Úvod do analýzy závislosti dvou kvantitativních znaků

Martina Litschmannová

martina.litschmannova@vsb.cz

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

Analýza závislosti dvou kvantitativních znaků

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v tydnu	praxe v oboru během studia	24	26

Jak popsat a vizualizovat závislost mezi výškou a hmotností respondentů?

Analýza závislosti dvou kvantitativních znaků

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v tydnu	praxe v oboru během studia	24	26

- Bodový graf
- Korelační koeficient

Jak popsat a vizualizovat závislost mezi výškou a hmotností respondentů?

Korelační koeficient

- Pearsonův korelační koeficient
- Spearmanův korelační koeficient
- Koeficient mnohonásobné korelace (Index determinace)
- Parciální korelační koeficient

Pearsonův korelační koeficient

- Pearsonův korelační koeficient je mírou lineární závislosti mezi X a Y pokud $(X_1; Y_1), \dots, (X_n; Y_n)$ je výběr z **dvourozměrného normálního rozdělení**.
- Zjistíme-li, že výběrový korelační koeficient $r \neq 0$, zpravidla nás zajímá, zda je indikovaná korelace statisticky významná, tj. velmi zjednodušeně řečeno, zda se korelační koeficient příslušných populačních dat statisticky významně liší od nuly.

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

kde

n … rozsah výběru,

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

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

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

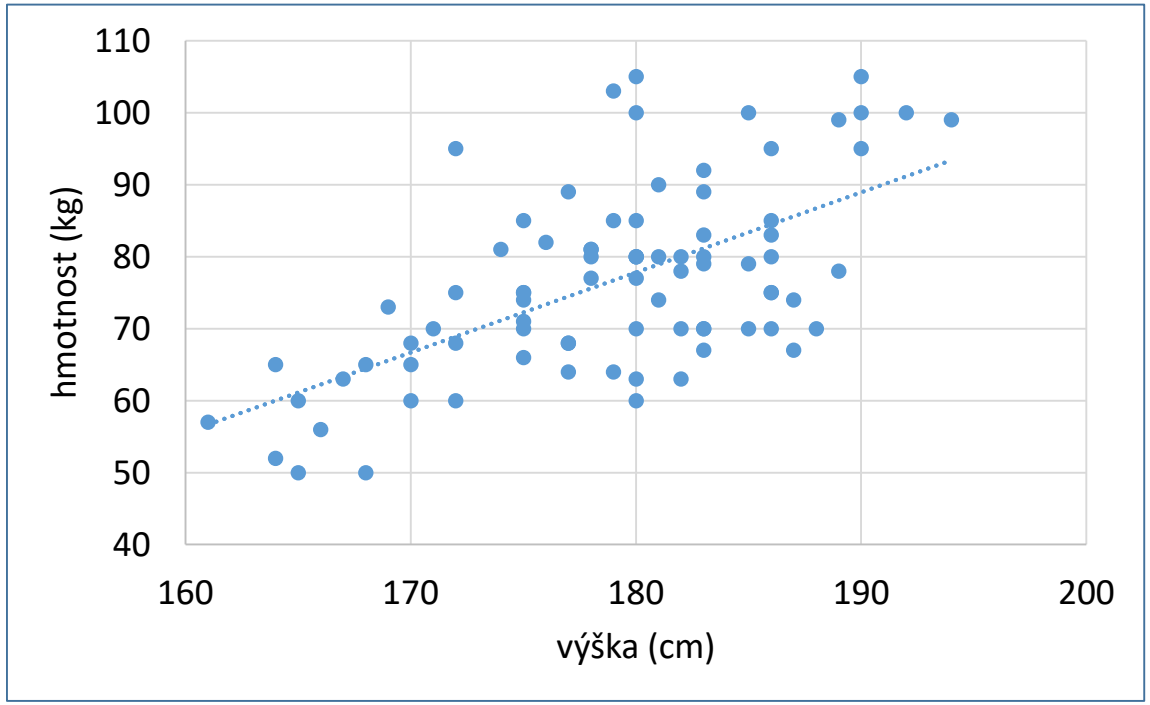
s_X (s_Y) … výběrová směrodatná odchylka znaku x (y)

Pearsonův korelační koeficient

- Hodnota Pearsonova korelačního koeficientu se pohybuje od -1 do 1.
- Hodnoty ± 1 nabývá tehdy, pokud všechny body $[x_i, y_i]$ leží na přímce.
- Nule je roven v případě, že veličiny jsou lineárně nezávislé.
- Při měření lineární závislosti je znaménko korelačního koeficientu kladné, když obě veličiny *X* a *Y* zároveň rostou nebo obě zároveň klesají, a záporné, když jedna z veličin roste, zatímco druhá klesá.
- Při užití Pearsonova korelačního koeficientu je vždy třeba posoudit, zda je jeho aplikace vhodná.

Analýza závislosti dvou kvantitativních znaků

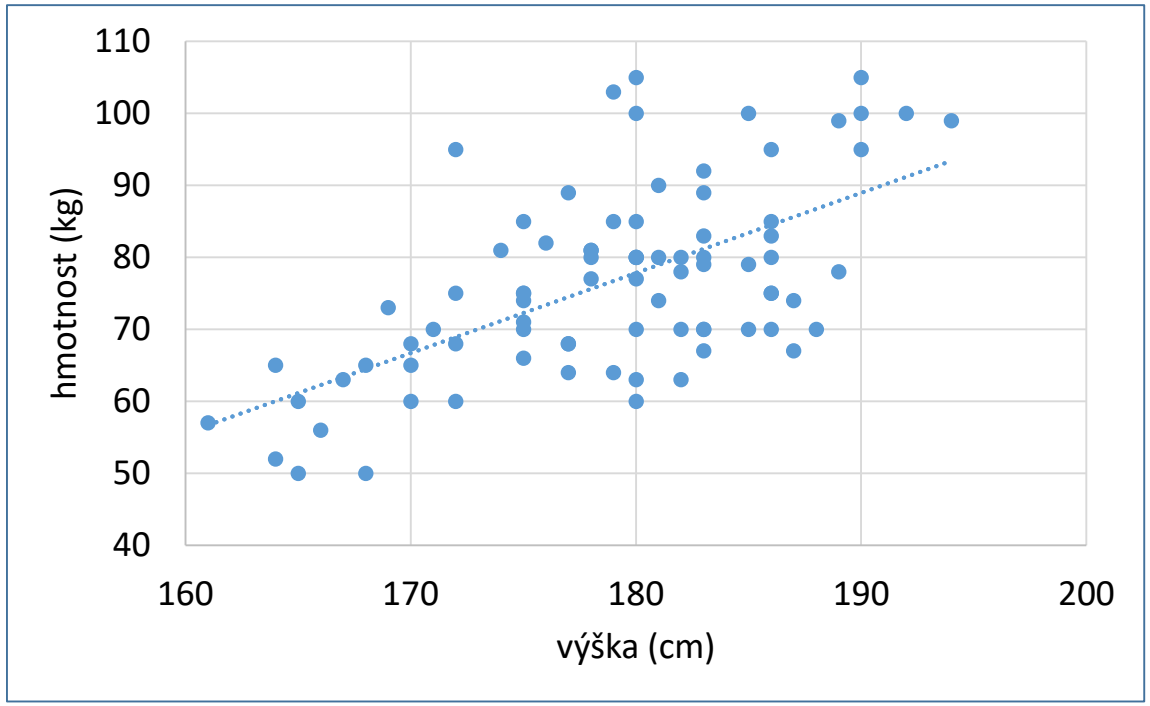
- **Vizualizace – bodový graf**



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).

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**



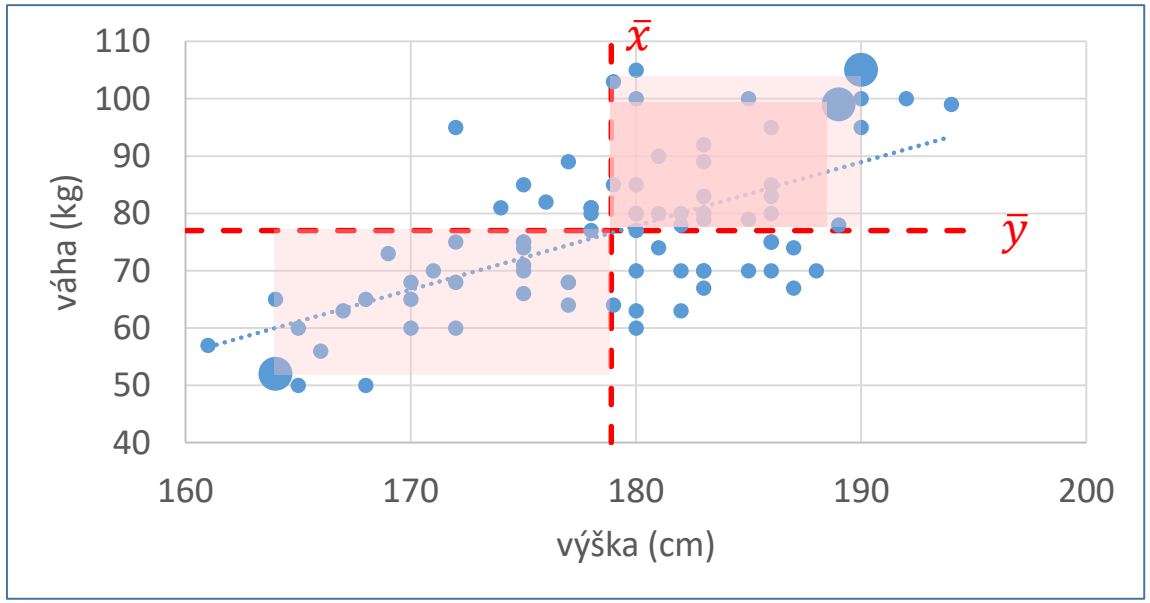
Jak „funguje“ Pearsonův korelační koeficient?

- **Pearsonův výběrový korelační koeficient**

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

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**



- **Pearsonův výběrový korelační koeficient**

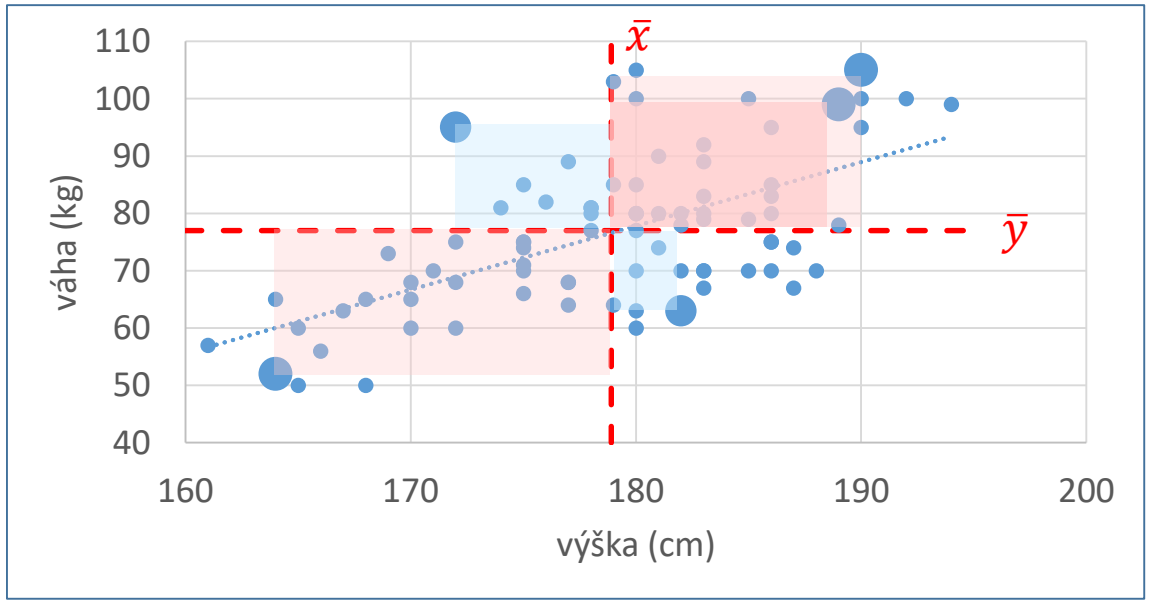
$$r = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_X \cdot s_Y}$$

Jak „funguje“ Pearsonův korelační koeficient?

- Nabývají-li pro danou statistickou jednotku oba znaky nadprůměrných hodnot, kor. koeficient se zvyšuje.
- Nabývají-li pro danou statistickou jednotku oba znaky podprůměrných hodnot, kor. koeficient se zvyšuje.

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**



- **Pearsonův výběrový korelační koeficient**

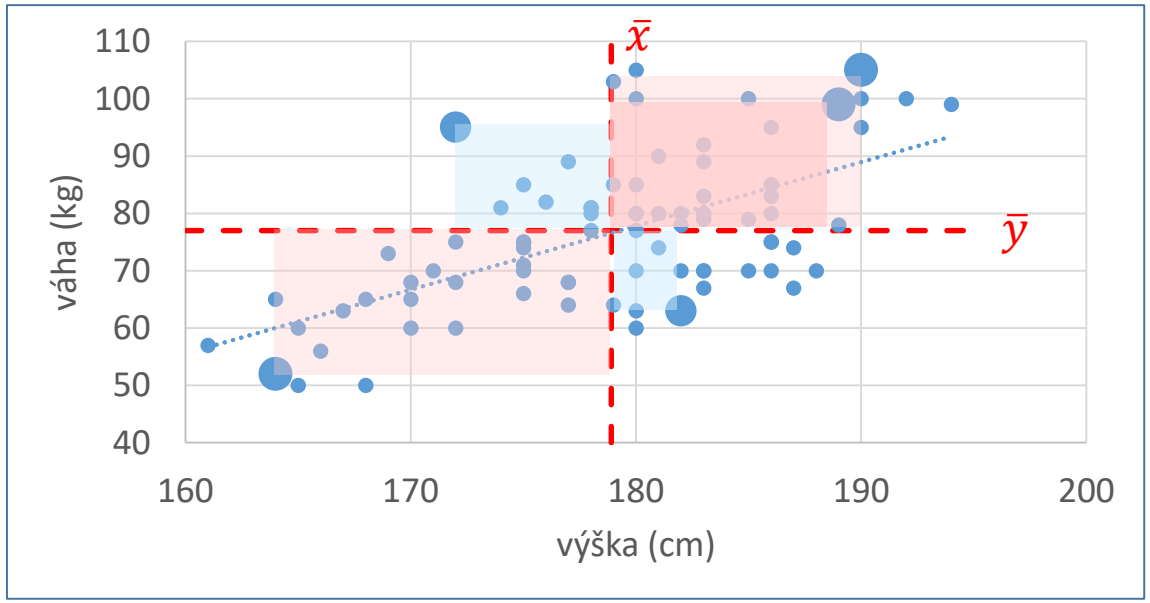
$$r = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_X \cdot s_Y}$$

Jak „funguje“ Pearsonův korelační koeficient?

- Nabývají-li pro danou statistickou jednotku oba znaky nadprůměrných (resp. podprůměrných) hodnot, korelační koeficient se zvyšuje.
- Nabývá-li pro danou statistickou jednotku jeden znak nadprůměrné hodnoty a druhý znak podprůměrné hodnoty, korelační koeficient se snižuje.

