Regrese

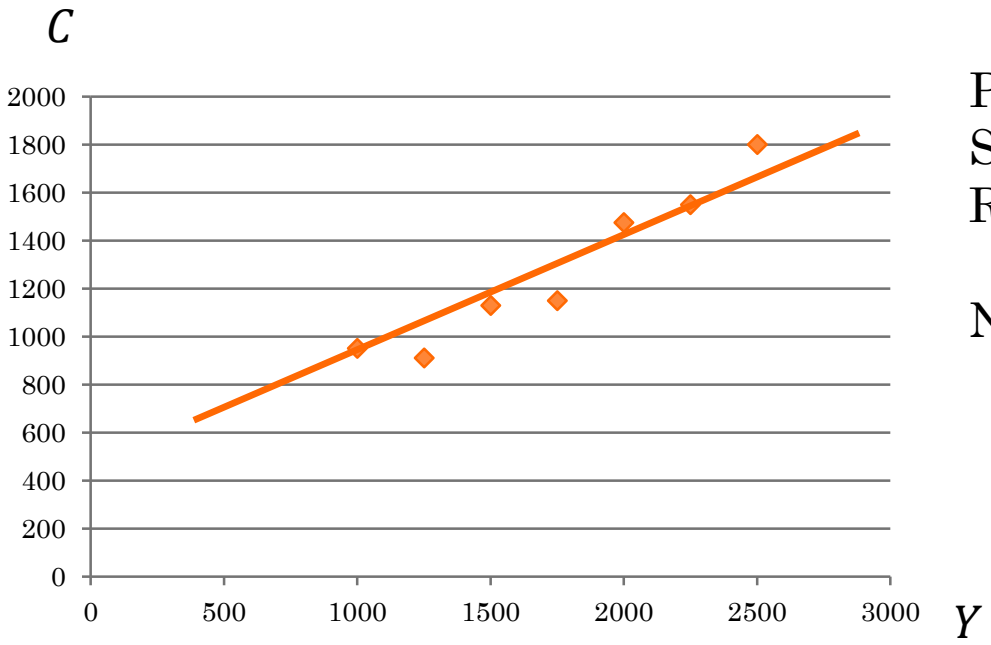
K čemu slouží regrese?

$$C = Ca + c.Y$$

$$C = 200 + 0{,}6.Y + e$$

C	Y
950	1000
910	1250
1130	1500
1150	1750
1475	2000
1550	2250
1800	2500

Budeme zjišťovat jak jedna proměnná (nezávislá)
Ovlivňuje jinou proměnnou (závislou)



Pozor na aplikaci regrese!!!
Striktní podmínky
Různé metody

Např. problém kauzality vztahů

$$Y = C + I + G + NX$$

Úvod

Pokoušíme se zjistit **příčinné/kauzalní** souvislosti

Spotřebu ovlivňuje velikost důchodu $C = Ca + c.Y$
Investice ovlivňuje velikost úrokové míry $I = Ia - bi$
Export ovlivňuje reálný měnový kurz a zahraniční HDP

Nejsou vztahy „vycucané“ z prstů

Chceme zjistit zda-li mezi proměnnými existují konkrétní vztahy
Například jak proměnná/proměnné (i, Y, R..)
Ovlivňuje jinou proměnnou (C,I,EX)
Detailně pochopit vztahy mezi nezávislou/mi a závislou proměnnou
A pokud možno vše popsat matematickou funkcí

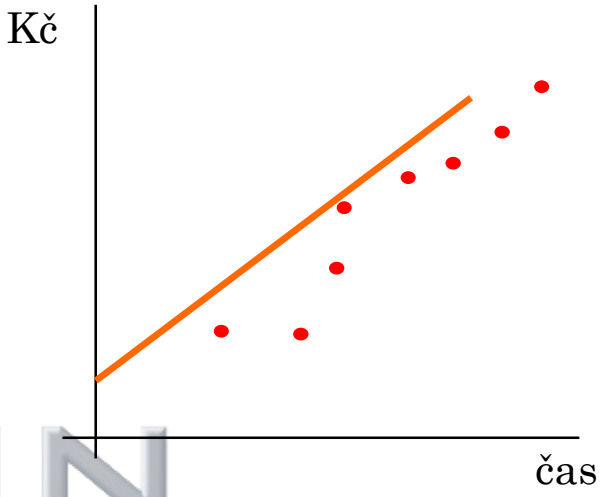
$$Q_x=20-0,54P_x+0,12P_y+0,2Y$$

Jsme schopni „dobře“ určit některé proměnné (příjem, hodnota majetku atd.)
Jak ale určit zda-li půjčit/nepůjčit peníze?
A které proměnné nejvíce ovlivní bankrot klienta?

Deterministický model

Jednoznačně existující vztah
Pravděpodobnost =1
Spoření (fixní sazba, poplatky)

$$y = \beta_0 + \beta_1 . x_1$$



y- závislá proměnná (vysvětlovaná proměnná)
x- nezávislá proměnná (vysvětlující proměnná)
ß- parametry ($ß_0$ absolutní člen, $ß_1$ sklon)

Stochastický model

Do modelu vstupuje nejistota (další neuvažované vlivy)
Například i chyby v měření

$$y = \beta_0 + \beta_1 . x_1 + \varepsilon$$

ε- náhodná chyba (náhodná veličina proto má pravděpodobnostní rozdělení)

Jednostranná závislost – **regresní analýzy**
Vzájemná závislost (**lineární**) – **korelační analýza**

$$C = Ca + c.Y$$

Cíl – snaha poznat a popsat příčinné vztahy mezi proměnnými

Výnos pole a množství hnojiva
Uvažujeme existenci lineárního vztahu (úvaha zemědělců)
– více hnojiva větší výnos
Jak ověřit tento vztah?
Dotážeme se všech zemědělců v ČR?

$$y = \beta_0 + \beta_1 . x_1$$

Získáme statistický soubor

- Pozorováním (n) statistických jednotek (sledujeme 100 zemědělců)
  snaha aby daty byla prostorově, časově a věcně vymezena

- Pozorováním určité statistické jednotky (HDP) v (n) časových intervalech

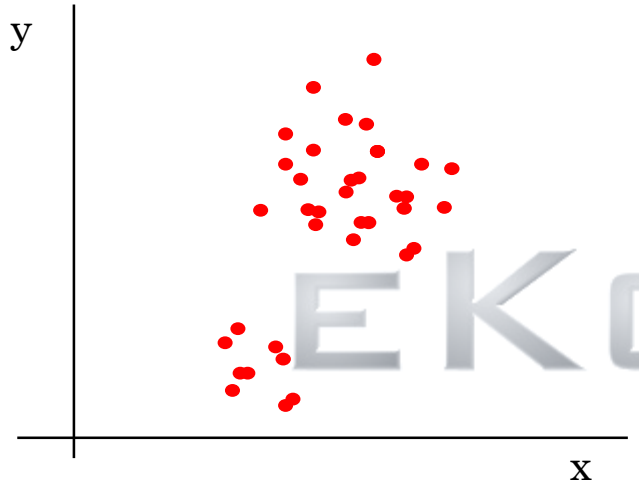
Snaha se co nejvíce přiblížit(aproximovat) empirickou regresní funkci
A hypotetickou regresní funkci
Co nejlépe by měla vyjadřovat charakter závislosti (lineární, logaritmická atd.)

