Regresní analýza; transformace dat

Pro řešení vztahů mezi proměnnými kontinuálního typu používáme ***korelační a regresní analýzy***. Korelace se používá pokud nelze určit "kauzalitu". Regresní analýza je určena pro řešení vztahů, kdy máme jednu závislou *(y)* a jednu či více závislých *(x)* proměnných. Předpoklad regresní analýzy je, že nezávislé proměnné jsou měřené s nulovou chybou (nebo je alespoň její chyba oproti závislé velmi malá). Pro jednu nezávislou proměnnou za předpokladu lineární závislosti je vztah roven $y=a+bx$. Regresní analýza hledá parametry *a* a *b*, kde *a* je absolutní člen (průsečík s osou y) a *b* je směrnicí regresní přímky.

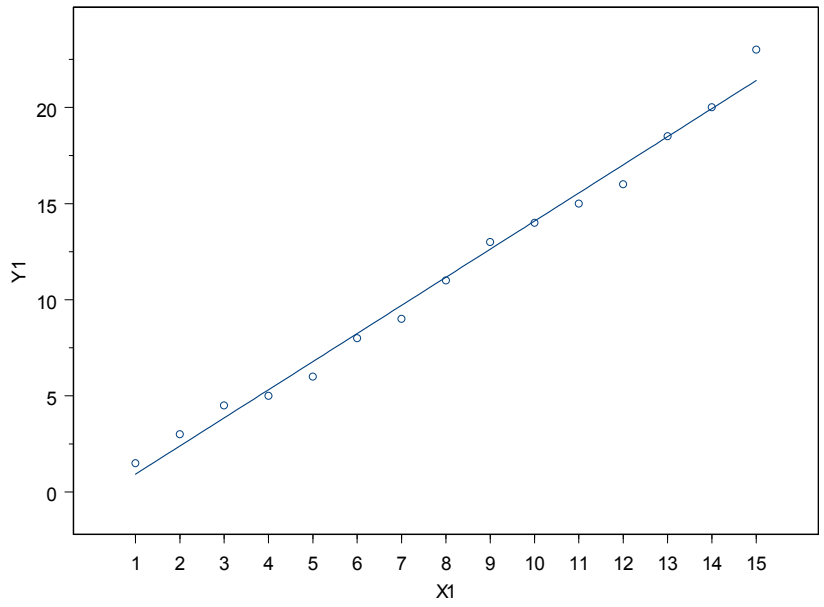
V příkladu máme k dispozici údaje o růstu rostliny v čase (počet dnů).

dny; X1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
výška; Y1	1.5	3	4.5	5	6	8	9	11	13	14	15	16	18.5	20	23

Pokud data zadáme přes "Commands" okno:

```
SDF1$X1<-c(1:15)          //vytvoření řady čísel od 1 do 15
SDF1$Y1<-c(1.5,3,4.5,5,6,8,9,11,13,14,15,16,18.5,20,23)
```

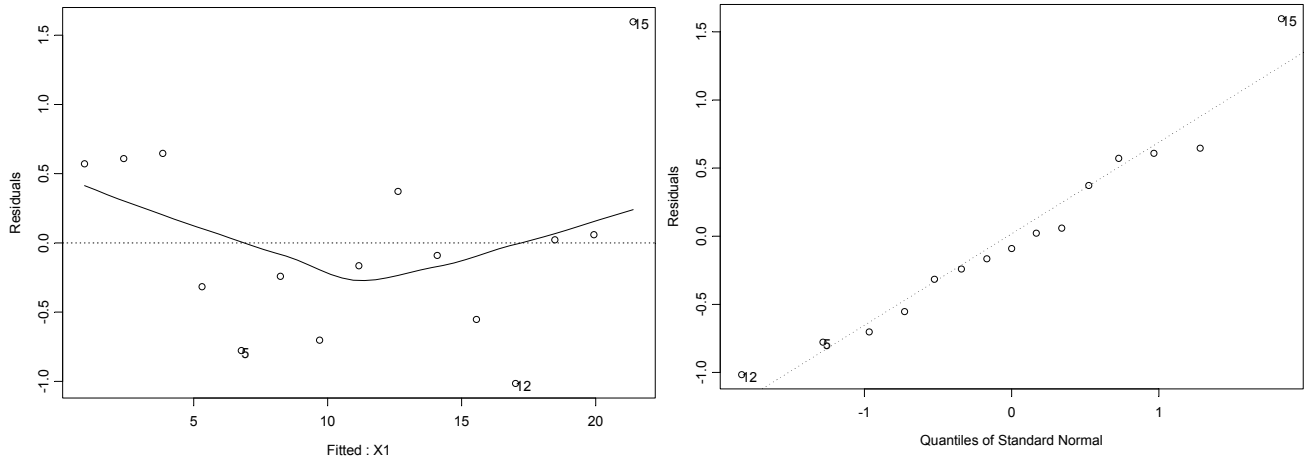
Nejdříve je vhodné si data zobrazit. K tomu použijeme graf s funkcí proloženou daty. K tomu slouží typ grafů "Fit...", pro proložení přímky vybereme "Fit - Linear Least Squares (x, y1, y2...)". Tím jsme dostali grafické zobrazení studovaného vztahu i s regresní přímkou. Stejný graf z příkazové řádky získáme zadáním: `plot (x, y)` a pak `abline (lm(y~x))`. `abline` je funkce pro vykreslení lineární závislosti daného modelu, v tomto případě lineární regrese (`lm`) se závislou proměnnou `y` a nezávislou `x` (viz níže).