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**



- **Pearsonův výběrový korelační koeficient**

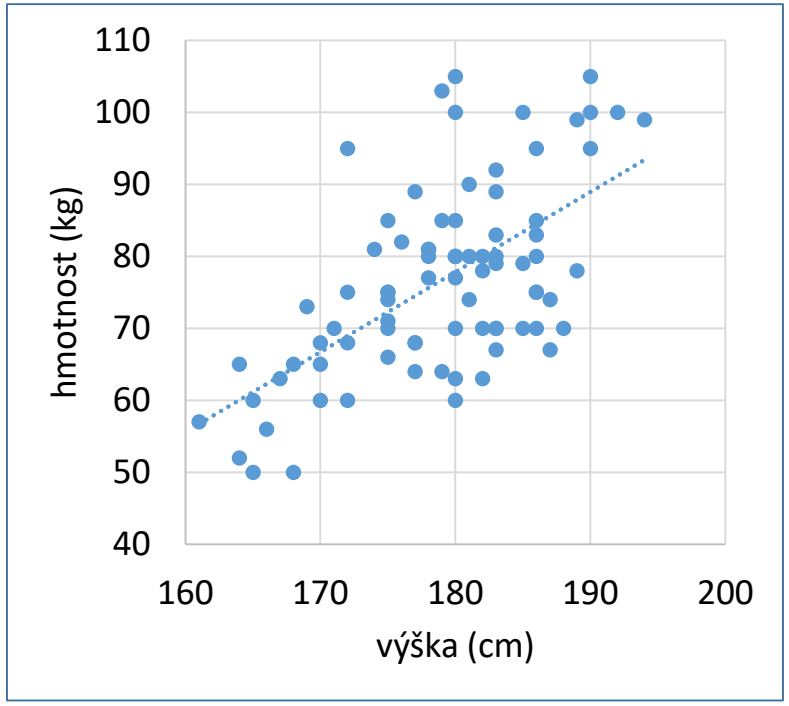
$$r = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_X \cdot s_Y}$$

Jak „funguje“ Pearsonův korelační koeficient?

- Lze ukázat, že korelační koeficient může nabývat hodnot z intervalu $\langle -1; 1 \rangle$.
- $|r| = 1 \iff \forall i \in \{1, 2, \dots, n\}: y_i = ax_i + b, kde\ a, b \in \mathbb{R}$ (mezi proměnnými je lineární závislost)
- Korelační koeficient (Pearsonův) je mírou **lineární** závislosti!!!

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**

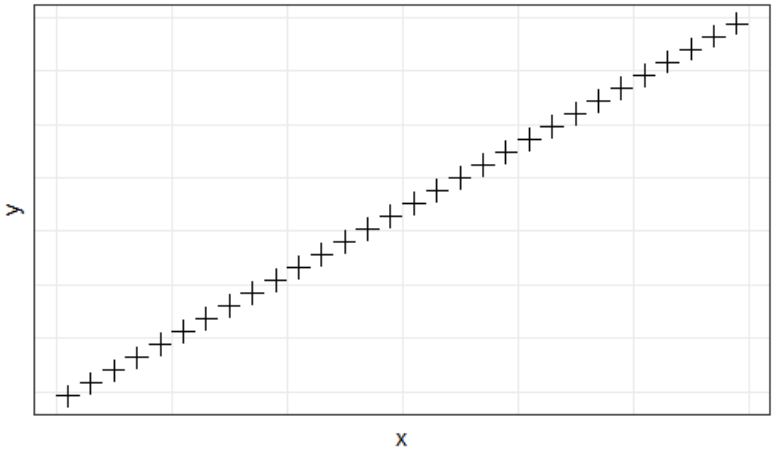


- **Pearsonův výběrový korelační koeficient**

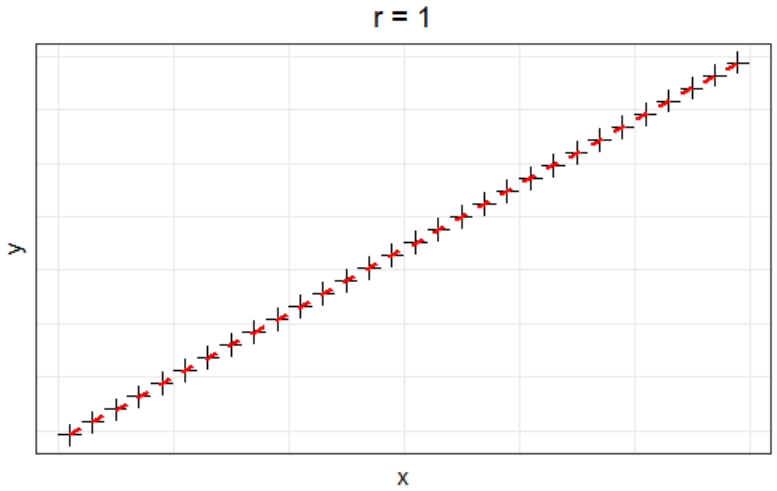
$$r = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_X \cdot s_Y}$$

