

ROBUST REGRESSION QUANTILES WITH CENSORED DATA

Michiel Debruyne and Mia Hubert

Key words: Robust statistics, regression quantiles, regression depth, censored data.

COMPSTAT 2004 section: Robustness.

Abstract: In this paper we propose a method to robustly estimate linear regression quantiles with censored data. We adjust the estimator recently developed by Portnoy by replacing the Koenker-Bassett regression quantiles with the regression depth quantiles. The resulting optimization problem is solved iteratively over a set of grid points. We show on some examples that, contrary to the Koenker-Bassett approach, this estimator can resist bad leverage points.

1 Introduction

We consider the linear regression setting, in which we have to model the response variable Y at a certain p -dimensional vector \mathbf{x} . For each $0 < \tau < 1$, denote $Q_\tau(Y|\mathbf{x})$ the τ th quantile of the conditional distribution of Y given \mathbf{x} . The regression quantile model states that this conditional distribution is linear in \mathbf{x} or

$$Q_\tau(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}_\tau. \quad (1)$$

We will try to estimate the parameters $\boldsymbol{\beta}_\tau$, given a data set which may include right censored data. This means that for every data point $1 \leq i \leq n$, the covariates $\mathbf{x}_i \in \mathbb{R}^p$ are measured, as well as censoring times c_i . Let y_i be the true response, possibly unobserved, then the observed responses (\tilde{y}_i, Δ_i) satisfy

$$\tilde{y}_i = \min\{y_i, c_i\} \quad \text{and} \quad \Delta_i = I(y_i \leq c_i)$$

with I the indicator function. So we only observe the smallest of y_i and c_i and we know whether y_i is censored or not. If there is no censoring, the observed \tilde{y}_i are all equal to the y_i .

In Section 2 we will briefly discuss the estimator recently developed by Portnoy [4]. This estimator is based on the classical Koenker-Bassett estimator and therefore is quite sensitive to leverage points. In Section 3 we will discuss a more robust estimator, the deepest regression estimator [5], which has been defined for non censored data. In Section 4 we will extend this estimator to the censored case along the lines of Portnoy's estimator. In Section 5 we provide an example.

2 The Koenker-Bassett estimator

In case of uncensored data, we can use the Koenker-Bassett estimator [3] to estimate regression quantiles. This estimator is defined as

$$\hat{\beta}_\tau = \arg \min_{\beta \in \mathbb{R}^p} \sum \rho_\tau(r_i(\beta))$$

with $r_i(\beta) = y_i - \mathbf{x}'_i\beta$ the i th residual and $\rho_\tau(u) = u(\tau - I(u < 0))$. For $\tau = 0.5$, $\hat{\beta}_\tau$ coincides with the L_1 -estimator which minimizes the sum of the absolute residuals.

In case of censored data a weighted version of this Koenker-Bassett estimator has been introduced by Portnoy [4]. The quantiles are iteratively estimated for $0 < \tau < 1$. At the start all observations have weight 1. When the i th censored observation is crossed for $\tau = \hat{\tau}_i$ (this means its residual with respect to the τ th regression quantile is negative if $\tau > \hat{\tau}_i$), weights are defined in two pseudo-observations. One pseudo-observation coincides with the most recently crossed observation and receives a weight $w_i(\tau) = (\tau - \hat{\tau}_i)/(1 - \hat{\tau}_i)$. Note that this weight is an estimate of the probability of the censored observation lying in between the $\hat{\tau}_i$ th and τ th quantile. The second pseudo-observation is put arbitrarily far away and receives a weight $1 - w_i(\tau)$, which is an estimate of the probability of the censored observation lying above the τ th quantile. In each step a weighted version of the objective function is minimized. More precisely, denote K the set of censored observations that have previously been crossed, then $\hat{\beta}_\tau$ is chosen to minimize

$$\sum_{i \notin K} \rho_\tau(r_i(\beta)) + \sum_{i \in K} \{w_i(\tau)\rho_\tau(c_i - \mathbf{x}'_i\beta) + (1 - w_i(\tau))\rho_\tau(y^* - \mathbf{x}'_i\beta)\}$$

over all hyperplanes β through p data points. The number y^* is any value sufficiently large to exceed all $\{\mathbf{x}'_i\beta : i \in K\}$. Iterative computation can be done using a simplex pivoting algorithm, as described in [4].

3 Regression depth quantiles

Regression depth has been introduced by Rousseeuw and Hubert [5]. For each $\theta \in \mathbb{R}^p$, the regression depth of θ with respect to the data set $Z_n = \{(\mathbf{x}_i, y_i)\}$ is defined as

$$rdepth(\theta, Z_n) = \min_{\lambda \in \mathbb{R}^p} (\#\{\mathbf{x}_i : sgn(r_i(\theta)) \neq sgn(\mathbf{x}'_i\lambda)\})$$

with $sgn(u) = -1$ if $u < 0$, $sgn(u) = 0$ if $u = 0$ and $sgn(u) = 1$ if $u > 0$.

