

## Quantile, Quantile Regression

- Quantiles
- Quantiles Based Volatility Forecasts
- Quantile Regression
- Definition and Applications
- Estimation Methods and Softwares
- Bayesian Quantile Regression

## 1. What is Quantile?

The term quantile is synonymous with percentile and percentage point.

The median is an example of a quantile.

More generally, the 25% and 75% sample quantiles can be defined as values that split the data into proportions of one- and three-quarters, and vice-versa.

Correspondingly, in the continuous case the population lower quartile and upper quartile are the solutions to the equations  $F(y) = \frac{1}{4}$  and  $F(y) = \frac{3}{4}$  respectively. Generally, in the continuous case the  $p$ th quantile of  $F$  is the value  $y$  which solves  $F(y) = p$ . Or it is the inverse function  $F^{-1}(p)$ .

## 1.1 Quantiles Based Volatility Forecasts

Pearson and Tukey (1965, *Biometrika*, 52, 533–546) found that the ratio of the standard deviation to the interval between symmetric quantiles,  $Q(\theta)$  and  $Q(1 - \theta)$ , in the tails of the distribution is remarkably constant for a variety of distributions.

$$\sigma = \frac{Q(0.99) - Q(0.01)}{4.65},$$

$$\sigma = \frac{Q(0.975) - Q(0.025)}{3.92},$$

$$\sigma = \frac{Q(0.95) - Q(0.005)}{3.25}$$

Taylor (2005, Management Sciences, 51, 712-725) proposed a two-step volatility forecasting

Use one step-ahead quantile forecasts,  $\hat{Q}(\theta)$  and  $\hat{Q}(1 - \theta)$ , to get one step-ahead volatility forecasts,

$$\hat{\sigma}_{t+1}^2 = \alpha + \beta(\hat{Q}_{t+1}(1 - \theta) - \hat{Q}_{t+1}(\theta))^2$$

Where  $\alpha$  and  $\beta$  are the parameters estimated by the LS regression.

In fact, for daily returns, if we assume that there is no autocorrelation between successive daily shocks, then, for the holding period of duration  $k$  days starting in period  $t + 1$ , the realised multiperiod variance can be calculated as

$$\sigma_{t,k}^2 = \sum_{i=1}^k \epsilon_{t+i}^2$$

then we can estimate regression via fitting model

$$\hat{\sigma}_{t,k}^2 = \alpha_k + \beta_k (\hat{Q}_{t+1}(1 - \theta) - \hat{Q}_{t+1}(\theta))^2$$

where  $\alpha_k$  and  $\beta_k$  are the parameters estimated by the LS regression.

## 2. What is Quantile Regression?

- Regression measures how the mean value of the response variable  $Y$  varies with co-variates  $X$ .
- The classical theory of **linear regression models** aims at estimating  $m(x) \equiv E(Y|X = x)$ .
- Method: Least-squares minimization or LS regression.
- Suitability: Gaussian errors, symmetric distributions.
- Weakness: Conditional skew distribution, outliers, tail behavior.

- Quantile regression measures how the quantiles of  $Y$  vary with  $X$ .

Let  $q_p(x)$  denote a  $p$ th ( $0 < p < 1$ ) quantile regression.

- Aim: explore a complete relationship between  $Y$  and  $X$  and check the tails of conditional distributions.
- Specific cases: median regression ( $p = 0.5$ ).

An example of a simple linear quantile regression: when  $(X, Y) \sim$  bivariate normal distribution  $N(0, 0, r, 1, 1)$ ,

$$q_p(x) = rx + (1 - r^2)\Phi^{-1}(p);$$

where  $\Phi^{-1}(p)$  denotes the inverse of standard normal distribution. See Fig. 0:  $q_{0.1}(x)$ ,  $q_{0.5}(x)$  and  $q_{0.9}(x)$  from  $N(0, 0, 0.75, 1, 1)$ .

