

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 jedynkami, 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 przyjmujemy, 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)#tworzymy 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)
```

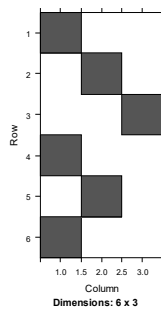
model0@pp\$Zt # transponowana macierz Z

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.

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

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

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

#uwzględnimy współczynniki a_i, b_i jako efekty losowe

#wyraz wolny i nachylenie są skorelowane

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

Linear mixed model fit by maximum likelihood [lmerMod']

Formula: Y ~ time + (time | 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')

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

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

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

#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')

modell@pp\$Zt # transponowana macierz Z

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

9 x 6 sparse Matrix of class "dgCMatrix"

	1	2	3	4	5	6
1.0	1	.	.	.	.	.
2.0	.	1	.	.	.	.
3.0	.	.	1	.	.	.
1.2	12	.	.	.	.	.
2.2	.	12	.	.	.	.
3.2	.	.	12	.	.	.
1.4	14	.	.	.	.	.
2.4	.	14	.	.	.	.
3.4	.	.	14	.	.	.

Macierz Z dla modeli:

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

modellbis 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)`

$$\mathbf{D} = \begin{bmatrix} \sigma_a^2 & & & & & \\ & \sigma_b^2 & & & & \\ & & \sigma_a^2 & & & \\ & & & \sigma_b^2 & & \\ & & & & \sigma_a^2 & \\ & & & & & \sigma_b^2 \end{bmatrix}$$

Macierz $\mathbf{Z}^T \mathbf{DZ}$ dla `model1=lmer(Y~time+(time|id),data=long_dane_n,REML=FALSE)`

$\mathbf{Z}^T \mathbf{DZ} =$

σ_a^2	0	0	$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 2\sigma_{ab}$	0	0
0	σ_a^2	0	0	$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 4\sigma_{ab}$	0
0	0	σ_a^2	0	0	$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 4\sigma_{ab}$
$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 2(\sigma_{ab} + 2\sigma_b^2)$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 2(\sigma_{ab} + 2\sigma_b^2)$	0	0
0	$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 2(\sigma_{ab} + 2\sigma_b^2)$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 4(\sigma_{ab} + 2\sigma_b^2)$	0
0	0	$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 2(\sigma_{ab} + 2\sigma_b^2)$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 4(\sigma_{ab} + 2\sigma_b^2)$
$\sigma_a^2 + 2\sigma_{ab}$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 2(\sigma_{ab} + 2\sigma_b^2)$	0	0	$\sigma_a^2 + 2\sigma_{ab} + 2(\sigma_{ab} + 2\sigma_b^2)$	0	0
0	$\sigma_a^2 + 4\sigma_{ab}$	0	0	$\sigma_a^2 + 4\sigma_{ab} + 2(\sigma_{ab} + 4\sigma_b^2)$	0	0	$\sigma_a^2 + 4\sigma_{ab} + 4(\sigma_{ab} + 4\sigma_b^2)$	0
0	0	$\sigma_a^2 + 4\sigma_{ab}$	0	0	$\sigma_a^2 + 4\sigma_{ab} + 2(\sigma_{ab} + 4\sigma_b^2)$	0	0	$\sigma_a^2 + 4\sigma_{ab} + 4(\sigma_{ab} + 4\sigma_b^2)$

Macierz $\mathbf{Z}^T \mathbf{DZ}$ dla `model1bis model1bis=lmer(Y~time+(time-1|id)+(1|id),data=long_dane_n,REML=FALSE)`

