Vliv odlehlých hodnot, korelační koeficient, mnohonásobná regrese

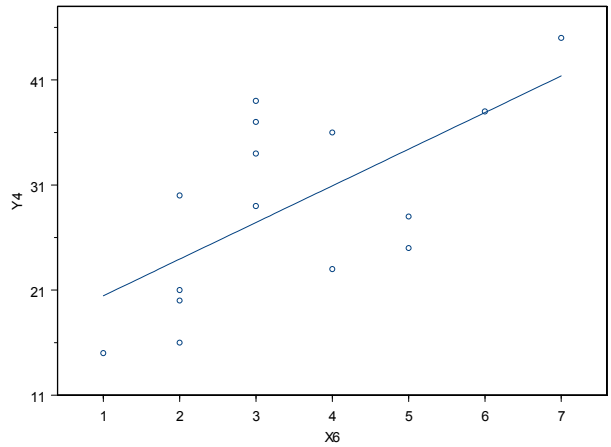
1. Vliv odlehlých hodnot

Na následujících dvou příkladech ukážeme jak ***odlehlé hodnoty*** (outliers) ovlivňují výsledek analýzy a jak je identifikovat.

Relativní velikostí listové plochy (plocha/hmotnost sušiny) a vysvětlujeme výšku rostliny.

Rel. vel. asimilační plochy; X6	1	2	2	2	4	5	5	3	2	3	4	3	6	3	7
Výška; Y4	15	16	20	21	23	25	28	29	30	34	36	37	38	39	45

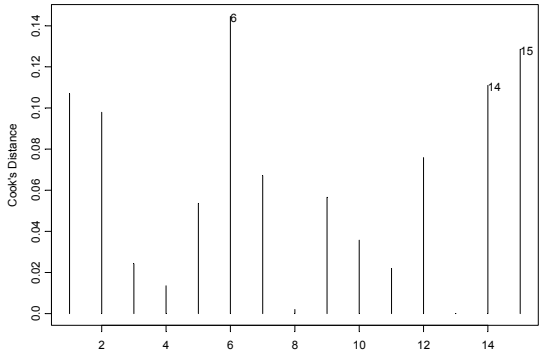
Graficky vyjádříme závislost.



Pokud spočítáme analýzu pak množství vysvětlené variability modelem je "pouze" okolo 42 % avšak celkově je model statisticky průkazný (0.0085).

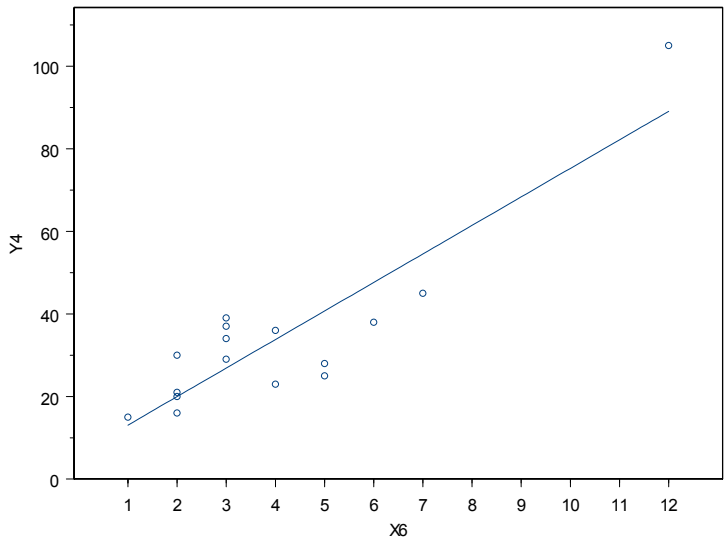
```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 16.9799  4.3156     3.9345  0.0017
         X6  3.4866  1.1269     3.0940  0.0085

Residual standard error: 7.103 on 13 degrees of freedom
Multiple R-Squared: 0.4241
F-statistic: 9.573 on 1 and 13 degrees of freedom, the p-value is 0.008545
```

Pokud bychom chtěli identifikovat odlehlá měření, která by mohla výrazněji ovlivňovat výsledek analýzy, kromě uložení hodnot reziduí a jejich vizuální kontrole (histogram...) můžeme v průběhu zadávání modelu na záložce "Plot" zaškrtnout "Cook's Distance" graf. V tomto grafu vynášíme míru vlivu každého měření na výslednou hodnotu regresního koeficientu. Hodnoty vyšší než 1 jsou považovány za výrazně ovlivňující regresní koeficient. V grafu jsou automaticky označeny tři nejvyšší hodnoty.



