Informační Bulletin
České Statistické Společnosti
č. 2. listopad 2000, ročník 11

Poznámka k příspěvkům ze Statistických dnů v Ostravě

Statistické dny pořádá každoročně Česká statistická společnost v různých místech České republiky s cílem umožnit setkání statistiků a uživatelů statistiky z regionu, poskytnout jim prostor pro presentování příspěvků i pro diskusi aktuálních otázek, a tak přispívat ke zvyšování statistické vzdělanosti v České republice.

Ostravské statistické dny se konaly ve dnech 20. -21. června 2000 na Přírodovědecké fakultě Ostravské university a na jejich organizaci se podílela i katedra aplikované matematiky VŠB-TU Ostrava. Celkový počet účastníků byl okolo třiceti. V programu převažovali přednášející z Ostravy a blízkého okolí, měli jsme však i potěšení z přednášek tří účastníků vzdálenějších. Většina účastníků připravila i písemné verse svých sdělení, která jsou uvedeny v následujících dvou číslech Bulletinu ČStS. Pokud jste zklamáni, že mezi nimi nenalézáte to, co právě vy jste si chtěli přečíst, obraťte na dotyčné autory, jistě mají něco v šuplíku. V Bulletinu nejsou texty těchto dvou přednášek:

- A. Bartkowiak (Wroclawská universita, Institut informaticzny):
  Recognizing shape and atypical observations in multivariate data.
- R. Briš (TUO, Fakulta elektrotechniky a informatiky):
  Efektivní postup pro implementaci provozních dat při verifikaci spolehlivosti vysoce spolehlivých prvků.

Pokračování na str. 23.

Vliv vybraných charakteristik struktury trhu práce na vývoj nezaměstnanosti v ČR[1]

Jana Hančlová, Milan Šimek
Ekonomická fakulta, VŠB-TU Ostrava

Abstrakt: Článek se zabývá zkoumáním modifikované Beveridgeovy křivky. Vzájemný bivariantní vztah mezi nezaměstnaností a volnými pracovními místy selhává při identifikaci dlouhodobého rovnovážného vztahu, jestliže toky nezaměstnaných jsou endogenní. Modifikace Beveridgeovy křivky spočívá v rozšíření bivariantního vztahu o další proměnné popisující strukturu trhu práce a zavedení náhodných složek, což nám umožňuje provést kointegrační analýzu při zkoumání dlouhodobých (ne)rovnovážných vztahů na trhu práce. Empirická studie je věnována českému trhu práce za období 1992 - 2000.

Klíčová slova: kointegrační analýza, nezaměstnanost, Beveridgeova křivka, struktura trhu práce

Úvod

Vzájemný inverzní vztah mezi nezaměstnaností a volnými pracovními místy lze charakterizovat Beveridgeovou křivkou a v případě exogenních toků nezaměstnaných je možné hledat dlouhodobý rovnovážný bivariantní vztah. Článek sleduje rozšířený dlouhodobý rovnovážný model trhu práce při respektování vlivu dlouhodobě nezaměstnaných nebo mladistvých nezaměstnaných. Empirická studie byla provedena na kvartálních časových řadách pro český trh práce od roku 1992 do konce prvního pololetí 2000. Při hledání dlouhodobého rovnovážného vztahu v modifikované Beveridgeově křivce bylo použito kointegrační analýzy a odhadu modelu korekce chyb.

Modifikace Beveridgeovy křivky

Řada ekonomů se ve svých empirických studiích zabývala nerovnovážnými modely toků ekonomicky aktivních osob na trhu práce. Absolutní přírůstek počtu zaměstnaných lze vyjádřit následujícím vztahem:

$$\Delta E_t = E_t - E_{t-1} \equiv H_t - Q_t \qquad (1)$$

kde H_t je počet přijatých zaměstnanců, Q_t počet ukončených pracovních poměrů.

[1] Tento příspěvek vznikl s podporou grantu GAČR č. 402/00/1165 a v rámci institucionálního výzkumného úkolu na EkF VŠB-TU Ostrava CEZ:J17/98:275100015

Funkce nově přijatých pracovníků lze vyjádřit jako homogenní Cobb-Douglasovu produkční funkci s konstantním výnosem:

$$H_t = \beta_0 \cdot U_{t-1}^{\alpha} \cdot V_{t-1}^{1-\alpha}, \tag{2}$$

kde β_0 je efektivnost hledání práce (podíl úspěšných kontaktů k celkovému počtu kontaktů při hledání práce za období *t*), U_t počet nezaměstnaných, V_t počet volných pracovních míst.

Počet ukončených pracovních poměrů je lineárně závislý na počtu zaměstnaných v čase *(t-1)* s parametrem δ_0. Absolutní změnu počtu zaměstnaných vyjadřuje *nerovnovážný model trhu práce:*

$$\ddot{A}E_t = \beta_0 \cdot U_{t-1}^{\alpha} \cdot V_{t-1}^{1-\alpha} - \delta_0 \cdot E_{t-1} \tag{3}$$

a relativní přírůstek zaměstnaných e_t

$$e_t = \beta_0 \cdot ue_{t-1}^{\alpha} \cdot ve_{t-1}^{1-\alpha} - \delta_0, \tag{4}$$

kde $e_t = \Delta E_t / E_{t-1}$, $ue_{t-1} = U_{t-1} / E_{t-1}$, $ve_{t-1} = V_{t-1} / E_{t-1}$.

Dlouhodobý rovnovážný model trhu práce, který je odvozen v logaritmické formě v rovnici (5) jako hyperbolický vztah, respektuje podmínky, že množství práce a zdroje pracovní síly jsou konstantní tj. $e_t = 0$.

$$\ln ue_{t-1} = -\frac{1-\alpha}{\alpha} \cdot \ln ve_{t-1} + \frac{1}{\alpha} \cdot \ln(\delta_0 / \beta_0) \quad . \tag{5}$$

Anthony (1999) zavedl ve své práci do tohoto deterministického modelu *náhodnou složku* ε_t v multiplikativní formě do rovnice (3):

$$\ddot{A}E_t = \beta_0 \cdot U_{t-1}^{\alpha} \cdot V_{t-1}^{1-\alpha} \cdot \left\{e^{\mu_{t-1}}\right\} - \delta_0 \cdot \left\{e^{\nu_{t-1}}\right\} \cdot E_{t-1} \tag{6}$$

resp. ve formě logaritmické

$$\ln ue_t = -\frac{1-\alpha}{\alpha} \cdot \ln ve_t + \frac{1}{\alpha} \cdot \ln(\delta_0 / \beta_0) + \frac{1}{\alpha} \cdot \varepsilon_t \tag{7}$$

kde $\varepsilon_t = \nu_t - \mu_t$, $\mu_t \sim N(0, \sigma_\mu^2)$, $\nu_t \sim N(0, \sigma_\nu^2)$.

Rovnice (7) umožňuje zkoumat existenci dlouhodobého rovnovážného vztahu mezi proměnnými ue_t a ve_t za předpokladu, že jsou tyto proměnné nestacionární a jsou integrovány řádu jedna[2] a náhodná ε_t je stacionární.

---

[2] Stacionární procesy se označují I(0). Pokud se nestacionární časová řada transformuje první diferencí na proces s krátkou pamětí tzn. stacionární, je původní časová řada integrována řádu jedna tj. I(1).

Bivariantní model trhu práce předpokládá, že parametry β_0 a δ_0 jsou fixní, tzn. toky mezi nezaměstnaností a pracovními místy jsou exogenní. V další části příspěvku rozšíříme bivariantní rovnovážný model, který bude předpokládat, že toky mezi nezaměstnaností a volnými pracovními místy jsou endogenní a jsou determinovány vývojem struktury trhu práce, kterou budou reprezentovat následující základní charakteristiky:
lg_t – počet registrovaných dlouhodobě nezaměstnaných (déle než 1 rok),
y_t – počet registrovaných mladistvých nezaměstnaných (15 –19 let).

Modifikovaný nerovnovážný model trhu práce lze nyní vyjádřit úpravou rovnice (6):

$$\Ä E_t = \beta_0 \cdot \lg_{t-1}^{\beta_1} \cdot y_{t-1}^{\beta_2} \cdot U_{t-1}^{\alpha} \cdot V_{t-1}^{1-\alpha} \cdot \left\{e^{\mu_{t-1}}\right\} - \delta_0 \cdot \left\{e^{\nu_{t-1}}\right\} \cdot E_{t-1} \quad (8)$$

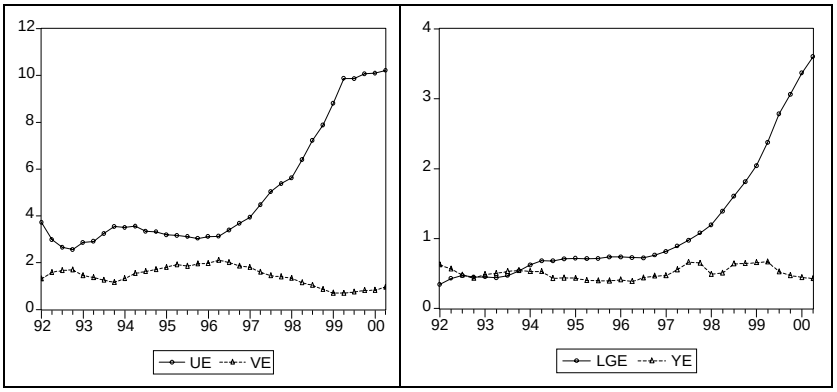
a tudíž *rozšířený rovnovážný model trhu práce* je dán vztahem:

$$\ln ue_t = -\frac{1-\alpha}{\alpha} \cdot \ln ve_t - \frac{\beta_1}{\alpha} \cdot \ln \lg e_t - \frac{\beta_2}{\alpha} \cdot \ln ye_t + \frac{1}{\alpha} \cdot \ln\left(\frac{\delta_0}{\beta_0}\right) + \frac{1}{\alpha} \cdot \varepsilon_t \quad (9)$$

kde $\varepsilon_t = \nu_t - \mu_t$, $\lg e_t = \frac{\lg_t}{E_t}$, $ye_t = \frac{y_t}{E_t}$.

Empirická studie

Testování bivariantního a rozšířeného rovnovážného modelu trhu práce bylo provedeno na kvartálních časových řadách v České republice od roku 1992 do 2. kvartálu 2000 (tj. 32 pozorování). Všechny časové řady byly sezónně očištěny multiplikativním modelem v softwarovém produktu SPSS. Následující obrázek 1 prezentuje vývoj časových řad ue_t a ve_t.



Obrázek 1: Vývoj poměrových ukazatelů ue_t a ve_t.

