Přednáška I
AKM I

Lukáš Frýd

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Máme 3 základní pilíře

Odhad regresní rovnice pomocí vhodné funkce

Ověření předpokladů

Statistická verifikace

Jaká je střední hodnota $E(X) = \mu$ výšky mužů v ČR?

Potřebujeme odhadnout střední hodnotu μ

Jak odhadneme μ ?

Aritmetický průměr
Medián
Jiný průměr atd.

Ovlivňuje důchod spotřebu?

$$C = \beta_0 + \beta_1 Y + \epsilon$$

Potřebujeme odhadnout podmíněnou střední hodnotu

$$E(C|Y) = \beta_0 + \beta_1 Y$$

Jak se v průměru mění C, když se mění Y

Jak odhadneme $E(C|Y)$?

Metoda nejmenších čtverců
Zobecněná metoda nejmenších čtverců
Dvoustupňová metoda nejmenších čtverců
Atd.

Který odhad použijeme a proč?
Požadujeme, aby měl nějaké vlastnosti?

$\tilde{X} - odhad\ \ E(X)$ $b - odhad\ \ \beta$

Požadavky na odhad

- **Nezkreslený (nestranný, nevychýlený) -**

$$E(\tilde{X}) = \mu \qquad E(b) = \beta$$

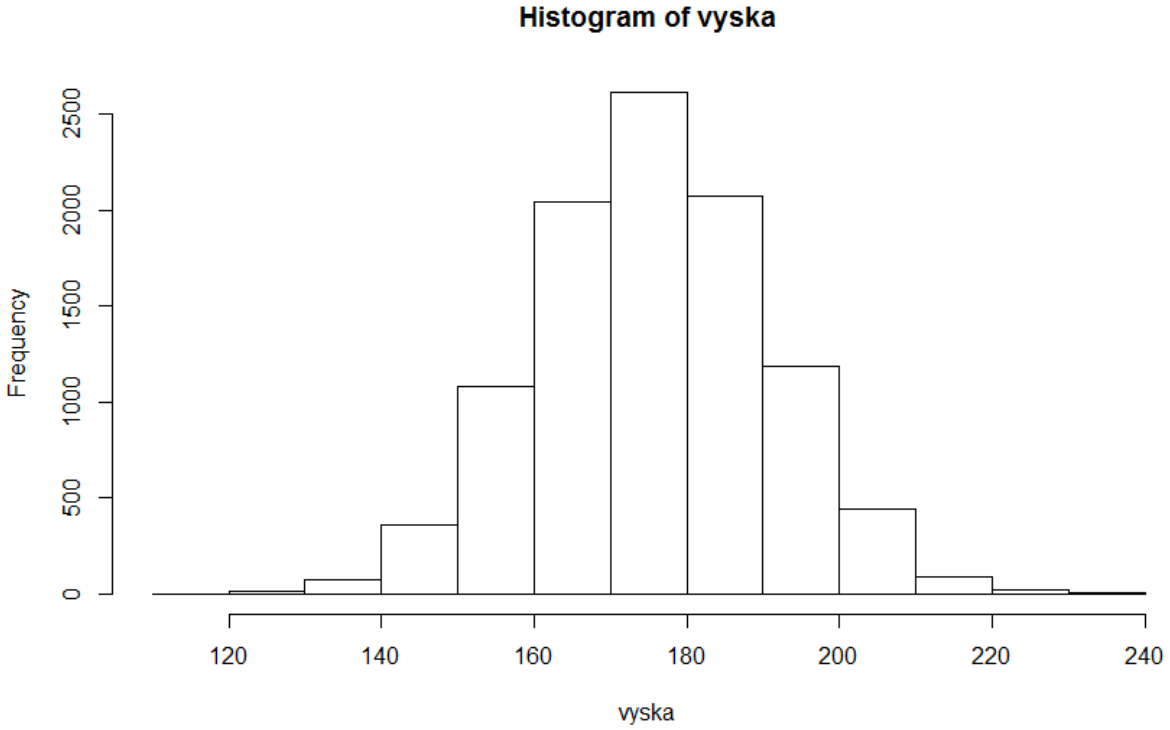
- **Konzistentní**

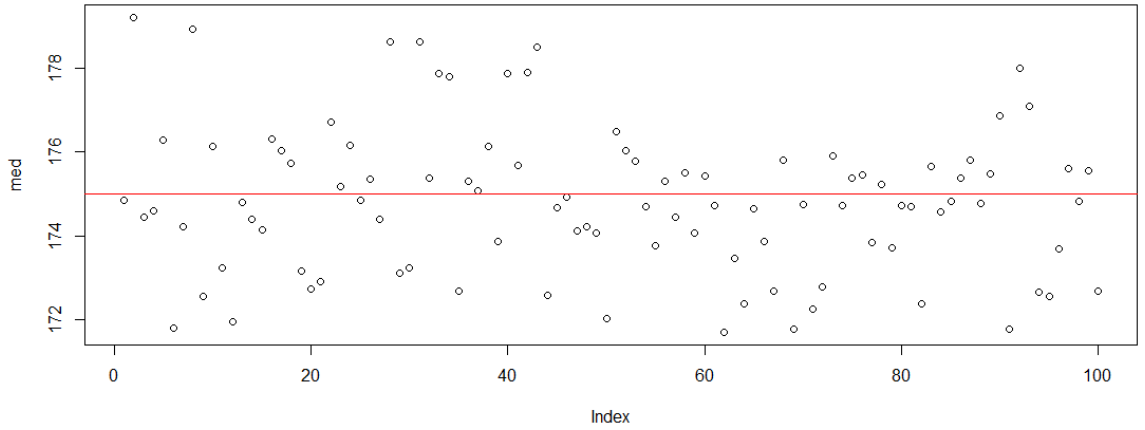
$$E(\tilde{X}) \xrightarrow{n\to\infty} \mu \qquad E(b) \xrightarrow{n\to\infty} \beta$$
$$\mathrm{Var}(\tilde{X}) \xrightarrow{n\to\infty} 0 \quad pak \tilde{X} \xrightarrow{P} \mu \qquad Var(b) \xrightarrow{n\to\infty} 0 \quad pak\ b \xrightarrow{P} \beta$$

- **Vydatný**

$$Var(\tilde{X}) < Var(\bar{X}) \qquad Var(b) < Var(\bar{b})$$

Jak zjistím, že daná metoda odhadu nám bude poskytovat odhady s danými vlastnostmi?



