Low Birth Weight Analysis Using Wild Bootstrap for Quantile Regression

Fabian Santi & Julian Sampedro

2021 (updated 2025)

Overview

(*i*) Introduce the Wild Btootrap method for quantile regression.

(*ii*) Perform a simulation study to see how the Wild Bootsrap compares to other bootstrap methods for quantile regression in the presence of homoskedasticity.

(*iii*) Model the variable birth weight using the introduced bootstrap method for quantile regression.

(*iv*) Analyse the data and comment our findings.

Notes: *Quantile regression is powerful because the incidence of a set of predictors on an outcome might be different for lower or higher quantiles. In this respect, it is a much better analysis option as compared to, say a linear or a polynomial regression.*

Introduction to Quantile regression

Ordinary Least Squares regression minimizes $\sum_{i=1}^{n} e_i^2$, while median regression, also known as Least Absolute Deviations (LAD) minimizes $\sum_{i=1}^{n} \mid e_i \mid$. As for Quantile regression, the idea is to minimize a sum that gives asymmetric penalties $(1-q) \mid e_i \mid$ for overprediction and $q \mid e_i \mid$ for underprediction. The quantile regression estimator for quantile q minimizes the objective function

$$Q(\beta_q) = \sum_{i:y_i \geq \mathbf{x}_i^T \boldsymbol{\beta}}^{n} q \mid y_i - \mathbf{x}_i^T \boldsymbol{\beta} \mid + \sum_{i:y_i \geq \mathbf{x}_i^T \boldsymbol{\beta}}^{n} (1-q) \mid y_i - \mathbf{x}_i^T \boldsymbol{\beta} \mid$$

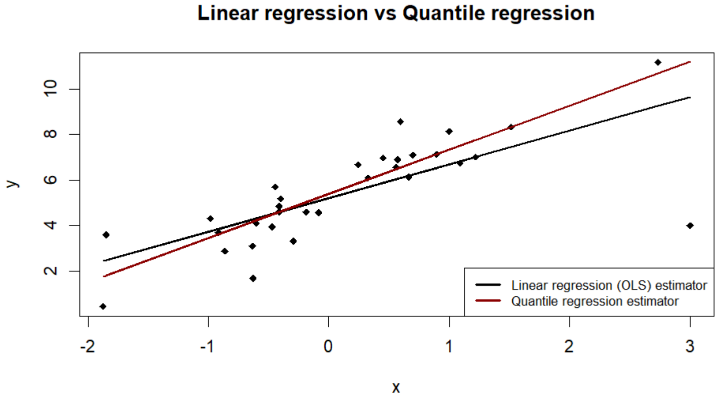
This nondifferentiable function is minimized numerically. Bootstrap confidence intervals for β_q are often used as theoretical confidence intervals may be hard or impossible to compute analytically.

Let us now compare the Linear regression (OLS) estimator to the Quantile regression estimator for the following simple model

$$y_i = 5 + 2x_i + e_i, \qquad (i = 1, \ldots, 30)$$

where x_i and e_i are drawn from a standard Normal distribution.

OLS vs Quantile regression estimator



Notes: *The Quantile regression estimator seems to be more robust to the presence of outliers in the dataset.*

When to use Quantile regression

(1) **Heteroskedasticity**: One could consider Quantile regression in the presence of heteroskedasticity (nonconstant variance of the residuals).

(2) **Asymmetric distribution of the response**: One could consider Quantile regression when the distribution of the response variable y is asymetric around its mean. In short, when we have a problem of skewness.

(3) **Outlying observations**: One could consider Quantile regression in the presence of outlying observations or influential observations in the dataset. In short, when the tails of the distribution of the response are thicker than those of a Normal distribution.

Wild Bootstrap

The Wild bootstrap (Wu and Liu, 1988) is suited when the model exhibits heteroscedasticity. The idea is to leave the regressors at their initial value, but to resample the response variable based on a modification of the residuals. For each replicate, we compute a new y based on $y_i^* = \hat{y}_i + \hat{e}_i w_i$. The full procedure is as follows:

(1) Fit a linear model to the data and denote the estimate of the parameter vector by $\hat{\boldsymbol{\beta}}$ and use $\hat{e}_i$ to represent the residuals.

(2) Generate w_i from an appropriate distribution satisfying the condition $e_i^* = w_i \mid \hat{e}_i \mid$.

(3) Calculate the bootstrapped sample as $y_i^* = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + e_i^*$.

(4) Refit the linear model to the bootstrap sample and denote the bootstrap estimate by $\hat{\boldsymbol{\beta}}^*$.

(5) Repeat steps 2–4 B times and estimate the variance of $\hat{\boldsymbol{\beta}}$ by the sample variance of the B copies of $\hat{\boldsymbol{\beta}}^*$.

