

A Quantile Approach to Ordinal Regression with Application to Aging Research

Hyokyoung (Grace) Hong

Baruch College, the City University of New York
Joint work with Xuming He, University of Illinois at Urbana-Champaign

July 29, 2009

- People are living longer;
- Quality of life at advanced ages is important;
- **Functional status** is frequently used as an ordinal variable;
- Can we help predict functional status well (statistically)?

Functional status is categorized based on ADL and IADL

Functional status : ability to perform self-care, self-maintenance, and physical activities;

Activity of Daily Living (ADLs)

bathing/ showering, dressing, eating, getting in/out of bed/chairs, walking, and using toilet

Instrumental ADL (IADLs)

preparing meals, shopping for groceries, managing money, using the telephone, doing heavy housework, doing light housework, getting outside, and managing medication.

Defining the functional status

Adapted from Anderson et al. (1988)'s classification of functional status (*FS*)

- 1=**Independent**: able to perform all ADL and IADL activities
- 2=**IADL disabled only**: unable to perform one or more IADLs but had no ADL disabilities
- 3=**Moderately ADL disabled**: unable to perform one or two ADL activities
- 4=**Severely ADL disabled**: unable to perform more than two ADL activities
- 5=**Death**: the end of the functional ability

The data source: the Second Longitudinal Study of Aging (LSOA II)

Purpose : providing the characteristics of aging

Measure change in health status, health-related behaviors and health care, and the causes and consequences of these changes.

Subjects

9,447 non-institutionalized, U.S. citizens aged 70 years and over

Period

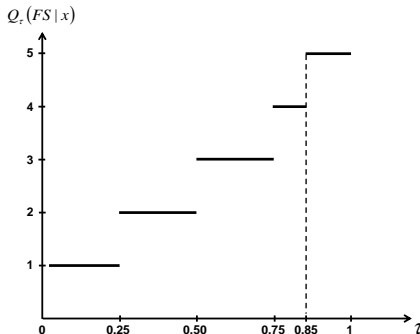
1994–1996 (Wave 1); 1997–1998 (Wave 2); 1999–2000 (Wave 3).

Main Objectives

- To develop more flexible ordinal regression models (beyond the ordered probit and logit models).
- To predict quantiles of the functional status, y , at 2-year time intervals, given available covariates, x .

Why do we care about quantiles?

- Knowing quantiles = knowing distribution.
- With 75% probability FS will be 4 or better
- The probability of death is 0.15.



Estimation problem of quantiles with ordinal response

- Typical quantile regression (Koenker and Bassett, 1978) handles continuous response variables.
Solution: Artificial smoothing of the data in linear models (Machado and Santos Silva (2005)).
- Linear quantile regression models may not be appropriate.
Solution: A proper transformation of the response variable y may induce linearity (Mu and He (2007)).

The proposed model

Introducing a random auxiliary variable

$$z_i = y_i + u_i,$$

where $y \in \{1, 2, 3, \dots\}$, and $u \sim U(0, 1)$ is random addition to y .
Then we adopt the following model :

$$\Lambda(z_i) = \mathbf{x}_i' \beta_0 + \epsilon_i,$$

for a monotone function Λ , where ϵ_i 's are independent noise.
- constraint in β_0 needed.

♠ This model generates large number of popular models, such as the probit, logit, and the Box-Cox transformation.

Transformed Ordinal Regression Quantile Estimator (TORQUE)

Method 1 (TORQUE with Quantile Regression)

We represent the Model $\Lambda(z_i) = x_i' \beta_0 + \epsilon_i$ as

$$Q_\tau(\Lambda(z_i)|x_i) = x_i' \beta_\tau,$$

for some coefficient β_τ , where the τ -th quantile of $\epsilon_i|x_i$ is 0.

Method 2 (TORQUE with Quantiles of Residuals)

An approach tailored to i.i.d. errors in the proposed model.

Let $F_\epsilon^{-1}(\tau) = \inf\{\epsilon : F(\epsilon) \leq \tau\}$. Then

$$Q_\tau(\Lambda(z_i)|x_i) = x_i' \beta + F_\epsilon^{-1}(\tau)$$

Computation of conditional quantiles

Proposition

$Q_\tau(y_i) = \lfloor Q_\tau(z_i) \rfloor$, where where $\lfloor x \rfloor$ is a function that returns the highest integer less than or equal to x .

- $\hat{Q}_\tau(z_i|x_i) = \hat{\Lambda}^{-1}(x_i'\hat{\beta}_n)$
- $\hat{Q}_\tau^m(y_i|x_i) = \lfloor \hat{Q}_\tau^m(z_i|x_i) \rfloor = \lfloor \frac{1}{m} \sum_{l=1}^m \hat{\Lambda}^{-1(l)}(x_i'\hat{\beta}_n^{(l)}) \rfloor, l = 1, \dots, m$

Studies

We specify the following settings throughout the studies.