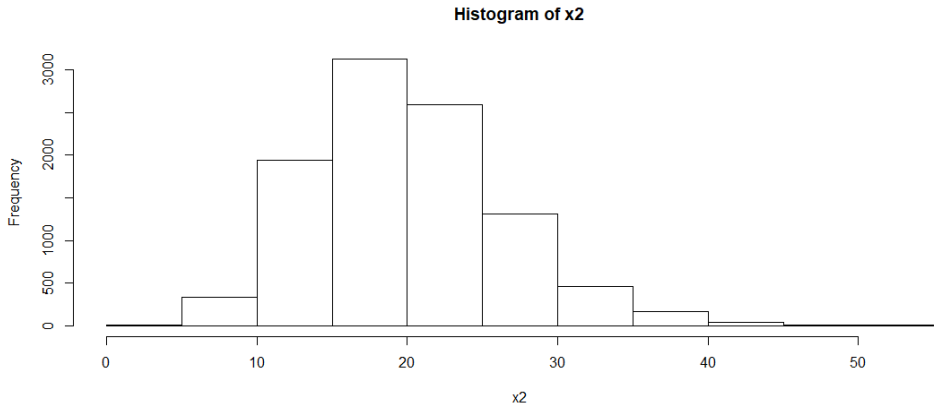


Pro odhad μ jsem použil medián

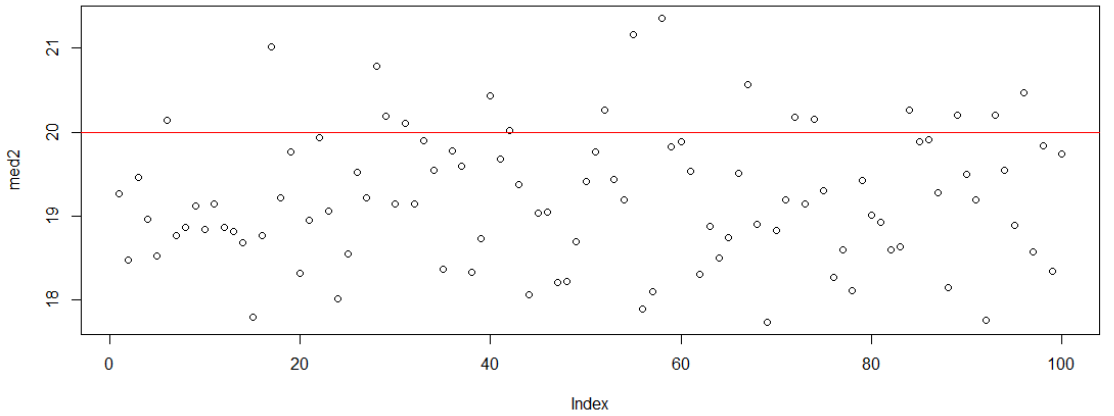
$$vyska \sim N(175{,}225)$$

NEZNÁM – VYGENEROVAL SEM DATA!!!!!!

V realitě existuje „nějaký neznámý“ proces, který určuje výšku

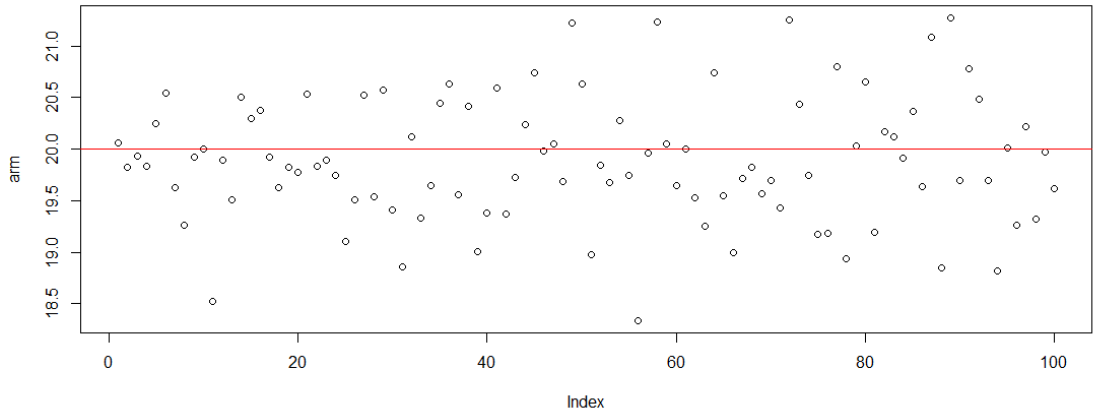


Median



$$x \sim \chi^2(20)$$

Aritmetický průměr



Jaký předpoklad musí být splněn, aby nám
Medián dával nezkreslený odhad střední hodnoty?

Náhodná veličina musí mít symetrické rozdělení!

Když se přeneseme zpět k naší regresi a odhadu podmíněné střední hodnoty.
Jaké podmínky musí být splněny, aby metoda nejmenších čtverců dávala odhady:

- **Nezkreslený (nestranný, nevychýlený)**
- **Konzistentní**
- **Vydatný**

Gauss-Markovovy předpoklady

Blue odhad

Nezkreslený odhad

1) Lineární v parametrech

2) $E(\varepsilon) = 0$

3) $E(\varepsilon|X)=E(\varepsilon) = 0$

4) Není dokonalá multikolinearita

5) $Var(\varepsilon|X) = \sigma^2 I$

Existuje více odhadů, které splňují 1,2,3,4
Hledáme ten nejlepší – s nejmenším rozptylem
Best linear unbiased estimators

```
set.seed(101)
sigma=500
N=10000
YD=rnorm(N,23000,1000)
e = rnorm(N, 0, sigma)        # vygenerujeme náhodnou slozku se strední hodnotou 0 a rozptylem 1
C=-1200+0.8*YD + e            # Populacní regresní funkce, proces, který generuje velikost wage
```

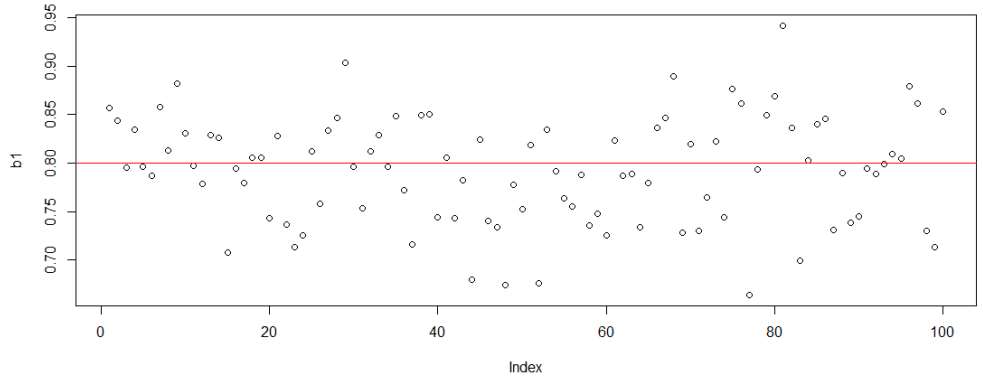
Musíme rozlišovat

$$C = \beta_0 + \beta_1 YD + \varepsilon$$

Populační regresní funkce

$$C = b_0 + b_1 . YD + e$$

výběrová regresní funkce (sample)



R-ko ukaž

Posledním krokem je statistická verifikace
(samozřejmě je nutná i ekonomická, ale k tomu Vám stačí selský rozum a znalost ekonomické teorie)

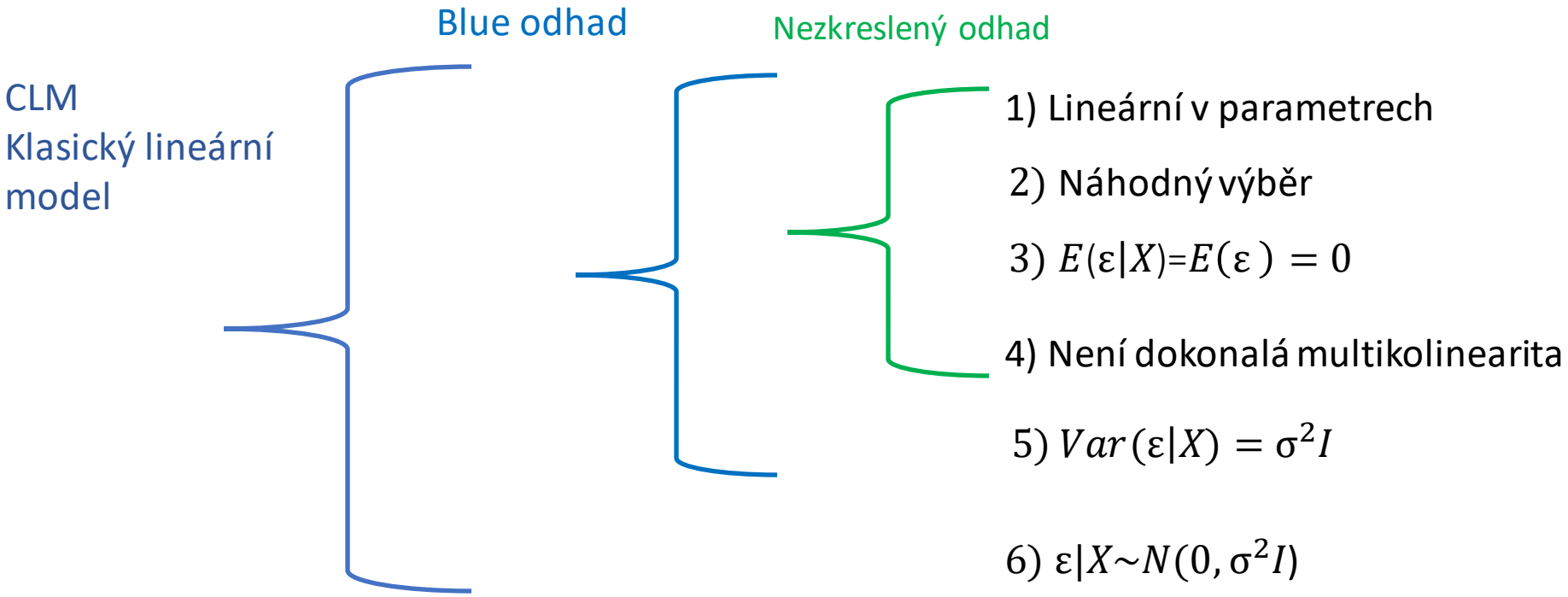
Když se vrátíme k odhadu průměrné výšky v populaci.
Vzneseme hypotézu, zdali je průměrná výška mužů rovna 180 cm.

