VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY | KATEDRA APLIKOVANÉ MATEMATIKY

MODAM 2023

Něco málo o závislostech

Ing. Martina Litschmannová, Ph. D.
Mgr. Adéla Kondé

Obsah

- Analýza závislosti dvou kvantitativních proměnných
  - ✓ Základní korelační koeficienty a vizualizace závislosti dvou kvantitativních proměnných
  - ✓ Interpretace korelačních koeficientů

- Regrese z pohledu statistiky
  - ✓ Základní pojmy
  - ✓ Metoda nejmenších čtverců
  - ✓ Interpretace regresních koeficientů
  - ✓ Extrapolace a interpolace

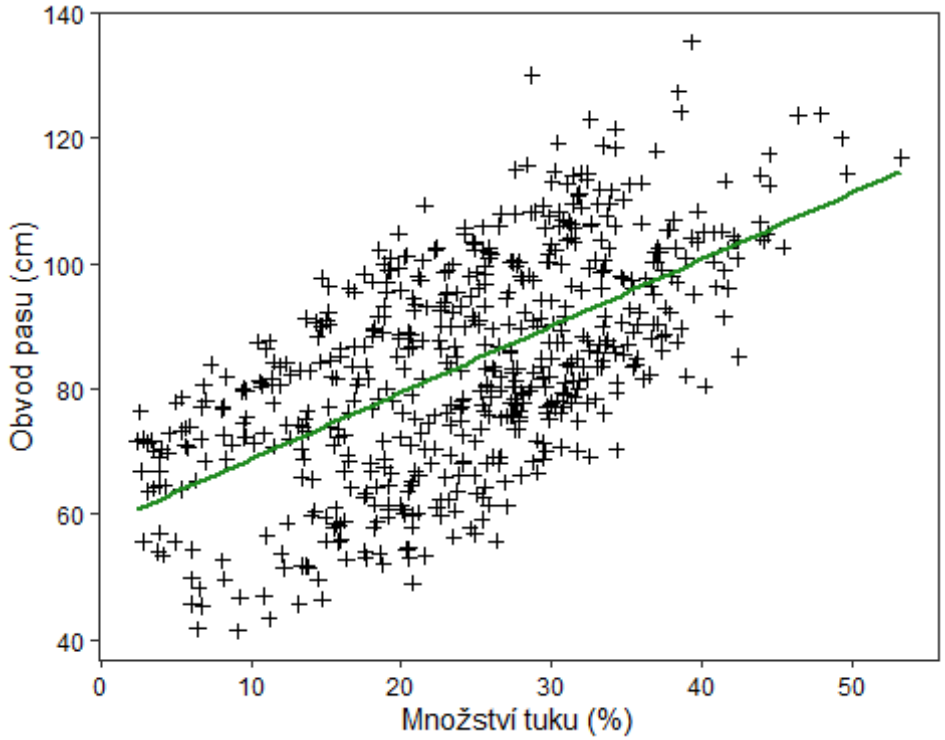
Analýza závislosti dvou kvantitativních proměnných

Analýza závislosti dvou numerických proměnných

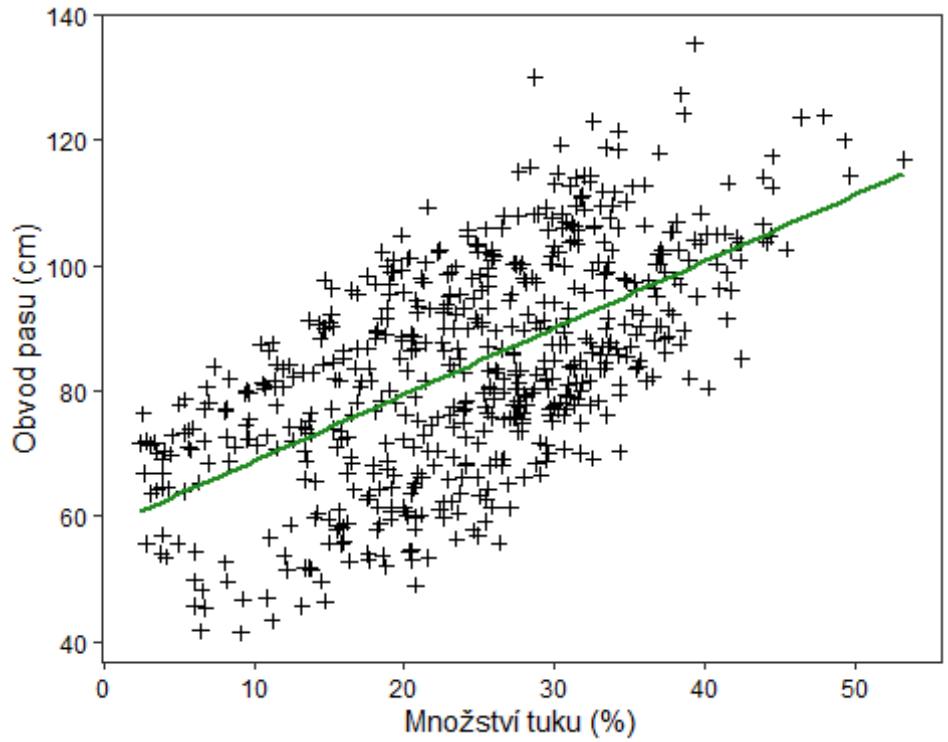
ID	Pohlaví	Rasa	Věk (roky)	BMI (kg/m^2)	Množství tuku (%)	Obvod pasu (cm)	Kvalita spánku	Kvalita spánku (dichotomizovaná)
1	muž	Negroidní	61	29,81	22,66	90,1	spíše špatná	špatná
2	žena	Mongoloidní	52	22,50	26,59	79,8	velmi dobrá	dobrá
3	muž	Negroidní	37	24,50	13,75	76,4	spíše dobrá	dobrá
4	žena	Mongoloidní	47	24,04	30,79	87,4	spíše špatná	špatná
5	muž	Europoidní	46	22,56	16,70	83,7	spíše dobrá	dobrá
6	žena	Negroidní	37	19,98	26,18	83,0	velmi dobrá	dobrá
7	žena	Negroidní	44	23,61	35,59	84,0	spíše dobrá	dobrá
8	muž	Mongoloidní	50	20,85	2,77	72,0	spíše dobrá	dobrá
9	muž	Negroidní	50	26,95	21,29	97,5	spíše špatná	špatná

Jak popsat a vizualizovat závislost mezi množstvím tuku a obvodem pasu?

Bodový graf

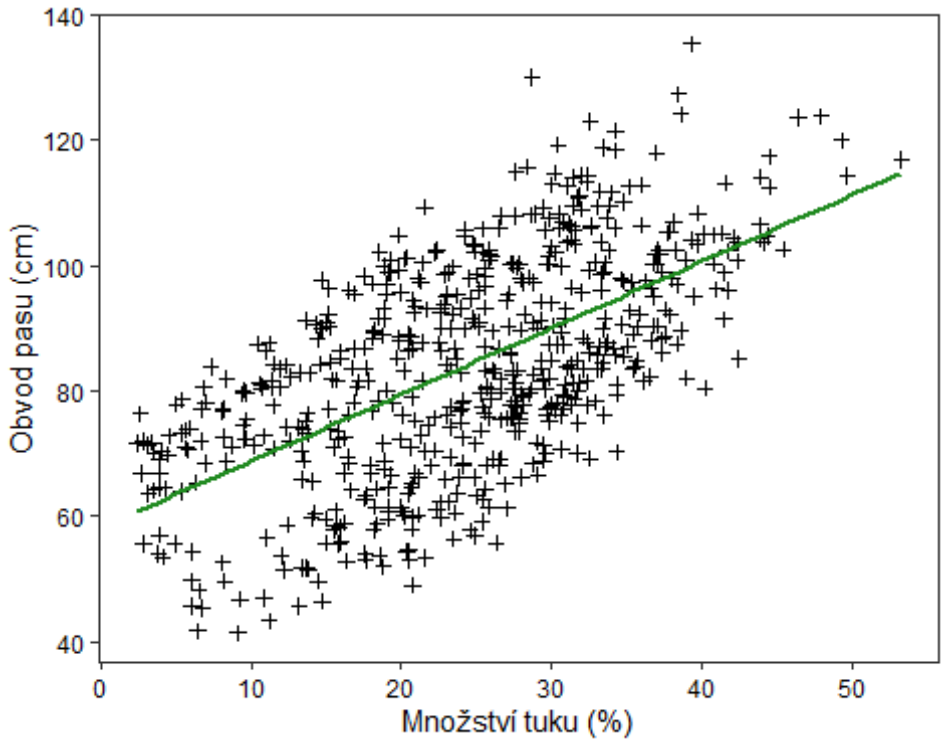


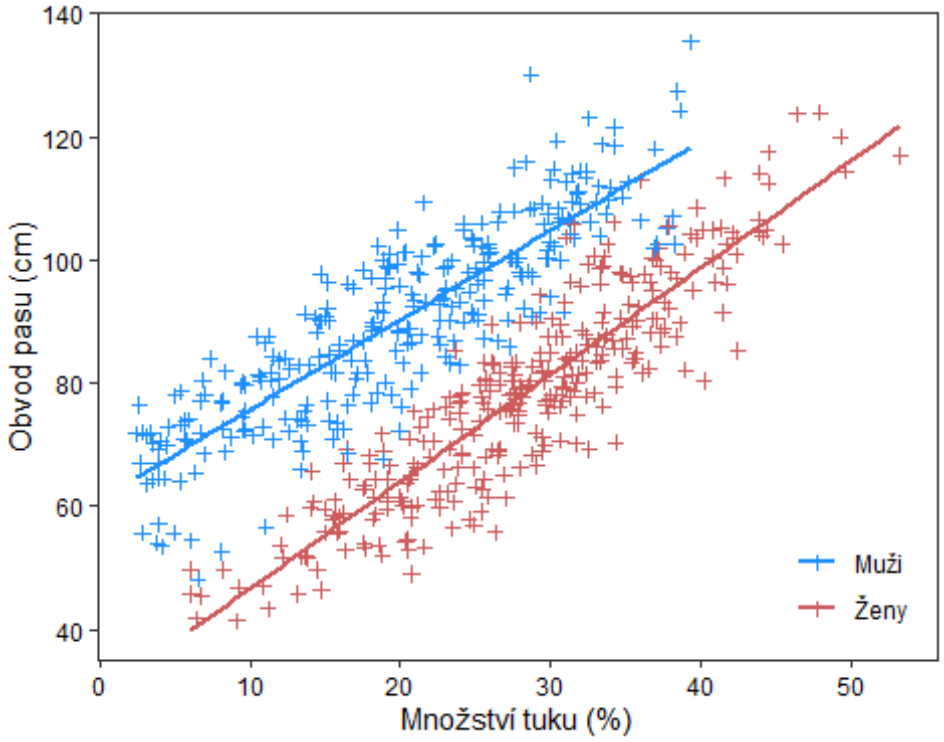
Bodový graf



Když chceme graficky prezentovat závislost dvou kvantitativních znaků, je pro snadnou interpretaci důležité rozhodnout, který ze znaků je vysvětlující (osa x) a který je vysvětlovaný (osa y).

Bodový graf





Když chceme graficky prezentovat závislost dvou kvantitativních znaků, je pro snadnou interpretaci důležité rozhodnout, který ze znaků je vysvětlující (osa x) a který je vysvětlovaný (osa y).

Pearsonův výběrový korelační koeficient

$$r_P(X,Y) = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_X \cdot s_Y}$$

n … rozsah výběru,

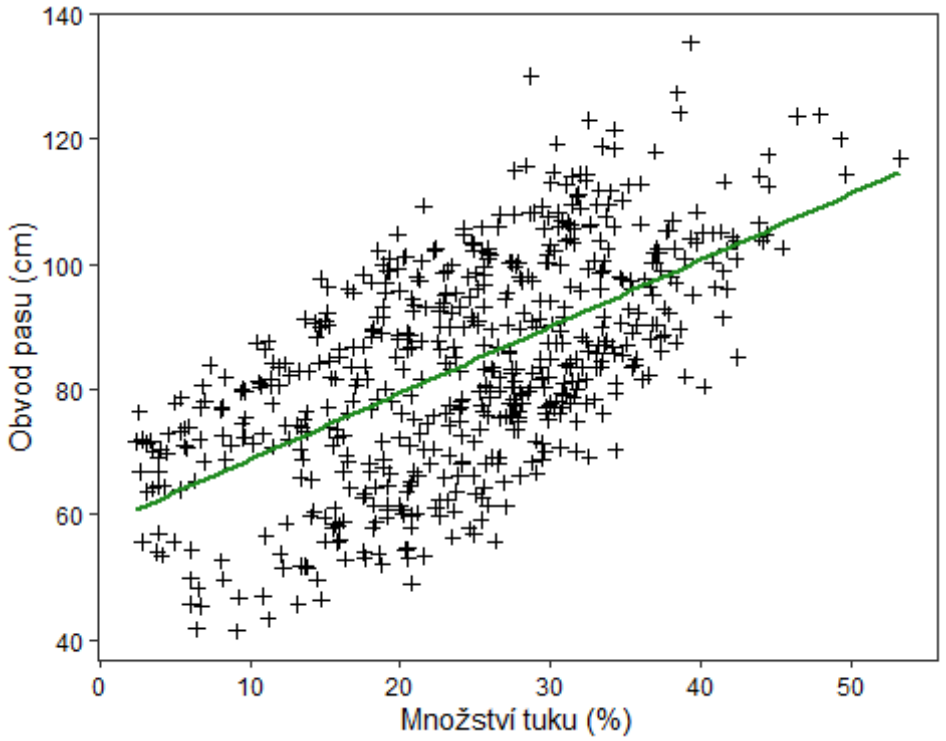
x_i … i-tá hodnota znaku x (vysvětlující proměnná),

y_i … i-tá hodnota znaku y (vysvětlovaná proměnná),

$\bar{x}$ ($\bar{y}$) … výběrový průměr znaku x (y),

s_X (s_Y) … výběrová směrodatná odchylka znaku x (y)

Pearsonův výběrový korelační koeficient

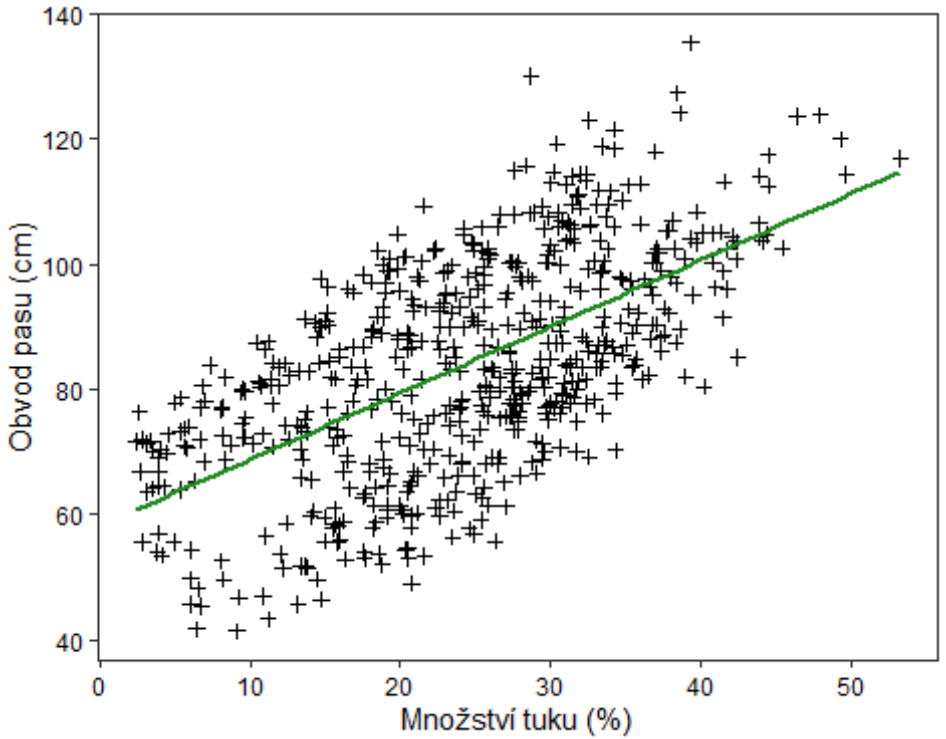


r_P = ?

Vlastnosti Pearsonova korelačního koeficientu

- $-1 \leq r_P(X,Y) \leq 1$
- $r_P(X,Y) = r_P(Y,X)$
- $r_P(X,X) = 1$
- Je-li $r_P(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_P(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_P(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_P(X,Y)| = 1$, pak je mezi X a Y lineární závislost.

Pearsonův výběrový korelační koeficient

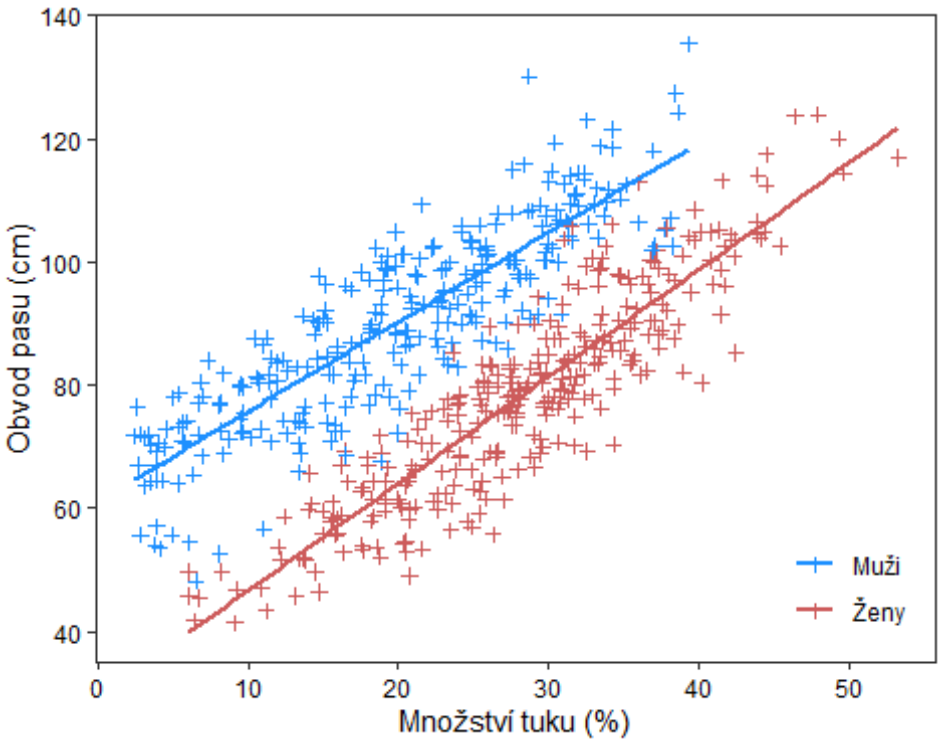


$r_P = 0{,}609$

Vlastnosti Pearsonova korelačního koeficientu

- $-1 \leq r_P(X,Y) \leq 1$
- $r_P(X,Y) = r_P(Y,X)$
- $r_P(X,X) = 1$
- Je-li $r_P(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_P(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_P(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_P(X,Y)| = 1$, pak je mezi X a Y lineární závislost.

Pearsonův výběrový korelační koeficient



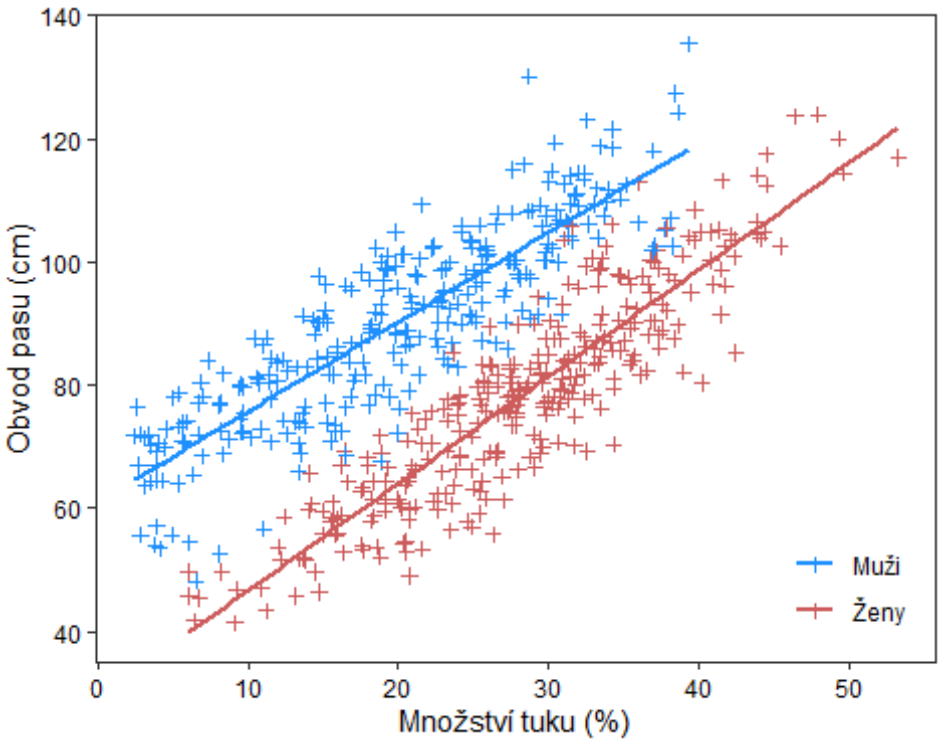
r_P = ?

r_P = ?

Vlastnosti Pearsonova korelačního koeficientu

- $-1 \leq r_P(X,Y) \leq 1$
- $r_P(X,Y) = r_P(Y,X)$
- $r_P(X,X) = 1$
- Je-li $r_P(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_P(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_P(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_P(X,Y)| = 1$, pak je mezi X a Y lineární závislost.

Pearsonův výběrový korelační koeficient

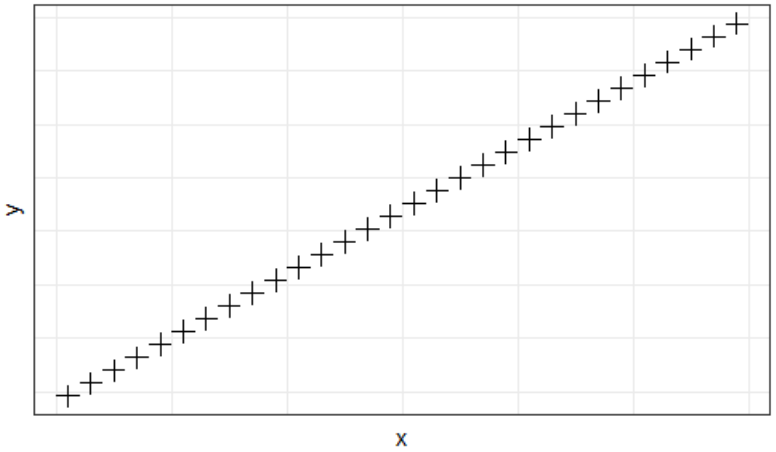


$r_P = 0{,}899$ $r_P = 0{,}865$

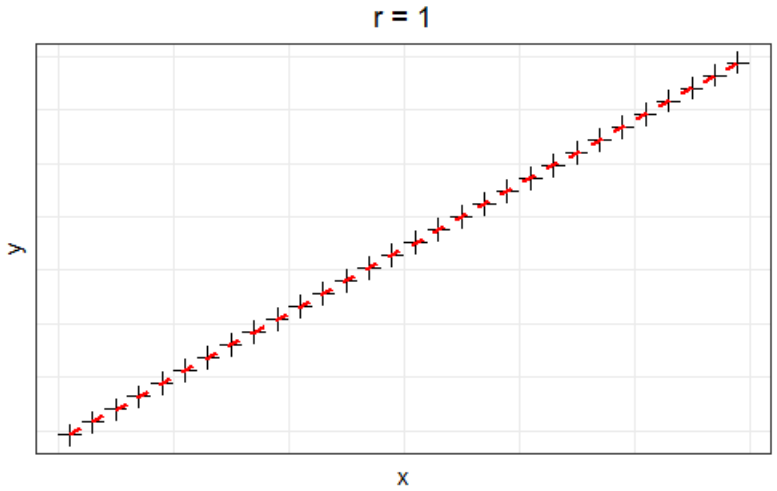
Vlastnosti Pearsonova korelačního koeficientu

- $-1 \leq r_P(X,Y) \leq 1$
- $r_P(X,Y) = r_P(Y,X)$
- $r_P(X,X) = 1$
- Je-li $r_P(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_P(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_P(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_P(X,Y)| = 1$, pak je mezi X a Y lineární závislost.

