Introduction to Quantile Regression and Applications

Cliff J. Huang
Emeritus Professor
Department of Economics
Vanderbilt University

Yung Lieh. Yang
Emeritus Professor
Department of Finance
Ling Tung University

November 2015

Outline

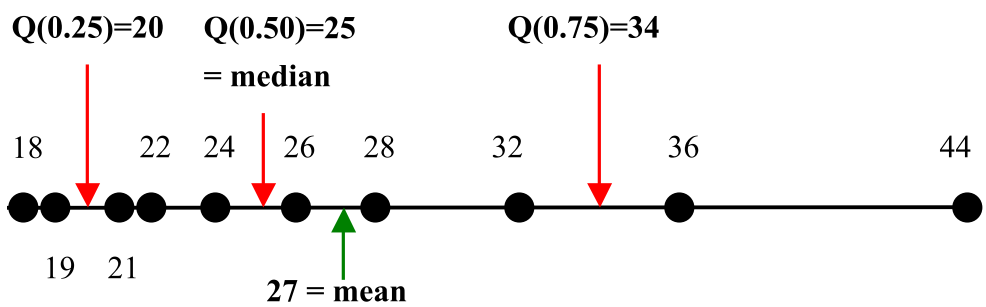
I. Mean, Quantiles, and Distribution
II. Quantile Regression
III. Fundamentals of Quantile Estimation
IV. Computation of Quantile Regression: R-Program
V. Conclusion Remarks

I. Mean, Quantiles, and Distribution

1. Distribution of a Random Variable
   - Example: Starting Salary of 10 Employees

     Y = 18, 19, 21, 22, 24, 26, 28, 32, 36, 44 (in thousands of NT$)

     ❖ Mean = 27



❖ Quartiles:

First Quartile (25 Percentile) = 20

Second Quartile (50 Percentile) = 25 = Median

Third Quartile (75 Percentile) = 34

❖ Distribution Skewness

Symmetric (No Skewness):

Median = Mean

Median – First Quartile = Third Quartile – Median

Positive Skewness (Skew to the Right)

Median < Mean

Median – First Quartile < Third Quartile – Median

Negative Skewness (Skew to the Left)

Median > Mean

Median – First Quartile > Third Quartile – Median

- Quantiles of order τ: $Q(\tau)$, $0 \le \tau \le 1$

❖ $Q(0.20) = 19$; $Q(0.25) = 20$; $Q(0.50) = 25$

$Q(0.75) = 34$; $Q(0.90) = 36$; $Q(1.00) = ?$

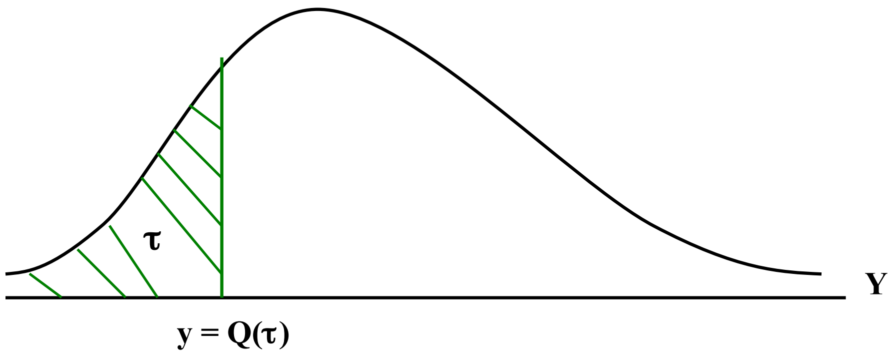
❖ Quantile as a Ranking:

$Q(\tau) < Q(\lambda)$ if $\tau < \lambda$

2. Quantile and Distribution Function

- Distribution Function

$$F(y) = P(Y \le y)$$



- Quantiles

❖ Quantiles of order τ:

$$F(y) = P(Y \le y) = \tau \quad \text{implies} \quad y = F^{-1}(\tau) = Q(\tau)$$

Examples:

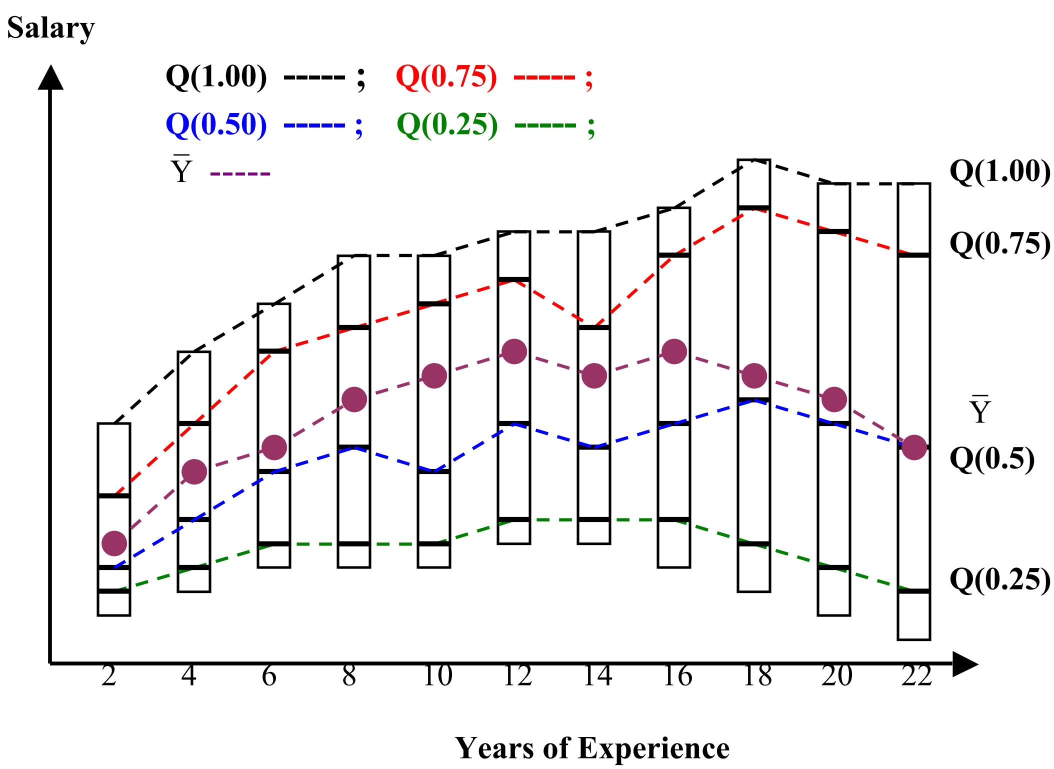
$$25 = F^{-1}(0.5) = Q(0.5)$$

$$34 = F^{-1}(0.75) = Q(0.75)$$

II. Quantile Regression

1. Classical Regression versus Quantile Regression
   - Example:

     y = salary; x = years of experience



   - Classical Regression

     $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

     ❖ Conditional Mean Regression

     $$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i \equiv \mu_i$$

   - Qunatile Regression of order $\tau \in [0, 1]$

     $$y_i = \beta_0^\tau + \beta_1^\tau x_i + \varepsilon_i$$

     ❖ Conditional Quantile Regression

     $$Q(\tau \mid x_i) = \beta_0^\tau + \beta_1^\tau x_i$$

2. Example: Quantile Regressions of Wage Determination in Taiwan
   - Data: 29,133 plant workers in 1996

     wage = hourly wage; experience = months on current job
   - Regressions:

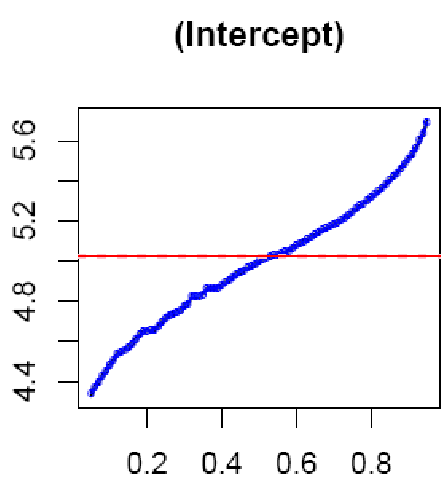
- Conditional Mean Regression

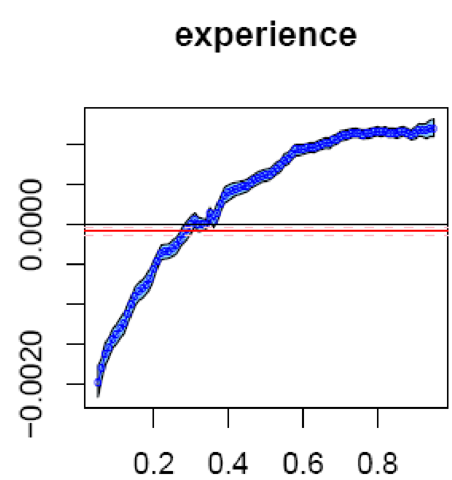
  $\ln(\text{wage}) = \beta_0 + \beta_1 (\text{experience}) + \varepsilon$

