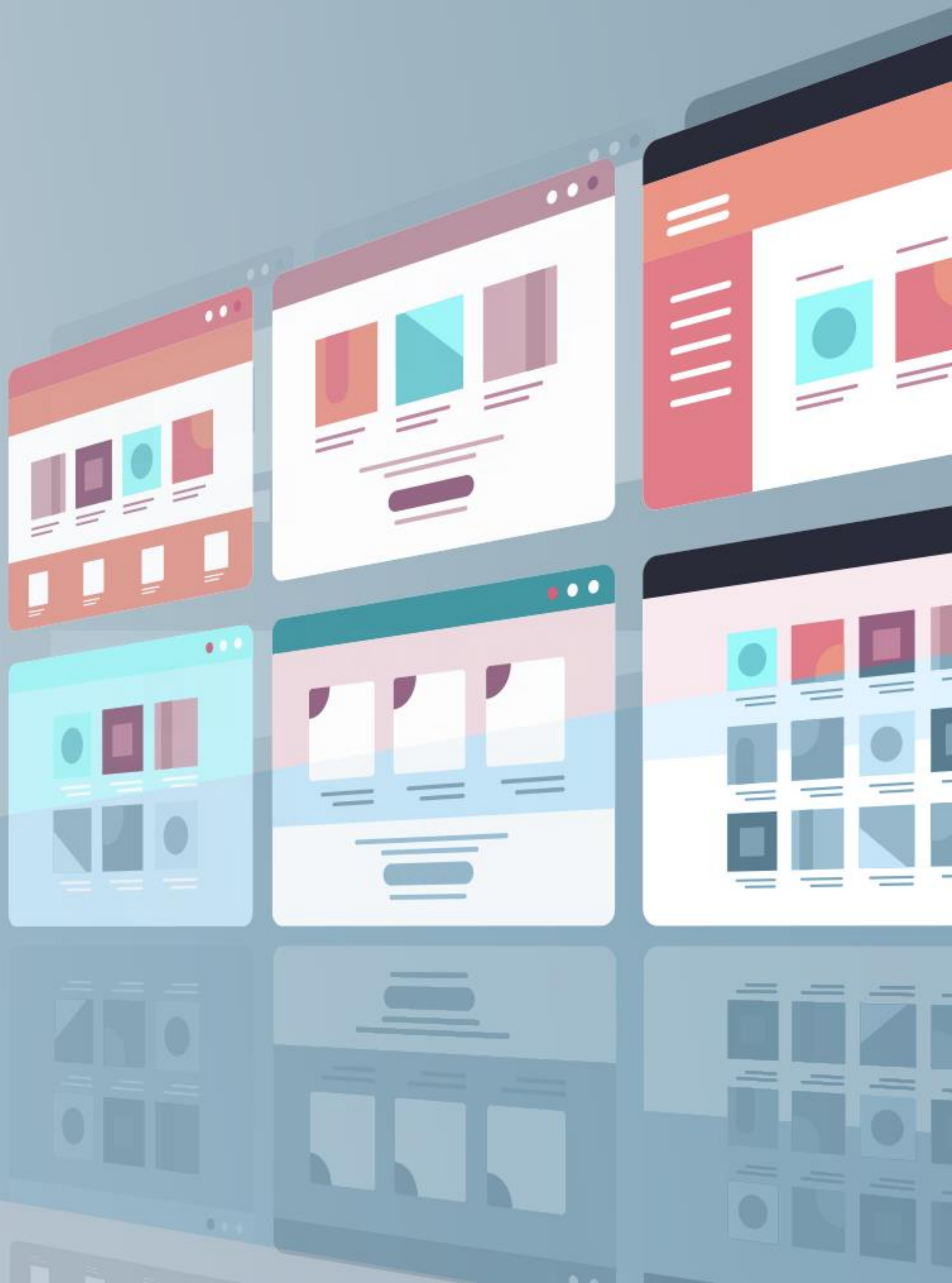
14

데이터분석방법론(1)

