

# Osiem historyjek o modelach liniowych i mieszanych

Przemyslaw.Biecek@gmail.com, MIM Uniwersytet Warszawski

## Plan prezentacji

- 1 politechnika, czyli ANOVA jednokierunkowa, testy post-hoc i analiza kontrastów,
- 2 mleko i krowy, czyli model mieszany z jednym efektem losowym,
- 3 mieszkania, czyli regresja prosta,
- 4 studium snu, czyli model mieszany z jedną zmienną grupującą,
- 5 mleko i byki, czyli model mieszany z zadaną strukturą korelacji,
- 6 miasta, czyli model mieszany dla danych przestrzennych,
- 7 węzły chłonne, czyli model hierarchiczny vs. crossed,
- 8 przeżycia 5-letnie, czyli regresja logistyczna i efekty losowe.

# Funkcje o których będziemy mówić

```
# z pakietu stats
lm(formula, data, subset, weights, contrasts=NULL, ...)

# z pakietu nlme
lme(fixed, data, random, correlation, weights, contrasts=NULL, ...)

# z pakietu MASS
glmmPQL(fixed, random, family, data, correlation, weights, verbos=TRUE, ...)

# z pakietu lme4
lmer(formula, data, REML=TRUE, start=NULL, verbose=FALSE, weights,
      contrasts=NULL, ...)

glmer(formula, data, family=gaussian, start=NULL, verbose=FALSE, weights,
       contrasts=NULL, ...)

nlmer(formula, data, start=NULL, verbose=FALSE, weights, contrasts=NULL, ...)
```

# ANOVA jednokierunkowa

W jednokierunkowej ANOVA porównujemy średnie wartości pewnej cechy ilościowej w rozbiciu na grupy określone przez zmienną jakościową. Przyjmujemy, że poziomy zmiennej jakościowej są charakterystyczne dla całej populacji, a nie tylko dla próby.

$$y_i = \mu + \mu_{g(i)} + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

Interesujące nas zagadnienia to:

- estymacja efektu grupy,
- test na istotność różnicy średnich,
- testy post-hoc,
- testy określonych kontrastów.

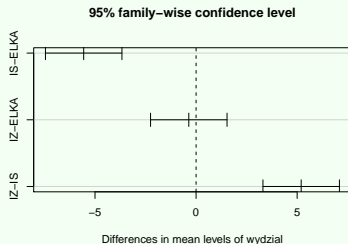
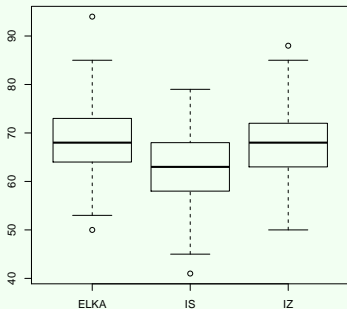
# ANOVA jednokierunkowa

```
> library(PBImisc)
> data(ocenyStudentow)
> summary(ocenyStudentow)
```

student	przedmiot	wydzial
S1 : 3	Algebra :150	ELKA:150
S10 : 3	Analiza I :150	IS :150
S100 : 3	Analiza II:150	IZ :150
S101 : 3		
S102 : 3		
S103 : 3		
(Other):432		

	wykladowca	oceny
Alicja (Brzytwa) Fiolkowska	:75	Min. :41.00
Bartek (Blizniak) Nowak	:75	1st Qu.:61.00
Jurek (Killer) Kowalski	:75	Median :67.00
Krzysiek (Kosa) Kwiatkowski	:75	Mean :66.31
Mariusz (Rozniczka) Wisniewski	:75	3rd Qu.:71.00
Marta (Niezle) Ziolk	:75	Max. :94.00

# ANOVA jednokierunkowa



Używamy funkcji `lm(stats)`, `HSD.test(agricolae)`, `TukeyHSD(stats)` do określenia efektu wydziału.

