Quantile Regression Confidence Intervals

Theoretical Exposition and Empirical Findings

Garth Tarr

School of Mathematics and Statistics
University of Sydney

25th September, 2009

Outline

What is Quantile Regression?

Quantile regression is a statistical technique intended to estimate, and conduct inference about, conditional quantile functions.

Outline

Quantiles

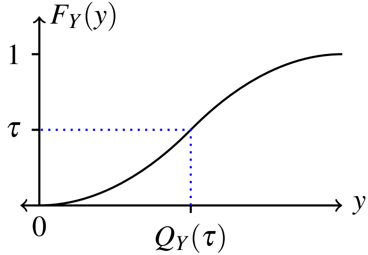
Quantile Regression

Confidence Intervals

Simulation Study

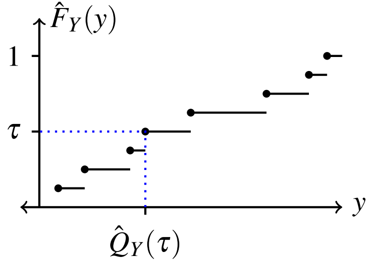
Conclusion

What is a Quantile Function?



The quantile function

$$Q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{y | F_Y(y) \geq \tau\}$$



The empirical quantile function

$$\hat{Q}_Y(\tau) = \hat{F}_Y^{-1}(\tau) = \inf\left\{ y \middle| \frac{\#(Y_i \leq y)}{n} \geq \tau \right\}$$

Quantile Estimation

- Historically estimation of $Q_Y(\tau)$ was accomplished by ranking.
- Koenker and Bassett (1978) proposed a method based on an optimisation problem:

$$\hat{Q}_Y(\tau) = \underset{\beta_\tau \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i \in \{Y_i \geq \beta_\tau\}} \tau \left| Y_i - \beta_\tau \right| + \sum_{i \in \{Y_i < \beta_\tau\}} (1-\tau) \left| Y_i - \beta_\tau \right| \right\}$$

where the solution is given by $\hat{Q}_Y(\tau) = \hat{\beta}_\tau$, the τth quantile of Y.

Introducing the check function

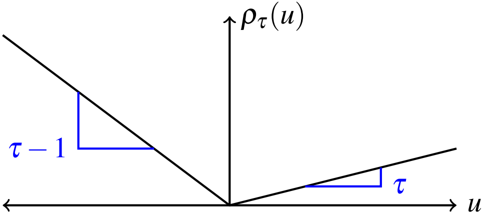
Condensed:

$$\rho_\tau(u) = u\big(\tau - \mathbb{I}(u < 0)\big)$$

Expanded:

$$\rho_\tau(u) = \begin{cases} u(\tau - 1) & ; \quad u < 0 \\ u\tau & ; \quad u \geq 0 \end{cases}$$

Graphically:



Reformulation of the objective function

Previously:

$$\hat{Q}_Y(\tau) = \underset{\beta_\tau \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i \in \{Y_i \geq \beta_\tau\}} \tau \left| Y_i - \beta_\tau \right| + \sum_{i \in \{Y_i < \beta_\tau\}} (1-\tau) \left| Y_i - \beta_\tau \right| \right\}$$

Using the check function:

$$\hat{Q}_Y(\tau) = \underset{\beta_\tau \in \mathbb{R}}{\operatorname{argmin}} \sum_i \rho_\tau (Y_i - \beta_\tau)$$

Outline

Formulating the problem

The vector of quantile regression coefficients, $\hat{\boldsymbol{\beta}}_\tau$, is found by solving

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta}).$$

Done by minimising the expected loss:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}} \left[\tau \int_{\mathbf{y}>X^T\boldsymbol{\beta}} \left|\mathbf{y} - X^T\boldsymbol{\beta}\right| dF_{Y|X}(y) + (1-\tau) \int_{\mathbf{y}<X^T\boldsymbol{\beta}} \left|\mathbf{y} - X^T\boldsymbol{\beta}\right| dF_{Y|X}(y)\right]$$

using linear programming techniques.

