
ポスター発表

[PB] ポスター B

2020年6月5日(金) 09:00 ~ 18:30 ポスター会場(2) (e-poster)

[PB-17] BMIと年齢のシミュレーションデータを用いた Quantile regressionの性能評価 Evaluating quantile regression using simulation data including BMI and Age

*西本 尚樹¹、田村 菜穂美²、伊藤 陽一² (1. 北海道大学病院 臨床研究開発センター、2. 統計数理研究所 医療健康データ科学研究センター)

*Naoki Nishimoto¹, Naomi Tamura², Ito. M. Yoichi² (1. Hokkaido University Hospital, Clinical Research and Medical Innovation Center, 2. The Institute of Statistical Mathematics, Research Center for Medical and Health Data Science)

BMI と年齢のシミュレーションデータを用いた Quantile regression の当てはめ

西本 尚樹^{*1}, 田村 菜穂美^{*2}, 伊藤 陽一^{*2}

^{*1} 北海道大学病院 臨床研究開発センター, ^{*2} 統計数理研究所 医療健康データ科学研究センター

Evaluating quantile regression using simulation data including BMI and Age

Naoki Nishimoto^{*1}, Naomi Tamura^{*2}, Yoichi M. Ito ^{*2}

^{*1} Clinical Research and Medical Innovation Center, Hokkaido University Hospital,

^{*2} Research Center for Medical and Health Data Science, the Institute of Statistical Mathematics

抄録:

本研究の目的は、サンプルデータを用いて、quantile regression で推定される値が、どの程度データから影響を受けるのかを明らかにすることとした。サンプルデータは、米国の the national center for health statistics が公開している national health and nutrition examination survey (NHANES) の 1999-2000 年及び 2001-2002 年に調査された男性のデータを用いた。モデルは、BMI の p% 分位点に対して、線形モデルを仮定した。外れ値の存在が示唆される部分で、90% 点以上の当てはめに大きなゆがみが見られた。外れ値の影響は受けにくいモデルではあるが、パラメータの設定に注意を要する。Quantile regression により、外れ値による影響は受けにくい、モデル自体のパラメータの設定によって 90% 点より大きい値の当てはめが影響を受けた。今後は、RWD の解析で柔軟な解析方法としての利用が期待される。

キーワード Quantile regression, BMI, Age, 欠測, simulation

1. はじめに

近年、臨床試験のような実験的な環境下で収集されたデータに対して、日常診療で収集されるデータを real-world data (以下、RWD) という概念が提唱されている[1]。Sherman らによって発表された論文では、日常診療実態下で収集されたデータをエビデンスの創出に直接利用することが求められている。しかし、RCT のように、治療や割付の方法が制御されていない日常診療下では、コホート研究のように、そもそも収集されていない変数などもあり得る。

電子カルテで収集されたデータは、質・量ともに膨大であるものの、選択バイアス・情報バイアスなどの、構造的な偏りを多分に含んでいる。さらに、年齢毎の BMI や、時系列推移により世代間の変化を可視化しようとしても、平均値や標準偏差は外れ値の影響を大きく受けるため、回帰分析などの手法を持って、調整した解析を行うにしても、外れ値

の存在により大きな影響を受ける。そこで、1978 年に Koenker らが、quantile regression (分位点回帰) という方法を発表した[2-4]。RWD を解析するとき、外れ値に影響されにくい方法の一つとして、quantile regression が有効であろうと考えられるが、データの特性や欠測から受ける影響を可視化した研究は少ない。そこで、本研究の目的は、サンプルデータを用いて、quantile regression で推定される値が、どの程度データから影響を受けるのかを明らかにすることとした。

2. 方法

1) サンプルデータ及び解析方法

(1) シミュレーションデータ

シミュレーションには、国民健康・栄養調査 (平成 29 年) の BMI を参考に、BMI 期待値を年齢の 3 乗項までの線形和とした。誤差の分布を正規分布、F 分布として、3000 人分の BMI データを人工的に発生させて、挙動を調べた。