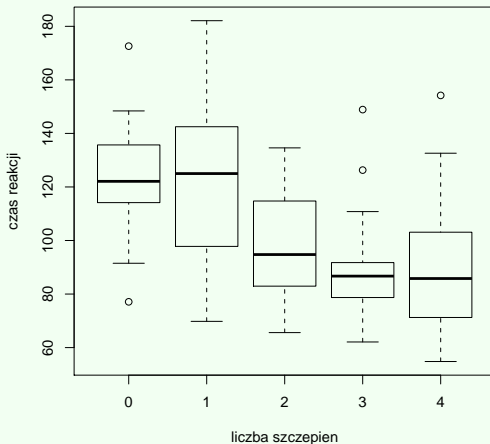
# ANOVA jednokierunkowa

```
> library(PBImisc)
> data(szczepienia)

> summary(szczepienia)
  czas.reakcji  l.szczepien
Min.   : 39.50   0:20
1st Qu.: 77.30   1:20
Median : 99.25   2:20
Mean   : 97.89   3:20
3rd Qu.:117.70   4:20
Max.   :154.70

> boxplot(czas.reakcji~l.szczepien, data=szczepienia)
```

# ANOVA jednokierunkowa



# Model mieszany z jednym efektem losowym

W model mieszanym z jednym czynnikiem losowym interesować nas będzie zróżnicowanie w wartościach średnich pewnej cechy ilościowej w rozbiciu na grupy określone przez zmienną jakościową. Poziomy zmiennej jakościowej nie muszą być reprezentatywne dla całej populacji, mogą być zależne od próby.

$$y_i = \mu + a_{g(i)} + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

$$a \sim \mathcal{N}(0, \sigma_a^2)$$

Interesujące nas zagadnienia to:

- estymacja zmienności w wartościach średnich,
- test na istotność średniej,
- test na istotność czynnika losowego.

## Estymacja / predykcja

- Parametrami modelu są  $\mu$ ,  $\sigma_0^2$  i  $\sigma_a^2$ , te wartości możemy estymować.
- Efekty losowych ( $a_i$ ) to zmienne losowe, nie estymujemy ich ale prognozujemy ich wartości używając wyestymowanych współczynników modelu, np. używając równań Henderson (Henderson 1984). Dla modelu

$$y = X\beta + Z\gamma + \varepsilon$$

ewktor efektów stałych i losowych możemy wyznaczyć rozwiązując układ równań

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \hat{\sigma}_e^2 \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix}$$

# Model mieszany z jednym efektem losowym

```

> library(PBImisc)
> library(lattice)
> library(RColorBrewer)
> library(lme4)
>
> data(mlecznoscKrow)
>
> summary(mlecznoscKrow)
      krowa      pomiar
krowaA : 4   Min.    :19.40
krowaB : 4   1st Qu.:24.65
krowaC : 4   Median  :27.30
krowaD : 4   Mean    :27.02
krowaE : 4   3rd Qu.:29.95
krowaF : 4   Max.    :32.00
(Other):16

> dotplot(krowa~pomiar, data=mlecznoscKrow, xlab="mlecznosc [kg/dzien]")

