Přednáška II
AKM I

Lukáš Frýd

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

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY

Víme, jak odhadnout neznámé parametry např. OLS

Jak ale můžeme vědět, že se náš odhad „pohybuje blízko“ populačnímu parametru, když jej neznáme?

Proto potřebujeme zavést předpoklady
Pro metodu nejmenších čtverců si zavádíme tzv. Gauss-Markov předpoklady

Zároveň si zavedeme i předpoklad o chování náhodné složky a získáme tak klasický lineární model (KLR)

V případě splnění předpokladů KLR budeme vědět, že metoda nejmenších čtverců je estimátor:

- Nevychýlený
- Konzistentní
- Vydatný

A budeme znát rozdělení odhadů parametrů („béčka“)

Předpoklady klasického lineárního modelu pro průřezová data

LRM 1. Předpokládáme linearitu v parametrech, tedy $y = X\beta + \epsilon$
To zároveň znamená, že máme správně specifikovaný tvar modelu

LRM 2. Výběrový vzorek je NÁHODNÝM výběrem z populace o rozsahu $n > k + 1$. Jednotlivé prvky ve výběrovém souboru na sobě nezávisí

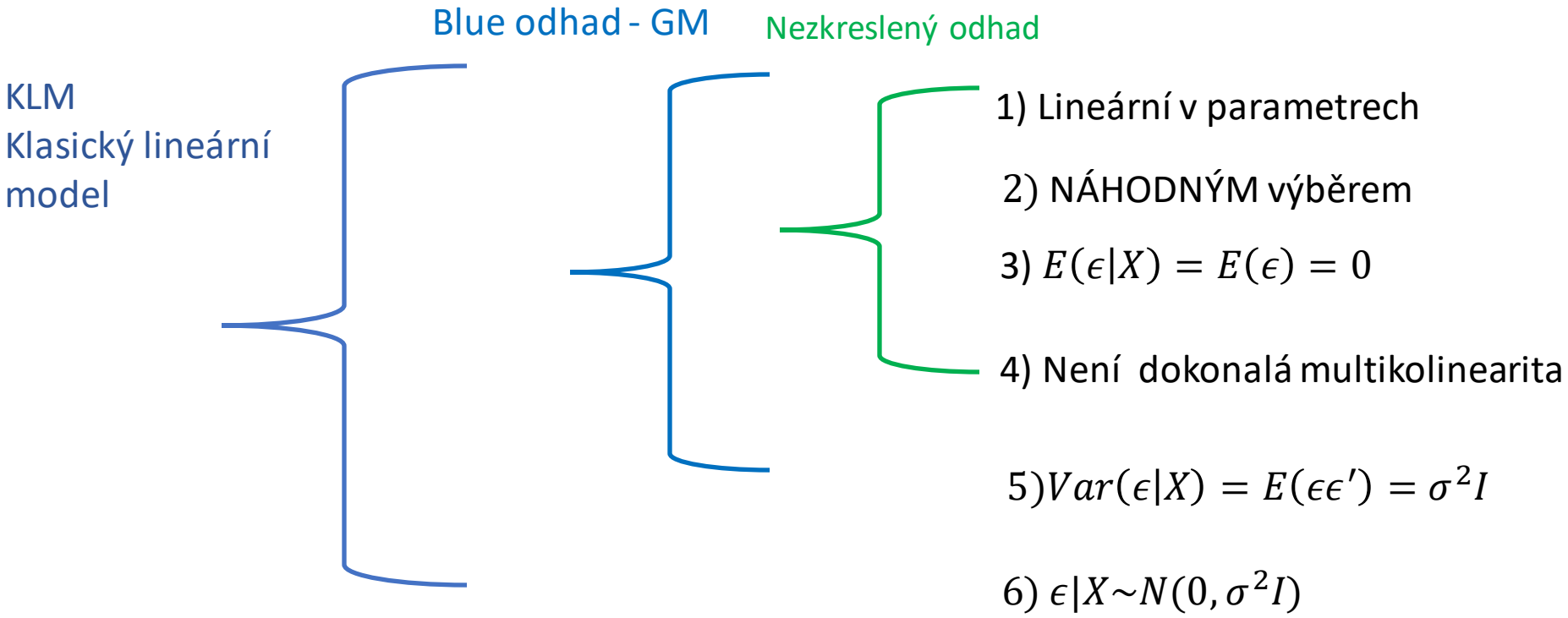
LRM 3. Střední hodnota náhodné složky nezávisí na hodnotách vysvětlujících proměnných $E(\epsilon|X) = E(\epsilon) = 0$

