
BSS
BIAŁOWIESKA SZKOŁA STATYSTYKI

Regresja

Współczynniki korelacji

r pearsona - najczęściej używany, parametryczny (zakłada rozkład normalny)

rho Spearmana - nieprarametryczny - korelacja rang

tau Kendalla

R2 a r:

współczynnik determinacji - procent wyjaśnionej zmienności

```
cor(sosny$AB,sosny$Age)

[1] 0.8144991
```

co to jest?

?cor

var, cov and cor compute the variance of x and the covariance or correlation of x and y if these are vectors. If x and y are matrices then the covariances (or correlations) between the columns of x and the columns of y are computed.

cov2cor scales a covariance matrix into the corresponding correlation matrix efficiently.

Usage

```
var(x, y = NULL, na.rm = FALSE, use)

cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
```

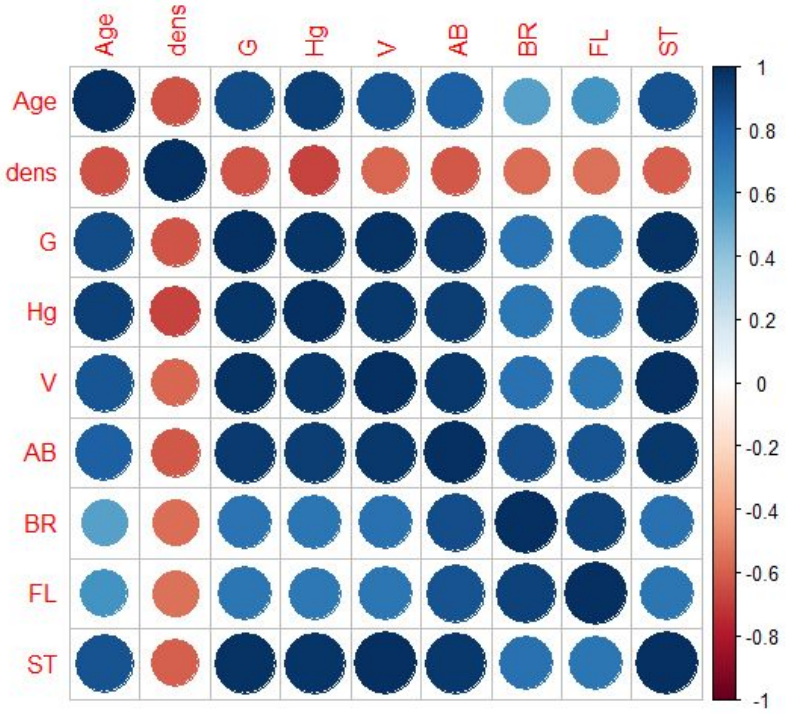
x a numeric vector, matrix or data frame.
y NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).
na.rm logical. Should missing values be removed?
use an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings "everything", "all.obs", **"complete.obs",** "na.or.complete", or "pairwise.complete.obs".
method a character string indicating which correlation coefficient (or covariance) is to be computed. One of "pearson" (default), "kendall", or "spearman": can be abbreviated.

```
> cor(sosny$AB,sosny$Age, method = 'pearson')

[1] 0.8144991

> cor(sosny$AB,sosny$Age, method = 'spearman')

[1] 0.8771537
```

macierz korelacji

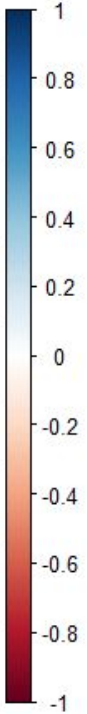
```
> cor(sosny[,c(5:13)])
             Age       dens          G         Hg          V
Age    1.0000000 -0.6369490  0.8984927  0.9308940  0.8521970
dens  -0.6369490  1.0000000 -0.6222580 -0.6751222 -0.5784278
G      0.8984927 -0.6222580  1.0000000  0.9766615  0.9812217
Hg     0.9308940 -0.6751222  0.9766615  1.0000000  0.9636563
V      0.8521970 -0.5784278  0.9812217  0.9636563  1.0000000
AB     0.8144991 -0.6186956  0.9576488  0.9441822  0.9626557
BR     0.5485267 -0.5598806  0.7362057  0.7238736  0.7405488
FL     0.5914534 -0.5479810  0.7225715  0.7191310  0.7230836
ST     0.8621775 -0.5917399  0.9847925  0.9700117  0.9955775
             AB          BR         FL         ST
Age    0.8144991  0.5485267  0.5914534  0.8621775
dens  -0.6186956 -0.5598806 -0.5479810 -0.5917399
G      0.9576488  0.7362057  0.7225715  0.9847925
Hg     0.9441822  0.7238736  0.7191310  0.9700117
V      0.9626557  0.7405488  0.7230836  0.9955775
AB     1.0000000  0.8822559  0.8623371  0.9669117
BR     0.8822559  1.0000000  0.9214561  0.7470985
FL     0.8623371  0.9214561  1.0000000  0.7253943
ST     0.9669117  0.7470985  0.7253943  1.0000000
>
```

```
library(corrplot)
corrplot(cor(sosny[,5:13]))
```



corrplot(cor(sosny[,5:13]),method='num')

	Age	dens	G	Hg	V	AB	BR	FL	ST
Age	1	-0.64	0.9	0.93	0.85	0.81	0.55	0.59	0.86
dens	-0.64	1	-0.62	-0.68	-0.58	-0.62	-0.56	-0.55	-0.59
G	0.9	-0.62	1	0.98	0.98	0.96	0.74	0.72	0.98
Hg	0.93	-0.68	0.98	1	0.96	0.94	0.72	0.72	0.97
V	0.85	-0.58	0.98	0.96	1	0.96	0.74	0.72	1
AB	0.81	-0.62	0.96	0.94	0.96	1	0.88	0.86	0.97
BR	0.55	-0.56	0.74	0.72	0.74	0.88	1	0.92	0.75
FL	0.59	-0.55	0.72	0.72	0.72	0.86	0.92	1	0.73
ST	0.86	-0.59	0.98	0.97	1	0.97	0.75	0.73	1



Regresja

zależność pomiędzy dwoma cechami

-przewidywanie (modelowanie) zmiennej zależnej

-wyjaśnianie procesów

co tak naprawdę chcemy osiągnąć?

z czego możemy zrezygnować?

Korelacja a regresja

korelacja - miara współzależności

regresja - opis zależności

np. masa~średnicy r=0,95 masa=10*średnica+2

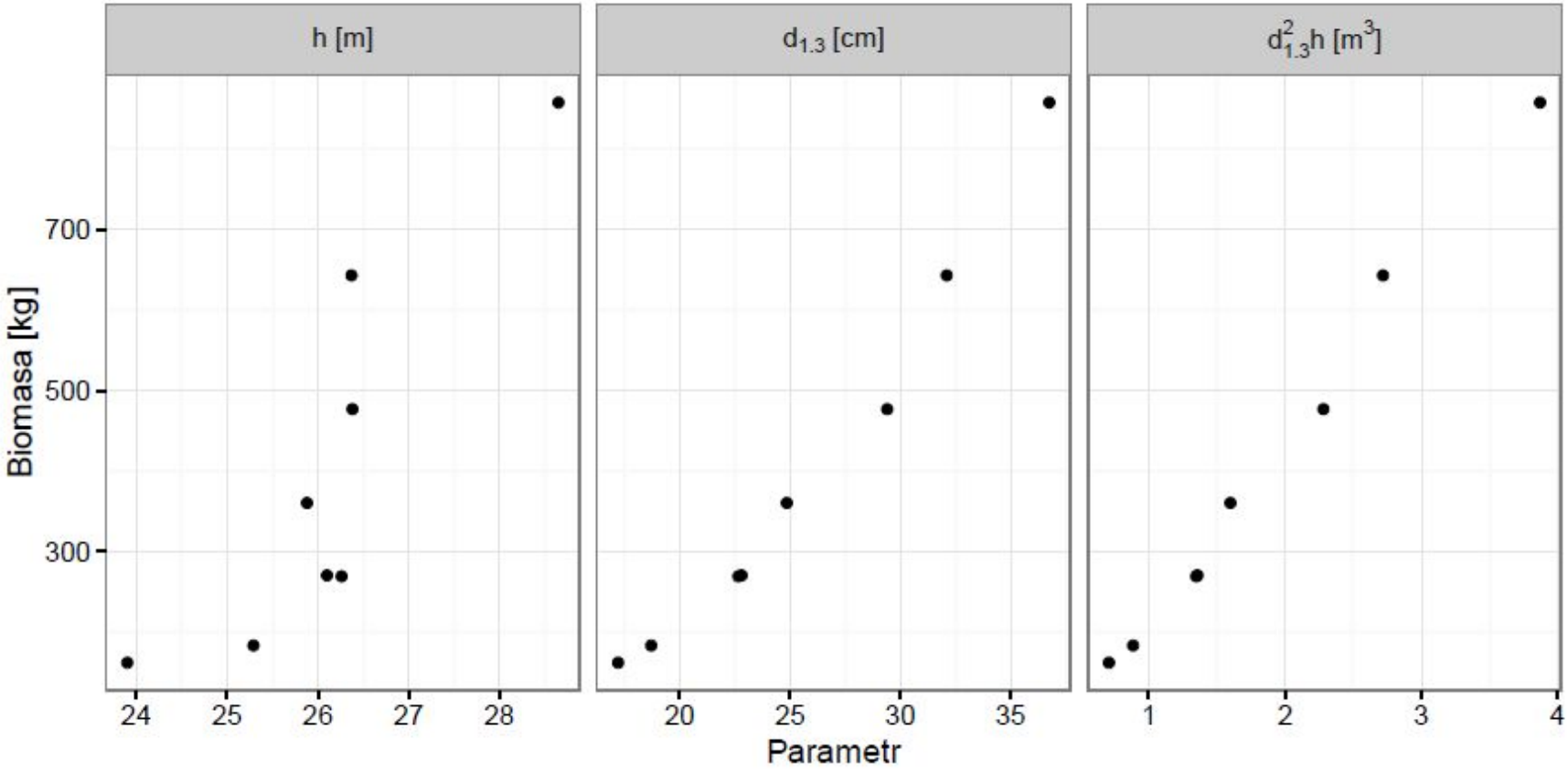
Regresja liniowa

Jak zmienia się masa drzewa wraz z przyrostem na grubość?

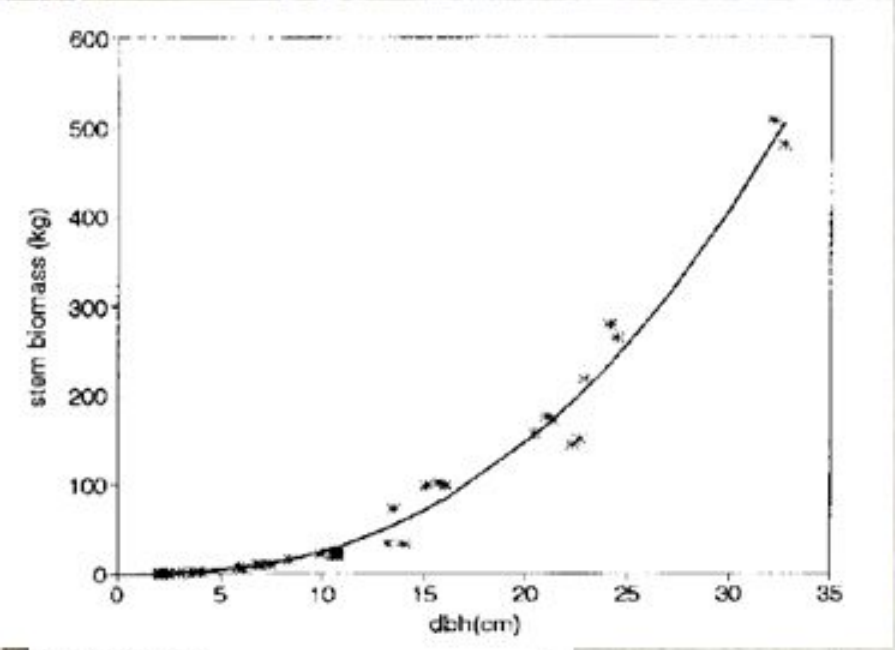
problem badawczy (poznanie tempa wzrostu)

problem aplikacyjny (możliwość estymacji)

?


h [m]
$d_{1.3}$ [cm]
$d^2_{1.3}h$ [m^3]
Biomasa [kg]
700
500
300
24
25
26
27
28
20
25
30
35
1
2
3
4
Parametr

Równania allometryczne



Ann Sci For (1997) 54, 39–50
© Elsevier/INRA

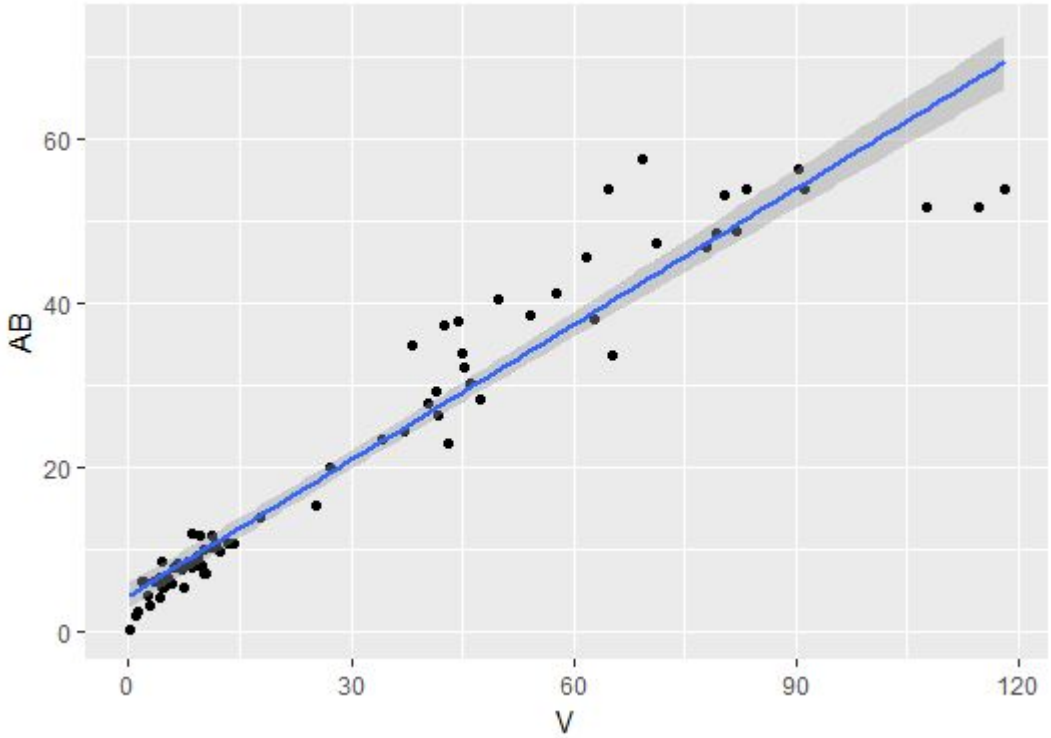
39

Original article

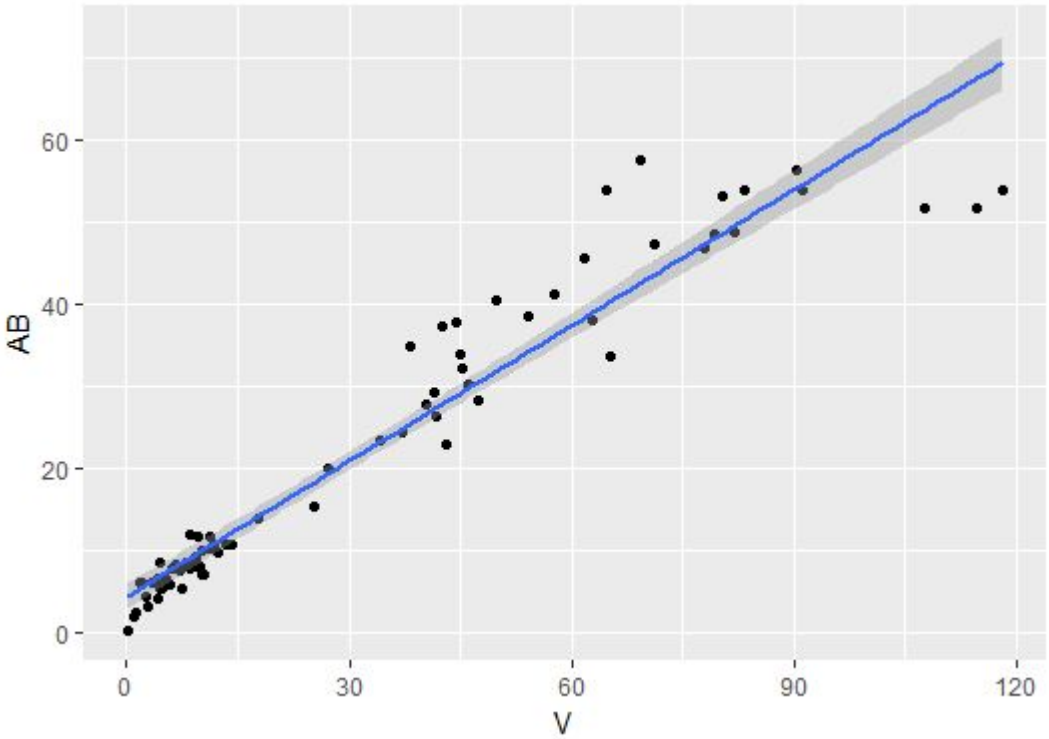
Allometric relationships for biomass and leaf area of beech (*Fagus sylvatica* L)

HH Bartelink

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



jak to się stało?



Model liniowy

y=a*x+b

a - współczynnik kierunkowy, slope, regression coefficient

b - wyraz wolny, intercept

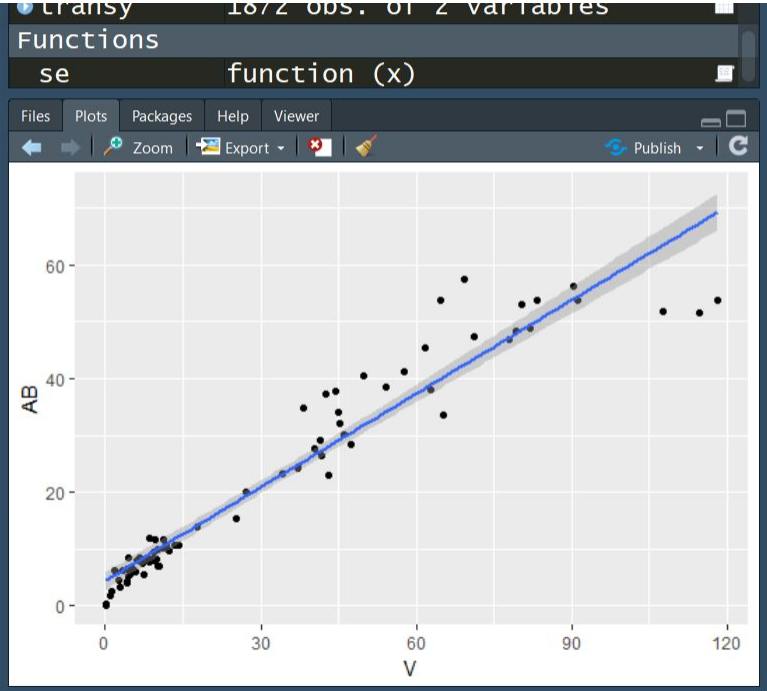
lm(V~AB,data=sosny)
Call:
lm(formula = AB ~ V, data = sosny)

Coefficients:
(Intercept) V
4.3984 0.5498

AB=4.3984+0.5498*V

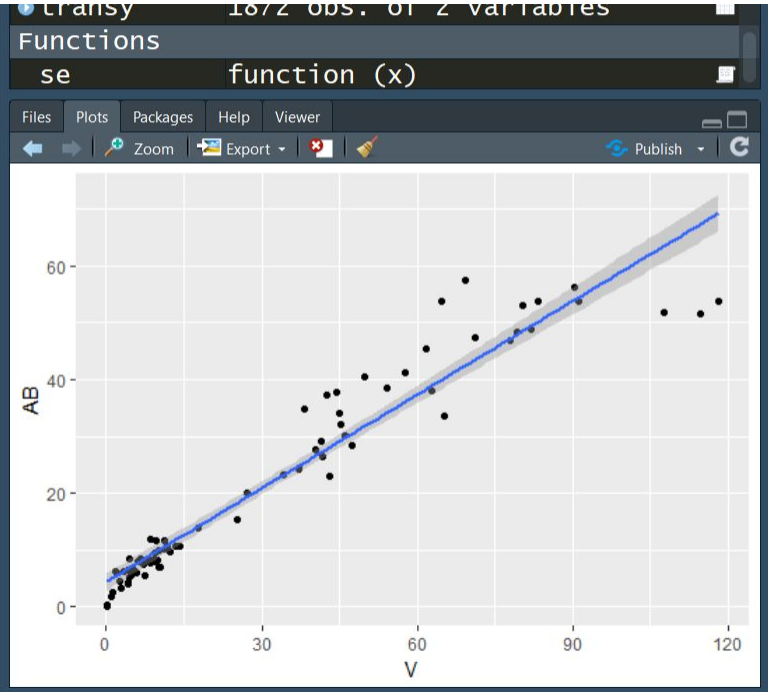
summary(lm(AB~V,data=sosny))

```
> 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

>
```