```

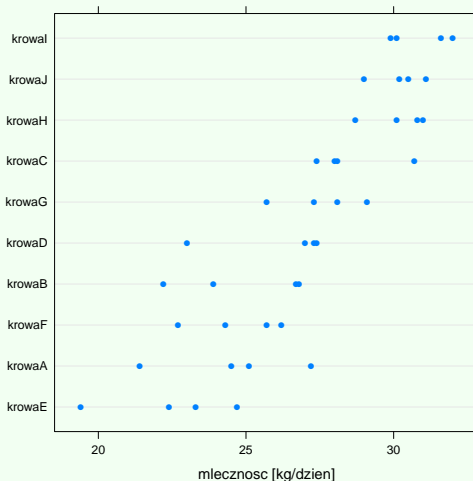
Mleczność krów

# Model mieszany z jednym efektem losowym



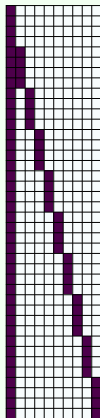
Mleczność krów

# Model mieszany z jednym efektem losowym

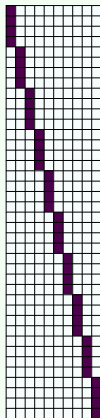


# Model mieszany z jednym efektem losowym

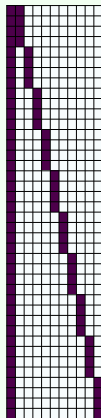
`lm(pomiar~krowa)`



`lm(pomiar~krowa-1)`



`nlme(pomiar~(1|krowa))`



# Regresja prosta

W prostym modelu regresji liniowej interesować nas będzie zależność pewnej cechy ilościowej od innej cechy ilościowej.

$$y_i = \mu + \beta x_i + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

Interesujące nas zagadnienia to:

- estymacja parametrów modelu  $\mu, \beta$ ,
- test na istotność tych współczynników modelu,
- diagnostyka reszt z modelu, badanie jakości dopasowania modelu.

# Regresja prosta

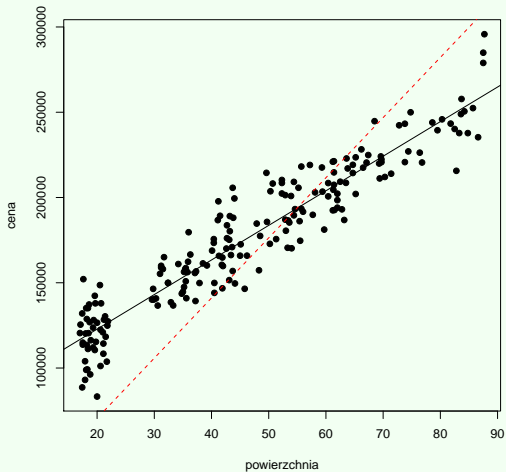
```

> library(PBImisc)
>
> data(mieszkania)
>
> summary(mieszkania)
      cena          pokoi      powierzchnia      dzielnica
Min.   : 83280   Min.   :1.00   Min.   :17.00   Biskupin   :65
1st Qu.:143304   1st Qu.:2.00   1st Qu.:31.15   Krzyki     :79
Median :174935   Median :3.00   Median :43.70   Srodmiescie:56
Mean   :175934   Mean   :2.55   Mean   :46.20
3rd Qu.:208741   3rd Qu.:3.00   3rd Qu.:61.40
Max.   :295762   Max.   :4.00   Max.   :87.70
      typ.budynku
kamienica :61
niski blok:63
wieszowiec :76

> plot(cena~powierzchnia, data=mieszkania, xlab="powierzchnia [m2]")

```

# Regresja prosta



# Model mieszany z jedną zmienną grupującą

W modelu mieszanym z jedną zmienną grupującą (model losowych współczynników) interesować zmienność współczynników regresji liniowej w grupach wyznaczonych przez inną zmienną jakościową. Poziomy tej zmiennej jakościowej nie muszą charakteryzować całej populacji.

$$y_i = \mu + \beta x_i + (m_{g(i)} + b_{g(i)} x_i) + \varepsilon_i,$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2),$$

$$(m_{g(i)}, b_{g(i)}) \sim \mathcal{N}(0, \Sigma).$$

Interesujące nas zagadnienia to:

- estymacja parametrów modelu  $\mu$ ,  $\beta$  i  $\Sigma$ ,
- test na istotność współczynników modelu,
- diagnostyka modelu.

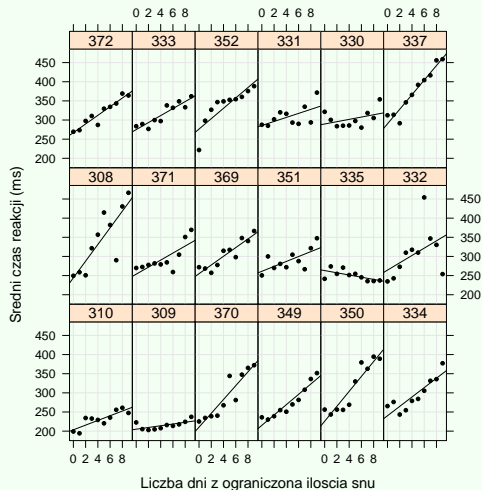
# Model mieszany z jedną zmienną grupującą

```
> library(lme4)
> data(sleepstudy)
> summary(sleepstudy)
```