LRM 4. Matice **X** má plnou hodnost. Tedy ve výběrovém souboru nelze vyjádřit některou z proměnných jako lineární kombinaci dalších proměnných. Nebude dokonalá multikolinearita

LRM 5. Rozptyl náhodné složky je konečný a nemění se v závislosti na X, $Var(\epsilon|X) = \sigma^2 I$, kde σ^2 je konstanta. Homoskedasticita

LRM 6. Náhodná složka se řídí normálním rozdělením. (podmíněným)
Spolu s předpoklady 3 a 5, tak platí, že $\epsilon|X \sim N(0, \sigma^2 I)$

Přepoklady KLM a Gauss – Markov teorém



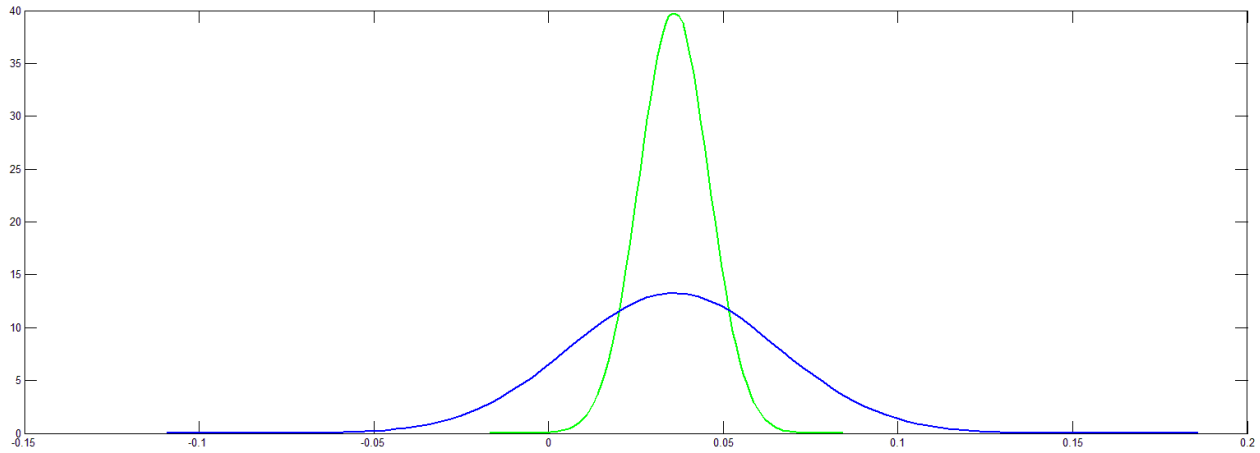
OLS je tak nezkreslenou, konzistentní a vydatnou statistikou (estimátorem)
Na co ale nezapomínat - máme pouze bodové odhady

OLS je BLUE

Best linear unbiased estimator

Po splnění GM již nemusíme hledat jiný estimátor v množině lineárních estimátorů. Jelikož víme, že buď najdeme stejně vydatný, nebo méně vydatný.

$$Var(b|X) < Var(\hat{b}|X)$$



LRM 1. Předpoklad linearity v parametrech

Cobb-Douglas produkční funkce
Není lineární v parametrech
f(..) – nelineární funkce

$$Y = A.K^{\beta_1}.L^{\beta_2} \qquad A = e^{\beta_0+\varepsilon}$$

$$y = f(\beta_0) + f(\beta_1)x + \varepsilon \qquad y = \frac{1}{\beta_0 + \beta_1 x} + \varepsilon$$

V parametrech je lineární

$$lnY = \beta_0+\beta_1 lnK + \beta_2 lnL + \varepsilon$$

$$y = \beta_0+\beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$y = \beta_0 + \beta_1 f(x) + \varepsilon$$

$$y = X\beta + \varepsilon$$

$$lnY = y$$

$$lnK = x_1$$

$$lnL = x_2$$

1) Lze vyřešit analyticky (na papíře)
2) Testy hypotéz – lineární kombinace náhodných veličin řídících se normálním rozdělením je stále náhodná veličina s normálním rozdělením

LRM2. Výběrový vzorek je NÁHODNÝM výběrem z populace

$$E(\varepsilon_i \varepsilon_j | X) = 0$$

$$Corr(\varepsilon_i, \varepsilon_j | X) = 0 \; kdy \; i \neq j \qquad Corr(\varepsilon_i, \varepsilon_j) = 0 \; kdy \; i \neq j$$

o rozsahu $n > k + 1$

Jaký proces generuje cenu nemovitostí v Praze? Co když vyberu data jen z Ořechovky?