Hledáme průběh závislosti (lineární, nelineární)
Intenzitu závislosti (silná/těsná)

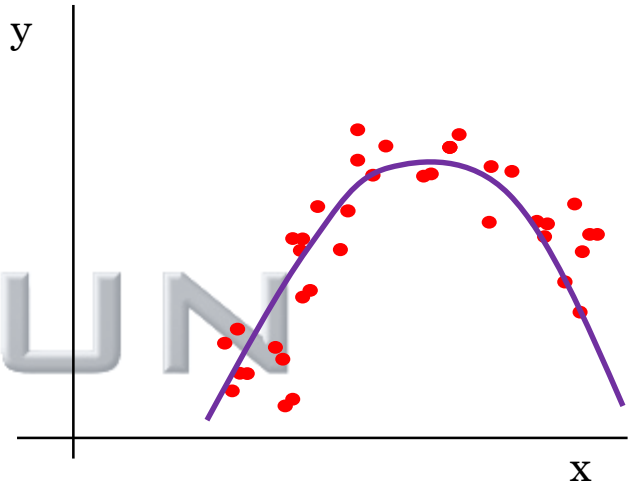
Snaha se co nejvíce přiblížit(aproximovat) empirickou regresní funkci
A hypotetickou regresní funkci
Co nejlépe by měla vyjadřovat charakter závislosti (lineární, logaritmická atd.)
Hledáme průběh závislosti (lineární, nelineární)
Intenzitu závislosti (silná/těsná)

Závislost a její intenzita

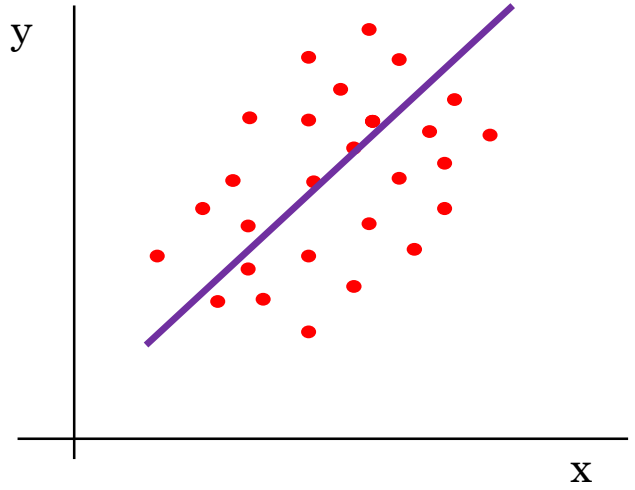
Nelineární



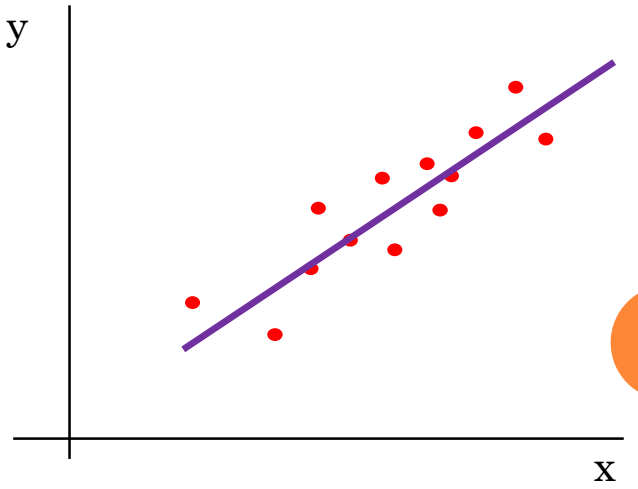
Nelineární závislost – silná



Lineární závislost – slabá



Lineární závislost – silná



Příklad

$$y = \beta_0 + \beta_1 . x_1 + \varepsilon$$

Máme pole a chceme zjistit co ovlivňuje výnos z pole
Myšlenka množství hnojiva

$$y = \beta_0 + \beta_1 . x_1$$

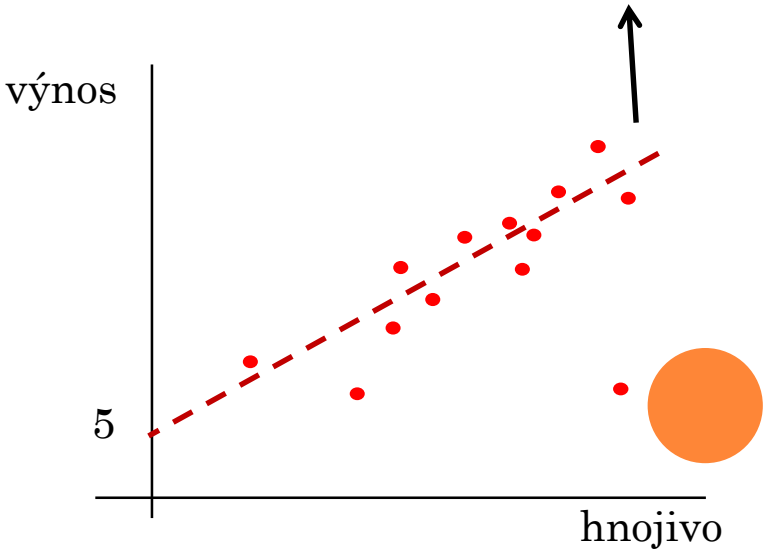
$$výnos = \beta_0 + \beta_1 . \text{hnojivo} + \varepsilon$$

výnos- závislá proměnná
hnojivo – množství hnojiva nezávislá proměnná
ε- ostatní faktory

