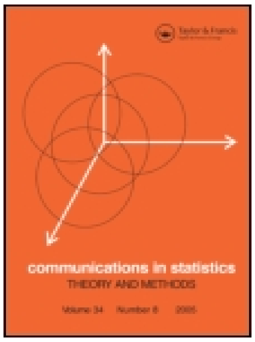


Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Multiple Case High Leverage Diagnosis in Regression Quantiles

Edmore Ranganai, Johan O. Van Vuuren & Tertius De Wet

To cite this article: Edmore Ranganai, Johan O. Van Vuuren & Tertius De Wet (2014) Multiple Case High Leverage Diagnosis in Regression Quantiles, Communications in Statistics - Theory and Methods, 43:16, 3343-3370, DOI: 10.1080/03610926.2012.715225

To link to this article: https://doi.org/10.1080/03610926.2012.715225

Communications in Statistics — Theory and Methods, 43: 3343–3370, 2014

ISSN: 0361-0926 print / 1532-415X online
DOI: 10.1080/03610926.2012.715225

Multiple Case High Leverage Diagnosis in Regression Quantiles

EDMORE RANGANAI,[1] JOHAN O. VAN VUUREN,[2]
AND TERTIUS DE WET[2]

[1]Department of Statistics, University of South Africa, Pretoria, South Africa
[2]Department of Statistics and Actuarial Science, Stellenbosch University, Stellenbosch, South Africa

Regression Quantiles (RQs) (see Koenker and Bassett, 1978) can be found as optimal solutions to a Linear Programming (LP) problem. Also, these optimal solutions correspond to specific elemental regressions (ERs). On the other hand, single case ordinary least squares (OLS) leverage statistics can be expressed as weighted averages of ER ones. Using this three-tier relationship amongst RQs, ERs, and OLS leverage statistics some relationships between single case leverage statistics and ER ones are explored and deduced. We build upon these results and propose a multiple-case RQ weighted predictive leverage statistic, T_J. We do this using an ER view of the well-known leverage relationship, $\sum_{i=1}^{n} h_i = p$, by summing the ER weighted predictive leverage statistics over all ERs (RQs included) instead of over observations, i.e., $\sum_{J=1}^{K} T_J = p$. As an ad-hoc cut-off value of this statistic we make use of the analog of the Hoaglin and Welsch (1978) one, i.e., high leverage points have $h_i > 2p/n$. So in the RQ weighted predictive leverage scenario, the cut-off value becomes $2p/K$, where K is the total number of ERs. We then apply this RQ high leverage diagnostic to well-known data sets in the literature. The cut-off value used generally seems too small. Some proposals of cut-off values based on some analytical bounds and a simulation study are therefore given and shown to be reasonable.

Keywords Regression quantile; Elemental regression; Predictive leverage.

Mathematics Subject Classification Primary 62J20; Secondary 62P99.

1. Introduction

The linear regression model is one of the most widely used tool in many applications via the ordinary least squares (OLS) procedure. An alternative and complementary procedure

Received July 19, 2011; Accepted July 19, 2012.
Address correspondence to Edmore Ranganai, Department of Statistics, University of South Africa, PO Box 392, UNISA 0003, Theo van Wijk Building, 7-125, Muckleneuk campus, Pretoria, South Africa; E-mail: rangae@unisa.ac.za
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

is quantile regression. The Koenker and Bassett (1978) Regression Quantiles (RQs) generalize the usual L_1 (50% RQ) estimator to all the quantile levels, so they are especially indispensible in applications where extremes are important. For an overview of such applications which include medical reference charts, survival analysis, finance, economics and environmental modelling see Yu et al. (2003). This emergence of quantile regression as a comprehensive approach to statistical analysis in the last three decades is partly attributed to its robustness to outliers (in the response variable) as RQs' influence functions are bounded in the response variable. However, their influence functions are unbounded in the predictor space, hence they are very susceptible to outliers in the predictor space (high leverage points). Some high-leverage points tend to create or obscure collinearity (see, e.g., Chatterjee and Hadi, 1988). Such leverage points are referred to as collinearity influential points.

Although some RQ high leverage diagnostics, constructed using robust multivariate location and scale estimates to circumvent the masking and swamping effects, have been reported in the literature (see Rousseeuw and van Zomeren, 1990; Rousseeuw and van Driessen, 1999), they have not received much attention at the multiple-case level. Multiple-case high leverage diagnostics are important since single case high leverage diagnostics fail in the presence of multiple-case effects (see, e.g., Kempthorne and Mendel, 1990). For instance, there may be situations where observations are individually influential (in the high leverage sense), but not jointly. Furthermore, joint influence is difficult to understand and detect since the computational demands may be huge (see, e.g., Barrett and Gray, 1997) and therefore practically infeasible.

In this article, we propose a multiple-case high leverage diagnostic for RQs using elemental regressions (ERs). An ER is based only on the minimum number of observations, p, to estimate the parameters of the given model (see, e.g., Hawkins, 1993).

ER procedures were introduced by Boscovich half a century before the advent of the OLS procedure (see, e.g., Stigler, 1986). However, they were out-competed by OLS due to the fact that the total number of ERs, given by

$$K = \binom{n}{p},$$

is often very large, resulting in extreme computational demands. For example, the Gunst and Mason data set (see Gunst and Mason, 1980) that we consider in this article has $n = 49$ and $p = 7$ giving $K = 85900\,584$. The K ERs consist of the set of all feasible solutions to the Linear Programming (LP) problem giving RQs as its optimal solutions (see Koenker and Bassett 1978, Theorem 3.1, p. 39). At the optimal solution the p basic observations at level zero correspond to a specific RQ. Thus, RQs can also be found through an exhaustive search over all ERs, but this procedure is not generally pursued in practice due to the availability of many efficient LP algorithms in the literature (see, e.g., the recent monograph by Koenker, 2005). Hence, there is an inherent relationship between these two sets of procedures. On the other hand, OLS leverage statistics can be expressed as weighted averages of ERs (RQs included) leverage statistics (see, e.g., Hawkins et al., 1984). These facts imply that there is an inherent relationship between leverage statistics based on ERs and those based on OLS. This three-tier relationship amongst RQs, ERs and OLS enables the single case OLS high leverage diagnostics to be extended to the ER scenario, and thus to RQs.

Although the high leverage multiple-case diagnostics developed here may be applicable to all ERs, we are only interested in those ERs that correspond to specific RQs as it is well-known that RQs are susceptible to high leverage points and their computation is practically feasible. However, we give analytical bounds and cut-off values that are applicable to a generic ER of size p. Summarizing, our motivations to exploit the above-mentioned three-tier relationship to address multiple-case high leverage diagnosis in RQs, namely, the following.

- It is well known that RQs have influence functions that are unbounded in the predictor space, and therefore have a high affinity for high leverage points unlike other ERs that do not correspond to RQs, which may not contain high leverage points and therefore unlikely to be high leverage multiple cases.
- The number (approximately, n) of unique RQs (which are easily computed using many efficient LP algorithms available in the literature) is usually much smaller than K, the total number of ERs, hence the extensive computations involving all the K subsets which may be practically infeasible (see, e.g., Hadi and Simonoff, 1993) are avoided.
- The determination of the subset size m (and hence K) is a subject of much debate in the literature since it is unknown and there are many proposals in the literature (see, e.g., Barrett and Gray,1997; Seaver et al., 1999). Using the ER procedure, taking $m = p$ ties in with the RQ setting, thus the dilemma of choosing the subset size is avoided.
- Subset diagnostics are not comparable across different subset sizes (see e.g., Barrett and Gray, 1997). So a different subset size $m \neq p$ is not useful for comparison purposes, hence we fix the subset size at $m = p$.

Using existing OLS-ER relationships in the next section we derive a RQ high leverage diagnostic analogy of the single case OLS one. The elemental regression weight (ERW) is the vehicle through which OLS leverage statistics are related to the ER leverage statistics. We briefly elaborate on the ERW. In Theorem 2.1 we give two expressions for the ERW and their proofs. To motivate the proposed RQ high leverage diagnostic we give Theorems 2.1 and 2.2 and their proofs. For an ad-hoc cut-off value of this statistic we make use of the analogue of the Hoaglin and Welsch (1978) one, i.e., high leverage points have $h_i > 2p\big/n$. Applications to well-known data sets show that this cut-off value is generally too small. Therefore some analytical bounds and reasonable cut-off values based on these bounds and a simulation study are given in Secs. 3 and 4, respectively. Applications are given in Sec. 5. We consider data sets with high leverage points that both create and hide collinearity as well as a control data set with very moderately high leverage points. Using as cut-off value the analogue of the single case one, it was found to be generally too small. However, cut-off values proposed in Sec. 3 and in a simulation study are found to be reasonable. Conclusions and further research are given in the last section.

2. High Leverage Diagnostics

Some important relationships between single case high leverage diagnostics and multiple-case (RQ) high leverage diagnostics using the three-tier relationship amongst RQs, ERs and OLS are explored and deduced leading to the development of a high leverage diagnostic for RQs (T_J in (2.6) below). We do this by deducing the relationship between the OLS leverage (h_i in (2.1)) and the ER predictive leverage ($h_{iJ},\ \ i \notin J$ in (2.3)) using the existing relationship between the jackknife OLS leverage statistic ($h_{(i)}$ in (2.2)) and (2.3); see Hawkins et al. (1984) for these results. In these relationships the so-called elemental regression weight (ERW, see (2.4) below) plays a pivotal role. As a precursor to developing a high leverage diagnostic for RQs, we first briefly elaborate on the existing leverage statistics h_i, $h_{(i)}$, and $h_{iJ},\ \ i \notin J$and their relationships. This enables us to define the relevant diagnostics in a natural way.

Let X denote the predictor matrix without the constant term and write $\tilde{\mathbf{X}} = \begin{pmatrix}\mathbf{1}_n & \mathbf{X}\end{pmatrix}$. Without loss of generality we assume that X is centered. The projection (hat) matrix $\mathbf{H} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'$, and its variants play a very important role in high leverage diagnostics, as we will now briefly discuss. Leverage values are given by the diagonal elements of $\mathbf{H}$ denoted by

$$h_i = \tilde{\mathbf{x}}_i' \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{x}}_i. \tag{2.1}$$

High leverage points have $h_i > 2p/n$ (see Hoaglin and Welsch, 1978).

The diagonal elements of one such variant of $\mathbf{H}$, $\mathbf{H}_{(i)} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}/_{(i)}\tilde{\mathbf{X}}_{(i)})^{-1}\tilde{\mathbf{X}}'$ give the usual jackknife leverage statistics

$$h_{(i)} = \tilde{\mathbf{x}}_i' \left(\tilde{\mathbf{X}}_{(i)}'\tilde{\mathbf{X}}_{(i)}\right)^{-1} \tilde{\mathbf{x}}_i, \tag{2.2}$$

where the subscript (i) denotes calculations with the i^{th} observation left out.

Let J denote a generic ER containing p observations and let I be its complement containing the remaining $n - p$ observations.

