



# Modele liniowe z efektami stałymi i losowymi

summary(sosny)

	Plot.ID	N	E
Forest - 102c	: 1	Min. :51.21	Min. :14.34
Forest - 106c - 1	: 1	1st Qu.:51.23	1st Qu.:17.59
Forest - 106c - 2	: 1	Median :51.65	Median :18.25
Forest - 107k	: 1	Mean :52.01	Mean :18.25
Forest - 111a-1	: 1	3rd Qu.:52.01	3rd Qu.:19.41
Forest - 111a-2	: 1	Max. :53.94	Max. :19.92
(Other)	:71		

	Soil.type	Age	dens	G
Haplic Podzol	:23	Min. : 3.00	Min. : 2717	Min. : 0.000
Spolic Technosols (Protic)	:20	1st Qu.: 6.00	1st Qu.: 4960	1st Qu.: 1.350
Albic Brunic Arenosol	:12	Median :10.00	Median : 6055	Median : 4.310
Albic Podzols	:12	Mean :10.26	Mean : 6924	Mean : 8.461
Haplic Brunic Arenosol	: 4	3rd Qu.:15.00	3rd Qu.: 8707	3rd Qu.:14.750
Gleyic Podzol	: 2	Max. :20.00	Max. :14000	Max. :29.790
(Other)	: 4			

	Hg	V	AB	BR
Min.	:0.334	Min. : 0.057	Min. : 0.1738	Min. : 0.038
1st Qu.:	1.824	1st Qu.: 6.178	1st Qu.: 6.9883	1st Qu.: 2.008
Median :	3.010	Median : 12.365	Median :11.7229	Median : 3.044
Mean :	3.844	Mean : 31.243	Mean :21.5752	Mean : 5.490
3rd Qu.:	5.598	3rd Qu.: 47.313	3rd Qu.:37.2202	3rd Qu.: 7.959
Max. :	9.096	Max. :118.178	Max. :57.5928	Max. :23.857

	FL	ST
Min.	: 0.1284	Min. : 0.057
1st Qu.:	2.5774	1st Qu.: 2.360
Median :	3.8398	Median : 4.404
Mean :	4.3740	Mean :11.785
3rd Qu.:	5.3993	3rd Qu.:18.756
Max. :	11.8670	Max. :41.173

sosny<-read.csv(,sosny.csv', sep=,;')

Zbiór danych: sosny

dane o biomase drzewostanów sosnowych w wieku 3-20 lat

N, E – koordynaty

Age – wiek [lata]

dens – zagęszczenie [szt. ha<sup>-1</sup>]

G – pole powierzchni drzew [m<sup>2</sup> ha<sup>-1</sup>]

Hg – wysokość górna d-stanu [m]

V – miąższość [m<sup>3</sup> ha<sup>-1</sup>]

AB – biomasa części nadziemnej [t ha<sup>-1</sup>]

BR – gałęzi

FL – liści

ST -pni

Forest Ecology and Management 409 (2018) 74–83



Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: [www.elsevier.com/locate/foreco](http://www.elsevier.com/locate/foreco)



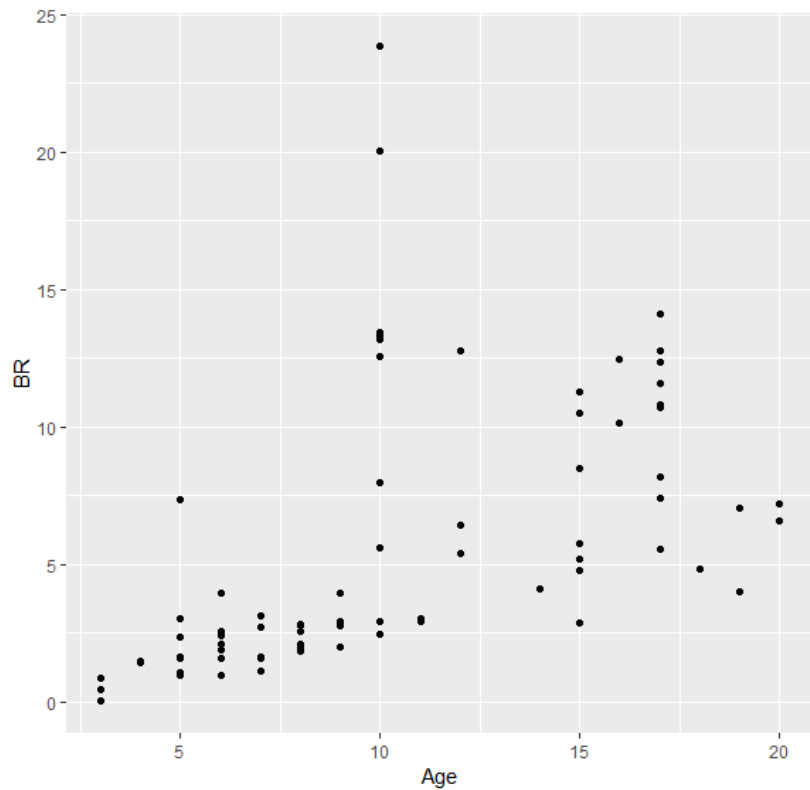
How do tree stand parameters affect young Scots pine biomass? – Allometric equations and biomass conversion and expansion factors

Andrzej M. Jagodziński<sup>a,b,\*</sup>, Marcin K. Dyderski<sup>a,b</sup>, Kamil Gęsikiewicz<sup>a</sup>, Paweł Horodecki<sup>a</sup>, Agnieszka Cysewska<sup>b</sup>, Sylwia Wierczyńska<sup>b</sup>, Karol Maciejczyk<sup>b</sup>

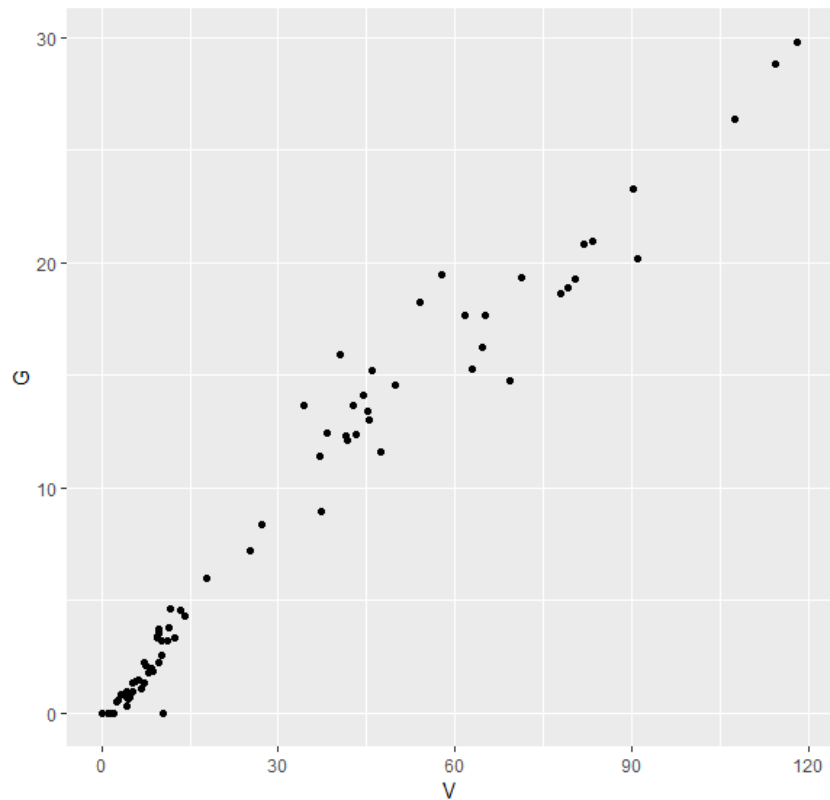
<sup>a</sup> Institute of Dendrology, Polish Academy of Sciences, Parkowa 5, 62-035 Kórnik, Poland

<sup>b</sup> Poznań University of Life Sciences, Faculty of Forestry, Department of Game Management and Forest Protection, Wojska Polskiego 71c, 60-625 Poznań, Poland



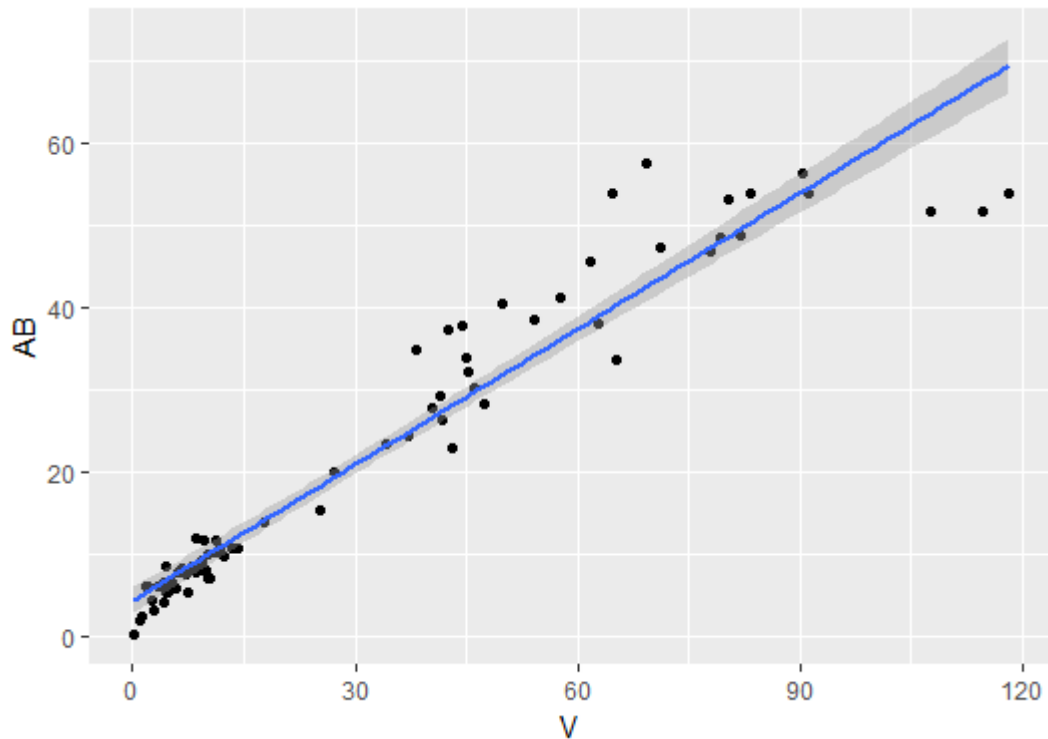


`ggplot(sosny, aes(x=Age,y=BR))+geom_point()`



`ggplot(sosny, aes(x=V,y=AB))+geom_point()`

```
ggplot(sosny, aes(x=V,y=AB))+geom_point()+geom_smooth(method='lm')
```



# Model liniowy

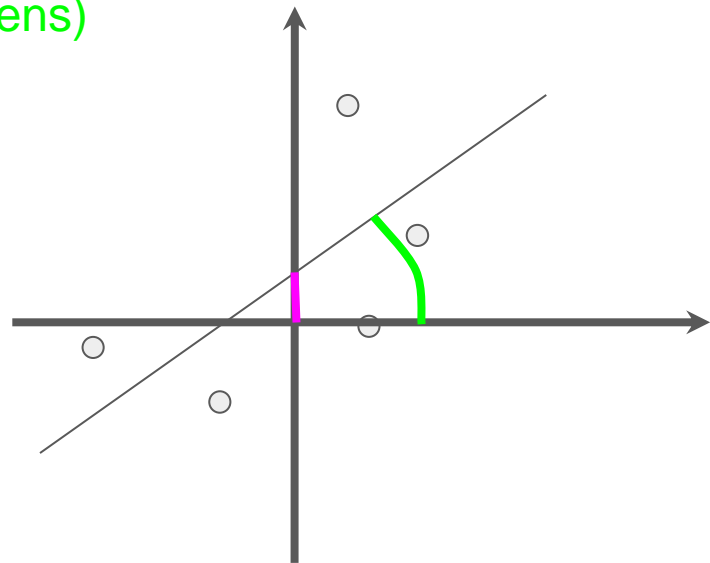
$$y=a*x+b$$

a - współczynnik kierunkowy, slope, regression coefficient, beta

nachylenie linii regresji (kąt – w zasadzie jego tangens)

b - wyraz wolny, intercept

punkt przecięcia z osią Y, położenie linii



# Biologiczne znaczenie - effect size!

<https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf>

<https://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108?needAccess=true>

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond “ $p < 0.05$ ”, The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913

nieistotny biologicznie efekt (3%) -  $p < 0.00001$  przy  $n=300$

istotny efekt (800%) -  $p > 0.05$  przy  $n=3$

# summary(lm(AB~V,data=sosny))

```
> summary(lm(AB~V,data=sosny))

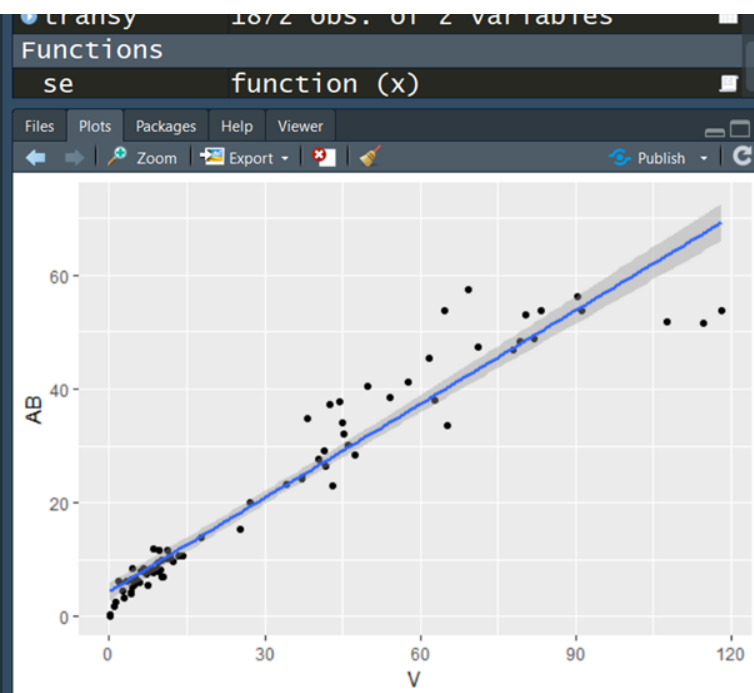
Call:
lm(formula = AB ~ V, data = sosny)

Residuals:
    Min       1Q   Median       3Q      Max
-15.7506  -1.7161  -0.3876   1.1526  15.1380

Coefficients:
            Estimate Std. Error t value Pr(> |t|)
(Intercept)  4.39843    0.79045   5.564 3.88e-07 ***
V            0.54978    0.01785  30.794 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

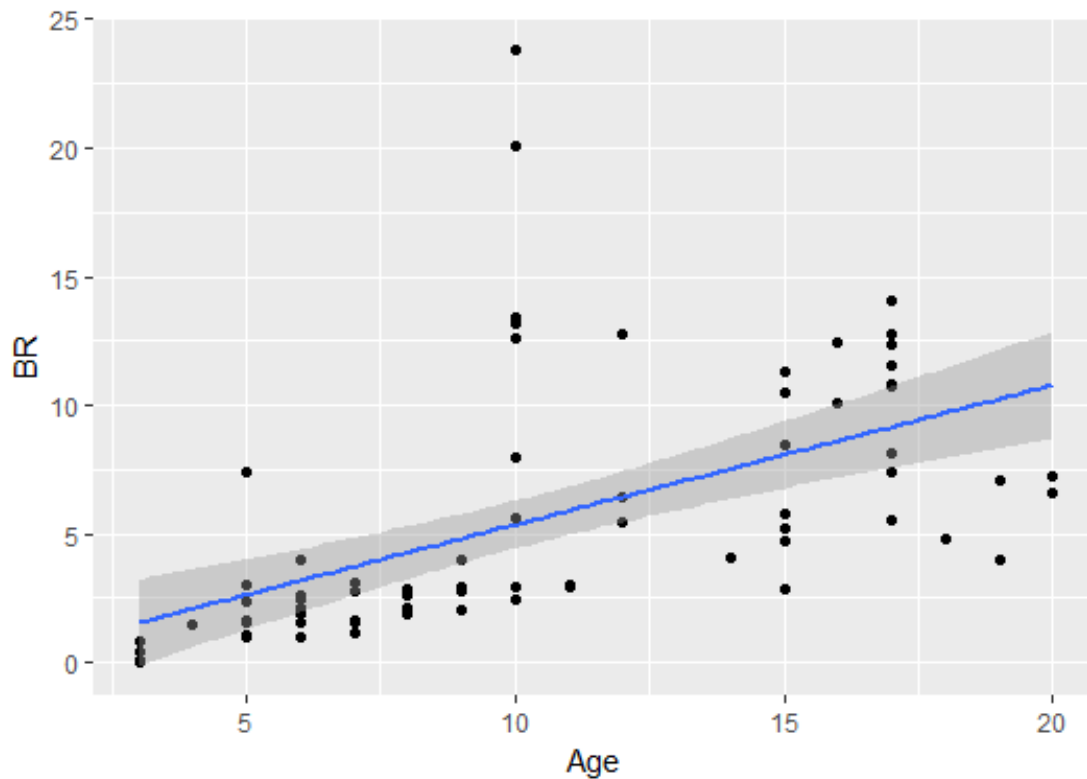
Residual standard error: 4.915 on 75 degrees of freedom
Multiple R-squared:  0.9267,    Adjusted R-squared:  0.9257
F-statistic: 948.3 on 1 and 75 DF,  p-value: < 2.2e-16

> |
```



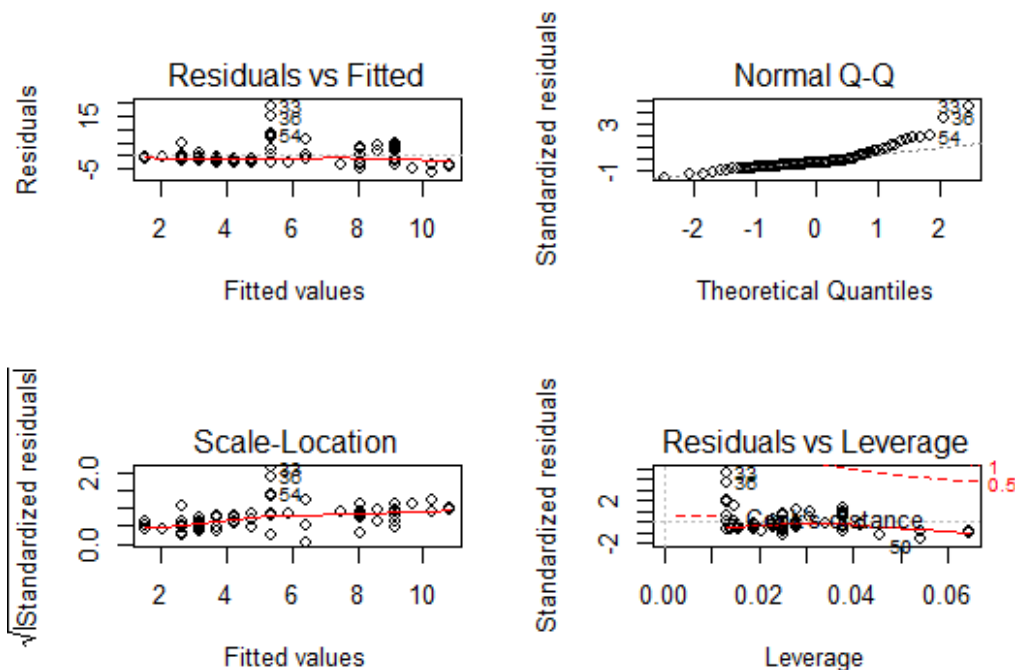


a tutaj? lepszy czy gorszy?

