Journal of Statistical Planning and Inference 140 (2010) 1852–1862

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A robust coefficient of determination for regression

Olivier Renaud [a,b,*], Maria-Pia Victoria-Feser [c,1]

[a] Methodology and Data Analysis, Psychology Department, University of Geneva, CH-1211 Geneva 5, Switzerland
[b] Distance Learning University, CH-3960 Sierre, Switzerland
[c] Faculty of Economics and Social Sciences, University of Geneva, CH-1211 Geneva 5, Switzerland

ARTICLE INFO

Article history:
Received 27 October 2009
Received in revised form
11 January 2010
Accepted 12 January 2010
Available online 20 January 2010

Keywords:
Consistency
Efficiency
Outliers
R-squared
Correlation

ABSTRACT

To assess the quality of the fit in a multiple linear regression, the coefficient of determination or R^2 is a very simple tool, yet the most used by practitioners. Indeed, it is reported in most statistical analyzes, and although it is not recommended as a final model selection tool, it provides an indication of the suitability of the chosen explanatory variables in predicting the response. In the classical setting, it is well known that the least-squares fit and coefficient of determination can be arbitrary and/or misleading in the presence of a single outlier. In many applied settings, the assumption of normality of the errors and the absence of outliers are difficult to establish. In these cases, robust procedures for estimation and inference in linear regression are available and provide a suitable alternative.

In this paper we present a companion robust coefficient of determination that has several desirable properties not shared by others. It is robust to deviations from the specified regression model (like the presence of outliers), it is efficient if the errors are normally distributed, it does not make any assumption on the distribution of the explanatory variables (and therefore no assumption on the unconditional distribution of the responses). We also show that it is a consistent estimator of the population coefficient of determination. A simulation study and two real datasets support the appropriateness of this estimator, compared with classical (least-squares) and several previously proposed robust R^2, even for small sample sizes.

1. Introduction

The regression model with response variable y and regressors $\mathbf{x}$ can be stated as $y|\mathbf{x} \sim (\mathbf{x}^T\boldsymbol{\beta}; \sigma^2)$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)$ a q + 1 dimensional vector that contains the regression parameters or slopes with β_0 the intercept and consequently $\mathbf{x}$=$(1, x_1, \ldots, x_q)^T$. For a sample of n observations, this amounts to postulating

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2) \quad \forall i = 1, \ldots, n. \tag{1}$$

The regression model is suitable if the underlying hypotheses are satisfied. In particular, the linearity of the relationship between the response and (possibly some function of) the explanatory variables holds, as well as the fact that the residual variance is constant.

* Corresponding author at: Methodology and Data Analysis, Psychology Department, University of Geneva, CH-1211 Geneva 5, Switzerland.
E-mail addresses: Olivier.Renaud@unige.ch (O. Renaud), maria-pia.victoriafeser@unige.ch (M.-P. Victoria-Feser).
[1] Partially supported by Swiss National Science Foundation, grant #PP001-106465.

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.01.008

Like for other models, given a dataset, it is important to be able to check the fit of the regression model. There exists in fact several measures for goodness-of-fit assessment and variable selection that are more or less routinely used in practice. These include for example the *F*-test, the Akaike (1973) *AIC* criterion, Mallows (1973) C_p or cross-validation (Stone, 1974; Shao, 1993). *AIC* and C_p belong to covariance penalty methods (see Efron, 2004) because they include a covariance penalty that corrects an estimated prediction error. More specifically, let $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ be the predicted value for y_i at the model with $p < q$ explanatory variables (based on a least-squares fit) and $\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ be the corresponding residual sum of squares, then Γ_p, the true C_p, can be written as

$$\Gamma_p = \frac{\mathbb{E}(\mathrm{RSS})}{n\sigma^2} + \frac{2}{n\sigma^2} \sum_{i=1}^{n} \mathbb{C}\mathrm{ov}(y_i, \hat{y}_i) - 1 \tag{2}$$

(see also Dupuis and Victoria-Feser, 2003, 2006). Efron (2004) shows the link between covariance penalty methods and cross-validation and related nonparametric bootstrap techniques.

More traditionally, for the linear regression model, a very simple fit indicator is, however, given by the coefficient of determination R^2. If $\mathbf{x}$ is considered as random, it can be defined as the (population) parameter that is the squared correlation between y and the best linear combination of the $\mathbf{x}$ (Anderson, 1984, p. 40):

$$\phi^2 = \max_{\boldsymbol{\beta}} \mathbb{C}\mathrm{orr}^2(y, \mathbf{x}^T \boldsymbol{\beta}). \tag{3}$$

If the regressors are considered as fixed, the same type of definition can be carried out (Helland, 1987); see also more details in the appendix. It is important to note that normality is *not* assumed, merely the existence of the second moments. Compared to covariance penalty methods, although the R^2 is solely based on the covariance penalty, it plays an important role in model fit assessment. It should certainly not be used as a unique model fit assessor, but can provide a reasonable and rapid model fit indication.

The R^2 is usually presented as the quantity that estimates the percentage of variance of the response variable explained by its (linear) relationship with the explanatory variables. It is computed by means of the ratio

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{4}$$

where ESS, TSS and RSS are respectively the explained, total and residual sum of squares. When there is an intercept term in the linear model, this coefficient of determination is actually equal to the square of the correlation coefficient between y_i and $\hat{y}_i$, i.e. (see e.g. Greene, 1997, p. 253)

$$R^2 = \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}} \right)^2, \tag{5}$$

with $\bar{\hat{y}}$ the mean predicted responses. Eq. (5) has a nice interpretation in that R^2 measures the goodness of fit of the regression model by its ability to predict the response variable, ability measured by the correlation. Further, this expression shows that the (unconditional) distribution of the response does not need to be Gaussian to allow for the interpretation of R^2. It also shows that it is a direct estimator of the population parameter (3). In finite samples, the R^2 is biased upward and is often adjusted, e.g. $R^2_{\mathrm{adj}} = 1 - (1 - R^2)((n-1)/(n-q))$.

In a classical framework, all the fit criteria work well as summary measures because the postulated model is assumed to be the exact generating process. But what happens if e.g. the errors are not normal, but instead the data have actually been generated through a model that is in a neighborhood of the assumed model? This is the fundamental assumption in robust statistics that consequently provides a set of robust estimators, testing procedures, goodness-of-fit measures built upon the hypothesis that the postulated model is only an approximated model. The most common types of model deviation are outliers, and robust procedures prevent the latter from biasing the regression parameter estimates, the testing procedures as well as the goodness-of-fit criteria. For example, Ronchetti (1982) (see also Ronchetti, 1997) proposes a robust version for the *AIC*, Ronchetti and Staudte (1994) propose a robust version for the C_p and Ronchetti et al. (1997) robustify the Shao (1993) cross-validation based method. Indeed, any goodness-of-fit criteria should, at its level, give a summary indication of the fit of the data to the postulated model. If the later is assumed to be only an approximation of the reality, then the measure should give an indication of the fit for the majority of the data, possibly leaving aside a few outlying observations. In other words, the (robust) goodness-of-fit criterion is used to choose a good model for the majority of the data rather than an "average" model for all the data.

