UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



INŠTRUMENTÁLNE PREMENNÉ

DIPLOMOVÁ PRÁCA

2022 Bc. Štefánia IVANIČOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

INŠTRUMENTÁLNE PREMENNÉ

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Ján Somorčík, PhD.

Bratislava 2022 **Bc. Štefánia IVANIČOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Štefánia Ivaničová
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Inštrumentálne premenné
Instrumental variables

Anotácia: V ekonometrii, ale aj v iných oblastiach, sa často vyskytujú situácie, keď klasické modely použité na dátach naznačujú koreláciu, ktorá však nemusí v skutočnosti znamenať príčinnú (t. j. kauzálnu) závislosť. Inštrumentálne premenné sú jedným z nástrojov, ako odhaliť skutočné kauzálne závislosti. Oproti klasickým modelom je však nutné použiť zložitejšie odhadovacie postupy, napr. dvojstupňovú metódu najmenších štvorcov.

Vedúci: Mgr. Ján Somorčík, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 13.01.2021

Dátum schválenia: 17.01.2021 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie V prvom rade chcem vyjadriť obrovskú vďaku vedúcemu tejto práce Mgr. Jánovi Somorčíkovi, PhD. za jeho neustálu podporu, vecné rady, zmysluplné konzultácie a za celkovú spoluprácu. Zároveň chcem poďakovať aj svojej rodine a priateľom, že mi boli stálou oporou a podporovali ma počas celého štúdia.

Abstrakt v štátnom jazyku

IVANIČOVÁ, Štefánia: Inštrumentálne premenné [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Ján Somorčík, PhD., Bratislava, 2022, 62 s.

V tejto diplomovej práci sa venujeme metóde inštrumentálnych premenných na inferenciu o parametroch lineárnej regresie. Táto spomínaná metóda sa využíva, ak regresory korelujú s náhodnou chybou. Vlastnosti tohto odhadu analyzujeme pomocou počítačových simulácií v štatistickom softvéri R. Práca sa ďalej venuje aj testom špecifikácie, pomocou ktorých vieme overiť relevantnosť použitia inštrumentálnych premenných. Taktiež sa zaoberáme aj testovaním lineárnych hypotéz o neznámych regresných parametroch modelu. Uvádzame viacero postupov testovania a porovnávame, v čom sa dané postupy vzájomne odlišujú. Uvedené testy aplikujeme na umelo generované dáta v softvéri R a zároveň navrhujeme možné riešenia pre ich nedostatky.

Kľúčové slová: Lineárna regresia, inštrumentálne premenné, nezávislosť premenných, lineárne hypotézy.

Abstract

IVANIČOVÁ, Štefánia: Instrumental variables [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Ján Somorčík, PhD., Bratislava, 2022, 62 p.

In this thesis, we discuss the instrumental variables method for inference about linear regression parameters. This previously mentioned method is used when the regressors are correlated with the disturbances. We analyze the features of this estimation using computer simulations in the statistical software R. The work also addresses specification tests, by means of which we can verify the relevance of the use of instrumental variables. We are also focusing on testing linear hypotheses about the unknown regression parameters of the model. We present several testing procedures and compare how the procedures differ from each other. We apply the above tests to randomly generated data in R software and also suggest possible solutions for their shortcomings.

Keywords: Linear regression, instrumental variables, independence of variables, linear hypotheses.

Obsah

Úvod **8**

1 Odhad parametrov lineárnej regresie **10**

1.1 Metóda najmenších štvorcov 10

1.2 Inštrumentálne premenné 15

1.3 Vlastnosti IV odhadu 24

1.4 Smearing 28

1.5 Chybné merania 30

2 Testy špecifikácie **35**

2.1 Hausmanov test a bootstrap 35

2.2 Wuov test 38

3 Testy o parametroch regresie **41**

3.1 Lineárne hypotézy o neznámych parametroch 41

3.2 Asymptotický t-test 52

3.3 Waldov test 55

Záver **60**

Zoznam použitej literatúry **62**

Úvod

Lineárna regresia je jedným z najčastejšie používaných štatistických nástrojov v rôznych vedeckých disciplínach, lekárskom výskume, marketingu, finančníctve a mnohých ďalších oblastiach. Používa sa napríklad na vytváranie odhadov a prognóz alebo vyhodnocovanie trendov. Rovnako ako pri väčšine štatistických analýz je cieľom regresie čo najjednoduchšie interpretovať pozorované údaje. Závislú premennú, ktorá nás zaujíma, modelujeme pomocou regresorov, regresných parametrov a náhodnej chyby. K tomu je potrebné predovšetkým odhadnúť neznáme regresné parametre lineárneho modelu. Regresné parametre je potrebné odhadnúť tak, aby model čo najlepšie zodpovedal pozorovaným údajom. Za platnosti klasických predpokladov v modeli lineárnej regresie je odhad metódou najmenších štvorcov pre neznáme parametre modelu najlepším lineárnym nevychýleným odhadom.

Jedným z hlavných predpokladov je nezávislosť regresorov a náhodnej chyby. Ak je táto podmienka porušená, potom sa odhad metódou najmenších štvorcov stáva vychýleným a nekonzistentným. Tento jav môže byť spôsobený napríklad chybnými meraniami jedného alebo viacerých regresorov a ďalšími inými dôvodmi. V takomto prípade sa dá skonštruovať konzistentný odhad pomocou takzvaných inštrumentálnych premenných, ktorým sa budeme venovať v našej práci.

V prvej kapitole predstavíme základy teórie inštrumentálnych premenných, ktoré budeme využívať vo zvyšných kapitolách práce. Metóda využívajúca inštrumentálne premenné je popísaná napríklad v [3] a [2], z ktorých budeme čerpať teoretické poznatky. Simulačne budeme skúmať vlastnosti odhadu pomocou inštrumentálnych premenných a vplyv parametrov modelu na tento odhad. Zameriame sa aj na chybné merania regresorov, ktoré spôsobujú závislosť medzi regresormi a náhodnou chybou.

V druhej kapitole práce sa budeme venovať testom špecifikácie, pomocou ktorých testujeme závislosť medzi regresormi a náhodnou chybou, a teda aj relevantnosť použitia inštrumentálnych premenných. Testom špecifikácie sa venujú autori v [3] a [6]. Pomocou simulácií porovnáme správanie testov na umelo generovaných dátach.

V tretej kapitole sa budeme venovať testom o parametroch lineárnej regresie, pre ktoré autori v [1] a [2] uvádzajú rozličné testové štatistiky. Znova porovnáme fungovanie a správanie testov pomocou simulačných experimentov, v ktorých vyrátame odhad chyby prvého druhu testov.

Na analýzu metód a testovanie hypotéz použijeme známy štatistický softvéri R. Prostredie R má zabudované mnoho užitočných funkcií, ktoré vedia do veľkej miery uľahčiť výpočty a grafické znázornenie výsledkov. Zaujímavý spôsob generovania dát, ktorý v práci využijeme, popísali autori článku [4].

1 Odhad parametrov lineárnej regresie

V úvode kapitoly popíšeme metódu najmenších štvorcov, ktorý poskytuje konzistentný a nevychýlený odhad pre neznáme parametre za platnosti základných predpokladov. Dôležitým predpokladom je nezávislosť regresných premenných a náhodnej chyby. Simuláciami preukážeme, že ak je tento predpoklad porušený, dostaneme metódou najmenších štvorcov vychýlený odhad. Ďalej predstavíme metódu, ktorá rieši tento problém, a poskytne konzistentný odhad využitím inštrumentálnych premenných. Bližšie budeme skúmať vlastnosti odhadu pomocou inštrumentálnych premenných. V tejto kapitole budeme vychádzať z [2] a [3]. Teoretický výklad doplníme vlastným simulačným štúdiom.

1.1 Metóda najmenších štvorcov

Majme model lineárnej regresie

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon. \tag{1}$$

To znamená, závislá premenná y je vyjadrená ako lineárna kombinácia regresných premenných $x_1, x_2, ..., x_K$ plus náhodná chyba ε. Parametre $\beta_1, \beta_2, ..., \beta_K$ označujú neznáme parametre modelu. Hodnoty závislej premennej y pre n pozorovaní môžeme vyjadriť ako

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{iK}\beta_K + \varepsilon_i, \qquad i=1,2,...,n. \tag{2}$$

Model (2) vieme zapísať aj v zjednodušenom tvare ako

$$y = X\beta + \varepsilon, \tag{3}$$

kde

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

a regresná matica X obsahujúca prvky regresných premenných je tvaru

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & . & . & . & x_{1,K} \\ x_{2,1} & x_{2,2} & . & . & . & x_{2,K} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ x_{n,1} & x_{n,2} & . & . & . & x_{n,K} \end{pmatrix}.$$

Pod značením X_j budeme rozumieť vektor

$$X_j = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{pmatrix}, \qquad j=1,2,...,K.$$

Ďalej vymenujme základné predpoklady modelu lineárnej regresie:

P1. Plná hodnosť matice X: Regresná matica X s rozmerom $n \times K$, kde $n > K$ má lineárne nezávislé stĺpce.

P2. Homoskedasticita a neautokorelovanosť: Náhodné chyby ε_i sú nezávislé a majú rovnakú, konečnú disperziu σ^2.

P3. Normálne rozdelenie: Náhodné chyby ε_i majú normálne rozdelenie $N(0,\sigma^2)$.

P4. Exogenita regresných premenných: Regresné premenné x_j a náhodná chyba ε sú nekorelované. Tomuto javu takisto hovoríme, že regresné premenné sú exogénne.

Predpoklad P4 nám hovorí, že neexistuje žiadna korelácia medzi náhodnými chybami a regresormi. Platí teda

$$cov(x_j, \varepsilon) = E(x_j.\varepsilon) - E(x_j).E(\varepsilon) = E(x_j.\varepsilon) = 0. \tag{4}$$

Za platnosti predpokladov P1−P4 môžeme aplikovať odhad metódou najmenších štvorcov a tak získať nevychýlený odhad pre neznáme parametre $\beta_1, ..., \beta_K$.

Veta 1.1. *Nech sú splnené predpoklady P1−P4. Metódou najmenších štvorcov dostaneme nasledovný odhad*

$$b_{LS} = (X^T X)^{-1} X^T y,$$

ktorý je nevychýleným odhadom pre β.

Dôkaz vety 1.1 možno nájsť napr. v [3, str. 55].

V nasledujúcich simuláciách budeme voliť $\sigma^2 = 1$ a počet simulácií $S = 10000$. Pokiaľ neuvedieme inak, predpokladajme, že $n = 50$ a $K = 3$ a pevne stanovme vektor β ako

$$\beta = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix},$$

ktorý sa budeme snažiť odhadnúť. V simuláciách budeme dáta umelo generovať v štatistickom softvéri R. Bude nás zaujímať odhad b_{LS}.

Simulácia 1.1. Predpokladajme, že matica X je pevne daná pre všetkých S simulácií taká, že

- X_1 je jednotkový vektor,
- X_2 je vektor vygenerovaných n hodnôt z rovnomerného delenia na intervale $[-10,10]$,
- X_3 je vektor vygenerovaných n hodnôt z rovnomerného delenia na intervale $[-2,2]$.

Simulačne preukážeme nevychýlenosť odhadu b_{LS} pre β pomocou metódy najmenších štvorcov. Keďže matica X je pevne daná, zostáva rovnaká v každej simulácií a predpoklad P4 je splnený. V nasledujúcich S simuláciách budeme náhodne generovať vektor ε tak, že $\varepsilon_1, ..., \varepsilon_n \sim N(0,1)$ sú nezávislé a rovnako rozdelené. Pomocou vektora ε vypočítame vektor y s využitím znalosti skutočného vektora β. V každej zo simulácií vypočítame odhad b_{LS} a na záver vypočítame priemer $S = 10000$ odhadov b_{LS} z jednotlivých simulácií.
Výsledok už s presnosťou na 2 desatinné miesta

$$\overline{b_{LS}} = \begin{pmatrix} 1{,}00 \\ 2{,}00 \\ 3{,}00 \end{pmatrix}$$

je v súlade s našim očakávaním, keďže b_{LS} je za daných predpokladov nevychýlený.

Simulácia 1.2. Na rozdiel od predchádzajúcej simulácie predpokladajme, že matica X je v každej z $S = 10000$ simulácií náhodná a jej stĺpce obsahujú realizácie z normálneho rozdelenia. Presnejšie

- X_1 je náhodný výber z $N(3,1)$,

- X_2 je náhodný výber z $N(2,1)$,
- X_3 je náhodný výber z $N(0,1)$

a stĺpce matice X sú nezávislé. Ďalej predpokladajme, že je splnená podmienka exogénnosti, teda regresory nie sú korelované s náhodnou chybou. Maticu X budeme v každom kole simulácie náhodne generovať nezávisle od vektora ε, aby bola splnená podmienka exogénnosti. Získaný priemer z S hodnôt b_{LS} s presnosťou na 2 desatinné miesta

$$\overline{b_{LS}} = \begin{pmatrix} 1{,}00 \\ 2{,}00 \\ 3{,}00 \end{pmatrix}$$

opäť potvrdzuje nevychýlenosť b_{LS}.

Simulácia 1.3. V porovnaní s predchádzajúcou simuláciou zmeňme predpoklad o nekorelovanosti medzi náhodnou chybou a regresormi. Teraz predpokladajme, že regresné premenné sú endogénne, teda korelujú s náhodnými chybami. Nech platí, že

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \varepsilon \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 3 \\ 2 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}2 & 0{,}2 & 0{,}4 \\ 0{,}2 & 1 & 0{,}2 & 0{,}4 \\ 0{,}2 & 0{,}2 & 1 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 1 \end{pmatrix} \right) \tag{5}$$

Dáta budeme generovať ako náhodný výber veľkosti n z viacrozmerného normálneho rozdelenia (5). V tomto prípade dostaneme

$$\overline{b_{LS}} = \begin{pmatrix} 0{,}97 \\ 2{,}11 \\ 3{,}39 \end{pmatrix},$$

čo naznačuje vychýlenosť b_{LS} ako odhadu pre β. Ak nie je splnená podmienka exogénnosti, neznáme parametre β je potrebné odhadnúť iným spôsobom.

Ak uvažujeme model, v ktorom regresory korelujú s náhodnou chybou, odhad b_{LS} pomocou metódy najmenších štvorcov už nie je nevychýlený. Ako je to uvedené aj v [3], odhad je taktiež nekonzistentný a platí nasledovné

$$\begin{aligned} \lim_{n\to\infty} b_{LS} &= \lim_{n\to\infty} \left(X^T X\right)^{-1} X^T y = \lim_{n\to\infty} \left(X^T X\right)^{-1} X^T \left(X\beta + \varepsilon\right) \\ &= \beta + \lim_{n\to\infty} \left(\frac{X^T X}{n}\right)^{-1} \lim_{n\to\infty} \left(\frac{X^T \varepsilon}{n}\right) = \beta + Q_{xx}^{-1}\gamma \neq \beta, \end{aligned} \tag{6}$$

kde matica

$$Q_{xx} = E\left(\begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} . \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}^T\right) \tag{7}$$

a vektor γ obsahuje kovariancie medzi regresnými premennými $x_1, ..., x_K$ a ε, čiže

$$\gamma = \begin{pmatrix} cov(x_1, \varepsilon) \\ \vdots \\ cov(x_K, \varepsilon) \end{pmatrix}.$$

Poznamenajme, že limita, ktorá vystupuje aj v rovnici (6) sa v celej práci chápe ako limita podľa pravdepodobnosti.

Simulácia 1.4. V jednotlivých simuláciách budeme dáta generovať ako náhodný výber z rozdelenia (5). Skúmali sme limitnú vlastnosť výchylky odhadu metódou najmenších štvorcov (6) v konečnej vzorke dát. V našom prípade je matica

$$Q_{xx} = Var\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) + E\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right) E\left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}\right)^T$$

$$= \begin{pmatrix} 1 & 0{,}2 & 0{,}2 \\ 0{,}2 & 1 & 0{,}2 \\ 0{,}2 & 0{,}2 & 1 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} . \begin{pmatrix} 3 & 2 & 0 \end{pmatrix}$$

a vektor

$$\gamma = \begin{pmatrix} 0{,}4 \\ 0{,}4 \\ 0{,}4 \end{pmatrix}.$$

Výsledky simulácií potvrdili, že odhadnutá výchylka odhadu b_{LS} je už s presnosťou na 2 desatinné miesta rovná výrazu $Q_{xx}^{-1}\gamma$:

$$\widehat{bias}(b_{LS}) = \begin{pmatrix} -0{,}03 \\ 0{,}11 \\ 0{,}39 \end{pmatrix} \doteq Q_{xx}^{-1}\gamma.$$

V simulácii 1.3, kedy nebola splnená podmienka (4), sme dostali vychýlený odhad metódou najmenších štvorcov. Riešením je prístup pomocou inštrumentálnych premenných.

