Vícenásobná lineární regrese

Multikolinearita

Co to je multikolinearita?

Silné (avšak nikoliv funkční) závislosti mezi vysvětlujícími proměnnými.

Příčiny multikolinearity:

1. Přeurčený regresní model obsahující nadměrný počet vysvětlujících proměnných.
2. Nevhodný plán experimentu. (Nevhodná volba kombinací hodnot vysvětlujících proměnných.)
3. Nevhodné rozmístění experimentálních bodů.
4. Fyzikální omezení v modelu nebo v datech. (Věcně zdůvodněná závislost vzájemně propojených veličin.)

Důsledky multikolinearity:

1. Multikolinearita má za následek nadhodnocení součtů čtverců regresních koeficientů, což může vést k přikládání větší důležitosti některé vysvětlující proměnné.

2. Multikolinearita zvyšuje rozptyly odhadů, což má za následek:
   a) Snížení přesnosti odhadů individuálních hodnot (rozšíření predikčních intervalů)
   b) Nízké hodnoty t_i pro dílčí t-testy (tj. některé (někdy dokonce všechny) regresní koeficienty se jeví statisticky nevýznamné i v případě jinak velmi kvalitního modelu (paradox - významný F-test, nevýznamné všechny dílčí t-testy).
   c) Nestabilní odhady regresních koeficientů, z čehož plynou numerické obtíže.

3. Multikolinearita komplikuje rozumnou interpretaci individuálního vlivu jednotlivých vysvětlujících proměnných.

4. Nelze odděleně sledovat vliv jednotlivých vysvětlujících proměnných.

Kritéria pro identifikaci multikolinearity:

1. Determinant korelační matice. Při silné vzájemné lineární závislosti vysvětlujících proměnných se determinant korelační matice málo liší od nuly.
2. Nejmenší charakteristické číslo. Nízká hodnota nejmenšího charakteristického čísla indikuje silnou lin. závislost vysvětlujících proměnných.
3. Index podmíněnosti matice $\mathbf{X}^T\mathbf{X}$ nebo korelační matice (tj. odmocnina poměru největšího a nejmenšího charakteristického čísla). Přibližně hodnoty nad 30 ukazují na existenci multikolinearity.
4. Jednoduché korelační koeficienty dvojic vysvětlujících proměnných. Hodnoty blízké ±1 naznačují multikolinearitu. Doporučuje se používat pouze s dalšími kritérii multikolinearity.
5. Vícenásobné korelační koeficienty j-té vysvětlující proměnné vzhledem k ostatním vysvětlujícím proměnným. Hodnoty blízké ±1 indikují silnou multikolinearitu.
6. Kritérium M (založeno na paradoxu F-testu a dílčích t-testu):

$$M = \frac{\frac{F}{\frac{1}{K}\sum_{i=1}^{K} t_i^2} - 1}{\frac{F}{\frac{1}{K}\sum_{i=1}^{K} t_i^2} + 1},$$

kde t_i jsou testová kritéria pro dílčí t-testy a F je testové kritérium pro celkový F-test. Orientačně – je-li (M>0,8), lineární závislost se označuje za silnou.

Poznámka: Odlehlá pozorování mohou znesnadnit či zcela znemožnit identifikaci multikolinearity.

Možnosti odstranění multikolinearity:

1. V případě přeurčeného regresního modelu – identifikaci a vypuštěním zbytečných vysvětlujících proměnných.
2. Je-li příčinou multikolinearity nevhodný plán experimentu, je možné nedostatky napravit a pořídit kvalitnější data.
3. Nejkomplikovanější (a bohužel i nejčastější) případ multikolinearity je způsoben fyzikálními závislostmi v modelu. Vypuštění proměnných z modelu může vést k systematickým chybám a ani pořízení nových dat většinou nepomůže. Jako jediné rozumné východisko se ukazuje opuštění třídy lineárních nezkreslených odhadů (statistický výzkum v posledních 40-ti letech, Bayesovská statistika…)

Příklad. Soubor Felicia2.sf3 obsahuje informace o 20 ojetých autech značky Felicia Combi – proměnné Stáří, Najeto (tis. km) a Cena (tis. Kč).

1) Zkonstruujte regresní model závislosti ceny auta na jeho stáří a na počtu najetých kilometrů.
2) Posuďte jeho kvalitu a použijte jej k odhadu ceny auta starého 6 let a s najetými 60 tis. km.