
BSS
BIAŁOWIESKA SZKOŁA STATYSTYKI

Wprowadzenie do regresji liniowej

Współczynniki korelacji

r Pearsona – najczęściej używany, parametryczny (zakłada rozkład normalny)

rho Spearmana – nieparametryczny, korelacja rang

tau Kendalla

R2 a *r:*

R2 to współczynnik determinacji – procent wyjaśnionej zmienności

```
soli.bss<-read.csv("soli.bss.csv", sep=";", dec=",")


cor(soli.bss$orto.shan, soli.bss$canopy.height)

[1] -0.638738
```

Co to jest? – miara współzależności

```
cor(soli.bss$orto.shan, soli.bss$canopy.height, method="pearson")
[1] -0.638738

cor(soli.bss$orto.shan, soli.bss$canopy.height, method="spearman")
[1] -0.5275458
```

Macierz korelacji

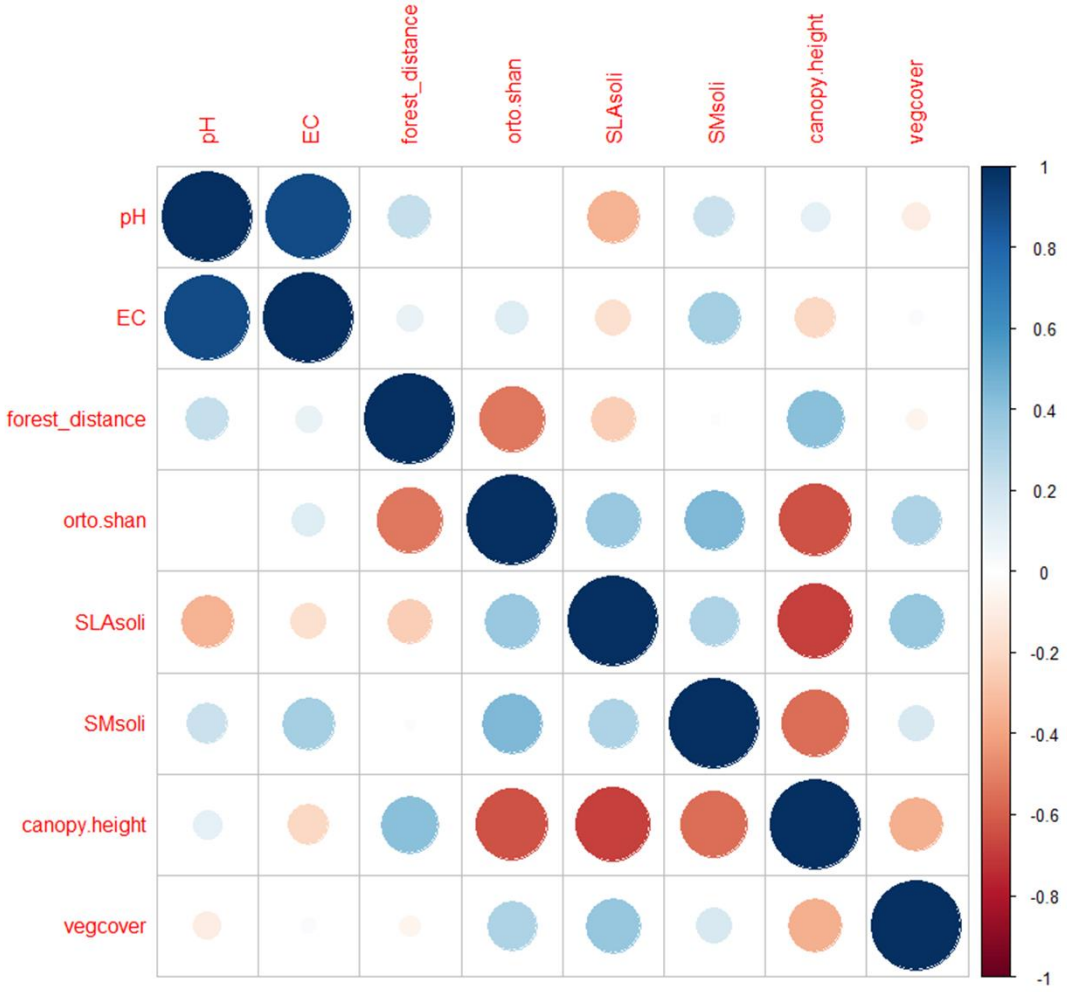
```
> cor(soli.bss[c(5,6,14,48,54:57)])
                          pH          EC forest_distance    orto.shan    SLAsoli      SMsoli
pH                1.000000000  0.89095618      0.23631083 -0.003861412 -0.3435727  0.21294876
EC                0.890956180  1.00000000      0.09205453  0.131965696 -0.1616592  0.33589774
forest_distance   0.236310833  0.09205453      1.00000000 -0.522595223 -0.2456127  0.01781984
orto.shan        -0.003861412  0.13196570     -0.52259522  1.000000000  0.3720923  0.44185726
SLAsoli          -0.343572725 -0.16165917     -0.24561272  0.372092313  1.0000000  0.30717664
SMsoli            0.212948762  0.33589774      0.01781984  0.441857256  0.3071766  1.00000000
canopy.height     0.102962439 -0.20373892      0.41212517 -0.638737964 -0.6840524 -0.55329898
vegcover         -0.098958732  0.02740244     -0.05549340  0.307354329  0.3802813  0.16198705
                 canopy.height    vegcover
pH                   0.1029624 -0.09895873
EC                  -0.2037389  0.02740244
forest_distance      0.4121252 -0.05549340
orto.shan           -0.6387380  0.30735433
SLAsoli             -0.6840524  0.38028134
SMsoli              -0.5532990  0.16198705
canopy.height        1.0000000 -0.35173972
vegcover            -0.3517397  1.00000000
>
```

Funkcja cor dla więcej niż dwóch zmiennych:

cor(soli.bss[c(5,6,14,48,54:57)])

Wizualizacja macierzy korelacji

```
library(corrplot)
corrplot(cor(soli.bss[,c(5,6,14,48,54:57)]))
```



Składnia:

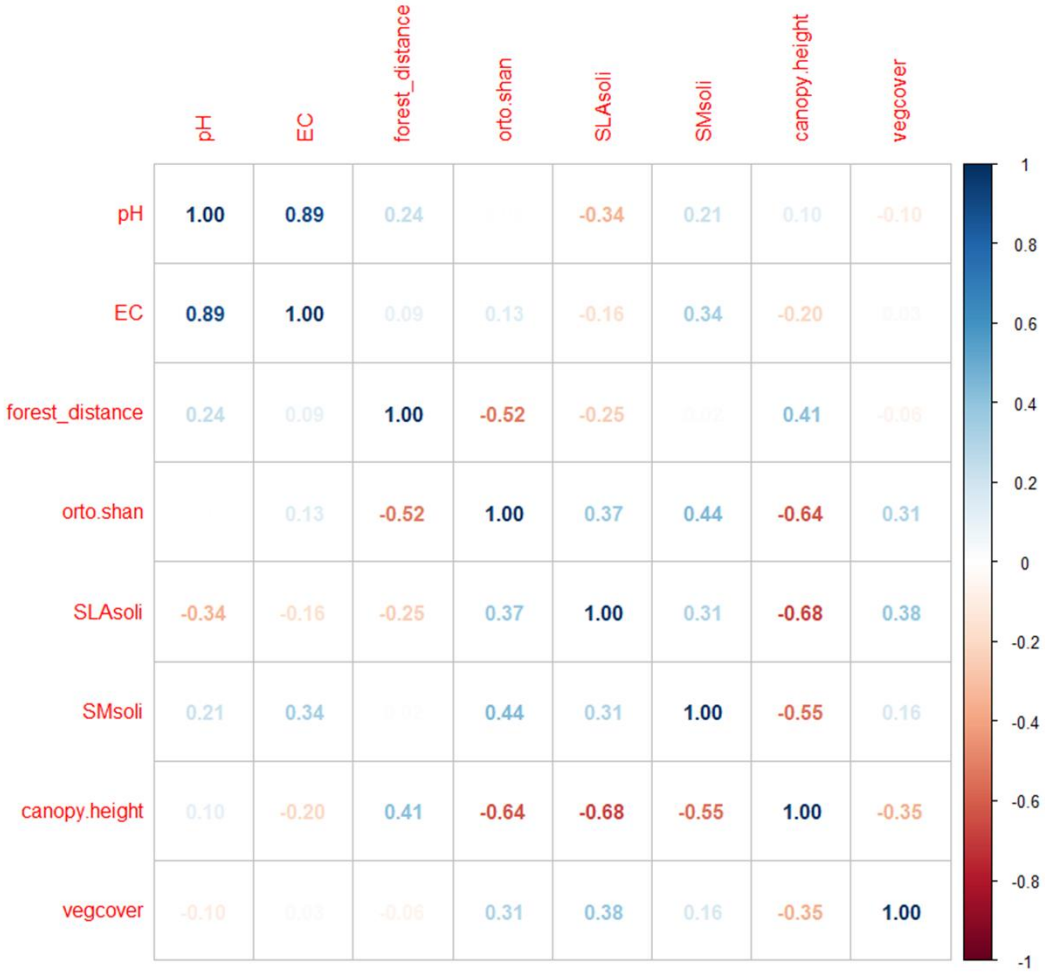
cor() dla więcej niż 2 zmiennych zwraca macierz korelacji

