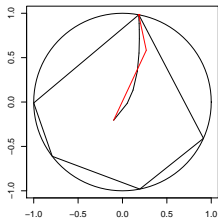


Quantile Regression Computation: From the Inside and the Outside

Roger Koenker

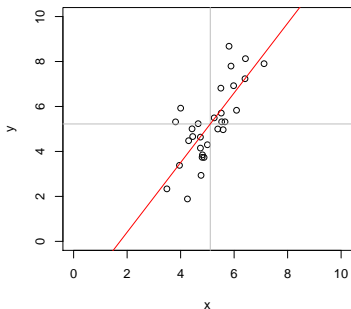
CEMMAP and University of Illinois, Urbana-Champaign

LSE: 17 May 2011



The Origin of Regression – Regression Through the Origin

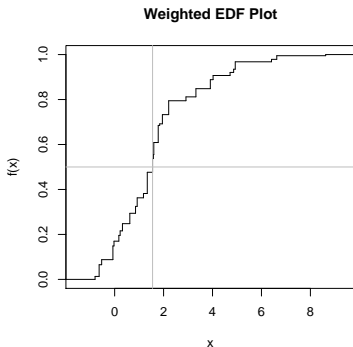
Find the line with mean residual zero that minimizes the sum of absolute residuals.



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 *Method de Situation*

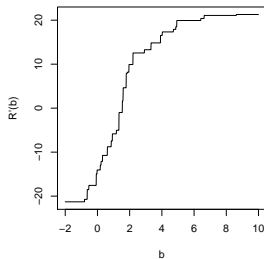
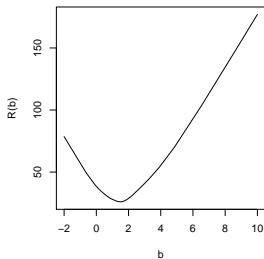
Algorithm: Order the n candidate slopes: $b_i = (y_i - \bar{y}) / (x_i - \bar{x})$ denoting them by $b_{(i)}$ with associated weights $w_{(i)}$ where $w_i = |x_i - \bar{x}|$. Find the weighted median of these slopes.



Method 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:

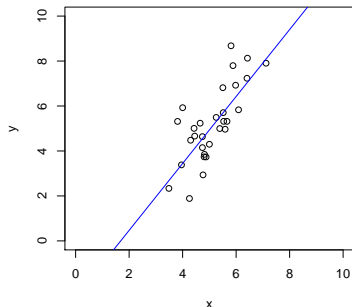
```
wquantile <- function(x, y, tau = 0.5) {  
  o <- order(y/x)  
  b <- (y/x)[o]  
  w <- abs(x[o])  
  k <- sum(cumsum(w) < ((tau - 0.5) * sum(x) + 0.5 * sum(w)))  
  list(coef = b[k + 1], k = ord[k+1])  
}
```

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?

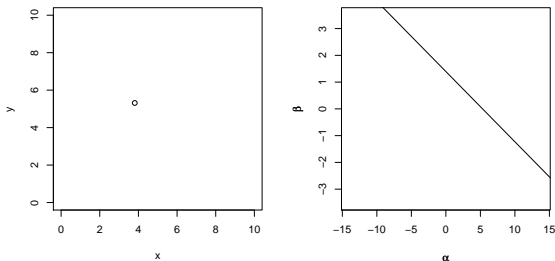
Problem: $\min_{\alpha, \beta} \sum_{i=1}^n |y_i - \alpha - x_i \beta|$



Edgeworth's (1888) Dual Plot: Anticipating Simplex

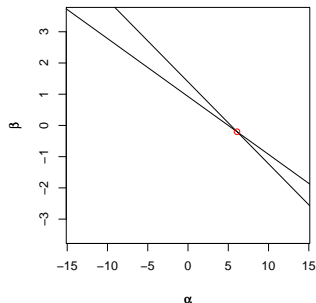
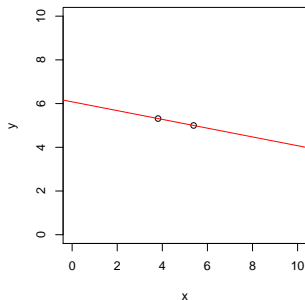
Points in sample space map to lines in parameter space.

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



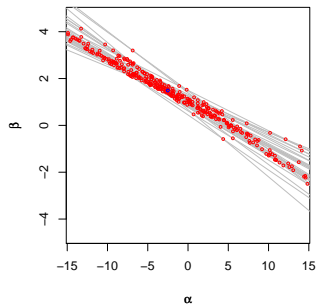
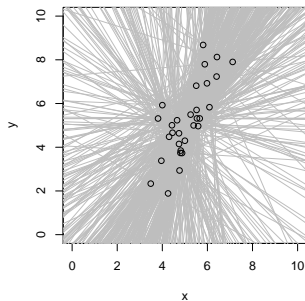
Edgeworth's (1888) Dual Plot: Anticipating Simplex

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



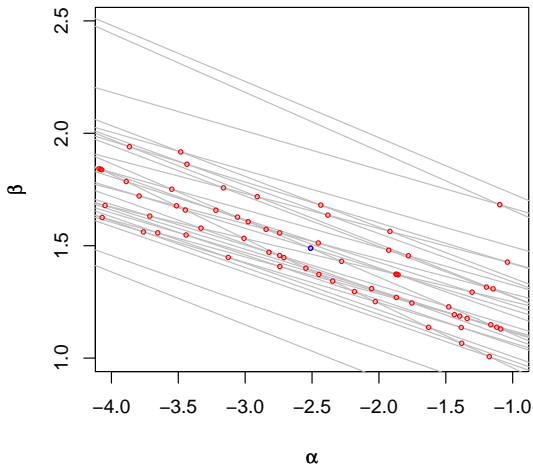
Edgeworth's (1888) Dual Plot: Anticipating Simplex

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: $\min_x \{c^\top x \mid Ax - b \in T, x \in S\}$

Dual: $\max_y \{b^\top y \mid c - A^\top y \in S^*, y \in T^*\}$

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

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

Linear Programming Duality

Primal: $\min_x \{c^\top x \mid Ax - b \in T, x \in S\}$

Dual: $\max_y \{b^\top y \mid c - A^\top y \in S^*, y \in T^*\}$

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

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

Note that for any feasible point (x, y)

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

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,

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

Quantile Regression Primal and Dual

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

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

with dual program:

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

Quantile Regression Primal and Dual

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

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

with dual program:

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

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

Quantile Regression Primal and Dual

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

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

with dual program:

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

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

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

Quantile Regression Dual

The dual problem for quantile regression may be formulated as:

$$\max_{\alpha} \{y^T \alpha \mid X^T \alpha = (1 - \tau)X^T \mathbf{1}, \alpha \in [0, 1]^n\}$$

What do these $\hat{\alpha}_i(\tau)$'s mean statistically?

They are regression rank scores (Gutenbrunner and Jurečková (1992)):

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

The integral $\int \hat{\alpha}_i(\tau) d\tau$ is something like the **rank** of the i th observation. It answers the question: On what quantile does the i th 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.

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:

$$\max\{e^T u \mid A^T x = u, e^T x = 1, x \geq 0\}$$

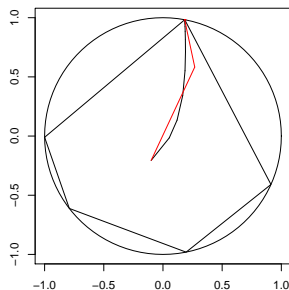
where e is vector of ones, and A has rows representing the n vertices. Eliminating u , setting $c = Ae$, we can reformulate the problem as:

$$\max\{c^T x \mid e^T x = 1, x \geq 0\},$$

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:

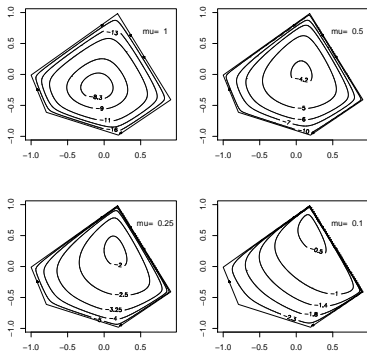
$$\max\{c^T x + \mu \sum_{i=1}^n \log x_i \mid e^T x = 1\}$$



Toy Story: From the Inside

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

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



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.

Globbering for Median Regression

Rather than solving $\min \sum |y_i - x_i \mathbf{b}|$ 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:

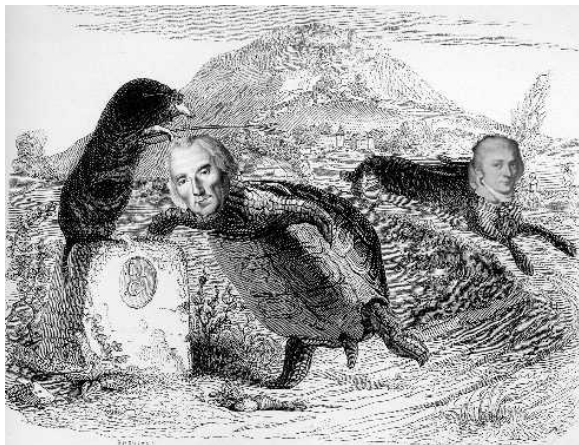
$$(x_L, y_L) = \left(\sum_{i \in J_L} x_i, -\infty \right), \quad (x_H, y_H) = \left(\sum_{i \in J_H} x_i, +\infty \right)$$

- 5 Solve the problem (with $m+2$ observations)

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

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

The Laplacian Tortoise and the Gaussian Hare



Retouched 18th century woodblock photo-print