Generalized Quantile Regression in Stata

Matthew Baker, Hunter College
David Powell, RAND
Travis Smith, University of Georgia

Stata Conference

August 1, 2014

Motivation

- Quantile regression techniques are useful in understanding the relationship between explanatory variables and the conditional distribution of the outcome variable.
- These techniques estimate conditional quantile treatment effects (QTEs).
- In conditional quantile models, the parameters of interest are assumed to vary based on a nonseparable disturbance term.
- As additional covariates are added, the interpretation of these parameters changes.
- Powell (2013a) and Powell (2013b) introduce estimators which allow the researcher to condition on additional covariates for the purposes of identification while maintaining the same structural quantile function.

Preview

- Powell (2013a) introduces a quantile panel data estimator with a nonadditive fixed effect (QRPD).
- Powell (2013b) introduces "Generalized Quantile Regression" (GQR).
    - Quantile regression (QR) and instrumental variable quantile regression (IVQR) are special cases of GQR.
- We have developed `genquantreg` to implement QRPD and GQR.

Outline

1. Conditional Quantile Estimators
2. Quantile Estimation with Panel Data
3. IVQR Framework / GQR Framework
4. GQR
5. `genquantreg`

Notation

1. D represents treatment (or policy) variables
2. X represents control variables
3. Z represents instruments
4. Y represents the outcome

Background: Quantile Estimation

- Cross-sectional Quantile Estimators (Koenker and Basset [1978], Chernozhukov and Hansen [2008]) allow parameters to vary based on a nonseparable disturbance term:

$$Y_i = D_i'\beta(U_i^*), \quad U_i^* \sim U(0,1),$$

and estimates the Structural Quantile Function (SQF)

$$S_Y(\tau|d_i) = d_i'\beta(\tau), \quad \tau \in (0,1).$$

- Interpret U^* as ability or "proneness" for the outcome variable. For reference, let's model $U^* = f(X,U)$ (where $U =$ "unobserved proneness" and $X =$ "observed proneness").
- For a given policy vector d_i, can predict distribution of Y_i.

Background: IV-QR

- Assumes all variables are treatment variables.
  - i.e., All variables that one wants to control for must be included in the quantile function itself.
- IV-QR assumes $U_i^* | Z_i \sim U(0,1)$.
- This assumption gives condition $P(Y_i \leq D_i' \beta(\tau) | Z_i) = \tau$.
- Moment condition: $E[Z_i (\mathbf{1}(Y_i \leq D_i' \beta(\tau)) - \tau)] = 0$.
- If one wants to add another variable (x_i), then must assume that $P(Y_i \leq D_i' \beta(\tau) + x_i \delta(\tau) | Z_i) = \tau$.

Motivating Example

- Consider studying the impact of job training (d_i) on the distribution of earnings (y_i). Assume that job training is randomized.
- A quantile regression of earnings on job training (`qreg y d, quan(90)`) for each quantile provides the distribution of $y_i|d_i$.
- You can interpret the result of the above quantile regression as the impact of job training on the 90^{th} quantile of the earnings distribution.
- But let's say that your data also contains a variable about each person's labor market ability (x_i) and you decide to control for that variable as well: `qreg y d x, quan(90)`
- The interpretation is different. You now have the effect of job training at the 90^{th} quantile of the distribution for a fixed level of labor market ability (i.e., people with high earnings given labor market ability).
- Some people with high earnings given their labor market ability are actually at the bottom of the earnings distribution.

Quantile Models with Fixed Effects

- Most quantile panel data estimators include an additive fixed effect: Koenker [2004], Harding and Lamarche [2009], Canay [2010], Galvao [2011], Ponomareva [2010].
- Additive fixed effect term assumes specification:

$$Y_{it} = \alpha_i + D'_{it}\beta(U_{it}), \quad U_{it} \sim U(0,1)$$

- Concern: An additive fixed effect means that we no longer have a completely nonadditive disturbance term.
- Parameters vary based only on U_{it}, not U^*_{it}.
- Assume α_i is known. Quantile models with additive fixed effects provide distribution of $Y_{it} - \alpha_i$ for a given D_{it}. Note that many observations at the top of the $Y_{it} - \alpha_i$ distribution are potentially at the bottom of the Y_{it} distribution.