Provedeme (n) náhodných výběrů – oslovíme n zemědělců
A zjistíme kolik hnojili a jaký měli výnos

$$výnos = 5 + 1{,}5 . \text{hnojivo} + e$$

Sklon 1,5



Když nebudeme hnojit výnos=*5*
Když se změní množství hnojiva o 1
Zvýší se výnos o *1,5.1=1,5*
Změna hnojiva o 2 – výnos=2.1,5=3

e- body neleží na čárkované přímce
Existují další faktory kromě hnojiva
Ovlivňující výnos

Jednoduchý lineární regresní model

Máme pouze jednu nezávisle proměnnou
Vztah mezi závisle proměnnou (y) a nezávisle proměnnou (x) je lineární

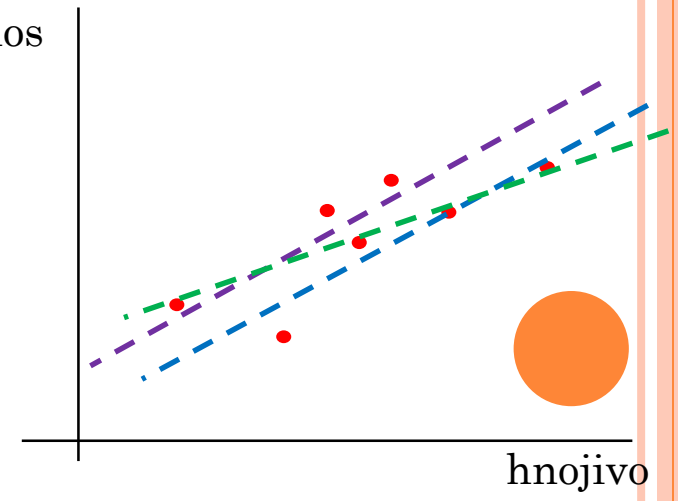
My získáme „nějaká“ data y a x (empirické/výběrové hodnoty) – co se naměřilo
Cílem je najít případný vztah mezi y a x a popsat jej

Výnos pole a množství hnojiva
My víme, že zde existuje lineární vztah – čím více hnojiva – tím větší výnos
Ale nevíme, jak přesně má daný vztah vypadat

Teoretická (hypotetická) regresní funkce – nepozorovatelná (η)
„ideální“ regresní funkce
Teoretický vztah – většinou neznáme ☺

$$y = \beta_0 + \beta_1 . x_1$$

Empirická regresní funkce je
Odhad teoretické regresní funkce



Teoretická a empirická regresní funkce

Pro každé pozorování (i) i=1,2…

$$y_i = \eta_i + \varepsilon_i \qquad \eta_i = \beta_0 + \beta_1 . x_i$$

$$y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i$$

Při neexistenci chyby (ε)
Model deterministický (pevná závislost)
η- předpis kdy *x* je přiřazeno *y* „přesně“
y=2.x

y_i- *i-tá* empirická hodnota vysvětlované proměnné (výnos pole)
η_i- *i-tá* hodnota teoretické regresní funkce (neznám)
ε_i- odchylka (náhodná chyba) y_i od η_i

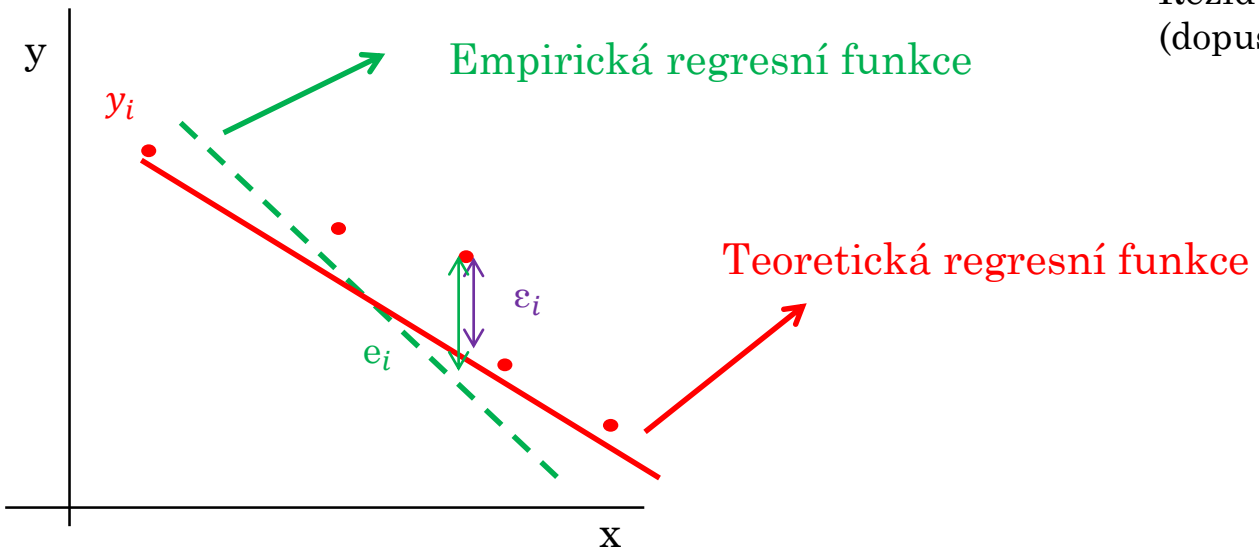
Odchylka

e_i-reziduum – rozdíl mezi empirickou regresní funkcí a empirickou hodnotou

Na *y* působí další náhodné proměnné než pouze (*x*)
Na pozorování působí náhodné chyby (nepřesné váhy)

$\varepsilon_i \neq e_i$

Reziduum je odhadem náhodné chyby (dopustili jsme se dalších chyb)