Problém nastává například u makroekonomických panelů. To se učí u nás na KEKO ☺

LRM 3. Střední hodnota náhodné složky nezávisí na hodnotách vysvětlujících proměnných

$$E(\epsilon|X) = E(\epsilon) = 0$$

Ukažme si co daná podmínka může znamenat:

$$E(\epsilon|X) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k$$

Kdybychom mohli pozorovat náhodnou složku, tak bychom mohli provést test:

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

Při nezamítnutí H0 bychom tak uvažovali, že:

$$E(\epsilon|X) = \alpha_0$$

Hodnota α_0 by byla střední hodnota náhodné složky $E(\epsilon) = \alpha_0$
V případě, že by se $\alpha_0 \neq 0$ nás už nemusí trápit, jelikož úrovňová konstanta nám zaručí, že $E(\epsilon) = 0$

Všechny vysvětlující proměnné (x) a jsou nekorelované s náhodnou chybou

Příklad regresní rovnice pro náhodnou složku byl však pouze jednou z možností.

$$E(\epsilon|X) = E(\epsilon) = 0$$

Nám přesně říká, že všechny vysvětlující proměnné (x) a jejich funkce jsou nekorelované s náhodnou chybou

$$E(\epsilon|X) \neq m(x) \qquad m(x) - je\ libovolná\ funkce, tedy\ nejenom\ lineární!$$

V případě, že se vzdáme požadavku na zkreslenost odhadu,
tak konzistenci nám zaručí podmínka, že $E(\epsilon' X) = 0$
Tedy, že náhodná složka není korelovaná s některým z regresorů.
Korelace postihuje pouze lineární vztahy, jedná se tedy o slabší předpoklad, než $E(\epsilon|X)$

S tímto se setkáte například u časových řad.
Nelze totiž vždy zajistit nevychýlený odhad, požadujeme však konzistentní!

Omitted variable biased

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v \qquad v \sim N(0, \Sigma_v)$$

$$x_2 = \alpha_0 + \alpha_1 x_1 + u \qquad u \sim N(0, \Sigma_v)$$

Co se stane, pokud nezahrneme x_2 do regrese. Budeme si myslet, že DGP vypadá následovně:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \qquad \epsilon = \beta_2 x_2 + v$$

$$b_1 = \beta_1 + x_1 \epsilon$$

$$b_1 = \beta_1 + x_1(\beta_2 x_2 + v)$$

$$E(b_1|x_1) = \beta_1 + E(x_1 v|x_1) + \beta_2 E(x_2 x_1|x_1)$$

$$E(b_1|x_1) = \beta_1 + x_1 \beta_2 E(x_2|x_1)$$

$$E(x_2|x_1) = ?$$

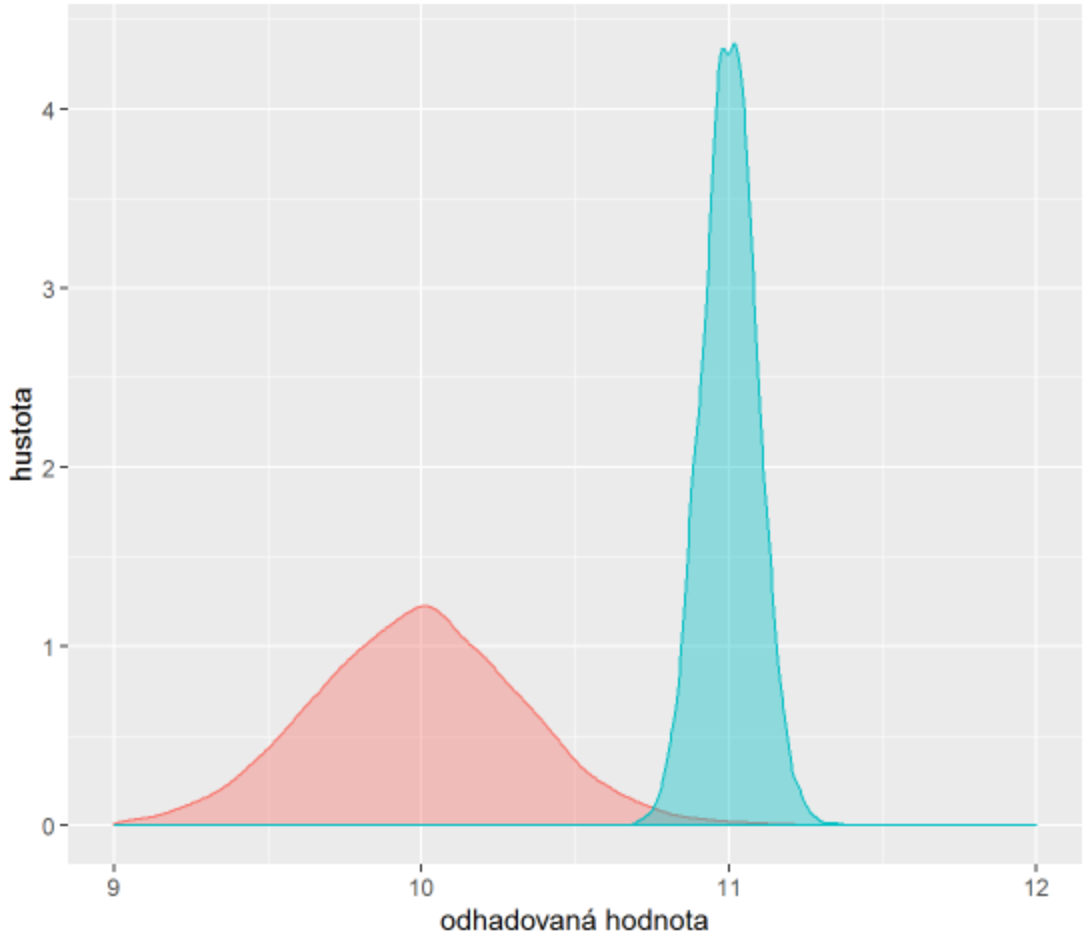
$$plim\ \widehat{b_1} = \beta_1 + \beta_2 \gamma_1$$

