Festschrift for Tarmo Pukkila on his 60th Birthday
Eds. E. P. Liski, J. Isotalo, J. Niemelä, S. Puntanen, and G. P. H. Styan
© Dept. of Mathematics, Statistics and Philosophy,
Univ. of Tampere, 2006, ISBN 978-951-44-6620-5, pages 243–259

On the role of the constant term in linear regression

Jarkko Isotalo Simo Puntanen George P. H. Styan

Abstract. In this paper we comment on the role of the constant term in linear regression, with particular emphasis in teaching statistics. The constant term corresponds to a variable whose observed values are all identical and hence its variance is zero. Hence students might wonder if such a variable is indeed a proper variable. We go through some important geometric considerations and comment on various models and the role of the constant term therein. A numerical example run in the Survo computing environment, see e.g., Mustonen (2001), illustrates our comments.

2000 MSC codes: 62J05, 62H12, 62H20.

Key words and phrases: BLUE; Centering; Coefficient of determination; Collinearity; OLSE; Orthogonal projector; R-squared.

1 ...Life gets troublesome

Let us begin by quoting the following e-mail correspondence which took place in May 1996, almost 10 years ago. (The mathematical symbols are coded in LaTeX.)

```
Date: Wed, 15 May 1996
From: "Simo Puntanen" <sjp@uta.fi>
To: "David A. Belsley" <belsley@bc.edu>
Subject: high correlation
```

> Dear Professor Belsley,
>
> I'd like to bother you with the following simple question: on page 20 of your book, you say that "... a high correlation surely implies a low angle..."
>
> I'm afraid that I have misunderstood something since it is easy to find an example where correlation is high but the angle is not low. Take $\mathbf{x} = (1, 1, 1.01)'$, $\mathbf{y} = (1, 0, -1)'$, and then $\mathrm{cor}(x, y) = -.866$ but $\cos(x, y) = -.004$. Clearly I have missed something?
>
> With best regards, Simo P.

```
Date: Wed, 15 May 1996 11:06:40 -0400 (EDT)
From: "David A. Belsley" <belsley@bc.edu>
```

Dear Simo P.,

Ah, you must be a mathematician or a statistician to think that $-.866$ is a high correlation. In the world of collinearity, that is pretty good; life doesn't get troublesome until the correlations are at least .96. And what I mean by a "high" correlation is really more like .99 or higher. You will note that this notion is made very clear as the monograph progresses.

Best wishes, David A. Belsley

```
Date: Wed, 15 May 1996 20:18:23 -0400 (EDT)
From: "David A. Belsley" <belsley@bc.edu>
```

Dear Simo P.,

I was entirely too quick and too flip in my previous answer to you. You are absolutely correct. It is quite possible for two variates to be perfectly correlated, yet also be orthogonal, i.e., perfectly conditioned. Your example comes very close. Letting $\mathbf{x} = (1, 1, \sqrt{2})'$ and $\mathbf{y} = (-1, -1, \sqrt{2})'$ comes right on the button. These have correlation 1 and cos 0.

I was also too quick in my book, giving away more than need be by following conventional wisdom, which, in this case, is wrong. It is indeed clear that correlation and conditioning are neither necessary nor sufficient to one another – which wonderfully strengthens my argument. For it is equally clear that two variates, such as exemplified above, do very well as regressands (without the constant term), even though they are perfectly correlated. [Note, however, that in situations like this, if you include the intercept (a column of 1s), the data become ill-conditioned and not suitable for regression.]

I thank you for pointing this out; strangely you are the first to do so. I know why I failed to pursue this issue further, because it was already clear that high correlation is not a suitable diagnostic for conditioning, and your example furthers this point.

Sincerely, David A. Belsley

2 Numerical illustration through Survo

The correspondence above is a lesson in itself. It illustrates how carefully we should proceed in the world of collinearity.

While teaching concepts related to regression and correlation, it is absolutely necessary to illustrate these concepts using geometric arguments. Students should realize that the sample correlation coefficient is the cosine between two specific vectors: these vectors are the centered values of corresponding variables. It is essential to observe that these vectors must be centered, or in geometric terms: the original vectors must be projected onto the plane which is orthogonal to the vector of ones; we denote this vector as $\mathbf{1}$ (or $\mathbf{1}_n$).

It is precisely due to this centering requirement that the concept of high correlation (absolute value) and high cosine (low angle) do not go hand-in-hand.

In his excellent book, Professor Belsley (1991, p. 20) has an illustrative example of the situation where the correlation is zero but the cosine is very close to 1. We go through this example using Survo, a statistical software developed by Professor Seppo Mustonen, see, e.g., Mustonen (2001). The calculations are shown in the example below.

The fundamental concept in Survo is the edit field. The user works with Survo by typing text in the edit field and by activating various commands and operations within the text; in the Example, these activated operations (except lines 11 and 12) are emphasized as a white text in a black background.

Between lines 20 and 24, we define a 3×2 matrix $\mathbf{B}$. Then, a 3×4 matrix $\mathbf{A}$ is created so that the first two columns of $\mathbf{A}$ comprise $\mathbf{B}$, while (in view of line 28) the last two columns of $\mathbf{A}$ include the values of variables U and V; they are functions of variables X and Y and a real number *eps*.

When line 36 is activated, the correlation matrix of all four variables will be calculated. We see at once that the correlation between U and V is zero. On line 45 we have calculated the cosines. In this situation $\cos(U, V)$ is extremely close to 1 (even though these variables are uncorrelated).

On line 54, the cosine $\cos(U, V)$ is expressed as a function of *eps*. We immediately confirm that $\cos(U, V)$ can go through all values when *eps* varies, but at the same time $\operatorname{cor}(U, V) = 0$ (except when *eps* $= 0$; then cor $= 0/0$).

