Analýza dat pro Neurovědy

RNDr. Eva Koriťáková, Ph.D.

doc. RNDr. Ladislav Dušek, Dr.

Blok 7

Jak hodnotit vztah kvantitativních proměnných a základy regresního modelování.

Osnova

1. Základy korelační analýzy
2. Parciální korelace
3. Základy regresní analýzy

1. Základy korelační analýzy

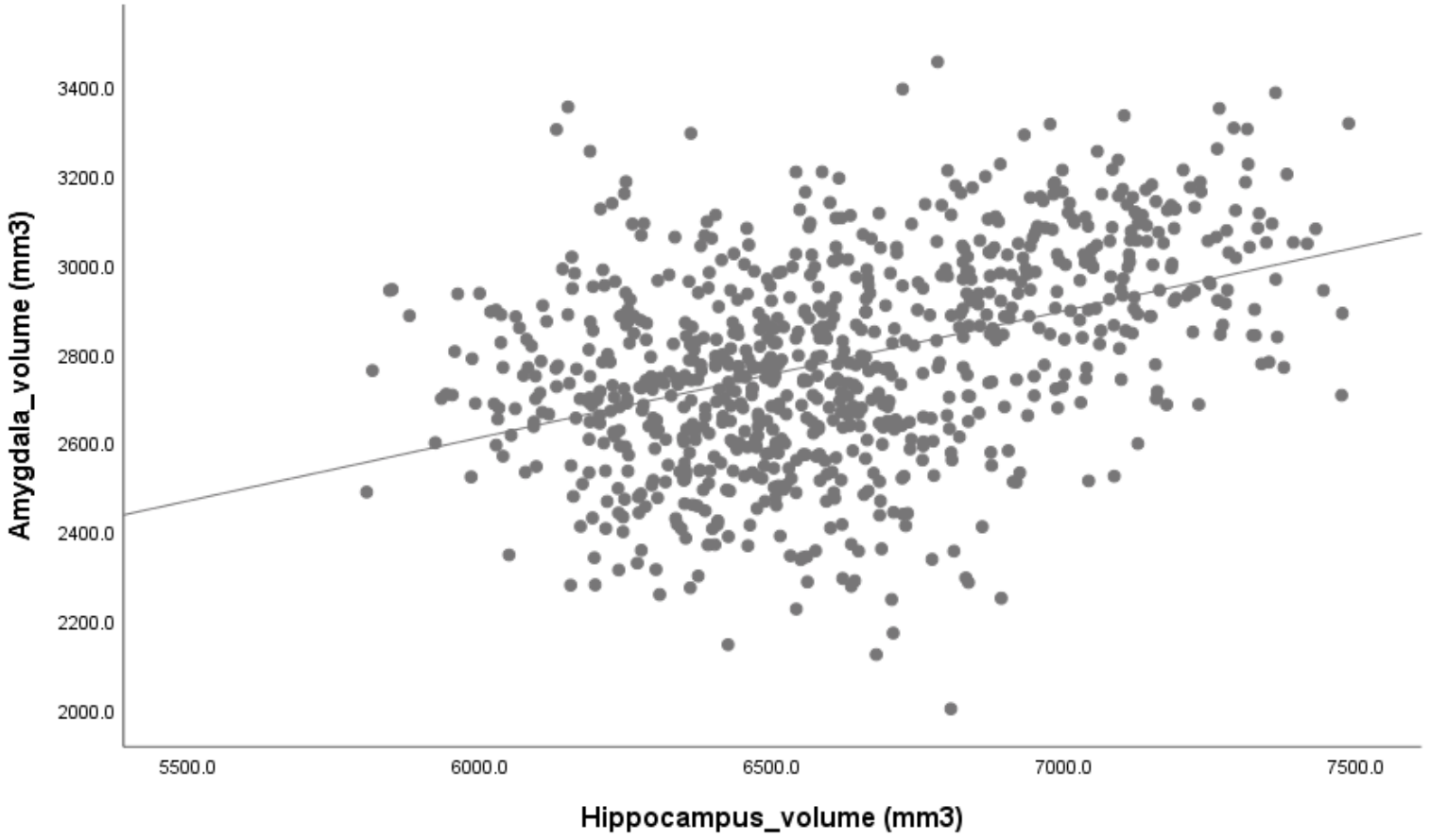
Motivace

- Zatím jsme se zabývali kvantitativní proměnnou v jedné skupině, kvantitativní proměnnou ve více skupinách, kategoriální proměnnou v jedné skupině, kategoriální proměnnou ve více skupinách, vztahem dvou kategoriálních proměnných.

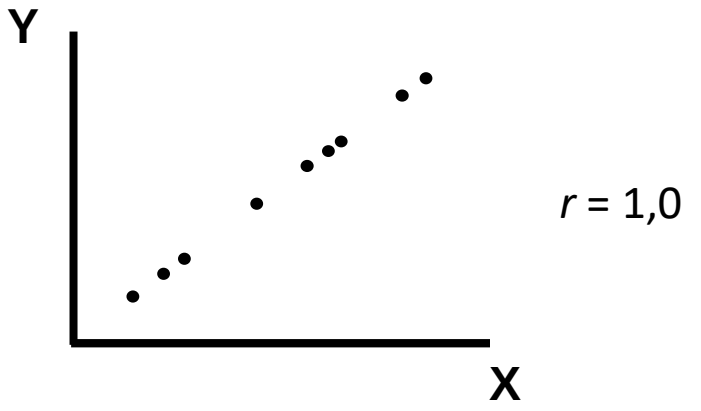
- Teď se chceme zabývat dvěma kvantitativními proměnnými:
  1. **Chceme zjistit, jestli mezi nimi existuje vztah** – např. jestli vyšší hodnoty jedné proměnné znamenají nižší hodnoty jiné proměnné.
  2. **Chceme jejich vztah kvantifikovat** – např. pro použití jedné proměnné na místo druhé proměnné.
  3. **Chceme predikovat hodnoty jedné proměnné na základě znalosti hodnot jiných proměnných.**

Jak hodnotit vztah dvou kvantitativních proměnných?

- Nejjednodušší formou je **bodový graf (x-y graf)**.
- Např. vztah objemu hipokampu a amygdaly:



Pearsonův korelační koeficient (*r*)



r = 1,0

r = -0,9

r = 0,4

r = 0,05

Korelace

- **Korelační koeficient** – kvantifikuje míru vztahu mezi dvěma kvantitativními proměnnými (X a Y).

- Standardní metodou je výpočet **Pearsonova korelačního koeficientu (r)**:
  - Charakterizuje **linearitu** vztahu mezi X a Y – jinak řečeno variabilitu kolem lineárního trendu.
  - Nabývá hodnot od -1 do 1.
  - Hodnota r je kladná (kladná korelace), když vyšší hodnoty X souvisí s vyššími hodnotami Y, a naopak je záporná (záporná korelace), když nižší hodnoty X souvisí s vyššími hodnotami Y.
  - Proměnné jsou nekorelované, pokud $r = 0$.
  - Hodnoty 1 nebo -1 získáme, když body x-y grafu leží na přímce.

- Lze statistickým testem **otestovat, zda jsou dvě kvantitativní proměnné nezávislé** – hypotézy mají tvar: $H_0: r = 0$ (tzn. korelační koeficient je roven nule) a $H_1: r \neq 0$.

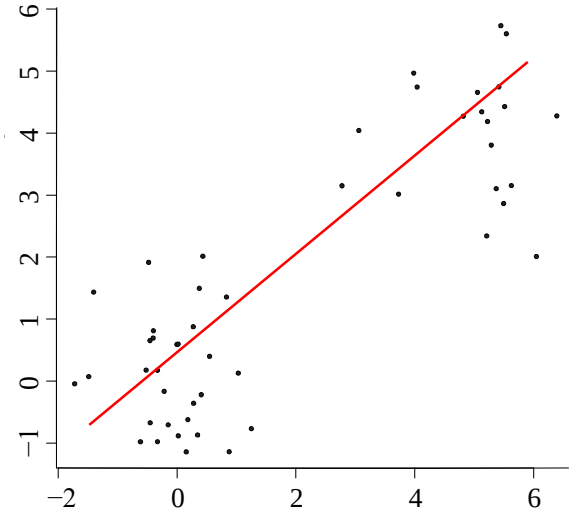
Síla vztahu podle hodnoty korelačního koeficientu

- Dle Evans, J. D. (1996). Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing je síla vztahu podle absolutní hodnoty korelačního koeficientu *r*:
  - 0,00-0,19: **velmi slabá korelace** („very weak correlation“)
  - 0,20-0,39: **slabá korelace** („weak correlation“)
  - 0,40-0,59: **střední korelace** („moderate correlation“)
  - 0,60-0,79: **silná korelace** („strong correlation“)
  - 0,80-1,00: **velmi silná korelace** („very strong correlation“)
- Tzn. např. korelace -0,23 by byla slabá záporná korelace (tzn. slabý vztah); korelace 0,84 by byla velmi silná kladná korelace (tzn. velmi silný vztah)
- V literatuře je však možno najít i jiné hranice (např. 0,1-0,3: slabá korelace, 0,3-0,5 střední korelace, >0,5 silná korelace [Cohen LH. Measurement of life events. In: Cohen LH, editor. Life Events and Psychological Functioning: Theoretical and Methodological Issues. Sage; Newbury Park, Calif.: 1988. pp. 11–30.], avšak hodnocení dle Evanse je zřejmě nejpoužívanější.

Pearsonův korelační koef. – problematické situace I.

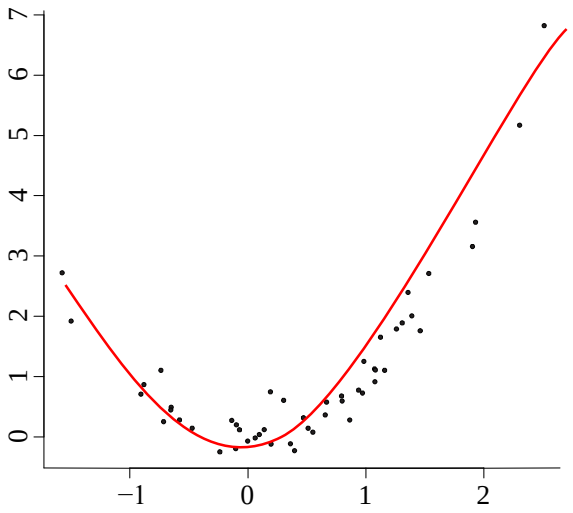
- Pearsonův korelační koeficient není vhodné počítat v situaci, kdy:
  - se v datech vyskytuje více skupin
  - proměnné mají nelineární vztah
  - se v datech vyskytují odlehlé hodnoty

Více skupin



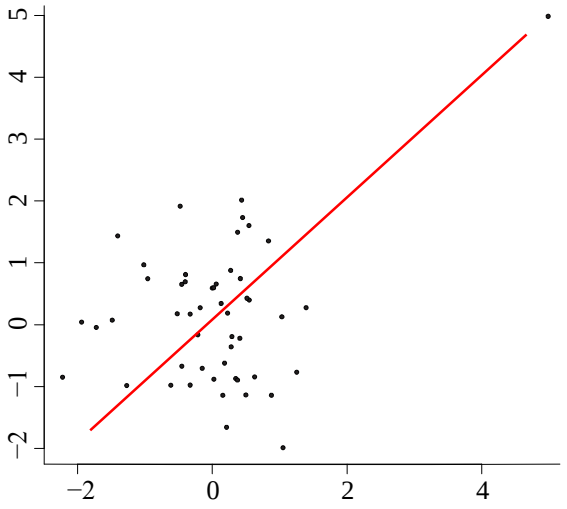
$r = 0{,}84$
($p < 0{,}001$)