Můžeme tedy konstatovat, že žádné měření není odlehlé natolik aby ovlivnilo nežádoucím způsobem výsledný regresní model

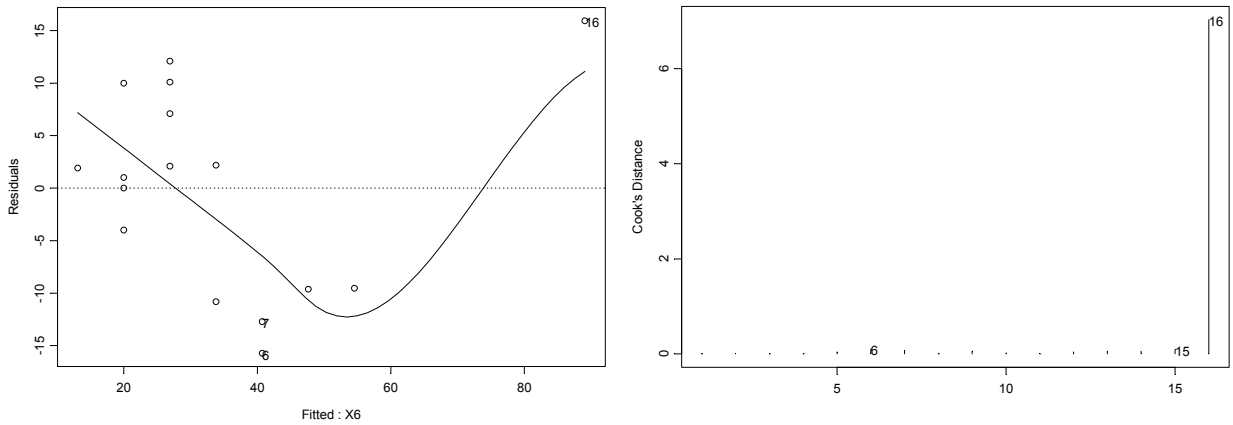
Vliv odlehlých hodnot ukážeme pokud přidáme na konec našeho datového souboru další měření (X6 = 12, Y4 = 105), které je výrazně vzdálené jak ve směru osy x tak i osy y.



Pokud spočteme regresi, pak hodnoty jednotlivých koeficientů se výrazně změní. Zvýší se jak signifikance celeého modelu ale i dvojnásobně se zvýší i hodnota determinačního koeficientu (R-Squared). Celkový regresní model je nyní lepší.

```
      *** Linear Model ***

Call: lm(formula = Y4 ~ X6, data = SDF3, na.action = na.exclude)
Residuals:
    Min      1Q Median     3Q    Max
 -15.72 -9.558  1.456 7.822 15.93

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 6.1829 4.5757     1.3512  0.1981
         X6 6.9074 0.9593     7.2002  0.0000

Residual standard error: 9.97 on 14 degrees of freedom
Multiple R-Squared: 0.7874
F-statistic: 51.84 on 1 and 14 degrees of freedom, the p-value is 4.563e-06
```

Přestože jde o odlehlé měření, hodnota reziduálu není vůči ostatním nijak extrémní a není z něj nějak patrné, že jde o odlehlé měření.



Avšak graf "Cook's Distance" ukazuje, že ona šestnáctá (odlehlá) hodnota výrazně překračuje 7. Tato jediná hodnota pak "fixuje" regresní přímku v levé horní části grafu a díky tomu se markantně zvýšila výpovědní síla celého regresního modelu. Jediné měření tedy může mít na celkový regresní model velmi podstatný vliv. Toto odlehlé měření však může být zcela v pořádku s podstatou studované závislosti. Přesto je však vhodné před definitivním závěrem jeho hodnotu ověřit. Může být výsledkem chyby při zadávání dat či chybného odečtu hodnot v terénu. V případě, že chceme z výpočtu regrese vyloučit nějakou hodnotu (např. šestnáctou hodnotu), pak v příkazovém řádku použijeme hranatých závorek a voláme: `model<-lm(y[-16]~x[-16]).`

