L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

On computation of regression quantiles: Making the Laplacian Tortoise faster

Stephen Portnoy

Department of Statistics, University of Illinois, Champaign, USA

Abstract: In "The Gaussian Hare and the Laplacian Tortoise", the authors present a two-pronged attacked on the computation of L_1 and other regression quantile estimators in linear models for large samples. The first prong involves the application of interior point linear programming methods, specifically designed to treat the absolute error and related regression quantile objective functions. The second prong applies a form of stochastic preprocessing, somewhat reminiscent of the O(n) algorithms for computing the median of a single sample. These ideas provide computational methods that are in theory faster than least squares as $n \to \infty$ (with probability tending to one), and in practice are faster than Splus least squares functions for n larger than 10^4 (and the number of parameters moderate). Here some issues concerning this algorithm are considered, and some improvements are proffered.

Key words: Linear models, regression quantiles, L_1-estimation, computation.

AMS subject classification: 62G05, 62J05, 68A20.

1 Introduction

Consider the standard linear model: there are observations $\{Y_i : i = 1, ..., n\}$, satisfying

$$Y_i = x_i'\beta + u_i \qquad i = 1, ..., n \tag{1}$$

where x_i are vectors in $\mathbf{R}^p$, $\beta \in \mathbf{R}^p$ is a vector of unknown parameters, and $\{u_i\}$ form an i.i.d. sequence of errors. Though we consider the model conditionally on $\{x_i\}$, we will generally assume that these design vectors are realizations of an independent random process. The traditional approach to

statistical analysis of (1) is to use least squares to estimate the conditional mean of Y given x: $E[Y|x] = x'\beta$. However, this provides an analysis only of the center of the conditional distribution. To get a more complete picture of the relationship between Y and x, Koenker and Bassett (1978) introduced regression quantiles as the solution to the problem: for each $\tau \in [0, 1]$, let $\hat{\beta}(\tau)$ achieve

$$\min_{b \in \mathbf{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'b) \tag{2}$$

where $\rho_\tau(u) = \tau u^+ + (1-\tau)u^-$. When $\tau = .5$, $\hat{\beta}(.5)$ is just the usual L_1 estimator, which corresponds to the conditional median estimator. Under model (1) (and more generally), the line $y = x'\hat{\beta}(\tau)$ estimates the conditional quantile of Y given x. These methods have been applied successfully in a wide variety of examples.

Computation of regression quantile estimators has depended on the recognition that the minimization problem (2) is equivalent to a linear program, *viz.*:

$$\min_{b \in \mathbf{R}^p} \sum_{i=1}^{n} (\tau u + (1-\tau)v) \tag{3}$$

subject to

$$Y_i = x_i'b + u - v, \quad u_i \geq 0, \quad v_i \geq 0 \quad i = 1, \ldots, n$$

This equivalence was first presented in the literature in the mid 1950's. However, it is not unlikely that it was known earlier but dismissed as having little or no computational implication until the power of the simplex algorithm was appreciated. In fact, in the early $19th$ century Gauss (1809) already recognized that the L_1 estimator could be characterized as having p zero residuals. As described in Portnoy and Koenker (1997), efficient algorithms based on Danzig's simplex algorithm were developed, and they proved quite effective for sample sizes $n < 1000$ (or so). However, these algorithms were extremely slow for sample sizes significantly larger than 1000. Thus, Portnoy and Koenker(1997) were led to develop a new computational approach based on two fundamental ideas.

The first idea involved replacing simplex approaches with interior point methods, which originated in the mid 1980's and have been under extremely active development since then. The traditional simplex method is based on the idea that the constraint set for a linear programming problem is a simplex: that is, a convex set defined as the set of convex combinations of a finite number of extreme points (*viz.*, the vertices of the constraint