(2) 用いたサンプルデータ

統計解析パッケージ SAS(SAS Institute Inc., Cary, NC, USA)に収録されている 3,264 人分の BMI と年齢のみが変数のサンプルデータ(BMI データ)を用いた。サンプルデータは、米国の the national center for health statistics が公開している national health and nutrition examination survey (NHANES) の 1999-2000 年 及び 2001-2002 年に調査された男性のデータであり、8250 人からさらにサンプリングが行われている [5-6]。

(3) 解析パッケージ

SAS9.4(SAS Institute Inc., Cary, NC, USA)及び JMP14.0、R3.6.1(The R Foundation for Statistical Computing)を用いた。

(4) 解析方法

SAS の HPQUANTSELECT プロシージャを用いた。R では、quantreg library(Ver.5.54)を使用した。モデルは、BMI の p%分位点に対して、線形モデルを仮定した。

$$Y_{p,i} = a + b \cdot X_{age} + c \cdot X_{age}^2 + d \cdot X_{age}^3 + e_i$$

上記のモデルから、年齢変数 X_{age} のベキ項を一つずつ削除したモデルを作った。シミュレーションでは、BMI データをランダムに 10%ずつ、50%まで削除したデータセットを作成し、quantile regression を 1000 回繰り返し、推定値のバラツキの変化を可視化した。

3. 結果

1) Quantile regression の当てはめ

年齢に対して BMI の 3%点から 97%点までの当てはめを Fig.1 に示す。

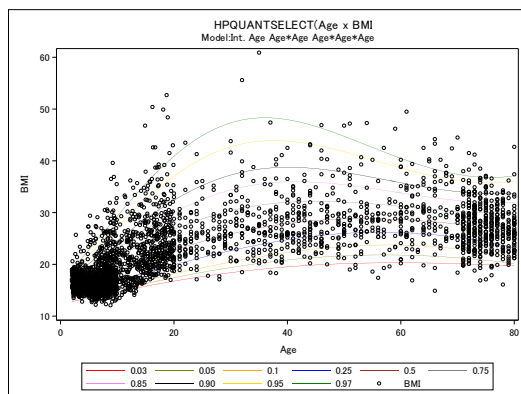


Fig1 quantile regression の 3 次関数の当てはめ。

発表では、その他の分位点についてモデルによる推定結果を示す。

4. 考察

1) モデルのパラメータ設定について

Fig.1 より、外れ値の存在が示唆される部分で、90%点以上の当てはめに大きなゆがみが見られた。外れ値の影響は受けにくいモデルではあるが、パラメータの設定に注意を要する。

5. 結語

Quantile regression により、外れ値による影響は受けにくい、モデル自体のパラメータの設定によって 90%点より大きい値の当てはめが影響を受けた。今後は、RWD の解析で柔軟な解析方法としての利用が期待される。

参 考 文 献

- [1] Rachel E. Sherman, et al. Real-world evidence - What is it and what can it tell us? NEJM. 375(23). 2293-2297. 2016.
- [2] Roger Koenker and Gilbert Bassett. Regression Quantiles. Econometrica, 46(1). 1978. 33-50.1978.
- [3] C.M. Hurvich, CL Tsai. Regression and time series model selection in small samples. Regression and time series model selection in small samples. Biometrika. 76(2). 297-307. 1989.
- [4] E.M. Gertz, J.D. Griffin. Using an iterative linear solver in an interior-point method for generating support vector machines. Comput Optim Appl. 47. 431-453. 2010.
- [5] SAS. SAS/STAT14.2 User's Guide The HPQUANTSELECT Procedure. SAS Institute Inc., Cary, NC, USA. 2016.
- [6] The National Center for Health Statistics. URL: <https://www.cdc.gov/nchs/index.htm>. (Accessed FEB.25 2020)