4EK211 Základy ekonometrie

Odhad klasického lineárního regresního modelu

Cvičení 2

Metodologický postup tvorby EM

1. **Specifikace modelu**

- určení proměnných
- určení vzájemných vazeb mezi proměnnými
- formulace hypotéz – v podobě algebraických vztahů (tj. jedné či více rovnic)
- specifikace náhodných vlivů

2. **Odhad parametrů (tj. kvantifikace)**

- využití disponibilní (tj. výběrové) informace z dat
- použití vhodné odhadové techniky

3. **Verifikace**

- jak se **odhadnutý model** shoduje s teorií a napozorovanými daty
  - **ekonomická** (jestli proměnné modelu mají správný směr a intenzitu)
  - **statistická** (ověření přesnosti a významnosti výsledků)
  - **ekonometrická** (zda byly dodrženy podmínky pro použití dané odhadové techniky a statistických testů)

Metodologický postup tvorby EM

4. **Využití**

- kvalitativní a kvantitativní analýza minulého vývoje
- předpovědi (predikce, prognózy)
- volba hospodářské politiky
  - analýza různých scénářů
  - simulační experimenty
- aplikace na mikroúrovni
  - banky (nesplácení úvěru, odhad ztráty dané nesplácením úvěru, odhad pravděpodobnosti podvodu)
  - marketing (odchod zákazníka, pořízení produktu – cross-sell)
  - cena komodity ve vztahu k poptávanému množství
  - příjem spotřebitele versus cena komodity
  - velikost prodeje versus prostředky na reklamu

Klasický lineární regresní model (KLRM)

Obecný model (maticový zápis)

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

$\boldsymbol{X}$ = matice (n x (k+1)) pozorování exogenních proměnných, včetně 1. sloupce úrovňové konstanty

$\boldsymbol{y}$ = sloupcový vektor (n x 1): n pozorování endogenních proměnných

$\boldsymbol{\beta}$ = sloupcový vektor ((k+1) x 1) regresních parametrů

$\boldsymbol{u}$ = sloupcový vektor náhodné složky, $u_i \sim N(0, \sigma^2)$, *iid*

$\boldsymbol{k}$ = počet vysvětlujících proměnných; k + 1: počet regresních parametrů (k + jedna úrovňová konstanta)

$\boldsymbol{n}$ = počet pozorování

Zápis KLRM po složkách

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- za předpokladu: $n > (k+1)$
- odhadnout koeficienty $\boldsymbol{\beta}$ – tj. určit $\boldsymbol{b}$ (resp. $\hat{\boldsymbol{\beta}}$ - odhady $\boldsymbol{\beta}$)

Klasický lineární regresní model (KLRM)

Obecně specifikovaný vs. odhadnutý model

$\boldsymbol{y} = \boldsymbol{X\beta} \rightarrow \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$

- model „naplníme“ daty a odhadneme

  $\rightarrow \boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e}$

  $\boldsymbol{X}$, $\boldsymbol{y}$ jsou pozorování, $\boldsymbol{b}$: odhad parametrů, $\boldsymbol{e}$: rezidua

  $\rightarrow \quad \hat{\boldsymbol{y}} = \boldsymbol{Xb}$

  kde $\hat{\boldsymbol{y}}$ jsou vyrovnané hodnoty, rezidua jsou rovna $\boldsymbol{0}$

Rezidua vs. náhodná složka

rezidua = rozdíl mezi napozorovanými a vyrovnanými hodnotami:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

náhodná složka = rozdíl mezi napozorovanými hodnotami a jejich střední hodnotou:

$$\boldsymbol{u} = \boldsymbol{y} - \mathrm{E}(\boldsymbol{y})$$

Klasický lineární regresní model (KLRM)

Bodová odhadová funkce „*b*"

- Založena na minimalizaci součtu čtverců
- součet čtverců reziduí: $\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$

  $\boldsymbol{b} \ldots \sum \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e} \rightarrow \mathbf{min}$
- kdy je funkce minimální:
  - první derivace funkce je nulová
  - druhá derivace funkce je kladná
  - fce více proměnných: matice 2. derivací je P.S.D.
- **metoda nejmenších čtverců (MNČ / OLS)**
- funkční předpis odhadové funkce:

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

  - poskytuje odhady:
    - **nestranné** (resp. nevychýlené)
    - **vydatné**

  $\boldsymbol{b}^{\mathrm{T}} = (b_0, b_1, b_2, \ldots, b_k)$

Klasický lineární regresní model (KLRM)