$\mathbf{Z}^T \mathbf{DZ} =$

σ_a^2	0	0	σ_a^2	0	0	σ_a^2	0	0
0	σ_a^2	0	0	σ_a^2	0	0	σ_a^2	0
0	0	σ_a^2	0	0	σ_a^2	0	0	σ_a^2
σ_a^2	0	0	$\sigma_a^2 + 4\sigma_b^2$	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0
0	σ_a^2	0	0	$\sigma_a^2 + 4\sigma_b^2$	0	0	$\sigma_a^2 + 8\sigma_b^2$	0
0	0	σ_a^2	0	0	$\sigma_a^2 + 4\sigma_b^2$	0	0	$\sigma_a^2 + 8\sigma_b^2$
σ_a^2	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0	$\sigma_a^2 + 16\sigma_b^2$	0	0
0	σ_a^2	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0	$\sigma_a^2 + 16\sigma_b^2$	0
0	0	σ_a^2	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0	$\sigma_a^2 + 16\sigma_b^2$

$V(\mathbf{y})$ dla `model1bis=lmer(Y~time+(time-1|id)+(1|id),data=long_dane_n,REML=FALSE)`

$V(\mathbf{y}) = \sigma_0^2 \mathbf{I} + \mathbf{Z}^T \mathbf{DZ} =$

$\sigma_0^2 + \sigma_a^2$	0	0	σ_a^2	0	0	σ_a^2	0	0
0	$\sigma_0^2 + \sigma_a^2$	0	0	σ_a^2	0	0	σ_a^2	0
0	0	$\sigma_0^2 + \sigma_a^2$	0	0	σ_a^2	0	0	σ_a^2
σ_a^2	0	0	$\sigma_0^2 + \sigma_a^2 + 4\sigma_b^2$	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0
0	σ_a^2	0	0	$\sigma_0^2 + \sigma_a^2 + 4\sigma_b^2$	0	0	$\sigma_a^2 + 8\sigma_b^2$	0
0	0	σ_a^2	0	0	$\sigma_0^2 + \sigma_a^2 + 4\sigma_b^2$	0	0	$\sigma_a^2 + 8\sigma_b^2$
σ_a^2	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0	$\sigma_0^2 + \sigma_a^2 + 16\sigma_b^2$	0	0
0	σ_a^2	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0	$\sigma_0^2 + \sigma_a^2 + 16\sigma_b^2$	0
0	0	σ_a^2	0	0	$\sigma_a^2 + 8\sigma_b^2$	0	0	$\sigma_0^2 + \sigma_a^2 + 16\sigma_b^2$

Przykład : funkcjonowanie nerki po przeszczepie.

W zbiorze danych **kidney** znajdują się dane o sprawności pracy nerki 334 pacjentów. Pomiary sprawności nerki zostały wykonane w 8 punktach czasowych po przeszczepie nerki. Dla każdego pacjenta mamy więc krótki ośmioelementowy wektor z pomiarami. Dla każdego pacjenta zostały zmierzone takie cechy jak wiek pacjenta , wiek dawcy nerki, liczba niezgodności antygenów.

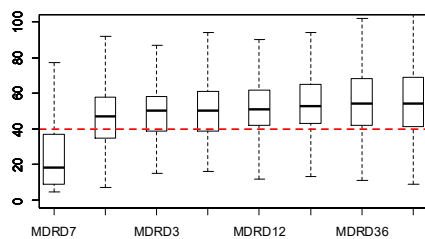
Zmienne

- recipient.age, donor.age – wiek biorcy i dawcy nerki w chwili wykonywania przeszczepu
- CIT (*cold ischemia time*) – czas w godzinach w którym nerka przebywała w lodzie
- disc.AB, disc.DR – liczba niezgodności a antygenach AB i DR - im większa zgodność immunologiczna, tym mniejsze ryzyko odrzucania nerki, mniejsza dawka leków immunosupresyjnych, które biorca będzie otrzymywał.
- therapy– terapia immunosupresyjna, która była stosowana dla pacjenta- jeden z trzech schematów leczenia.
- diabetes– informacja czy pacjent jest cukrzykiem
- bpl.drugs – liczba leków obniżających ciśnienie krwi
- MDRDxx– wartość MDRD (*Modification of Diet in Renal Disease*) w 7 dniu po przeszczepie oraz w 3,6,12,24,36 i 60 miesiącu po przeszczepie. Wartość MDRD wyznacz się na podstawie ilości kreatyniny, płci , wieku, rasy i jest oceną wydajności nerek czyli filtracji kłębuszkowej GFR(*glomerular filtration rate*). Im większa jest wartość MDRD tym sprawniejsza jest nerka

Dane są dostępne w tzw. postaci szerokiej tzn. każdy wiersz zawiera dane dotyczące jednego pacjenta. Aby móc skorzystać z danych w funkcji `lmer()` należy zmienić format danych z **szerokiego** na **wąski**. Służy do tego funkcja `reshape()`.

Plik kidney.R

