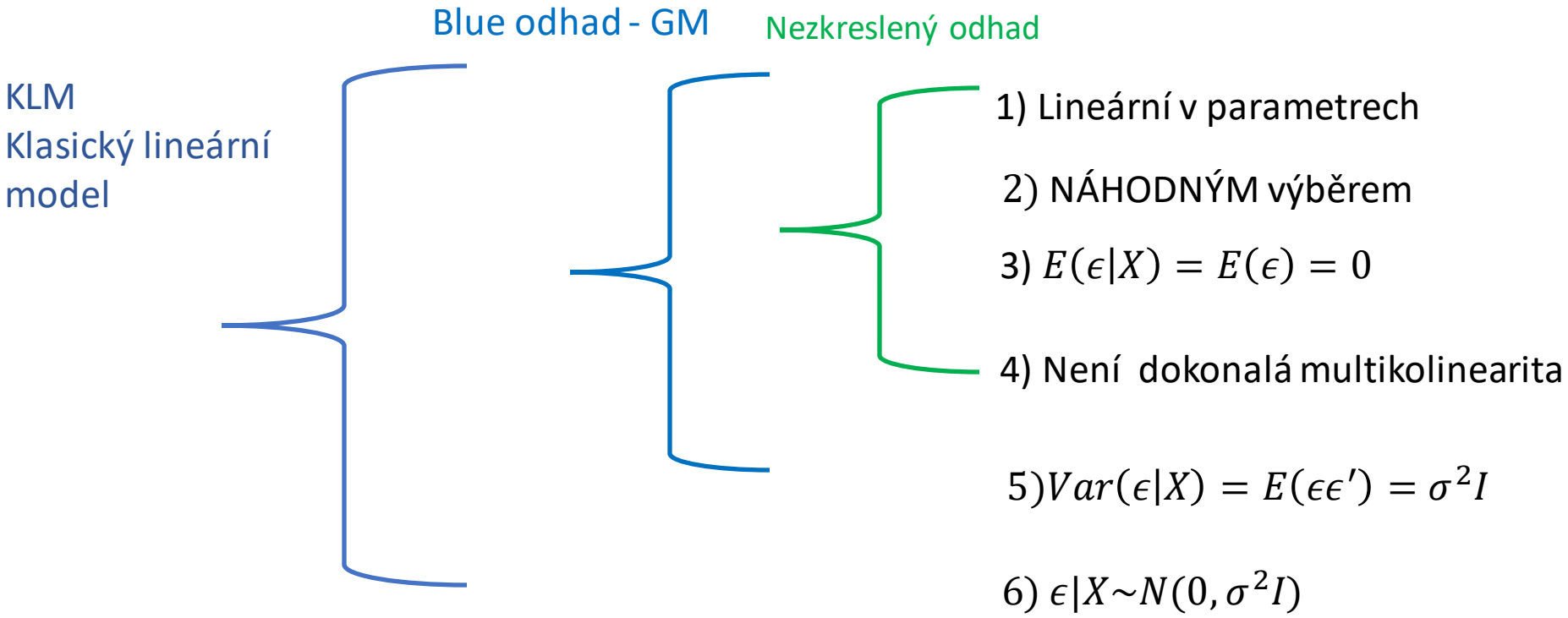
Přednáška VII
AKM I

Lukáš Frýd

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

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Přepoklady KLM a Gauss – Markov teorém



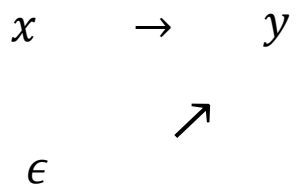
Potom platí, že:

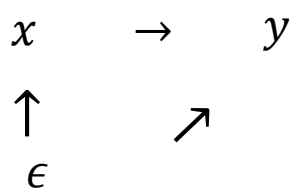
$$b \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Problém endogenity