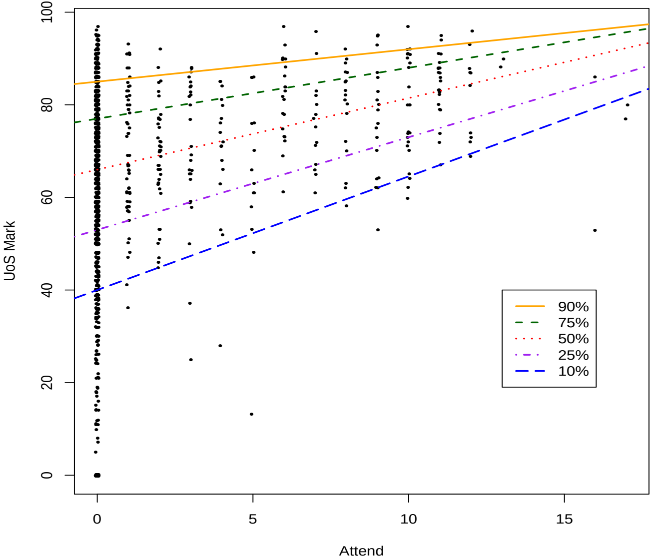
SLR Example

Regressing final UoS mark on attendance at a voluntary supplementary program, Peer Assisted Study Sessions:

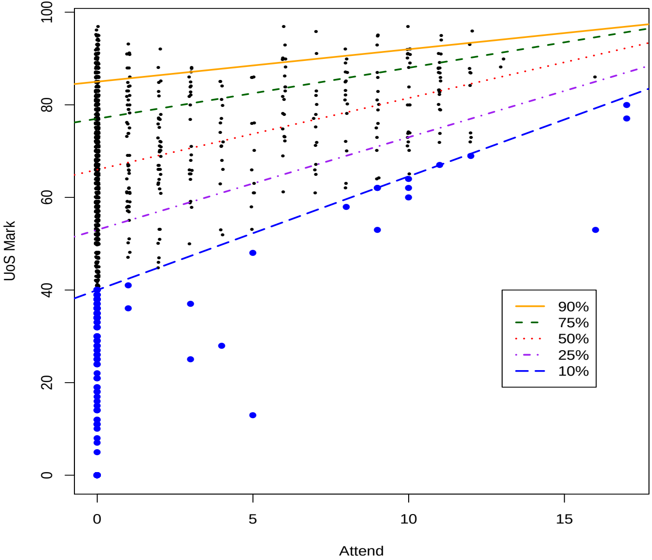
$$\texttt{UoSmark}_i = \beta_0 + \beta_1 \texttt{Attend}_i + \varepsilon_i$$

Covariates	0.10	0.25	0.50	0.75	0.90
(Intercept)	40.000 (1.341)	53.000 (0.707)	66.000 (0.559)	77.000 (0.502)	85.000 (0.538)
Attend	2.455 (0.191)	2.000 (0.202)	1.545 (0.107)	1.100 (0.116)	0.700 (0.084)

