View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Papers in Economics

Estimators of the Multiple Correlation Coefficient: local robustness and confidence intervals.

Christophe Croux and Catherine Dehon *
Université Libre de Bruxelles

Abstract: Many robust regression estimators are defined by minimizing a measure of spread of the residuals. An accompanying R^2-measure, or multiple correlation coefficient, is then easily obtained. In this paper, local robustness properties of these robust R^2-coefficients are investigated. It is also shown how confidence intervals for the population multiple correlation coefficient can be constructed in the case of multivariate normality.

Key words: Influence function, Multiple correlation coefficient, Regression analysis, R^2-measure, Robustness.

*ECARES, Faculté SOCO, and Institut de Statistique, Université Libre de Bruxelles, CP-114, Av. F.D. Roosevelt 50, B-1050 Brussels, Belgium.

1 Introduction

The R^2-statistic, or squared multiple correlation coefficient, is a standard tool in applied regression analysis. Although it is not always thoughtfully used, it remains an informative summary measure of the predictive power of the selected regression model. Cautionary notes about R^2 can be found in Kvalseth (1985) and Willet & Singer (1988). We work in a standard linear regression model with intercept:

$$y_i = \alpha + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i \qquad i = 1, \cdots, n \tag{1.1}$$

where n is the sample size, $x_i = (x_{i1}, \cdots, x_{ip})^t$ the vector containing the explanatory variables and y_i the response variable. Suppose that the errors ε_i are independent of the explanatory variables and i.i.d. according to a distribution $F_\sigma(x) = F_0(x/\sigma)$, where σ is the residual scale parameter, and F_0 verifies

(F) F_0 is is symmetric and unimodal around zero, with strictly positive density.

With $r_i(\hat{\beta}_{\mathrm{LS}}) = y_i - x_i^t\hat{\beta}_{\mathrm{LS}} - \hat{\alpha}_{\mathrm{LS}}$ the residuals from the Least Squares (LS) fit, the classical R^2 coefficient is defined as

$$R^2_{\mathrm{LS}} = 1 - \frac{\sum_{i=1}^n r_i(\hat{\beta}_{\mathrm{LS}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{1.2}$$

with $\bar{y}$ the sample average of the dependent variable. We see that the numerator in (1.2) equals the variance of the residuals in the full model, while the denominator is the variance of the residuals in the reduced model:

$$y_i = \alpha_0 + \varepsilon_i \qquad i = 1, \cdots, n. \tag{1.3}$$

Indeed, the LS estimator of α_0 in (1.3) equals $\bar{y}$.

Since the LS estimator is very vulnerable in presence of outliers, it is not surprising that R^2_{LS} inherits this problem. This is illustrated in Figure 1, where the value of R^2_{LS} decreases drastically when one single outlier, represented by an inverted triangle, is added to the sample. Although the fitted LS line is not enormously altered, the R^2_{LS} decreases from 83% to 30%.

Let us do now an experiment allowing us to study the influence of an outlier in the y-direction (called a vertical outlier) or in the x-direction (called a leverage point) on R^2_{LS}. A data set was generated and its multiple correlation coefficient computed. Afterwards a

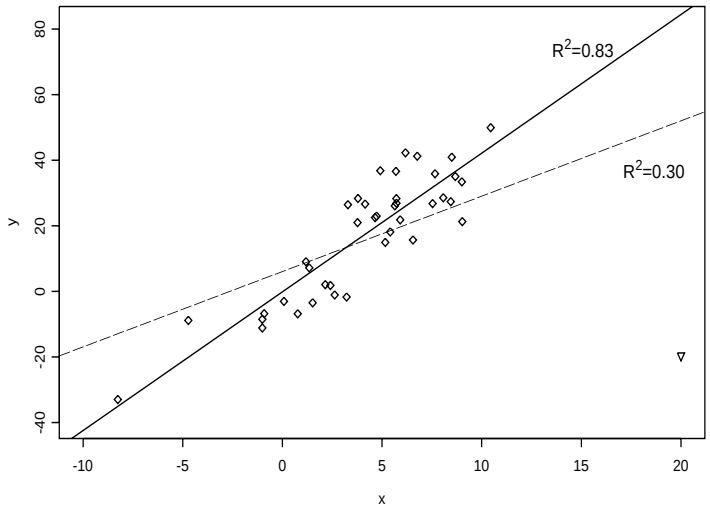

Figure 1: Effect of adding one single outlier, indicated by ∇, on R^2_{LS}.

point with coordinates $(0, y)$ is added, where the value of y ranges in the interval $[-5, 5]$, as is seen in the upper left panel of Figure 2. The lower left panel shows how this affects the value of R^2_{LS}. A similar exercise is done for leverage points, by letting vary the value x of an outlier placed at $(x, 0)$. The value of R^2_{LS} did decrease when adding these points, and by letting x or y tend to infinity it will even tend to zero. Figure 2 confirms that R^2_{LS} is highly sensible to single outliers. Later on in this paper we will quantify the influence of adding observations by computing influence functions. We will find that there can also be an increase in R^2 when adding outliers.

In Section 2 we will introduce R^2-measures which go along with robust regression fits. Influence functions will be computed in Section 3. We will focus on regression estimators based on the minimization of M-estimators of residual scale. The Most B-robust measure of multiple correlation which minimizes the maximal influence that an observation can have on R^2, is obtained. Section 4 explains how confidence intervals for the population multiple correlation coefficient can be constructed in the case of multivariate normality, while Section 5 concludes.

