Asociace / souvislost

- Hlavním účelem analýzy dat s dvěma proměnnými je zjistit zda je mezi proměnnými souvislost a popsat povahu této souvislosti
- Souvislost existuje, pokud nějaká hodnota jedné proměnné se vyskytne pravděpodobněji s určitými hodnotami jinými proměnnými
  - př. 1. Přežít určité období je pravděpodobnější pro kuřáky než pro nekuřáky. 2. V případě vyššího užití energie, má úroveň CO2 v atmosféře tendenci být vyšší? 3. Daňoví poplatníci vyšší příjmové skupiny mají tendenci být pravděpodobněji kontrolováni než poplatníci nižších příjmových skupin. 4. Pokud dám první koš, mám vyšší pravděpodobnost dát druhý koš, než pokud ho nedám.
- 2 otázky:
  - Existuje souvislost?
  - Jak je silná?

Závislá a nezávislá proměnná

- Jak výsledek závislé proměnné závisí / je vysvětlen hodnotou nezávisle proměnné
- Závislá proměnná = výsledek který je porovnáván
- Nezávislá
  - Kategorická = definuje skupiny které srovnáváme vzhledem k hodnotám závislé proměnné
  - Kvantitativní = definuje jak změna mezi různými numerickými hodnotami souvisí s hodnotami výsledné proměnné

Souvislost podle typu proměnné

- 3 základní typy situací:
  - Obě proměnné kategorické
    - Zobrazujeme pomocí kontingenčních tabulek a asociaci zkoumáme prostřednictvím podmíněných proporcí/pravděpodobností
  - Kvantitativní a kategorická proměnná
    - Srovnáváme kategorie nezávisle proměnné (pohlaví) podle velikosti závislé proměnné (př. příjem) na základě měr centrální tendence a variability (např. průměrný příjem)
  - Obě kvantitativní
    - Analyzujeme jak se výsledek závisle proměnné mění když se mění hodnota nezávisle proměnné
    - Zobrazujeme bodové rozptýlení (regulujeme extrémní/odlehlé hodnoty)

Asociace mezi kategorickými proměnnými

- Př. typ jídla (bio/běžné) vs. úroveň pesticidů (vysoká/nízká)
  - Proces: „křížení“ *(crosstabulation)* / třídění 2. stupně
  - Způsob zobrazení (výsledek): kontingenční tabulka
    - Obsahuje kombinace kategorií obou proměnných
    - Každá kombinace = buňka
    - Podmíněné proporce vs. marginální proporce
- Asociaci zjistíme srovnáním podmíněných proporcí – je proporce potravin s obsahem pesticidů stejná u bio potravin a běžných potravin? Ano = nezávislost, Ne = souvislost

Míry měření asociace mezi kategorickými proměnnými

- Míra asociace je statistika která sumarizuje sílu závislosti mezi dvěma proměnnými
  - Rozdíl v podmíněných proporcích
    - (-1 až 1) , 0 = nezávislost (např. 0.6 - 0.6 = 0), 1 a -1 extrémní souvislost (např. 0 - 1)
  - Poměr podmíněných proporcí (relativní riziko=RR, také „risk ratio“)
    - (0 až nekonečno), 1 = nezávislost, čím dále od 1 tím větší závislost, nicméně RR = 4 (např. 0.8/0.2) a RR=0.25 (0.2/0.8) představují stejně silný vztah
  - Poměr šancí (OR, také „odds ratio“)
    - (0 až nekonečno), 1 = nezávislost
  - Statistiky založené na chí-kvadrátu (Phí, Cramerovo V)…naučíme se později

Relativní riziko

Příjem	*Prošlo kontrolou*	*Neprošlo kontrolou*	*Celkem*
Pod 200tis	**0,01 (1260)**	0,99 (132147)	1 (133407)
200tis-1mil	**0,03 (131)**	0,97 (4311)	1 (4442)
Více než 1mil	**0,07 (22)**	0,93 (371)	1 (393)

- RR (pod 200tis. Vs. 200-1.mil.) = 0,01 / 0,03 = 0,3333
- Nebo RR (200-1.mil. Vs. pod 200tis.) = 0,03 / 0,01 = 3

	2koš	*2mimo*	*celkem*
1koš	**0,8 (41)**	0,2 (10)	1 (51)
1mimo	**0,8 (10)**	0,2 (3)	1 (13)

- RR = 0,8 / 0,8 = 1

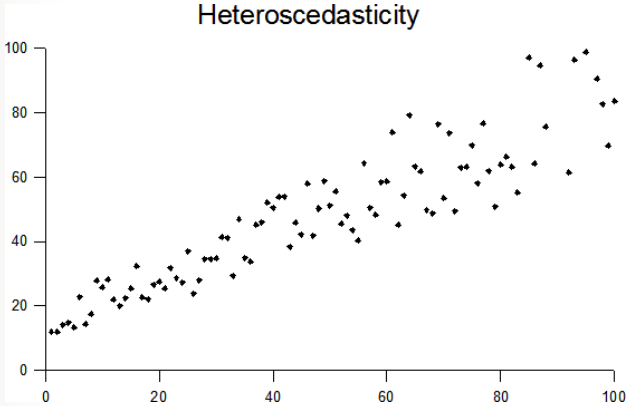
Poměr šancí (odds ratio)

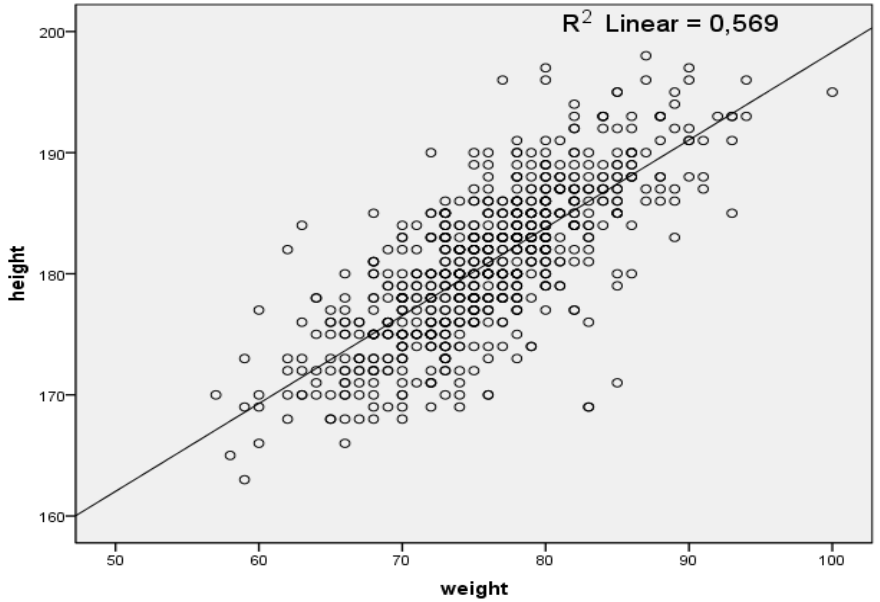
Příjem	*Prošlo kontrolou*	*Neprošlo kontrolou*	*Celkem*
Pod 200tis	0,0091 (1260)	0,9559 (132147)	0,9650 (133407)
200tis-1mil	0,0009 (131)	0,0312 (4311)	0,0321 (4442)
Více než 1mil	0,0002 (22)	0,0027 (371)	0,0029 (393)
Celkem	0,0102 (1413)	0,9898 (136829)	1 (138242)

- Šance (odds) projít kontrolou spíše než neprojít kontrolou v případě příjmu pod 200tis = 1260 / 132147 = 0,0095348
- Šance (odds) projít kontrolou spíše než neprojít kontrolou v případě příjmu 200tis.-1mil. = 131 / 4311 = 0,030387
- Poměr šancí (OR, odds ratio) = 0,0095348 / 0,030387 = 0,3137
  - Alternativní výpočet do kříže (1260*4311)/(131*132147)=5431860/17311257=0,3137

