

APLIKOVANÉ KVANTITATIVNÍ METODY

Vypracovali:

Dominik Stroukal a Martin Regéczy

ÚVOD

Pod pojmem „výnosy ze vzdělání“ si můžeme představit (i) soukromé výnosy, (ii) společenské výnosy a (iii) výnos z produktivity práce. Základním komponentem měření všech těchto druhů výnosů je měření dopadu vzdělání na výdělek, což je velmi oblíbeným tématem mnoha akademických prací, pravděpodobně z důvodu své vhodnosti k využití ekonometrických nástrojů. [Blundell *et al*, 2001]

V této práci se zaměříme na kvantifikaci dopadu dalšího roku vzdělání na výdělek. Za použití kontrolních proměnných pro rasu, pracovní zkušenosti a lokalitu jsme zjistili, že každý další rok vzdělání přináší o 6.5% vyšší hodinovou mzdu. Přikládáme i odhady za použití instrumentálních proměnných a zpracování, které pro každý rok vzdělání obsahuje binární proměnnou, z čehož si můžeme odvodit problematiku samotné ideje výnosů ze vzdělání.

Tato práce má následující strukturu: započneme (I) teoretickým rámcem, kde prezentujeme verbální model a naši představu o výnosech investice do vzdělání a (II) popisem dat, statistickým popisem použitých proměnných a základním přehledem systému vzdělání v USA.

Třetí sekce obsahuje kompletní těžiště celé práce, tedy (III) Model, testy a interpretaci, kterou jsme rozdělili do následujících podčástí (a) základní model, obsahující základní problémy vznikající při odhadu; (b) použité kontrolní proměnné a jejich zdůvodnění; (c) možné další proměnné, nezahrnuté, či nezjištěné; (d) funkční transformace; (e) testy a interpretace výsledků; (f) odhad bez jedné proměnné; (g) instrumentální proměnné; a část věnovaná (h) Gauss-Markovým předpokladům.

Tato hlavní část je následována (IV) kritikou modelu a alternativním postupem, který poukazuje na chybný předpoklad linearitu výnosů ze vzdělání, a konečně závěrem.

Úvod 1

I Teoretický rámec 2

Verbální model 2

Výnosnost investice do vzdělání 2

II Data 3

Přehled použitých proměnných 3

Vzdělání v USA 4

III Model – Testy – Interpretace 4

Základní model 4

Kontrolní proměnné 5

Možné další proměnné 6

Funkční transformace 6

Testy a Interpretace Výsledků 6

Odhad bez jedné proměnné 7

Instrumentální proměnné 7

GM předpoklady 9

Multikolinearita 9

Heteroskedasticita 10

IV. Kritika modelu a alternativní postup 11

Závěr (předposlední) 12

Použitá literatura 12

poznámka: v III. části jsme se z důvodů logické návaznosti odklonili o předepsané struktury práce. Zamýšlenou část IV. a V. jsme zahrnuli přímo do III a odpovídá jim část „Odhad bez jedné proměnné“ a „Instrumentální proměnné“.

I TEORETICKÝ RÁMEC

VERBÁLNÍ MODEL

Jedním z přínosů marginální revoluce k ekonomii trhů práce byla teze, že rovnovážná mzda (na dokonalém trhu) je rovna meznímu produktu pracovníka. Oprostíme se od květnatosti této teorie, protože se nám hodí pouze ona myšlenka, že výše mzdy je zcela určena prací pracovníka. Pokud je schopnější získá vyšší mzdu za kratší čas, pokud je méně schopný dostane i menší mzdu.

Pokud se tedy snažíme vysvětlit výši mzdy, hlavním faktorem je kvalita pracovníka, což je něco velmi těžce uchopitelného a ještě hůře měřitelného. Zaměstnavatel, který najímá nového pracovníka, nevidí, jestli je pracovník kvalitní nebo nikoli, a snaží se vybrat nejlepšího pracovníka nejefektivnějším způsobem. [Thales a Sunstein, 2008] Reaguje v závislosti na dostupných signálech, které pomáhají překonat informační asymetrii. Jedním ze signálů kvality je vzdělání. Předpokládáme, že vzdělání je investicí do lidského kapitálu¹, který je následně efektivnější. **Mzda by tedy měla pozitivně záviset na výši vzdělání**, ale přesný efekt vzdělání bude nižší z důvodu pozitivního vychýlení způsobeného kvalitními jedinci, kteří by získávali vyšší mzdu i bez vzdělání.

Další z důležitých faktorů ovlivňujících mzdu jsou pracovní zkušenosti, které jsou oproti teoretickému vzdělání velmi praktické. Při záměru maximalizovat mzdu se mohou zpočátku jevit zkušenosti jako substitut, kdy několik let vzdělání ve výnosech nahradí několik let zkušeností. Uvědomujeme si, že **zkušenosti budou ovlivňovat mzdu zpočátku pozitivně a později negativně**. Toto bude ještě markantnější v našem omezeném data setu, kde mnoho pracovních zkušeností znamená spíše málo vzdělání a podřadnou práci.

Pohlaví, barvu pleti a další vrozené aspekty lze jen těžko ovlivnit. Domníváme se, že tyto aspekty mají vliv na výši mzdy jakožto jeden z identifikátorů diskriminace. Předpokládáme existenci diskriminace černochů a tudíž, že **občané tmavé barvy pleti mají nižší mzdy**. Dalším aspektem, který si uvědomujeme je možný nižší výnos afro-američanů ze vzdělání, který je zkoumán v tzv. „white-black“ gap.

Tyto faktory jsou dále ovlivněni místem, kde se daná práce nachází. Je velký rozdíl mezi Prahou a Karvinou i mezi centrem a periferií. Předpokládáme, že **mzda bude vyšší v centru měst a na severu USA**, z důvodu shlukování větších firem v centrech měst a větší průmyslové síle severu nad spíše zemědělsky založeným jihem.

Náš ekonomický model by tedy vypadal následovně²:

$$wage_i = f(educ_i, black_i, exper_i, smsa_i, south_i)$$

(+)(-)(+)³(+)(-)

VÝNOSNOST INVESTICE DO VZDĚLÁNÍ

Výnosností investice do vzdělání rozumíme změnu (kladnou i zápornou) ve výnosech jednotlivce v závislosti na zvýšení vzdělání o jednotku. Za jednotku je zvolen rok vzdělání a změnou ve výnosech pak změna mzdy, která je měřena v centech.