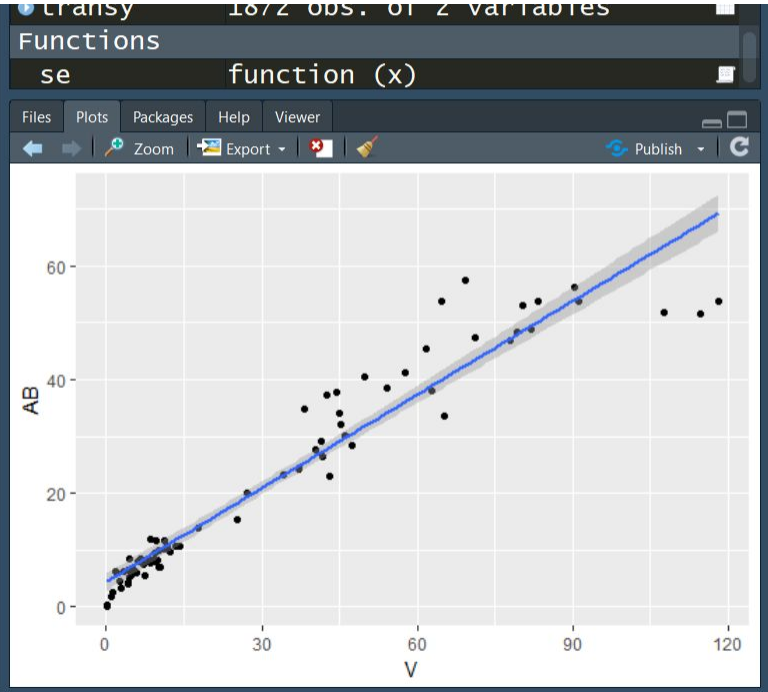
summary(lm(AB~V,data=sosny))

```
> 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

>
```

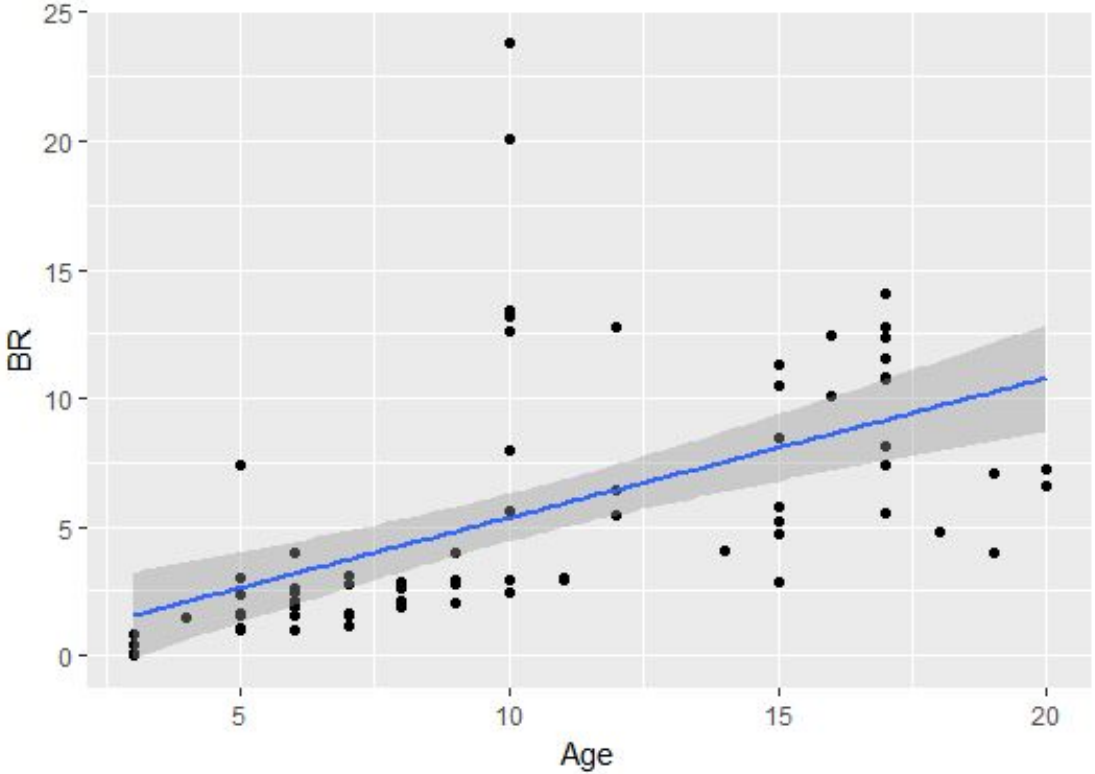


summary(lm(AB~V,data=sosny))

```
> 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

>
```



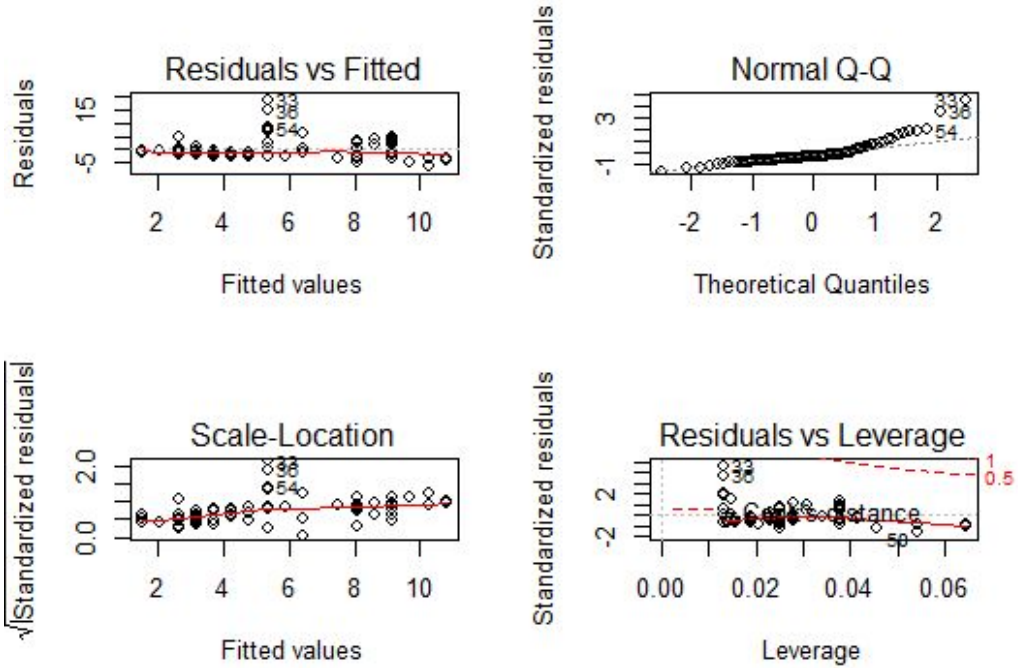
a tutaj? lepszy czy gorszy?

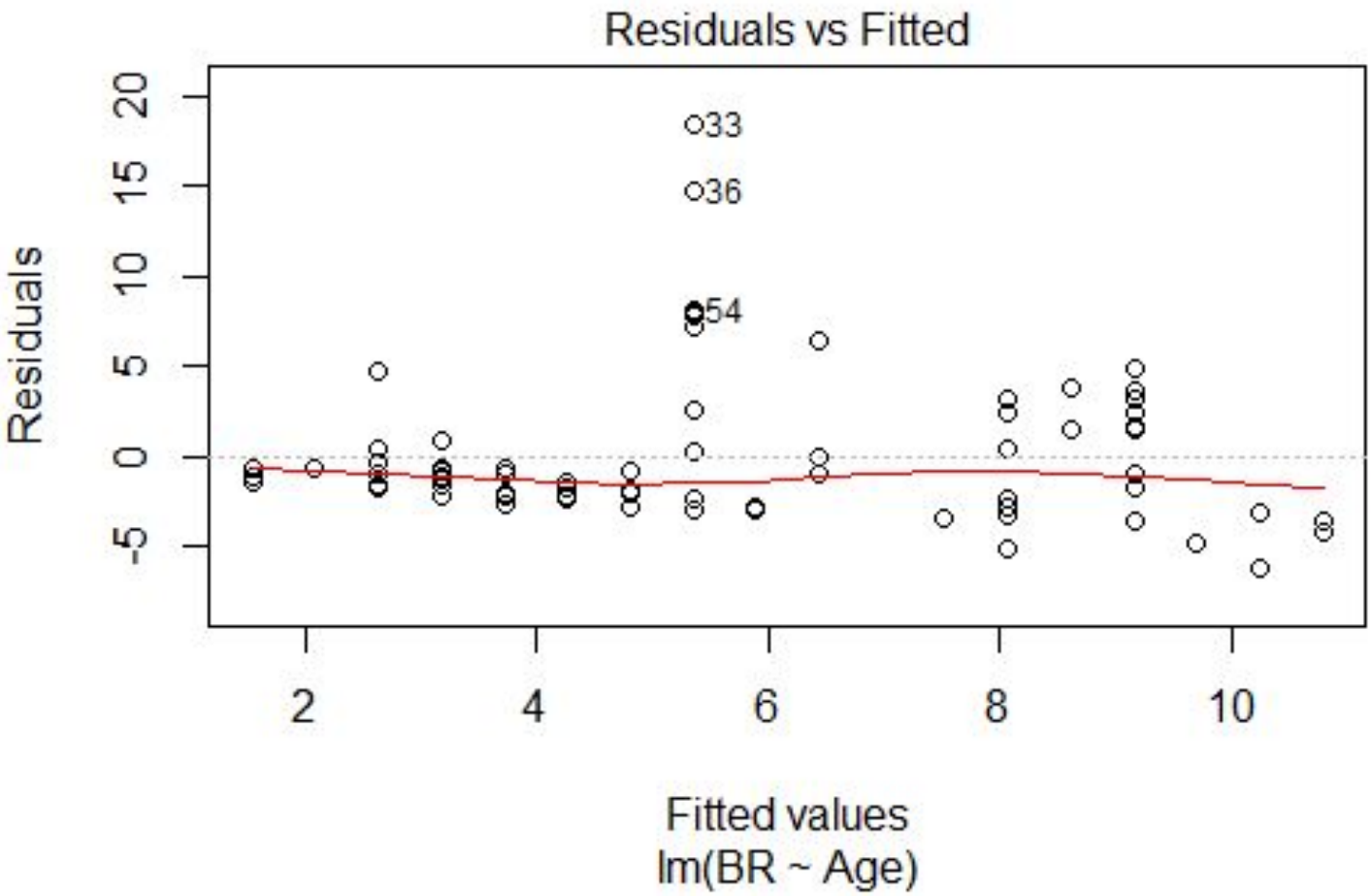


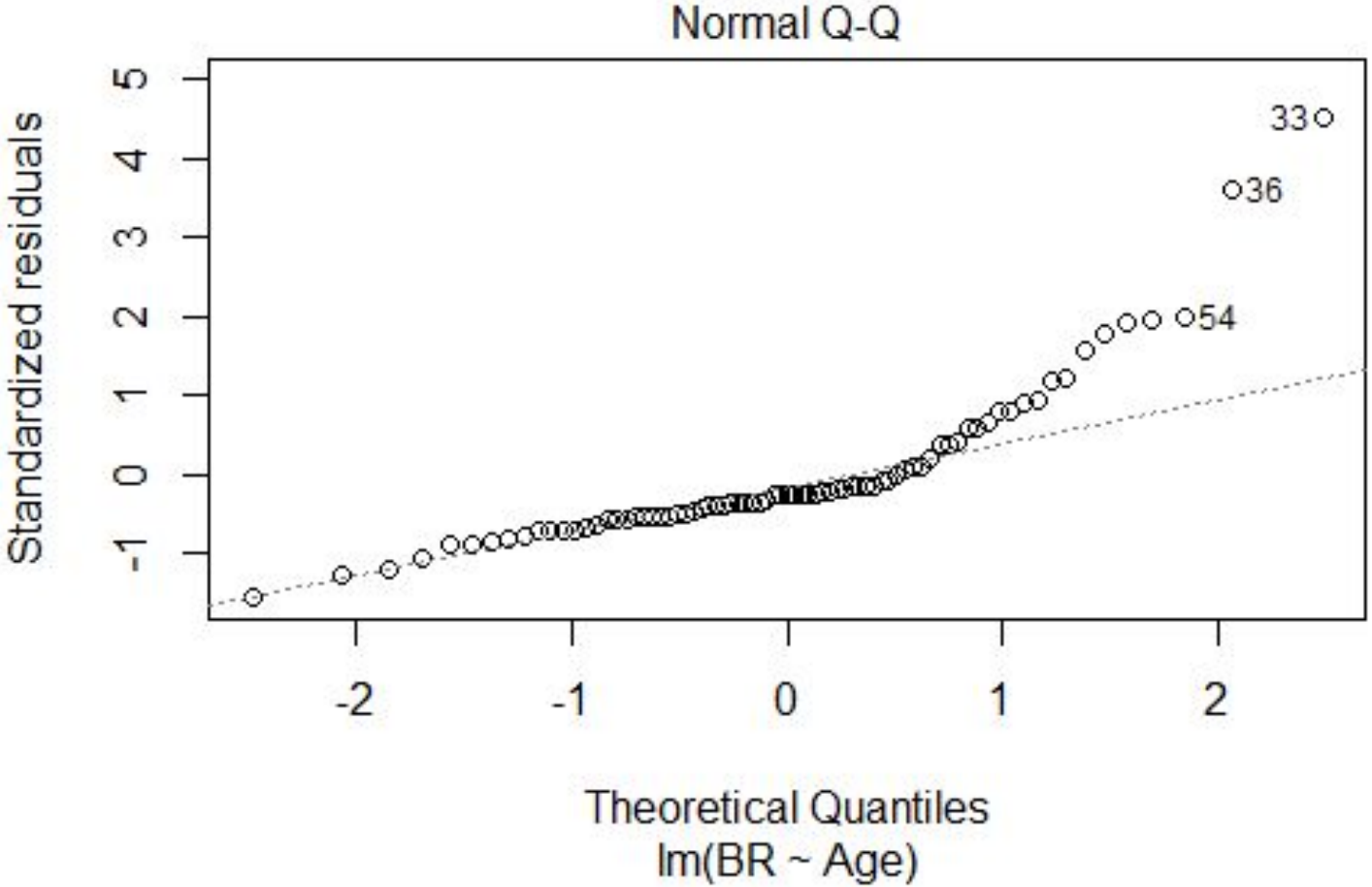
```
)
> summary(lm(BR~Age,data=sosny))

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

>
```

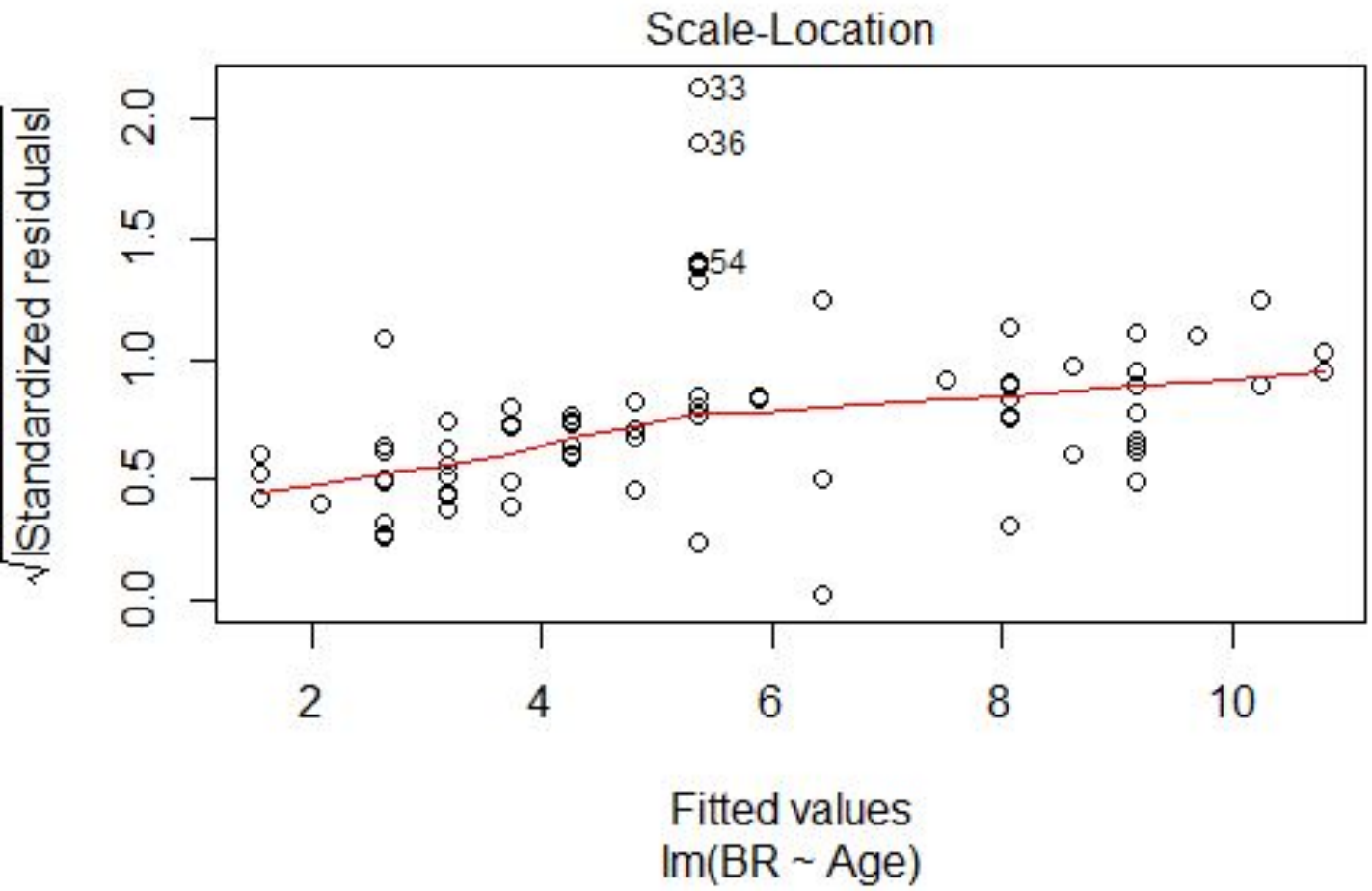
Diagnostyka modeli

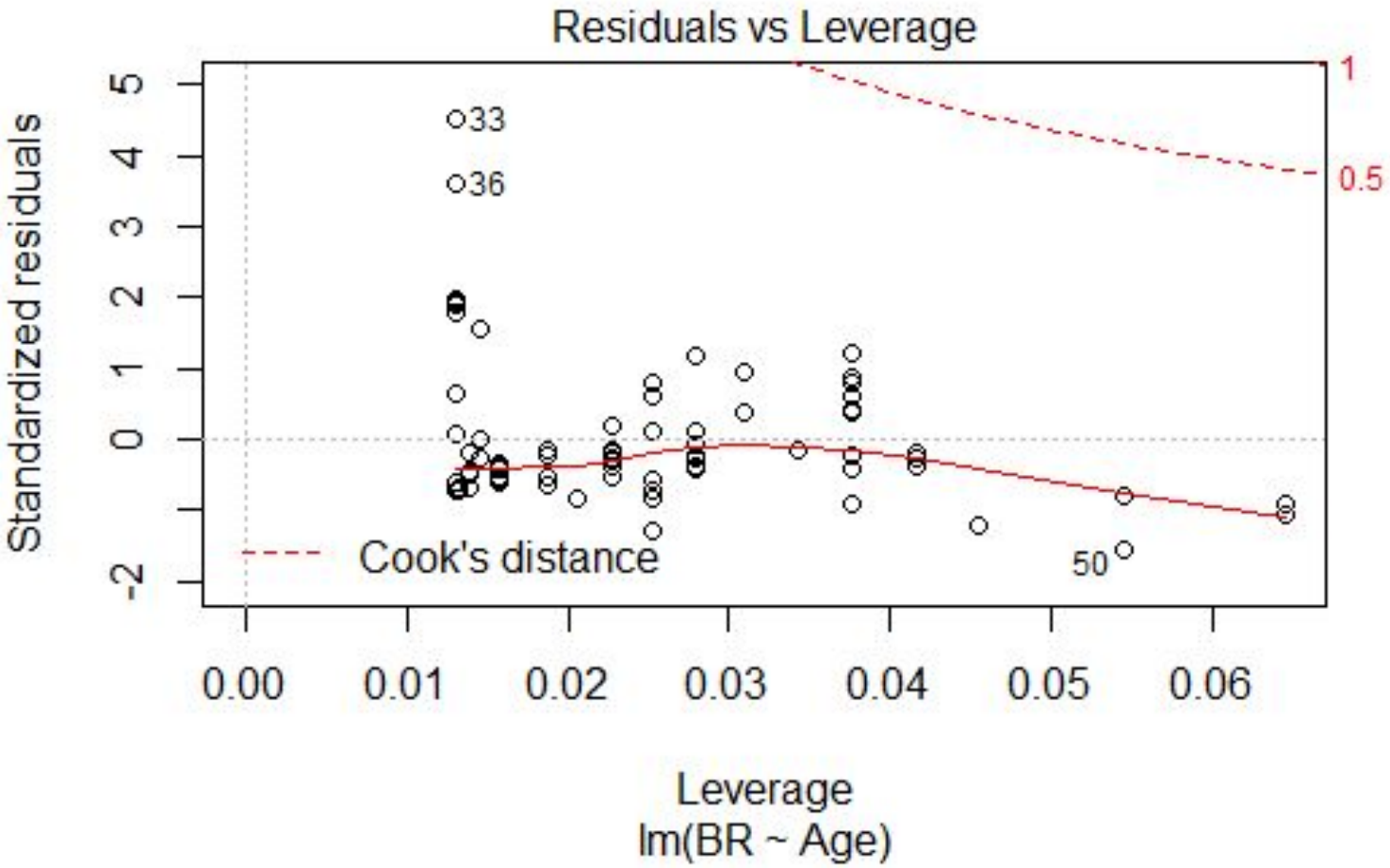
```
par(mfrow=c(2,2)) #podział wykresu na 4
plot(lm(BR~Age,data=sosny))
```