The RQ (ER) analogues of the above jackknife leverage statistics, are the diagonal elements of the analogue of matrix $\mathbf{H}_{(i)}$, $\mathbf{H}_{(I)} \equiv \mathbf{H}_J = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J)^{-1}\tilde{\mathbf{X}}'$, where subscript (I) denotes calculations with set I left out. These elements are given by

$$h_{iJ} = \begin{cases} \tilde{\mathbf{x}}_i' \left(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J\right)^{-1} \tilde{\mathbf{x}}_i, & i \notin J \ (i \in I) \\ 1 & i \in J, \end{cases}, \tag{2.3}$$

for non singular $\tilde{\mathbf{X}}_J$, which are the diagonal elements of the matrix $\mathbf{H}_{IJ} = \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J)^{-1}\tilde{\mathbf{X}}_I'$.

Remark 2.1. Hawkins et al. (1984) refer to h_{iJ} as the *residual freedom*, to "convey the impression of its property of measuring the extent to which the elemental set J fails to predict Y_i, $i \in I$".

For a given ER J, the corresponding elemental regression weight (ERW) is given by

$$\omega_J = \frac{|\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J|}{|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}|}. \tag{2.4}$$

Note that $0 \leq \omega_J \leq 1$, $\sum_J \omega_J = 1$ and by the Cauchy-Binet Theorem (Aitken, 1964, p. 86), it follows that

$$\omega_J = \frac{|\tilde{\mathbf{X}}_J|^2}{\sum_J |\tilde{\mathbf{X}}_J|^2}.$$

A tighter upper bound for the ERW is derived in Sec. 3. These weights play a pivotal role in the construction of least squares estimators. Jacobi, in 1841 showed that the least squares estimators can be expressed as weighted averages of the elemental regressions (see Sheynin, 1973; Mayo and Gray, 1997).

We now prove Theorem 2.1 below, which shows that the ERW can be expressed as the determinant of the elemental regression matrix (ERM), $(\mathbf{I}_{n-p} + \mathbf{H}_{IJ})^{-1} \equiv \mathbf{I}_{n-p} - \mathbf{H}_I$, where $\mathbf{H}_I \equiv \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I'$ is the sub-matrix of $\mathbf{H}$.

Theorem 2.1. *For $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$ non singular, we have*

$$\text{(i)} \quad \omega_J = |\mathbf{I}_{n-p} + \mathbf{H}_{IJ}|^{-1}$$

and

$$\text{(ii)} \quad \omega_J = |\mathbf{I}_{n-p} - \mathbf{H}_I|.$$

Proof of (i). We have $|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}|\& = |\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J + \tilde{\mathbf{X}}_I'\mathbf{I}_{n-p}\tilde{\mathbf{X}}_I|$, from the well known result (see, e.g., Harville, 1997) $|\mathbf{R} + \mathbf{STU}| = |\mathbf{R}||\mathbf{T}||\mathbf{T}^{-1} + \mathbf{UR}^{-1}\mathbf{S}|$, where $\mathbf{R}$ and $\mathbf{T}$ are non singular, follows that

$$\begin{aligned} |\tilde{\mathbf{X}}'\tilde{\mathbf{X}}| &= |\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J||\mathbf{I}_{n-p}||\mathbf{I}_{n-p} + \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J)^{-1}\tilde{\mathbf{X}}_I'| \\ &= |\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J||\mathbf{I}_{n-p} + \mathbf{H}_{IJ}|. \end{aligned}$$

Finally, substituting the above last expression into (2.4) follows (i). □

Proof of (ii). Let $\mathbf{Z} = \left(\begin{smallmatrix} \mathbf{I}_{n-p} & \tilde{\mathbf{X}}_I \\ \tilde{\mathbf{X}}_I' & \tilde{\mathbf{X}}'\tilde{\mathbf{X}} \end{smallmatrix}\right)$. Now, for any $k \times k$ non singular matrix $\mathbf{A}_{11}$, $k \times m$ matrix $\mathbf{A}_{12}$, $m \times k$ matrix $\mathbf{A}_{21}$, and $m \times m$ nonsingular matrix $\mathbf{A}_{22}$, it is well known that the determinant of the partitioned matrix is given by

$$\begin{aligned} \begin{vmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{vmatrix} &= |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}| \\ &= |\mathbf{A}_{11}|\left|\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right|. \end{aligned}$$

Applying the first form to $\mathbf{Z}$, gives

$$\begin{aligned} |\mathbf{Z}| &= |\tilde{\mathbf{X}}'\tilde{\mathbf{X}}||\mathbf{I}_{n-p} - \tilde{\mathbf{X}}_I\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}_I'| \\ &= |\tilde{\mathbf{X}}'\tilde{\mathbf{X}}||\mathbf{I}_{n-p} - \mathbf{H}_I|, \end{aligned}$$

and using the second form, gives

$$\begin{aligned} |\mathbf{Z}| &= |\tilde{\mathbf{X}}'\tilde{\mathbf{X}} - \tilde{\mathbf{X}}_I'\tilde{\mathbf{X}}_I| \\ &= |\tilde{\mathbf{X}}_{(I)}'\tilde{\mathbf{X}}_{(I)}| \\ &= |\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J|. \end{aligned}$$

From these two expressions and (2.4), (ii) follows. This completes the proof of the theorem. □

Remark 2.2. Making use of the above results it will be shown that the RQ weighted predictive leverage can be written in matrix form (see (2.7) and (2.8)).

We now prove Theorem 2.2 below which gives alternative expressions of leverage statistics. These expressions and some intermediate results (in the proofs) are central to deducing a RQ weighted predictive leverage statistic.

Theorem 2.2. *For $\tilde{\mathbf{X}}_J$ non singular,*

(i)

$$h_{(i)} = \frac{\sum_{J \not\ni i} \omega_J h_{iJ}}{(n-p)\sum_{J \not\ni i} \omega_J} = \frac{\tilde{R}_i}{(n-p)}.$$

(ii)

$$h_i = \frac{\tilde{R}_i}{(n-p)+\tilde{R}_i}.$$

(iii)

$$h_i = \frac{\sum_{J \not\ni i} \omega_J h_{iJ}}{(n-p)},$$

where $\tilde{R}_i = \frac{\sum_{J \not\ni i} \omega_J h_{iJ}}{\sum_{J \not\ni i} \omega_J}$.

Proof of (i). For any $\mathbf{d} \in \Re^p$, $|\mathbf{I}_p + \mathbf{dd}'| = 1+\mathbf{d}'\mathbf{d}$. Similarly for any $(p \times p)$ symmetric and nonsingular matrix $\mathbf{A}$, $\mathbf{d} \in \Re^p$, $|\mathbf{A} + \mathbf{dd}'| = |\mathbf{A}|\,(1 + \mathbf{d}'\mathbf{A}^{-1}\mathbf{d})$ (see Graybill, 1983, p. 231; Schott, 2005, p. 92). Hence, □

$$\begin{aligned} 1 + h_{(i)} &= 1 + \tilde{\mathbf{x}}_i' \left(\tilde{\mathbf{X}}_{(i)}'\tilde{\mathbf{X}}_{(i)}\right)^{-1} \mathbf{x}_i \\ &= \frac{|\left(\tilde{\mathbf{X}}_{(i)}'\tilde{\mathbf{X}}_{(i)}\right) + \tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i|}{|\left(\tilde{\mathbf{X}}_{(i)}'\tilde{\mathbf{X}}_{(i)}\right)|} = \frac{|\tilde{\mathbf{X}}'\tilde{\mathbf{X}}|}{|\left(\tilde{\mathbf{X}}_{(i)}'\tilde{\mathbf{X}}_{(i)}\right)|}. \end{aligned}$$

Thus, invoking the Cauchy-Binet Theorem, we have

$$\begin{aligned} 1 + h_{(i)} &= \frac{\sum_J |\tilde{\mathbf{X}}_J|^2}{\sum_{J \not\ni i} |\tilde{\mathbf{X}}_J|^2} \\ h_{(i)} &= \frac{\sum_{Ji} |\tilde{\mathbf{X}}_J|^2}{\sum_{J \not\ni i} |\tilde{\mathbf{X}}_J|^2}. \end{aligned} \tag{2.2.1}$$

Analogously for the ER J, we have

$$\begin{aligned} 1 + h_{iJ} &= 1 + \tilde{\mathbf{x}}_i' \left(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J\right)^{-1} \tilde{\mathbf{x}}_i \\ &= \frac{|\left(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J\right) + \tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i|}{|\left(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J\right)|}. \end{aligned}$$

Similarly, again invoking the Cauchy-Binet Theorem to the numerator, we have

$$1 + h_{iJ} = \frac{\sum_{H\subset\{J,i\}} |\tilde{\mathbf{X}}_H|^2}{|\tilde{\mathbf{X}}_J|^2},$$

where H is an ER (a subset of size p) from the observations in the set $\{J, i\}$.
Thus,

$$h_{iJ} = \frac{\sum_{H\subset\{J,i\}, H i} |\tilde{\mathbf{X}}_H|^2}{|\tilde{\mathbf{X}}_J|^2}.$$

Multiplying both sides by $|\tilde{\mathbf{X}}_J|^2$ and summing over $J \not\ni i$,

$$\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2 h_{iJ} = \sum_{J\not\ni i} \sum_{H\subset\{J,i\}, H i} |\tilde{\mathbf{X}}_H|^2.$$

Now, for any given set H, set J, which does not include i, must contain the other $p-1$ elements of H. Thus, J will have only one free element, which will run over the $n-p$ values outside H. Consequently, any term $|\tilde{\mathbf{X}}_H|^2$ will appear $n-p$ times, i.e.,

$$\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2 h_{iJ} = (n-p) \sum_{J i} |\tilde{\mathbf{X}}_J|^2. \tag{2.2.2}$$

(For this result, see also Hawkins et al., 1984.)
Dividing both sides by $(n-p)\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2$, the above equation becomes

$$\frac{\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2 h_{iJ}}{(n-p)\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2} = \frac{\sum_{J i} |\tilde{\mathbf{X}}_J|^2}{\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2}.$$

Thus, using (2.2.1)

$$h_{(i)} = \frac{\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2 h_{iJ}}{(n-p)\sum_{J\not\ni i} |\tilde{\mathbf{X}}_J|^2}.$$

Dividing the numerator and denominator by $\sum_J |\tilde{\mathbf{X}}_J|^2$, the relation becomes

$$h_{(i)} = \frac{\sum_{J\not\ni i} \omega_J h_{iJ}}{(n-p)\sum_{J\not\ni i} \omega_J}. \tag{2.2.3}$$

The first result of (i) is proved.
Dividing the numerator and denominator by $\sum_{J\not\ni i} \omega_J$, the relation becomes

$$h_{(i)} = \frac{\tilde{R}_i}{n-p}, \tag{2.2.4}$$

where $\tilde{R}_i = \frac{\sum_{J\not\ni i} \omega_J h_{iJ}}{\sum_{J\not\ni i} \omega_J}$, proving the second result of (i).