Možné příčiny

- Vynechání proměnné, např. nepozorujeme jí (omitted variable biased)
- Simultálnost (vzájemné propojení)
- Chyby v měření
- Dynamický model – časové řady a zpožděná závisle proměnná + autokorelovaná náhodná složka





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:

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

$b_1 = \beta_1 + x_1 \epsilon$

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

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

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

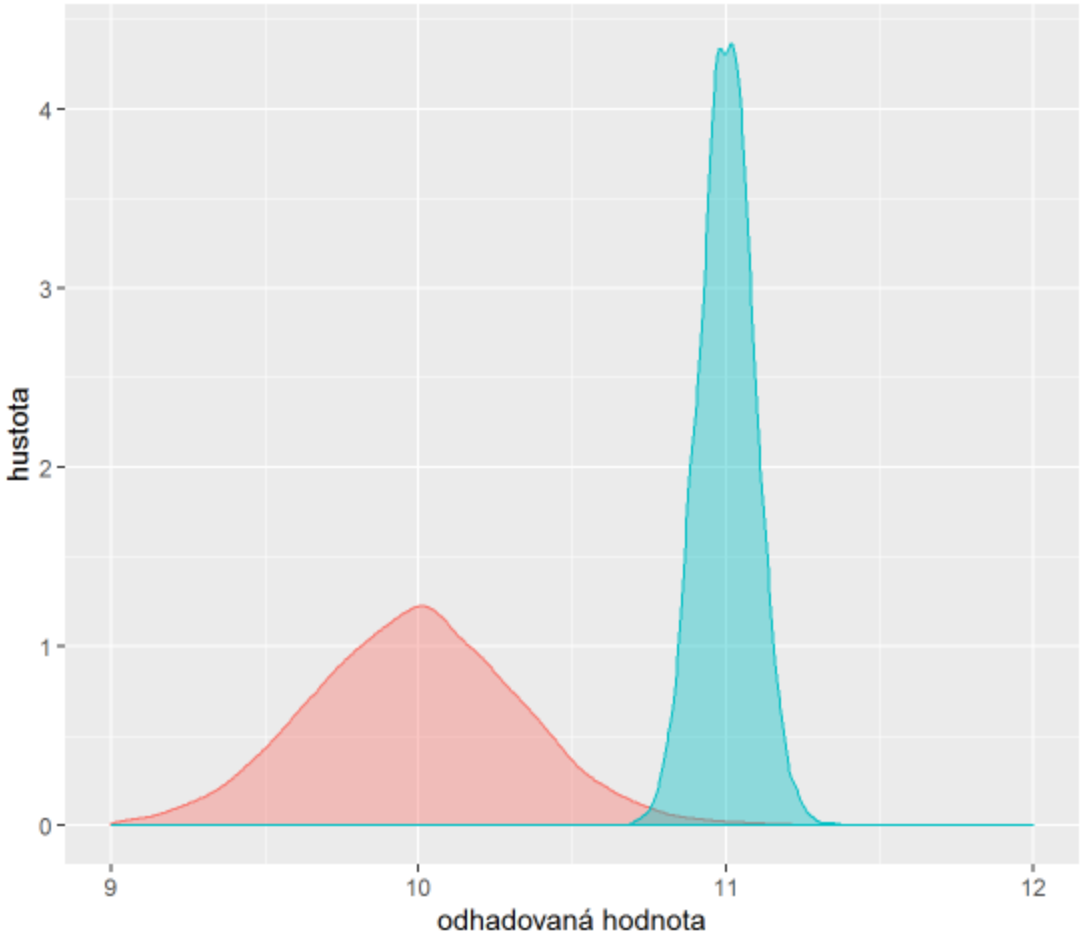
$E(x_2|x_1) = ?$

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

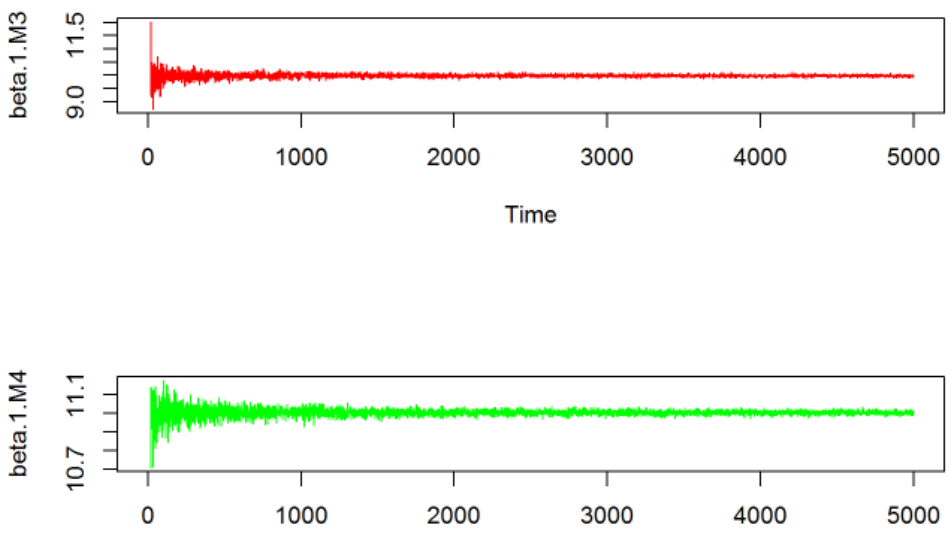
$\gamma_1 = cov(x_1, x_2)/Var(x_1)$

$cov(x_1, x_2) \neq 0$

Dopady endogenity



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



Jaké máme možnosti řešení?

Co vlastně znamená endogenita proměnné.
Proměnná je vysvětlena v rámci modelu

Obecné řešení problému s endogenitou a konzistentní odhady

1) Specifikovat dodatečnou rovnici(e) pro endogenní proměnné
2) Použití instrumentální proměnné

$$crime\_rate = \beta_0 + \beta_1 PolPC + \beta_3 IncPC + \epsilon$$

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 47.1123513 18.4730418  2.5503 0.0124718 *
polpc       18.4786908  4.5374153  4.0725 0.0001006 ***
pcinc        0.0013534  0.0015275  0.8860 0.3779959
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$PolPC = \alpha_0 + \alpha_1 crime_{rate} + .. + \epsilon$$

