Commentary

分位点回帰について
――理論と応用――

元山　斉（信州大学学術研究院社会科学系講師）

1　はじめに

社会科学の多くの分野の分析では，複数の変数間の相互関係を数式で表し，その数式を決定するパラメータをデータから推定をして解釈を行い，パラメータの値について検定を行うのが一般的であった。そのなかでも，従来，最も多く分析手法として用いられてきたのが回帰分析である。

最小２乗法に基づいた通常の回帰分析は，説明変数を与えたときの条件付き平均値を求めることで，説明変数が目的変数の分布に与える影響を平均値で評価するのに対して，本稿で紹介するKoenker and Bassett（1978）によって導入された分位点回帰は条件付きの分位点，すなわち中央値や四分位数・十分位数などの分布の位置を推定することで，目的変数の分布に対する説明変数の影響を分布の様々な位置で評価する手法である。本稿では，分位点回帰の基本的な考え方を中心に紹介したいと思う。

最初に，なぜ分位点回帰を考えることに意味があるかを考えたい。分位点回帰のメリットは大きく２つ，分布の裾の情報をみることで従来できなかった分析を可能にすること，分位点をみることで外れ値に対して頑健な分析を可能にすることが挙げられる。それらについて順に述べていきたい。

市町村の人口，企業の資本金・売上高・従業員数などの社会・経済的規模を表すデータは，しばしば規模が大きい集団ほど散らばり具合が多く，そのようなデータを回帰分析で分析する際は，通常回帰分析で仮定する誤差項の分散均一性の条件が満足されない。

分散が不均一であるときには，説明変数の（条件付き）分布に与える影響が，分布のどの位置かによって異なってくる。**図１**，**図２**は分散不均一性によって説明変数が被説明変数の分布に与える影響が，分散が均一か否かによって，異なることを図示したものである。

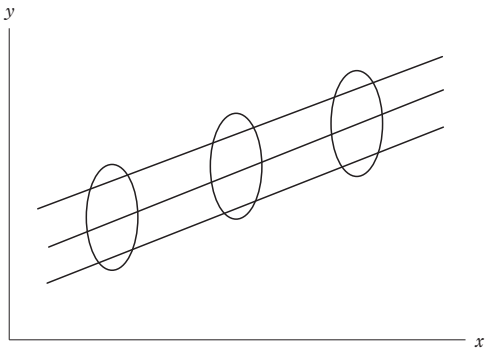

図１　分散が均一のケース

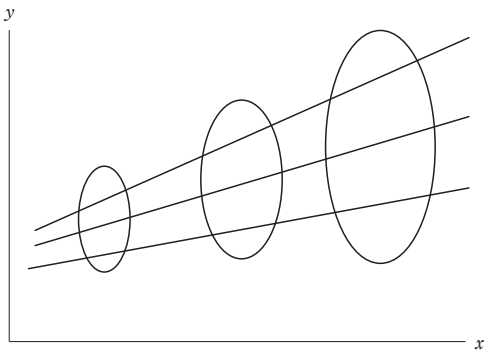

図２　分散が不均一のケース

この図において横軸xは説明変数を表しており，縦軸yは目的変数を表している。楕円は，説明変数を与えたときの，誤差項の散らばり具合，すなわち目的変数の分布の散らばり具合の大きさをイメージで表したものである。斜めに引いた直線は，説明変数を与えたときの目的変数の条件付き分布の分位点と説明変数の関係のグラフ，すなわち本稿で取り上げる分位点回帰の式を表している。

この図からわかるように，誤差項が対称に分布して分散が均一なケースでは，古典的な回帰分析で得られる説明変数を与えたときの目的変数の平均的な動きや，後ほど紹介する最小絶対偏差回帰で得られる目的変数の条件付き中央値と，分位点回帰で得られる説明変数を与えたときの目的変数の分布の分位点との関係に違いはみられない。しかし，分散不均一のケースでは，条件付き分布の挙動は分位点ごとに異なっており，最小2乗法で得られる条件付きの平均値だけでは，分布の挙動を正確に測ることができないことがわかる。

