(Conditional) Quantile Regression

In essence

- (Conditional) Quantile regression is estimating the coefficients as a conditional quantile function. Recall that OLS is a conditional mean function. Here 'quantile' means 'percentile'. E.g., the 10th quantile is the same as the 10th percentile.

- There are two kinds of quantile regression models: conditional quantile regression models and unconditional quantile regression models. Unless specified otherwise, "quantile regression" usually indicates conditional quantile regression models. In this class, we will focus on the conditional quantile regression models.

- By using quantile regression, you can model the entire distribution of the data rather than estimating only the mean (= OLS).

- Understanding the mathematical logic behind the quantile regression fully will not be easy, but the estimation of the quantile regression using Stata and the interpretation of the results is relatively easy.

- The coefficients estimated in quantile regression for the quantile point q quantifies the expected change in the distribution of y for the quantile point q as x increases by 1 unit net of other covariates.

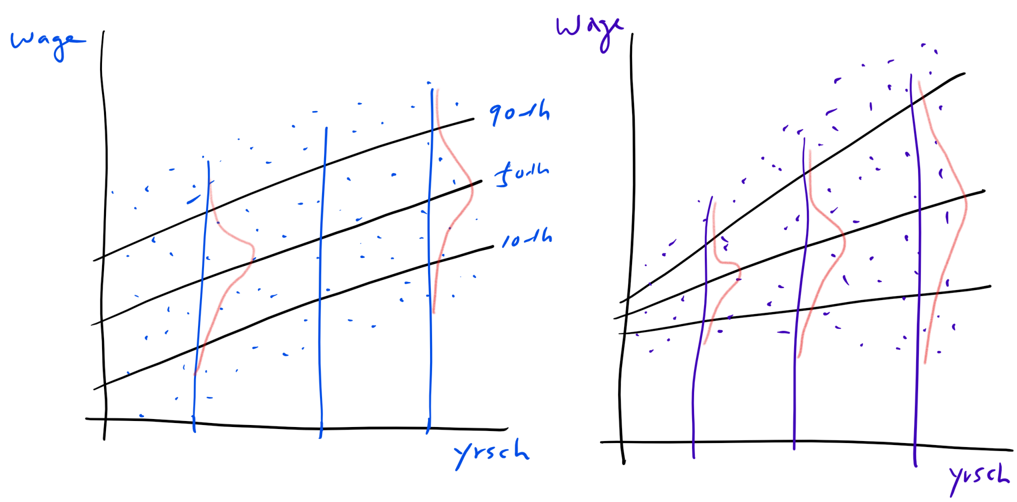
- <u>Caution</u>: the unit change in the conditional quantile regression models is the change in the distribution y, not the change in the expected outcome for individuals.

Graphic Explanation

- In the following graph, the left side is the usual assumption in OLS, homoscedasticity across individual variable x (= years of schooling in the example). The gaps between the less educated and the highly educated are identical across quantile points. That is, gap(high earner among BA - high earner among HSG) = gap(low earner among BA - low earner among HSG).

