

APLIKACE ROBUSTNÍ REGRESE V ANALÝZE KOMPARATIVNÍCH CENOVÝCH HLADIN ZEMÍ EVROPSKÉ UNIE

Dagmar Blatná, Vysoká škola ekonomická v Praze*

1. Úvod

Komparativní cenové hladiny (dále *KC*) je ukazatel používaný EUROSTATem ke srovnání cenové hladiny jednotlivých zemí EU vzhledem k průměru EU 27. Podle definice EUROSTATu jsou komparativní cenové hladiny poměry mezi paritami kupní síly a devizovými kursy každé země. Paritou kupní síly je poměr měnové konverze, který přepočítává hodnoty ekonomických ukazatelů, vyjádřené v domácí měně na měnu společnou, která se nazývá standard kupní síly (PPS). Vyrovnáním kupní síly různých národních měn umožňuje tento standard srovnání ukazatelů mezi jednotlivými státy. Poměr je koncipován vzhledem k průměru (EU 27 = 100). Pokud je tento poměr za určitou zemi vyšší/nížší než 100, znamená to, že daná země je relativně dražší/levnější než průměr EU 27.

Teoretickým základem přibližování se cenových hladin evropských zemí je teorie nominální konvergence. V užším pojetí se nominální konvergencí rozumí sblížování ekonomik i z hlediska cenových charakteristik, např. z hlediska hrubého domácího produktu na hlavu ve „společné“ měně. V širším pojetí pak lze nominální konvergenci chápat jako sblížování různých makroekonomických veličin v absolutních hodnotách i tempech růstu.¹ Články věnované problémům nominální konvergence diskutují a navrhuji řešení, jak realizovat zvýšení cenové hladiny, zejména, je-li vhodnější zvyšovat ji cestou vyšší inflace nebo zhodnocováním měny, diskutují problematiku monetární politiky a rovnováhy v konvergenčním procesu, strukturální změny, srovnání *KC* různých zemí apod.²

Článek si neklade za cíl přispět k rozvoji ekonomické teorie v této oblasti, ale statistickou analýzou komparativních cenových úrovní zemí EU přispět k hlubšímu poznání vztahů mezi *KC* a ukazateli charakterizujícími ekonomiku evropských zemí z různých hledisek.

* Příspěvek je zpracován jako součást výzkumného projektu VŠE IG410040.

1 Nominální konvergence je chápána odlišně různými autory – přehled názorů viz např. Vintrová (2006), Vintrová, Žďárek (2006).

2 Viz např. Frait a Komárek (2001), Šaroch a Žák (2004), Žďárek (2006, 2010), Vintrová (2010), Vintrová a Žďárek (2006).

Cílem článku je ukázat, jak lze využít některých klasických, neparametrických a robustních statistických metod, zejména robustní regrese při analýze reálných dat evropských zemí a ukázat na některé situace, se kterými se můžeme setkat při analýze tak heterogenního souboru dat. Při analýze dat evropských zemí můžeme očekávat, že soubor bude obsahovat i odlehlá pozorování. Aplikace pouze klasických statistických metod, např. klasické lineární LM regrese, pak může poskytnout zkreslenou informaci o existenci závislosti i o hodnotách regresních parametrů. V takových situacích mohou být metody robustní regrese užitečným statistickým nástrojem při analýze závislosti, protože jsou méně citlivé na existenci odlehlých pozorování a současně poskytují dobré odhady regresních parametrů v případech, kdy jsou splněny podmínky pro použití klasické lineární regrese a k odhadům parametrů použijeme metodu nejmenších čtverců.

Tabulka 1

Komparativní cenové hladiny zemí EU v letech 2004–2009

Č.	Země EU	2004	2005	2006	2007	2008	2009
1	Belgie	106.7	106.4	107.7	108.3	111.1	113.9
2	Bulharsko	42.0	43.2	44.9	46.2	50.2	52.7
3	Česká republika	55.4	58.1	61.3	62.4	72.8	70.6
4	Dánsko	139.5	140.4	138.4	137.2	141.2	144.6
5	Německo	104.7	103.3	102.6	101.9	103.7	106.4
6	Estonsko	63.0	64.7	68.5	73.1	78.0	75.1
7	Irsko	125.9	123.4	124.5	124.5	127.6	125.0
8	Řecko	87.6	88.2	89.0	90.7	94.0	97.4
9	Španělsko	91.0	91.1	91.8	92.8	95.4	97.4
10	Francie	109.9	108.2	108.5	108.1	110.7	114.3
11	Italie	104.9	104.7	104.2	102.9	105.6	106.5
12	Kypr	91.2	90.3	90.3	88.1	90.5	91.2
13	Lotyšsko	56.1	57.0	60.7	66.6	72.6	74.8
14	Litva	53.5	54.8	57.3	60.0	64.6	67.8
15	Lucembursko	103.0	111.6	111.3	115.3	119.1	121.3
16	Maďarsko	62.0	63.3	60.6	66.7	68.1	65.5
17	Malta	73.2	73.0	74.8	75.5	78.8	81.4
18	Nizozemsko	106.1	104.7	104	101.9	104.0	108.5
19	Rakousko	103.3	102.5	101.9	102.2	105.1	107.9
20	Polsko	53.2	61.1	62.4	62.0	69.1	58.6
21	Portugalsko	87.4	85.1	84.9	85.7	87.0	89.3
22	Rumunsko	43.3	54.4	57.6	63.8	60.9	57.5
23	Slovinsko	75.5	76.0	76.7	79.0	82.3	85.5
24	Slovensko	54.9	55.4	58.0	63.2	70.1	73.7
25	Finsko	123.8	123.6	122.7	119.9	124.3	126.4
26	Švédsko	121.4	119.0	118.5	115.7	114.5	107.0
27	Spojené království	108.5	109.7	110.6	112.6	101.1	92.7

Zdroj: data EUROSTAT.

2. Komparativní cenové hladiny v zemích EU a postavení ČR v rámci EU

EUROSTAT každoročně přepočítává hodnoty komparativních cenových hladin vždy na aktuální stav, nyní pro 27 evropských zemí sdružených v EU. Z tohoto důvodu nejsou disponibilní data prezentovaná v jednotlivých letech porovnatelná s dříve publikovanými daty, rovněž některé dříve provedené analýzy vycházející z EU 25=100, nemusí poskytnout vzájemně srovnatelné výsledky.

Tabulka 1 uvádí komparativní cenové hladiny zemí EU v letech 2004 (vstup ČR do EU) do roku 2009.

Základní charakteristiky *KC* zemí EU ve sledovaných letech udává tabulka 2, přehlednější informaci podávají krabicové grafy na obrázku 1.

Tabulka 2

Základní popisné charakteristiky *KC* zemí EU

	2004	2005	2006	2007	2008	2009
Medián	91,00	90,30	90,30	90,70	94,00	92,70
Směrodatná odchylka	28,04	28,62	25,61	23,95	22,94	23,81
Variační koeficient	32,26 %	30,28%	28,90%	26,65%	24,76 %	25,59
Minimum	42,00	43,00	44,90	46,20	50,20	52,70
Maximum	139,50	140,40	138,40	137,40	141,20	144,60
Variační rozpětí	97,50	97,40	93,50	91,20	91,00	91,90
Dolní kvartil	56,10	61,10	61,30	66,60	72,60	73,70
Horní kvartil	106,70	108,20	108,20	108,30	110,70	108,50
Kvartilové rozpětí	50,60	47,10	46,90	41,70	38,10	34,80
Šikmost	-0,04	0,03	0,04	0,05	0,15	0,12
Špičatost	-1,16	-1,13	-1,14	-1,01	-0,70	-0,69

Zdroj: vlastní výpočty

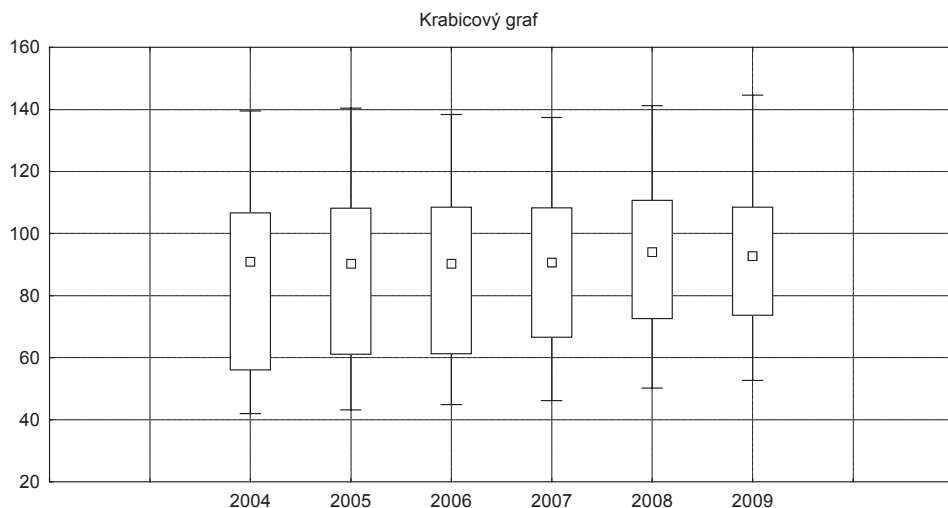
Průměr hodnot *KC* zemí EU se v celém období do roku 2009 stále zvyšoval, v roce 2009 došlo k poklesu, medián se v letech 2004–2007 udržoval na stejné úrovni, v posledních letech se zvýšil, na mediální úrovni se střídaly země Španělsko, Kypr, Řecko a v posledním roce se na mediální úroveň propadlo Spojené království. Česká republika je ve všech sledovaných obdobích na úrovni kolem dolního kvartilu hodnot *KC* (20. – 24. místo). Nejnižší *KC* má ve sledovaném období stále stejná země – Bulharsko, na nízké úrovni *KC* se udržují Rumunsko, Litva a Polsko. V celém období je zemí s nejvyšším *KC* Dánsko, na dalších místech se střídají Finsko a Irsko. Je zajímavé, že největší posun v pořadí zaznamenalo ve sledovaném období Lucembursko, které se posunulo ze 14. místa v r. 2004 na 4. místo, které si udržuje od roku 2005. Naproti tomu největší pokles *KC* zaznamenalo Spojené království (pokles z 6. místa v roce 2007 na 14. místo v roce 2009). Variabilita *KC* se dlouhodobě snižovala, což dokazuje, že dochází v souladu s teorií nominální konvergence ke zmenšování rozdílnosti cenových hladin v zemích EU a sbližování ekonomik, v roce 2009 se

variabilita *KC* zemí EU mírně zvýšila. Zajímavé je stále zužování kvartilového rozpětí *KC* i při zvýšení ostatních ukazatelů variability.