Za základní problematiku výnosnosti je (z pochopitelných důvodů špatné měřitelnosti) absence výnosů nepeněžních, což může být flexibilnější pracovní doba, příjemnější šéf, dobrý

¹ Plně si uvědomujeme problém self-selekce, kdy kvalitnější jedinci, kteří jsou zároveň vzdělání by měli vyšší mzdy i bez tohoto signálu. Budeme jej ovšem z důvodu zjednodušení ignorovat.

² Pro jednoduchost opouštíme od podmínek prvního a druhého řádu.

³ Závislost má podobu klesajících mezních výnosů z praxe:

$$\frac{\partial wage}{\partial exper} > 0 \wedge \left(\frac{\partial wage}{\partial exper}\right)' > 0$$

společenský status dané pozice, hezké toalety, apod.⁴. Práce se tak vztahuje pouze na výnosy ryze mzdové.

Naším cílem je změřit ze zadaných dat co nejpřesněji procentuální⁵ změnu hodinové mzdy v závislosti na zvýšení vzdělání o jeden dodatečný rok.

II DATA

V této části diskutujeme (i) použité proměnné, jejich interpretaci a uzavřeme ji (ii) přehledem systému vzdělání v USA v maximální stručnosti pouze pro nastínění problémů, které přináší pro naše zkoumání.

PŘEHLED POUŽITÝCH PROMĚNNÝCH

Zadaná data pochází z průzkumu dotazujícího kohortu 1000⁶ mladých mužů v letech 1975 a 1985. Předpokládáme, že výzkum byl proveden v USA. Některé otázky směřovaly i před rok 1975, jmenovitě ke věku 14 let každého respondenta (v době prvního kola dotazování v roce 1975 to byl věk nejmladších respondentů). Data z roku 1975 používáme až později při modelu s instrumentálními proměnnými, následující jsou výhradně z pozdějšího dotazování.

V našem zkoumání používáme následující proměnné (statistický popis v tabulce 1): *wage* - hodinová mzda v centech z roku 1985; *educ* - vzdělání (v letech) v roce 1985; *exper* - pracovní zkušenost dotazovaného, vypočítané jako věk – *educ* – 6; *black* (=1, pokud je dotazovaný tmavé barvy pleti); *smsa* (=1, pokud bydlel v roce 1985 v městské/příměstské oblasti); a *south* (=1, pokud bydlel v roce 1985 na jihu).

Tabulka 1 - Statistický popis použitého data setu

	n	min	max	průměr	sm. odch.	Q1	medián	Q3
wage	1000	100	2244	560.50	244.44	385	529	693
educ	1000	3	18	13.27	2.66	12	13	16
black	1000	0	1	0.22	-	-	-	-
smsa	1000	0	1	0.73	-	-	-	-
south	1000	0	1	0.4	-	-	-	-
exper	1000	0	21	8.92	4.08	6	8	11
near2	1000	0	1	0.44	-	-	-	-
near4	1000	0	1	0.69	-	-	-	-
sinmom14	1000	0	1	0.11	-	-	-	-
libcrd14	998	0	1	0.68	-	-	-	-
fatheduc	766	0	18	7.62	5.37	2	8	12
motheduc	888	0	18	9.15	4.46	8	11	12

⁴ Mezi další diskutované výnosy patří osobní nepeněžní výnosy či výnosy společenské [Blundell, Dearden a Sianesi, 2001].

⁵ více viz 3c)

⁶ Uvedený vzorek obsahuje 1000 pozorování. Protože jsme obeznámeni s faktem, že jde jen o vybraných 1000 pozorování z většího vzorku, pro důkladnější a přesnější analýzu by bylo lepší vzorek rozšířit, tedy použít všechna pozorování. Zároveň nejsme obeznámeni se způsobem výběru dat, tedy způsobem jakým byli respondenti vybráni a dotazováni. Při takovém výzkumu je totiž podstatné si uvědomit možné vychýlení dat samotnými respondenty. Je možné, že chudší si přidávají a bohatí ubírají nebo naopak v závislosti na kulturních specifikách [Thaler a Sunstein, 2008]. Opomenutí adresovat toto systematické vychýlení u této a dalších charakteristik již během sběru dat zamlžuje výsledná data a činí, z nich vytvořené odhady, náchylné k chybě.

Z deskriptivních statistik je zřejmé, že 50% jednotlivců ve vzorku má mzdu v úzkém rozmezí mezi 385 a 693 centy. Data set byl otestován na přítomnost odlehlých pozorování, s kterými jsme se rozhodli nic nedělat⁷ z důvodu neznalosti postupu sběru dat, který mohl některé nepřesnosti způsobit⁸. Zajímavým zjištěním může být viditelný posun v množství vzdělání za jednu generaci. Znamená to, že naše výsledky jsou časově specifické a ztrácejí na obecnosti.

Dataset obsahuje i chybějící pozorování, ale nikoli u základní sady zkoumaných proměnných. U instrumentálních proměnných chybí pozorování minimálně nebo vůbec, proto jsme se rozhodli je nenahrazovat průměrem či mediánem, ale tato pozorování vynechat. Pozorování chybí u některých instrumentálních proměnných, ale i za použití všech proměnných je 73% pozorování kompletní.

VZDĚLÁNÍ V USA

Základní škola (tzv.K6) v USA je šest tříd, na které navazuje mnoho různě dlouhých a specializovaných větví vzdělání, které se spojuje až po 12ti letech vzdělávání, kdy studenti ukončují střední školu (s tzv. highschool diploma nebo ekvivalent). Po střední škole mohou absolvovat 2-3 roky na college nebo 4 roky na univerzitě. Po 4 letech na univerzitě je velmi oblíbené před dalším studiem pracovat a pak jej dokončit dvěma⁹ lety studia navazujícího.

Z toho je jasné, že se jedná o diskrétní veličinu s nelineárním vývojem a zdá se nám nevhodné ji tímto způsobem používat jako vysvětlující proměnnou se snahou získat spojitě výnosy ze vzdělání¹⁰. Je jasné, že tyto výnosy jsou skokové a nikoli rovnoměrně rozdělené a člověk, s nedokončenou střední školou nebude mít z tohoto nedokončeného vzdělání přehnané výnosy ze vzdělání.

III MODEL – TESTY – INTERPRETACE

V této sekci nejprve objasníme (i) základní model, který bude postupně rozšířen v (ii) kompletní model přidáním kontrolních a (iii) instrumentálních proměnných. (iv) Výsledky testů jsou shrnuty v tabulce xy a následovány (v) interpretací výsledků.

