Robust Regression

유제진

Department of Statistics, University

INDEX

Quantile Regression

Quantile

- r-th Quantile of random variable Y with Cdf F_y

$$Q_Y(\tau) = F_y^{-1}(\tau) = Inf\{y : F_Y(y) \geq \tau\}, \tau \in (0,1)$$

- Here, we define the loss function as

$$\rho_\tau(y) = y * (\tau - I(y < 0)) \text{ where } I : \text{Indicator Function}$$

Also known as "Check Function".

Quantile Regression

Quantile

- CDF Function : function returns probabilities of X being smaller than or equal to some value x.

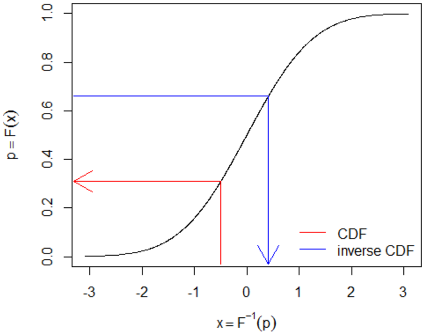
$$Pr(X \leq x) = F(x)$$

- The inverse of CDF = Quantile Function

$$F^{-1}(p) = x$$

where x would return some value P.

Quantile Regression



- Gummbel Distribution

$$F(x) = e^{-e^{-x}}$$

$$\Longleftrightarrow$$

$$F^{-1}(p) = -ln(-lnp)$$

Quantile Regression

Generalized Inverse Distribution Function

- Not every function has an inverse.
- So condition "Monotonically increasing" is needed.
- Also, to be a function, one to one condition need to be satisfied.
- CDF satisfy this condition.
- But what happens when it comes to discrete random variable CDF?
- Discrete random variable CDF is not continuous, but increasing.

Quantile Regression

Generalized Inverse Distribution Function

- Generalized Inverse Distribution only requires non-decreasing condition.

$$Q_Y(\tau) = F_y^{-1}(\tau) = Inf\{y : F_Y(y) \geq \tau\}, \tau \in (0, 1)$$

- For given probability value τ,
- Look for some y that results in $F(y)$ returning value greater or equal then τ.
- But since there could be multiple values of y that meet this condition.
  e.g. $F(y) \geq 0$ is true for any y
- So use Infimum to take the smallest among y.

Quantile Regression

Derivation of Check Function

- Find "Location" x^* relative to a distribution or set of data F.
- mean

$$L_F(\bar{x}) = \int_R (x - \bar{x})^2 dF(x)$$

  mean minimized the expected squared residual.
- L_F is for loss function determined by F.
- To show x^* minimized any function begins with,
- Demonstrating the function's value does not decrease when x^* is changed by a little bit.
- Such a value is called a critical point of the function.

Quantile Regression

Derivation of Check Function

- What kind of loss function Λ would result in a percentile $F^{-1}(\alpha)$ being a critical point?

$$
\begin{aligned}
L_F(F^{-1}(\alpha)) &= \int_R \Lambda(x - F^{-1}(\alpha))dF(x) \\
&= \int_0^1 \Lambda(F^{-1}(u) - F^{-1}(\alpha))du
\end{aligned}
$$

- For this to be critical point, derivative must be zero.

Quantile Regression

Derivation of Check Function

$$0 = L'_F(x^*) = L_F(F^{-1}(\alpha)) = -\int_0^1 \Lambda'(F^{-1}(u) - F^{-1}(\alpha))du$$
$$= -\int_0^{\alpha} \Lambda'(F^{-1}(u) - F^{-1}(\alpha))du - \int_{\alpha}^1 \Lambda'(F^{-1}(u) - F^{-1}(\alpha))du$$

- On the left hand side, argument of Λ' is negative.
- On the right hand side, argument of Λ' is positive.
- Other than that, we have little control over the values of these integrals.
- Because F could be any distribution function.