In case of simple regression this definition can be interpreted as follows. A line θ can be rotated around any point v lying on θ , until it becomes a vertical line. This can be done clockwise or counterclockwise. Both ways, the minimum number of data points that is passed is counted. Finally this can be repeated for every point v and the overall minimum of crossed data

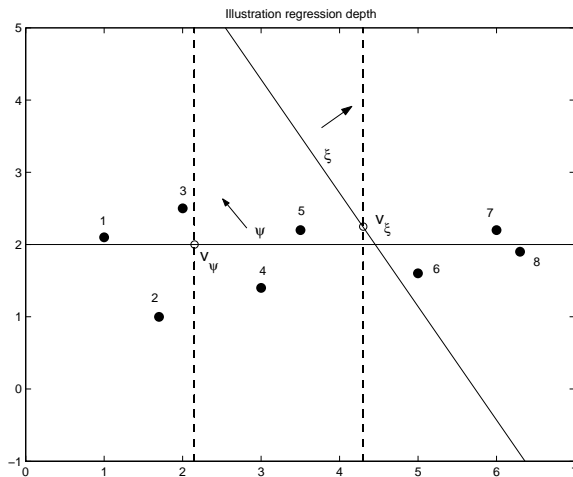


Figure 1: Example with 8 data points. The regression depth of the line ξ is 0, whereas the depth of the line ψ is 3.

points is retained. This is exactly the regression depth of θ . An illustration is given in Figure 1. The line ξ has regression depth 0, since it can be rotated clockwise around the point v_ξ without encountering any data points. The regression depth of the line ψ is 3, since we encounter 3 data points if we rotate it counterclockwise around v_ψ and we can find no lower number than this.

For general p , a hyperplane with high regression depth is well surrounded by data points as we will always find a large number of observations when it is rotated. So it can be expected that hyperplanes with high regression depth are quite good fits. Rousseeuw and Hubert ([5]) defined the deepest regression as

$$DR(Z_n) = \arg \max_{\theta \in \mathbb{R}^p} rdepth(\theta, Z_n).$$

It was shown in [2] that DR is a consistent estimator for the conditional median β_τ with $\tau = 0.5$.

The deepest regression has a breakdown value of 33%, which means that it can resist up to 33% of outliers in the data set. This is not the case for the Koenker-Bassett estimator, which can be heavily influenced by even one single outlier. More specifically, the estimator can resist vertical outliers but not leverage points which are outlying in the \mathbf{x} -space.

For general τ the deepest regression can be generalized similar to the Koenker-Bassett estimator. This yields the regression depth quantiles:

$$\hat{\beta}_\tau = \arg \max_{\beta \in \mathbb{R}^p} \inf_{\lambda \in \mathbb{R}^p} \left(\tau \#\{y_i : (r_i(\beta) > 0, \mathbf{x}'_i \lambda < 0)\} \right. \tag{2}$$

$$\left. + (1 - \tau) \#\{y_i : (r_i(\beta) < 0, \mathbf{x}'_i \lambda > 0)\} \right)$$

4 Depth quantiles with censored data

The ideas to obtain Koenker-Bassett regression quantiles for censored data can now be extended to the regression depth quantiles. A schematic overview of the algorithm is as follows.

STEP 1 Choose a set of grid points $\{0 < \tau_1 < \dots < \tau_m < 1\}$. The number of grid points m needed to find good estimates is quite low: Portnoy already suggested \sqrt{n} in [4] and this was also sufficient in our examples.

Estimate the τ_1 th regression quantile using the regression depth quantile for uncensored data (2). Crossed censored observations can be ignored since they almost do not contain any information.

STEP 2 Suppose we have estimated the τ_1 th regression quantile $\hat{\beta}_{\tau_1}$. Then we also know the set of crossed censored observations $K_{\tau_1} = \{(\mathbf{x}_i, c_i) : c_i - \mathbf{x}'_i \hat{\beta}_{\tau_1} \leq 0\}$. For each of these crossed censored observations a number $\hat{\tau}_i$ has been given following equation (3) that will be explained in step 3 of the algorithm. The according weight is

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i}.$$

STEP 3 Estimate the τ_{l+1} th regression quantile using a weighted version of (2).

