ELSEVIER

Statistics & Probability Letters 20 (1994) 295–306

STATISTICS & PROBABILITY LETTERS

Robust measures of association in the correlation model

Lee D. Witt[a], Joseph W. McKean[b,*], Joshua D. Naranjo[b]

[a]*Davenport College, Devenport, IA, USA*
[b]*Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, MI 49008-5152, USA*

Received June 1993; revised September 1993

Abstract

In the correlation model, the classical coefficient of multiple determination $\bar{R}^2$ is a measure of association between the dependent random variable Y and the random vector of independent variables $\boldsymbol{x}$. Slight departures from normality, however, can have a pronounced effect on the measure. In the regression model robust estimates of the regression coefficients are less sensitive to outlying points than least squares estimates. These estimates are often obtained by minimizing an objective (dispersion) function. For such robust estimates, the proportion of explained dispersion is a natural analogue to the statistic R^2. Although this statistic is generally not robust, it leads to robust statistics which are consistent estimates of functionals in the correlation model. These functionals are robust measures of association between Y and $\boldsymbol{x}$. They are efficiently robust and have bounded influence provided the robust estimator on which they are based has bounded influence.

Keywords: Bounded influence; Coefficients of multiple determination; M-Estimates; R^2; R-Estimates; Robust

1. Introduction

In the correlation model, the classical coefficient of multiple determination $\bar{R}^2$ is a measure of association between the dependent random variable Y and the random vector of independent variables $\boldsymbol{x}$. This measure is between 0 and 1 and is 0 if and only if Y and $\boldsymbol{x}$ are independent. Under the multivariate normal model, the maximum likelihood estimate of $\bar{R}^2$ is R^2, the proportion of explained variation in the least squares fit of the regression model. The measure $\bar{R}^2$ is not robust. Slight departures form normality can have a pronounced effect on the measure.

In the regression model, sensitivity of the least-squares estimates of regression coefficients to outliers can be overcome by using robust estimators of the coefficients in place of least-squares estimators. Consider robust estimates which are obtained by minimizing an objective function, i.e., by minimizing a dispersion of the residuals which is the analogue of minimizing the sums of squares of the residuals for least-squares fits. As we propose in Section 3, the natural analogue of R^2 for such a robust fit is R_1 the proportion of dispersion

*Corresponding author.

SSDI 0167-7152(93)E0191-U

accounted for. Although, R_1 is generally not robust, it leads to an efficiently robust statistic which we label as R_2. If R_2 is based on bounded influence estimates then it also has bounded influence.

Under the correlation model R_1 and R_2 are consistent estimates of measures of association, $\bar{R}_1$ and $\bar{R}_2$, respectively, between Y and $\boldsymbol{x}$. The properties of these measures are presented in Sections 3 and 4. As with $\bar{R}^2$ these measures are between 0 and 1 and they are 0 if and only if Y and $\boldsymbol{x}$ are independent. In Section 5, we illustrate the robustness properties of these measures over families of contaminated normal distributions.

McKean and Sievers (1987) presented similar measures for fits based on least absolute deviations. Our presentation here, though, is for any robust estimating procedure which is based on the minimization of an objective function. It includes regular robust estimates and bounded influence estimates. Much of our discussion is presented in terms of the class of R-estimates because for these estimates simple closed formed expressions for these measures can often be obtained. In fact, they are one-to-one functions of $\bar{R}^2$ at the multivariate normal model. But our discussion is completely general for any robust estimator, as we show in Section 6 for the class of M-estimates.

2. Notation

We are concerned with the linear model defined by

$$Y = \alpha + \boldsymbol{x}'\beta + e, \tag{2.1}$$

where $\boldsymbol{x}$ is a p-dimensional random vector with distribution function M and density function m, e is a random variable with distribution function F and density f, and $\boldsymbol{x}$ and e are independent. Let H and h denote the joint distribution function and joint density of Y and $\boldsymbol{x}$. It follows that

$$h(\boldsymbol{x}, y) = f(y - \alpha - \boldsymbol{x}'\beta)m(\boldsymbol{x}). \tag{2.2}$$

Denote the distribution and density function of Y by G and g. Following Arnold (1981), we will call this the correlation model. The hypotheses of interest are:

$$H_0\text{: } Y \text{ and } \boldsymbol{x} \text{ are independent versus } H_A\text{: } Y \text{ and } \boldsymbol{x} \text{ are dependent.} \tag{2.3}$$

By (2.2) this is equivalent to the hypotheses H_0: $\beta = 0$ versus H_A: $\beta \neq 0$. For the theory below, we make the following assumptions:

(F.1) f is absolutely continuous and $f > 0$ a.e. Lebesgue,

(F.2) $\mathrm{Var}(e) = \sigma^2 < \infty$,

(M.1) $E[\boldsymbol{x}\boldsymbol{x}'] = \Sigma$, $\Sigma > 0$.

Without loss of generality assume that $\mathrm{E}[\boldsymbol{x}] = 0$ and $E(e) = 0$. We are interested in measures of association between Y and $\boldsymbol{x}$ and estimates of such measures.