Several robust R^2 have been proposed in the literature (see next section). However, no simulations are reported that assess their properties in some way, and as it will be shown through two examples and by means of simulations, the available robust R^2 can be misleading in assessing the fit of the regression model to the (majority of) data. In this paper, we propose an alternative robust estimator that we present as a robust version of the correlation coefficient between y_i and $\hat{y}_i$ that makes no assumption on the (unconditional) distribution of the responses. We show that it can be reformulated as a weighted version of (4). We also show how to obtain a consistent estimator of the population coefficient of determination. A simulation study and the analysis of two datasets illustrate that it better represents the fit of the model to the data than the available robust R^2 and that it is robust to model misspecification such as outliers, contrarily to the classical (least-squares) R^2.

2. Robust coefficients of determination

It is rather obvious that the R^2 (4) can be driven by extreme observations, not only through the LS estimator $\hat{\boldsymbol{\beta}}$ used to compute the predicted responses, but also through the average response $\bar{y}$ and the possible large residuals or deviations $y_i - \bar{y}$. For the slope parameter, one can choose a robust estimator $\hat{\boldsymbol{\beta}}$, for example in the class of M-estimators (Huber, 1964) defined generally (for the regression model) by the solution in $\boldsymbol{\beta}$ of

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i),$$

where $r_i = (y_i - \mathbf{x}_i^T\boldsymbol{\beta})/\sigma$, or alternatively by the solution in $\boldsymbol{\beta}$ of

$$\sum_{i=1}^{n} \psi(r_i)\mathbf{x}_i = \mathbf{0}, \tag{6}$$

with $\psi(z) = \partial/\partial z \rho(z)$. The functions ρ and ψ actually generalize, respectively, the log-likelihood and the score associated with the postulated model. For the resulting M-estimator to be robust to small amounts of model deviations (small amounts of outliers), a sufficient condition is that the function ψ is bounded. Redescending ψ functions improve the robustness of the M-estimator to larger amounts. Such a function is given by the popular Tukey's (Beaton and Tukey, 1974) biweight function:

$$\psi_{[bi]}(r_i; c) = \begin{cases} r_i\left(\left(\frac{r_i}{c}\right)^2 - 1\right)^2 & \text{if } |r_i| \le c, \\ 0 & \text{if } |r_i| > c. \end{cases}$$

Note that it can be written as

$$\psi_{[bi]}(r_i; c) = w(r_i, c) r_i \quad \text{with } w(r_i, c) = \begin{cases} \left(\left(\frac{r_i}{c}\right)^2 - 1\right)^2 & \text{if } |r_i| \le c, \\ 0 & \text{if } |r_i| > c. \end{cases} \tag{7}$$

To compute $\hat{\boldsymbol{\beta}}$, one needs an estimate of σ and an iterative procedure with a suitable starting point. One first computes a starting consistent (very robust) estimator $\hat{\boldsymbol{\beta}}^0$ together with a robust residual scale estimator $\hat{\sigma}^2$ and uses $\hat{\boldsymbol{\beta}}^0$ as starting point in an iterative procedure to find the solution $\hat{\boldsymbol{\beta}}_{bi}$ in $\boldsymbol{\beta}$ of $\sum_{i=1}^{n} \psi_{[bi]}(r_i; c)\mathbf{x}_i = \mathbf{0}$ with σ^2 replaced by $\hat{\sigma}^2$. This estimator is actually an MM-estimator (Yohai, 1987). The efficiency with respect to the LS estimator can be chosen with a suitable value for c in (7) (see also Yohai et al., 1991). At the (exact) regression model, c=4.685 leads to an MM-estimator with 95% efficiency compared to the LS estimator.

Given then a choice for the $\rho-$function, Maronna et al. (2006, p. 171), propose a robust equivalent for the R^2 (4) given by

$$R^2_{\rho} = 1 - \frac{\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)}{\sum_{i=1}^{n} \rho\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right)}, \tag{8}$$

where $\hat{\mu}$ is a robust location estimate of $\mathbb{E}(y)$, solution of

$$\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mu}{\hat{\sigma}}\right) \tag{9}$$

and $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are robust estimators for $\boldsymbol{\beta}$ and σ (for the full model) based on the chosen ρ function. Independently, Croux and Dehon (2003) propose a class of robust R^2 which generalizes (4) given by

$$R^2_S = 1 - \frac{S_n(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})}{S_n(y_i - \hat{\mu}),} \tag{10}$$

where S_n is a robust scale estimator. For example, using the L_1 regression estimator one defines

$$R^2_{L_1} = 1 - \left(\frac{\sum_{i=1}^{n} |y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{L_1}|}{\sum_{i=1}^{n} |y_i - \text{median}_i y_i|}\right)^2 \tag{11}$$

and for the least median of squares (LMS) regression estimator (Rousseeuw, 1984) one has

$$R^2_{\text{LMS}} = 1 - \left(\frac{\text{median}_i |y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\text{LMS}}|}{\text{SHORT}}\right)^2, \tag{12}$$

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

Although (8) and (10) are direct generalizations of (4) to the robust framework, they suffer from an important drawback: in practice they are often biased. We will illustrate this point with a simulation study and two datasets in Sections 3 and 4. One possible reason why this phenomenon happens is that in computing R^2_ρ or R^2_S, one uses two "models": the regression model through $y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, and a location model through the term in the denominator ($y_i - \hat{\mu}, y_i - \text{median}_i y_i$ or SHORT). These two quantities are not influenced by model deviations in the same way, so that bounding these quantities directly and separately is not necessarily appropriate in the regression model framework. More specifically, suppose that the regressor $\mathbf{x}$ is a dummy variable (representing a category of a categorical variable) or binary, and hence the response is split into two groups, then when the two groups are very different, i.e. the response is bimodal, a robust location estimator (9) will consider most of the responses (i.e. residuals) as extreme. Similarly, the SHORT might exclude totally the responses of one category, underestimating the total variability and thus the coefficient of variation, see the second row of Figs. 1–3. It should be noted that Croux and Dehon (2003) made it very clear that their proposed robust R^2_S class is suitable under the assumption that the marginal distribution of the response variable is a location-scale transformation of the conditional or error's distribution. This means in practice that it is symmetric and unimodal, so that the dummy variable example above does not satisfy this rather strong hypothesis.

To remedy this problem, we propose instead to measure the goodness of fit of the model by its ability to predict the response using the correlation between the response and its prediction (5), which in the robust case is