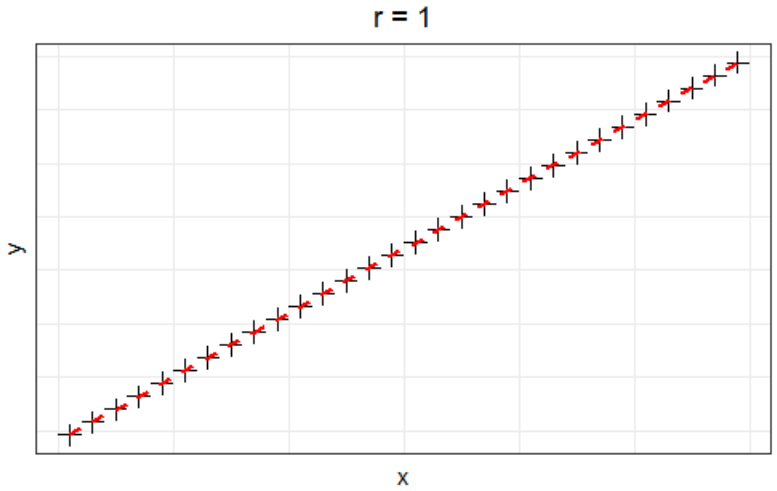
Pearsonův korelační koeficient

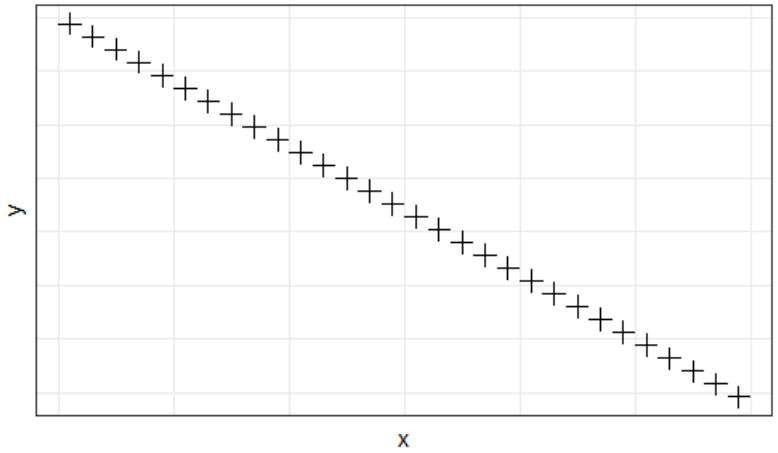


Pearsonův korelační koeficient

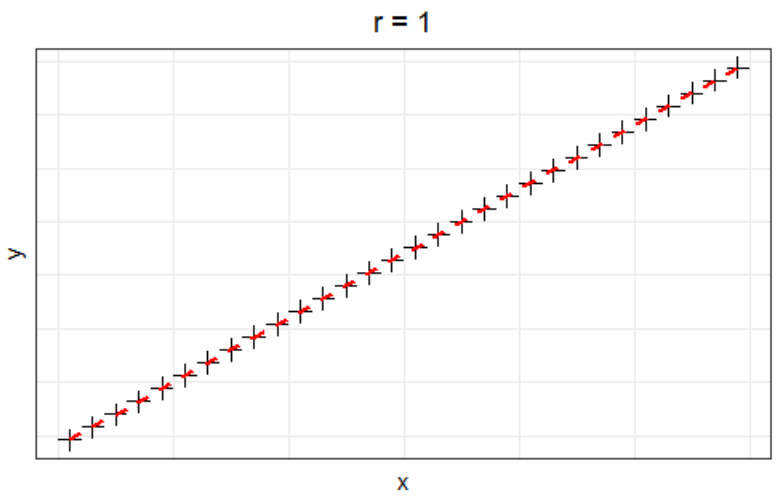


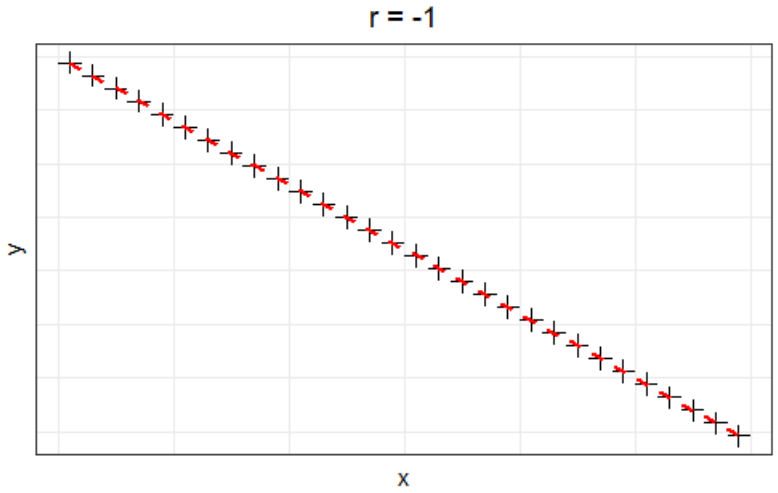
Pearsonův korelační koeficient



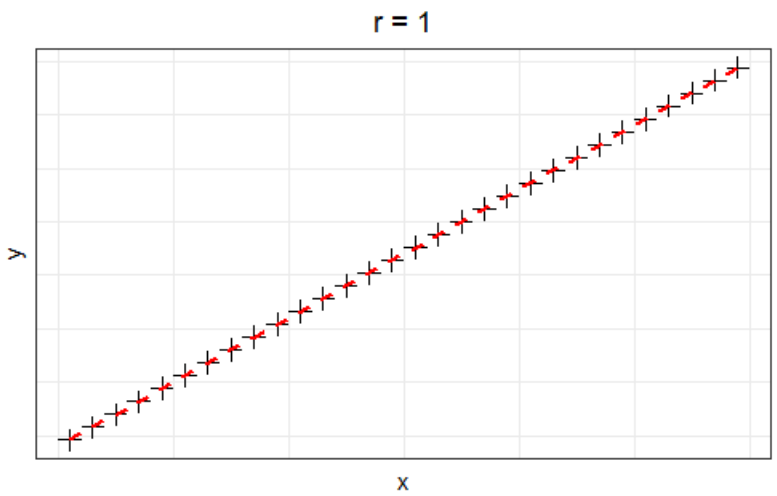


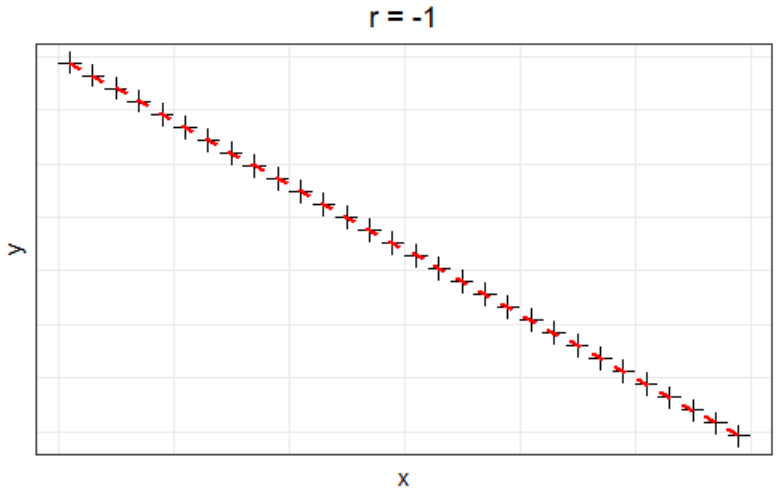
Pearsonův korelační koeficient

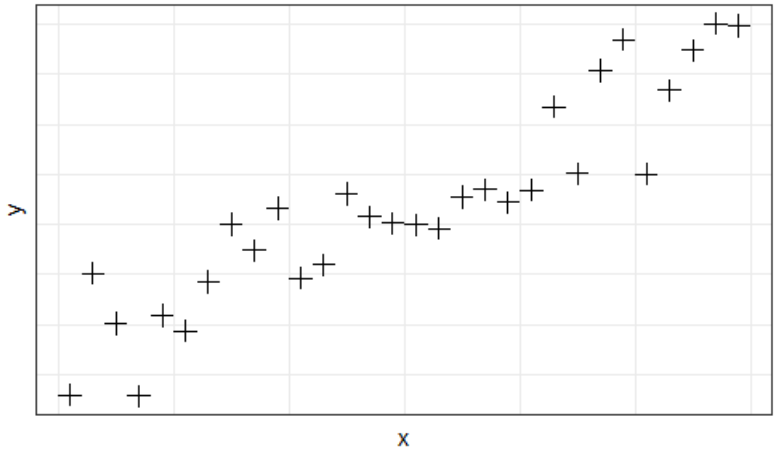




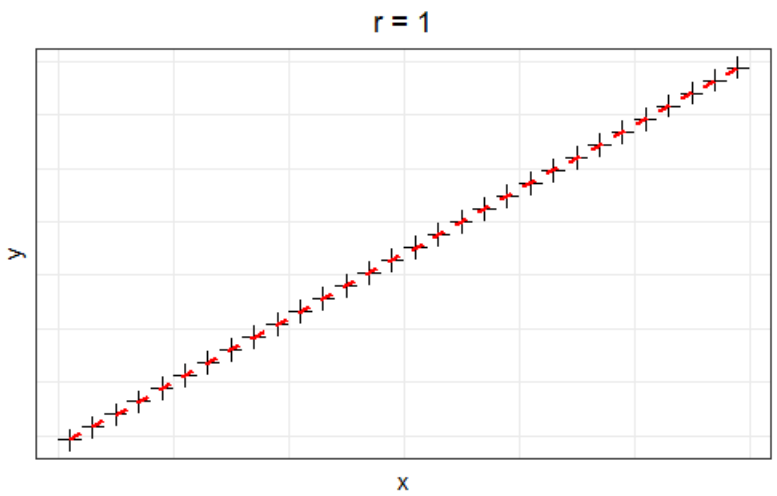
Pearsonův korelační koeficient

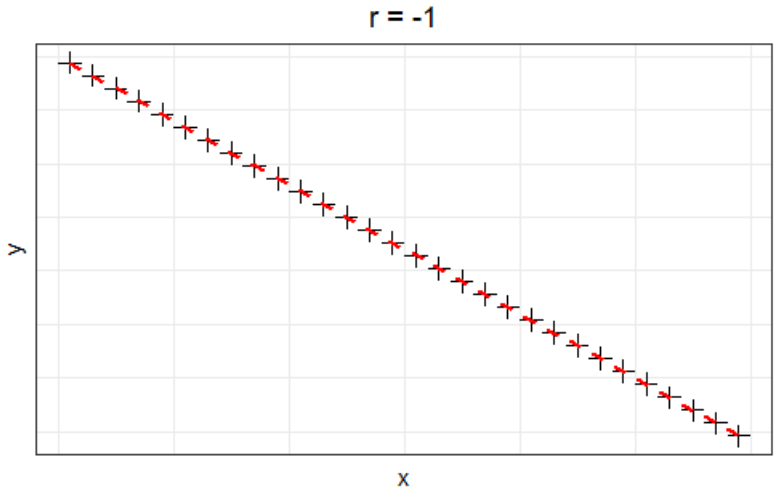


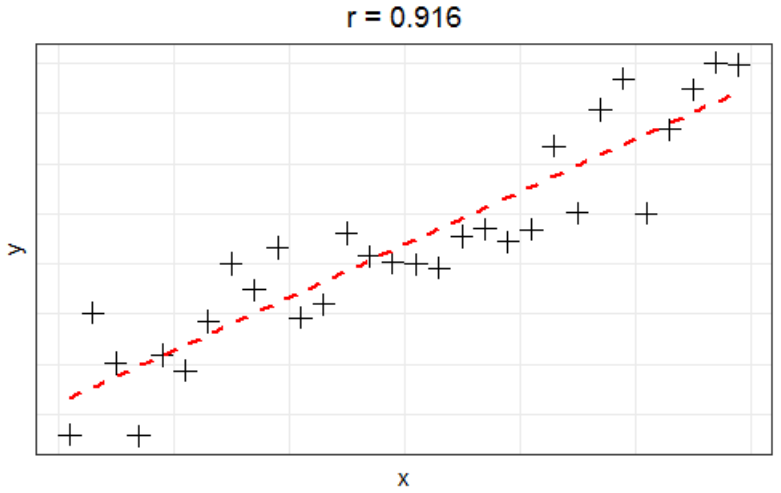




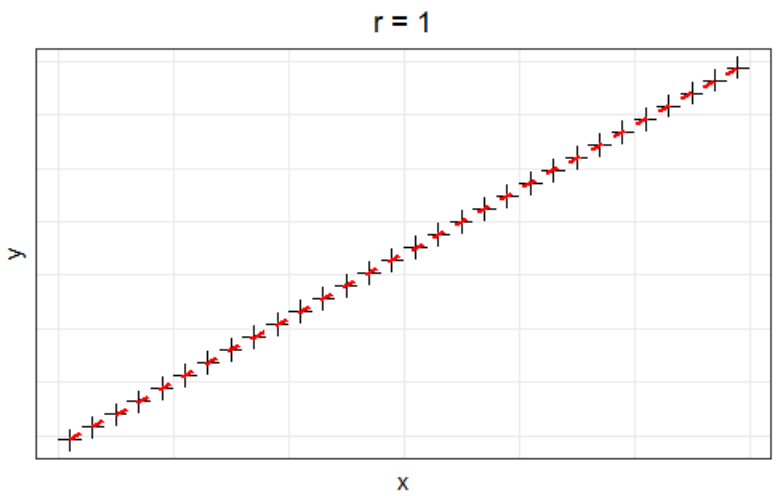
Pearsonův korelační koeficient

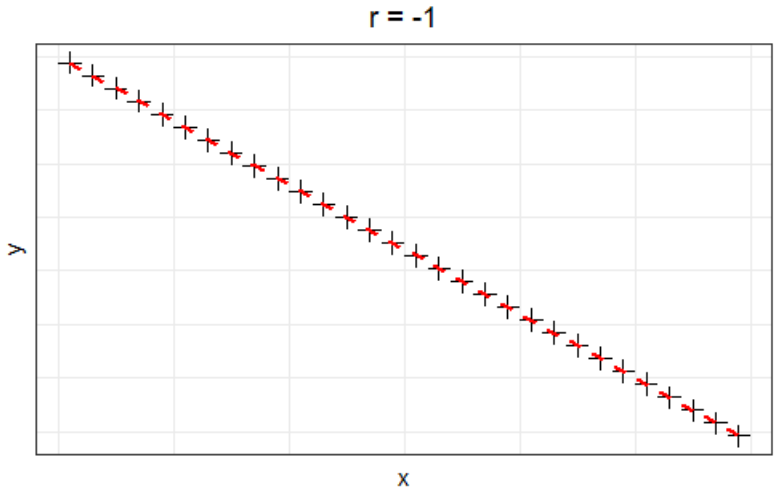


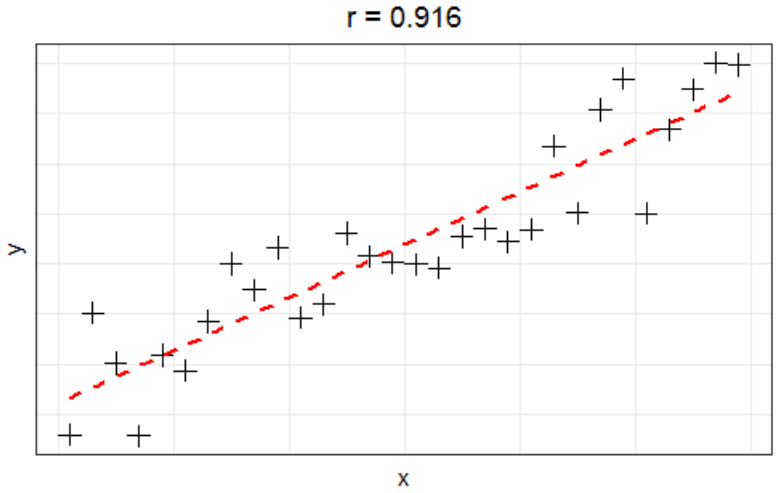


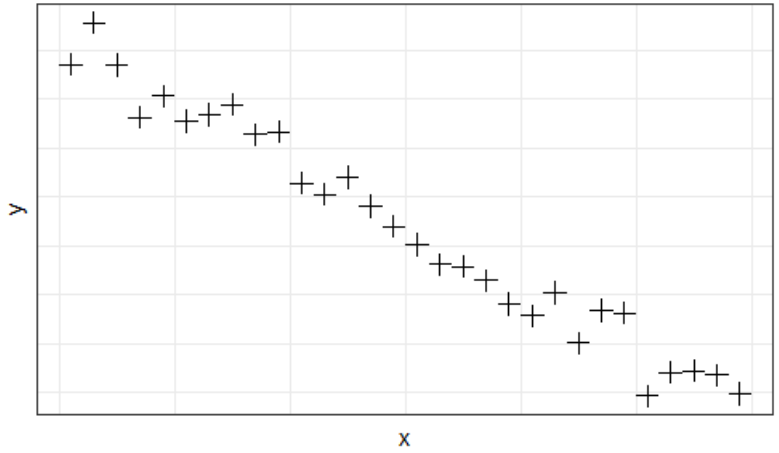


Pearsonův korelační koeficient

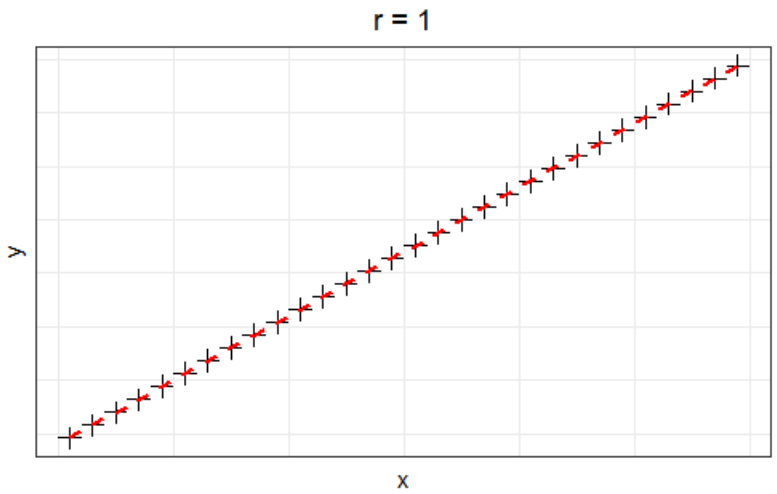


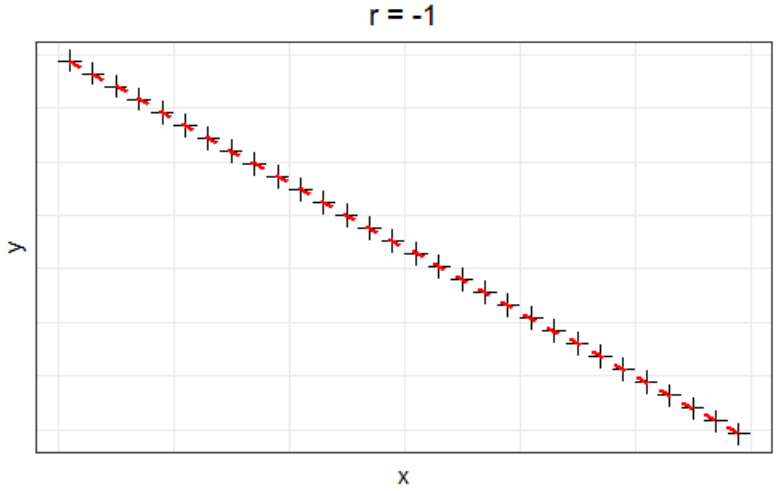


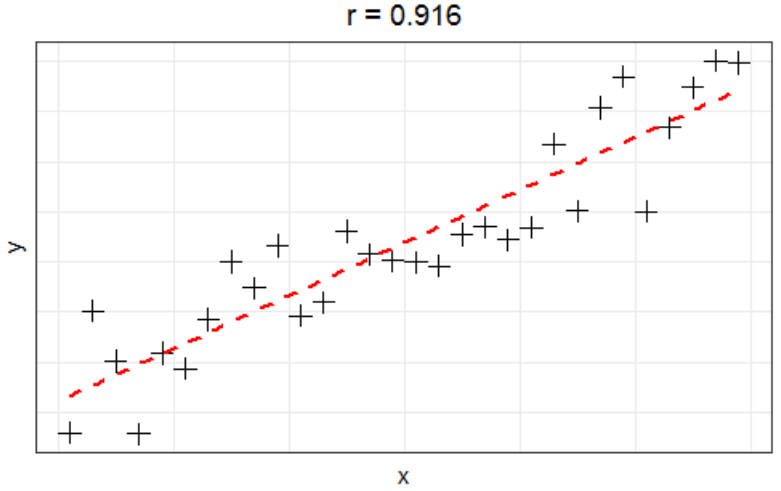


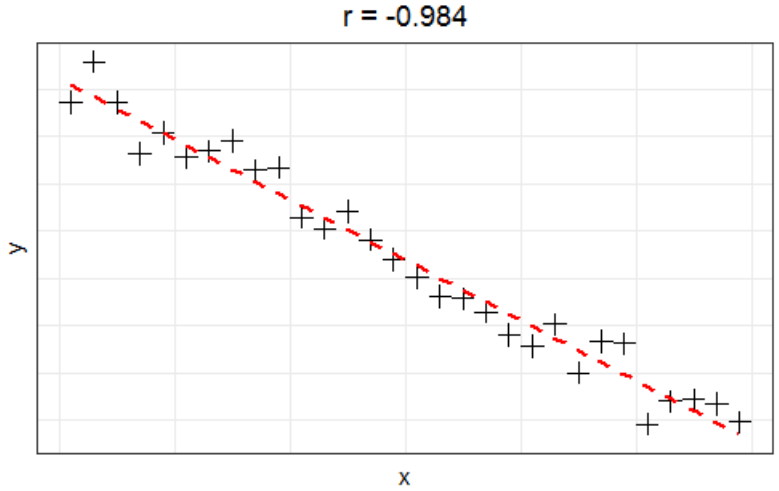


Pearsonův korelační koeficient

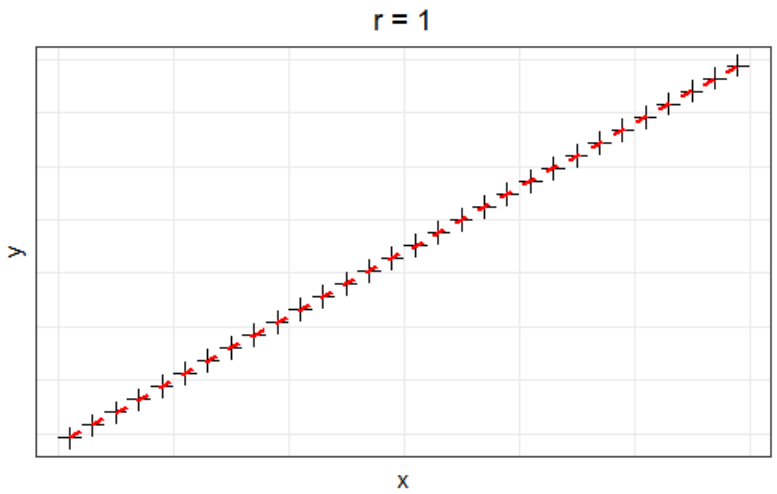


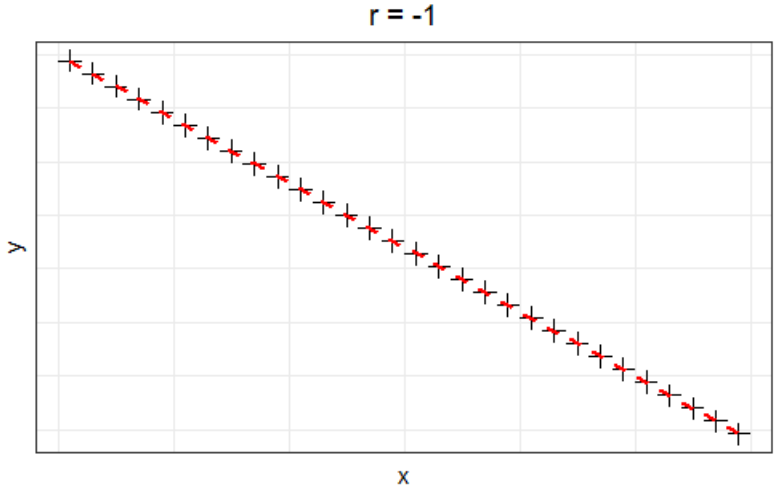


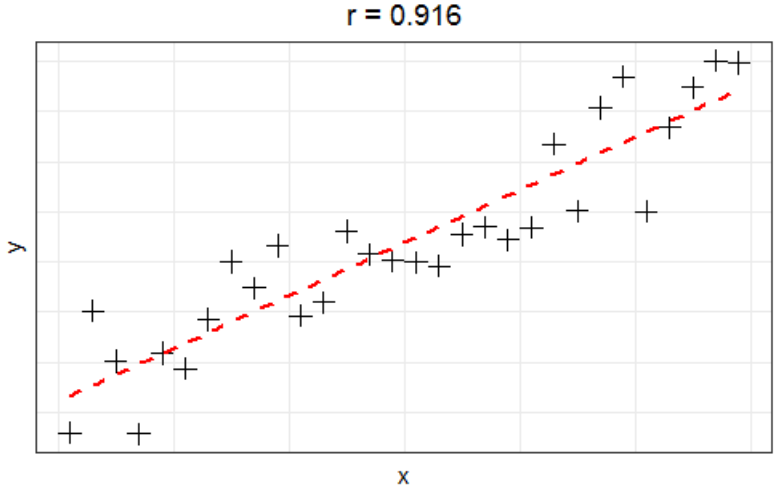


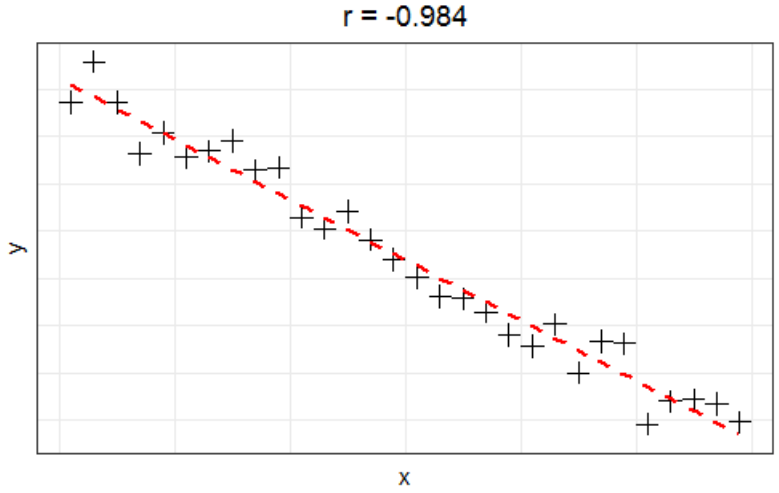


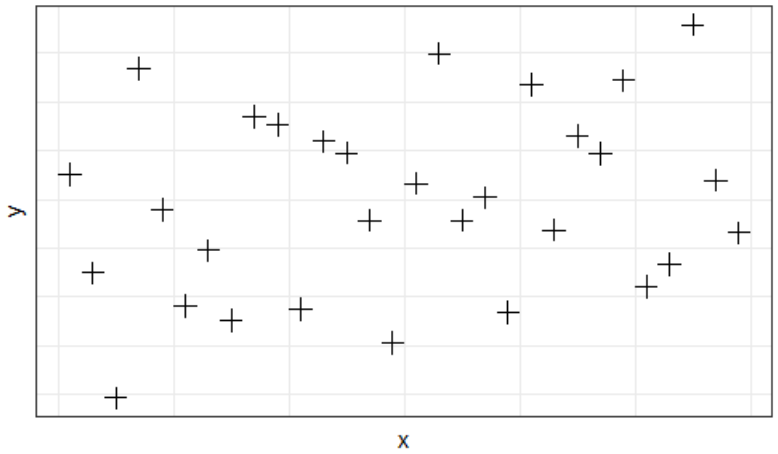
Pearsonův korelační koeficient



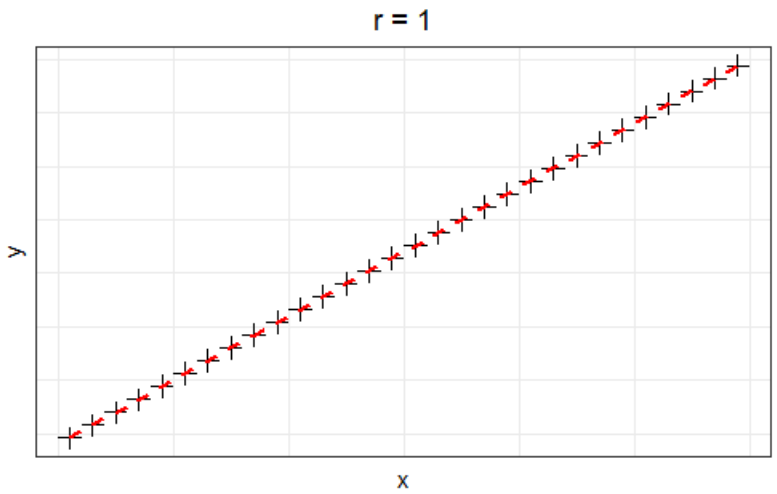


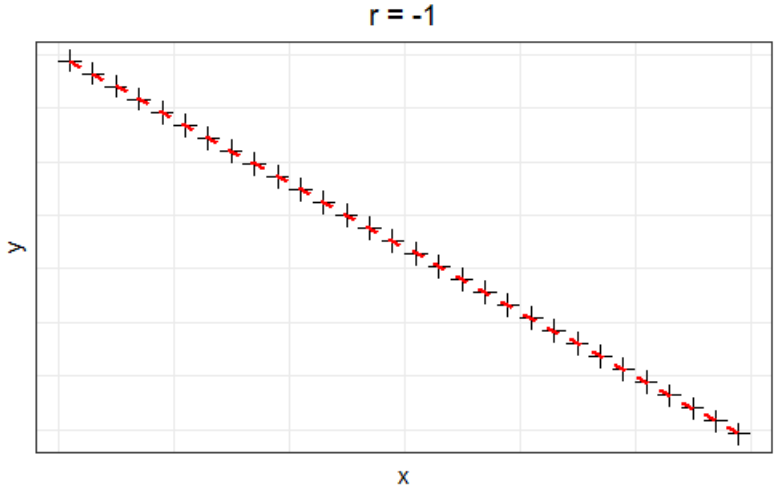


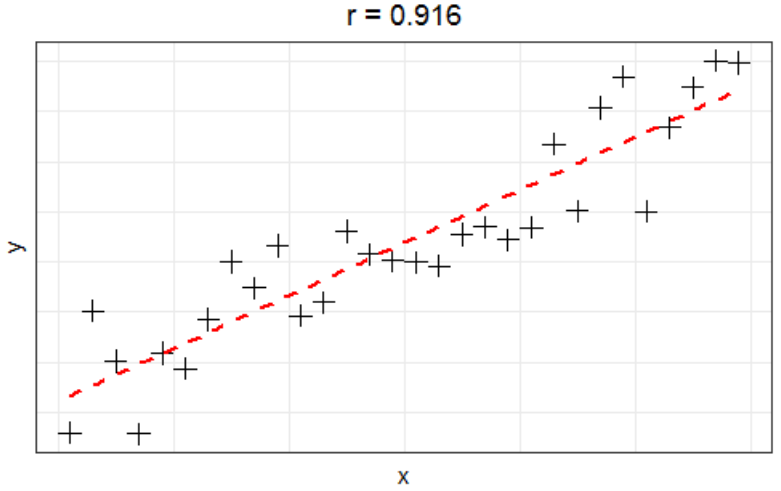


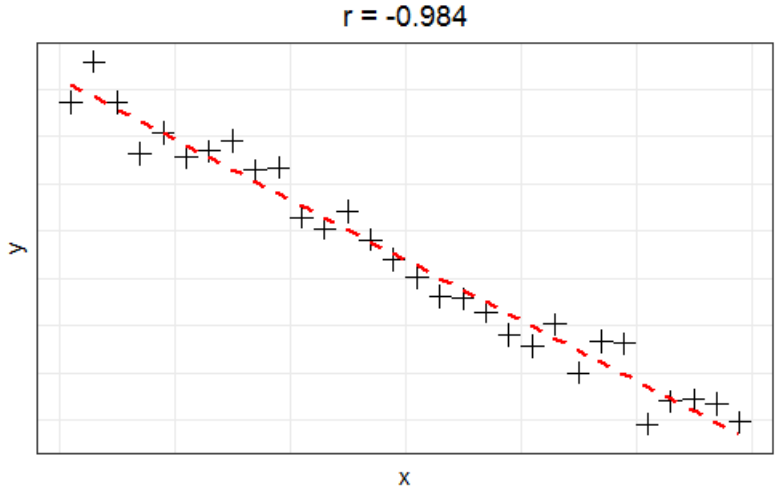


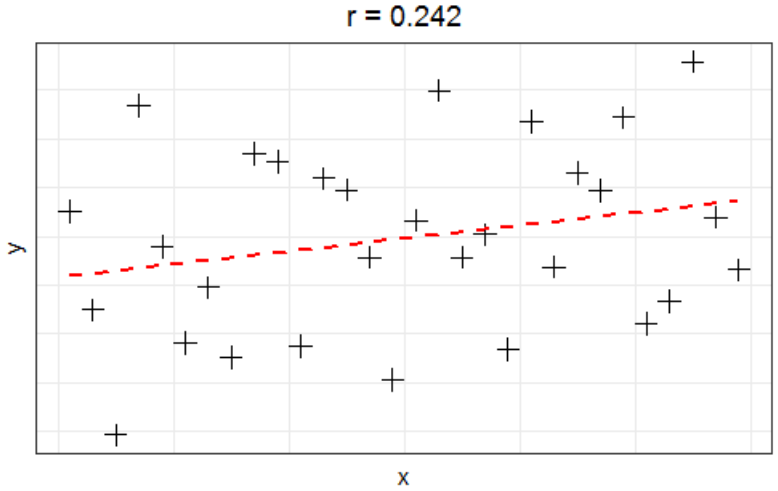
Pearsonův korelační koeficient



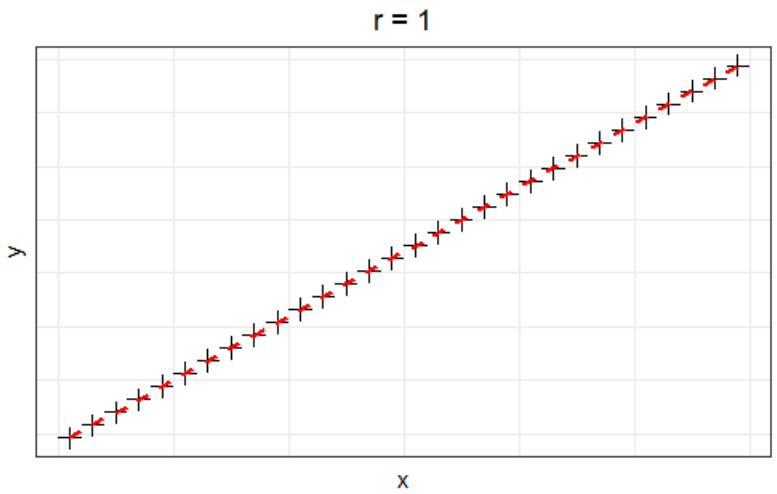


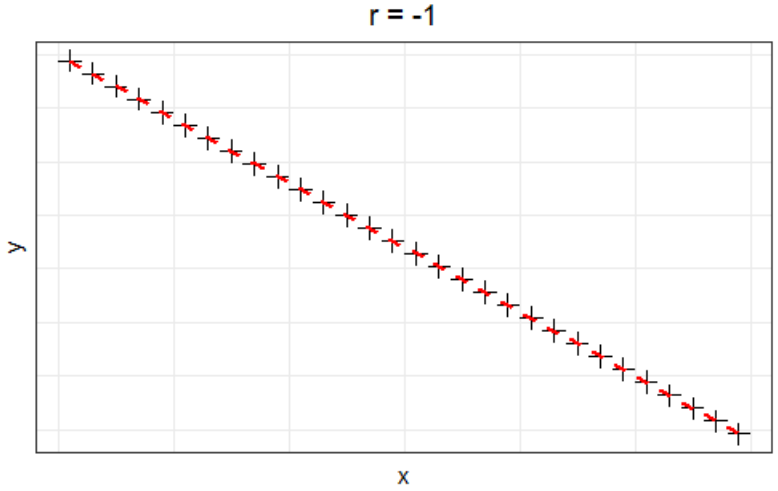


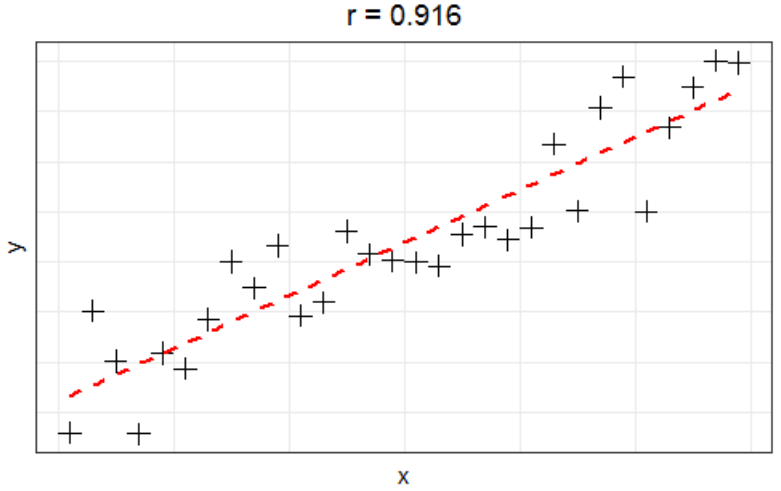


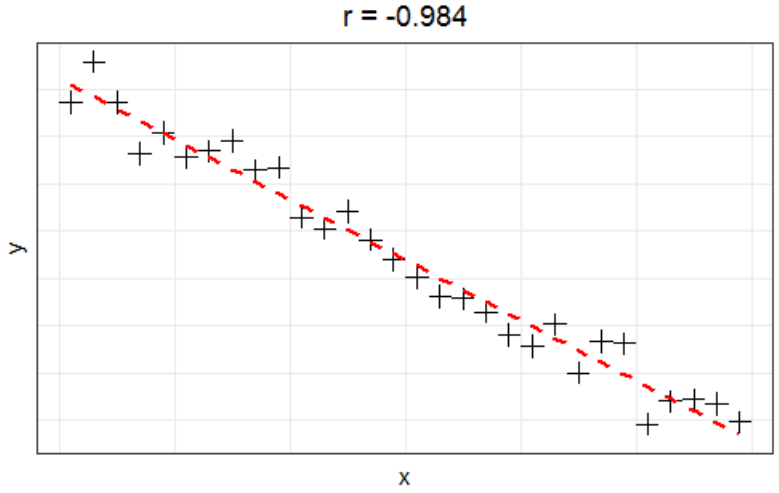


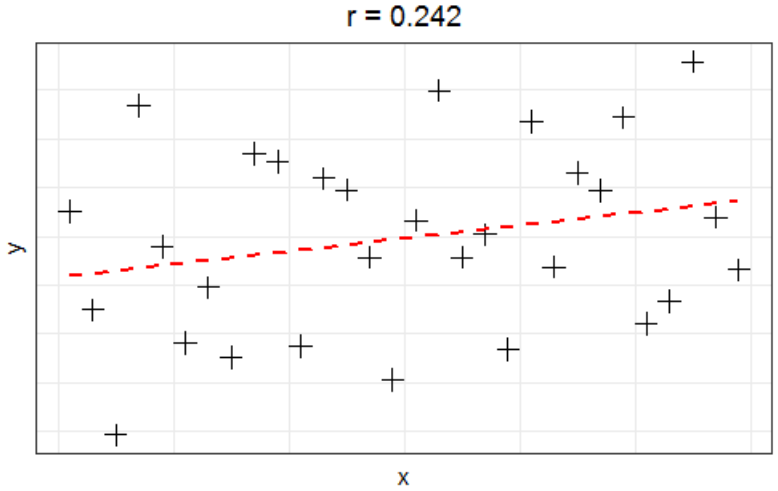
Pearsonův korelační koeficient

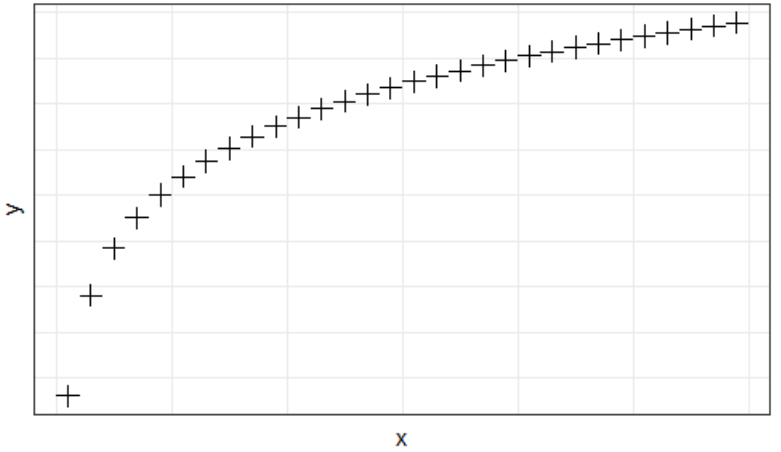




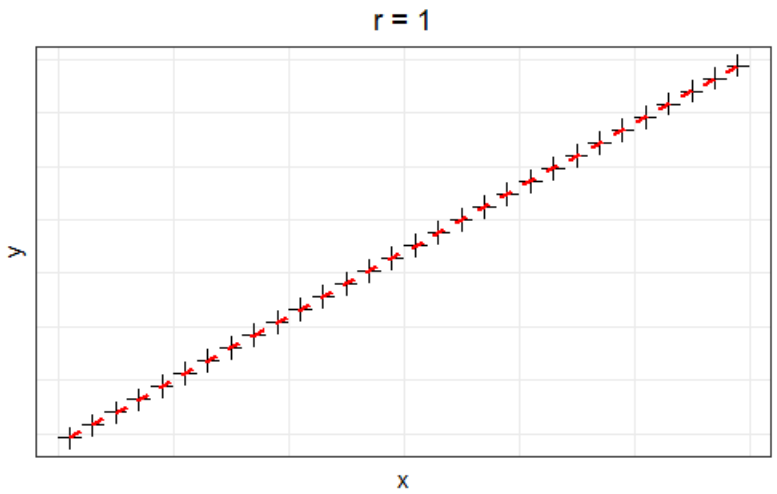


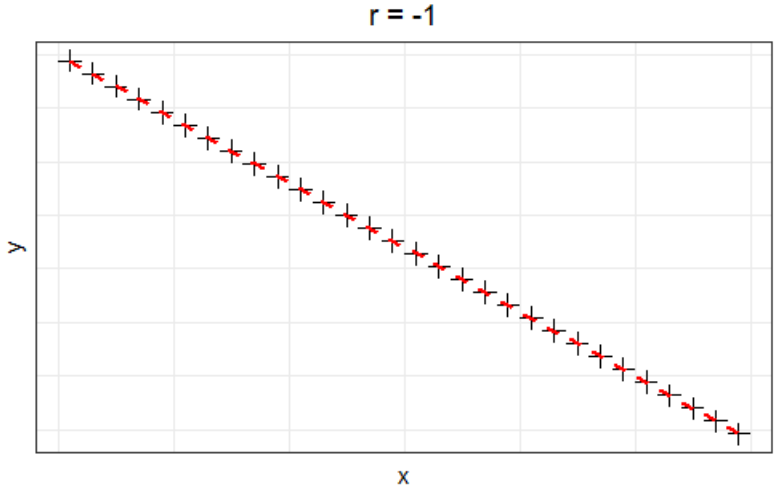


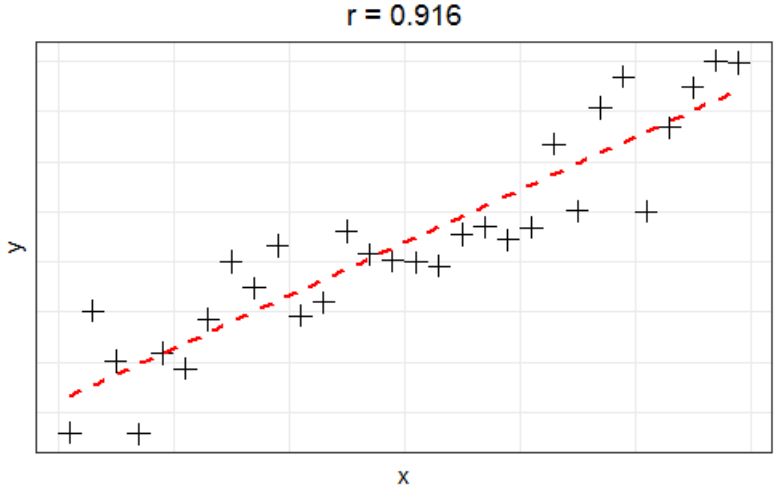


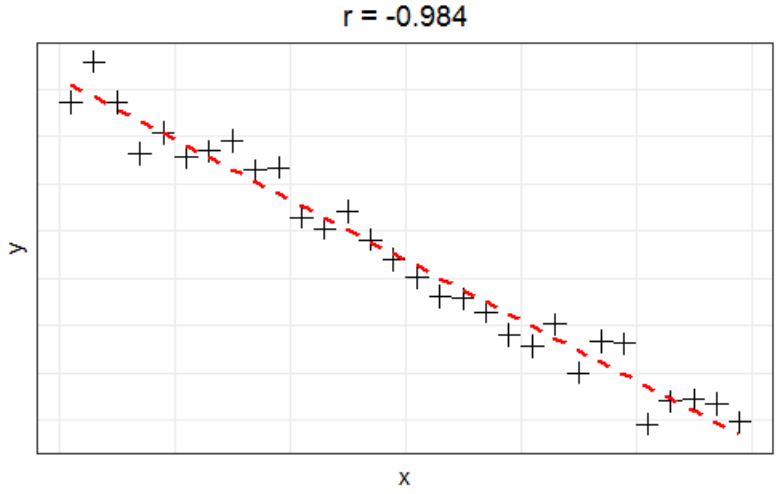