V roce 2004 převažovaly státy s vyšší hodnotou *KC* než je průměr (standardizovaná šikmost měla zápornou hodnotu), v průběhu let se její hodnoty zvyšovaly (od roku 2005 je kladná), což bylo zjevně způsobeno nově přistupujícími zeměmi, jejichž *KC* byly nižší než průměr zemí EU. Zvyšování hodnoty standardizované špičatosti rozdělení *KC* také dokumentuje konvergenci *KC* evropských zemí.

Obrázek 1

Krabicové grafy *KC* zemí EU v letech 2004–2009



Zdroj: vlastní výpočty

Je třeba zdůraznit, že nominální konvergence (sbližování cenových hladin) není výrazem pouze růstu cen v domácí ekonomice, ale je spojena s trendem v relativních cenách vzhledem k ostatním zemím EU.

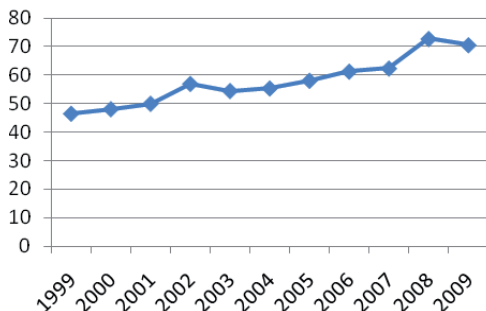
Významnost rozdílů *KC* v roce vstupu ČR do EU (rok 2004) a letech 2008 a 2009 byla posouzena párovými *t*-testy i neparametrickými testy Wilcoxonovým a znaménkovým, které potvrdily rozdílnost *KC* v roce 2004 od roku 2008 i od roku 2009, rozdíly mezi *KC* mezi roky 2008 a 2009 nejsou statisticky významné.

Nyní nabídneme stručný pohled na postavení ČR v rámci EU a na vývoj úrovně *KC* v ČR. Na obrázku 2 je zobrazen vývoj *KC* ČR od roku 1997 (zdroj EUROSTAT – přepočítaná data na EU 27=100).

ČR se stala členem EU v roce 2004, od té doby měla *KC* v ČR stoupající tendenci (viz obrázek 2), k výraznému růstu došlo v roce 2008 (ČR se dostala na 20. místo v EU), v roce 2009 došlo k mírnému poklesu *KC* (pokles na 22. místo). Dlouhodobý rostoucí trend znamená, že se *KC* ČR v posledních 10 letech (pomalu) přibližuje k průměru Evropské unie. Průměrný koeficient růstu *KC* v období členství v EU byl v ČR 1,0394 (průměrný růst 3,94 % za rok).

Obrázek 2

Komparativní cenové hladiny ČR v letech 1999–2009



Zdroj: data EUROSTAT, vlastní výpočty.

3. Data a ukazatele pro analýzu

Pro analýzu závislosti komparativní cenové hladiny (*KC*) byly vybrány ukazatele z různých oblastí ekonomiky za rok 2008. Výběr ukazatelů byl do značné míry determinován jejich dostupností – celá řada vhodných ukazatelů z dalších oblastí ekonomiky nebyla v době zpracování článku k dispozici.³ Zdrojem dat je EUROSTAT, výpočty jsou vlastní. K výpočtům byly použity programové pakety SAS 9.1, S-PLUS 6.2, Statistica a Statgraphics.

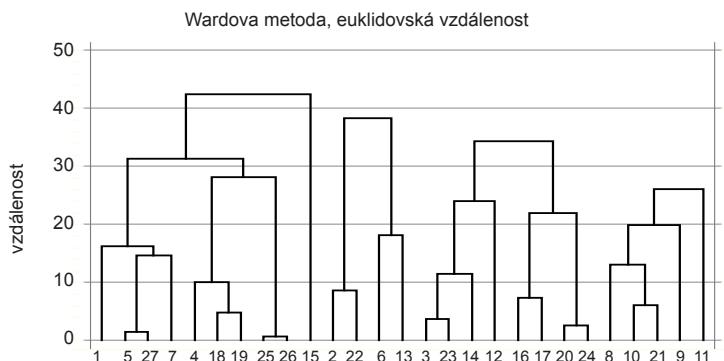
Ukazatele použité v analýze:

- HDP na obyvatele v paritě kupní síly (PPS), (EU 27 =100), (*HDP*)
- Míra růstu reálného HDP (*rHDP*)
- Produktivita práce na zaměstnanou osobu (HDP v PPS na zaměstnanou osobu), (*PP*)
- Celkové investice (% HDP), (*TI*)
- Celková státní podpora (% HDP), (*SP*)
- Celkový vládní dluh (% z HDP), (*VD*)
- Míra inflace (relativní přírůstek harmonizovaného indexu spotřebitelských cen), (*MI*)
- Celková míra zaměstnanosti v % (*MZ*)
- Celková míra dlouhodobé nezaměstnanosti (podíl z aktivního obyvatelstva), (*MN*)
- Energetická náročnost ekonomiky (hrubá domácí spotřeba energie dělená HDP), (*EN*)
- Cena elektrické energie (E/kWh), (*CE*)
- Hrubé domácí výdaje na výzkum a vývoj (% z HDP), (*IV*)
- Procento domácností, které mají přístup k internetu (*IN*).

3 I když údaje o *KC* byly v době zpracování článku k dispozici do roku 2009, data pro mnohé z ukazatelů nebyla uveřejněna, proto je analýza provedena na datech roku 2008. I když o hodnotách některých ukazatelů (zejména v některých zemích) lze mít pochybnosti, jedná se o oficiální data, s nimiž pracuje EUROSTAT.

Roztřídění zemí EU z hlediska všech uvažovaných ukazatelů lze provést mnoha způsoby a podle různých kritérií. V analýze bylo provedeno roztřídění shlukovou analýzou (teorie viz Hebák, 2005). Byla použita Wardova metoda s Euklidovskou vzdáleností. Dendrogram je uveden na obrázku 3.

Obrázek 3
Dendrogram



Zdroj: data EUROSTAT, vlastní zpracování.

Z hlediska všech uvažovaných ukazatelů byly země EU roztříděny do čtyř shluků:

1. Belgie, Dánsko, Německo, Irsko, Lucembursko, Nizozemsko, Rakousko, Finsko, Švédsko a Spojené království
2. Bulharsko, Estonsko, Litva, Rumunsko
3. ČR, Lotyšsko, Maďarsko, Malta, Polsko, Portugalsko, Slovinsko, Slovensko
4. Řecko, Španělsko, Francie, Itálie, Kypr.

Zajímavé je postavení země 15 (Lucembursko), které je zjevně odlehlým pozorováním i ve skupině vyspělých zemí. Základní popisné statistiky vytvořených shluků zemí uvádí tabulka 3.

Tabulka 3
Průměry a směrodatné odchylky KC shluků

Skupina	<i>n</i>	Průměr	Sm. odchylka
1	10	115.060	13.062
2	4	65.425	12.408
3	8	74.100	7.789
4	5	99.240	8.520

Zdroj: EUROSTAT, vlastní výpočty

Pomocí analýzy rozptylu bylo ověřeno, jestli lze skupiny zemí vytvořené shlukovou analýzou považovat za navzájem statisticky významně odlišné i z hlediska *KC*. Bartlettovým testem byla ověřena přípustnost použití analýzy rozptylu, ale vzhledem k malému počtu jednotek ve skupinách byla současně použita neparametrická analýza rozptylu – Kruskallův-Wallisův test a mediánový test (viz Antoch, 1992; Blatná, 1996). Závěry všech testů (viz tabulka 4) jsou stejné – vytvořené skupiny zemí nelze z hlediska *KC* považovat za shodné.

Tabulka 4

Výsledky analýzy rozptylu a neparametrických testů

ANOVA		Kruskal-Wallis Test		Median One-Way Analysis	
F	30,97	Chi-Square	20.953	Chi-Square	21.371
DF	3;23	DF	3	DF	3
Pr > F	<0.0001	Pr > Chi-Square	0.0001	Pr > Chi-Square	<.0001

Zdroj: EUROSTAT, vlastní výpočty.

Z hlediska výše *KC* se statisticky významně odlišují skupiny 1-2, 1-3, 2-4 a 3-4. Proto pro další analýzu byly sloučeny skupiny 1+4 (skupina I) a 2+3 (skupina II).

Skupinu I nyní tvoří „staré“ členské státy EU, ze zemí, které přistoupily později, je zde pouze Kypr, skupina II zahrnuje země, které přistoupily do EU později, ze „starších“ členských zemí je v této skupině zahrnuto Portugalsko a Malta:

- I. Belgie, Dánsko, Německo, Irsko, Lucembursko, Nizozemsko, Rakousko, Finsko, Švédsko, Spojené království, Řecko, Španělsko, Francie, Itálie, Kypr.
- II. Bulharsko, Estonsko, Litva, Rumunsko, Česká republika, Lotyšsko, Maďarsko, Malta, Polsko, Portugalsko, Slovinsko, Slovensko.

Vytvořené skupiny zemí tvoří celky, které lze dobře interpretovat i když zařazení některých zemí bylo poněkud překvapivé.⁴

4. Analýza závislosti komparativních cenových hladin na vybraných ukazatelích

Analýzy závislosti *KC* zemí EU na vybraných ukazatelích byly provedeny za tři skupiny zemí:

- A. Všechny členské země EU ($n = 27$),
- B. Země skupiny I ($n = 15$),
- C. Země skupiny II ($n = 12$).