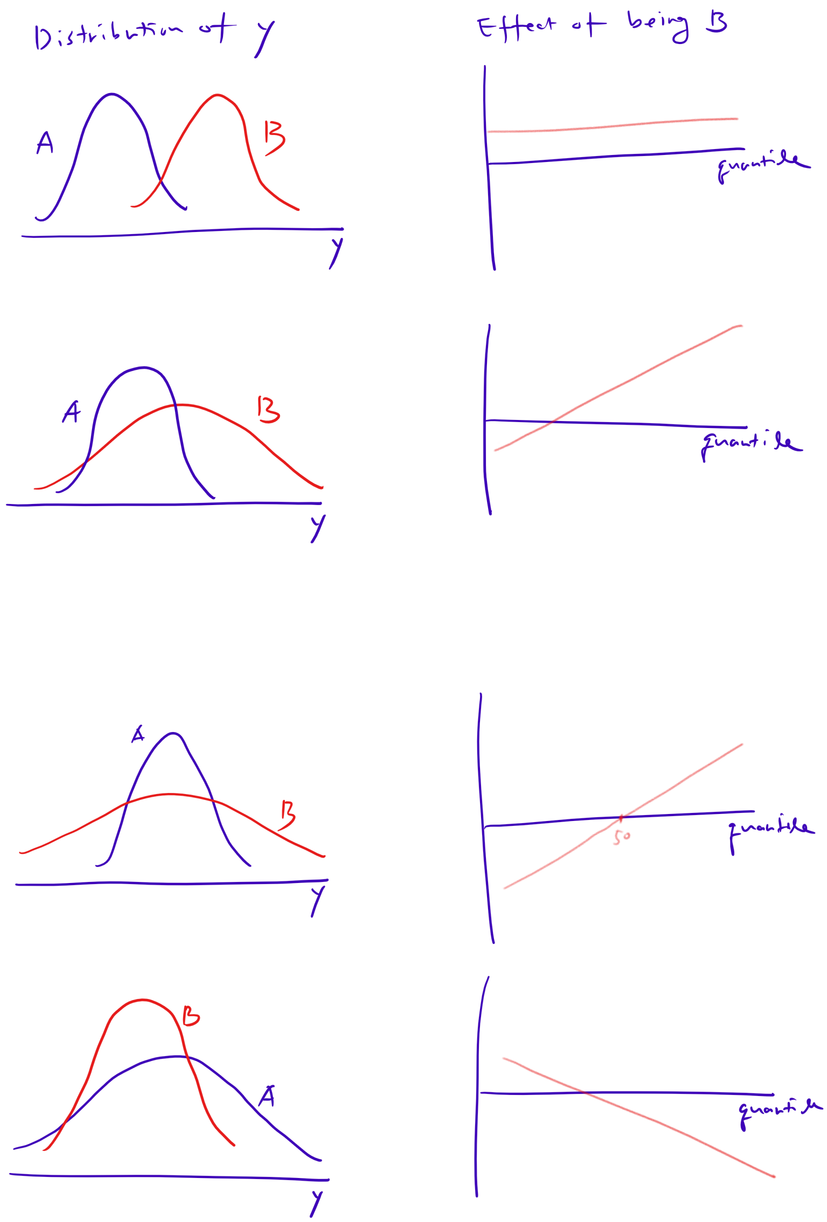
  The right side graph shows that gap(high earner among BA - high earner among HSG) is much larger than gap(low earner among BA - low earner among HSG). At the 90th percentile, gap(BA - HSG) is substantial, while at the 10th percentile, gap(BA - HSG) is negligible. The large gap(BA - HSG) at the 90th percentile indicates that high earners among BA earn much more than high earners among HSG, while the small gap(BA - HSG) at the 10th percentile indicates that low earners among BA earn equally low income with the lower earners among HSG.

  OLS cannot measure this.