$$H_0: \mu = 180$$
$$H_1: \mu \neq 180$$

U regrese budeme například testovat:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Gauss-markovovy předpoklady + chování náhodné složky



Proč GM?
Po splnění předpokladů GM – nemusíme hledat/použít jiných metod pro odhad
Nenajdeme lepší maximálně stejně „ dobré“
ve skupině lineárních modelů tedy, ale to nám je zatím jedno

Body 1-5 jsou GM předpoklady
Předpoklad 6 již není GM!!!
1-6 klasický lineární model

Proč $\varepsilon|X\sim N(0,\sigma^2 I)$?
Jelikož pak bude platit $b\sim N(\beta,\Sigma)$

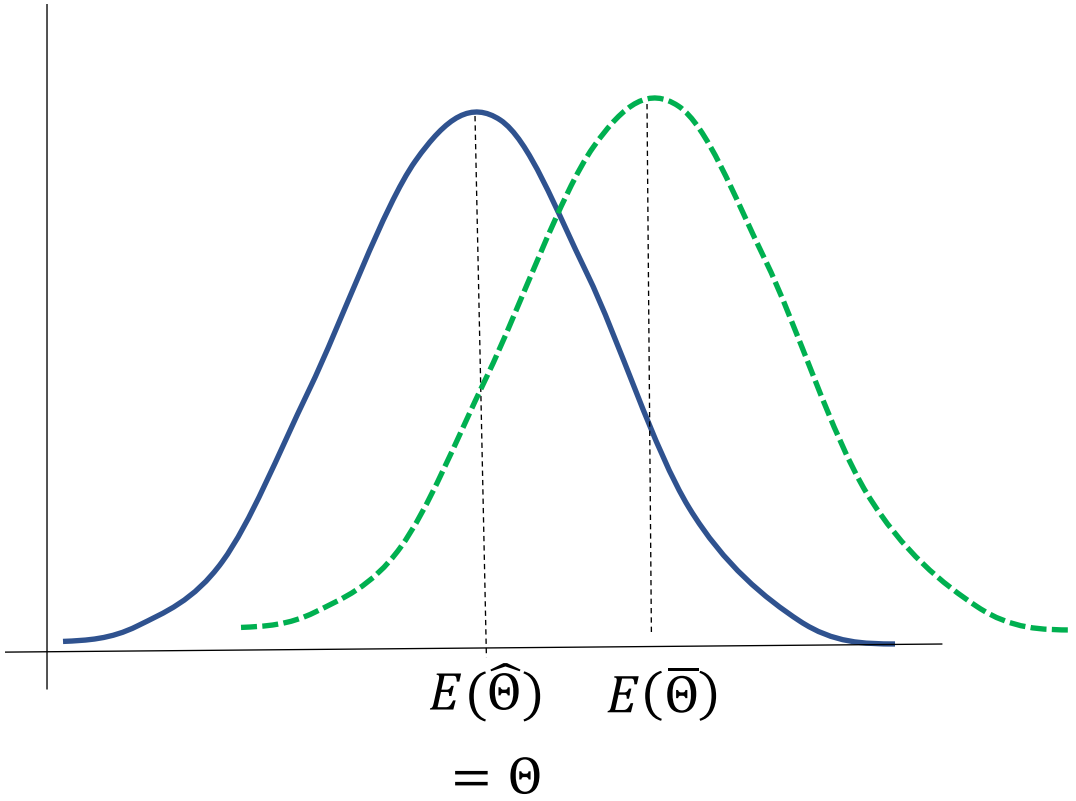
Nezkreslenost je vlastností „estimátoru“
Neříká, že konkrétní odhad je „blízko“ neznámemu parametru
Musíme doufat, že výběrový vzorek je „co nejvíce podobný“ populaci

$\widehat{\Theta} - odhad, výběrový\ rozptyl, aritmetický\ průměr, b - z\ OLS$

$\Theta -$ populační rozptyl, střední hodnota, β

$E(\widehat{\Theta}) > \Theta -$ odhady jsou nadhodnoceny

$E(\widehat{\Theta}) < \Theta -$ odhady jsou podhodnoceny



metoda poskytuje nezkreslený odhad

V reálném životě samozřejmě neznáme hodnotu Θ. Ani si nemůžeme dělat *n* výběrů.
Tak jak si můžeme vědět, že nejsme úplně mimo?
Nemužeme, ale víme, že když jsou SPLNĚNY DANÉ PŘEDPOKLADY tak se pohybujeme kolem hodnoty skutečného parametru
Proto tak záleží na rozptylu daného odhadu.

Jako estimátor střední hodnoty populačního rozdělení E(X)
Použijeme aritmetický průměr výběrového souboru

$$E(\bar{X}) = \mu?$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right)$$

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) \qquad E(X_i) = \mu$$

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu$$

$$E(\bar{X}) = \frac{1}{n}n.\mu = \mu$$

NEzkreslený odhad

Předpokládáme výběr z $výška\ X \sim N(\mu, \sigma^2)$
Každý 1 výběr ze základního souboru, tedy prvek výběrového souboru můžeme chápat jako 1 z možných hodnot náhodné veličiny X (X_i)
Každá z těchto *n* náhodných veličin má stejné rozdělení $X_i \sim N(\mu, \sigma^2)$

$$\hat{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_i \qquad E(\hat{X}) = \mu?$$

