Bi7491 Regresní modelování

Lineární regresní model I

Definice a zadání

Co byste po dnešní hodině měli vědět a umět?

- Vědět, jak se definuje lineární regresní model
- Vysvětlit předpoklady regresních modelů
- Umět použít v lineárním regresním modelu různé typy prediktorů
- Vědět, co je multikolinearita, jak ji zjistit a jak se s ní vypořádat

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Definice lineárního regresního modelu

RESEARCH

Vitamin D supplementation to prevent acute respiratory tract infections: systematic review and meta-analysis of individual participant data

Adrian R Martineau,[1,2] David A Jolliffe,[1] Richard L Hooper,[1] Lauren Greenberg,[1] John F Aloia,[3] Peter Bergman,[4] Gal Dubnov-Raz,[5] Susanna Esposito,[6] Davaasambuu Ganmaa,[7] Adit A Ginde,[8] Emma C Goodall,[9] Cameron C Grant,[10] Christopher J Griffiths,[1,2,11] Wim Janssens,[12] Ilkka Laaksi,[13] Semira Manaseki-Holland,[14] David Mauger,[15] David R Murdoch,[16] Rachel Neale,[17] Judy R Rees,[18] Steve Simpson,Jr[19] Iwona Stelmach,[20] Geeta Trilok Kumar,[21] Mitsuyoshi Urashima,[22] Carlos A Camargo Jr[23]

WHAT IS ALREADY KNOWN ON THIS TOPIC

Randomised controlled trials of vitamin D supplementation for the prevention of acute respiratory tract infection have yielded conflicting results

Individual participant data (IPD) meta-analysis has the potential to identify factors that may explain this heterogeneity, but this has not previously been performed

WHAT THIS STUDY ADDS

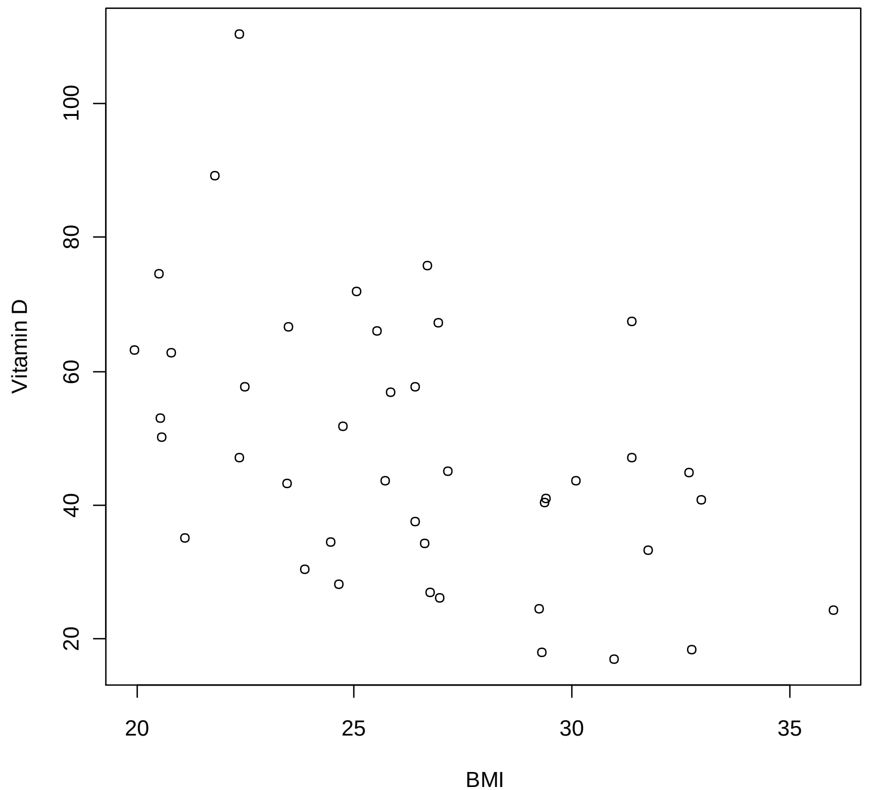
Meta-analysis of IPD from 10 933 participants in 25 randomised controlled trials showed an overall protective effect of vitamin D supplementation against acute respiratory tract infection (number needed to treat (NNT)=33)

Benefit was greater in those receiving daily or weekly vitamin D without additional bolus doses (NNT=20), and the protective effects against acute respiratory tract infection in this group were strongest in those with profound vitamin D deficiency at baseline (NNT=4)

These findings support the introduction of public health measures such as food fortification to improve vitamin D status, particularly in settings where profound vitamin D deficiency is common

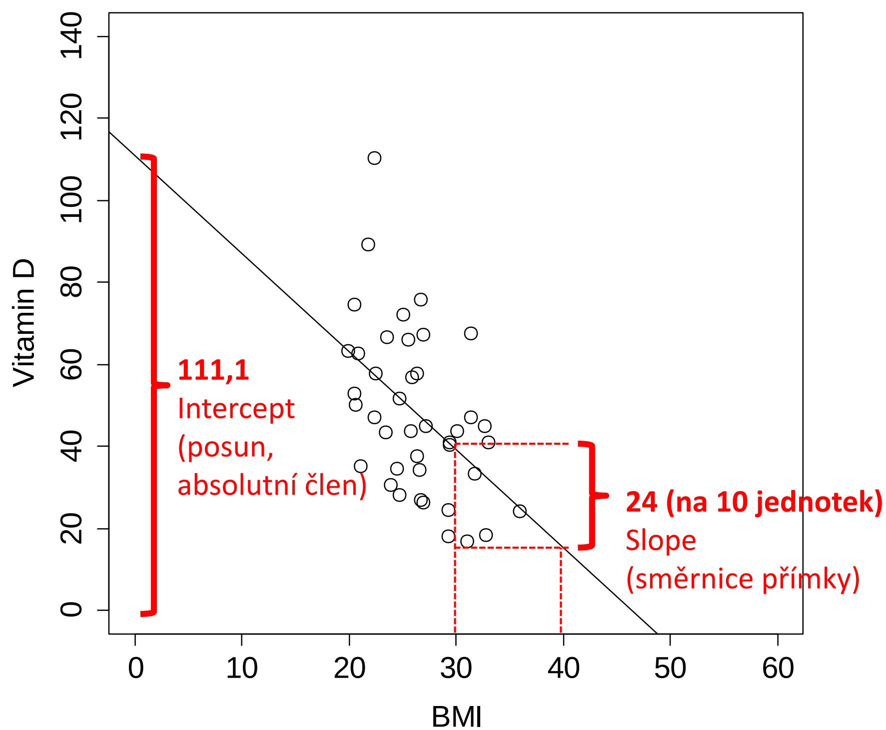
thebmj | *BMJ* 2017;356:i6583 | doi: 10.1136/bmj.i6583

Jak popsat vztah mezi dvěma kvantitativními proměnnými?



intuitivně jsme schopni nakreslit přímku vedoucí mezi pozorováními...

