Regrese

28. listopadu 2013

Pokud chceme daty proložit vhodnou regresní křivku, musíme obvykle splnit tři úkoly:

1. Ukázat, že data jsou opravdu závislá.
2. Provést regresi.
3. Ukázat, že zvolená křivka je pro data opravdu vhodná.

Těmto třem úkolům se budeme postupně věnovat.

Korelační koeficient

Pro porozumění dalšímu textu musíme znát pojem „regresní koeficient", který udává sílu lineární závislosti mezi dvěmi veličinami.

Jak si příslušný vzorec snadno zapamatovat? Již na střední škole jsme se naučili vzorec pro výpočet úhlu mezi dvěma vektory:

$$\cos\alpha = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\cdot|\vec{v}|}.$$

Kosinus úhlu je jedna pro vektory ukazující stejným směrem, tj. v případě, že jeden vektor je kladným násobkem druhého.

Kosinus úhlu je mínus jedna pro vektory ukazující směrem opačným, tj. v případě, že jeden vektor je záporným násobkem druhého.

Tato skutečnost platí pro vektory o dvou, třech, nebo třeba tisíci položkách.

Nyní stačí vzít za vektor $\vec{u}$ data x, od nichž odečtu jejich střední hodnotu a podobně pro vektor $\vec{v}$. Tedy:

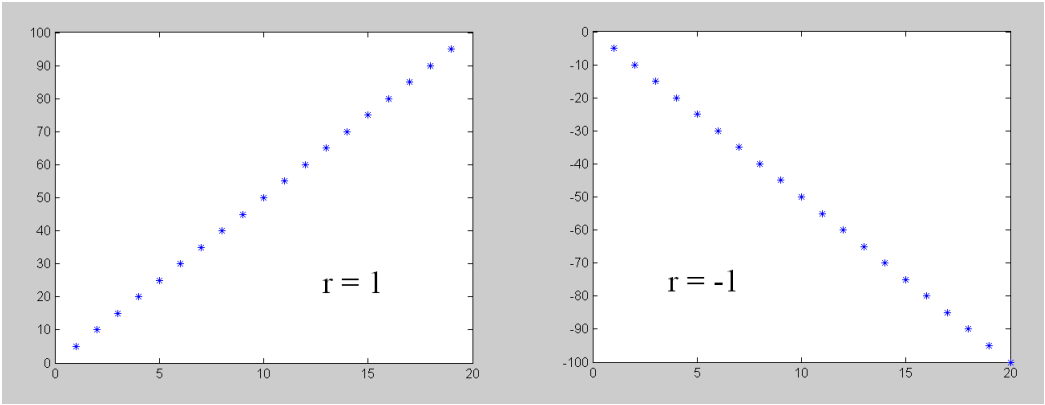
$$\begin{aligned}\vec{u} &= x-\overline{x},\\ \vec{v} &= y-\overline{y}.\end{aligned}$$

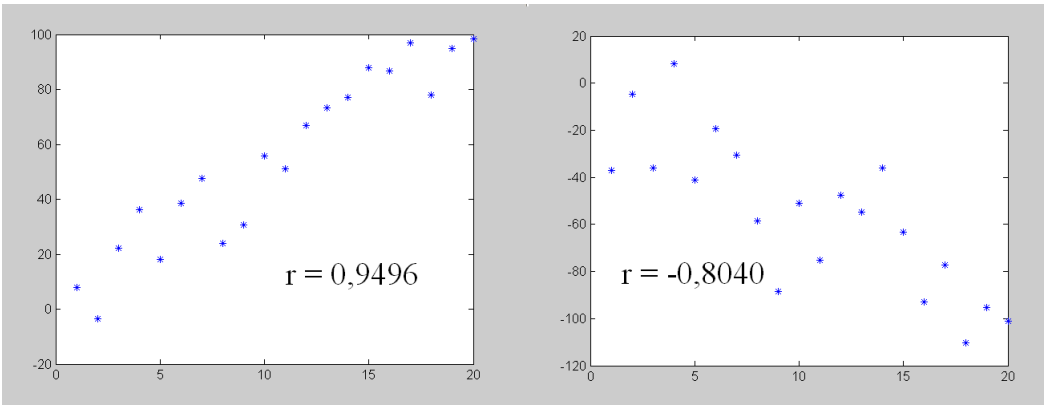
Za korelační koeficient r mezi veličinami x a y pak prohlásíme právě onen kosinus úhlu. Tedy:

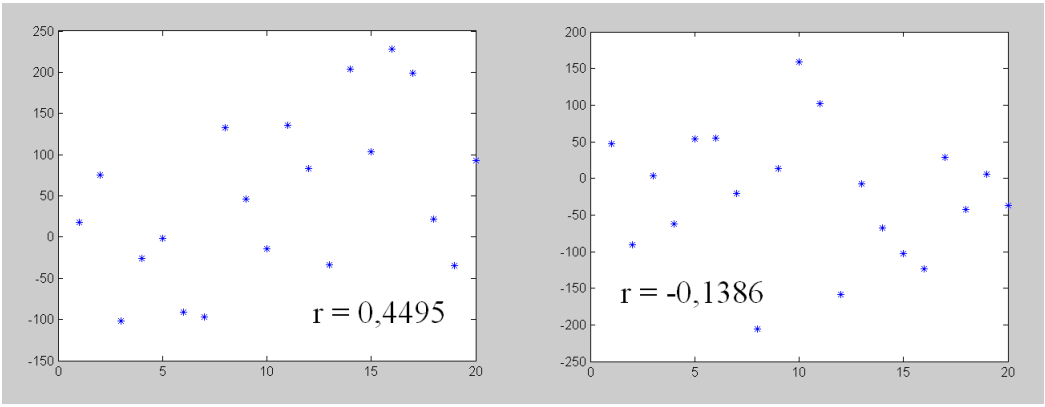
$$r = \frac{\vec{u}\cdot\vec{v}}{|\vec{u}|\cdot|\vec{v}|} = \frac{\sum(x-\overline{x})\cdot(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2}\cdot\sqrt{\sum(y-\overline{y})^2}}.$$

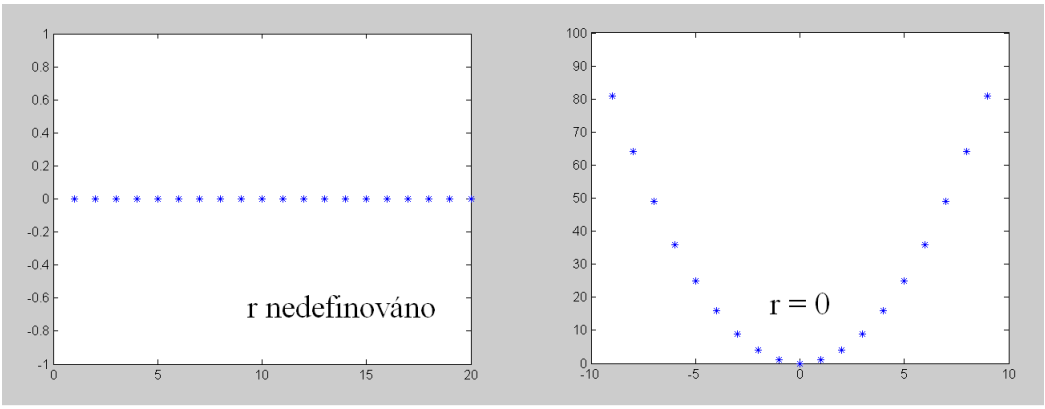
Korelační koeficient tedy nabývá hodnot mezi 1 a -1.

Na následujících obrázcích máme korelační koeficienty pro některé případy veličin x a y:


r = 1
r = -1


r = 0,9496
r = -0,8040


r = 0,4495
r = -0,1386


r nedefinováno
r = 0

Závislost či nezávislost dat

Prokládat daty jakoukoliv křivku má smysl pouze v případě, že jsou data závislá. My budeme daty prokládat přímku, proto budeme ověřovat závislost lineární.