Quantile Regression

통계데이터과학과 장영재 교수



한국방송통신대학교 대학원

학습목차

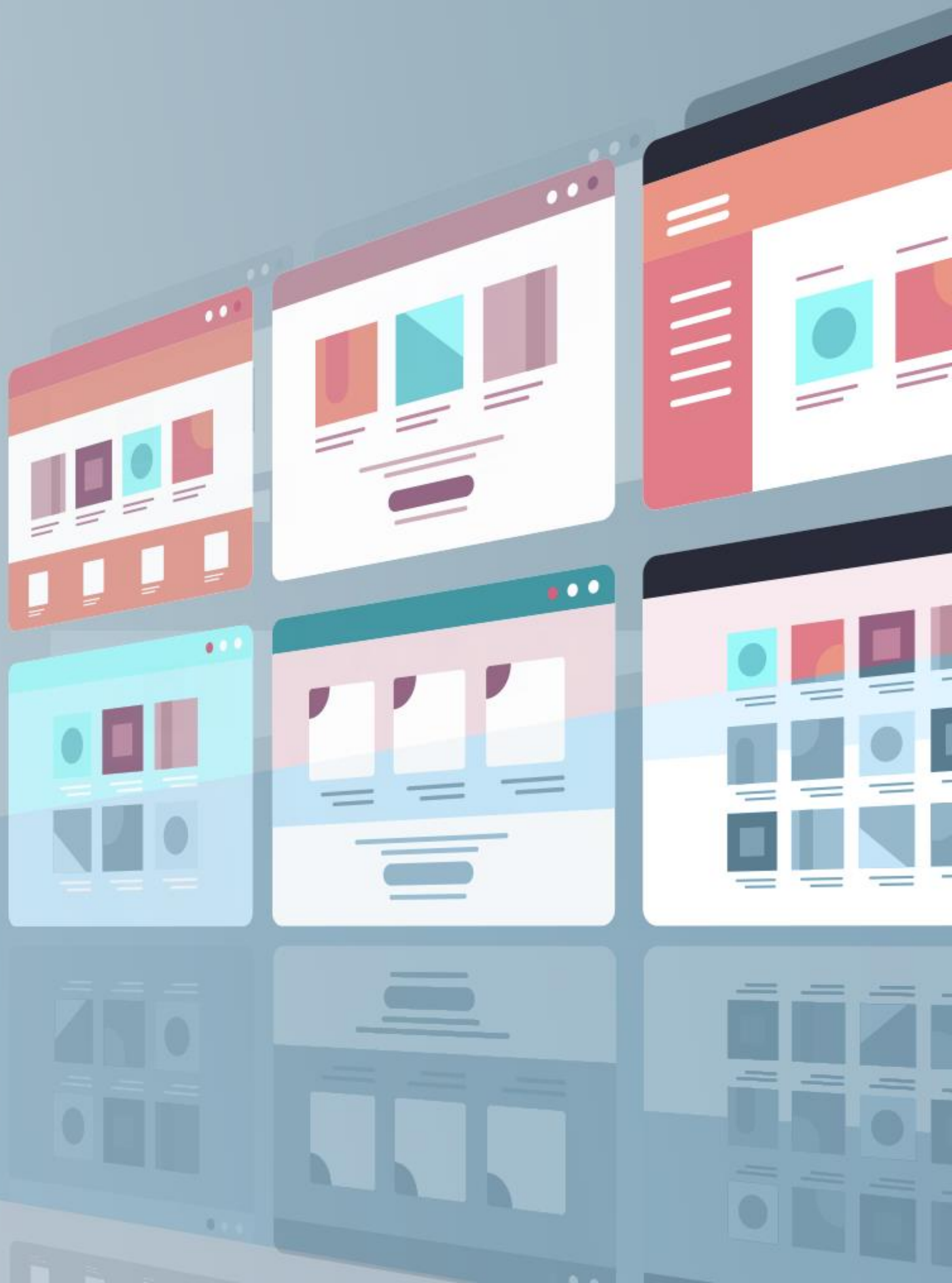
1 Introduction

2 The Model

3 Examples

01

Introduction



1. Regression and Quantile Regression

◆ Quantile regression (Koenker, 1978)

① Recall : Regression – relationship between a response variable y and a set of covariates x_i's.

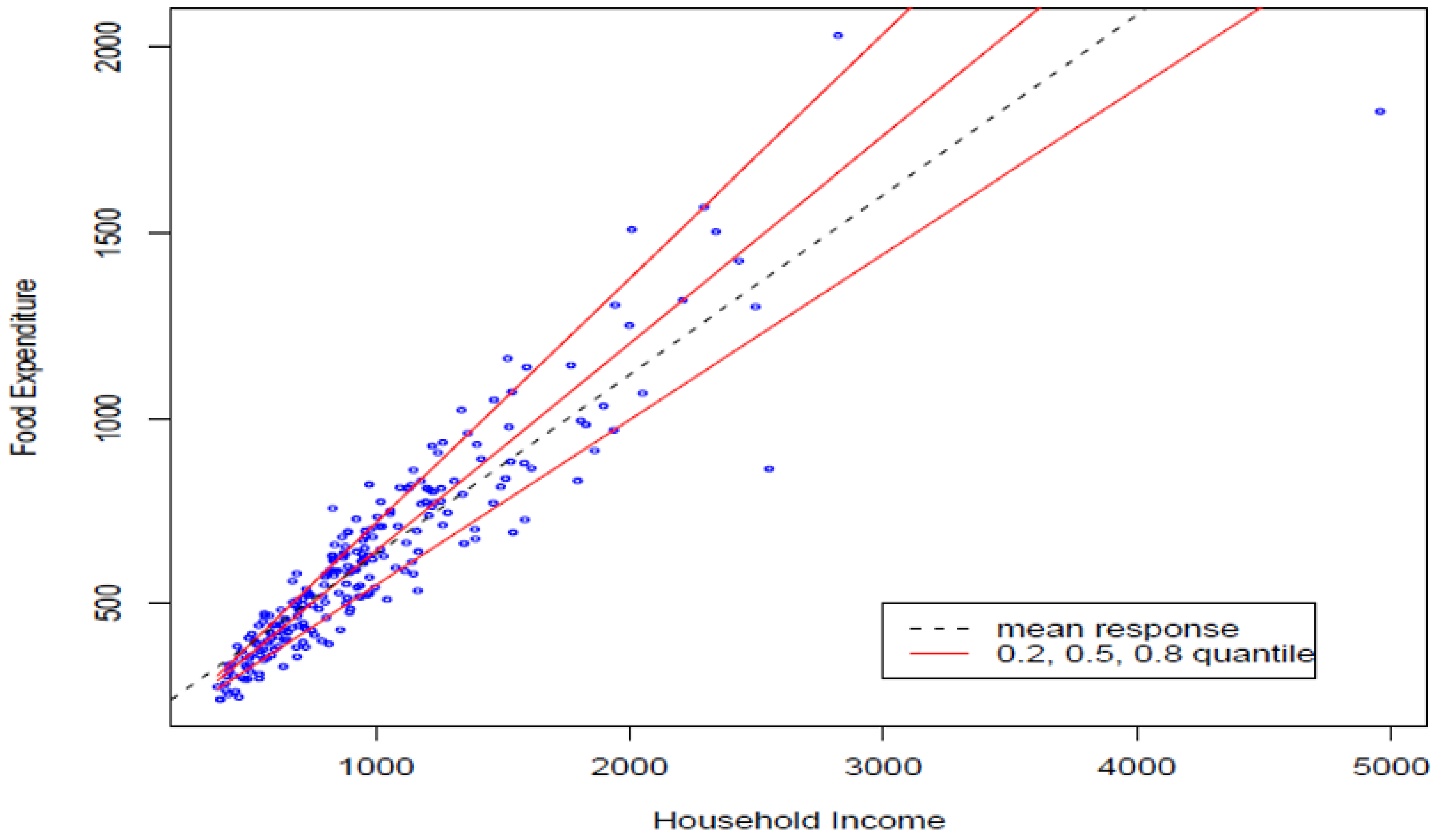
- Focus on how the mean of y changes with x_i's – Least squares.
- A mean curve.