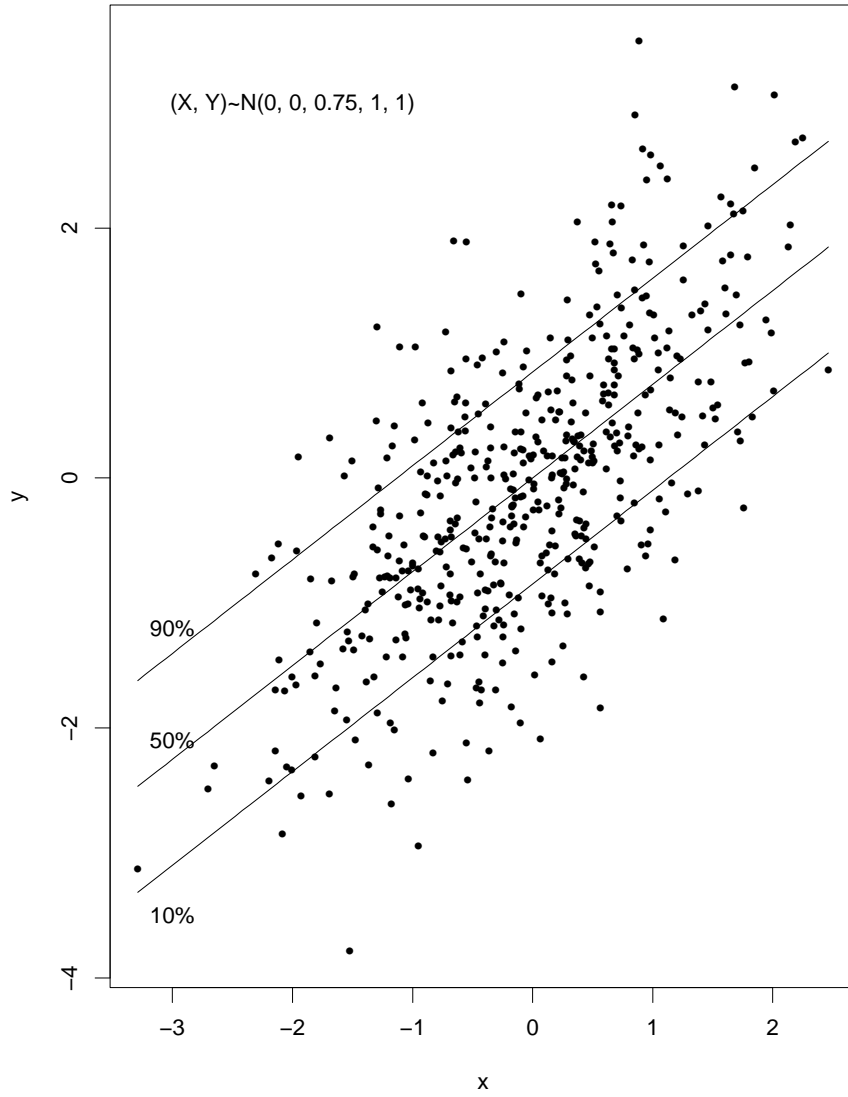
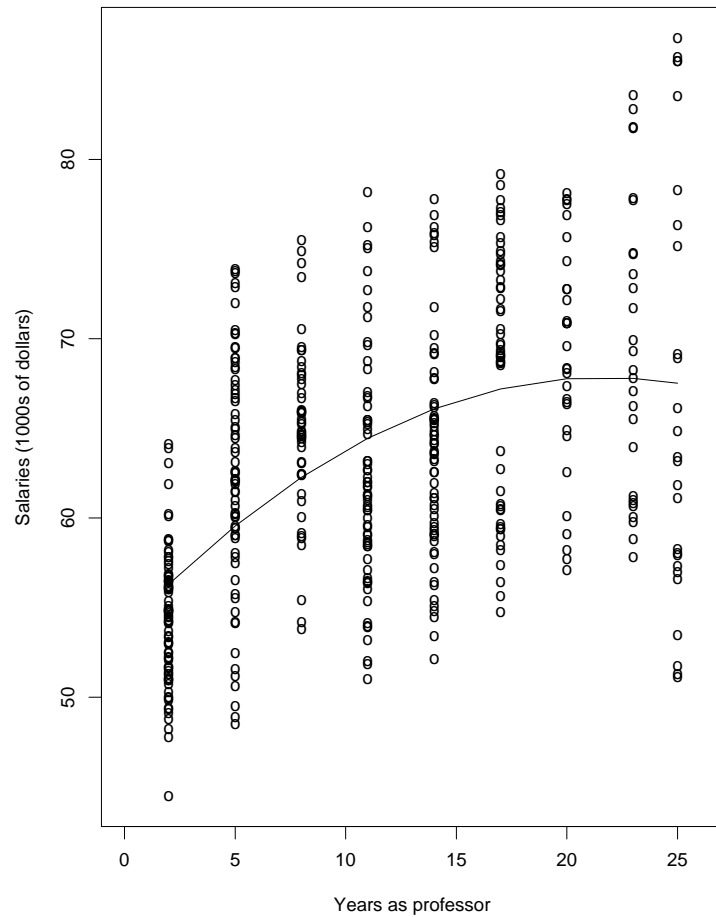


Fig. 0

**Example 2.1:** Some years ago a university union wished to examine the relationship between the earnings of professors and the number of years they had been professor. The union collected data on the salaries of 459 US statistics professors and the number of years for which they had been professors during the period from 1980 to 1990; see Bailar (1991, Amstat News, 182). A standard linear regression model for this is

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon, \quad (1)$$

where  $\mathbf{x} = (1, x)^T$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ ,  $y$  is salary,  $x$  is the number of years as professor and  $\epsilon$  is a Gaussian-error. In Figure 1 we present the data, together with the best fitting quadratic regression curve.

