Mikroekonometria 5

Mikołaj Czajkowski
Wiktor Budziński

Regresja kwantylowa

- W KMRL modelujemy warunkową średnią zmiennej objaśnianej:

$$E\left(y_i \mid \mathbf{X}_i\right) = \mu\left(\mathbf{X}_i\right)$$

- Pokazaliśmy, że można modelować równocześnie również jej drugi moment, to jest warunkową wariancję: $\operatorname{var}\left(y_i \mid \mathbf{Z}_i\right) = \sigma^2\left(\mathbf{Z}_i\right)$
  - W przypadku rozkładu normalnego, te dwa momenty w sposób jednoznaczny definiują rozkład zmiennej objaśnianej
  - W praktyce, zmienne ciągłe mogą mieć różne inne rozkłady
  - Czasem chcielibyśmy modelować inne charakterystyki rozkładu, aby lepiej zrozumieć co się dzieje w danych
  - Rozwiązanie – regresja kwantylowa

Regresja kwantylowa

- Kwantylem rzędu τ nazywamy taką wartość λ_τ, dla której zachodzi: $F(\lambda_\tau) = \tau$
  - Taka wartość, że wartości mniejsze lub równe od niej przyjmowane są z prawdopodobieństwem τ, a większe z prawdopodobieństwem $1-\tau$
- Kwantyl można policzyć korzystając ze wzoru: $\lambda_\tau = F^{-1}(\tau)$
- W regresji kwantylowej chcemy zobaczyć, jak różne kwantyle zmiennej objaśnianej zależą od wybranych zmiennych objaśniających
  - Definiuje się tzw. kwantyle warunkowe: $Q_\tau(\mathrm{y}_i \mid \mathbf{X}_i) = F_{y_i}^{-1}(\tau \mid \mathbf{X}_i)$
  - Zazwyczaj, analogicznie jak w regresji liniowej, zakłada się zależność liniową: $Q_\tau(\mathrm{y}_i \mid \mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta}_\tau$
  - Uwaga: w tej notacji parametry mogą być różne dla każdego kwantyla

Regresja kwantylowa – estymacja

- Najprostszym wariantem regresji kwantylowej jest tzw. regresja na medianie
  - Staramy się znaleźć postać funkcyjną warunkowej mediany
  - Parametry znajdujemy minimalizując następującą funkcję celu:

$$Q\left(\boldsymbol{\beta}_{0,5}\right) = 0,5\sum_{i=1}^{N}\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{0,5}\right|$$

  - Bardziej ogólnie, aby oszacować parametry dowolnego kwantyla minimalizujemy funkcję:

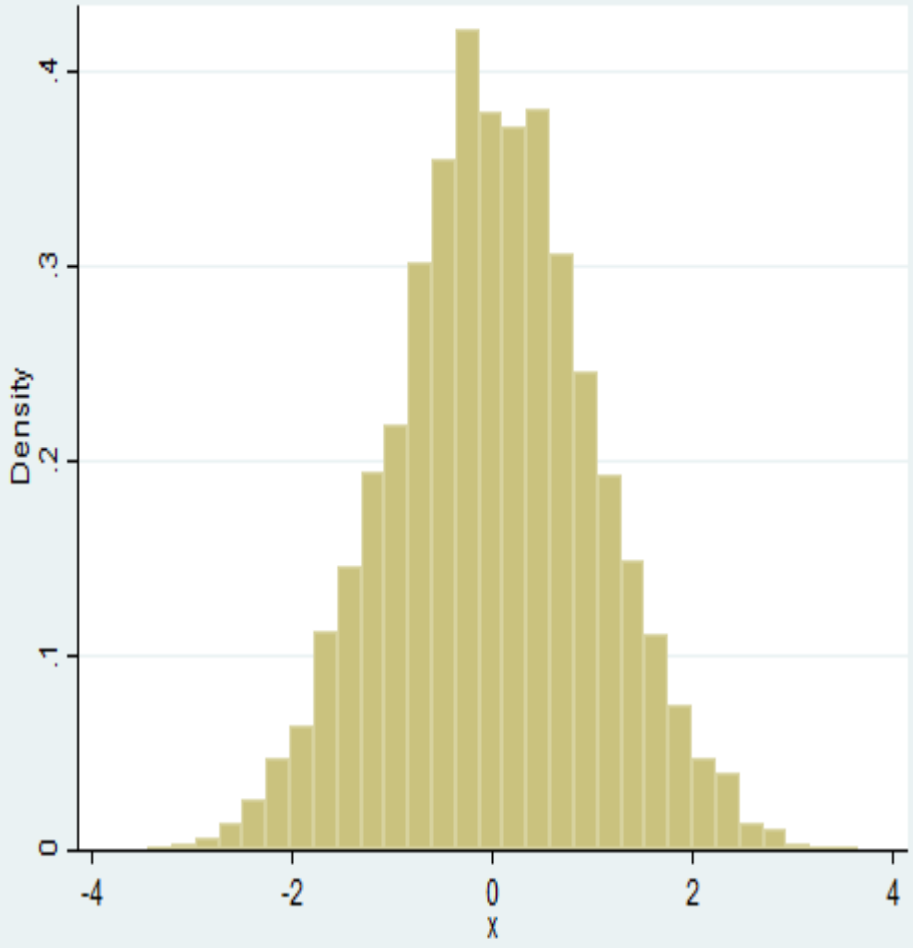
$$Q\left(\boldsymbol{\beta}_{\tau}\right) = \sum_{i:y_i \geq \mathbf{X}_i\boldsymbol{\beta}_{\tau}} \tau\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{\tau}\right| + \sum_{i:y_i \leq \mathbf{X}_i\boldsymbol{\beta}_{\tau}} \left(1-\tau\right)\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{\tau}\right|$$