- Qunatile Regression

  $\ln(\text{wage}) = \beta_0^\tau + \beta_1^\tau (\text{experience}) + \varepsilon$

τ	β_0^τ	β_1^τ
0.1	4.47987	-0.00137
0.2	4.65334	-0.00056
0.3	4.77952	0.00000
0.4	4.88014	0.00040
0.5	4.99839	0.00062
0.6	5.07736	0.00094
0.7	5.18110	0.00111
0.8	5.31419	0.00117
0.9	5.50766	0.00115
0.95	5.69128	0.00121
Conditional Mean	5.02400	-0.000086





III. Fundamentals of Quantile Regression Estimation

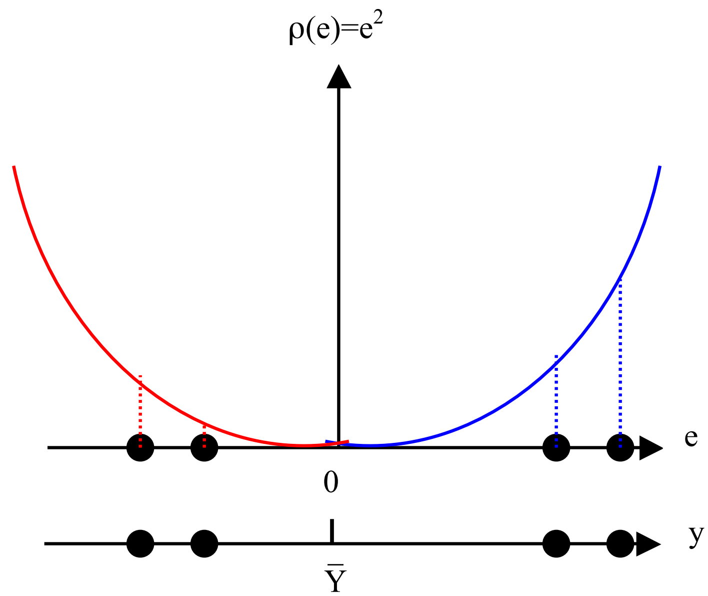
1. Estimation of Mean and Quantile

- Sample Observations: $y = \left(y_1, y_2, \ldots, y_n\right)$
- Population Mean and Population Quantile

  $E\left(y_i\right) = \mu;\ \ Q\left(\tau\right) = F^{-1}(\tau)$
- Least-square Estimation of Population Mean
  - ❖ Model: $y_i = \mu + \varepsilon_i = \bar{Y} + e_i$

    Error: $e_i = y_i - \bar{Y}$
  - ❖ Loss Function: $\rho\left(e_i\right) = e_i^2 = \left(y_i - \bar{Y}\right)^2$
  - ❖ Least-square method:

$$\min_{\bar{Y}} \sum_{i=1}^{n} \rho\left(e_i\right) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \bar{Y}\right)^2; \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$



- Quantile Estimation of $Q\left(\tau\right)$
  - ❖ Model: $y_i = Q\left(\tau\right) + \varepsilon_i = \hat{Q} + e_i$

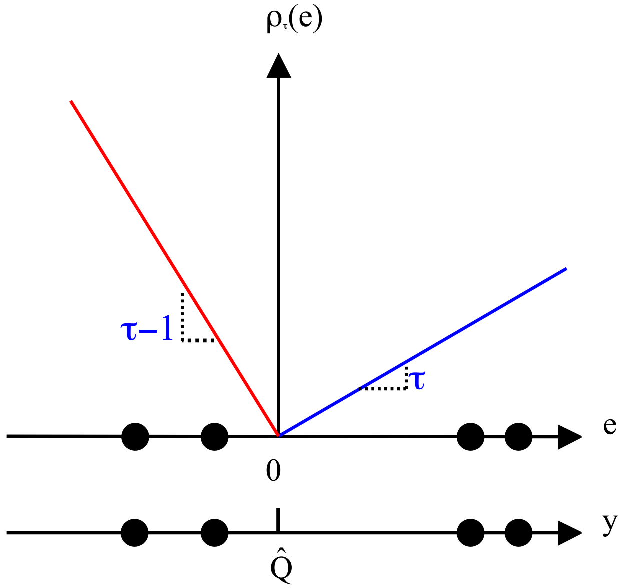
    Error: $e_i = y_i - \hat{Q}$

    Example: $\tau = 0.25$ implies $100\,\tau$ % of $e_i = y_i - \hat{Q} \le 0$

    $100\,(1-\tau)$ % of $e_i = y_i - \hat{Q} \ge 0$

❖ Loss Function:

$$\rho_\tau(e_i) = \tau \times e_i = \tau \times \left(y_i - \hat{Q}\right) \quad \text{if } e_i \ge 0$$
$$= (\tau - 1) \times e_i = (\tau - 1) \times \left(y_i - \hat{Q}\right) \quad \text{if } e_i \le 0$$
$$\rho_\tau(e_i) = \left(\tau - I(e_i < 0)\right) \times e_i = \left(\tau - I\left(y_i - \hat{Q} < 0\right)\right) \times \left(y_i - \hat{Q}\right)$$



❖ Quantile Estimation Method

$$\min_{\hat{Q}} \sum_{i=1}^{n} \rho_\tau(e_i) = \sum_{i=1}^{n} \left(\tau - I\left(y_i - \hat{Q} < 0\right)\right) \times \left(y_i - \hat{Q}\right)$$

Note: (1) For any i, there is the constraint

$$\text{If } \left(y_i - \hat{Q}\right) \begin{cases} \ge \\ \le \end{cases} 0, \text{ then } \rho_\tau(e_i) = \begin{cases} \tau \times \left(y_i - \hat{Q}\right) \\ (1 - \tau) \times \left(\hat{Q} - y_i\right) \end{cases}$$

i.e, for any i,

$$\hat{Q} + \left(y_i - \hat{Q}\right) - \left(\hat{Q} - y_i\right) = y_i$$

(2) Let $u_i = \left(y_i - \hat{Q}\right) \ge 0$, $v_i = \left(\hat{Q} - y_i\right) \ge 0$ be two "slack" variables,

then for any i,

$$\rho_\tau(e_i) = \tau \times u_i + (1 - \tau) \times v_i$$

subject to $\hat{Q} + u_i - v_i = y_i$

❖ Linear Programming Estimation of Quantiles

$$\min_{\hat{Q},u,v} \sum_{i=1}^{n} \left(\tau \times u_i + (1-\tau) \times v_i\right)$$

$$\text{subject to } \hat{Q} + u_i - v_i = y_i, \quad i = 1, 2, \ldots, n$$

$$u_i, v_i \geq 0$$

2. Estimation of Classical Regression and Quantile Regression
   - Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
   - Classical Regression

     ❖ Conditional Mean: $E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$

     ❖ Least-square Method: $\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$

   - Quantile Regression:

     ❖ Conditional Quantile: $Q(\tau \mid x_i) = \beta_0^{\tau} + \beta_1^{\tau} x_i$

     ❖ Quantile Method: $\min_{\hat{\beta}_0^{\tau}, \hat{\beta}_1^{\tau}} \sum_{i=1}^{n} \rho_{\tau}(e_i) = \sum_{i=1}^{n} \left(\tau - I(e_i < 0)\right) \times (e_i)$

$$e_i = y_i - \hat{\beta}_0^{\tau} - \hat{\beta}_1^{\tau} x_i$$

IV. Computation of Quantile Regression: R-Program

- Free Download of **R-Program** at
  http://www.r-project.org
- Example on Wage Determination in Taiwan
  - ❖ Observations: 29,133 plant workers in 1996
  - ❖ Variables: wage = hourly wage; lnwage = ln(wage); gender = 1 male, 0 female;
    age = year of age; married = 1 married, 0 unmarried;
    education = year of education;
    experience = months of working experience on current job
- Use of R-Program
  - ❖ Data file: wage.dat

wage	lnwage	gender	age	married	education	experience
270.83	5.60	1	33	0	9	2
175	5.16	1	36	1	12	125
116.07	4.75	0	28	1	12	26
...	…	..	..	..	…	….