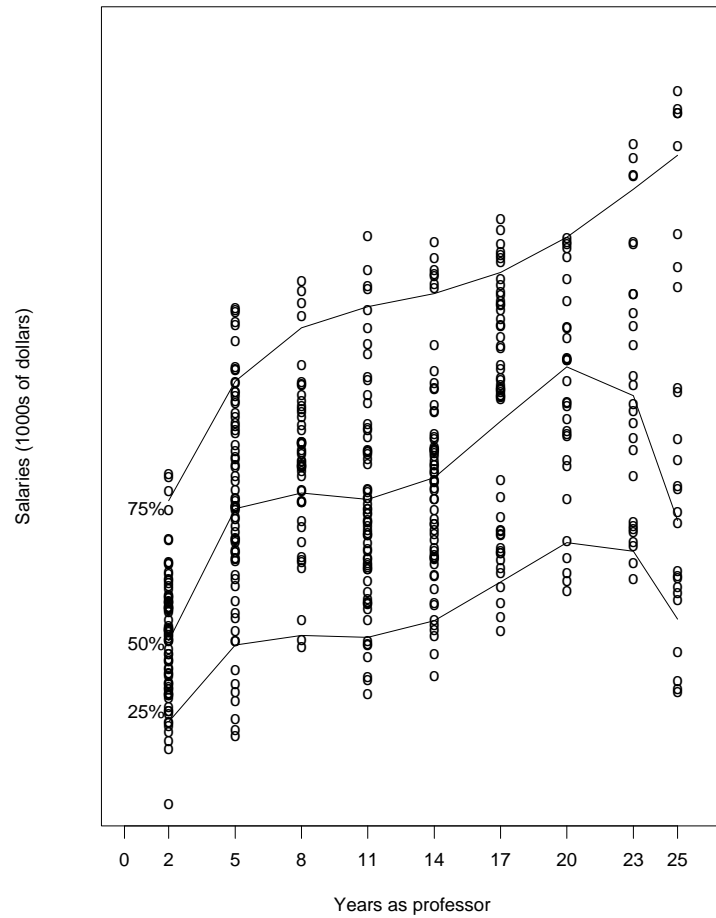


**Fig. 1.** The salaries of 469 US statistics professors as a function of years as professor with the fitted quadratic regression curve

Unfortunately, the curves in Figure 1 provides an inadequate picture of salary distribution in the sense that the change in shape of the salary distribution with years as professor is not displayed. This is simply because the standard regression fit models only the average relationship between salary and years as professor.

To give a more complete picture of the relationship between salary and years as professor, we present in Figure 2 the 25%, 50% and 75% sample quantiles. The resulting curves are called quantile regression curves.

We see that “the rich got richer and the poor got poorer.”



**Fig. 2.** The salaries of 469 US statistics professors as a function of years as professor. Three quantile regression curves with  $p = 0.25$ , 0.5 and 0.75 are shown. We see that “the rich got richer and the poor got poorer.”

## 2.2. Definitions of Quantile Regression

(1) In terms of conditional distribution:

$$q_p(\mathbf{x}) = \inf\{y : F(y|\mathbf{x}) \geq p\},$$

where  $F(y|\mathbf{x})$  is the conditional distribution of a response variable  $Y$  given the value  $\mathbf{X} = \mathbf{x}$ .

Hence, one can estimate  $q_p(\mathbf{x})$  via estimating  $F(y|\mathbf{x})$ .

(2) In terms of “check function”:

Least-squares regression estimation is the value of  $\theta$  that minimizes the expected square loss function  $E[(Y - \theta)^2 | \mathbf{x} = \mathbf{x}]$  and the associated loss function is  $r(u) = u^2$ .

Median regression estimates the conditional median of  $Y$  given  $X = x$ , and corresponds to the minimization of  $E[|Y - \theta| | \mathbf{x} = \mathbf{x}]$  over  $\theta$ . An associated loss function is  $r(u) = |u|$ . However, it is more convenient to take the loss function to be  $\rho_{0.5}(u) = 0.5|u|$ . Estimation proceeds by minimizing  $\sum_{i=1}^n \rho_{0.5}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$  over  $\boldsymbol{\beta}$ .

Note that

$$\rho_{0.5}(u) = 0.5 u I_{[0,\infty)}(u) - (1 - 0.5) u I_{(-\infty,0)}(u),$$

where

$$I_A(u) = \begin{cases} 1 & u \in A \\ 0 & \text{otherwise,} \end{cases}$$

is the usual indicator function of the set  $A$ . This definition may be generalized by replacing 0.5 by  $p$  ( $0 < p < 1$ ) to obtain a characterization of 100 $p$ th quantile regression  $q_p(x)$  at  $x$  as the value of  $\theta$  that minimizes

$$E[\rho_p(Y - \theta) | x = x],$$

where

$$\rho_p(u) = p u I_{[0,\infty)}(u) - (1 - p) u I_{(-\infty,0)}(u).$$

is called the “check function.”

(3) In terms of model error assumption:  
Give a regression model

$$Y = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon,$$

we assume that  $Q_{\epsilon}(p) = 0$  instead of  $E(\epsilon) = 0$ .

For example, CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles by Engle and Manganelli (2004, JBES)

$$y_t = f(y_{t-1}, x_{t-1}, \dots, y_1, x_1; \beta^0) + \epsilon_{t,p},$$

where

$$Q_{\epsilon_{t,p}}(p|\Omega_t) = 0$$

with  $\Omega_t = [y_{t-1}, x_{t-1}, \dots, y_1, x_1, f_1(\beta^0)]$ .

