

Procedamos de manera similar al ejemplo guía 1

```
In [78]: rt <- aggregate(data.matrix(Rail),by=list(Rail$Rail),mean)
rt
```

A data.frame: 6 × 3

Group.1	Rail	travel
<ord>	<dbl>	<dbl>
2	1	31.66667
5	2	50.00000
1	3	54.00000
6	4	82.66667
3	5	84.66667
4	6	96.00000

Ahora es posible ajustar el modelo (3) y calcular  $\hat{\sigma}_b$ ,  $\hat{\sigma}$ :

```
In [79]: m0 <- lm(travel ~ 1, rt)
```

```
In [80]: model.matrix(m0)
```

A matrix: 6 × 1 of  
type dbl

(Intercept)	
1	1
2	1
3	1
4	1
5	1
6	1

```
In [81]: sig <- summary(m1)$sigma
```

```
In [82]: sigb <- (summary(m0)$sigma^2 - sig^2/3)^0.5
```

```
In [83]: sigb
```

24.8054653476025

```
In [84]: sig
```

4.02077936060494

- Por lo tanto, hay una cantidad bastante grande de variabilidad entre rieles, mientras que el error de medición es relativamente pequeño.
- En este caso, el intercepto del modelo,  $\alpha$ , se confunde con los efectos aleatorios,  $b_j$ , por lo que  $\alpha$  debe estimarse a partir del ajuste del modelo (3).

In [85]: `summary(m0)`

Call:

`lm(formula = travel ~ 1, data = rt)`

Residuals:

1	2	3	4	5	6
-34.83	-16.50	-12.50	16.17	18.17	29.50

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	66.50	10.17	6.538	0.00125 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

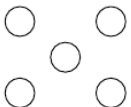
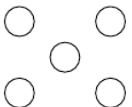
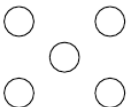
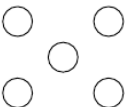
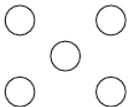
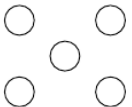
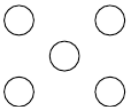
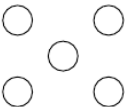
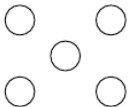
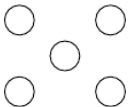
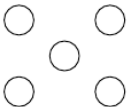
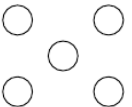
Residual standard error: 24.91 on 5 degrees of freedom

- La verificación del modelo se realiza observando las gráficas residuales, desde los ajustes hasta los datos originales y los datos agregados, ya que, aproximadamente, estos deberían verse como muestras de i.i.d. variables aleatorias normales.
- Sin embargo, tendría que haber una violación realmente grotesca del supuesto de normalidad para la  $b_j$  antes de poder detectarlo en el examen de 6 residuos (muestra muy pequeña).

## Ejemplo guía 3: Un modelo con dos factores

Consideremos ahora un experimento en el que cada observación se agrupa según dos factores.

En la figura se muestra un diagrama esquemático de dicho diseño.

		Factor A			
		1	2	3	4
Factor B	1				
	2				
	3				

Supongamos que un factor se va a modelar como un efecto fijo y el otro como un efecto aleatorio.

Entonces, un modelo para la  $k$ -ésima observación en el nivel  $i$  de efecto fijo  $A$  y el nivel  $j$  de efecto aleatorio  $B$  es

$$y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \epsilon_{ijk}, \quad (4)$$

donde:

$$b_j \sim N(0, \sigma_b^2),$$

$$(\alpha b)_{ij} \sim N(0, \sigma_{\alpha\beta}^2),$$

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

y todas las v.a. son mutuamente independientes.

Ademas,  $\mu$  es la media poblacional global,  $\alpha_i$  son los  $I$  efectos fijos para el factor  $A$ , y  $b_j$  representan los  $J$  efectos aleatorios para el factor  $B$ .

Por otro lado,  $(\alpha b)_{ij}$  son las  $IJ$  interacciones.

### Pruebas estadísticas del error

Probar  $H_0 : \sigma_{\alpha\beta}^2 = 0$  es equivalente a probar

$$H_0 : (\alpha b)_{ij} = 0, \forall ij,$$

en un marco de efectos fijos.

Por lo tanto, esta hipótesis puede probarse mediante la comparación habitual de la prueba de relación ANOVA/F de modelos con y sin términos de interacción.

Si  $RSS_1$  ahora denota la suma residual de cuadrados del ajuste (4) por mínimos cuadrados, entonces

$$\hat{\sigma}^2 = RSS_1 / (n - IJ)$$

$$\text{nota: } n - p = n - 1 - I - J - (I - 1)(J - 1)$$

En un contexto de efectos puramente fijos sólo tiene sentido probar los efectos principales si los términos de interacción no son significativos y, por tanto, pueden tratarse como cero.