Let $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$ be a random sample from the above model. Define the $n \times p$ matrix $\boldsymbol{X}$ to be the matrix whose ith row is the vector $\boldsymbol{x}_i$. Let $\boldsymbol{X}_c = (\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{X}$ where $\boldsymbol{H}_1$ denotes the projection matrix onto the space spanned by an $n \times 1$ vector of ones. Under (M.1), Arnold (1980) showed that $\lim_{n\to\infty} \max \mathrm{diag}\{\boldsymbol{X}_c(\boldsymbol{X}_c'\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c'\} \to 0$ with probability 1; that is, Huber's (1973) design condition holds with probability 1.

We motivate the classical and robust measures of association from the viewpoint of the regression model where $\boldsymbol{x}$ is treated as a nonstochastic variable. For this model $\boldsymbol{X}$ is the design matrix and $\boldsymbol{X}_c$ is the centered design matrix. Although the regression model supplies the motivation, all the results presented in this paper are under the correlation model (2.1).

2.1. Classical model and estimates

For motivation, we first consider the multivariate normal model. The classical population coefficient of multiple determination (CMD) is defined by

$$\bar{R}^2 = \frac{\beta' \Sigma \beta}{\sigma^2 + \beta' \Sigma \beta}; \tag{2.4}$$

see Arnold (1981). Note that $\bar{R}^2$ is a measure of association between Y and $\boldsymbol{x}$. It lies between 0 and 1 and it is 0 if and only if Y and $\boldsymbol{x}$ are independent.

Under the multivariate normal model, the maximum likelihood estimate of $\bar{R}^2$ can be obtained in the following way. Treat $\boldsymbol{x}_i$ as nonstochastic and fit by least squares the model $Y_i = \alpha + \boldsymbol{x}_i'\beta + e_i$, which will be called the full model. The residual amount of variation is $SSE = \sum_{i=1}^n (Y_i - \hat{\alpha}_{LS} - \boldsymbol{x}_i'\hat{\beta}_{LS})^2$, where $\hat{\beta}_{LS}$ and $\hat{\alpha}_{LS}$ are the least-squares estimates. Next fit the reduced model defined as the full model subject to H_0: $\beta = 0$. The residual amount of variation is $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$. The reduction in variation in fitting the full model over the reduced model is $SSR = SST - SSE$. The maximum likelihood estimate of $\bar{R}^2$ is the proportion of explained variation, i.e.,

$$R^2 = \frac{SSR}{SST}. \tag{2.5}$$

The least-squares test statistic for H_0 versus H_A is $F_{LS} = (SSR/p)/\hat{\sigma}^2_{LS}$ where $\hat{\sigma}^2_{LS} = SSE/(n - p - 1)$. Recall that R^2 can be expressed as

$$R^2 = \frac{SSR}{SSR + (n - p - 1)\hat{\sigma}^2}. \tag{2.6}$$

Under (F.2) and (M.1) and not necessarily multivariate normality, R^2 is a consistent estimate of $\bar{R}^2$; $\sqrt{n}(\hat{\beta}_{LS} - \beta) \xrightarrow{\mathcal{D}} N_p(0, \sigma^2 \Sigma^{-1})$; and under the sequence of contiguous alternatives, H_n: $\beta = \theta/\sqrt{n}$ where $\theta \neq 0$, $pF_{LS} \xrightarrow{\mathcal{D}} \chi^2(p, \delta_{LS})$ with noncentrality parameter $\delta_{LS} = \theta' \Sigma \theta/\sigma^2$. See Arnold (1980) for details.

3. Robust analysis of correlation model

To motivate our measures of association for the R-analysis, we propose analogues of R^2 based on R-fits of the linear model which in turn lead us to the population coefficients of determination. We present coefficients associated with regular R-estimates and bounded influence R-estimates. Further properties of these coefficients are derived in Section 4.

3.1. R-estimates

The objective function for regular R-estimates is the dispersion function proposed by Jaeckel (1972),

$$D(\beta) = \sum_{i=1}^n a(R(Y_i - \boldsymbol{x}_i'\beta))(Y_i - \boldsymbol{x}_i'\beta), \tag{3.1}$$

where $R(u_i)$ denotes the rank of u_i among $u_1, \ldots, u_n$ and $a(1) \leqslant \cdots \leqslant a(n)$ is a set of scores such that $\sum a(i) = 0$. We will consider scores generated as $a(i) = \varphi(i/(n + 1))$ where $\varphi(u)$ is a function defined on the interval (0, 1) and which satisfies the conditions

(S.1) φ is nondecreasing, $\int \varphi = 0$ and $\int \varphi^2 = 1$,

(S.2) φ is bounded.

Jaeckel showed that D is a nonnegative convex function of β. Further properties of D along with a k-step procedure to minimize it can be found in McKean and Hettmansperger (1978). A recent discussion on families of scores suitable for asymmetric as well as symmetric error distributions can be found in McKean and Sievers (1989). We will make use of the Wilcoxon score function which is given by the linear function $\varphi_W(u) = \sqrt{12}(u - \frac{1}{2})$.