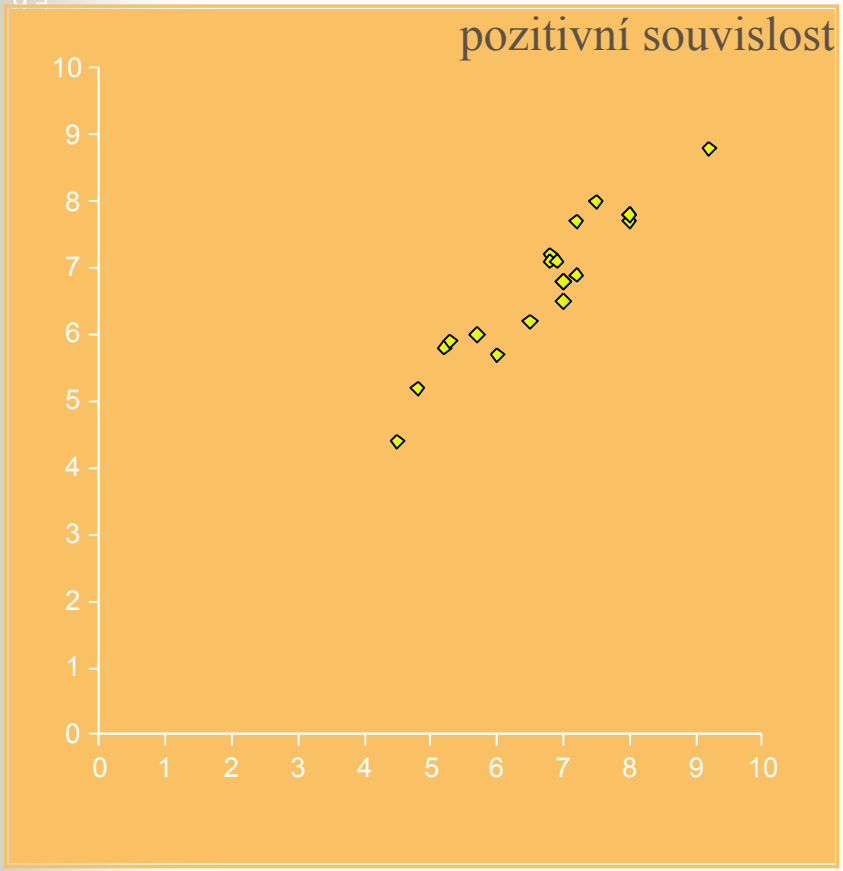
Souvislost mezi dvěma kvantitativními proměnnými

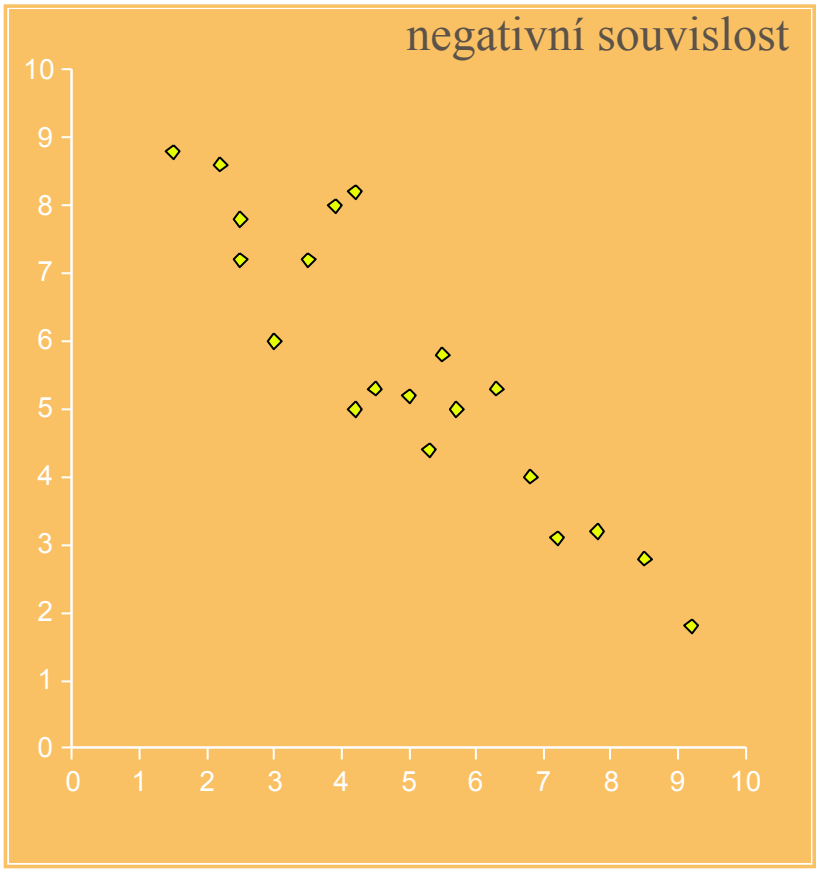
- Prozkoumáme bodové rozptýlení…
  - Bodové rozptýlení: bod = hodnoty na proměnných x a y pro daného člověka
    - Čeho si lze všimnout:
      - Jasný trend: Lidé s vyšší váhou mají obecně vyšší výšku.
      - Různost ve výšce v rozsahu cca 10 až 15 cm je relativně konstantní bez ohledu na váhu člověka (homogenní rozptyl)
      - Člověk s velkou vahou (83kg) a malou výškou (169cm) je relativně neobvyklý případ vzhledem k celkovému trendu



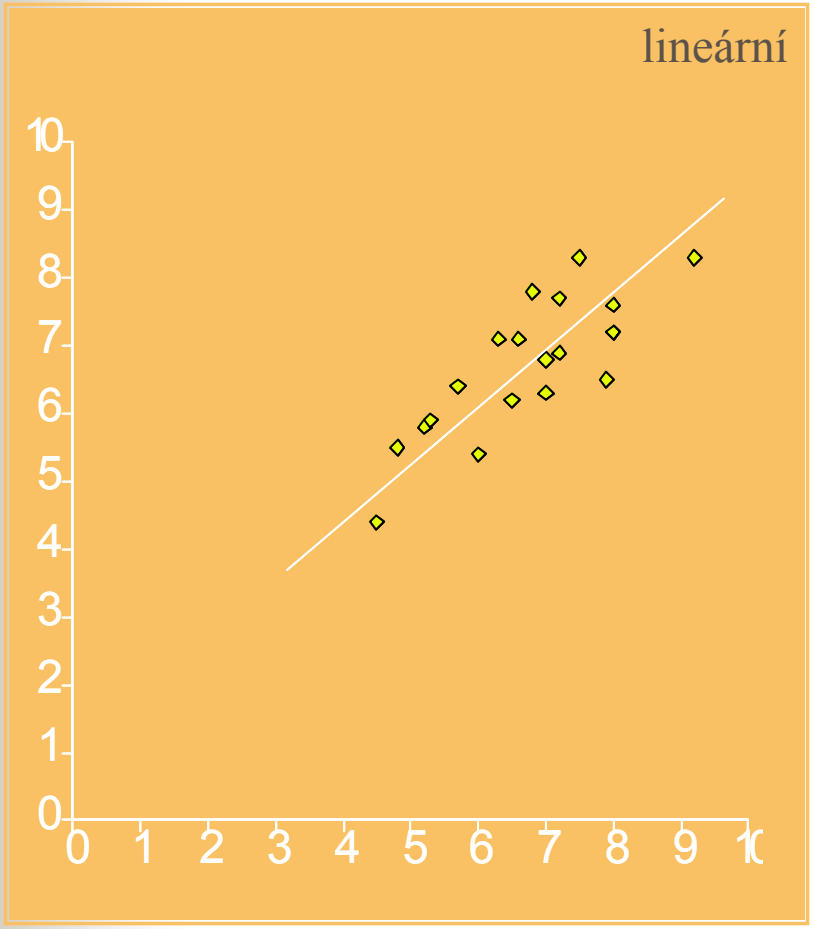


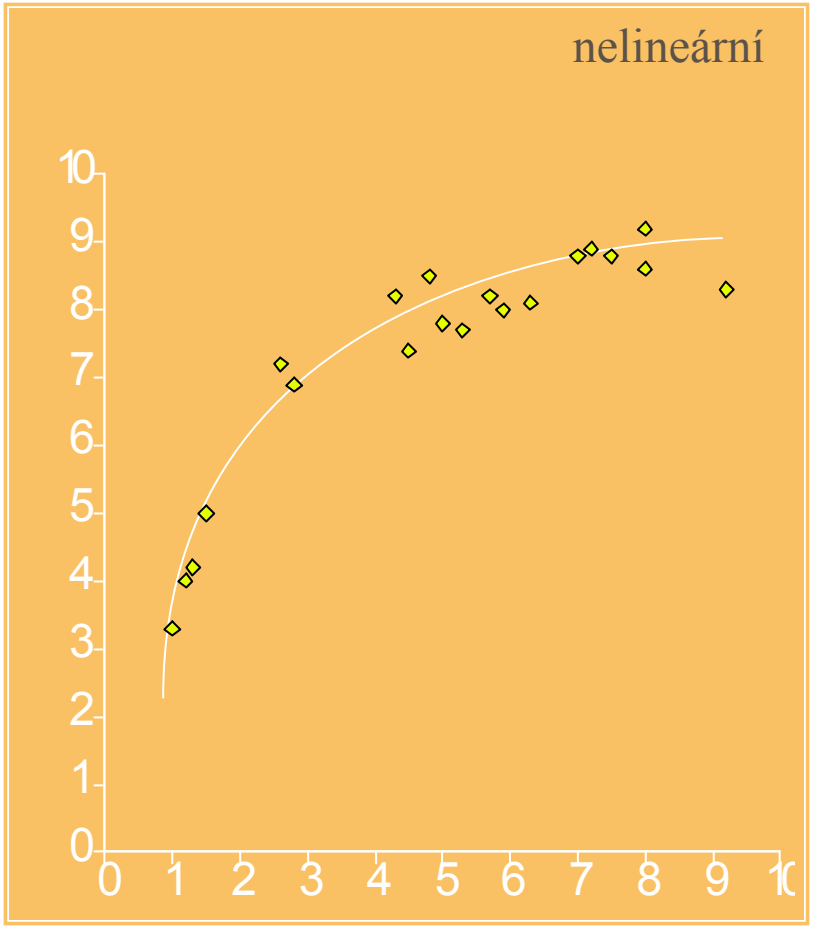
- Prozkoumáme bodové rozptýlení…(2)
- Vypadá vztah pozitivně nebo negativně?
  - Pozitivní souvislost=když vysoké hodnoty x mají tendenci vyskytovat se s vysokými hodnotami y (když x jde nahoru, y má tendenci jít také nahoru)
  - Negativní souvislost=když vysoké hodnoty x mají tendenci vyskytovat se s nízkými hodnotami y (když x jde nahoru, y má tendenci jít dolů)



