Modele z interakcjami i wielomianami

Ekonometria WNE

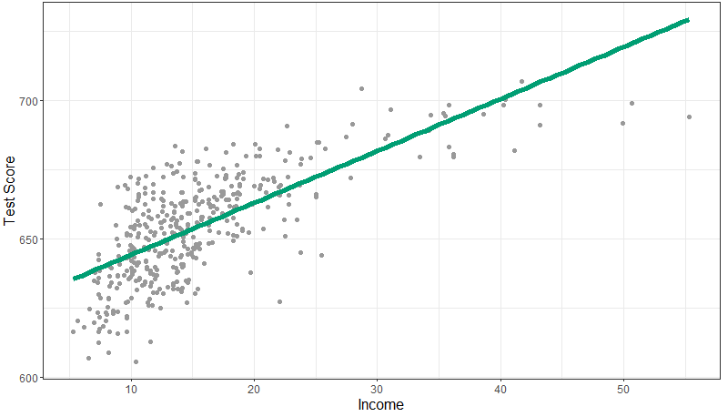
Sebastian Zalas

University of Warsaw
s.zalas@uw.edu.pl

Nieliniowość w linowym modelu?

- TAK, a w jaki sposób?:
  - umieszczenie wielomianu zmiennej objaśniającej
  - 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 x

- Przybliżenie y za pomocą wielomianu x:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \ldots + \beta_k x_1^k + \varepsilon$$

- model nadal jest linowy w stosunku do **parametrów** (*linear in parameters*), model szacujemy korzystając z MNK

- parametry takiego modelu są trudniejsze do interpretacji, ale możliwe $\Rightarrow$ efekt cząstkowy

Model z wielomianem x

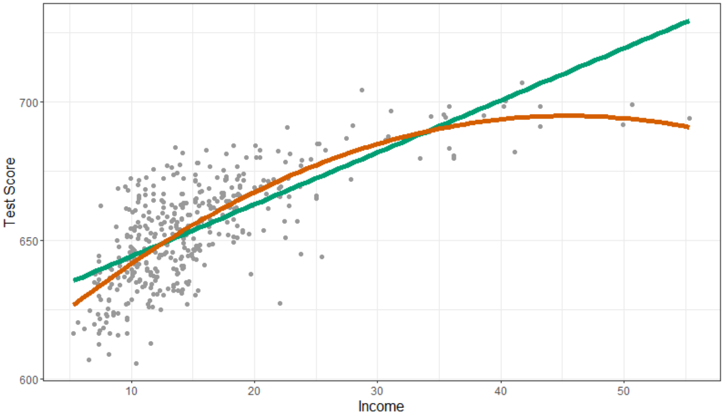
- Szacujemy relację między *TestScore* i *Income* (średni dochód w dystrykcie per capita)
- oszacowanie modelu z kwadratem dochodu

$$TestScore = 607.302 + 3.85099 Income - 0.04231 Income^2 + \varepsilon$$

- oszacowanie modelu z sześcianem dochodu

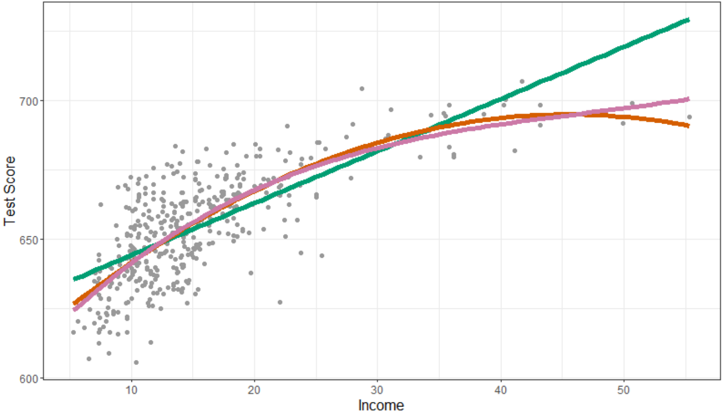
$$TestScore = 600.079 + 5.018677 Income - 0.09580521 Income^2 + 0.0006854851 Income^3 + \varepsilon$$

Model z x, x^2



Model z kwadratem zm. zależnej jest lepiej dopasowany do danych

Model z x, x^2 oraz x^3



Wielomian 3. stopnia nie poprawia sytuacji...