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

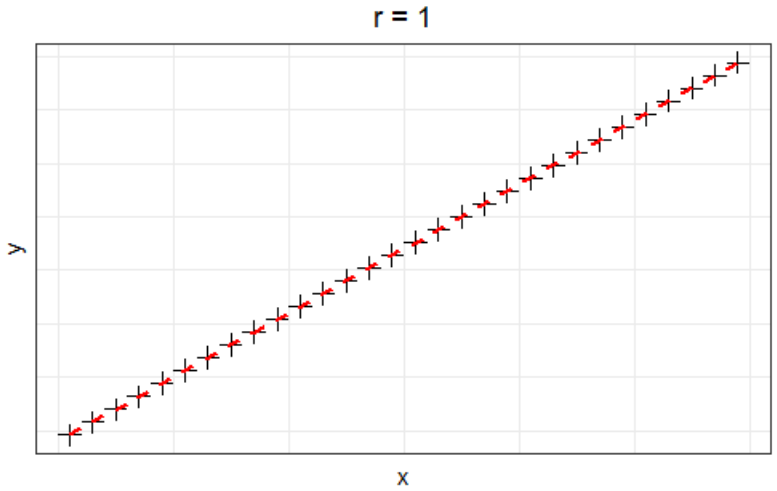
Pearsonův korelační koeficient

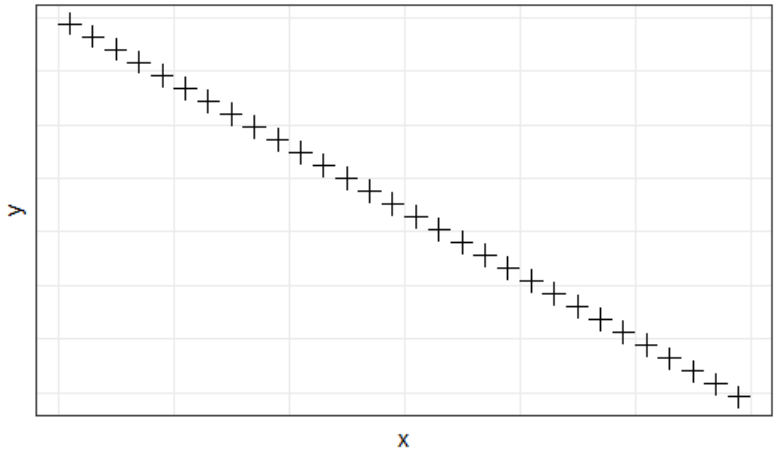


Pearsonův korelační koeficient

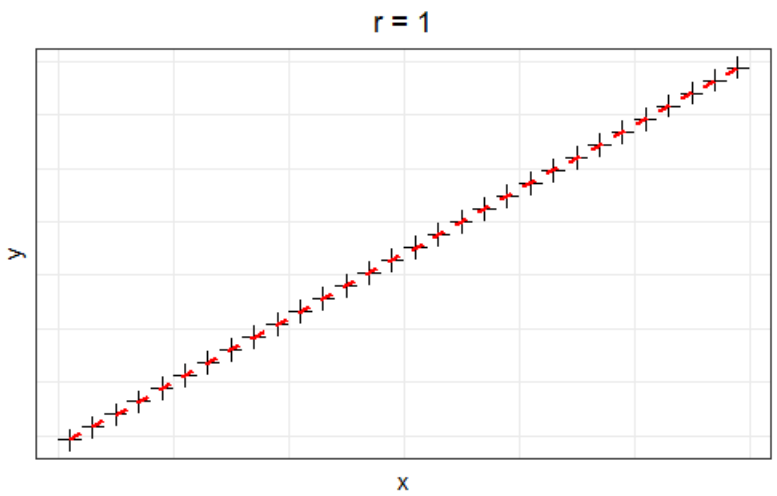


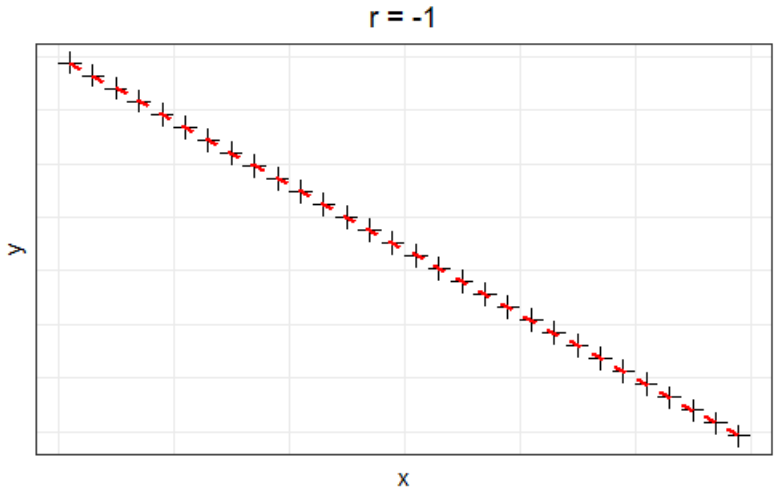
Pearsonův korelační koeficient



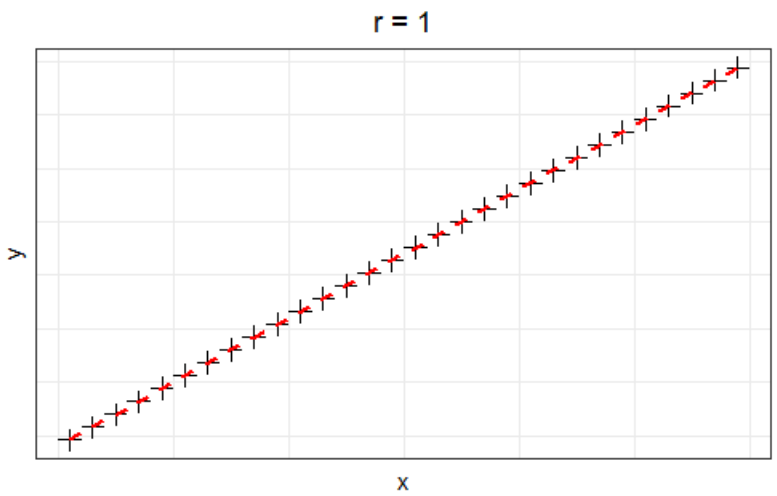


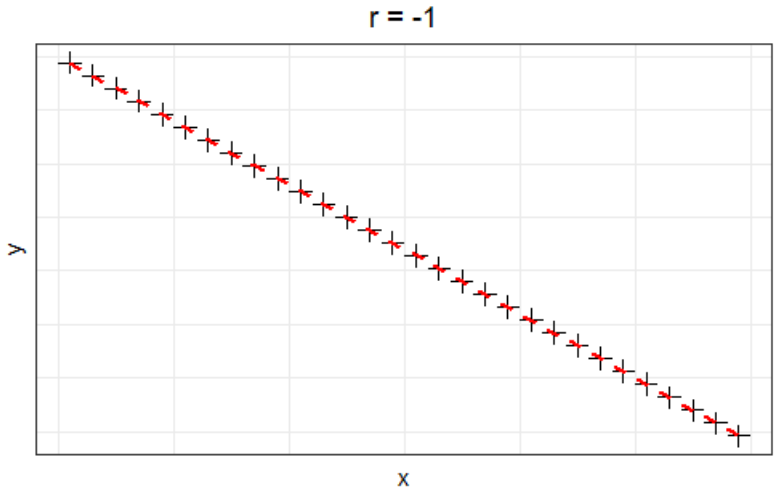
Pearsonův korelační koeficient

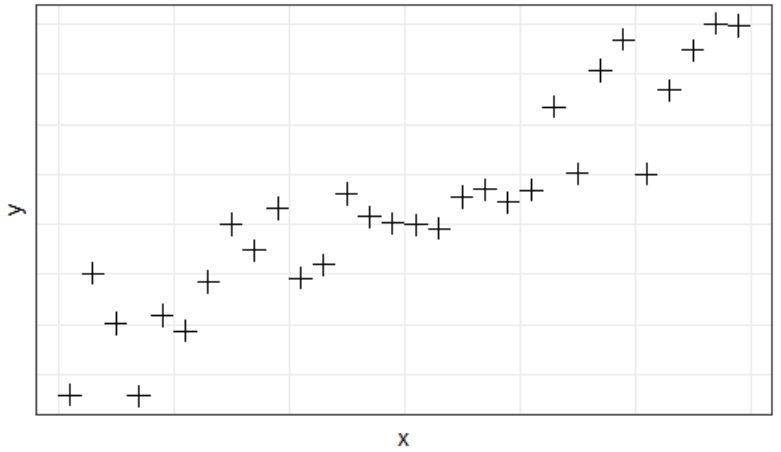




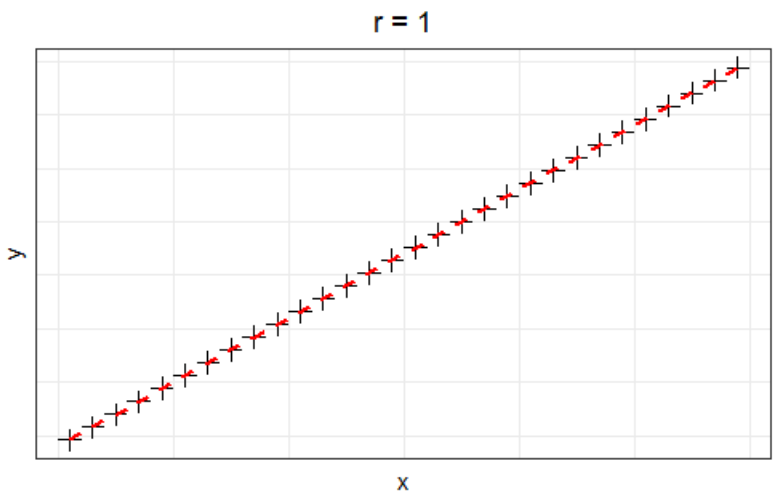
Pearsonův korelační koeficient

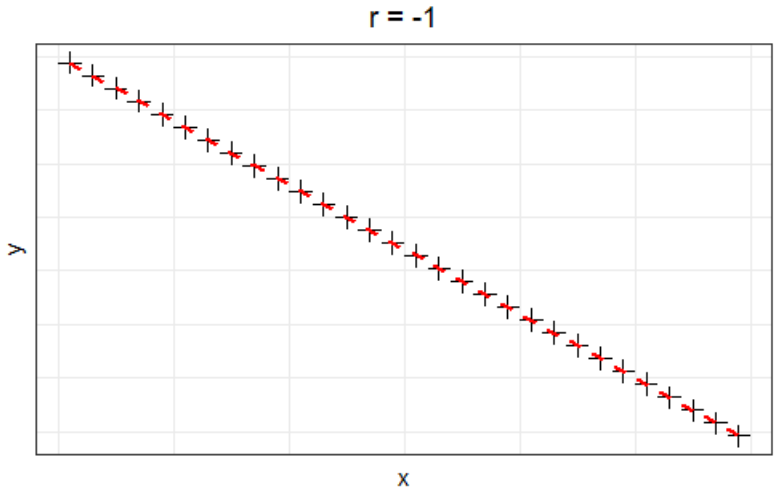


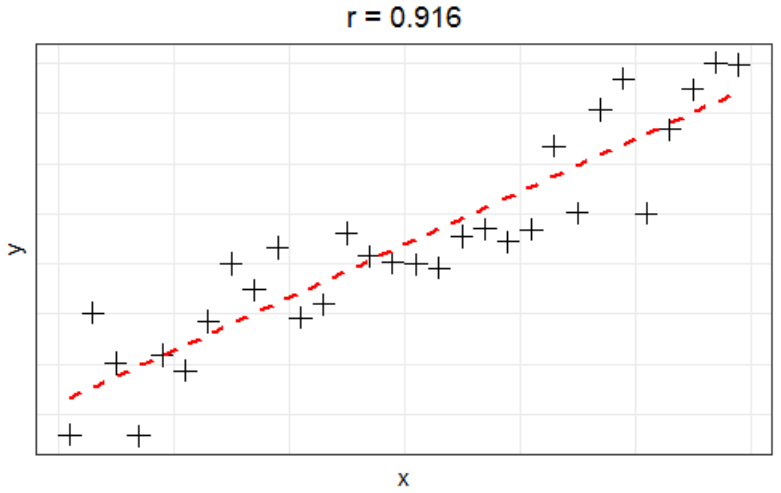




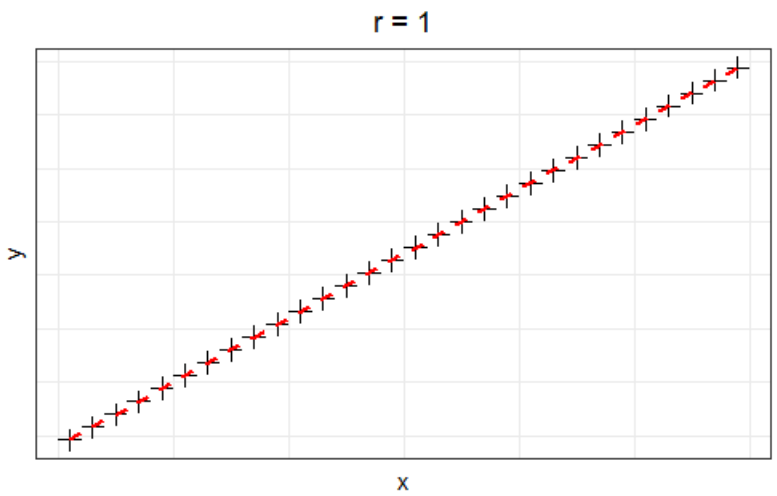
Pearsonův korelační koeficient

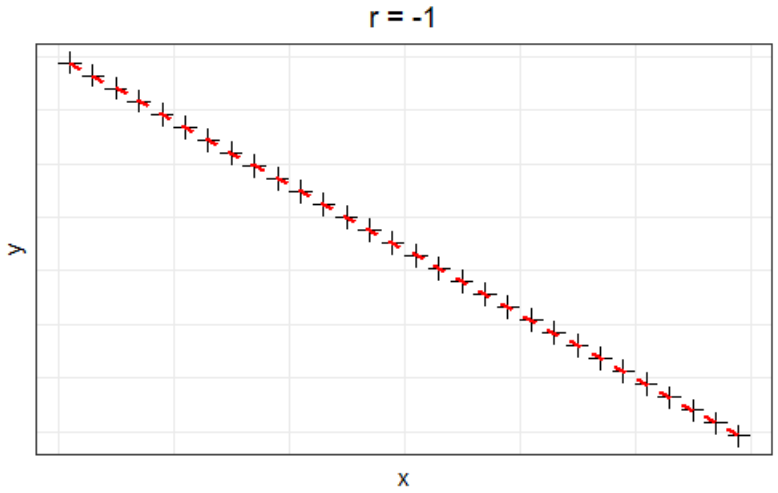


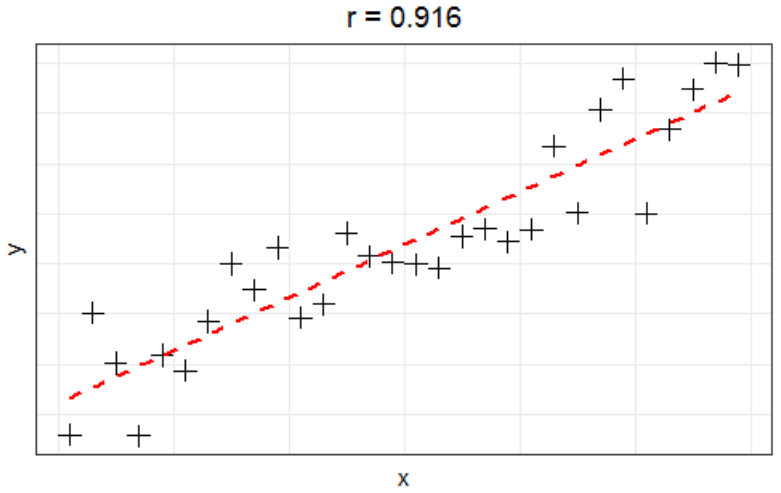


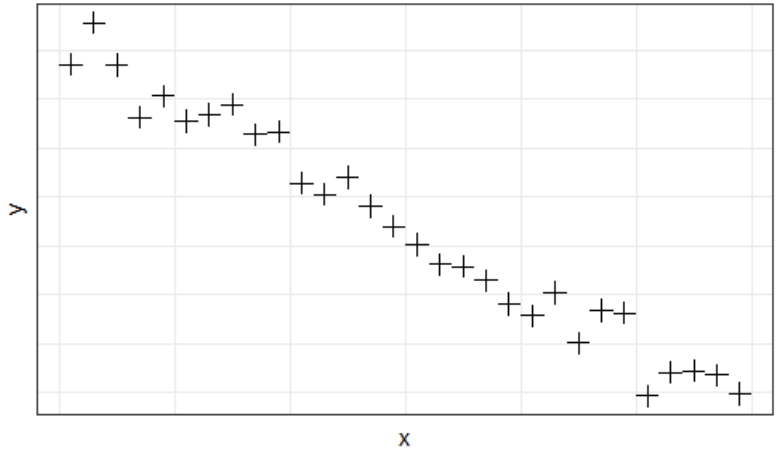


Pearsonův korelační koeficient

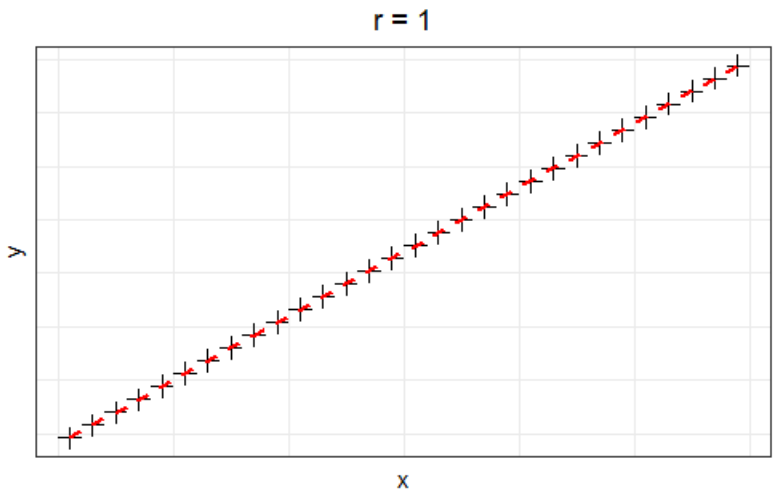


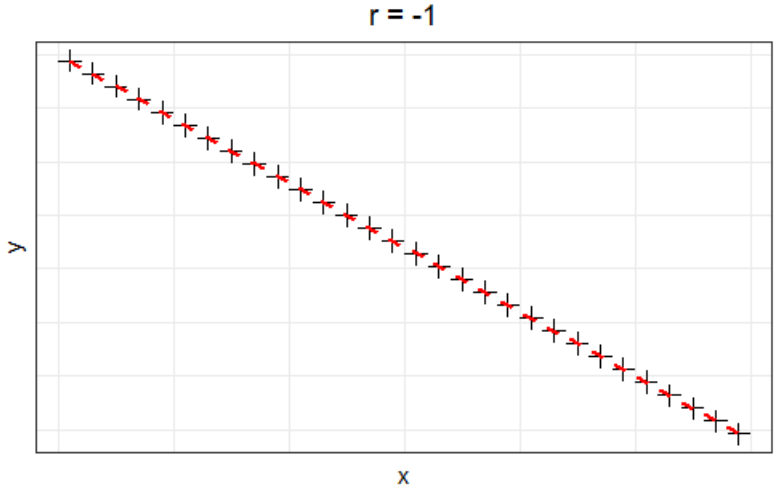


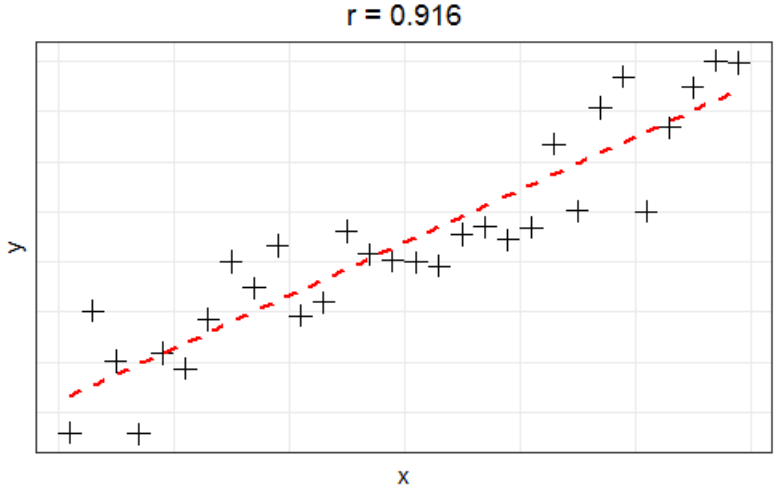