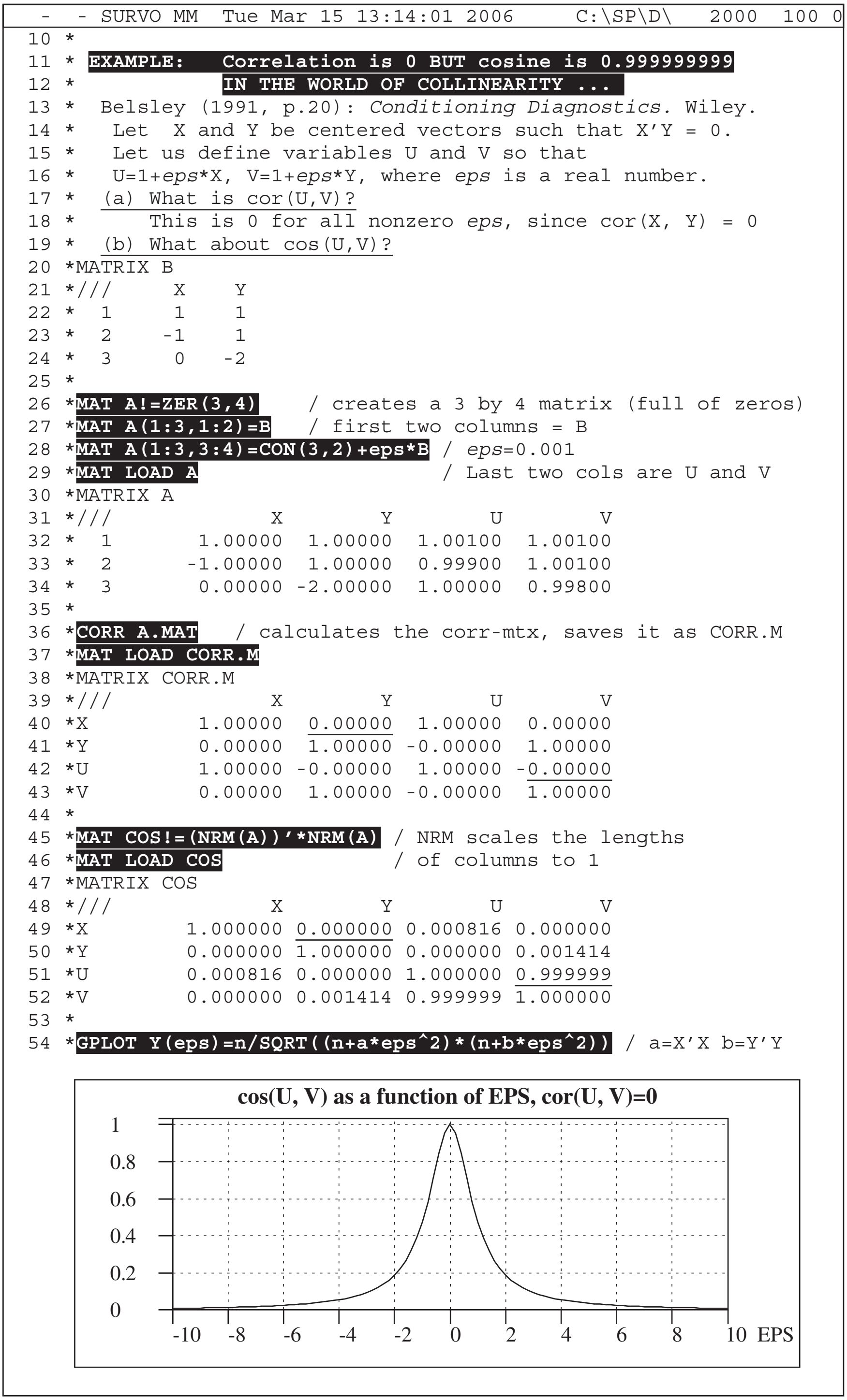
3 Introduction

The linear model that we are considering can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \cdots + \beta_k\mathbf{x}_k + \boldsymbol{\varepsilon}, \tag{3.1}$$

or in other notation, $\mathcal{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$, where $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, and $\operatorname{cov}(\mathbf{y}) = \operatorname{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. Vector $\mathbf{y}$ is an $n \times 1$ observable random vector, $\boldsymbol{\varepsilon}$ is an $n \times 1$ random error vector, $\mathbf{X}$ is a known $n \times p$ model matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and σ^2 is an unknown nonzero constant. By $\mathrm{E}(\cdot)$ and $\operatorname{cov}(\cdot)$ we denote the expectation vector and the covariance matrix, respectively.

We will use the symbols $\mathbf{A}'$, $\mathbf{A}^-$, $\mathbf{A}^+$, $\mathcal{C}(\mathbf{A})$, $\mathcal{C}(\mathbf{A})^\perp$ and $\mathrm{r}(\mathbf{A})$ to denote, respectively, the transpose, a generalized inverse, the Moore–Penrose inverse, the column space, the orthogonal complement of the column space, and the rank of the matrix $\mathbf{A}$. Furthermore we will write $\mathbf{P}_\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ to denote the orthogonal projector (with respect to the standard inner product)



```
 -  - SURVO MM  Tue Mar 15 13:14:01 2006     C:\SP\D\   2000  100 0
*
* EXAMPLE:   Correlation is 0 BUT cosine is 0.999999999
*                     IN THE WORLD OF COLLINEARITY ...
*  Belsley (1991, p.20): Conditioning Diagnostics. Wiley.
*   Let  X and Y be centered vectors such that X'Y = 0.
*   Let us define variables U and V so that
*   U=1+eps*X, V=1+eps*Y, where eps is a real number.
*  (a) What is cor(U,V)?
*      This is 0 for all nonzero eps, since cor(X, Y) = 0
*  (b) What about cos(U,V)?
*MATRIX B
*///     X    Y
*  1     1    1
*  2    -1    1
*  3     0   -2
*
*MAT A!=ZER(3,4)     / creates a 3 by 4 matrix (full of zeros)
*MAT A(1:3,1:2)=B    / first two columns = B
*MAT A(1:3,3:4)=CON(3,2)+eps*B / eps=0.001
*MAT LOAD A                    / Last two cols are U and V
*MATRIX A
*///             X        Y        U        V
*  1       1.00000  1.00000  1.00100  1.00100
*  2      -1.00000  1.00000  0.99900  1.00100
*  3       0.00000 -2.00000  1.00000  0.99800
*
*CORR A.MAT   / calculates the corr-mtx, saves it as CORR.M
*MAT LOAD CORR.M
*MATRIX CORR.M
*///             X        Y        U        V
*X         1.00000  0.00000  1.00000  0.00000
*Y         0.00000  1.00000 -0.00000  1.00000
*U         1.00000 -0.00000  1.00000 -0.00000
*V         0.00000  1.00000 -0.00000  1.00000
*
*MAT COS!=(NRM(A))'*NRM(A) / NRM scales the lengths
*MAT LOAD COS              / of columns to 1
*MATRIX COS
*///             X        Y        U        V
*X        1.000000 0.000000 0.000816 0.000000
*Y        0.000000 1.000000 0.000000 0.001414
*U        0.000816 0.000000 1.000000 0.999999
*V        0.000000 0.001414 0.999999 1.000000
*
*GPLOT Y(eps)=n/SQRT((n+a*eps^2)*(n+b*eps^2)) / a=X'X b=Y'Y
```

onto $C(\mathbf{A})$. By $(\mathbf{A} : \mathbf{B})$ we denote the partitioned matrix with $\mathbf{A}$ and $\mathbf{B}$ as submatrices.

Our model matrix above is

$$\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_0), \qquad p = k + 1, \tag{3.2}$$

where $\mathbf{X}_0$ is an $n \times k$ matrix. Denoting $\mathbf{J} = \mathbf{P_1} = \frac{1}{n}\mathbf{1}\mathbf{1}'$ and $\mathbf{C} = \mathbf{I} - \mathbf{J}$, where $\mathbf{C}$ is a centering matrix, we get

$$\mathbf{T} = (\mathbf{X}_0 : \mathbf{y})'\mathbf{C}(\mathbf{X}_0 : \mathbf{y}) = \begin{pmatrix} \mathbf{X}_0'\mathbf{C}\mathbf{X}_0 & \mathbf{X}_0'\mathbf{C}\mathbf{y} \\ \mathbf{y}'\mathbf{C}\mathbf{X}_0 & \mathbf{y}'\mathbf{C}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{T}_1 & \mathbf{t}_2 \\ \mathbf{t}_2' & t_{yy} \end{pmatrix}, \tag{3.3}$$

and hence the sample covariance matrix and sample correlation matrix of variables $x_1, x_2, \ldots, x_k, y$ are

$$\mathbf{S} = \frac{1}{n-1}\mathbf{T} = \begin{pmatrix} \mathbf{S}_1 & \mathbf{s}_2 \\ \mathbf{s}_2' & s_y^2 \end{pmatrix}, \qquad \mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{r}_2 \\ \mathbf{r}_2' & 1 \end{pmatrix}. \tag{3.4}$$

