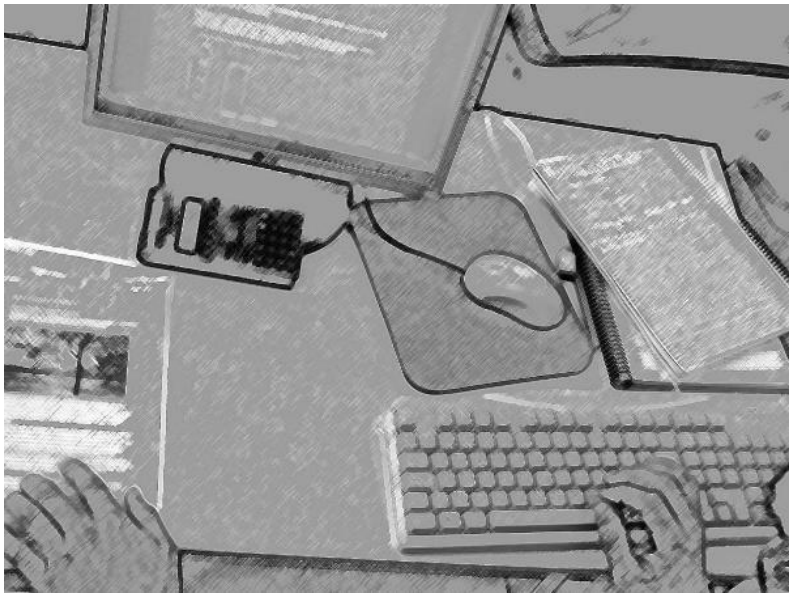
Estadística y modelos lineales usando R



Clase 4.2
Modelos mixtos

Adriana Pérez
Grupo de Bioestadística Aplicada
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Efecto de la fertilización y el tratamiento del suelo sobre *Pinus taeda*

En la provincia de Corrientes se efectuó un ensayo en plantaciones de *Pinus taeda* para comparar el efecto de distintas dosis de fertilización con superfosfato triple:

I. Sin fertilización

II. 100 g/planta de SFT

III. 200 g/planta de SFT

utilizando dos métodos de tratamiento del suelo:

Descompactado hasta los 20 cm de profundidad (método tradicional)

Descompactado hasta los 50 cm de profundidad

Se dispone de un campo en el SE de Corrientes, en el que se definen 30 parcelas conteniendo 10 plantines cada una. Se deciden asignar 5 parcelas a cada tratamiento. Se midió la altura total al año de cada planta.

pinus.csv

Variaciones en rasgos del cedro amargo

- 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

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

```
lmer(largo ~ 1 + (1 | poblacion/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?

Control de maíz resistente a glifosato

- Se denomina maíz “guacho” a aquel que brota en los campos luego de la cosecha, como resultado de pérdidas en la recolección del grano. Este maíz es indeseable y considerado maleza, ya que estará presente en el barbecho o incluso en el cultivo de verano posterior. El problema se agrava en el caso de cultivos transgénicos, capaces de soportar herbicidas de amplio espectro.
- Se planea un ensayo a fin de detectar potenciales herbicidas para la erradicación del maíz “guacho” transgénico tolerante al glifosato.
- Los tratamientos herbicidas que se planea ensayar son: I) Cletodim 0.6 l/ha, II) Sal triazolamina 2 l/ha, III) Haloxifop 0.5 l/ha, IV) Testigo sin tratar.
- El ensayo se llevará a cabo en 9 localidades. En cada localidad se delimitarán 3 bloques y en cada bloque se aplicarán los 4 tratamientos
- Se medirá el rendimiento (en ton/ha)

localidades.csv

Variaciones en rasgos del cedro amargo

- ¿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

- ¿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)
```