set). In the regression quantile problem, the vertices are just the β-values defined as having p zero residuals (which are often called "elemental' solutions). One proceeds iteratively by evaluating the linear objective function at a vertex and finding the adjacent vertex in the direction of steepest descent of the objective function. The solution is found after a finite number of such steps (called "pivots") when the algorithm reaches a point where no descent direction remains. Unfortunately, in large sample regression quantile problems, the number of vertices is of order n^p, and this tends to require the algorithm to pass through a very large number of vertices before reaching the solution. Fortunately, under moderate distributional assumptions in model (1), it is possible to show that with probability tending to one, each pivot moves a fraction $1/n$ towards the solution. Thus, since each pivot takes $\mathcal{O}(np^2)$ operations, the simplex algorithm might be expected to take $\mathcal{O}_p(n^2p^2)$ operations (much larger for large n than the $\mathcal{O}(np^2)$ operations required for least squares algorithms). In fact, this rate should be reducible to $\mathcal{O}_p(n^{3/2}p^2)$, since one can generally obtain a initial estimate within $\mathcal{O}_p(n^{-1/2})$ of the solution (*e.g.*, the least squares estimate will work as an initial estimate of the median for symmetric distributions). Nonetheless, this is substantially longer than least squares, especially since the constants implicit in the big-O terms are larger for simplex pivoting than for least squares methods.

Modern interior point methods escape this problem by avoiding the boundary of the constraint set. Consider the canonical linear program

$$\min_x \{c'x \mid Ax = b,\ x \geq 0\} \tag{4}$$

(where c is a vector, A is a matrix, and the inequalities are taken coordinatewise). Associate with this problem the following logarithmic barrier reformulation, which severely penalizes points close to the boundary:

$$\min_x \{B(x,\mu) \mid Ax = b\} \tag{5}$$

where

$$B(x,\mu) = c'x - \mu \sum \log x_k.$$

In effect, (5) replaces the inequality constraints in (4) by the penalty term of the log barrier. Solving (5) with a sequence of parameters μ such that $\mu \to 0$, we obtain in the limit a solution to the original problem (4). For each μ, problem (5) can be solved relatively effectively by iterative Newton or quasi-Newton methods: at each trial solution, one approximates the problem locally by a quadratic minimization problem and moves as far as possible towards the solution of the approximating problem without leaving

the constraint set (that is, remaining in the interior). The historical context of these methods and the details of our use of this approach for regression quantiles is presented in Portnoy and Koenker (1997). It appears that such methods can be as reliable as simplex methods, and are substantially faster for large n.

Nonetheless, the interior point algorithms were still much slower than least squares for very large n. In fact, the best rates for complexity available from the interior point literature are of the order $\mathcal{O}_p(n^{5/4}p^3)$ computer operations for random problems, though it has been conjectured that the $n^{5/4}$ factor can be reduced to $n \log n$. The second approach to accelerating the computation of regression quantiles is based on a stochastic preprocessing step that begins with a much smaller random subset of the data. To describe this approach, consider the L_1 problem:

$$\min_b \sum_{i=1}^{n} |Y_i - x_i'b| \,. \tag{6}$$

Suppose for the moment that we "knew" that a certain subset J_H of the observations $N = \{1, \ldots, n\}$ would fall above the optimal median plane and another subset J_L would fall below. Then, since knowing the sign of a residuals permits the replacement of the absolute value by the appropriate sign,

$$\sum_{i=1}^{n} |Y_i - x_i'b| \;=\; \sum_{i \in N \setminus J_L \cup J_H} |Y_i - x_i'b| - \sum_{i \in J_L} (Y_i - x_i'b) + \sum_{i \in J_H} (Y_i - x_i'b) \,.$$

It follows that

$$\sum_{i=1}^{n} |Y_i - x_i'b| \;=\; \sum_{i \in N \setminus J_L \cup J_H} |Y_i - x_i'b| + |Y_L - x_L'b| + |Y_H - x_H'b| \tag{7}$$

where $x_K = \sum_{i \in J_K} x_i$, and $Y_K = \sum_{i \in J_K} Y_i$ for $K \in \{H, L\}$. We will refer to these combined "pseudo-observations" as "globs" in what follows. It is not hard to show that minimizing (7), under our provisional hypothesis on the signs of the residuals, yields exactly the same solution as (6), but the revision has reduced effective sample size by $\#\{J_L \cup J_H\} - 2$ (essentially by the number of observations in the globs).