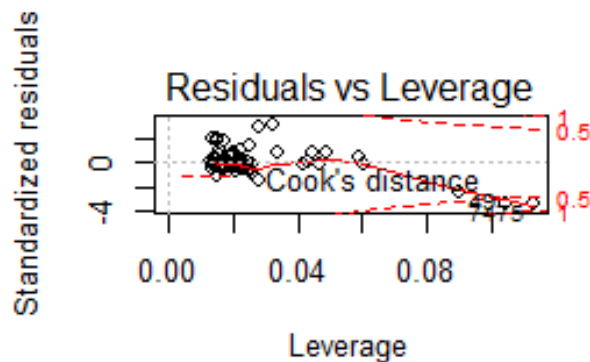
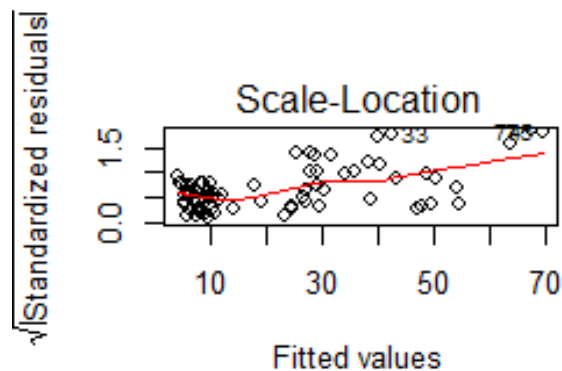
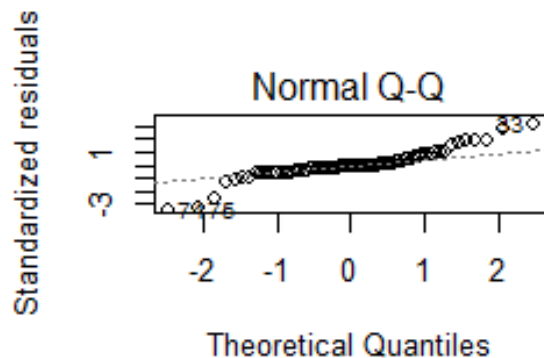
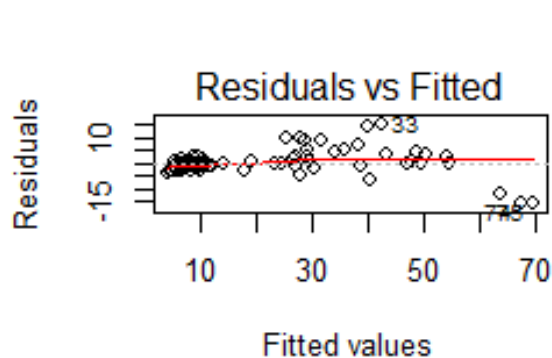


# Diagnostyka modeli – pakiet bazowy

```
par(mfrow=c(2,2)) #podział wykresu na 4  
plot(lm(BR~Age,data=sosny))
```

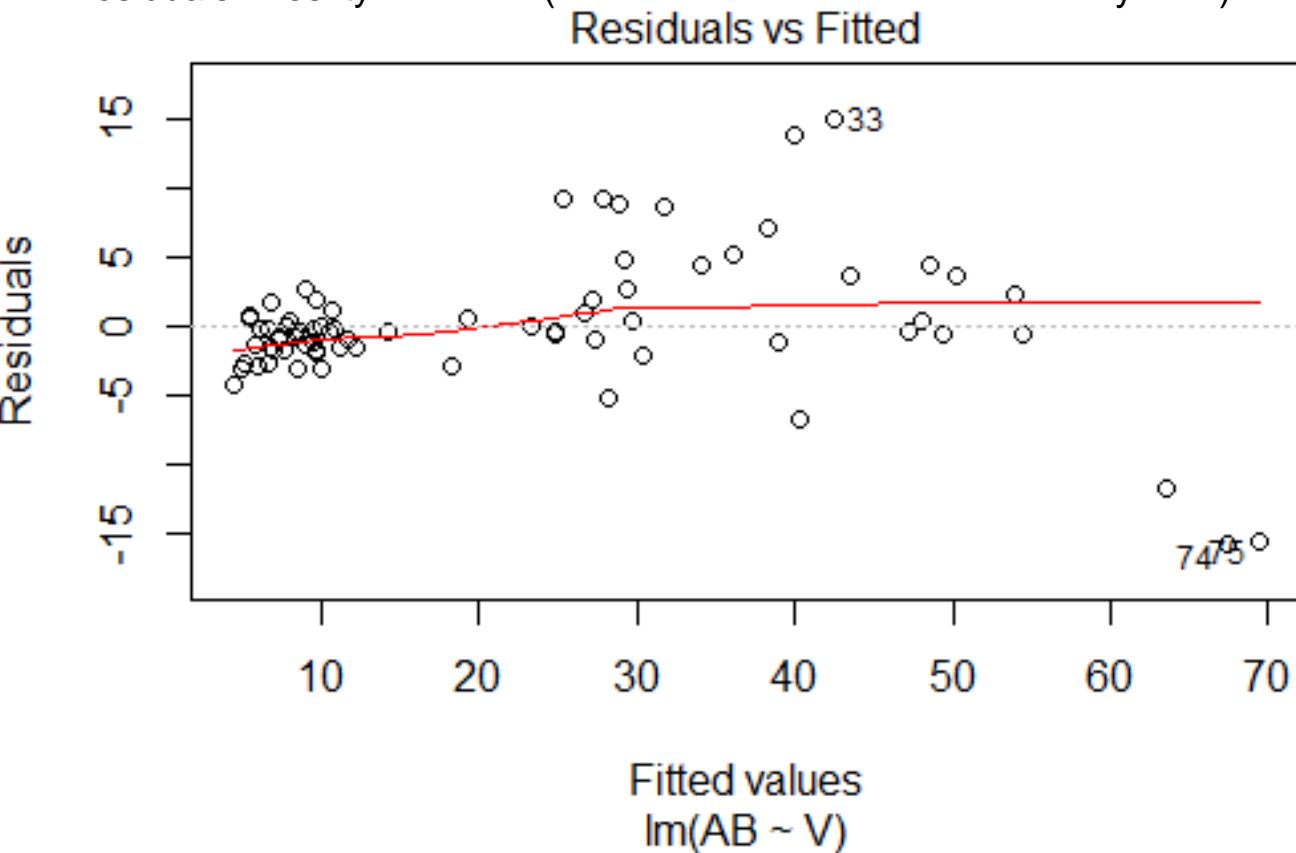


# Jak wygląda dobry model? $AB \sim V \dots$

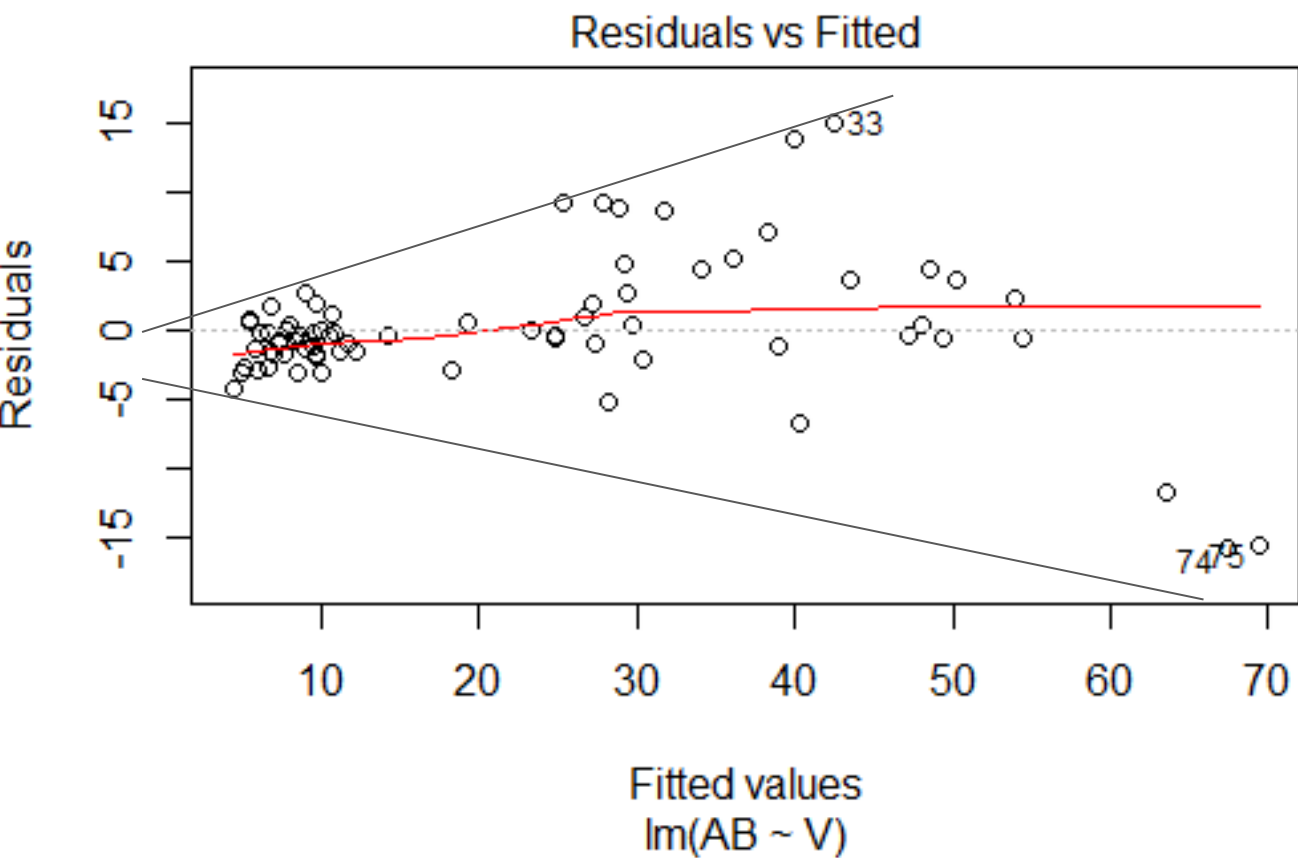


Heteroskedastyczność – większy rozrzut residuals przy większych wartościach zmiennej zależnej

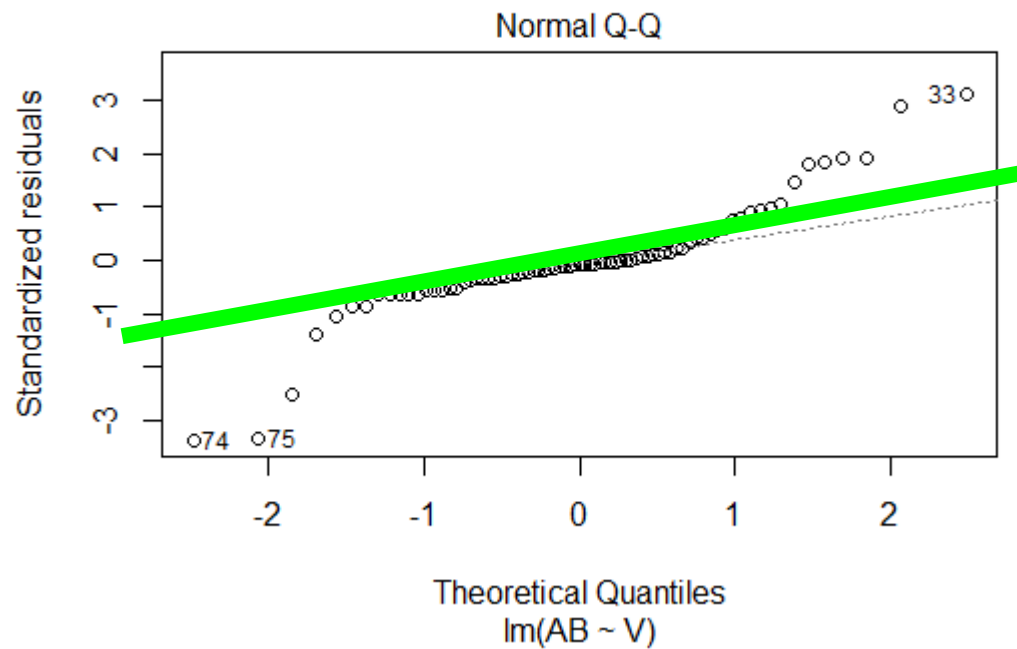
Residuals – reszty z modelu (wartość modelowana minus rzeczywista)



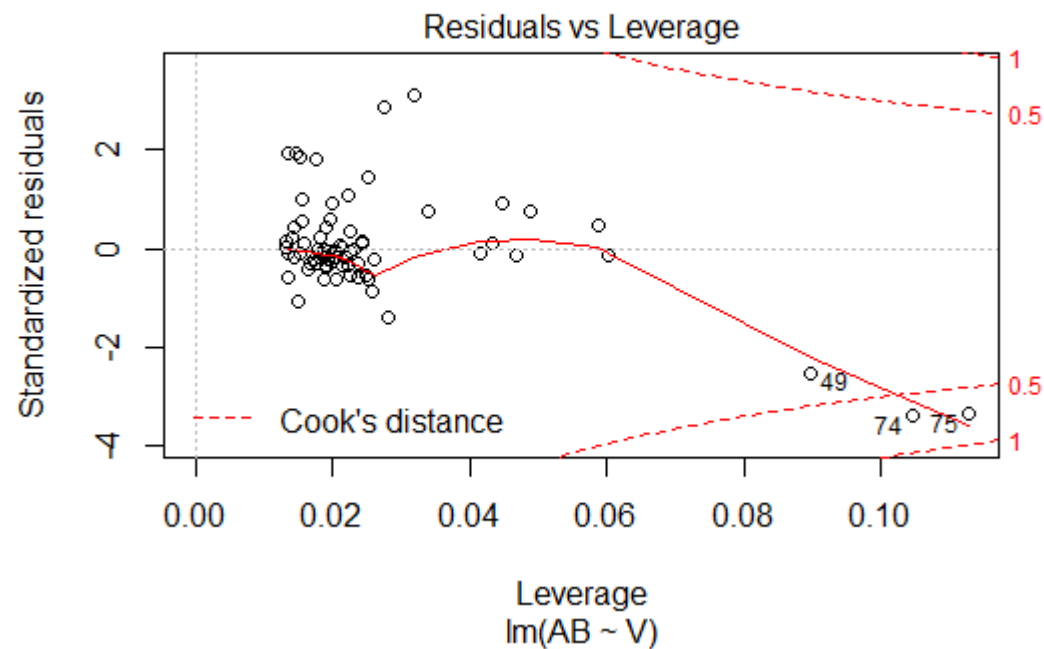
heteroskedastyczność!



# normalność reszt



# Obserwacje odstające



# Co się stało?

model gorszy wg  $R^2$  ( $BR \sim age$ ) okazał się mieć mniej problemów

model lepszy ( $AB \sim V$ ) - heteroskedastyczność



# Prostszy sposób oceny jakości modeli

- pakiet DHARMa

```
model<-lm(BR~Age,data=sosny)
```

```
library(DHARMa)
```

```
model.sr<-simulateResiduals(model)
```

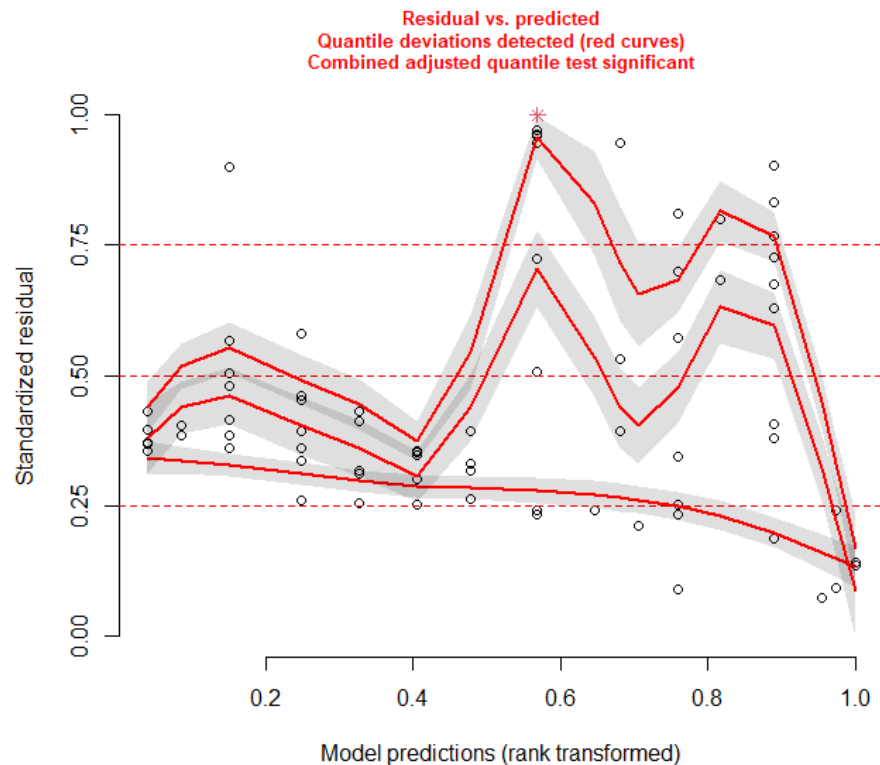
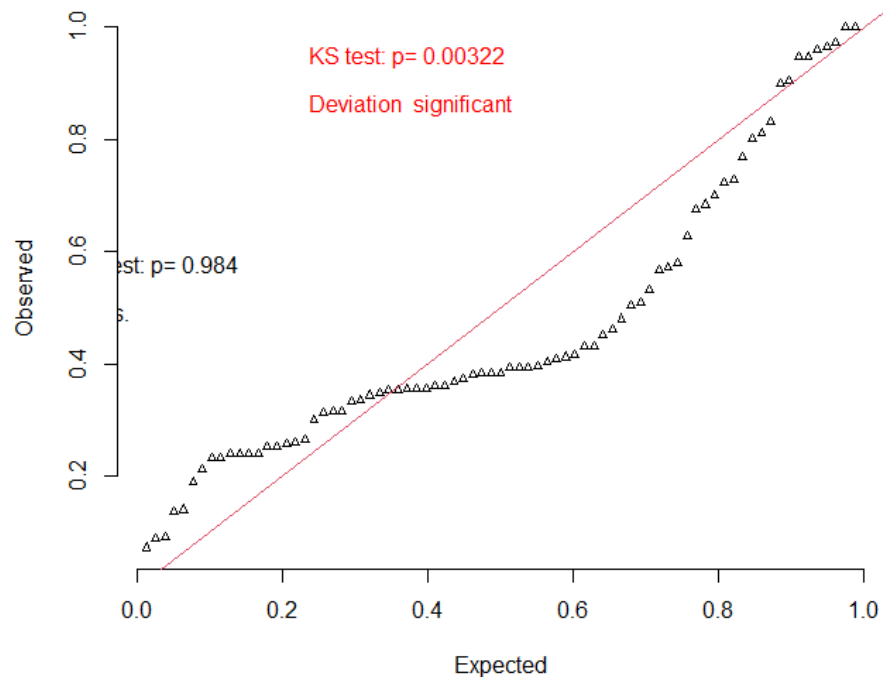
```
plot(model.sr)
```

- <https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html>

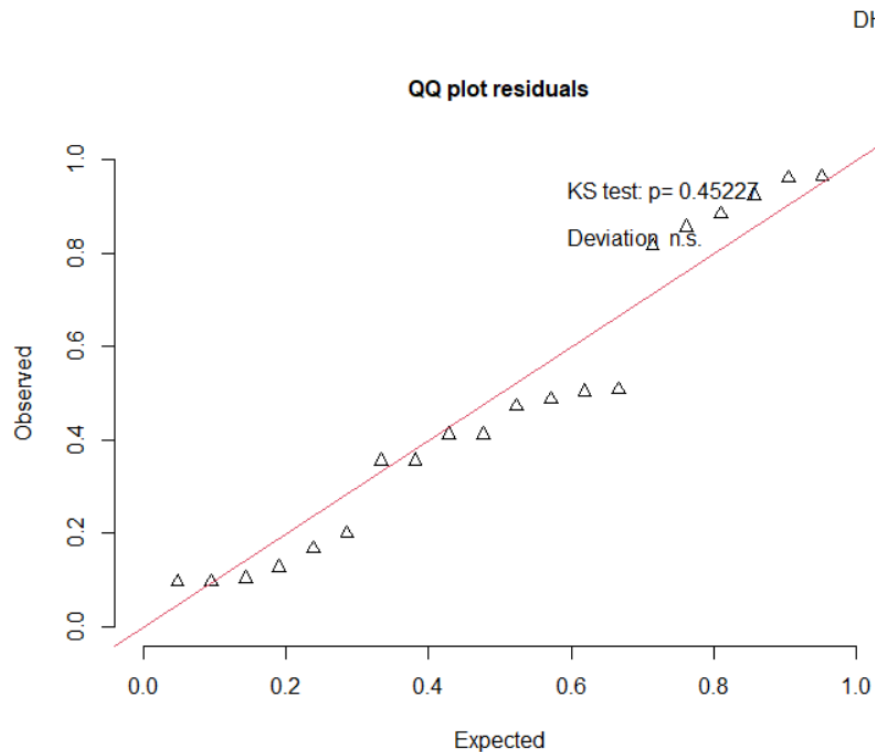
# Problem z rozkładem reszt

DHARMA residual diagnostics

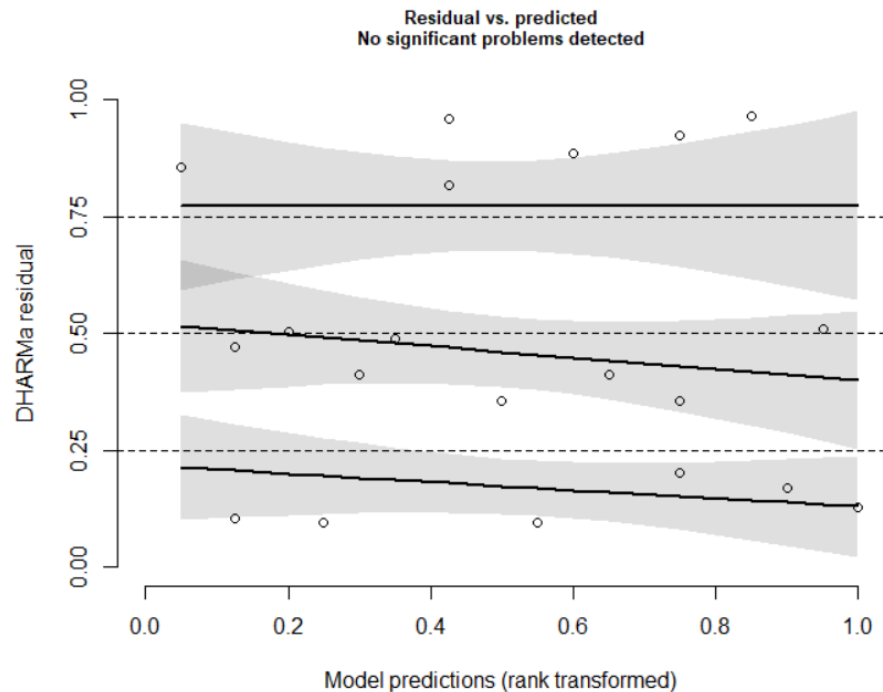
QQ plot residuals



# przykład poprawnego dopasowania



DHARMA residual



# Czy można tylko oceniać na podstawie statystyk opisowych

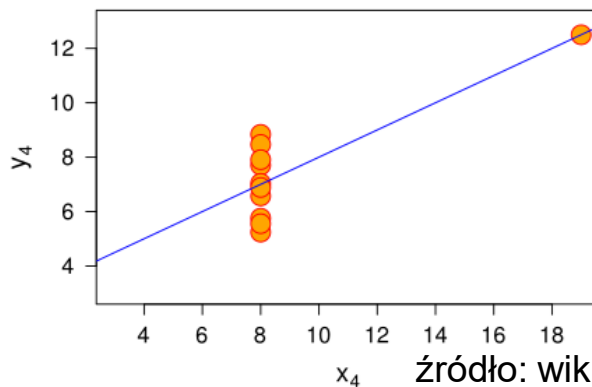
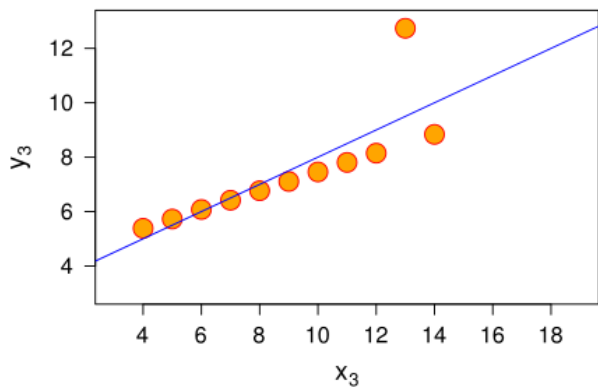
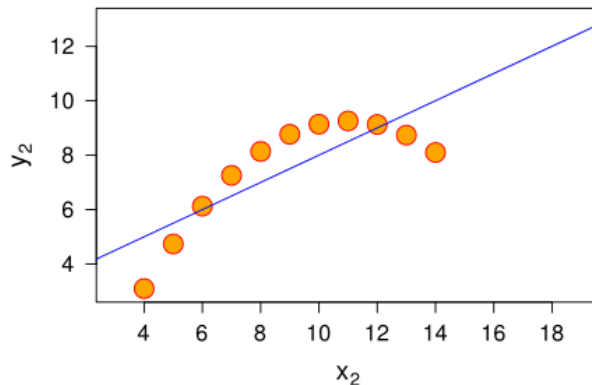
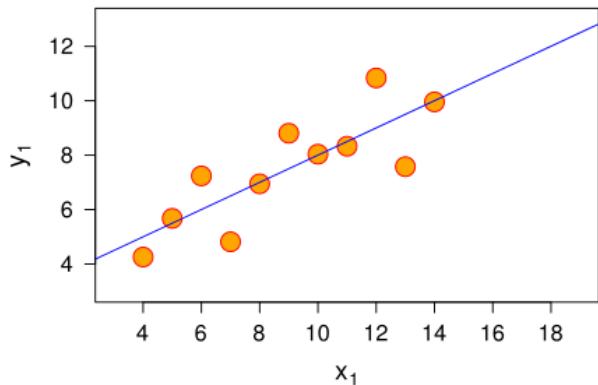
NIE!