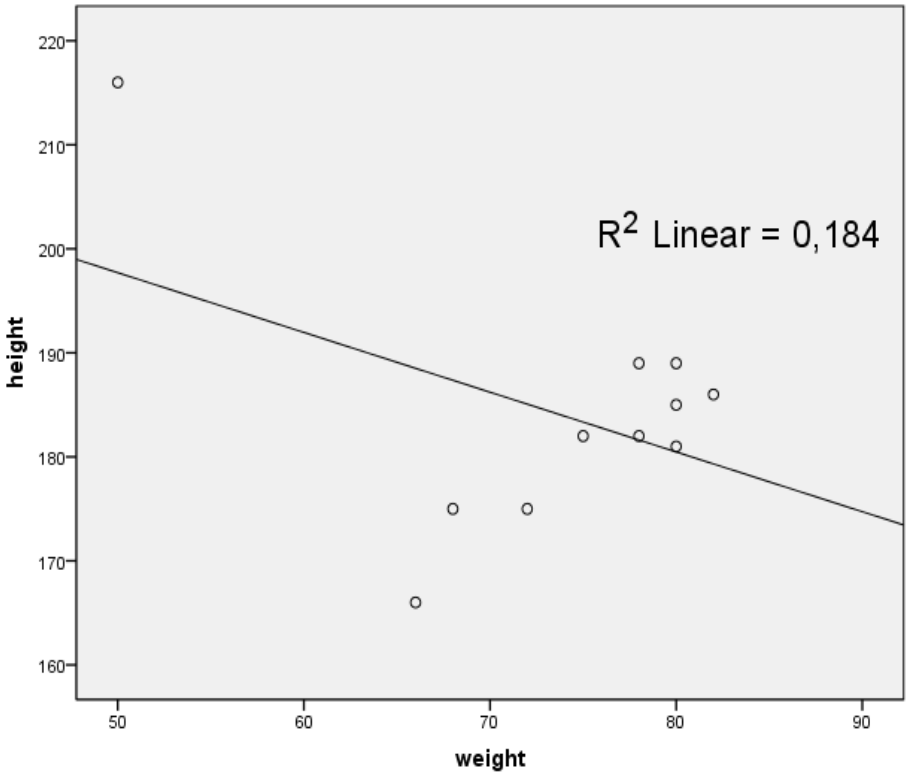


- Prozkoumáme bodové rozptýlení…(3)
  - Je trend v datech lineární tj. dá se aproximovat přímkou? Pokud ano, jak blízko přímky body leží?





- Prozkoumáme bodové rozptýlení…(4)
  - Jsou některá pozorování neobvyklá, odporující celkovému trendu?
  - Př. váha=50kg, výška=215cm



Měření síly souvislosti: dvě kvantitativní proměnné

- koeficient korelace r = -1 až 1
  - mezní hodnoty -1 a 1 značí absolutní souvislost
  - hodnota 0 značí absolutní nezávislost
- různé druhy korelačních koeficientů, použití se liší podle druhu dat, typu závislosti a typu rozložení
  - nejčastěji používané:
    - Pearsonův koeficient součinové korelace
    - Spearmanův koeficient pořadové korelace

Pearsonův koeficient součinové korelace

- Indikuje směr a sílu lineární souvislosti mezi dvěma kvantitativními proměnnými
- Nabývá hodnot -1 až 1
  - Negativní hodnota r indikuje negativní souvislost, pozitivní r značí pozitivní souvislost
  - Čím více se hodnota blíží +-1, tím blíže přímce body přiléhají a tím silnější je souvislost
  - Čím blíže 0, tím slabší souvislost
- Hodnota koeficientu nezávisí na jednotce měření (např. pokud změníme jednotku z centimetrů na milimetry, korelace se nemění)
- Dvě proměnné mají stejnou korelaci, bez ohledu na to, kterou z nich vnímáme jako závislou resp. nezávislou

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

Výpočet Pearsonova korelačního koeficientu

- Na pearsonův korelační koeficient lze pohlížet jako na průměrný násobek všech z-skórů pro proměnné x a y
  - $r = \Sigma (z_x * z_y) / n - 1$
- Při využití hodnot z tabulky: 7,35 / 9 = 0,816

x	y	Zx	Zy	Zx*Zy
2	0	-1,34164	-1,64317	2,205
2	2	-1,34164	-0,54772	0,735
3	1	-0,67082	-1,09545	0,735
3	3	-0,67082	0	0
4	2	0	-0,54772	0
4	4	0	0,54772	0
5	3	0,67082	0	0
5	5	0,67082	1,09545	0,735
6	4	1,34164	0,54772	0,735
6	6	1,34164	1,64317	2,205
			suma =	7,35

- Následující snímek nabízí alternativní výpočet se stejným výsledkem…

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

korelace mezi x a y, neboli R xy = cov(x,y) / s(x) * s(y),

(Databáze korelace a regrese.sav)

$$Cov\ (x,y) = \Sigma\ dx * dy / n - 1 = \Sigma\ (xi - \bar{x})*(yi - \bar{y}) / n - 1$$

$$\bar{X} = \Sigma\ x_i / n = 40 / 10 = 4$$

$$\bar{Y} = \Sigma\ y_i / n = 30 / 10 = 3$$

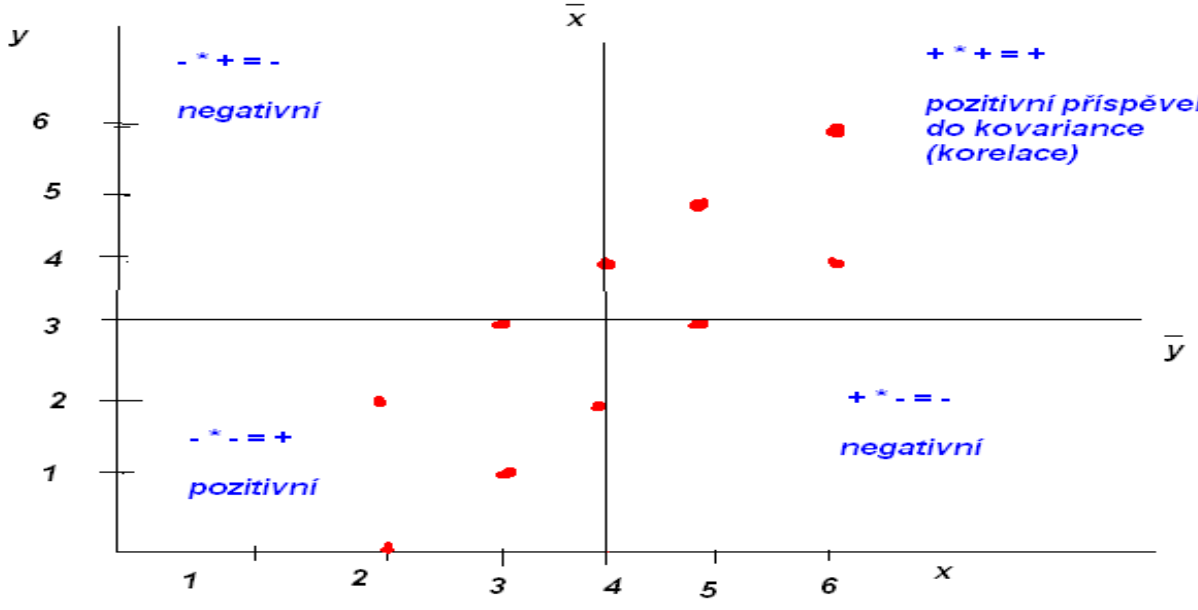
$$Cov\ (x,y) = \Sigma\ dx * dy / n - 1 = \Sigma\ (xi - \bar{x})*(yi - \bar{y}) / n - 1 = (-2*-3) + (-2*-1) + (-1*-2) + 0 + 0 + 0 + 0 + (1*2) + (2*1) + (2*3) = 20 / 9 = 2,22$$

$$s(x) * s(y) = \sqrt{}\ var\ (x) * \sqrt{}\ var\ (y)$$

$$var\ (x) = \Sigma\ (x_i - \bar{x})^2 / n - 1 = 20 / 9 = 2,22$$

$$var\ (y) = \Sigma\ (y_i - \bar{y})^2 / n - 1 = 30 / 9 = 3,33$$

$$R\ xy = cov(x,y) / s(x)\ s(y) = 2,22 / \sqrt{}\ 2,22 * \sqrt{}\ 3,33 = 2,22 / 2,72 = 0,816$$



