



Grażyna Trzpiot

Uniwersytet Ekonomiczny w Katowicach
Wydział Informatyki i Komunikacji
Katedra Demografii i Statystyki Ekonomicznej
grazyna.trzpiot@ue.katowice.pl

GEOMETRYCZNE WŁASNOŚCI REGRESJI KWANTYLOWEJ

Streszczenie: Zastosowania miar porządkowych, w tym kwantyli, znajdujemy w różnych obszarach zastosowań, w szczególności w statystyce odpornej. Odejście od klasycznego podejścia bazującego na momentach zmiennej losowej wynika zazwyczaj z analizowanego zbioru danych, który nie spełnia założeń modeli. Dwa pierwsze momenty zmiennej losowej, na których budujemy dalej modele regresji, nie są adekwatne w opisie zbiorów danych z obserwacjami odstającymi. Zbiory danych z asymetrycznymi rozkładami również nie powinny być analizowane z wykorzystaniem modeli regresji estymowanych MNK. Celem artykułu jest prezentacja własności geometrycznych regresji kwantylowej. To podejście metodologiczne wykorzystuje wartości rozkładu, wyznaczając zbiór kwantyli oraz zbiór modeli regresji.

Słowa kluczowe: kwantyle, regresja kwantylowa, własności modeli regresji kwantylowych.

JEL Classification: C14, C20, C21.

Wprowadzenie

Regresja kwantylowa pozwala na estymację wpływu zmian wartości rozkładu zmiennej objaśniającej na rozkład zmiennej objaśnianej. Podstawowe zasady przyświecające temu podejściu znajdujemy w pracy autorów Koenker i Bassett [1978] oraz podjęte raz jeszcze w pracy Koenker [2005]. Wykorzystanie tego podejścia jest coraz powszechniejsze, zwłaszcza w analizach zbiorów o zależnościach nieliniowych czy o obserwacjach odstających, dla których podejmuje się oceny ryzyka wystąpienia zdarzeń.

W ilościowym zarządzaniu ryzykiem miary ryzyka służą do określania preferencyjnego porządku wśród pozycji finansowych z losowym wynikiem. Pozycja finansowa jest postrzegana jako zmienna losowa, która odwzorowuje każdy stan natury x w rzeczywistą wartość. Wartość zmiennej losowej odpowiada wynagrodzeniu (lub jest to stopa zwrotu) zapewnianemu przez instytucję finansową, gdy wystąpi stan x . Miary ryzyka mają na celu uwzględnienie kompromisu między wartościami, jakie może zajmować pozycja finansowa, a ryzykiem lub zmiennością tych wartości. Wybór miary ryzyka określa profil ryzyka inwestycyjnego. Miara ryzyka, jaką jest wartość zagrożona (VaR), kwantyl rozkładu prawdopodobieństwa dla stopy zwrotu z inwestycji, została wykorzystana w instytucjach finansowych [RiskMetrics, 1995]. Miara ta została skrytykowana za brak wykrywania niekorzystnych wartości stóp zwrotu w ogonie rozkładu prawdopodobieństwa [Donnelly, Embrechts, 2010; Trzpiot, 2007a; 2007b; 2009a]. Wprowadzono klasy ryzyka, które spełniają pewne pożądane właściwości. Klasa wypukłych miar ryzyka [Föllmer, Schied, 2002] obejmuje monotonne i wypukłe odwzorowania, które posiadają własność niezmienniczości translacji. Koherentne miary ryzyka [Artzner i in., 1999] stanowią podklasę wypukłych miar ryzyka, są to miary dodatnio homogeniczne. Mogą być wyrażone jako najgorszy oczekiwany wynik portfela, gdy miara prawdopodobieństwa stóp zwrotów z aktywów zmienia się w zbiorze niepewnych stóp zwrotu [Artzner i in., 1999; Trzpiot, 2016]. Warunkowa wartość zagrożona ($CVaR$), czyli oczekiwana wartość portfela na ogonie rozkładu stopy zwrotu portfela, która leży poza ustalonym kwantylem rozkładu, jest koherentną miarą ryzyka. Optymalny portfel wykorzystaniem $CVaR$ można wyznaczyć, wykorzystując różne estymatory [Rockafellar, Uryasev, 2000; Trzpiot, 2007a; 2007b; 2008; 2009a; 2009b]. W artykule podejmujemy problem geometrycznych własności regresji kwantylowej celem wskazania zalet metodologii, która jest wykorzystywana w wielu analizach, a w szczególności w opisie ryzyka, w tym finansowego.

1. Funkcja kwantyli rozkładu zmiennej losowej

Rozważamy przestrzeń probabilistyczną (Ω, F, P) oraz klasę zmiennych losowych $L^1(\Omega, F, P)$. Dystrybuenta zmiennej losowej $Y \in L^1(\Omega, F, P)$ będzie zapisana jako $F_Y(y) = P(Y \leq y)$.

