Zadanie k cvičeniu o multikolinearite:

1. Otvorte si prednášku k multikolinearite a pozrite sa na to čo je to multikolinearita a aké sú jej dôsledky.

Je to teda závislosť medzi vysvetľujúcimi premennými resp. vzniká keď sú vysvetľujúce premenné navzájom lineárnou kombináciou. Jej dôsledkom je nepresný odhad, resp. veľké štandardné chyby parametrov, navyše odhady parametrov sú veľmi citlivé na zmeny v údajoch.

Keďže z povahy dát často vyplýva istá miera kolinearity vysvetľujúcich premenných, pri multikolinearite zisťujeme nie to či sa vyskytuje, ale či znehodnocuje model významnou mierou.

Skúste sa pozrieť na to čo sa stane v prípade perfektnej multikolienarity:

Napríklad keď nedodržíte postup pri tvorbe umelých premenných. V údajoch ktoré máte na hárku jedna si vytvorte tri umelé premenné. D1 D2 D3, D1 bude mať od začiatku po polku údajov hodnotu jedna, potom nula, D2 opačne, od začiatku do polky nula potom jedna, D3 je rovná jednej stále (resp. ako lokujúca konštanta). Teda D3 je kombináciou premenných D1 a D2.

Skúste odhadnúť model kde závislá premenná bude y a nezávislé premenné len D1 D2 a D3.

Čo sa stalo?

Model je znehodnotený perfektnou multikolinearitou, preto niektoré parametre nemohli byť vôbec odhadnuté kvôli nekonečne veľkým štandardným chybám.

Skúste teraz odhadnúť model tak ako ho máte, teda vysvetľovaná premenná y a vysvetľujúce premenné X1 až x5. Zinterpretujte výsledok. Podľa čoho môžete vidieť že sa v modeli vyskytuje multikolinearita? Zvyčajne sa to dá zistiť na základe vysokého R^2 a nevýznamnosti niektorej z premenných (ak by bol model veľmi dobrý tak ako napovedá koeficient determinácie, mali by byť všetky premenné významné). Toto teda naznačuje že model je znehodnotený multikolinearitou.

Metódy zisťovania multikolinearity:

1. Korelačná matica

Otvorte si údaje/data analysis vyberte položku correlation

V okne ktoré sa vám otvorí vyberte do údajov len vysvetľujúce premenné (s názvami alebo bez, ak s názvami nezabudnite zaškrtnúť labels) a vyberte si kam chcete výstup.

Výsledkom bude korelačná matica v ktorej sa nachádzajú korelačné koeficienty merajúce závislosť medzi každou dvojicou premenných. Na diagonále korelačnej matice sú jednotky keďže je to závislosť medzi tými istými prvkami matice. Čím je bližšia jednej, tým je závislosť silnejšia, čím bližšia nule tým slabšia, pri mínus jednotke je to silná negatívna závislosť. V exceli sú prvky len pod diagonálou, predpokladá sa že nad diagonálou budú tie isté hodnoty (závislosť medzi x1 a x2 je to isté ako medzi x2 a x1).

Ak sa v matici vyskytujú hodnoty korelácie vyššie ako 0,8 znamená to silné závislosti, teda v modeli bude multikolinearita. Na druhej strane, ak sa v matici nevyskytnú závislosti nemusí to nevyhnutne znamenať že sa v modeli významná multikolinearita nevyskytuje.

Vyskúšajte si postup Farrar Glauber

Tento postup v praxi nie je veľmi používaný lebo vám povie len či sa jedná alebo nejedná o významnú multikolinearitu, ale neoznačí premennú ktorá sa na nej významným spôsobom podieľa.

Najprv doplňte trojuholníkovú korelačnú maticu na štvorec tak že stĺpce ktoré máte v spodnej časti transponujete do riadkov hornej časti tak aby vznikol zrkadlový odraz matice cez diagonálu.

Vypočítajte R cez funkciu MDETERM(do funkcie označte celý rozsah doplnenej korelačnej matice- len hodnoty)

Dopočítajte testovaciu charakteristiku podľa vzorca v prednáške.

Dopočítajte kritickú hodnotu pomocou funkcie chiinv(0,05, ½\*k\*(k-1)), kde k je počet odhadovaných parametrov v regresnom modeli, teda

Nulová hypotéza v tomto prípade hovorí že model nie je znehodnotený významnou multikolinearitou, alternatívna hypotéza že model je znehodnotený významnou multikolinearitou. V tomto prípade bude testovacia charakteristika (asi okolo 130) významne vyššia ako kritická hodnota, takže výsledok je významnjé zenohodnotenie modelu multikolinearitou.