4 Např., ve světle posledního vývoje je otázkou, zda je správné zařazení Řecka do skupiny I. Tato diskuse nám ale nepřísluší, neboť výpočty jsou provedeny na základě oficiálních dat EUROSTATu, i když se objevily pochybnosti o reálnosti dat dodaných Řeckem.

První informaci o existenci závislosti *KC* na vybraných ukazatelích podává korelační analýza. Vzhledem k předpokládané existenci odlehlých pozorování (a někdy i nelinearitě) závislostí, byly vedle klasických korelačních koeficientů vypočteny i robustní korelační koeficienty i neparametrické Spearmanovy koeficienty pořadové korelace. Tabulka 5 uvádí párové korelační koeficienty lineárních závislostí mezi skupinou vybraných ukazatelů pro soubor zemí celé EU: v horní polovině matice jsou klasické korelační koeficienty, ve spodní polovině Spearmanovy koeficienty pořadové korelace (ověření *t*-testem provedeno na 5% hladině významnosti).

Tabulka 5

Korelační koeficienty a Spearmanovy koeficienty mezi sledovanými ukazateli – skupina A (27 zemí EU)

	KC	IN	HDP	PP	SP	MN	INF	CE	MZ	TI	HDP	VD	VV	EN
KC	1,000	0,723	0,727	0,841	0,284	-0,18	-0,63	0,566	0,514	-0,56	-0,51	0,303	0,722	-0,79
IN	0,714	1,000	0,662	0,672	0,415	-0,29	-0,45	0,347	0,611	-0,35	-0,41	-0,01	0,733	-0,61
HDP	0,924	0,779	1,000	0,919	0,602	-0,23	-0,49	0,555	0,285	-0,50	-0,31	0,06	0,450	-0,59
PP	0,889	0,697	0,932	1,000	0,475	-0,09	-0,67	0,660	0,226	-0,63	-0,34	0,339	0,552	-0,75
SP	0,278	0,493	0,307	0,311	1,000	-0,08	0,053	0,328	0,024	-0,26	-0,33	-0,02	0,081	-0,26
MN	-0,14	-0,25	-0,28	-0,08	-0,28	1,000	0,006	-0,09	-0,49	0,113	-0,04	0,202	-0,18	0,02
INF	-0,69	-0,54	-0,73	-0,68	-0,14	-0,70	1,000	-0,64	-0,15	0,701	-0,06	-0,56	-0,44	0,668
CE	0,562	0,319	0,635	0,649	0,257	-0,04	-0,57	1,000	0,071	-0,56	-0,21	0,537	0,175	-0,61
MZ	0,468	0,584	0,544	0,298	0,146	-0,50	-0,41	0,080	1,000	-0,08	-0,31	-0,22	0,625	-0,33
TI	-0,55	-0,51	-0,60	-0,59	-0,37	0,000	0,606	-0,54	-0,95	1,000	0,245	-0,61	-0,36	0,726
rHDP	-0,56	-0,39	-0,42	-0,38	-0,29	-0,14	0,295	-0,26	-0,29	0,298	1,000	-0,13	-0,32	0,492
VD	0,321	0,031	0,269	0,421	0,101	0,369	-0,52	0,525	-0,16	-0,58	-0,16	1,000	0,666	-0,53
VV	0,732	0,765	0,718	0,624	0,312	-0,17	-0,56	0,226	0,633	-0,38	-0,44	0,163	1,000	-0,48
EN	-0,83	-0,60	-0,82	-0,76	-0,26	0,073	0,772	-0,04	-0,39	0,707	0,531	-0,50	-0,58	1,000

Zdroj: EUROSTAT, vlastní výpočty. Tučně jsou vyznačeny statisticky významné hodnoty korelačních koeficientů ověřené *t*-testy – použita 5% hladina významnosti.

Následující tabulka 6 uvádí hodnoty korelačních koeficientů mezi *KC* a sledovanými ukazateli pro všechny tři soubory zemí. Je zřejmé, že mezi sledovanými ukazateli často existuje silná multikolinearita.

Z porovnání jednotlivých typů korelačních koeficientů je vidět, že hodnoty klasických a robustních měr se liší, často i z hlediska významnosti, v některých případech i znaménka. Z toho je možno usuzovat na existenci odlehlých pozorování, které se v souborech vyskytují a porušují předpoklady použití klasických měr závislosti a za následek mohou mít zavádějící informaci o existenci a směru závislosti.

Tabulka 6

Korelační koeficienty, Spearmanovy koeficienty a robustní korelační koeficienty KC a skupiny vybraných ukazatelů.

	Skupina A (27 zemí EU)			Skupina B		Skupina C	
	Korelační koeficient	Spearmanův koeficient	Robustní koeficient korelace	Korelační koeficient	Robustní koeficient korelace	Korelační koeficient	Robustní koeficient korelace
<i>IN</i>	0,723	0,714	0,693	0,575	0,488	0,733	0,737
<i>HDP</i>	0,727	0,924	0,936	0,338	0,345	0,824	0,857
<i>PP</i>	0,841	0,889	0,953	0,367	0,497	0,745	0,832
<i>SP</i>	0,284	0,278	0,119	0,248	-0,181	0,017	0,024
<i>MN</i>	-0,181	-0,136	0,075	-0,272	0,010	0,027	0,027
<i>MI</i>	-0,632	-0,690	-0,843	-0,141	0,349	-0,461	-0,461
<i>CE</i>	0,566	0,562	0,635	-0,053	0,054	0,334	0,334
<i>MZ</i>	0,514	0,468	0,533	0,304	-0,211	0,337	0,337
<i>TI</i>	-0,560	-0,555	-0,681	-0,149	0,619	-0,492	-0,492
<i>rHDP</i>	-0,513	-0,566	-0,511	-0,666	-0,443	-0,517	-0,785
<i>VD</i>	0,303	0,321	0,481	-0,444	-0,302	0,388	0,878
<i>VV</i>	0,722	0,732	0,746	0,545	0,512	0,652	0,589
<i>EN</i>	-0,791	-0,826	-0,837	-0,402	0,321	-0,826	-0,859

Zdroj: EUROSTAT, vlastní výpočty. Tučně jsou vyznačeny korelační koeficienty, u nichž byla *t*-testy na 5% hladině významnosti ověřena statisticky významná hodnota koeficientu.

Z tabulky 6 je zřejmé, že *KC* ve vytvořených skupinách závisí mnohdy na jiných ukazatelích, což potvrzuje oprávněnost rozdělení celého souboru zemí EU do skupin. *KC* zemí EU jako celku (skupina A) silně koreluje s produktivitou práce, výši HDP/obyvatele, energetickou náročností země, vybaveností domácností internetem a podílem HDP vydávaným na výzkum a vývoj, významné jsou korelační koeficienty s mírou inflace, cenou energie, celkovými investicemi a růstem HDP. U skupiny „starších“ členských zemí (skupina B) je nejtěsnější závislosti mezi *KC* a růstem HDP (záporná, takže s růstem HDP klesá *KC* země), robustní korelace potvrzuje silnou závislost *KC* na celkových investicích země (*TI*). Naproti tomu u skupiny méně vyspělých zemí (skupina C) je *KC* silně závislá na vyspělosti země (charakterizované produktivitou práce a HDP/obyvatele a energetickou náročností ekonomiky (*EN*)), v případě robustní míry i na růstu HDP a celkovém vládním dluhu, který naopak není významným faktorem změny úrovně *KC* ve vyspělých zemích (skupina B).

Některé ze zjištěných závislostí jsou známé z ekonomické teorie (např. Šaroch, 2004; Frait, 2001), zejména závislost komparativních cenových úrovní a HDP/obyvatele a závislost *KC* a produktivity práce. V provedené analýze se překvapivě neprojevila ani u jedné z vytvořených skupin zemí B a C statisticky významná závislost *KC* na míře inflace, i když u celého souboru 27 zemí, je korelace *KC* a inflace statisticky významná (na 5% hladině významnosti).

Lze předpokládat, že ve všech souborech zemí existují takové, které se od většiny zemí své skupiny mohou v některých sledovaných závislostech významně odlišovat

(ze statistického hlediska se jedná o odlehlá pozorování), což může mít významný vliv na odhady získané klasickou lineární LS regresí. Metod identifikace odlehlých pozorování v regresi existuje celá řada,⁵ v případě, když analýza je provedena robustní regresí, je vhodné použít robustní nástroje identifikace, které jsou schopné identifikovat extrémní odlehlá pozorování (outliers), vybočující pozorování (leverage points) i vlivná pozorování (současně extrémní i vybočující odlehlá pozorování). K identifikaci byly využity nástroje robustní LTS regrese – robustní vzdálenost, robustní standardizovaná rezidua, regresní diagnostické grafy, graf standardizovaných reziduí od robustního regresního modelu, klasický i robustní Q-Q graf zobrazující odchylky (rezidua) od lineárního regresního modelu a posuzující normalitu rozdělení reziduí.⁶

Regresní analýza *KC* byla provedena pomocí klasické metody nejmenších čtverců (LS regrese), robustní LTS-regrese a robustní MM-regrese. V následujícím textu uvedeme pouze vybrané výsledky této rozsáhlé analýzy. Cílem je ukázat, jaký vliv mohou mít odlehlá pozorování na hodnoty regresních parametrů získané klasickou MNČ a porovnání jejich hodnot s odhady získanými robustními metodami odhadu parametrů s vysokým bodem selhání. Z tohoto důvodu jsou uvedeny různé typy výsledných regresních modelů vyhovujících ze statistického hlediska.⁷ Jako příklad možné grafické diagnostiky odlehlých pozorování je u vybraných modelů uveden i některý z příslušných diagnostických grafů.

4.1 Regresní analýza *KC* pro soubor 27 zemí EU (soubor A)

Ze statistického hlediska je vyhovujících celá řada různých regresních modelů. Je tedy možné modelovat závislost *KC* různými modely. V následujícím textu uvedeme některé z nich.⁸

Jeden z nejlepších výsledných lineárních regresních modelů je model závislosti *KC* na HDP/obyvatele (*HDP*) a na míře zaměstnanosti (*MZ*). Tabulka 7 uvádí robustní identifikaci odlehlých pozorování, obrázky 4 a 5 grafickou identifikaci (popis viz Teoretický dodatek).