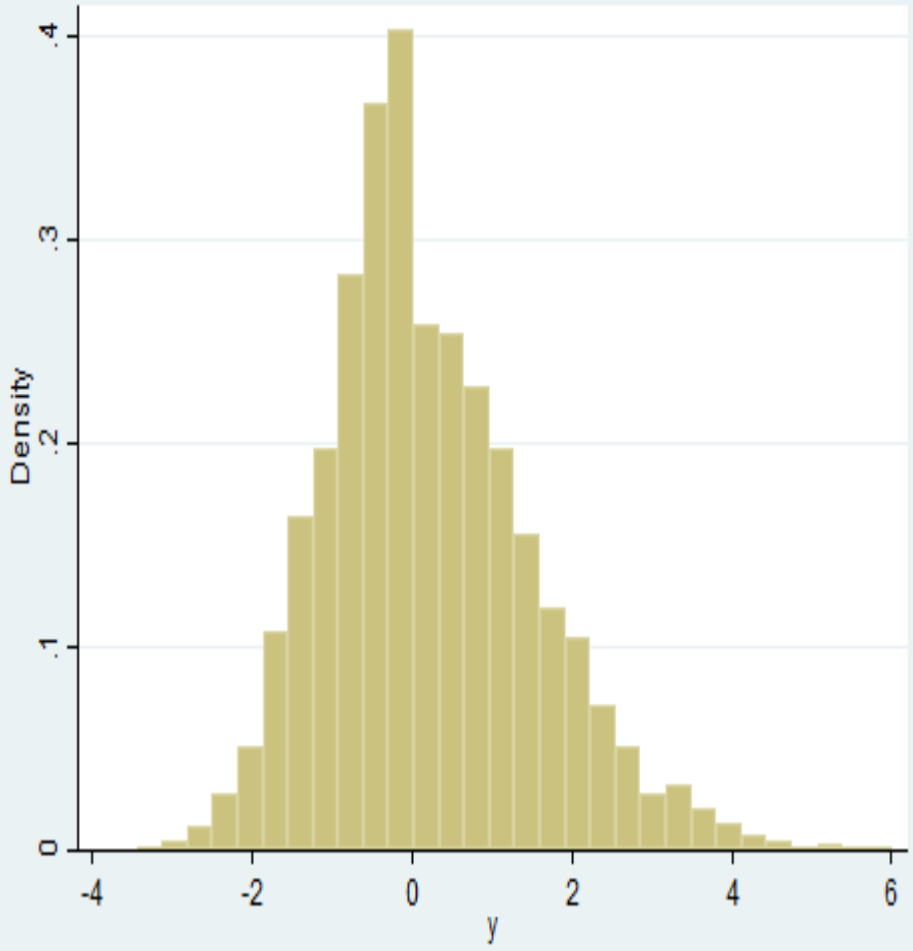
  - W literaturze model ten jest również nazywany estymatorem LAD (ang. *Least Absolute Deviations*)
  - Do optymalizacji używa się metod programowania liniowego (metoda simpleks)

