Vícenásobná regresní a korelační analýza [1]

[1]Tyto materiály byly vytvořeny za pomoci grantu FRVŠ číslo 1145/2004.

O vícenásobné závislosti mluvíme tehdy, jestliže je závisle proměnná y závislá na více nezávislých proměnných $x_1, x_2, \ldots, x_p$. Řada pojmů a metod je analogická jako při prosté regresní analýze. Rozdíl je pouze v určitém zevšeobecnění na více proměnných.

Rozeznáváme pojmy jako mnohonásobná závislost, mnohonásobná nezávislost, korelační závislost, korelační nezávislost. Vysvětleme tyto pojmy.

Mnohonásobnou závislostí rozumíme, závislost, kdy dochází se změnou nezávisle proměnných ke změně podmíněného rozdělení závisle proměnné.

Mnohonásobnou nezávislostí je pak takový stav mezi proměnnými, kdy se změnou hodnot nezávisle proměnných nedochází ke změně podmíněného rozdělení četností závisle proměnné.

Za korelační závislost označíme takový stav, kdy se změnou hodnot nezávisle proměnných dochází ke změně podmíněných průměrů závisle proměnné.

Korelační nezávislostí naopak rozumíme takový stav, kdy se změnou nezávisle proměnných nedochází ke změně podmíněných průměrů závisle proměnné.

Nejjednodušším případem vícenásobné závislosti je trojnásobná lineární regrese. Ta je vyjádřená rovnicí roviny v trojrozměrném euklidovském prostoru, tedy $\hat{y} = b_0 + bx_1 + bx_2$.

Odhady jednotlivých regresních koeficientů, zde se přesněji nazývají parciální regresní koeficienty. Lze je získat opět prostřednictvím metody nejmenších čtverců. Při výkladu využijeme maticovou symboliku. Usnadní nám to náš postup a umožní nám to odvodit obecný postup, kterým můžeme odhadovat i regresní koeficienty u nelineárních funkcí v proměnných [1] :

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min \tag{1}$$

Dále předpokládejme, že $\hat{y}_i = b_0 + b_1 f_1 + b_2 f_2 + \cdots + b_p f_p$, kde symboly (regersory - nezávisle proměnné) f_i, $i = 1, 2, \ldots, p$ představují libovolné známé funkce vysvětlujících proměnných. Speciálním případem jsou modely typu:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \ , \tag{2}$$

kde jsou regresory přímo vysvětlující proměnné ($f_i = x_i$), tj. modely lineární z hlediska parametrů i z hlediska vysvětlujících proměnných. Model můžeme zapsat pro všech n pozorování pomocí maticové symboliky následovně:

$$\hat{\mathbf{y}} = \mathbf{F}\ \boldsymbol{b} \tag{3}$$

[1] ne však odhadovat regresní koeficienty u funkcí nelineárních v parametrech. Řešením těchto úloh se zabývat nebudeme.

kde:

$$\mathbf{F} = \begin{bmatrix} 1 & f_{11} & \cdots & f_{1p} \\ 1 & f_{21} & \cdots & f_{2p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & f_{n1} & \cdots & f_{np} \end{bmatrix}, \tag{4}$$

Vektor $\boldsymbol{b}^T$ pak představuje vektor hledaných regresních regresních koeficientů, tedy:

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \tag{5}$$

Pokud hodláme minimalizovat funkci S, pak musíme tuto funkci derivovat a takto upravenou funkci položit rovno nule. Tím je splněn nutný předpoklad. Odhad jednotlivých složek vektoru $\mathbf{b}$ tj. odhad skutečných regresních koeficientů β_i, $i = 1, 2, \cdots, p$ získáme takto:

$$\begin{aligned} S = \textstyle\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})^2 = \epsilon^T\epsilon = (\mathbf{y} - \mathbf{F}\boldsymbol{b})^T(\mathbf{y} - \mathbf{F}\boldsymbol{b}) \\ = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{F}\boldsymbol{b} - (\mathbf{F}\boldsymbol{b})^T\mathbf{y} + (\mathbf{F}\boldsymbol{b})^T\mathbf{F}\boldsymbol{b} \\ = \mathbf{y}^T\mathbf{y} - 2(\mathbf{F}\boldsymbol{b})^T\mathbf{y} + \boldsymbol{b}^T\mathbf{F}^T\mathbf{F}\boldsymbol{b} \end{aligned}$$

Vzhledem k tomu, že se snažíme o minimalizaci celkového součtu čtverců reziduí, budeme derivovat funkci S dle vektoru regresních koeficientů a řešit následující rovnici:

$$\frac{\partial S}{\partial \boldsymbol{b}} = -2\mathbf{F}^T\mathbf{y} + 2\mathbf{F}^T\mathbf{F}\boldsymbol{b} = 0$$

Lze tedy psát

$$\begin{aligned} 2\mathbf{F}^T\mathbf{F}\boldsymbol{b} = 2\mathbf{F}^T\mathbf{y} \qquad |(\mathbf{F^T F})^{-\mathbf{I}} \\ \boldsymbol{b} = (\mathbf{F}^T\mathbf{F})^{-I}\mathbf{F}^T\mathbf{y}\ , \end{aligned}$$

kde symbol $\boldsymbol{b}$ je vektorem odhadnutých regresních parametrů.

