Analýza dat pro Neurovědy

RNDr. Eva Janoušová

doc. RNDr. Ladislav Dušek, Dr.

Blok 7

Jak hodnotit vztah spojitých proměnných a základy regresního modelování.

Osnova

1. Základy korelační analýzy
2. Základy regresní analýzy

1. Základy korelační analýzy

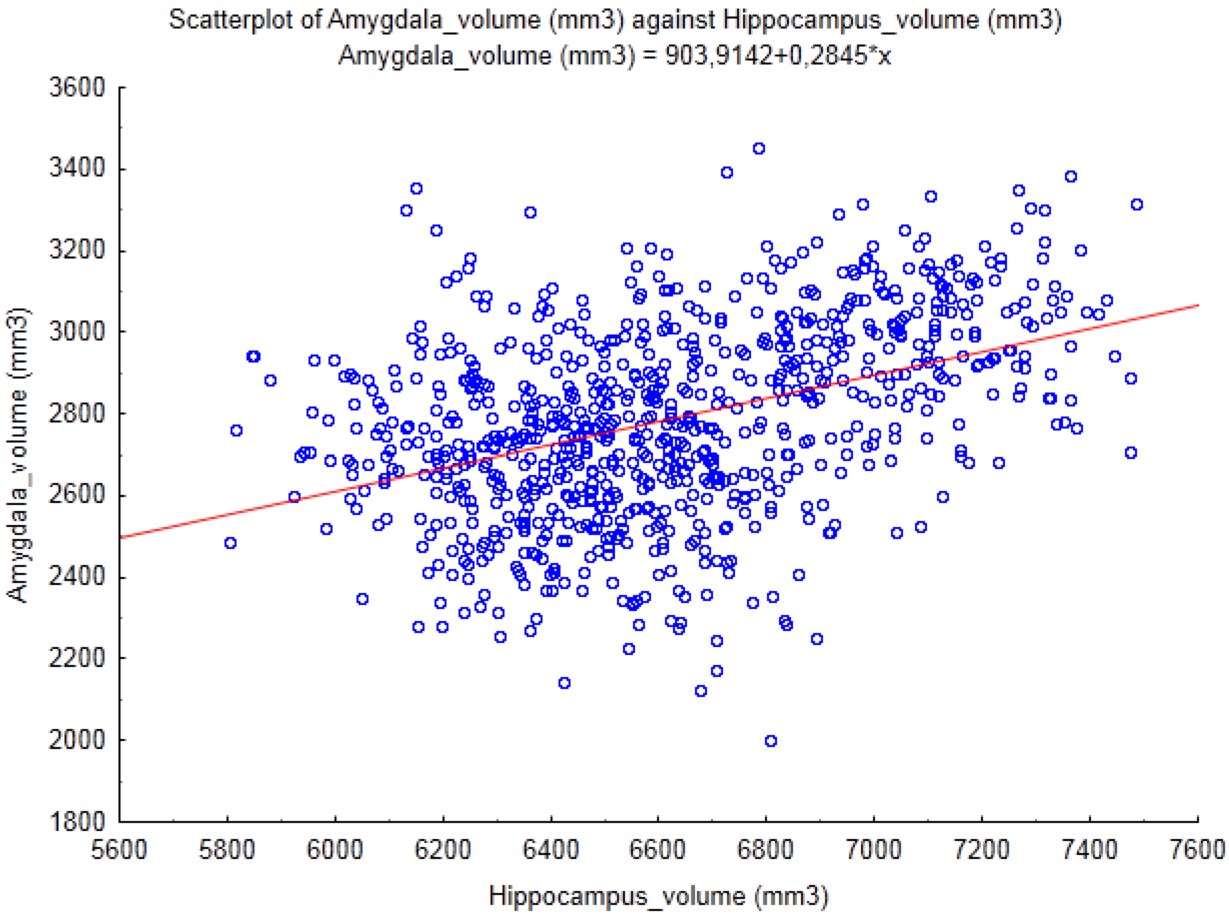
Motivace

- Zatím jsme se zabývali spojitou proměnnou v jedné skupině, spojitou proměnnou ve více skupinách, diskrétní proměnnou v jedné skupině, diskrétní proměnnou ve více skupinách, vztahem dvou diskrétních proměnných.

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

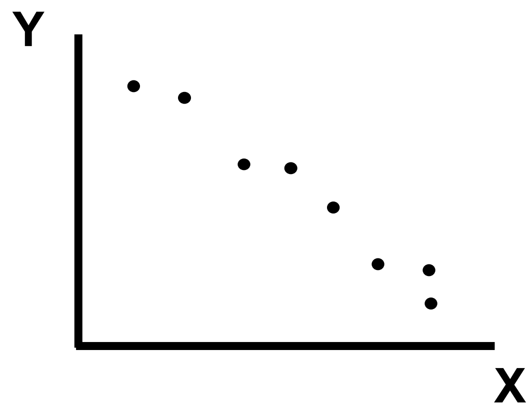
Jak hodnotit vztah dvou spojitých proměnných?

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

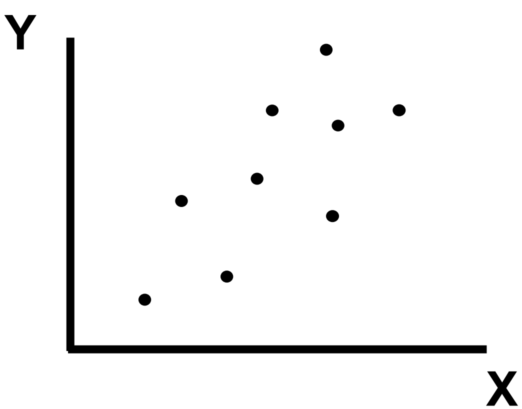


Pearsonův korelační koeficient (r)

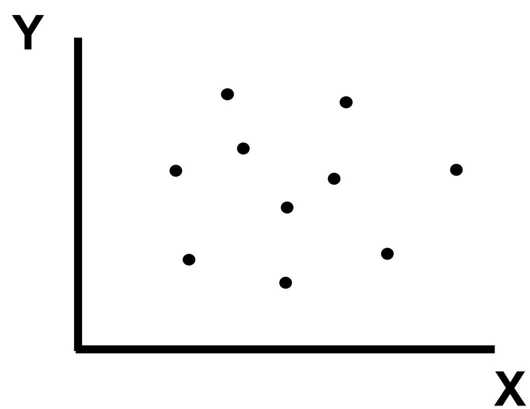
r = 1,0



r = -0,9



r = 0,4



r = 0,05

Korelace

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

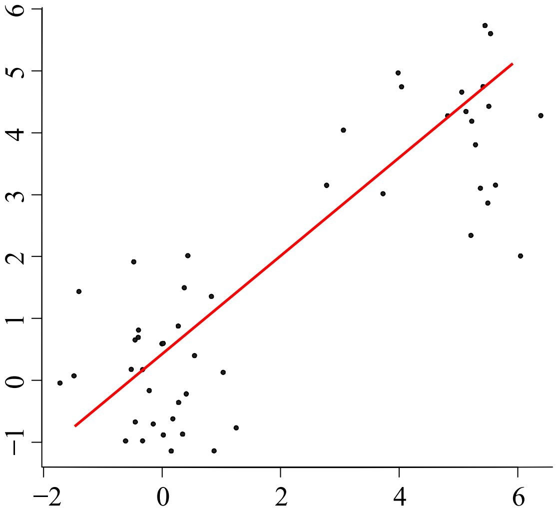
- Standardní metodou je výpočet **Pearsonova korelačního koeficientu (r)**:
  - Charakterizuje **linearitu** vztahu mezi X a Y – jinak řečeno variabilitu kolem lineárního trendu.
  - Nabývá hodnot od -1 do 1.
  - Hodnota r je kladná (kladná korelace), když vyšší hodnoty X souvisí s vyššími hodnotami 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 spojité proměnné nezávislé** – hypotézy mají tvar: $H_0\colon r = 0$ (tzn. korelační koeficient je roven nule) a $H_1\colon r \neq 0$.