O té nám mnoho říká již korelační koeficient, ten ale nezohledňuje množství dat. Je něco jiného $r = 0,5$ pro tři body a pro 1000 bodů.

Proto provedeme test (ne)závislosti dat. Testů, které se k tomuto používají, je celá řada, my se naučíme test Pearsonův.

Pearsonův test nezávislosti

Nulová hypotéza: Data x a y jsou lineárně nezávislá.

Směrování testu: Test je oboustranný.

Testová statistika:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}},$$

kde r je regresní koeficient a n počet změřených bodů..

Rozdělení statistiky: Studentovo s $n-2$ stupni volnosti - $St\,(n-2)$.

Příklad

Naměřili jsme tato data:

x	3	5	8	9	11	14
y	115	112	123	128	131	144

Jsou tato data lineárně závislá?

Provedeme Pearsonův test na hladině významnosti $\alpha = 5\%$.

Nejprve spočteme korelační koeficient: $r = 0,9619$.

Dosadíme do testové statistiky:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,9619}{\sqrt{\frac{1-0,9619^2}{6-2}}} = 7,0365.$$

V tabulkách Studentova rozdělené se 4 stupni volnosti najdeme kvantily, které odříznou 2,5% vlevo a 2,5% vpravo:

$t_{0,025}\,(4) = -2,776$,

$t_{0,975}\,(4) = 2,776$.

Obor přijetí je mezi těmito kvantily. Statistika ale padla do kritického oboru, proto nulovou hypotézu na hladině významnosti 5% zamítáme:

Data x a y jsou lineárně lineárně závislá a pravděpodobnost, že se v tomto mýlíme, je menší než 5%.

Příklad

Testujte stejné zadání pomocí Matlabu. Hladinu významnosti volte $\alpha = 5\%$.

Je potřeba mít nainstalovaný statistický balíček. Pokud jej nemáte, najděte si na stránkách Ivana Nagyho (http://www.fd.cvut.cz/personal/nagyivan/) v části PRP materiál Úvod do Matlabu a postupujte podle kapitolky 1.8.

Pomocí příkazu hlp, volba 5, zjistíme, že test nezávislosti má tvar: struc=cor_test(x,y,alt,method).

Dosadíme:

struc=cor_test([3,5,8,9,11,14],[115,112,123,128,131,144],'<>','p')

Zápis '<>' znamená, že se jedná o oboustranný test, volba 'p', že se jedná o Pearsonův test.

Funkce nám vypíše kromě jiného hodnotu statistiky $t = 7.0349$ (mírně odlišná díky našemu zaokrouhlování) a zejména p-hodnotu 0.0022.

P-hodnota je menší, než naše zvolená hladina významnosti 5%, proto nulovou hypotézu zamítáme:

Data x a y jsou lineárně lineárně závislá a pravděpodobnost, že se v tomto mýlíme, je 0,22%.

Poznámka

Tento závěr testu je pro nás žádoucí. Ověříme tím totiž na příslušné hladině významnosti, resp. s příslušnou p-hodnotou, že data jsou opravdu lineárně závislá a že má smysl pokračovat lineární regresí.