Hledání konkrétního tvaru regresní funkce

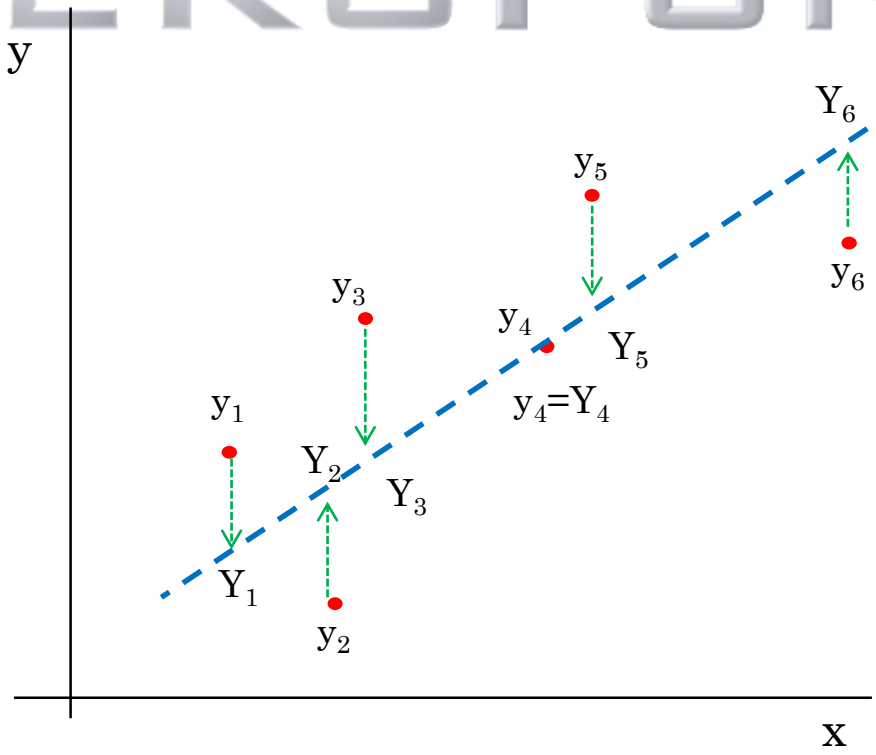
$$y_i = \eta_i + \varepsilon_i$$

Červené body značí empirické (napozorované) hodnoty
Musíme najít „vhodnou“ přímku

$$y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i$$

Každou empirickou hodnotu y_i nahradíme určitou „**vyrovnanou**“ hodnotou Y_i
Která bude ležet na zvolené empirické (výběrové) regresní přímce

$$Y_i = b_0 + b_1 . x_i$$



Problém je, že takových přímek může existovat nekonečně mnoho

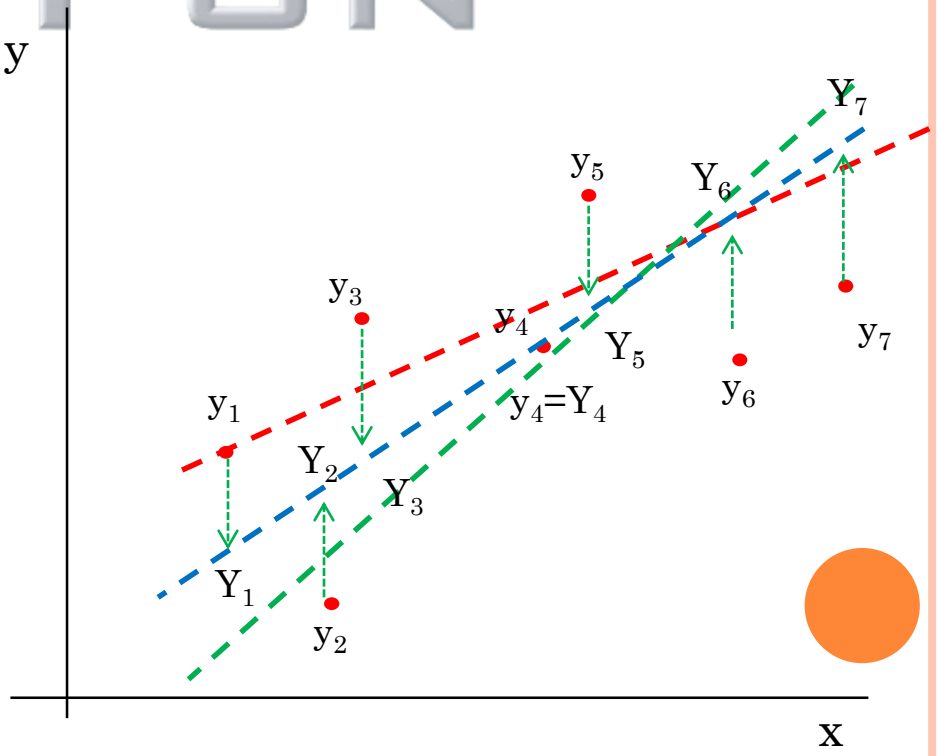
Musíme najít kritérium – nejlépe vystihne danou závislost

Zelené šipky představují odchylku skutečné hodnoty od „vyrovnané“ hodnoty
Když už musí existovat odchylky – ideální by bylo
jejich vzájemné vykompenzování

$$\sum_{i=1}^{n}(y_i - Y_i) = \sum_{i=1}^{n} e_i = 0$$

Kladné a záporné odchylky
Se „požerou“

e_i-reziduum
Rozdíl mezi empirickou regresní funkcí
a empirickou hodnotou



Součet čtverců odchylek empirických hodnot y_i od hodnot teoretických η_i byl minimální
Metoda nejmenších čtverců (MNČ, OLS)