For a specified score function φ the R-estimate of β is a value $\hat{\beta}_R$ which minimizes the dispersion function D. Under the correlation model (2.1), the influence function of $\hat{\beta}_R$ is $IF(\boldsymbol{x}, y;\ \hat{\beta}_R) = \tau\varphi(F(y))\,\Sigma^{-1}\boldsymbol{x}$ where τ is the scale parameter

$$\tau = \left[\int_0^1 \varphi(u)\left(-\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}\right)\mathrm{d}u\right]^{-1}, \tag{3.2}$$

see Hettmansperger (1984) and Witt (1989). Note that this influence function is bounded in the Y-space but not in the $\boldsymbol{x}$-space.

Under the correlation model (2.1) and the assumptions (F.1), (S.1) and (M.1) $\sqrt{n}(\hat{\beta}_R - \beta) \xrightarrow{\mathscr{D}} \mathrm{N}_p(0, \tau^2\Sigma^{-1})$. By the note in Section 2 concerning Huber's condition, this limiting distribution follows by using a conditional argument and Heiler and Willers's (1988) asymptotic theory for R-estimates in the nonstochastic case.

A consistent estimate of τ is given in Koul et al. (1987) for the non-stochastic case under the conditions (F.1), (S.1) and (S.2). Once again using the conditional argument it would be consistent for the correlation model.

3.2. *R-coefficients of multiple determination*

For the above dispersion function the drop in dispersion when passing from the reduced to the full model is given by

$$RD = D(0) - D(\hat{\beta}_R), \tag{3.3}$$

hence, the proportion of dispersion accounted for by fitting β is

$$R_1 = RD/D(0). \tag{3.4}$$

Note that R_1 is quite analogous to R^2 in that dispersion is substituted for variation. This is a natural CMD for any robust estimate and, as we shall show in Section 4, the population CMD for which R_1 is a consistent estimate does satisfy interesting properties. However, the influence function of the denominator is not bounded in the y-space (see Witt, 1989); hence the statistic R_1 is not robust. See also Section 5.

In order to obtain a CMD which is robust we first consider a robust test statistic for the hypothesis (2.3), given by

$$F_R = \frac{RD/p}{\hat{\tau}/2}, \tag{3.5}$$

which was proposed by McKean and Hettmansperger (1976) for regular R-estimates. Employing the usual conditional argument with the results in McKean and Hettmansperger (1976), it follows that $pF_R \xrightarrow{\mathscr{D}} \chi^2(p, \delta_R)$ *a.e. M* under H_n: $\beta = \theta/\sqrt{n}$ where the noncentrality parameter δ_R is given by $\delta = \theta'\Sigma\theta/\tau^2$. The ARE of the R-test F_R to the least-squares test F_{LS} is the ratio of noncentrality parameters, σ^2/τ^2. This is the usual ARE of rank tests to tests based on least squares in simple location models. Thus the test statistic F_R is efficiently robust.

Proceeding as in Hampel et al. (1986), Witt (1989) showed that the influence function of $\sqrt{pF_R}$ is given by $IF(\boldsymbol{x}, y, \sqrt{pF_R}) = |\varphi(F(y))|\sqrt{\boldsymbol{x}\Sigma^{-1}\boldsymbol{x}}$, which is bounded in the y-space. Hence the test statistic is bias robust.

Consider the relationship between the classical F-test and R^2 given by expression (2.6). In the same way but using the robust test F_R, we can define a second R-coefficient of multiple determination

$$R_2 = \frac{RD}{RD + (n - p - 1)(\hat{\tau}/2)}. \tag{3.6}$$

It follows from the above discussion on the R-test statistic that the influence function of R_2 has bounded influence in the Y-space. Note that like R^2, both R_1 and R_2 are invariant to affine transformations of the form $a + bY_i$ and $\boldsymbol{c} + \boldsymbol{D}\boldsymbol{x}_i$.

The functionals that, respectively, correspond to the statistics $D(0)$ and $D(\hat{\beta}_R)$ are $\overline{D}_y = \int \varphi(G(y))\, y \mathrm{d}G(y)$ and $\overline{D}_e = \int \varphi(F(e))e\, \mathrm{d}F(e)$. The population CMDs associated with R_1 and R_2 are:

$$\overline{R}_1 = \overline{RD}/\overline{D}_y, \tag{3.7}$$

$$\overline{R}_2 = \overline{RD}/(\overline{RD} + (\tau/2)), \tag{3.8}$$

where $\overline{RD} = \overline{D}_y - \overline{D}_e$. The properties of these parameters are discussed in the next section. The consistency of R_1 and R_2 is given in the following theorem:

Theorem 3.1. *Under the correlation model* (2.1) *and the assumptions* (*F*.1), (*S*.1), (*S*.2) *and* (*M*.1),

$$R_i \xrightarrow{P} \overline{R}_i \text{ a.e. } M, \quad i = 1, 2.$$

Proof. Conditionally given X, David (1970) showed that $(1/n)D(0) \xrightarrow{P} \overline{D}_y$ and McKean and Hettmansperger (1976) showed that $(1/n)D(\hat{\beta}_R) \xrightarrow{P} \overline{D}_e$. Hence these results hold unconditionally *a.e. M*. The consistency of $\hat{\tau}$ was discussed above. The result then follows.

