Statystyczna Analiza Danych – laboratorium

Weryfikacja założeń KMRL

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 8
28/29 kwietnia 2022

Idea zajęć – co i po co będziemy robić?

- Jeśli założenia Klasycznego Modelu Regresji Liniowej nie są spełnione, to estymator MNK może nie być nieobciążony, zgodny lub efektywny
- Uzyskane przez nas oszacowania i wnioskowanie statystyczne mogą nie być prawidłowe
- Wielu problemom możemy przeciwdziałać, czasem będzie konieczne zastosowanie bardziej zaawansowanych technik analizy danych
- Pokażemy dziś, jak oceniać, czy model spełnia założenia KMRL

Przypomnienie – założenia KMRL

- Liniowość: $y = X\beta + \varepsilon$
- X są nielosowe lub losowe
- $E(\varepsilon) = 0$
- Homoskedastyczność: $Var(\varepsilon) = \sigma^2 I$
- Brak autokorelacji: $\forall_{i \neq j} Cov(\varepsilon_i, \varepsilon_j) = 0$
- $\star$ Składnik losowy ma rozkład normalny (założenie dodatkowe)

Założenia, których spełnienia nie sprawdzamy

- Losowość lub nielosowość X wpływają jedynie na postać dowodów właściwości estymatora MNK
- W modelach ze stałą założenie o wartości oczekiwanej błędu losowego nie jest restrykcyjne
- Odchylenia od wartości oczekiwanej składnika losowego są wtedy przejmowane przez stałą

Konsekwencje niespełnienia założenia o liniowości

- Podważa interpretację oszacowanych współczynników
- Niemożliwe jest udowodnienie właściwości estymatora MNK, takich jak nieobciążoność czy efektywność

Jak naprawić

Przebudować model, aby uwzględniał nieliniowość relacji:

- Zmienne w modelu mogą wymagać transformacji: logarytmowania, potęgowania, etc (*transformacja Boxa-Coxa, opisana w scenariuszu!*)
- wprowadzenie interakcji (iloczynów) zmiennych x
- zastosowanie innej formy funkcyjnej: np. modelu schodkowego lub krzywej łamanej

Konsekwencje rozkładu składnika losowego odbiegającego od normalnego

- Założenie jest niezbędne do wyprowadzenia rozkładów statystyk testowych oraz prawidłowego wnioskowania statystycznego
- Można znaleźć estymator nieliniowy, który będzie mieć niższą wariancję niż estymator MNK

Jak naprawić

- Próba o dużej liczebności, rozkład reszt przypomina krzywą dzwonową – rozkłady statystyk bliskie standardowym (CTG)
- Próba o małej liczebności – sprawdzić obecność obserwacji odstających, popracować nad formą funkcyjną modelu, (w ostateczności) powiększyć próbę

Niesferyczność błędów losowych

- Jeśli założenia o homoskedastyczności lub braku autokorelacji są niespełnione, mówimy o niesferyczności składnika losowego
- Macierz wariancji-kowariancji dla składnika losowego ma wtedy postać dowolnej macierzy symetrycznej i dodatnio półokreślonej

Konsekwencje niesferyczności błędu losowego #1

- Estymator MNK nadal nieobciązony i zgodny, ale nieefektywny
- Estymator wariancji składnika losowego obciążony w próbach o małej liczebności, zgodny w próbach o dużej liczebności
- Standardowy estymator macierzy wariancji-kowariancji dla $\hat{\beta}$ obciążony, niezgodny

Konsekwencje niesferyczności błędu losowego #2

- Estymator macierzy wariancji-kowariancji dla $\hat{\beta}$ używany jest przy konstrukcji prawie wszystkich statystyk testowych, będą one nieprawidłowe
- Wnioskowanie statystyczne może być nieprawidłowe

Jak naprawić – heteroskedastyczność

- Sprawdzić, czy nie wynika z pominięcia istotnej zmiennej
- Zastosować estymację z wykorzystaniem odpornej macierzy wariancji kowariancji
- Stosowalna Uogólniona MNK

Jak naprawić – autokorelacja

- Autokorelacja jest poważnym problemem, jeżeli nie możemy zmienić kolejności obserwacji w próbie
- Próba przekrojowa – pojawia się w bardzo dziwnych przypadkach, wystarczy spermutować zbiór danych
- Szeregi czasowe – różnicowanie, zastosowanie właściwych technik analizy danych
- Panele – zastosowanie właściwych technik analizy danych