Reaction	Days	Subject
Min. :194.3	Min. :0.0	308 : 10
1st Qu.:255.4	1st Qu.:2.0	309 : 10
Median :288.7	Median :4.5	310 : 10
Mean :298.5	Mean :4.5	330 : 10
3rd Qu.:336.8	3rd Qu.:7.0	331 : 10
Max. :466.4	Max. :9.0	332 : 10
		(Other):120

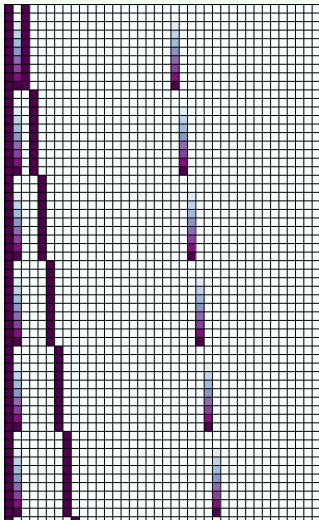
```
>
> xyplot(Reaction ~ Days | Subject, sleepstudy,
+       type = c("g","p","r"), index = function(x,y) coef(lm(y ~ x))[1],
+       xlab = "Liczba dni z ograniczona iloscia snu",
+       ylab = "Sredni czas reakcji (ms)", aspect = "xy")
```

# Model mieszany z jedną zmienną grupującą



# Model mieszany z jedną zmienną grupującą

`lmer(Reaction ~ Days + (Days|Subject))`



# Znana struktura korelacji

W model mieszanym z jednym czynnikiem losowym efekty losowe nie muszą być iid (poprzednio to zakładaliśmy).

W pewnych sytuacjach nie dość, że wiemy, że efekty losowe nie są iid, to znamy ich macierz korelacji (podobnie może być dla realizacji  $\varepsilon$ ).

$$y_i = \mu + a_{g(i)} + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

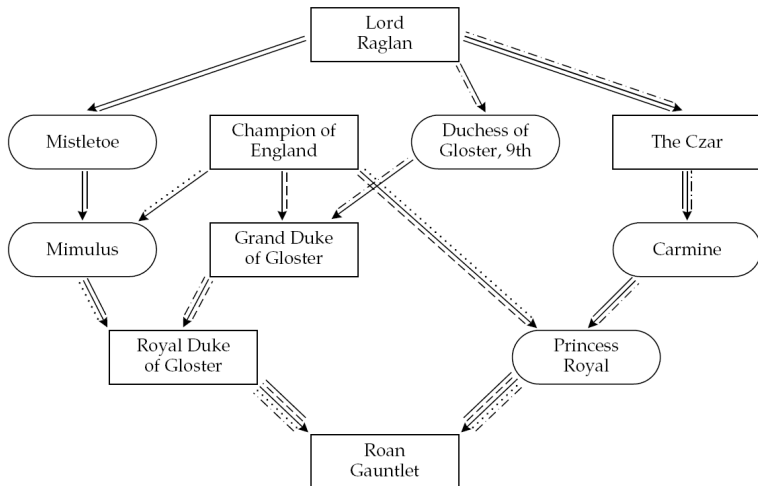
$$a \sim \mathcal{N}(0, \sigma_a^2)$$

$$\text{Cor}(a_i, a_j) = \rho(i, j)$$

Interesujące nas zagadnienia to:

- estymacja parametrów modelu, w tym  $\sigma_a^2$ ,
- predykcja efektów losowych,
- testy.