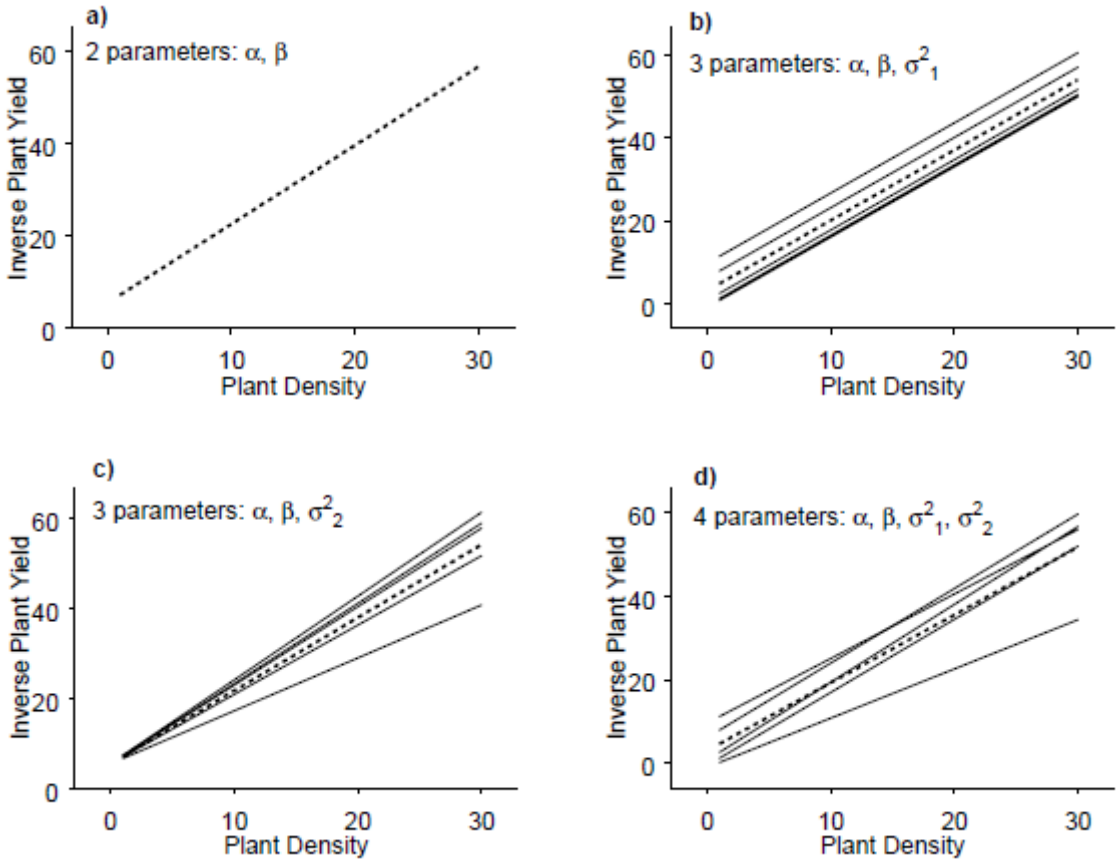
- ¿Y si de cada ejemplar se registró el pH del suelo e interesa saber si el largo del ejemplar se asocia con el pH?

```
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?

http://mfviz.com/hierarchical-models/

Modelos lineales mixtos



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 + (|Factor_aleatorio), data)
d <-lmer (Y ~ X + (1+X|Factor_aleatorio),data)
```

Modelos con intercepto y pendiente aleatorios

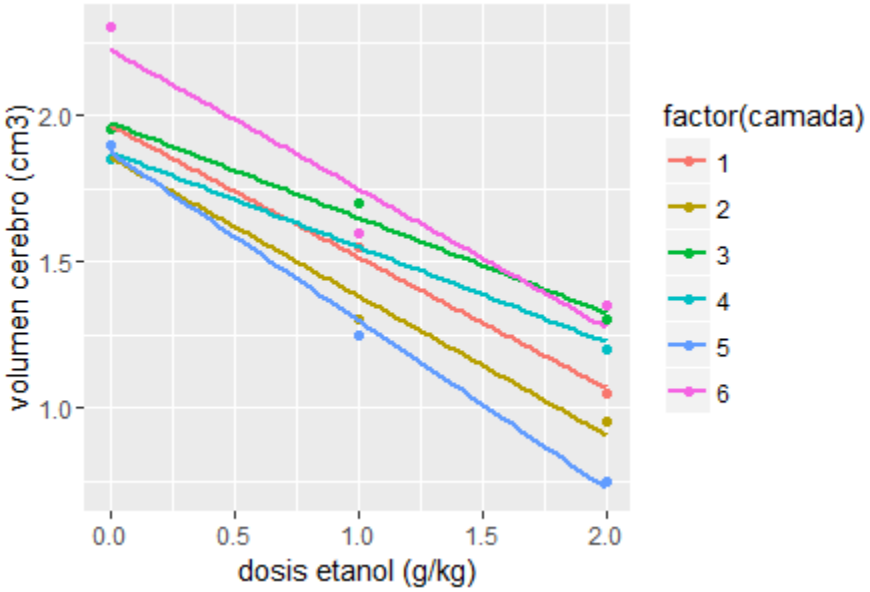
- ¿Y si se sospecha que el “efecto” del pH sobre el largo del ejemplar entre poblaciones?