We will use the notation $\mathbf{H} = \mathbf{P_X}$ (= hat matrix), and $\mathbf{M} = \mathbf{I} - \mathbf{H}$, thereby obtaining the ordinary least squares (OLS) estimator of $\mathbf{X}\boldsymbol{\beta}$ as

$$\mathrm{OLSE}(\mathbf{X}\boldsymbol{\beta}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{P_X}\mathbf{y}, \tag{3.5}$$

where $\hat{\boldsymbol{\beta}}$ is any solution to the normal equation $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$. The corresponding vector of residuals is

$$\mathrm{res}(\mathcal{M}) = \mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \mathbf{M}\mathbf{y}. \tag{3.6}$$

One particular model deserves special attention: if the model matrix $\mathbf{X}$ has only one column and that column is $\mathbf{1}$, then we have the very simple basic model $\mathcal{M}_0 = \{\mathbf{y}, \mathbf{1}\beta_0, \sigma^2\mathbf{I}\}$. Under $\mathcal{M}_0$ we have $\mathrm{OLSE}(\mathbf{1}\beta_0 \mid \mathcal{M}_0) = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$; the residual vector is the centered $\mathbf{y}$:

$$\mathrm{res}(\mathcal{M}_0) = \mathbf{y} - \mathbf{J}\mathbf{y} = \mathbf{C}\mathbf{y} = \tilde{\mathbf{y}}. \tag{3.7}$$

The four orthogonal projectors

$$\mathbf{H} = \mathbf{P_X}, \qquad \mathbf{M} = \mathbf{P}_{\mathbf{X}^\perp} = \mathbf{I} - \mathbf{H}, \qquad \mathbf{J} = \mathbf{P_1}, \qquad \mathbf{C} = \mathbf{P}_{\mathbf{1}^\perp} = \mathbf{I} - \mathbf{J} \tag{3.8}$$

play crucial roles in many considerations related to linear regression.

As emphasized in Section 2, the sample correlation coefficient between variables x and y whose values are the elements of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is the cosine between the corresponding centered vectors:

$$\mathrm{cor_s}(x, y) = \mathrm{cor_d}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{C}\mathbf{x}, \mathbf{C}\mathbf{y}) = r_{xy} = \frac{\mathbf{x}'\mathbf{C}\mathbf{y}}{\sqrt{\mathbf{x}'\mathbf{C}\mathbf{x} \cdot \mathbf{y}'\mathbf{C}\mathbf{y}}}. \tag{3.9}$$

Note that the correlation $\mathrm{cor_s}(x, y)$ refers to a sample correlation coefficient when the arguments are the variables x and y, while in the correlation

$\mathrm{cor_d}(\mathbf{x},\mathbf{y})$ the arguments are the vectors (data) comprising the observed values of variables x and y.

Taking a look at the model matrix $\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k)$, we see that the first column there looks just as any other column, but there is one big difference: all other variables represented in the model matrix $\mathbf{X}$ have a nonzero sample variance. (Of course we can request that there be no multiples of $\mathbf{1}$ in $\mathbf{X}_0$.) Belsley (1991, p. 196) writes:

> "Much confusion surrounding centering arises because of some commonly held misconceptions about the 'constant term'. This section [6.8] aims at several of these issues with the goal of showing that, for most of the part, despite much practice to the contrary, the constant is most reasonably viewed as just another element in a regression analysis that plays no role different from any other 'variate'."

We recall that variables (columns) $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are said to be exactly collinear if one of the $\mathbf{u}_i$ is an exact linear combination of the others. This is exact collinearity, i.e., linear dependency, and by the term *collinearity* or *near dependency* we mean inexact collinear relations.

The book by Belsley (2001) offers a thorough discussion on the vector $\mathbf{1}$ and the collinearity. For example, Belsley (1991, p. 176) notes that "there is general tendency to confuse the two notions of collinearity and correlation, many practitioners thinking them to be the same." In this article, there is no space to consider such concepts, like condition number, in further detail and we refer to the interesting Chapter 6 in the book by Belsley (1991).

Inspired by Belsley's remarks, we now consider several features related to centering. Before that, we summarize (for clarity) our comments above and present three helpful lemmas that will be needed later on.

Proposition 1. *It is possible that*

(a) $\cos(\mathbf{x},\mathbf{y})$ *is high, but* $\mathrm{cor_d}(\mathbf{x},\mathbf{y}) = 0$*,*

(b) $\cos(\mathbf{x},\mathbf{y}) = 0$*, but* $\mathrm{cor_d}(\mathbf{x},\mathbf{y}) = 1$*.*

Moreover, let $\mathbf{x} \notin \mathcal{C}(\mathbf{1})$ *and* $\mathbf{y} \notin \mathcal{C}(\mathbf{1})$*. Then*

$$\mathrm{cor_d}(\mathbf{x},\mathbf{y}) = 0 \iff \mathbf{y} \in \mathcal{C}(\mathbf{Cx})^{\perp} = \mathcal{C}(\mathbf{1} : \mathbf{x})^{\perp} \oplus \mathcal{C}(\mathbf{1}). \tag{3.10}$$

Lemma 1. *The orthogonal projector* (*with respect to the standard inner product*) *onto* $\mathcal{C}(\mathbf{A} : \mathbf{B})$ *can be decomposed as* $\mathbf{P}_{(\mathbf{A}:\mathbf{B})} = \mathbf{P}_{\mathbf{A}} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A}})\mathbf{B}}$.

Lemma 2. *The rank of a partitioned matrix* $(\mathbf{A} : \mathbf{B})$ *can be expressed as*

$$\mathrm{r}(\mathbf{A} : \mathbf{B}) = \mathrm{r}(\mathbf{A}) + \mathrm{r}[(\mathbf{I} - \mathbf{P}_{\mathbf{A}})\mathbf{B}],$$

while the rank of the matrix product $\mathbf{AB}$ *is*

$$\mathrm{r}(\mathbf{AB}) = \mathrm{r}(\mathbf{A}) - \dim \mathcal{C}(\mathbf{A}') \cap \mathcal{C}(\mathbf{B}^{\perp}).$$

Lemma 3. *The following three statements are equivalent:*

(a) $\mathbf{P_A} - \mathbf{P_B}$ *is orthogonal projector,*

(b) $\mathbf{P_A P_B} = \mathbf{P_B P_A} = \mathbf{P_B}$,

(c) $\mathcal{C}(\mathbf{B}) \subset \mathcal{C}(\mathbf{A})$.

If any of the above conditions hold, then $\mathbf{P_A} - \mathbf{P_B} = \mathbf{P}_{\mathcal{C}(\mathbf{A}) \cap \mathcal{C}(\mathbf{B})^{\perp}}$.

For more about these three lemmas, see Marsaglia and Styan (1974) and Isotalo et al. (2005, Th. 1 and 4, Cor. 3.1).

4 Centered vector as a residual

Under $\mathcal{M}_0 = \{\mathbf{y}, \mathbf{1}\beta, \sigma^2\mathbf{I}\}$, the residual vector is the centered $\mathbf{y}$. What happens here is that we "eliminate" (we will return to this much-used phrase later on) the effect of the column vector of ones $\mathbf{1}$ from $\mathbf{y}$, and, what is left is just the residual which in this case is the centered $\mathbf{y}$, i.e., $\tilde{\mathbf{y}}$. Correspondingly, the centered $\mathbf{x}$ is the residual of $\mathbf{x}$ after the elimination of $\mathbf{1}$.