# Zadana struktura korelacji



# Znana struktura korelacji

Model poligeniczny opisuje wpływ małych genów mających niewielki addytywny wpływ na badaną cechę. Estymacja wpływu każdego z tych genów przy obecnych rozmiarach próby jest niemożliwa, chcemy jednak uwzględnić te efekty w przeprowadzanych analizach. Jeżeli więc model opisujący wpływ genów na ceche jest postaci

$$Y_i = \mu + \sum_l x_{i,l} \beta_l + a_i + \varepsilon_i$$

gdzie  $i$  to numer osobnika,  $\beta_l$  oznacza efekt  $l$ ty efekt stały (np. QTL),  $a_i$  odpowiada wpływowi poligenicznemu,  $\varepsilon_i$  to szum środowiskowy iid. o rozkładzie normalnym.

Korelację pomiędzy osobnikami można wyznaczyć z macierzy pokrewieństwa, nie wdając się w szczegóły dla populacji outbred przy założeniu równowadze Hardyego-Weinberga można policzyć ją ze wzoru

$$\text{Cov}(a_i, a_j) = 2\Phi^{i,j} \sigma_a^2 + \Delta_7^{i,j} \sigma_d^2$$

gdzie  $\Phi^{i,j}$  to współczynnik pokrewieństwa pomiędzy osobnikami  $i$  i  $j$ , a  $\Delta_7^{i,j}$  to skondensowany współczynnik Jacquarda (jeżeli nie ma inbredu to  $\Delta_7^{i,j} = \Phi^{i,j} - \rho$ )

# Zadana struktura korelacji

```
library(kinship)
library(MASS)
```

```
id      = c(1,2,3,4,5,6,7,8,9,10,11)
dadid   = c(0,1,0,1,1,3,3,5,7,3,9)
momid   = c(0,0,0,0,0,2,4,0,6,8,10)
sex     = c(1,2,1,2,1,2,1,2,1,2,1)
```

```
families <- makefamid(id,dadid,momid)
```

```
(kmatrix <- makekinship(families,id,dadid,momid))
heatmap(as.matrix(kmatrix), symm=T)
```

```
dieta    = rnorm(11)
mleko    = mvrnorm(1, dieta, kmatrix)
```

```
byki <- data.frame(mleko, dieta, id)
```

```
%(model <- lmekin(mleko~dieta, data=byki, random=~1|id, varlist=list(kmatrix)))
```

# Zależność przestrzenna

W wielu zagadnieniach możemy uznać, że efekty losowe  $\varepsilon_i$  są iid lub, że  $\varepsilon_i$  nie są iid. Dla wielu zagadnień możemy wyliczyć lub oszacować macierz korelacji pomiędzy efektami losowymi.

Przykładowe odstępstwa od iid to zależności typu AR, ARMA itp.

Ciekawą klasą są zależności wynikające z rozmieszczenia przestrzennego obiektów. Pytaniem jest jak zamienić informację o zależności przestrzennej na macierz korelacji. Jest wiele modeli na to, np. dwa popularniejsze to:

- model wykładniczy  $\rho(s, r) = 1 - \exp(-s/p)$ ,
- model Gaussowski  $\rho(s, r) = 1 - \exp(-(s/p)^2)$

# Zależność przestrzenna

```

> library(maps)
> library(MASS)
> library(nlme)
>
> data(world.cities)
>
> (dane = world.cities[world.cities[,2]=="Poland",] )
      name country.etc   pop  lat  long capital
936   Aleksandrow Kujawski   Poland 12214 52.88 18.70      0
937   Aleksandrow Lodzki   Poland 20311 51.82 19.30      0
1455          Andrychow   Poland 21880 49.86 19.34      0
2375          Augustow   Poland 29780 53.84 23.00      0
3436          Barlinek   Poland 14365 53.00 15.20      0
3503          Bartoszyce   Poland 25629 54.25 20.81      0
3802          Bedzin     Poland 57952 50.34 19.13      0

>
> map("world", "Poland")
> map.cities(country = "Poland", capitals = 0, cex=dane[,3])