ZÁKLADNÍ MODEL

Hledáme výnosy ze vzdělání, tedy zjišťujeme výši konkrétního koeficientu u jednoho z n faktorů ovlivňující mzdu, což si můžeme představit následovně:¹¹

$$wage_i = \alpha educ_i + \beta X_i^T + \varepsilon_i$$

Před dalším rozšířením modelu se chvíli zastavíme u dvou strukturálních problémů, které plynou z podstaty veličin, které používáme.

Prvním problémem při vysvětlování mzdy není splnění předpokladu normality náhodné složky, protože (a) cena práce nemůže být mzda záporná a z důvodu státní regulace (b) ani nižší než minimální mzda stanovená na federální, či státní úrovni¹².

⁷ V části IV. nicméně diskutujeme výsledky robustního odhadu získaného kvantilovou regresi.

⁸ Například z důvodu existence hodinových výnosů pod minimální mzdou, která je diskutována v části III věnující se základnímu modelu

⁹ Maximální počet let vzdělání je 18, z důvodu chybějících dodatečných informací předpokládáme neexistenci doktorského studia.

¹⁰ Z tohoto důvodu jsme připojili samostatnou sekci s alternativním řešením daného problému, který naráží na problém nespecifikace dokončeného vzdělání v data setu, ale nese zajímavé informace o rozdělení výnosu ze vzdělání mezi jednotlivými odstudovanými roky.

¹¹ Jedná se o hypotetický model, kde α vysvětluje výnosy ze vzdělání, β je vektor koeficientů a X vektor všech pozorovatelných vysvětlujících proměnných.

Druhým problémem je předpoklad lineární vlastnosti výnosů ze vzdělání, tedy, že výnos z první třídy je stejný jako z posledního roku navazujícího studia, který není příliš realistický.

Jednou z možností jak řešit oba problémy je logaritmická transformace mzdy, která byla zvolena, aby vyřešila oba tyto problémy, protože následně (1) náhodná složka je bližší požadavku normality a (2) výnosy ze vzdělání jsou v procentech a tudíž blíže realitě. [Wooldridge, 2006]

KONTROLNÍ PROMĚNNÉ

V základu odhadujeme následující rovnici populace:

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + \varepsilon_i$$

Je zřejmé, že v případě odhadu tohoto modelu získáme velmi vychýlený model z důvodu podspecifikovanosti. Existuje zde totiž mnoho proměnných, které nám odhad výnosu ze vzdělání vychylují (viz. verbální model). Z tohoto důvodu jsme zvolili následující model obsahující několik dodatečných vysvětlujících proměnných, které mají za cíl snížit šum v náhodné složce.

$$\ln(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 black_i + \beta_3 exper_i + \beta_4 exper_i^2 + \beta_5 smsa_i + \beta_6 south_i + \varepsilon_i$$

Vysvětlovanou proměnnou *wage* a základní vysvětlovanou *educ* jsme již rozebrali.

Binární proměnná *black* (1 pro občana tmavé pleti, 0 jiné barvy) byla zvolena z důvodu, že v případě diskriminace je hned po ženách velkým tématem problém černochů. Tento problém byl v daném období ještě vyšší, než nyní. Při začleňování této proměnné do modelu vycházíme z předpokladu diskriminace a již nás nezajímá jakého druhu je, ale pouze že tento jev existuje a projevuje se rozdílnou mzdou. *Zajímavé by bylo mít binární proměnnou i za hispánské obyvatelstvo, či kolik procent má daná menšina zastoupení v daném městě*, protože i čím je zastoupení menší, tím je menšina více diskriminována [Wooldridge, 2006]. Zvažovali jsme též rozdělit výnosy ze vzdělání pro bílé a černé. Ač mnoho výzkumů [Fryer *et al*, 2004 a 2006] zkoumá black-white gap a je tedy možností, že černí mají ze vzdělání méně¹³.

Naši zkušenosti je blízke, že vysokoškolské vzdělání v českých podmínkách bez praxe nese méně, než středoškolské s praxí. Zařadili jsme tedy do našeho modelu proměnnou *exper* jakožto proxy proměnnou pro délku praxe. Ačkoliv to činí velmi nedokonale je tím nejlepším, co máme. *Získat přesnější data, která by jasně vystihovala pracovní zkušenost, či relevantnost dané pracovní zkušenosti pro zaměstnavatele, či dokonce tenure, tedy počet let pro daného zaměstnavatele by bylo jistě výzkumu jenom prospěšné. U exper jsme si uvědomili, že se nemůže jednat o lineární trend.*

¹² Bohužel zvolený data set postrádá informaci, ve kterém státu byla vyplácena a kterou sazbu tedy použít pro případnou analýzu jakožto kontrolní proměnná (rozdíl v minimální mzdě má totiž jistě vliv na výši nominální mzdy). Z důvodů minimálních rozdílů od federální sazby napříč americkými státy můžeme nicméně předpokládat, že minimální mzda byla 3.35\$ (History of Federal Minimum Wage Rates Under the Fair Labor Standards Act, 1938 – 2009 - <http://www.dol.gov/whd/minwage/chart.htm>). Zde ovšem narážíme na pozorování, která jsou hluboko pod naší minimální mzdou (celkově skoro 10% pozorování je pod touto výší). Jedním z vysvětlení může být, že respondenti odpovídali v měsíčním, či ročním vyjádření mzdy, která se následně přepočítala na hodinovou mzdu. Všechna pozorování pod minimální mzdou jsou buď lehce ponižená (může být způsobené zaokrouhlením/započtením zdanění), či se jedná o snížené úvazky, které nebyly při sběru dat zohledněny.

¹³ Tento předpoklad jsme zkusili odhadnout pomocí regrese zařazením proměnné *black*educ*, ale jevila se jako nevýznamná a tudíž jsme ji z modelu vyloučili, abychom jej dále nepřespecifikovali.

Očekávali jsme klesající výnosy z praxe, a proto jsme umožnili zakřivení. První rok, či dva pracovní zkušenosti se nám zdají velkou výhodou, ale další dva, když jich má člověk již 10 je malým přínosem. V případě daného věkového rozložení jako je v našem datasetu značí vysoká hodnota *exper* vysoký věk nebo opravdu dlouhou pracovní zkušenost, která ovšem musela být na úkor vzdělání. Tudíž pozice, kterou daný člověk zastává, bude spíše hůře placená, čímž vysvětlujeme následující negativní vliv při vyšších hodnotách.