For the Wilcoxon scores, $\varphi_W(u) = \sqrt{12}(u - 1/2)$, $\overline{D}_y = \sqrt{3/4}E|Y_1 - Y_2|$ where Y_1, Y_2 are i.i.d. with distribution function G. Likewise, $\overline{D}_e = \sqrt{3/4}E|e_1 - e_2|$ where e_1, e_2 are i.i.d. with distribution function F. Finally $\tau = (\sqrt{12}\int f^2)^{-1}$. Hence for Wilcoxon scores these coefficients of multiple determination simplify to

$$\overline{R}_{W1} = \frac{E|Y_1 - Y_2| - E|e_1 - e_2|}{E|Y_1 - Y_2|}, \tag{3.9}$$

$$\overline{R}_{W2} = \frac{E|Y_1 - Y_2| - E|e_1 - e_2|}{E|Y_1 - Y_2| - E|e_1 - e_2| + (1/6\int f^2)}. \tag{3.10}$$

3.3. Bounded influence R-estimates

For the Wilcoxon scores, the dispersion function (3.1) can be expressed as $D(\beta) = (\sqrt{3}/(n+1))$ $\sum_{i<j}|(Y_i - \boldsymbol{x}_i'\beta) - (Y_j - \boldsymbol{x}_j'\beta)|$. Sievers (1983) extended the above Wilcoxon estimates to a class of R-estimates which have bounded influence in both the Y- and the $\boldsymbol{x}$-spaces by using the objective function

$$D_B(\beta) = \frac{\sqrt{3}}{n+1}\sum_{i<j} b(\boldsymbol{x}_i)b(\boldsymbol{x}_j)|(Y_i - \boldsymbol{x}_i'\beta) - (Y_j - \boldsymbol{x}_j'\beta)|, \tag{3.11}$$

where the $b(\boldsymbol{x}_i)$'s are a set of specified weights. Let $\hat{\beta}_B$ denote a minimizing value of this objective function. Naranjo and Hettmansperger (1994) showed that the influence function of these estimates is

$IF(x, y; \hat{\beta}_B) = \tau\varphi_W(F(y))C^{-1}\int(x - w)b(x)b(w)\,dM(w)$, where φ_W denotes the Wilcoxon score function and C is the matrix

$$C = 2^{-1}\int\int(x_2 - x_1)(x_2 - x_1)'b(x_1)b(x_2)\,dM(x_1)\,dM(x_2).$$

For appropriate weights the influence function is bounded in both the Y- and x-spaces. Further $\sqrt{n}(\hat{\beta}_B - \beta) \xrightarrow{\mathcal{D}} N_p(0, \tau^2 C^{-1}VC^{-1})$, *a.e.* M, where

$$V = \int\left[\int(x_2 - x_1)b(x_1)b(x_2)\,dM(x_2)\right]\left[\int(x_2 - x_1)b(x_1)b(x_2)\,dM(x_2)\right]'dM(x_1).$$

Naranjo (1989) obtained a test of the hypothesis of independence (2.3) in terms of the reduction in the dispersion function (3.11). Its influence function is bounded in both the Y- and x-spaces. Sample coefficients of multiple determination R_{B1} and R_{B2} can be formulated for these bounded influence R-estimates in the same way as R_1 and R_2 were formulated for regular R-estimates. As in the case of regular R-estimates, the influence function of the denominator of R_{B1} will be unbounded in the Y-space; however, the influence function of R_{B2} will be bounded in both spaces. The functionals corresponding to $D_B(0)$ and $D_B(\hat{\beta}_R)$ for the bounded influence R-estimates are

$$\overline{D}_{By} = \frac{\sqrt{3}}{2}\int\int b(x_1)b(x_2)|y_1 - y_2|\,dH(x_1, y_1)\,dH(x_2, y_2), \tag{3.12}$$

$$\overline{D}_{Be} = \frac{\sqrt{3}}{2}\int b(x_1)b(x_2)\,dM(x_1)\,dM(x_2)E|e_1 - e_2|. \tag{3.13}$$

Let $\overline{RD}_B = \overline{D}_{By} - \overline{D}_{Be}$. In the same way as for regular R-estimates, we can define population coefficients of multiple determination $\overline{R}_{B1}$ and $\overline{R}_{B2}$ for the bounded influence estimates. Unlike the regular R-estimates, however, these parameters are not as readily interpretable.

4. Properties of robust coefficients of multiple determination

In this section we explore further properties of the population coefficients of multiple determination proposed in Section 3. We will first consider those coefficients that were formulated from consideration of regular R-estimates and then briefly summarize the results for those coefficients that were formulated under bounded influence estimates.

To show that $\overline{R}_1$ and $\overline{R}_2$ are indeed measures of association we have the following two theorems. The proof of the first theorem is quite similar to corresponding proofs of properties of the dispersion function for the nonstochastic model found in Jaeckel (1972).