### 3. Quantile regression estimation is also robust for contaminated data

Suppose that a random batch of mixed “good” and “bad” pairs of independent observations  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ , is available for estimating the conditional mean  $E[Y|X = x]$ . We assume that a pair  $(x_i, y_i)$  is “bad” with probability  $\pi$  and “good” with probability  $1 - \pi$ , and that the pairs  $(x_i, y_i)$  are distributed as

$$(X, Y) \sim \begin{cases} N(0, 0, r, 1, 1) & \text{if } (x_i, y_i) \text{ is good} \\ N(0, 0, r, k, k) & \text{if } (x_i, y_i) \text{ is bad,} \end{cases}$$

where  $N(\mu_1, \mu_2, r, \sigma_1^2, \sigma_2^2)$  denotes a bivariate normal distribution.

That is,  $\{x_i, y_i\}_{i=1}^n$  are independent realizations from the common underlying “contaminated density” :

$$f(x, y) = (1 - \pi) f_1(x, y) + \pi f_2(x, y),$$

where  $f_1$  and  $f_2$  are the density functions for  $N(0, 0, r, 1, 1)$  and  $N(0, 0, r, k, k)$ , respectively. For a genuine contaminated distribution, we assume  $k \neq 1$ . A simple theoretic investigation presented by Yu and Jones (1998, JASA) established that the variance of a typical kernel smoothing estimation of  $E[Y|X = x]$  is much greater than the variance of any  $p$ th smooth quantile regression function. This is further evidence that quantile regression is much more stable than mean regression for analyzing this kind of “contaminated data” .

## 4. Value at risk, tails of distribution and quantiles

VaR, which estimates how much a certain portfolio can lose within a given time period for a given confidence level, is simply a particular quantile of future portfolio values.

Suppose  $Y_t^h$  is the return of a stock price over period  $[t, t + h)$ , then the CVaR can be defined as a level of return over the period which is exceeded with probability  $p$ :  $CVaR = \inf_v \{v : Pr(Y_t^h \leq v) \geq 1 - p\}$ . This can be written as

$$CVaR = F_{Y_t^h}^{-1}(1 - p|X_t),$$

where  $X_t$  is taken to be a “state process” or “information” vector, such as lagged return, interest rates, prices of securities, market indexes and so on.

## 5. Applications to detecting heteroscedasticity

We know that there is evidence of heteroscedasticity in daily, weekly and monthly returns in a market index on the Stock Exchange, also in many other finance cases. One can use ARCH (autoregressive conditional heteroskedasticity) or GARCH models to test heteroscedasticity.

But we want to say that quantile regression curves plots can provide a useful descriptive tool to detect heteroscedasticity.

That is, for a regression model  $Y = x^T \beta + \epsilon$ , if the distribution of  $\epsilon$  does not depend on the value of the covariate  $X$ , all regression quantiles will be parallel.

Any non-parallelism clearly indicates heteroscedasticity.

## 6. Estimation methods and algorithms

### Parametric quantile regression model

If we measure the relationship between a  $p$  quantile of the response  $Y$  and covariates  $\mathbf{x}$  or measure the effect of  $\mathbf{x}$  by a simple parametric model, we could assume that  $q_p(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$ .

Then given data set  $\{X_i, Y_i\}_{i=1}^n$ , we estimate the parameters  $\boldsymbol{\beta}$  from the minimization of

$$\sum_{i=1}^n \rho_p(y_i - \mathbf{x}^T \boldsymbol{\beta}).$$

## Software:

S-function “quantreg” in R, Matlab, StatLib and Qreg in GAUSS Library.

See <http://cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf>.

## Box-cox transformation quantile model

Conditional on  $X$ , assume that there exist  $\lambda$  such as  $g(\lambda; Y) = \frac{Y^\lambda - 1}{\lambda}$ , if  $\lambda \neq 0$ ,  $\log(Y)$ , if  $\lambda = 0$ , is normal, then the  $p$ th quantile of  $Y$  is given by  $h(\lambda; \beta_0 + \beta_1 x)$ , where  $h(\lambda; z)$  be the inverse function of  $g(\lambda; y)$  with respect to  $y$ .

## Nonparametric quantile regression model

Several methods have been explored for fitting conditional distribution so far. For example,  $F(y|x)$  may be estimated by  $\hat{F}(y|x) = \sum_{i=1}^n w_i(x) I(Y_i < y)$  with weight function  $w(\cdot)$ .

## Kernel fitting “check function”:

Typically, we estimate the quantile function by  $\min_a \sum_1^n \rho_p(Y_i - a)K(\frac{X_i - x}{h})$ , where  $K(\cdot)$  is a probability density function is called kernel function and  $h$  is a bandwidth or smoothing parameter.