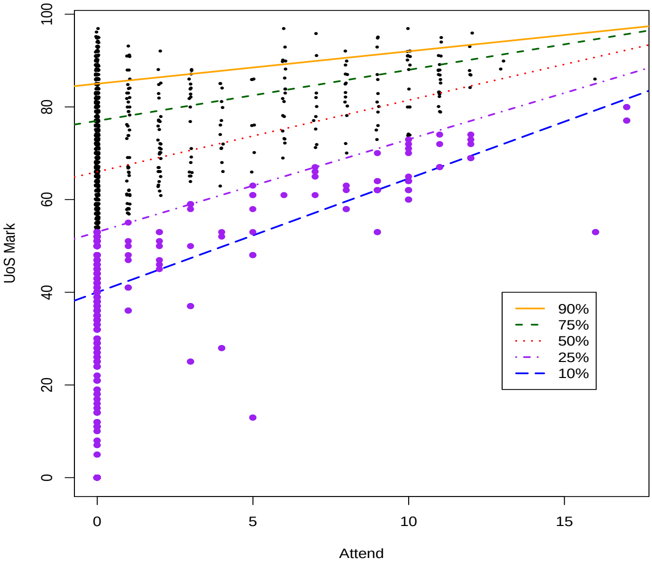
SLR Example



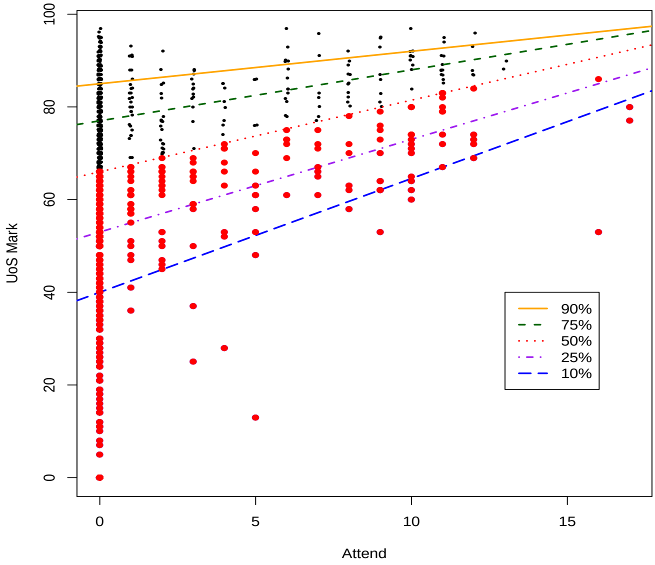
SLR Example



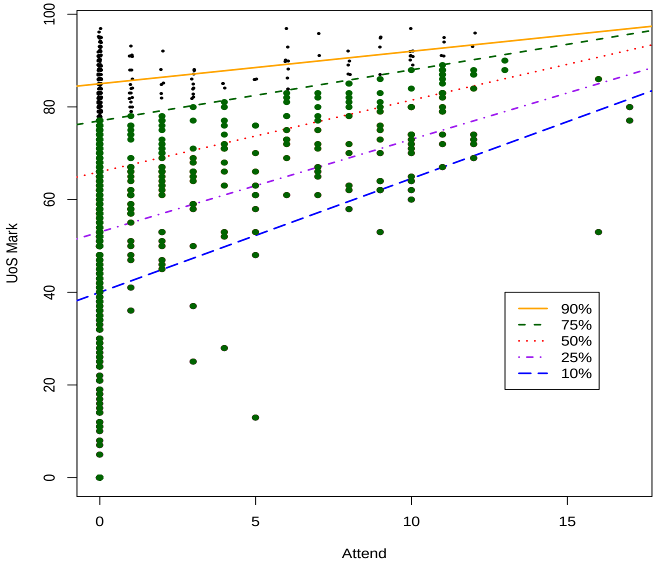
SLR Example



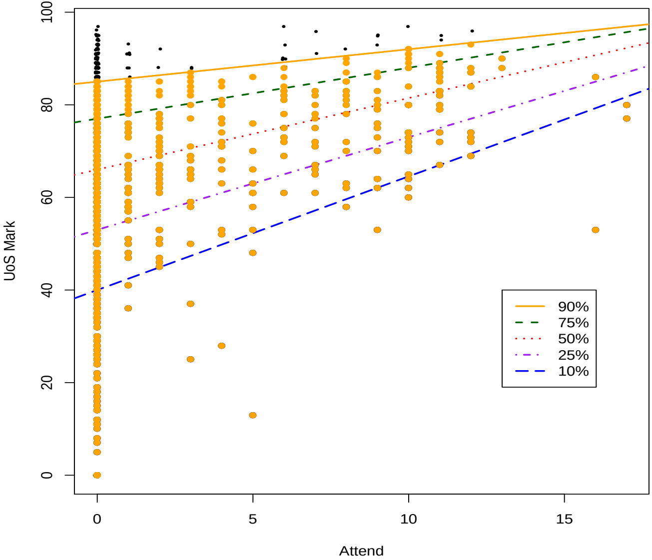
SLR Example



SLR Example



SLR Example



Outline

Quantiles

Quantile Regression

Confidence Intervals

Simulation Study

Conclusion

Confidence Intervals for Quantile Regression Estimates

1. Direct Estimation
   1.1 Independent and identically distributed errors (iid)
   1.2 Not iid errors (nid)
2. Resampling Methods
   2.1 xy-bootstrap (xy)
   2.2 Parzen Wei Ying Approach (pwy)
   2.3 Markov Chain Marginal Bootstrap (mcmb)
   2.4 Generalised Bootstrap (wxy)
3. Rank Score Method
   3.1 Independent and identically distributed errors (riid)
   3.2 Not iid errors (rnid)

Rank Score Method

- Regression Rank Scores come about from the linear programming method used to find quantile regression estimates.
- Using standard rank test theory, these can be manipulated to form a test statistic.
- The riid CI is found by "inverting" the test statistic.
- The non iid (rnid) CI is similar, though it allows for $\varepsilon_i \sim F_i$ type considerations in the construction of its test statistic.
- These CIs aren't necessarily symmetric.

Rank Score Method

- Regression Rank Scores come about from the linear programming method used to find quantile regression estimates.
- Using standard rank test theory, these can be manipulated to form a test statistic.
- The riid CI is found by “inverting” the test statistic.
- The non iid (rnid) CI is similar, though it allows for $\varepsilon_i \sim F_i$ type considerations in the construction of its test statistic.
- These CIs aren’t necessarily symmetric.

Rank Score Method

- Regression Rank Scores come about from the linear programming method used to find quantile regression estimates.
- Using standard rank test theory, these can be manipulated to form a test statistic.
- The riid CI is found by “inverting” the test statistic.
- The non iid (rnid) CI is similar, though it allows for $\varepsilon_i \sim F_i$ type considerations in the construction of its test statistic.
- These CIs aren’t necessarily symmetric.

