MODELE Z INTERAKCJAMI I WIELOMIANAMI

EKONOMETRIA WNE 2022/23

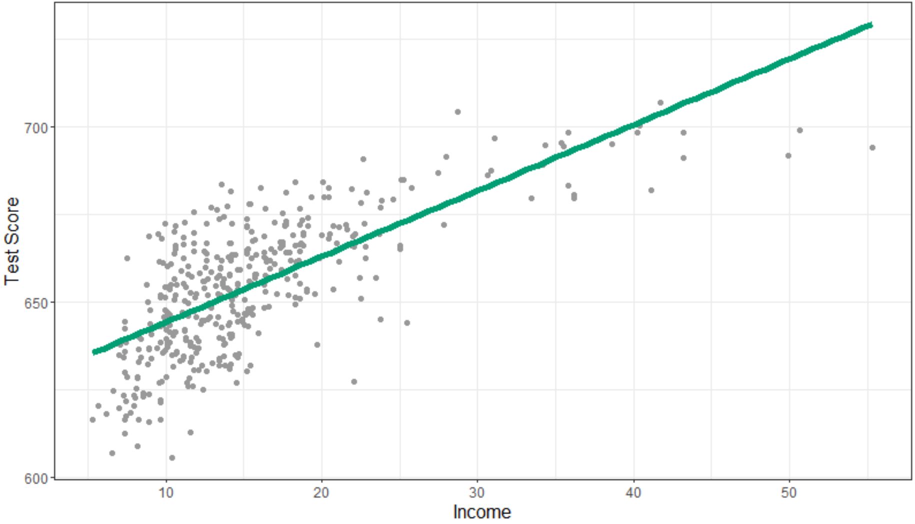
Sebastian Zalas

1 grudnia 2022

WPROWADZENIE

- Sposoby uchwycenia nieliniowej relacji między zm. zależną i niezależną:
  - umieszczenie wielomianu zm. niezależnej
  - interakcje
- Przykład: TestScore vs Income, STratio
  - dane `CASchools` - dostępne w pakiecie `AER`
  - TestScore - łączny wynik testu z matematyki i czytania; średnia w dystrykcie
  - Income - średni dochód w dystrykcie
  - STratio - stosunek liczby nauczycieli do liczby uczniów

WYNIK TESTU VS DOCHÓD



Wyraźnie relacja pomiędzy TestScore oraz Income jest nieliniowa

MODEL Z WIELOMIANEM ZM. ZALEŻNEJ

- Jak uchwycić nieliniową relację?

⇒ Przybliżenie zm. zależnej wielomianem:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{1,i}^2 + \beta_3 x_{1,i}^3 + \ldots + \beta_k x_{1,i}^k \varepsilon$$

- to zwykły model liniowy - mimo podniesionych do potęgi regresorów
- parametry takiego modelu są trudne do interpretacji, ale interpretowanie wyników regresji jest możliwe

MODEL Z WIELOMIANEM ZM. NIEZALEŻNEJ - PRZYKŁAD

- Relacja między *TestScore* i *Income* (średni dochód w dystrykcie per capita)