Zadanie 1. Wykorzystując dane me.medexp3.dta przygotuj model regresji kwantylowej

1. Przygotuj model regresji kwantylowej w którym logarytm wydatków na leki jest objaśniany przez to, czy ktoś posiada dodatkowe ubezpieczenie od pracodawcy (*hi_empun*), liczbę chorób przewlekłych (*totchr*), wiek (*age*), płeć (*female*), kolor skóry (*blhisp*) oraz logarytm dochodu (*linc*)
   - Domyślnie Stata liczy regresję na medianie
   - Policz model regresji kwantylowej dla kwantyli innych rzędów np. 0,25 i 0,75

Przykładowa zmiana kształtu rozkładu





Zalety regresji kwantylowej

- Brak założeń na temat rozkładu zmiennej zależnej
- Można modelować cały rozkład zmiennej, a nie tylko wybrane momenty
- Łatwość wykrycia heteroskedastyczności
- Bardziej odporna na występowanie outlierów niż KMRL dla średniej
- Kwantyl zmiennej przekształconej przy pomocy ściśle rosnącej funkcji $g(x)$ jest równy przekształconemu kwantylowi oryginalnej zmiennej:

$$Q_\tau\left(g\left(y_i\right)|\mathbf{X}_i\right)=g\left(Q_\tau\left(\mathrm{y}_i|\mathbf{X}_i\right)\right)$$

Wady regresji kwantylowej

- Wzory analityczne na macierz kowariancji estymatora regresji kwantylowej są skomplikowane, a przez to rzadko stosowane
  - W praktyce stosuje się metody symulacyjne np. bootstrap, przez co model liczy się dłużej niż np. regresja liniowa
  - Współcześnie nie jest to duży problem

Regresja kwantylowa a heteroskedastyczność

- Regresja kwantylowa może zostać wykorzystana do analizy heteroskedastyczności
  - Parametry różniące się między kwantylami wskazują na istotną heteroskedastyczność

2. Policz regresję kwantylową dla kwantyli rzędu 0.1, 0.25, 0.5, 0.75, 0.9 wykorzystując polecenie `sqreg`
3. Przetestuj czy w analizowanych danych występuje heteroskedastyczność
4. Przeanalizuj graficznie, czy parametry zmieniają się między kwantylami

Metody symulacyjne – Monte Carlo

- Metoda Monte-Carlo
  - Wykorzystanie mocy obliczeniowej komputerów, aby poznać charakterystyki zmiennych losowych poprzez wielokrotne próbkowanie (zamiast rozwiązań analitycznych)
  - Jak w kasynie – wielokrotnie gramy i obserwujemy nasze wyniki, żeby ocenić np. jaka jest wartość oczekiwana jakiegoś zagrania
  - Np. w celu szacowania wartości skomplikowanych całek

Metody symulacyjne – symulowane dane

- Wykorzystanie symulowanych danych
  - Zakładamy istnienie jakiegoś procesu generującego dane (DGP, *data generating process*)
  - Takie dane pozwalają testować m.in.:
    - Czy model ekonometryczny działa prawidłowo?
      - Może programista się pomylił?
    - Konsekwencje odstępstw od przyjętych założeń (np. dla KMRL)
    - Zachowanie modelu dla małych prób
  - Zaletą jest to, że dokładnie wiemy, jaki jest DGP
  - Odpowiedź na powyższe pytania nie zawsze jest możliwa przy pomocy narzędzi analitycznych
  - Proces może być zaimplementowany jako symulacja Monte Carlo, aby obserwować wariancję oszacowań