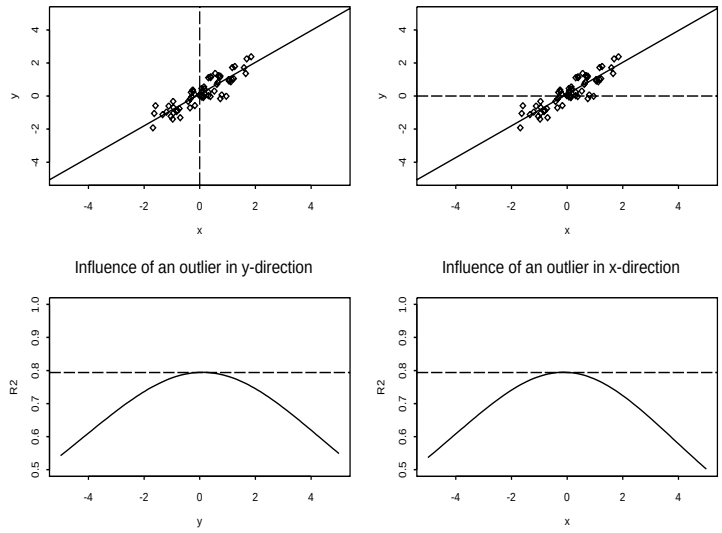

Figure 2: Effect of adding one observation $(0, y)$ *to a data cloud on the value of* R^2_{LS} *(left panel) and effect of adding one observation* $(x, 0)$ *on the value of* R^2_{LS} *(right panel). The horizontal dashed line indicates the value of* R^2_{LS} *at the original sample.*

2 A robust measure of R^2

We introduce the robust R^2-measure as in Anderson-Sprecher (1994). Consider a regression estimator defined by

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)}{\operatorname{argmin}} S_n(r_1(\alpha, \beta), \ldots, r_n(\alpha, \beta)) \tag{2.1}$$

with $r_i(\alpha, \beta) = y_i - x_i^t\beta - \alpha$ and S_n a residual scale estimator verifying

$$S_n(ae_1, \ldots, ae_n) = |a|S_n(e_1, \ldots, e_n)$$

for all $e_1, \ldots, e_n$ and $a \in I\!R$. Now we fit the regression model (1.3) having only an intercept term, yielding

$$\hat{\alpha}_0 = \underset{\alpha_0}{\operatorname{argmin}} S_n(y_1 - \alpha_0, \ldots, y_n - \alpha_0).$$

By comparing the estimated residual scales in the full model and in the reduced model (1.3), we define by analogy with the classical formula (1.2)

$$R^2_S = 1 - \frac{S^2_n\left(r_1(\hat{\alpha}, \hat{\beta}), \ldots, r_n(\hat{\alpha}, \hat{\beta})\right)}{S^2_n\left(y_1 - \hat{\alpha}_0, \ldots, y_n - \hat{\alpha}_0\right)}. \tag{2.2}$$

Since (1.3) is a submodel of (1.1), and by the definition of the regression estimators (2.1) one readily sees that

$$0 \leq R_S^2 \leq 1.$$

For every residual scale estimator S we have another regression estimator and another measure of multiple correlation. Values of R_S^2 close to zero indicate that the explanatory variables are not explaining much of the dispersion of y, while a value of R^2 close to one indicates that the dispersion of the residuals is almost zero, so an excellent fit. This interpretation is standard, but the meaning of "dispersion" depends on the selected measure of scale S.

We will focus on the class of M-estimators of residual scale, where $S_n(e_1, \ldots, e_n)$ is defined as the solution of the following equation in s:

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\frac{e_i}{s}) = b.$$

The function ρ is supposed to verify

(R) The function ρ is symmetric and non decreasing on $[0, \infty[$. Furthermore, $\rho(0) = 0$ and ρ is almost everywhere continuously differentiable.

The number b is a selected constant, chosen as $E_{F_0}[\rho(\varepsilon)]$ to ensure consistent estimation of σ at the model.

If we take $\rho(u) = u^2$ (and $b = 1$), then S_n is a sum of squares and we get the classical R^2_{LS} for (2.2). Instead of a squared loss function one could take $\rho(u) = |u|$, which leads to the minimization of the sum of absolute values of the residuals and yields the L_1 regression estimator. The associated R^2_{L1} is given by

$$R^2_{\mathrm{L1}} = 1 - \left(\frac{\sum_{i=1}^{n} |y_i - x_i^t \hat{\beta}_{\mathrm{L1}} - \hat{\alpha}_{\mathrm{L1}}|}{\sum_{i=1}^{n} |y_i - \mathrm{median}_i\, y_i|} \right)^2 .$$

The above measure, and a variant of it, have been studied by McKean & Sievers (1987). A smooth and bounded ρ function is the Tukey Biweight

$$\rho_c(u) = \min\left(\frac{c^2}{6}, \frac{u^2}{2} - \frac{u^4}{2c^2} + \frac{u^6}{6c^4}\right). \tag{2.3}$$

The resulting estimator is then an S-estimator (Rousseeuw & Yohai 1984), which we call the Biweight S-estimator (BS). The constant c in (2.3) determines the breakdown point of the estimator, which is the maximal fraction of contamination that an estimator can withstand.

By default, the 50% breakdown version is taken, but we will also consider the 25% breakdown Biweight S-estimator (BS25).

Taking a step function

$$\rho_c(u) = I(|u| > c),$$

yields the $(1-b)$ quantile of the absolute values of the residuals:

$$S_n(e_1, \ldots, e_n) = c^{-1}|e|_{([[(1-b)n])}.$$

The value of b is determined by $b = E_{F_0}[\rho_c(\varepsilon)] = 2(1 - F_0(c))$. For $b = 0.5$ we minimize the median of the absolute values of the residuals, and get the *Least Median of Squares* (LMS) regression estimator of Rousseeuw (1984). For other values of b we will speak about the Least Quantile of Squares estimator (Rousseeuw & Leroy 1987, p. 124). For the LMS, as already noticed by Anderson-Sprecher (1994), (2.2) results in

$$R^2_{\text{LMS}} = 1 - \left(\frac{\text{median}_i \, |y_i - x_i^t \hat{\beta}_{\text{LMS}} - \hat{\alpha}_{\text{LMS}}|}{\text{SHORT}} \right)^2,$$

where SHORT stands for half of the length of the shortest interval covering half of the y_i observations.

Let us repeat the experiment described in Section 1 for the L_1 estimator, the Biweight S-estimator and the Least Median of Squares. Figure 3 shows the results: the L_1 estimator is much more robust than LS, and suffers more from leverage points than from vertical outliers. The LMS and Biweight S-estimator show a very robust behavior: there is only a slight loss in R^2_S, becoming constant when the outlier moves further away from the origin. The curve for LMS is quite erratic when the added point is close to the fitted regression line, but for R^2_{BS} we obtain a very smooth curve. In the next section we will generalize and formalize these findings by computing the associated influence functions.