1.2 Inštrumentálne premenné

Konzistentný odhad pre model obsahujúci endogénne regresné premenné možno získať použitím inštrumentálnych premenných. Metóda inštrumentálnych premenných doplní regresné premenné novými premennými, ktoré nekorelujú s náhodnou chybou a zároveň nesú dostatočné množstvo informácií vyjadrenej pôvodnými regresormi, čiže mali by s nimi byť v ideálnom prípade čo najviac korelované. Zavedieme dva predpisy pre výpočet odhadu pomocou inštrumentálnych premenných a to vzhľadom na počet inštrumentálnych premenných. Naznačíme odvodenie týchto vzťahov, pričom pri odvodení budeme vychádzať z [3]. Podrobne sa budeme venovať konzistentnosti odhadu. V závere kapitoly vysvetlíme súvis odhadu metódou najmenších štvorcov s odhadom pomocou inštrumentálnych premenných.

Teraz predpokladajme, že existuje ďalšia sada L premenných, $z_1, z_2, ..., z_L$, ktoré majú nasledujúce dve vlastnosti:

- sú nekorelované s náhodnou chybou ε a
- sú korelované s regresnými premennými $x_1, x_2, ..., x_K$.

Matica inštrumentálnych premenných Z pre n pozorovaní má tvar

$$Z = \begin{pmatrix} z_{1,1} & z_{1,2} & . & . & . & z_{1,L} \\ z_{2,1} & z_{2,2} & . & . & . & z_{2,L} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ z_{n,1} & z_{n,2} & . & . & . & z_{n,L} \end{pmatrix}.$$

V predchádzajúcej kapitole sme definovali maticu Q_{xx} v (7) a ďalej definujme matice

$$Q_{zz} = E\left(\begin{pmatrix} z_1 \\ \vdots \\ z_L \end{pmatrix} . \begin{pmatrix} z_1 \\ \vdots \\ z_L \end{pmatrix}^T\right), \qquad Q_{xz} = E\left(\begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} . \begin{pmatrix} z_1 \\ \vdots \\ z_L \end{pmatrix}^T\right). \tag{8}$$

Zmeňme základné predpoklady modelu lineárnej regresie. Predpoklad P4 nahradíme predpokladom P4′ a pridáme ďalšie predpoklady týkajúci sa inštrumentálnych premenných:

P4'. Endogenita regresných premenných: Aspoň jedna z regresných premenných x_j koreluje s náhodnou chybou ε. Regresné premenné, ktoré korelujú s náhodnou chybou sa nazývajú endogénne premenné.

P5. Plná hodnosť matice Z: Matica Z s rozmerom $n \times L$, kde $n > L$ má lineárne nezávislé stĺpce.

P6. Exogenita inštrumentálnych premenných: Inštrumentálne premenné z_j a náhodná chyba ε sú nekorelované.

P7. Relevantnosť regresných a inštrumentálnych premenných: Inštrumentálne premenné z_j korelujú s regresnými premennými x_k.

P8 Matice Q_{xx}, Q_{zz} a Q_{xz} sú matice s konečnými prvkami.

P9. Matica Q_{xz} má lineárne nezávislé riadky, to znamená, že jej hodnosť je rovná počtu regresných premenných K.

Teraz z predpokladu P4′ vyplýva, že

$$cov(x_j, \varepsilon) = E(x_j.\varepsilon) - E(x_j).E(\varepsilon) = E(x_j.\varepsilon) = \gamma_j \tag{9}$$

a z predpokladu P6 vyplýva, že

$$cov(z_j, \varepsilon) = E(z_j.\varepsilon) - E(z_j).E(\varepsilon) = E(z_j.\varepsilon) = 0. \tag{10}$$

Najprv sa zaoberajme prípadom, keď $L = K$, to znamená, že existuje rovnaký počet inštrumentálnych premenných ako regresných premenných v modeli. Prenásobením rovnosti (3) maticou Z^T dostaneme

$$Z^T y = Z^T X\beta + Z^T \varepsilon$$

a z toho

$$Z^T \varepsilon = Z^T y - Z^T X\beta.$$

Keďže sú inštrumentálne premenné nekorelované s náhodnou chybou, a teda $E(z_j.\varepsilon) = 0$ a zároveň všetky premenné majú konečnú disperziu, vyplýva nasledovné

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \lim_{n\to\infty} \frac{Z^T\varepsilon}{n} = \lim_{n\to\infty} \frac{Z^T y}{n} - \lim_{n\to\infty} \frac{Z^T X\beta}{n}.$$

Z toho

$$\left(\lim_{n\to\infty}\frac{Z^TX}{n}\right)\beta=\lim_{n\to\infty}\frac{Z^Ty}{n}.$$

Keďže predpokladáme, že $L = K$, matica Z^TX je štvorcová a navyše ak je splnený predpoklad P9, potom platí nasledovné

$$\beta=\left[\lim_{n\to\infty}\frac{Z^TX}{n}\right]^{-1}\lim_{n\to\infty}\frac{Z^Ty}{n}. \tag{11}$$

Na základe vzťahu (11) je určený predpis odhadu pomocou inštrumentálnych premenných.

Definícia 1.2. *Nech $K = L$. Potom odhad pomocou inštrumentálnych premenných pre β je definovaný ako*

$$b_{IV}=(Z^TX)^{-1}Z^Ty. \tag{12}$$

Simulácia 1.5. Nech $L = K = 3$. V každej simulácií budeme dáta umelo generovať v štatistickom softvéri R ako náhodný výber z rozdelenia

$$\begin{pmatrix} x_1\\ x_2\\ x_3\\ z_1\\ z_2\\ z_3\\ \varepsilon \end{pmatrix} \sim N_7\left(\begin{pmatrix} 3\\ 2\\ 0\\ 0\\ 0\\ 0\\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}2 & 0{,}2 & 0{,}4 & 0{,}2 & 0{,}15 & 0{,}4\\ 0{,}2 & 1 & 0{,}2 & 0{,}2 & 0{,}3 & 0{,}1 & 0{,}4\\ 0{,}2 & 0{,}2 & 1 & 0{,}15 & 0{,}25 & 0{,}35 & 0{,}4\\ 0{,}4 & 0{,}2 & 0{,}15 & 1 & 0{,}3 & 0{,}3 & 0\\ 0{,}2 & 0{,}3 & 0{,}25 & 0{,}3 & 1 & 0{,}3 & 0\\ 0{,}15 & 0{,}1 & 0{,}35 & 0{,}3 & 0{,}3 & 1 & 0\\ 0{,}4 & 0{,}4 & 0{,}4 & 0 & 0 & 0 & 1 \end{pmatrix}\right) \tag{13}$$

Takto definovaná kovariančná matica zohľadní požadované vlastnosti inštrumentálnych premenných. Videli sme, že odhad metódou najmenších štvorcov je vychýlený. Ak teraz použijeme prístup pomocou inštrumentálnych premenných, dostaneme priemer $S = 10000$ odhadov b_{IV}:

$$\overline{b_{IV}}=\begin{pmatrix} 0{,}69\\ 2{,}59\\ 2{,}48 \end{pmatrix},$$

ktorý naznačuje vychýlenosť odhadu b_{IV}. Problém spôsobuje počet inštrumentálnych premenných. Ak $L = K$, potom priemer odhadov b_{IV} nekonverguje ku $E(b_{IV})$ napriek tomu,

že odhad je konzistentný. Konzistentnosť odhadu znamená, že odhad b_{IV} konverguje s rastúcim počtom pozorovaní k β. Podrobnejšie sa tejto vlastnosti odhadu budeme venovať na str. 18. Ukážeme, že predpoklad P9 musí byť splnený, aby sme dostali konzistentný odhad.

Ak máme v modeli rovnaký počet inštrumentálnych premenných v Z ako regresných premenných v X, potom priemer odhadov pomocou inštrumentálnych premenných získaných z nezávislých simulácií nekonverguje s rastúcim počtom simulácií. Ako sa aj v [2] spomína, dôvodom je, že $m-$tý moment odhadu pomocou inštrumentálnych premenných existuje iba, ak $m < L - K + 1$. Túto podmienku môžeme formálne zapísať ako

$$\exists \quad E(b_{IV}^m) \iff m < L - K + 1. \tag{14}$$

To znamená, že ak $L = K$, potom neexistuje $E(b_{IV})$, a teda $\overline{b_{IV}}$ nekonverguje s rastúcim počtom simulácií. S jednou ďalšou inštrumentálnou premennou v modeli bude existovať $E(b_{IV})$, ale nebude existovať konečná disperzia odhadu, atď. Tento výsledok je veľmi dôležitý, napriek tomu W. H. Greene v knihe [2] sa týmto problémom vôbec nezaoberá.

Simulácia 1.6. Pre ukážku, opakovane vypočítame priemer z b_{IV} získaných zo simulácií popísaných v simulácii 1.5 s rôznymi iniciálnymi nastaveniami náhodného generátora. Dostávame nasledovné výsledky

$$\overline{b_{IV}} = \begin{pmatrix} -3{,}54 \\ 3{,}26 \\ 4{,}29 \end{pmatrix}, \quad \overline{b_{IV}} = \begin{pmatrix} 7{,}48 \\ 31{,}09 \\ -48{,}01 \end{pmatrix}, \quad \overline{b_{IV}} = \begin{pmatrix} -15{,}17 \\ 23{,}16 \\ 8{,}36 \end{pmatrix}.$$

Je to v súlade so spomínaným nekonvergovaním $\overline{b_{IV}}$. Inak povedané, nemožno sa spoľahnúť, že ak je počet simulácií S veľké, tak $\overline{b_{IV}}$ sa stabilizuje okolo nejakej hodnoty.

Teraz sa bližšie pozrieme na konzistentnosť odhadu b_{IV}, keď $L = K$. Napriek tomu, že $E(b_{IV})$ neexistuje, ak máme k dispozícii rozsiahly počet pozorovaní, môžeme dostať konzistentný odhad. Pri odvodení rovnice (11) sme využili predpoklad P9, aby sme dostali konzistentný odhad. Ak je táto podmienka porušená, aj napriek veľkému počtu pozorovaní dostaneme odhad b_{IV}, ktorý je vychýlený od vektora β.

Analógiu toho, čo sa deje pri výpočte odhadu b_{IV} vysvetlíme najprv na príklade s maticou rozmeru 1×1. Zoberme $a = 3$, potom jej inverzná hodnota je $a^{-1} \approx 0{,}3333$.

Ak zmeníme hodnou a o jednu desatinu, napríklad na $a^* = 2{,}9$ alebo $a^* = 3{,}1$, potom jej inverzná hodnota $a^{*-1} \approx 0{,}3448$, resp. $a^{*-1} \approx 0{,}3226$ je rádovo rovnaká. To znamená, malá zmena hodnoty a sa prejavila malou zmenou jej inverznej hodnoty. Teraz, ak zvolíme $a^* = 0{,}1$ blízke singulárnej hodnote $a = 0$, potom $a^{*-1} = 10$. Urobíme rovnakú zmenu, čiže zmeníme hodnotu $a = 0$ o jednu desatinu na $a^* = -0{,}1$, a dostaneme inverznú hodnotu $a^{*-1} = -10$. Samotná hodnota $a = 0$ sa zmenila len o jednu desatinu ale jej inverzná hodnota sa zmenila výrazne. Dôvodom je, že pre a blízke singulárnej hodnote je inverzia numericky nestabilná.

Tento istý jav pozorujeme aj pre maticu $Z^T X$ pri výpočte odhadu b_{IV}, ktorý možno zapísať v tvare

$$b_{IV} = \left(\frac{Z^T X}{n}\right)^{-1} . \frac{Z^T y}{n}.$$

Samotná matica $(Z^T X)/n$ konverguje k limitnej hodnote

$$Q_{xz} = \lim_{n\to\infty} \frac{Z^T X}{n}.$$

Problém nastáva, ak matica Q_{xz} je singulárna, a teda inverzia matice $(Z^T X)/n$ je numericky nestabilná. Pri výpočte odhadu používame maticu $(Z^T X)/n$, ktorá v závislosti od dát sa bude mierne líšiť. Inverzia matice $(Z^T X)/n$ sa bude potom výrazne líšiť už pri malej zmene hodnôt matice $Z^T X$. To znamená, že ak matica Q_{xz} nie je regulárna, potom konzistentnosť odhadu nie je zabezpečená. Numericky si tento problém nemusíme všimnúť, keďže pre konkrétne dáta je matica $Z^T X$ síce blízka singulárnej matici ale regulárna.

Simulácia 1.7. Uvažujme rovnaké generovanie dát, ako v simulácii 1.5. To znamená, že

$$Q_{xz} = \begin{pmatrix} 0{,}4 & 0{,}2 & 0{,}15 \\ 0{,}2 & 0{,}3 & 0{,}1 \\ 0{,}15 & 0{,}25 & 0{,}35 \end{pmatrix}.$$

Videli sme v simulácii 1.6, že pre $n = 50$ bol priemer z $S = 10000$ odhadov b_{IV} oscilujúci, keďže stredná hodnota odhadu neexistuje. Kovariančná matica Q_{xz} je regulárna, a teda predpoklad P9 je splnený. To znamená, že so zvyšujúcim sa počtom pozorovaní n by mal odhad b_{IV} konvergovať ku skutočnej hodnote β. Porovnajme hodnoty odhadu pre

narastajúci počet pozorovaní n s presnosťou na 2 desatinné miesta. Dostávame výsledky

$$n = 50: \quad b_{IV} = \begin{pmatrix} -0{,}53 \\ 6{,}19 \\ 1{,}87 \end{pmatrix}, \qquad n = 100: \quad b_{IV} = \begin{pmatrix} 0{,}77 \\ 2{,}19 \\ 2{,}53 \end{pmatrix},$$

$$n = 500: \quad b_{IV} = \begin{pmatrix} 0{,}77 \\ 2{,}28 \\ 2{,}82 \end{pmatrix}, \qquad n = 5000: \quad b_{IV} = \begin{pmatrix} 0{,}97 \\ 2{,}03 \\ 3{,}02 \end{pmatrix}.$$

Vidíme, že hodnota odhadu sa s narastajúcim n stabilizuje okolo hodnoty β, čo naznačuje konzistentnosť odhadu b_{IV}. Teraz porušme predpoklad P9 a dáta generujme ako náhodný výber z rozdelenia (13) ale so singulárnou maticou

$$Q_{xz} = \begin{pmatrix} 0{,}25 & 0{,}25 & 0{,}25 \\ 0{,}25 & 0{,}25 & 0{,}25 \\ 0{,}25 & 0{,}25 & 0{,}25 \end{pmatrix}.$$

Náhodne vygenerujme dve rôzne sady dát s počtom pozorovaní $n = 5000$. Výsledky výpočtov uvedieme s presnosťou na 2 desatinné miesta. Ak pre obe sady dát vypočítame maticu $(Z^T X)/n$ dostaneme podobný výsledok blízky matici Q_{xz}

$$\frac{Z^T X}{n} = \begin{pmatrix} 0{,}25 & 0{,}24 & 0{,}25 \\ 0{,}27 & 0{,}26 & 0{,}26 \\ 0{,}23 & 0{,}24 & 0{,}24 \end{pmatrix}, \qquad \frac{Z^T X}{n} = \begin{pmatrix} 0{,}22 & 0{,}22 & 0{,}23 \\ 0{,}22 & 0{,}22 & 0{,}25 \\ 0{,}25 & 0{,}24 & 0{,}26 \end{pmatrix}.$$

Teraz, ak vypočítame inverziu matíc dostaneme výrazne rozdielne výsledky

$$\left(\frac{Z^T X}{n}\right)^{-1} = \begin{pmatrix} 63{,}67 & 19{,}34 & -87{,}80 \\ -91{,}89 & 48{,}23 & 45{,}30 \\ 28{,}82 & -65{,}71 & 44{,}05 \end{pmatrix}$$

$$\left(\frac{Z^T X}{n}\right)^{-1} = \begin{pmatrix} -107{,}42 & -109{,}28 & 199{,}58 \\ 120{,}47 & 27{,}68 & -133{,}61 \\ -7{,}29 & 78{,}23 & -64{,}10 \end{pmatrix}.$$

Pre maticu $(Z^T y)/n$ dostaneme podobné výsledky

$$\left(\frac{Z^T y}{n}\right) = \begin{pmatrix} 1{,}51 \\ 1{,}56 \\ 1{,}43 \end{pmatrix}, \qquad \left(\frac{Z^T y}{n}\right) = \begin{pmatrix} 1{,}36 \\ 1{,}42 \\ 1{,}52 \end{pmatrix}.$$

Na záver, samotné odhady sú rovné

$$b_{IV} = \begin{pmatrix} 1{,}09 \\ 1{,}12 \\ 3{,}80 \end{pmatrix}, \qquad b_{IV} = \begin{pmatrix} 1{,}86 \\ 0{,}38 \\ 3{,}67 \end{pmatrix}.$$

Vidíme, že sa odhady pre rôzne sady dát líšia a sú vychýlené od β. Výsledky potvrdili, že ak predpoklad P9 nie je splnený, potom odhad b_{IV} nie je konzistentný.

J. DiNardo a J. Johnston v [2] majú predpoklad P9 zaradený medzi predpokladmi modelu. W. Greene v [3] túto podmienku síce spomína pri odvodení odhadu b_{IV} ale vo zvyšnom teoretickom výklade ju nezdôrazňuje, napriek tomu, že je to kľúčový predpoklad pre konzistentnosť odhadu. Predpoklad P9 nevyplýva z predpokladov P1−P8 preto je potrebné ju uviesť medzi predpokladmi pre odhad pomocou inštrumentálnych premenných.