Proof of (ii). Using $h_i = h_{(i)}/(1 + h_{(i)})$ (see Hawkins et al., 1984), and substituting $h_{(i)}$ from (2.2.4), (ii) follows immediately. □

Proof of (iii). Using $h_i = h_{(i)}/(1 + h_{(i)})$ again with (2.2.3) gives

$$h_i = \frac{\sum_{J \not\ni i} \omega_J h_{iJ}}{(n-p)\sum_{J \not\ni i} \omega_J + \sum_{J \not\ni i} \omega_J h_{iJ}} \tag{2.2.5}$$

Dividing (2.2.2) by $\sum_J |\tilde{\mathbf{X}}_J|^2$, the relation becomes

$$\sum_{J \not\ni i} \omega_J h_{iJ} = (n-p)\sum_{Ji} \omega_J.$$

Substituting $(n-p)\sum_{Ji} \omega_J$ for $\sum_{J \not\ni i} \omega_J h_{iJ}$ the denominator of (2.2.5) becomes

$$(n-p)\left(\sum_{J \not\ni i} \omega_J + \sum_{Ji} \omega_J\right) \text{ but } \sum_{J \not\ni i} \omega_J + \sum_{Ji} \omega_J = 1.$$

Thus,

$$h_i = \frac{\sum_{J \not\ni i} \omega_J h_{iJ}}{(n-p)}, \tag{2.5}$$

proving (iii). □

Remark 2.3. In (i) we rewrote the crude expression for $h_{(i)}$ given by Hawkins et al. (1984) in terms of the ERW and $\tilde{R}_i$. In (ii) and (iii) we extended the same ideas to h_i. However, for the purpose of determining multiple leverage points, it is more convenient to use (iii).

Now summing the individual quantities $\omega_J h_{iJ}\big/(n-p)$, over $J \not\ni i$ gives the usual ordinary least squares (OLS) leverage diagnostic h_i which is not useful in the RQ scenario. However, an alternative way of summing the individual quantities $\omega_J h_{iJ}/(n-p)$, $J \not\ni i$ in this expression is over i (the observation not contained in ER J), suggesting as a RQ weighted predictive leverage (RQWPL) statistic, the quantity

$$T_J = \frac{\sum_{i \notin J} \omega_J h_{iJ}}{n-p}. \tag{2.6}$$

This statistic is very useful in the RQ scenario as one can only sum over the quantities corresponding to a particular ER and thus RQ as schematically depicted in Table 1 below. In this, use is made of Theorem 2.2 (iii) and the well-known fact that $\sum_i h_i = p$, implying that

$$\sum_{J=1}^{K} T_J = \sum_{i=1}^{n} h_i = p.$$

Summarizing, there are therefore two ways of summing the quantities $\omega_J h_{iJ}/(n-p)$ in Table 1.

- h_i, OLS leverage of observation i is obtained as the sum of the quantities $\omega_J h_{iJ}/(n-p)$ over the ERs not containing observation i. These quantities can then be viewed as the contribution of ER J to the OLS leverage of observation i.
- In the same way, T_J is obtained as the sum of the quantities $\omega_J h_{iJ}/(n-p)$ over the observations not contained in ER J. These quantities can then be viewed as the contribution of observation i (not in J) to the weighted predictive leverage of ER J.

Table 1
Relationship between ER weighted predictive leverage, T_J, and OLS Leverage, h_i

	ELEMENTAL SETS						
Obs	1	2	.	.	.	K	Leverage
1	$\frac{\omega_1 h_{11}}{n-p}$	$\frac{\omega_2 h_{12}}{n-p}$	.	.	.	$\frac{\omega_K h_{1K}}{n-p}$	h_1
2	$\frac{\omega_1 h_{21}}{n-p}$	$\frac{\omega_2 h_{22}}{n-p}$	.	.	.	$\frac{\omega_K h_{2K}}{n-p}$	h_2
3	$\vdots$	$\vdots$	.	.	.	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	.	.	.	$\vdots$	$\vdots$
n	$\frac{\omega_1 h_{n1}}{n-p}$	$\frac{\omega_2 h_{n2}}{n-p}$	.	.	.	$\frac{\omega_K h_{nK}}{n-p}$	h_n
Sum	T_1	T_2	.	.	.	T_K	p

From (2.6) and Theorem 2.1, the RQWPL statistic, T_J can be expressed in matrix notation as follows:

$$T_J = \frac{trace(\mathbf{H}_{IJ})}{(n-p)|\mathbf{I}_{n-p} + \mathbf{H}_{IJ}|} \tag{2.7}$$

$$= \frac{|\mathbf{I}_{n-p} - \mathbf{H}_I|trace(\mathbf{H}_{IJ})}{n-p}, \tag{2.8}$$

where $|\mathbf{I}_{n-p} - \mathbf{H}_I| = \omega_J$ the ERW and $trace(\mathbf{H}_{I\,J}) = \sum_{i \notin J} h_{iJ}$ the sum of the RQ predictive leverage.Since

$$\sum_{i=1}^{n} h_i = \sum_{i=1}^{K} T_J = p,$$

an intuitively appealing idea is to extend the single case Hoaglin and Welsch (1978) procedure of identifying high leverage points by $h_i > 2p/n$ to the multiple-case by flagging ERs (RQs) having $T_J > 2pK^{-1}$. In the applications in Sec. 3, we take this as our ad-hoc cut-off value. Also, due to the small values of T_J, it is convenient to use the rule, flag $-\ln(T_J) < -\ln(2pK^{-1})$. We will adopt this approach in Sec. 5 on applications. However, in these applications it is clearly evident that the cut-off value is generally small.

For comparison with existing multiple-case high leverage diagnostics in the literature, the measures of choice would be $||\mathbf{H}_I|| \equiv ||\tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'_I||$ and $||\mathbf{H}_J|| \equiv ||\tilde{\mathbf{X}}_J(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'_J||$, the Frobenius matrix norms of the sub-matrices of $\mathbf{H}$, since they are not measure-specific and easy to compute (see Barrett and Gray, 1997). The measure $||\mathbf{H}_I||$ quantifies the leverage for observations outside subset Jcorresponding to a RQ, i.e., observations $i \in I$. Since RQs have a high affinity for leverage points, we are more interested interested in the measure $||\mathbf{H}_J||$, quantifying the leverage for observations in subset J corresponding to a RQ. Simulation results show that $||\mathbf{H}_I||$ is negatively correlated with $||\mathbf{H}_J||$ and T_J while $||\mathbf{H}_J||$ and T_J are positively correlated. Therefore, we will use $||\mathbf{H}_J||$ for comparison purposes.

Since the Hoaglin and Welsch (1978) multiple-case analog cut-off values of single case ones for T_J, are practically too small and the computation of all ERs is practically infeasible, some analytical bounds could be useful, if the interest is in all ERs. Therefore in the next section we derive some bounds for T_J.

3. Some Bounds and Cut-off Values

The RQWPL leverage statistic T_J is the scaled product of the ERW and $trace(\mathbf{H}_{IJ})$. We first find the bounds of these two statistics which are the building blocks of RQWPL, separately starting with the ERW. The ERW is the determinant of the elemental regression matrix (ERM), and the ERM is a sub-matrix of the projection matrix, $\mathbf{I} - \mathbf{H}$. It is well known that $\mathbf{I}-\mathbf{H}$ is idempotent and hence, its sub-matrix $\mathbf{I}_{n-p} - \mathbf{H}_I$ is positive semidefinite definite with eigenvalues contained in [0, 1). Positive semidefinite definite matrices are rich in determinantal relationships. Using these relationships, it can be shown that

$$1 - trace(\mathbf{I}_{n-p} - \mathbf{H}_I) \le |\mathbf{I}_{n-p} - \mathbf{H}_I| = \omega_J \le 1 - trace(\mathbf{H}_I) + \frac{m-1}{2m}[trace(\mathbf{H}_I)]^2 \tag{3.1}$$

(see, Barrett and Gray, 1996).
An alternative upper bound for the ERW to that in (3.1) is given by the following lemma.

Lemma 3.1. *For any elemental set J from a full rank* $\tilde{\mathbf{X}}$, $0 \le \omega_J \le \frac{p}{n}$.

Proof. For the lower bound, since $\mathbf{I}_{n-p} - \mathbf{H}_I$ is positive semidefinite, $0 \le |\mathbf{I}_{n-p} - \mathbf{H}_I| = \omega_J$.
Now, we consider the upper bound.
Let, as before, $\tilde{\mathbf{X}} = (\mathbf{1} \quad \mathbf{X})$, and denote by $\mathbf{H}$ the projection matrix of $\tilde{\mathbf{X}}$, i.e.,

$$\begin{aligned}\mathbf{H} &= \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}' \\ &= \mathbf{H}_0 + \mathbf{H}_1,\end{aligned}$$

where

$$\mathbf{H}_0 = \frac{1}{n}\mathbf{J}_n,$$

with $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}'_n$ and

$$\mathbf{H}_1 = (\mathbf{I}_n - \mathbf{H}_0)\mathbf{X}(\mathbf{X}'(\mathbf{I}_n - \mathbf{H}_0)\mathbf{X}_I)^{-1}\mathbf{X}'(\mathbf{I}_n - \mathbf{H}_0)$$

(see Chatterjee and Hadi, 1988, Property 2.4, p. 16).
Consider the following equality:

$$(\mathbf{I}_n - \mathbf{H}_0) = (\mathbf{I}_n - \mathbf{H}) + (\mathbf{H} - \mathbf{H}_0).$$

Clearly, all three matrices (each one in brackets) are non negative definite (since they are projection matrices).
Now, taking the corresponding $(n-p) \times (n-p)$ sub-matrices indexed by I from the above three matrices, we can write

$$(\mathbf{I}_n - \mathbf{H}_0)_I = (\mathbf{I}_n - \mathbf{H})_I + (\mathbf{H} - \mathbf{H}_0)_I.$$

Again, all three matrices are non-negative definite since they are sub-matrices of projection matrices. Using Corollary 18.1.8 in Harville (1997), p. 418, we have

$$|(\mathbf{I}_n - \mathbf{H}_0)_I| \geq |(\mathbf{I}_n - \mathbf{H})_I|.$$

Clearly,

$$(\mathbf{I}_n - \mathbf{H}_0)_I = \mathbf{I}_{n-p} - \tfrac{1}{n}\mathbf{J}_{n-p}.$$

Also, since

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{\mathbf{X}}_J \\ \tilde{\mathbf{X}}_I \end{pmatrix},$$

we have

$$\begin{aligned} \mathbf{H} &= \begin{pmatrix} \tilde{\mathbf{X}}_J \\ \tilde{\mathbf{X}}_I \end{pmatrix} (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \begin{pmatrix} \tilde{\mathbf{X}}'_J & \tilde{\mathbf{X}}'_I \end{pmatrix} \\ &= \begin{pmatrix} \tilde{\mathbf{X}}_J(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'_J & \tilde{\mathbf{X}}_J(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'_I \\ \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'_J & \tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'_I \end{pmatrix} \end{aligned}$$

From which

$$(\mathbf{H} - \mathbf{H}_0)_I = \mathbf{H}_I - \mathbf{H}_{0I}$$

and

$$(\mathbf{I}_n - \mathbf{H})_I = \mathbf{I}_{n-p} - \mathbf{H}_I,$$

with

$$\mathbf{H}_{0I} = \frac{1}{n}\mathbf{J}_{n-p}.$$

The above inequality can then be written as

$$\left|\mathbf{I}_{n-p} - \frac{1}{n}\mathbf{J}_{n-p}\right| \geq |\mathbf{I}_{n-p} - \mathbf{H}_I| = \omega_J$$

(see Theorem 2.1).
Thus,

$$\omega_J \leq \left|\mathbf{I}_{n-p} - \frac{1}{n}\mathbf{J}_{n-p}\right|.$$

The proof is completed by calculation of the determinant on the right-hand side.