2. Korelační koeficient

V případě, že nemůžeme jednoznačně určit, která z proměnných je vysvětlující a která vysvětlovaná (např. délka vs. šířka křídla motýla) není správné počítat regresi. Za těchto podmínek ale můžeme stanovit míru těsnosti vztahu obou proměnných (korelaci). Výsledkem této analýzy je korelační matice (symetrická dle diagonály), která udává hodnotu korelačního koeficientu r (Pearsonův nebo Spearmanův koef. používaný pro pořadí). Dále pro každý korelační koeficient je možné spočítat hladinu významnosti t testu nulové hypotézy, kdy korelační koeficient je roven nule.

výška rostliny; Y4	15	16	20	21	23	25	28	29	30	34	36	37	38	39	45
X4	1	2	2	2	4	5	5	3	2	3	4	3	6	3	7
X5	1	1	2	1	3	5	5	2	0	1	2	0	1	0	2
X6	1	2	3	4	5	6	8	9	12	13	14	15	16	17	20

X4 - počet dní od vytvoření prvního listu, X5 - relativní příkon radiace, X6 - relativní velikost asimilační plochy

V S+ spočítáme korelační matici pro proměnné X1, X2, X3, Y4, v menu Statistics>Data Summaries>Correlations... Pokud chceme spočítat testovací statistiky pro jednotlivé korelační koeficienty musíme použít příkazové okno. Pro výpočet voláme funkci `cor.test` s parametry: první proměnná, druhá proměnná; další parametry jsou volitelné:

`alternative = "two.sided"` (default)/`"greater"/"less"` (můžeme zkrátit a použít pouze první písmena, např. `a="g"`), `method = "pearson" (default)/"kendall"/"spearman"`.
Výsledná korelační matice je tedy

```
      ***  Correlations for data in:  SDF3 ***

numeric matrix: 4 rows, 4 columns.
           y4        x4         x5         x6
y4  1.0000000 0.6512199 -0.1891716  0.9917707
x4  0.6512199 1.0000000  0.4795634  0.5905029
x5 -0.1891716 0.4795634  1.0000000 -0.2862575
x6  0.9917707 0.5905029 -0.2862575  1.0000000
```

Testovací statistiky pro jednotlivé parametry spočteme dle výše uvedeného návodu. Počítáme oboustranou hypotézu pro Pearsonův koeficient, takže vyplníme pouze testované proměnné.

```
>cor.test (SDF3$y4, SDF3$x4)


      Pearson's product-moment correlation

data:  SDF3$y4 and SDF3$x4
t = 3.094, df = 13, p-value = 0.0085
alternative hypothesis: true coef is not equal to 0
sample estimates:
       cor
 0.6512199
```

Výsledek říká, že korelační koeficient pro proměnné y4 a x4 a hladina významnosti je 0.0085. Pokud srovnáme dosaženou hladinu významnosti pro tento korelační koeficient s hladinou významnosti regresního koeficientu z prvního příkladu na regresi v tomto dokumentu, zjistíme, že je shodná (0.0085) Pro jednoduchou lineární regresi platí, že hodnota Pearsonova korelačního koeficientu závislé a nezávislé proměnné umocněná na druhou se rovná R-Squared ($0.65122^2 = 0.424087$).

Pokud data zdvojíme (zkopírujeme hodnoty na počet měření 30) můžeme sledovat jak bude ovlivněna hodnota korelačního koeficientu (opět jen příklad, není to návod jak zpřesňovat odhady!!!) a odpovídající hladiny významnosti.

```
      ***  Correlations for data in:  SDF3 ***

numeric matrix: 4 rows, 4 columns.
           y4        x4         x5         x6
y4  1.0000000 0.6512199 -0.1891716  0.9917707
x4  0.6512199 1.0000000  0.4795634  0.5905029
x5 -0.1891716 0.4795634  1.0000000 -0.2862575
x6  0.9917707 0.5905029 -0.2862575  1.0000000

> cor.test (SDF3$y4, SDF3$x4)

      Pearson's product-moment correlation

data:  SDF3$y4 and SDF3$x4
t = 4.5408, df = 28, p-value = 0.0001
alternative hypothesis: true coef is not equal to 0
sample estimates:
       cor
```