Wrzucamy wynik cor() w funkcję corrplot()

Jest wiele opcji wizualizacji danych za pomocą tego pakietu, np.:

corrplot(cor(soli.bss[,c(5,6,14,48,54:57)]), method="num")

#https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html



Korelacja a regresja

Korelacja – miara współzależności
Regresja – opis zależności

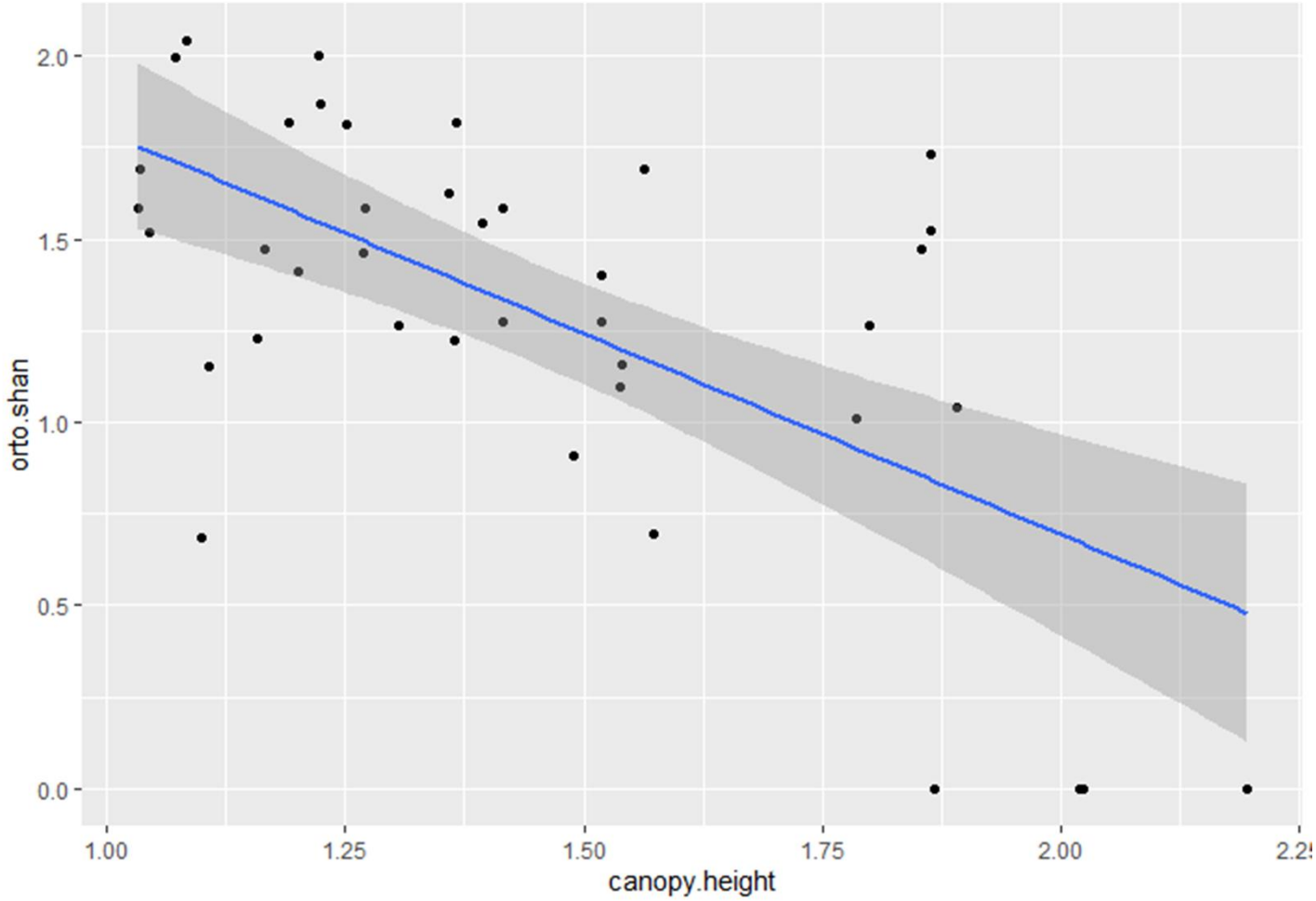
Kompromis pomiędzy dwoma cechami
Przewidywanie (modelowanie) zmiennej zależnej
Wyjaśnianie procesów (procent wyjaśnionej zmienności – R2)

Regresja liniowa:

Jak zmienia się różnorodność taksonomiczna prostoskrzydłych wraz ze zmieniającą się wysokością roślin?

Problem badawczy – poznanie wpływu roślinności na różnorodność biologiczną owadów
Problem aplikacyjny – możliwości estymacji

```
soli.bss<-read.csv("soli.bss.csv", sep=";", dec=",")

ggplot(soli.bss, aes(x=canopy.height, y=orto.shan)) + geom_point() +
geom_smooth(method="lm")
```

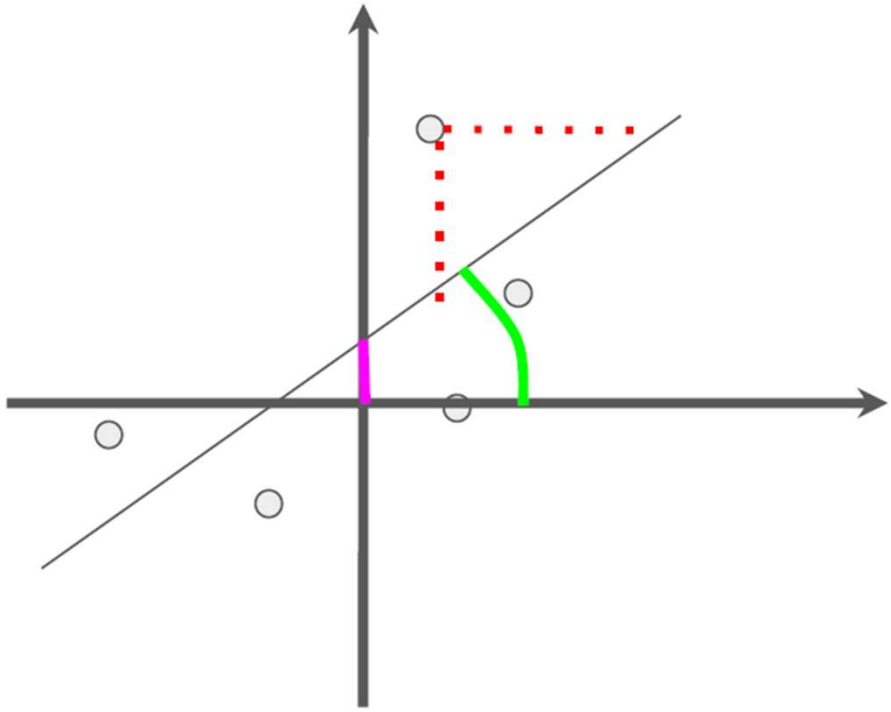


Model liniowy

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 lini



Zapis matematyczny:

Y=ax+b

Zapis w R:

Y~Z ~ -tylda (pod Esc)

Y - zmienna zależna, odpowiedź (response). zmienna modelowana, coś co chcemy wymodelować

X - zmienna niezależna predyktor coś, co ma nam wyjaśniać Y
*ale nie parametr (parametr to a)

Założenie: rozkład normalny zmiennej zależnej
(lub zbliżony do normalnego)

mod1<-lm(orto.shan~canopy.height, data=soli.bss)
summary(mod1)

```
> mod1<-lm(orto.shan~canopy.height, data=soli.bss)
> summary(mod1)

Call:
lm(formula = orto.shan ~ canopy.height, data = soli.bss)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99640 -0.19926  0.01136  0.28560  0.88558

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.8862     0.3176   9.089 4.54e-11 ***
canopy.height -1.0964     0.2142  -5.117 9.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4261 on 38 degrees of freedom
Multiple R-squared:  0.408,     Adjusted R-squared:  0.3924
F-statistic: 26.19 on 1 and 38 DF,  p-value: 9.196e-06
```

Model liniowy z predyktorem kategorycznym

Pytanie badawcze:

Czy inwazja nawłoci kanadyjskiej (wyrażona za pomocą klas % pokrycia poletek tym gatunkiem obcym) wpływa na różnorodność taksonomiczną (wskaźnik Shannona) prostoskrzydłych?