Now,

$$\left|\mathbf{I}_{n-p} - \frac{1}{n}\mathbf{J}_{n-p}\right| = \begin{vmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{vmatrix}$$
$$= 1 - \frac{(n-p)}{n}$$
$$= \frac{p}{n}$$

(see Searle, 1982, p. 322).
This completes the proof of the lemma. □

Comparing the upper bounds in (3.1) and the one given by Lemma 3.1, we have

$$1 - trace(\mathbf{H}_I) + \frac{m-1}{2m}[trace(\mathbf{H}_I)]^2 - \frac{p}{n}$$
$$= \frac{m-1}{2m}\left\{\left(trace(\mathbf{H}_I) - \frac{m}{m-1}\right)^2\right\} - \frac{n(m-2) - 2p(m-1)}{2n(m-1)}.$$
$$\geq 0, \text{ for } m > 2 \text{ and } n \geq 2p[(m-1)/(m-2)],$$

which is usually the case in practice. Therefore, we make use of the tighter upper bound given by Lemma 3.1, for $m > 2$ and $n \geq 2p[(m-1)/(m-2)]$.

The bounds for $trace(\mathbf{H}_{IJ})$ follow from the fact that $\mathbf{H}_{IJ}$ is positive semidefinite and so can be written as a product of two positive semidefinite matrices as given by Lemma 3.2 below.

Lemma 3.2. *For any elemental set J from a full rank* $\tilde{\mathbf{X}}$, $\mathbf{H}_{IJ} = \mathbf{H}_I(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}$, *for non singular (positive definite)* $\mathbf{I}_{n-p} - \mathbf{H}_I$.

Proof. This result can be realized from evaluating $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} - \tilde{\mathbf{X}}_I'\tilde{\mathbf{X}}_I)^{-1}$ (see Chatterjee and Hadi, 1988, p. 192), i.e.,

$$(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J)^{-1} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}} - \tilde{\mathbf{X}}_I'\tilde{\mathbf{X}}_I)^{-1}$$
$$= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} + (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I'(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}\tilde{\mathbf{X}}_I(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}.$$

Post-multiplying by $\tilde{\mathbf{X}}_I'$, we have

$$(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J)^{-1}\tilde{\mathbf{X}}_I' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I'\{\mathbf{I}_{n-p} + (\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}\mathbf{H}_I\}.$$

Adding and subtracting $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I'(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}$ yields

$$(\tilde{\mathbf{X}}_J'\tilde{\mathbf{X}}_J)^{-1}\tilde{\mathbf{X}}_I' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_I'(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}\mathbf{H}_I,$$

which on pre-multiplying by $\mathbf{X}_I$ the proof is completed. □

In terms of the Frobenius inner product and using Lemma 3.2, we have

$$\begin{aligned}\left\langle \mathbf{H}_I, (\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1} \right\rangle_F &= trace(\mathbf{H}_I(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}) \\ &= trace(\mathbf{H}_{IJ}),\end{aligned}$$

and from the Cauchy-Schwartz inequality we have

$$\left\langle \mathbf{H}_I, (\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1} \right\rangle = trace(\mathbf{H}_{IJ}) \leq \|\mathbf{H}_I\| \left\|(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}\right\| \tag{3.3}$$

which can also be deduced from the well-known relationship

$$\cos[\mathbf{H}_I, (\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}] = \frac{trace(\mathbf{H}_{IJ})}{\|\mathbf{H}_I\| \left\|(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}\right\|},$$

where $||\cdot||$ denotes the Frobenius matrix norm.
Using the result from Yang et al. (2001) for the product of the trace of positive semidefinite matrices and Lemma (3.2) we have

$$trace(\mathbf{H}_{IJ}) \leq trace(\mathbf{H}_I)trace[(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}]. \tag{3.4}$$

Also,

$$\begin{aligned}&\sum_{i=1}^{n-p} \lambda_i(\mathbf{H}_I)\lambda_{n-i+1}([(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}]) \leq trace(\mathbf{H}_{IJ}) \\ &\qquad \leq \sum_{i=1}^{n-p} \lambda_i(\mathbf{H}_I)\lambda_i([(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}])\end{aligned} \tag{3.5}$$

where $\{\lambda_i(\cdot)\}_1^{n-p}$ are the real eigenvalues in decreasing order (see Schott, 2005, p. 128) for this result.

Now, the upper bounds of the RQWPL statistic follow from the above results. From Lemma 3.1 and (3.3), (3.4), and (3.5) we have

$$\frac{p\,\|\mathbf{H}_I\| \left\|(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}\right\|}{n(n-p)}, \tag{3.6}$$

$$\frac{p\{trace(\mathbf{H}_I)trace[(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}]\}}{n(n-p)} \tag{3.7}$$

and

$$\frac{p\sum_{i=1}^{n-p} \lambda_i(\mathbf{H}_I)\lambda_i([(\mathbf{I}_{n-p} - \mathbf{H}_I)^{-1}])}{n(n-p)} \tag{3.8}$$

as the upper bounds for T_J, respectively. We propose cut-off values as c/K multiplied by each of the above three upper bounds, namely

$$\frac{cp\,\|\mathbf{H}_I\|\,\left\|(\mathbf{I}_{n-p}-\mathbf{H}_I)^{-1}\right\|}{Kn(n-p)}, \tag{3.9}$$

$$\frac{cp\{trace(\mathbf{H}_I)trace[(\mathbf{I}_{n-p}-\mathbf{H}_I)^{-1}]\}}{Kn(n-p)} \tag{3.10}$$

and

$$\frac{cp\sum_{i=1}^{n-p}\lambda_i(\mathbf{H}_I)\lambda_i([(\mathbf{I}_{n-p}-\mathbf{H}_I)^{-1}])}{Kn(n-p)} \tag{3.11}$$

where c is a constant determined from simulation studies in the next section since the factor 2 from the analogue of the Hoaglin and Welsch (1978) methodology is generally too small for application purposes as we will show in Sec. 5.

4. Simulation Study

The cut-off values for the RQWPL statistic, T_J based on the Hoaglin and Welsch (1978) methodology with factor 2, i.e., $2pK^{-1}$ are generally too small. Therefore, we determine factors $c > 2$ in order to compute reasonable cut-off values for the RQ case using simulation studies. The simulation design comprises the following scenarios:

- error distributions
  - Gaussian;
- design matrices choices
  - $D1 - x_{ij}$ iid $N(0, 1)$ for $i = 1, ...n$ and $j = 2, 3, ..., p$,
  - $D2$−As in $D1$, but one point is moved 10 units in the X space,
  - $D3$−As in $D1$, but two points are moved 10 units in the X space,
  - $D4$−As in $D1$, but one point is moved 100 units in the X space,
  - $D5$−As in $D1$, but two points are moved 100 units in the X space;
- covariates: $p = 3, 4, 5, 6, 7, 8$;
- choices of τ : $n - 2\ \tau$ levels which approximately corresponds to the entire RQ solution set;
- sample size: $n = 20$; and
- number of simulation runs: 200.

The design matrix $D1$ is orthogonalized so that $\mathbf{X}'\mathbf{X} = n\mathbf{I}$ while design matrices $D2$ to $D5$, were generated as in de Jongh et al. (1988). We chose the Gaussian distribution as the appropriate error term distribution. This would avoid the distortion brought about by the interplay between the two antagonistic forces namely, the RQs exclusion of outliers (in the case of heavy tailed error term distributions) and their affinity for high leverage points.

Figure 1 below gives the scatter plot representation of matrices D2–D5.

There is only one influential point, *viz.*, observation 3, across all design matrices and distributions. The influence of this point increases as observation 5's leverage increases. However, the addition of another high leverage point (observation 7), reduces the influence of this point.

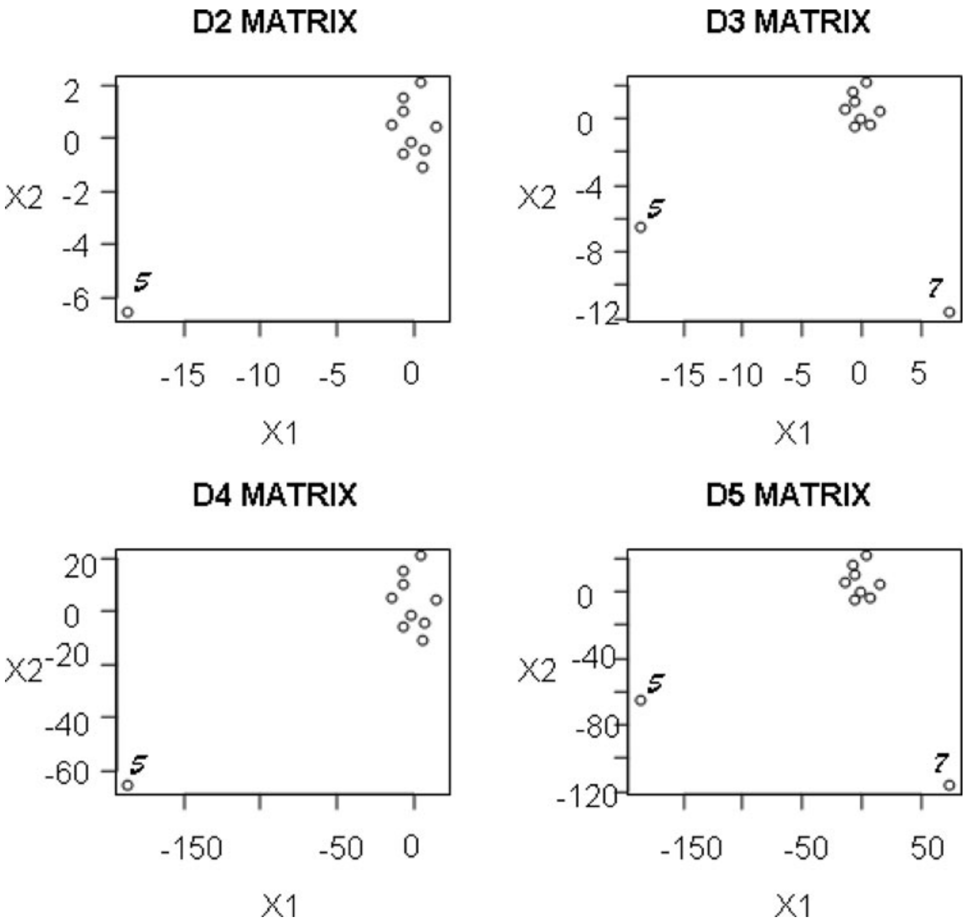


Figure 1. Scatter plots for design matrices D2–D5.

Remark 4.1. Note that D2 and D4 are similar (high leverage observation, 5 creates collinearity) but differ in the severity leverage while D3 and D5 are also similar (high leverage observation, 7 hides the collinearity) but also differ in the severity of leverage.

We start with the orthogonal design. At this design we want to find the minimum value of c for which we would flag no T_J's corresponding to RQs using cpK^{-1} as the cut-off value. This is so because D1 is our "ideal" design and we are not expected to flag any RQ at this design matrix.