Z obrázku 1 vyplývá relativně stabilizovaný vývoj „míry“ nezaměstnanosti i počtu volných pracovních míst v období 1992–1996. V roce 1997 v souvislosti s prohlubujícími se problémy české ekonomiky nastal prudký zlom ve stávajících trendech. V případě „míry“ nezaměstnanosti došlo k intenzivnímu růstu až ke hranici 10% s určitou stagnací v první polovině roku 2000. K poklesu počtu volných pracovních míst docházelo již v roce 1996 s nejnižší úrovní ke konci roku 1999. V roce 2000 došlo k mírnému nárůstu, což může souviset s určitým ekonomickým oživením v České republice. Koncem roku 1998 začal narůstat počet dlouhodobě nezaměstnaných, což bylo důsledkem prohlubujících se problémů v ekonomice a na českém trhu práce. Počet mladistvých nezaměstnaných se mírně rozkolísal od roku 1997 a již v roce 2000 se ustálil na původní úrovni.

Všechny sledované časové řady v logaritmické formě byly testovány na přítomnost jednotkového kořene prostřednictvím ADF (rozšířeného Dickey-Fullerova) testu a PP (Phillips–Perronova) testu. Výsledky shrnuje tabulka 1. Je zřejmé, že všechny časové řady jsou nestacionární a jsou integrovány řádem 1 pro obě testovací statistiky. Na základě získaných výsledků můžeme přistoupit ke kointegrační analýze.

Tabulka 1: Testování nestacionárnosti časových řad

proměnná	ADF statistika	PP statistika	proměnná	ADF statistika	PP statistika
$yue_t = \ln ue_t$	-1,407	-2,464	$due_t = yue_t - yue_{t-1}$	-2,449	-3,325
$yve_t = \ln ve_t$	-1,786	-1,633	$dve_t = yve_t - yve_{t-1}$	-2,436	-3,092
$y\lg e_t = \ln\lg e_t$	-1,851	-0,321	$d\lg e_t = y\lg e_t - y\lg e_{t-1}$	-1,839	-2,542
$yye_t = \ln ye_t$	-2,642	-2,382	$dye_t = yye_t - yye_{t-1}$	-4,766	-4,540
5% kritická hodnota	-4,271	-4,261	**5% kritická hodnota**	-1,521	-1,952

Testy kointegrace budou aplikovány pro následující modely:

- Model_1 – bivariantní vztah yue_t a yve_t podle odvozené rovnice (5),
- Model_2 – rozšířený model - yue_t, yve_t, $ylge_t$,
- Model_3 – rozšířený model - yue_t, yve_t, yye_t,
- Model_4 – modifikovaný model - yue_t, yve_t, $ylge_t$, yye_t.

Testy kointegrace, které byly realizovány prostřednictvím statistického softwaru Eviews 3.1, sledovaly Johansenovu metodu (Johansen (1991, 1995)), která vychází z restriktivních vektorových autoregresivních modelů.

Model_1 testoval dlouhodobý rovnovážný vztah mezi *yue* a *yve*. Johansenův test nezamítl hypotézu, že *neexistuje žádný dlouhodobý rovnovážný vztah mezi přírůstkem časové řady* ue_t *a* ve_t *na 5% hladině významnosti.* Tento výsledek potvrzuje, že nelze sledovat toky jako exogenní, protože jsou determinovány dalšími faktory (viz např. Gottvald, 1999).

Model_2 (yue_t, yve_t, $ylge_t$) – rozšířil bivariantní model o proměnnou podílu dlouhodobě nezaměstnaných logaritmickém vyjádření. Johansenův test - indikoval jednu kointegrační rovnici:

$$EQ1 = -2{,}348 + 0{,}052 \cdot trend + yue_t - 0{,}685 \cdot yve_t - 1{,}821^{*} \cdot y\,lg\,e_t \qquad (10)$$

Dosažené výsledky ukazují, že z dlouhodobého pohledu existuje statisticky významný pozitivní vliv míry dlouhodobě nezaměstnaných na „míru“ nezaměstnanosti (1,821). Tento vztah lze označit za statisticky významnější než je vliv hlášených volných pracovních míst.

Model korekce chyb byl odhadnout:

$$\begin{bmatrix} \Ä\, yue_t \\ \Ä\, yve_t \\ \Ä\, y\lg e_t \end{bmatrix} = \begin{bmatrix} 0{,}009 \\ -0{,}012 \\ -0{,}007 \end{bmatrix} + \begin{bmatrix} 0{,}004^{*} \\ -0{,}003 \\ 0{,}004^{*} \end{bmatrix} \cdot trend + \qquad (11)$$

$$+ \begin{bmatrix} -0{,}143 & -0{,}127 & -0{,}574^{*} \\ 0{,}459 & 0{,}391 & 0{,}668^{*} \\ -0{,}146 & 0{,}103 & 0{,}136 \end{bmatrix} \cdot \begin{bmatrix} \Ä\, yue_{t-1} \\ \Ä\, yve_{t-1} \\ \Ä\, y\lg e_{t-1} \end{bmatrix} + \begin{bmatrix} 0{,}188^{*} \\ -0{,}220 \\ 0{,}222^{*} \end{bmatrix} \cdot EQ1$$

Z krátkodobého hlediska působí na změnu nezaměstnanosti negativně přírůstek dlouhodobě nezaměstnaných z předcházejícího období (-0,574). Odchylka od dlouhodobé rovnováhy na trhu práce může být vyrovnána tempem 18.8% (viz. koeficient zatížení).

Model_3 (yue_t, yve_t, yye_t) – modifikoval bivariantní model o logaritmickou proměnnou podílu nezaměstnaných mladistvých ve věku 15 – 19 let. Johansenův test opět zaznamenal jednu statisticky významnou kointegrační rovnici:

$$EQ2 = -1{,}484^{*} - 0{,}033^{*} \cdot trend + yue_t + 0{,}616^{*} \cdot yve_t - 0{,}553^{*} \cdot yye_t \qquad (12)$$

Kointegrační rovnice *EQ2* empiricky dokazuje dlouhodobý statisticky významný negativní vliv volných pracovních míst na „míru“ nezaměstnanosti (-0,616), který je ve srovnání s faktorem mladistvých nepatrně silnější, ale má opačný směr působení (0,553).

Model korekce chyby lze opět prostřednictvím softwaru Eviews pro délku zpoždění jednoho kvartálu odhadnout v rovnici (13). Z krátkodobého

---

[*] Odhadnuté koeficienty jsou statisticky významné na 5% hladině významnosti.

hlediska je „míra“ nezaměstnanosti statisticky významně ovlivněna pouze přírůstkem volných pracovních míst negativně (-0.334).

$$\begin{bmatrix} \Ä\, yue_t \\ \Ä yve_t \\ \Ä yye_t \end{bmatrix} = \begin{bmatrix} 0{,}022^{*} \\ -0{,}006 \\ -0{,}032 \end{bmatrix} + \begin{bmatrix} 0{,}321 & -0{,}334^{*} & -0{,}149 \\ 0{,}034 & 0{,}692^{*} & 0{,}034 \\ -0{,}146 & 0{,}103 & 0{,}136 \end{bmatrix} \cdot \begin{bmatrix} \Ä\, yue_{t-1} \\ \Ä\, yve_{t-1} \\ \Ä\, yye_{t-1} \end{bmatrix} + \begin{bmatrix} -0{,}013 \\ -0{,}142 \\ 0{,}849^{*} \end{bmatrix} \cdot EQ\,2 \qquad (13)$$

Model_4 (yue_t, yve_t ,$ylge_t$, yye_t) – zahrnul do bivariantního modelu obě proměnné popisující strukturu registrované nezaměstnanosti. Johansenův test, který je vhodný při zkoumání dlouhodobého rovnovážného vztahu pro vícerozměrné modely indikoval dvě statisticky významné normalizované kointegrační rovnice:

$$EQ3 = -2{,}564 + 0{,}002 \cdot trend + yue_t - 0{,}711^{*} \cdot y\,lg\,e_t - 1{,}409 \cdot yye_t \qquad (14)$$

$$EQ4 = 0{,}266^{*} - 0{,}091^{*} \cdot trend + yve_t + 1{,}505^{*} \cdot y\,lg\,e_t - 1{,}659 \cdot yye_t$$

Model korekce chyby lze opět prostřednictvím software Eviews pro délku zpoždění 1 kvartálu odhadnout v následující formě:

$$\begin{bmatrix} \Ä yue_t \\ \Ä yve_t \\ \Ä y\lg e_t \\ \Ä yye_t \end{bmatrix} = \begin{bmatrix} 0{,}040 & -0{,}154 & -0{,}525^{*} & -0{,}103 \\ 0{,}471 & 0{,}424 & 0{,}667 & -0{,}018 \\ -0{,}044 & 0{,}063 & 0{,}156 & -0{,}072 \\ 0{,}212 & -0{,}753^{*} & -0{,}205 & -0{,}076 \end{bmatrix} \cdot \begin{bmatrix} \Ä yue_{t-1} \\ \Ä yve_{t-1} \\ \Ä y\lg e_{t-1} \\ \Ä yye_{t-1} \end{bmatrix} + \begin{bmatrix} 0{,}111^{*} \\ -0{,}134 \\ 0{,}133^{*} \\ -0{,}112 \end{bmatrix} \cdot EQ5 + \begin{bmatrix} 0{,}071^{*} \\ -0{,}072 \\ 0{,}057^{*} \\ -0{,}013 \end{bmatrix} \qquad (15)$$

kde

$$EQ5 = -2{,}845^{*} + 0{,}098 \cdot trend + yue_t - 1{,}057 \cdot yve_t - 2{,}301^{*} \cdot y\,lg\,e_t + 0{,}344 \cdot yye_t \qquad (16)$$

Dlouhodobý rovnovážný vztah mezi „mírou“ nezaměstnanosti a dlouhodobě nezaměstnanými je statisticky významný s pozitivním působením, ale krátkodobě je vliv opačný. Další proměnná popisující strukturu trhu práce (mladiství nezaměstnaní) není statisticky významná nejen v případě

[*] Odhadnuté koeficienty jsou statisticky významné na 5% hladině významnosti.

dlouhodobého pohledu, ale i z hlediska krátkodobého odchýlení se od dlouhodobě rovnovážného vztahu. Statisticky významná rychlost přizpůsobení se rovnovážnému vztahu byla prokázána pro faktory nezaměstnanost (0,111) a dlouhodobě nezaměstnaní (0,133).

Shrnutí a závěr

Johansenův test kointegrace pro model_1 ukázal, že neexistuje statisticky významná dlouhodobá rovnováha pro změny vzájemného vztahu podílových časových řad nezaměstnaných a volných pracovních pro český trh práce za období 1992 – polovina roku 2000 (tzn. model hledání práce nezahrnuje pouze exogenní vztah, nýbrž také endogenní vztah).