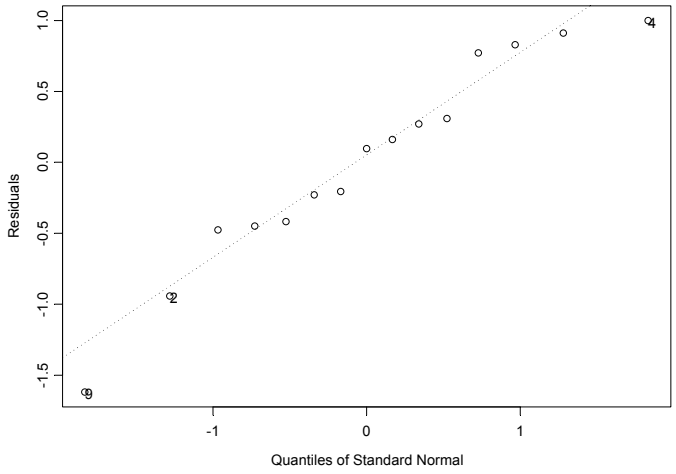
```
0.6512199
```

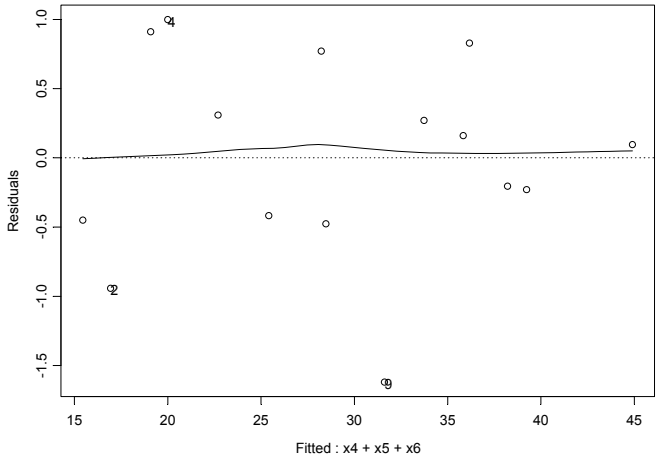
Zvýšením počtu měření se korelační koeficient nezměnil, avšak získali jsme mnohem větší jistotu pro odhad, tzn. zvýšila se hladina dosažené pravděpodobnosti. Výsledkem jsou shodné hodnoty r s "lepšími" hladinami pravděpodobnosti příslušných t testů. T test srovnává hodnotu korelačního koeficientu ku střední chybě jeho odhadu, která klesá s rostoucím počtem měření. Tento vztah platí i obecně (nejen při zkopírování dat). S větším počtem měření , za předpokladu opravdu náhodného a nezávislého výběru, se korelační koeficient výrazně nezmění. Měříme totiž stále stejnou skutečnost, ale odhad korelačního koeficientu je podstatně přesnější.

3. Mnohonásobná regrese, postupný výběr (stepwise regression)
Jestliže máme jednu vysvětlovanou a několik vysvětlujících proměnných pro analýzu jejich vztahu používáme mnohonásobnou lineární regresi. Cílem analýzy je stanovit hodnoty parciálních regresních koeficientů b_i pro každou vysvětlující proměnnou. Jejich kombinace pak vytváří regersní rovnici.

výška rostliny; Y4	15	16	20	21	23	25	28	29	30	34	36	37	38	39	45
X4	1	2	2	2	4	5	5	3	2	3	4	3	6	3	7
X5	1	1	2	1	3	5	5	2	0	1	2	0	1	0	2
X6	1	2	3	4	5	6	8	9	12	13	14	15	16	17	20

Výpočet regrese opět voláme z menu Statistics>Regression>Linear... Zadáme proměnné (obecný vzorec je: y4~x4+x5+x6), zvolíme jaké typy výstupů budeme chtít (Long Output). Z grafů vybereme pro kontrolu reziduí Residuals vs. Fit, Residuals Normal QQ a graf Partial Residuals. Po výběru posledně zmíněného grafu se aktivuje podokno s možnostmi jeho nastavení. Vzhledem ke snadnější čitelnosti jednotlivých grafů pro parciální regresní koeficienty je vhodnější nemít zaškrtnuté "Common Y-axis Scale".