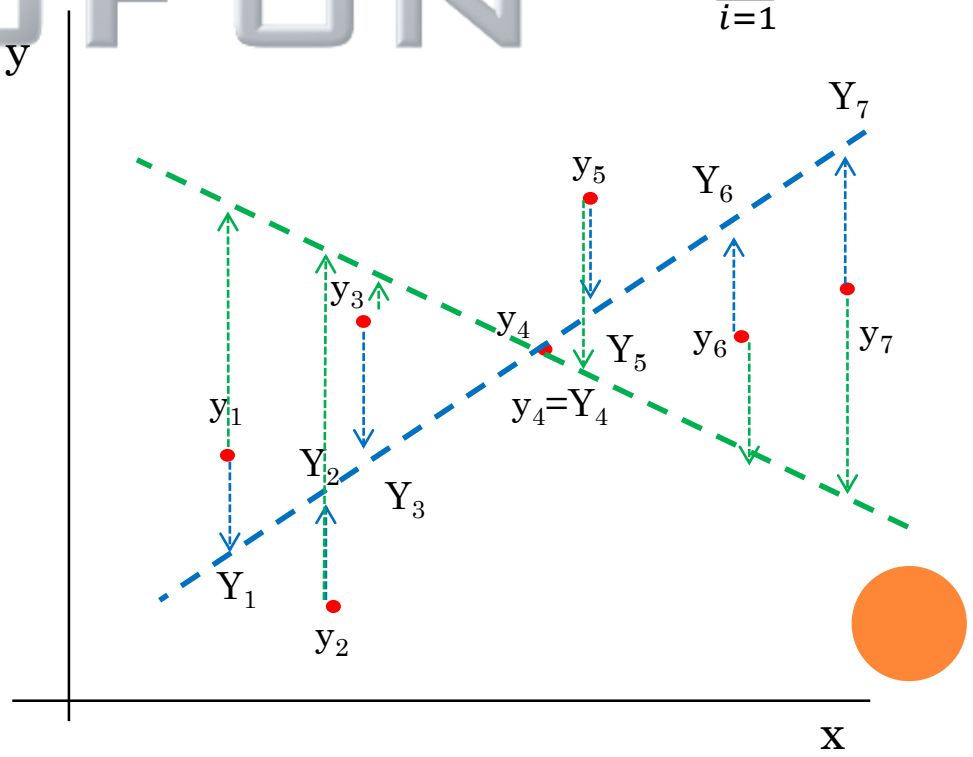
$$y_i = \eta_i + \varepsilon_i$$

$$\sum_{i=1}^{n} e_i = 0$$

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \eta_i)^2 \ldots min$$

Reziduum e je odhadem ε
A Y je odhadem η
Musí platit, že:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - Y_i)^2 \ldots min$$



Přímková regrese

$$y_i = \eta_i + \varepsilon_i$$

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \eta_i)^2 \ldots min$$

$$\eta = \beta_0 + \beta_1 . x$$

$$Y = b_0 + b_1 . x$$

b_0 *je odhad* β_0

b_1 *je odhad* β_1

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad Q\ min$$

$$\frac{\partial Q}{\partial \beta_0} = 0 \qquad \frac{\partial Q}{\partial \beta_1} = 0 \qquad Q = \sum_{i=1}^{2} (y_i - b_0 - b_1 x_i)^2 = (y_1 - b_0 - b_1 x_1)^2 + (y_2 - b_0 - b_1 x_2)^2$$

$$\frac{\partial Q}{\partial b_0} = 2.(y_1 - b_0 - b_1 x_1).(-1) + 2.(y_2 - b_0 - b_1 x_2).(-1) = 0$$

$$\frac{\partial Q}{\partial b_0} = 2.\sum_{i=1}^{2} (y_i - b_0 - b_1 x_i).(-1) = 0$$

$$\frac{\partial Q}{\partial b_1} = 2.(y_1 - b_0 - b_1 x_1).(-x_1) + 2.(y_2 - b_0 - b_1 x_2).(-x_2) = 0$$

$$\frac{\partial Q}{\partial b_1} = 2.\sum_{i=1}^{2} (y_i - b_0 - b_1 x_i).(-x_i) = 0$$

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \eta_i)^2 \dots min \qquad Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q}{\partial b_0} = 2. \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i). (-1) = 0$$

$$\frac{\partial Q}{\partial b_0} = 2. (y_1 - b_0 - b_1 x_1). (-1) + 2. (y_2 - b_0 - b_1 x_2). (-1) = 0$$

$$\frac{\partial Q}{\partial b_1} = 2. \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i). (-x_i) = 0$$

$$\frac{\partial Q}{\partial b_1} = 2. (y_1 - b_0 - b_1 x_1). (-x_1) + 2. (y_2 - b_0 - b_1 x_2). (-x_2) = 0$$

Normální rovnice

$$\sum_{i=1}^{n} y_i = n. b_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i . x_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

$$\left(\begin{array}{cc|c} n & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i^2 & \sum y_i x_i \end{array} \right)$$

$$b_0 = \begin{vmatrix} \sum y_i & \sum x_i \\ \sum y_i x_i & \sum x_i^2 \\ \hline n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}$$

$$b_1 = \begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum y_i x_i \\ \hline n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}$$

$E(Y|X)=\bar{y}+.b_{xy}(x-\bar{x})$ $Y=\bar{y}+b_{xy}.(x-\bar{x})$ $Y=b_0+b_1.x$

Regresní koeficient (výběrový regresní koeficient) $b_{xy}=\frac{s_{xy}}{s_x^2}$

Směrnice (sklon) regresní přímky

$b_{xy}=\frac{cov(x,y)}{Var(x)}$

Průměrná změna závisle proměnné y
Při jednotkové změně nezávisle proměnné x

Může nabýt libovolných hodnot!!!

Jednodušší postup pro **přímkovou regresi**!!!!

$b_0=\bar{y}-b_1\bar{x}$

Přímková regrese je lineární regresní funkce (lineární v parametrech)
Obráceně nemusí platit!!!

$cov(x,y)>0$

$cov(x,y)<0$

$cov(x,y)=0$

Lineární nezávislost

Linearizace modelu

Linearita v parametrech

$$lny = lnb_0 + b_1 lnx \qquad OK$$

$$y = b_0 x^{b_1} \qquad Není\ OK\ :)$$

Vzpomeňte na matice
Lineární algebra – pro praktičnost je výhodnější mít lineární model

Některé nelineární modely se dají linearizovat

$$Q = 5 - 2lnP$$

$$lnQ = 100 - 0{,}04P$$

$$lnQ = 7 - 0{,}01lnP$$

Linearizující transformace

$y = b_0 x^{b_1}$ $\quad lny = lnb_0 + b_1 lnx$

$y = \frac{b_0}{x^{b_1}}$ $\quad lny = lnb_0 - b_1 lnx$

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 \Delta x$
level-log	y	log(x)	$\Delta y = (\beta_1/100)\%\Delta x$
log-level	log(y)	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	log(y)	log(x)	$\%\Delta y = \beta_1 \%\Delta x$

Další typy regresních funkcí

Parabolická regrese

Není vícenásobná regrese!!!

$$\eta = \beta_0 + \beta_1 . x + \beta_2 . x^2$$

Aplikujeme MNČ
Interpretace výsledků

ei = yi □ b0 □ b1xi

Polynomická regrese

$$\eta = \beta_0 + \beta_1 . x + \beta_2 . x^2 + \cdots + \beta_p . x^p$$

Lineární v parametrech
Nelineární v

Hyperbolická regrese

$$\eta = \beta_0 + \frac{\beta_1}{x}$$

Logaritmická regrese

Lineární v parametrech $\eta = \beta_0 + \beta_1 log x$
Nelineární v

Interpretace výsledků

Exponenciální regrese

Nelineární v parametrech
Nelze použít MNČ

$$\eta = \beta_0 . \beta_1^x$$

Logaritmická transformace – zlogaritmujeme (linearizujeme)

$$\log \eta = \log \beta_0 + x . \log \beta_1$$

Interpretace výsledků

Zdánlivá regrese (spurious regression)

Někdy nastane situace, že regresní model vykazuje vysoké R2
Přesto se jedná o nesmyslný vztah

Váha dětí a znalost gramatiky
Čím jsou děti těžší, tím mají lepší gramatiku
Zapomínáme na stáří dětí!!!