Doteraz sme predpokladali, že počet inštrumentálnych premenných je rovnaký ako počet regresných premenných. Teraz sa budeme zaoberať prípadom, kedy $L > K$. Ak matica Z obsahuje viac premenných ako matica X, nemôžeme použiť odhad (12), pretože matica $Z^T X$ s rozmerom $L \times K$ sa nedá invertovať. V prípade, kedy $L > K$ sa používa dvojstupňová metóda najmenších štvorcov. Už zo samotného názvu odhadu vyplýva, že na výpočet odhadu sa dvakrát aplikuje metóda najmenších štvorcov.

V prvom kroku sa počíta matica

$$\widehat{X} = Z(Z^T Z)^{-1} Z^T X = P_Z X, \tag{15}$$

kde matica P_Z je projektor na stĺpcový priestor matice Z. V druhom kroku sa vypočíta odhad pre β ako

$$b_{2LS} = \left(\widehat{X}^T \widehat{X}\right)^{-1} \widehat{X}^T y. \tag{16}$$

Ako to aj neskôr uvedieme, regresné premenné v matici X, ktoré nie sú korelované s náhodnou chybou sa v odhade použijú aj ako inštrumentálne premenné v matici Z.

Poznamenajme, že odhad pomocou inštrumentálnych premenných v modeli

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

je prakticky odhad metódou najmenších štvorcov v modeli

$$y = \widehat{X}\beta + \varepsilon.$$

V nasledujúcej vete ukážeme odvodenie ekvivalentného predpisu pre odhad dvojstupňovou metódou najmenších štvorcov.

Veta 1.3. *Pre odhad b_{2LS} platí*

$$b_{2LS} = \left(\widehat{X}^T X\right)^{-1} \widehat{X}^T y. \tag{17}$$

Dôkaz. Z vlastností projekčných matíc platia nasledovné vzťahy:

$$\begin{aligned} b_{2LS} &= \left(\widehat{X}^T \widehat{X}\right)^{-1} \widehat{X}^T y = \left((P_Z X)^T P_Z X\right)^{-1} \widehat{X}^T y = \left(X^T {P_Z}^T P_Z X\right)^{-1} \widehat{X}^T y \\ &= \left(X^T {P_Z}^T X\right)^{-1} \widehat{X}^T y = \left(\widehat{X}^T X\right)^{-1} \widehat{X}^T y. \end{aligned}$$

Využili sme idempotentnú a symetrickú vlastnosť projekčnej matice P_Z. □

Simulácia 1.8. Teraz predpokladajme, že $K = 3 < L = 4$. Dáta budeme teraz generovať ako náhodný výber z viacrozmerného normálneho rozdelenia

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ z_1 \\ z_2 \\ z_3 \\ z_4 \\ \varepsilon \end{pmatrix} \sim N_8 \left(\begin{pmatrix} 3 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}2 & 0{,}2 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}2 & 1 & 0{,}2 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}2 & 0{,}2 & 1 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}25 & 0{,}25 & 0{,}25 & 1 & 0{,}3 & 0{,}3 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 1 & 0{,}3 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 0{,}3 & 1 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 0{,}3 & 0{,}3 & 1 & 0 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right) \tag{18}$$

Na takto generované dáta sme aplikovali dvojstupňovú metódu najmenších štvorcov. Priemer z odhadov b_{2LS} získaných dvojstupňovou metódou najmenších štvorcov s presnosťou na 2 desatinné miesta

$$\overline{b_{2LS}} = \begin{pmatrix} 0{,}98 \\ 2{,}03 \\ 3{,}13 \end{pmatrix}$$

naznačuje v porovnaní s priemerom odhadu b_{LS} v simulácii 1.5 relatívne menšiu výchylku b_{2LS} ako odhadu pre β.

Všimli sme si, že napriek tomu, že podmienka P9 je v simulácii 1.8 porušená, výsledky naznačujú, že odhad b_{2LS} je nevychýleným odhadom pre β. V zdrojoch, ktoré sme študovali, sme o tom nenašli žiadne zmienky a vzhľadom na rozsah práce nechávame tento problém otvorený.

Na záver ukážeme za akých podmienok nastane ekvivalencia medzi odhadom dvojstupňovou metódou najmenších štvorcov b_{2LS} a odhadom metódou najmenších štvorcov b_{LS}. Taktiež dokážeme, že odhad b_{IV} daný vzťahom (12) je len špeciálnym prípadom odhadu dvojstupňovou metódou najmenších štvorcov b_{2LS}. Tieto vzťahy sú uvedené v [2], pričom dopĺňame vlastné dôkazy.

Veta 1.4. *Nech $L = n$. Potom odhad dvojstupňovou metódou najmenších štvorcov b_{2LS} je rovný odhadu metódou najmenších štvorcov b_{LS}, ktorý je vychýlený.*

Dôkaz. Podľa vety 1.3 platí, že

$$b_{2LS} = (\widehat{X}^T X)^{-1} \widehat{X}^T y.$$

Potom dostávame:

$$\begin{aligned} b_{2LS} = (\widehat{X}^T X)^{-1} \widehat{X}^T y = [X^T Z (Z^T Z)^{-1} Z^T X]^{-1} X^T Z (Z^T Z)^{-1} Z^T y = \\ = [X^T Z Z^{-1} (Z^T)^{-1} Z^T X]^{-1} X^T Z Z^{-1} (Z^T)^{-1} Z^T y = (X^T X)^{-1} X^T y = b_{LS}. \end{aligned}$$

Využili sme, že v takomto prípade je matica Z štvorcová, navyše regulárna matica. □

Veta 1.5. *Nech $L = K$. Potom je odhad pomocou inštrumentálnych premenných b_{IV} ekvivalentný odhadu dvojstupňovou metódou najmenších štvorcov b_{2LS}. To znamená, že odhad b_{IV} je len špeciálnym prípadom odhadu b_{2LS} pre $L = K$.*

Dôkaz. Podľa vety 1.3 platí, že

$$b_{2LS} = (\widehat{X}^T X)^{-1} \widehat{X}^T y.$$

Potom dostávame:

$$\begin{aligned} b_{2LS} = (\widehat{X}^T X)^{-1} \widehat{X}^T y = [X^T Z (Z^T Z)^{-1} Z^T X]^{-1} X^T Z (Z^T Z)^{-1} Z^T y = \\ = (Z^T X)^{-1} [X^T Z (Z^T Z)^{-1}]^{-1} X^T Z (Z^T Z)^{-1} Z^T y = (Z^T X)^{-1} Z^T y = b_{IV}. \end{aligned}$$

Využili sme, že v takomto prípade sú matice $X^T Z$ a $Z^T X$ štvorcové, regulárne matice. □

Odteraz budeme odhad pomocou inštrumentálnych premenných pre $L \geq K$ označovať všeobecne ako b_{IV} a za všeobecný predpis odhadu budeme považovať vzorec (16).

1.3 Vlastnosti IV odhadu

Nasleduje niekoľko simulácií, ktorými budeme skúmať vplyv jednotlivých parametrov modelu na presnosť odhadu pomocou inštrumentálnych premenných. Vysvetlíme, ako súvisia s teóriou spracovanou v [2] a [3]. Navyše sa budeme venovať aj vplyvu korelačných koeficientov medzi regresnými a inštrumentálnymi premennými. V závere podkapitoly predstavíme model, ktorý obsahuje exogénne aj endogénne regresné premenné zároveň a vysvetlíme ako to ovplyvní odhad b_{LS} pre β.

Ako prvé nás bude zaujímať vplyv korelačných koeficientov medzi regresnými premennými v X a inštrumentálnymi premennými v Z na odhad pomocou inštrumentálnych premenných b_{IV}. Intuitívne očakávame, že ak sa zvyšuje väzba medzi premennými v X a premennými v Z, potom sa nám podarí dosiahnuť presnejší odhad.

Simulácia 1.9. Naďalej uvažujme počet pozorovaní $n = 50$. Rovnako ako v predošlých simuláciách zostáva počet regresných premenných $K = 3$ a zvoľme počet inštrumentálnych premenných $L = 4$. Nech je teraz kovariančná matica vektora
$(x_1, x_2, x_3, z_1, z_2, z_3, z_4, \varepsilon)$ definovaná ako

$$\begin{pmatrix} 1 & 0{,}5 & 0{,}5 & & & & & 0{,}4 \\ 0{,}5 & 1 & 0{,}5 & & & cor_{xz} & & 0{,}4 \\ 0{,}5 & 0{,}5 & 1 & & & & & 0{,}4 \\ & & & 1 & 0{,}5 & 0{,}5 & 0{,}5 & 0 \\ & & & 0{,}5 & 1 & 0{,}5 & 0{,}5 & 0 \\ & cor_{xz} & & 0{,}5 & 0{,}5 & 1 & 0{,}5 & 0 \\ & & & 0{,}5 & 0{,}5 & 0{,}5 & 1 & 0 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Postupne budeme zvyšovať korelačné koeficienty medzi regresnými a inštrumentálnymi premennými a sledovať, ako sa mení presnosť odhadu b_{IV}. Výsledky uvedieme s presnosťou na 2 desatinné miesta. Pre jednoduchosť sme zvolili všetky korelačné koeficienty medzi regresnými a inštrumentálnymi premennými rovnaké. Pre maticu

$$cor_{xz} = \begin{pmatrix} \rho_{xz} & \rho_{xz} & \rho_{xz} \\ \rho_{xz} & \rho_{xz} & \rho_{xz} \\ \rho_{xz} & \rho_{xz} & \rho_{xz} \\ \rho_{xz} & \rho_{xz} & \rho_{xz} \end{pmatrix}$$

najprv zvoľme $\rho_{xz} = 0{,}1$, potom je priemer odhadov pomocou inštrumentálnych premenných

$$\overline{b_{IV}} = \begin{pmatrix} 0{,}97 \\ 2{,}08 \\ 3{,}31 \end{pmatrix}.$$

Teraz zvýšime korelačné koeficienty na $\rho_{xz} = 0{,}3$, a dostávame nový priemer odhadov

$$\overline{b_{IV}} = \begin{pmatrix} 0{,}99 \\ 2{,}03 \\ 3{,}09 \end{pmatrix},$$

ktorý je v porovnaní s predchádzajúcim priemerom odhadov menej vychýlený od skutočného vektora β. Nakoniec zvolíme korelačné koeficienty $\rho_{xz} = 0{,}5$, pre ktoré je priemer odhadov

$$\overline{b_{IV}} = \begin{pmatrix} 1{,}00 \\ 2{,}00 \\ 3{,}05 \end{pmatrix}.$$

Vidíme, že posledný z priemerov odhadov je najpresnejší. Výsledky teda potvrdili kladný vzťah medzi zvyšujúcimi sa korelačnými koeficientami medzi regresnými a inštrumentálnymi premennými a presnosťou odhadu.

Ďalej nás bude zaujímať vplyv parametra n, teda vplyv počtu pozorovaní, na presnosť odhadu pomocou inštrumentálnych premenných. Opäť môžeme intuitívne očakávať, že ak disponujeme väčším počtom dát, a teda máme k dispozícii viac informácií, potom vieme dosiahnuť presnejší odhad pre neznáme parametre modelu.

Simulácia 1.10. V ďalšom budeme simulačne skúmať vplyv počtu pozorovaní na presnosť odhadu b_{IV}. Presnosť odhadu overíme pomocou absolútnej hodnoty výchylky odhadu, ktorú odhadneme ako

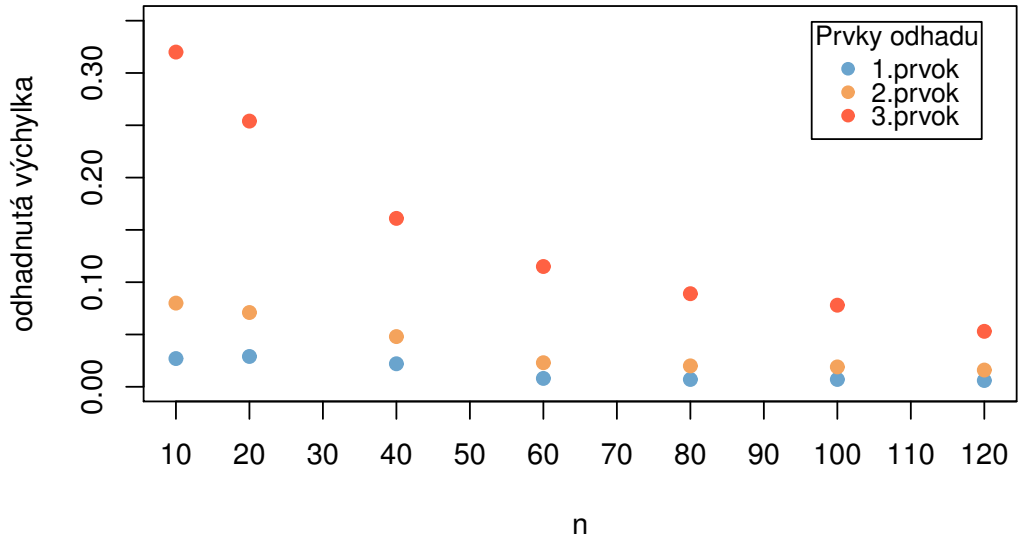
$$|\widehat{bias}(b_{IV})| = |\overline{b_{IV}} - \beta|,$$

kde aplikovanie absolútnej hodnoty sa chápe po súradniciach. Presnosť odhadu b_{IV} budeme porovnávať pre každý prvok odhadu zvlášť. Výpočty budeme postupne realizovať pre $n \in \{10, 20, 40, 60, 80, 100, 120\}$. Aby sme zabezpečili menšiu kolísavosť odhadu, zvolíme $K = 3$ a $L = 5$. Potom bude podľa (14) existovať nielen stredná hodnota odhadu ale

aj jeho konečná disperzia. Dáta budeme generovať ako náhodný výber z rozdelenia

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ \varepsilon \end{pmatrix} \sim N_9 \left(\begin{pmatrix} 3 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}2 & 0{,}2 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}2 & 1 & 0{,}2 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}2 & 0{,}2 & 1 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}25 & 0{,}25 & 0{,}25 & 1 & 0{,}3 & 0{,}3 & 0{,}3 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 1 & 0{,}3 & 0{,}3 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 0{,}3 & 1 & 0{,}3 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 0{,}3 & 0{,}3 & 1 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 0{,}3 & 0{,}3 & 0{,}3 & 1 & 0 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right) \tag{19}$$

a výsledky následne graficky znázornime.



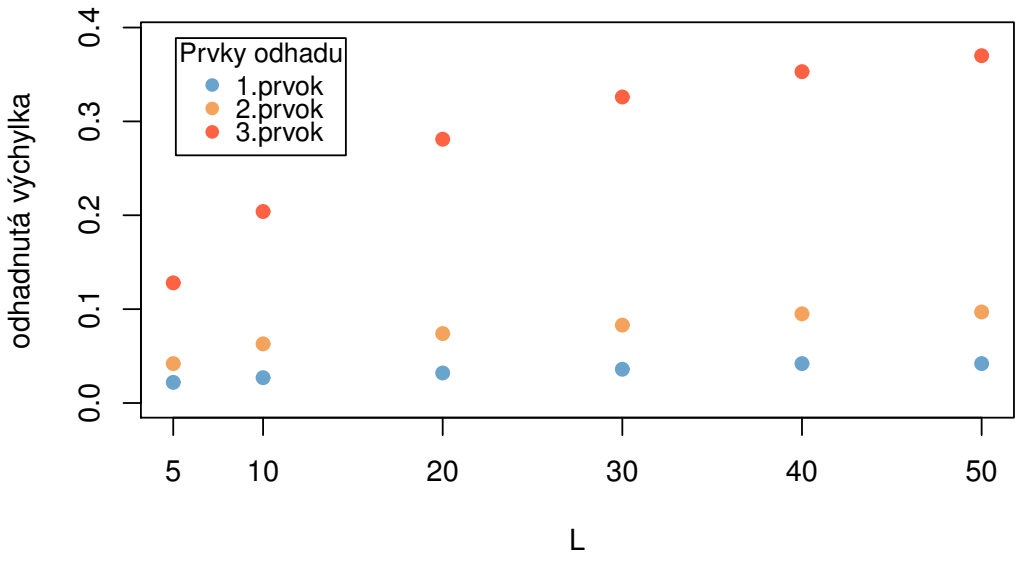
Obr. 1: Vplyv parametra n na odhadnuté absolútne výchylky prvkov odhadu b_{IV}.

Na obrázku 1 sú pre rôzne hodnoty parametra n znázornené rozličnými farbami jednotlivé prvky vektora odhadnutej výchylky odhadu v absolútnej hodnote. Získané výsledky sú v súlade s našim očakávaním. Vidíme, že s nárastom počtu pozorovaní n sa všetky prvky odhadnutej výchylky odhadu zmenšujú. To znamená, že s rastúcim počtom pozorovaní vieme získať presnejší odhad neznámych parametrov β. Ako sa aj v [3] spomína, súvisí to s

tým, že s rastúcim počtom pozorovaní sa stredná hodnota odhadu b_{IV} blíži ku skutočnému vektoru β.