Quantile Regression

Derivation of Check Function

- Now, make Λ' depend only on the sign of its argument, otherwise it must be constant.
- This implies Λ will be piecewise linear, potentially with different slopes to the left and right of zero.
- Moreover, Rescaling Λ by a constant will not change its properties.
- So may feel free to set the left hand slope to -1.

Quantile Regression

Derivation of Check Function

- Let $\tau > 0$ be the right hand slope.
- Then final equation simplifies to

$$0 = \alpha - \tau(1-\alpha) \implies \tau = \frac{\alpha}{1-\alpha}$$

- To conclude,

$$\Lambda(x) = \begin{cases} -x \\ \dfrac{\alpha}{1-\alpha} x \end{cases} \implies \Lambda(x) = \begin{cases} -(1-\alpha)x & x \leq 0 \\ \alpha x & x \geq 0 \end{cases}$$

Drawing Check function

Check Function

```
rho <- function(u) {u * (tau - ifelse(u <
0,1,0) )}
tau <- .25; curve(rho,-2,2,lty=1,lwd=3)
tau <- .50; curve(rho,
-2,2,lty=2,col="blue",add=T,lwd=3)
tau <- .90; curve(rho,
-2,2,lty=3,col="red",add=T, lwd=3)
abline(v=0,lty=5,col="gray",lwd=3)
legend("bottomleft",c(".25",".5",".9"),
lty=1:3,col=c("black","blue","red"),cex=.6)
```

Drawing Check function

Data Plot

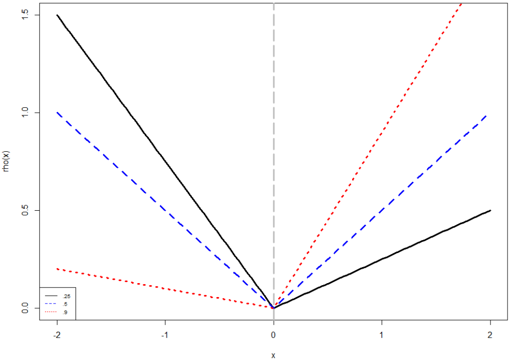


Figure 1: Loss of check function

Building Quantile Regression

Median

- Consider simple $\{y_1 \cdots y_n\}$

$$\min_{\mu}\{\sum_{i=1}^{n} |y_i - \mu|\} \iff \min_{\mu,a,b}\{\sum_{i=1}^{n} a_i + b_i\}$$

subject to $a_i, b_i \geq 0$ and $y_i - \mu = a_i - b_i \quad \forall 1, \cdots, n$

- Turn into Linear Programming Problem.
- To illustrate, consider a lognormal sample.

Building Quantile Regression

Median

- Pre-built Function

```
n=101
set.seed(32420)
y = rlnorm(n)
median(y)
[1] 1.03435
```

Building Quantile Regression

Median

- Solve by Linear Programming

```
library(lpSolve)
A1 = cbind(diag(2*n),0)
A2 = cbind(diag(n), -diag(n), 1)
r = lp("min", c(rep(1,2*n),0),
        rbind(A1, A2),c(rep(">=", 2*n),
        rep("=", n)), c(rep(0,2*n), y))
tail(r$solution,1)
[1] 1.03435
```

Building Quantile Regression

Quantile

- Now Change the equation into quantile regression,

$$\min_{\mu,a,b}\{\sum_{i=1}^{n}\tau a_i + (1-\tau)b_i\}$$

subject to $a_i, b_i \geq 0$ and $y_i - [\beta_0^\tau + \beta_1^\tau x_i] = a_i - b_i \quad \forall\, 1, \cdots, n$

Building Quantile Regression

Quantile

- Pre-built Function

```
tau = .3
quantile(y,tau)
30% 0.5775248
```

- Solve by Linear Programming

```
A1 = cbind(diag(2*n),0)
A2 = cbind(diag(n), -diag(n), 1)
r = lp("min", c(rep(tau,n),rep(1-tau,n),0),
        rbind(A1, A2),c(rep(">=", 2*n),
        rep("=", n)), c(rep(0,2*n), y))
tail(r$solution,1)
30% 0.5775248
```

Building Quantile Regression

Quantile

- Other than above programming, there are several more approaches.
- If interested, check out more on this page.
  Quantile Regression Computation: From the Inside and the Outside
- Instead building your own function, we can use R package "quantreg".

