4B: Quantile regression

© A. Colin Cameron
U. of Calif. - Davis

.

OeNB Summer School 2010
Microeconometrics
Central Bank of Austria

Based on
A. Colin Cameron and Pravin K. Trivedi,
Microeconometrics: Methods and Applications (MMA), ch.4.6.
Microeconometrics using Stata (MUS), ch.7.
Data examples are from MUS.

Aug 30 - Sept 3, 2010

1. Introduction

- Rather than model $E[y|\mathbf{x}]$ we model $\text{Median}[y|\mathbf{x}]$ or more generally $\text{Quantile}_q[y|\mathbf{x}]$.
- Quantile regression is easy to implement
  - In Stata replace `regress y x` with `qreg y x, quantile()`
  - This uses a linear model to approximate the conditional quantile
- Challenge is to understand what is going on.
  - Begin with comparing two distributions (treatment and control groups) before one or more regressors.
  - Emphasize difference between conditional quantiles and unconditional quantiles.

Outline

2. Comparison of two distributions

- Question: How do we compare medical expenditures with and without supplementary insurance?
- Data are from Cameron and Trivedi (2009, ch.7).
- MEPS 2003 65 and older who are in Medicare
- $y = \texttt{ltotexp} = \log(\text{total medical expenditure in 2003})$
  - $N = 2955$ after drop 109 with zero expenditure
- $d = \texttt{suppins} = 1$ if have supplementary medical insurance
  - 58% have supplementary insurance
  - may cover pharmaceutical drugs (not covered by Medicare)
  - may cover copays and coinsurance under regular Medicare

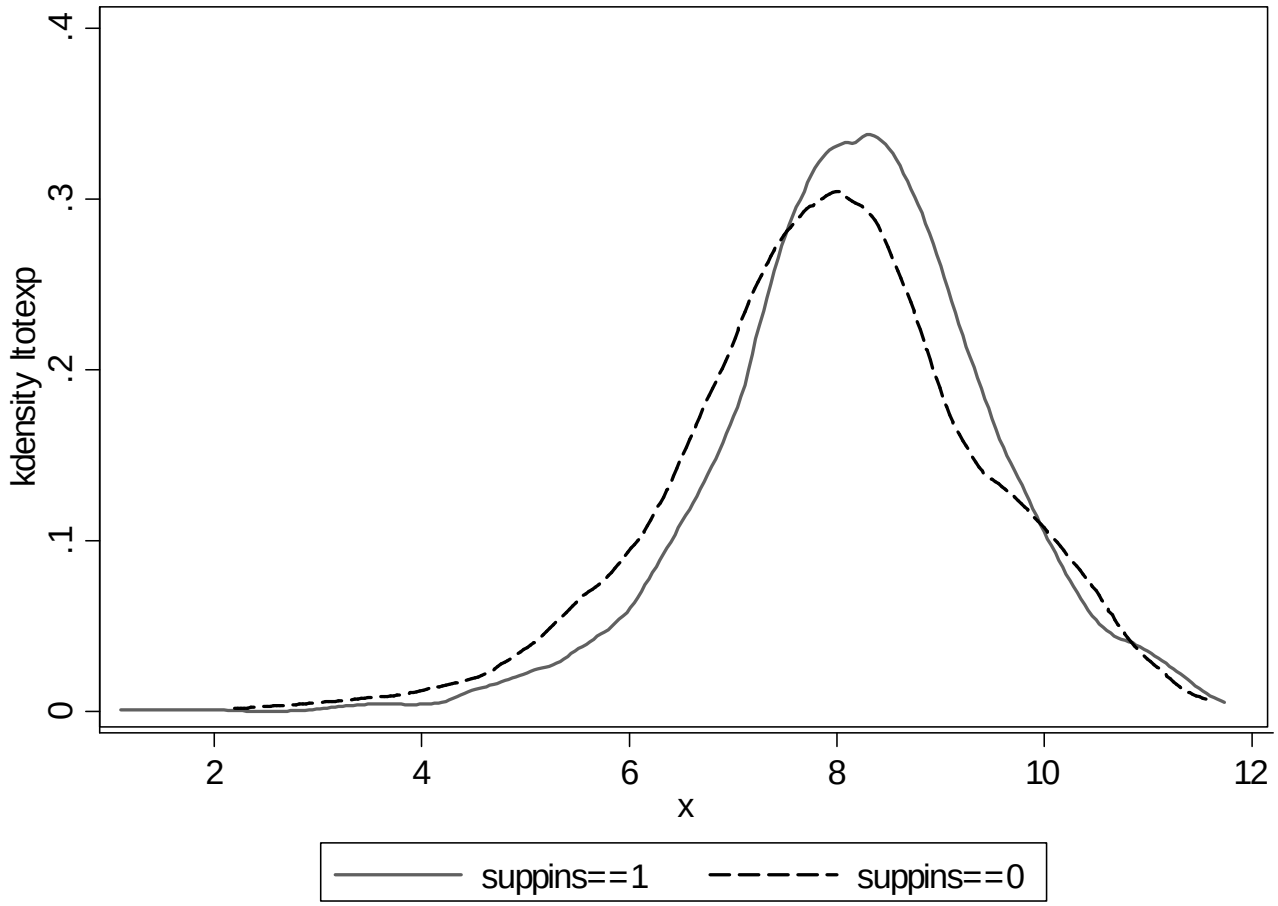
Compare summary statistics

- Sample means are substantially higher with supplementary insurance.
- Standard deviations are higher in levels but not logs.