Outline

Purpose

To build on a paper by Kocherginsky, He, and Mu (2005). In particular focusing on:

1. Small sample size analysis $n \leq 200$.
2. Comment on the variability of confidence intervals.
3. Provide an objective analysis.

KPIs:

The key performance indicators are coverage, confidence interval length and standard deviation.

Purpose

To build on a paper by Kocherginsky, He, and Mu (2005). In particular focusing on:

1. Small sample size analysis $n \leq 200$.
2. Comment on the variability of confidence intervals.
3. Provide an objective analysis.

KPIs:

The key performance indicators are coverage, confidence interval length and standard deviation.

Purpose

To build on a paper by Kocherginsky, He, and Mu (2005). In particular focusing on:

1. Small sample size analysis $n \leq 200$.
2. Comment on the variability of confidence intervals.
3. Provide an objective analysis.

KPIs:

The key performance indicators are coverage, confidence interval length and standard deviation.

Purpose

To build on a paper by Kocherginsky, He, and Mu (2005). In particular focusing on:

1. Small sample size analysis $n \leq 200$.
2. Comment on the variability of confidence intervals.
3. Provide an objective analysis.

KPIs:

The key performance indicators are coverage, confidence interval length and standard deviation.

Method

1. Define the model

   `for(i in 1:1000){`

   1.1 Generate the relevant random data
   1.2 Construct the dependent variable
   1.3 Generate confidence interval estimates using each of the 8 different methods.

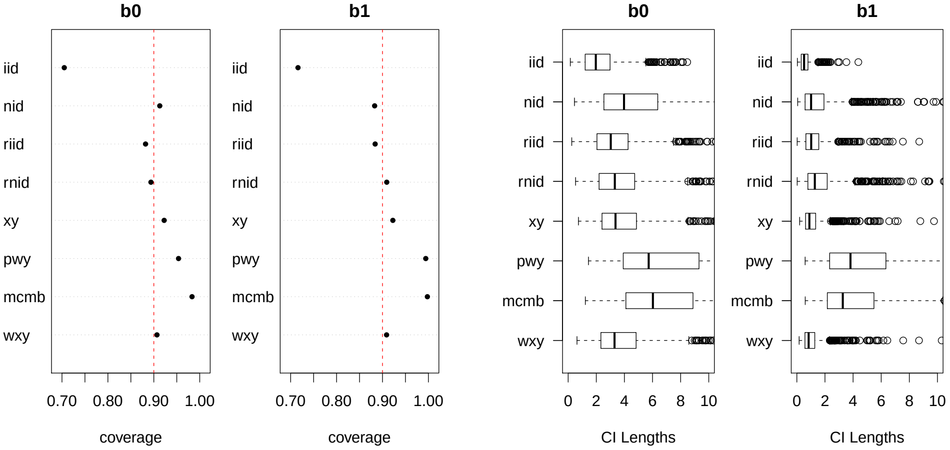
   `}`

2. Plot coverages and lengths and output summary table.

This basic process was done for each model, over $\tau = \{0.1, 0.2, \ldots, 0.9\}$ and $n = 50, 100, 150, 200$.

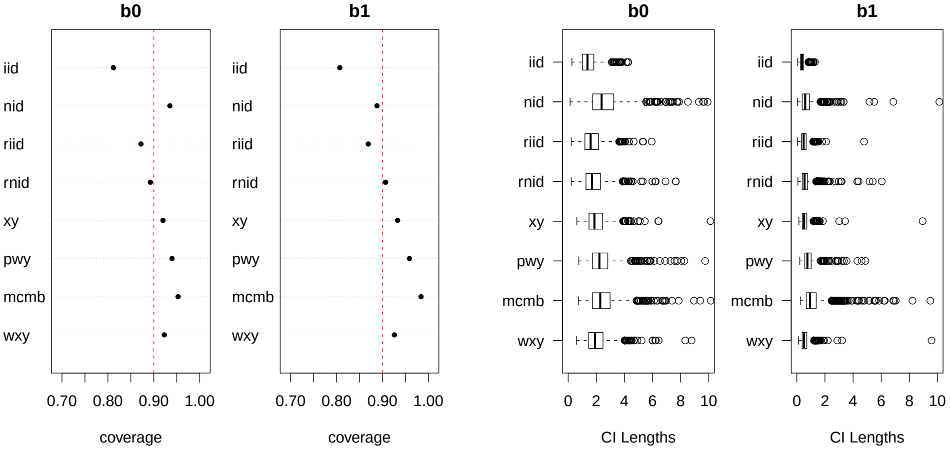
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.1$$