Pearsonův korelační koef. – problematické situace I.

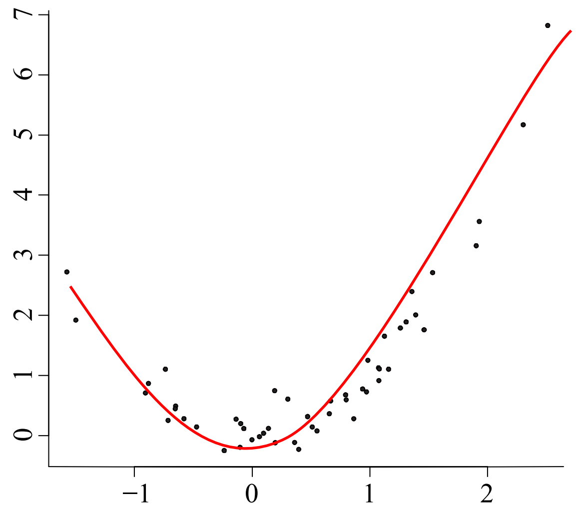
- Pearsonův korelační koeficient není vhodné počítat v situaci, kdy:
  - se v datech vyskytuje více skupin
  - proměnné mají nelineární vztah
  - se v datech vyskytují odlehlé hodnoty

Více skupin



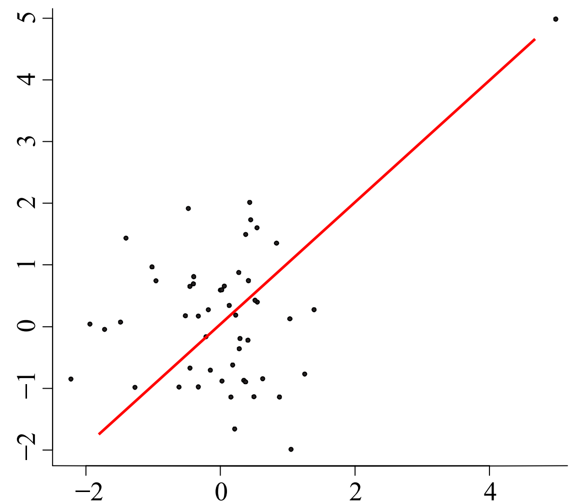
r = 0,84
(p < 0,001)

Nelineární vztah



r = 0,58
(p < 0,001)

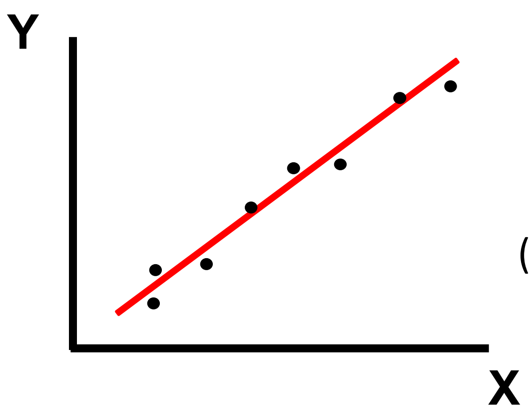
Odlehlá hodnota



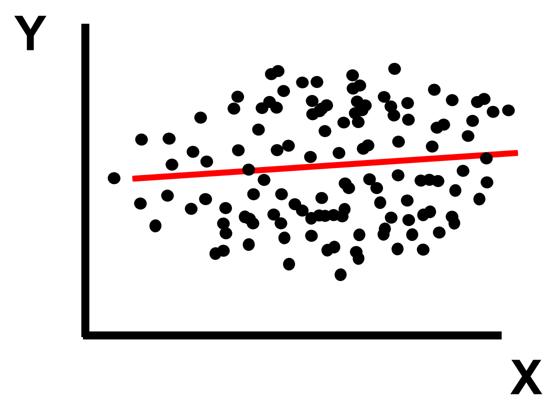
r = 0,36
(p = 0,009)

Pearsonův korelační koef. – problematické situace II.

- Problém velikosti vzorku:



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

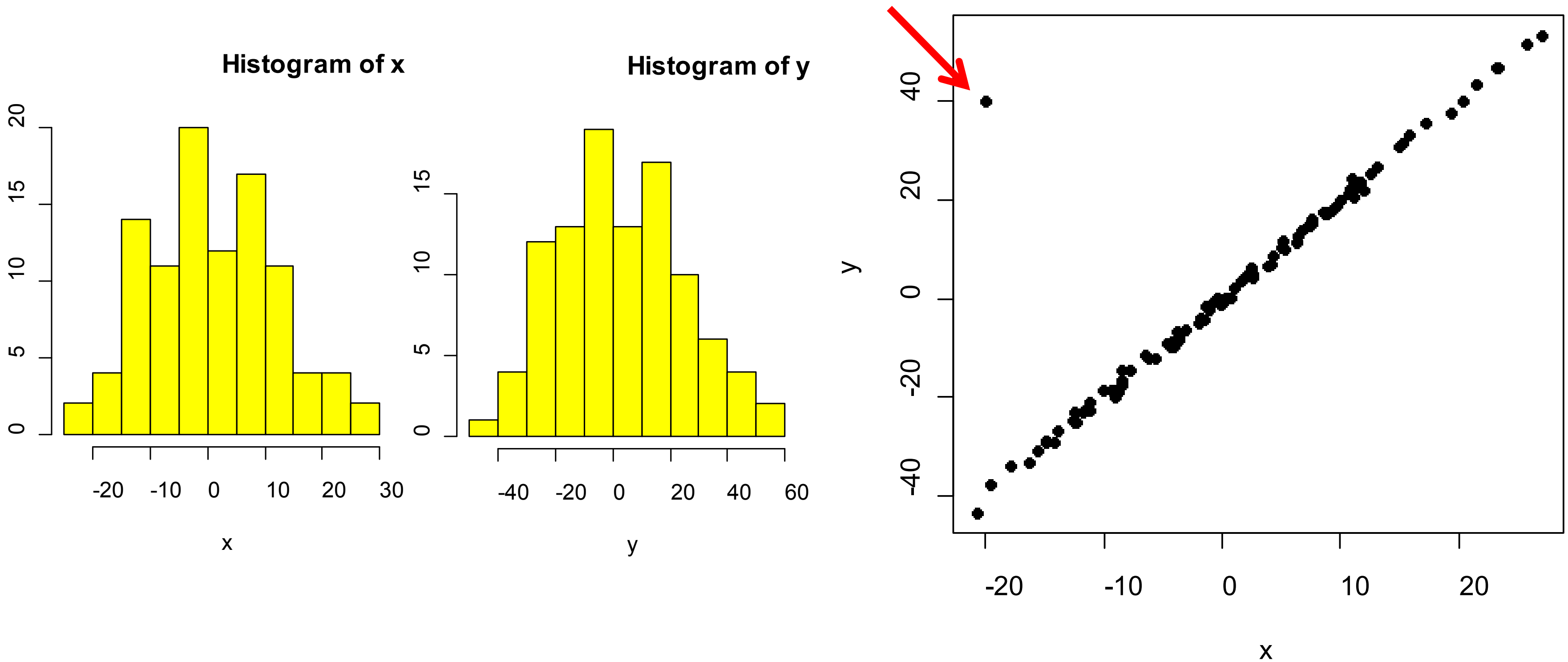


r = 0,012
(p < 0,008)