Grafy rozložení reziduálů ukazují, že použitý lineární model je v pořádku a odpovídá závislosti studovaných proměnných.

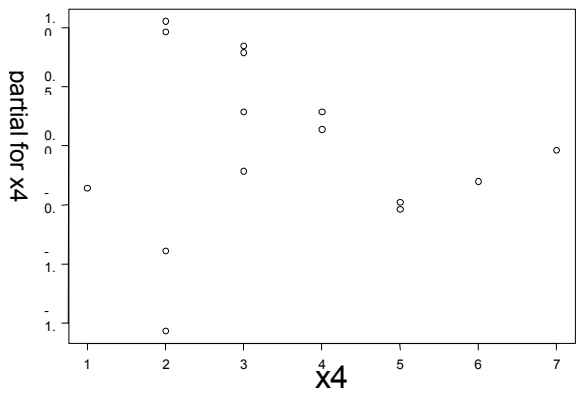
Výsledná tabulka regrese ukazuje, že pro proměnné x5 a x6 je parciální regersní koeficient odlišný od 0, ale pro proměnnou x4 již ne. Pokud zkombinujeme jednotlivé regresní koeficienty je výsledná regresní rovnice tvaru:

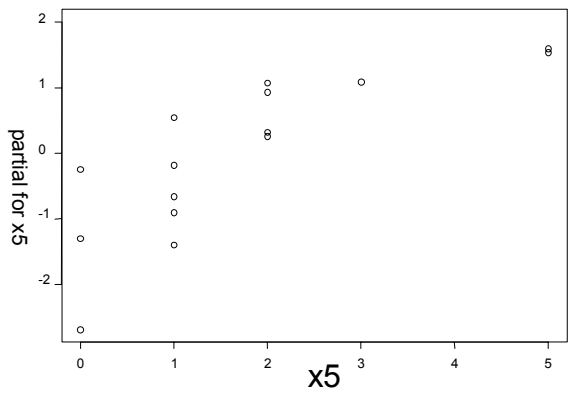
`y = 13.3409 - 0.0374x4 + 0.6174x5 + 1.5295x6`, avšak vliv proměnné x4 je zřejmě minimální. Model vysvětluje více jak 99 % variability v datech a je vysoce průkazný (2.879e-012)

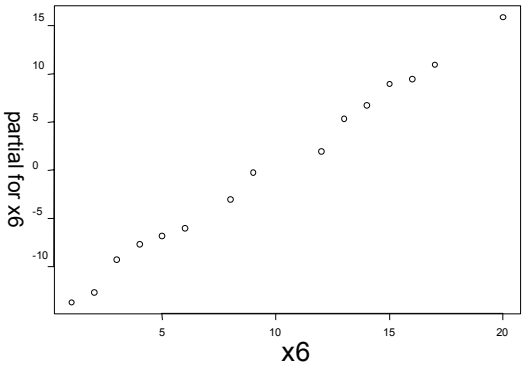
```
      *** Linear Model ***

Call: lm(formula = y4 ~ x4 + x5 + x6, data = SDF3, na.action =
      na.exclude)
Residuals:
   Min      1Q  Median    3Q     Max
 -1.62 -0.4341 0.09585 0.54 0.9985

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  13.3409    0.5466    24.4070   0.0000
         x4  -0.0374    0.2986    -0.1253   0.9025
         x5   0.6174    0.2682     2.3016   0.0419
         x6   1.5295    0.0761    20.1029   0.0000

Residual standard error: 0.8271 on 11 degrees of freedom
Multiple R-Squared: 0.9934
F-statistic: 551.3 on 3 and 11 degrees of freedom, the p-value is 2.879e-
012
```

Problémem mnohonásobné regrese je možnost, že vysvětlující proměnné jsou navzájem závislé. Pokud to tak je, pak jednotlivé proměnné postihují (částečně či úplně) stejnou část variability závislé proměnné. Pokud je tedy vysvětlující potenciál jedné nezávislé proměnné obsažen v jiné nebo v kombinaci jiných nezávislých proměnných, pak je tato proměnná pro daný model zcela nadbytečná. V našem příkladě je velmi pravděpodobné, že proměnná x4 není pro regresní model nezbytná.