```

# Zależność przestrzenna

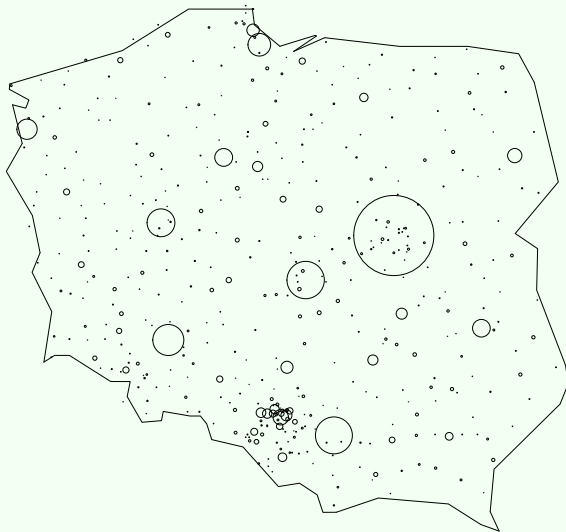
```
> summary(dane)
      miasto      kraj      pop      x
Bialystok   : 1 Length:23   Min.   : 157981   Min.   :49.82
Bielsko-Biala: 1 Class :character 1st Qu.: 202401   1st Qu.:50.30
Bydgoszcz   : 1 Mode  :character Median : 253850   Median :51.24
Bytom       : 1      Mean  : 391722   Mean   :51.68
Cracow      : 1      3rd Qu.: 437747   3rd Qu.:53.07
Czestochowa : 1      Max.   :1634441   Max.   :54.52
(Other)     :17

      y      stolica      zarobki
Min.   :14.53   Min.   :0.00000   Min.   :-1.526983
1st Qu.:18.62   1st Qu.:0.00000   1st Qu.: -0.832104
Median :19.05   Median :0.00000   Median : 0.101643
Mean   :19.37   Mean   :0.04348   Mean   : 0.002358
3rd Qu.:20.57   3rd Qu.:0.00000   3rd Qu.: 0.482895
Max.   :23.16   Max.   :1.00000   Max.   : 2.158704

> cs1Exp <- corExp( 1, form = ~ x + y)
> cs1Exp <- Initialize( cs1Exp, dane)
> mKor = corMatrix( cs1Exp )
> mKor[1:10,1:10]
      Bialystok Bielsko-Biala Bydgoszcz Bytom
Bialystok 1.000000000 0.005075057 0.00579918 0.006195357
Bielsko-Biala 0.005075057 1.000000000 0.03142976 0.578001507
Bydgoszcz 0.005799180 0.031429762 1.00000000 0.054337437
Bytom 0.006195357 0.578001507 0.05433744 1.000000000
Cracow 0.011778950 0.390191998 0.02655567 0.336447988
Czestochowa 0.009513121 0.370379522 0.07675038 0.600554459
Gdansk 0.009262773 0.010478019 0.24886044 0.017969498
Gdynia 0.007976277 0.008838143 0.22459593 0.015187564
Gliwice 0.004954557 0.537901060 0.05577283 0.784028034
Katowice 0.006452980 0.643378847 0.04816606 0.867511332
```

Zarobki/ceny mieszkań/temperatura a rozmieszczenie przestrzenne

# Zależność przestrzenna



# Model hierarchiczny i efekty zagnieżdżone

Jeżeli w modelu występują dwie zmienne grupujące, poziomy tych zmiennych mogą być w relacji „crossed” lub „nested”.

$$y_i = \mu + a_{g1(i)} + b_{g2(i)} + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

