

Technical Report 78

(Ab)Using Regression for Data Adjustment

Lutz Dümbgen  
September 2016

Abstract

In various economic applications, people want to compare  $n$  units with respect to certain quantities  $Y_1, Y_2, \dots, Y_n$  measuring their performance. The latter, however, is often influenced by certain factors which are beyond control of the units, and one would like to extract an adjusted performance from the data. Specifically, let  $X_i \in \mathcal{X}$  summarize the factors of the  $i$ -th unit. Then one could think of a model equation  $Y_i = f_o(X_i) + \varepsilon_i$  with a regression function  $f_o : \mathcal{X} \rightarrow \mathbb{R}$  describing the unavoidable influence of the factors  $X_i$  and  $\varepsilon_i$  being the adjusted performance of the  $i$ -th unit. Now a common proposal is to estimate  $f_o$  via regression methods by a function  $\hat{f}$  depending on the current data  $(X_i, Y_i)$ , possibly augmented by additional past data, and to use the residuals  $\hat{\varepsilon}_i := Y_i - \hat{f}(X_i)$  as surrogates for the adjusted performances  $\varepsilon_i$ . In the present report we discuss this approach, its potential pitfalls and (mis)interpretation. In particular, an unavoidable property of the residuals  $\hat{\varepsilon}_i$  is that they measure only parts of the adjusted performance while the remaining parts get hidden in the estimated function  $\hat{f}$ . Possible alternatives are mentioned briefly.

## 1 Introduction

This report is motivated by various consulting cases all of which involved a variant of the following method: Suppose that one wants to compare  $n$  units with respect to certain quantities  $Y_1, Y_2, \dots, Y_n$  measuring their performance. Two specific examples are:

- The units are hospitals, and for a certain type of diseases,  $Y_i$  measures the mean success of treatment within the  $i$ -th hospital.
- The units are service areas of a big postal service, and  $Y_i$  measures the mean delivery time for an item in service area  $i$ . (Here one aims for low values of  $Y_i$ .)

As usual in regression, we refer to the  $Y_i$  as ‘responses’, although in the present context ‘raw performance measures’ would be appropriate, too. Typically the performance is influenced by certain factors which cannot be controlled by the units, and one would like a fair comparison, taking such differences into account. In the example of hospitals, some hospitals may tend to get more problematic cases than others or the mix of clients may vary from hospital to hospital with respect to important factors such as age, social background

or initial diagnosis (in case of considering various diseases simultaneously). In the example of postal service areas, the population density is certainly negatively correlated with the delivery times. In general let  $X_i \in \mathcal{X}$  be a tuple containing the potentially relevant factors of the  $i$ -th unit. Then one could think of a model equation

$$Y_i = f_o(X_i) + \varepsilon_i \quad \text{for } 1 \leq i \leq n \quad (1)$$

with a regression function  $f_o : \mathcal{X} \rightarrow \mathbb{R}$  describing the unavoidable influence of the tuple  $X_i$ , while  $\varepsilon_i$  is the adjusted performance (‘true performance’) of the  $i$ -th unit. One could also think of a multiplicative model with  $Y_i = f_o(X_i) \cdot \varepsilon_i$  with positive quantities  $f_o(X_i)$  and  $\varepsilon_i$ , but then a log-transformation would lead us to the additive equation (1) with  $\log Y_i$ ,  $\log f_o(X_i)$  and  $\log \varepsilon_i$  in place of  $Y_i$ ,  $f_o(X_i)$  and  $\varepsilon_i$ , respectively.

If we were able to estimate the regression function  $f_o$  with sufficient accuracy by a function  $\hat{f}$  from the actual data  $(X_1, Y_1), \dots, (X_n, Y_n)$ , possibly augmented by additional past data, one could use the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{f}(X_i)$$

as surrogates for the adjusted performances  $\varepsilon_i$ . A standard way to estimate  $f_o$  would be via some regression method, for instance, least squares estimation of  $f_o$  under the assumption that it belongs to a given family  $\mathcal{F}$  of regression functions  $f : \mathcal{X} \rightarrow \mathbb{R}$ . Thus we estimate  $f_o$  by a function  $\hat{f} \in \mathcal{F}$  such that

$$\sum_{i=1}^n (Y_i - \hat{f}(X_i))^2$$

becomes minimal. In this report we focus on this approach with linear models  $\mathcal{F}$ . That means,  $\mathcal{F}$  is assumed to be a finite-dimensional linear space of functions on  $\mathcal{X}$ . The simplest case is multiple linear regression: Suppose that  $x \in \mathcal{X}$  stands for a tuple of  $K$  numerical or  $\{0, 1\}$ -valued covariables  $x^{(1)}, \dots, x^{(K)}$ . Then one could consider functions

$$f(x) = a + \sum_{k=1}^K b_k x^{(k)}$$

with real constants  $a, b_1, \dots, b_K$ . The set  $\mathcal{F}$  of all such functions  $f$  is a linear space of dimension  $K + 1$ . Alternatively one could consider multiple quadratic regression involving functions of the form

$$f(x) = a + \sum_{k=1}^K b_k x^{(k)} + \sum_{k=1}^K \sum_{\ell=k}^K g_{k,\ell} x^{(k)} x^{(\ell)}$$

with real constants  $a, b_k$  and  $g_{k,\ell}$ ,  $1 \leq k \leq \ell \leq K$ . This corresponds to a function space  $\mathcal{F}$  of dimension  $(K + 1)(K + 2)/2$ . For a general introduction to regression methods we refer to Ryan (1997) or Dümbgen (2015).

In Section 2 we discuss (non-)identifiability of  $f_o$  in (1) and its implications. In particular we contrast the present (ab)use of regression with common applications and mention briefly an alternative approach called data envelopment analysis.

In Section 3 we discuss the potential impact of replacing the  $\varepsilon_i$  with the residuals  $\widehat{\varepsilon}_i$  under the rather optimistic assumption that the  $\varepsilon_i$  may be viewed as realisations of independent Gaussian random variables with mean 0 and a common standard deviation  $\sigma > 0$ . It is shown that ranking of units via  $\widehat{\varepsilon}_i$  is strongly influenced by the observations' leverages.

Finally, in Section 4 we describe a potential alternative method which is feasible whenever for each unit  $i$  we have several observations ('cases')

$$(X_{ij}, Y_{ij}), \quad 1 \leq j \leq J_i.$$

For instance, in the example of hospitals, we may have data of  $J_i$  patients or treatments in hospital  $i$ , and the tuples  $X_{ij}$  may be case-specific rather than hospital-specific. Then a possible alternative to model equation (1) is given by

$$Y_{ij} = f_o(X_{ij}) + a_i + \epsilon_{ij} \tag{2}$$

with a regression function  $f_o : \mathcal{X} \rightarrow \mathbb{R}$  as before, unit-specific parameters  $a_i$  measuring their performances and random errors  $\epsilon_{ij}$ .

Some technical arguments and proofs are deferred to Section 5.

## 2 Non-identifiability and its implications

For the moment let us view  $\varepsilon = (\varepsilon_i)_{i=1}^n$  as a fixed  $n$ -dimensional vector. Without further assumptions on  $\varepsilon$ , the function  $f_o$  is not well-defined through (1). We could replace  $f_o$  and  $\varepsilon$  with  $f_o + \Delta$  and  $(\varepsilon_i - \Delta(X_i))_{i=1}^n$ , respectively, where  $\Delta$  is an arbitrary function in our model space  $\mathcal{F}$ , and (1) would remain true.

For  $x \in \mathcal{X}$  the value  $f_o(x)$  could be interpreted as the average response over all units, if all of them were forced to work under conditions as specified by  $x$ . Thus a natural additional requirement seems to be that

$$\sum_{i=1}^n \varepsilon_i = 0. \tag{3}$$

But (3) alone does not alleviate our identifiability problem. Suppose that  $\mathcal{F}$  contains all constant functions. For any non-constant function  $\Delta \in \mathcal{F}$  we could replace  $f_o$  and  $\varepsilon$  with  $f_o + \Delta - c$  and  $(\varepsilon_i - \Delta(X_i) + c)_{i=1}^n$ , respectively, where  $c := n^{-1} \sum_{i=1}^n \Delta(X_i)$ . Then both (1) and (3) would remain true.

