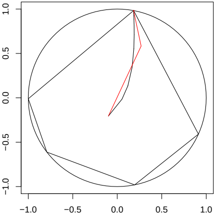
Quantile Regression Computation: From Outside, Inside and the Proximal

Roger Koenker

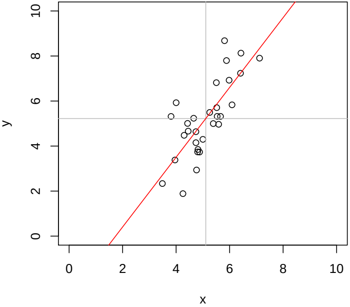
University of Illinois, Urbana-Champaign

University of Minho 12-14 June 2017



The Origin of Regression – Regression Through the Origin

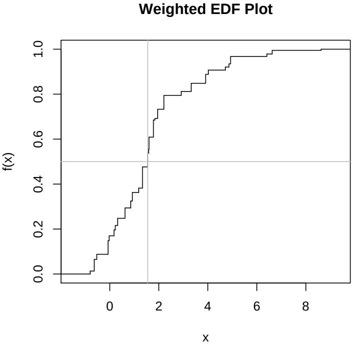
In 1755, well before Gauss and Legendre argued over who had invented the method of least squares, the Croatian Jesuit Rudjer Boscovich proposed to estimate linear relationships by finding the line with mean residual zero that minimized the sum of absolute residuals.



Problem: $\min_{\alpha,\beta} \sum_{i=1}^{n} |y_i - \alpha - x_i\beta| \quad \text{s.t.} \quad \bar{y} = \alpha + \bar{x}\beta.$

Boscovich/Laplace *Methode de Situation*

Algorithm: Order the n candidate slopes: $b_i = (y_i - \bar{y})/(x_i - \bar{x})$ denoting them by $b_{(i)}$ with associated weights $w_{(i)}$ where $w_i = |x_i - \bar{x}|$. Find the weighted median of these slopes.

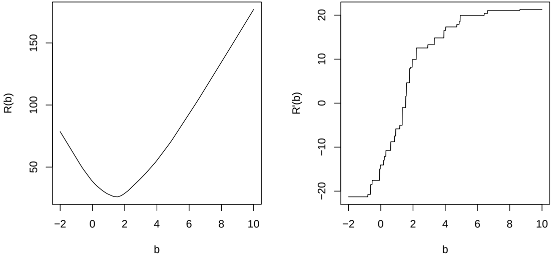


Variant: Theil's (1950) median of pairwise slopes.

Methode de Situation via Optimization

$$R(b) = \sum |\tilde{y}_i - \tilde{x}_i b| = \sum |\tilde{y}_i/\tilde{x}_i - b| \cdot |\tilde{x}_i|.$$

$$R'(b) = -\sum \text{sgn}(\tilde{y}_i/\tilde{x}_i - b) \cdot |\tilde{x}_i|.$$



Quantile Regression through the Origin in R

This can be easily generalized to compute quantile regression estimates:

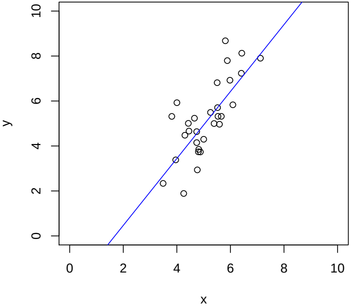
```
wquantile <- function(x, y, tau = 0.5) {
        o <- order(y/x)
        b <- (y/x)[o]
        w <- abs(x[o])
        k <- sum(cumsum(w) < ((tau - 0.5) * sum(x) + 0.5 * sum(w)))
        list(coef = b[k + 1], k = o[k+1])
}
```

Warning: When $\bar{x} = 0$ then τ is irrelevant. Why?

Edgeworth's (1888) Plural Median

What if we want to estimate both α and β by median regression?

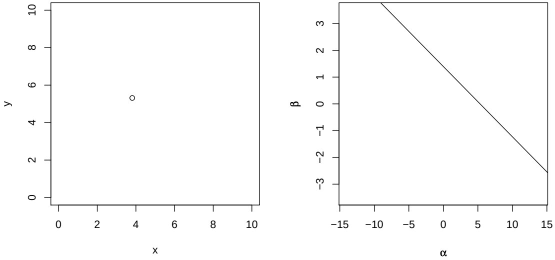
Problem: $\min_{\alpha,\beta} \sum_{i=1}^{n} |y_i - \alpha - x_i\beta|$



Edgeworth's (1888) Dual Plot: Anticipating Simplex

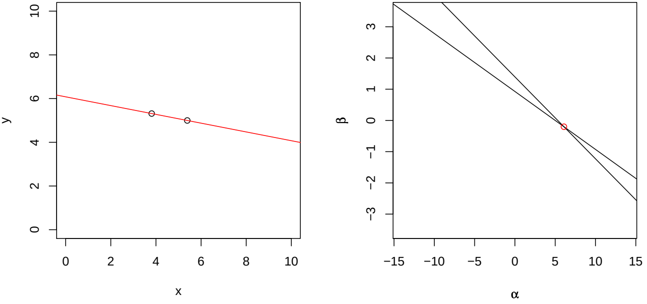
Points in sample space map to lines in parameter space.

$$(x_i, y_i) \mapsto \{(\alpha, \beta) : \alpha = y_i - x_i\beta\}$$



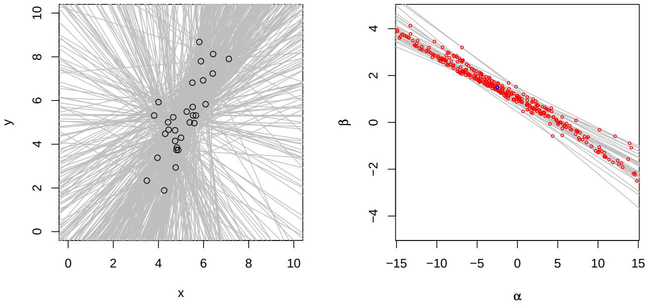
Edgeworth's (1888) Dual Plot: Anticipating Simplex

Lines through pairs of points in sample space map to points in parameter space.



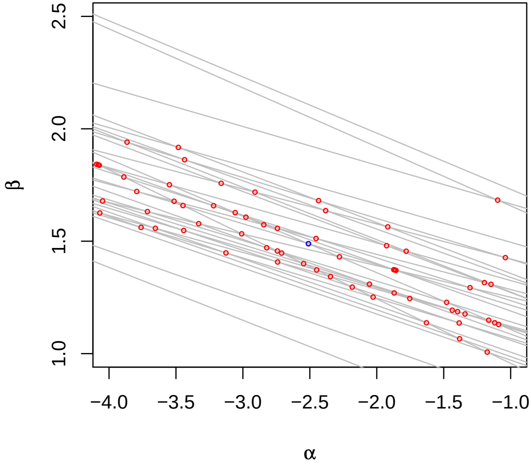
Edgeworth's (1888) Dual Plot: Anticipating Simplex

All pairs of observations produce $\binom{n}{2}$ points in dual plot.



Edgeworth's (1888) Dual Plot: Anticipating Simplex

Follow path of steepest descent through points in the dual plot.



Barrodale-Roberts Implementation of Edgeworth

```
rqx<- function(x, y, tau = 0.5, max.it = 50) { # Barrodale and Roberts -- lite
        p <- ncol(x); n <- nrow(x)
        h <- sample(1:n, size = p) #Phase I -- find a random (!) initial basis
        it <- 0
        repeat {
                it <- it + 1
                Xhinv <- solve(x[h,  ])
                bh <- Xhinv %*% y[h]
                rh <- y - x %*% bh
        #find direction of steepest descent along one of the edges
                g <-  - t(Xhinv) %*% t(x[ - h,  ]) %*% c(tau - (rh[ - h] < 0))
                g <- c(g + (1 - tau),  - g + tau)
                ming <- min(g)
                if(ming >= 0 || it > max.it) break
                h.out <- seq(along = g)[g == ming]
                sigma <- ifelse(h.out <= p, 1, -1)
                if(sigma < 0) h.out <- h.out - p
                d <- sigma * Xhinv[, h.out]
        #find step length by one-dimensional wquantile minimization
                xh <- x %*% d
                step <- wquantile(xh, rh, tau)
                h.in <- step$k
                h <- c(h[ - h.out], h.in)
        }
        if(it > max.it) warning("non-optimal solution: max.it exceeded")
        return(bh)
}
```

Linear Programming Duality

Primal: $\min_x \{c^\top x | Ax - b \in T,\ x \in S\}$
Dual: $\max_y \{b^\top y | c - A^\top y \in S^*,\ y \in T^*\}$

The sets S and T are closed convex cones, with dual cones S^* and T^*. A cone K^* is dual to K if:

$$K^* = \{y \in \mathbb{R}^n | x^\top y \geqslant 0 \text{ if } x \in K\}$$

Note that for any feasible point (x, y)

$$b^\top y \leqslant y^\top Ax \leqslant c^\top x$$

while optimality implies that

$$b^\top y = c^\top x.$$

Quantile Regression Primal and Dual

Splitting the QR "residual" into positive and negative parts, yields the primal linear program,

$$\min_{(b,u,v)} \{\tau 1^\top u + (1-\tau) 1^\top v \mid Xb + u - v - y \in \{0\}, \quad (b,u,v) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}\}.$$

with dual program:

$$\max_d \{y^\top d \mid X^\top d \in \{0\}, \quad \tau 1 - d \in \mathbb{R}_+^n, \quad (1-\tau) 1 + d \in \mathbb{R}_+^n\},$$

$$\max_d \{y^\top d \mid X^\top d = 0, \; d \in [\tau - 1, \tau]^n\},$$

$$\max_a \{y^\top a \mid X^\top a = (1-\tau) X^\top 1, \quad a \in [0,1]^n\}$$

Quantile Regression Dual

The dual problem for quantile regression may be formulated as:

$$\max_a \{y^\top a | X^\top a = (1-\tau)X^\top 1,\ a \in [0,1]^n\}$$

What do these $\hat{a}_i(\tau)$'s mean statistically?
They are regression rank scores (Gutenbrunner and Jurečková (1992)):

$$\hat{a}_i(\tau) \in \begin{cases} \{1\} & \text{if} \quad y_i > x_i^\top \hat{\beta}(\tau) \\ (0,1) & \text{if} \quad y_i = x_i^\top \hat{\beta}(\tau) \\ \{0\} & \text{if} \quad y_i < x_i^\top \hat{\beta}(\tau) \end{cases}$$

The integral $\int \hat{a}_i(\tau) d\tau$ is something like the rank of the ith observation. It answers the question: On what quantile does the ith observation lie?

Linear Programming: The Inside Story

The Simplex Method (Edgeworth/Dantzig/Kantorovich) moves from vertex to vertex on the outside of the constraint set until it finds an optimum.

Interior point methods (Frisch/Karmarker/et al) take Newton type steps toward the optimal vertex from inside the constraint set.

A toy problem: Given a polygon inscribed in a circle, find the point on the polygon that maximizes the sum of its coordinates:

$$\max\{e^\top u | A^\top x = u,\ e^\top x = 1,\ x \geqslant 0\}$$

were e is vector of ones, and A has rows representing the n vertices.

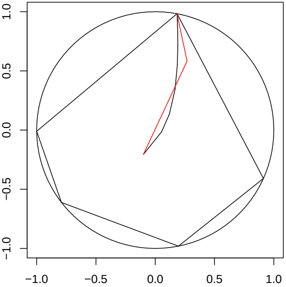
Eliminating u, setting $c = Ae$, we can reformulate the problem as:

$$\max\{c^\top x | e^\top x = 1, \quad x \geqslant 0\},$$

Toy Story: From the Inside

Simplex goes around the outside of the polygon; interior point methods tunnel from the inside, solving a sequence of problems of the form:

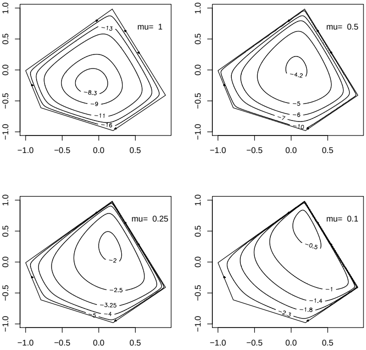
$$\max\{c^\top x + \mu \sum_{i=1}^{n} \log x_i | e^\top x = 1\}$$



Toy Story: From the Inside

By letting $\mu \to 0$ we get a sequence of smooth problems whose solutions approach the solution of the LP:

$$\max\{c^\top x + \mu \sum_{i=1}^{n} \log x_i | e^\top x = 1\}$$



Implementation: Meketon's Affine Scaling Algorithm

```
meketon <- function (x, y, eps = 1e-04, beta = 0.97) {
   f <- lm.fit(x,y)
   n <- length(y)
   w <- rep(0, n)
   d <- rep(1, n)
   its <- 0
   while(sum(abs(f$resid)) - crossprod(y, w) > eps) {
       its <- its + 1
       s <- f$resid * d
       alpha <- max(pmax(s/(1 - w), -s/(1 + w)))
       w <- w + (beta/alpha) * s
       d <- pmin(1 - w, 1 + w)^2
       f <- lm.wfit(x,y,d)
       }
   list(coef = f$coef, iterations = its)
   }
```

Mehrotra Primal-Dual Predictor-Corrector Algorithm

The algorithms implemented in `quantreg` for R are based on Mehrotra's Predictor-Corrector approach. Although somewhat more complicated than Meketon this has several advantages:

- Better numerical stability and efficiency due to better central path following,
- Easily generalized to incorporate linear inequality constraints.
- Easily generalized to exploit sparsity of the design matrix.

These features are all incorporated into various versions of the algorithm in `quantreg`, and coded in Fortran.

Back to Basics

Which is easier to compute: the median or the mean?

```
> x <- rnorm(100000000) # n = 10^8
> system.time(mean(x))
   user  system elapsed
 10.277   0.035  10.320
> system.time(kuantile(x,.5))
   user  system elapsed
  5.372   3.342   8.756
```

`kuantile` is a `quantreg` implementation of the Floyd-Rivest (1975) algorithm. For the median it requires $1.5n + O((n \log n)^{1/2})$ comparisons.

Portnoy and Koenker (1997) propose a similar strategy for “preprocessing” quantile regression problems to improve efficiency for large problems.

Globbing for Median Regression

Rather than solving $\min \sum |y_i - x_i b|$ consider:

1. Preliminary estimation using random $m = n^{2/3}$ subset,
2. Construct confidence band $x_i^\top \hat{\beta} \pm \kappa \|\hat{V}^{1/2} x_i\|$.
3. Find $J_L = \{i | y_i \text{ below band }\}$, and $J_H = \{i | y_i \text{ above band }\}$,
4. Glob observations together to form pseudo observations:

$$(x_L, y_L) = (\sum_{i \in J_L} x_i, -\infty), \quad (x_H, y_H) = (\sum_{i \in J_H} x_i, +\infty)$$

5. Solve the problem (with m+2 observations)

$$\min \sum |y_i - x_i b| + |y_L - x_L b| + |y_H - x_H b|$$

6. Verify that globbed observations have the correct predicted signs.

Proximal Algorithms for Large p Problems

Given a closed, proper convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ the proximal operator, $P_f : \mathbb{R}^n \to \mathbb{R}^n$ of f is defined as,

$$P_f(v) = \text{argmin}_x\{f(x) + \tfrac{1}{2}\|x - v\|_2^2\}.$$

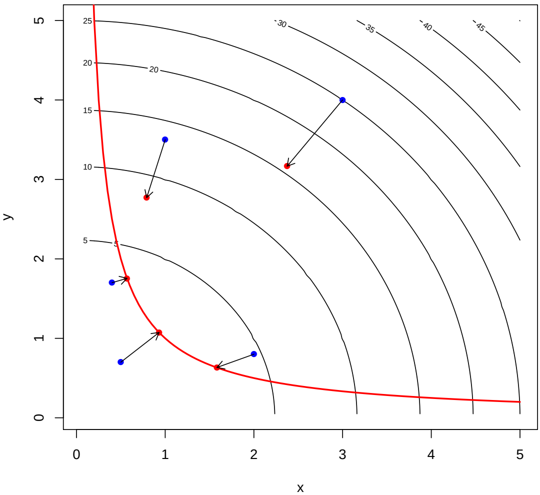
View v as an initial point and $P_f(v)$ as a half-hearted attempt to minimize f, while constrained not to venture too far away from v.
The corresponding Moreau envelope of f is

$$M_f(v) = \inf_x\{f(x) + \tfrac{1}{2}\|x - v\|_2^2\}.$$

thus evaluating M_f at $v = x$ we have,

$$M_f(x) = f(P_f(x)) + \tfrac{1}{2}\|x - P_f(x)\|_2^2\}.$$

A Toy Example:



Proximal Operators as (Regularized) Gradient Steps

Rescaling f by $\lambda \in \mathbb{R}$,

$$M_{\lambda f}(x) = f(P_{\lambda f}(x)) + \frac{1}{2\lambda}\|x - P_{\lambda f}(x)\|_2^2\}.$$

so

$$\nabla M_{\lambda f}(x) = \lambda^{-1}(x - P_{\lambda f}(x)),$$

or

$$P_{\lambda f}(x) = x - \lambda \nabla M_{\lambda f}(x).$$

So $P_{\lambda f}$ may be interpreted as a gradient step of length λ for $M_{\lambda f}$.
Unlike f, which may have a nasty subgradient, M_f has a nice gradient:

$$M_f = (f^* + \frac{1}{2}\| \cdot \|_2^2)^*$$

where $f^*(y) = \sup_x\{y^\top x - f(x)\}$ is the convex conjugate of f.

Proximal Operators and Fixed Point Iteration

The gradient step interpretation of P_f suggests the fixed point iteration:

$$x^{k+1} = P_{\lambda f}(x^k).$$

While this may not be a contraction, it is "firmly non-expansive" and therefore convergent.

In additively separable problems of the form

$$\min_x \{f(x) + g(x)\},$$

with f and g convex, this may be extended to the ADMM algorithm:

$$\begin{aligned} x^{k+1} &= P_{\lambda f}(z^k - u^k) \\ z^{k+1} &= P_{\lambda g}(x^k - u^k) \\ u^{k+1} &= (u^k + x^k - z^k) \end{aligned}$$

Alternating Direction Method of Multipliers, Parikh and Boyd (2013).

The Proximal Operator Graph Solver

A further extension that encompasses many currently relevant statistical problems is:

$$\min_{(x,y)} \{f(y) + g(x) \mid y = Ax\},$$

where (x, y) is constrained to the graph $\mathcal{G} = \{(x, y) \in \mathbb{R}^{n+m} \mid y = Ax\}$. The modified ADMM algorithm becomes:

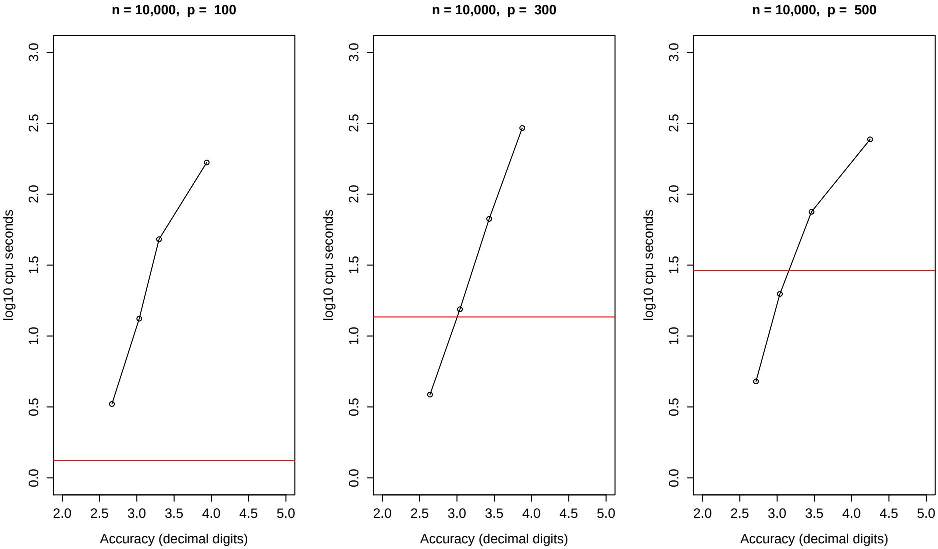
$$\begin{aligned}
(x^{k+1/2}, y^{k+1/2}) &= (P_{\lambda g}(x^k - \tilde{x}^k), P_{\lambda f}(y^k - \tilde{y}^k)) \\
(x^{k+1}, y^{k+1}) &= \Pi_A(x^{k+1/2} - \tilde{x}^k, y^{k+1/2} - \tilde{y}^k) \\
(\tilde{x}^{k+1}, \tilde{y}^{k+1}) &= (\tilde{x}^k + x^{k+1/2} - x^{k+1}, \tilde{y}^{k+1/2} + y^{k+1/2} - y^{k+1})
\end{aligned}$$

where Π_A denotes the (Euclidean) projection into graph $\mathcal{G}$. This has been elegantly implemented by Fougner and Boyd (2015) and made available by Fougner in the R package `POGS`.

When Is POGS Most Attractive?

- f and g must:
    - Be closed, proper convex
    - Be additively (block) separable
    - Have easily computable proximal operators
- A should be:
    - Not too thin
    - Not too sparse
- Other Problem Aspects
    - Available parallelizable hardware, cluster, GPUs, etc.
    - Not too stringent accuracy requirement

POGS Performance – Large p Quantile Regression



Global Quantile Regression?

Usually quantile regression is local, so solutions,

$$\hat{\beta}(\tau) = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top b)$$

are sensitive only to $\{y_i\}$ near $Q(\tau | x_i)$, the τth conditional quantile function of $Y_i | X = x_i$.

But recently there has been more interest in jointly estimating several $\beta(\tau_i)$:

$$\{\hat{\beta}(\tau) \mid \tau \in \mathcal{T}\} = \text{argmin} \sum_{\tau \in \mathcal{T}} \sum_{i=1}^{n} w_\tau \rho_\tau(y_i - x_i^\top b_\tau)$$

This is sometimes called "composite quantile regression" as in Zou and Yuan (2008). Constraints need to be imposed on the $\beta(\tau)$ otherwise the problem separates.

Example 1: Choquet Portfolios

Bassett, Koenker and Kordas (2004) proposed estimating portfolio weights $\pi \in \mathbb{R}^p$ by solving:

$$\min_{\pi \in \mathbb{R}^p,\ \xi \in \mathbb{R}^m} \{ \sum_{k=1}^{m} \sum_{i=1}^{n} w_{\tau_k} \rho_{\tau_k} (x_i^\top \pi - \xi_{\tau_k}) \mid \bar{x}^\top \pi = \mu_0 \}$$

where $x_i \in \mathbb{R}^p : i = 1, \cdots, n$ denote historical returns, and μ_0 is a required mean rate of return. This approach replaces the traditional Markowitz use of variance as a measure of risk with a lower-tail expectation measure.

- The number of assets, p, is potentially quite large in these problems.
- Linear inequality constraints can easily be added to the problem to prohibit short sales, etc.
- Interior point methods are fine, but POGS may have advantages in larger problems.

Example 2: Smoothing the Quantile Regression Process

Let $\tau_1, \cdots, \tau_m \subset (0,1)$ denote an equally spaced grid and consider

$$\min_{\beta(\tau) \in \mathbb{R}^{mp}} \{ \sum_{k=1}^{m} \sum_{i=1}^{n} w_{\tau_k} \rho_{\tau_k}(y_i - x_i^\top \beta(\tau_k)) \mid \sum_{k} (\Delta^2 \beta(\tau_k))^2 \leqslant M \}.$$

Imposes a conventional L_2 roughness penalty on the quantile regression coefficients.

- Implemented recently in POGS by Shenoy, Gorinevsky and Boyd (2015) for forecasting load in a large power grid setting,
- Smoothing, or borrowing strength from adjacent quantiles, can be expected to improve performance,
- Many gory details of implementation remain to be studied.

Conclusions and Lingering Doubts

- Optimization can replace sorting
- Simplex is just steepest descent at successive vertices
- Log barriers revive Newton method for linear inequality constraints
- Proximal algorithms revive gradient methods
- Statistical vs computational accuracy?
- Quantile models as global likelihoods?
- Multivariate, IV, extensions?