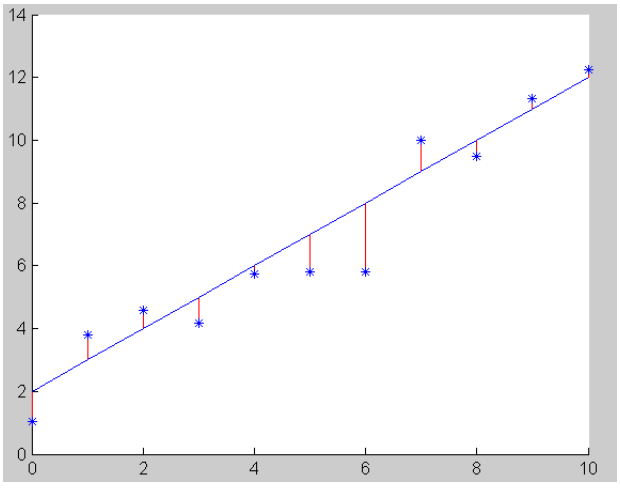
Metoda nejmenších čtverců

Dříve, než s našimi daty provedeme lineární regresi, ukážeme si, jak se příslušné vzorce odvozují metodou nejmenších čtverců. Užitečné je to v tom, že stejnou metodou můžeme odvodit vzorce pro výpočet parametrů pro libovolnou křivku. Např. parabolu, exponencielu, logaritmus, nějakou s-křivku, atd.

Myšlenka je následující: Body, které sedí na přímce, splňují rovnici: $y_n = A \cdot x_n + B$. Reálná data ale obvykle přesně na přímce nesedí, ale jsou rozmístěna v okolí přímky. To popíšeme tak, že y_n vzniká nejen vlivem x_n, ale i vlivem náhodné veličiny e_n:

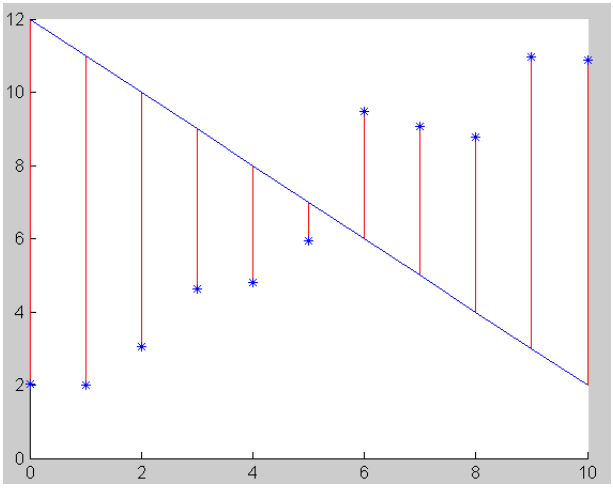
$$y_n = A \cdot x_n + B + e_n.$$

Náhodné odchylky e_n jsou na obrázku vyznačeny červeně:



Data (modré hvězdičky) známe, ale onu „správnou" přímku (modrá čára) neznáme. Z důvodů, které jsou skryty hlouběji v teorii, tipujeme, že ona „správná" přímka bude právě ta, pro niž bude součet čtverců odchylek e_n nejmenší. [1] Odtud název metody „metoda nejmenších čtverců".

O tom, že to není hloupá myšlenka, nás přesvědčí další obrázek, kde jsme se pokusili proložit daty zcela špatnou přímku a vidíme, že chyby e_n (a tudíž i jejich čtverce) jsou podstatně větší.



Nyní z principu nejmenších čtverců odvodíme vzorce pro parametry A a B regresní přímky $y_n = A \cdot x_n + B$.

Součet čtverců chyb e_n si označíme S:

$$S = \sum e_n^2.$$

Chyby e_n si můžeme vyjádřit za pomoci dat takto:

$$e_n = y_n - A \cdot x_n - B.$$

S je tedy funkcí dvou neznámých parametrů A a B. Hodnoty x_n a y_n známe z měření.

$$S(A, B) = \sum (y_n - A \cdot x_n - B)^2 .$$

S chceme mít co nejmenší. Minimum této funkce najdeme tak, že zjistíme, kde jsou obě parciální derivace nulové:

$$\begin{aligned} \frac{\partial S}{\partial A} &= 0, \\ \frac{\partial S}{\partial B} &= 0. \end{aligned}$$

Spočteme derivace:

$$\begin{aligned} \frac{\partial S}{\partial A} &= \sum 2 \cdot (y_n - A \cdot x_n - B) \cdot (-x_n) = \\ &= -2 \sum x_n \cdot y_n + 2A \sum x_n^2 + 2B \sum x_n = 0, \end{aligned}$$

