Lineární modely
(stručné shrnutí)

(Obecné) lineární modely

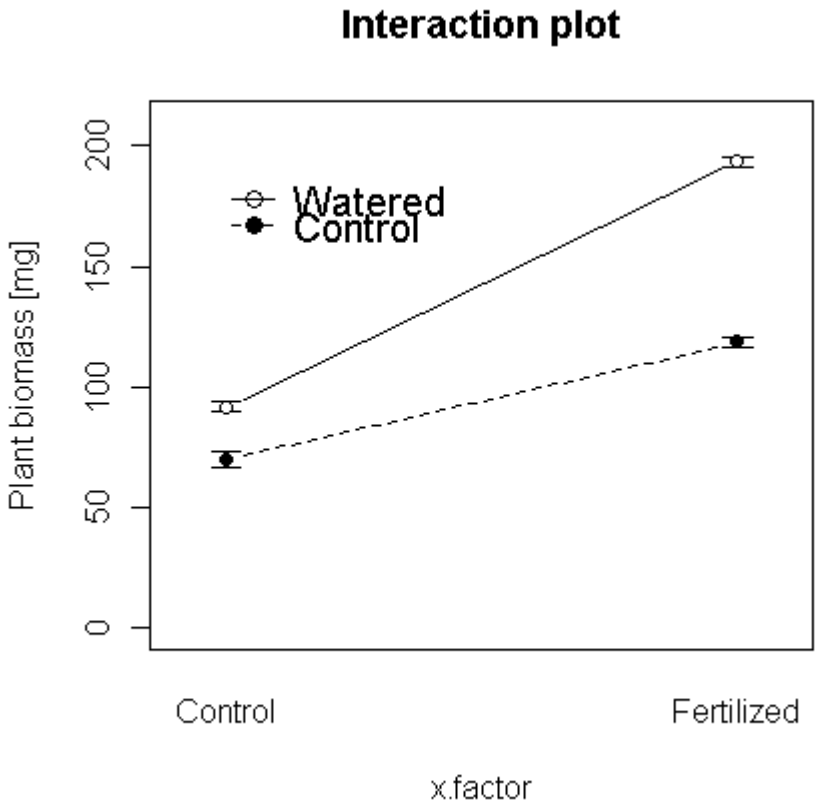
- **Jedna odpověď**, jeden a více prediktorů
  - Prediktory kvantitativní i kategoriální
  - ANOVA s prediktorem o *n* úrovních je analogická mnohonásobné regresi o *n*-1 prediktorech
- Rovnice: $y = a + bx1 + cx2 + \ldots + \varepsilon$
  - y: odpověď
  - a: intercept
  - b,c: regresní koeficienty prediktorů
  - ε: reziduály (residua): předpokládá se, že všechny pocházejí ze stejného normálního rozložení $N(0, \sigma)$

Modely s více prediktory

- Dvoucestná (Vícecestná) ANOVA
  - odpověď ~ faktor.1 + faktor.2 + ...
- Mnohonásobná regrese
  - odpověď ~ lin.prediktor.1 + lin.prediktor.2 + ...
- Lineární model (Analýza kovariance): kategoriální i lineární prediktory
- Aditivní efekty prediktorů vs. interakce – odpověď na faktor.1 (prediktor.1) závisí na hodnotě faktoru.2 (prediktoru.2)
  - aditivitu lze statisticky testovat a případně zamítnout ve prospěch **interakce**

Interakce mezi prediktory

- Průkazná interakce znamená vzájemné ovlivňování vlivu prediktorů na odpověď
  - $y = a + bx_1 + cx_2 + dx_1x_2 + \varepsilon$
- Test interakce H0: d = 0
  - d > 0: pozitivní int., vyšší hodnoty než aditivita
  - d < 0 : negativní int., nižší hodnoty než aditivita
- $df_{int} = df_{x1} * df_{x2}$
- Jak napsat
  - × (Alt + 0215) formálne v odborných textech
  - : v R pouze interakční člen
  - * v R znamená additivitu dohromady s interakcí
  - rovnici nahoře v R napíšeme takto
    $y \sim x_1 * x_2$
- Interaction plot
  Znázornění interakce
  - Pokud int. není přítomna, budou čáry paralelní
- **Interakce neznamená korelaci prediktorů!**



