Problem równoczesności w MNK

- O problemie równoczesności mówimy, gdy występuje korelacja między wartością oczekiwaną $\boldsymbol{\varepsilon}_i$ i równoczesnym $\boldsymbol{x}_i$

- Model liniowy

$$
\begin{aligned}
\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\
\mathrm{E}\left(\boldsymbol{u}\right) &= \mathbf{0}
\end{aligned}
$$

- Powiedzmy, że występuje w nim zależność między wartością oczekiwaną błędu losowego a zmiennymi objaśniającymi:

$$
\mathrm{E}\left(\boldsymbol{\varepsilon}\middle|\,\boldsymbol{X}\right) \neq \mathbf{0}
$$

- W takiej sytuacji

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{b}\right) &= \mathrm{E}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}\middle|\,\boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\,\mathrm{E}\left[\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right] \\
&= \boldsymbol{\beta}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\,\mathrm{E}\left[\boldsymbol{\varepsilon}|\,\boldsymbol{X}\right] \neq \boldsymbol{\beta}
\end{aligned}$$

- Estymator MNK jest więc obciążony!

Wynik Aby estymator nie był obciążony, nie może być związku między wartością oczekiwaną błędu losowego $\boldsymbol{\varepsilon}_i$ a wszystkimi $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$

- Jednak jeśli $\mathrm{E}\left(\boldsymbol{x}_i'\boldsymbol{\varepsilon}_i\right) = 0$ estymator obiążony może być estymatorem zgodnym

- W dużych próbach

$$
\begin{aligned}
\operatorname{plim} \frac{1}{n} \boldsymbol{X}'\boldsymbol{X} &= \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \mathrm{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i'\right) \\
\operatorname{plim} \frac{1}{n} \boldsymbol{X}'\boldsymbol{\varepsilon} &= \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i' \boldsymbol{\varepsilon}_i = \mathrm{E}\left(\boldsymbol{x}_i' \boldsymbol{\varepsilon}_i\right)
\end{aligned}
$$

- Estymator MNK ma granicę według prawdopodobieństwa równą

$$
\begin{aligned}
\operatorname{plim} \boldsymbol{b} &= \operatorname{plim} \left[\left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{y} \right] = \operatorname{plim} \left[\left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}' \left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \right) \right] \\
&= \boldsymbol{\beta} + \operatorname{plim} \left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} \right)^{-1} \left(\frac{1}{n} \boldsymbol{X}'\boldsymbol{\varepsilon} \right) \\
&= \boldsymbol{\beta} + \left[\mathrm{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' \right) \right]^{-1} \mathrm{E}\left(\boldsymbol{x}_i' \boldsymbol{\varepsilon}_i \right) \neq \boldsymbol{\beta}
\end{aligned}
$$

Wynik Estymator MNK parametru $\boldsymbol{\beta}$ jest estymatorem zgodnym jeśli $\mathrm{E}\left(\boldsymbol{x}_i'\boldsymbol{\varepsilon}_i\right) = \mathrm{Cov}\left(\boldsymbol{x}_i, \boldsymbol{\varepsilon}_i\right) = 0$.

- Zauważmy, że wymóg ten jest znacznie słabszy niż warunek konieczny do nieobciążoności - wystarczy jedynie brak korelacji między ε_i i równoczesnym $\boldsymbol{x}_i$.

Przykład Chcemy wyestymować wpływ edukacji na poziom zarobków. Jednak w badaniach nie jesteśmy w stanie uchwycić zdolności danej osoby. Zwykle zdolności wpływają zarówno na poziom edukacji jak i na poziom zarobków. Jeśli zamiast modelu (dane PGSS lrincome - logarytm dochodu z pracy)

$$\log\left(lrincome\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\text{educ} + \boldsymbol{\beta}_2\text{sex} + \boldsymbol{\beta}_3\text{age} + \boldsymbol{\beta}_4\text{age}^2 + \boldsymbol{\beta}_5\text{zdolności} + \boldsymbol{\varepsilon}$$

wyestymujemy model

$$\log(wage) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{educ} + \boldsymbol{\beta}_2 \text{sex} + \boldsymbol{\beta}_3 \text{age} + \boldsymbol{\beta}_4 \text{age}^2 + u$$

to przy założeniu, że osiągnięty poziom wykształcenia zależy od zdolności uzyskujemy:

$$\operatorname{Cov}(\text{educ}, u) = \operatorname{Cov}(\text{educ}, \boldsymbol{\beta}_5 \text{zdolności} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta}_5 \operatorname{Cov}(\text{educ}, \text{zdolności}) \neq 0$$

Ponieważ zmienna pominięta jest dodatnio zdolności są dodanio skorelowane z edukacją i wpływają dodatnio na zarobki więc oczekujemy, że obciążenie oszacowania współczynnika $\boldsymbol{\beta}_4$ jest dodatnie.