Simulation study: overview (1/3)

We generate data from the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + 3^{-1/2}\left(2 + \left(1 + (x_{i1} - 8)^2 + x_{i2}\right)/10\right) e_i$$

where $(\beta_0, \beta_1, \beta_2) = (1, 1, 1)$ and $e_i \sim t_3$. In addition, x_{i1} are drawn from a standard Log-Normal distribution and x_{i2} are choosen to be 1 for the first 80% of the observations and 0 for the rest. We consider median regression with a samples size of $n = 50$.mIn addition, we use the following weight distribution:

$$g(w) = \begin{cases} -w(-2\tau - 1/4 \leq w \leq -2\tau + 1/4) \\ w(2(1-\tau) - 1/4 \leq w \leq 2(1-\tau) + 1/4) \end{cases}$$

Note: *(after Xingdong Feng, Xuming He and Jianhua Hu). We use the function rq() of the R package 'quantreg' to fit a median regression model.*

Simulation study: partial code (2/3)

```
for(j in 1:MC_sim){
  n = 50;  b0 = b1 = b2 = 1; e = rt(n,3)
  x1 = rlnorm(n); x2 <- numeric(n)
  x2[1:(0.8*n)] = 1; x2[((0.8*n)+1):n] = 0

  y = numeric(n)
  y = b0 + b1*x1 + b2*x2 + 3^(-1/2)*(2+(1+(x1-8)^2 + x2)/10)*e
  M1 = rq(y~x1+x2)
  boot_sim = 1000

  coef_boot_BN = matrix(rep(0, boot_sim*3),boot_sim, 3)
  w_BN = numeric(n)

  #Bootstrap BN
  for(i in 1:boot_sim){
    u = runif(n)
    w_BN[u<=0.5] = 1
    w_BN[u>0.5] = -1
    # Correction for residuals in finite sample
    x = cbind(rep(1,length(x1)), x1,x2)
    r = M1$residuals
    f0 = akj(r, z = 0)$dens
    r = r + hat(x) * (tau - I(r < 0))/f0

    y_boot_BN = M1$fitted.values + w_BN * abs(r)
    coef_boot_BN[i,] = rq(y_boot_BN~x1+x2)$coefficients
  }
  # BN
  MC_quant = quantile(coef_boot_BN[,1], c(0.05,0.95))
  MC_length_BN[j,1] = MC_quant[2] - MC_quant[1]
  MC_coverage_BN[j,1] = (MC_quant[2]>=b0 && MC_quant[1]<=b0)
}
```

Simulation study: results (3/3)

Comparison of the coverage rate of the 90% confidence intervals for $n = 50$. The bootstrap t-intervals are used for the Wild bootstrap method with weights as described earlier (CG) and weights generated from a Bernoulli distribution with equal probabilities at -1 and 1 (BN). We use $M = 1,000$ Monte Carlo simulations and $B = 1,000$ bootstrap copies.

	β_0		β_1		β_2	
Method	Coverage (%)	Length (SE)	Coverage (%)	Length (SE)	Coverage (%)	Length (SE)
CG	87.4	5.6 (0.29)	88.8	1.2 (0.09)	90.9	5.6 (0.24)
BN	87.3	5.6 (0.29)	88.7	1.2 (0.09)	90.7	5.6 (0.24)
PB	91.7	6.8 (0.36)	94.4	1.5 (0.11)	93.6	6.9 (0.32)
ND	90.7	3.4 (0.19)	86.9	0.8 (0.09)	91.9	3.4 (0.17)
RK	85.8	5.8 (0.33)	84.6	1.4 (0.13)	88.5	6.4 (0.34)

Note: *Coverage Probabilities and Interval Lengths for β_0, β_1, and β_2. Quote from the authors: "The proposed Wild bootstrap methods perform better overall than other methods, especially for the slope parameters β_1 and β_2. In particular, other reported resampling methods, PB (paired bootstrap), ND (noral approximation) and RK (rank score method) tend to be overly conservative".*

Median regression on low birthweight

We will fit a Quantile regression model to the 'birthwt' dataset from the package 'MASS'. There are $n = 189$ observations and 10 variables in the dataset. The variable 'bwt' represents the birth weight in grams and is our response variable. We will use as covariates the variables 'age', 'race' and 'smoke'. 'age' is the mother's age in years, 'race' is the mother's race ($1 =$ white, $2 =$ black, $3 =$ other) and 'smoke' is a binary variable taking value 1 if the mother smoked during pregnancy.

```
data(birthwt, package = 'MASS')
data = birthwt
head(data)
#    low age lwt race smoke ptl ht ui ftv  bwt
# 85   0  19 182    2     0   0  0  1   0 2523
# 86   0  33 155    3     0   0  0  0   3 2551
# 87   0  20 105    1     1   0  0  0   1 2557
# 88   0  21 108    1     1   0  0  1   2 2594
# 89   0  18 107    1     1   0  0  1   0 2600
# 91   0  21 124    3     0   0  0  0   0 2622
dim(data)
# [1] 189  10
class(data)
# [1] "data.frame"
```