Nelineární vztah



$r = 0{,}58$
($p < 0{,}001$)

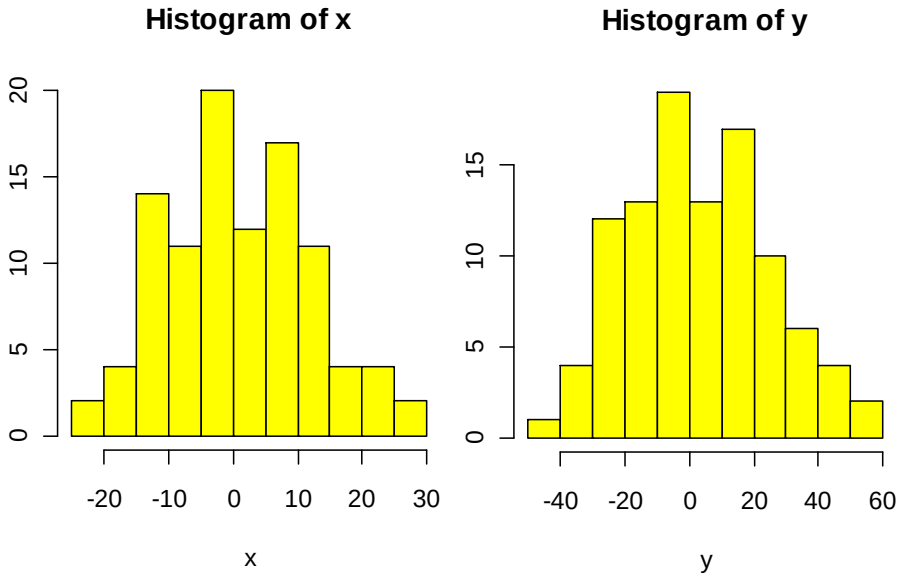
Odlehlá hodnota

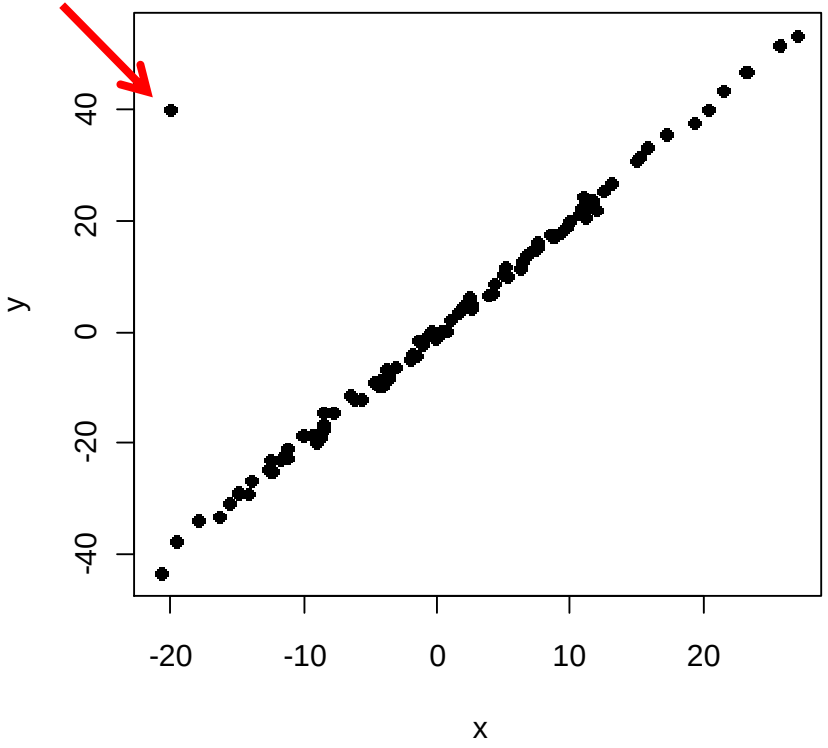


$r = 0{,}36$
($p = 0{,}009$)

Pearsonův korelační koef. – problematické situace II.

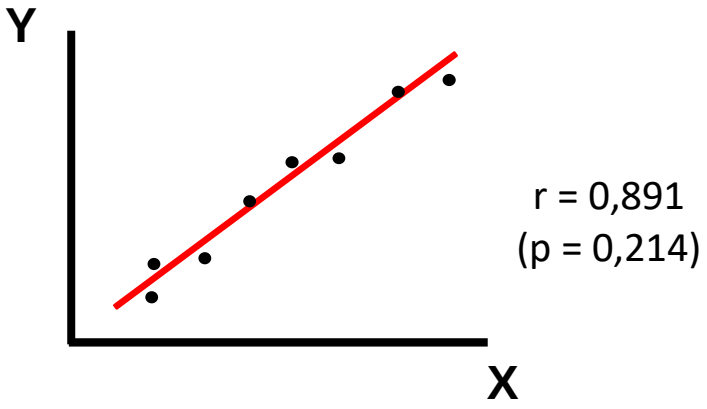
- Při srovnání dvou kvantitativních proměnných je nutné vykreslovat bodový graf, protože histogramy pro jednotlivé proměnné zvlášť nám nemusejí odhalit odlehlé hodnoty!



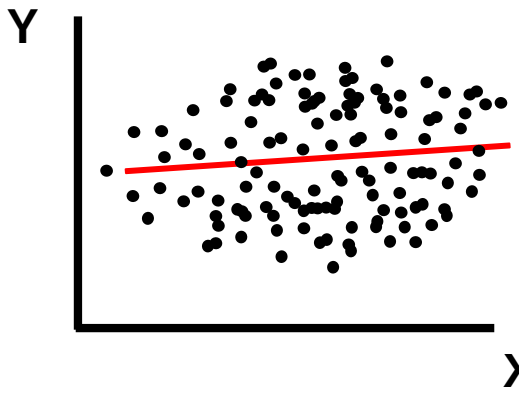


Pearsonův korelační koef. – problematické situace III.

- Problém velikosti vzorku:



r = 0,891
(p = 0,214)



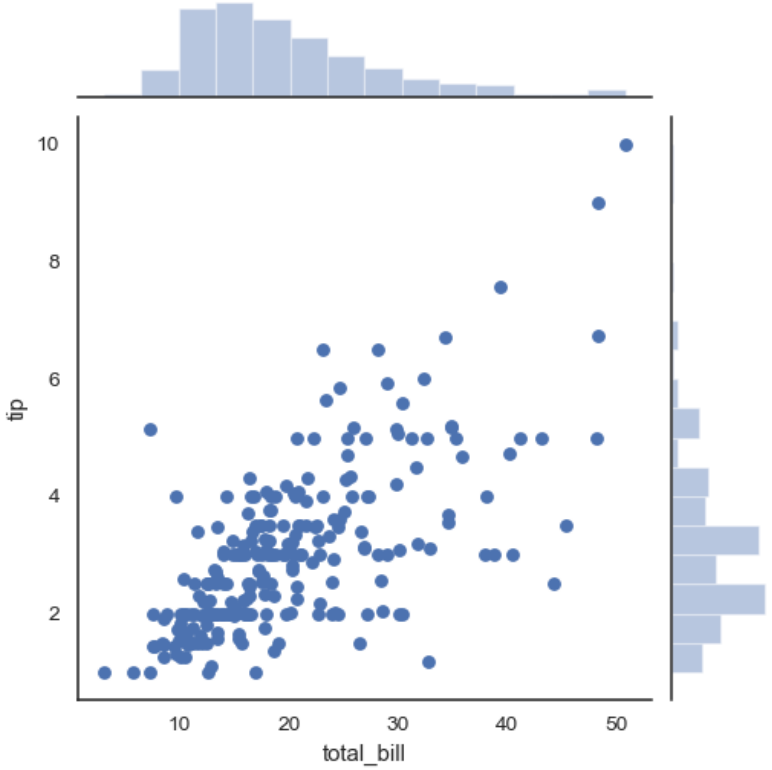
r = 0,012
(p = 0,008)

- Test na ověření, zda je Pearsonův korelační koeficient různý od nuly, je parametrický test – předpoklad normality srovnávaných kvantitativních proměnných!

Pearsonův korelační koef. – předpoklad normality

- předpoklad normality proměnných je tedy nutný proto, abychom mohli věřit vypočítané p-hodnotě

- pokud nás zajímá pouze hodnota korelačního koeficientu, můžeme počítat Pearsonův korelační koeficient i v situaci, kdy není splněn předpoklad normality (tzn. rozdělení proměnných je zešikmené)



Zdroj: https://stackoverflow.com/questions/55111214/change-position-of-marginal-axis-in-seaborn-jointplot

- porušení normality však nesmí být z důvodu, že jsou v datech odlehlé hodnoty, jinak by byla zkreslená nejen p-hodnota, ale i korelační koeficient!

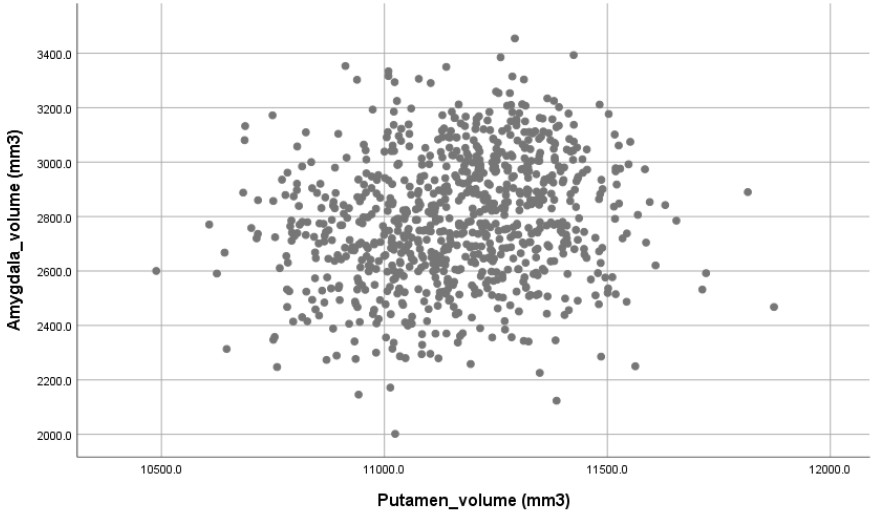
Pearsonův korelační koeficient – příklad

- **Příklad:** Ověřte, zda existuje vztah objemu putamenu a amygdaly v souboru 833 subjektů.
- **Řešení:**

Correlations

		Amygdala_volume (mm3)	Putamen_volume (mm3)
Amygdala_volume (mm3)	Pearson Correlation	1	,174**
	Sig. (2-tailed)		,000
	N	833	833
Putamen_volume (mm3)	Pearson Correlation	,174**	1
	Sig. (2-tailed)	,000	
	N	833	833