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

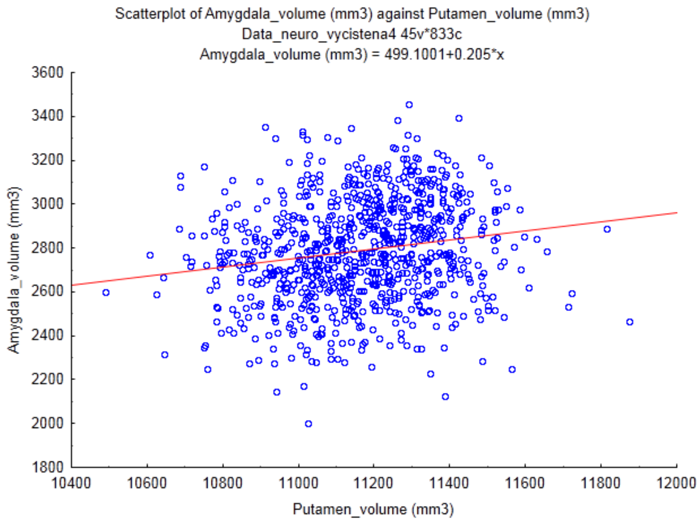
Pearsonův korelační koef. – problematické situace III.

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



Pearsonův korelační koeficient

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



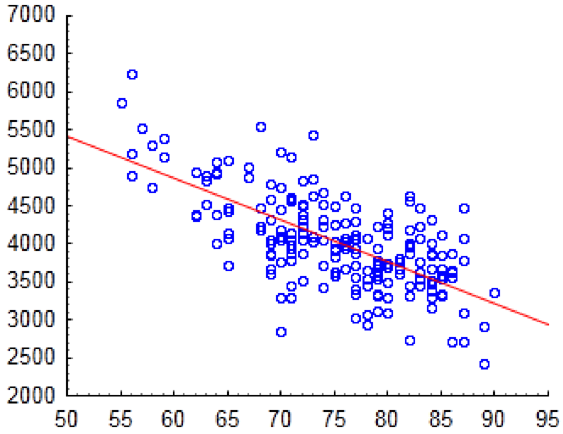
Variable	Correlations (Data_neuro_vycistena4) Marked correlations are significant at p < .05000 N=833 (Casewise deletion of missing data) Putamen_volume (mm3)	Amygdala_volume (mm3)
Putamen_volume (mm3)	1.0000	.1742
	p= ---	p=.000
Amygdala_volume (mm3)	.1742	1.0000
	p=.000	p= ---

Úkol 1.

- **Zadání:** Ověřte, zda existuje vztah objemu nucleus caudatus a věku u pacientů s AD, pacientů s MCI a u kontrol. Nezapomeňte ověřit normalitu srovnávaných proměnných.

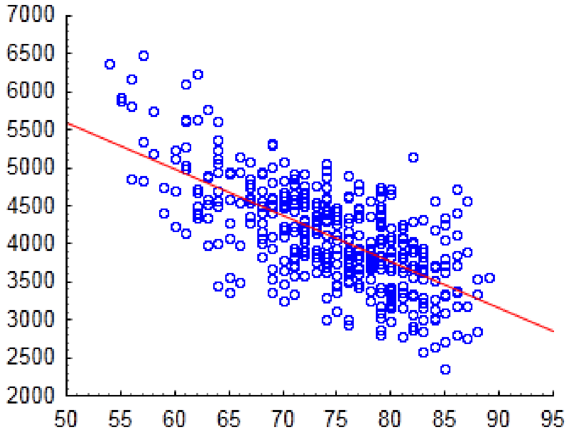
- **Řešení:**

AD



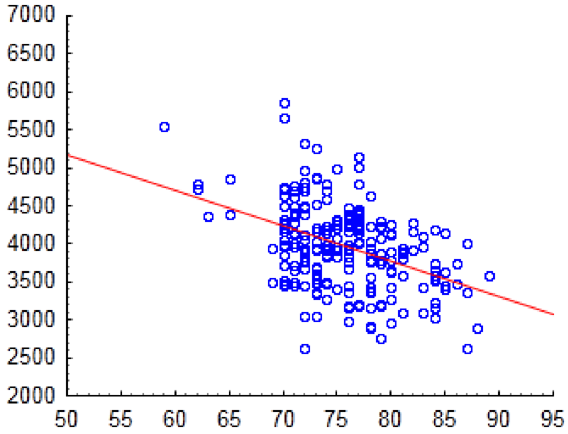
r = -0,68
(*p* < 0,001)

MCI



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

CN



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

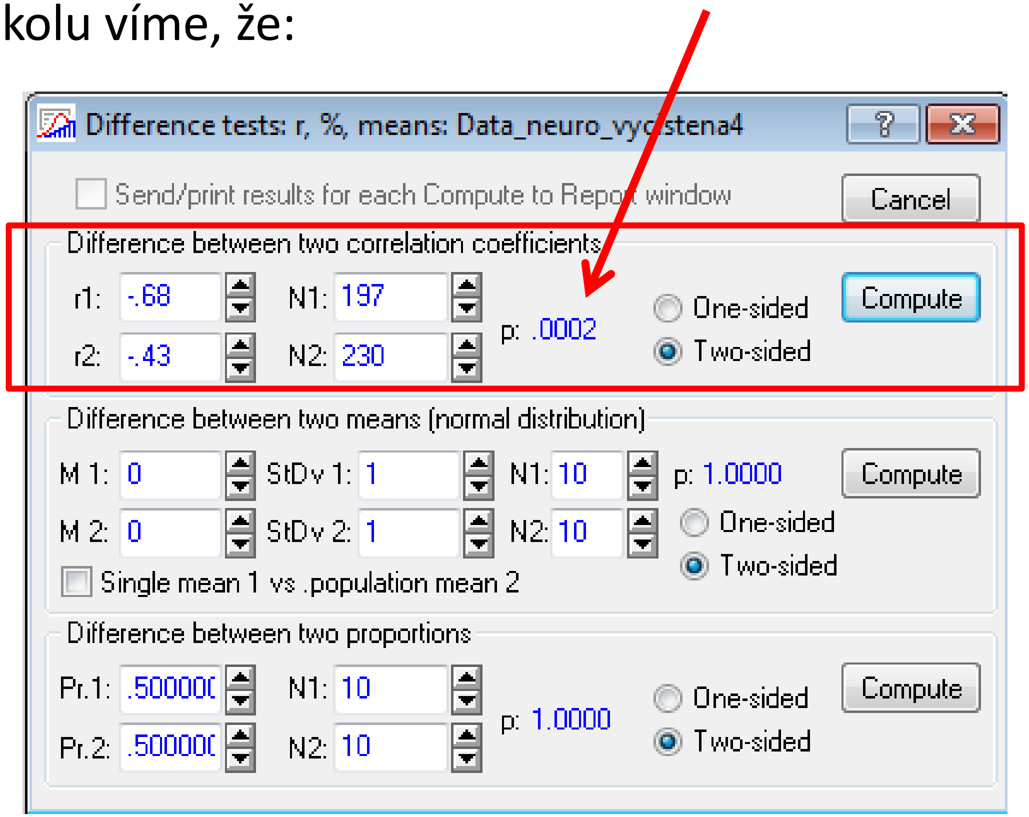
Srovnání dvou korelačních koeficientů

- **Příklad:** Srovnejte korelační koeficienty objemu nucleus caudatus a věku u pacientů s AD a kontrolních subjektů.