Odvození bodové odhadové funkce vektoru „*b*"

***f*(*b*) … ∑ e$^{\mathrm{T}}$e → min**

$$\boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}$$

$$\frac{\partial \boldsymbol{e}^T\boldsymbol{e}}{\partial \boldsymbol{b}^T} = \frac{\partial(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b})}{\partial \boldsymbol{b}^T} = 0 - 2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = 0$$

$$2\boldsymbol{X}^T\boldsymbol{y} = 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}$$

$$(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$$

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

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

Klasický lineární regresní model (KLRM) - příklad

Příklad 1:

$y_i = (1, 4, 7, 9)$

$x_i = (4, 3, 1, 1)$

$y_i = \beta_0 + \beta_1 x_i + u_i,\ i = 1, 2, \ldots, 4$

- stanovte odhad parametrů β_0 a β_1, aby součet čtverců odchylek vyrovnaných hodnot od hodnot napozorovaných byl minimální
- napište odhadnutou regresní nadrovinu
- vypočítejte vyrovnané hodnoty
- vypočítejte rezidua e_i
- proveďte výpočty v programu MS Excel (využijte maticové počty)
- proveďte výpočty v programu GiveWin2 (s modulem PcGive)
- **řešení:** $b_0 = 31/3 = 10,33$

  $b_1 = -61/27 = -2,26$

Klasický lineární regresní model (KLRM) - příklad

Příklad 2:

$y_i = (5, 4, 6, 4, 3)$

$x_i = (3, 2, 3, 2, 3)$

$y_i = \beta_0 + \beta_1 x_i + u_i, i = 1, 2,...,5$

- stanovte odhad parametrů β_0 a β_1, aby součet čtverců odchylek vyrovnaných hodnot od hodnot napozorovaných byl minimální
- napište odhadnutou regresní nadrovinu
- vypočítejte vyrovnané hodnoty
- vypočítejte rezidua e_i
- proveďte výpočty v programu MS Excel (využijte maticové počty)
- proveďte výpočty v programu GiveWin (s modulem PcGive)
- **řešení:** $b_0 = 2{,}67$

  $b_1 = 0{,}67$

Klasický lineární regresní model (KLRM)

Gaussovy-Markovovy předpoklady

1. E($\boldsymbol{u}$) = $\boldsymbol{0}$
   – náhodné vlivy mají nulovou stření hodnotu (chyba není systematická)

2. E($\boldsymbol{u}\,\boldsymbol{u}^T$) = $\boldsymbol{\sigma}^2\,\boldsymbol{I}_n$; n – počet pozorování
   – kovarianční matice: rozptyl vícerozměrné náhodné složky
   – konečný a konstantní rozptyl na diagonále: homoskedasticita
   – náhodné složky jsou sériově (vzájemně) nezávislé: nuly mimo diagonálu (porušení: autokorelace)

3. $\boldsymbol{X}$ je nestochastická matice → E($\boldsymbol{X}^T\boldsymbol{u}$) = $\boldsymbol{0}$
   – veškerá „náhodnost“ $\boldsymbol{y}$ pochází z náhodné složky

4. $\boldsymbol{X}$ má plnou sloupcovou hodnost
   – Sloupce matice $\boldsymbol{X}$ (vysvětlující proměnné modelu vč. Sloupce konstanty) jsou vzájemně lineárně nezávislé → $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ existuje
   Porušení: perfektní či silná multikolinearita

Klasifikace vlastností odhadových metod:

Vlastnosti bodového odhadu, *n* < 30

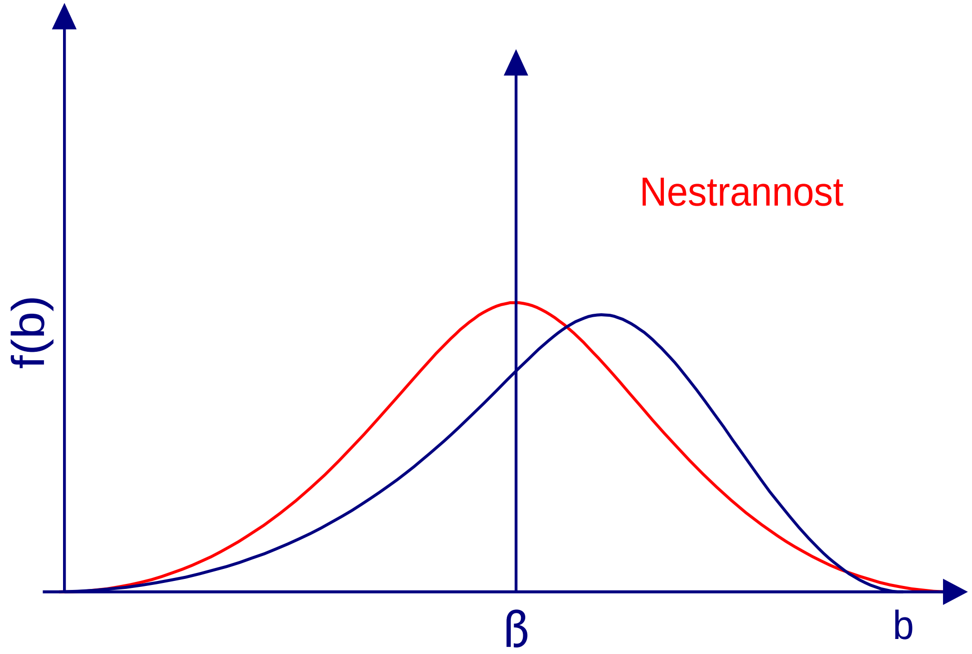
- **nestrannost (nevychýlenost / nezkreslenost)** odhadu:

  E(***b***) = ***β***

  b – získáme z více výběrových vzorků

  pokud E(***b***) > ***β*** – odhady jsou nadhodnoceny,

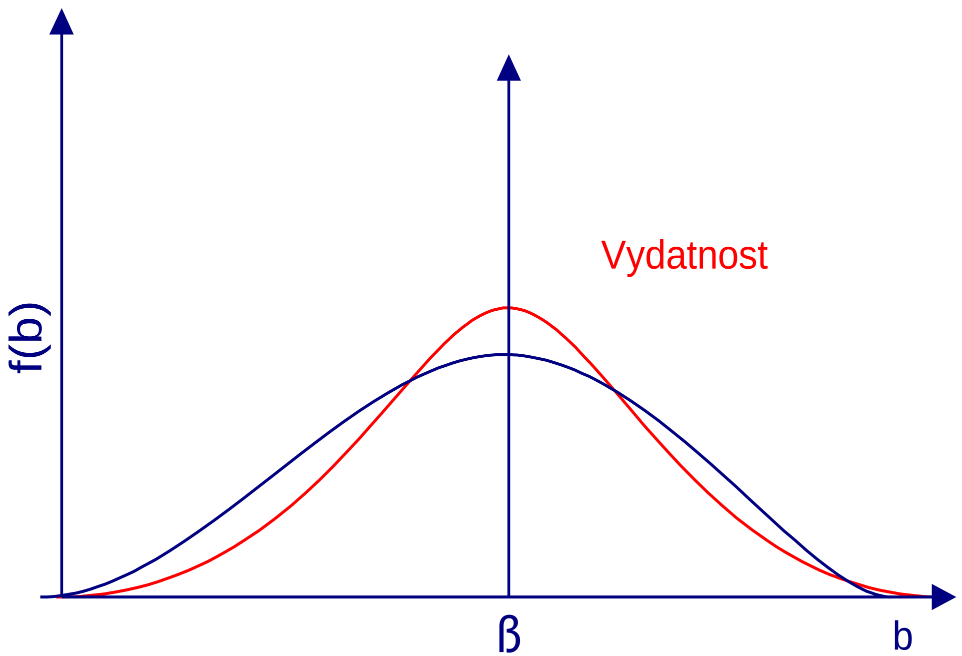
  E(***b***) < ***β*** – odhady jsou podhodnoceny



Klasifikace vlastností odhadových metod:

Vlastnosti bodového odhadu, *n* < 30

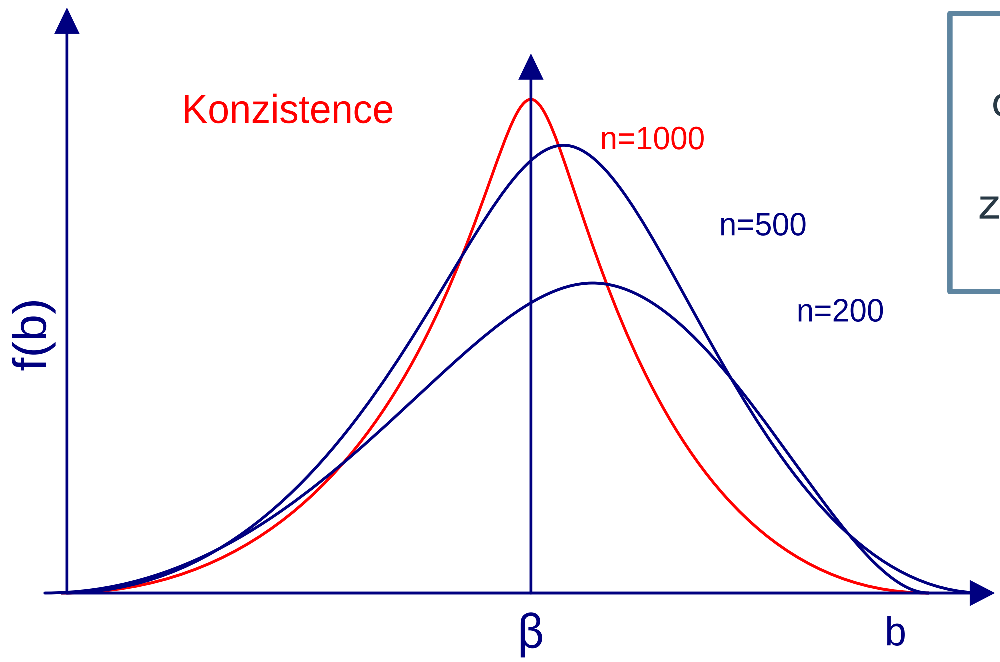
- **vydatnost** odhadu: standardní chyba regresního koeficientu s_b **musí být minimální** ze všech jiných postupů
  - to způsobuje, že intervalové odhady jsou nejmenší
  - jako nevychýlený odhad může sloužit více statistik, z nichž nejvhodnější je ta, která má minimální rozptyl



Klasifikace vlastností odhadových metod:

<u>Vlastnosti bodového odhadu, *n* ≥ 30</u>

- **konzistentní** – bodový odhad ***b*** je konzistentním odhadem, jestliže jeho hodnota s rostoucím počtem pozorování *n* konverguje ke skutečnému = populačnímu parametru:
- $\underset{n\to\infty}{p\,lim}\, b = \beta$, tj. pro $n \to \infty$: $b_j \to \beta_j \;\wedge\; VAR(b_j) \to 0$

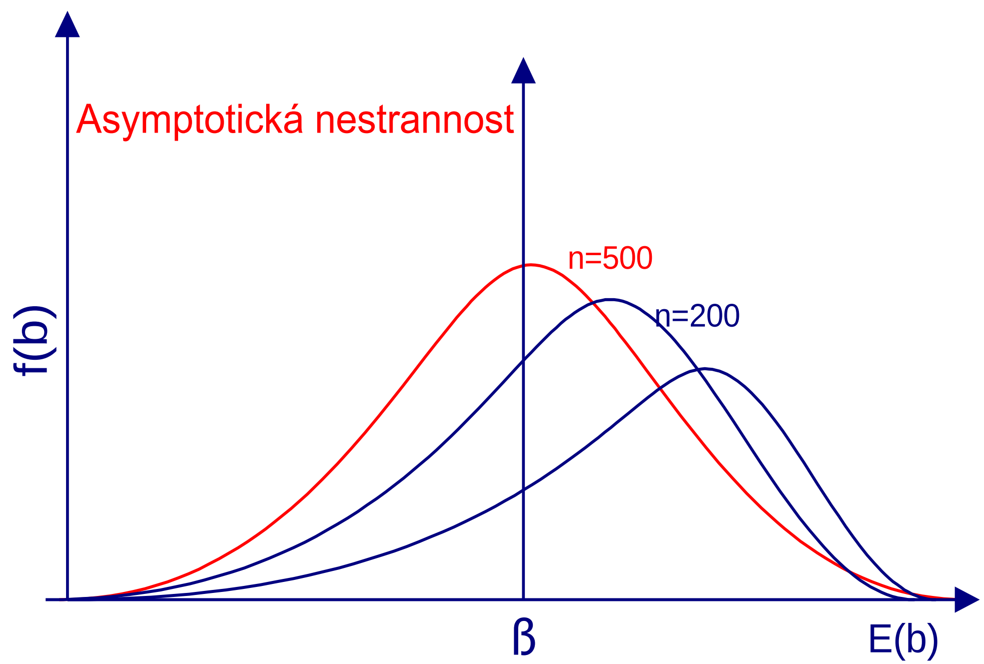


Pro konzistentní odhadové funkce růst *n* vede ke zlepšení vlastností odhadu.