- The sample size is 1000 for each data set, and a total of 100 data sets will be generated in each study.
- $y = 1, 2, 3, 4$

The studies are chosen based on the proposed model $\Lambda(z_i) = x_i' \beta_0 + \epsilon_i$ with some transformation of z , $\Lambda(z)$, and some distribution F .

Studies

Study 1

$$2z = x_1 + x_2 + 5 + \epsilon,$$

where $x_1 \sim N(0.5, 0.5)$, $x_2 \sim N(0.5, 0.5)$, and $\epsilon \sim N(0, 1)$

Study 2

$$3z = x_1 + x_2 + 3 + \epsilon,$$

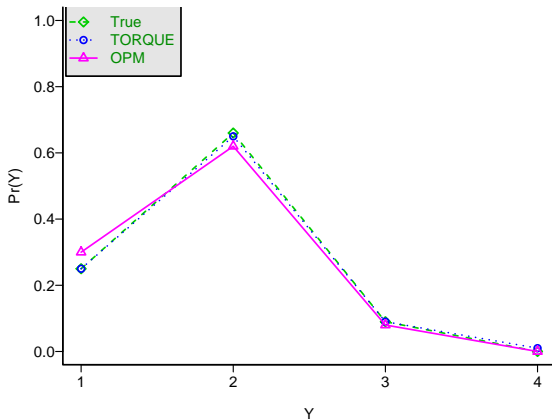
where $x_1 \sim U(0, 5)$, $x_2 \sim U(0, 5)$, and $\epsilon \sim LN(0, 0.75)$

Study 3

$$\log(7z) = x_1 + x_2 + \epsilon,$$

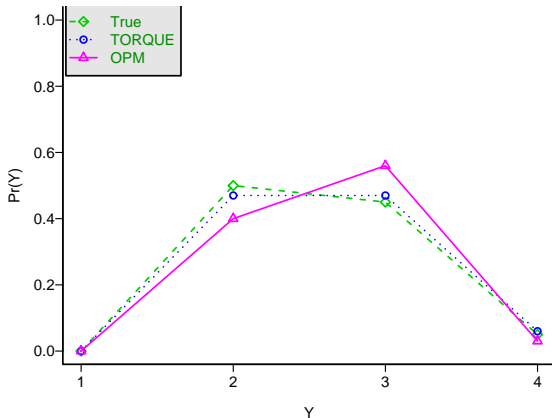
where $x_1 \sim U(0, 5)$, $x_2 \sim U(0, 5)$, and $\epsilon \sim LN(0, 0.75)$

Study 1: $\hat{P}(Y = j)$, ($j = 1, 2, 3, 4$) at the mean covariates



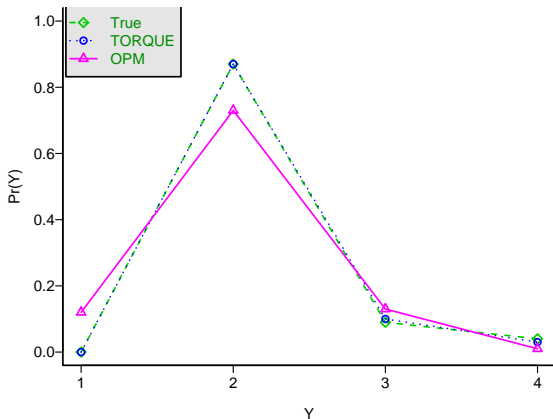
The TORQUE model predicts the true probability as accurately as the OPM does.

Study 2: $\hat{P}(Y = j)$, ($j = 1, 2, 3, 4$) at the mean covariates



The OPM does not predict the true probability well, whereas the TORQUE model does.

Study 3: $\hat{P}(Y = j)$, ($j = 1, 2, 3, 4$) at the mean covariates



The OPM does not predict the true probability well, whereas the TORQUE model does.

D_2 for comparing quality of the prediction performance

- $\mathbf{p} = (p_1, \dots, p_4)$, where $p_j = P(y = j)$ and $j = 1, 2, 3, 4$.
- $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_4)$, where \hat{p}_j is an estimator of $P(y = j)$.

Definition

- M_1 data set and M_2 subject.
- $j = 1, \dots, m$, where $1 < \dots < m$.
- $D_2 = \sum_{h=1}^{M_1} \sum_{i=1}^{M_2} S(\mathbf{p}_{hi}, \hat{\mathbf{p}}_{hi}) / M_1 M_2$,
where $S(\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^m |X_j - Y_j|$
- D_2 measures how the model fits the true probability accurately.

Summary of D_2 for 100 data sets

- The smaller the value becomes, the better the model is in predicting the true probabilities.