② A single mean curve may not be informative enough for some cases

- A more complete view – Conditional quantile functions

1. Regression and Quantile Regression

◆ Comparison



2. Least Squares vs Least Absolute Deviation

Consider some linear model, $y_i = \beta_0 + x_i^T \beta + \varepsilon_i$, and define

$$(\hat{\beta}_0^{ols}, \hat{\beta}^{ols}) = argmin \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 \right\}$$

$$(\hat{\beta}_0^{lad}, \hat{\beta}^{lad}) = argmin \left\{ \sum_{i=1}^{n} \left| y_i - \beta_0 - x_i^T \beta \right| \right\}$$

02

The Model

1. Quantiles

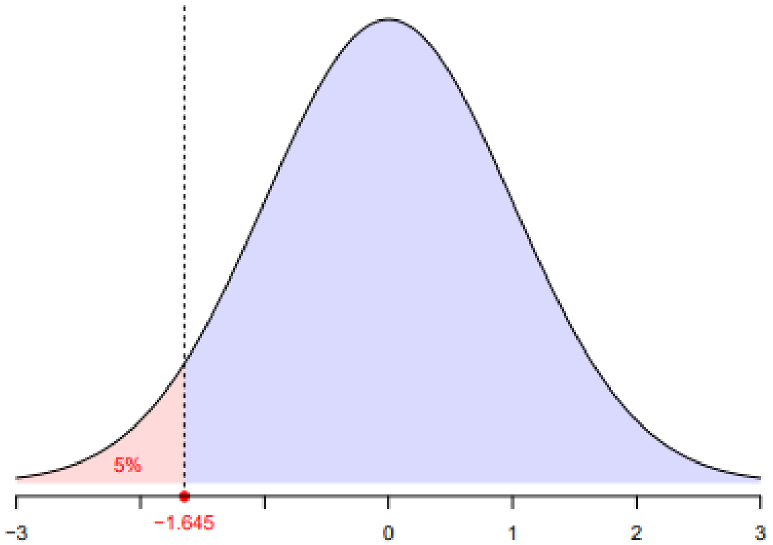
Let Y denote a random variable with cumulative distribution function F,

$F(y) = \mathbb{P}[Y \le y]$.

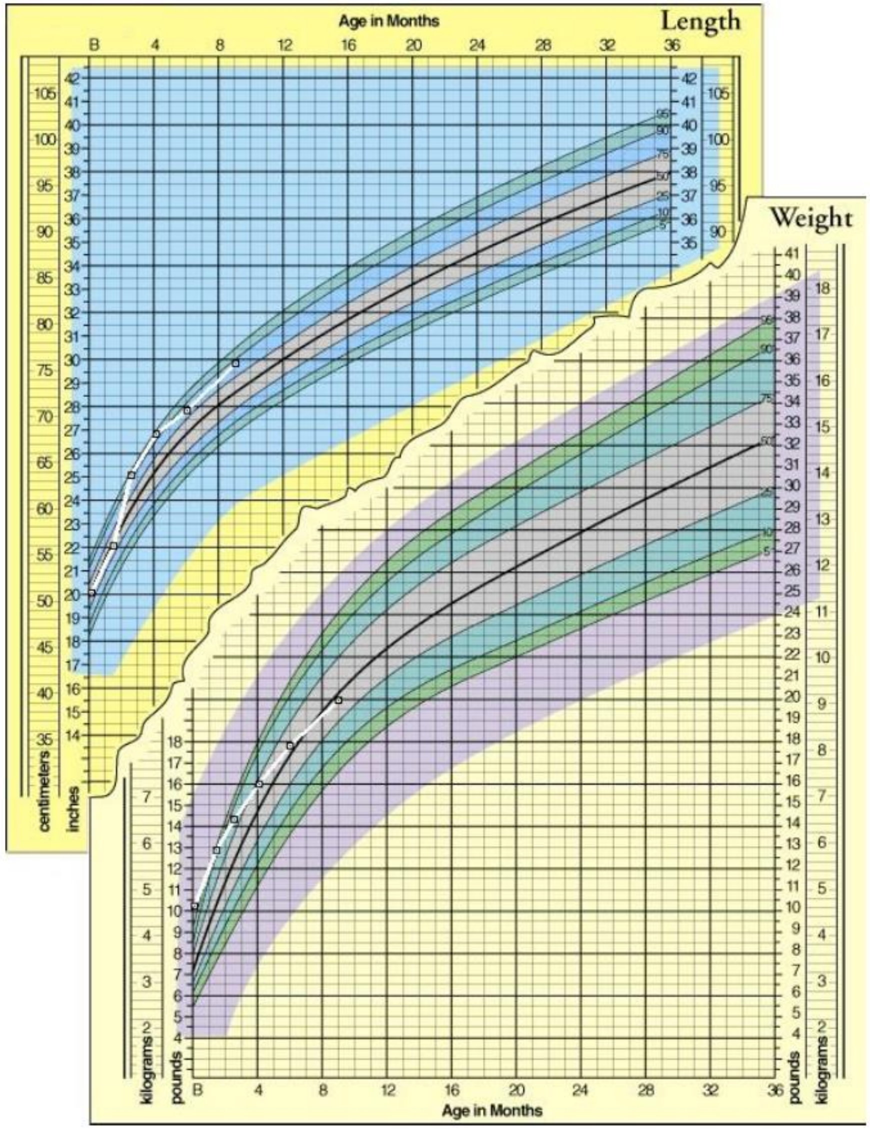
The quantile is $Q_\tau(x) = inf\,\{x \in R, F(x) > \tau\}$.

2. Quantiles and Quantile Regressions

Quantiles are important quantities in many areas (inequalities, risk, health, sports, etc).



