

Model mieszany w analizie powtarzalnych pomiarów (longitudinal data)

Dane zawierające pomiary powtarzalne w czasie nazywa się często **danymi panelowymi**. Nazwa ta bierze się stąd, że grupa osób, które poddane są powtarzalnym pomiarom, nazywana jest **panelem**. Rozważmy na początek przypadek, w którym nie mamy zmiennej grupującej obserwacje, ale z sytuacją, gdy dla każdej jednostki eksperymentalnej rejestrujemy serię pomiarów pewnej cechy. Interesuje nas wnioskowanie o zmianach wartości obserwowanej zmiennej w czasie. Część pomiarów dotyczy tego samego obiektu (jednostki eksperymentalnej), więc nie możemy traktować wszystkich pomiarów jako obserwacji zmiennych niezależnych. Aby rozróżnić pomiary zależne i niezależne uznajemy identyfikator jednostki eksperymentalnej jako zmienną grupującą. Dla danego osobnika obserwujemy więc serię czasową pomiarów. Pomiary mogą być wykonywane w tych samych punktach czasowych (**pomiary synchroniczne**) lub w różnych punktach (**pomiary asynchroniczne**). Przyjmijmy pewną postać dynamiki zmian interesujących nas pomiarów. Możemy założyć, że wartości danej cechy zmieniają się w czasie w sposób liniowy, kwadratowy lub inny funkcyjny. Dla prostoty założymy dla obiektu i liniową zmienność badanej cechy w czasie:

$$y_{ij} = a_i + b_i t_{ij} + \varepsilon_{ij}, \quad \text{gdzie } \varepsilon_{ij} \sim N(0, \sigma_0^2).$$

Jeżeli wartości a_i, b_i są stałe dla populacji, to zostałyby uwzględnione w modelu jako efekty stałe. Jeżeli spodziewamy się, że zależność liniowa może mieć różny charakter dla różnych osobników, efekty te możemy opisać jako efekty losowe. Załóżmy, że

$$a_i \sim N(a, \sigma_a^2) \quad \text{i} \quad b_i \sim N(b, \sigma_b^2).$$

Tak więc współczynniki opisujące zmianę w czasie cechy y są superpozycją pewnych stałych niezmiennych dla populacji a i b oraz osobniczych odchyłeń $u_{1i} = a_i - a$, $u_{2i} = b_i - b$.

Ogólnie model mieszany zapisujemy w postaci $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$. W przypadku powtarzalnych pomiarów wektor \mathbf{y} składa się z pomiarów badanej cechy y_{ij} dla każdego osobnika w każdej chwili czasu. Macierz efektów stałych można rozłożyć na 3 macierze $\mathbf{X} = [\mathbf{X}_a, \mathbf{X}_b, \mathbf{X}_c]$. Macierz \mathbf{X}_a ma jedną kolumnę wypełnioną jedynekami, macierz \mathbf{X}_b ma jedną kolumnę w której dla obserwacji y_{ij} wpisany jest czas t_{ij} , w którym dana obserwacja została zmierzona. Macierz \mathbf{X}_c koduje wszystkie pozostałe efekty stałe. Podobnie wektor $\boldsymbol{\beta}$ można przedstawić jako $\boldsymbol{\beta} = [a, b, \boldsymbol{\beta}_c]$. **Przyjmijmy osobnika (jednostkę eksperymentalną) jako zmienną grupującą o liczbie poziomów (powiedzmy k) równą liczbie osobników.** Ponieważ w modelu mamy 2 efekty losowe: jeden opisujący osobnicze odstępstwo od średniej a , oraz drugi opisujący osobnicze odstępstwo od współczynnika liniowego trendu b , macierz \mathbf{Z} będzie się składała z $q=2k$ kolumn. Bez straty ogólności możemy założyć, że

początkowe k kolumn macierzy koduje efekty $u_{1i} = a_i - a$ a kolejne k kolumn efekty $u_{2i} = b_i - b$.