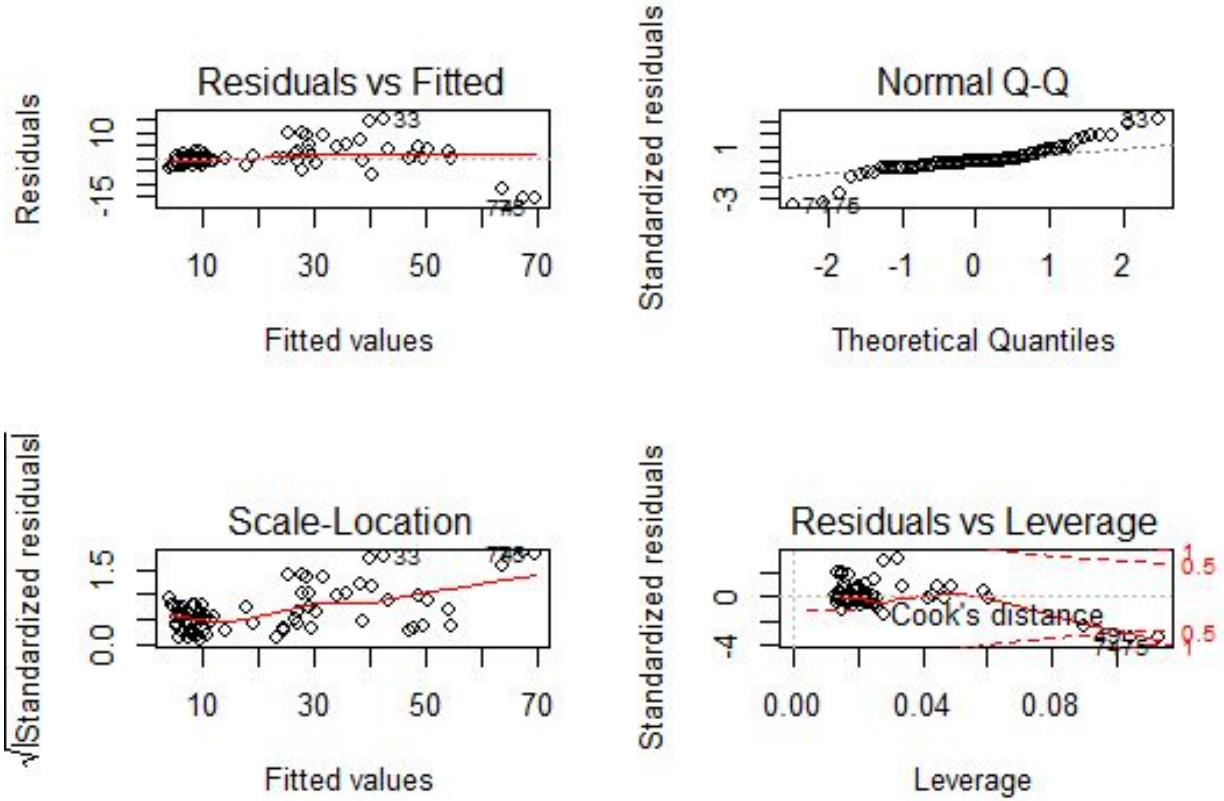

Residuals vs Fitted
Residuals
20
15
10
5
0
-5
33
36
54
2
4
6
8
10
Fitted values
lm(BR ~ Age)


Normal Q-Q
Standardized residuals
-1
0
1
2
3
4
5
33
36
54
-2
-1
0
1
2
Theoretical Quantiles
lm(BR ~ Age)

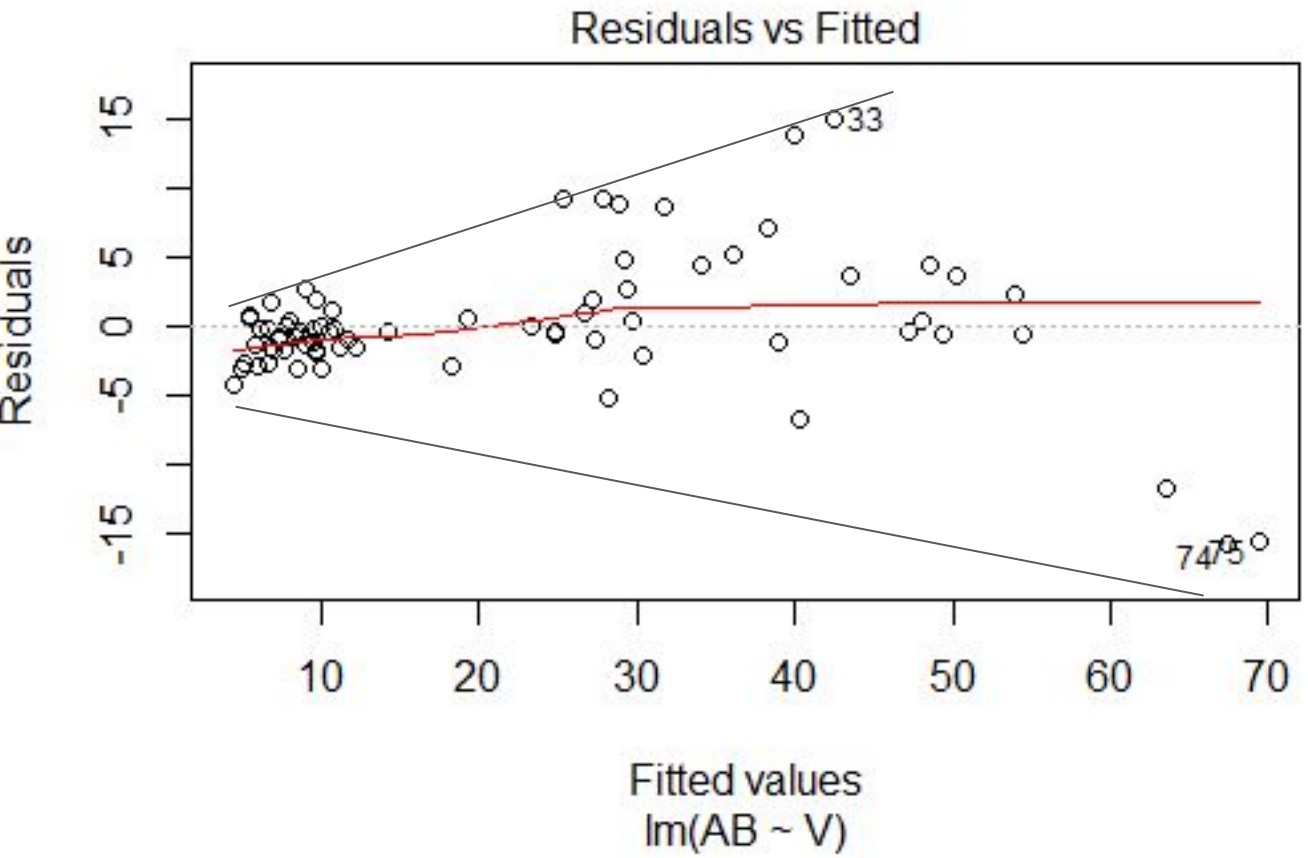

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
33
36
54
2
4
6
8
10
Fitted values
lm(BR ~ Age)

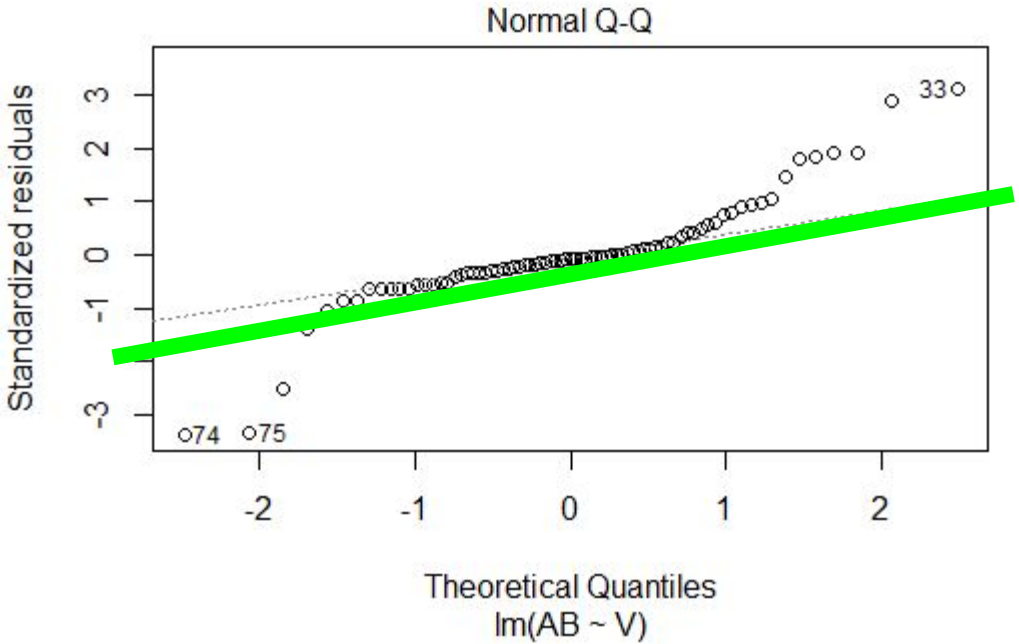

Residuals vs Leverage
Standardized residuals
Leverage
lm(BR ~ Age)
Cook's distance
1
0.5
33
36
50
-2
0
1
2
3
4
5
0.00
0.01
0.02
0.03
0.04
0.05
0.06

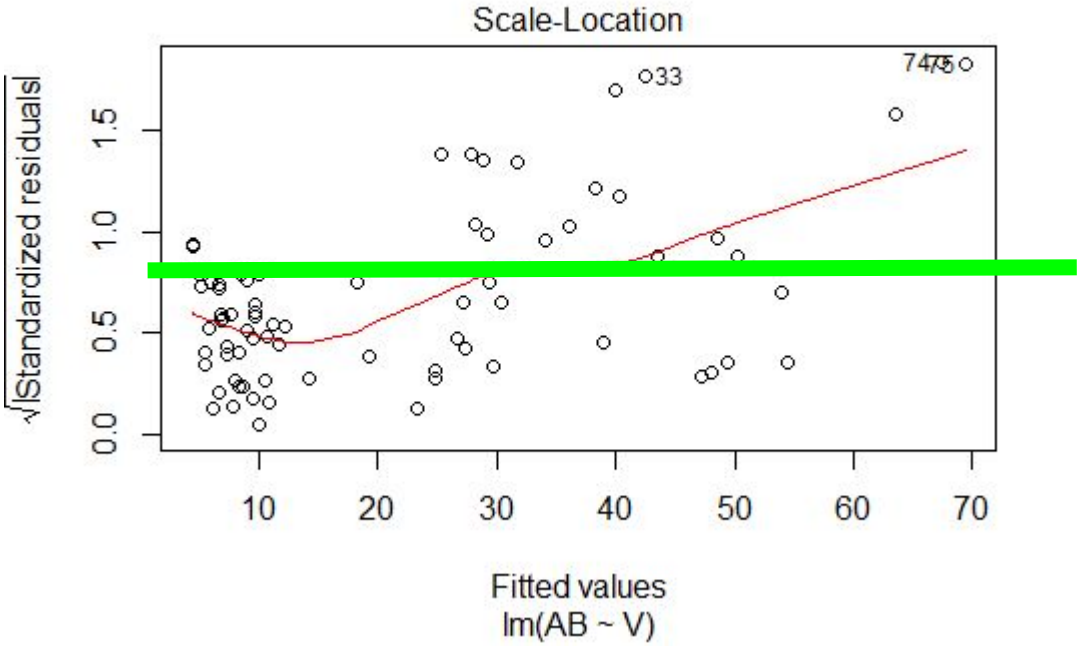
Jak wygląda dobry model? AB~V...

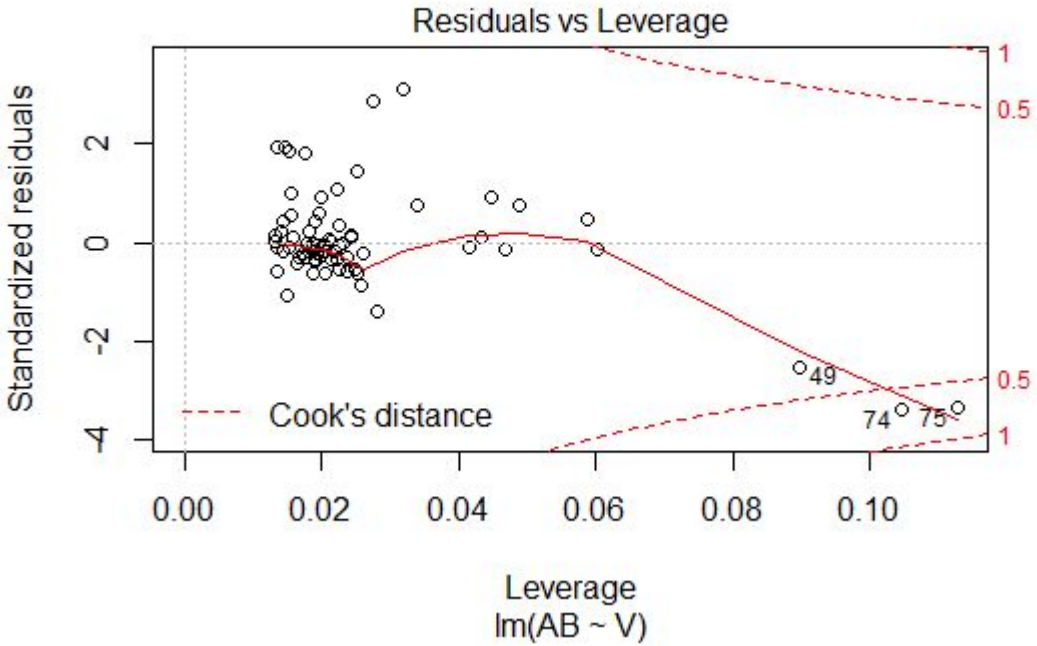


heteroskedastyczność!




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(AB ~ V)
33
74
75
-3
-1
0
1
2
3
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
33
74
75
10
20
30
40
50
60
70
Fitted values
lm(AB ~ V)


Residuals vs Leverage
Standardized residuals
Leverage
lm(AB ~ V)
Cook's distance
1
0.5
0.5
1
49
74
75
0.00
0.02
0.04
0.06
0.08
0.10
2
0
-2
-4

Co się stało?

model gorszy wg R2 (BR~age) okazał się mieć mniej problemów

model lepszy (AB~V) - heteroskedastyczność

Czy można tylko oceniać na podstawie R2?

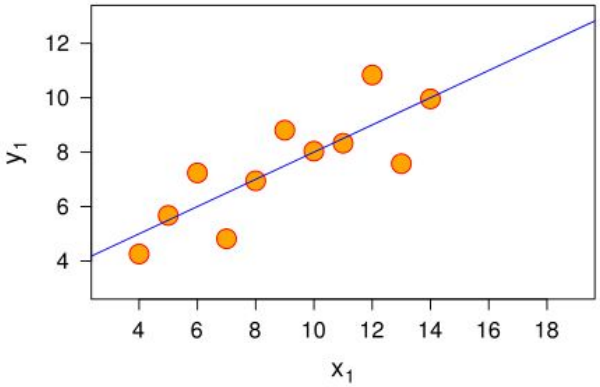
NIE!

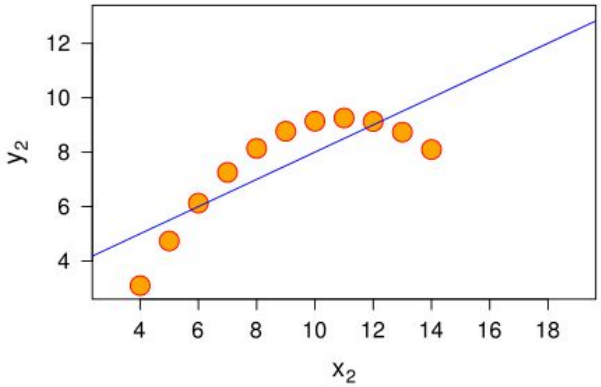
R2 nie uwzględnia kształtu rozkładu, heteroskedastyczności, outlierów…

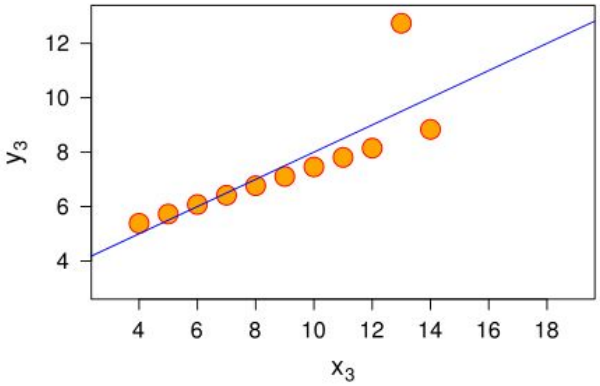
kwadrat Anscombe’a

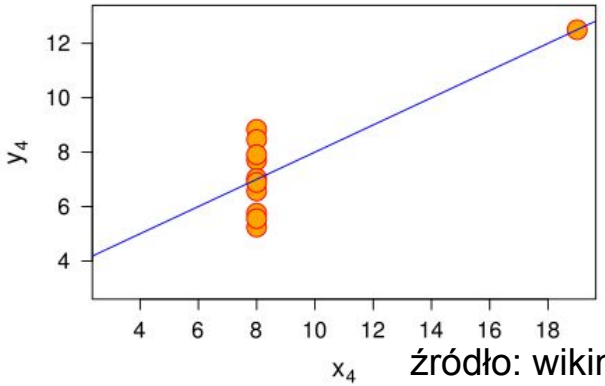
Kwadrat Ascombe’a

średnia y=7,5 średnia x=9
współczynnik r2=0,816
równanie regresji:
y=3+0,5*x









źródło: wikimedia.org

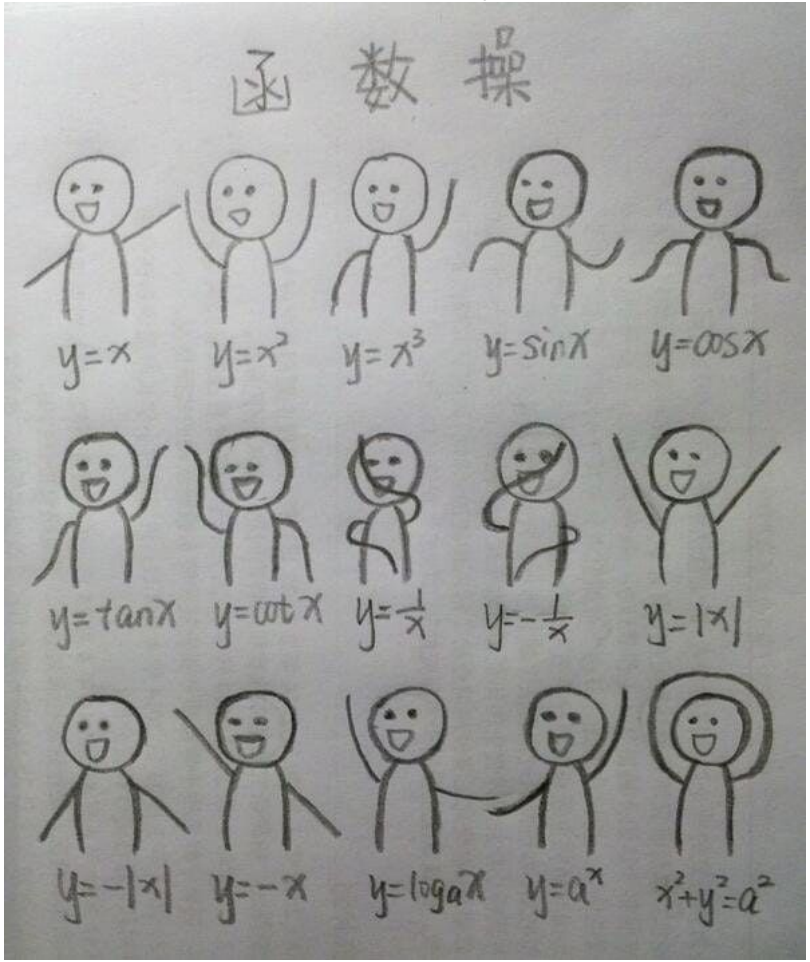
Co zrobić jak nie jest tak różowo?

transformacje (log, skalowanie…)