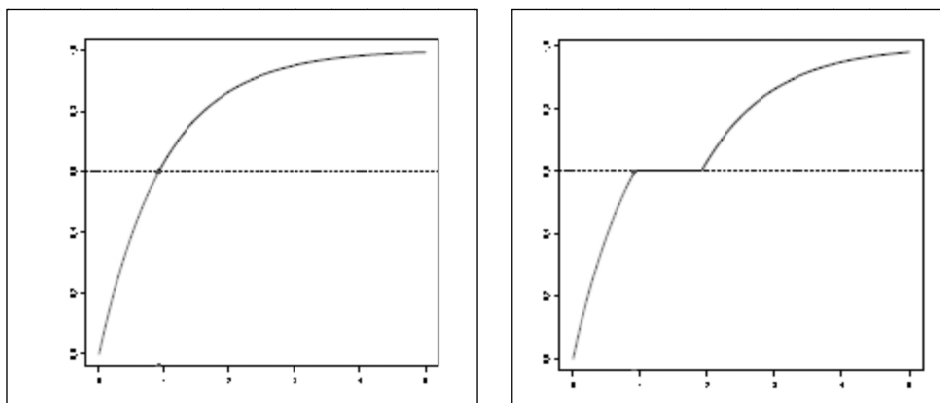
Funkcja kwantyli zdefiniowana jest następująco:

$$Q_Y(u) = \inf \{ y \in \mathbb{R} : F_Y(y) \geq u \},$$

co można zapisać równoważnie $Q_X(u) = F_X^{-1}(u)$. Możemy rozważać:

$$Q_Y(u) = \sup \{y \in \mathbb{R}: F_Y(y) < u\}.$$

Zatem dla dowolnej rosnącej transformacji t , $Q_{t(Y)}(\tau) = t(Q_Y(\tau))$.



Rys. 1. Przykładowe wykresy funkcji dystrybucyjnej i funkcji kwantyli

Źródło: Opracowanie własne.

Poprzez definicję przytoczoną powyżej otrzymujemy zapis dla rozkładów warunkowych $F(y|\mathbf{x}) = P[Y \leq y | \mathbf{X} = \mathbf{x}]$ i równoważnie $Q_{Y|\mathbf{x}}(u) = F^{-1}(u|\mathbf{x})$.

2. Estymacja kwantyli rozkładu zmiennej losowej

W badaniach niezwykle istotny jest problem estymacji parametrów populacji, w tym kwantyli. Rozważmy próbę prostą $\{y_1, \dots, y_n\}$ pochodzącą z populacji z dystrybucją F .

Zapiszemy $Q_\tau = \operatorname{argmin}\{E[R_\tau^q(Y - q)]\}$, gdzie Y ma dystrybucję F oraz $\hat{Q}_\tau = \operatorname{argmin}\{\sum_{i=1}^n R_\tau^q(y_i - q)\}$, dla funkcji ryzyka:

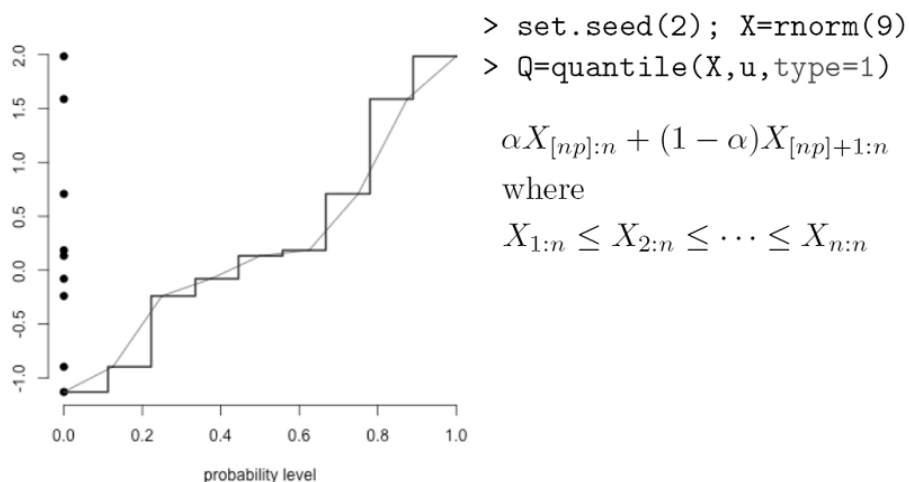
$$R_\tau^q(u) = u \cdot (\tau - \mathbf{1}(u < 0)), \quad \tau \in [0, 1].$$

