Ekonometria

Metodologia budowy modelu

Jerzy Mycielski

WNE, UW

2023

Hipotezy proste i złożone

- Najprostszy przypadek: ustalenie zbioru zmiennych objaśniających w modelu

1. Hipoteza złożona $H_0 : \beta_1 = \ldots = \beta_K = 0$, poziomie istotności α.
2. K hipotez prostych $H_1 : \beta_1 = 0; \ldots; H_K : \beta_K = 0$, poziom istotności *dla każdej* z nich α.

- W drugiej procedurze odrzucamy H_0 zostaje odrzucona, gdy odrzucona choćby jedna z hipotez $H_1, \ldots, H_K$.
- Procedury te nie są równoważne!!!

Obciążania Lovella

- Załóżmy, że statystyki testowe dla każdej z hipotez niezależne od siebie
- Poziom istotności w drugiej procedurze jest równy prawdopodobieństwu α^*, że jedna lub więcej z hipotez $H_1, \ldots, H_K$ zostanie odrzucona (przy prawdziwym H_0):

$$\alpha^* = 1 - (1 - \alpha)^K$$

- Różnicę między nominalnym poziomem istotności α i prawdziwym poziomem istotności α^* nazywamy obciążeniem Lovella.
- Dla omawianego przypadku $\lim_{K \longrightarrow \infty} \alpha^* = 1$
- Prawdopodobieństwo błędu drugiego rodzaju zbliża się do 1.
- **Wniosek**: powinniśmy testować hipotezy złożone a nie proste

Obciążania Lovella - przykład

Przykład

Badanie związku między prestiżem wykonywanego zawodu a znakiem zodiaku - baza PGSS rok 1997

siops	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Izodiac_2	-1.185876	1.188894	-1.00	0.319	-3.517406	1.145655
_Izodiac_3	-1.163471	1.214544	-0.96	0.338	-3.545303	1.21836
_Izodiac_4	-1.943503	1.224624	-1.59	0.113	-4.345101	.4580955
_Izodiac_5	-.8442572	1.231682	-0.69	0.493	-3.259698	1.571183
_Izodiac_6	.489734	1.248656	0.39	0.695	-1.958995	2.938463
_Izodiac_7	-1.735876	1.233491	-1.41	0.159	-4.154865	.6831132
_Izodiac_8	-2.516932	1.25468	-2.01	0.045	-4.977474	-.0563895
_Izodiac_9	-1.77159	1.303026	-1.36	0.174	-4.326944	.7837636
_Izodiac_10	-1.710475	1.208141	-1.42	0.157	-4.079749	.6587993
_Izodiac_11	-.0627988	1.216184	-0.05	0.959	-2.447846	2.322248
_Izodiac_12	-.4462717	1.186128	-0.38	0.707	-2.772377	1.879833
_cons	39.48588	.8659397	45.60	0.000	37.78769	41.18406

Obciążania Lovella - przykład c. d.

- Przy 11 zmiennych objaśniających p-stwo, że 1 wyjdzie istotna na poziomie 5% wynosi $1-(0.95)^{11}=0.44$. Powinniśmi jednak patrzeć na statystykę na łączną istotność zmiennych.

```
      Source |       SS       df       MS              Number of obs =    2118
-------------+------------------------------           F( 11,  2106) =    1.08
       Model |  1571.42612    11   142.85692            Prob > F      =  0.3764
    Residual |  279516.147  2106  132.723717            R-squared     =  0.0056
-------------+------------------------------           Adj R-squared =  0.0004
       Total |  281087.573  2117  132.776369            Root MSE      =  11.521
```

Upraszczanie modelu

- Przykład: dowolny model
  - Jak ustalić listę istotnych zmiennych?
- Przykład: model wielomianowy

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i$$

  - Jak ustalić p?
- Przykład: PKD
  - Jak szczegółowy podział na sektory powinien być uwzględniony w modelu?