Reakce růstu rostlin na zalévání a hnojení

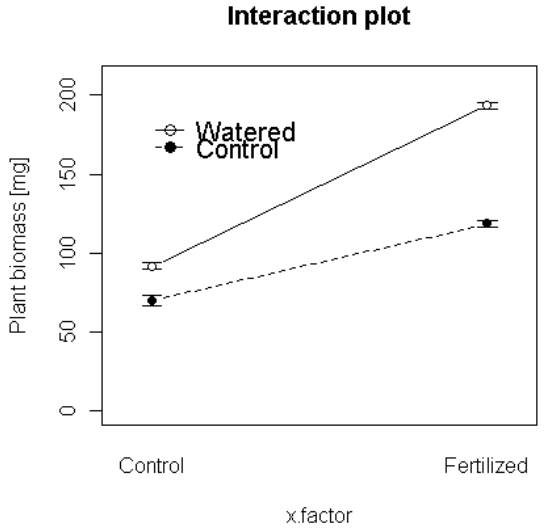
data

biomass	watering	fertil
65	Control	Control
58	Control	Control
74	Control	Control
65	Control	Control
81	Control	Control
78	Control	Control
92	Watered	Control
86	Watered	Control
94	Watered	Control
100	Watered	Control
89	Watered	Control
90	Watered	Control
110	Control	Fertilized
118	Control	Fertilized
128	Control	Fertilized
121	Control	Fertilized
119	Control	Fertilized
116	Control	Fertilized
185	Watered	Fertilized
196	Watered	Fertilized
201	Watered	Fertilized
195	Watered	Fertilized
193	Watered	Fertilized
191	Watered	Fertilized

Experimentální design: 24 květináčů s rostlinami rozdělenými náhodně do 4 skupin s faktoriálním uspořádáním hnojení a velikosti zálivky

Otázka: Jaký je vliv hnojení a zalévání na produkci nadzemní biomasy rostlin?

```
aov.1<-aov(biomass~watering*fertil, data=plants)
summary(aov.1)
                 Df Sum Sq Mean Sq F value Pr(>F)
watering          1  13968   13968  336.38 5.604e-14 ***
fertil            1  33825   33825  814.57 < 2.2e-16 ***
watering:fertil   1   4240    4240  102.11 2.654e-09 ***
Residuals        20    831      42
```



Reakce růstu rostlin na zalévání a hnojení

data

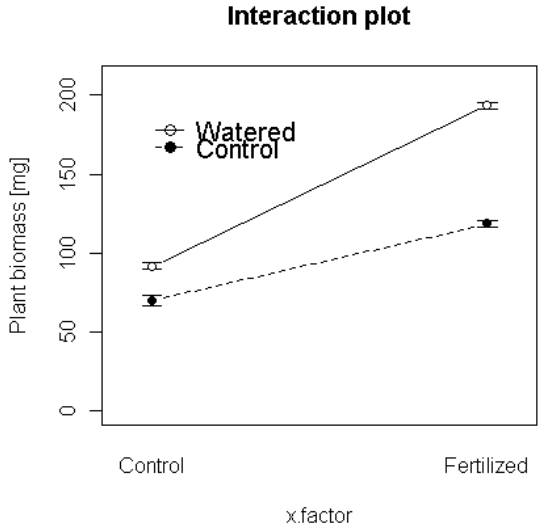
biomass	watering	fertil
65	Control	Control
58	Control	Control
74	Control	Control
65	Control	Control
81	Control	Control
78	Control	Control
92	Watered	Control
86	Watered	Control
94	Watered	Control
100	Watered	Control
89	Watered	Control
90	Watered	Control
110	Control	Fertilized
118	Control	Fertilized
128	Control	Fertilized
121	Control	Fertilized
119	Control	Fertilized
116	Control	Fertilized
185	Watered	Fertilized
196	Watered	Fertilized
201	Watered	Fertilized
195	Watered	Fertilized
193	Watered	Fertilized
191	Watered	Fertilized

Experimentální design: 24 květináčů s rostlinami rozdělenými náhodně do 4 skupin s faktoriálním uspořádáním hnojení a velikosti zálivky

Otázka: Jaký je vliv hnojení a zalévání na produkci nadzemní biomasy rostlin?