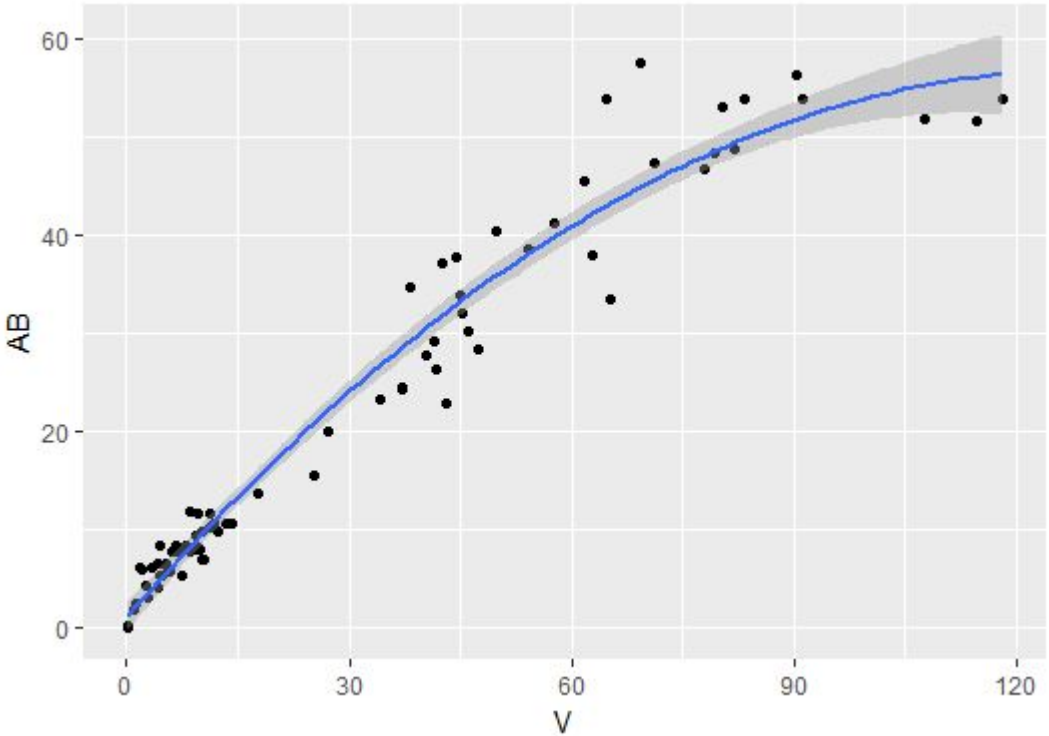
inny typ rozkładu

inny typ modelu

https://www.reddit.com/r/funny/comments/21h32s/dancing_math/



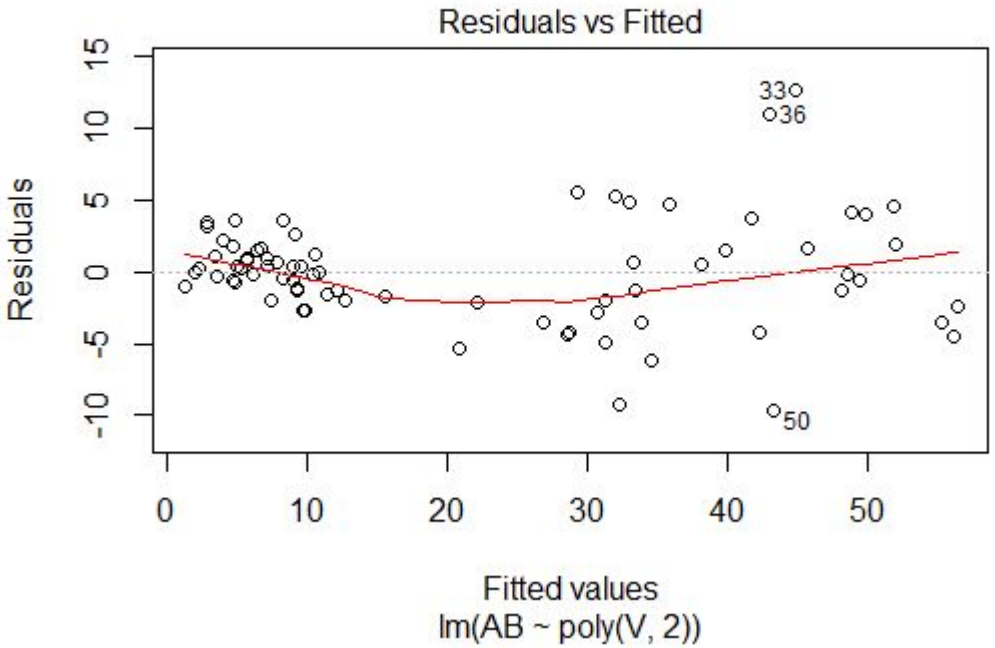
Parabole tańczą y=ax2+bx+c

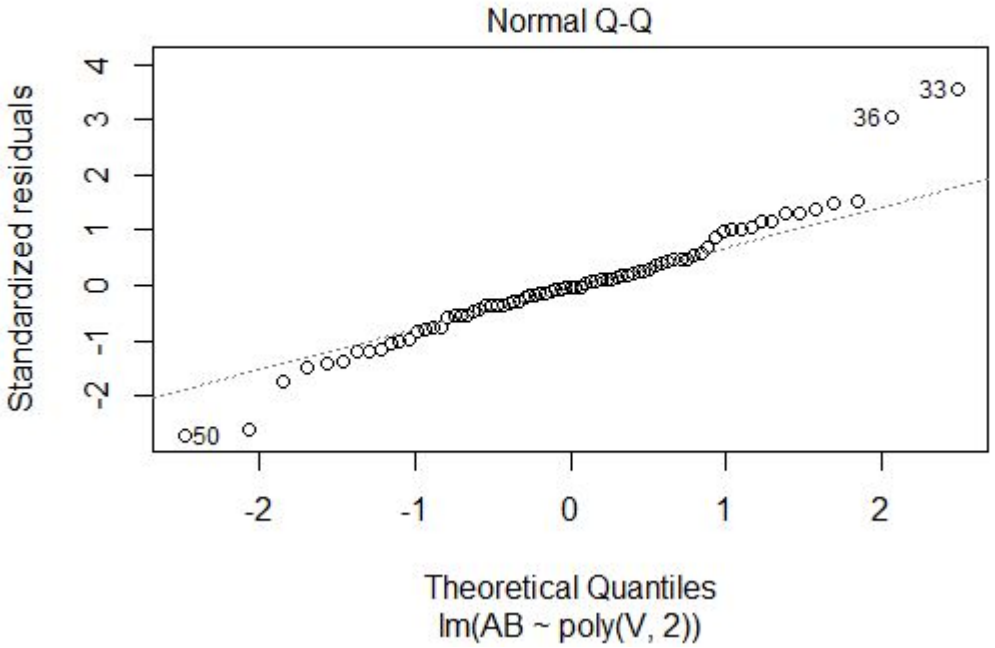


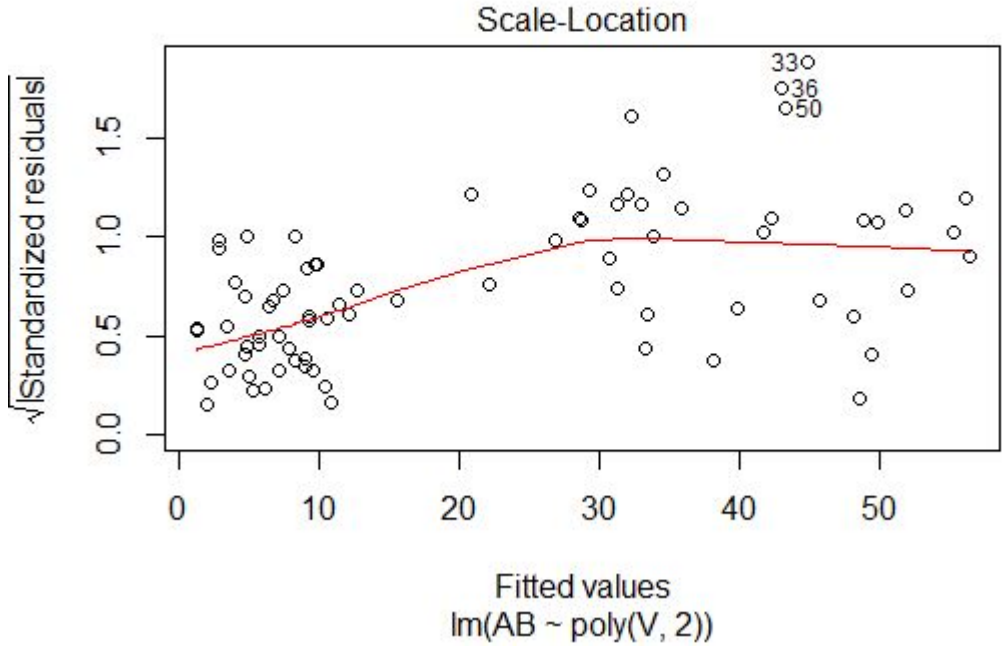
ggplot(sosny, aes(x=V,y=AB))+geom_point()+
geom_smooth(method='lm',**formula=y~poly(x,2)**)

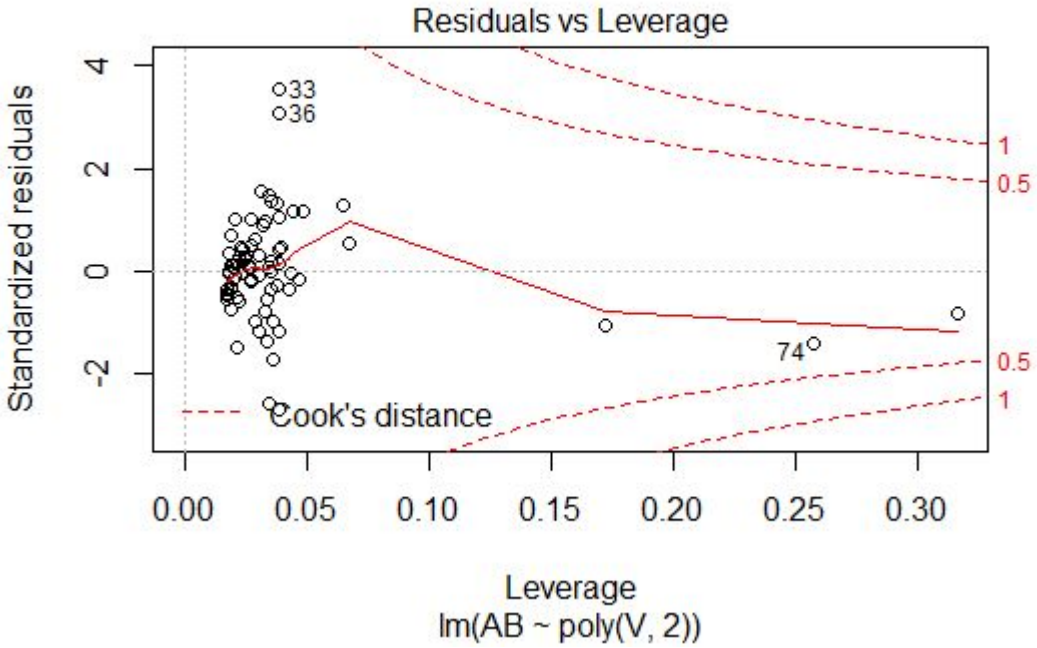
```
> summary(lm(AB~poly(V,2),data=sosny))

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


Residuals vs Fitted
Residuals
Fitted values
lm(AB ~ poly(V, 2))
33
36
50
15
10
5
0
-5
-10
0
10
20
30
40
50


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(AB ~ poly(V, 2))
33
36
50
-2
-1
0
1
2
3
4


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
33
36
50
0
10
20
30
40
50
Fitted values
lm(AB ~ poly(V, 2))


Residuals vs Leverage
Standardized residuals
4
2
0
-2
33
36
74
1
0.5
0.5
1
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(AB ~ poly(V, 2))

Lepiej:)

Jak głęboko wchodzić?

Cel: predykcja/eksploracja?

Zastosowanie i oczekiwana dokładność

Przyzwolenie na błędy?

Ilość analizowanych danych…

inne miary jakości

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~1, czyli wstawiamy wszędzie średnią

porównanie jakości dwóch modeli między sobą

```
> 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
```

for all biomass components and volume. For each equation we calculated ten regression models:

$W = a \times D^b$

$W = a + b \times D^2$

$W = a + b \times \log(D)$

$W = a + (b/D)$

$W = a \times (D^2H)^b$

$W = a \times D^b \times H^c$

$W = a + b \times \log(D^2H)$

$W = a + b \times D^2 + c \times H$ (8)

$W = a + b \times (D^2H)$ (9)

$W = a + b \times D^2 + c \times H^2$ (10)

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: 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[a,b,*], Marcin K. Dyderski[a,b], Kamil Gęsikiewicz[a], Paweł Horodecki[a], Agnieszka Cysewska[b], Sylwia Wierczyńska[b], Karol Maciejczyk[b]

[a] Institute of Dendrology, Polish Academy of Sciences, Parkowa 5, 62-035 Kórnik, Poland

[b] Poznań University of Life Sciences, Faculty of Forestry, Department of Game Management and Forest Protection, Wojska Polskiego 71c, 60-625 Poznań, Poland

implementacja

liniowy - lm() nieliniowy - nls()

model.liniowy<-lm(masa~D,data=dane)

model.liniowy<-lm(masa~log(D), data=dane)

modelnieliniowy<-nls(masa~a*D^b, data=dane, start=list(a=1,b=-1)

formuła, dane, (start)

model nieliniowy

```
nls(AB~a*V^b,data=sosny,start=list(a=1,b=1))
Nonlinear regression model
  model: AB ~ a * V^b
   data: sosny
     a      b
2.0119 0.7233
 residual sum-of-squares: 1297

Number of iterations to convergence: 6
Achieved convergence tolerance: 1.784e-06
```

nls(AB~a*V^b,data=sosny,start=list(a=1,b=exp(99999999)))

Error in numericDeriv(form[[3L]], names(ind), env) :

Brakuje wartości lub wyprodukowano wartości nieskończone podczas wyliczania modelu

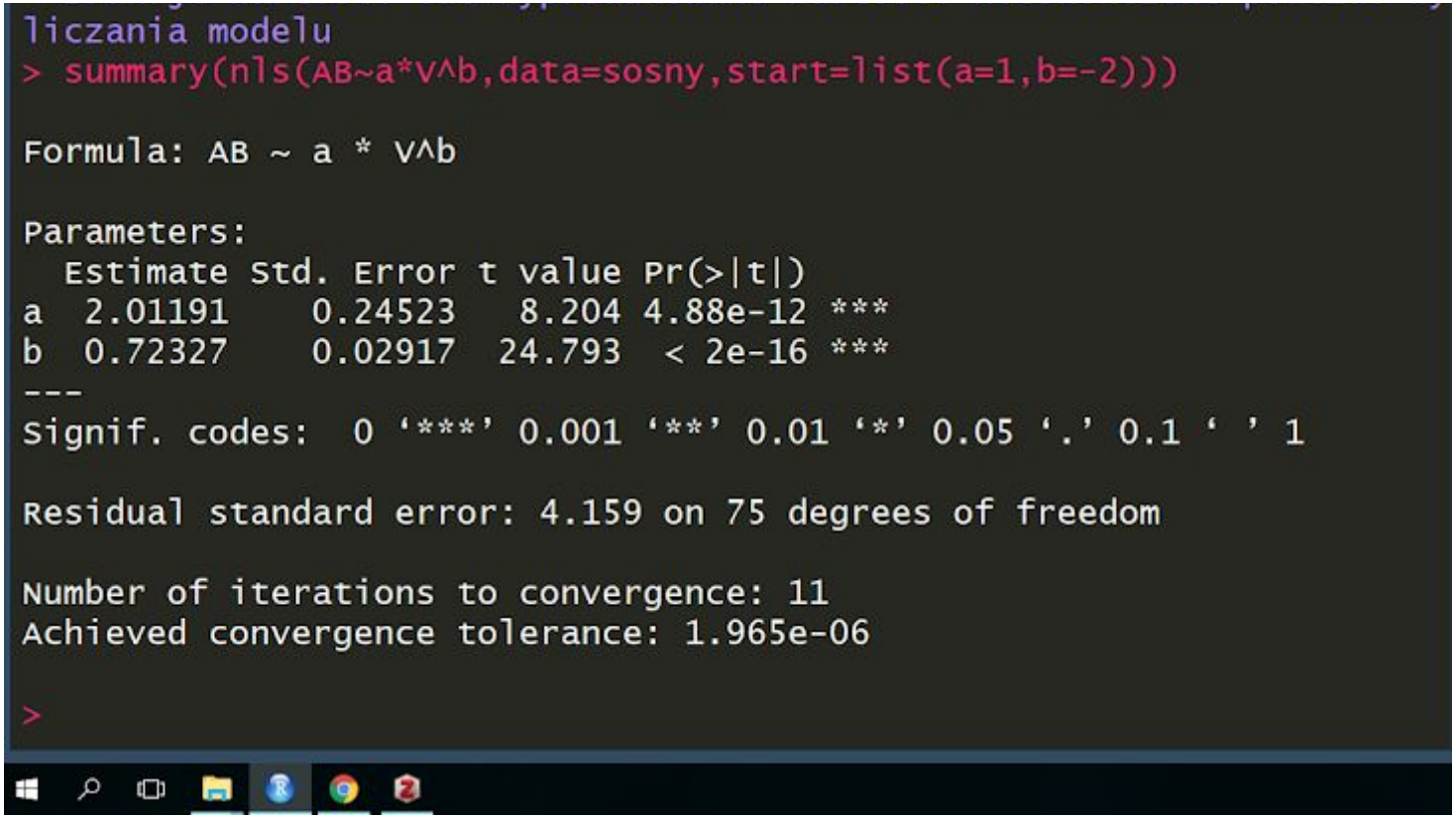
Parametry startowe

literatura

brute force - dopasowanie metodą prób i błędów

dobra rada - zacząć od 1, -1, potem małe i duże cyfry

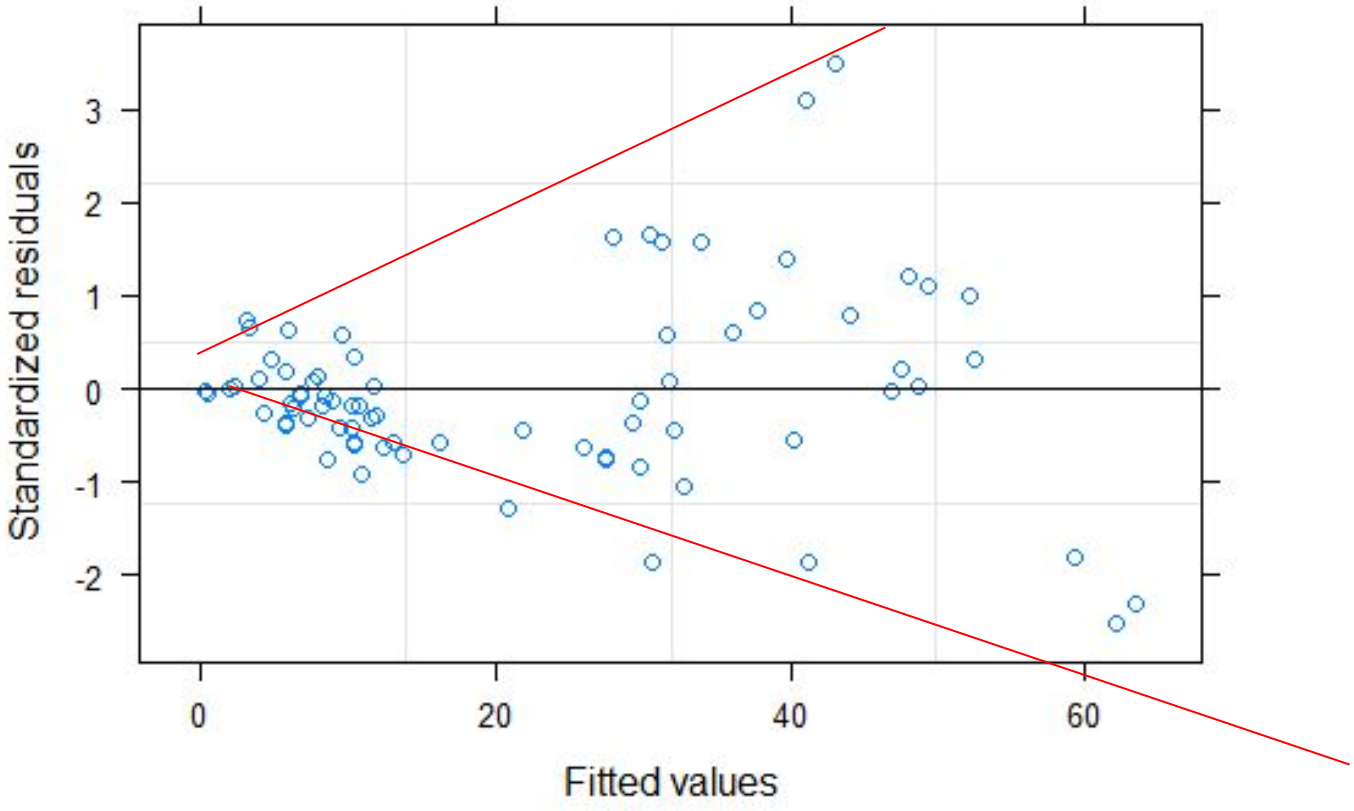
nie ma r2 :(i co ja teraz zrobię?



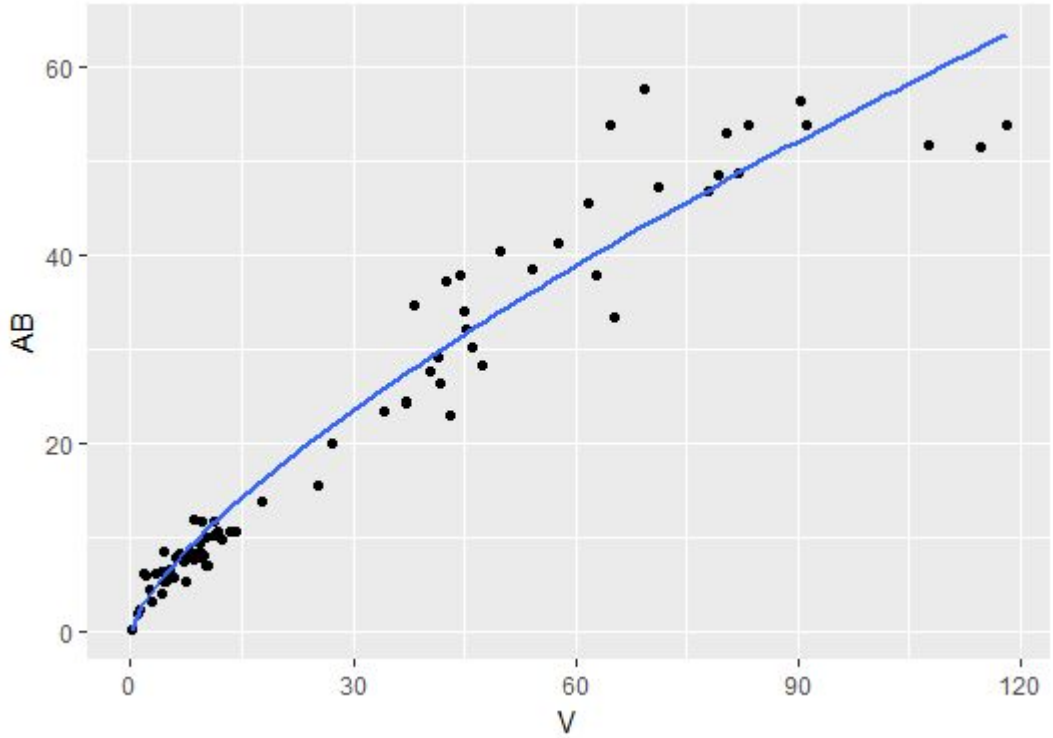
policzę AIC i porównam z modelem zerowym