For each design matrix we give the proportions of the T_J's flagged and show the display depicted graphically. The graphical representation gives us some idea of the distribution of the T_J 's since we plot the τ levels (and p sizes) against the quantiles of the simulated T_J 's.

4.1 Results for the D1 Matrix

Based on the orthogonal ("clean") D1 we do not flag any T_J's at the given c for a given size of p. Graphically, the plot of the τ levels (and p sizes) against the quantiles of the 200 simulated T_J values, $\underset{1\le i\le 200}{quant}(T_{Ji})$ at 18 τ levels (at p predictors) is as in Fig. 2 below.

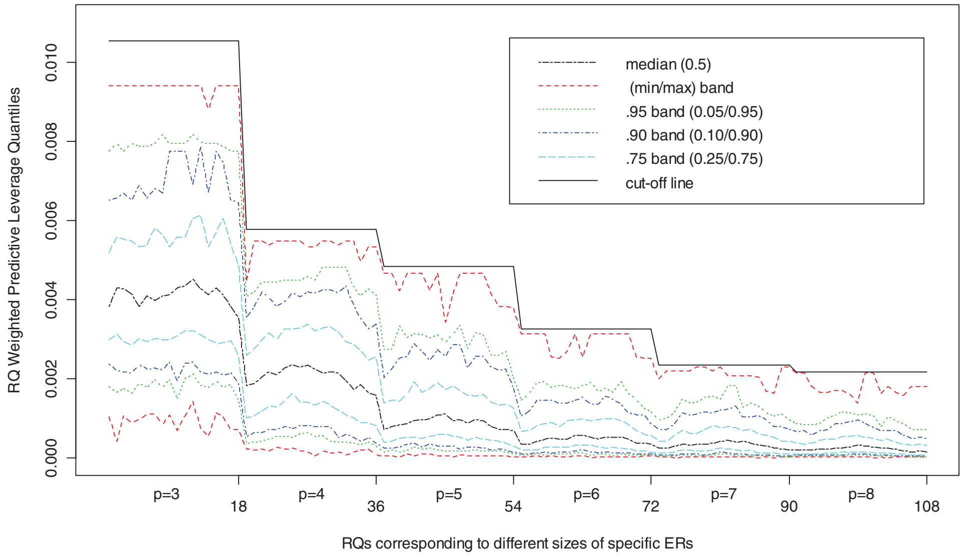


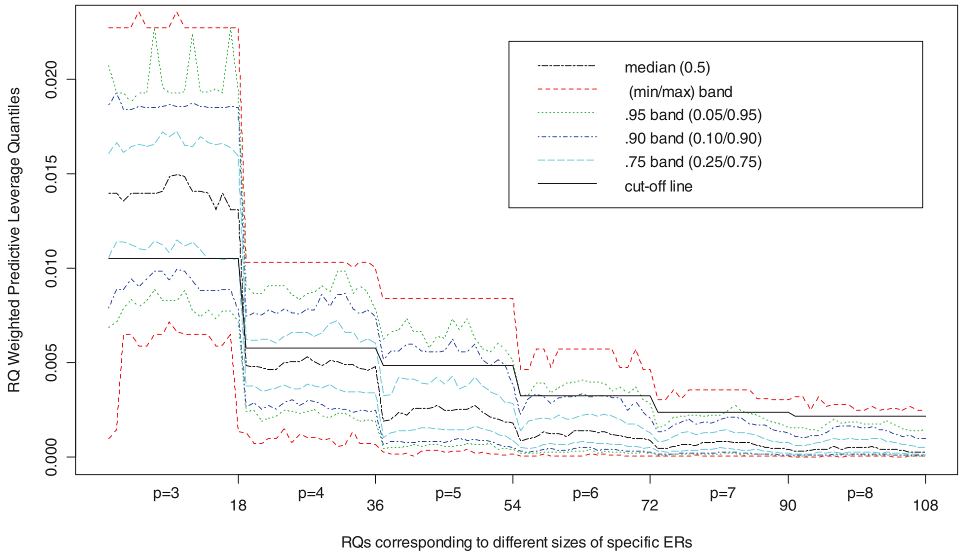
Figure 2. RQ Leverage based on D1 under the Normal distribution at different sizes of p.

At D1 under the Gaussian distribution, all the optimal solutions to the LP problem giving RQs consists of 18 unique ERs corresponding to RQs at the given sizes of p, i.e., {3, 4, ..., 8}, thus the horizontal axis consists of 18 τ levels per given size of p. The vertical axis consists of the (min/max), 95%, 90%, 75% bands, the median and the cut-off values for the 200 simulated T_J values as shown in the legend.

Based on D1 ("clean" case) we take values of c to be the set $\{4, 7, 15, 21, 26, 34\}$ corresponding to the sizes of p in the set $\{3, 4, 5, 6, 7, 8\}$, respectively. Hence, for these sizes of p, the threshold values are $\{4pK^{-1}, 7pK^{-1}, 15pK^{-1}, 21pK^{-1}, 26pK^{-1}, 34pK^{-1}\}$, respectively.

Remark 4.2. For the graphical representations that will follow and the various statistics we are investigating, the horizontal axis will stay the "same" as in Fig. 2, i.e., we will fix the τ's at equally spaced 18 levels at the given p sizes. This helps us to carry out convenient comparisons among the different scenarios as each scenario may result in a different number of unique RQs.

Note that the distribution of the T_J's generally decreases in an approximate stepwise fashion. That is, as K increases (or as p approaches $\frac{n}{2}$), the statistic T_J generally becomes smaller and smaller. The determined new cut-off values are given by the solid line. Note that the cut-off values flag no RQs at this "clean" case as expected. It is interesting to see if these cut-off values will result in the flagging of RQs in the presence of different leverage scenarios. We only give results for D2 and D3 since the flagging pattern of the RQs at D2 is similar to that at D4 and that at D3 similar to that at D5 except that it more pronounced at the latter design matrices.



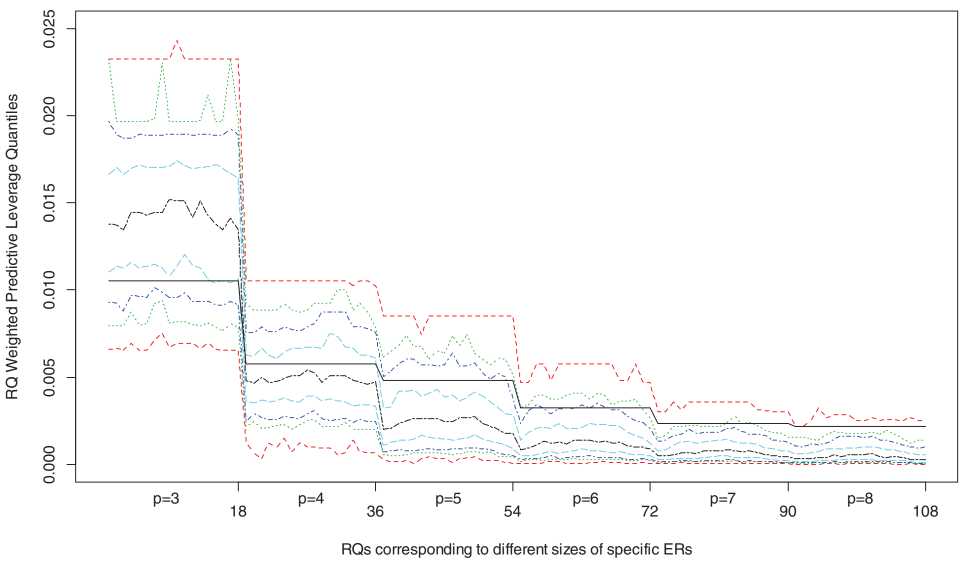


Figure 3. RQ Leverage based on D2 and D3 under the Normal distribution at different sizes of p. Upper panel D2; lower panel D3.

4.2 Results for D2 and D3 Matrices

In order to see the effect of the degree of leverage on the distribution of the T_J we give the graphs of the τ levels (and p sizes) against the quantiles of the 200 simulated T_Js at 18 τ levels, i.e., 200×18 T_Js. We give D2 and D3 results together in Fig. 3 below to see the differences that may arise due to collinearity inducing high leverage points and collinearity hiding high leverage points, respectively. Since the D2 and D4 matrices are similar, we present the results for D2 only.

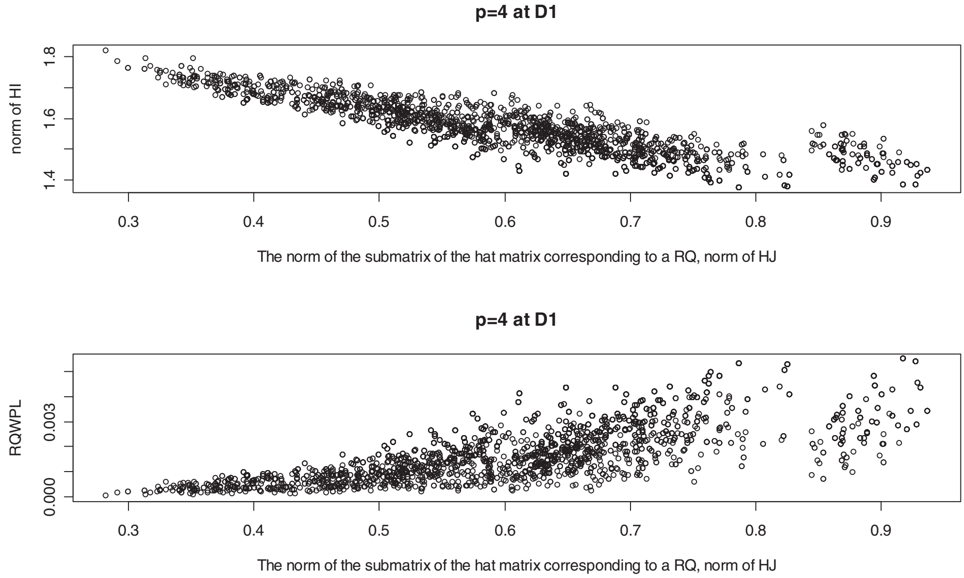


Figure 4. Plot of $||\mathbf{H}_I||$ vs. $||\mathbf{H}_J||$, upper panel; plot of T_J vs. $||\mathbf{H}_J||$, lower panel at $p = 4$ and D1.

On average we flag 78% for $p = 3$, 37% for $p = 4$, 15% for $p = 5$, 8% for $p = 6$, 0.0% for $p = 7$, 6% for $p = 8$ of the RQs using the threshold values based on c for this design matrix.

Again since the D3 and D5 matrices are similar, we present the results for D3 only.

On average, we flag 98% for $p = 3$, 100% for $p = 4$, 63% for $p = 5$, 37% for $p = 6$, 23% for $p = 7$, 22% for $p = 8$ of the RQs using the cut-off values based on c for this design matrix.

On average, we flag 100% for $p = 3$, 100% for $p = 4$, 64% for $p = 5$, 38% for $p = 6$, 22% for $p = 7$, 23% for $p = 8$ of the RQs using the threshold values based on c for this design matrix.

At D3 T_Js tend to be larger and more RQs are flagged than at D2 and generally, the proportion of RQs being flagged generally decreases as p increases.

