Linear Rank Regression

(Robust Estimation of Regression Parameters)

S. Sawyer — April 25, 2003 rev April 13, 2009

1. Introduction. Consider paired data (Y_i, X_i) for a regression

$$Y_i = \mu + \beta X_i + e_i, \qquad 1 \le i \le n \tag{1.1}$$

The errors e_i in (1.1) are assumed to be independent and identically distributed, but are not necessarily normal and may be heavy-tailed.

Assume for convenience that β is one dimensional. Then (1.1) is a simple linear regression. However, most of the following extends more-or-less easily to higher-dimensional β, in which case (1.1) is a multiple regression.

Given β, define $R_i(\beta)$ as the rank (or midrank) of $Y_i - \beta X_i$ among $\{\, Y_j - \beta X_j \,\}$. Thus $1 \le R_i(\beta) \le n$. The *rank-regression estimator* $\widehat{\beta}$ is any value of β that minimizes the sum

$$D(\beta) = \sum_{i=1}^{n} R_i^c(\beta)(Y_i - \beta X_i) \tag{1.2}$$

where

$$R_i^c(\beta) = R_i(\beta) - (n+1)/2 \tag{1.3}$$

are the centered ranks or midranks.

Since $\sum_{i=1}^n R_i^c(\beta) = 0$ in (1.3), we can subtract a constant from $Y_i - \beta X_i$ in (1.2) without affecting $D(\beta)$. That is,

$$\begin{aligned} D(\beta, \mu) \;&=\; \sum_{i=1}^{n} R_i^c(\beta)(Y_i - \beta X_i - \mu) \\ &=\; \sum_{i=1}^{n} R_i^c(\beta)(Y_i - \beta X_i) \;=\; D(\beta) \end{aligned} \tag{1.4}$$

for all μ. Since

$$D(\beta) \;=\; \sum_{i=1}^{n} R_i^c(\beta)(Y_i - \beta X_i - \overline{\mu}), \qquad \overline{\mu} = \overline{Y} - \beta \overline{X}$$

and $\sum_{i=1}^n (Y_i - \beta X_i - \overline{\mu}) = 0$, and since $Y_i - \beta X_i < Y_j - \beta X_j$ implies $R_i^c(\beta) < R_j^c(\beta)$, it follows that $D(\beta) > 0$ for all β unless $Y_i - \beta X_i$ is constant.

As mentioned above, the *rank regression* slope estimator for β in (1.1) is any solution of

$$\min_{\beta} D(\beta) = D(\widehat{\beta}) \tag{1.5}$$

In particular, both $D(\beta)$ in (1.2) and $\widehat{\beta}$ in (1.5) are functions of the *residuals* $Y_i - \mu - \beta X_i$ in (1.4) and (1.1).

The classical least-squares estimators of μ and β are found by minimizing

$$C(\beta,\mu) = \sum_{i=1}^{n} \big(Y_i - \mu - \beta X_i\big)^2 \tag{1.6}$$

instead of (1.5). The least-squares estimator $\widehat{\beta}_c$ from (1.6) is

$$\widehat{\beta}_c \;=\; \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{1.7}$$

There is an algorithm for finding $\widehat{\beta}$ in (1.5) that is nearly as simple (see below).

Remarks: (1) The system (1.1)–(1.2) has a natural generation to the multiple regression

$$Y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + e_i, \qquad 1 \le i \le n, \quad p \ge 2 \tag{1.8}$$

for which $\beta = (\beta_1, \ldots, \beta_p)$ is vector valued. The analog of the classical estimator $\widehat{\beta}_c$ in (1.7) is $\widehat{\beta}_c = (X'X)^{-1}X'Y$ where X is the $n \times (p+1)$ matrix implicit on the right-hand side of (1.8).

For vector-valued β, the function $D(\beta)$ in (1.2) is piecewise linear, continuous, and convex. (See below for a proof of this in the one-dimensional case.) Thus the minimum value $\widehat{\beta}$ can be found by any routine that minimizes piece-wise linear continuous convex functions, for example for the simplex method in dynamic programming. There is a particularly easy algorithm for one dimension (see below).