Quantile Model with Nonadditive Fixed Effects (QRPD)

- Let $U_{it}^* = f(\alpha_i, U_{it}), \qquad U_{it}^* \sim U(0,1)$
-
$$Y_{it} = D_{it}'\beta(U_{it}^*), \quad U_{it}^* \sim U(0,1)$$
- SQF is same as quantile regression (QR) and instrumental variables quantile regression (IV-QR):
$$S_Y(\tau|d_{it}) = d_{it}'\beta(\tau), \quad \tau \in (0,1).$$
- Note that including an additive fixed effect term causes bias even if D_{it} randomly assigned.
- QRPD assumes $U_{it}^* \sim U(0,1)$ but makes no such assumptions on conditional distribution. Instead, it uses pairwise comparisons.

Model

Assumptions:

A1 Potential Outcomes and Monotonicity: $Y_{it} = D_{it}'\beta(U_{it}^*)$ where $D_{it}'\beta(U_{it}^*)$ is increasing in $U_{it}^* \sim U(0,1)$.

A2 Independence: $E\Big[[\mathbb{1}(U_{it}^* \leq \tau) - \mathbb{1}(U_{is}^* \leq \tau)] \,|Z_i \Big] = 0$ for all s, t.

Moment Conditions

MC1

$$E\left\{(Z_{it}-Z_{is})\left[\mathbf{1}\left(Y_{it} \leq D_{it}'\beta(\tau)\right) - \mathbf{1}\left(Y_{is} \leq D_{is}'\beta(\tau)\right)\right]\right\}$$

$$\Rightarrow E\left\{(Z_{it}-Z_{is})E\left[\mathbf{1}(U_{it}^{*} \leq \tau) - \mathbf{1}(U_{is}^{*} \leq \tau)\,|Z_i\right]\right\} = 0$$

Moment Conditions

MC1

$$E\left\{(Z_{it}-Z_{is})\left[\mathbf{1}\left(Y_{it}\leq D_{it}'\beta(\tau)\right)-\mathbf{1}\left(Y_{is}\leq D_{is}'\beta(\tau)\right)\right]\right\}$$

MC2

$$E[\mathbf{1}(Y_{it}\leq D_{it}'\beta(\tau))-\tau]=0$$

Simulation

$$
\begin{aligned}
t &\in \{0,1\} \\
\text{Fixed Effect:} \quad \alpha_i &\sim U(0,1) \\
U_{it} &\sim U(0,1) \\
\text{Total Disturbance:} \quad U^*_{it} \equiv F(\alpha_i + U_{it}) \Rightarrow U^*_{it} &\sim U(0,1) \\
\text{Year Effect:} \quad \delta_0 = 1, \delta_1 &= 2 \\
\psi_{it} &\sim U(0,1) \\
\text{Instrument:} \quad Z_{it} &= \alpha_i + \psi_{it} \\
\text{Policy Variable:} \quad D_{it} &= Z_{it} + U_{it} \\
\text{Outcome:} \quad Y_{it} &= U^*_{it}(\delta_t + D_{it}) \\
N = 500, T &= 2
\end{aligned}
$$