```
library(foreign)
download.file('http://fmwww.bc.edu/ec-
p/data/wooldridge/crime2.dta','crime2.dta',mode='wb')
crime2 <- read.dta('crime2.dta')
attach(crime2)
regrese=lm(crmrte~polpc+pcinc)
library(sandwich)
library(lmtest)
coeftest(regrese, vcov = vcovHC(regrese, type = "HC1"))
```

Vzájemná vazba – model simultánní rovnice

$y_1 = \alpha_1 y_2 + \beta_1 z_1 + \epsilon_1$ $\quad y_2 = \alpha_2 y_1 + \beta_2 z_2 + \epsilon_2$

$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + \epsilon_1) + \beta_2 z_2 + \epsilon_2$

$y_2 = \alpha_2\alpha_1 y_2 + \alpha_2\beta_1 z_1 + \boldsymbol{\alpha_2\epsilon_1} + \beta_2 z_2 + \epsilon_2$

$\boldsymbol{Cov(y_2, \epsilon_1) \neq 0}$

Endogenní proměnné

$Cov(y_1, \epsilon_2) \neq 0$

Není splněno GM

$Cov(z_1, \epsilon_1) = 0$

$Cov(z_2, \epsilon_2) = 0$

Exogenní proměnné

„**Simultaneity bias**“
Odhad bude zkreslený a nebude konzistentní

$y_2 = \alpha_2\alpha_1 y_2 + \alpha_2\beta_1 z_1 + \boldsymbol{\alpha_2\epsilon_1} + \beta_2 z_2 + \epsilon_2$ $\quad Cov(z_2, \epsilon_2) = 0$ $\quad Cov(z_1, \epsilon_2) = 0$ $\quad Cov(y_2, \epsilon_2) = 0$

Mezi proměnnými již neexistuje zpětná vazba
všechny endogenní proměnné jako funkce pouze predeterminovaných proměnných

Jednoduchý keynesiánský model

$$C_t = \beta_0 + \beta_1 Y_t + \epsilon_{t1}$$

$$Y_t = C_t + I_t$$

Máme 2 endogenní proměnné C_t, Y_t
A 1 exogenní I_t- předpokládáme, že je exo.