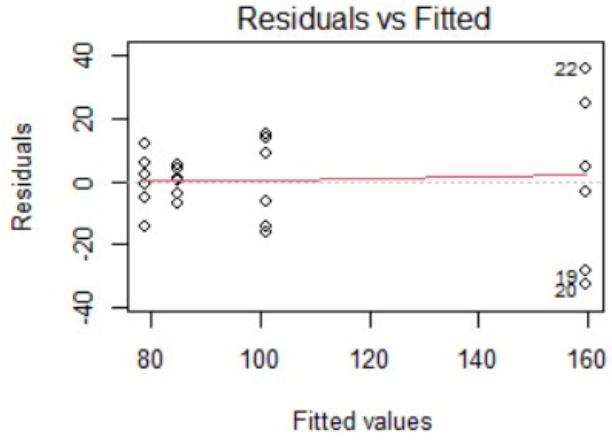
```
aov.1<-aov(biomass~watering*fertil, data=plants)
summary(aov.1)
                 Df Sum Sq Mean Sq F value Pr(>F)
watering          1  13968   13968  336.38 5.604e-14 ***
fertil            1  33825   33825  814.57 < 2.2e-16 ***
watering:fertil   1   4240    4240  102.11 2.654e-09 ***
Residuals        20    831      42
```

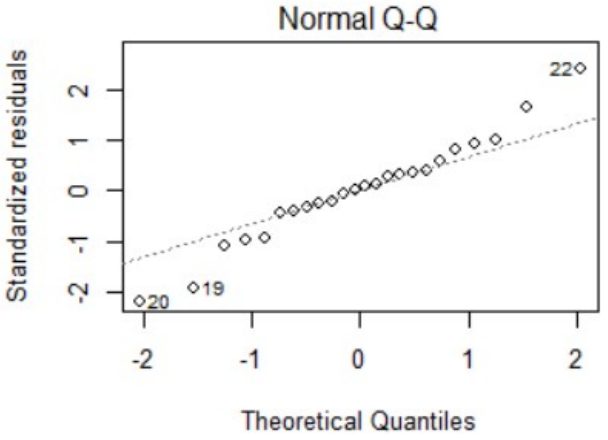
Závěr: Hnojení i zálevání mají průkazný pozitivní vliv růst rostlin. Mezi prediktory je navíc pozitivní interakce.

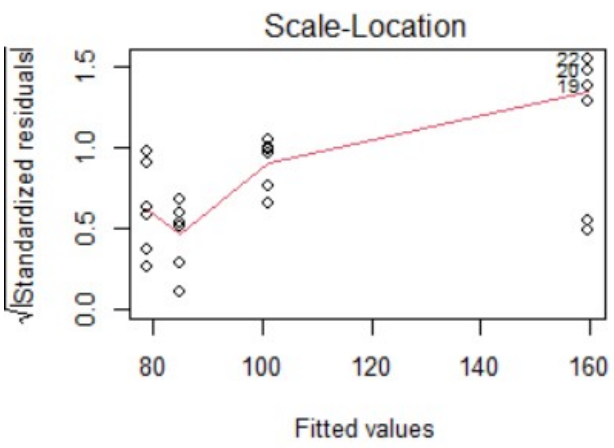


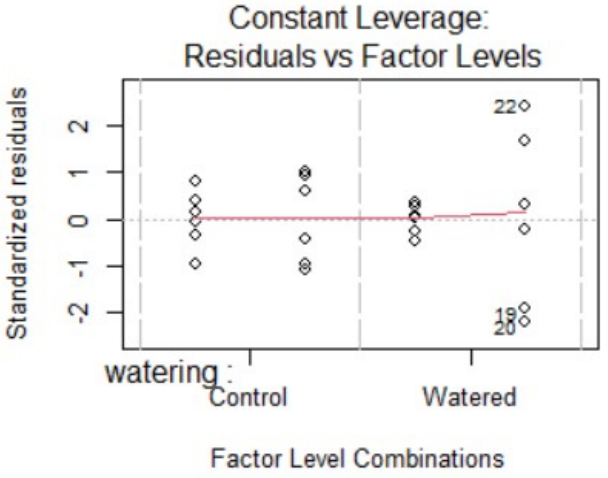
Regresní diagnostika

- Umožňuje vyhodnotit naplnění předpokladů regresních metod (normalita reziduálů, homogenita variancí) a zhodnotit vliv outlierů
- plot(lm.object) nebo plot(aov.object)