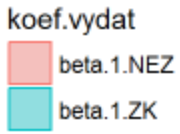
$$\gamma_1 = cov(x_1, x_2)/Var(x_1)$$

$$cov(x_1, x_2) \neq 0$$

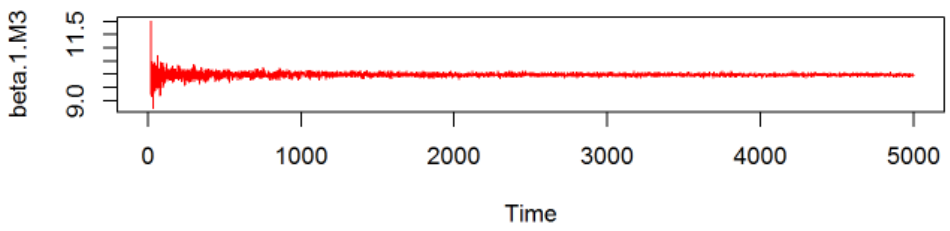
DOPLNĚK R-ko I

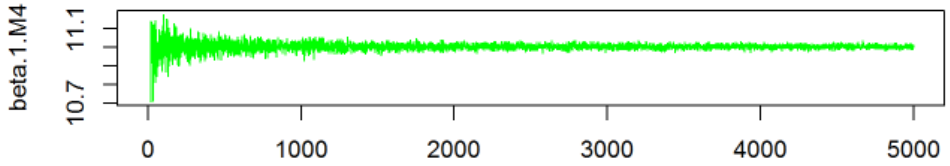
```
N=10000
v=rnorm(N,0,1)
w=rnorm(N,0,1)
IQ=100+10*v  # IQ obsahuje  slozku v, která je i v náhodné  sloZce
educ=runif(N,  min=9, max=20)
expert=runif(N,  min=0, max=30)
 e =0+10*v+10*sqrt(1-0.8^2)*w  # vygenereujeme  náhodnou  sloZku, která je korelovaná s náhodnou  slozkou v promenné  IQ
wage2  = -50  +10*IQ+  50*educ+3*expert  + e
```





Ukázka jak vypadá nevychýlený a konzistentní odhad
Vs.
Vychýlený a nekonzistentní odhad





LRM 4. Lineární nezávislost nezávislých proměnných

Žádnou vysvětlující (nezávislou) proměnnou nemůžeme napsat jako lineární kombinaci jiných proměnných
Mluvíme o perfektní kolinearitě

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

$$příjem = mzda + kap.vynos$$

$$y = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \beta_3 . x_3 + \varepsilon$$

$$x_3 = x_1 + x_2$$

X3 získáme jako lineární kombinaci X1 a X2

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 4 & 2 \end{pmatrix} . \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 4 \\ 1 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 & 9 \\ 6 & 12 & 18 \\ 9 & 18 & 29 \end{pmatrix}$$

$$\begin{pmatrix} 3 & 6 & 9 \\ 6 & 12 & 18 \\ 9 & 18 & 29 \end{pmatrix}^{-1} = nelze, singularni\ matice$$

