Statystyczna Analiza Danych – laboratorium

Regresja liniowa – zmienne kategoryczne, problemy z danymi

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 7
20/21 kwietnia 2023

Idea zajęć – co i po co będziemy robić?

- W tym omówieniu jest część treści, których nie ma w scenariuszu (ale są ciekawe i ważne)
- Praca ze zmiennymi kategorycznymi różni się od pracy ze zmiennymi ilościowymi, dowiemy się, jak dołączyć takie zmienne do modelu regresji
- Porozmawiamy o problemach: zmiennych pominiętych, współliniowości i równoczesności w odniesieniu do zadań z laboratorium

Przypomnienie – model liniowy

$$y = X\beta + \varepsilon$$

- y – zmienna objaśniana (endogeniczna, zależna) o rozkładzie (quasi)ciągłym
- X – macierz zmiennych objaśniających (egzogenicznych, niezależnych)
- ε – składnik losowy o rozkładzie normalnym
- β – wektor parametrów

Przypomnienie – oszacowanie

$$y = X\hat{\beta} + e$$

- y – zmienna objaśniana (endogeniczna, zależna) o rozkładzie (quasi)ciągłym, n obserwacji
- X – macierz k zmiennych objaśniających (egzogenicznych, niezależnych)
- e – wektor reszt (residua, oszacowanie składnika losowego)
- $\hat{\beta}$ – wektor oszacowań parametrów, uzyskany MNK: $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{y} = X\hat{\beta}$ – wartość dopasowana

Zadanie – zmienne jakościowe

- Stwórz model liniowy wyjaśniający wagę za pomocą wszystkich zmiennych poza zmienną Sport. Zinterpretuj wyniki.
- Czy z modelu wynika, że sportowcy różnych płci różnią się wagą?
- Czy taki model przewiduje, że zależność pomiędzy wzrostem a wagą może być różna dla różnych płci?

Dodawanie zmiennych jakościowych do regresji

- **Nie możemy oszacować regresji, w której są wszystkie poziomy zmiennej jakościowej i stała** – *dummy variable trap*
- Stała bywa "cenna" w modelu, dlatego z reguły usuwamy jeden z poziomów zmiennej jakościowej – poziom bazowy
- (Możliwe jest oszacowanie regresji, w której są wszystkie poziomy zmiennej jakościowej, bez stałej, wtedy nie usuwamy poziomu bazowego!)

Interpretacja oszacowań przy zmiennych jakościowych

- Interpretujemy w odniesieniu do poziomu bazowego *b*:
- Model na poziomach:
  - Obiekty poziomu *a* przeciętnie cechują się wielkością *y* o $\hat{\beta}$ większą/mniejszą niż obiekty poziomu *b ceteris paribus*
- Model ze zlogarytmowanym *y*, $\hat{\beta}$ niskie (!):
  - Obiekty poziomu *a* przeciętnie cechują się wielkością *y* o $100\% * \hat{\beta}$ większą/mniejszą niż obiekty poziomu *b ceteris paribus*
- Nie rozważamy modelu na logarytmach – dlaczego?

Współliniowość

- **Współliniowość dokładna** – występuje np. gdy włączymy do modelu wszystkie poziomy zmiennej jakościowej i stałą. Zmienne x tworzą wtedy kombinację liniową, macierz $X^T X$ przestaje być odwracalna
- Łatwa do usunięcia, z reguły wszystkie pakiety statystyczne ten problem kontrolują
- **Współliniowość niedokładna** – wystepuje, gdy pomiędzy zmiennymi objaśniającymi zachodzi silna korelacja
- Trudniejsza do zauważenia i (czasem) naprawienia

VIF

$$VIF_k = \frac{1}{1 - R_k^2}$$

- Współczynnik inflacji wariancji, pozwala sprawdzić, czy występuje problem współliniowości niedokładnej
- Badana zmienna x_k staje się zmienną objaśnianą w regresji na pozostałych zmiennych x.
- R_k^2 – R^2 z regresji zmiennej x_k na pozostałe zmienne objaśniające

```
regresja <- lm(y ~ x1 + x2, dane)
library(car)
vif(regresja) # VIF dla kazdego x
mean(vif(regresja)) # sredni VIF
```

VIF

$$VIF_k = \frac{1}{1 - R_k^2}$$

- Jeśli x_k jest wysoce skorelowany z innymi zmiennymi objaśniającymi, jego zmienność będzie przez ich zmienność dobrze tłumaczona i R_k^2 będzie wysokie
- $VIF > 10$ sugeruje występowanie silnej niedokładnej współliniowości
- Najczęściej usuwamy zmienną o najwyższej wartości VIF