```
library(PBImisc)
#wczytujemy zbiór danych kidney
data(kidney)
#krótkie graficzne i tekstowe przedstawienie zmienności funkcji nerki w czasie
boxplotInTime(kidney[,9:16],colnames(kidney)[9:16],ylim=c(0,100)) #począć na rysunek
abline(c(40,0),col="red",lwd=2,lt=2)# MDRD poniżej 40 oznacza złe funkcjonowanie nerki
```



```
colMeans(kidney[,9:16])
```

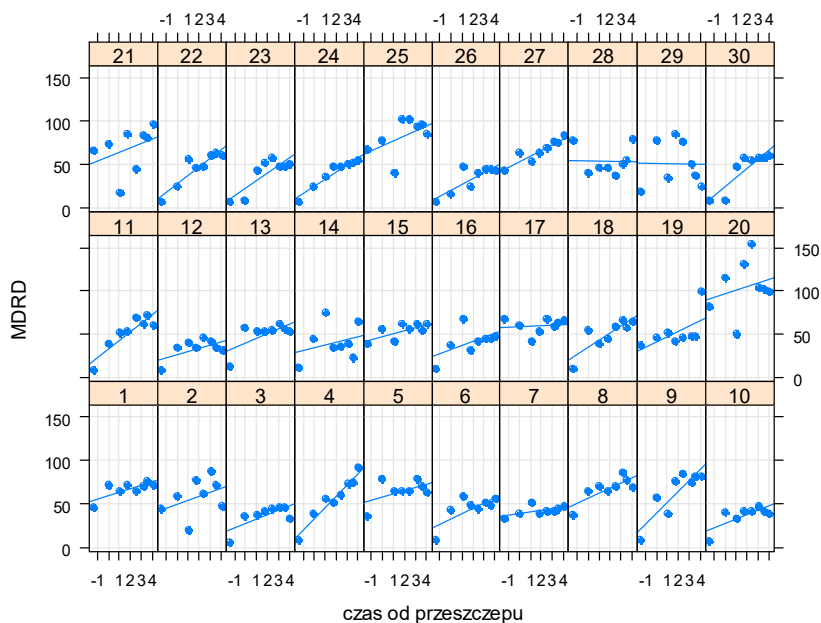
```
#przekształcanie formatu danych z szerokiego na wąski
```

```
lkidney=reshape(kidney,direction="long",varying=list(9:16),v.names=c("MDRD"),times=c(1/4,1,3,6,12,24,36,60))
```

```
library(lattice)
```

```
#wykres zależności MDRD~log(time) dla pierwszych 30 pacjentów
```

```
xyplot(MDRD~log(time)|factor(id),subset(lkidney,id<31.),type=c("g","p","r"),xlab="czas od przeszczepu",ylab="MDRD",aspect="xy",pch=19)
```



```
#czas od przeszczepu będziemy analizować po zlogarytmowaniu
```

```
lkidney$time=log(lkidney$time)
```

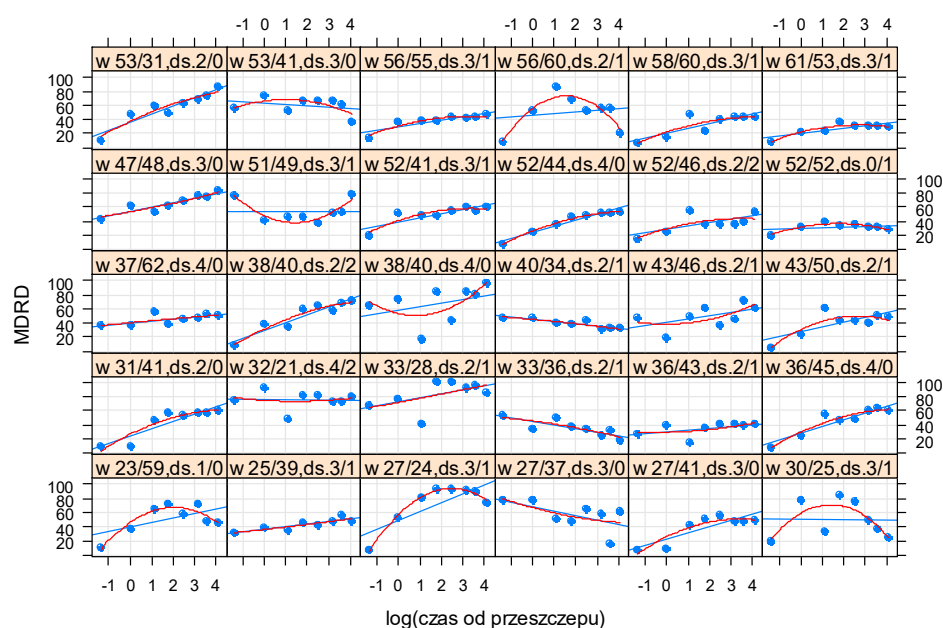
```
#do wykresu dodamy opis wieku biorcy/dawcy oraz informację o liczbie niezgodności w antygenach AB/DR
```