Regresní analýzu najdeme v S+ přes menu Statistics>Regression>Linear... Sestavíme obecný vzorec pro náš příklad, tedy "Dependent" je Y1 a "Independent" X1. Pokud budeme chtít s modelem pracovat později, pak si zvolíme pod jakým jménem jej uložíme (Save Model Object). Na záložce "Results" zaškrtneme požadované výstupy (Long Output a ANOVA Table). Pokud budeme dále pracovat s reziduály, zaškrtneme je a vybereme cílové místo kam

budou uloženy. Stejně tak i fitované hodnoty. Na záložce "Predict" můžeme rovnou propojit spočtenou regresi s dalším datovým souborem a uložit výsledek nafitovaných hodnot. Grafy nalezneme na záložce "Plot". Zde bychom měli vybrat grafy pro následnou analýzu reziduí (Residuals vs. Fit, Residuals Normal QQ). Další grafy (histogram) můžeme vytvořit z uložených hodnot reziduí.
Při práci pomocí příkazového řádku postupujeme takto: podle zvoleného jména (např. model1) zadáme `model1<-lm(SDF1$Y1~SDF1$X1)`. `lm` určuje že data budou fitována lineárním modelem, v případě že máme v proměnných některé prázdné buňky, pak je třeba zadat doplňující parametr na `model1<-lm(SDF1$Y1~SDF1$X1, na.action = na.exclude)`. Výsledek regrese uvidíme po zadání: `summary (model1)`. Další obecné funkce pro práci s modely viz. příloha.

Z výsledků S+ nejprve zobrazí grafy. Jedním z předpokladů regresní analýzy, podobně jako u ANOVy, je normalita rozdělení reziduálů. Použití lineárního modelu je oprávněné tehdy, je-li studovaná závislost v datech (přibližně) lineární. Pokud je to pravda, pak rozdělení reziduálů nevykazuje žádný trend a blíží se normalitě. Vizuální kontrola rozdělení reziduálů je důležitá pro ověření správnosti použitého regresního modelu. Pokud máme model uložený, jak z nabídky tak i z příkazové řádky, pro diagnostické grafy voláme `plot(model1)`.
Ideální rozložení reziduí vůči predikované nebo vysvětlované proměnné je v rovnoměrném pásu okolo nulové hodnoty.



Po prohlídce rozdělení reziduí v tomto případě můžeme použitý lineární model považovat za vhodný. Stejně tak i shrnutí rozdělení reziduálů ve výsledné tabulce ukazuje na symetrii.

Výsledkem analýzy je rovnice `y = -0.533+1.4625x`. Hodnota Intercept odpovídá absolutnímu členu, parametru *a*. Základní otázkou regresní analýzy je zda se regresní koeficient b průkazně liší od nuly, zda existuje statistická závislost mezi proměnnými *x* a *y*. Náš test (t-test) zamítá nulovou hypotézu, že regresní koeficient není odlišný od nuly. Parametr R-Squared uvádí, kolik variability v datech bylo vysvětleno modelem, v tomto případě se tedy jedná o téměř 99 %.

```
Call: lm(formula = Y1 ~ X1, data = SDF3, na.action = na.exclude)
Residuals:
    Min      1Q   Median      3Q   Max
```