```
> 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
> AIC(nls(AB~a*V^b,data=sosny,start=list(a=1,b=-2)))
[1] 441.9817
```

diagnostyka modelu:


Standardized residuals
3
2
1
0
-1
-2
0
20
40
60
Fitted values

```
ggplot(sosny, aes(y=AB,x=V))+geom_point()
+geom_smooth(method='nls',formula=y~a*x^b,
method.args=list(start=list(a=1,b=1)),se=F)
```



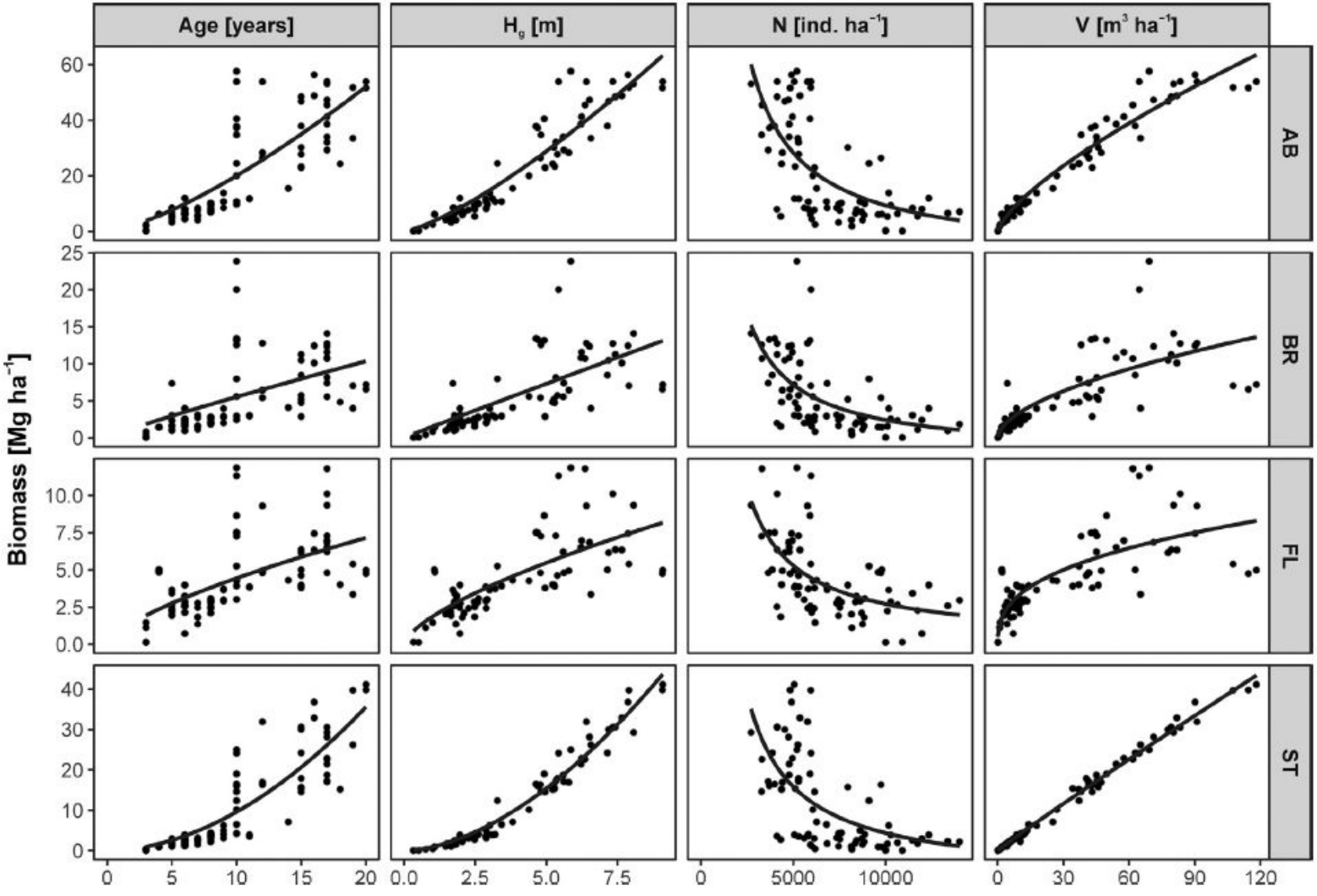
dlaczego tak?
za dużo punktów małych!
więcej nie znaczy lepiej;)

$$BCEF = a \times z^{b} \quad \text{(e. g. Peichl and Arain, 2007)} \tag{11}$$

$$BCEF = a + b \times e^{-z \times c} \quad \text{(e. g. Lehtonen et al. , 2004; Peichl and Arain, 2007; Jagodziski et al. , 2017)} \tag{12}$$

$$BCEF = a + b/z \quad \text{(e. g. Teobaldelli et al. ,2009; Wojtan et al. , 2011)} \tag{13}$$

$$BCEF = a + b/z^{c} \quad \text{(e. g. Teobaldelli et al. , 2009)} \tag{14}$$


Age [years]
H_g [m]
N [ind. ha^{-1}]
V [m^3 ha^{-1}]
AB
BR
FL
ST
Biomass [Mg ha^{-1}]
0
20
40
60
0
5
10
15
25
0.0
2.5
5.0
7.5
10.0
0
10
20
30
40
0
5000
10000
0
30
60
90
120

jak prezentować wyniki?

Table 7
Relationships between tree stand characteristics (predictors) and BCEFs for particular biomass components [Mg m^{-3}].

Biomass component	Predictor	Model type (Eq. no.)	a	SE	b	SE	c	SE	RMSE	R^2	AIC	AIC_0
AB	A	(12)	0.61691	0.06542	5.33448	1.03961	0.35147	0.05317	< 0.0001	0.709	44.473	134.220
	H_g	(12)	0.62412	0.05746	3.26428	0.35316	0.87235	0.10597	< 0.0001	0.729	39.084	–
	N	(11)	0.00376	0.00537	0.63451	0.15950	–	–	0.0119	0.186	120.599	–
	V	(12)	0.67209	0.04559	2.09454	0.15847	0.22621	0.02754	< 0.0001	0.764	28.410	–
BR	A	(12)	0.07917	0.07343	0.99264	0.22166	0.18209	0.06920	< 0.0001	0.438	−42.693	−2.843
	H_g	(12)	0.10830	0.06310	0.69410	0.13060	0.45800	0.18110	< 0.0001	0.378	−34.930	–
	N	(13)	0.46840	0.07462	−1080.0	422.1	–	–	< 0.0001	0.081	−7.283	–
	V	(12)	0.16099	0.03062	0.55175	0.08779	0.14350	0.03963	< 0.0001	0.428	−41.339	–
FL	A	(12)	0.09562	0.05773	5.67080	1.45618	0.42193	0.07180	< 0.0001	0.637	42.380	115.452
	H_g	(12)	0.08925	0.05133	3.09855	0.38049	1.01347	0.13006	< 0.0001	0.682	31.089	–
	N	(13)	0.91070	0.15630	−3192.17130	884.35630	–	–	< 0.0001	0.150	32.470	–
	V	(12)	0.12295	0.04423	1.86903	0.16629	0.26766	0.03813	< 0.0001	0.700	27.958	–
ST	A	(12)	0.37645	0.01009	6.30763	9.27127	1.11956	0.48019	< 0.0001	0.316	−174.461	−149.642
	H_g	(12)	0.38054	0.00548	3.65437	1.24671	5.30087	0.92841	< 0.0001	0.736	−246.734	–
	N	(11)	0.14574	0.08838	0.11308	0.06883	–	–	< 0.0001	0.035	−150.317	–
	V	(13)	0.37748	0.00526	0.03547	0.00240	–	–	< 0.0001	0.747	−252.212	–

skubany, skąd masz R2?

R2 dotyczy tylko modeli liniowych

w pozostałych przypadkach można podać (zaznaczając to wyraźnie) - **pseudo-R2**

bo czym jest tak naprawdę R2?

1-suma kwadratów odchyleń modelu/suma kwadratów odchyleń rzeczywistych

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

wikimedia.org

*Jak obliczyć R2 ręcznie?

```
model<-nls(AB~a*V^b,data=sosny,start=list(a=1,b=-2))

1-(sum(residuals(model)^2)/sum((sosny$AB-mean(sosny$AB))^2))
```

czyli 1 -

suma kwadratów reszt (reziduów): sum(residuals(model)^2)

przez

suma kwadratów odchyleń (sosny$AB - mean(sosny$AB))^2

*RMSE - pierwiastek średniego błędu kwadratowego

jakie są odchyły

sqrt(sum (**zmienna-**predict(model))^2 / length(**zmienna**)

pierwiastek sumy kwadratów odchyleń / liczba obserwacji -

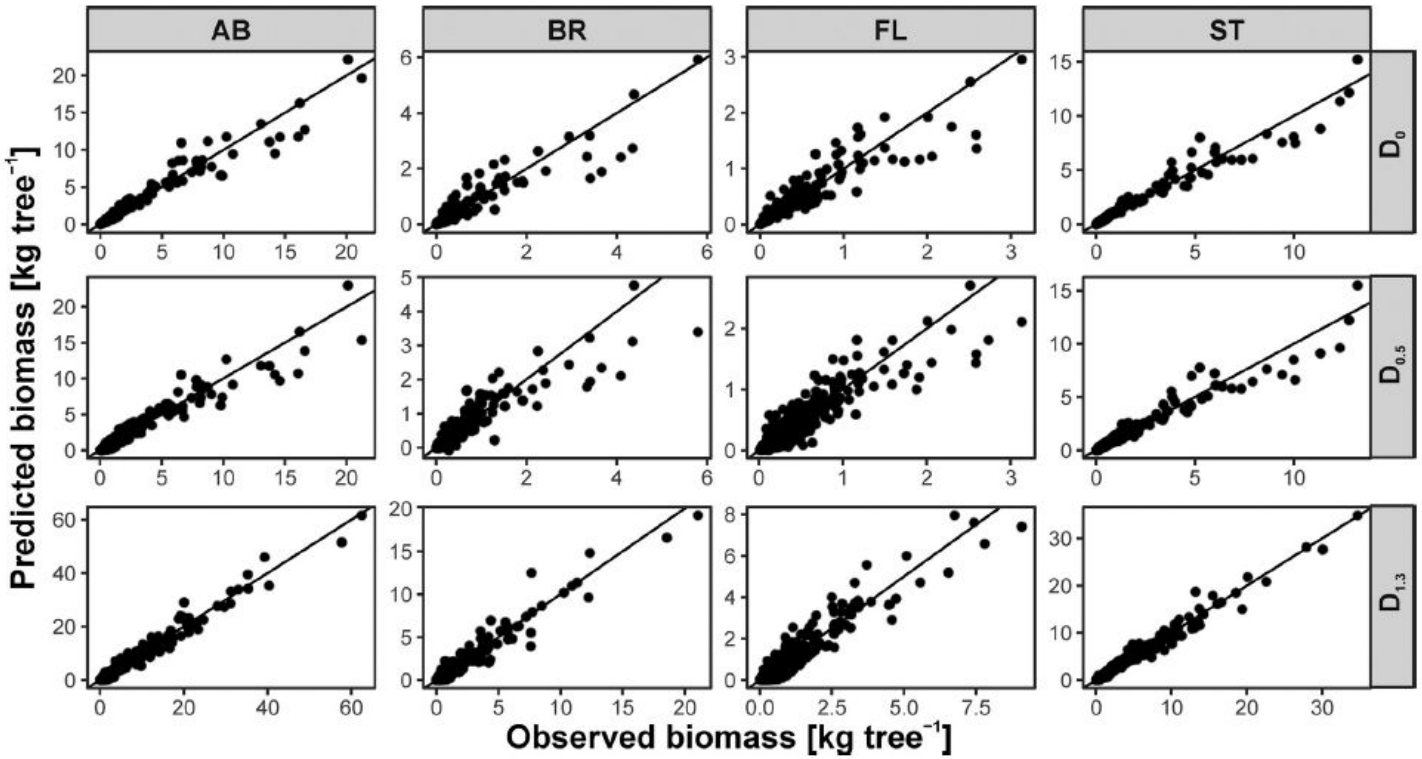
One function to rule the all - predict

Modele są po to, aby je wykorzystywać

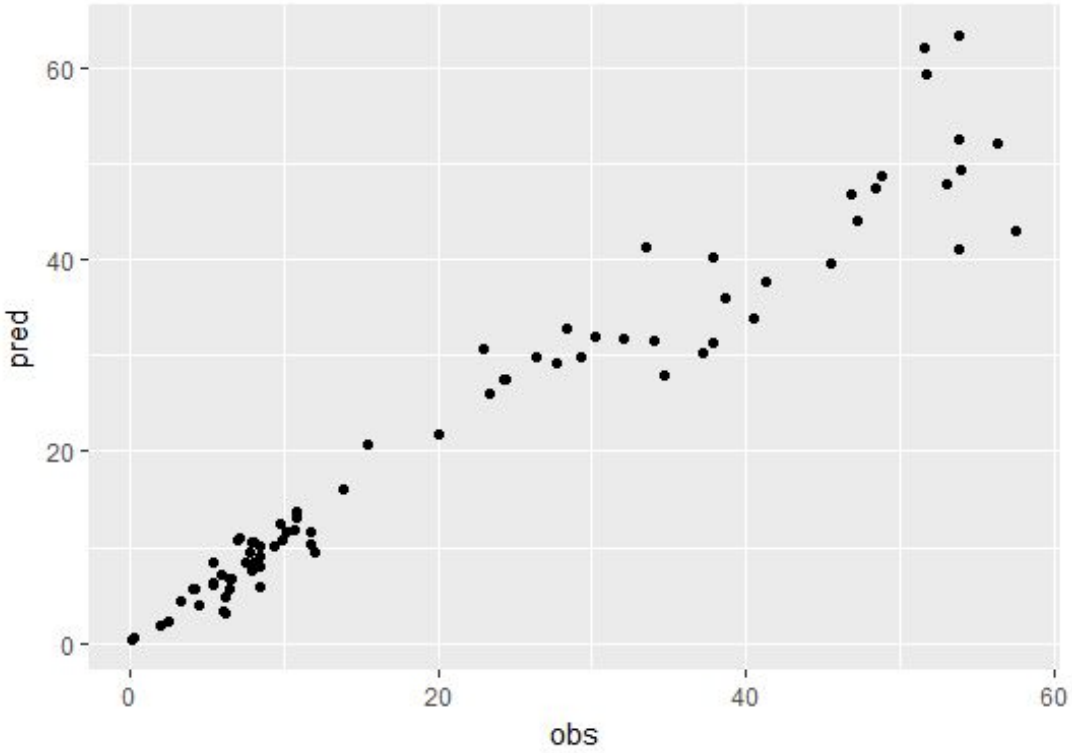
```
model<-nls(AB~a*V^b,data=sosny,start=list(a=1,b=-2))

nowedane<-data.frame(V=c(1,2,3,10))

predict(model,nowedane)

[1]  2.011908  3.321489  4.453419 10.638339
```

Inny sposób oceny - predicted vs. observed

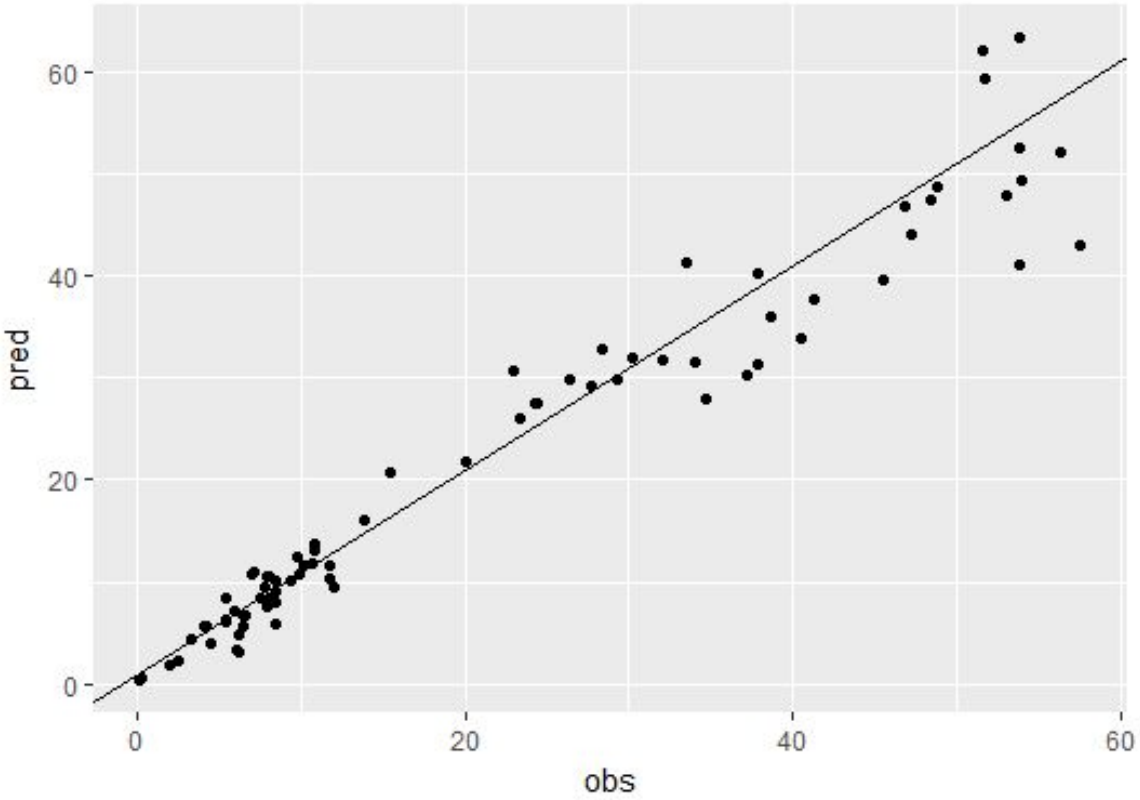
Forest Ecology and Management 409 (2018) 74–83



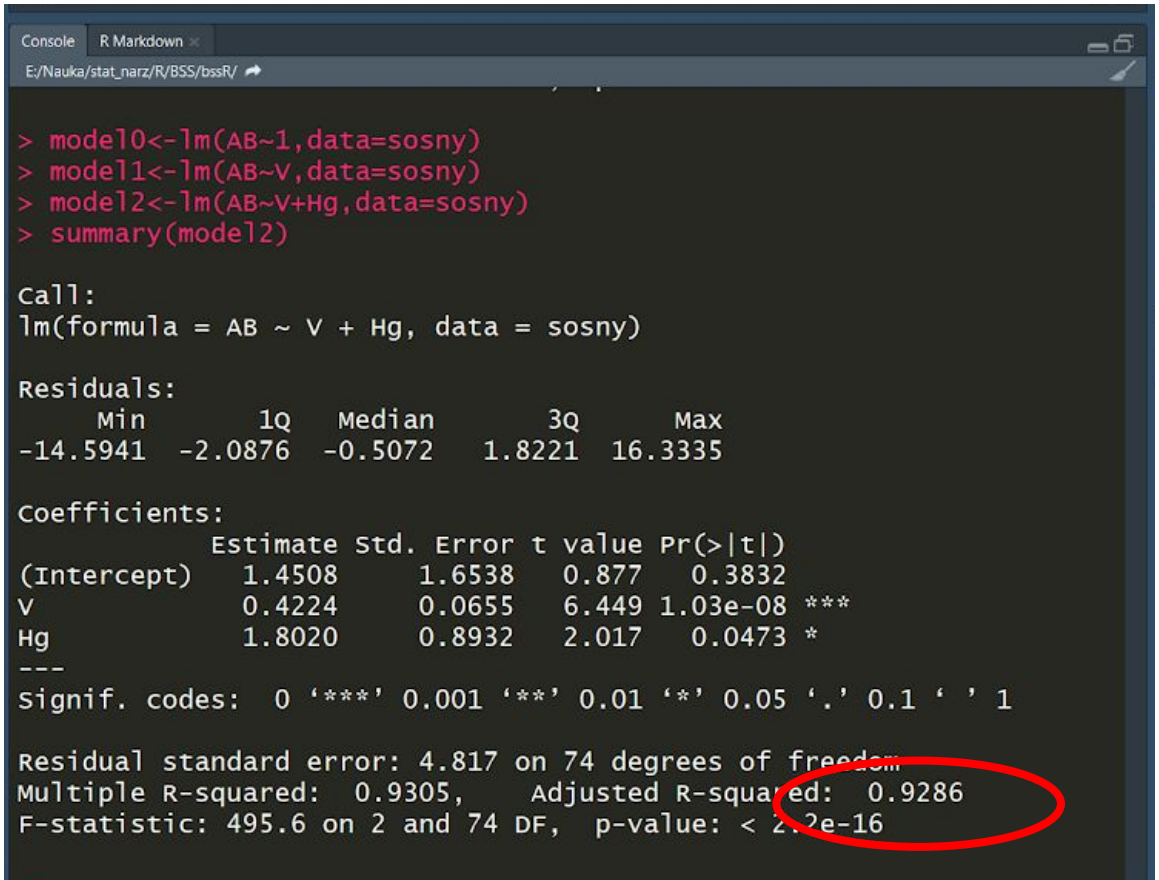
```
data.frame(obs=sosny$AB, pred=predict(model,sosny$V))%>%
ggplot(aes(x=obs,y=pred))+geom_point()
```



```
data.frame(obs=sosny$AB, pred=predict(model,sosny$V))%>%
ggplot(aes(x=obs,y=pred))+geom_point()+geom_abline(intercept=1)
```



Poprawianie jakości modelu



urosło nieznacznie, co z AIC?