Rozšířená Beveridgeova křivka indikovala statisticky významný dlouhodobě rovnovážný vztah mezi nezaměstnanými a volnými pracovními místy při zahrnutí dalších strukturálních proměnných. Výsledky dalších modelů dokumentují, že problematika dlouhodobé nezaměstnanosti patří ke klíčovým problémům českého trhu práce. Počet a struktura volných pracovních míst na trhu práce zde působí jako významný determinant procesu snižování nezaměstnanosti i dlouhodobé nezaměstnanosti a z hlediska přizpůsobení se dlouhodobě rovnovážnému vztahu na trhu práce patří mezi klíčové faktory, což potvrzuje základní Beveridgeův vztah.

Literatura

Anthony J.D.F. (1999): The relationship between unemployment and vacancies in Australia. Applied Economics, No. 31, pp. 641-652.

Bewley, R.A. (1979): The dynamic behaviour of unemployment and unfilled vacancies in Great Britain: 1958-1971, Applied Economics, No. 11, pp. 303-308.

Eviews User's Guide (1998). Quantitative Micro Software.

Gottvald, J. (1999): Structural Adjustment of the Czech Labour Market. In: Proceedings of Papers No. 06-2-1, EALE Conference, 1999, Regensburg, Germany

Johansen,S. (1991): Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. Econometrica, No. 59, pp. 1551-1580.

Johansen,S. (1995):Likelihood-based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press.

Ekonometrický model OKD-1

Miroslav Liška

Přírodovědecká fakulta, Ostravská univerzita

Abstrakt: Článek pojednává o ekonometrickém modelu OKD-1 pro analýzu a prognózu výrobní, technické a ekonomické činnosti Ostravsko-karvinských dolů (OKD). Představuje ho soustava 114 simultánních rovnic, které vyjadřují regresní a technicko-bilanční vztahy mezi 205 proměnnými. Jedná se o rekurzívní model ekonometricko-normativního typu, respektující nelineární a dynamický charakter modelovaných vztahů. Základem ekonometrického modelu OKD-1 je soustava vzájemně provázaných produkčních funkcí. Hlavní produkční funkcí ekonometrického modelu OKD-1 je produkční funkce pro těžbu uhlí. Výsledným tvarem je vícefaktorová produkční funkce, vyjadřující závislost průměrné denní těžby dohromady na 12 výrobních faktorech včetně ukazatelů důlně-geologických podmínek.

Klíčová slova: ekonometrický model, soustava simultánních rovnic, produkční funkce, Ostravsko-karvinské doly (OKD), těžba uhlí, důlně-geologické podmínky.

Úvod

Ekonometrický model OKD-1 představuje aplikační výsledek [1], který přesahuje svým významem hranice Ostravy a časový horizont svého vzniku. Řada poznatků a zkušeností zůstává doposud aktuální a některé návraty k ekonometrickému modelu OKD-1 v 90.létech potvrzují platnost a použitelnost dosažených výsledků.

Původní oblastí ekonometrického modelování jsou makroekonomické problémy [3]. Objevují se také případy využívání na podnikové úrovni (např. modely CHEPOS, SIGMA, VÍTKOVICE) [2].

Ostravsko-karvinské doly (OKD) jsou důležitým podnikem, jenž plní hlavní výrobní úkoly v odvětví hlubinné těžby černého uhlí. Koncepce sestavení první verze ekonometrického modelu činnosti OKD (označ. OKD-1) byla zvolena tak, aby co nejlépe formulovala obsahové a věcné stránky podstaty

vztahů mezi hlavními výrobními faktory důlních i povrchových výrobních procesů OKD. Bylo nutno vyřešit následující dílčí úkoly:

- identifikovat základní vztahy mezi rozhodujícími ukazateli výrobní, technické a ekonomické činnosti OKD
- formalizovat a kvantifikovat tyto vztahy pomoci prostředků ekonometrického modelování na základě regresní analýzy dosavadního vývoje
- pro tyto účely vytvořit bázi dat časových řad vymezeného souboru ukazatelů
- na základě výsledků identifikace, analýzy a kvantifikace příslušných vztahů mezi ukazateli vytvořit a ověřit ekonometrický model pro analýzy a prognózy činnosti OKD, umožňující variantní výpočty možného vývoje výrobně-technických a ekonomických ukazatelů podle zvolených scénářů

Soustava simultánních rovnic ekonometrického modelu OKD-1

Výsledná struktura ekonometrického modelu OKD-1 je vyjádřením přijaté úrovně desagregace. Vzhledem k specifickým podmínkám činnosti OKD byla v modelu uplatněna desagregace na 5 skupin podniků a organizací:

- důlní podniky
- koksovny
- strojírenské podniky
- stavební podniky
- ostatní podniky a organizace

Se zřetelem na předpokládané využití modelu se ukázalo vhodné vyjádřit vztahy mezi následujícími skupinami ukazatelů (obr. 1):

- pracovníci
- základní prostředky
- důlně-geologické podmínky těžby uhlí
- produkce, výkony
- materiálové náklady a služby
- produktivita práce a průměrné výdělky
- zásoby, náklady, zisk a rentabilita
- investice

Z charakteru základních kauzálních a technicko-bilančních vztahů vyplývá, že pracovní síly a důlně-geologické podmínky mají exogenní charakter. Základní prostředky jsou ze značné části výsledkem vlastních investičních

dodávek. Pracovní síly, základní prostředky a důlně-geologické podmínky rozhodujícím způsobem ovlivňují produkci a výkony. Z vývoje výkonů se odvozuje vývoj materiálových nákladů, zásob, nákladů, zisku, rentability atd. Počty pracovníků mají vliv na produktivitu práce a společně s důlně-geologickými podmínkami též na průměrné výdělky, na nichž potom závisí mzdové náklady.

Ekonometrický model OKD-1 tvoří 114 vzájemně provázaných simultánních rovnic, které vyjadřují regresní a technicko-bilanční vztahy mezi 205 proměnnými. Z celkového počtu 114 rovnic je 20 regresních a 94 deterministických. Vyjadřují vztahy mezi 205 proměnnými, z toho je 91 exogenních a 114 endogenních.

Hypotetické kauzální vztahy mezi ukazateli jsou vyjádřeny regresními rovnicemi, zpravidla o větším počtu faktorů zaručujícím větší vypovídací schopnosti ekonometrického modelu i vyšší stabilitu.

Kromě regresních rovnic obsahuje model OKD-1 deterministické rovnice vyjadřující vztahy normativního charakteru, definiční vztahy (např. produktivita práce), různé technicko-bilanční vztahy apod.

Produkční funkce ekonometrického modelu OKD-1

Základna ekonometrického modelu OKD-1 je formulována jako soustava vzájemně provázaných produkčních funkcí. Výchozím tvarem je klasická Cobbova-Douglasova substituční produkční funkce. Pro praktické účely bylo třeba zohlednit působení dalších faktorů a charakteristik výrobních procesů OKD.

Vazby mezi produkčními funkcemi vyjadřují předávky výsledné produkce mezi skupinami podniků:

- předávku uhlí ke koksování
- výrobu důlních strojů a zařízení
- důlně stavební práce

Hlavní produkční funkcí ekonometrického modelu OKD-1 je produkční funkce pro těžbu uhlí. Její konstrukce vychází jednak z obecných teoretických a empirických poznatků o modelování výrobního procesu, jednak se snaží zobrazit působení specifických faktorů hlubinného dobývání uhlí v OKD. Výsledným tvarem je *komplexní* vícefaktorová produkční funkce vyjadřující závislost průměrné denní těžby celkem na 12 výrobních faktorech včetně různých ukazatelů důlně-geologických podmínek. Finální

tvary produkčních funkcí ekonometrického modelu OKD-1 obsahují faktory, pro které ještě nelze nalézt jednoznačnou podporu v rámci obecně uznávaných principů teorie produkčních funkcí. Jedná se zejména o některé novější teoretické poznatky, podle nichž se specifické faktory zahrnují do jediné produkční funkce současně.

Dynamické vlastnosti ekonometrického modelu OKD-1

Pro model OKD-1 byl zvolen rekurzívní tvar. Jedná se o rekurzívní model tzv. ekonometricko-normativního typu, respektující nelineární a dynamický charakter značné části modelovaných vztahů. Zpětné vazby lze zachytit pomocí interdependentní soustavy rovnic, v praxi ekonometrického modelování se však dává přednost zachycení těchto vazeb pomocí časově zpožděných proměnných v rekurzívních soustavách. Časově zpožděné proměnné dodávají soustavě rovnic ekonometrického modelu dynamický charakter.

Mezi ekonomickými veličinami (např. při investičním procesu) se vyskytují časové posuny mezi působením určitých faktorů a vyvolanými změnami ovlivňovaných veličin. S tím souvisí určitá setrvačnost ve vývoji ovlivňovaných veličin, které se postupně přizpůsobují změnám příslušných faktorů. Adekvátní vyjádření si proto vyžaduje použití dynamizovaných ekonometrických modelů a tedy přechod k jinému modelovému zobrazení zachycujícímu dynamiku sledovaných jevů. Uspokojivým řešením v tomto smyslu může být např. Koyckova transformace.

Datová báze a časové řady ekonometrického modelu OKD-1

Pro tvorbu ekonometrického modelu OKD-1 bylo nutno vytvořit datovou bázi. Odhad a testování parametrů regresních rovnic spolu s ověřením fungování ekonometrického modelu vyžaduje vytvoření rozsáhlého souboru dostatečně dlouhých a vnitřně i vzájemně konzistentních časových řad všech modelových proměnných.

V případě ekonometrického modelu OKD-1 byly časové řady sestaveny za 16 let, 1970-1985, pro časově posunuté proměnné také za rok 1969.

Bylo nutné provést množství přepočtů: agregace údajů za skupiny podniků, rekonstrukce některých ukazatelů za starší léta, přepočty údajů na srovnatelnou metodickou základnu (základní prostředky, výkony, materiálové náklady, finanční ukazatele atd.), přepočty na srovnatelné ceny

(hrubá výroba) a v některých případech i aproximace chybějících údajů (výroba zboží, některé složky investic, likvidace základních prostředků).

Výsledky ověření ekonometrického modelu

Předpokladem úspěšné aplikace modelu pro zpracování analýz a prognóz je, aby model dobře *reprodukoval* skutečný vývoj v retrospektivě. Celkové fungování ekonometrického modelu OKD-1 bylo ověřeno dynamickou simulací skutečného vývoje v období 1971-1985. Podle výsledků této simulace průměrná absolutní odchylka modelem generovaných hodnot od skutečného vývoje 114 výstupních proměnných za 15 let dosahuje 0,82%. Agregátní koeficient korelace R mezi modelem generovanými a skutečnými hodnotami endogenních proměnných za období 1971-1985 dosahuje 0,9962. Podle koeficientu determinace R^2 model vysvětluje 99,24% rozptylu hodnot souboru endogenních proměnných. Tyto koeficienty jsou přitom příznivé rovněž za všechny proměnné i léta. Vyhovující jsou rovněž hodnoty průměrných absolutních odchylek proměnných (vyjádřených v procentech k průměrným skutečným hodnotám). Průměrná kompenzovaná odchylka vyjadřující. systematické vychýlení modelem generovaných hodnot od skutečných dosahuje 0,18%.