To find J_L and J_H, consider computing a preliminary estimate $\hat{\beta}$ based on a subsample of m observations. Compute a simultaneous confidence band for $x_i'\beta$ based on this estimate for each $i \in N$. Under plausible sampling assumptions the length of each interval is proportional to $1/\sqrt{m}$, so if M denotes the number of Y_i falling inside the band, $M = \mathcal{O}_p(n/\sqrt{m})$. Take J_L, J_H to be composed of the indices of the observations falling

outside the band. So we may now create the "globbed" observations $(Y_K, x_K), K \in \{L, H\}$ and reestimate based on $M + 2$ observations. Finally, we must check to verify that, in fact, all the observations in J_H, J_L have the predicted residual signs. If so, we are done; if not, we must repeat the process. If the coverage probability of the bands is P, presumably near 1, then the expected number of repetitions of this process is the expectation of a geometric random variable, Z, with expectation P^{-1}. We will call each repetition a cycle. As described in Portnoy and Koenker (1997), it is possible to show that the optimal choice for the initial subsample size is $m = \mathcal{O}(n^{2/3})$, and that the resulting size of the globbed sample is then $M = \mathcal{O}_p(n^{2/3})$. Under modest distributional assumptions, this provides an algorithm with complexity $\mathcal{O}_p(n^{2/3}(\log n)^2 p^3) + \mathcal{O}(np)$, where the last term comes from the computation of the globs and the checking of residuals. For n large and p moderate, this is strictly better than the rate $\mathcal{O}(np^2)$ for least squares; and in fact the algorithm in Portnoy and Koenker (1997) is essentially as fast as Splus least squares algorithms for $n < 10^6$ and p moderate; and it is undoubtedly strictly faster for n much larger than this range.

The algorithm presented in Portnoy and Koenker (1997) can be described formally, but briefly, as follows:

```
k ← 0
l ← 0
m ← [2n^{2/3}]
while(k is small){
    k = k + 1
    solve for initial rq using first m observations
    compute confidence interval for this solution
    reorder globbed sample as first M observations
    while(l is small){
        l = l + 1
        solve for new rq using the globbed sample
        check residual signs of globbed observations
        if no bad signs: return optimal solution
        if only few bad: adjust globs, reorder sample, update M, continue
        if too many bad: increase m and break to outer loop
        }
    }
```

Here, "rq" is the regression quantile problem, and an interior point

method is used to "solve" this problem. The remainder of the paper describes several approaches to improving this algorithm and to obtaining a better understanding of its performance.

2 Some simple improvements

Three relatively simple and straightforward improvements in the algorithm above have been made and tested. The first concerns the choice of the initial random subsample. The algorithm above assumes that the sample comes already randomized. For samples that are not simulated, this requires and initial random permutation of all the data – a rather time-consuming task if n is very large. Note that the above algorithm does not permit simply taking an initial random subsample of size m, since m increases with each cycle in order to ensure termination of the iterations. One could take a random subsample of size $m(k)$ (where k is the cycle number) at the beginning of each cycle, but this is still rather time-consuming. A faster approach is to form the "random" subsample by taking a random observation from each consecutive n/m observations. Although this is slightly different from taking a fully random permutation, is appears to be sufficiently random in all cases checked so far. It is extremely quick, requiring only m random uniforms and no sorting. It has the added advantage for real data of avoiding any large gaps in the sampling (as can occur with fully random permutations). The current implementation uses a simple multiplicative congruential generator. It appears to provide a modest improvement in timings at the cost only of keeping track of an addition random seed.

The second relatively simple improvement involves choice of the simultaneous confidence bands used to determine the residual signs for the first subsample. The algorithm above was originally programmed using the traditional Scheffé bands of the form

$$x_i'\hat{\beta} \pm \left(c\, x_i'(X'X)^{-1}x_i\right)^{1/2} /\hat{\sigma}_\tau$$

where c is a constant (from F-tables) and $\hat{\sigma}_\tau^2$ is an estimate of $\tau(1-\tau)/f^2(F^{-1}(\tau))$. Unfortunately, these bands require np^2 operations, a value that can make the algorithm very time-consuming and, in fact, does not even attain the complexity rate claimed above. To get a faster approach, note that it is preferable to choose the constant c to optimize the speed of the algorithm rather than to attain a given coverage probability. Thus, different conservative alternatives might prove better. One method that seems to work well is based on the inequality,

$$|x_i'\hat{\beta}| \leq \max_j \left\{ \left|\hat{\beta}_j\right| / s_j \right\} \times \sum_{j=1}^{p} |x_{ij}|\, s_j \,,$$