En el caso de efectos mixtos, debido a que la interacción es un efecto aleatorio, es posible hacer inferencias sobre los efectos principales, independientemente de que los términos de la interacción sean significativos o no.

Esto se puede hacer promediando los  $K$  datos en cada nivel de la interacción.

El promediado, junto con el modelo (4), implica el siguiente modelo para los promedios:

$$\bar{y}_{ij} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \frac{1}{K} \sum_{k=1}^K \epsilon_{ijk}$$

Definiendo

$$e_{ij} = (\alpha b)_{ij} + \frac{1}{K} \sum_{k=1}^K \epsilon_{ijk}$$

Es claro ver  $e_{ij}$  es normal con media cero, puesto que cada elemento  $(i, j)$  es la suma de v.a. normales con media cero.

Ademas, por ser mutuamente independientes  $(\alpha b)_{ij}$  y  $\epsilon_{ijk}$

$$\text{var}(e_{ij}) = \sigma_{\alpha b}^2 + \sigma^2 / K$$

Por lo que el modelo simplificado es:

$$\bar{y}_{ij} = \mu + \alpha_i + b_j + e_{ij}, \quad e_{ij} \sim N(0, \sigma_{\alpha b}^2 + \sigma^2 / K), \quad (5)$$

La hipótesis nula  $H_0 : \alpha_i = 0, \forall i$ , se realiza comparando el modelo (5) con el modelo:

$$\bar{y}_{ij} = \mu + b_j + e_{ij}, \quad (6)$$

mediante el estadístico  $F$  (como es usual).

De manera equivalente, se puede probar  $H_0 : \sigma_b = 0$  a través de  $H_0 : b_j = 0, \forall j$ , mediante el estadístico F entre el modelo (5) y el modelo:

$$\bar{y}_{ij} = \mu + \alpha_i + e_{ij}, \quad (7)$$

La suma residual de cuadrados para el modelo (5), denotada como  $RSS_2$ , es útil para la estimación insesgada de la varianza de la interacción:

$$\hat{\sigma}_{\alpha b}^2 + \hat{\sigma}^2/K = RSS_2/(IJ - I - J + 1)$$

$$\text{nota: } n - p = IJ - 1 - (I - 1) - (J - 1)$$

Por lo tanto

$$\hat{\sigma}_{\alpha b}^2 = RSS_2/(IJ - I - J + 1) - \hat{\sigma}^2/K$$

Promediando nuevamente, pero ahora sobre los niveles del factor  $B$ :

$$\bar{y}_j = \mu + \frac{1}{I} \sum_{i=1}^I \alpha_i + b_j + \frac{1}{I} \sum_{i=1}^I e_{ij}$$

Definiendo

$$\tilde{\mu} = \mu + \frac{1}{I} \sum_{i=1}^I \alpha_i$$

$$e_j = b_j + \frac{1}{I} \sum_{i=1}^I e_{ij}$$

El modelo se convierte en:

$$\bar{y}_j = \tilde{\mu} + e_j, \quad e_{ij} \sim N(0, \sigma_b^2 + \sigma_{\alpha b}^2/I + \sigma^2/(IK)) \quad (8)$$

Entonces, definiendo  $RSS_3$  como la suma residual de cuadrados del modelo (8), un estimador insesgado de  $\sigma_b^2$  está dado por

$$\hat{\sigma}_b^2 = RSS_3/(J - 1) - \hat{\sigma}_{\alpha b}^2/I + \sigma^2/(IK)$$

## Datos

- Consideremos ahora un ejemplo práctico.
- El marco de datos `Machines`, del paquete `nlme`, contiene datos de un experimento industrial que compara 3 tipos de máquinas diferentes.
- El objetivo del experimento es determinar qué tipo de máquina daba como resultado la mayor productividad de los trabajadores.
- Se seleccionaron al azar 6 trabajadores para participar en la prueba, y cada trabajador operó cada máquina 3 veces (presumiblemente después de un período apropiado de

capacitación diseñado para eliminar cualquier "efecto de aprendizaje").