Outlier Protection on Median Regression

Basic Setting

- Import necessary packages.

```
library(MASS)
library(quantreg)
set.seed(32420)
```

- Same seed will be used repeatedly.
- $n_1 = 1000$ and $n_2 = 200$ for this trial.

OUTLIER PROTECTION ON MEDIAN REGRESSION

Generate Data Sample

- Generate n_1 of Good Observation.

```
data1 <- mvrnorm(n=n1,mu=c(0,0),Sigma =
matrix(c(1,0.8,0.8,1),ncol=2))
```

- Generate n_2 of Bad Observation.

```
data2 <- mvrnorm(n=n2,mu=c(1.5,-1.5), Sigma
= .2*diag(c(1,1))))
```

Outlier Protection on Median Regression

Generate Data Sample

- Bind the Data and turn it into data.frame.

```
data <- rbind(data1,data2)
data <- data.frame(data)
```

- Distinguish the Good and Bad using Indicator vector.

```
names(data) <- c("X","Y")
ind <- c(rep(1,n1),rep(2,n2))
```

Outlier Protection on Median Regression

Drawing Scatter Plot

- Draw plot of the generated Data Sample.

```
plot(Y ~ X,data,pch=c("x","o")[ind],
     col=c("black","red")[ind],
     main=paste("N1 =",n1,"N2 =",n2))
```

Outlier Protection on Median Regression

Data Plot

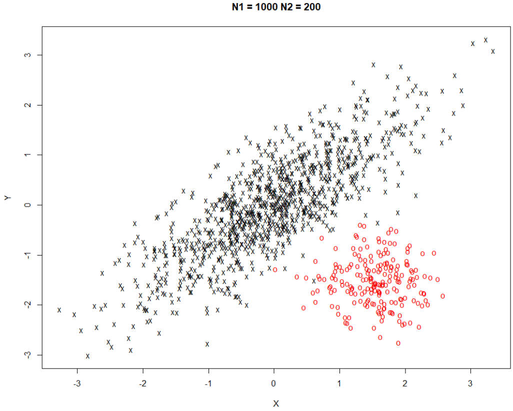


Figure 2: Scatter Plot

Outlier Protection on Median Regression

Data Plot

- Fit and Draw Quantile Regression of τ=0.5.

```
r1 <- rq(Y~X,data=data,tau=0.5)
abline(r1)
summary(r1)
```

- See Summary if you need.

Outlier Protection on Median Regression

Drawing Quantile Regression Plot

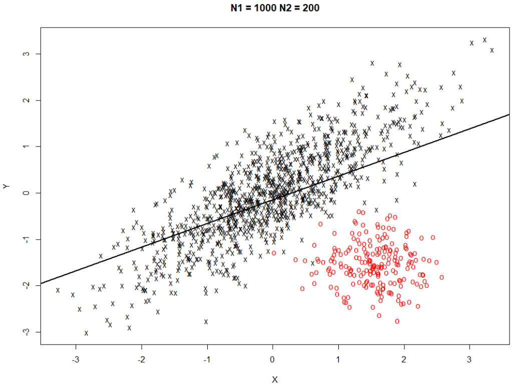


Figure 3: Plot with QR

Outlier Protection on Median Regression

Comparison Data Plot

- Draw Ordinary Least Square.

```
abline(lm(Y~X,data),lty=2, col="red")
```

- Draw Ordinary Least Square on Good.

```
abline(lm(Y~X,data,subset= 1:n1),
          lty=1,col='blue'))
```

- Draw Topleft Index.

```
legend("topleft",c("L1","ols","ols on good"),
        inset=0.02, lty=c(1,2,1),
        col=c("black","red","blue"),cex=.9)
```