Wówczas dla $n \rightarrow \infty$:

$$\sqrt{n}(\hat{Q}_\tau - Q_\tau) \rightarrow N(0, \tau(1-\tau)/f^2(Q_\tau),$$

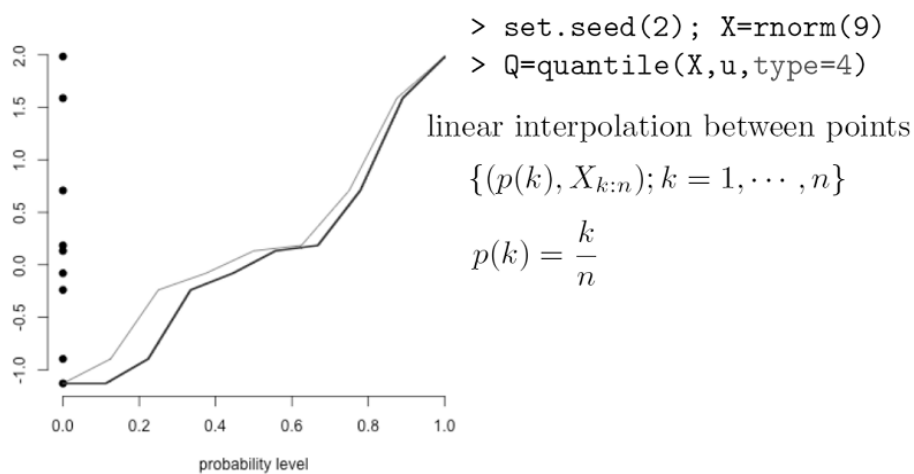
gdzie f jest funkcja gęstości zmiennej Y .

Estymacje empirycznych kwantyli na podstawie próby, które możemy wykonać z wykorzystaniem Minitab i SPSS, zostały opisane w pracy Hyndman i Fan [1996]. Na kolejnych rysunkach przedstawiono metody estymacji i wyniki estymacji różnych opisanych w tej pracy algorytmów wraz z kodami w R (rys. 2-4).



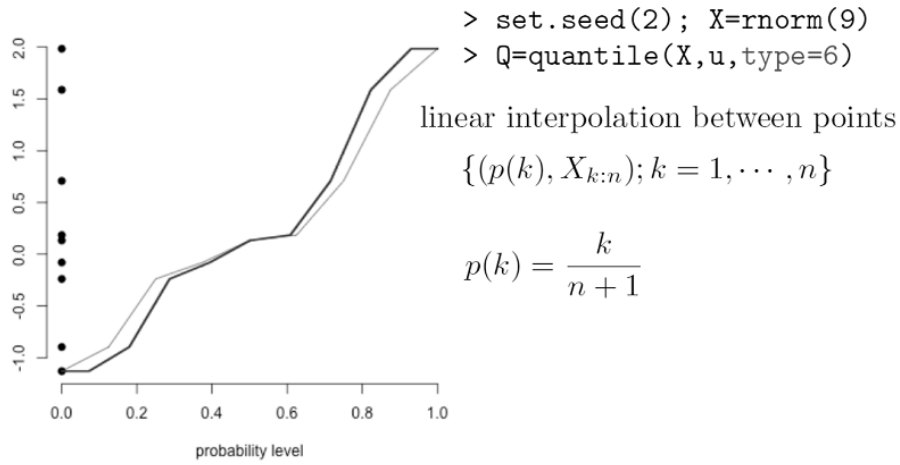
Rys. 2. Odwrotność empirycznej dystrybuanty

Źródło: Hyndman, Fan [1996].



Rys. 3. Liniowa interpolacja empirycznej dystrybuanty

Źródło: Hyndman, Fan [1996].



Rys. 4. Liniowa interpolacja empirycznej dystrybuanty

Źródło: Hyndman, Fan [1996].

3. Modele regresji i metody estymacji

Modele regresji mają na celu wskazanie wpływu zbioru zmiennych objaśniających na wybraną zmienną objaśnianą. Klasyczny model regresji liniowej jest wartością oczekiwaną rozkładu warunkowego zmiennej objaśnianej. Model regresji kwantylowej jest rozkładem warunkowym kwantyla odpowiedniego rzędu.

Rozważmy model liniowy $y = \beta_0 + \mathbf{x}^T \beta + \varepsilon$, prawie wszędzie, wówczas:

$$E[Y|X=\mathbf{x}] = \mathbf{x}^T \beta.$$

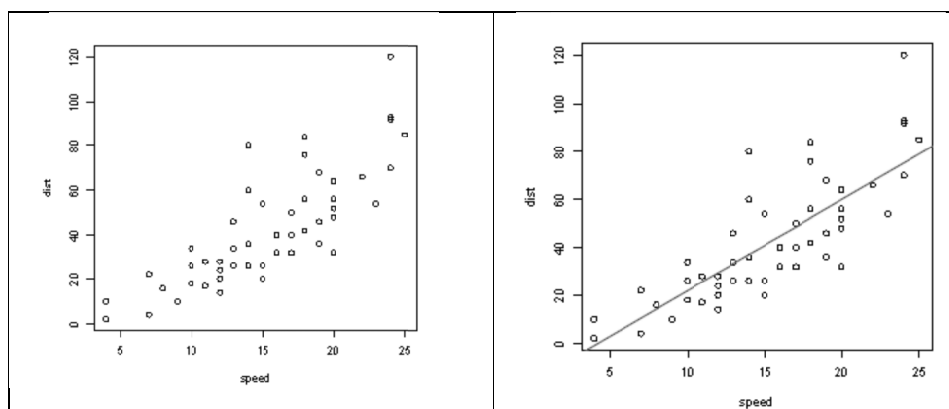
Następnie rozważmy funkcję regresji względem rozkładu warunkowego dowolnego kwantyla rzędu τ .