Při odhadu spotřební fce pomocí OLS, nedostaneme konzistentní odhady

$$C_t = \beta_0 + \beta_1 (C_t + I_t) + \epsilon_{t1}$$

$$C_t = \frac{\beta_0}{1-\beta_1} + \frac{\beta_1}{1-\beta_1} I_t + \frac{\epsilon_{t1}}{1-\beta_1}$$

Nyní již můžeme použít OLS

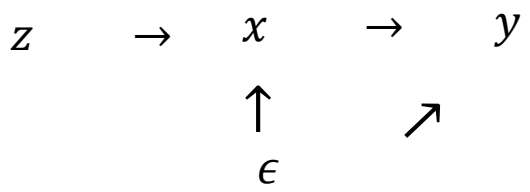
Instrumentální proměnná

Představme si, že najdeme proměnnou z, pro kterou platí:

1) Je korelovaná s x - $E(z'x) \neq 0$
2) Je exogenní - $E(z'\epsilon) = 0$

Tuto proměnnou *z* nazveme instrumentální proměnnou a nebo instrumentem

Podmínka 1) se nazývá podmínka relevance
Podmínka 2) se nazývá podmínkou exogenity



Zatím předpokládejme, že máme 1 instrument a 1 endogenní regresor

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + \epsilon$$

abil nejsme schopni měřit a dá se předpokládat, že $Corr(educ, abil) > 0$

Pokud vynecháme $abil$

$$\log(wage) = \beta_0 + \beta_1 educ + \epsilon$$

Dostaneme nekonzistentní odhady

Nyní je třeba rozlišit dva druhy proměnných. Proxy proměnná a instrumentální proměnná.
Proxy proměnnou použijeme, pokud nejsme schopni získat data o regresoru, zde abil.
Požadujeme, aby byla proxy proměnná s abil. korelovaná.
To je například IQ

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 IQ + \epsilon$$

Bude IQ i vhodná IV proměnná?

Podmínka 1) se nazývá podmínka relevance – OK
Podmínka 2) se nazývá podmínkou exogenity - NE

Problém identifikace

Zatím jsme předpokládali, že máme máme jednu „problematickou“ proměnnou a jednu instrumentální proměnnou

Mohou však nastat tři situace

1) Počet instrumentálních proměnných je více než endogenních regresorů – Over-identified
2) Počet instrumentálních proměnných je menší, než počet endogenních regresorů – Under - identifiend
3) Počet instrumentálních proměnných je roven počtu endogenních regresorů – Just – identifed

Případ 2) nemusíme dále řešit, jelikož v takové případě získáme nekonečně mnoho řešení

Počet instrumentálních proměnných je roven počtu endogenních regresorů – Just – identifed

$$\log(wage) = \beta_0 + \beta_1 educ + \epsilon$$

Využijeme například vzdělání otce.
Dá se očekávat, že vzdělání otce ovlivní i vzdělání daného jedince a nebude mít vliv na jeho schopnosti
Obecně je však velký problém najít vhodnou IV