Metodologia od ogólnego do szczegółowego

- Cel: znalezienie prawidłowej formy modelu
- Ustalając prawidłową formę modelu można często sformułować ciąg zagnieżdżonych modeli
- **Przykład:** ustalanie zbioru zmiennych objaśniających
  - $H_0^0 : \beta_i \neq 0$ dla $i = 1, \ldots, K$
  - $H_0^1 : \beta_1 = 0$
  - $H_0^2 : \beta_1 = \beta_2 = 0$
  - $\vdots$
  - $H_0^K : \beta_1 = \ldots = \beta_K = 0$
- Model A jest zagnieżdżony w modelu B, jeśli jest szczególnym przypadkiem modelu B w tym sensie, że model B staje się modelem A w przypadku gdy prawdą są pewne ograniczenia narzucone na parametry modelu B .
- Hipotezy $H_0^1, H_0^2, \ldots, H_0^K$ testujemy sekwencyjnie dla $i = 1, \ldots, K$
- Zaczynamy od modelu najbardziej ogólnego i przechodzimy do modeli coraz bardziej szczegółowych (z coraz większą ilością ograniczeń)

Metodologia od ogólnego do szczegółowego

Testowanie

- **Pytanie**: Jaka powinna być hipotezy alternatywne?
- Dwie możliwości testowania:
  - testujemy H_0^i przy hipotezie alternatywnej H_0^{i-1} (po kolei eliminujemy zmienne - statystyka t)
  - testujemy H_0^i przy hipotezie alternatywnej H_0^0 (testujemy hipotezę o niestotności coraz większego zbioru zmiennych - statystyka F)
- W pierwszym podejściu pojawia się problem z obciążeniem Lovella
- W drugim przypadku za każdym razem poprawny poziom istotności α - to podejście jest poprawne
- Kończymy upraszczanie gdy hipoteza H_0^i odrzucona

- Błędna procedura: od szczegółowego do ogólnego
- **Przykład:**
  - Zaczynamy od modelu z małą ilością zmiennych
  - Dodajemy kolejne zmienne i testujemy ich istotność
  - Problem: całe wnioskowanie statystyczne dla modeli ze zbyt małą ilością zmiennych jest błędne z racji na występowanie problemu zmiennych pominiętych
  - **Wniosek**: nie da się udowodnić, że przy dużej próbie ustalimy w ten sposób prawidłowy zbiór zmiennych
- Model uzyskany w rezultacie procedury od ogólnego do szczegółowego zależeć może od uszeregowania hipotez
- Uszeregowanie to powinno mieć jakieś teoretycze uzasadnienie

Kryteria informacyjne (selekcja modelu)

- W przypadku modeli szacowanych metodą *MNW* niemożliwe jest zdefiniowanie $\overline{R}^2$.
- Czy istnieją miary dopasownia, które uwzględniałaby ilość traconych w procesie estymacji stopni swobody?
- Statystyki takie istnieją i noszą ogólną nazwę kryteriów informacyjnych.
- Jedną z takich statystyk jest skorygowane R^2.

- Najbardziej popularnymi z nich jest kryterium informacyjne Akaike *AIC* (**A**kaike **I**nformation **C**riterion) i Bayesowskie kryterium informacyjne Schwartza *BIC* (**B**ayes **I**nformation **C**riterion) - określane skrótami SC, SBC, SIC
- Wzory dla MNK i ogólniejsze dla MNW

$$BIC = \log\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{2}\right) + n^{-1}K\log(n) = -\frac{2\ell\left(\widehat{\theta}\right)}{n} + \frac{K\log(n)}{n}$$

$$AIC = \log\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{2}\right) + \frac{2K}{n} = -\frac{2\ell\left(\widehat{\theta}\right)}{n} + \frac{2K}{n}$$

- Za najlepszy uznaje się ten model, dla którego kryterium informacyjne uzyskuje najniższą wartość.
- Pokazano, że dla pewnych ogólnych założeń, dla $n \longrightarrow \infty$ wybrany na podstawie *BIC* model zawierać będzie poprawny zbiór zmiennych objaśniających.
- W przypadku *AIC* okazuje się, że zbiór ten nawet dla $n \longrightarrow \infty$ może być zbyt duży.