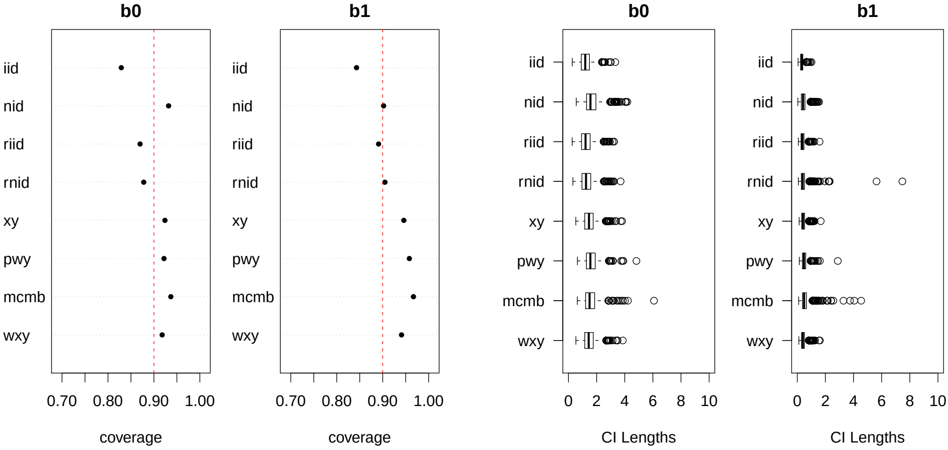
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.2$$



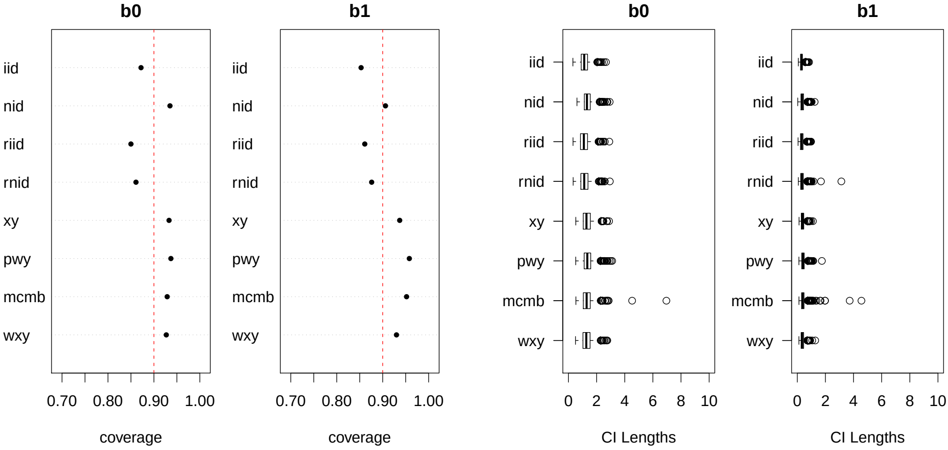
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.3$$



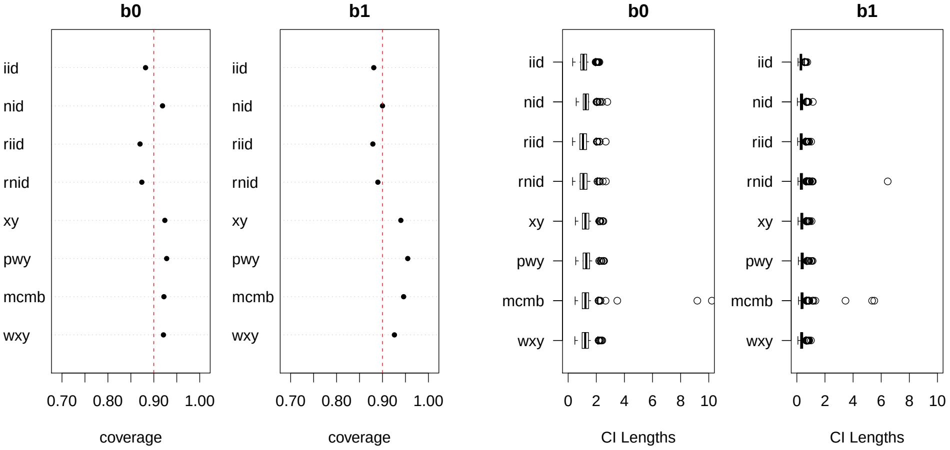
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.4$$