Source	SS	df	MS
Model	192.55502	4	48.1387549
Residual	4815.96513	4222	1.14068336
Total	5008.52015	4226	1.18516804

Number of obs =	4227
F(4, 4222) =	42.20
Prob > F =	0.0000
R-squared =	0.0384
Adj R-squared =	0.0375
Root MSE =	1.068

lrincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0415494	.0052952	7.85	0.000	.0311681	.0519307
sex	-.3073812	.033083	-9.29	0.000	-.3722412	-.2425211
age	.052235	.0096668	5.40	0.000	.0332831	.071187
age2	-.0005858	.0001157	-5.06	0.000	-.0008127	-.0003589
_cons	4.788569	.2033733	23.55	0.000	4.389851	5.187288

Metoda Zmiennych Instrumentalnych

- Załóżmy, że jesteśmy w stanie znaleźć takie zmienne, od których nie zależy równoczesny błąd losowy:

$$\mathrm{E}\left(\boldsymbol{\varepsilon}_i \middle| \boldsymbol{z}_i\right) = 0$$

- Zgodnie z terminologią stosowaną w GMM, zmienne $\boldsymbol{z}_i$ nazywać będzimy instrumentami.

- Instrumentami mogą być również elementy $\boldsymbol{x}_i$

- Zgodnie z tym co powiedziano na temat GMM warunek $\mathrm{E}\left(\boldsymbol{\varepsilon}_i \middle| \boldsymbol{z}_i\right) = 0$ implikuje, że

$$\mathrm{E}\left(\boldsymbol{z}_i'\boldsymbol{\varepsilon}_i\right) = 0$$

- Spróbujmy teraz wyprowadzić estymator GMM dla tego przypadku

$$\boldsymbol{\varepsilon}_i = y_i - \boldsymbol{x}_i\boldsymbol{\beta}$$

- A więc ograniczenie narzucone na momenty teoretyczne ma postać:

$$\mathrm{E}\left[\boldsymbol{z}_i'\left(y_i - \boldsymbol{x}_i\boldsymbol{\beta}\right)\right] = 0$$

- Analogiczne ograniczenie narzucone na momenty empiryczne następujące:

$$\begin{aligned} \sum_{i=1}^{n}\frac{1}{n}\boldsymbol{z}_i'\left(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}\right) &= \sum_{i=1}^{n}\frac{1}{n}\boldsymbol{z}_i'y_i - \sum_{i=1}^{n}\frac{1}{n}\boldsymbol{z}_i'\boldsymbol{x}_i\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0 \end{aligned}$$

- Jeśli liczba elementów $\boldsymbol{z}_i$ będzie równa liczbie elementów $\boldsymbol{x}_i$

- liczba rownań będzie równa liczbie niewiadomowych zawartych w $\widehat{\boldsymbol{\beta}}$
- $\widehat{\boldsymbol{\beta}}$ będzie dokładnie zidentyfikowana
- macierz $\boldsymbol{Z}'\boldsymbol{X}$ będzie $K \times K$

- W takim przypadku można będzie policzyć odwrotność $\boldsymbol{Z}'\boldsymbol{X}$ i uzyskać wzór na $\widehat{\boldsymbol{\beta}}$

- Estymator MZI (IV - **I**ntrumental **V**ariable) jest wtedy dany wzorem:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \boldsymbol{Z}'\boldsymbol{y}$$

- **Zgodność** estymatora MZI

$$
\begin{aligned}
\operatorname{plim}\left(\widehat{\boldsymbol{\beta}}\right) &= \operatorname{plim}\left[\left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}\right] = \operatorname{plim}\left[\left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)\right] \\
&= \boldsymbol{\beta}+\operatorname{plim}\left[\left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{\varepsilon}\right] \\
&= \boldsymbol{\beta}+\left(\operatorname{plim}\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\operatorname{plim}\frac{1}{n}\boldsymbol{Z}'\boldsymbol{\varepsilon}\right) \\
&= \boldsymbol{\beta}+\left[\mathrm{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i\right)\right]^{-1}\underbrace{\mathrm{E}\left(\boldsymbol{z}_i'\boldsymbol{\varepsilon}_i\right)}_{0} = \boldsymbol{\beta}
\end{aligned}
$$

Uwaga Wyprowadzenie to jest prawidłowe jedynie, gdy $\mathrm{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i\right)$ jest macierzą nieosobliwą: estymator MZI jest zgodny jeśli nie ma korelacji między instrumentami i równoczesnymi błędami losowymi, ale *jest* korelacja między instrumentami i zmiennymi objaśniającymi.