3 Influence functions

Let S be the scale functional associated with S_n. We require that S is Fisher consistent at the error model distribution, meaning that $S(F_\sigma) = \sigma$. The scale functional representing M-estimators of scale is defined as the solution of

$$\int \rho(\frac{e}{S(F)}) dF(e) = b,$$

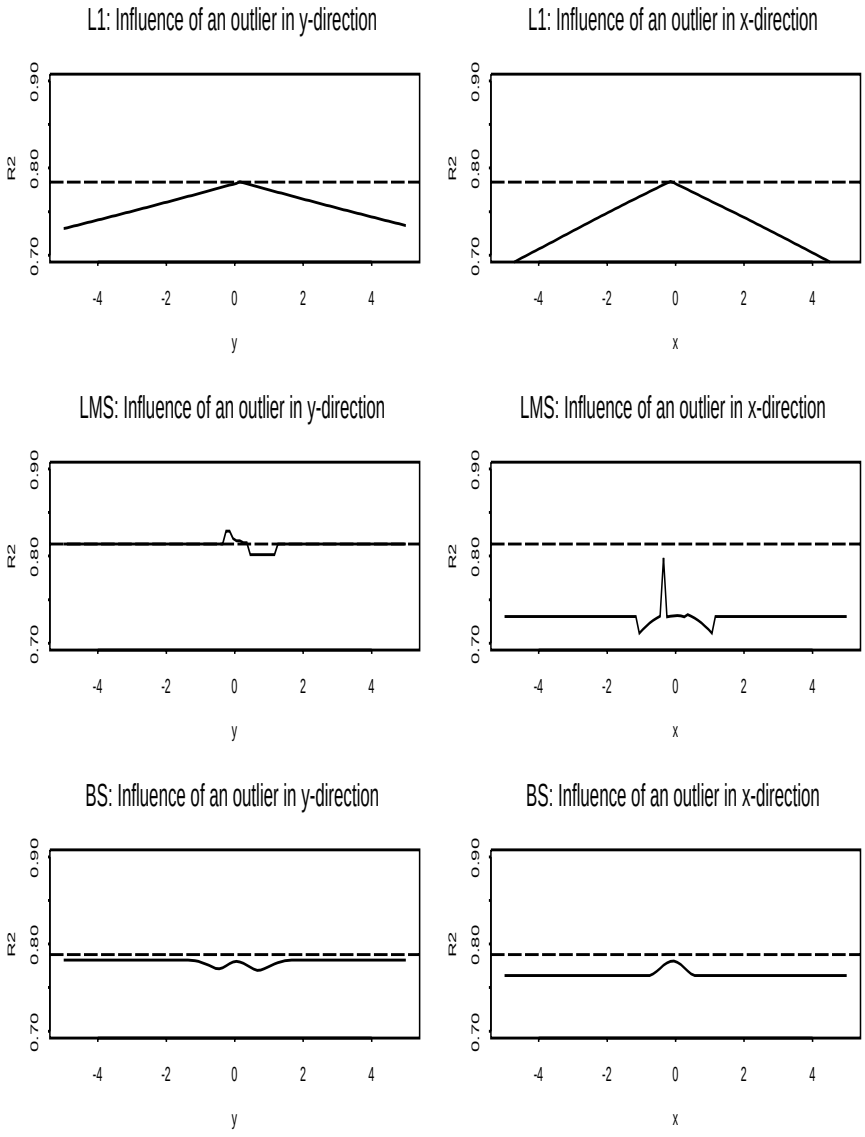

Figure 3: Effect of adding one observation $(0, y)$ *to a data cloud on the value of* R^2_{L1}, R^2_{BS} *and* R^2_{LMS} *(left figures) and effect of adding one observation* $(x, 0)$ *(right figures), as in Figure 2. The horizontal dashed lines indicate the value of* R^2_{L1}, R^2_{BS} *and* R^2_{LMS} *at the original sample.*

for any univariate distribution F. We need to set $b = E_{F_0}[\rho(Y)]$ to obtain Fisher consistency at the error model distribution. Furthermore, we will denote $S(Z) \equiv S(F)$ for any random variable $Z \sim F$.

In the regression context we have a stochastic variable (X, Y) coming from a $(p+1)$-dimensional distribution H. The regression functional corresponding to (2.1) is given by

$$(a(H), b(H)) = \underset{\alpha,\beta}{\operatorname{argmin}}\, S(Y - \beta^t X - \alpha).$$

The dispersion of the residuals is measured by the regression scale functional

$$S_1(H) = S(Y - b(H)^t X - a(H)).$$

The functional estimating the intercept parameter in (1.3) equals

$$a_0(H) = \underset{\alpha_0}{\operatorname{argmin}}\, S(Y - \alpha_0)$$

with associated scale

$$S_0(H) = S(Y - a_0(H)).$$

Finally, we obtain as functional representation for the R_S^2 statistic

$$R_S^2(H) = 1 - \frac{S_1^2(H)}{S_0^2(H)} = 1 - \frac{S^2(Y - b(H)^t X - a(H))}{S^2(Y - a_0(H))}. \tag{3.1}$$

If the denominator in the above expression equals zero, then by convention $R_S^2(H) = 1$.

The R_S^2 measure has a nice invariance property. We will write $R_S^2(X, Y) = R_S^2(H)$ whenever $(X, Y) \sim H$. (All proofs are in the Appendix.)

Proposition 1. *Let A be a non-singular $p \times p$ matrix , b a p-dimensional vector, and $\lambda \neq 0$ and γ two scalars. The R_S^2-functional defined in (3.1) verifies*

$$R_S^2(AX + b, \lambda Y + \gamma) = R_S^2(X, Y).$$

Recall that we supposed that the good data are generated by the model $Y = \alpha + \beta^t X + \varepsilon$, with ε independent of X, and where the errors followed a distribution $F_\sigma(x) = F_0(x/\sigma)$ with F_0 satisfying (F). We require consistency of all involved regression functionals, so that $a(H) = \alpha$ and $b(H) = \beta$, yielding $S_1(H) = S(\varepsilon) = \sigma$ and

$$R_S^2(H) = 1 - \frac{\sigma^2}{S_0^2(Y)}.$$

The above expression yields another value of $R_S^2(H)$ for every other scale measure. To make a comparison between approaches based on different scale measures easier, we will add an extra condition:

(H) The distribution of the response variable is a location-scale transformation of the error-term model distribution, so $P_H(Y \leq y) = F_0(\frac{y-\mu_Y}{\sigma_Y})$.