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

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

A co se pak v průměru děje? Aplikujeme střední hodnotu E

$$E(Z'y) = E(Z'X\beta) + E(Z'\epsilon)$$

$$E(Z'y) = \beta E(Z'X)$$

$$E(Z'X)^{-1}E(Z'y) = \beta$$

Odhad je pak:

$$b_{iv} = (Z'X)^{-1}(Z'y)$$

Trocha matematiky

$$b_{iv} = (Z'X)^{-1}(Z'y)$$

$$b_{iv} = (Z'X)^{-1}Z'(X\beta + \epsilon)$$

$$b_{iv} = (Z'X)^{-1}Z'X\beta + (Z'X)^{-1}Z'\epsilon$$

$$b_{iv} = \beta + (Z'X)^{-1}Z'\epsilon$$

$$plim\, b_{iv} = \beta + plim\left(\frac{Z'X}{n}\right)^{-1} plim\left(\frac{Z'\epsilon}{n}\right)$$

$$plim\, b_{iv} = \beta$$

$$plim\, b_{ols} = \beta + plim\left(\frac{X'X}{n}\right)^{-1} plim\left(\frac{X'\epsilon}{n}\right)$$

$$plim\, b_{ols} \neq \beta$$

Statistické vlastnosti IV

Pro jednoduchost, budeme předpokládat přímkovou regresi.
Za předpokladu homoskedasticity, má b_{iv} rozptyl.

$$Var(b_{IV1}) = \frac{\sigma_\epsilon^2}{\sum(x_1 - \bar{x}_1)R_{x,z}^2}$$

$\sigma_\epsilon^2 - rozptyl\ náhodné\ složky\ \epsilon$
$R_{x,z}^2 - koeficient\ determinace\ z\ regrese$

$$x = \alpha_0 + \alpha_1 z + v$$

Chceme, aby α_1 bylo stat. Významné. Proč?

IV odhad má asymptoticky normální rozdělení – můžeme použít t,LM testy.
Ty však také mají pouze asymptotické rozdělení při platnosti H0

F test se u IV nepoužívá, jelikož u IV neplatí SST=SSR+SSE a tedy ani R2 nepoužíváme k vyhodnocení modelu

Počet instrumentálních proměnných je více než endogenních proměnných

Jedna z možností je vybrat jednu z IV.
Máme však lepší možnost – můžeme získat vydatnější odhad!

Pokud pro jednotlivé IV platí podmínka exogenity a relevance, tak pak jsou podmínky splněny i pro jejich lineární kombinaci
Nejlepším instrumentem (poskytující nejnižší rozptyl) je pak tato lineární kombinace

Tím se dostáváme k metodě odhadu (estimátoru) dvoustupnové metody nejmenších čtverců 2SLS
Cílem je nahradit endogenní regresor nafitovanou hodnotou, získanou z lineární kombinace IV

Stupeň I

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \epsilon$$

1) $Cov(y_2, \epsilon) \neq 0$
2) $Cov(x_1, \epsilon) = 0$

Dále máme exogenní proměnné z_1, z_2
Předpokládejme, že z_1, z_2 jsou korelována s y_2 a nekorelována s ϵ
Potom libovolná lineární kombinace z_1, z_2, x_1 bude též nekorelována s ϵ
Cílem je najít takovou lineární kombinaci, která bude nejvíce korelována s y_2

Odhadneme pomocí OLS $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_1 + v$

$\pi_1 \neq 0 \; nebo \; \pi_2 \neq 0$
$testování \; pomocí \; F - testu$

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 x_1$$

Stupeň II

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \epsilon$$

1) $Cov(y_2, \epsilon) \neq 0$
2) $Cov(x_1, \epsilon) = 0$

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

$$y_1 = \beta_0 + \beta_1 \boldsymbol{\hat{y}_2} + \beta_2 x_1 + \epsilon$$

1) Provedeme odhad endogenní proměnné na IV a ostatních exogenních regresorů
2) Získáme nafitované hodnoty
3) Nahradíme endogenní regresor, nafitovanou (vyrovnanou) hodnotou z bodu 2
4) Odhadneme upravenou původní rovnici

Jinak pozor, pokud si sami budete dělat odhad pomocí 1—4, tak odhady *Var(b)* nebudou konzistentní.
Když použijete „nějaký“ software pro IV, tak ten již poskytuje konzistentní odhady *Var(b)*