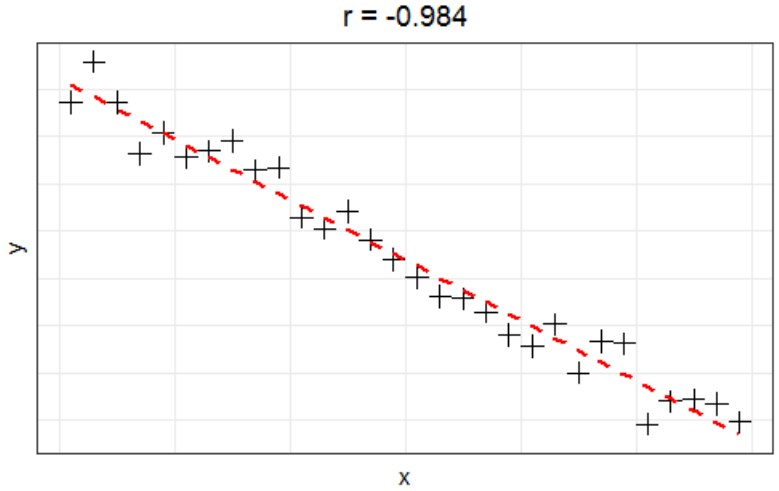


Pearsonův korelační koeficient

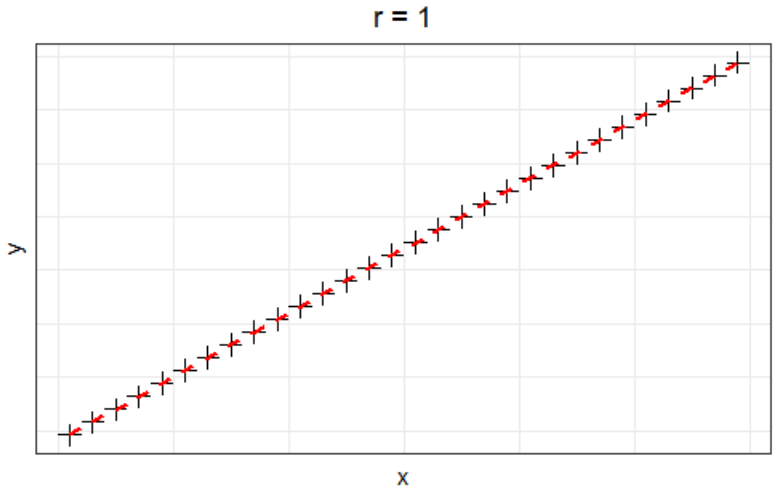


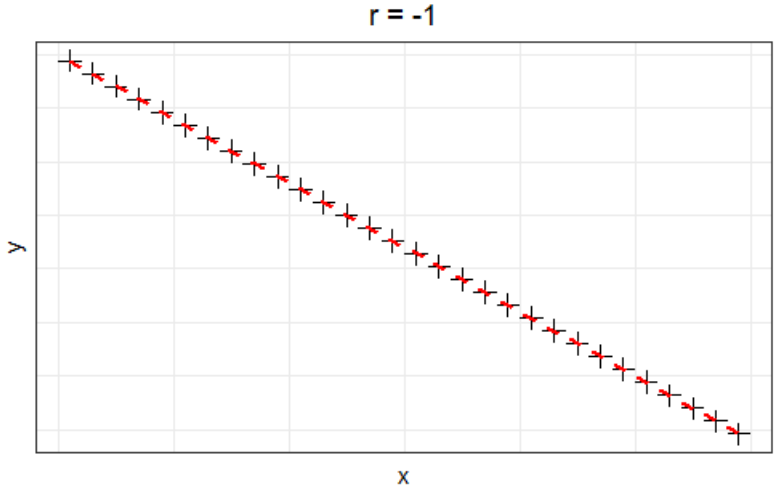


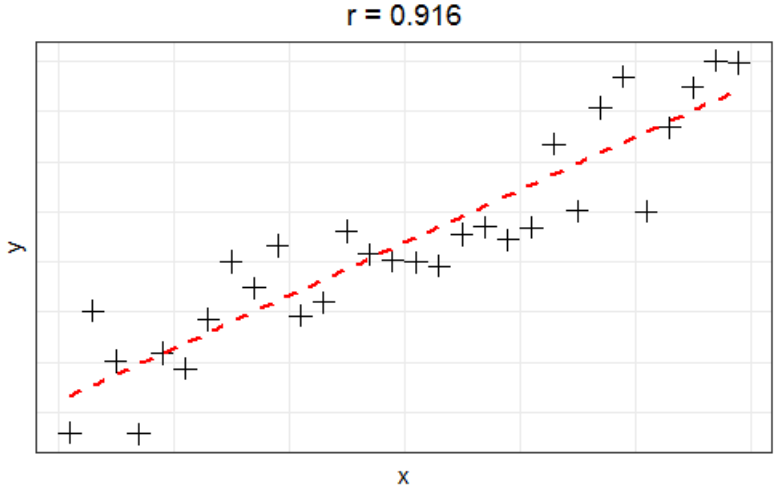


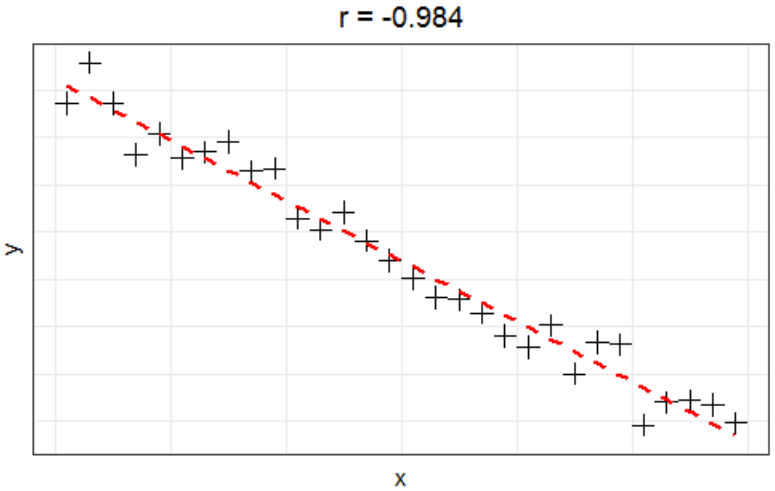


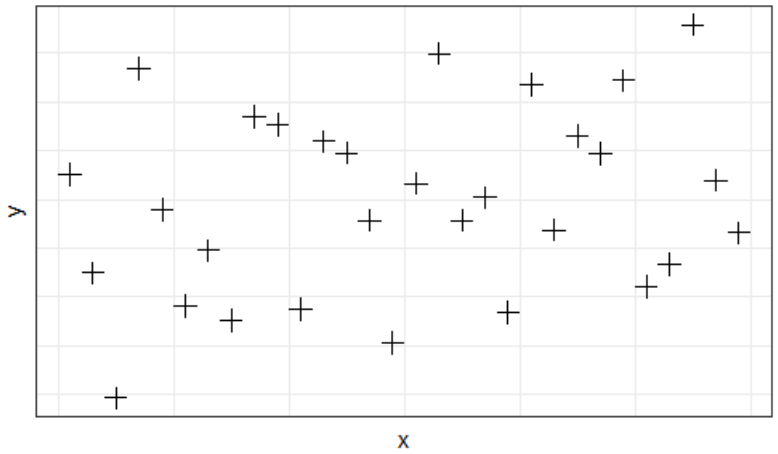
Pearsonův korelační koeficient



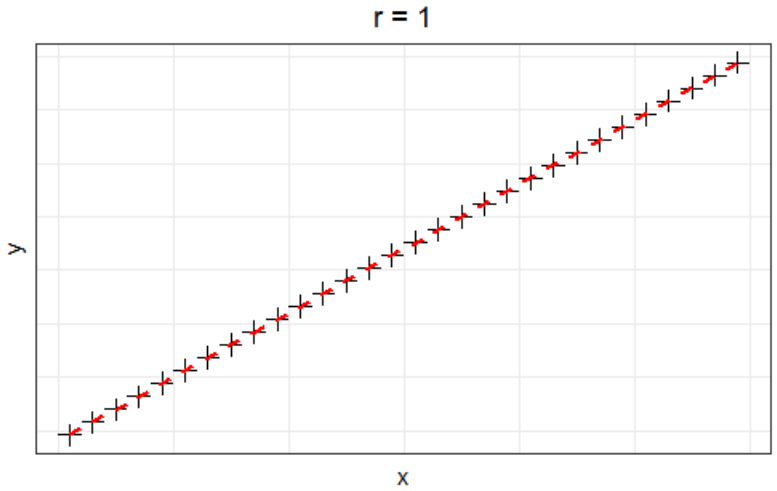


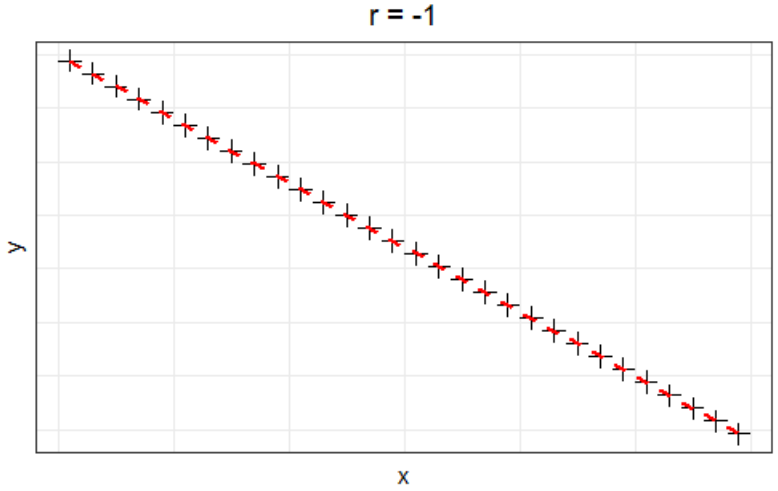


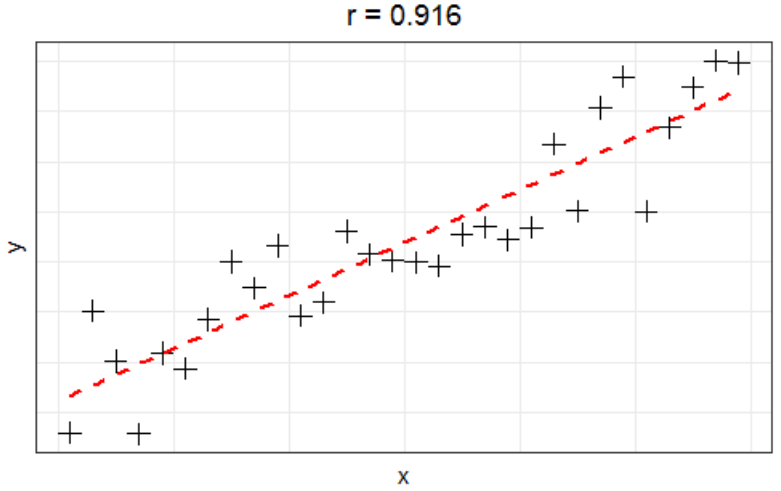


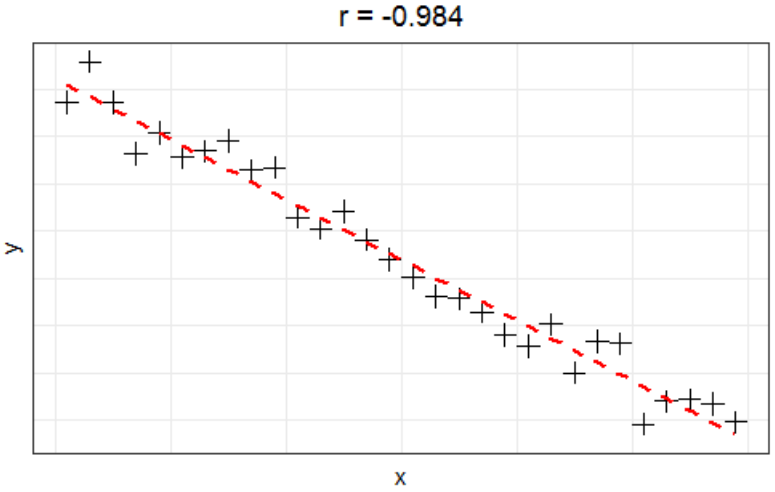


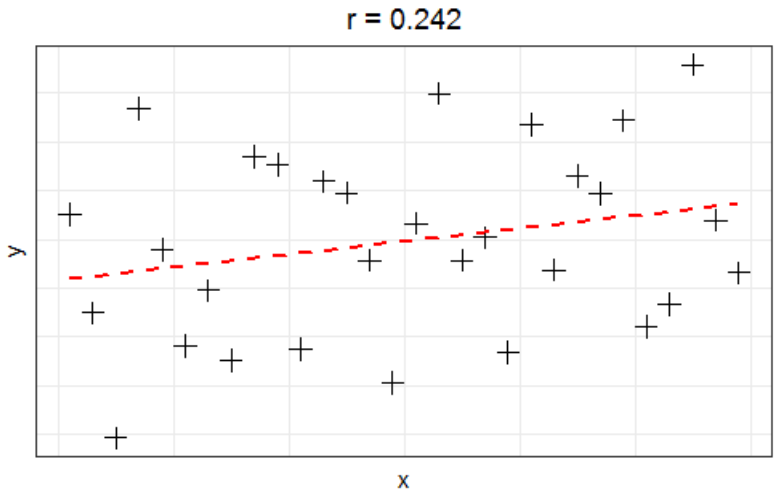
Pearsonův korelační koeficient



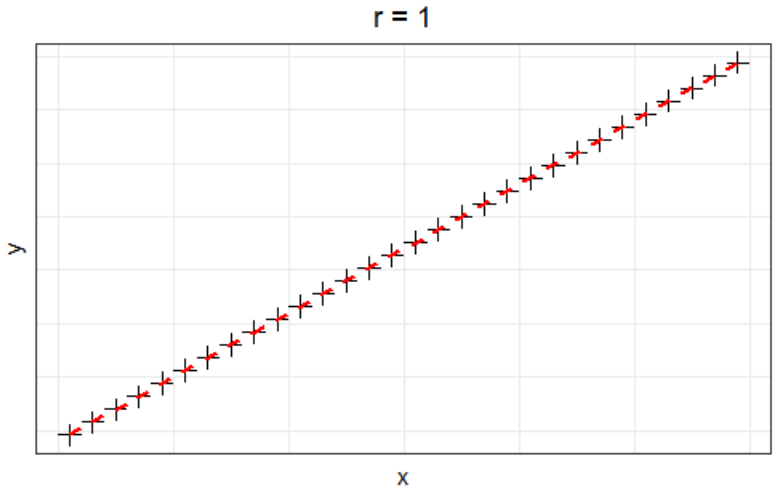


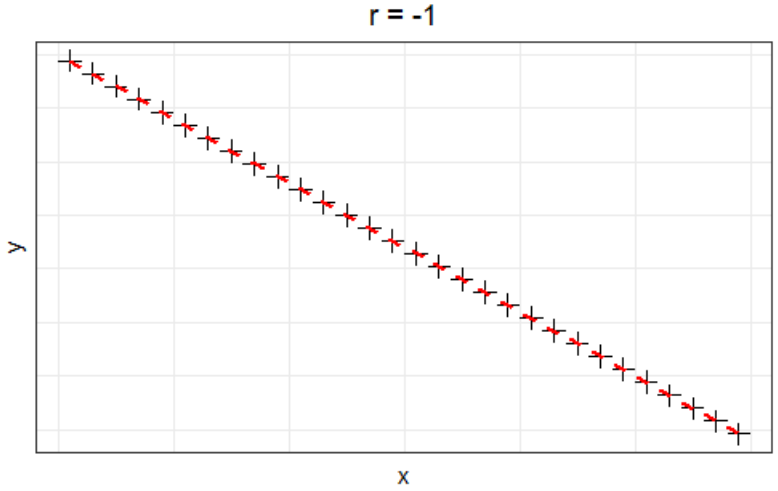


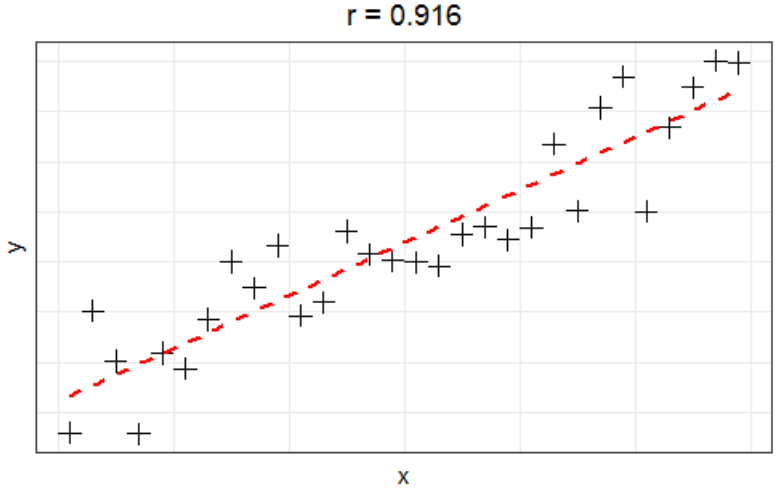


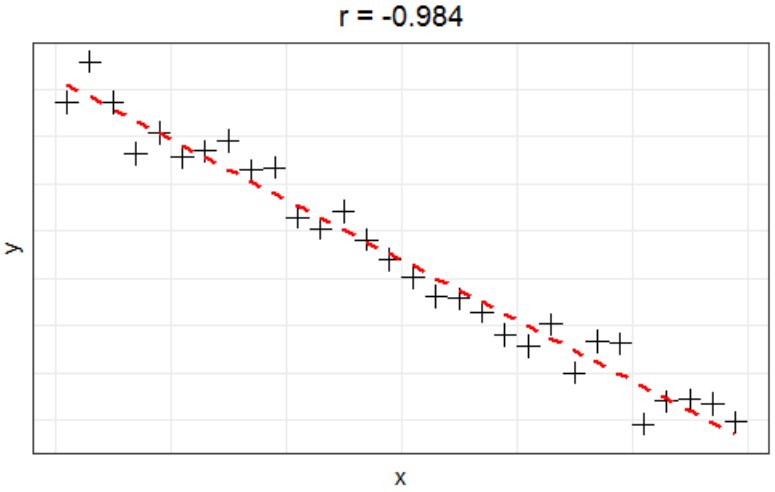


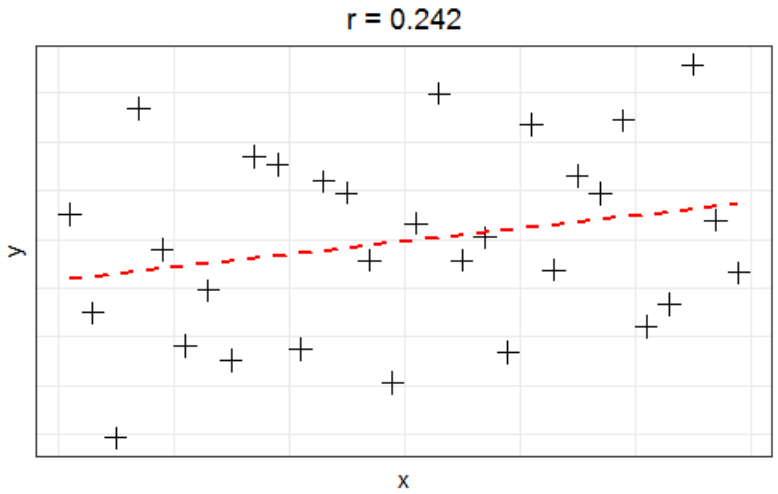
Pearsonův korelační koeficient

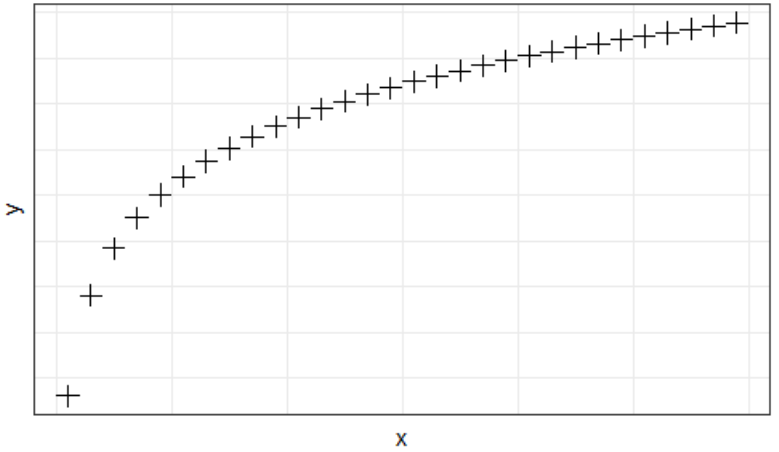




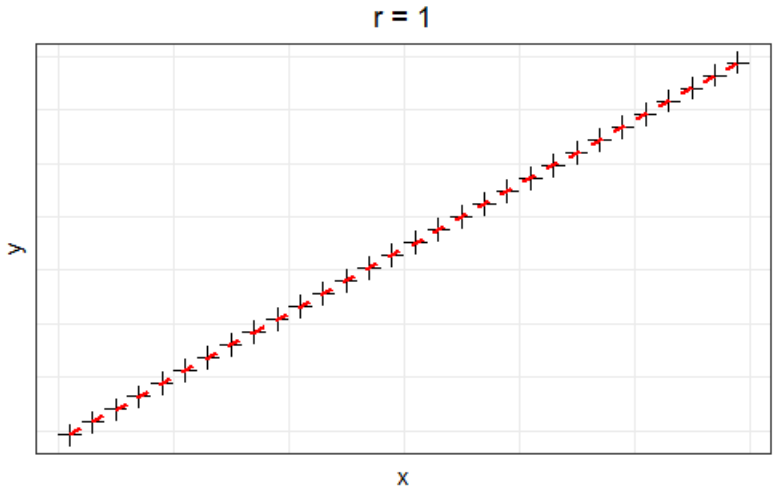


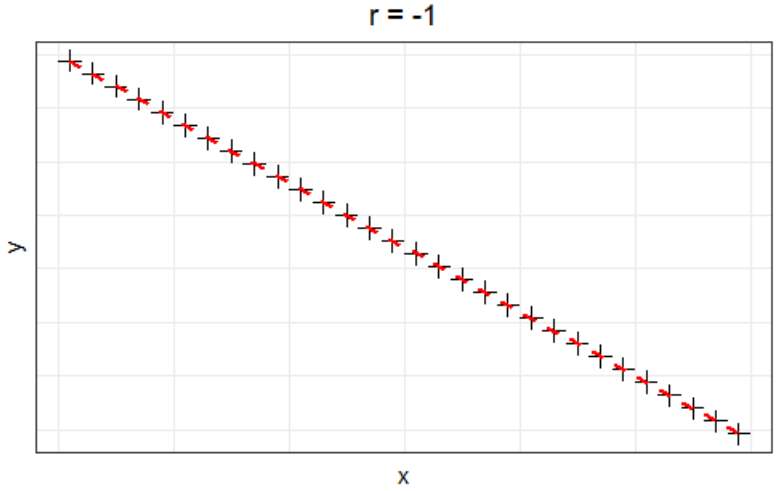


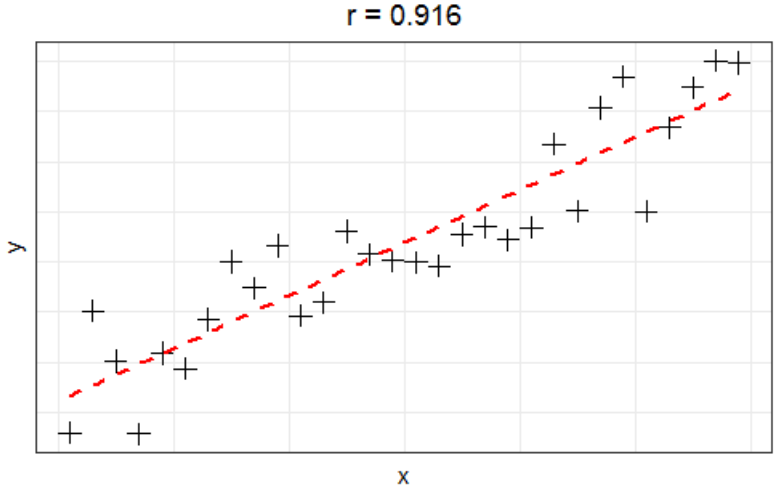