Recall that the least squares estimator of  $f_o$  is a function  $\hat{f} \in \mathcal{F}$  minimising the sum of squares  $\sum_{i=1}^n (Y_i - \hat{f}(X_i))^2$ . Geometrically speaking, the vector  $\hat{f}(\mathbf{X}) = (\hat{f}(X_i))_{i=1}^n$  is the orthogonal projection of  $\mathbf{Y} = (Y_i)_{i=1}^n$  onto the linear subspace

$$\mathcal{F}(\mathbf{X}) := \{f(\mathbf{X}) : f \in \mathcal{F}\}$$

of  $\mathbb{R}^n$ , where  $f(\mathbf{X}) := (f(X_i))_{i=1}^n \in \mathbb{R}^n$ . In particular,  $\hat{\varepsilon} = (\hat{\varepsilon}_i)_{i=1}^n$  is the orthogonal projection of  $\mathbf{Y}$  onto the space  $\mathcal{F}(\mathbf{X})^\perp$  of all vectors  $\mathbf{v} \in \mathbb{R}^n$  which are perpendicular to  $\mathcal{F}(\mathbf{X})$ . If  $f_o \in \mathcal{F}$ , then  $\hat{\varepsilon} = (\hat{\varepsilon}_i)_{i=1}^n$  is the orthogonal projection of  $\varepsilon = (\varepsilon_i)_{i=1}^n$  onto the space  $\mathcal{F}(\mathbf{X})^\perp$ . Then  $\hat{f}(\mathbf{X}) = f_o(\mathbf{X})$  and  $\hat{\varepsilon} = \varepsilon$  if, and only if,

$$\varepsilon \perp \mathcal{F}(\mathbf{X}). \tag{4}$$

**Violations of (4).** This assumption is often hard to justify. One can easily imagine situations in which it is violated: In the setting of hospitals, suppose that each tuple  $X_i$  includes a quantity  $X_i^{(1)}$  measuring average severity of cases treated in hospital  $i$ . Hospitals which tend to treat the more difficult cases could also hire particularly experienced or highly qualified personnel, despite the higher costs. Presumably this would result in a positive sample correlation for the pairs  $(X_i^{(1)}, \varepsilon_i)$ ,  $1 \leq i \leq n$ , a clear violation of (4) in the setting of multiple linear or quadratic regression. In the setting of postal service areas, suppose that  $X_i$  includes a measure  $X_i^{(1)}$  of population density. It may happen that service areas in low density regions take extra efforts to accelerate deliveries, trying to alleviate this unfavourable condition. This would result in a positive sample correlation for the pairs  $(X_i^{(1)}, \varepsilon_i)$  if  $Y_i$  stands for average delivery time, again a violation of (4). But in both settings the positive effect of the extra efforts would be eliminated by the regression method in that it appears in  $\hat{f}(\mathbf{X})$  rather than in  $\hat{\varepsilon}$ .

Alternatively, suppose that the first component  $x^{(1)}$  of a tuple  $x \in \mathcal{X}$  is a categorical variable with values in  $\{1, \dots, L\}$ . Thus the  $n$  units may be divided into  $L$  different categories. If this factor  $x^{(1)}$  is considered to have a potential impact on the units' performance, one should require  $\mathcal{F}$  to contain at least all functions

$$f(x) = \sum_{\ell=1}^L a_\ell 1_{[x^{(1)}=\ell]}$$

with real constants  $a_1, \dots, a_L$ . But then the residuals will automatically satisfy the equations

$$\sum_{i: X_i^{(1)}=\ell} \hat{\varepsilon}_i = 0 \quad \text{for } 1 \leq \ell \leq L.$$

It may happen that the units in a particular category  $\ell_o$  perform above or below average, meaning that  $\sum_{i: X_i^{(1)}=\ell_o} \varepsilon_i$  is strictly positive or negative, on top of potential differences

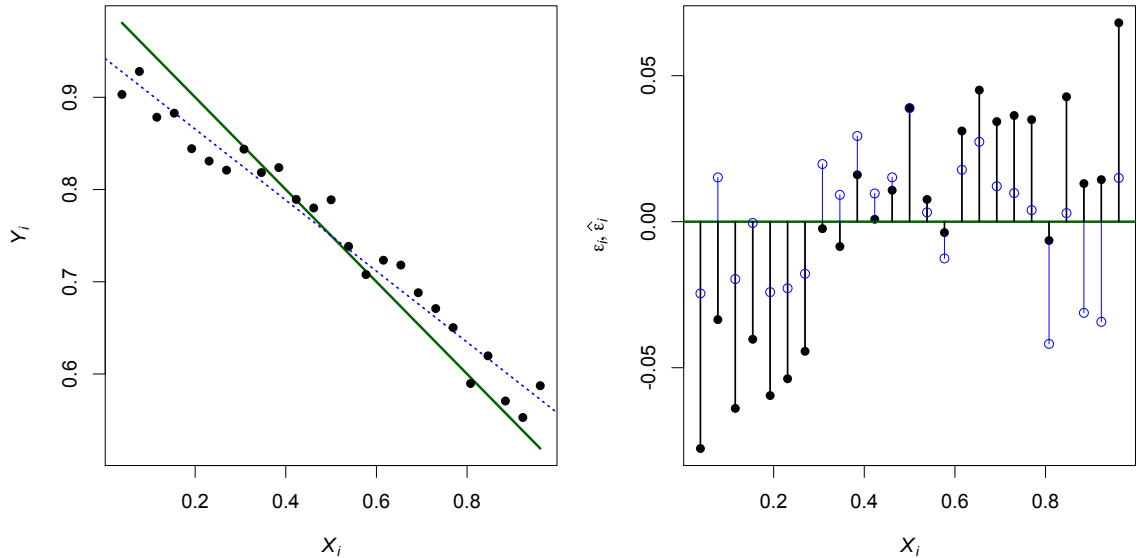


Figure 1: Simple example: Left panel: Raw data  $(X_i, Y_i)$  with  $f_o$  (line) and  $\hat{f}$  (dashed line). Right panel: Adjusted performances  $\varepsilon_i$  (bullets) and residuals  $\hat{\varepsilon}_i$  (circles).

between the sums  $\sum_{i: X_i^{(1)}=\ell} f_o(X_i)$ ,  $1 \leq \ell \leq L$ , which one intends to adjust for. But the regression residuals won't reflect such patterns.

**A simple example.** To illustrate these potential problems, we simulated  $n = 25$  observation pairs  $(X_i, Y_i)$  with univariate  $X_i$  such that  $f_o(x) = 1 - x/2$ . The sample correlation between  $\mathbf{X} = (X_i)_{i=1}^n$  and  $\mathbf{Y}$  equals  $-0.981$ , but  $\varepsilon$  and  $\mathbf{X} = (X_i)_{i=1}^n$  have sample correlation  $0.831$ . The left panel of Figure 1 shows the data pairs  $(X_i, Y_i)$ , the function  $f_o$  (thick line) as well as the fitted function  $\hat{f}(x) = \hat{a} + \hat{b}x$  (dashed line). Here  $\mathcal{F}$  was taken to be the set of all affine functions  $f(x) = a + bx$ . Due to the negative sample correlation between  $\mathbf{X}$  and  $\varepsilon$ , the estimated function  $\hat{f}$  with slope  $\hat{b} = -0.385$  is less steep than the true regression function  $f_o$  with slope  $-0.5$ .

In the right panel we see the adjusted performances  $\varepsilon_i$  and their estimators, i.e. the residuals  $\hat{\varepsilon}_i$ . For 20 observations, the signs of  $\varepsilon_i$  and  $\hat{\varepsilon}_i$  were identical, for the other 5 observations they differed! In Figure 2 one sees the pairs  $(R_i, \hat{R}_i)$ , where  $R_i$  is the rank of  $\varepsilon_i$  within  $\varepsilon$  while  $\hat{R}_i$  is either the rank of  $\hat{\varepsilon}_i$  within  $\hat{\varepsilon}$  (bullets) or of  $Y_i$  within  $\mathbf{Y}$  (circles). Obviously, the differences  $\hat{R}_i - R_i$  may be substantial. Replacing the raw values  $Y_i$  with the residuals  $\hat{\varepsilon}_i$  results in an improvement for 18 and no change for 3 observations; but for 4 observations the estimated rank gets worse.