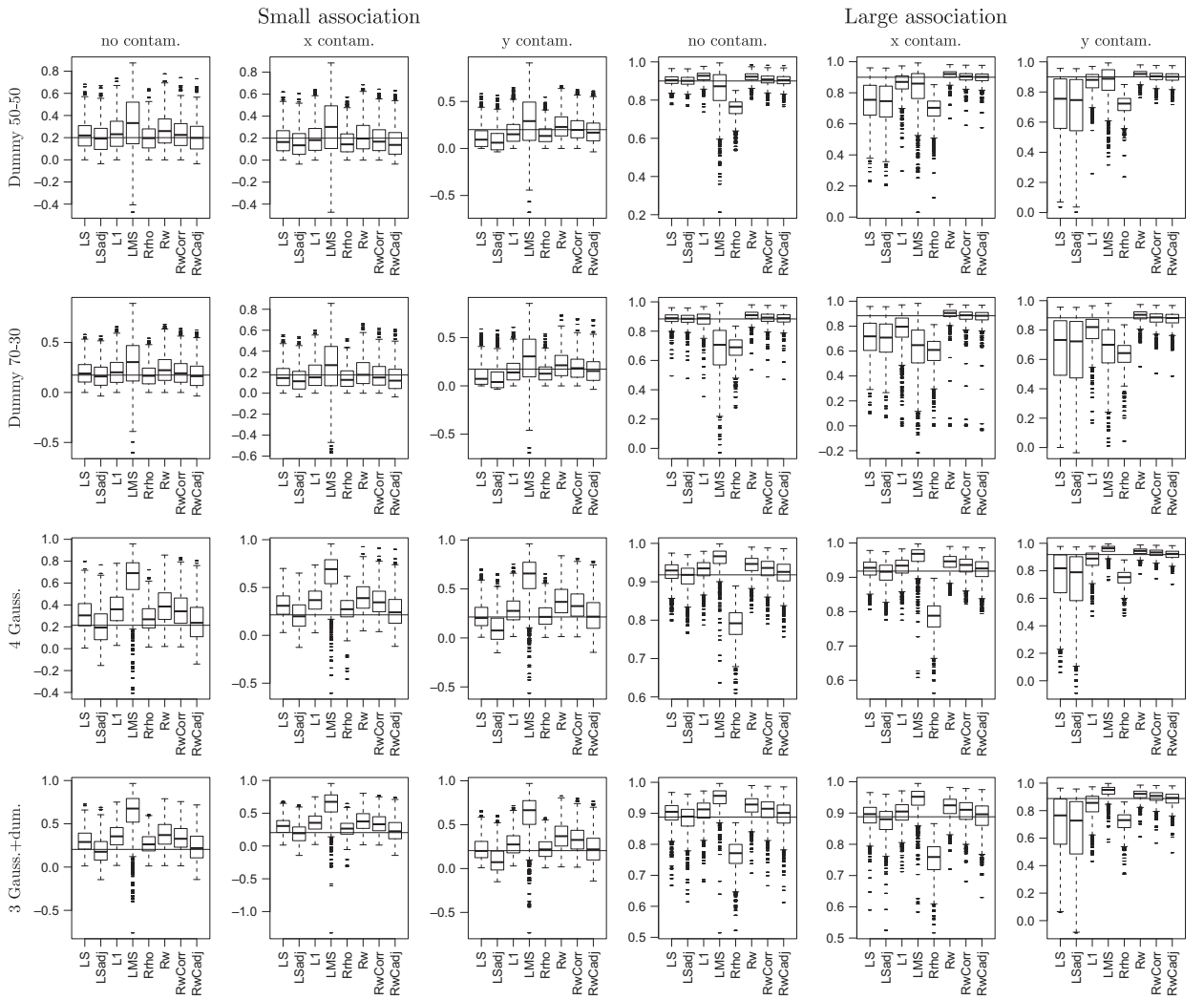

Fig. 1. Boxplots of R^2 for samples of size $n=30$. Four settings for the x's: a unique dummy with probability 50–50% (Dummy 50–50), with probability 70–30% (Dummy 70–30), four-dimensional correlated Gaussian (4 Gauss.) and three-dimensional correlated Gaussian plus an uncorrelated dummy (3 Gauss.+dum.). Eight estimators: least-squares R^2 (LS and LSadj), S-type robust estimators using the L_1 criterion (L1) and the LMS criterion (LMS), S-Plus/R implemented robust (Rrho), and the proposed robust in the uncorrected (Rw) version and corrected version (RwCorr and RwCadj). The strength of the association is either small or large, the data are either not contaminated (no contam.), or the x's are contaminated (x contam.), or the responses are contaminated (y contam.). Overall the proposed corrected method seems to be the only method that is consistent.

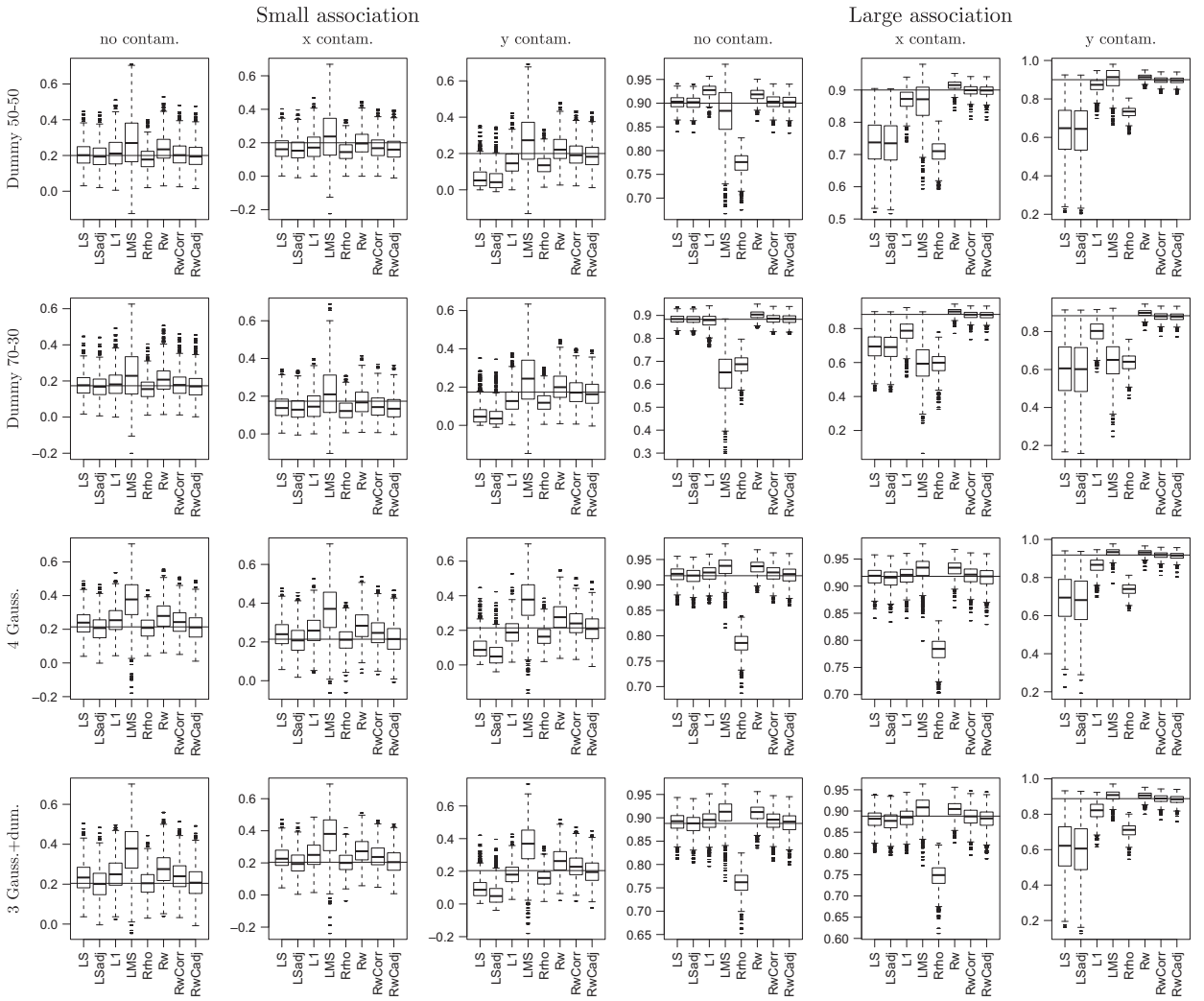

Fig. 2. Same settings as in Fig. 1, but with n=100 observations for each sample.