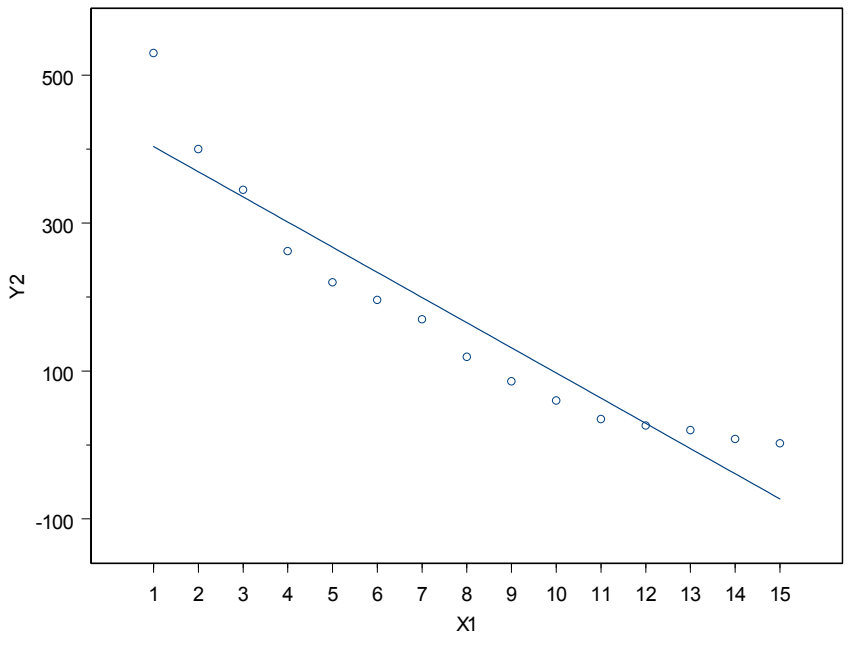
```
 -1.017 -0.4354 -0.09167 0.4708 1.596

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  -0.5333   0.3824    -1.3946   0.1865
         X1   1.4625   0.0421    34.7710   0.0000

Residual standard error: 0.7038 on 13 degrees of freedom
Multiple R-Squared: 0.9894
F-statistic: 1209 on 1 and 13 degrees of freedom, the p-value is 3.264e-014

Analysis of Variance Table

Response: Y1

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
       X1  1  598.8938 598.8938 1209.025 3.264056e-014
Residuals 13    6.4396   0.4954
```

Součástí regresní analýzy je také analýza variance příslušného regresního modelu, která testuje zda model vysvětluje signifikantní množství variability v datech. V příkazovém řádku zadáváme `anova (model1)`. V tabulce je množství variance vysvětlené modelem a zbytková variance. Tato analýza je testem průkaznosti modelu. R-squared je podíl sumy čtverců modelu ku celkové sumě čtverců. Pro jednoduchou regresi je tato analýza testem, zda se regresní koeficient průkazně liší od nuly (je tedy analogií příslušného T-testu). Pokud má model více nezávislých proměnných, pak analýza variance testuje průkaznost celého modelu.

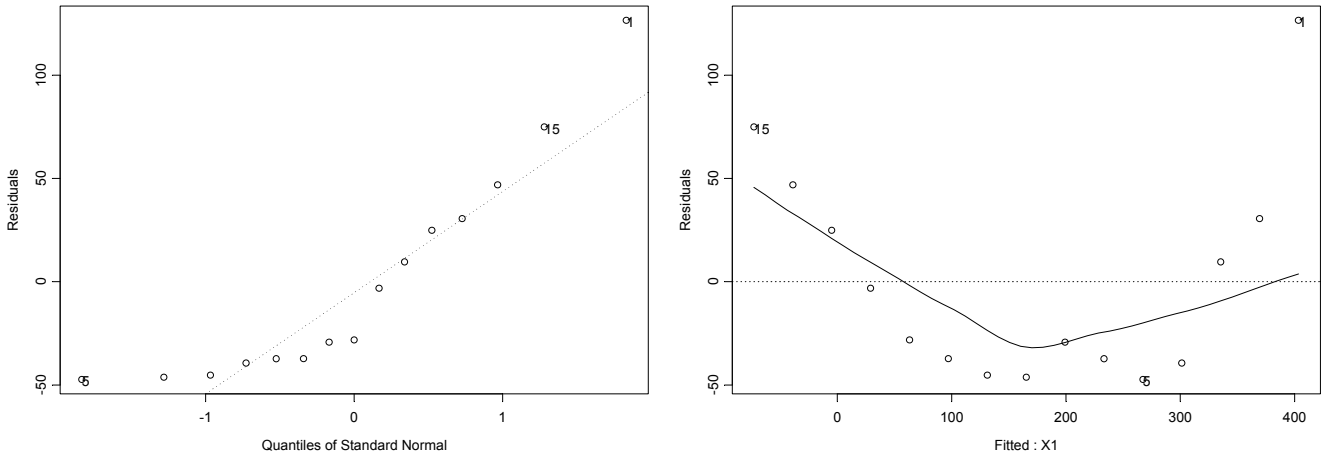
V dalším příkladě studujeme závislost počtu semen na vzdálenosti od mateřské rostliny.