  - ❖ Program file:

```
> library(quantreg)
## change directory
{
dat=read.table("wage.dat",header=T)
attach(dat)
ls <- lm(lnwage~gender+married+education+experience)
summary(ls)
qt <- summary(rq(lnwage~gender+married+education+experience,
                  tau=5:95/100))
postscript("wage.ps", horizontal = FALSE, width = 6.5, height = 3.5)
plot(qt, nrow =1, ncol = 2)
dev.off()
}
```

  - ❖ Outputs:

(1) Least-square outputs:

Coefficients:

(Intercept)	gender	married	education	experience
4.078e+00	3.074e-01	2.242e-01	5.660e-02	-2.925e-05

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.078e+00	1.099e-02	370.912	<2e-16 ***
gender	3.074e-01	6.071e-03	50.641	<2e-16 ***
married	2.242e-01	6.344e-03	35.340	<2e-16 ***
education	5.660e-02	7.833e-04	72.260	<2e-16 ***
experience	-2.925e-05	2.814e-05	-1.039	0.299

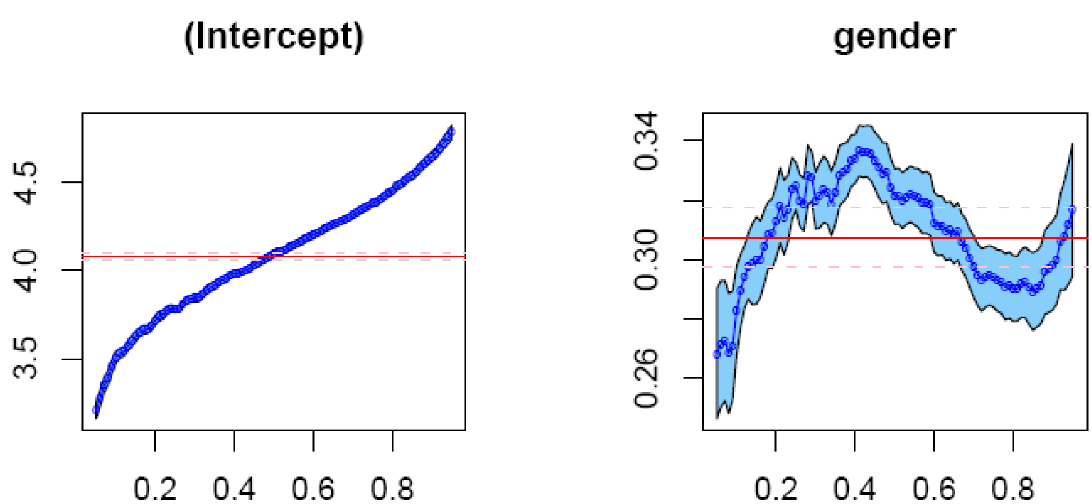
(2) Quantile Outputs:

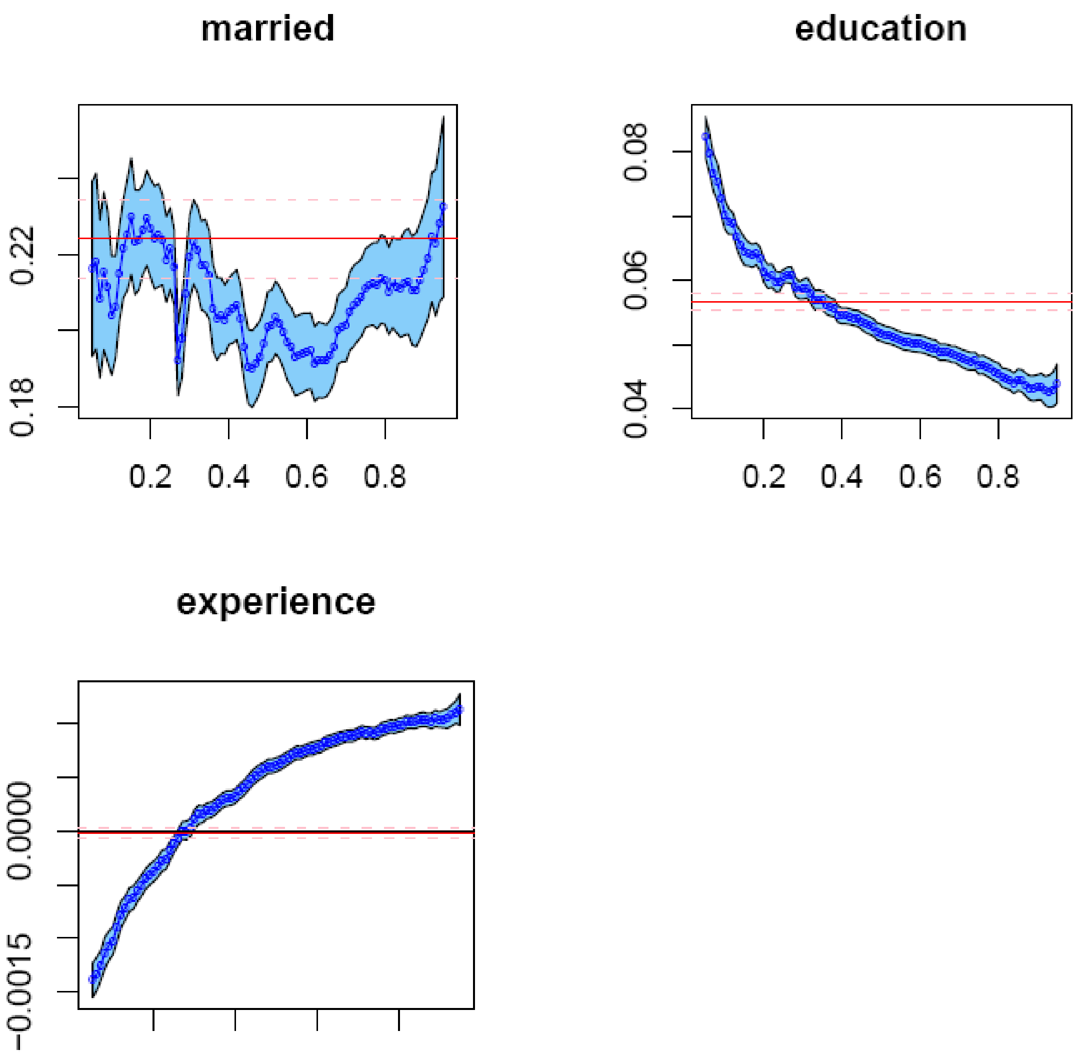
tau: [1] 0.05

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	3.20542	0.02845	112.66064	0.00000
gender	0.26821	0.01336	20.08050	0.00000
married	0.21626	0.01395	15.49835	0.00000
education	0.08225	0.00195	42.17080	0.00000
experience	-0.00139	0.00010	-14.33464	0.00000

(3) Graphs of $y_i = \beta\tau$




married
0.18
0.22
0.2
0.4
0.6
0.8
education
0.04
0.06
0.08
0.2
0.4
0.6
0.8
experience
-0.0015
0.0000
0.2
0.4
0.6
0.8

V. Conclusion Remarks on Quantile Regression

1. Regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Homocedastic Case: $\varepsilon_i \sim \text{iid}$

$$Q_Y(\tau \mid x_i) = \beta_0 + \beta_1 x_i + Q_\varepsilon(\tau) = \left(\beta_0 + Q_\varepsilon(\tau)\right) + \beta_1 x_i$$
$$= \beta_0^\tau + \beta_1 x_i$$

Intercept varies with τ, but not slope.

- Heteroscedastic Case: $\varepsilon_i = h(x_i) w_i = x_i \ w_i$

$$Q_Y(\tau \mid x_i) = \beta_0 + \beta_1 x_i + x_i Q_w(\tau) = \beta_0 + \left(\beta_1 + Q_w(\tau)\right) x_i$$
$$= \beta_0 + \beta_1^\tau x_i$$

Slope varies with τ, but not intercept.

2. Interpreting Quantile Regression Coefficients:

- Classical Regression:
  - ❖ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \ \ E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$

$$\frac{\partial E(y_i \mid x_i)}{\partial x_i} = \beta_1$$

  - ❖ Transformation: $\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i; \ \ E(\ln(y_i) \mid x_i) = \beta_0 + \beta_1 x_i;$

$$\frac{\partial E(\ln(y_i) \mid x_i)}{\partial x_i} = \beta_1, \ \text{but} \ \frac{\partial E(y_i \mid x_i)}{\partial x_i} \neq e^{\beta_1}$$

- Quantile Regression:
  - ❖ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \ \ Q_y(\tau \mid x_i) = \beta_0^\tau + \beta_1^\tau x_i$

$$\frac{\partial Q_y(\tau \mid x_i)}{\partial x_i} = \beta_1^\tau$$

  - ❖ Transformation: $\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i; \ \ Q_{\ln(y)}(\tau \mid x_i) = \beta_0^\tau + \beta_1^\tau x_i$

$$\frac{\partial Q_{\ln(y)}(\tau \mid x_i)}{\partial x_i} = \beta_1^\tau \ \text{ and } \ \frac{\partial Q_y(\tau \mid x_i)}{\partial x_i} = e^{\beta_1^\tau}$$