Używając powyższych oznaczeń dla dwóch komponentów model mieszany można zapisać w postaci

$$\mathbf{y} = \mathbf{X}_a \mathbf{a} + \mathbf{X}_b \mathbf{b} + \mathbf{X}_c \boldsymbol{\beta}_c + \mathbf{Z}_1 \mathbf{u}_1 + \mathbf{Z}_2 \mathbf{u}_2 + \boldsymbol{\varepsilon}.$$

Jak zwykle przyjmiemy, że $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I})$ natomiast o wektorach \mathbf{u}_1 i \mathbf{u}_2 założymy, że są realizacjami dwuwymiarowej zmiennej losowej (u_1, u_2) o rozkładzie normalnym. Załóżmy na początek, że zmienne (u_1, u_2) są niezależne, więc macierz wariancji wektora $(\mathbf{u}_1, \mathbf{u}_2)$ to macierz diagonalna

$$\mathbf{D} = \text{diag}(\underbrace{\sigma_a^2, \dots, \sigma_a^2}_k, \underbrace{\sigma_b^2, \dots, \sigma_b^2}_k), \quad .$$

Macierz wariancji wektora obserwacji jest postaci

$$V(\mathbf{y}) = \mathbf{V} = \mathbf{V}_0 + \mathbf{V}_1 + \mathbf{V}_2 = \sigma_0^2 \mathbf{I} + \sigma_a^2 \mathbf{Z}_1 \mathbf{Z}_1^T + \sigma_b^2 \mathbf{Z}_2 \mathbf{Z}_2^T.$$

Dla przykładu przyjmijmy, że mamy zbiór obserwacji dla trzech pacjentów. Dla pierwszego pacjenta wykonaliśmy pomiary w chwilach czasowych $t_1=0,2,4$, dla drugiego pacjenta w chwilach $t_2=0,2$, a dla trzeciego pacjenta tylko w chwili $t_3=0$. Zakładając powyższą kolejność obserwacji w wektorze odpowiedzi tzn.(najpierw kolejne obserwacje dla pierwszego pacjenta potem drugiego itd.) macierze \mathbf{Z}_1 i \mathbf{Z}_2 mają postać:

$$\mathbf{Z}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Z}_2 = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 0 & 0 \\ 4 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Natomiast macierze

$$\mathbf{V}_1 = \sigma_a^2 \mathbf{Z}_1 \mathbf{Z}_1^T = \begin{bmatrix} \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 & \sigma_a^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & 0 \\ 0 & 0 & 0 & \sigma_a^2 & \sigma_a^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_a^2 \end{bmatrix}$$

$$\mathbf{V}_2 = \sigma_b^2 \mathbf{Z}_2 \mathbf{Z}_2^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4\sigma_b^2 & 8\sigma_b^2 & 0 & 0 & 0 \\ 0 & 8\sigma_b^2 & 16\sigma_b^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4\sigma_b^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

mają klatkową strukturę.

Zobaczmy jak powyższy model zapisać w pakiecie R - [plik panel.R](#)

```
id<-1:3
y0<-c(10,10,10)+rnorm(3)
y2<-c(c(11,11)+rnorm(2),NA)
y4<-c(c(12)+rnorm(1),NA,NA)
dane<-data.frame(id,y0,y2,y4)#tworzmy ramkę danych w postaci szerokiej
  id   y0   y2   y4
1  1  8.640037 10.91195 13.052
2  2  7.505190 10.90964   NA
3  3 11.119275   NA   NA

summary(dane)
#przekształcanie formatu danych z szerokiego na wąski
#służy do tego funkcja reshape()
#zmienne 2:4 zostaną przekształcone w nową zmienną o nazwie Y i zostanie utworzona zmienna time
#z opisem chwil czasowych
long_dane=reshape(dane,direction="long",varying=list(2:4),v.names=c("Y"),times=c(0,2,4))
  id   time   Y
1.0  1     0  8.640037
2.0  2     0  7.505190
3.0  3     0 11.119275
1.2  1     2 10.911947
2.2  2     2 10.909636
3.2  3     2   NA
1.4  1     4 13.051999
2.4  2     4   NA
3.4  3     4   NA

#usuwamy wiersze z brakującymi danymi
dane1=na.omit(long_dane)
  id   time   Y
1.0  1     0  8.640037
2.0  2     0  7.505190
3.0  3     0 11.119275
1.2  1     2 10.911947
2.2  2     2 10.909636
1.4  1     4 13.051999
```

Model z jednym efektem losowym, efektem pacjenta

$$y = \beta_0 + Z_{id}u_{id} + \varepsilon \quad u_{id} \sim N(\mathbf{0}, \sigma_{id}^2 \mathbf{I}_3), \quad \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_7)$$