- 5 Např. pomocí diagonálních prvků projekční (hat) matice, standardizovaných reziduí, jackknife reziduí, DFBETA, DFFIT, Cookovy vzdálenosti, pomocí 0-1 umělých proměnných atd., podrobněji viz např. Hebák (2005), Hušek (2007), Zvára (2009), Rao (1995), Hadi (1993), Swallow (1996), Maronna (2006), Rousseeuw (2003).
- 6 Uvedené nástroje robustní regrese jsou implementované v systémech SAS a S-Plus, které byly použity pro výpočty. Stručný popis principu použité robustní regrese a robustní identifikace odlehlých pozorování je uveden v Teoretickém dodatku článku, blíže viz např. Rousseeuw (2003), Rousseeuw (1990), Swallow (1996), Chen (2002), Olive (2002).
- 7 Tzn. z hlediska významnosti *t*-testů regresních parametrů, *F*-testů, ověření normality reziduí grafickou diagnostikou pomocí Q-Q- grafů, posouzením homoskedasticity, u robustních modelů robustních *t*-testů, *F*-testů, Waldových testů, robustní vzdáleností, Akaikeho a Schwarzovým informačním kritériem. Vzorce jsou uvedeny v teoretickém dodatku článku, blíže viz např. Hebák (2005), Hušek (2007), Rao (1995), Rousseeuw (2003).
- 8 Jsou uvedeny vždy jen klasický lineární LS model a robustní MM model, i když výpočty byly provedeny také LTS regresí, neboť použitý software neposkytuje u LTS regrese testování významnosti parametrů.

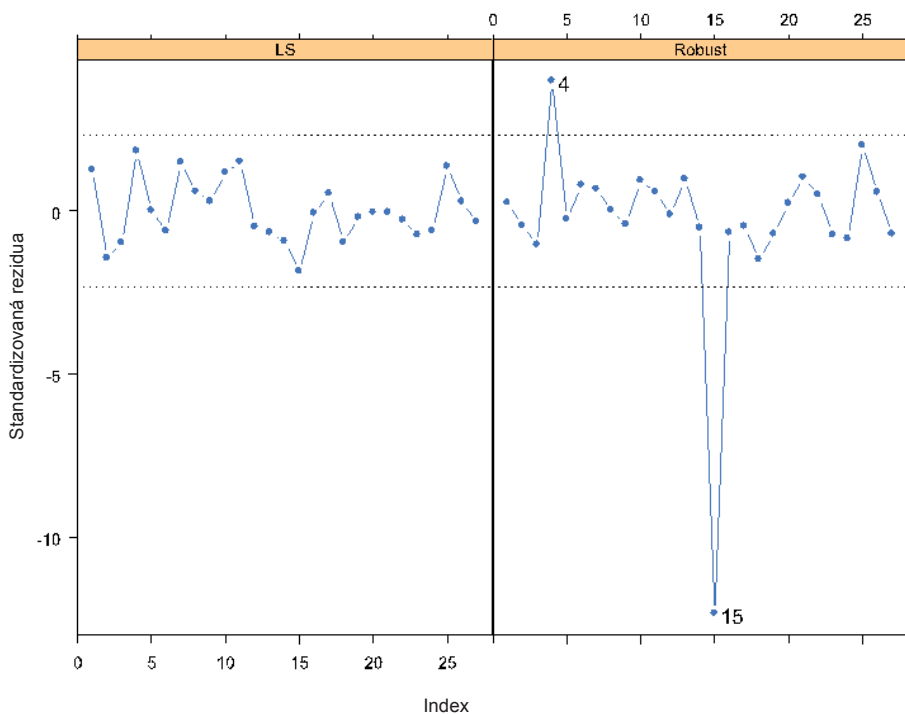
Tabulka 7

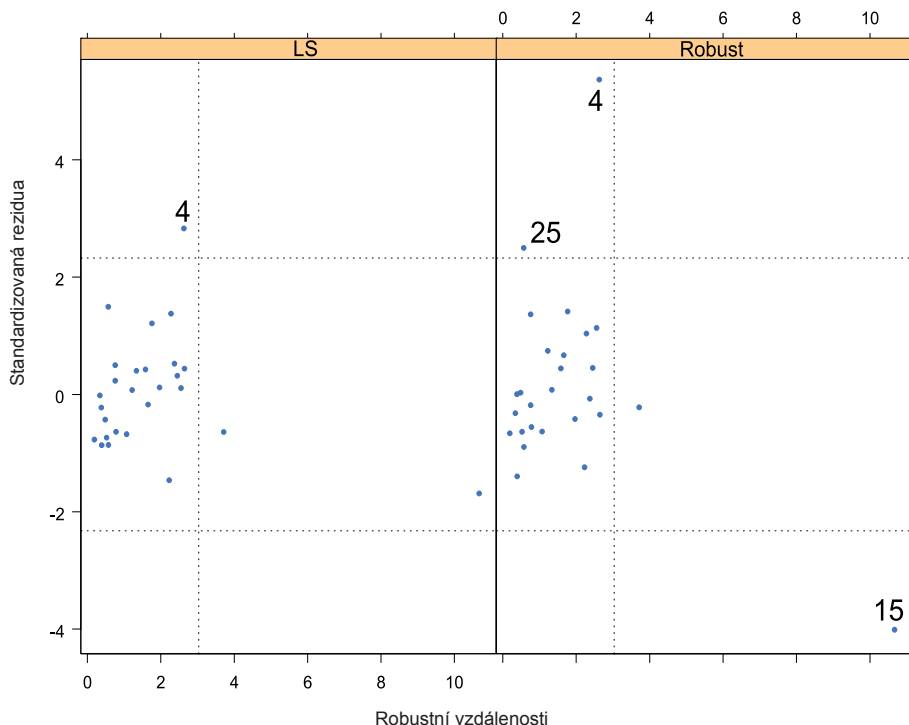
Robustní diagnostika odlehlých pozorování

Země	Mahalanobisova vzdálenost	Robustní vzdálenost	Vybočující pozorování	Standardizovaná robustní rezidua	Extrémní hodnoty
4 Dánsko	1,958	2,402		5,876	*
15 Lucembursko	2,694	4,174	*	-13,451	*
17 Malta	1,856	3,115	*	-0,837	
25 Finsko	0,849	0,935		2,586	*

Zdroj: EUROSTAT, vlastní výpočty.

Obrázek 4

Grafy standardizovaných reziduí lineárního regresního modelu *KC-HDP+MZ*



Zdroj: vlastní zpracování.

Z tabulky 7 je vidět, že robustní diagnostikou byla identifikována tři extrémní pozorování a dvě vybočující pozorování. Rychlejší informaci o existenci odlehlých pozorování lze získat z grafů. Na obrázku 4 je graf standardizovaných reziduí („index“ označuje číslo zemí EU seřazených podle anglické abecedy), na obrázku 5 je robustní diagnostický graf. V levé části obou grafů 4 a 5 je klasická a v pravé části robustní diagnostika. Jak je zřejmé z obrázku 4, žádné standardizované reziduum od klasického lineárního regresního modelu neleží mimo 5% toleranční meze, u robustní regrese jsou statisticky významná standardizovaná rezidua u zemí 4 (Dánsko) a 15 (Finsko), z obrázku 5 je vidět, že klasickou regresní diagnostikou bylo identifikováno jedno extrémní pozorování 4 (Dánsko), robustní diagnostika odhalila i odlehlá pozorování 15 (Lucembursko) a 25 (Finsko). Přitom země 15 je vlivné pozorování (současně extrémní i vybočující), což můžeme interpretovat tak, že hodnoty ukazatelů Dánska, Finska a zejména Lucemburska mají podstatný vliv na odhady parametrů regresní funkce získané klasickou regresí a můžeme předpokládat, že výsledný klasický a robustní lineární regresní model se budou z hlediska hodnot odhadnutých parametrů lišit. Tento fakt je vidět z tabulky 8, kde jsou uvedeny výsledné regresní lineární LS a MM modely.

Tabulka 8

Regresní modely $KC \sim HDP + MZ$ získané robustní MM regresí a klasickou lineární regresí (LS)

Metoda odhadu parametrů	Parametry regresní funkce	Odhady regresních parametrů	Směr.odch. odhadů regr. parametrů	t-test	P(> t) (p-hodnota)
MM	abs.člen	-10.626	15.307	-0.694	0.494
LS	abs.člen	-48.551	22.262	-2.181	0.039
MM	HDP	0.688	0.052	13.135	0.000
LS	HDP	0.579	0.067	8.588	0.000
MM	MZ	0.577	0.243	2.373	0.026
LS	MZ	1.313	0.343	3.832	0.001

Zdroj: vlastní výpočty.

Modifikovaný index determinace je u LS modelu 81,94 %, u robustního MM modelu 69,94 %. Je nutno upozornit, že hodnoty obou indexů determinace nejsou přímo srovnatelné, oba ale vyjadřují statisticky významnou závislost (vzorec robustního indexu determinace viz Teoretický dodatek – vzorec 5). Vzhledem k prokázané existenci odlehklých pozorování a porušení předpokladu normality reziduí je vhodné preferovat robustní model. Základní charakteristiky pro posouzení vhodnosti tohoto robustního modelu jsou v tabulce 11.

Jako příklad vhodného modelu, u něhož nebyla identifikována odlehlá pozorování ani klasickou ani lineární robustní regresí a nebylo zjištěno porušení předpokladu normality reziduí, je model s vysvětlujícími proměnnými vybavenost internetem (*IN*) a míra inflace (*MI*). Odhady regresních parametrů u obou metod jsou v takovém případě stejné, jak je vidět z tabulky 9, modifikovaný index determinace je 0,638 u LS modelu a 0,604 u MM modelu.