$$E(\hat{X}) = E\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right)$$

$$E(\hat{X}) = \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i\right)$$

$$E(\hat{X}) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i)$$

$$E(\hat{X}) = \frac{1}{n-1}\sum_{i=1}^{n} \mu$$

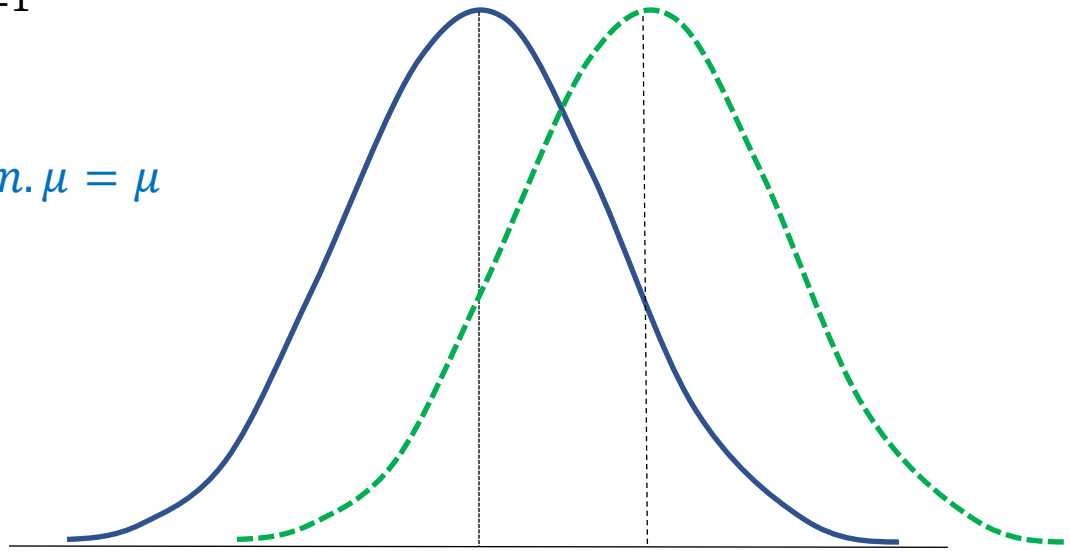
$$E(\bar{\hat{X}}) = \frac{1}{n-1}n.\mu = \frac{n}{n-1}\mu$$

$$E(\hat{X}) \neq \mu$$

Zkreslený odhad

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\bar{X}) = \frac{1}{n} n.\mu = \mu$$



$$E(\bar{X}) = \mu \qquad E(\hat{X})$$

$$E(\hat{X}) \neq \mu$$

zkreslený odhad

$$\hat{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_i$$

$$E(\hat{X}) = \frac{1}{n-1} n.\mu = \frac{n}{n-1}\mu$$

$$E(\bar{X}) < E(\hat{X})$$

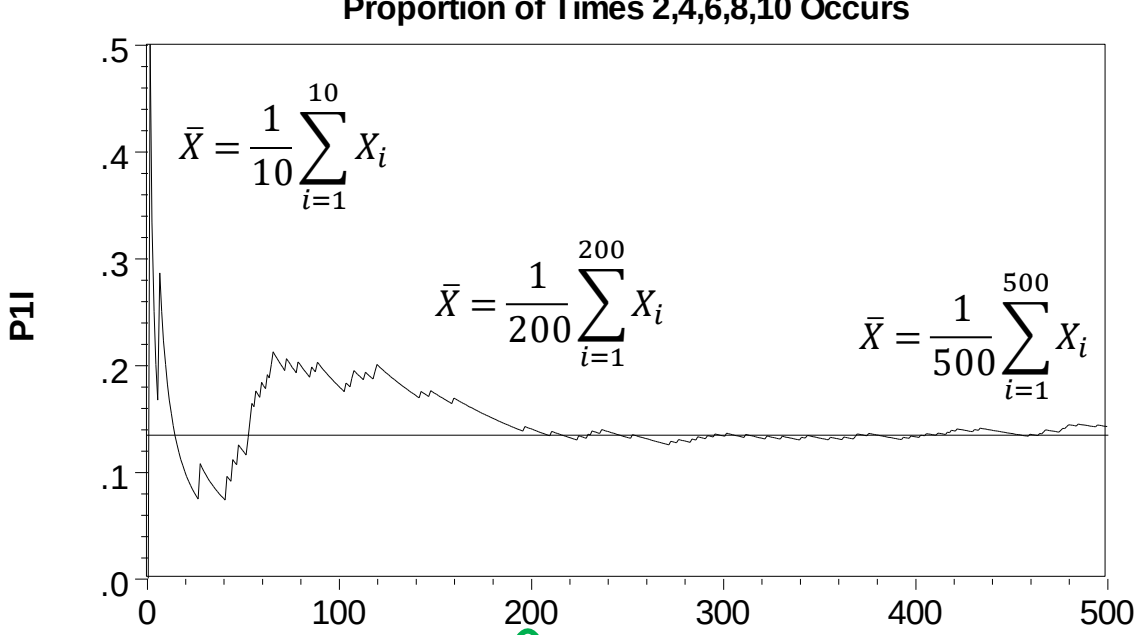
Konzistentní odhad

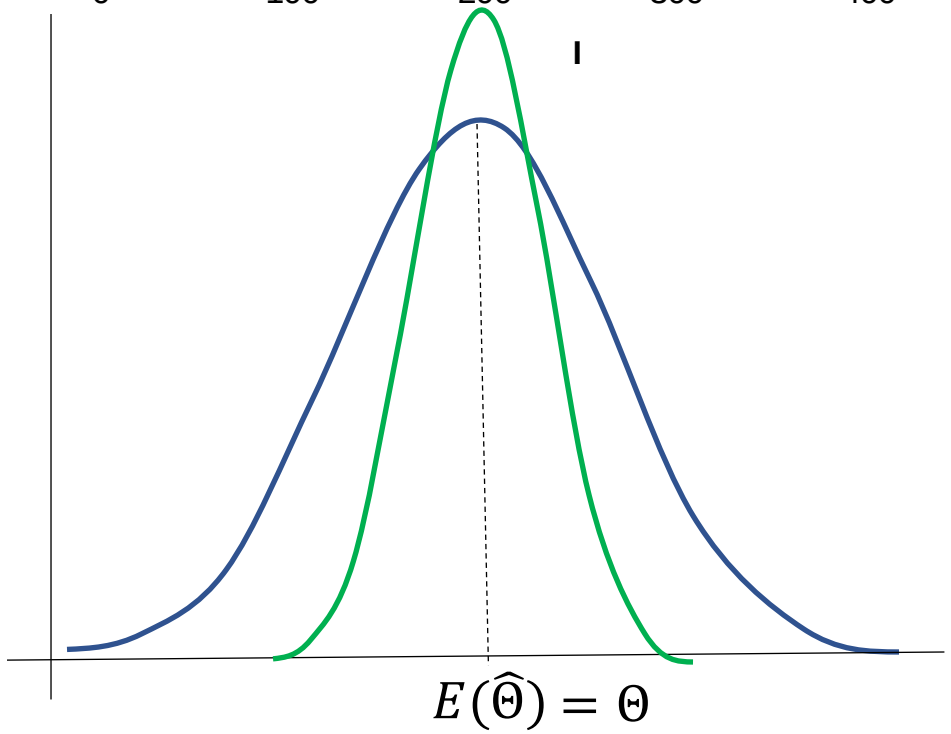
Nezkreslenost pro konečný výběr
Vybíráme *n*-vzorků o konkrétní velikosti

Konzistentní odhad „large sample“ „asymptotic“
Daný vzorek rozšiřujeme do nekonečna
Zpřesní se odhad pokud budeme zvyšovat velikost vzorku?
Bude menší rozptyl?