Quantiles of the $N(0,1)$ distribution.



3. A First Model for Conditional Quantiles

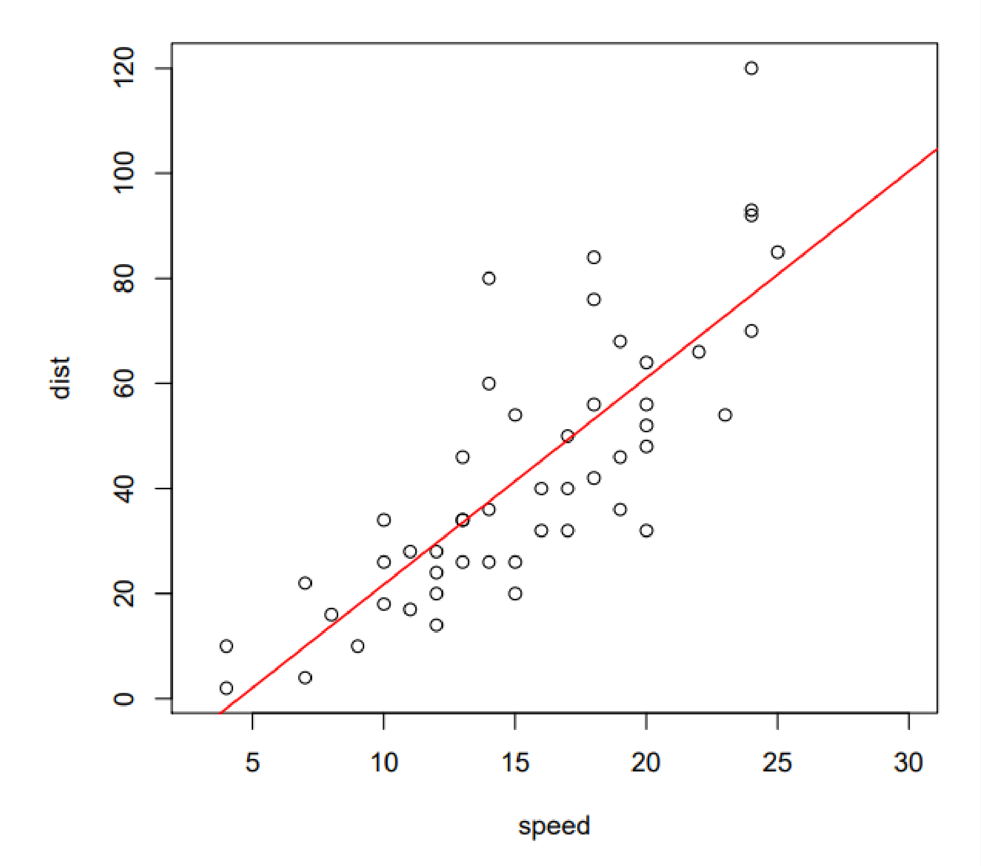
Consider a location model, $y = \beta_0 + x^T\beta + \varepsilon \qquad i.e.$

$$\mathbb{E}[Y|X = x] = x^T\beta$$

Then one can consider

$$Q(\tau\,|X = x) = \beta_0 + Q_\varepsilon\,(\tau\,) + \,x^T\beta$$

where $Q_\varepsilon\,(\cdot)$ is the quantile function of the residuals.



4. OLS vs. Median Regression (Least Absolute Deviation)

Consider some linear model, $y_i = \beta_0 + x_i^T\beta + \varepsilon_i,$ and define

$$(\hat{\beta}_0^{ols}, \hat{\beta}^{ols}) = argmin\left\{\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2\right\}$$

$$(\hat{\beta}_0^{lad}, \hat{\beta}^{lad}) = argmin\left\{\sum_{i=1}^{n}\left|y_i - \beta_0 - x_i^T\beta\right|\right\}$$

Assume that $\varepsilon \mid X$ has a symmetric distribution, $\mathbb{E}[\,\varepsilon \mid X\,] = median[\,\varepsilon \mid X\,] = 0,$ then $(\hat{\beta}_0^{ols}, \hat{\beta}^{ols})$ and $(\hat{\beta}_0^{lad}, \hat{\beta}^{lad})$ are consistent estimators of (β_0, β).

Assume that $\varepsilon \mid X$ does not have a symmetric distribution, but $\mathbb{E}[\,\varepsilon \mid X\,] = 0,$ then $\hat{\beta}^{ols}$ and $\hat{\beta}^{lad}$ are consistent estimators of the slopes β.

If $median[\,\varepsilon \mid X\,] = \gamma$, then $\hat{\beta}_0^{lad}$ converges to $\beta_0 + \gamma$

6. Quantile regression

In OLS regression, we try to evaluate $\mathbb{E}[\,Y \mid X = x\,] = \int_{\mathbb{R}} \; y dF_{Y|X=x}(y)$

In quantile regression, we try to evaluate

$$Q_\tau(Y|X = x) = \inf\{y : F_{Y|X=x}(y) \geq \tau\}$$

as introduced in Newey & Powell (1987) Asymmetric Least Squares Estimation and Testing.

Li & Racine (2007) Nonparametric Econometrics: Theory and Practice suggested

$$\hat{Q}_\tau(Y|X = x) = \inf\{y : \hat{F}_{Y|X=x}(y) \geq \tau\}$$

Where $\hat{F}_{Y|X=x}(y)$ can be some kernel-based estimator.