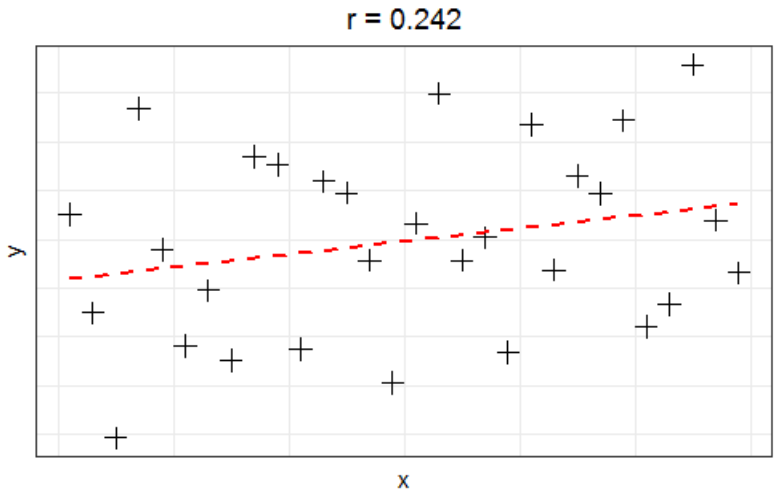
Pearsonův korelační koeficient

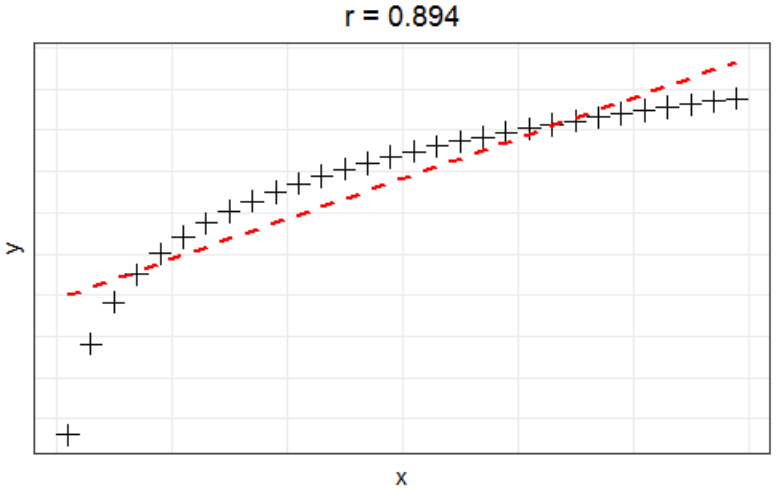




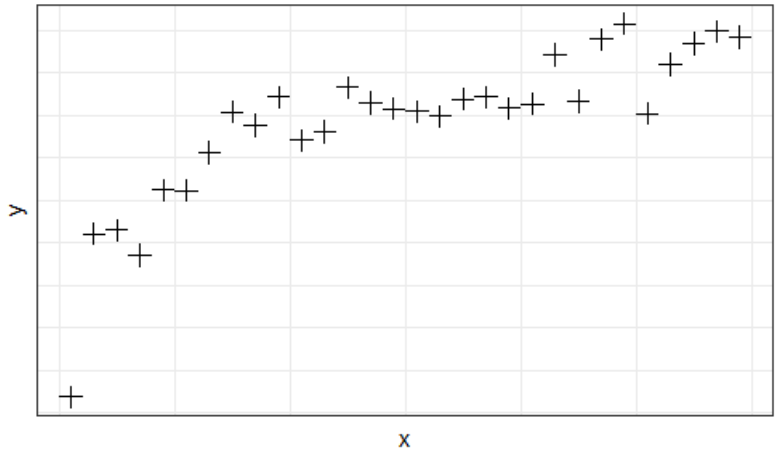




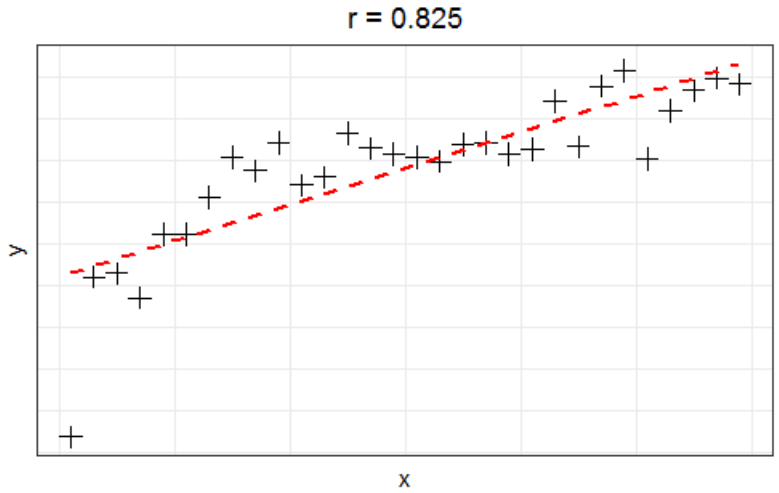




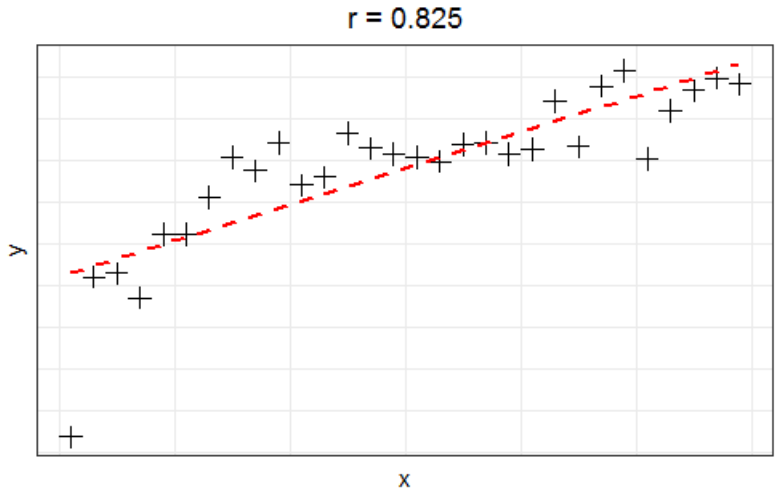
Pearsonův korelační koeficient

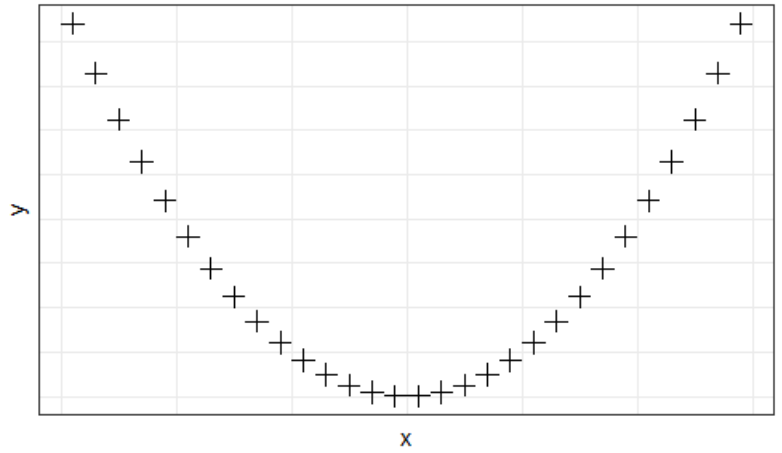


Pearsonův korelační koeficient

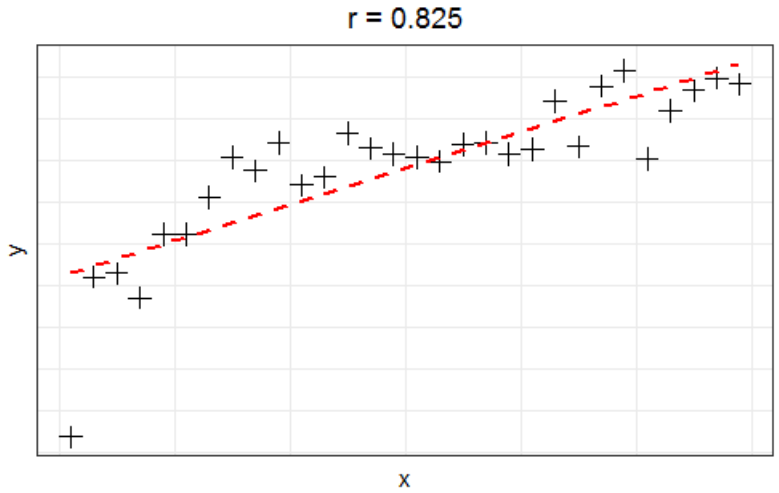


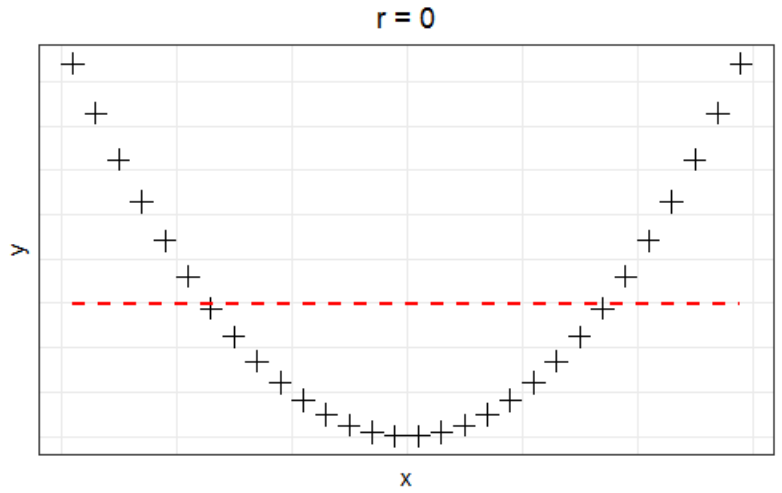
Pearsonův korelační koeficient



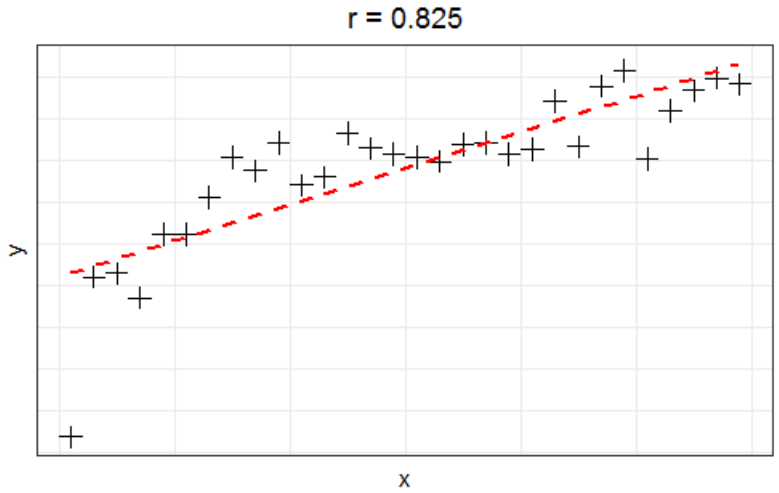


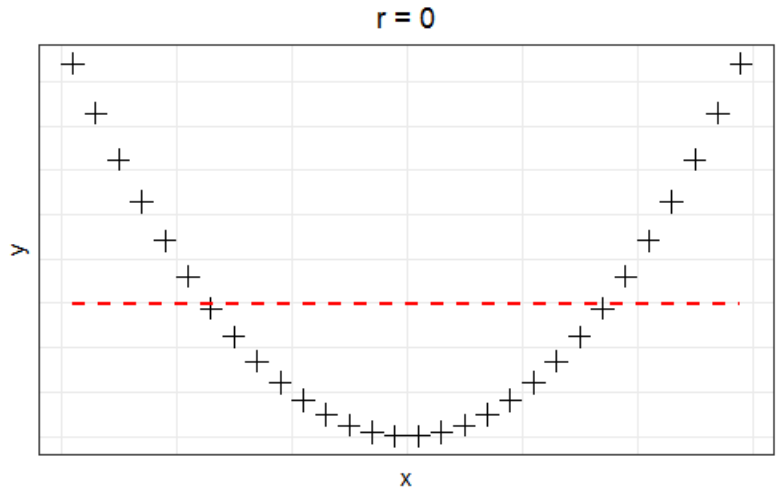
Pearsonův korelační koeficient

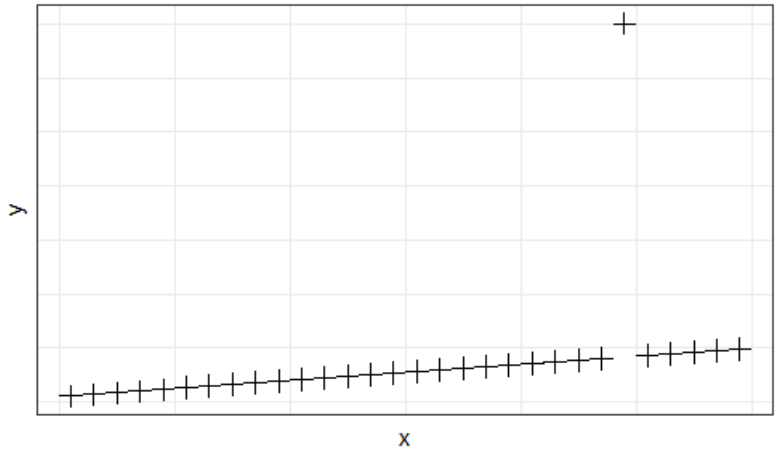




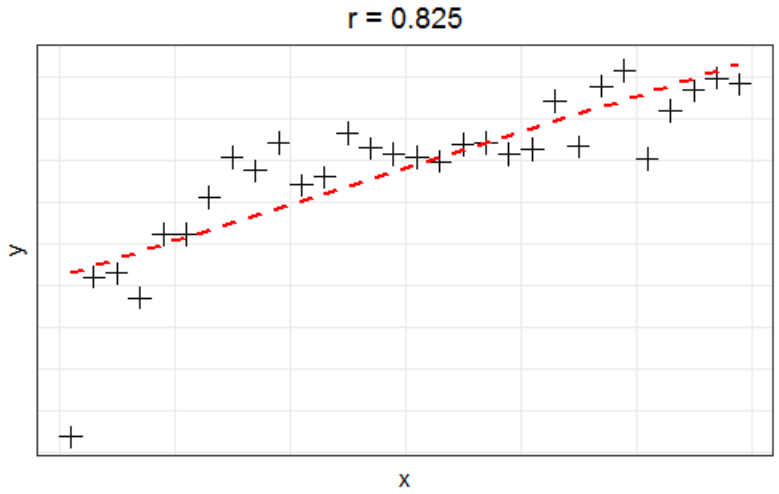
Pearsonův korelační koeficient

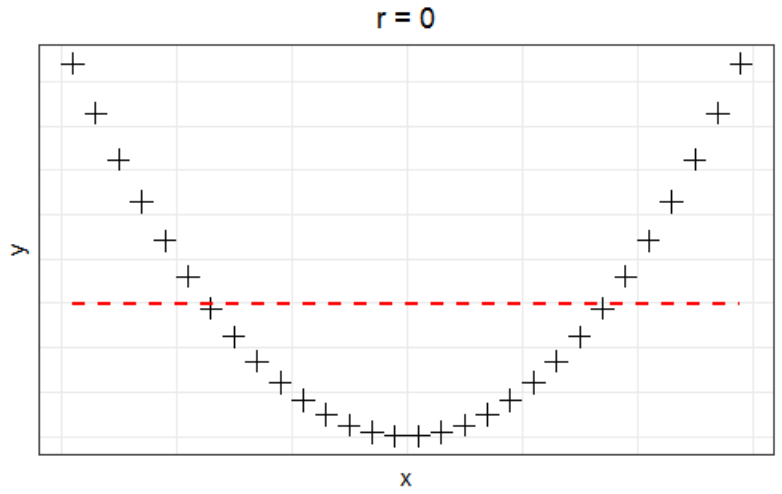


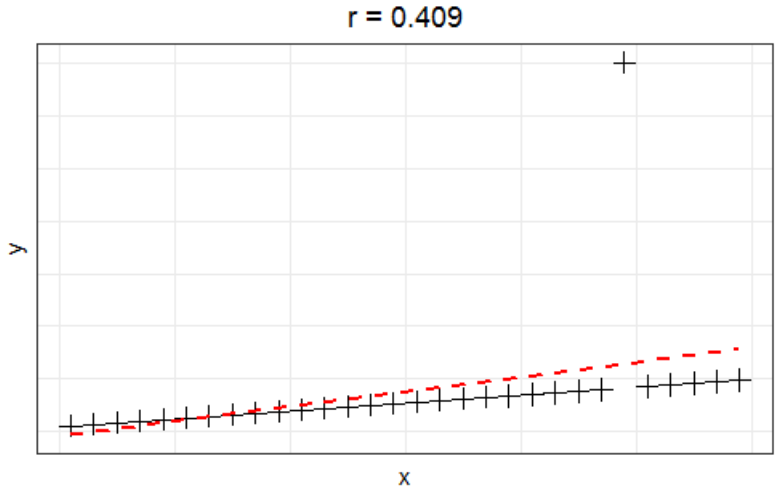


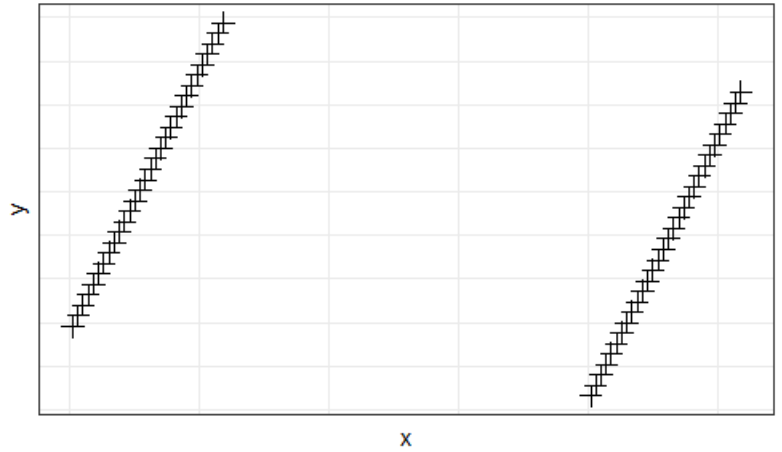


Pearsonův korelační koeficient

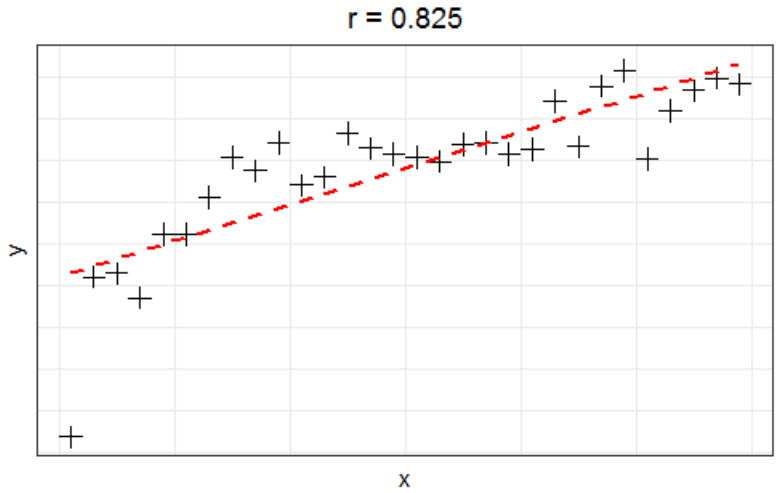


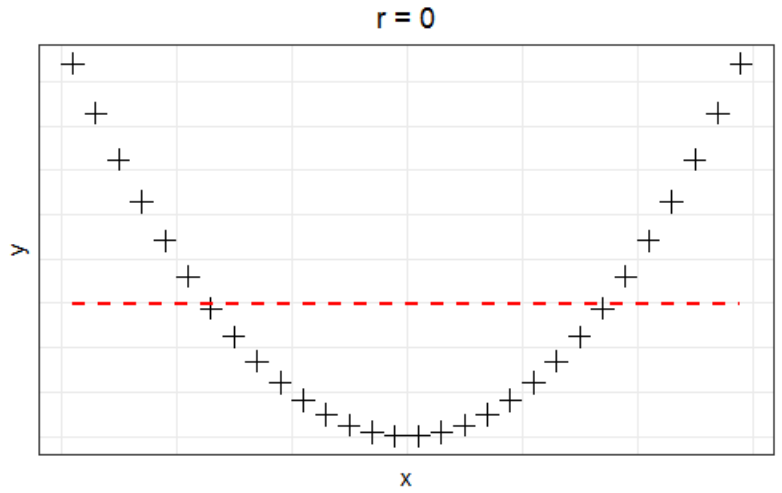


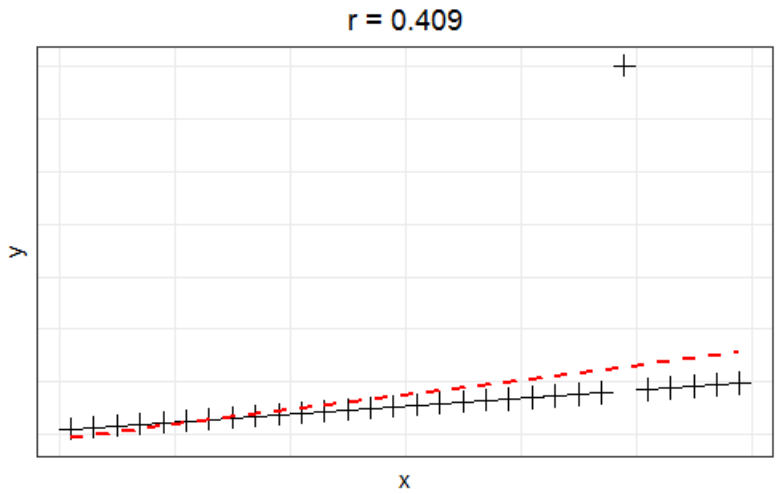


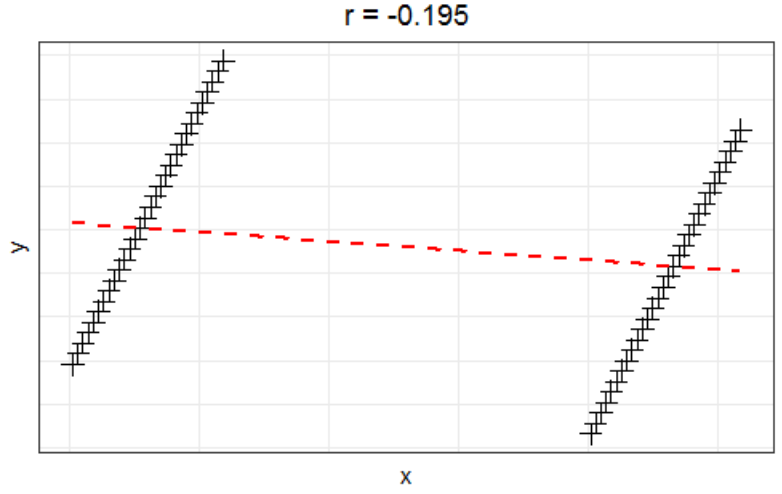


Pearsonův korelační koeficient









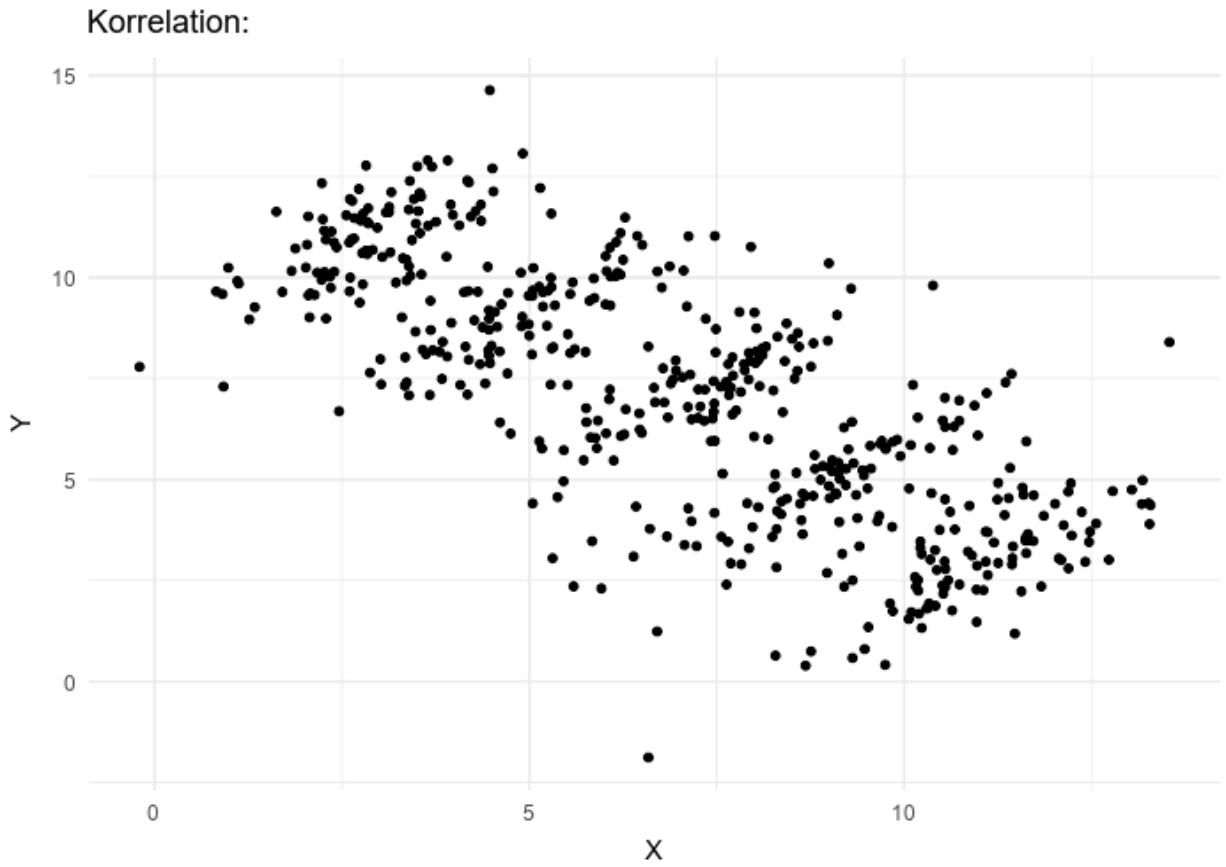
Simpsonův paradox

- Simpsonův paradox je jev v oblasti pravděpodobnosti a statistiky, kdy lze při analýze jednotlivých sledovaných skupin (tříd) pozorovat jistý trend, který ale zmizí, nebo se zcela obrátí ve chvíli, kdy skupiny spojíme.

- Jinými slovy, jedná se o situaci, kdy se závislost mezi dvěma znaky kvalitativně změní, jestliže uvážíme vliv znaku třetího (skrytého, tzv. confounderu).

- Důvodem je silná závislost mezi jedním z dvou analyzovaných znaků a confounderem.

Simpsonův paradox



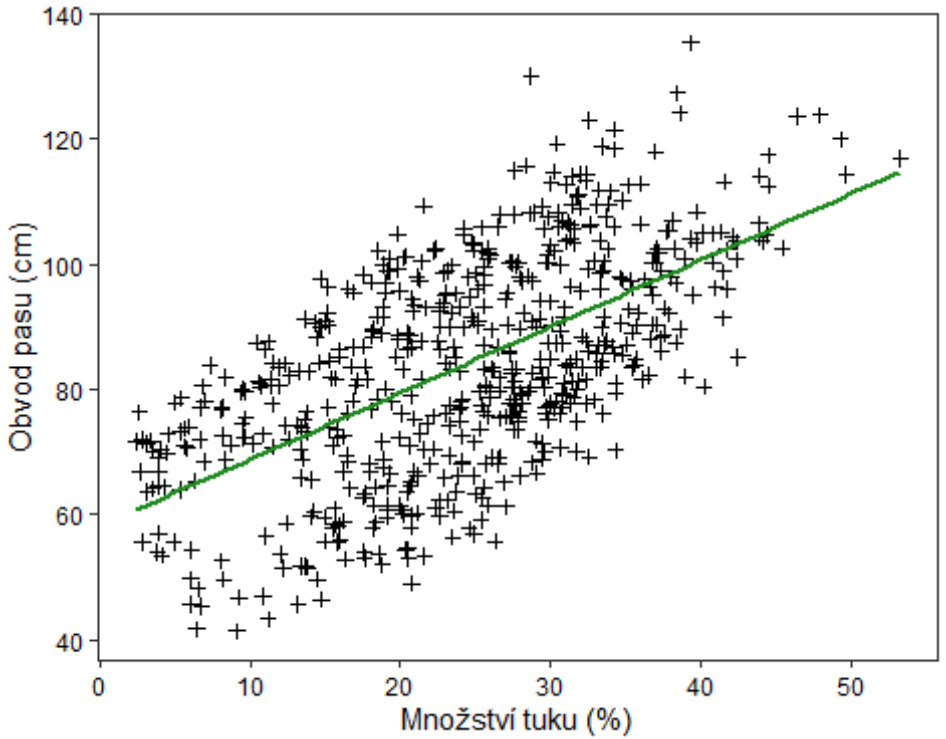
Zdroj: By Pace~svwiki - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=62007681

Spearmanův korelační koeficient

- Mějme náhodný výběr $(X_1; Y_1), \dots, (X_n; Y_n)$ z dvourozměrného rozdělení. Nechť $R_{X_1}, \dots, R_{X_n}$ jsou pořadí veličin $X_1, \dots, X_n$, nechť $R_{Y_1}, \dots, R_{Y_n}$ jsou pořadí veličin $Y_1, \dots, Y_n$.