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

Multikolinearita

$$příjem = mzda + kap.vynos$$

$$x_3 = x_1 + x_2$$

Někdy jen kolinearita
Nezávislé proměnné jsou „těsně“ **lineárně** spojené
Vysoký stupeň korelace

$$y = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \varepsilon$$

$$Corr(x_1, x_2) - "vysoká"$$

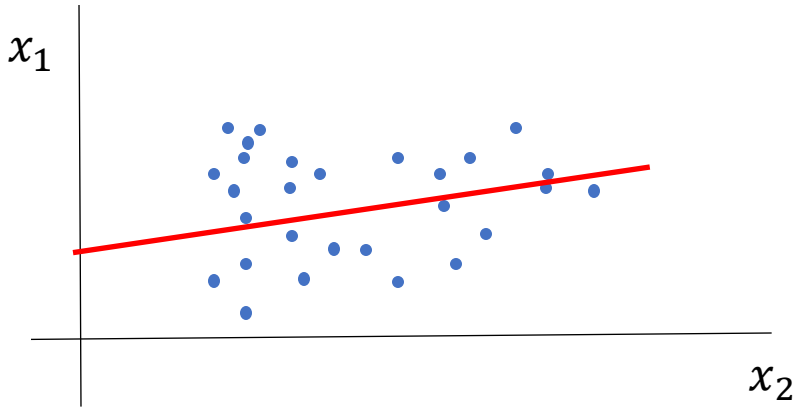
Kolinearita pro 2 proměnné
Multikolinearita pro více jak 2 proměnné
Pro jednoduchost budeme říkat multikolinearita

Třeba rozlišovat **perfektní multikolinearitu** (kolinearitu) a **multikolinearitu** (kolinearitu)
Perfektní multikolinearita je porušení GM, multikolinearita NENÍ!

Nelze jasně definovat kdy nastává multikolinearita

Příčiny:

- Časové řady a společný trend
- Zpožděné proměnné



Pozor R_j^2 není R^2 z výstupu programu!!

Čím větší bude R_j^2 - tím větší bude rozptyl odhadu

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

Multikolinearita problém

Ale nezapomínat na problém s malým výběrovým vzorekem!!

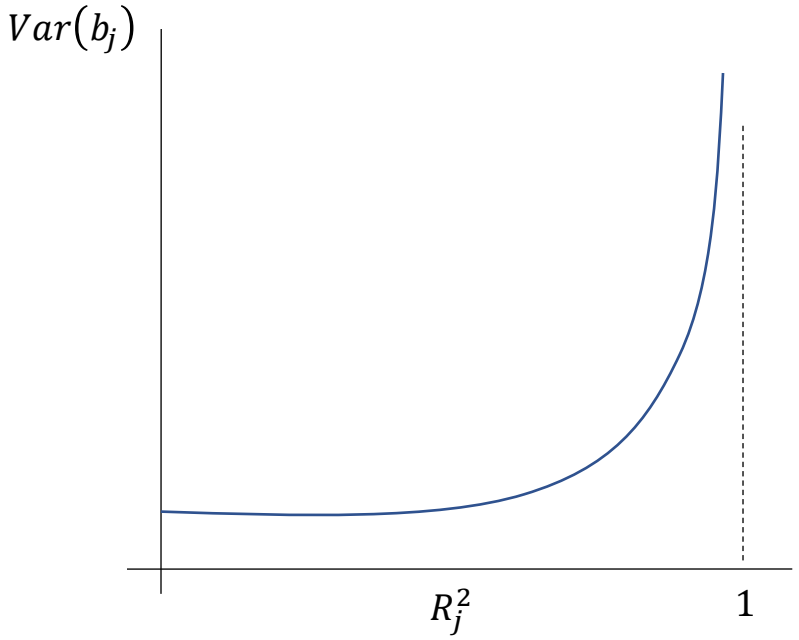
$R_j^2 - R^2 z\ regrese\ x_j\ na\ všechny\ ostatní\ X + intercept$

