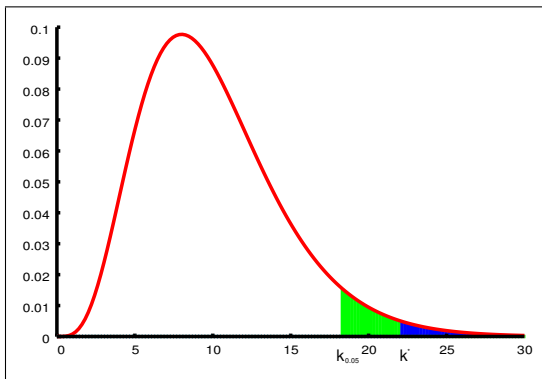


3. Testowanie hipotez

3.1. Testowanie hipotez przy pomocy programów ekonometrycznych

Przejdziemy teraz do testowania hipotez statystycznych. Tradycyjna metoda testowania hipotez polega na porównywaniu wartości statystyki testowej k^* z wartością krytyczną k_{α} zdefiniowaną jako liczba, dla której, jeśli H_0 jest prawdziwe, $F(k_{\alpha}) = 1 - \alpha$. Rysunek 1 ilustruje tę procedurę testowania dla rozkładu χ^2_{10} . Załóżmy, że uzyskana statystyka $k^* = 22$ a $\alpha = 0,05$. Dla takiego poziomu istotności wartość krytyczna $k_{0,05} = 18,3$. Ponieważ $k^* > k_{0,05}$, statystyka k^* wpada do obszaru krytycznego i odrzucamy hipotezę zerową. Na rysunku 2 pokazano przypadek, kiedy $k^* = 12$. Teraz $k^* < k_{0,05}$, statystyka wpada do obszaru przyjęć i tym samym nie ma podstaw do odrzucenia H_0 .



Rysunek 1: Testowanie hipotez $k_{0,05} < k^*$

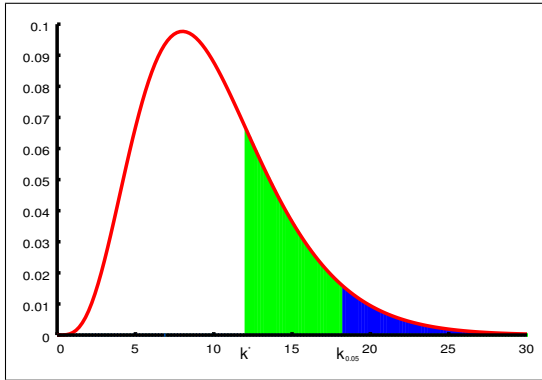
Wadą tradycyjnej procedury testowania jest to, że przed przeprowadzeniem testu musimy znaleźć w tablicach wartość krytyczną k_α . W czasach kiedy nie istniały komputery procedura ta była jedynym wygodnym sposobem testowania hipotez. Obecnie, dla standardowych rozkładów, istnieją procedury numeryczne, za pomocą których znaleźć można wartości dystrybuant. Można więc łatwo policzyć $\alpha^* = 1 - F(k^*)$. Zauważmy, że α odpowiada polu pod funkcją gęstości dla $x \in (k_\alpha, \infty)$ a α^* polu pod tą funkcją dla $x \in (k^*, \infty)$. Na rysunku 1 α jest sumą pola zakreśkowanego jaśniej i pola zakreśkowanego ciemniej podczas gdy α^* odpowiada polu zakreśkowanemu ciemniej. Łatwo zauważyć, że dla $k^* > k_\alpha$ mamy $\alpha^* < \alpha$ a dla $k^* < k_\alpha$ mamy, że $\alpha^* > \alpha$. Wynika z tego, że dla $\alpha^* < \alpha$ odrzucamy hipotezę zerową, a dla $\alpha^* > \alpha$ nie ma podstaw do odrzucenia H_0 . Wartość α^* można interpretować jako prawdopodobieństwo, że wartość statystyki testowej osiągnie większą lub równą tej, którą uzyskaliśmy w wyniku testowania. Wydaje się oczywiste, że uzyskanie bardzo mało prawdopodobnych, przy założeniu prawdziwości H_0 , wartości α^* będzie nas skłaniało do odrzucenia tej hipotezy.

Większość współczesnych pakietów ekonometrycznych automatycznie liczy α^* używając do tego rozkładu prawdopodobieństwa odpowiedniego dla konkretnej statystyki testowej. Aby przeprowadzić test wystarczy porównać uzyskane α^* z założonym poziomem istotności α . Dodatkową zaletą takiego podejścia jest to, że łatwo sprawdzić jak zmiana założonego poziomu istotności wpłynie na rezultat wnioskowania statystycznego.