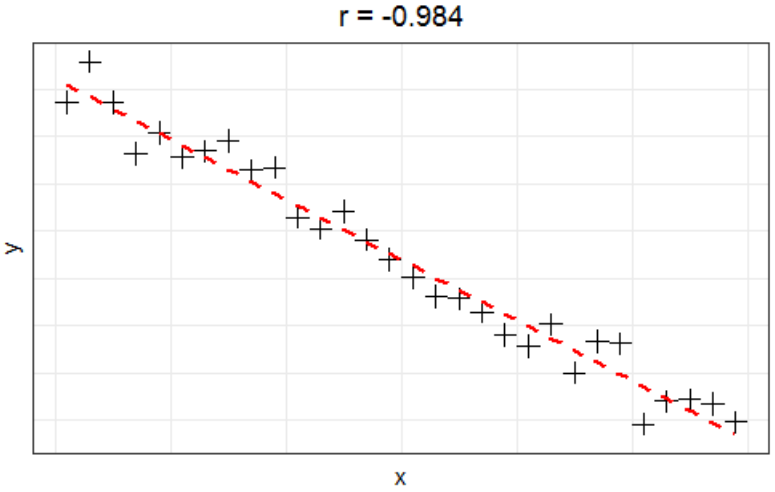


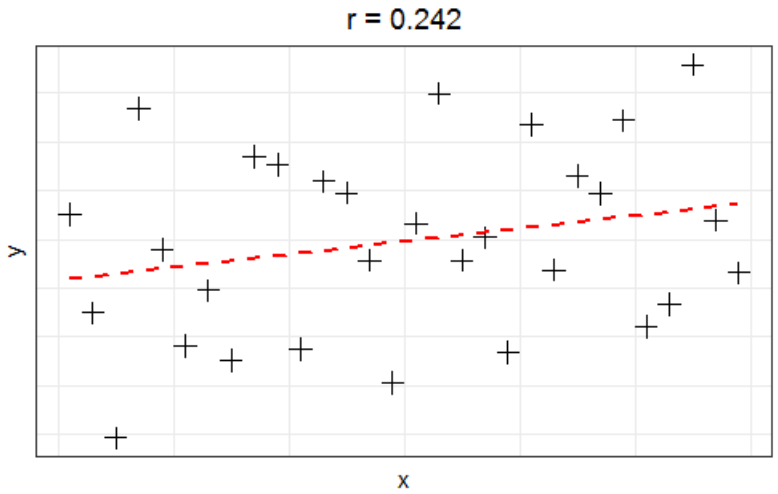
Pearsonův korelační koeficient

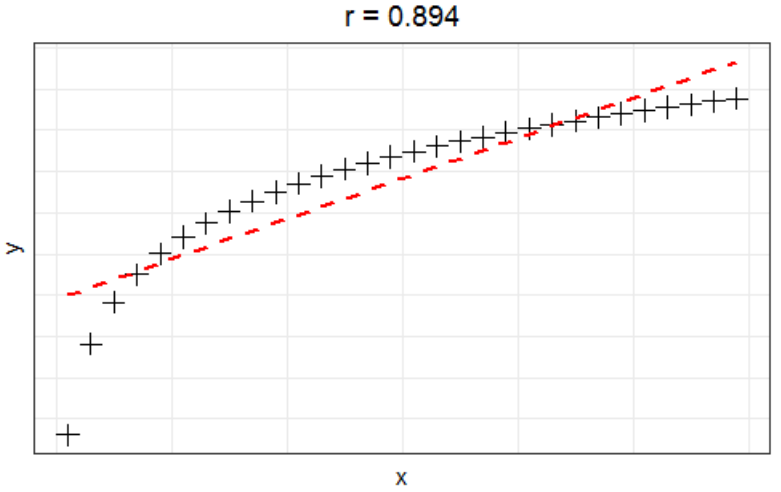




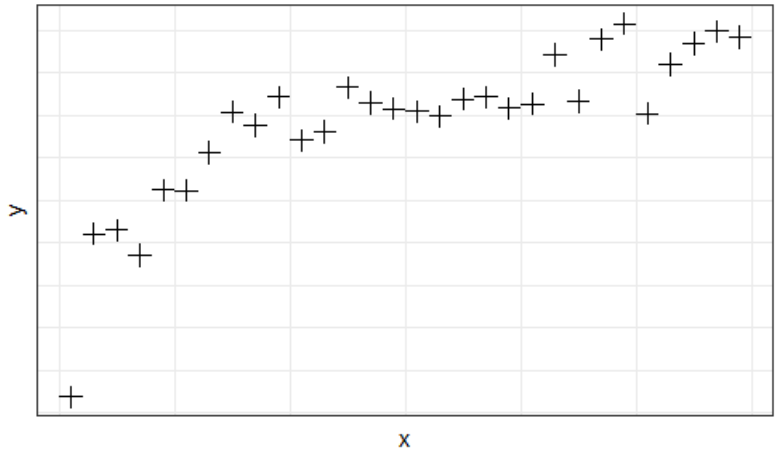




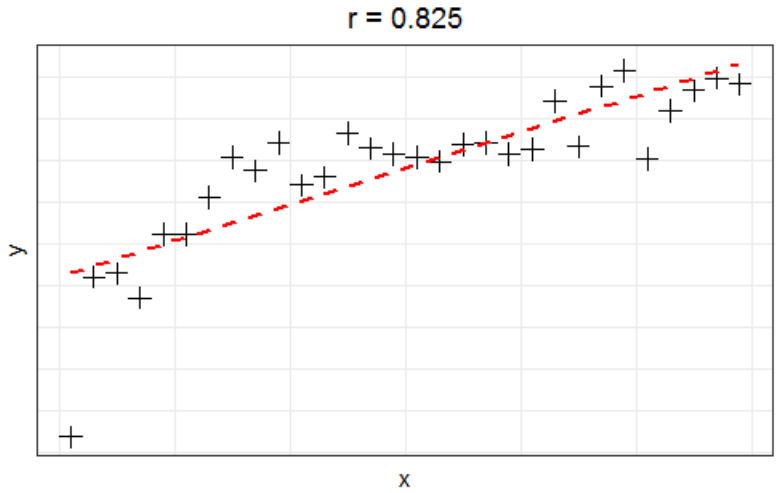




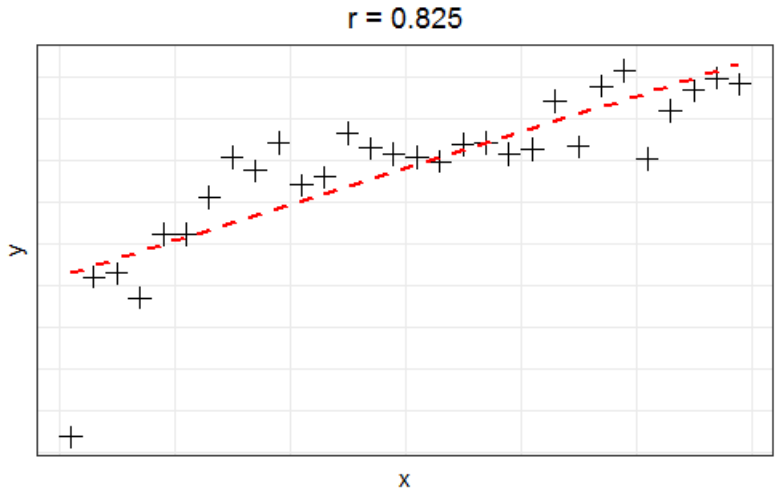
Pearsonův korelační koeficient

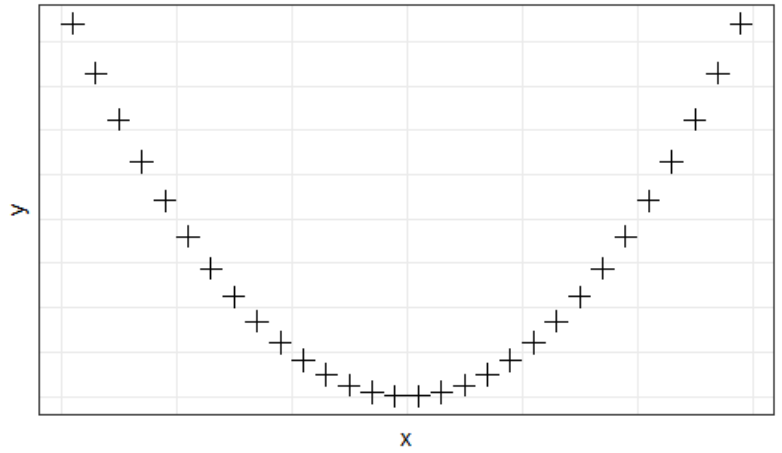


Pearsonův korelační koeficient

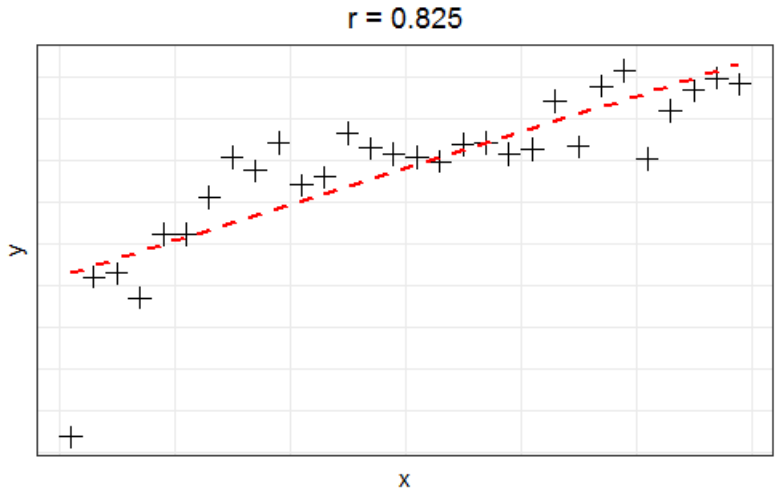


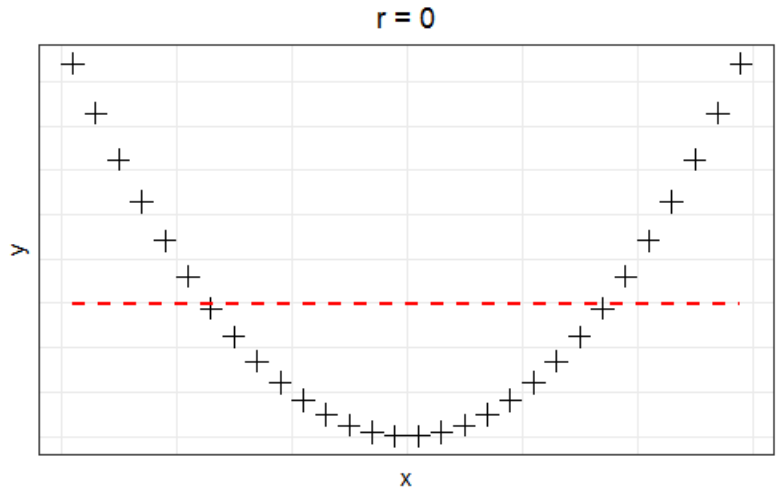
Pearsonův korelační koeficient



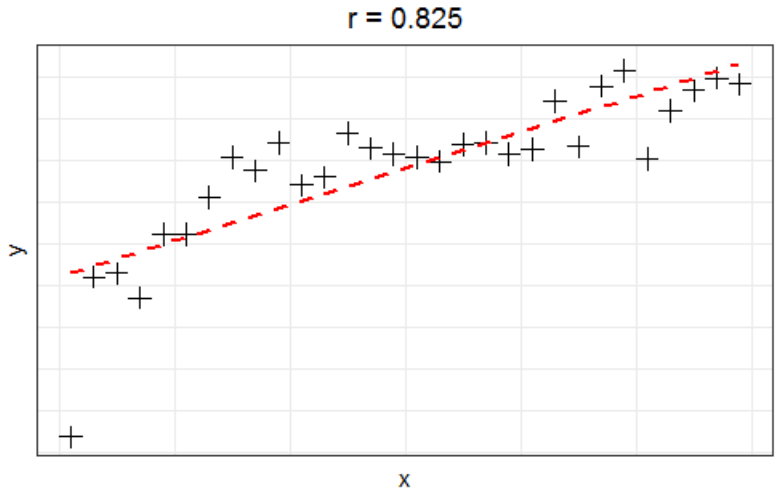


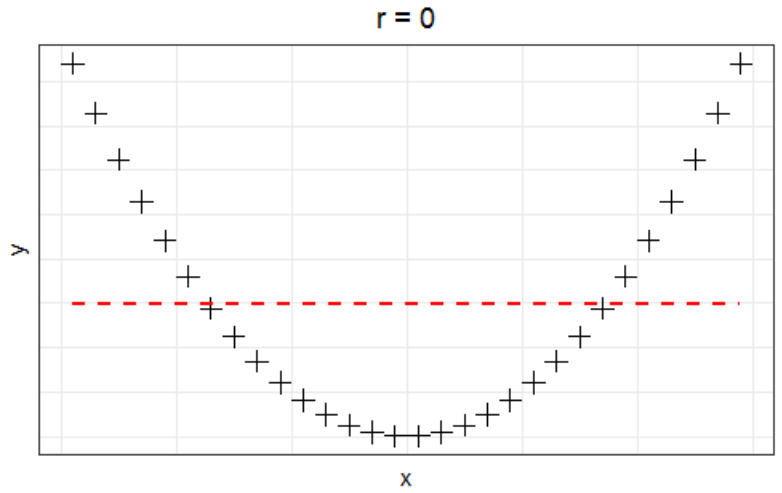
Pearsonův korelační koeficient

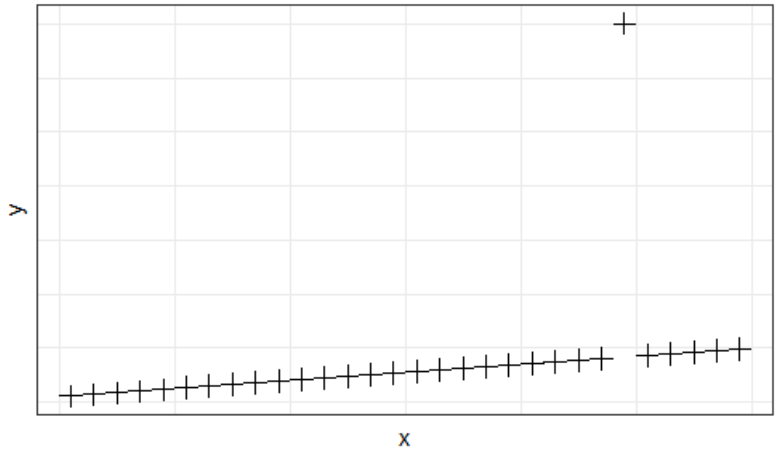




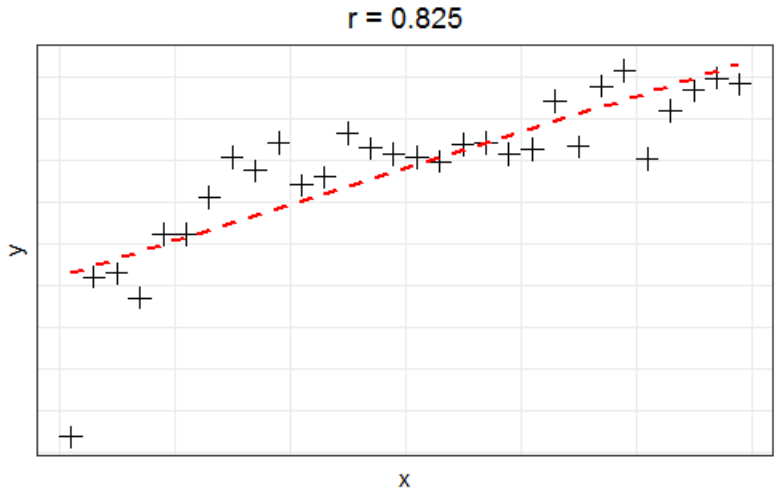
Pearsonův korelační koeficient

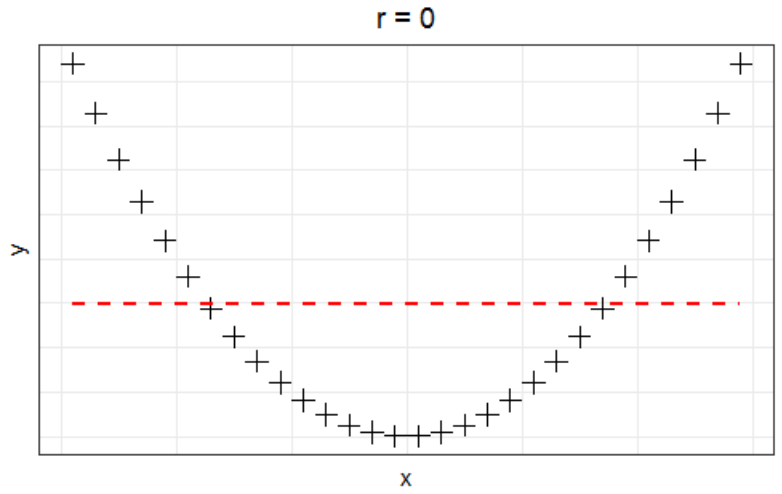


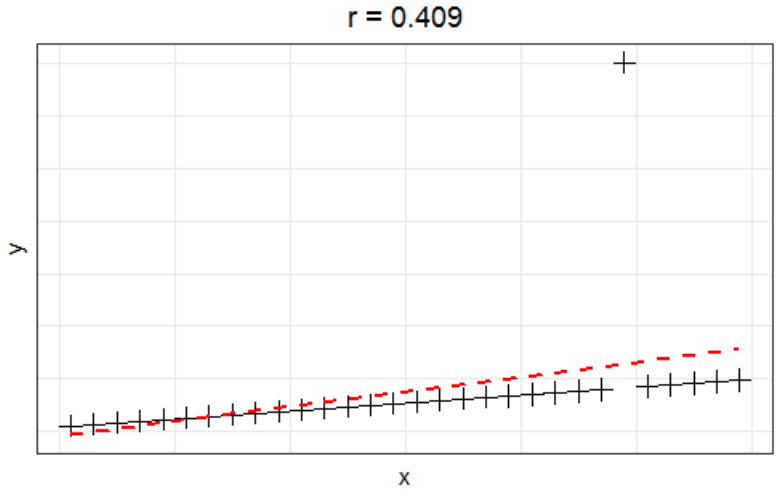




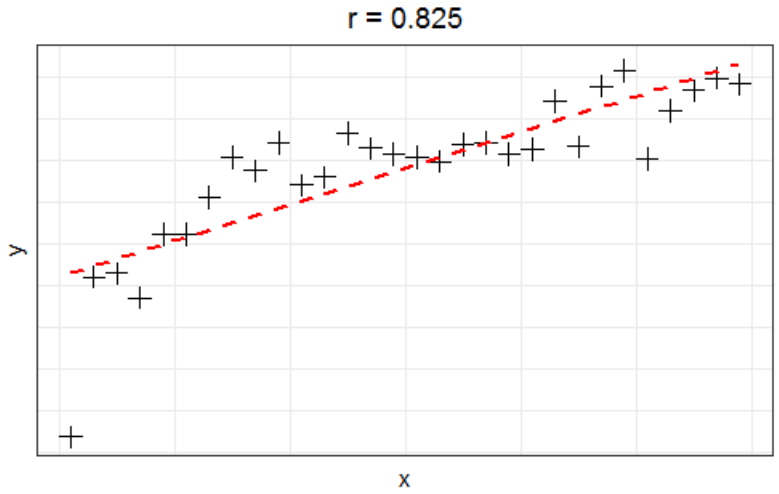
Pearsonův korelační koeficient

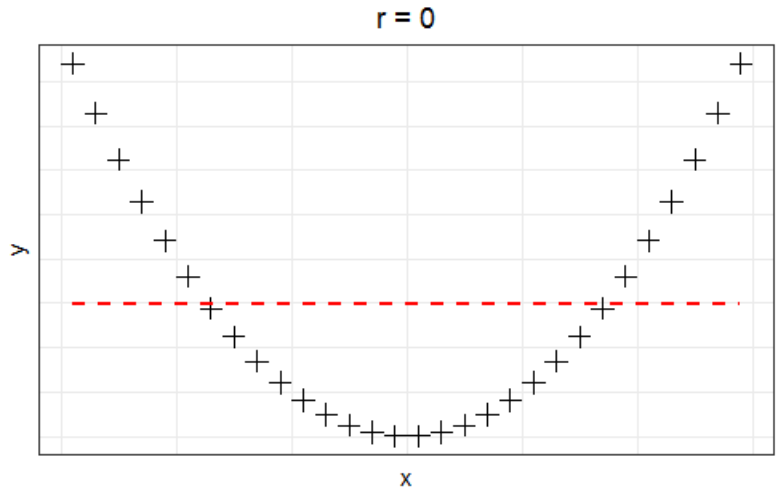


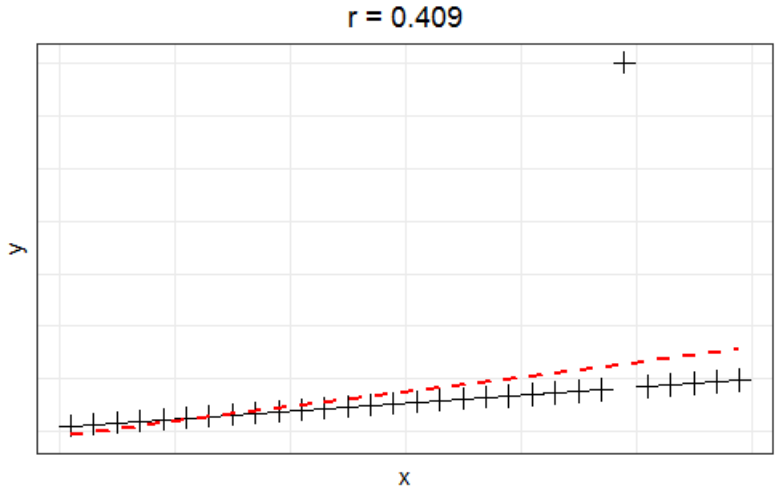


