Zaawansowana Ekonometria

Paweł Struski

Uniwersytet Warszawski

2 marca 2026

Plan na dzisiaj

- Metoda Zmiennych Instrumentalnych (MZI) z ang. *Instrumental Variables (IV)*

Wstęp do MZI: zmienne endogeniczne

Rozważmy klasyczny model regresji liniowej z jedną zmienną:

$$y = \beta_0 + \beta_1 x + u$$

Definicja: Estymator $\hat{\beta}$ jest **zgodny** (z ang. *consistent*) jeśli ten estymator jest zbieżny według prawdopodobieństwa do prawdziwej wartości parametru tzn.: $\text{plim}_{n\to\infty} \hat{\beta} = \beta$

Jeśli cov(x, u) = 0, to estymator MNK jest zgodny.

Jeśli cov(x, u) $\neq$ 0, to:

- mówimy, że zmienna x jest **endogeniczna**,
- estymator MNK **nie** jest zgodny.

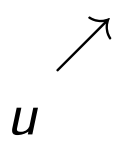
Wstęp do MZI: przyczyny endogeniczności

Przyczyny endogeniczności zmiennej x mogą być m.in. następujące:

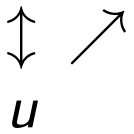
1. Błąd pomiarowy zmiennej x
2. Pominięcie innej istotnej zmiennej objaśniającej x_p (omitted variable), która jest skorelowana zarówno ze zmienną x jak i zmienną objaśnianą y.
3. Symultaniczność: występuje sprzężenie zwrotne między zmiennymi x i y, czyli gdy zmiana y prowadzi do zmian wartości x.

MZI: główna idea

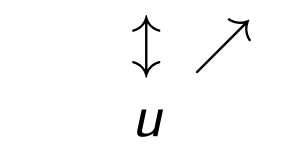
MNK z egzogenicznymi zmiennymi:



MNK z problemem pominiętych zmiennych:



MZI: załóżmy, że istnieje zmienna z, taka że:



Założenia wymagane dla instrumentu (jedna zmienna)

$$y = \beta_0 + \beta_1 x_1 + u$$

Zmienna x_1 jest endogeniczna, używamy instrumentu z_1

1. Istotność (*relevance*): $cov(z_1, x_1) \neq 0$
2. Egzogeniczność: $cov(z_1, u) = 0$

Dwustopniowa Metoda Najmniejszych Kwadratów (2MNK / 2SLS)

Równanie strukturalne:

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \ldots + \beta_k x_k}_{\text{zmienne } \textbf{endogeniczne}\text{: } \mathrm{Cov}(x_j,u)\neq 0} + \underbrace{\beta_{k+1} x_{k+1} + \ldots + \beta_{k+r} x_{k+r}}_{\text{zmienne } \textbf{egzogeniczne}\text{: } \mathrm{Cov}(x_j,u)=0} + u$$

Etap I – Regresja pomocnicza (dla każdej zmiennej endogenicznej x_j):

$$x_j = \delta_0 + \delta_1 x_{k+1} + \ldots + \delta_r x_{k+r} + \delta_{r+1} z_1 + \ldots + \delta_{r+m} z_m + v_j$$

gdzie $z_1, \ldots, z_m$ to **instrumenty** (zmienne wykluczone z równania strukturalnego).
Wyznaczamy wartości dopasowane $\hat{x}_j$.

Etap II – Regresja właściwa:

$$y = \beta_0 + \beta_1 \hat{x}_1 + \ldots + \beta_k \hat{x}_k + \beta_{k+1} x_{k+1} + \ldots + \beta_{k+r} x_{k+r} + e$$

Zadanie 1

Rozważano prosty model, w którym średnią uczniów kończących liceum (GPA) wyjaśniano za pomocą informacji o tym, czy posiadają oni komputer czy nie (zmienna zerojedynkowa PC).

$$GPA = \beta_0 + \beta_1 PC + \epsilon$$

1A. Dlaczego zmienna PC prawdopodobnie jest skorelowana z błędem losowym?

1B. Rozważ czy prawdopodobnym jest, że fakt posiadania komputera jest powiązany z dochodem rodziców ucznia. Czy to oznacza, że dochód rodziców jest dobrym kandydatem na zmienną instrumentalną?

Zadanie 2

Skonstruowano model dla logarytmu wysokości zarobków w zależności od poziomu edukacji (*wykształcenie*), zdolności danej osoby (*zdolności*) oraz informacji o tym czy członkowie rodziny danej osoby zasiadają w zarządzie firmy (zmienna zero-jedynkowa *koneksje*).

$$log(zarobki) = \beta_0 + \beta_1 wyksztalcenie + \beta_2 zdolnosci + \beta_3 konkesje + \epsilon_1$$

Z powodu braku dostępu do danych o poziomie zdolności pominięto tą zmienną w modelu i oszacowano prostszy model:

$$log(zarobki) = \beta_0 + \beta_1 wyksztalcenie + \beta_3 konkesje + \epsilon_2$$

2A. Jakie będą konsekwencje pominięcia w modelu zmiennej *zdolności*?

Zadanie 2 c.d.

W celu oszacowania modelu zdecydowano się wykorzystać metodę zmiennych instrumentalnych.