W przypadku zilustrowanym na rysunku 1 $\alpha^* = F_{\chi_{10}^2}(22) = 0,015$ co oznacza, że odrzucamy hipotezę zerową dla 5% poziomu istotności, ponieważ $\alpha^* < 0,05$. Analogicznie dla rysunku 2 $\alpha^* = F_{\chi_{10}^2}(12) = 0,285$ i $\alpha^* > \alpha$ i nie ma podstaw do odrzucenia H_0 . Zauważmy, że w przypadku pierwszym hipotezę zerową odrzucamy na poziomie istotności $\alpha = 0,05$ ale nie ma podstaw, by ją odrzucić dla poziomu istotności $\alpha = 0,01$, skoro $\alpha^* = 0,015 > 0,01$.

3.2. Testowanie hipotez prostych

Twierdzenie 3.1 *Estymator \mathbf{b} w KML ma rozkład normalny o średniej β i wariancji $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.*



Rysunek 2: Testowanie hipotez $k_{0,05} > k^*$

Dowód.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon}$$

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

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$$

wynika z tego, że

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

■

Wniosek 3.2 Estymator $\delta' \mathbf{b}$ w KMRL ma rozkład normalny o średniej $\delta' \boldsymbol{\beta}$ i wariancji $\sigma^2 \delta' (\mathbf{X}'\mathbf{X})^{-1}$.

Wniosek 3.3 Element b_i wektora b ma rozkład normalny o średniej β_i i wariancji $\Sigma_{ii} = \sigma^2 (\mathbf{X}'\mathbf{X})_{ii}^{-1}$, gdzie $(\mathbf{X}'\mathbf{X})_{ii}^{-1}$ oznacza i -ty element diagonalny macierzy $(\mathbf{X}'\mathbf{X})^{-1}$.

Dowód. Wniosek ten otrzymujemy z poprzedniego lematu, gdy przyjmiemy $\boldsymbol{\delta}' = [0 \ \dots \ 0 \ 1]$ i 1 jest na i -tym miejscu. ■

Lemat 3.4 Suma kwadratów reszt podzielona przez wariancję błędu losowego ma rozkład

$$\frac{\sum_{i=1}^n e_i^2}{\sigma^2} = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} \sim \chi_{n-K}^2$$

Dowód. Wiemy z założeń *KMRL*, że $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ a wcześniej dowiedliśmy $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ i \mathbf{M} jest macierzą idempotentną oraz $\text{Rank}(\mathbf{M}) = n - K$. Tym samym spełnione są założenia twierdzenia 17.78. ■

Lemat 3.5 Estymator *MNK* \mathbf{b} i wektor reszt \mathbf{e} są niezależne (nieskorelowane).

Dowód.

$$\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{M}\boldsymbol{\varepsilon}$$

$$\begin{aligned} \text{Cov}(\mathbf{b}, \mathbf{e}) &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \text{Var}(\boldsymbol{\varepsilon}) \mathbf{M}' = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' (\sigma^2 \mathbf{I}) \mathbf{M}' \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{M}' = \mathbf{0} \end{aligned}$$

■

Twierdzenie 3.6 Statystyka t dana wzorem $t = \frac{b_i - \beta_i}{\sqrt{\mathbf{S}_{ii}}}$ ma rozkład *t*-studenta z $n - K$ stopniami swobody.

Dowód.

$$t = \frac{b_i - \beta_i}{\sqrt{\mathbf{S}_{ii}}} = \frac{b_i - \beta_i}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{ii}^{-1}}} = \frac{\frac{b_i - \beta_i}{\sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})_{ii}^{-1}}}}{\sqrt{\frac{\mathbf{e}'\mathbf{e}}{\sigma^2 (n-K)}}}$$

Z wniosku 3.3 i lematu 3.4 wynika, że

$$\frac{b_i - \beta_i}{\sqrt{\Sigma_{ii}}} = \frac{b_i - \beta_i}{\sigma \sqrt{(\mathbf{X}'\mathbf{X})_{ii}^{-1}}} \sim N(\mathbf{0}, 1)$$

a

$$\frac{\mathbf{e}'\mathbf{e}}{\sigma^2} \sim \chi_{n-K}^2$$

a z lematu 3.5, że \mathbf{e} i b_i są niezależne. Spełnione są więc założenia twierdzenie 17.76. ■

Wniosek 3.7 *Hipotezę statystyczną o tym, że poszczególne elementy wektora \mathbf{b} równają się zeru można testować za pomocą statystyki $t = \frac{b_i}{\sqrt{\mathbf{S}_{ii}}}$.*

Literatura: Steward (1991) str. 44-49, Green (1997) str. 263-267, Goldberger (1972) str. 227-229.

3.3. Zmienne pominięte i zmienne nieistotne

Załóżmy, że mamy dwa modele

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (3.1)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (3.2)$$

Zauważmy, że model (3.1) jest równoważny modelowi (3.2) pod warunkiem, że prawdziwa jest hipoteza zerowa, że prawdziwe jest ograniczenie:

$$H_0 : \boldsymbol{\gamma} = \mathbf{0} \quad (3.3)$$

Z tego powodu model (3.1) nazywamy niekiedy modelem z ograniczeniami (*restricted model*). Dla estymatorów $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ podamy warunki pierwszego rzędu maksymalizacji sumy kwadratów reszt.

• Zmienne pominięte

Jeśli wyestymujemy model (3.1) a w rzeczywistości $\gamma \neq \mathbf{0}$, co oznacza, że prawdziwy jest model (3.2), to pojawi się problem zmiennych pominiętych. Estymator $\tilde{\beta}$ w (3.1) sytuacji równy

$$\begin{aligned}\tilde{\beta} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{u}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\gamma + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}\end{aligned}$$

Wartość oczekiwana tego estymatora jest równa

$$\mathbb{E}(\tilde{\beta}) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\gamma + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbb{E}(\mathbf{u}) = \beta + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Z}\gamma$$

co implikuje, że estymator ten jest obciążony. Wariancja estymatora jest równa

$$\text{Var}(\tilde{\beta}) = \mathbb{E}\left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Zauważmy, że istnieją dwa ważne przypadki, dla których pominięcie zmiennej nie powoduje obciążenia estymatora. Pierwszy przypadek jest trywialny i zachodzi dla $\gamma = \mathbf{0}$. Bardziej interesująca sytuacja zachodzi, gdy $\mathbf{X}'\mathbf{Z} = \mathbf{0}$, to jest dla przypadku, kiedy \mathbf{X} i \mathbf{Z} są ortogonalne. Równość taka będzie zachodzić w przybliżeniu¹, gdy $\text{Cov}(\mathbf{X}, \mathbf{Z}) = 0$. Wynika z tego, że problem związany z obciążeniem estymatora parametrów na skutek pominięcia ważnych zmiennych objaśniających pojawia się jedynie wtedy, gdy zmienne pominięte są skorelowane ze zmiennymi, które zostały uwzględnione w modelu.

Przykład 3.8 *Zbudowano prosty model liniowy, w którym zmienną objaśnianą była stopa przyrostu naturalnego na określonym terenie a zmienną objaśniającą ilość bocianów zamieszkujących na tym terenie. Stwierdzono, że ilość bocianów istotnie wpływa na ilość rodzących się dzieci. Czyżby istotnie bociany przynosiły dzieci?*

¹Dla zmiennych objaśniających w postaci odchyłeń od średnich oryginalnych zmiennych.

Odpowiedź: W Polsce znacznie wyższy przyrost naturalny odnotowuje się na wsi niż w mieście. Na wsi mieszka też znacznie więcej bocianów. W modelu pominięto ważną zmienną związaną z tym, czy dany teren jest w dominującej części terenem wiejskim czy miejskim. Gdyby zmienną to uwzględniono w modelu to parametr przed zmienną "ilość bocianów mieszkających w okolicy" spadłby prawdopodobnie do zera.

• Zmienne nieistotne

W przypadku, kiedy $\gamma = \mathbf{0}$ i prawdziwym modelem jest model (3.1) a estymujemy model (2.10), to wśród regresorów pojawiają się zmienne, które nie są istotne w prawdziwym modelu. Na podstawie wniosku z twierdzenia FWL wiemy, że estymator $\hat{\beta}$ w modelu (3.2) jest równy

$$\hat{\beta} = (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z(\mathbf{X}\beta + \varepsilon) = \beta + (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\varepsilon$$

Wartość oczekiwana estymatora wynosi

$$\mathbb{E}(\hat{\beta}) = \beta + (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbb{E}(\varepsilon) = \beta$$

co oznacza, że jest estymator jest w dalszym ciągu nieobciążony. Wariancję tego estymatora znajdujemy w sposób następujący

$$\begin{aligned}\text{Var}(\hat{\beta}) &= \mathbb{E}\left\{\left[(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\varepsilon\right]\left[(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\varepsilon\right]'\right\} \\ &= (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbb{E}(\varepsilon\varepsilon')\mathbf{M}_Z\mathbf{X}(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\end{aligned}$$

Zauważmy, że estymator $\hat{\beta}$ jest estymatorem liniowym, ponieważ $\hat{\beta} = (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{y}$. Z twierdzenia Gaussa-Markowa wiemy, że jeśli model (3.1) jest prawdziwy i spełnione są pozostałe założenia KMRL, to estymatorem o minimalnej wariancji jest estymator $\tilde{\beta}$. Wynika z tego, że $\text{Var}(\hat{\beta}) \geq \text{Var}(\tilde{\beta})$, co oznacza, że estymator $\hat{\beta}$ jest estymatorem nieefektywnym.

Literatura: Steward (1991) str. 58-66, Green (1997) str. 401-404, Goldberger (1972) str. 255-262, Theil (1979)

3.4. Testowanie łącznej nieistotności zmiennych

Estymator γ do wyprowadzonej postaci estymatora $\hat{\beta}$

$$\hat{\gamma} = \gamma + (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{M}_X\mathbf{u}$$

ma rozkład

$$\hat{\gamma} \sim N\left(\gamma, \sigma^2 (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1}\right)$$

Wynika z tego, że

$$\frac{(\hat{\gamma}-\gamma)' (\mathbf{Z}'\mathbf{M}_X\mathbf{Z}) (\hat{\gamma}-\gamma)}{\sigma^2} \sim \chi_g^2$$

z drugiej strony wiemy, że

$$\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma^2} \sim \chi_{n-K}^2$$

oraz $\hat{\gamma}$, $\hat{\mathbf{u}}$ są niezależne, ponieważ w MNK \mathbf{b} i \mathbf{e} są niezależne. Można więc sformułować następujące twierdzenie

Twierdzenie 3.9 *Do testowania hipotezy, że $H_0 : \gamma = \mathbf{0}$ w modelu (3.2) można zastosować statystykę*

$$\frac{\hat{\gamma}'(\mathbf{Z}'\mathbf{M}_X\mathbf{Z})\hat{\gamma}}{\frac{g}{n-K}} = \frac{S_R - S}{\frac{S}{n-K}} \sim F(g, n-K)$$

gdzie $S_R = \mathbf{e}'\mathbf{e}$ a $S = \hat{\mathbf{u}}'\hat{\mathbf{u}}$.

Wniosek 3.10 Do testowania hipotezy, że wszystkie zmienne w modelu poza stałą są nieistotne można wykorzystać statystykę

$$\frac{n - K}{K - 1} \frac{R^2}{1 - R^2} \sim F(K, n - K)$$

Dowód. Dla takiej hipotezy $S = RSS$, $S_R = TSS$ a ilość ograniczeń $g = K - 1$.

$$\begin{aligned} \frac{(S_R - S)/(K - 1)}{S/(n - K)} &= \frac{(TSS - RSS)/(K - 1)}{RSS/(n - K)} = \frac{n - K}{K - 1} \frac{ESS}{RSS} \\ &= \frac{n - K}{K - 1} \frac{R^2}{1 - R^2} \end{aligned}$$

ponieważ

$$\frac{R^2}{1 - R^2} = \frac{ESS/TSS}{1 - ESS/TSS} = \frac{ESS/TSS}{(TSS - ESS)/TSS} = \frac{ESS}{RSS}$$

■

Inny sposób testowania hipotezy $H_0 : \gamma = \mathbf{0}$ można wyprowadzić zauważając, że przyjmując reszty z regresji \mathbf{y} na \mathbf{X} za zmienną zależną i przeprowadzając regresję tych reszt na pełnej liście zmiennych otrzymujemy następujące estymatory β^* , γ^* :

$$\beta^* = (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1} \mathbf{X}'\mathbf{M}_Z\mathbf{e} = (\mathbf{X}'\mathbf{M}_Z\mathbf{X})^{-1} \mathbf{X}'\mathbf{M}_Z(\mathbf{y} - \mathbf{X}\tilde{\beta}) = \hat{\beta} - \tilde{\beta}$$

$$\gamma^* = (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{M}_X\mathbf{e} = (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{M}_X(\mathbf{y} - \mathbf{X}\tilde{\beta}) = (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1} \mathbf{Z}'\mathbf{M}_X\mathbf{y} = \hat{\gamma}$$

Ponieważ

$$\begin{aligned} \mathbf{e} - \mathbf{X}\beta^* - \mathbf{Z}\gamma^* &= \mathbf{y} - \mathbf{X}\tilde{\beta} - \mathbf{X}\beta^* - \mathbf{Z}\gamma^* = \mathbf{y} - \mathbf{X}\tilde{\beta} - \mathbf{X}(\hat{\beta} - \tilde{\beta}) - \mathbf{Z}\hat{\gamma} \\ &= \mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\gamma} = \hat{\mathbf{u}} \end{aligned}$$

więc reszty z regresji e na X , Z są równe \hat{u} . W takiej regresji $TSS = e'e$ a $RSS = \hat{u}'\hat{u}$. Skorzystamy teraz z tego, że

$$\frac{e'e - \hat{u}'\hat{u}}{\sigma^2} \sim \chi_g^2$$

i zamiast σ^2 zastosujemy zgodny² estymator $\hat{\sigma}^2$ równy $\frac{e'e}{n}$. Możemy teraz sformułować następujące twierdzenie

Twierdzenie 3.11 *Do testowania hipotezy, że $H_0 : \gamma = \mathbf{0}$ w modelu (3.2) można zastosować statystykę*

$$\frac{S_R - S}{\frac{S_R}{n}} = \frac{(e'e - \hat{u}'\hat{u})}{(e'e/n)} = n \left(\frac{TSS - RSS}{TSS} \right) = nR^2 \xrightarrow{D} \chi_g^2$$

gdzie R^2 jest współczynnikiem określoności dla regresji zmiennych w modelu (3.2) na resztach z wyestymowanego modelu 3.1.

Wniosek 3.12 *Gdy hipotezą zerową jest, że wszystkie współczynniki poza stałą są równe 0, do testowania można wykorzystać statystykę*

$$nR^2 \xrightarrow{D} \chi_K^2$$

gdzie R^2 jest współczynnikiem określoności dla regresji bez ograniczeń.

Wadą statystyki nR^2 jest to, że jej rozkład został wyprowadzony jedynie dla dużych prób.

Literatura: Steward (1991) str. 65-71, Green (1997) str. 268-270

3.5. Testowanie hipotez złożonych

Dla modelu

²Dowód zgodności tego estymatora podamy w kontekście własności estymatorów MNK w dużych próbach.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

można sformułować hipotezę dotyczącą wektora parametrów $\boldsymbol{\beta}$ w postaci układu g równań liniowych

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

przy czym zakładamy, że wszystkie równania w tym układzie równań są liniowo niezależne (macierz \mathbf{H} ma pełen rząd). Podzielmy macierz \mathbf{H} w taki sposób, że $\mathbf{H} = (\mathbf{H}_1, \mathbf{H}_2)$ i \mathbf{H}_2 jest nieosobliwe. Podzielmy odpowiednio $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ i $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. Wtedy

$$\mathbf{H}\boldsymbol{\beta} = \mathbf{H}_1\boldsymbol{\beta}_1 + \mathbf{H}_2\boldsymbol{\beta}_2$$

i $\boldsymbol{\beta}_2 = \mathbf{H}_2^{-1}(\mathbf{h} - \mathbf{H}_1\boldsymbol{\beta}_1)$. Podstawmy ten wynik do modelu zapisanego w postaci

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

z otrzymamy

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\mathbf{H}_2^{-1}(\mathbf{h} - \mathbf{H}_1\boldsymbol{\beta}_1) + \mathbf{u}$$

tak, że przeniesieniu na prawą stronę $\mathbf{X}_2\mathbf{H}_2^{-1}\mathbf{h}$ dostajemy

$$(\mathbf{y} - \mathbf{X}_2\mathbf{H}_2^{-1}\mathbf{h}) = (\mathbf{X}_1 - \mathbf{X}_2\mathbf{H}_2^{-1}\mathbf{H}_1)\boldsymbol{\beta}_1 + \mathbf{u}$$

W modelu tym po prawej znajdują się wyłącznie elementy, których wartości są nam z góry znane, a prawa strona jest funkcją liniową względem parametru $\boldsymbol{\beta}_1$. Możemy teraz zdefiniować model³:

³Ten model także można nazwać modelem z ograniczeniami.

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}_1 + \mathbf{u},$$

gdzie $\mathbf{y}^* = \mathbf{y} - \mathbf{X}_2 \mathbf{H}_2^{-1} \mathbf{h}$ a $\mathbf{X}^* = \mathbf{X}_1 - \mathbf{X}_2 \mathbf{H}_2^{-1} \mathbf{H}_1$ i przeprowadzić za pomocą *MNK* estymację $\boldsymbol{\beta}_1$. Estymator *MNK* parametru $\boldsymbol{\beta}_2$ otrzymujemy ze wzoru $\mathbf{b}_2 = \mathbf{H}_2^{-1} (\mathbf{h} - \mathbf{H}_1 \mathbf{b}_1)$, gdzie \mathbf{b}_1 jest estymatorem $\boldsymbol{\beta}_1$ otrzymanym z poprzedniej regresji.

Opisany powyżej sposób znajdowania estymatora $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2)$, polegający na bezpośrednim podstawianiu ograniczeń do równania regresji, jest równoważny do przeprowadzenia minimalizacji z ograniczeniami z użyciem następującej funkcji Lagrange'a

$$S_R(\mathbf{b}_R) = (\mathbf{y} - \mathbf{X}\mathbf{b}_R)'(\mathbf{y} - \mathbf{X}\mathbf{b}_R) - \boldsymbol{\lambda}'(\mathbf{H}\mathbf{b}_R - \mathbf{h})$$

Warunki pierwszego rzędu dla takiej minimalizacji są następujące

$$\frac{\partial S_R}{\partial \mathbf{b}_R} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}_R - \mathbf{H}'\boldsymbol{\lambda} = \mathbf{0}$$

$$\frac{\partial S_R}{\partial \boldsymbol{\lambda}} = -(\mathbf{H}\mathbf{b}_R - \mathbf{h}) = \mathbf{0}$$

mnożąc obustronnie pierwszy warunek przez $\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}$, rozwiązując dla $\boldsymbol{\lambda}$ i wykorzystując drugi warunek otrzymujemy

$$\boldsymbol{\lambda} = -2[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}']^{-1}(\mathbf{H}\mathbf{b} - \mathbf{h})$$

gdzie $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Po podstawieniu $\boldsymbol{\lambda}$ do pierwszego warunku i rozwiązaniu powstałego równania ze względu na \mathbf{b}_R dostajemy

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\left[\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\right]^{-1}(\mathbf{H}\mathbf{b} - \mathbf{h})$$

Jeśli H_0 jest prawdziwe to $E(\mathbf{b}_R) = \boldsymbol{\beta}$ i estymator jest nieobciążony. Jeśli jednak H_0 jest fałszywe, to otrzymamy estymator obciążony i obciążenie będzie równe

$$E(\mathbf{b}_R) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} (\mathbf{H}\boldsymbol{\beta} - \mathbf{h})$$

Jeśli H_0 jest prawdziwa, to wariancja \mathbf{b}_R wynosi

$$\text{Var}(\mathbf{b}_R) = \text{Var}(\mathbf{b}) - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} \mathbf{H} (\mathbf{X}'\mathbf{X})^{-1}$$

Ponieważ $\text{Var}(\mathbf{b})$ i $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} \mathbf{H} (\mathbf{X}'\mathbf{X})^{-1}$ są dodatnio określone więc wariancja \mathbf{b}_R jest mniejsza od wariancji \mathbf{b} .

Aby wprowadzić rozkład statystyki testowej zauważmy, że

$$\begin{aligned} \mathbf{b} &\sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right) \\ (\mathbf{H}\mathbf{b} - \mathbf{h}) &\sim N\left[\mathbf{0}, \sigma^2 \mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}'\right] \end{aligned} \quad (3.4)$$

Reszty dla równania z ograniczeniami są dane wzorem

$$\mathbf{e}_R = \mathbf{y} - \mathbf{X}\mathbf{b}_R = \mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} (\mathbf{H}\mathbf{b} - \mathbf{h})$$

co implikuje, że

$$\mathbf{e}_R = \mathbf{e} + \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} (\mathbf{H}\mathbf{b} - \mathbf{h}) \quad (3.5)$$

i

$$\mathbf{e}'_R \mathbf{e}_R = \mathbf{e}'\mathbf{e} + (\mathbf{H}\mathbf{b} - \mathbf{h})' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} (\mathbf{H}\mathbf{b} - \mathbf{h})$$

$$S_R - S = (\mathbf{H}\mathbf{b} - \mathbf{h})' \left[\mathbf{H} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{H}' \right]^{-1} (\mathbf{H}\mathbf{b} - \mathbf{h}) \quad (3.6)$$

W efekcie

$$\frac{S_R - S}{\sigma^2} \sim \chi_g^2$$

Ponieważ $(S_R - S)$ jako funkcja \mathbf{b} jest niezależne od S , które jest funkcją \mathbf{e} , więc możemy sformułować następującą twierdzenie:

Twierdzenie 3.13 *Do testowania hipotezy zerowej $H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ można zastosować statystykę*

$$F = \frac{\frac{S_R - S}{g}}{\frac{S}{n - K}} \sim F(g, n - K) \quad (3.7)$$

gdzie S_R jest sumą kwadratów reszt w modelu z ograniczeniami a S jest sumą kwadratów reszt w modelu bez ograniczeń.

Po to, by uzyskać S_R można wyestymować model, w którym ograniczenia podstawiono do równania regresji.

Twierdzenie 3.14 *Do testowania ograniczeń $\mathbf{H}\boldsymbol{\beta} = \mathbf{h}$ można zastosować statystykę*

$$nR^2 \xrightarrow{D} \chi_g^2$$

gdzie R^2 jest współczynnikiem określoności dla regresji reszt z modelu z ograniczeniami na zmiennych w modelu bez ograniczeń.

Uwaga 3.15 *Twierdzenia 3.9 i 3.11 są szczególnymi przypadkiem twierdzeń 3.13 i 3.14.*

Literatura: Steward (1991) str. 72-74, Chow (1995) str. 64-68, Theil (1979) str. 155-162.

3.6. Metodologia testowania hipotez

W kontekście procedur służących do testowania hipotez prostych i złożonych oraz problemu współliniowości omówimy problem metodologicznie poprawnego testowania hipotez.

Rozpatrzmy przypadek hipotezy złożonej $H_0 : \beta_1 = \dots = \beta_K = 0$ testowanej przy złożonym poziomie istotności α . Pojawia się pytanie, czy testowanie tej hipotezy jest równoważne testowaniu szeregu hipotez prostych $H_0^1 : \beta_1 = 0; \dots; H_0^K : \beta_K = 0$, przy czym założony poziom istotności dla każdej z tych hipotez wynosi α . Zakładamy przy tym, że w drugim przypadku hipoteza H_0 zostaje odrzucona, gdy odrzucona zostaje choćby jedna z hipotez prostych H_0^1, \dots, H_0^K . Sprzeczna z intuicją odpowiedź na to pytanie brzmi, że procedury te nie są równoważne.

Dla najprostszego przypadku, kiedy statystyki testowe dla każdej z hipotez prostych są od siebie niezależne prawdziwy poziom istotności jest równy prawdopodobieństwu α^* , że jedna lub więcej z hipotez H_0^1, \dots, H_0^K zostanie odrzucona. Prawdopodobieństwo α^* można policzyć zauważając, że prawdopodobieństwa zdarzenia odwrotnego⁴ wynosi $(1 - \alpha)^K$. W konsekwencji

$$\alpha^* = 1 - (1 - \alpha)^K$$

Różnicę między założonym poziomem istotności α i prawdopodobieństwem α^* nazywamy obciążeniem Lovella. Dla omawianego przypadku $\lim_{K \rightarrow \infty} \alpha^* = 1$, co oznacza, że dla dużej ilości testowanych hipotez prostych prawdopodobieństwo błędu drugiego rodzaju zbliża się do 1. Problem ten związany jest z tak zwanym przekopywaniem danych (*data mining*). Jeśli wyjdziemy od wystarczająco dużego zbioru zmiennych wyjściowych i badać będziemy kolejno istotność zmiennych, to prawie zawsze znajdziemy pewne zmienne o istotnych statystykach t , nawet wtedy, gdy wszystkie zmienne w danym zbiorze są w rzeczywistości nieistotne.

Przykład 3.16 Grupa socjologów postanowiła przetestować hipotezę, że fakt urodzenia się pod konkretnym znakiem Zodiaku ma wpływ na losy respondentów. Do tego celu przeanalizowano

⁴To jest zdarzenia, że nie będzie podstaw do odrzucenia H_0^1, \dots, H_0^K .

bazę danych zawierającą datę urodzenia respondenta i jego dochody. Data urodzenia posłużyła do stworzenia 12 zmiennych zero-jedynkowych determinujących znak Zodiaku, pod którym urodził się respondent. Zmienną zależną były dochody, które stanowić miały miarę sukcesu zawodowego. Na poziomie istotności $\alpha = 0,1$ stwierdzono, że znaki Zodiaku mają istotny wpływ na kariery zawodowe respondentów, ponieważ istotna okazała się zmienna zero-jedynkowa oznaczająca, że respondent urodził się pod znakiem Lwa.

Oczywiście wynik ten był najprawdopodobniej rezultatem błędnej procedury testownia. Prawdziwy poziom istotności dla założonego α wynosi $\alpha^* = 1 - (0.9)^{12} \approx 0,72$, a więc w przypadku, kiedy prawdziwa jest hipoteza zerowa w 72% przypadków uzyskujemy co najmniej jedną statystycznie istotną statystykę t . Prawidłowa procedura testowania powinna polegać na łącznym przetestowaniu, przy użyciu statystyki F , hipotezy o nieistotności wszystkich zmiennych zero-jedynkowych.

Kolejne testownie hipotez prostych zamiast testowania hipotezy łącznej nie zawsze musi prowadzić do wyższego prawdopodobieństwa odrzucenia hipotezy zerowej. Dobrym przykładem jest tu model, w którym występuje współliniowość między x_1 i x_2 . W takim przypadku statystyki t dla x_1 i x_2 mogą być niskie ponieważ wariancje estymatorów b_1 i b_2 są wysokie. Jednak spadek sumy kwadratów reszt w przypadku usunięcia zmiennych x_1 i x_2 może być duży pod warunkiem, że obie te zmienne są silnie skorelowane z y . W takim przypadku może się zdarzyć, że nie będzie podstaw do odrzucenia odrzucenia $H_0^1 : \beta_1 = 0$ i $H_0^2 : \beta_2 = 0$ na poziomie istotności α ale odrzucimy, na tym samym poziomie istotności, hipotezę $H_0 : \beta_1 = \beta_2 = 0$.

Dodatkowe trudności pojawiają się jeśli hipotezy proste testowane są w ramach modeli z ograniczeniami. Powiedzmy, że szukając w pewnym zbiorze danych zmiennych egzogenicznych mających objaśniać zmienną y , testujemy kolejno istotność zmiennych w modelach szczegółowych $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, gdzie x_i jest kolejną testowaną zmienną. Jeśli w rzeczywistości na y_i wpływa więcej niż jedna zmienna z tego zbioru danych, wtedy estymatory b_1 dla *wszystkich* analizowanych modeli szczegółowych są asymptotycznie obciążone ze względu na istnienie zmiennych pominiętych. Tym samym *wszystkie* statystyki t będą miały nietypowe rozkłady i nawet dla hipotez prostych

procedura ta będzie dawać nieprawidłowe wyniki.

Wnioskiem z powyższych rozważań jest, że przy testowaniu hipotez należy unikać wielokrotnego testowania hipotez prostych zamiast testowania hipotez złożonych. Najczęściej jednak poza hipotezą $H_0 : \beta_1 = \dots = \beta_K = 0$ interesują nas hipotezy $H_0^1 : \beta_1 = 0; \dots; H_0^K : \beta_K = 0$. Inaczej mówiąc, interesuje nas nie tylko to, czy którakolwiek zmienne w modelu jest istotna ale także to, które z nich są istotne. Proponowanym obecnie rozwiązaniem tego problemu jest metodologia od ogólnego do szczegółowego (general to specific). Mówimy, że hipotezy są zagnieżdzone jeśli można uszeregować tak, że H_0^1 zawiera najmniej ograniczeń, H_0^2 zawiera ograniczenia zawarte w H_0^1 plus pewne dodatkowe ograniczenia i tak dalej aż do H_0^K zawierającej najwięcej ograniczeń. Sytuację taką zapisujemy jako $H_0^1 \subset H_0^2 \subset \dots \subset H_0^K$. W takim przypadku możemy sekwencyjnie testować hipotezy od H_0^1 aż do momentu, kiedy hipoteza H_0^i zostanie odrzucona.

Przykład 3.17 *Przypuśćmy, że modelujemy poziom dochodu gospodarstwa domowego na podstawie następujących charakterystyk demograficzno-społecznych: miejsce zamieszkania (miasto, wieś), płeć głowy gospodarstwa (mężczyzna, kobieta) oraz poziomu wykształcenia głowy (podstawowe, średnie, wyższe). Ćacznie w takim modelu będziemy mieć 7 zmiennych zero-jedynkowych. Załóżmy, że interesują nas następujące hipotezy:*

- H_0^1 : dla poziomu dochodu gospodarstwa nie ma znaczenia płeć głowy gospodarstwa.
- H_0^2 : dla poziomu dochodu gospodarstwa nie ma znaczenia płeć głowy gospodarstwa i miejsce zamieszkania.
- H_0^3 : dla poziomu dochodu gospodarstwa nie mają znaczenia wszystkie charakterystyki.

W modelu tym $H_0^1 \subset H_0^2 \subset H_0^3$. Stosując metodologię od ogólnego do szczegółowego powinniśmy najpierw przetestować hipotezę łączną, że 2 zmienne zero-jedynkowe związane z płcią są nieistotne. W przypadku braku podstaw do odrzucenia tej hipotezy testujemy hipotezę łączną, że 4 zmienne zero-jedynkowe związane z miejscem zamieszkania i płcią są nieistotne. Jeśli z kolei nie

ma podstaw do odrzucenia tej hipotezy, to powinniśmy przetestować hipotezę o łącznej nieistotności wszystkich zmiennych.

Jeśli przekopujemy duże ilości zmiennych, wtedy przetestowanie części hipotez łącznych może być niemożliwe z racji zbyt małej ilości stopni swobody. Z tego powodu należy zaczynać od zbioru danych, których ewentualna przydatność w objaśnianiu danego zjawiska da się teoretycznie lub intuicyjnie uzasadnić. Sposób uszeregowania hipotez zagnieżdżonych może mieć także istotny wpływ na wynik testowania. Uszeregowanie to powinno, jeśli to możliwe, mieć jakieś uzasadnienie.

Literatura: Charemza i Deadman (1997) str. 23-36 i 75-102