Tabulka 9

Regresní modely $KC \sim IN + MI$ pro soubor 27 zemí EU

Metoda odhadu parametrů	Parametry regresní funkce	Odhady regresních parametrů	Směr.odch. odhadů regr. Parametrů	t-test	P(> t) (p-hodnota)
MM	abs.člen	62.425	15.860	3.936	0.001
LS	abs.člen	62.425	14.559	4.287	0.000
MM	IN	0.771	0.211	3.660	0.001
LS	IN	0.771	0.193	3.987	0.001
MM	MI	-2.696	1.057	-2.549	0.017
LS	MI	-2.696	0.971	-2.777	0.011

Zdroj: vlastní výpočty

V tabulce 10 jsou uvedeny další vyhovující lineární regresní modely pro závisle proměnnou *KC*. V případě identifikovaného odlehlého pozorování je uvedeno jméno země, tučně označené země jsou vlivná pozorování. Tabulka 11 uvádí charakteristiky k posouzení vhodnosti robustních modelů.

Tabulka 10

Další akceptovatelné regresní modely pro závisle proměnnou komparativní cenové hladiny zemí souboru A

Diagnostikovaná odlehlá pozorování	Regresní rovnice		Indexy determ.
E: 6,13 V: 2,6,13,22,24	LS	$KC = 58,447 + 0,332 HDP + 8,531 VV - 0,034 EN$	0,856
	MM	$KC = 53,684 + 0,352 HDP + 3,377 VV - 0,025 EN$	0,769
E: - V: 2,3,6,22	LS	$KC = 79,835 + 0,536 IN - 0,064 EN$	0,719
	MM	$KC = 79,835 + 0,536 IN - 0,064 EN$	0,665
E: 4,7,15 V: 8,11,25,26	LS	$KC = 56,809 + 0,216 VD + 17,160 VV$	0,587
	MM	$KC = 51,087 + 0,319 VD + 14,851 VV$	0,599
E: -	LS	$KC = 5,5168 - 3,977 MI + 1,632 CE$	0,575
V: -	MM	$KC = 5,5168 - 3,977 MI + 1,632 CE$	0,531
E: 6, 13 V: 2,6,13, 22,24	LS	$KC = 45,201 - 2,066 rHDP + 0,869 IN$	0,577
	MM	$KC = 67,621 - 5,166 rHDP + 0,616 IN$	0,523

Zdroj: vlastní výpočty. E značí extrémní pozorování, V značí vybočující pozorování

Pokud jsou v modelu diagnostikována vlivná odlehlá pozorování a parametry LS a MM modelu se liší (viz zejména modely 1 a 5), je vhodné dát přednost robustnímu modelu.

Tabulka 11

Základní charakteristiky pro posouzení vhodnosti robustních modelů

Model	Indexy determinace	AICR	BICR	Deviance
$KC = -10,626 + 0,6885 HDP + 0,5776 MZ$	0,699	22,573	30,274	2250,37
$KC = 53,684 + 0,352 HDP + 0,377 VV - 0,025 EN$	0,769	21,169	31,026	1659,13
$KC = 79,835 + 0,536 IN - 0,064 EN$	0,665	21,656	27,727	3310,59
$KC = 51,087 + 0,319 VD + 14,851 VV$	0,599	22,937	30,519	3527,51
$KC = 5,5168 - 3,977 MI + 1,632 CE$	0,531	21,289	27,647	3290,66
$KC = 67,621 - 5,166 rHDP + 0,616 IN$	0,523	22,251	29,216	4773,09

Zdroj: vlastní výpočty

4.2 Regresní analýza KC zemí skupiny B

Ve skupině B bylo nalezeno jen několik vyhovujících regresních lineárních modelů, indexy determinace u většiny modelů jsou poměrně nízké, často mají nevyhovující testy významnosti a je porušena normalita reziduí. Z vyhovujících modelů se jako nejlepší jeví robustní model $KC \sim rHDP + VD + EN + TI$, kterým lze vysvětlit 61 % variability KC. V modelu bylo robustní diagnostikou identifikováno pět vybočujících pozorování a tři extrémní pozorování (Dánsko, Španělsko, Kypr), dvě z nich jsou vlivnými body (Španělsko, Kypr). V tabulce 12 je robustní diagnostika, na obrázku 6

robustní diagnostický graf tohoto modelu, na obrázku 7 normální Q-Q graf reziduí (diagnostikuje normalitu rozdělení reziduí).

Tabulka 12

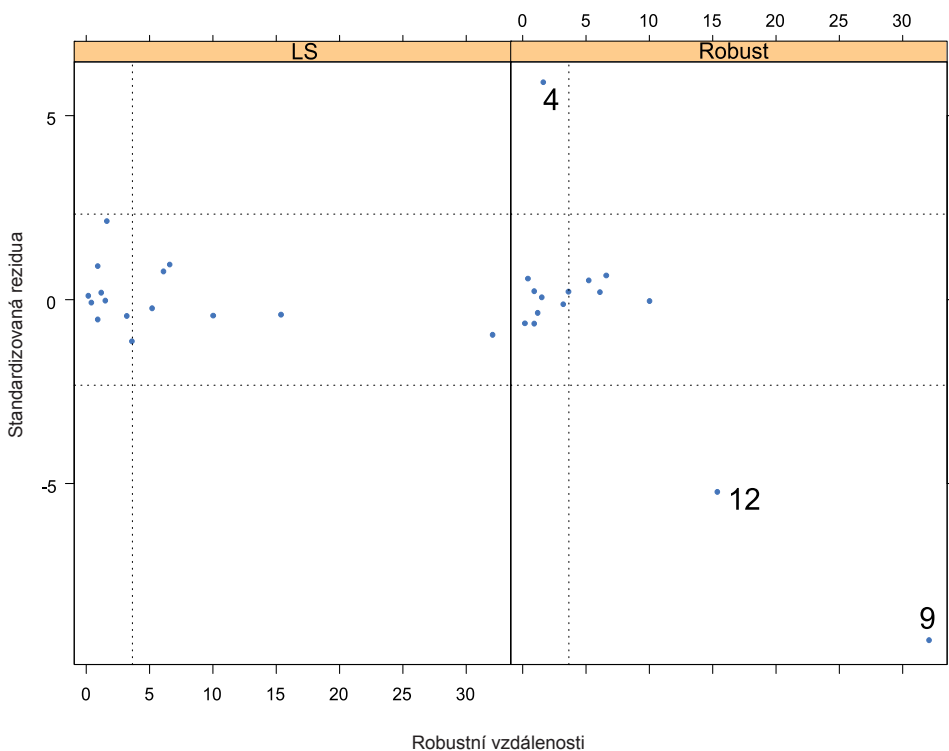
Robustní diagnostika odlehklých pozorování modelu $KC \sim rHDP + VD + EN + TI$

Země	Mahalanobisova vzdálenost	Robustní vzdálenost	Vybočující pozorování	Standardizovaná robustní rezidua	Extrémní hodnoty
4 Dánsko	1,813	2,460		10,618	*
7 Irsko	2,296	8,148	*	0,852	
9 Španělsko	2,987	8,506	*	-31,887	*
12 Kypr	2,021	14,584	*	-9,569	*
19 Rakousko	1,979	8,225	*	0,268	
25 Finsko	2,394	6,139	*	1,086	

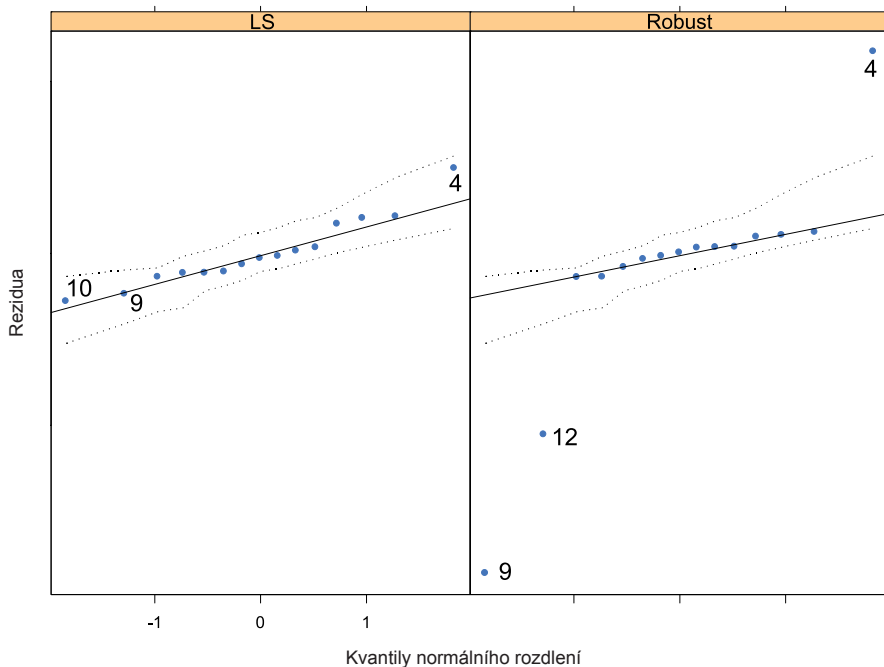
Zdroj: vlastní výpočty.

Obrázek 6

Robustní diagnostický graf modelu $KC \sim rHDP + VD + EN + TI$



Obrázek 7

Normální Q-Q graf modelu $KC \sim rHDP + VD + EN + TI$ 

Zdroj: vlastní zpracování.

Z obou grafů je zřejmé, že klasickými diagnostickými nástroji nebyla odlehlá pozorování identifikována. LS i MM modely jsou uvedeny v tabulce 13.