Graficky je významnost jednotlivých nezávislých proměnných pro celkový model vyjádřena grafy "Partial Residuals". Každý graf vynáší hodnoty parciálních reziduí jednotlivých proměnných na sebe samé, kdy parciální rezidua jsou získána odfiltrováním vlivu zbývajících nezávislých proměnných. Čím více se tato závislost blíží lineární, tím větší je vysvětlující síla dané proměnné.

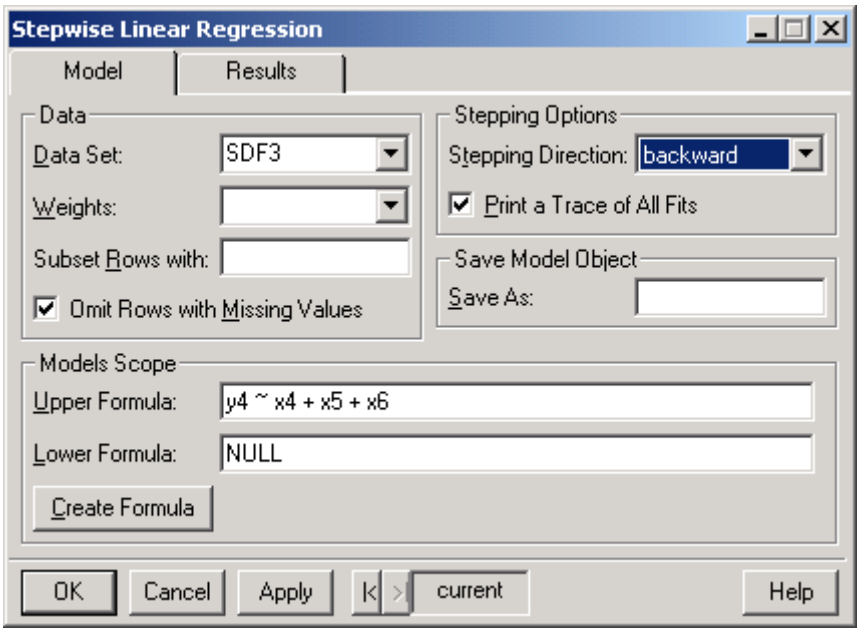






U proměnné x6 je závislost jasně lineární a u proměnné x5 je poměrně zřetelná pro proměnnou x4 jsou rezidua téměř rovnoměrně rozprostřena. To ukazuje na minimální vysvětlující sílu této proměnné.

Předchozí odstavce nás vedou k úvaze, zda je nezbytně nutné v regresním modelu zahrnout všechny tři nezávisle proměnné *x*. V S+ máme možnost vybrat regresi s postupným výběrem či postupným přidáváním proměnných. Z menu volíme Statistics>Regression>Stepwise...., v poli "Upper Formula" zadáme "nejsložitější" model a v poli "Lower Formula" necháme nulový model. Máme k dispozici tři varianty postupného výběru, od nejjednoduššího k nejsložitějšímu, od nejsložitějšího k nejjednoduššímu a kombinaci obou. Pro náš příklad použijeme "backward" Stepping Direction.



Pro srovnání modelů slouží statistika Cp. Tabulka

```
       Df Sum of Sq      RSS      Cp
<none>                 7.5246  12.9971
    x4  1    0.0107   7.5354  11.6397
    x5  1    3.6237  11.1484  15.2527
    x6  1  276.4454 283.9701 288.0744
```

uvádí hodnoty Cp statistiky pro modely, kde v prvním sloupci je uveden faktor, který je z vyšší varianty modelu vyloučen. (první řádek uvádí model se všemi proměnnými, druhý variantu bez proměnné x4 atd.
Poté se z analýzy vyloučí proměnná jejíž vyloučení snížilo statistiku Cp pod hodnotu Cp statistiky pokud žádná proměnná nebyla vyloučena. Tedy v tomto případě vyloučíme proměnnou x4 a znovu přepočítáme statistiku Cp bez proměnné x4.

```
       Df Sum of Sq      RSS       Cp
<none>                 7.535   11.640
    x5  1    11.133   18.668   21.404
    x6  1  1090.640 1098.176 1100.912
```