given by

$$R_w^2 = \left(\frac{\sum_{i=1}^n w_i (y_i - \overline{y}_w)(\hat{y}_i - \overline{\hat{y}}_w)}{\sqrt{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2 \sum_{i=1}^n w_i (\hat{y}_i - \overline{\hat{y}}_w)^2}} \right)^2, \tag{13}$$

where $\overline{y}_w = (1/\sum w_i)\sum w_i y_i$, $\overline{\hat{y}}_w = (1/\sum w_i)\sum w_i \hat{y}_i$ and the weights w_i and the predicted values $\hat{y}_i$ are those produced by the robust regression estimator, for example Tukey's biweight $w_i = w(r_i, c)$, with $r_i = (y_i - \hat{y}_i)/\hat{\sigma}^2$ as defined in the paragraph below (7) and $\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{bi}$. Since these weights are used to estimate the model (1), they downweight observations based on their residuals, not on their y value. This is a different rationale as for example simply using a robust correlation estimator such as one in the class proposed by Huber (1981), or the one proposed by Gnanadesikan and Kettenring (1972) (and studied by Devlin et al., 1981). Indeed, in both cases, like with R_ρ^2 and R_S^2, observations are downweighted (or even trimmed) when $y_i - \overline{y}$ and/or $\hat{y}_i - \overline{\hat{y}}$ is large (with respect to the underlying normal model), and as explained above with the dummy variable example, with respect to the regression model, the "wrong" observations can be pinpointed. It is therefore important to base a weighting scheme on the regression model as is done with our proposed robust R_w^2 in (13).

With the same weights and predictions, another robust R_w^2 can be based on its explained versus total sum of squares version (4):

$$\tilde{R}_w^2 = 1 - \frac{\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2}. \tag{14}$$

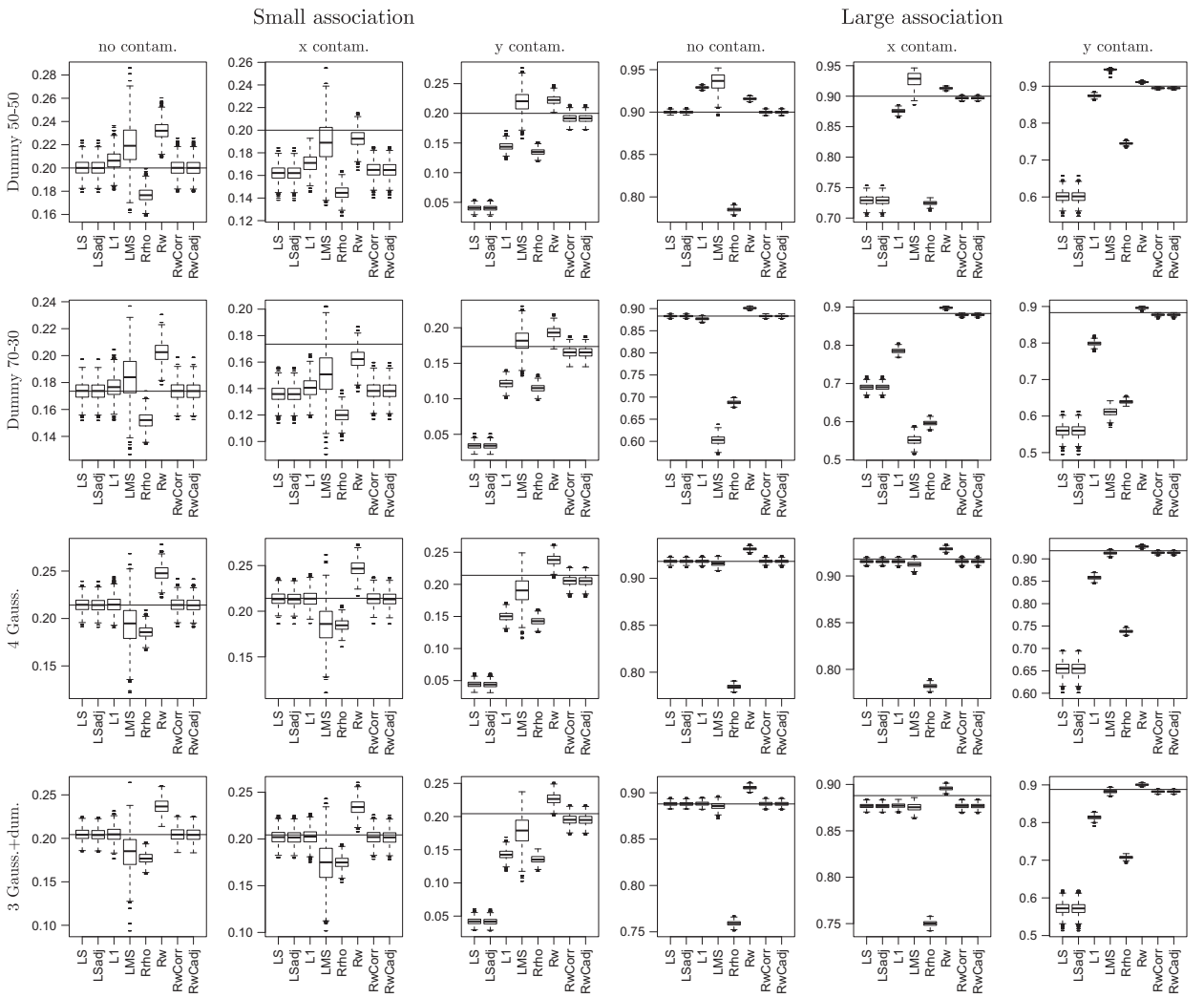

Fig. 3. Same settings as in Fig. 1, but with $n=10{,}000$ observations for each sample.

It turns out, as will be shown in Theorem 1, that the two robust coefficients of determinations are equal. A more general formulation is given by

$$\check{R}^2_{w,a} = \frac{\sum_{i=1}^n w_i(\hat{y}_i - \bar{\hat{y}}_w)^2}{\sum_{i=1}^n w_i(\hat{y}_i - \bar{\hat{y}}_w)^2 + a\sum_{i=1}^n w_i(y_i - \hat{y}_i)^2}, \tag{15}$$

which allows for a possible correction factor a for consistency considerations.

Theorem 1. *The two robust coefficients of determinations given in* (13) and (14) *are equal to* (15) *with* $a=1$. *Moreover, with no assumption on the distribution of the explanatory variables* (*and therefore no assumption on the unconditional distribution of the responses*), *but under the assumption of normality of the errors and for a consistent estimator of the residual scale* $\hat{\sigma}$, *a consistent estimator of the population coefficient of determination* (3) *is obtained when* $a = \mathbb{E}(w(r))/\mathbb{E}(\psi(r))$ *in* (15). *In particular, for the biweight with* $c=4.685$ (95% efficiency), $a=1.2076$.

The proof of this theorem is given in the appendix. As a robust and consistent $\hat{\sigma}$, we propose to use the *S*-estimator with biweight $\rho-$function (see Rousseeuw and Yohai, 1984). For small samples and a relatively large number of covariates, using the same rationale than for the classical R^2, the robust coefficient might benefit of being adjusted, hence leading to the adjusted coefficient

$$\check{R}^2_{adj,w,a} = 1-(1-\check{R}^2_{w,a})\left(\frac{n-1}{n-q}\right). \tag{16}$$

A simulation study and two datasets provide further evidence on the appropriateness of the proposed robust R^2.

3. Simulation study

A simulation study was carried out to compare all the R^2 estimators presented in this paper. The aim is to evaluate their consistency, their variability and their robustness.