$$Q(\tau|X=\mathbf{x}) = \beta_0 + Q_\varepsilon(\tau) + \mathbf{x}^T \beta.$$

Estymacje przeprowadzamy odpowiednio metodą najmniejszych kwadratów (rys. 5 i 6) oraz metodą najmniejszych bezwzględnych odchyleń¹. Dla modelu liniowego: $y_i = \beta_0 + \mathbf{x}_i^T \beta + \varepsilon_i$ wyznaczamy rozwiązania:

$$\begin{aligned} \left(\hat{\beta}_0^{MNK}, \hat{\beta}^{MNK} \right) &= \operatorname{argmin} \left(\sum_{i=1}^n (y_i - \beta_0 - \mathbf{x}_i^T \beta)^2 \right), \\ \left(\hat{\beta}_0^{LAD}, \hat{\beta}^{LAD} \right) &= \operatorname{argmin} \left(\sum_{i=1}^n |y_i - \beta_0 - \mathbf{x}_i^T \beta| \right). \end{aligned}$$

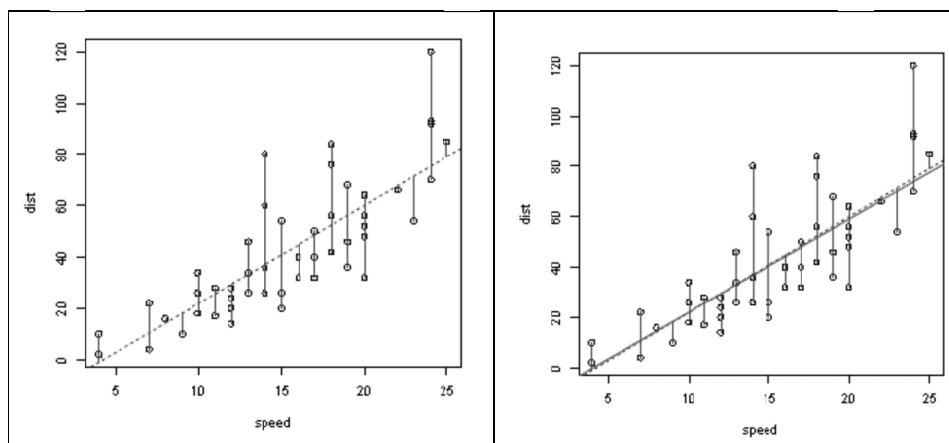
¹ LAD – *Least Absolute Deviation* – metoda najmniejszych bezwzględnych odchyleń.



Rys. 5. Zbiór obserwacji i liniowa funkcja regresji

Źródło: Opracowanie własne.

Rozważając zadanie takie, że rozkład $\varepsilon|X$ jest symetryczny oraz $E[\varepsilon|X] = M_\varepsilon$ $[\varepsilon|X] = 0$, wówczas $(\hat{\beta}_0^{MNK}, \hat{\beta}^{MNK})$ oraz $(\hat{\beta}_0^{LAD}, \hat{\beta}^{LAD})$ są zgodnymi estymatorami parametrów (β_0, β) . Rozważając przypadek taki, że rozkład $\varepsilon|X$ nie jest symetryczny, ale $E[\varepsilon|X] = 0$, wówczas $\hat{\beta}^{MNK}$ oraz $\hat{\beta}^{LAD}$ są zgodnymi estymatorami β . Jeżeli $M_\varepsilon[\varepsilon|X] = \gamma$, wówczas $\hat{\beta}_0^{LAD}$ jest zbieżne do $\beta_0 + \gamma$.



Rys. 6. Reszty regresji oraz liniowa funkcja regresji

Źródło: Opracowanie własne.

Estymator wyznaczony metodą najmniejszych kwadratów, zapisany jako $\hat{\beta}^{MNK}$, jest rozwiązaniem równania równoważnego:

$$\hat{\beta}^{MNK} = \operatorname{argmin} \{E([Y|X=x] - x^T \beta)^2\}.$$

W swojej pracy Angrist, Chernozhukov i Fernandez-Val [2006] udowodnili, że:

$$\hat{\beta}_\tau = \operatorname{argmin} \{E[\omega_\tau(\beta)(Q_\tau[Y|X=x] - x^T \beta)^2]\},$$

gdzie:

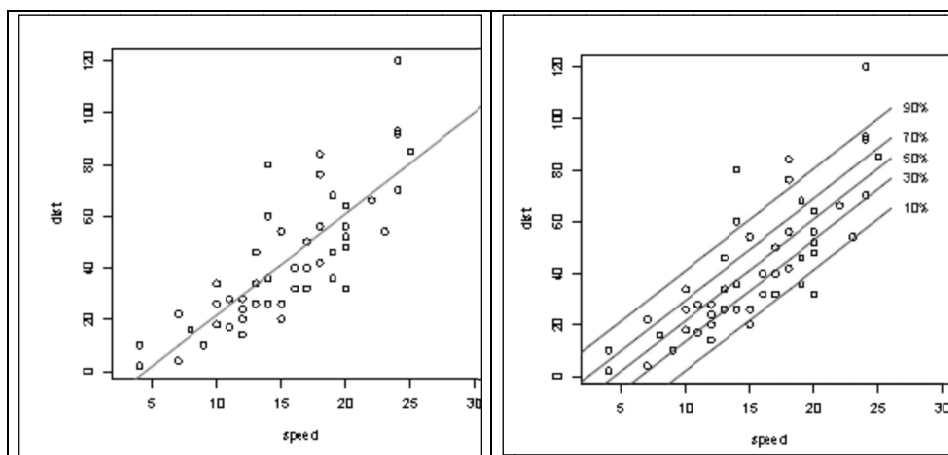
$$\omega_\tau(\beta) = \int_0^1 (1-u)f_{Y|X}(ux^T \beta + (1-u)Q_\tau[Y|X=x])du.$$

Tak zaproponowana metoda pozwala na wyznaczenie $\hat{\beta}_\tau$, które jest najlepszą ważoną średniokwadratową aproksymacją regresji kwantylowej, w której wagi zależą od średniej warunkowej funkcji gęstości Y względem $x^T \beta$.

Aby estymatory były zgodne, muszą być spełnione pewne warunki:

- obserwacje (Y_i, X_i) muszą być (warunkowo) niezależne oraz o tym samym rozkładzie,
- regresory (zmienne objaśniające) muszą mieć ograniczony drugi moment składnik losowy, ε jest zmienną losową ciągłą X_i , z medianą wynoszącą 0 (scentrowana),
- własność „lokalnej identyfikacji” $[f_\varepsilon(0)XX^T]$ jest dodatnio określona.

Zatem regresja kwantylowa może być postrzegana jako iteracyjna metoda najmniejszych kwadratów. Wybieramy początkowe $\beta^{(0)}$, np. β^{MNK} . W kroku k przyjmujemy $\varepsilon_i^{(k)} = y_i - x_i^T \beta^{(k-1)}$, a następnie definiujemy wagi $\omega_\tau^{(k)} = R_\tau(\varepsilon_i^{(k)})$ i wyznaczmy ważoną MNK, aby estymować $\beta^{(k)}$.



Rys. 7. Modele warunkowej średniej oraz modele warunkowych kwantyli

Źródło: Opracowanie własne.

W klasycznym zadaniu regresji, w szczególności dla zmiennych losowych ciągłych, rozwiązujemy zadanie $E[Y | X = x] = \int_R y dF_{Y|X=x}(y)$. Wyznaczając regresje kwantylowe, próbujemy rozwiązać zadanie (rys. 7):

$$Q_{Y|X=x}(u) = \inf\{y : F_{Y|X=x}(y) \geq u\}.$$

Jak zaproponowano w pracach Newey, Powell [1987] oraz Li, Racine [2007]:

$$\hat{Q}_{Y|X=x}(u) = \inf\{y : \hat{F}_{Y|X=x}(y) \geq u\},$$

gdzie $\hat{F}_{Y|X=x}(y)$ może być estymatorem jądrowym dystrybucyjnego rozkładu warunkowego.

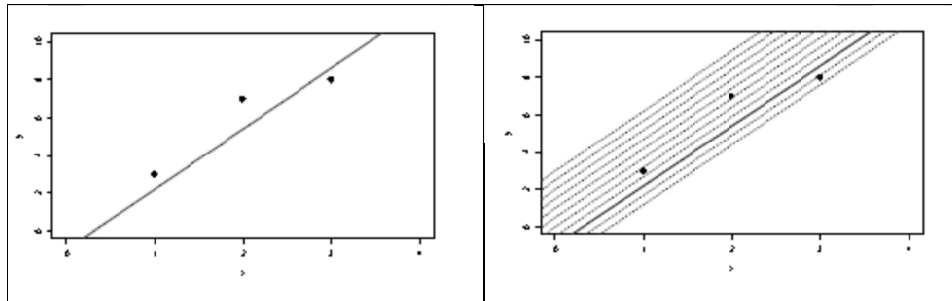
4. Wybrane własności regresji kwantylowej

