Ekonometria

Dekompozycja wariancji zmiennej zależnej, R^2 i jego własności cz. II

Natalia Nehrebecka
Stanisław Cichocki

Wykład 5

Plan wykładu

- Dobroć dopasowania równania regresji. Współczynnik determinacji R^2
  - Dekompozycja wariancji zmiennej zależnej
  - Współczynnik determinacji R^2

Uwaga!

- R^2 jest WYŁĄCZNIE statystyką opisową i nie należy jej stosować do porównywania modeli.

- Przy szacowaniu kilku modeli dla danej zmiennej zależnej z różną liczbą zmiennych objaśniających na podstawie identycznego zbioru danych, korzystanie ze współczynnika determinacji R^2 dla wyboru modelu lepiej dopasowanego do danych empirycznych staje się problematyczne.
- Gdy bowiem dodajemy do równania dalsze zmienne objaśniające to zawsze wzrasta R^2 niezależnie od prawdziwej ważności tych nowododanych zmiennych.

$$placa_i = \beta_1 + \beta_2 wiek_i + \varepsilon_i \qquad R^2 = 5\%$$

$$placa_i = \beta_1 + \beta_2 wiek_i + \beta_3 plec_i + \varepsilon_i \qquad R^2 = 7\%$$

Uwaga!

- Gdy bowiem dodajemy do równania dalsze zmienne objaśniające to zawsze wzrasta R^2 niezależnie od prawdziwej ważności tych nowododanych zmiennych.

- Wiąże się to z **ogólnymi własnościami optymalizacji**.
- Jeśli, poprzez narzucenia ograniczeń, zmniejszymy zbiór, na którym minimalizujemy funkcję celu, to uzyskana w minimum wartość funkcji celu będzie większa lub równa wartości funkcji w minimum dla minimalizacji bez ograniczeń.

Minimalizacja z ograniczeniami i bez ograniczeń

- Analizuje dwie regresji:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + \varepsilon_i$$

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + x_{K+1i}\beta_{K+1} + \varepsilon_i$$

- Model pierwszy można uzyskać z modelu drugiego, jeśli narzucimy

ograniczenie $\beta_{K+1} = 0$

Minimalizacja z ograniczeniami i bez ograniczeń

- Estymator MNK parametrów ***b*** dla **modelu drugiego**:

$$\min_{b_1,b_2,\ldots,b_K,b_{K+1}} s(b_1,b_2,\ldots,b_K,b_{K+1})$$

- Estymator MNK parametrów ***b**** **dla modelu pierwszego**:

$$\min_{b_1,b_2,\ldots,b_K,b_{K+1}} s(b_1,b_2,\ldots,b_K,b_{K+1}) \quad s.t. \quad b_{K+1}=0$$

- S(***b****) – jest wartością funkcji w minimum z warunkami pobocznymi
- S(***b***) – jest wartością funkcji w minimum bez tych warunków

$$S(\boldsymbol{b}^*)=RSS^* \geq S(\boldsymbol{b})=RSS$$

$$TSS^*=TSS$$

Uwaga!

- Nawet dodając do modelu całkowicie bezsensowną zmienną uzyskujemy polepszenie dopasowania.

- Co więcej, jeśli K=N (*liczba parametrów = liczbie obserwacji*), to $\boldsymbol{R^{2=}1}$

- Z tego powodu za miarę dobroci dopasowania zaproponowano nie $\boldsymbol{R^2}$, a tak zwany „skorygowany współczynnik determinacji” $\overline{R}^2$.

Skorygowany współczynnik determinacji $\overline{R}^2$

- $\overline{R}^2$ jest skorygowany ze względu na tak zwaną liczbę stopni swobody, to znaczy ze względu na różnicę między liczbą obserwacji **N**
  a liczbą zmiennych objaśniających **K**.

$$\overline{R}^2 = 1 - \frac{N-1}{N-K}(1-R^2)$$

Zadanie 1

- Oszacowano dwa modele za pomocą MNK na próbie liczącej 12 obserwacji i otrzymano następujące wyniki:

$$a)\hat{y}_i = 1{,}5 + 2x_{2i}, \quad R^2 = 0{,}8$$

$$b)\hat{y}_i = 2 - 2{,}64x_{2i} + 3x_{3i}, \quad R^2 = 0{,}81$$

- Który z powyższych modeli należy wybrać i dlaczego?

Zadanie 2

- Oszacowano dwa modele za pomocą MNK na próbie liczącej 1083 obserwacji:
- **Model A**: ln_zarobki=f(płeć, wiek)
- **Model B**: ln_zarobki=f(płeć, wiek, wiek^2)

Zadanie 2

Model A

Source	SS	df	MS		
Model	28.9189338	2	14.4594669	Prob > F	0.0000
Residual	340.860238	1080	.315611331	R-squared	0.0782
Total	369.779172	1082	.341755242	Root MSE	.56179

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

Zadanie 2

Model B

Source	SS	df	MS		
Model	34.9709245	3	11.6569748	Number of obs	1083
Residual	334.808247	1079	.310294946	F(3, 1079)	37.57
Total	369.779172	1082	.341755242	Prob > F	0.0000
				R-squared	0.0946
				Adj R-squared	0.0921
				Root MSE	.55704

ln_zarobki	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
plec	-.3216684	.0339394	-9.48	0.000	-.3882631	-.2550737
wiek	.0520709	.0108737	4.79	0.000	.0307349	.0734069
wiek2	-.0006115	.0001385	-4.42	0.000	-.0008831	-.0003398
_cons	5.62267	.2055893	27.35	0.000	5.21927	6.02607

Zadanie 2

Porównanie modeli !

Model	R^2	K	$\overline{R}^2$
A	**0.0782**	3	$\overline{R}^2 = 1 - \frac{1083-1}{1083-3}(1-0{,}0782) = 0{,}0765$
B	**0.0946**	4	$\overline{R}^2 = 1 - \frac{1083-1}{1083-4}(1-0{,}0946) = 0{,}0921$

Pytania teoretyczne

1. Wyjaśnić, dlaczego R^2 nie można używać do porównania modeli.

Dziękuję za uwagę