Ďalej budeme simulačne skúmať vplyv počtu inštrumentálnych premenných na výchylku a na disperziu odhadu pomocou inštrumentálnych premenných b_{IV}. V [2] sa uvádza, že minimálny počet inštrumentálnych premenných L potrebných na výpočet odhadu je rovný počtu regresných premenných K a maximálny počet inštrumentálnych premenných, ktoré môžeme zvoliť je rovný počtu pozorovaní n.

Simulácia 1.11. Výchylky odhadov budeme počítať rovnako, ako v simulácii 1.10. Kolísavosť odhadov b_{IV} budeme porovnávať pomocou ich výberových kovariančných matíc, presnejšie pomocou determinantov týchto matíc. Výpočty budeme postupne opakovať pre $L \in \{5, 10, 20, 30, 40, 50\}$, keďže maximálne môžeme zvoliť $n = 50$ inštrumentálnych premenných. Dáta budeme generovať ako náhodný výber z rozdelenia (19), pričom dimenzia vektora stredných hodnôt a dimenzia kovariančnej matice sa budú líšiť v závislosti od voľby parametra L.



Obr. 2: Vplyv parametra L na odhadnuté absolútne výchylky prvkov odhadu b_{IV}.

Na obrázku 2 sú pre rôzne hodnoty parametra L znázornené rozličnými farbami jednotlivé prvky vektora odhadnutej výchylky odhadu v absolútnej hodnote. Výsledky v tabuľke 1,

L	5	10	20	30	40	50
$det(\widehat{cov}(b_{IV}))$	0,025335	0,000145	0,000009	0,000002	0,000001	0

Tabuľka 1: Vplyv parametra L na kolísavosť odhadu b_{IV}.

zaokrúhlené na 6 desatinných miest, a na obrázku 2 naznačujú, že s nárastom počtu inštrumentálnych premenných výchylka odhadu b_{IV} rastie a naopak disperzia odhadu b_{IV} klesá. Pokles disperzie odhadu s rastúcim počtom inštrumentálnych premenných vyplýva z tvrdenia (14). Pre veľký počet inštrumentálnych premenných L existuje stredná hodnota vysokého momentu pre odhad b_{IV}, a teda pravdepodobnosť, že odhad b_{IV} nadobudne nejakú extrémnu hodnotu je malá. V súhrne, ak zvýšime počet inštrumentálnych premenných dostaneme síce odhad, ktorý je menej rozkolísaný, ale zároveň je viac vychýlený od skutočného vektora β.

1.4 Smearing

Prirodzene môže nastať situácia, kedy budú v modeli lineárnej regresie iba niektoré regresné premenné korelované s náhodnou chybou ε. Ani v takomto prípade nie je vhodné aplikovať metódu najmenších štvorcov pre neznáme parametre týkajúce sa exogénnych premenných. Vychýlenosť odhadu spôsobená prítomnosťou aspoň jednej endogénnej premennej sa prenesie na odhad všetkých neznámych parametrov modelu. Teoretické poznatky budeme čerpať z [2] a pridáme výsledky vlastných simulácií.

Uvažujme model, v ktorom sú regresné premenné rozdelené na dve sady premenných na základe toho, či korelujú s náhodnou chybou. Regresnú maticu X môžeme potom zapísať ako

$$X = [X_C \;\; X_N],$$

kde matica X_C s rozmerom $n \times K_1$ obsahuje pozorovania K_1 regresných premenných a matica X_N s rozmerom $n \times K_2$ obsahuje pozorovania K_2 regresných premenných. Ďalej predpokladajme, že K_1 regresných premenných je exogénnych, teda nie sú korelované s náhodnou chybou. Naopak K_2 regresných premenných je endogénnych, čiže korelujú s náhodnou chybou. Maticu Z možno potom zvoliť ako

$$Z = [X_C \;\; Z_N],$$

čiže K_1 exogénnych premenných sa použije aj ako inštrumentálne premenné v matici Z. Zvyšné premenné v matici Z sú nejaké inštrumentálne premenné pre K_2 endogénnych premenných. Zároveň platí, že

$$\widehat{X} = [X_C \ \widehat{X}_N], \qquad \widehat{X}_N = Z(Z^T Z)^{-1} Z^T X_N.$$

Simulácia 1.12. Zvoľme teraz regresné premenné v modeli tak, že premenné x_1 a x_2 sú exogénne premenné a premenná x_3 je endogénna premenná a uvažujme generovanie dát ako náhodný výber z viacrozmerného normálneho rozdelenia

$$\begin{pmatrix} x_1 = z_1 \\ x_2 = z_2 \\ x_3 \\ z_3 \\ z_4 \\ \varepsilon \end{pmatrix} \sim N_6 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}1 & 0{,}2 & 0{,}25 & 0{,}25 & 0 \\ 0{,}1 & 1 & 0{,}3 & 0{,}25 & 0{,}25 & 0 \\ 0{,}2 & 0{,}3 & 1 & 0{,}25 & 0{,}25 & 0{,}4 \\ 0{,}25 & 0{,}25 & 0{,}25 & 1 & 0{,}3 & 0 \\ 0{,}25 & 0{,}25 & 0{,}25 & 0{,}3 & 1 & 0 \\ 0 & 0 & 0{,}4 & 0 & 0 & 1 \end{pmatrix} \right) \tag{20}$$

Ako sme už ukázali pri odvodení vzťahu (6), výchylka odhadu b_{LS} metódou najmenších štvorcov je asymptoticky rovná výrazu

$$Q_{xx}^{-1}\gamma = \begin{pmatrix} 1 & 0{,}1 & 0{,}2 \\ 0{,}1 & 1 & 0{,}3 \\ 0{,}2 & 0{,}3 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ 0{,}4 \end{pmatrix}.$$

To znamená, že v tomto prípade výchylka odhadu pomocou metódy najmenších štvorcov závisí od posledného stĺpca matice Q_{xx}^{-1}. Nie je dôvod, aby sme očakávali, že niektorý z prvkov posledného stĺpca matice Q_{xx}^{-1} sa bude rovnať nule. Z toho vyplýva, že aj keď iba jedna z premenných v matici X koreluje s náhodnou chybou, všetky prvky odhadu b_{LS} budú vychýlené, nie len odhad koeficientu pri endogénnej premennej. Tento jav sa nazýva *smearing* (v angl. preklade rozfŕkanie, vo význame výchylka v dôsledku aspoň jednej endogénnej premennej je rozfŕknutá vo všetkých prvkoch odhadu metódou najmenších štvorcov). Na odhadnutej výchylke odhadu metódou najmenších štvorcov s presnosťou na 2 desatinné miesta

$$\widehat{bias}(b_{LS}) = \begin{pmatrix} -0{,}08 \\ -0{,}13 \\ 0{,}45 \end{pmatrix}$$

vidíme, že v dôsledku endogénnej premennej x_3 sú všetky prvky odhadu vychýlené. Pre porovnanie, pomocou inštrumentálnych premenných získame odhad, ktorého odhadnutá výchylka je

$$\widehat{bias}(b_{IV}) = \begin{pmatrix} -0{,}03 \\ -0{,}04 \\ 0{,}17 \end{pmatrix}.$$

Vidíme, že sme dostali presnejší odhad a v porovnaní s odhadom metódou najmenších štvorcov sa výrazne neprejaví vplyv prítomnosti endogénnej premennej na zvyšné prvky odhadu.

1.5 Chybné merania

Doteraz sme implicitne predpokladali, že dáta použité na odhad neznámych parametrov v modeli sú presnými meraniami náhodných premenných. V praxi k tomu dochádza len zriedka a často máme k dispozícií len chybne namerané údaje. Taktiež sa môžeme stretnúť s modelmi, v ktorých nie všetky premenné majú určenú presnú mierku na základe ktorej ich môžeme merať, ako napríklad vzdelanie. Chybné merania spôsobujú, že odhad pre β metódou najmenších štvorcov sa stáva vychýleným. V tejto časti práce budeme vychádzať z teoretických výsledkov odvodených v [3]. Uvedieme vlastnú verziu odvodenia vzťahov, kde všetky kroky podrobne odôvodníme. Teóriu doplníme vlastnými simulačnými experimentami zamerané na aplikovanie teórie na umelo generovaných dátach.

Najprv uvažujme jednoduchý model lineárnej regresie s jednou regresnou exogénnou premennou x

$$y = x\beta + \varepsilon, \tag{21}$$

ktorý spĺňa základné predpoklady modelu lineárnej regresie definované v prvej kapitole práce. Ak by sme mali presné merania premenných x a y, potom metódou najmenších štvorcov by sme dostali nevychýlený odhad pre β. Predpokladajme ale, že regresná premenná x je chybne nameraná a poznáme len

$$x' = x + v, \qquad v \sim N(0,\sigma_v^2) \tag{22}$$

pričom premenné x a v sú nezávislé. Dosadením vzťahu (22) do rovnice (21) dostaneme

$$y = (x' - v)\beta + \varepsilon = x'\beta + (\varepsilon - v\beta) = x'\beta + w. \tag{23}$$

Dostávame sa k tomu, že v modeli (23) regresor x' koreluje s náhodnou chybou w, keďže

$$cov(x',w) = cov(x + v,\varepsilon - v\beta) = -\sigma_v^2\beta.$$

Podmienka (4) je porušená, tzn. odhad metódou najmenších štvorcov

$$b' = \left(X'^T X'\right)^{-1} X'^T y \tag{24}$$

je vychýleným odhadom pre β. V tomto prípade je X' vektor s rozmerom $n \times 1$ obsahujúci pozorovania náhodnej premennej x'. Ďalej pre odhad b' platí

$$\lim_{n\to\infty} b' = \lim_{n\to\infty} \left[\left(\frac{X'^T X'}{n} \right)^{-1} \frac{X'^T y}{n} \right]. \tag{25}$$

Keďže x' a y sú vektory s rozmerom $n \times 1$, potom limitu (25) môžeme vyjadriť aj ako

$$\lim_{n\to\infty} b' = \frac{\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} x_i' y_i}{\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} x_i'^2} = \frac{\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} (x_i + v_i)(\beta x_i + \varepsilon_i)}{\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} (x_i + v_i)^2}. \tag{26}$$

Ďalej využijeme, že premenné x, ε a v sú navzájom nezávislé a tým zjednodušíme (26) na

$$\lim_{n\to\infty} b' = \frac{\beta \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} x_i^2}{\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} x_i^2 + \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{i=1}^{n} v_i^2} = \frac{\beta Q}{Q + \sigma_v^2} = \frac{\beta}{1 + \frac{\sigma_v^2}{Q}}, \tag{27}$$

kde

$$Q = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x^2 = E\left(x^2\right) = Var(x) + [E(x)]^2. \tag{28}$$

Zo vzťahu (27) je zrejmé, že čím väčšia je disperzia chyby merania σ_v^2, tým je odhad b' viac vychýlený od vektora β, a to smerom nadol.

Pomocou simulácií sa pozrieme na to aké výsledky dostaneme, ak túto teóriu aplikujeme v konečnej vzorke dát. Zameriame sa na skutočnú výchylku odhadu b_{LS} od skutočného vektora β a na asymptotickú vlastnosť odhadu (27).

Simulácia 1.13. Premennú x budeme v simuláciách generovať ako náhodný výber z $N(3,1)$ a premennú x' dopočítame na základe vzťahu (22). Zvoľme $\sigma_v^2 = 1$ a $\beta = 2$. Predpokladajme, že máme k dispozícií iba chybne namerané údaje x'. Najprv budeme výpočty v jednotlivých simuláciách robiť pre $n = 50$. Potom priemer odhadov metódou najmenších štvorcov je rovný

$$\overline{b'} = 1{,}74.$$

To naznačuje, že odhad b' počítaný z chybne nameraných údajov je podhodnoteným odhadom pre β. Teraz zvoľme počet pozorovaní $n = 5000$ a dostaneme priemer $S = 10000$ nasimulovaných odhadov rovný

$$\overline{b'} = 1{,}82.$$

Tento výsledok už s presnosťou na 2 desatinné miesta sa zhoduje s limitnou hodnotou

$$\frac{\beta}{1 + \frac{\sigma_v{}^2}{Q}} = 1{,}82,$$

kde na základe vzťahu (28) je $Q = 10$ pre zvolené parametre v týchto simuláciách. Výsledky zo simulácií naznačujú, že s rastúcim počtom pozorovaní sa odhad b' blíži k limitnej hodnote (27).

Teraz uvažujme všeobecný model (3) s K regresnými premennými a predpokladajme, že každá regresná premenná je meraná s chybou, tak že

$$x_j^{'} = x_j + v_j, \qquad v_j \sim N(0,\sigma_{v_j}^2), \quad j = 1, ..., K. \tag{29}$$

Taktiež predpokladajme, že premenné x_j a v_j sú nezávislé pre $j = 1, ..., K$. Keďže máme k dispozícií iba chybné merania, odhad pre β vypočítame ako

$$b' = \left(X'^T X'\right)^{-1} X'^T y. \tag{30}$$

Teraz X' označuje maticu s rozmerom $n \times K$ obsahujúcu pozorovania premenných $x_j^{'}$. Potom zovšeobecnením predchádzajúcich výsledkov sa dá ukázať, že

$$\lim_{n\to\infty} b' = (Q + \Sigma_v)^{-1} Q\beta = \beta - (Q + \Sigma_v)^{-1} \Sigma_v \beta, \tag{31}$$

kde

$$Q = E\left(\begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} . \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}^T\right), \quad \Sigma_v = \begin{pmatrix} \sigma_{v_1}^2 & 0 & . & . & . & 0 \\ 0 & \sigma_{v_2}^2 & 0 & . & . & . \\ . & 0 & . & . & . & . \\ . & . & . & . & . & 0 \\ 0 & . & . & . & 0 & \sigma_{v_K}^2 \end{pmatrix}.$$

Simulácia 1.14. Znovu overíme teóriu na konečnej vzorke dát. Pre $K = 3$ budeme jednotlivé regresné premenné nezávisle generovať ako náhodný výber z rozdelenia $N(2,1)$. Vypočítame odhad metódou najmenších štvorcov pre

$$\beta = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Predpokladajme, že poznáme len chybne namerané údaje X', pričom zvolíme $\sigma^2_{v_j} = j$ pre $j = 1, 2, 3$. Ak nastavíme počet pozorovaní $n = 10$, potom dostaneme priemer odhadov zo simulácií

$$\overline{b'} = \begin{pmatrix} 2{,}07 \\ 1{,}72 \\ 1{,}54 \end{pmatrix},$$

ktorý je vychýleným odhadom pre β. Teraz ak zvolíme $n = 500$ dostaneme priemer odhadov

$$\overline{b'} = \begin{pmatrix} 2{,}03 \\ 1{,}69 \\ 1{,}52 \end{pmatrix},$$

ktorý s presnosťou na 2 desatinné miesta sa zhoduje s limitnou hodnotou (31).

Pozrime sa ešte na situáciu, keď iba jedna premenná v modeli (3) je chybne nameraná. Aj v takomto prípade sa vychýlenosť prejaví na odhade každého neznámeho parametra v modeli. Bez ujmu na všeobecnosti predpokladajme, že prvá premenná x_1 je chybne meraná :

$$x_1^{'} = x_1 + v, \qquad v \sim N(0,\sigma_v^2).$$

Zhrnieme výsledky pre tento špeciálny prípad. Dá sa odvodiť, že pre chybne nameranú premennú x_1 bude pre odhad $b_1^{'}$ prislúchajúci neznámemu parametru β_1 platiť

$$\lim_{n\to\infty} b_1^{'} = \frac{\beta_1}{1 + \sigma_v^2 Q_{11}^{-1}}, \tag{32}$$

a pre zvyšné premenné, ktoré nie sú chybne merané, bude pre odhady platiť

$$\lim_{n\to\infty} b_k^{'} = \beta_k - \beta_1 \frac{\sigma_v^2 Q_{k1}^{-1}}{1 + \sigma_v^2 Q_{k1}^{-1}} \qquad k = 2, ..., K, \tag{33}$$

kde Q_{ij}^{-1} označuje (i,j)-ty prvok matice Q^{-1}.

Ak máme viac ako jednu premennú meranú s chybou, potom sa odvodenie týchto výsledkov výrazne skomplikuje, lebo závisí od veľkého množstva parametrov.

Simulácia 1.15. Opäť zvoľme nezávislé generovanie regresných premenných (x_1, x_2, x_3) ako realizácie z rozdelenia $N(2,1)$. Predpokladajme, že premennú x_1 máme chybne nameranú a $\sigma_v^2 = 3$. Vypočítané priemery odhadov neznámych parametrov zo simulácií porovnáme s hodnotami (32) a (33), ku ktorým by sa mali limitne blížiť. Pre $n = 10$ dostaneme priemer odhadov

$$\overline{b'} = \begin{pmatrix} 0{,}33 \\ 2{,}30 \\ 3{,}31 \end{pmatrix}$$

a pre $n = 500$

$$\overline{b'} = \begin{pmatrix} 0{,}32 \\ 2{,}30 \\ 3{,}32 \end{pmatrix}.$$

Výsledky potvrdili, že chybné meranie jednej premennej má vplyv na všetky prvky odhadu. Odhady neznámych parametrov sa stávajú vychýleným a pre $n = 500$ sú výsledky s presnosťou na 2 desatinné miesta v súlade limitnými hodnotami (32) a (33).