Możemy zaobserwować, że regresja medianowa zawsze ma dwie obserwacje, na których budujemy model. Zaczynamy od linii regresji:

$$y_i = \beta_0 + \beta_1 x_i.$$

Wprowadzamy transformację: $y_i = (\beta_0 \pm \varepsilon) + \beta_1 x_i$, a następnie wyznaczamy:

$$\text{minimum} \sum_{i=1}^n |y_i - (\beta_0 + \beta_1 x_i)|.$$

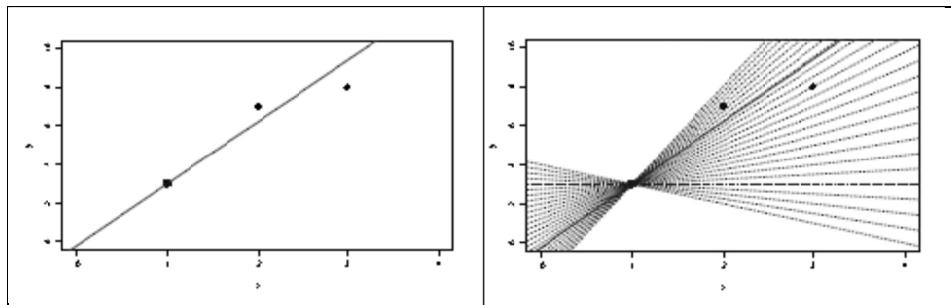


Rys. 8. Własności geometryczne regresji medianowej – przesunięcie

Źródło: Opracowanie własne.

Przesunięcie w górę od wyznaczonej linii obniża wartość sumy o ε , aż minniemy punkt po lewej, a następne przesunięcie będzie powiększać sumę (rys. 8). Musimy minąć pewien punkt (suma jest liniowa w ε). Rozważmy teraz obrót linii względem punktu wspierającego (rys. 9). Jeżeli obrót będzie w górę (zgodnie z kierunkiem osi OY), wówczas powiększa się wartość sumy (duży wpływ

punktu po prawej) Jeżeli obrót będzie w dół (kierunek przeciwny do skierowania osi OY), wówczas pomniejszymy wartość sumy (aż osiągniemy punkt po prawej).



Rys. 9. Własności geometryczne regresji medianowej – rotacja

Źródło: Opracowanie własne

Jeżeli obrót będzie w dół (kierunek przeciwny do skierowania osi OY), wówczas pomniejszymy wartość sumy (aż osiągniemy punkt po prawej). Zatem regresja medianowa posiada zawsze dwa punkty wspierające².

Rozważymy wybrane transformacje regresji kwantylowej:

- 1) Równoważnik skali: Dla dowolnego $\alpha > 0$ oraz $\tau \in [0, 1]$:

$$\hat{\beta}_{\tau}(\alpha Y, X) = \alpha \hat{\beta}_{\tau}(Y, X) \text{ and } \hat{\beta}_{\tau}(-\alpha Y, X) = -\alpha \hat{\beta}_{1-\tau}(Y, X);$$

- 2) Równoważność po reparametryzacji:

Macierz A ma wymiary $p \times p$ oraz $\tau \in [0, 1]$:

$$\hat{\beta}_{\tau}(Y, XA) = A^{-1} \hat{\beta}_{\tau}(Y, X).$$

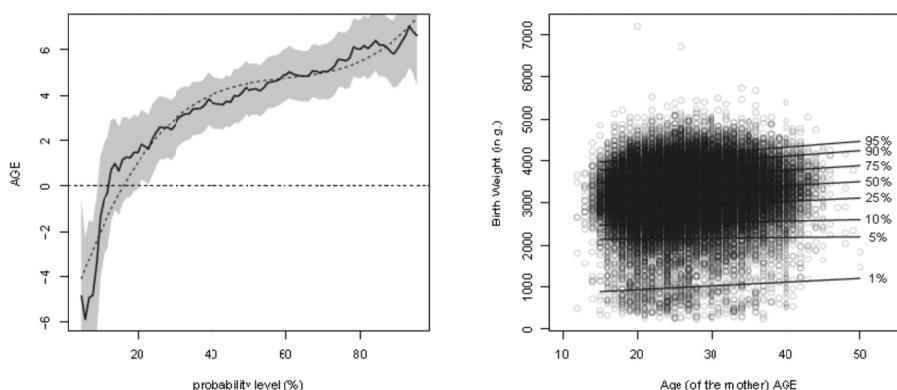
Wykorzystamy dla ilustracji wizualizacji omówionego odwzorowania $\tau \mapsto \hat{\beta}_{\tau}$ pracę Abrevaya [2001]³.

² 1 > library (quantreg),

2 > fit <- rq(dist ~ speed, data = cars, tau = .5),

3 > which (predict (fit) == cars \$ dist.

