BEBR

FACULTY WORKING PAPER NO. 90-1666

A Note on Computing Dual Regression Quantiles and Regression Rank Scores Remark on Algorithm 229

Roger Koenker
Vasco d'Orey

College of Commerce and Business Administration
Bureau of Economic and Business Research
University of Illinois Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 90-1666

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

July 1990

A Note on Computing Dual Regression Quantiles
and Regression Rank Scores Remark
on Algorithm 229

Roger Koenker
Department of Economics
University of Illinois
Champaign, IL 61820

Vasco d'Orey
Department of Economcis
Universidade Nova de Lisboa
Travessa Estavao Pinto 1000

This research was conducted while Koenker was visiting the Department of Statistics, Charles University, Prague, and the Faculty of Economics, Universidade Nova de Lisboa to whom he wishes to express his thanks for their hospitality. Support of NSF grant DMS 88-02555 is also gratefully acknowledged.

ABSTRACT

A slight modification of Algorithm 229 is described to compute the dual regression quantile statistics which are essential to the construction of the regression rank score statistics introduced by Jureckova and Gutenbrunner [1]. The latter statistics appear to generate a natural new approach to rank estimation and testing for the linear model.

1. Introduction

In [5] we described an algorithm to compute the "regression quantile" statistics introduced in [4]. The algorithm employs parametric linear programming to solve the problem

$$\min_{b \in \mathbf{R}^p} \sum_{i=1}^{n} \rho_t(y_i - x_i' b) \qquad \text{(P)}$$

where $\rho_t(u) = t\,|u|^+ + (1-t)|u|^-$. The corresponding dual problem

$$\max_{a \in A} \sum_{i=1}^{n} y_i a_i \qquad \text{(D)}$$

where $A = \{a \in [0,1]^n \mid X'a = (1-t)X'1\}$, has recently been studied by Jureckova and Gutenbrunner[1] and Jureckova, Gutenbrunner, Koenker and Portnoy[2] and appears to be of substantial independent statistical interest. Indeed the dual solution $\hat{a}_n(t)$ serves as a foundation for a new theory of linear rank statistics for the linear regression model. In this note we will briefly describe some (slight) modifications of our original algorithm required to compute $\hat{a}_n(t)$ and sketch briefly how it might be used.

Recall that a solution $\hat{b}_n(t)$ to the (primal) problem (P) is computed first at the arbitrary point $t = 1/n$, and subsequently the algorithm finds a sequence of breakpoints t_i: $i = 1,2,...,J_n$ at which the solution vector $\hat{b}_n(t)$ is modified by one simplex pivot operation. The sample path $\hat{b}_n(t)$ is piecewise constant with jumps at these breakpoints. (Portnoy[7] has recently shown that $EJ_n = O(n \log n)$ as $n \to \infty$, under mild regularity conditions on the process generating X and y.) In contrast, the components of $\hat{a}_n(t)$ are piecewise linear and continuous with *kinks* at the points where $\hat{b}_n(t)$ jumps. To describe the computation in more detail, we must recall some further notation from [4].

Let h denote the index set of "basic observations" at t such that

$$r_i(t) \equiv y_i - x_i' \hat{b}_n(t) = 0 \qquad i \in h$$

Baring degeneracies, there are exactly p elements of h at each t at which the solution is evaluated. By complementary slackness,

$$\hat{a}_n(t) = \begin{cases} 1 & \text{if } r_i(t) > 0 \\ 0 & \text{if } r_i(t) < 0 \end{cases}$$

so partitioning $\hat{a} = (\hat{a}_h, \hat{a}_{\bar{h}})'$ we have from the dual equality constraint,

$$X_h a_h = (1-t)X'1 - X_{\bar{h}} a_{\bar{h}}$$

where X has been reordered and partitioned to conform with a. Fortuituously however, $\hat{a}$ is exactly what the tableau in the primal problem computes to test the optimality of $\hat{b}_n(t)$. Recall the famous maxim of linear programming: "dual feasibility implies primal optimality." Thus to obtain $\hat{a}_n(t)$ we need only to create an appropriate array and copy the appropriate values to it from the tableau of the primal problem. Details are provided in the following section.

2. Implementation

To modify the algorithm we have added the array DSOL to the calling sequence declaring it REAL DSOL (M,NSOL) and in place of the the statement

```
330 CONTINUE
```

we have inserted the statements:

```
      KD=ABS(WA(M4,I))-N
      DSOL(KD,LSOL)=ONE+WA(M1,I)/TWO
      IF(WA(M4,I).LT.ZERO)DSOL(KD,LSOL)=ONE-DSOL(KD,LSOL)
  330 CONTINUE
      DO 335 I=KL,M
      KD=ABS(WA(I,N2))-N
      IF(WA(I,N2).LT.ZERO)DSOL(KD,LSOL)=ZERO
      IF(WA(I,N2).GE.ZERO)DSOL(KD,LSOL)=ONE
  335 CONTINUE
```

and we have replaced the group of statements

```
      DO 350 I=2,LSOL
      SOL(1,I-1)=SOL(1,I)
  350 CONTINUE
      LSOL=LSOL-1
```

with the statements

```
      DO 350 I=1,M
      DSOL(I,1)=ONE
      DSOL(I,LSOL)=ZERO
  350 CONTINUE
      SOL(1,1)=ZERO
```

One consequence of the latter substitution is to add a redundant column to the array SOL which contains the primal solution: the last $p+1$ elements of the last 2 columns are now identical and contain the vector

$$\begin{pmatrix} \bar{x}'\hat{b}_n(t) \\ \hat{b}_n(t) \end{pmatrix}$$

for $t \in [t_{J_n}, 1]$ where J_n is SOL(1,LSOL-1). The columns of the new array DSOL correspond to the evaluation of $\hat{a}_n(t)$ at $t =$ SOL(1,I), I=1,LSOL. Since SOL(1,1)=0 and SOL(1,LSOL)=1, the first and last column of DSOL are ones and zeros respectively.

3. An Example

The use of the resulting array is illustrated by the computation of the Wilcoxon scores which in the present notation would be

$$s_i = \int_0^1 \hat{a}_{ni}(t)dt = ½(\sum_{j=1}^{J_n} (\hat{a}_{ni}(t_{j+1}) + \hat{a}_{ni}(t_j))(t_{j+1} - t_j) - 1). \tag{3.1}$$

Since $\hat{a}_n(t)$ is piecewise linear the trapezoidal rule is exact. In the location model where the design matrix, X, is simply an n-vector of ones, $\hat{a}_n(t)$ takes the form

$$\hat{a}_{ni}(t) = \hat{a}_n(R_i,t) = \begin{cases} 1 & t \le (R_i-1)/n \\ R_i - tn & (R_i-1)/n < t \le R_i/n \\ 0 & R_i/n < t \end{cases}$$

where R_i is the rank of the ith observation and the function $a_n(j,t)$ is exactly as introduced by Hajek and Sidek[3, section 3.5] The connection of $\hat{a}_n(t)$ to the ranks is immmediate from the relation

$$n\int_0^1 \hat{a}_{ni}(t)dt + \frac{1}{2} = R_i, \quad i=1,...,n$$

Thus, just as solving the problem dual to (D) yields the sample quantiles or order statistics in the location model, solving (D) itself finds the ranks, with the aid of (3.1). The regression rank scores in the linear model are of course not so simple, but they preserve several important characteristics from the location case. Sample paths $\{\hat{a}_{ni}(t):0 \le t \le 1\}$ are (i) continuous, (ii) piecewise linear, (iii) $\hat{a}_{ni}(0)==1$ and $\hat{a}_{ni}(1)\equiv 0$, and (iv) $0 \le \hat{a}_{ni}(t) \le 1$. However they are not generally monotone decreasing as in the location case. To illustrate we have computed $\hat{a}_n(t)$ for the well-known stackloss data and the sample paths for each of the 21 observations appear as Figure 1. While several of these figures have the characteriestic shape we might have expected from the location model, others notably {10,12,16,19} are quite different. Observations like these which have a prolonged transit from one to zero, tend to be influential design points. Indeed Portnoy [6] has suggested using the length of the interval for which $\hat{a}_{ni}(t) \in (0,1)$ as a diagnostic for "outlyingness".

Computing the Wilcoxon regression rank scores yields the scores reported above each panel and it is clear from these scores that the observations {4,9,21} are unusual as is generally recognized in other robust analyses of this data.

REFERENCES

[1] Gutenbrunner, C. and Jureckova, J. (1990). Regression quantile and regression rank score process in the linear model and derived statistics, to appear, *Annals of Statistics.*

[2] Gutenbrunner, C., Jureckova, J., Koenker, R. and Portnoy, S. (1990). A new approach to rank tests for the linear model, in preparation.

[3] Hajek J. and Sidak, Z. (1967) *Theory of Rank Tests.* Academia, Prague.

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

[5] Koenker, R.W. and d'Orey (1987). Computing Regression Quantiles, *Applied Statistics, 36*, 383-393.

[6] Portnoy, S. (1987). Using regression quantiles to identify outliers, *Statistical Data Analysis Based on the L_1 Norm and Related Methods* (ed: Y. Dodge), North Holland, Amsterdam, 345-356.

[7] Portnoy, S. (1989). Asymptotic behavior of the number of regression quantile breakpoints, forthcoming in *J. Sci. Statist. Computing.*

Figure 1

Regression Rankscores for the Stackloss Data