Multikolinearita

Problém ***multikolinearity*** se váže k matici vysvětlujících proměnných X, jejíž sloupce mohou být někdy (přesně nebo přibližně) lineárně závislé. Tím je zřejmě porušen předpoklad o maximální možné hodnosti k matice X v klasickém (ale i zobecněném) lineárním regresním modelu. Pokud existuje pouze jeden lineární vztah mezi vysvětlujícími proměnnými, mluvíme o kolinearitě, pokud takovýchto vztahů existuje více, hovoří se o multikolinearitě. Pojem multikolinearita se nicméně používá často i v prvním jmenovaném případě.

Negativní důsledky multikolinearity :

V důsledku toho, že matice X nemá plnou hodnost (nebo její sloupce budou téměř lineárně závislé) bude momentová matice $X'X$ singulární (vzácně) nebo téměř singulární (častěji). V důsledku toho bude mít tato matice determinant nulové nebo velmi malé hodnoty a následně toho matice $(X'X)^{-1}$ bude mít značně velké prvky. Pak mají rozptyly parametrů velmi značné hodnoty (a tedy t-statistiky regresních koeficientů signalizují zpravidla nevýznamnost těchto koeficientů). Odhady regresních koeficientů mohou být současně vzájemně silně závislé. Nelze proto rozlišit individuální významnost jednotlivých vysvětlujících proměnných. To nemusí vadit při predikčním využití modelu, zpravidla to však komplikuje analytické uplatnění.

Multikolinearita (exaktní=přesná) je téměř vždy důsledkem poruchy ve struktuře modelu a nemůže být napravena vzetím dodatečných pozorování z téhož vzorku (jde-li o hodnoty vzaté z téhož základního souboru).

Multikolinearita (přibližná) je častěji důsledkem poruchy ve struktuře modelu a jen zřídkakdy může být napravena vzetím dodatečných pozorování z téhož vzorku (z téhož základního souboru).

Přesná multikolinearita vzniká vzácně, a to zpravidla pro neuvědomění si potřeby vyhnout se zařazení skupiny přesně lineárně závislých proměnných do rovnice :

1) zařazením nezávisle proměnné, která má stejnou hodnotu ve všech pozorováních.
2) zařazením vysvětlující proměnné, která vzniká zprůměrováním některých dalších nezávisle proměnných
3) při chybném zařazení **všech** umělých (např. sezónně chápaných) proměnných (jednu je třeba vždy z rovnice vypustit).

Postupy zmírňující negativní důsledky multikolinearity

1) vynechání problémové proměnné = zjednodušení modelu

V principu jde o velmi jednoduchý postup, který může vést k úplné eliminaci problému. Není však vždy zřejmé, kterou proměnnou (proměnné) obětujeme za tímto účelem.

Přednost : Jednoduchost postupu okamžitě řešícího problém
Slabina : V případě, že vynechávané proměnné dle ekonomické teorie mají své opodstatnění v regresní rovnici, dopouštíme se prohřešku ve specifikaci modelu (současně tím porušíme nestrannost odhadu parametrů zjednodušeného modelu).

2) transformace některé z vysvětlujících proměnných

Transformací (např. použitím přechodem k diferencím nebo relativním přírůstkům) některých vysvětlujících proměnných obvykle dosáhneme eliminaci nebo zmírnění multikolinearity (pokud lineárně závisle proměnné ponecháme beze změny). Na druhé straně však použití relativních přírůstků může vést k heteroskedasticitě (náhodných složek) **a použití diferencí analogicky zase k autokorelaci** náhodných složek, **takže řešením jednoho neduhu vyvoláme neduh jiný. Nicméně, tento postup lze vcelku vhodně uplatnit, pokud zařazení transformovaných proměnných neodporuje ekonomické teorii.**

3) aplikace metody hlavních komponent

Jde o postup (známý především z prostředí faktorové analýzy), kterým se množina původních vysvětlujících proměnných transformuje na jinou množinu vysvětlujících proměnných, přičemž individuální přínos každé z těchto transformovaných proměnných (tzv. hlavních komponent) k vysvětlení závisle proměnné je ve statistickém slova smyslu **nezávislý na jiných těchto transformovaných proměnných. Původně korelované vysvětlující proměnné se takto převádějí na nekorelované (dokonce ortonormální) nové vysvětlující proměnné. Obdobně se transformují i parametry modelu. Počet původních a transformovaných proměnných je stejný, transformované proměnné však nemusíme použít všechny.**

Přednost: Přechodem k ortogonálním proměnným (ne nutně ke všem, ale jen k těm, které mají největší vliv na závisle proměnnou) docílíme úplnou eliminaci multikolinearity.