Of course, if we use R_S^2 as a descriptive measure, then we do not need to have (H). We only use it to simplify the theoretical setting. The advantage is now that

$$R_S^2(H) = 1 - \frac{\sigma^2}{\sigma_Y^2} := \wp^2 \tag{3.2}$$

yields a parameter independent of S, denoted by $\wp^2$, and called the squared population multiple correlation coefficient.

The standard example is H equal to a multivariate normal distribution. Take $X \sim N_p(\mu_X, \Sigma_{XX})$, $Y \sim N(\mu_Y, \sigma_Y^2)$ and $\mathrm{Cov}(X, Y) = \Sigma_{XY}$. We can write $y = \alpha + \beta^t x + \varepsilon$ where $\beta = \Sigma_{XX}^{-1}\Sigma_{XY}$, $\alpha = \mu_Y - \beta^t \mu_X$ and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Easy calculus (cfr. Johnson & Wichern 1998, p. 429) shows that

$$R_S^2(H) = \wp^2 = \frac{\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}}{\sigma_Y^2}. \tag{3.3}$$

Now we are ready to compute the influence function of the R_S^2 functional. Let $(x, y) \in I\!R^p \times I\!R$ be an individual observation, which can be interpreted as a contaminating point. The influence function of the R_S^2 functional at the model distribution H is defined as

$$IF((x, y), R_S^2, H) = \lim_{\varepsilon \downarrow 0} \frac{R_S^2((1-\varepsilon)H + \varepsilon\Delta_{(x,y)}) - R_S^2(H)}{\varepsilon}$$

where $\Delta_{(x,y)}$ is a distribution which puts all its mass at (x, y). The influence function measures the effect of a possible outlier (x, y) at the R_S^2 statistic. It gives the amount of change in the R_S^2 estimator caused by an infinitesimal amount of contamination. For more details about influence functions, see Hampel, Ronchetti, Rousseeuw, & Stahel (1986). We first derive the influence function using any scale functional S verifying a certain smoothness condition.

Theorem 1. *Let the model distribution H verify (H). Take $(X, Y) \sim H$ and denote ε the error term of the model. Assume that S has the property that $(a, b) \longrightarrow S(\varepsilon + b^t X + a)$ is differentiable with partial derivatives equal to zero at the origin $(0, 0)$. Then*

$$IF((x, y), R_S^2, H) = 2(1 - \wp^2)\left(IF(\frac{y - \mu_Y}{\sigma_Y}, S, F_0) - IF(\frac{y - \beta^t x - \alpha}{\sigma}, S, F_0)\right). \tag{3.4}$$

The influence function of the R_S^2 functional using M-scale estimators follows now from Theorem 1 by checking the condition on S and by using the well-known expression of the influence function of an M-estimator of scale.

Proposition 2. *Let $(X, Y) \sim H$ where the distribution H verifies (H). Then*

$$IF((x,y), R_S^2, H) = \frac{2(1-\wp^2)}{E_{F_0}[\rho'(\varepsilon)\varepsilon]} \left(\rho(\frac{y-\mu_Y}{\sigma_Y}) - \rho(\frac{y-\beta^t x-\alpha}{\sigma}) \right).$$

It readily follows that the influence function is bounded if the ρ-function is bounded. In Figure 4 we pictured some influence functions for a bivariate normal distribution H, with associated regression parameters $\alpha = 0$ and $\beta = 1$. We see that R_{LS}^2 is non-robust, since it has an unbounded influence function. (This was already noticed by Romanazzi 1992). Along the regression line $y = \alpha + \beta x$, where the good leverage points are, the influence grows the fastest to infinity. For L_1 we have an unbounded influence for R_{L1}^2, but only in the x direction. Indeed, it is known that L_1 is robust with respect to vertical outliers, but breaks down in presence of large leverage points. Good leverage points may let the influence on the R^2-statistic tend to infinity, while bad leverage points (with huge x values and far away from the regression line) can have an unbounded negative influence on the multiple correlation coefficient for the L_1 estimator. The influence function for R_{LMS}^2 is bounded and discontinuous, see Figure 4. Again, good leverage points increase the value of R_{LMS}^2, but only to a certain extent. Note that in a large zone outliers have zero influence, even when they are bad leverage points. This is because they have a similar influence on the spread of y as on the spread of the residuals. Bad leverage points with an outlying y value are harmless, and give a zero value for $IF((x,y), R_{\mathrm{LMS}}^2, H)$. On the other hand, bad leverage points with non outlying y values have a negative, but bounded influence. For the Biweight S-estimator we obtain an influence function which looks like a smoothed version of the LMS result, and has the same characteristics.

The gross-error sensitivity of the R_S^2-functional is defined as the maximal influence that an observation can have. Consider the following two extreme cases[1], where S is an M-estimator of scale.

1. The observation (x, y) follows perfectly the regression line $y = x^t\beta + \alpha$. If we let x

[1]Take $\beta \neq 0$, otherwise $IF \equiv 0$.