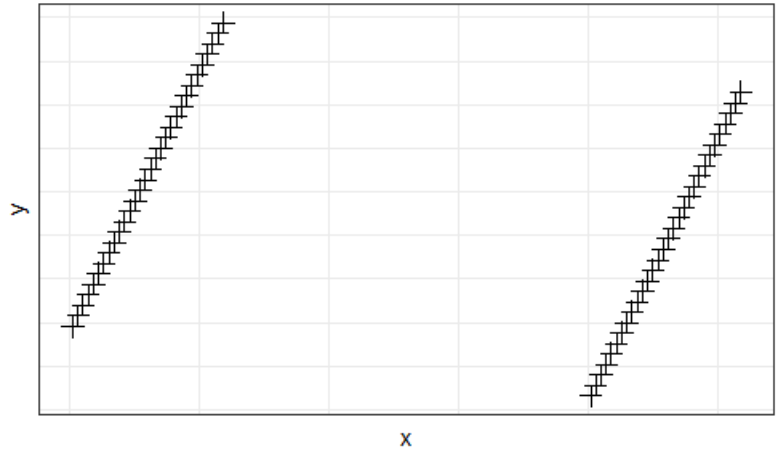


Pearsonův korelační koeficient

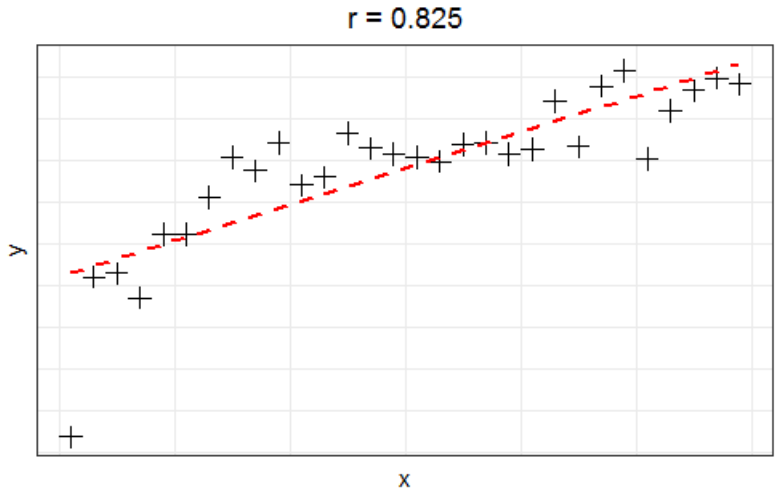


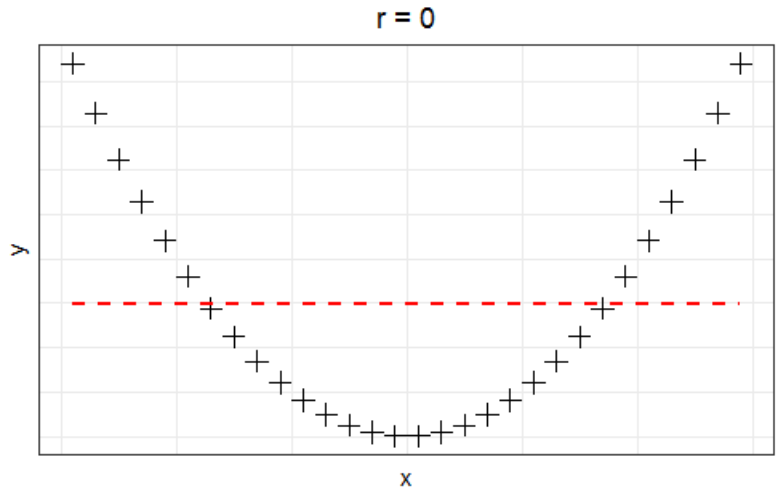


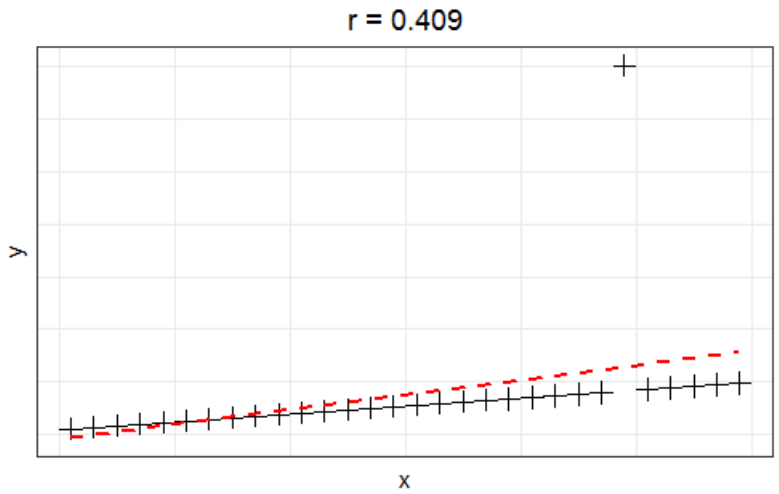


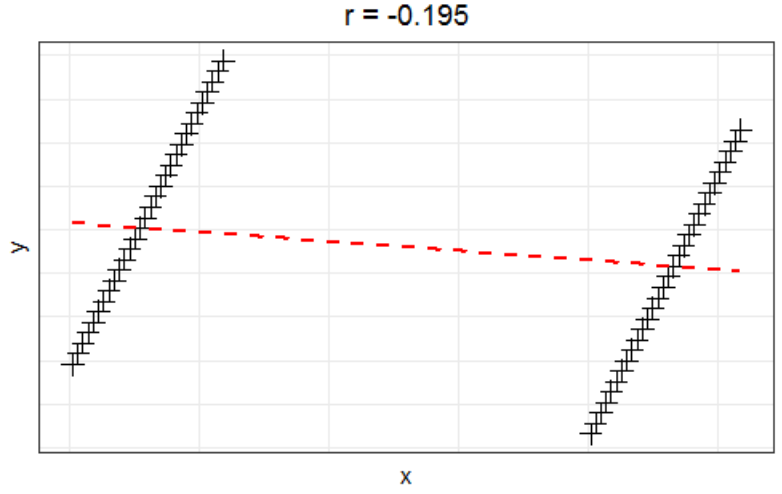


Pearsonův korelační koeficient

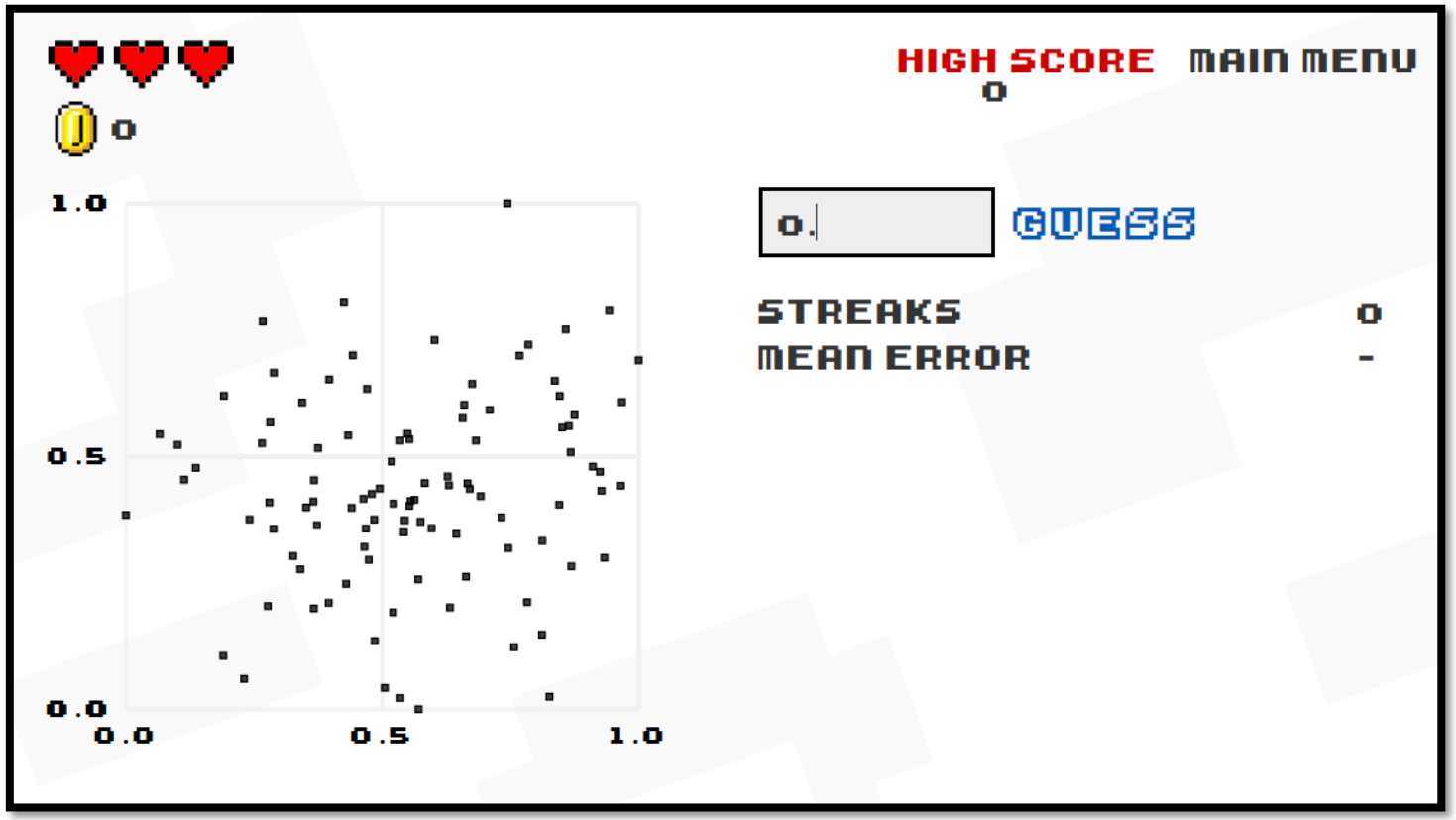






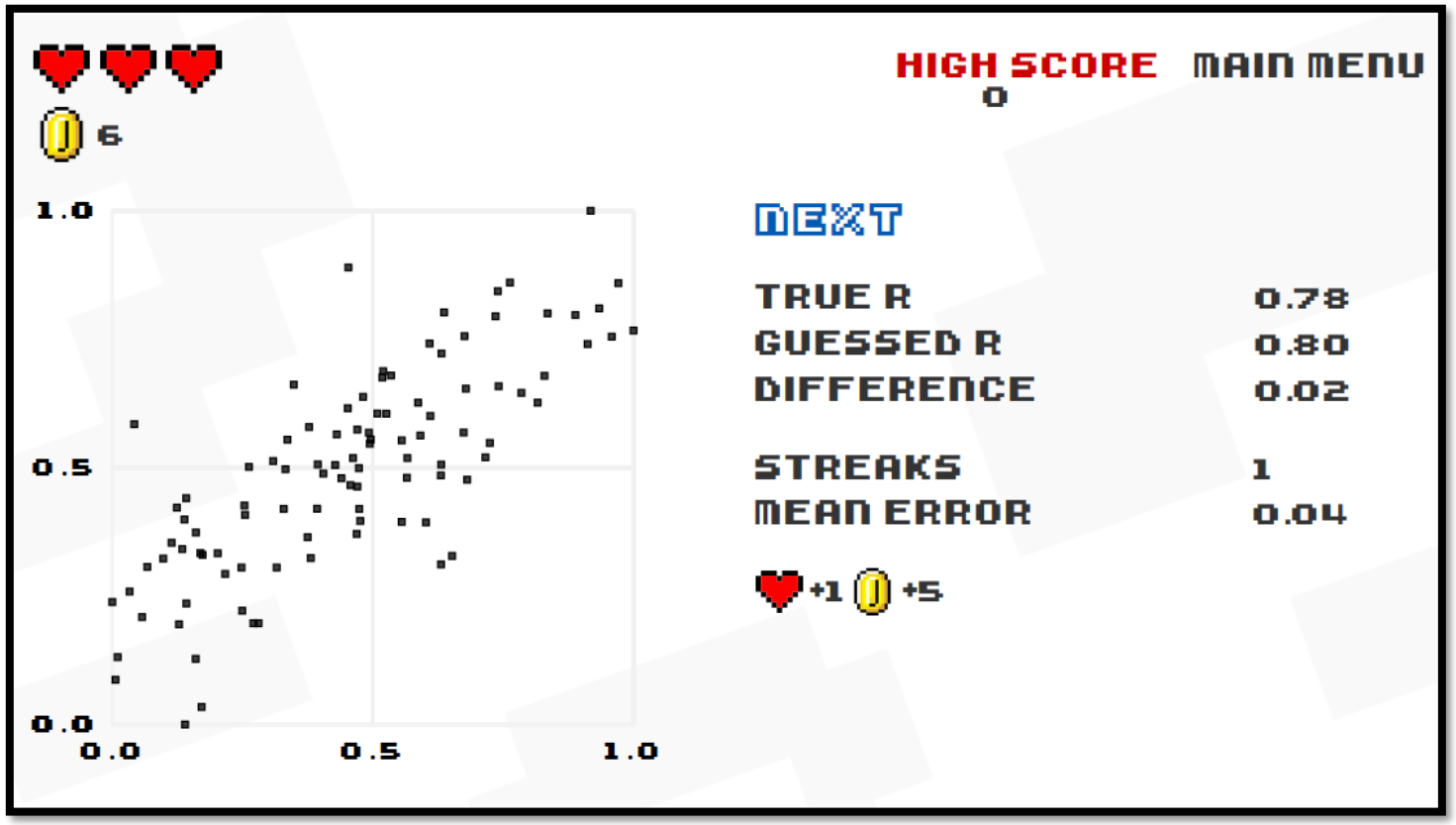


Pearsonův korelační koeficient - hra



Guess the Correlation Game
http://guessthecorrelation.com/

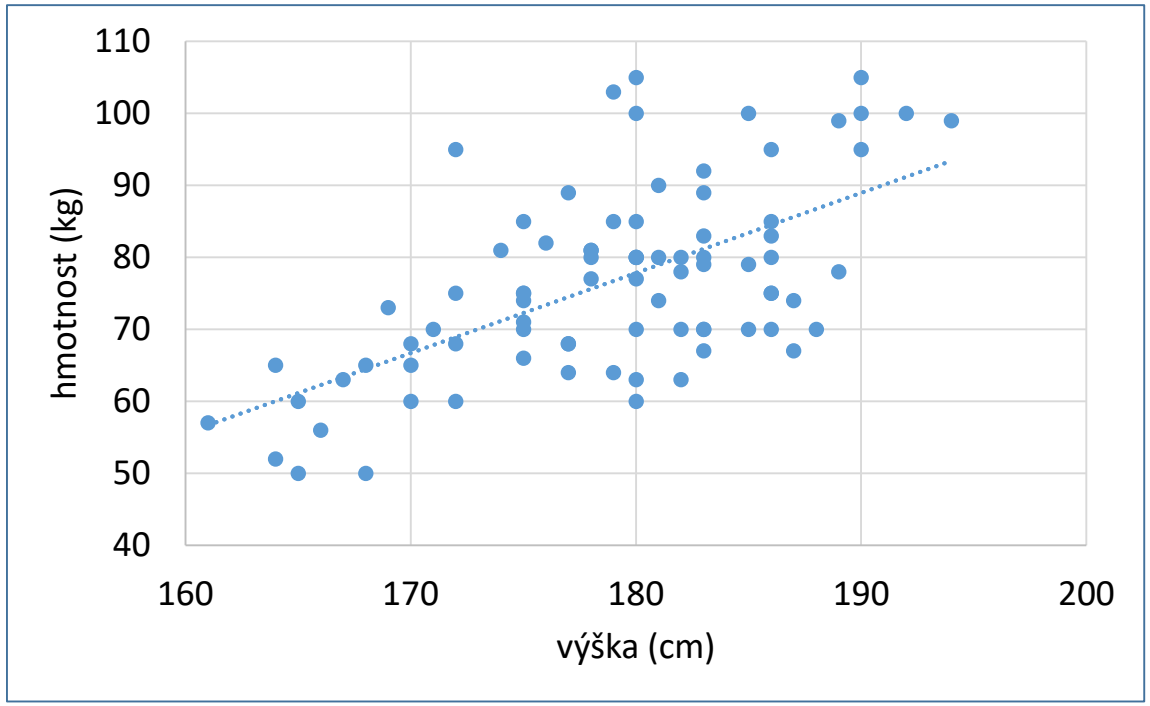
Pearsonův korelační koeficient - hra



Guess the Correlation Game
http://guessthecorrelation.com/

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**



- **Pearsonův výběrový korelační koeficient**

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

$$\boldsymbol{r = 0,6146}$$

Analýza závislosti dvou kvantitativních znaků

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v tydnu	praxe v oboru během studia	24	26

Tipněte si, které dva znaky jsou negativně korelovány.