Outlier Protection on Median Regression

Data Plot

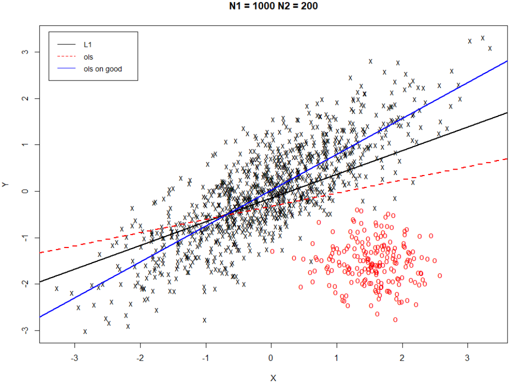


Figure 4: Final Comparison Plot

Outlier Protection on Median Regression

Comparison Data Plot

- Produce 4 simulation plot with difference amount of bad observation.
- By observing several graph, we can guess how much outlier can be bear-ed on median regression.
- In this simulation, 1000 good observations could bear till 371 bad observations.

Outlier Protection on Median Regression

Merge into One Function

```
qresult <- function(n1,n2){
  data <- mvrnorm(n=n1,mu=c(0,0),
          Sigma = matrix(c(1,0.8,0.8,1),ncol=2))
  data <- rbind(data,mvrnorm(n=n2,mu=c(1.5,-1.5),
          Sigma = .2*diag(c(1,1))))
  data <- data.frame(data) names(data) <- c("X","Y")
  ind <- c(rep(1,n1),rep(2,n2))
  plot(Y~X,data,pch=c("x","o")[ind], col=c("black","red")[ind],
       main=paste("N1 =",n1,"N2 =",n2))
  summary(r1 <- rq(Y~X,data=data,tau=0.5))
  abline(r1)
  abline(lm(Y~X,data),lty=2, col="red")
  abline(lm(Y~X,data,subset= 1:n1),lty=1,col='blue')
  legend("topleft",c("L1","ols","ols on good"),
          inset=0.02, lty=c(1,2,1),
          col=c("black","red","blue"),cex=.9)}
```