Uogólniony estymator estymator MZI

- Uogólniony estymator MZI (GIV - **G**eneralized **I**ntrumental **V**ariable) używamy wtedy, gdy liczba instrumentów jest większa od liczby zmiennych w modelu

- W tym przypadku w układzie równań

$$\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0$$

  - liczba równań będzie większa niż liczba niewiadomowych zawartych w $\widehat{\boldsymbol{\beta}}$
  - to $\widehat{\boldsymbol{\beta}}$ będzie przeidentyfikowana
  - macierz $\boldsymbol{Z}'\boldsymbol{X}$ będzie $G \times K$ gdzie G liczba instrumentów $G > K$

- W tym przypadku nie da się policzyć odwrotności $\boldsymbol{Z}'\boldsymbol{X}$ - macierz ta nie jest kwadratowa

- Załóżmy, że

$$\boldsymbol{x}_i = \boldsymbol{z}_i \boldsymbol{B} + \boldsymbol{u}_i$$
$$\mathrm{E}\left(\boldsymbol{u}_i | \boldsymbol{z}_i\right) = 0$$

- Wynika z tego, że
  - $\mathrm{E}\left(\boldsymbol{x}_i | \boldsymbol{z}_i\right) = \boldsymbol{z}_i \boldsymbol{B}$

- Przypomnijmy sobie, że

$$\mathrm{E}\left(\boldsymbol{\varepsilon}_i | \boldsymbol{z}_i\right) = \mathrm{E}\left[f_i\left(\boldsymbol{\beta}\right) | \boldsymbol{z}_i\right] = 0$$

gdzie $f_i\left(\boldsymbol{\beta}\right) = y_i - \boldsymbol{x}_i \boldsymbol{\beta}$

- Dla tego problemu optymalne intrumenty będą można prosto policzyć ze wzoru:

$$\begin{aligned} \boldsymbol{J}_i &= \mathrm{E}\left(\left.\frac{\partial \boldsymbol{f}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right| \boldsymbol{z}_i\right) = \mathrm{E}\left(\left.\frac{\partial (y_i - \boldsymbol{x}_i \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right| \boldsymbol{z}_i\right) \\ &= \mathrm{E}\left(-\boldsymbol{x}_i \middle| \boldsymbol{z}_i\right) = \mathrm{E}\left(-\boldsymbol{z}_i \boldsymbol{B} - \boldsymbol{u}_i \middle| \boldsymbol{z}_i\right) = -\boldsymbol{z}_i \boldsymbol{B}. \end{aligned}$$

- Wektorem optymalnych instrumentów jest

$$\boldsymbol{W}_i = \boldsymbol{z}_i \, \boldsymbol{B}.$$

- Ponieważ w praktyce $\boldsymbol{B}$ jest nieznane, więc zastępujemy je przez estymator MNK postaci $\widehat{\boldsymbol{B}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{X}$.

- Uzyskany w ten sposób optymalne instrumenty mają postać

$$\widehat{\boldsymbol{W}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}$$

- Postać estymatora można parametru $\widehat{\boldsymbol{\beta}}$ ma teraz postać:

$$\begin{aligned}\boldsymbol{m}\left(\widehat{\boldsymbol{\beta}}\right) &= \frac{1}{n}\widehat{\boldsymbol{W}}'\boldsymbol{e}\left(\widehat{\boldsymbol{\beta}}\right) = \frac{1}{n}\widehat{\boldsymbol{W}}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = 0 \\ \boldsymbol{b}_{MZI} &= \left(\widehat{\boldsymbol{W}}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{W}}'\boldsymbol{y} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_Z\boldsymbol{y}\end{aligned}$$

gdzie

$$\boldsymbol{P}_Z = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$$

- Pokazanie zgodności i postaci macierzy wariancji przebiega tak samo jak dla zwykłego estymatora MZI

Testowanie w MZI

- Test egzogeniczności jest wariantem testu Hausmana
- Test Hausmana można przeprowadzić, jeśli mamy dwa estymatory o następujących właściwościach

1. jeśli H_0 jest prowdziwa to $\widehat{\boldsymbol{\theta}}$ jest zgodny a $\widetilde{\boldsymbol{\theta}}$ jest asymptotycznie efektywny
2. jeśli H_0 jest fałszywa jest to $\widehat{\boldsymbol{\theta}}$ jest dalej zgodny, ale $\widetilde{\boldsymbol{\theta}}$ jest asymptotycznie obciążony

- Test Hausmana polega na sprawdzeniu, czy oszacowania parametrów uzyskanych za pomocą estymatora $\widehat{\boldsymbol{\theta}}$ i $\widetilde{\boldsymbol{\theta}}$ są do siebie zbliżone.

- W naszym przypadku H_0 brzmi, że $\boldsymbol{X}$ jest egzogeniczne.

- W takim przypadku rolę estymatora $\widehat{\boldsymbol{\theta}}$ pełni estymator MZI a estymatora $\widetilde{\boldsymbol{\theta}}$ estymator MNK.