Theorem 4.1. *Suppose f and g satisfy the conditions* (F.1) *and their first moments are finite then $\overline{D}_y > 0$ and $\overline{D}_e > 0$.*

Proof. It sufficies to show it for $\overline{D}_y$ since the proof for $\overline{D}_e$ is the same. The function φ is increasing and $\int\varphi = 0$; hence, φ must take on both negative and positive values. Thus the set $A = \{y: \varphi(G(y)) < 0\}$ is

not empty and is bounded above. Let $y_0 = \sup A$. Then

$$\overline{D}_y = \int_{-\infty}^{y_0} \varphi(G(y))(y - y_0)\mathrm{d}G(y) + \int_{y_0}^{\infty} \varphi(G(y))(y - y_0)\mathrm{d}G(y). \tag{4.1}$$

Since both integrands are nonnegative, it follows that $\overline{D}_y \geqslant 0$. If $\overline{D}_y = 0$ then it follows from (F.1) that $\varphi(G(y)) = 0$ for all $y \neq y_0$ which contradicts the facts that φ takes on both positive and negative values and that G is absolutely continuous.

Theorem 4.2. *Suppose f and g satisfy the conditions* (F.1) *and* (F.2) *and that φ^{-1} exists. Then $\overline{RD}$ is a strictly convex function of β and has a minimum value of* 0 *at $\beta = 0$.*

Proof. We will show that the first derivative of $\overline{RD}$ is zero at $\beta = 0$ and that its second derivative is positive definite. Note first that the distribution function, G, and density, g, of Y can be expressed as $G(y) = \int F(y - \beta' \boldsymbol{x})\mathrm{d}M(\boldsymbol{x})$ and $g(y) = \int f(y - \beta' \boldsymbol{x})\mathrm{d}M(\boldsymbol{x})$. We have

$$\begin{aligned}\frac{\partial \overline{RD}}{\partial \beta} = &- \iiint \varphi'[G(y)]\, y f(y - \beta' \boldsymbol{x}) f(y - \beta' \boldsymbol{u})\boldsymbol{u}\,\mathrm{d}M(\boldsymbol{x})\,\mathrm{d}M(\boldsymbol{u})\mathrm{d}y \\ &- \iint \varphi[G(y)]\, y f'(y - \beta' \boldsymbol{x})\boldsymbol{x}\mathrm{d}M(\boldsymbol{x})\mathrm{d}y.\end{aligned} \tag{4.2}$$

Since $E[\boldsymbol{x}] = \boldsymbol{0}$, both terms on the right-hand side of the above expression are $\boldsymbol{0}$ at $\beta = 0$. Before obtaining the second derivative, we rewrite the first term of (4.2) as

$$\begin{aligned}&- \int \left[\iint \varphi'[G(y)]\, y f(y - \beta' \boldsymbol{x}) f(y - \beta' \boldsymbol{u})\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{x}) \right] \boldsymbol{u}\,\mathrm{d}M(\boldsymbol{u}) \\ &\quad = - \int \left[\int \varphi'[G(y)]\, g(y) y f(y - \beta' \boldsymbol{u})\,\mathrm{d}y \right] \boldsymbol{u}\mathrm{d}M(\boldsymbol{u}).\end{aligned}$$

Next integrate by parts the expression in brackets with respect to y using $\mathrm{d}v = \varphi'[G(y)]g(y)\mathrm{d}y$ and $t = y f(y - \beta' \boldsymbol{u})$. Since φ is bounded and f has a finite second moment this leads to

$$\begin{aligned}\frac{\partial \overline{RD}}{\partial \beta} = &\iint \varphi[G(y)] f(y - \beta' \boldsymbol{u})\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{u}) + \iint \varphi[G(y)]\, y f'(y - \beta' \boldsymbol{u})\,\boldsymbol{u}\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{u}) \\ &- \iint \varphi[G(y)]\, y f'(y - \beta' \boldsymbol{x})\boldsymbol{x}\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{x}) \\ = &\iint \varphi[G(y)] f(y - \beta' \boldsymbol{u})\,\boldsymbol{u}\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{u}).\end{aligned}$$

Hence the second derivative of $\overline{RD}$ is

$$\begin{aligned}\frac{\partial^2 \overline{RD}}{\partial \beta\, \partial \beta'} = &- \iint \varphi[G(y)] f'(y - \beta' \boldsymbol{x})\,\boldsymbol{x}\boldsymbol{x}'\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{x}) \\ &- \iiint \varphi'[G(y)] f(y - \beta' \boldsymbol{x}) f(y - \beta' \boldsymbol{u})\boldsymbol{x}\boldsymbol{u}'\,\mathrm{d}y\,\mathrm{d}M(\boldsymbol{x})\,\mathrm{d}M(\boldsymbol{u}).\end{aligned} \tag{4.3}$$

Now integrate the first term on the right-hand side of (4.3) by parts with respect to y by using $dt = f'(y - \beta'x)dy$ and $v = \varphi[G(y)]$. This leads to

$$\frac{\partial^2 \overline{RD}}{\partial\beta\,\partial\beta'} = -\int\int\int \varphi'[G(y)]f(y-\beta'x)f(y-\beta'u)x(u-x)'\,dy\,dM(x)\,dM(u). \tag{4.4}$$