$$\begin{aligned} \hat{\beta}_{\tau_{l+1}} = \arg \max_{\beta \in \mathbb{R}^p} \inf_{\lambda \in \mathbb{R}^p} & \left(\tau \#\{\tilde{y}_i \notin K_{\tau_l} : (r_i(\beta) > 0, \mathbf{x}'_i \lambda < 0)\} \right. \\ & + (1 - \tau) \#\{\tilde{y}_i \notin K_{\tau_l} : (r_i(\beta) < 0, \mathbf{x}'_i \lambda > 0)\} \\ & + \tau w_i(\tau) \#\{\tilde{y}_i \in K_{\tau_l} : (r_i(\beta) > 0, \mathbf{x}'_i \lambda < 0)\} \\ & + (1 - \tau)w_i(\tau) \#\{\tilde{y}_i \in K_{\tau_l} : (r_i(\beta) < 0, \mathbf{x}'_i \lambda > 0)\} \\ & \left. + \tau (1 - w_i(\tau)) \#\{\tilde{y}_i \in K_{\tau_l} : (\mathbf{x}'_i \lambda < 0)\} \right). \end{aligned}$$

The maximization is performed on a random grid of β and λ vectors, similar to the algorithms described in [1]. Consider the set $K_{\tau_{l+1}} = \{(\mathbf{x}_i, c_i) : c_i - \mathbf{x}'_i \hat{\beta}_{\tau_{l+1}} \leq 0\}$.

IF $K_{\tau_{l+1}} = K_{\tau_l}$,

the current estimate $\hat{\beta}_{\tau_l}$ was found using the correct weights and is therefore a correct solution.

IF $K_{\tau_{l+1}} \neq K_{\tau_l}$,

then the weights should be changed. Observations in $K_{\tau_l} \setminus K_{\tau_{l+1}}$ are censored observations that were crossed but are not anymore. These receive weight 1 again. Observations from $K_{\tau_{l+1}} \setminus K_{\tau_l}$ are censored observations that are

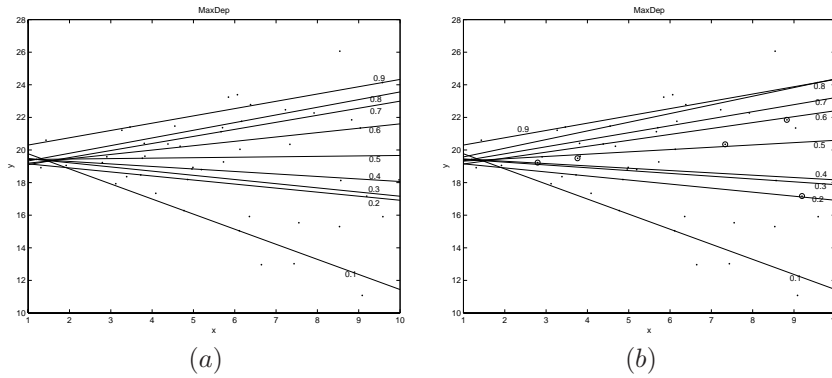


Figure 2: Regression depth quantiles for a simulated example of 50 data points (a) without censoring, (b) with 5 censored (= circled) data points.

crossed just now, during the transition from τ_l to τ_{l+1} . We define the number

$$\hat{\tau}_i = \tau_l \tag{3}$$

for each of these observations. Their weight is then

$$w_i(\tau) = \frac{\tau - \hat{\tau}_i}{1 - \hat{\tau}_i} = \frac{\tau - \tau_l}{1 - \tau_l}.$$

The remaining weight $1 - w_i(\tau)$ is assigned to a pseudo-observation arbitrarily far away. Thus we find a new set of crossed censored observations. The regression quantile $\hat{\beta}_{\tau_{l+1}}$ is then recomputed with this new set of weights. We repeat this step until we find an estimate for which the weights remain the same.

STEP 4 The algorithm stops when we have dealt with the last grid point τ_m .

□

Remark that it is possible that no convergence is obtained in step 3. In the examples and simulations we studied so far, this rarely occurred. Hence it does not seem too much of a problem. If it happens, we just skip the grid point τ_{l+1} and continue with the next one τ_{l+2} .

Although the algorithm performs very well in our examples and simulations, its computational complexity is very high since we have to solve a nested optimization problem in each grid point. Therefore in our presentation we will discuss some possible improvements using updating methods, which seem to speed up the computation time.