Jak popsat vztah mezi dvěma kvantitativními proměnnými?



Metoda nejmenších čtverců – minimalizuje vzdálenosti přímky od bodů

koncentrace vitaminu D = 111,1 – 2,4BMI

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Model s jednou spojitou proměnnou

absolutní člen, posun

směrnice (sklon) regresní přímky

$$Y_i \approx \beta_0 + \beta_1 x_i$$

$$i = 1, \ldots, n$$

počet pozorování

proč tady není = ??

IBA MUNI MED
Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Lineární regresní model

Stochastická složka

Rezidua

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$i = 1,\ldots,n$$

Pro rezidua musí platit:

1. jsou **nesystematické** $E\varepsilon_i = 0$
2. jsou **homogenní v rozptylu** $D\varepsilon_i = \sigma^2 > 0$
3. jsou **nekorelované** $C(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

Vícenásobná (*víceprediktorová*) regrese

lze zapojit více vysvětlujících proměnných

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

$$i = 1, \ldots, n$$

lze zapsat jako vztah pro střední hodnotu (a vynechat rezidua)

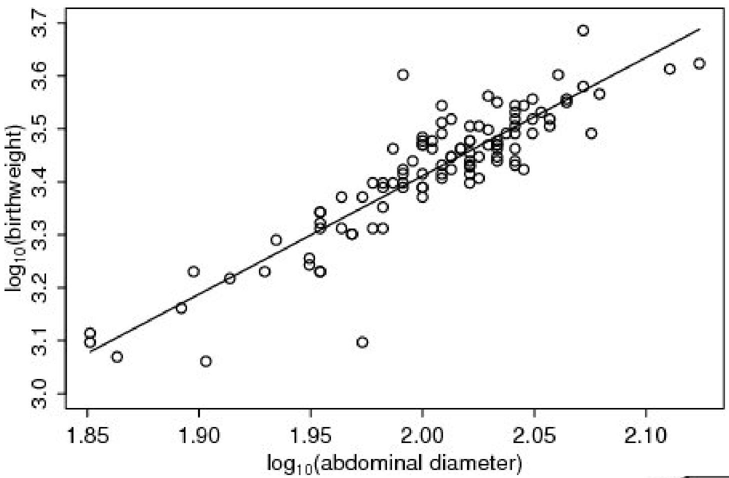
$$EY_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

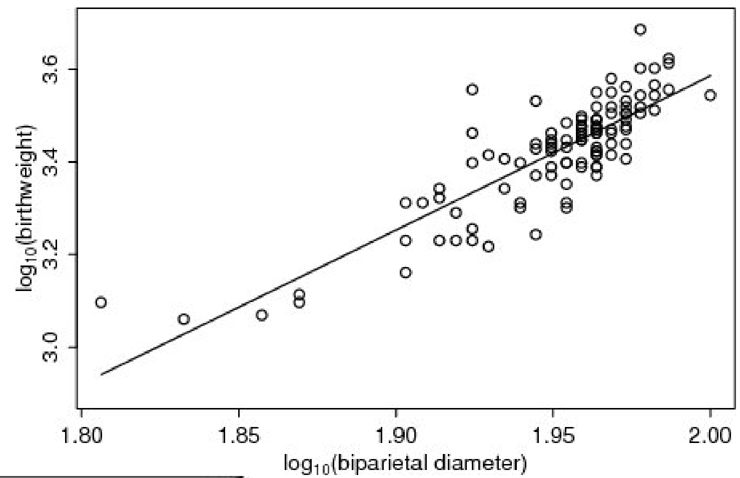
$$i = 1, \ldots, n$$

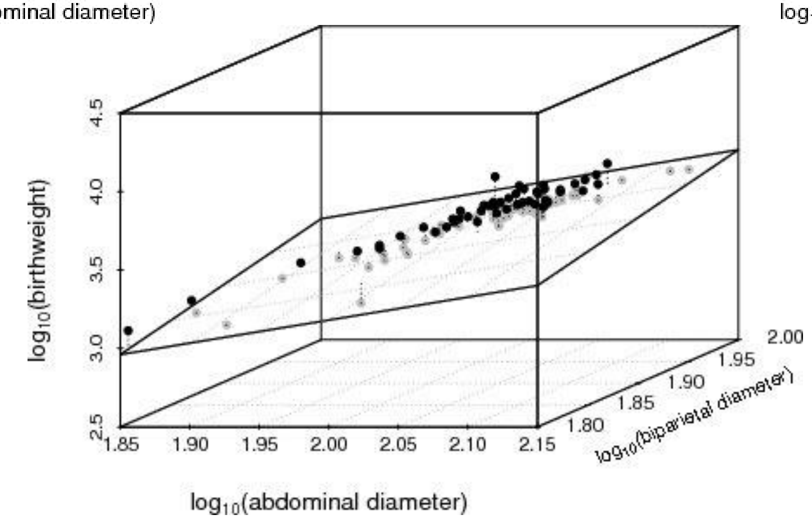
kde $ß_j$ jsou neznámé **parametry** (j = 0,...,p)
počet **prediktorů** je p
počet **parametrů** je p+1=k
počet **pozorování** je n.

Víceprediktorová regrese

Model pro predikci porodní hmotnosti dle UZ markerů







IBA MUNI MED
Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Rozepsané...

jednotlivá pozorování

$$\begin{cases} EY_1 = \beta_0 + \beta_1 x_{11} + \ldots + \beta_p x_{1p} \\ EY_2 = \beta_0 + \beta_1 x_{21} + \ldots + \beta_p x_{2p} \\ \vdots \\ EY_n = \beta_0 + \beta_1 x_{n1} + \ldots + \beta_p x_{np} \end{cases}$$

IBA MUNI MED Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Maticový zápis

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\text{závisle proměnná}} = \underbrace{\underbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}_{\text{matice plánu}} \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}_{\text{regresní koeficienty}}}_{\text{systematická složka}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\text{náhodná složka}}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Normální lineární regresní model

- <u>Náhodná složka modelu</u> je reprezentována náhodnými chybami ε_i
  Rozdělení těchto náhodných veličin ε_i je **normální**
- Rozptyl je všude stejný, pozorování jsou nezávislá

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, i = 1, \ldots, n$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

- předpokladem sestrojení statistik pro testy v tomto modelu

IBA MUNI MED
Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Odhad neznámých parametrů
Parametry β