Tabulka 13

Regresní modely $KC \sim rHDP + VD + EN + TI$ pro soubor zemí B

Metoda odhadu parametrů	Parametry regresní funkce	Odhady regresních parametrů	Směr. odch. odhadů regr. parametrů	t-test	P(> t) (p-hodnota)
MM	abs. člen	64.186	8.922	7.194	0.000
LS	abs. člen	133.313	24.464	5.449	0.000
MM	HDP	-4.091	0.629	-6.500	0.000
LS	HDP	-6.249	2.347	-2.663	0.024
MM	VD	-0.228	0.026	-8.668	0.000
LS	VD	-0.210	0.108	-1.947	0.080
MM	EN	0.102	0.032	3.180	0.009
LS	EN	0.078	0.116	0.674	0.515
MM	TI	2.228	0.541	4.119	0.002
LS	TI	-0.953	1.163	-0.819	0.432

Zdroj: vlastní výpočty

Jak je zřejmé z tabulky 13, oba modely se liší. V souvislosti s existencí vlivných bodů, jsou hodnoty regresních parametrů odlišné, u proměnné *TI* dokonce má dílčí regresní koeficient opačné znaménko. Klasický model obsahuje i proměnné s nevýznamnými *t*-testy regresních parametrů. Index determinace robustního modelu je 0,6106, modifikovaný index determinace u LS modelu 0,6073, LS model má i významně vyšší odhady reziduálních směrodatných odchylek. Robustní model je jednoznačně lepším modelem. Charakteristiky vhodnosti robustního modelu jsou obsaženy v tabulce 15.

V tabulce 14 jsou uvedeny další regresní modely popisující závislost *KC* skupiny B na vybraných ukazatelích, u nichž byla robustní diagnostikou identifikována odlehlá pozorování, preferujeme v těchto případech výsledky robustní regrese, zejména v situaci, kdy bylo identifikováno vlivné odlehlé pozorování (model *KC~rHDP+VV*). V tabulce 15 jsou charakteristiky vhodnosti uvedených robustních modelů.

Tabulka 14

Další akceptovatelné regresní modely pro závisle proměnnou *KC* pro skupinu zemí B

Diagnostikovaná odlehlá pozorování	Regresní rovnice		R-sq.
E: 4 Dánsko, 7 Irsko, V: 7 Irsko, 11 Itálie, 12 Kypr	LS	$KC = 100,627 - 5,028 \text{ rHDP} + 6,194 \text{ VV}$	0,632
	MM	$KC = 100,746 - 4,585 \text{ rHDP} + 5,653 \text{ VV}$	0,528
E: 4 Dánsko, V: 15 Lucembursko, 12 Kypr	LS	$KC = 67,964 + 7,379 \text{ VV} + 0,237 \text{ PP}$	0,412
	MM	$KC = 63,055 + 5,632 \text{ VV} + 0,266 \text{ PP}$	0,568
E: 4 Dánsko, 7 Irsko, 5 Finsko V: 15 Lucembursko, 12 Kypr,	LS	$KC = 118,754 - 5,845 \text{ rHDP} - 0,197 \text{ VD} + 0,037 \text{ EN}$	0,581
	MM	$KC = 96,461 - 5,344 \text{ rHDP} - 0,203 \text{ VD} + 0,181 \text{ EN}$	0,505

Zdroj: vlastní výpočty. E značí extrémní pozorování, V značí vybočující pozorování.

Tabulka 15

Základní charakteristiky pro posouzení vhodnosti robustních modelů skupiny B

Model	R-sq.	AICR	BICR	Deviance
$KC = 64,186 - 4,091 \text{ rHDP} - 0,228 \text{ VD} + 0,102 \text{ EN} + 2,228 \text{ TI}$	0,611	7,740	18,680	907,87
$KC = 100,746 - 4,693 \text{ rHDP} + 5,751 \text{ VV}$	0,528	9,656	15,259	849,690
$KC = 63,055 + 5,632 \text{ VV} + 0,266 \text{ PP}$	0,568	11,558	18,140	714,643
$KC = 96,461 - 5,344 \text{ rHDP} - 0,203 \text{ VD} + 0,181 \text{ EN}$	0,506	10,683	18,780	886,96

Zdroj: vlastní výpočty

4.3 Regresní analýza *KC* zemí skupiny C

U této skupiny zemí je většina vyhovujících modelů shodných pro klasickou lineární i robustní regresi, neboť nebyla identifikována vlivná pozorování ani porušení předpokladu normality reziduí, tudíž klasické i robustní lineární regresní modely mají stejné hodnoty regresních parametrů. V některých případech robustní a klasická diagnostika odhalila stejná odlehlá pozorování.

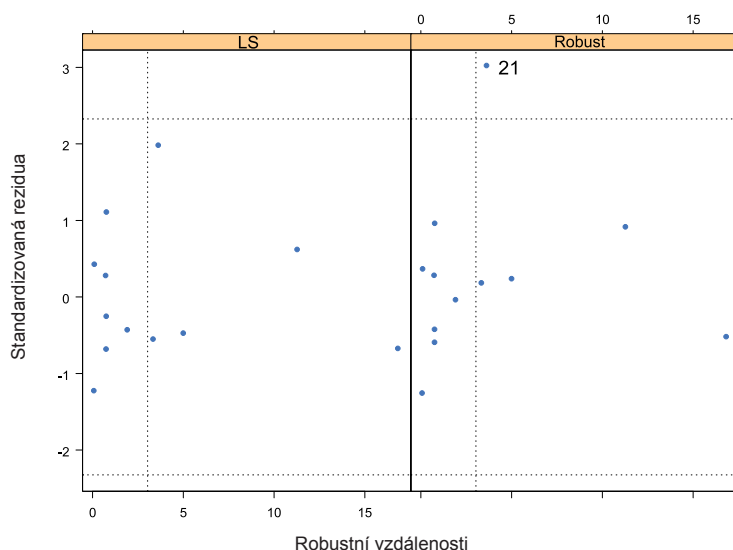
Nejvyšší procento variability *KC* této skupiny (93,9 % u *LS* a 72,9 % u *MM* modelu) lze vysvětlit modelem s vysvětlujícími proměnnými *VV* (hrubé domácí výdaje na výzkum a vývoj) a *EN* (energetická náročnost ekonomiky). Klasická regresní diagnostika neodhalila žádné odlehle pozorování ani porušení předpokladu normality reziduí, oba regresní modely mají stejný tvar

$$KC = 79,782 + 9,354 VV - 0,032 EN. \quad (1)$$

Pro skupinu C je rovněž vhodný model s vysvětlujícími proměnnými *IN* a *VV* (viz tabulka 16). Robustní diagnostikou bylo identifikováno 5 vybočujících a jedno extrémní pozorování, které je současně vlivným bodem – 21 (Portugalsko), které klasická diagnostika neodhalila, proto je vhodné dát přednost robustnímu modelu. Tento fakt je vidět rovněž z grafické identifikace na obrázku 8. Modifikovaný index determinace *LS* modelu je 0,756, u robustního modelu 0,631.

Obrázek 8

Robustní diagnostický graf modelu $KC \sim IN + VV$



Tabulka 16

Regresní modely $KC \sim IN + VV$ pro soubor zemí C

Metoda odhadu parametrů	Parametry regresní funkce	Odhady regresních parametrů	Směr.odch. odhadů regr. parametrů	t-test	P(> t) (p-hodnota)
MM	abs. člen	34.269	5.213	6.574	0.0001
LS	abs. člen	35.104	7.506	4.676	0.0012
MM	IN	0.618	0.109	5.627	0.0003
LS	IN	0.540	0.155	3.493	0.0068
MM	VV	6.300	2.715	2.320	0.0486
LS	VV	10.810	3.816	2.833	0.0196

Zdroj: vlastní výpočty

Další vyhovující lineární regresní modely pro skupinu zemí C jsou uvedené v tabulce 17. Dva modely jsou bez identifikovaných vlivných odlehlých pozorování, byla identifikována jen vybočující pozorování, regresní rovnice jsou stejné pro robustní a klasickou regresi. V takové situaci je možno po ověření normality reziduí použít klasický i robustní regresní model (u uvedených modelů mají rezidua přibližně normální rozdělení). Poslední dva modely obsahují jedno vlivné pozorování (Bulharsko), model 4 i dvě vybočující pozorování. V těchto případech je opět vhodné dát přednost robustním regresním modelům.

Tabulka 17

Další akceptovatelné regresní modely pro závisle proměnnou *KC* pro skupinu zemí C

Diagnostikovaná odlehlá pozorování	č.	Regresní rovnice		R-sq.
V: Bulharsko, Portugalsko, Slovinsko	1	LS MM	$KC=104,72-2,8734 MI-2,638 rHDP-0,176VD$ $KC=104,72-2,8734 MI-2,638 rHDP-0,176VD$	0,868 0,680
V: Bulharsko, Portugalsko, Slovinsko	2	LS MM	$KC = 76,782 + 9,354 VV -0,032 EN$ $KC=76,782 + 9,354 VV -0,032 EN$	0,839 0,729
E: Bulharsko V: Bulharsko	3	LS MM	$KC = -19,888 +0,868 MZ + 0,543 PP$ $KC = -6,776 +0,818 MZ + 0,412 PP$	0,745 0,682
E: Bulharsko V: Bulharsko, Rumunsko, Portugalsko	4	LS MM	$KC =19,878+ 1,492 MZ - 1,695 TI$ $KC = 21,104 + 1,321 MZ -1,239 TI$	0,677 0,634

Zdroj: vlastní výpočty. E značí extrémní pozorování, V značí vybočující pozorování.

4.4 Porovnání regresních modelů pro skupiny zemí

Porovnáme-li výsledné vyhovující lineární regresní modely pro skupiny zemí A, B, C, je vidět, že v každé skupině je pro modelování *KC* vhodné použít jinou kombinaci vysvětlujících proměnných. To ukazuje na fakt, že není možné použít stejný regresní model k popisu závislosti *KC* zemí EU jako celku a *KC* vytvořených skupin zemí.

Pro ověření tohoto závěru byly pro skupiny zemí B a C propočteny všechny regresní modely s podmnožinou vysvětlujících proměnných vhodných pro celý soubor zemí EU. Ani v jednom případě nebyl výsledný regresní model (ani LS ani MM) akceptovatelný pro některou ze skupin zemí B a C. Ani v případech, kdy v dané závislosti nebyla identifikována odlehlá pozorování, nebyly statisticky významné *t*-testy u dílčích regresních koeficientů, byl nízký index determinace, v řadě případů byla zamítnuta normalita reziduí.

Rovněž žádná kombinace vysvětlujících proměnných vhodných pro jednu skupinu zemí (B nebo C), nebyla vhodnou kombinací vysvětlujících proměnných pro druhou skupinu.