Testowanie działania modeli ekonometrycznych

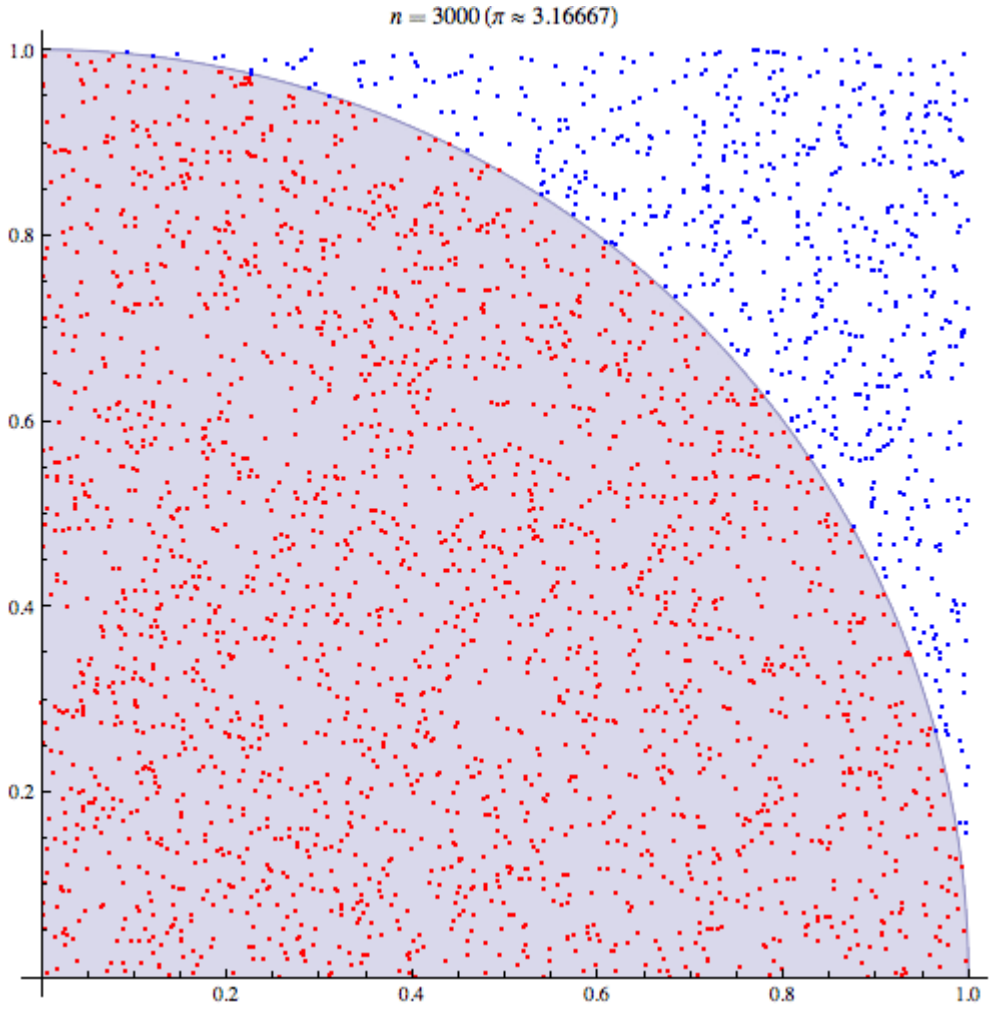
- Algorytm symulacji Monte Carlo może wyglądać w następujący sposób

1. Określ DGP (specyfikację, wartości parametrów)
2. Wygeneruj wartości zmiennych objaśniających i błędy losowe
   - Wykorzystaj odpowiednie rozkłady
3. Wygeneruj wartości zmiennej objaśnianej
4. Dokonaj estymacji modeli
5. Zapisz wyniki
6. Powtórz punkty 2-5 dużą liczbę razy
7. Przeprowadź analizę zapisanych wyników
   - Średnia jako wartość oczekiwana
   - Odchylenie standardowe jako błąd standardowy

Zadanie 2. Testowanie działania modeli ekonometrycznych

1. Sprawdź czy regresja liniowa jest w Stacie prawidłowo zaprogramowana
2. Sprawdź rozkład statystyki testu RESET oraz p-value tego testu, kiedy hipoteza 0 jest lub nie jest spełniona
3. Sprawdź czy endogeniczność obciąża oszacowania regresji liniowej
4. Sprawdź czy model 2MNK rozwiązuje problem endogeniczności