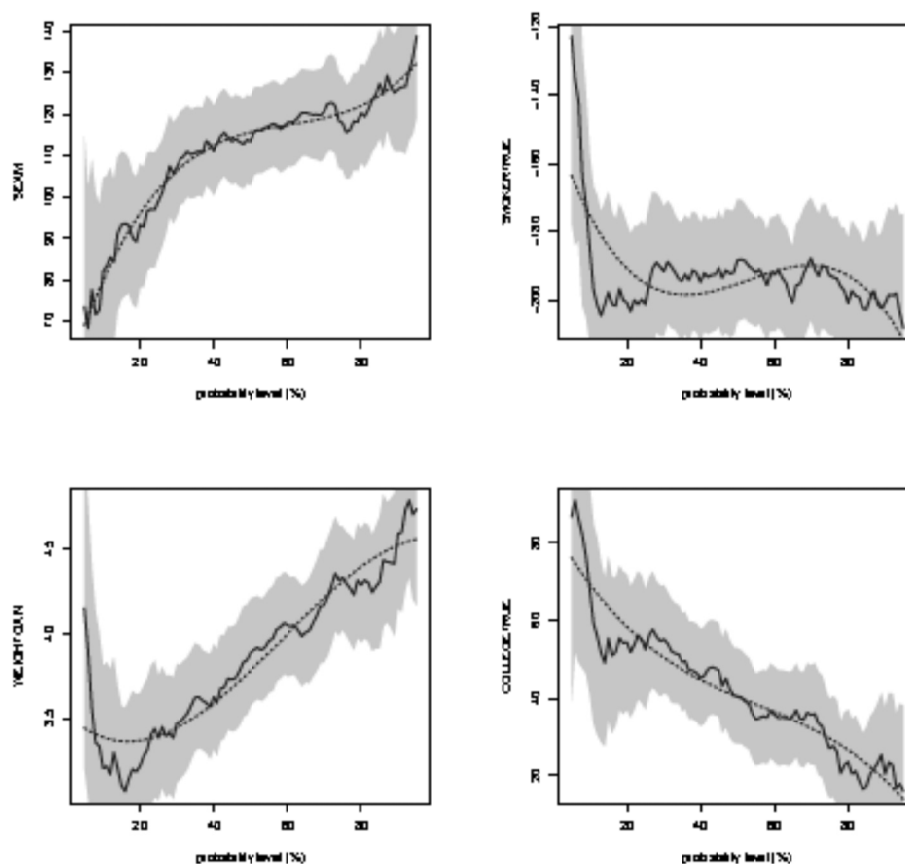
³ Badania dla dużej próby noworodków w latach 1992-1996 w USA.



Rys. 10. Regresja kwantylowa dla zmiennej objaśnianej wiek matki oraz zbiór danych: wiek matki oraz waga noworodka

Źródło: [www 1].

Wykorzystywano zbiór zmiennych takich jak: płeć noworodka, rasa, wykształcenie matki, opieka zdrowotna, waga noworodka, wiek matki, palenie papierosów przez matkę oraz wyznaczono regresje kwantylowe dla różnych przekrojów zmiennych. Uzyskano znaczącą istotność modeli dla dolnych kwantyli oraz mały wpływ dla wysokich kwantyli (rys. 10 i 11): płeć, palacz, waga noworodka, wykształcenie (SEXM, SMOKERTRUE, WEITHGAIN, COLLEGETRUE), regresja kwantylowa dla zmiennej objaśnianej wiek matki. Dopasowanie MNK nie jest w tym zbiorze liniowe, dla zmiennych objaśnianych (rys. 11) dopasowują się funkcje wielomianowe, które są trudne w interpretacjach. Zbiór danych jest dostępny online, stanowi dobry przykład sytuacji, gdy regresja kwantylowa jest bardziej użyteczna, odnosi się bowiem do rozkładów badanych zmiennych. Interpretacja modeli regresji kwantylowych jest analogiczna, jak dla klasycznej regresji, ale odnosi się do wartości warunkowego kwantyla rozkładu [Machado, Mata, 2005; d'Haultfoeuille, Givord, 2014].



Rys. 11. Wpływ zmiennych objaśniających takich jak płeć, palacz, waga noworodka, wykształcenie (SEXM, SMOKERTRUE, WEITHGAIN, COLLEGETRUE), regresja kwantylowa dla zmiennej objaśnianej: wiek matki

Źródło: [www 1].

Podsumowanie

W pracy omówiono wybrane własności funkcji kwantylowych oraz regresji kwantylowych w powiązaniu z różnymi metodami estymacji regresji kwantylowej. Wskazano na różne metody wykorzystywane w estymacji modelu, jak również możliwości wizualizacji wyników z wykorzystaniem programów komputerowych. Regresja kwantylowa jest metodą stosowaną z uwagą w różnych zastosowaniach. Kojarzona jest z miarami ryzyka na ogonach rozkładów i temu zagadnieniu poświęca się w większości prace badawcze [Trzpiot, Kręzolek,

2009, Trzpiot, Majewska, 2010] W tym artykule wykorzystano dostępną bazę danych, aby wskazać możliwość pracy na zbiorach niekoniecznie posiadających wartości odstające. Niezależną ścieżką badawczą są zastosowania regresji kwantylowej z parametrem przestrzennym [Trzpiot, 2014; Trzpiot, Orwat-Acedańska, 2016]. Prace aplikacyjne w powiązaniu z miarami ryzyka to osobne rozważania, często modele mają wówczas charakter modeli dynamicznych w odniesieniu do szeregów czasowych.