Simulation Results

	IVQR			IVQRFE			IVQRPD		
Quantile	Mean Bias	MAD	RMSE	Mean Bias	MAD	RMSE	Mean Bias	MAD	RMSE
5	0.56057	0.55	0.56753	0.39750	0.41	0.42170	-0.00544	0.05	0.07027
10	0.70229	0.70	0.70723	0.34740	0.36	0.37478	-0.01025	0.06	0.09861
15	0.80304	0.80	0.80664	0.29736	0.31	0.32898	-0.00941	0.08	0.11788
20	0.87783	0.88	0.88058	0.24750	0.26	0.28468	-0.01046	0.09	0.13316
25	0.93577	0.93	0.93802	0.19762	0.21	0.24270	0.00099	0.11	0.14822
30	0.98169	0.98	0.98365	0.14765	0.16	0.20403	0.00181	0.11	0.16042
35	1.01647	1.02	1.01806	0.09748	0.13	0.17123	0.00337	0.12	0.16867
40	1.04178	1.04	1.04303	0.04731	0.10	0.14851	0.00291	0.12	0.17832
45	1.06114	1.06	1.06216	-0.00259	0.09	0.14093	0.00773	0.13	0.18106
50	1.06906	1.07	1.06987	-0.05266	0.10	0.15030	0.00852	0.13	0.18329
55	1.06489	1.07	1.06563	-0.10259	0.11	0.17430	0.00442	0.13	0.18429
60	1.04540	1.05	1.04602	-0.15269	0.15	0.20768	0.00167	0.13	0.18474
65	1.00899	1.01	1.00952	-0.20252	0.19	0.24663	-0.00151	0.12	0.18685
70	0.96410	0.96	0.96461	-0.25235	0.24	0.28898	-0.00279	0.12	0.18217
75	0.91812	0.92	0.91867	-0.30238	0.29	0.33360	-0.00361	0.12	0.18069
80	0.86625	0.87	0.86687	-0.35251	0.34	0.37954	-0.00390	0.12	0.17601
85	0.79638	0.80	0.79722	-0.40264	0.39	0.42653	-0.00539	0.12	0.16687
90	0.70683	0.71	0.70813	-0.45260	0.44	0.47395	-0.00672	0.10	0.15145
95	0.58787	0.59	0.59085	-0.50250	0.49	0.52185	-0.01127	0.09	0.12454

Generalized Quantile Regression (GQR)

- Let D_i represent policy variables, X_i represent control variables, Z_i represent instruments. Let $U_i^* = f(X_i, U_i)$ be the disturbance term.
- Conditional quantile models require policy variables and control variables to be included in Structural Quantile Function and assume underlying equation is

$$Y_i = D_i'\beta(U_i) + X_i'\delta(U_i)$$

Comparison

1. Conditional Quantile (without covariates) assumptions:
   - $U_i^*|Z_i \sim U(0,1), \quad U_i^* \sim U(0,1)$
   - $P(Y_i \leq D_i'\beta(\tau)|Z_i) = \tau$
2. Conditional Quantile (with covariates) assumptions:
   - $U_i|Z_i, X_i \sim U(0,1), \quad U_i \sim U(0,1)$
   - $P(Y_i \leq D_i'\beta(\tau) + X_i'\delta(\tau)|Z_i, X_i) = \tau$
3. GQR assumptions:
   - $U_i^*|Z_i, X_i \sim U_i^*|X_i, \quad U_i^* \sim U(0,1)$
   - $P(Y_i \leq D_i'\beta(\tau)|Z_i, X_i) = P(Y_i \leq D_i'\beta(\tau)|X_i) \equiv \tau_{X_i}$
   - $E\left[\tau_{X_i}\right] = \tau$

Model

Assumptions:

A1 Potential Outcomes and Monotonicity:
$Y_i = D_i' \beta(U_i^*)$ where $D_i' \beta(U_i^*)$ is increasing in $U_i^* \sim U(0,1)$.

A2 Conditional Independence:

(a) $P(U_i^* \leq \tau | Z_i, X_i) = P(U_i^* \leq \tau | X_i)$.
(b) $E[Z_i(\widehat{\tau}_{X_i} - \tau_{X_i})] = 0$.