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

Implica una interacción trans-nivel

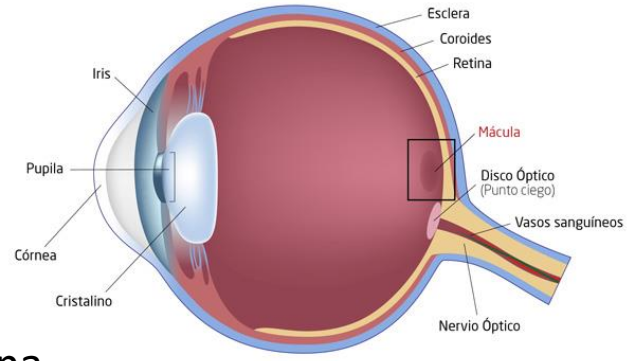
En el ej de DBA:

Implica interacción tratamiento x bloque



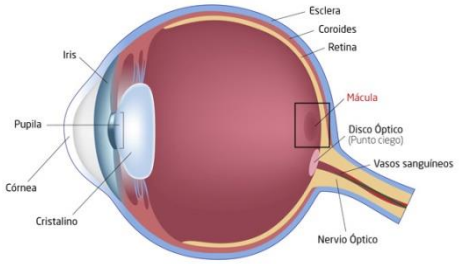
DISEÑO DE MEDIDAS REPETIDAS

Tratamiento de la degeneración macular



- La degeneración macular asociada a la edad (DMAE) es una enfermedad ocasionada por daño o deterioro por envejecimiento de la mácula, un área pequeña en la retina responsable de la visión central
- Las causas de la degeneración macular incluyen la acumulación de depósitos así como el crecimiento de vasos sanguíneos anormales por debajo de la retina.
- Se llevó a cabo un ensayo clínico en Individuos con diagnóstico de DMAE, con el objetivo de evaluar el efecto de un nuevo fármaco que prevendría la formación de vasos sanguíneos.
- Para ello se seleccionaron 240 Individuos que fueron aleatorizados en dos grupos balanceados. Al primer grupo se le suministró el nuevo fármaco, mientras que al segundo grupo el fármaco tradicional.
- Se midió el área de la lesión macular (en mm^2) antes del comienzo de la experiencia (basal) y luego a los 3, 6 y 9 meses de comenzado el estudio.

Tratamiento de la degeneración macular



- UE
- VR
- Tipo, potencial distribución de probabilidades
- VE

- Modelo

macula.csv

Diseño de medidas repetidas

- 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. Eso se hace mediante distintos modelos para la matriz de covarianza

¿Cómo modelamos datos correlacionados?

- Modelos Marginales (efectos fijos + estructura de correlación residual)