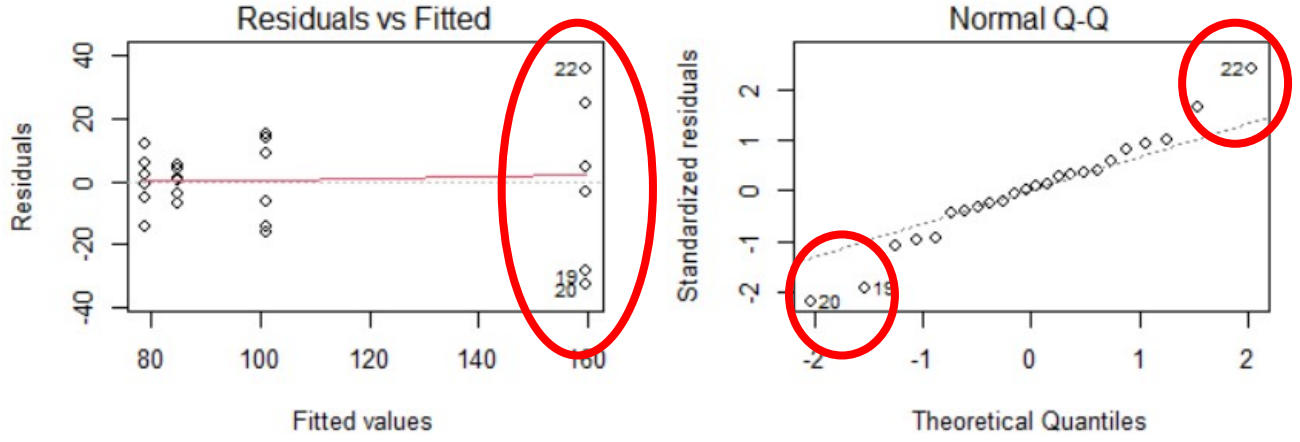


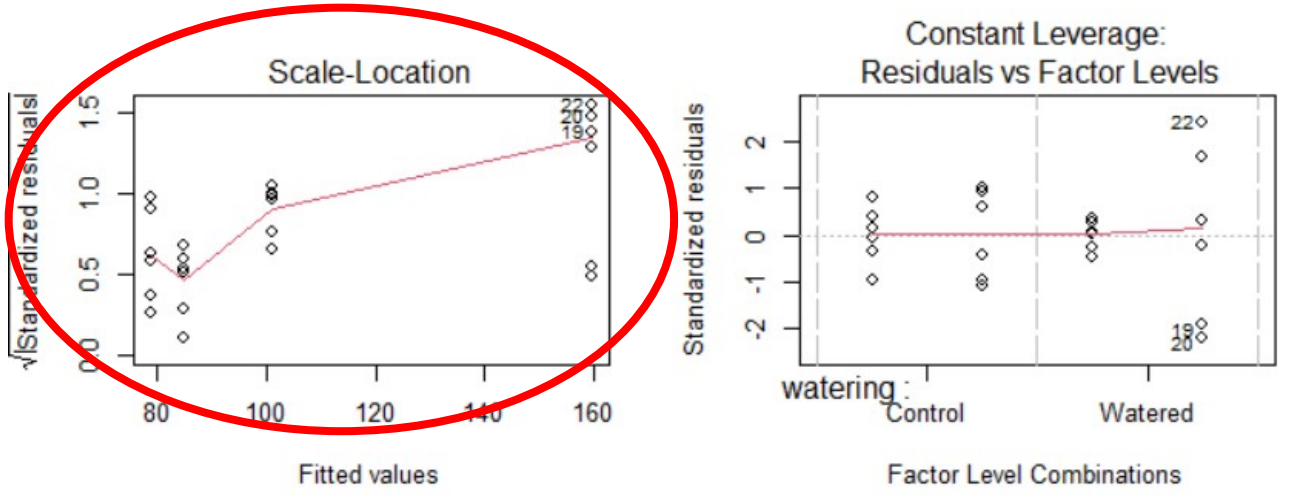




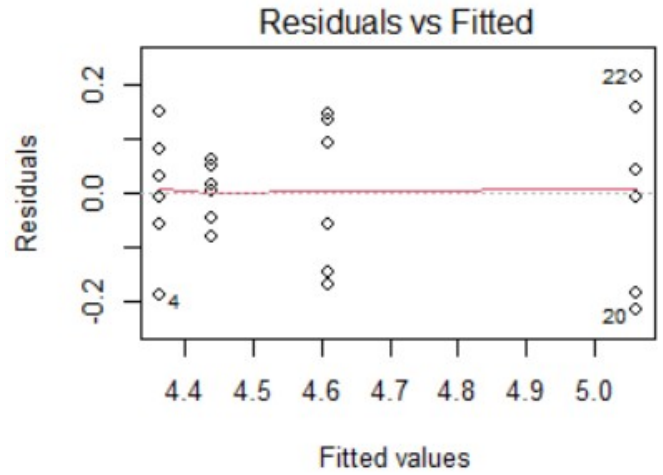
Regresní diagnostika

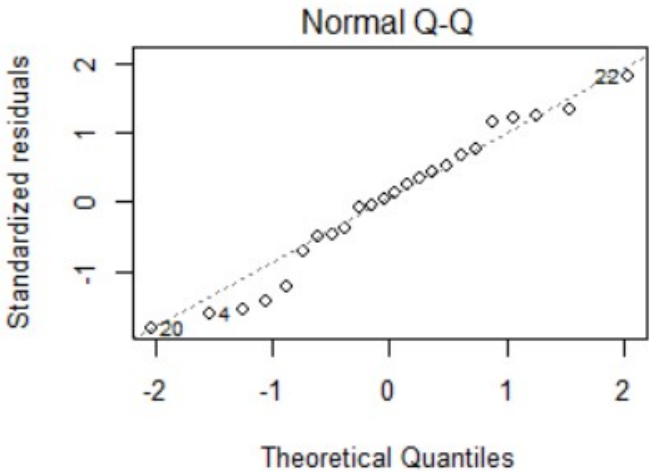
- Umožňuje vyhodnotit naplnění předpokladů regresních metod (normalita reziduálů, homogenita variancí) a zhodnotit vliv outlierů
- plot(lm.object) nebo plot(aov.object)