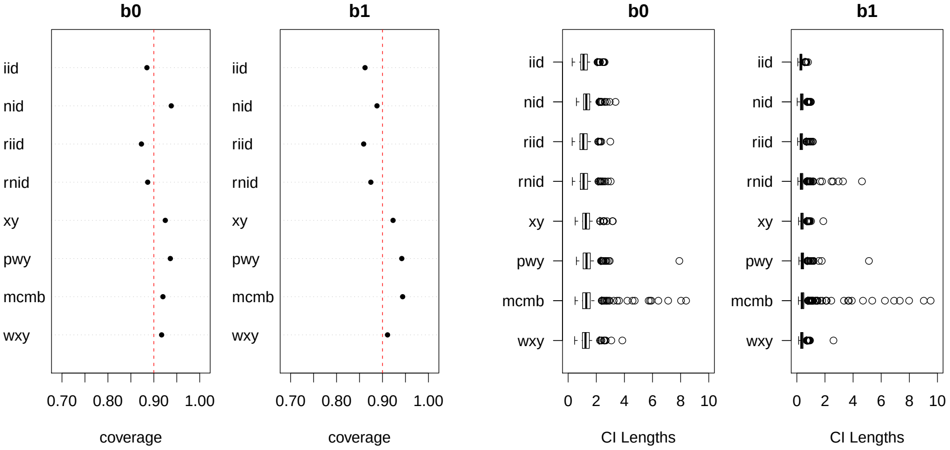
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.5$$



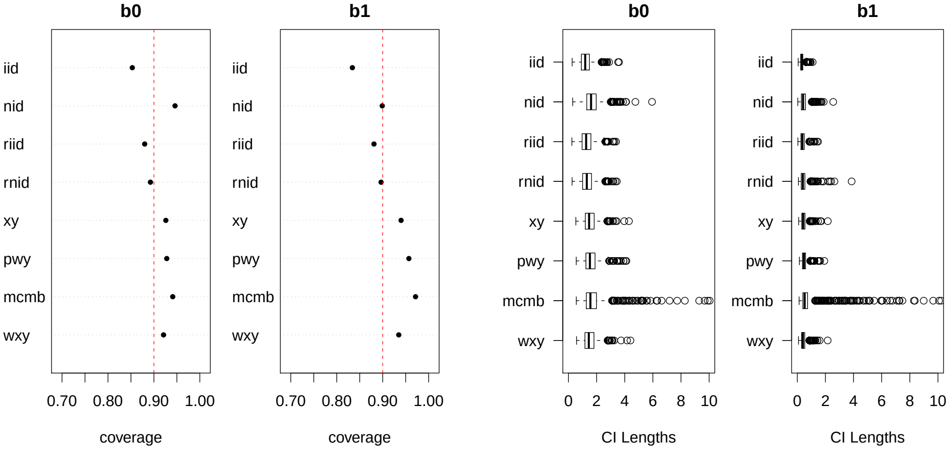
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.6$$



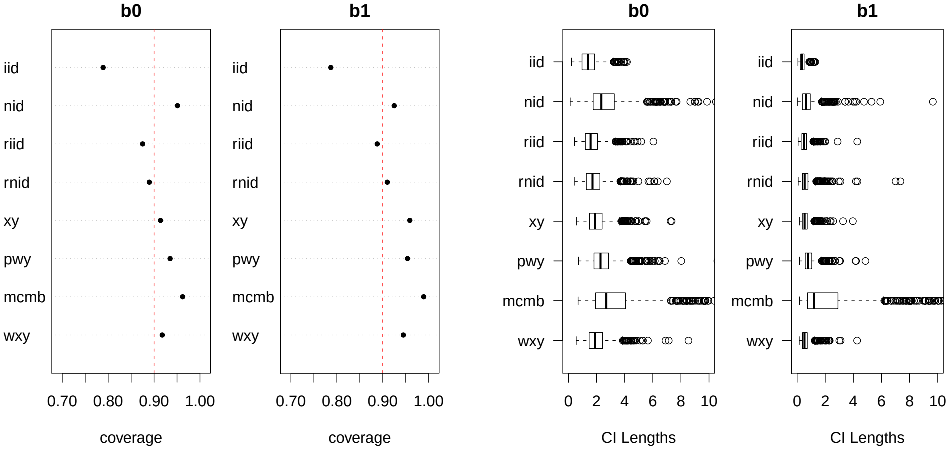
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.7$$