Pro ilustraci uvádí tabulka 18 srovnání modelů všech tří skupin pro dva modely

vhodné pro skupinu A (celá EU), z toho první byl s identifikovanými odlehlými pozorováními, druhý model bez odlehlých pozorování. V tabulce jsou * označeny modely s identifikovanými odlehlými pozorováními, kurzívou jsou napsány dílčí regresní parametry, u nichž nejsou statisticky významné hodnoty (na 5% hladině významnosti). Jak je z tabulky 18 zřejmé, regresní modely skupin B a C by se stejnými vysvětlujícími proměnnými, které jsou v modelu pro celý soubor zemí EU, byly nepoužitelné, což opět potvrzuje, že není možné aplikovat stejný model k popisu závislosti *KC* zemí EU jako celku a *KC* vytvořených skupin zemí.

Tabulka 18

Regresní modely pro skupiny B a C se stejnými vysvětlujícími proměnnými jako u modelu pro skupinu A

Soubor A*				R-sq
MM	-10,63 + 0,688 <i>HDP</i> + 0,578 <i>MZ</i>			0,699
LS	-48,551 + 0,579 <i>HDP</i> + 1,313 <i>MZ</i>			0,819
	Soubor B*	R-sq	Soubor C	R-sq
MM	81,62 + 0,793 <i>HDP</i> - 0,958 <i>MZ</i>	0,256	22,96 + 0,529 <i>HDP</i> + 0,214 <i>MZ</i>	0,606
LS	41,736 + 0,115 <i>HDP</i> + 0,782 <i>MZ</i>	0,223	22,96 + 0,529 <i>HDP</i> + 0,214 <i>MZ</i>	0,689
Soubor A				R-sq
MM	62,425 + 0,771 <i>IN</i> - 2,696 <i>MI</i>			0,604
LS	62,425 + 0,771 <i>IN</i> - 2,696 <i>MI</i>			0,638
	Soubor B	R-sq	Soubor C*	R-sq
MM	53,086 + 0,592 <i>IN</i> + 5,017 <i>MI</i>	0,268	38,74 + 0,676 <i>IN</i> - 0,230 <i>MI</i>	0,540
LS	53,086 + 0,592 <i>IN</i> + 5,017 <i>MI</i>	0,369	48,224 + 0,598 <i>IN</i> - 0,795 <i>MI</i>	0,629

Zdroj: vlastní výpočty

5. Závěr

KC jsou základním ukazatelem k porovnání cenové hladiny a k charakterizování nominální konvergence zemí EU. Soubor 27 zemí EU byl rozdělen pomocí shlukové analýzy a následné analýzy rozptylu do dvou relativně homogenních skupin zemí. V souboru evropských zemí se při analýze pomocí lineární regrese vyskytují odlehlá pozorování, která způsobují, že odhady parametrů získané klasickou regresí mohou poskytnout zkreslenou informaci o hodnotách regresních parametrů, mnohdy dokonce mohou nelogicky kvantifikovat opačný směr závislosti. V případech porušení předpokladů normality reziduí a identifikovaných extrémních a vybočujících pozorování, je vhodné použít robustní metodu odhadu parametrů s vysokým bodem selhání (např. LTS, MM), která není citlivá na existenci odlehlých pozorování.

Robustní LTS regrese (používá odhady parametrů metodou nejmenších useknutých čtverců), je schopna odhalit vlivné body v datech i v případech, kdy klasická regresní diagnostika selhává, zejména v případech vícenásobných vztahů, kde často může dojít k „maskování“ odlehlých pozorování.

Výsledky získané klasickou lineární a robustní regresí se mnohdy liší, což je právě důsledek existence odlehlých pozorování. Pokud byla v souboru identifikována vlivná odlehlá pozorování, je vhodnější dát přednost výsledkům získaným robustní regresí s vysokým bodem selhání. Pokud nastane situace, že stejná odlehlá pozorování jsou identifikována jak klasickou, tak i robustní regresí, jsou získané regresní odhady oběma metodami blízké a po ověření dalších předpokladů (zejména normality a homoskedasticity reziduí) lze použít výsledky klasické i robustní regrese. V případě, kdy neidentifikujeme vlivná odlehlá pozorování a je ověřena normalita reziduí, postačí aplikace klasické lineární regrese.

Roztřídění zemí EU do homogennějších skupin se ukázalo být oprávněné, protože *KC* v celém souboru 27 zemí EU a ve vytvořených skupinách jsou ovlivňovány jinými ukazateli a závislost *KC* musí být popsána modely s odlišnými vysvětlujícími proměnnými. Bylo ověřeno, že modely vytvořené pro jednu skupinu zemí nejsou aplikovatelné pro druhou skupinu zemí. Rovněž regresní modely vhodné pro celý soubor 27 zemí EU, nejsou vhodné ani pro skupinu „starších“ vyspělejších členských zemí EU, ani pro skupinu zemí, které přistoupily do EU později. Analýzu *KC* úrovní je proto vhodné provádět odděleně pro skupiny zemí.

Cílem článku nebylo přispět k rozvoji ekonomické teorie nominální konvergence, ale ukázat na některé závislosti a vztahy mezi komparativními cenovými hladinami členských zemí EU a ukazateli charakterizujícími ekonomiku těchto zemí z různých hledisek a dát tak i podnět k dalšímu zkoumání a k rozvoji ekonomické teorie nominální i reálné konvergence. Zejména bude vhodné provést dynamickou analýzu vztahu *KC* a ekonomických ukazatelů, která v této rozsáhlé analýze nebyla provedena a ověřit, jestli i v dynamice je nutné modelovat závislost *KC* jinými modely pro skupiny zemí nebo zda lze použít jeden model pro země EU jako celek.

Analýza byla provedena na datech roku 2008, takže vliv krize se v datech dosud nemohl projevit.

Teoretický dodatek.

Základní myšlenka robustní regrese spočívá v nalezení postupů, které modifikují metodu nejmenších čtverců tak, aby byla redukována citlivost MNČ na odlehlá pozorování ale současně byly zachovány dobré vlastnosti odhadů získaných MNČ. V případě nekontaminovaných dat jsou robustní regresní funkce blízké regresním funkcím získaným klasickou lineární regresí. Je chybné tvrzení, že robustní postupy ignorují odlehlá pozorování, naopak, robustní regresní metody s vysokým bodem selhání jsou schopné existenci odlehlých (extrémních i vybočujících) pozorování odhalovat (viz např. Rousseeuw, 2003; Yohai, 1987; Rousseeuw, 1984).

Odhad metodou nejmenších useknutých čtverců (the Least Trimmed Squares estimator), který navrhl Rousseeuw v roce 1984 získáme minimalizací součtu čtverců h reziduí

$$\sum_{i=1}^h e_{(i)}^2 = \min, \quad (2)$$

kde je i -tá pořadová statistika čtverců reziduí (uspořádaných ve vzestupném pořadí).

(Např. systém SAS, použitý k výpočtům, užívá celočíselnou hodnotu h ležící mezi $\left\lfloor \frac{n}{2} \right\rfloor + 1$ a $\left\lfloor \frac{3n+p+1}{4} \right\rfloor$. Nastavená hodnota (default) $h = \left\lfloor \frac{3n+p+1}{4} \right\rfloor$ odpovídá bodu selhání kolem 25%).

LTS – odhad má relativně nízkou asymptotickou vydatnost a je převážně využíván jako počáteční odhad v odhadech metodou MM a v robustní identifikaci odlehlých pozorování (viz Rousseeuw, 2003; Ruppert, 1990; Chen, 2002).

MM – odhady jsou definovány pomocí třístupňového postupu (podrobněji viz např. Yohai, 1987; Rousseeuw, 2003; Maronna, 2006; Chen, 2002). V prvním stupni jsou vypočteny počáteční regresní odhady parametrů $\hat{\beta}_j$, které jsou konzistentní s vysokým bodem selhání, ale nemusí být vydatné. Ve druhém stupni je vypočten robustní odhad směrodatné odchylky chyb na základě reziduálních odchylek vypočtených z modelu s počátečními odhady $\hat{\sigma}$. Ve třetím stupni jsou iterativní procedurou vypočteny *M*-odhady regresních parametrů s vhodnou neklesající funkcí ψ řešením vztahu

$$\sum_{i=1}^n \mathbf{x}_i \psi \left(\frac{y_i - \mathbf{x}_i^T \hat{\beta}}{\hat{\sigma}} \right) = 0, \quad (3)$$

kde $\hat{\sigma}$ je robustní odhad směrodatné odchylky reziduí (vypočtený ve 2. kroku), $\psi = \rho'$ je derivace vhodné ztrátové funkce ρ . (Byla aplikována Tukeyho ztrátová funkce ρ , která je implementována v obou použitých softwarech SAS a S-Plus – viz např. Rousseeuw (2003), Maronna (2006), Chen (2002) a uvedené internetové zdroje).

Identifikace odlehlých pozorování

Pozorování, u nichž se odlehlé hodnoty vyskytují u vysvětlované (závisle) proměnné y se označují jako extrémní odlehlá pozorování (*outliers*), odlehlé hodnoty u vysvětlujících proměnných (nezávisle proměnných) se nazývají vybočující pozorování (*leverage points*), přičemž tzv. „dobrá vybočující pozorování“ (*good leverage points*) jsou vybočující pozorování, která nejsou současně extrémními odlehlými pozorováními. Dobrá vybočující pozorování mají omezený vliv na kvalitu regresních odhadů, někdy mohou dokonce i zlepšit jejich přesnost. Tzv. „špatná odlehlá pozorování“ (*bad leverage points*) jsou taková, která mají hodnoty vysvětlující proměnné vzdálené od většiny ostatních pozorování. Existence špatných odlehlých pozorování výrazně snižuje přesnost regresních odhadů, mnohdy i způsobuje mylnou informaci o směru závislosti. Pozorování, která jsou současně extrémní a vybočující jsou vlivná pozorování (*influential points*). Podrobněji viz např. Rousseeuw (2003), Maronna (2006), Hubert (2008).