where s_j is $\hat{\sigma}$ times the diagonal element of the $(X'X)^{-1}$ matrix, and $\hat{\sigma}$ computed as for the Scheffé intervals. This approach provides conservative (though not "exact") confidence bands with width $c_q \sum_{j=1}^{p} |x_{ij}|\, s_j$. Note that this requires only $\mathcal{O}(np)$ operations; thus providing the rate required for the result of Section 1. Choice of the constant, c_q, is somewhat problematic, but some experimentation with simulated data showed that c_q could be taken conservatively to be approximately one, and that the algorithm was remarkably independent of the precise value of c_q. Although our initial experience with this approach is extremely promising, the next improvement discussed below permits the selection of c_q to be replaced by a simpler selection of an alternative parameter.

A third improvement concerns the size of the globbed sample. If the constant, c, in the simultaneous confidence intervals is fixed, the size of the globbed sample, M, is random. Since it is optimal to have M approximately equal to m, where m is the initial sample size, we have tried to choose the c so that M is near m with high probability (this is possible under model (1) as $n \to \infty$). In practice, M varies significantly, and this leads to difficulties. If M is too small, the final estimates are likely to be wrong, thus requiring extra cycles. If M is too large, the interior point algorithm takes much more time that it should. Since in the limit, M will be very near its mean, which is a constant times m, it seems reasonable to fix the sample size $M \equiv am$, where a is a constant near 1. The globbed sample consists of the $M = am$ observations with the smallest values of $|r_i|/z_i$, where r_i is the residual and z_i is the constant in the confidence intervals depending only on the design matrix. That is, $z_i = (x_i'(X'X)^{-1}x_i)^{1/2}$ for Scheffé intervals, and $z_i = \sum_{j=1}^{p} |x_{ij}|\, s_j$ for the conservative intervals described above. This is asymptotically equivalent to fixing c, but always provides an appropriate globbed sample size. In a small scale simulation study, it appeared that $a = .8$ was a good choice over a variety of data set sizes and distribution assumptions. The lack of randomness in M provides a much more reliable and faster algorithm, whose performance varies substantially less from trial to trial. The result that $a < 1$ probably arises from the fact the the globbed sample is not like a random sample (as discussed in Section 4). Thus the interior point algorithm takes somewhat more time on a globbed sample than on a random subsample of the same size (perhaps 20 to 40 per cent more time, depending of the specific problem).