ANOVA :

Analiza wariancji - sprawdza czy są różnice między którąkolwiek z par poziomów zmiennej grupującej

summary(aov(orto.shan~Sol_class, data=soli.bss))

```
> summary(aov(orto.shan~Sol_class, data=soli.bss))
            Df Sum Sq Mean Sq F value  Pr(>F)
Sol_class    3  3.775  1.2582   5.749 0.00255 **
Residuals   36  7.879  0.2189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Są różnice* -tylko pomiędzy którym z wariantów?

Wiadomo tylko, że jest różnica

Aby sprawdzić co się różni -test *post hoc* (test *posteriori*)

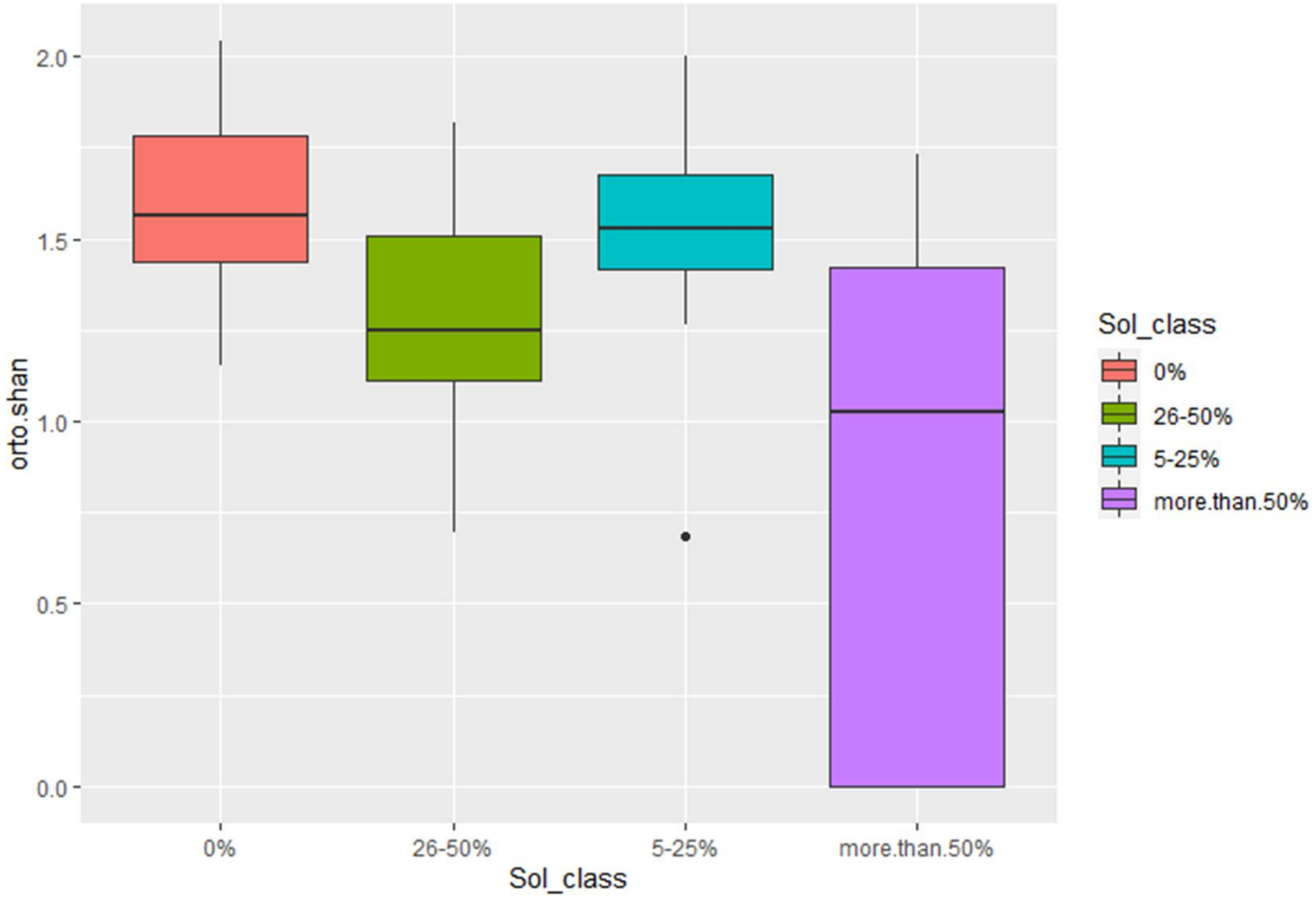
Test *post hoc* wykonujemy tylko jeśli wyszły istotne różnice

```
library(agricolae)
an1<-aov(orto.shan~Sol_class,
data=soli.bss)
HSD.test(an1, 'Sol_class',console = T)
```

```
Study: an1 ~ "Sol_class"

HSD Test for orto.shan

Mean Square Error:  0.2188612

Sol_class,  means

              orto.shan       std  r       Min      Max
0%            1.5980023 0.2958730 10 1.1537419 2.041517
26-50%        1.2846909 0.3649729 10 0.6931472 1.820076
5-25%         1.5057679 0.3617439 10 0.6849548 2.002049
more.than.50% 0.8039236 0.7237681 10 0.0000000 1.729346

Alpha: 0.05 ; DF Error: 36
Critical Value of Studentized Range: 3.808798

Minimun Significant Difference: 0.5634721

Treatments with the same letter are not significantly different.

              orto.shan groups
0%            1.5980023      a
5-25%         1.5057679      a
26-50%        1.2846909     ab
more.than.50% 0.8039236      b
```

Pokażmy to na boxplocie:

ggplot(soli.bss, aes(x=Sol_class, y=orto.shan, fill=Sol_class))+geom_boxplot()



Model liniowy

Model liniowy - daje nam informacje o wpływie współczynników i postać modelu
Poziom referencyjny (pierwszy poziom zmiennej grupującej) - różnice w stosunku do niego

```
model<-lm(orto.shan~Sol_class, data=soli.bss)
summary(model)
```

Czy różnorodność prostoskrzydłych różni się pomiędzy klasami pokrycia poletek przez nawłoć?

Są gwiazdki, ale jak to zinterpretować?

orto.shan=1.598 jeśli Sol_class0% (poziom referencyjny)

```
> summary(mod1)

Call:
lm(formula = orto.shan ~ Sol_class, data = soli.bss)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82081 -0.20053 -0.01164  0.31489  0.92542

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.59800    0.14794  10.802 7.61e-13 ***
Sol_class26-50%      -0.31331    0.20922  -1.498 0.142972
Sol_class5-25%       -0.09223    0.20922  -0.441 0.661956
Sol_classmore.than.50% -0.79408    0.20922  -3.795 0.000545 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4678 on 36 degrees of freedom
Multiple R-squared:  0.3239,    Adjusted R-squared:  0.2676
F-statistic: 5.749 on 3 and 36 DF,  p-value: 0.00255
```

Dla Sol_class5-25%: orto.shan średnio mniejsze o 0.09 w porównaniu do referencji
Dla Sol_class26-50%: orto.shan średnio mniejsze o 0.31 w porównaniu do referencji
Dla Sol_classmore.than.50%: orto.shan średnio mniejsze o 0.79 w porównaniu do referencji

Model liniowy z więcej niż jedna zmienną

Model z więcej niż jedną zmienną trudniej pokazać

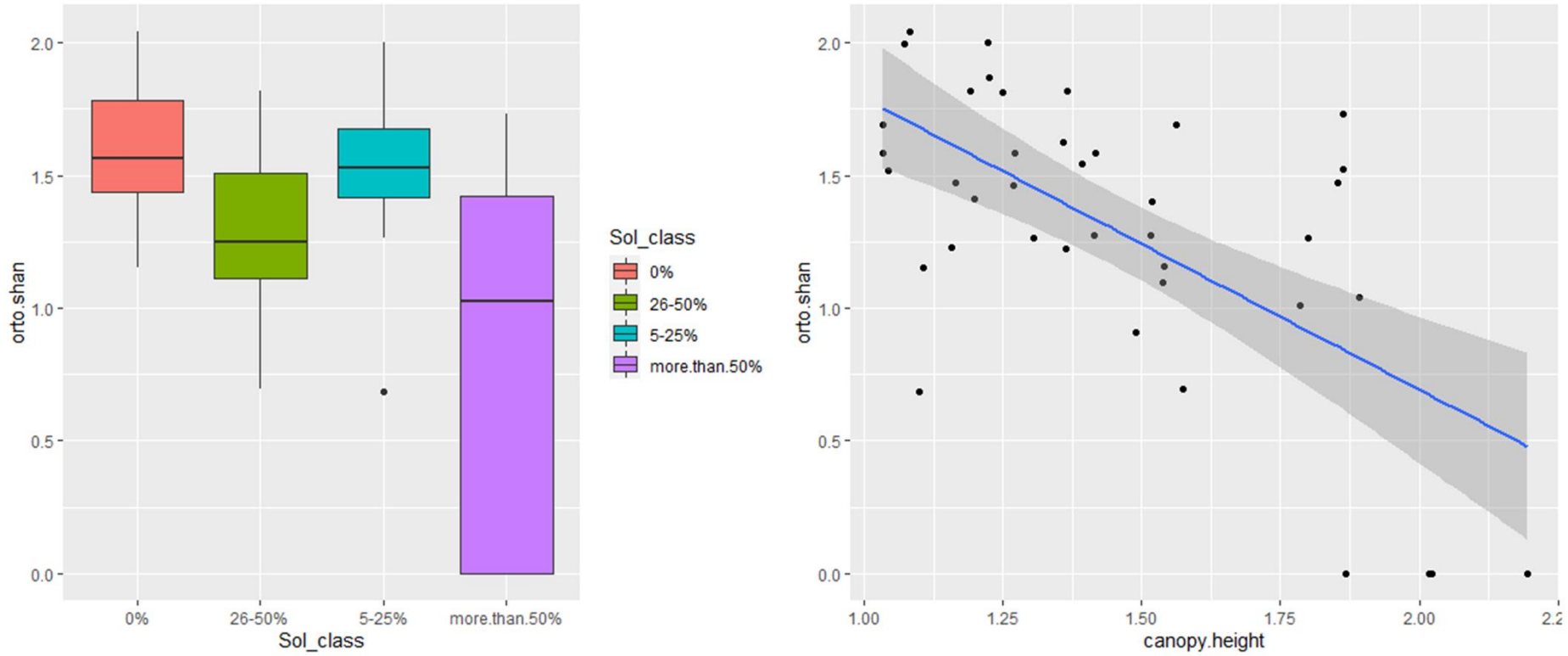
Rozważmy sytuację, gdy na różnorodność prostoskrzydłych oprócz % pokrycia nawłoci wpływ może mieć również wysokość roślin