$$a \sim \mathcal{N}(0, \sigma_a^2)$$

$$b \sim \mathcal{N}(0, \sigma_b^2)$$

$$\text{Cor}(a_i, a_j) = \rho(i, j)$$

Efekty nested vs. crossed

# Model hierarchiczny i efekty zagnieżdżone

```

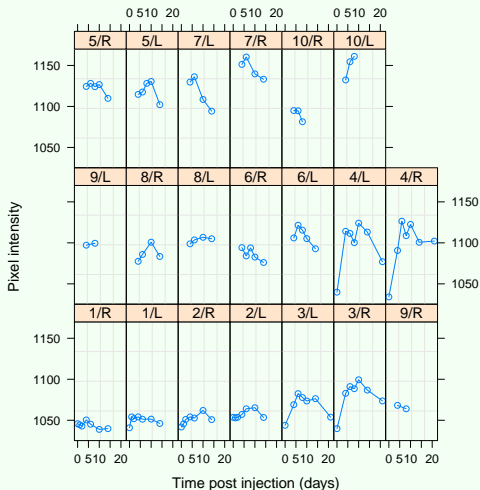
> library(nlme)
> data(Pixel)
> summary(Pixel)
      Dog      Side      day      pixel
1      :14    L:51    Min.   : 0.00    Min.   :1034
2      :14    R:51    1st Qu.: 4.00    1st Qu.:1054
3      :14                    Median : 6.00    Median :1088
4      :14                    Mean   : 7.49    Mean   :1087
5      :10                    3rd Qu.:10.00   3rd Qu.:1110
6      :10                    Max.   :21.00   Max.   :1161
(Other):26

> class(Pixel)
[1] "nmGroupedData" "groupedData"   "data.frame"
> plot(Pixel)

```

Efekty nested vs. crossed

# Model hierarchiczny i efekty zagnieżdżone



# Regresja logistyczna i efekty losowe

Efekty losowe można wprowadzać też do innych zadań regresji, np. do regresji nieliniowej lub do uogólnionych modeli liniowych (mieszanych). W tym przypadku model ma postać np.

$$y_i \sim F(\theta_i)$$

$$\theta_i = \mu + a_{g1(i)} + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_0^2)$$

$$a \sim \mathcal{N}(0, \sigma_a^2)$$

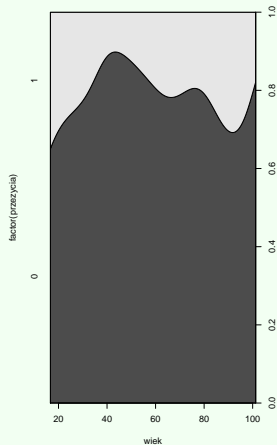
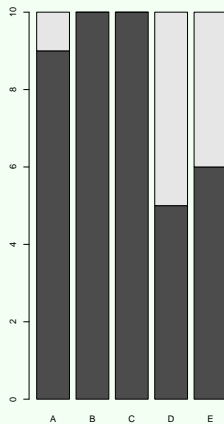
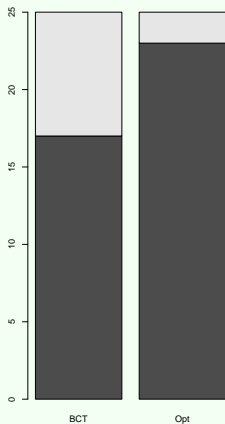
# Regresja logistyczna i efekty losowe

```
> library(PBImisc)
> library(YaleToolkit)
> data(przezycia)
> summary(przezycia)
```

przezycia	terapia	wiek	lekarz
Min. :0.0	BCT:25	Min. : 16.40	A:10
1st Qu.:0.0	Opt:25	1st Qu.: 45.77	B:10
Median :0.0		Median : 55.55	C:10
Mean :0.2		Mean : 56.38	D:10
3rd Qu.:0.0		3rd Qu.: 70.92	E:10
Max. :1.0		Max. :101.50	

```
> gpairs(przezycia)
```

# Regresja logistyczna i efekty losowe



# Literatura



„Practical Regression and Anova using R”, Faraway.



„Extending the Linear Model with R”, Faraway.



„Linear Mixed Models”, Fox.



„SAS for Mixed Models”, Littell, Milliken, Stroup, Wolfinger, Schabenberger.



„Generalized, Linear, and Mixed Models”, McCulloch, Charles, Searle.



„Mixed-Effects Models in S and S-PLUS” Pinheiro, Bates.



„Computing Gaussian Likelihoods and their Derivatives for General Linear Mixed Models”  
Wolfinger, Tobias, Sall.



„lme for SAS PROC MIXED Users” Bates, Pinheiro, Jose.