```
lkidney$opis=factor(with(lkidney,paste("w
```

```
",recipient.age,"/",donor.age,"ds.",discrepancy.AB,"/",discrepancy.DR,sep=""))))
```

```
#funkcja panel.z.kwadratem() dorysuję do wykresu xyplot krzywą trendu kwadratowego
panel.z.kwadratem<-function(...){
  nx=list(...)$x
  ny=list(...)$y
  cc=lm(ny~nx+I(nx^2))$coef
  panel.xyplot(...)
  panel.curve(x^2*cc[3]+x*cc[2]+cc[1],from=min(nx),to=max(nx),col="red")
}
```

```
#funkcją xyplot{lattice} rysujemy sprawność nerki dla wybranych pacjentów
xyplot(MDRD~time|opis,lkidney,type=c("g","p","r"),xlab="log(czas od
przeszczepu)",ylab="MDRD",pch=19,panel=panel.z.kwadratem,subset=id>20 & id <51)
```



Pytania badawcze.

- Jak uwzględnić to, że 8 pomiarów jest wykonywanych dla tego samego pacjenta? Pomiary wykonane dla tego samego pacjenta dzielą wspólny efekt pacjenta
- Jak uwzględnić w funkcji nerki po przeszczepie. Czy przyjęcie liniowego trendu nie jest zbyt uproszczeniem?
- Jakie zmienne mają wpływ na funkcję nerki po przeszczepie?

Zacznijmy od pierwszego pytania. Ponieważ niektórzy pacjenci mają sprawniejszą nerkę a inni mniej więc uwzględnimy średni stan nerki jako efekt losowy

Model z jednym efektem losowym, efektem pacjenta

$$y_{MDRD} = \beta_0 + Z_{id}u_{id} + \varepsilon \quad u_{id} \sim N(\mathbf{0}, \sigma_{id}^2 \mathbf{I}_{334}), \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_{2672})$$

Macierz Z_{id} to macierz kodująca który pomiar był wykonany dla którego pacjenta

```
#model z jednym efektem losowym, efektem pacjenta /nerki
library(lme4)
model0=lmer(MDRD~1+(1|id),data=lkidney,REML=FALSE)
summary(model0)
```

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: MDRD ~ 1 + (1 | id)
Data: lkidney

AIC	BIC	logLik	deviance	df.resid
23222.0	23239.7	-11608.0	23216.0	2669

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.6366	-0.3947	0.0924	0.5396	4.4835

Random effects:

Groups	Name	Variance	Std.Dev.
id	(Intercept)	131.8	11.48
Residual		286.5	16.93

Number of obs: 2672, groups: id, 334

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	48.4804	0.7083	68.44

Model `model0` uwzględnia różnice między średnimi (w czasie tzn. z 8 pomiarów) wartościami MDRD dla różnych pacjentów.

Przyjmijmy, że zmianę MDRD dla pacjenta i opiszemy równaniem liniowym

$$y_{i,time} = a_i + b_i x_{i,time},$$

gdzie $x_{i,time}$ oznacza logarytm z czasu od przeszczepu.

Rozważymy dwa modele: w jednym uwzględnimy oba współczynniki jako losowe a w drugim jako stałe

Model uwzględniający te współczynniki jako losowe jest postaci

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

$$(u_{0,id}, u_{time,id}) \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_{id}^2 & \sigma_{id,time} \\ \sigma_{id,time} & \sigma_{time}^2 \end{bmatrix}\right), \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_{2672})$$

W tym modelu średnia wartość MDRD dla każdego pacjenta jest rozbita na dwa efekty, efekt stały β_0 i efekt losowy $u_{0,id}$ tak, że $a_i = \beta_0 + u_{0,id}$, a współczynnik liniowej zależności pomiędzy logarytmem czasu a MDRD jest rozbity na β_{time} i $u_{time,id}$ tak, że $b_i = \beta_{time} + u_{time,id}$

Dopuszczamy tu istnienie zależności pomiędzy $u_{0,id}$ i $u_{time,id}$