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} \left(R_{X_i} - R_{Y_i}\right)^2$$

Spearmanův korelační koeficient

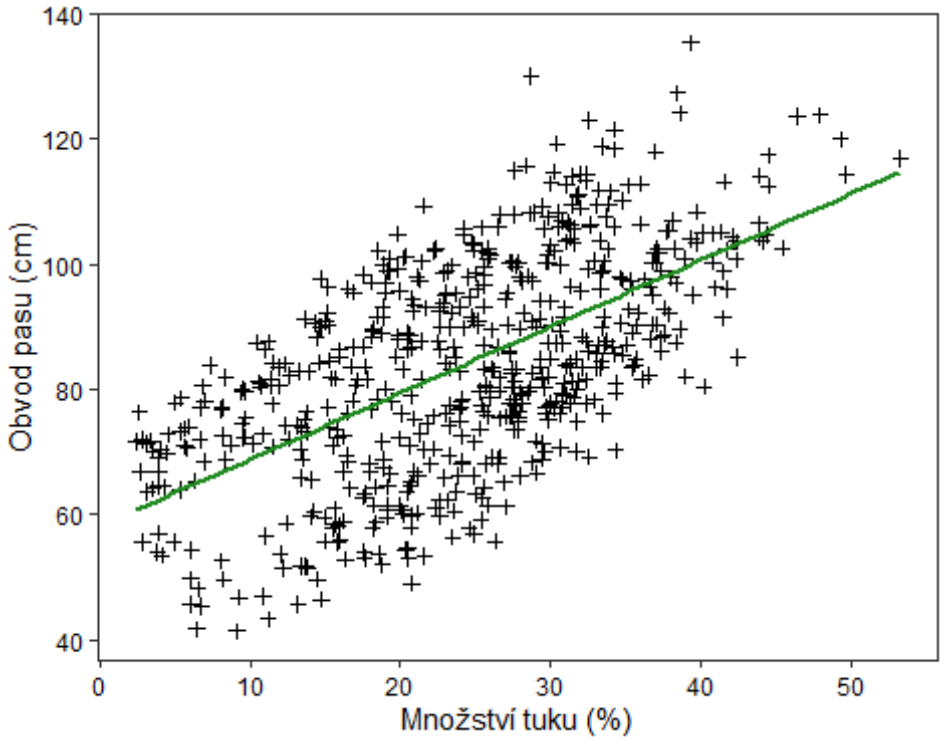


r_S = ?

Vlastnosti Spearmanova korelačního koeficientu

- $-1 \leq r_S(X,Y) \leq 1$
- $r_S(X,Y) = r_S(Y,X)$
- $r_S(X,X) = 1$
- Je-li $r_S(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_S(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_S(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_S(X,Y)| = 1$, pak je mezi X a Y monotónní závislost.

Spearmanův korelační koeficient

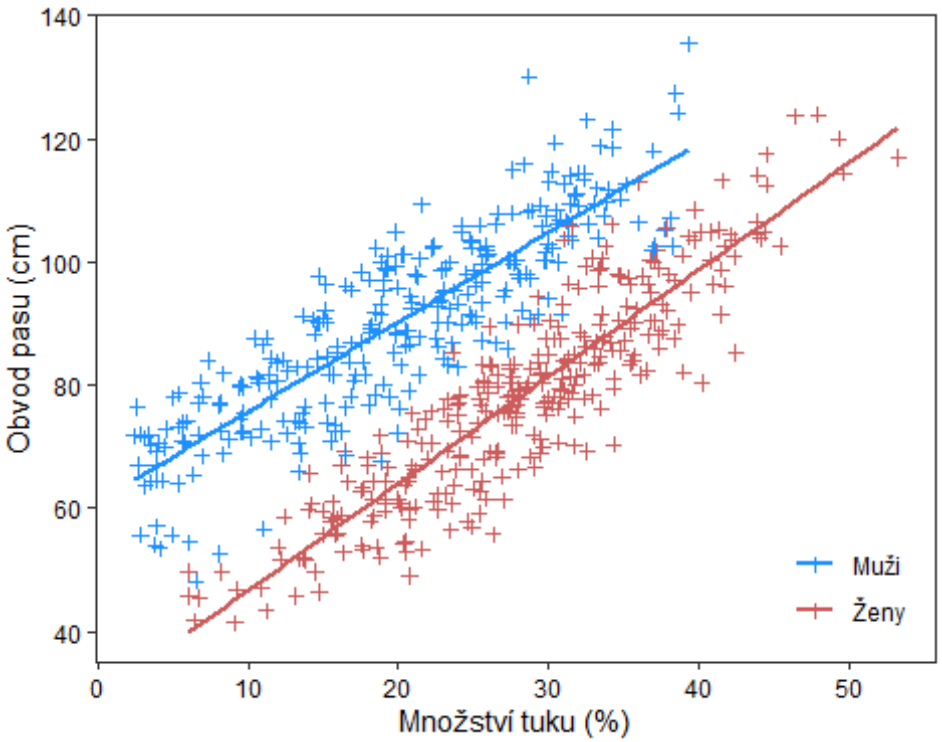


$r_S = 0{,}602$

Vlastnosti Spearmanova korelačního koeficientu

- $-1 \leq r_S(X,Y) \leq 1$
- $r_S(X,Y) = r_S(Y,X)$
- $r_S(X,X) = 1$
- Je-li $r_S(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_S(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_S(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_S(X,Y)| = 1$, pak je mezi X a Y monotónní závislost.

Spearmanův korelační koeficient



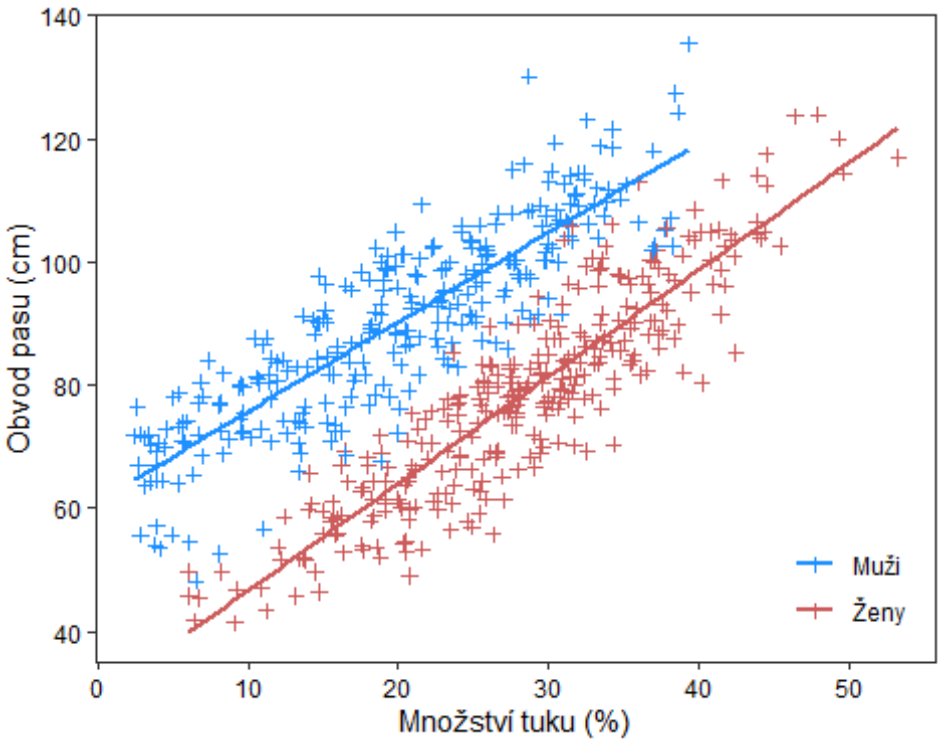
r_S = ?

r_S = ?

Vlastnosti Spearmanova korelačního koeficientu

- $-1 \leq r_S(X,Y) \leq 1$
- $r_S(X,Y) = r_S(Y,X)$
- $r_S(X,X) = 1$
- Je-li $r_S(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_S(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_S(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_S(X,Y)| = 1$, pak je mezi X a Y monotónní závislost.

Spearmanův korelační koeficient

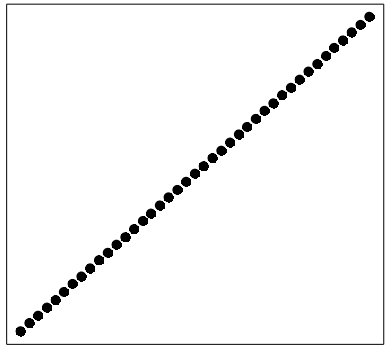


r_S = 0,897 r_S = 0,878

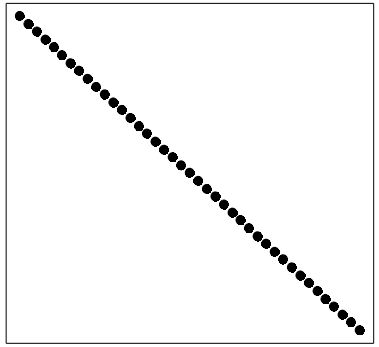
Vlastnosti Spearmanova korelačního koeficientu

- $-1 \leq r_S(X,Y) \leq 1$
- $r_S(X,Y) = r_S(Y,X)$
- $r_S(X,X) = 1$
- Je-li $r_S(X,Y) = 0$, říkáme, že X, Y jsou nekorelované znaky.
- Je-li $r_S(X,Y) > 0$, říkáme, že X, Y jsou pozitivně korelované (s rostoucím X roste Y).
- Je-li $r_S(X,Y) < 0$, říkáme, že X, Y jsou negativně korelované (s rostoucím X klesá Y).
- Je-li $|r_S(X,Y)| = 1$, pak je mezi X a Y monotónní závislost.