[1] Nebudeme tu řešit, proč zrovna druhá mocnina a ne třeba čtvrtá či proč ne absolutní hodnota.

$$\begin{aligned}\frac{\partial S}{\partial B} &= \sum 2 \cdot (y_n - A \cdot x_n - B) \cdot (-1) = \\ &= -2 \sum y_n + 2A \sum x_n + 2B \sum 1 = \\ &= -2 \sum y_n + 2A \sum x_n + 2Bn = 0.\end{aligned}$$

Máme tedy dvě rovnice pro dvě neznámé:

$$\begin{aligned}A \sum x_n^2 + B \sum x_n &= \sum x_n \cdot y_n, \\ A \sum x_n + Bn &= \sum y_n.\end{aligned}$$

Odtud již snadno vyjádříme A i B:

$$\begin{aligned}A &= \frac{\sum x_n \sum y_n - n \sum x_n \cdot y_n}{\left(\sum x_n\right)^2 - n \sum x_n^2}, \\ B &= \frac{\sum x_n \sum x_n \cdot y_n - \sum x_n^2 \sum y_n}{\left(\sum x_n\right)^2 - n \sum x_n^2}.\end{aligned}$$

Příklad

Metodou nejmenších čtverců odvoďte vzorec pro bodový odhad pro parametr A, pokud víte, že veličiny x a y spolu souvisí přímou úměrností $y_n = A \cdot x_n + e_n$.

V tomto případě víme, že přímka musí procházet nulou. Nemůžeme tedy použít vzorec odvozený výše, kde je B obecné, ale musíme si vzorec odvodit:

Součet čtverců chyb e_n si označíme S:

$$S = \sum e_n^2.$$

Chyby e_n si můžeme vyjádřit za pomoci dat takto:

$$e_n = y_n - A \cdot x_n.$$

S je tedy funkcí neznámého parametru A. Hodnoty x_n a y_n známe z měření.

$$S(A) = \sum \left(y_n - A \cdot x_n\right)^2.$$

S chceme mít co nejmenší. Minimum této funkce najdeme tak, že zjistíme, kde je derivace nulová: [2]

$$\frac{dS}{dA} = 0.$$

Spočteme derivaci:

$$\begin{aligned}\frac{dS}{dA} &= \sum 2\left(y_n - A \cdot x_n\right) \cdot (-x_n) = \\ &= \sum \left(-2x_n y_n + 2Ax_n^2\right) =\end{aligned}$$

[2] Pokud máme více parametrů, používáme parciální derivace. Pokud je parametr jediný, použijeme derivaci obyčejnou.

$$= \quad -2\sum x_n y_n + 2A\sum x_n^2 = 0.$$

Vyjádříme A:

$$A = \frac{\sum x_n y_n}{\sum x_n^2}.$$

Lineární regrese

Nyní již máme vše připraveno. Ověřili jsme, že data jsou opravdu lineárně závislá, rozvážíme si, zda hledaná lineární závislost je přímá úměra $y_n = A \cdot x_n$ nebo obecná přímka $y_n = A \cdot x_n + B$ a dosadíme do příslušných vzorců.

Vzorce, které jsme zde uvedli, jsou pro bodový odhad. To znamená, že nám dají tu nejpravděpodobnější hodnotu parametrů s tím, že ty skutečně správné hodnoty jsou nejspíše trošku jinde.

Bodovým odhadem můžeme odhadovat i výstupní veličinu y. To uděláme velmi jednoduše tak, že dosadíme příslušné x do rovnice se spočtenými parametry, ale bez šumu.

Kromě bodových odhadů můžeme mít i odhady intervalové, kdy pro parametry i výstup získáme příslušné intervaly spolehlivosti. Vzorce pro intervalový odhad ani jejich odvození zde uvádět nebudeme.

