

Diagnostyka w Pakiecie Stata

Karol Kuhl

Zgodnie z twierdzeniem Gaussa-Markowa, estymator MNK w KMRL jest liniowym estymatorem efektywnym i nieobciążonym, co po angielsku opisuje się za pomocą wyrażenia „*BLUE*” – *Best Linear Unbiased Estimator*. Po oszacowaniu parametrów modelu za pomocą MNK, można mieć pewność, że otrzymane estymatory mają takie pożądane właściwości, ale pod warunkiem, że spełnione są założenia KMRL. Dlatego kolejnym krokiem, po oszacowaniu parametrów modelu, powinno być sprawdzenie, czy są one spełnione. Przytaczane poniżej wyniki zostały uzyskane na podstawie obliczeń przeprowadzonych na zbiorze danych `diagnostyka.dta`. Punktem wyjścia było oszacowanie modelu (zwanego dalej pierwotnym) za pomocą polecenia:

```
reg y x1 x2 x3
```

i uzyskanie wyników:

Source	SS	df	MS			
Model	1192044.85	3	397348.282	Number of obs =	100	
Residual	264796.546	96	2758.29736	F(3, 96) =	144.06	
				Prob > F =	0.0000	
				R-squared =	0.8182	
				Adj R-squared =	0.8126	
				Root MSE =	52.519	
Total	1456841.39	99	14715.5696			

	y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	x1	-2.186126	.1864091	-11.73	0.000	-2.556146	-1.816107
	x2	3.127254	.1825507	17.13	0.000	2.764893	3.489614
	x3	.2374342	.183834	1.29	0.200	-.1274735	.6023419
	_cons	488.7165	17.36733	28.14	0.000	454.2426	523.1904

1 Normalność rozkładu składnika losowego

Założeniem podlegającym sprawdzeniu jest normalność rozkładu składnika losowego, którego realizacjami są reszty z oszacowanego modelu – służy do tego np. test Jarque’a-Bery (JB). **Hipoteza zerowa tego testu mówi o tym, że składnik losowy, ma rozkład normalny. Hipoteza alternatywna – że składnik losowy, nie ma rozkładu normalnego.** Statystyka testująca w tym teście ma rozkład χ^2 z 2 stopniami swobody. Po oszacowaniu modelu należy wygenerować nową zmienną zawierającą reszty z modelu. Najłatwiej w tym celu posłużyć się poleceniem:

```
predict e, residual
```

Mając reszty, test JB wywołuje się za pomocą polecenia:

```
sktest e
```

i otrzymuje następujące wyniki:

Skewness/Kurtosis tests for Normality				
Variable	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint Prob>chi2
e	0.187	0.658	1.99	0.3705

Test JB bada na ile jednocześnie skośność i kurtoza rozkładu empirycznego zmiennej różni się od wartości teoretycznych. Wartość statystyki testującej znajduje się w kolumnie przedostatniej (1.99), a odpowiadające jej prawdopodobieństwo (0.3705) – w kolumnie ostatniej. W powyższym przykładzie nie ma podstaw do odrzucenia hipotezy zerowej mówiącej o normalności rozkładu składnika losowego.

2 Błąd specyfikacji

Założeniem podlegającym sprawdzeniu jest odpowiedni dobór zmiennych objaśniających do modelu oraz odpowiedni dobór ich formy funkcyjnej. Do weryfikacji tego założenia służy test *RESET*, polegający na ponownym oszacowaniu parametrów modelu, w którym do zbioru regresorów dołączone zostały nowe zmienne będące naturalnymi potęgami oszacowanych wartości teoretycznych. **Hipoteza zerowa tego testu mówi o tym, że liniowa specyfikacja modelu jest właściwa. Hipoteza alternatywna – że liniowa specyfikacja modelu nie jest właściwa.** W pakiecie Stata wyniki tego testu uzyskuje się za pomocą polecenia:

```
ovtest, rhs
```

które wyświetli następujące wyniki:

```
Ramsey RESET test using powers of the fitted values of y
Ho: model has no omitted variables
F(3, 93) = 4.00
Prob > F = 0.0100
```

W tym przypadku mała wartość prawdopodobieństwa (0.010) nakazuje odrzucić hipotezę zerową mówiącą o właściwej specyfikacji modelu.

3 Heteroskedastyczność (test Breuscha-Pagana)

Do weryfikacji założenia o stałości wariancji służy m.in. test Breuscha-Pagana (BP), w którym szacowana jest regresja ze zmienną objaśnianą będącą kwadratem unormowanych reszt (podzielonych przez odchylenie standardowe). **Hipoteza zerowa tego testu mówi o tym, że składnik losowy modelu jest homoskedastyczny. Hipoteza alternatywna – że składnik losowy modelu jest heteroskedastyczny.** Statystka testująca ma rozkład χ^2 o $m - 1$ stopniach swobody, gdzie m to liczba zmiennych objaśniających w drugim modelu. W pakiecie Stata wyniki tego testu otrzymuje się wpisując polecenie:

```
hettest, rhs
```

Przykładowe wyniki są następujące:

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of y
chi2(1) = 5.66
Prob > chi2 = 0.0174
```

W tym przypadku mała wartość prawdopodobieństwa (0.0174) nakazuje odrzucić hipotezę zerową mówiącą o homoskedastyczności składnika losowego.

4 Heteroskedastyczność (test White'a)

Innym testem służącym do weryfikacji założenia o stałości wariancji jest test White'a, w którym szacowana jest regresja ze zmienną objaśnianą będącą resztami z pierwotnego modelu i zmiennymi objaśniającymi będącymi kwadratami i iloczynami krzyżowymi zmiennych objaśnianych z pierwotnego modelu. **Hipoteza zerowa tego testu mówi o tym, że składnik losowy modelu jest homoskedastyczny. Hipoteza alternatywna – że składnik losowy modelu jest heteroskedastyczny.** Statystyka testująca ma rozkład χ^2 o $m - 1$ stopniach swobody, gdzie m to liczba zmiennych objaśniających w drugim modelu. W pakiecie Stata wyniki tego testu otrzymuje się wpisując polecenie:

```
imtest, white
```

Przykładowe wyniki są następujące:

```
White's test for Ho: homoskedasticity
      against Ha: unrestricted heteroskedasticity
      chi2(9)      =      16.19
      Prob > chi2  =      0.0630
```

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	16.19	9	0.0630
Skewness	3.66	3	0.3000
Kurtosis	0.77	1	0.3801
Total	20.63	13	0.0806

Przy poziomie istotności $\alpha = 5\%$ hipoteza o homoskedastyczności składnika losowego jest przyjmowana, co wynika z wartości prawdopodobieństwa (0.0630). Dolna część wyników wyświetlanych w przypadku tego testu przez pakiet Stata dotyczy innego testu.

5 Autokorelacja składnika losowego (test Durbina-Watsona)

Inną ważną kwestią diagnostyczną jest brak autokorelacji składnika losowego. Hipotezę o autokorelacji pierwszego rzędu weryfikuje się za pomocą testu Durbina-Watsona (DW). **Hipoteza zerowa tego testu mówi o tym, że składnik losowy modelu nie podlega autokorelacji pierwszego rzędu. Hipoteza alternatywna – że składnik losowy podlega autokorelacji pierwszego rzędu.** Autokorelacja najczęściej występuje w przypadku szeregów czasowych – wtedy chronologia wydarzeń pozwala jednoznacznie uporządkować obserwacje. Aby w pakiecie Stata policzyć statystykę DW, należy określić, która zmienna opisuje przebieg czasu. Zmienna `t` w zbiorze `diagnostyka.dta` opisuje przebieg czasu i fakt ten należy zgłosić pakietowi za pomocą polecenia:

```
tsset t
```

Wtedy, po oszacowaniu modelu regresji, można użyć polecenia:

```
dwstat
```

otrzymując:

```
Durbin-Watson d-statistic( 4, 100) = 2.277351
```

Prawdopodobieństwo odpowiadające tej statystyce nie jest podawane, natomiast do odczytania z tablic statystycznych są wartości krytyczne tego testu. Dla poziomu istotności $\alpha = 5\%$, 3 regresorów i 100 obserwacji, wartości krytyczne wynoszą: $d_L = 1.613$ i $d_U = 1.736$. W związku z tym:

1. W przedziale $(0, 1.613)$ hipoteza zerowa jest odrzucana na rzecz alternatywnej, mówiącej o tym, że autokorelacja pierwszego rzędu jest dodatnia.
2. W przedziale $[1.613, 1.736)$ nie można podjąć decyzji odnośnie przyjęcia albo odrzucenia hipotezy zerowej.
3. W przedziale $[1.736, 2.264)$ hipoteza zerowa jest przyjmowana.
4. W przedziale $(2.264, 2.387)$ nie można podjąć decyzji odnośnie przyjęcia albo odrzucenia hipotezy zerowej.
5. W przedziale $(2.387, 4)$ hipoteza zerowa jest odrzucana na rzecz alternatywnej, mówiącej o tym, że autokorelacja pierwszego rzędu jest ujemna.

W analizowanym przypadku wartość krytyczna należy do przedziału niekonkluzywności, w związku z czym nie można za pomocą tego testu stwierdzić, czy autokorelacja pierwszego rzędu występuje, czy nie.

6 Autokorelacja skł. losowego (test Breuscha-Godfrey'a)

Testowanie autokorelacji (dowolnego rzędu) składnika losowego odbywa się za pomocą testu Breuscha-Godfrey'a. Polega on na oszacowaniu modelu regresji, w którym zmienna objaśnianą są reszty z modelu pierwotnego, a zmiennymi objaśniającymi – opóźnione reszty z modelu pierwotnego. autokorelacji pierwszego rzędu weryfikuje się za pomocą testu Durбина-Watsona (DW). **Hipoteza zerowa tego testu mówi o tym, że składnik losowy modelu nie podlega autokorelacji żadnego rzędu. Hipoteza alternatywna – że składnik losowy podlega autokorelacji któregoś rzędu.** W celu zweryfikowania, czy w modelu zachodzi autokorelacja I, II i III rzędu należy wpisać:

```
. bgodfrey, lags(1 2 3)
```

Otrzymane wyniki:

```
Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	2.176	1	0.1402
2	2.947	2	0.2291
3	3.429	3	0.3300

H0: no serial correlation

świadczą o braku auto korelacji wymienionych rzędów.

7 Test stabilności strukturalnej

W celu stwierdzenia, czy współczynniki regresji są takie same dla wszystkich obserwacji w zbiorze, należy przeprowadzić test Chow'a. W odróżnieniu od testów opisanych wcześniej nie jest on wywoływany za pomocą jednego polecenia. Można jednak w celu jego przeprowadzenia posłużyć się odpowiednio zkonstruowanymi interakcjami zmiennych z modelu i zmiennych binarnych opisujących podgrupy obserwacji. **Hipoteza zerowa tego testu mówi o tym, że współczynniki regresji w różnych grupach obserwacji są takie same. Hipoteza alternatywna – że współczynniki regresji w różnych grupach obserwacji różnią się od siebie.** Przykładowo, w celu sprawdzenia, czy parametry regresji są takie same dla pierwszych 50 obserwacji, jak dla pozostałych 50 obserwacji, należy utworzyć następujące zmienne:

```

gen d=0
gen dx1=0
gen dx2=0
gen dx3=0
replace d=1 if t>50
replace dx1=x1 if t>50
replace dx2=x2 if t>50
replace dx3=x3 if t>50

```

Zmienna d rozróżnia grupy obserwacji, a zmienne dx tworzą odpowiednie interakcje. Następnie należy oszacować model regresji:

```
reg y x1 x2 x3 d dx1 dx2 dx3
```

otrzymując:

Source	SS	df	MS			
Model	1201331.71	7	171618.816	Number of obs =	100	
Residual	255509.678	92	2777.27911	F(7, 92) =	61.79	
Total	1456841.39	99	14715.5696	Prob > F =	0.0000	
				R-squared =	0.8246	
				Adj R-squared =	0.8113	
				Root MSE =	52.7	

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	-2.311932	.2639053	-8.76	0.000	-2.836071	-1.787793
x2	3.206111	.2791789	11.48	0.000	2.651637	3.760584
x3	.1160369	.2712475	0.43	0.670	-.4226842	.6547579
d	-26.44456	35.72695	-0.74	0.461	-97.40137	44.51225
dx1	.2429014	.3759222	0.65	0.520	-.5037126	.9895154
dx2	-.2238238	.3735186	-0.60	0.550	-.9656639	.5180164
dx3	.2179172	.3731992	0.58	0.561	-.5232886	.959123
_cons	504.472	27.01627	18.67	0.000	450.8154	558.1286

Wyniki testu Chowa otrzymuje się testując łączną nieistotność zmiennych: d , $dx1$, $dx2$, $dx3$, za pomocą polecenia:

```
test (d=0) (dx1=0) (dx2=0) (dx3=0)
```

otrzymując:

```

( 1) d = 0
( 2) dx1 = 0
( 3) dx2 = 0
( 4) dx3 = 0
      F( 4, 92) = 0.84
      Prob > F = 0.5058

```

W tym przypadku widać, że hipoteza zerowa jest przyjmowana i, że oszacowania parametrów w obydwu podpróbach nie są od siebie istotnie różne.