		$\texttt{Suppins} = 1$	$\texttt{Suppins} = 0$	
Means	Levels	7470	6420	+16%
	Logs	8.17	7.91	+26%
St.Devs.	Levels	12300	11200	+10%
	Logs	1.30	1.45	−15%

- But where is action: High expenditures? Low expenditures.

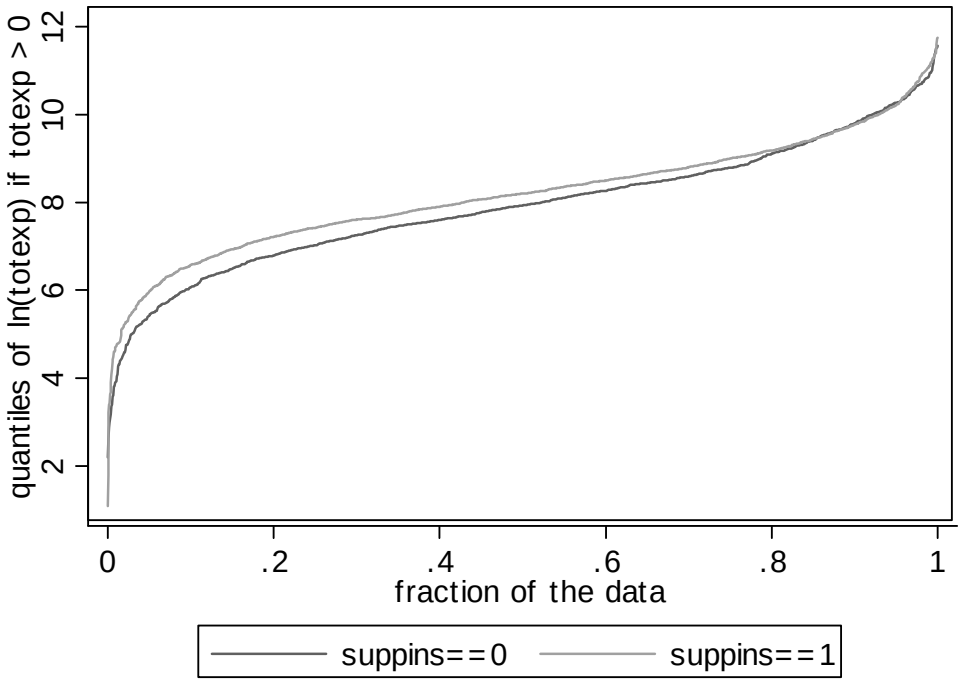
Compare densities



It looks like more action at lower levels of expenditures

Compare quantile plots (inverse cdfs)

- NOTE: This is cdf with the axes reversed



- Biggest difference is around the .1 quantile (the 10th percentile)
  - around 0.5 (a lot on log scale)
- By the .9 quantile (the 90th percentile) there is little difference.

Compare percentiles / quantiles

- Obtain percentiles of the upper curve in previous slide

```
. centile ltotexp if suppins==1, centile(10 50 90)

  Variable   Obs   Percentile   Centile   [95% Conf. Interval]
   ltotexp  1748           10   6.571299    6.457681, 6.673633
                           50   8.202071    8.146281, 8.258152
                           90   9.771977    9.665329, 9.886245
```

- Obtain percentiles of the lower curve in previous slide

```
. centile ltotexp if suppins==0, centile(10 50 90)

  Variable   Obs   Percentile   Centile   [95% Conf. Interval]
   ltotexp  1748           10   6.056784    5.880274, 6.27851
                           50   7.929846    7.843799, 8.019941
                           90   9.796142     9.65716, 9.96981
```

- The difference between the two:

Variable	Obs	Percentile	Difference			
ltotexp	1748	10	6.571299	−	6.056784	= .514515
		50	8.202071	−	7.929846	= .272225
		90	9.771977	−	9.796142	= −.024165

- Same as from eyeballing the differences in the two cdf's.
- Wouldn't it be nice to have standard errors / confidence intervals for these differences?