社会保障政策・経済政策の分析や立案を行う際には，その政策が与える効果を評価する必要がある。たとえば，就学促進の政策の効果を測定するためには，就学年数が所得水準に与える影響を知る必要がある。しかしながら，就学年数が異なるグループ間の所得の分散が異なっているときは，上記の図からも想像されるように，就学年数が所得に与える影響は，所得が低いグループと所得が高いグループでは異なっているため，古典的な回帰分析がもたらす結論だけでは誤った解釈に陥る可能性がある。分位点回帰では，説明変数を与えたときの目的変数の分布の裾の挙動について知ることができるので，教育が所得に与える影響を所得水準の違いごとに評価することができる。

応用例1：Chamberlain（1994）はアメリカの製造業における労働者の賃金に労働組合加入が与える影響を分析した。Chamberlain（1994）では，従来の分析で用いられてきた最小2乗法による回帰と分位点回帰による比較を行っている。通常の回帰では労働組合に入っている労働者は平均的に15.8% 賃金が高いという結論が得られたが，分位点回帰を行ったところ，労働組合に入っていることが与える影響は，第1十分位点に対しては28% であるが，上の位の分位点になるにつれて，影響は単調に減少して，第9十分位点に対しては0.3% にまで落ちることが確認される。よって，最小2乗法による分析は主に賃金が低いグループの影響を捉えていると考えられる。

分位点回帰のメリットのもう1つは，分位点が平均値と比べて極端に大きい値または小さい値である外れ値に対して頑健であり家計の所得・資産額などの社会・経済データの多くにみられる歪んだ分布をもつデータの分析に適していることが挙げられる。以下は，架空の数値例である。

例（架空の数値例）：ある企業に所属する7人の年間所得（単位万円）

330，280，230，240，290，340，1,580

この数値は，1人だけ1,580万円と大きな値があるが，それ以外は大体300万円前後の値である。このデータで平均値を求めると470万円となり，1,580万円の影響を受けて，それ以外の値のすべてよりも大きな値となっており，データの代表値としては不適切であると考えられる。

この問題に対処する1つの方法は，平均に大きな影響を与える少数だが極端な値である外れ値を取り除くことであろう。上のデータから一番大きい1,580万円と一番小さい230万円を除いて平均を取ると296万円となり，この値はデータの中心を表す代表値として，まずまず妥当なものといえよう。このような計算方法はトリム平均（trimmed mean）と呼ばれているもので，身近な例としては体操競技における評価が挙げられる。体操競技などでは1人の人の競技に何人かの審判員が点数をつけ，その点数のなかから最低点と最高点を除いて平均点を採用する。このような方式の歴史は古く，18世紀のフランスでは土地の価格の算定に用いられていたとのことである（竹内，1980: 21）。

中央値（median）は，このようなトリム平均の特殊なケースであり，データの中心を表す代表値のなかで一番外れ値に対して頑健なものと考えられる。データのなかから大きいデータと小さいデータを除いていって最後に残る値はデータの中央に位置する290万円であり，これはデータの中央値となっている。以上の議論からもわかるように，大きい値または小さい値の外れ値の数がデータの数 n の半分 $n/2$ を超えないかぎり，中央値は影響を受けないという意味で頑健である。同様に，データの小さいほうから数えて 100τ% の位置にある τ-分位点は，外れ値の数が $n\ \min\{\tau, 1-\tau\}$ を超えないかぎり影響を受けない頑健な指標である。