2B. Wskaż które zmienne w modelu są endogeniczne, a które egzogeniczne.

Zmienna endogeniczna = zmienna objaśniająca skorelowana z błędem losowym
Zmienna egzogeniczna = zmienna objaśniająca nieskorelowana z błędem losowym

Zadanie 2 c.d.

2C. Rozważ zasadność zaproponowanych instrumentów. Jakie warunki powinniśmy sprawdzić?

- Liczba lat edukacji ojca
- dzień urodziń
- wynik testu IQ

Zadanie 3

Oszacowano model za pomocą metody zmiennych instrumentalnych i uzyskano następujące oszacowania:

$$\underset{}{lwage} = \underset{.486}{-1.434} + \underset{.014}{.082}\,exper + \underset{.026}{.227}\,educ$$

gdzie *lwage* oznacza wysokość godzinowej stawki, *exper* oznacza liczbę lat doświadczenia w zawodzie, *educ* oznacza liczbę lat edukacji. Zakładamy, że zmienna *exper* jest egzogeniczna, a zmienna *educ* jest endogeniczna.

3A. Jako zmienne instrumentalne wykorzystano liczbę lat edukacji matki (motheduc), liczbę lat edukacji ojca (fatheduc) oraz liczbę rodzieństwa (sibs). Jaką postać ma regresja pomocnicza? Jaką hipotezę należy przetestować, aby sprawdzić czy *motheduc*, fatheduc i *sibs* to dobrzy kandydaci na instrumenty?

Zadanie 3 c.d.

Następnie przeprowadzono testy diagnostyczne i uzyskano wyniki:
Hausman-Wu F(1,1226) = 17.8689 (p = 0.0000)
Sargan (score) chi2(2) = 3.53983 (p = 0.1703)

3B. Jaka jest H0 w teście Hausmana-Wu? Jaki wniosek należy wyciągnąć na podstawie uzyskanego p-value?
3C. Jaka jest H0 w teście Sargana? Jaki wniosek należy wyciągnąć na podstawie uzyskanego p-value?

Zadanie 4

Badacze zastanawiali się nad związkiem między wagą dziecka przy urodzeniu a paleniem papierosów przez matkę podczas ciąży. Oszacowano następujący model za pomocą MNK (model 1) i MZI (model 2).

$$lbwght = \beta_0 + \beta_1 cigs + \beta_2 male + \beta_3 parity + \beta_4 lfaminc + \epsilon$$

gdzie lbwhgt oznacza wagę dziecka przy urodzeniu (w logarytmie), cigs oznacza liczbę papierosów palonych dziennie, male to zmienna zerojedynkowa oznaczająca płeć dziecka (1=mężczyzna, 0=kobieta), parity oznacza dotychczasową liczbę zdrowych urodzeń danej kobiety, lfaminc oznacza dochód w rodzinie w setkach dolarów (w logarytmie). Jako zmienną instrumentalną wykorzystano cenę paczki papierosów w stanie, w którym mieszka matka.

Zadanie 4 c.d.

- **4A.** Podaj interpretację współczynnika przy zmiennej *cigs* w modelu 2. Czy wpływ zmiennej jest zgodny z oczekiwaniami?
- **4B.** Co można powiedzieć o błędach standardowych oszacowań w obu modelach?

lbwght	Model 1 (MNK)	Model 2 (MZI)
cigs	-0.0042***	0.0399
	(0.0009)	(0.5431)
male	0.0262***	0.0298*
	(0.0101)	(0.0178)
parity	0.0147***	-0.0012
	(0.0057)	(0.0219)
lfaminc	0.0180***	0.0636
	(0.0056)	(0.0570)
_cons	4.6757***	4.469***
	(0.0219)	(0.2588)
N	1,388	1,388
R^2	0.035	.
Prob > F	0.000	0.049

Zadanie 4 c.d.

4C. W następnej kolejności przeprowadzono Test Hausmana-Wu i uzyskano wynik:

Wu-Hausman $F(1,1382) = 1.91858$ $(p = 0.1662)$

Jaki wniosek możemy wyciągnąć na podstawie uzyskanego wyniku?

Zadanie 5

Rozważano następujący model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_1 + \beta_4 z_2 + \beta_5 z_3 + \epsilon$$

W którym zmienne x_1 i x_2 są endogeniczne, a zmienne z_1, z_2, z_3 są egzogeniczne oraz $E(\epsilon) = 0$.

1. Jako instrument dla zmiennych x_1 oraz x_2 zaproponowano zmienną z_4. Czy to podejście jest właściwe?
2. Jako instrument dla zmiennej x_1 zaproponowano zmienną z_4 oraz z_5, a jako instrument dla zmiennej x_2 zmienną z_6. Czy to podejście jest właściwe?
3. Jako instrument dla zmiennej x_1 zaproponowano zmienną z_4, a jako instrument dla zmiennej x_2 zmienną z_2. Czy to podejście jest właściwe?

Podsumowanie

- Zakończyliśmy pierwszy blok materiału.
  - testowanie hipotez i przypomnienie twierdzenia Gauss'a-Markov'a
  - dobór zmiennych do modelu i porównywanie/selekcja modeli
  - metoda zmiennych instrumentalnych
- Za tydzień laboratorium (sala A102) - przypomnijcie sobie proszę hasło do konta na WNE.