or estimating  $F(y|x)$  by

$$\hat{F}(y|x) = \sum_{i=1}^n \Omega(\frac{Y-y}{b})K(\frac{X_i-x}{h}),$$

$$\text{where } \Omega(\frac{Y-y}{b}) = \int_{-\infty}^{\frac{y-Y}{b}} W(u)du.$$

**Qreg in GAUSS Library** can carry out this computation.

## Semi-parametric quantile regression model

Generally, if we assume that the underlying quantile function of  $Y$  on explanatory variable  $\mathbf{x}$  and a scalar  $t$ . With model of the form

$$q_p = \mathbf{x}^T \boldsymbol{\beta} + g(t)$$

then we usually fit the  $\boldsymbol{\beta}$  and  $g(\cdot)$  by

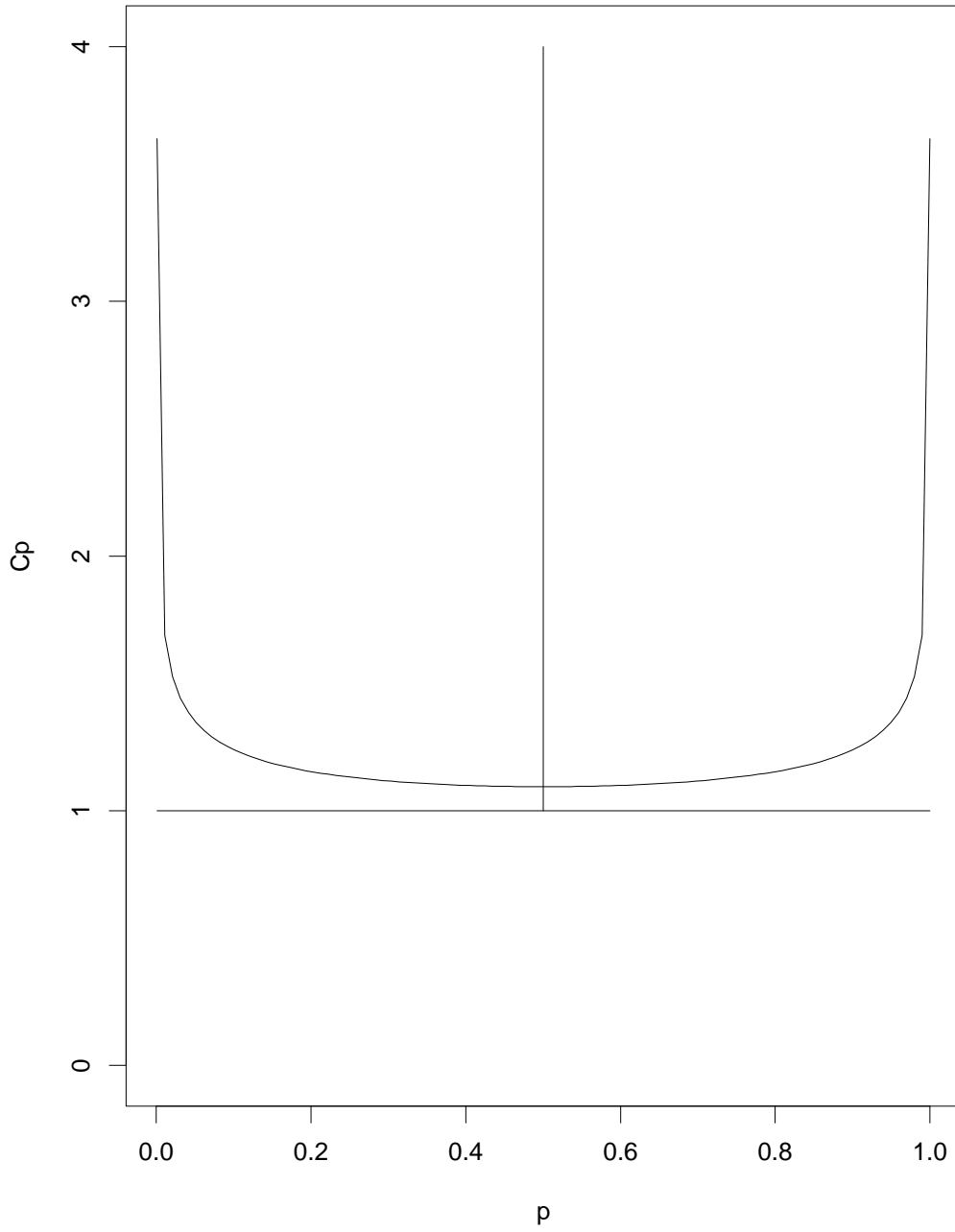
$$\sum_{i=1}^n \rho_p(Y_i - \mathbf{x}_i^T \boldsymbol{\beta} - g(t)) + \alpha \int g''(t)^2 dt.$$

`nlrq` in **R** can carry out this computation.

The **bandwidth selection** in kernel smoothing is given as follows:

(a) Select  $h_{mean}$ , the smoothing parameter for mean regression;

(b) Use  $h_p = h_{mean} \left\{ \frac{p(1-p)}{\phi(\Phi^{-1}(p))^2} \right\}^{1/5}$  to obtain all other  $h_p$ s from  $h_{mean}$ .



**Fig. 3** Graphically, the relation of  $p$  and  $h_p$

## Bayesian quantile regression

It is now generally recognised that a point forecast is not sufficient for well-informed decision-making in the face of an uncertainty future. Instead, a density forecast which is an estimate of the complete probability distribution of possible future values of the variable of interests is one of ways to deal with uncertainty. Bayesian inference, which uses Bayes' formula to compute the conditional posterior density or probability in terms of prior information and usually integrating out the unknown parameters, is particularly effective for dealing with parameter uncertainty.

Also, Bayesian inference enables making exact inference (posterior density) while some of classical inferences are based on asymptotic results, and Bayesian inference incorporate the variation of parameters (parameter uncertainty).