5 Examples

Let us look at an example in case of simple regression. Data were simulated from a heteroscedastic linear regression model. More precisely data were

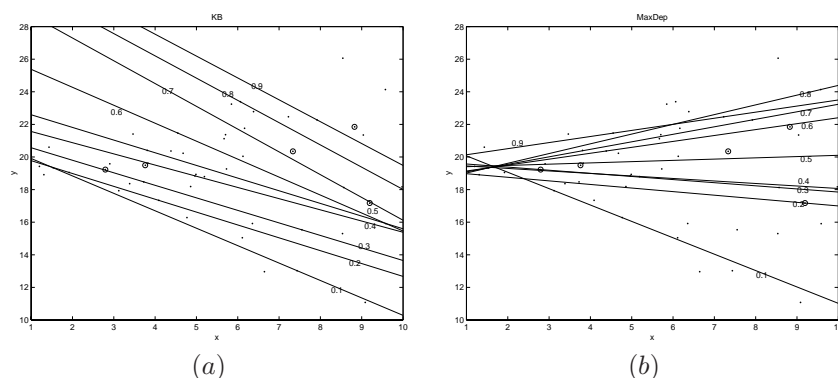


Figure 3: (a) Koenker-Bassett quantiles, (b) Regression depth quantiles for the simulated example with one leverage point.

generated according to the model

$$Y = 20 + 0.5xN(0, 1)$$

with the x uniformly distributed in $[1, 10]$. Figure 2(a) shows the resulting deciles for 50 uncensored data points. Then we have randomly chosen 5 data points to be censored with $c_i = y_i$ and obtained Figure 2(b). We see that the regression quantiles are slightly higher than in Figure 2(a) (this is especially clear for the 0.5-quantile), just as one expects.

Figure 3 shows the results when adding one leverage point at $(-50, 100)$. The Koenker-Bassett quantiles in Figure 3(a) are heavily affected, but the depth quantiles in Figure 3(b) only change slightly. This demonstrates the robustness of the latter.

A more realistic example is shown in Figure 4. 50 datapoints were generated according to the model

$$Y = 20 + (0.5N(0, 1) + 2)x$$

with the x uniformly distributed in $[1, 10]$. Three outliers were added with coordinates around $(1, 35)$. Note that the x - nor the Y -value of these outliers is outlying. Yet, since these points do not follow the linear model, they have a bad impact on the Koenker-Bassett regression quantiles. The 0.7-quantile is already slightly biased and the 0.8-quantile is very badly estimated. The depth quantiles perform better. The 0.7- and 0.8-quantile are well estimated and even the 0.9-quantile is not completely tilted towards the outliers.

In our presentation we will further show the efficiency and the robustness of the algorithm by several simulation studies in varying dimensions p .

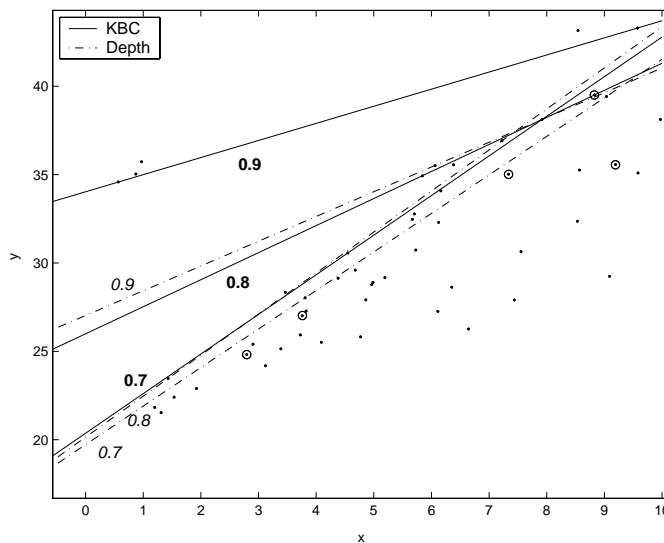


Figure 4: Comparison between Koenker-Bassett quantiles (solid lines) and depth quantiles (dashed-dotted lines) for a simulated example with three outliers.

References

- [1] Adrover J., Maronna R.A., Yohai V.J. (2004). *Robust regression quantiles*. Journal of Statistical Planning and Inference **122**, 187–202.
- [2] Bai Z.D., He X. (2000). *Asymptotic distributions of the maximal depth estimators for regression and multivariate location*. The Annals of Statistics **27**, 1616–1637.
- [3] Koenker R., Bassett G.J. (1978). *Regression quantiles*. Econometrica **46**, 33–50.
- [4] Portnoy S. (2003). *Censored regression quantiles*. Journal of the American Statistical Association **98**, 1001–1012.
- [5] Rousseeuw P.J., Hubert M. (1999). *Regression depth*. Journal of the American Statistical Association **94**, 388–402.

Address: M. Debruyne, M. Hubert, K.U.Leuven, Department of Mathematics, W. De Croylaan 54, B-3001 Leuven, Belgium

E-mail: {michiel.debruyne,mia.hubert}@wis.kuleuven.ac.be