$R^2$  nie uwzględnia kształtu rozkładu, heteroskedastyczności, outlierów...

kwadrat Anscombe'a

# Kwadrat Ascombe'a

średnia  $y=7,5$  średnia  $x=9$   
współczynnik  $r^2=0,816$   
równanie regresji:  
 $y=3+0,5 \cdot x$



# Co zrobić jak nie jest tak różowo?

transformacje (log, skalowanie...)

inny typ modelu

włączenie struktur zależności (modele z efektami losowymi – po przerwie)

inny typ rozkładu – uogólnienie LM na inne – GLM (jutro:)

# Transformacje

wielomian 2. stopnia ( $x^2$ ) – `poly(x,2)`

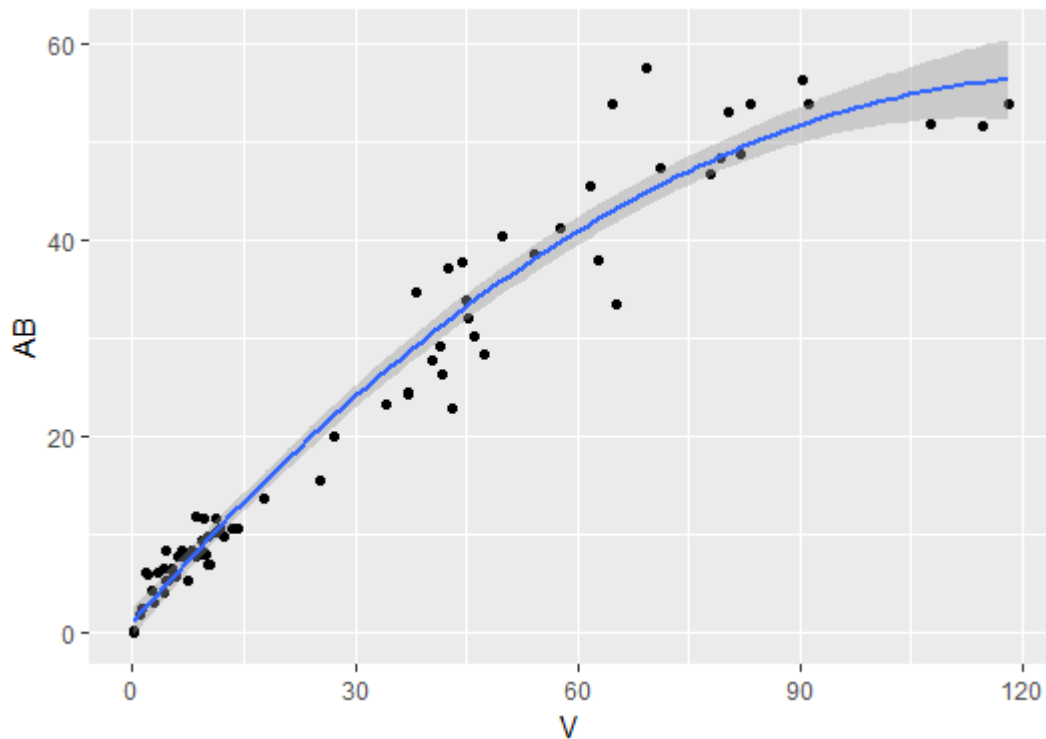
logarytm - `log(x)`

pierwiastek - `sqrt(x)`

skalowanie i centrowanie `scale(x)`

coś za coś – lepiej dopasowany model vs. trudniejsza interpretacja

# Równania kwadratowe $y=ax^2+bx+c$



Wielomian drugiego stopnia  
Model kwadratowy  
Parabola  
polynomial

model:  
`lm(y~poly(x,2), data)`

```
ggplot(sosny, aes(x=V,y=AB))+geom_point()+  
geom_smooth(method='lm',formula=y~poly(x,2))
```



```
> summary(lm(AB~poly(V,2),data=sosny))
```

Call:

```
lm(formula = AB ~ poly(V, 2), data = sosny)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-9.6997	-1.9382	-0.1007	1.6296	12.7176

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	21.5752	0.4149	52.004	< 2e-16	***
poly(V, 2)1	151.3404	3.6405	41.571	< 2e-16	***
poly(V, 2)2	-28.8228	3.6405	-7.917	1.86e-11	***

---

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

Residual standard error: 3.64 on 74 degrees of freedom

Multiple R-squared: 0.9603, Adjusted R-squared: 0.9592

F-statistic: 895.4 on 2 and 74 DF, p-value: < 2.2e-16

```
> | AB=-28,8228*V^2+151,3404*V+21,5752
```

# model z logarytmem

```
mod.log<-lm(log(AB)~V,data=sosny)
summary(mod.log)
```

Call:  
lm(formula = log(AB) ~ V, data = sosny)

Residuals:

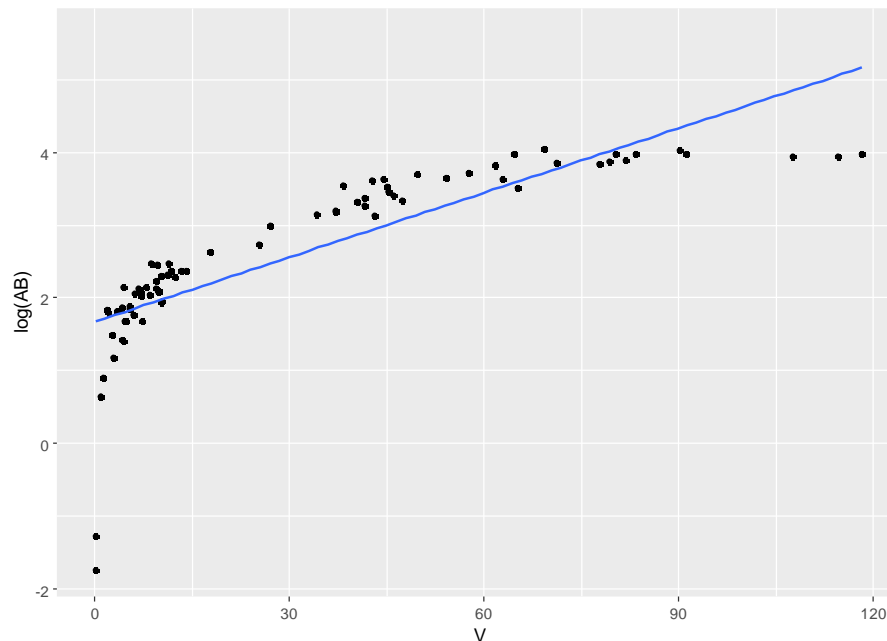
	Min	1Q	Median	3Q	Max
	-3.4248	-0.1294	0.1824	0.3694	0.7462

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.673302	0.109323	15.31	<2e-16 ***
V	0.029567	0.002469	11.97	<2e-16 ***

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

Residual standard error: 0.6797 on 75 degrees of freedom  
Multiple R-squared: 0.6566, Adjusted R-squared: 0.652  
F-statistic: 143.4 on 1 and 75 DF, p-value: < 2.2e-16



```
> mod.log<-lm(log(AB)~log(V),data=sosny)
> summary(mod.log)
```

Call:  
lm(formula = log(AB) ~ log(V), data = sosny)

Residuals:

Min	1Q	Median	3Q	Max
-0.39549	-0.13620	-0.02235	0.10353	0.76807

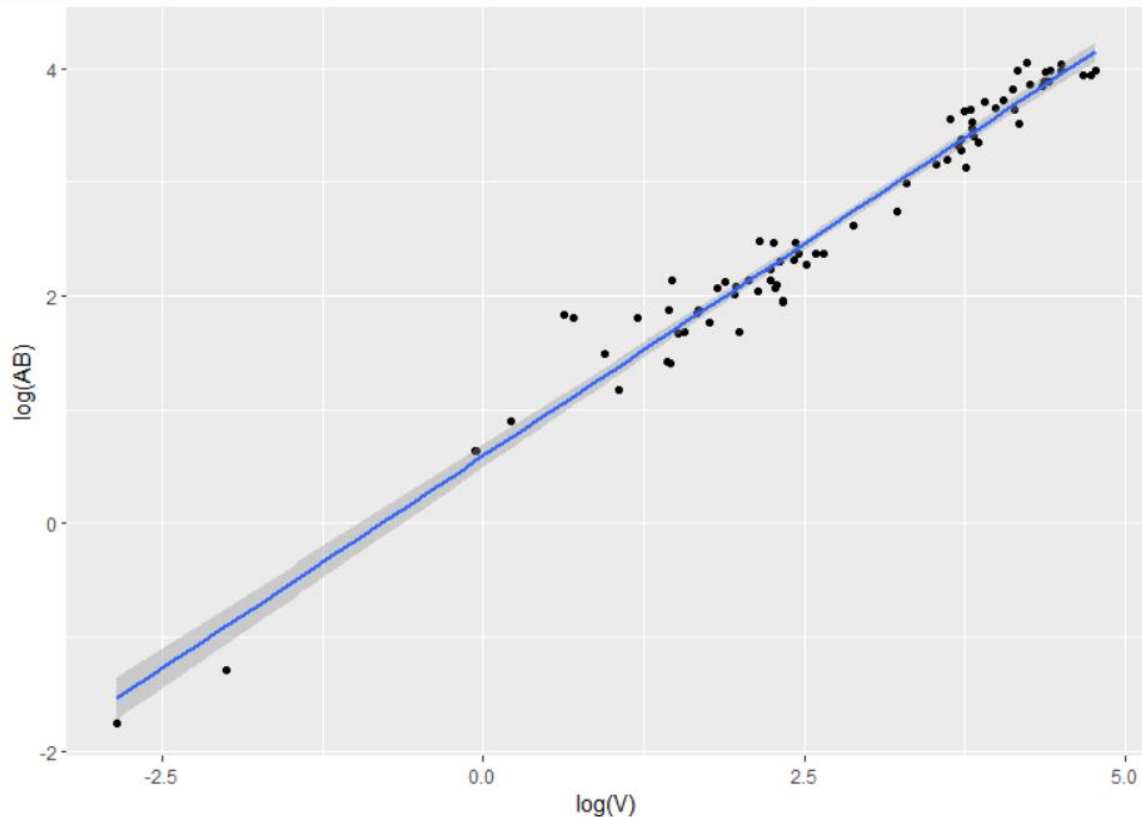
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.59510	0.05019	11.86	<2e-16 ***
log(V)	0.74524	0.01629	45.75	<2e-16 ***

---

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

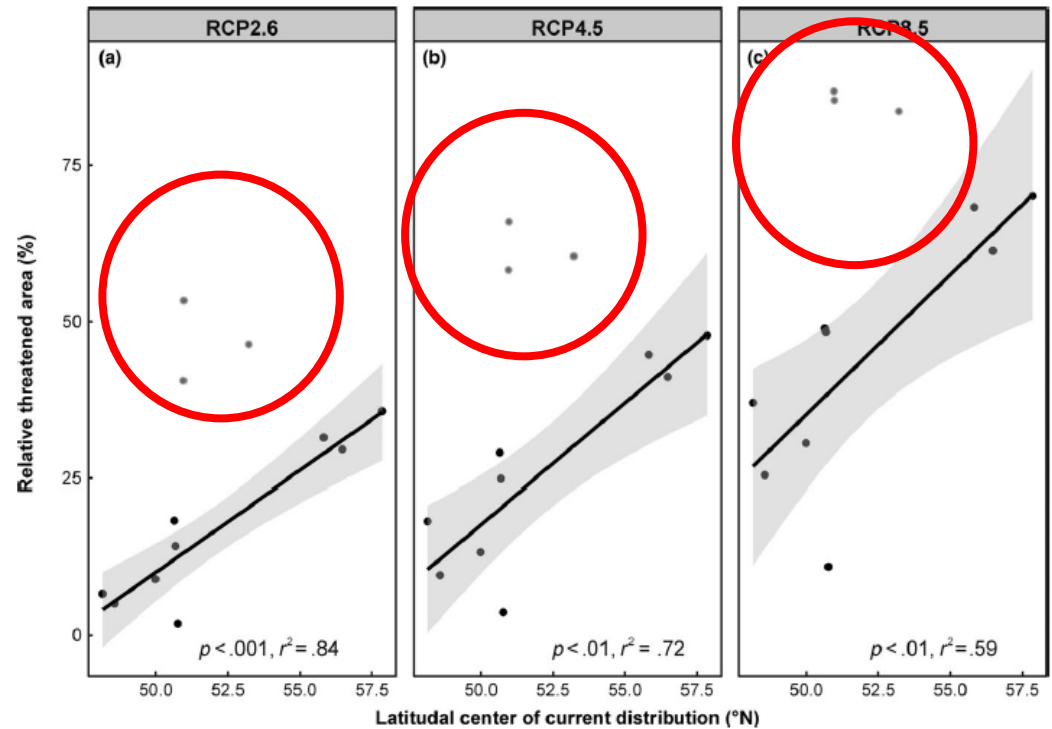
Residual standard error: 0.2157 on 75 degrees of freedom  
Multiple R-squared: 0.9654, Adjusted R-squared: 0.9649  
F-statistic: 2093 on 1 and 75 DF, p-value: < 2.2e-16



# Outliery

obserwacje odstające

wyjątki




**FIGURE 7** The relationship between threatened proportion of current distribution and latitudinal center of distribution for each climate change scenario: optimistic (RCP2.6; panel a), moderate (RCP4.5; panel b), and pessimistic (RCP8.5; panel c) calculated for averaged GCMs outputs. Three species outlying from the trend—*L. decidua*, *P. menziesii*, and *Q. rubra* were removed from the regression analysis and are represented by gray dots. Gray areas around the regression lines represent standard errors (SE) of each model

Received: 20 July 2017 | Accepted: 30 August 2017  
DOI: 10.1111/gcb.13925

PRIMARY RESEARCH ARTICLE

WILEY | Globa

How much does climate change threaten European forest tree species distributions?

Marcin K. Dyderski<sup>1,2</sup> | Sonia Paż<sup>3</sup> | Lee E. Frelich<sup>4</sup> | Andrzej M. Jagodziński<sup>1,2</sup> 

# Formalna podstawa?

Outlier wynikający z błędu pomiaru

Outlier wynikający ze struktury danych (filogeneza, biologia, układ doświadczenia)

Outlier będący poprawnym pomiarem

Usunąć? Zostawić? Co jeśli Recenzent się uprze przy konieczności uzasadnienia?

# Statystyki liczbowe

For inspection of outliers we checked Cook's distances and leverage values using *hat* statistics implemented in the `base::influence()` function. To identify outliers we used Bonferroni *p*-values for Studentized residuals t-tests using the `car::outlierTest()` function. After visual interpretation of diagnostic plots both for calcareous and granite grasslands we adopted 0.15 as threshold leverage values, as most of the observations had relatively low leverage values, ranging from 0.07 in calcareous and from 0.05 to 0.15 in granite grasslands. After that, we excluded two outliers: plot no. 71 (elevation 1039 m a.s.l.) from calcareous and plot no. 57 (elevation 2123 m a.s.l.) from granite grasslands due to high leverage values (0.20 and 0.24, respectively) and due to the `car::outlierTest()`  $p < 0.05$  in both cases.

Źródło: Czortek et al. 2018. Cessation of livestock grazing and windthrow drive a shift in plant species composition in the Western Tatra Mts. *Tuexenia* 38: 177–196. [https://www.zobodat.at/pdf/Tuexenia\\_NS\\_38\\_0177-0196.pdf](https://www.zobodat.at/pdf/Tuexenia_NS_38_0177-0196.pdf)

# Rozważmy taki model

```
> model<-lm(BR~Age, data=sosny)  
> summary(model)
```

Call:  
lm(formula = BR ~ Age, data = sosny)

Residuals:

Min	1Q	Median	3Q	Max
-6.2556	-2.2313	-1.0834	0.7993	18.5083

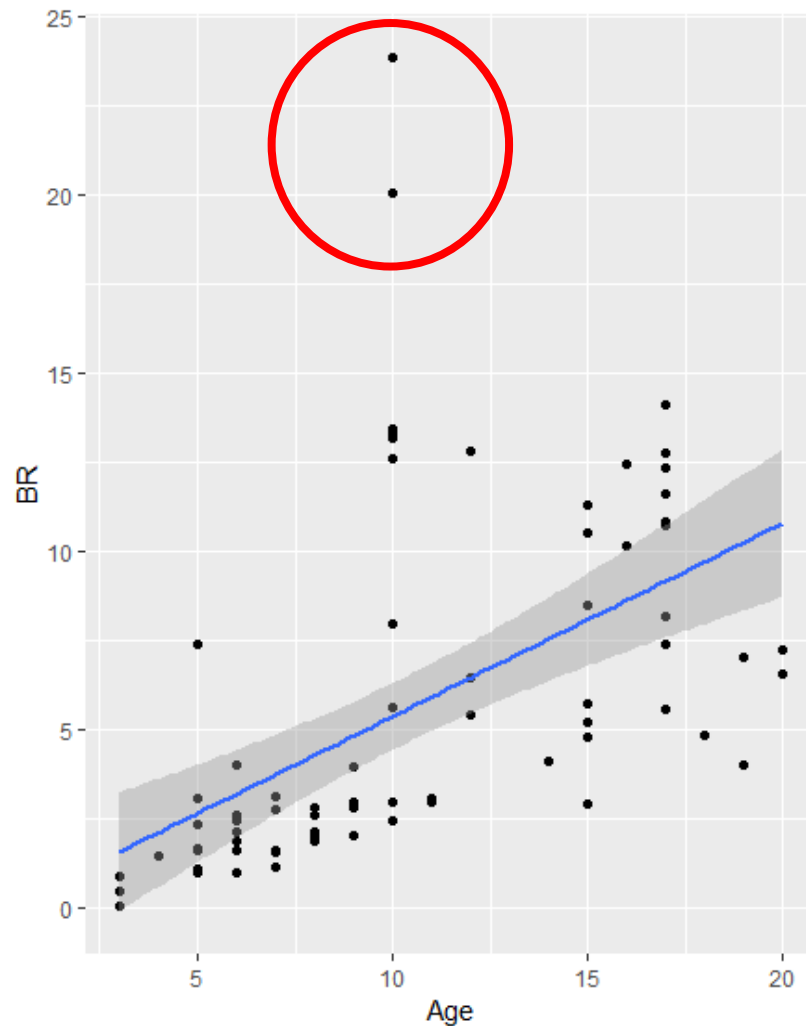
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.08740	1.08768	-0.080	0.936
Age	0.54361	0.09568	5.681	2.41e-07 ***

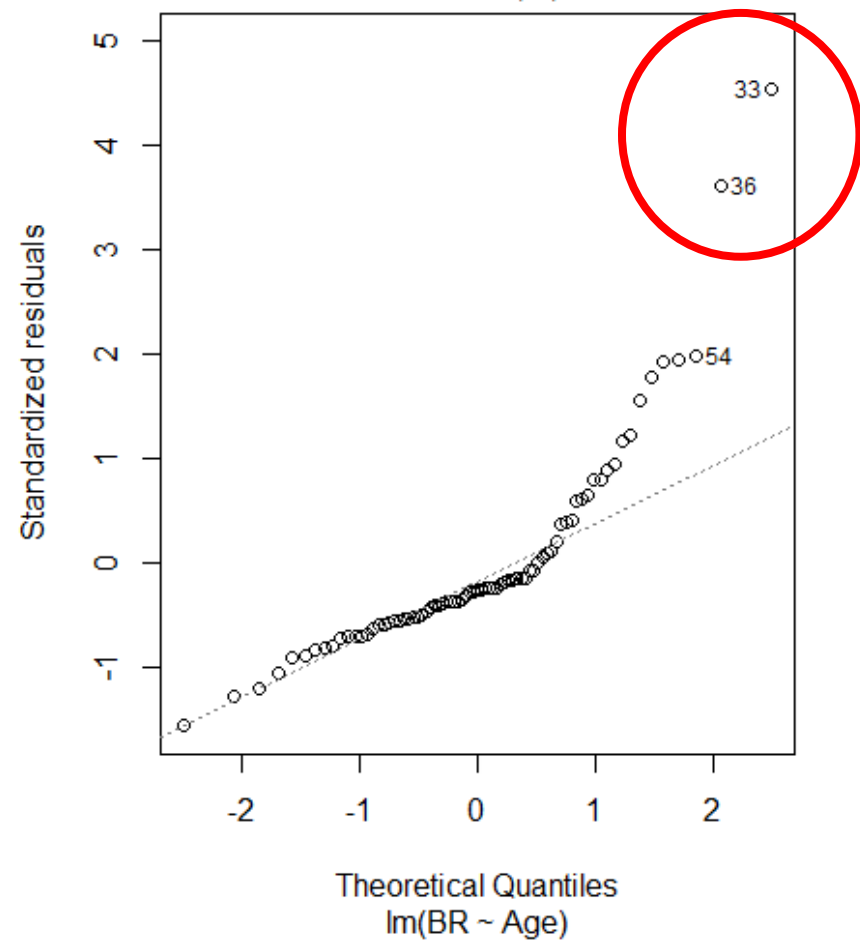
---

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

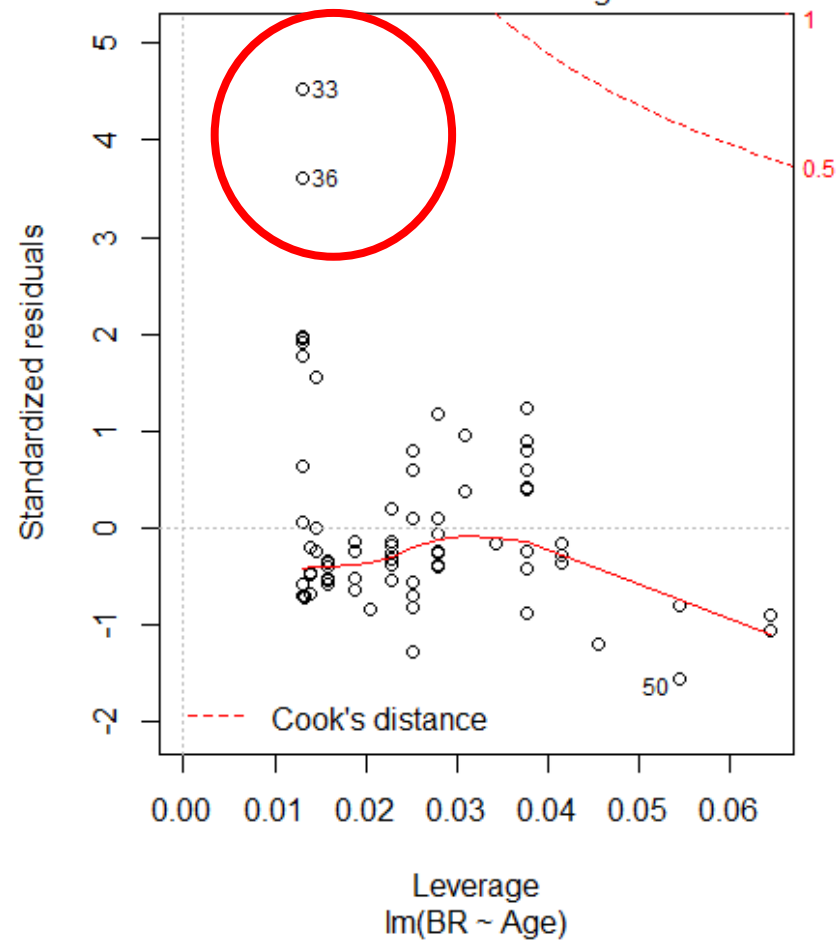
Residual standard error: 4.11 on 75 degrees of freedom  
Multiple R-squared: 0.3009, Adjusted R-squared: 0.2916  
F-statistic: 32.28 on 1 and 75 DF, p-value: 2.411e-07



Normal Q-Q



Residuals vs Leverage