```
#uwzględnimy współczynniki a_i ,b_i jako efekty losowe dopuszczając ich skorelowanie
modell=lmer(MDRD~time+(time|id),data=lkidney,REML=FALSE)
summary(modell)
```

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

Formula: MDRD ~ time + (time | id)

Data: lkidney

AIC	BIC	logLik	deviance	df.resid
22445.1	22480.5	-11216.6	22433.1	2666

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.3317	-0.4637	-0.0046	0.4961	3.8846

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
id	(Intercept)	135.112	11.624	
	time	8.375	2.894	-0.16
Residual		185.263	13.611	

Number of obs: 2672, groups: id, 334

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	40.1081	0.7425	54.02
time	4.5119	0.2181	20.69

Correlation of Fixed Effects:

(Intr)
time -0.355

Korelacja pomiędzy $u_{0,id}$ i $u_{time,id}$ została oceniona na -0.16 co można interpretować następująco:

im wyższa sprawność początkowa tym mniejsza jej poprawa jej funkcjonowania w czasie. Korelacja ta jest słaba.

Jeżeli chcemy rozważyć model w którym $u_{0,id}$ i $u_{time,id}$ mają być traktowane niezależnie, czyli

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

$$(u_{0,id}, u_{time,id}) \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_{id}^2 & 0 \\ 0 & \sigma_{time}^2 \end{bmatrix}\right), \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_{2672})$$

musimy użyć poniższej formuły

```
#uwzględnimy współczynniki a_i ,b_i jako efekty losowe wykluczając ich skorelowanie
modellbis=lmer(MDRD~time+(time-1|id)+(1|id),data=lkidney,REML=FALSE)
summary(modellbis)
```

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

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

Data: lkidney

AIC	BIC	logLik	deviance	df.resid
-----	-----	--------	----------	----------

22446.2 22475.6 -11218.1 22436.2 2667

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.4154	-0.4560	-0.0077	0.4974	3.8352

Random effects:

Groups	Name	Variance	Std.Dev.
id	time	7.543	2.747
id.1	(Intercept)	125.799	11.216
Residual		186.844	13.669

Number of obs: 2672, groups: id, 334

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	40.1081	0.7243	55.37
time	4.5119	0.2127	21.21

Correlation of Fixed Effects:

(Intr)
time -0.273

#składowa stała w modelu mieszanym
fixef(model1)

(Intercept)	time
40.108126	4.511859

#składowa losowa w modelu mieszanym
head(ranef(model1)\$id)

(Intercept)	time
1 15.5033695	0.14873551
2 7.9479289	0.34348993
3 -10.0745004	-0.08182755
4 -0.7259365	4.62145998
5 13.8444517	0.07262906
6 -4.1490093	0.64730425

#współczynniki a_i i b_i dla modelu mieszanego
współczynnikiLosowe=t(t(ranef(model1)\$id)+fixef(model1))
head(współczynnikiLosowe)

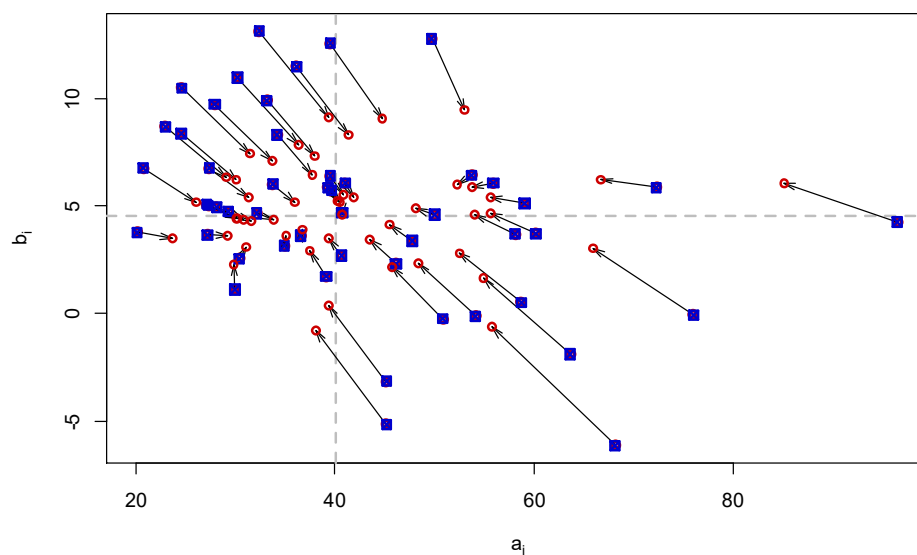
(Intercept)	time
1 55.61150	4.660595
2 48.05605	4.855349
3 30.03363	4.430032
4 39.38219	9.133319
5 53.95258	4.584489
6 35.95912	5.159164

#model w którym a_i oraz b_i są uwzględnione jako efekty stałe
model1stały=lm(MDRD~time:factor(id)+factor(id)-1,data=lkidney)
współczynnikiStałe=matrix(coef(model1stały),ncol=2)