3 On the interior point algorithm

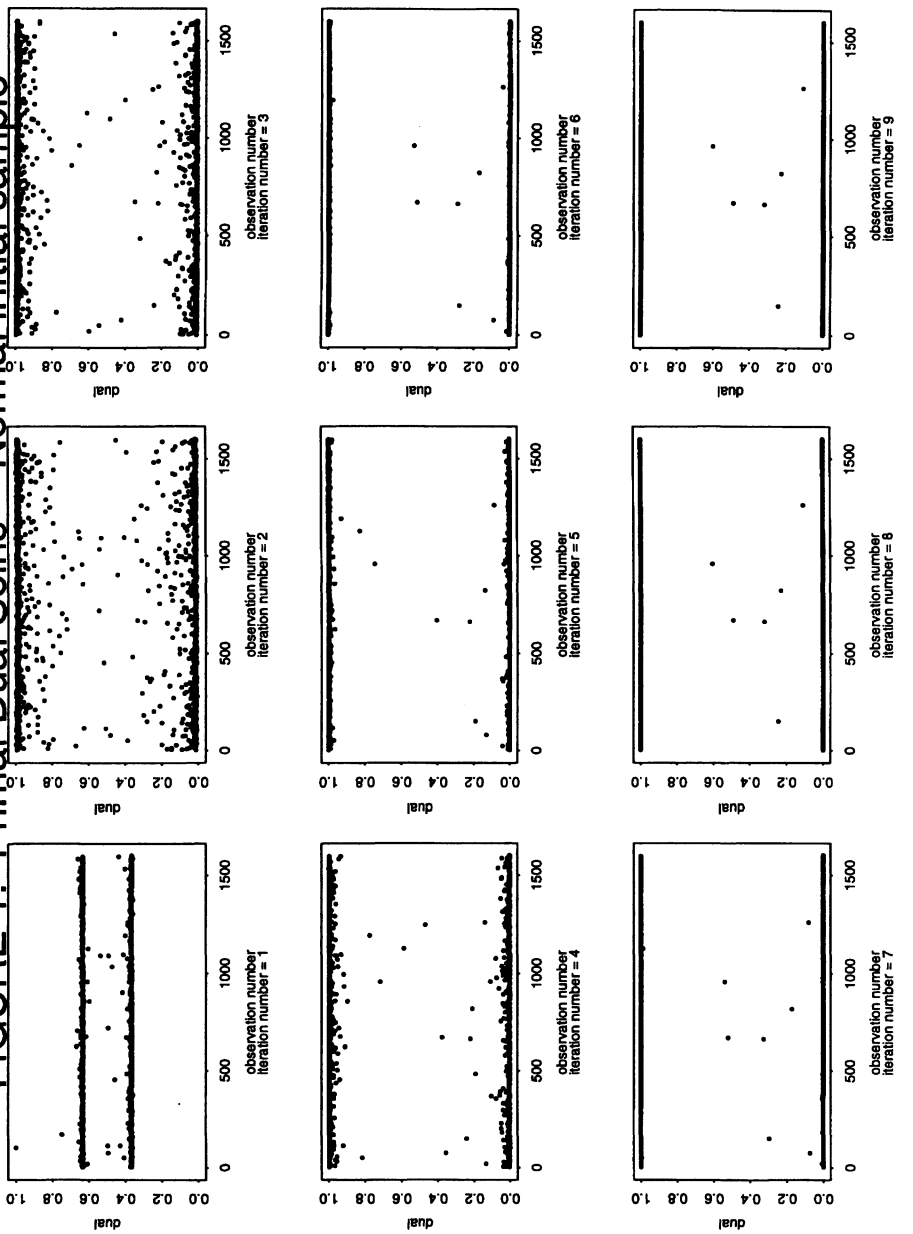
For $n < 10^6$, the algorithm of Section 1 spends almost all of its time in the interior point optimization steps for the random and globbed subsamples. However, the full advantage of the preprocessing step isn't realized until the operations of complexity $\mathcal{O}(np)$ dominate (that is, until most of the time is spend in creating the globs and in checking residuals). Thus, improvements in the interior point algorithm should provide the most important sources of faster performance (for $n < 10^6$). Furthermore, the use of nonlinear optimization methods requires the specification of a number of performance parameters. These include: (*i*) choice of starting values for the variables over which optimization is carried out, (*ii*) choice of direction at each step (for example, use of gradient direction for steepest descent, or use of the Newton direction to solve the approximating quadratic, or some combination of these), (*iii*) specification of the distance to move along the chosen descent direction, (*iv*) selection of a method for updating the penalty coefficient μ in equation (5), and (*v*) selection of stopping criteria. Specification of these parameters provides ample room for fine-tuning the algorithm to provide better performance.

Interior point methods for L_1 problems generally replace the linear program in equation (3) by a related problem, called the dual problem

$$\max\{Y'a \mid X'a = \frac{1}{2}X'e, \quad a \in [0,1]^n\}. \tag{8}$$

where Y is the vector of response observations and X is the design matrix. Here the coordinates of a, $\{a_i\}$, correspond to the signs of the residuals, $r_i \equiv Y_i - x_i'b^*$, at the final solution, a^*, where b^* is the optimal solution to equation (3). Precisely, at a solution, $a_i = 1$ if $r_i > 0$, $a_i = 0$ if $r_i < 0$, and a_i is strictly between 0 and 1 if $r_i = 0$. In the regression quantile setting, the values $a_i(\tau)$ $(0 \leq \tau \leq 1)$ are exact analogues of the rank functions of Hájek and Šidák (1967): see Gutenbrunner and Jurečková (1992). Thus, the solution b^* to the primal problem (3) can be determined from the solution to (8); and, in fact, the objective functions are the same at the solutions. The "primal-dual" algorithm of Portnoy and Koenker (1997) proceeds by establishing first order conditions depending on both a and b, and solving for both simultaneously. That is, in each iteration, both a and b are moved in a Newton direction toward the solution. At each iteration, the difference between the primal and dual objective functions is positive until the solution is reached. At the solution, this difference, called the "duality gap" becomes zero, and this provides a reliable test for convergence of the algorithm. Although each iteration of the primal-dual algorithm is somewhat more complicated, the number of iterations tends

to be somewhat smaller than other approaches, and the algorithm appears to be quite robust to idiosyncrasies in the data.