Outlier Protection on Median Regression

Data Plot

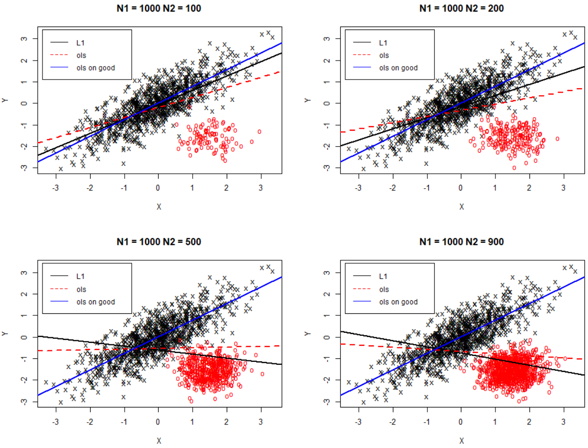


Figure 5: Final Comparison Plot

Outlier Protection on Median Regression

Data Plot

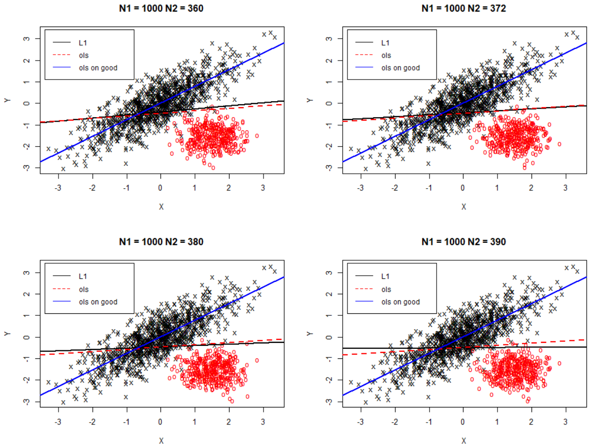


Figure 6: Final Comparison Plot

Outlier Protection on Median Regression

Heteroscedasticity

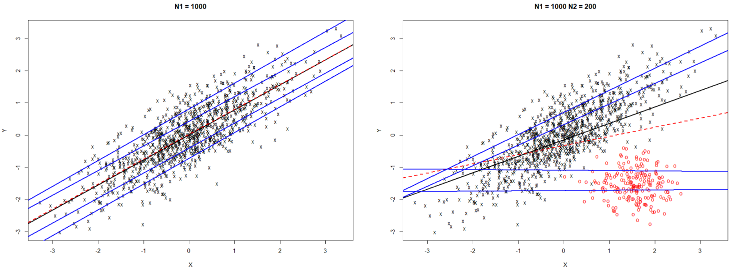
- Simply to check, Heteroscedasticity, check out the Quantile Regression line.
- If they look like having different slope, probably there will be Heteroscedasticity.

Predictive Interval

- For different quantiles, there will be different predictive intervals obvioulsy.
- By plotting quantiles, the accuracy of OLS interval can be checked.

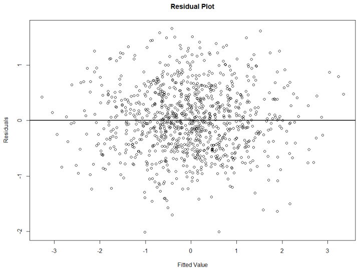
Outlier Protection on Median Regression

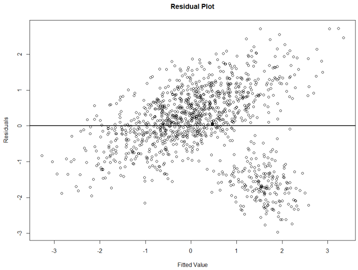
Heteroscedasticity



Outlier Protection on Median Regression

Heteroscedasticity - Residual Plot





Outlier Protection on Median Regression

Predictive Intervals

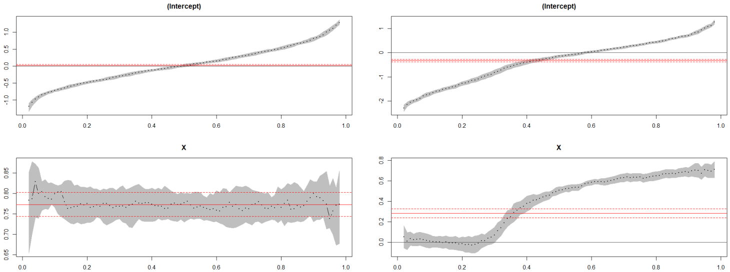
- Summary Plot of the Intercept and Coefficient.

```
plot(summary(rq(Y~X,data1,
        tau=2:98/100)))
```

- The horizontal line is the OLS estimate.
- Dashed lines for confidence interval for OLS estimate.

Outlier Protection on Median Regression

Predictive Intervals



Quantile Regression on Engel Data

Engel's Law

- As income rises, the proportion of income spent on food falls, even if absolute expenditure on food rises.
- Empirical Law based on observation.
- This example shows expenditures on food as a function of income for 19th century Belgian households.

Quantile Regression on Engel Data

Plot Data

- Engel Data is pre-included in the R package 'quantreg'.

```
data(engel)
```

- Draw the Scatter plot.

```
plot(foodexp~income, engel,
          cex=.5,xlab="Household Income",
          ylab="Food Expenditure")
```

