

Testowanie hipotez statystycznych

Wykład 9

Natalia Nehrebecka Stanisław Cichocki

28 listopada 2018

Plan zajęć

- 1 Rozkład estymatorów MNK w KMRL
 - Rozkład estymatora b
- 2 Testowanie hipotez prostych
 - Hipotezy proste - test t
 - Badanie istotności zmiennych w modelu
- 3 Przedziały ufności
 - dla parametrów
- 4 Testowanie hipotez łącznych
 - Hipotezy łączne - test F
- 5 Pytania teoretyczne

Dodatkowe założenie

Oprócz założeń o:

- braku autokorelacji i homoskedastyczności: $Var(\varepsilon) = \sigma^2 \mathbf{I}$
- zerowej wartości oczekiwanej: $E(\varepsilon) = \mathbf{0}$

Dochodzi założenie o:

- normalności rozkładu błędów losowych.

Reasumując:

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Wiemy już że:

- $\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$
- $E(b) = \beta$ oraz $\text{Var}(b) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Stąd:

$$b \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

- Hipoteza teoretyczna powinna być **zgodna z danymi**, aby móc ją uwzględnić w procesie estymacji
- Niskie prawdopodobieństwo uzyskania obserwowanego odchylenia oszacowań od narzuconych przez teorię ograniczeń \Rightarrow odrzucamy hipotezę teoretyczną
- Aby określić to prawdopodobieństwo, niezbędna jest znajomość **rozkładu** estymatorów i statystyk testujących

Testowanie hipotez przy użyciu statystyki t

Testowanie hipotez statystycznych polega na badaniu prawdopodobieństwa uzyskania otrzymanej wartości statystyki testowej przy założonej hipotezie zerowej H_0 .

Hipotezy proste dotyczą pojedynczego parametru modelu lub kombinacji liniowej parametrów

Rozkład statystyki t

$$t = \frac{b_k - \beta_k}{\hat{se}(b_k)}$$

$$t \sim t_{N-K}$$

Przykład (1/2)

Założmy, że teoria mówi, że pewien parametr modelu, β_k , jest równy określonej wartości, β_k^* , $\beta_k = \beta_k^*$

Jeżeli:

- spełnione są założenia KMRL
- błąd losowy ma rozkład normalny
- teoria jest słuszna / hipoteza zerowa H_0 jest prawdziwa

Przykład (2/2)

Wtedy:

- statystyka testowa:

$$t = \frac{b_k - \beta_k^*}{\hat{se}(b_k)} \sim t_{N-K}$$

- statystyka krytyczna (odczytujemy z tablic rozkładu t-Studenta):

$$t^* = t \left(\underbrace{N - K}_{\text{Stopni swobody}}, \underbrace{1 - \frac{\alpha}{2}}_{\text{Rzad kwantyla}} \right)$$

gdzie: α - poziom istotności

⇒ Jeśli $|t| \geq t^*$ - odrzucamy H_0

⇒ Jeśli $|t| < t^*$ - nie ma podstaw do odrzucenia H_0

Hipotezy dwustronne

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases}$$

Jeśli H_0 jest prawdziwa, wówczas model ma postać:

$$\mathbf{y} = \beta_0 + \cdots + \underbrace{\beta_k}_{0} \mathbf{x}_k + \cdots + \beta_K \mathbf{x}_K + \varepsilon$$

zmienna x_k nie ma znaczenia dla wyjaśnienia zmienności y

Statystyka testowa

- statystyka testowa:

$$t = \frac{b_k}{\hat{se}(b_k)}$$

czyli jest to stosunek wielkości **estymatora parametru** przez estymator jego **odchylenia standardowego**

- statystyka krytyczna (odczytujemy z tablic rozkładu t-Studenta):

$$t^* = t \left(N - K, 1 - \frac{\alpha}{2} \right)$$

- Jeśli $|t| \geq t^*$ - *odrzucaamy* H_0
- Jeśli $|t| < t^*$ - nie ma podstaw do odrzucenia H_0

Wnioskowanie statystyczne

hipotezy dwustronne

$$P(|t| > t^*) = 2[1 - F_{t_{N-K}}(t^*)] = \alpha$$

Obecnie, zamiast stosować wartości krytyczne, oblicza się **p-value** (policzony poziom istotności):

$$2[1 - F_{t_{N-K}}(t)] = p - value$$

- Jeśli $p - value$ poniżej określonego poziomu istotności (np. 0,05) - odrzucamy H_0
- W przeciwnym przypadku - nie ma podstaw do odrzucenia H_0