- Odhad metodou nejmenších čtverců

$$\hat{\boldsymbol{\beta}}_{\mathbf{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- nejlepší, nestranný, lineární odhad $\boldsymbol{\beta}$ (BLUE)
- lze ukázat, že rozptyl tohoto odhadu je

$$D\hat{\boldsymbol{\beta}}_{\mathbf{OLS}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Odhad neznámých parametrů
Reziduální součet čtverců

$$S_e = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS})$$
$$= (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$$
$$= (Y_1 - \hat{Y}_1)^2 + \ldots + (Y_n - \hat{Y}_n)^2$$

$$\Big\downarrow \text{ lze ukázat}$$

$$S_e = \mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'_{OLS}\mathbf{X}'\mathbf{Y}$$

Odhad neznámých parametrů

Rozptyl σ^2

$$s^2 = \frac{S_e}{n-k}$$

reziduální součet čtverců

stupně volnosti modelu

Statistické testy v lineárním regresním modelu

- **Testování lineární kombinace parametrů**
- Hypotéza:

$$H_0: \mathbf{c}'\boldsymbol{\beta} = x \qquad x \text{ ... konstanta}$$
$$H_1: \mathbf{c}'\boldsymbol{\beta} \neq x$$

- Testová statistika:

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}_{\mathbf{OLS}} - \mathbf{c}'\boldsymbol{\beta}}{s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \sim t(n-k)$$

- Speciálním případem je klasický t-test

Statistické testy v lineárním regresním modelu

- **Testování více parametrů zároveň**
- Předpokládá blokové označení parametrů:

$$\boldsymbol{\beta} = (\underbrace{\beta_1,\ldots,\beta_m}_{=\boldsymbol{\beta}_1'}, \underbrace{\beta_{m+1},\ldots,\beta_k}_{=\boldsymbol{\beta}_2'})'$$

- Obdobně i pro odhad

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \qquad \hat{\boldsymbol{\beta}}_{OLS} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{OLS,1} \\ \hat{\boldsymbol{\beta}}_{OLS,2} \end{pmatrix} \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$$

Statistické testy
v lineárním regresním modelu

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \qquad \hat{\boldsymbol{\beta}}_{OLS} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{OLS,1} \\ \hat{\boldsymbol{\beta}}_{OLS,2} \end{pmatrix} \qquad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$$

- Hypotéza $H_0\colon \boldsymbol{\beta}_2 = \boldsymbol{x}$ $\quad$ $\boldsymbol{x}$ ... konstantní vektor
  $H_1\colon \boldsymbol{\beta}_2 \neq \boldsymbol{x}$
- Statistikou je

$$F = \frac{1}{s^2(k-m)}(\hat{\boldsymbol{\beta}}_{\mathbf{OLS,2}} - \boldsymbol{\beta}_2)'\mathbf{V}_{\mathbf{22}}^{-1}(\hat{\boldsymbol{\beta}}_{\mathbf{OLS,2}} - \boldsymbol{\beta}_2) \sim F(k-m, n-k)$$

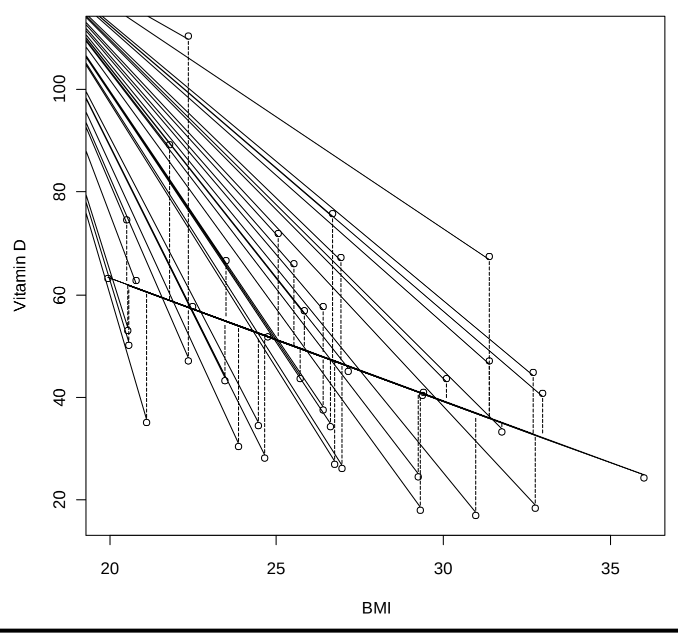
- Speciálním případem je klasická analýza rozptylu

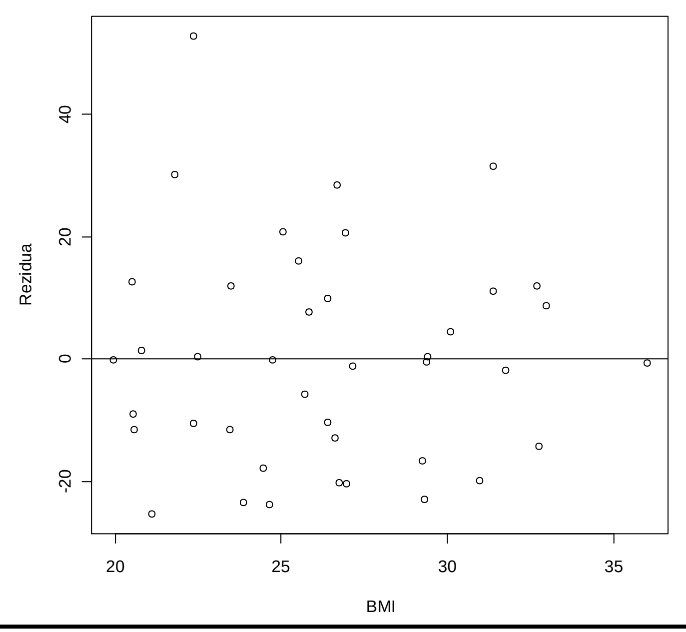
Analýza reziduí

- V lineárním modelu jsou rezidua rozdíly mezi pozorovanými a odhadnutými (očekávanými) hodnotami závisle proměnné:

$$\mathbf{r} = \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$$

- Hodnocení reziduí je nesmírně důležité pro posouzení splnění předpokladů modelu