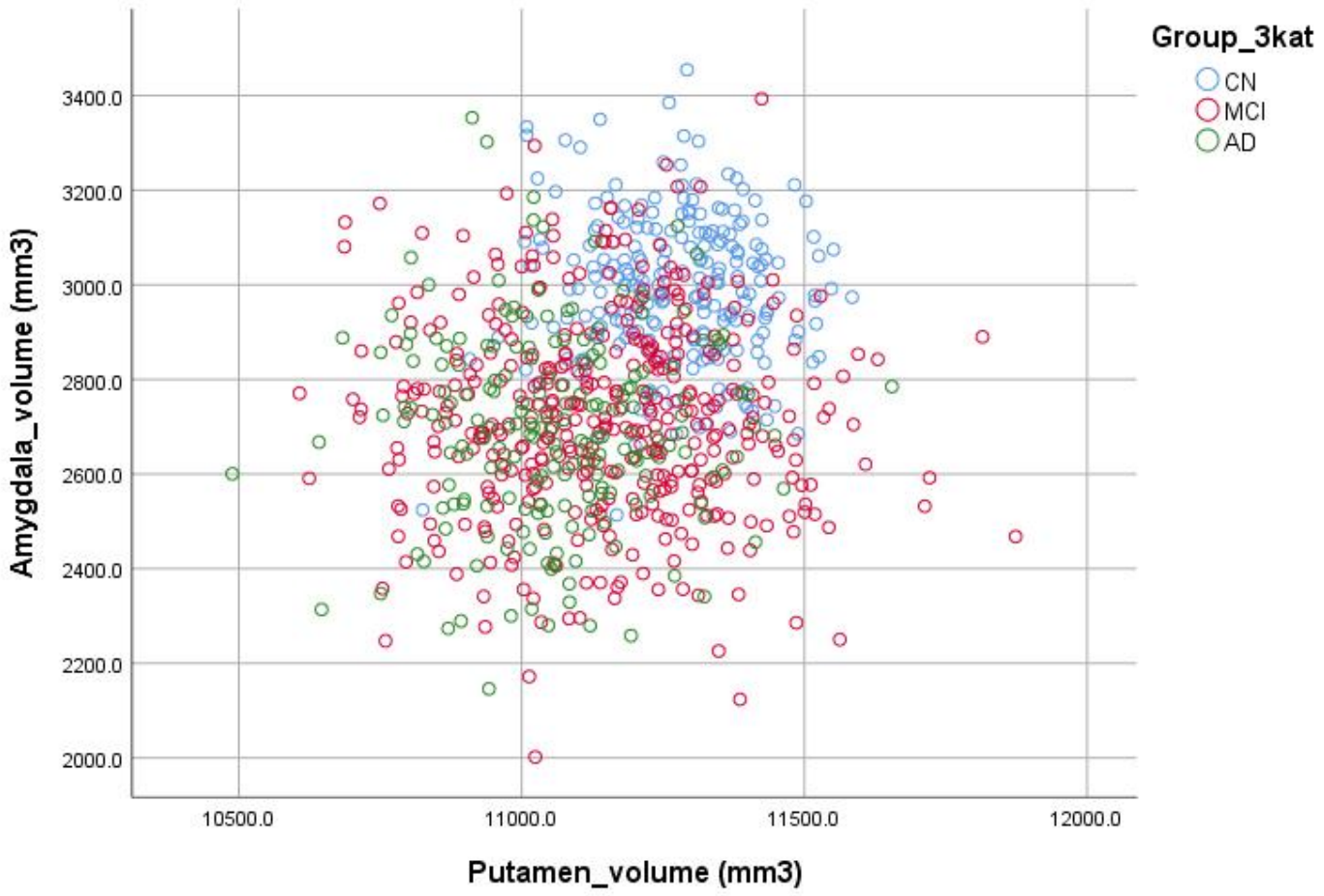
**. Correlation is significant at the 0.01 level (2-tailed).



Velmi slabá korelace (r=0,17), avšak statisticky významná (je to artefakt velkého N).

Je to však vyhodnoceno správně?

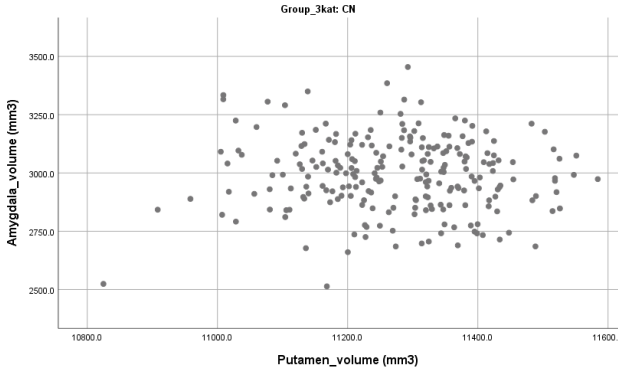
Pearsonův korelační koeficient – příklad – pokračování



Počítat korelační koeficient celkově pro všech 833 subjektů je zavádějící!
Kontrolní subjekty nejsou „promíchaní“ s ostatními subjekty, budou tudíž celkový korelační koeficient zkreslovat!

Pearsonův korelační koeficient – příklad – pokračování

Musíme tedy hodnotit vztah objemu putamenu a amygdaly u každé skupiny zvlášť!



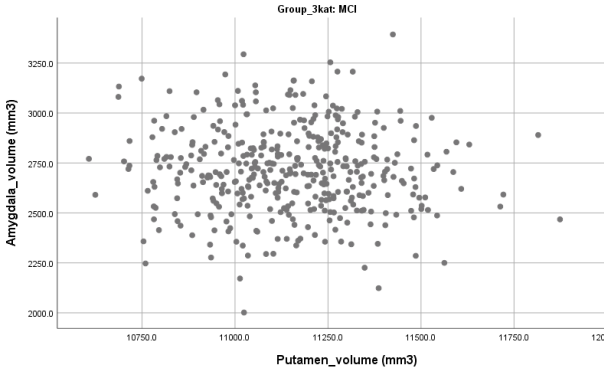
Group_3kat = CN

Correlations[a]

		Putamen_volume (mm3)	Amygdala_volume (mm3)
Putamen_volume (mm3)	Pearson Correlation	1	-.038
	Sig. (2-tailed)		.571
	N	230	230
Amygdala_volume (mm3)	Pearson Correlation	-.038	1
	Sig. (2-tailed)	.571	
	N	230	230

a. Group_3kat = CN

r = -0,038; p = 0,571



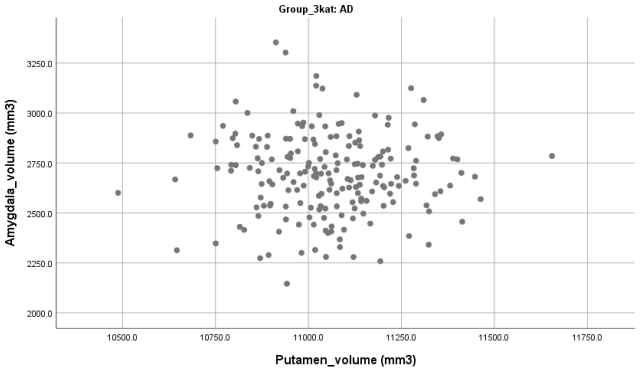
Group_3kat = MCI

Correlations[a]

		Putamen_volume (mm3)	Amygdala_volume (mm3)
Putamen_volume (mm3)	Pearson Correlation	1	-.044
	Sig. (2-tailed)		.372
	N	406	406
Amygdala_volume (mm3)	Pearson Correlation	-.044	1
	Sig. (2-tailed)	.372	
	N	406	406

a. Group_3kat = MCI

r = -0,044; p = 0,372



Group_3kat = AD

Correlations[a]

		Putamen_volume (mm3)	Amygdala_volume (mm3)
Putamen_volume (mm3)	Pearson Correlation	1	.005
	Sig. (2-tailed)		.946
	N	197	197
Amygdala_volume (mm3)	Pearson Correlation	.005	1
	Sig. (2-tailed)	.946	
	N	197	197

a. Group_3kat = AD

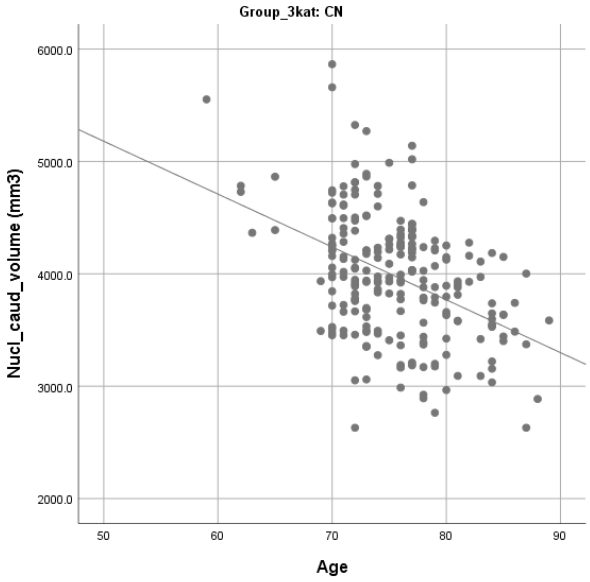
r = 0,005; p = 0,946

Závěr: Korelace je téměř nulová u všech podskupin a statisticky nevýznamná – není vztah mezi objemem putamenu a amygdaly. Dokonce u kontrol a pac. s MCI je záporná. Velmi slabý vztah při hodnocení všech subjektů současně (r=0,17) byl tedy opravdu pouze artefakt vzniklý přítomností více skupin!

Úkol 1.

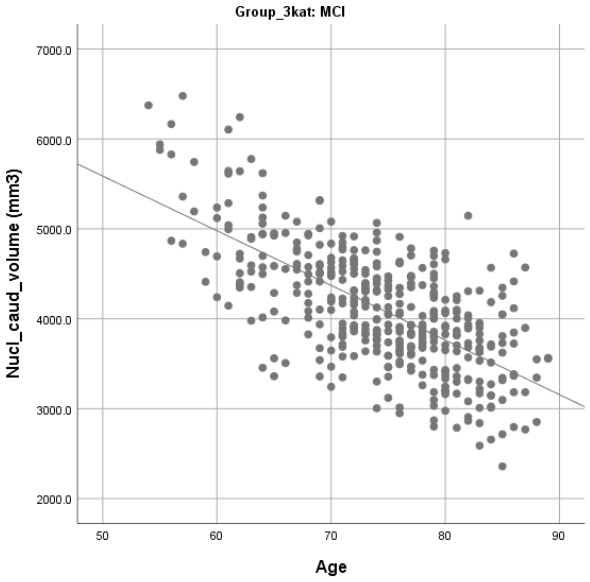
- **Zadání:** Ověřte, zda existuje vztah objemu nucleus caudatus a věku u pacientů s AD, pacientů s MCI a u kontrol.
- **Řešení:**

CN



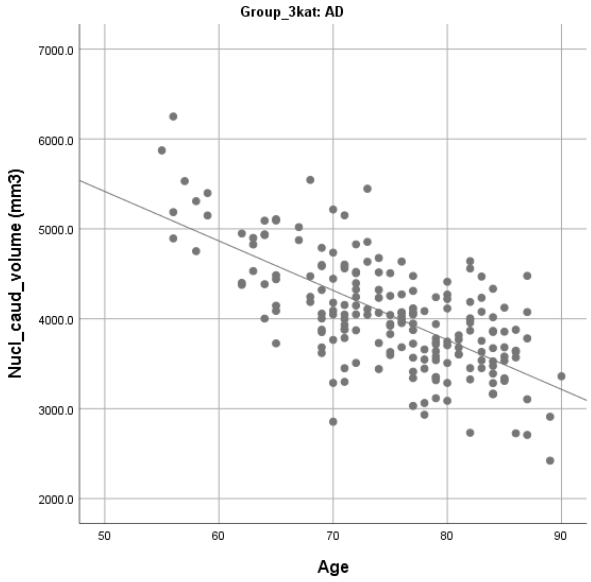
r = -0,43
(p < 0,001)

MCI



r = -0,67
(p < 0,001)

AD



r = -0,68
(p < 0,001)

Závěr: S rostoucím věkem statisticky významně klesá objem nucleus caudatus u všech skupin subjektů. U pacientů s MCI a AD je korelace (vztah) mezi objemem nucleus caudatus a věkem silná, u kontrol střední.