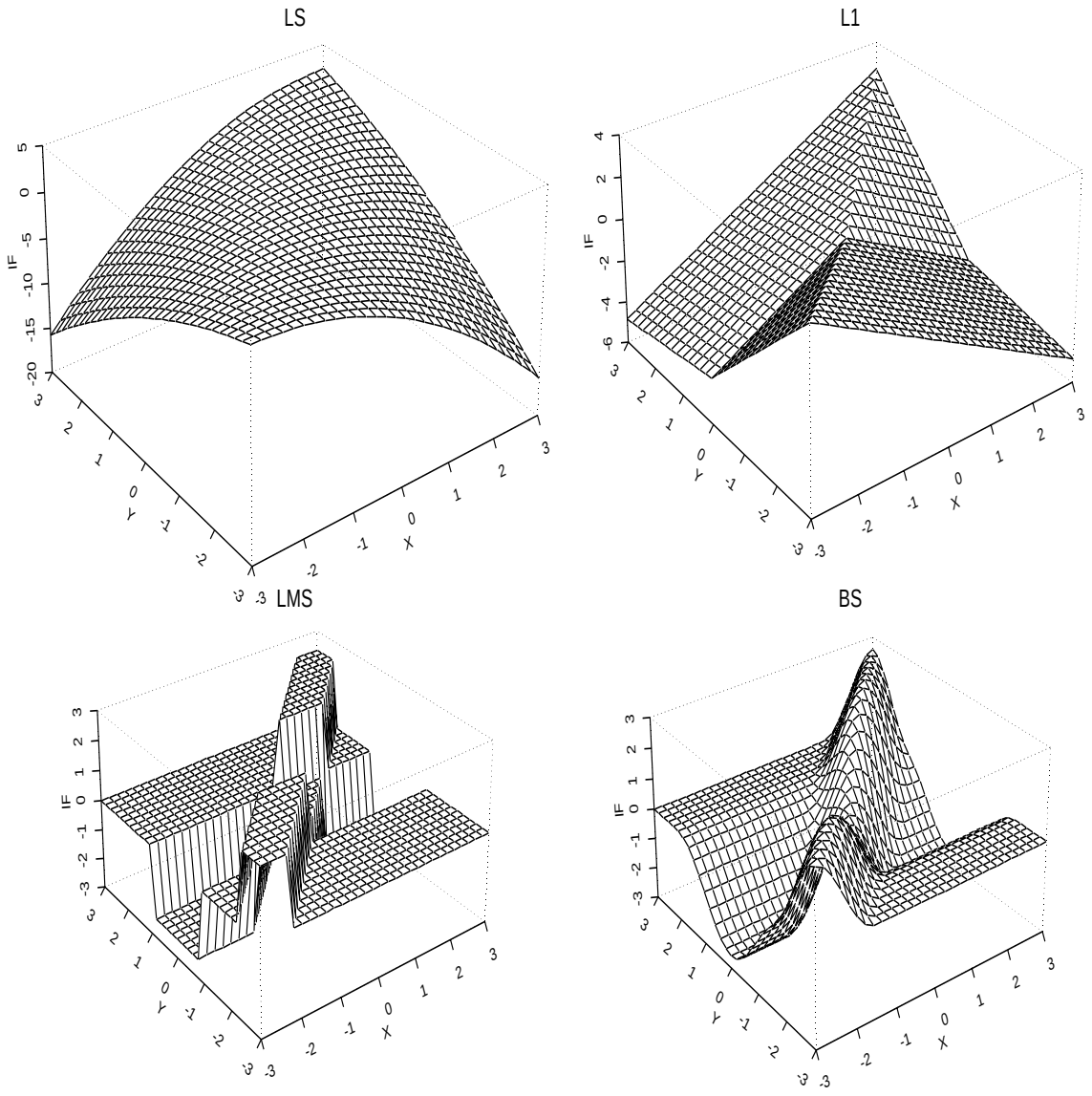

Figure 4: Plots of $IF((x,y),R_S^2,H)$ in the classical case, for the L_1 estimator, for the LMS and for the Biweight S-estimator. The distribution H is bivariate normal with $\wp^2 = 0.5$.

tend to infinity, then the influence on the R_S^2 functional will become maximal

$$\gamma^+ = \sup_{(x,y)} IF((x,y), R_S^2, H) = \frac{2(1-\wp^2)}{E_{F_0}[\rho'(\varepsilon)\varepsilon]}\rho(\infty).$$

We conclude that good leverage points have maximal and positive influence on the multiple correlation coefficient.

2. Take $y = \mu_Y$ and let x tend to infinity, then we reach

$$\gamma^- = \inf_{(x,y)} IF((x,y), R_S^2, H) = \frac{-2(1-\wp^2)}{E_{F_0}[\rho'(\varepsilon)\varepsilon]}\rho(\infty).$$

So well chosen bad leverage points will decrease the R_S^2. Note that not all bad leverage points will decrease the value of the multiple correlation coefficient, as can be seen from Figure 4.

We conclude that the gross-error sensitivity of R_S^2 is given by

$$\gamma^*(R_S^2, H) = \max(|\gamma^+|, |\gamma^-|) = \frac{2(1-\wp^2)}{E_{F_0}[\rho'(\varepsilon)\varepsilon]}\rho(\infty).$$

Note that $\gamma^*(R_S^2, H)$ coincides (except by a constant factor) with the gross-error sensitivity for the class of M-estimators with general scale as defined in Yohai & Zamar (1997). The next theorem tells us which ρ-function yields the lowest gross-error sensitivity. We say that it is Most B-Robust (cfr. Hampel, Ronchetti, Rousseeuw & Stahel 1986, p. 133) within the class of R_S^2 measures based on an M-estimator of scale. It is an immediate consequence of Theorem 3.2 of Yohai and Zamar (1997).

Theorem 2. *Denote $\Lambda(x) = -f_0'(x)/f_0(x)$ the score function associated with the distribution F_0 satisfying (F). Suppose that $x\Lambda(x) - 1$ has a unique positive root c^*. Among all ρ-function satisfying (R), the function*

$$\rho(y) = I(|\, y| > c^*)$$

minimizes $\gamma^*(R_S^2, H)$.

As we have seen in Section 2, this ρ-function is associated with a Least Quantile of Squares estimator. For a Gaussian error distribution, $c^* = 1$ and the optimal quantile equals $2\Phi(1) - 1 = 0.68$, corresponding with a gross-error sensitivity of $(1-\wp^2)\sqrt{2\pi e}$. It is

somehow surprising that the LMS is less locally robust than an LQS-estimator with quantile 68%. (This optimal quantile of 68% has also been obtained by Yohai & Zamar (1997) but in a different context). We will call this most B-Robust estimator the LQS32, since it has a breakdown point of 32%. In the next table, we give the values of the gross-error sensitivity relative to its minimum value

$$\frac{\gamma^*(R_S^2, H)}{(1-\wp^2)\sqrt{2\pi e}} \tag{3.5}$$

for different scale measures S and H a multivariate normal distribution.

Table 1: Relative gross-error sensitivities (3.5) for several estimators.

	LS	L1	LMS	BS	BS25	LQS32	LTS
relative gross-error sensitivity	∞	∞	1.129	1.243	1.387	1	1.543

We see that the LMS is close to the minimal value of $\gamma^*(R_S^2, H)$. The Biweight S-estimator with 50% breakdown point is neither losing much gross-error sensitivity w.r.t. the optimal procedure, but the loss increase if the breakdown value decreases to 25%.