Legenda

Xi = hodnota X pro jednotlivá individua

X = průměr pro x

d = absolutní odchylka

var(x)=rozptyl x

s(x)=směrodatná odchylka

Cov (x,y)=kovariance mezi x a y

R (x,y)= korelace mezi x a y

Korelace výpočet s příkladem

Předpoklady použití Pearsonova korelačního koeficientu

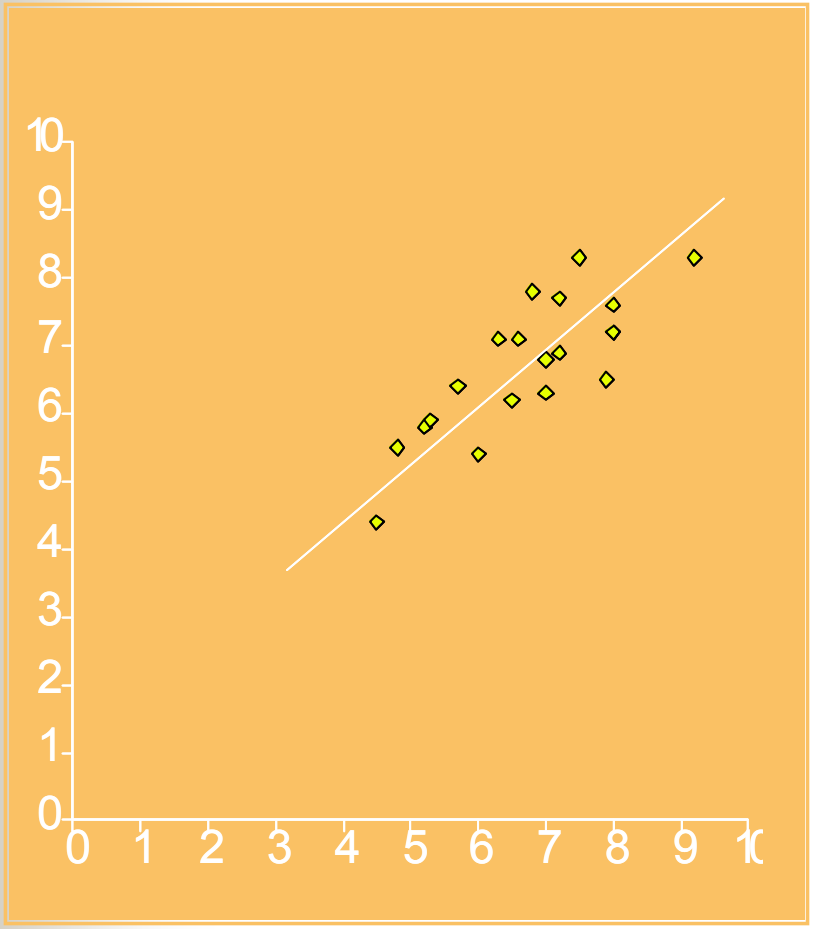
- 1) nejméně intervalová data
- 2) normální rozložení v populaci
- 3) neexistence extrémních případů
- 4) linearita vztahu

2, 3 a 4 třeba ověřit / otestovat

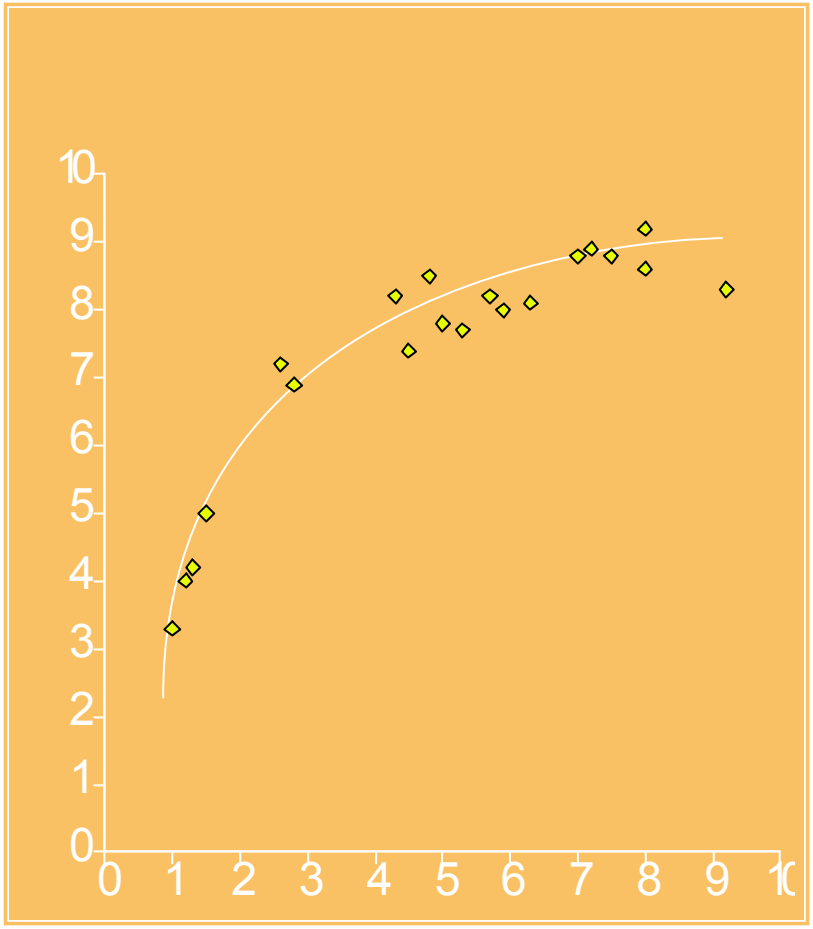
Není-li jeden z předpokladů naplněn a máme-li alespoň ordinální data, používáme Spearmanův koeficient