Srovnání dvou korelačních koeficientů

- **Příklad:** Srovnejte korelační koeficienty objemu nucleus caudatus a věku u pacientů s AD a kontrolních subjektů.
- Použijte webový kalkulátor: http://vassarstats.net/rdiff.html

- **Postup:**

Z předchozího úkolu víme, že:
$r_1 = -0{,}68$
$N_1 = 197$
$r_2 = -0{,}43$
$N_2 = 230$

Sample A		Sample B		
r_a =	-0.68	r_b =	-0.43	Reset
n_a =	197	n_b =	230	Calculate
z =	-3.78			
P	one-tailed	0.0001		
	two-tailed	0.0002		

Závěr: Korelační koeficienty se u pacientů s AD a kontrol statisticky významně liší.

Srovnání korelačního koeficientu s referenční hodnotou

- **Příklad:** Srovnejte korelační koeficient objemu nucleus caudatus a věku u pacientů s MCI s hodnotou -0,62, jež byla zjištěna při populačním průzkumu.
- Použijte webový kalkulátor: http://vassarstats.net/rpop.html

- **Postup:**

Z předchozího úkolu víme, že:
$r_1 = -0,67$
$N_1 = 406$

Populační průzkum:
$r_2 = -0,62$

Observed for Sample		Hypothetical for Population		
r =	-0.67	rho =	-0.62	Reset
n =	406			Calculate
z =	-1.72			
P	one-tailed	0.042716		
	two-tailed	0.085432		

Závěr: Korelační koeficient se statisticky významně neliší od hodnoty -0,62 zjištěné při populačním průzkumu.

Poznámka

- Korelace dvou náhodných veličin se často interpretuje pomocí druhé mocniny Pearsonova korelačního koeficientu: r^2.

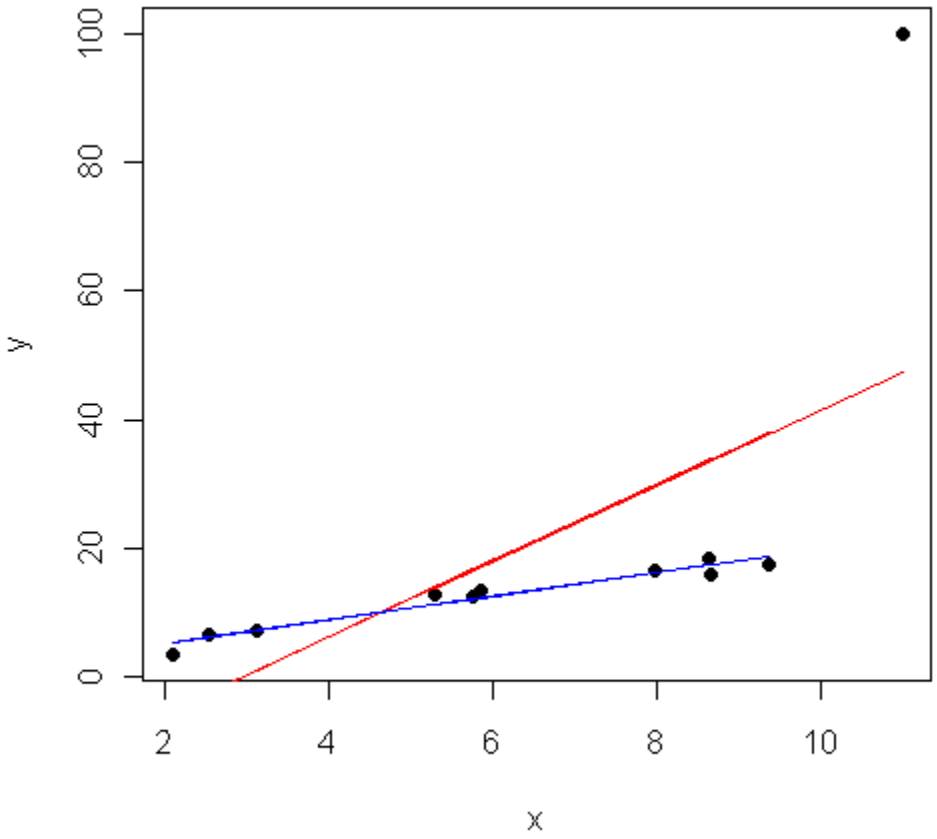
- Hodnota r^2 vyjadřuje, kolik % své variability sdílí jedna veličina s druhou, jinak řečeno, kolik % variability jedné veličiny může být predikováno pomocí té druhé.

- S hodnotou r^2 se setkáte v lineárních modelech.

Spearmanův korelační koeficient (r_s)

- Pearsonův korelační koeficient je náchylný k odlehlým hodnotám.
- **Spearmanův korelační koeficient** stejně jako řada dalších neparametrických metod **pracuje pouze s pořadími** pozorovaných hodnot.
- Hodnoty Spearmanova korelačního koeficientu r_s se pohybují stejně jako u Pearsonova korelačního koeficientu r od -1 do 1.
- Pozor! Spearmanův korelační koeficient je stejně jako Pearsonův korelační koeficient nevhodné použít při nelineárním vztahu proměnných nebo při více skupinách!

Srovnání Pearsonova a Spearmanova korelačního koeficientu



Pearsonův korelační koeficient:

$r = 0{,}65$

$(p = 0{,}029)$

Spearmanův korelační koeficient:

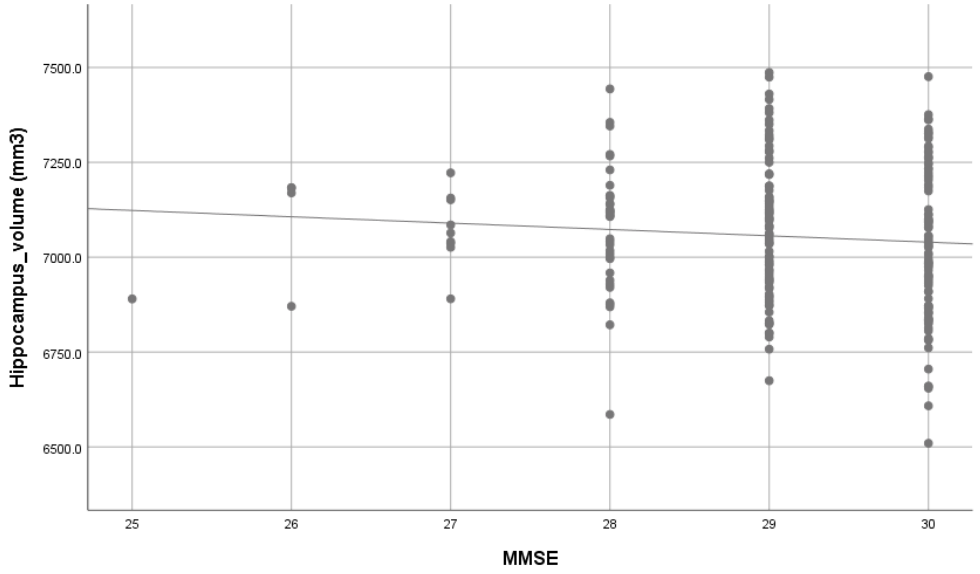
$r_S = 0{,}95$

$(p < 0{,}001)$

Spearmanův korelační koeficient není náchylný k odlehlým hodnotám.

Spearmanův korelační koeficient

- **Příklad:** Zjistěte, zda existuje vztah objemu hipokampu a MMSE skóre u kontrol.
- **Řešení:**



Correlations

			MMSE	Hippocampus_volume (mm3)
Spearman's rho	MMSE	Correlation Coefficient	1.000	-.121
		Sig. (2-tailed)	.	.066
		N	230	230
	Hippocampus_volume (mm3)	Correlation Coefficient	-.121	1.000
		Sig. (2-tailed)	.066	.
		N	230	230

Závěr: Vztah je velmi slabý a statisticky nevýznamný.

Úkol 2.

- **Zadání:** Zjistěte, zda existuje vztah objemu všech dalších pěti mozkových sktruktur s MMSE skóre u kontrol (nezapomeňte vykreslit bodové grafy).
- **Řešení:**

Correlations

		MMSE	Amygdala_volume (mm3)	Thalamus_volume (mm3)	Pallidum_volume (mm3)	Putamen_volume (mm3)	Nucl_caud_volume (mm3)
MMSE	Pearson Correlation	1	.049	-.048	-.152*	.031	.039
	Sig. (2-tailed)		.461	.471	.021	.636	.555
	N	230	230	230	230	230	230
Amygdala_volume (mm3)	Pearson Correlation	.049	1	-.035	-.042	-.038	-.009
	Sig. (2-tailed)	.461		.600	.525	.571	.894
	N	230	230	230	230	230	230
Thalamus_volume (mm3)	Pearson Correlation	-.048	-.035	1	.080	-.103	.124
	Sig. (2-tailed)	.471	.600		.229	.119	.060
	N	230	230	230	230	230	230
Pallidum_volume (mm3)	Pearson Correlation	-.152*	-.042	.080	1	.073	-.077
	Sig. (2-tailed)	.021	.525	.229		.268	.247
	N	230	230	230	230	230	230
Putamen_volume (mm3)	Pearson Correlation	.031	-.038	-.103	.073	1	-.111
	Sig. (2-tailed)	.636	.571	.119	.268		.094
	N	230	230	230	230	230	230
Nucl_caud_volume (mm3)	Pearson Correlation	.039	-.009	.124	-.077	-.111	1
	Sig. (2-tailed)	.555	.894	.060	.247	.094	
	N	230	230	230	230	230	230

*. Correlation is significant at the 0.05 level (2-tailed).

Statisticky významný vztah je pouze mezi MMSE a objemem pallida, nicméně i tento vztah je velmi slabý (a tudíž klinicky nevýznamný).