4 Confidence intervals for $\wp^2$

Thanks to the preceding results we can construct a confidence interval for $\wp$, the population parameter of multiple correlation. Herefore we use the influence function as an heuristic tool to compute the asymptotic variance: if the functional R_S^2 is "sufficiently regular" then

$$\sqrt{n}\left(R_S^2 - \wp^2\right) \xrightarrow{d} N(0, ASV(R_S^2)),$$

where

$$ASV(R_S^2) = E_H[IF((X,Y), R_S^2, H)^2]$$

(Hampel, Ronchetti, Rousseeuw & Stahel 1986, p. 226). We require in this Section that H is a multivariate normal distribution, and Φ will denote the cumulative distribution function of a univariate standard normal.

For the Least Squares procedure, Theorem 1 gives $ASV(R_{\mathrm{LS}}^2) = 4\wp^2(1-\wp^2)^2$ for $0 < \wp < 1$. Since this asymptotic variance depends on $\wp$ it is customary to apply the Fisher transformation, yielding

$$\sqrt{n}\left(\tanh^{-1}(R_{\mathrm{LS}}) - \tanh^{-1}(\wp)\right) \xrightarrow{d} N(0,1),$$

and as 95% confidence interval for $\tanh^{-1}(\wp)$

$$\tanh^{-1}(R_{\mathrm{LS}}) \pm \frac{1.96}{\sqrt{n}}.$$

In the classical case even exact finite-sample distributional results exist, see already Fisher (1928).

For the robust case, $ASV(R_S^2)$ depends in a complicated way on $\wp$. Also here we apply the Fisher transformation, hoping that it will stabilize the variance:

$$\sqrt{n}(\tanh^{-1}(R_S) - \tanh^{-1}(\wp)) \xrightarrow{d} N(0, \frac{A(\wp)}{\wp^2(E_\Phi[\rho'(\varepsilon)\varepsilon])^2}),$$

where $A(\wp) = E[\{\rho(V\wp + U\sqrt{1-\wp^2}) - \rho(U)\}^2]$ with U and V two independent standard normal variables. In practice, we will use as 95% confidence interval for $\tanh^{-1}(\wp)$:

$$\tanh^{-1}(R_S) \pm 1.96\frac{W(R_S)}{\sqrt{n}} \tag{4.1}$$

with

$$W(R_S) = \sqrt{\frac{A(R_S)}{R_S^2 E_\Phi[\rho'(\varepsilon)\varepsilon]^2}}. \tag{4.2}$$

The quantity $A(R_S)$, and hence $W(R_S)$, can easily be computed using numerical integration. In Figure 5, $W(R_S)$ is plotted for a grid of R_S values and for the LS, L_1, BS, and LMS based method. It is constant and equal to 1 for LS. For L_1 and BS, we obtain slowly decreasing lines, showing that we get an almost constant variance for the Fisher tranformed estimators of multiple correlation. For LMS the Fisher tranformation is not succeeding in stabilizing the asymptotic variance, and very large confidence intervals are obtained for smaller values of $\wp$. Figure 5 cleary shows the differences in efficiency between the different approaches: the width of the confidence intervals for L_1 are quite close to those of LS; for the BS estimator we already lose quite some efficiency, while the inefficiency of the LMS procedure is almost prohibitive[2]. In order to obtain smaller confidence intervals while not losing too much robustness, the BS25 could be used.

By fitting curves to the points in Figure 5, easy to implement and extremely accurate approximations of $W(R_S)$ have been obtained. They are reported in Table 2 (also for BS25

[2]It is well known that the LMS regression estimator has a convergence rate slower that $n^{1/2}$, and can therefore be said to have a zero efficiency. But the corresponding estimator of multiple correlation will still be asymptotically normal for $\wp > 0$.

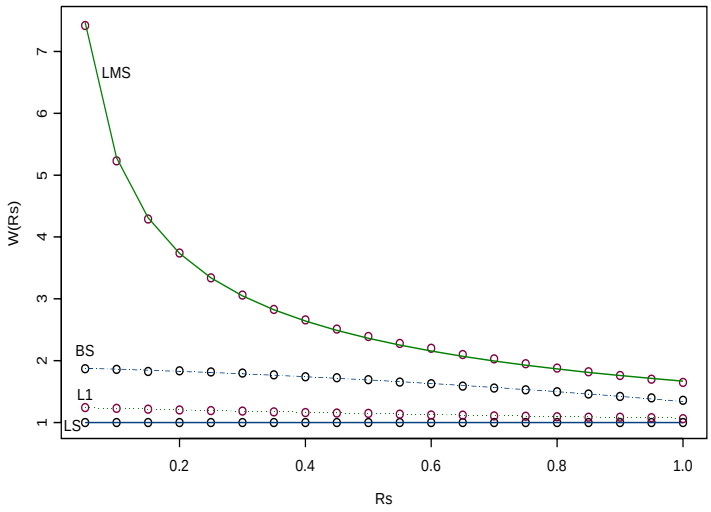

Figure 5: $W(R_S)$, determining the width of the confidence interval for $\tanh^{-1}(\wp)$*, as a function of R_S, for the LS, L_1, Biweight S, and LMS based procedures.*

and LQS32) and pictured in Figure 5. Finally we recall that the formulas for $W(R_S)$ have been derived under the assumption of multivariate normality, which is a quite restrictive condition.

5 Conclusion

The classical coefficient of determination is very sensitive to outliers. In this paper we studied the R_S^2 statistic associated with robust regression estimators based on minimizing estimators of scale. By computing influence functions we saw to which extent observations

Table 2: Approximations to $W(R_S)$ defined in (4.2) for several regression estimators.

LS	$W(R_S) = 1$
L_1	$W(R_S) \approx 1.25 - 0.24R_S + 0.06R_S^2$
BS	$W(R_S) \approx 1.89 - 0.26R_S - 0.29R_S^2$
BS25	$W(R_S) \approx 1.15 - 0.01R_S - 0.07R_S^2$
LMS	$W(R_S) \approx 1.67/\sqrt{R_S}$
LQS32	$W(R_S) \approx 1.29/\sqrt{R_S}$
LTS	$W(R_S) \approx 3.78 - 2.51R_S + 0.50R_S^2$

can increase or decrease R_S^2. We also found the estimator of squared multiple correlation having the lowest gross-error sensitivity for the class of M-estimators of scale. Easy rules to construct confidence intervals around the R_S^2-coefficients were obtained in case of multivariate normality.