```
as.data.frame(influence(model))
      hat coefficients..Intercept. coefficients.Age  sigma  wt.res
1  0.04155579    -3.933644e-02  2.900932e-03  4.136538 -0.706535320
2  0.04155579    -8.295232e-02  6.117460e-03  4.133604 -1.489935320
3  0.04155579    -8.381529e-02  6.181101e-03  4.133525 -1.505435320
4  0.04155579    -6.202961e-02  4.574479e-03  4.135273 -1.114135320
5  0.04155579    -6.202961e-02  4.574479e-03  4.135273 -1.114135320
6  0.03422738    -3.269323e-02  2.320786e-03  4.136651 -0.660547628
7  0.03422738    -3.136678e-02  2.226626e-03  4.136709 -0.633747628
8  0.02798310    -1.267398e-02  8.554936e-04  4.137246 -0.291659937
9  0.02798310    1.797893e-02  -1.213578e-03  4.137101  0.413740063
....
30 0.01575502    -5.745441e-02  2.754013e-03  4.129255 -2.212896862
31 0.01384724    -1.722269e-02  5.882398e-04  4.136193 -0.849509170
32 0.01384724    -3.814717e-02  1.302914e-03  4.131521 -1.881609170
33 0.01302358    2.706275e-01  -2.640268e-03  3.525303 18.508278522
34 0.01302358    3.444622e-03  -3.360607e-05  4.137297  0.235578522
35 0.01302358    1.143815e-01  -1.115917e-03  4.034866  7.822578522
36 0.01302358    2.149178e-01  -2.096759e-03  3.762979 14.698278522
37 0.01302358    -4.253562e-02  4.149817e-04  4.123362 -2.909021478
38 0.01302358    3.816302e-02  -3.723222e-04  4.126102  2.609978522
39 0.01302358    -3.526558e-02  3.440544e-04  4.127752 -2.411821478
40 0.01462865    2.132902e-02  6.081039e-03  4.069968  6.352053905
41 0.01462865    -3.423447e-03  -9.760465e-04  4.135665 -1.019546095
42 0.01462865    5.553506e-06  1.583340e-06  4.137388  0.001653905
43 0.02057022    2.738037e-02  -7.104097e-03  4.117704 -3.431870712
44 0.02516719    3.927169e-02  -7.544299e-03  4.123642 -2.862183020
45 0.02516719    -3.319243e-02  6.376441e-03  4.127573  2.419116980
46 0.02516719    -5.488588e-03  1.054387e-03  4.137120  0.400016980
```

zwraca nam wartości statystyki `hat`,  
 wartości współczynników przy odrzuceniu  
 obserwacji, wartość SD po odrzuceniu  
 obserwacji oraz ważone reszty  
 (`wt_residuals`)

można coś zobaczyć, ale nie zawsze;)

mówi o wpływie na współczynniki modelu  
 (`coefficients`)

```
library(car)
```

```
outlierTest(model)
```

	rstudent	unadjusted	p-value	Bonferonni	p
--	----------	------------	---------	------------	---

33	5.284650	1.2239e-06	0.00009424
----	----------	------------	------------

36	3.931709	1.8820e-04	0.01449200
----	----------	------------	------------

test t-studenta na istotność średniego przesunięcia od linii regresji - daje p-value surowe oraz po poprawce Bonferonniego

Wskazuje które obserwacje można roważyc jako wpływowe (ew. usunąć)

```
model<-lm(BR~Age, data=sosny)
summary(model)
```

Call:  
lm(formula = BR ~ Age, data = sosny)

Residuals:

Min	1Q	Median	3Q	Max
-6.2556	-2.2313	-1.0834	0.7993	18.5083

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.08740	0.08768	-0.080	0.936
Age	0.54361	0.09568	5.681	2.41e-07 ***

---

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

Residual standard error: 4.11 on 75 degrees of freedom  
Multiple R-squared: 0.3009, Adjusted R-squared: 0.2916  
F-statistic: 32.28 on 1 and 75 DF, p-value: 2.411e-07

```
model2<-lm(BR~Age, data=sosny[-c(33,36),])
summary(model2)
```

Call:  
lm(formula = BR ~ Age, data = sosny[-c(33, 36), ])

Residuals:

Min	1Q	Median	3Q	Max
-5.8548	-1.7888	-0.6365	0.8509	8.5087

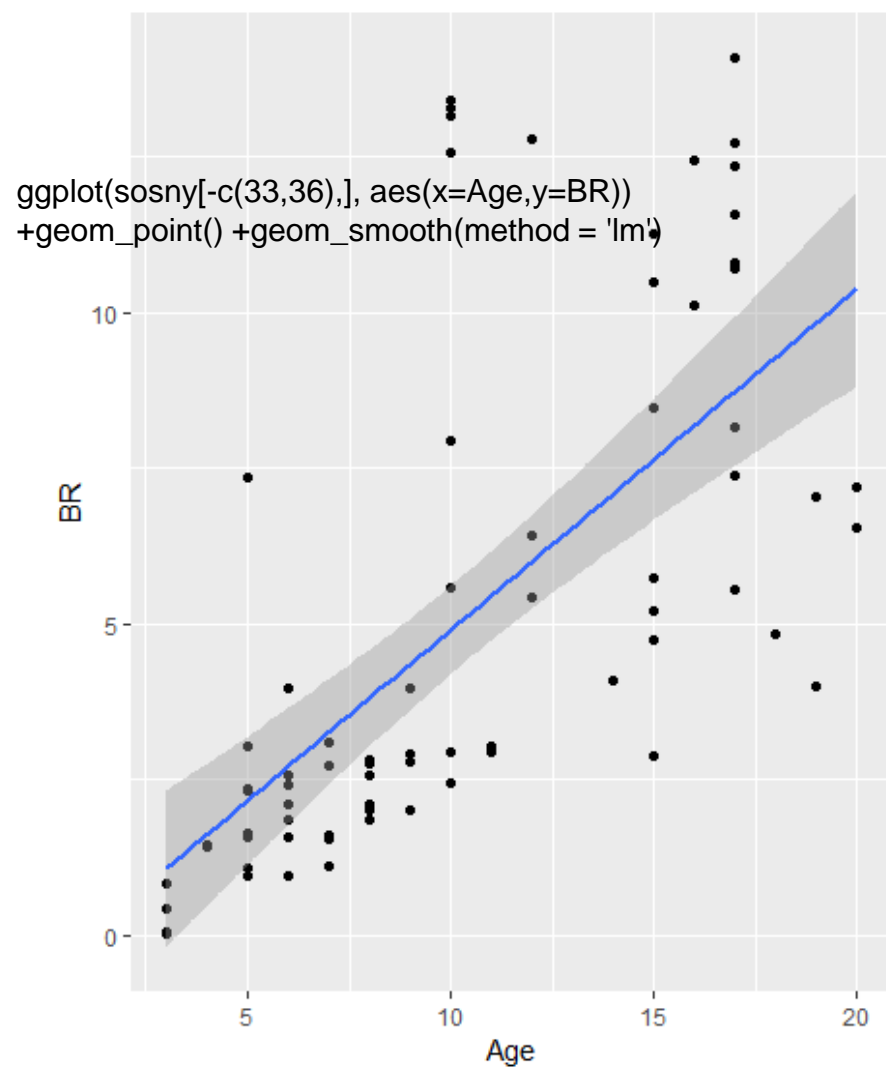
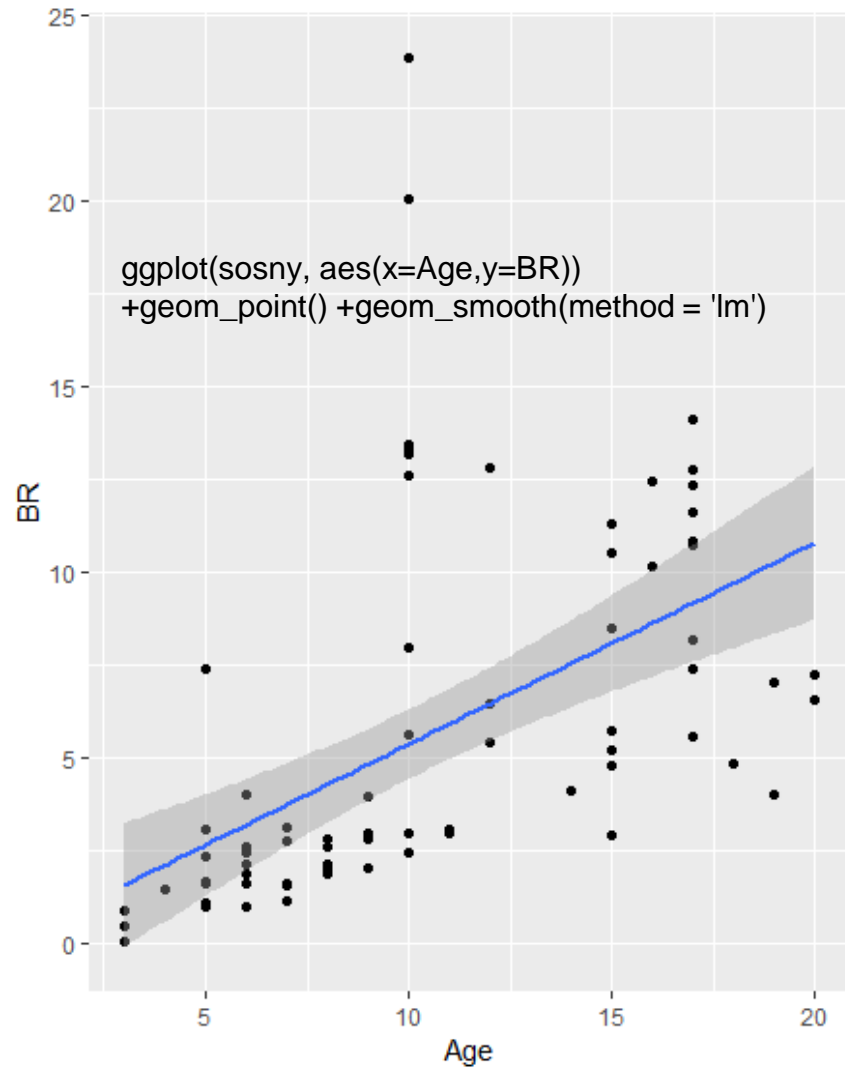
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.57944	0.81816	-0.708	0.481
Age	0.54841	0.07176	7.643	6.6e-11 ***

---

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

Residual standard error: 3.082 on 73 degrees of freedom  
Multiple R-squared: 0.4445, Adjusted R-squared: 0.4369  
F-statistic: 58.41 on 1 and 73 DF, p-value: 6.598e-11



# Modele z większą liczbą ością zmiennych

proste zależności są rzadko spotykane w przyrodzie

często chcemy zbadać wpływ kilku cech

po udanym wypadzie w teren mamy aż nadto danych

czy możemy wrzucić je wszystkie na raz i zobaczyć co wyjdzie?

- Która zmienna jest kluczowa?
- Ile i które zmienne mają wpływ na to co modelujemy?
- Jak uzasadnić ich wybór i dopasowanie?

# Po pierwsze – dlaczego?

- Hipotezy badawcze
  - weryfikowalne w sposób statystyczny
  - uzasadnione merytorycznie (w intro lub w MM)
- Czy mam prawo zakładać że coś ma wpływ?
  - Zależność bogactwa gatunkowego roślin od numeru powierzchni badawczej
  - Zależność liczby roślin od fazy księżyca
  - Zależność różnorodności gatunkowej grzybów od masy nasion
- Dobrze postawienie hipotezy i uzasadnienie zmiennych – połowa sukcesu

summary(sosny)

	Plot.ID	N	E
Forest - 102c	: 1	Min. :51.21	Min. :14.34
Forest - 106c - 1	: 1	1st Qu.:51.23	1st Qu.:17.59
Forest - 106c - 2	: 1	Median :51.65	Median :18.25
Forest - 107k	: 1	Mean :52.01	Mean :18.25
Forest - 111a-1	: 1	3rd Qu.:52.01	3rd Qu.:19.41
Forest - 111a-2	: 1	Max. :53.94	Max. :19.92
(Other)	:71		

	Soil.type	Age	dens	G
Haplic Podzol	:23	Min. : 3.00	Min. : 2717	Min. : 0.000
Spolic Technosols (Protic)	:20	1st Qu.: 6.00	1st Qu.: 4960	1st Qu.: 1.350
Albic Brunic Arenosol	:12	Median :10.00	Median : 6055	Median : 4.310
Albic Podzols	:12	Mean :10.26	Mean : 6924	Mean : 8.461
Haplic Brunic Arenosol	: 4	3rd Qu.:15.00	3rd Qu.: 8707	3rd Qu.:14.750
Gleyic Podzol	: 2	Max. :20.00	Max. :14000	Max. :29.790
(Other)	: 4			

	Hg	V	AB	BR
Min.	:0.334	Min. : 0.057	Min. : 0.1738	Min. : 0.038
1st Qu.:	1.824	1st Qu.: 6.178	1st Qu.: 6.9883	1st Qu.: 2.008
Median :	3.010	Median : 12.365	Median :11.7229	Median : 3.044
Mean :	3.844	Mean : 31.243	Mean :21.5752	Mean : 5.490
3rd Qu.:	5.598	3rd Qu.: 47.313	3rd Qu.:37.2202	3rd Qu.: 7.959
Max. :	9.096	Max. :118.178	Max. :57.5928	Max. :23.857

	FL	ST
Min.	: 0.1284	Min. : 0.057
1st Qu.:	2.5774	1st Qu.: 2.360
Median :	3.8398	Median : 4.404
Mean :	4.3740	Mean :11.785
3rd Qu.:	5.3993	3rd Qu.:18.756
Max. :	11.8670	Max. :41.173

sosny<-read.csv('sosny.csv', sep=,;')

Zbiór danych: sosny

dane o biomasie drzewostanów sosnowych w wieku 3-20 lat

N, E – koordynaty

Age – wiek [lata]

dens – zagęszczenie [szt. ha<sup>-1</sup>]

G – pole powierzchni drzew [m<sup>2</sup> ha<sup>-1</sup>]

Hg – wysokość górna d-stanu [m]

V – miąższość [m<sup>3</sup> ha<sup>-1</sup>]

AB – biomasa części nadziemnej [t ha<sup>-1</sup>]

BR – gałęzi

FL – liści

ST -pni

Co wpływa na AB?

koordynaty – nie bardzo

Soil.type – może, ale założmy że nie

BR, FL i ST są składowymi AB

## Po drugie – jak wybrać?

- Im więcej zmiennych tym lepiej?
- Czy zmienne są zależne od siebie czy nie?
- Czy dodanie kolejnych zmiennych jest uzasadnione?



# AIC - kryterium informacyjne Akaikego

miara jakości dopasowania modelu

wartości można porównywać w ramach tej samej zmiennej objaśnianej

do czego może to służyć?

porównanie jakości modeli z modelem zerowym

model zerowy - intercept-only,  $Y \sim 1$ , czyli wstawiamy wszędzie średnią

porównanie jakości modeli pomiędzy sobą

Czy dodanie kolejnej zmiennej jest usprawiedliwione? Zasada parsymonii

# Dla modeli z jedną zmienną – porównanie jakości

```
> AIC(lm(AB~1,data=sosny)) #model zerowy
```

```
[1] 666.9126
```

```
> AIC(lm(AB~V,data=sosny)) #model liniowy
```

```
[1] 467.6902
```

```
> AIC(lm(AB~poly(V,2),data=sosny)) #model kwadratowy
```

```
[1] 422.4431
```

# Dodanie zmiennej poprawia dopasowanie modelu

```
Console R Markdown
E:/Nauka/stat_narz/R/BSS/bssRV

> model0<-lm(AB~1,data=sosny)
> model1<-lm(AB~V,data=sosny)
> model2<-lm(AB~V+Hg,data=sosny)
> summary(model2)

call:
lm(formula = AB ~ V + Hg, data = sosny)

Residuals:
    Min       1Q   Median       3Q      Max
-14.5941  -2.0876  -0.5072   1.8221  16.3335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4508     1.6538   0.877   0.3832
V              0.4224     0.0655   6.449 1.03e-08 ***
Hg            1.8020     0.8932   2.017  0.0473 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.817 on 74 degrees of freedom
Multiple R-squared:  0.9305,    Adjusted R-squared:  0.9286
F-statistic: 495.6 on 2 and 74 DF, p-value: < 2.2e-16
```

# urościło nieznacznie, co z AIC?

```
> AIC(model1,model2,model0)
```

	df	AIC
model1	3	467.6902
<u>model2</u>	<u>4</u>	<u>465.5678</u>
model0	2	666.9126

```
model3<-lm(AB~V+Hg+dens,data=sosny)
```

```
AIC(model3)
```

```
[1] 464.6941
```

```
model4<-lm(AB~V+Hg+E,data=sosny)
```

```
AIC(model4)
```

```
[1] 466.0577
```

# Współliniowość zmiennych i VIF (variance inflation factor)

Nie mówmy dwa razy o tym samym

Rule of Thumb  $VIF > 5 \Rightarrow$  problem, ale...

<https://statisticalhorizons.com/multicollinearity>

<https://pdfs.semanticscholar.org/ed1f/4466a0982f3e8de202de01ecceb473d11893.pdf>

z czego wynika VIF? czy ma to biologiczne znaczenia dla badanej cechy?

# library(car)

```
> vif(model1)
```

```
Error in vif.default(model1) : model contains fewer than 2 terms
```

```
> vif(model2)
```

V	Hg
---	----

14.01216	14.01216
----------	----------

```
> vif(model3)
```

V	Hg	E
---	----	---

15.573167	16.907578	1.359008
-----------	-----------	----------

```
> vif(model4)
```

V	Hg	E
---	----	---

15.573167	16.907578	1.359008
-----------	-----------	----------

# Dlaczego tak?

$$V=G*H*f$$

f - wskaźnik kształtu którego nie znamy

V - miąższość

G - pole powierzchni przekroju pierśnicowego

H - wysokość

cechy są ze sobą silnie związane



```
model5<-lm(AB~V+dens,data=sosny)
```

```
Call:
lm(formula = AB ~ V + dens, data = sosny)

Residuals:
    Min       1Q   Median       3Q      Max
-14.537  -2.552  -0.056   1.745  15.193

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.8163784   2.2893626   4.288 5.37e-05 ***
V             0.5190664   0.0211514  24.540 < 2e-16 ***
dens        -0.0006439   0.0002565  -2.510  0.0142 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.75 on 74 degrees of freedom
Multiple R-squared:  0.9325,    Adjusted R-squared:  0.9306
F-statistic: 510.8 on 2 and 74 DF,  p-value: < 2.2e-16

> AIC(model5)
[1] 463.3966
> vif(model5)
      V      dens
1.502807 1.502807
> |
```

# Jak dobierać zmienne?

- Wg hipotez
- Nie współliniowe (VIF)
- Minimalizując AIC

## W praktyce:

- Budujemy model globalny – wszystkie zmienne które mogą mieć wpływ
- Usuwamy zmienne współliniowe (w oparciu o VIF)
- Redukujemy model aby zminimalizować AIC

# Selekcja modeli – funkcja step()

```
> step(lm(AB~V+Hg+G+dens+Age, sosny))
```

```
Start:  AIC=232.44
```

```
AB ~ V + Hg + G + dens + Age
```

	Df	Sum of Sq	RSS	AIC
- dens	1	21.067	1369.4	231.63
<none>			1348.3	232.44
- V	1	57.497	1405.8	233.65
- G	1	109.161	1457.5	236.43
- Hg	1	123.181	1471.5	237.17
- Age	1	272.744	1621.1	244.62

```
Step:  AIC=231.63
```

```
AB ~ V + Hg + G + Age
```

	Df	Sum of Sq	RSS	AIC
<none>			1369.4	231.63
- V	1	41.015	1410.4	231.90
- G	1	118.331	1487.7	236.01
- Hg	1	216.220	1585.6	240.92
- Age	1	311.616	1681.0	245.42

```
call:
```

```
lm(formula = AB ~ V + Hg + G + Age, data = sosny)
```

```
Coefficients:
```

```
(Intercept)
```

```
V
```

```
Hg
```

```
G
```

```
Age
```

dlaczego niebezpieczne?

pozwała szybko i bez refleksy  
wybrać model o najmniejszym  
AIC, bez sprawdzania  
pozostałych parametrów.

Pomocne, ale zawsze należy  
obejrzeć model

Nie chroni przed  
interkorelacją, outlierami,  
brakiem logiki

# Lepsza opcja: library(MuMIn)

wybór modelu w oparciu o AIC – wszystkie kombinacje

tworzymy model globalny - ze wszystkimi zmiennymi które mogą wejść do modelu

funkcja dredge() buduje podzbiór wszystkich możliwych kombinacji zmiennych

i oblicza AICc (domyślnie, AICc = AIC z korektą dla małych prób)

dłużej się wykonuje ale warto

```
library(MuMIn)
global.model<-lm(AB~V+G+Hg+dens+Age,
data=sosny,na.action = na.fail)
```

#na.action=na.fail jest kluczowe!