Oblíbenou frázi realitních makléřů je „*Location, location, location*“. Vliv lokace je velmi podstatný na mzdu i například v České republice. Stejně vzdělaní lidé berou diametrálně odlišné platy v Praze a na Moravě. Velikost platu je tedy silně ovlivněna lokací, z toho důvodu jsme zahrnuli *smsa* a *south*. Předpokládali jsme, že historicky průmyslovější sever bude mít vyšší mzdy, než chudší a více zemědělsky založený jih. Zvažovali jsme pro lokaci využít dummy proměnné z roku 75, ale nakonec jsme odmítli předpoklad rigidity mobility práce v USA, která je dávana Evropě za příklad a ještě je umocněna faktem, že jsou to data před studiem na college/university, tedy před přesunem za vyšším vzděláním, které může lokalitou, ve které v 85 získává plat velmi zamíchat. K lepší využitelnosti obou proměnných by bylo potřeba znát přesnější vymezení „jihu“ či „příměstské oblasti“, které ze zadaných dat nebyly známy. Rozdělení dle jednotlivých regionů jako v případě roku 75 by velmi pomohlo lépe lokalizovat významně odlišné regiony.

MOŽNÉ DALŠÍ PROMĚNNÉ

Vliv na mzdu má v reálném světě mnohem více faktorů, než ty, které jsme vybrali a provádíme s nimi analýzu. V našem modelu, který zkoumá vliv vzdělání na mzdu, bychom tedy v ideálním případě počítali s talentem obecně či snad s talentem k dané jednotlivé práci. Model by se dal blíže specifikovat, pokud by existovaly dummy proměnné pro místa zaměstnání v roce 1985. Rozdělení regionů například na jednotlivé státy by mohlo vysvětlit i rozdíly mezi jednotlivými státy na jihu, které jsou v našich datech zbytečně agregovány.

FUNKČNÍ TRANSFORMACE

V otázce transformací jsme zvolili (i) logaritmickou transformaci mzdy. Výhody, které nám to nese pro naše odhady a při interpretaci, tedy vyšší normalita náhodné složky u prvního a vyšší reálnost procentuálních výnosů ze vzdělání, shrneme v III/i; (ii) možnost zakřivení *exper*², které nám poskytuje lepší odhad skutečných výnosů z praxe, shrneme v III/ii.

TESTY A INTERPRETACE VÝSLEDKŮ

Výše představený model jsme odhadli metodou nejmenších čtverců. Následně jsme vybrali sadu instrumentů pro *educ* a dvoustupňovou metodou nejmenších čtverců jsme odhadli koeficienty jednotlivých proměnných. Výsledky testů jsou uvedeny v tabulce 5 v příloze.

Metodou nejmenších čtverců jsme zjistili, že každý další rok vzdělání, *ceteris paribus*, přidává ke mzdě 0.65%. To je následně ovlivňováno barvou pleti, kdy člověk s tmavou barvou kůže, *c.p.*, má v průměru o 18.2% nižší mzdu, než člověk s bílou barvou kůže. Vliv pracovní zkušenosti má námi odhadovaný, tedy nelineární průběh, tedy s klesajícími výnosy z praxe¹⁴. Vliv lokace je opět stejný jak jsme předpokládali a má, *c.p.*, následující vliv: o 14.7% vyšší mzda pro lidi žijící v centru a o 14.1% nižší mzda pro ty žijící na jihu na rozdíl od těch žijících na okraji, či na severu.

¹⁴ Výnos z dodatečného roku praxe se přibližně rovná $0.68 * \text{exper} - 2 * 0.002 * \text{exper}$. Bod zlomu nastává při 170 letech, zdaleka mimo náš pozorovaný úsek.

Použití různých instrumentálních proměnných přineslo různé výsledky pro výnosy ze vzdělání. Výnosy se zmenšily z 0.65 na rozmezí od 0.32 do 0.53 pro použití různých kombinací instrumentů. Původní koeficient 0.65 je až dvakrát vyšší. To je konzistentní s předpokladem chybějících proměnných v původním modelu (například talent, motivace, atd. viz výše). Všechny směrodatné odchylky u education se více než zdvojnásobily, což lze předpokládat [Wooldridge, 2006].

Protože jsou data pouze z jednoho zkoumání a instrumentální proměnné nesplňují předpoklad nulové korelace se mzdou, potom nelze přesně určit, zdali je blíže realitě hodnota 0.65 z MNČ či hodnoty nižší při použití instrumentálních proměnných.

ODHAD BEZ JEDNÉ PROMĚNNÉ

Provedli jsme následující odhad bez proměnné *exper*, tedy i její mocniny (rozdíly shrnuje tabulka 5).

Rozhodli jsme se vybrat proměnnou, která je nejvíce korelovaná se vzděláním. Dle specifikace proměnné *exper* je to zřetelně právě tato proměnná. Jejím vynecháním z modelu vidíme takřka nulové změny u *msa*, *south* a *black*. Ovšem proměnná *educ* klesla o více než polovinu, zatímco směrodatná odchylka se takřka nezměnila. To je způsobeno tím, že část variability, která byla vysvětlená proměnnou *exper* se rozdělila mezi *educ* a náhodnou složku, čímž zvýšila nepřesnost modelu a vyšší rozptyl koeficientu *educ*. Jelikož proměnná *exper* by měla být součástí modelu je i odhad proměnné *educ* vychýlený z důvodu podspecifikovanosti modelu. Vynechání také způsobilo pád koeficientu determinace z 0.274 na 0.192. Model se zhoršil nejen vzhledem k R^2 ale také se zvýšila p-hodnota testovací statistiky F.

Vynechání proměnné *exper* vychyluje koeficient pro *educ* natolik, že je nemožné s tímto modelem pracovat. Proměnná *exper* do modelu patří, i přes výše diskutované vyhrady. Vynechání proměnné upozorňuje na sníženou kvalitu modelu, a protože je zřejmé, že podobný efekt by mohlo mít opomenutí schopností či talentu. Této skutečnosti jsme si vědomi a hledali jsme vhodnou proxy proměnnou, která by tuto nepozorovatelnou vlastnost zastoupila. Jako proxy schopností či talentu se z dostupných dat nabízí IQ, případně KWW¹⁵. Jako proxy je možné také použít jejich kombinaci.