Całkowanie przy pomocy metody Monte Carlo



Całkowanie przy pomocy metody Monte Carlo

- Załóżmy, że chcemy policzyć całkę

$$\int_b^a f(y)\,dy$$

- Jeżeli potrafimy zapisać $f(y) = h(y)g(y)$, gdzie $g(y)$ to gęstość znanego ciągłego rozkładu, to możemy ją przybliżyć jako:

$$\int_b^a f(y)\,dy \approx \frac{1}{R}\sum_{r=1}^{R} h(y^r)$$

- Gdzie y^r jest wylosowane z rozkładu o gęstości $g(y)$

Zadanie 3. Całkowanie przy pomocy metody Monte Carlo

- Oszacuj używając metody Monte Carlo:

1. $$\int_0^1 \frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right) dy$$

2. $$\int_{-\infty}^{\infty} \exp\left(-\exp(y)\right) \frac{1}{\sqrt{2\pi}} \exp\left(-y^2/2\right) dy$$

Bootstrap

- Bootstrap to metoda symulacyjna, wykorzystująca podpróby posiadanej próby, zamiast wielu prób z populacji
- Może być wykorzystywana np. do wnioskowania statystycznego – kiedy nie znamy teoretycznych właściwości danej statystyki/estymatora
  - Jeśli mielibyśmy 1000 prób z pewnej populacji, to dla każdej z nich moglibyśmy policzyć daną statystykę i w efekcie otrzymać jej 1000 oszacowań
  - Średnią z tych 1000 oszacowań moglibyśmy traktować jako jej najlepsze oszacowanie, a wariancję jako miarę niepewności
  - Wykorzystując te charakterystyki moglibyśmy przeprowadzić wnioskowanie statystyczne

Bootstrap

- Z próby wielkości *N* losujemy ze zwracaniem *B* „sztucznych” prób również o długości *N*
  - Przykładowo, mamy próbę 5 obserwacji: X_1, X_2, X_3, X_4, X_5
  - „Sztuczne” próby wyglądałyby np. tak:

    $$X_1, X_1, X_3, X_4, X_4$$
  - Albo

    $$X_1, X_2, X_3, X_3, X_3$$
- Dla każdej z wylosowanych prób „sztucznych” obliczamy wartość interesującej nas statystyki/estymatora
- Wnioskowanie statystyczne przeprowadzamy analizując rozkład *B* oszacowań tej statystyki/estymatora

Zadanie 4. Bootstrap i Jackknife

- Jackknife
  - Podobna (wcześniejsza) metoda, polegająca na doborze obserwacji opierającym się na pominięciu (jednej lub więcej) obserwacji z próby
  - Np. przeprowadzamy regresję *n* razy, za każdym razem pomijając kolejną z *n* obserwacji

1. Wygeneruj sztuczne dane dla regresji z heteroskedastycznością
2. Porównaj błędy standardowe z KMRL, macierzy White'a, bootstrapu oraz jackknife

Praca domowa ME.5

Wyjątkowo: termin oddania do przyszłej środy (31.03)

1. Przeanalizuj, wykorzystując metodę Monte Carlo, działanie testu Breusha-Pagana (na heteroskedastyczność), kiedy model ma niepoprawną formę funkcyjną
   1. W pierwszym kroku przeprowadź symulację, w której model MNK ma poprawną formę funkcyjną oraz nie ma heteroskedastyczności
      1. Jaki rozkład ma p-value testu B-P? W jakim procencie wygenerowanych zbiorów danych test dał „zły wynik” (np. p-value mniejsze niż 5%)?
   2. W drugim kroku przeprowadź symulację, w której model MNK ma **niepoprawną** formę funkcyjną oraz nie ma heteroskedastyczności.
      1. Porównaj różne formy złej formy funkcyjnej np. pominięty kwadrat zmiennej, albo modelowanie y, kiedy powinniśmy modelować log(y)
      2. Jaki rozkład ma p-value testu B-P? W jakim procencie wygenerowanych zbiorów danych test dał „zły wynik” (np. p-value mniejsze niż 5%)?
      3. Porównaj wyniki z wynikami z pierwszego kroku
   3. Co na podstawie przeprowadzonych symulacji można powiedzieć o działaniu testu B-P kiedy założenie o liniowej formie funkcyjnej jest niespełnione?

- Do przygotowania w grupach trzyosobowych

set seed 10+”Nr indeksu”