```
> AIC(model1,model2,model0)
        df      AIC
model1  3 467.6902
model2  4 465.5678
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
```

VIF - variance inflation factor

Rule of Thumb VIF>10 => 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
```

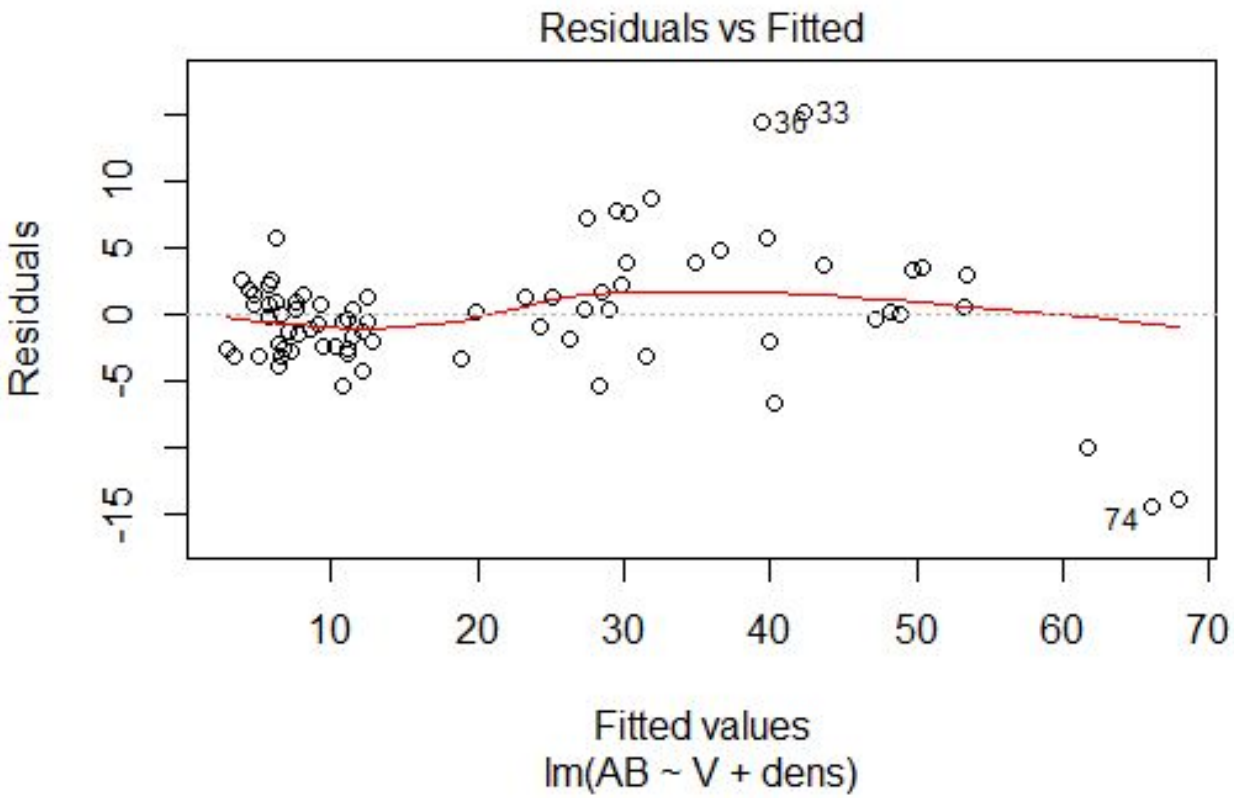
Dlaczego tak?

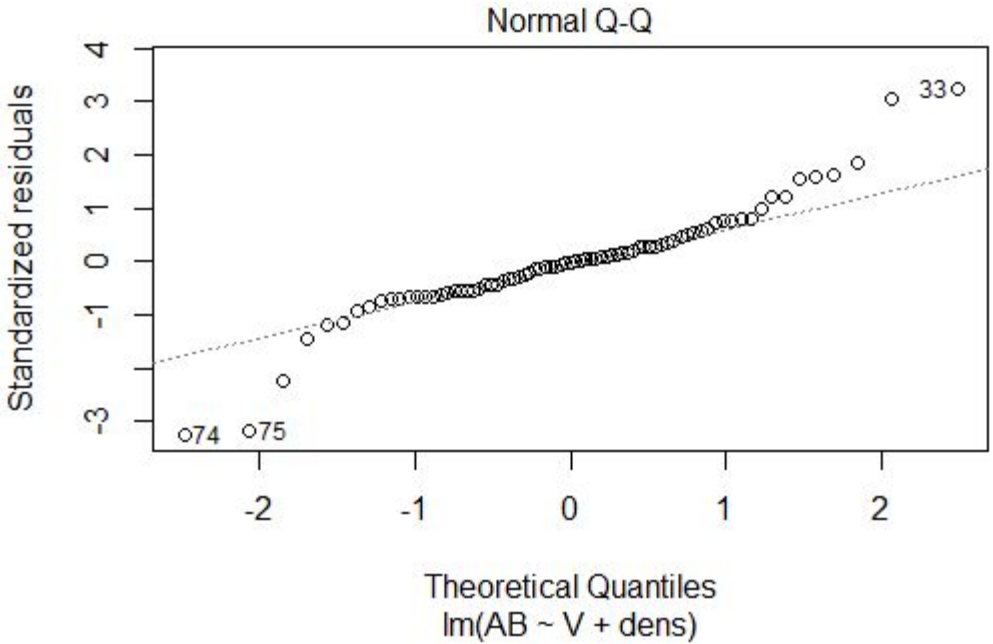
V=G*H*f

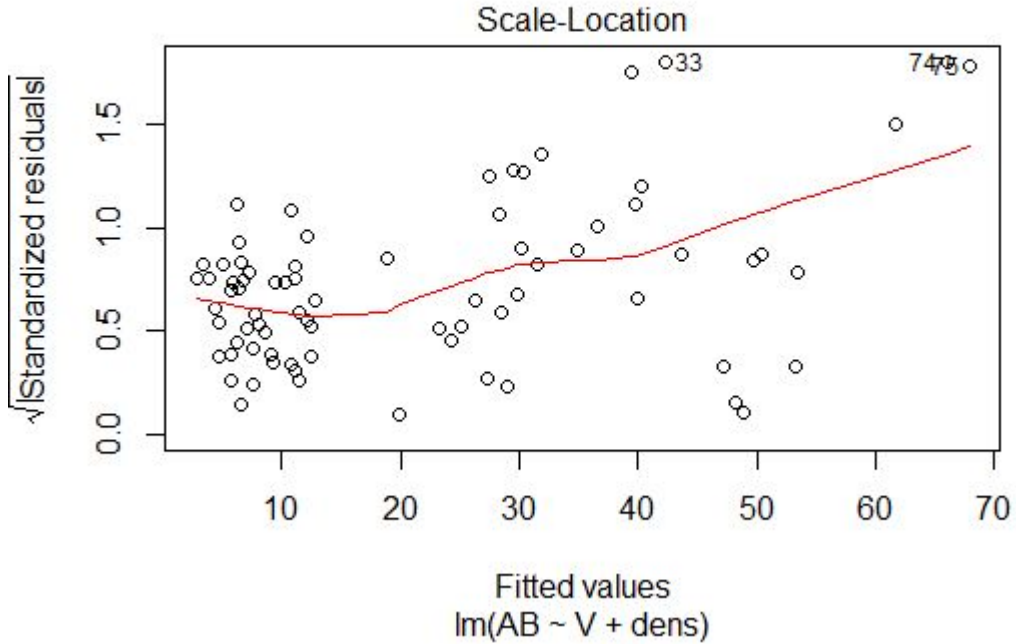
f - wskaźnik kształtu którego nie znamy

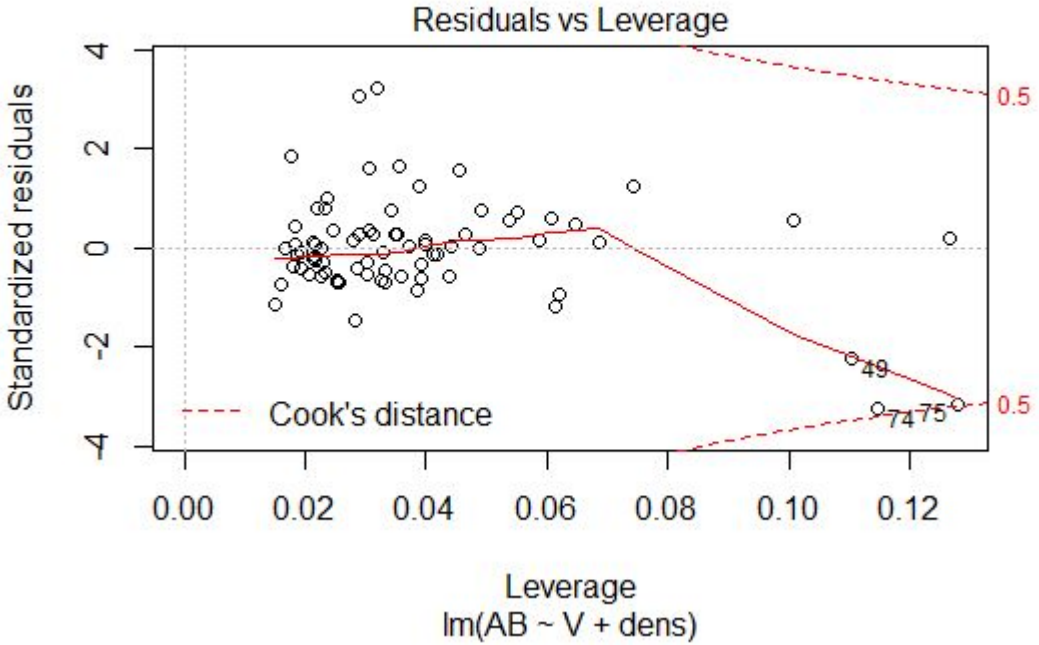
model5<-lm(AB~V+dens,data=sosny)

```
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
>
```


Residuals vs Fitted
Residuals
-15
-5
0
5
10
10
20
30
40
50
60
70
Fitted values
lm(AB ~ V + dens)
33
74


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-3
33
74
75
-2
-1
0
1
2
Theoretical Quantiles
lm(AB ~ V + dens)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
33
10
20
30
40
50
60
70
Fitted values
lm(AB ~ V + dens)


Residuals vs Leverage
Standardized residuals
Leverage
lm(AB ~ V + dens)
Cook's distance
0.5
0.5
49
74
75
4
2
0
-2
-4
0.00
0.02
0.04
0.06
0.08
0.10
0.12

i tak dalej...

Co trzeba sprawdzać?

VIFy (chyba że się zna korelacje między cechami i unika współliniowości)

AIC (pamiętamy o modelu zerowym)

R2

wykresy diagnostyczne… - chyba że nie zależy nam na dokładności

niebezpieczne narzędzie

```
> 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
```

GLM

generalized linear model - nie ma R2, przy założeniu rozkładu normalnego - rozszerzony lm

po co? do innych rozkładów

GLM z innymi rozkładami

glm(formula, data, family=**’rozklad’**)

family=’poisson’ - rozkład Poissona (dyskretny - liczby naturalne)

family=binomial(link=’logit’) - regresja logistyczna

regresja logistyczna

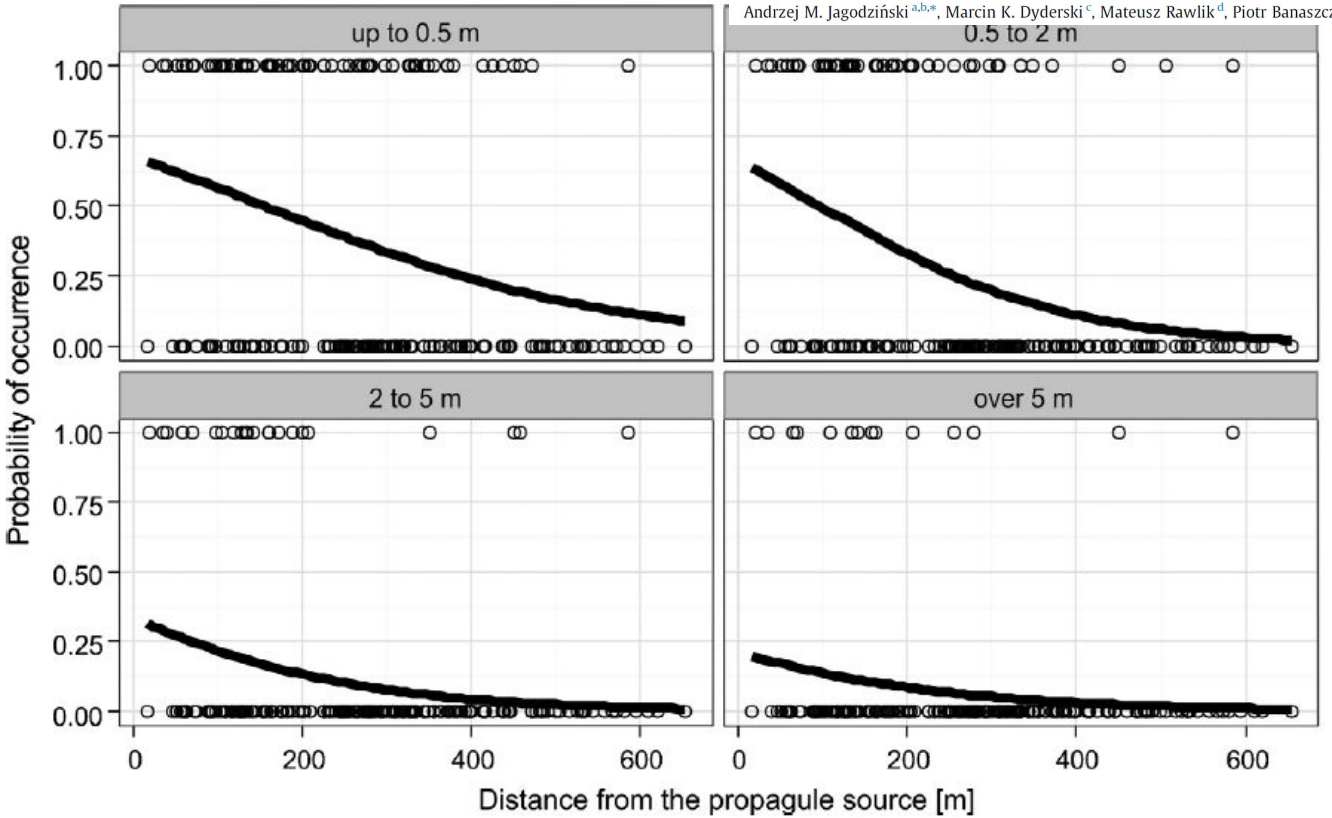
Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

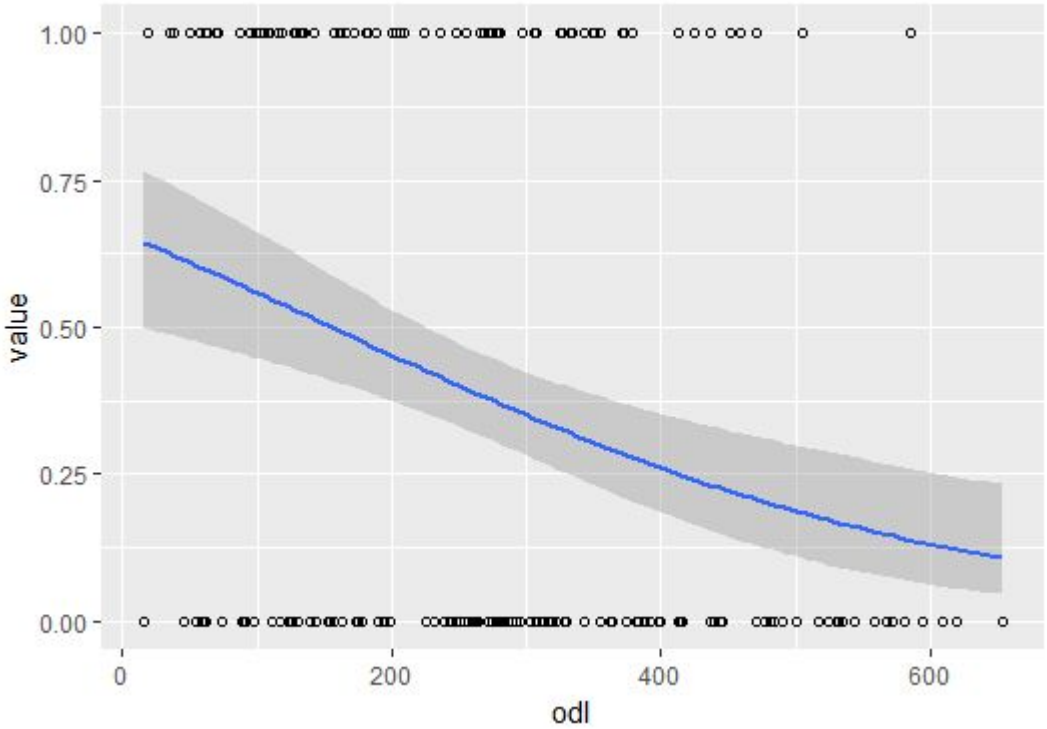
Plantation of coniferous trees modifies risk and size of *Padus serotina* (Ehrh.) Borkh. invasion – Evidence from a Rogów Arboretum case study

Andrzej M. Jagodziński [a,b,*], Marcin K. Dyderski [c], Mateusz Rawlik [d], Piotr Banaszczak [e]



```
> model<-glm(szans~odl,family=binomial(link="logit"))
> summary(model)

Call:
glm(formula = szans ~ odl, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4374  -0.9748  -0.7137   1.1419   1.9901

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.661478   0.318846   2.075 0.038024 *
odl         -0.004263   0.001120  -3.806 0.000141 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 268.48  on 200  degrees of freedom
Residual deviance: 252.23  on 199  degrees of freedom
AIC: 256.23

Number of Fisher Scoring iterations: 4

> predict(model)
```

```
ggplot(data=fig4df, aes(x=odl, y=value))+geom_point(shape=1)
+geom_smooth(method="glm", method.args=c(family="binomial"))
```



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

p-value zależy od n!

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

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

zmienne kategoryczne - nieliczbowe

analiza kowariancji - ANCOVA

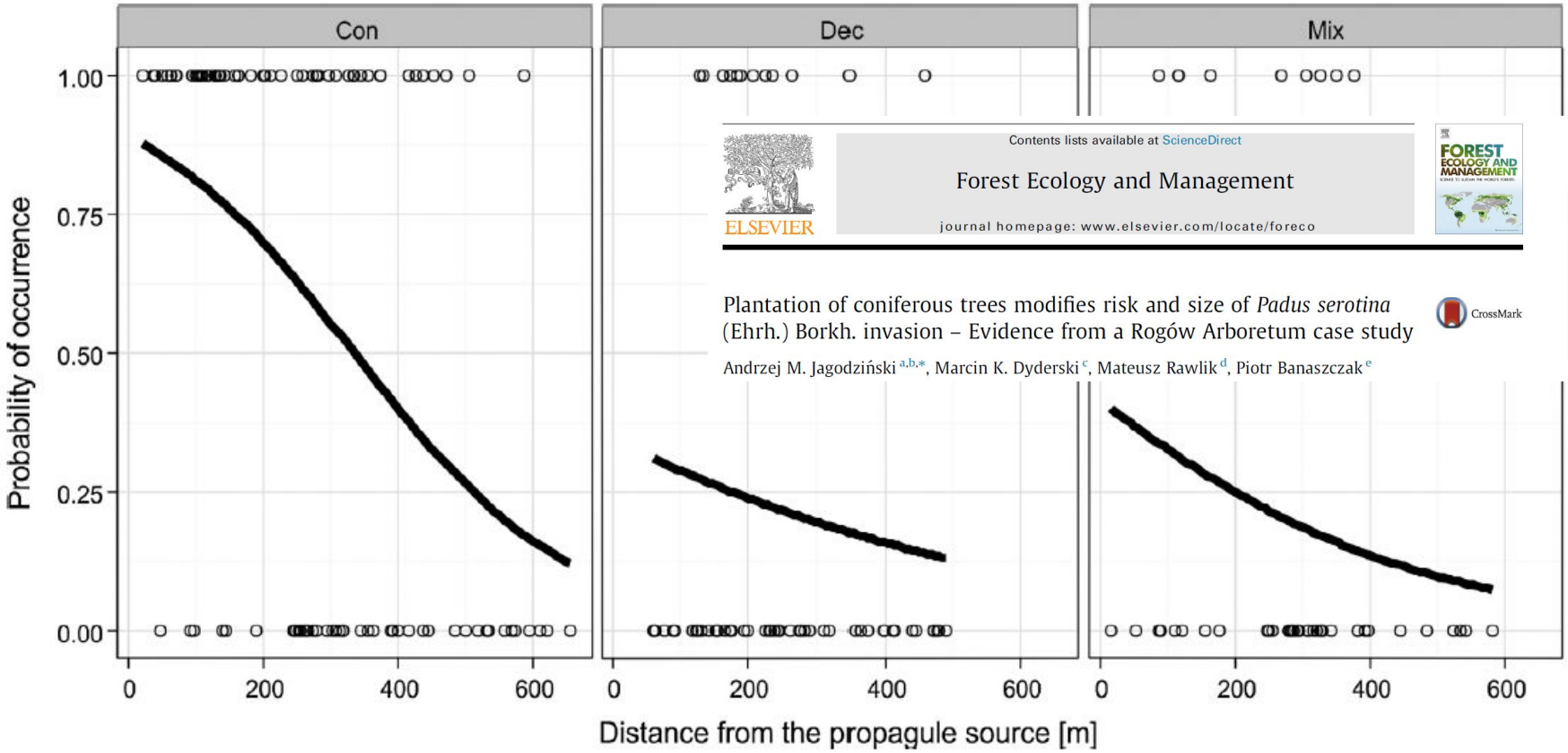
zmienna liczbowa - np. odległość od źródła nasion

zmienna kategoryczna - np. typ drzewostanu

```
model0<-glm(prunus~1,dane)

model1<-glm(prunus~odl,dane)

model2<-glm(prunus~typ,dane)

model3<-glm(prunus~odl+typ,dane) #ancova

model4<-glm(prunus~odl*typ,dane) #ancova
```

znaki w formule: + addytywność (wspólne oddziaływanie)

: interakcja *addytywność i interackja



Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco

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

Andrzej M. Jagodziński [a,b,*], Marcin K. Dyderski [c], Mateusz Rawlik [d], Piotr Banaszczak [e]

Fig. 4. Effect of coniferous (Con), deciduous (Dec) and mixed (Mix) tree stands on the probability of colonisation by black cherry ($p < 0.001$).

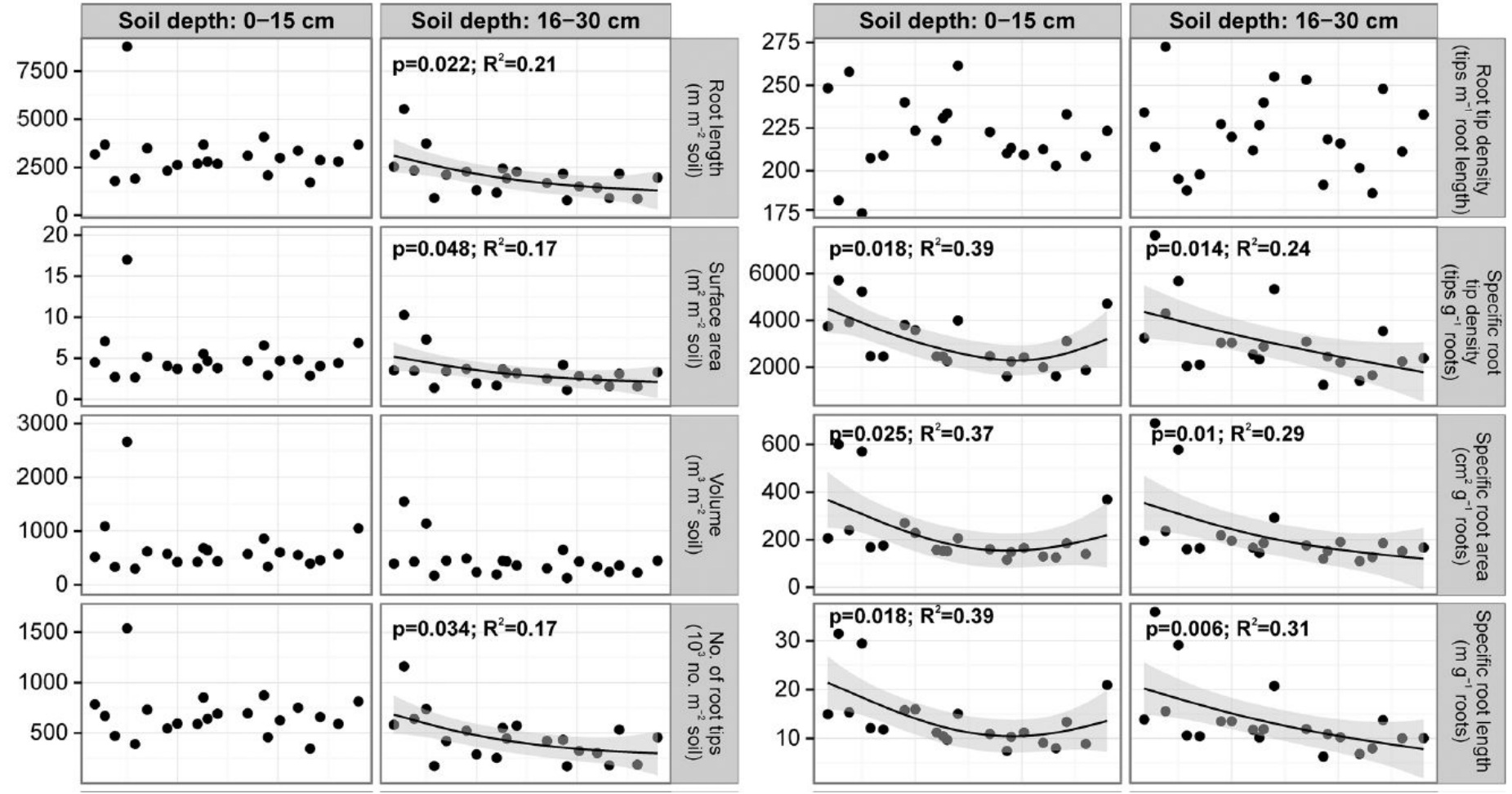
GAM - addytywność

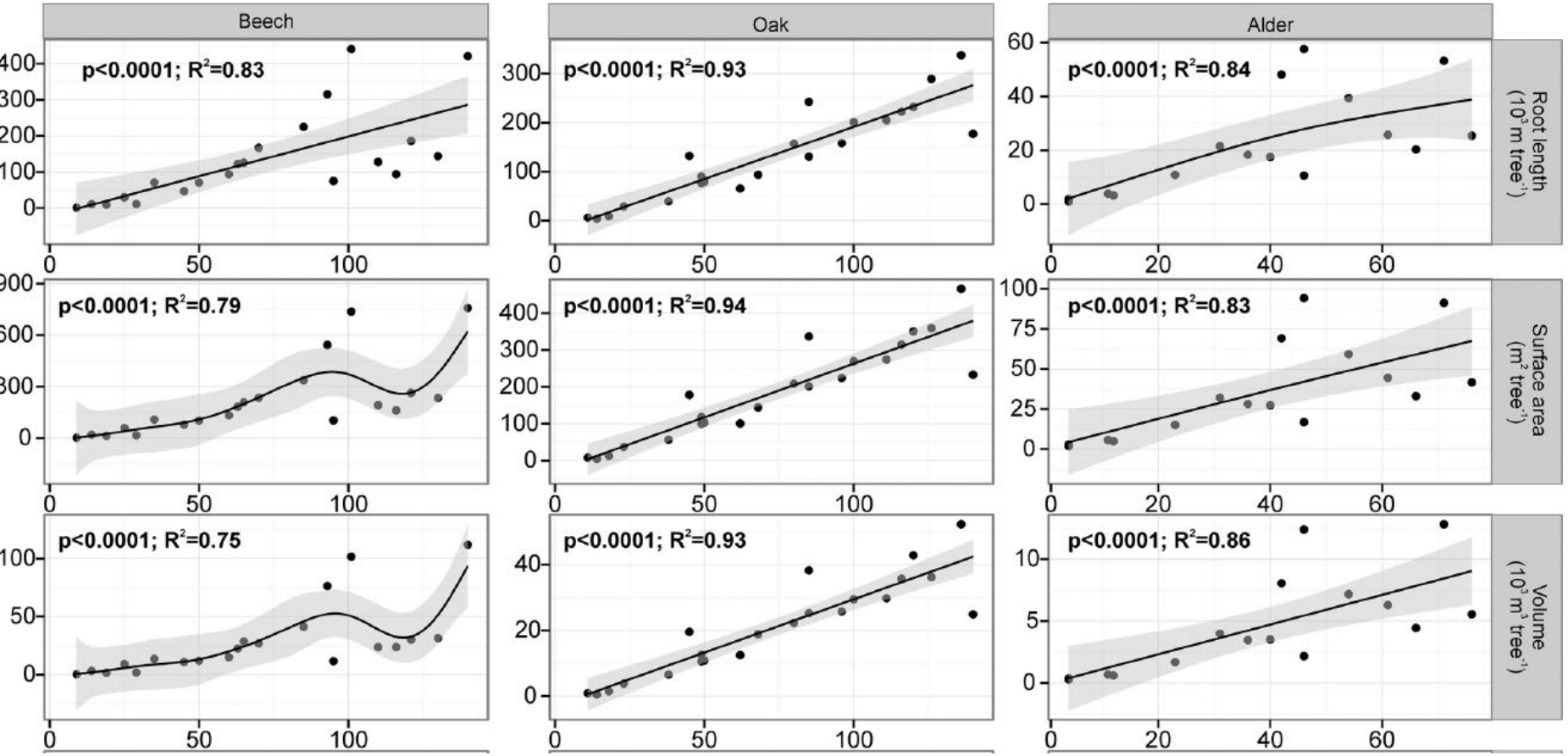
PLOS ONE

RESEARCH ARTICLE

Tree Age Effects on Fine Root Biomass and Morphology over Chronosequences of *Fagus sylvatica*, *Quercus robur* and *Alnus glutinosa* Stands

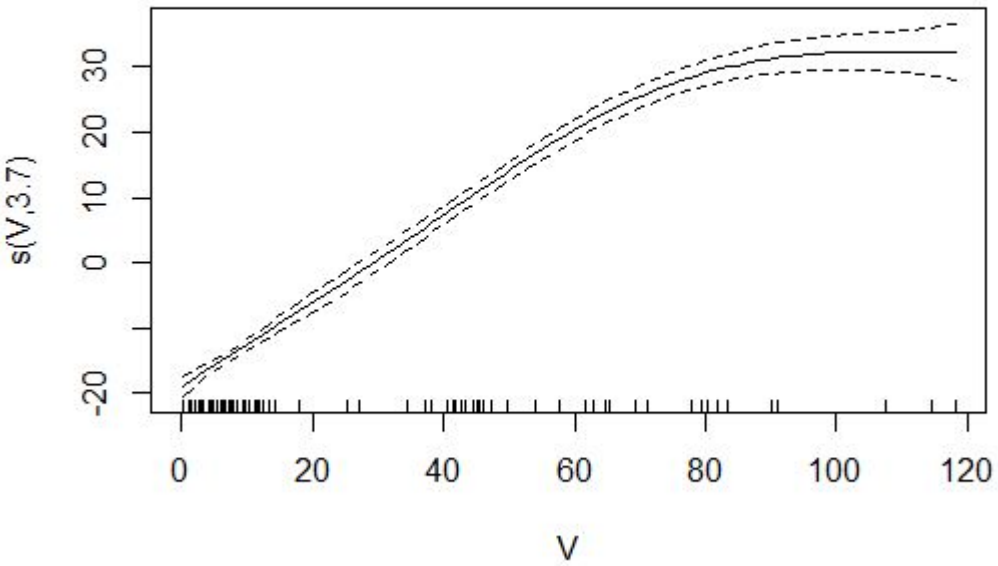
Andrzej M. Jagodzinski[1,2]*, Jędrzej Ziółkowski[2], Aleksandra Warnkowska[2], Hubert Prais[2]


Soil depth: 0–15 cm
Soil depth: 16–30 cm
Soil depth: 0–15 cm
Soil depth: 16–30 cm
Root length (m m⁻² soil)
p=0.022; R²=0.21
Root tip density (tips m⁻¹ root length)
Surface area (m² m⁻² soil)
p=0.048; R²=0.17
p=0.018; R²=0.39
p=0.014; R²=0.24
Specific root tip density (tips g⁻¹ roots)
Volume (m³ m⁻² soil)
p=0.025; R²=0.37
p=0.01; R²=0.29
Specific root area (cm² g⁻¹ roots)
No. of root tips (10³ no. m⁻² soil)
p=0.034; R²=0.17
p=0.018; R²=0.39
p=0.006; R²=0.31
Specific root length (m g⁻¹ roots)

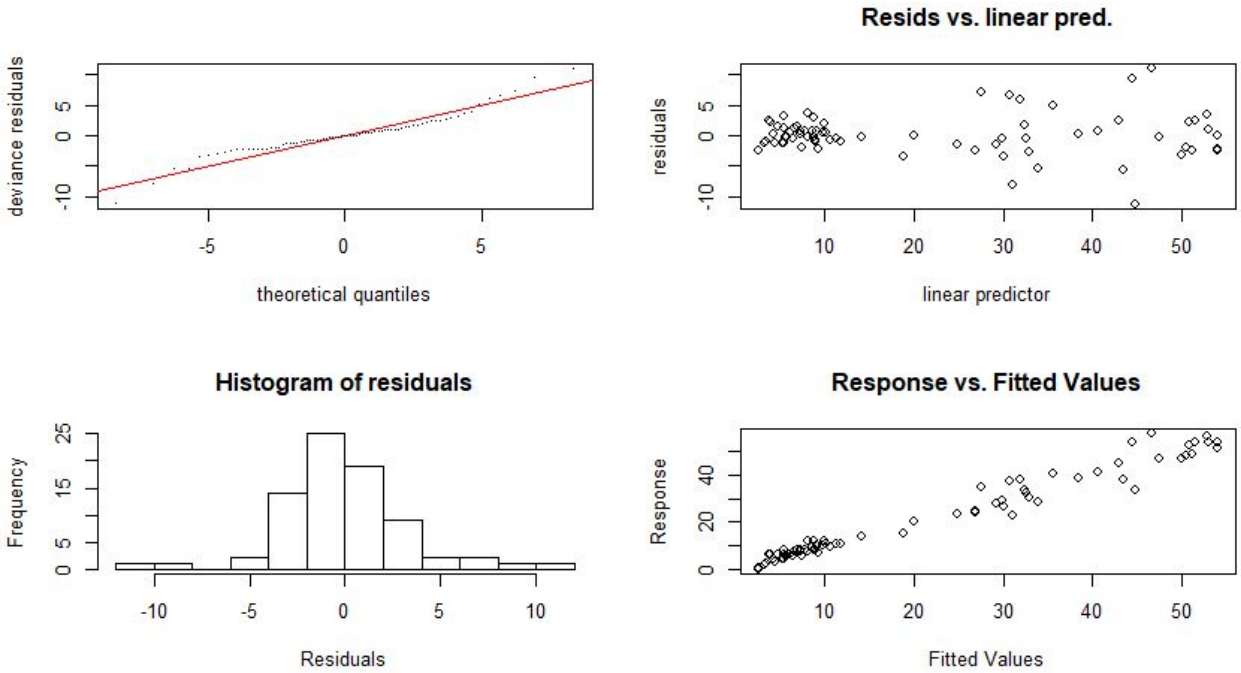

Beech
Oak
Alder
p<0.0001; R²=0.83
p<0.0001; R²=0.93
p<0.0001; R²=0.84
p<0.0001; R²=0.79
p<0.0001; R²=0.94
p<0.0001; R²=0.83
p<0.0001; R²=0.75
p<0.0001; R²=0.93
p<0.0001; R²=0.86
Root length (10³ m tree⁻¹)
Surface area (m² tree⁻¹)
Volume (10³ m³ tree⁻¹)

```
> library(mgcv)
> model1<-gam(AB~s(V),data=sosny)
> summary(model1)

Family: gaussian
Link function: identity

Formula:
AB ~ s(V)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.5752     0.3845   56.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df     F p-value
s(V) 3.701  4.546 460.4  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.965   Deviance explained = 96.7%
GCV = 12.125  Scale est. = 11.385    n = 77
>
```

plot(model1)



```
par(mfrow=c(2,2))
gam.check(model1)
```



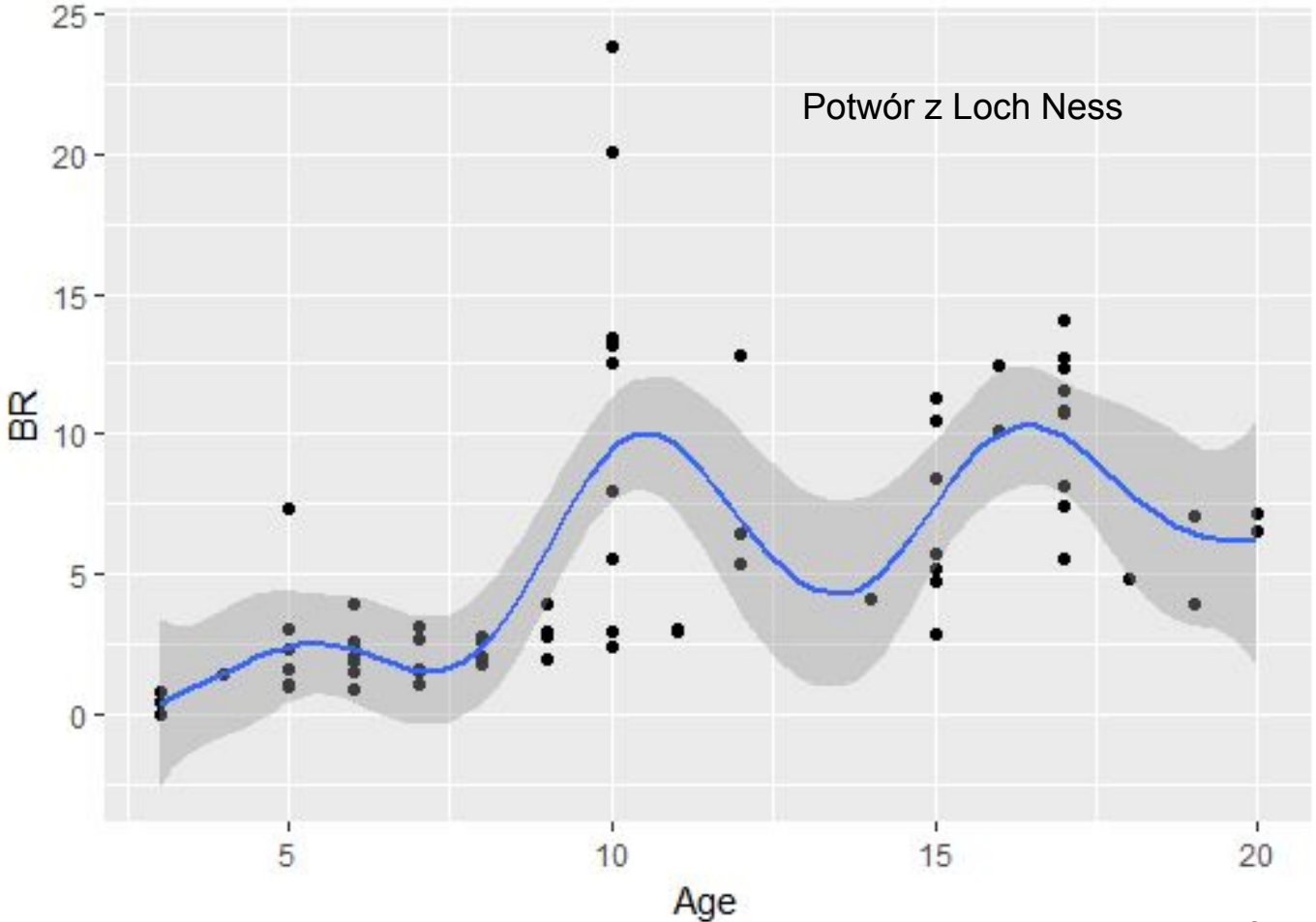
```
> AIC(model1)
[1] 412.3526
> AIC(model0)
[1] 666.9126
```

poprzednie modele

> 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
> AIC(nls(AB~a*V^b,data=sosny,start=list(a=1,b=-2))) #model potęgowy
[1] 441.9817

>AIC(gam(AB~s(V),data=sosny)) #GAM
[1] 412.3526

```
> plot(modelbr)
> modelbr<-gam(BR~s(Age),data=sosny)
> summary(modelbr)

Family: gaussian
Link function: identity

Formula:
BR ~ s(Age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.4899     0.3988   13.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
        edf Ref.df     F  p-value
s(Age) 7.751  8.566 9.005 3.45e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.486   Deviance explained = 53.9%
GCV = 13.814  Scale est. = 12.244    n = 77
> |
```




```
ggplot(sosny, aes(y=BR,x=Age))+geom_point()+geom_smooth(method='gam',formula=y~s(x))
```

Czy jest to biologicznie uzasadnione?

Co się dzieje w wieku 10 lat?

Czy coś takiego przejdzie?