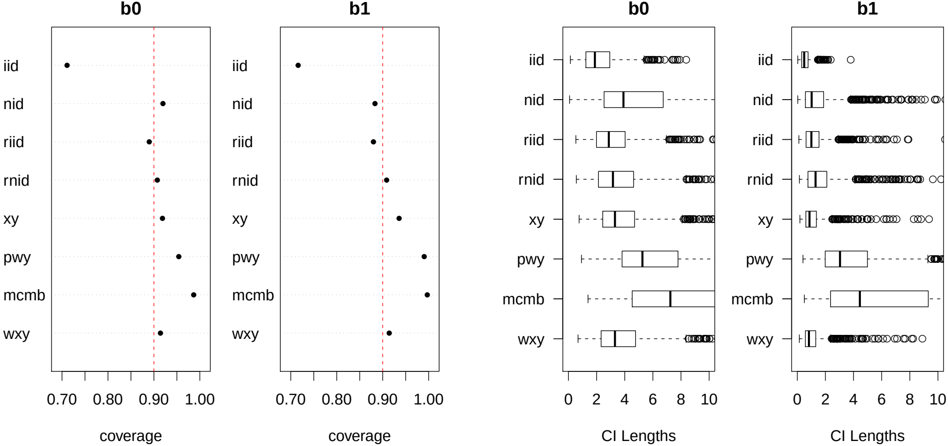
Heavy Tailed Covariates and Errors

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.8$$



Heavy Tailed Covariates and Errors

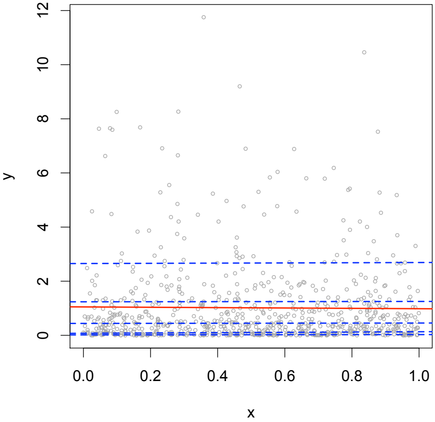
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad x_i \sim \chi^2_3,\ \varepsilon_i \sim t_2 \text{ with } n = 50 \text{ and } \tau = 0.9$$

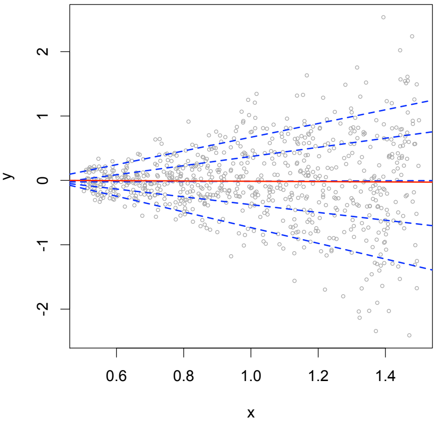


Outline

Summary

Quantile Regression is useful whenever there's a need to gain a broader understanding of the full conditional quantile function of the response variable.





Summary

Careful consideration should be given to which method you will use for confidence interval construction:

- Confident that the iid assumption holds and looking at moderate τ with large n? Difference is immaterial.
- iid with heavy tailed covariates or errors? Use riid or rnid.
- Not sure if the iid assumption holds? Use nid; rnid; xy or wxy.
- Correlation amongst the covariates in a non iid setting? Use nid or wxy.
- Looking at extreme τ in models with heteroskedasticity? Be extremely cautious about your inferences.

Summary

Careful consideration should be given to which method you will use for confidence interval construction:

- Confident that the iid assumption holds and looking at moderate τ with large n? Difference is immaterial.
- iid with heavy tailed covariates or errors? Use riid or rnid.
- Not sure if the iid assumption holds? Use nid; rnid; xy or wxy.
- Correlation amongst the covariates in a non iid setting? Use nid or wxy.
- Looking at extreme τ in models with heteroskedasticity? Be extremely cautious about your inferences.

Summary

Careful consideration should be given to which method you will use for confidence interval construction:

- Confident that the iid assumption holds and looking at moderate τ with large n? Difference is immaterial.
- iid with heavy tailed covariates or errors? Use riid or rnid.
- Not sure if the iid assumption holds? Use nid; rnid; xy or wxy.
- Correlation amongst the covariates in a non iid setting? Use nid or wxy.
- Looking at extreme τ in models with heteroskedasticity? Be extremely cautious about your inferences.

Summary

Careful consideration should be given to which method you will use for confidence interval construction:

- Confident that the iid assumption holds and looking at moderate τ with large n? Difference is immaterial.
- iid with heavy tailed covariates or errors? Use riid or rnid.
- Not sure if the iid assumption holds? Use nid; rnid; xy or wxy.
- Correlation amongst the covariates in a non iid setting? Use nid or wxy.
- Looking at extreme τ in models with heteroskedasticity? Be extremely cautious about your inferences.