```
dred<-dredge(global.model)
dred
```

mamy selekcje modeli i ich współczynniki,  
AICc - AIC dla małych prób  
delta - różnica w AICc w stosunku do najlepszego modelu

weights - wagi Akaiego  
można je interpretować luźno jako  
“prawdopodobieństwo że dany model jest najlepszym modelem pod względem AIC w zbiorze modeli-kandydatów”

źródło interpretacji:  
<https://link.springer.com/article/10.3758/BF03206482>

Global model call: lm(formula = AB ~ V + G + Hg + dens + Age, data = sosny, na.action = na.fail)

---

Model selection table

	(Intrc)	Age	dens	G	Hg	V	df	logLik	AICc	delta	weight
14	4.644	-1.57100		1.4680	5.3670			5 -221.209	453.3	0.00	0.319
30	4.582	-1.32400		1.0300	4.5360	0.1418	6	-220.073	453.3	0.08	0.306
32	7.889	-1.26000	-0.0002998	0.9929	3.8280	0.1785	7	-219.476	454.6	1.31	0.165
16	6.102	-1.57100	-0.0001315	1.5010	5.1510		6	-221.084	455.4	2.10	0.111
26	2.274	-1.08700			5.3560	0.3158	5	-223.264	457.4	4.11	0.041
28	6.320	-1.02100	-0.0003578		4.4750	0.3522	6	-222.474	458.1	4.88	0.028
24	13.380	-0.70170	-0.0006240	1.1820		0.3108	6	-222.842	458.9	5.62	0.019
22	7.419	-0.59970		1.3700		0.2789	5	-225.718	462.3	9.02	0.004
19	9.816	-0.0006439				0.5191	4	-227.698	464.0	10.69	0.002
23	8.397	-0.0005050	0.5347			0.3889	5	-226.571	464.0	10.72	0.001
20	12.590	-0.29060	-0.0007630			0.5520	5	-226.816	464.5	11.21	0.001
21	4.039		0.7695		0.3529	4	-228.354	465.3	12.00	0.001	
27	7.277	-0.0005090			0.8580	0.4649	5	-227.347	465.5	12.28	0.001
25	1.451			1.8020	0.4224		4	-228.784	466.1	12.86	0.001
31	8.219	-0.0004971	0.5214	0.0720	0.3876		6	-226.569	466.3	13.07	0.000
29	2.575		0.5440	0.9591	0.3428		5	-227.967	466.8	13.51	0.000
8	13.640	-0.97780	-0.0004926	2.5260			5	-228.529	467.9	14.64	0.000
17	4.398				0.5498	3	-230.845	468.0	14.75	0.000	
6	8.811	-0.87260		2.5670			4	-230.136	468.8	15.56	0.000
18	4.878	-0.07853				0.5602	4	-230.779	470.1	16.85	0.000
10	-2.010	-1.76700			10.8500		4	-232.947	474.4	21.18	0.000
12	-3.863	-1.76100	0.0001866		10.9800		5	-232.750	476.3	23.08	0.000
5	3.828		2.0980			3	-235.591	477.5	24.25	0.000	
13	1.549		1.6860	1.5000			4	-234.788	478.1	24.87	0.000
7	6.040	-0.0002576	2.0470				4	-235.195	478.9	25.68	0.000
15	3.014	-0.0001322	1.7190	1.2830			5	-234.699	480.2	26.98	0.000
9	-6.688		7.3530			3	-245.955	498.2	44.97	0.000	
11	-9.035	0.0002385		7.5340			4	-245.725	500.0	46.74	0.000
4	3.070	2.58900	-0.0011640				4	-287.564	583.7	130.42	0.000
2	-9.012	2.98100				3	-289.534	585.4	132.13	0.000	
3	51.240	-0.0042850				3	-312.879	632.1	178.82	0.000	
1	21.580					2	-331.456	667.1	213.81	0.000	

# Średnie brzegowe/ odpowiedzi brzegowe

- Model z więcej niż jedną zmienną trudniej pokazać
- Jak zmienia się wartość oczekiwana przy założeniu że pozostałe parametry są stałe?
- $\text{Masa} = 2.5 * \text{wysokość} + \text{typ1} * 0.2 + \text{typ2} * 0.5 + \text{typ3} * (-0.3)$
- Wiemy jak typ modyfikuje masę przy danej wysokości, ale jak pokazać średnie wartości z modelu?
- Co jak jest więcej zmiennych?

marginal means, emmeans, marginal response, partial dependence

# Średnie brzegowe – marginal means – średnie z modelu

```
model<-lm(AB~V+type,data=sosny)
library(multcomp);library(emmeans)
```

```
summary(model)
```

```
Call:
lm(formula = AB ~ V + type, data = sosny)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.360	-2.046	-0.242	1.509	16.415

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.94087	0.79884	4.933	4.93e-06 ***
V	0.53795	0.01845	29.155	< 2e-16 ***
typepostagric	5.03635	1.96117	2.568	0.0123 *
typepostind	1.42232	1.33221	1.068	0.2892

---

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

Residual standard error: 4.764 on 73 degrees of freedom  
Multiple R-squared: 0.933, Adjusted R-squared: 0.9302  
F-statistic: 338.7 on 3 and 73 DF, p-value: < 2.2e-16

```
emmeans(model, 'type')
```

type	emmean	SE	df	lower.CL	upper.CL
forest	20.7	0.69	73	19.4	22.1
postagric	25.8	1.82	73	22.2	29.4
postind	22.2	1.10	73	20.0	24.4

Confidence level used: 0.95

```
cld(emmeans(model, 'type'))
```

type	emmean	SE	df	lower.CL	upper.CL	.group
forest	20.7	0.69	73	19.4	22.1	1
postind	22.2	1.10	73	20.0	24.4	12
postagric	25.8	1.82	73	22.2	29.4	2

Confidence level used: 0.95

P value adjustment: tukey method for comparing a family of 3 estimates  
significance level used: alpha = 0.05

NOTE: Compact letter displays can be misleading  
because they show NON-findings rather than findings.  
Consider using 'pairs()', 'pwpp()', or 'pwpm()' instead.

# Odpowiedź brzegowa - ggeffects

```
mod11<-lm(AB~V+N+dens, sosny)
> summary(mod11)
```

Weźmy model z trzema zmiennymi

Call:  
lm(formula = AB ~ V + N + dens, data = sosny)

Residuals:

Min	1Q	Median	3Q	Max
-15.2934	-2.1345	-0.2878	1.3569	15.3180

Coefficients:

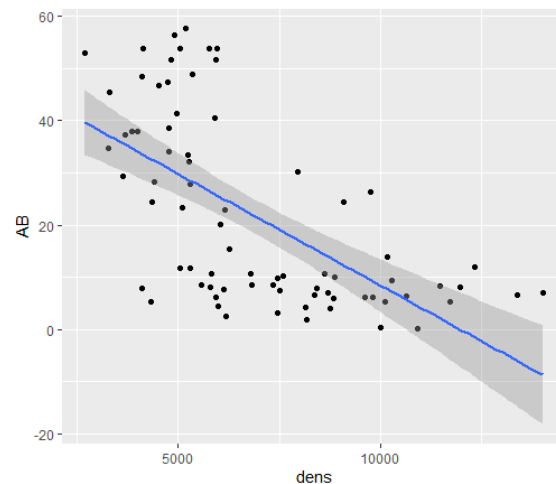
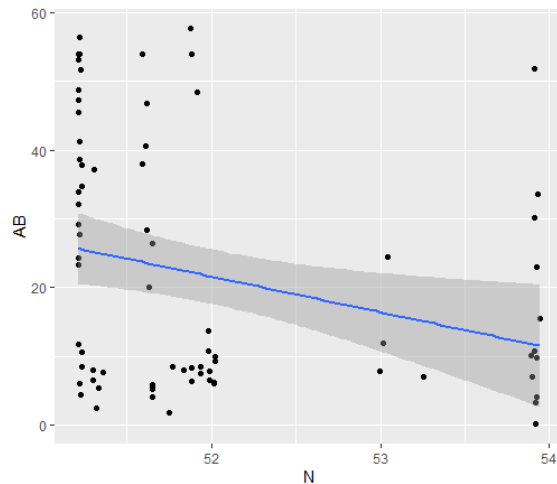
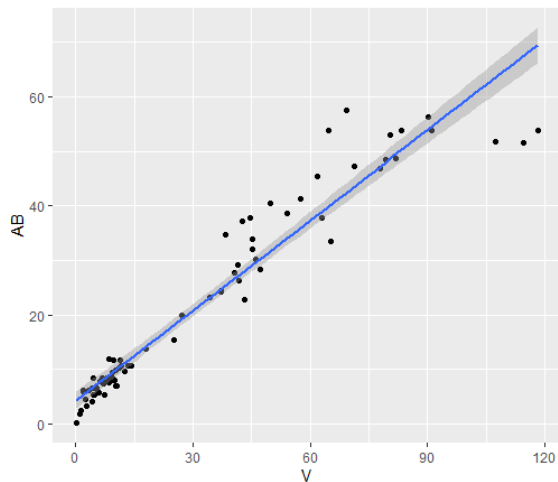
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	83.6055072	29.5867642	2.826	0.00608 **
V	0.5192881	0.0204384	25.407	< 2e-16 ***
N	-1.4437916	0.5772877	-2.501	0.01463 *
dens	-0.0004569	0.0002589	-1.765	0.08175 .

---

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



# Summary jedno, obrazki drugie



Dlaczego?

Każdy obrazek osobno, bez całościowego uwzględnienia modelu

# Marginal response

- Odpowiedź brzegowa, zakładająca średni poziom pozostałych zmiennych
- $y = ax_1 + bx_2 + cx_3 + d$ , podstawiamy pod  $x_1$  poszczególne wartości
- pod  $x_2$  i pod  $x_3$  podstawiamy średnią wartość z zestawu danych
- mamy więc obraz jak wraz ze zmianą  $x_1$  zmieni się  $y$  przy założeniu że pozostałe czynniki się nie zmieniają
- *ceteris paribus* – wszystkie bez jednego się nie zmieniają

```
ggpredict(mod11)
```

```
$V
```

```
# Predicted values of AB
```

```
  V | Predicted |      95% CI
```

```
-----
```

```
  0 |    5.75 | [ 3.90, 7.60]
```

```
 15 |   13.54 | [12.13, 14.95]
```

```
 30 |   21.33 | [20.20, 22.45]
```

```
 45 |   29.12 | [27.98, 30.25]
```

```
 60 |   36.91 | [35.49, 38.32]
```

```
 75 |   44.69 | [42.83, 46.55]
```

```
 90 |   52.48 | [50.11, 54.86]
```

```
120 |   68.06 | [64.58, 71.55]
```

```
Adjusted for:
```

```
*   N =  52.01
```

```
•   dens = 6055.00
```

```
$N
```

```
# Predicted values of AB
```

```
  N | Predicted |      95% CI
```

```
-----
```

```
51.00 |    23.43 | [21.93, 24.93]
```

```
51.50 |    22.71 | [21.51, 23.90]
```

```
52.00 |    21.99 | [20.87, 23.10]
```

```
52.50 |    21.26 | [19.96, 22.57]
```

```
53.00 |    20.54 | [18.87, 22.21]
```

```
53.50 |    19.82 | [17.69, 21.95]
```

```
54.00 |    19.10 | [16.47, 21.72]
```

```
Adjusted for:
```

```
*   V =  31.24
```

```
*   dens = 6055.00
```

```
$dens
```

```
# Predicted values of AB
```

```
 dens | Predicted |      95% CI
```

```
-----
```

```
2700 |    23.51 | [21.13, 25.88]
```

```
4100 |    22.87 | [21.10, 24.63]
```

```
5500 |    22.23 | [20.97, 23.48]
```

```
6900 |    21.59 | [20.56, 22.61]
```

```
8400 |    20.90 | [19.63, 22.17]
```

```
9800 |    20.26 | [18.48, 22.04]
```

```
11200 |    19.62 | [17.22, 22.02]
```

```
14000 |    18.34 | [14.61, 22.08]
```

```
Adjusted for:
```

```
*   V = 31.24
```

```
*   N = 52.01
```

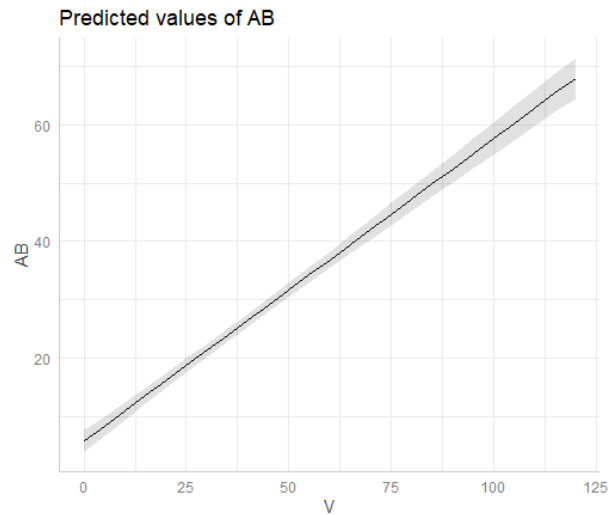
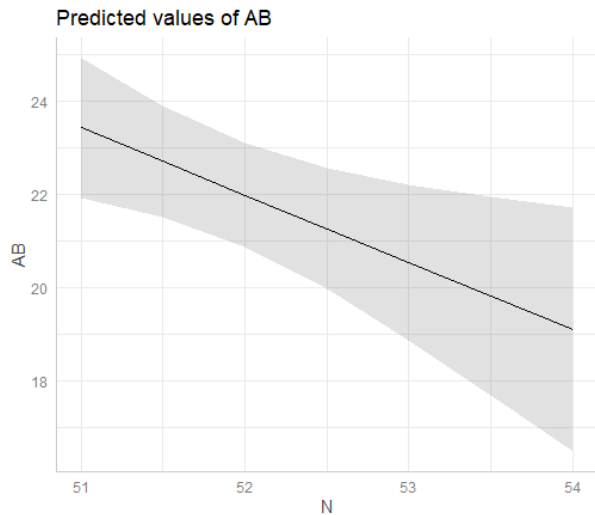
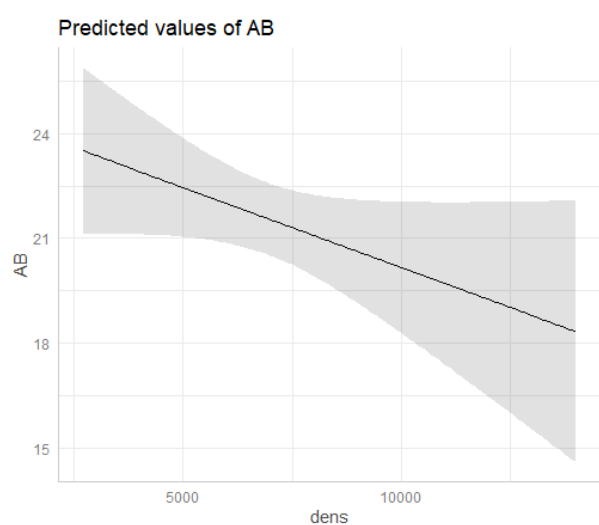
```
attr(,"class")
```

```
[1] "ggalleffects" "list"
```

```
attr(,"model.name")
```

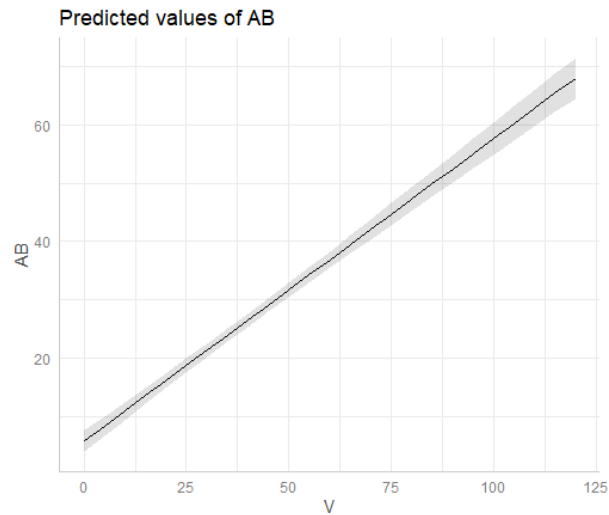
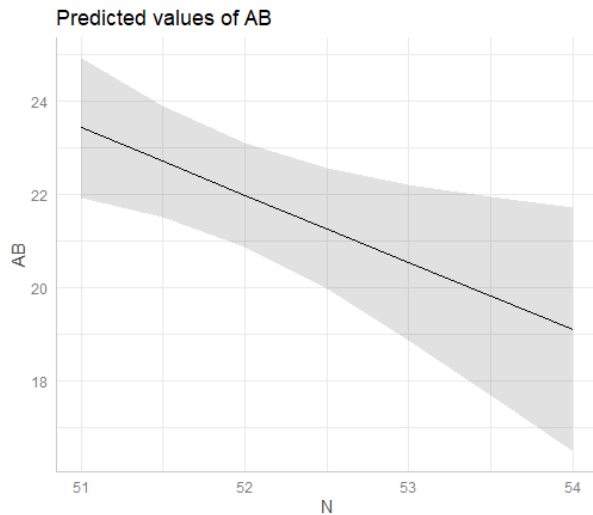
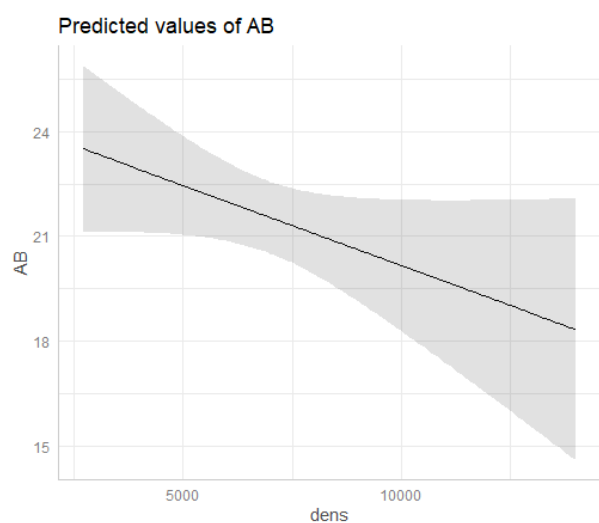
```
[1] "mod11"
```

`plot(ggpredict(mod1 1))`, trzy osobne wykresy



fajnie ale nie widać efektów – różne skale osi

# plot(ggpredict(mod1 1))



fajnie ale nie widać efektów – różne skale osi

# Jak zrobić na jednym ggplotcie ręcznie?

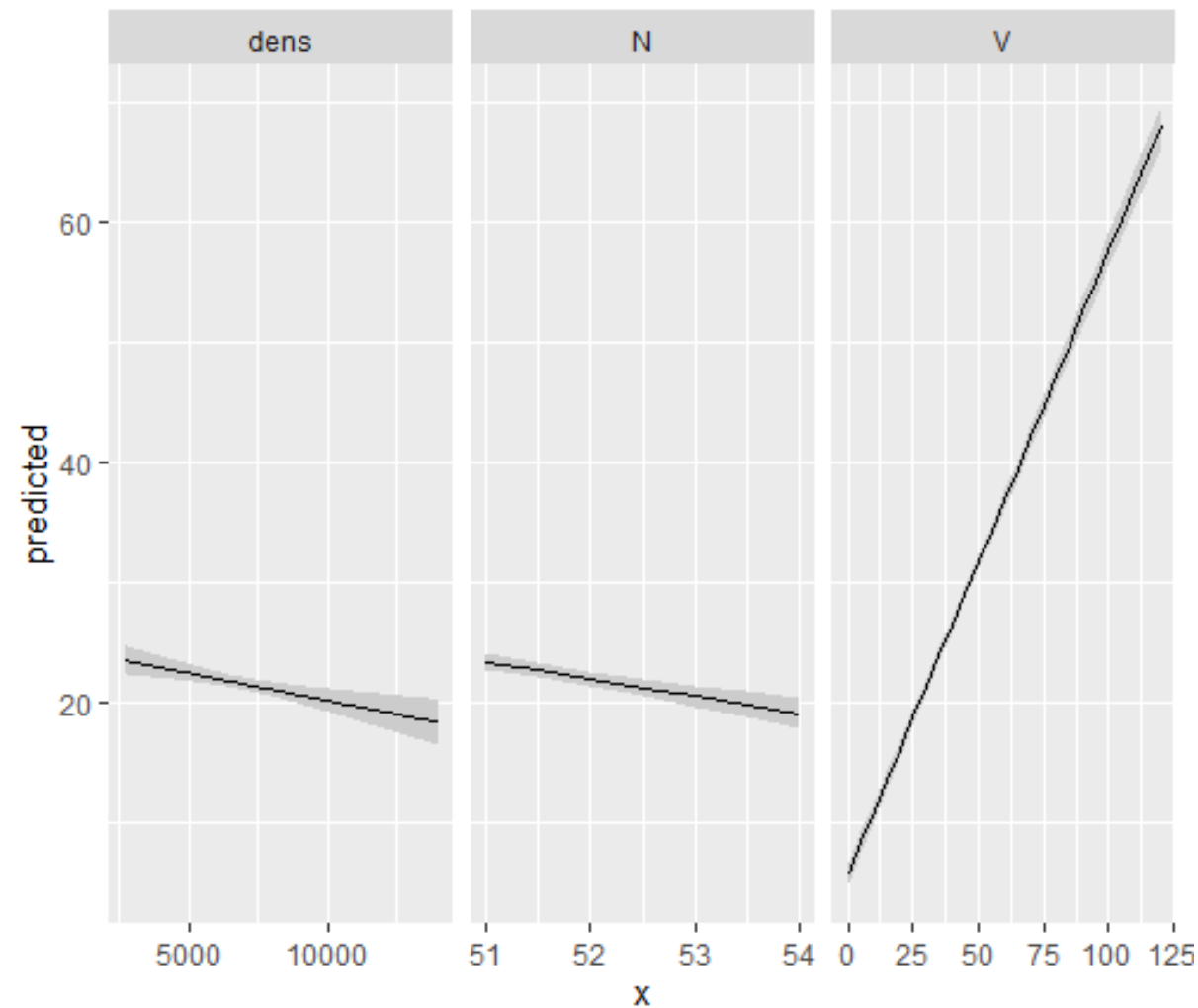
```
ggp1<-as.data.frame(ggpredict(mod11, 'V'))  
#stwórz ggpredict dla jednej zmiennej  
ggp1$group<-'V' #w kolumnie group wpisz nazwę zmiennej  
ggp2<-as.data.frame(ggpredict(mod11, 'N'))  
ggp2$group<-'N'  
ggp3<-as.data.frame(ggpredict(mod11, 'dens'))  
ggp3$group<-'dens'
```

```
ggp<-bind_rows(list(ggp1, ggp2, ggp3))  
#złącz listę tabel w jedno
```

# Jak to wygląda?

```
head(ggp)
```

	x	predicted	std.error	conf.low	conf.high	group
1	0	5.748008	0.9440649	3.897675	7.598342	V
2	5	8.344449	0.8635288	6.651963	10.036934	V
3	10	10.940889	0.7880304	9.396378	12.485400	V
4	15	13.537330	0.7191579	12.127806	14.946853	V
5	20	16.133770	0.6589922	14.842169	17.425371	V
6	25	18.730210	0.6101146	17.534408	19.926013	V



```
ggplot(ggp, aes(x=x, y=predicted))  
+geom_ribbon(aes (ymin=predicted-  
std.error, ymax=predicted+std.error ),  
fill='gray80') +geom_line()  
+facet_wrap(~group, scales='free_x')
```