Median regression on low birthweight

For a median ($\tau = 0.5$) regression model, we get the following estimates. We also display the 95% confidence interval for the intercept using the Wild bootstrap.

```
quantile.model <- rq(bwt ~ age + race + smoke,
                     tau = seq(0.05, 0.95, by = 0.05),
                     data = data)
summary(quantile.model)
# tau: [1] 0.5
# Coefficients:
#           coefficients     lower bd   upper bd
# (Intercept) 3248.20000   2620.25015 4011.99351
# age           17.60000    -19.56983   33.77936
# race        -283.20000   -362.99101  -42.80247
# smoke       -512.80000   -701.82717 -241.68587

# Wild bootstrap
quantile(boot.rq(cbind(rep(1,length(bwt)),age,race,smoke), bwt,
                 tau = 0.5, R = 1000, bsmethod = "wild")$B[,1], c(0.025,
                      0.975))
#       2.5%    97.5%
#   2509.913 4006.367
```

Note: *We also modeled different quantiles. Some results are presented next.*

Results

Modeling of *birthweight* (in grams) and main results:

$$Q_{bwt_i}(\tau \mid \text{age}_i, \text{race}_i, \text{smoke}_i) = \beta_0(\tau)+\beta_1(\tau)\,\text{age}_i+\beta_2(\tau)\,\text{race}_i+\beta_3(\tau)\,\text{smoke}_i.$$

with $age : mother'sageinyears$, $race : mother'srace(1 = white, 2 = black, 3 = other)$ and $smoke : smokingstatusduringpregnancy$. The regression coefficients are obtained by minimizing this asymmetric absolute deviation function:

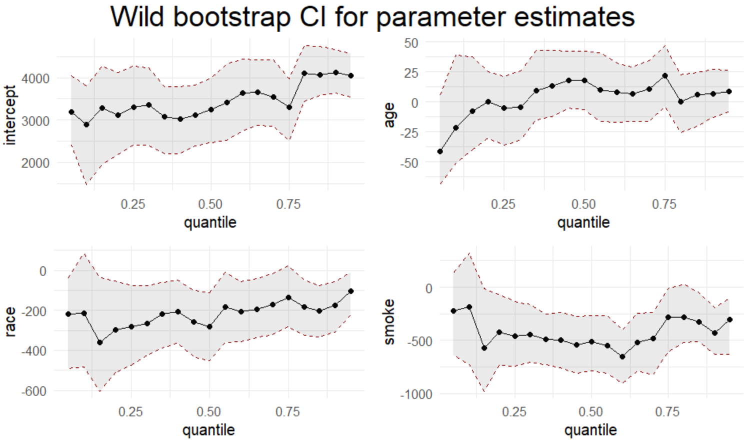
$$\hat{\beta}(\tau) = \arg\min_{\beta\in\mathbb{R}^4} \sum_{i=1}^{n} \rho_\tau(bwt_i - \beta_0 - \beta_1\,\text{age}_i - \beta_2\,\text{race}_i - \beta_3\,\text{smoke}_i)\,.$$

Below is a sumary of some of the results obtained.

	β_0	β_1	β_2	β_3
$Q_{bwt_i}(\tau = 0.25 \mid \text{predictors})$	3300.64	-5.09	-281.09	-456.82
$Q_{bwt_i}(\tau = 0.5 \mid \text{predictors})$	3248.20	17.60	-283.20	-512.80
$Q_{bwt_i}(\tau = 0.75 \mid \text{predictors})$	3296.40	21.87	-136.13	-279.07

Note: *These are partial results and confidence intervals have been computed but are not reported here.*

Estimated parameters



Note: *Estimated parameters across quantiles and* 95% *confidence intervals using the Wild bootstrap.*

Discussion of the results

(*i*) It is clear that heteroskedasticity is present in the dataset.

(*ii*) The intercept represents the estimated quantiles of birth weight for all covariates being equal to zero. The intercept is much higher for upper quantiles than for lower quantiles of birth weight.

(*iii*) Age has a more significant impact on low birth weight for the upper quantiles of birth weight than on lower quantiles.

(*iv*) Smoking during pregnancy has a negative impact on birth weight which is more important for the lower and upper quantiles than for quantiles close to the median. However, from the graph, the confidence intervals seem relatively large, which tends to indicate that the effects are not so important. Indeed, for smoking during pregnancy, at $\tau = 0.25$, we have about -400 g. (±300 g.) and at $\tau = 0.75$, we have about -250 g. (±250 g.) for instance. Therefore, the confidence intervals overlap and it is difficult to conclude that the effect is significatively different for different quantiles.

(*v*) Variability in parameter estimates across quantiles would not be captured by Ordinary Least Squares (OLS) regression.

References

Xingdong Feng, Xuming He and Jianhua Hu, 2011. Wild bootstrap for quantile regression.

Tutorial on quantile regression, Presentation by Xuming He and Ying Wei

Rizzo, M.L. , 2019. Statistical Computing with R, Second Edition (2nd ed.). Chapman and Hall/CRC.

The R Project for Statistical Computing