2. Parciální korelace

Motivace

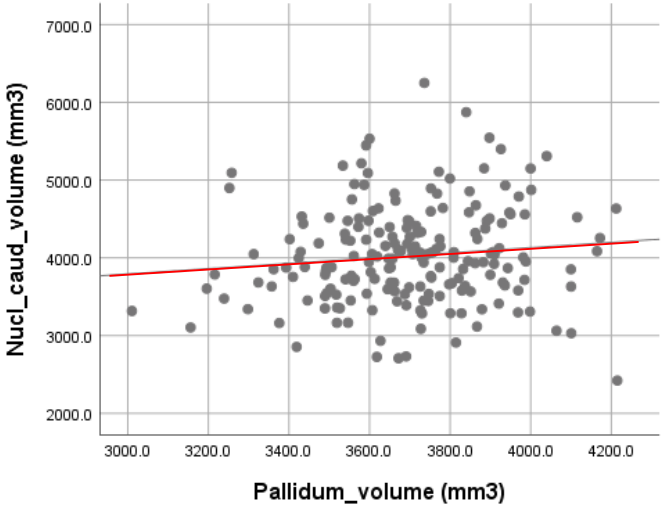
- Chceme hodnotit vztah dvou kvantitativních proměnných. Tento vztah ale může být ovlivněn jinou proměnnou či proměnnými – Pearsonův či Spearmanův korelační koeficient bude zavádějící.
- Cílem parciální korelace je **odstranění vlivu tzv. zavádějící proměnné (či proměnných) při vyhodnocení vztahu 2 kvantitativních proměnných**.

- Příklad: Chceme hodnotit vztah objemu hipokampu a objemu amygdaly u pacientů s Alzheimerovou chorobou. Z literatury však víme, že na objem hipokampu i objem amygdaly má vliv délka vzdělání. Použijeme tedy parciální korelaci na výpočet vztahu objemů obou struktur, v rámci něhož vliv délky vzdělání odstraníme.

- **Zavádějící proměnné (tzv. kovariáty)**, které by mohly zkreslit analýzu, **vyčteme z literatury či zjistíme z dat** (jedná se o charakteristiky subjektů, které mají statisticky významný vztah s hodnocenými proměnnými – viz. další slide).

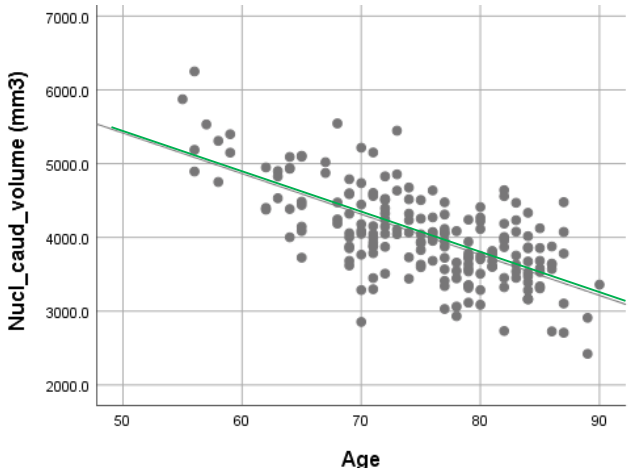
Parciální korelace – příklad

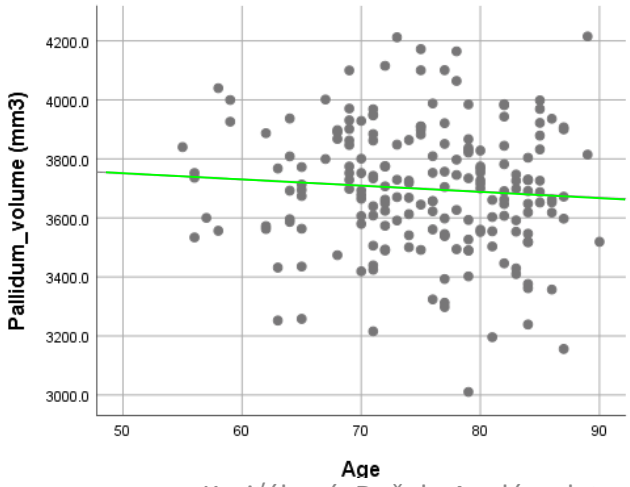
- **Příklad:** Hodnocení vztahu objemu pallida a objemu nukleus caudatus u pacientů s Alzheimerovou chorobou s odstraněním vlivu věku.



Correlations

		Age	Pallidum_volume (mm3)	Nucl_caud_volume (mm3)
Age	Pearson Correlation	1	-.075	-.676
	Sig. (2-tailed)		.293	.000
	N	197	197	197
Pallidum_volume (mm3)	Pearson Correlation	-.075	1	.113
	Sig. (2-tailed)	.293		.113
	N	197	197	197
Nucl_caud_volume (mm3)	Pearson Correlation	-.676	.113	1
	Sig. (2-tailed)	.000	.113	
	N	197	197	197





Parciální korelace – příklad – pokračování

- Korelace pallida s věkem je nevýznamná, negativní a velmi slabá, avšak objem nucleus caudatus je významně a silně negativně korelovaný s věkem.
- Věk tedy může zkreslit hodnocení vztahu objemu těchto dvou struktur → použijeme parciální korelaci:

Partial Corr

Correlations

Control Variables			Pallidum_volume (mm3)	Nucl_caud_volume (mm3)
Age	Pallidum_volume (mm3)	Correlation	1.000	.085
		Significance (2-tailed)	.	.237
		df	0	194
	Nucl_caud_volume (mm3)	Correlation	.085	1.000
		Significance (2-tailed)	.237	.
		df	194	0

Závěr: Vztah objemu pallida a nucleus caudatus je velmi slabý a statisticky nevýznamný (0,085, p=0,237).

Korelace je nižší než Pearsonův korelační koeficient (0,113, p=0,113). Věk tedy opravdu hodnocení vztahu objemu pallida a nukleus caudatus částečně zkreslil.

Parciální korelace – poznámka

- Je možno spočítat parametrickou i neparametrickou parciální korelaci.

- Výpočet v SPSS:
  - parametrickou parciální korelaci lze „vyklikat“ či použít syntax
  - neparametrickou parciální korelaci lze spočítat pouze použitím syntaxu

3. Základy regresní analýzy

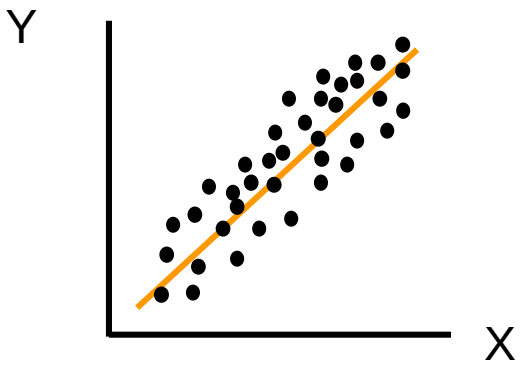
Motivace

- Cílem regresní analýzy je popsat závislost hodnot jedné proměnné na hodnotách druhé proměnné.
- Např. závislost objemu hipokampu na věku.

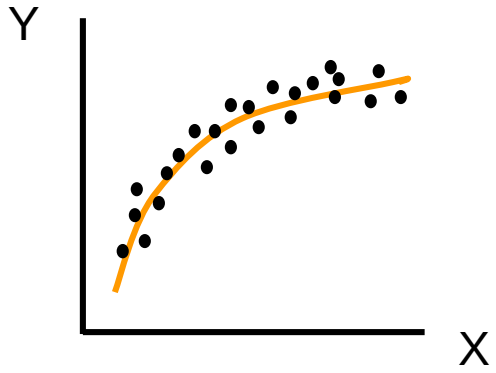
- Dva problémy:
  - Vybrat správnou funkci k popisu dané závislosti.
  - Stanovit konkrétní parametry daného typu funkce.

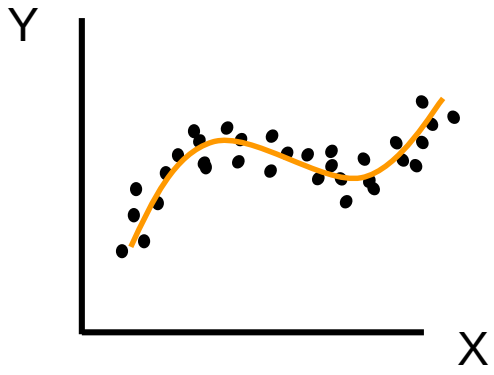
Příklady závislostí

Lineární



Nelineární

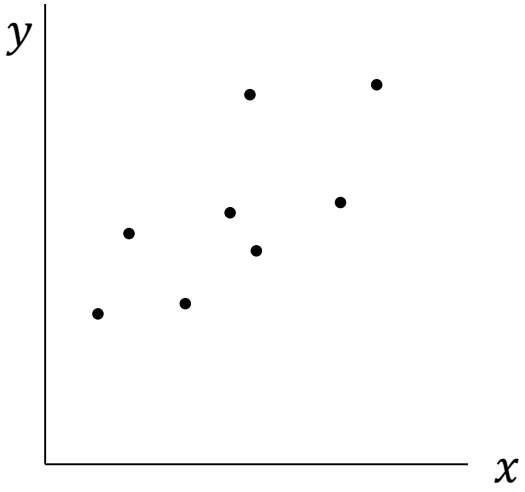




Lineární regrese

Obecný zápis: $\boldsymbol{y} = \boldsymbol{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Zápis, pokud máme pouze jednu nezávisle proměnnou: $\boldsymbol{y} = \beta_0 + \beta_1 * \boldsymbol{x} + \boldsymbol{\varepsilon}$



y – závisle proměnná (vysvětlovaná proměnná)

x – nezávisle proměnná (vysvětlující proměnná, regresor)

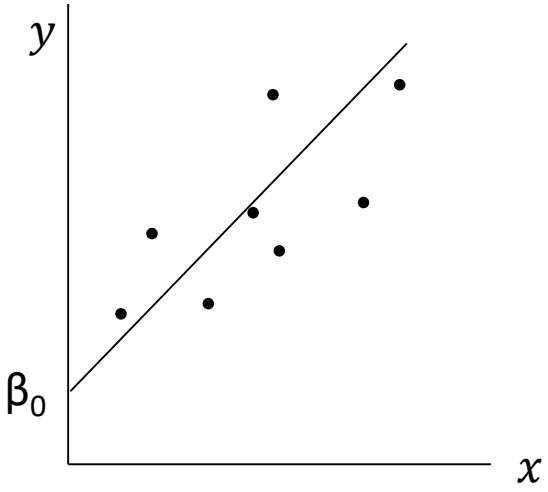
ε – náhodná složka modelu přímky (rezidua přímky)

β_0 – intercept
β_1 – regresní koeficient – „sklon regresní přímky“

Lineární regrese

Obecný zápis: $\boldsymbol{y} = \boldsymbol{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Zápis, pokud máme pouze jednu nezávisle proměnnou: $\boldsymbol{y} = \beta_0 + \beta_1 * \boldsymbol{x} + \boldsymbol{\varepsilon}$



y – závisle proměnná (vysvětlovaná proměnná)

x – nezávisle proměnná (vysvětlující proměnná, regresor)