- widać effect size
- widać co wychodzi z modelu



# Dodać rzeczywiste wartości ręcznie (czasem trzeba)

- trzeba zmeltować tabelkę z danymi

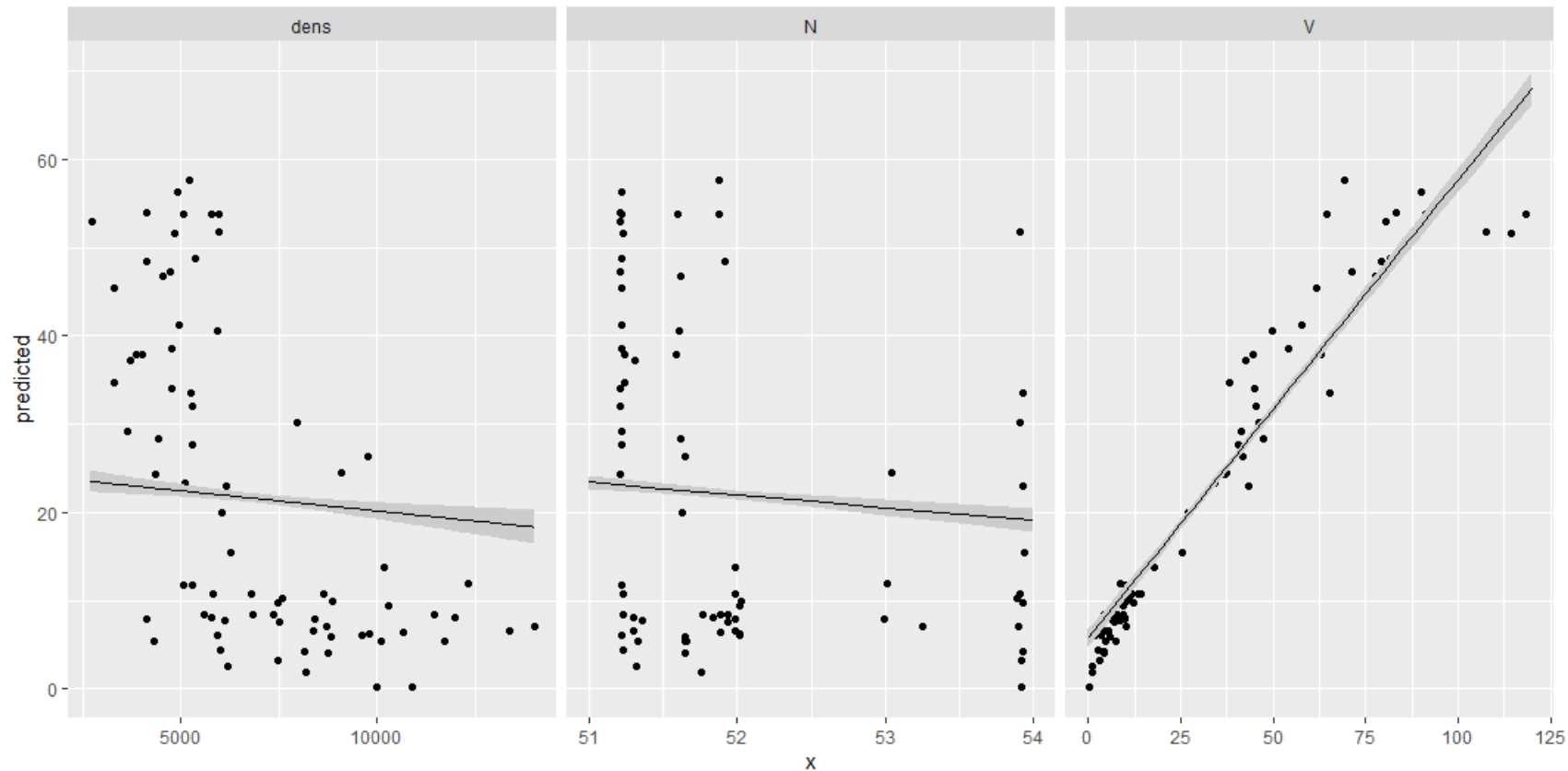
```
gg.points<-reshape2::melt(sosny[,c(2,6,9,10)], id.vars='AB')
```

```
> head(gg.points)
```

	AB	variable	value
1	2.4730	N	51.3192
2	0.2778	N	53.9168
3	0.1738	N	53.9161
4	1.8986	N	51.7522
5	1.8986	N	51.7523
6	6.2242	N	52.0138

```
colnames(gg.points)<-c('predicted','group','x')
```

```
#zmiana nazw kolumn na odpowiadające w tabeli ggp
```



```
ggplot(ggp, aes(x=x, y=predicted)) +geom_point(data=gg.points)
+geom_ribbon(aes(ymin=predicted-std.error ,ymax=predicted+std.error ),fill='gray80')
+geom_line()+facet_wrap(~group, scales='free_x')
```

#każdy geom\_ może mieć swoją własną tabelę wejściową, definiowaną argumentem data=nazwa.tabeli

# Przykład: wpływ inwazyjnych drzew na mszaki

- dataset – pokrycie mszaków epigeicznych w Wielkopolskim PN
- budujemy model globalny

```
bssmodel<-lm(Cover.moss.layer....~soilpH+soilCN+DIFN+type,  
data=plots,na.action=na.fail)
```

```
car::vif(bssmodel)
```

GVIF Df GVIF<sup>1/(2\*Df)</sup>

soilpH 2.069323 1 1.438514

soilCN 2.015716 1 1.419759

DIFN 1.527215 1 1.235805

type 3.500151 8 1.081447

Folia Geobot (2020) 55:351–363  
<https://doi.org/10.1007/s12224-020-09384-2>

**Impacts of alien tree species on the abundance and diversity  
of terricolous bryophytes**

Marcin K. Dyderski  • Andrzej M. Jagodziński 



```
> dredge(bssmodel)
```

Fixed term is "(Intercept)"

Global model call: `lm(formula = Cover.moss.layer.... ~ soilpH + soilCN + DIFN + type, data = plots, na.action = na.fail)`

```
Model selection table
  (Intrc)  DIFN      soilCN      soilpH type df  logLik   AICc  delta weight
12  0.016380 3.503 -0.0039580      + 12 127.928 -230.1   0.00  0.399
10 -0.078880 3.435      + 11 126.618 -229.7   0.33  0.338
16  0.061090 3.515 -0.0042590 -9.279e-03 + 13 128.114 -228.1   1.94  0.151
14 -0.062430 3.438      -4.076e-03 + 12 126.655 -227.5   2.55  0.112
 2 -0.075550 4.121      3 96.136 -186.1  43.91  0.000
 4 -0.055590 4.170 -0.0009355      4 96.222 -184.2  45.83  0.000
 6 -0.075120 4.121      -9.312e-05 4 96.137 -184.1  46.00  0.000
 8 -0.029210 4.173 -0.0013070 -4.000e-03 5 96.257 -182.2  47.87  0.000
 9  0.004750      + 10 73.461 -125.7 104.39  0.000
13 -0.006928      2.880e-03 + 11 73.472 -123.4 106.63  0.000
11 -0.003788      0.0003487 + 11 73.467 -123.4 106.64  0.000
15 -0.020230      0.0004542 3.418e-03 + 12 73.481 -121.2 108.89  0.000
 3 -0.166600      0.0114000      3 30.367 -54.6 175.45  0.000
 7 -0.184100      0.0116400 2.651e-03 4 30.374 -52.5 177.53  0.000
 5  0.263000      -3.622e-02 3 25.133 -44.1 185.92  0.000
 1  0.102600      2 23.221 -42.4 187.68  0.000
Models ranked by AICc(x)
> |
```

Model finalny: DIFN, soilC:N oraz type

```
lm(formula = Cover.moss.layer.... ~ soilCN + DIFN + type, data = plots,  
na.action = na.fail)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.38648	-0.04396	0.00196	0.03929	0.52344

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.016382	0.072649	0.225	0.82186	
soilCN	-0.003958	0.002512	-1.576	0.11685	
DIFN	3.502543	0.296747	11.803	< 2e-16	***
typePinus plantation	0.007056	0.043492	0.162	0.87131	
typePinus plantation - Ps	0.091420	0.048747	1.875	0.06240	.
typePinus poor	0.243618	0.061564	3.957	0.00011	***
typePinus poor - Ps	0.152591	0.049856	3.061	0.00256	**
typeQ. petraea	-0.081435	0.047951	-1.698	0.09123	.
typeQ. rubra	-0.041634	0.053871	-0.773	0.44065	
typeQuercus-Acer-Tilia	-0.012353	0.050637	-0.244	0.80755	
typeRobinia	-0.005330	0.059131	-0.090	0.92829	

---

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

Residual standard error: 0.1254 on 175 degrees of freedom

Multiple R-squared: 0.6756, Adjusted R-squared: 0.6571

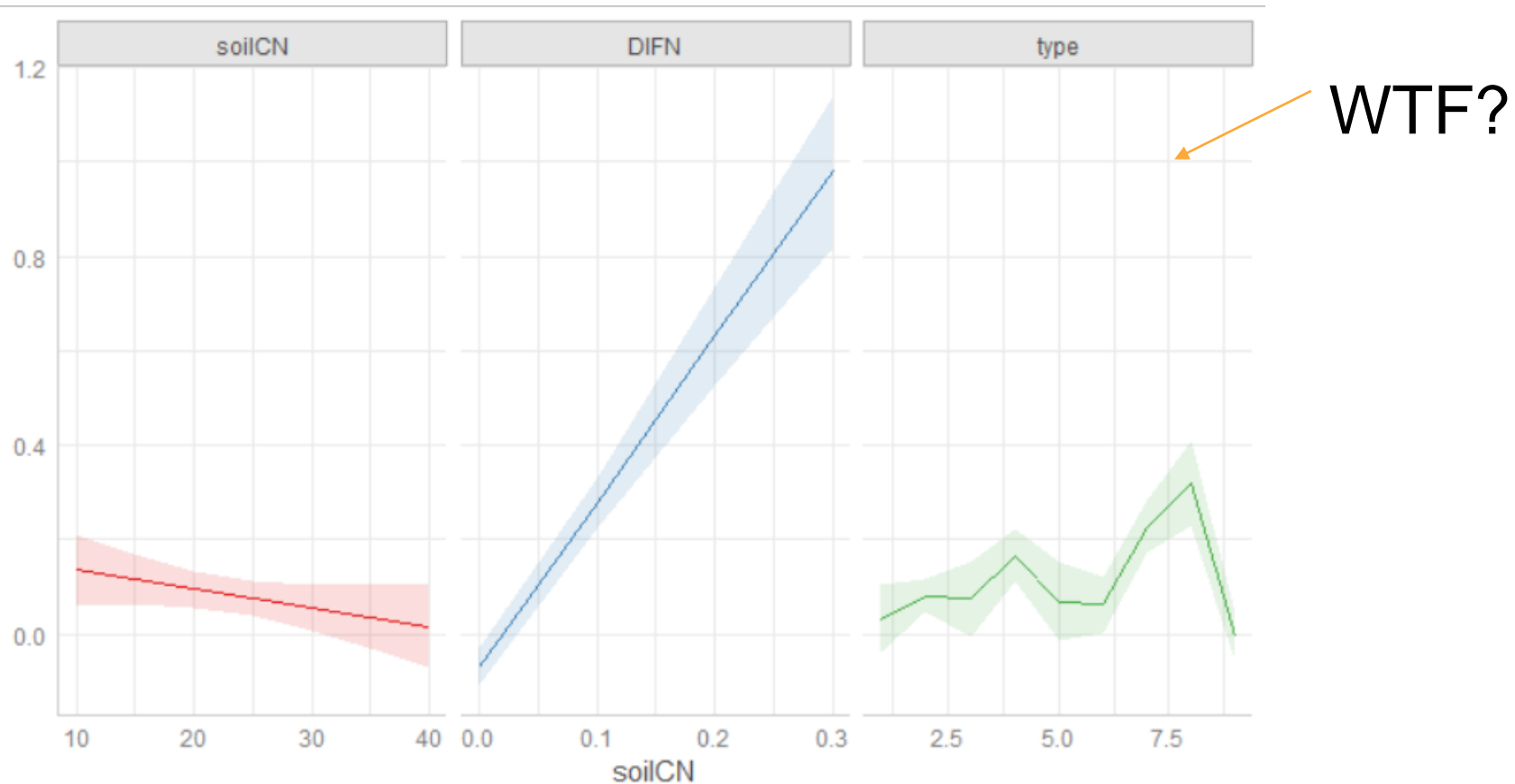
F-statistic: 36.45 on 10 and 175 DF, p-value: < 2.2e-16

No fajnie, ale co z tego?



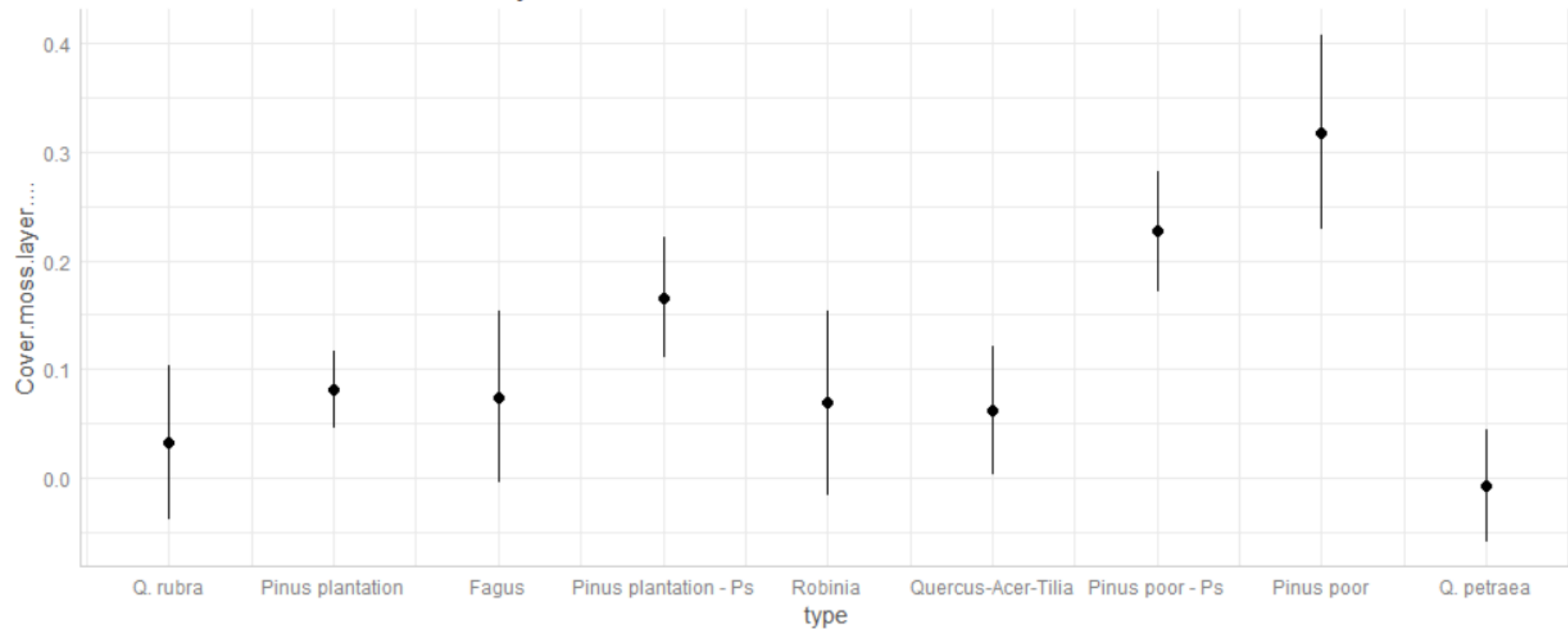
Wyszukaj





`plot(ggpredict(bssmodel2),facets=T)` #tak wyglądało  
kiedyś, teraz zmienili i nie działa, trzeba ręcznie

Predicted values of Cover.moss.layer....



```
plot(ggpredict(bssmodel2,'type'))
```





# Model z interakcją

znaki w formule: + addytywność (wspólne oddziaływanie)

: interakcja \*addytywność i interakcja

```
> an2<-aov(lm(AB~type:Soil.type,data=sosny))
> summary(an2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type:Soil.type	10	7125	712.5	2.673	0.0083 **
Residuals	66	17590	266.5		

---

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

```
> an2<-aov(lm(AB~type*Soil.type,data=sosny))
> summary(an2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	2	3767	1883.4	7.067	0.00166 **
Soil.type	7	3326	475.1	1.783	0.10554
type:Soil.type	1	32	32.2	0.121	0.72917
Residuals	66	17590	266.5		

---

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

# Co znaczy interakcja?

Różne różnice pomiędzy grupami jednego czynnika w grupach drugiego czynnika

np. gatunki iglaste w ramach jednej grupy będą miały większy przyrost biomasy niż w ramach drugiej grupy

Czy sprawdzać wszystkie możliwe interakcje (full-factorial)?

Nie, bo testujemy coś dla czego nie ma uzasadnienia?

Tak, bo może za kilkanaście lat ktoś będzie miał uzasadnienie?

Tylko te które możemy uzasadnić biologicznie (hipotezy)?

# Przykład – zagęszczenie czarernchy amerykańskiej

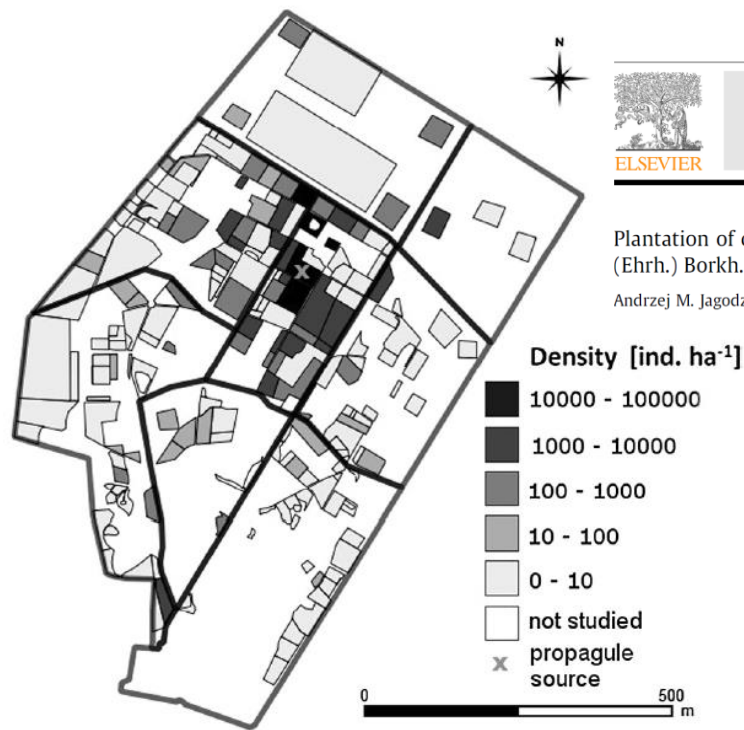


Fig. 1. Spatial distribution of natural black cherry regeneration density (all height classes) in Rogów Arboretum.



Plantation of coniferous trees modifies risk and size of *Padus serotina* (Ehrh.) Borkh. invasion – Evidence from a Rogów Arboretum case study

Andrzej M. Jagodziński<sup>a,b,\*</sup>, Marcin K. Dyderski<sup>c</sup>, Mateusz Rawlik<sup>d</sup>, Piotr Banaszcak<sup>e</sup>

	df	AIC
mconi	7	934.0510
mcon1	5	953.7521
mcon2	5	960.6979

$\log_1 p(x)$  to  $\log(x+1)$

pase – zagęszczenie czarernchy

odl – odległość od źródła nasion

igl – typ d-stanu (3 poziomy)