Consider two vectors $\mathbf{x}$ and $\mathbf{y}$ and the model $\mathcal{M}_{xy} = \{\mathbf{y}, \mathbf{x}\beta, \sigma^2\mathbf{I}\}$. A very natural measure for the "goodness" of this model would be

$$R^2_{xy} = \frac{\|\mathbf{P_x y}\|^2}{\|\mathbf{y}\|^2} = \frac{\|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2} = \frac{\mathbf{y}' \cdot \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' \cdot \mathbf{y}}{\mathbf{y}'\mathbf{y}} = \cos^2(\mathbf{x}, \mathbf{y}). \tag{4.1}$$

Note that the goodness measure above can also be expressed as

$$R^2_{xy} = \frac{\|[\mathbf{I} - (\mathbf{I} - \mathbf{P_x})]\mathbf{y}\|^2}{\|\mathbf{y}\|^2} = 1 - \frac{\|(\mathbf{I} - \mathbf{P_x})\mathbf{y}\|^2}{\|\mathbf{y}\|^2} = 1 - \frac{\mathrm{SSE}(\mathcal{M}_{xy})}{\|\mathbf{y}\|^2}, \tag{4.2}$$

where $\mathrm{SSE}(\mathcal{M}_{xy})$ refers to the sum of squares of errors under $\mathcal{M}_{xy}$.

The quantity R^2_{xy} defined here is now usually called the *coefficient of determination*, see Section 8 below. To distinguish between the situations of simple linear regression (one regressor) and multiple linear regression (more than one regressor), the terms *coefficient of simple determination* and *coefficient of multiple determination* have been used, respectively, see Puntanen and Styan (2006, page 6).

If we now eliminate the effect of $\mathbf{1}$ from $\mathbf{y}$ and $\mathbf{x}$, i.e., we center them, then we can consider the model $\mathcal{M}_{xy\cdot 1} = \{\tilde{\mathbf{y}}, \tilde{\mathbf{x}}\beta, \#\}$, where we have deliberately left the covariance matrix unnotated. Now a natural measure for the "goodness" of the model $\mathcal{M}_{xy\cdot 1}$ would be

$$R^2_{xy\cdot 1} = \frac{\|\mathbf{P}_{\tilde{\mathbf{x}}}\tilde{\mathbf{y}}\|^2}{\|\tilde{\mathbf{y}}\|^2} = \frac{\|\hat{\tilde{\mathbf{y}}}\|^2}{\|\tilde{\mathbf{y}}\|^2} = \frac{\tilde{\mathbf{y}}' \cdot \tilde{\mathbf{x}}(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1}\tilde{\mathbf{x}}' \cdot \tilde{\mathbf{y}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} = \cos^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}). \tag{4.3}$$

Obviously $R^2_{xy\cdot 1} = \mathrm{cor_d}^2(\mathbf{x}, \mathbf{y}) = r^2_{xy}$, and hence we have shown that r^2_{xy} can be interpreted as a "measure of goodness" when $\mathbf{y}$ is regressed on $\mathbf{x}$ after the elimination of $\mathbf{1}$. We later discuss a more general corresponding situation. Note further that corresponding to (4.2), we have

$$R^2_{xy\cdot 1} = 1 - \frac{\|(\mathbf{I} - \mathbf{P}_{\tilde{\mathbf{x}}})\tilde{\mathbf{y}}\|^2}{\|\tilde{\mathbf{y}}\|^2} = 1 - \frac{\mathrm{SSE}(\mathcal{M}_{xy\cdot 1})}{\mathrm{SST}(\mathcal{M}_{xy\cdot 1})}, \tag{4.4}$$

where $\mathrm{SSE}(\mathcal{M}_{xy\cdot 1})$ refers to the sum of squares of errors under $\mathcal{M}_{xy\cdot 1}$ and $\mathrm{SST}(\mathcal{M}_{xy\cdot 1}) = \mathbf{y}'\mathbf{C}\mathbf{y}$.

5 OLSE of the constant term

In this section we simply introduce (in a handy way) the formula for the OLSE of the constant term. For that purpose we partition the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0\mathbf{1} + \mathbf{X}_0\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}. \tag{5.1}$$

We recall that in the partitioned linear model $\mathcal{M}_{12}\colon \mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, we have (assuming $\mathbf{X}$ to have full column rank)

$$\hat{\boldsymbol{\beta}}_1(\mathcal{M}_{12}) = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}, \qquad \hat{\boldsymbol{\beta}}_2(\mathcal{M}_{12}) = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}, \tag{5.2}$$

where $\mathbf{M}_i = \mathbf{I} - \mathbf{P}_i$, $\mathbf{P}_i = \mathbf{P}_{\mathbf{X}_i}$. Putting $\mathbf{X}_1 = \mathbf{1}$, $\mathbf{X}_2 = \mathbf{X}_0$ yields

$$\hat{\beta}_0 = [\mathbf{1}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{1}]^{-1}\mathbf{1}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}, \tag{5.3a}$$

$$\hat{\boldsymbol{\beta}}_{(2)} = (\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{C}\mathbf{y} = (\tilde{\mathbf{X}}_0'\tilde{\mathbf{X}}_0)^{-1}\tilde{\mathbf{X}}_0'\mathbf{y} = \mathbf{T}_1^{-1}\mathbf{t}_2 = \mathbf{S}_1^{-1}\mathbf{s}_2, \tag{5.3b}$$

where $\tilde{\mathbf{X}}_0$ refers to the centered $\mathbf{X}_0$.

Usually the intercept $\hat{\beta}_0$ is not expressed as in (5.3a); the most common expression is

$$\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{(2)} = \bar{y} - (\bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_k\hat{\beta}_k), \tag{5.4}$$

where $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}_0'\mathbf{1} = (\bar{x}_1, \dots, \bar{x}_k)'$. To confirm (5.4), we use Lemma 1 which gives the decomposition

$$\mathbf{P}_{(\mathbf{A}:\mathbf{B})} = \mathbf{P}_{\mathbf{A}} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{A}})\mathbf{B}}. \tag{5.5}$$

Substituting $\mathbf{A} = \mathbf{1}$, $\mathbf{B} = \mathbf{X}_0$ into (5.5) yields

$$\begin{aligned}\mathbf{H}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{1}\hat{\beta}_0 + \mathbf{X}_0\hat{\boldsymbol{\beta}}_{(2)} \\ &= \mathbf{J}\mathbf{y} + \mathbf{C}\mathbf{X}_0(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{C}\mathbf{y} = \mathbf{J}\mathbf{y} + \mathbf{C}\mathbf{X}_0\hat{\boldsymbol{\beta}}_{(2)},\end{aligned} \tag{5.6}$$

i.e., $\mathbf{1}\hat{\beta}_0 = \mathbf{J}\mathbf{y} - \mathbf{J}\mathbf{X}_0\hat{\boldsymbol{\beta}}_{(2)}$, from which (5.4) follows.

Note that from (5.3a) we can immediately conclude the following:

$$\hat{\beta}_0 = 0 \quad\Longleftrightarrow\quad \mathcal{C}(\mathbf{X}_0 : \mathbf{y}) \subset \mathcal{C}(\mathbf{1})^{\perp} \quad\text{or}\quad \mathbf{y} \in \mathcal{C}(\mathbf{X}_0), \tag{5.7}$$

and

$$\operatorname{var}(\hat{\beta}_0) = \frac{\sigma^2}{\mathbf{y}'(\mathbf{I} - \mathbf{P}_{\mathbf{X}_0})\mathbf{y}}. \tag{5.8}$$

The result (5.8) implies, as shown (by other means) by Seber and Lee (2003, pp. 251–252, and Ex. 9d, No. 1), that the variance of $\hat{\beta}_0$ does not depend on the scale used to measure the variables in $\mathbf{X}_0$; changing the scale means the multiplication $\mathbf{X}_0\mathbf{D}$, where $\mathbf{D}$ is a positive definite diagonal matrix.