Definicja obejmowania

Definicja

Model **A** obejmuje model **B** jeśli

1. Wszystko to co można wyjaśnić za pomocą modelu **B** można wyjaśnić także za pomocą modelu **A**
2. Istnieją elementy, które można wyjaśnić za pomocą modelu **A** ale nie można wyjaśnić za pomocą modelu **B**

Przykład

Teoria względności obejmuje fizykę klasyczną w tym sensie, że wszystkie zjawiska, które można wyjaśnić za pomocą fizyki klasycznej można wyjaśnić też za pomocą teorii względności ale pewne zjawiska (n.p. istnienie czarnych dziur) można wyjaśnić jedynie przy użyciu teorii względności.

Testy obejmowania

- Testy obejmowania stosujemy w przypadku modeli niezagnieżdżonych:

$$H_0 : \boldsymbol{y} = \boldsymbol{X}_A \boldsymbol{\beta}_A + \boldsymbol{\varepsilon}_1 \quad \text{(A)}$$
$$H_1 : \boldsymbol{y} = \boldsymbol{X}_B \boldsymbol{\beta}_B + \boldsymbol{\varepsilon}_2 \quad \text{(B)}$$

przy czym $\boldsymbol{X}_A \subsetneq \boldsymbol{X}_B$ i $\boldsymbol{X}_B \subsetneq \boldsymbol{X}_A$

Prosty test obejmowania

- Konstruujemy sztuczny model, który w którym zagnieżdżone są modele **A** i **B**

$$\boldsymbol{y} = \overline{\boldsymbol{X}}_A \overline{\boldsymbol{\beta}}_A + \overline{\boldsymbol{X}}_B \overline{\boldsymbol{\beta}}_B + \boldsymbol{W}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

- $\overline{\boldsymbol{X}}_A$ zmienne, które należą do $\boldsymbol{X}_A$, ale nie należą do $\boldsymbol{X}_B$,
- $\overline{\boldsymbol{X}}_B$ zmienne, które należą do $\boldsymbol{X}_B$, ale nie należą do $\boldsymbol{X}_A$
- $\boldsymbol{W}$ jest macierzą zmiennych, które należą do $\boldsymbol{X}_A$ i $\boldsymbol{X}_B$.
- Testujemy dwie hipotezy : $H_0^* : \overline{\boldsymbol{\beta}}_B = 0$ oraz $H_0^{**} : \overline{\boldsymbol{\beta}}_A = 0$
- Cztery możliwe przypadki:
  1. H_0^* odrzucona, H_0^{**} odrzucona - model **A** nie obejmuje modelu **B** i model **B** nie obejmuje modelu **A**
  2. H_0^* nie odrzucona, H_0^{**} odrzucona - model **A** obejmuje model **B**
  3. H_0^* odrzucona, H_0^{**} nie odrzucona - model **B** obejmuje model **A**
  4. H_0^* nie odrzucona, H_0^{**} nie odrzucona - model **A** jest równoważny modelowi **B**

Test J

- Formułujemy następujący sztuczny model, w którym zagnieżdżone są modele **A** i **B**

$$\boldsymbol{y} = (1-\lambda)\,\boldsymbol{X}_A\boldsymbol{\beta}_A + \lambda\boldsymbol{X}_B\boldsymbol{\beta}_B + \boldsymbol{\varepsilon}$$

- Tak sformułowany model jest modelem nieliniowym
- Pokazano jednak, że hipotezę H_0: $\lambda = 0$ można przetestować w sposób następujący

1. przeprowadzamy regresję $\boldsymbol{y}$ na $\boldsymbol{X}_B$ uzyskujemy wartości dopasowane $\widehat{\boldsymbol{y}}_B = \boldsymbol{X}_B\boldsymbol{b}_B$
2. przeprowadzamy regresję $\boldsymbol{y}$ na $\boldsymbol{X}_A$ i $\widehat{\boldsymbol{y}}_B$
3. Testujemy istotność parametru stojącego przy $\widehat{\boldsymbol{y}}_B$

- Jeśli hipoteza H_0: $\lambda = 0$ nie może zostać odrzucona, to model **A** obejmuje model **B**