- Jeśli H_0 jest prawdziwe i $\boldsymbol{X}$ jest ogzogeniczne to łatwo pokazać, że

$$\sqrt{n}\left(\boldsymbol{b}_{MZI} - \boldsymbol{b}_{MNK}\right) \xrightarrow{D} N\left(0, \boldsymbol{\Sigma}_{b_{MZI}-b_{MNK}}\right)$$

co oznacza, że statystyka

$$\left(\boldsymbol{b}_{MZI} - \boldsymbol{b}_{MNK}\right)' \left(\widetilde{\boldsymbol{\Sigma}}_{b_{MZI}-b_{MNK}}\right)^{-1} \left(\boldsymbol{b}_{MZI} - \boldsymbol{b}_{MNK}\right) \xrightarrow{D} \boldsymbol{\chi}^2_g$$

gdzie g jest ilością zmiennych, których egzogeniczność badamy.

- Test na poprawność instrumentów Sargana

- Jest to test na prawdziwość ograniczeń przeidentyfikowujących
- Przeprowadzamy go badając

$$\frac{\widetilde{\boldsymbol{e}}'\boldsymbol{P}_W\widetilde{\boldsymbol{e}}}{\widetilde{\boldsymbol{\sigma}}^2} \xrightarrow{D} \boldsymbol{\chi}^2_{p-K}$$

gdzie $\widetilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{MZI}$, $\widetilde{\boldsymbol{\sigma}}^2 = \frac{\widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}}{n}$

- Statystyka ta służy do badania prawidłowości doboru instrumentów.
- Jeśli H_0 jest odrzucone to znaczy, że założenie o braku skorelowania zmiennych instrumentalnych i zaburzeń losowych jest fałszywe.

Przykład kontynuacja. Za instrumenty dla edukacji osiągniety poziom przez matkę i ojca poziom wykształcenia. Zakładamy, że edukacja rodziców nie mają znaczenia dla poziomu płac (jest nieskorelowana z u) Uzuskane za pomocą MZI oszacowania mają postać:

```
Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =    3862
-------------+------------------------------           F(  4,  3857) =   37.77
       Model |  36.0635538     4  9.01588846           Prob > F      =  0.0000
    Residual |  4503.18032  3857  1.16753444           R-squared     =  0.0079
-------------+------------------------------           Adj R-squared =  0.0069
       Total |  4539.24387  3861  1.17566534           Root MSE      =  1.0805

------------------------------------------------------------------------------
    lrincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1005262   .0131433     7.65   0.000     .0747576    .1262948
         sex |  -.3317927   .0358404    -9.26   0.000    -.4020608   -.2615247
         age |   .0496855   .0103803     4.79   0.000     .0293342    .0700369
        age2 |  -.0005467   .0001253    -4.36   0.000    -.0007924   -.0003011
       _cons |   4.200114   .2384509    17.61   0.000     3.732612    4.667616
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   sex age age2 _Ipadeg_1 _Ipadeg_2 _Ipadeg_3 _Ipadeg_4 _Ipadeg_5
               _Ipadeg_6 _Ipadeg_7 _Ipadeg_8 _Ipadeg_9 _Ipadeg_98 _Imadeg_1
               _Imadeg_2 _Imadeg_3 _Imadeg_4 _Imadeg_5 _Imadeg_6 _Imadeg_7
               _Imadeg_8 _Imadeg_9 _Imadeg_98
------------------------------------------------------------------------------
```

- Zauważmy, że oszacowany współczynnik dla edukacji jest teraz znacznie wyższy!

- Wynik testu sargana

```
Tests of overidentifying restrictions:
Sargan N*R-sq test       12.918  Chi-sq(19)   P-value = 0.8427
Basmann test             12.881  Chi-sq(19)   P-value = 0.8446
```

- Instrumenty wydają się być poprawne.

- Wynik testu Hausmana:

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      MZI          MNK         Difference          S.E.
-------------+----------------------------------------------------------------
        educ |    .1005262     .0415494        .0589768        .0118632
         sex |   -.3317927    -.3073812       -.0244116        .0126693
         age |    .0496855      .052235       -.0025495        .0034388
        age2 |   -.0005467    -.0005858        .0000391        .0000441
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from ivreg
          B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                 chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                         =       28.59
               Prob>chi2 =      0.0000
```

- Wyniki uzyskane dla MZI i MNK istotnie się różnią - odrzucamy H_0, że w modelu nie występuje równoczesność

- Uwaga: uzyskany wynik nie jest zgodny z intuicją - obciążenie oszacowania MNK współczynnika przy edukacji powinno być dodatnie. Zgodny estymator MZI dał jednak jeszcze wyższe oszacowanie tego współczynnka.