Klasifikace vlastností odhadových metod:

Vlastnosti bodového odhadu, *n* ≥ 30

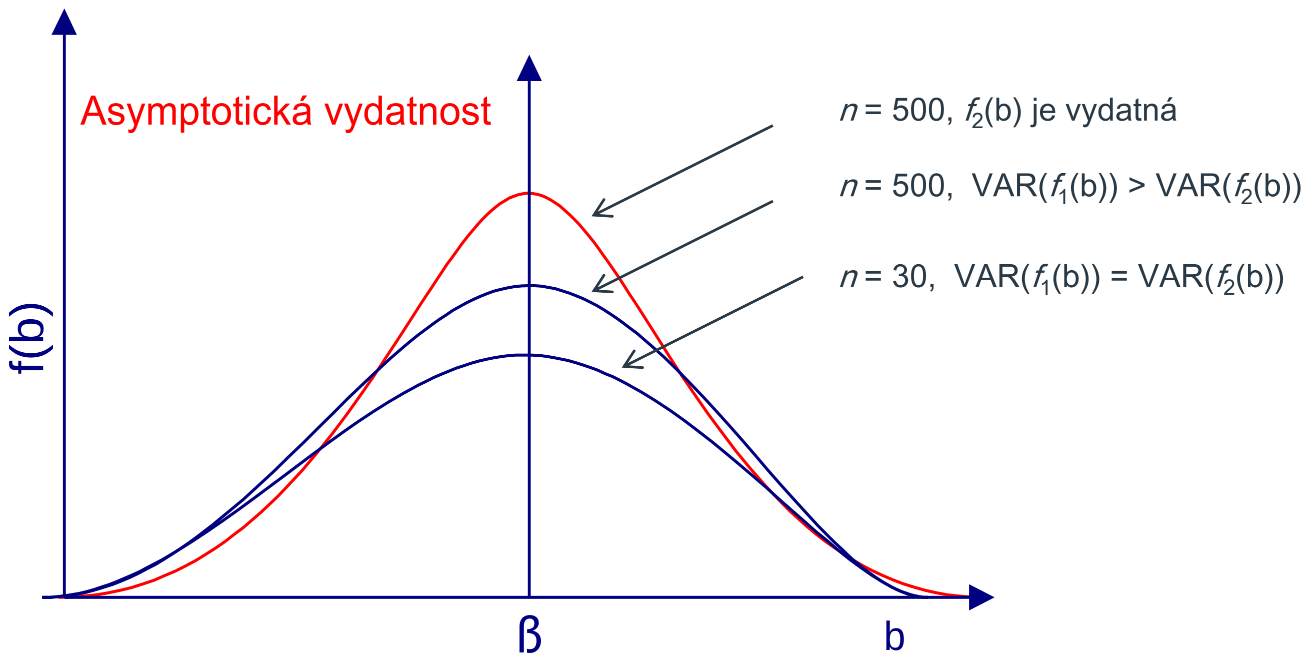
- **asymptoticky nestranný** – je to slabší vlastnost, (pokud je odhad konzistentní je vždy také asymptoticky nestranný)
- $\underset{n\to\infty}{p\,lim}\, E(b) = \beta$, tj. pro $n \to \infty$: $E(b_j) \to \beta_j \;\wedge\; VAR(b_j) \to k \neq 0$



Klasifikace vlastností odhadových metod:

Vlastnosti bodového odhadu, *n* ≥ 30

- **asymptotická vydatnost** – rozptyl konverguje k nule rychleji než s použitím jiné odhadové funkce



Klasický lineární regresní model (KLRM)

Vlastnosti bodového odhadu – maticové odvození

- Mějme KLRM: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$; $\boldsymbol{u} \sim N(\boldsymbol{0}; \sigma^2 \boldsymbol{I}_n)$
- Východisko pro popis teoretických vlastností (střední hodnoty a rozptylu) odhadové funkce MNČ:

$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$		odhadová funkce MNČ
$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X\beta} + \boldsymbol{u})$		dosadili jsme $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$
$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}$		roznásobena závorka
$= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}$		výsledný tvar po eliminaci: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_{nxn}$. Tento tvar je vhodný pro další popis vlastností MNČ.