```
mconi<-lm(log1p(pase)~log(odl)*igl,data=dane)
mcon1<-lm(log1p(pase)~log(odl)+igl,data=dane)
mcon2<-lm(log1p(pase)~log(odl):igl,data=dane)
AIC(mconi,mcon1,mcon2)
```

Call:

lm(formula = log1p(pase) ~ log(odl) \* igl, data = dane)

Residuals:

Min	1Q	Median	3Q	Max
-7.9034	-1.3385	-0.8346	1.4789	6.6448

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	19.0511	1.8077	10.539	< 2e-16 ***
log(odl)	-2.9116	0.3306	-8.806	7.03e-16 ***
iglLIS	-15.5681	3.5483	-4.387	1.88e-05 ***
iglMIX	-15.9376	3.4419	-4.630	6.65e-06 ***
log(odl):iglLIS	2.5064	0.6543	3.831	0.000172 ***
log(odl):iglMIX	2.5157	0.6241	4.031	7.96e-05 ***

---

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

Residual standard error: 2.423 on 195 degrees of freedom

Multiple R-squared: 0.3771, Adjusted R-squared: 0.3611

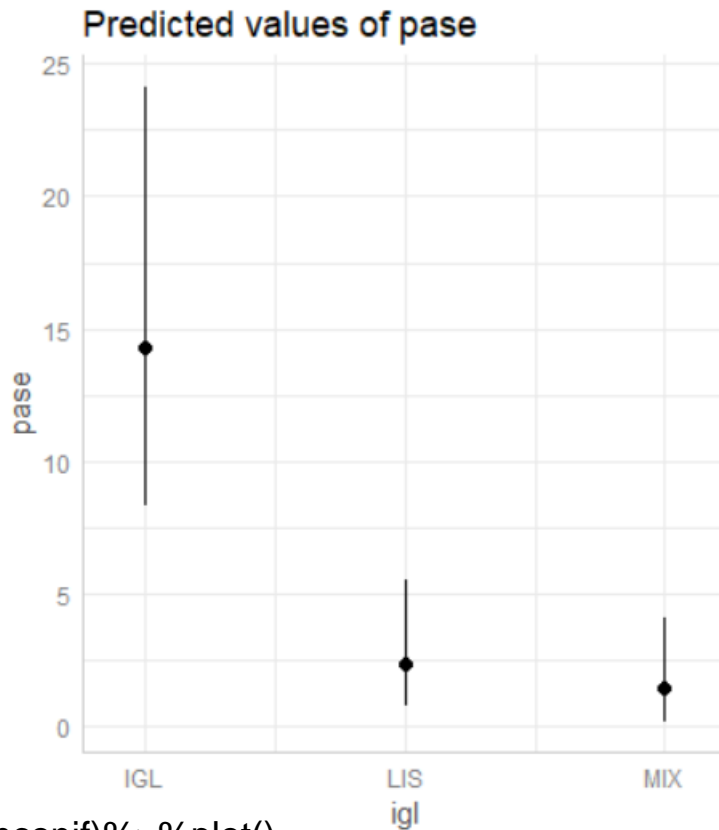
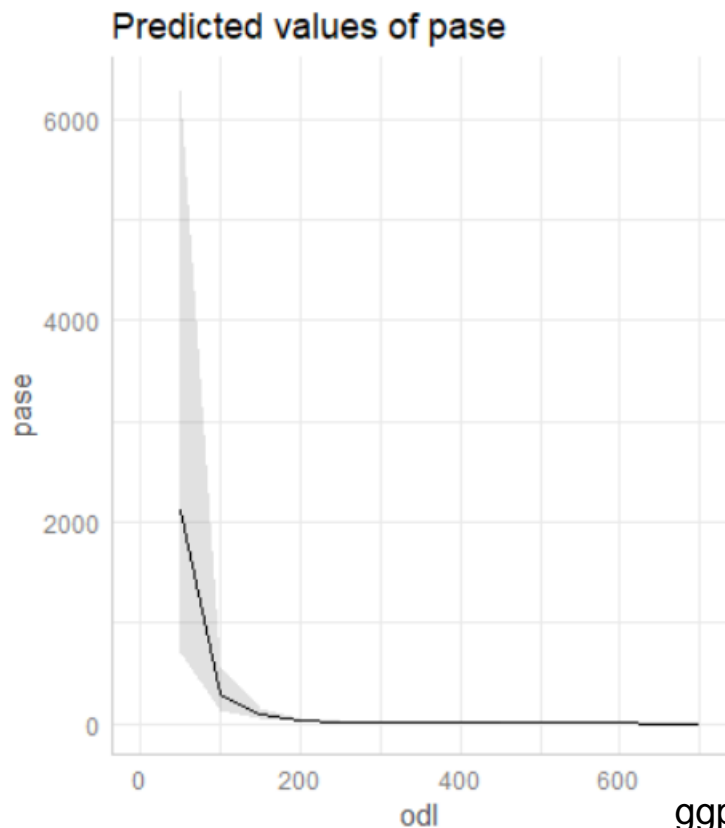
F-statistic: 23.61 on 5 and 195 DF, p-value: < 2.2e-16

Co wynika z tabelki?

Na zagęszczenie czeremchy wpływa:  
typ drzewostanu (igl) – w LIS i MIX mniej niż w iglastych (referencja)  
odległość od źródła nasion – negatywnie

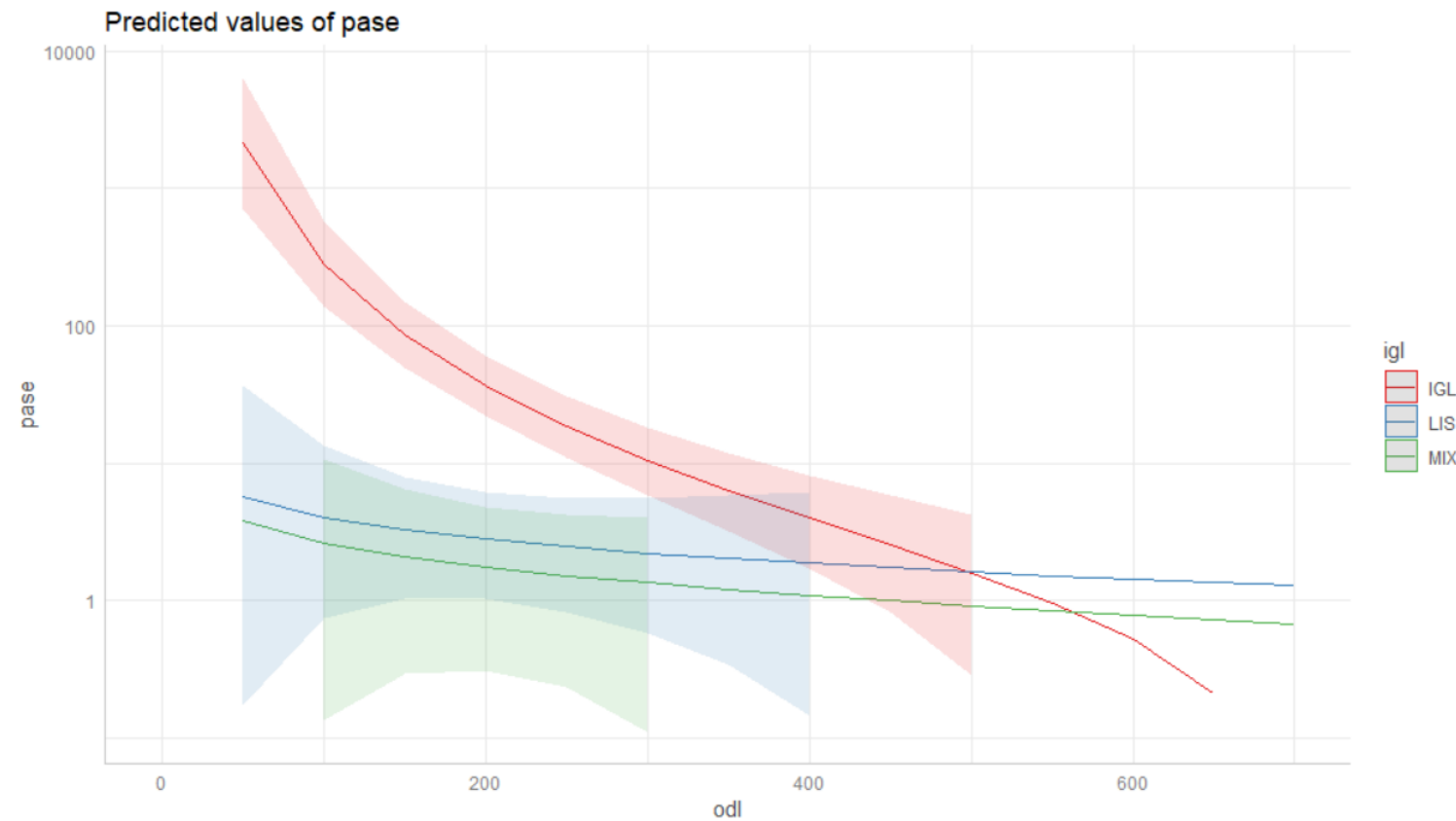
i interakcja – za coraz większą odległość na LIS i MIX są punkty dodatnie – łagodzące negatywny efekt odległości i d-stanu

# Tylko zależność addytywna – bez interakcji



```
ggpredict(mconif, c('odl','igl'))%>%plot()+scale_y_log10()
```

# Z interakcją



# Modele liniowe z efektami losowymi i stałymi (LMM)

a.k.a. modele mieszane, modele z efektem losowym, powtarzalne pomiary,

linear mixed-effect models/ generalized linear mixed-effects model,

# Po co nam modele?

- przybliżenie procesu
- estymacja – szacowanie wartości oczekiwanej
- uwzględnianie dodatkowych czynników – co by było gdyby?
- zgodność z założeniami – niezależność obserwacji



# efekt stały a efekt losowy

stały - związany z działaniem czynnika

losowy - związany z elementami które powinny być niezależne, a mogą mieć wpływ, np. wariant doświadczenia, powtórzenie, termin, itp.

Przykład

**3** terminy badań na **6** blokach po **10** poletek **4** gatunków drzew - światło i odczyn

światło, odczyn, gatunek drzewa - efekty stałe

termin badań, poletko, blok - efekty losowe

# Po co efekty losowe

niezależność obserwacji - możliwość wnioskowania i ekstrapolacji

zależność czasowa, przestrzenna, filogenetyczna, osobnicza...

uwypuklenie trendu z uwzględnieniem kontekstu

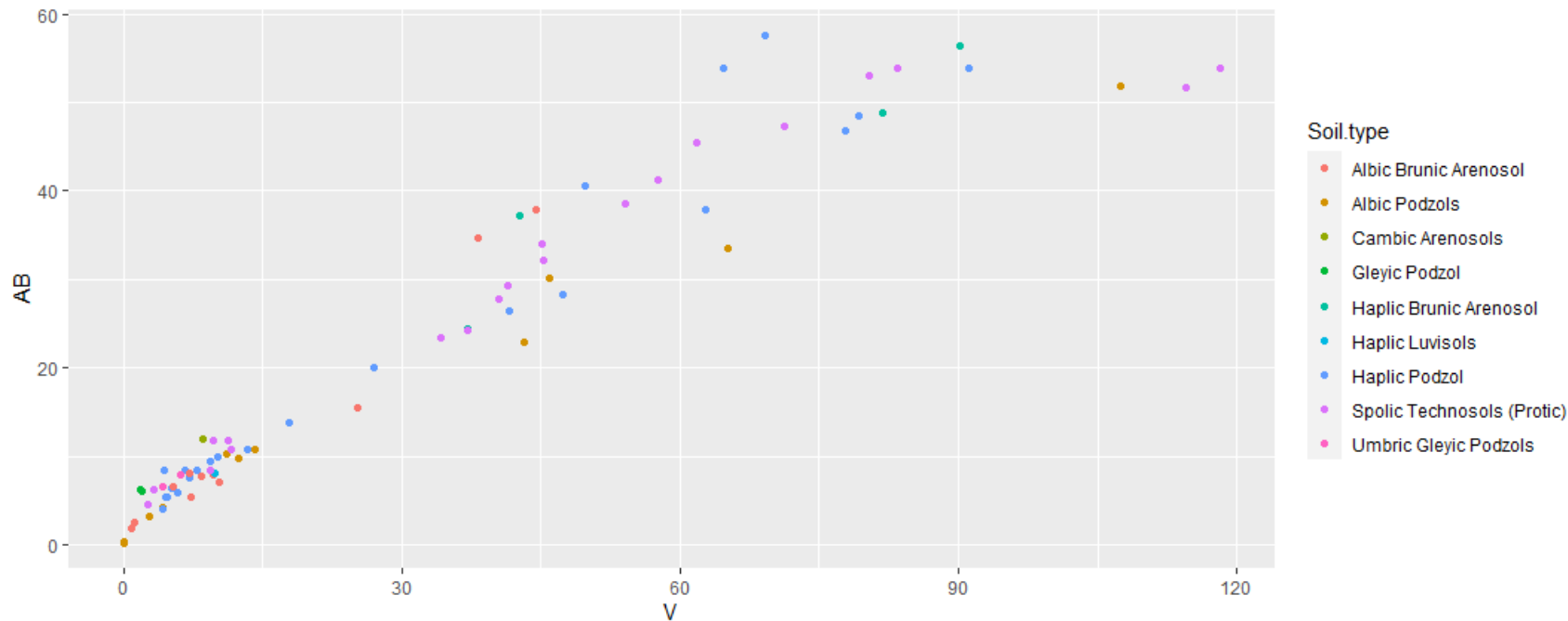
ocena zmienności na różnych poziomach hierarchii

“a co jeśli jakieś poletko jest odwiedzane przez psy i koty - to też może mieć wpływ na zbiorowiska mszaków epifitycznych przez dopływ biogenów i wydrapywanie?” - plot-specific effects

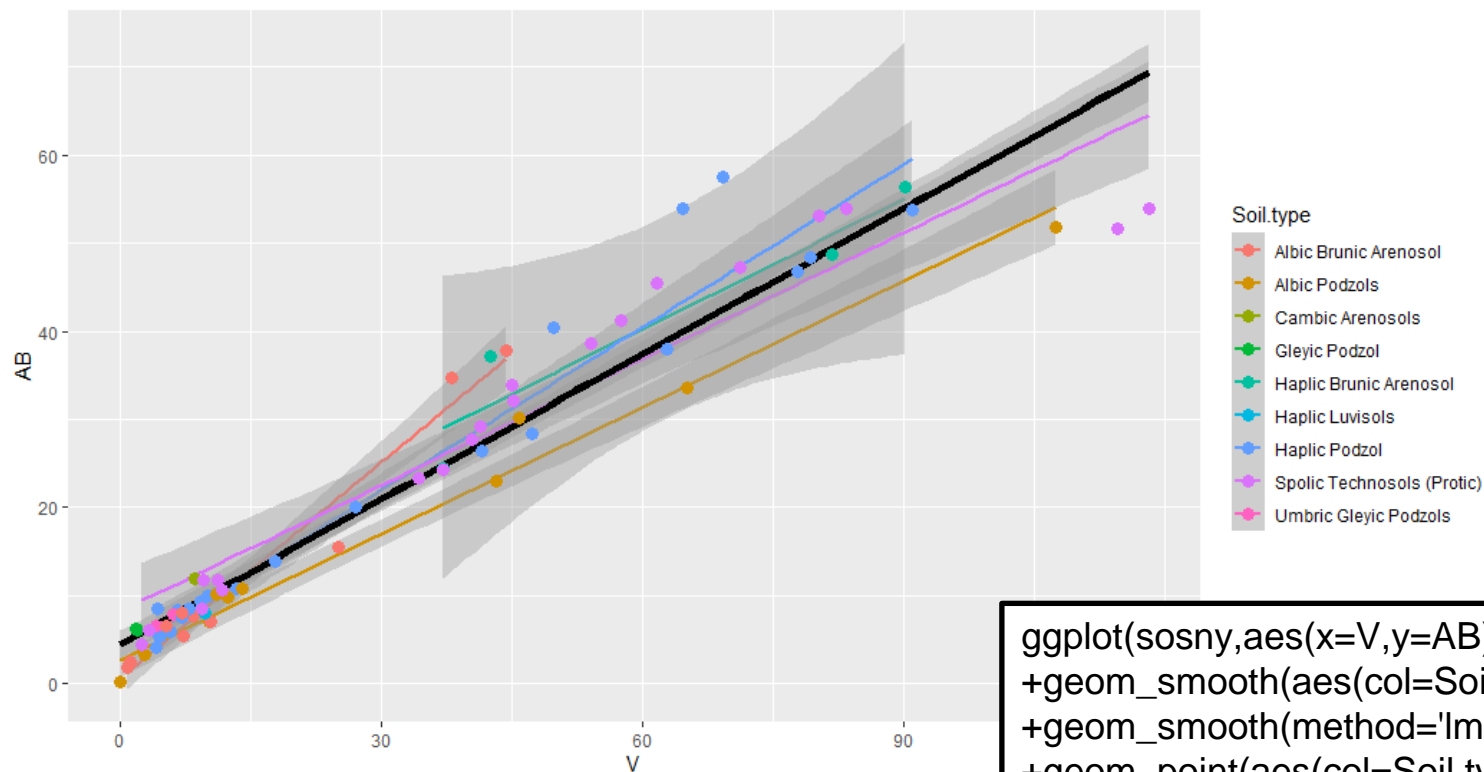
# Przykłady

- przestrzenna: po 4 poletka na 10 transektach – nr transektu jako RE
- czasowa: 20 niezależnych poletek w 3 terminach – termin jako RE
- filogenetyczna: 100 gatunków z 40 rodzin – gatunek:rodzina jako RE
- osobnicza: kleszcze z 20 saren, pięć części ciała na sarnę – ID sarny jako RE

# Sosny – różne wzorce dla różnych typów gleb



# Jak wygląda linia trendu dla całości i dla typów?



```
ggplot(sosny,aes(x=V,y=AB))  
+geom_smooth(aes(col=Soil.type),method='lm')  
+geom_smooth(method='lm',lwd=2,col='black')  
+geom_point(aes(col=Soil.type),size=3)
```

**model<-lm(AB~V, sosny)**

- Model liniowy
- $AB=a*AB+b$

summary(model)

Call:

lm(formula = AB ~ V, data = sosny)

Residuals:

Min	1Q	Median	3Q	Max
-15.7506	-1.7161	-0.3876	1.1526	15.1380

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.39843	0.79045	5.564	3.88e-07 ***
V	0.54978	0.01785	30.794	< 2e-16 ***

---

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

Residual standard error: 4.915 on 75 degrees of freedom

Multiple R-squared: 0.9267, Adjusted R-squared: 0.9257

F-statistic: 948.3 on 1 and 75 DF, p-value: < 2.2e-16

**mod.mix<-lmer(AB~V+(1|Soil.type), sosny)**

- Model liniowy z efektem losowym i stałym
- $AB=a*AB+(b+u_1)$ , gdzie  $u_1$  dla każdego poziomu Soil.type

summary(mod.mix)

Linear mixed model fit by REML. t-tests use Satterthwaite's method

[lmerModLmerTest]

Formula: AB ~ V + (1 | Soil.type)

Data: sosny

REML criterion at convergence: 466

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.2871	-0.3629	-0.1629	0.2790	3.0615

Random effects:

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

Soil.type	(Intercept)	1.755	1.325
-----------	-------------	-------	-------

Residual		22.777	4.772
----------	--	--------	-------

Number of obs: 77, groups: Soil.type, 9

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t )
(Intercept)	4.47608	0.94963	11.51329	4.713	0.000562 ***
V	0.54579	0.01812	73.36032	30.120	< 2e-16 ***

# Prosty przykład na początek

```
hotspots<-read.csv('datasety/hotspots.csv',sep=';')  
mod<-lm(plants~mammals,data=hotspots)  
summary(mod)
```

Call:

```
lm(formula = plants ~ mammals, data = hotspots)
```

Residuals:

Min	1Q	Median	3Q	Max
-631.69	-192.69	-82.69	133.72	1294.67

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	48.746	28.716	1.698	0.0903.
mammals	31.824	3.114	10.219	<2e-16 ***

---

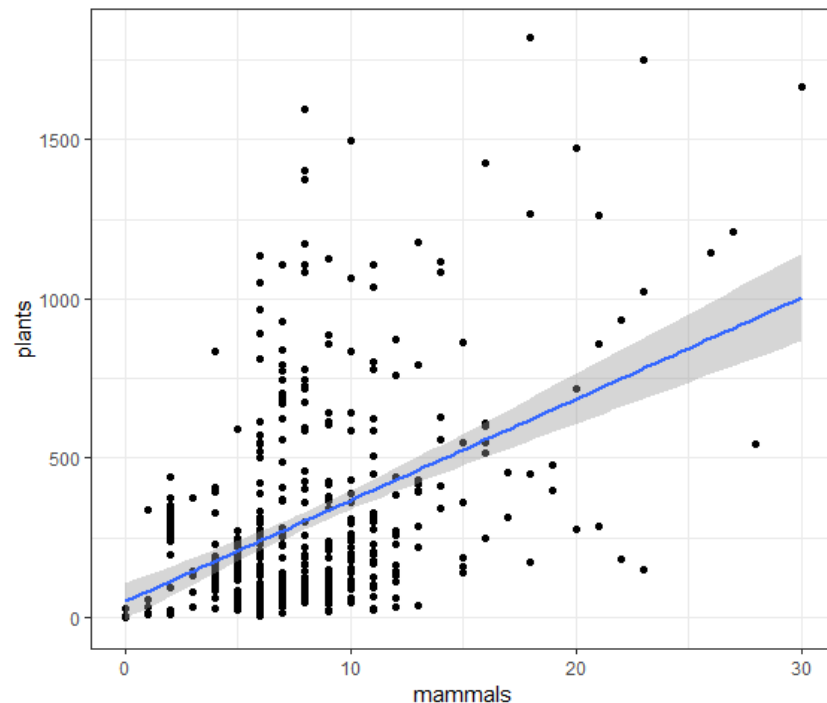
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 294.8 on 447 degrees of freedom

(160 observations deleted due to missingness)

Multiple R-squared: 0.1894, Adjusted R-squared: 0.1876

F-statistic: 104.4 on 1 and 447 DF, p-value: < 2.2e-16



```
ggplot(hotspots, aes(x=mammals,y=plants))+geom_point()  
+theme_bw()+geom_smooth(method='lm')
```

```
mod<-lm(plants~mammals+mainl,data=hotspots)
```

Call:  
lm(formula = plants ~ mammals + mainl, data = hotspots)

Residuals:

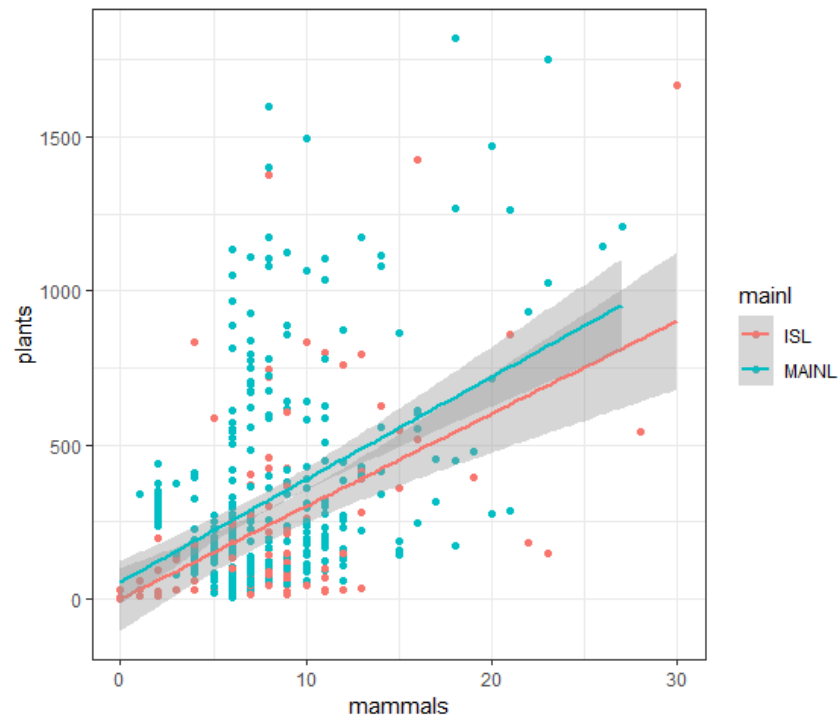
Min	1Q	Median	3Q	Max
-574.32	-203.29	-80.29	130.46	1275.97

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-18.392	39.241	-0.469	0.640
mammals	32.249	3.101	10.400	<2e-16 ***
mainlMAINL	82.437	33.056	2.494	0.013 *

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

Residual standard error: 293 on 446 degrees of freedom  
(160 observations deleted due to missingness)  
Multiple R-squared: 0.2005, Adjusted R-squared: 0.1969  
F-statistic: 55.93 on 2 and 446 DF, p-value: < 2.2e-16



```
ggplot(hotspots, aes(x=mammals,y=plants,col=mainl))  
+geom_point()+theme_bw()+geom_smooth(method='lm')
```



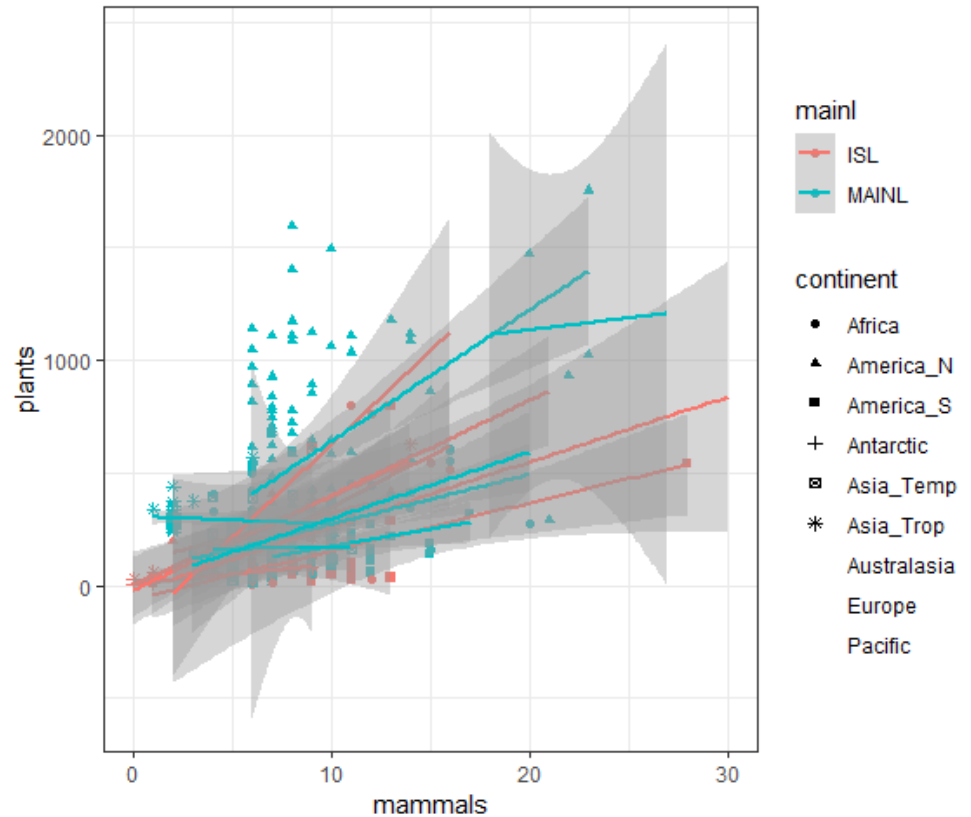
czy dołożyć kontynent?

Czy robi nam różnice?

O co nam chodzi?

różne nachylenia i różne położenia

slope i intercept zależny od continent