3. Quantile regression on a single indicator variable

- Quantile regress y on an intercept gives percentiles of y.

```
. qreg ltotexp if suppins==1, q(.1)
Iteration  1:  WLS sum of weighted deviations =  1594.6984

Iteration  1: sum of abs. weighted deviations =  1595.9127
Iteration  2: sum of abs. weighted deviations =  857.32593

.1 Quantile regression                                  Number of obs =      1748
  Raw sum of deviations 857.3417 (about 6.5624442)
  Min sum of deviations 857.3259                        Pseudo R2     =    0.0000
```

ltotexp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons	6.572282	.0616965	106.53	0.000	6.451276	6.693289

- Almost same as `centile ltotexp if suppins==1, centile(10)`

- Quantile regress y on an intercept and d gives difference across d of percentiles of y
  - so reproduces earlier

```
. qreg ltotexp suppins, q(.1)
Iteration  1:  WLS sum of weighted deviations =  2813.7766

Iteration  1: sum of abs. weighted deviations =  2814.6925
Iteration  2: sum of abs. weighted deviations =  2076.1058
Iteration  3: sum of abs. weighted deviations =  1516.3577

.1 Quantile regression                          Number of obs =      2955
  Raw sum of deviations 1534.441 (about 6.3613024)
  Min sum of deviations 1516.358                Pseudo R2     =    0.0118
```

ltotexp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	.5154982	.1052626	4.90	0.000	.3091027	.7218936
_cons	6.056784	.080816	74.95	0.000	5.898323	6.215246

- Here 6.572282 (previous slide) - 6.056784 (not given) = .515498.
- And the constant is the 10th percentile when suppins=0.

- Simultaneously estimate several quantile differences (sqreg)
  - with heteroskedastic robust standard errors:

```
. sqreg ltotexp suppins, q(.1 .5 .9) reps(400) nodots

Simultaneous quantile regression                Number of obs =      2955
  bootstrap(400) SEs                            .10 Pseudo R2 =    0.0118
                                                .50 Pseudo R2 =    0.0058
                                                .90 Pseudo R2 =    0.0000
```

ltotexp		Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10							
	suppins	.5154982	.1111779	4.64	0.000	.297504	.7334923
	_cons	6.056784	.1002135	60.44	0.000	5.860289	6.25328
q50							
	suppins	.2715392	.0512858	5.29	0.000	.1709796	.3720988
	_cons	7.929846	.0414162	191.47	0.000	7.848639	8.011054
q90							
	suppins	-.0227232	.0948319	-0.24	0.811	-.2086665	.1632201
	_cons	9.794621	.0787969	124.30	0.000	9.640118	9.949123

- Difference is .51 at 10th percentiles and -.02 at 90th percentiles.
- Also note more precision in estimation at the median (center of data).

4. Quantile regression on a single continuous variable

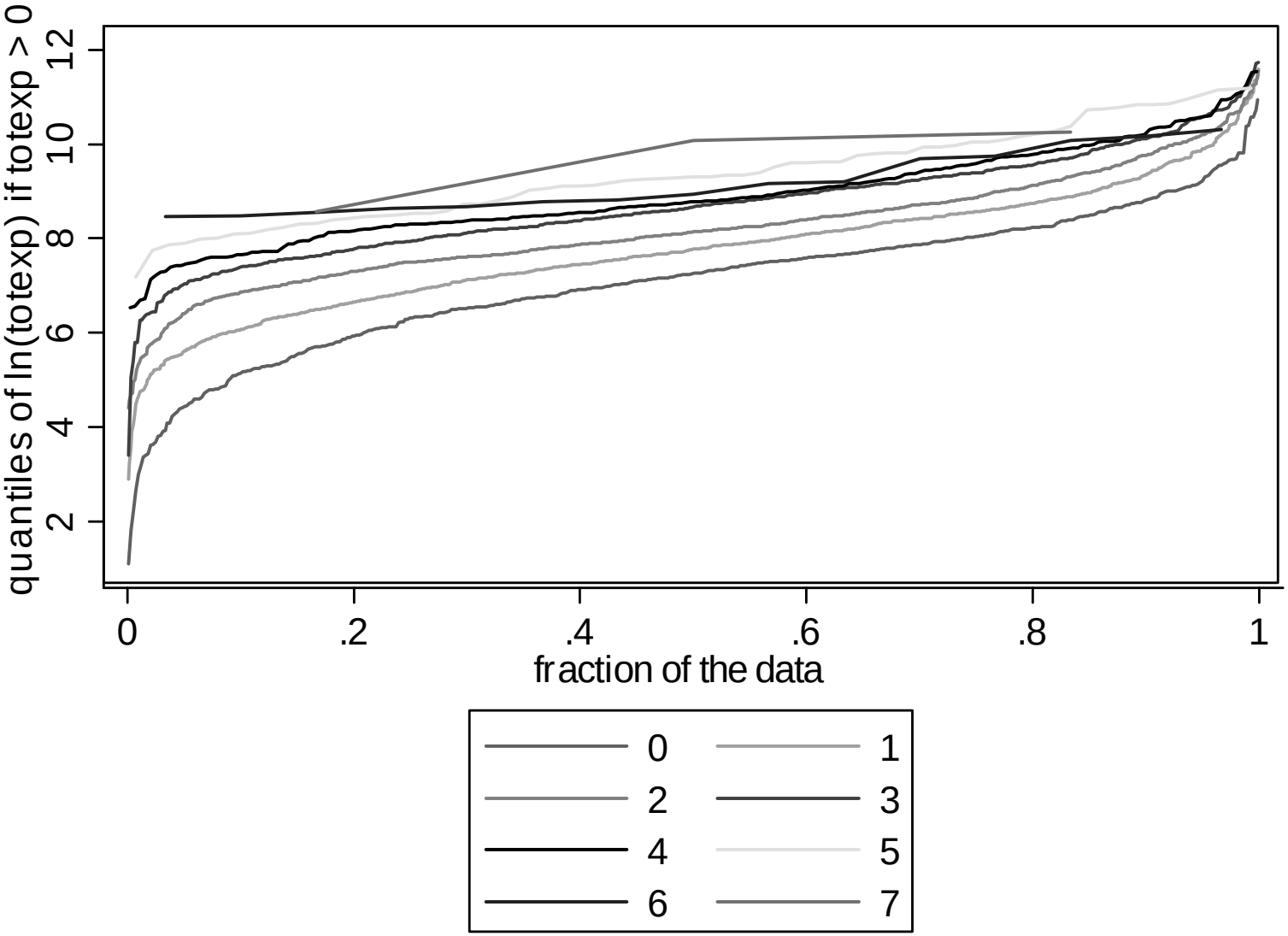
- Regressor `totchr` = number of chronic conditions
  - For the moment ignore `suppins`

```
. tabulate totchr
```

