round

Testowanie hipotez statystycznych

Wykład 9

Natalia Nehrebecka Stanisław Cichocki

13 grudnia 2014

Plan zajeć

Dodatkowe założenie

Oprócz założeń o:

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

Dochodzi założenie o:

- normalności rozkładu błedów losowych.

Reasumujac:

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

Wiemy już że:

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

Stad:

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

Wiemy już, że:

- $\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}_X\varepsilon$
- macierz $\mathbf{M}_X$ - symetryczna i idempotentna
- rzad macierzy $\mathbf{M}_X = N - K$

Stad:

$$\frac{\sum_{i=1}^{N} e_i^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\varepsilon'\mathbf{M}_X\varepsilon}{\sigma^2} \sim \chi^2_{N-K}$$

Brak korelacji miedzy b a e

$$cov(\mathbf{b}, \mathbf{e}) = 0$$

co implikuje, że:

$$cov(\mathbf{b}, \mathbf{e}'\mathbf{e}) = 0$$

Testowanie hipotez

Hipotezy proste dotycza pojedynczego parametru modelu lub kombinacji liniowej parametrów

Rozkład statystyki t

$$t = \frac{b_k - \beta_k}{\hat{se}(b_k)} = \dots = \frac{\frac{b_k-\beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}}}{\sqrt{\frac{\frac{\mathbf{e}'\mathbf{e}}{\sigma^2}}{N-K}}}$$

Ponieważ: $\frac{b_k-\beta_k}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} \sim N(0,1)$ oraz $\frac{\mathbf{e}'\mathbf{e}}{\sigma^2} \sim \chi^2_{N-K}$, stad

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

Przykład (1/2)

Załóżmy, że teoria mówi, że pewien parametr modelu, β_k, jest równy określonej wartości, $\beta_k^\star$, $\beta_k = \beta_k^\star$
Jeżeli:

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

Przykład (2/2)

Wtedy:

- statystyka testowa:

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

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

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

gdzie: α- poziom istotności
$\Rightarrow$Jeśli $|t| \geq t^\star$ - *odrzucamy* H_0
$\Rightarrow$Jeśli $|t| < t^\star$ - *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 brak podstaw do odrzucenia H_0, wówczas model ma postać:

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

zmienna $\mathbf{x}_k$ nie ma znaczenia dla wyjaśnienia zmienności $\mathbf{y}$

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*):

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

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

Wnioskowanie statystyczne
hipotezy dwustronne

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

  gdzie: $t^{\star}-$ statystyka krytyczna.
- Obecnie, zamiast stosować wartości krytyczne, oblicza sie **p-value** (*policzony poziom istotności*):

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

  gdzie: $t-$statystyka testowa.
- 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
-------------+------------------------------           F(  2,  1080) =   45.81
       Model |  28.9189338     2  14.4594669            Prob > F      =  0.0000
    Residual |  340.860238  1080  .315611331            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
```

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

$$P(|t| < t^\star) = P\left(\left|\frac{b_k - \beta_k}{\hat{se}(b_k)}\right| < t^\star\right) =$$

$$= P\left(b_k - \hat{se}(b_k)t^\star < \beta_k < b_k + \hat{se}(b_k)t^\star\right) = 1-\alpha$$

gdzie:

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

Przykład

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

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

Source	SS	df	MS
Model	28.9189338	2	14.4594669
Residual	340.860238	1080	.315611331
Total	369.779172	1082	.341755242

Number of obs	1083
F(2, 1080)	45.81
Prob > F	0.0000
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
```

Hipotezy łaczne sa ważne z punktu widzenia:

- rozważań teoretycznych
- doboru zmiennych do modelu

Uwaga:

Hipotezy łaczne nie sa równoważne iloczynowi hipotez prostych!

Typowa hipoteza łaczna

dana jest układem równań:

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

gdzie: **H** - macierz o pełnym rzedzie wierszowym $= g$.
Liczba równań w tym układzie nazywana jest **liczba 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 $\mathbf{H}$ i $\mathbf{h}$ za pomoca których hipoteze H_0 można zapisać jako $\mathbf{H}\beta = \mathbf{h}$.

1. (*) Udowodnić, że rozkład sumy kwadratów reszt jest rozkładem χ^2_{N-K} niezależnym od rozkładu b.
2. Wyprowadzić rozkład małopróbkowy estymatora MNK. Jakie założenie, poza standardowymi KMRL, należy w tym przypadku przyjać?
3. Jaka postać ma statystyka służaca do testowania hipotezy o tym, że $\beta_k = \beta_k^\star$?
4. Majac 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 $1 - \alpha$.