```
ggplot(hotspots, aes(x=mammals,y=plants,col=mainl,shape=continent))  
+geom_point()+theme_bw()+geom_smooth(method='lm')
```

# Dodajemy losowy intercept

normalnie  $\text{plants} = a * \text{mammals} + b$   
teraz  
 $\text{plants} = a * \text{mammals} + b + u_2$   
 $u_2$  - losowy intercept

```
mod2<-lmer(plants~mammals+(1|continent),hotspots)
```

```
summary(mod2)
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']

Formula:  $\text{plants} \sim \text{mammals} + (1 \mid \text{continent})$

Data: hotspots

REML criterion at convergence: 6242.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.8175	-0.5247	-0.0337	0.3143	4.3142

Random effects:

Groups	Name	Variance	Std.Dev.
continent	(Intercept)	21998	148.3
Residual		62794	250.6

Number of obs: 449, groups: continent, 9

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t )
(Intercept)	30.684	57.399	12.012	0.535	0.603
mammals	36.700	3.055	441.698	12.011	<2e-16 ***

---

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

Correlation of Fixed Effects:

(Intr)  
mammals -0.421

O co chodzi z zapisem z kreską?

mammals|continent oznacza że mammals ma losowy slope i intercept dla każdego continent dla mammals

normalnie  $\text{plants} = a * \text{mammals} + b$

teraz

$\text{plants} = (a + u_1) * \text{mammals} + b + u_2$

$u_1$  - losowy slope,  $u_2$  - losowy intercept

```
mod3<-lmer(plants~mammals+(mammals|continent),hotspots)
```

```
summary(mod3)
```

Linear mixed model fit by REML. t-tests use Satterthwaite's method

[lmerModLmerTest]

Formula: plants ~ mammals + (mammals | continent)

Data: hotspots

REML criterion at convergence: 6220.9

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.9728	-0.5086	-0.0979	0.3215	4.4747

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
continent	(Intercept)	7088.0	84.19	
	mammals	357.5	18.91	-0.41
Residual		59102.2	243.11	

Number of obs: 449, groups: continent, 9

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t )
(Intercept)	42.905	40.798	8.532	1.052	0.3218
mammals	33.215	7.429	5.960	4.471	0.0043 **

---

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

Correlation of Fixed Effects:

(Intr)  
mammals -0.549

# Jak wyciągnąć wartości

```
coef(mod2)
```

```
$continent
```

```
(Intercept) mammals
```

```
Africa      -52.99065 36.69969  
America_N    223.33035 36.69969  
America_S   -179.55621 36.69969  
Antarctic   -58.33926 36.69969  
Asia_Temp   -34.16470 36.69969  
Asia_Trop    161.22696 36.69969  
Australasia 159.63777 36.69969  
Europe      -80.81611 36.69969  
Pacific     137.82662 36.69969
```

```
attr("class")
```

```
[1] "coef.mer"
```

```
coef(mod3)
```

```
$continent
```

```
(Intercept) mammals
```

```
Africa      25.878318 26.99799  
America_N    52.495373 57.20669  
America_S     4.901014 16.86633  
Antarctic    19.770299 22.02693  
Asia_Temp    58.166601 19.92525  
Asia_Trop   186.721396 22.31839  
Australasia  17.782051 48.76824  
Europe       24.924300 27.16040  
Pacific     -4.492198 57.66827
```

```
attr("class")
```

```
[1] "coef.mer"
```

# Co otrzymujemy?

Model bez efektów losowych - część uogólnioną

Efekty losowe - czynniki warunkujące przebieg krzywych dla kontynentów

Informacje o zmienności tych efektów w ramach modelu:

SD i variance

Który lepszy?

```
> AIC(mod, mod2, mod3)
  df    AIC
mod  4 6380.127
mod2 4 6250.657
mod3 6 6232.945
```

# Ile procent wyjaśnia?

więcej o metodzie i praca do cytowania r2c/r2m  
<https://doi.org/10.1111/j.2041-210x.2012.00261.x>

```
> library(MuMIn)
> r.squaredGLMM(mod2)
      R2m      R2c
[1,] 0.2410538 0.4379521
> r.squaredGLMM(mod3)
      R2m      R2c
[1,] 0.2039087 0.4537038
```

R2m - marginal R2; R2c - conditional R2

R2m - % zmienności wyjaśnionej przez fixed effects

R2c - % zmienności wyjaśnionej przez fixed + random effects

random effects - R2c-R2m

# Przykład – co decyduje o cechach roślin?

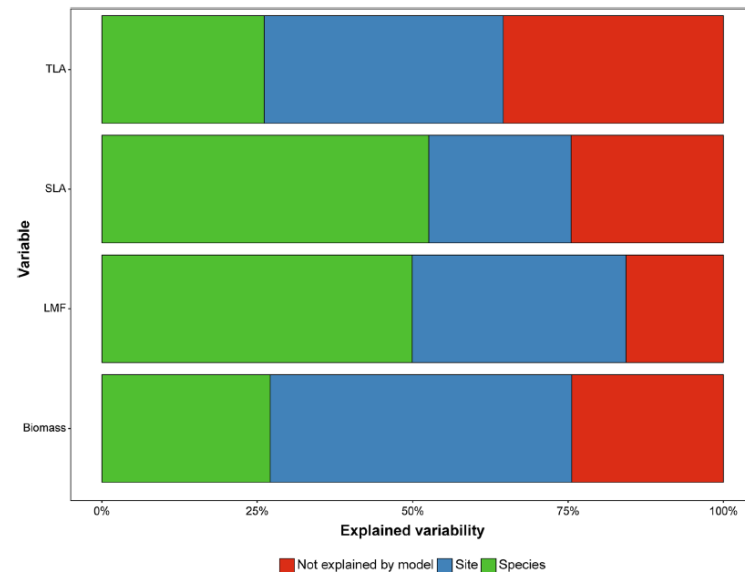
**Table 1.** Comparison of Linear Mixed Models for the Parameters Studied

Response	Fixed effects	AIC	$R_m^2$	$R_c^2$
SLA	Intercept only	61,395.630	0.000	0.53
	Species	57,457.420	0.526	0.75
	Habitat	61,360.600	0.027	0.53
	Species $\times$ habitat	56,054.150	0.577	0.81
LMF	Intercept only	– 1679.004	0.000	0.45
	Species	– 5438.193	0.499	0.84
	Habitat	– 1675.422	0.027	0.45
	Species $\times$ habitat	– 6299.425	0.544	0.85
TLA	Intercept only	52,286.910	0.000	0.52
	Species	50,374.540	0.261	0.64
	Habitat	52,267.520	0.009	0.53
	Species $\times$ habitat	49,219.380	0.359	0.74
Aboveground biomass	Intercept only	5102.774	0.000	0.54
	Species	3600.211	0.271	0.75
	Habitat	5102.840	0.024	0.54
	Species $\times$ habitat	3023.296	0.371	0.75

*Intercept only is a null model; in all models collection site was a random effect.*

Ecossystems (2020) 23: 555–569  
<https://doi.org/10.1007/s10021-019-00421-6>

**ECOSYSTEMS**  
 © 2019 The Author(s)



## Leaf Traits and Aboveground Biomass Variability of Forest Understory Herbaceous Plant Species

Sonia Paż-Dyderska,<sup>1,2</sup> Marcin K. Dyderski,<sup>2</sup> Piotr Szwaczka,<sup>1</sup>  
 Marta Brzezicha,<sup>1</sup> Karolina Bigos,<sup>1</sup> and Andrzej M. Jagodziński<sup>1,2\*</sup>

# LMM w układzie hierarchicznym

4 obiekty, po 50 poletek w każdym

`lmer(biomasa~land.use.history+invasion+(1|variant:plot), dane)`

plot zagnieżdżony w variant - czyli efekty losowe na poziomie plot i variant  
zakładamy że dla każdego mamy random intercept

SD tych modyfikacji interceptów wynosi 0,0837

R2c i R2m - z `r.squaredGLMM()`



Table 2

Differences in dry biomass and biomass allocation of *V. myrtillus* shoots (ramets) between types of pine forest assessed using linear mixed effect models.

Dry shoot mass (n = 400)	Random effects	Variance	SD	Mixed model parameters	–
–	Plots nested in variant	0.0070	0.0837	R <sub>m</sub> <sup>2</sup>	0.1159
–	Residuals	0.1557	0.3946	R <sub>c</sub> <sup>2</sup>	0.1540
–	Fixed effects	Estimate	SE	t	Pr(>  t )
–	(Intercept)	0.745	0.041	18.110	< 0.001
–	Land use history – recent forest	–0.243	0.048	–5.110	< 0.001
–	Invasion – <i>Q. rubra</i>	–0.162	0.048	–3.401	0.002



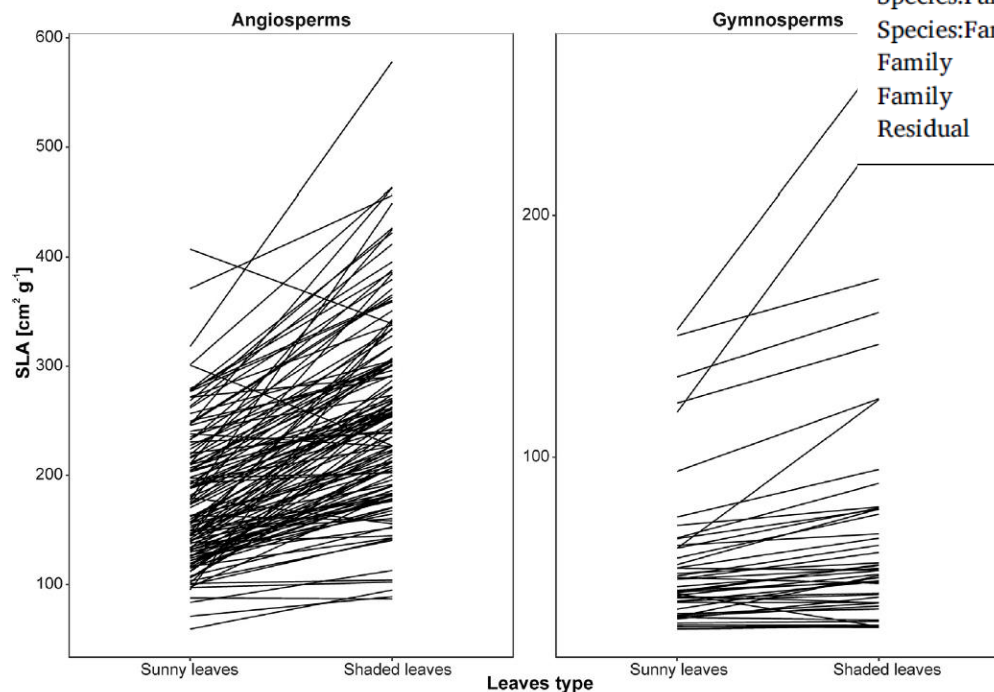


Fig. 2. SLA differences between sunny and shaded parts of the crown.

**Table 1**

Parameters of the model explaining SLA variability.

Fixed effects	Estimate	SE	t	Pr(≥ t)
Intercept	161.27	9.725	16.58	< 0.0001
Light variant (shade)	69.851	6.276	11.13	< 0.0001
Random effect	Variable	SD	Model performance	
Species:Family	Intercept	38.28	R <sub>m</sub> <sup>2</sup>	0.1216
Species:Family	Light variant (shade)	23.55	R <sub>c</sub> <sup>2</sup>	0.8302
Family	Intercept	51.36	AIC	6307.526
Family	Light variant (shade)	28.26	AIC <sub>0</sub>	7041.933
Residual		41.31		

Forest Ecology and Management 472 (2020) 118254



Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: [www.elsevier.com/locate/foreco](http://www.elsevier.com/locate/foreco)

On the sunny side of the crown – quantification of intra-canopy SLA variation among 179 taxa

Sonia Paż-Dyderska<sup>a,\*</sup>, Marcin K. Dyderski<sup>a</sup>, Kinga Nowak<sup>a</sup>, Andrzej M. Jagodziński<sup>a,b</sup>

<sup>a</sup> Institute of Dendrology, Polish Academy of Sciences, Parkowa 5, 62-035 Kórnik, Poland

<sup>b</sup> Poznań University of Life Sciences, Faculty of Forestry, Department of Game Management and Forest Protection, Wojska Polskiego 71c, 60-625 Poznań, Poland

# więcej o modelach mieszanych

<https://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf>

<https://www.r-bloggers.com/linear-mixed-models-in-r/>

<https://www.r-bloggers.com/getting-started-with-mixed-effect-models-in-r/>

<http://www.biecek.pl/WZUR/PrzemekBiecek2009.pdf>

<https://libra.ibuk.pl/book/39524> - podręcznik P. Biecka

# Przykład: wpływ obcych gatunków drzew na biodiv

- 9 typów lasu, schemat blokowy
- spisy florystyczne, miary różnorodności o rozkładzie ~normalnym
- Hipoteza: różnice między typami lasów
- Efekt stały: typ lasu, efekt losowy: blok (random intercept)

- Model globalny:

```
mo.shan<-lmer(shan~type+(1|Blok),data=div,na.action = na.fail)
```

```
dredge(mo.shan)
```

Model selection table

	(Intrc)	type	df	logLik	AICc	delta	weight
2	1.950	+	11	-118.811	261.1	0.00	0.96
1	2.563		3	-130.688	267.5	6.37	0.04

Biol Invasions (2021) 23:235–252  
<https://doi.org/10.1007/s10530-020-02367-6>

ORIGINAL PAPER



**Impacts of invasive trees on alpha and beta diversity of temperate forest understories**

Marcin K. Dyderski · Andrzej M. Jagodziński

summary(mo.shan)

Formula: shan ~ type + (1 | Blok)  
Data: div

REML criterion at convergence: 237.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.14574	-0.59704	0.04685	0.61263	2.30812

Random effects:

Groups	Name	Variance	Std.Dev.
Blok	(Intercept)	0.02135	0.1461
Residual		0.17997	0.4242

Number of obs: 186, groups: Blok, 21

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t )
(Intercept)	1.9502	0.1436	165.5642	13.584	< 2e-16 ***
typePinus plantation	0.6566	0.1533	176.8780	4.283	3.02e-05 ***
typePinus plantation - Ps	0.9551	0.1755	173.7198	5.442	1.77e-07 ***
typePinus poor	0.3610	0.2064	163.7470	1.749	0.082189 .
typePinus poor - Ps	0.7134	0.1741	165.6991	4.097	6.54e-05 ***
typeQ. petraea	0.6022	0.1689	174.3143	3.566	0.000467 ***
typeQ. rubra	0.2504	0.1873	174.1426	1.337	0.183022
typeQuercus-Acer-Tilia	0.6165	0.1714	170.3561	3.598	0.000421 ***
typeRobinia	0.8096	0.1982	176.9397	4.085	6.67e-05 ***

anova(mo.shan)

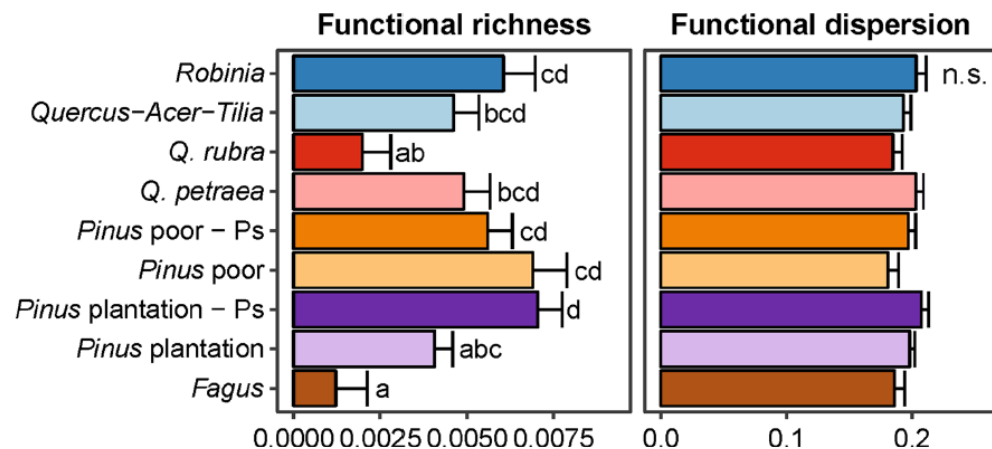
Type III Analysis of Variance Table with  
Satterthwaite's method

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
type	8.2356	1.0294	8	155.77	5.7201	2.202e-06 ***

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

cld(emmeans(mo.shan, ~type,type='resp'))

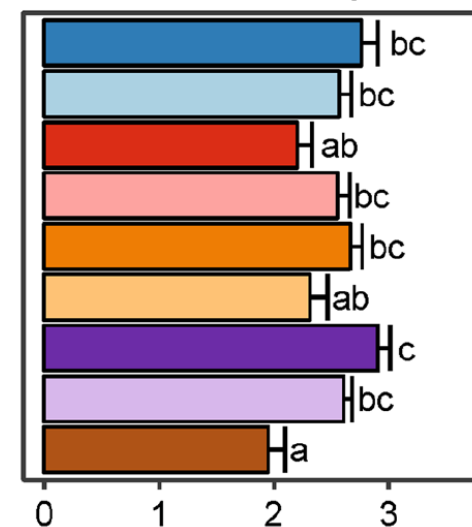
type	emmean	SE	df	lower.CL	upper.CL	group
Fagus	1.95	0.145	165.4	1.66	2.24	1
Q. rubra	2.20	0.130	169.7	1.94	2.46	12
Pinus poor	2.31	0.153	131.1	2.01	2.61	12
Q. petraea	2.55	0.106	61.3	2.34	2.76	23
Quercus-Acer-Tilia	2.57	0.105	74.1	2.36	2.78	23
Pinus plantation	2.61	0.072	57.2	2.46	2.75	23
Pinus poor - Ps	2.66	0.105	90.7	2.46	2.87	23
Robinia	2.76	0.144	163.2	2.47	3.05	23
Pinus plantation - Ps	2.91	0.107	123.5	2.69	3.12	3



**Table 4** Analysis of variance for alpha diversity indices among forest types studied, assessed using linear mixed-effects models (with study plot block as a random intercept)

Response	Variable	df	Sum of Squares	Mean Square	F	Pr(> F)	Block random effects SD
Species richness	Forest type	8	160.3678	20.0460	20.046	< 0.0001	0.265
Shannon diversity index	Forest type	8	8.2356	1.0294	5.720	< 0.0001	0.146
Faith's phylogenetic diversity	Forest type	8	17,473,654.8014	2,184,206.8502	10.821	< 0.0001	286.098
Mean pairwise distance	Forest type	8	184,069.1143	23,008.6393	3.327	0.0014	62.986
Functional dispersion	Forest type	8	0.0086	0.0011	1.978	0.0525	0.008
Functional richness	Forest type	8	0.0003	0.0000	6.703	< 0.0001	0.002

### Shannon's diversity index



# Podsumowanie

obrazki, obrazki, obrazki!

-pomogą dobrać narzędzie i typ rozkładu

biologiczne znaczenie (effect size) > rozkład błędów > AIC >  $R^2$  > p-value

effect size - o ile rośnie nam Y na jednostkę X?

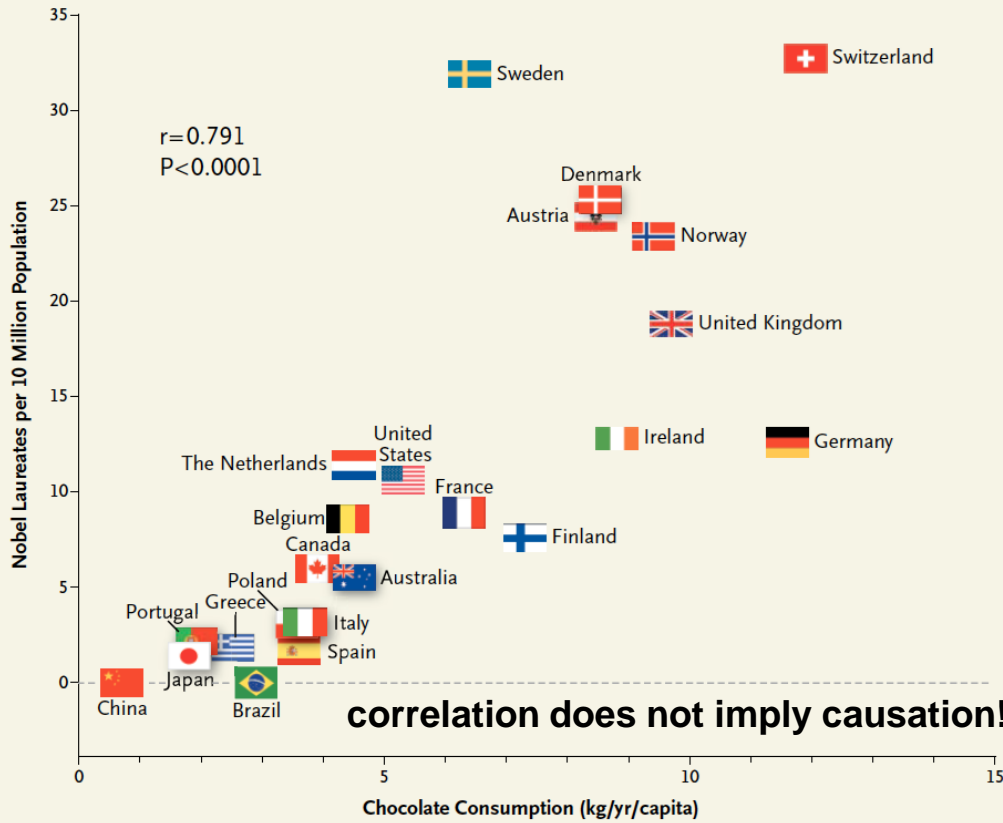
p-value - pomocniczo, **nie podajemy że coś jest istotne statystycznie**

nie  $p < 0.05$ , tylko dokładna wartość – podawać całe summary!

nie można za dużo predyktorów na raz - dwóch strażników - VIF i AIC

# Schemat pracy z modelami

- Czy może być model liniowy? Założenia (jutro)
- Model globalny – wszystkie zmienne które mają **sens biologiczny** (hipotezy)
- Sprawdzamy VIFy – czy nie są współliniowe *car::vif()*
- Selekcja modeli w oparciu o AIC/AICc *stats::step()* lub *MuMIn::dredge()*
- Model finalny:
  - *summary()* lub *anova()* do tabeli (pamiętaj o efektach losowych)
  - przy modelach mieszanych  $R^2_m$  i  $R^2_c$
  - ocena effect size (wykresy) *ggeffects::ggpredict()*
- Czy nas to satysfakcjonuje? Czy ma biologiczny sens?



**Figure 1.** Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

<https://blogs.scientificamerican.com/the-curious-wavefunction/chocolate-consumption-and-nobel-prizes-a-bizarre-juxtaposition-if-there-ever-was-one/>





BIĄŁOWIESKA SZKOŁA STATYSTYKI