V tomto případě již žádná varianta odebrání proměnné nesníží Cp statistiku pod Cp statistiku celkového modelu a tak výběr proměnných skončí. Dále uvádím výpis celé "Stepwise" analýzy.

```
       *** Stepwise Regression ***

       *** Stepwise Model Comparisons ***
Start:  AIC= 12.9971
 y4 ~ x4 + x5 + x6

Single term deletions

Model:
y4 ~ x4 + x5 + x6

scale:  0.6840561

       Df Sum of Sq      RSS       Cp
<none>                 7.5246  12.9971
    x4  1    0.0107   7.5354  11.6397
    x5  1    3.6237  11.1484  15.2527
    x6  1  276.4454 283.9701 288.0744

Step:  AIC= 11.6397
 y4 ~ x5 + x6

Single term deletions

Model:
y4 ~ x5 + x6

scale:  0.6840561

       Df Sum of Sq      RSS       Cp
<none>                 7.535   11.640
    x5  1    11.133   18.668   21.404
    x6  1  1090.640 1098.176 1100.912

       *** Linear Model ***

Call: lm(formula = y4 ~ x5 + x6, data = SDF3, na.action = na.exclude
      )
Residuals:
    Min      1Q  Median      3Q     Max
 -1.595 -0.4318 0.05643 0.5424 0.9859

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 13.3398  0.5236    25.4751  0.0000
         x5  0.5892  0.1399     4.2106  0.0012
         x6  1.5213  0.0365    41.6754  0.0000

Residual standard error: 0.7924 on 12 degrees of freedom
Multiple R-Squared: 0.9934
F-statistic: 900.9 on 2 and 12 degrees of freedom, the p-value is 8.382e-014
```

Výsledek analýzy je v poslední části. Výsledný model popisuje více jak 99 % variability dat.

V případě, kdy chceme provést výběr jednotlivých proměnných sami na základě Cp statistiky v příkazovém řádku, postupujeme takto. Nejprve vytvoříme "nejsložitější" model, který zahrnuje všechny dostupné proměnné: `model<-lm(y4~x4+x5+x6)`. Pak voláme funkci `drop1`,

která postupně odebere jednotlivé proměnné a spočítá pro výsledné modely statistiku Cp. Voláme `drop1 (model)` a výsledkem je stejná tabulka jako v automatickém výběru:

```
       Df Sum of Sq      RSS       Cp
<none>                7.5246  12.9971
    x4  1    0.0107   7.5354  11.6397
    x5  1    3.6237  11.1484  15.2527
    x6  1  276.4454 283.9701 288.0744
```

Nejnižší hodnotu Cp statistiky má model bez proměnné X4 a tak z původního modelu vyjmeme požadovanou proměnnou. Použijeme funkci update. Ta má syntaxi update (původní model, co z něj měním), dále používá tzv. tečkovou konvenci. Vysvětlení na příkladu, odebírám z modelu `model` proměnnou `X4` a nový model ukládám pod jménem `model.1`: `model.1<-update(model, .~. -X4)`, kdy tečky okolo vlnovky značí původní levou a pravou stranu závislosti a mínus X4 značí odebrání proměnné.
Opět použiji funkci drop1 na změněný model model.1:

```
       Df Sum of Sq      RSS       Cp
<none>                 7.535   11.640
    x5  1    11.133   18.668   21.404
    x6  1  1090.640 1098.176 1100.912
```

A vidíme, že již není třeba odebírat další proměnné a dostali jsme stejný výsledek jako při automatickém výběru.
Na závěr ještě můžeme ověřit, zda se náš zjednodušený model liší od původního komplikovaného. Voláme: `anova (model, model.1)` a pokud nám vyjde, že rozdíly mezi modely jsou neprůkazné, můžeme dále pracovat se zjednodušeným modelem.

Tento materiál byl vytvořen na základě příkladů a textů P. Sklenáře pro kurz biostatistiky v NCSS.