V analýze byly použity následující nástroje identifikace odlehlých pozorování, které jsou rovněž implementovány v použitém softwaru SAS 9.1 a S-Plus 6.2 a byly využity v analýze:

- *robustní vzdálenost*

$$RD(x_i) = \sqrt{[x_i - \mathbf{T}(\mathbf{X})^T \mathbf{C}(\mathbf{X})^{-1}] [X_i - \mathbf{T}(X)]}, \quad (4)$$

kde $\mathbf{T}(\mathbf{X})$ a $\mathbf{C}(\mathbf{X})$ jsou matice robustních měř polohy a měřítka vysvětlujících proměnných.

Vybočující pozorování (leverage points) mají hodnoty $RD(x_i) > \sqrt{\chi_{p-1;1-\alpha}^2}$ (kde p je počet parametrů regresní funkce).

- *Standardizovaná rezidua* $r_i / \hat{\sigma}$ založená na robustních odhadech umožní identifikovat extrémní odlehlá pozorování (outliers); ($\hat{\sigma}$ je robustní odhad parametru měřítka – podrobněji viz např. Rousseeuw, 2003; Chen, 2002).

- *Robustní regresní diagnostický graf* (graf standardizovaných reziduí versus robustní vzdálenosti $RD(x_i)$). V grafu jsou vyznačeny horizontální a vertikální čáry sloužící k identifikaci. Body, jejichž standardizovaná rezidua od robustní

regresní funkce leží za hranicí danou hodnotou $\sqrt{\chi_{1;1-\alpha/2}^2}$, považujeme za regresní extrémní odlehlá pozorování (outliers), zatímco pozorování, pro něž

robustní vzdálenost $RD(x_i)$ přesahuje hodnotu $\sqrt{\chi_{p;1-\alpha/2}^2}$ identifikují vybočující pozorování (leverage points). (Podrobněji viz např. Rousseeuw, 2003; Chen, 2002).

Metody výběru vhodného robustního modelu

Robustní diagnostická kritéria implementovaná v SAS 9.2 (blíže viz. např. Maronna, 2006; Rousseeuw, 2003; Chen, 2002):

- *Robustní index determinace R-squared*

$$R^2 = \frac{\sum_{i=1}^n \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum_{i=1}^n \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right)}{\sum_{i=1}^n \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} = 1 - \frac{\sum_{i=1}^n \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right)}{\sum_{i=1}^n \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)}, \quad (5)$$

kde $\hat{\beta}$ je MM odhad parametru β , $\hat{\mu}$ je MM odhad parametru polohy, $\hat{s}$ je MM-odhad parametru měřítka v úplném modelu, ρ je ztrátová funkce.

- *Významnost robustních testů parametrů regresního modelu*: robustní dílčí t -testy, robustní F -testy a robustní Waldovy testy (s testovým kritériem s rozdělením chí-kvadrát).
- *Robustní vzdálenost (deviance)*

$$D = 2(\hat{s})^2 \sum \rho\left(\frac{y_i - x_i^T \hat{\beta}}{\hat{s}}\right) \quad (6)$$

kde $\hat{\beta}$ je MM odhad parametru β , $\hat{s}$ je MM odhad parametru měřítka.

- *Robustní Akaikeho informační kritérium (AICR)*

$$AICR(p; \alpha, \rho) = 2 \sum_{i=1}^n \rho(r_{i,p}) + \alpha p = 2 \sum_{i=1}^n \rho \left(\frac{y_i - x_i^T \hat{\beta}}{\hat{\sigma}} \right) + \alpha p, \quad (7)$$

- *Robustní Schwarzovo Bayesovské informační kritérium (BICR)*

$$BICR = 2 \sum_{i=1}^n \rho \left(\frac{y_i - x_i^T \hat{\beta}}{\hat{\sigma}} \right) + p \ln(n) \quad (8)$$

kde $(y_i - x_i^T \hat{\beta})$ jsou regresní rezidua spojená s *MM*-odhady parametrů, $\hat{\sigma}$ je robustní odhad σ , p je počet parametrů.

Literatura

- ANTOCH, J.; VORLÍČKOVÁ, D. 1992. *Vybrané metody statistické analýzy dat*. Praha: Academia, 1992. ISBN 80-200-0204-9.
- BLATNÁ, D. 1996. *Neparametrické metody*. Praha: VŠE v Praze, 1996. ISBN 80-7089-607-3.
- CHEN, C. 2002. Robust Regression and Outlier Detection with the ROBUSTREG procedure. SUGI Paper 265-27. SAS Institute Inc., Cary, NC. 2002.
- FRAIT, J.; KOMÁREK, L. 2001. Na cestě do Evropské unie: nominální a reálná konvergence v tranzitivních ekonomikách. *Finance a úvěr*. 2001, Vol. 51, No. 6, pp. 314–330. ISSN 0015-1920.
- HADI, A. S.; SIMONOFF, J. S. 1993. Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of the American Statistical Association*. 1993, Vol 88, No. 424, pp. 1264–1272.
- HEBÁK, P.; HUSTOPECKÝ, J.; PECÁKOVÁ, I.; PRŮŠA, M.; ŘEZANKOVÁ, H.; SVOBODOVÁ, A.; VLACH, P. 2005. *Vícerozměrné statistické metody* (3). Praha: Informatorium, 2005. ISBN 80-7333-039-3.
- HEBÁK, P.; HUSTOPECKÝ, J.; MALÁ, I. 2005. *Vícerozměrné statistické metody* (2). Praha: Informatorium, 2005. ISBN 80-7333-036-9.
- HUBERT, M.; ROUSSEEUW, P. J.; VAN AELST. 2008. High-Breakdown Robust Multivariate Methods. *Statistical Science*. 2008, Vol 23, No. 1, pp. 92–119. ISSN0883-4237
- HUŠEK, R. 2007. *Ekonometrická analýza*. Praha: Nakladatelství Oeconomica, 2007. ISBN 978-80-245-1300-3.
- MARONNA, R. A.; MARTIN, R. D.; YOHAI, V. J. 2006. *Robust Statistics. Theory and Methods*. London: J Wiley, 2006. ISBN-13 978-0-470-01092-1
- OLIVE, D. J. 2002. Applications of Robust Distances for Regression. *Technometrics*. 2002, No. 44, pp. 64–71.
- RAO, C. R.; TOUTENBURG, H. 1995. *Linear Models. Least Squares and Alternatives*. New York: Springer-Verlag, 1995. ISBN 0-387-95462-8.
- ROUSSEEUW, P. J. 1984. Least Median of Squares Regression. *Journal of the American Statistical Association*. 1984, No. 79, pp. 878–880.
- ROUSSEEUW, P. J.; LEROY, A. M. 2003. *Robust Regression and Outlier Detection*. New Jersey: J. Wiley, 2003. ISBN 0-471-48855-0.
- ROUSSEEUW, P. J.; VAN ZOMEREN, B. C. 1990. Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*. 1990, No. 85, pp. 633–639.
- RUPPERT, D.; CARROLL, R. J. 1990. Trimmed Least Squares Estimation in the Linear Model. *Journal of the American Statistical Association*. 1990, No. 75, pp. 828–838.

- SWALLOW, W. H.; KIANIFARD, F. 1996. Using robust scale estimates in detecting multiple outliers in linear regression. *Biometrics*. 1996, No. 52, pp. 545–556.
- ŠAROCH, S.; ŽÁK, M. (ed.) 2004. *Česká ekonomika a ekonomická teorie*. 1. vyd. Praha : Academia, 2004. 266 s. ISBN 80-200-1129-3.
- VINTROVÁ, R.; ŽDÁREK, V. 2006. Konvergence České republiky a Slovenské republiky – současný stav a vybrané problémy. *Ekonomický časopis*. 2006, č. 5, s. 468–489. ISSN 0013-3035.
- VINTROVÁ, R.; ŽDÁREK, V. 2006. *Nové členské země Evropské unie a příprava na přijetí společné měny*. Praha: Bulletin CES VŠEM, č. 17/2006, s. 1–3. ISSN 1801-1578.
- VINTROVÁ, R. 2010. *Cenová a mzdová konvergence nových členských zemí EU*. Praha: Bulletin CES VŠEM, 1/2010, s. 5–7. ISSN 1801-6871.
- UCLA. Academy Technology Services. Regression with SAS. <http://www.ats.ucla.edu/stat/sas>
- YOHAI, V. J. 1987. High Breakdown-point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*. 1987, Vol. 15, No 20, pp. 642–656.
- ZVÁRA, K. 1989. *Regresní analýza*. Praha: Academia, 1989. ISBN 80-200-0125-5.
- ŽDÁREK, V. 2006. *Nominální konvergence v České republice*. Praha: Bulletin CES VŠEM, č. 23/ 2006, s. 5–6. ISSN 1801-6871.
- ŽDÁREK, V. 2010. *Cenová konvergence nových členských zemí EU- strukturální pohled*. Praha: Bulletin CES VŠEM, č. 2/2010. ISSN 1801-1578.
- ŽDÁREK, V.; ŠINDEL, J. 2007. Real and Nominal Convergence and the New EU Members States – Actual State and Implication. *Prague Economic Papers*. 2007, Vol. 14, No. 3, pp.195–219. ISSN 1210-0455.

ROBUST REGRESSION IN ANALYSIS OF COMPARATIVE PRICE LEVELS OF EU COUNTRIES

Dagmar Blatná, University of Economics, Prague, nám. W. Churchilla 4, CZ – 130 67 Praha 3 (blatna@vse.cz).

Abstract

The values of comparative price levels vary greatly in individual EU countries and depend on many different economic factors. The EU countries were divided into two distinguishable groups. Both OLS and robust regressions were used to analyze the influence of the comparative price levels on selected indicators. If outliers and leverage points were identified using robust outliers' detection and the results of the OLS and robust fits differed significantly, the robust fits are preferred. When differences were not significant, the OLS fits can be used. Models for 27 EU countries and groups of countries differ significantly with respect to indicators included.

Keywords

Comparative price levels, EU countries, OLS regression, robust regression, robust outliers' detection

JEL Classification

E31, O57, O11