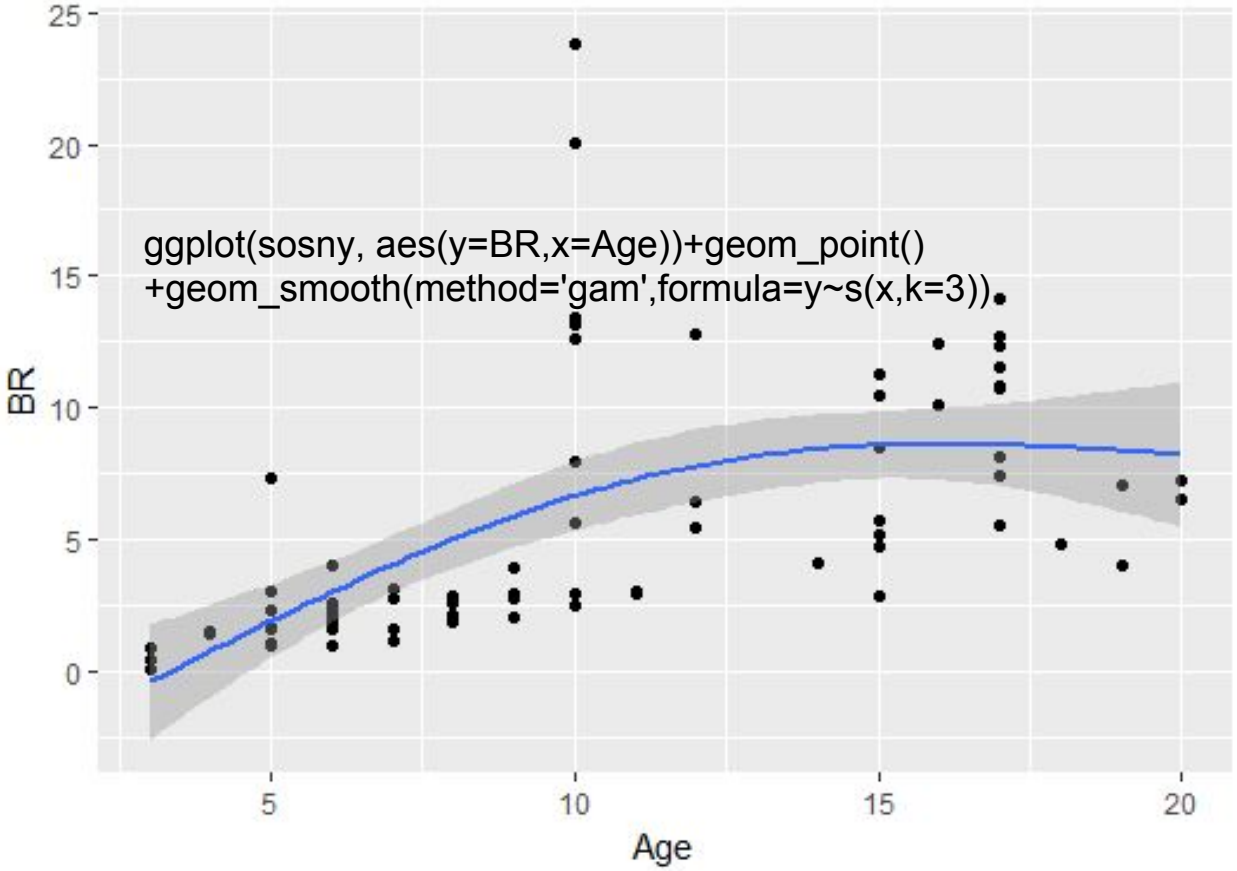
- na obronie doktoratu - zależy od audytorium
- w dobrym czasopiśmie - nie bardzo

im większy stopień wielomianu/głębokość splinu tym lepsze dopasowanie

na czym nam zależy?

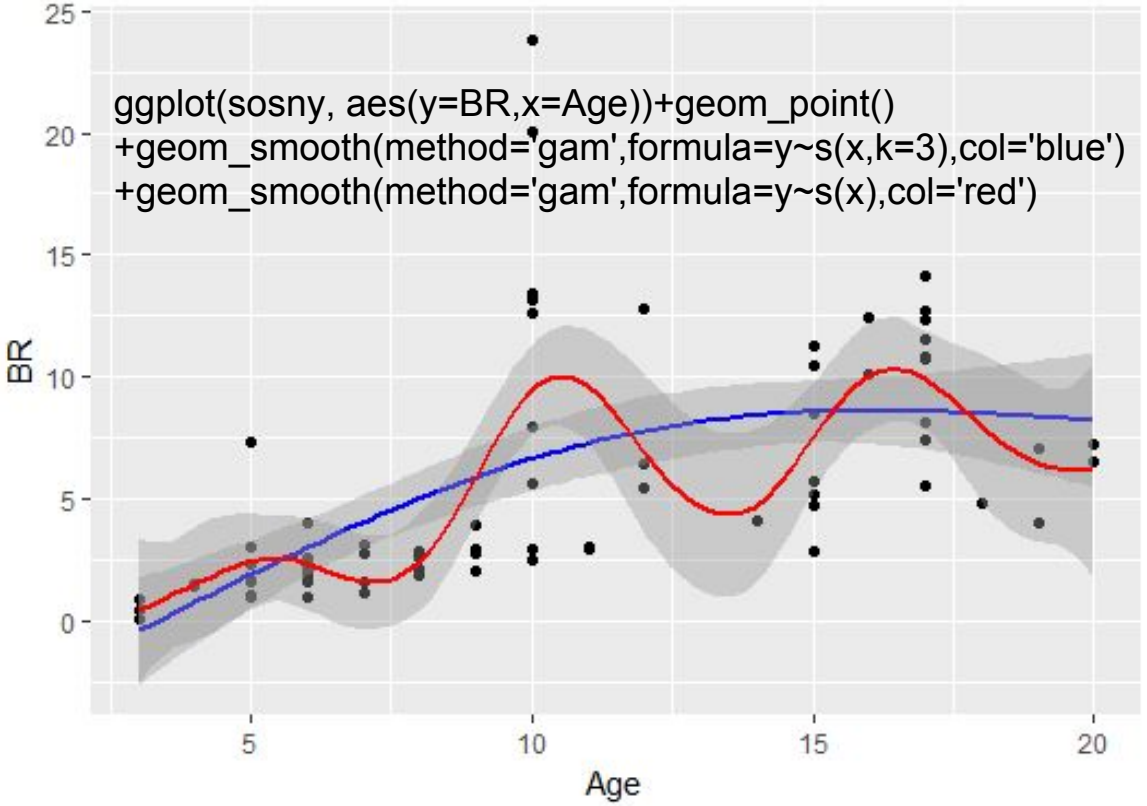
overfitting - model dobrze działa na zbiorze treningowym

rozwiązanie - ograniczyć spline



```
',formula=y~s(x,k=3))
> modelbr<-gam(BR~s(Age,k=3),data=sosny)
> summary(modelbr)

Family: gaussian
Link function: identity

Formula:
BR ~ s(Age, k = 3)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.4899     0.4488   12.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
        edf Ref.df     F p-value
s(Age) 1.87  1.983 22.44 6.6e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.349   Deviance explained = 36.5%
GCV = 16.112  Scale est. = 15.512    n = 77
> 
```

Tracimy trochę R2 ale model jest bardziej sensowny



modele mieszane

kiedy? jeśli oprócz efektów stałych występują efekty losowe

-układ doświadczenia

-zmienność osobnicza

…

-wola recenzentów;)

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

https://link.springer.com/article/10.1007/s10530-018-1706-3

Przykład

udział liści w biomasie młodego pokolenia

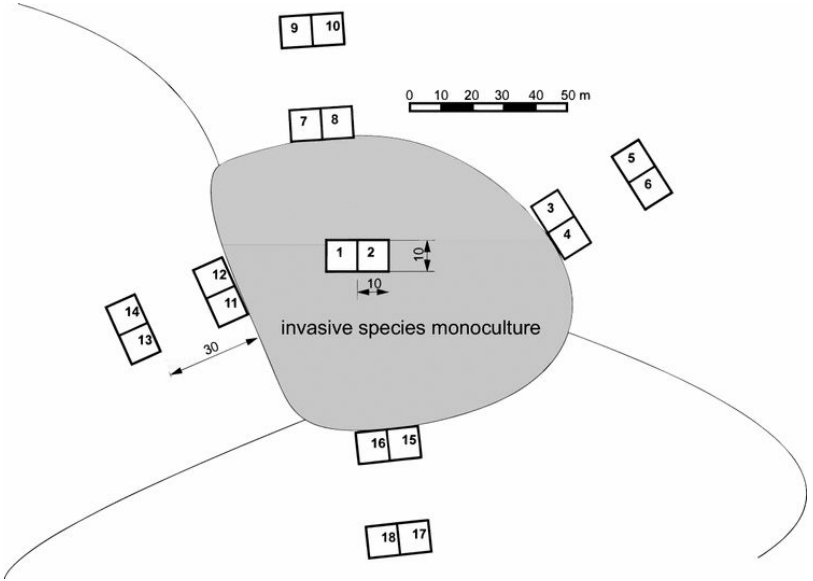
układ powierzchni badawczych - hierarchia

mixed?

wg Roberts et al. 2017 - może być konieczny

https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881

klastrowania brak - sprawdzone inną metodą



A może jednak?

4 gatunki, 2 klasy wieku, DIFN - światło + układ poletek (blok, plot)

```
> zwykly<-lm(lmf~gat+wiek+DIFN,data=mix)
> summary(zwykly)

Call:
lm(formula = lmf ~ gat + wiek + DIFN, data = mix)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40724 -0.06906 -0.00117  0.06392  0.31410

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.304005   0.010155  29.937  < 2e-16 ***
gatApse      -0.013414   0.011163  -1.202     0.23
gatPser       0.076652   0.011704   6.549 9.21e-11 ***
gatRpse       0.097873   0.013457   7.273 7.05e-13 ***
wiekseedling  0.134403   0.007248  18.545  < 2e-16 ***
DIFN         -0.800482   0.102510  -7.809 1.44e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1046 on 1007 degrees of freedom
Multiple R-squared:  0.3566,    Adjusted R-squared:  0.3534
F-statistic: 111.6 on 5 and 1007 DF,  p-value: < 2.2e-16

>
```

```
> mieszany<-lmer(lmf~gat+wiek+DIFN+(pow|blok),data=mix)
> summary(mieszany)
Linear mixed model fit by REML t-tests use Satterthwaite
  approximations to degrees of freedom [lmerMod]
Formula: lmf ~ gat + wiek + DIFN + (pow | blok)
   Data: mix

REML criterion at convergence: -1771

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9339 -0.6478  0.0255  0.5964  2.9677

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 blok     (Intercept) 4.423e-03 0.06650
          pow         1.318e-05 0.00363  -0.74
 Residual             9.205e-03 0.09594
Number of obs: 1013, groups:  blok, 21

Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)  0.314516   0.014885  59.100000  21.130  < 2e-16 ***
gatApse     -0.007777   0.011325 983.100000  -0.687    0.492
gatPser      0.082059   0.012502 972.800000   6.564 8.51e-11 ***
gatRpse      0.111816   0.013036 988.000000   8.577  < 2e-16 ***
```

```
Fixed effects:
             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)  0.314516   0.014885  59.100000  21.130  < 2e-16 ***
gatApse     -0.007777   0.011325 983.100000  -0.687    0.492
gatPser      0.082059   0.012502 972.800000   6.564 8.51e-11 ***
gatRpse      0.111816   0.013036 988.000000   8.577  < 2e-16 ***
wiekseedling 0.120891   0.006968 990.900000  17.350  < 2e-16 ***
DIFN        -0.762006   0.112451 713.700000  -6.776 2.58e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) gatAps gatPsr gatRps wksdln
gatApse     -0.522
gatPser     -0.504  0.777
gatRpse     -0.384  0.576  0.520
wiekseedlng -0.191 -0.220 -0.090 -0.183
DIFN        -0.274 -0.030 -0.117  0.007  0.003
>
```

formuła modeli mieszanych

fixed effects - podajemy normalnie

random effects - np. (1|blok)

random effects zagnieżdżone (pow|blok) - poletko w ramach bloku

jakie są wartości SE efektów losowych?

który model lepszy?

> AIC(zwykly,mieszany)

	df	AIC
zwykly	7	-1690.901
mieszany	**10**	**-1750.960**

Ile procent wyjaśnia?

```
> library(MuMIn) #polskie dzieło;)
> r.squaredGLMM(mieszany)
      R2m       R2c
0.3186558 0.4684045
```

R2m - marginal R2; R2c - conditional R2

R2m - % zmienności wyjaśnionej przez fixed effects

R2c - % zmienności wyjaśnionej przez fixed + random effects

random effects - R2c-R2m ~ 33%

Z czym związane poletka?

dominant w drzewostanie, żyzność…

efekty niebadane, mające wpływ na wynik

na ile są uwzględnione, na ile są ważne?

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

Outliery

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

PRIMARY RESEARCH ARTICLE

WILEY Global Change Biology

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

Marcin K. Dyderski[1,2] | Sonia Paź[3] | Lee E. Frelich[4] | Andrzej M. Jagodziński[1,2]

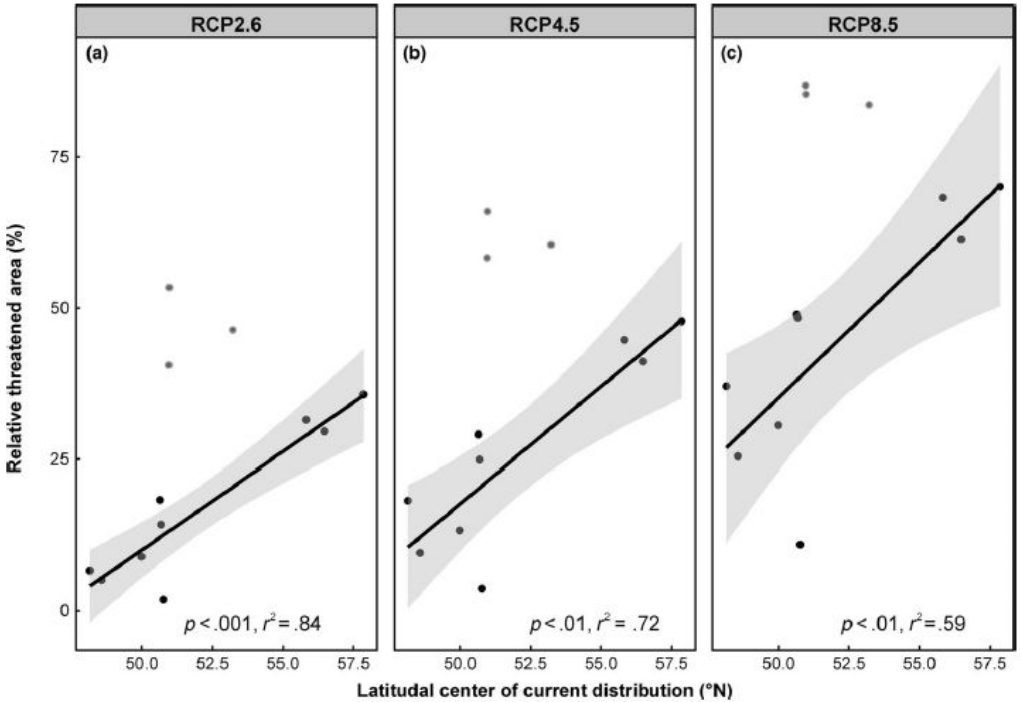


FIGURE 7 The relationship between threatened proportion of current distribution and latitudinal center of distribution for each climate

Wysublimowanie czy prostota?

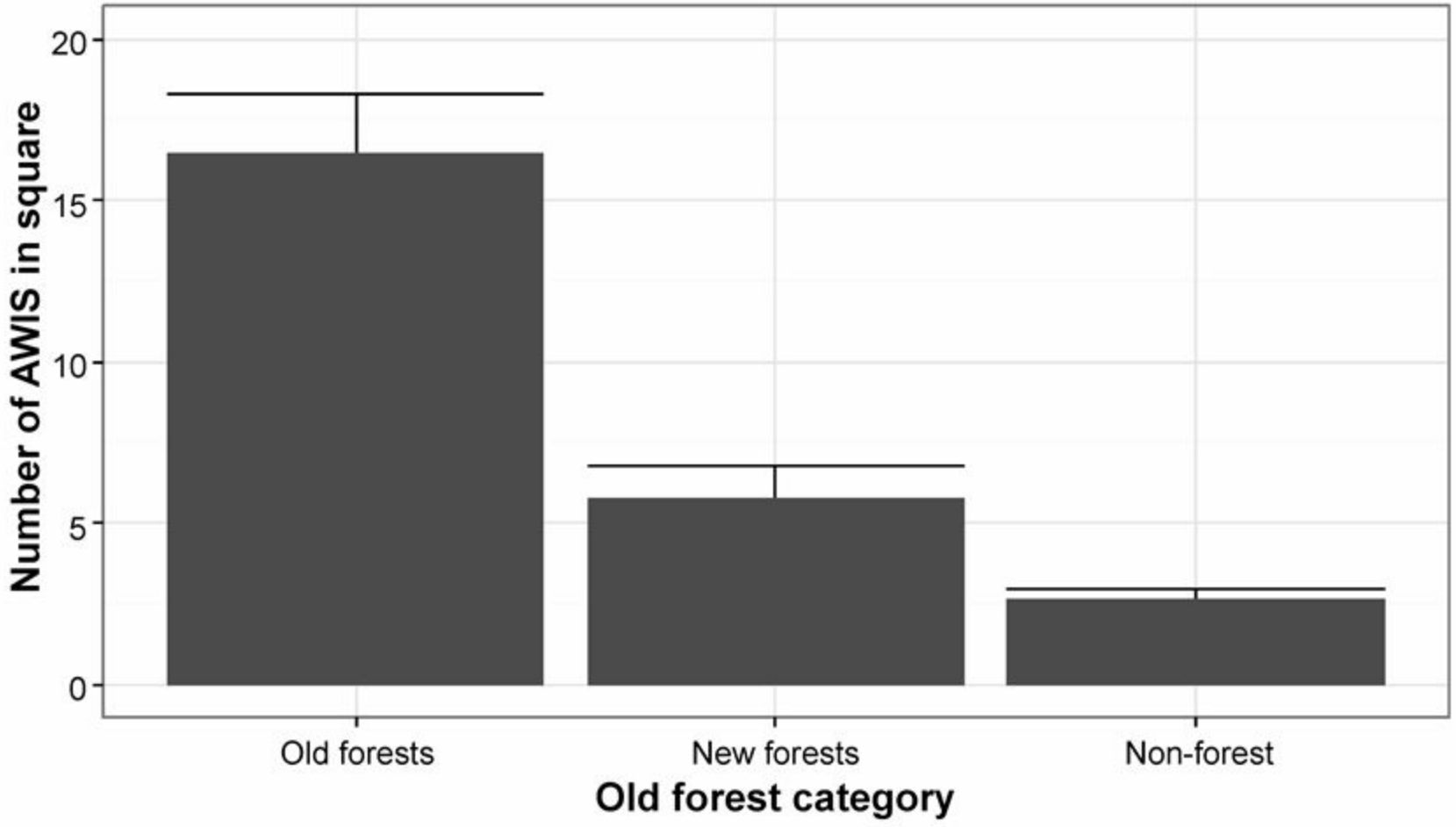
Cel

Użytkownik

Jakość danych

Widzimisię recenzentów

ANOVA vs. Spatially-explicit Poisson GLM


Number of AWIS in square
20
15
10
5
0
Old forests
New forests
Non-forest
Old forest category

Podsumowanie

obrazki, obrazki, obrazki!

-pomogą dobrać narzędzie i typ rozkładu

biologiczne znaczenie (effect size)> rozkład błędów > AIC > p > R2

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

correlation does not imply causation!

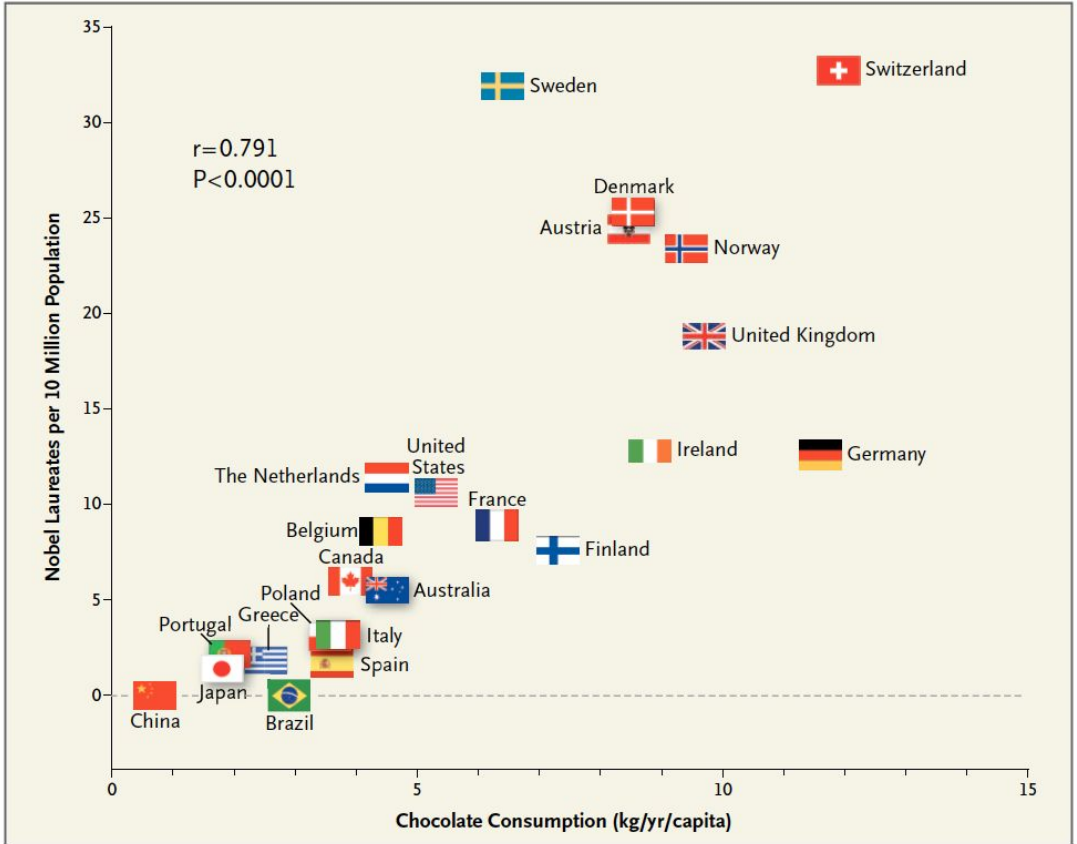


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

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


BSS
BIAŁOWIESKA SZKOŁA STATYSTYKI