Moment Conditions

MC1

$$E\left\{ Z_i \left[\mathbf{1}\left(Y_i \leq D_i' \beta(\tau) \right) - \hat{\tau}_{X_i} \right] \right\} = 0$$

MC2

$$E[\mathbf{1}(Y_i \leq D_i' \beta(\tau)) - \tau] = 0$$

Estimation

- Use both moment conditions.
- Estimation simplifies if confine set of possible coefficients to

$$\mathcal{B} \equiv \left\{ b \quad | \quad \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(Y_i \leq D_i' b\right) = \tau \right\}.$$

- For a given b, estimate

$$\hat{\tau}_{X_i}(b) = \widehat{P}\left(Y_i \leq D_i' b | X_i\right).$$

- Estimation uses GMM with

$$g_i(b) = Z_i \left[\mathbf{1}\left(Y_i \leq D_i' b\right) - \hat{\tau}_{X_i}(b) \right],$$

$$\widehat{\beta(\tau)} = \arg\min_{b \in \mathcal{B}} \hat{g}(b)' \hat{A} \hat{g}(b)$$

Simulation

$$\begin{aligned}
\text{Observed Skill:} \quad & X_i \sim U(0,1), \\
\text{Unobserved Skill:} \quad & U_i \sim U(0,0.1), \\
\text{Total Disturbance:} \quad & U_i^* \equiv F_{X_i+U_i}(X_i + U_i) \Rightarrow U_i^* \sim U(0,1), \\
\text{Policy Variable:} \quad & D_i \sim U(0,1), \\
\text{Outcome:} \quad & Y_i = U_i^*(1 + D_i).
\end{aligned}$$

Simulation Results

Table: Simulation Results: Policy Variable Randomly-Assigned

	QR (conditional)			QR (unconditional)		
Quantile	Mean Bias	MAD	RMSE	Mean Bias	MAD	RMSE
5	0.40555	0.40555	0.41231	0.00120	0.04159	0.05007
10	0.37051	0.37051	0.37440	0.00166	0.05665	0.06928
15	0.32667	0.32667	0.32933	0.00436	0.06725	0.08252
20	0.27998	0.27998	0.28215	0.00305	0.07552	0.09295
25	0.23430	0.23430	0.23605	0.00272	0.08028	0.09881
30	0.18773	0.18773	0.18938	0.00303	0.08514	0.10510
35	0.14101	0.14101	0.14283	0.00406	0.08675	0.10875
40	0.09373	0.09373	0.09606	0.00408	0.09021	0.11247
45	0.04702	0.04722	0.05121	0.00281	0.09162	0.11369
50	0.00005	0.01655	0.02059	0.00446	0.09243	0.11460
55	-0.04703	0.04722	0.05127	0.00364	0.09156	0.11297
60	-0.09393	0.09393	0.09614	0.00374	0.09027	0.11148
65	-0.14057	0.14057	0.14241	0.00400	0.08791	0.10953
70	-0.18723	0.18723	0.18893	0.00371	0.08493	0.10515
75	-0.23423	0.23423	0.23604	0.00035	0.07901	0.09851
80	-0.28087	0.28087	0.28305	-0.00055	0.07203	0.08940
85	-0.32529	0.32529	0.32809	-0.00029	0.06454	0.07939
90	-0.36743	0.36743	0.37142	0.00040	0.05400	0.06637
95	-0.40129	0.40129	0.40843	0.00044	0.04085	0.04906

Results based on 1000 replications, N=500. MAD is Mean Absolute Deviation, RMSE is Root Mean Squared Error.