6. Quantile regression

Summary

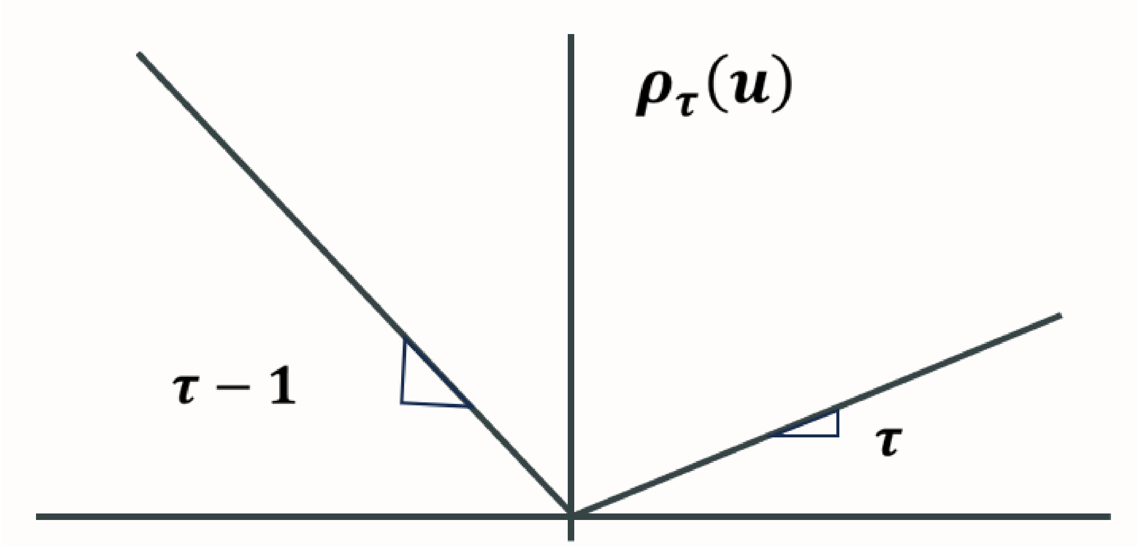
- Let Y be a dependent variable, X a (d-dimensional) predictor variable,
  - $Q_\tau\,(Y|X=x) = inf\,\{y\text{:}\,\mathrm{F}(y|\mathrm{X}=x\} \geq \tau\}$

    where $\mathrm{F}(y|\mathrm{X}=x) = \mathrm{P}(\mathrm{Y} \leq y|\mathrm{X}=x)$.
  - Conditional quantile which minimizes the expected loss E(ρ_τ)
  - $\beta_\tau = argmin_{\beta \in R^d} E\left(\rho_\tau(\mathrm{Y} - \mathrm{x}'\upbeta)\right)$,

    where $\rho_\tau(u) = u\left(\tau - I(u<0)\right)$.

7. Asymmetric Loss Function

- Loss function : asymmetric in general.
- Example of $\tau = 0.3$: large negative errors are more heavily penalized (than positive errors).



8. Estimates of Quantile Regression

- The estimate for a given quantile τ is $\hat{Q}_{\tau}\,(\mathrm{X} = x) = x'\hat{\beta}_{\tau}(x)$,

Where $\hat{\beta}_{\tau}(x) = argmin_{\beta \in R^{d}} \sum \rho_{\tau}(y_i - x_i'\beta)$,

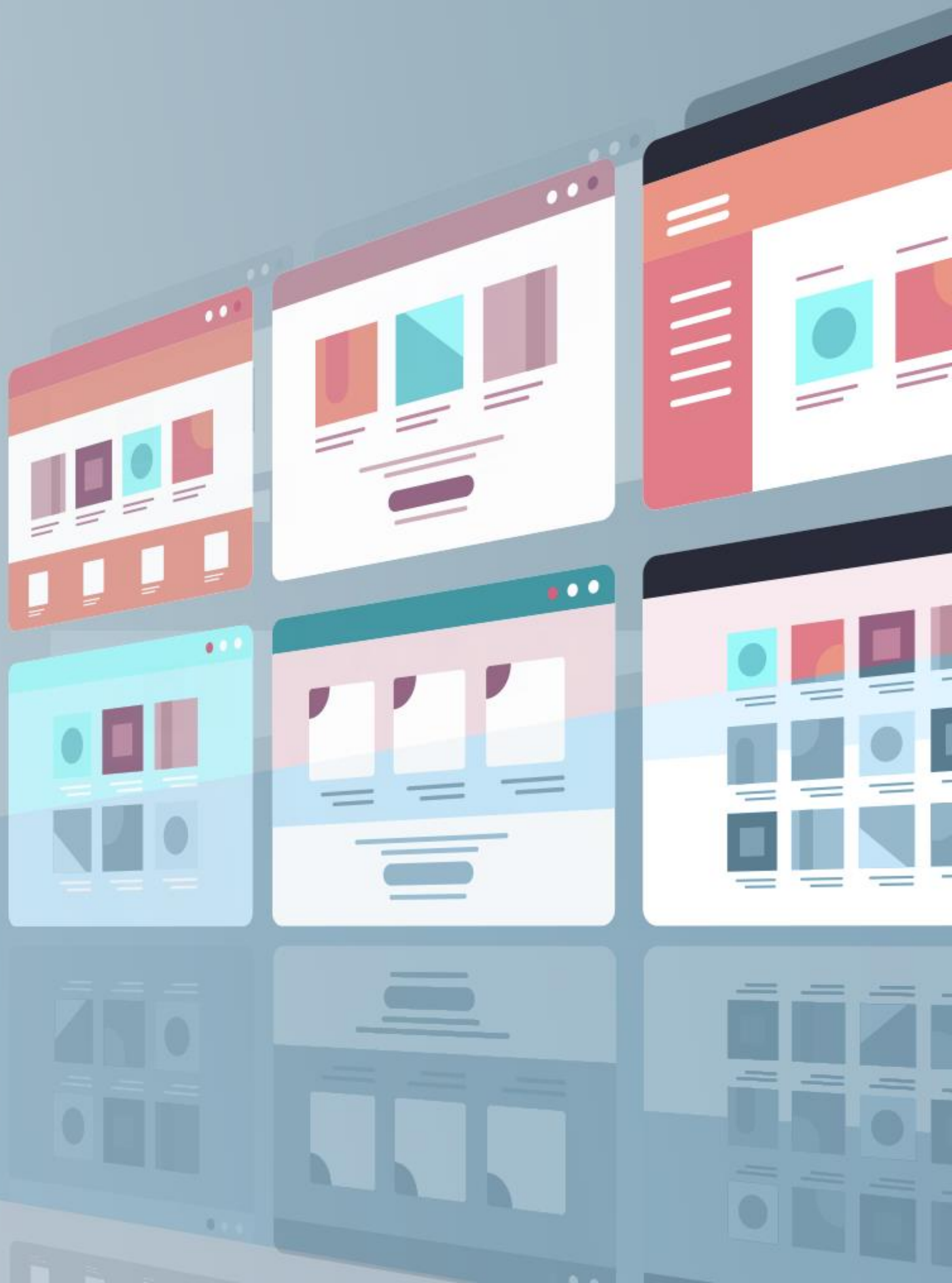
with $\rho_{\tau}(u) = u(\tau - I(u < 0))$