- **Postup:**

Z předchozího úkolu víme, že:

r_1 = -0,68
N_1 = 197
r_2 = -0,43
N_2 = 230



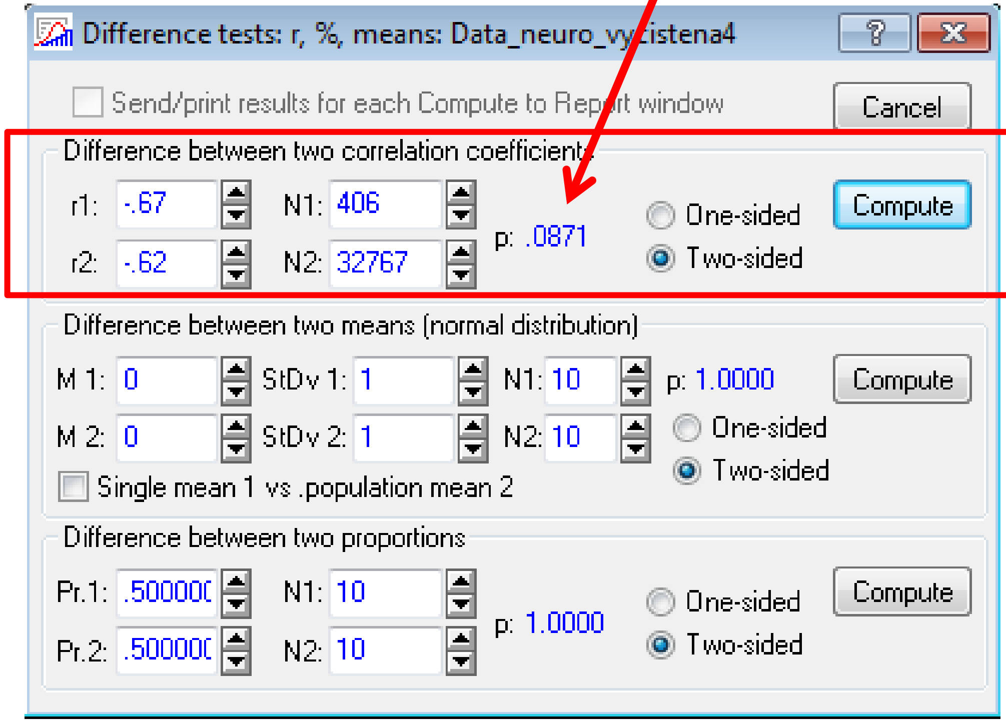
Srovnání korelačního koeficientu s referenční hodnotou

- **Příklad:** Srovnejte korelační koeficient objemu nucleus caudatus a věku u pacientů s MCI s hodnotou -0,62, jež byla zjištěna při populačním průzkumu.

- **Postup:**

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

Populační průzkum:
r_2 = -0,62
N_2 = 32767 (co největší N)



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.

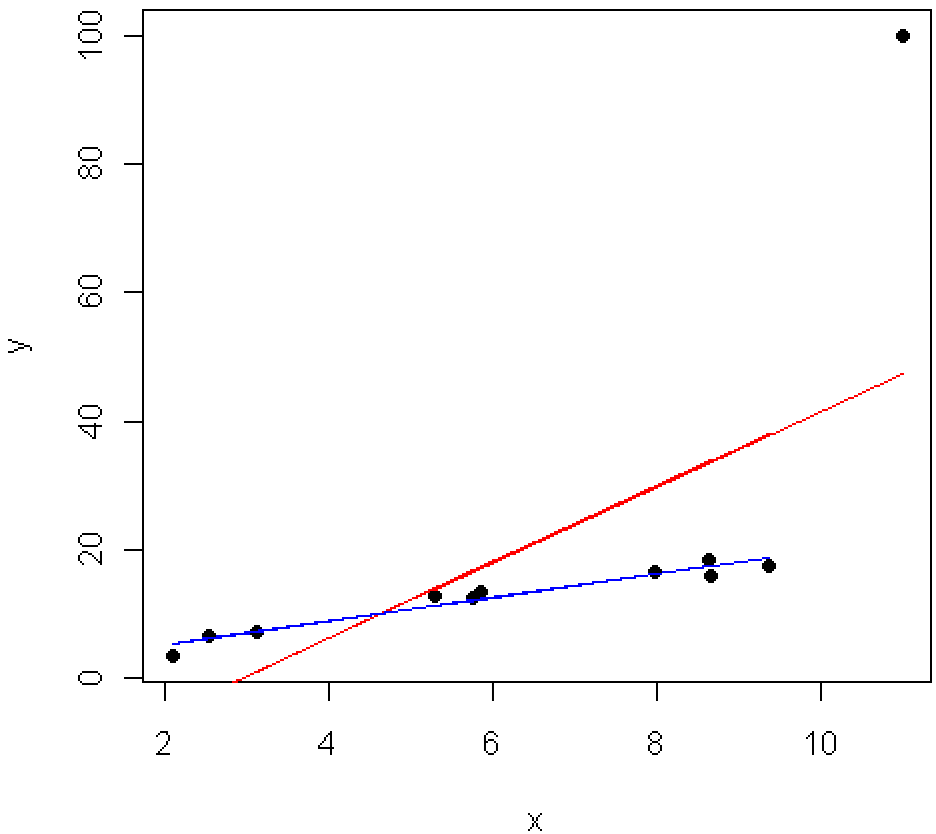
Spearmanův korelační koeficient (r_s)

- Pearsonův korelační koeficient je náchylný k odlehlým hodnotám a obecně odchylkám od normality.

- **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.

Srovnání Pearsonova a Spearmanova korelačního koeficientu



Pearsonův korelační koeficient:

r = 0,65

(p = 0,029)

Spearmanův korelační koeficient:

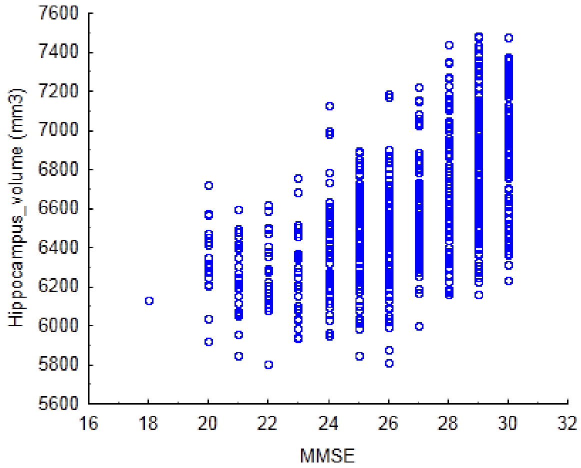
r_S = 0,95

(p < 0,001)

Spearmanův korelační koeficient není náchylný k odlehlým hodnotám.