IQ jsme zavrhlí, protože měří pouze abstraktní inteligenci, nikoliv praktickou, sociální či emoční¹⁶. IQ nedokáže změřit kreativitu a neintelektuální schopnosti, mnohdy důležité pro výkon povolání. Současně vzniká v souvislosti s dotazováním na IQ problém self-selekce, kdy ti s nízkým IQ mají tendenci svůj výsledek neuvádět (případně si své IQ ani nezjistí) a naopak. Uvedené hodnoty IQ tedy mohou být vychýlené a nepoužíváme je.

U proměnné KWW neznáme přesně způsob jeho měření. Navíc se nedomníváme, že by bylo možné sestavit test, jehož výsledky by byly kvalitní proxy pro tak různorodou vlastnost, jako jsou schopnosti.

Protože jsme nenalezli vhodnou proxy proměnnou pro schopnosti a je tedy možné, že náš odhad je vychýlený, pokusíme nyní odhad zpřesnit pomocí instrumentálních proměnných.

INSTRUMENTÁLNÍ PROMĚNNÉ

Pro instrumentální proměnné hledáme takové proměnné, které vykazují korelaci k *educ* a zároveň nejsou korelovány s vysvětlovanou proměnnou *wage*, konkrétně s nevysvětlenou

¹⁵ Souhrnné hodnocení z testu ohledně pracovních činností a trhu práce (pracovní náplň a kvalifikační nároky 10 různých zaměstnání, poměr platového ohodnocení 8 párů zaměstnání), tzv. „Knowledge of World of Work Score“ (odtud jméno proměnné).

¹⁶ Více o druzích IQ například na <http://www.mensa.cz/mensa.cz/site/casopis/zajimavosti/intelligence.html>

variabilitou ukrytou v náhodné složce náhodnou složkou. Implicitně předpokládáme, že *educ* koreluje s náhodnou složkou, tedy existuje zde nevysvětlená variabilita způsobená talentem daného člověka, a splňuje tak podmínku nutnou k provedení následné analýzy. Je tedy zřejmé, že získat IV, která by odpovídala těmto podmínkám je velmi náročné, až nemožné. Obzvláště v případě, kdy i malá korelace IV s náhodnou složkou způsobuje velké vychýlení odhadů. [Wooldridge, 2006]

Z důvodu neznalosti náhodné složky, kontrolujeme splnění podmínek logikou ekonomickou a pomocí korelačních matic.

Jako první jsme zvažovali použít proměnné *near2* a *near4*, ale zavrhlí jsme je v důsledku neprůkazné ekonomické verifikace. Tyto instrumenty pro vzdálenost od škol neodpovídají výše diskutovaným podmínkám; tedy výše mzdy není nezávislá na vzdálenosti od školy. Domníváme se, že úzký vztah mezi výší mzdy a vzdáleností od školy je způsoben velikostí či vyspělostí města či regionu. Je zřejmé, že člověk žijící v blízkosti vysoké školy je pravděpodobně obyvatelem velkého města či vyspělého regionu. Obyvatelé vyspělých oblastí potom mají obecně vyšší mzdy způsobené například širšími pracovními příležitostmi, vyšší specializací práce či relativně vyššími životními náklady.

Přestože se vzdáleností od školy někteří autoři pracují [Card, 1995], my jsme se rozhodli i vzhledem k následujícímu testu korelace proměnné *near2* a *near4* jako instrumenty zavrhnout. Tabulka 3 ukazuje nejen na korelaci mezi mzdou a vzdáleností od školy, dokonce u obou proměnných je korelace ke mzdě vyšší než ke vzdělání.

Tabulka 3 - Korelační koeficienty pro vzdálenost

<i>wage</i>	<i>educ</i>	<i>nearc2</i>	<i>nearc4</i>	
1	0.28	0.08	0.16	<i>wage</i>
	1	0.01	0,13	<i>educ</i>
		1	0.08	<i>nearc2</i>
			1	<i>nearc4</i>

5% kritická hodnota (oboustranná) = 0,06 pro n = 1000

Současně z MNČ na *educ* za pomoci všech proměnných základního modelu a proměnných *near2* respektive *near4* vidíme, že pokud při ostatních neměnných faktorech muž bydlel blízko dvouleté školy, dosáhl o 0.03 roku vzdělání více, respektive téměř třetinu roku vzdělání v případě čtyřleté školy.

$$educ = 16,66 + 0,29near4 \dots RSQ=.442, p\text{-hodnota} (near4) = .05$$

$$educ = 16,64 + 0,03near2 \dots RSQ=.439, p\text{-hodnota} (near2) = .81$$

Proměnná *near2* vychází jak statisticky (p-hodnota 0.81) tak prakticky (pouze 0.03 roku vzdělání – téměř týden navíc) nevýznamně. Proměnnou *near4* by bylo možné z těchto podmínek použít, zavrhuje ji však z důvodů silné korelace s *wage*.

Pro využití metody instrumentálních proměnných jsme tedy zvolili proměnné *sinmom14*, *fatheduc*, *motheduc* a *libcrd14*. Domníváme se, že vzdělání rodičů má pozitivní vliv na vzdělání dětí, přičemž by nemělo významně ovlivňovat výši mzdy. Z našich dat vyplývá, že mzdu ovlivňuje, ale v menší míře než v jaké ovlivňuje¹⁷ vzdělání. Obdobně jsme uvažovali i u kartičky do knihovny ve 14 letech, kde vlastnění kartičky by mělo zvyšovat výnosy ze vzdělání, ale již nikoli mzdu. Zvláštní instrumentální proměnou pak je *sinmom14*, tedy pokud respondent ve 14 letech žil se svobodnou matkou. Domníváme se, že tento fakt může negativně korelovat s výší vzdělání, protože je těžší zabezpečit rodinu a respondent nastupuje dříve do zaměstnání. Talent, či nepozorované schopnosti, mohou být ovšem negativně ovlivněny i možnou nižší péčí matky, která musela pracovat místo

¹⁷ Důvodem můžou být konexe, „vyšlapané cestičky“ a motivace překonat výdělek rodičů.

rozvíjení potenciálu svého dítěte¹⁸. Z dat však nelze zjistit, jestli žil respondent se svobodnou matkou i dále, či se stav změnil. Přes všechny výhrady jsme po ověření našich hypotéz v korelační matici proměnných (viz tabulka 4) tyto instrumenty použili.