vzdálenost; X1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
počet semen; Y2	530	400	345	262	220	196	170	119	86	60	35	26	20	8	2

Pokud si závislost vyneseme do grafu stejně jako v předchozím příkladě, pak získáme závislost, kdy odchylky od přímky se ke koncům zvyšují.



Spočteme-li lineární regresi pro tento vztah získáme tyto grafy reziduí.

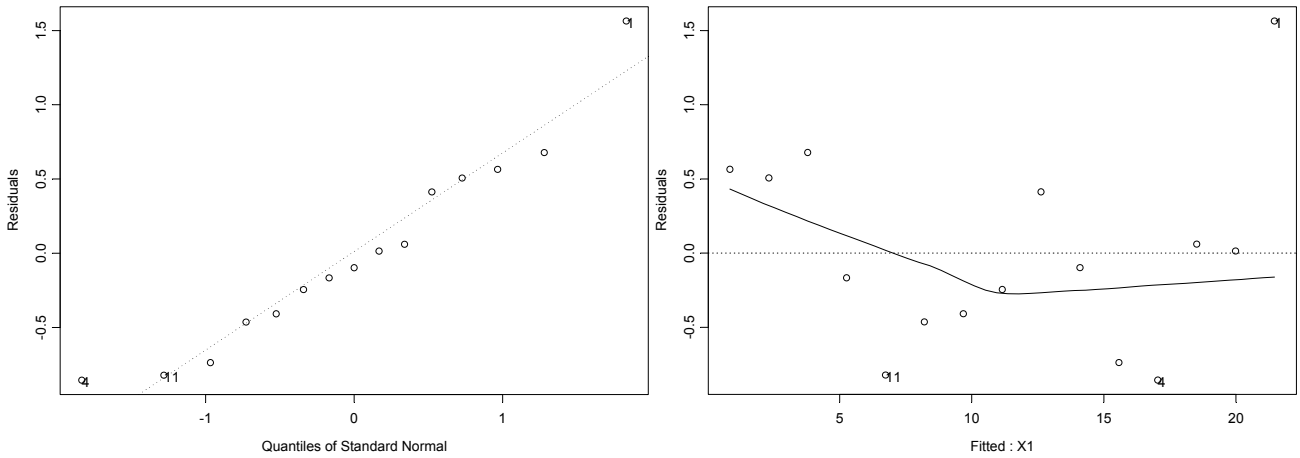


Analýza reziduálů nám ukazuje, že předpoklad lineární závislosti byl chybný. Jejich rozdělení není náhodné a jejich uspořádání vykazuje trend, kdy jejich spojnice tvoří parabolu. Přestože výsledky regresní analýzy i analýzy variance modelu jsou průkazné, použití lineárního modelu je v tomto případě chybné.

```
      *** Linear Model ***

Call: lm(formula = Y2 ~ X1, data = SDF3, na.action = na.exclude)
Residuals:
    Min      1Q  Median      3Q     Max
 -47.35  -38.35  -28.18   27.72  126.5

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  437.4952   29.3346    14.9139    0.0000
         X1  -34.0286    3.2264   -10.5470    0.0000

Residual standard error: 53.99 on 13 degrees of freedom
Multiple R-Squared: 0.8954
F-statistic: 111.2 on 1 and 13 degrees of freedom, the p-value is 9.666e-
008

Analysis of Variance Table

Response: Y2

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
       X1  1  324224.2 324224.2 111.2388 9.666232e-008
Residuals 13   37890.7   2914.7
```

V našem příkladě bychom spíše mohli očekávat, že s rostoucí vzdáleností od zdroje bude semen ubývat podle kvadratické závislosti. Můžeme buď použít kvadratický model a nebo data linearizovat a poté s takto upravenými daty provést lineární regresi. Na tomto příkladě použijeme transformaci dat.
Pro lineární regresi použijeme tedy jako závislou proměnnou odmocninu z Y2. Pro transformování proměnných můžeme použít nabídku v okně "Create Formula" a poté

"Transform". Výsledný vzorec bude tedy `sqrt(Y2) ~ X1`. To samé platí i pro zadání dat z příkazového řádku `model2<-lm(sqrt(Y2)~X1)`.
Jak je vidět z grafů reziduí, závislost již mnohem více odpovídá lineárnímu vztahu a její použití je nyní opravněné. Rezidua se vyskytují v "požadovaném" pasu okolo nulové hodnoty.