分位点回帰は，条件付きの平均値ではなく条件付きの分位点を分析するものであるので，極端な値の影響を受けない頑健な分析を可能にするものと考えられる。

応用例2：Koenker and Hallock（2001）では，

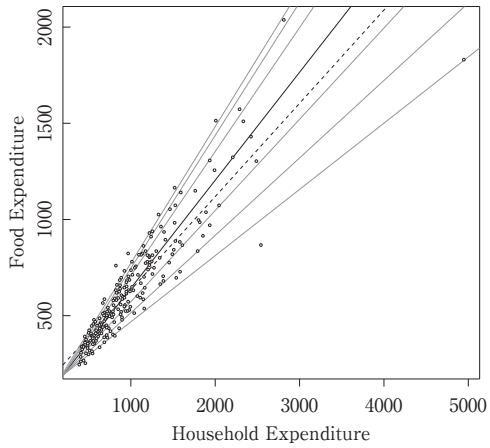

図３　エンゲル曲線の推定

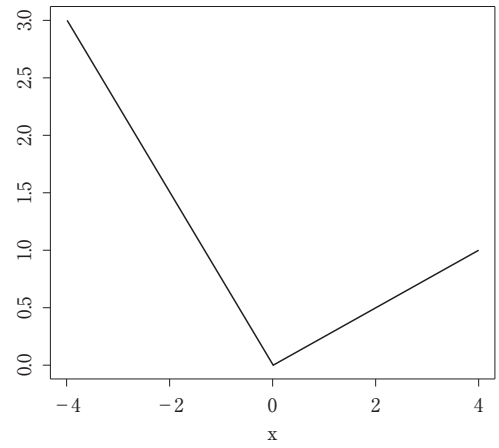

図４　チェック関数（$\tau=0.25$ のケース）

Engel の 1857 年の論文「ザクセン王国の生産及び消費事情」において用いられた，当時の勤労者世帯 235 世帯の家計データに基づくエンゲル曲線を推定している。**図 3** は R 言語の quantreg パッケージと Koenker のホームページ（http://www.econ.uiuc.edu/~roger/）で公開されているプログラムを用いて計算した結果のグラフである。

破線で示されているのが最小２乗法による回帰直線であり，下からそれぞれ 5%，10%，25%，50%，75%，90%，95% の分位点を回帰式で表している（条件付き中央値に対応する 50% 分位点は太い線で表示されている）。

このグラフの形状から，前述の分散不均一性の影響による，回帰係数の傾きの違いがみてとれる。分位点が下位から上位に上がるにつれて，傾きが急になっており，食費に対する支出のばらつきが高額支出者の方が大きいことが示唆される。また，最小２乗法の破線が外れ値の影響を受けて下方にあるのに対して，50% 分位点（条件付き中央値）を表す式は影響を受けていないことがみてとれる。

2　分位点回帰の推定方法

本節では，分位点回帰のパラメータの推定方法について述べる。議論を行うため，本節以降では回帰式を

$$y_i = X_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n$$

と表す。ここで，y_i はスカラー値の目的変数であり，X_i は k×1 の説明変数ベクトル，β は k×1 のパラメータベクトル，′はベクトルの転置を表す。

通常の最小２乗法の回帰では，データ $x_1, \ldots, x_n$ の平均値 $\bar{x}$ が２乗損失関数 $\Sigma_{i=1}^{n}(x_i-a)^2$ を最小にする a であることに対応して，$\Sigma_{i=1}^{n}(y_i-X_i'\beta)^2$ を最小にするベクトル β を求めてパラメータの推定値とするが，分位点回帰のパラメータの推定は，分位点を与えるチェック関数（check function）または非対称絶対損失関数（asymmetric absolute loss function）と呼ばれる損失関数で評価した損失

$$\sum_{i=1}^{n}\rho_\tau(y_i - X_i'\beta)$$

を最小にする値として計算される。

最初にチェック関数で評価した損失を最小にする値が，分位点であることを示そう。チェック関数 $\rho_\tau(u)$ とは，τ $(0\leqq\tau\leqq1)$ に対して

$$\rho_\tau(u) = \begin{cases} (\tau-1)u & u\leqq 0 \\ \tau u & u>0 \end{cases}$$