- Sum of error losses related with $\rho_{\tau}(u)$,

SEL = $\sum \rho_{\tau}\left(y_i - x_i'\hat{\beta}\right)$

03

Examples



1. Geometric Properties of the Quantile Regression

Observe that the median regression will always have two supporting observations.

Start with some regression line, $y_i = \beta_0 + \beta_1 x_i$

Consider small translations $y_i = (\beta_0 \pm \epsilon) + \beta_1 x_i$

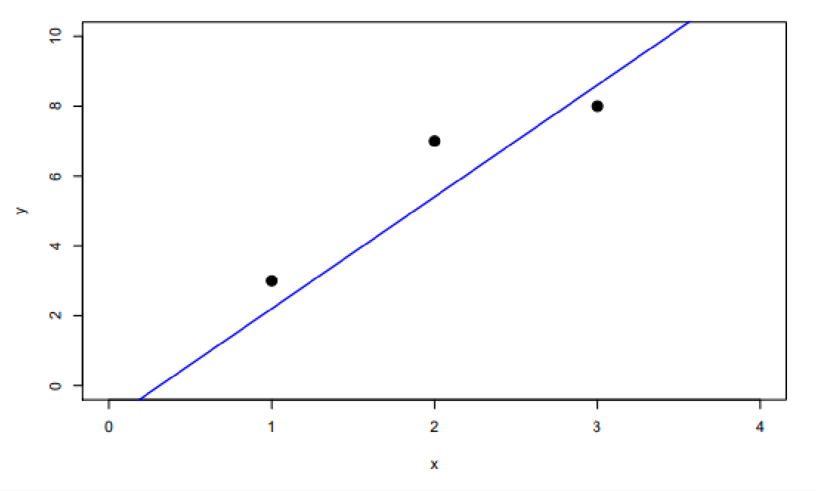
We minimize $\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$

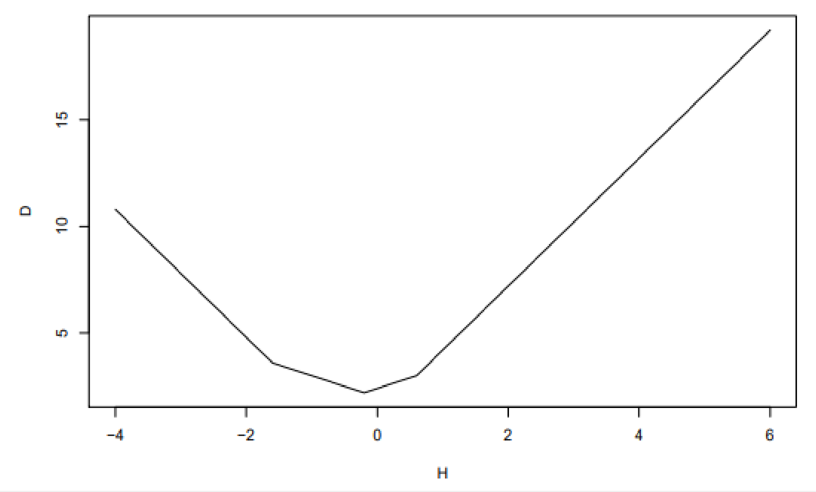
From line blue, a shift up decrease the sum by ϵ until we meet point on the left

an additional shift up will increase the sum

We will necessarily pass through one point

(observe that the sum is piecewise linear in ϵ)





1. Geometric Properties of the Quantile Regression

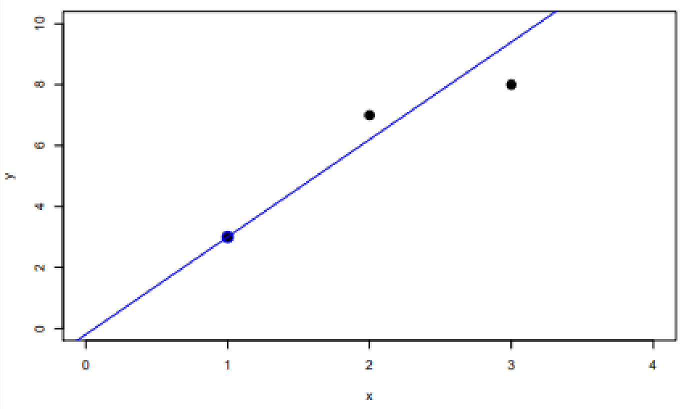
Consider now rotations of the line around the support point

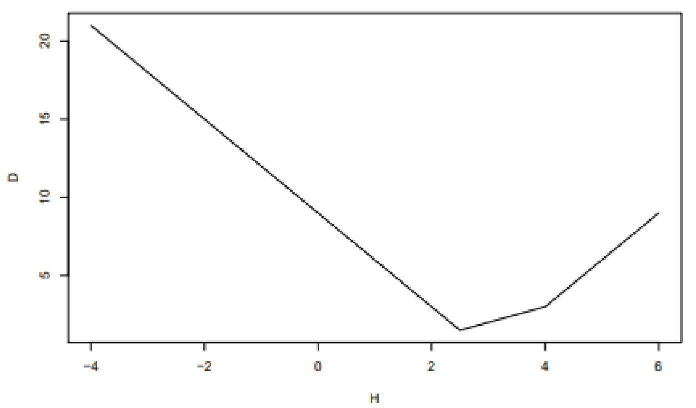
If we rotate up, we increase the sum of absolute difference (large impact on the point on the right)