6 Does centering 'get rid' of the constant term?

Belsley (1991, p. 199) writes:

> "It is generally thought that centering 'gets rid' of the constant term. But this is not the case; centering merely redistributes the constant among all the variates so that it continues to be present, but not explicitly."

To see what is behind this remark, we will consider the following models:

$$\mathcal{M}_{12} = \{\mathbf{y},\ (\mathbf{1} : \mathbf{X}_0)\boldsymbol{\beta},\ \sigma^2\mathbf{I}\}, \qquad \mathcal{M}_{12\cdot 1} = \{\mathbf{C}\mathbf{y},\ \mathbf{C}\mathbf{X}_0\boldsymbol{\beta}_{(2)},\ \sigma^2\mathbf{C}\}, \tag{6.1a}$$

$$\mathcal{M}_c = \{\mathbf{y},\ (\mathbf{1} : \mathbf{C}\mathbf{X}_0)\boldsymbol{\beta},\ \sigma^2\mathbf{I}\}, \qquad \mathcal{M}_r = \{\mathbf{y},\ \mathbf{C}\mathbf{X}_0\boldsymbol{\beta}_{(2)},\ \sigma^2\mathbf{I}\}. \tag{6.1b}$$

The model $\mathcal{M}_{12}$ is the full model and all other models are various versions of it. In all these versions we have done something related to the constant term: we have centered something. The above models frequently appear in practice (and in teaching regression in statistics courses). We now review some properties of these models.

We first note that the above models are special versions of the following more general models:

$$\mathscr{M}_{12} = \{\mathbf{y},\ \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\ \mathbf{V}\}, \qquad \mathscr{M}_{12\cdot 1} = \{\mathbf{M}_1\mathbf{y},\ \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\ \mathbf{M}_1\mathbf{V}\mathbf{M}_1\}, \tag{6.2a}$$

$$\mathscr{M}_c = \{\mathbf{y},\ (\mathbf{X}_1 : \mathbf{M}_1\mathbf{X}_2)\boldsymbol{\beta},\ \mathbf{V}\}, \qquad \mathscr{M}_r = \{\mathbf{y},\ \mathbf{M}_1\mathbf{X}_2\boldsymbol{\beta}_2,\ \mathbf{V}\}. \tag{6.2b}$$

These models have recently been studied, for example, by Bhimasankaram et al. (1998), Groß and Puntanen (2000), and Chu et al. (2004, 2005).

We call $\mathscr{M}_{12\cdot 1}$ a reduced model. It is obtained by premultiplying the full model equation

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \tag{6.3}$$

by the orthogonal projector $\mathbf{M}_1$. In the reduced model, the response variable is the residual when $\mathbf{y}$ is explained by the variables represented by $\mathbf{X}_1$, and the explanatory variables are the residuals of the $\mathbf{X}_2$ "after elimination of $\mathbf{X}_1$". Therefore, in the case when $\mathbf{X}_2 = \mathbf{x}_k$, the (squared) multiple correlation coefficient in the corresponding reduced model is the (squared) partial correlation between $\mathbf{y}$ and $\mathbf{x}_k$ after the elimination of all other $\mathbf{x}$-variables. The plots of residuals $\mathbf{M}_1\mathbf{y}$ and $\mathbf{M}_1\mathbf{x}_k$ are called added variable plots.

Now, let us return to $\mathcal{M}$-models. We assume that $\mathbf{X}$ has full column rank and thereby that $\hat{\boldsymbol{\beta}}(\mathcal{M}_{12})$ is unique. In view of Lemma 1, we have

$$p = \mathrm{r}(\mathbf{X}) = \mathrm{r}(\mathbf{1} : \mathbf{X}_0) = \mathrm{r}(\mathbf{1} : \mathbf{C}\mathbf{X}_0) = 1 + \mathrm{r}(\mathbf{C}\mathbf{X}_0), \tag{6.4}$$

and hence $\mathrm{r}(\mathbf{X}) = p \implies \mathrm{r}(\mathbf{C}\mathbf{X}_0) = k = p - 1$, and so $\mathbf{C}\mathbf{X}_0$ has also full column rank and hence $\beta_{(2)}$ is estimable under $\mathcal{M}_c$, etc.

Taking a look at the four models in (6.1), we immediately observe an interesting feature which we may state as a Proposition:

Proposition 2. *Consider the models defined in* (6.1)*, and let* $\mathbf{X}$ *have full column rank. Then* $\hat{\boldsymbol{\beta}}_{(2)}$ *is the same in each model, i.e.,*

$$\hat{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_{12}) = \hat{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_{12\cdot 1}) = \hat{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_r) = \hat{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_c). \tag{6.5}$$

Moreover, the residuals under the models $\mathcal{M}_{12}$*,* $\mathcal{M}_{12\cdot 1}$*, and* $\mathcal{M}_c$ *are identical.*

Proof. The first two equalities in (6.5) are obvious. Model $\mathcal{M}_c$ is a reparameterization of $\mathcal{M}_{12}$. This is seen from $(\mathbf{1} : \mathbf{C}\mathbf{X}_0) = (\mathbf{1} : \mathbf{X}_0)\mathbf{A}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & -\mathbf{1}^{+}\mathbf{X}_0 \\ \mathbf{0} & \mathbf{I}_{k-1} \end{pmatrix} = \begin{pmatrix} 1 & -(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{X}_0 \\ \mathbf{0} & \mathbf{I}_{k-1} \end{pmatrix} = \begin{pmatrix} 1 & -\bar{\mathbf{x}}' \\ \mathbf{0} & \mathbf{I}_{k-1} \end{pmatrix}. \tag{6.6}$$

It is easy to confirm that $\hat{\beta}_0(\mathcal{M}_c) = \bar{y}$ and $\hat{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_c) = \hat{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_{12})$. Since $\mathcal{C}(\mathbf{1} : \mathbf{C}\mathbf{X}_0) = \mathcal{C}(\mathbf{1} : \mathbf{X}_0)$, the residual vectors under $\mathcal{M}_c$ and $\mathcal{M}_{12}$ are identical:

$$\mathrm{res}(\mathcal{M}_c) = \mathrm{res}(\mathcal{M}_{12}) = \mathbf{y} - \mathbf{H}\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{e}. \tag{6.7}$$

The residual vector under $\mathcal{M}_{12\cdot 1}$ becomes (in view of Lemma 1)

$$\mathrm{res}(\mathcal{M}_{12\cdot 1}) = \mathbf{C}\mathbf{y} - \mathbf{P}_{\mathbf{C}\mathbf{X}_0}\mathbf{C}\mathbf{y} = \mathbf{y} - (\mathbf{P}_{\mathbf{1}}\mathbf{y} + \mathbf{P}_{(\mathbf{I}-\mathbf{P}_{\mathbf{1}})\mathbf{X}_0}\mathbf{y}) = \mathbf{M}\mathbf{y}. \tag{6.8}$$

□

The first equality in (6.5) is known as (a special case of) the Frisch–Waugh–Lovell Theorem, see, e.g., Frisch and Waugh (1933) and Lovell (1963), and further extensions by Groß and Puntanen (2000, 2005).