で表される関数であり，グラフの形状が**図 4** のようにチェック・マークに似ていることから，そのように呼ばれる。$\Sigma_{i=1}^{n}\rho_\tau(x_i-a)$ を最小にする a は $x_1, \ldots, x_n$ の τ 分位点である。

チェック関数は，$\tau=0.5$ のとき

$$\rho_{0.5}(u) = 0.5|u| = \begin{cases} -0.5u & u\leqq 0 \\ 0.5u & u>0 \end{cases}$$

と絶対値関数の 0.5 倍となる。$\Sigma_{i=1}^{n}\rho_{0.5}(x_i-a) = 0.5\,\Sigma_{i=1}^{n}|x_i-a|$ を最小にする a は $x_1, \ldots, x_n$ の中央値（0.5 分位点）である。

$x_1, \ldots, x_n$ の τ 分位点$(0\leqq\tau\leqq1)$が $\Sigma_{i=1}^{n}\rho_\tau(x_i-a)$

を最小にすることの厳密な証明は，Manski（1988）の pp.54–56 や Koenker（2005）の pp.5–6 で行われているが，ここではいくつかの例に限定して直感的な議論を行う。はじめに，$\tau=0.5$ のとき $\Sigma_{i=1}^{n}\rho_{0.5}(x_i-a)=0.5\ \Sigma_{i=1}^{n}|x_i-a|$ を最小にする a が $x_1, \ldots, x_n$ の中央値であることを示す。絶対値関数は 0 となる点では微分できないが，連続型データを念頭に $|x_i-a|=0$ となる可能性を無視すると 0 となる点以外では，場合分けをすると微分ができて，

$$(a>x_i \text{ のとき}) \quad \frac{d}{da}|x_i-a|=\frac{d}{da}(a-x_i)=1$$

$$(a<x_i \text{ のとき}) \quad \frac{d}{da}|x_i-a|=\frac{d}{da}(x_i-a)=-1$$

となるので，a が x_i を小さい順に並べたときの真ん中の点（すなわち中央値）で $\Sigma_{i=1}^{n}\rho_{0.5}(x_i-a)=0.5\ \Sigma_{i=1}^{n}|x_i-a|$ は最小になることがわかる。

次に，$\tau=0.25$ のとき $\Sigma_{i=1}^{n}\rho(x_i-a)$ を最小にする a は $x_1, \ldots, x_n$ の 0.25 分位点（第 1 四分位点）であることを確認する。

$$\rho_{0.25}(x_i-a)=\begin{cases}-0.75(x_i-a)=0.75|x_i-a|, & x_i\leq a\\ 0.25(x_i-a)=0.25|x_i-a|, & x_i>a\end{cases}$$

であるので，先ほどと同様に

$$(a>x_i \text{ のとき}) \quad \frac{d}{da}|x_i-a|=\frac{d}{da}(a-x_i)=1$$

$$(a<x_i \text{ のとき}) \quad \frac{d}{da}|x_i-a|=\frac{d}{da}(x_i-a)=-1$$

となることから

$$\frac{d}{da}\Sigma_{i=1}^{n}\rho_{0.25}(x_i-a)=0.75\times(a>x_i \text{ の } x_i \text{ の数})-0.25\times(a<x_i \text{ の } x_i \text{ の数})$$

と表わされるので，$(a>x_i$ の x_i の数$)$：$(a<x_i$ の x_i の数$)=1:3$ となる a すなわち，a が $x_1, \ldots, x_n$ の第 1 四分位点のときに $\Sigma_{i=1}^{n}\rho_{0.25}(x_i-a)$ は最小となる。

同様に，一般の τ のときも同様に a が τ 分位点であるときに $\Sigma_{i=1}^{n}\rho_{\tau}(x_i-a)$ は最小になることを示すことができる。