Now, we give simulation comparisons of the proposed multiple-case leverage diagnostic T_J with those in the literature, $||\mathbf{H}_I||$ and $||\mathbf{H}_J||$ for some high leverage scenarios at D1, the control design matrix and $p = 4$, and analogues of the Hocking and Pendleton and Gunst and Mason data sets, design matrices D3 and D4, respectively. At the orthogonal design, D1 $||\mathbf{H}_I||$ is negatively correlated with $||\mathbf{H}_J||$ while $||\mathbf{H}_J||$ and T_J are positively correlated as shown in Fig. 4 below.

Like the Hocking and Pendleton data set, design matrices D3 and D5 have a high leverage point that hides collinearity. Figure 5 below shows results for D3.

At design matrix D3 the correlation pattern is as at D1. However, a higher and more drastic flagging rate at lower values of p depicted in Fig. 2 is also evident. Only 4 values are very small for both T_J and $||\mathbf{H}_J||$.

The analog of the Gunst and Mason data set are the design matrices D2 and D4. We give results for D4 only in Fig. 6 below.

At $p = 7$, a more gradual increase in both T_J and $||\mathbf{H}_J||$ is synonymous with a more gradual and less drastic flagging rate as depicted at larger values of p in Fig. 3.

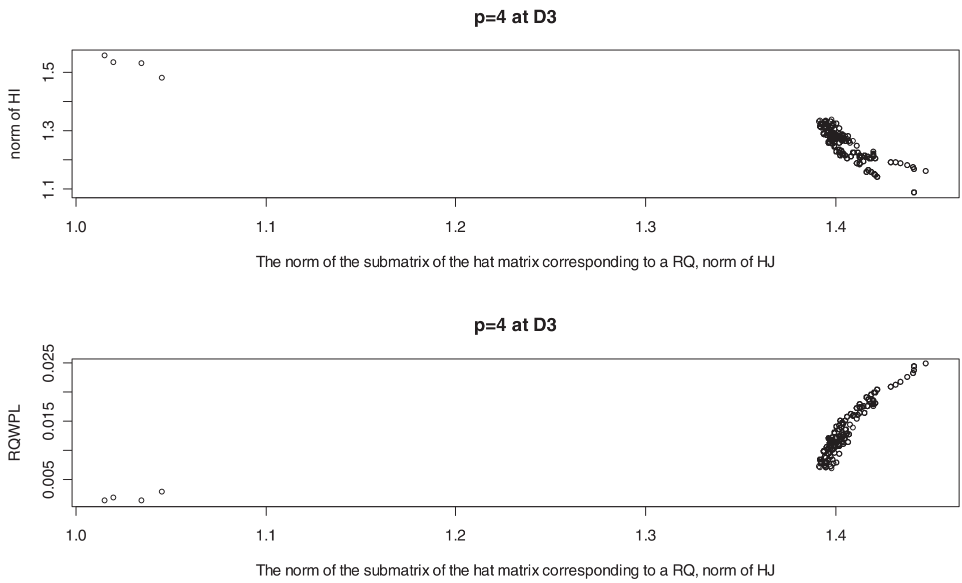


Figure 5. Plot of $||\mathbf{H}_I||$ vs. $||\mathbf{H}_J||$, upper panel; plot of T_J vs. $||\mathbf{H}_J||$, lower panel at $p = 4$ and D3.

These simulation results show that the RQWPL statistic T_J compares very well to the existing multiple-case diagnostic $||\mathbf{H}_J||$.

In the next section we apply the cut-off values of the RQWPL statistic based on the analogue of the Hoaglin and Welsch (1978) one as well as those proposed in section 3 based on the simulation to the above-mentioned 3 data sets. In order to find out the single case high leverage points involved in the high leverage multiple-cases, we first apply the

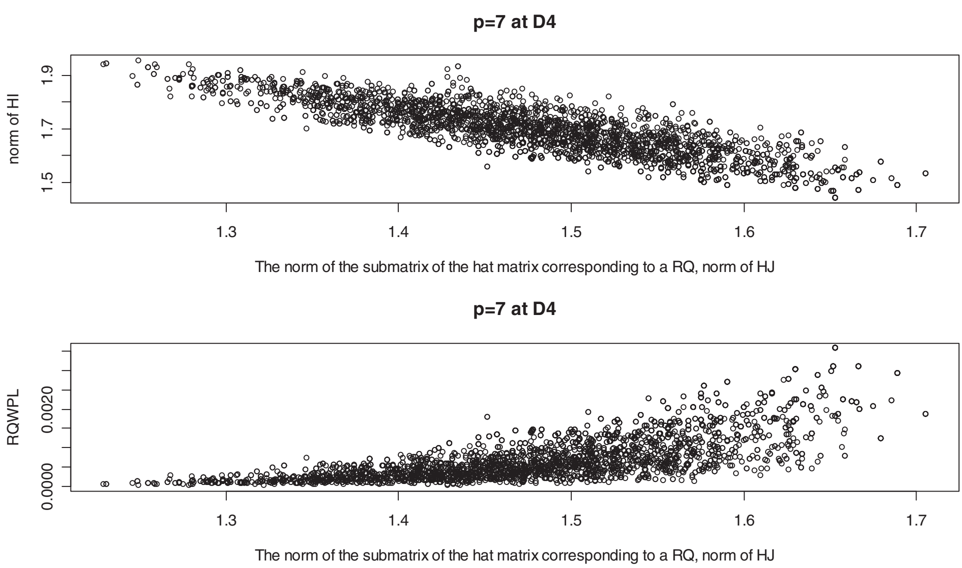


Figure 6. Plot of $||\mathbf{H}_I||$ vs. $||\mathbf{H}_J||$, upper panel; plot of T_J vs. $||\mathbf{H}_J||$, lower panel at $p = 7$ and D4.

robust and multivariate location and scale estimates of Rousseeuw and van Driessen (1999) in order to circumvent the masking and swamping effects that might be inherent in the data sets.

5. Applications

It is wellknown that the usual OLS single case diagnostics suffer from the masking and swamping effects. Therefore, we make use of the robust and multivariate location and scale estimates computed with the minimum covariance determinant (MCD) method of Rousseeuw and van Driessen (1999) to expose all the single case high leverage points and outliers.

In this section we consider three data sets. First, we consider the Gunst and Mason data set (see Gunst and Mason, 1980). It has 6 predictors and 49 observations, of which high leverage observations 17 and 39 induce collinearity, while high leverage observation 27 obscures collinearity. Other high leverage observations are 3, 8, 20, 24, 26, 33, 37, 43, and 46. Graphically, some of these high leverage points are shown to be more severe as well as being outliers in Fig. 7 below.

Second, we consider the Hocking and Pendleton data set (see Hocking and Pendleton, 1983) which has 3 predictors and 26 observations of which 24 is a high leverage point that hides collinearity. The other points with notable relatively high leverage are 11 and 18 which are also outliers. Observation 17 is an outlier. The plot revealing these high leverage points and outliers is shown in Fig. 8 below.

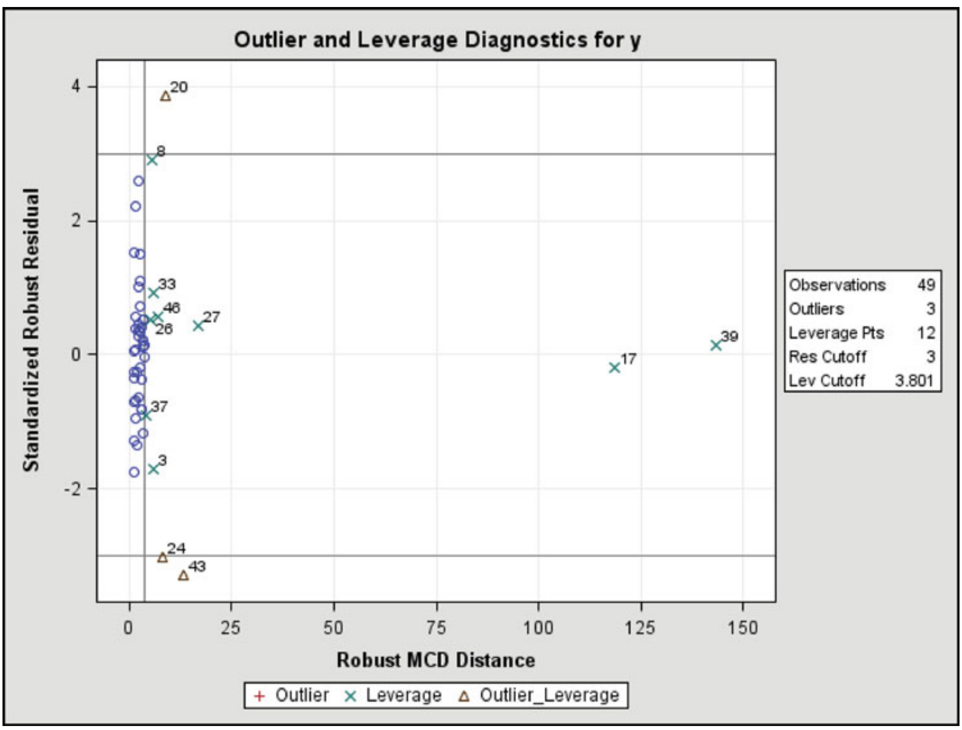


Figure 7. High leverage and outlier diagnosis for the Gunst and Mason data set.

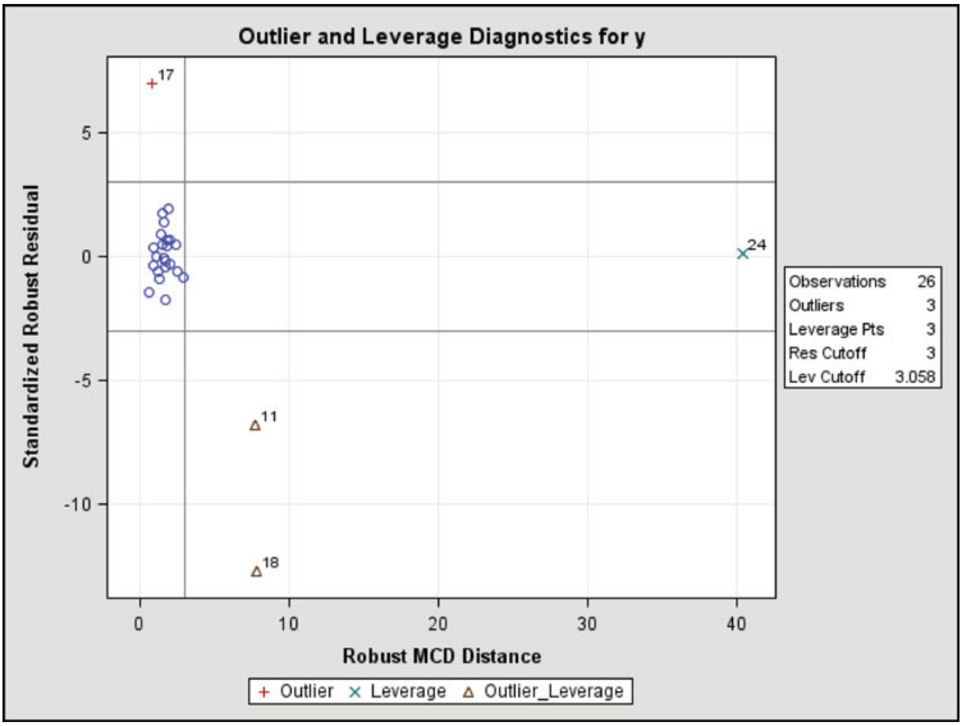


Figure 8. High leverage and outlier diagnosis for the Hocking and Pendleton data set.