Of course, there are some limitations to our study. Focus in the paper was on regression estimators based on the minimization of M-estimators. But Theorem 1 could also be applicable to the class of τ-estimators (Yohai & Zamar 1988), which are based on minimizing a τ-estimator of scale. The measure (2.2) can also be computed for the *Least Trimmed Squares* or LTS estimator of Rousseeuw (1984). For this estimator we verified that Proposition 2 applies by putting $\rho(u) = \max(u^2, c)$ for the appropriate constant c. The relative gross-error sensitivity at Gaussian distributions of LTS equals 1.543, which is rather large compared to the other estimators we considered. Confidence intervals can be constructed using formula (4.1), with $W(R_{LTS})$ given in Table 2.

Besides that, there exist many regression estimator not defined by minimizing a scale measure. As an example we take the MM-estimator (Yohai 1987), defined as the minimizer of

$$\sum_{i=1}^{n} \rho(\frac{y_i - \beta^t x_i - \alpha}{\hat{\sigma}^0})$$

with $\hat{\sigma}^0$ an inital high breakdown estimate of residual scale. An accompanying R^2-statistic is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \rho(\frac{y_i - \hat{\beta}^t x_i - \hat{\alpha}}{\hat{\sigma}^0})}{\sum_{i=1}^{n} \rho(\frac{y_i - \hat{\mu}_Y}{\hat{\sigma}^0})}.$$

and implemented in S-plus 2000 (1999, p. 285). The above statistic is surely a measure of explained variation of the response variable, but it is not an estimator of the multiple correlation coefficient. If the data come from a multivariate normal, then this coefficient tends to

$$1 - \frac{E_\Phi[\rho(\varepsilon)]}{E_\Phi[\rho(\varepsilon/\sqrt{1-\wp^2})]}$$

which only equals $\wp^2$ in case of a quadratic loss-function. Depending upon the value of $\wp^2$, there may be quite a bias (also when no outliers are present) when using this approach.

The R_S^2-statistic could be used to test the model (1.1) against the null model (1.3). The associated F-test is given by

$$F = \frac{R_S^2/(n-p)}{(1-R_S^2)/(p-1)}.$$

Properties and finite-sample behavior of the above test statistic are studied in Croux & Dehon (2001). Note that many proposals for robust tests in the linear model have already been made in the literature (e.g. Markatou & He 1994), but the estimation of the multiple correlation coefficient received much less attention.

Multiple correlation analysis could also be seen as special case of canonical correlation analysis, which has been robustified by Croux & Dehon (2001). This approach consists of estimating robustly the covariance matrices appearing in expression (3.3). These multiple correlation measures, however, are no longer interpretable as "the percentage of spread of y explained by x," and it is much less natural to use them as an R^2-measure in a regression context.

Acknowledgment We wish to thank the referee for useful remarks, and especially for suggesting equation (3.4) to us.

6 Appendix

Proof of Proposition 1: By (3.1) we may write

$$\begin{aligned} R_S^2(AX+b, \lambda Y+\gamma) &= 1 - \frac{\min_{\tilde{\alpha},\tilde{\beta}} S^2(\lambda Y+\gamma-(A^t\tilde{\beta})^t X - b^t\tilde{\beta}-\tilde{\alpha})}{\min_{\tilde{\alpha}_0} S^2(\lambda Y+\gamma-\tilde{\alpha}_0)} \\ &= 1 - \frac{\min_{\tilde{\alpha},\tilde{\beta}} S^2(Y-(\frac{A^t\tilde{\beta}}{\lambda})^t X - \frac{b^t\tilde{\beta}+\tilde{\alpha}-\gamma}{\lambda})}{\min_{\tilde{\alpha}_0} S^2(Y-\frac{\tilde{\alpha}_0-\gamma}{\lambda})} \\ &= 1 - \frac{\min_{\alpha,\beta} S^2(Y-\beta^t X-\alpha)}{\min_{\alpha_0} S^2(Y-\alpha_0)} \end{aligned}$$

where we set $\alpha = (\tilde{\alpha}-\gamma+b^t\tilde{\beta})/\lambda$, $\beta = A^t\tilde{\beta}/\lambda$, and $\alpha_0 = (\tilde{\alpha}_0-\gamma)/\lambda$. Noting that there is a one-to-one relationship between (α,β) and $(\tilde{\alpha},\tilde{\beta})$ finishes the proof. □

Proof of Theorem 1: By definition of the influence function and (3.1) we have for any $(x,y) \in I\!R^p \times I\!R$:

$$\begin{aligned} IF((x,y), R_S^2, H) &= \frac{\partial}{\partial\varepsilon}\left(1 - \frac{S_1^2(H_\varepsilon)}{S_0^2(H_\varepsilon)}\right)|_{\varepsilon=0} \\ &= \frac{2\sigma}{\sigma_Y^3}\left(\sigma\frac{\partial}{\partial\varepsilon}S_0(H_\varepsilon)\,|_{\varepsilon=0} - \sigma_Y\frac{\partial}{\partial\varepsilon}S_1(H_\varepsilon)\,|_{\varepsilon=0}\right), \end{aligned} \tag{6.1}$$

since $S_0(H) = \sigma_Y$ and $S_1(H) = \sigma$. Now we compute the derivative at $\varepsilon = 0$ of

$$S_1(H_\varepsilon) = S\left(B_\varepsilon(Y - b(H_\varepsilon)^t X - a(H_\varepsilon)) + (1-B_\varepsilon)(y - b(H_\varepsilon)x - a(H_\varepsilon))\right)$$

where B_ε is a Bernoulli variable being equal to 1 with probability $(1-\varepsilon)$ and zero otherwise. Denote $g_1(\varepsilon) = (\varepsilon, a(H_\varepsilon), b(H_\varepsilon))$ and

$$g_2(\varepsilon, a, b) = S\left(B_\varepsilon(Y - b^t X - a) + (1 - B_\varepsilon)(y - b^t x - a)\right)$$

for all $\varepsilon > 0, a \in I\!R, b \in I\!R^p$. Since $S_1(H_\varepsilon) = g_2(g_1(\varepsilon))$, the chain rules yields