Tabulka 4 - Korelační koeficienty instrumentálních proměnných

<i>wage</i>	<i>educ</i>	<i>fatheduc</i>	<i>motheduc</i>	<i>sinmom14</i>	<i>libcrd14</i>	
1	0,2764	0,1419	0,1695	-0,091	0,1682	<i>wage</i>
	1	0,434	0,4327	-0,1584	0,3433	<i>educ</i>
		1	0,6072	-0,0495	0,3268	<i>fatheduc</i>
			1	-0,1753	0,3612	<i>motheduc</i>
				1	-0,0567	<i>sinmom14</i>
					1	<i>libcrd14</i>

5% kritická hodnota (oboustranná) = 0,06 pro n = 1000, chybějící hodnoty vynechány

Rovnice po aplikaci metody nejmenších čtverců na vysvětlení *educ* pomocí instrumentálních proměnných, tedy 1 stupeň metody 2SLS¹⁹:

$$educ = 13,82 + 0,09fatheduc + 0,12motheduc \text{ (RSQ=.433)}$$

$$educ = 14,53 + 0,16fatheduc. \text{ (RSQ=.428)}$$

$$educ = 14,49 + 0,17motheduc \text{ (RSQ=.464)}$$

$$educ = 16 + 0,82libcrd14 \text{ (RSQ=.457)}$$

$$educ = 15,97 + 0,83libcrd14 - 0,47sinmom14 \text{ (RSQ=.459)}$$

$$educ = 16,64 - 0,42sinmom14 \text{ (RSQ=.441)}$$

$$educ = 13,7 + 0,09fatheduc + 0,11motheduc + 0,42libcrd14 - 0,82sinmom14 \text{ (RSQ=.442)}$$

Z těchto rovnic vyplývá vliv instrumentů na vzdělání v různých kombinacích. Zajímavé jsou koeficienty u vzdělání rodičů, z kterých například v poslední rovnici vyplývá, že zhruba s každým desátým rokem vzdělání rodičů dosáhne jejich syn jednoho roku vzdělání navíc oproti rodičům bez vzdělání za jinak nezměněných okolností. Vliv *libcrd14* a *sinmom14* je stále diskutabilní, přičemž minimálně znaménka u koeficientů vychází podle předpokladů. *Sinmom14* jako samotný instrument jsme následně vyřadily a nepokračovali v testování z důvodu nevýznamnosti v při vysvětlování *educ*. Při kombinaci s *libcrd14* již významný byl a dle výstupu odchytil část variability a zpřesnil odhad *educ*, při vysvětlování mzdy.

Základním nedostatkem modelu s instrumentálními proměnnými je nenulová korelace zvolených instrumentů se mzdou, tedy porušení základní podmínky pro použití instrumentálních proměnných [Wooldridge, 2006].

GM PŘEDPOKLADY

Porušením Gauss-Markovových předpokladů vzniká v modelu autokorelace, heteroskedasticita či multikolinearita, přičemž *autokorelace nemá v našem modelu smysl testovat*, protože data, která používáme, jsou průřezová.

MULTIKOLINEARITA

Multikolinearita se v modelu může vyskytovat, pokud je mezi minimálně dvěma proměnnými nějaká vazba. Určitá kolinearita se vyskytuje mezi všemi proměnnými, proto je nutné testovat její sílu, tedy jestli je vzájemná vazba mezi proměnnými únosná.

¹⁸ Tento efekt by způsoboval nevhodnost použití této proměnné jako instrumentu, protože se jedná o korelovanost instrumentu s náhodnou složkou. Tato možnost je potvrzena i faktem, že *libcrd14* je častější u rodičů s lepším vzděláním. Výsledky probitového modelu vysvětlujícího vliv vzdělání matky i otce na pravděpodobnost vlastnění kartičky do knihovny ukázal významnou závislost na 0.01 procentní hladině významnosti u každého koeficientu. Kartička je tedy korelovaná se vzděláním rodičů, tedy jakožto IV vlastní všechny neduhy obou.

¹⁹ Rovnice druhého stupně jsou uvedeny v tabulce 10.

Provedli jsme test multikolinearity pomocí pomocných regresí vysvětlujících proměnných a jejich vzájemných vztahů. U testu následně porovnáváme dílčí koeficienty determinace s původním v plném modelu.

Oproti původnímu R^2 (0.27) vychází vyšší dílčí R^2 u modelu s vysvětlovanou proměnnou *educ* (0.47) a *exper* (0.44). Multikolinearita je v modelu tedy neúnosná.

Důvodem této silné závislosti u proměnných *educ* a *exper* je pravděpodobně jejich vzájemná závislost. Korelační koeficient mezi vzděláním a zkušenostmi

$$\text{corr}(\text{educ}, \text{exper}) = -0.62$$

je vysoký, protože samotná proměnná *exper* je vypočítávána přímo z *educ* dle:

$$\text{exper} = \text{age} - \text{educ} - 6$$

Vztah mezi *educ*, *exper* a konstantou je tedy určován věkem, který se navíc pohybuje pouze v rozmezí od 24 do 34 let. Dalším důvodem může být špatný vzorek dat. Protože neznáme klíč, podle kterého byla data vybrána, můžeme se domnívat, že vzorek 1000 pozorování není náhodný a tento fakt způsobuje multikolinearitu.

Tato závislost může negativně prostupovat do koeficientů a zkreslovat jejich hodnoty. Model by podle nás nebylo vhodné přetransformulovat, ale použít vhodné funkční transformace, případně rozšířit vzorek dat. Vliv zkušenosti na mzdu je nezanedbatelný a v modelu důležitý, přestože jsme přesvědčeni, že jeho formulace výše uvedeným způsobem je nevhodná. Ideální by bylo nahradit *exper* pomocí zkušenosti pro daného zaměstnavatele nebo například léty relevantní pracovní zkušenosti.

HETEROSKEDASTICITA

Pokud by v modelu byla přítomna heteroskedasticita, potom rozptyl náhodné složky by nebyl shodný pro všechna pozorování, tedy měnil by se v závislosti na hodnotách vysvětlujících proměnných.

Provedli jsme optický test na grafu skutečných a fitovaných hodnot, který napovídá na potvrzení hypotézy o homoskedasticitě. Graf 1 (viz. níže) nemá viditelnou tendenci se rozšiřovat ani zužovat.

Abychom se vyhnuli možnému optickému klamu, rozhodli jsme se pro heteroskedasticitu otestovat pomocí Whitova testu, který prokázal, že heteroskedasticita v modelu není přítomna (p -value = 0.199). Metodu nejmenších čtverců, jakožto BLUE, pro odhad regresních parametrů tedy nezamítáme.

