GALTON, EDGEWORTH, FRISCH, AND PROSPECTS FOR QUANTILE REGRESSION IN ECONOMETRICS

ROGER KOENKER

Abstract. The work of three leading figures in the early history of econometrics is used to motivate some recent developments in the theory and application of quantile regression. We stress not only the robustness advantages of this form of semiparametric statistical method, but also the opportunity to recover a more complete description of the statistical relationship between variables. A recent proposal for a more X-robust form of quantile regression based on maximal depth ideas is described along with an interesting historical antecedent. Finally, the notorious computational burden of median regression, and quantile regression more generally, is addressed. It is argued that recent developments in interior point methods for linear programming together with some new preprocessing ideas make it possible to compute quantile regressions as quickly as least squares regressions throughout the entire range of problem sizes encountered in econometrics.

1. Galton's Regression to the Mean

Arguably, the most important statistical graphic ever produced is Galton's (1885) figure illustrating "regression to the mean", reproduced below as Figure 1.1. In it Galton plots childrens' height versus parents' height for a sample of 928 children. He begins by dividing the plane into one inch squares and entering the frequency counts for each square. The resulting "histogram" appeared too rough so he smoothed the plot by averaging the counts within each group of four adjacent squares and plotting the averaged count at the intersection of the cell boundaries. Not content to invent "regression" in one plot, he managed to invent bivariate kernel density estimation, too! After smoothing, the counts appeared more regular and he enlisted the help of the Cambridge mathematician, J.H. Dickson, to draw elliptical contours corresponding to level curves of the underlying population density.

1991 *Mathematics Subject Classification.* C12, C13.

Key words . Least absolute error regression, quantile regression, regression depth, interior point methods, linear programming .

Version: September 16, 2012. Department of Economics, University of Illinois, Champaign, IL 61820. This research was partially supported by NSF grant SBR-9617206, and was prepared for the Conference on Principles of Econometrics in Madison, May 2-3, 1998. The author wishes to thank Art Goldberger, Chuck Manski, and Peter Rousseeuw for very helpful comments on an earlier draft.

Now suppose we wished to predict children's height based on parental height, say the average height of the parents which we will call, following Galton, the height of the midparent. what would we do? One approach, given the graphical apparatus at hand would be to find the "most likely" value of the child's height given the parents height, that is for any given value of the mid-parent height we could ask, what value of the child's height puts us on the highest possible contour of the joint density. This obviously yields a locus of tangencies of the ellipses with horizontal lines in the figure. These conditional modes, given the joint normality implicit in the elliptical contours, are also the conditional medians and means. The slope of the line describing this locus of tangencies is roughly 2/3 so a child with midparent 3 inches taller than average can expected to be (will most probably be) only 2 inches taller than average.

Galton termed this regression towards the mean, and paraphrasing Lincoln we might strengthen this to regression of the mean, to the mean, and for the mean. Stigler (1997) provides a fascinating guide to Galton's own thinking about this idea, and to the illusive nature of its reception in subsequent statistical research. It is a remarkable feature of the conditional densities of jointly Gaussian random variables that the conditioning induces what we may call a "pure location shift". In Galton's original example the height of the midparent alters only the location of the center of the conditional density of the child's height; dispersion and shape of the conditional density as invariant to the height of the midparent. This is, of course, the essential feature of the classical linear regression model – the entire effect of the covariates on the response is captured by the location shift

$$E(Y|X = x) = x'\beta$$

while the remaining randomness of Y given X may be modeled as an additive error *independent* of X.

Of course, we are all aware that the attempt to compare random variables only in terms of means is fraught with difficulties, and we have all suffered the indignities of "humor" about our professional complacency in the face of simultaneously frozen and roasted extremities. Galton (1889) offered his own warning about this, chiding his statistical colleagues who,

> limited their inquiries to Averages, and do not seem to revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of a native of one of our flat English counties, whose retrospect of Switzerland was that, if the mountains could be thrown into its lakes, two nuisances would be got rid of at once. [*Natural Inheritance,* p. 62]

A bit earlier, Galton had conducted experiments on sweet peas to investigate the heritability of size. Remarkably, this prior analysis was conducted in terms of conditional medians and "probable deviations," a term he used to designate half the interquartile range. He concluded from these experiments that the median weight

of "filial seeds" was approximately linear in the weight of the parent seeds, with a slope less than unity, a phenomenon he described as "reversion". And he noted that the probable deviation of filial seeds was approximately constant over the range of parental sizes. Hald (1998) offers a detailed discussion of this work.

Ironically, it is the indispensable tool Galton provided us that is probably most responsible for narrowing the scope of statistical investigations to the comparison of means. So, compounding the irony, it seems fitting that – as a resident of one of the flatter U.S. counties – I have spent a large fraction of my professional energy arguing that we don't need to throw the mountains into the lakes with least squares regression. We *can* take a more comprehensive view of the statistical relationship between variables by expanding the scope of the linear model to include quantile regression.

As an elementary introductory illustration, I would like to reconsider an AR(1) model for daily temperature in Melbourne, Australia, considered recently by Hyndman, Bashtannyk and Grunwald (1996) using modal nonparametric regression ideas introduced by Collomb, Härdle, and Hassan (1987) and is also discussed by Scott (1992). The approach adopted here will be somewhat different, but the conclusions are rather similar. In Figure 1.2, we see the data, 10 years of daily maximum temperature data depicted as an AR(1) scatter plot. Not surprisingly, one's first "unit root" impression of the plot is that today's maximum temperature bears a strong resemblance to yesterday's maximum. But closer examination of the plot reveals that this impression is based primarily on the left half of the plot where the central tendency of the scatter does follow the 45° line quite closely. However, in the right half, when yesterday was warm the pattern is more complicated. It appears that *either* there is another hot day, or there is a dramatic cooling off, but a mild cooling off appears to be infrequent. In the language of conditional densities, if today is hot, tomorrow's temperature appears bimodal with one mode centered (roughly) at today's temperature and the other mode centered at about 20° centigrade.

This form of mean reversion has a natural meteorological explanation as high pressure systems bringing hot weather from the interior of the continent eventually terminate with a cold front creating a rapid drop in temperature. This sort of dynamic does not seem entirely implausible in some econometric settings, and yet the linear time-series models we typically consider are not capable of accommodating this behavior. This is clearly a case in which the conditioning variables influence not only the location, but also the scale and shape of the conditional density.

In Figure 1.3 we illustrate a family of estimated conditional quantile functions superimposed on the original scatter plot. Each curve is specified as a series expansion in B-splines

$$Q_{Y_t}(\tau|Y_{t-1}) = \sum_{i=1}^{p} \phi_i(Y_{t-1})\beta_i(\tau) \equiv x_t'\beta(\tau)$$

and the parameters are estimated by minimizing

$$\mathcal{R}_\tau(b) = \sum_{t=1}^{T} \rho_\tau(y_t - x_t'b)$$

as in Koenker and Bassett (1978) where $\rho_\tau(u) = u(\tau - I(u < 0))$. Selection of p and the knot locations for the B-splines are described in He and Ng (1998). Given a family of such conditional quantile functions it is straightforward to estimate conditional densities at any desired value of the conditioning variables. Related smoothing spline methods for quantile regression are discussed in Koenker, Ng, and Portnoy (1995).

If there is one message which I would extract from this example, it would be this: *there is more to econometric life than can be captured by the philosophy of the Gaussian location shift.* In the next section, I would like to describe briefly some early contributions of F.Y. Edgeworth to the theory of median regression and mention some related recent developments on alternative methods for quantile regression. In the third section I will describe some recent developments in the theory and practice of computing quantile regression estimates, developments which may be traced back to work in the 1950's by Ragnar Frisch, and which dramatically improve the computational feasibility of these methods in large problems. Some mildly polemical comments conclude the paper.

2. Edgeworth's Plural Median

In 1887, six years after publishing his pathbreaking *Mathematical Psychics*, Edgeworth began a series of papers "On a new method of reducing observations related to several quantities." These papers have been somewhat neglected in subsequent accounts of Edgeworth's contributions to statistics, notably Stigler (1978, 1986) so this occasion constitutes an opportunity to remedy this neglect.

2.1. **Absolute** *vs.* **Squared Errors.**

In fact, Edgeworth's method was not *entirely* new. In the 1760's Rudjer Boscovich had proposed estimating the ellipticity of the earth using five observations on the length of one degree of latitude, y_i, at various latitudes, θ_i (at Rome, Paris, Cape Hope, Quito, and a measurement in Lapland) by solving the problem

$$\min \sum |y_i - \alpha - \beta \sin^2 \theta_i|.$$

subject to the constraint that the mean residual, $n^{-1} \sum (y_i - \hat{\alpha} - \hat{\beta} \sin^2 \theta_i)$, equalled zero. Somewhat later Laplace showed that this problem could be solved by computing a weighted median. He also provided an astonishingly complete theory of the asymptotic behavior of the weighted median for the scalar case. This early history is described in considerable detail by Sheynin (1973), Stigler (1986) and Farebrother(1987).

Edgeworth's new method, which he called the "plural median" was intended to revive the Boscovich method as a direct competitor to the least squares approach which

had reigned supreme since it was introduced by Gauss and Legendre at the end of the 18th century. Edgeworth (1888) proposed dropping the zero mean constraint on the residuals, arguing that it conflicted with the median intent of the absolute error approach. Appealing to the results of Laplace, he conjectured that the resulting plural median would be more accurate than the corresponding least squares estimator when the observations were more "discordant", i.e. heavier tailed, than those from the Gaussian law of errors. Unfortunately, Edgeworth was unable to provide any mathematical support for his assertion that the superiority of the median over the mean in such discordant circumstances could be extended from Laplace's scalar context to his plural median. His argument rested entirely on the, admittedly rather compelling, analogy between the optimization problems for the univariate median and its plural counterpart in the regression context.

Laplace(1818) had considered the simple, scalar regression-though-the-origin model, in our (anachronistic) notation,

$$y_i = x_i\beta + u_i.$$

He showed that if the $\{u_i\}$ are iid with variance, σ^2, and have strictly positive density at the median, $f(0)$; the least squares estimator, $\hat{\beta} = x'y/x'x$ was asymptotically normal with mean, β and variance $\sigma^2/x'x$, while the ℓ_1 estimator, $\tilde{\beta}$, which minimized $\sum |y_i - x_i b|$ was asymptotically normal with mean β and variance $\omega^2/x'x$ where $\omega^2 = 1/(4f^2(0))$. If the u_i's were strictly Gaussian, of course, this implied that the asymptotic relative efficiency of the two estimators was

$$\text{ARE} = \text{Avar}(\tilde{\beta})/\text{Avar}(\hat{\beta}) = \omega^2/\sigma^2 = \pi/2 \approx 1.57$$

which would imply, in turn, that confidence intervals for β based on $\tilde{\beta}$ would be about 25 percent wider than those based on $\hat{\beta}$. Edgeworth (1888) anticipating work by Kolmogorov (1931) and Tukey(1960) showed that if, instead, the u_i's came from a scale mixture of Gaussians, i.e.

$$u_i \sim F(u) = (1-\epsilon)\Phi(u) + \epsilon\Phi(u/\sigma)$$

for some $\epsilon \in (0,1)$ and $\sigma > 1$, that $\tilde{\beta}$ "may be ever so much better than" $\hat{\beta}$.

Edgeworth's motivation for this line of inquiry goes back at least as far as his paper on "The method of least-squares" (Edgeworth (1883)) in which he considers the relative merits of ℓ_1, ℓ_2, and ℓ_∞ loss as measures of statistical performance.

> It is here submitted that these three criteria are equally right and equally wrong. The probable error, the mean error, the mean square of error, are forms divined to resemble in an essential feature the real object of which they are the imperfect symbols – the quantity of evil, the diminution of pleasure incurred by error. The proper symbol, it is submitted, for the quantity of evil incurred by a simple error is not any power of the error, nor any definite function at all, but an almost arbitrary function, restricted only by the conditions that it

> should vanish when the independent variable, the error, vanishes, and continually increase with the increase of the error.

Edgeworth asks whether the army shoemaker who makes a greater proportion of "exact fits" but occasionally makes a terribly painful mistake is better than the one who is more consistently inaccurate. He quotes Horace to the effect that one "who seldom tells a lie, but when he does, 'lies like a man', may do more harm than the habitual dealer in white lies." He concludes with a more "dignified example:" Which is better, he inquires, an instrument of observation whose errors follow the standard normal distribution or one whose errors arise from the Cauchy density,

$$f(u) = \frac{1}{\pi c} \frac{1}{1 + (u/c)^2}$$

"where c is small"? Clearly, the answer depends crucially on the form of the loss function. It is then a small step to the question: how should we estimate the central tendency of such measurements given a sample of observations from either instrument, a question which leads him to an extremely enlightened consideration of weighted least squares estimation. This discussion verges remarkably close to the enunciation of the principle of maximum likelihood.

2.2. **Computation of the Plural Median.** Writing near the end of his career, Edgeworth (1923), offers a reprise on his earlier work, addressing primarily the computational problems posed by the plural median. His 1888 paper had described a geometric approach to computing the plural median which brought Edgeworth to the brink of the simplex algorithm:

> The method may be illustrated thus:–Let $C - R$ (where C is a constant, [and R denotes the ℓ_1 objective function]) represent the height of a surface, which will resemble the roof of an irregularly built slated house. Get on this roof somewhere near the top, and moving continually upwards along some one of the edges, or *arrétes*, climb up to the top. The highest position will in general consist of a solitary pinnacle. But occasionally there will be, instead of a single point, a horizontal ridge, or even a flat surface.

Supplemented by a slightly more explicit rule for choosing the edges of ascent at each vertex, this description would fit nicely into modern textbooks of linear programming. Edgeworth's geometric approach to the computation of the plural median employed a device which has been recently termed the "dual plot" in the work of Rousseeuw and Hubert (1998). In the bivariate regression version of the dual plot, each point, (x_i, y_i) appears as a line in parameter space, that is,

$$a = y_i - bx_i$$

so all the points on this line in (a, b)-space have ith residual zero, and intersections of such lines correspond to points which have two zero residuals – basic solutions in the

terminology of linear programming. Edgeworth's strategy was to choose a point of intersection like this, and then to try to move to an adjacent intersection which reduced the sum of the absolute residuals. He choose the following rather unfortunate example to illustrate the technique: given the data $\{(6, y_1), (6, y_2), (1, y_3), (1, y_4), (1, y_5)\}$ with the y's ordered so that $y_1 < ... < y_5$ we obtain the dual plot illustrated in Figure 2.1. Starting at point A he computed the directional derivative of R along the paths Ap, Aq, Ar, and AB and determined that only the AB direction was a direction of descent. Moving to B, he again computed the directional derivatives of R in the directions Bs and BC and determined that R was increasing toward the point s and flat in the direction of C. From this, he concluded, quite correctly, that the plural median solution for this example consists of the entire interval connecting BC. Implicitly, there was a recognition in this approach that the problem was convex, and therefore any such path of descent would eventually arrive at a global minimum.

Gill, Murray and Wright (1991) distinguish between *direct* and *iterative* algorithm in the following way:

> ...we consider as *direct* a computation procedure that produces one and only one estimate of the solution, without needing to perform *a posteriori* tests to verify that the solution has been found... In contrast, an iterative method generates a sequence of trial estimates of the solution, called *iterates*. An iterative method includes several elements: an initial estimate of the solution; computable tests to verify whether or not an alleged solution is correct; and a procedure for generating the next iterate in the sequence if the current estimate fails the test.

Edgeworth's description falls short of this definition of a complete *iterative* algorithm, but it captures the essential ingredients of later algorithms based on simplex, or exterior point methods. Its chief failing, in my view, was not the omission of explicit rules to choose the next edge, or the lack of an explicit stopping rule; it was the iterative nature of the method itself. Iteration is rather like a voyage of exploration of the 15th century, sailing into the Atlantic perhaps even believing that the world was flat, not knowing when, or even if, the voyage would end. Direct algorithms like Gaussian elimination, on the other hand, made least squares like a trip along a familiar road; at each step one knew exactly how much more effort was necessary. Only with the emergence of computers in the 1940's were researchers able to transfer the risk, or uncertainty, of the iterative approach to the machine, and the spirit of adventure blossomed as investigators put down their pencils, poured their coffee, and watched the tapes whir and the lights flicker.

Edgeworth's example highlights an aspect of median regression which continues to perplex the unwary. The nonuniqueness of the solution, consisting of the entire line segment BC, should not be surprising in view of the fact that we have replicated design points at $x = 1$ and $x = 6$ with an even number of points (2) at the latter. Medians of even numbers of observations are inherently nonunique, and the median

regression solution in this "two sample problem" consists simply in connecting any median of sample 1, at $x = 1$, with any median at sample 2, at $x = 6$, a procedure which translated back into parameter space yields the line segment BC. Edgeworth argued that one could always select a central point from the set of solutions, but more importantly, he argued that the nonuniqueness was "apt to be slight and neglectable for it tends to be of an order which is insignificant in comparison to the probable error to which the solution is liable." This point which has been made by a number of other subsequent authors, including Bassett and Koenker (1978), is worth reiterating. Dupačová (1992) provides a detailed discussion of this issue.

Having argued, by analogy to Laplace's asymptotic theory for the scalar regression case, that median regression could be, under certain circumstances, more accurate than "the ordinary method" of least squares, Edgeworth concludes his 1888 paper with the startling claim that,

> On the other hand, the labour of extracting the former is rather less: I should think, in the case of many unknown variables. At the same time that labour is more "skilled". There may be needed the attention of a mathematician; and, in the case of many unknowns, some power of hypergeometrical conception. Perhaps the balance of advantage might be affected by an *à priori* knowledge of an approximate solution.

While a good starting value is always welcome, this seems to be symptomatic of a rather virulent case of wishful thinking. Certainly, it did nothing to impede the hegemony of least squares methods over the next century. I would like to return Edgeworth's conjecture regarding the computability of the median regression estimator in the next section, but before doing so I would like to consider the dual plot more carefully in light of some interesting recent developments in quantile regression.

Edgeworth's adoption of the dual plot as a computational device for median regression was his response to criticism of his prior computational proposal. The dual plot is an extremely valuable instrument for focusing attention on the geometric essentials of the problem – movement from one adjacent vertex to another in much the same spirit as the modern simplex algorithm. However, the dual plot had two apparent disadvantages. The first was that it seemed inherently limited to the case of bivariate regression and therefore failed to suggest a natural extension to the multivariate case. This could have been remedied easily by carefully writing down the algebra corresponding to the geometric interpretation of the algorithm. The second disadvantage was that it introduced an irresistible temptation to invent variations on the original principle of minimizing the sum of absolute residuals.

2.3. **Regression Depth: Prospects and Antecedants.** One of these variations was described by Bowley (1902) in some work on the distribution of English wages:

> ... we may with advantage apply Prof. Edgeworth's "double median" method and find the point, line or small area, such that, whether we

> proceed from it to the left, or right, or up, or down, we always intersect the same number of lines before we are clear of the network.

This idea has been independently rediscovered, more precisely formulated and greatly elaborated in recent work of Rousseeuw and Hubert (1998). Their formulation in terms of the dual plot is:

> The [regression depth] of a fit θ is (in dual space) the smallest number of lines L_i that need to be removed to set θ free, i.e. so that it lies in the exterior of the remaining arrangement of lines.

In the space of the original observations, where we have data $Z_n = \{(x_i, y_i) : i = 1, ..., n\} \in \mathbf{R}^2$ and the model

$$y_i = \theta_1 x_i + \theta_2 + u_i \tag{2.1}$$

Rousseeuw and Hubert formulate the following complementary definitions:

Definition 2.1. *A candidate fit $\theta = (\theta_1, \theta_2)$ to Z_n is called a nonfit iff there exists a real number, $v_\theta = v$ which does not coincide with any x_i and such that*

$$r_i(\theta) < 0 \;\; \textit{for all}\; x_i < v \;\textit{and}\; r_i(\theta) > 0 \;\; \textit{for all}\; x_i > v$$

or

$$r_i(\theta) > 0 \;\; \textit{for all}\; x_i < v \;\textit{and}\; r_i(\theta) < 0 \;\; \textit{for all}\; x_i > v$$

where $r_i(\theta) = y_i - \theta_1 x_i - \theta_2$.

Definition 2.2. *The regression depth of a fit $\theta = (\theta_1, \theta_2)$ relative to a data set $Z_n \in \mathbf{R}^2$ is the smallest number of observations that need to be removed to make θ a nonfit.*

A mechanical description of regression depth in the "primal" or data-space plot is also provided by Rousseeuw and Hubert: the existence of v_θ for any nonfit θ, corresponds to a point on the line $y = \theta_1 x + \theta_2$ about which one could rotate the line to the vertical without encountering any observations.

Suppose that we have ordered the observations so that $x_1 < x_2 < \ldots < x_n$, and define

$$L^+(t) = \sum_{i \le nt} I(r_i \ge 0), \qquad R^+(t) = \sum_{i > nt} I(r_i \ge 0)$$

with

$$L^-(t) = nt - L^+(t), \qquad R^-(t) = n(1-t) - R^+(t).$$

We may view the current fit $y = \theta_1 x + \theta_2$ as dividing the plane into an upper half and lower half, and the vertical line at $x = x_{([nt])}$ as dividing the plane into right and left halves. Then L^+, R^+, L^-, R^- are simply the observation counts in each of the corresponding four "quadrants". Regression depth as defined by Rousseeuw and Hubert is then given by

$$d(\theta) = \min_t \{\min\{L^+(t) + R^-(t), R^+(t) + L^-(t)\}\}$$

which can be easily seen by imagining the mechanical analogy of rotating the fit through the point $x = x_{([nt])}$ to the vertical. Rotation clockwise passes through $L^+(t) + R^-(t)$ points, while rotation counterclockwise passes through $R^+(t) + L^-(t)$. The smaller of these quantities, when minimized over t gives the depth of the point θ.

Bowley's description, if interpreted a bit generously to include *all* directions, not just the canonical ones, characterizes what Rousseeuw and Hubert call the "deepest line" or maximal regression depth estimator. Contrary to Bowley's claim, it is not Edgeworth's double median, but it is a quite distinct and extremely intriguing alternative. As Rousseeuw and Hubert elegantly demonstrate it may be viewed as a natural extension to regression of the halfspace depth ideas of Tukey (1975), Donoho and Gasko (1992).

In Figure 2.2 we illustrate an empirical example taken from Bowley (1902) in which we have six observations, and therefore six lines in the dual plot. In this example the median regression solution is represented by the point D, which is "exposed" on the outer edge of the dual plot. The maximum regression depth is attained by the points $\{A, B, C\}$ all of which have depth three. Note that other points of the triangle ABC have only regression depth 2, thus illustrating the potentially disconnected nature of the maximal depth solution set.

Unlike the conventional median regression (ℓ_1) estimator which has a breakdown point of $1/n$ in the (x, y)-contamination model, and only marginally better breakdown properties in the fixed-x, y-contamination model, as discussed in He, Jurečková , Koenker, and Portnoy (1993) and Mizera and Muller (1997), the deepest line estimator has breakdown point $1/3$. It shares the equivariance properties of the ℓ_1 estimator, but exhibits a somewhat greater tendency toward non-uniqueness.

It is worth remarking in this connection that Theil's (1950) earliest papers also deal with a variant of this type which is usually described as the "median of pairwise slopes" and may be viewed geometrically in the dual plot by projecting all the the intersections onto the axis of the "slope" parameter and then choosing the median of these projected values. In the preceding example Theil's estimator is the point B. We might also note that the dual plot plays a prominent role in recent work in political science by King (1997) about the so-called ecological inference problem.

The contrast between the deepest line estimator and the usual median regression estimator is, perhaps, most clearly seen in their asymptotic behavior, which has been recently studied by He and Portnoy (1998). It is well known that

$$\hat{\theta} = \arg\min_{\theta \in \mathbf{R}^2} \sum |y_i - \theta_1 x_i - \theta_2|$$

satisfies, under mild conditions given in Bassett and Koenker (1978) and related work by numerous subsequent authors,

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, \omega^2 D^{-1})$$

where $\theta_0 = (\theta_1, \theta_2)'$, $\omega^2 = 1/(4f^2(0))$ and

$$\lim_{n\to\infty} n^{-1}X'X \to D$$

with $X = (x_i, 1)_{i=1}^n$.

In contrast, the deepest line estimator may be formulated as

$$\tilde{\beta}_n = \text{ argmin } \max_{x_{(1)}\le a\le x_{(n)}} |D_n(b,a)|$$

where

$$D_n(b,a) = \sum \text{ sgn } \{(y_i - \theta_1 x_i - \theta_2)(x_i - a)\}.$$

To formulate an asymptotic theory for the maximum regression depth estimator He and Portnoy (1997) assume that the sequence $\{x_i\}$ satisfies the conditions:

A1.) $\sum x_i^2 = O(n)$

A2.) $n^{-1}\sum x_i \text{ sgn } (x_i - x_{[tn]}) \to g_1(t)$ uniformly from $t \in (0,1)$, with $g_1''(t) < 0$ for all t.

In addition, they assume,

A3.) The $\{u_i\}$'s are iid random variables with median zero, bounded density f, $f(0) > 0$, and that f is Lipschitz in a neighborhood of zero.

When the $\{x_i\}$'s are iid from distribution function G with positive density on its entire support, they note that

$$g_1(t) = \int_t^1 G^{-1}(u)du - \int_0^t G^{-1}(u)du$$

so $g_1'(t) = -2G^{-1}(t)$ and therefore, (A2) follows immediately from the Kolmogorov strong law and the monotonicity of G^{-1}. Now let $g_0(t) = 1 - 2t$ denote the limit of $n^{-1}\sum \text{ sgn } (z_i - z_{[nt]})$ and set $g(t) = (g_0(t), g_1(t))'$. He and Portnoy prove the following theorem.

Theorem 2.1. *Under conditions A1 - 3, $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a random variable whose distribution is that of the unique minimizer of the function*

$$h(\delta) = \max_t |2B(t) - B(1) + 2f(0)g(t)'\delta|$$

where $B(t)$ is standard Brownian motion.

Unfortunately, it is rather difficult to compare the asymptotic performance of the maximal depth estimator with the more familiar median regression estimator even in this simple iid-error bivariate setting. Whether the efficiency bound result of Newey and Powell (1990) applies to the maximum depth estimator is not entirely clear given the non-Gaussian limiting distribution of the maximal depth estimator. Even under non-iid error conditions, as long as the conditional median function is linear in parameters, both approaches can be shown to be $\sqrt{n}$-consistent for the same parameter; this is already quite remarkable. We would expect that the improved robustness of the maximal depth estimator would come at the price of some efficiency

loss under the idealized conditions A1-3 where influential design observations are highly desirable. He and Portnoy provide a very limited evaluation of the asymptotic relative efficiency of the two estimates which is reported in Table 2.1.

Given that the maximal depth estimator consistently estimates the linear conditional median function under essentially similar conditions to those required by the ℓ_1-estimator, it is natural to ask whether it is possible to estimate the parameters of other linear conditional quantile models using similar methods. A simple reweighting of the maximal depth objective function analogous to the asymmetric reweighting of the absolute value loss in Koenker and Bassett (1978) allows us to answer this question affirmatively.

Asymmetrically reweighting positive and negative residuals as in Koenker and Bassett (1978) suggests the quantile regression depth function

$$d_\tau(\theta) = \min_t\{\min\{\tau L^+(t) + (1-\tau)R^-(t), \tau R^+(t) + (1-\tau)L^-(t)\}\}$$

and essentially the same asymptotic analysis of He and Portnoy shows that the minimizer

$$\hat{\theta}_n(\tau) = \text{argmin } d_\tau(\theta)$$

is a $\sqrt{n}$ consistent estimator of the parameters of the linear τ^{th} conditional quantile function.

Thus, regression depth provides an alternative "influence robust" approach to quantile regression estimation which could be compared to the earlier GM-type weighting proposals of Antoch and Jurečková (1985) and DeJongh, DeWet and Welsh (1988). Extending the regression depth idea beyond the bivariate model poses some challenges particularly on the asymptotic and algorithmic fronts, but the basic conceptual apparatus is already provided by Rousseeuw and Hubert (1998), and Rousseeuw and Struyf (1998). It would also be valuable to explore connections to the maximum score estimator for binary response models, Manski(1985), both from a computational and a statistical point of view. It is a pleasant irony that what can only be regarded as misguided attempt by Bowley (1902) to implement the method of his intellectual inspiration, Francis Edgeworth, could be independently rediscovered nearly a century later, and developed into a valuable complement to Edgeworth's proposed methods.

3. Frisch's Radar

In this Section, I would like to briefly describe some recent developments which hold the promise of eventually vindicating Edgeworth's second conjecture that median regression, and thus quantile regression in general, can be made "less labourious" than the method of least squares. Since this claim is bound to appear farfetched to many readers, it is perhaps worth pausing to consider its validity in the simplest possible setting: Which is easier to compute – the median or the mean?

It will be immediately clear that the median is "easier" from the viewpoint of hand computation, especially if the observations are provided to high precision. This was

a part of the motivation for related "median polish" methods for robust ANOVA introduced by Tukey. However, few of us do any computing by hand these days, so this is hardly compelling evidence for Edgeworth's claim.

On the computer, it is clear that we can compute the mean in $\mathcal{O}(n)$ operations (n additions and 1 division), while naively it would appear that computing the median requires $n \log n$ operations (comparisons) to sort the observations. Further reflection suggests, however, that we don't really need to fully sort *all* the observations, a careful partial sort would suffice. Several proposals along this line have been made, culminating in the algorithm of Floyd and Rivest (1975), which showed that the median can also be computed in $\mathcal{O}(n)$ operations. Even the casual reader of their paper would quickly grant that Floyd and Rivest's algorithm is "more skilled" than the one line program need to implement the sample mean computation, but once implemented the skill is reified. And since the comparisons of the median algorithm are generally quicker than the additions of the mean algorithm, it is not implausible that the handicraft superiority of the median can be rescued.

In the regression setting, the development of the simplex method of solving linear programming problems by Dantzig and others in the late 1940's was rapidly recognized as an effective tool for minimizing sums of absolute errors. Refinements of simplex designed for the ℓ_1 regression problem including the widely implemented Barrodale and Roberts (1974) algorithm have proven to be quite competitive with least squares in computational speed for problems up to a few thousand observations. In Figure 3.1 we illustrate this based on experience in Splus on a Sparc 20. Note that up to about 3000 observations for $p = 4$ parameters, or up to about n=1000, for $p = 8$, or $n = 300$ for $p = 16$, the Splus function `l1fit` implementing the algorithm of Barrodale and Roberts is actually faster than the QR decomposition algorithm for least squares embodied the the Splus function `lm()`. However in larger problems the simplex approach founders badly, exhibiting quadratic growth in cpu-time with n. By the time that we reach $n = 100,000$, with $p = 16$ for example, `l1fit` requires nearly an hour of Sparc 20 time while the equivalent least squares computation takes about 10 seconds.

The parable which has evolved in Urbana to describe the experience reported in Figure 3.1 we call the "Gaussian Hare and the Laplacian Tortoise." At least from a purely computational vantage point, the least-squares methods championed by Gauss seems destined to triumph over the inevitably slower absolute error methods of Laplace. The house that the tortoise carries around on his back to protect himself against inclement statistical weather must come at a price. Or does it? The hare may frolic in the flowers allowing the tortoise an advantage in the sprints, but the plodding tortoise can't win the longer races. Or can it? In Portnoy and Koenker (1997), we explore two new approaches to absolute error computation which, taken together, provide some reason for optimism about Edgeworth's conjecture. I will briefly describe both approaches, relegating many details to the original paper.

Consider the median regression problem,

$$\min_{b \in \mathbf{R}^p} \sum_{i=1}^{n} |y_i - x_i' b| \tag{3.1}$$

which may be formulated as the linear program,

$$\min\{\ e'u + e'v \ \mid\ y = Xb + u - v, (u,v) \in \mathbf{R}^{2n}_+\}, \tag{3.2}$$

where e denotes an n-vector of ones. Note that we have simply decomposed the regression residual vector into its positive and negative parts, calling them u and v, and written the original problem as one of minimizing a linear function of the $2n$-vector (u, v) subject to n linear equality constraints and $2n$ linear inequality constraints. This "primal" linear program formulation of the ℓ_1-regression problem has an associated "dual" formulation in which we maximize with respect to a vector, $d \in \mathbf{R}^n$, which may be viewed as the vector of Lagrange multipliers associated with the equality constraints of the primal problem. This dual formulation is,

$$\max\{y'd \ \mid\ X'd = 0, \quad d \in [-1,1]^n\}, \tag{3.3}$$

or equivalently, setting $a = d + \frac{1}{2}e$,

$$\max\{y'a \ \mid\ X'a = \tfrac{1}{2}X'e, \ \ a \in [0,1]^n\}. \tag{3.4}$$

The simplex approach to solving this problem may be briefly described as follows. A p-element subset of $\mathcal{N} = \{1, 2, ..., n\}$ will be denoted by h, and $X(h), y(h)$ will denote the submatrix and subvector of X, y with the corresponding rows and elements identified by h.

Recognizing that solutions to (3.1) may be characterized as planes which pass through precisely $p = \dim(b)$ observations, or as convex combinations of such "basic" solutions, we can begin with any such solution, which we may write in the primal formulation as,

$$b(h) = X(h)^{-1} y(h). \tag{3.5}$$

We may regard any such "basic" primal solution as an extreme point of the polyhedral, convex constraint set. In the dual formulation since the index set h identifies the active constraints of the primal problem, i.e. those observations for which both u_i and v_i are zero, $a(h)$ lies in the interior of the p dimension unit cube, and the complement of h corresponds to coordinates of a which lie on the boundary: if $u_i > 0$ then $a_i = 1$, while if $v_i > 0$ then $a_i = 0$. A natural algorithmic strategy is then to move to the adjacent vertex of the constraint set in the direction of steepest descent. This transition involves two stages: the first chooses a descent direction by considering the removal of each of the current basic observations and computing the gradient in the resulting direction, then having selected the direction of steepest descent and thus an observation to be removed from the currently active "basic" set, find the maximal step length in the chosen direction by searching over the remaining $n - p$

available observations for a new element to introduce into the "basic" set. Each of these transitions involves an elementary "simplex pivot" matrix operation to update the current basis. The iteration continues in this manner until no direction is found at which point the current $b(h)$ can be declared optimal. This is, in effect, just the natural iteration proposed by Edgeworth for the simple bivariate regression problem using the dual plot.

To illustrate the shortcomings of the simplex method, or indeed of any method which travels around the exterior of the constraint set like this, one need only imagine the number of vertices in a typical median regression problem, which is of order, $\binom{n}{p} = \mathcal{O}(n^p)$. A poor starting point in a moderately large problem may entail an enormous number of pivots. Even in the Barrodale and Roberts (1974) formulation which is an enormous improvement over conventional simplex algorithms in ℓ_1-type problems, we observe linear growth, in n, in the number of iterations (pivots), and in the effort per iteration, yielding the quadratic growth in cpu-time observed in Figure 3.1.

Although prior work in the Soviet literature offered theoretical support for the idea that linear programs could be solved in polynomial time, thus avoiding certain pathological behavior of simplex methods, the paper of Karmarker (1984) constituted a watershed in the numerical analysis of linear programming. It offered not only a cogent argument for the polynomiality of interior point methods of solving LP's, but also provided for the first time direct evidence that interior point methods were demonstrably faster than simplex in specific, large, practical problems.

But it is an interesting irony, illustrating the spasmodic progress of science, that the most fruitful practical formulation of the interior point revolution of Karmarker (1984) can be traced back to a series of Oslo working papers by Ragnar Frisch in the early 1950's. The basic idea of Frisch (1956) was to replace the linear inequality constraints of the LP, by what he called a log barrier, or potential, function. Thus, in place of the canonical linear program,

$$\min \ \{c'x \ \mid \ Ax = b, \ \ x \geq 0\}, \tag{3.6}$$

we may associate the logarithmic barrier reformulation

$$\min \ \{B(x,\mu) \ \mid \ Ax = b\} \tag{3.7}$$

where

$$B(x,\mu) = c'x - \mu \sum \log x_k. \tag{3.8}$$

In effect, (3.7) replaces the inequality constraints in (3.6) by the penalty term of the log barrier. Solving (3.7) with a sequence of parameters μ such that $\mu \to 0$ we obtain in the limit a solution to the original problem (3.6). The salient virtue of the log barrier formulation is that, unlike the original formulation, it yields a differentiable objective function which is consequently attackable by Newton's method and under easily verifiable conditions inherits the quadratic convergence of Newton's

method. This approach was elaborated in Fiacco and McCormick (1968) for general constrained optimization problems, but was only revived as a linear programming tool after its close connection to the approach of Karmarkar (1984) was pointed out by Gill, et al. (1986). Frisch (1956) described it in the following vivid terms,

> My method is altogether different than simplex. In this method we work systematically from the interior of the admissible region and employ a logarithmic potential as a guide – a sort of radar – in order to avoid crossing the boundary.

See Wright(1992), Gonzaga (1992), and Wright(1996) for excellent introductions to the interior point literature.

Quantile regression, as introduced in Koenker and Bassett (1978), places asymmetric weight on positive and negative residuals, and solves the slightly modified ℓ_1 problem,

$$\min_{b\in\mathbf{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b) \tag{3.9}$$

where $\rho_\tau(r) = r(\tau - I(r<0))$ for $\tau \in (0,1)$. This yields the modified linear program,

$$\min\{\tau e'u + (1-\tau)e'v \mid y = Xb + u - v, (u,v) \in \mathbf{R}^{2n}_+\}, \tag{3.10}$$

and has dual formulation,

$$\max\{y'a \mid X'a = (1-\tau)X'e,\ \ a \in [0,1]^n\}. \tag{3.11}$$

The dual formulation of the quantile regression problem fits nicely into the standard formulations of interior point methods for linear programs with bounded variables. The function $a(\tau)$ that maps $[0,1]$ to $[0,1]^n$ plays a crucial role in connecting the statistical theory of quantile regression to the classical theory of rank tests as described in Gutenbrunner and Jurečková (1992) and Gutenbrunner, Jurečková , Koenker and Portnoy (1993).

Adding slack variables s, and the constraint $a+s = e$, we obtain the barrier function

$$B(a,s,\mu) = y'a + \mu \sum_{i=1}^{n} (\log a_i + \log s_i), \tag{3.12}$$

which should be maximized subject to the constraints, $X'a = (1-\tau)X'e$ and $a+s=e$. The Newton step, δ_a, solving

$$\max y'\delta_a + \mu\delta_a'(A^{-1} - S^{-1})e - \tfrac{1}{2}\mu\delta_a'(A^{-2} + S^{-2})\delta_a \tag{3.13}$$

subject to $X'\delta_a = 0$, satisfies

$$y + \mu(A^{-1} - S^{-1})e - \mu(A^{-2} + S^{-2})\delta_a = X\delta_b \tag{3.14}$$

for some $\delta_b \in \mathbf{R}^p$, and δ_a such that $X'\delta_a = 0$. Here we follow a notational convention of the interior point literature denoting $A = \operatorname{diag}(a)$, and $S = \operatorname{diag}(s)$. Multiplying

through by $X'(A^{-2}+S^{-2})^{-1}$ and using the constraint, we can solve explicitly for the vector δ_b,

$$\delta_b = (X'WX)^{-1}X'W(y+\mu(A^{-1}-S^{-1})e) \tag{3.15}$$

where $W = (A^{-2}+S^{-2})^{-1}$. The vector δ_b is the vector of Lagrange multipliers on the equality constraints of the dual formulation, and since the dual of the dual is the primal, it provides a search direction in the primal space for the vector b. Setting $\mu = 0$ in (3.15) yields a version of the affine scaling interior point algorithm for this problem. Given the direction δ_b we can solve for δ_a, δ_s, etc. and compute a step length which takes us a fixed proportion of the distance to the boundary of the constraint set. Then taking this step, updating, and continuing the iteration provides a simple implementation of Frisch's radar. In Portnoy and Koenker (1997) we describe a slightly more complicated version of the interior point approach due to Mehrotra (1992) which has been widely implemented and seems to provide a somewhat more efficient and more robust approach. Particularly on large problems this interior point approach performs vastly better than earlier simplex implementations. For example, problems with $n = 100,000$ and $p = 16$ which take about one hour using Barrodale and Roberts algorithm, can be done in about one minute using the new interior point approach. Problems of this size are not atypical of current practice in labor economics for example, and particularly when researchers are considering bootstrapping strategies the difference in performance can be crucial.

Still, the gentle reader, imbued with the Gaussian faith, may be thinking: "a whole minute? it can't take more than 10 seconds to compute the least squares estimate for a problem like that on a Sparc 20." This is quite correct, so interior point methods are not sufficient to rescue Laplace's tortoise from yet another century of humiliation. Formal computational complexity results indicate that for large problems primal-dual implementations of interior point methods for solving ℓ_1 problems require $\mathcal{O}(np^3 \log^2 n)$ operations which is considerably better than the quadratic in n behavior of simplex, but still inferior to the $\mathcal{O}(np^2)$ behavior of least squares.

Fortunately, further gains are possible from careful preprocessing of ℓ_1 type problems. Preprocessing rests on an extremely simple idea which is closely connected to the ideal of partial sorting underlying the $\mathcal{O}(n)$ median algorithm of Floyd and Rivest. If, by preliminary estimation, or some other form of statistical necromancy, we could determine the signs of the residuals for a significant group of observations, we could then combine observations with positive residuals into a single "globbed" observation, and similarly glob together the negative observations, so that the original problem,

$$\min \sum_{i=1}^{n} |y_i - x_i'b| \tag{3.16}$$

would be equivalent to,

$$\min \sum_{i \in N \setminus J_L \cup J_H}^{n} |y_i - x_i'b| + |y_L - x_L'b| + |y_H - x_H'b| \tag{3.17}$$

where $N = \{1, 2, ..., n\}$, $x_K = \sum_{i \in J_K} x_i$ for $K \in \{K, L\}$ and y_L and y_H can be chosen arbitrarily small and large respectively, to ensure that the corresponding residuals on the globbed observations remain negative and positive. In this process we have reduced the problem of n original observations to $n - \sharp\{J_L, J_H\} + 2$ observations so if the cardinality of the J-sets is large we have gained substantially. Under plausible sampling assumptions we can, based on a preliminary subsample of m observations, make a prediction region for $\{x_i\beta : i = 1, 2, ..., n\}$ of width $\mathcal{O}(p/\sqrt{m})$, so assigning observations above this region to J_H and observations below this region to J_L, we would have $M = \mathcal{O}_p(np/\sqrt{m})$ observations falling inside the region. This is illustrated in Figure 3.2.

Minimizing the computational effort required to compute the preliminary fit based on m observations plus the effort required for the solution of the globbed problem (3.2) with M observations, we obtain $m^* = \mathcal{O}((np)^{2/3})$, which under our claimed performance of the underlying interior point algorithm yields a complexity for the full problem of

$$C = \mathcal{O}_p(n^{2/3}p^3 \log^2 n) + \mathcal{O}(np^2), \tag{3.18}$$

where the first term comes from the solution of the two median regression problems of size $\mathcal{O}(n^{2/3})$ and the second term arises from the computation of the confidence band.

Further details are provided in Portnoy and Koenker(1997) and I will comment only briefly here on the important fact that any implementation of this preprocessing approach must verify that the solution to the globbed problem actually agrees with the predicted signs based on the confidence region. The simultaneous confidence region can be chosen to assure this with arbitrarily high probability, and the eventuality that we may need to repeat the cycle to remedy some inaccurately predicted signs introduces another multiplicative factor which does not affect the orders in probability in the complexity computation.

The crucial consequence of the formal complexity theory and the extensive concomitant empirical testing of our implementation of the algorithm is that the computational effort required for quantile regression can be made comparable with the effort required for least squares over the full range of currently plausible problem dimensions. In the final empirical example of Portnoy and Koenker (1997), we compare timings for a typical large econometric application of quantile regression with $n = 113,547$ and $p = 6$. With the new algorithm, quantile regression estimates take about 10 seconds on a Sparc-Ultra, comparable to the least squares time of 8 seconds. Interior point methods applied to the full problem before preprocessing requires

about a minute for these problems. Simplex solution of the same quantile regression problems requires approximately an hour on the same machine.

4. Future Prospects

Galton's admonition that we should revel in the "charms of variety" of matters statistical, not throw the mountains into the lakes, certainly asks that we venture beyond "regression to the mean." As Mosteller and Tukey (1977) put it in their influential text:

> What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

Quantile regression offers a means to accomplish this important task. Models for conditional quantile functions offer a number of advantages over more familiar models for conditional moments. By allowing the investigator to focus attention *locally* on specific segments of the conditional distribution, we achieve an inherent robustness and a natural interpretability which are inherited from the behavior of the ordinary sample quantiles. The fact that quantiles commute with monotonic transformations of the response is also particularly convenient as Powell (1991) and others have emphasized. By replacing a monolithic model of conditional central tendency with a family of models for conditional quantiles we are able to achieve considerably greater flexibility and a much more complete view of the effect of the covariates on the distribution of the response.

Edgeworth's improbable conjectures that the plural median has asymptotic behavior like Laplace's scalar (weighted) median, and that it could be made "less labourious" than least-squares computation are both almost fully vindicated. There are still many important problems which remain, of course. Progress, as the foregoing historical remarks have indicated, has been highly episodic. But I would like to think that now we have reached a critical stage in the research process, and ℓ_1 methods and quantile regression methods more generally, will finally enjoy a sustained development.

In concluding, I can not resist one final quotation from Frisch (1963), who, late in his career, was exploring solving systems of nonlinear equations in large general equilibrium macro models by minimizing sums of absolute versus sums of squared errors. He comments about this choice:

> I have also an intuitive feeling that *even in principle the sum of absolute values is better than the sum of squares.* I cannot substantiate this feeling theoretically, but I can point to some empirical evidence... At this writing my preference is therefore for the sum of absolute values even though this will entail discontinuities in the derivatives. The discontinuities which will occur when we work with absolute values [are] after all not very serious, they can be handled by choosing between the forward and the backward derivatives, an operation which the machine can do very quickly whether it concerns a total or a partial derivative. [my italics]

Now *this* is a principle of econometrics worth defending!