2 Testy špecifikácie

Ak nemáme informáciu o tom, či sú regresné premenné v modeli korelované s náhodnými chybami, je potrebné si to overiť pomocou vhodného testu. Hlavným dôvodom je, že ak sa v modeli lineárnej regresie nenachádzajú žiadne endogénne premenné, potom variančná matica odhadu metódou najmenších štvorcov b_{LS} je „menšia“ než variančná matica odhadu pomocou inštrumentálnych premenných b_{IV}. Ak je IV odhad nepotrebný, tzn. γ z (6) je nulový vektor, potom použitie odhadu metódou najmenších štvorcov je aj výpočtovo menej náročné a nie je potrebné navyše hľadať vhodné inštrumentálne premenné. V tejto časti práce budeme čerpať najmä z [3] a [6] a doplníme vlastné ilustrácie a závery praktického fungovania testov na základe výsledkov simulácií v štatistickom softvéri R.

2.1 Hausmanov test a bootstrap

Prvý z testov na overenie závislosti medzi regresnými premennými a náhodnými chybami, ktoré v práci uvedieme, je Hausmanov test špecifikácie. V modeli lineárnej regresie (3) chceme testovať nasledovné hypotézy

$$H_0: \quad \gamma = 0 \qquad vs \qquad H_1: \quad \gamma \neq 0, \tag{34}$$

kde γ je vektor obsahujúci kovariancie medzi premennými $x_1, ..., x_K$ a ε. Za platnosti nulovej hypotézy sú oba odhady b_{LS} a b_{IV} konzistentnými odhadmi pre β. Ak platí alternatívna hypotéza, potom je iba odhad b_{IV} konzistentným odhadom pre β. Na základe tejto myšlienky bola stanovená testová štatistika pre danú nulovú hypotézu. Označme

$$d = b_{IV} - b_{LS}.$$

Ak H_0 platí, tak d očakávame blízke nulovému vektoru a naopak, ak H_0 neplatí, tak d očakávame výrazne odlišné od nulového vektora. Príslušná testová štatistika sa stanoví v kontexte myšlienky Waldovej testovej štatistiky ako

$$H = \frac{d^T[(\widehat{X}^T\widehat{X})^{-1} - (X^TX)^{-1}]^{-1}d}{s^2} \sim \chi^2_{K^*}, \tag{35}$$

kde s^2 je spoločný odhad pre σ^2 definovaný ako

$$s^2 = \frac{(y - Xb_{LS})^T(y - Xb_{LS})}{n - K} \tag{36}$$

a K^* predstavuje počet premenných v modeli, ktoré podozrievame z endogénnosti. Inverzná matica k matici v hranatých zátvorkách v čitateli testovej štatistiky (35) nemusí existovať, preto je potrebné počítať zovšeobecnenú inverznú maticu. Podrobné odvodenie testovej štatistiky a porovnanie variančných matíc odhadov je dostupné v [3].

Bootstrap je často využívaná štatistická metóda, ktorá predstavuje alternatívny prístup k testovaniu hypotéz. Hausmanov test je možné vylepšiť aplikovaním bootstrapovej metódy, najmä ak máme k dispozícii malý počet pozorovaní alebo ak sú regresné premenné slabo korelované s inštrumentálnymi premennými. V nasledujúcom budeme čerpať z [6].

Postup pre Hausmanov test založený na bootstrape je nasledovný:

1. Z pôvodnej vzorky údajov (X, Z, y) zostrojíme novú simulovanú vzorku (X, Z, y^*) takto:

   Najprv vypočítame vektor rezíduí

   $$u = y - Xb_{LS},$$

   kde b_{LS} je odhad metódou najmenších štvorcov v pôvodnej vzorke údajov. Potom zo sady dát $(u_1, u_2, ..., u_n)$ vytvoríme náhodným výberom s vrátením novú sadu $(u_1^*, u_2^*, ..., u_n^*)$, ktorú následne použijeme na výpočet y^* :

   $$y^* = Xb_{LS} + u^*.$$

2. Zo vzorky dát (X, Z, y^*) vypočítame hodnotu testovej štatistiky (35) Hausmanovho testu a označíme ju H^*.

3. Prvé dva kroky postupu zopakujeme B-krát, a tak získame hodnoty bootstrapových testových štatistík $H_1^*, H_2^*, ..., H_B^*$. Týchto B hodnôt získaných zo simulovaných vzoriek použijeme na aproximáciu rozdelenia testovej štatistiky H za platnosti nulovej hypotézy. Hodnotu testovej štatistiky vypočítame aj z pôvodnej vzorky údajov (X, Z, y) a označíme ju H.

4. Hodnoty B odhadov zoradíme vzostupne. Nulovú hypotézu na hladine významnosti $(\alpha \times 100)\%$ zamietame, ak

   $$H > H_{1-\alpha}^*,$$

kde $H^*_{1-\alpha}$ je $((1-\alpha)\times B)$-ta hodnota v rámci hodnôt $H^*_1, H^*_2, ..., H^*_B$ zoradených podľa veľkosti od najmenšej po najväčšiu.

Nevýhodou tohto testu je, že výpočtový čas je veľmi dlhý, pretože testová štatistika sa počíta až B-krát. Bootstrapovú verziu Hausmanovho testu je vhodné použiť, najmä ak máme k dispozícií malý počet pozorovaní, keďže testová štatistika má len asymptoticky χ^2-rozdelenie. Zakomponovanie bootstrapovej metódy do testu umožňuje aproximovať rozdelenie testovej štatistiky pomocou empirickej distribučnej funkcie.

Simulácia 2.1. Uvažujme generovanie dát ako náhodný výber z rozdelenia

$$\begin{pmatrix} x_1 = z_1 \\ x_2 = z_2 \\ x_3 \\ z_3 \\ \varepsilon \end{pmatrix} \sim N_5 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0{,}2 & 0{,}2 & 0{,}25 & \\ 0{,}2 & 1 & 0{,}2 & 0{,}25 & \gamma \\ 0{,}2 & 0{,}2 & 1 & 0{,}25 & \\ 0{,}25 & 0{,}25 & 0{,}25 & 1 & 0 \\ & \gamma^T & & 0 & 1 \end{pmatrix} \right). \tag{37}$$

To znamená, že vektor γ má tvar

$$\gamma = \begin{pmatrix} 0 \\ 0 \\ \rho_{x\varepsilon} \end{pmatrix}.$$

Postupne budeme zvyšovať korelačný koeficient $\rho_{x\varepsilon}$ medzi regresorom x_3 a náhodnou chybou ε, čím porušíme platnosť nulovej hypotézy definovanej v (34). Nastavme počet simulácií $S = 10000$ a počet realizácií bootstrapového výberu $B = 1000$. Fungovanie vyššie odvodených testov porovnáme na základe odhadov pravdepodobnosti chyby prvého druhu pre rôzne voľby počtu pozorovaní n. Budeme vychádzať z toho, že pravdepodobnosť chyby prvého druhu testu by mala byť ideálne rovná nominálnej hladine významnosti testu α. Testy budeme realizovať na hladine významnosti $\alpha = 5\%$. Pre výpočet odhadu pravdepodobnosti chyby prvého druhu je potrebné dáta generovať za platnosti nulovej hypotézy (34). To vieme docieliť voľbou parametra $\rho_{x\varepsilon} = 0$. Taktiež budeme simulovať aj silu testov, čiže fungovanie za neplatnosti nulovej hypotézy. Postupným zvyšovaním korelačného koeficientu $\rho_{x\varepsilon}$ by mala empirická sila testov narastať. Pri výpočte testovej štatistiky (35) aplikujeme Moore-Penroseovu zovšeobecnenú inverziu.

	n=10		n=50	
$\rho_{x\varepsilon}$	Hasumanov test	Bootstrapová verzia Hausmanovho testu	Hasumanov test	Bootstrapová verzia Hausmanovho testu
0	3,65	5,09	4,68	5,36
0,1	3,77	5,07	5,52	5,22
0,2	3,61	5,07	5,8	6,65
0,3	4,47	5,28	8,02	8,48
0,4	4,77	5,43	11,07	11,27
0,5	5,21	6,07	15,97	16,18
0,6	4,86	7,31	24,42	26,02
0,7	6,18	9,02	39,5	40,59
0,8	9,81	12,56	65,7	66,1
0,9	22,18	27,62	97,72	97,74

Tabuľka 2: Empirická sila testov vyjadrená v percentách na hladine významnosti $\alpha = 5\%$.

Výsledky v tabuľke 2 zaokrúhlené na 2 desatinné miesta naznačujú, že pre $n = 10$ za platnosti nulovej hypotézy Hausmanov test zamieta nulovú hypotézu len v 3,65% simuláciách, zatiaľ čo bootstrapová verzia testu zamieta v 5,09% simuláciách. Aplikovanie bootstrapovej metódy na testovanie danej nulovej hypotézy má za následok výrazné priblíženie pravdepodobnosti chyby prvého druhu na približne požadovanú hodnotu 5%. To znamená, že pre malý počet dát $n = 10$ nie je vhodné použiť Hausmanov test, ktorý je príliš konzervatívny, ale vhodnou alternatívou je bootstrapová verzia tohto testu. S väčším počtom dát $n = 50$ sa však výsledky zlepšujú, čo je v súlade s asymptotickou vlastnosťou testovej štatistiky (35).

2.2 Wuov test

Alternatívny prístup testovania nezávislosti medzi regresnými premennými a náhodnými chybami je test, ktorý navrhol D. Wu. Hlavná myšlienka testu je uvedená v [3], kde je predstavený ako asymptoticky ekvivalentný test k Hausmanovmu testu. Využíva sa F-štatistika na testovanie signifikantnosti premenných v rozšírenom lineárnom regresnom

modeli. Základnú myšlienku testu z [3] doplníme o vlastný rozširujúci komentár a postup výpočtov podrobne popíšeme.

V rozšírenom regresnom modeli

$$y = X\beta + \widehat{X}^*\nu + \varepsilon^* \tag{38}$$

sa testuje signifikantnosť (čiže nenulovosť) prvkov vektora ν. Vektor ν obsahuje korelácie medzi regresnými premennými podozrivými z endogénnosti a náhodnou chybou ε. Matica X^* obsahuje regresné premenné z X neobsiahnuté v Z a

$$\widehat{X}^* = P_Z X^*.$$

Model (38) sa dá prepísať do tvaru

$$y = \left(X \;\; \widehat{X}^*\right) \begin{pmatrix} \beta \\ \nu \end{pmatrix} + \varepsilon^* = W \begin{pmatrix} \beta \\ \nu \end{pmatrix} + \varepsilon^*. \tag{39}$$

Submodel, ktorý predstavuje model postačujúci na modelovanie dát za platnosti nulovej hypotézy, má potom tvar základného regresného modelu

$$y = X\beta + \varepsilon. \tag{40}$$

Príslušná testová štatistika je definovaná ako

$$F = \frac{\dfrac{RSS_0 - RSS_1}{K^*}}{\dfrac{RSS_1}{n - K - K^*}} \sim F_{K^*, n-K-K^*}, \tag{41}$$

kde K^* predstavuje počet možných endogénnych premenných v modeli, resp. počet zložiek vektora ν. Za platnosti nulovej hypotézy sa testová štatistika (41) asymptoticky riadi F-rozdelením s K^* a $n \breve{} K \breve{} K^*$ stupňami voľnosti. Súčty štvorcov rezíduí, ktoré vstupujú do testovej štatistiky sa vypočítajú pomocou odhadu metódou najmenších štvorcov ako

$$RSS_0 = (y - Xb_{LS0})^T (y - Xb_{LS0})$$
$$RSS_1 = (y - Wb_{LS1})^T (y - Wb_{LS1}),$$

kde b_{LS0} je odhad metódou najmenších štvorcov v modeli (40) a b_{LS1} je odhad metódou najmenších štvorcov v modeli (39).

Simulácia 2.2. Teraz zopakujeme simulačný experiment popísaný v simulácii 2.1 a porovnáme empirickú silu Hausmanovho testu a Wuovho testu na hladine významnosti $\alpha = 5\%$. Rovnako ako v predošlých simuláciách zvolíme $K = 3$. Porovnanie testov realizujeme pre rôzne nastavenia počtu inštrumentálnych premenných L.

	Hausmanov test			Wuov test		
$\rho_{x\varepsilon}$	$L = K$	$L = K + 1$	$L = K + 2$	$L = K$	$L = K + 1$	$L = K + 2$
0	4,68	4,64	4,65	4,85	4,72	4,79
0,1	5,52	5,5	5,26	5,63	5,55	5,37
0,2	5,8	6,72	6,69	5,95	6,86	6,82
0,3	8,02	8,66	9,96	8,22	8,82	10,12
0,4	11,07	12,87	13,76	11,3	13,03	13,98
0,5	15,97	18,93	20,62	16,17	19,18	20,88
0,6	24,42	29,09	32,36	24,78	29,45	32,72
0,7	39,5	47,88	50,7	39,92	48,33	51,11
0,8	65,7	73,99	77,68	66,03	74,22	77,87
0,9	97,72	95,42	95,57	97,74	95,45	95,62

Tabuľka 3: Empirická sila testov vyjadrená v percentách na hladine významnosti $\alpha = 5\%$.

Výsledky v tabuľke 3 zaokrúhlené na 2 desatinné miesta naznačujú, že porovnávané testy majú takmer rovnakú empirickú silu. To je v súlade s teóriou, keďže testové štatistiky týchto testov sú asymptoticky ekvivalentné.

3 Testy o parametroch regresie

Táto kapitola práce bude zameraná na testovanie lineárnych hypotéz o neznámych parametroch modelu lineárnej regresie. Počas simulačných štúdií sme zistili, že rôzne zdroje uvádzajú rôzne postupy testovania, ktoré sa líšia v tvare testovej štatistiky. Bližšie sa zameriame na príčinu rozdielov v jednotlivých testových štatistikách. Taktiež navrhneme úpravu testových štatistík pre nedobre fungujúce testy. V ďalšom budeme čerpať z [1], [2] a [5], pričom budeme simulačne skúmať fungovanie testov na umelo generovaných dátach v štatistickom softvéri R.

3.1 Lineárne hypotézy o neznámych parametroch

Uvažujme model lineárnej regresie (3), v ktorom regresné premenné korelujú s náhodnou chybou a nulovú hypotézu o lineárnych ohraničeniach

$$H_0: \quad R\beta = r, \tag{42}$$

kde R je matica s rozmerom $q \times K$, r je vektor s rozmerom $q \times 1$ a parameter q označuje počet lineárnych ohraničení. Keďže máme model obsahujúci endogénne regresné premenné, nie je korektné použiť bežný postup testovania lineárnych hypotéz o β v modeli lineárnej regresie. Ako prvé sme sa zamerali na postup testovania popísaný v knihe [2] autorov J. Johnston a J. DiNardo. Naše výsledky simulácií poukážu na to, že tento test nie je vždy vhodný na testovanie nulovej hypotézy v danom modeli.

V [2] je bez odvodenia uvedený nasledovný postup testovania lineárnych hypotéz:

1. Odhad b_{IV0} parametrov β pomocou inštrumentálnych premenných v modeli zohľadňujúcom lineárne ohraničenia, ktorý môžeme formálne zapísať ako

$$y = X\beta + \varepsilon \quad \wedge \quad R\beta = r \tag{43}$$

   a výpočet príslušného súčtu štvorcov rezíduí

$$RSS_0 = \left(y - \widehat{X} b_{IV0}\right)^T \left(y - \widehat{X} b_{IV0}\right). \tag{44}$$

2. Odhad b_{IV1} parametrov β pomocou inštrumentálnych premenných v modeli bez lineárnych ohraničení, teda v modeli

$$y = X\beta + \varepsilon \tag{45}$$

a výpočet príslušného súčtu štvorcov rezíduí

$$RSS_1 = \left(y - \widehat{X}b_{IV1}\right)^T \left(y - \widehat{X}b_{IV1}\right). \tag{46}$$

3. Výpočet súčtu štvorcov rezíduí pomocou odhadu b_{IV1} a matice X

$$RSS = \left(y - Xb_{IV1}\right)^T \left(y - Xb_{IV1}\right). \tag{47}$$

Potom testová štatistika pre nulovú hypotézu je definovaná ako

$$F_{JD} = \frac{\frac{RSS_0 - RSS_1}{q}}{\frac{RSS}{n-K}}. \tag{48}$$

Na záver testovú štatistiku (48) porovnáme s F-rozdelením s q a $n-K$ stupňami voľnosti.

V simulačných experimentoch sme narazili na situáciu, keď vyššie uvedený test nefunguje správne a správa sa extrémne konzervatívne. Tento jav môžeme pozorovať aj na nasledujúcom príklade.