Koeficient determinace

Celková variabilita výsledku:

$$S_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Reziduální součet čtverců:

$$S_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Nevyčerpaná variabilita

Koeficient determinace = vyčerpaná variabilita výsledku modelem

$$R^2 = 1 - \frac{S_e}{S_T}$$

Lineární regresní model I

Prediktory různých datových typů

Matice plánu

- představuje matici nezávislých proměnných - prediktorů

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}^{\text{závisle proměnná}} = \overbrace{\underbrace{\begin{pmatrix} 1\, x_{11} & \cdots & x_{1p} \\ \vdots \;\; \vdots & & \vdots \\ 1\, x_{n1} & \cdots & x_{np} \end{pmatrix}}_{\text{matice plánu}} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}}^{\text{systematická složka}} + \overbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}^{\text{náhodná složka}}$$

$$EY_1 = \beta_0 + \beta_1 x_{11} + \ldots + \beta_p x_{1p}$$
$$EY_2 = \beta_0 + \beta_1 x_{21} + \ldots + \beta_p x_{2p}$$
$$\vdots$$
$$EY_n = \beta_0 + \beta_1 x_{n1} + \ldots + \beta_p x_{np}$$

Matice plánu

- představuje matici nezávislých proměnných – prediktorů
- promítají se do ní
  - konstanta – absolutní člen
  - spojité proměnné
  - kategoriální proměnné

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Konstanta – absolutní člen

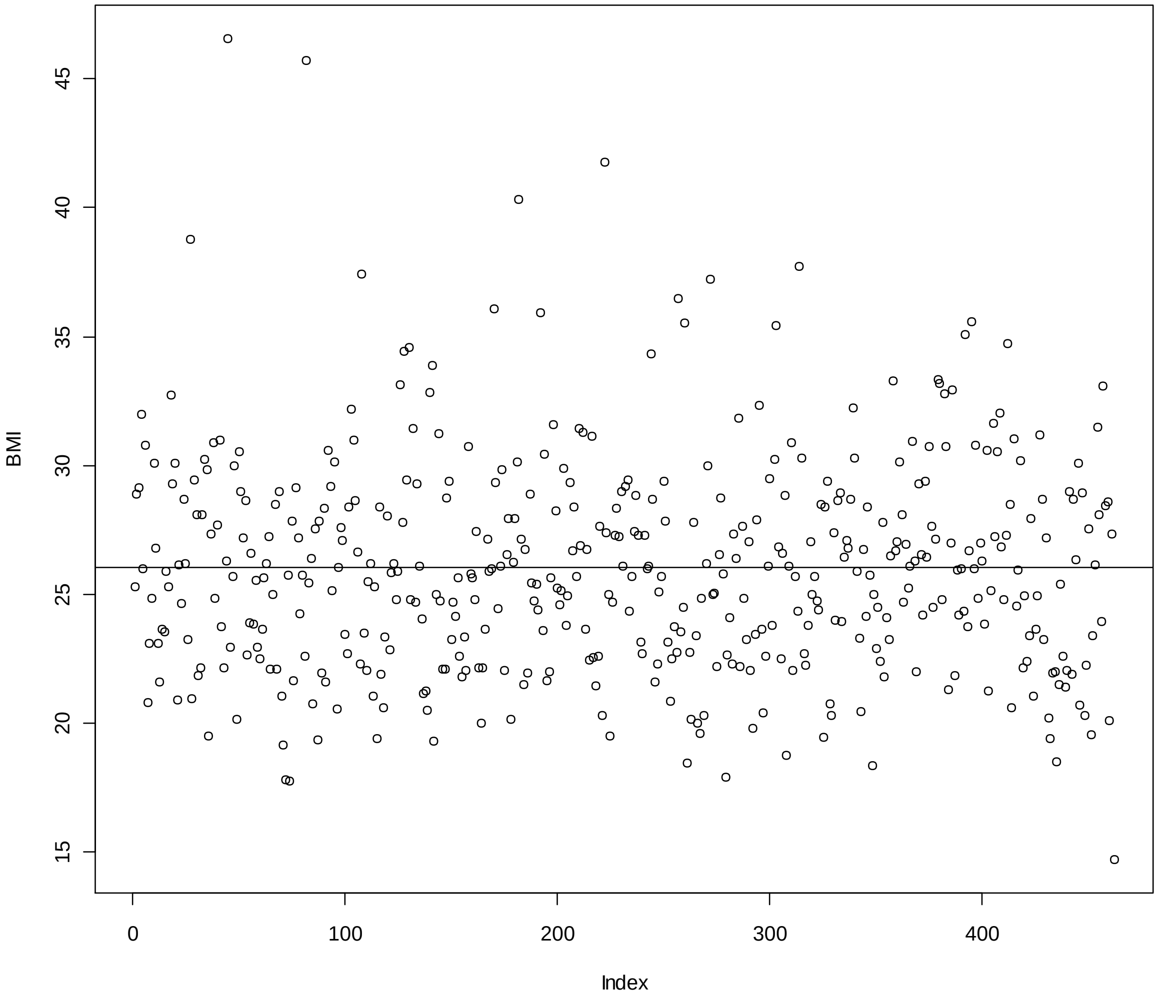
$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$EY_1 = \beta_0$$
$$EY_2 = \beta_0$$
$$\vdots$$
$$EY_n = \beta_0$$

- Předpokládáme stejnou střední hodnotu pro celý soubor – odhadli jsme výběrový průměr
- Sloupec jedniček budeme v matici plánu uvažovat téměř vždy