# of chronic problems	Freq.	Percent	Cum.
0	466	15.77	15.77
1	865	29.27	45.04
2	809	27.38	72.42
3	506	17.12	89.54
4	222	7.51	97.06
5	69	2.34	99.39
6	15	0.51	99.90
7	3	0.10	100.00
Total	2,955	100.00	

- Not strictly continuous, but not binary.

Inverse cdfs (quantile plot) for each value of totchr



Percentiles of y for each value of the regressor

- `bysort totchr: centile ltotexp, centile(10)`

totchr	N	10th percentile	Change	Cum change
0	466	5.142216		
1	865	6.065638	0.92	0.92
2	809	6.863803	0.80	1.72
3	506	7.378726	0.52	2.24
4	222	7.654755	0.28	2.52
5	69	8.094684	0.44	2.95
6	15	8.472199	0.38	3.33
7	3	8.565602	0.09	3.42

Using qreg with full set of dummies

- Following gives same results

```
. quietly tabulate totchr, generate(dtotchr)

. drop dtotchr1

. qreg ltotexp dtotchr*, q(.1) nolog

.1 Quantile regression                          Number of obs =      2955
  Raw sum of deviations 1534.441 (about 6.3613024)
  Min sum of deviations 1282.308                Pseudo R2     =    0.1643
```

ltotexp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dtotchr2	.9209313	.0963249	9.56	0.000	.7320604	1.109802
dtotchr3	1.716309	.0974044	17.62	0.000	1.525322	1.907297
dtotchr4	2.236495	.1075276	20.80	0.000	2.025658	2.447332
dtotchr5	2.516852	.1366955	18.41	0.000	2.248824	2.784881
dtotchr6	2.947189	.2056589	14.33	0.000	2.54394	3.350439
dtotchr7	3.335108	.3710754	8.99	0.000	2.607515	4.062701
dtotchr8	3.418108	.2757883	12.39	0.000	2.877351	3.958865
_cons	5.147494	.0775648	66.36	0.000	4.995408	5.299581

Conditional quantile specified to be linear in x

- Specify the 10th conditional quantile to be linear in x

```
. qreg ltotexp totchr, q(.1) nolog

.1 Quantile regression                                  Number of obs =      2955
  Raw sum of deviations 1534.441 (about 6.3613024)
  Min sum of deviations 1306.327                        Pseudo R2     =    0.1487

------------------------------------------------------------------------------
     ltotexp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     totchr |   .5674899   .0374614    15.15   0.000     .4940367    .6409431
      _cons |   5.500936   .0788182    69.79   0.000     5.346392     5.65548
------------------------------------------------------------------------------
```

- The 10th conditional quantile of ltotexp increases by 0.57 with each extra chronic condition - enormous effect.
  - A crude estimate from earlier analysis compares .1 quantile at totchr=7 with that at =0: get $(8.56 - 5.15)/(7 - 0) = 0.49$.