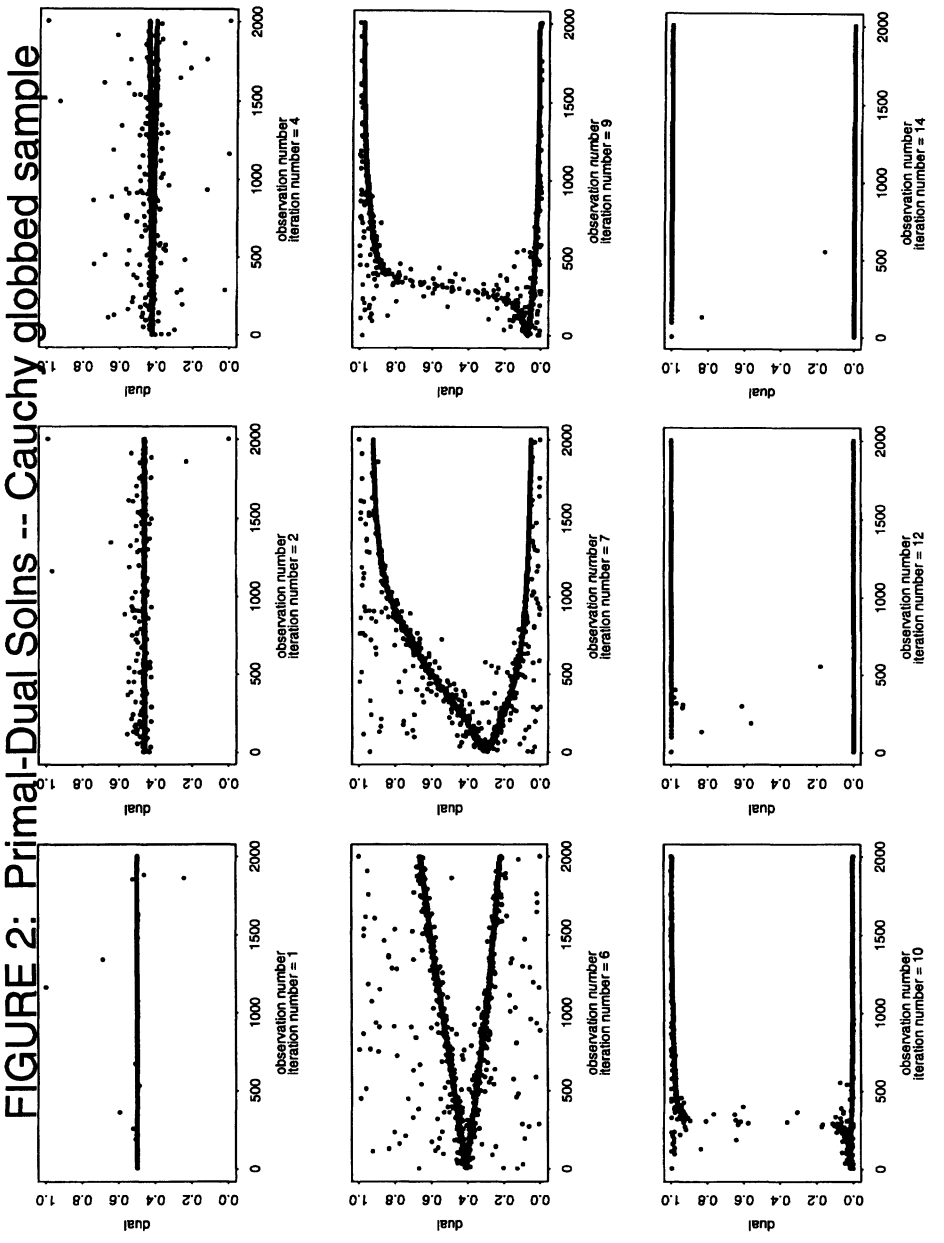


The algorithm in Portnoy and Koenker (1997) chooses the initial starting values as follows: b_0 is chosen to be the least squares estimator, and a_0 is chosen to be a constant vector with all n elements equal to .5 (halfway