Spearmanův korelační koeficient

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



Variable	Spearman Rank Order Correlations (Data_neuro_vycistena4) MD pairwise deleted Marked correlations are significant at p <,05000 MMSE	Hippocampus_volume (mm3)
MMSE	1,000000	0,626892
Hippocampus_volume (mm3)	0,626892	1,000000

Pair of Variables	Spearman Rank Order Correlations (Data_neuro_vycistena4) MD pairwise deleted Marked correlations are significant at p <,05000 **Valid N**	Spearman R	t(N-2)	p-value
MMSE & Hippocampus_volume (mm3)	833	0,626892	23,19513	0,00

Úkol 2.

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

- **Řešení:**

Pair of Variables	Valid N	Spearman R	t(N-2)	p-value
	Spearman Rank Order Correlations (Data_neuro_vycistena4) MD pairwise deleted Marked correlations are significant at p <,05000			
MMSE & Amygdala_volume (mm3)	833	0,338742	10,37852	0,000000
MMSE & Thalamus_volume (mm3)	833	-0,000759	-0,02187	0,982557
MMSE & Pallidum_volume (mm3)	833	0,039167	1,12992	0,258834
MMSE & Putamen_volume (mm3)	833	0,324925	9,90402	0,000000
MMSE & Nucl_caud_volume (mm3)	833	0,011837	0,34124	0,733012

2. Základy regresní analýzy

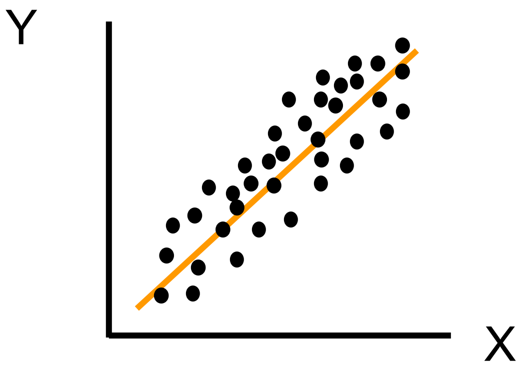
Motivace

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

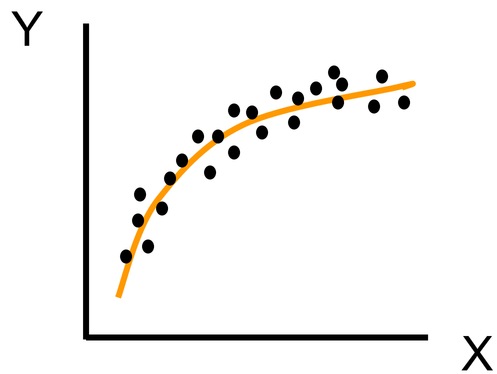
- Dva problémy:
  – Vybrat správnou funkci k popisu dané závislosti.
  – Stanovit konkrétní parametry daného typu funkce.

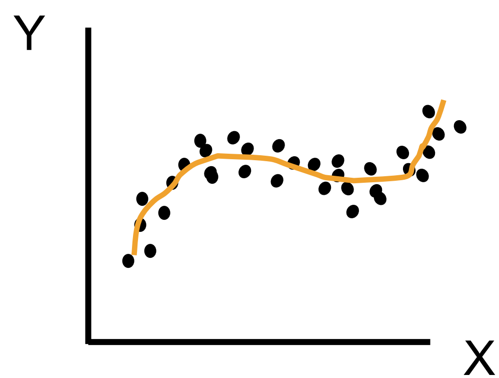
Příklady závislostí

Lineární



Nelineární

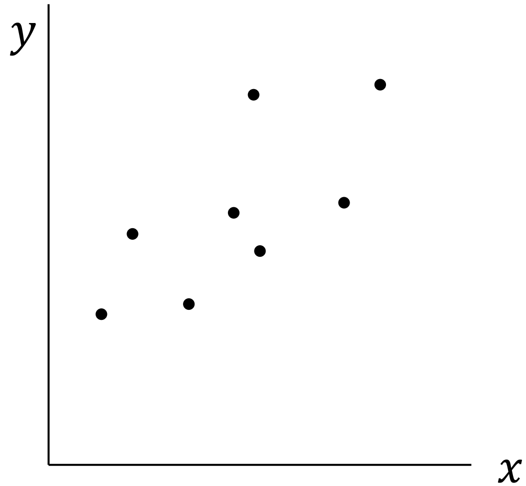




Lineární regrese

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)

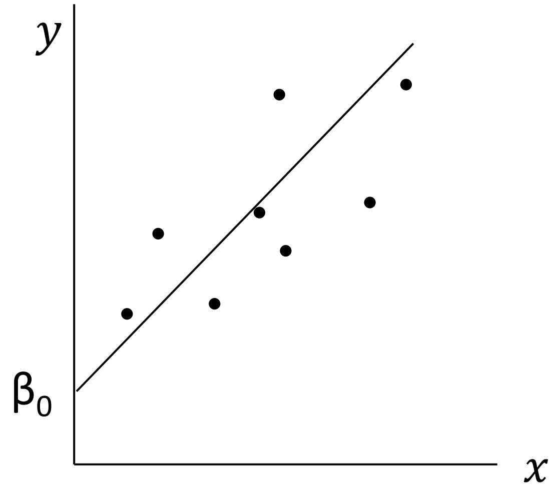
ε – náhodná složka modelu přímky (rezidua přímky)

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)

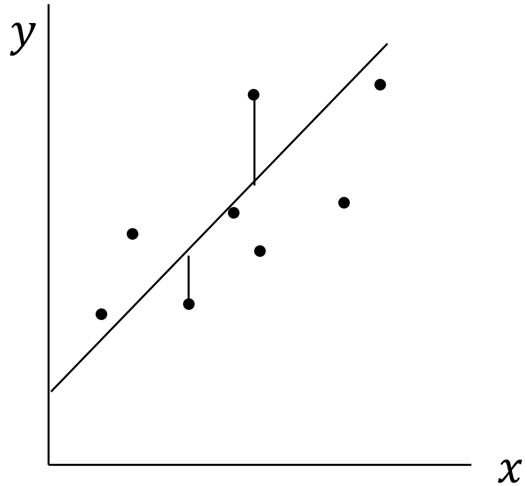
$\boldsymbol{\varepsilon}$ – náhodná složka modelu přímky (rezidua přímky)

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)

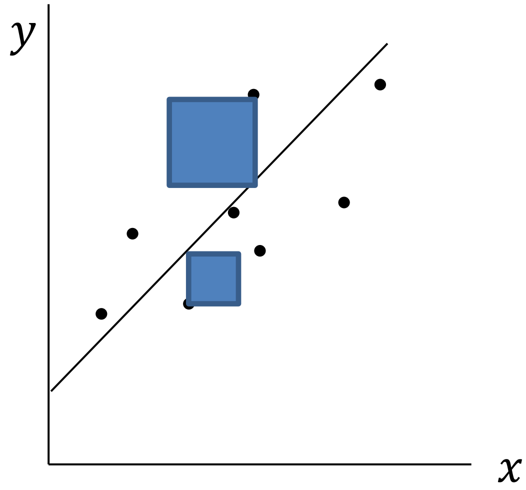
ε – náhodná složka modelu přímky (rezidua přímky)