Matematicky vlastně pouze nahradíme Z za X

$$b_{iv} = (Z'X)^{-1}(Z'y)$$

$$b_{2SLS} = \left(\hat{X}'X\right)^{-1}(\hat{X}'y)$$

Weak instrument

Je logické, že čím silnější vazba mezi z a x existuje, tím lépe.
Přesnější řečeno, rozptyl odhadu *b*, bude menší. Pro jednoduchost opět použijeme pouze příklad přímkové regrese:

$$Var(b_1) = \frac{\sigma_\epsilon^2}{\sum(x_1 - \bar{x}_1)R_{x,z}^2} \qquad R_{x,z}^2 \to 1 \text{ tak } Var(b) \to 0$$

Ale co když $R_{x,z}^2 \to 0$ tak $Var(b) \to$?

Čím bude slabší vztah mezi endogenním regresorem a IV, tím větší rozptyl bude mít odhad IV
Toto je problém „weak instrument“

$$plim\ b_{1,iv} = \beta_1 + \frac{corr(z,\epsilon)}{corr(z,x)} \times \frac{\sigma_\epsilon}{\sigma_x} \qquad \frac{corr(z,\epsilon)}{corr(z,x)} > corr(x,\epsilon)$$

$$plim\ b_{1,ols} = \beta_1 + corr(x,\epsilon) \times \frac{\sigma_\epsilon}{\sigma_x}$$