Basically, the posterior probability or density is proportional to likelihood function and prior probability.

$$\pi(\boldsymbol{\beta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta}).$$

$$0 < \int \pi(\boldsymbol{\beta}|\mathbf{y}) < \infty.$$

Where the likelihood function:

$$L(\mathbf{y}|\boldsymbol{\beta}) = p^n (1 - p)^n \exp \left\{ - \sum_{i=1}^n \rho_p(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \right\}. \quad (2)$$

**Priors:** a standard conjugate prior distribution is not available, but many priors, including improper uniform prior ( $\pi(\beta) \propto 1$ ), normal prior and exponential prior for  $\beta$  results in a proper joint posterior distribution.

## Posterior is Proper ?

First, the posterior is proper if and only if

$$0 < \int_{R^{k+1}} \pi(\beta|\mathbf{y}) d\beta < \infty, \quad (3)$$

or, equivalently, if and only if,

$$0 < \int_{R^{k+1}} L(\mathbf{y}|\beta) \pi(\beta) d\beta < \infty.$$

Moreover, we require that all posterior moments exist. That is,

$$E\left[\left(\prod_{j=0}^k |\beta_j|^{r_j}\right) | \mathbf{y}\right] < \infty, \quad (4)$$

where  $(r_0, \dots, r_k)$  denotes the order of the moments of  $\beta = (\beta_0, \dots, \beta_k)$ .

**Theorem 1:** Assume that the prior for  $\beta$  is improper and uniform, that is,  $\pi(\beta) \propto 1$ , then all posterior moments of  $\beta$  exist., i.e., equation (2) holds.

**Theorem 2:** When the elements of  $\beta$  are assumed prior independent, and each  $\pi(\beta_i) \propto \exp(-\frac{|\beta_i - \mu_i|}{\lambda_i})$ , a double-exponential with fixed  $\mu_i$  and  $\lambda_i > 0$ , all posterior moments of  $\beta$  exist.

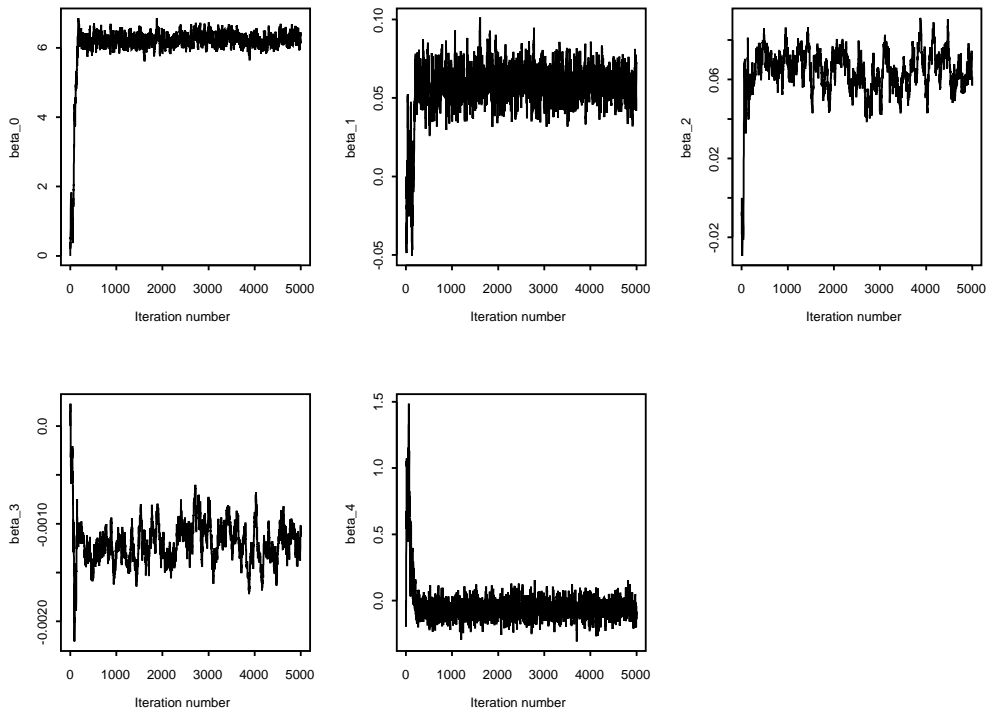
**Theorem 3:** Assume that the prior for  $\beta$  is multivariate normal  $N(\mu, \Sigma)$  with fixed  $\mu$  and  $\Sigma$ , then all posterior moments of  $\beta$  exist.

In particular, when the elements of  $\beta$  are assumed a prior independent and univariate normal, all posterior moments of  $\beta$  exist.

## Algorithm: MCMC (Markov chain Monte Carlo)

In this case an MCMC scheme would construct a Markov chain with equilibrium distribution the posterior  $\pi(\beta|\mathbf{y})$ . After running the Markov chain for a certain *burn-in* period so that it can reach equilibrium, one obtains samples from  $\pi(\beta|\mathbf{y})$ .

One popular method for constructing a Markov chain is via the Metropolis-Hastings (MH) algorithm.