ここまででチェック関数を損失関数として用いることで分位点が求められることを紹介してきたが，分位点回帰においては，その対応から

$$\sum_{i=1}^{n}\rho_{\tau}(y_i-X_i'\beta)$$

を最小にする β を求めることでパラメータ β の推定値を得ることができる。

$\tau=0.5$ のときは，最小絶対偏差回帰（least absolute deviation regression: LAD）と呼ばれ，説明変数を与えたときの目的変数の条件付き中央値を求める手法として，古くから知られた手法である。最小絶対偏差回帰は，前節で行った中央値と平均値の比較から，最小 2 乗法に対して外れ値に頑健な回帰を行っていると解釈することも可能ではあるが，目的変数の分布が左右対称であるなどの条件を満足しないかぎり，条件付き平均値と条件付き中央値は一致しないので，本来は別のものを推定していることを心に留めたほうがよい。古典的な回帰と最小絶対偏差回帰の解釈におけるこれらの留意点については，Wooldridge（2013）の 9.6 節の議論を参照されたい。

3 分位点回帰の回帰係数の解釈，あてはまりの評価，係数の検定

本節では分位点回帰を行った際の係数の解釈，あてはまりの評価，回帰係数の有意性検定について紹介する。

分位点回帰の回帰係数は，通常の回帰と同様に説明変数が 1 単位増加をしたときの影響として解釈が可能である。通常の回帰との違いは，説明変数との間に記述されている関係が目的変数の（条件付き）分布の平均値であるか分位点であるかの違いにすぎない。

また重回帰分析のときの係数の解釈についても，通常の回帰と同様に解釈ができる。最小 2 乗法による通常の回帰分析においては，回帰係数は「他の説明変数の影響を取り除いたうえでのその変数の影響」と解釈されることはよく知られたことである。この解釈の数理的根拠は残差回帰と呼ばれる「ある説明変数と目的変数を他の説明変数に回帰した後で，残差同士を回帰させたときの単回帰の係数が重回帰の係数と一致する」という結果であり，加重最小 2 乗法においても同様の結果が得られている。

分位点回帰における回帰係数についても，Angrist et al.（2006）が示したように，分位点回帰推定量はある種の加重最小 2 乗法の解として得られることから，通常の回帰分析と同様の解釈が可能である。すなわち，分位点回帰の回帰係数は他の説明変数の影響を除いた後，その変数が被説明

数の分位点に与える影響を表していると解釈することができる。

このように，分位点回帰の回帰係数は，その推定対象や推定方法の違いにもかかわらず，通常の回帰分析と同様の解釈が可能であるという意味で汎用性の高いものであるが，政策提言を行ううえでは注意しなければならない点があるのでここで紹介する。Angrist and Pischke（2009）の 7. 1. 3 節で注意されていることだが，分位点回帰係数は，あくまでも条件付き分布の分位点に対する効果を表したものであり，個人に対する効果を表したものではない。すなわち，ある職業訓練や就学年数の追加が，賃金分布の下位の分位点を押し上げる効果があったとしても，現在の貧しい人たちの状況が職業訓練を受けることや就学年数が伸びることで，分位点回帰式の係数で評価されるだけ改善することは保証されるとはいえない。なぜなら，このような職業訓練や就学年数を増加させる政策を行ったときと行わなかったときの条件付き分布において，同じ個人が相対的に同じ位置にいるとは限らないからである。しかしながら，政策実施前と同様の個人でないとしても，賃金分布の下位グループの賃金を押し上げるという意味において，政策の効果を評価することはできる。

最後に回帰式のあてはまりの評価方法と係数の有意検定について，簡単に紹介する。通常の最小 2 乗法に基づいた回帰分析では，回帰式のあてはまりは，説明変数が 1 つの単回帰分析のときは決定係数や自由度修正済みの決定係数で行うのが一般的である。