Simulácia 3.1. Uvažujme model s $K = 3$ regresnými premennými, $L = 6$ inštrumentálnymi premennými a počtom pozorovaní $n = 1000$. Počet simulácií ponecháme $S = 10000$. Dáta budeme generovať ako náhodný výber z viacrozmerného normálneho rozdelenia, pričom stredné hodnoty všetkých premenných $(x_1, x_2, x_3, z_1, z_2, ..., z_6, \varepsilon)$ zvolíme rovné 0 a ich disperzie rovné 1. Kovariancie medzi jednotlivými premennými nastavíme na

$$cov(x_i, x_j) = cov(z_i, z_j) = cov(x_i, \varepsilon) = 0{,}4, \qquad i \neq j$$

a kovariancie medzi náhodnou chybou ε a inštrumentálnymi premennými $z_1, ..., z_6$ sú z predpokladu nulové. Keďže uvažujeme nulové stredné hodnoty náhodných premenných modelu, platí nasledovné

$$Q_{xz} = cov\left(\begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}, \begin{pmatrix} z_1 \\ \vdots \\ z_L \end{pmatrix}\right). \tag{49}$$

Podľa predpokladu P9 musí mať matica Q_{xz} v (49) lineárne nezávislé riadky. Ak by sme zvolili $cov(x_i, z_j) = 0{,}4$, tak by matica

$$Q_{xz} = \begin{pmatrix} 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \end{pmatrix} \tag{50}$$

mala lineárne závislé riady. Riadky matice v (50) preto náhodne zašumíme, aby sme dostali maticu Q_{xz} s hodnosťou $K = 3$, teda maticu s lineárne nezávislými riadkami. Ku každému prvku matice v (50) pripočítame náhodné číslo z rovnomerného rozdelenia na intervale $[-0{,}05, 0{,}05]$. Dostaneme tak maticu

$$Q_{xz} \doteq \begin{pmatrix} 0{,}40 & 0{,}42 & 0{,}38 & 0{,}39 & 0{,}36 & 0{,}39 \\ 0{,}38 & 0{,}36 & 0{,}38 & 0{,}42 & 0{,}41 & 0{,}36 \\ 0{,}39 & 0{,}37 & 0{,}41 & 0{,}41 & 0{,}39 & 0{,}38 \end{pmatrix}. \tag{51}$$

Riadky matice Q_{xz} v (51), ktoré predstavujú vektory, sú síce lineárne nezávislé ale sú takmer zhodné. Predpoklad P9 je však momentálne splnený. Definujme lineárne ohraničenia pre vektor β, ktorých platnosť budeme testovať v nulovej hypotéze

$$H_0: \quad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{52}$$

alebo ekvivalentne testujeme

$$H_0: \quad \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{53}$$

Definujme vektor β tak, aby bola splnená nulová hypotéza, napr.

$$\beta = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Model (43) zohľadňujúci dané lineárne ohraničenia má v tomto prípade tvar

$$y = X_1\beta_1 + \varepsilon. \tag{54}$$

Odhad b_{IV0} vypočítame ako

$$b_{IV0} = \begin{pmatrix} b_1 \\ 0 \\ 0 \end{pmatrix},$$

kde b_1 je odhad pre β_1 pomocou inštrumentálnych premenných v modeli (54). Po aplikovaní predchádzajúceho postupu testovania dostávame odhad pravdepodobnosti chyby prvého druhu testu na hladine významnosti $\alpha = 5\%$ rovný 0,29%. Test sa správa konzervatívne, tzn. nulovú hypotézu zamieta veľmi zriedkavo. Výsledné hodnoty testovej štatistiky boli teda často výrazne blízke k nule, a preto test nulovú hypotézu len zriedka zamietol. Dôvodom, ktorý sme si všimli, je, že odhad disperzie chýb

$$\widehat{\sigma^2}_{JD} = \frac{RSS}{n-K} = \frac{(y - Xb_{IV1})^T (y - Xb_{IV1})}{n-K}, \tag{55}$$

ktorý je súčasťou testovej štatistiky F_{JD} v (48), vychádza v priemere nadhodnotený. V simuláciách máme nastavený parameter $\sigma^2 = 1$. Očakávame, že ak je odhad daný výrazom (55) nevychýleným odhadom pre σ^2, potom priemer zo $\widehat{\sigma^2}_{JD}$ vypočítaný v simuláciách bude približne rovný hodnote 1. Pre $S = 10000$ simulácií je priemer zo $\widehat{\sigma^2}_{JD}$ zaokrúhlený na 2 desatinné miesta rovný 1,43, čo naznačuje vychýlenosť tohto odhadu. O nevhodnosti odhadu $\widehat{\sigma^2}_{JD}$ sa vieme presvedčiť aj dosadením skutočnej hodnoty $\sigma^2 = 1$ do testovej štatistiky F_{JD} v simuláciách. V takom prípade dostaneme odhad pravdepodobnosti chyby prvého druhu rovný 5,07%.

Pri zavedení postupu testovania pomocou F_{JD} sa autori odvolávajú na knihu [1]. V [1] je uvedený podrobnejší návod ako testovať lineárne hypotézy v modeli s endogénnymi premennými. Na odvodenie asymptoticky platnej testovej štatistiky sa využíva verzia Gauss-Newtonovej regresie s inštrumentálnymi premennými (IVGNR). V ďalšom budeme vychádzať z [1].

Nulovú a alternatívnu hypotézu môžeme vo všeobecnosti zapísať ako

$$H_0: \quad y = X\beta + \varepsilon \quad \wedge \quad R\beta = r \tag{56}$$

$$H_1: \quad y = X\beta + \varepsilon. \tag{57}$$

V metóde IVGNR sa skonštruujú modely

$$IVGNR_0: \quad y - X\beta' = \widehat{X}\beta + \varepsilon \quad \wedge \quad R\beta = r \tag{58}$$

$$IVGNR_1: \quad y - X\beta' = \widehat{X}\beta + \varepsilon. \tag{59}$$

Ďalej sa určia odhady metódou najmenších štvorcov v modeloch (58) a (59) a následne sa vypočítajú príslušné súčty štvorcov rezíduí RSS_0 a RSS_1 potrebné na výpočet testovej štatistiky

$$F_{IVGNR} = \frac{\frac{RSS_0 - RSS_1}{q}}{\frac{RSS_1}{n-K}}. \tag{60}$$

Tieto výpočty sa dajú zjednodušiť využitím niekoľkých vlastností IVGNR. Prvou vlastnosťou IVGNR je, že hodnota rozdielu $RSS_0 - RSS_1$ v čitateli testovej štatistiky F_{IVGNR} nezávisí od voľby vektora β'. To znamená, že na výpočet súčtov štvorcov rezíduí v tomto prípade môžeme zvoliť β' v modeloch (58) a (59) ako nulový vektor. Dôsledkom toho je, že čitateľ testovej štatistiky F_{IVGNR} v (60) sa zhoduje s čitateľom testovej štatistiky F_{JD} v (48). Pre súčet štvorcov rezíduí RSS_1, ktorý vstupuje do menovateľa testovej štatistiky samostatne táto vlastnosť neplatí a pri jeho výpočte treba v modeli (59) zvoliť vektor β' vhodne. Je nutné, aby vektor β' bol konzistentným odhadom pre β za platnosti nulovej hypotézy inak nedostaneme konzistentný odhad pre σ^2 a testová štatistika F_{IVGNR} sa nebude asymptoticky riadiť rozdelením $F_{q,n-K}$.

Ďalšia vlastnosť sa týka odhadu metódou najmenších štvorcov v modeli (59), označme ho ako b. Tento odhad sa dá vypočítať jednoduchšie pomocou vzťahu

$$b = (\widehat{X}^T X)^{-1}\widehat{X}^T(y - X\beta') = (\widehat{X}^T X)^{-1}\widehat{X}^T y - (\widehat{X}^T X)^{-1}\widehat{X}^T X\beta' = b_{IV1} - \beta', \tag{61}$$

kde b_{IV1} je odhad pre β pomocou inštrumentálnych premenných v modeli (57). Tu vypočítaný odhad b_{IV1} je samozrejme totožný s odhadom b_{IV1} z predchádzajúceho testu pomocou F_{JD}. Poznamenajme, že pri odvodení vzťahu pre odhad b v (61) sme použili vetu 1.3.

Potom príslušné rezíduá dané odhadom b v modeli (59) vypočítame ako $y - X\beta' - \widehat{X}b$. To sa dá upraviť na tvar obsahujúci rezíduá $y - Xb_{IV1}$ využívané pri výpočte $\widehat{\sigma^2}_{JD}$:

$$y - X\beta' - \widehat{X}b = y - X\beta' - \widehat{X}(b_{IV1} - \beta') = y - X\beta' - P_Z X b_{IV1} + P_Z X\beta'$$
$$= y - Xb_{IV1} + Xb_{IV1} - P_Z X b_{IV1} - X\beta' + P_Z X\beta' = y - Xb_{IV1} + (I - P_Z)X(b_{IV1} - \beta').$$

Odhad disperzie σ^2 je potom daný vzťahom

$$\widehat{\sigma^2}_{IVGNR} = \frac{(y - Xb_{IV1} + (I - P_Z)X(b_{IV1} - \beta'))^T(y - Xb_{IV1} + (I - P_Z)X(b_{IV1} - \beta'))}{n - K}. \tag{62}$$

Porovnajúc $\widehat{\sigma^2}_{JD}$ v (55) a $\widehat{\sigma^2}_{IVGNR}$ v (62) vidíme, že testová štatistika F_{IVGNR} sa teda líši od predchádzajúcej testovej štatistiky F_{JD} iba v odhade pre σ^2 a to o výraz

$$(I - P_Z)X(b_{IV1} - \beta').$$

Takáto úprava testovej štatistiky eliminuje problém, na ktorý sme narazili pri aplikovaní predošlého testu pomocou F_{JD} v simulácii 3.1. Odhad $\widehat{\sigma^2}_{IVGNR}$ daný vzťahom (62) sa javí byť lepším odhadom disperzie σ^2, ako to preukážu aj simulácie.

Zhrnutie postupu testu pomocou IVGNR:

1. Odhad b_{IV0} parametrov β metódou najmenších štvorcov v modeli (58), ak β' je nulový vektor, tzn. odhad pomocou inštrumentálnych premenných v modeli (56) a výpočet príslušného súčtu štvorcov rezíduí

$$RSS_0 = \left(y - \widehat{X}b_{IV0}\right)^T \left(y - \widehat{X}b_{IV0}\right). \tag{63}$$

2. Odhad b_{IV1} parametrov β metódou najmenších štvorcov v modeli (59), ak β' je nulový vektor, tzn. odhad pomocou inštrumentálnych premenných v modeli (57) a výpočet príslušného súčtu štvorcov rezíduí

$$RSS_1 = \left(y - \widehat{X}b_{IV1}\right)^T \left(y - \widehat{X}b_{IV1}\right). \tag{64}$$

3. Výpočet vektora β' ako konzistentného odhadu pre β za platnosti nulovej hypotézy a následne výpočet vektora b na základe vzťahu

$$b = b_{IV1} - \beta'.$$

   Vektor β' závisí od konkrétneho lineárneho ohraničenia, ktoré je definovaný v nulovej hypotéze. Nižšie zavedieme explicitný vzorec pre β' pre prípad, kedy testujeme nulovosť niektorých prvkov vektora β.

4. Výpočet súčtu štvorcov rezíduí, ktorý vstupuje do menovateľa testovej štatistiky F_{IVGNR} na základe vzťahu

$$RSS = \left(y - X\beta' - \widehat{X}b\right)^T \left(y - X\beta' - \widehat{X}b\right). \tag{65}$$

Testová štatistika

$$F_{IVGNR} = \frac{\frac{RSS_0 - RSS_1}{q}}{\frac{RSS}{n-K}} \tag{66}$$

sa za platnosti nulovej hypotézy asymptoticky riadi rozdelením $F_{q,n-K}$.

Simulácia 3.2. Teraz aplikujeme postup testovania pomocou F_{IVGNR} na model definovaný v simulácii 3.1. Zmena nastáva vo výpočte menovateľa testovej štatistiky. Aby bol test vhodný na testovanie danej nulovej hypotézy, a teda mal pravdepodobnosť chyby prvého druhu rovnú 5%, použijeme odhad disperzie $\widehat{\sigma^2}_{IVGNR}$. K tomu je potrebné určiť vektor β'. Vektor β' má byť konzistentným odhadom pre β za platnosti nulovej hypotézy, preto zvolíme $\beta' = b_{IV0}$. Výpočet odhadu b_{IV0} sme už popísali v simulácii 3.1. Odhad pravdepodobnosti chyby prvého druhu testu pomocou F_{IVGNR} na hladine významnosti $\alpha = 5\%$ je teraz rovný 5,22%. Priemer z odhadov disperzie $\widehat{\sigma^2}_{IVGNR}$ s presnosťou na 2 desatinné miesta je rovný 1,00. To naznačuje, že odhad $\widehat{\sigma^2}_{IVGNR}$ je nevychýleným odhadom pre disperziu náhodnej chyby σ^2 a upravená testová štatistika F_{IVGNR} sa asymptoticky riadi rozdelením $F_{q,n\breve{}K}$.

Skúmali sme príčiny toho, prečo test pomocou F_{JD} v simulácii 3.1 funguje nesprávne. Došli sme k záveru, že problém spôsobuje matica Q_{xz} definovaná v (8), na ktorú sú testová štatistika F_{JD} a odhad $\widehat{\sigma^2}_{JD}$ veľmi citliví. Riadky matice Q_{xz} boli v simulácii 3.1 formálne lineárne nezávislé, ale v skutočnosti boli takmer závislé. Predpoklad P9 sme síce dodržali, ale takpovediac „veľmi tesne“. Pozrieme sa ešte na ďalšie dva scenáre. Konkrétne, že aké výsledky dostaneme, ak Q_{xz} má jednoznačne lineárne nezávislé riadky, a naopak, ak predpoklad P9 porušíme.

Simulácia 3.3. Uvažujme rovnakú nulovú hypotézu a generovanie dát ako v simulácii 3.1, s tým rozdielom, že maticu Q_{xz} zvolíme inak. Zašumenie prvkov matice Q_{xz} v (50) zväčšíme, aby riadky matice Q_{xz} boli „jednoznačnejšie“ lineárne nezávislé. K prvkom matice Q_{xz} v (50) pripočítame náhodné číslo z rovnomerného rozdelenia na intervale $[-0{,}3, 0{,}3]$. Dostaneme tak maticu

$$Q_{xz} \doteq \begin{pmatrix} 0{,}40 & 0{,}52 & 0{,}26 & 0{,}36 & 0{,}17 & 0{,}36 \\ 0{,}28 & 0{,}15 & 0{,}26 & 0{,}49 & 0{,}46 & 0{,}13 \\ 0{,}36 & 0{,}24 & 0{,}47 & 0{,}44 & 0{,}31 & 0{,}26 \end{pmatrix}, \tag{67}$$

ktorá má výrazne lineárne nezávislé riadky. Ak na testovanie nulovej hypotézy použijeme testovú štatistiku F_{JD}, dostaneme odhad pravdepodobnosti chyby prvého druhu rovný 4,67%, blízky nominálnej hladine významnosti testu $\alpha = 5\%$. Priemer zo $\widehat{\sigma^2}_{JD}$ rovný 1,01 naznačuje, že odhad $\widehat{\sigma^2}_{JD}$ je v tomto prípade nevychýleným odhadom pre disperziu σ^2. Postup testovania pomocou F_{IVGNR} funguje stále ideálne, keďže odhad pravdepodobnosti chyby prvého druhu vyšiel rovný 5,24% a priemer z odhadov $\widehat{\sigma^2}_{IVGNR}$ je rovný 1,00.

Výsledky v simulácii 3.3 naznačujú, že ak má matica Q_{xz} definovaná v (8) hodnosť rovnú K a nie je „blízka“ matici, ktorá má hodnosť menšiu ako K, potom sú testové štatistiky F_{JD} a F_{IVGNR} rovnocenné, v zmysle, že majú rovnakú pravdepodobnosť chyby prvého druhu.

Simulácia 3.4. Tretí scenár pre testovanie nulovej hypotézy popísanej v simulácii 3.1 bude obsahovať maticu Q_{xz} s lineárne závislými riadkami. To predstavuje akúsi extrémnu situáciu, keď matica Q_{xz} nespĺňa predpoklad P9. Uvažujme teda maticu s lineárne závislými riadkami

$$Q_{xz} = \begin{pmatrix} 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \end{pmatrix}.$$

Potom test pomocou F_{JD} má odhad pravdepodobnosti chyby prvého druhu rovný 0,02%, čiže sa správa extrémne konzervatívne. Je to dôsledkom nadhodnoteného odhadu $\widehat{\sigma^2}_{JD}$, ktorého priemer z $S = 10000$ simulácií je rovný 1,97. Test pomocou F_{IVGNR} má odhad pravdepodobnosti chyby prvého druhu rovný 5,06%, to znamená, že funguje ideálne aj keď je predpoklad P9 porušený.