⇒ Oszacowanie modelu z kwadratem dochodu

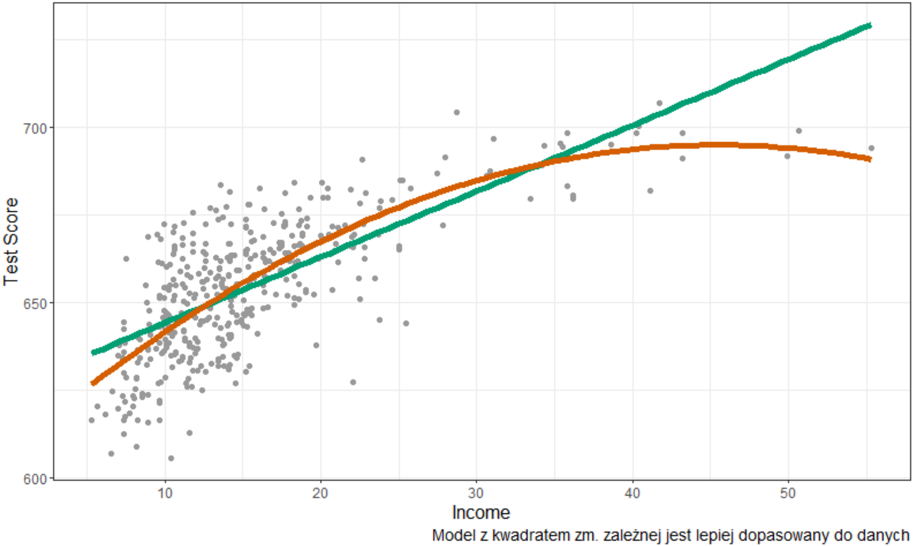
$$TestScore_i = 607.302 + 3.85099 Income_i - 0.04231 Income_i^2 + \varepsilon_i$$

⇒ Oszacowanie modelu z sześcianem dochodu

$$TestScore_i = 600.079 + 5.018677 Income_i - 0.09580521 Income_i^2 + 0.0006854851 Income_i^3 + \varepsilon_i$$

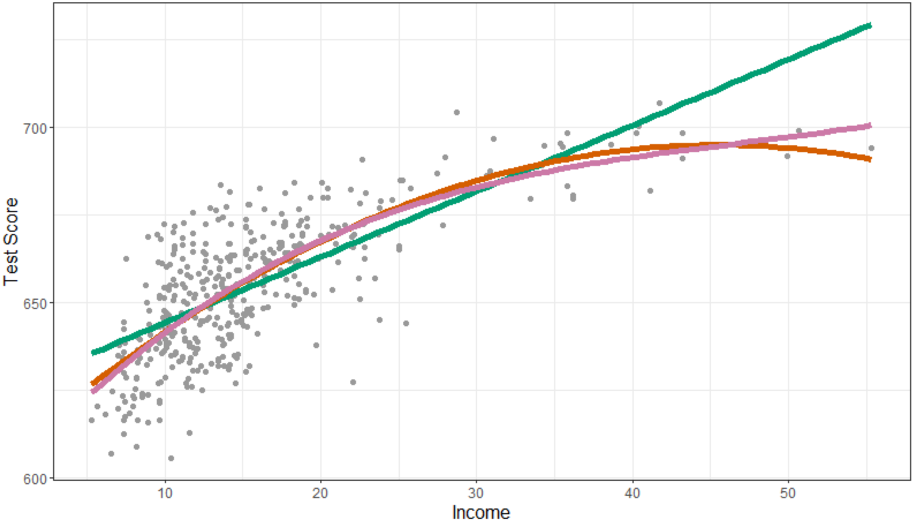
MODEL Z WIELOMIANEM ZM. NIEZALEŻNEJ - INTERPRETACJA

- Zrób wykres wartości dopasowanych



MODEL Z WIELOMIANEM ZM. NIEZALEŻNEJ - INTERPRETACJA

- Czy model z wielomianem 3. stopnia jest lepszy?



Wielomian 3. stopnia nie poprawia sytuacji...

MODEL Z WIELOMIANEM ZM. NIEZALEŻNEJ - INTERPRETACJA

- Policz zmiany (dla modelu z kwadratem):

$$\Delta y_i = \beta_1 \Delta x_{1,i} + \beta_2 \Delta x_{1,i}^2$$

w przybliżeniu:

$$\Delta y_i \approx \beta_1 + 2\beta_2 x_{1,i}$$

- Przykład:

$$\Delta Test\hat{S}core_i = 3.85099 - 2 \times 0.04231 Income_i$$

- dla zmiany dochodu o 1 jeden tysiąc, przy dochodzie wynoszącym 10 tysięcy:

$$\begin{aligned}\Delta Test\hat{S}core_i &= 3.85099 - 2 \times 0.04231 \times 10 \\ &= 3.005\end{aligned}$$