Příklad

Zjistěte bodové odhady parametrů přímky z následujících dat. Spočtěte i bodový odhad výstupu pro $x = 10$.

x	3	5	8	9	11	14
y	115	112	123	128	131	144

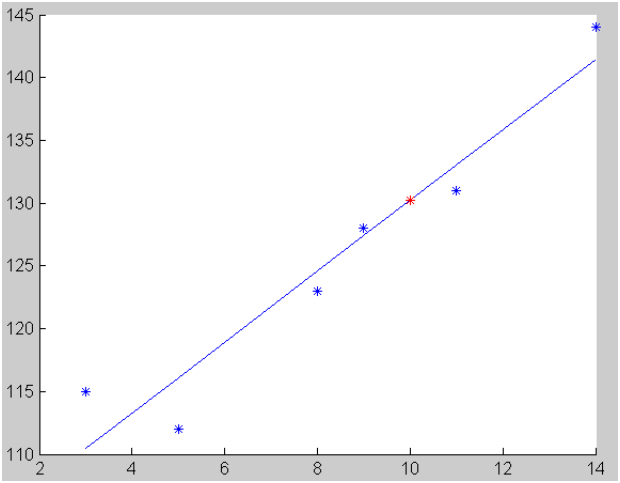
Pro tato data jsme již ověřili lineární závislost, můžeme proto rovnou přistoupit k lineární regresi. Jako model volíme obecnou přímku $y_n = A \cdot x_n + B + e_n$.

Dosadíme do vzorce:

$$\begin{aligned} A &= \frac{\sum x_n \sum y_n - n\sum x_n \cdot y_n}{\left(\sum x_n\right)^2 - n\sum x_n^2} = 2,8109, \\ B &= \frac{\sum x_n \sum x_n \cdot y_n - \sum x_n^2 \sum y_n}{\left(\sum x_n\right)^2 - n\sum x_n^2} = 102,0756. \end{aligned}$$

Tyto výsledky jsem si poctivě spočetl dle těchto vzorců v Matlabu, mohl jsem ale rovnou použít funkci ze statistického balíčku *lin_ reg(x,y)*. Bodový odhad výstupu pro $x = 10$ spočtu dosazením: $y = A \cdot x + B = A \cdot 10 + B = 130,1849$.

Výsledky i regresní přímku jsem vykreslil:

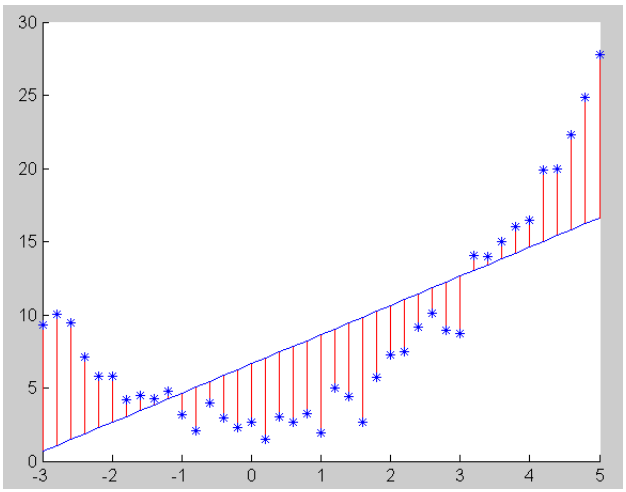


Ověření vhodnosti regresní křivky

Často jsme v situaci, kdy nevíme, jakou přesně křivkou máme data proložit. Pomoci nám může teorie, důležité je také data vykreslit a „vidět". Zda byla vybraná křivka opravdu vhodná, můžeme zjistit pohledem, nebo můžeme provést test pořadové nezávislosti pro rezidua (chyby) e_n.

Test pořadové nezávislosti reziduí

Následující obrázek nám ilustruje, že pokud křivku nezvolíme dobře, bude sérií reziduí se stejným znaménkem příliš málo a test pořadové nezávislosti neprojde.