**Partial performance.** The previous considerations and simple example illustrate the dilemma we have to deal with: If the sample correlation between  $f_o(\mathbf{X})$  and  $\mathbf{Y}$  is sub-

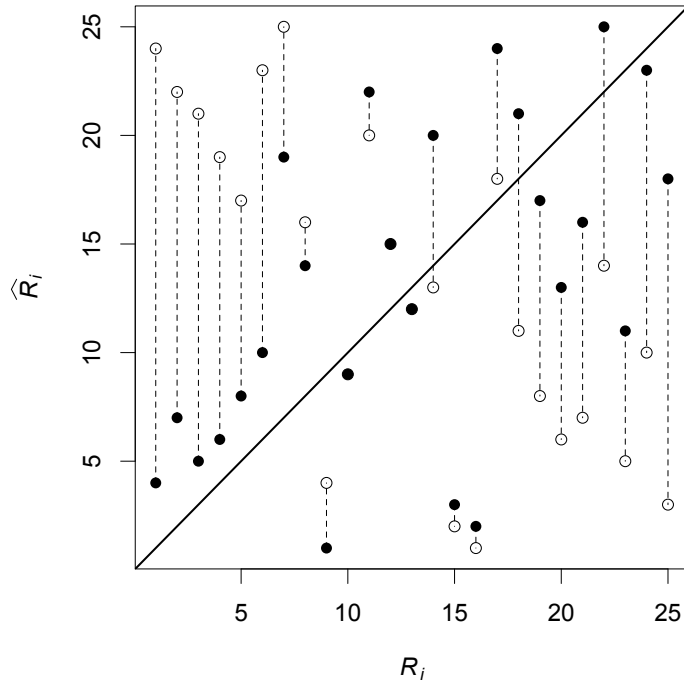


Figure 2: Simple example: Ranks  $R_i$  of the adjusted performances  $\varepsilon_i$  and estimated ranks  $\hat{R}_i$  based on residuals  $\hat{\varepsilon}_i$  (bullets) or raw data  $Y_i$  (circles).

stantial, it is better to estimate  $f_o$  than to ignore it. But then  $\hat{\varepsilon}$  measures only *partial performance*. That means, roughly saying, the vector  $\hat{\varepsilon}$  contains only those aspects of the adjusted performances which are not correlated with  $\mathbf{X}$ . Precisely,  $\hat{\varepsilon}$  is the orthogonal projection of  $\varepsilon$  onto the linear space  $\mathcal{F}(\mathbf{X})^\perp$ .

Whenever one uses  $\hat{\varepsilon}$  to quantify performance of the  $n$  units, one should keep in mind that substantial parts of the adjusted performances  $\varepsilon_i$  may be missing, and even the sign of  $\hat{\varepsilon}_i$  may differ from the sign of  $\varepsilon_i$ . This is particularly important whenever analyses of  $\hat{\varepsilon}$  are used to create incentives for better performance. If the different units are figuring out ways to improve performance, it may happen that self-evident steps are somehow related to  $\mathbf{X}$ . In that case, even if these measures are successful, the resulting improvements may disappear in the regression adjustment. Somehow this gives a new twist to the well-known phenomenon of ‘regression towards mediocrity’.

**Further ‘leakage’ due to misspecified models.** Even in case of (4), the estimator  $\hat{\varepsilon}$  may contain systematic errors due to model-misspecification. In general,

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

with  $\boldsymbol{\mu} = f_o(\mathbf{X})$  if the  $X_i \in \mathcal{X}$  contain all relevant covariables. Even then it may happen that  $f_o \notin \mathcal{F}$ . If we missed some important covariables  $X'_i \in \mathcal{X}'$ , then  $\boldsymbol{\mu} = (f_*(X_i, X'_i))_{i=1}^n$

with some regression function  $f_* : \mathcal{X} \times \mathcal{X}' \rightarrow \mathbb{R}$ . In both cases it is likely that  $\boldsymbol{\mu} \notin \mathcal{F}(\mathbf{X})$ , and

$$\begin{aligned}\widehat{\boldsymbol{\varepsilon}} &= (\mathbf{I} - \mathbf{H})\boldsymbol{\mu} + (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon} \\ &= (\mathbf{I} - \mathbf{H})\boldsymbol{\mu} + \boldsymbol{\varepsilon} \quad \text{under (4)}.\end{aligned}$$

Here  $\mathbf{I} \in \mathbb{R}^{n \times n}$  denotes the identity matrix, and  $\mathbf{H} \in \mathbb{R}^{n \times n}$  describes the orthogonal projection onto  $\mathcal{F}(\mathbf{X})$ , see also Section 3.

**Performances and random errors.** Note that we treated the vector  $\boldsymbol{\varepsilon}$  as fixed rather than random in the preceding considerations. Now one could think of

$$\varepsilon_i = \pi_i + \delta_i$$

with adjusted performances  $\pi_i$  and truly random errors  $\delta_i$  representing random fluctuations which are neither related to the  $X_i$  nor to the units' abilities or failures. Without more advanced sampling schemes and certain assumptions on the  $\pi_i$ , however, there is no possibility to estimate the latter quantities.

In traditional regression applications, people are mainly interested in  $f_o$  as a means to describe the relation between  $X_i$  and  $Y_i$  and to predict  $Y_o$  from  $X_o$  for future observations  $(X_o, Y_o)$  which are independent of the given pairs  $(X_i, Y_i)$ ,  $1 \leq i \leq n$ . Here the  $\varepsilon_i$  represent measurement or sampling errors which are modelled as random and considered to be an unavoidable nuisance. The residuals  $\widehat{\varepsilon}_i$  are only used to estimate certain properties of the random errors  $\varepsilon_i$  and to validate or falsify certain assumptions on these. Using regression as described previously to create ‘corrected values’  $\widehat{\varepsilon}_i$  as a surrogate for the raw values  $Y_i$  is not what the method is designed for; therefore the provocative word ‘abusing’ in the title of this report.

**Data envelopement analysis and quantile regression.** One should mention here an established method of benchmarking, called data envelopement analysis (DEA), initiated by Farrel (1957) and Charnes et al. (1978). Very roughly saying, in that approach one assumes that the  $\varepsilon_i$  are non-positive (if higher values of  $Y_i$  mean better performance), and  $f_o(x)$  is the maximally achievable performance under the circumstances described by  $x \in \mathcal{X}$ . The deviations  $\varepsilon_i$  are then estimated via a linear optimization method. The main reasons for using a regression approach rather than DEA seem to be the higher complexity of DEA, which makes it more difficult to communicate it to laymen, and the known sensitivity of DEA to errors in the data. Moreover, normal quantile-quantile plots of the residuals  $\widehat{\varepsilon}_i$  often indicate a Gaussian distribution, whereas the DEA paradigm would predict a non-symmetric, left-skewed distribution.

If the residuals show indeed a left-skewed distribution, a possible compromise between least squares regression and DEA would be regression quantiles (cf. Koenker and Bassett, 1978). That means, for a given parameter  $\gamma \in (0, 1)$  one determines a function  $\widehat{f}_\gamma$  minimising

$$\sum_{i=1}^n \rho_\gamma(Y_i - f(X_i))$$

over all  $f \in \mathcal{F}$ , where

$$\rho_\gamma(t) := (1_{[t \geq 0]} \gamma + 1_{[t \leq 0]} (1 - \gamma)) |t|$$

for  $t \in \mathbb{R}$ . An advantage of this approach would be that the estimator  $\widehat{f}_\gamma$  is less sensitive to outliers in the  $Y_i$  than the least squares estimator  $\widehat{f}$ . For  $\gamma$  close to one, the estimated function  $\widehat{f}_\gamma$  would be close to the envelope function  $f_o$ .