Ďalšou možnosťou ako zistiť multikolinearitu v modeli sú parciálne regresie, teda pri 5 vysvetľujúcich premenných spravíte 5 regresií kde sa jednotlivé vysvetľujúce premenné budú striedať v úlohe závisle premennej (y teraz neriešime, skúmame len vzťahy medzi vysvetľujúcimi premennými). Ak by v modeli nebola multikolinearita všetky regresie by mali byť nevýznamné bez závislosti. Takže odhadnete nasledujúce modely:

Závislosť x1 od x2x3x4x5

Závislosť x2 od x1x3x4x5

Závislosť x3 of x1x2x4x5

Závislosť x4 od x1x2x3x5

Závislosť x5 od x1x2x3x4

Nezabudnite si vždy premiestniť vysvetľujúce premenné v danej regresii tak aby boli pri sebe a dali sa označiť naraz v regresii do jedného bloku.

Keďže všetky regresie vyšli významné naznačuje to významnú multikolinearitu v modeli. Z týchto modelov budeme potrebovať koeficienty determinácie pre výpočet špeciálnych mier multikolinearity.

Vytvorte si tabuľku kde v prvom stĺpci budete mať označenie x1 až x5

Do druhého stĺpca si skopírujte koeficienty determinácie (r square) z parciálnych regresií. Ku každému iksu vždy ten kde bola daná premenná použitá ako vysvetľujúca, tá premenná kde je najvyšší by sa mala najvýznamnejším spôsobom podieľať na multikolinearite.

Pre všetky premenné môžeme dopočítať špeciálne miery multikolinearity

VIF inflačný faktor rozptylu =1/(1-rsquare) podľa jednoduchého vzorca s použitím parciálnych koeficientov determinácie pre každú premennú vypočítajte vif, hodnoty vyššie ako 10 indikujú významný problém s multikolinearitou, premenná s najvačšou hodnotou vif sa na nej podieľa najviac.

Podobne TOL

TOL – miery tolerancie -=1/VIF fungujú opačne hodnoty menšie ako 0,1 znamnajú významnú multikolinearitu, čím nižšie tým sa daná premenná podieľa významnejšou mierou na multikolinearite. Výsledkom je teda významná multikolienarita v modeli.

Naimportujte si údaje do gretlu. Odhadnite si v gretli rovnaký model (použite hárok údaje pre gretl). V odhadnutom modely hore na lište v menu analysis/ collinearity

Gretl automaticky vypočíta VIF (najčastejšie používané miery kolinearity, sú súčasťou v podstate každého softvéru), takže presne to isté čo sme robili v exceli aj s popisom

V spodnej tabuľke gretl zrátal lambdy a conf (confidence index alebo aj ci) Multikolinearitu značí hodnota ci vyššia ako 30, na základe prekryvu variability (hodnoty vpravo pod premennými) ak sa pri premenných vyskytujú hodnoty vyššie ako 0,5 tak sa premenné významnou mierou podieľajú na multikolinearite. Takže hľadať hodnoty vyššie ako 30 pri ci a v danom riadku hodnoty prekryvu variability premenných viac ako 0,5. Takto sa identifikujú premenné ktoré sa podieľajú na multikolinearite.

Ako vyriešiť problém s multikolinearitou?

Najjednoduchším riešením čo sa ponúka je vynechanie problematickej premennej čo sa však veľmi v praxi neodporúča. Skúste v gretli odhadnúť model bez premennej s najvyšším vif a pozrite si znovu analýzu kolinearity. Ako sa zmenila? Vynechaním premennej ale vypustíme z modelu dodatočnú informáciu ktorá by mohla byť potrebná. Ďalšou možnosťou je neriešiť multikolinearitu, čo závisí hlavne od účelu použitia modelu. Ak je účelom modelu napr. predikcia, je možné použiť aj model znehodnotený multikolinearitou (vysoké rsqare je výhodou).

Možnosťou by bola modifikácia premenných tak aby sa významná kolinearita znížila, prípadne vytvorenie komponentov z premenných čo by však bolo ťažko interpretovateľné. Najčastejšie sa odporúča zvýšiť rozsah údajov. Keďže multikolinearita vyplýva z výberovej povahy údajov a jej dôsledky sú rovnaké ako prípad nedostatočného rozsahu pozorovaní, odpoveď na otázku ako riešiť multikolinearitu je rovnaká ako na otázku čo s malým počtom pozorovaní, zväčšiť rozsah použitých údajov napríklad využitím panelových dát, teda kombináciou časových radov a prierezových údajov.