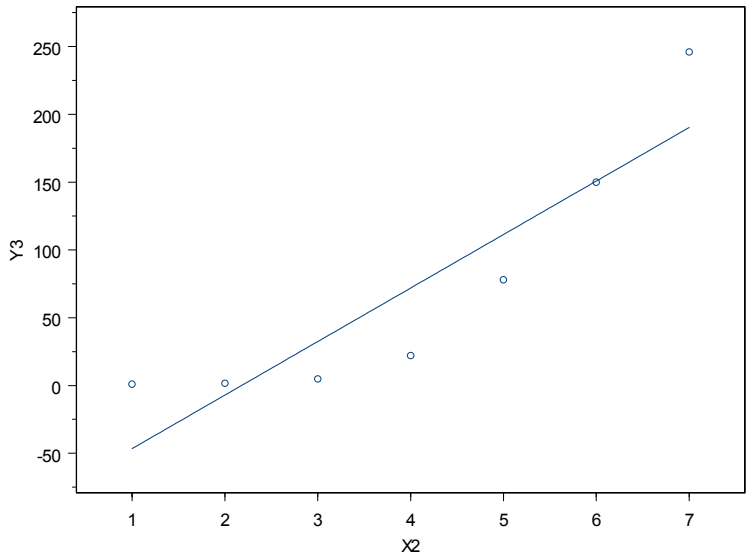
Regresní model má nyní podobu `y=22.9-1.47x`. Regresní koeficient je opět průkazně odlišný od nuly. Oproti netransformované závislosti se ale výrazně zvýšil koeficient determinace R-squared na 99 %.

```
      *** Linear Model ***

Call: lm(formula = sqrt(Y2) ~ X1, data = SDF3, na.action =
      na.exclude)
Residuals:
     Min      1Q   Median     3Q   Max
 -0.8557 -0.4363 -0.09811 0.4594 1.564

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  22.9302   0.3718    61.6816   0.0000
         X1  -1.4720   0.0409   -36.0018   0.0000

Residual standard error: 0.6842 on 13 degrees of freedom
Multiple R-Squared: 0.9901
F-statistic: 1296 on 1 and 13 degrees of freedom, the p-value is 2.087e-014

Analysis of Variance Table

Response: sqrt(Y2)

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
       X1  1  606.7102 606.7102 1296.132 2.087219e-014
Residuals 13    6.0852   0.4681
```

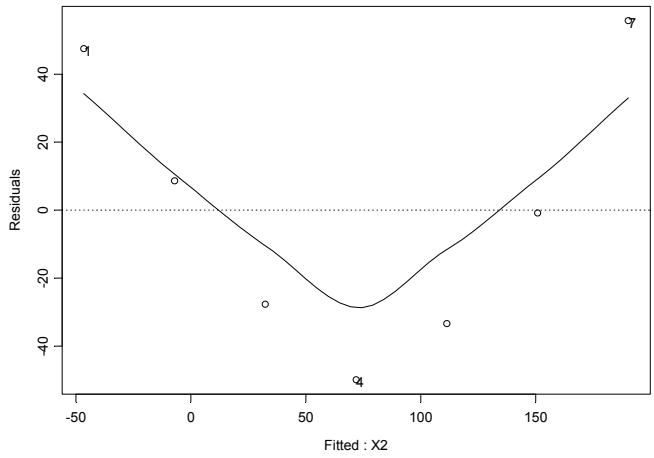
V tomto příkladu si ukážeme další linearizaci dat. Máme k dispozici data o hmotnosti jakési pěstované houby na Petriho misce v čase.

Počet dní; X2	1	2	3	4	5	6	7
Hnotnost; Y3	0.92	1.55	4.75	22	78	150	246

Zobrazením vztahu lehce zjistíme, že závislost mezi proměnnými není lineární.