We used four different settings for the explanatory variables: a unique dummy explanatory variable, generated with probability of 50% of being 0 (Dummy 50–50); a unique dummy explanatory variable, generated with probability of 70% of being 0 (Dummy 70–30); four correlated Gaussian variables (4 Gauss.); three correlated Gaussian plus an uncorrelated dummy variable (3 Gauss.+dum.). In all cases, y is generated according to model (1). The sample size is either $n=30$ for Fig. 1 or $n=100$ for Fig. 2 or $n=10{,}000$ for Fig. 3.

Eight estimators are tested. The first two are least-squares R^2 (Eq. (4), LS) and (LSadj), respectively, without and with adjustment for small sample. The middle ones are S-type robust estimators using, respectively, the L_1 criterion (Eq. (11), L1) and the LMS criterion (Eq. (12), LMS), S-Plus/R implemented robust R^2_ρ (Eq. (8), Rrho). The last three are the proposed robust $R^2_w = \check{R}^2_{w,1}$ (Eq. (13), Rw) in the uncorrected version, $\check{R}^2_{w,1.2}$ (Eq. (15), RwCorr) and $\check{R}^2_{adj,w,1.2}$ (Eq. (16), RwCadj) in the corrected version with $a=1.2$, respectively, without and with adjustment for small sample. Two additional characteristics of the simulation were varied. The strength of the association is either small (ϕ^2 in (3) close to 0.2) for the boxplots on the left side of the panel or strong (ϕ^2 around 0.9) for the boxplots on the right side of the panel. Finally, in the "no contam." columns all the errors are normally distributed with unit variance; in the "y contam." columns, the errors have a probability of 5% to have a standard deviation 10 times larger (hence of 10); in the "x contam." for the two first rows, the value of the dummy variable has a probability of 5% to be switched (0–1 or conversely); for the two last rows, the value of the first explanatory Gaussian variable has a probability of 5% to have a standard deviation 10 times larger. In each case, 1000 samples have been generated. The horizontal line displays the population ϕ^2.

If we first look at the simulation with no contamination (columns with "no contam."), we see that in all cases only the estimators of the coefficient of determination by least-squares (LS) and the proposed corrected robust (RwCadj) are unbiased. For the small sample size (30) and several predictors (two last rows of Fig. 1), the adjusted versions (LSadj and RwCadj) reveal to be useful. Moreover, the proposed robust coefficient does not seem to be more variable than the LS, indicating a good efficiency at the model. All the other methods are somehow biased at the model, sometimes to a large extent. Concerning contamination on the response (columns with "y contam."), even 5% of contamination hamper the least-squares estimate up to an important amount, the other robust methods are also biased. Only the proposed corrected robust (RwCadj and possibly RwCorr) stays unbiased in all the settings. A 5% of contamination in not unrealistic in practice and the proposed method brings a good safeguard. Finally, if one contaminates the explanatory variables (columns with "x contam."), the same conclusions hold, except for sample size $n=10{,}000$, small association and only one dummy (two first rows of Fig. 3). In that case, the contamination seems to perturb even the proposed corrected robust (RwCorr). Actually, this design (small association and contaminated dummy variables) appears to be a difficult one also for the robust approach, even with smaller sample sizes. The boxplots in the 2 first rows of the second column of Fig. 1 and Fig. 2 show that the R^2 are not quite centered, but their variance is relatively large, while with $n=10{,}000$ the bias appears because the variance becomes small enough. It therefore seems that the breakdown of the R^2 is reached at 5% of contaminated data in that type of setting, and consequently, the proposed corrected robust (RwCorr) shows the same performance as the classical R^2 based on the LS.

However, as a general conclusion of this simulation study, overall the proposed corrected robust estimator $\check{R}^2_{w,1.2}$ (Eq. (15), RwCorr), or for small samples and a relatively large number of covariates its adjusted version ($\check{R}^2_{adj,w,1.2}$, Eq. (16), RwCadj), are the only one that meet the three aims of consistency, efficiency and robustness. Empirically, the proposed coefficient is remarkably unbiased, even for small sample sizes.

4. Application to real datasets

The first dataset we consider here is about the measurement or estimation of glomerular filtration rate (GFR) by means of Serum Creatinine (CR). Several models have been proposed to explain the logarithm of the GFR by means of the CR and possibly other explanatory variables such as AGE and SEX (see e.g. Rule et al., 2004). The data we consider here are the GFR and CR measured on a random sample of 30 men out of the 180 in the Brochner-Mortensen et al. (1977) study of renal function, and analyzed in Ingelfinger et al. (1987, Table 9b-2, p. 229) as well as in Heritier et al. (2009, Chapter 3). For the purpose of comparing different (robust) versions of coefficients of determination, we will consider the following (simple) model:

$$\log(\mathrm{GFR}_i) = \beta_0 + \frac{\beta_1}{\mathrm{CR}_i} + \varepsilon_i. \tag{17}$$

The parameter $\boldsymbol{\beta}$ is estimated using the biweight *MM*-estimator and the estimated regression lines together with the LS estimated regression lines with and without observation no. 2 are presented in Fig. 4; see also Heritier et al. (2009, Chapter 3). The corresponding R^2 are for the LS on all the data of 0.73, for the LS without observation no. 2 of 0.85, for the robust R^2_ρ

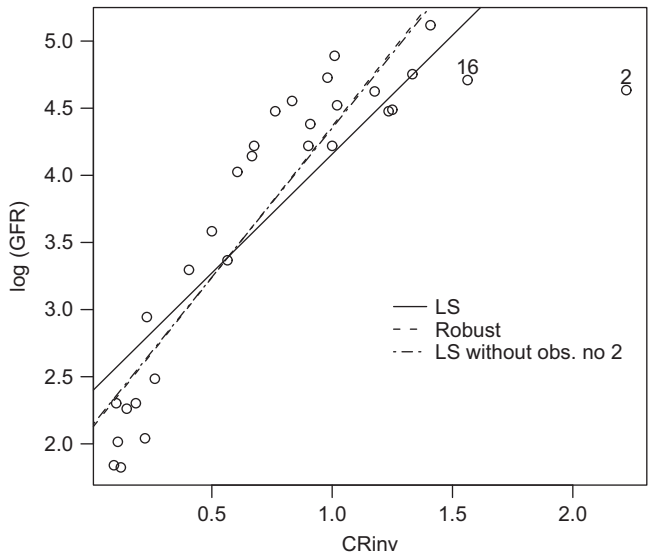

Fig. 4. Estimated regression lines for the model $\log(\mathrm{GFR}) = \beta_0 + \beta_1 \mathrm{CR}^{-1}$.