IV. KRITIKA MODELU A ALTERNATIVNÍ POSTUP

První věcí, která nás při žádosti o tento úkol zarazila, byla vlastně samotná proměnná *educ*, ke které máme hledat výnosy. Zpočátku jsme zkoušeli vytvořit transformace, které by odpovídaly rostoucímu výnosu ze vzdělání, ale pak nám to došlo. Máme sice 13 různých zástupců vzdělání, ale pořád se jedná jenom o několik málo různých stupňů vzdělání, kdybychom měli údaj o nejvyšším dokončeném vzdělání, tak bychom se dostali ke skutečné proměnné *educ*, která očividně není spojitá. Dokonce jak je vidno z následujícího postupu je přichází výnos většinou až se získáním onoho diplomu, či titulu. K tomuto bychom velmi rádi podotkli, že mít data o typech škol (státní/soukromé, college/university) apod., by dále velmi pomohlo při vysvětlování skutečného výnosu ze vzdělání.

Následující model je velmi podobný nejširšímu modelu v tabulce 10 bez instrumentálních proměnných, které jsme nechtěli zkoušet u binární vysvětlované. Koeficient u *educ6* až *educ18* vyjadřuje, o kolik více mají při dané úrovni vzdělání oproti těm, kteří nemají dokončen první stupeň základní školy (tzv. K-6, tedy *educ* < 6). Zajímavý je onen propad po dokončení základní školy, kdy dokončení střední školy je až za 4 roky. Propad může být způsoben faktem, že být vyhozen např. ve třetím roce univerzity znamená nemít diplom a ztratit nejenom 3 roky vzdělání, ale i potencionální praxe.

Absolutní a mezní procentuální přínosy v jednotlivých letech nad prvních 6 let školní docházky můžete vidět v grafu 2.

Druhou věcí je existence odlehlých pozorování, které mohou činit model prezentovaný v předchozí části náchylný chybám. Robustnější metodou nežli OLS je například kvantilová regrese [Koenker a Hallock, 2001], která namísto průměrné hodnoty koeficientu odhaduje hodnotu kvantilovou. Touto metodou námi získaný mediánový odhad vyjadřuje hranici, která odděluje přesně polovinu nižších a vyšších odhadů. Je zřejmé, že v tomto odhadu je vliv odlehlých pozorování minimální.

Zjistili jsme, že odhad metou OLS je oproti mediánovému odhadu vychýlený v *educ* o 0.004 dolu z důvodu pochybných pozorování, kdy je mzda nižší než minimální. Odhady provedené touto metodou jsou uvedeny v hranatých závorkách. Oproti odhadu z OLS je vidět logický propad v 17tém roce vzdělání, kdy se znovu jedná o rok, který nepřináší titul, ale v OLS se neprojevil.

$$\begin{aligned} \ln(wage) = & 5.024[5.013] - .182[.162]black - .146[.154]smsa - .137[.127]south + .071[.055]exper - \\ & (.149[.175]) \quad (.031[.036]) \quad (.027[.032]) \quad (.025[.03]) \quad (.013[.015]) \\ & - .002[.001]exper^2 + .348[.456]educ6^{20} + .533[.437]educ7 + .553[.555]educ8 + \\ & (<.001[<.001]) \quad (.224[.263]) \quad (.189[.222]) \quad (.156[.183]) \\ & + .406[.362]educ9 + .45[.493]educ10 + .582[.659]educ11 + .682[.746]educ12 + \\ & (.151[.178]) \quad (.146[.17]) \quad (.147[.172]) \quad (.14[.164]) \\ & + .698[.764]educ13 + .755[.88]educ14 + .771[.85]educ15 + .939[1.02]educ16 + \\ & (.144[.169]) \quad (.145[.17]) \quad (.149[.174]) \quad (.144[.17]) \\ & + .953[1.016]educ17 + 1.069[1.138]educ18 \\ & (.151[.176]) \quad (.149[.174]) \end{aligned}$$

²⁰ Jediný koeficient, který nevyšel statisticky významný v obou odhadech. Proměnná *educ7* vyšla v kvantilové regresi významná na pěti a *educ9* na deseti procentní hladině významnosti. Všechny ostatní významné na 0.1% hladině významnosti v obou regresích.

ZÁVĚR

Odhadem modelu vysvětlující mzdu pomocí vzdělání a kontrolních proměnných jsme zjistili, že dodatečný rok vzdělání přidá ke mzdě 0.65%. Faktory ovlivňující výši mzdy jsou následně barva kůže, která znevýhodňuje tmavé o 18.2% nižší mzdou; pracovní zkušenosti, které mají klesající výnosy; místo, kde respondenti žijící v centru mají o 14.7% vyšší mzdy, než na okraji a mimo město, a respondenti žijící na severu o 14.1% vyšší mzdy, než respondenti žijící na jihu.

Použitím instrumentálních proměnných jsme získali sadu odhadů parametru výnosů ze vzdělání, které se lišili od 0.38% až po 0.53%, ale u všech dostupných instrumentů pochybujeme o jejich vhodnosti z důvodů nesplnění předpokladů pro využití metody instrumentálních proměnných.

Alternativní postup, který byl nastíněn v poslední části, ukazuje na některé chyby při pokusu modelovat výnosy ze vzdělání jako spojitou veličinu s lineárním trendem. Výnosy ze vzdělání jsou sice rostoucí, ale vytváří logické vrcholy po dokončení daného stupně vzdělání. Vhodnější pro získání odhadů je navrženo využít kvantilové regrese, která je robustnější metodou v případě odlehklých pozorování, které jsou velmi častým jevem u mezd.

Pro další výzkum bychom doporučili lepší informace ke sbíraným datům, či dokonce lepší postup samotného sběru. Problémy, které s daným data setem vznikly jako (i) existence nižší mzdy než minimální; (ii) chybějící informace o nejvyšším dosaženém vzdělání; a (iii) neošetření možného vychýlení sbíraných dat respondenty; tvoří při následném ekonometrickém zpracování závažné problémy a vychýlené výsledky.

POUŽITÁ LITERATURA

BLUNDELL, Richard, DEARDEN, Lorraine, SIANESI, Barbara. Estimating the Returns to Education: Models, Methods and Results. London: London School of Economics. 2001

CARD, David. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In CHRISTOFIDES, L.N., GRANT, E.K, SWIDINSKY, R. *Aspects of Labor Market Behaviour : Essays in Honour of John Vanderkamp*. Toronto: Toronto Press, 1995.