Table: Comparison of D_2

Study 1		Study 2		Study 3	
TORQUE	OPM	TORQUE	OPM	TORQUE	OPM
0.09(5)*	0.05(95)	0.06(100)	0.26(0)	0.06(100)	0.28(0)

* The numbers in parenthesis represents percentage of smaller distances within a data set.

Covariates

In selecting appropriate explanatory variable we refer to Anderson *et al.* (1988) and Lee *et al.* (2006) for details.

Continuous variables

age, education

Binary variables

sex, race, cardiovascular diseases (CVD), musculoskeletal diseases (MSD), diabetes, lung disease, smoking, # of conditions (1 = # of condition > 2, 0 = otherwise), self-rated health (SRH) (1 = excellent/very good, 2 = good/fair/poor)

Checking the Normal Error Assumption

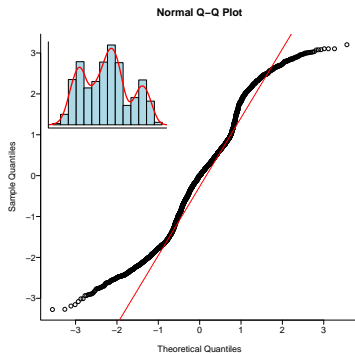


Figure: Q-Q plot for the Pearson residuals from the TORQUE model in the LSOA II data.

Estimation result for Λ

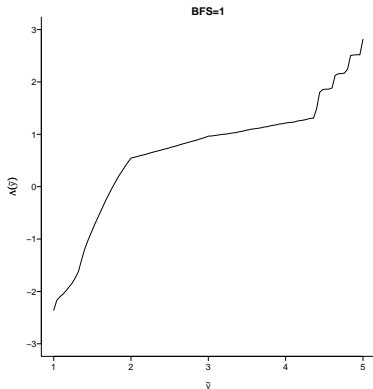


Figure: Estimation result for Λ of the group at BFS=1

Length of the Prediction Intervals (PIs) of each model

- We use as a 50% prediction interval (PI), $(\hat{Q}_{.25}(y_i|x_i), \hat{Q}_{.75}(y_i|x_i))$.
- About 25% subjects died after the second follow-up.
→ Not meaningful to consider $\tau > 0.75$.
- Length of PI: $L = |\hat{Q}_{.75}(y_i|x_i) - \hat{Q}_{.25}(y_i|x_i)| \in \{0, 1, \dots, 4\}$
- Longer length implies that the prediction is not so informative.

Length of the Prediction Intervals (PIs) of each model

Table: Length comparison

Estimation data set (N = 2692)					
Method	L = 0	L = 1	L = 2	L = 3	L = 4
OPM	35%	31%	9%	5%	19%
TORQUE(M1)	24%	30%	42%	4%	1%
Validation data set (N = 1778)					
OPM	35%	29%	10%	6%	20%
TORQUE(M1)	26%	27%	43%	2%	3%

% of total N

- The OPM gives many $L = 4$.

The mean coverage and length of the PIs

Table: Comparison of coverage and length

Method	OPM		TORQUE	
data sets	Est	Val	Est	Val
\bar{C}	0.84	0.84	0.80	0.79
\bar{L}	1.36	1.44	1.24	1.25

- TORQUE is closer to the nominal level, and both are conservative.
- TORQUE has a shorter length.

Advantage of the TORQUE model: a summary

- Weak assumption needed about the error distribution.
- Weak assumption needed about the transformation of z , Λ .

- Prediction intervals are more informative than OPM's.

♣ Jittering is not cheating...

Existing model for predicting four-year mortality risk of the elderly

Table: Lee et al. JAMA 2006: Prognostic index

<i>Patient characteristic</i>	<i>Points</i>
Age (years) 60 – 64	1
65 – 69	2
70 – 74	3
75 – 79	4
80 – 84	5
≥ 85	7
Male sex	2
Diabetes	1
Cancer (not including minor skin cancer)	2
Chronic lung disease (limits activities or individual requires aided oxygen)	2
Heart failure	2
Body mass index < 25 kg/m ²	1
Current smoker	2
Functional difficulties : Bathing	2
Managing money or finances	2
Walking several blocks	2
Pulling or pushing large objects	1
Point total: 0 to 5	Predicted 4-yr mortality risk: < 4%
6 to 9	15%
10 to 13	42%
≥ 14	64%

(Adapted from Lee, et al. JAMA 2006;295:805.)

Comparison with the existing models

- TORQUE incorporates a transformation of an additive index to be more flexible.
- Method 1 allows differential effects at different quantile levels of the model, and is more robust against model mis-specification.
- ?

Future Research

- Variable selection under the proposed model;
- Confidence intervals on quantile estimation;
- Extension to multi-index models.
- Thank you for your attention!