(2) A natural generalization of the least-squares estimator $\widehat{\beta}_c$ in (1.7) is to minimize

$$E(\beta,\mu) = \sum_{i=1}^{n} \big|Y_i - \mu - \beta X_i\big| \tag{1.9}$$

instead of $C(\beta,\mu)$ in (1.6). The parameter estimates $\widehat{\beta}_1, \widehat{\mu}_1$ at the minimum of (1.9) do not seem to be as easy to analyze as for the rank regression model (1.2).

Theil's estimator for the slope in (1.1) is

$$\widehat{\beta}_T = \text{median}\left\{ \frac{Y_j - Y_i}{X_j - X_i} : 1 \le i < j \le n \right\} \tag{1.10}$$

(see Hollander and Wolfe, 1999, p421, in the references). If the values X_i are equally spaced, $\widehat{\beta}_T$ and the rank-regression estimator $\widehat{\beta}$ from (1.2) can be shown to be asymptotically equally powerful for estimating β (Hollander and Wolfe, 1999, p456–457). If the X_i are not equally spaced, the rank-regression estimator $\widehat{\beta}$ is asymptotically more powerful (that is, more accurate given the same sample size).

2. A Simple Algorithm for Finding $\widehat{\boldsymbol{\beta}}$ in (1.2). First, notice that the function $D(\beta)$ in (1.2) is a linear function of β except at values of β for which the ranks $R_i^c(\beta)$ change. These values correspond to pairs of integers (i,j) $(i \ne j)$ for which $Y_j - \beta X_j = Y_i - \beta X_i$, or equivalently (if $X_i \ne X_j$) if $\beta = (Y_j - Y_i)/(X_j - X_i)$ for some i and j. Let W_k be the sorted difference quotients

$$\begin{aligned} \{\, W_k : 1 \le k \le N \,\} \ = & \\ & \text{(sorted)}\,\{\, (Y_j - Y_i)/(X_j - X_i) : 1 \le i < j \le n,\ X_i \ne X_j \,\} \end{aligned} \tag{2.1}$$

For completeness, set $W_0 = -\infty$ and $W_{N+1} = \infty$. Then $D(\beta)$ is linear in each interval (W_k, W_{k+1}) $(0 \le k \le N)$. Since the midranks R_i and R_j are the same if $Y_i - \beta X_i = Y_j - \beta X_j$, it follows that $D(\beta)$ is continuous at each $\beta = W_k$, and hence is continuous (and piecewise linear) for all β.

Not consider a point of discontinuity $\beta = W_k$ in the slope of $D(\beta)$. Then there exist integers i, j such that for sufficiently small $\epsilon > 0$

$$\begin{aligned} Y_i - (W_k - \epsilon)X_i &< Y_j - (W_k - \epsilon)X_j \\ Y_i - W_k X_i &= Y_j - W_k X_j \\ Y_i - (W_k + \epsilon)X_i &> Y_j - (W_k + \epsilon)X_j \end{aligned} \tag{2.2}$$

That is, $Y_i - \beta X_i$ crosses $Y_j - \beta X_j$ from below at $\beta = W_k$. This implies that $X_i < X_j$, and also that $R_i(Y - \beta X) < R_j(Y - \beta X)$ at $\beta = W_k - \epsilon$. Thus R_i increases by one and R_j decreases by one as β crosses through $\beta = W_k$ from below. This means that the slope of $D(\beta)$ increases by $-X_i - (-X_j) = X_j - X_i > 0$.

Thus the slope of $D(\beta)$ always increases as β crosses through $\beta = W_k$ from below, and the slope of $D(\beta)$ is increasing for $-\infty < \beta < \infty$. Hence

$D(\beta)$ is convex as well as being piecewise linear and continuous. Since $D(\beta)$ is convex, continuous, and piecewise linear, $D(\beta)$ attains its minimum either at a unique node $\beta = W_k$ or else on a unique interval (W_{k-1}, W_k).

By the definition (2.1), the differences $Y_i - \beta X_i$ have the same relative order for $\beta < W_1$, which is the same relative order for $\beta \to -\infty$, which is the same order as the X_i. Similarly, $Y_i - \beta X_i$ have the opposite order of X_i if $\beta > W_N$. Thus

$$\begin{aligned} R_i(Y - \beta X) &= R_i(X), && \beta < W_1 \\ &= n + 1 - R_i(X), && \beta > W_N \end{aligned}$$

In particular by (1.2)

$$\begin{aligned} \mathrm{Slope}\big(D(\beta)\big) &= -\sum_{i=1}^{n} R_i^c(X) X_i, && \beta < W_1 \\ &= \sum_{i=1}^{n} R_i^c(X) X_i, && \beta > W_N \end{aligned}$$

Since $\sum_{i=1}^{n} R_i^c(X)\overline{X} = 0$, it follows that

$$Q = \sum_{i=1}^{n} R_i^c(X)\, X_i > 0 \tag{2.3}$$

unless the X_i are constant. We have now proven

Theorem 2.1. Let (i_k, j_k) be the integers (i, j) corresponding to k in the definition of W_k in (2.1). Define $S_0 = -Q$ for Q in (2.3) and

$$\begin{aligned} S_k &= -Q + \sum_{p=1}^{k} |X_{j_p} - X_{i_p}| \\ k_0 &= \min\{\, k : S_k > 0\} \end{aligned} \tag{2.4}$$

for $1 \le k \le N$. Then S_k is the slope of $D(\beta)$ for $W_k < \beta < W_{k+1}$. The rank-regression estimator $\widehat{\beta}$ defined by the minimum of $D(\beta)$ in (1.5) is

$$\widehat{\beta} = W_{k_0} = \frac{Y_{j_{k_0}} - Y_{i_{k_0}}}{X_{j_{k_0}} - X_{i_{k_0}}} \qquad \text{if } S_{k_0-1} < 0 < S_{k_0} \quad \text{and} \tag{2.5a}$$

$$\widehat{\beta} = \frac{W_{k_0-1} + W_{k_0}}{2} \qquad \text{if } S_{k_0-1} = 0 < S_{k_0} \tag{2.5b}$$

Remarks: (1) Theorem 2.1 gives a simple algorithm for estimating $\widehat{\beta}$. The most time-consuming part of the algorithm is sorting the difference quotients $(Y_j - Y_i)/(X_j - X_i)$ in (2.1).

(2) Since $\widehat{\beta} = W_{k_0}$ where k_0 depends on S_k, the estimator $\widehat{\beta}$ can be viewed as a "weighted median" of the difference quotients $W_k = (Y_j - Y_i)/(X_j - X_i)$ (Hollander and Wolfe, 1999).

3. A Numerical Example. Suppose that $n = 5$ and

	1	2	3	4	5
Y_i :	6.19	2.15	-2.15	11.68	3.85
X_i :	0.10	0.20	0.30	0.40	0.50

Then the ranks $R_i(X) = 1, 2, 3, 4, 5$ and the centered ranks $R_i^c(X) = R_i(X) - (n+1)/2 = -2, -1, 0, 1, 2$. Hence Q in (2.3) is $Q = 0.10(-2) + 0.2(-1) + 0.3(0) + 0.4(1) + 0.5(2) = 1.00$.

For $n = 5$, there are $N = n(n-1)/2 = 10$ difference quotients $D_k = (Y_j - Y_i)/(X_j - X_i)$. In lexicographical order (i then j), these are

$-40.38(1,2)$	$-41.70(1,3)$	$18.30(1,4)$	$-5.86(1,5)$	$-43.03(2,3)$
$47.64(2,4)$	$5.65(2,5)$	$138.30(3,4)$	$29.99(3,5)$	$-78.33(4,5)$

(The Y_i in the table were rounded to two significant figures after the decimal point.) The number W_k in (2.1) are the sorted values D_k:

$-78.33(4,5)$	$-43.03(2,3)$	$-41.70(1,3)$	$-40.38(1,2)$	$-5.86(1,5)$
$5.65(2,5)$	$18.30(1,4)$	$29.99(3,5)$	$47.64(2,4)$	$138.30(3,4)$

Then $\beta = W_k$ will be the minimum of $D(\beta)$ if $S_{k-1} < 0 < S_k$, where S_k are the numbers in (2.4). The first seven points $\beta = W_k$ along with S_k (which is the slope *just after* W_k) are:

k :	1	2	3	4	5	6	7
W_k :	-78.33	-43.03	-41.70	-40.38	-5.86	5.65	18.30
i, j :	4, 5	2, 3	1, 3	1, 2	1, 5	2, 5	1, 4
S_k :	-0.90	-0.80	-0.60	-0.50	-0.10	0.20	0.50

Note $S_5 = -0.10 < 0 < S_6 = 0.20$. Thus $D(\beta)$ is minimized at $\beta = W_6 = 5.65$ and the rank-regression estimator is $\widehat{\beta} = W_6 = 5.65$.

4. Bootstrap Confidence Intervals for $\boldsymbol{\beta}$. In general, there are two ways to bootstrap a regression in order to get confidence intervals for model parameters. Which is preferable depends on how you view the regression. The two methods often give similar results.

Bootstrapping Residual Values: If the covariates X_i are assumed to be known and fixed, you can bootstrap the residuals of the regression. To do this, carry out the following steps:

First, calculate $\widehat{\beta}$ by (2.4)–(2.5) and define "residuals"

$$r_a = Y_a - \widehat{\beta} X_a, \qquad 1 \le a \le n \tag{4.1}$$

(These are not quite the same as classical residuals, since they do not contain an estimate of the intercept parameter μ in (1.1).)

Second, for each of a large number of "bootstrap replications", define a "bootstrap resample of residuals" $\{\, r_i^* : 1 \le i \le n \,\}$ by sampling n values from the set $\{\, r_a : 1 \le a \le n \,\}$ with replacement. That is, each r_i^* is chosen so that it has probability $1/n$ of being equal to r_a for each value r_a in (4.1).

Third, define "bootstrap resampled" values Y_i^* $(1 \le i \le n)$ by

$$Y_i^* = \widehat{\beta} X_i + r_i^* \tag{4.2}$$

The variables X_i stay the same. Define W_k^* by (2.1) with Y_i^* in place of Y_i and $\widehat{\beta}^*$ by (2.5) with W_k^* in place of W_k and ${k_0}^*$ in place of k_0. While k_0 is determined only by the X_i, it also depends on the order of the difference quotients $(Y_j - Y_i)/(X_j - X_i)$.

Fourth, for some number B, collect values $\widehat{\beta}^{*j}$ for $1 \le j \le B$ by carrying out the steps in the two preceding paragraphs B times in sequence. Sort $\widehat{\beta}^{*j}$ to determine the sorted sequence $\widehat{\beta}^{*(j)}$. The classical 95% bootstrap confidence interval for β is the interval $(\widehat{\beta}^{*(0.025n)}, \widehat{\beta}^{(0.975n+1)})$. The usual rule of thumb for this confidence interval is $B \ge 1000$, so that $0.025n \ge 25$.

Some C code that carries out the first few steps above is

```
betahat = getrankbeta(nn,yy,xx);
/* Find the residuals for Y = beta X + e */
for (i=0; i<nn; i++)
  res[i] = yy[i] - betahat*xx[i];
/* For `nboot´ replicated samples */
for (ns=0; ns<nboot; ns++)
 { /* Form (yystar[i],xx[i]) (0 <= i < nn) by */
   /*   bootstrapping the residuals of yy[i] */
   for (i=0; i<nn; i++)
    { int b=nrand(nn);
      yystar[i] = betahat*xx[i] + res[b]; }
   /* Find and store the bootstrapped estimates betahat^* */
   bootbetas[ns] = getrankbeta(nn,yystar,xx); }
```

Here `nn` is the sample size, `getrankbeta()` is a function that returns the rank-regression estimator of beta, `res[]` is an array that stores the residuals of the regression $Y = \beta X + e$, `nboot` is the number of bootstrap replications of the sample, `yystar[]` is an array that holds a single bootstrapped sample of `yy` values, `nrand(nn)` is a function that returns a random integer in `0,1,2,...,nn-1`, and `bootbetas[]` is an array that holds the `nboot` rank-regression estimated values $\widehat{\beta}^{*j}$.

Bootstrapping Observations: Alternatively, if the data is viewed as random *pairs* of data (Y_i, X_i), you can bootstrap the (vector-valued) *observations* (Y_i, X_i). To do this, carry out the following steps:

First, for each of a large number of "bootstrap replications", define a "bootstrap resample of observations" $\{ (Y_i^*, X_i^*) : 1 \le i \le n \}$ by sampling n paired values from the set $\{ (Y_a, X_a) : 1 \le a \le n \}$ with replacement. That is, for each pair (Y_i^*, X_i^*), choose b with probability $1/n$ of being any of the integers $a = 1, \ldots, n$ and set $(Y_i^*, X_i^*) = (Y_b, X_b)$ (or $Y_i^* = Y_b$, $X_i^* = X_b$).

Second, define W_k^* by (2.1) with (Y_i^*, X_i^*) in place of (Y_i, X_i) and $\widehat{\beta}^*$ by (2.5) with W_k^* in place of W_k and ${k_0}^*$ in place of k_0.

Third, for some number B, collect values $\widehat{\beta}^{*j}$ for $1 \le j \le B$ by carrying out the steps in the two preceding paragraphs B times in sequence. Confidence intervals for β can be obtained from $\{ \widehat{\beta}^{*j} \}$ as in the preceding subsection.

Some C code that carries out the first few steps above is

```
betahat = getrankbeta(nn,yy,xx);
/* For `nboot´ replicated samples */
for (ns=0; ns<nboot; ns++)
 { /* Form (yystar[i],xxstar[i]) (0 <= i < nn) by */
   /*   bootstrapping PAIRS of values (yy[b],xx[b]) */
   for (i=0; i<nn; i++)
    { int b=nrand(nn);
      yystar[i] = yy[b];
      xxstar[i] = xx[b]; }
   /* Find and store the bootstrapped estimate betahat^* */
   bootbetas[ns] = getrankbeta(nn,yystar,xxstar); }
```

where `nn`, `getrankbeta()`, `nboot`, etc. are the same as before and `xxstar[]` is an array that holds the X components of the bootstrapped pairs.

Rationale of bootstrap approximations and the independence of the $\widehat{\boldsymbol{\beta}}^{*j}$: Both bootstrap methods make the implicit assumption that the $\widehat{\beta}^{*j}$ can be treated as independent. This can be justified by the fact that they are determined by independent samples (in the first case) of r_i^* drawn from the empirical distribution of the residuals r_a in (4.1) and in the second case by independent samples (Y_i^*, X_i^*) from the pairs (Y_i, X_i). In either case, the $\widehat{\beta}^{*j}$ are independent given the observed values (Y_i, X_i) $(1 \le i \le n)$ with a conditional mean $E(\widehat{\beta}^{*j})$ (conditional on the observations (Y_i, X_i)). This should be close to β if the empirical distribution of the r_i^* is close to the error distribution e_i, or else if the empirical distribution of the (Y_i^*, X_i^*) matches that of the pairs (Y, X) in the original regression model (1.1). In this conditional sense, the $\widehat{\beta}^{*j}$ can be viewed as independent estimators of β that, hopefully, have at most a small bias.

If the conditional distribution of an estimator $\widehat{\beta}$ of a parameter β given β is symmetrically distributed about β, then the middle 95% of the distribution of $\widehat{\beta}$ given β is a 95% confidence interval for β. (*Exercise:* Prove that.) Of course, this conclusion without some assumption about the relationship of the distribution of $\widehat{\beta}$ to β: If $\widehat{\beta} < \beta$ with probability one, then the entire range of the distribution of $\widehat{\beta}$ will be less than β. However, most reasonable estimators are approximately unbiased (that is, $E(\widehat{\beta}) = \beta$) and the middle 95% of the range of their distribution is a reasonable approximate 95% confidence interval for the parameter.

Sampling the middle 95% of the distribution of the $\widehat{\beta}^{*j}$ is thought to be generally reasonable if the number of bootstrap replications $B \ge 1000$, although $B = 10{,}000$ or $B = 100{,}000$ should work even better. Alternatively, if there are $B \ge 50$ replications, you can treat the values $\widehat{\beta}^{*j}$ as B independent observations with mean β and construct a classical Student-t or normal-theory 95% confidence interval for β. This often works as well as the middle 95% of the distribution of the $\widehat{\beta}^{*j}$.

References:

Hettmansperger, T. P., and J. W. McKean (1977) A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* **19**, p275–284.

Hettmansperger, T. P., and J. W. McKean (1998) *Robust Nonparametric Statistical Methods.* Arnold, London.

Hollander, M., and D. A. Wolfe (1999) *Nonparametric statistical methods*, 2nd edition. John Wiley & Sons, New York.

Jaeckel, L. A. (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.* **43**, p1449–1458.