Ukazuje sa, že obe postupy testovania sú správne, ale testová štatistika F_{JD} je veľmi citlivá na to, či má matica Q_{xz} definovaná v (8) lineárne nezávislé riadky. Pokiaľ matica Q_{xz} má lineárne závislé riadky, tak sa testová štatistika F_{JD} správa extrémne konzervatívne. Pokiaľ ich má lineárne nezávislé ale takmer zhodné, tak test pomocou F_{JD} nefunguje ideálne napriek tomu, že všetky predpoklady modelu sú splnené.

Momentálne je fungovanie testov podložené simuláciami s konkrétnym nastavením parametrov lineárneho regresného modelu. Preto urobíme ďalší simulačný experiment s iným nastavením parametrov, pričom dáta budeme generovať z iného ako normálneho rozdelenia. K takémuto generovaniu dát nás motivoval článok [4].

Simulácia 3.5. Dáta v tejto simulácii budeme generovať novým spôsobom, ktorý nám umožní generovať pozorovania regresných a inštrumentálnych premenných z iného ako normálneho rozdelenia. Nesmieme zabudnúť, že náhodné chyby majú z predpokladu normálne rozdelenie $N(0, \sigma^2)$. Spôsob generovania dát môžeme popísať nasledovnými rovnicami

$$X = Z\pi + Z'\pi' + U, \tag{68}$$

$$y = X\beta + Z'\tau + V. \tag{69}$$

Náhodná matica Z' predstavuje maticu akýchsi pomocných inštrumentálnych premenných a má rozmer $n \times L'$. Matica π s rozmerom $L \times K$, matica π' s rozmerom $L' \times K$ a vektor τ s rozmerom $L' \times 1$ sú matice pevných parametrov a môžeme ich voliť ľubovoľne. Ich voľba však ovplyvní aj väzby medzi premennými modelu. Matica U s rozmerom $n \times K$ a vektor V s rozmerom $n \times 1$ predstavujú matice náhodných šumov. Z rovnice (69) vyplýva, že

$$\varepsilon = Z'\tau + V,$$

z čoho podľa (68) vidno, že ε koreluje s X ale nie so Z, ktoré naopak koreluje s X. Aby sme zachovali normálne rozdelenie náhodných chýb, vektor V a stĺpce matice Z' budeme generovať nezávisle ako náhodný výber z normálneho rozdelenia $N(0, 1)$. Zvoľme parameter $L' = 1$ a $\tau = 2$. Z toho vyplýva, že vektor náhodných chýb ε bude mať rozdelenie $N(0, 5)$. Matice π a π' nastavme ako matice samých jednotiek. Inštrumentálne premenné môžeme generovať z ľubovoľného rozdelenia, preto stĺpce matice Z budeme generovať nezávisle ako náhodný výber z rovnomerného rozdelenia $R(-1, 1)$. To znamená, že regresné premenné v matici X budú zmesou normálneho a rovnomerného rozdelenia. Stĺpce matice U budeme generovať nezávisle ako náhodný výber z normálneho rozdelenia $N(0, 1)$. Nastavme počet simulácií $S = 10000$ a počet pozorovaní $n = 1000$. Počet regresných premenných zvýšme na $K = 4$ a počet inštrumentálnych premenných na $L = 8$. Na takto generovaných dátach budeme testovať nulovú hypotézu

$$H_0 : \quad \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

pomocou testových štatistík F_{JD} a F_{IVGNR}. Kvôli porovnaniu chyby prvého druhu testov,

chceme dáta generovať za platnosti nulovej hypotézy, preto vektor β zvolíme

$$\beta = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Prepíšeme maticový zápis v (68) do vektorového tvaru

$$x^T = z^T\pi + z'^T\pi' + u^T, \tag{70}$$

kde $x = (x_1, ..., x_K)^T$, $z = (z_1, ..., z_L)^T$, $z' = (z'_1, ..., z'_{L'})^T$ a $u = (u_1, ..., u_K)^T$. Potom maticu Q_{xz} vieme zapísať ako

$$Q_{xz} = cov(x,z) + E(x).(E(z))^T.$$

Premenné modelu sme zvolili tak, že regresné a inštrumentálne premenné majú nulovú strednú hodnotu, preto

$$Q_{xz} = cov(x,z). \tag{71}$$

Ďalej dosadíme výraz v (70) do rovnice (71) a využijeme že z je nezávislé s u a z', potom

$$\begin{aligned} Q_{xz} &= cov(x,z) = cov(\pi^T z + \pi'^T z' + u,z) = \pi^T cov(z,z) + \pi'^T cov(z',z) + cov(u,z) \\ &= \pi^T cov(z,z) = \pi^T var(z). \end{aligned} \tag{72}$$

To znamená, že reguláciou π vieme meniť Q_{xz}. V našom prípade je matica Q_{xz} rovná

$$Q_{xz} = \pi^T var(z) = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 & 1/3 \end{pmatrix}, \tag{73}$$

keďže z má rozdelenie $R(-1,1)$, a teda varianciu rovnú 1/3. Riadky matice v (73) sú lineárne závislé, a teda podmienka P9 je porušená. Ak na testovanie nulovej hypotézy použijeme testovú štatistiku F_{JD}, dostaneme odhad pravdepodobnosti chyby prvého druhu na hladine významnosti $\alpha = 5\%$ rovný 0%. Priemer z odhadov $\widehat{\sigma^2}_{JD}$ vyšiel rovný 9,94, čo je výrazne vychýlené od skutočnej hodnoty disperzie $\sigma^2 = 5$. Výsledky znova naznačujú, že v dôsledku nadhodnoteného odhadu $\widehat{\sigma^2}_{JD}$ sa test správa konzervatívne a potvrdili

nefunkčnosť testu. Použitím testovej štatistiky F_{IVGNR} dostaneme odhad pravdepodobnosti chyby prvého druhu rovný 4,94% a priemer z odhadov $\widehat{\sigma^2}_{IVGNR}$ vyšiel rovný 4,98. To naznačuje, že testová štatistika F_{IVGNR} je vhodná na testovanie hypotéz o lineárnych ohraničení pre β aj v prípade, že regresné a inštrumentálne premenné nie sú z normálneho rozdelenia.

Ešte sa pozrieme na situáciu, keď je predpoklad P9 pre maticu Q_{xz} splnený. Maticu Q_{xz} v (73) vieme priamo meniť tým, že zmeníme maticu π. V každom kole simulácie zašumíme prvky matice π tak, že k nim pripočítame náhodné číslo z rovnomerného rozdelenia na intervale $[-0{,}8, 0{,}8]$. Dostaneme maticu Q_{xz}, ktorá má hodnosť $K = 4$, a teda má lineárne nezávislé riadky. Potom použitím testovej štatistiky F_{JD} dostaneme odhad pravdepodobnosti chyby prvého druhu rovný 4,6% a použitím testovej štatistiky F_{IVGNR} dostaneme odhad pravdepodobnosti chyby prvého druhu rovný 5,13%. Z výsledkov vidíme, že aj keď sú dáta z iného ako normálneho rozdelenia, tak asymptotické správanie testovej štatistiky F_{JD} sa zlepšilo a pravdepodobnosť chyby prvého druhu testu sa priblížila k nominálnej hladine významnosti testu $\alpha = 5\%$.

Autori v [1] a v [3] varujú čitateľov pred použitím matice $\widehat{X}$ pri výpočte rezíduí v odhade pre σ^2, čo sa na prvý pohľad môže zdať vhodnou voľbou. Neuvádzajú však, aký to má vplyv na testovanie hypotéz o β. V nasledovnej simulácii budeme skúmať vplyv odhadu

$$\widehat{\sigma^2} = \frac{(y - \widehat{X} b_{IV})^T (y - \widehat{X} b_{IV})}{n - K} \tag{74}$$

na správanie skúmaných testov.

Simulácia 3.6. Zopakujme simulačný experiment popísaný v simulácii 3.3. Videli sme, že obe testové štatistiky F_{JD} a F_{IVGNR} boli vhodné na testovanie danej nulovej hypotézy. Už sme ukázali, že F_{JD} a F_{IVGNR} sa líšia iba v menovateli, teda v odhade pre σ^2. Teraz v menovateli testovej štatistiky použijeme odhad $\widehat{\sigma^2}$ daný vzťahom (74). Potom odhad pravdepodobnosti chyby prvého druhu s presnosťou na 2 desatinné miesta je rovný 0,07% a priemer z odhadov $\widehat{\sigma^2}$ je rovný 2,44. Výsledky naznačujú, že $\widehat{\sigma^2}$ v (74) je vysoko nadhodnotený odhad pre $\sigma^2 = 1$, v dôsledku čoho sa test správa extrémne konzervatívne.

3.2 Asymptotický t-test

Na testovanie lineárneho ohraničenia pre jeden neznámy parameter sa používa asymptotický t-test. Postup testu je odvodený aj v [1], avšak v tomto prípade autori neaplikovali IVGNR metodiku na odvodenie testovej štatistiky. Dôsledkom toho je opäť nekonzistentný odhad pre σ^2 v prípade, keď riadky matice Q_{xz} definovanej v (8) sú blízke lineárne závislým vektorom. Na základe rešerše literatúry sme usúdili, že sa nenašli autori, ktorí by sa zaoberali riešením tohto problému. Preto na základe skúseností z predchádzajúcej kapitoly navrhneme úpravu testovej štatistiky. Fungovanie testov opäť ilustrujeme na umelo generovaných dátach v štatistickom softvéri R.

Na testovanie ohraničenia týkajúceho sa jedného neznámeho parametra β_i je najjednoduchší spôsob použiť asymptotický t-test. Uvažujme opäť obvyklý model lineárnej regresie (3) a odhad pomocou inštrumentálnych premenných b_{IV}. V predchádzajúcej kapitole sme tento odhad značili ako b_{IV1}, teraz kvôli zavedeniu indexovania prvkov odhadu použijeme značenie b_{IV}. Nech máme danú nulovú hypotézu

$$H_0: \quad \beta_i = \beta_0.$$

Bez ujmu na všeobecnosti predpokladajme, že testujeme nulové ohraničenie $\beta_i = 0$. Ako sa aj v [1] spomína, odhad b_{IV} má asymptotické normálne rozdelenie. Potom príslušná testová štatistika má tvar

$$T = \frac{b_{IV\,i} - 0}{\sqrt{Var\,(b_{IV})_{ii}}}, \tag{75}$$

kde $b_{IV\,i}$ je i-ty prvok odhadu pomocou inštrumentálnych premenných b_{IV} a $Var\,(b_{IV})_{ii}$ je i-ty diagonálny prvok variančnej matice odhadu b_{IV}. Testová štatistika (75) sa asymptoticky riadi štandardným normálnym rozdelením $N(0,1)$.

Ostáva definovať odhad variančnej matice pre odhad b_{IV}. V [1] a tiež v [3] je definovaný ako

$$Var\,(b_{IV}) = \widehat{\sigma^2}[X^T Z(Z^T Z)^{-1} Z^T X]^{-1} \tag{76}$$

a odhad pre σ^2 je definovaný ako

$$\widehat{\sigma^2}_{IV} = \frac{(y - Xb_{IV})^T (y - Xb_{IV})}{n}. \tag{77}$$

Odhad $\widehat{\sigma^2}_{IV}$ je až na menovateľ (chýba $-K$) zhodný s odhadom $\widehat{\sigma^2}_{JD}$ z (55).

Ak na testovanie nulovej hypotézy použijeme odhad disperzie (77), opäť narazíme na rovnaký problém ako v prípade testovania lineárnych hypotéz pomocou F_{JD}. Ak sú riadky matice Q_{xz} blízke lineárne závislým vektorom, potom odhad disperzie $\widehat{\sigma^2}_{IV}$ je vychýlený a tým pádom sa testová štatistika (75) neriadi asymptoticky rozdelením $N(0,1)$. Na vylepšenie asymptotického správania testovej štatistiky aplikujeme úpravu pre odhad disperzie, ktorú sme popísali v predchádzajúcej kapitole. Namiesto odhadu $\widehat{\sigma^2}_{IV}$ pre disperziu náhodných chýb navrhujeme použiť odhad $\widehat{\sigma^2}_{IVGNR}$ definovaný vzťahom (62). Vektor β' je definovaný ako konzistentný odhad pre β za platnosti nulovej hypotézy. To znamená, vektor β' závisí od daného ohraničenia pre parameter β_i. V nasledujúcej simulácii zavedieme explicitný vzorec pre β' pre prípad, keď testujeme nulové ohraničenie.

Simulácia 3.7. Uvažujme počet regresných premenných $K = 2$, počet inštrumentálnych premenných $L = 4$ a počet pozorovaní $n = 1000$. Dáta budeme generovať ako náhodný výber z viacrozmerného normálneho rozdelenia, pričom stredné hodnoty premenných modelu $(x_1, x_2, z_1, ..., z_4, \varepsilon)$ zvolíme rovné 0 a ich disperzie rovné 1. Kovariancie medzi premennými modelu nastavíme na

$$cov(x_i, x_j) = cov(z_i, z_j) = cov(x_i, \varepsilon) = 0{,}4, \qquad i \neq j$$

a kovariancie medzi náhodnou chybou ε a inštrumentálnymi premennými $z_1, ..., z_4$ sú z predpokladu nulové. Keďže uvažujeme nulové stredné hodnoty premenných modelu, potom matica Q_{xz} je rovná

$$Q_{xz} = cov\left(\begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix}, \begin{pmatrix} z_1 \\ \vdots \\ z_L \end{pmatrix}\right),$$

ktorá má mať podľa predpokladu P9 lineárne nezávislé riadky. Aby sme tento predpoklad dodržali, maticu Q_{xz} zvolíme tak, že riadky matice

$$\begin{pmatrix} 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \end{pmatrix} \tag{78}$$

náhodne zašumíme. Ku každému prvku matice v (78) pripočítame náhodné číslo z rovnomerného rozdelenia na intervale $[-0{,}05, 0{,}05]$. Zašumenie spôsobí, že riadky matice

$$Q_{xz} \doteq \begin{pmatrix} 0{,}37 & 0{,}41 & 0{,}44 & 0{,}36 \\ 0{,}42 & 0{,}37 & 0{,}44 & 0{,}43 \end{pmatrix}$$

sú lineárne nezávislé ale zároveň sú takmer zhodné s lineárne závislými vektormi. Formálne sú predpoklady modelu dodržané. Pomocou $S = 10000$ simulácií budeme testovať nulovú hypotézu

$$H_0 : \quad \beta_2 = 0. \tag{79}$$

Vektor β zvolíme tak, aby bola splnená nulová hypotéza, napr.

$$\beta = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Ak v testovej štatistike (75) použijeme odhad disperzie $\widehat{\sigma^2}_{IV}$, potom odhad pravdepodobnosti chyby prvého druhu testu na hladine významnosti $\alpha = 5\%$ je rovná 2,94%. Priemer z odhadov $\widehat{\sigma^2}_{IV}$ s presnosťou na 2 desatinné miesta je rovný 1,11. To naznačuje, že odhad $\widehat{\sigma^2}_{IV}$ je vychýlený od skutočnej disperzie $\sigma^2 = 1$. Test sa správa konzervatívne a je potrebné zvoliť iný odhad pre σ^2. Teraz použijeme odhad disperzie $\widehat{\sigma^2}_{IVGNR}$. Definujme submodel

$$y = X_1\beta_1 + \varepsilon, \tag{80}$$

ktorý predstavuje model spĺňajúci systém daných lineárnych ohraničení pre β. Kľúčové je vypočítať vektor β'. Vieme, že β' má byť konzistentný odhad pre β za platnosti nulovej hypotézy. Preto zvolíme

$$\beta' = \begin{pmatrix} b_1 \\ 0 \end{pmatrix},$$

kde b_1 je odhady pre β_1 pomocou inštrumentálnych premenných v modeli (80). Odhad $\widehat{\sigma^2}_{IVGNR}$ dopočítame na základe vzorca (62). Použitím upravenej testovej štatistiky dostaneme odhad pravdepodobnosti chyby prvého druhu testu na hladine významnosti $\alpha = 5\%$ rovný 4,97%. Priemer z $S = 10000$ odhadov disperzie $\widehat{\sigma^2}_{IVGNR}$ s presnosťou na 2 desatinné miesta je rovný 1,00. Nami navrhnutá testová štatistika obsahujúca odhad $\widehat{\sigma^2}_{IVGNR}$ dáva výrazne lepšie výsledky v porovnaní s testovou štatistikou obsahujúcou odhad $\widehat{\sigma^2}_{IV}$.

Teraz zväčšíme zašumenie riadkov matice Q_{xz} v (78) tak, že ku každému prvku tejto matice pripočítame náhodné číslo z rovnomerného rozdelenia na intervale $[-0{,}3, 0{,}3]$. Dostaneme tak maticu

$$Q_{xz} \doteq \begin{pmatrix} 0{,}21 & 0{,}44 & 0{,}67 & 0{,}18 \\ 0{,}52 & 0{,}20 & 0{,}67 & 0{,}60 \end{pmatrix},$$

ktorej riadky už nie sú blízke lineárne závislým vektorom. Potom pre testovú štatistiku T obsahujúcu odhad $\widehat{\sigma^2}_{IV}$ dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 4,97% a priemer z odhadov $\widehat{\sigma^2}_{IV}$ rovný 1,00. Výsledky naznačujú, že v tomto prípade je odhad $\widehat{\sigma^2}_{IV}$ nevychýleným odhadom pre σ^2. Ak v simuláciách použijeme odhad $\widehat{\sigma^2}_{IVGNR}$, tak test naďalej funguje ideálne. Pre testovú štatistiku T obsahujúcu odhad $\widehat{\sigma^2}_{IVGNR}$ dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 5,12%.