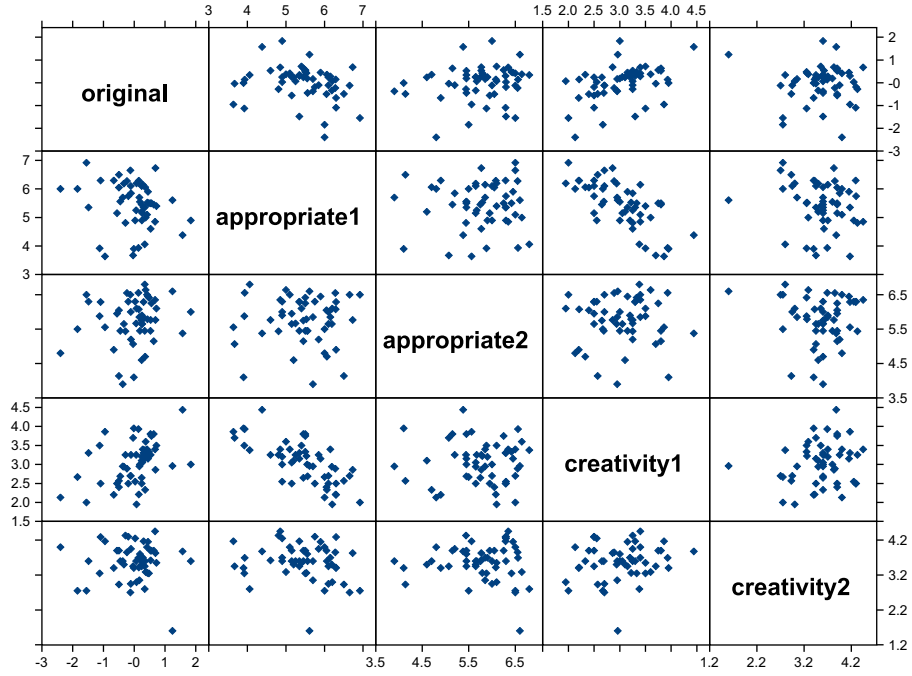

Fig. 5. Matrix plot of 5 scores for 54 children.

in (8) of 0.68 and for the robust $R_w^2 = \check{R}_{w,1}^2$ in (13) of 0.86 in the uncorrected version and of $\check{R}_{w,1.2}^2 = 0.84$ in the corrected version (a = 1.2). The adjusted versions for small sample are less than 1% smaller than the above (non-adjusted) version, due to the fact that there is a unique predictor. Since the coefficient of determination is an indicator of the fit of the regression model to the data, it should reflect how the data (or the majority of the data for the robust version) evolve around the regression line. The robust estimator clearly leads to a better fit (except for the two extreme observations no 2 and 16) than the LS estimator, but this is not reflected by the R^2 in (8). Without observation no. 2, the R^2 corresponding to the LS is the largest and it is also the case for the robust $\check{R}_{w,1.2}^2$, and hence better reflects the nature of the robust fit as illustrated in Fig. 4. Note finally that a small curvature is visible in Fig. 4. However, both classical and robust R^2 are not built to measure the appropriateness of the linearity assumption, but instead to what extend the responses are well predicted by their estimated linear predictor.

The second dataset is in the field of psychology, where two tasks have been assigned individually to 54 primary school children (Fürst et al., in press). Originality, appropriateness and creativeness scores have been graded and standardized. The matrix plot is given in Fig. 5. This dataset is interesting as there are no clear far outliers, but still many points lie outside the core of the majority of points, as exemplified in the scatter plot between *original* and *appropriate1*. It would then be almost impossible to adopt a strategy to decide which observation(s) to remove before fitting a regression model. In this situation, only a robust approach can handle the situation by weighting appropriately each observation. Concerning the multiple regression of *original* on the four other variables, diagnostics show indeed that the residuals are not Gaussian and many observations may be influencing the results. The robust fit gives a residual scale estimate of 0.47 while the LS residual scale estimate is 0.70. However, the original robust $R^2_\rho = 0.16$ is substantially smaller than the LS R^2=0.23. This seems somehow contradictory, since a better fit (higher R^2) is interpreted as smaller unexplained error. The robust coefficients of determination proposed in this paper give a value of $\check{R}^2_{w,1} = 0.36$ in the uncorrected version and of $\check{R}^2_{w,1.2} = 0.32$ in the corrected version (a=1.2). This example seems to support the simulation results showing that R^2_ρ and the R^2 are probably biased downward in the presence of outliers. The same comparison can be made on coefficients adjusted for small sample. Here, R^2_{adj}=0.17 and $\check{R}^2_{\text{adj},w,1.2} = 0.26$, which shows that after adjusting for small sample, the proposed robust R^2 gives again a more adequate indication of fit.

5. Conclusion

Assessing the model fit is an important step in data analysis. This can be achieved by means of the computation (estimation) of some criterion such as the coefficient of determination for the regression model. The chosen measure should indicate to what extend the data fit the postulated model, and when the later is supposed to be just an approximation of the reality, in particular by allowing the presence in the data of a few extreme observations, then the fit should be measured for the majority of the data. In this paper we illustrate this point by means of the analysis of two real datasets, propose an alternative robust estimator for the coefficient of determination and show analytically and by means of simulations its consistency, efficiency, robustness and unbiasedness even for small samples.

Appendix A. Proof of Theorem 1

To show the equality of the three forms, we need to define the following quantities. Let first $y_i^* = \sqrt{w_i} y_i$, $\mathbf{y}^* = [\mathbf{y}_i^*]$, $\mathbf{x}_i^* = \sqrt{w_i}\mathbf{x}_i$ and $\mathbf{X}^* = \mathbf{X}^*[\mathbf{x}_i^*]$ with weights w_i given in (7). With this notation and (7), Eq. (6) can thus be written as

$$\mathbf{X}^{*T}(\mathbf{X}^*\hat{\boldsymbol{\beta}}_{bi} - \mathbf{y}^*) = \mathbf{X}^{*T}(\hat{\mathbf{y}}^* - \mathbf{y}^*) = \mathbf{0}, \tag{18}$$

where $\hat{\mathbf{y}}^* = [\sqrt{w_i}\mathbf{x}_i\hat{\boldsymbol{\beta}}_{bi}] = \mathbf{X}^*\hat{\boldsymbol{\beta}}_{bi}$. Define $\mathbf{w}^{1/2} = [\sqrt{w_i}]$, and the projection matrix

$$M_w = I_n - \frac{1}{\sum_i w_i}\mathbf{w}^{1/2}\mathbf{w}^{T/2}.$$

Then we have

$$M_w\mathbf{y}^* = \mathbf{y}^* - \frac{1}{\sum_i w_i}\mathbf{w}^{1/2}\mathbf{w}^{T/2}\mathbf{y}^* = \mathbf{y}^* - \mathbf{w}^{1/2}\overline{y}_w,$$

and similarly $M_w\hat{\mathbf{y}}^* = \hat{\mathbf{y}}^* - \mathbf{w}^{1/2}\overline{\hat{y}}_w$. We also have that

$$M_w(\mathbf{y}^* - \hat{\mathbf{y}}^*) = (\mathbf{y}^* - \hat{\mathbf{y}}^*) - \frac{1}{\sum_i w_i}\mathbf{w}^{1/2}\mathbf{w}^{T/2}(\mathbf{y}^* - \hat{\mathbf{y}}^*) = (\mathbf{y}^* - \hat{\mathbf{y}}^*). \tag{19}$$

The last equation is due to the fact that the first column of $\mathbf{X}^*$ is $\mathbf{w}^{1/2}$, so that the first row in (18) is $\mathbf{w}^{T/2}(\hat{\mathbf{y}}^* - \mathbf{y}^*) = 0$. With the above, the matrix form of the coefficients (13)–(15) are

$$R^2_w = \frac{[\hat{\mathbf{y}}^{*T}M_w\mathbf{y}^*]^2}{\mathbf{y}^{*T}M_w\mathbf{y}^*\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^*},$$

$$\tilde{R}^2_w = \frac{\mathbf{y}^{*T}M_w\mathbf{y}^* - (\mathbf{y}^* - \hat{\mathbf{y}}^*)^T M_w(\mathbf{y}^* - \hat{\mathbf{y}}^*)}{\mathbf{y}^{*T}M_w\mathbf{y}^*} = \frac{2\hat{\mathbf{y}}^{*T}M_w\mathbf{y}^* - \hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^*}{\mathbf{y}^{*T}M_w\mathbf{y}^*},$$

$$\check{R}^2_{w,a} = \frac{\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^*}{\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^* + a\cdot(\mathbf{y}^* - \hat{\mathbf{y}}^*)^T M_w(\mathbf{y}^* - \hat{\mathbf{y}}^*)}.$$

Showing that $\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^* = \hat{\mathbf{y}}^{*T}M_w\mathbf{y}^*$ will prove that these three terms are equal (with a=1 for the last one). Using (19) then (18), we have

$$\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^* = \hat{\mathbf{y}}^{*T}M_w\mathbf{y}^* - \hat{\mathbf{y}}^{*T}M_w(\mathbf{y}^* - \hat{\mathbf{y}}^*) = \hat{\mathbf{y}}^{*T}M_w\mathbf{y}^* - \hat{\boldsymbol{\beta}}_{bi}^T\mathbf{X}^{*T}(\mathbf{y}^* - \hat{\mathbf{y}}^*) = \hat{\mathbf{y}}^{*T}M_w\mathbf{y}^*,$$

which proves the first statement of the theorem.