#model z jednym efektem losowym, efektem pacjenta

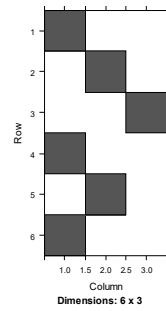
```
library(lme4)
model0=lmer(Y~1+(1|id),data=dane1,REML=FALSE)
summary(model0)
```

model0@pp\$Zt # transponowana macierz Z

t(model0@pp\$Zt) # macierz Z

```
6 x 3 sparse Matrix of class "dgCMatrix"
  1 2 3
1.0 1 . .
2.0 . 1 .
3.0 . . 1
1.2 1 . .
2.2 . 1 .
```

```
1.4      1 ..
image(getME(model0,"Z"))
```

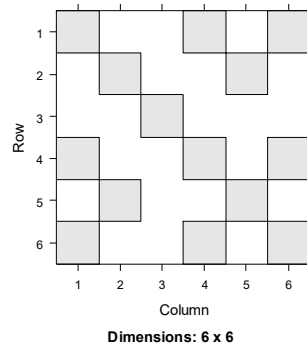


Widać że najpierw są przedstawione dane dla kolejnych pacjentów w chwili 0 potem z zachowaniem kolejności w chwili 2 i w końcu w chwili 4.

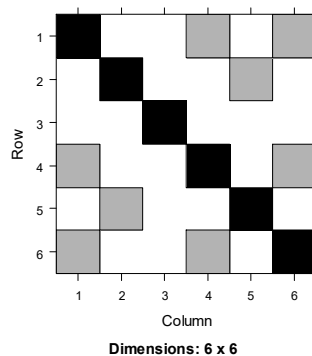
```
t(model0@pp$Zt)%*%model0@pp$Zt # macierz Z*t(Z)
```

```
6 x 6 sparse Matrix of class "dgCMatrix"
  1.0 2.0 3.0 1.2 2.2 1.4
1.0  1 . . 1 . 1
2.0 . 1 . . 1 .
3.0 . . 1 . . .
1.2 1 . . 1 . 1
2.2 . 1 . . 1 .
1.4 1 . . 1 . 1
```

```
image(t(model0@pp$Zt)%*%model0@pp$Zt) #graficzny obraz macierzy Z*t(Z)
```



```
image(diag(6)+t(model0@pp$Zt)%*%model0@pp$Zt) #graficzny obraz macierzy I+Z*t(Z)
```



Rozważamy więc model

$$\mathbf{y} = \mathbf{X}_a \mathbf{a} + \mathbf{Z}_1 \mathbf{u}_1 + \boldsymbol{\varepsilon} \text{ z macierzami } \mathbf{Z} = \mathbf{Z}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad V(\mathbf{u}_1) = \mathbf{D} = \begin{bmatrix} \sigma_a^2 & 0 & 0 \\ 0 & \sigma_a^2 & 0 \\ 0 & 0 & \sigma_a^2 \end{bmatrix}.$$

$$V(\mathbf{y}) = \sigma_0^2 \mathbf{I} + \mathbf{Z}^T \mathbf{D} \mathbf{Z} = \begin{bmatrix} \sigma_0^2 + \sigma_a^2 & 0 & 0 & \sigma_a^2 & 0 & \sigma_a^2 \\ 0 & \sigma_0^2 + \sigma_a^2 & 0 & 0 & \sigma_a^2 & 0 \\ 0 & 0 & \sigma_0^2 + \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & 0 & 0 & \sigma_0^2 + \sigma_a^2 & 0 & \sigma_a^2 \\ 0 & \sigma_a^2 & 0 & 0 & \sigma_0^2 + \sigma_a^2 & 0 \\ \sigma_a^2 & 0 & 0 & \sigma_a^2 & 0 & \sigma_0^2 + \sigma_a^2 \end{bmatrix}$$

```
model0@pp$Xwts # transponowana macierz X
[1] 1 1 1 1 1 1
```

Aby zobaczyć jak zapisywane są dane przypadku przyjęcia modelu liniowego wzrostu w czasie musimy zwiększyć liczbę danych aby mieć przynajmniej po jednej obserwacji dla rozważnych chwil czasowych.

```
id<-1:3
y0n<-c(10,10,10)+rnorm(3)
y2n<-c(11,11,11)+rnorm(3)
y4n<-c(12,12,12)+rnorm(3)
dane_n<-data.frame(id,y0n,y2n,y4n)#tworzymy ramkę danych w postaci szerokiej
long_dane_n=reshape(dane_n,direction="long",varying=list(2:4),v.names=c("Y"),times=c(0,2,4))
```

	id	time	Y
1.0	1	0	10.523015
2.0	2	0	9.372300
3.0	3	0	8.602420
1.2	1	2	10.746595
2.2	2	2	9.149214
3.2	3	2	11.495028
1.4	1	4	11.926965
2.4	2	4	10.571576
3.4	3	4	11.331615