Na záver zopakujeme výpočty s maticou

$$Q_{xz} = \begin{pmatrix} 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \end{pmatrix},$$

ktorá má lineárne závislé riadky, a teda podmienka P9 je porušená. Pre testovú štatistiku s odhadom $\widehat{\sigma^2}_{IV}$ dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 0,28%, a teda test sa správa extrémne konzervatívne. Ak v simuláciách použijeme nami navrhnutú úpravu, teda použijeme odhad $\widehat{\sigma^2}_{IVGNR}$, dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 5,18%. Napriek tomu, že podmienka P9 je porušená, testová štatistika s odhadom $\widehat{\sigma^2}_{IVGNR}$ sa ukazuje byť vhodná na testovanie hypotézy o lineárnom ohraničení pre β.

3.3 Waldov test

Zovšeobecnením t-testu je Waldov test. Testovať lineárne ohraničenia pre viac ako jeden neznámy parameter modelu vieme aj pomocou Waldovej testovej štatistiky.

Predpokladajme, že chceme testovať nulovú hypotézu o lineárnych ohraničeniach (42) pre β. Waldova testová štatistika je definovaná ako

$$\begin{aligned} W &= (Rb_{IV} - r)^T \left(RVar\left(b_{IV}\right) R^T \right)^{-1} (Rb_{IV} - r) \\ &= \frac{(Rb_{IV} - r)^T \left(R[X^T Z (Z^T Z)^{-1} Z^T X]^{-1} R^T \right)^{-1} (Rb_{IV} - r)}{\widehat{\sigma^2}_{IV}} \end{aligned} \tag{81}$$

a za platnosti nulovej hypotézy má asymptoticky rozdelenie χ^2 s q stupňami voľnosti.

Rovnako ako v prípade testovania lineárnych hypotéz pomocou testovej štatistiky F_{JD} je aj testová štatistika W, resp. odhad $\widehat{\sigma^2}_{IV}$, citlivý na hodnosť matice Q_{xz}. Analogicky, často vychýlený odhad $\widehat{\sigma^2}_{IV}$ daný vzťahom (77) navrhujeme nahradiť odhadom $\widehat{\sigma^2}_{IVGNR}$ definovaný vzťahom (62). Ako to uvidíme v nasledujúcej simulácii, táto úprava odhadu

pre σ^2 „zlepší“ asymptotické správanie Waldovej testovej štatistiky (81), čiže za platnosti nulovej hypotézy bude bližšie k χ^2_q.

Simulácia 3.8. Zopakujme simulačné experimenty popísané v simuláciách 3.1-3.4, pričom na testovanie nulovej hypotézy použijeme Waldovu testovú štatistiku (81). V prvom simulačnom scenári má matica

$$Q_{xz} \doteq \begin{pmatrix} 0{,}40 & 0{,}42 & 0{,}38 & 0{,}39 & 0{,}36 & 0{,}39 \\ 0{,}38 & 0{,}36 & 0{,}38 & 0{,}42 & 0{,}41 & 0{,}36 \\ 0{,}39 & 0{,}37 & 0{,}41 & 0{,}41 & 0{,}39 & 0{,}38 \end{pmatrix}$$

lineárne nezávislé riadky, blízke lineárne závislým vektorom. To znamená, že všetky podmienky modelu, vrátane podmienky P9 sú formálne dodržané. Ak v testovej štatistike (81) použijeme odhad disperzie $\widehat{\sigma^2}_{IV}$, potom odhad pravdepodobnosti chyby prvého druhu testu na hladine významnosti $\alpha = 5\%$ je rovný 0,29% a priemer z odhadov $\widehat{\sigma^2}_{IV}$ s presnosťou na 2 desatinné miesta je rovný 1,43. Znova sa opakuje ten istý jav ako v predošlých nefunkčných testoch. Odhad disperzie $\widehat{\sigma^2}_{IV}$ je nadhodnotený a dôsledkom toho sa testová štatistika W neriadi ani približne rozdelením χ^2. Riešenie, ktoré navrhujeme, je použiť vo Waldovej testovej štatistike odhad disperzie $\widehat{\sigma^2}_{IVGNR}$. Výpočet vektora β' sme už popísali v simulácií 3.2. Použitím upravenej testovej štatistiky W dostávame zo simulácií odhad pravdepodobnosti chyby prvého druhu testu rovný 5,12%. Priemer z odhadov $\widehat{\sigma^2}_{IVGNR}$ s presnosťou na 2 desatinné miesta je rovný 1,00. To sa zhoduje so skutočnou hodnotou disperzie. Výsledky naznačujú, že odhad $\widehat{\sigma^2}_{IVGNR}$ je jednoznačne lepším odhadom pre σ^2.

V druhom simulačnom scenári zvolíme maticu

$$Q_{xz} \doteq \begin{pmatrix} 0{,}40 & 0{,}52 & 0{,}26 & 0{,}36 & 0{,}17 & 0{,}36 \\ 0{,}28 & 0{,}15 & 0{,}26 & 0{,}49 & 0{,}46 & 0{,}13 \\ 0{,}36 & 0{,}24 & 0{,}47 & 0{,}44 & 0{,}31 & 0{,}26 \end{pmatrix}$$

tak, aby mala jednoznačne lineárne nezávislé riadky. Potom pre Waldovu testovú štatistiku s odhadom $\widehat{\sigma^2}_{IV}$ dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 4,76% a pre Waldovu testovú štatistiku s odhadom $\widehat{\sigma^2}_{IVGNR}$ dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 5,3%. Obe prístupy testovania nulovej hypotézy sa ukazujú byť v tomto prípade vhodným.

Tretí simulačný scenár je navrhnutý tak, aby bola podmienka P9 porušená, a teda

maticu Q_{xz} zvolíme

$$Q_{xz} = \begin{pmatrix} 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \end{pmatrix}.$$

Dostávame analogické výsledky ako pre testovú štatistiku F_{JD}. Výsledky uvedieme s presnosťou na 2 desatinné miesta. Pre testovú štatistiku W s odhadom $\widehat{\sigma^2}_{IV}$ dostaneme odhad pravdepodobnosti chyby prvého druhu testu rovný 0,02% a priemer z odhadov $\widehat{\sigma^2}_{IV}$ rovný 1,97. Odhad $\widehat{\sigma^2}_{IV}$ vypočítaný v simuláciách je výrazne vychýlený od skutočnej hodnoty disperzie $\sigma^2 = 1$ a Waldov test sa správa extrémne konzervatívne. Ak vo Waldovej testovej štatistike W použijeme nami navrhnutý odhad $\widehat{\sigma^2}_{IVGNR}$, ukazuje sa, že test funguje ideálne. Odhad pravdepodobnosti chyby prvého druhu testu rovný 5,09% a priemer z odhadov $\widehat{\sigma^2}_{IVGNR}$ je rovný 1,00.

Výsledky naznačujú, že pri testovaní lineárnych hypotéz o β musíme obzvlášť dávať pozor nato, akú hodnosť má matica Q_{xz}. Odhad $\widehat{\sigma^2}_{IV}$ je vysoko citlivý na to, či má matica Q_{xz} lineárne nezávislé riadky. Ak sú riadky matice Q_{xz} „dokonale“ lineárne závislé alebo sú blízke lineárne závislým vektorom, potom Waldova testová štatistika W obsahujúca odhad $\widehat{\sigma^2}_{IV}$ nefunguje ideálne. V takom prípade nami navrhnutý odhad $\widehat{\sigma^2}_{IVGNR}$ sa javí byť lepším odhadom pre disperziu σ^2 a vylepší asymptotické správanie Waldovej testovej štatistiky, čiže za platnosti nulovej hypotézy bude bližšie k χ^2_q.

Je známe, že v klasickej lineárnej regresii s odhadom metódou najmenších štvorcov existuje súvis medzi Fisherovou−Snedecorovou štatistikou F a Waldovou štatistikou W. To nás motivovalo skúmať vzťah týchto testových štatistík aj v lineárnej regresii s endogénnymi regresnými premennými a odhadom pomocou inštrumentálnych premenných. Dospeli sme k záveru, že aj pre odhad pomocou inštrumentálnych premenných sa dajú odvodiť rovnaké cenné výsledky pre testovanie lineárnych hypotéz. Súvis medzi testovými štatistikami popíšeme v nasledujúcej vete, ktorú následne aj dokážeme.

Veta 3.1. *Nech sú dané testové štatistiky*

$$F = \frac{\dfrac{RSS_0 - RSS_1}{q}}{\widehat{\sigma^2}},$$

$$W = \frac{(Rb_{IV} - r)^T \left(R[X^T Z (Z^T Z)^{-1} Z^T X]^{-1} R^T\right)^{-1} (Rb_{IV} - r)}{\widehat{\sigma^2}},$$

kde RSS_0, RSS_1 boli definované v (44), (46). *Pre testové štatistiky F a W platí*

$$q.F = W. \tag{82}$$

Dôkaz. Na to, aby sme dokázali rovnosť v (82), treba dokázať, že

$$RSS_0 - RSS_1 = (Rb_{IV} - r)^T \Big(R[X^T Z(Z^T Z)^{-1} Z^T X]^{-1} R^T\Big)^{-1} (Rb_{IV} - r)\,.$$

Uvedomme si, že odhadovanie pomocou inštrumentálnych premenných v modeloch

$$y = X\beta + \varepsilon \quad \wedge \quad R\beta = r$$

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

je to isté, ako odhadovanie metódou najmenších štvorcov v modeloch

$$y = \widehat{X}\beta + \varepsilon \quad \wedge \quad R\beta = r$$

$$y = \widehat{X}\beta + \varepsilon.$$

To znamená, že metóda inštrumentálnych premenných je bežná metóda najmenších štvorcov, ak v modeli namiesto X použijeme $\widehat{X}$. Potom na základe známeho tvrdenia z [5] vieme, že

$$RSS_0 - RSS_1 = (Rb_{IV} - r)^T \Big(R\Big[\widehat{X}^T \widehat{X}\Big]^{-1} R^T\Big)^{-1} (Rb_{IV} - r).$$

Upravme ešte výraz

$$\begin{aligned}&\Big(R\Big[\widehat{X}^T \widehat{X}\Big]^{-1} R^T\Big)^{-1} = \Big(R\Big[X^T P_Z P_Z X\Big]^{-1} R^T\Big)^{-1} = \Big(R\Big[X^T P_Z X\Big]^{-1} R^T\Big)^{-1} \\ &= \Big(R[X^T Z(Z^T Z)^{-1} Z^T X]^{-1} R^T\Big)^{-1},\end{aligned}$$

a tým je daná veta dokázaná. □

Poznamenajme, že $\widehat{\sigma^2}$ v menovateli testových štatistík F a W označuje odhad pre σ^2 a aby platila rovnosť (82), treba ho zvoliť rovnaký pre obe testové štatistiky. Vieme, že testovú štatistiku F porovnávame s rozdelením $F_{q,n-K}$ a testovú štatistiku W porovnávame s rozdelením χ_q. Treba si ale uvedomiť, že $q.F_{q,n-K} \dot{\sim} \chi_q$ pre $n \to \infty$. To znamená, že testy pomocou F a W sú asymptoticky ekvivalentné.

Pomocou [5] sa dá ukázať ďalšie analogické tvrdenie, aké platí v klasickej lineárnej regresii s odhadom metódou najmenších štvorcov. Dá sa odvodiť explicitný vzorec

$$b_{IV0} = b_{IV1} - \Big(\widehat{X}^T \widehat{X}\Big)^{-1} R^T \Big[R\Big(\widehat{X}^T \widehat{X}\Big)^{-1} R^T\Big]^{-1} (Rb_{IV1} - r)$$

pre odhad b_{IV0} definovaný v (44). To znamená, že nám stačí vypočítať odhad b_{IV1} definovaný v (46) a na základe toho vieme dopočítať aj odhad b_{IV0}. Tento výsledok môže výpočty zjavne uľahčiť, hlavne v prípade, kedy testujeme nejaký zložitejší systém lineárnych ohraničení pre β.

Záver

Cieľom diplomovej práce bolo podrobne skúmať teóriu inštrumentálnych premenných a rozšíriť ju vlastnými zisteniami. To sme docielili pomocou výsledkov vlastných simulácií. Prácu sme rozdelili na tri kapitoly.

V úvode prvej kapitoly práce sme simulačne preukázali vychýlenosť odhadu metódou najmenších štvorcov, ak regresné premenné korelujú s náhodnou chybou a tiež sme vysvetlili, ktoré vlastnosti premenných majú vplyv na veľkosť tejto výchylky. Následne sme predstavili odhad pomocou inštrumentálnych premenných a to najskôr pre prípad, keď máme rovnaký počet inštrumentálnych premenných ako regresných. Skúmali sme konzistentnosť tohto odhadu, ktorá vo všeobecnosti nie je zabezpečená. Prínosom tejto časti práce bolo aj vlastné zdôvodnenie nekonzistentného správania odhadu. Sformulovali sme Vetu 1.5, ktorá hovorí, že odhad pomocou inštrumentálnych premenných je len špeciálnym prípadom dvojstupňovej metódy najmenších štvorcov a následne sme túto vetu dokázali. Ďalej sme sa venovali vlastnostiam odhadu metódou inštrumentálnych premenných. Simulačne sme skúmali, ako počet pozorovaní, počet inštrumentálnych premenných či väzba medzi regresnými a inštrumentálnymi premennými ovplyvní odhad a vysvetlili sme súvis výsledkov s teoretickými poznatkami. V závere kapitoly sme analyzovali model obsahujúci chybné merania regresných premenných. Limitné vlastnosti odhadu metódou najmenších štvorcov sme overili na simuláciách v R.

Druhá kapitola práce bola zameraná na testy špecifikácie. Porovnali sme správanie viacerých testov pomocou odhadu pravdepodobnosti chyby prvého druhu a takisto pomocou empirickej sily testov, čiže fungovanie za neplatnosti nulovej hypotézy. Hausmanov test sme navrhli vylepšiť bootstrapovou metódou, ktorú odporúčame použiť najmä v prípade malého počtu pozorovaní. Na základe výsledkov vlastných simulácií sme ukázali, že Hausmanov test a Wuov test sú asymptoticky ekvivalentné.

V poslednej kapitole práce sme sa venovali testovaniu lineárnych hypotéz o parametroch regresie. Zvlášť sme sa venovali klasickému F-testu pomocou submodelov, Waldovho testu a asymptotickému t-testu z [2]. Pre rôzne scenáre sme pomocou odhadu pravdepod. chyby prvého druhu skúmali, či sa testové štatistiky riadia príslušnými rozdeleniami. Narazili sme na situáciu, keď testy nefungovali ideálne a správali sa extrémne konzervatívne. Dôvod, ktorý sme si všimli, bol, že odhad disperzie náhodnej chyby je nadhodnotený,

ak matica zmiešaných momentov regresorov a inštrumentálnych premenných má riadky „blízke“ lineárne závislým vektorom. Pre klasický F-test sme našli v [1] vylepšenie pomocou metódy IVGNR, ktoré tento problém vyriešilo, ale treba poznamenať, že autor problém nespomína. Odvodili sme, v čom sa testová štatistika pomocou IVGNR líší od klasickej testovej štatistiky. Pre Waldov test a asymptotický t-test sme navrhli podobné vylepšenie pomocou metódy IVGNR, ktoré zlepšilo asymptotické správanie testových štatistík. Hlavný problém, ktorý sme riešili, bola extrémna konzervatívnosť testov z [2] aj keď všetky podmienky regresného modelu boli formálne dodržané. Nami naštudované zdroje sa týmto problémom nezaoberali. Skúmali sme pravdepod. chyby prvého druhu, čiže správanie za platnosti nulovej hypotézy. Ako možné rozšírenie tejto práce by sa ďalej dala skúmať sila testov, čiže správanie za neplatnosti nulovej hypotézy. Na záver sme vo Vete 3.1 sformulovali súvis medzi klasickým F-testom pomocou submodelov a Waldovým testom obsahujúce odhad metódou inštrumentálnych premenných a následne sme ju dokázali pomocou poznatkov z [5].

Zoznam použitej literatúry

[1] Davidson, R., MacKinnon, J. G.: *Econometric Theory and Methods*, Oxford University Press, New York, 2004

[2] DiNardo, J., Johnston, J.: *Econometric Methods. 4th Edition*, McGraw-Hill, New York, 1997

[3] Greene, W. H.: *Econometric Analysis. 7th Edition*, Prentice Hall, Upper Saddle River, 2012

[4] Nakamura, A., Nakamura, M.: *On the Relationships Among Several Specification Error Tests Presented by Durbin, Wu, and Hausman*, Econometrica 49 (1981), 1583–1588

[5] Rosa, S.: *Ekonometria*, cvičenia, FMFI UK, Bratislava, 2019

[6] Wong, K.: *Bootstrapping Hausman's exogeneity test*, Economics Letters 53 (1996), 139–143