For the consistency, we first need to rewrite the parameter ϕ^2 in (3). Let $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\beta}}$ be $\mathbf{x}$ and $\boldsymbol{\beta}$ without the first (intercept) element. Then, the solution of (3) is $\tilde{\boldsymbol{\beta}} = \Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\Sigma_{y\tilde{\mathbf{x}}}$ (see Anderson, 1984, p. 40) and ϕ^2 can be written in the following equivalent forms (see e.g. Gurland, 1968):

$$\phi^2 = \frac{\tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\tilde{\boldsymbol{\beta}}}{\Sigma_{yy}} = \frac{\tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\tilde{\boldsymbol{\beta}}}{\tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\tilde{\boldsymbol{\beta}}+\sigma^2} \quad \left(= \frac{(\Sigma_{y\tilde{\mathbf{x}}}^T\tilde{\boldsymbol{\beta}})^2}{\Sigma_{yy}\tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\tilde{\boldsymbol{\beta}}} = \frac{\Sigma_{y\tilde{\mathbf{x}}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\Sigma_{y\tilde{\mathbf{x}}}}{\Sigma_{yy}}\right), \tag{20}$$

where the Σ are the covariance matrix of the indices. The second equality comes from model (1). Note that normality is *not* assumed here, merely the existence of the second moments. If the regressors are considered as fixed, using a sequence of models as in Eq. (1) indexed by n, can lead to the same results, see Helland (1987).

Let us now rewrite $\check{R}_{w,a}^2$ as

$$\begin{aligned}\check{R}_{w,a}^2 &= \frac{\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^*}{\hat{\mathbf{y}}^{*T}M_w\hat{\mathbf{y}}^* + a\cdot(\mathbf{y}^*-\hat{\mathbf{y}}^*)^TM_w(\mathbf{y}^*-\hat{\mathbf{y}}^*)} = \frac{\hat{\boldsymbol{\beta}}_{bi}^T\mathbf{X}^{*T}M_w\mathbf{X}^*\hat{\boldsymbol{\beta}}_{bi}}{\hat{\boldsymbol{\beta}}_{bi}^T\mathbf{X}^{*T}M_w\mathbf{X}^*\hat{\boldsymbol{\beta}}_{bi} + a\cdot(\mathbf{y}^*-\hat{\mathbf{y}}^*)^TM_w(\mathbf{y}^*-\hat{\mathbf{y}}^*)}\\ &= \frac{\hat{\boldsymbol{\beta}}_{bi}^T\left(\frac{\sum_i w_i}{n^2}\mathbf{X}^{*T}M_w\mathbf{X}^*\right)\hat{\boldsymbol{\beta}}_{bi}}{\hat{\boldsymbol{\beta}}_{bi}^T\left(\frac{\sum_i w_i}{n^2}\mathbf{X}^{*T}M_w\mathbf{X}^*\right)\hat{\boldsymbol{\beta}}_{bi} + \frac{a\sum_i w_i}{n^2}\cdot(\mathbf{y}^*-\hat{\mathbf{y}}^*)^TM_w(\mathbf{y}^*-\hat{\mathbf{y}}^*).}\end{aligned} \tag{21}$$

We will show the convergence in probability of each term separately to the terms in the second expression of (20). First, since $\hat{\boldsymbol{\beta}}_{bi}$ is (Fisher) consistent (Yohai, 1987), it converges in probability to $\boldsymbol{\beta}$. By assumption, we also have the same property for $\hat{\sigma}$. Second, given that the weights $w_i := w_i(\hat{\boldsymbol{\theta}})$ with $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_{bi}^T, \hat{\sigma})^T$ can be approximated by a Taylor series expansion as $w_i(\hat{\boldsymbol{\theta}}) = w_i(\boldsymbol{\theta}) + (\partial/\partial\boldsymbol{\theta})w_i(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ with

$$\frac{\partial}{\partial\boldsymbol{\theta}}w_i(\boldsymbol{\theta}) = \frac{4}{c}\left(1-\left(\frac{\varepsilon_i}{\sigma c}\right)^2\right)\left(\frac{\varepsilon_i}{\sigma c}\right)\begin{bmatrix}\mathbf{x}_i\\ \frac{1}{\sigma^2}\varepsilon_i\end{bmatrix}$$

and given that the observations and the errors ε_i are *iid*, the weights w_i are asymptotically independent. Note that a similar argument holds for the residuals r_i.

For the middle term, we then have

$$\left\{\frac{\sum_i w_i}{n^2}\mathbf{X}^{*T}M_w\mathbf{X}^*\right\}_{j,l} = \frac{\sum_i w_i}{n^2}\left(\mathbf{x}_j^{*T}\mathbf{x}_l^* - \sum_i w_i\overline{x}_{jw}\overline{x}_{lw}\right) = \frac{1}{n^2}\left(\sum_i\sum_k w_iw_kx_{kj}x_{kl} - \sum_i\sum_k w_iw_kx_{ij}x_{kl}\right) = \frac{1}{n^2}\sum_i\sum_{k\neq i} w_iw_k(x_{kj}-x_{ij})x_{kl},$$

which expectation converges to

$$\frac{1}{n^2}\sum_i\sum_{k\neq i}\mathbb{E}^2(w(r))(x_{kj}-x_{ij})x_{kl} = \frac{1}{n^2}\sum_i\sum_k\mathbb{E}^2(w(r))(x_{kj}-x_{ij})x_{kl} = \frac{\mathbb{E}^2(w(r))}{n^2}\left(n\sum_k x_{kj}x_{kl} - \sum_i x_{ij}\sum_k x_{kl}\right) = \mathbb{E}^2(w(r))\hat{\Sigma}_{j,l}. \tag{22}$$

If X is random, the computation leads to $\mathbb{E}^2(w(r))\Sigma_{j,l}$, since the w_i are independent of X.