- Draw the OLS on the plot.

```
abline(lm(foodexp~income,engel))
```

Quantile Regression on Engel Data

Data Plot

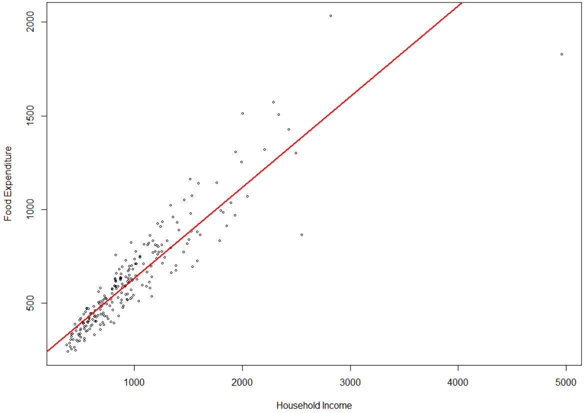


Figure 10: Scatter Plot and OLS

Quantile Regression on Engel Data

Plot Quantile Regression

- Draw Median Regression on the plot.

```
abline(rq(foodexp~income,engel,tau=.5),
col="blue")
```

- Draw the 10-20-75-90 Quantile Regression on the plot.

```
taus <- c(.1,.25,.75,.90)
for( i in 1:length(taus)){
abline(rq(foodexp~income,engel,tau=taus[i]),)
       col="gray")}
```

Quantile Regression on Engel Data

Data Plot

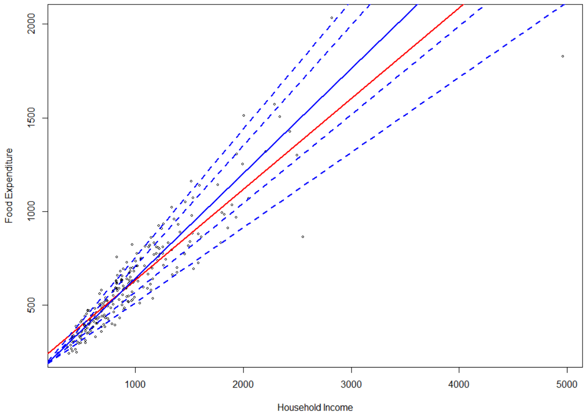


Figure 11: Quantile Regression

Quantile Regression on Engel Data

Plot Summary

- Summary Plot of the Intercept and Coefficient.

```
plot(summary(rq(foodexp~income,engel,
        tau=2:98/100)))
```

- The horizontal line is the OLS estimate.
- Dashed lines for confidence interval for OLS estimate.

Quantile Regression on Engel Data

Data Plot

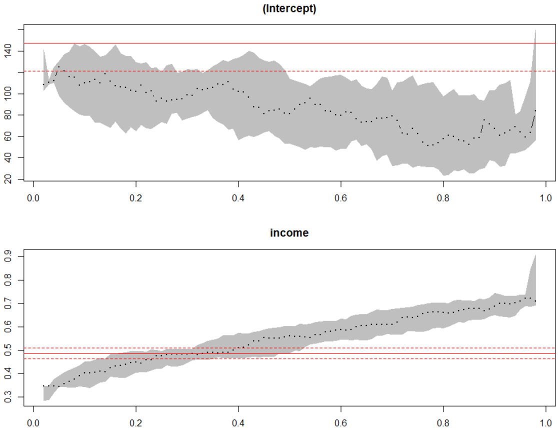


Figure 12: Summary Plot of Quantile Regression

Quantile Regression on Engel Data

Log Transformation Case of Engel Data

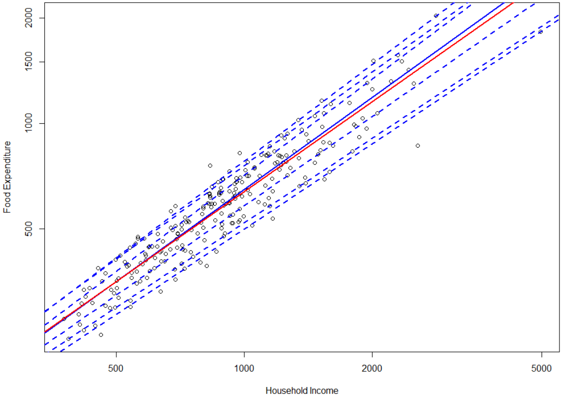


Figure 13: Summary Plot of Log10 Quantile Regression

Quantile Regression on Wage Data

Wage Data from ISLR

- R package 'ISLR' contains only data, for use of the book "Introduction to Statistical Learning with applications in R".
- Load Wage Data from the package 'ISLR'.

```
library(ISLR)
data(Wage)
```

- Rest of procedure will be same as the previous case.

Quantile Regression on Wage Data

Data Plot

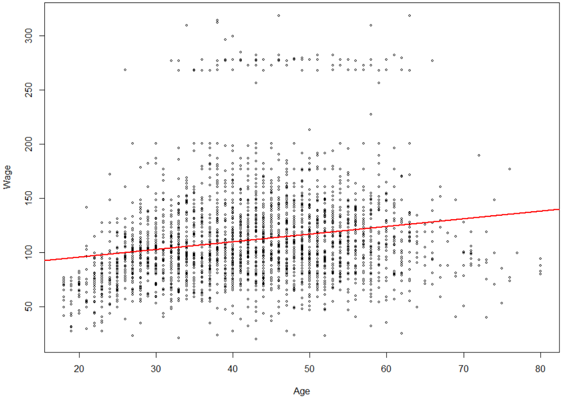


Figure 14: Scatter Plot and OLS

Quantile Regression on Wage Data

Data Plot

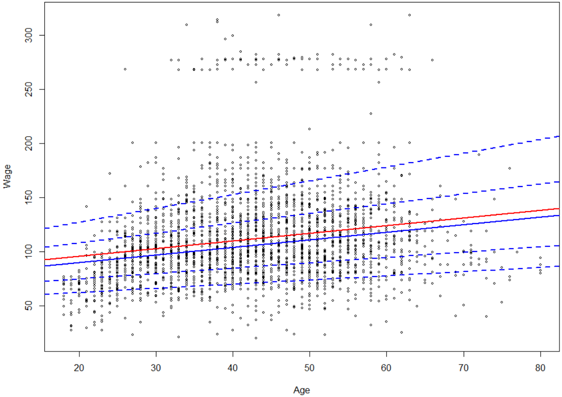


Figure 15: Quantile Regression

Quantile Regression on Wage Data

Data Plot

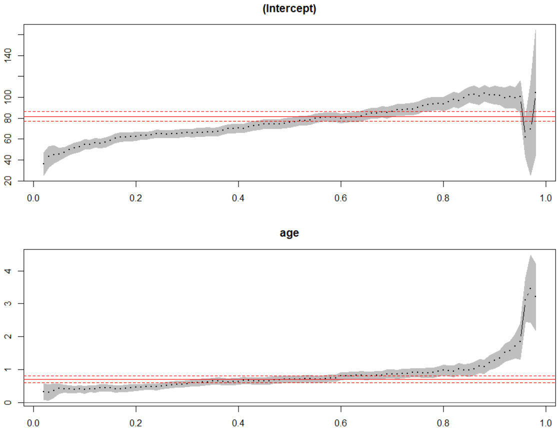


Figure 16: Summary Plot of Quantile Regression

Quantile Regression on Wage Data

Log Transformation Case of Wage Data

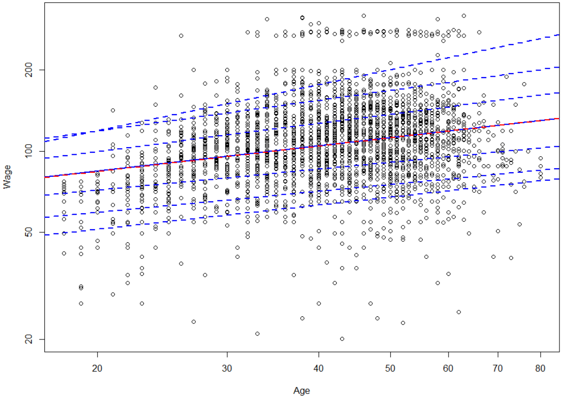


Figure 17: Summary Plot of Log10 Quantile Regression