Kromě retrospektivního ověření celkového fungování. modelu byl ekonometrický model OKD-1 ověřen též aplikací na výpočet experimentální prognózy na léta *1986-1990*. Za tím účelem bylo třeba sestavit vstupní hodnoty všech exogenních proměnných na prognózované období, zejména pro některé ukazatele související s očekávaným vývojem důlně-geologických podmínek, dále byly navrženy hodnoty exogenních složek investic, koeficienty likvidace základních prostředků, podíly stavu základních prostředků skupin podniků na průměrném stavu základních prostředků nedůlních organizací atd. Po dosazení těchto vstupních dat se výsledky pro endogenní proměnné prognostické aplikace ekonometrického modelu OKD-1 v podstatě shodovaly s hodnotami základních agregátních ukazatelů uvedených v *Návrhu koncepce rozvoje OKD,* zpracovaným obvyklými metodami v příslušných odborných útvarech generálního ředitelství. Významné odchylky této varianty modelové prognózy měly v podstatě dvojí vysvětlení: buď naznačovaly, že do vstupních dat pro prognózu nebyly promítnuty všechny předpoklady o měnících se podmínkách dalšího rozvoje, nebo signalizovaly, že v údajích návrhu existují některé rezervy a nekonzistence. Experimentální modelová prognóza ukázala, že ekonometrický model OKD-1 je vhodným nástrojem pro výpočty dalších variant prognóz s odlišnými vstupními předpoklady. Simulační výpočty

podle aktuálních dat v *90.létech* byly stále pozoruhodné a reflektovaly mj. *silnou setrvačnost* působení hlavních faktorů výrobně-technické a ekonomické činnosti OKD.

LITERATURA

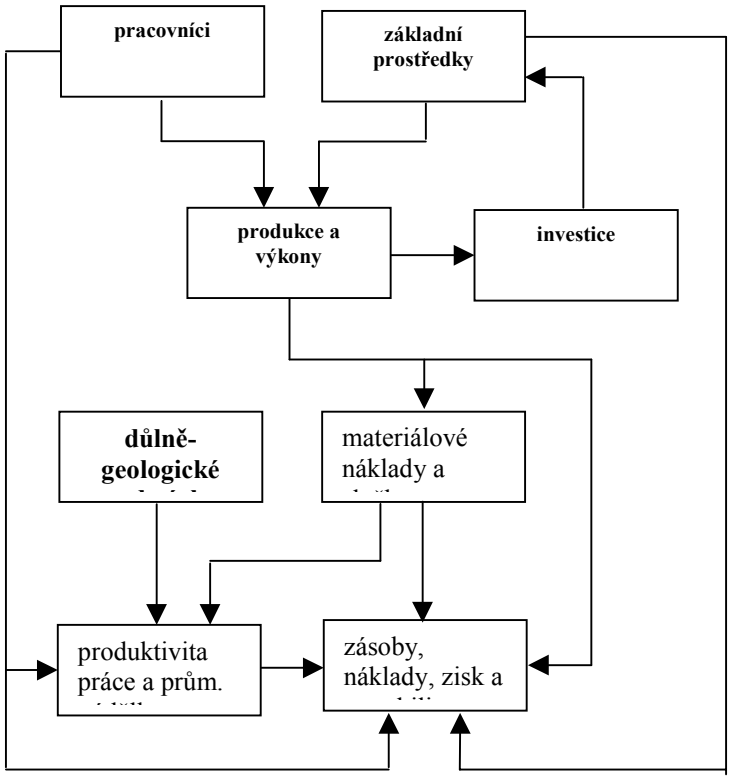
[1] LIŠKA, M. *Ekonometrický model činnosti OKD (kandidátská disertační práce)*. Praha: Ekonomický ústav ČSAV, 1987.

[2] ŠUJAN, I. *Vývoj a aplikácie ekonometrických modelov v ČSSR*. Informačné systémy 12, 1986. č.4.

[3] HUŠEK, R. *Základy ekonometrické analýzy I, II.* Praha: Vysoká škola ekonomická, 1997.

RNDr.Miroslav Liška,CSc. Katedra informatiky, Přírodovědecká fakulta, Ostravská universita, e-mail: liska@osu.cz

Schéma základních vztahů ekonometrického modelu OKD-1



Teorie a praxe dolování znalostí z dat

Jana Šarmanová

Abstract: The great expansion of data mining in the last decade may lead to expectations that are more optimistic than the practice usually shows afterwards. The methods used for mining in marketing are often satisfied with simple or less precise conclusions since the speed of processing is usually the most important criterion. For classical research, more exact methods and complex analyses are needed, which is not easy to carry out in a fully automatic way. In this paper, we use two examples from practice to suggest some problems of practical data mining. The experience was obtained during analysing data on assisted reproduction, which was done as a part of grant IGA MZ no. 4916-3.

Úvod

V posledních letech se mnoho napsalo o multioborové disciplině, nazývané dolováním znalostí z dat (Data Mining). Vznikla spojením a rozšířením metod a technologií převzatých z databází, matematické statisticky, metod analýzy mnohorozměrných dat, ekonomie a mnoha dalších „uživatelských" oborů. V heslech si zopakujme její základní pojmy.

Definice dolování znalostí: Proces netriviálního získávání implicitní, dříve neznámé a potencionálně užitečné informace.

Odkud se doluje – jak je možno ke znalostem přijít

neautomatizovaně

- pozorováním, pamatováním skutečností světa, zobecňováním, odvozováním
- od experta - člověka s nadprůměrným vzděláním, zkušeností, schopností zobecňovat a formulovat znalosti ve svém oboru
- z textové, obrazové a zvukové informace již expertem zpracované automaticky
- z dat speciálně pro tento účel nasbíraných nebo z databází původně sloužících jiným účelům
- z textových a multimediálních dokumentů, které se obvykle různými technikami převádí na předcházející případ

V dalším se budeme zabývat jen automatickým získáváním znalostí z dat**.**

Kdo doluje - potřeby dolování z hlediska uživatelů

- *průzkum - marketing, bankovnictví, výroba, pojišťovnictví, ...*
- výzkum – medicína, biologie, hutnictví, …
- sociologický průzkum – veřejné mínění, sčítání lidu, lokální věcné problémy, …

Proč doluje – jaké důvody vedou k dolování znalostí

- komerční - získání obchodních výhod
- výzkumné - získání nových odborných znalostí, hypotéz
- průzkumné, sociologické - získání politických výhod

Jak doluje - přehled klasických metod automatického dolování

- analýzy asociací ASOC
- analýzy příčin a následků IMPL
- analýzy shlukovací SHLUK
- hledání rozhodovacích stromů STROM
- agregace, časové a dimenzionální řady
- …

Čím doluje – jaké SW prostředky jsou k dispozici

- použitím řady hotových SW produktů, specializovaných na některé metody
- použitím speciálního hotového SW systému určeného pro dolování z dat
- spoluprací s programátory, kteří potřebný SW implementují na míru problému
- sám sobě být analytikem datovým i problémovým, programátorem i interpretem

Teoretický postup při dolování znalostí

Obdobně, jako u každé složité činnosti, jsou pro projekty dolování znalostí definována mnohá teoretická pravidla: etapy projektu, řešitelské týmy, metody, algoritmy, SW nástroje, způsoby prezentace a interpretace výsledků, doporučení dalších postupů. Prakticky všechny se shodují v následujících etapách a následujících typech spoluřešitelů:

1. Formulace problému
2. Věcná analýza úlohy, dělí na datovou analýzu a problémovou**.**
3. Sběr nebo výběr údajů
4. Hrubá filtrace
5. Předzpracování dat

6. Vlastní analýzy
7. Prezentace výsledků
8. Interpretace a vyhodnocení výsledků.

V současném softwarovém světě je nabízena řada produktů, jejichž slogany zní velmi slibně jak z hlediska automatických postupů, tak uživatelského prostředí. Žádný z nich neslibuje dřinu, nutnost mnoho studovat, dlouho analyzovat, znovu a znovu analyzovat, probírat se množstvím banalit často krásně barevně provedených, 95% svých pracně získaných výsledků zahodit a jen při nemalém štěstí se dobrat malého množství skutečně nových znalostí. A to vše za ceny produktů desetitisícové, statisícové i milionové.

Protože se léta zabývám teoretickými metodami dolování dat i praktickým získáváním znalostí v rámci různých druhů výzkumů a sociologických průzkumů, pokusím se zformulovat některé praktické zkušenosti do pragmatických rad začínajícím analytikům. Snad to bude ku prospěchu i uživatelům, jejich datům i z nich získaným výsledkům.

Profese matematika-analytika a programátora, dlouholetá neexistence vhodného SW nebo jeho vysoká cena, praktická potřeba mnoha uživatelů ve výzkumu, to vše bylo již před mnoha léty důvodem k rozhodnutí vybudovat vlastní programový systém pro analýzy dat. Hluboko v době, kdy se ještě data nedolovala, ale analyzovala. Systém byl několikrát implementován v novém prostředí a vždy rozšiřován o nové metody, pokrývající celý uvedený cyklus projektu.

Cílem bylo nejen nabídnout případnému uživateli metody předzpracování, dolování a prezentací výsledků, ale i automatizovaného experta-analytika. Ten za pomoci metadat (popisu dat, dat o datech) a vlastních zabudovaných analytických pravidel bude průvodcem uživateli, bude mu doporučovat metody na míru jeho datům a pomůže mu výsledky interpretovat. V krajním případě bude umět provést všechny analýzy automaticky sám. Praxe se ukázala opět složitější, než několik školních testovacích příkladů.

Praktický postup nad konkrétními daty

Následující 2 příklady jsou voleny jako protiklady v mnohém směru: oblastí výzkumu, typem dat, zdánlivou složitostí problému, dobou zpracování, zkušenostmi analytika, rozsahem dat, úspěšností výsledků. První uvádíme jako příklad rychlého a úspěšného dolování z dat obtížných pouze analytikem. Druhý jako dlouhé, pracné, zatím málo objevné dolování z dat rozsáhlých a slibných. Oba případy byly řešeny pomocí vlastního SW.

Příklad 1. Technologický problém

Při válcování trub za studena na poutnických válcovacích stolicích je možno nastavovat různé technologické parametry, které vedou ke kvalitativně rozdílnému průběhu zpracování i výsledkům válcování. Cílem bylo vysledovat vliv ovlivnitelných parametrů na optimální průběh válcování. Praktické testování všech možných kombinací a sledování výsledků je příliš pracné, zdlouhavé a drahé. Není možno je proto podle potřeby opakovat.

