METODA NAJMNIEJSZYCH KWADRATÓW

Założenia MNK:

(Z1) zmienne objaśniające są nielosowe i nieskorelowane ze składnikiem losowym,

(Z2) $r(X) = k + 1 \ \leq n$ (liczba obserwacji n jest większa od liczby szacowanych parametrów),

(Z3) wartość oczekiwana składnika losowego wynosi 0,

(Z4) $var(\varepsilon) = D^2(\varepsilon) = \sigma^2 I, \quad \sigma^2 < \infty$ (macierz wariancji-kowariancji reszt jest macierzą diagonalną, w której wszystkie elementy na głównej przekątnej są równe σ^2, tzn. każdy składnik losowy ma jednakową wariancję i składniki losowe nie są ze sobą skorelowane).

Twierdzwnie Gaussa-Markowa

Przy spełnionych założeniach Z1 - Z4 estymator α wyznaczony za pomocą MNK jest estymatorem: liniowym, zgodnym, nieobciążonym i najefektywniejszym w klasie liniowych i nieobciążonych estymatorów wektora parametrów modelu (jest BLUE - Best Linear Unbiased Estimator).

$\alpha = (X^T X)^{-1} \cdot X^T Y$

NORMALNOŚĆ ROZKŁADU SKŁADNIKA LOSOWEGO – TEST JARQUE-BERA

Jeśli składnik losowy jednorównaniowego liniowego modelu ekonometrycznego ma rozkład normalny, to estymator uzyskany za pomocą MNK ma własności, użyteczne w konstruowaniu testów statystycznych w celu sprawdzenie różnych cech modelu. Pozytywna ocena tego założenia ma więc zasadnicze znaczenie w procesie weryfikacji modelu.

Jednym z testów służących do weryfikacji rozkładu normalnego jest test Jarque-Bera. Hipotezy testu:

H_0: składnik losowy modelu ma rozkład normalny,

$H_1: \sim H_0$

ISTOTNOŚĆ POJEDYNCZEJ ZMIENNEJ OBJAŚNIAJĄCEJ – TEST t-STUDENTA

Hipotezy testu:

H_0: $\alpha_j = 0$ (zmienna objaśniająca X_j jest nieistotna, brak wpływu zmiennej X_j na Y)
H_1: $\alpha_j \neq 0$ (zmienna objaśniająca X_j jest istotna)

Jeśli reszty z modelu mają rozkład normalny oraz prawdziwa jest H_0, to zmienna losowa $t = \frac{\alpha_j}{S_{\alpha_j}}$ ma rozkład t-Studenta z n-(k+1) stopniami swobody. Z tablic rozkładu t-Studenta odczytywana jest wartość krytyczna t_*. Jeżeli $|t| > t_*$, to należy odrzucić H_0.

ISTOTNOŚĆ PODZBIORU ZMIENNYCH OBJAŚNIAJĄCYCH – UOGÓLNIONY TEST WALDA

Za pomocą testu Walda można zweryfikować hipotezę zerową o istotności części lub wszystkich zmiennych objaśniających w modelu.

Hipotezy testu (przy testowaniu istotności wszystkich zmiennych w modelu):

H_0: $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ (brak wpływu zmiennych $X_1, X_2, \ldots, X_k$ na Y)
H_1: $co\ najmniej\ jeden\ z\ parametrów\ \alpha_j, j = 1, \ldots, k\ jest\ różny\ od\ zera$

W teście tym wyznacza się statystykę F, która w przypadku spełnienia założenia dotyczącego rozkładu normalnego reszt ma rozkład F-Snedecora.

Wyznaczanie statystyki:

1. Budujemy modele:
   Model podstawowy: $Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k + \varepsilon$
   Model rozszerzony: $Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k + \alpha_{k+1} X_{k+1} + \cdots + \alpha_{k+m} X_{k+m} + \varepsilon$
2. Wyznaczamy reszty z modeli: e – reszty z modelu podstawowego (z mniejszą liczbą zmiennych), r – reszty z modelu rozszerzonego (z większą liczbą zmiennych).
3. Obliczamy wartość statystyki F:
   $$F = \frac{e^T e - r^T r}{r^T r} \cdot \frac{n - (k+1) - m}{m}$$
   Statystyka F ma rozkład F-Snedecora o stopniach swobody $r_1 = m$, $r_2 = n - (k+1)$. Jeżeli $F > F^*$, to odrzucamy H_0, gdzie:
   H_0: $\alpha_{k+1} = \cdots = \alpha_{k+m} = 0$
   H_1: $co\ najmniej\ jeden\ z\ parametrów\ \alpha_j, j = k+1, \ldots, k+m\ jest\ różny\ od\ zera$

WSPÓŁCZYNNIK DETERMINACJI $\boldsymbol{R^2}$ ($\boldsymbol{R^2} \in [\mathbf{0}, \mathbf{1}]$)

Współczynnik determinacji pozwala ocenić, w jakim stopniu model objaśnia zmienność zmiennej Y.

Współczynnik determinacji zwykły

- W modelu musi występować wyraz wolny.
- Interpretacja jest poprawna pod warunkiem, że badane związki są liniowe.

Współczynnik determinacji skorygowany

- Stosowany, gdy liczba szacowanych parametrów jest niewiele mniejsza od liczby dostępnych obserwacji (zwykły współczynnik determinacji może być wówczas zawyżony).

Współczynnik determinacji niescentrowany

- Stosowany, gdy w modelu nie występuje wyraz wolny

STABILNOŚĆ POSTACI ANALITYCZNEJ MODELU – TEST RAMSEYA

Test ten upewnia nas czy wybrana liniowa postać modelu jest dobrze dobrana do opisu zmienności danej zmiennej objaśnianej w zależności od wartości zmiennych objaśniających. Test bada stabilność postaci analitycznej modelu. Podobnej informacji dostarcza test liczby serii.

Hipotezy testu:

H_0: *wybór postaci analitycznej modelu jest prawidłowy*
H_1: *wybór postaci analitycznej modelu nie jest prawidłowy*

Opis kolejnych czynności:

1. Szacujemy parametry modelu postaci:
   $y_t = \alpha_0 + \alpha_1 x_{1t} + \cdots + \alpha_k x_{kt} + \varepsilon_t, \;\; t = 1,2,\dots,n$
2. Obliczamy wartości teoretyczne zmiennej objaśnianej $\hat{y}_t$ oraz współczynnik determinacji modelu R_I^2.
3. Szacujemy parametry modelu postaci:
   $y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \beta_{k+1}\hat{y}_t^2 + \beta_{k+2}\hat{y}_t^3 + \mu_t$
4. Dla nowego modelu wyznaczamy współczynnik determinacji R_{II}^2.
5. Badamy czy przyrost wartości współczynnika determinacji jest statystycznie istotny. Wyznaczamy wartość statystyki:
   $$F = \frac{R_{II}^2 - R_I^2}{1 - R_{II}^2} \cdot \frac{n - (k+3)}{2}$$
   Statystyka F ma rozkład F-Snedecora o stopniach swobody $r_1 = 2, r_2 = n - (k+3)$. Jeżeli $F > F^*$, to odrzucamy H_0.

Jeśli w teście tym odrzucona zostanie hipoteza zerowa, to należy zmienić postać analityczną modelu.

STABILNOŚĆ PARAMETRÓW MODELU – TEST CHOWA

Test ten służy do weryfikacji hipotezy o stabilności parametrów modelu. Stabilność parametrów oznacza, że oceny parametrów uzyskane dla obserwacji z różnych okresów nie różnią się istotnie.

Hipotezy testu:

$H_0: \alpha = \beta = \delta$ (*parametry modelu są stabilne*)
H_1: *parametry modelu nie są stabilne*

Opis kolejnych czynności:

1. Szacujemy parametry modelu postaci:
   $y_t = \alpha_0 + \alpha_1 x_{1t} + \cdots + \alpha_k x_{kt} + \varepsilon_t, \;\; t = 1,2,\dots,n$
2. Obliczamy resztową sumę kwadratów: $RSK = \sum_{t=1}^{n} e_t^2$.
3. Dzielimy okres obserwacji na dwa podokresy. Podział ten może być podziałem subiektywnym albo wynikającym z analizy zjawiska bądź procesu opisywanego przez model.

4. Szacujemy odpowiednie modele dla podzielonych danych:
   $y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_{1t}, \quad t = 1,2, \dots, n_1$
   $y_t = \delta_0 + \delta_1 x_{1t} + \cdots + \delta_k x_{kt} + \varepsilon_{2t}, \quad t = n_1, n_1 + 1, \dots, n$
   Zakładamy, że składniki losowe mają rozkład normalny.
5. Obliczamy resztową sumę kwadratów dla obu nowych modeli: RSK_1, RSK_2.
6. Wyznaczamy wartość statystyki:
   $$F = \frac{RSK - (RSK_1 + RSK_2)}{RSK_1 + RSK_2} \cdot \frac{n - 2(k+1)}{k+1}$$
   Statystyka F ma rozkład F-Snedecora o stopniach swobody $r_1 = k + 1, r_2 = n - 2(k + 1)$.
   Jeżeli $F > F^*$, to odrzucamy H_0.

LINIOWOŚĆ MODELU EKONOMETRYCZNEGO – TEST LICZBY SERII

Sformułowanie jednorównaniowego modelu ekonometrycznego postaci:
$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k + \varepsilon$$
jest równoważne z przyjęciem założenia o liniowej zależności zmiennej objaśnianej od zmiennych objaśniających. Weryfikacja tego założenia jest niezbędna do prawidłowej interpretacji współczynnika determinacji.

Test liczby serii

H_0: postać modelu jest dobrze dobrana, model jest liniowy

H_1: $\sim H_0$

Rozważmy ciąg reszt, w którym rozróżnia się dwa rodzaje elementów: reszty dodatnie i reszty ujemne. Reszty równe zero pomijamy. Serią nazywamy każdy podciąg jednakowych elementów, który poprzedzony jest i po którym następuje element różny od elementu podciągu.

Przykładowo ciąg: ++-++---+--+-- ma 8 serii.

Procedura testowa:

1) Za pomocą MNK szacujemy parametry modelu liniowego i obliczamy reszty.
2) Porządkujemy ciąg reszt (według rosnących wartości wybranej zmiennej objaśniającej, a w przypadku szeregów czasowych według numerów okresów obserwacji).
3) Wyznaczamy liczbę symboli „+” (n_1), liczbę symboli „-” (n_2) oraz liczbę serii (S).
4) Jeśli $n_1, n_2 \leq 20$, to z tablic rozkładu liczby serii odczytujemy S_1^* dla $\frac{1}{2}\alpha$ i S_2^* dla $1 - \frac{1}{2}\alpha$. Jeśli $S_1^* < S < S_2^*$, to nie ma podstaw do odrzucenia H_0. W przeciwnym przypadku odrzucamy H_0.
5) Jeśli $n_1, n_2 > 20$, to korzystając ze wzorów:
   $E(S) = \frac{2n_1n_2}{n} + 1, \quad D^2(S) = \frac{2n_1n_2(2n_1n_2 - n)}{(n-1)n^2}$
   wyznaczamy statystykę Z, która ma rozkład $N(0,1)$: $\quad Z = \frac{S - E(S)}{D(S)}$

PROGNOZA PUNKTOWA

Do prognozowania posługujemy się modelem oszacowanym na podstawie danych z okresów t=1,2,…,n, mającym postać:

$$y_t = \alpha_0 + \alpha_1 x_{1t} + \cdots + \alpha_k x_{kt} = \mathbf{Xa}$$

Symbolem τ oznaczamy okres prognozy, $\tau > n$. Przyjmujemy, że wektor zmiennych objaśniających dla okresu τ ma postać:

$$\mathbf{x}_\tau^\mathrm{T} = [1 \quad x_{1\tau} \quad x_{2\tau} \quad \cdots \quad x_{k\tau}]$$

Prognozę punktową wartości zmiennej objaśnianej w okresie τ wyznaczamy jako:

$$y_\tau^P = \mathbf{x}_\tau^\mathrm{T}\mathbf{a}$$

gdzie $\mathbf{a}$ to wektor parametrów modelu.

Ocena ex post prognozy punktowej:

- Błąd prognozy ex post: $y_\tau - y_\tau^P$
- Względny błąd prognozy ex post: $\frac{y_\tau - y_\tau^P}{y_\tau}$
- Średni błąd prognozy ex post: $ME = \frac{1}{S}\sum_{\tau=1}^{S}(y_\tau - y_\tau^P)$
- Średni absolutny błąd: $MAE = \frac{1}{S}\sum_{\tau=1}^{S}|y_\tau - y_\tau^P|$
- Pierwiastek błędu średniokwadratowego: $RMSE = \sqrt{\frac{1}{S}\sum_{\tau=1}^{S}(y_\tau - y_\tau^P)^2}$
- Średni absolutny błąd procentowy: $MAPE = \frac{1}{S}\sum_{\tau=1}^{S}\left|\frac{y_\tau - y_\tau^P}{y_\tau}\right| \cdot 100$

PROGNOZA PRZEDZIAŁOWA

Zakładamy, że pozytywnie zweryfikowana została hipoteza o normalności rozkładu składnika losowego modelu. Zmienna losowa $u = \frac{|y_\tau - y_\tau^P|}{S_\tau^P}$ ma wtedy rozkład t-Studenta z $n-(k+1)$ stopniami swobody.

S_τ^P – średni błąd prognozy ex ante:

$$S_\tau^P = \sqrt{S^2(\mathbf{x}_\tau^\mathrm{T}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{x}_\tau + 1)}$$

S^2 – wariancja składnika losowego:

$$S^2 = \frac{\mathbf{e}^\mathrm{T}\mathbf{e}}{n-(k+1)}$$

Konstruujemy przedział ufności:

$$P\{|u| < t_*\} = 1 - \alpha$$

czyli:

$$P\{y_\tau^P - t_* S_\tau^P < y_\tau < y_\tau^P + t_* S_\tau^P\} = 1 - \alpha$$

gdzie $1 - \alpha$ jest przyjętym poziomem ufności, a t_* jest odpowiednią wartością z rozkładu t-Studenta. Otrzymujemy przedział liczbowy $(y_\tau^P - t_* S_\tau^P,\ y_\tau^P + t_* S_\tau^P)$, stanowiący przedziałową prognozę zmiennej Y na okres τ.