```
mod1<-lm(orto.shan~canopy.height+Sol_class, data=soli.bss)
summary(mod1)
```

```
Call:
lm(formula = orto.shan ~ canopy.height + Sol_class, data = soli.bss)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10516 -0.22617  0.01402  0.29766  0.84987

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.2011     0.6926   4.622    5e-05 ***
canopy.height        -1.4170     0.5997  -2.363   0.0238 *
Sol_class26-50%       0.1340     0.2733   0.490   0.6270
Sol_class5-25%        0.1463     0.2214   0.661   0.5131
Sol_classmore.than.50%  0.3182     0.5103   0.624   0.5370
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4406 on 35 degrees of freedom
Multiple R-squared:  0.4169,    Adjusted R-squared:  0.3503
F-statistic: 6.256 on 4 and 35 DF,  p-value: 0.0006596
```

Pokażmy to na obrazkach:

```
library(gridExtra)
g1<-ggplot(soli.bss, aes(x=Sol_class, y=orto.shan, fill=Sol_class))+geom_boxplot()
g2<-ggplot(soli.bss, aes(x=canopy.height,
y=orto.shan))+geom_point()+geom_smooth(method="lm")
grid.arrange(g1,g2, ncol=2, nrow=1)
```



Każdy obrazek osobno, bez całościowego uwzględnienia modelu

Czy takie coś przejdzie?

Na obronie doktoratu być może
W części czasopism być może
Niemniej, taka wizualizacja modelu pokazuje tylko pojedynczy wpływ jednego predyktora na zmienną objaśnianą, podczas gdy w modelu mamy dwa predyktory, które działają na zmienną objaśnianą jednocześnie

My pokazaliśmy tylko wpływ pojedynczego predyktora (bez uwzględnienia wpływu drugiego, który również jest w modelu – mamy więc pewne przekłamanie/zakrzywienie rzeczywistości

I co teraz zrobić?

Z pomocą przyjdą nam średnie brzegowe/odpowiedzi brzegowe

marginal means, emmeans, marginal response, partial dependence

Średnie brzegowe/odpowiedzi brzegowe

Średnie brzegowe – marginal means – średnie z modelu

Jak zmienia się wpływ inwazji nawłoci kanadyjskiej na prostoskrzydłe przy założeniu, że slope dla wysokości roślin jest na stałym/średnim poziomie?

Czyli, że się nie zmienia

```
library(emmeans)

mod1<-lm(orto.shan~canopy.height+Sol_class,
data=soli.bss)
```

```
Sol_class      emmean    SE df lower.CL upper.CL
0%               1.15 0.236 35    0.670     1.63
26-50%           1.28 0.139 35    1.000     1.57
5-25%            1.29 0.165 35    0.959     1.63
more.than.50%    1.47 0.313 35    0.831     2.10

Confidence level used: 0.95
```

Aby uzyskać literki (tutaj cyferki) z testu Tukeya:

```
library(multcomp)
library(multcompView)

emy.orto1<-cld(emmeans(mod1, ~Sol_class, type="response"))
emy.orto1
```

```
 Sol_class      emmean    SE df lower.CL upper.CL .group
 0%               1.15 0.236 35    0.670     1.63  1
 26-50%           1.28 0.139 35    1.000     1.57  1
 5-25%            1.29 0.165 35    0.959     1.63  1
 more.than.50%    1.47 0.313 35    0.831     2.10  1

Confidence level used: 0.95
P value adjustment: tukey method for comparing a family of 4 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.
>
```

Niby wpływ nieistotny statystycznie, ale być może istotny biologicznie!!!

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

Np.

Nieistotny biologicznie efekt (3%); p<0.00001 przy n=300

Istotny efekt (800%); p>0.05 przy n=3

Jak pokazać to na obrazku?

Tworzymy data.frame z obiektu, w którym siedzi wynik naszego emmeansa

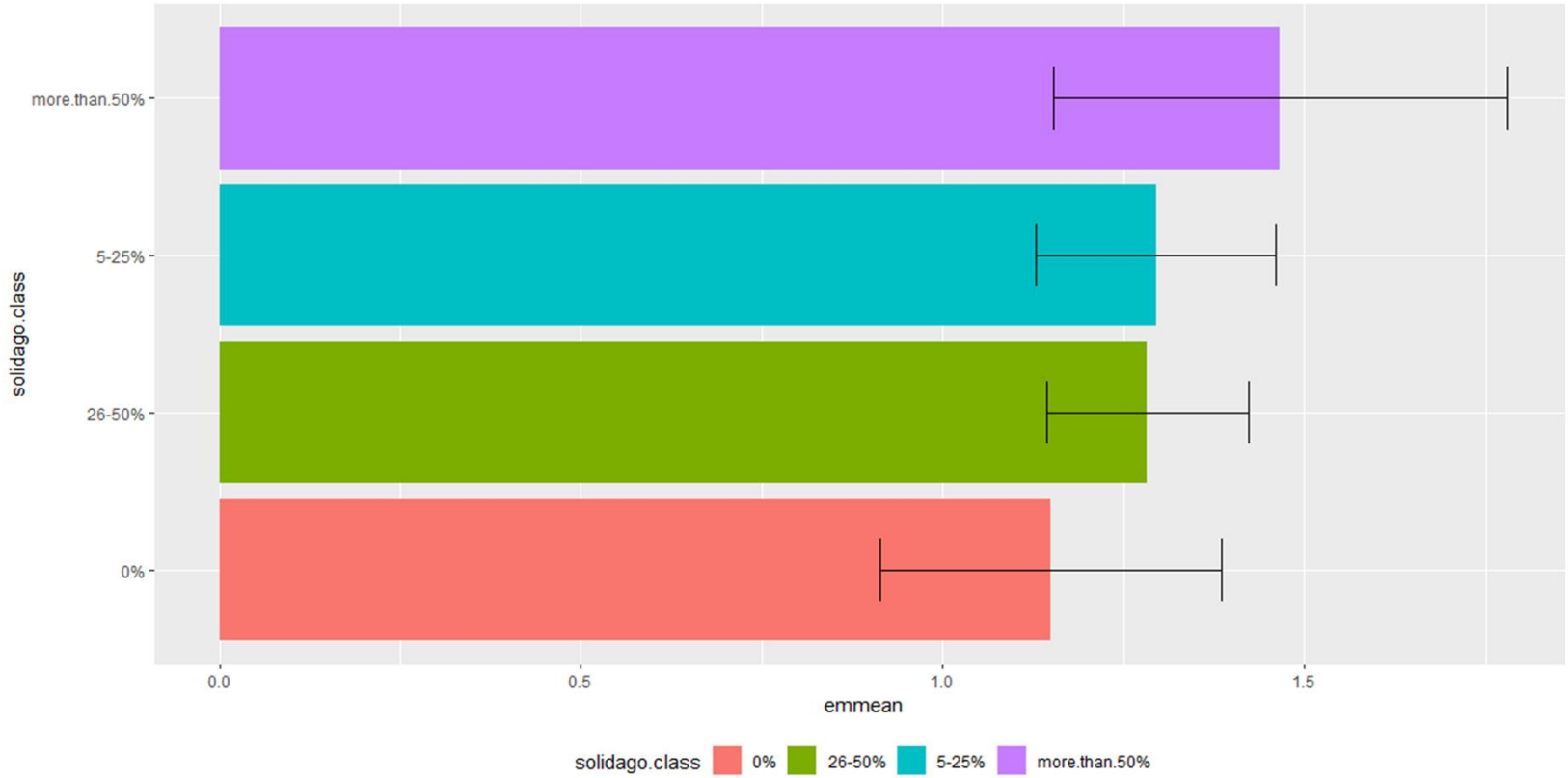
ggplot2 zje emmeansa przygotowanego w takiej postaci:

```
em.orto1<-data.frame(solidago.class=emy.orto1$Sol_class,
                     emmean=emy.orto1$emmean,
                     se=emy.orto1$SE)

emem1<-ggplot(em.orto1, aes(x=solidago.class, y=emmean,
fill=solidago.class)) +
  geom_bar(stat="identity") +
  geom_errorbar(aes(ymin=emmean-se, ymax=emmean+se), width=.4)+
  coord_flip()+
  theme(legend.position = "bottom")
```

Obrazek pokazuje średnie brzegowe (średnie odpowiedzi orto.shan) z modelu przy różnych Sol_class

Innymi słowy: obrazek tego, jak wraz ze zmianą x1 zmieni się y przy założeniu, że x2 się nie zmieni (wpływ canopy.height na prostoskrzydłe jest na stałym/średnim poziomie)

Zasada *ceteris paribus* – wszystkie bez jednego się nie zmieniają



A co z wpływem canopy.height na pasikoniki, przy założeniu, że wpływ solidago.class jest na stałym (średnim) poziomie?

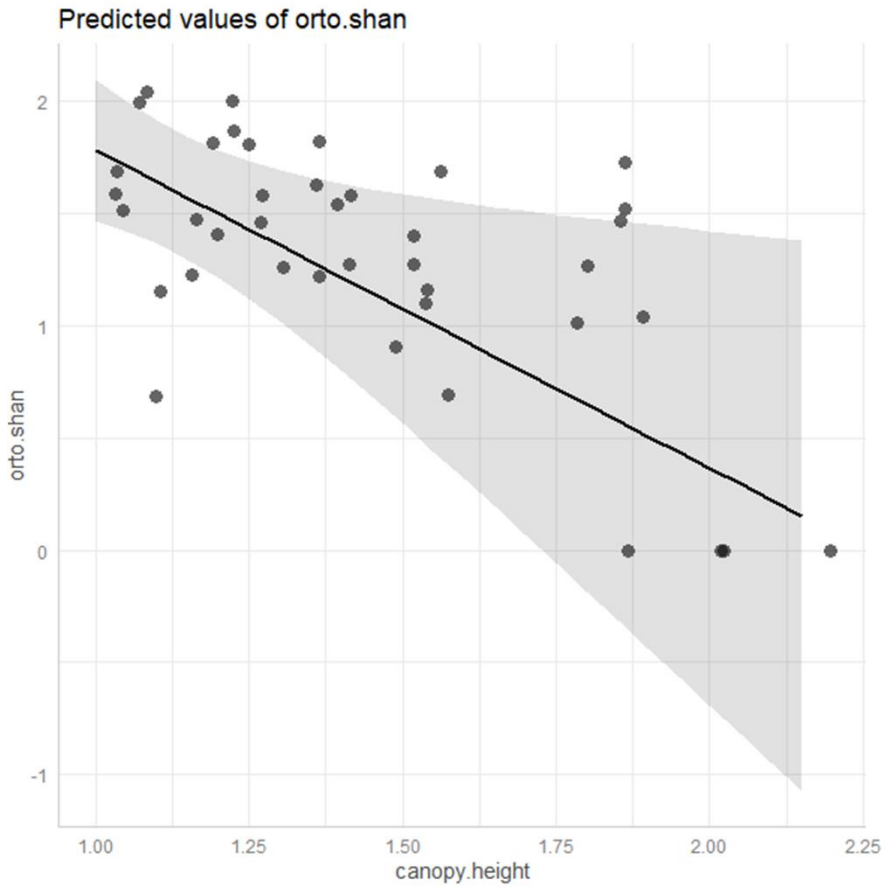
Z pomocą może przyjść ggpredict:

```
library(ggeffects)
data.gg.pred<-ggpredict(mod1)
data.gg.pred
```

```
$canopy.height
# Predicted values of orto.shan

canopy.height | Predicted |         95% CI
-------------------------------------------
         1.00 |      1.78 | [ 1.47, 2.10]
         1.15 |      1.57 | [ 1.30, 1.85]
         1.30 |      1.36 | [ 1.02, 1.70]
         1.45 |      1.15 | [ 0.68, 1.61]
         1.55 |      1.00 | [ 0.44, 1.57]
         1.70 |      0.79 | [ 0.07, 1.51]
         1.85 |      0.58 | [-0.31, 1.47]
         2.15 |      0.15 | [-1.07, 1.38]

Adjusted for:
* Sol_class = 0%

$Sol_class
# Predicted values of orto.shan

Sol_class      | Predicted |         95% CI
-------------------------------------------
more.than.50% |      1.47 | [0.85, 2.08]
5-25%          |      1.29 | [0.97, 1.62]
26-50%         |      1.28 | [1.01, 1.56]
0%             |      1.15 | [0.69, 1.61]

Adjusted for:
* canopy.height = 1.45

attr(,"class")
[1] "ggalleffects" "list"
attr(,"model.name")
[1] "mod2"
```

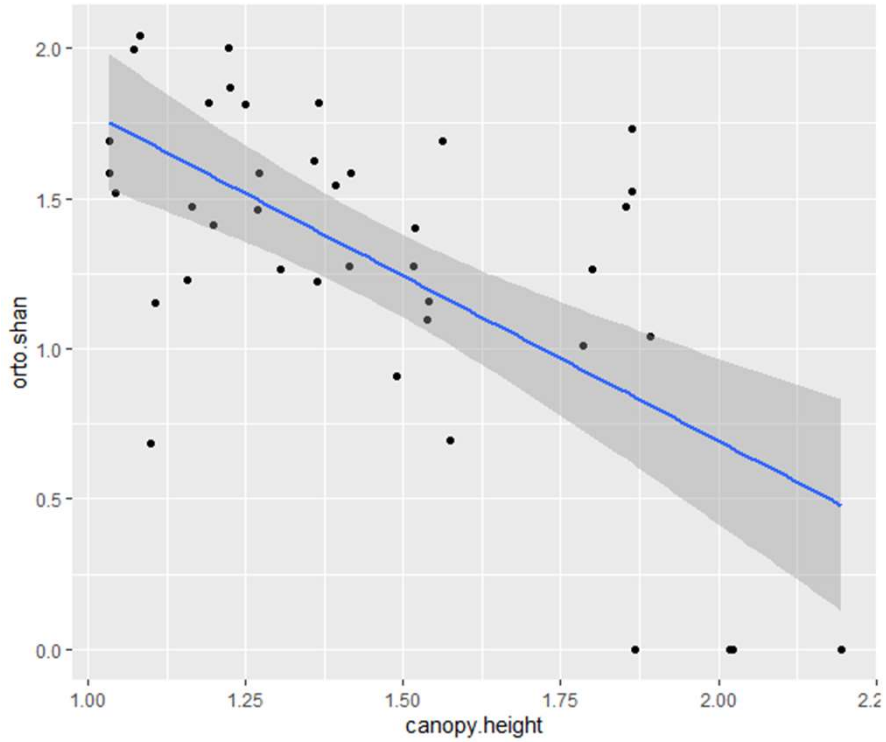
Obrazek dla canopy.height, przy założeniu, że pokrycie Solidago jest na średnim poziomie:

```
p1 <-plot(data.gg.pred, add.data=TRUE, dot.size = 3, dot.alpha = 0.6, dodge=0.5,
line.size=1, jitter=0)$canopy.height
```

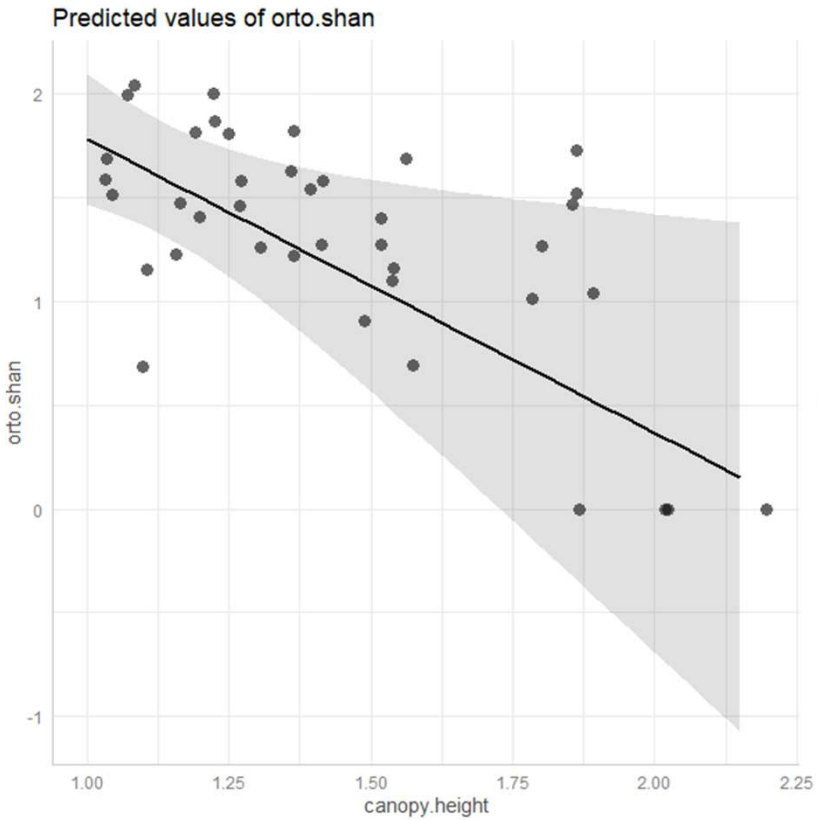


A teraz porównajmy obrazki dla canopy.height

Bez uwzględnienia wpływu nawłoci:



Z uwzględnieniem wpływu nawłoci:



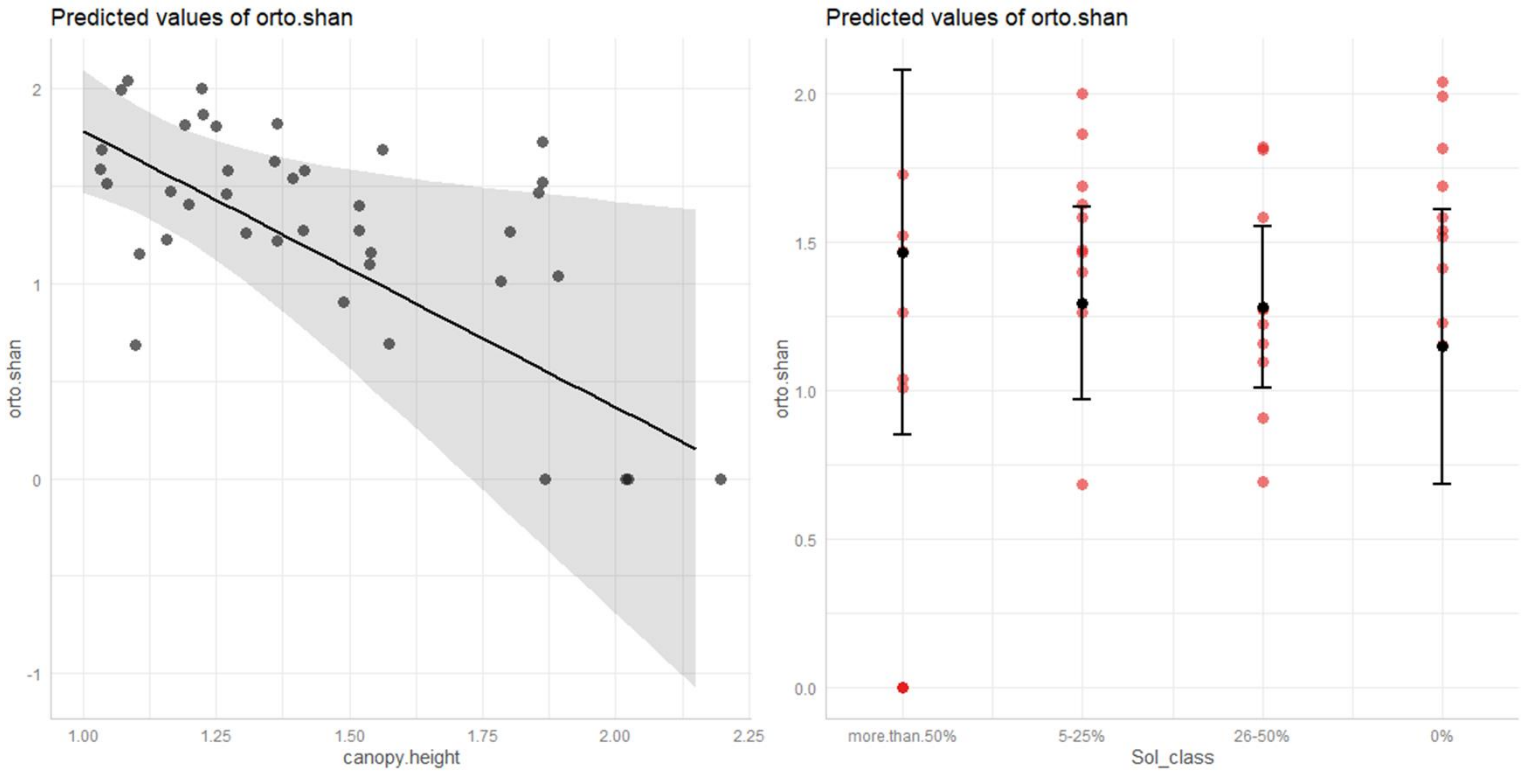
Układ punktów taki sam, ale zmieniło się slope oraz se!!!!

Czyli pokazaliśmy wpływ canopy.height na orto.shan z uwzględnieniem pokrycia nawłoci w modelu

Jak zatem pokazać na obrazku cały model? Dwie opcje:

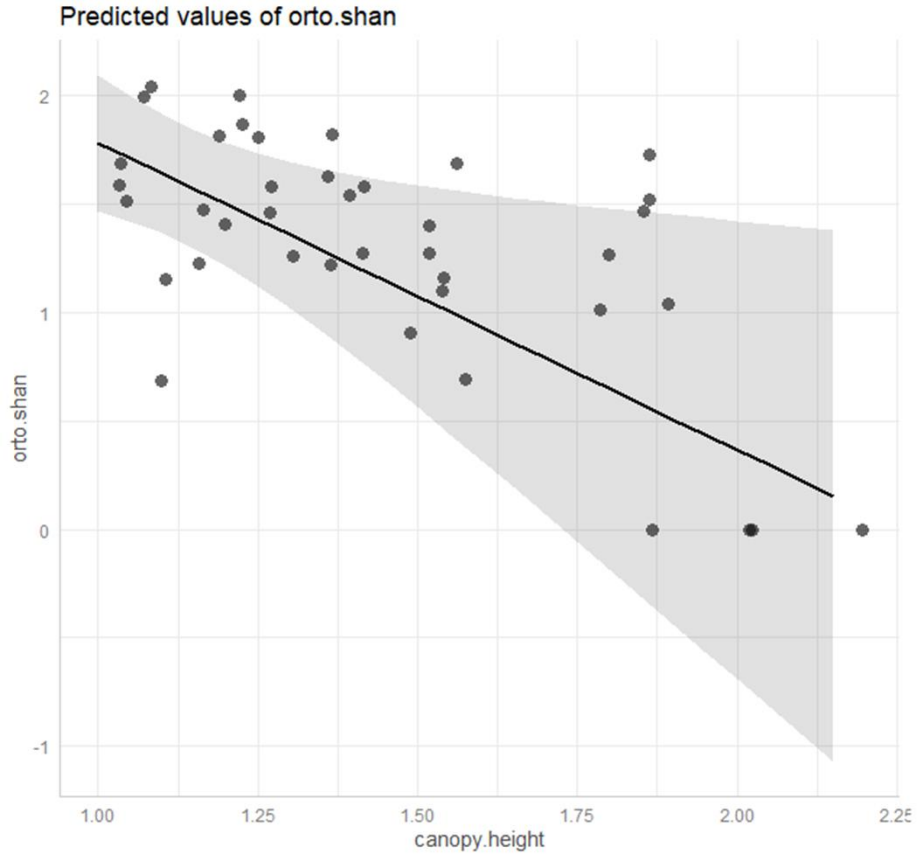
1 – wyciągamy oba obrazki z ggpredict:

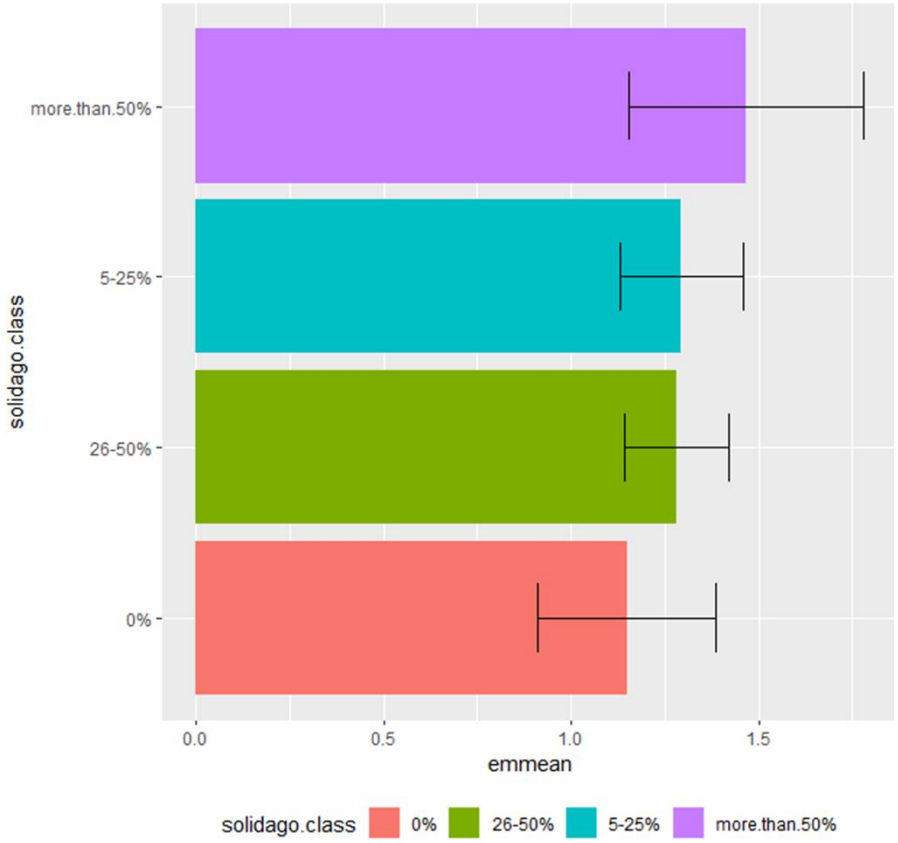
```
p1<-plot(data.gg.pred, add.data=TRUE, dot.size = 3, dot.alpha = 0.6, dodge=0.5,
line.size=1, jitter=0)$canopy.height
p2<-plot(data.gg.pred, add.data=TRUE, dot.size = 3, dot.alpha = 0.6, dodge=0.5,
line.size=1, jitter=0)$Sol_class
grid.arrange(p1, p2, ncol=2, nrow=1)
```



2 – obrazek dla zmiennej kategorycznej wyciągamy z emmeans,
a dla zmiennej ciągłej z ggpredict

grid.arrange(p1, emem1, ncol=2,
nrow=1)





Która opcja lepsza?

Model z interakcją

Kiedy?

Gdy zakładamy różne różnice pomiędzy grupami jednego czynnika w grupach drugiego czynnika

Np. siła i kierunek zależności pomiędzy canopy.height a różnorodnością pasikoników może być różna dla różnych klas pokrycia nawłoci

Co z tego może wynikać? – że np. dla niskich pokryć nawłoci zależność pomiędzy orto.shan a canopy.height może być pozytywna ale słaba, a dla wyższych pokryć nawłoci zależność ta może być negatywna, ale silna

Uzasadnienie biologiczne dla interakcji – zwiększające się pokrycie nawłoci może wpływać np. na dostępność roślin pokarmowych (lub czatowni dla drapieżców) dla prostoskrzydłych

Czy tak jest naprawdę – sprawdźmy to!

Model z interakcją

Znaki w formule:
+ addytywność (wspólne oddziaływanie)
: interakcja
*addytywność i interakcja

Sprawdźmy addytywność i interakcję:

```
mod2<-lm(orto.shan~canopy.height+Sol_class+Sol_class*canopy.height,
data=soli.bss)
summary(mod2)
```

To, co wypluwa funkcja summary raczej mało czytelne

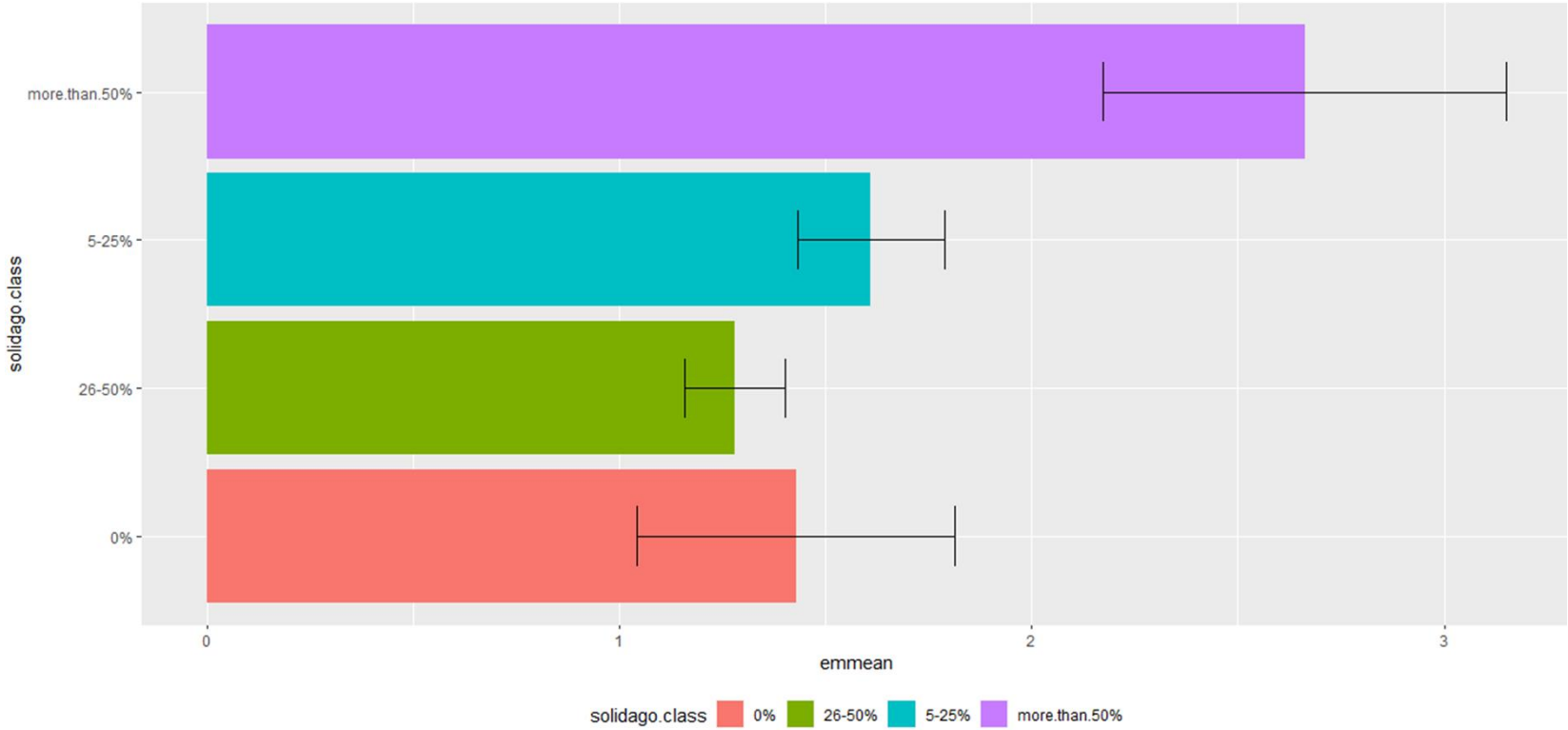
```
Call:
lm(formula = orto.shan ~ canopy.height + Sol_class + Sol_class *
    canopy.height, data = soli.bss)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99750 -0.24905  0.02163  0.21930  0.71351

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                          2.200      1.310   1.679   0.1029
canopy.height                       -0.532      1.153  -0.461   0.6478
Sol_class26-50%                      3.263      2.240   1.457   0.1550
Sol_class5-25%                      -1.602      1.739  -0.921   0.3640
Sol_classmore.than.50%               6.220      2.346   2.652   0.0124 *
canopy.height:Sol_class26-50%       -2.356      1.703  -1.383   0.1762
canopy.height:Sol_class5-25%         1.230      1.448   0.850   0.4017
canopy.height:Sol_classmore.than.50% -3.442      1.535  -2.242   0.0320 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3836 on 32 degrees of freedom
Multiple R-squared:  0.5959,    Adjusted R-squared:  0.5076
F-statistic: 6.743 on 7 and 32 DF,  p-value: 6.098e-05
```

W publikacji można pokazać to w tabelce, ale kluczowe dla opisu wyników oraz ich wizualizacji będzie zrobienie emmeans i ggpredict

Krok1 – pokazujemy pojedynczy wpływ samej nawłoci (bez interakcji):