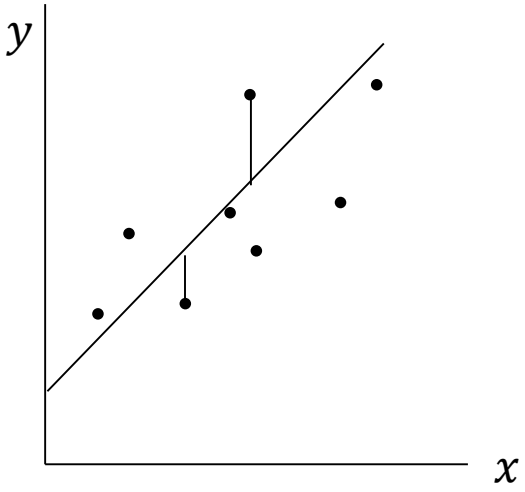
ε – náhodná složka modelu přímky (rezidua přímky)

β_0 – intercept
β_1 – regresní koeficient – „sklon regresní přímky“

Lineární regrese

Obecný zápis: $\boldsymbol{y} = \boldsymbol{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Zápis, pokud máme pouze jednu nezávisle proměnnou: $\boldsymbol{y} = \beta_0 + \beta_1 * \boldsymbol{x} + \boldsymbol{\varepsilon}$



y – závisle proměnná (vysvětlovaná proměnná)

x – nezávisle proměnná (vysvětlující proměnná, regresor)

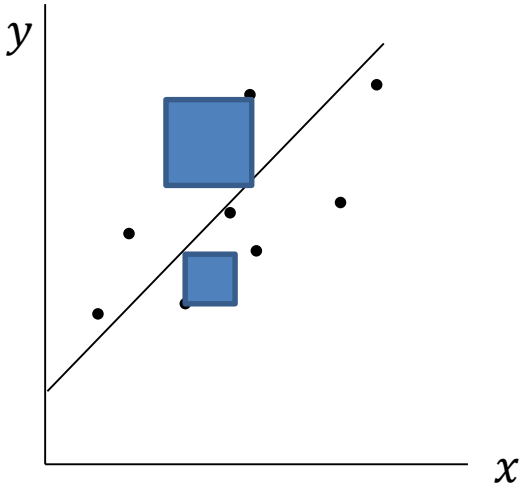
ε – náhodná složka modelu přímky (rezidua přímky)

β_0 – intercept
β_1 – regresní koeficient – „sklon regresní přímky“

Lineární regrese

Obecný zápis: $\boldsymbol{y} = \boldsymbol{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Zápis, pokud máme pouze jednu nezávisle proměnnou: $\boldsymbol{y} = \beta_0 + \beta_1 * \boldsymbol{x} + \boldsymbol{\varepsilon}$



y – závisle proměnná (vysvětlovaná proměnná)

x – nezávisle proměnná (vysvětlující proměnná, regresor)

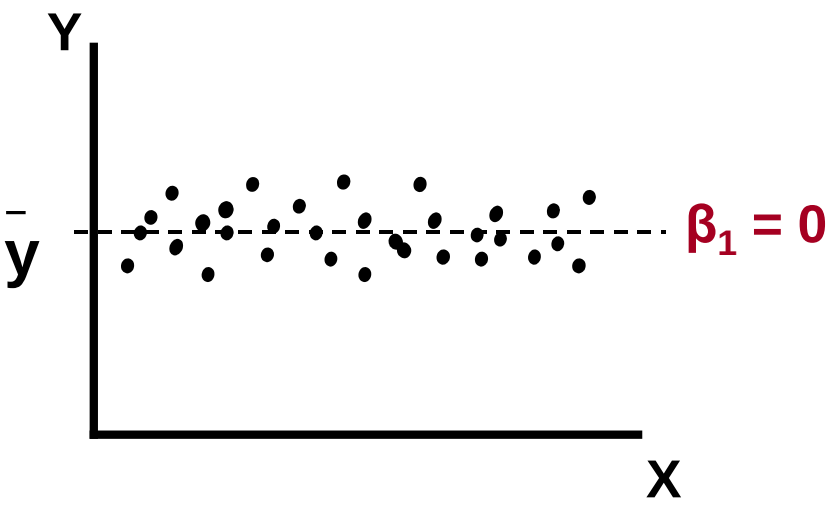
ε – náhodná složka modelu přímky (rezidua přímky)

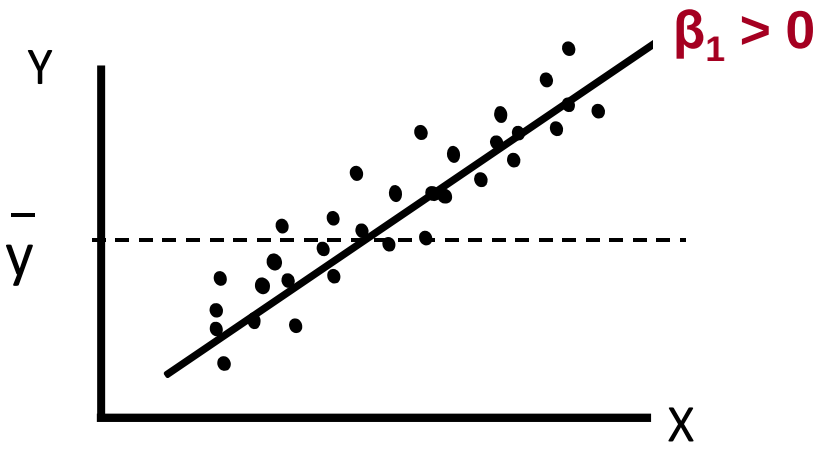
β_0 – intercept
β_1 – regresní koeficient – „sklon regresní přímky“

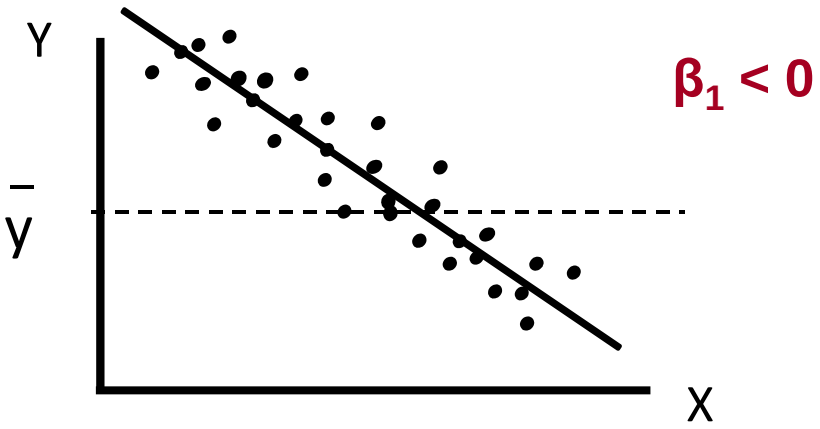
Odhad koeficientů β metodou nejmenších čtverců:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

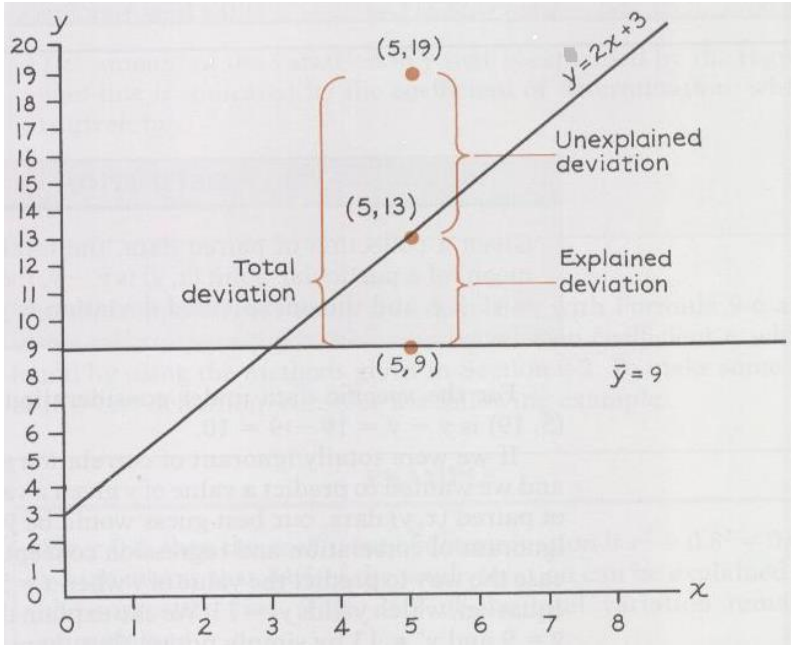
Lineární regrese - příklady







Lineární regrese



Převzato z přednášek
RNDr. Marie Budíkové, Dr.

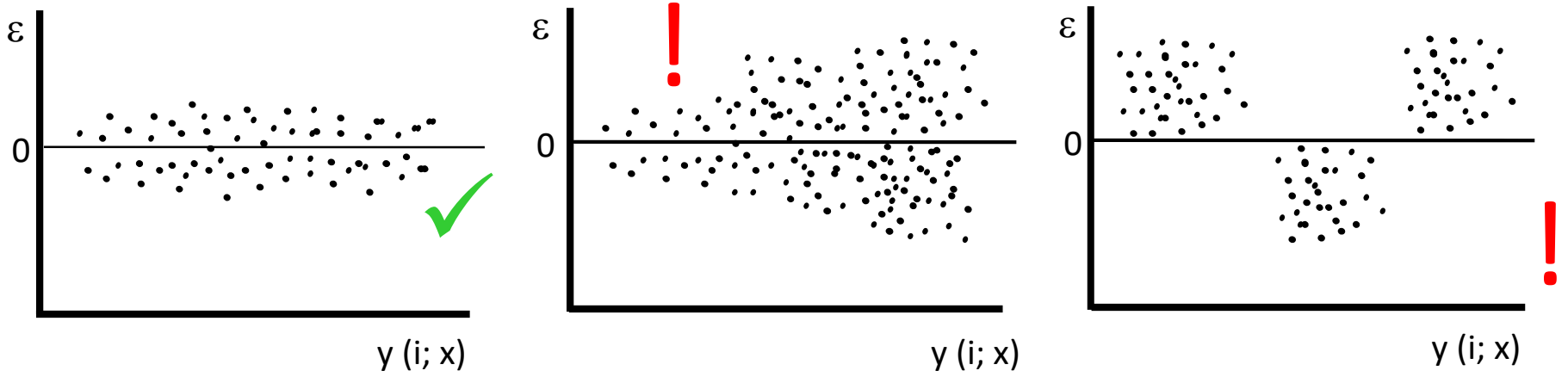
Testování významnosti modelu jako celku – celkový F-test:

zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	S_R	p	S_R/p	$\frac{S_R/p}{S_E/(n-p-1)}$
reziduální	S_E	n-p-1	$S_E/(n-p-1)$	-
celkový	S_T	n-1	-	-