Galileův pokus aneb jednoduchý příklad regresní analýzy na závěr?

Galileo se zabýval studiem pohybu tělesa. K tomuto studiu si sestrojil jednoduché zařízení. Na stůl umístnil nakloněnou rovinu s drážkou. Pokus spočíval v opakovaném vypouštění bronzové koule v jisté výšce, označme tuto výšku jako x a měřil vzdálenost dopadu stříbrné koule od hrany stolu. Výška stolu Galileova stolu činila 500 punti. Galileo naměřil tato data [punti[2]]:

```
      x   y
[1,]  100 253
[2,]  200 337
[3,]  300 395
[4,]  450 451
```

[2] Jedno punti je rovno 169/180 mm

```
[5,]  600 495
[6,]  800 534
[7,] 1000 573
```

Pokusíme se proložit Galileova data prostým regresním modelem, tedy přímkou. Uvažujme nejprve přímku procházející počátkem soustavy souřadnic. Uvažujme tedy model

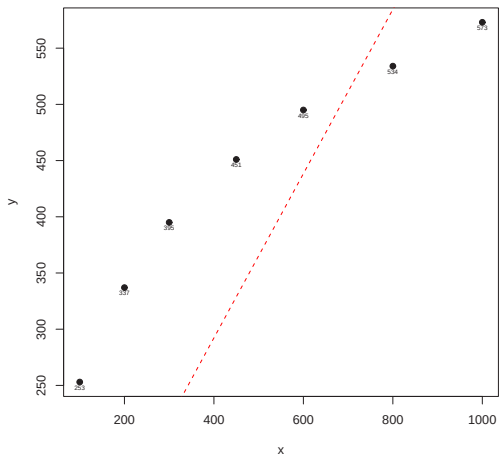
$$y_i = \beta_1 x_i + \varepsilon_i \tag{6}$$

Pomocí metody nejmenších čtverců odhadneme regresní koeficient β_1. Pomocí softwarového prostředí `R` získáme tyto výsledky:

```
Call: lm(formula = y ~ x - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-157.566    3.104  122.245  177.887  190.887

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   0.7306     0.1017   7.186 0.000367 ***
---
Signif. codes:  0 ,,***, 0.001 ,,**, 0.01 ,,*, 0.05 ,,., 0.1 ,, , 1

Residual standard error: 155.6 on 6 degrees of freedom Multiple
R-Squared: 0.8959,     Adjusted R-squared: 0.8786 F-statistic:
51.64 on 1 and 6 DF,  p-value: 0.0003671
```

Pokud regresní rovnici znázorníme získáme tento výstup

Obrázek 1: Prostá lineární regrese bez absolutního členu



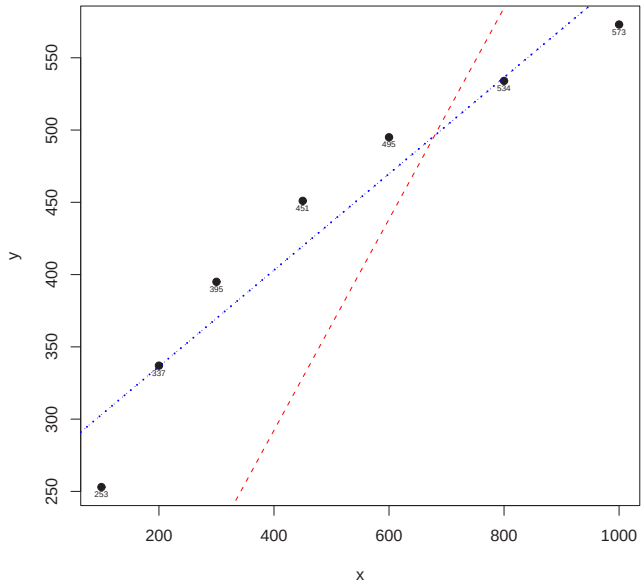
Vidíme, že regresní rovnice nevystihuje dobře empirická data. Index determinace činí pouze 87,87 % a reziduální součet čtverců činí 145268,16. Přistoupíme proto k jinému regresnímu modelu který zahrnuje i absolutní člen tj. využijeme modelu