### 3 Ranking errors in a best case scenario

Let us set aside all reservations towards estimating  $f_o$  by the least squares estimator  $\widehat{f}$ . Rather than considering  $\varepsilon$  to be a fixed vector satisfying (4), let us assume that its components  $\varepsilon_i$  are independent and identically distributed random variables with centered Gaussian distribution,

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

with unknown standard deviation  $\sigma > 0$ . Under the additional assumption that  $f_o$  is contained in our model  $\mathcal{F}$ , the estimator  $\widehat{f}$  and the residuals  $\widehat{\varepsilon}_i$  are unbiased in the sense that

$$\mathbb{E} \widehat{f}(x) = f_o(x) \quad \text{and} \quad \mathbb{E}(\widehat{\varepsilon}_i - \varepsilon_i) = 0.$$

**Leverage.** Using the  $\widehat{\varepsilon}_i$  as surrogates for the  $\varepsilon_i$ , however, is problematic because of the well-known phenomenon of *leverage*: Observations  $(X_i, Y_i)$  with “exotic” part  $X_i$  tend to produce residuals  $\widehat{\varepsilon}_i$  with smaller modulus than the adjusted performances  $\varepsilon_i$ . Precisely, let  $\mathbf{H} \in \mathbb{R}^{n \times n}$  be the so-called hat matrix for the given model  $\mathcal{F}$ . That means, if  $f_1, f_2, \dots, f_p$  are basis functions of  $\mathcal{F}$ , and if the corresponding design matrix

$$\mathbf{D} = [f_1(\mathbf{X}), f_2(\mathbf{X}), \dots, f_p(\mathbf{X})] \in \mathbb{R}^{n \times p}$$

has full rank  $p < n$ , then

$$\mathbf{H} = \mathbf{D}(\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top$$

describes the orthogonal projection from  $\mathbb{R}^n$  onto the  $p$ -dimensional model space  $\mathcal{F}(\mathbf{X})$ . The matrix  $\mathbf{H}$  does not depend on the particular choice of basis functions  $f_1, \dots, f_p$  but



on the tuple  $\mathbf{X}$  of covariable vectors  $X_i$  and the model  $\mathcal{F}$ . Now

$$\widehat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}.$$

Elementary calculations and the fact that  $\mathbf{H}^\top = \mathbf{H} = \mathbf{H}^2$  show that

$$\mathbb{E}(\widehat{\varepsilon}_i^2) = (1 - H_{ii})\sigma^2 \quad \text{and} \quad \text{Corr}(\widehat{\varepsilon}_i, \varepsilon_i) = \sqrt{1 - H_{ii}}.$$

Thus the modulus of  $\widehat{\varepsilon}_i$  may be systematically too small and the correlation of  $\widehat{\varepsilon}_i$  and  $\varepsilon_i$  substantially smaller than 1 in case of high values of the leverage  $H_{ii} \in [0, 1]$  of the  $i$ -th observation part  $X_i$ . Note also that

$$\sum_{i=1}^p H_{ii} = p$$

whence

$$\max_{i=1,2,\dots,n} H_{ii} \geq \frac{p}{n}.$$

Thus a rather complex model with high dimension  $p$  relative to the sample size will automatically yield observations with high leverages  $H_{ii}$ . But even in case of  $p \ll n$  some leverages may be substantial.

**Ranking errors.** Suppose one uses the estimated performances  $\widehat{\varepsilon}_i$  to rank the units. That means, one computes the rank  $\widehat{R}_i$  of  $\widehat{\varepsilon}_i$  within  $\widehat{\boldsymbol{\varepsilon}}$  as a proxy for the rank  $R_i$  of  $\varepsilon_i$  within  $\boldsymbol{\varepsilon}$ ,

$$R_i = \sum_{j=1}^n 1_{[\varepsilon_j \leq \varepsilon_i]}.$$

In what follows we derive explicit expressions for the root mean squared ranking errors,

$$\sqrt{\mathbb{E}((\widehat{R}_i - R_i)^2)}.$$

First of all, under a mild regularity condition on the hat matrix  $\mathbf{H}$ , the residuals  $\widehat{\varepsilon}_i$  are pairwise different:

**Lemma 1.** *For arbitrary indices  $1 \leq i < j \leq n$ ,*

$$\mathbb{P}(\widehat{\varepsilon}_i = \widehat{\varepsilon}_j) = \begin{cases} 1 & \text{if } H_{ii} = H_{jj} = H_{ij} + 1, \\ 0 & \text{else.} \end{cases}$$

*The condition  $H_{ii} = H_{jj} = H_{ij} + 1$  implies that  $H_{ii} \geq 1/2$ .*

This lemma remains valid if the errors  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  are only assumed to be independent with continuous distributions. An immediate consequence of Lemma 1 is that the residuals

$\widehat{\varepsilon}_i$  are pairwise different almost surely, whenever  $H_{ii} \geq 1/2$  for at most one index  $i$ . In particular, with probability one,

$$\widehat{R}_i = \sum_{j=1}^n 1_{[\widehat{\varepsilon}_j \leq \widehat{\varepsilon}_i]}.$$

Here is a first main result about the ranks  $R_i$  and  $\widehat{R}_i$ :

**Theorem 1.** *Suppose that  $H_{ii} \geq 1/2$  for at most one index  $i \in \{1, 2, \dots, n\}$ . Then for arbitrary indices  $i, j \in \{1, 2, \dots, n\}$ ,*

$$\begin{aligned} \mathbb{E}((\widehat{R}_i - R_i)(\widehat{R}_j - R_j)) = & \\ & \frac{1}{2\pi} \sum_{k,\ell=1}^n \left( \arcsin\left(\frac{\Delta_{kl,ij}}{2}\right) + \arcsin\left(\frac{\Delta_{kl,ij} - H_{kl,ij}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,jj})}}\right) \right. \\ & \left. - \arcsin\left(\frac{\Delta_{kl,ij} - H_{kl,ij}}{\sqrt{2(2 - H_{kk,ii})}}\right) - \arcsin\left(\frac{\Delta_{kl,ij} - H_{kl,ij}}{\sqrt{2(2 - H_{\ell\ell,jj})}}\right) \right), \end{aligned}$$

where  $\Delta_{kl,ij} := \delta_{kl} + \delta_{ij} - \delta_{kj} - \delta_{il}$  and  $H_{kl,ij} := H_{kl} + H_{ij} - H_{kj} - H_{il}$ . In particular,

$$\begin{aligned} \mathbb{E}((\widehat{R}_i - R_i)^2) = & \frac{1}{2\pi} \sum_{k=1}^n \left( \pi - 2 \arccos\left(\sqrt{H_{kk,ii}/2}\right) \right) + \\ & \frac{1}{\pi} \sum_{1 \leq k < \ell \leq n} \left( \frac{\pi}{6} + \arcsin\left(\frac{1 - H_{kl,ii}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,ii})}}\right) \right. \\ & \left. - \arcsin\left(\frac{1 - H_{kl,ii}}{\sqrt{2(2 - H_{kk,ii})}}\right) - \arcsin\left(\frac{1 - H_{kl,ii}}{\sqrt{2(2 - H_{\ell\ell,ii})}}\right) \right). \end{aligned}$$

Here and throughout  $\delta_{st}$  denotes Kronecker's symbol,  $\delta_{st} = 1_{[s=t]}$ . Theorem 1 is useful for exact numerical calculations. Numerical experiments reveal also that the rank distortions are closely related to the leverages  $H_{ii}$ . Here is a theoretical result about the rank distortions in case of small maximal leverage:

**Theorem 2.** *Suppose that the column space of  $\mathbf{D}$  contains the constant vectors, i.e.  $\mathbf{H}\mathbf{1} = \mathbf{1} := (1)_{i=1}^n$ . Then, as  $\eta := \max_{i=1, \dots, n} H_{ii} \rightarrow 0$ ,*

$$\mathbb{E}((\widehat{R}_i - R_i)(\widehat{R}_j - R_j)) = \frac{n^2 H_{ij} - n}{2\sqrt{4 - \delta_{ij}} \pi} + O(n\eta^{1/2} + n^2\eta^2)$$

uniformly in  $i, j \in \{1, 2, \dots, n\}$ .