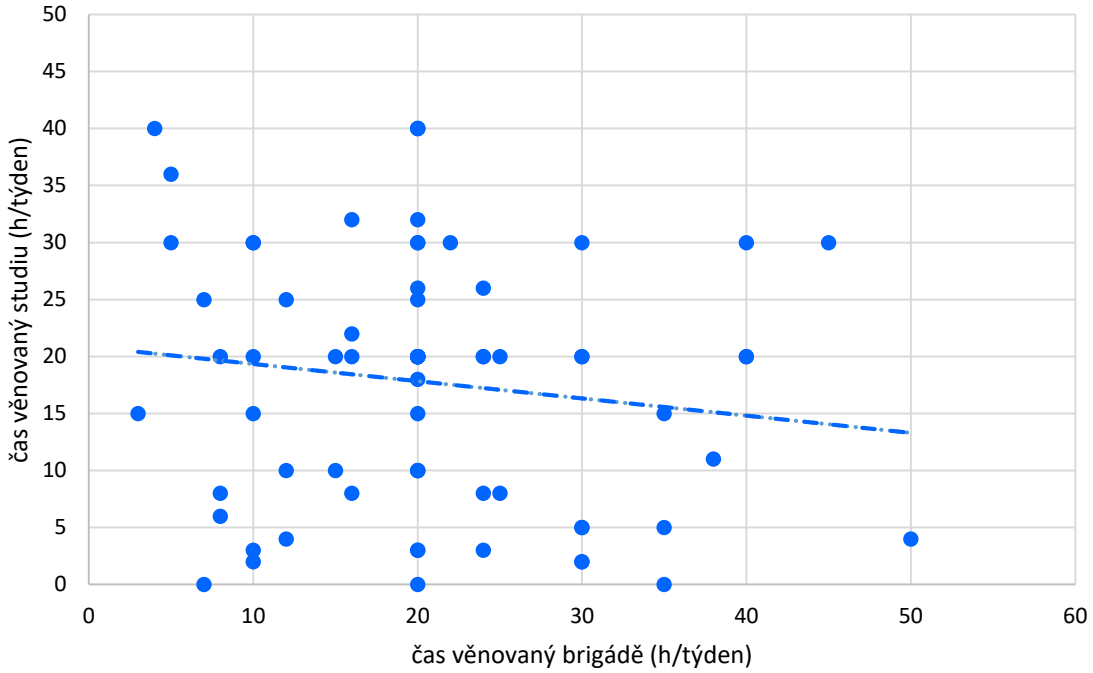
Analýza závislosti dvou kvantitativních znaků

Časová značka	Pohlaví	Výška (cm)	Váha (kg)	Přivyděláváte si v rámci prezenčního studia na brigádách?	Jak často brigádu máte?	Jak byste svou brigádu charakterizoval(a)?	Kolik času týdně obvykle věnujete brigádě?	Kolik času týdně obvykle věnujete studiu?
ID	pohlaví	výška (cm)	váha (kg)	brigáda	frekvence brigády	charakteristika brigády	čas věnovaný brigádě (h/týden)	čas věnovaný studiu (h/týden)
1.4.2016 10:38	muž	180	70	ano	každý pracovní den	praxe v oboru během studia	20	15
1.4.2016 10:41	muž	186	85	ano	nepravidelně	kancelářská práce a na ní navazující práce manuální při realizaci projektů	30	20
1.4.2016 10:41	muž	172	75	ano	nepravidelně	praxe v oboru během studia	5	36
1.4.2016 10:45	žena	166	56	ano	Různě, 2-3 týdně	Hlídání dětí	12	10
1.4.2016 10:52	žena	188	70	ano	3 dny v tydnu	praxe v oboru během studia	24	26

Tipněte si, které dva znaky jsou negativně korelovány.

