Metody Ekonometryczne
Endogeniczność i Metoda Zmiennych Instrumentalnych (IV)

Jakub Mućk
Szkoła Główna Handlowa w Warszawie

Enodgeniczność

Założenia:

1. $rz(\mathbf{X}) = k + 1 \leq n$
2. $\mathbb{E}(\mathbf{X}\varepsilon) = 0$
3. $\mathbb{E}(\varepsilon) = 0$
4. $Var(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon^T) = I\sigma^2$
5. $\varepsilon_i \sim N(0, \sigma^2)$

Twierdzenie Gaussa - Markowa

Estymator $\hat{\beta}$ uzyskany Klasyczną Metodą Najmniejszych Kwadratów jest estymatorem **BLUE** [***best linear unbiased estimator***], tj. zgodnym, nieobciążonym i najefektywniejszy w klasie liniowych estymatorów wektora β.

Założenia:

1. $rz(\mathbf{X}) = k + 1 \leq n$
2. $\mathbb{E}(\mathbf{X}\varepsilon) = 0$
3. $\mathbb{E}(\varepsilon) = 0$
4. $Var(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon^T) = I\sigma^2$
5. $\varepsilon_i \sim N(0, \sigma^2)$

Twierdzenie Gaussa - Markowa

Estymator $\hat{\beta}$ uzyskany Klasyczną Metodą Najmniejszych Kwadratów jest estymatorem **BLUE** [***best linear unbiased estimator***], tj. zgodnym, nieobciążonym i najefektywniejszy w klasie liniowych estymatorów wektora β.

- Zgodnie z założeniem (2) zmienne objaśniające są egzogeniczne.
- Endogeniczność polega na niespełnieniu warunku (2). W tym przypadku

$$\mathbb{E}(\mathbf{X}\varepsilon) \neq 0. \tag{1}$$

Zmienna Endogeniczna

Zmienna objaśniająca jest **endogeniczna (*endogenous*)** jeżeli jest skorelowana ze składnikiem losowym, tj., $\mathbb{E}(\varepsilon|x) \neq 0$.

Niezgodność i obciążoność estymatora MNK

Endogeniczność prowadzi do niezgodności estymatora MNK i jego obciążoności.

Zmienna Endogeniczna

Zmienna objaśniająca jest **endogeniczna (*endogenous*)** jeżeli jest skorelowana ze składnikiem losowym, tj., $\mathbb{E}(\varepsilon|x) \neq 0$.

Niezgodność i obciążoność estymatora MNK

Endogeniczność prowadzi do niezgodności estymatora MNK i jego obciążoności.

Standardowe przypadki kiedy zmienna objaśniająca jest endogeniczna:

1. Błąd pomiaru (*measurement error*).
2. Problem pominiętych zmiennych (*omitted variable bias*).
3. Współprzyczynowość (*simultaneity causality*).

- Załóżmy, że **prawdziwy** DGP (*data generating process*) dla konsumpcji (c) można zapisać następująco:
$$c = \alpha + \beta inc^* + \varepsilon, \tag{2}$$
gdzie inc^* oznacza **dochód permanentny.**
- Zazwyczaj posiadamy dane o inc, ale nie o **dochodzie permanentnym**. W takim przypadku staramy się aproksymować **dochód permanentny** dochodem bieżącym.
$$inc^* = inc + \eta, \tag{3}$$
gdzi eη to losowy błąd pomiaru $\eta \sim \mathcal{N}\left(0, \sigma_\eta^2\right)$.
- Bieżący dochód (inc) jest zatem **proxy variable** dochodu permenentnego (inc^*).
- Wykorzystując prawdziwy proces generujący dane (2):
$$c = \alpha + \beta\left(inc + \eta\right) + \varepsilon = \alpha + \beta inc + \beta\eta + \varepsilon = \alpha + \beta inc + \nu, \tag{4}$$
gdzie $\nu = \varepsilon + \beta\eta$.
- Kluczowe jest tutaj zbadanie kowariancji między dochodem bieżącym inc a składnikiem losowym (ν):
$$\mathrm{cov}(inc, \nu) = \mathbb{E}\left(inc\nu\right) = \mathbb{E}\left((inc^* - \eta)(\varepsilon + \beta\eta)\right) = \mathbb{E}\left(\beta\eta^2\right) = -\sigma_\eta^2\beta \neq 0. \tag{5}$$
- Zatem **błąd pomiaru (*measurement error*)** może prowadzić do endogeniczności.

- **Ekonomia pracy (labor economics)**: zwrot z edukacji.
- Załóżmy, że **prawdziwy** proces generujący dane dla logarytmu płac (w) to:

$$w = \alpha + \rho\mathcal{S} + \beta\mathcal{A} + \varepsilon, \tag{6}$$

gdzie $\mathcal{S}$ to stopień wykształcenia (lub liczba lat edukacji), a $\mathcal{A}$ mierzy zdolności indywidualne oraz motywcje.

- **Problem:** dane o $\mathcal{A}$ są niedostępne.
- Rozważmy jednak bardziej aplikacyjną (i przez to uproszczoną wersję) zależności (7):

$$w = \alpha + \rho\mathcal{S} + \eta, \tag{7}$$

gdzie składnik losowy η uwzględnia indywidualne umiejętności i predyspozycje ($\mathcal{A}$), tj.

$$\eta = \varepsilon + \beta\mathcal{A} \tag{8}$$

- Wtedy estymator MNK dla ρ, a więc zwrotu z edukacji, może zostać uproszczony do :

$$\hat{\rho}^{OLS} = \text{cov}(w, \mathcal{S})/Var(\mathcal{S}). \tag{9}$$

- Wykorzystując jednak *prawdziwy* proces generujący dane dla płac w:

$$\hat{\rho}^{OLS} = \frac{\text{cov}(\alpha + \rho\mathcal{S} + \beta\mathcal{A} + \varepsilon, \mathcal{S})}{Var(\mathcal{S})}, \tag{10}$$

- Po przekształceniach uzyskujemy:

$$\hat{\rho}^{OLS} = \frac{1}{Var(\mathcal{S})}\mathbb{E}\left[\left(\alpha + \rho\mathcal{S} + \varepsilon\right)\mathcal{S} + \beta\mathcal{A}\mathcal{S}\right] = \rho + \underbrace{\beta\frac{\text{cov}(\mathcal{A},\mathcal{S})}{Var(\mathcal{S})}}_{=\text{bias}} \neq \rho. \tag{11}$$

- Estymator MNK zwrotu z edukacji ($\hat{\rho}^{OLS}$) będzie zatem obciążony w górę (*upward biased* jeżeli znaki prawdziwego parametru β i komponent $\text{cov}(\mathcal{A},\mathcal{S})/Var(\mathcal{S})$ są takie same.

- Przykład: **(Keyensowski)model konsumpcji**:

$$c = \alpha + \beta y + \varepsilon \tag{12}$$
$$y = c + i \tag{13}$$

gdzie c to konsumpcja, y odpowiada zagregowanemu produktowi, i to inwestycje a ε to składnik losowy, t.j., $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

- W powyższym systemie równań mamy dwie zmienne endogeniczne (c oraz y) i jedną zmienną egzogeniczną (i).
- **Postać zredukowana (*reduced form*)** to taka postać modelu, w której zmienne endogeniczne są determinowane jedynie przez zmienne egzogeniczne oraz zaburzenia losowe. W naszym przypadku:

$$\begin{aligned} y &= c + i \\ y &= \alpha + \beta y + \varepsilon + i \\ (1-\beta)y &= \alpha + i + \varepsilon \\ y &= \frac{\alpha}{(1-\beta)} + \frac{1}{(1-\beta)} i + \frac{1}{(1-\beta)} \varepsilon. \end{aligned}$$

Współprzyczynowość II

SGH

- Ogólny zapis estymatora MNK dla krańcowej skłonności do konsumpcji (β) można wyprowadzić korzystając z tożsamości (12):

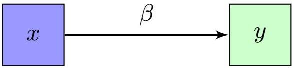
$$\hat{\beta}^{OLS} = \beta + \underbrace{\frac{\sum (y - \bar{y})\,\varepsilon}{\sum (y - \bar{y})^2}}_{=0 \text{ jezeli } \mathbb{E}(y|\varepsilon)=0} . \qquad (14)$$

- Ale z formy zredukowanej naszego modelu wiemy, że y zależy od ε. W takim przypadku $\hat{\beta}^{OLS} \neq \beta$, a więc estymator MNK będzie niezgodny.