**A numerical example.** Suppose that  $n = 70$ , and let  $\mathbf{D}$  be equal to

$$\mathbf{D}^{(1)} := \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \text{or} \quad \mathbf{D}^{(2)} := \begin{bmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{bmatrix},$$

the design matrix for simple linear or quadratic regression, where  $X_1 < X_2 < \dots < X_n$  are equispaced numbers. The maximal leverage  $\max_i H_{ii}$  is equal to 0.0559 for  $\mathbf{D}^{(1)}$  and 0.1215 for  $\mathbf{D}^{(2)}$ . Figure 3 shows the root mean squared ranking errors  $\sqrt{\mathbb{E}((\hat{R}_i - R_i)^2)}$  for both cases. In addition the approximations  $n\sqrt{(H_{ii} - n^{-1})/(2\pi\sqrt{3})}$  and  $n\sqrt{H_{ii}/(2\pi\sqrt{3})}$  are shown as lines.

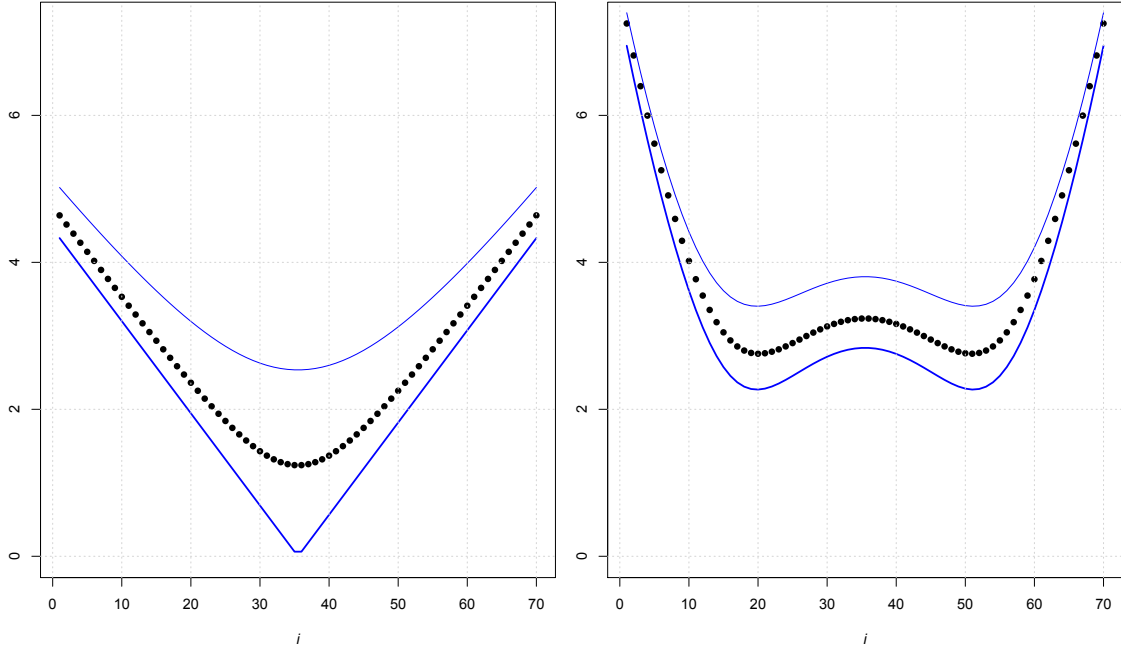


Figure 3: Root mean squared ranking errors for simple linear regression (left) and quadratic regression (right) with  $n = 70$  equispaced  $X$ -values.

**Heuristics.** Now we present heuristic arguments to approximate the rank distortions which also indicates what may happen in non-Gaussian settings. Presumably these arguments could be made rigorous by applying similar techniques and arguments as Koul (1969, 1992), Loynes (1980) and Mammen (1996). As in Theorem 2, asymptotic statements are meant as  $\eta = \max_{i=1, \dots, n} H_{ii} \rightarrow 0$ . We assume that the errors  $\varepsilon_1, \dots, \varepsilon_n$  are independent and identically distributed with finite standard deviation  $\sigma$  and distribution function  $F_\varepsilon$  with bounded and uniformly continuous density  $f_\varepsilon$ .

One can easily deduce from  $\mathbf{H}^\top = \mathbf{H} = \mathbf{H}^2$  that

$$\mathbb{E}((\mathbf{H}\varepsilon)_i(\mathbf{H}\varepsilon)_j) = \sigma^2 H_{ij}, \quad (5)$$

whereas the Cauchy-Schwarz inequality implies that

$$|\mathbb{E}((\mathbf{H}\varepsilon)_i(\mathbf{H}\varepsilon)_j)| \leq \text{Std}((\mathbf{H}\varepsilon)_i)\text{Std}((\mathbf{H}\varepsilon)_j) = \sigma^2 \sqrt{H_{ii}H_{jj}}.$$

Hence

$$|H_{ij}| \leq \eta \quad \text{for } 1 \leq i, j \leq n. \quad (6)$$

Pretending that the empirical c.d.f.  $\check{F}$  of the errors  $\varepsilon_i$  and the empirical c.d.f.  $\hat{F}$  of the residuals  $\hat{\varepsilon}_i$  are sufficiently close to  $F$ , we write

$$\begin{aligned} R_i &= n\check{F}(\varepsilon_i) \approx nF_\varepsilon(\varepsilon_i), \\ \hat{R}_i &= n\hat{F}(\hat{\varepsilon}_i) \approx nF_\varepsilon(\hat{\varepsilon}_i) = nF_\varepsilon(\varepsilon_i - (\mathbf{H}\boldsymbol{\varepsilon})_i). \end{aligned}$$

But  $(\mathbf{H}\boldsymbol{\varepsilon})_i$  is quite small, precisely,

$$\mathbb{E}((\mathbf{H}\boldsymbol{\varepsilon})_i^2) = \sigma^2 H_{ii} \leq \sigma^2 \eta$$

by (5). Hence we write

$$\hat{R}_i - R_i \approx -nf_\varepsilon(\varepsilon_i)(\mathbf{H}\boldsymbol{\varepsilon})_i.$$

Moreover, for  $i, j \in \{1, 2, \dots, n\}$  and  $\ell \in \{i, j\}$ ,

$$(\mathbf{H}\boldsymbol{\varepsilon})_\ell = \sum_{k=1}^n H_{\ell k} \varepsilon_k \approx \sum_{k \notin \{i, j\}} H_{\ell k} \varepsilon_k,$$

because  $\sum_{k \in \{i, j\}} H_{\ell k} \varepsilon_k$  is very small in the sense that

$$\mathbb{E}\left(\left(\sum_{k \in \{i, j\}} H_{\ell k} \varepsilon_k\right)^2\right) = \sigma^2 \sum_{k \in \{i, j\}} H_{\ell k}^2 \leq 2\sigma^2 \eta^2$$

by (6). Thus we pretend that the random pairs  $(\varepsilon_i, \varepsilon_j)$  and  $((\mathbf{H}\boldsymbol{\varepsilon})_i, (\mathbf{H}\boldsymbol{\varepsilon})_j)$  are stochastically independent and conjecture that

$$\begin{aligned} \mathbb{E}((\hat{R}_i - R_i)(\hat{R}_j - R_j)) &\approx n^2 \mathbb{E}(f_\varepsilon(\varepsilon_i)f_\varepsilon(\varepsilon_j)(\mathbf{H}\boldsymbol{\varepsilon})_i(\mathbf{H}\boldsymbol{\varepsilon})_j) \\ &\approx n^2 \mathbb{E}(f_\varepsilon(\varepsilon_i)f_\varepsilon(\varepsilon_j)) \mathbb{E}((\mathbf{H}\boldsymbol{\varepsilon})_i(\mathbf{H}\boldsymbol{\varepsilon})_j) \\ &= n^2 \sigma^2 \mathbb{E}(f_\varepsilon(\varepsilon_i)f_\varepsilon(\varepsilon_j)) H_{ij}. \end{aligned} \quad (7)$$