Literatura

- Abrevaya J. (2001), *The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes*, "Empirical Economics" March, Vol. 26, No. 1, s. 247-257.
- Angrist J., Chernozhukov V., Fernandez-Val I. (2006), *Quantile Regression under Misspecification*, "Econometrica" March, Vol. 74, Iss. 2, s. 539-563.
- Artzner P., Delbaen F., Eber J.-M., Heath D. (1999), *Coherent Risk Measures*, "Mathematical Finance", No. 9(3), s. 203-228.
- d'Haultfoeuille X., Givord P. (2014), *La régression quantile en pratique*, "Économie et Statistique", No. 471, s. 85-113.
- Donnelly C., Embrechts P. (2010), *The Devil is in the Tails: Actuarial Mathematics and the Subprime Mortgage Crisis*, "ASTIN Bulletin", May, Vol. 40(01), s. 1-33.
- Föllmer H., Schied A. (2002), *Convex Measures of Risk and Trading Constraints*, "Finance & Stochastics", No. 6(4), s. 429-447.
- Hyndman R., Fan Y. (1996), *Sample Quantiles in Statistical Packages*, "American Statistician", No. 50, s. 361-365.
- Koenker R. (2005), *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press, Cambridge.
- Koenker R., Bassett G. (1978), *Regression Quantiles*, "Econometrica", No. 46, s. 33-50.
- Li Q., Racine J.S. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press, Princeton.
- Machado J., Mata J. (2005), *Counterfactual Decomposition of Changes in Wage Distributions using Quantile Regression*, "Journal of Applied Econometrics" May/June, Vol. 20, Iss. 4, s. 445-465.
- Newey W.K., Powell J.L. (1987), *Asymmetric Least Squares Estimation and Testing*, "Econometrica", Vol. 55, No. 4, s. 819-847.
- RiskMetrics (1995), *Technical Document. Technical Report*, Morgan Guarantee Trust Company, Global Research, New York.
- Rockafellar R.T., Uryasev S. (2000), *Optimization of Conditional Value-at-Risk*, "The Journal of Risk", No. 2(3), s. 21-41.

- Trzpiot G. (2007a), *Decomposition of Risk and Quantile Risk Measures* [w:] *Dynamiczne Modele Ekonometryczne*, Prace Naukowe Uniwersytetu Mikołaja Kopernika w Toruniu, s. 35-42.
- Trzpiot G. (2007b), *Regresja kwantylowa a estymacja VaR*, „Prace Naukowe Akademii Ekonomicznej we Wrocławiu”, nr 1176, s. 465-471.
- Trzpiot G. (2008), *Implementacja metodologii regresji kwantylowej w estymacji VaR*, „Studia i Prace”, nr 9, Uniwersytet Szczeciński, Szczecin, s. 316-323.
- Trzpiot G. (2009a), *Application Weighted VaR in Capital Allocation*, „Polish Journal of Environmental Studies”, Vol. 18, No. 5B, s. 203-208.
- Trzpiot G. (2009b), *Estimation Methods for Quantile Regression*, „Studia Ekonomiczne. Zeszyty Naukowe Akademii Ekonomicznej w Katowicach”, nr 53, s. 81-90.
- Trzpiot G. (2014), *Some Properties of Spatial Quantiles*, „Acta Universitatis Lodzensis. Folia Oeconomica. Spatial Econometrics”, nr 5(307), Wydawnictwo Uniwersytetu Łódzkiego, s. 141-152.
- Trzpiot G. (2016), *Semi-parametric Risk Measures*, „Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach”, nr 288(5), s. 108-120.
- Trzpiot G., Krężolek D. (2009), *Quantiles Ratio Risk Measures for Stable Distributions Models in Finance*, „Studia Ekonomiczne. Zeszyty Naukowe Akademii Ekonomicznej w Katowicach”, nr 53, s. 109-120.
- Trzpiot G., Majewska J. (2010), *Estimation of Value at Risk: Extreme Value and Robust Approaches*, „Operation Research and Decisions”, vol. 20, nr 1, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, s. 131-143.
- Trzpiot G., Orwat-Acedańska A. (2016), *Spatial Quantile Regression in Healthy Life Years in the UE Countries*, „Comparative Economic Research”, vol. 19, nr 5, Uniwersytet Łódzki, s. 179-199.
- [www 1] <http://freakonometrics.free.fr/natality2005.txt> (dostęp: 15.06.2017).

GEOMETRIC PROPERTIES OF THE QUANTILE REGRESSION

Summary: The use of ordinal measures, including quantiles, is found in various areas of application, in particular robust statistics. The retreat from the classical approach based on the moments of random variables is usually the result of a data set that does not meet the assumptions of the models. The first two moments of the random variable, on which we are building the regression models, are not adequate in describing the sets of observation data sets. Data sets with asymmetric distributions should also not be analyzed using regression models estimated by MNK. The aim of this paper is to present the geometrical properties of quantile regression. This methodological approach uses the values of the distribution of the random variables by determining the set of quantiles and the set of regression models.

Keywords: quantiles, quantile regression, property of quantile regression models.