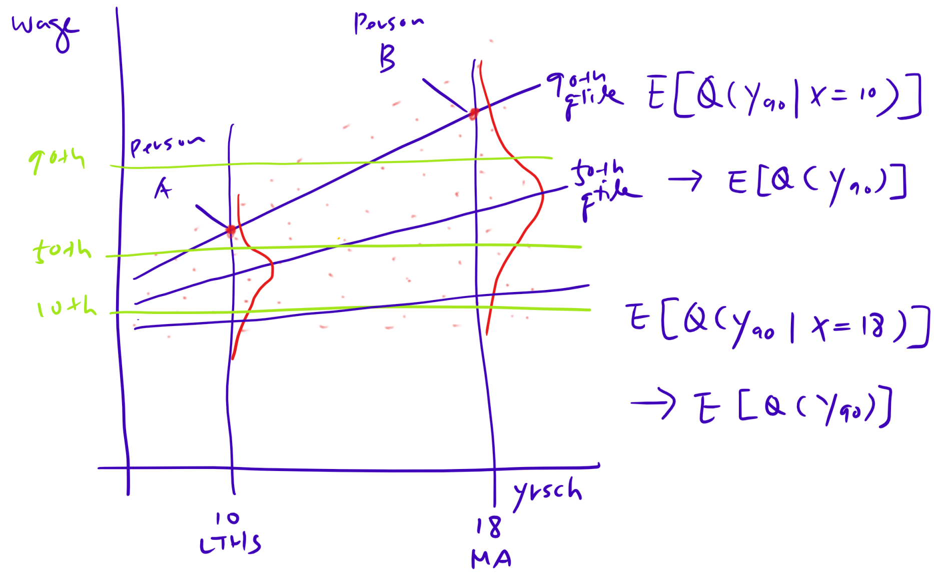


- Consider the following two distributions, A and B. In the 1st example, the shapes of A and B are almost identical. The only difference between two distributions is their mean points. The gaps between lower locations from each distribution, A and B, and the gaps between higher locations are identical. If you draw the effect of B across quantile points, you will get the graph on the right side.

  Look at the 2nd example in which B has a larger variance than A. The lower points of B are lower than the lower points of A, and the higher points of B are higher than the higher points of A. Then, the gaps between A and B across quantiles will look like the graph on the right side. You will be able to figure out what are going on in the 3rd and 4th examples.



- In the following graph, the green lines show the 10th, 50th, and 90th percentile points for the entire distribution. Three blue lines show the lines of 10th, 50th, and 90th percentile points conditional on education.



Slightly Technical Explanation

- For OLS, $E(y|x)$ is the expected y given x. Thus, $E(y|x)$ is the conditional mean. The given condition is a set of x.
- For quantile regression, $Q_q(y|x)$ is the expected quantile q given x. Thus, $Q_q(y|x)$ is the conditional quantile point for the distribution y. The given condition is a set of x.
- The quantile q is between 0 and 1 (thus, $q \in (0, 1)$). It is the point by which the dependent variable y is split into proportions q below and $1 - q$ above. $F(y_q) = q$ (as a function of y_q, we compute q) and $y_q = F^{-1}(q)$ (the inverse function of q, we can estimate y_q).
- $$Q_q(y|x) = x\prime\beta_q = \beta_{0q} + \beta_{1q}x_1 + \beta_{2q}x_2 + \epsilon \tag{1}$$
- For OLS, it minimize the error squared. I.e., minimize $\sum e^2$. For quantile regression, it minimize the sum that gives asymmetric penalties $(1 - q)|e|$ for overprediction and $q|e|$ for underprediction. Although its computation requires linear programming methods, the quantile regression estimator is asymptotically normally distributed.
- Just as regression models conditional moments, such as predictions of the conditional mean function, we may use quantile regression to model conditional quantiles of the joint distribution of y and x.

- Let $\hat{y}(x)$ denote the predictor function and $e(x) = y - \hat{y}(x)$ denote the prediction error. Then $L(e(x)) = L(y - \hat{y}(x))$ denotes the loss associated with the prediction errors. If $L(e) = e^2$, we have squared error loss, and least squares is the optimal predictor. If $L(e) = |e|$, the optimal predictor is the conditional median, $med(y|x)$, and the optimal predictor is that $\hat{\beta}$ which minimizes $\sum |y - x\prime\beta|$.

- Both the squared-error and absolute-error loss functions are symmetric; the sign of the prediction error is not relevant. If the quantile q differs from 0.5, there is an asymmetric penalty, with increasing asymmetry as q approaches 0 or 1.

- More formally, quantile regression estimators minimize the following function:

$$Q(\beta_q) = \sum_{i:y_i \geq x_i\prime\beta}^{N} q|y_i - x_i\prime\beta_q| + \sum_{i:y_i < x_i\prime\beta}^{N} (1-q)|y_i - x_i\prime\beta_q| \tag{2}$$

- Standard errors are estimated using bootstrapping method.

Differences from OLS

- While OLS is sensitive to outliers, quantile regression is much more robust to the outliers.

- Quantile regression estimates are semiparametric as quantile regression avoids assumptions about the parametric distribution of the error process. That is, unlike OLS, quantile regression does not assume normal distribution of error terms. Therefore, there are no such problems like nonnormality, heteroscadastisity. While OLS can be inefficient if the errors are highly non-normal, quantile regression is more robust to non-normal errors and outliers.

- Quantile regression provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean.