Przykład

```
. xi: regress ln_dochod wiek i.plec
i.plec          _Iplec_0-1      (naturally coded; _Iplec_0 omitted)
```

Source	SS	df	MS	Number of obs =	1083
Model	28.9189338	2	14.4594669	F(2, 1080) =	45.81
Residual	340.860238	1080	.315611331	Prob > F =	0.0000
Total	369.779172	1082	.341755242	R-squared =	0.0782
				Adj R-squared =	0.0765
				Root MSE =	.56179

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
wiek	.00463	.0017007	2.72	0.007	.001293 .007967
_Iplec_1	-.3191872	.0342242	-9.33	0.000	-.3863407 -.2520337
_cons	6.479705	.0684582	94.65	0.000	6.345379 6.614031

```
. matrix list e(V)
```

```
symmetric e(V) [3,3]
      wiek      _Iplec_1      _cons
wiek   2.892e-06
_Iplec_1 -3.562e-06   .0011713
_cons  -.00010919  -.00043014   .00468652
```

- Jaki jest przedział, w którym z określonym prawdopodobieństwem znajdzie się nieznana wartość parametru β_k .
- Odpowiedź na to pytanie uzyskamy wyznaczając tak zwany przedział ufności.
- Przedział ufności dla nieznanego parametru β_k na poziomie ufności $1 - \alpha$ można skonstruować następująco:

$$\begin{aligned} P(|t| < t^*) &= P\left(\left|\frac{b_k - \beta_k}{\hat{se}(b_k)}\right| < t^*\right) = \\ &= P(b_k - \hat{se}(b_k)t^* < \beta_k < b_k + \hat{se}(b_k)t^*) = 1 - \alpha \end{aligned}$$

gdzie:

$$t^* = t\left(N - K, 1 - \frac{\alpha}{2}\right)$$

Przykład

- Policzyć 95%-procentowy przedział ufności dla β_{plec} .

```
. xi: regress ln_dochod wiek i.plec
i.plec _Iplec_0-1 (naturally coded; _Iplec_0 omitted)
```

Source	SS	df	MS	Number of obs =	1083
Model	28.9189338	2	14.4594669	F(2, 1080) =	45.81
Residual	340.860238	1080	.315611331	Prob > F =	0.0000
				R-squared =	0.0782
				Adj R-squared =	0.0765
Total	369.779172	1082	.341755242	Root MSE =	.56179

ln_dochod	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wiek	.00463	.0017007	2.72	0.007	.001293	.007967
_Iplec_1	-.3191872	.0342242	-9.33	0.000	-.3863407	-.2520337
_cons	6.479705	.0684582	94.65	0.000	6.345379	6.614031

```
. matrix list e(V)
```

```
symmetric e(V) [3,3]
      wiek      _Iplec_1      _cons
wiek      2.892e-06
_Iplec_1  -3.562e-06      .0011713
_cons     -.00010919  -.00043014      .00468652
```

Hipotezy łączne są ważne z punktu widzenia:

- rozważań teoretycznych
- doboru zmiennych do modelu

Uwaga:

Hipotezy łączne nie są równoważne iloczynowi hipotez prostych!

Typowa hipoteza łączna

dana jest układem równań:

$$H_0 : \mathbf{H}\beta = \mathbf{h}$$

gdzie: \mathbf{H} - macierz o pełnym rzędzie wierszowym = g .

Liczba równań w tym układzie nazywana jest **liczbą ograniczeń**

Układ równań:

- zawiera równania liniowo niezależne
- nie jest sprzeczny

Zadanie

W modelu:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

testowana jest hipoteza H_0 postaci:

$$H_0 : \begin{cases} \beta_0 = 0 \\ \beta_1 = \beta_2 \\ \beta_2 + \beta_3 = 1 \end{cases}$$

Znaleźć macierze H i h za pomocą których hipotezę H_0 można zapisać jako $H\beta = h$.

- 1 Wyprowadzić rozkład małopróbkowy estymatora MNK. Jakie założenie, poza standardowymi KMRL, należy w tym przypadku przyjąć?
- 2 Jaką postać ma statystyka służąca do testowania hipotezy o tym, że $\beta_k = \beta_k^*$?
- 3 Mając oszacowanie b_k oraz oszacowanie odchylenia standardowego tego oszacowania $\hat{se}(b_k)$ wyjaśnić w jaki sposób należy zbudować przedział ufności dla β_k . Ilość obserwacji wynosi N , ilość szacowanych parametrów K , a poziom ufności $1 - \alpha$.