Finally, we consider the Hald data set (Montgomery and Peck, 1982) which has 4 predictors and 13 observations with relatively very moderate high leverage points, 3 and 10, and no outliers

Since the high leverage points in the first two data sets change the eigenstructure of the design matrices thereby, inducing or hiding collinearity, they are called collinearity-influential points. The Hald data set will be used as more or less of the control since it has very moderate high leverage points (see Fig. 9 below).

Tables 2–4 below give the RQ high leverage view at different $0 < \tau < 1$ levels for the Gunst and Mason, Hocking and Pendleton, and Hald data sets, respectively, using the cut-off value based on the Hoaglin and Welsch (1978) methodology. The τ levels correspond to the entire solution set of the LP giving unique RQs. The tables are sorted in ascending order of $-\ln(T_J)$ (descending order of T_J).

High leverage points at both the single case and the RQ case are shown in bold in all three tables. Flagged values for T_J, i.e., $-\ln(T_J) < -\ln(2pK^{-1})$ are sorted from smallest value of $-\ln(2pK^{-1})$ to the largest. We discuss the flagging percentages of the three data sets below.

Discussion

The flagging rate of the Gunst and Mason data set, given in Table 2, is second highest with 90.70% (having $-\ln(T_J) < 15.630$). In this data set, the most severe influential weighted leverage subgroups are based on the highest single leverage points 27 (obscuring collinearity), 39 (inducing collinearity), and the moderate leverage point 20 which apparently form

Table 2
RQs that are flagged due to high multiple-case weighted leverage[1]

Dataset	τ	$-\ln(T_J)$	ERs Corresponding to RQs						
Gunst and Mason		Flagged							
cut-off = 15.630	0.772	**8.759**	**8**	10	**20**	**26**	**27**	**39**	**46**
	0.957	**8.860**	10	**20**	25	**26**	**27**	**39**	**46**
	0.814	**8.888**	10	**20**	**26**	**27**	**33**	**39**	**46**
	0.748	**8.933**	**8**	10	16	**20**	**27**	**39**	**46**
	0.875	**8.979**	10	**20**	**26**	**27**	**39**	41	**46**
	0.845	**9.020**	10	**20**	**26**	**27**	**39**	45	**46**
	0.937	**9.020**	10	**20**	**26**	**27**	**39**	42	**46**
	0.726	**9.324**	10	16	**20**	**27**	31	**39**	**46**
	0.698	**9.328**	1	10	16	**20**	**27**	**39**	**46**
	0.635	**11.081**	1	10	16	**27**	35	**39**	**46**
	0.588	**11.483**	10	13	15	**27**	35	**39**	**46**
	0.629	**11.561**	10	13	**27**	35	**39**	44	**46**
	0.674	**11.810**	1	10	16	**27**	**39**	46	**48**
	0.580	**11.910**	9	10	13	15	**27**	**39**	**46**
	0.107	**12.180**	**3**	5	**20**	32	37	**39**	**49**
	0.567	**12.789**	13	15	**17**	29	38	**39**	**46**
	0.532	**13.005**	13	**17**	22	29	38	**39**	**46**
	0.244	**13.032**	**17**	**20**	21	30	32	34	**39**
	0.196	**13.036**	**17**	19	23	32	**37**	**39**	49
	0.580	**13.036**	9	10	13	15	**17**	**39**	**46**
	0.260	**13.250**	**17**	**20**	30	32	34	39	47
	0.106	**13.309**	**3**	5	**20**	37	**39**	40	49
	0.373	**13.346**	2	9	**17**	23	29	**39**	**46**
	0.382	**13.377**	9	14	**17**	23	29	**39**	**46**
	0.519	**13.702**	9	11	13	**17**	38	**39**	**46**
	0.057	**13.757**	**3**	5	**20**	**39**	40	**43**	49
	0.140	**13.954**	**3**	5	6	32	**37**	**39**	49
	0.060	**14.058**	**3**	5	37	**39**	40	**43**	49
	0.496	**14.143**	9	11	**17**	28	38	39	**46**
	0.175	**14.236**	**3**	19	32	34	**37**	**39**	49
	0.469	**14.286**	9	14	**17**	28	29	38	**46**
	0.449	**14.445**	9	14	**17**	18	28	29	**46**
	0.221	**14.592**	**17**	21	30	32	34	**39**	49
	0.147	**14.639**	**3**	6	32	34	**37**	**39**	49
	0.270	**14.755**	**17**	**20**	30	32	34	38	47
	0.205	**14.796**	17	21	23	32	34	**39**	49
	0.178	**14.871**	19	23	32	34	**37**	**39**	49
	0.285	**14.977**	9	**17**	**20**	30	32	38	47
	0.054	**15.206**	**3**	5	**20**	37	40	**43**	49
		Not Flagged							
	0.366	15.661	2	9	**17**	23	29	38	**39**
	0.342	15.863	2	9	**17**	23	36	38	**39**
	0.050	16.609	**3**	5	**20**	24	32	40	**43**
	0.323	16.890	9	**17**	23	30	36	38[1]	47

[1]high leverage points at both the single case and the RQ case are shown in bold

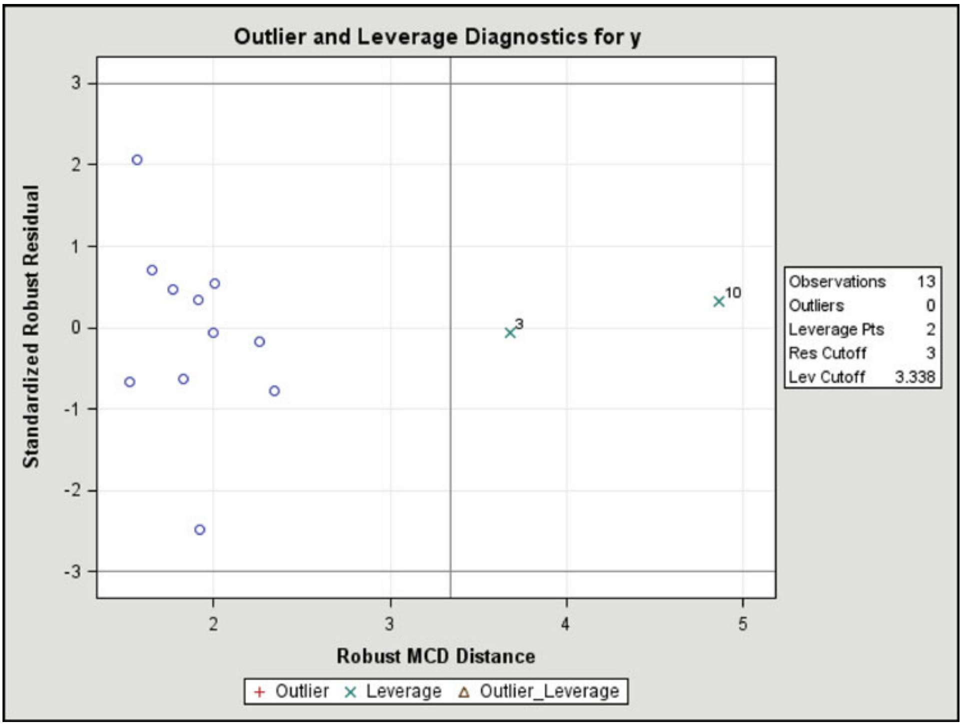


Figure 9. High leverage and outlier diagnosis for the Hald data set.

the "anchor points" to which three more obscure points are added to create the high leverage subset. Some values of $-\ln(T_J)$ are very far away from the cut-off value indicating that the most severe high leverage subsets occur in this data set.

The flagging rate of the Hocking and Pendleton data set, given in Table 4, is the highest with 96.00% (having $-\ln(T_J) < 7.533$). In this data set, the single most important anchor leverage point is the collinearity-obscuring point 24 which combines with the moderate leverage point 8 and low leverage values 6 and 16 to create the highest leverage subset. The values of $-\ln(T_J)$ are relatively closer to the cut-off value compared to the Gunst and Mason data set, indicating that these are not severe high leverage subsets. The Hocking and Pendleton data set with the single case high leverage point, 24, that hides collinearity has a higher flagging rate of high leverage RQs than the Gunst and Mason data set with single case high leverage points that both induce and hide collinearity. However, the difference in the flagging rate between the first two data sets is marginal.

As expected, the flagging rate was least in the Hald data set, given in Table 4, which is the control (but still quite high at 70.00%). This flagging rate is rather too high for the control data set although it is wellknown that there may be situations where observations are individually not high leverage points, but have high leverage jointly.

In the Hald data set, there are no severe high leverage subsets in this data set. Although there is a high flagging rate in this data set, there are no extremely high leverage RQs compared to the other two data sets as the values of $-\ln(T_J)$ are pretty close to the cut-off value. However, the subsets that are flagged are based mostly on the relatively larger single case leverage points, 3 and 10 in conjunction with points 11 and 12.

Table 3
RQs that are flagged due to high multiple-case weighted leverage[1]

Dataset	τ	$-\ln(T_J)$	ERs Corresponding to RQs			
Hocking cut-off =		Flagged				
7.533	0.957	**5.535**	6	8	16	**24**
	0.123	**5.574**	8	**11**	19	**24**
	0.622	**5.613**	8	9	10	**24**
	0.73	**5.717**	8	10	15	**24**
	0.205	**5.745**	8	13	14	**24**
	0.723	**5.784**	8	9	15	**24**
	0.955	**5.805**	2	6	8	**24**
	0.305	**5.822**	1	14	16	**24**
	0.631	**5.949**	8	9	**24**	25
	0.739	**5.982**	6	8	21	**24**
	0.253	**5.999**	1	14	**24**	26
	0.186	**6.024**	8	12	13	**24**
	0.828	**6.028**	6	8	22	**24**
	0.551	**6.039**	3	8	10	**24**
	0.469	**6.067**	10	14	16	**24**
	0.402	**6.121**	1	14	23	**24**
	0.259	**6.189**	1	5	14	**24**
	0.684	**6.211**	9	15	**24**	25
	0.537	**6.299**	7	10	14	**24**
	0.366	**6.388**	1	4	16	**24**
	0.545	**6.684**	3	9	10	**24**
	0.766	**6.769**	6	21	22	**24**
	0.441	**6.888**	14	16	23	**24**
	0.085	**7.389**	8	**11**	16	**18**
		Not Flagged				
	0.093	7.811	8	**11**	16	19[1]

[1]high leverage points at both the single case and the RQ case are shown in bold.

The high inclusion rate of single case high leverage points (shown in bold in Tables 2 and 3) and single observations with notably relatively high leverage points, 3 and 10 (in Table 4) in ERs corresponding to specific RQs is clearly evident in all data sets.