- A continuación se produce el gráfico de resultados:

```
In [86]: library(nlme)
```

```
In [87]: data(Machines)
```

```
In [88]: Machines
```



A nffGroupedData: 54 × 3

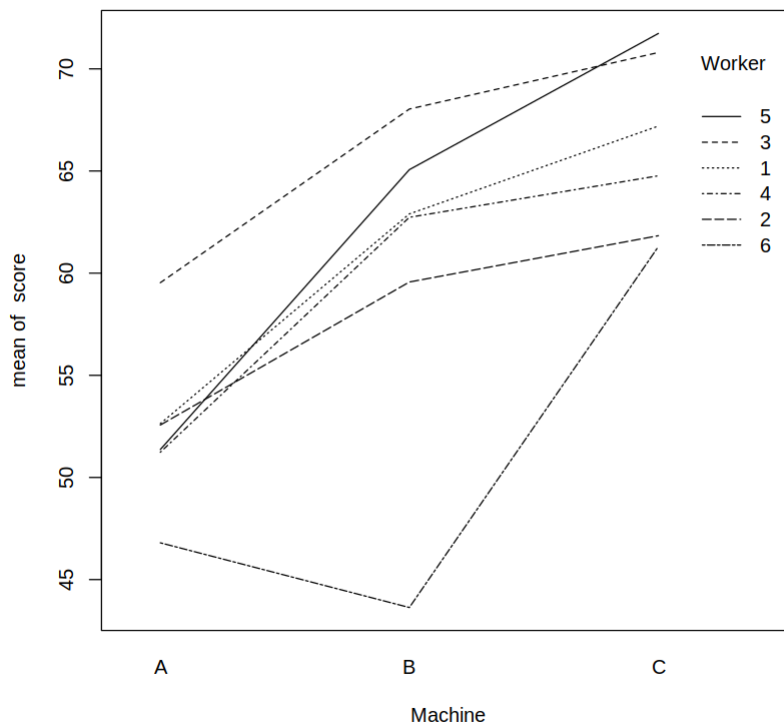
	Worker	Machine	score
	<ord>	<fct>	<dbl>
1	1	A	52.0
2	1	A	52.8
3	1	A	53.1
4	2	A	51.8
5	2	A	52.8
6	2	A	53.1
7	3	A	60.0
8	3	A	60.2
9	3	A	58.4
10	4	A	51.1
11	4	A	52.3
12	4	A	50.3
13	5	A	50.9
14	5	A	51.8
15	5	A	51.4
16	6	A	46.4
17	6	A	44.8
18	6	A	49.2
19	1	B	62.1
20	1	B	62.6
21	1	B	64.0
22	2	B	59.7
23	2	B	60.0
24	2	B	59.0
25	3	B	68.6
26	3	B	65.8
27	3	B	69.7
28	4	B	63.2
29	4	B	62.8
30	4	B	62.2
31	5	B	64.8
32	5	B	65.0

	Worker	Machine	score
	<ord>	<fct>	<dbl>
33	5	B	65.4
34	6	B	43.7
35	6	B	44.2
36	6	B	43.0
37	1	C	67.5
38	1	C	67.2
39	1	C	66.9
40	2	C	61.5
41	2	C	61.7
42	2	C	62.3
43	3	C	70.8
44	3	C	70.6
45	3	C	71.0
46	4	C	64.1
47	4	C	66.2
48	4	C	64.0
49	5	C	72.1
50	5	C	72.0
51	5	C	71.1
52	6	C	62.0
53	6	C	61.4
54	6	C	60.5

```
In [111... names(Machines)
```

```
'Worker' · 'Machine' · 'score'
```

```
In [114... interaction.plot(Machine,Worker,score)
```



- Si el experimento se repitiera en otro lugar (con diferentes trabajadores), esperaríamos que las estimaciones de los efectos de la máquina fueran bastante cercanas a los resultados obtenidos en el experimento actual, mientras que los efectos de los trabajadores individuales serían bastante diferentes (aunque esperamos que tenga una variabilidad similar).
- Por lo tanto, el modelo (4) es apropiado, donde  $\alpha_i$  representa los efectos fijos de la máquina,  $b_j$  representa los efectos aleatorios del trabajador y  $(\alpha b)_{ij}$  representa la interacción trabajador-máquina (es decir, el hecho de que diferentes trabajadores pueden trabajar mejor en diferentes máquinas).
- Ajustando el modelo completo, podemos probar inmediatamente
- $H_0 : \sigma_{\alpha\beta}^2 = 0$

```
In [97]: m0 <- lm(score ~ Worker + Machine, Machines)
         m1 <- lm(score ~ Worker*Machine, Machines)
```

```
In [98]: anova(m0, m1)
```

A anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	46	459.81667	NA	NA	NA	NA
2	36	33.28667	10	426.53	46.12982	1.64125e-17

Debemos aceptar  $H_1 : \sigma_{\alpha\beta}^2 \neq 0$

Existe evidencia muy sólida de una interacción entre la máquina y el trabajador.