Klasický lineární regresní model (KLRM)

Nestrannost bodového odhadu MNČ

$\boldsymbol{b} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}$	Obě strany výchozí rovnice vyjádříme jako střední hodnoty:
$E(\boldsymbol{b}) = E(\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u})$	Pro N.V. A a B platí: $E(A + B) = E(A) + E(B)$. Protože $\boldsymbol{\beta}$ je vektor skutečných (deterministických ale přímo nepozorovatelných) parametrů, lze psát $E(\boldsymbol{\beta}) = \boldsymbol{\beta}$:
$= \boldsymbol{\beta} + E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}]$	$\boldsymbol{X}$ je deterministická matice, opět můžeme využít $E(\boldsymbol{X}) = \boldsymbol{X}$ a přepíšeme:
$= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{u})$	V KLRM je střední hodnota náhodné složky nulová /nulový vektor/. Pro $E(\boldsymbol{u}) = \boldsymbol{0}$:
$= \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{0}$	Je zřejmé, že při násobení nulovým vektorem platí: $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{0} = \boldsymbol{0}$. Proto:
$E(\boldsymbol{b}) = \boldsymbol{\beta}$	Neboli, odhadová fce MNČ je nestranná (při dodržení uvedených podmínek, resp. Gaussových-Markovových předpokladů).

Klasický lineární regresní model (KLRM)

Rozptyl bodového odhadu MNČ

$\boldsymbol{b} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}$	Výchozí rovnici vyjádříme pro rozptyl:
$VAR(\boldsymbol{b}) = VAR(\boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u})$	Protože $\boldsymbol{\beta}$ je vektor deterministických (fakticky konstantních) parametrů, $VAR(\boldsymbol{\beta}) = \boldsymbol{0}$:
$= VAR[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{u}]$	Analogicky k $VAR(cX) = c^2 VAR(X)$ pro konstantu c a N.V. X , konstruujeme maticovou „kvadratickou formu“ typu: $VAR(\boldsymbol{Av}) = \boldsymbol{A}\, VAR(\boldsymbol{v})\, \boldsymbol{A}^T$, kde $\boldsymbol{A}$ je deterministická matice a $\boldsymbol{v}$ je vektor N.V.:
$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T VAR(\boldsymbol{u})[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T$	Pro rozptyl *iid* náhodné složky v KLRM platí $VAR(\boldsymbol{u}) = \sigma^2\boldsymbol{I}_n$:
$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T$	Výraz v hranaté závorce upravíme pomocí věty o transpozici součinu matic: $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$
$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\,\boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^T$	Protože $\boldsymbol{X}^T\boldsymbol{X}$ je symetrická matice, pro $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ platí: $[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.
$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\,\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	$\boldsymbol{I}_n$ vypustíme (násobení jednotkovou maticí nemá vliv na výsledek). Protože $c\cdot\boldsymbol{A} = \boldsymbol{A}\cdot c$, lze rozptyl σ^2 (číslo) předsunout ve výrazu před matice:
$= \sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	Protože $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_n$, tento výraz opět vypustíme:
$VAR(\boldsymbol{b}) = \sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	Je kovarianční matice udávající rozptyl vektoru $\boldsymbol{b}$ (vektor bodových odhadů b_j).

Klasický lineární regresní model (KLRM)

Rozptyl bodového odhadu MNČ

- Pro KLRM: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$; $\boldsymbol{u} \sim N(\boldsymbol{0}; \sigma^2 \boldsymbol{I}_n)$ tedy platí:

  $\boldsymbol{b} \sim N(\boldsymbol{\beta}; \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ … při odhadu MNČ

- Protože neznáme skutečný rozptyl u_j (resp. nepozorovatelné náhodné složky $\boldsymbol{u}$), tj.: σ^2, pracujeme s odhadnutým rozptylem, vypočteným pomocí reziduální složky odhadnutého modelu:

$$\text{odh. } \sigma^2 = s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-k-1}$$

kde $n-k-1$: počet stupňů volnosti odhadu MNČ,
n: počet pozorování,
$k+1$: počet regresních parametrů LRM (včetně konst.)