Slabina: Velmi často nejsou nové proměnné nijak rozumně obsahově interpretovatelné (jde o lineární kombinaci původních proměnných), takže přitažlivost techniky postupu může na druhé straně znamenat snížení obsahové průhlednosti kauzálního vztahu.

4) aplikace hřebenové (ridge) regrese [1]

Jde o postup, kterým se uměle zvětší všechny prvky na diagonále matice $(X'X)^{-1}$ o stejnou konstantu (c) za účelem dosažení vyšší hodnoty determinantu této matice. Odhady regresních koeficientů přitom získáme jako

$$_{RLS}\hat{\beta} = (X'X + c.I_k)^{-1}X'y$$

Přednost : Determinant matice $X'X$ lze zvýšit přidáním konstanty c o libovolnou hodnotu, takže lze dosáhnout výrazného zvýšení hodnoty determinantu matice $X'X + c.I_k$ a následně snížení směrodatných odchylek parametrů (tudíž se posílí významnost t-statistik regresních koeficientů)

[1] Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Non-orthogonal Problems. *Technometrics.* 12/1970 s.55-67.

<u>Slabina</u> : Jde o umělý postup, který vede k tomu, že získané odhady nejsou nestranné (mají však v jistém rozsahu aditivního zvýšení diagonály menší rozptyly). Existují nicméně jistá vodítka, jak volit konstantu c aditivně přidávanou k diagonále matice $X'X$, aby přednost převýšila slabinu.[2]

Obvykle se pro hodnotu konstanty c doporučuje tato volba:

$$c = \frac{k.s^2}{{}_{OLS}\hat{\beta}' \cdot {}_{OLS}\hat{\beta}}$$ **, kde**

kde k je počet vysvětlujících proměnných rovnice

s^2 je (metodou OLS) odhadnutý rozptyl náhodných složek ε

${}_{OLS}\hat{\beta}$ je odhad regresních parametrů β pořízený estimátorem OLS

<u>5) nasazením Steinovy odhadové funkce</u>[3]
založené na principu kritéria minima kvadratické ztrátové funkce ve tvaru

$$Min\,({}_{OLS}\hat{\beta} - \beta)' X^{*}{}' X^{*} ({}_{OLS}\hat{\beta} - \beta)$$

kde X^* je matice normovaných vysvětlujících proměnných (uložených opět po sloupcích), tedy s vlastnostmi $\sum_i x_{ij} = 0$ a $\sum_i x_{ij}^2 = 1$ při vypuštění absolutního členu z regresního vztahu.
Steinova odhadová funkce neposkytuje nestranné odhady, avšak odhady pomocí ní získané mají zpravidla menší rozptyl a metoda je robustnější vůči OLS při nedodržení předpokladu (přesné) normality náhodných složek.

<u>6) využitím apriorní (mimomodelové) statistické informace</u>

Nejčastěji lze uplatnit různé ad hoc postupy podle povahy řešené úlohy, jako např.

(a) Položením omezení na některé parametry modelu
(b) uplatněním „smíšeného" (*mixed*) odhadu kombinací průřezových dat a údajů časových řad. Např. v analýze spotřebitelské poptávky se závisle proměnnou spotřeba daného statku jde o využití (neznámé u časových řad) **informace o příjmové a cenových pružnostech poptávky z průřezových dat. Informaci o příjmové pružnosti z výběrového šetření o příjmech a výdajích lze např. uplatnit k eliminaci vlivu změn příjmu na poptávku a v analýze založené na vzorku časových řad se omezíme toliko na odhad cenových pružností poptávky. Nutným předpokladem ovšem je, aby v obou dílčích úlohách byl koeficient příjmové pružnosti poptávky stejně chápán (u některých druhů zboží se mohou podstatně lišit příjmové pružnosti v krátko- a střednědobém časovém horizontu.**

---

[2] Hoerl, A.E., Kennard, R.W., Baldwin, K.F.: Ridge Regression. Some Simulation. *Commun.Statistics.* 4/1975 s. 105-123

[3] Stein, C. M.: Inadmissibility of the Usual estimator for the Mean of a Multivariate Normal Distribution. Proc. of the J.Bekeley Symp. on Mathematical Statistics and Probability. 1956 s.197-206.