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{7}$$

Výsledky regresní analýzy pro druhý model

```
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5        6        7
-50.0462   0.6201  25.2864  31.2859  25.2853  -2.3821 -30.0495

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 269.71246   24.31239  11.094 0.000104 ***
x             0.33334    0.04203   7.931 0.000513 ***
---
Signif. codes:  0 ,,***, 0.001 ,,**, 0.01 ,,*, 0.05 ,,., 0.1 ,, , 1

Residual standard error: 33.68 on 5 degrees of freedom
Multiple R-Squared: 0.9264,     Adjusted R-squared: 0.9116
F-statistic: 62.91 on 1 and 5 DF,  p-value: 0.0005132
```

Obrázek 2: Prostá lineární regrese s absolutním členem



Vidíme, že ani tato regresní funkce nepopisuje data příliš dobře, třebaže index determinace činí 92,64 % a došlo k dosti podstatnému snížení hodnoty reziduálního součtu čtverců. Jeho hodnota činí 5671,712.
Přistoupíme k dalšímu regresnímu modelu. Vzhledem k hodnotám by mohl být adekvátním modelem kvadratický regresní model

$$\hat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \tag{8}$$

Výsledky regresní analýzy pro kvadratický regresní model

```
Call:
lm(formula = y ~ x + I(x^2))

Residuals:
      1       2       3       4       5       6       7
-14.308   9.170  13.523   1.940  -6.177 -12.607   8.458

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.999e+02  1.676e+01  11.928 0.000283 ***
x            7.083e-01  7.482e-02   9.467 0.000695 ***
I(x^2)      -3.437e-04  6.678e-05  -5.147 0.006760 **
---
Signif. codes:  0 ,,***, 0.001 ,,**, 0.01 ,,*, 0.05 ,,., 0.1 ,, , 1

Residual standard error: 13.64 on 4 degrees of freedom
Multiple R-Squared: 0.9903,     Adjusted R-squared: 0.9855
F-statistic:   205 on 2 and 4 DF,  p-value: 9.333e-05
```

Z grafu kvadratické regresní funkce vidíme, že kvadratická funkce s regresními parametry výborně vystihuje Galileova data. Index determinace dosáhl dokonce hodnoty 98,55 %, což znamená, že náš model vysvětluje 98,55 % rozptýlenosti naměřených vodorovných vzdáleností. Reziduální součet čtverců činí 744,1984. Tento výsledek lze považovat za velmi dobrý. Všiměte si, že jsou signifikantní všechny regresní koeficienty a to dokonce na hladině $\alpha = 0.01$

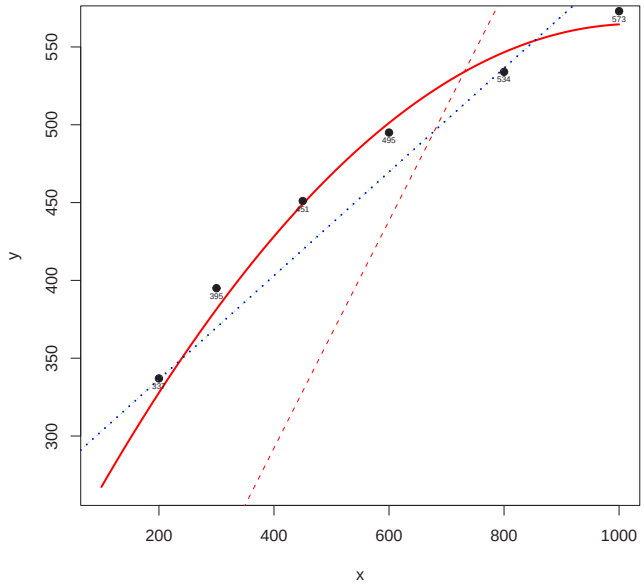
Pokusme se přidat ještě kubický člen. Bude popisovat odhadnutá regresní funkce data lépe? Model obecně zapíšeme takto:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \ . \tag{9}$$

Výsledky regresní analýzy pro případ polynomu třetího stupně jsou uvedeny níže. Všiměte si, že i kubický člen je statisticky významný:

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3))