V tomto případě, pro 41 dat, máme jen tři série znamének. Kladnou, zápornou a kladnou. Dosadíme do příslušné statistiky:

$$z = \frac{2b - (n-2)}{\sqrt{n-1}} = \frac{2 \cdot 3 - (41-2)}{\sqrt{41-1}} = -5,2178.$$

Test je levostranný, provedeme jej na hladině významnosti 1%. Příslušný kvantil najdeme v tabulce: $z_{0,01} = -z_{0,99} = -2,326$. Obor přijetí je tedy $\langle -2,326, \infty)$.

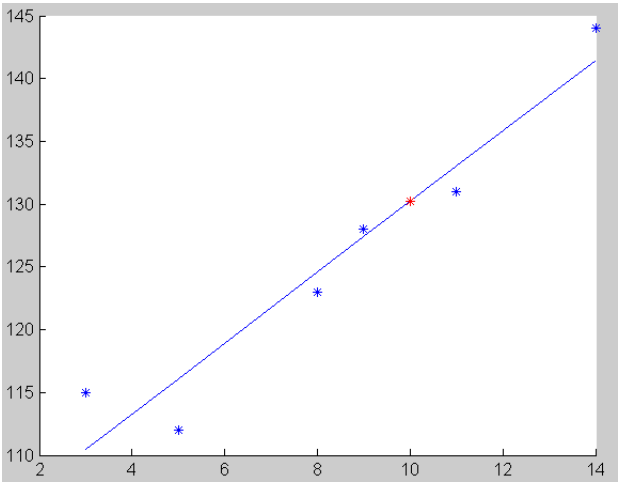
Statistika padla do kritického oboru, proto hypotézu, že data jsou pořadově nezávislá, respektive, že odchylky od přímky jsou pouze náhodné, respektive, že naše přímka je vhodnou regresní

křivkou, zamítáme. Pravděpodobnost, že se v tomto mýlíme, je menší než 1%.

Příklad

Ověřte, že přímka, kterou jste spočetli lineární regresí v minulém příkladě je pro data opravdu vhodnou křivkou.

Nejsnazší je vyjít z obrázku a vyznačit si znaménka reziduí:



Získáme tuto posloupnost znamének: +, -, -, +, -, +. Máme tedy pět sérií: $b = 5$. Počet dat je: $n = 6$. Dosadíme do příslušné statistiky pro test pořadové nezávislosti:

$$z = \frac{2b - (n-2)}{\sqrt{n-1}} = 2,6833.$$

Test je levostranný, provedeme jej na hladině významnosti 5%. Příslušný kvantil najdeme v tabulce: $z_{0,05} = -z_{0,95} = -1,645$. Obor přijetí je tedy $\langle -1,645, \infty)$.

Statistika padla do oboru přijetí, proto hypotézu, že data jsou pořadově nezávislá, respektive, že odchylky od přímky jsou pouze náhodné, respektive, že naše přímka je vhodnou regresní křivkou, nezamítáme.

Převod nelineární regrese na lineární

Často potřebujeme daty prokládat jinou křivku než přímku. Můžeme si pro ten či onen případ odvodit příslušné vzorce metodou nejmenších čtverců, často se ale volí jiný postup, a nelineární případ se převede na lineární:

Příkladem může být nepřímá úměrnost:

$$y = \frac{k}{x}.$$

Pokud místo dat x vezmu jejich převrácenou hodnotu $X = \frac{1}{x}$, získávám rovnici přímé úměrnosti:

$$y = k \cdot X.$$

Nyní mohu použít vzorec, který jsme si již odvodili.

Jiným příkladem může být exponenciála:

$$y = b \cdot e^{ax}.$$

Rovnici zlogaritmujeme:

$$\ln y = \ln b \cdot e^{ax} = \ln b + \ln e^{ax} = \ln b + ax \ln e = ax + \ln b.$$

Pokud místo hodnot y vezmu $Y = \ln y$ a konstantu $\ln b$ si označím B, dostávám obecnou lineární závislost

$$Y = ax + B.$$

Použiji standardní vzorce a spočtu konstanty a a B. Konstantu malé b spočtu z velkého B snadno: $\ln b = B \rightarrow b = e^B$.