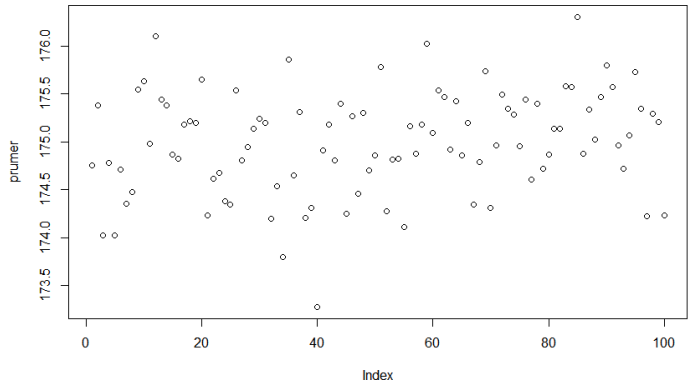
S rostoucím počtem pozorování (n)
konverguje hodnota odhadu
ke skutečné hodnotě parametru základního souboru

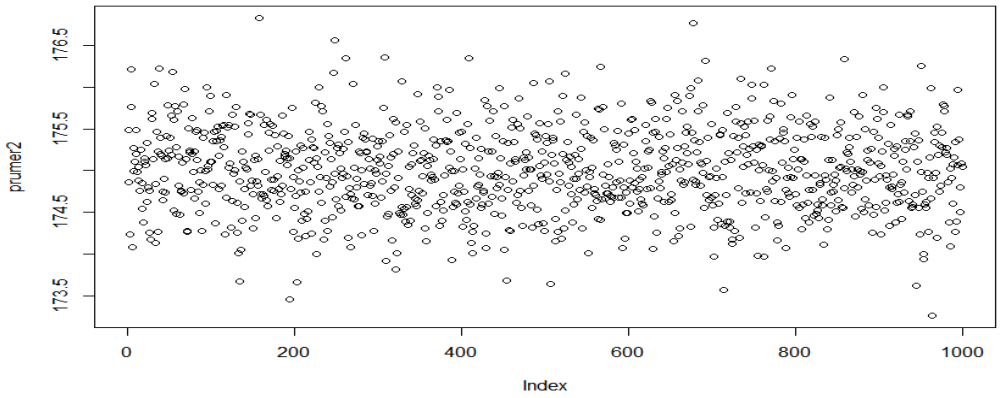
Například u časových řad, nebudou dané estimátory
poskytovat nevychýlený odhad.
Budou však konzistentní.





Populační průměr (střední hodnota) je 175 cm

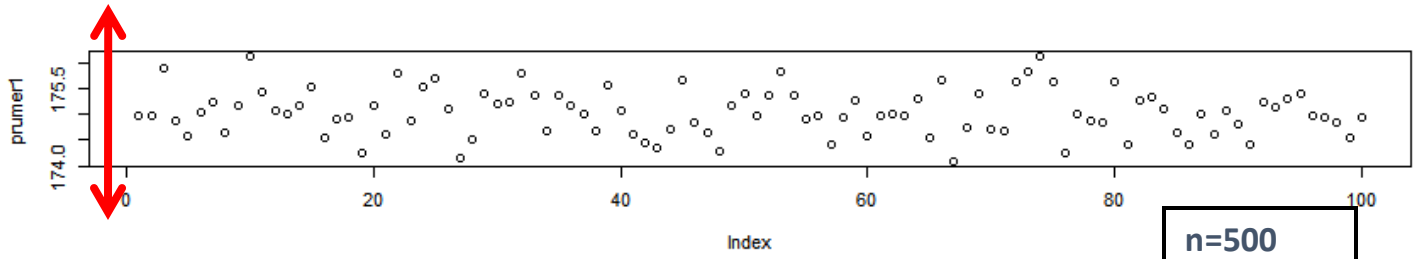


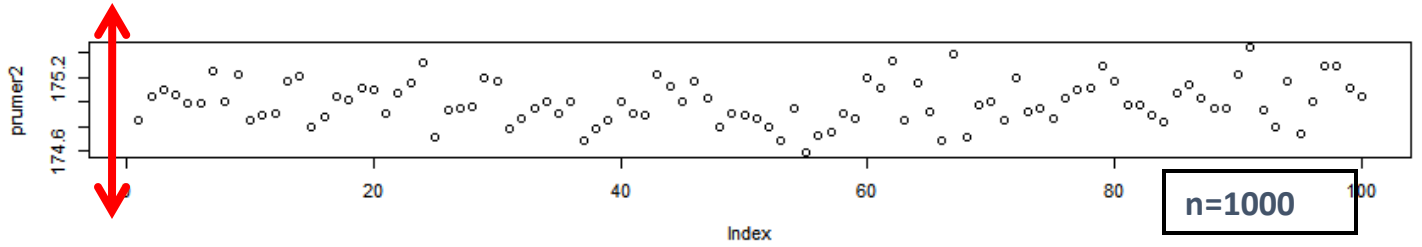


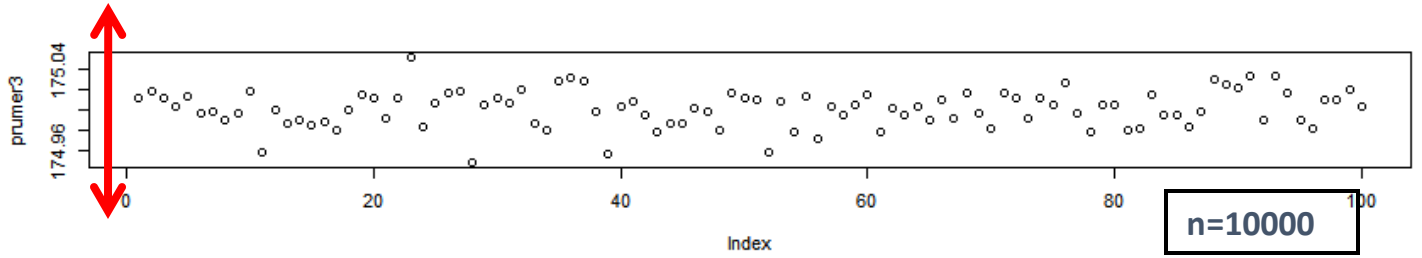
Provedu 100-krát výběr o velikosti 500 lidí

a výpočtu pro každý výběr průměru

Provedu 1000-krát výběr o velikosti 500 lidí





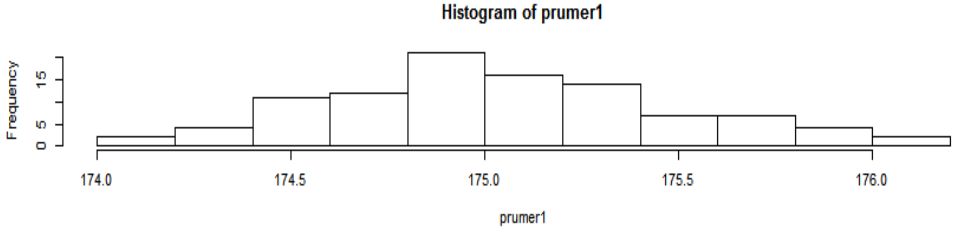


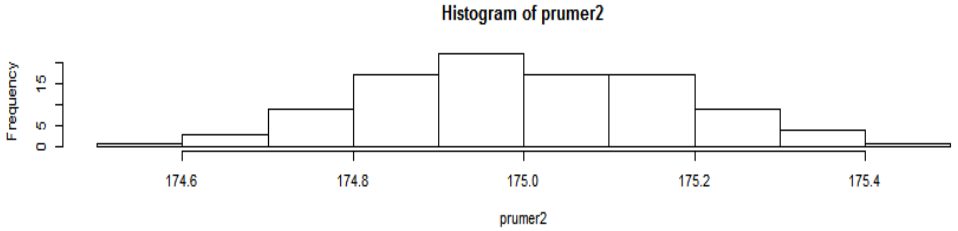
Provedu 100-krát výběr o velikosti 500 lidí