between the bounds 0 and 1). In an effort to see how this algorithm works, the values of the dual variables, a_i, were plotted against i at each iteration. Plots were made for several random data sets with sample size 1000 or 2000, values near typical initial sample sizes for problems with n between 10^4 and 10^5. One picture of successive iterations on normal data with $p = 6$ appears in Figure 1. The first plot gives the a-values after one iteration. Note that the values have separated from the initial constant .5 and have begun to move toward 0 and 1. Remarkably, the vast majority of the values move exactly the same fraction of the distance to the extreme limits, 0 and 1. The values quickly approach the limiting values, except for the observations with zero residuals at the solution. The last iteration represents a very minor fine tuning of the next to last one, which already identified the zero residuals. This suggests the possibility of stopping a bit earlier, but in practice early stopping provides only a modest improvement in timings (at the cost of potentially less reliable performance).

Similar plots were made when the primal-dual algorithm was applied to the globbed sample. Here the a_i-values were plotted against $|r_i|/z_i$, where r_i are the residuals and the z_i-values are defined in the third improvement discussed in Section 2. That is, the a_i are plotted against the order at which observations enter the globbed sample. A picture for Cauchy data with $p = 3$ is given in Figure 2. At first glance, the results appear even more remarkable. With the residual ordering taken into account, it is clear that the a_i-values for smaller residuals tend to their limits much more slowly than those for larger residuals (among the globbed sample, for which all but the last two observations – the globs – have small residuals). Again, the a_i-values tend to fall along lines, but the lines are not symmetric about .5, and the become highly curved after iteration 6.

It is possible that the use of the least squares estimator as the initial starting value for applying the interior point algorithm to the globbed sample might be responsible for some of the oddities in the plots. An alternative that should be somewhat better and might be somewhat faster is to use the $\hat{\beta}$ from the solution to the initial random subsample as the starting value for the globbed sample. It turns out that this does not affect the plots of the a_i-values for the specific example plotted in Figure 2. However, in various simulation experiments, use of the initial $\hat{\beta}$ as the starting value for applying the interior point algorithm to the globbed samples appeared to provide a modest but definite improvement. These figures give some tantalizing hints as to why this is so, but significantly better understanding of the performance of interior point methods should provide even more substantial improvements.