**Fig. 4.** Typical time series plot of model parameter estimation based on (5000) iterations from Metropolis algorithm.

R function below for computation is based on uniform prior:

```

BayesQR_function(p, s, r) {
  prob <- function(y, x, p, beta)
  {
    v <- x %*% beta
    u <- (y - v)
    ind <- ifelse(u < 0, 1, 0)
    z <- - u * (p - ind)
    return(sum(z))
  }
  accept <- function(y, x, p, denom, thold, thres,
    beta)
  {
    prob <- function(y, x, p, beta)
    {
      v <- x %*% beta
      u <- (y - v)
      ind <- ifelse(u < 0, 1, 0)
      z <- - u * (p - ind)
    }
  }
}

```

```

        return(sum(z))
    }
    num <- prob(y, x, p, beta)
#denom <- prob(y, x, p, thold)
    mh <- (num - denom)
    mh <- exp(mh)
    mh <- min(mh, 1)
    q <- runif(1, 0, 1)
    if(mh > q) {
        th <- thnew
        den <- num
    }
    else {
        th <- thold
        den <- denom
    }
    list(th = th, den = den)
}
y <- dataY
nr <- dim(dataX.matrix)[1]
x <- cbind(rep(1, nr), dataX.matrix )

```

```
av1 <- mean(x[, 2])
av2 <- mean(x[, 3])
av3 <- mean(x[, 4])
x[, 2] <- x[, 2] - av1
x[, 3] <- x[, 3] - av2
x[, 4] <- x[, 4] - av3
b0 <- vector(mode = "numeric", length = r)
b1 <- vector(mode = "numeric", length = r)
b2 <- vector(mode = "numeric", length = r)
b3 <- vector(mode = "numeric", length = r)
b0old <- 0
b1old <- 0
b2old <- 0
b3old <- 0
sig0 <- 1.5
sig1 <- 0.3
sig2 <- 1
sig3 <- 0.25
beta <- c(b0old, b1old, b2old, b3old)
denom <- prob(y, x, p, beta)
b0[1] <- b0old
```

```

b1[1] <- b1old
b2[1] <- b2old
b3[1] <- b3old
for(i in 2:r) {
    thnew <- rnorm(1, b0old, sig0)
    beta <- c(thnew, b1old, b2old, b3old)
    acc <- accept(y, x, p, denom, b0old,
                 thnew, beta)
    b0old <- acc$th
    denom <- acc$den
    thnew <- rnorm(1, b1old, sig1)
    beta <- c(b0old, thnew, b2old, b3old)
    acc <- accept(y, x, p, denom, b1old,
                 thnew, beta)
    b1old <- acc$th
    denom <- acc$den
    b1[i] <- b1old
    thnew <- rnorm(1, b2old, sig2)
    beta <- c(b0old, b1old, thnew, b3old)
    acc <- accept(y, x, p, denom, b2old,
                 thnew, beta)

```

```

    b2old <- acc$th
    denom <- acc$den
    b2[i] <- b2old
    thnew <- rnorm(1, b3old, sig3)
    beta <- c(b0old, b1old, b2old, thnew)
    acc <- accept(y, x, p, denom, b3old,
                 thnew, beta)
    b3old <- acc$th
    denom <- acc$den
    b3[i] <- b3old
    b0[i] <- b0old - b1old * av1 - b2old *
              av2 - b3old * av3
  }

#par(mfrow = c(2, 2)) #plot(1:r, b0, xlab = "Iteration
ylab = "beta_0", type = "l") #plot(1:r, b1, xlab =
number", ylab = "beta_1", type = "l") #plot(1:r, b
"Iteration number", ylab = "beta_2", type = "l") #
xlab = "Iteration number", ylab = "beta_3", type =
c(summary(b0[(s + 1):r]), stdev(b0[(s + 1):r])) #s
c(summary(b1[(s + 1):r]), stdev(b1[(s + 1):r])) #s
c(summary(b2[(s + 1):r]), stdev(b2[(s + 1):r])) #s