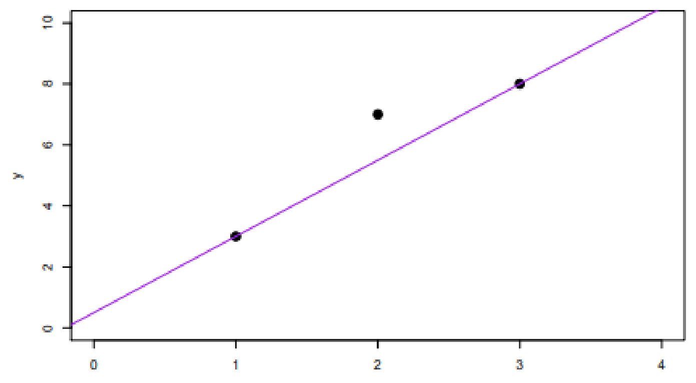
If we rotate down, we decrease the sum, until we reach the point on the right

Thus, the median regression will always have two supporting observations.

```
> library(quantreg)
> fit <- rq(dist~speed, data=cars, tau=.5)
> which(predict(fit)==cars$dist)
 1 21 46
 1 21 46
```





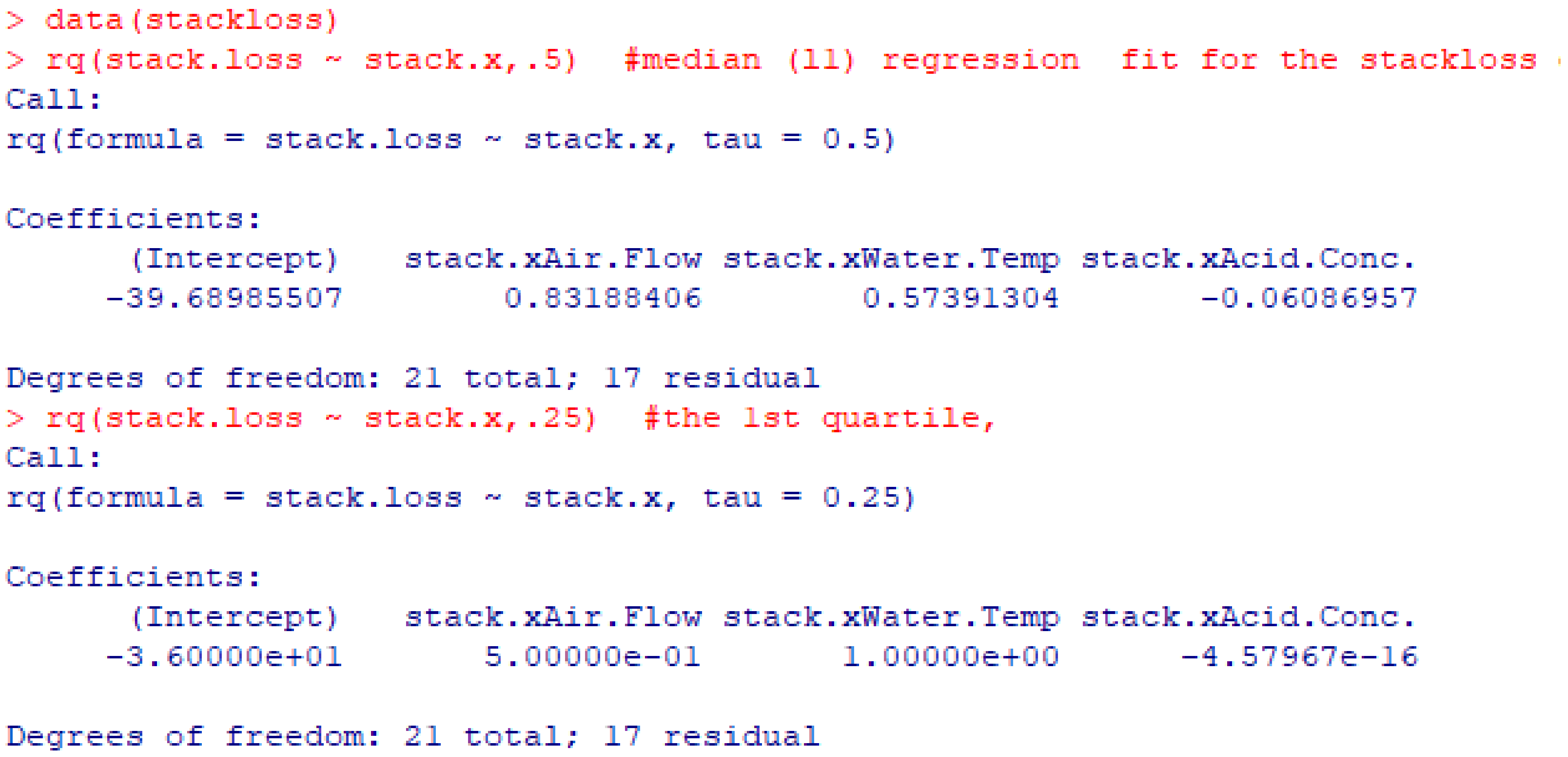


3. Example (R : quantreg package and rq function)

◆ Stackloss data

```
> install.packages("quantreg")
> library(quantreg)
> data(stackloss)  # Operational data of a plant for the
oxidation of ammonia to nitric acid
> rq(stack.loss ~ stack.x,.5)  #median (l1) regression
> rq(stack.loss ~ stack.x,.25)  #the 1st quartile
> rq(stack.loss ~ stack.x, tau=-1)   #full set of quantiles
```