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)

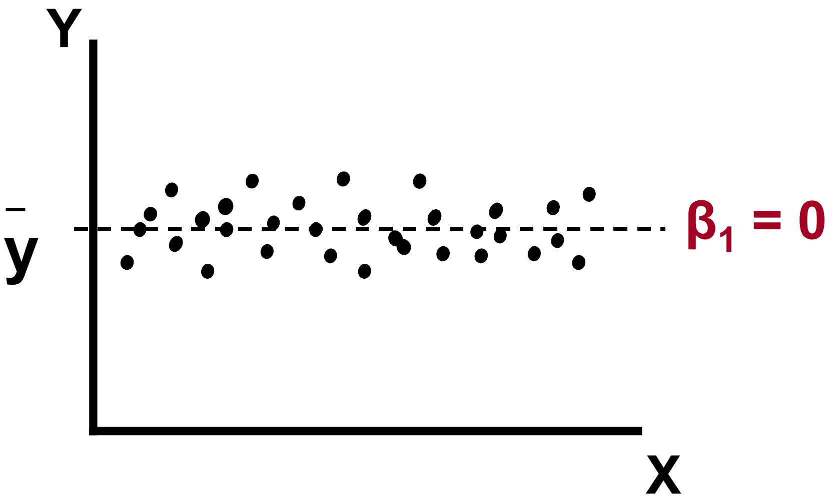
ε – náhodná složka modelu přímky (rezidua přímky)

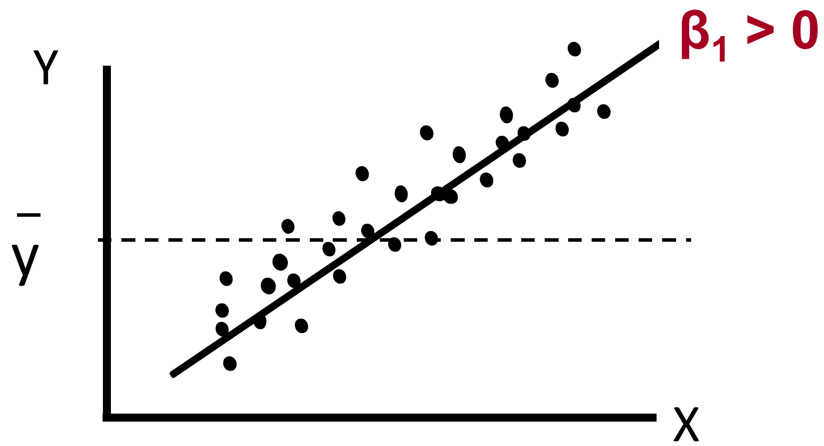
β_0 – intercept
β_1 – regresní koeficient – „sklon regresní přímky“

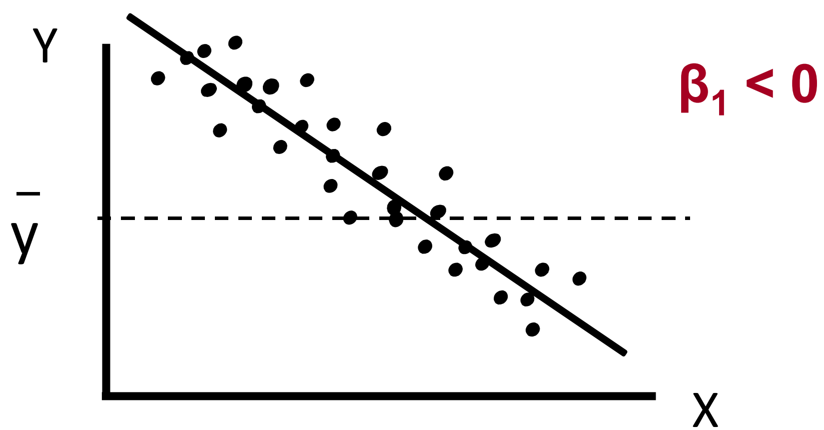
Odhad koeficientů β metodou nejmenších čtverců:

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

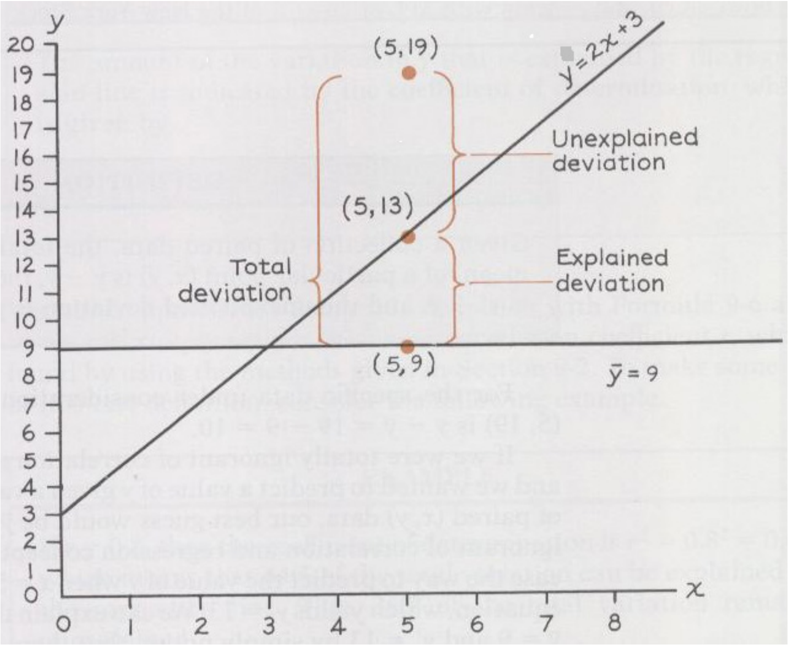
Lineární regrese - příklady







Lineární regrese



Převzato z přednášek
RNDr. Marie Budíkové, Dr.

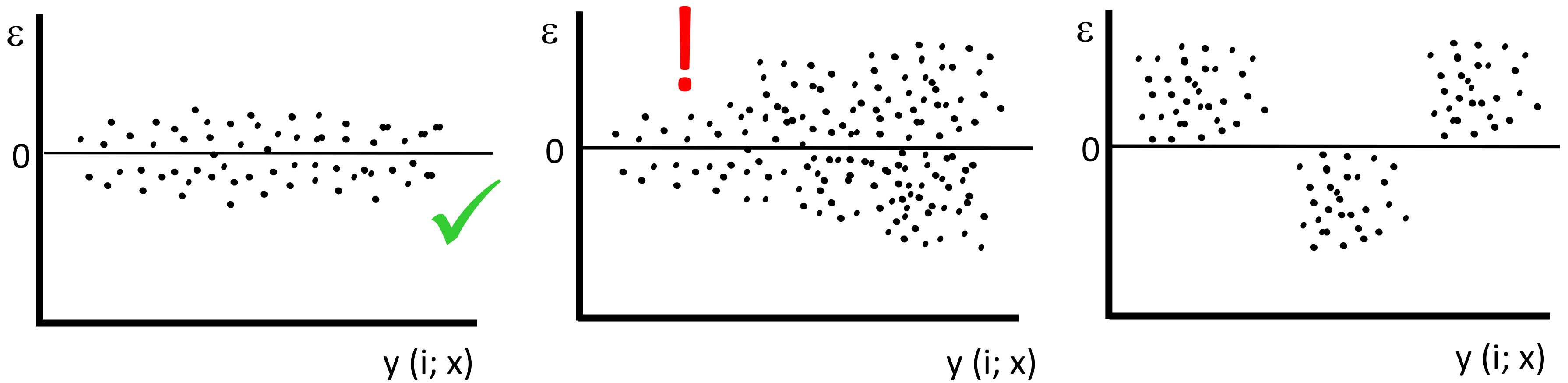
Testování významnosti modelu jako celku – celkový F-test:

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

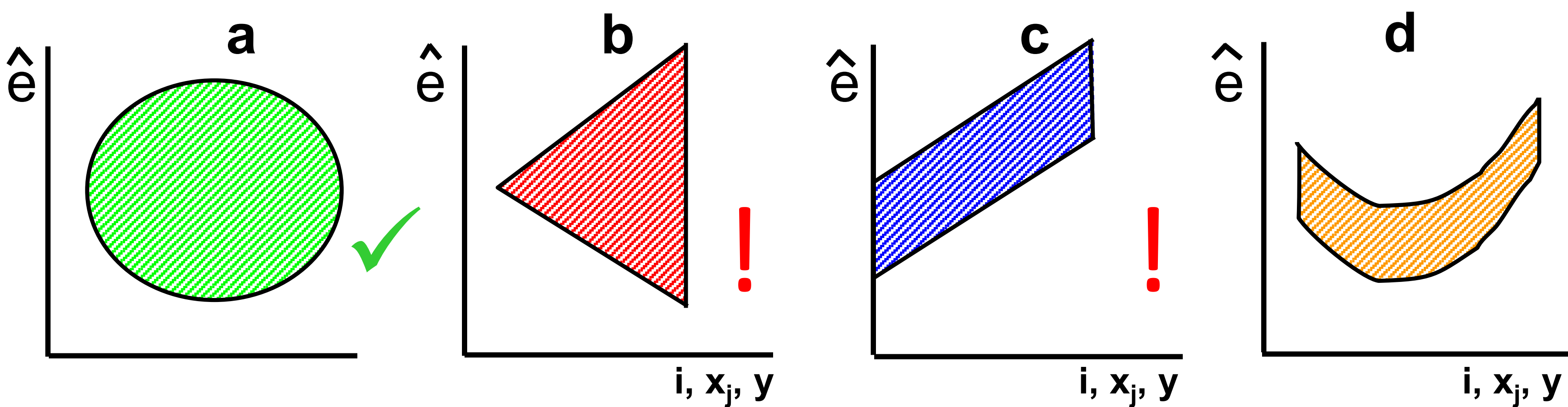
n ... počet subjektů; p ... počet proměnných

Regresní analýza v grafech

Grafy residuí modelů (příklady)



Obecné tvary residuí modelů (schéma)

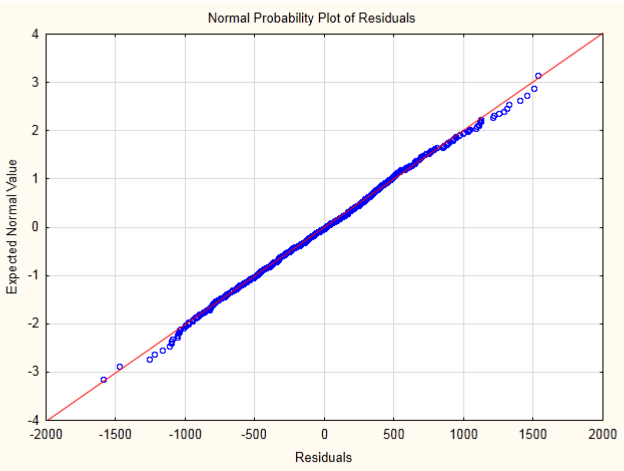


Lineární regrese – příklad I

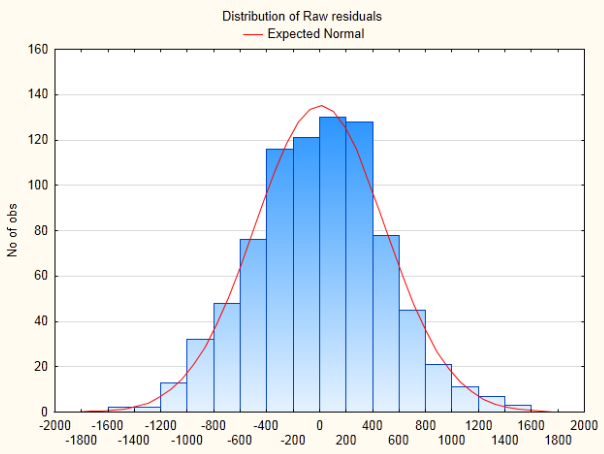
- **Příklad:** Proveďte regresní analýzu, v níž budete modelovat závislost objemu nucleus caudatus na věku.

	Regression Summary for Dependent Variable: Nucl_caud_volume (mm3) (Data_neuro_vycistena4) R= ,62657661 R2= ,39259825 Adjusted R2= ,39186732 F(1,831)=537,12 p<0,0000 Std.Error of estimate: 494,97					
N=833	b*	Std.Err. of b*	b	Std.Err. of b	t(831)	p-value
Intercept			8348,848	186,0558	44,8728	0,00
Age	-0,626577	0,027036	-57,369	2,4754	-23,1759	0,00

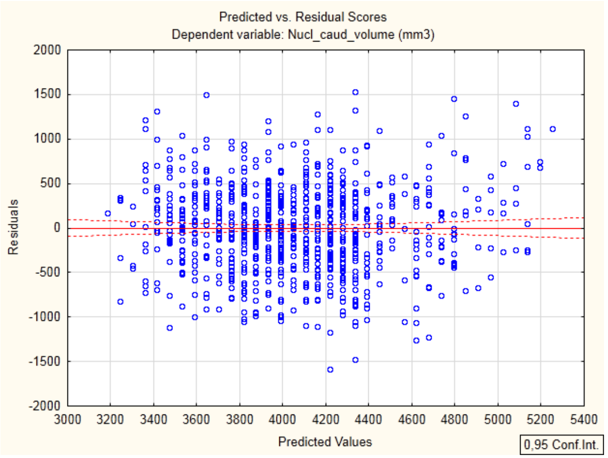
Q-Q graf reziduí



Histogram reziduí



Bodový graf reziduí vs. predikované hodnoty



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.