Residuals:
       1        2        3        4        5        6        7
-2.40359  3.58091  1.89175 -4.46885 -0.08044  2.32159 -0.84138
```

Obrázek 3: Kvadratická regrese funkce s absolutním členem



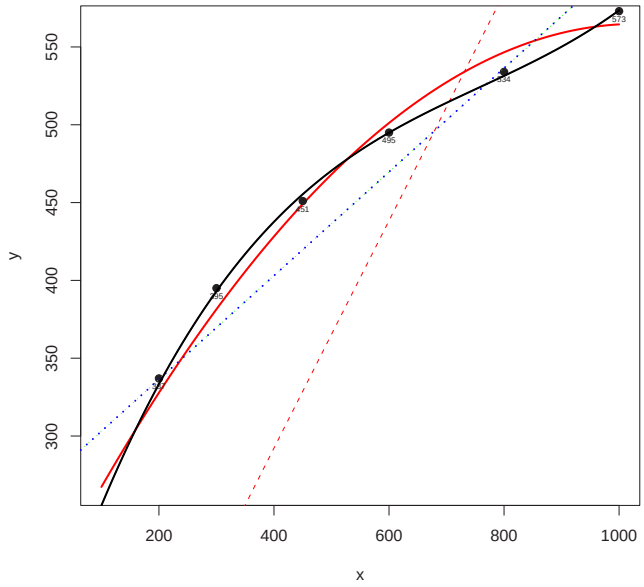

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.558e+02  8.326e+00  18.710 0.000333 ***
x            1.115e+00  6.567e-02  16.983 0.000445 ***
I(x^2)      -1.245e-03  1.384e-04  -8.994 0.002902 **
I(x^3)       5.477e-07  8.327e-08   6.577 0.007150 **
---
Signif. codes:  0 ,,***, 0.001 ,,**, 0.01 ,,*, 0.05 ,,., 0.1 ,, , 1

Residual standard error: 4.011 on 3 degrees of freedom
Multiple R-Squared: 0.9994,     Adjusted R-squared: 0.9987
F-statistic:  1595 on 3 and 3 DF,  p-value: 2.662e-05
```

Z výsledků je patrné, že je tento model ještě lepším než předchozí kvadratický. Index determinace činí 99,94 %, reziduální součet čtverců činí $4.011^2.3 = 48.264363$. Prostě paráda co více si přát. Pro úplnost znázorníme ještě průběh funkce spolu s empirickými daty. Jistě bychom byli spokojeni. Reziduální suma čtverců je relativně malá, index determinace činí 99,94 %. Spokojeni však být nemůžeme,

neboť je to celé úplně špatně!!

Proč? Jistě jste si všimli(viz obrázek 4), že graf průběhu polynomické regrese vykazuje ve své "horní části" inflexi, která z hlediska přírodních zákonů, fyziky

Obrázek 4: Funkce polynomické regrese 3 stupně s absolutním členem



nemá opodstatnění. Model je tedy zcela chybný. Z fyzikálního hlediska by byla jediným správným modelem funkce popisující zákony pohybu po nakloněné rovině a šikmého vrhu mající tvar:

$$y_i = \sqrt{x_i^2 \sin^2\alpha + 4d \cdot x_i \cos^2\alpha} - x_i \sin 2\alpha \tag{10}$$

Symbol α představuje úhel nakloněné roviny po které byla vypouštěna koule, symbol d pak výšku stolu.

Pokusme se tedy dospět k výsledku jinou cestou. Víme, že Galileův stůl měl výšku 500 punti, po dosazení se správná regresní rovnice (10) zjednoduší:

$$y_i = \sqrt{x_i^2 \sin^2\alpha + 2000 \cdot x_i \cos^2\alpha} - x_i \sin 2\alpha \ . \tag{11}$$

Pomocí software `R` a při využití Gauss-Newtonova algoritmu se pokusíme získat odhad neznámého parametru α. Ten představuje, jak jistě víte úhel, který svírala nakloněná rovina s deskou stolu. Řešení je tedy následující:

```
nls(y~sqrt(x^2*(sin(2*a))^2+4*500*x*(cos(a))^2)-x*sin(2*a),
start=c(a=0.5203),trace=TRUE)

31357.71 :  0.5203
2576.557 :  0.6101944
2485.28 :  0.6157443
2485.263 :  0.6158206
```

```
2485.263 :  0.6158214
Nonlinear regression model
  model:  y ~ sqrt(x^2 * (sin(2 * a))^2 +
  4 * 500 * x * (cos(a))^2) - x * sin(2 * a)
   data:  parent.frame()
        a
0.6158214
 residual sum-of-squares:  2485.263
```

Řešením jsme získali odhad $\hat{\alpha} = 0,6158214, tj. \doteq 35,3°$. Dále můžeme odečíst reziduální sumu čtverců, dosahuje hodnoty 2485,263.

I když je to dobře, můžete si všimnout toho, že jsme v případě správného regresního modelu obdrželi větší reziduální součet čtverců, než v předchozích evidentně chybných modelech! Graf této funkce spolu s ostatními regresními modely je uveden zde:

Obrázek 5: Graf správné regresní funkce