Vzájemný vztah přes třetí proměnnou

Možnost existence krátkodobého vztahu např. stochastický trend atd.

Dávat si na zdánlivou regresi VELKÝ pozor

Zájemci si mohou vyhledat termín kointegrace časových řad

Interpolační a extrapolační odhady

Vzniklý model musíme testovat

Interpolační odhad
Do vzniklého modelu dosazujeme vysvětlující proměnné z oblasti měření

$$výnos = 5 + 1{,}5.\text{hnojivo} + u$$

Extrapolační odhad
Do vzniklého modelu dosazujeme hodnoty mimo interval měření

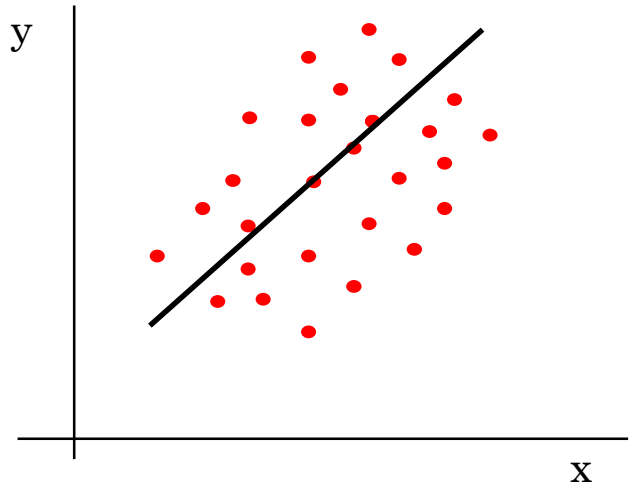
Máme hodnoty z intervalu (0;1000)
A chceme predikovat chování pro hodnoty z intervalu (1000;1500)

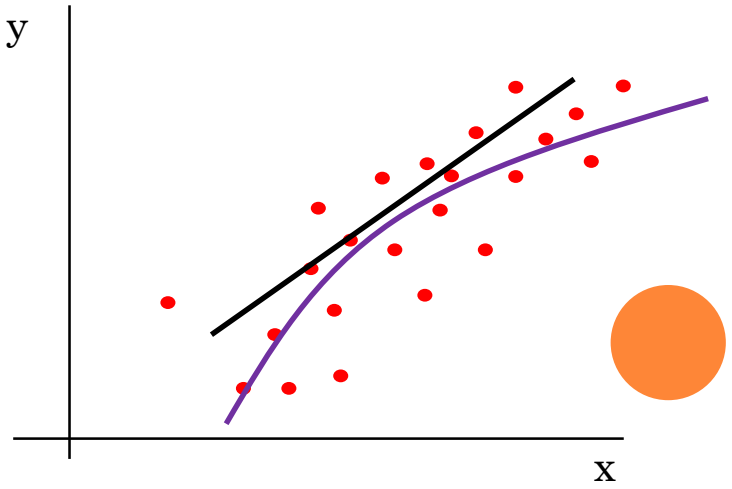
Kvalita regresní funkce a intenzita závislosti

Zjistíme případný vztah lineární/nelineární
Přímková regrese, parabolická atd.

Je však daný model „kvalitní“?
Regresní model bude tím lepší
čím více budou empirické hodnoty vysvětlované proměnné
soustředěny (nalepany) kolem odhadnuté regresní funkce

Cílem kapitoly je objasnit si nástroje na měření kvality regresního modelu





Index korelace

Empirický rozptyl (ER)

$$s_y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Teoretický rozptyl (TR)

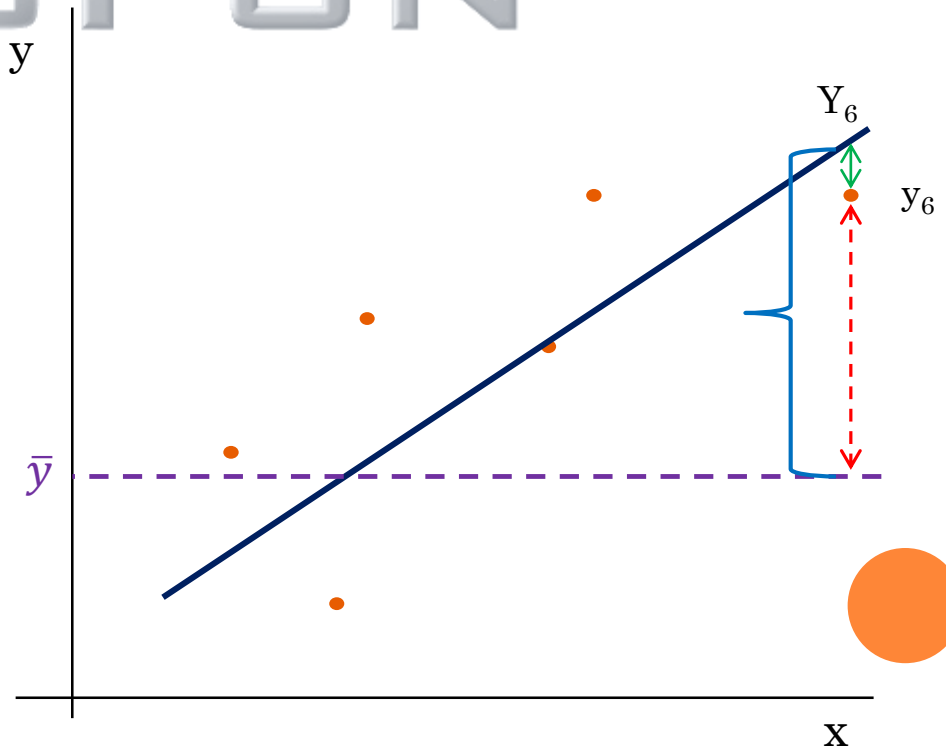
$$s_Y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (Y_i - \bar{y})^2$$