We have, however, the following identity:

$$\begin{aligned}
&\int\int \varphi'[G(y)]f(y-\beta'x)f(y-\beta'u)(u-x)(u-x)'\,dy\,dM(x)\,dM(u)\\
&\quad = \int\int \varphi'[G(y)]f(y-\beta'x)f(y-\beta'u)u(u-x)'\,dy\,dM(x)\,dM(u)\\
&\quad\quad - \int\int \varphi'[G(y)]f(y-\beta'x)f(y-\beta'u)x(u-x)'\,dy\,dM(x)\,dM(u).
\end{aligned}$$

Since the two integrals on the right-hand side of the last expression are negatives of each other, this combined with expression (4.3) leads to

$$2\frac{\partial^2 \overline{RD}}{\partial\beta\,\partial\beta'} = \int\int \varphi'[G(y)]f(y-\beta'x)f(y-\beta'u)(u-x)\,(u-x)'\,dy\,dM(x)\,dM(u).$$

Since the distribution functions f and M are continuous and the score function is increasing, it follows that the right side of this last expression is a positive definite matrix.

It follows from these theorems that the $\overline{R}_i$'s satisfy similar properties of association as $\overline{R}^2$ does. We have $0 \leqslant \overline{R}_i \leqslant 1$. By Theorem 4.2, $\overline{R}_i = 0$ if and only if $\beta = 0$ if and only if Y and x are independent.

Further, understanding of $\overline{R}_i$ can be gleaned from their direct relationship with R^2 for the multivariate normal model.

Theorem 4.3. *Assume that the multivariate normal model holds. Then*

$$\overline{R}_1 = 1 - \sqrt{1-\overline{R}^2} \tag{4.5}$$

$$\overline{R}_2 = \frac{1-\sqrt{1-\overline{R}^2}}{1-\sqrt{1-\overline{R}^2}+\tau^2/2\sigma_y\sigma}, \tag{4.6}$$

where σ_y^2 denotes the variance of Y.

Proof. Note that $\sigma_y^2 = \sigma^2 + \beta'\Sigma\beta$. Further, the distribution function of Y is $G(y) = \Phi((y - \alpha - \beta'E(x))/\sigma_y)$ where Φ is the standard normal distribution function. Let $T = \int \varphi[\Phi(t)]t\,d\Phi(t)$. It then follows that $\overline{D}_y = \sigma_y T$, $\overline{D}_e = \sigma T$, and $T = \sigma/\tau$. The results follow immediately from these relationships and from the fact that $\overline{R}^2 = 1 - (\sigma^2/\sigma_y^2)$.

It can be shown directly that the $\overline{R}_i$'s are one-to-one increasing functions of $\overline{R}^2$. Hence, for this model the parameters $\overline{R}^2$, $\overline{R}_1$, and $\overline{R}_2$ are equivalent.

Under Wilcoxon scores, we obtain $\overline{R}^2 = \overline{R}_1^{*2} = \overline{R}_2^{*2}$ where

$$\overline{R}_1^{*2} = 1 - (1 - \overline{R}_1)^2$$

$$\overline{R}_2^{*2} = 1 - \left[\frac{1 - \overline{R}_2}{1 - \overline{R}_2(1 - (\pi/6))}\right]^2.$$

Note that $\overline{R}^2$, $\overline{R}_1^{*2}$, and $\overline{R}_2^{*2}$ are on the same scale. This will be useful for the comparisons found in Section 5. □

4.1. *Bounded influence CMDs*

It is clear from their definitions that both $\overline{D}_{By}$ an $\overline{D}_{Be}$ are strictly positive. Similarly we can show that $\overline{RD}_B(\beta)$ is a strictly convex function of β with a minimum value of 0 at $\beta = 0$. Thus $0 \leqslant \overline{R}_{Bi} \leqslant 1$ for $i = 1, 2$ and $\overline{R}_{Bi} = 0$ if and only if Y and $\boldsymbol{x}$ are independent. Due to the weights, though, simple expressions for these CMDs are not obtainable.

5. Behavior of CMDs under contamination

We evaluated the CMDs for two situations:

Situation 1. The random error e has a contaminated normal distribution with proportion of contamination ε and the ratio of contaminated variance to uncontaminated σ_c^2. The random variable x has a univariate normal N(0, 1) distribution, and $\beta = 1$.

Situation 2. The random error e has a N(0, 1) distribution while the random variable x has a contaminated normal distribution with proportion of contamination ε and the ratio of contaminated variance to uncontaminated σ_c^2.

In both cases Y and x are dependent. Without loss of generality, we took $\alpha = 0$ in (2.1). We used Wilcoxon scores and, for the bounded influence CMDs, we used the simple weight function $b(x) = 1$ or $2/|x|$ depending on whether $|x| < 2$ or $|x| > 2$. Closed form expressions for $\overline{R}_1$ and $\overline{R}_2$ were obtained while $\overline{R}_{B1}$ and $\overline{R}_{B2}$ were evaluated by numerical integration routines found in NAG (1983).