- czyli zmiana dochodu o 1 tysiąc, jest związana ze wzrostem wyniku testu średnio o 3.005 punktu (w przybliżeniu)

MODEL Z WIELOMIANEM ZM. NIEZALEŻNEJ - INTERPRETACJA

- Zmiany zależą od poziomu zmiennej zależnej
- Możemy policzyć zmiany wartości dopasowanych dla wybranych, konkretnych przedziałów
- Przykład:

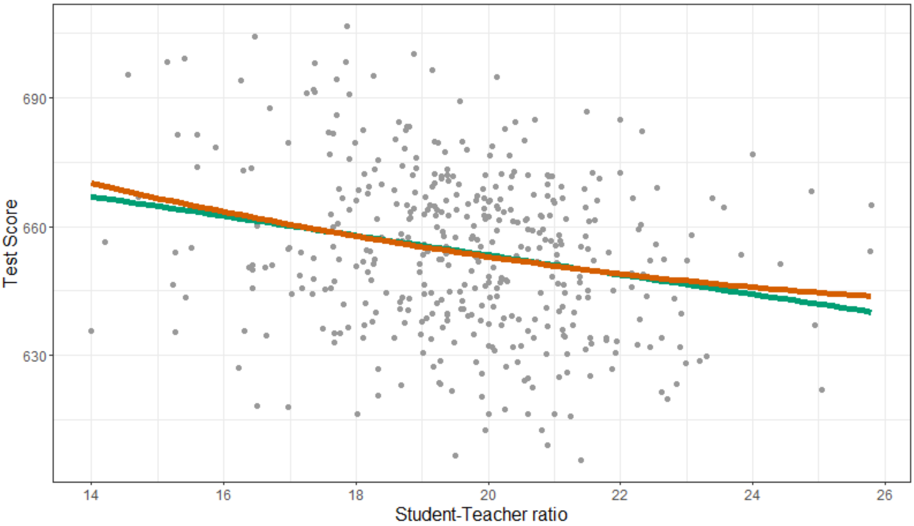
Zmiana w *Income* o 1000 per capita	$\Delta TestScore$
z 5 na 6 tys. pc	3.3856
z 10 na 11 tys. pc	2.9625
z 25 na 26 tys. pc	1.6933
z 45 na 46 tys. pc	0.0009

MODEL Z WIELOMIANEM ZM. NIEZALEŻNEJ - PODSUMOWANIE

- Model z wielomianem zm. zależnej estymujemy MNK
- Współczynniki mają skomplikowaną interpretację
- Aby zinterpretować oszacowania:
  - zrób wykres
  - oblicz przewidywane $\frac{\Delta y}{\Delta x}$ dla różnych wartości x.
- Sposoby wyboru stopnia wielomianu k: sprawdź wrażliwość oszacowań w różnych wariantach, osąd własny, a także korzystamy ze statystyk t oraz F (w przyszłości)

INTERAKCJE

- Czy dostrzegamy tu jakiś rodzaj nieliniowości?



Na pierwszy rzut oka, relacja między Test Score a Student-Teacher ratio nie wydaje się nieliniowa...

INTERAKCJE

- Przyjrzyjmy się modelowi:

$$TestScore = \beta_0 + \beta_1 STratio + \varepsilon$$

  gdzie

  - *TestScore* - średni wyniku testu uczniów w Kalifornii w danym dystrykcie
  - *STratio* - stosunek liczby nauczycieli do liczby uczniów

- Zwykle mniej uczniów na jednego nauczyciela, oznacza bardziej indywidualne podejście...

$\Rightarrow$ ... ale czy ta zależność jest zawsze taka sama? NIE!