Model z dwoma efektami losowym, efektem pacjenta i efektem nachylenia prostej

$$y = \beta_0 + X_{time} \beta_{time} + Z_{0,id} u_{0,id} + Z_{time,id} u_{time,id} + \varepsilon$$

$$(u_{0,id}, u_{time,id}) \sim N(\mathbf{0}, \begin{bmatrix} \sigma_{id}^2 & \sigma_{id,time} \\ \sigma_{id,time} & \sigma_{time}^2 \end{bmatrix}), \quad \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_9)$$

```
#uwzględnimy współczynniki a_i, b_i jako efekty losowe
#wyraz wolny i nachylenie są skorelowane
model1=lmer(Y~time+(time|id),data=long_dane_n,REML=FALSE)
summary(model1)
```

Linear mixed model fit by maximum likelihood [`lmerMod`]
 Formula: $Y \sim \text{time} + (\text{time} \mid \text{id})$
 Data: `long_dane_n`

AIC	BIC	logLik	deviance	df.resid
32.9	34.0	-10.4	20.9	3

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.41611	-0.43512	0.02556	0.37766	1.49662

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
id	(Intercept)	0.112040	0.33472	
	time	0.001014	0.03184	1.00
	Residual	0.469017	0.68485	

Number of obs: 9, groups: id, 3

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	9.5245	0.4094	23.263
time	0.4444	0.1410	3.152

Correlation of Fixed Effects:

(Intr)
 time -0.616

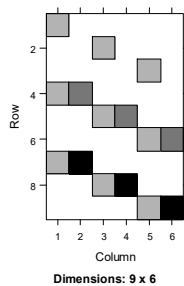
optimizer (nloptwrap) convergence code: 0 (OK)

boundary (singular) fit: see help('isSingular')

`t(modell@pp$Zt) # macierz Z`

```
9 x 6 sparse Matrix of class "dgCMatrix"
  1 1 2 2 3 3
1.0 1 . . . .
2.0 . 1 . . .
3.0 . . 1 . .
1.2 1 2 . . .
2.2 . . 1 2 .
3.2 . . . 1 2
1.4 1 4 . . .
2.4 . . 1 4 .
3.4 . . . 1 4
```

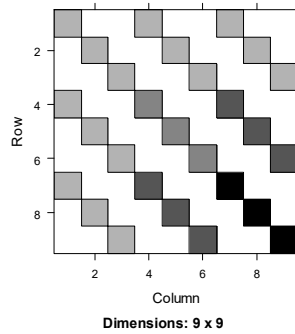
`image(getME(modell,"Z"))`



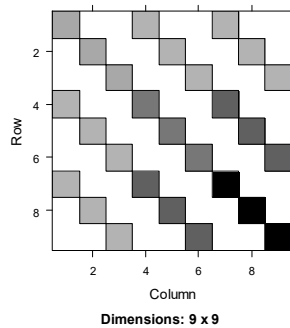
`t(modell@pp$Zt)%*%modell@pp$Zt # macierz Z*t(Z)`

```
9 x 9 sparse Matrix of class "dgCMatrix"
  1.0 2.0 3.0 1.2 2.2 3.2 1.4 2.4 3.4
1.0 1 . . 1 . . 1 . .
2.0 . 1 . . 1 . . 1 . .
3.0 . . 1 . . 1 . . 1 . .
1.2 1 . . 5 . . 9 . .
2.2 . 1 . . 5 . . 9 . .
3.2 . . 1 . . 5 . . 9 . .
1.4 1 . . 9 . . 17 . .
2.4 . 1 . . 9 . . 17 . .
3.4 . . 1 . . 9 . . 17 . .
```

image(t(modell@pp\$Zt)%*%modell@pp\$Zt)#graficzny obraz macierzy $Z^*t(Z)$



image(diag(9)+t(modell@pp\$Zt)%*%modell@pp\$Zt)#graficzny obraz macierzy $I+Z^*t(Z)$