- Less effect at higher quantiles but still big.

```
. sqreg ltotexp totchr, q(.1 .5 .9) reps(100) nodots

Simultaneous quantile regression                Number of obs =      2955
  bootstrap(100) SEs                            .10 Pseudo R2 =    0.1487
                                                .50 Pseudo R2 =    0.0903
                                                .90 Pseudo R2 =    0.0646
```

ltotexp	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10						
totchr	.5674899	.0322534	17.59	0.000	.5042484	.6307313
_cons	5.500936	.0801445	68.64	0.000	5.343791	5.65808
q50						
totchr	.3932115	.0183942	21.38	0.000	.3571448	.4292782
_cons	7.347944	.0496718	147.93	0.000	7.250549	7.445339
q90						
totchr	.3762154	.0247912	15.18	0.000	.3276056	.4248252
_cons	8.956737	.0762571	117.45	0.000	8.807215	9.10626

5. Multivariate conditional quantile regression

- Now both regressors: similar results to when each in isolation
  - this is surprising

```
. sqreg ltotexp suppins totchr, q(.1 .5 .9) reps(100) nodots

Simultaneous quantile regression                Number of obs =      2955
  bootstrap(100) SEs                            .10 Pseudo R2 =    0.1588
                                                .50 Pseudo R2 =    0.0956
                                                .90 Pseudo R2 =    0.0646
```

ltotexp	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10						
suppins	.3785672	.0703368	5.38	0.000	.2406531	.5164814
totchr	.5533481	.0227236	24.35	0.000	.5087924	.5979037
_cons	5.335844	.0782309	68.21	0.000	5.182451	5.489237
q50						
suppins	.2595026	.0584407	4.44	0.000	.1449139	.3740912
totchr	.3957717	.0162626	24.34	0.000	.3638846	.4276588
_cons	7.18299	.0606674	118.40	0.000	7.064035	7.301944
q90						
suppins	.035327	.0825533	0.43	0.669	-.1265408	.1971948
totchr	.3792658	.0217453	17.44	0.000	.3366283	.4219032
_cons	8.928353	.092908	96.10	0.000	8.746182	9.110524

- With many regressors still similar results
  - perhaps due to insignificant / low R^2
  - $R^2 = 1 -$ (sum deviations about est. quantile) / (sum deviations about raw quantile)

```
. sqreg ltotexp suppins totchr age female white, q(.1 .5 .9) reps(100) nodots

Simultaneous quantile regression                    Number of obs =      2955
  bootstrap(100) SEs                                .10 Pseudo R2 =    0.1640
                                                    .50 Pseudo R2 =    0.1009
                                                    .90 Pseudo R2 =    0.0687
```

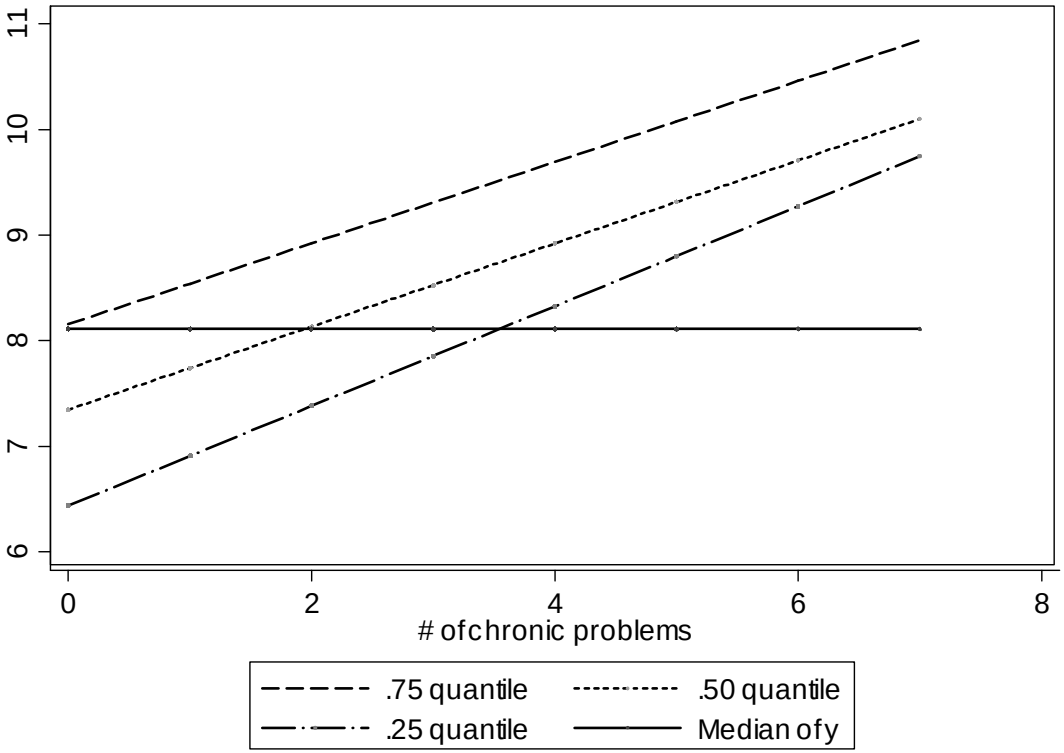
ltotexp	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q10						
suppins	.3957205	.0689863	5.74	0.000	.2604543	.5309866
totchr	.5391863	.0256054	21.06	0.000	.4889801	.5893926
age	.0192688	.0040498	4.76	0.000	.0113281	.0272095
female	-.0127282	.0723526	-0.18	0.860	-.1545949	.1291384
white	.0734392	.1680173	0.44	0.662	-.2560038	.4028823
_cons	3.867043	.3454958	11.19	0.000	3.189606	4.54448
q50						
suppins	.2769771	.0519143	5.34	0.000	.1751852	.3787689
totchr	.3942664	.0173185	22.77	0.000	.3603088	.4282239
age	.0148666	.0046507	3.20	0.001	.0057476	.0239855
female	-.0880967	.0470521	-1.87	0.061	-.1803551	.0041617
white	.4987456	.2050219	2.43	0.015	.0967452	.9007461
_cons	5.648891	.4273174	13.22	0.000	4.81102	6.486761
q90						
suppins	-.0142829	.0903319	-0.16	0.874	-.1914029	.1628371
totchr	.3579523	.0324396	11.03	0.000	.2943457	.421559
age	.0059236	.0066611	0.89	0.374	-.0071374	.0189846
female	-.1576334	.0919006	-1.72	0.086	-.3378292	.0225624
white	.3052239	.2255461	1.35	0.176	-.1370198	.7474676
_cons	8.32264	.4997803	16.65	0.000	7.342687	9.302594