Předpoklad linearity vztahu

lineární

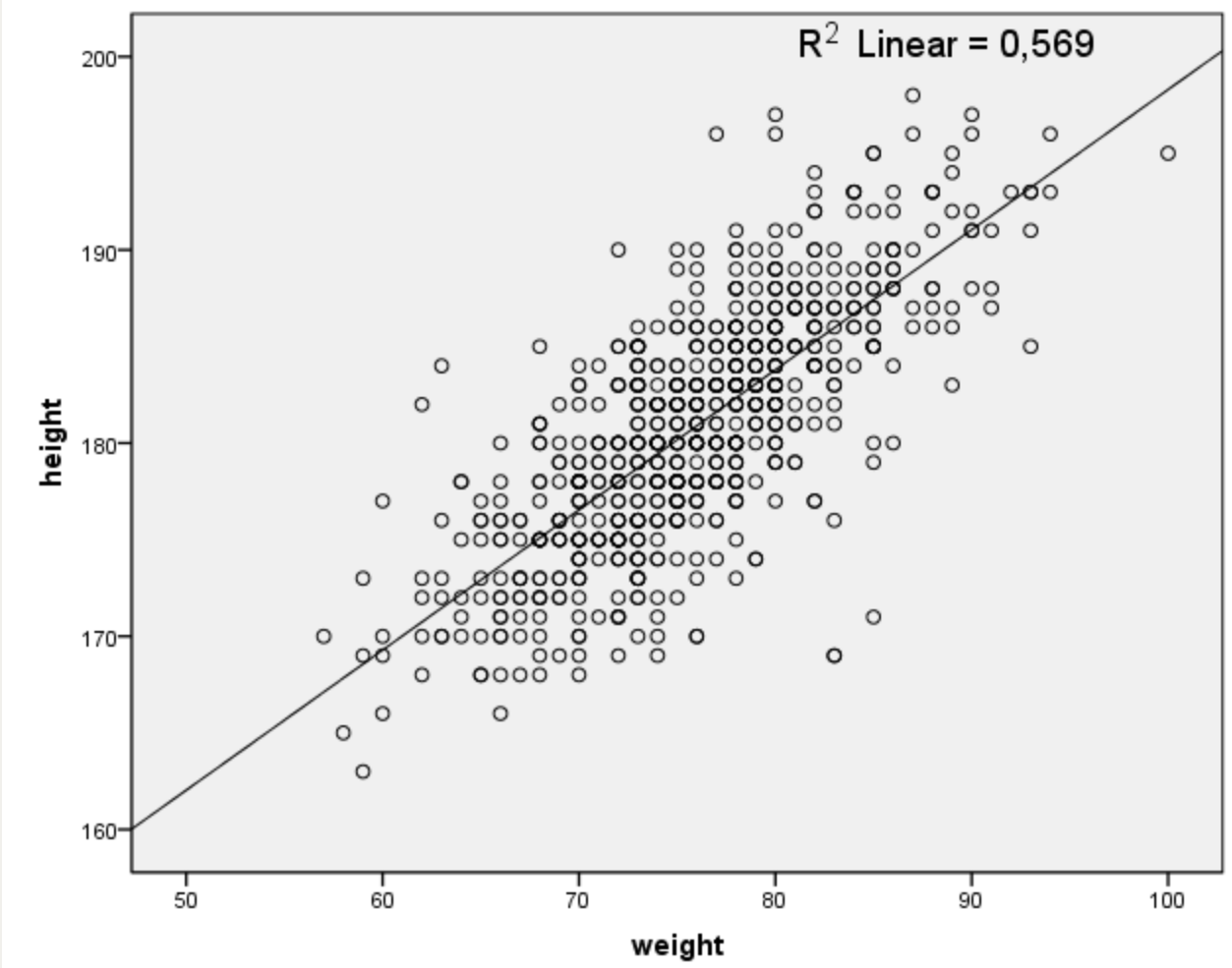


nelineární



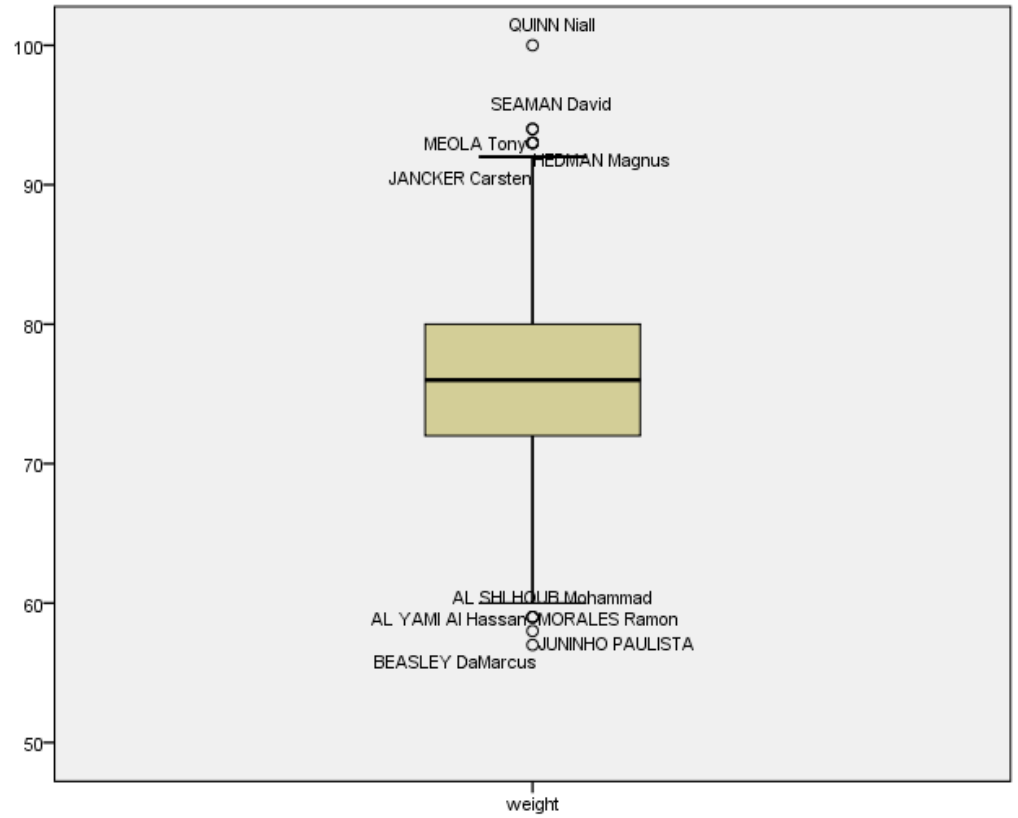
Ověřování předpokladu linearity vztahu

- Nejlépe pomocí bodového rozptýlení (scatterplot)



Ověřování předpokladu neexistence extrémních hodnot

Např. pomocí krabicového diagramu (boxplot) nebo jiného zobrazení extrémních hodnot…

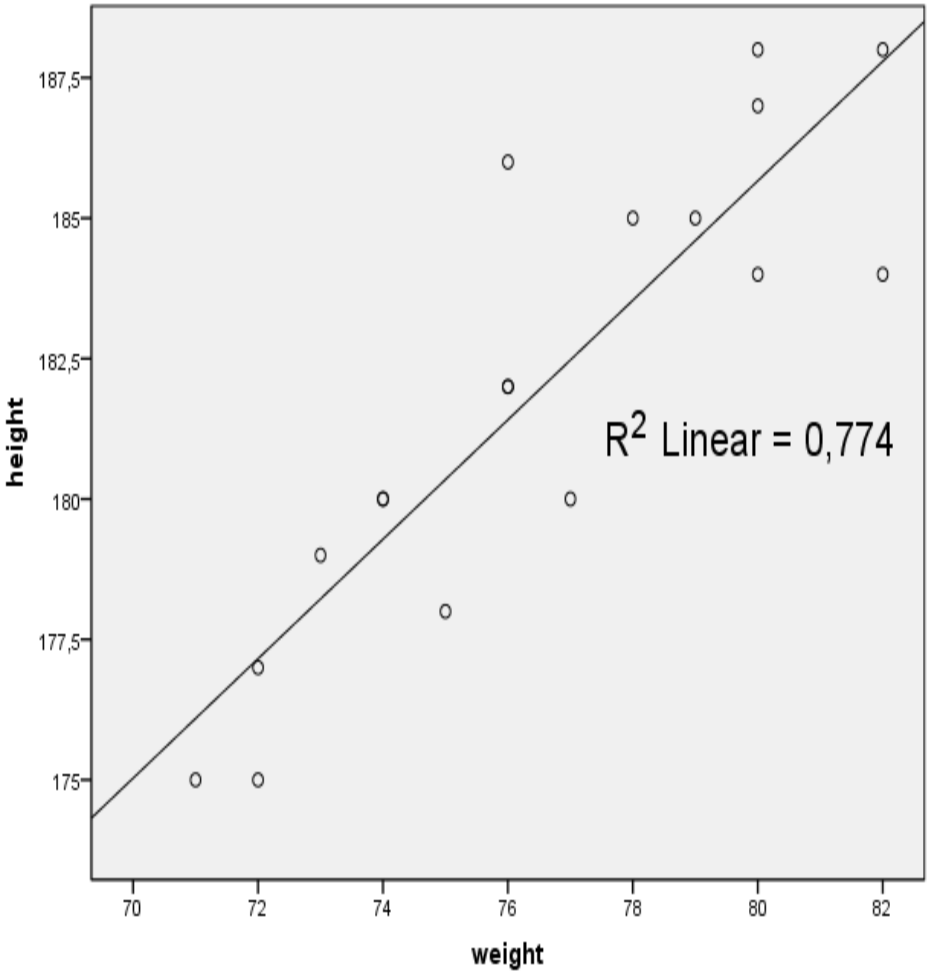


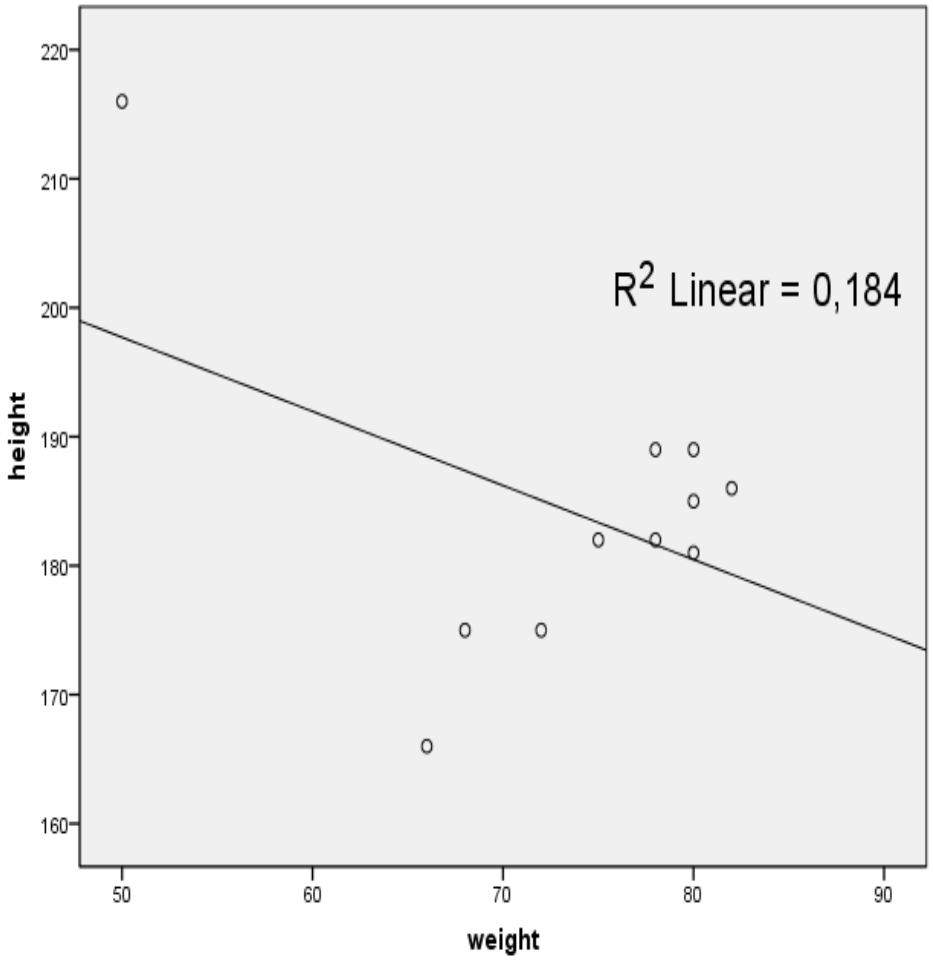
Extreme Values

			Case Number	name	Value
weight	Highest	1	316	QUINN Niall	100
		2	229	JAMES David	94
		3	610	DABANOVIC Mladen	94
		4	208	SEAMAN David	93
		5	285	JANCKER Carsten	93[a]
	Lowest	1	730	BEASLEY DaMarcus	57
		2	65	JUNINHO PAULISTA	58
		3	421	MORALES Ramon	59
		4	411	AL YAMI Al Hassan	59
		5	401	AL SHLHOUB Mohammad	59

a. Only a partial list of cases with the value 93 are shown in the table of upper extremes.

Jak nenaplnění předpokladu neexistence extrémních hodnot ovlivní Pearsonův r?

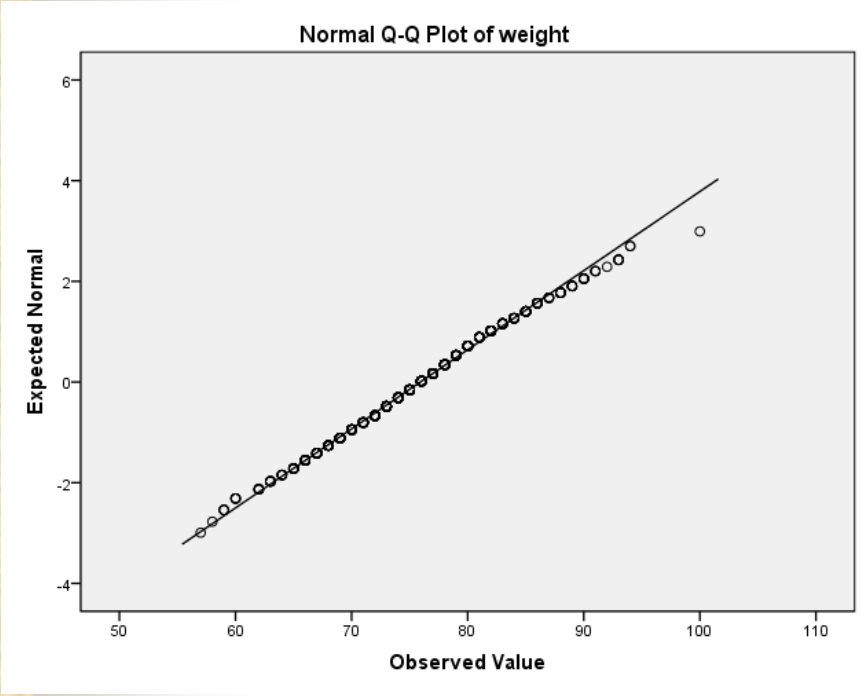




Ověřování předpokladu normality

a) Graficky – pozorované hodnoty ve vzorku vs. očekávané hodnoty pokud je populace normálně rozložená

b) Kolmogorov-Smirnov test normality rozložení



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
weight	,060	723	,000	,994	723	,009

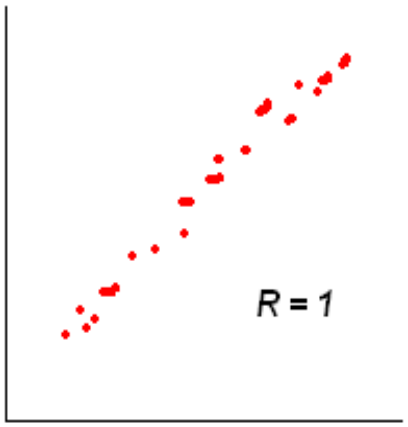
a. Lilliefors Significance Correction

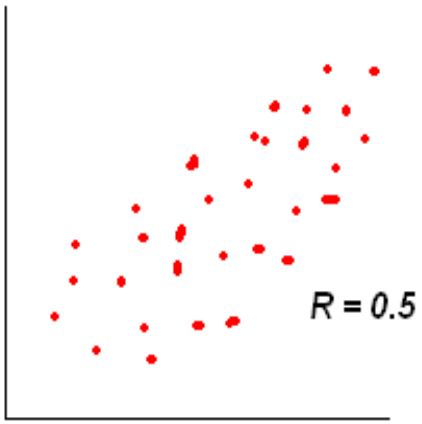
Hypothesis Test Summary

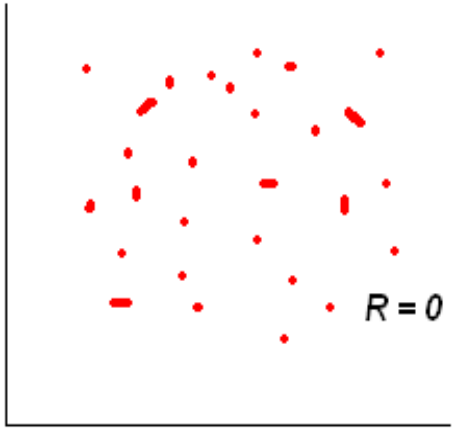
	Null Hypothesis	Test	Sig.	Decision
1	The distribution of weight is normal with mean 75.918 and standard deviation 6.364.	One-Sample Kolmogorov-Smirnov Test	.011	Reject the null hypothesis.

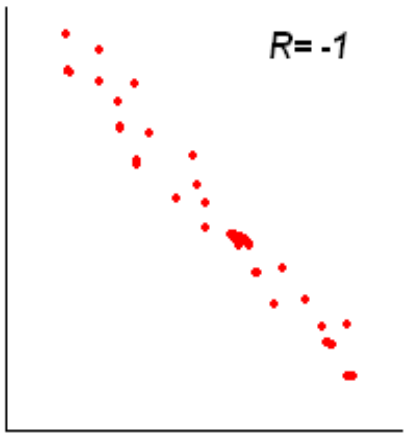
Asymptotic significances are displayed. The significance level is .05.