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Konstanta – absolutní člen



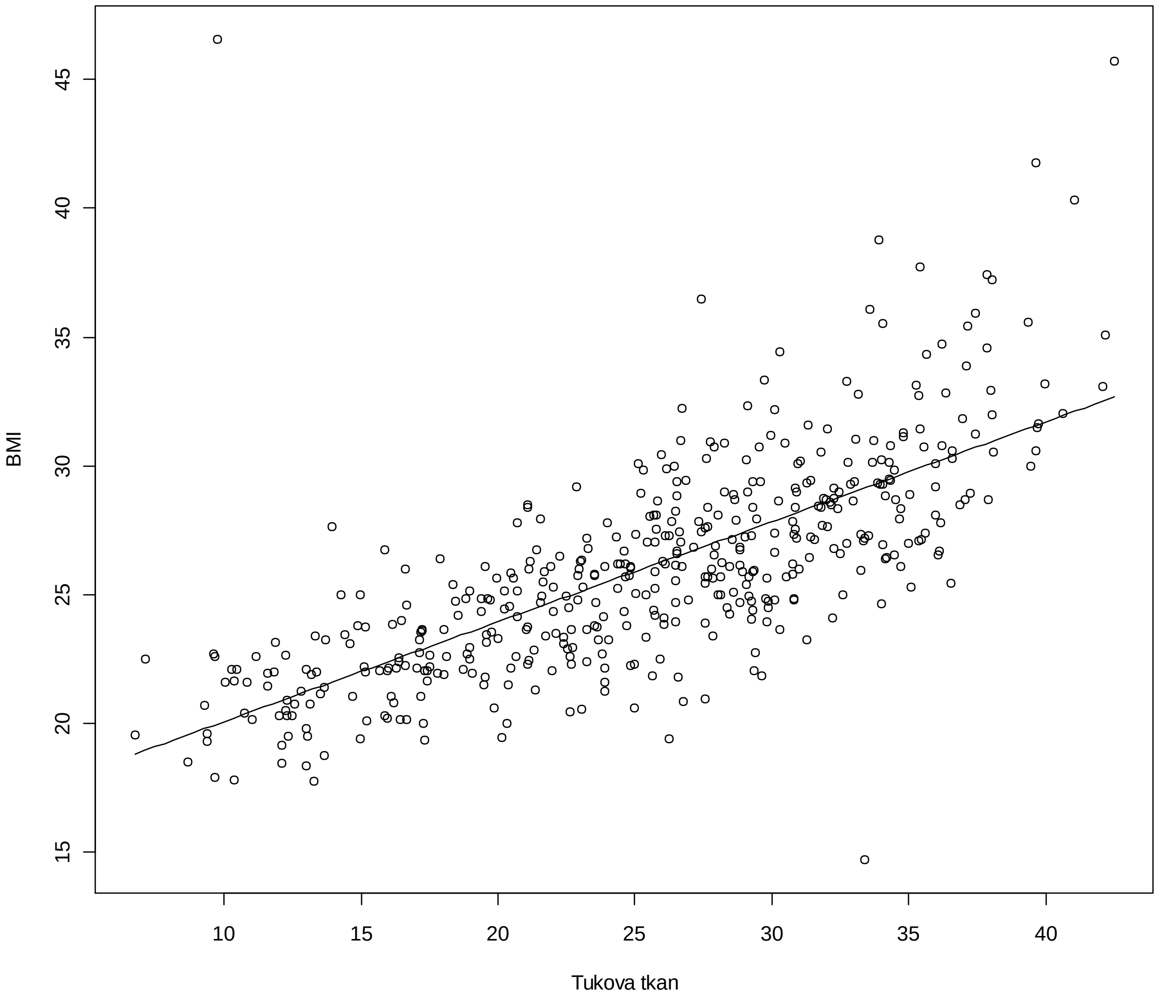
Spojité prediktory

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\begin{aligned} EY_1 &= \beta_0 + \beta_1 x_1 \\ EY_2 &= \beta_0 + \beta_1 x_2 \\ &\vdots \\ EY_n &= \beta_0 + \beta_1 x_n \end{aligned}$$

- Střední hodnota se lineárně mění v závislosti na prediktoru

Spojité prediktory



Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

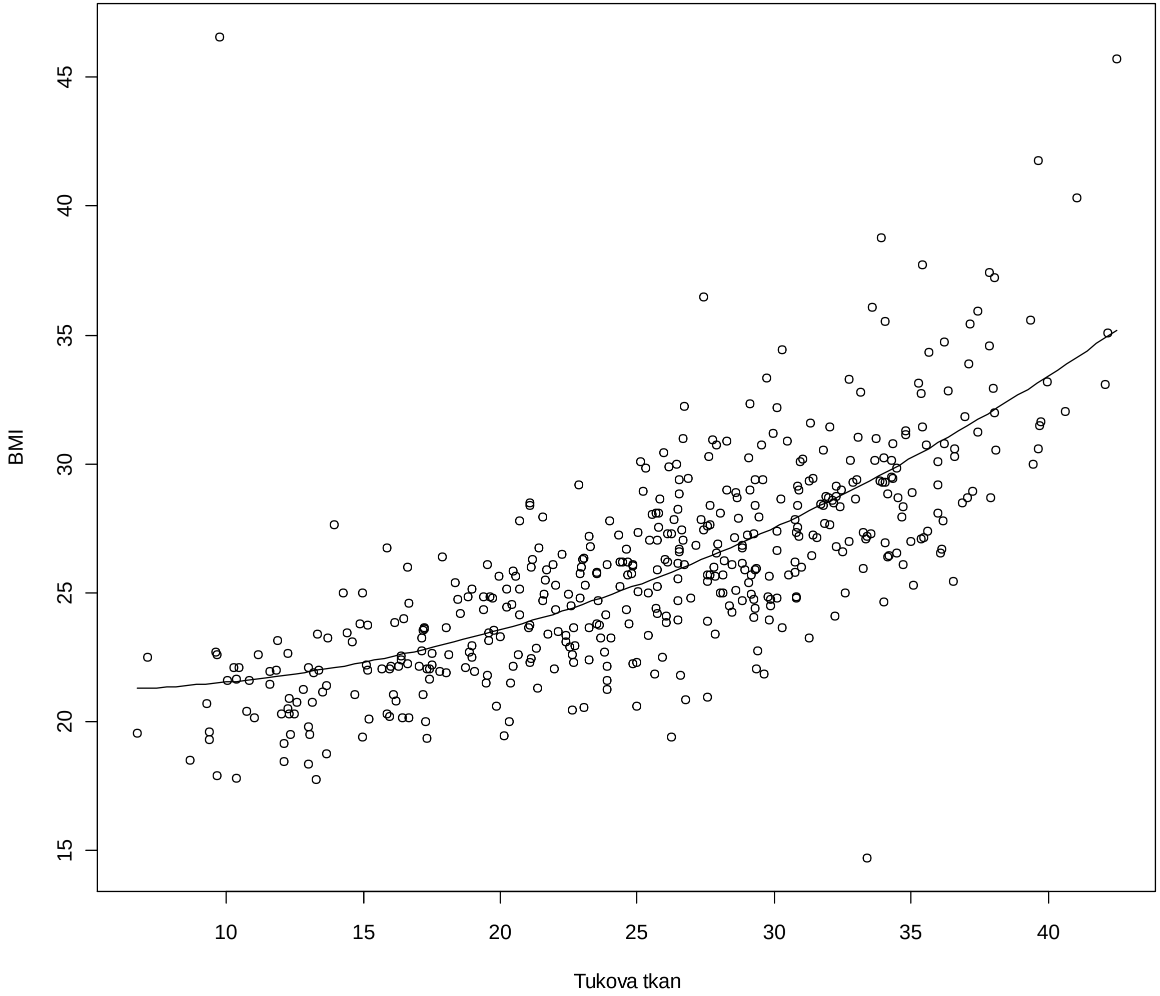
Spojité prediktory – jak na polynom?