決定係数を例にとると，最小 2 乗法による回帰では回帰式で説明できる変動を全体の変動で割った値，書き換えると，全体の変動において回帰式で説明できなかった変動を割った値を 1 から引いた

$$R^2=\frac{\Sigma_{i=1}^{n}(X_i'\beta-\bar{y})^2}{\Sigma_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\Sigma_{i=1}^{n}(y_i-X_i'\beta)^2}{\Sigma_{i=1}^{n}(y_i-\bar{y})^2}$$

（ここで $\bar{y}$ は $y_1, \ldots y_n$ の平均値）

が大きく 1 に近ければ近いほどあてはまりがよく，小さく 0 に近ければ近いほどあてはまりがよくないと評価が行われる。

分位点回帰の場合も，この類推から，2 乗で損失を評価するのではなく，チェック関数で損失を評価した pseudo R^2

$$pseudo\ R^2(\tau)=1-\frac{\Sigma_{i=1}^{n}\rho_\tau(y_i-X_i'\beta)}{\Sigma_{i=1}^{n}\rho_\tau(y_i-Q_\tau(y))}$$

（ここで $Q_\tau(y)$ は $y_1, \ldots y_n$ の τ-分位点）

を通常の回帰と同様に解釈することで，分位点回帰の式のあてはまりを評価することができる（Hao and Naiman, 2007）。

また，分位点回帰推定量は，漸近的に正規分布に従うことから（Koenker and Bassett, 1978; Koenker, 2005），係数の有意性の検定ができる。しかしながら，漸近分布の分散を求めることは必ずしも容易ではなく，ブートストラップ法による推定量を含むいくつかの推定量が提案されており，ソフトウェアでも分散計算については複数のオプションを用意してあるものがある。

4 おわりに

本稿では，社会科学データの分析において，最近，急速に需要が高まっている分位点回帰について，基本的な考え方に焦点を絞って紹介した。さらに詳しく分位点回帰について知りたい方は，文献に挙げた Hao and Naiman（2007）や Koenker（2005）などを参照されたい。分位点回帰は，本稿で取り上げたエンゲル曲線や賃金関数の推定以外にも，出生児の体重に影響を与える要因の分析や損害保険の保険料や金融資産の分布の分析などの広範囲のテーマで用いられ始めている。これは分位点回帰が社会科学データの分析において強力なツールであることが認識されていることを反映していると思われる。今後の理論・実証研究のさらなる発展を祈念したい。

文献

Angrist, J., V. Chernozhukov and I. Fernandez-Val, 2006, “Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure,” *Econometrica*, 74: 539–563.

Angrist, J. and J-S. Pischke, 2009, *Mostly Harmless Econometrics: An Empricist's Companion*, Princeton, NJ: Princeton University Press.（大森義明・小原美紀・田中隆一・野口晴子訳，2013，『「ほとんど無害」な計量経済学：応用経済学のための実証分析ガイド』NTT 出版。）

Chamberlain, G., 1994, “Quantile Regression, Censoring and the Structure of Wage,” in C.A. Sims ed., *Ad-*

vances in Econometrics: Vol.2.: Sixth World Congress (Econometric Society Monographs), Cambridge; New York: Cambridge University Press, 171–209.
Hao, L. and D. Q. Naiman, 2007, *Quantile Regression* (Quantitative Applications in the Social Sciences), Sage Publications, Inc.
Koenker, R., 2005, *Quantile Regression* (Econometric Society Monographs), Cambridge: Cambridge University Press.
Koenker, R. and G. Bassett, 1978, "Regression Quantiles," *Econometrica*, 46(1): 33–50.
Koenker, R. and K.F. Hallock, 2001, "Quantile Regression," *Journal of Economic Perspectives*, 15(4): 143–56.
Manski, C.F.,. 1988,. *Analog Estimation Methods in Econometrics*, London; New York: Chapman & Hall.
竹内啓, 1980, 『現象と行動のなかの統計数理』新曜社。
Wooldridge, J. M., 2013, *Introductory Econometrics: A Modern Approach*, 5th ed., Australia: South-Western Cengage Learning.