Charakteristiky průběhu válcování trub za studena. Vlastnosti modelu byly definovány s odborníky na válcování trub. Z provedených měření byly zadány pro různé kombinace níže uvedených parametrů a jim odpovídající výsledky tyto nastavitelné technologické parametry:

- n [ot//min] počet otáček klikového hřídele
- p [mm] posuv na otáčku
- o [m] počet odválcovaných metrů (=>opotřebení nástrojů)
- m typ maziva
- k typ kalibrace (typ trnu)

Výstupem experimentů je axiální síla ve tvaru grafického záznamu, na papírových rolích zaznamenaný průběh několika válcovacích cyklů pro každý vzorek.

Analýza. Teoretický průběh axiální síly je znám. Ideální stav válcování má všechny cykly shodné, při špatně nastavených podmínkách tahová složka axiální síly postupně narůstá v průběhu několika cyklů, potom se trn "utrhne" a proces začíná znovu. Na teoretických křivkách je vytipováno několik charakteristických bodů, které společně určují typ průběhu axiální síly, od optimální 1 po nejhorší 6, způsobující až přetržení trubky.

Vlastní zpracování. Digitalizovaná, filtrovaná a standardizovaná data tvoří tedy n-tice, každá n-tice popisuje jeden cyklus jednoho měření. Tyto n-tice byly pomocí shlukové analýzy rozděleny do skupin vzájemně si podobných cyklů. Porovnáním s teoretickým průběhem byly výsledkům přiřazeny typy teoretických křivek, každému měření pak nehorší typ křivky, kterého dosáhl některý jeho cyklus. Posledním krokem bylo nalezení souvislostí mezi množinami nastavených parametrů a výsledným typem křivky. K hledání hypotéz o příčinách různého průběhu procesu válcování byla použita procedura IMPL s použitím kvantifikátorů implikačních.

Slovně interpretované výsledné hypotézy*:* Pro dosažení nejlepší křivky je třeba následující nastavení:

- Počet otáček může být nastaven libovolně, důležitý je vzájemný vztah otáček a posuvu.
- Vhodnější jsou posuvy vyšších hodnot (od 4.5 mm), nejlepší je 6 mm.
- Nejvhodnější kombinace otáček a posuvu jsou: 80 / 7.5, 120 / 6, 160 / 4.5, 180 / 6
- Vliv nastavení trnu se výrazně neprojevil.
- Vliv mazání nebyl průkazný
- Významný je naopak vliv kalibrace: při stejném mazání a nastavení trnu byla nejlepší křivka dosažena pouze na kalibracích typů 1-3, kalibrace 4 a 5 souvisely vždy s horšími typy křivek.

Křivka typu 2 vzniká většinou při mírném nedodržení výše doporučených kombinací. Naopak nehorších křivek dosáhly všechny ty kombinace, které se nedoporučují pro dosažení křivek dobrých – to vše naznačuje, že nově nalezené vztahy skutečně existují. Závěrem poznamenejme, že některé důležité výsledky dosažené popsaným zpracováním nebyly známy ani dlouholetým praktikům ve válcování, ani odborníkům teoreticky se zabývajících tímto problémem. Následné praktické ověřovací zkoušky daly těmto výsledkům zapravdu.

Příklad 2. Medicínský problém - Asistovaná reprodukce

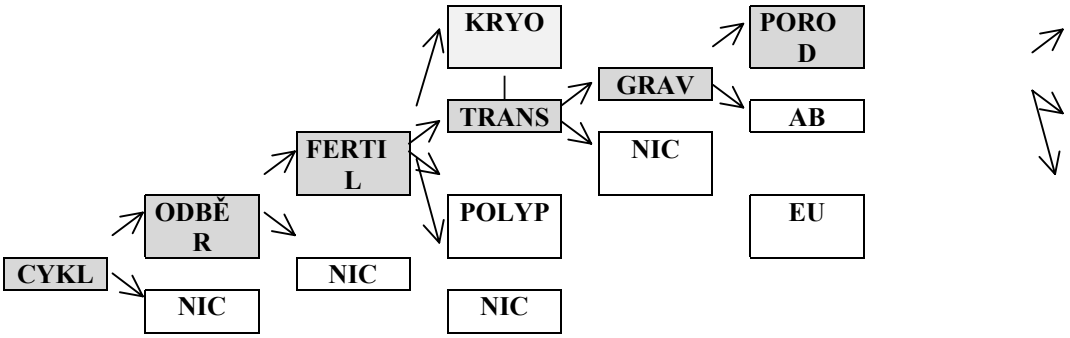
Ve Fakultní porodnici v Brně se již 15 let provádí asistovaná reprodukce (AR), po celou dobu se o průběhu a výsledcích léčby vedou záznamy v databázi, která má 3 základní evidence (mimo řadu číselníků). Celkem jsou uloženy údaje o více než 8500 výkonech u 3500 pacientek. Je sledováno celkem asi 180 atributů. V rámci grantu MZ se tato data v průběhu letošního roku analyzují.

Data sbíraná od roku 1984

Karta	[4470 x 15]	*objekty:* pacienti, výkony, výsledky	
Anamnézy	[3161 x 65]	*atributy:*	numerické, textové, datumové
Výkony	[8516 x 58]		neovlivnitelné, ovlivnitelné
Gravidity	[666 x 54]		průběžné, výsledné

Cílem je najít souvislosti mezi neovlivnitelnými i ovlivnitelnými atributy na straně jedné a výsledkem léčby na straně druhé a na základě této znalostiurčovat ovlivnitelné atributy léčby.

Analýza. Definovaný cíl nabízí použití analýzy příčin a následků, analýzy asociací, případně konstrukci rozhodovacího stromu. Při rozdělování atributů na ovlivnitelné, neovlivnitelné, výsledné se ukazuje, že celý cyklus AR se dělí na dílčí výkony, každý z nich má své vstupy a výstupy a celkově se dá znázornit schématem na obr. 1. Odtud pak plyne řada dílčích analýz, z nichž každá závisí na předcházejícím průběhu tohoto cyklu i případně cyklů (u téže pacientky) předcházejících a na nově vstupujících atributech. Ideálem je vysledovat, jak mohou ovlivnitelné parametry (podávané léky, volba dne výkonu, metoda výkonu apod.) v souvislosti s objektivními skutečnostmi (věk pacientky, diagnóza apod.) ovlivnit úspěšnost výsledku (odebrání dostatečného počtu vajíček, jejich oplodnění, otěhotnění pacientky, donošení plodu). Na základě výsledných pravidel pak doporučovat hodnoty ovlivnitelných parametrů, aby bylo dosahováno optimálních možných výsledků.



Obr. 1

Vlastní zpracování. Prvním a pak celým zpracováním se prolínajícím problémem jsou zdrojová data. Jejich strukturu původně navrženou jen pro evidenci a občasné vyhledávání údajů bylo nutno upravit – odstranit redundance, rozdělit kumulované údaje, překódovat některé číselníky, kategorizovat některé hodnoty apod. Nejen struktura, i obsah dat byl mnohdy chybný, redundandní údaje nekonzistentní, i při analýzách se objevovaly stále nové nesrovnalosti, negativně ovlivňující výsledky a tak se znovu dohledávala a opravovala data.

Další velký problém nastal při vlastní analýze. Podle typu problému se jeví základní vhodnou metodou analýza příčin a následků – IMPL. Ovšem

v datech, kde průměrná úspěšnost výsledků výkonů (po etapách) je 17% až 50%, vychází klasickou metodou IMPL (při nastavené minimální spolehlivosti hypotézy např. 70%) jen mnohé výsledky typu Jestliže podm pak pacientka neotěhotní s vahou V % a prakticky žádný využitelný přímý pozitivní výsledek. Jednou možností jeinterpretovat negativní výsledek jako doplněk hledaného pozitivního (když za podm pacientka neotěhotní s vahou V%, pak za stejné podmínky otěhotní s vahou 100-V) a dále zkoumat, zda tento výsledek je statisticky významný.

Podstatou tohoto problému je fakt, že v těchto datech nestačí hledat jen hypotézy tvaru implikace s vysokou spolehlivostí (a dostatečnou podporou, počtem výskytů tohoto případu), ale všechny statisticky významné odchylky od průměrných hodnot. Pro tuto úlohu bylo potřeba vyvinout novou metodu hledání hypotéz (kterou jsme nazvali GRIMPL = grupované implikace) a její dostatečně rychlou implementaci.

Metoda GRIMPL testuje všechny teoretické kombinace hodnot možných příčin a pro každou skupinu vypočítá hodnoty předem (analytikem) definovaných ukazatelů úspěšnosti. Proti metodě IMPL navíc vypočte statistickou významnost odchylek těchto hodnot od průměrných hodnot ukazatelů za celý soubor. Metoda tedy testuje výrazy tvaru

Jestliže A=a $\wedge$ B=b $\wedge$… , pak U1=u1 (S1) $\wedge$U2=u2 (S2) $\wedge$… s podporou P

kde A,B,C,… jsou atributy = možné příčiny, a,b,c,… jejich aktuální hodnoty, U1,U2,… jsou definované ukazatele, u1,u2,… jejich hodnoty pro testovanou skupinu charakterizovanou levou stranou implikace. Hodnoty S1, S2, … buď numericky nebo symbolicky označují, zda je ukazatel Ui průměrný nebo neprůměrný vzhledem k základnímu souboru.

Praxí vydolovaná pravidla pro dolování dat

Data i plán zpracování u druhého příkladu se zdají velmi transparentní. Přesto se postupně objevuje tolik nových i věčně zelených problémů, že se je pokusíme zformulovat do bezpochyby neúplného seznamu následujících pravidel a zkušeností. Skutečné výsledky získané z těchto analýz přesahují rámec tohoto příspěvku a patří do jiné – medicínské oblasti.

- Neslibuj jakékoliv výsledky předem, až data ukáží, co je v nich.
- Bez (živého) experta to (pořád ještě) nejde.
- I v každé analýze (nejen v programu) je chyba.
- Cykly, cykly, cykly = chyby, nápady, upřesnění, výběry.
- I když jsi hotov, nikdy nejsi hotov.
- Nikdy nejsi hotový analytik.