Model z wielomianem x

- Policz zmiany (dla modelu z kwadratem):

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_1^2$$

- W przybliżeniu

$$\Delta y \approx \beta_1 + 2\beta_2 x_1$$

- równoważność efektu cząstkowego - pochodna po x_1

Model z wielomianem X

- Przykład: model z kwadratem dochodu

$$\Delta\widehat{TestScore_i} = 3.85099 - 2 \times 0.04231 Income_i$$

- dla zmiany dochodu o 1 tys. \$ pc., przy dochodzie pc. wynoszącym 10 tys. \$

$$\begin{aligned}\Delta\widehat{TestScore_i} &= 3.85099 - 2 \times 0.04231 \times 10 \\ &= 3.005\end{aligned}$$

- Zmiana dochodu o 1 tys. \$ pc. przy dochodzie pc. wynoszącym 10 tys. \$, jest związana ze wzrostem wyniku testu średnio o 3.005 punktu (w przybliżeniu)

Model z wielomianem z x

- Interpretacja dokonywana w okolicy danego punktu
- Możemy policzyć zmiany wartości dopasowanych dla wybranych, konkretnych przedziałów

$\Delta Income$ o 1 tys. $ 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

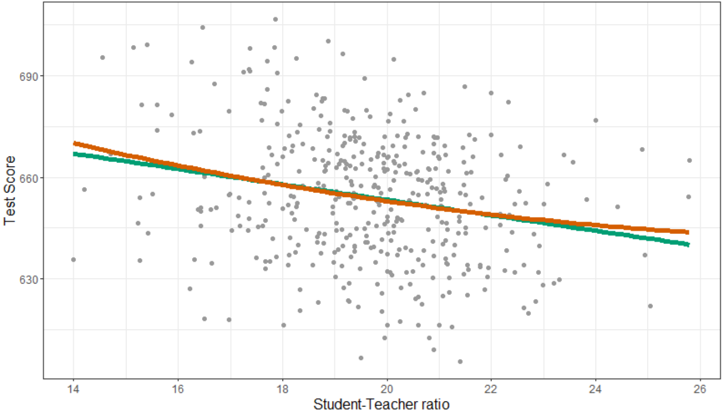
MODEL Z WIELOMIANEM x

- 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)

- Im wyższy będzie stopień wielomianu, tym bardziej model stanie się podatny na odstające obserwacje

Czy dostrzegamy tu nieliniowość?



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

Interakcje

- Przyjrzyjmy się modelowi:

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

  - *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 $\Rightarrow \uparrow$ *TestScore*

- Ta zależność (reprezentowna przez β_1) zależy także od innych czynników

Interakcje

- Jakich? Kontekst Kalifornii: ile uczniów zna j. angielski? Przez imigrację z Meksyku te odsetki różnią się między dystryktami
- $\frac{\Delta TestScore}{\Delta STratio}$ może różnić się w zależności od procenta uczniów uczących się angielskiego, *PEL*
- Bardziej ogólnie $\frac{\Delta Y}{\Delta X_1}$ może zależeć od X_2. Jak modelować takie interakcje?
  - zmienne binarne vs ciągłe

INTERAKCJE - ZMIENNE BINARNE

- Model ze zmiennymi binarnymi D_1, D_2

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \varepsilon$$

- 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 łączną zależność między D_1 i D_2 a y, musimy uwzględnić interakcję, czyli $D_1 \times D_2$:

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + \varepsilon$$

INTERAKCJE - ZMIENNE BINARNE - INTERPRETACJA

- Intuicja: porównaj przypadki:

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 \times \beta_2 D_2 + \varepsilon$$

- Ustalamy wartości D_2, zmieniamy D_1:

$$\mathbb{E}(y|D_1 = 0, D_2 = d_2) = \beta_0 + \beta_2 d_2 \quad (1)$$

$$\mathbb{E}(y|D_1 = 1, D_2 = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \quad (2)$$

odejmijmy (2) - (1) :

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

- to jak D_1 wpływa na y zależy od d_2 (wartość zm. D_2). β_3 mierzy wpływ zmiany D_1 gdy $D_2 = 1$

Interakcje - zmienne binarne - przykład

- Zależność *TestScore*, *STratio*; *HEL* oraz *HSTr*
  - $HEL = 1$ jeżeli odsetek uczących się angielskiego ≥ 10; 0 wpp.
  - $HSTr = 1$ jeżeli $STratio \geq 20$, 0 wpp.
- Oszacowanie:

$$TestScore = 664.1 - 18.2HEL - 1.9HSTr - 3.5(HSTr \times HEL)$$

- Spróbujmy zinterpretować wyniki

Interakcje - zmienne binarne - przykład

- gdy $HEL = 0$

$$\mathbb{E}[TestScore|HEL = 0, HSTR = 0] = 664.1$$

$$\mathbb{E}[TestScore|HEL = 0, HSTR = 1] = 664.1 - 1.9 \times 1$$

gdy odsetek uczących się angielskiego jest niski, zwiększęszenie stosunku l. uczniów do nauczycieli zmniejszyło średnio wynik o 1.9

- gdy $HEL = 1$

$$\mathbb{E}[TestScore|HEL = 1, HSTR = 0] = 664.1 - 18.2$$

$$\mathbb{E}[TestScore|HEL = 1, HSTR = 1] = 664.1 - 18.2 - 1.9 - 3.5$$

- Zmniejszenie liczby uczniów do nauczycieli ma silniejszy związek z wynikiem testu, gdy procent uczących się angielskiego jest duży

Interakcja zm. binarnej i ciągłej

- Model z interakcją x (zm. ciągłej) oraz D (zm. binarna)

$$y = \beta_0 + \beta_1 D + \beta_2 x + \beta_3 D \times x + \varepsilon$$

- Stosując interakcję mamy dwie zagnieżdżone regresje:
  - przypadek z $D_i = 0$

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

  - przypadek z $D = 1$

$$y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x + \varepsilon$$

- czy relacja między $y \sim x$ zależy od D?

INTERAKCJA ZM. BINARNEJ I CIĄGŁEJ

- Co mierzy β_3?

$$y = \beta_0 + \beta_1 D + \beta_2 x + \beta_3 D \times x + \varepsilon$$

- Ustalmy D, zmieńmy x

$$\Delta y = \beta_2 \Delta x + \beta_3 D \Delta x$$
$$\frac{\Delta y}{\Delta x} = \beta_2 + \beta_3 D$$

- Co mierzy β_3: różnicę nachylenia $y \sim x$ przy uwzględnieniu D

Interakcja zm. binarnej i ciągłej - przykład

- Zależność $TestScore$, $STratio$ i HEL ($HEL = 1$ jeżeli odsetek uczących się angielskiego ≥ 10)

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

- Spójrzmy na dwa przypadki:
  - gdy $HEL = 0$: $TestScore = 682.2 - 0.97STratio$
  - gdy $HEL = 1$:
    $TestScore = 682.2 - 0.97STratio + 5.6 - 1.28STratio = 687.8 - 2.25STratio$

- Obniżenie liczby uczniów przypadających na jednego nauczyciela ma o 1.28 silniejszy wpływ na wyniki uczniów, gdy udział uczniów uczących się angielskiego jest duży.

Pytania? Wątpliwości?

Dziękuję!

e: s.zalas@uw.edu.pl