Summary

Careful consideration should be given to which method you will use for confidence interval construction:

- Confident that the iid assumption holds and looking at moderate τ with large n? Difference is immaterial.
- iid with heavy tailed covariates or errors? Use riid or rnid.
- Not sure if the iid assumption holds? Use nid; rnid; xy or wxy.
- Correlation amongst the covariates in a non iid setting? Use nid or wxy.
- Looking at extreme τ in models with heteroskedasticity? Be extremely cautious about your inferences.

Key References

Hendricks, W. and Koenker, R. (1992).
Hierarchical spline models for conditional quantiles and demand for electricity.
JASA, 87(417):58–68.

Kocherginsky, M., He, X., and Mu, Y. (2005).
Practical confidence intervals for regression quantiles.
J. Comput. Graph. Statist., 14(1):41–55.

Koenker, R. (2005).
Quantile Regression.
Cambridge University Press, Cambridge.

Koenker, R. and Bassett, Gilbert, J. (1978).
Regression quantiles.
Econometrica, 46(1):33–50.

Koenker, R. and Machado, J. A. F. (1999).
Goodness of fit and related inference processes for quantile regression regression.
JASA, 94(448):1296–1310.

This Slide Intentionally Left Blank.

Some Properties of Quantile Regression Estimates

1. Full conditional distribution estimation
2. Equivariance to monotone transformations

$$Q_\tau\big(h(Y)|X\big) = h\big(Q_\tau(Y|X)\big)$$

   where $h(\cdot)$ is a monotone function.

3. Robustness

$$\hat{\boldsymbol{\beta}}_\tau(\mathbf{y}, X) = \hat{\boldsymbol{\beta}}_\tau\big(\mathbf{y}^\star, X\big)$$

   where

   - $\mathbf{y}^\star = X\hat{\boldsymbol{\beta}}_\tau(\mathbf{y}, X) + D\hat{\mathbf{u}}$
   - D is a diagonal matrix with non-negative elements d_i and
   - $\hat{\mathbf{u}} = \mathbf{y} - X\hat{\boldsymbol{\beta}}_\tau(\mathbf{y}, X)$.

Direct Estimation

1. The iid model comes from the original paper of Koenker and Bassett (1978). Standard normal asymptotic results apply.
2. This basic result has been extended to the non-iid (nid) setting, allowing for $\varepsilon_i \sim F_i$:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau) \xrightarrow{\mathscr{D}} \mathscr{N}(\mathbf{0}, \tau(1-\tau)B_n^{-1}\Omega_n B_n^{-1}),$$

where $\Omega_n = \lim_{n\to\infty} n^{-1}X^TX$ and

$$B_n(\tau) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T f_i(F_i^{-1}(\tau))$$

with

$$\hat{f}_i\big(F_i^{-1}(\tau)\big) = \frac{2h_n}{\mathbf{x}_i^T(\hat{\boldsymbol{\beta}}_{\tau^+} - \hat{\boldsymbol{\beta}}_{\tau^-})}$$

and h_n is a bandwidth; $\tau^\pm = (\tau \pm h_n)$.

Resampling Methods

1. Standard design matrix bootstrap (xy):

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}_\tau) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}^*_{\tau,b} - \bar{\boldsymbol{\beta}}_\tau\right)\left(\hat{\boldsymbol{\beta}}^*_{\tau,b} - \bar{\boldsymbol{\beta}}_\tau\right)^T$$

2. Parzen, Wei and Ying Bootstrap (pwy) exploits the asymptotically pivotal subgradient and bootstraps the estimating equation.
3. Markov Chain Marginal Bootstrap (mcmb) takes the k dimensional traditional bootstrap problem and breaks it down into k, 1 dimensional problems.
4. Generalised Bootstrap (wxy) weights the original objective function with a unit exponential weight.

How to estimate in R

```
require(quantreg)
fit = qr(y~x, tau=0.5)
```

To implement the different methods use:

```
summary.rq(fit, se = ____)
```

- Direct estimation: `se="iid"` or `se="nid"`.
- Rank Inversion: `se="rank"` with `iid=TRUE` or `FALSE`
- Resampling: `se="boot"` with `bsmethod="xy"` or `"pwy"` or `"mcmb"` and specify `R=` the number of resamples.

To plot the estimated coefficients over a range of τ use:

```
plot(summary(rq(y~x,tau=1:99/50),se=____))
```

A Multivariate Example

Extend the previous example to the a multivariate regression model:

$$\texttt{UoSmark}_i = \beta_0 + \beta_1 \texttt{Attend}_i + \beta_2 \texttt{Gender}_i + \beta_3 \texttt{DInt}_i$$
$$+ \beta_4 \texttt{Law}_i + \beta_5 \texttt{Other}_i + \varepsilon_i$$

A Multivariate Example