Las mediciones de Area 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

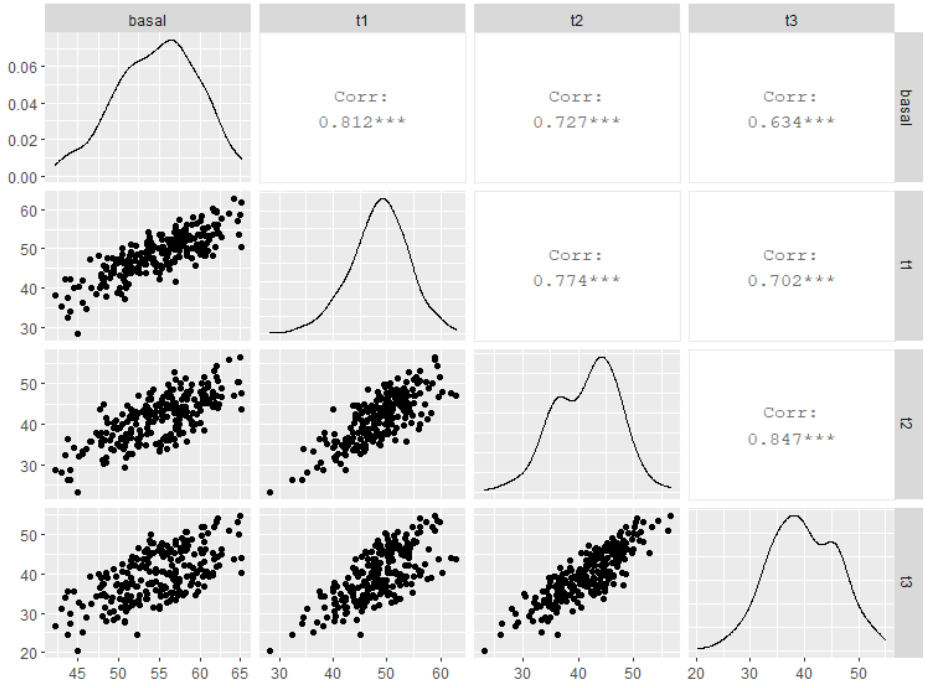
- Modelos Condicionales, sujeto-específicos (efectos fijos + efectos aleatorios)

Las mediciones de Area 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

Asociación en medidas repetidas

Podemos pensarlos como distintas variables

	Individuo	tratamiento	basal	t1	t2	t3
1	1	Tradicional	47.70806	40.03951	43.70996	36.39499
2	2	Tradicional	53.52273	50.96614	43.59097	46.26738
3	3	Tradicional	53.31746	43.76525	44.59296	41.47438
4	4	Tradicional	47.80227	40.62783	35.77816	36.76756
5	5	Tradicional	51.11634	47.03972	44.51514	40.16199
6	6	Tradicional	49.61367	47.34000	39.20886	39.20688
7	7	Tradicional	55.28245	53.72793	45.16866	47.74754
8	8	Tradicional	56.30818	44.71728	44.06577	40.13363
9	9	Tradicional	60.27624	50.70777	48.04782	45.19015
10	10	Tradicional	51.31588	48.24014	41.68702	41.74201
11	11	Tradicional	64.96432	58.81887	56.60106	54.89153
12	12	Tradicional	60.44182	49.63510	46.02121	45.22272
13	13	Tradicional	43.72147	37.35296	36.32079	35.65234
14	14	Tradicional	51.03838	44.67703	44.68511	36.52921
15	15	Tradicional	56.45349	52.22815	46.88627	46.34510
16	16	Tradicional	52.04741	50.84277	43.17814	44.17750



Formato “wide” vs “long”

Matriz de covarianza Σ

	Individuo	tratamiento	basal	t1	t2	t3
1	1	Tradicional	47.70806	40.03951	43.70996	36.39499
2	2	Tradicional	53.52273	50.96514	43.59097	46.26738
3	3	Tradicional	53.31746	43.76525	44.59296	41.47438
4	4	Tradicional	47.80227	40.62783	35.77816	36.76756
5	5	Tradicional	51.11634	47.03972	44.51514	40.16199
6	6	Tradicional	49.61367	47.34000	39.20886	39.20688
7	7	Tradicional	55.28245	53.72793	45.16866	47.74754
8	8	Tradicional	56.30818	44.71728	44.06577	40.13363
9	9	Tradicional	60.27624	50.70777	48.04782	45.19015

$$\Sigma = \begin{array}{c} \\ T1 \\ T2 \\ T3 \\ T4 \end{array} \begin{array}{c} \begin{array}{cccc} T1 & T2 & T3 & T4 \end{array} \\ \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix} \end{array}$$

- ✓ en la diagonal principal, las varianzas de cada variable σ^2_i . En el resto, las covarianzas σ_{ij} entre pares de variables
- ✓ Matriz cuadrada y simétrica ($\sigma_{12} = \sigma_{21}$)
- ✓ Matriz no estandarizada, tiene unidades
- ✓ $\sum$: matriz poblacional, S: matriz muestral