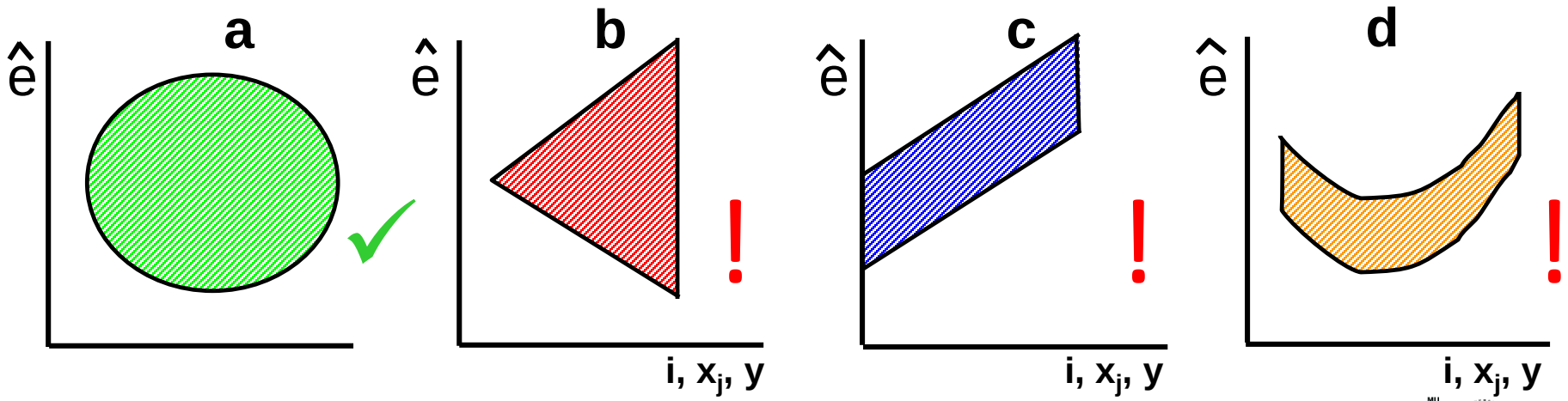
n ... počet subjektů; p ... počet proměnných

Regresní analýza v grafech

Grafy residuí modelů (příklady)



Obecné tvary residuí modelů (schéma)



Lineární regrese – příklad I

- **Příklad:** Proveďte regresní analýzu, v níž budete modelovat závislost objemu nucleus caudatus na věku.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,627[a]	,393	,392	494,9657860

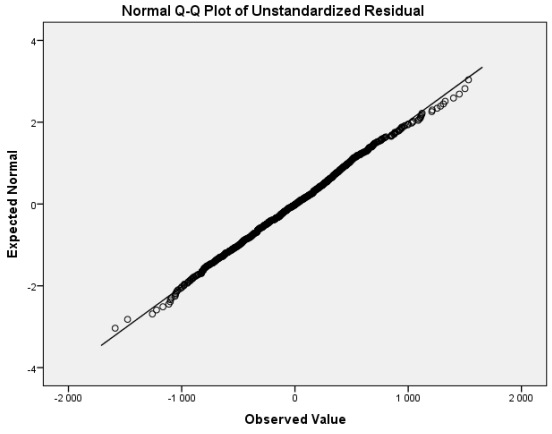
a. Predictors: (Constant), Age

Coefficients[a]

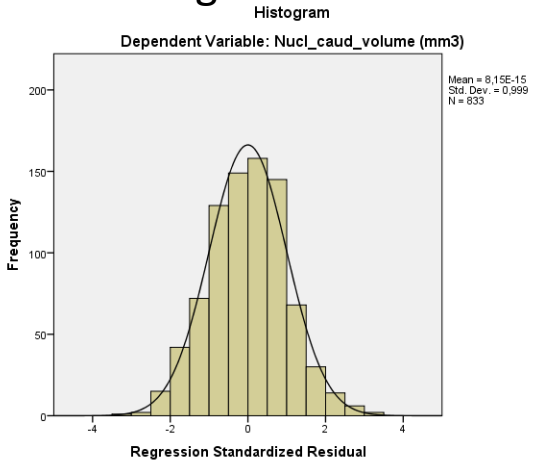
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8348,848	186,056		44,873	,000
	Age	-57,369	2,475	-,627	-23,176	,000

a. Dependent Variable: Nucl_caud_volume (mm3)

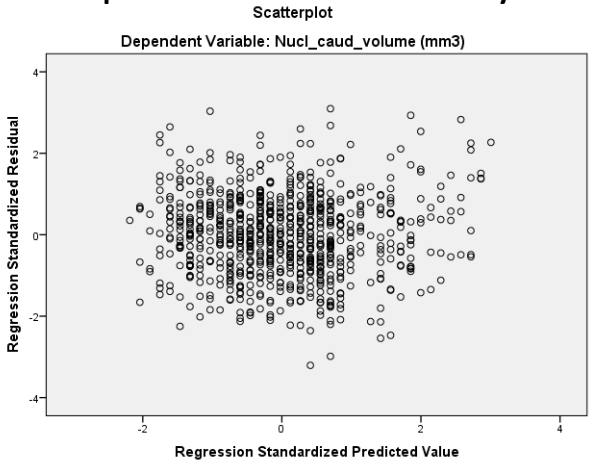
Q-Q graf reziduí



Histogram reziduí



Bodový graf reziduí vs. predikované hodnoty



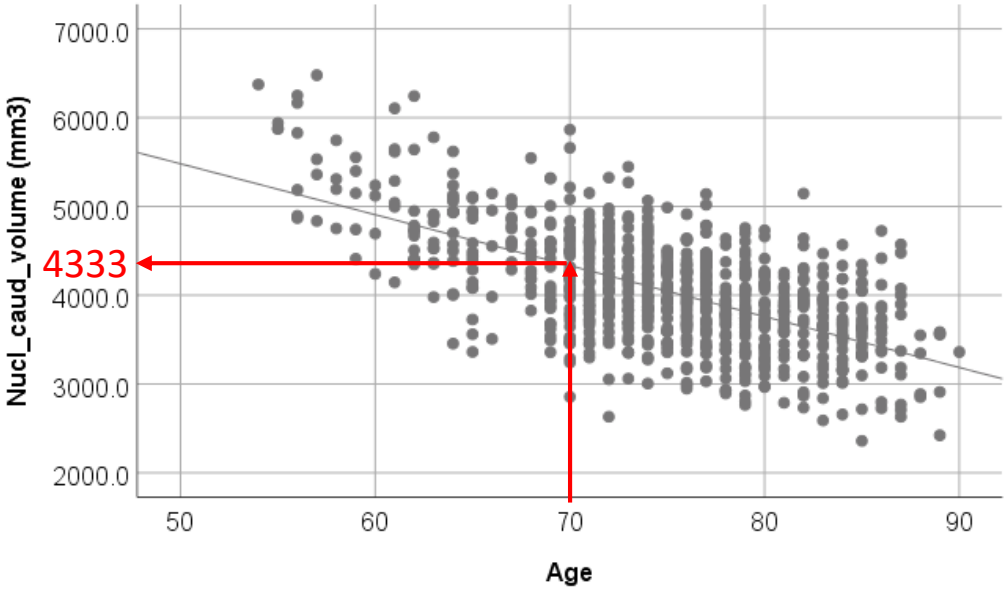
Lineární regrese – příklad I – pokračování

Jaký bude přibližně objem nucleus caudatus pro pacienta s věkem 70?

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error
1	(Constant)	8348,848	186,056
	Age	-57,369	2,475

a. Dependent Variable: Nucl_caud_volume (mm3)



Ruční výpočet:

8348,848 - 57,369 * 70 = 4333,018

Závěr: Odhad objemu nucleus caudatus pro pacienta s věkem 70 je přibližně 4333 mm^3.

Poznámka: Bylo by vhodné doplnit i intervaly spolehlivosti pro tento odhad, ruční výpočet je však vcelku komplikovaný.

Lineární regrese – příklad II

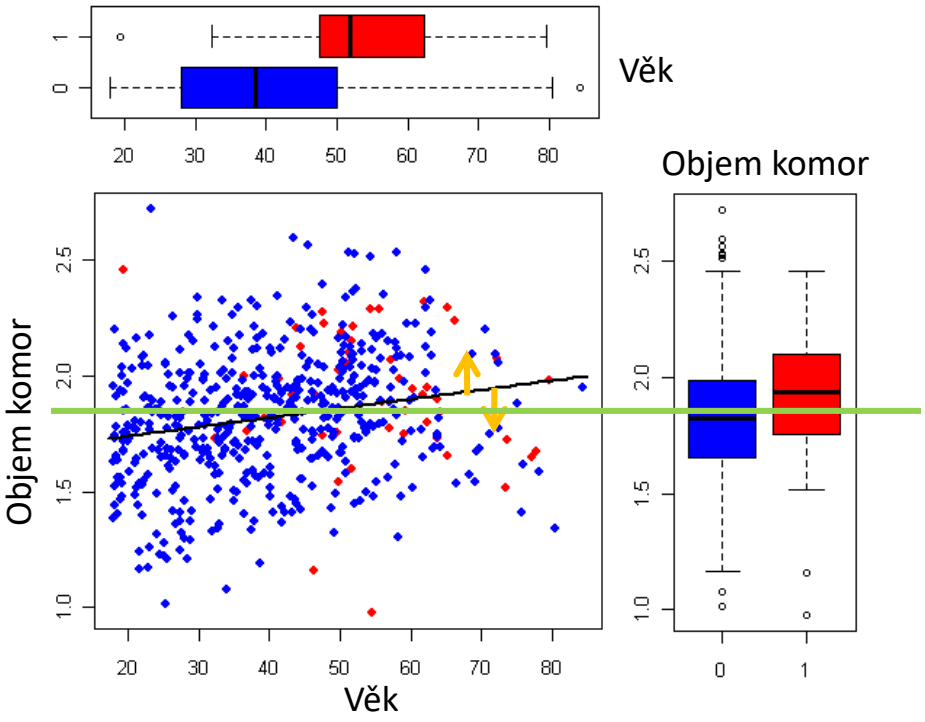
- **Příklad:** Chceme zjistit, zda se liší objem nucleus caudatus podle typu onemocnění (pacienti s AD, pacienti s MCI, kontroly). Srovnávané skupiny subjektů však obsahují jiný poměr mužů a žen a liší se i věkovým složením. Odstraňte vliv věku a pohlaví, aby výsledek srovnání objemu nucleus caudatus podle typu onemocnění nebyl ovlivněn tím, že skupiny nejsou srovnatelné.