Interpretation of conditional quantile regression

- q10 coefficient of `suppins`: Holding the number of chronic conditions, age, gender and race constant, if we compare people with and without supplementary health insurance, the 10^{th} percentile of `ltotexp` is 0.396 higher for those with supplementary health insurance.
- This is within variation (conditional on **x**)
  - it is **not** saying that if we were to give everyone in the sample supplementary health insurance and compare that to the case where no-one had supplementary health insurance then the 10^{th} percentile of `ltotexp` is 0.396 higher.
- It is not an unconditional quantile effect.

Unconditional quantile regression

- The following diagram is from conditional quantile regression of y (`ltotexp`) on x (`totchr`) at quantiles 0.25, 0.50, 0.75.
- If we wanted the effect of change x on the unconditional median of y this is complicated: weighted sum of change x at low values of x at .75 quantile, change x at high values of x at .25 quantile, etc.



Theory: quantiles defined

- Quantiles and percentiles are synonymous
  - the .99 quantile is the 99^{th} percentile.
- The median, the middle value of a set of ranked data, is the best-known specific quantile.
  - The sample median is an estimator of the population median.
- Let $F(y) = \Pr[Y \leq y]$ define the cumulative distribution function. Then $F(y_{med}) = 0.5$ has solution the median $y_{med} = F^{-1}(0.5)$.
- The q^{th} quantile of y, $q \in (0, 1)$, is that value of y that splits the data into proportions q below and $1 - q$ above.
  - So $F(y_q) = q$ and $y_q = F^{-1}(q)$.
  - If $y_{.99} = 200$ then $\Pr[Y \leq 200] = 0.99$.
- The median y_{med} minimizes $\sum_i |y_i - y_{med}|$

Theory: Conditional quantiles

- The q^{th} quantile α_q minimizes $\sum_{i:y_i \geq \boldsymbol{\alpha}_q}^{N} q|y_i - \alpha_q| + \sum_{i:y_i < \boldsymbol{\alpha}_q}^{N} (1-q)|y_i - \alpha_q|$
- Now introduce regressors.
- Define the conditional quantile regression function, $Q_q(y|\mathbf{x})$, where the conditional quantile is taken to be linear in $\mathbf{x}$.
- The q^{th} quantile regression estimator $\widehat{\boldsymbol{\beta}}_q$ minimizes over $\boldsymbol{\beta}_q$

$$Q(\boldsymbol{\beta}_q) = \sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}_q}^{N} q|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}_q}^{N} (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q|.$$

  - This is a linear programming problem.
- The special case $q = .5$ is least absolute deviations (LAD) or median regression where we minimize $Q(\boldsymbol{\beta}_{.5}) = \sum_{i}^{N} |y_i - \mathbf{x}_i'\boldsymbol{\beta}_{.5}|$.

Theory: Asymptotic distribution

- The asymptotic distribution is

$$\widehat{\boldsymbol{\beta}}_q \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}_q, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}]$$

where

$$\begin{aligned} \mathbf{A} &= \textstyle\sum_i q(1-q)\mathbf{x}_i\mathbf{x}_i' \\ \mathbf{B} &= \textstyle\sum_i f_{u_q}(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i' \end{aligned}$$

and $f_{u_q}(0|\mathbf{x})$ is the conditional density of $u_q = y - \mathbf{x}'\boldsymbol{\beta}_q$ evaluated at $u_q = 0$.

- As expected, precision drops as q moves away from 0.5.
- It is simpler to use a paired bootstrap as $f_{u_q}(0|\mathbf{x}_i)$ is awkward to estimate.
- Wooldridge (2010, pp.454-457) presents the asymptotic theory.