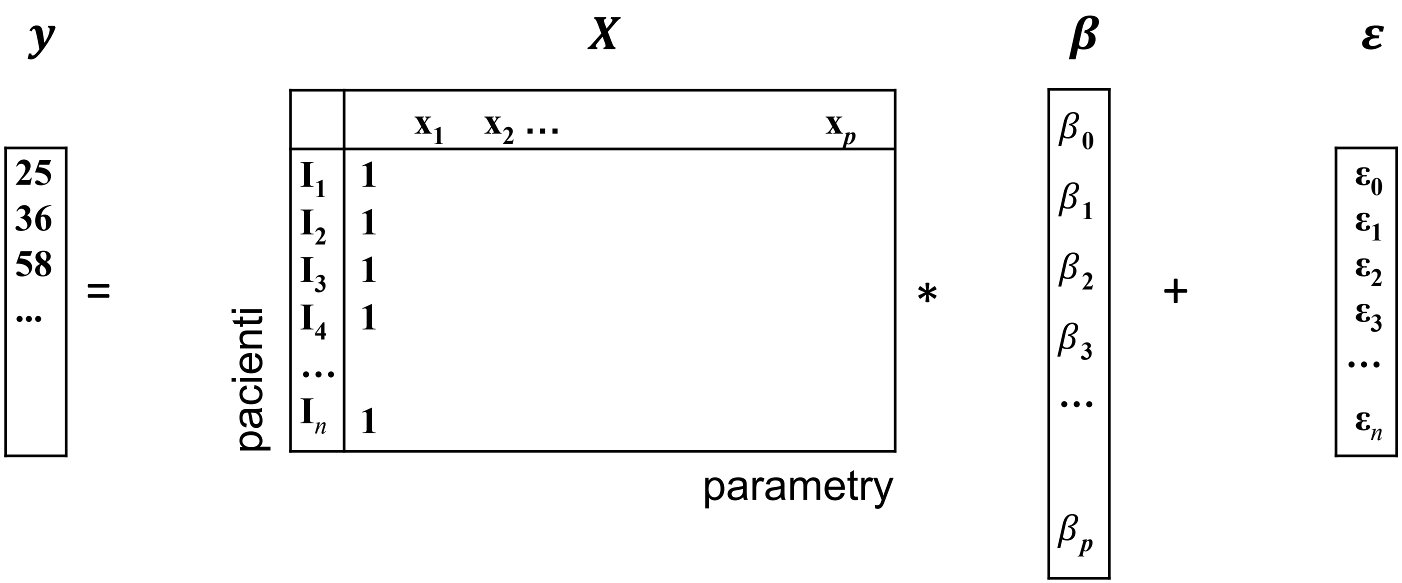
Data: Spreadsheet1 (8v by 833c)

Data_neuro_vycistena4

	1 Predicted	2 Residuals	3 StandardPredicted	4 StandardResidual	5 StdErrorPredicted	6 MahalanobisDistanc	7 DeletedResidual	8 CookDistance
1	3543.61	-15.89	-1.29	-0.03	31.38	2.35	-15.95	0.00
2	3967.80	-194.34	-0.22	-0.39	26.73	1.43	-194.91	0.00
3	3831.22	463.23	-0.56	0.94	24.44	1.03	464.36	0.00
4	3220.00	365.00	-2.10	0.74	44.91	5.85	368.03	0.00
5	4255.41	-532.15	0.50	-1.08	27.70	1.61	-533.82	0.00
6	4312.93	-343.56	0.65	-0.69	28.54	1.77	-344.71	0.00
7	3277.52	-391.29	-1.95	-0.79	42.93	5.26	-394.25	0.00
8	3392.57	348.66	-1.66	0.70	39.14	4.20	350.85	0.00
9	3507.61	229.79	-1.38	0.46	35.63	3.31	230.99	0.00
10	3335.05	-704.48	-1.81	-1.42	41.00	4.71	-709.34	0.00
11	3852.75	-960.02	-0.51	-1.94	27.92	1.65	-963.08	0.00
12	3543.61	7.62	-1.29	0.02	31.38	2.35	7.65	0.00
13	4348.93	285.38	0.74	0.58	25.97	1.29	286.17	0.00
14	3831.22	50.65	-0.56	0.10	24.44	1.03	50.77	0.00
15	4291.41	-837.95	0.59	-1.69	24.83	1.10	-840.07	0.00

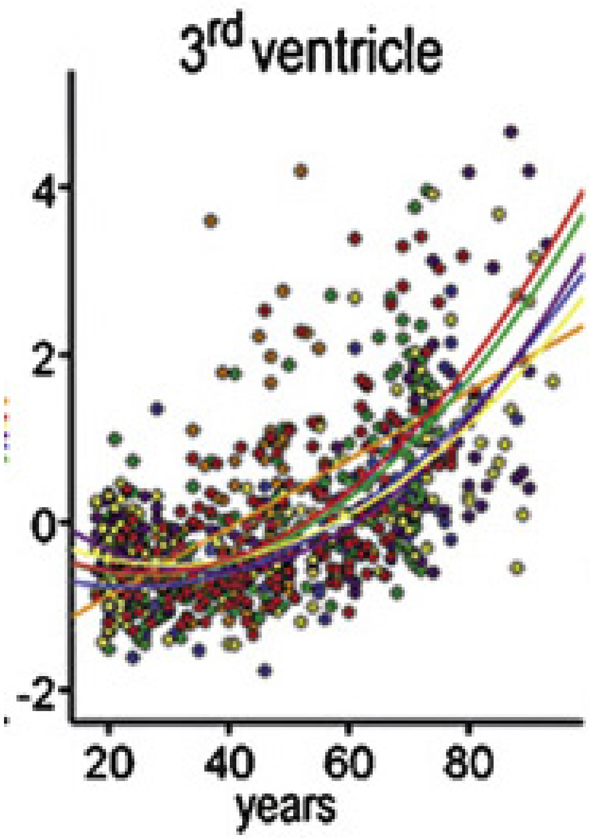
Vícenásobná lineární regrese

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



***X* – matice plánu (design matice)**

Kvadratická závislost objemu mozkové struktury na věku



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

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

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

Kategoriální data jako prediktory v regresi

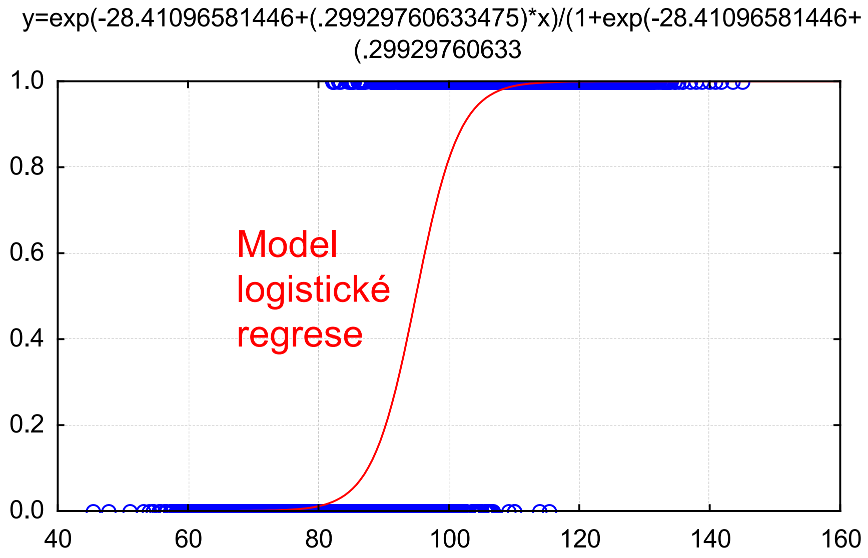
- Kategoriální a ordinální data mohou do analýzy vstupovat jako binární proměnné
- Kategoriální data (nelze seřadit) -> dummies
- Ordinální data (lze seřadit)
  - Dummies
  - Definice referenční kategorie (obvykle kategorie s nejnižším rizikem pro hodnocený endpoint

- 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 spojitých 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 spojité 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“

FOND ROZVOJE VYSOKÝCH ŠKOL