Now consider the special case of  $F = \Phi(\sigma^{-1}\cdot)$  and  $f = \sigma^{-1}\phi(\sigma^{-1}\cdot)$  with the standard Gaussian c.d.f.  $\Phi$  and density  $\phi$ . For  $i \neq j$ ,

$$\begin{aligned} \sigma^2 \mathbb{E}(f_\varepsilon(\varepsilon_i)f_\varepsilon(\varepsilon_j)) &= (\sigma \mathbb{E}f_\varepsilon(\varepsilon_i))^2 \\ &= \left(\sigma^{-1} \int \phi(\sigma^{-1}x)^2 dx\right)^2 = \left((2\pi)^{-1/2} \int \phi(\sqrt{2}x) dx\right)^2 \\ &= (4\pi)^{-1}, \end{aligned}$$

and

$$\begin{aligned} \sigma^2 \mathbb{E}(f_\varepsilon(\varepsilon_i)^2) &= \sigma^{-1} \int \phi(\sigma^{-1}x)^3 dx = (2\pi)^{-1} \int \phi(\sqrt{3}x) dx \\ &= (2\sqrt{3}\pi)^{-1}. \end{aligned}$$

Hence the conjectured approximation (7) equals

$$\frac{n^2 H_{ij}}{2\sqrt{4 - \delta_{ij}} \pi}.$$

## 4 Alternative methods for case-by-case data

As indicated in the introduction, suppose that for each unit  $i$  we have several observations ('cases')

$$(X_{ij}, Y_{ij}), \quad 1 \leq j \leq J_i.$$

For instance, in the example of hospitals, we may have data of  $J_i$  patients or treatments in hospital  $i$ , and the tuples  $X_{ij}$  may be case-specific rather than hospital-specific. In the example of service areas of a postal service, cases could be items to be delivered with the tuple  $X_{ij}$  describing the type, size and weight of the item and characteristics of the receiver or the neighborhood he or she is living in. A potential problem, though, would be the determination of the single delivery times  $Y_{ij}$ .

Then a possible alternative to model equation (1) is given by (2), i.e.

$$Y_{ij} = f_o(X_{ij}) + a_i + \epsilon_{ij}$$

with a regression function  $f_o : \mathcal{X} \rightarrow \mathbb{R}$  as before, unit-specific parameters  $a_i$  measuring their performances and random errors  $\epsilon_{ij}$  with mean zero. Identifiability can be achieved by requiring that

$$\sum_{i=1}^n \sum_{j=1}^{J_i} f_o(X_{ij}) = 0.$$

Then the parameter  $a_i$  would be the expected performance of unit  $i$  if it was dealing with a randomly chosen case from all  $J_+ := \sum_{i=1}^n J_i$  cases.

For instance, if any  $x \in \mathcal{X}$  is a tuple  $(x^{(k)})_{k=1}^K$  of  $K$  numerical or  $\{0, 1\}$ -valued covariables, one could think about multiple linear regression:

$$f_o(x) := \sum_{k=1}^K \beta_k (x^{(k)} - \bar{X}^{(k)})$$

with the overall averages

$$\bar{X}^{(k)} := \frac{1}{J_+} \sum_{i=1}^n \sum_{j=1}^{J_i} X_{ij}^{(k)}$$

and certain real parameters  $\beta_k$ . Alternatively one could consider multiple quadratic regression:

$$f_o(x) := \sum_{k=1}^K \beta_k (x^{(k)} - \bar{X}^{(k)}) + \sum_{k=1}^K \sum_{\ell=k}^K \beta_{k,\ell} (x^{(k)} x^{(\ell)} - \bar{X}^{(k,\ell)})$$

with the average  $\bar{X}^{(k,\ell)}$  of all  $J_+$  products  $X_{ij}^{(k)} X_{ij}^{(\ell)}$  and certain real parameters  $\beta_k, \beta_{k,\ell}$ .

An important restriction is that we do not allow for interactions between units and the  $K$  covariables. That means, the impact of the  $K$  covariables is the same for all  $n$  units. Without such an assumption we would encounter similar identifiability problems as in the simpler regression setting discussed before.

Now the  $\varepsilon_{ij}$  are really considered as random errors, and the performance parameters  $a_i$  may be estimated via least squares estimators  $\hat{a}_i$ . In addition, one could determine standard errors for these estimators and single or simultaneous confidence bounds for the underlying parameters  $a_i$ . Depending on standard residual diagnostics, the latter could be based on a generalisation of Tukey's method for linear models with homoscedastic Gaussian errors. Alternatively, if homoscedasticity is not plausible, one could apply a suitable variant of the wild bootstrap (Mammen 1993); see the lecture notes of Dümbgen (2015) for more details.

## 5 Technical arguments

**Proof of Lemma 1.** We write  $\hat{\varepsilon} = \mathbf{G}\varepsilon$  with the companion hat matrix  $\mathbf{G} := \mathbf{I} - \mathbf{H}$  describing the orthogonal projection on  $\mathcal{F}(\mathbf{X})^\perp$ . Since

$$\hat{\varepsilon}_i - \hat{\varepsilon}_j = \sum_{k=1}^n (G_{ik} - G_{jk})\varepsilon_k,$$

we may conclude that  $\hat{\varepsilon}_i - \hat{\varepsilon}_j$  has a continuous distribution whenever  $G_{ik} \neq G_{jk}$  for at least one index  $k$ , and this implies that  $\mathbb{P}(\hat{\varepsilon}_i = \hat{\varepsilon}_j) = 0$ . Otherwise,

$$G_{ik} = G_{jk} \quad \text{for } k = 1, 2, \dots, n, \tag{8}$$

and  $\mathbb{P}(\hat{\varepsilon}_i = \hat{\varepsilon}_j) = 1$ . Since  $\mathbf{G}^\top = \mathbf{G}$ , condition (8) implies that

$$G_{ii} = G_{jj} = G_{ij}. \tag{9}$$

On the other hand, since  $\mathbf{G}^\top \mathbf{G} = \mathbf{G}$ ,

$$\sum_{k=1}^n (G_{ik} - G_{jk})^2 = G_{ii} + G_{jj} - 2G_{ij}.$$

Thus conditions (8) and (9) are equivalent. Since  $\mathbf{H} = \mathbf{I} - \mathbf{G}$ , condition (9) is equivalent to

$$H_{ii} = H_{jj} = 1 + H_{ij}. \tag{10}$$

Finally, denoting the columns of  $\mathbf{H}$  with  $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n$ , it follows from  $\mathbf{H}^\top \mathbf{H} = \mathbf{H}$  and the Cauchy-Schwarz inequality that  $|H_{ij}| = |\mathbf{h}_i^\top \mathbf{h}_j| \leq \|\mathbf{h}_i\| \|\mathbf{h}_j\| = \sqrt{H_{ii} H_{jj}}$ . Thus condition (10) implies that  $H_{ii} \geq 1 - |H_{ij}| \geq 1 - \sqrt{H_{ii} H_{jj}} = 1 - H_{ii}$ , whence  $H_{ii} \geq 1/2$ .  $\square$

A key ingredient for the proof of Theorem 1 is an elementary equality for bivariate Gaussian distributions which is well-known in the literature on robust correlation measures. For the reader's convenience we include a proof.