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & & \vdots \\ 1 & x_n & x_1^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$EY_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$
$$EY_2 = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2$$
$$\vdots$$
$$EY_n = \beta_0 + \beta_1 x_n + \beta_2 x_n^2$$

- Předpokládáme, že očekávaná hodnota opisuje parabolu
- Lze přidat flexibilitu přidáním další mocniny
- Pozor na multikolineritu a „znesmyslnění“ koeficientů
- NAJEDNOU UŽ LINEÁRNÍ MODELY NEJSOU TAK DOCELA LINEÁRNÍ
- Stále však musí platit, že lineární prediktor je lineární kombinací parametrů modelu

Spojité prediktory – jak na polynom?



Kategoriální prediktory

- Většinou nelze uvažovat jako kvantitativní
  - to by předpokládalo linearitu a stejné rozdíly mezi následujícími skupinami
- Je potřeba vytvořit tzv. dummy proměnné
- Vždy vytváříme o jednu proměnnou méně, než je hodnot kategoriálního prediktoru

Kategoriální prediktory

Nominální kódování

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

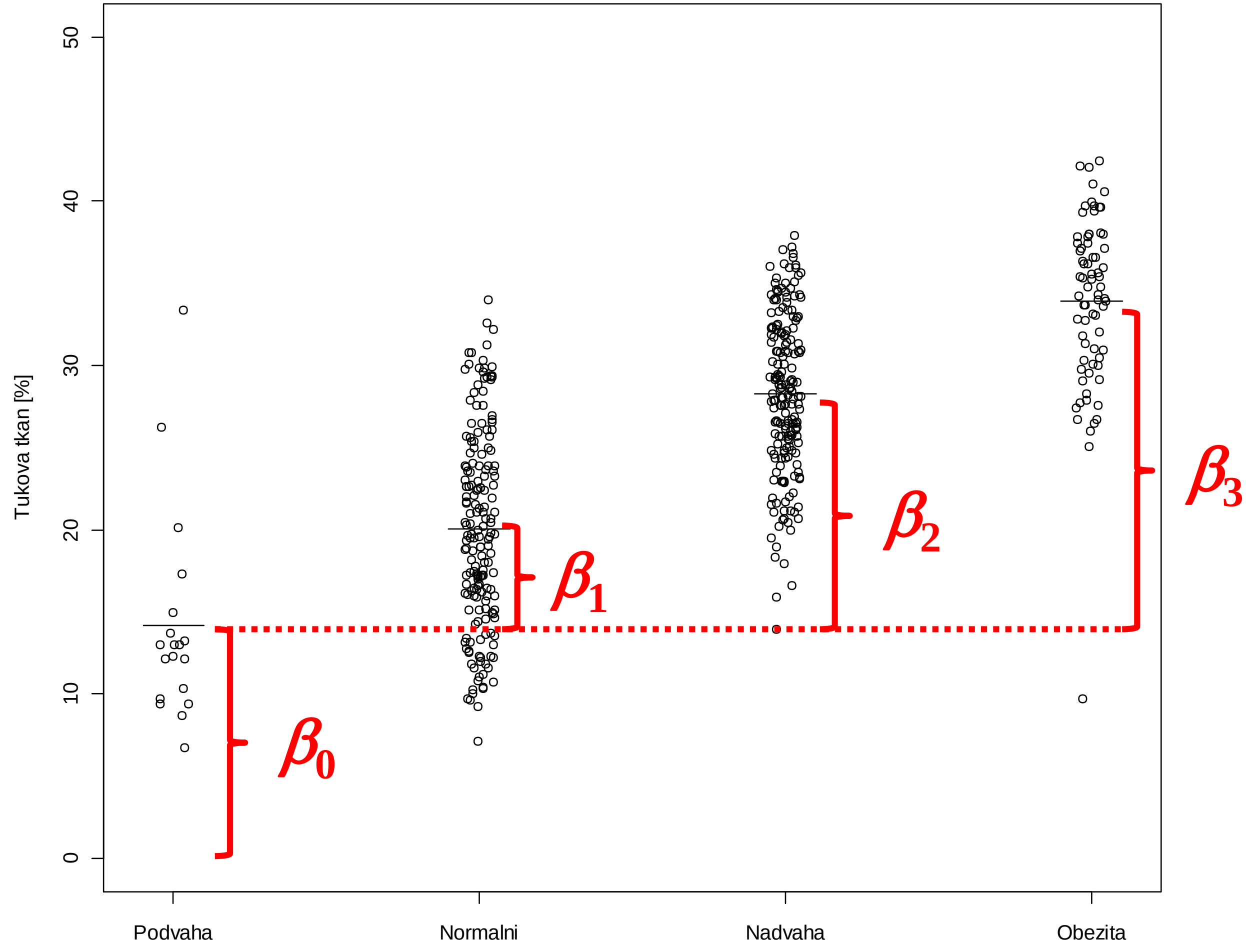
Původní	Nové proměnné		
BMI kateg.	**Normální váha**	**Nadváha**	**Obezita**
Podváha	0	0	0
Normální váha	1	0	0
Nadváha	0	1	0
Obezita	0	0	1

$$EY_i = \beta_0$$
$$EY_i = \beta_0 + \beta_1$$
$$EY_i = \beta_0 + \beta_2$$
$$EY_i = \beta_0 + \beta_3$$

IBA MUNI MED
Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Kategoriální prediktory