Analýza závislosti dvou kvantitativních znaků

- **Vizualizace – bodový graf**



- **Pearsonův výběrový korelační koeficient**

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

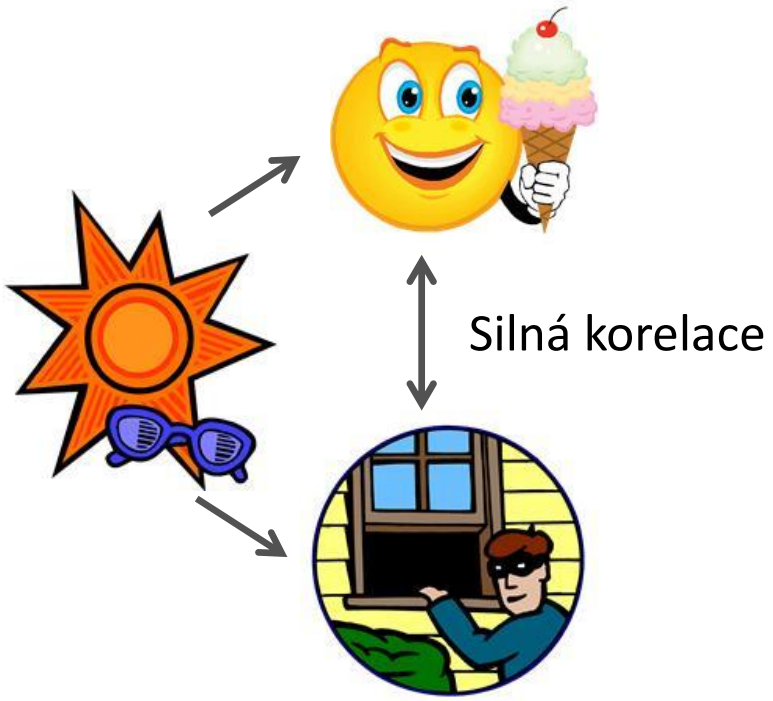
$$\boldsymbol{r = -0,1370}$$

Pozor na falešnou (zdánlivou) korelaci!

Pokud jsou dvě náhodné veličiny korelované, znamená to pouze to, že jsou lineárně závislé.

Nelze z toho však ještě usoudit, že by jedna z nich musela být příčinou a druhá následkem.

To samotná korelovanost nedovoluje rozhodnout.

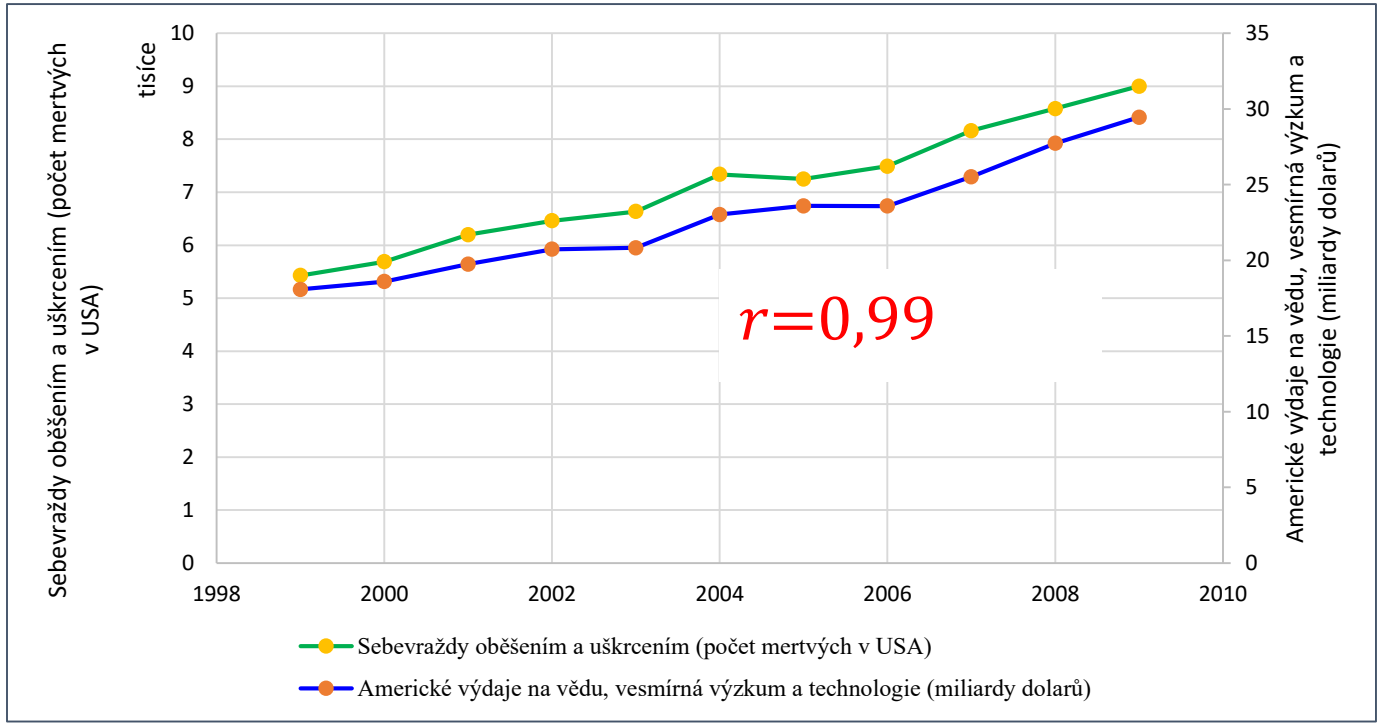


Pozor na falešnou (zdánlivou) korelaci!

Pokud jsou dvě náhodné veličiny korelované, znamená to pouze to, že jsou lineárně závislé.

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 hodnocení „síly“ korelace!

V praxi se zpravidla hodnota koeficientu korelace 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!

Spearmanův korelační koeficient

- Neparametrický korelační koeficient, který je robustní vůči odlehlým hodnotám a obecně odchylkám od normality.
- Spearmanův korelační koeficient je mírou monotónní závislosti mezi X a Y (nemusí jít o závislost lineární).
- Zjistíme-li, že výběrový korelační koeficient $r_{SP} \neq 0$, zpravidla nás zajímá, zda je indikovaná korelace statisticky významná, tj. velmi zjednodušeně řečeno, zda se korelační koeficient příslušných populačních dat statisticky významně liší od nuly.

Spearmanův korelační koeficient

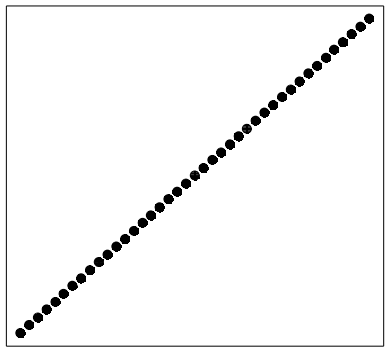
Mějme náhodný výběr $(X_1; Y_1), \dots, (X_n; Y_n)$ z dvourozměrného rozdělení. Nechť $R_{X_1}, \dots, R_{X_n}$ jsou pořadí veličin $X_1, \dots, X_n$ a nechť $R_{Y_1}, \dots, R_{Y_n}$ jsou pořadí veličin $Y_1, \dots, Y_n$.

$$r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} \left(R_{X_1} - R_{Y_1}\right)^2$$

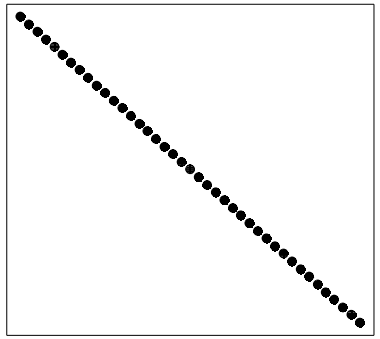
Pokud se v náhodných výběrech, z nichž je r_S počítán, vyskytuje mnoho shod (tj. stejně velkých pozorování), doporučuje se používat korigovaný Spearmanův korelační koeficient $r_{S_{korig}}$.

$$r_{S_{korig}} = 1 - \frac{6}{n^3 - n - T_X - T_Y} \sum_{i=1}^{n} \left(R_{X_1} - R_{Y_1}\right)^2$$

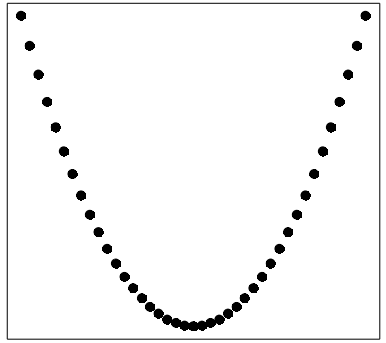
Srovnání Pearsonova a Spearmanova korel. koeficientu



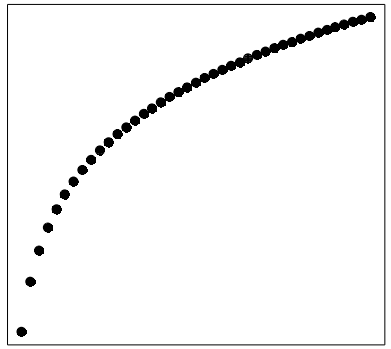
$\rho(X,Y) = 1,000$
$\rho_S(X,Y) = 1,000$



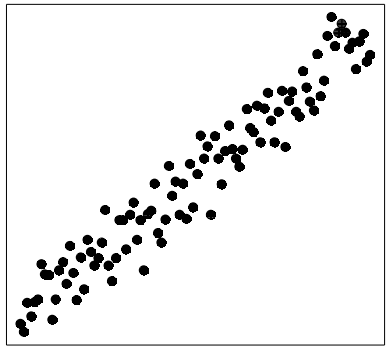
$\rho(X,Y) = -1,000$
$\rho_S(X,Y) = -1,000$



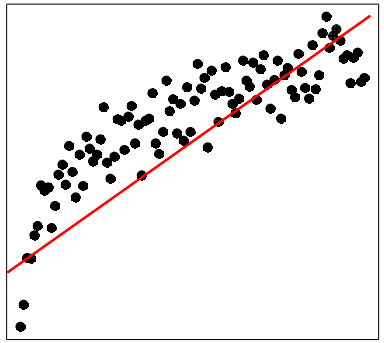
$\rho(X,Y) = 0,000$
$\rho_S(X,Y) = 0,000$



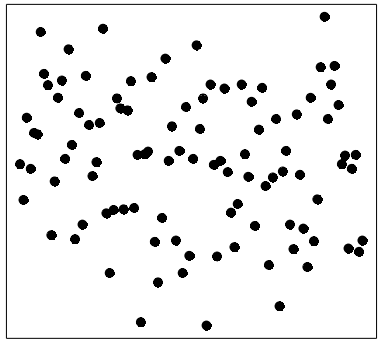
$\rho(X,Y) = 0,934$
$\rho_S(X,Y) = 1,000$



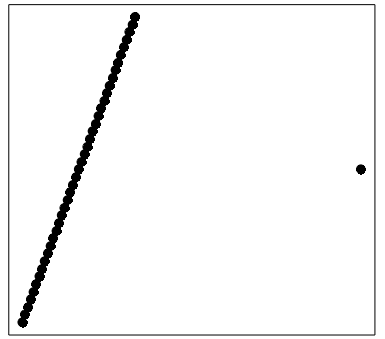
$\rho(X,Y) = 0,967$
$\rho_S(X,Y) = 0,981$



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



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



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

Korelační matice

Nechť $\boldsymbol{X} = (X_1, \dots, X_k)^T$ je náhodný vektor, potom matici

$$R(\boldsymbol{X}) = \begin{pmatrix} r(X_1, X_1) & \cdots & r(X_1, X_k) \\ \vdots & \ddots & \vdots \\ r(X_k, X_1) & \cdots & r(X_k, X_k) \end{pmatrix}$$

nazýváme korelační matici náhodného vektoru $\boldsymbol{X}$.

Koeficient mnohonásobné korelace

Nechť $\boldsymbol{Z} = (Y, X_1, \dots, X_k)^T$. Označme $\boldsymbol{X} = (X_1, \dots, X_k)^T$. Pak koeficient mnohonásobné korelace $r(Y \cdot \boldsymbol{X})$ mezi náhodnou veličinou Y a náhodným vektorem $\boldsymbol{X}$ vypočteme dle

$$r(Y \cdot \boldsymbol{X}) = \sqrt{1 - \frac{\det(R(\boldsymbol{Z}))}{\det(R(\boldsymbol{X}))}}$$

- Druhou mocninu $r(Y \cdot \boldsymbol{X})$ nazýváme index determinace a značíme ji R^2.
- Index determinace je standardním výstupem lineární regrese.

Parciální korelační koeficient

- Korelační koeficient $r(X,Y)$ mezi náhodnými veličinami X a Y může být vysoký proto, že obě náhodné veličiny jsou silně závislé na náhodném vektoru $\boldsymbol{Z} = (Z_1, \dots, Z_n)^T$.
- Proto nás mnohdy zajímá, jaká je korelace mezi X a Y při vyloučení vlivu, který je způsoben vlivem náhodného vektoru $\boldsymbol{Z}$, tj. parciální korelační koeficient $r(X, Y \cdot \boldsymbol{Z})$.

$$r(X, Y \cdot \boldsymbol{Z}) = \frac{r(X,Y) - r(X \cdot \boldsymbol{Z}) \cdot r(Y \cdot \boldsymbol{Z})}{\sqrt{\left(1 - r^2(X \cdot \boldsymbol{Z})\right) \cdot \left(1 - r^2(Y \cdot \boldsymbol{Z})\right)}}$$

Příklad:

H … tělesná hmotnost, V … tělesná výška, S … výkon ve skoku vysokém (dle M. Sebery)

$$\left.\begin{array}{l} r(H,V) = 0{,}91 \\ r(V,S) = 0{,}86 \\ r(H,S) = 0{,}69 \end{array}\right\} \quad r(H, S \cdot V) = \frac{0{,}69 - 0{,}91 \cdot 0{,}86}{\sqrt{(1-0{,}91^2)(1-0{,}86^2)}} = -0{,}44$$

Děkuji za pozornost!

martina.litschmannova@vsb.cz

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