**Lemma 2.** *Let  $\mathbf{Y}$  be a random vector with distribution  $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ , where  $\Sigma_{11}, \Sigma_{22} > 0$ . Then*

$$\mathbb{P}(Y_1 \leq 0 \text{ and } Y_2 \leq 0) = \frac{\pi/2 + \arcsin(\rho)}{2\pi} \quad \text{with } \rho := \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}.$$

**Proof of Lemma 2.** Since the probability in question does not change when we replace  $Y_i$  with  $\Sigma_{ii}^{-1/2}Y_i$ , we may assume without loss of generality that  $\Sigma_{11} = \Sigma_{22} = 1$  and  $\Sigma_{12} = \rho$ . If  $\mathbf{Z}$  denotes a random vector with standard Gaussian distribution on  $\mathbb{R}^2$ , then  $\mathbf{Y}$  has the same distribution as  $[Z_1, \rho Z_1 + \bar{\rho} Z_2]^\top$ , where  $\bar{\rho} := \sqrt{1 - \rho^2}$ . Now we write  $\rho = \sin(\alpha)$  and  $\bar{\rho} = \cos(\alpha)$  with  $\alpha := \arcsin(\rho) \in [-\pi/2, \pi/2]$ , and  $\mathbf{Z} = [R \cos(\Theta), R \sin(\Theta)]^\top$ , where  $R := \|\mathbf{Z}\| > 0$  almost surely, and  $\Theta$  is uniformly distributed on  $[0, 2\pi]$ . Then

$$\begin{aligned} \mathbb{P}(Y_1 \leq 0 \text{ and } Y_2 \leq 0) &= \mathbb{P}(\cos(\Theta) \leq 0 \text{ and } \sin(\alpha) \cos(\Theta) + \cos(\alpha) \sin(\Theta) \leq 0) \\ &= \mathbb{P}(\cos(\Theta) \leq 0 \text{ and } \sin(\alpha + \Theta) \leq 0) \\ &= \mathbb{P}(\Theta \in [\pi/2, 3\pi/2] \text{ and } \alpha + \Theta \in [\pi, 2\pi] + 2\pi\mathbb{Z}) \\ &= \mathbb{P}(\Theta \in [\pi - \alpha, 3\pi/2]) \\ &= \frac{\pi/2 + \alpha}{2\pi}. \end{aligned} \quad \square$$

**Proof of Theorem 1.** According to Lemma 1,

$$\begin{aligned} R_i &= 1 + \sum_{k \neq i} 1_{[\varepsilon_k \leq \varepsilon_i]} = 1 + \sum_{k \neq i} 1_{[\mathbf{a}_{ki}^\top \boldsymbol{\varepsilon} \leq 0]} \quad \text{and} \\ \widehat{R}_i &= 1 + \sum_{k \neq i} 1_{[\widehat{\varepsilon}_k \leq \widehat{\varepsilon}_i]} = 1 + \sum_{k \neq i} 1_{[\widehat{\mathbf{a}}_{ki}^\top \boldsymbol{\varepsilon} \leq 0]} \end{aligned}$$

almost surely, where  $\mathbf{a}_{ki} := \mathbf{e}_k - \mathbf{e}_i$  and  $\widehat{\mathbf{a}}_{ki} := \mathbf{G}(\mathbf{e}_k - \mathbf{e}_i)$  with the standard basis  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$  of  $\mathbb{R}^n$ . Consequently it follows from Lemma 2 that

$$\begin{aligned} &\mathbb{E}((\widehat{R}_i - R_i)(\widehat{R}_j - R_j)) \\ &= \sum_{k \neq i, \ell \neq j} \mathbb{E}(1_{[\mathbf{a}_{ki}^\top \boldsymbol{\varepsilon} \leq 0]} - 1_{[\widehat{\mathbf{a}}_{ki}^\top \boldsymbol{\varepsilon} \leq 0]})(1_{[\mathbf{a}_{\ell j}^\top \boldsymbol{\varepsilon} \leq 0]} - 1_{[\widehat{\mathbf{a}}_{\ell j}^\top \boldsymbol{\varepsilon} \leq 0]}) \\ &= \sum_{k \neq i, \ell \neq j} \left( \mathbb{P}(\mathbf{a}_{ki}^\top \boldsymbol{\varepsilon} \leq 0, \mathbf{a}_{\ell j}^\top \boldsymbol{\varepsilon} \leq 0) + \mathbb{P}(\widehat{\mathbf{a}}_{ki}^\top \boldsymbol{\varepsilon} \leq 0, \widehat{\mathbf{a}}_{\ell j}^\top \boldsymbol{\varepsilon} \leq 0) \right. \\ &\quad \left. - \mathbb{P}(\mathbf{a}_{ki}^\top \boldsymbol{\varepsilon} \leq 0, \widehat{\mathbf{a}}_{\ell j}^\top \boldsymbol{\varepsilon} \leq 0) - \mathbb{P}(\widehat{\mathbf{a}}_{ki}^\top \boldsymbol{\varepsilon} \leq 0, \mathbf{a}_{\ell j}^\top \boldsymbol{\varepsilon} \leq 0) \right) \\ &= \frac{1}{2\pi} \sum_{k \neq i, \ell \neq j} \left( \arcsin(\cos(\mathbf{a}_{ki}, \mathbf{a}_{\ell j})) + \arcsin(\cos(\widehat{\mathbf{a}}_{ki}, \widehat{\mathbf{a}}_{\ell j})) \right. \\ &\quad \left. - \arcsin(\cos(\mathbf{a}_{ki}, \widehat{\mathbf{a}}_{\ell j})) - \arcsin(\cos(\widehat{\mathbf{a}}_{ki}, \mathbf{a}_{\ell j})) \right), \end{aligned}$$

where

$$\cos(\mathbf{v}, \mathbf{w}) := \frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\| \|\mathbf{w}\|} \quad \text{for } \mathbf{v}, \mathbf{w} \in \mathbb{R}^n \setminus \{\mathbf{0}\}.$$

Note that

$$\mathbf{a}_{ki}^\top \mathbf{a}_{\ell j} = (\mathbf{e}_k - \mathbf{e}_i)^\top (\mathbf{e}_\ell - \mathbf{e}_j) = \delta_{k\ell} + \delta_{ij} - \delta_{kj} - \delta_{i\ell} =: \Delta_{k\ell,ij},$$

and with

$$H_{k\ell,ij} := (\mathbf{e}_k - \mathbf{e}_i)^\top \mathbf{H} (\mathbf{e}_\ell - \mathbf{e}_j) = H_{k\ell} + H_{ij} - H_{kj} - H_{i\ell},$$

we may write

$$\begin{aligned} \mathbf{a}_{ki}^\top \widehat{\mathbf{a}}_{\ell j} &= (\mathbf{e}_k - \mathbf{e}_i)^\top \mathbf{G} (\mathbf{e}_\ell - \mathbf{e}_j) = G_{k\ell} + G_{ij} - G_{kj} - G_{i\ell} \\ &= \Delta_{k\ell,ij} - H_{k\ell,ij}, \\ \widehat{\mathbf{a}}_{ki}^\top \widehat{\mathbf{a}}_{\ell j} &= (\mathbf{e}_k - \mathbf{e}_i)^\top \mathbf{G}^\top \mathbf{G} (\mathbf{e}_\ell - \mathbf{e}_j) = (\mathbf{e}_k - \mathbf{e}_i)^\top \mathbf{G} (\mathbf{e}_\ell - \mathbf{e}_j) \\ &= \Delta_{k\ell,ij} - H_{k\ell,ij}. \end{aligned}$$

Hence we obtain the formula

$$\begin{aligned} &\mathbb{E}((\widehat{R}_i - R_i)(\widehat{R}_j - R_j)) \\ &= \frac{1}{2\pi} \sum_{k \neq i, \ell \neq j} \left( \arcsin\left(\frac{\Delta_{k\ell,ij}}{2}\right) + \arcsin\left(\frac{\Delta_{k\ell,ij} - H_{k\ell,ij}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,jj})}}\right) \right. \\ &\quad \left. - \arcsin\left(\frac{\Delta_{k\ell,ij} - H_{k\ell,ij}}{\sqrt{2(2 - H_{\ell\ell,jj})}}\right) - \arcsin\left(\frac{\Delta_{k\ell,ij} - H_{k\ell,ij}}{\sqrt{2(2 - H_{kk,ii})}}\right) \right). \end{aligned}$$

But the restriction to indices  $k \neq i$  and  $\ell \neq j$  is superfluous, because  $\Delta_{k\ell,ij} = H_{k\ell,ij} = 0$  whenever  $k = i$  or  $\ell = j$ . This yields the first asserted formula.