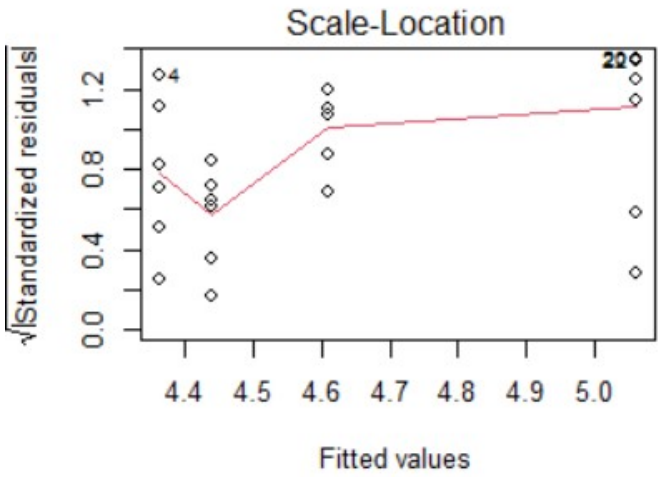


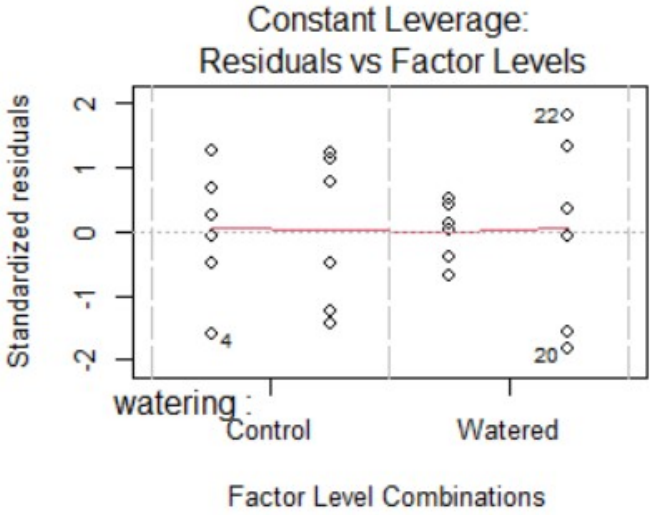


Regresní diagnostika po log-transformaci odpovědi









Testování vlivu jednotlivých prediktorů / výběr prediktorů do modelu

- Cíle
  - ukázat, které prediktory mají průkazný vliv
    - lze testovat jednoduché (marginální) efekty, tj. párové korelace mezi jednotlivými prediktory a odpovědí
  - sestavit minimální adekvátní model, který bude zahrnovat pouze průkazné prediktory
    - kondicionální (parciální) efekty prediktorů
- Statistické testy tohle moc neumí
  - umí otestovat model, porovnat kvalitu více modelů
    - test signifikance
    - AIC

Akaike information Criterion

- Kvantifikuje množství informaci v odpovědi, kterou vysvětluje model
  - umožňuje porovnání dvou modelů, které se liší počtem stupňu volnosti (modelu)
  - nižší AIC značí lepší fit než vyšší AIC (absolutní hodnoty nejsou důležité)
- AIC = 2k – 2log(L) , (log = přirozený log), k je počet parametrů (tedy df model v lin. modelu)
  - v lin. mod. AIC = 2k – 2 log (n/RSS) + C, kde RSS je reziduální suma čtverců, C je konstanta (lze ignorovat)
- Různé názory na možnost kombinovat s F-testem průkaznosti

Výběr prediktorů do modelu

- Postupný
  - Forward selection: k nulovému modelu přidávám postupně prediktory
    - vhodnější pro observační data
  - Backward selection: ze saturovaného modelu odebírám nevýznamné prediktory
    - vhodnější pro experimentální data
  - Oba směry: zvažují v každém kroku přidání I odebrání prediktorů

Problém s mnohonásobným paralelním testováním

- Provedu-li více testů s pravděpodobností chyby I. druhu 0.05, pravděpodobnost, že udělám chybu aspoň jednou velmi vzrůstá.
  - pro dva testy p = 0.05 + 0.05 – 0.05 x 0.05 = 0.0975
- Řešení: různé korekce (Holm, Bonferroni, false detection rate – FDR) upravují p-hodnoty nahoru, čímž kontrolují/redukují riziko chyby
- Alternativa: “protected multiple testing”
  - spočtu test saturovaného modelu se všemi prediktory. Je-li průkazný, další korekci kvůli multiple testing už neřeším
  - Je-li ovšem neprůkazný, tak s testováním končím se závěrem, že regrese není průkazná.
- Skutečný problém hlavně v analýzách “velkých” dat z databází.

Lineární model pro experimentální data

Sample	seedlings	treatment	productivity	temperature
1	7	control	714	7.2
2	5	control	518	4.5
3	12	control	379	7.4
4	8	control	686	5.4
5	13	control	703	5
6	12	control	775	6
7	7	control	651	6.2
8	5	control	630	7.6
9	7	control	470	8.4
10	6	control	557	4.7
11	19	grazing	394	6.8

... to be continued

Q: Jaký je vliv obhospodařování louky (kosení, pastva) na počet semenáčků, které se objeví na jaře? Ovlivňují semenáčky ještě další faktory prostředí?