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

Interpretace R^2 - kolik variability v x_1

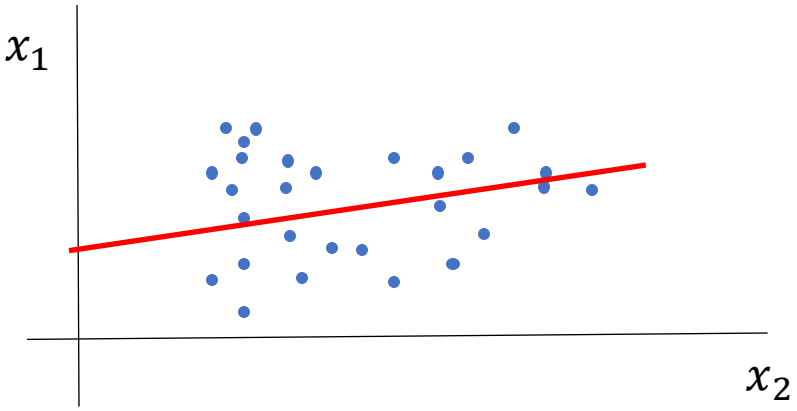
Se nám podařilo vysvětlit pomocí x_2

Čím více – tím větší je vzájemný vztah



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

Problém u testování hypotéz



Pokud existuje silná závislost mezi NP – nemůžeme izolovat jednotlivé efekty

Vidíme pouze společné působení

Detekce

Většinou se nejedná o vlastnost základního souboru, ale výběrového
Neděláme statistické testy – pouze jí měříme

Vždy najdete určitou kovarianci – otázkou je jak je vysoká

Korelační matice – částečně
Pokud bude vysoké R^2 a odhady parametrů nesignifikantní – zbystřit
Špatné znaménko, nebo „vysoká“ hodnota parametru (teorie)
VIF-Variance inflation factor

$$y = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \beta_3 . x_3 + \epsilon$$

$$Cov(x_1, x_2) - relativně\ nízká$$

$$x_1 = \gamma_0 + \gamma_2 . x_2 + + \gamma_3 . x_3 + v$$

$$R_1^2 \to 1$$

$$multikolinearita!$$

Variance inflation factor

$$y = \beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \beta_3 . x_3 + \epsilon$$

$$VIF = \frac{1}{1 - R_j^2}$$

R_1^2 $\quad x_1 = \gamma_0 + \gamma_2 . x_2 + \gamma_3 . x_3 + v$

R_2^2 $\quad x_2 = \gamma_0 + \gamma_1 . x_1 + \gamma_3 . x_3 + v$

R_3^2 $\quad x_3 = \gamma_0 + \gamma_1 . x_1 + \gamma_2 . x_2 + v$

$$VIF = \frac{1}{1 - R_j^2} \geq 5$$

Vysoké R2 představuje
Vysokou úroveň multikolinearity

VIF ≥ 5 se obvykle považuje za neúnosnou multikolinearitu

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

Řešení

Zamyslet se nad modelem ☺
Vynecháme některou z proměnných?
Pokud je populační model správně,
vypuštění relevantní proměnné povede ke zkresleným odhadům

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v \qquad x_2 = \alpha_0 + \alpha_1 x_1 + u$$

$$y = \beta_0 + \beta_1 x_1 + \epsilon \qquad \epsilon = \beta_2 x_2 + v$$

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

Vynechání silně korelované proměnné – pokud to jde a má to smysl

Zvýšit výběrový soubor (sample) více dat + snížení rozptylu *Var(b)*

Shrnout proměnné do 1 proměnné – faktor pokud to jde

Pokud se zdají být t-test a odhady „v normě“ – spíše ponechat tvar modelu

F-test

Pro hodnocení významnosti jednotlivých parametrů použijeme ***t-test***
Pro hodnocení významnosti 2 a více parametrů ($\beta_{0,1,..,k}$) (**NAJEDNOU** (i celého modelu) použijeme **F-test**
Zdali **množina nezávislých proměnných** má parciální efekt na (y)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