- Przyjrzyjmy się modelowi:

$$TestScore = \beta_0 + \beta_1 STratio + \varepsilon$$

  gdzie
  - *TestScore* - średni wyniku testu uczniów w Kalifornii w danym dystrykcie
  - *STratio* - stosunek liczby nauczycieli do liczby uczniów
- Zwykle mniej uczniów na jednego nauczyciela, oznacza bardziej indywidualne podejście...

$\Rightarrow$ ... ale czy ta zależność jest zawsze taka sama? NIE!

INTERAKCJE

- Przyjrzyjmy się modelowi:

$$TestScore = \beta_0 + \beta_1 STratio + \varepsilon$$

  gdzie
  - *TestScore* - średni wyniku testu uczniów w Kalifornii w danym dystrykcie
  - *STratio* - stosunek liczby nauczycieli do liczby uczniów
- Zwykle mniej uczniów na jednego nauczyciela, oznacza bardziej indywidualne podejście...

⇒ ... ale czy ta zależność jest zawsze taka sama? NIE!

INTERAKCJE

- Być może zmniejszenie klasy jest zróżnicowana w zależności od okoliczności...
- Być może uczniowie uczący się angielskiego mogą bardziej potrzebować atencji nauczyciela..
- To znaczy: $\frac{\Delta TestScore}{\Delta STratio}$ może zależeć od Pc_EL (procent uczniów uczących się angielskiego)
- Bardziej ogólnie $\frac{\Delta Y}{\Delta X_1}$ może zależeć od X_2
- Jak modelować takie interakcje?

$\Rightarrow$ Najpierw zmienne binarne, później zmienne ciągłe.

INTERAKCJE - ZMIENNE BINARNE

$$y_i = \beta_0 + \beta_1 D_{1,i} + \beta_2 D_{2,i} + u_i$$

- $D_{1,i}$, $D_{2,i}$ są zmiennymi zero-jedynkowymi
- Spójrzmy na β_1
  - mierzy efekt zmiany z $D_1 = 0$ na $D_2 = 1$
  - w takiej specyfikacji ten efekt jest niezależny od wartości D_2
- Aby zbadać zależność między zmianą D_1 a D_2, musimy uwzględnić interakcję $D_1 \times D_2$ jako regresor:

$$y_i = \beta_0 + \beta_1 D_{1,i} + \beta_2 D_{2,i} + \beta_3 (D_{1,i} \times D_{2,i}) + u_i$$

INTERAKCJE - ZMIENNE BINARNE

$$y_i = \beta_0 + \beta_1 D_{1,i} + \beta_2 D_{2,i} + u_i$$

- $D_{1,i}$, $D_{2,i}$ są zmiennymi zero-jedynkowymi
- Spójrzmy na β_1
  - mierzy efekt zmiany z $D_1 = 0$ na $D_2 = 1$
  - w takiej specyfikacji ten efekt jest niezależny od wartości D_2
- Aby zbadać zależność między zmianą D_1 a D_2, musimy uwzględnić interakcję $D_1 \times D_2$ jako regresor:

$$y_i = \beta_0 + \beta_1 D_{1,i} + \beta_2 D_{2,i} + \beta_3 (D_{1,i} \times D_{2,i}) + u_i$$

INTERAKCJE - ZMIENNE BINARNE - INTERPRETACJA

$$y_i = \beta_0 + \beta_1 D_{1,i} + \beta_2 D_{2,i} + \beta_3 D_{1,i} \times \beta_2 D_{2,i} + u_i$$

- Ogólna zasada: porównaj przypadki:

$$\mathbb{E}(y_i | D_{1,i} = 0, D_{2,i} = d_2) = \beta_0 + \beta_2 d_2 \quad (1)$$

$$\mathbb{E}(y_i | D_{1,i} = 1, D_{2,i} = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \quad (2)$$

  odejmijmy (2) - (1) :

$$\mathbb{E}(y_i | D_{1,i} = 1, D_{2,i} = d_2) - \mathbb{E}(y_i | D_{1,i} = 0, D_{2,i} = d_2) = \beta_1 + \beta_3 d_2$$