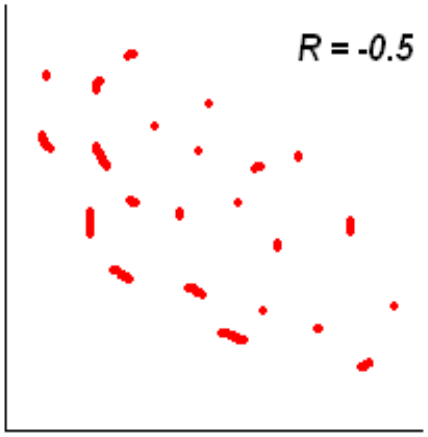
Korelační koeficient a bodové rozptýlení proměnných x a y

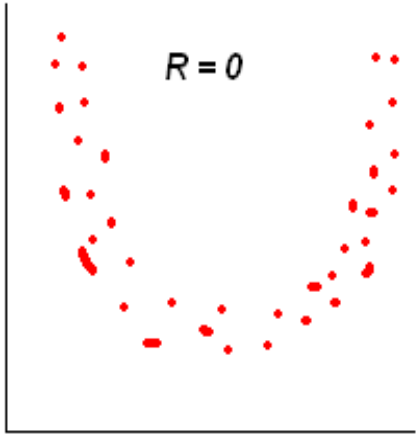




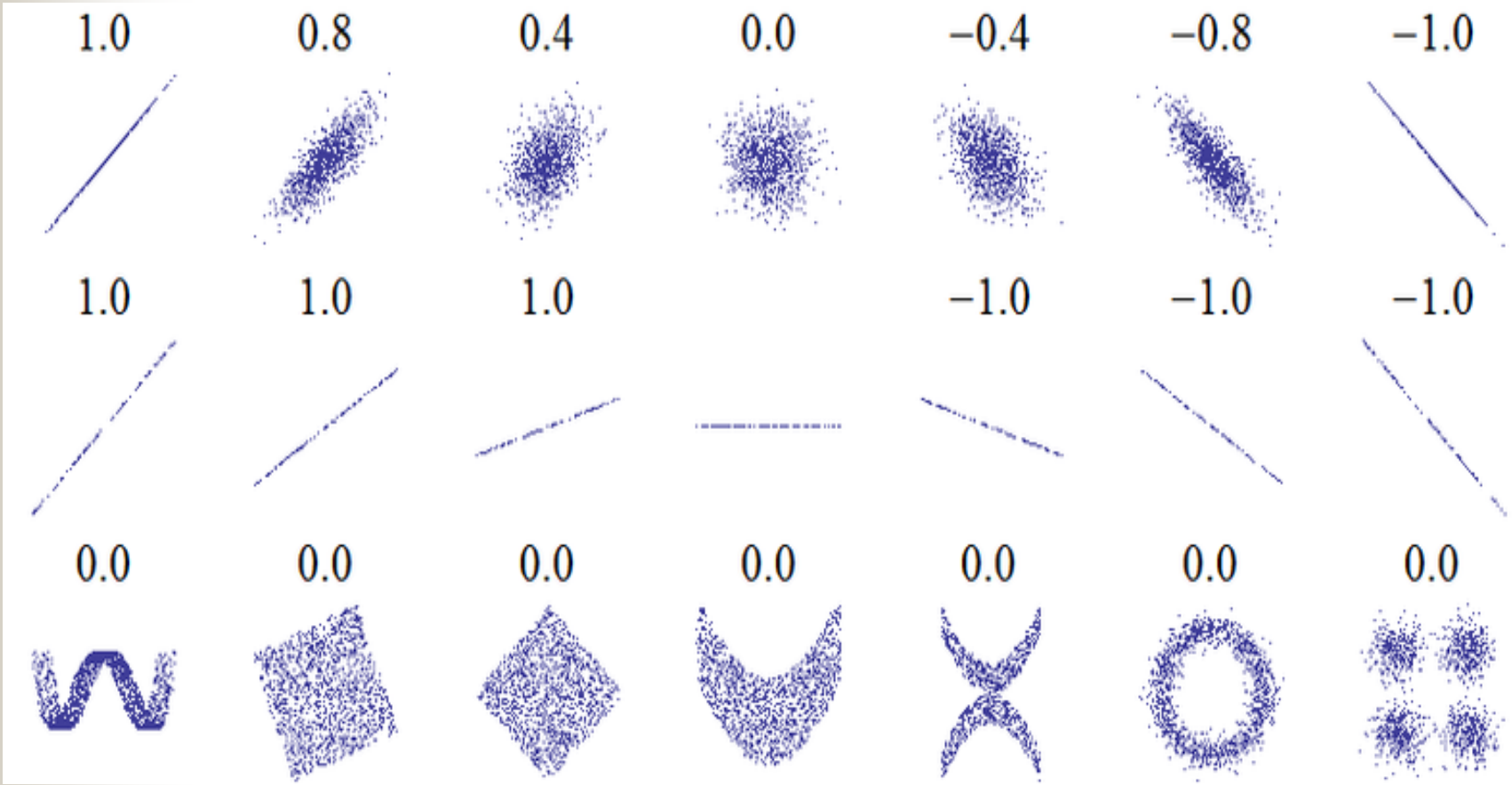








…další příklady



(zdroj: wikipedia)

Úvod do regresní analýzy

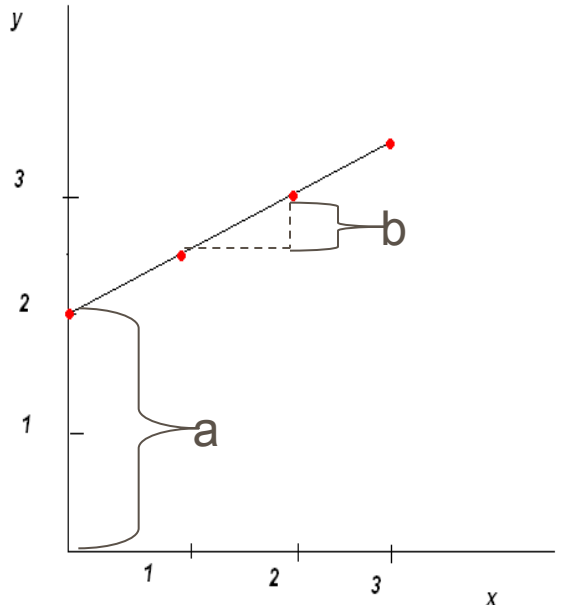
- Účel regresní analýzy: Predikce výsledku závislé proměnné na základě nezávisle proměnné
- Regresní přímka predikuje (odhaduje) hodnotu závislé proměnné (např. váha) jako lineární funkci hodnoty nezávisle proměnné (např. výška)
- Predikovanou hodnotu proměnné y označujeme y^.
- Rovnice regresní přímky má obecný tvar: y^ = a + bx,

  kdy a=úrovňová konstanta a b=sklon
- Úrovňová konstanta (a) je predikovaná

  (průměrná) hodnota y když x = 0
- Sklon (b) představuje množství

  o které se změní y ^

  pokud se x zvýší o jednotku



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

Výpočet regresních koeficientů

- Sklon b je roven součinu korelačního koeficientu xy s podílem směrodatných odchylek x a y
  - $b = r\ (Sy/Sx) = 0{,}816*(1{,}826/1{,}491)=0{,}816*1{,}225=1$
- Úrovňová konstanta a je rovna rozdílu mezi průměrnou hodnotou y a součinem sklonu s průměrnou hodnotou x
  - $a= y_{průměr} - b(x_{průměr}) = 3 - 1(4) = -1$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1,000	1,061		-,943	,373
	x	1,000	,250	,816	4,000	,004