$$\frac{\partial}{\partial\varepsilon} S_1(H_\varepsilon)|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon} g_2(\varepsilon, \alpha, \beta)|_{\varepsilon=0} + \frac{\partial}{\partial a} g_2(0, a, \beta)|_{a=\alpha} \frac{\partial}{\partial\varepsilon} a(H_\varepsilon)|_{\varepsilon=0} + \frac{\partial}{\partial b} g_2(0, \alpha, b)|_{b=\beta} \frac{\partial}{\partial\varepsilon} b(H_\varepsilon)|_{\varepsilon=0}, \tag{6.2}$$

where we used $a(H) = \alpha$ and $b(H) = \beta$. Since $g_2(0, a, \beta) = S(\varepsilon + (\alpha - a))$ and $g_2(0, \alpha, b) = S(\varepsilon + (\beta - b)^t X)$, where ε stands now for the error term, we have that the last two terms in (6.2) are equal to zero (herefore we use the condition on S). Equation (6.2) simplifies then to

$$\begin{aligned} \frac{\partial}{\partial\varepsilon} S_1(H_\varepsilon)|_{\varepsilon=0} &= \frac{\partial}{\partial\varepsilon} S((1-\varepsilon)F_\sigma + \varepsilon\Delta_{y-\beta^t x-\alpha})|_{\varepsilon=0} \\ &= \sigma \frac{\partial}{\partial\varepsilon} S((1-\varepsilon)F_0 + \varepsilon\Delta_{(y-\beta^t x-\alpha)/\sigma})|_{\varepsilon=0} \\ &= \sigma IF(\frac{y - \beta^t x - \alpha}{\sigma}, S, F_0) \end{aligned} \tag{6.3}$$

where F_σ is the error distribution, and by using scale equivariance of S.

In a similar way, we can show that

$$\frac{\partial}{\partial\varepsilon} S_0(H_\varepsilon)|_{\varepsilon=0} = \sigma_Y IF(\frac{y - \mu_Y}{\sigma_Y}, S, F_0). \tag{6.4}$$

Inserting (6.3) and (6.4) into (6.1) yields

$$IF((x, y), R_S^2, H) = \frac{2\sigma^2}{\sigma_Y^2} \left(IF(\frac{y - \mu_Y}{\sigma_Y}, S, F_0) - IF(\frac{y - \beta^t x - \alpha}{\sigma}, S, F_0) \right)$$

By (3.2) the results follows. □

Proof of Proposition 2: We will show that S satisfies the condition of Theorem 1. The result follows then immediately since it is well known that

$$IF(x, S, F_0) = \frac{\rho(x) - E_{F_0}[\rho(x)]}{E_{F_0}[\rho'(x)x]}$$

for M-estimators of scale (e.g. Huber 1981, page 109). Take $a \in I\!R$ and $b \in I\!R^p$, then we have that $s(a, b) = S(\varepsilon + b^t X + a)$ verifies

$$\int \rho(\frac{\varepsilon + b^t x + a}{s(a, b)}) dF_\sigma(\varepsilon) dG(x) = \text{constant}, \tag{6.5}$$

where $X \sim G$ and $\varepsilon \sim F_\sigma$ is independent of X. Differentiating both sides of (6.5) with respect to b and evaluating at $(a, b) = (0, 0)$ yields

$$\int \rho'(\frac{\varepsilon}{\sigma})(\frac{x}{\sigma} - \frac{\varepsilon}{\sigma^2}\frac{\partial}{\partial b}s(a,b)|_{b=0})dF_\sigma(\varepsilon)dG(x) = 0, \tag{6.6}$$

since $\sigma = s(0, 0)$. By condition (F), $E_{F_\sigma}[\rho'(\frac{\varepsilon}{\sigma})] = 0$, and (6.6) reduces to

$$E_{F_\sigma}[\rho'(\frac{\varepsilon}{\sigma})\frac{\varepsilon}{\sigma}]\frac{\partial}{\partial b}s(a,b)|_{b=0} = 0. \tag{6.7}$$

By symmetry, $E_{F_0}[\rho'(\varepsilon)\varepsilon] > 0$, so that (6.7) implies $\frac{\partial}{\partial b}s(a,b)|_{b=0} = 0$. In a similar way, it can be shown that $\frac{\partial}{\partial a}s(a,b)|_{a=0} = 0$. □

7 References

Anderson-Sprecher, R. A. (1994), "Model Comparison and R^2", *The American Statistician*, 48, 113–117.

Croux, C., and Dehon, C. (2001), "Analyse canonique basée sur des estimateurs robustes de la matrice de covariance", *Revue de Statistique Appliquée*, to appear.

Croux, C., and Dehon, C. "The F-test for high breakdown robust regression", in preparation.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley and Sons.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley.

Johnson, R. A. and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, fourth edition, Prentice Hall International Editions.

Fisher, R.A. (1928), *Proceedings of the Royal Society London A* 121, 654-673.

Kvalseth, T. O. (1985), "Cautionary note about R^2", *The American Statistician*, 39, 279–285.

Markatou, M., and He, X. (1994), 'Bounded influence and high breakdown point testing procedures in linear models," *Journal of the American Statistical Association*, 89, 543–549.

McKean, J. W., and Sievers, G. L. (1987), "Coefficients of determination for least absolute deviation analysis," *Statistics and Probability Letters*, 5 49–54.

Romanazzi, M. (1992), "Influence in canonical correlation analysis", *Psychometrika*, 57, 237–259.

Rousseeuw, P. J. (1984), "Least Median of Squares regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.

Rousseeuw, P. J., and Yohai, V. J. (1984), Robust regression by means of S-estimators. In: Franke, J., Härdle W., Martin, R.D. (Eds.), Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics 26, New York: Springer Verlag.

S-Plus 2000 Guide to Statistics, Volume 1, (1999), Data Analysis Products Division, Mathsoft, Seatlle, WA.

Yohai, V. J. (1987), "High breakdown point and high efficiency robust estimates for regression", *The Annals of Statistics,* 15, 642-656.

Yohai, V. J., and Zamar, R. H. (1988), "High breakdown-point estimates of regression by means of the minimization of an efficient scale," *Journal of the American Statistical Association*, 83 406–413.

Yohai, V. J., and Zamar, R. H. (1997), "Optimal locally robust M-estimates of regression," *Journal of Statistical Planning and Inference,* 64, 2, 309–323.

Willet, J. B., and Singer, J. D. (1988), "Another cautionary note about R^2 : its use in weighted least squares regression analysis," *The American Statistician*, 42, 236–238.