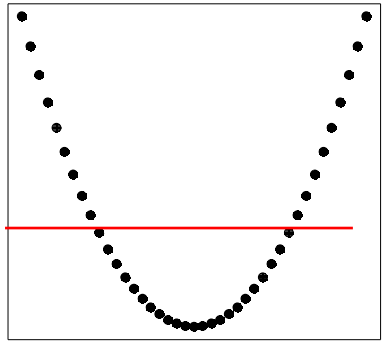
Srovnání Pearsonova a Spearmanova korelačního koeficientu



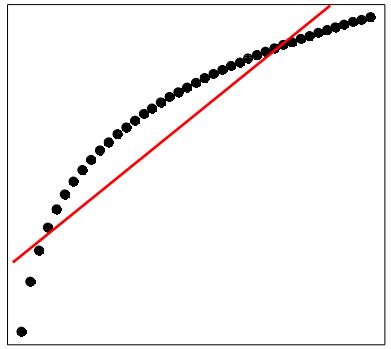
$\rho(X,Y) = 1{,}000$



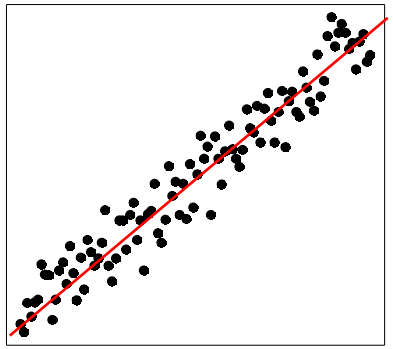
$\rho(X,Y) = -1{,}000$



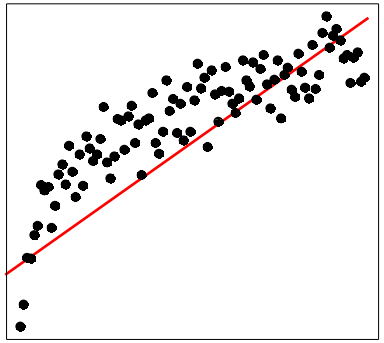
$\rho(X,Y) = 0{,}000$



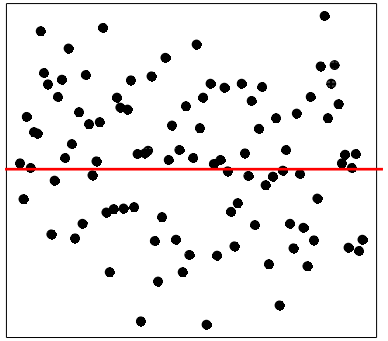
$\rho(X,Y) = 0{,}934$



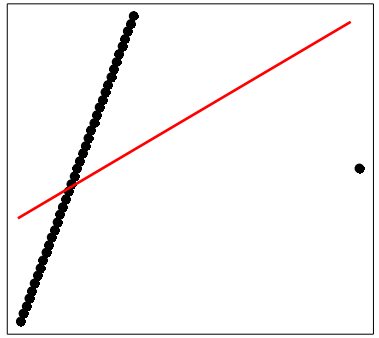
$\rho(X,Y) = 0{,}967$



$\rho(X,Y) = 0{,}857$

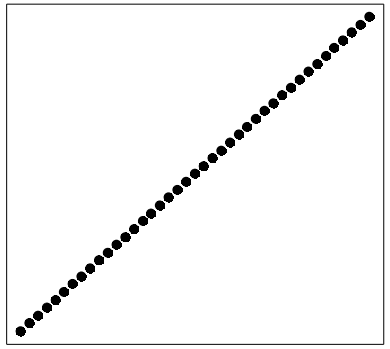


$\rho(X,Y) = -0{,}143$

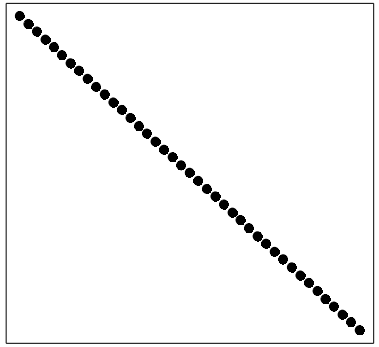


$\rho(X,Y) = 0{,}608$

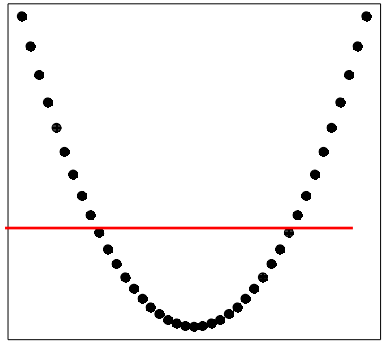
Srovnání Pearsonova a Spearmanova korelačního koeficientu



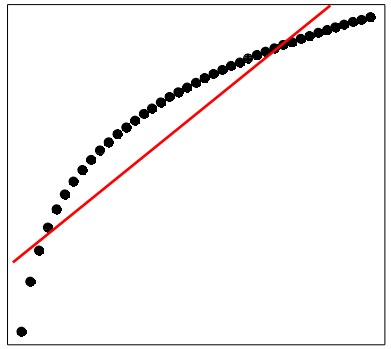
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



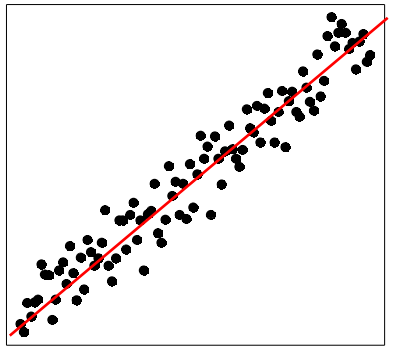
$\rho(X,Y) = -1{,}000$



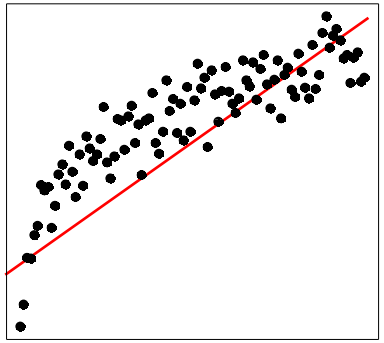
$\rho(X,Y) = 0{,}000$



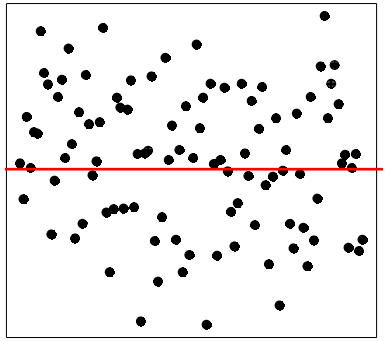
$\rho(X,Y) = 0{,}934$



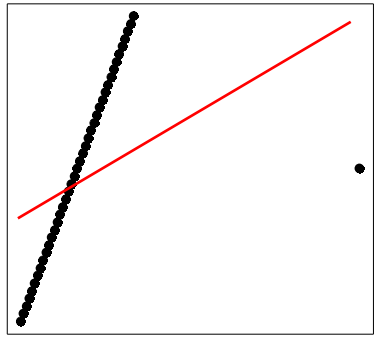
$\rho(X,Y) = 0{,}967$



$\rho(X,Y) = 0{,}857$

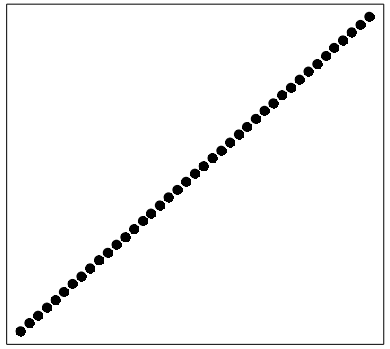


$\rho(X,Y) = -0{,}143$

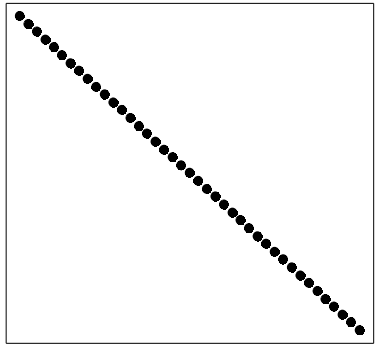


$\rho(X,Y) = 0{,}608$

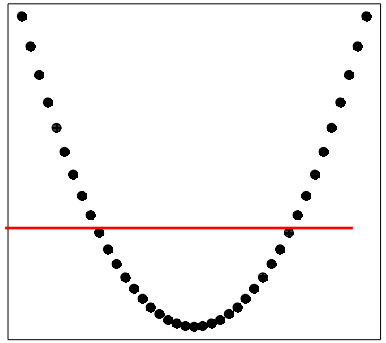
Srovnání Pearsonova a Spearmanova korelačního koeficientu



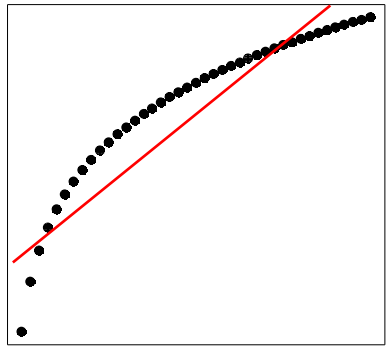
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



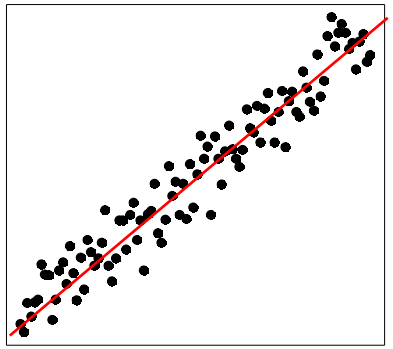
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



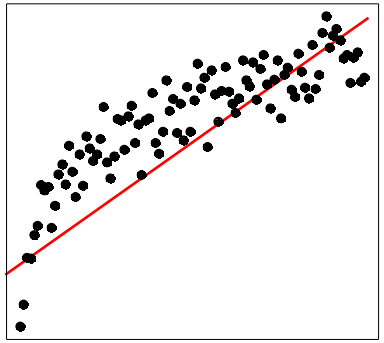
$\rho(X,Y) = 0{,}000$



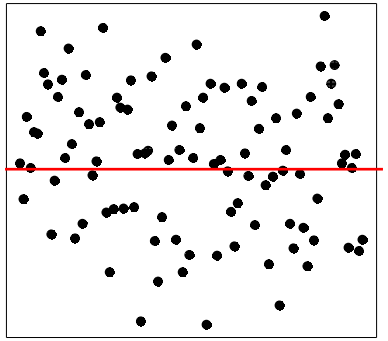
$\rho(X,Y) = 0{,}934$



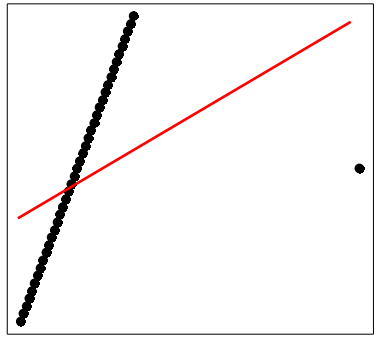
$\rho(X,Y) = 0{,}967$



$\rho(X,Y) = 0{,}857$

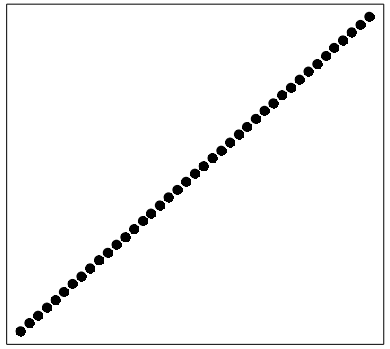


$\rho(X,Y) = -0{,}143$

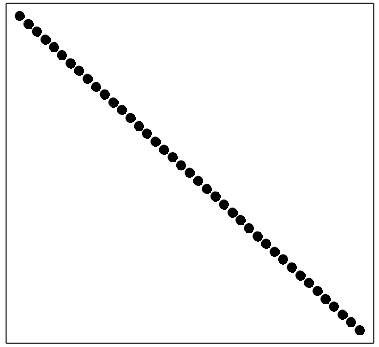


$\rho(X,Y) = 0{,}608$

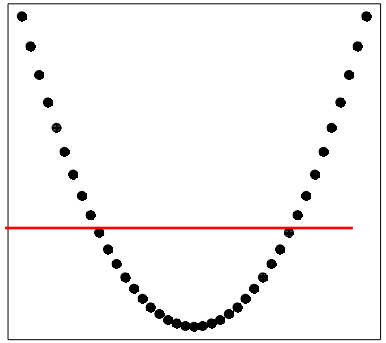
Srovnání Pearsonova a Spearmanova korelačního koeficientu



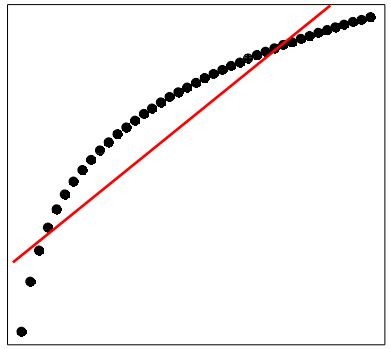
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



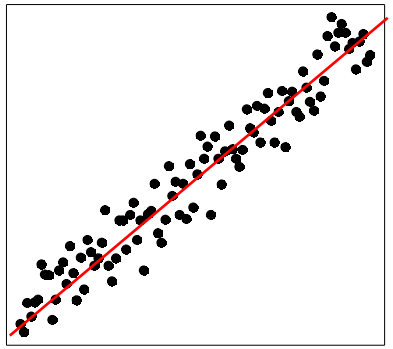
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



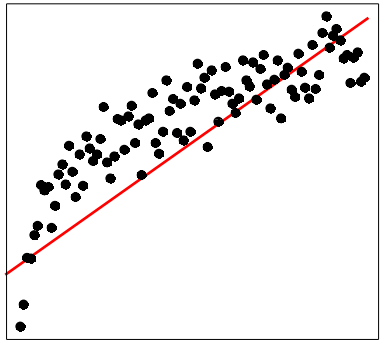
$\rho(X,Y) = 0{,}000$
$\rho_S(X,Y) = 0{,}000$



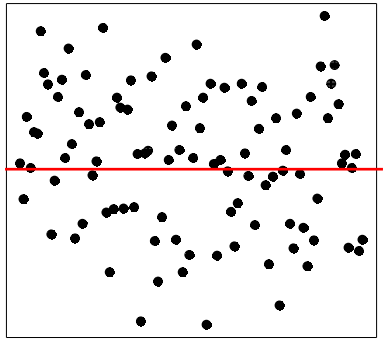
$\rho(X,Y) = 0{,}934$



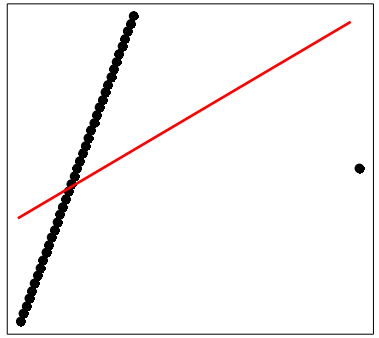
$\rho(X,Y) = 0{,}967$



$\rho(X,Y) = 0{,}857$

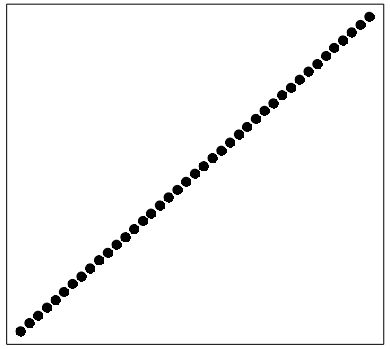


$\rho(X,Y) = -0{,}143$

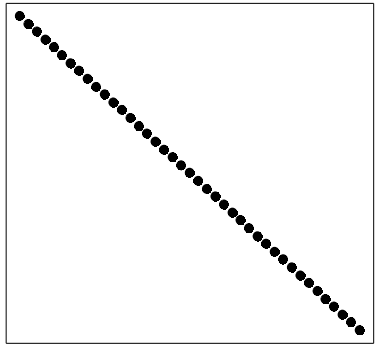


$\rho(X,Y) = 0{,}608$

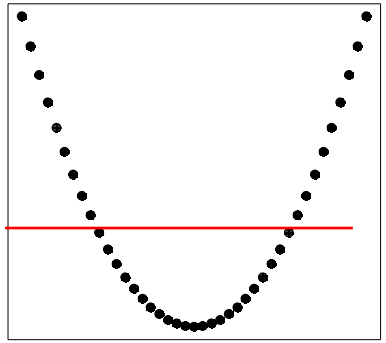
Srovnání Pearsonova a Spearmanova korelačního koeficientu



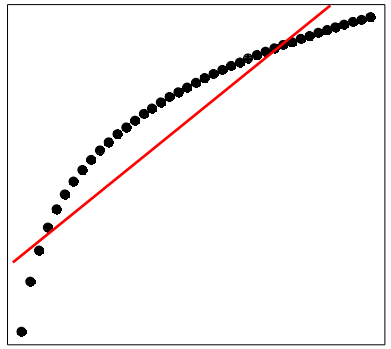
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



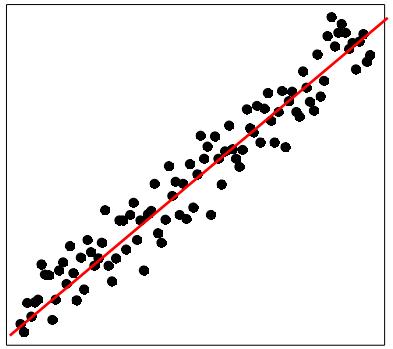
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



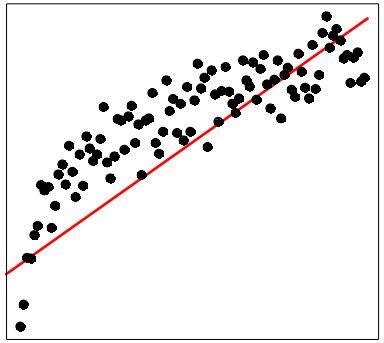
$\rho(X,Y) = 0{,}000$
$\rho_S(X,Y) = 0{,}000$



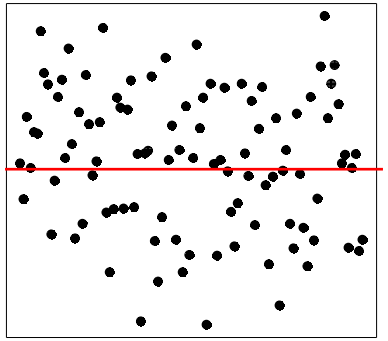
$\rho(X,Y) = 0{,}934$
$\rho_S(X,Y) = 1{,}000$



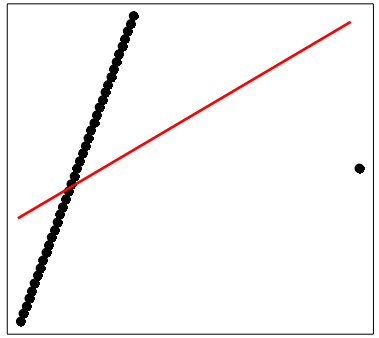
$\rho(X,Y) = 0{,}967$



$\rho(X,Y) = 0{,}857$

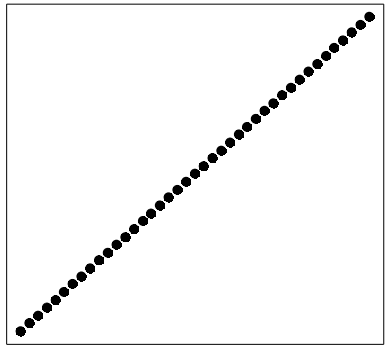


$\rho(X,Y) = -0{,}143$

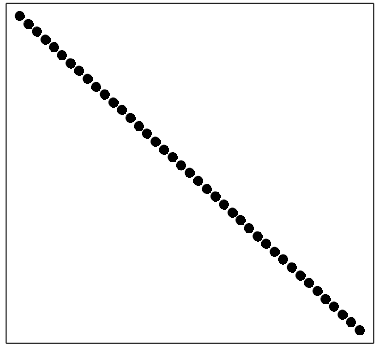


$\rho(X,Y) = 0{,}608$

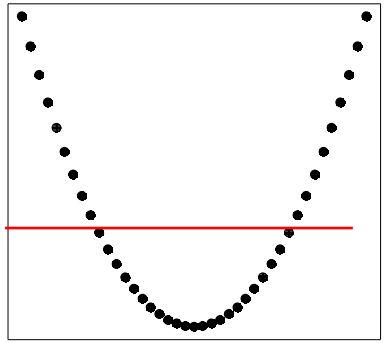
Srovnání Pearsonova a Spearmanova korelačního koeficientu



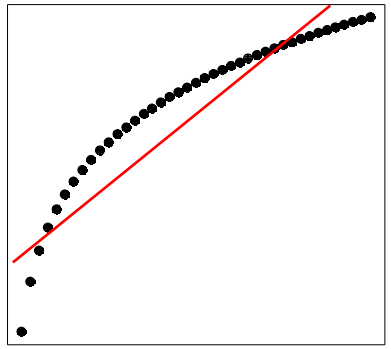
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



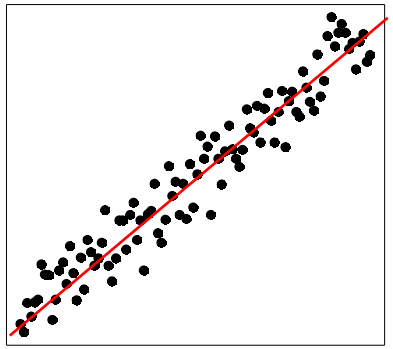
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



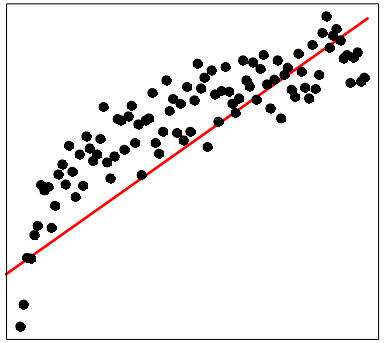
$\rho(X,Y) = 0{,}000$
$\rho_S(X,Y) = 0{,}000$



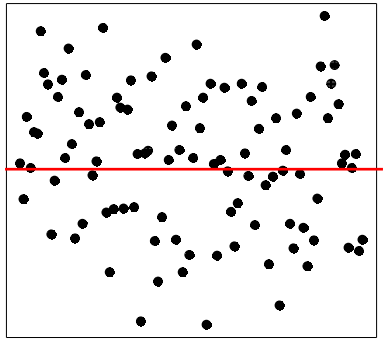
$\rho(X,Y) = 0{,}934$
$\rho_S(X,Y) = 1{,}000$



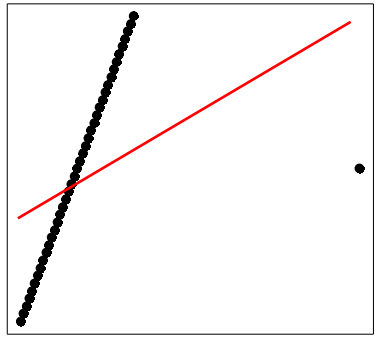
$\rho(X,Y) = 0{,}967$
$\rho_S(X,Y) = 0{,}981$



$\rho(X,Y) = 0{,}857$

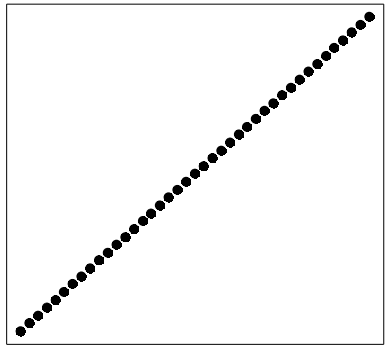


$\rho(X,Y) = -0{,}143$

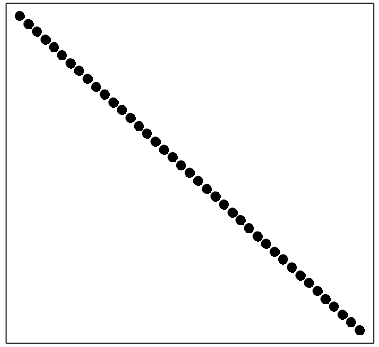


$\rho(X,Y) = 0{,}608$

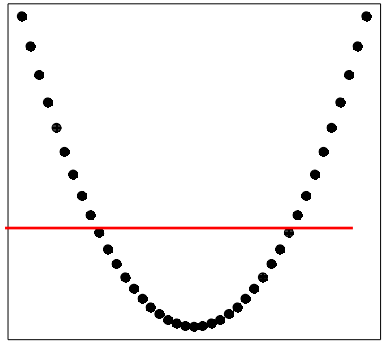
Srovnání Pearsonova a Spearmanova korelačního koeficientu



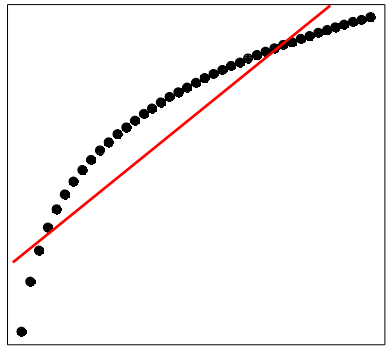
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



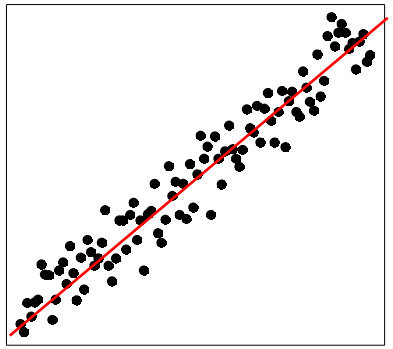
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



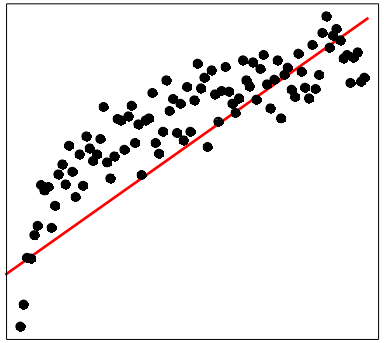
$\rho(X,Y) = 0{,}000$
$\rho_S(X,Y) = 0{,}000$



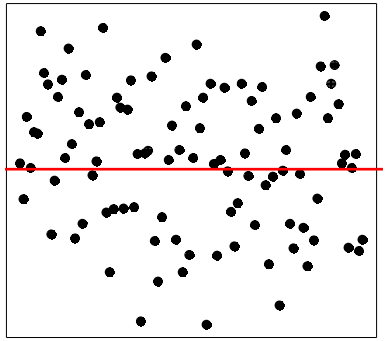
$\rho(X,Y) = 0{,}934$
$\rho_S(X,Y) = 1{,}000$



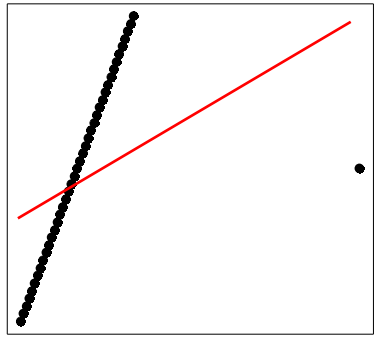
$\rho(X,Y) = 0{,}967$
$\rho_S(X,Y) = 0{,}981$



$\rho(X,Y) = 0{,}857$
$\rho_S(X,Y) = 0{,}893$

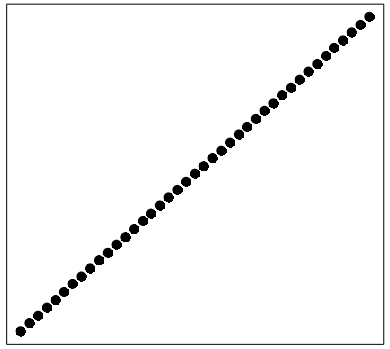


$\rho(X,Y) = -0{,}143$

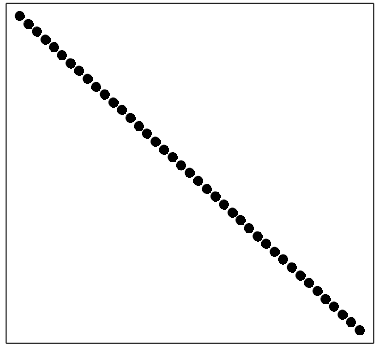


$\rho(X,Y) = 0{,}608$

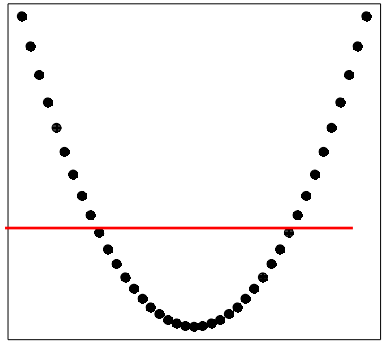
Srovnání Pearsonova a Spearmanova korelačního koeficientu



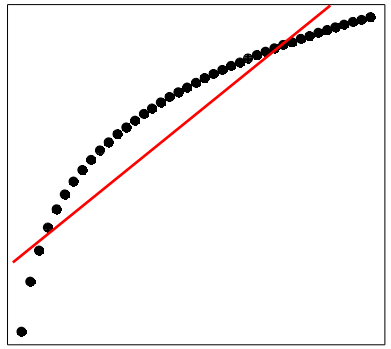
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