3. Example



```
> data(stackloss)
> rq(stack.loss ~ stack.x,.5)  #median (l1) regression  fit for the stackloss
Call:
rq(formula = stack.loss ~ stack.x, tau = 0.5)

Coefficients:
      (Intercept)   stack.xAir.Flow stack.xWater.Temp stack.xAcid.Conc.
     -39.68985507        0.83188406        0.57391304       -0.06086957

Degrees of freedom: 21 total; 17 residual
> rq(stack.loss ~ stack.x,.25)  #the 1st quartile,
Call:
rq(formula = stack.loss ~ stack.x, tau = 0.25)

Coefficients:
      (Intercept)   stack.xAir.Flow stack.xWater.Temp stack.xAcid.Conc.
     -3.60000e+01       5.00000e-01       1.00000e+00      -4.57967e-16

Degrees of freedom: 21 total; 17 residual
```

3. Example

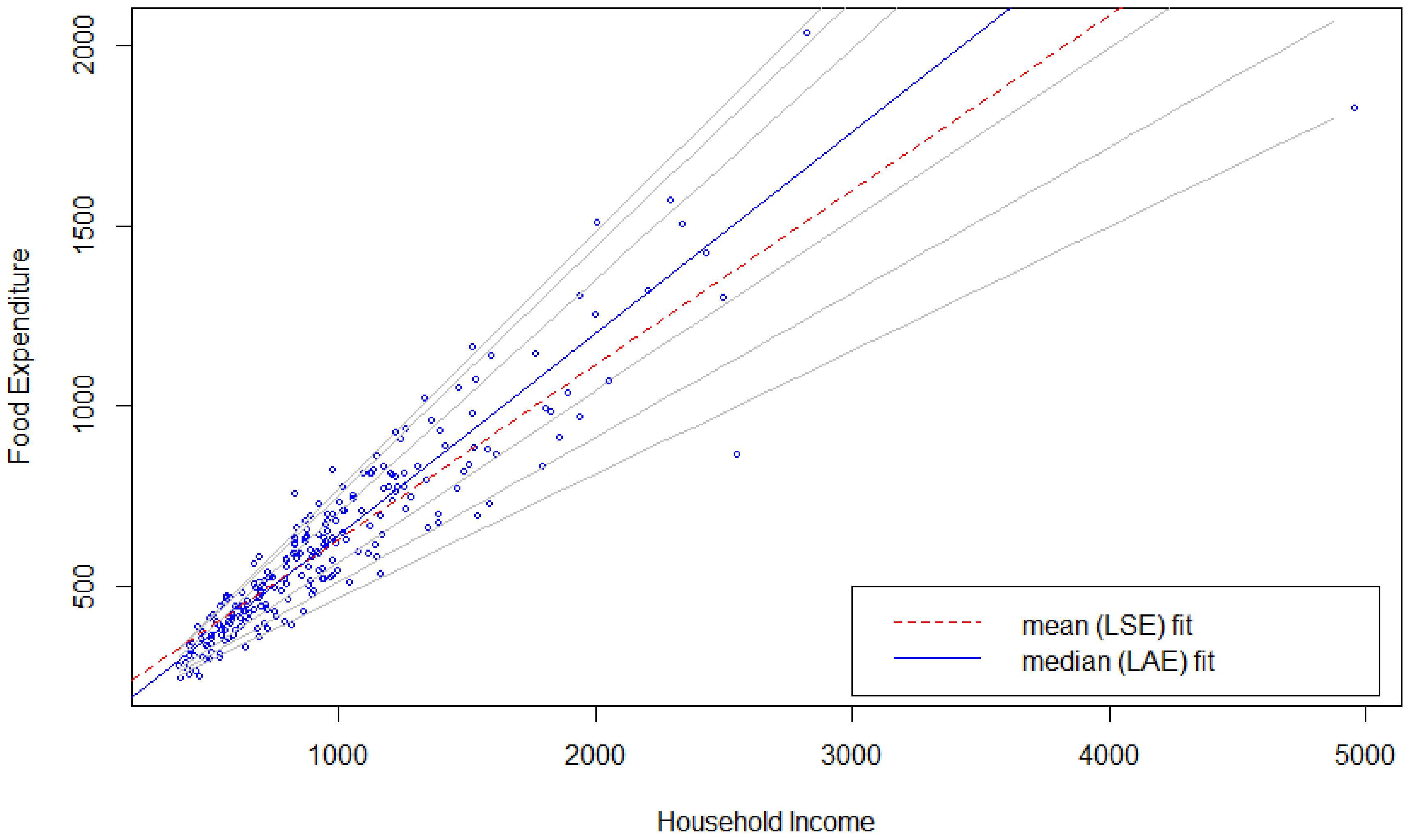
◆ Engel data

```
#plot of engel data and some rq lines see KB(1982) for
references to data

> install.packages("quantreg")
> library(quantreg)
> data(engel)
> attach(engel)
> plot(income,foodexp,xlab="Household Income",ylab="Food
Expenditure",type = "n", cex=.5)
> points(income,foodexp,cex=.5,col="blue")
> taus <- c(.05,.1,.25,.75,.9,.95)
> xx <- seq(min(income),max(income),100)
> f <- coef(rq((foodexp)~(income),tau=taus))
```

3. Example

```
> yy <- cbind(1,xx)%*%f
for(i in 1:length(taus)){
        lines(xx,yy[,i],col = "gray")
        }
> abline(lm(foodexp ~ income),col="red",lty = 2)
> abline(rq(foodexp ~ income), col="blue")
> legend(3000,500,c("mean (LSE) fit", "median (LAE) fit"),
      col = c("red","blue"),lty = c(2,1))

#Example of plotting of coefficients and their confidence bands
> plot(summary(rq(foodexp~income,tau = 1:49/50,data=engel)))
```

3. Example


Food Expenditure
500
1000
1500
2000
1000
2000
3000
4000
5000
Household Income
mean (LSE) fit
median (LAE) fit

다음시간 안내

끝

수고하셨습니다.