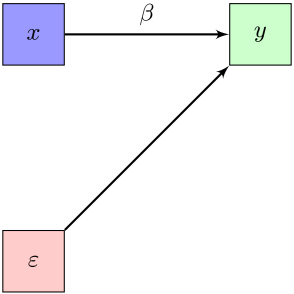
Endogeniczność jest również charakterystyczna dla estymatora MNK w przypadku specyficznych modeli ekonometrycznych:

- wielorównaniowych (jednoczesnych) modeli ekonometrycznych (*simultaneous equations model*).
- model Koycka.
- dynamicznych modeli panelowych z efektami indywidualnymi.

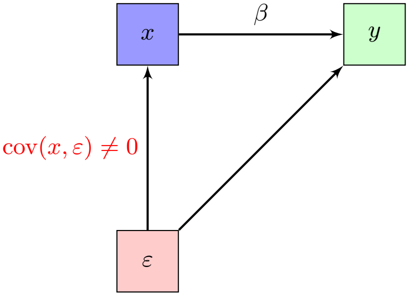
Metoda Zmiennych Instrumentalnych



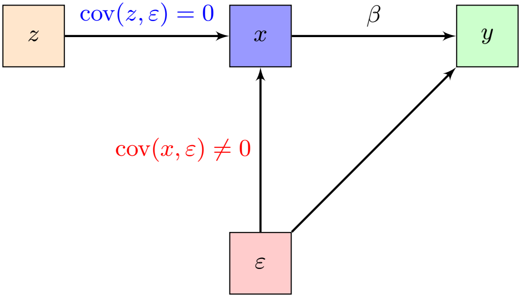
- x – zmienna objaśniająca;
- y – zmienna objaśniana;



- x – zmienna objaśniająca;
- y – zmienna objaśniana;
- ε – składnik losowy;



- x – zmienna objaśniająca;
- y – zmienna objaśniana;
- ε – składnik losowy;



- x – zmienna objaśniająca;
- y – zmienna objaśniana;
- ε – składnik losowy;
- z – zmienna instumentalna (instrument).

- Rozważmy prostą zależność liniową:

$$y = \alpha + \beta x + \varepsilon \quad \text{and} \quad \text{cov}(\varepsilon, x) \neq 0. \tag{15}$$

- Oszacowania uzyskane MNK β będą niezgodne.
- **Metoda zmiennych instrumentalnych (*Instrumental variables* IV)** pozwala *podzielić* zmienność endogenicznej zmiennej (x) na dwa komponenty:
  1. komponent, który potencjalnie **nie** jest skorelowany ze składnikiem losowym (ε),
  2. komponent, który może być skorelowany ze składnikiem losowym (ε).
- Powyższy zabieg jest możliwy dzięki wykorzystaniu **zmiennej instrumentalnej (instrument, z)**, która nie jest skorelowana ze składnikiem losowym ε.
- Instrument (z) pozwala zatem na zidentyfikowanie tej części zmienności zmiennej endogenicznej, która nie jest skorelowana ze składnikiem losowym ε i dzięki temu można uzyskać zgodne oszacowania β.

Zmienna instrumentalna powinna spełniać następujące warunki:

- **[Instrument relevance] Zmienna instrumetalna powinna być mocna.**, tj.

$$\operatorname{cov}(z, x) \neq 0 \tag{16}$$

  gdzie z to instrument, a x to zmienna endogeniczna.
  Oznacza to, że instrumenty powinny być skorelowane ze zmienną endogeniczną.

- **[Instrument exogeneity] Zmienna instrumetalna powinna być egzogeniczna.** tj.

$$\operatorname{cov}(z, \varepsilon) = 0 \tag{17}$$

  a więc zmienna instrumentalna **nie** może być skorelowana ze składnikiem losowym.

Podwójna MNK

- **Postać strukturalna:**
$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{18}$$
gdzie x to zmienna endogeniczna.
- Zakładamy, że posiadamy zmienną instrumentalną z.
- **Krok pierwszy** to regresja x zmiennej endogenicznej względem instrumentu z:
$$x = \pi_0 + \pi_1 z + \eta. \tag{19}$$
Na podstawie oszczowań MNK parametrów wektora π wyznaczamy wartości teoretyczne dla zmiennej endogenicznej, tj. $\hat{x}$., tj.
$$\hat{x} = \hat{\pi}_0 + \hat{\pi}_1 z. \tag{20}$$
- **Krok pierwszy** podwójnej metody najmniejszych kwadratów jest nazwyany **formą zredukowaną (*reduced form*)** dla zmiennej instrumentalnej.
- **Krok drugi** polega na wykorzystaniu liniowej projekcji z postaci zredukowanej, a więc $\hat{x}$ w szacowaniu parametrów **postaci strukturalnej**:
$$y = \beta_0 + \beta_1 \hat{x} + \varepsilon. \tag{21}$$
- Estymatorem podwójnej metody najmniejszych kwadratów $\hat{\beta}^{2SLS}$ będą zatem parametry powyższego modelu szacowane MNK.

- **Podwójna metoda najmniejszych kwadratów (2SLS, *Two Stages Least Squares*)** polega na dwustopniowym szacowaniu wektora nieznanych parametrów β:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon, \tag{22}$$

gdzie x_i to zmienne endogeniczne, a w_j to zmienne egzogeniczne.

- Zakładamy, że mamy k zmiennych endogenicznych, m instrumentów i r zmiennych egzogenicznych.

- **Krok pierwszy** to regresja/e x_i zmiennych endogenicznych względem instrumentów $(z_1, \ldots, z_m)$ oraz pozostałych zmiennych egzogenicznych $(w_1, \ldots, w_r)$:

$$\forall_{i \in 1,\ldots,k} \quad x_i = \pi_0 + \pi_1 z_1 + \ldots + \pi_m z_m + \pi_{m+1} w_1 + \ldots + \pi_{m+r} w_r + \eta. \tag{23}$$

Na podstawie oszcowań MNK parametrów wektora π wyznaczamy wartości teoretyczne dla endogenicznych, tj. $\hat{x}_i$., tj.

$$\forall_{i \in 1,\ldots,k} \quad \hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \ldots + \hat{\pi}_m z_m + \hat{\pi}_{m+1} w_1 + \ldots + \hat{\pi}_{m+r} w_r. \tag{24}$$

- **Krok pierwszy** podwójnej metody najmniejszych kwadratów jest nazwany **formą zredukowaną (*reduced form*)** dla zmiennej instrumentalnej.

- **Krok drugi** polega na wykorzystaniu liniowej projekcji z postaci zredukowanej, a więc $\hat{x}_i$ w szacowaniu parametrów **postaci strukturalnej**:

$$y = \beta_0 + \beta_1 \hat{x}_1 + \ldots + \beta_k \hat{x}_k + \beta_{k+1} w_1 + \ldots + \beta_{k+r} w_r + \varepsilon. \tag{25}$$

- Estymatorem podwójnej metody najmniejszych kwadratów $\hat{\beta}^{2SLS}$ będą zatem parametry modelu (25) szacowane MNK.

W metodzie zmiennych instumentalnych parametry przy zmiennych endogenicznych będą:

1. **Dokładnie identyfikowalne (*exactly identified*)** jeżeli $m = k$.
2. **Nadmiernie identyfikowalne (*overidentified*)** jeżeli $m > k$.
3. **Brak identyfikowalności (*underidentified*)** będzie równoznaczny sytuacji, gdy $m < k$.

gdzie:
k – liczba zmiennych endogenicznych,
m – liczba instrumentów.

- Dla uproszczenia: 1 zmienna endeogeniczna (x) oraz 1 zmienna instrumentalna (z).
- Wtedy asymptotyczne obciążenie estymatorów OLS i 2SLS(IV):

$$\text{plim}\left(\hat{\beta}^{OLS}\right) = \beta + cor(x,e)\frac{\sigma_e}{\sigma_x}, \tag{26}$$

$$\text{plim}\left(\hat{\beta}^{IV}\right) = \beta + \frac{cor(z,e)}{cor(z,x)}\frac{\sigma_e}{\sigma_z}. \tag{27}$$

- Zatem, estymator IV jest zgodny i nieobciążony, gdy:

$$\frac{cor(z,e)}{cor(z,x)} = 0, \tag{28}$$

  a więc gdy instrumenty są **mocne** i **egzogeniczne**.
- W przypadku słabych instrumentów, estymator IV charakteryzuje się wysoką wariancją.

- Kluczowym założeniem 2SLS, czy szerzej IV, jest korelacja zmiennych endogenicznych z instrumentami. Ujmując innymi słowami, instrumenty muszą być mocne.
- Moc instrumentów (*instrument relevance*) można analizować za pomocą statystyki $\mathcal{F}$ testu Walda dla regrrsji w pierwszym kroku:

$$x_i = \pi_0 + \pi_1 z_1 + \ldots + \pi_m z_m + \pi_{m+1} w_1 + \ldots + \pi_{m+r} w_r + \eta. \tag{29}$$

analizując hipotezę:

$$\mathcal{H}_0 : \qquad \pi_1 = \ldots = \pi_m = 0. \tag{30}$$

- Jako *rule of thumb* przyjmuje się wartość 10 dla statystyki $\mathcal{F}$.

- **Test Hausmana** pozwala na porównanie dwóch estymatorów:
  - β^1 charakteryzuje się wyższą efektywnością (niższą wariancją),
  - β^2 charakteryzuje się niższą efektywnością.
- **Hipoteza zerowa** testu Hausmana postuluje zgodność **obu estymatorów**.
- Natomiast w **hipotezie alternatywnej** zakłada się **brak zgodności estymatora o wyższej efektywności** β^1 przy jednoczesnej zgodności estymatora o niższej efektywności, β^2.
- Formalna statystyka testu Hausmana:

$$\mathcal{H} = \left(\beta^2 - \beta^1\right)^T \left[Var(\beta^2) - Var(\beta^1)\right]^{-1} \left(\beta^2 - \beta^1\right), \qquad (31)$$

  ma rozkład χ^2 z n stopniami swobody odpowiadającymi rzędowi macierzy $Var(\beta^2) - Var(\beta^1)$, a więc zazwyczaj (ale nie zawsze) liczbie parametrów wektora β.
- Statystyką Hausmana $\mathcal{H}$ można zapytać czy różnice pomiedzy estymatorami są statystycznie istotne.
- W przypadku metody zmiennych instrumentalnych można porównać oszacowania 2SLS z oszacowaniami OLS aby zbadać endogeniczność zmiennych.
  - Hipoteza alternatywna odnosi się do sytuacji, w której estymator OLS jest niezgodny (endogeniczność) a estymator 2SLS (IV) jest zgodny.
  - Warto tutaj zauważyć, że estymator 2SLS będzie zawsze mniej efektywny niż estymator OLS.
  - Odrzucenie $\mathcal{H}_0$ może wskazywać na problem endogeniczności.

- W przypadku gdy liczba instrumentów jest większa od liczby zmiennych endogenicznych, tj. jeżeli $m > k$, można zbadać zasadność (*validity*) nadmiarowych restrykcji.
- Procedura jest następująca:
  1. Oszacuj parametry strukturalnego równania (modelu dla dugiego kroku). Wyznacz reszty $\hat{e}$.
  2. Oszacuj parametry regresji pomocnicznej, w której zmienną objaśnianą są reszty $\hat{e}$, a zmiennymi objąśniającymi wszystkie zmienne egzogeniczne.
  3. Na podstawie R^2 można wyznaczyć statystykę testową, tj. $nR^2 \sim \chi^2_{m-k}$.

Zmienna objaśniana	Endogeniczna zmienna objaśniająca	Zmienna instrumentalna	Badanie
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Crime	Police	Electoral cycles	Levitt (1997)

- **Problem doboru instrumentów**.
- **Utrata efektywności instrumentów**.
  - Mniejsza precyzja oszacowań jest w szczególności znacząca gdy instrumenty są słabe.
- **Obciążoność (*bias*) oszacowań IV może być znacząca w małych próbach.**
- **Niesferyczność macierzy wariancji-kowariancji**
  - Heteroskedastyczność/ autokorelacja składnika losowego powinna być uwzględniona w oszacowaniach IV. W przeciwnym przypadku, testy statystyczne służące diagnostyce założeń IV mogą być niewiarygodne.