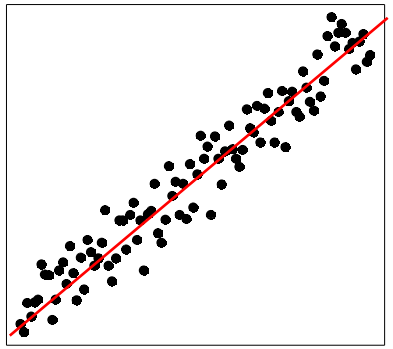
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



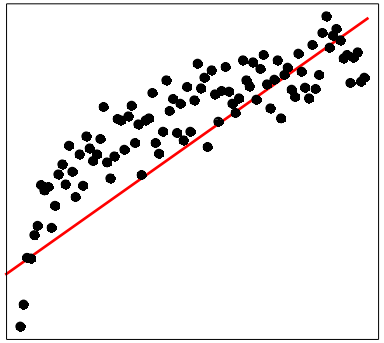
$\rho(X,Y) = 0{,}000$
$\rho_S(X,Y) = 0{,}000$



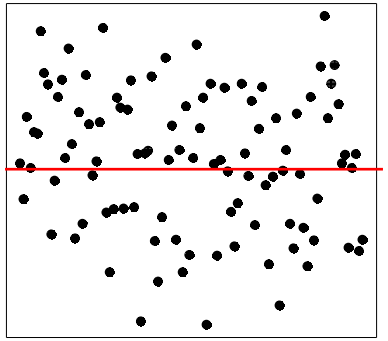
$\rho(X,Y) = 0{,}934$
$\rho_S(X,Y) = 1{,}000$



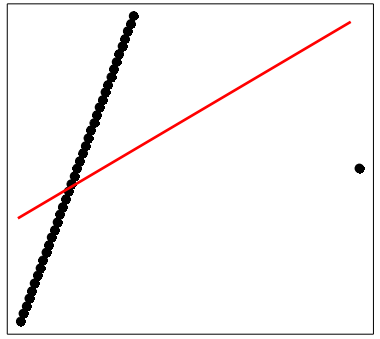
$\rho(X,Y) = 0{,}967$
$\rho_S(X,Y) = 0{,}981$



$\rho(X,Y) = 0{,}857$
$\rho_S(X,Y) = 0{,}893$

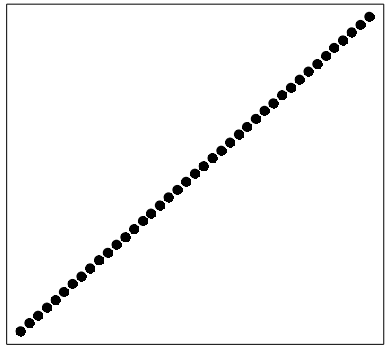


$\rho(X,Y) = -0{,}143$
$\rho_S(X,Y) = -0{,}178$

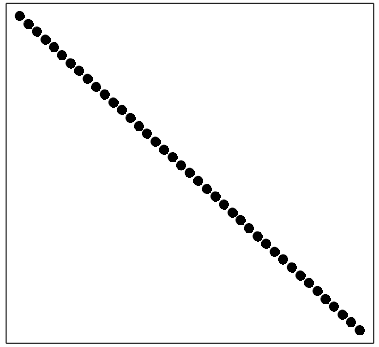


$\rho(X,Y) = 0{,}608$

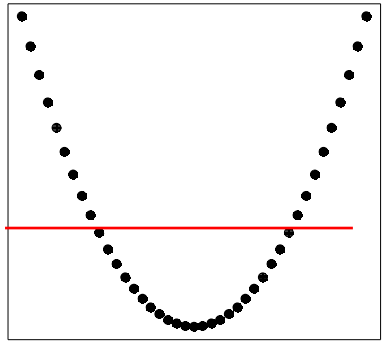
Srovnání Pearsonova a Spearmanova korelačního koeficientu



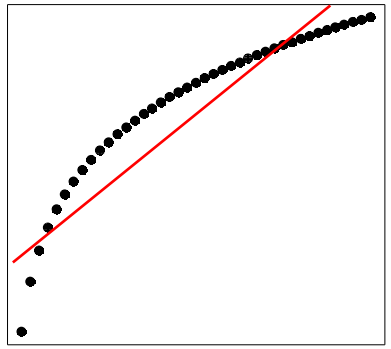
$\rho(X,Y) = 1{,}000$
$\rho_S(X,Y) = 1{,}000$



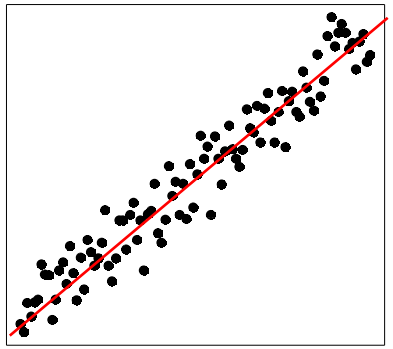
$\rho(X,Y) = -1{,}000$
$\rho_S(X,Y) = -1{,}000$



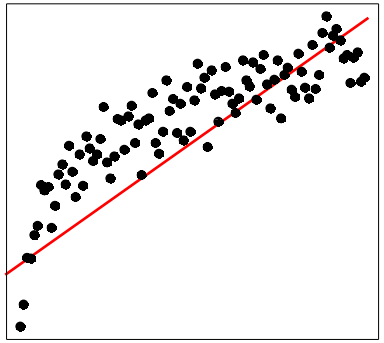
$\rho(X,Y) = 0{,}000$
$\rho_S(X,Y) = 0{,}000$



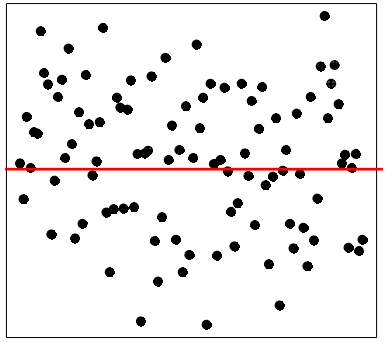
$\rho(X,Y) = 0{,}934$
$\rho_S(X,Y) = 1{,}000$



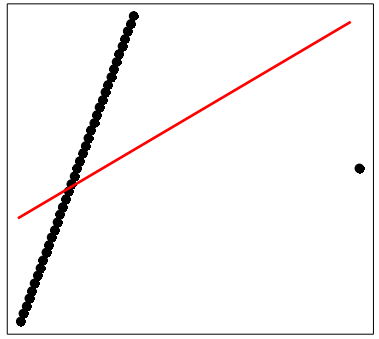
$\rho(X,Y) = 0{,}967$
$\rho_S(X,Y) = 0{,}981$



$\rho(X,Y) = 0{,}857$
$\rho_S(X,Y) = 0{,}893$



$\rho(X,Y) = -0{,}143$
$\rho_S(X,Y) = -0{,}178$



$\rho(X,Y) = 0{,}608$
$\rho_S(X,Y) = 0{,}911$

Analýza závislosti dvou numerických proměnných

ID	Pohlaví	Rasa	Věk (roky)	BMI (kg/m^2)	Množství tuku (%)	Obvod pasu (cm)	Kvalita spánku	Kvalita spánku (dichotomizovaná)
1	muž	Negroidní	61	29,81	22,66	90,1	spíše špatná	špatná
2	žena	Mongoloidní	52	22,50	26,59	79,8	velmi dobrá	dobrá
3	muž	Negroidní	37	24,50	13,75	76,4	spíše dobrá	dobrá
4	žena	Mongoloidní	47	24,04	30,79	87,4	spíše špatná	špatná
5	muž	Europoidní	46	22,56	16,70	83,7	spíše dobrá	dobrá
6	žena	Negroidní	37	19,98	26,18	83,0	velmi dobrá	dobrá
7	žena	Negroidní	44	23,61	35,59	84,0	spíše dobrá	dobrá
8	muž	Mongoloidní	50	20,85	2,77	72,0	spíše dobrá	dobrá
9	muž	Negroidní	50	26,95	21,29	97,5	spíše špatná	špatná

Jsou některé znaky negativně korelované?

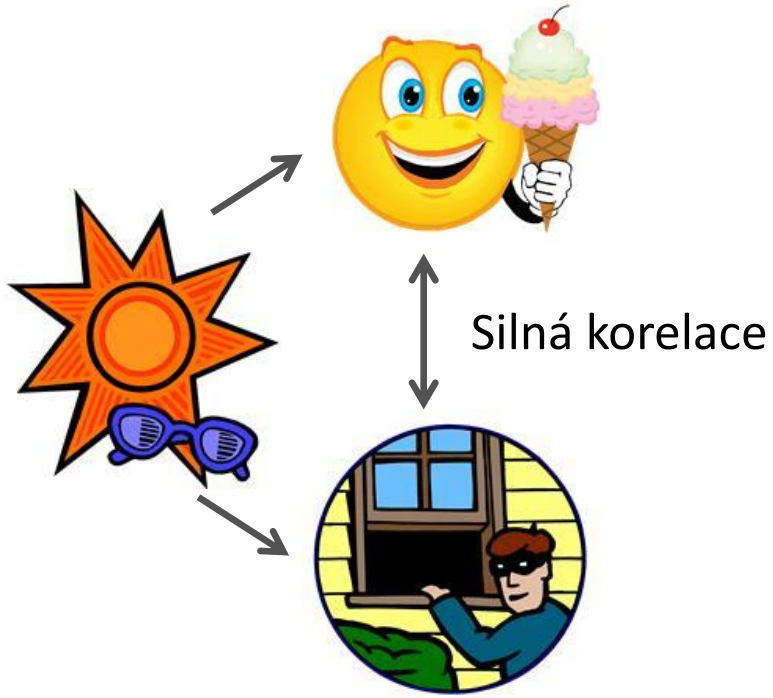
Analýza závislosti dvou numerických proměnných

ID	Pohlaví	Rasa	Věk (roky)	BMI (kg/m^2)	Množství tuku (%)	Obvod pasu (cm)	Kvalita spánku	Kvalita spánku (dichotomizovaná)
1	muž	Negroidní	61	29,81	22,66	90,1	spíše špatná	špatná
2	žena	Mongoloidní	52	22,50	26,59	79,8	velmi dobrá	dobrá
3	muž	Negroidní	37	24,50	13,75	76,4	spíše dobrá	dobrá
4	žena	Mongoloidní	47	24,04	30,79	87,4	spíše špatná	špatná
5	muž	Europoidní	46	22,56	16,70	83,7	spíše dobrá	dobrá
6	žena	Negroidní	37	19,98	26,18	83,0	velmi dobrá	dobrá
7	žena	Negroidní	44	23,61	35,59	84,0	spíše dobrá	dobrá
8	muž	Mongoloidní	50	20,85	2,77	72,0	spíše dobrá	dobrá
9	muž	Negroidní	50	26,95	21,29	97,5	spíše špatná	špatná

Dokázali bychom vymyslet další numerický statistický znak,
který by mohl být **negativně korelovaný** s některým znakem v datech?

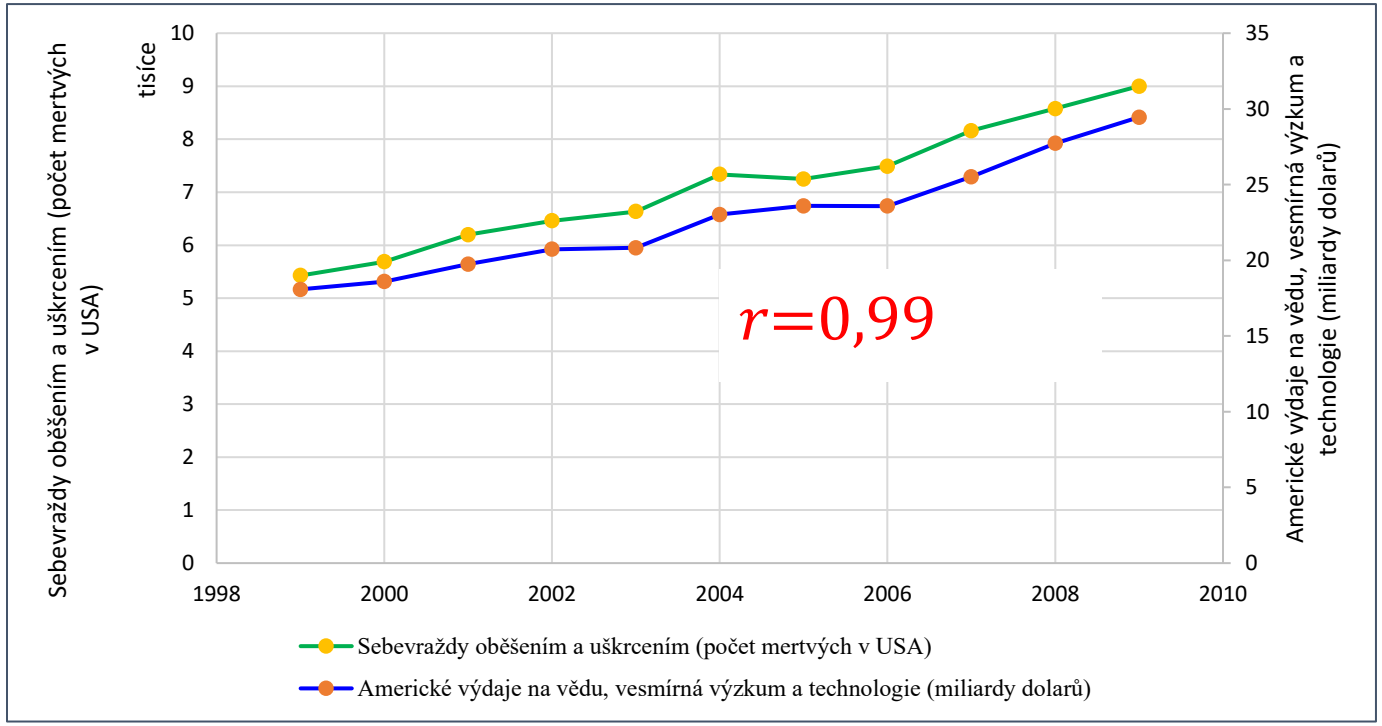
Pozor na falešnou (zdánlivou) korelaci!

- Pokud jsou dvě náhodné veličiny korelované, znamená to pouze to, že jsou lineárně závislé (příp. je mezi nimi monotónní závislost).
- Nelze z toho však ještě usoudit, že by jedna z nich musela být příčinou a druhá následkem.
- To samotná korelovanost nedovoluje rozhodnout.



Pozor na falešnou (zdánlivou) korelaci!

- Pokud jsou dvě náhodné veličiny korelované, znamená to pouze to, že jsou lineárně závislé (příp. je mezi nimi monotónní závislost).
- Nelze z toho však ještě usoudit, že by jedna z nich musela být příčinou a druhá následkem.
- To samotná korelovanost nedovoluje rozhodnout.



Pozor na falešnou (zdánlivou) korelaci!

ZPRÁVY / ZAHRANIČÍ

K Nobelově ceně dopomáhá čokoláda, naznačuje studie

12. 10. 2012 10:36 AKTUALIZOVÁNO

Mezi počtem nobelistů v přepočtu na obyvatele a konzumaci čokolády je souvislost

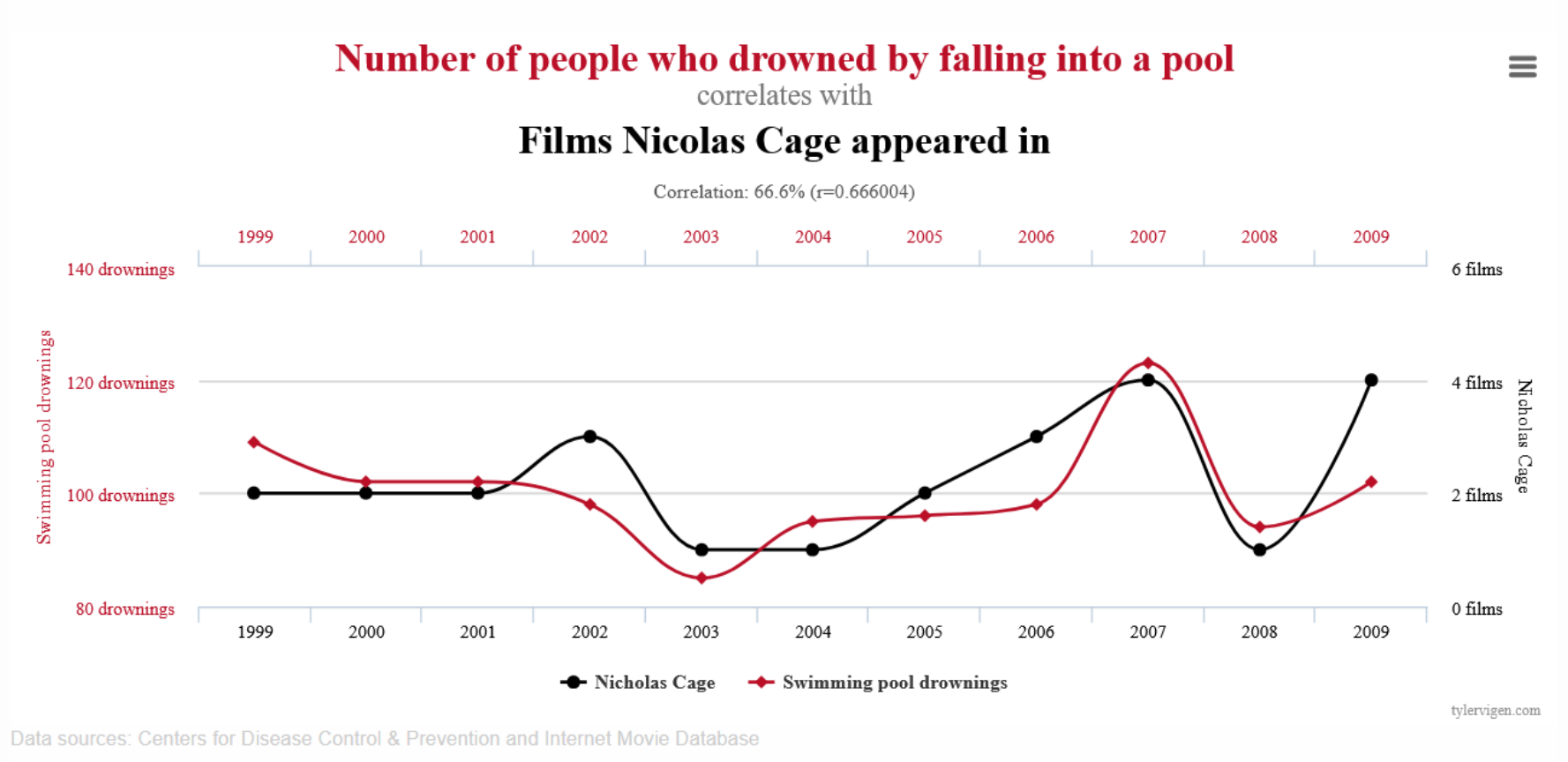
New York - Počet nositelů Nobelovy ceny v přepočtu na jednoho obyvatele se v jednotlivých zemích odvíjí od spotřeby čokolády.

To není úryvek z reklamy na sladkosti, ale závěr studie publikované v jednom z nejprestižnějších světových lékařských časopisů New England Journal of Medicine.

Nejvíce laureátů Nobelovy ceny mají Švýcaři, kteří jsou zároveň největšími jedlíky oblíbené sladké pochutiny.

Zdroj: http://zpravy.aktualne.cz/zahranici/k-nobelove-cene-dopomaha-cokolada-naznacuje-studie/r~i:article:760147/

Pozor na falešnou (zdánlivou) korelaci!



Zdroj: https://www.tylervigen.com/spurious-correlations

Pozor na hodnocení „síly“ korelace!

V praxi se zpravidla hodnota korelačního koeficientu interpretuje takto:

Korelační koeficient	**Typ lineární závislosti**
$\|r\| = 0{,}0$	neexistující
$\|r\| \in (0{,}0; 0{,}3\rangle$	velmi slabá
$\|r\| \in (0{,}3; 0{,}7\rangle$	středně silná
$\|r\| \in (0{,}7; 1{,}0)$	těsná
$\|r\| = 1{,}0$	funkční

- Mezi proudem a napětím na odporu byl zjištěn korelační koeficient 0,6.
- Mezi školním prospěchem a pocitem deprese u dětí byl zjištěn korelační koeficient 0,6.

Výsledky interpretujte!

Shrnutí

Pearsonův korelační koeficient

- Je mírou lineární závislosti.

Spearmanův korelační koeficient

- Je mírou monotónní závislosti.

- Vizualizace je nezbytnou součástí korelační analýzy!

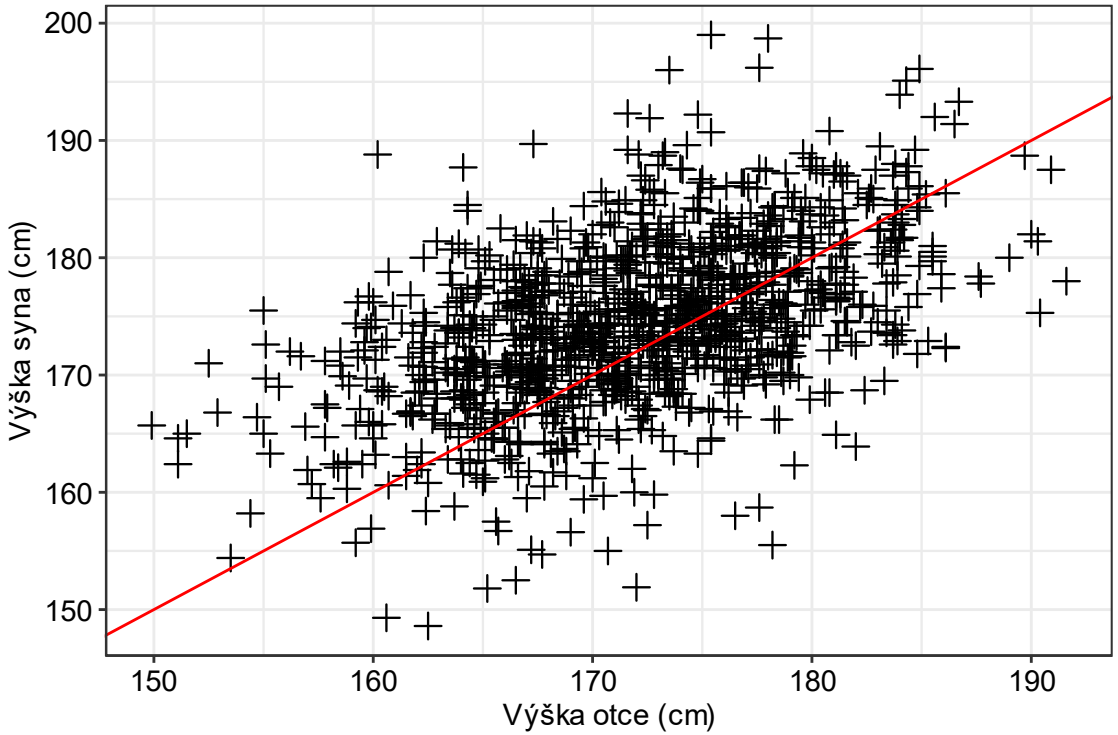
Regrese z pohledu statistiky

Co je to regrese?

- **Regrese nemoci** – stav, kdy nemoc ustoupila a vykazuje jen lehčí příznaky, může se však zase vrátit jako recidiva nemoci.
- **Regresní hypnóza** – vracení se do minulosti, případně i před dobu narození do minulých životů.
- **Regrese (psychologie)** – návrat k již překonaným, méně dospělým či dětštějším formám chování jako jeden z obranných mechanismů, nebo řízené vzpomínání, kdy se pacient pod vedením terapeuta vrací do minulosti.
- **Softwarová regrese** – oprava stavu, kdy úpravou zdrojového kódu přestalo fungovat něco, co do úpravy fungovalo v pořádku.
- **Regresní analýza (statistika)** – soubor parametrických i neparametrických statistických metod, které odhadují hodnotu náhodné veličiny na základě znalosti hodnot jiných veličin.

Historické okénko

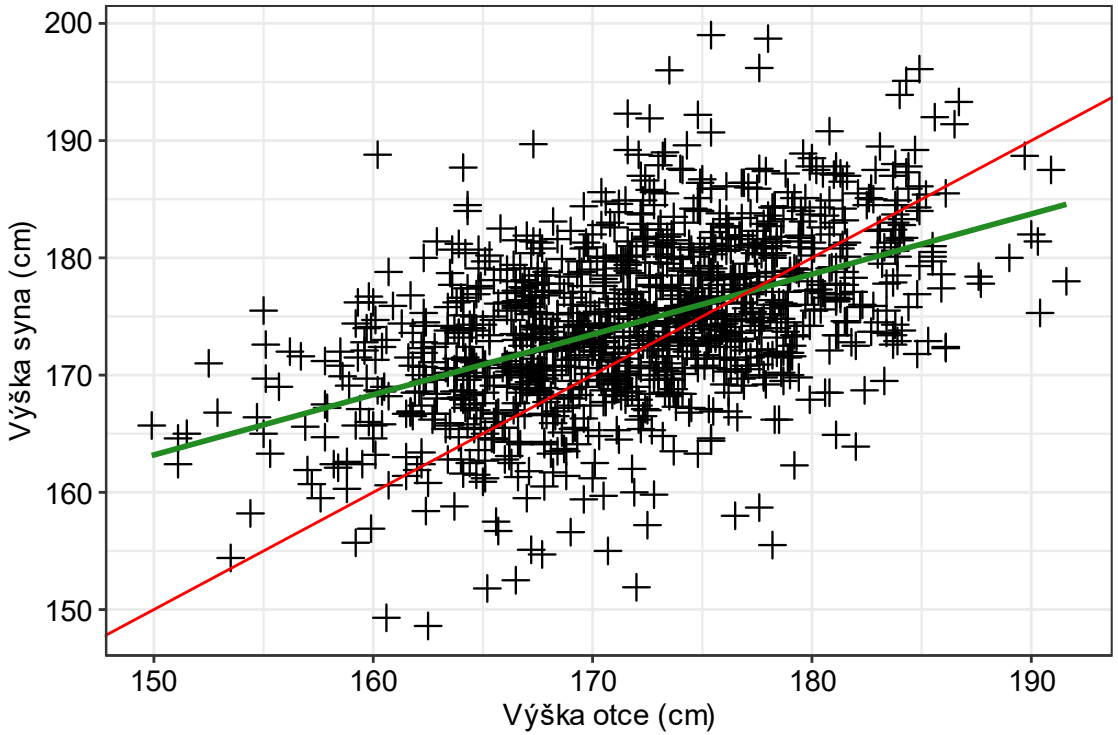
- Francis Galton (1822 – 1911) – zakladatel regresní analýzy
- Datový soubor father.son (R) obsahuje záznamy o výšce 1078 otců a jejich synů
- Původní očekávání: syn bude cca stejně vysoký jako otec



průměrná výška synů: *174,5 cm*

Historické okénko

- Francis Galton (1822 – 1911) – zakladatel regresní analýzy
- Datový soubor father.son (R) obsahuje záznamy o výšce 1078 otců a jejich synů
- Regrese k průměru (původní očekávání: syn bude cca stejně vysoký jako otec)



průměrná výška synů: *174,5 cm*

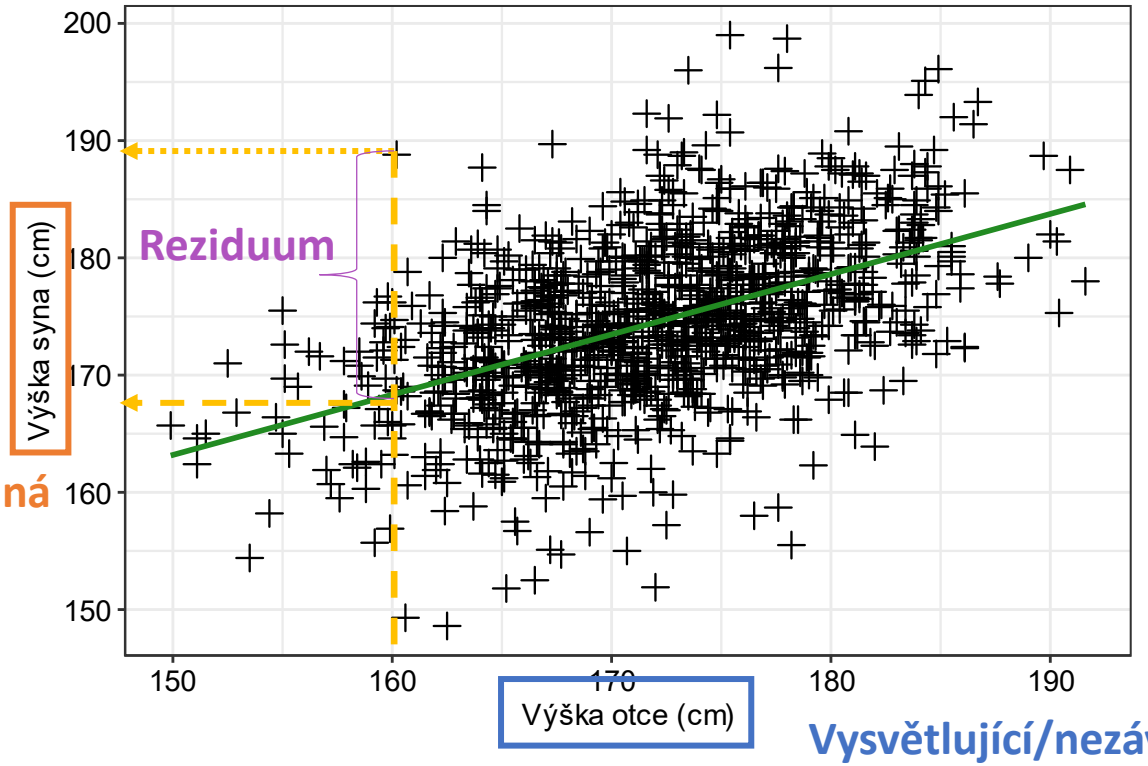
Základní pojmy

- Regresní analýza - statistické metody, které se používají k odhadu hodnoty náhodné veličiny na základě znalosti hodnot jiných náhodných veličin.