For Situation 1, Table 1 displays these parameters for several values of ε and for $\sigma_c^2 = 9$ and 100. For ease of interpretation we rescaled the robust CMDs as discussed in Section 4. Thus at the normal ($\varepsilon = 0$) we have $\overline{R}_1^{*2} = \overline{R}_2^{*2} = \overline{R}^2$ with the common value of 0.5 in these situations. Although this is not necessarily true of $\overline{R}_{B1}^*$ and $\overline{R}_{B2}^*$, it does put them on a similar scale to $\overline{R}^2$. As either ε or σ change, the amount of dependence between Y and $\boldsymbol{x}$ changes; hence all the coefficients change somewhat. However, R^2 decays as the percentage of contamination increases, and the decay is rapid in the case where $\sigma_c^2 = 100$. This is true also, to a lesser degree, for $\overline{R}_1^*$ and $\overline{R}_{B1}^*$ which is predictable since their denominators have unbounded influence in the Y-space. The coefficients $\overline{R}_2^*$ and $\overline{R}_{B2}^*$ show robustness to the increase in contamination. For instance when $\sigma_c^2 = 100$, R^2 decays 0.44 units while these robust coefficients decay only 0.14 units.

Table 2 shows the CMDs for Situation 2. In this situation, all the coefficients except $\overline{R}_{B1}^*$ and $\overline{R}_{B2}^*$ inflate, rapidly for $\sigma_c^2 = 100$, as ε increases. In terms of fitting data, this corresponds to least squares and regular R-fits sensitivity to outliers in the x-space. Such points draw these fits towards them resulting in larger R^2, R_1, and R_2. On the other hand, the bounded influence R-estimate is not as sensitive to these outliers.

Table 1
Coefficients of multiple determination under contaminated errors (e)

CMD	$e \sim CN(\varepsilon, \sigma_c^2 = 9)$ ε						$e \sim CN(\varepsilon, \sigma_c^2 = 100)$ ε					
	0.00	0.01	0.02	0.05	0.10	0.15	0.00	0.01	0.02	0.05	0.10	0.15
$\overline{R}^2$	0.50	0.48	0.46	0.42	0.36	0.31	0.50	0.33	0.25	0.14	0.08	0.06
$\overline{R}_1^*$	0.50	0.50	0.48	0.45	0.41	0.38	0.50	0.47	0.42	0.34	0.26	0.19
$\overline{R}_2^*$	0.50	0.50	0.49	0.47	0.44	0.42	0.50	0.49	0.47	0.45	0.40	0.36
$\overline{R}_{B1}^*$	0.50	0.48	0.48	0.44	0.41	0.38	0.50	0.45	0.42	0.34	0.26	0.21
$\overline{R}_{B2}^*$	0.49	0.49	0.47	0.46	0.43	0.40	0.49	0.47	0.46	0.44	0.39	0.36

Table 2
Coefficients of multiple determination under contaminated x

CMD	$x \sim CN(\varepsilon, \sigma_c^2 = 9)$ ε						$x \sim CN(\varepsilon, \sigma_c^2 = 100)$ ε					
	0.00	0.01	0.02	0.05	0.10	0.15	0.00	0.01	0.02	0.05	0.10	0.15
$\overline{R}^2$	0.50	0.52	0.54	0.58	0.64	0.69	0.50	0.67	0.75	0.86	0.92	0.94
$\overline{R}_1^*$	0.50	0.51	0.52	0.56	0.62	0.66	0.50	0.58	0.63	0.75	0.85	0.90
$\overline{R}_2^*$	0.50	0.51	0.53	0.57	0.62	0.66	0.50	0.57	0.63	0.74	0.85	0.90
$\overline{R}_{B1}^*$	0.50	0.50	0.51	0.52	0.56	0.59	0.50	0.51	0.52	0.55	0.62	0.66
$\overline{R}_{B2}^*$	0.49	0.50	0.50	0.52	0.55	0.57	0.49	0.50	0.51	0.53	0.58	0.62

Example. A simple data set found on page 19 of Rousseeuw and Leroy (1987) illustrates the statistics R_1 and R_2. The response is the monthly payment made by a Belgian insurance firm over the course of a year and the predictor is time, 1:12. From a plot of the data, there seems to be little if any linear relationship between the response and time. This is confirmed by the statistics R_1^{*2} and R_2^{*2} which had values 0.001 and 0.003, respectively. There is one large outlier in this data set which occurs at month 12 and has a pronounced affect on the LS fit. This results in $R^2 = 0.185$ which would indicate some relationship between monthly payment and time. If the outlier is removed then $R^2 = 0.0341$, similar to the results of the robust statistics on the original data. On the outlier deleted data, $R_1^{*2} = 0.02$ and $R_2^{*2} = 0.03$.

6. Robust CMDs based on M-estimates

Although much of the above discussion is based on the class of R-estimates, the development is completely general and a parallel development can be made for any robust estimation procedure which is based on minimizing an objective function. In this section we demonstrate this development for the class of M-estimates. In general an M-estimate is found by minimizing a function of the form, $\sum_{i=1}^n t(\boldsymbol{x}_i, (y_i - \boldsymbol{x}_i'\beta)/\sigma)$. In practice σ is replaced by an initial estimate of scale, often the MAD. See Hampel et al. (1986) for details. Regular M-estimates are obtained by taking $t(\boldsymbol{x}_i, y/\sigma) = \rho(y/\sigma)$ for some function ρ. One that is frequently used is Huber's ρ-function which is given by $\rho(z) = z^2/2$ or $c|z| - c^2/2$ depending on whether $|z| < c$ or

$|z| \geqslant c$, respectively, where c is a specified constant. The M-estimates based on Huber's ρ-function have bounded influence in the Y-space but not in the x-space. Hampel et al. (1986) discuss M-estimates of the above form which bound influence in both spaces.