- Quantile regression originally proposed as a diagnostic (Koenker and Bassett (1978))
  - with errors heteroskedastic (variance increasing in x) the quantile regression lines fan out.
- Quantile regression is used to look at $\boldsymbol{\beta}$ (marginal effects) at different points of the distribution of y (Buchinsky (1994))
  - But note: it is the conditional distribution controlling for all **x** (or within **x** variation)
  - e.g. For $q = .1$ look at effect of schooling on earnings not at low levels of earnings (or of schooling), but at low levels of earnings conditional on schooling, age, gender, race, ...
- Quantile regression is particularly useful for censoring (Powell 1984))
  - Suppose negative values of y are recorded as 0.
  - Then we cannot estimate $\mathrm{E}[y]$
  - But we can estimate $\mathrm{Median}[y]$ if less than one-half are censored.
  - This carries over to censored regression: censored LAD.

Intuition: First-order conditions

- Define the indicator function

$$\mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \begin{cases} 1 & y_i - \mathbf{x}_i'\boldsymbol{\beta} \geq 0 \\ 0 & y_i - \mathbf{x}_i'\boldsymbol{\beta} < 0 \end{cases}$$

- Then writing $\boldsymbol{\beta}_q$ more simply as $\boldsymbol{\beta}$

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}}^{N} q|y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}}^{N} (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}| \\ &= \sum_{i=1}^{N} \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})] \times q \times (y_i - \mathbf{x}_i'\boldsymbol{\beta}) \\ &\quad + \sum_{i=1}^{N} [1 - \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})] \times (1-q) \times \{-(y_i - \mathbf{x}_i'\boldsymbol{\beta})\} \\ &= \sum_{i=1}^{N} [q - 1 + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})](y_i - \mathbf{x}_i'\boldsymbol{\beta}). \end{aligned}$$

- F.o.c. (ignore derivative of $\mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})$ as contribution negligible)

$$\partial Q(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = -\sum_{i=1}^{N} [q - 1 + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i = \mathbf{0}.$$

- Now verify that with only regressor the intercept we get the usual sample quantile.
- With $\mathbf{x}_i'\boldsymbol{\beta} = \beta$ we have f.o.c.

$$\begin{aligned} & \textstyle\sum_{i=1}^{N}[q - 1 + \mathbf{1}(y_i - \beta_q) = 0 \\ \Rightarrow \quad & \textstyle\sum_{i=1}^{N}\mathbf{1}(y_i - \beta_q) = (1-q)/N \end{aligned}$$

- The q^{th} quantile estimate β_q sets the fraction of observations with $y_i > \beta_q$ to $(1-q)/N$.
- So q/N of the y_i are less than β_q!

Intuition: The B matrix

- Given $Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} q_i(\boldsymbol{\beta})$ and $\partial q_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = -[q - 1 + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i$

$$\begin{aligned} \mathbf{B} &= \mathrm{E}\left[\sum_{i=1}^{N} \frac{\partial q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right] \\ &= \mathrm{E}\left[\sum_{i=1}^{N} [q - 1 + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]^2 \mathbf{x}_i\mathbf{x}_i'\right] \\ &= \sum_{i=1}^{N} q(1-q)\mathbf{x}_i\mathbf{x}_i' \end{aligned}$$

- Intuition: From the f.o.c. $\sum_{i=1}^{N} \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = N(1-q)$ so

$$\begin{aligned} & \sum_{i=1}^{N} [q - 1 + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]^2 \\ = & \sum_{i=1}^{N} (q-1)^2 + 2(q-1)\mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \\ = & N(q-1)^2 + 2(q-1)N(1-q) + N(1-q) \\ = & Nq(1-q) \text{ upon simplification} \end{aligned}$$

Intuition: The A Matrix

- The indicator function $\mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})$
  - changes sign only if $y_i - \mathbf{x}_i'\boldsymbol{\beta} = 0$
  - with derivative 1 and
  - probability $f_{y_i - \mathbf{x}_i'\boldsymbol{\beta}}(0|\mathbf{x}_i)$, the conditional density of $y_i - \mathbf{x}_i'\boldsymbol{\beta}$ evaluated at 0.
- So

$$
\begin{aligned}
\mathbf{A} &= \mathrm{E}\left[\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \frac{\partial q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right] \\
&= -\mathrm{E}\left[\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} [q - 1 + \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i'\right] \\
&= -\sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})]\mathbf{x}_i' \\
&= -\sum_{i=1}^{N} f_{y_i - \mathbf{x}_i'\boldsymbol{\beta}}(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'.
\end{aligned}
$$

6. Further details

- Topics include
  - Linearity
  - Misspecification
  - Two-stage least absolute deviations
  - Count data
  - Unconditional quantile regression

Linearity

- The linearity of the conditional quantile is a restriction / approximation.
- It is not a restriction if the model is fully saturated, meaning a full set of interactive dummies for all values taken by the regressors.
  - For the earlier example this would be $8 \times 2 = 16$ indicator variables as 8 values of `totchr` and two of `suppins`.
  - Not possible with continuous regressors taking many values.
- Quantiles without regression are "equivariant to monotone transformations"
  - e.g. .9 quantile of y is log of the .9 quantile of $\exp(y)$
  - whereas $\mathsf{E}[y] \neq \ln(\mathsf{E}[\exp(y)])$.
- This carries over to quantile regression in the special case that the model is fully saturated.