- Furthermore, quantile regression is invariant to monotonic transformations, such as $\log(y)$, so the quantiles of $\log(y)$, a monotone transform of y, are $h(Q_q(y))$, and the inverse transformation may be used to translate the results back to y. As we discussed previously, this is not possible for OLS. Mean of log-transformed y is not equal to the log-transformed mean. OLS computes the geometric mean when the dependent variable is log-transformed. Unlike OLS, quantile of log-transformed y can be reconverted to original y without an issue.

Unconditional Quantile Regression Models

- Firpo, Fortin, & Lemieux. 2009. "Unconditional Quantile Regressions." *Econometrica* 77(3):953-973.

- You can download and install a user-written Stata program, "rifreg" from https://faculty.arts.ubc.ca/nfortin/datahead.html.

- Key Idea: The results of conditional quantile regression (CQR) show how much y will change for distribution y as x changes by 1 unit at quantile point q for a given set of other covariates. But it does not quantify how much y will shift at quantile q for the entire population distribution as x changes by 1 unit. Nor does it quantify the expected change in individual outcomes. For example, what will be the value of the 10th quantile point if everyone has college education? Except special cases, CQR cannot answer this question. Firpo et al.'s unconditional quantile regression models (UQR) makes this possible by using so-called "recentered influence function."
- The interpretation of UQR is the same as OLS.
- If you're interested in the showing the group difference across quantiles, CQR is enough. If you're interested in the net effect of x on y for the entire population, use UQR.

Example 1. Estimating QR without a covariate is the same as computing percentile points with standard errors.

```
. sum earning, d

                           earning
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2770.333       .8541667
 5%     9879.678       .8541667
10%     15295.54       .9571984       Obs             204,402
25%     25783.88       3.322048       Sum of Wgt.     204,402
50%        40000                      Mean           52904.92
75%        60000       684085.2
90%        94000       684085.2       Variance       2.83e+09
95%       137654       740851.2       Skewness       3.666459
99%     321032.7       745147.7       Kurtosis       20.81034

. qreg earning, q(25)

.25 Quantile regression                            Number of obs =    204,402
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   25783.88   44.54687   578.80   0.000     25696.56    25871.19
------------------------------------------------------------------------------

. qreg earning, q(50)

Median regression                                  Number of obs =    204,402
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |      40000   51.10088   782.77   0.000     39899.84    40100.16
------------------------------------------------------------------------------

. qreg earning, q(75)

.75 Quantile regression                            Number of obs =    204,402
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |      60000   116.4434   515.27   0.000     59771.77    60228.23
------------------------------------------------------------------------------
```

Example 2. Quantile regression between 2 groups.

(1) Quantile Earnings Estimates for whites

```
. qreg earning if white==1, q(25)
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   25211.85   42.43177   594.17   0.000     25128.68    25295.01
------------------------------------------------------------------------------

. qreg earning if white==1, q(50)
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |      39400   64.47439   611.10   0.000     39273.63    39526.37
------------------------------------------------------------------------------

. qreg earning if white==1, q(75)
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |      60000   75.53552   794.33   0.000     59851.95    60148.05
------------------------------------------------------------------------------
```

(2) Quantile Earnings Estimates for Asian Americans

```
. qreg earning if white==0, q(25)
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   29416.25   257.3963   114.28   0.000     28911.72    29920.78
------------------------------------------------------------------------------

. qreg earning if white==0, q(50)
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   45547.84   323.3055   140.88   0.000     44914.12    46181.56
------------------------------------------------------------------------------

. qreg earning if white==0, q(75)
------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |      70000   489.6096   142.97   0.000      69040.3     70959.7
------------------------------------------------------------------------------
```