```
head(współczynnikiStałe)
      [,1] [,2]
[1,] 60.12820 3.703253
[2,] 49.97823 4.592431
[3,] 27.07673 5.064779
[4,] 32.35810 13.124720
[5,] 58.07457 3.678262
[6,] 33.73900 6.001253
```

```
#równoważnie możemy dopasować listę modeli liniowych używając funkcji lmList;
#inny zapis ale te same wyniki co w modelu stałym
lmList(MDRD~time|id,lkidney)
```

```
#wykresy współczynników dla 50 pacjentów
plot(współczynnikiStałe[1:50,1],współczynnikiStałe[1:50,2],cex=1,pch=21,lwd=3,col="red3",xlab=expression(a[i]),ylab=expression(b[i]))
abline(v=fixef(model1)[1],col="grey",lwd=2,lty=2)
abline(h=fixef(model1)[2],col="grey",lwd=2,lty=2)
arrows(współczynnikiStałe[1:50,1],współczynnikiStałe[1:50,2],
współczynnikiLosowe[1:50,1],współczynnikiLosowe[1:50,2],length=0.1,angle=15,code=2,lwd=1)
points(współczynnikiStałe[1:50,1],współczynnikiStałe[1:50,2],cex=1,pch=7,lwd=2,col="blue3")
points(współczynnikiLosowe[1:50,1],współczynnikiLosowe[1:50,2],cex=1,pch=21,lwd=2,col="red3")
```



Model mieszany uwzględniający kwadratowy trend

$$y_{MDRD} = \beta_0 + X_{time}\beta_{1,time} + X_{time}^2\beta_{2,time} + Z_{0,id}u_{0,id} + Z_{time,id}u_{1,time,id} + Z_{time,id}^2u_{2,time,id} + \varepsilon$$

$$(u_{0,id}, u_{1,time,id}, u_{2,time,id}) \sim N(\mathbf{0}, \begin{bmatrix} \sigma_{id}^2 & \sigma_{id,1,time} & \sigma_{id,2,time} \\ \sigma_{id,1,time} & \sigma_{1,time}^2 & \sigma_{id,1,2,time} \\ \sigma_{id,2,time} & \sigma_{id,1,2,time} & \sigma_{2,time}^2 \end{bmatrix}), \quad \varepsilon \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_{2672})$$

#model z kwadratowym trendem

```
model2=lmer(MDRD~time+I(time^2)+(time+I(time^2)|id),data=lkidney,REML=FALSE,
control = lmerControl(optimizer="Nelder_Mead"))
summary(model2)
```

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: MDRD ~ time + I(time^2) + (time + I(time^2) | id)
Data: lkidney

AIC	BIC	logLik	deviance	df.resid
22059.4	22118.3	-11019.7	22039.4	2662

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.4478	-0.4584	-0.0372	0.4584	4.0756

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
id	(Intercept)	157.138	12.535	
	time	28.775	5.364	0.06
	I(time^2)	2.394	1.547	-0.25 -0.80
Residual		134.095	11.580	

Number of obs: 2672, groups: id, 334

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	42.1031	0.7662	54.95
time	8.8684	0.3901	22.73
I(time^2)	-1.5441	0.1158	-13.33

Correlation of Fixed Effects:

	(Intr)	time
time	0.019	
I(time^2)	-0.254	-0.829

```
#test ilorazu wiarygodności porównujący kolejno opracowane modele
anova(model2,model1,model0)
anova(model2,model1)
anova(model1,model0)
```

Data: lkidney

Models:

model0: MDRD ~ 1 + (1 | id)

```

model1: MDRD ~ time + (time | id)
model2: MDRD ~ time + I(time^2) + (time + I(time^2) | id)
  npar  AIC    BIC   logLik      deviance    Chisq   Df   Pr(>Chisq)
model0   3 23222  23240 -11608      23216          3          3   < 2.2e-16 ***
model1   6 22445  22481 -11217      22433      782.84    3   < 2.2e-16 ***
model2  10 22059  22118 -11020      22039      393.72    4   < 2.2e-16 ***

```

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

Uwaga. Najlepsze dopasowanie uzyskujemy z modelu2. Modele 1 i 2 poprawiają funkcję wiarygodności o kilkaset punktów. Do duża zmiana

#czynniki wpływające na funkcję nerki