```
emy.orto2<-cld(emmeans(mod2, ~Sol_class, type="response"))
em.orto2<-data.frame(solidago.class=emy.orto2$Sol_class,
                     emmean=emy.orto2$emmean,
                     se=emy.orto2$SE)
emem2<-ggplot(em.orto2, aes(x=solidago.class, y=emmean, fill=solidago.class)) +
  geom_bar(stat="identity") +
  geom_errorbar(aes(ymin=emmean-se, ymax=emmean+se), width=.4)+
  coord_flip()+
  theme(legend.position = "bottom")
```

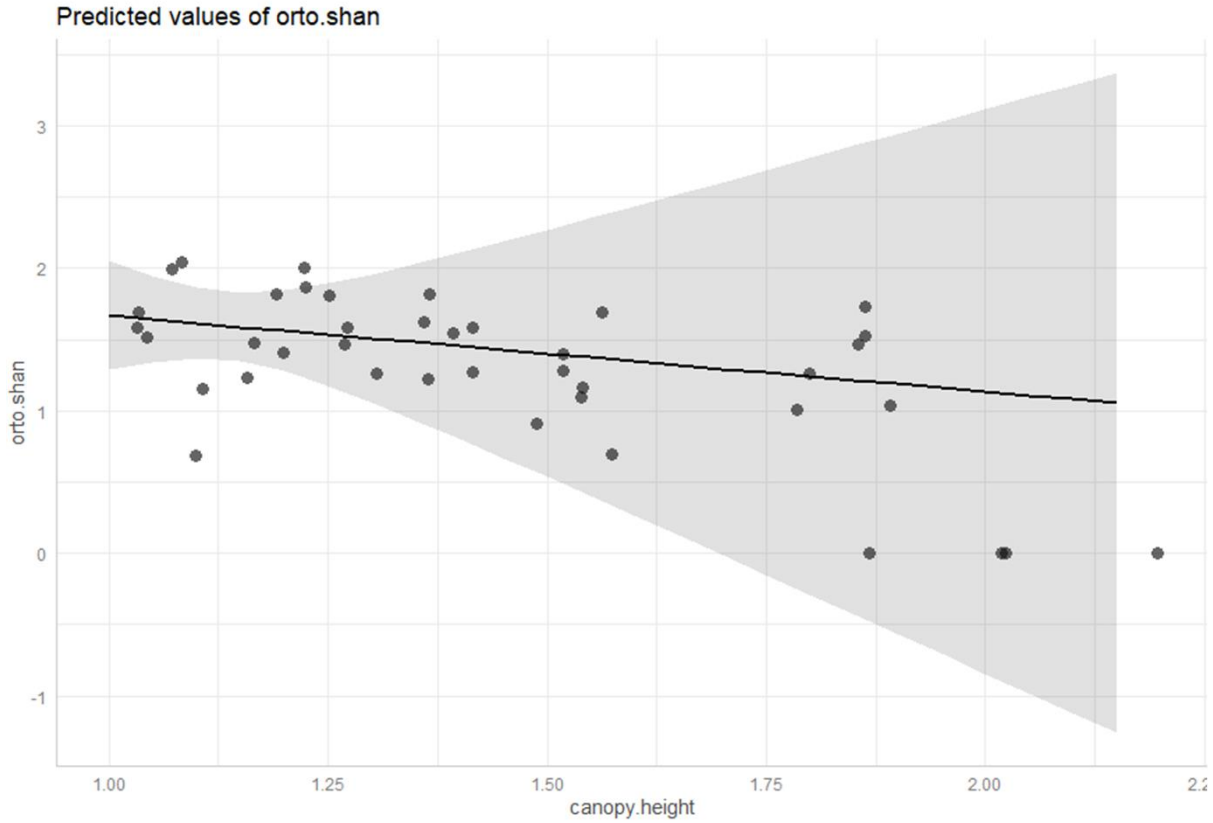
emem2



Widać, że w porównaniu do mod2 (bez interakcji) wpływ nawłoci się zmienił!

Krok2 – pokazujemy pojedynczy wpływ samego canopy.height (bez interakcji):

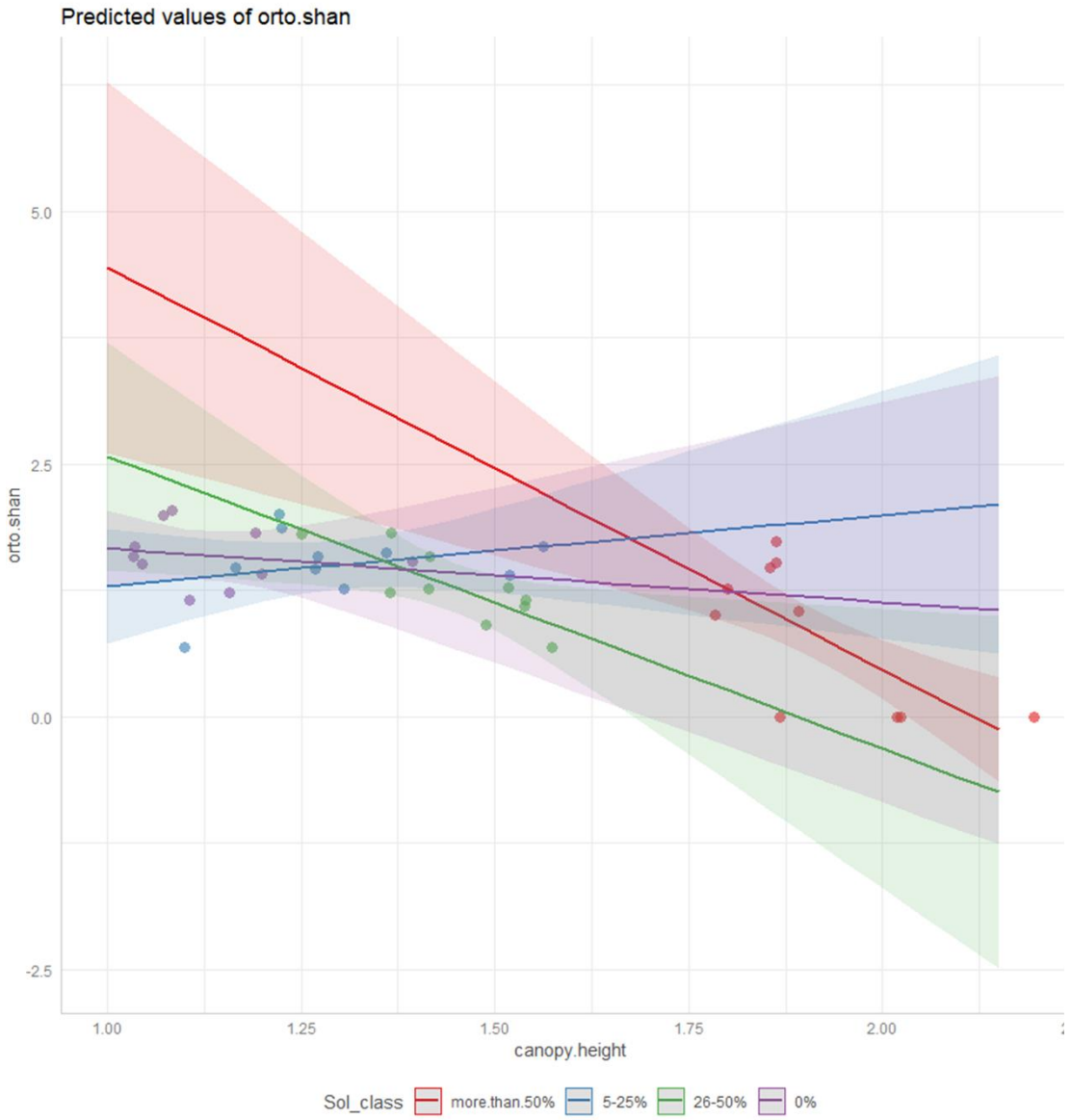
```
data.gg.pred2<-ggpredict(mod2)
obr2<-plot(data.gg.pred2, add.data=TRUE, dot.size = 3, dot.alpha = 0.6, dodge=0.5,
    line.size=1, jitter=0)$canopy.height
obr2
```



Widać, że zależność znowu wygląda inaczej, niż w mod2 (bez interakcji)

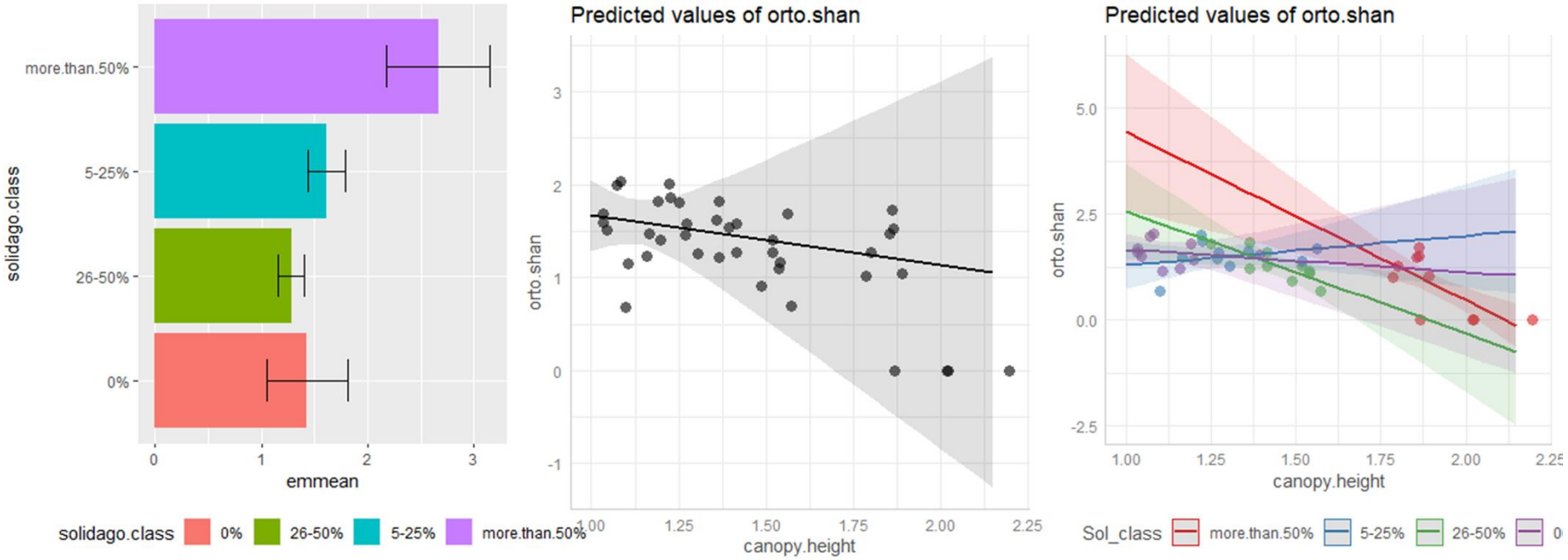
Krok3 – pokazujemy wpływ i nawłoci i canopy.height (interakcja pomiędzy predyktorami):

```
data.gg.pred22<-ggpredict(mod2,
terms = c("canopy.height",
"Sol_class"))
obr3<-plot(data.gg.pred22,
add.data=TRUE, dot.size = 3,
dot.alpha = 0.6, dodge=0.5,
    line.size=1,
jitter=0)+theme(legend.position =
"bottom")
obr3
```



Wizualizacja całego modelu:

grid.arrange(emem2, obr2, obr3, ncol=3, nrow=1)



Wnioski:

- Dla różnorodności prostoskrzydłych duże znaczenie mają większe pokrycia nawłoci per se
- Samo canopy.height nie kształtuje różnorodności Orthoptera
- Interakcje pomiędzy canopy.height a nawłocią bardzo ważne:

1. Najsilniejsza negatywna zależność dla największych pokryć
2. Dla małych pokryć nawłoci zależność pozytywna, ale słaba
3. Dla plotów bez nawłoci brak zależności
4. Model z interakcją pozwolił na identyfikację efektów per capita bezpośrednich i pośrednich inwazji Solidago canadensis


BSS
BIAŁOWIESKA SZKOŁA STATYSTYKI