	Nucl_caud_volumemm3	Hippocampus_volume_24mm3	hip_rozdil	PRE_1	RES_1	ZPR_1	ZRE_1
1	3527,724137000...	.	.	3543,61046	-15,88633	-1,28509	-,03210
2	3773,458262000...	.	.	3967,79607	-194,33781	-,21954	-,39265
3	4294,449622000...	.	.	3831,22444	463,22519	-,56261	,93591
4	3585,004603000...	.	.	3219,99974	365,00486	-2,09800	,73747
5	3723,259473000...	.	.	4255,41004	-532,15057	,50294	-1,07517
6	3969,370347000...	.	.	4312,93283	-343,56249	,64744	-,69414
7	2886,235913000...	.	.	3277,52254	-391,28662	-1,95350	-,79057
8	3741,225598000...	.	.	3392,56812	348,65747	-1,66451	,70444
9	3737,405432000...	.	.	3507,61371	229,79172	-1,37551	,46428
10	2630,569601000...	.	.	3335,04533	-704,47573	-1,80900	-1,42334
11	2892,733743000...	.	.	3852,75048	-960,01674	-,50854	-1,93964
12	3551,229211000...	.	.	3543,61046	7,61875	-1,28509	,01539

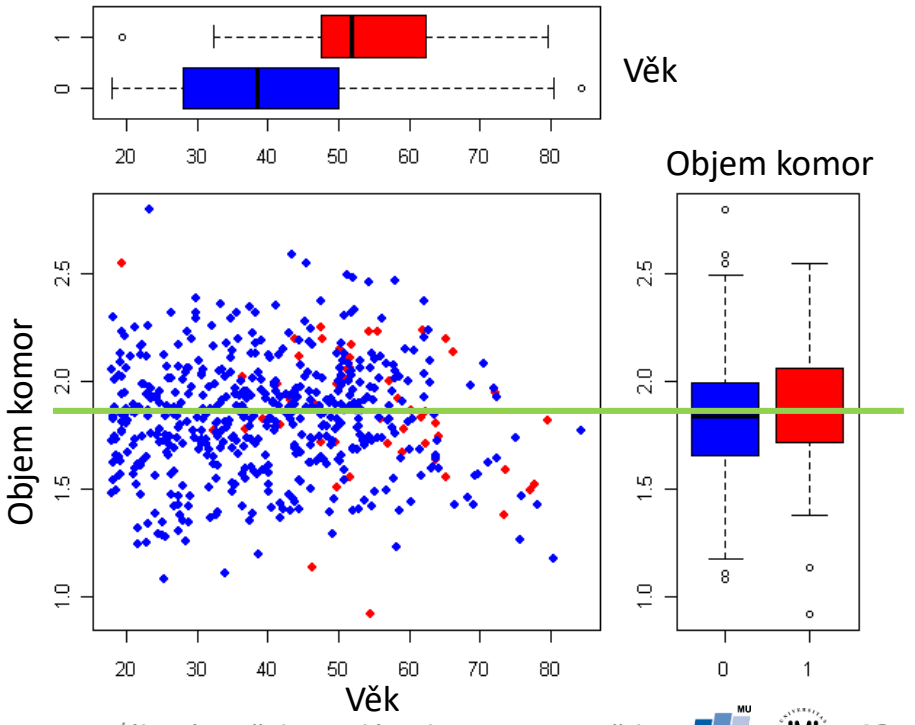
Odstranění vlivu kovariát (tzv. adjustace)

1. V prvním kroku definujeme regresní model vztahu kovariáty (např. věku) a dané proměnné
2. Pro každého pacienta je vypočteno jeho reziduum od regresní přímky ↑↓
3. Reziduum (představující hodnotu parametru po odečtení vlivu věku, jeho průměr je 0) je přičteno k průměrné hodnotě parametru —
4. Výsledná adjustovaná hodnota má odečten vliv věku, ale zároveň není změněna číselná hodnota parametru

Původní data

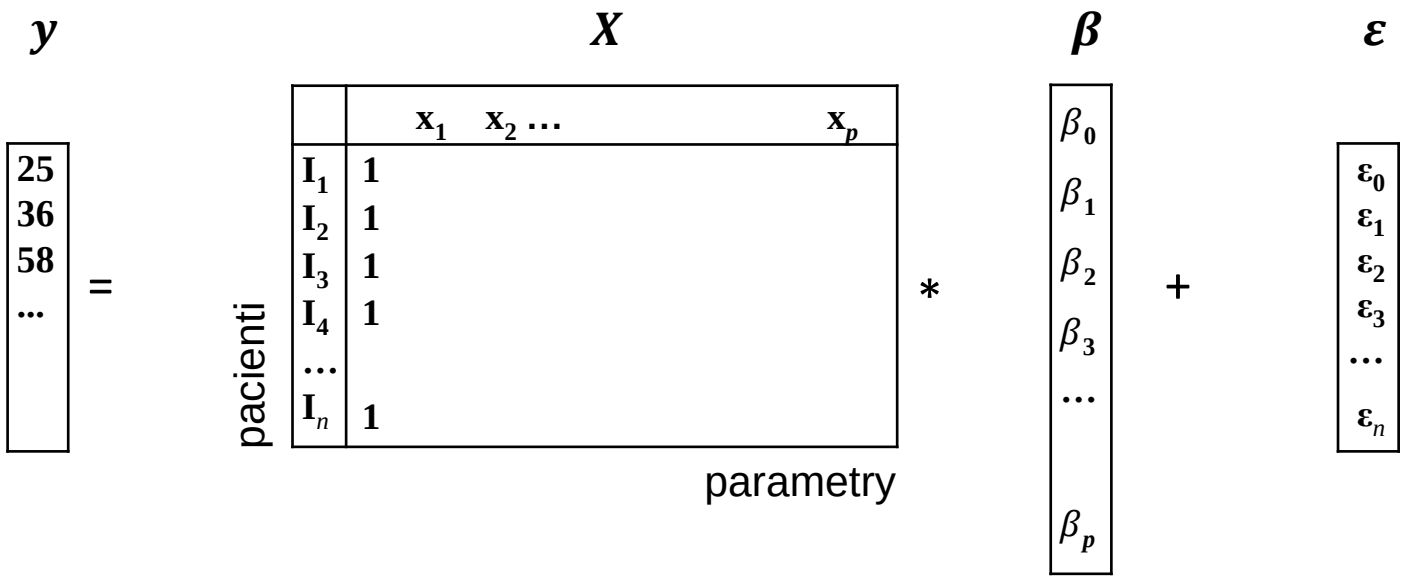


Adjustovaná data



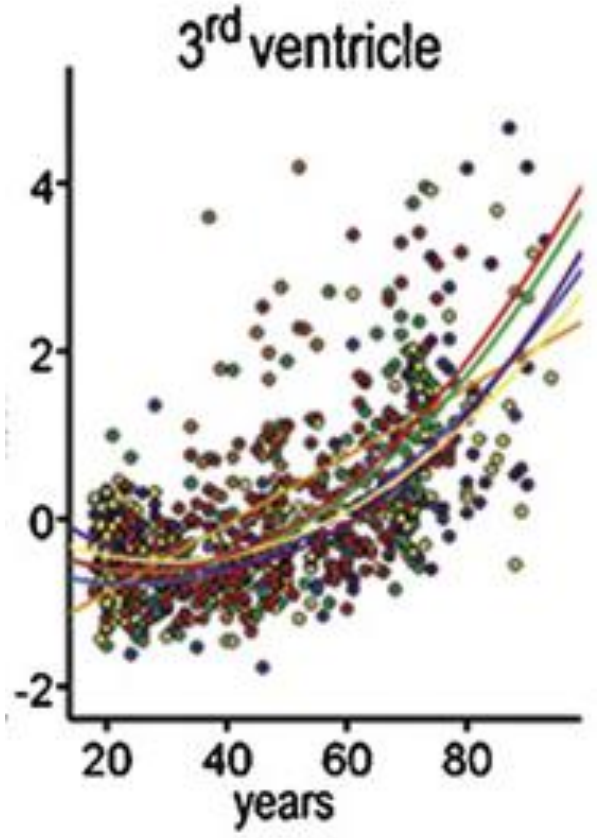
Vícenásobná lineární regrese

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$



X – matice plánu (design matice)

Kvadratická závislost objemu mozkové struktury na věku



Převzato z: Walhovd et al. 2011, Neurobiol. of aging

$$y = \beta_0 + \beta_1 * x + \beta_2 * x^2 + \varepsilon$$

$$\underset{y}{\begin{bmatrix} 1.5 \\ 2.6 \\ -0.8 \\ \dots \end{bmatrix}} = \underset{X}{\begin{array}{c|cc:c} & & \text{věk} & \text{věk*věk} \\ \hline I_1 & 1 & & \\ I_2 & 1 & & \\ I_3 & 1 & & \\ I_4 & 1 & & \\ \dots & \dots & & \\ I_n & 1 & & \end{array}} * \underset{\beta}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}} + \underset{\varepsilon}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \dots \\ \varepsilon_n \end{bmatrix}}$$

pacienti

parametry

Kategoriální data jako prediktory v regresi

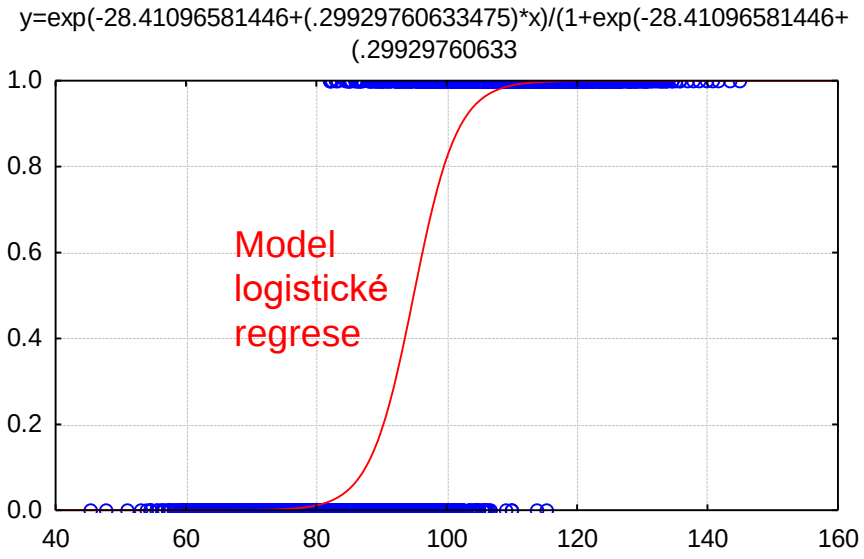
- Kategoriální a ordinální data mohou do analýzy vstupovat jako binární proměnné
- Kategoriální data (nelze seřadit) -> dummies
- Ordinální data (lze seřadit)
  - Dummies
  - Definice referenční kategorie (obvykle kategorie s nejnižším rizikem pro hodnocený endpoint

- Příklad: Stádium karcinomu

Původní	**Dummies**				**Vzhledem k referenci**		
Stádium	Stádium I	Stádium II	Stádium III	Stádium IV	Stád. II ref	Stád. III ref	Stád. IV ref
I	1	0	0	0	0	0	0
I	1	0	0	0	0	0	0
I	1	0	0	0	0	0	0
II	0	1	0	0	1		
II	0	1	0	0	1		
III	0	0	1	0		1	
III	0	0	1	0		1	
IV	0	0	0	1			1
IV	0	0	0	1			1

Logistická regrese

- Standardní metoda pro analýzu binárních charakteristik (pacient/kontrolní subjekt, zemřelý/žijící, s nežádoucími účinky/bez n. ú. apod.) bez vlivu času
- Modeluje závislost výskytu události (nežádoucího účinku, úmrtí, onemocnění) na binárních, kategoriálních nebo kvantitativních proměnných
- Výsledkem rovnice je pravděpodobnost, že u daného pacienta nastane hodnocená událost
- Alternativou jsou např. rozhodovací stromy, neuronové sítě a další klasifikační metody



Příklad logistické regrese: predikce binární charakteristiky (osa y) za pomoci kvantitativní proměnné (osa x)

Poděkování…

Příprava výukových materiálů předmětu „DSAN01 Analýza dat pro Neurovědy “ byla finančně podporována prostředky projektu FRVŠ č. 942/2013 „Inovace materiálů pro interaktivní výuku a samostudium předmětu Analýza dat pro Neurovědy“