Residuální rozptyl (RR)

$$s_{(y-Y)}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - Y_i)^2$$

Při použití MNČ platí mezi rozptyly vztah:

$$s_y^2 = s_Y^2 + s_{(y-Y)}^2$$



Empirický rozptyl (ER)
Teoretický rozptyl (TR)
Residuální rozptyl (RR)

$$s_y^2 = \frac{1}{n}.\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad s_Y^2 = \frac{1}{n}.\sum_{i=1}^{n}(Y_i - \bar{y})^2 \qquad s_{(y-Y)}^2 = \frac{1}{n}.\sum_{i=1}^{n}(y_i - Y_i)^2$$

$$s_y^2 = s_Y^2 + s_{(y-Y)}^2$$

Funkční závislost

$$s_y^2 = s_Y^2$$

Všechny empirické hodnoty (y_i)
jsou zároveň vyrovnanými hodnotami (Y_i)
„čím lepší závislosti, tím více se ER a TR blíží“

Úplná nezávislost $s_y^2 = s_{(y-Y)}^2$

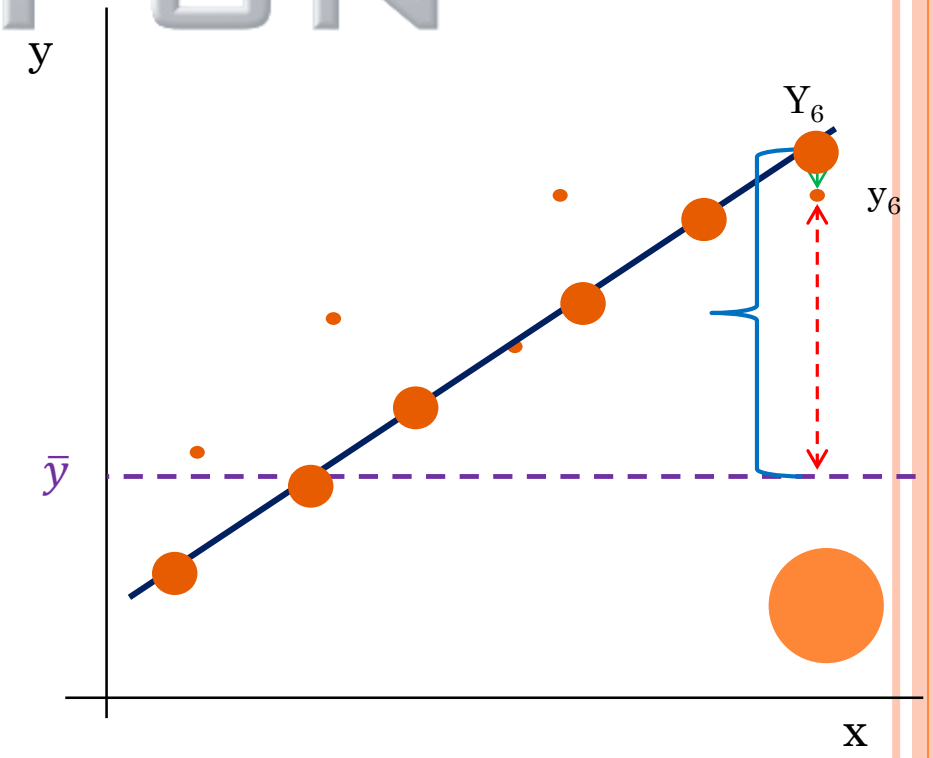
Empirický rozptyl shodný s reziduálním
„čím horší závislost, tím se ER a RR blíží“

Hodnocení stochastického modelu

Zvolený model bude tím kvalitnější
Čím bude podíl teoretického rozptylu
Na celkovém rozptylu větší!!!

$$\frac{s_Y^2}{s_y^2}$$

Tím silnější bude závislost y na x



$$s_y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$s_Y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (Y_i - \bar{y})^2$$

Index determinace R2

$$I_{yx}^2 = \frac{s_Y^2}{s_y^2}$$

$$s_y^2 = s_Y^2 + s_{(y-Y)}^2$$

$$R^2 = \frac{"vysvětlený\ rozptyl"}{celkový\ rozptyl}$$

Index nabývá hodnot 0-1
R^2=1 představuje funkční závislost
R^2=0 představuje nezávislost

x	y
0	2
1	2,2
2	2,4
3	2,6
4	2,8
5	3

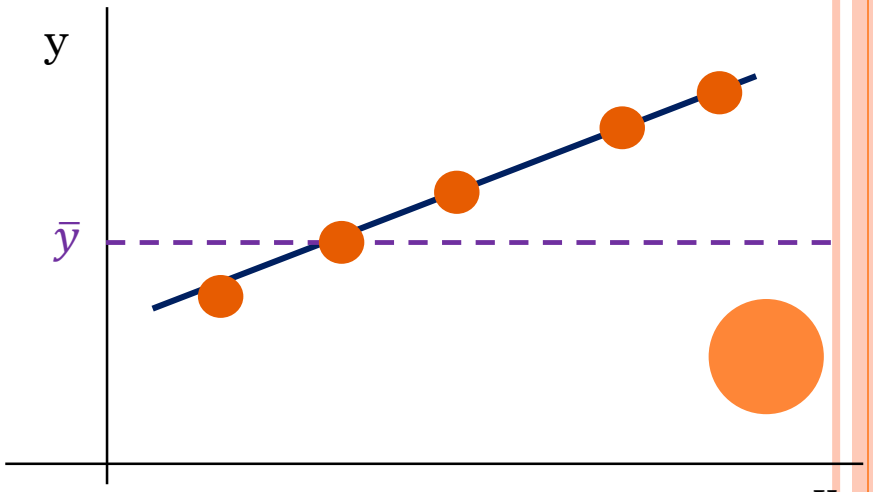
$$Y_i = 2 + 0{,}3 . x_i$$

$\bar{x}$	$\bar{y}$
2,5	2,75

Vynásobeno 100 udává v % tu část rozptylu kterou se podařilo vysvětlit regresní funkcí

$$\frac{s_Y^2}{s_y^2} = 1 - \frac{s_{(y-Y)}^2}{s_y^2}$$

Relativní část, která se nepodařila vysvětlit modelem



Index korelace

$$I_{yx} = \sqrt{\frac{s_Y^2}{s_y^2}}$$

Index determinace <0,1>

$$I_{yx}^2 = \frac{s_Y^2}{s_y^2}$$

Funkční závislost – R2=1
Nezávislost –R2=0

Převedením na % - vyjadřuje tu část rozptylu vysvětlované proměnné (y) kterou se podařilo vysvětlit pomocí regresní funkce