Premultiplying the equation

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}_0\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon} \tag{6.9}$$

with $\mathbf{J} = \mathbf{P}_{\mathbf{1}}$ shows that

$$\beta_0\mathbf{1} = \bar{y}\mathbf{1} - \mathbf{P}_{\mathbf{1}}\mathbf{X}_0\boldsymbol{\beta}_{(2)} - \bar{\varepsilon}\mathbf{1}. \tag{6.10}$$

If (6.9) is premultiplied with $\mathbf{C}$ we obtain the reduced model $\mathcal{M}_{12\cdot 1}$ which can be written as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_0\boldsymbol{\beta}_{(2)} + \tilde{\boldsymbol{\varepsilon}}, \tag{6.11}$$

$$\mathbf{y} - \bar{y}\mathbf{1} = (\mathbf{X}_0 - \mathbf{P}_{\mathbf{1}}\mathbf{X}_0)\boldsymbol{\beta}_{(2)} + (\boldsymbol{\varepsilon} - \bar{\varepsilon}\mathbf{1}). \tag{6.12}$$

Belsley (1991, p. 199), referring to the above equations, observes that

> "In the form (6.11), there appears to be no constant term, but in the more transparent form (6.12) we see through to the fact that the information in (6.10) is indeed present; it is just incorporated into the other variates – the constant term is still there."

Belsley (1991, p. 199) also remarks (and proves, somewhat tediously) that "**1** does not even play a unique role in such 'reductive' transformations of the regression, for we may similarly transform (6.9) with any of the explanatory variates $\mathbf{x}_i$". Indeed this is so, and the explanation is simply the Frisch–Waugh–Lovell Theorem, which is not referred to by Belsley.

It is worth noting that in the reduced model $\mathcal{M}_{12\cdot 1}$, the covariance matrix $\sigma^2\mathbf{C}$ is singular. It is well known that a linear model with a singular covariance matrix requires specific attention.

We recall that a linear unbiased estimator $\mathbf{Gy}$ is the BLUE of an estimable parametric function $\mathbf{K}'\boldsymbol{\beta}$ if it has the smallest covariance matrix (in the Löwner sense) among all unbiased linear estimators of $\mathbf{K}'\boldsymbol{\beta}$.

It is well known that $\hat{\boldsymbol{\beta}}_{(2)}$ is the BLUE of $\boldsymbol{\beta}_{(2)}$ under $\mathcal{M}_{12}$. But we may wonder what is the BLUE of $\boldsymbol{\beta}_{(2)}$ under $\mathcal{M}_{12\cdot 1}$, denoted as $\tilde{\boldsymbol{\beta}}_{(2)}(\mathcal{M}_{12\cdot 1})$. "Luckily" it appears to be equal to the corresponding OLSE. To prove this, we can use, for example, the following general result: Under the general linear model $\{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \mathbf{V}\}$, OLSE($\mathbf{X}\boldsymbol{\beta}$) = BLUE($\mathbf{X}\boldsymbol{\beta}$) if and only if

$$\mathcal{C}(\mathbf{VX}) \subset \mathcal{C}(\mathbf{X}), \tag{6.13}$$

see, e.g., Rao (1967), Zyskind (1967), and Puntanen and Styan (1989). Substituting (6.13) into the model $\mathcal{M}_{12\cdot 1}$, we obtain the column space inclusion $\mathcal{C}(\mathbf{C}\cdot\mathbf{CX}_0) \subset \mathcal{C}(\mathbf{CX}_0)$, which clearly holds. Hence we may state the following proposition.

Proposition 3. *Let* $\mathbf{X}$ *have full column rank. Then*

$$\mathrm{OLSE}(\boldsymbol{\beta}_{(2)} \mid \mathcal{M}_{12\cdot 1}) = \mathrm{BLUE}(\boldsymbol{\beta}_{(2)} \mid \mathcal{M}_{12\cdot 1}). \tag{6.14}$$

The model $\mathcal{M}_r$ may seem to be obscure, but we note that Groß and Puntanen (2000, p. 133) "...rather like to think of model $\mathcal{M}_r$ as a source of estimators whose properties under the model $\mathcal{M}_{12}$ are investigated..." .

7 Rank of the sample correlation matrix

Let the $n \times p$ model matrix $\mathbf{X}$ be partitioned as $\mathbf{X} = (\mathbf{1} : \mathbf{x}_1 : \ldots : \mathbf{x}_k) = (\mathbf{1} : \mathbf{X}_0)$, where $p = k + 1$. The sample covariance and correlation matrices of x-variables are

$$\mathbf{S}_1 = \tfrac{1}{n-1}\mathbf{X}_0'\mathbf{C}\mathbf{X}_0 = \tfrac{1}{n-1}\mathbf{T}_1, \qquad \mathbf{R}_1 = [\mathrm{diag}(\mathbf{T}_1)]^{-\frac{1}{2}}\mathbf{T}_1[\mathrm{diag}(\mathbf{T}_1)]^{-\frac{1}{2}}. \tag{7.1}$$

We assume that all x-variables have nonzero variances, that is, $\mathbf{x}_i \notin \mathcal{C}(\mathbf{1})$, $i = 1, \ldots, k$. Then Lemma 1 now immediately gives the following result where the vector $\mathbf{1}$ has a crucial role:

Proposition 4. *Assume that all x-variables have nonzero variances. Then the rank of the model matrix* $\mathbf{X} = (\mathbf{1} : \mathbf{X}_0)$ *can be expressed as*

$$\begin{aligned} \mathrm{r}(\mathbf{X}) &= 1 + \mathrm{r}(\mathbf{X}_0) - \dim \mathcal{C}(\mathbf{1}) \cap \mathcal{C}(\mathbf{X}_0) = 1 + \mathrm{r}(\mathbf{C}\mathbf{X}_0) \\ &= 1 + \mathrm{r}(\mathbf{T}_1) = 1 + \mathrm{r}(\mathbf{S}_1) = 1 + \mathrm{r}(\mathbf{R}_1), \end{aligned} \tag{7.2}$$

and moreover,

$$\det(\mathbf{R}_1) \neq 0 \iff \mathrm{r}(\mathbf{X}) = k + 1 \iff \mathrm{r}(\mathbf{X}_0) = k \quad \text{and} \quad \mathbf{1} \notin \mathcal{C}(\mathbf{X}_0). \tag{7.3}$$

It is noteworthy that $r_{ij} = 1$ for some $i \neq j \implies \det(\mathbf{R}_1) = 0$ (but not vice versa). It is also easy to conclude that the correlation matrix $\mathbf{R}_1$ is singular if and only if (at least) one column, $\mathbf{x}_k$ say (for notational simplicity), of $\mathbf{X}_0$, is a linear combination of vectors $\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_{k-1}$.

8 Coefficient of determination

A place where the constant vector plays a fundamental role is in the concept of the coefficient of determination (squared sample multiple correlation coefficient), denoted as $R^2 = R^2_{y\cdot\mathbf{x}}$, see also §4 above. Let us consider the model

$$\mathcal{M}_{12} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\} = \{\mathbf{y}, \mathbf{1}\beta_0 + \mathbf{X}_0\boldsymbol{\beta}_{(2)}, \sigma^2\mathbf{I}\}, \tag{8.1}$$

and denote

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \quad \mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{8.2}$$

Then, according to many textbooks on regression, e.g., Seber and Lee (2003, § 4.4),

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}, \tag{8.3}$$

and the coefficient of determination can be defined as

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}. \tag{8.4}$$

Some statisticians prefer to adjust R^2 by dividing numerator and denominator in $1 - R^2$ by the corresponding degrees of freedom:

$$R^2_{\mathrm{adj}} = 1 - \frac{\mathrm{SSE}\,/\,(n-p)}{\mathrm{SST}\,/\,(n-1)}. \tag{8.5}$$

This adjusted R^2_{adj}, which is also known as Fisher's A statistic, is used in subset selection, see, e.g., Miller (2002, page 161), Puntanen and Styan (2006, page 9).

Why is R^2 so popular? For an interesting discussion about the "hot air" in R^2, see, McGuirk and Driscoll (1995, 1996) and Lavergne (1996).