KVALITNÍ PŘÍPRAVA DAT JE 90% ÚSPĚCHU

Bez dobrých metadat to nejde

- pořádek pomáhá zvládnout data
- předpočítané hodnoty urychlují analýzy (NULL, min, max, …)
- dokumentace, protokol, evidence i detailů, chyb, návratů, důvodů

Problém integrity a konzistence

- párování souborů
- kontrola atributů: NULL-nula; chybějící dohledat, dořešit; opakované prověřit; špatně kódované opravit;

Problém kategorizace

- nutnost opakované kategorizace, nutnost verzí atributů

Možnosti odvozených atributů

- využití položek datumových, textových; doplnit chybějící odvoditelné; respektovat profesní zvyklosti, …

Nutnost transformací

- standardizace, normalizace, hlavní komponenty**, …**

Výběry – projekce a selekce

- jinak pro SHLUK, IMPL a ASSOC, STROM, …

VLASTNÍ ANALÝZY

Princip: rozděl a pokus se panovat

Problém exploze výsledků

- nutnost redukce redundance
- automatizovat i zpracování výsledků
- nutnost podpory databází výsledků

I negativní výsledek může být výsledek - nejhorší NE může být nejlepší ANO

Problém malých četností - co je dost významné?

Bez statistiky to často nejde - významnost výsledných hypotéz je nutno otestovat

Se statistikou to často nejde - statistické charakteristiky někdy necharakterizují

Problém rychlých algoritmů - všechny cesty vedoucí k cíli nevedou stejně rychle

Problém rozsáhlých dat - vzorkování, reprezentanti, nové přístupy k algoritmům

Prezentace výsledků - ani geniální objevy nedojdou ocenění bez slovního polopatizmu, obrázků, grafů

Závěr

Nová definice dolování znalostí: Proces hledání zlatých zrnek na poušti,když nevíme, jestli tam jsou, a musí to být rychle.

Nic nového: **Nic není snadno a rychle. Každý dobrý nápad je vítán**

Literatura

Šarmanová,J.: Metody automatizovaného získávání znalostí z databází. In: Sb. Moderní matematické metody v inženýrství 1998, str. 190-194.

Šarmanová,J.: Proces dolování znalostí z databází. In: Sb. ASIS'98, Krnov, str.227-236.

Hudeček,R.-Ventruba,P.-Šarman,Z.-Šarmanová,J.-Solar,T.: Analýza faktorů ovlivňujících výsledky asistované reprodukce prostřednictvím moderních metod výpočetní techniky. In: Sb. 8. symposium SAR, Brno 1998.

Šarmanová, J.: Expertní řízení procesu generování hypotéz z dat. In: Sb. Medzinárodnej konferencie Systémová integrácia¨99. Stará Lesná 13.-15.10.1999, str. 80-86.

Pokračování ze strany 1.

Za organizátory Ostravských statistických dnů děkuji všem přednášejícím, obzvlášť těm, kteří dodali text k publikaci v Bulletinu a trpělivě se snažili splnit všechny editační požadavky či nápravu různých jiných nedostatků. Pro vzájemné uklidnění poznamenávám, že nikdo z nás autorů neuspěl s první a mnohý ani s druhou versí. České statistické společnosti děkuji za její příspěvek na organizování této akce, čtyřem pražským zástupcům výboru za to, že vážili cestu do vzdálené Ostravy a významným podílem přispěli k důstojnému a přátelskému průběhu formální i neformální časti Statistických dnů. Děkuji spoluorganizátorům Martině Litschmannové z VŠB-TUO, Ivanovi Křivému z PřF OU, a také Natálii Bosové a Simoně Polochové, které se pečlivě postarali o technické a administrativní zabezpečení akce. Organizátorům příštích Statistických dnů ČStS přeji hodně štěstí.

Josef Tvrdík, editor

HIGH BREAKDOWN POINT REGRESSION ANALYSIS OF THE CZECH ECONOMY

Jan Ámos Víšek, Prague[1]

Abstract

The least trimmed squares-estimator of regression coefficients is recalled and its advantages and disadvantages are discussed. Then it is employed for an analysis of the Czech economy, from the point of view of its export possibilities. Results revealed at least two things.

Firstly, that high breakdown point estimators may be efficiently used as diagnostic tools. It is in some sense a result, by an example confirming a general expectation that it should work.

Secondly, the results indicate that the industries may be divided into two groups, one containing industries behaving like in (stabilized) market economy and another which gathers industries which work like under centrally planned economy. The result is not too much surprising (after the analysis), as the Czech economy was (and still surely is) under a transition. What is much more valuable is that the example shows that a direct division of industries into two groups is not possible and the approach via an regression analysis is (probably) the only one having a hope to succeed.

AMS classification: 62F35, 62J20
Key words: Regression, high breakdown point, the least trimmed squares, economic application.

INTRODUCTION AND NOTATION

Sometimes we meet with a situation when the OLS-analysis fails. It gives e. g. for any combination of explanatory variables too small value of coefficient of determination and/or other characteristics as normal plot, Durbin-Watson etc. are not satisfactory. It may be just a consequence of our inappropriate idea about relations among variables in question. Sometimes however it may be caused by the fact that our data consist of two (or more) subpopulations and/or it may be a result of the presence

[1]Research was supported by Grant Agency of Academy of Sciences of the Czech Republic, number A 2075803.

of a contamination of data. In such a case we would like either rid of the contamination or to divide the data on the respective subpopulations. However a characteristic (or a factor, or how we shall call it) which indicates for each single observation whether it belongs into the first or the second subpopulation (if they are only two), need not be either available or may be just that, what we have not yet unfortunately recognized in data. Maybe, and our numerical example below demonstrates it, that no simple factor indicating how to divide the data exists at all. Of course, there can be a relation between some factors (e. g. between explanatory variables) which can a posteriori "confirm" that the division of data can be justified. But as we shall see from our example, even in the case when we knew that "confirmating" relation, we cannot divide data according to it, because this relation does not distinguish uniquely all observations.

In the above described situation, we can try to find (the largest "homogeneous") subpopulation by means of some techniques from the offer of modern robust statistics (or, econometrics, as you want). Of course, having found some subpopulation with a satisfactory model (from statistical point of view), we should subject the result to heuristic discussion whether the result is acceptable or senseless from a point of view of given branch of science, the data came from.

A large collection of methods is nowadays supplied by robust regression. The methods with high (and usually easy controllable) breakdown point can serve excellently for the purpose, as we shall try to demonstrate below. First of all, let us recall that the breakdown point is one of characteristics of the point estimators, indicating an upper limit of the level of contamination for which the estimator is still able to give reasonable results. From the demonstration we are going to present below it will be quite clear what is meant by "reasonable". So let us recall only the definition of breakdown point. In order to be able to do it, let us denote by π the Prokhorov metric.

DEFINITION 1 *The estimator T_n of the parameter $\theta \in \Theta$ has breakdown point*

$$\varepsilon^* = \sup_{0 \le \varepsilon \le 1 \subset} \{\varepsilon \ : \ \exists K(\varepsilon) \subset \Theta, K(\varepsilon) \ compact \ : \\ \pi(F,G) < \varepsilon \Rightarrow G(T_n \in K(\varepsilon)) \to 1 \ for \ n \to \infty\} .$$

As it follows from that what was already said, the advantage of robust methods in comparison with OLS is that they work even for contami-

nated data. Of course, models that are found by such a method, give different weights to observation, and in an extreme case they depress some observations completely. On the other hand, let us realize that "classic" weighted least squares do make sometimes precisely the same.

Another advantage, in this case in comparison with "classic" diagnostics, is that they need not lose the information that is carried by leverage points in the case when these leverage points do not represent contamination. Of course, due to the fact that some kind of a *law of the preservation of a mass* applies anytime, these advantages should be paid for by something. In the case of robust procedures, we luckily repay only partially, by a complexity of estimation, and of course by a loss of efficiency. And the objections of "classic" statisticians and econometricians speak about a considerable loss of efficiency.

Firstly, the loss of efficiency, in the case that there is no contamination, is much smaller than it is commonly assumed. Moreover, we may (and we strongly recommend it) anytime calculate the results by several methods and in the case that the results are not significantly different, which indicates no or low contamination, we can accept that one which we assume to be most efficient. Of course, the best way is to do it by all available methods on given computer.

Secondly, the loss of efficiency of classic methods, in the case when the contamination is present, is however deteriorating. The fact was known already to Sir Ronald Alyner Fisher but it was later discreetly forgotten (see Fisher (1922) for a shocking example of the inefficiency of the most commonly used classic methods of estimating location and scale; for a lot of other nice examples, see Hampel et al. (1986)).

Much more serious disadvantage of the methods with a high breakdown point is however their computational complexity. Since the corresponding estimators are defined implicitly by an extremal problem, no close formula for their evaluation is usually available (or even such formula cannot exist at all). Hence an invention of a special *trick* which allows to establish an effective algorithm for evaluation of an efficient approximation to precise value to the estimator has to be found. Moreover, even when the trick is "discovered", it still needs usually an elaborate implementation and another step of invention, namely an invention of a way how to verify that the evaluation has really given an efficient approximation (for larger discussion see Hettmansperger & Sheather (1992), Víšek (1994), (1996 a) and (2000 a)). On the other hand, this disadvantage (with evaluation of the estimator) is irrelevant for the potential

user whenever the method is implemented and available. But there are still other difficulty, namely that the application of such a method needs somewhat more time then the OLS and that the interpretation of results requires much more care and knowledge. We may compare it with the situation when an aspirin can treat a simple disease and may be used by a layman but the employment of a more complicated treatment requires usually presence of a physician. And this disadvantage cannot be quickly and easily removed but it should not lead to (hasty) rejection of such methods, similarly as more complicated medical treatment were not simply rejected due to their higher requirements on physicians. At the end of this discussion we should also admit that some robust estimator may have still one another disadvantage, namely high subsample sensitivity. For details see Víšek (1996 b) and (2000 b).

What concerns the time, which is spend for the performance of one task, it may be very different. One can imagine that in the case of two, relatively separated clouds of data, the evaluation can be very quick[2]. On the other hand, sometimes the structure of data (and especially, of contamination) may be very complicated and evaluation takes rather large time (for details see Víšek (1996 a)).

An intensive study of the robust method has begun at the middle of sixties but it lasted as long as up to the middle of eighties when the first methods for estimating regression coefficients, having breakdown point as high as 50%, have appeared. It was a consequence of the fact that the methods which were a straightforward generalization of the robust estimators of location parameter, namely M-estimators, suffered by disadvantage of dependence of the breakdown point on the dimension of the underlying model (see famous result of Maronna, Bustos and Yohai (1979)). At a first glance the roots of the problem are in bad results of M-estimators under the presence of leverage points which, however, can be removed (or at least substantially depressed) by employment of redescending ψ-functions. Much more difficult however is, that the M-estimators are not generally scale- and regression-equivariant without a studentization of residuals by a scale-invariant and regression-equivariant estimator of scale of random fluctuations, see Bickel (1975), Jurečková & Sen (1993) or Víšek (1999 a).

[2]Let us stress that even in such situation the clouds need not be visible (or recognizable) even by good graphical editor allowing to rotate data, just due to the fact that we are able to draw every time only three dimensions from multidimensional data.