For M-estimates, tests of H_0: $\beta = 0$ can readily be formed based on the drop in the M-objective function similar to the drop dispersion for R-estimates. These were proposed by Schrader and Hettmansperger (1980) for regular M-estimates and were generalized to the bounded influence M-estimates by Hampel et al. (1986).

Using these estimates and tests, a development of CMDs based on M-estimates can be formulated which parallels the development based on R-estimates found in Sections 3 and 4; see Witt (1989) for details. In general these will be zero if and only if x and Y are independent. However, unlike their R-counterparts, these coefficients are not easily interpretable parameters, even under the multivariate normal model. For regular M-estimates with an unbounded ρ-function, the M-analogue of R_1 will not be robust but the analogue of R_2 will be robust in the y-space. Analogues of $\overline{R}_2$ based on bounded influence M-estimates will be robust in both spaces, similar to those based on the bounded influence R-estimates.

7. Conclusion

The classical CMD $\overline{R}^2$ is quite sensitive to departures from normality. We have shown that functionals corresponding to robust versions of the R^2 statistic are: (i) measures of association, and (ii) more robust against departures from normality. The study found in Section 5 verifies these findings for families of contaminated normal distributions and further shows that $\overline{R}_1$ and $\overline{R}_2$ are more stable than $\overline{R}^2$ over variations in the error distribution. If there is contamination in only the Y-space then $\overline{R}_2$ is recommended while if there is contamination in the x-space then $\overline{R}_{B2}$ is recommended.

The framework presented here is quite general, and analogous results are available for M-estimators or other procedures which are based on minimizing an objective function. One advantage of the rank-based measures $\overline{R}_1$ and $\overline{R}_2$ in this regard is their interpretability: under a multivariate normal distribution, they are one-to-one functions of the classical $\overline{R}^2$.

In a regression setting, like R^2, the statistics R_1 and R_2 can be used for selecting predictors. We are planning a study on this use.

Acknowledgement

The research of Drs. McKean and Naranjo was partially supported by NSF Grant DMS-9103916. We appreciate the helpful comments from an anonymous referee.

References

Arnold, S.F. (1980), Asymptotic validity of F-tests for the ordinary linear model and the multiple correlation model, *J. Amer. Statist. Assoc.* **75**, 890–894.

Arnold, S.F. (1981), *The Theory of Linear Models and Multivariate Analysis* (Wiley, New York).

David, H.A. (1970), *Order Statistics* (Wiley, New York).

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986), *Robust Statistics* (Wiley, New York).

Heiler, S. and R. Willers (1988), Asymptotic normality of R-estimates in the linear model, *Statistics* **19**, 173–184.

Hettmansperger, T.P. (1984), *Statistical Inference Based on Ranks* (Wiley, New York).

Huber, P.J. (1973), Robust regression: asymptotics, conjectures, and Monte Carlo, *Ann. Statist.* **1**, 799–821.

Jaeckel, L.A. (1972), Estimating regression coefficients by minimizing the dispersion of the residuals, *Ann. Math. Statist.* **43**, 1449–1458.

Koul, H.L., G.L. Sievers and J.W. McKean (1987), An estimator of the scale parameter for the rank analysis of linear models under general score functions, *Scand. J. Statist.* **14**, 131–141.

McKean, J.W. and T.P. Hettmansperger (1976), Tests of hypotheses of the general linear model based on ranks, *Comm. Statist. Part A-Theory and Methods*, **5**, 693–709.

McKean, J.W. and T.P. Hettmansperger (1978), A robust analysis of the general linear model based on one step R-estimates, *Biometrika* **65**, 571–579.

McKean, J.W. and G.L. Sievers (1987), Coefficients of determination for least absolute deviation analysis, *Statist. Probab. Lett.* **5**, 49–54.

McKean, J.W. and G.L. Sievers (1989), Rank scores suitable for the analysis of linear models under asymmetric error distributions, *Technometrics* **31**, 207–218.

Numerical Algorithms Group, Inc. (1983), Library Manual Mark 15, Numerical Algorithms Groups, Oxford.

Naranjo, J.D. (1989), Bounded-influence regression: A modified Wilcoxon procedure, Ph.D. Thesis, Department of Statistics, The Pennsylvania State University, unpublished.

Naranjo, J.D. and T.P. Hettmansperger (1994), Bounded-influence rank regression, *J. Roy. Statist. Soc. B* **56**, 209–220.

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

Schrader, R.M. and T.P. Hettmansperger (1980), Robust analysis of variance based upon a likelihood ratio criterion, *Biometrika* **67**, 93–101.

Sievers, G.L. (1983), A weighted dispersion function for estimation in linear models, *Comm. Statist. Theory Methods* **12**, 1161–1179.

Witt, L.D. (1989), Coefficients of multiple determination based on rank estimates, Ph.D. Thesis, Department of Mathematics and Statistics, Western Michigan University, unpublished.