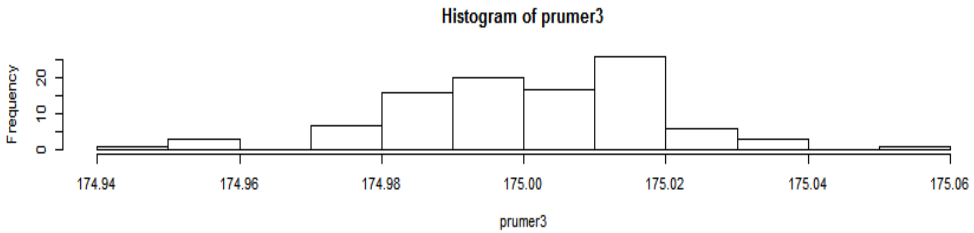
a výpočtu pro každý výběr průměru

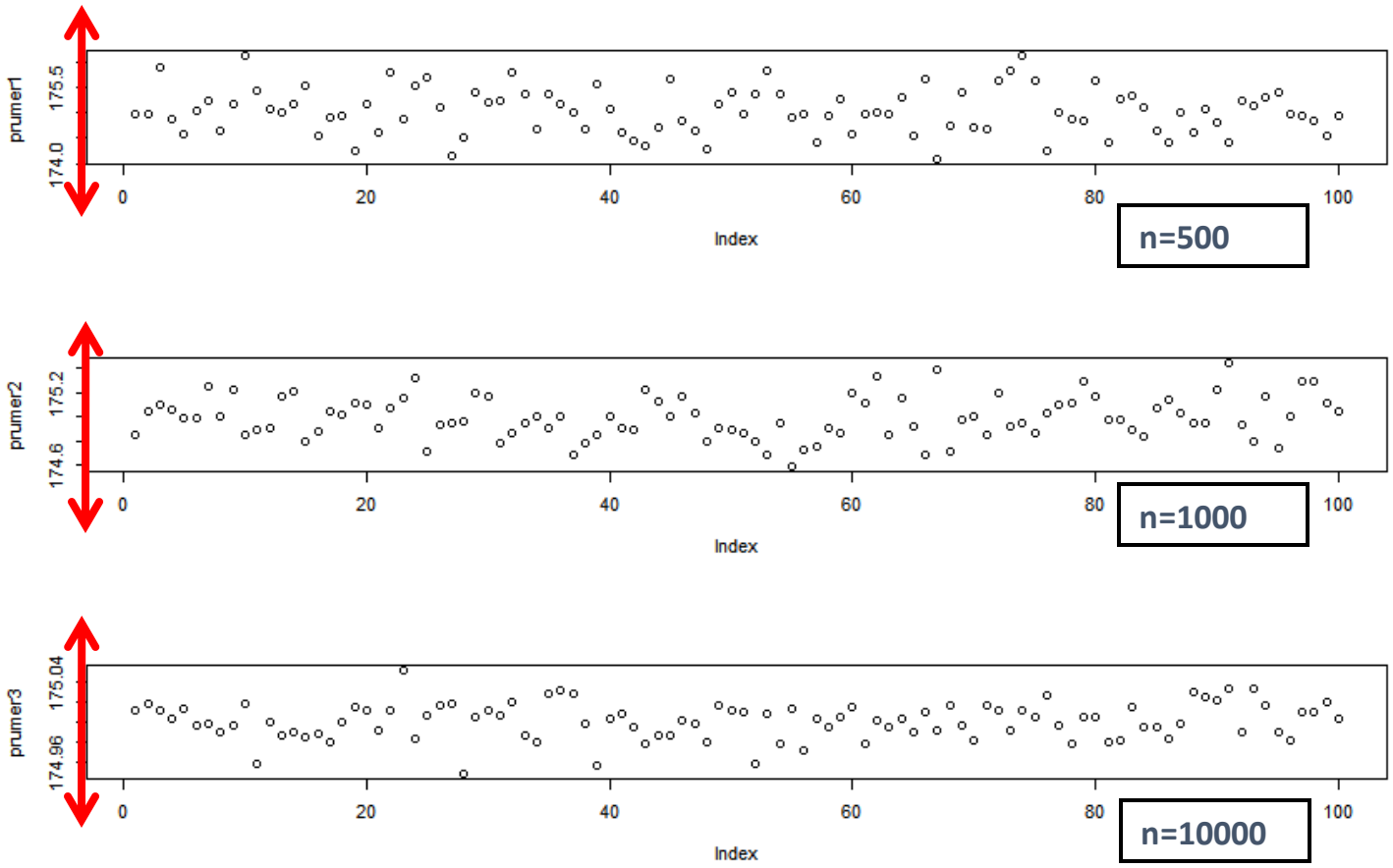
Provedu 100-krát výběr o velikosti 1000 lidí

Provedu 100-krát výběr o velikosti 10 000 lid


Histogram of prumer1
Frequency
0
5
15
174.0
174.5
175.0
175.5
176.0
prumer1


Histogram of prumer2
Frequency
0
5
15
174.6
174.8
175.0
175.2
175.4
prumer2


Histogram of prumer3
Frequency
0
10
20
174.94
174.96
174.98
175.00
175.02
175.04
175.06
prumer3


prumer1
174.0
175.5
0
20
40
60
80
100
Index
n=500
prumer2
174.6
175.2
0
20
40
60
80
100
Index
n=1000
prumer3
174.96
175.04
0
20
40
60
80
100
Index
n=10000

$$\bar{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_i$$

$$E(\bar{X}) = E\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i\right)$$

$$E(\bar{X}) = \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i\right)$$

$$E(\bar{X}) = \frac{1}{n-1}\sum_{i=1}^{n} E(X_i)$$

$$E(\bar{X}) = \frac{1}{n-1}\sum_{i=1}^{n} \mu$$

$$E(\bar{X}) = \frac{1}{n-1} n.\mu = \frac{n}{n-1}\mu$$

$$E(\bar{X}) \neq \mu$$

Zkreslený odhad

$$n \to \infty \qquad E(\bar{X}) = \frac{n}{n-1}\mu = \mu$$

Daná metoda poskytuje
Sice zkreslené, ale konzistentní odhady

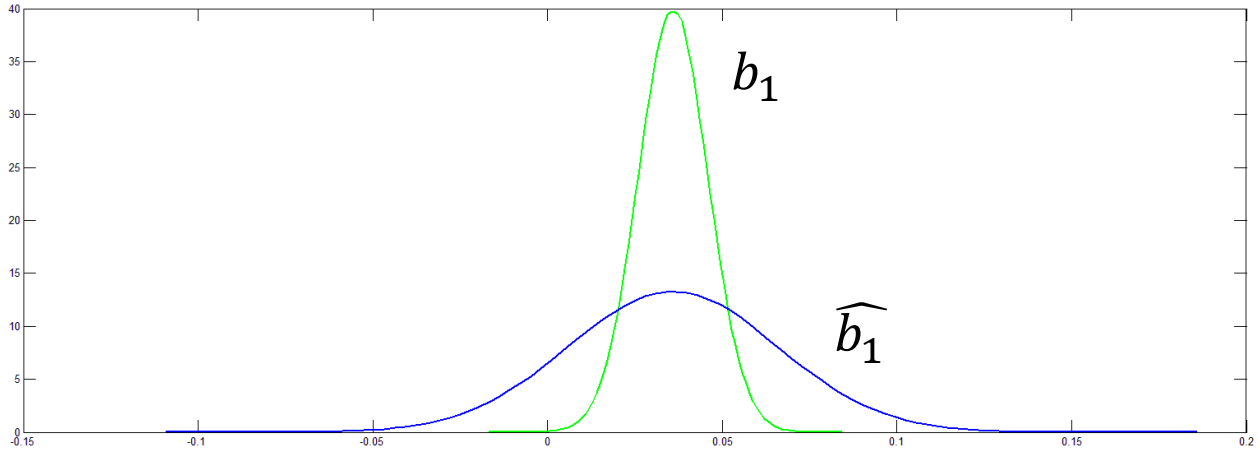
Proč to řešíme?
Někdy nenajdeme nezkreslený odhad
Chceme, aby byl alespoň konzistentní

Vydatný(eficientní, efficient) odhad

Získám 2 NEZKRESLENÉ odhady parametru beta b_1 *a* $\widehat{b_1}$
Tedy střední hodnota obou odhadů je shodná
Který odhad použít?
Bude nás zajímat také rozptyl odhadu
Pro $\widehat{b_1}$ - hodnoty dosahují i záporných hodnot – značný rozptyl

b_1 je vydatnější než $\widehat{b_1}$

Existuje více nezkreslených statistik – my hledáme tu nejvydatnější
S nejmenším rozptylem!!!