Pozorovaná hodnota

Vyrovnaná/predikovaná hodnota

Vysvětlovaná/závislá proměnná



Vysvětlující/nezávislá proměnná/prediktor/regresor

Základní pojmy

Regresní model

- Statistický model vysvětlující vztah mezi tzv. závisle proměnnou a nezávislými proměnnými (prediktory).
- Při daných hodnotách nezávisle proměnných umožňuje odhadnout hodnotu závisle proměnné.

Závisle proměnná (vysvětlovaná proměnná)

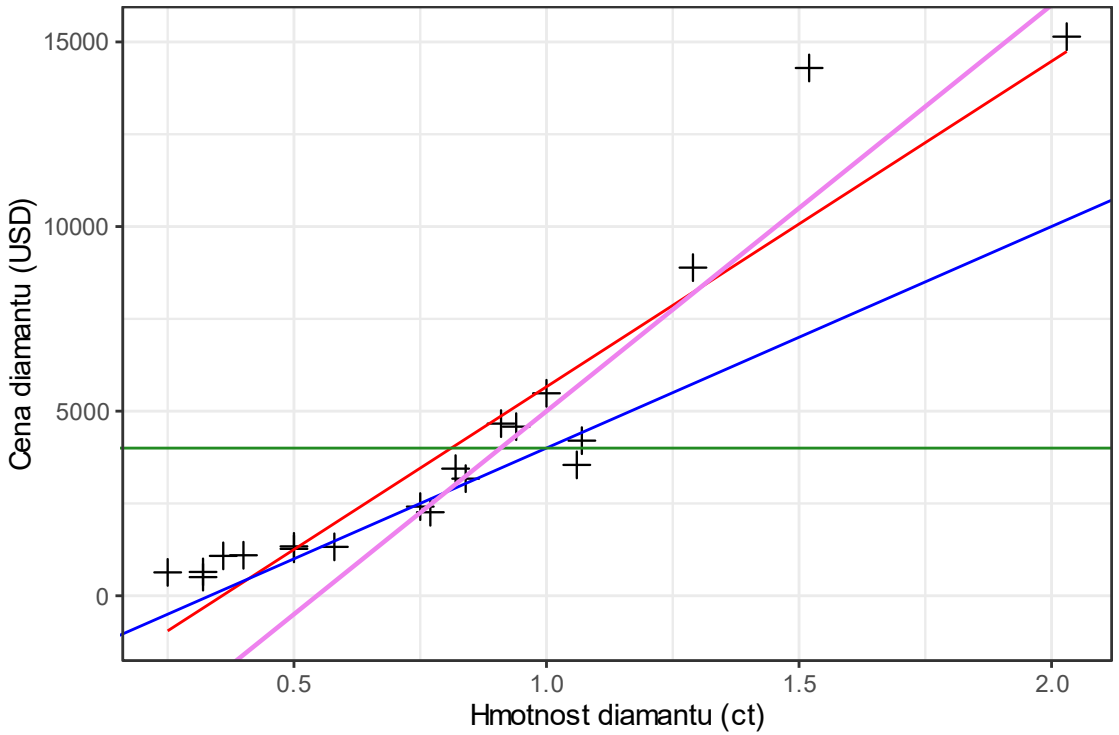
- Proměnná, jejíž hodnotu chceme odhadovat.

Nezávisle proměnné (vysvětlující proměnné, regresory, prediktory)

- Proměnné ovlivňující hodnoty závisle proměnné.

Jak odhadnout „nejlepší“ regresní přímku?

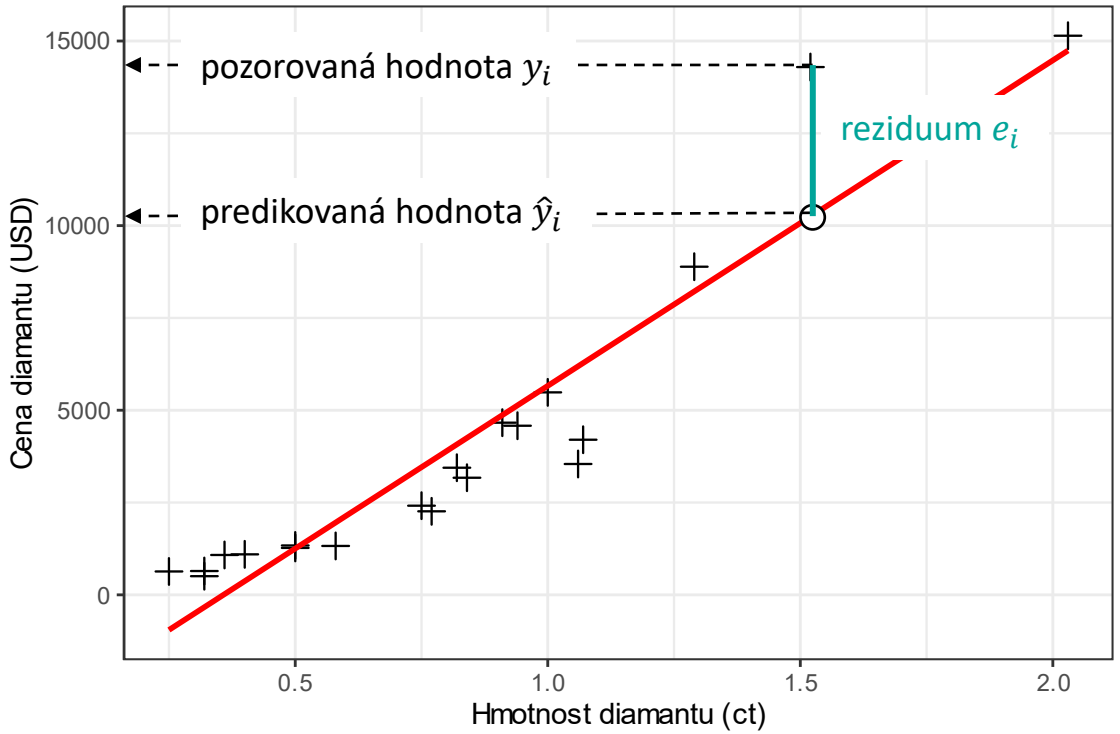
- Která přímka nejlépe vystihuje závislost mezi prodejní cenou diamantu a jeho karátovou hmotností?



a) fialová
b) modrá
c) červená
d) zelená

Jak odhadnout „nejlepší“ regresní přímku?

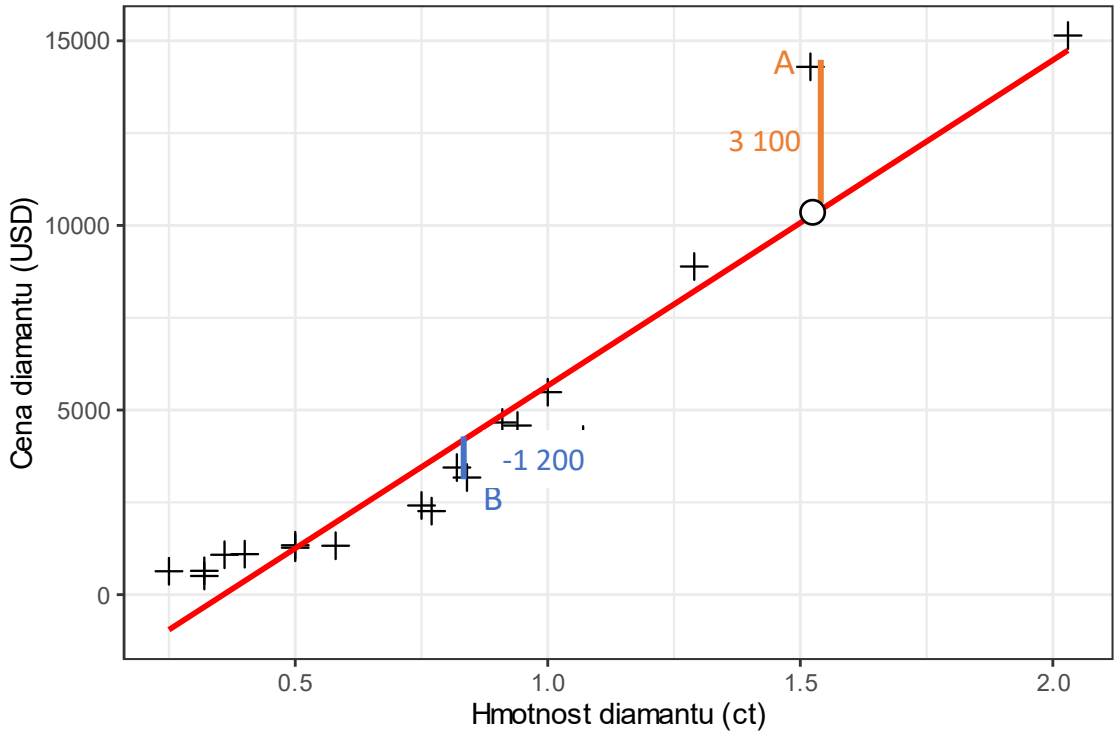
- Rezidua i-tého pozorování $e_i = y_i - \hat{y}_i$



- Reziduum je rozdíl mezi pozorovanou hodnotou a predikovanou hodnotou závisle proměnné.
- Body nad/pod regresní přímkou mají kladná/záporná rezidua.

Jak odhadnout „nejlepší“ regresní přímku?

- Rezidua i-tého pozorování $e_i = y_i - \hat{y}_i$



- Reziduum je rozdíl mezi pozorovanou hodnotou a predikovanou hodnotou závisle proměnné.
- Body nad/pod regresní přímkou mají kladná/záporná rezidua.
- Diamant A byl prodán za cenu o 3 100 USD převyšující predikovanou cenu.
- Diamant B byl prodán za cenu o 1 200 USD nižší než byla predikovaná cena.

Jak odhadnout „nejlepší“ regresní přímku?

Cílem je, aby regresní přímka $\hat{y} = b_0 + b_1 x$ byla určena tak, aby rezidua byla co nejmenší.

- **Možnost 1**: Minimalizovat součet absolutních hodnot reziduí, tj.

$$|e_1| + |e_2| + \cdots + |e_n| = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - b_0 - b_1 x_i|.$$

  - ✓ Obtížné určit bez využití výpočetní techniky.
  - ✓ Přímka je méně citlivá na odlehlá pozorování.

- **Možnost 2** (metoda nejmenších čtverců): Minimalizovat součet čtverců reziduí, tj.

$$e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

  - ✓ Jednoduché určit “ručně“ i pomocí výpočetní techniky.
  - ✓ Přímka je citlivější na odlehlá pozorování.

Metoda nejmenších čtverců (MNČ)

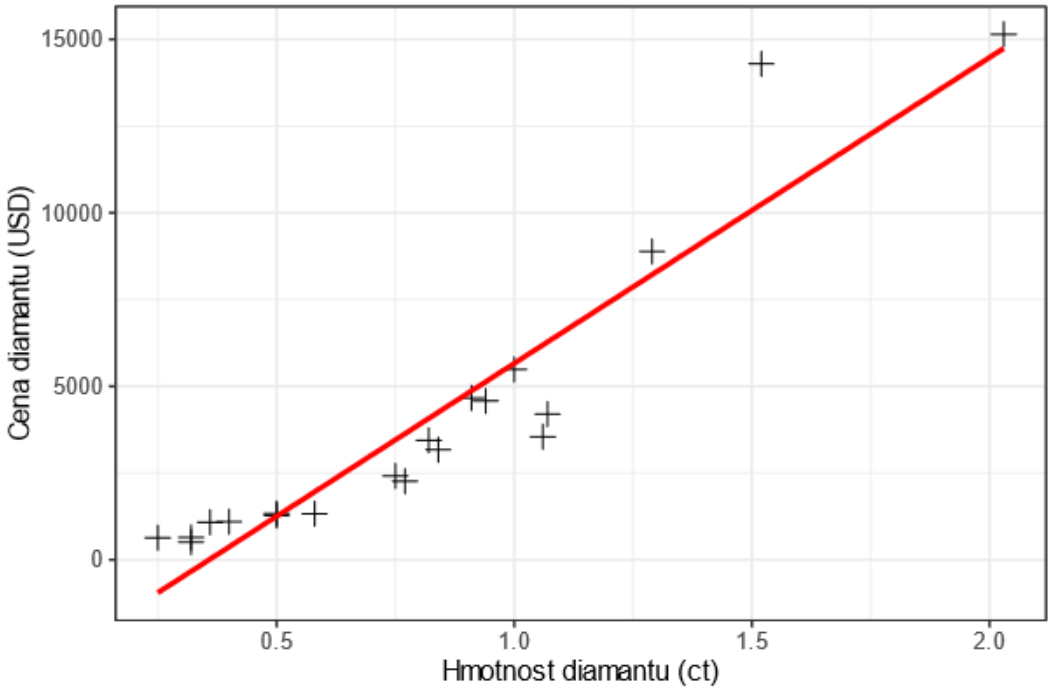
Regresní přímka $\hat{y} = b_0 + b_1 x$ určena MNČ je přímka, jejíž koeficienty jsou určeny tak, aby minimalizovaly součet čtverců reziduí, tj.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

- b_1 … směrnice (angl. slope), $b_1 = r \cdot \frac{s_Y}{s_X}$
  (r … Pearsonův korelační koeficient mezi X a Y,
  s_X (s_Y) … směrodatná odchylka X (Y))
- b_0 … konstanta/posun (angl. intercept), $b_0 = \bar{y} - b_1 \cdot \bar{x}$
  ($\bar{x}$ ($\bar{y}$) … průměr X (Y))

Metoda nejmenších čtverců (MNČ)

- Regresní přímka



	průměr	směr. odchylka	Pearsonův korel. koef.
Hmotnost diamantu (x)	0,812	0,449	0,940
Cena diamantu (y)	3 999	4 213	

Zaokrouhleno!!!

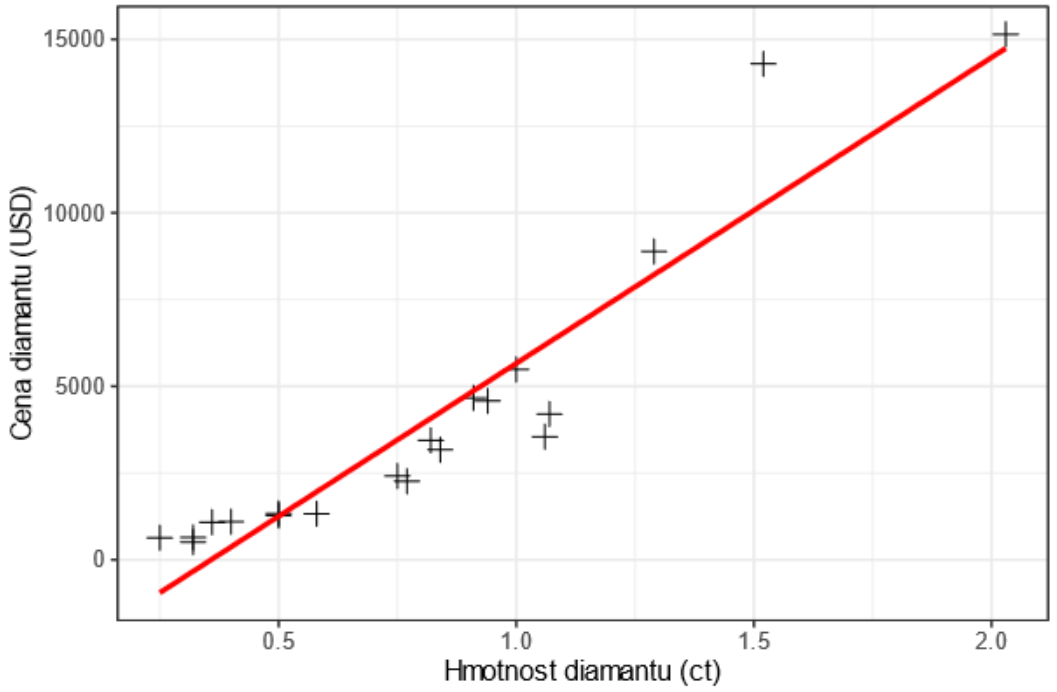
$$b_1 = r \cdot \frac{s_Y}{s_X} \cong 0{,}940 \cdot \frac{4\,213}{0{,}449} \cong 8\,820$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \cong 3\,999 - 8\,820 \cdot 0{,}812 \cong -3\,163$$

$$\widehat{Cena} = -3\,163 + 8\,820 \cdot Hmotnost$$

Metoda nejmenších čtverců (MNČ)

- Regresní přímka



„Ruční“ výpočet:

$$\widehat{Cena} = -3\,163 + 8\,820 \cdot \textit{Karátová hmotnost}$$

Výpočet s využitím software R:

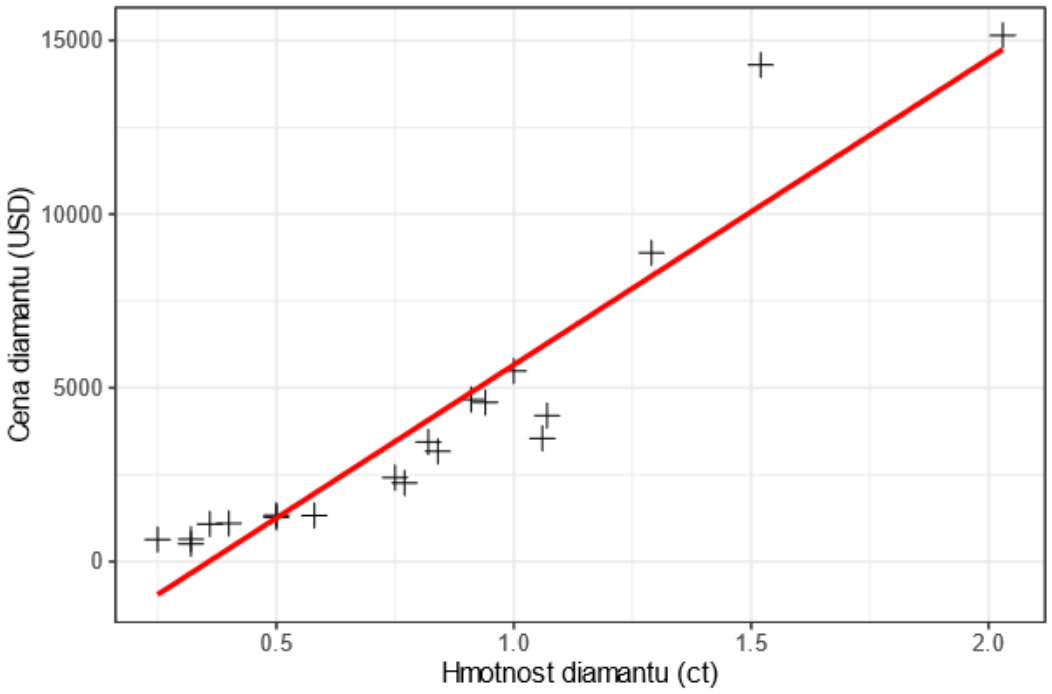
```
> lm(price~carat,data = diamonds_sample)

Call:
lm(formula = price ~ carat, data = diamonds_sample)

Coefficients:
(Intercept)        carat
      -3155         8816
```

$$\widehat{Cena} = -3\,155 + 8\,816 \cdot \textit{Hmotnost}$$

Rozdíl je způsoben zaokrouhlovací chybou!

Interpretace směrnice

- Regresní přímka: $\hat{y} = b_0 + b_1 \cdot x$
- Směrnice b_1 je **odhadem** toho, jak se **v průměru** změní predikovaná hodnota $\hat{y}$ při jednotkové změně x.



Důkaz:

$\widehat{y_1} = b_0 + b_1 \cdot x$

$\widehat{y_2} = b_0 + b_1 \cdot (x + 1)$ … x se zvýšilo o 1

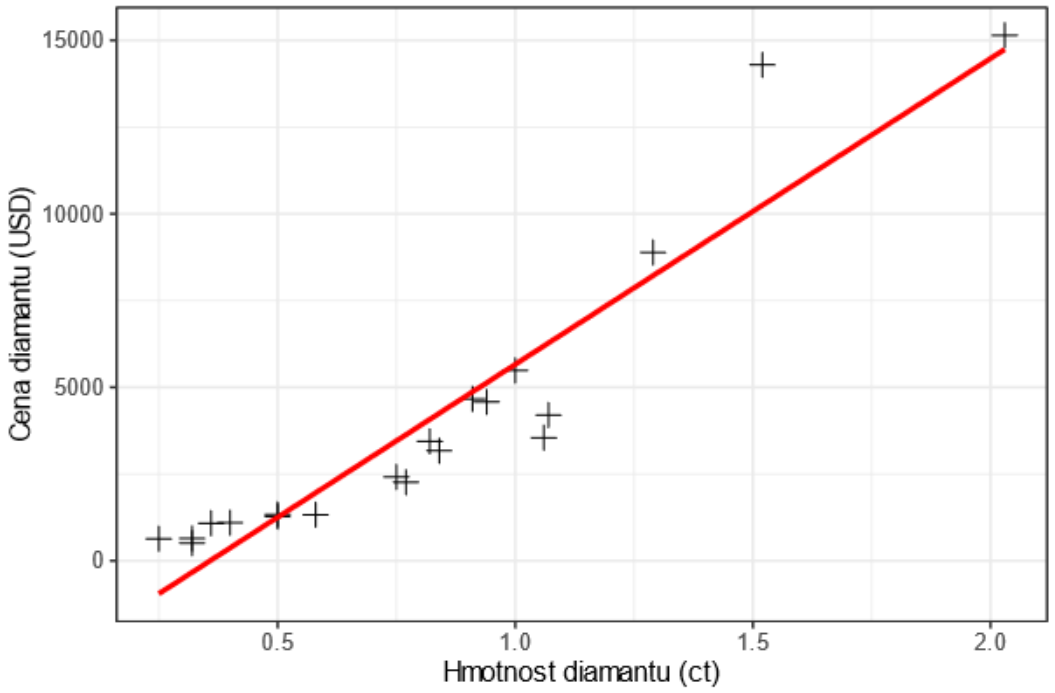
$\widehat{y_2} - \widehat{y_1} = b_1$

$$\widehat{Cena} = -3\,155 + 8\,816 \cdot Hmotnost$$

- Odhadujeme, že prodejní cena diamantů, jejichž hmotnost se liší o jeden karát, se bude v průměru lišit o 8 816 USD.

Interpretace konstanty/posunu

- Regresní přímka: $\hat{y} = b_0 + b_1 \cdot x$
- Konstanta b_o je **odhadem** průměrné predikované hodnoty závisle proměnné pro případ, kdy je nezávisle proměnná (prediktor) rovna nule.



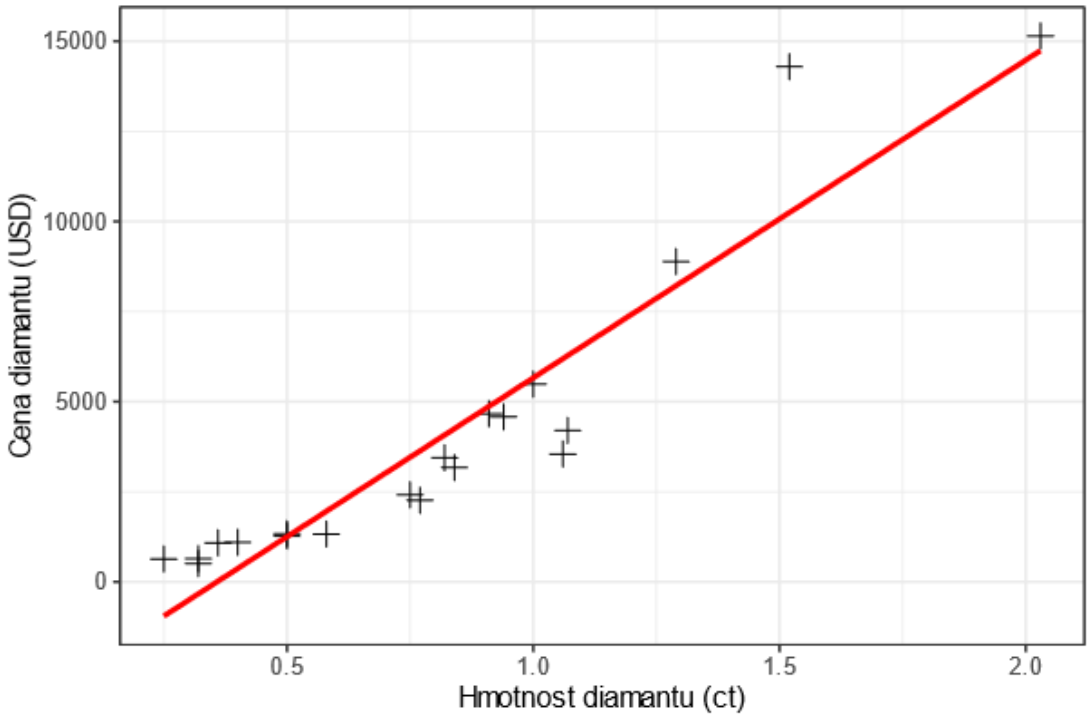
$$\widehat{Cena} = -3\,155 + 8\,816 \cdot Hmotnost$$

- Odhadujeme, že průměrná prodejní cena diamantů, jejichž hmotnost bude nula karátů bude -3 155 USD.

???

Predikce

- Predikci (odhad) závisle proměnné ($\hat{y}$) pro danou hodnotu nezávisle proměnné (x) získáváme dosazením dané hodnoty nezávisle proměnné do nalezené regresní funkce.



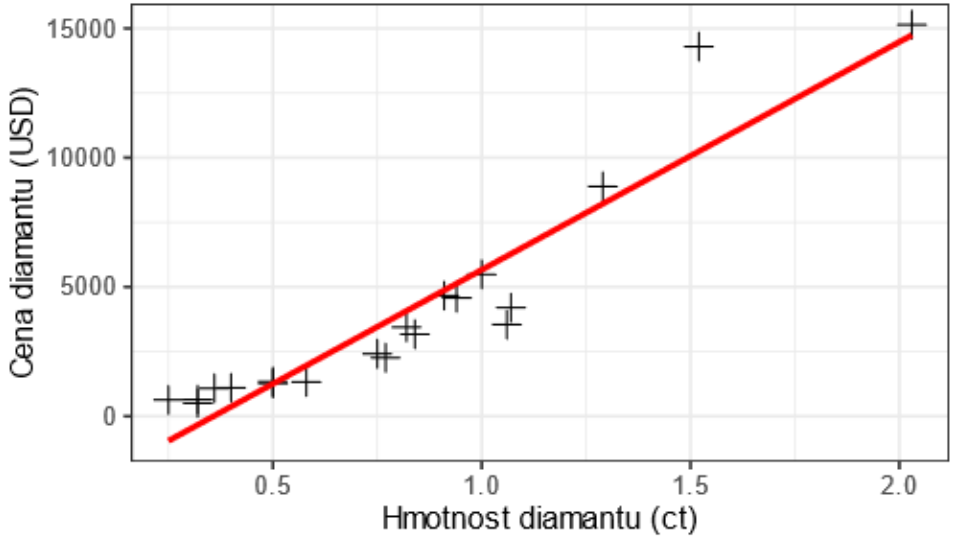
$$\widehat{Cena} = -3\,155 + 8\,816 \cdot Hmotnost$$

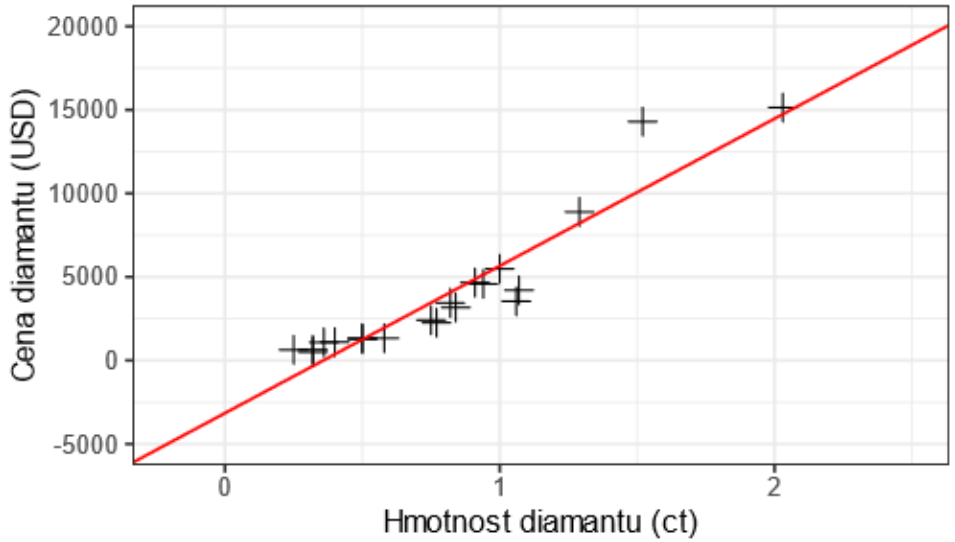
- Odhadněte průměrnou prodejní cenu diamantů s hmotností jeden karát.

$$\widehat{Cena}_{Hmotnost=1} = -3\,155 + 8\,816 \cdot 1 = \mathbf{5\,661\ USD}$$

Interpolace vs. Extrapolace

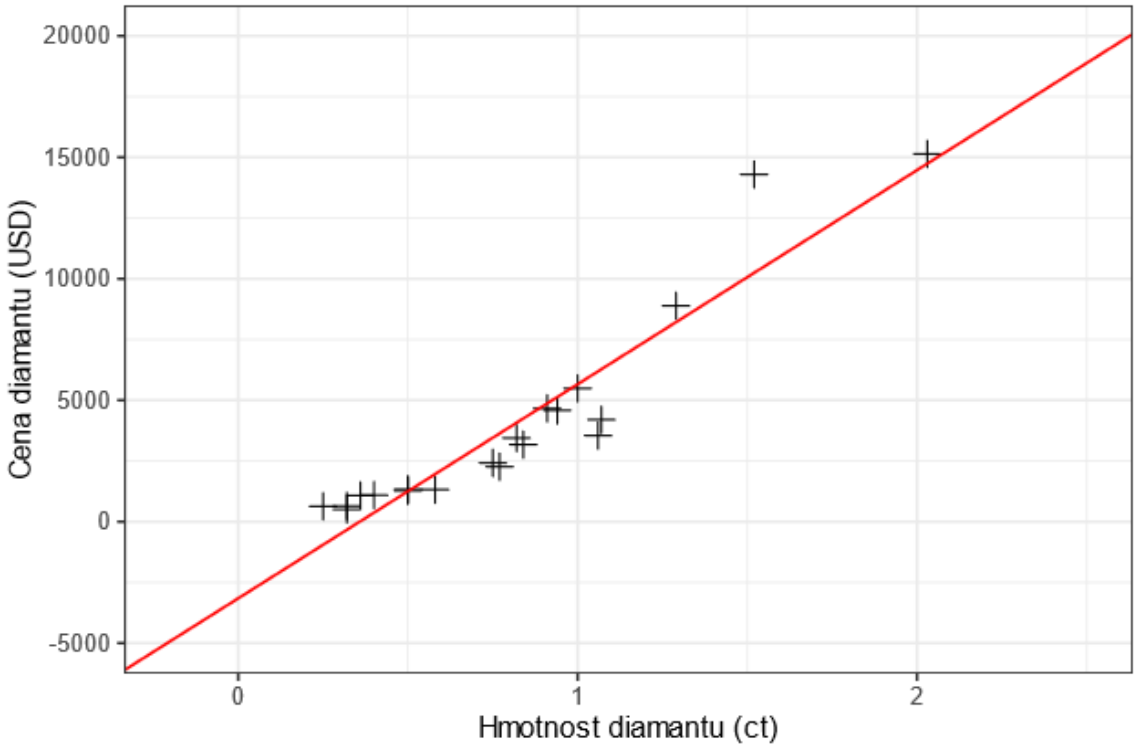
- Interpolace je predikce pro hodnoty nezávisle proměnné náležící do intervalu $\langle \min(x_i), \max(x_i) \rangle$, tj. pro hodnoty nezávisle proměnné spadající do rozpětí, v němž byla naměřena data, na jejichž základě byla regresní přímka odhadnuta.
- Extrapolace je predikce pro hodnoty nezávisle proměnné ležící mimo interval $\langle \min(x_i), \max(x_i) \rangle$.
  - ✓ **Při použití extrapolace je nutné být velmi obezřetný!!!** Nemáme žádné informace o tom, jak se závisle proměnná chová pro případ, že nezávisle proměnná je mimo interval $\langle \min(x_i), \max(x_i) \rangle$.





Interpretace konstanty/posunu

- Regresní přímka: $\hat{y} = b_0 + b_1 \cdot x$
- Konstanta b_o je **odhadem** průměrné predikované hodnoty závisle proměnné pro případ, kdy je nezávisle proměnná (regresor) rovna nule.



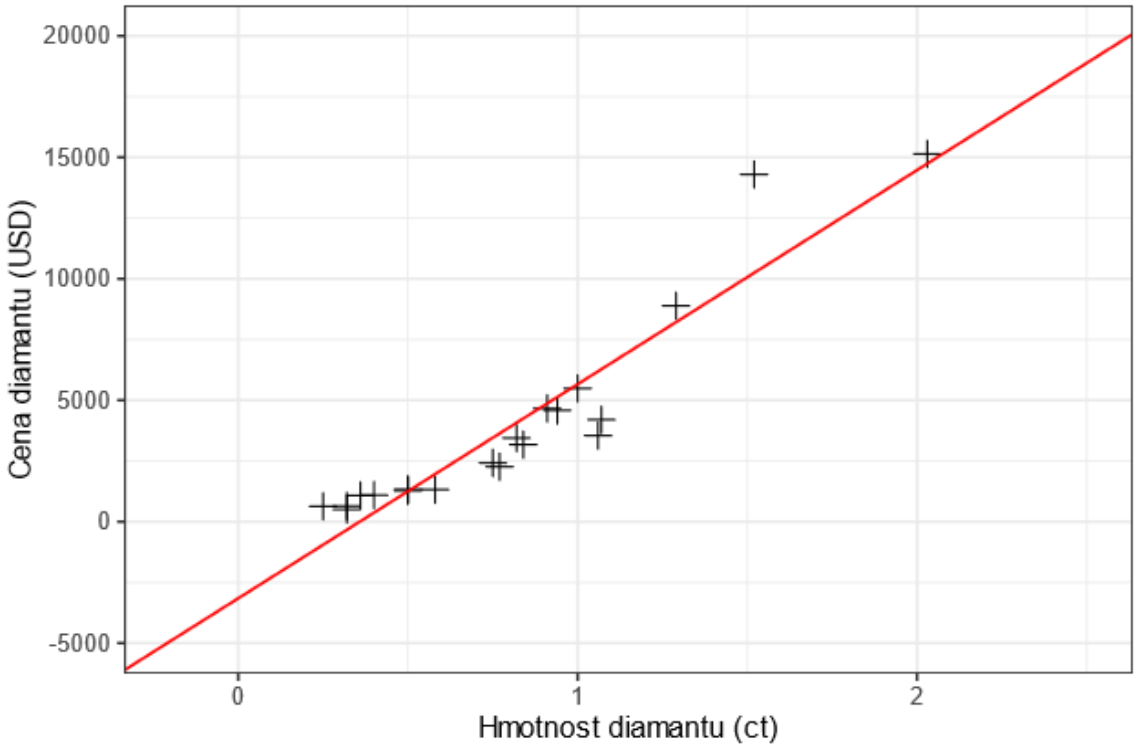
$$\widehat{Cena} = -3\,155 + 8\,816 \cdot Hmotnost$$

- Odhadujeme, že průměrná prodejní cena diamantů, jejichž hmotnost bude nula karátů bude -3 155 USD.

???

Interpretace konstanty/posunu

- Regresní přímka: $\hat{y} = b_0 + b_1 \cdot x$
- Konstanta b_o je **odhadem** průměrné predikované hodnoty závisle proměnné pro případ, kdy je nezávisle proměnná (regresor) rovna nule.



$$\widehat{Cena} = -3\,155 + 8\,816 \cdot Hmotnost$$

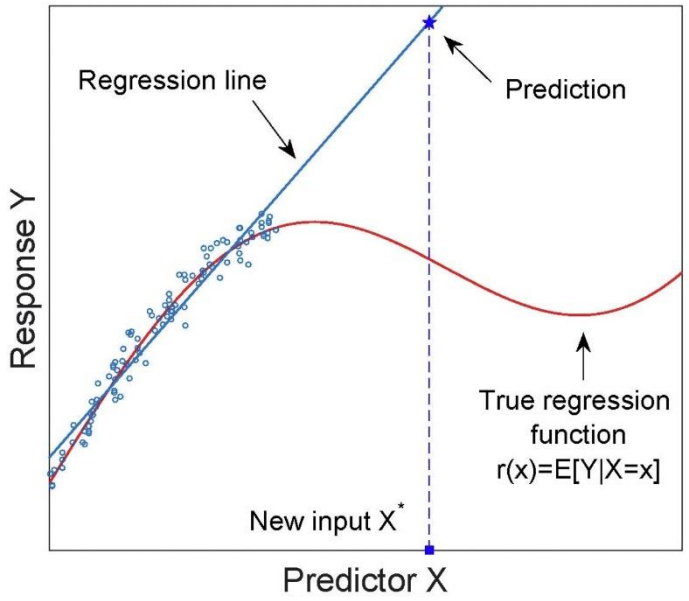
- Odhadujeme, že průměrná prodejní cena diamantů, jejichž hmotnost bude nula karátů bude -3 155 USD.

Predikce!!!

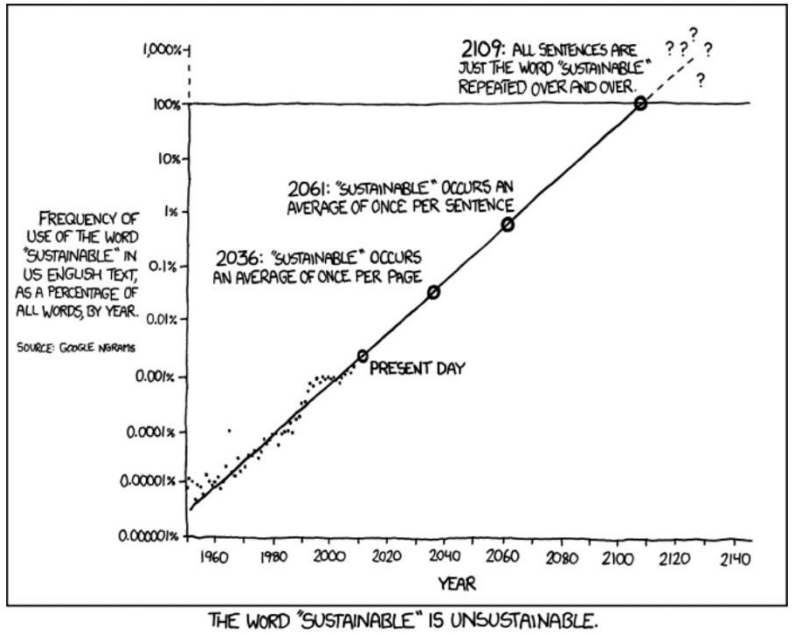
Je zřejmé, že model není vhodný pro predikci ceny drahokamu o hmotnosti nula karátů, tj. interpretace konstanty je v tomto případě nemožná.

Predikce

- Interpolace – predikce „uvnitř intervalu naměřených dat“
- Extrapolace – predikce „vně intervalu naměřených dat“
  - ✓ **POZOR!!!** Extrapolaci můžeme považovat za důvěryhodnou pouze v případě, že jsme přesvědčeni o platnosti používaného modelu v oblasti predikce!!!

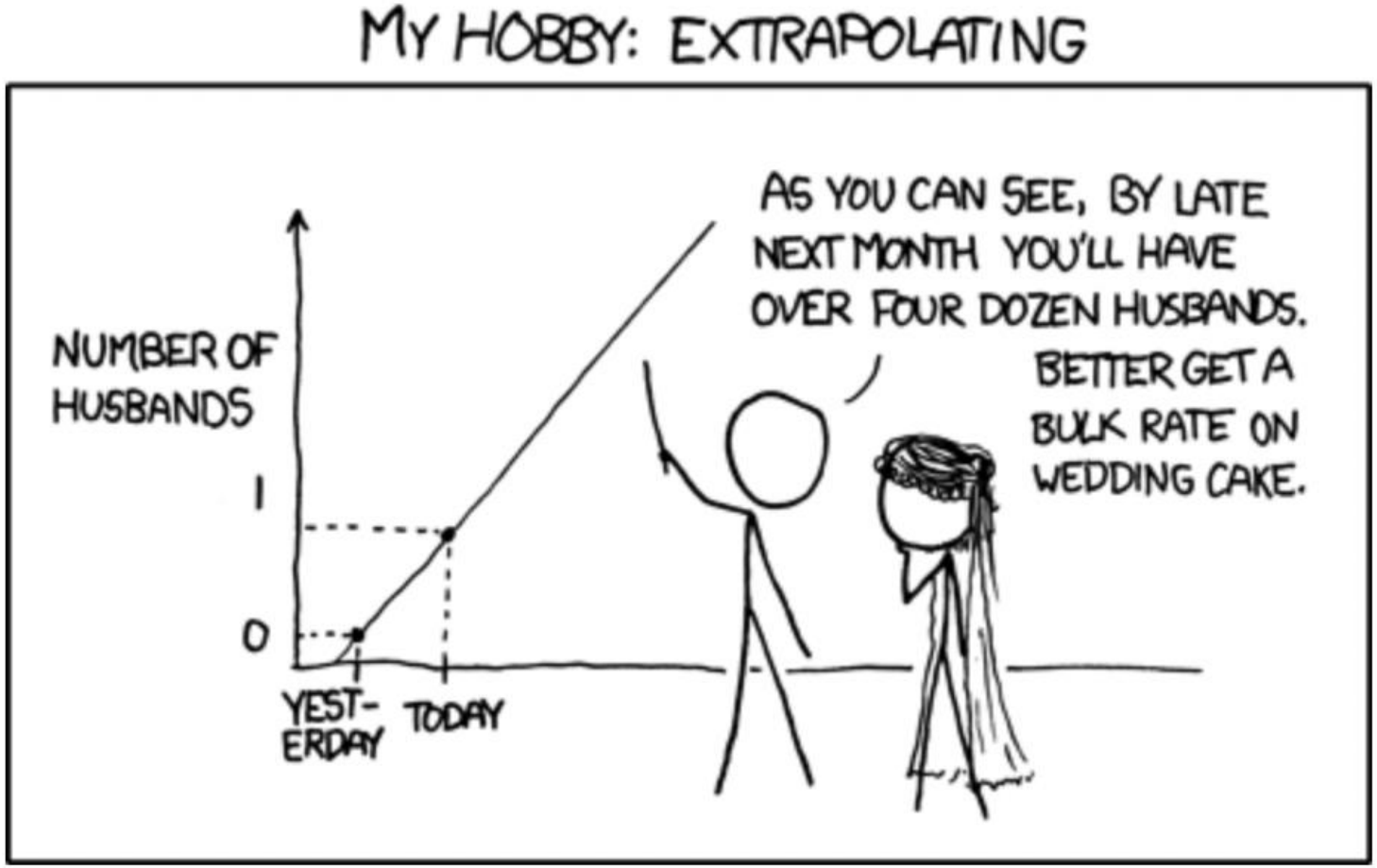


Zdroj: https://stats.stackexchange.com/questions/219579/what-is-wrong-with-extrapolation



Zdroj: https://xkcd.com/1007/

Extrapolace



Zdroj: https://xkcd.com/605/

Extrapolace

BBC Home News Sport Radio TV Weather Languages Search

[an error occurred while processing this directive]

Low graphics | Accessibility help

BBC NEWS

Watch One-Minute World News

News services
Your news when you want it

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.

An Oxford University study found that women are running faster than they have ever done over 100m.

Women are set to become the dominant sprinters

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

A team led by Dr Tatem, from the Department of Zoology at Oxford University, calculated that by 2156, a woman sprinter could cover the 100m in 8.079 seconds.

That would put women ahead of their male colleagues, who are expected to manage a best result of 8.098.

The mathematical formula used by the scientists indicated that

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news
Video and Audio
Programmes
Have Your Say
In Pictures
Country Profiles

SEE ALSO:
- Top sprinters may have key gene 27 Aug 03 | Health
- How to eat like an Olympian 20 Aug 04 | Health

RELATED BBC LINKS:
- Athletics

TOP UK STORIES
- Major manhunt for Afghan soldier
- Unemployment dips to 2.47 million
- PM condemns sympathy for Moat

News feeds

Zdroj: http://news.bbc.co.uk/2/hi/uk_news/3702650.stm

Extrapolace

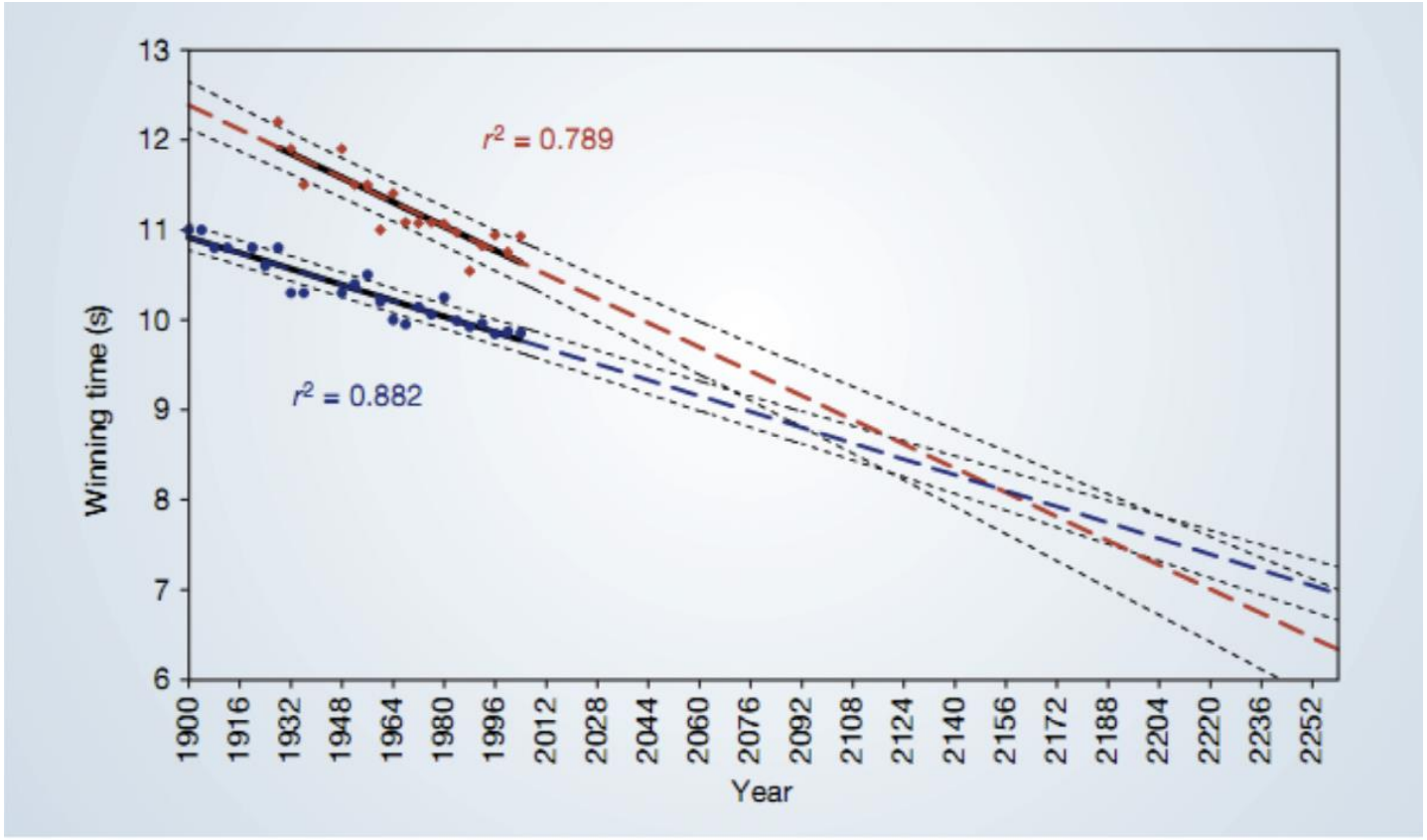
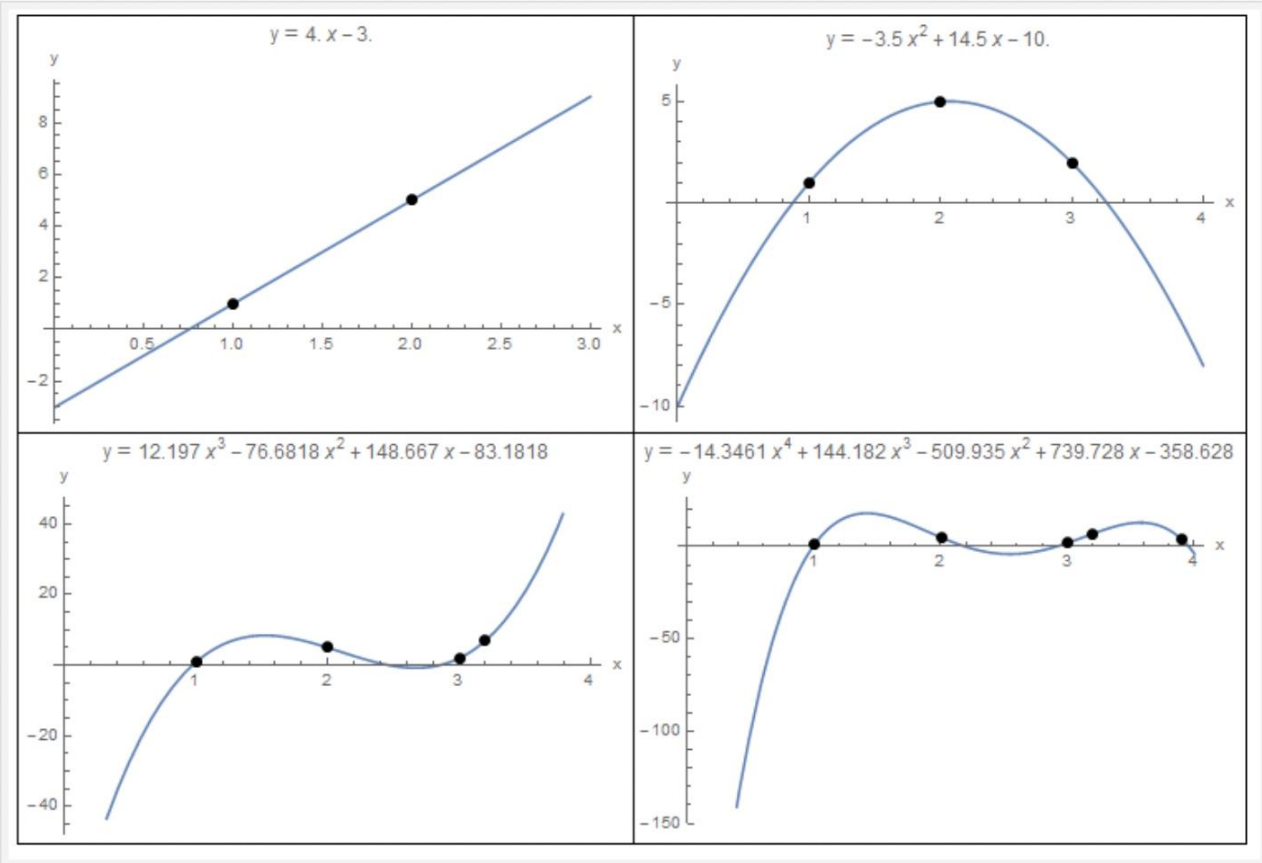


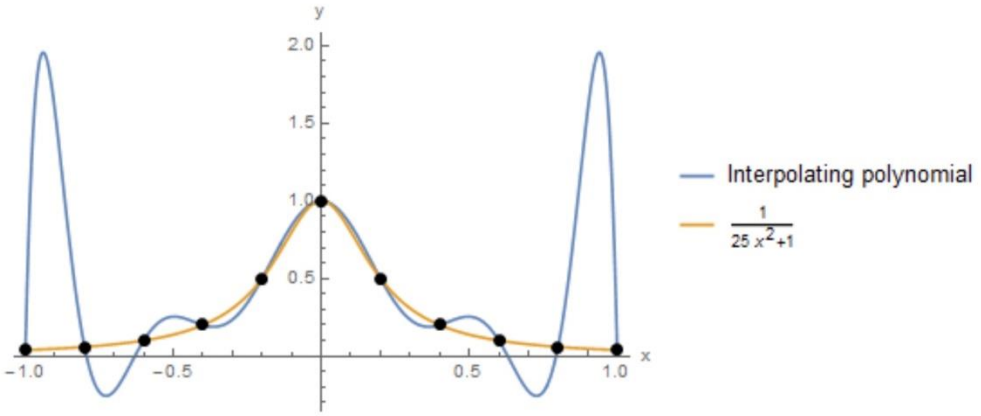
Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Zdroj: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3173856/

Interpolace

Pozor na „rozumnou“ volbu regresní funkce!





Každá n-tice bodů jednoznačně definuje polynom (n-1). řádu!

Zdroj: https://engcourses-uofa.ca/introduction-to-numerical-analysis/polynomial-interpolation/

Děkujeme za pozornost!

martina.litschmannova@vsb.cz

adela.konde@vsb.cz

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY | KATEDRA APLIKOVANÉ MATEMATIKY