So it seems that it is preferable to use an estimator with high (and moreover controllable) breakdown point, which is scale- and regression-equivariant. One such estimator is so called the *Least Trimmed Squares*. The definition is rather simple. Let us consider for any size $n \in N$ (N-set of all integers) a linear regression model

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j^0 + \varepsilon_i, \qquad i = 1, 2, ..., n$$

where Y_i is the value of response (dependent) variable for the i-th case, x_{ij} is the value of the j-th explanatory (independent) variable (again for the i-th case), β_j^0 is the j-th coordinate of (true) vector of regression coefficients β^0 and finally e_i's are independent identically distributed random variables. Further, for any $\beta \in R^p$ (R^p - p-dimensional Euclidian space) denote the i-th residual by $r_i(\beta) = Y_i - \sum_{j=1}^{p} x_{ij}\beta_j$ and by $r_{(i)}^2(\beta)$ the i-th order statistic among the squared residuals, i.e. we then have

$$r_{(1)}^2(\beta) \leq r_{(2)}^2(\beta) \leq ... \leq r_{(n)}^2(\beta).$$

Finally, let us define for an integer h, $\frac{n}{2} \leq h \leq n$ the *Least Trimmed Squares* estimator as

$$\hat{\beta}^{(LTS,n,h)} = \underset{\beta \in R^p}{\arg\min} \sum_{i=1}^{h} r_{(i)}^2(\beta) \tag{1}$$

in difference with OLS which are given by

$$\hat{\beta}^{(LS,n)} = \underset{\beta \in R^p}{\arg\min} \sum_{t=1}^{n} r_i^2(\beta).$$

In both cases *arg min* means argument which minimizes the expression which stays behind this sign. So that in our former case we are looking for such $\beta \in R^p$ for which the sum of the h smallest squared residuals (among all residuals evaluated for given β) is minimal among all possible sums, i.e. among sums evaluated for all $\beta \in R^p$. The heuristics, which are behind the definition, are straightforward. The estimator takes in fact into account only a subset of data containing h cases, for which the sum of squared residuals is the smallest one among all subsets of size h. In turn, it means that the presence of contamination or another subpopulation up to the size $n-h$ is irrelevant for the result of estimation.

Let us stress that there is another estimator, namely $\hat{\beta}^{(TLS,n)}$, which is sometimes falsely assumed to be the same as $\hat{\beta}^{(LTS,n)}$. But as the orders of letters of superindeces indicate, $\hat{\beta}^{(TLS,n)}$ is evaluated on a subpopulation which is obtained by trimming off $n-h$ cases according to an external rule (see e.g. Ruppert & Carroll (1980)) while $\hat{\beta}^{(LTS,n)}$ defines the subpopulation itself implicitly. It means that the former is usually much easier to evaluate than the latter. On the other hand, the former lacks some good (and plausible) properties of the latter, namely an ability to cope with leverage points without any loss of useful information, i.e. without any loss of efficiency with respect to the true underlying model. It is of course rather complicated to prove that $\hat{\beta}^{(LTS,n)}$ is consistent, asymptotically normal and it has breakdown point approximately equal to h/n. But it is not important for applications. What matters is the fact that there is an algorithm which appeared to be effective for its evaluation in the sense that it gave good approximation in the case when we were able to find the precise solution of the extremal problem (1), see again Čížek & Víšek (2000) and Víšek (1994), (1996 a) and (2000 a). We shall leave aside all other details about $\hat{\beta}^{(LTS,n)}$ (some of them may be found either in Hampel et al. (1986)) or in Rousseeuw & Leroy (1987) or in Víšek (1996 a) and (2000 a); in the last two references the algorithm for the evaluation and its testing is described and results on examples are also discussed).

RESULTS

Now, let us turn to the models, which can be established by means of $\hat{\beta}^{(LTS,n)}$ for data describing the Czech economy in 1994. First of all, let us give an explanation of the abbreviations.

X/S	-	export per sales
US/VA	-	number of university students per value added
HS/VA	-	number of high school students per value added
K/VA	-	capital per value added
CR_3	-	market power (concentration in given industry)
$TFPW$	-	total factor productivity related to wages
BAL	-	Balsa index
DP	-	price development after opening-up

These abbreviations hint that for each of 91 industries of the Czech economy we had at hand data about the export, sail, the human capital, the capital equipment, Ballasa index (which is one of characteristics recording relation between the export and the import into the industry in question; in the case of Balasa index it is difference between the export and import divided by the sum of export and import), the total factor productivity, the value added, the wages, the research and development, the foreigner direct investments, the debts etc. In total we had in our disposal about 25 explanatory variables.

After a rather long and woeful search[3] for a model, we have arrived at

$$\frac{X_i}{S_i} = \beta_0 + \beta_1 \cdot \frac{US_i}{VA_i} + \beta_2 \cdot \frac{HS_i}{VA_i} + \beta_3 \cdot \frac{K_i}{VA_i} + \beta_4 \cdot CR_{3i}$$

$$+ \beta_5 \cdot TFPW_i + \beta_6 \cdot BAL_i + \beta_7 \cdot DP_i + \varepsilon_i \qquad (2)$$

which appeared to hold approximately for a half of our data. In the Table 1 - 6 we give results for $h = 52, 53, ..., 57$. These values were selected due to the fact that the values of the estimates of coefficients were (relatively) stable, coefficients of determinations (which are collected in Table 7) of corresponding models were sufficiently high and indication of normality of residuals in corresponding subpopulations of data was satisfactory. Further, let us stress that the corresponding subpopulations (which will be called in what follows *main* subpopulations) were nested. Finally, the fluctuations of the estimates of regression coefficients for h below 48 and over 60 was (very) large and the corresponding subpopulations were not nested and the other characteristics, as coefficient of determination or normal graph were poor.

Moreover, it appeared that the *complementary* subpopulation (after deletion of a few points) allows also satisfactory regression models, of course with different coefficients (see Table 9 below). Finally, it occurred that this division of data has the following sense. The *main* subpopulation has a model for relationship between capital and labour which can

[3] We had to experiment not only with the many combinations of explanatory variables but for any selected p-tuple of explanatory variables we had to evaluate the estimate of regression coefficients for several values of h and also check other characteristics which are monitored during regression analysis. It was tiresome, monotone work.

be called *market-economy-like* (due to respective production function), while the *complementary* subpopulation (even without deletion of any industry) has a model for relation between labour and capital may be interpreted as valid in *central planned economy* (more details are given below; to save the space we shall write the heads of the second and the third columns of the next tables as *Estimates* and *Standards* instead of *Estimates of regression coefficients* and *Standard error of estimates*, respectively).

Table 1
Estimates of regression coefficients of model (2) for the main subpopulations of size 52

Item	Estimates	Standard	t-value	P-value
intercept	0.8663	0.5615	1.5428	0.130045
VS/PH	0.1175	0.0208	5.6592	0.000001
SZ/PH	0.1852	0.0373	4.9678	0.000011
K/PH	-0.241	0.0189	-12.7374	0
CR3	1.3439	0.3113	4.3176	0.000088
TFPW	-0.9103	0.2209	- 4.1201	0.000164
BAL	0.3357	0.2008	1.6718	0.101659
DP	1.1237	0.1572	7.1475	0

Table 2
Estimates of regression coefficients of model (2) for the main subpopulations of size 53

Item	Estimates	Standard	t-value	P-value
intercept	0.8376	0.575	1.4567	0.152135
VS/PH	0.119	0.0213	0.5601	0.000001
SZ/PH	0.1876	0.0382	4.9169	0.000012
K/PH	-0.2417	0.0194	-12.4709	0
CR3	1.4048	0.317	4.4319	0.000059
TFPW	-0.8967	0.2262	-3.9643	0.000261
BAL	0.3725	0.2046	1.8204	0.075359
DP	1.097	0.1603	6.8419	0

Table 3
Estimates of regression coefficients of model (2) for the main subpopulations of size 54

Item	Estimates	Standard	t-value	P-value
intercept	0.7284	0.5874	1.24	0.221252
VS/PH	0.1151	0.0217	5.2987	0.000003
SZ/PH	0.189	0.0392	4.8238	0.000016
K/PH	-0.2449	0.0198	-12.3556	0
CR3	1.4864	0.3224	4.6106	0.000032
TFPW	-0.9796	0.2278	-4.3007	0.000088
BAL	0.2716	0.2027	1.3399	0.186866
DP	1.1807	0.1581	7.4671	0

Table 4
Estimates of regression coefficients of model (2) for the main subpopulations of size 55

Item	Estimates	Standard	t-value	P-value
intercept	0.5801	0.6025	0.9629	0.340537
VS/PH	0.1135	0.0224	5.0594	0.000007
SZ/PH	0.2011	0.04	5.0229	0.000008
K/PH	-0.2476	0.0204	-12.1078	0
CR3	1.2749	0.3157	4.0382	0.000197
TFPW	-0.9374	0.2345	-3.998	0.000224
BAL	0.1068	0.1923	0.5553	0.581324
DP	1.244	0.1603	7.7621	0

Table 5
Estimates of regression coefficients of model (2) for the main subpopulations of size 56

Item	Estimates	Standard	t-value	P-value
intercept	0.7421	0.6217	11.937	0.238447
VS/PH	0.1065	0.0231	4.6146	0.00003
SZ/PH	0.1924	0.0414	4.6467	0.000027
K/PH	-0.2452	0.0212	-11.5519	0

Item	Estimates	Standard	t-value	P-value
CR3	1.3457	0.3265	4.1215	0.000148
TFPW	-10.264	0.2401	-4.2747	0.00009
BAL	0.0879	0.1998	0.4401	0.661844
DP	1.2492	0.1666	7.4982	0

Table 6
Estimates of regression coefficients of model (2) for the main subpopulations of size 57

Item	Estimates	Standard	t-value	P-value
intercept	0.9832	0.6306	1.559	0.12544
VS/PH	0.124	0.0238	5.1997	0.00000
SZ/PH	0.2039	0.0428	4.7662	0.00001
K/PH	-0.2512	0.0219	-11.4816	
CR3	1.1532	0.3367	3.4247	0.00125
TFPW	-10.187	0.2496	-4.0812	0.00016
BAL	0.2008	0.2036	0.9863	0.328851
DP	1.1017	0.1654	6.6623	0

In the next table we have collected estimates of regression coefficients from the previous tables (Table 1 - 6) to enable easier judgment about stability of these estimates.