In the special case of  $i = j$ , note that  $\Delta_{k\ell,ii} = 1 + \delta_{k\ell}$  if  $i \notin \{k, \ell\}$ . If we replace  $\Delta_{k\ell,ii}$  with  $1 + \delta_{k\ell}$  in our formula for  $\mathbb{E}((\widehat{R}_i - R_i)^2)$ , we end up with the expression

$$\begin{aligned} &\frac{1}{2\pi} \sum_{k, \ell=1}^n \left( \arcsin\left(\frac{1 + \delta_{k\ell}}{2}\right) + \arcsin\left(\frac{1 + \delta_{k\ell} - H_{k\ell,ii}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,ii})}}\right) \right. \\ &\quad \left. - \arcsin\left(\frac{1 + \delta_{k\ell} - H_{k\ell,ii}}{\sqrt{2(2 - H_{\ell\ell,ii})}}\right) - \arcsin\left(\frac{1 + \delta_{k\ell} - H_{k\ell,ii}}{\sqrt{2(2 - H_{kk,ii})}}\right) \right). \end{aligned}$$

But for  $k = i$  or  $\ell = i$  the corresponding summands are equal to zero, because  $k = i$  implies that  $H_{k\ell,ii} = H_{kk,ii} = 0$ , and  $\ell = i$  implies that  $H_{k\ell,ii} = H_{\ell\ell,ii} = 0$ . Distinguishing the cases  $k = \ell$  and  $k \neq \ell$  yields

$$\begin{aligned} \mathbb{E}((\widehat{R}_i - R_i)^2) &= \frac{1}{2\pi} \sum_{k=1}^n \left( \pi - 2 \arcsin\left(\sqrt{1 - H_{kk,ii}/2}\right) \right) + \\ &\quad \frac{1}{\pi} \sum_{1 \leq k < \ell \leq n} \left( \frac{\pi}{6} + \arcsin\left(\frac{1 - H_{k\ell,ii}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,ii})}}\right) \right. \\ &\quad \left. - \arcsin\left(\frac{1 - H_{k\ell,ii}}{\sqrt{2(2 - H_{\ell\ell,ii})}}\right) - \arcsin\left(\frac{1 - H_{k\ell,ii}}{\sqrt{2(2 - H_{kk,ii})}}\right) \right). \end{aligned}$$



Finally the assertion follows from the well-known fact that  $\arcsin(\sqrt{1-t}) = \arccos(\sqrt{t})$  for  $0 \leq t \leq 1$ .  $\square$

**Proof of Theorem 2.** First recall that  $|H_{k\ell}| \leq \sqrt{H_{kk}H_{\ell\ell}} \leq \eta$ , whence  $|H_{k\ell,ij}| \leq 4\eta$ . Furthermore,  $\Delta_{k\ell,ij} = \delta_{ij}$  whenever  $\{k, \ell\} \cap \{i, j\} = \emptyset$  and  $k \neq \ell$ , i.e.  $\Delta_{k\ell,ij} \neq \delta_{ij}$  for at most  $n+2$  index pairs  $(k, \ell)$ . Elementary calculus shows that

$$|\arcsin(x) - \arcsin(y)| \leq C\sqrt{|x-y|}$$

for some constant  $C$ , the optimal one being  $\pi/\sqrt{2}$ . Hence

$$\left| \arcsin\left(\frac{d}{2}\right) + \arcsin\left(\frac{d - H_{k\ell,ij}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,jj})}}\right) - \arcsin\left(\frac{d - H_{k\ell,ij}}{\sqrt{2(2 - H_{kk,ii})}}\right) - \arcsin\left(\frac{d - H_{k\ell,ij}}{\sqrt{2(2 - H_{\ell\ell,jj})}}\right) \right| = O(\eta^{1/2})$$

uniformly in  $k, \ell, i, j \in \{1, 2, \dots, n\}$  and  $d \in \{\Delta_{k\ell,ij}, \delta_{ij}\}$ . Consequently,

$$\begin{aligned} & \mathbb{E}((\widehat{R}_i - R_i)(\widehat{R}_j - R_j)) \\ &= \frac{1}{2\pi} \sum_{k,\ell=1}^n \left( \arcsin\left(\frac{\delta_{ij}}{2}\right) + \arcsin\left(\frac{\delta_{ij} - H_{k\ell,ij}}{\sqrt{(2 - H_{kk,ii})(2 - H_{\ell\ell,jj})}}\right) - \arcsin\left(\frac{\delta_{ij} - H_{k\ell,ij}}{\sqrt{2(2 - H_{kk,ii})}}\right) - \arcsin\left(\frac{\delta_{ij} - H_{k\ell,ij}}{\sqrt{2(2 - H_{\ell\ell,jj})}}\right) \right) + O(n\eta^{1/2}) \end{aligned}$$

uniformly in  $i, j \in \{1, 2, \dots, n\}$ . But for  $d \in [0, 1]$  and  $x, y, z \in [-4\eta, 4\eta]$ ,

$$\begin{aligned} & \arcsin\left(\frac{d}{2}\right) + \arcsin\left(\frac{d-x}{\sqrt{(2-y)(2-z)}}\right) - \arcsin\left(\frac{d-x}{\sqrt{2(2-y)}}\right) - \arcsin\left(\frac{d-x}{\sqrt{2(2-z)}}\right) \\ &= \arcsin\left(\frac{d}{2}\right) - \arcsin\left(\frac{d-x}{2}\right) \\ & \quad + \arcsin\left(\frac{d-x}{2}\right) - \arcsin\left(\frac{d-x}{\sqrt{2(2-y)}}\right) \\ & \quad + \arcsin\left(\frac{d-x}{\sqrt{(2-y)(2-z)}}\right) - \arcsin\left(\frac{d-x}{\sqrt{2(2-z)}}\right) \\ &= \arcsin\left(\frac{d}{2}\right) - \arcsin\left(\frac{d-x}{2}\right) \\ & \quad + \arcsin\left(\frac{d-x}{2}\right) - \arcsin\left(\frac{d-x}{2}\left(1 + \frac{y}{4} + O(\eta^2)\right)\right) \\ & \quad + \arcsin\left(\frac{d-x}{\sqrt{2(2-z)}}\left(1 + \frac{y}{4} + O(\eta^2)\right)\right) - \arcsin\left(\frac{d-x}{\sqrt{2(2-z)}}\right) \\ &= \arcsin'\left(\frac{d}{2} + O(\eta)\right)\frac{x}{2} \\ & \quad + \arcsin'\left(\frac{d}{2} + O(\eta)\right)\left(\frac{-dy}{8} + O(\eta^2)\right) + \arcsin'\left(\frac{d}{2} + O(\eta)\right)\left(\frac{dy}{8} + O(\eta^2)\right) \\ &= \frac{x}{\sqrt{4-d^2}} + O(\eta^2). \end{aligned}$$

Consequently,

$$\mathbb{E}((\widehat{R}_i - R_i)(\widehat{R}_j - R_j)) = \frac{1}{2\pi\sqrt{4 - \delta_{ij}}} \sum_{k,\ell=1}^n H_{k\ell,ij} + O(n\eta^{1/2} + n^2\eta^2).$$

Finally it follows from  $\mathbf{H}\mathbf{1} = \mathbf{H}^\top\mathbf{1} = \mathbf{1}$  that

$$\sum_{k,\ell=1}^n H_{k\ell,ij} = \sum_{k,\ell=1}^n (H_{k\ell} + H_{ij} - H_{kj} - H_{i\ell}) = n^2 H_{ij} - n. \quad \square$$

**Acknowledgement.** Constructive comments of Werner Stahel, Anja Mühlemann and Christof Strähl are gratefully acknowledged.

## References

- [1] A. CHARNES, W.W. COOPER and E. RHODES (1978). Measuring the efficiency of decision making units. *Europ. J. Oper. Res.* **2**, 429–444.
- [2] L. DÜMBGEN (2015). *Lineare Modelle und Regression*. Lecture notes, University of Bern.
- [3] M.J. FARRELL (1957). The measurement of productive efficiency. *J. Royal Statist. Soc. A* **120**, 253–281.
- [4] R. KOENKER and G. BASSET (1978). Regression quantiles. *Econometrica* **46**(1), 33–50.
- [5] H. KOUL (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.* **40**, 1950–1979.
- [6] H. KOUL (1992). *Weighted Empiricals and Linear Models*. IMS, Hayward, CA.
- [7] R.M. LOYNES (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.* **8**, 285–298.
- [8] E. MAMMEN (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21**(1), 255–285.
- [9] E. MAMMEN (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* **24**, 307–335.
- [10] T.P. RYAN (1997). *Modern Regression Methods*. Wiley, New York.