Misspecification

- What if the conditional quantile function is nonlinear?
- Recall that OLS gives the minimum mean-squared error linear approximation to the conditional expectation function even when the linear model is misspecified.
- Qualitatively similarly, linear quantile regression minimizes a (weighted) mean-squared error loss function for specification error in the quantile function.
  - Angrist, Chernozhukov, and Fernandez-Val (2006).

Two-stage LAD

- Two-stage least absolute deviations and quantile regression
  - Two-stage least absolute deviations is absolute error loss analog of IV estimation with squared error loss
    - ★ Amemiya (1981), Powell (1983).
  - More generally can estimate quantile effects under endogeneity
    - ★ Portney and Chen (1986)
- However, this assumes a constant effect of the endogenous regressor across observations.
  - The modern treatment effects literature considers endogeneity when the endogeneity varies across observations (LATE in the linear IV framework)
- Instrumental variable quantile regression
  - Provides consistent estimates when the treatment is endogenous and has heterogeneous effects
    - ★ Chernozhukov and Hansen (2005, 2006).

Count data

- Quantiles of a count variable are not unique since the c.d.f. is discontinuous, with discrete jumps between flat sections.
- So for count data $y = 0, 1, 2, ...$
  - jitter count y to continuous $z = y + u$ where $u \sim \text{Uniform}[0, 1]$
  - the q^{th} quantile is $q + \exp(\mathbf{x}'\boldsymbol{\beta})$
  - to reduce noise from jittering repeat with multiple draws of u and average $\widehat{\beta}$
  - post-estimation transform back from z quantiles to y quantiles
    - ★ Machado and Santos Silva (2007)
    - ★ Stata add-on qcount

Unconditional quantile regression

- Conditional quantile coefficients estimate the change in $q_\tau(y|\mathbf{x})$ when $\mathbf{x}$ changes
  - if $\Delta x_j = 1$ then what is change in the τ^{th} conditional quantile of y given $\mathbf{x}$?
  - e.g. in τ^{th} quantile of earnings given education, race, gender, age, ...
- Unconditional quantile coefficients estimate the change in $q_\tau(y)$ when $\mathbf{x}$ changes
  - if $\Delta x_j = 1$ then what is change in the τ^{th} quantile of y?
  - this combines both within individual variation (given $\mathbf{x}$) and between variation across ($\mathbf{x}'s$ of all individuals)
  - e.g. in τ^{th} quantile of earnings.

- Firpo, Lemieux and Fortin (2009) provide methods to do this
  - Based on an approximation to the influence function
  - RIF-OLS estimates by OLS $\widehat{\text{RIF}}_\tau(Y) = \mathbf{x}'\gamma_\tau + error$
  - where $\widehat{\gamma}_\tau$ gives the unconditional quantile effect
  - $\widehat{\text{RIF}}_\tau(Y) = \widehat{q}_\tau + (\tau - \mathbf{1}[Y \leq \widehat{q}_\tau])/\widehat{f}_Y(\widehat{q}_\tau)$ is the recentered influence function
  - $\widehat{q}_\tau$ is the usual estimate of the $\tau - th$ quantile of
  - $\widehat{f}_Y(\widehat{q}_\tau)$ is a kernel density estimate of the density of Y evaluated at $\widehat{q}_\tau$.
- Stata add-on `rifreg` at
  - http://faculty.arts.ubc.ca/nfortin/datahead.html

7. Some References

- The material is generally not covered in graduate level texts. Exceptions are
  - CT(2005) MMA chapter 5.6 and CT(2009) MUS chapter 7
- A book-length treatment is
  - Koenker (2005), *Quantile Regression*, Cambridge University Press.
- Articles include
  - Bitler, M., J. Gelbach, H. Hoynes (2006), "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 988-1012.
  - Buchinsky, M. (1994), "Changes in the US wage structure 1963-1987: An application of quantile regression," *Econometrica*, 62, 405-458.
  - Buchinsky. M. (1998), "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research," *Journal of Human Resources*, 33, 88-126.

- More recent literature
  - Machado, J.A.F. and J. Mata (2005), "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of Applied Econometrics*, 20, 445–465.
  - Angrist, J., V. Chernozhukov, and Ivan Fernandez-Val (2006), "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.
  - Chernozhukov, V., and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.
  - Chernozhukov, V., and C. Hansen (2006), "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.
  - Machado, J.A.F. and J.M.C. Santos Silva (2005), "Quantiles for Counts," *JASA*, 100, 1226-1237.
  - Firpo, S., N.M. Fortin and T. Lemieux (2009), "Unconditional Quantile Regression," Econometrica, 953-973 and NBER Tech WP 339.