Let $\mathbf{r}^* = [\sqrt{w_i}r_i] = (\mathbf{y}^*-\hat{\mathbf{y}}^*)/\hat{\sigma}$. For the second term in the denominator of (21), we do the same computation:

$$\frac{a\hat{\sigma}^2\sum_i w_i}{n^2}\mathbf{r}^{*T}M_w\mathbf{r}^* = \frac{a\hat{\sigma}^2}{n^2}\left(\sum_i\sum_{k\neq i} w_iw_kr_kr_k - \sum_i\sum_{k\neq i} w_iw_kr_ir_k\right) = \frac{a\hat{\sigma}^2}{n^2}\left(\sum_i\sum_{k\neq i} w_i(r_k\psi(r_k)) - \sum_i\sum_{k\neq i}\psi(r_i)\psi(r_k)\right).$$

Since the r_i are asymptotically independent, the two expressions in the summand have a covariance that converges to zero, which implies that the expectation of their product converges to the product of the two expectations. The expectation of the above equation thus converges to

$$\frac{a\sigma^2}{n^2}\left(\sum_i\sum_{k\neq i}\mathbb{E}(w(r))\mathbb{E}(r\psi(r)) - \sum_i\sum_{k\neq i}\mathbb{E}^2(\psi(r))\right) = \frac{a\sigma^2(n-1)}{n}\mathbb{E}(w(r))\mathbb{E}(r\psi(r)),$$

since $\mathbb{E}(\psi(\mathbb{r})) = 0$ by symmetry.

All the above terms have a variance that converges to 0, and multiple applications of the multivariate version of Slutsky theorem proves the convergence of (21) to

$$\frac{\mathbb{E}^2(w(r))\boldsymbol{\beta}^T\Sigma_{\mathbf{xx}}\boldsymbol{\beta}}{\mathbb{E}^2(w(r))\boldsymbol{\beta}^T\Sigma_{\mathbf{xx}}\boldsymbol{\beta} + a\mathbb{E}(w(r))\mathbb{E}(r\psi(r))\sigma^2} = \frac{\mathbb{E}^2(w(r))\tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\tilde{\boldsymbol{\beta}}}{\mathbb{E}^2(w(r))\tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}\tilde{\boldsymbol{\beta}} + a\mathbb{E}(w(r))\mathbb{E}(r\psi(r))\sigma^2.}$$

The last equality comes from the fact that the first line and row of $\Sigma_{\mathbf{xx}}$ (corresponding to the intercept) contain only zeros. It is clear now that if $a = \mathbb{E}(\mathbb{w}(\mathbb{r}))/\mathbb{E}(\mathbb{r}\psi(\mathbb{r}))$, the above is equal to the second expression of (20) and hence $\check{R}_{w,a}^2$ is consistent.

Finally, for the computation of a for the biweight, we have

$$\mathbb{E}(w(r)) = \int_{-c}^{c} \left(\left(\frac{r}{c}\right)^2 - 1 \right)^2 d\Phi(r) = \frac{1}{c^4} \int_{-c}^{c} r^4 \, d\Phi(r) - 2\frac{1}{c^2} \int_{-c}^{c} r^2 \, d\Phi(r) + \int_{-c}^{c} d\Phi(r),$$

where Φ is the standard normal distribution, and

$$\mathbb{E}(r\psi(r)) = \int_{-c}^{c} r^2 \left(\left(\frac{r}{c}\right)^2 - 1 \right)^2 d\Phi(r) = \frac{1}{c^4} \int_{-c}^{c} r^6 \, d\Phi(r) - 2\frac{1}{c^2} \int_{-c}^{c} r^4 \, d\Phi(r) + \int_{-c}^{c} r^2 \, d\Phi(r).$$

The truncated moments of the standard normal distribution $\int_{-c}^{c} r^k \, d\Phi(r)$ can be found in Heritier et al. (2009, Appendix B). Replacing these in the expressions above for $c = 4.685$, leads to $a = 1.2076$.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Anderson, T.W., 1984. An Introduction to Multivariate Statistical Analysis, second ed. Wiley, New York.

Beaton, A.E., Tukey, J.W., 1974. The fitting of power series, meaning polynomials, illustrated on band-spectroscopicdata. Technometrics 16, 147–185.

Brochner-Mortensen, J., Jensen, S., Rodbro, P., 1977. Assessment of renal function from plasma creatinine in adult patients. Scandinavian Journal of Urology and Nephrology 11, 263–270.

Croux, C., Dehon, C., 2003. Estimators of the multiple correlation coefficient: local robustness and confidence intervals. Statistical Papers 44, 315–334.

Devlin, S.J., Gnanadesikan, R., Kettenring, J.R., 1981. Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association 76, 354–362.

Dupuis, D., Victoria-Feser, M.-P., 2003. A prediction error criterion for choosing the lower quantile in Pareto index estimation. Cahiers de recherche HEC no. 2003.10, University of Geneva.

Dupuis, D.J., Victoria-Feser, M.-P., 2006. A robust prediction error criterion for Pareto modeling of upper tails. Canadian Journal of Statistics 34, 639–658.

Efron, B., 2004. The estimation of prediction error: covariance penalties and cross-validation. Journal of the American Statistical Association 99, 619–632.

Fürst, G., Lubart, T., Ghisletta, P., 2010. Originalité, caractère approprié et créativité des idées chez les enfants de 8-11 ans. In: De Ribaupierre, A., Ghisletta, P., Lecerf, T., Roulin, J.-L. (Eds.), Identité et spécificité de la psychologie différentielle. Presses Universitaires de Rennes, Rennes, pp. 263–268.

Gnanadesikan, R., Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 29, 81–124.

Greene, W., 1997. Econometric Analysis, third ed. Prentice-Hall, Englewood Cliffs, NJ.

Gurland, J., 1968. A relatively simple form of the distribution of the multiple correlation coefficient. Journal of the Royal Statistical Society, Series B: Methodological 30, 276–283.

Helland, I.S., 1987. On the interpretation and use of R^2 in regression analysis. Biometrics 43, 61–69.

Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.-P., 2009. Robust Methods in Biostatistics. Wiley, New York.

Huber, P.J., 1964. Robust estimation of a location parameter. Annals of Mathematical Statistics 35, 73–101.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Ingelfinger, J.A., Mosteller, F., Thibodeau, L.A., Ware, J.H., 1987. Biostatistics in Clinical Medicine, second ed. MacMillan, New York.

Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–675.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, Chichester, West Sussex, UK.

Ronchetti, E., 1982. Robust testing in linear models: the infinitesimal approach. Ph.D. Thesis, ETH, Zurich, Switzerland.

Ronchetti, E., 1997. Robustness aspects of model choice. Statistica Sinica 7, 327–338.

Ronchetti, E., Field, C., Blanchard, W., 1997. Robust linear model selection by cross-validation. Journal of the American Statistical Association 92, 1017–1023.

Ronchetti, E., Staudte, R.G., 1994. A robust version of Mallows's C_p. Journal of the American Statistical Association 89, 550–559.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.

Rousseeuw, P.J., Yohai, V.J., 1984. Robust regression by means of S-estimators. In: Franke, J., Härdle, W., Martin, D. (Eds.), Robust and Nonlinear Time Series Analysis. Springer, New York, pp. 256–272.

Rule, A.D., Larson, T.S., Bergstralh, E.J., Slezak, J.M., Jacobsen, S.J., Cosio, F.G., 2004. Using serum creatinine to estimate glomerular filtration rate: accuracy in good health and in chronic kidney disease. Annals of Internal Medicine 141 (12), 929–938.

Shao, J., 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88, 486–494.

Stone, M., 1974. Cross-validation choice and assessment of statistical predictions. Journal of the Royal Statistical Society, Series B 36, 1–147.

Yohai, V.J., 1987. High breakdown point and high efficiency robust estimates for regression. The Annals of Statistics 15, 642–656.

Yohai, V.J., Stahel, W.A., Zamar, R.H., 1991. A procedure for robust estimation and inference in linear regression. In: Stahel, W.A., Weisberg, S. (Eds.), Directions in Robust Statistics and Diagnostics, Part II, The IMA Volumes in Mathematics and its Applications, vol. 34. Springer, Berlin, pp. 365–374.