Prostředky k diagnostikování multikolinearity

(A) F-test dílčích koeficientů determinace. Pro každou z vysvětlujících proměnných formulujeme regresi, v níž tato proměnná vystupuje jako závisle proměnná a je „vysvětlována" ostatními k-1 nezávisle proměnnými regresní rovnice. Pro pevně zvolenou j^* – tou proměnnou x_{j^*} pak testujeme (standardním F-testem) statistickou významnost skupiny ostatních k-1 proměnných obvyklým F-testem, založeným na koeficientu determinace:

$$F_J^* = \frac{\dfrac{R_{j*}^2}{k-2}}{\dfrac{1-R_{j*}^2}{(T-k+1)}} \qquad j^* = 2,3,\ldots,k$$

Je-li většina hodnot F_{j^*} statisticky významná, zamítá se hypotéza o ortogonalitě vysvětlujících proměnných, což je však podstatně přísnější podmínka než nepřítomnost významné kolinearity.

(B) Rovněž *Lawrence Klein* navrhuje jako test únosnosti multikolinearity test spočívající v posouzení velikosti R^2 (koeficient determinace závislosti vysvětlované proměnné y na k vysvětlujících proměnných $x_1, x_2, \ldots, x_k$) vůči koeficientům determinace $R_{j^*}^2$ (koeficient determinace závislosti pevně zvolené vysvětlující proměnné x_{j^*} na k-1 „vysvětlujících" proměnných $x_1, x_2, \ldots, x_k$ mimo x_{j^*}). Za vážný problém je však multikolinearita považována až tehdy, platí-li pro některé j^* nerovnost $R^2 < R_{j^*}^2$

tzn. pokud pro některou z vysvětlujících proměnných (vzatou jako závisle proměnná) má korelační koeficient $R_{j^*}^2$ vyšší hodnotu než standardní R^2.

(C) *Farrar a Glauber*[4] doporučují nepřímý **test multikolinearity založený na velikosti determinantu normované momentové matice $R = X^{*\prime} X^*$, kde prvky X^* získáme jako normované hodnoty původních prvků matice X, tzn.**

$$x_{tj}^* = \frac{x_{tj} - \bar{x}_{,j}}{s_{x_j}} \quad , \quad \textbf{kde } \bar{x}_{.j} = \frac{\sum_{t=1}^{T} x_{tj}}{T} \quad , \quad s_{x_j} = \left[\frac{\sum_{t=1}^{T} (x_{tj} - \bar{x}_{.j})^2}{T} \right]^{1/2}$$

Prvky této momentové matice mají při daném způsobu normování charakter korelační matice, načež lze pomocí χ^2- testu testovat, zda se determinant této matice významně liší od jedné.

Pro případ multikolinearity platí, že $|R^*| = 0$

---

[4] Farrar, D,E., Glauber R.R.: Multicollinearity in Regression Analysis: The Problem Revisited. *Review of Economics and Statistics* 49/1969 s. 92-107.

Pro případ ortogonality naopak platí $\left|R^*\right| = 1$

Hodnota mezi 0 a 1 (zjištěná v praxi téměř vždy) tedy svědčí o jistém stupni kolinearity ; ten však může být docela únosný. Testování nulové hypotézy tvaru $H_0: \ \left|R^*\right| = 1$ **se provádí** $\chi^2 -$**testem, avšak zamítnutí této nulové hypotézy neznamená, že by byla v daném případě multikolinearita fatálním problémem. Ani zde tedy test nesměřuje bezprostředně k ověřování multikolinearity, nýbrž k testování opačné situace, tj. stavu kdy jsou vysvětlující proměnné navzájem ortogonální** (tzn. kdy žádná neobsahuje ani kousek informace, kterou obsahují ostatní)**.**