Design: 30 experimentálních ploch (10 na každý typ obhospodařování) náhodně rozmístěné v krajině (různé lokality). Control = opuštěná nekosená louka. Zaznamenána byla i produktivita a průměrná teplota lokality.

```
lm.full<-lm(seedlings~treatment*productivity*temperature, data=seedl)
anova(lm.full)
# Analysis of Variance Table
#
# Response: seedlings
#                                    Df  Sum Sq Mean Sq F value    Pr(>F)
# treatment                           2 618.07 309.033 31.5710 1.301e-06 ***
# productivity                        1 290.61 290.605 29.6883 3.553e-05 ***
# temperature                         1   1.88   1.882  0.1922    0.6663
# treatment:productivity              2   5.02   2.509  0.2563    0.7767
# treatment:temperature               2   6.87   3.434  0.3508    0.7088
# productivity:temperature            1   0.25   0.250  0.0256    0.8747
# treatment:productivity:temperature  2   6.48   3.242  0.3312    0.7223
# Residuals                          18 176.19   9.789
# ---
#    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tuto tabulku lze použít jako výstup testu jednotlivých prediktorů a jejich interakcí

Lineární model pro experimentální data

Další krok: odstraňte neprůkazné prediktory z modelu (pomocí backward selekce; ponechejte I neprůkazné main efekty pokud je průkazná interakce)

```
lm.final<-lm(seedlings~treatment+productivity, data=seedl)
summary(lm.final)
Call:
lm(formula = seedlings ~ treatment + productivity, data = seedl)

Residuals:
   Min     1Q Median     3Q    Max
-4.698 -1.840 -0.315  1.975  4.741

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.315470   2.437909   9.154 1.29e-09 ***
treatmentgrazing  9.581024   1.251180   7.658 3.98e-08 ***
treatmentmowing   2.173362   1.268726   1.713   0.0986 .
productivity     -0.024764   0.003996  -6.198 1.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.75 on 26 degrees of freedom
Multiple R-squared: 0.8221,   Adjusted R-squared: 0.8015
F-statistic: 40.04 on 3 and 26 DF,  p-value: 6.857e-10
```

Koeficienty prediktorů (nebo kontrasty pro faktory)

Celkový test modelu

Lineární model pro experimentální data

Další krok: odstraňte neprůkazné prediktory z modelu (pomocí backward selekce; ponechejte I neprůkazné main efekty pokud je průkazná interakce)

```
lm.final<-lm(seedlings~treatment+productivity, data=seedl)
summary(lm.final)
Call:
lm(formula = seedlings ~ treatment + productivity, data = seedl)

Residuals:
   Min     1Q Median     3Q    Max
-4.698 -1.840 -0.315  1.975  4.741

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      22.315470   2.437909   9.154 1.29e-09 ***
treatmentgrazing  9.581024   1.251180   7.658 3.98e-08 ***
treatmentmowing   2.173362   1.268726   1.713   0.0986 .
productivity     -0.024764   0.003996  -6.198 1.48e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.75 on 26 degrees of freedom
Multiple R-squared: 0.8221,   Adjusted R-squared: 0.8015
F-statistic: 40.04 on 3 and 26 DF,  p-value: 6.857e-10
```

Koeficienty prediktorů (nebo kontrasty pro faktory)

Celkový test modelu

Závěr: Průkazný vliv na počet semenáčků má způsob obhospodařování a produktivita mají. Jejich efekty jsou aditivní. Produktivita snižuje počet semenáčků. Obhospodařivání jejich počet zvyšuje. Průkazné zvýšení je ale způsobené pouze pastvou.

Lineární model pro observační data

- Typicky mnoho potenciálních prediktorů
  - mnohdy nemožnost fitovat saturovaný model (potenciálních prediktorů více než pozorování)
- Forward selection **s korekcí** – akceptovatelná varianta
- Korelované potenciální prediktory: těžko řešitelný problém
  - např. v ČR: nadmořská výška, průměrná teplota a úhrn srážek
  - Možnost prezentovat jednoduché korelace spolu s finální lineárním modelem

Výběr prediktorů do lineárního modelu

- Spousta možností jak to udělat v zásadě dobře
  - postupný výběr s korekcí (ať už jakoukoliv)
  - využití “test protection”
- Jednoznačně špatně – Statistical fishing
  - Forward selekce bez korekce pro observační data
    - speciálně při použití mnoha potenciálních prediktorů
  - Úpravy testů “aby to vyšlo průkazně”