(3) Quantile regression on earnings: Dep = earnings, Independent = Asian.
The coefficients of Asian Americans should be equal to (2) - (1).
The constant shows the expected quantile estimates for the reference group.

```
. qreg earning Asian, q(25)

.25 Quantile regression                           Number of obs =    204,402
  Raw sum of deviations 1.87e+09 (about 25783.875)
  Min sum of deviations 1.87e+09                  Pseudo R2     =     0.0007

------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Asian |   4204.402   170.1813    24.71   0.000     3870.851    4537.953
       _cons |   25211.85   44.98773   560.42   0.000     25123.67    25300.02
------------------------------------------------------------------------------

. qreg earning Asian, q(50)

Median regression                                 Number of obs =    204,402
  Raw sum of deviations 2.91e+09 (about 40000)
  Min sum of deviations 2.91e+09                  Pseudo R2     =     0.0012

------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Asian |   6147.836   241.9535    25.41   0.000     5673.613    6622.059
       _cons |      39400   63.96085   616.00   0.000     39274.64    39525.36
------------------------------------------------------------------------------

. qreg earning Asian, q(75)

.75 Quantile regression                           Number of obs =    204,402
  Raw sum of deviations 3.13e+09 (about 60000)
  Min sum of deviations 3.12e+09                  Pseudo R2     =     0.0017

------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Asian |      10000   305.5854    32.72   0.000      9401.06    10598.94
       _cons |      60000   80.78206   742.74   0.000     59841.67    60158.33
------------------------------------------------------------------------------
```

Example 3. Quantile regression coefficients estimated can be directly transformed into log and anti-log without bias.

```
. qreg earning i.Asian##i.edu, q(90)

.9 Quantile regression                              Number of obs =    204,402
  Raw sum of deviations 2.50e+09 (about 94000)
  Min sum of deviations 2.09e+09                    Pseudo R2     =     0.1607

------------------------------------------------------------------------------
     earning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.Asian |   5694.902   9866.125     0.58   0.564    -13642.46    25032.27
             |
         edu |
          2  |    8593.77   1732.837     4.96   0.000     5197.451    11990.09
          3  |   23432.25   1733.239    13.52   0.000     20035.14    26829.36
          4  |   73432.25   1808.311    40.61   0.000        69888     76976.5
          5  |   163995.3   1982.974    82.70   0.000     160108.7    167881.8
             |
   Asian#edu |
        1 2  |  -4856.422   10856.66    -0.45   0.655    -26135.21    16422.36
        1 3  |   -9260.91   10352.43    -0.89   0.371    -29551.42     11029.6
        1 4  |   -30694.9   10244.58    -3.00   0.003    -50774.02   -10615.79
        1 5  |  -22686.43   10482.92    -2.16   0.030     -43232.7   -2140.166
             |
       _cons |   51567.75    1544.88    33.38   0.000     48539.82    54595.68
------------------------------------------------------------------------------

. qreg lnearning i.Asian##i.edu, q(90)
Iteration  1:  WLS sum of weighted deviations =  50137.587

.9 Quantile regression                              Number of obs =    204,402
  Raw sum of deviations 28943.93 (about 11.45105)
  Min sum of deviations 24886.29                    Pseudo R2     =     0.1402

------------------------------------------------------------------------------
   lnearning |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     1.Asian |   .1047525   .0748318     1.40   0.162    -.0419159     .251421
             |
         edu |
          2  |   .1541367   .0131431    11.73   0.000     .1283765    .1798968
          3  |   .3745918   .0131461    28.49   0.000     .3488257    .4003579
          4  |    .885417   .0137155    64.56   0.000     .8585349    .9122991
          5  |   1.430357   .0150403    95.10   0.000     1.400878    1.459836
             |
   Asian#edu |
        1 2  |  -.0909119   .0823447    -1.10   0.270    -.2523054    .0704817
        1 3  |  -.1534672   .0785202    -1.95   0.051     -.307365    .0004306
        1 4  |  -.3278961   .0777022    -4.22   0.000    -.4801905   -.1756017
        1 5  |  -.1868572     .07951    -2.35   0.019    -.3426948   -.0310196
```

```
             |
       _cons |   10.85065   .0117175   926.02   0.000     10.82769    10.87362
------------------------------------------------------------------------------
```

In the above example, the constant (= the expected 90th percentile point for whites without high school diploma) is 10.85065 in the 2nd model. When anti-log is taken, exp(10.85065) = 51567.66 which is almost identical to the constant of the 1st model.
For the expected 90th percentile point for whites without high school diploma from the 2nd model is 10.85065 + .1047525 = 10.9554. By taking anti-log, it is 57262.43. From the 1st model, it is 51567.75 + 5694.902 = 57262.65.