Ahora se puede estimar  $\sigma^2$ :

```
In [99]: summary(m1)$sigma^2
```

0.92462962962963

Para examinar los principales efectos podemos agregar en cada nivel de la interacción

```
In [101]: Mach <- aggregate(data.matrix(Machines),by= list(Machines$Worker,Machines$Machine),FUN= function(x){
Mach$Worker <- as.factor(Mach$Worker)
Mach$Machine <- as.factor(Mach$Machine)
})
```

y ajustar el modelo

$$\bar{y}_{ij} + \mu + \alpha_i + b_j + \epsilon_{ij}$$

a los datos resultantes.

```
In [102]: m0 <- lm(score ~ Worker + Machine,Mach)
```

```
In [103]: anova(m0)
```

A anova: 3 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Worker	5	413.9650	82.79300	5.823248	0.0089494552
Machine	2	585.0878	292.54389	20.576083	0.0002855485
Residuals	10	142.1767	14.21767	NA	NA

Los valores p muy bajos indican nuevamente que  $H_0 : \sigma_b^2 = 0$  y  $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$  deben rechazarse en favor de las alternativas obvias.

Hay pruebas sólidas de diferencias entre tipos de máquinas y de variabilidad entre trabajadores.

Al examinar las estimaciones de efectos fijos, utilizando métodos estándar de efectos fijos, se indica que la máquina C conduce a un aumento sustancial de la productividad.

```
In [104... summary(m0)
```

Call:

```
lm(formula = score ~ Worker + Machine, data = Mach)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-7.6167	-1.3375	0.8056	1.8222	4.1000

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	43.283	2.514	17.219	9.23e-09	***
Worker2	7.411	3.079	2.407	0.03686	*
Worker3	9.000	3.079	2.923	0.01521	*
Worker4	10.333	3.079	3.356	0.00729	**
Worker5	15.544	3.079	5.049	0.00050	***
Worker6	12.144	3.079	3.945	0.00275	**
Machine2	7.967	2.177	3.660	0.00439	**
Machine3	13.917	2.177	6.393	7.91e-05	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.771 on 10 degrees of freedom

Multiple R-squared: 0.8754, Adjusted R-squared: 0.7882

F-statistic: 10.04 on 7 and 10 DF, p-value: 0.0008036

La estimación de  $\sigma_{\alpha b}^2$ , la varianza de la interacción, es sencilla:

```
In [105... summary(m0)$sigma^2 - (summary(m1)$sigma^2)/3
```

```
13.9094567901235
```

Agregando una vez más, podemos estimar el componente de la varianza de los trabajadores,  $\sigma_b^2$ :

```
In [106... M <- aggregate(data.matrix(Mach), by=list(Mach$Worker), mean)
```

```
In [107... m00 <- lm(score ~ 1, M)
```

```
In [108... summary(m00)$sigma^2 - (summary(m0)$sigma^2)/3
```

```
22.8584444444443
```

## Ejemplo guía 4: Usando librerías especializadas

librería nlme

Existen varios paquetes para modelado lineal mixto en R, de los cuales `nlme` y `lme4` son particularmente dignos de mención.

La principal función de interés de ajuste del modelo se llama `lme`.

Una llamada a la función `lme` es similar a una llamada a `lm`, excepto que también se debe proporcionar al modelo un argumento adicional que especifique la estructura de efectos aleatorios.

Específicamente, `lme` supone que sus datos están agrupados de acuerdo con los niveles de algunos factores y que se requiere la misma estructura de efectos aleatorios para cada grupo, con efectos aleatorios independientes entre grupos.

Suponiendo solo un nivel de agrupación, el modelo para los datos en el  $i$ -ésimo grupo es entonces:

$$y_i = X_i\beta + Z_ib_i + \epsilon_i,$$

$$b_i \sim N(0, \Sigma_\theta),$$

$$\epsilon_i \sim N(0, \Sigma_\theta\sigma^2)$$

Una llamada de ejemplo a `lme` se parece a esto:

```
lme(y ~ x + z, dat, ~ x | g)
```

donde la respuesta es `y`, los efectos fijos dependen de `x` y `z`, los efectos aleatorios dependen solo de `x`, los datos se agrupan según el factor `g` y todos los datos están en el marco de datos `dat`.

Una forma alternativa de especificar el mismo modelo es:

```
lme(y ~ x + z, dat, list(g = ~ x))
```

El ejemplo 2 se puede ajustar a través de `lme` con la siguiente sintaxis:

```
In [109... library(nlme)
```

```
In [110... lme(travel ~ l, Rail, list(Rail = ~ l))
```

Linear mixed-effects model fit by REML

Data: Rail

Log-restricted-likelihood: -61.0885

Fixed: travel ~ 1

(Intercept)

66.5

Random effects:

Formula: ~1 | Rail

(Intercept) Residual

StdDev: 24.80547 4.020779

Number of Observations: 18

Number of Groups: 6