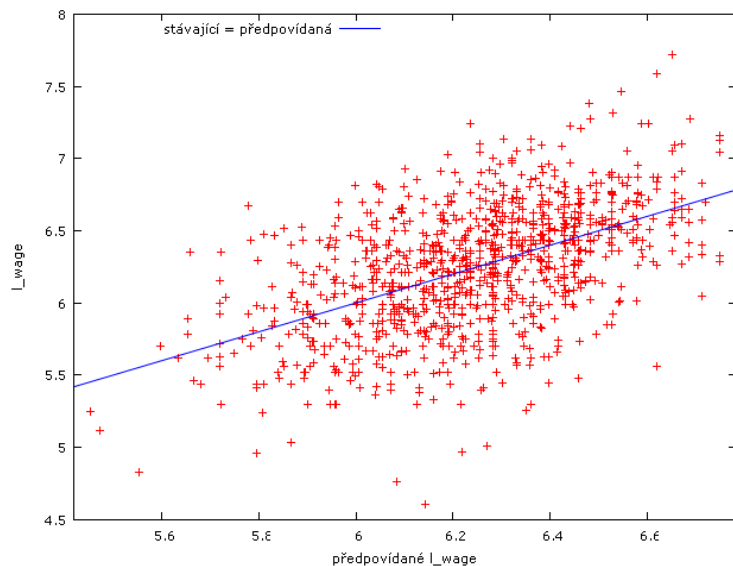
FRYER, Roland, LEVITT, Steven. Understanding the Black-White Test Score Gap in the First Two Years of School. *In The Review of Economics and Statistics*, 86(2). 2004. s. 447-464.

FRYER, Roland, LEVITT, Steven. The Black-White Test Score Gap Through Third Grade. *In American Law and Economic Review*, 8(2). 2006. s. 249-281.

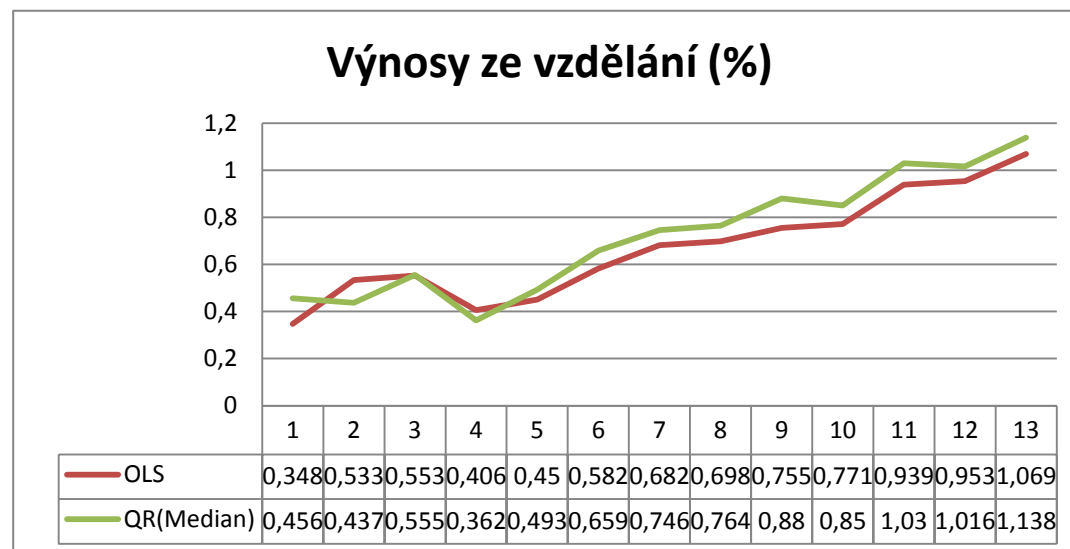
KOENKER, Roger, HALLOCK, Kevin F. Quantile Regression. *Journal of Economic Perspectives*, vol. 15 no. 4. 2001. s. 143-156

THALER, Richard H., SUNSTEIN, Cass R. Nudge : Improving decisions about health, wealth and happiness. Penguin Books, 2008. 306 s.

WOOLDRIDGE, Jeffrey M. Introductory Econometrics : A Modern Approach. 3rd International Student Edition, 2006. 890 s. ISBN 0-324-32348-4.



Graf 2 - Heteroskedasticita



Graf 1 - Výnosy ze vzdělání

Tabulka 5 - Výsledky regresí vysvětlujících mzdu pomocí let vzdělání a kontrolních proměnných s instrumenty pro vzdělání

	OLS		Med.regrese	Výsledky 2SLS s následujícími instrumenty				
				libcrd14	fatheduc (RSQ: 0.18)	motheduc (RSQ: 0.19)	fath&mothedu c (RSQ: 0.22)	all instruments
const	4.912*** (.114)	5.853*** (.071)	4.937*** (.137)	5.562*** (.213)	5.425*** (.219)	5.269*** (.192)	5.181*** (.213)	5.19*** (.21)
educ	.065*** (.006)	.028*** (.005)	.068*** (.007)	.032** (.015)	.038*** (.014)	0.5*** (.013)	.053*** (.014)	.053*** (.014)
black	-.182*** (.031)	-.196*** (.032)	-.184*** (.044)	-.238*** (.033)	-.177*** (.043)	-.236 (.036)	-.179*** (.046)	-.173*** (.046)
exper	.068*** (.012)	-	.05*** (.015)	.042*** (.012)	.047*** (.014)	.048*** (.013)	.051*** (.014)	.05*** (.014)
experSQ	-.002*** (.001)	-	-.001 (.001)	-.001** ¥**	-.001* ¥*	-.001** ¥**	-.001* ¥*	-.001* ¥*
smsa	.147*** (.027)	.145*** (.028)	.15*** (.034)	.178*** (.029)	.162*** (.033)	.167*** (.03)	.163*** (.034)	.159*** (.034)
south	-.141*** (.026)	-.149*** (.027)	-.127*** (.031)	-.154*** (.027)	-.141*** (.0313)	-.14*** (.029)	-.128*** (.032)	-.124*** (.032)
R-SQR, (n)	.274 (1000)	.192 (1000)	(1000)	.189 (998)	.152 (766)	.191 (888)	.165 (730)	.166 (730)

Pozn: Vysvětlovaná mzda je logaritmičsky transformována. Údaje jsou zaokrouhleny na tři desetinné místa. Hodnoty v závorkách jsou směrodatné odchylky; ***/**/* uaději hladinu významnosti 5/1/.01; ¥ pro data < .001.

Koeficienty v závorkách značí zápornou hodnotu. Hodnotu koeficientu determinace u 2SLS neuvádíme z důvodu ztráty vypovídací hodnoty [Wooldridge, 521]