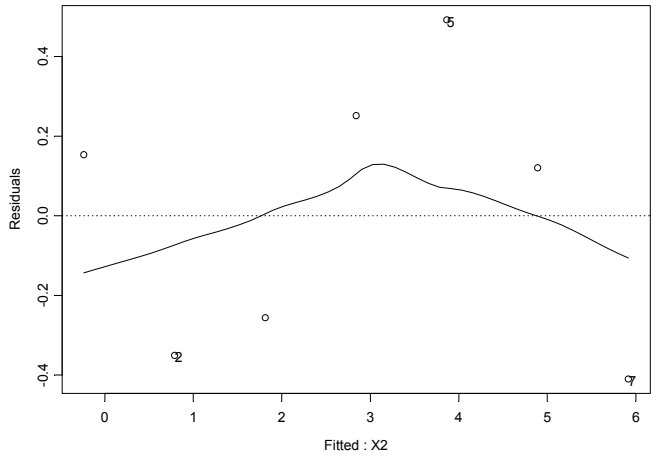
Graf reziduí ukazuje také na nelinearitu ve vztahu. Data tvoří jakési V a je tedy opět třeba tento vztah zlinearizovat.



Náš předchozí vztah předpokládal lineární závislost růstu kolonie houby na čase. Avšak ve skutečnosti je tento růst exponenciální (do té doby, než je dostupný prostor a živiny vyčerpán). Rovnice exponenciálního růstu je: $Y = e^{(a+bX)}$. Pomocí logaritmické transformace

tento vztah linearizujeme do formy *ln Y = a+bX*. Graf pro takto transformovanou proměnnou pak dobře odpovídá lineárnímu vztahu. Vytvoříme další sloupec pro transformovanou proměnnou Y3; z menu voláme Data>Transform..., vybereme zdrojovou proměnnou, typ transformace (log - logaritmus s přirozeným základem, sqrt - odmocnina). Ze získané "nové" proměnné vytvoříme graf a je patrné, že použitá transformace závislost zlinearizovala.

Nyní tedy můžeme spočítat lineární regresi s modelem buď použitím netransformované proměnné a zápisu `log(Y3) ~ X2` a nebo s transformovanou proměnnou jako závislou.



Rezidua jsou již mnohem lépe rozloženy a vzhledem k malému počtu měření, jsou ve výsledku vypsány rezidua pro všechny hodnoty.

```
      *** Linear Model ***

Call: lm(formula = log(Y3) ~ X2, data = SDF3, na.action = na.exclude
      )
Residuals:
      1       2       3      4      5      6       7
 0.1531 -0.3506 -0.2561 0.2515 0.4918 0.1204 -0.4102

Coefficients:
              Value Std. Error  t value Pr(>|t|)
(Intercept)  -1.2618   0.3162    -3.9906   0.0104
         X2   1.0253   0.0707    14.5019   0.0000

Residual standard error: 0.3741 on 5 degrees of freedom
Multiple R-Squared: 0.9768
F-statistic: 210.3 on 1 and 5 degrees of freedom, the p-value is 0.00002814

Analysis of Variance Table

Response: log(Y3)

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
       X2  1  29.43691 29.43691 210.3046 0.0000281389
Residuals  5   0.69986  0.13997
```

Procento vysvětlené variability modelem je téměř 98 % a celý model je výrazně signifikantní. Takže závěrem jsme mohli použít lineární regresi, alternativou (avšak náročnější) k tomuto postupu by bylo použití netransformovaných hodnot a nelineárního modelu.

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

Příloha: funkce pro práci s modely I.

`y~x`	model kde y je závislá proměnná a x nezávislá (model zahrnuje absolutní člen)
`y~x-1`	model bez absolutního členu
`lm`	fituje data lineární regresí s normálním rozložením chyb
`aov`	fituje analýzu variance
`glm`	fituje data zobecněným lineárním modelem (podrobnosti viz. manuály)
`abline`	vytvoří graf lineární závislosti; abline(pokus1), abline(mean(x),0)
`summary`	výstup souhrnu dle použitého modelu; např. summary (pokus1)
`plot`	diagnostické grafy
`update`	změna zadaného modelu
`coef`	vypíše odhady parametrů; coef (pokus1)
`fitted`	seznam fitovaných hodnot
`resid`	seznam reziduí

funkce lze kombinovat:
vytvoření histogramu reziduí modelu `hist(resid(pokus1)); hist(resid(lm(y~x)))`
vyjmutí "outlieru" (desátá hodnota) z analýzy `lm(y[-10]~x[-10])`