$H0: \beta_2 = 0\,, \beta_3 = 0, \beta_4 = 0$ **Restrikce nemá efekt na (y)**

$H1: neplatí\ H0$

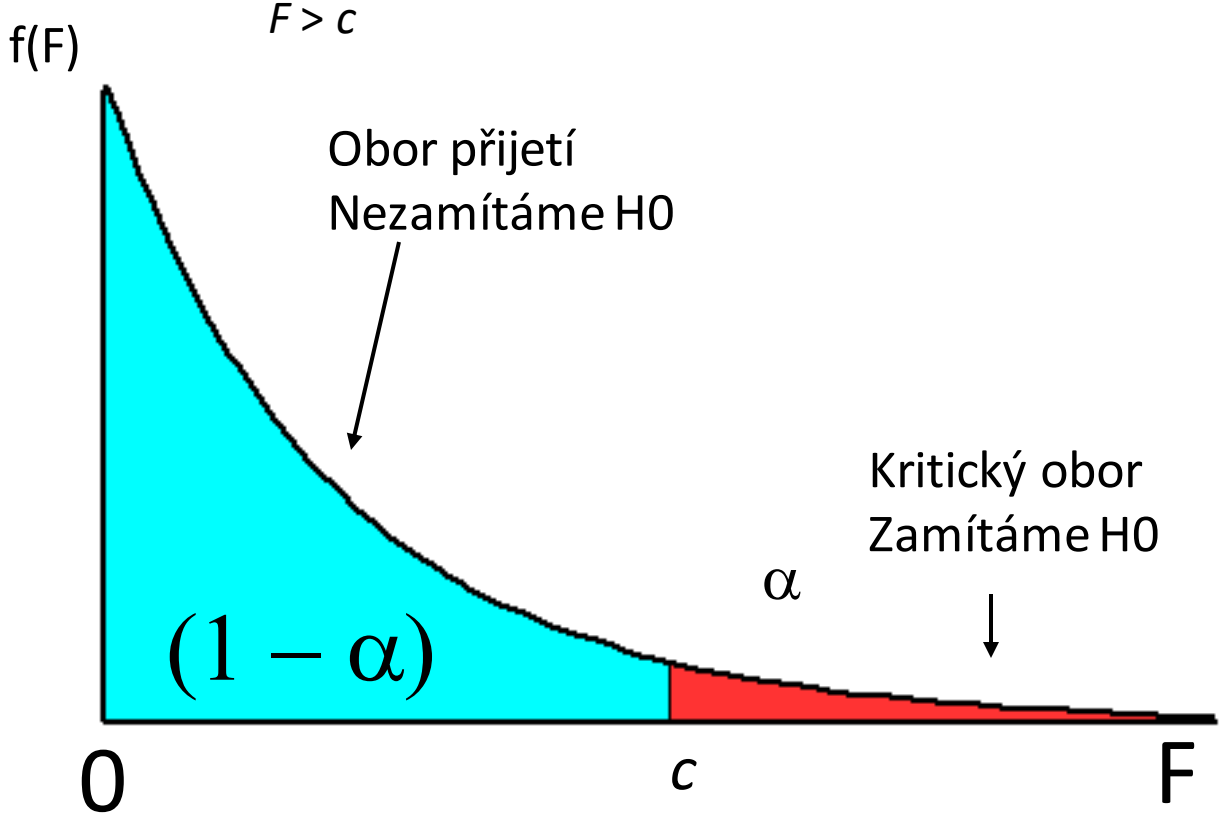
Omezený vs. Neomezený model

$$y = \beta_0 + \beta_1 x_1 + \beta_5 x_5 + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Nebudeme konkretizovat H1
Pokud nebude naplněna **ALESPOŇ** jedna restrikce
některá beta $\neq 0$
Zamítáme H0

Zamítneme H0 na α hladině významnosti pokud
F > c



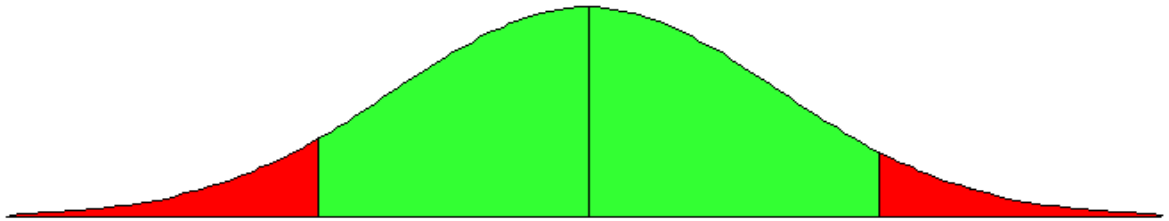
Nešvar 1

Regrese s „velkým množstvím“ proměnných

Ztrácíme stupně volnosti!