```

modelPełny=lmer(MDRD~donor.age+recipient.age+therapy+diabetes+bpl.drugs+discrepancy.AB+discrepancy.DR+time+I(time^2)+(time+I(time^2)|id),data=lkidney,REML=FALSE)
summary(modelPełny)

```

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

Formula: MDRD ~ donor.age + recipient.age + therapy + diabetes + bpl.drugs + discrepancy.AB + discrepancy.DR + time + I(time^2) + (time + I(time^2) | id)

Data: lkidney

AIC	BIC	logLik	deviance	df.resid
21983.1	22089.2	-10973.6	21947.1	2654

Scaled residuals:

Min	1Q	Median	3Q	Max
-4.3696	-0.4479	-0.0326	0.4542	4.1354

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
id	(Intercept)	126.381	11.242	
	time	28.744	5.361	-0.01
	I(time^2)	2.392	1.547	-0.25 -0.80
	Residual	134.112	11.581	

Number of obs: 2672, groups: id, 334

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	68.80617	3.41022	20.176
donor.age	-0.30158	0.05014	-6.015
recipient.age	-0.14409	0.05866	-2.456
therapycm	-1.41539	1.56787	-0.903
therapytc	-4.75594	1.47255	-3.230
diabetes	2.76084	1.48438	1.860
bpl.drugs	-2.14907	0.54829	-3.920
discrepancy.AB	-0.79380	0.71060	-1.117
discrepancy.DR	-0.38084	1.15305	-0.330
time	8.86844	0.38997	22.742
I(time^2)	-1.54412	0.11580	-13.335

Correlation of Fixed Effects:

	(Intr)	donr.g	repnt.	thrpvc	thrpvt	diabts	bpl.dr	dsc.AB	dsc.DR	time
donor.age		-0.331								
recipient.g			-0.573							

```

therapycm -0.132 -0.145 0.114
therapytc -0.077 -0.268 0.135 0.399
diabetes 0.139 -0.015 -0.244 -0.027 -0.004
bpl.drugs -0.231 -0.106 -0.065 -0.046 -0.104 -0.100
dscrpnycy.AB -0.399 -0.076 0.009 -0.002 0.046 0.004 -0.062
dscrpnycy.DR -0.264 0.010 0.023 -0.004 -0.069 -0.086 0.140 -0.058
time -0.007 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
I(time^2) -0.053 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 -0.829
optimizer (nloptwrap) convergence code: 0 (OK)
Model failed to converge with max|grad| = 0.0215838 (tol = 0.002, component 1)

```

Z dodanych zmiennych najbardziej istotne, to **wiek dawcy** (im starszy dawca tym gorsza praca nerki) i **liczba leków na nadciśnienie** (im więcej leków tym gorsze wyniki MDRD)

Zbudujemy model końcowy w który uwzględnimy tylko istotne zmienne uwzględnione jako efekty stałe i kwadratowy trend

```

modelKońcowy=lmer(MDRD~donor.age+bpl.drugs+time+I(time^2)+(time+I(time^2)|id),data=lkidney,REML=FALSE)
summary(modelKońcowy)

```

```

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: MDRD ~ donor.age + bpl.drugs + time + I(time^2) + (time + I(time^2) | id)
Data: lkidney

```

```

AIC      BIC    logLik deviance df.resid
21988.1  22058.8 -10982.0 21964.1    2660

```

Scaled residuals:

```

   Min      1Q  Median      3Q      Max
-4.3473 -0.4539 -0.0354  0.4511  4.1245

```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
id	(Intercept)	133.624	11.560	
	time	28.748	5.362	0.04
	I(time^2)	2.391	1.546	-0.29 -0.80
Residual		134.118	11.581	

Number of obs: 2672, groups: id, 334

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	62.40830	2.30581	27.066
donor.age	-0.36131	0.04838	-7.468
bpl.drugs	-2.29584	0.54945	-4.178
time	8.86844	0.38998	22.741
I(time^2)	-1.54412	0.11579	-13.335

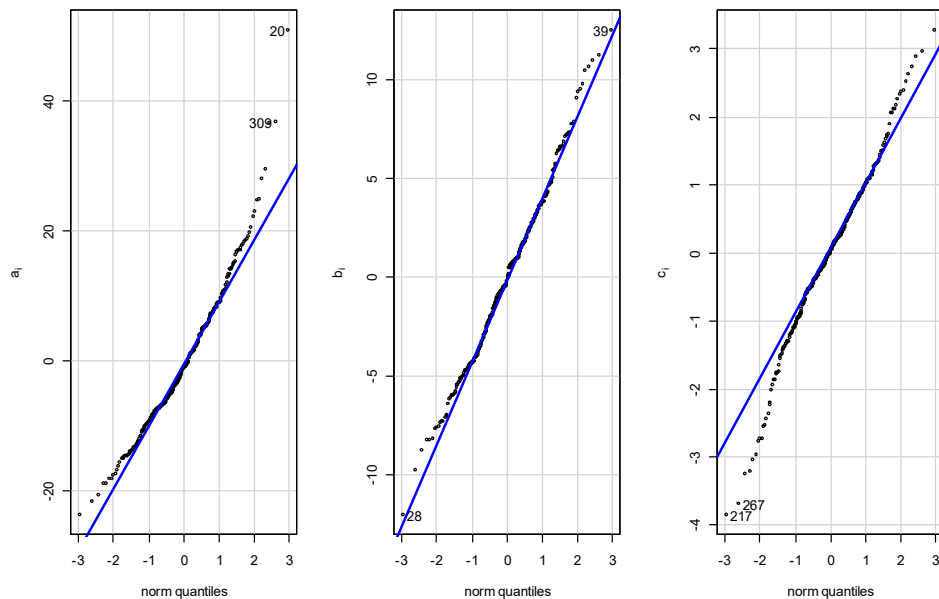
Correlation of Fixed Effects:

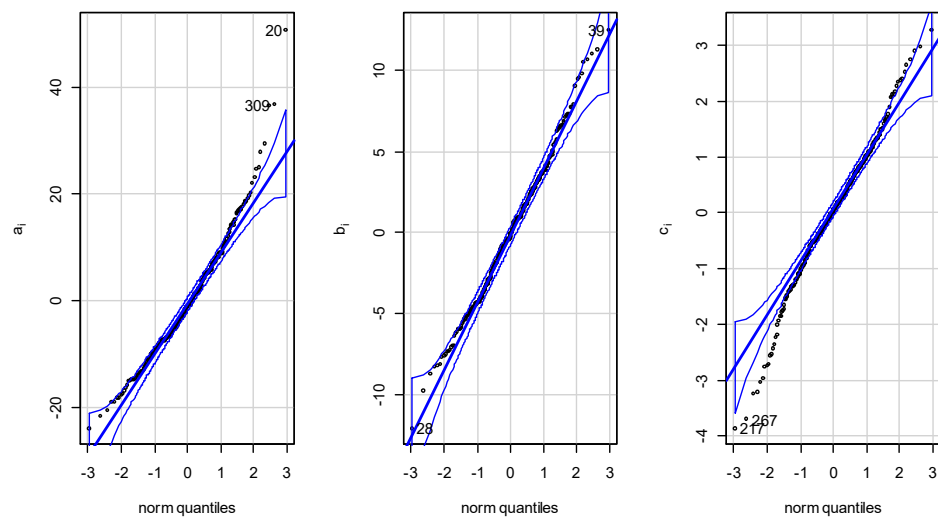
	(Intr)	donr.g	bpl.dr	time
donor.age	-0.784			
bpl.drugs	-0.399	-0.166		
time	0.000	.000	0.000	

I(time^2) -0.089 0.000 0.000 -0.829

```
#diagnostyka modelu
#wariancja zmiennej MDRD
(var1=var(lkidney$MDRD))
[1] 418.4605
#wariancja resztowa po dopasowaniu Modelu końcowego
(var2=var(residuals(modelKońcowy)))
[1] 99.81622
#jaki procent wyjściowej zmienności jest wyjaśniona przez model końcowy
(1-var2/var1)*100
[1] 76.1468
#wykresy diagnostyczne dla efektów losowych modelu modelKońcowy;
library(car)
par(mfrow=c(1,3))
qqPlot(ranef(modelKońcowy)$id[,1],envelope=FALSE,ylab=expression(a[i]))
qqPlot(ranef(modelKońcowy)$id[,2],envelope=FALSE,ylab=expression(b[i]))
qqPlot(ranef(modelKońcowy)$id[,3],envelope=FALSE,ylab=expression(c[i]))
qqPlot(ranef(modelKońcowy)$id[,3],envelope=FALSE,ylab=expression(c[i]))
qqPlot(ranef(modelKońcowy)$id[,1],envelope=TRUE,ylab=expression(a[i]))
qqPlot(ranef(modelKońcowy)$id[,2],envelope=TRUE,ylab=expression(b[i]))
qqPlot(ranef(modelKońcowy)$id[,3],envelope=TRUE,ylab=expression(c[i]))

par(mfrow=c(1,1))
```





Dla efektów b_i zgodność z rozkładem normalnym jest duża natomiast dla pozostałych efektów widać słabe dopasowanie w ogonach.