The application of the regression quantile procedure to the above-mentioned 3 data sets confirms the RQs' high affinity for high leverage points and exclusion of outliers, e.g., observation 17 in the Hocking and Pendleton data set is never included. Also, the cut-off values deduced from the Hoaglin and Welsch (1978) methodology are generally too small in practice as evidenced by a very low flagging rate. Therefore, we apply the cut-off values proposed in Sec. 3 and based on the simulation study of Sec. 4 to the three data sets. Figure 10 below shows The RQWPL statistic T_J values and cut-off values (3.9), (3.10), Hoaglin and Welsch (1978) (HW) analog, $2pK^{-1}$, and the one based on simulation for the Gunst and Mason, Hocking and Pendleton, and the Hald data sets. We left out bound 3.11 as it was again generally very small in the applications.

Table 4
RQs that are flagged due to high multiple-case weighted leverage[1]

Dataset	τ	$-\ln(T_J)$	ERs Corresponding to RQs				
Hald cut-off =		Flagged					
4.857	0.514	**4.092**	1	**3**	**10**	11	12
	0.426	**4.123**	1	**3**	5	7	**10**
	0.337	**4.213**	**3**	5	7	8	**10**
	0.429	**4.406**	1	**3**	5	**10**	12
	0.752	**4.605**	1	9	**10**	11	12
	0.825	**4.646**	1	2	**10**	11	12
	0.547	**4.689**	1	5	**10**	11	12
		Not Flagged					
	0.253	5.021	**3**	4	5	7	8
	0.23	5.319	1	**3**	4	5	8
	0.86	5.809	1	2	5	11	12[1]

[1]high leverage points at both the single case and the RQ case are shown in bold

Bounds (3.9) and (3.10) exhibit a very similar flagging rate. Though these two bounds exhibit a lower flagging rate compare to the simulation bound they exhibit a higher flagging rate than the Hoaglin and Welsch (1978) (HW) analog, $2pK^{-1}$.

Comparison of the RQWPL statistic T_J to $||\mathbf{H}_J||$ shown in Fig. 11 below is consistent with the simulation results

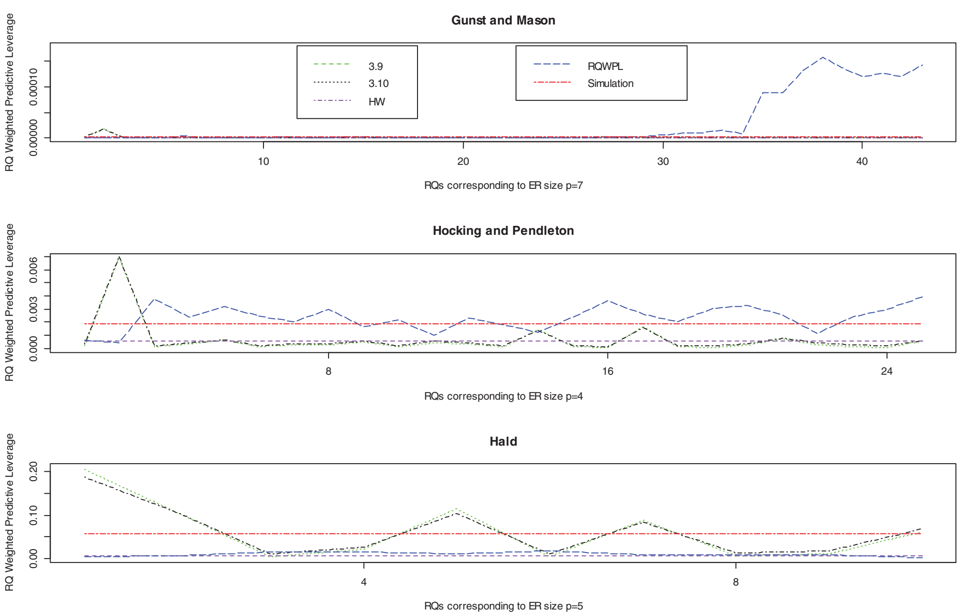


Figure 10. The RQWPL values and cut-off values (3.9), (3.10), Hoaglin and Welsch (1978) (HW) analog and the one based on simulation for the Gunst and Mason, Hocking and Pendleton, and the Hald data sets.

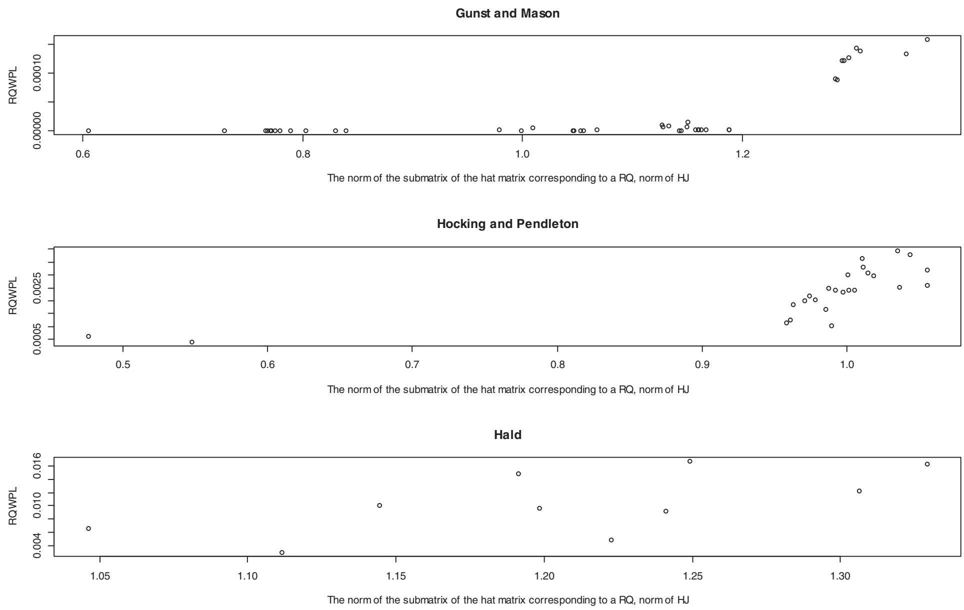


Figure 11. Comparison of T_J and $||\mathbf{H}_J||$ for the Gunst and Mason, Hocking and Pendleton, and the Hald data sets.

However, note that Gunst and Mason data set is not quite like D7 as it contains some outlier which may be responsible for the some sort of distortion of the flagging rate and multiple-case high leverage pattern.

7. Conclusions

In this article, it was shown that an extension of the Hoaglin and Welsch (1978) methodology of identifying single case high leverage points by $h_i > 2p/n$ to the multiple-case level by flagging ERs (RQs) having $T_J > 2pK^{-1}$, is mathematically tractable. However, this leads to cut-off values which are generally small. Therefore, we proposed some analytical bounds and more reasonable cut-off values based on these bounds and a simulation study. We adjusted the given form of the cut-off value, by flagging RQs with $T_J > cpK^{-1}$, for some factor $c > 2$. This resulted in a lower reasonable flagging rate. The other cut-off values were deduced by multiplying the analytical bounds by the factor cK^{-1}. These cut-off values generally exhibit a moderate flagging rate between that from the simulation study, cpK^{-1} and the Hoaglin and Welsch (1978) analogue $2pK^{-1}$.

We chose to compare the RQWPL statistic T_J to the well-known multiple-case high leverage diagnostic, $||\mathbf{H}_I||$ (and $||\mathbf{H}_J||$) as they are not measure specific. From the simulation study the RQWPL statistic T_J is negatively correlated with $||\mathbf{H}_I||$ but positively correlated with $||\mathbf{H}_J||$. Therefore, we can deduce that T_J measures the multiple-case leverage of the RQ corresponding to ER J as does $||\mathbf{H}_J||$ as opposed to $||\mathbf{H}_I||$ which measures the multiple-case leverage of the complement of subset J, I.

The RQs' affinity for leverage points was evident as well as its exclusion of outliers.

A second approach to determine the cut-off values for T_J could be based on deriving its distribution under some regularity conditions such as those used in the derivation of the distributions of h_i and $h_{(i)}$ in Chatterjee and Hadi (1988).

The RQWPL statistic T_J could be incorporated in determinantal subset influence measures (see, e.g., Barrett and Gray, 1996). Both proposals are currently being investigated by the authors.

Acknowledgment

The authors would like to thank the reviewers for their comments and suggestions that led to a substantial improvement of the article.

Funding

This research was supported by the University of South Africa's Research fund and the University of KwaZulu-Natal Competitive Research Grant.

References

Aitken, A. G. (1964). *Determinants and Matrices*. Edinburgh: Oliver and Boyd.

Chatterjee, S., Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons.

Barrett, B. E., Gray, J. B. (1996). Computation of determinantal subset influence in regression. *Statist. Comput.* 6:131–138.

Barrett, B. E., Gray, J. B. (1997). Leverage, residual and interaction diagnostics for subsets of cases in least squares regression. *Computat. Statisti. Data Anal.* 26:39–52.

De Jongh, P. J., De Wet, T., Welsh, A. H. (1988). Mallows-type bounded-influence regression trimmed means. *J. Amer. Statist. Assoc.* 83:805–810.

Graybill, F. A. (1983). *Matrices with Applications in Statistics*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Gunst, R. F., Mason, R. L. (1980). *Regression Analysis and Its Application: A Data Oriented Approach*. New York: Marcel Dekker.

Hadi, A. S., Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J. Amer. Statist. Assoc.* 88:1264–1272.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York: Springer.

Hawkins, D. M., Bradu, D., Kass, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26:197–208.

Hawkins, D. M. (1993). The accuracy of elemental set approximations for regression. *J. Amer. Statist. Assoc.* 88:580–589.

Hoaglin, D. C., Welsch, R. E. (1978). The hat matrix and Anova. *Ameri. Statistician* 32:17–22.

Hocking, R. R., Pendleton, O. J. (1983). The regression dilemma. *Commun. Statist. Part A-Theor. Meth.* 12:497–527.

Kempthorne, P. J., Mendel, M. B. (1990). Comment. *J. Amer. Statist. Assoc.* 85:647–648.

Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica* 46:33–50.

Koenker, R. (2005). *Quantile regression*. Econometric Society Monographs. New York: Cambridge University Press.

Mayo, M. S., Gray, J. B. (1997). Elemental subsets: the building blocks of regression. *Amer. Statistician* 51:122–129.

Montgomery, D. C., Peck, E. A. (1982). *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *J. Amer. Statist. Assoc.* 85:633–651.

Rousseeuw, P. J., Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41:212–223.

Schott, J. R. (2005). *Matrix Analysis for Statistics.* 2nd ed. Hoboken, NJ: John Wiley & Sons.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley & Sons.
Seaver, B., Trantis, K., Reeves, C. (1999). The identification of influential subsets in regression using a fuzzy clustering strategy. *Technometrics, Amer. Statist. Assoc. Amer. Soc. Qual.* 41:340–351.
Sheynin, O. B. (1973). R. J. Boscovich's work on probability. *Arch. History Exact Sci.* 9:306–324.
Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Cambridge, MA: Harvard University Press.
Yang, X. M., Yang, X. Q., Teo, K. L. (2001). A matrix trace inequality. *J. Mathemat. Anal. Applic.* 263:327–331.
Yu, K., Lu, Z., Stander, J. (2003). Quantile regression: applications and current research areas. *Statistician*, 52:331–350.