Konsekwencje współliniowości

- Występowanie współliniowości może utrudnić wnioskowanie statystyczne
- W przypadku występowania silnej współliniowości między dwoma zmiennymi, wysoka wariancja oszacowań dla tych zmiennych obniża wartość statystyk t do poziomu, przy którym oszacowania przy zmiennych wydają się nieistotne
- Możemy nie być w stanie rozróżnić wpływu zmiennej na y – jeśli usuniemy ją z modelu, jej wpływ zostanie "przechwycony" przez silnie skorelowaną zmienną
- Jeśli usuniemy z modelu zmienną, która w istotny sposób tłumaczyła y, oszacowania przy zmiennych, które są z nią skorelowane będą obciążone

Skrót myślowy

- Mimo że istotność badamy w odniesieniu do współczynników modelu, popularnym skrótem myślowym jest mówienie o:
  - **zmiennej istotnej** – jeśli odrzucimy H_0 w teście t
  - **zmiennej nieistotnej** – jeśli nie odrzucimy H_0 w teście t

Zmienne pominięte

- Istnieją dwa przypadki, w których pominięcie zmiennej x_k nie wywoła obciążenia estymatora:
  1. $\beta_k = 0$ – tej zmiennej nie powinno być w modelu
  2. $x_i^T x_k = 0$ – zmienne są ortogonalne
- Jeśli w modelu pominiemy zmienną, która w istotny sposób tłumaczy y i jest skorelowana z pozostałymi zmiennymi, estymator MNK przestanie być estymatorem nieobciążonym

Konsekwencje **pominięcia** zmiennych **istotnych**

- Oszacowania przy zmiennych skorelowanych ze zmienną pominiętą będą obciążone (czasem można wysnuć wnioski na temat kierunku obciążenia przy zmiennej zawartej w modelu)
- Wnioskowanie statystyczne jest utrudnione (nieprawidłowe wartości statystyki testowej w teście t)
- Pominięcie zmiennej może "odbić się" w zachowaniu wariancji błędu losowego – może nie być już homoskedastyczna

Porównanie Z1 i Z2

- Gdy przyjrzymy się tabeli korelacji, okazuje się, że wzrost jest skorelowany ze zmiennymi X.Bfat i SSF
- Pominięcie tych zmiennych (w Z1) sprawiało, że współczynnik przy zmiennej wzrost wydawał się istotny statystycznie
- Po wprowadzeniu zmiennych do modelu, współczynnik przy zmiennej wzrost już nie był istotny statystycznie
- Zmienił się też znak oszacowania!

Konsekwencje **pozostawienia** zmiennych **nieistotnych**

- Estymator MNK nadal nieobciążony
- Estymator MNK z takiego modelu będzie mieć jednak wyższą wariancję, niż estymator z modelu bez tych zmiennych – nie jest efektywny
- Precyzja oszacowań przy pozostałych zmiennych jest niższa – wnioskowanie z testu t może być nieprawidłowe

Konsekwencje **usuwania** zmiennych **nieistotnych**

- W uzyskanym modelu jakość dopasowania będzie niższa, ale model będzie prostszy
- Usuwanie zmiennych nieistotnych poprawi dokładność oszacowań przy zmiennych istotnych
- Usuwanie tylko na podstawie testu t może być wadliwe – jeśli usuwamy więcej zmiennych, powinniśmy testować hipotezę o łącznej nieistotności

Równoczesność

- Jedną z właściwości hiperpłaszczyzny regresji jest, że reszty są ortogonalne do zmiennych objaśniających
- W modelu z równoczesnością **nie jest spełnione**, że:

$$Cov(x_i, \varepsilon_i) = E(x_i^T \varepsilon_i) = 0$$

- Równoczesność często pojawia się, gdy w modelu występują sprzężenia zwrotne – zmienna x zależy od y, które zależy od zmiennej x

Konsekwencje równoczesności

- Estymator MNK z reguły nie jest zgodny
- Oznacza to, że nawet na podstawie dużej próby będziemy otrzymywać błędne oszacowania parametrów
- Wnioskowanie statystyczne na podstawie testu t utrudnione lub niepoprawne

Oryginalne zadanie Z2

- W oryginalnym sformułowaniu Z2 byli Państwo poproszeni o wykorzystanie wszystkich zmiennych ilościowych w bazie
- Jedną z tych zmiennych było $BMI = \frac{waga}{wzrost^2}$
- Wprowadzenie jej jako jednej ze zmiennych objaśniających, w regresji, w której zmienną objaśniają była waga wprowadzało problem równoczesności (sprzężenie zwrotne)