We take a quick look at the equality (8.3). In matrix terms, we can write

$$\mathrm{SST} = \|(\mathbf{I}-\mathbf{J})\mathbf{y}\|^2, \qquad \mathrm{SSR} = \|(\mathbf{H}-\mathbf{J})\mathbf{y}\|^2, \qquad \mathrm{SSE} = \|(\mathbf{I}-\mathbf{H})\mathbf{y}\|^2, \tag{8.6}$$

and we trivially have $(\mathbf{I}-\mathbf{J})\mathbf{y} = (\mathbf{H}-\mathbf{J})\mathbf{y} + (\mathbf{I}-\mathbf{H})\mathbf{y}$. However, the equation

$$\|(\mathbf{I}-\mathbf{J})\mathbf{y}\|^2 = \|(\mathbf{H}-\mathbf{J})\mathbf{y}\|^2 + \|(\mathbf{I}-\mathbf{H})\mathbf{y}\|^2 \tag{8.7}$$

holds if and only if $(\mathbf{H}-\mathbf{J})\mathbf{y}$ and $(\mathbf{I}-\mathbf{H})\mathbf{y}$ are orthogonal, i.e.,

$$\mathbf{y}'(\mathbf{H}-\mathbf{J})(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{JH}-\mathbf{J})\mathbf{y} = 0. \tag{8.8}$$

Now (8.8) holds (for all $\mathbf{y}$) if and only if $\mathbf{J} = \mathbf{JH}$, and thereby $\mathbf{J} = \mathbf{JH} = \mathbf{HJ}$. Lemma 3 immediately gives the following Proposition:

Proposition 5. *Consider the model* $\mathcal{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$. *Then the decomposition* $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$ *holds* (*for all* $\mathbf{y}$) *if and only if* $\mathbf{1} \in C(\mathbf{X})$.

Note that for $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$ to hold, it is not necessary that the vector $\mathbf{1}$ be explicitly a column of $\mathbf{X}$; it is enough if $\mathbf{1} \in C(\mathbf{X})$. In this situation, in view of Lemma 3, $\mathbf{H}-\mathbf{J}$ is an orthogonal projector:

$$\mathbf{H}-\mathbf{J} = \mathbf{P}_{C(\mathbf{X})\cap C(\mathbf{1})^\perp} = \mathbf{P}_{\mathbf{CX}_0}. \tag{8.9}$$

When $\mathbf{1} \notin C(\mathbf{X})$, it makes no sense to use (8.4) as a statistic to describe how well OLS fits the data. When $\mathbf{1} \notin C(\mathbf{X})$, R^2 as defined in (8.4), can be even negative. In the no-intercept model, it is natural [see, e.g., Searle (1982, p. 379)] to consider the decomposition

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}, \qquad \mathrm{SST}_* = \mathrm{SSR}_* + \mathrm{SSE}. \tag{8.10}$$

Then the coefficient of determination can be defined as

$$R^2_* = \frac{\mathrm{SSR}_*}{\mathrm{SST}_*} = \frac{\mathbf{y}'\mathbf{H}\mathbf{y}}{\mathbf{y}'\mathbf{y}} = 1 - \frac{\mathrm{SSE}}{\mathbf{y}'\mathbf{y}} = \cos^2(\mathbf{y}, \mathbf{H}\mathbf{y}). \tag{8.11}$$

However, there are some disadvantages in (8.11) as pointed out by Weisberg (2005, p. 84): "The quantity in (8.11) is not invariant under location change, so, for example, if units are changed from Fahrenheit to Celsius, you will get different value for (8.11)." See also Puntanen and Styan (2006, page 15).

We may end up with R^2 via different routes. One very natural approach in which the vector $\mathbf{1}$ has an important role is to consider the simple basic model where the only explanatory variable is a constant: $\mathcal{M}_0 = \{\mathbf{y}, \mathbf{1}\beta_0, \sigma^2\mathbf{I}\}$. Under $\mathcal{M}_0$ we have $\mathrm{OLSE}(\mathbf{1}\beta_0) = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1}$, while the residual vector is the centered $\mathbf{y}$, that is, $\mathbf{C}\mathbf{y}$, and hence the residual sum of squares under $\mathcal{M}_0$ is

$$\mathrm{SSE}_0 = \mathbf{y}'(\mathbf{I}-\mathbf{J})\mathbf{y} = \mathbf{y}'\mathbf{C}\mathbf{y} = t_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathrm{SST}. \tag{8.12}$$

We may want to compare the full model $\mathcal{M}_{12}$ and the simple basic model $\mathcal{M}_0$ by means of residual sum of squares: how much benefit is gained in the residual sum of squares when also using x-variables as explanatory variables. The change in SSE when moving from $\mathcal{M}_0$ to $\mathcal{M}_{12}$ is

$$\mathrm{SSE}_0 - \mathrm{SSE} = \mathbf{y}'(\mathbf{I}-\mathbf{J})\mathbf{y} - \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{H}-\mathbf{J})\mathbf{y} = \mathrm{SSR}, \tag{8.13}$$

which is called "sum of squares due to regression". The value of SSR tells the reduction in SSE when using $\mathcal{M}_{12}$ instead $\mathcal{M}_0$, but it is definitely more informative to study the *relative* reduction in SSE, that is, we have reasons to calculate the ratio

$$\frac{\mathrm{SSE}_0 - \mathrm{SSE}}{\mathrm{SSE}_0} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = R^2. \tag{8.14}$$

A fundamental property of R^2 defined above is that it equals the square of the multiple correlation coefficient between the $\mathbf{y}$ and $\mathbf{Hy}$. This is a well-known result but we state it here since it nicely illustrates the important role of $\mathbf{1}$.

Proposition 6. *Consider the model* $\mathcal{M} = \{\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}\}$*, where* $\mathbf{1} \in C(\mathbf{X})$ *and let* R^2 *be defined as in* (8.14)*. Then*

$$R = \mathrm{cor_d}(\mathbf{y}, \mathbf{Hy}) = \cos[(\mathbf{I}-\mathbf{J})\mathbf{y}, (\mathbf{I}-\mathbf{J})\mathbf{Hy}] = \cos(\mathbf{Cy}, \mathbf{CHy}). \tag{8.15}$$

Proof. In view of $\mathbf{1} \in C(\mathbf{X}) \iff \mathbf{H1} = \mathbf{1} \iff \mathbf{J} = \mathbf{HJ} = \mathbf{JH}$, we observe that $(\mathbf{I}-\mathbf{J})\mathbf{H} = \mathbf{H}-\mathbf{J}$. Hence our claim is

$$R = \mathrm{cor_d}(\mathbf{y}, \mathbf{Hy}) = \cos[\mathbf{Cy}, (\mathbf{H}-\mathbf{J})\mathbf{y}] := \frac{a}{\sqrt{b \cdot c}}, \tag{8.16}$$

which indeed is true since $a = \mathbf{y}'(\mathbf{I}-\mathbf{J})\mathbf{Hy} = \mathbf{y}'(\mathbf{H}-\mathbf{J})\mathbf{y}$, $b = \mathbf{y}'\mathbf{Cy}$, and $c = \mathbf{y}'(\mathbf{H}-\mathbf{J})\mathbf{y}$. □

The geometry behind decomposition SST = SSR + SSE is illustrated in Figure 1.