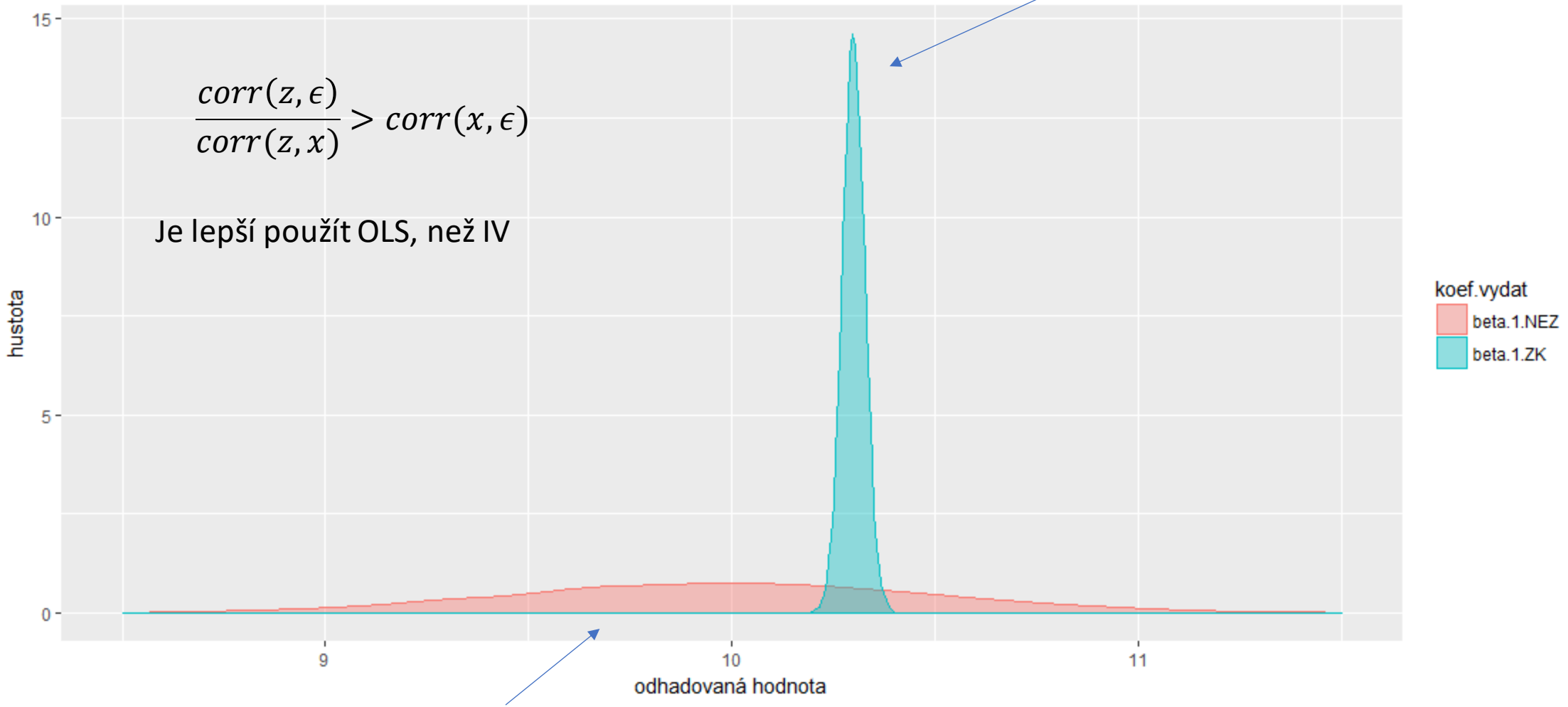
Je lepší použít OLS, než IV

Bias vs variance

OLS

$$\frac{corr(z,\epsilon)}{corr(z,x)} > corr(x,\epsilon)$$

Je lepší použít OLS, než IV



$Weak\ instrument$

Testování endogenity

Jak máme zjistit, zdali použít 2SLS a nebo OLS?

$$Var(b_{OLS,1}) = \frac{\sigma_\epsilon^2}{\sum(x_1 - \bar{x}_1)} \qquad Var(b_{IV,1}) = \frac{\sigma_\epsilon^2}{\sum(x_1 - \bar{x}_1)R_{x,z}^2}$$

Pokud nebude přítomna endogenita, tak by se měly odhady lišit pouze z důvodu sampling error
V případě endogenity, by však měl být rozdíl mezi odhady významný.
Hrubě řečeno testujeme následující hypotézu:

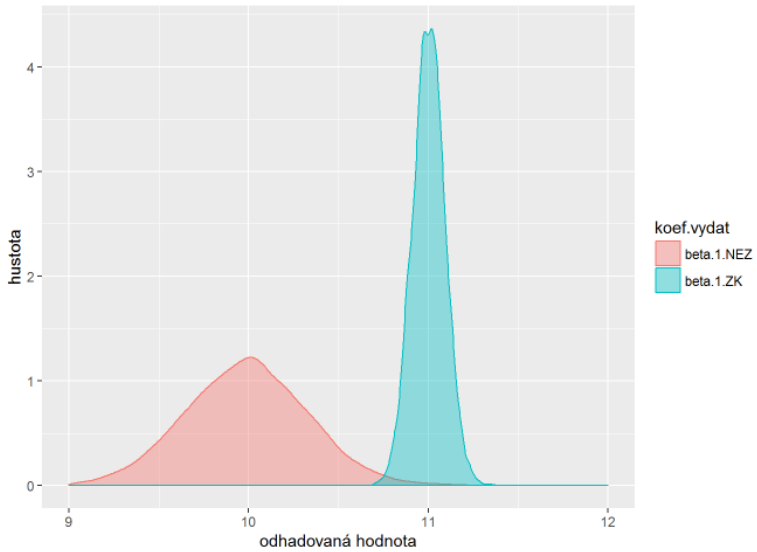
$$H_0: b_{2SLS} - b_{OLS} = 0$$
$$H_1: b_{2SLS} - b_{OLS} \neq 0$$

Hausmann test

Předpoklady:

- Není hetero, ani autokorelace
- Odhady se řídí normálním rozdělením
- Při H0 $\chi^2(k)$ *rozdělení*

Metoda	H_0: exogenita	H_1: exogenita
OLS	Konzistetní, vydatný	Nekonzistetní
2SLS	Konzistetní, není vydatný	Konzistentní



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