Estructura de la matriz de covarianza para medidas repetidas

- Simple
- Simetría compuesta `corCompSymm`
- Autoregresiva de orden 1 (AR1) `corAR1`
- Desestructurada `corSymm`
- Autoregresiva continua de orden 1 o Interdependencia de primer orden `corCAR1`

Requieren tiempos igualmente espaciados

- Se pueden combinar con varianzas heterogéneas

Estructura simple de la matriz de covarianza Σ

- 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.

$$\begin{array}{c} \\ T1 \\ T2 \\ T3 \\ T4 \end{array} \begin{array}{c} \begin{array}{cccc} T1 & T2 & T3 & T4 \end{array} \\ \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \end{array}$$

Suponiendo homocedasticidad

Más parsimoniosa; más restringida

$$\begin{array}{c} \\ T1 \\ T2 \\ T3 \\ T4 \end{array} \begin{array}{c} \begin{array}{cccc} T1 & T2 & T3 & T4 \end{array} \\ \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix} \end{array}$$

No suponiendo homocedasticidad (varident)

de parámetros?

Estructura de simetría compuesta

- 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:

$$\sigma_{Y_1Y_2} = \rho_{Y_1Y_2}\sigma_{Y_1}\sigma_{Y_2} = \rho\sigma^2$$

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

Matriz de simetría compuesta
`corCompSymm`

- Asume igual correlación entre cualquier par de MR
- Poco realista en DMR: las observaciones adyacentes estarán más fuertemente asociadas que las más alejadas en el tiempo. Puede funcionar para tiempos cortos.

de parámetros?

Estructura autoregresiva de primer orden

- Supongamos que la correlación entre tiempos disminuye exponencialmente según la distancia entre tiempos $\triangle t$: $\rho_{t_i, t_{i+\Delta t}} = \rho^{\Delta t}$
- Supongamos que la correlación entre las observaciones de dos tiempos con la misma diferencia de tiempo es siempre la misma, ρ

$$\sigma^2 \begin{array}{c} \begin{matrix} T1 & T2 & T3 & T4 \end{matrix} \\ \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \end{array}$$

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

- No hay restricciones sobre los parámetros de la matriz
- Es la menos parsimoniosa, con menores restricciones (mayor cantidad de parámetros)

$$
\begin{array}{c} \\ T1 \\ T2 \\ T3 \\ T4 \end{array}
\begin{array}{c}
\begin{array}{cccc} T1 & T2 & T3 & T4 \end{array} \\
\begin{bmatrix}
\sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\
\sigma_{12} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{43} \\
\sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2
\end{bmatrix}
\end{array}
$$

Matriz de correlación desestructurada
`corSymm`

Modelos marginales
Modelamos la estructura de covarianza

- 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

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

$$i = 1 a\ 3,\ j = 1\ a\ 9, k = 1\ a\ 72$$

$$\varepsilon_{ijk} \approx N(0, \Sigma_k)$$

- ✓ 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
- ✓ ε_{ijk} es el error aleatorio

Modelos marginales

```
library(nlme)
gls
```

#Modelo 1: Simetría compuesta.

```
m1<-gls(Area ~Tratamiento*tiempo,  correlation = corCompSymm(form
= ~ 1 | Individuo), bd)
```

#Modelo 2: Simetría compuesta. varianzas distintas

```
m2<-gls(Area ~Tratamiento*tiempo,  correlation = corCompSymm(form
= ~ 1 | Individuo), bd, weights=varIdent(form= ~ 1|tiempo ))
```

#Modelo 3: AR1, varianzas iguales

```
m3<-gls(Area ~Tratamiento*tiempo,  correlation = corAR1(form = ~ 1
| Individuo), bd)
```

#Modelo 4: AR1, varianzas distintas

```
m4<-gls(Area ~Tratamiento*tiempo,  correlation = corAR1(form = ~ 1
| Individuo), bd, weights=varIdent(form= ~ 1|tiempo ))
```

#Modelo 5: matriz desestructurada

```
m5<-gls(Area ~Tratamiento*tiempo,  correlation = corSymm(form = ~
1 | Individuo), bd)
```

- ¿Son paralelos los perfiles de respuesta en los grupos?
- El nuevo tratamiento ¿es más efectivo que el método tradicional?
- ¿Cuál es la magnitud del efecto del nuevo tratamiento comparado con el tradicional a los 9 meses (t3) de iniciado el tratamiento?