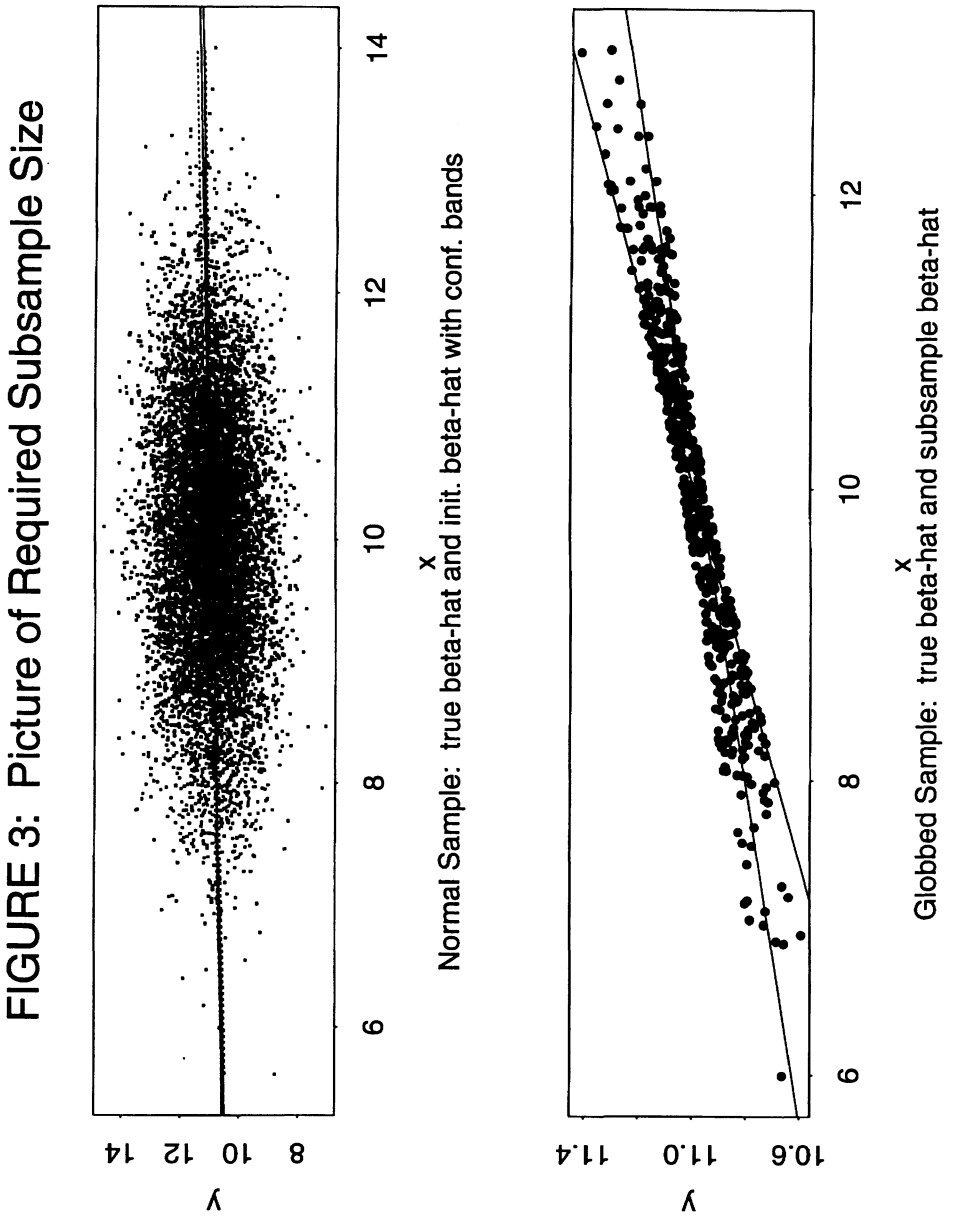


4 On applying the algorithm recursively

An obvious source of improved performance for very large samples would clearly be to apply the algorithm recursively. Since the algorithm gives the fastest computation of regression quantiles, the use of interior point meth-

ods should be replaced by the full algorithm using preprocessing whenever the subsample sizes are sufficiently large to make this replacement noticeably faster. Generally, this would occur when n is somewhat larger than 10^5. Unfortunately, there is a serious problem with this replacement for solving the globbed problem. The stochastic preprocessing step assumes that the sample is a random one. The globbed sample is far from random – it is chosen to consist of the smaller residuals from the initial sample plus the two globs. Thus, the preprocessing step cannot be expected to provide any real reduction in sample size for the globbed sample. Figure 3 should make this clear. The first graph in Figure 3 plots a Normal sample of size 10,000 together with the the whole-sample L_1 line, the initial subsample L_1 line and the confidence bands (based on the subsample). The solution to the initial subsample should differ from the correct sample regression quantile by an error of order $\mathcal{O}_p(m^{-1/2})$ (where m is the initial subsample size). This error is of the same order as the width of the confidence bands. Therefore, as the first graph shows, the lines and confidence bands are extremely close together on the scale of the data. The second plot of Figure 3 shows just the globbed sample together with the correct L_1 line and an L_1 line based on a random subsample of the globbed sample. Clearly, any confidence band about this subsample line that contains the correct line must contain all but a relatively modest fraction of the globbed sample. That is, since the residuals should be roughly uniformly distributed between the bands (assuming the random errors have a smooth density near the desired quantile), it is clear that an appreciable fraction of them must have their residuals from a subsample estimate differ in sign from those of the correct estimate. Thus, for the globbed sample, it would be impossible to replace the problem by one with a sample size smaller than a constant fraction of the globbed sample size. This contrasts markedly with the reduction from n to $n^{2/3}$ that preprocessing affords for random samples.

It is possible to use the preprocessed algorithm for the initial random sample. A few simulations with $n > 10^6$ and $p \leq 4$ were tried with this modification, and a modest improvement (about 20%) was obtained. Unfortunately, computer space limitations precluded more extensive testing, which remains to be done.



References

[1] Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium ... ; Werke.* Gotha (1871), Second Book, Third part, Section 186.

[2] Gutenbrunner, C. and Jurečková, J. (1992). Regression rank scores and

regression quantiles. *Ann. Statist.* **20** 305-330.

[3] Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* Prague: Academia.

[4] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46** 33-50.

[5] Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: computability of squared error and absolute-error estimators. *Statistical Science.* To appear.