Analýza dat pro Neurovědy

RNDr. Eva Koriťáková, Ph.D.

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

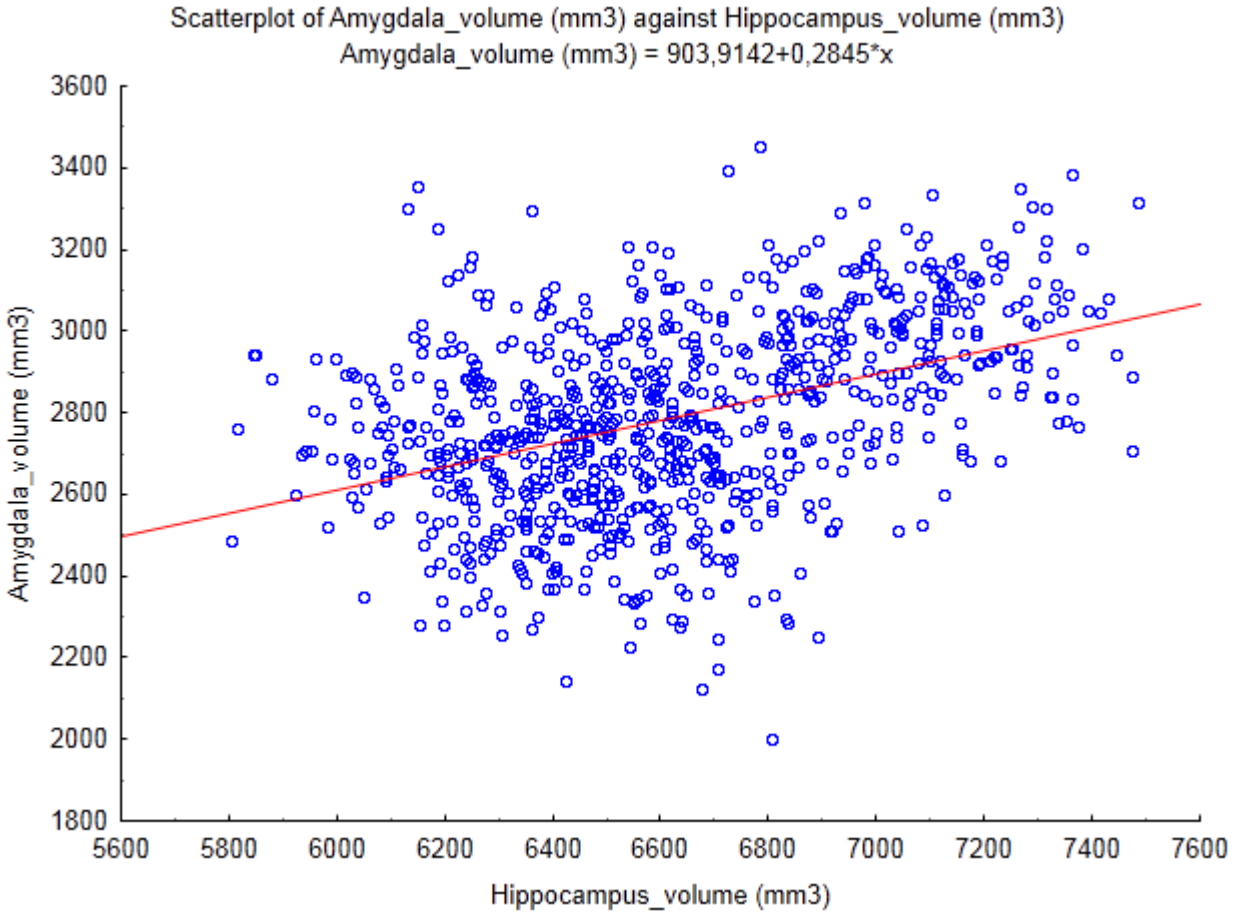
Motivace

- 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.

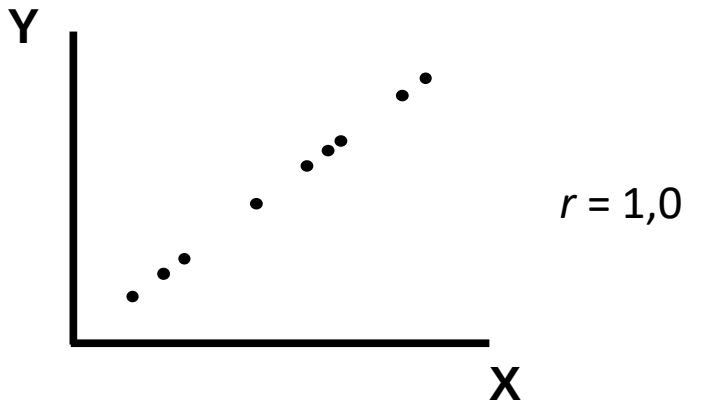
- Teď se chceme zabývat dvěma spojitými proměnnými:
  1. **Chceme zjistit, jestli mezi nimi existuje vztah** – např. jestli vyšší hodnoty jedné proměnné znamenají nižší hodnoty jiné proměnné.
  2. **Chceme kvantifikovat vztah mezi dvěma spojitými proměnnými** – např. pro použití jedné proměnné na místo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

Jak hodnotit vztah dvou spojitých proměnných?

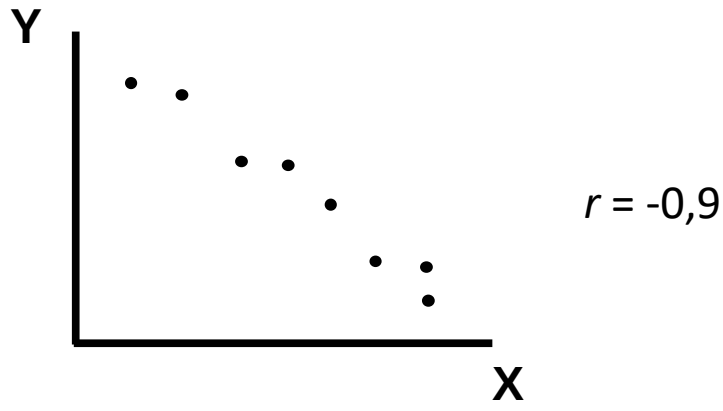
- Nejjednodušší formou je **bodový graf (x-y graf)**.
- Např. vztah objemu hipokampu a amygdaly:



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).

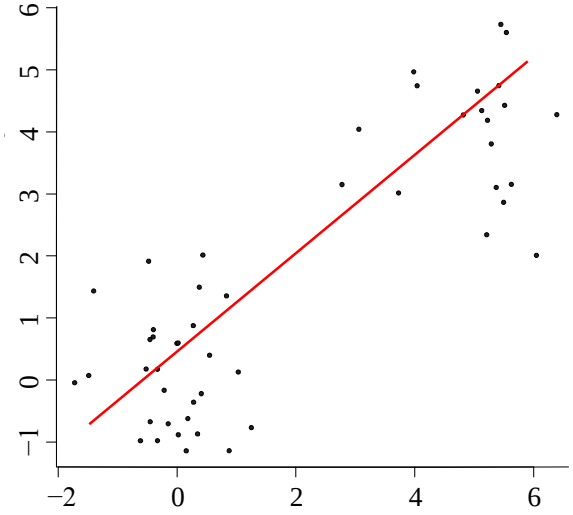
- Standardní metodou je výpočet **Pearsonova korelačního koeficientu (r)**:
  - Charakterizuje **linearitu** vztahu mezi X a Y – jinak řečeno variabilitu kolem lineárního trendu.
  - Nabývá hodnot od -1 do 1.
  - Hodnota r je kladná (kladná korelace), když vyšší hodnoty X souvisí s vyššími hodnotami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

Pearsonův korelační koef. – problematické situace I.

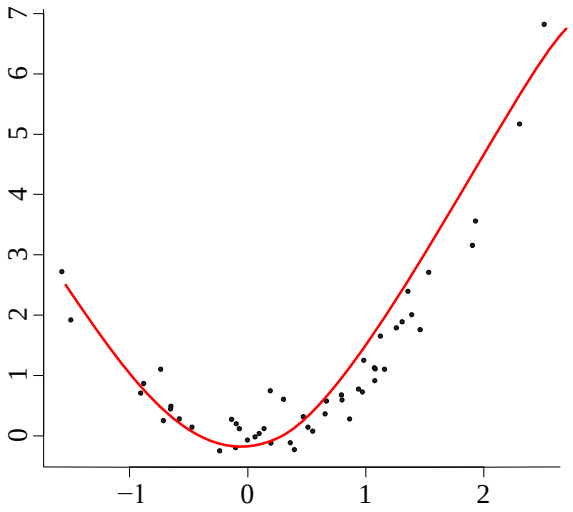
- Pearsonův korelační koeficient není vhodné počítat v situaci, kdy:
  - se v datech vyskytuje více skupin
  - proměnné mají nelineární vztah
  - se v datech vyskytují odlehlé hodnoty

Více skupin



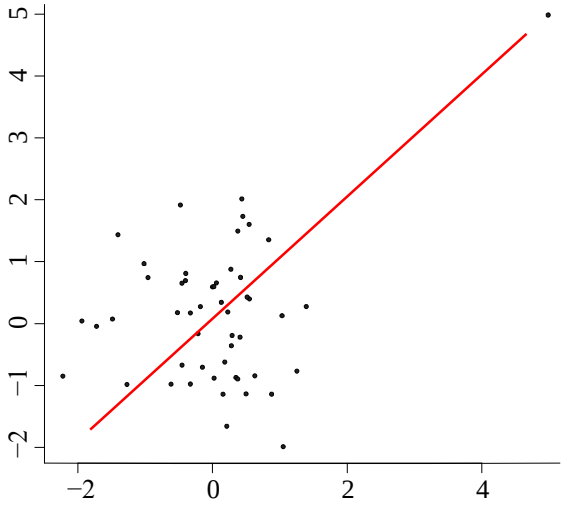
$r = 0{,}84$
($p < 0{,}001$)

Nelineární vztah



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

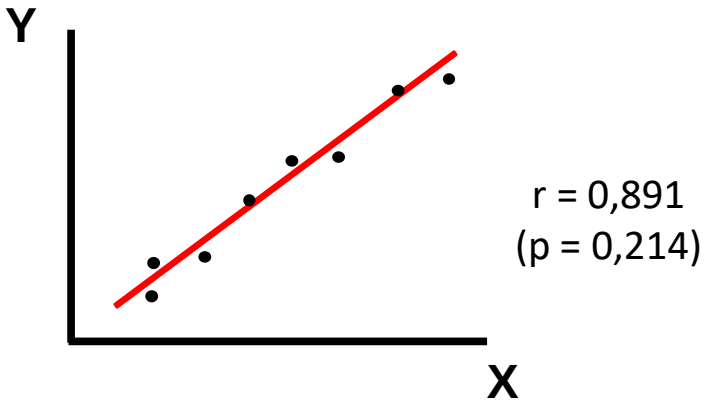
Odlehlá hodnota



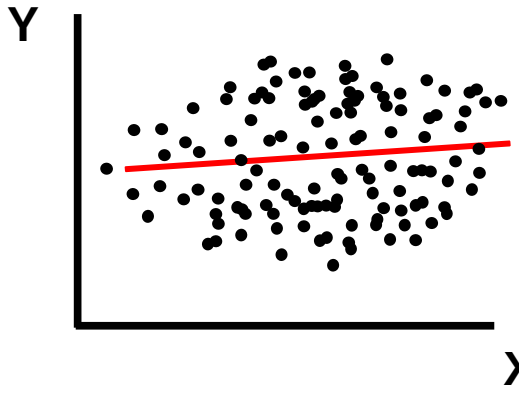
$r = 0{,}36$
($p = 0{,}009$)

Pearsonův korelační koef. – problematické situace II.

- Problém velikosti vzorku:



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

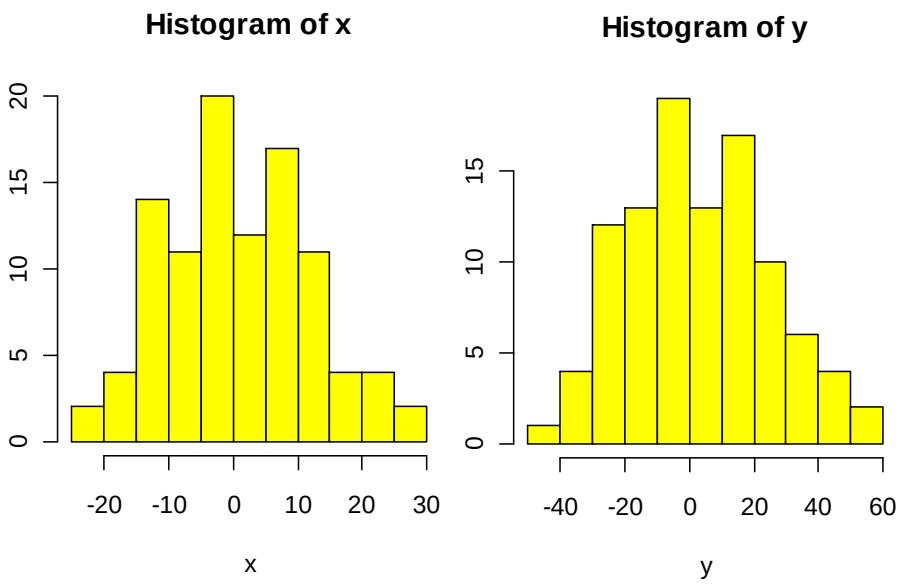


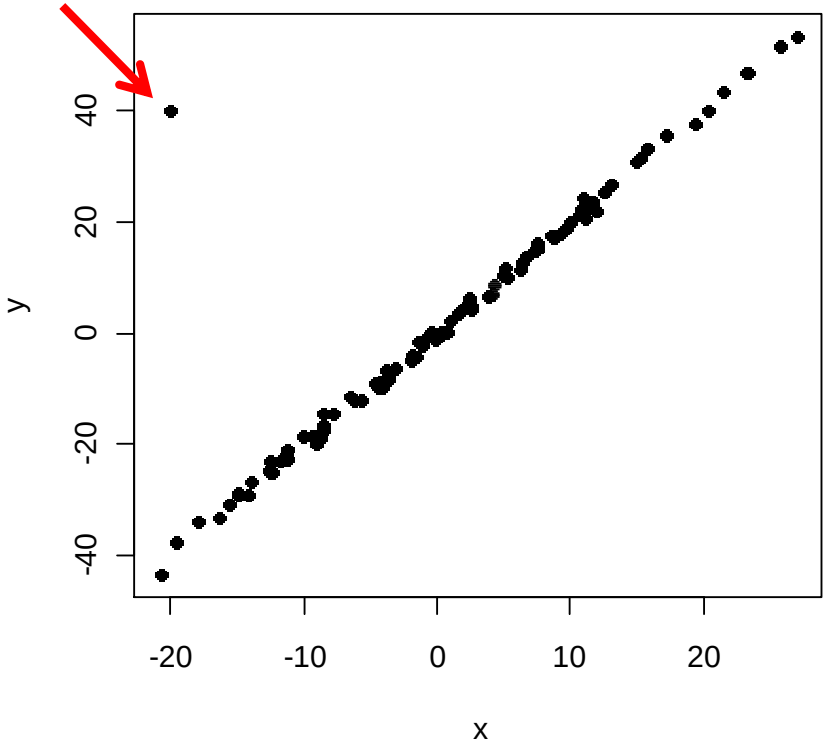
r = 0,012
(p = 0,008)

- 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!





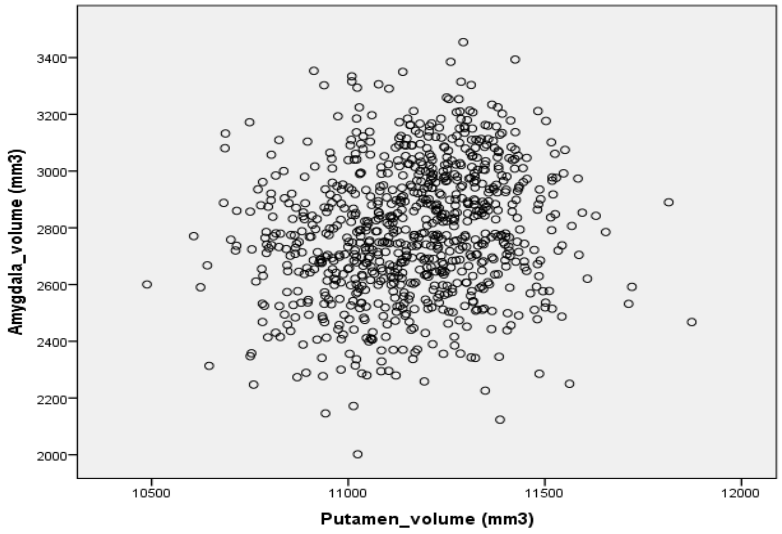
Pearsonův korelační koeficient

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

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

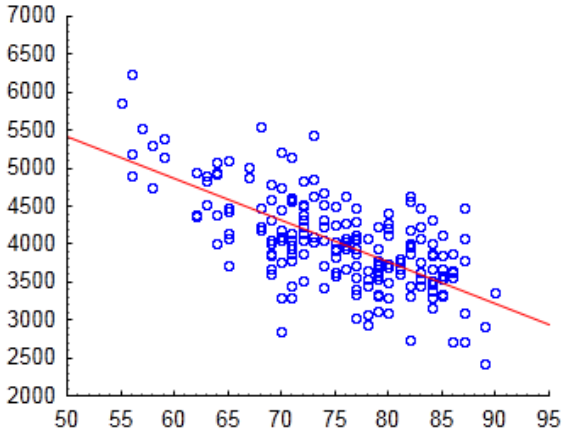
**. Correlation is significant at the 0.01 level (2-tailed).



Úkol 1.

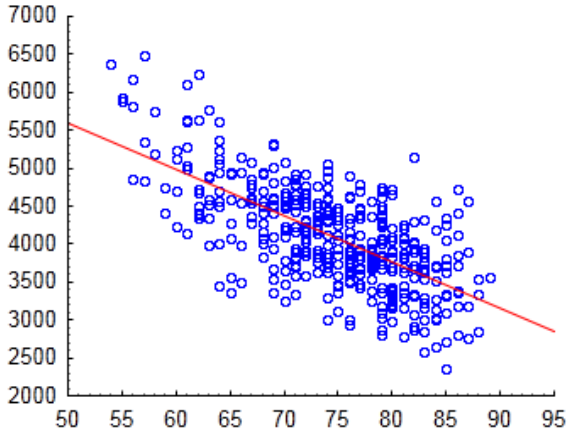
- **Zadání:** Ověřte, zda existuje vztah objemu nucleus caudatus a věku u pacientů s AD, pacientů s MCI a u kontrol. Nezapomeňte ověřit normalitu srovnávaných proměnných.
- **Řešení:**

AD



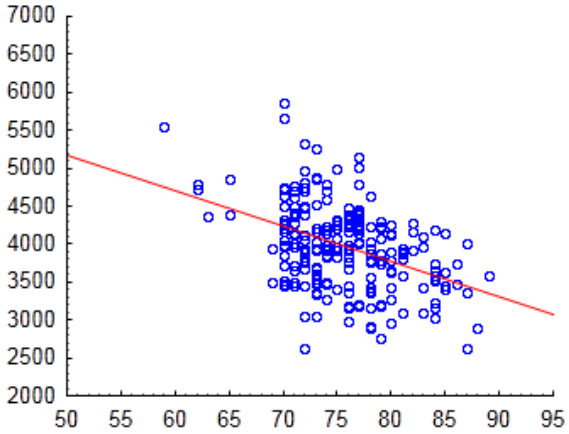
r = -0,68
(p < 0,001)

MCI



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

CN



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

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.

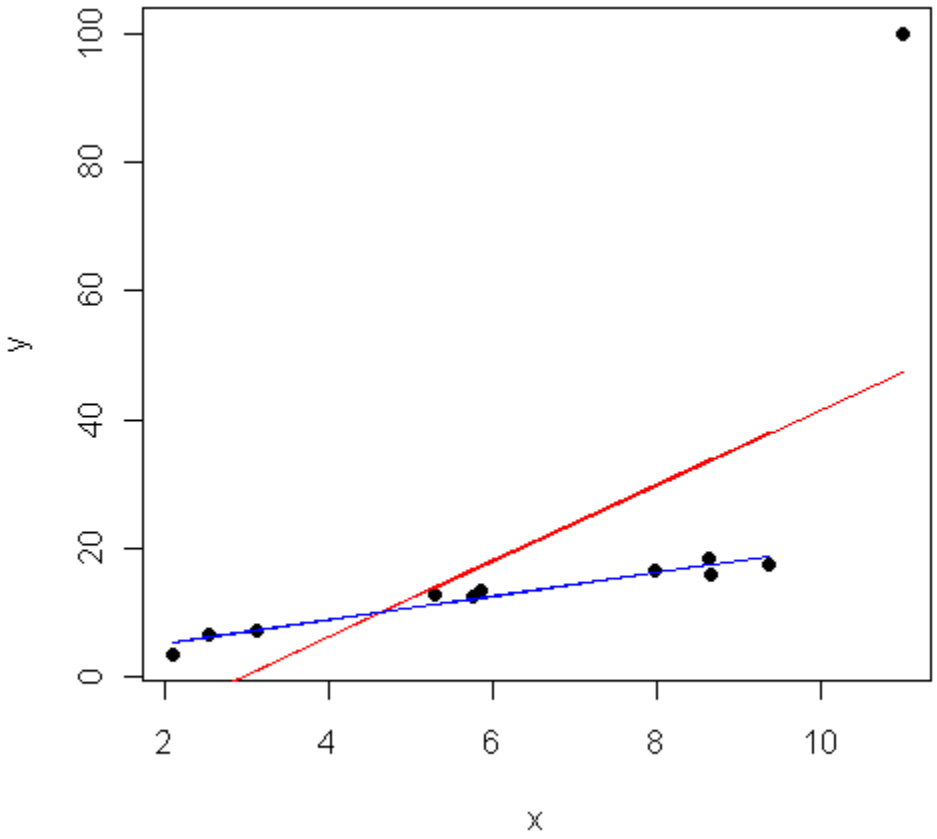
Spearmanův korelační koeficient (r_s)

- 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:

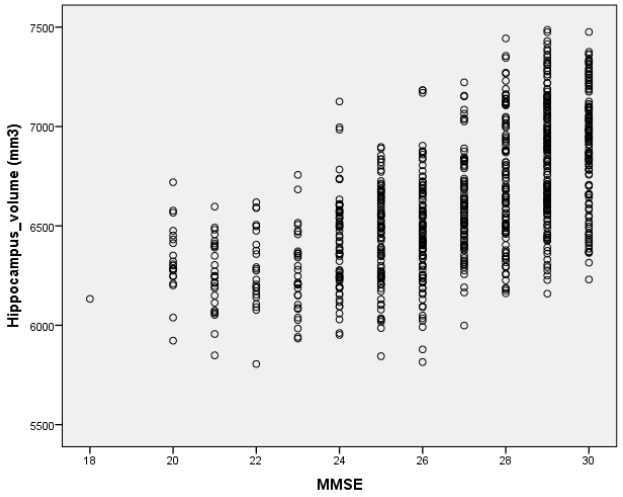
$r_S = 0{,}95$

$(p < 0{,}001)$

Spearmanův korelační koeficient není náchylný k odlehlým hodnotám.

Spearmanův korelační koeficient

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



Correlations

			MMSE	Hippocampus_volume (mm3)
Spearman's rho	MMSE	Correlation Coefficient	1,000	,627**
		Sig. (2-tailed)	.	,000
		N	833	833
	Hippocampus_volume (mm3)	Correlation Coefficient	,627**	1,000
		Sig. (2-tailed)	,000	.
		N	833	833

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

Úkol 2.

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

Correlations

			MMSE	Amygdala_volume (mm3)	Thalamus_volume (mm3)	Pallidum_volume (mm3)	Putamen_volume (mm3)	Nucl_caud_volume (mm3)
Spearman's rho	MMSE	Correlation Coefficient	1,000	,339**	-,001	,039	,325**	,012
		Sig. (2-tailed)	.	,000	,983	,259	,000	,733
		N	833	833	833	833	833	833
	Amygdala_volume (mm3)	Correlation Coefficient	,339**	1,000	,019	,066	,198**	-,064
		Sig. (2-tailed)	,000	.	,582	,056	,000	,065
		N	833	833	833	833	833	833
	Thalamus_volume (mm3)	Correlation Coefficient	-,001	,019	1,000	-,038	-,002	,007
		Sig. (2-tailed)	,983	,582	.	,274	,949	,836
		N	833	833	833	833	833	833
	Pallidum_volume (mm3)	Correlation Coefficient	,039	,066	-,038	1,000	-,001	-,025
		Sig. (2-tailed)	,259	,056	,274	.	,969	,464
		N	833	833	833	833	833	833
	Putamen_volume (mm3)	Correlation Coefficient	,325**	,198**	-,002	-,001	1,000	-,069*
		Sig. (2-tailed)	,000	,000	,949	,969	.	,046
		N	833	833	833	833	833	833
	Nucl_caud_volume (mm3)	Correlation Coefficient	,012	-,064	,007	-,025	-,069*	1,000
		Sig. (2-tailed)	,733	,065	,836	,464	,046	.
		N	833	833	833	833	833	833

2. Základy regresní analýzy

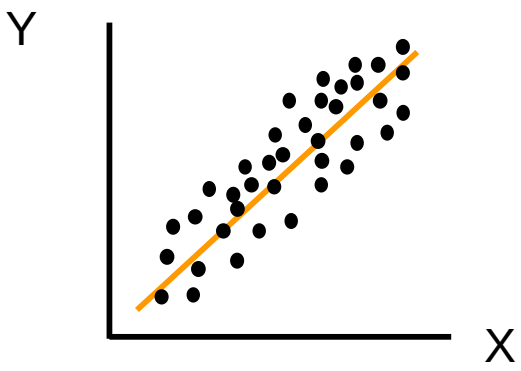
Motivace

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

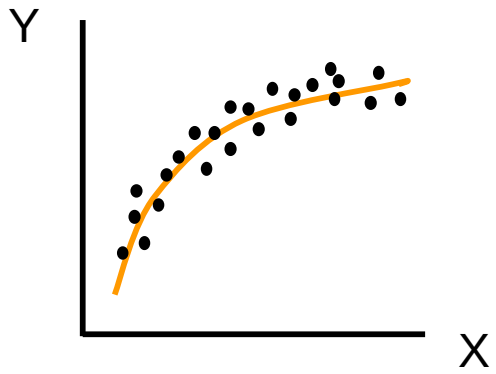
- Dva problémy:
  - Vybrat správnou funkci k popisu dané závislosti.
  - Stanovit konkrétní parametry daného typu funkce.

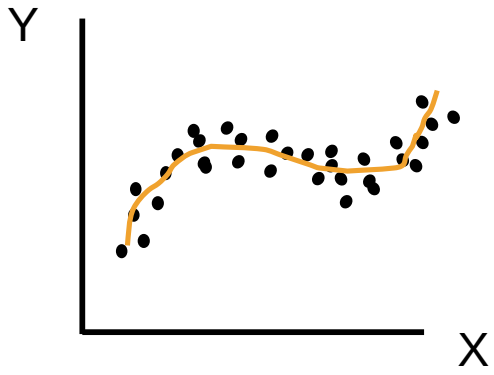
Příklady závislostí

Lineární



Nelineární

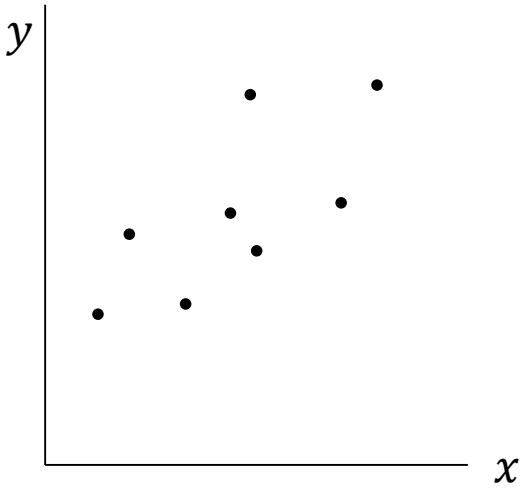




Lineární regrese

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)

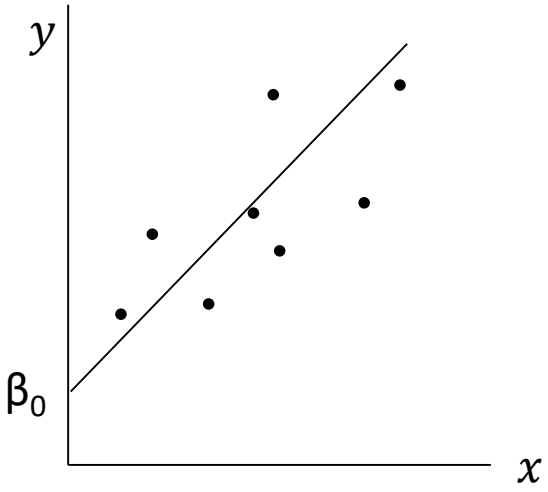
ε – náhodná složka modelu přímky (rezidua přímky)

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)

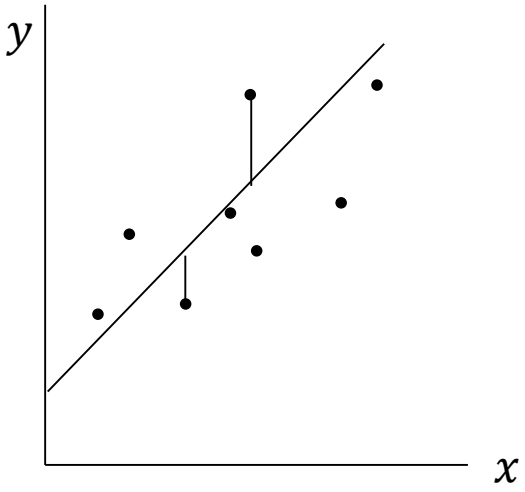
ε – náhodná složka modelu přímky (rezidua přímky)

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)

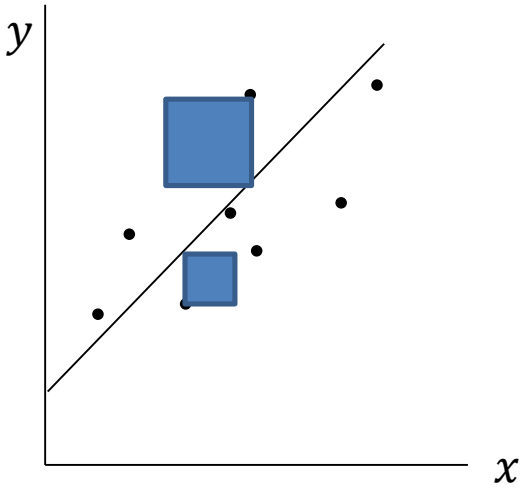
ε – náhodná složka modelu přímky (rezidua přímky)

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: $\boldsymbol{y} = \beta_0 + \beta_1 * \boldsymbol{x} + \boldsymbol{\varepsilon}$



Odhad koeficientů β metodou nejmenších čtverců:

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

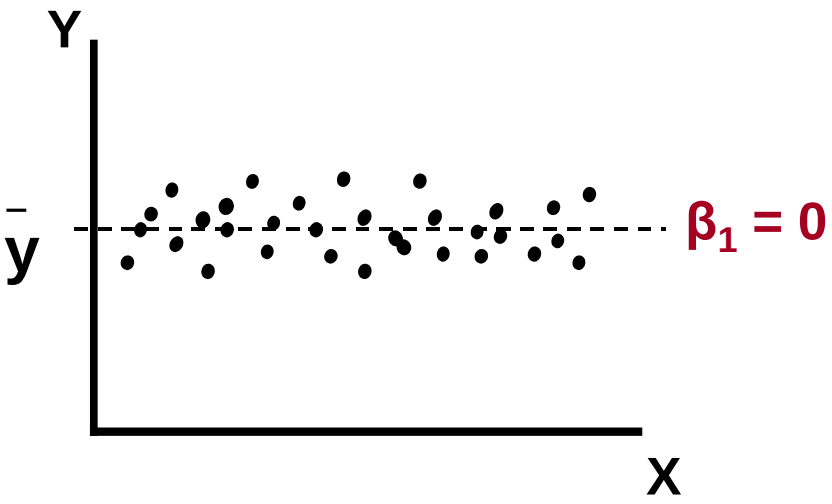
y – závisle proměnná (vysvětlovaná proměnná)

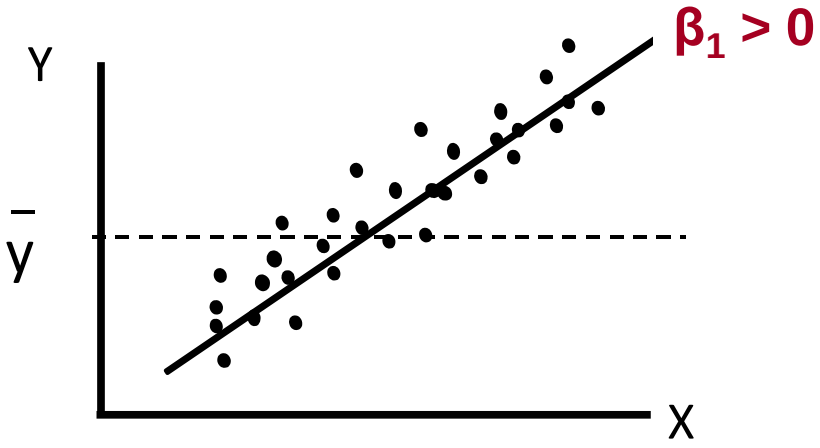
x – nezávisle proměnná (vysvětlující proměnná, regresor)

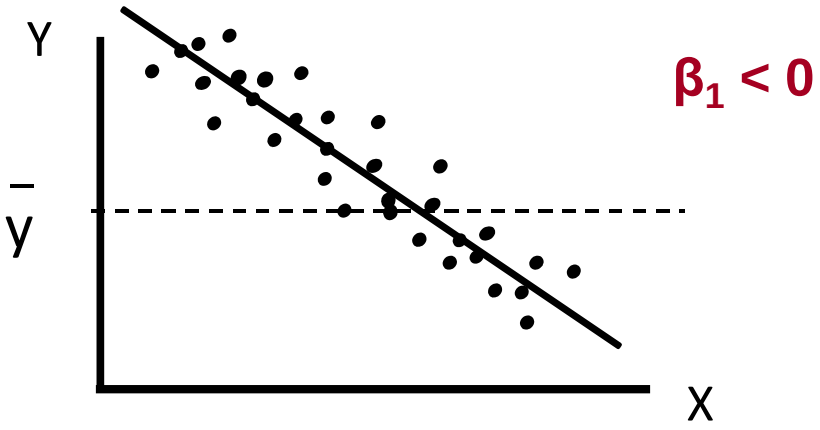
ε – náhodná složka modelu přímky (rezidua přímky)

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

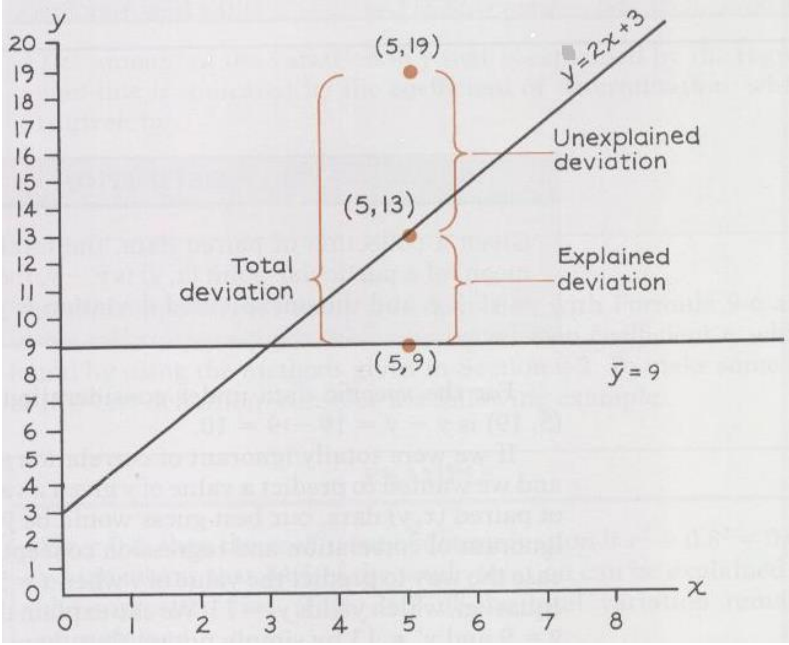
Lineární regrese - příklady







Lineární regrese



Převzato z přednášek
RNDr. Marie Budíkové, Dr.

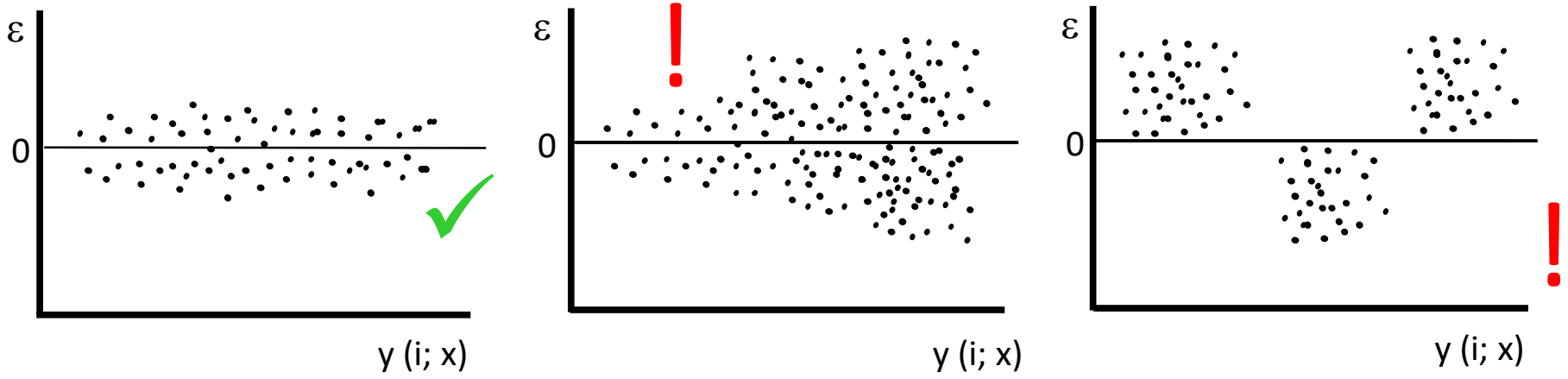
Testování významnosti modelu jako celku – celkový F-test:

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	-	-

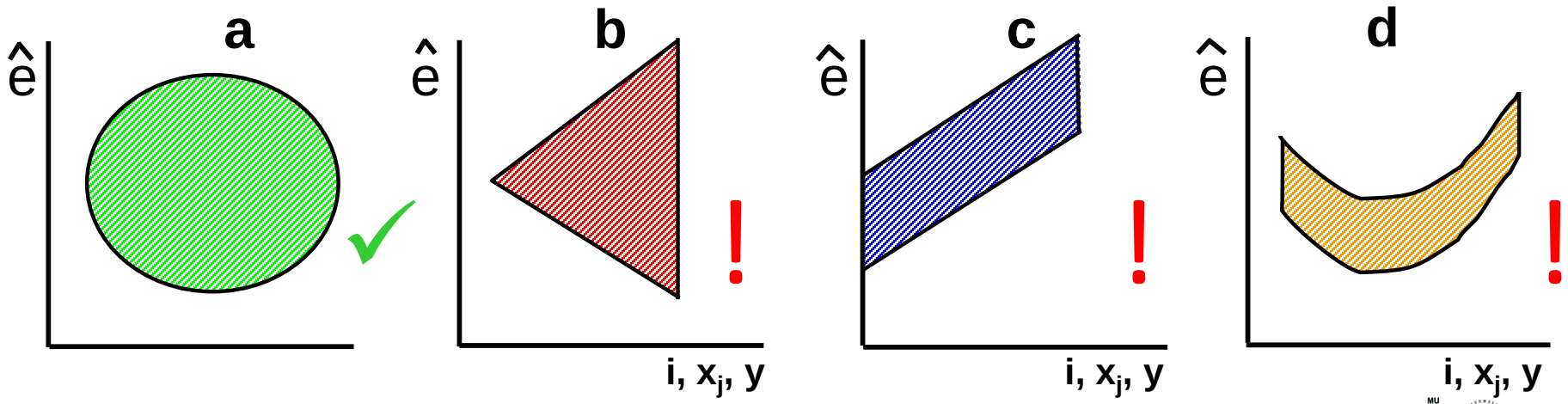
n ... počet subjektů; p ... počet proměnných

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

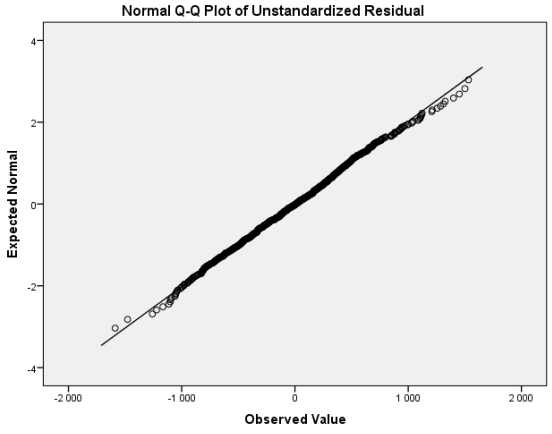
a. Predictors: (Constant), Age

Coefficients[a]

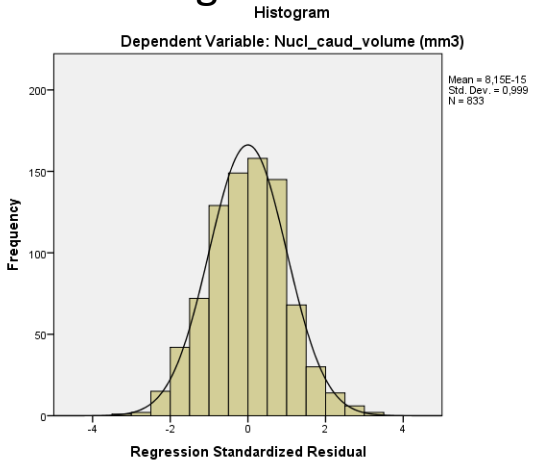
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8348,848	186,056		44,873	,000
	Age	-57,369	2,475	-,627	-23,176	,000

a. Dependent Variable: Nucl_caud_volume (mm3)

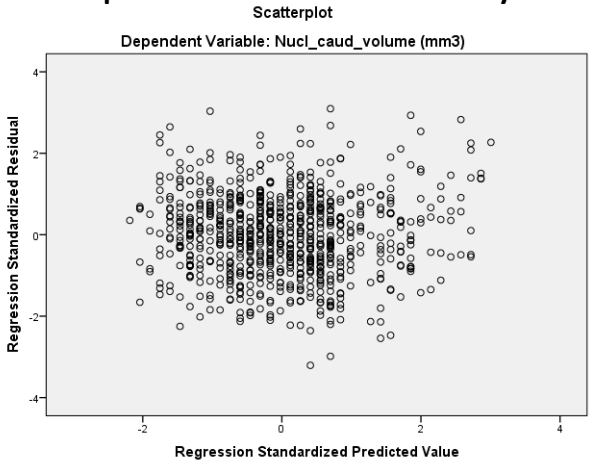
Q-Q graf reziduí



Histogram reziduí



Bodový graf reziduí vs. predikované hodnoty



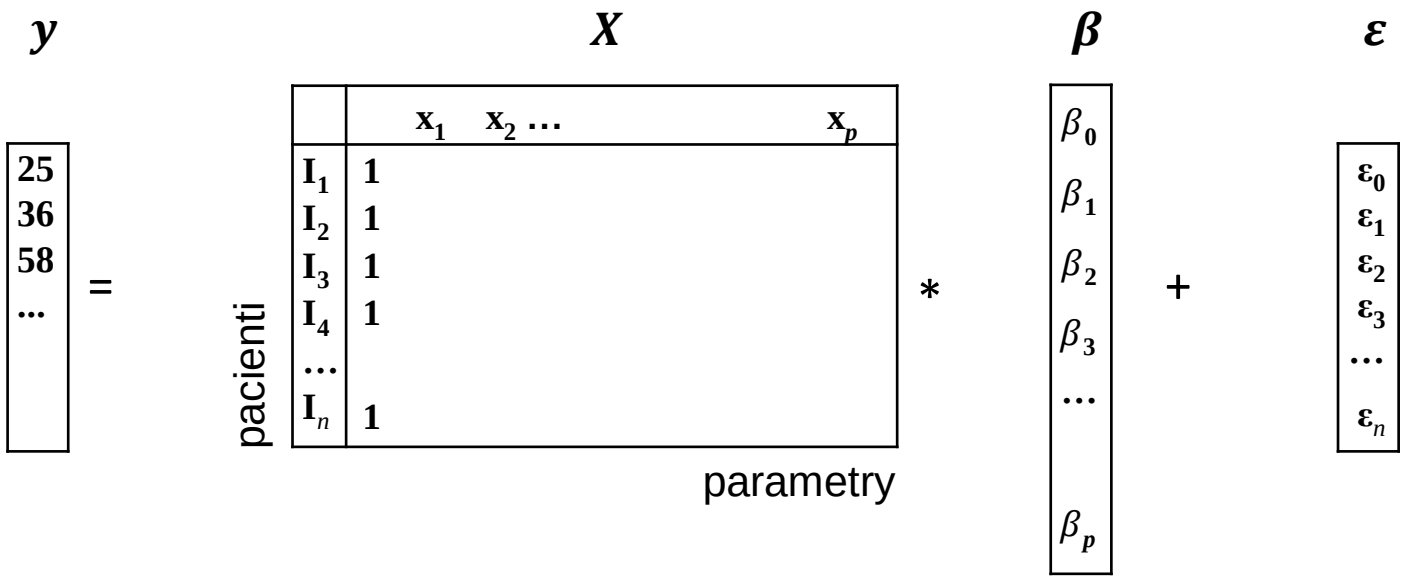
Lineární regrese – příklad II

- **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

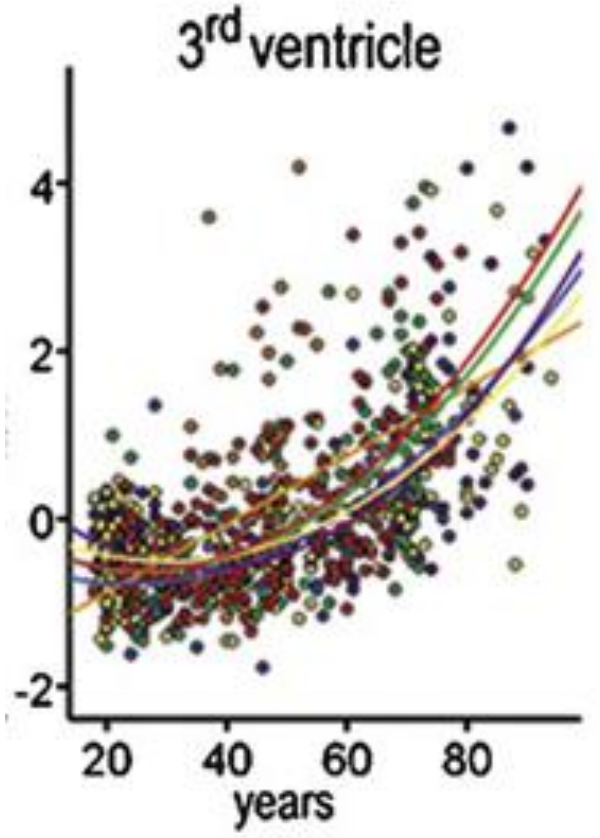
Vícenásobná lineární regrese

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



X – matice plánu (design matice)

Kvadratická závislost objemu mozkové struktury na věku



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

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

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

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

parametry

Kategoriální data jako prediktory v regresi

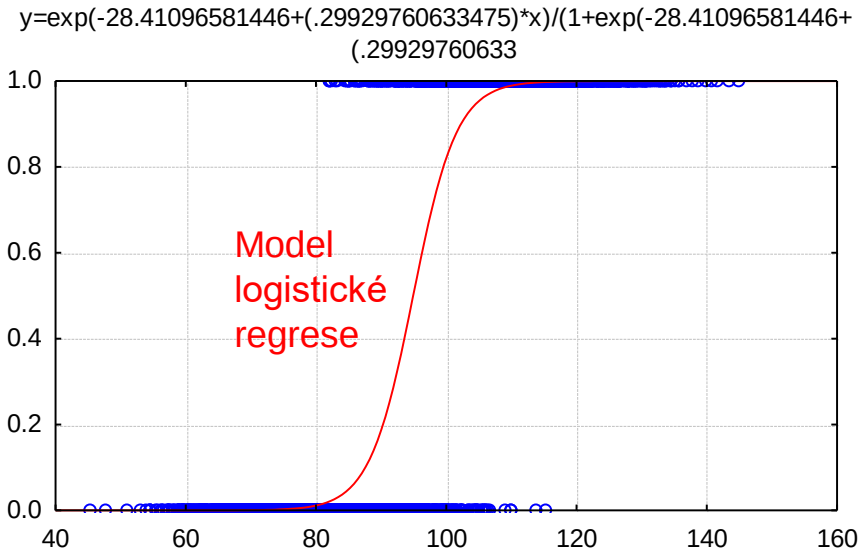
- Kategoriální a ordinální data mohou do analýzy vstupovat jako binární proměnné
- Kategoriální data (nelze seřadit) -> dummies
- Ordinální data (lze seřadit)
  - Dummies
  - Definice referenční kategorie (obvykle kategorie s nejnižším rizikem pro hodnocený endpoint

- 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 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“