Model z dwoma efektami losowym, efektem pacjenta i efektem nachylenia prostej ale oba efekty są niezależne

$$y = \beta_0 + X_{time}\beta_{time} + Z_{0,id}u_{0,id} + Z_{time,id}u_{time,id} + \varepsilon$$

$$(u_{0,id}, u_{time,id}) \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_{id}^2 & 0 \\ 0 & \sigma_{time}^2 \end{bmatrix}\right), \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_9)$$

#wyraz wolny i nachylenie są teraz nieskorelowane

```
modellbis=lmer(Y~time+(time-1|id)+(1|id),data=long_dane_n,REML=FALSE)
#modellbis=lmer(Y~time+(time||id),data=long_dane_n,REML=FALSE) #Inaczej
summary(modellbis)
```

Linear mixed model fit by maximum likelihood ['lmerMod']

Formula: Y ~ time + (time - 1 | id) + (1 | id)

Data: long_dane_n

AIC	BIC	logLik	deviance	df.resid
35.6	36.6	-12.8	25.6	4

Scaled residuals:

Min	1Q	Median	3Q	Max
-1.0862	-0.7158	-0.3459	0.7351	1.4045

Random effects:

Groups	Name	Variance	Std.Dev.
--------	------	----------	----------

```

id          time          0.0000 0.0000
id.1       (Intercept)   0.2234 0.4727
Residual                   0.8247 0.9082
Number of obs: 9, groups: id, 3

```

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	10.1996	0.5510	18.512
time	0.2768	0.1854	1.493

Correlation of Fixed Effects:

```

(Intr)
time -0.673
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

```

```
t(model1bis@pp$Zt) # macierz Z
```

```
9 x 6 sparse Matrix of class "dgCMatrix"
```

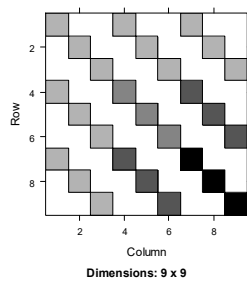
```

 1 2 3 1 2 3
1.0 . . . 1 . .
2.0 . . . . 1 .
3.0 . . . . . 1
1.2 2 . . 1 . .
2.2 . 2 . . 1 .
3.2 . . 2 . . 1
1.4 4 . . 1 . .
2.4 . 4 . . 1 .
3.4 . . 4 . . 1

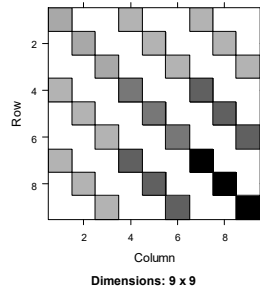
```

```
t(model1bis@pp$Zt)%*%model1bis@pp$Zt # macierz Z*t(Z)
```

```
image(t(model1bis@pp$Zt)%*%model1bis@pp$Zt)#graficzny obraz Z*t(Z)
```



```
image(diag(9))+t(model1bis@pp$Zt)%*%model1bis@pp$Zt #graficzny obraz I+Z*t(Z)
```



Macierz Z dla modeli:

`modell=lmer(Y~time+(time|id),data=long_dane_n,REML=FALSE)`

`modellbis=lmer(Y~time+(time-1|id)+(1|id),data=long_dane_n,REML=FALSE)`

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

Macierz D dla:

`modell=lmer(Y~time+(time|id),data=long_dane_n,REML=FALSE)`

$$V(\mathbf{u}) = \mathbf{D} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} & & & & \\ \sigma_{ab} & \sigma_b^2 & & & & \\ & & \sigma_a^2 & \sigma_{ab} & & \\ & & \sigma_{ab} & \sigma_b^2 & & \\ & & & & \sigma_a^2 & \sigma_{ab} \\ & & & & \sigma_{ab} & \sigma_b^2 \end{bmatrix}, \mathbf{u} = \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \\ u_{13} \\ u_{23} \end{bmatrix}, u_{1i} = a_i - a, u_{2i} = b_i - b.$$

Macierz D dla:

`modellbis=lmer(Y~time+(time-1|id)+(1|id),data=long_dane_n,REML=FALSE)`

Przykład 1: funkcjonowanie nerki po przeszczepie.

Przemysław Biecek -Analiza danych z programem R

Przykład 2: Beat the Blues

Handbook of Statistical Analyses Using R . Everit Hothorn ch. 10