NECHCI PO VÁS DŮKAZY NEZKRESLENOSTI, ALE MUSÍTE SI UVĚDOMIT, JAK DŮLEŽITÉ JSOU PŘEDPOKLADY. NEJEN V EKONOMETRII, ALE I V EKONOMII. POKUD NEJSOU SPLNĚNY PŘEDPOKLADY MODELU, ESTIMÁTORU ATD. TAK VLASTNĚ NEVÍTE, JAK SE DANÝ MODEL, ESTIMÁTOR CHOVÁ. MŮŽETE TAK BÝT ÚPLNĚ MIMO REALITU

Metoda nejmenších čtverců

Výběrový soubor – „proženeme nějakou funkcí“ (estimator) – získáme odhad (bodový, intervalový)

Výška studentů VŠE — Estimátor — Odhadneme střední hodnotu výšky dospělých lidí

$$\bar{X} = \frac{1}{n}\sum X_i$$

Vztah mezi C a YD — $b = (X'X)^{-1}X'y$ — Odhadneme parametry podmíněné střední hodnoty

Ok odhadujeme parametry podmíněné střední hodnoty, ale co to znamená?

$$E(y|x) = \beta_0 + \beta_1 . x$$

Vliv x_i na vývoj střední hodnoty y_i

měří změnu střední hodnoty (*y*) - *tedy* $E(y|x)$
v závislosti na změně *x*

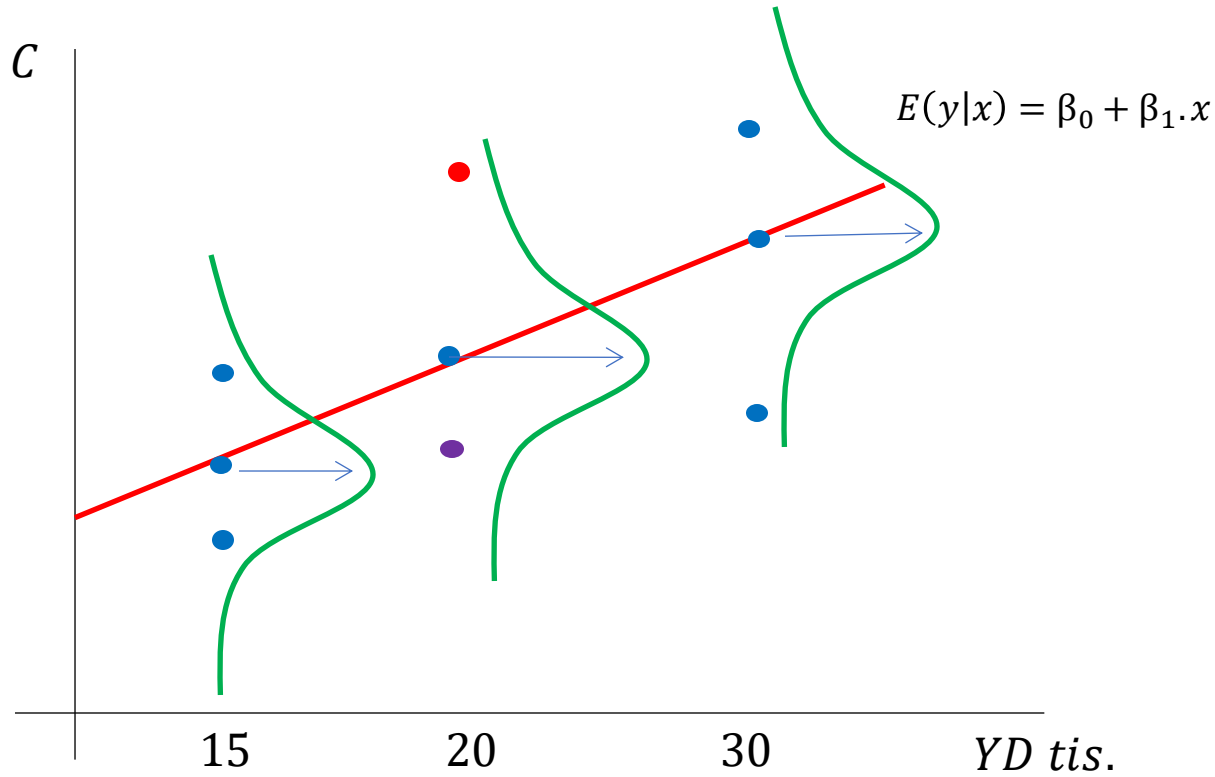
$$C = -1800 + 0.75YD + e$$

Markéta má plat 18 000 a její spotřeba je 15 000
Měla by však být

$$-1800 + 0.75 \times 18000 = 11\,700$$

Ne?

Myšlenka „v průměru“



Přímková regrese

$$y = \beta_0 + \beta_1 . x + \epsilon$$

$$\hat{y} = b_0 + b_1 . x$$

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \dots min$$

b_0 *je odhad* β_0

b_1 *je odhad* β_1

$$Q\ min - hledáme\ extrém\ - minimum$$

Tedy takové $b_{0,1}$, které budou minimalizovat funkci Q

Index *i* představuje *i-té* pozorování
Mzdu, vzdělání Natáliye

Hledáme takové hodnoty $b_{0,1}$
Které budou minimalizovat vzdálenost $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

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

$$y = X\beta + \epsilon$$

$$b = (X'X)^{-1}X'(X\beta + \epsilon)$$

$$b = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon$$

$$b = \beta + (X'X)^{-1}X'\epsilon$$

$$E(b) = \beta + E[(X'X)^{-1}X'\epsilon]$$

$$E(b) = \beta + (X'X)^{-1}X'E(\epsilon)$$

$$pokud\ E(\epsilon) = 0$$

$$E(b) = \beta - nezkreslený\ odhad$$

$$Var(b) = Var(\boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon)$$

$$Var(b) = Var[(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon]$$

$$Var(AX) = AVar(X)A'$$

$$Var(b) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}Var(\epsilon)\left((\boldsymbol{X'X})^{-1}\boldsymbol{X'}\right)'$$

$$Var(b) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}Var(\boldsymbol{\epsilon})\boldsymbol{X}(\boldsymbol{X'X})^{-1}$$

$$Var(b) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\boldsymbol{\sigma^2 I}\ \boldsymbol{X}(\boldsymbol{X'X})^{-1}$$

$$(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\ \boldsymbol{X} = \boldsymbol{I}$$

$$\boldsymbol{\sigma^2 - konstanta, lze\ vytknout\ před}$$

$$Var(b) = \boldsymbol{\sigma^2}(\boldsymbol{X'X})^{-1}$$

Odvodíme nyní Gauss-Markov předpoklady

Abychom získali nezkreslený odhad:

Model je lineární v parametrech- musí platit, jinak bychom měli nelineární metodu nejmenších čtverců

$$E(b) = \beta + (X'X)^{-1}X'E(\epsilon)$$

$$pokud\ E(\epsilon) = 0$$

$$E(b) = \beta - nezkreslený\ odhad$$

A máme předpoklad!