```

```

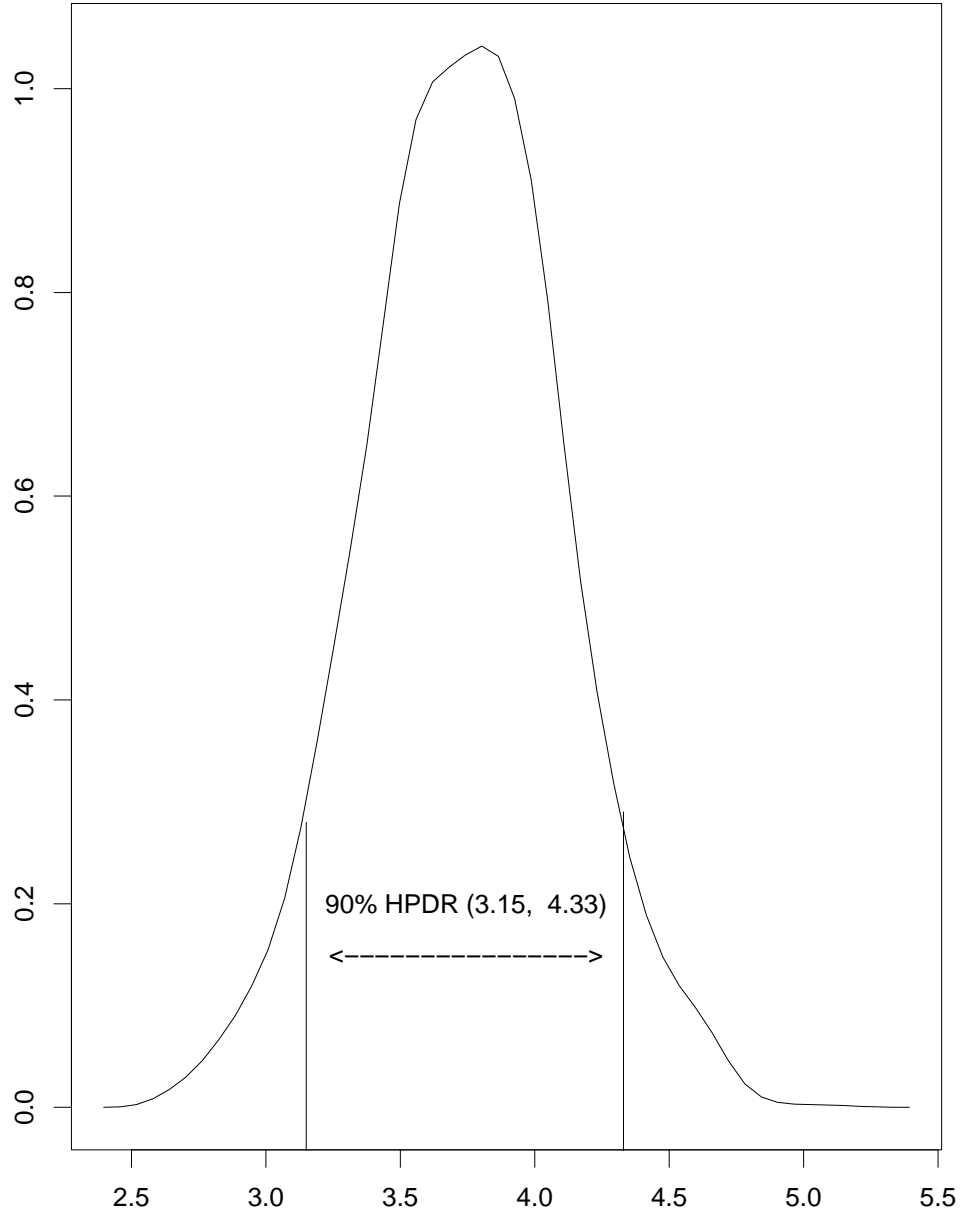
c(summary(b3[(s + 1):r]), stdev(b3[(s + 1):r])) #h
1):r]) #hist(b1[(s + 1):r]) #hist(b2[(s + 1):r]) #
1):r]) #par(mfrow = c(1, 1)) #stat <- rbind(stat0,
stat3) #return(stat)
    ball <- c(b0[(s + 1):r], b1[(s + 1):r], b2[(s
        1):r], b3[(s + 1):r])
    n <- r - s #credint(n, ball)
    return(ball)
}

```

## More reasons for Bayesian inference

- obtaining the highest posterior density (HPD) region (see Zellner, 1971 for details) of parameters;
- providing better prediction than classical approaches.

Bayesian approach provides a simple and easy method for obtaining the full posterior distributions (not only single values) of parameters, and obtaining the highest posterior density (HPD) region which can be used to build hypothesis tests. Formally, we are interested in testing the null hypothesis  $H_0 : \beta = c$ . Under Bayesian inference, we use the posterior density of  $\beta$  to construct an exact interval such as that  $Pr\{a < \beta - c < b | Data\} = 1 - \alpha$ , where  $\alpha$  is the significance level. If the value of  $\beta$  under the null hypothesis ( $H_0$ ) falls outside of the interval  $(a, b)$ , the null hypothesis will be rejected. Figure 5 shows the typical posterior distribution of a parameter  $\beta$  and its HPD region. The figure clearly shows, for example, that the null hypothesis  $H_0 : \beta = 0$  is rejected at the 10 percent level.



**Fig. 5.** Estimated density and its HPD

**Prediction:** For example, for the model

$$Y = X'\beta + u,$$

the actual value of  $Z_{T+1}$  in time  $T + 1$  is given by  $Z_{T+1} = X'_{T+1}\beta + u_{T+1}$ , and its prediction based on up to time  $T$  data can be written as  $\hat{Z}_{T+1} = X'_{T+1}\hat{\beta}$ , so that the forecast error is given by  $\epsilon_{T+1} = Z_{T+1} - \hat{Z}_{T+1} = X_{T+1}(\beta - \hat{\beta}) + u_{T+1}$ . Any point prediction based on classical methods is therefore subject to two types of uncertainty which relate to (i) to the estimation of  $\beta$  and (ii) the distribution of  $u_{T+1}$ . But Bayesian prediction can integrate over the parameters to get the marginal density of the as yet unobserved data, say  $h(z|I)$ , where  $I$  denotes the past sample and prior information. In this case, the integration over parameters  $\beta$  to obtain a marginal predictive density,  $h(z|I)$ , is a very useful way to get rid of parameter uncertainty by averaging the conditional density using the posterior density as a weight function.