R2=0,8 – 100.0,8=80%
80% hodnot se nám podařilo vysvětlit pomocí konkrétního typu reg. fce

Index korelace

$$I_{yx} = \sqrt{\frac{s_Y^2}{s_y^2}}$$

Koeficient korelace

Zvláštní případ indexu korelace

Měří těsnost závislosti dané **LINEÁRNÍ** regresní funkce

$$I_{yx} = \sqrt{\frac{s_Y^2}{s_y^2}} \quad \rightarrow \quad r_{yx} = r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 . s_y^2}}$$

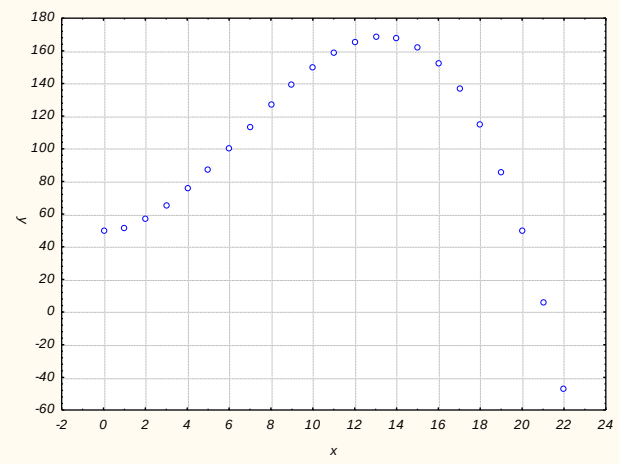
r_{xy}- koeficient korelace
s_{xy}- kovariance
$s^2_{(x,y)}$- rozptyly

Koeficient korelace <-1,1>

r_{xy}=-1
Nepřímá **lineární** závislost –

r_{xy}=1
Přímá **lineární** závislost –

r_{xy}=0 **lineární** nezávislost



$$-0{,}1x^3 + 2x^2 + 50$$

x	y
0	50
1	51,9
2	57,2
3	65,3
4	75,6
5	87,5
6	100,4
7	113,7
8	126,8
9	139,1
10	150
11	158,9
12	165,2
13	168,3
14	167,6
15	162,5
16	152,4
17	136,7
18	114,8
19	86,1
20	50
21	5,9
22	-46,8

$$r_{yx} = -0{,}02$$

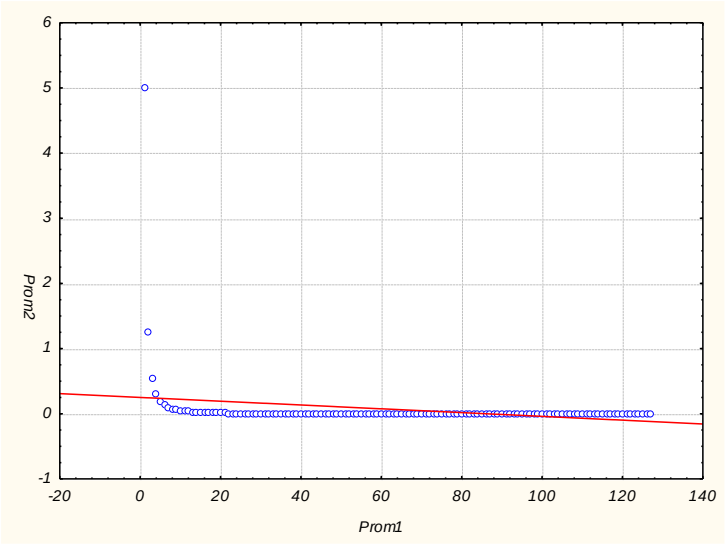
r_{xy}=0

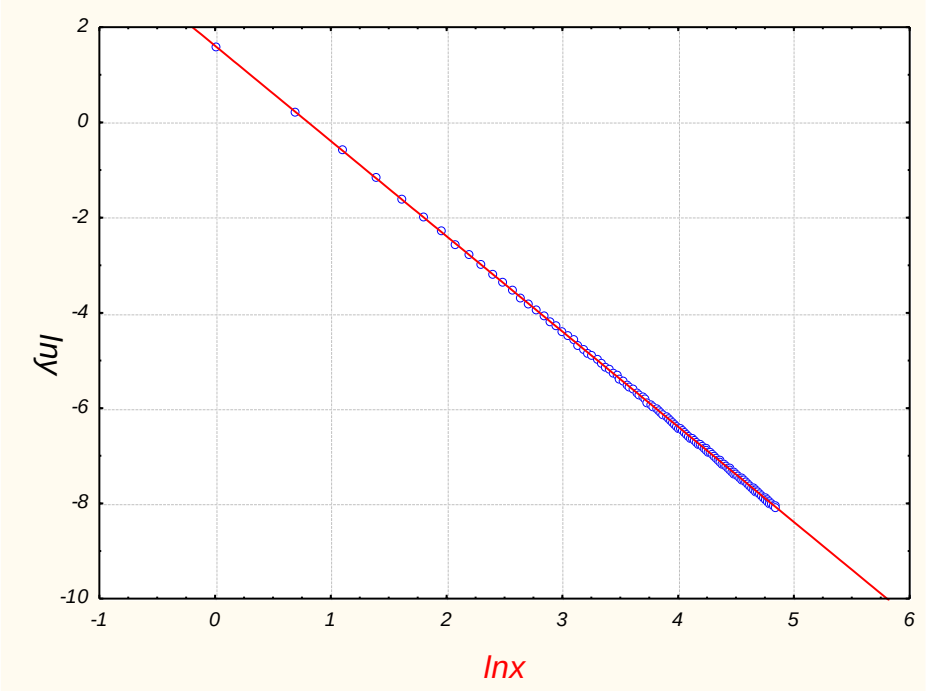
Nemusí znamenat nezávislost
Může se jednat o silnou závislost
Ale NELINEÁRNÍ!!!

$$y = \frac{5}{x^2}$$

$$lny = ln5 - 2lnx$$

$$r_{yx} = -1$$





x	y
1	5
2	1,25
3	0,555555556
4	0,3125
5	0,2
6	0,138888889
7	0,102040816
8	0,078125
9	0,061728395
10	0,05
11	0,041322314
12	0,034722222
13	0,029585799
14	0,025510204
15	0,022222222
16	0,01953125
17	0,017301038
18	0,015432099
19	0,013850416
20	0,0125

$$r_{yx} = -0{,}23$$