Simulation Results

Table: Simulation Results: Policy Variable Randomly-Assigned

	GQR (logit)			GQR (probit)		
Quantile	Mean Bias	MAD	RMSE	Mean Bias	MAD	RMSE
5	-0.00121	0.02397	0.02954	-0.00330	0.02396	0.02974
10	-0.00025	0.02491	0.03037	-0.00006	0.02528	0.03110
15	0.00056	0.02582	0.03144	0.00038	0.02556	0.03094
20	0.00083	0.02599	0.03175	0.00088	0.02602	0.03157
25	-0.00015	0.02517	0.03057	0.00053	0.02523	0.03068
30	0.00061	0.02455	0.03013	0.00006	0.02540	0.03125
35	0.00098	0.02540	0.03129	0.00003	0.02609	0.03199
40	0.00062	0.02560	0.03151	0.00013	0.02547	0.03135
45	-0.00016	0.02508	0.03052	-0.00104	0.02512	0.03086
50	0.00103	0.02437	0.03006	0.00073	0.02539	0.03135
55	0.00033	0.02561	0.03077	0.00030	0.02542	0.03068
60	-0.00010	0.02588	0.03144	-0.00067	0.02581	0.03133
65	-0.00033	0.02515	0.03054	-0.00022	0.02502	0.03012
70	0.00117	0.02521	0.03125	0.00083	0.02509	0.03126
75	-0.00004	0.02374	0.02941	-0.00011	0.02435	0.02999
80	-0.00037	0.02469	0.03000	0.00039	0.02515	0.03080
85	0.00066	0.02580	0.03136	0.00042	0.02564	0.03103
90	-0.00015	0.02475	0.03081	0.00050	0.02454	0.03056
95	-0.00304	0.02520	0.03081	-0.00128	0.02460	0.03012

Results based on 1000 replications, N=500. MAD is Mean Absolute Deviation, RMSE is Root Mean Squared Error.

Simulation II

$$
\begin{aligned}
\text{Observed Skill:} \quad & X_i \sim U(0,1), \\
\text{Unobserved Skill:} \quad & U_i \sim U(0,0.1), \\
\text{Total Disturbance:} \quad & U_i^* \equiv F_{X_i+U_i}(X_i+U_i) \Rightarrow U_i^* \sim U(0,1), \\
& \psi_i \sim U(0,1), \\
\text{Policy Variable:} \quad & D_i = X_i + \psi_i, \\
\text{Outcome:} \quad & Y_i = U_i^*(1+D_i).
\end{aligned}
$$

Simulation Results

Table: Simulation Results

	QR (conditional)			QR (unconditional)		
Quantile	Mean Bias	MAD	RMSE	Mean Bias	MAD	RMSE
5	0.40938	0.40938	0.41082	1.09017	1.09017	1.10298
10	0.36755	0.36755	0.36862	1.10850	1.10850	1.11502
15	0.32340	0.32340	0.32447	1.10497	1.10497	1.10947
20	0.27840	0.27840	0.27968	1.09717	1.09717	1.10016
25	0.23338	0.23338	0.23488	1.08256	1.08256	1.08475
30	0.18708	0.18708	0.18894	1.06411	1.06411	1.06602
35	0.14144	0.14144	0.14384	1.04266	1.04266	1.04421
40	0.09507	0.09509	0.09859	1.02061	1.02061	1.02188
45	0.04898	0.04976	0.05602	0.99611	0.99611	0.99717
50	0.00160	0.02305	0.02887	0.96943	0.96943	0.97033
55	-0.04627	0.04854	0.05625	0.94097	0.94097	0.94172
60	-0.09483	0.09488	0.10114	0.91168	0.91168	0.91231
65	-0.14223	0.14223	0.14768	0.88244	0.88244	0.88296
70	-0.19153	0.19153	0.19658	0.85331	0.85331	0.85374
75	-0.24021	0.24021	0.24560	0.82262	0.82262	0.82302
80	-0.28817	0.28817	0.29378	0.79468	0.79468	0.79509
85	-0.33448	0.33448	0.34085	0.77663	0.77663	0.77720
90	-0.38074	0.38074	0.38907	0.77256	0.77256	0.77332
95	-0.41806	0.41806	0.43170	0.78999	0.78999	0.79197

Results based on 1000 replications, N=500. MAD is Mean Absolute Deviation, RMSE is Root Mean Squared Error.

Simulation Results

Table: Simulation Results

	GQR (logit)			GQR (probit)		
Quantile	Mean Bias	MAD	RMSE	Mean Bias	MAD	RMSE
5	-0.00252	0.02536	0.03149	-0.00497	0.02515	0.03100
10	-0.00076	0.02654	0.03254	-0.00146	0.02692	0.03319
15	-0.00004	0.02786	0.03381	-0.00024	0.02804	0.03394
20	-0.00024	0.02952	0.03567	-0.00120	0.03078	0.03743
25	-0.00176	0.02968	0.03593	-0.00176	0.03052	0.03673
30	-0.00133	0.03011	0.03704	-0.00213	0.03153	0.03850
35	-0.00066	0.03214	0.03936	-0.00249	0.03353	0.04061
40	-0.00121	0.03373	0.04080	-0.00191	0.03441	0.04192
45	-0.00165	0.03315	0.03993	-0.00328	0.03530	0.04291
50	-0.00106	0.03364	0.04128	-0.00173	0.03491	0.04311
55	-0.00187	0.03605	0.04326	-0.00334	0.03758	0.04585
60	-0.00172	0.03692	0.04474	-0.00385	0.04015	0.04838
65	-0.00222	0.03786	0.04525	-0.00405	0.03909	0.04759
70	-0.00106	0.03842	0.04694	-0.00158	0.04120	0.05038
75	-0.00323	0.03711	0.04487	-0.00415	0.04105	0.05008
80	-0.00243	0.03957	0.04822	-0.00267	0.04241	0.05250
85	-0.00054	0.04130	0.04976	-0.00294	0.04458	0.05452
90	-0.00483	0.04063	0.05009	-0.00290	0.04482	0.05560
95	-0.01141	0.04293	0.05195	-0.00331	0.04461	0.05429

Results based on 1000 replications, N=500. MAD is Mean Absolute Deviation, RMSE is Root Mean Squared Error.

genquantreg

- Implements QRPD and GQR.
- Documentation and Package forthcoming.

genquantreg syntax

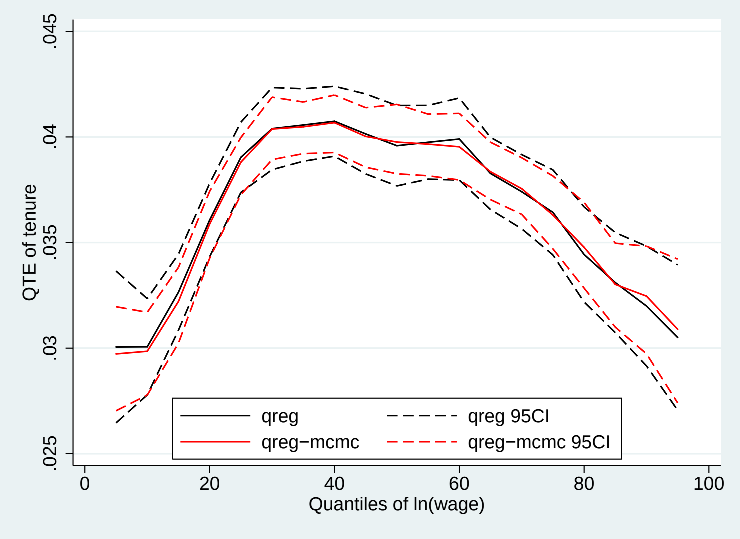
genquantreg *varlist* [if] [in] [, – variable list should include dependent variable + treatment variables

- PRONEness(*varlist*) – control variables
- INSTRuments(*varlist*) – instruments
- FIX(*varname*) – implements QRPD, fixed effects based on given variable
- TECHnique(*string*) – probit, logit or linear
- TAU(*real 50*) – quantile

genquantreg details

- User specifies instruments, which are same as treatment variables when they are conditionally exogenous.
- Optimization builds on amcmc() wrapper developed by Matt Baker.
- If no variables included in PRONEness and FIX not specified, estimator is QR or IV-QR.

Comparing genquantreg to qreg



Conclusion / Next Steps

- GQR and QRPD generalize traditional quantile estimators.
- `genquantreg` provides a flexible way to estimate quantile treatment effects.
- Documentation and package forthcoming.