Nominální kódování



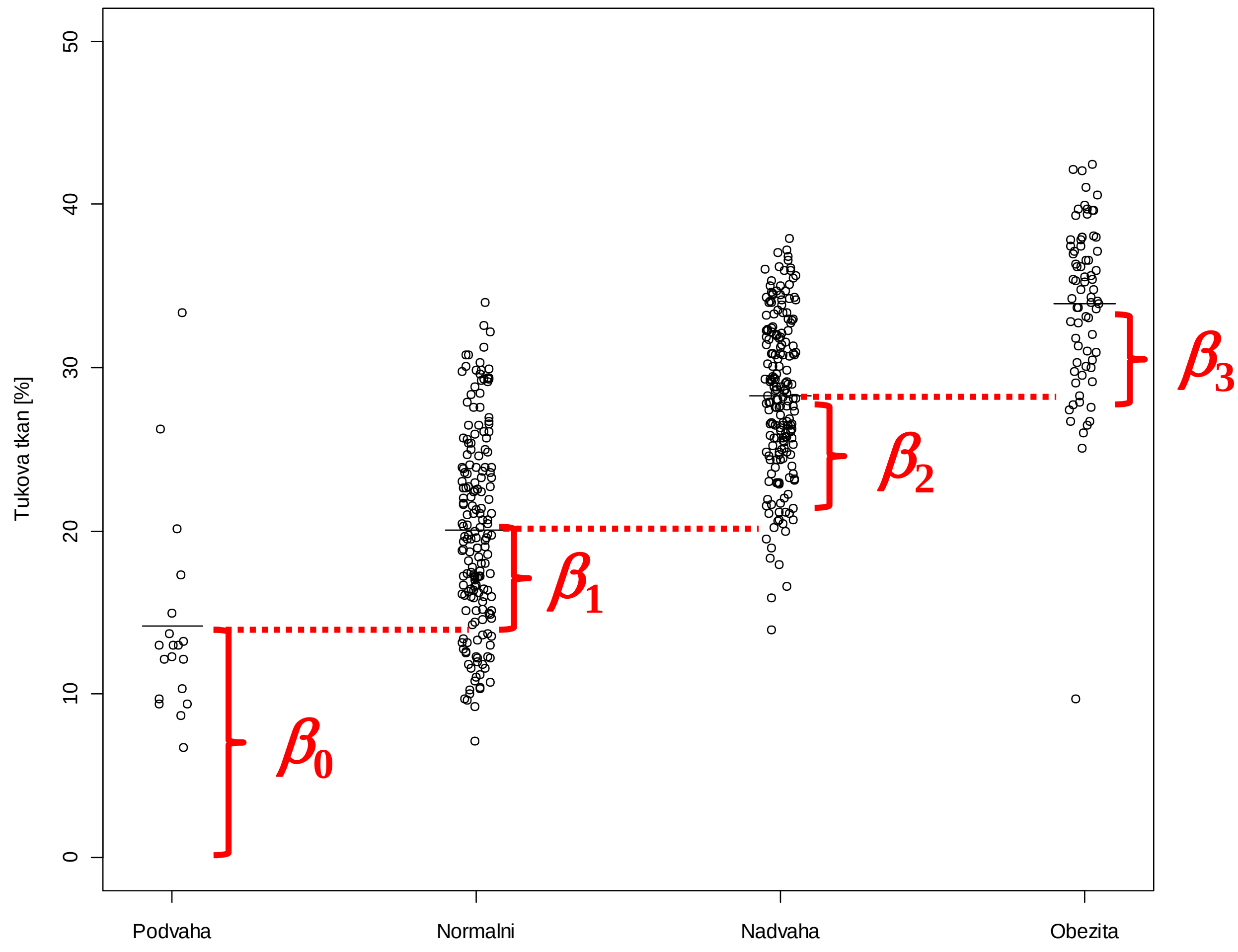
Kategoriální prediktory

Ordinální kódování

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Původní	Nové proměnné			
BMI kateg.	**Normální váha**	**Nadváha**	**Obezita**	
Podváha	0	0	0	$EY_i = \beta_0$
Normální váha	1	0	0	$EY_i = \beta_0 + \beta_1$
Nadváha	1	1	0	$EY_i = \beta_0 + \beta_1 + \beta_2$
Obezita	1	1	1	$EY_i = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Kategoriální prediktory
Ordinální kódování



Lineární regresní model I

Klasické modely novým pohledem

Co už známe, ale jinak...

t-test

- jedná se o lineární model s jedním kategoriálním prediktorem se dvěma hodnotami

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hdashline 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \qquad \begin{matrix} EY_i = \mu \\ \\ EY_i = \mu + \alpha \end{matrix}$$

$$H_0:\ \alpha = 0$$
$$H_1:\ \alpha \neq 0$$

Použijeme výše zmíněný t-test pro lineární modely...

Co už známe, ale jinak...
analýza rozptylu

- jedná se o lineární model s jedním kategoriálním prediktorem s m hodnotami

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 0 \\ \hdashline 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & \cdots & 0 \\ \hdashline \vdots & \vdots & \cdots & \vdots \\ \hdashline 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix}$$

$$EY_i = \mu$$

$$EY_i = \mu + \alpha_1$$

$$EY_i = \mu + \alpha_{m-1}$$

$$H_0: \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$H_1: \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{m-1} \end{pmatrix} \neq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Použijeme výše zmíněný F-test pro lineární modely...

Lineární regresní model I

Předpoklady lineárního regresního modelu

Předpoklady lineární regrese

LINEARITA

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip} + \varepsilon_i, i = 1, \ldots, n$$

ADITIVITA

$\varepsilon_i \sim N(0, \sigma^2)$ **ROZLOŽENÍ REZIDUÍ**

$C(\varepsilon_i, \varepsilon_j) = 0, i \neq j$ **NEZÁVISLOST POZOROVÁNÍ**

Naučíme se s nimi vypořádat...

V prediktorech ... polynomiální zadání

LINEARITA V parametrech ... linkovací funkce

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip} + \varepsilon_i, i = 1, \ldots, n$$

ADITIVITA

Interakce

$$\varepsilon_i \sim N(0, \sigma^2)$$

ROZLOŽENÍ REZIDUÍ

Logistická/Poissonova regrese

$$C(\varepsilon_i, \varepsilon_j) = 0, i \neq j$$

NEZÁVISLOST POZOROVÁNÍ

Korelační struktura – smíšený model

IBA MUNI MED
Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Lineární regresní model I

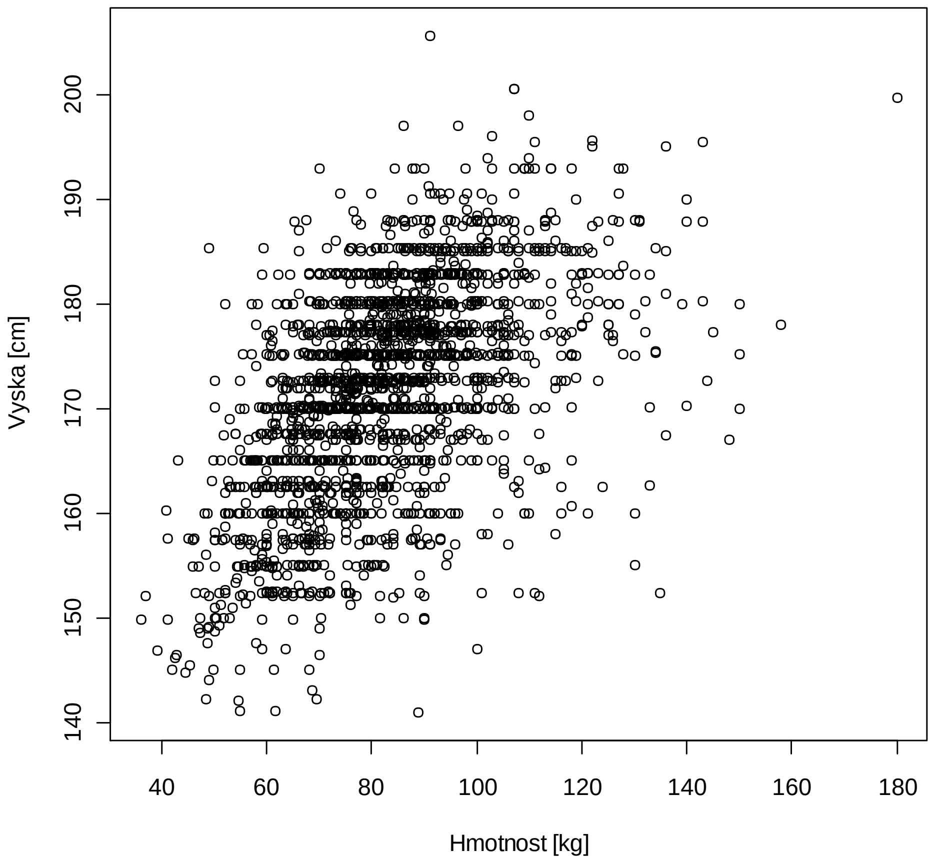
Multikolinearita

Co je multikolinearita?