- Wpływ D_1 zależy od d_2 (wartość zm. D_2)
- β_3 = wpływ zmiany D_1 gdy $D_2 = 1$

INTERAKCJE - ZMIENNE BINARNE - PRZYKŁAD

- Zależność *TestScore*, *STratio*, *hiEL* oraz *hiSTr*
- *hiEL* = 1 jeżeli *PctEL* $\geq$ 10
- *hiSTr* = 1 jeżeli *STratio* $\geq$ 20
- Oszacowanie:

$$\widehat{TestScore} = 664.1 - 18.2 hiEL - 1.9 hiSTr - 3.5 (hiSTr \times hiEL)$$

- "Wpływ" *hiSTr* gdy *hiEL* = 0 wynosi –1.9
- "Wpływ" *hiSTr* gdy *hiEL* = 1 wynosi –1.9 – 3.5 = –5.4
- Zmniejszenie liczby uczniów do nauczycieli ma większy efekt, gdy procent uczących się angielskiego jest duży

INTERAKCJE - ZMIENNE BINARNA I CIĄGŁA

- Interakcja między zmienną ciągłą i binarną:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + u_i$$

- D_i to zm. binarna, x_i to zm. ciągła
- w powyższym modelu wpływ x_i na y_i, czyli β_2, nie zależy od D_i
- aby sprawdzić jak x_i wpływa na y_i w zależności D_i trzeba wprowadzić do modelu interakcję, $D_i \times x_i$ jako zm. objaśniającą:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 D_i \times x_i + u_i$$

INTERAKCJE - ZMIENNE BINARNA I CIĄGŁA

- Dwie linie regresji:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 D_i \times x_i + u_i$$

- przypadek z $D_i = 0$

$$y_i = \beta_0 + \beta_2 x_i + u_i$$

- przypadek z $D_i = 1$

$$y_i = \beta_0 + \beta_1 + \beta_2 x_i + \beta_3 x_i + u_i$$

$$y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_i + u_i$$

INTERAKCJE - ZMIENNE BINARNA I CIĄGŁA - INTERPRETACJA

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 D_i \times x_i + u_i$$

- Ogólna zasada: porównaj przypadki:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \beta_3 D_i \times x_i \tag{3}$$

- Zmieńmy x_i

$$y_i + \Delta y_i = \beta_0 + \beta_1 D_i + \beta_2 (x_i + \Delta x_i) + \beta_3 D_i \times (x_i + \Delta x_i) \tag{4}$$

- odejmijmy (4) - (3)

$$\Delta y_i = \beta_2 \Delta x_i + \beta_3 D_i \Delta x_i$$

$$\frac{\Delta y_i}{\Delta x_i} = \beta_2 + \beta_3 D_i$$

INTERAKCJE - ZMIENNE BINARNA I CIĄGŁA - PRZYKŁAD

- Zależność *TestScore*, *STratio* i *hiEL* (*hiEL* = 1 jeżeli *PctEL* $\geq$ 10)
- Oszacowanie:

$$TestScore = 682.2 - 0.97STratio + 5.6hiEL - 1.28(STratio \times hiEL)$$

- gdy *hiEL* = 0:

$$TestScore = 682.2 - 0.97STratio$$

- gdy *hiEL* = 1:

$$\begin{aligned} TestScore &= 682.2 - 0.97STR + 5.6 - 1.28STratio \\ &= 687.8 - 2.25STratio \end{aligned}$$

- w efekcie otrzymujemy dwie regresje, każda dla jednego przypadku zmiennej *hiEL*
- Obniżenie liczby uczniów przypadających na jednego nauczyciela ma silniejszy wpływ na wyniki uczniów, gdy udział uczniów uczących się angielskiego jest duży.

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl