

B. Testování hypotéz a oblastí spolehlivosti v jednorovnicovém modelu

B1. Testování jednoho regresního koeficientu

Z předchozích již odvozených vztahů víme, že platí :

$$SSE = e'e = \varepsilon' M \varepsilon \quad , \text{ kde}$$

$M = I_T - X(X'X)^{-1}X'$ je idempotentní matice mající hodnotu $T-k$, neboť – jak praví příslušná věta z lineární algebry – hodnota idempotentní matice je rovna její stopě.

Vzhledem k tomu, že v klasickém normálním lineárním regresním modelu je vektor náhodných složek ε normovaným normálním vektorem, každá $\varepsilon_j \approx N(0, \sigma^2)$ pro $j = 1, 2, \dots, T$, je SSE idempotentní kvadratickou formou o hodnotě $T-k$, (přesně řečeno : kvadratickou formou s idempotentní maticí M). Platí tedy, že

Výraz (kvadratická forma s náhodnými proměnnými) $\frac{SSE}{\sigma^2}$ je rozdělen jako χ^2_{T-k} neboli že výraz SSE je rozdělen jako $\sigma^2 \cdot \chi^2_{T-k}$.

Víme, že pro odhad rozptylu platí vztah $s^2 = \hat{\sigma}^2 = \frac{SSE}{(T-k)}$; a proto tedy

$$\text{výraz } \frac{s^2}{\sigma^2} \text{ je rozdělen jako } \frac{\chi^2_{T-k}}{(T-k)}$$

Uvažujme dále podíl

$$\frac{b_j - \beta_j}{s_{bj}} = \frac{(b_j - \beta_j)}{\frac{s_{bj}}{\sigma_{bj}}} \quad , \text{ kde}$$

s_{bj} je odhadnutá směrodatná odchylka parametru b_j , pro který je

$s_{bj} = s \sqrt{v^{jj}}$ kde $\sqrt{v^{jj}}$ označuje druhou odmocninu z prvku ležícího na j -tém místě hlavní diagonály matice $(X'X)^{-1}$. Výraz na pravé straně v čitateli má normované normální rozdělení $N(0,1)$, zatímco proměnná ve jmenovateli má charakter „ $\frac{s}{\sigma}$ “,

což je druhá odmocnina proměnné mající $\frac{\chi^2_{T-k}}{(T-k)}$ rozdělení.

V důkazu Věty 2 jsme ukázali, že lineární forma $b - \beta = (X'X)^{-1}X'\varepsilon$ je nezávislá na kvadratické formě $SSE = e'e = \varepsilon' M \varepsilon$. To vyplývá ze skutečnosti, že součin matic obou těchto forem, tj.

$$M = I_T - X(X'X)^{-1}X' \quad \text{a} \quad N = X(X'X)^{-1}X' \quad \text{dává nulovou matici :}$$

$$(I_T - X(X'X)^{-1}X')(X(X'X)^{-1}X') = 0 .$$

Proto je lineární forma $b - \beta$ nezávislá také na veličině $\frac{s}{\sigma} = \sqrt{\frac{SSE}{\sigma^2(T-k)}}$.

Odtud je již zřejmé, že veličina

$\frac{b_j - \beta_j}{s_{b_j}}$ má rozdělení t_{T-k} , tedy Studentovo t-rozdělení o $T - k$ stupních volnosti.

(poznamenejme, že tato statistika je vhodná i pro malé výběry tj. pro $T < 30$)

Uvedený výsledek je též základem pro možnost testování hypotézy, že $\beta_j = \beta_j^*$ pro nějakou konkrétní hodnotu β_j^* s použitím Studentova t-rozdělení resp. následně pro konstrukci intervalu spolehlivosti pro parametr β_j . Parametr β_j je zřejmě střední hodnotou normálně rozděleného b_j .

(Test hypotézy i konstrukce intervalu spolehlivosti je obdobná jako při odhadu neznámé střední hodnoty výběrového průměru T nezávislých stejně a normálně rozdělených náhodných veličin při stejném ale neznámém rozptylu).

Jako zvláštní případ se hypotéza verbálně vyjádřená jako „***y se nemění, když x_j se mění***“ vyjádří jako hypotéza, že podmíněná střední hodnota y není ovlivněna hodnotou x_j , což je ekvivalentní hypotéze, že $\beta_j = 0$.

Tzv. t-poměr b_j/s_{b_j} je právě statistikou vhodnou pro testování uvedené hypotézy: Překročí-li (z empirických hodnot spočtený) t-poměr (teoretickou) kritickou hodnotu t-rozdělení o $T-k$ stupních volnosti na hladině významnosti α (např. $\alpha = 0,05$), zamítáme s pravděpodobností $(1 - \alpha)$ - neboli $100 \cdot (1 - \alpha) \%$ - nulovou hypotézu o (skutečné) nulové hodnotě regresního koeficientu β_j . Jinými slovy to znamená, že – s toutéž pravděpodobností posuzováno – je j -tá vysvětlující proměnná do regresní rovnice zařazena oprávněně.

B2. Testování více než jednoho regresního koeficientu

Uvažujme dále hypotézu o celém vektoru β ve tvaru $\beta = \beta^*$. Tato hypotéza odpovídá vyšetření otázky, zda celá skupina zahrnutých vysvětlujících proměnných nabude určených hodnot.

Nejčastějším testovaným případem bývá hypotéza vyjádřená ve tvaru

$$\beta = 0 \quad ,$$

což odpovídá vyšetřování, zda skupina použitých vysvětlujících proměnných (vzata jako celek) se vyznačuje statisticky významným přínosem pro vysvětlení závisle proměnné (pro hypotetické hodnoty to znamená podmínku $\beta^* = 0$).

Poznámka: Je užitečné říci, že v přímé podobě se takovýto test sice aplikuje velmi často, avšak jeho vypovídací hodnota není zvláště u ekonomických regresních vztahů příliš vysoká. Opačné zjištění (tj. nevýznamnost všech zahrnutých vysvětlujících proměnných) je poměrně vzácné, zvláště v případech, kdy regresní rovnice obsahuje větší (cca více než 3) počet vysvětlujících proměnných.

Pro praxi užitečnějším nasazením tohoto testu je případ, kdy testujeme významnost určité podskupiny z celého souboru vysvětlujících veličin (tzn. v počtu 2 až T-1). Zde má obdobně konstruovaný test svůj význam mj. proto, že můžeme variantně zkoumat přínos různých podskupin vysvětlujících proměnných. Přes toto konstatování (a pro jednoduchost) formulujeme test v původní podobě :

Je patrné, že test hypotézy $\beta = \beta^*$ (např. $\beta = 0$) bude založen na rozdílu $b - \beta^*$, tedy na rozdílu vypočtené a domnělé (hypotetické) hodnoty.

Za předpokladu platnosti nulové hypotézy $H_0 : \beta = \beta^*$ bude platit, že

$b - \beta^* = b - \beta = (X'X)^{-1} X' \varepsilon$. V konstrukci testu využijeme statistiku

$$Q = (b - \beta^*)' X' X (b - \beta^*)$$

Za předpokladu platnosti H_0 (tj. pokud platí hypotéza $\beta = \beta^*$), dostaneme

$$Q = \varepsilon' X (X' X)^{-1} X' X (X' X)^{-1} X' \varepsilon = \varepsilon' X (X' X)^{-1} X' \varepsilon = \varepsilon' [I_T - M] \varepsilon = \varepsilon' N \varepsilon ,$$

kde opět $M = I_T - X(X'X)^{-1}X'$, $N = X(X'X)^{-1}X'$.

A) Víme již, že $N = I_T - M$ je idempotentní matice a že její hodnota je

$$h[N] = \text{tr}[N] = \text{tr}[X(X'X)^{-1}X'] = \text{tr}[X'X(X'X)^{-1}] = \text{tr}I_k = k .$$

Tedy Q je kvadratická forma v proměnných ε s idempotentní maticí N o hodnotě k obsahující náhodné veličiny s normálním rozdělením $\varepsilon \approx N(0, \sigma^2)$. (Zdůrazněme, že tak je tomu pouze za předpokladu platnosti nulové hypotézy $\beta = \beta^*$).

Stejně tak z předchozího víme, že za platnosti téže hypotézy $\beta = \beta^*$ bude veličina

$Q = (b - \beta^*)' X' X (b - \beta^*)$ rozdělena jako $\sigma^2 \cdot \chi^2_k$ a následně výraz

$$\frac{(b - \beta^*)' X' X (b - \beta^*)}{k} = \frac{Q}{k} = \frac{\varepsilon' N \varepsilon}{k} \quad \text{bude mít rozdělení} \quad \frac{\sigma^2 \cdot \chi^2_k}{k} .$$

B) Z Věty 2 dále víme, že výraz $SSE = e'e$ (součet čtverců reziduí) lze zapsat jako $SSE = \varepsilon' M \varepsilon$ a že je představován kvadratickou formou v proměnných ε s maticí M, která je symetrická a idempotentní a má hodnotu $T - k$. V důsledku toho má výraz

$SSE = e'e = \varepsilon' M \varepsilon = \varepsilon' [I_T - X(X'X)^{-1}X'] \varepsilon$ rozdělení $\sigma^2 \cdot \chi^2_{T-k}$ a tedy výraz

$$\frac{SSE}{T - k} = \frac{\varepsilon' M \varepsilon}{T - k} \quad \text{má rozdělení} \quad \frac{\sigma^2 \cdot \chi^2_{T-k}}{T - k} .$$

Konečně víme, že

$$M.N = M[I_T - M] = M - M^2 = M - M = 0_T .$$

a tedy, že kvadratické formy Q a SSE jsou lineárně nezávislé (v důsledku ortogonalit matic M a N).

Lze tedy vyslovit tvrzení umožňující otestovat významnost celého souboru vysvětlujících proměnných jako celku pomocí F -rozdělení (odvozeného jako podíl dvou nezávislých náhodných veličin majících χ^2 -rozdělení dělených svými stupni volnosti).

TVRZENÍ: Za platnosti nulové hypotézy $\beta = \beta^*$ bude podílová veličina

$$\frac{Q/k}{SSE/(T-k)}$$

rozdělena jako $F^k_{(T-k)}$, tedy bude mít Fisher-Snedecorovo rozdělení o k a $(T-k)$ stupních volnosti.

Toto tvrzení je základem pro testování hypotézy $\beta = \beta^*$, založíme-li tento test na F -rozdělení. Je zřejmé, že v tomto případě je adekvátní jednostranný test a že oblast zamítnutí nulové hypotézy bude tvořena vysokými hodnotami podílu

$$(*) \quad \frac{Q/k}{SSE/(T-k)} = \frac{(b - \beta^*)' X' X (b - \beta^*)/k}{e' e / (T-k)} = \frac{\varepsilon' N \varepsilon / k}{\varepsilon' M \varepsilon / (T-k)}$$

kteří takto odpovídají vysokým hodnotám Q tj. velkým odchylkám b od β^* .

Je zřejmé, že čím je větší rozdíl $b - \beta^*$, tím je číselník v předchozích výrazech větší a (empiricky spočtená) F -statistika nabývá větší hodnotu – tento případ mluví proti platnosti hypotézy $b = \beta^*$ ve prospěch alternativy $b \neq \beta^*$.

Jestliže má nulová hypotéza (nejčastější) tvar $\beta^* = 0$, lze psát výrazy v čitateli ve tvaru

$$\frac{b' X' X b}{k} = \frac{(Xb)' Xb}{k} = \frac{\hat{y}' \hat{y}}{k}$$

neboli jde o skalární součin vyrovnaných hodnot. Čím je tento skalární součin větší, tím (při neměnicích se hodnotách reziduí e a jejich skalárního součinu $e'e$) je větší pravděpodobnost zamítnutí nulové hypotézy (o nulových hodnotách β^* nebo jinými slovy o nevýznamnosti zvolených vysvětlujících proměnných jako celku).

Poznámka: Všimněme si, že rozdělení výrazu SSE není nijak závislé na hypotetické hodnotě vektoru β^* . Jmenovatel výrazu (*) má tedy vždy χ^2 -rozdělení, zatímco číselník má χ^2 -rozdělení pouze tehdy, platí-li nulová hypotéza $\beta = \beta^*$.

Podíl Q/SSE se používá i v definici koeficientu determinace jako ústřední v ekonometrii používané míry pro vyjádření shody modelu s pozorovanými daty (tzv. „goodness of fit“ testy)

Koeficient determinace (jako vůbec nejčastěji v ekonometrii užívaná míra shody modelu s daty) je definován vztahem

$$R^2 = 1 - \frac{e'e}{y'y} = \frac{\hat{y}' \hat{y}}{y'y} = \frac{b' X' X b}{y'y}$$

pokud předpokládáme, že vysvětlovaná veličina y má nulovou střední hodnotu (jinak je třeba výrazy o střední hodnoty upravit) .

Koeficient R^2 lze tedy interpretovat jako podíl součtu čtverců (centrovaných) vyrovnaných hodnot a součtu čtverců pozorovaných hodnot (závisle proměnné).

Statistickou významnost koeficientu R^2 lze testovat pomocí podílu

$$F_{R^2} = \frac{R^2 / k}{(1 - R^2) / (T - k)}$$

Je-li tento podíl větší než teoretická (v tabulkách uvedená) teoretická hodnota F^* na zvolené hladině významnosti při daných stupních volnosti, zamítneme nulovou hypotézu o nevýznamnosti R^2 ve prospěch tvrzení, že R^2 je v kontextu uvažované regresní rovnice dostatečně vysoký.

Jak známo, s přidáním každé nové vysvětlující proměnné k souboru již existujících vysvětlujících veličin nemůže hodnota R^2 klesnout, ať je přidávána $k+1$ -tá vysvětlující veličina statisticky významná či ne. Z tohoto důvodu má pro posuzování hodnot R^2 u dvou různých specifikací regresních rovnic pro tutéž vysvětlovanou proměnnou vliv počet vysvětlujících proměnných v uvažovaných specifikacích.

Vyjádříme-li totiž (při centrovaných hodnotách vysvětlované proměnné) výrazy vyskytující se v F statistice jako

$$R^2 = \frac{b' X' X b}{y' y} = \frac{Q}{y' y} \qquad 1 - R^2 = \frac{e' e}{y' y} = \frac{SSE}{y' y} \qquad , \text{ pak zřejmě}$$

$$\frac{R^2 / k}{(1 - R^2) / (T - k)} = \frac{\frac{Q}{y' y \cdot k}}{\frac{SSE}{y' y (T - k)}} = \frac{Q / k}{SSE / (T - k)} = F_{R^2}$$

Znamená to tedy , že testovací statistikou konvenčního F-testu vlastně přímo testujeme statistickou významnost koeficientu determinace R^2 .