Note that if *all* variables $x_1, \dots, x_k$, y are random variables with the covariance matrix

$$\mathrm{cov}\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\sigma}_2\\ \boldsymbol{\sigma}_2' & \sigma_y^2\end{pmatrix}, \tag{8.17}$$

then the population multiple correlation (squared) is defined as

$$\mathcal{R}^2 = \max_{\mathbf{a}} \mathrm{cor}^2(y, \mathbf{a}'\mathbf{x}) = \frac{\boldsymbol{\sigma}_2'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_2}{\sigma_y^2} = 1 - \frac{\sigma_y^2 - \boldsymbol{\sigma}_2'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\sigma}_2}{\sigma_y^2}. \tag{8.18}$$

It is worth emphasizing (see, e.g., Weisberg 2005, § 4.4), that when the x-variables are controllable, fixed, then R^2 is not to be considered as an estimate of something like $\mathcal{R}^2$: for a linear model where $\mathbf{X}$ is fixed, there exists no

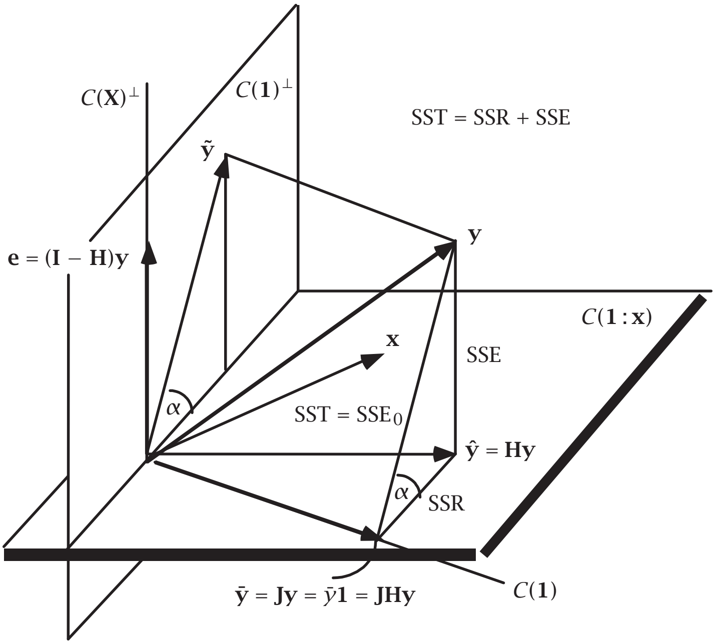

$C(\mathbf{X})^{\perp}$
$C(\mathbf{1})^{\perp}$
SST = SSR + SSE
$\tilde{\mathbf{y}}$
$\mathbf{y}$
$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$
$C(\mathbf{1} : \mathbf{x})$
$\mathbf{x}$
SSE
α
$\text{SST} = \text{SSE}_0$
$\hat{\mathbf{y}} = \mathbf{Hy}$
α
SSR
$\bar{\mathbf{y}} = \mathbf{Jy} = \bar{y}\mathbf{1} = \mathbf{JHy}$
$C(\mathbf{1})$

Figure 1. Illustration of SST = SSR + SSE.

such a parameter as $\mathcal{R}$. The sample value R^2 is merely a descriptive measure how well the OLS fits the data.

We complete this section by noting that in view of (3.3) and (8.9), we can express R^2, corresponding to (8.18), as

$$R^2 = \frac{\mathbf{y}'(\mathbf{H} - \mathbf{J})\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{J})\mathbf{y}} = \frac{\mathbf{t}_2'\mathbf{T}_1^{-1}\mathbf{t}_2}{t_{yy}}. \tag{8.19}$$

Acknowledgements

We wish to thank Professor Belsley for allowing us to publish the e-mail correspondence in the Section 1. The research of the third author was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Noted added in proof

As we were reading the final page proofs of this paper we discovered the recent article by Friedman and Wall (2005), and the related Letter to the Editor by Christensen (2006), and reply by Friedman (2006). We plan to comment on this work in a further paper.

References

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: Wiley.

Bhimasankaram, P., Shah, K. R. and Saha Ray, R. (1998). On a singular partitioned linear model and some associated reduced models. *Journal of Combinatorics, Information & System Sciences*, 23, 415–421.

Christensen, R. (2006). Comment on Friedman and Wall (2005). Letter to the Editor. *The American Statistician*, 60, 101–102.

Chu, K. L., Isotalo, J., Puntanen, S. and Styan, G. P. H. (2004). On decomposing the Watson efficiency of ordinary least squares in a partitioned weakly singular linear model. *Sankhyā*, 66, 634–651.

Chu, K. L., Isotalo, J., Puntanen, S. and Styan, G. P. H. (2005). Some further results concerning the decomposition of the Watson efficiency in partitioned linear models. *Sankhyā*, 67, 74–89.

Friedman, L. (2006). Reply to Christensen (2006). Letter to the Editor. *The American Statistician*, 60, 102–103.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59, 127–136.

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1, 387–401.

Groß, J. and Puntanen, S. (2000). Estimation under a general partitioned linear model. *Linear Algebra and Its Applications*, 321, 131–144.

Groß, J. and Puntanen, S. (2005). Extensions of the Frisch–Waugh–Lovell Theorem. *Discussiones Mathematicae – Probability and Statistics*, 25, 39–49.

Isotalo, J., Puntanen, S. and Styan, G. P. H. (2005). Matrix tricks for linear statistical models: our personal Top Sixteen. Report A 363, Dept. of Mathematics, Statistics & Philosophy, University of Tampere.

Lavergne, P. (1996). The hot air in R^2: comment [on McQuirk and Driscoll (1995)]. *American Journal of Agricultural Economics*, 78, 712–714.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58, 993–1010.

Marsaglia, G. and Styan, G. P. H. (1974). Equalities and inequalities for ranks of matrices. *Linear and Multilinear Algebra*, 2, 269–292.

McQuirk, A. and Driscoll, P. (1995). The hot air in R^2 and consistent measures of explained variation. *American Journal of Agricultural Economics*, 77, 319–328. [See also comment by Lavergne (1996) and Reply by McQuirk and Driscoll (1996).]

McQuirk, A. and Driscoll, P. (1996). The hot air in R^2: reply [to Lavergne (1996)]. *American Journal of Agricultural Economics*, 78, 715–717.

Mustonen, Seppo (2001). *SURVO MM: Computing Environment for Creative Processing of Text and Numerical Data.* http://www.survo.fi/english/index.html

Puntanen, S. and Styan, G. P. H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43, 153–164.

Puntanen, S. and Styan, G. P. H. (2006). Some easy matrix tricks useful in teaching linear statistical models, with some comments on subset selection criteria in multiple linear regression. Report 2006-04, Dept. of Mathematics and Statistics, McGill University.

Rao, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability: Berkeley, California, 1965/1966*, vol. 1. Eds. L. M. Le Cam and J. Neyman. Berkeley: Univ. of California Press, 355–372.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* New York: Wiley.

Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis.* Second Edition. New York: Wiley.

Weisberg, S. (2005). *Applied Linear Regression.* Third Edition. New York: Wiley.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Annals of Mathematical Statistics*, 38, 1092–1109.

Jarkko Isotalo
Department of Mathematics, Statistics and Philosophy
FI-33014 University of Tampere, Finland
jarkko.isotalo@uta.fi

Simo Puntanen
Department of Mathematics, Statistics and Philosophy
FI-33014 University of Tampere, Finland
sjp@uta.fi
http://www.uta.fi/~sjp/

George P. H. Styan
Department of Mathematics and Statistics
McGill University, Burnside Hall Room 1005
805 rue Sherbrooke Street West
Montréal (Québec), Canada H3A 2K6
styan@math.mcgill.ca
http://www.math.mcgill.ca/styan/