Table 7
Estimates of coefficients collected from Tables 1 - 6

Number of cases	52	53	54	55	56	57
VS/PH	0.118	0.119	0.115	0.114	0.107	0.124
SZ/PH	0.185	0.188	0.189	0.201	0.192	0.204
K/PH	-0.24	-0.24	-0.25	-0.25	-0.25	-0.25
CR3	1.344	1.405	1.486	1.275	1.346	1.153
TFPW	-0.91	-0.900	-0.98	-0.94	-10.3	-10.2
BAL	0.336	0.373	0.272	0.107	0.088	0.201
DP	1.124	1.097	1.181	1.124	1.249	1.102

Anybody who sometimes performed regression analysis and was interested in changes of model caused by deletion of influential point(s), knows that we meet with such behavior (as demonstrated by Table 7) only in the case when there are no influential points, i.e. when data are (on rather high level) homogeneous. This fact and the arguments we gave in previous, supports our belief that we have divided data into two reasonable parts (other arguments will be given below). Before we shall continue let us give summarized results for all subpopulations in question.

Table 8
Sum of squares, estimates of scale, coefficients of determination and Durbin-Watson statistics of models, estimates of which are given in Table 1 - 6

Number of cases	52	53	54	55	56	57
Sum of squares	12.37	13.28	14.31	15.62	17.24	18.88
Estimates of scale	0.281	0.295	0.311	0.332	0.359	0.385
Coefficients of determination	0.860	0.851	0.854	0.845	0.832	0.824
Durbin-Watson	1.839	1.747	1.659	1.605	1.637	1.607

In the next two tables the results of robust regression analysis (as estimates of regression coefficients, sum of squares, estimate of scale, coefficient of determination and Durbin-Watson statistic are given for *complementary* subpopulation which consisted of 37 cases. (In what follows we shall use this word *complementary* every time for the population which is *complementary* to the *main* one.) From these 37 cases 4 were deleted since they "represented" contamination (in the sense explained above), so that h was equal to 33. These four industries were: production of textile and ready-made garment goods (174), agrochemistry (242), production of musical instruments and records (363+223) and the remainder of processing industry including weapons (296+366)[4].

[4] Numbers in round parentheses are codes of the industries in OKEC.

Table 9

Estimates of regression coefficients of model (2) for the complementary subpopulations of size 37 from which $\hat{\beta}^{(LTS,n,h)}$ selected a subpopulation containing 33 cases

Estimates of regression coefficients in model (2)

h = 33

Item	Estimate of coefficients	Standard error of estimate	t-value	P-value
intercept	-15.0073	3.4658	-4.3301	0.000211
VS/PH	0.098	0.0705	1.3891	0.177068
SZ/PH	0.9203	0.2418	3.8062	0.000814
K/PH	1.3243	0.2641	5.0137	0.000036
CR3	1.5501	1.1318	1.3696	0.182991
DP	1.0762	0.4548	2.3663	0.026028
TFPW	3.3645	0.9949	3.5828	0.001434
BAL	-0.0793	0.6304	-0.1257	0.900952
DP	1.076	0.4548	2.366	0.026

Table 10

Sum of square, estimate of scale, coefficient of determination and Durbin-Watson statistic of model, estimate of which is given in Table 9

Sum of squares	55.445
Scale estimate	1.4892
Coefficient of determination	0.6564
Durbin-Watson statistic	1.842

Tables 11 and 12 (given below) "prove" that above proposed division of data exhibit some sense. It seems quite acceptable that capital and labour are in (stabilized) market economy the substitutes, each of other, i.e. their product is approximately constant. This fact is formally expressed (in economic theories) e. g. by Cobb-Douglass production function, see Kmenta (1986). It is usually written in a form

$$Q = \mu \cdot L^{\lambda} \cdot K^{1-\lambda}.$$

In other words, we may expect e.g. that the labour will be proportional to one over capital or to one over square root of capital or to a combination of both these functions or to some other appropriate function of hyperbolic character. Roughly speaking, labour times some increasing function of capital should be constant. But the last formulation shows that there is a problem. In fact we have in our data industries of different sizes, so that we would need both variables, labour as well as capital, to standardize on a "unit" level. We have used for this purpose the explanatory variable wages (for standardizing capital) and variable sale for labour. And it appeared that for the *main* subpopulations the model

$$\frac{K_i}{W_i} = a_1 + b_1 \cdot \frac{S_i}{L_i} + \varepsilon_i^{(1)}, \quad i = 1, 2, ..., n \tag{3}$$

has satisfactorily high coefficient of determination (for smaller sample sizes, i. e. for 52, 53 and 54) as Table 11 shows[5].

On the other hand, for the *complementary* subpopulations the direct proportionality between labour and capital, i.e. the model

$$\frac{K_i}{W_i} = a_2 + b_2 \cdot \frac{L_i}{S_i} + \varepsilon_i^{(2)}, \quad i = 1, 2, ..., n \tag{4}$$

(4) is valid, see Table 12. Moreover, coefficients of determination of model (4) for *main* subpopulations are substantially smaller than for model (3), see again Table 11. Similarly, coefficients of determination of model (3) for *complementary* subpopulations are really small.

Table 11
Coefficients of determination of models (3) and (4) (the last but one and the last row, respectively) for *main* subpopulations of given sizes

Number of cases	52	53	54	55	56	57
Coefficients of determination	0.620	0.615	0.489	0.235	0.148	0.161
Coefficients of determination	0.308	0.306	0.255	0.034	0.022	0.011

[5]Recent study of possible models for *foreigner direct investment* indicates that presumably the numbers 52, 53 or 54 represent the upper limit of the size of a group of industries in the Czech economy having *market economy* character.

Table 12
Coefficients of determination of models (4) and (3) (the last but one and the last row, respectively) for *complementary* subpopulations of given sizes

Number of cases	39	38	37	36	35	34
Coefficients of determination	0.709	0.711	0.708	0.699	0.694	0.687
Coefficients of determination	0.008	0.008	0.005	0.003	0	0.000

Steeply decreasing values of coefficients of determination for *main* subpopulations indicate that we are presumably really on an "upper" bound of the size of subpopulation which still exhibit property which was called above *(stabilized) market economy*. Probably it would be reasonable to restrict even somewhat more size of the *main* subpopulations (and of course, increase sizes of *complementary* subpopulations).

CONCLUSIONS

The results of our analysis hinted that the robust methods of estimating regression coefficients in linear regression model are really powerful tools for finding "true " (hidden) structure of data. Not having at hand these tools we had to conclude (may be with a despair) that the classic least squares are not able to find the determinants of exports, although it is intuitively clear that some factors which have an influence on the amount of exported goods should exist. We have used the *least trimmed squares* but there is of course plenty others estimators with high breakdown point which may contribute to the analysis by their results. E. g. there is available very reliable algorithm by Boček *&* Lachout (1993) for the *least median of squares*. Applying more estimators may lead to the *diversity of estimates* (for details see Víšek (1997) or (2000 a)) which may be at the first glance surprising but then giving more possibilities.

The results gathered in Table 11 and 12 then confirm that there may be two kinds of industries, first ones which behave like industries in (stabilized) market economy and the rest which still work as under centrally planned economy. Of course we should speak probably about a main feature of given industry because it seems rather unlikely that all factories in an industry which fell in our analysis into the *main* subpop-

ulation behave reasonably. Similarly a factory falling into an industry belonging into *complementary* subpopulation need not behave inevitably as in centrally planned economy. It hints that it may be (much) more interesting to search for similar division (on *(stabilized) market economy* and *centrally planned economy*) on the level of factories. Let us hope that also in the Czech republic, similarly as in most countries all over the world, corresponding data will be soon available.

References

[1] Bickel, P. J. (1975): One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc. 70, 428–433.*

[2] Boček, P., P. Lachout (1993): Linear programming approach to LMS-estimation. *Memorial volume of Comput. Statist. & Data Analysis 19(1995), 129 - 134.*

[3] Čížek, P., J. Á. Víšek (2000): Least trimmed squares. *XPLORE, Application Guide, 49 - 64.* Springer Verlag, (2000b), Berlin, eds. W. Hardle, Z. Hlavka, S. Klinke, ISBN 3-540-67545-0.

[4] Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A 222, pp. 309-368.*

[5] Hampel, F.R., E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel (1986): *Robust Statistics - The Approach Based on Influence Functions.* New York: J.Wiley Sons.

[6] Hettmansperger, T. P., S. J. Sheather (1992): A Cautionary Note on the Method of Least Median Squares. *The American Statistician 46, 79–83.*

[7] Jurečková, J., P. K. Sen (1993): Regression rank scores scale statistics and studentization in linear models. *Proceedings of the Fifth Prague Symposium on Asymptotic Statistics, Physica Verlag, 111-121.*

[8] Kmenta, J. (1986): *Elements of econometrics*, Macmillan Publishing Company, New York.

[9] Maronna, R.A., O. H. Bustos, V. J. Yohai (1979): Bias- and efficiency-robustness of general M-estimators for regression with random carriers. *In Smoothing Techniques for Curve Estimation.*

Eds. T. Gasser and M. Rosenblatt, New York: Springer-Verlag, 91 - 116.

[10] Rousseeuw, P.J., A. M. Leroy (1987): *Robust Regression and Outlier Detection.* New York: J.Wiley *&* Sons.

[11] Ruppert, D., R. J. Carroll (1980): Trimmed least squares estimation in linear model. *J. American Statist. Ass.,75 ,(372), pp. 828–838.*

[12] Víšek, J. Á. (1994): A cautionary note on the method of Least Median of Squares reconsidered. *Transactions of the Twelfth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague, 1994, 254 - 259.*

[13] Víšek, J. Á. (1996 a): On high breakdown point estimation. *Computational Statistics (1996) 11:137-146, Berlin.*

[14] Víšek, J. Á. (1996 b): Sensitivity analysis of M-estimates. *Annals of the Institute of Statistical Mathematics, 48(1996), 469-495.*

[15] Víšek, J. Á. (1997): Contamination level and sensitivity of robust tests. *Handbook of Statistics, volume 15, 633 - 642, eds. G. S. Maddala & C. R. Rao, 1997, Amsterdam: Elsevier Science B. V., ISBN 0-444-82172-4*

[16] Víšek, J. Á. (1999 a): Robust estimation of regression model. *Bulletin of the Czech Econometric Society, Volume 9/1999, 57 - 79.*

[17] Víšek, J. Á. (1999 b): The least trimmed squares - random carriers. *Bulletin of the Czech Econometric Society, Volume 10/1999, 1 - 30.*

[18] Víšek, J. Á. (2000 a): On the diversity of estimates. *Computational Statistics and Data Analysis 34, (2000), 67 - 89.*

[19] Víšek, J. Á. (2000 b): Regression with high breakdown point. Submitted to *ROBUST 2000*

Affiliation: *Víšek J. Á.*

Department of Macroeconomics and Econometrics, Institute of Economic Studies, Faculty of Social Sciences, Charles University

&

Department of Stochastic Informatics, Institute of Information Theory and Automation, Academy of Sciences of Czech Republic

Mailing address:

Opletalova ulice 26, CZ - 110 01 Prague 1, Czech Republic,

e-mail: visek@mbox.fsv.cuni.cz