- Kdyby spolu proměnné nesouvisely, tak by víceprediktorová regrese pozbývala smyslu...
- Problém však představuje vysoká korelace mezi prediktory, neboť znemožňuje odhadnutí účinku jednotlivých prediktorů

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

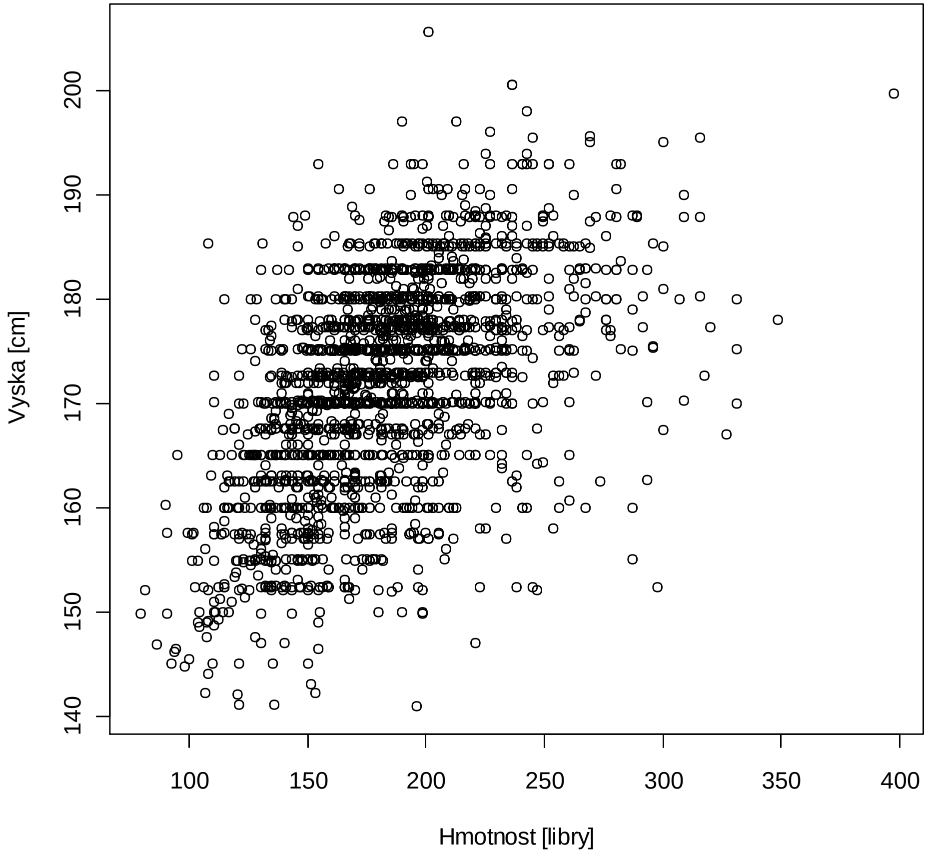
Příklad



Výška = 147 + HmotnostKg x 0,3

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Příklad



Výška = 147 + HmotnostLb x 0,14

Příklad

- Co kdybychom dali do modelu obě proměnné?

- Výška = β_0+ HmotnostKg · β_1 + HmotnostLb · β_2
- Výška = 147 + HmotnostKg · 0,3+ HmotnostLb · 0
- Výška = 147 + HmotnostKg · 0 + HmotnostLb · 0,14

- Výška = β_0+ (HmotnostLb · 0,45) · β_1+ HmotnostLb · β_2
- Výška = β_0+ HmotnostLb · (0,45 · β_1+ β_2)
- tedy kterékoliv koeficienty, které řeší **0,45 · $\boldsymbol{\beta_1}$+ $\boldsymbol{\beta_2}$ = 0,14**
- **a těch je nekonečně mnoho…**

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Problémy s multikolinearitou

- může se objevit divné chování
  - velké změny parametrů při odebrání/přidání prediktoru
  - obrovské směrodatné odchylky
  - extrémní odlehlé hodnoty
- software může upozornit na numerickou nestabilitu
- prediktory v automatických metodách jsou vybírány náhodně
- je obtížné skutečně odhadnout efekt
- může být i dobrý model na predikci, ale nepoužitelný na odhad efektu kovariát

Jak najít dva korelované prediktory?

- Jak najít korelované proměnné?
  - Dvě proměnné – xy-graf, korelační matice
  - Korelační matice odhadnutých koeficientů

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Jak najít více korelovaných prediktorů?

- Ze dvou a více prediktorů lze spočítat jiný

Tolerance

Variance inflation factor (nafouknutí rozptylu)

- převrácená hodnota tolerance
- nad 4 znepokojivé, nad 10 závažné
- výpočet pro i-tý parametr

$$VIF_i = \frac{1}{1 - R_i^2}$$

- kde R_i^2 je čtverec vícenásobné korelace mezi i-tým sloupcem matice plánu a ostatními sloupci (koeficient determinace modelu vysvětlující daný prediktor ostatními prediktory)

Řešení

- Vypustit část korelovaných proměnných
  - ty, které obsahují chybějící data, hůře se měří, nebo jsou z jiných důvodů nedůvěryhodné
- Vytvoření a/nebo proměnné
- Zkobinovat prediktory do jednoho skóre
  - např. věk + výška + váha -> věk + BMI

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita

Lineární regresní model I

Závěr

Co byste po dnešní hodině měli vědět a umět?

- Vědět, jak se definuje lineární regresní model
- Vysvětlit předpoklady regresních modelů
- Umět použít v lineárním regresním modelu různé typy prediktorů
- Vědět, co je multikolinearita, jak ji zjistit a jak se s ní vypořádat

Institut biostatistiky a analýz
Lékařská fakulta, Masarykova univerzita