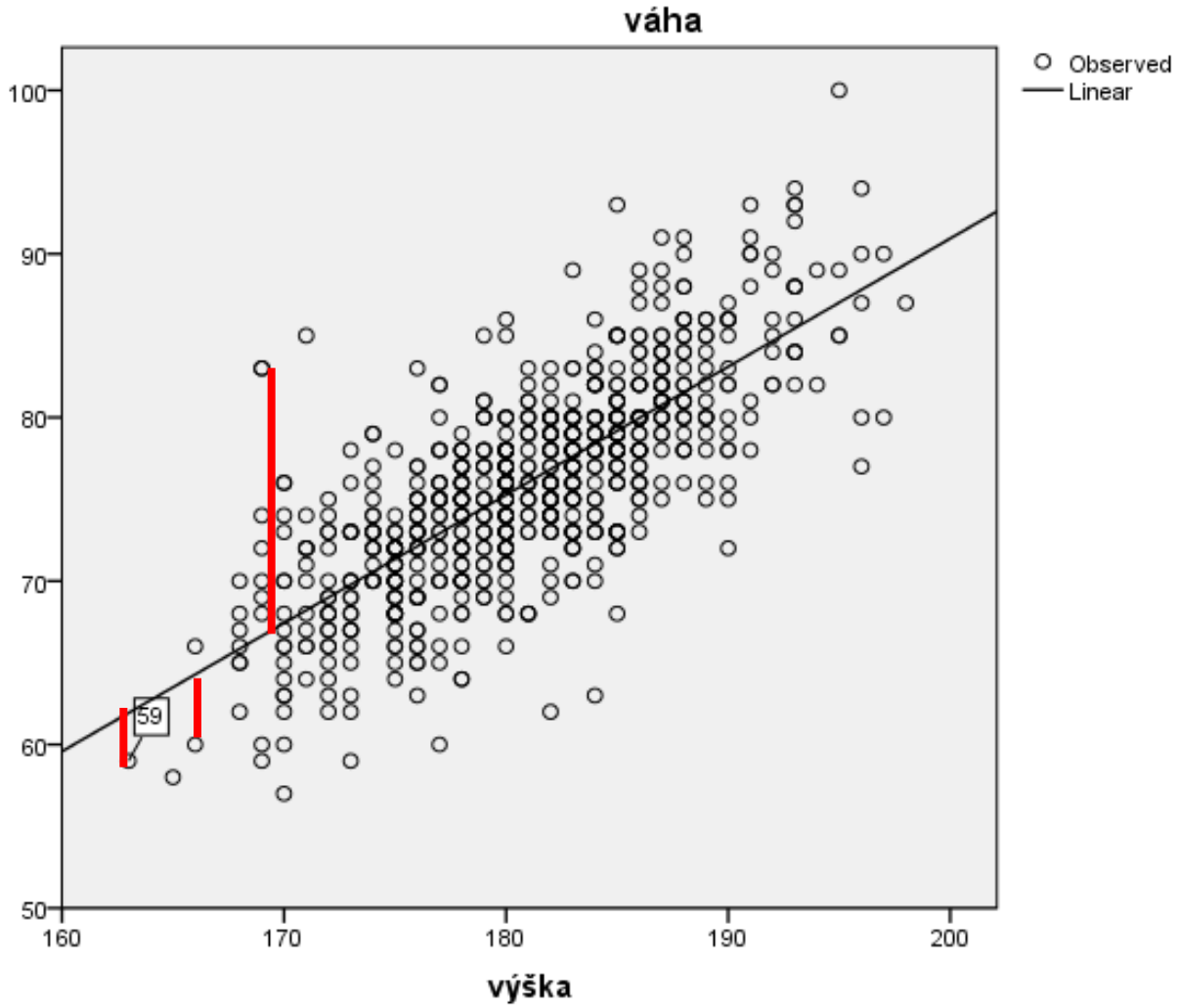
a. Dependent Variable: y

Interpretace regresních koeficientů: příklad fotbalistů (databáze fotbal_korelace.sav)

- Př. Odhadujeme váhu fotbalisty na základě jeho výšky
  - Dostali jsme regresní rovnici y^ = - 65.85 + 0,784x,
    - kde y^ =odhad váhy (v kg) a x=výška (v cm)
  - Interpretace úrovňové konstanty a=-65.85: Fotbalisté kteří měří 0 cm váží v průměru -65.85kg
  - Interpretace sklonu b=0,784: S každým centimetrem navíc roste váha fotbalisty o 0,784kg. Např. fotbalisté, kteří měří 163cm váží v průměru yi = -65.85 + 0,784 *(163) = 62kg

Reziduál a metoda nejmenších čtverců

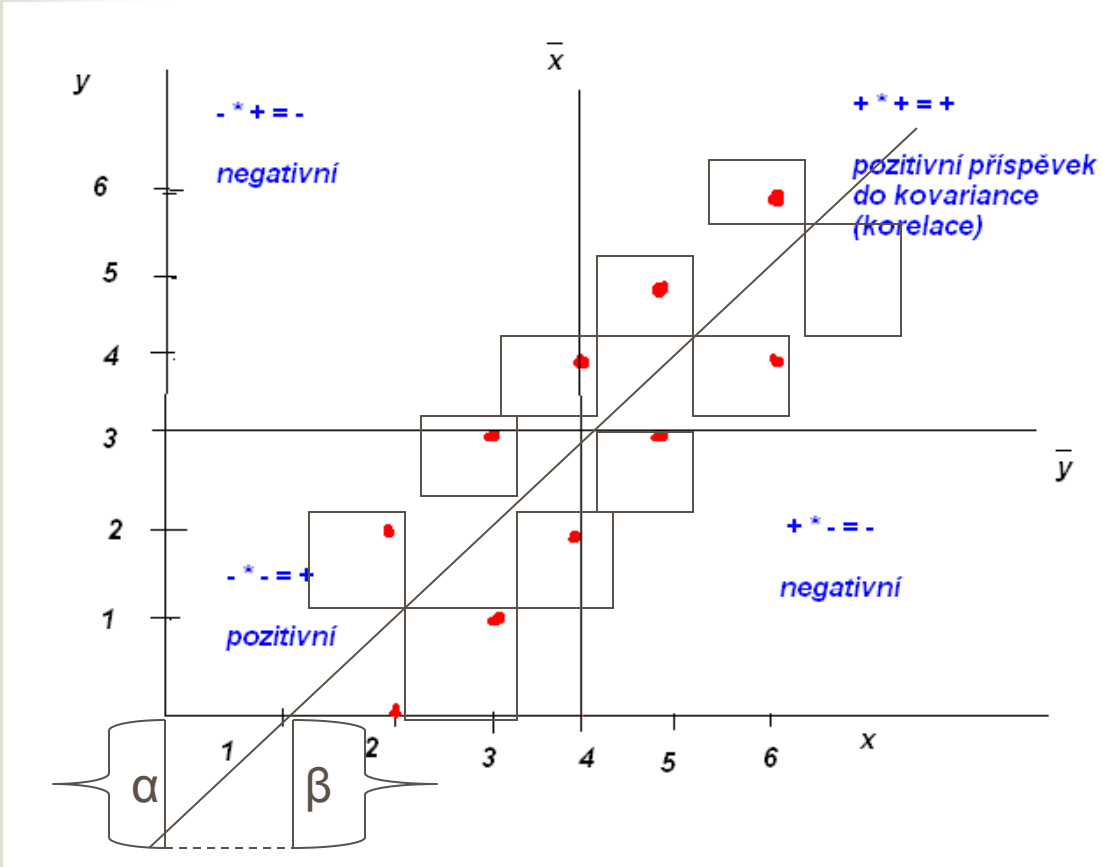
- V reálném světě kde „vše souvisí se vším“ a data tvoří velký počet případů nelze najít takové koeficienty, které by přesně vyhovovaly každému jednotlivému případu – predikce bude zatížena chybou (každý bod leží v různé vertikální vzdálenosti od přímky)
  - Např. regresní model predikuje člověku s výškou 163 cm váhu 62kg, ačkoli jeho skutečná váha je 59kg (viz další snímek), neboť váha souvisí kromě výšky i s jinými faktory, které jsme v regresní analýze ignorovali
- Absolutní hodnota reziduálu (e) každého člověka
  - představuje rozdíl mezi skutečnou hodnotou závislé proměnné (např. y = 59kg) a jejím odhadem (např. y^ = 62kg)
    - Ideální stav: $\hat{y} = a + bx$
    - skutečný stav: $y = a + bx + e$
    - $y - \hat{y} = e$
  - v bodovém rozptýlení se jedná o vertikální vzdálenost mezi bodem a přímkou u každého člověka (viz další snímek)
- Volí se takový tvar regresní rovnice, jehož použitím dosáhneme nejmenší celkové chyby tj. nejmenší sumy všech druhých mocnin reziduálů u všech lidí ve vzorku = metoda nejmenších čtverců


váha
Observed
Linear
100
90
80
70
60
50
160
170
180
190
200
výška
59

Regresní přímka a metoda nejmenších čtverců

(databáze korelace a regrese.sav)

Regresní přímka je položena tak, aby součet všech čtverců (=součet všech rozdílů mezi odhadovanými y a skutečnými y umocněných na druhou) byl nejmenší možný



Regresní rovnice: y = -1 + 1x + e

Koeficient determinace R^2

- Jako predikce hodnoty závislé proměnné y by mohl posloužit i průměrná hodnota y
- Pokud však mezi x a y je souvislost, pak nám regresní přímka umožňuje predikovat y přesněji než za použití pouhého průměru y
- Síla vztahu mezi x a y je dána tím, jak moc přesněji můžeme predikovat y použijeme-li regresní rovnici namísto pouhého průměru y
- Druhá mocnina korelačního koeficientu udává, o kolik menší je chyba predikce za použití regresní rovnice (y - y^) ve srovnání s použitím průměru (y – yprůměr).
- Př. $r^2 = 0,816*0,816=0,67$ udává, že průměrná chyba predikce y použitím regresní rovnice je o 67% menší, než v případě použití průměru
  - Nebo také jinak, že 67% rozptylu v proměnné y je vysvětleno lineárním vztahem mezi x a y (rozptyl predikovaných hodnot y z regresní rovnice představuje 67% rozptylu pozorovaných hodnot y)

Nepravá korelace

- Korelace neznamená kauzalitu (příčinný vztah mezi X a Y)
- Kdykoli pozorujeme vztah mezi X a Y, je možné že existuje třetí proměnná Z která je zodpovědná za tento vztah
- Př. znamená vysoká korelace mezi počtem čápů a počtem dětí, že čápi nosí děti?