Podívejte se jak vypadá 95% kvantil pro $t(5)$ a $t(30)$

Více na cvičení



$$t_i = \frac{b_i - \beta_i}{se\,(b_i)} \sim t(n-k-1)$$

Hladina významnosti α

$$P(zamítneme\ H_0 | H_0\ platí) = \alpha$$

Na co nezapomínat

Multikolinearita nemá dopad na nezkreslenost odhadu parametrů E(b)
Po splnění všech GM- odhad bude BLUE
Určitá multikolinearita bude téměř vždy
Problém spíše výběrového souboru, než populace

Problém ve vysokém rozptylu odhadu
Testy hypotéz budou naznačovat nesignifikantní parametr – přestože vše daný parametr bude součástí reg. fce
Může nás tak „navést“ na nesprávný model

4) Žádná vysvětlující proměnná není lineární kombinací jiných VP

$$sav = \beta_0 + \beta_1 . inc + \beta_2 . inc^2 + \epsilon$$

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

LRM 5. Homoskedasticita

Více příští hodinu

Nás samozřejmě zajímá jaký rozptyl má výběrové rozdělení pro b
A tedy i rozptyl. Odhad rozptylu je dán:

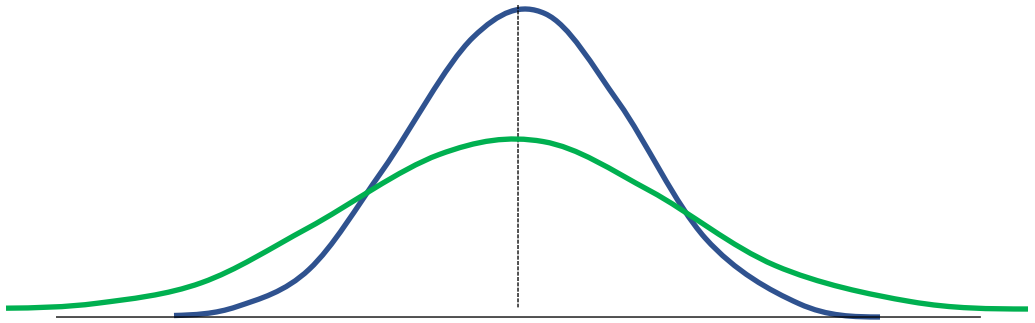
$$Var(b_j) = \frac{\sigma^2}{\sum(x_j - \bar{x_j})^2 (1 - R_j^2)}$$

σ^2 –rozptyl náhodné složky, neznám, musím také odhadnout

$$\widehat{\sigma^2} = \frac{1}{n-k-1} \sum e^2$$

Konzistentní odhad rozptylu b_j při platnosti GM je pak:

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



$E(b_j) = \beta_j$

$E(b_j) = \beta_j$

Požaduji, aby platilo: $E(\widehat{\sigma^2}) = \sigma^2$

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

Odhad rozptylu náhodné složky a $Var(b)$ je nezkreslený a konzistentní pokud platí $Var(\epsilon|X) = \sigma^2 I$
Tedy požaduji homoskedasticitu

Rozptyl $Var(b_j)$ se vyskytuje jak u t-testu, tak u testů pro hodnocení sdružené významnosti (F, Wald)
Tedy pokud není splněn předpoklad homoskedasticity, tak nelze brát výsledky těchto testů jako relevantní!!!

Gauss-Markovovův teorém

1) Lineární v parametrech

2) Výběrový vzorek je NÁHODNÝM výběrem z populace

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

4) Není dokonalá multikolinearita

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

Při splnění těchto 5 předpokladů
Získáme pomocí metody OLS tzv. BLUE odhad
Best Linear Unbiased Estimator

NIC NEŘÍKÁME O TYPU ROZDĚLENÍ NÁHODNÉ SLOŽKY!!!!

LRM 6. Náhodná složka se řídí normálním rozdělením

Vlastnosti pro „malé“ vzorky (finite sample properties)
Vlastnosti pro „velké“ vzorky (large sample properties)

Nás zajímá jak vypadá výběrové rozdělení odhadu β
Pro (finite sample properties) proto musíme zavést předpoklad o tvaru rozdělení náhodné složky
Pokud platí pro náhodnou složku, že $\epsilon|X \sim N(0, \sigma^2 I)$, tak potom:

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

V případě (large sample properties) již nemusíme zavádět takto silný předpoklad (v praxi často nereálný), jelikož můžeme použít některou z centrálních limitních vět.

Důsledky normality náhodné složky

$$t_i = \frac{b_i - \beta_i}{se(b_i)} \sim t(n-k-1)$$

t-test bude mít díky normalitě náhodné složky právě studentovo rozdělení
To samé platí pro F-test, který bude mít právě Fisherovo rozdělení

Pokud nebude mít náhodná složka normální rozdělení, tak testy budou mít pouze asymptotické *F,t* rozdělení

Což může být problém v případě, kdy máte málo pozorování! A viz také ztráta stupňů volnosti.

Graf R

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