Dále máme matici $(X'X)^{-1}$.
Abychom mohli vytvořit inverzní matici, musí platit, že čtvercová matice $(X'X)$ o velikosti $n \times n$ je tzv. regulární matice
Tedy hodnost matice je rovna n.
Řádky i sloupce jsou lineárně nezávislé
Kdyby tato podmínka neplatila, tak nejsme schopni spočítat $b = (X'X)^{-1}X'\boldsymbol{y}$

Navíc chceme mít BLUE odhad – nejlepší s nejmenším rozptylem

$$Var(b) = \boldsymbol{(X'X)^{-1}X'}Var(\boldsymbol{\epsilon})\boldsymbol{X(X'X)^{-1}}$$

$$Var(b) = \boldsymbol{(X'X)^{-1}X'\sigma^2 I\ X(X'X)^{-1}}$$

$$Var(b) = \boldsymbol{\sigma^2(X'X)^{-1}}$$

Pokud $Var(\boldsymbol{\epsilon}) \neq \boldsymbol{\sigma^2 I}$, tak potom ani
$Var(b) \neq \boldsymbol{\sigma^2(X'X)^{-1}}$
OLS již nebude BLUE. Bude existovat nějaká dodatečná informace o tvaru $Var(\boldsymbol{\epsilon})$, kterou Se kterou bude pracovat jiná metoda. Pamatujete jaká?

Proč nás tak zajímá rozptyl a co jej ovlivňuje?

$$Var(b) = \boldsymbol{\sigma^2}(\boldsymbol{X'X})^{-1}$$

$$Var(b) = \widehat{\sigma^2}(\boldsymbol{X'X})^{-1}$$

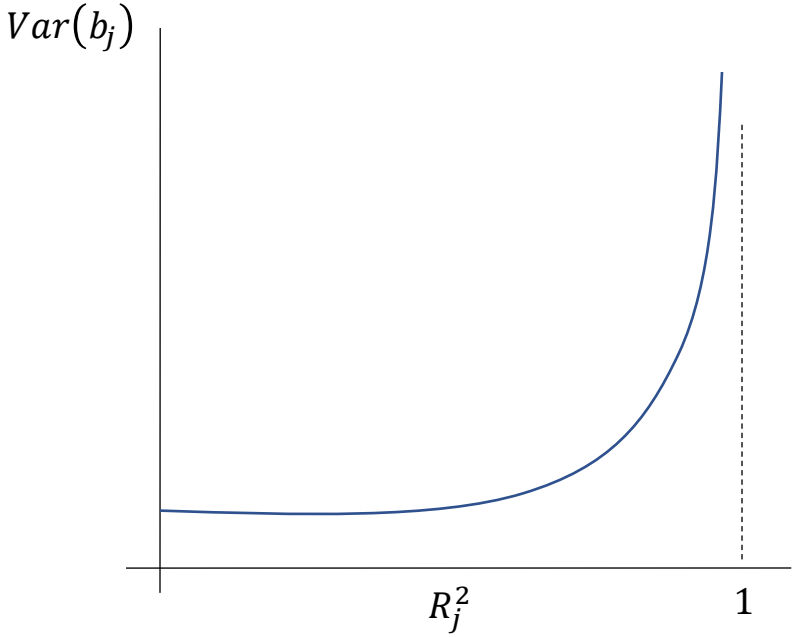
$$Var(b_j) = \frac{\sigma^2}{\sum(x_j - \bar{x_j})^2(1 - R_j^2)}$$

$$Var(b_j) = \frac{\widehat{\sigma^2}}{\sum(x_j - \bar{x_j})^2(1 - R_j^2)}$$

Pozor R_j^2 není R^2 z výstupu programu!!
Čím větší bude R_j^2 - tím větší bude rozptyl odhadu

$R_j^2 - R^2 z\ regrese\ x_j\ na\ všechny\ ostatní\ X + intercept$

$y = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \epsilon \qquad x_1 = \gamma_0 + \gamma_2.x_2 + v \qquad R_1^2$



$R_j^2 \to 1\ Var(b_j) \to \infty$

Problém u testování hypotéz

Interpretace R^2 - kolik variability v x_1
Se nám podařilo vysvětlit pomocí x_2
Čím více – tím větší je vzájemný vztah

- Počet pozorování
- Varibialita x
- R_j^2
- Rozptyl náhodné složky σ^2

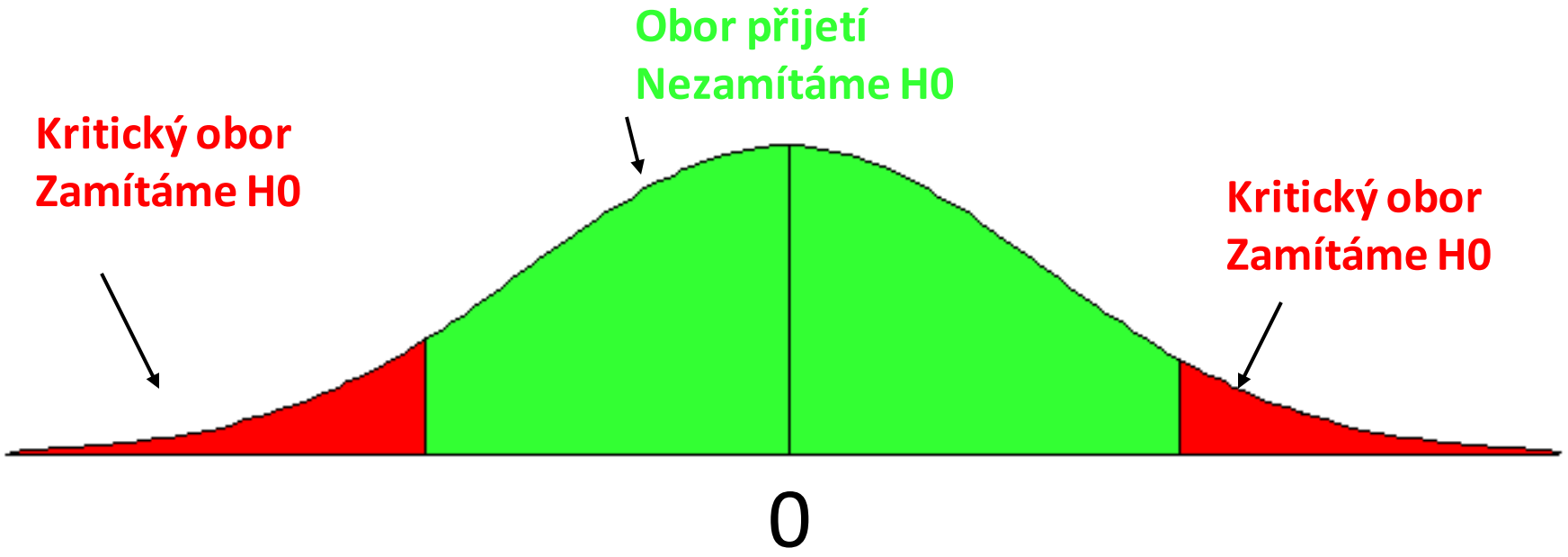
$$spotřeba = \beta_0 + \beta_1 . mzda + \beta_2 . kap.vynos + \beta_3 . urok.mira \ + \varepsilon$$

$$t = \frac{b_3}{\sqrt{Var(b_3)}} \sim t(n-k-1)$$

Jaké hodnoty svědčí pro H1?
„okrajové“ nebo blízko nuly?

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$



$$b \sim iid(\beta, \sigma^2 (X'X)^{-1})$$

$$\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} \sim iid \left[\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \sigma^2 \begin{pmatrix} & & \\ & (X'X)^{-1} & \\ & & \end{pmatrix} \right]$$

$$E(\varepsilon|x) = 0$$

$$neplést!!!$$

$$E(y|x) - regresní\ funkce$$

$$Var(y|x) = skedastická\ funkce$$

Oživit si

Koeficient determinace vs. Upravený koeficient determinace
Log modely umět interpretovat
Dummy proměnné
Náhodná složka vs. Residuum

$$y = b_0 + b_1 x + e \; vs. \; \hat{y} = b_0 + b_1 x$$

OPRAVDU CHÁPAT TESTY HYPOTÉZ!!!!

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání