GEOMETRY OF BASIC PROPERTIES ON LINEAR REGRESSION AND OF THE MALLOWS'S C_P STATISTICS

Leandro da Silva PEREIRA[1]
Lucas Monteiro CHAVES[1]
Devanil Jaques de SOUZA[1]
Carlos José dos REIS[1]

■ ABSTRACT: A fully geometric approach, employing only vectorial subspaces and orthogonal projections, is applied to the theory of linear models. Basic results, usually proved in textbooks using non trivial matricial algebra, are demonstrated using only geometry. As a rather unusual application, the method is applied to the construction of Mallows's C_p statistic.

■ KEYWORDS: Geometric Methods; Mallows's C_p; linear models; orthogonal projections.

1 Introduction

The multiple regression model can be expressed in matricial form by $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1}+\boldsymbol{\varepsilon}_{n\times 1}$, with $\mathbf{Y}$ the data vector, $\mathbf{X}$ the design matrix, $\boldsymbol{\beta}$ the vector of parameters and $\boldsymbol{\varepsilon}$ the vector of errors. This formulation admits an interpretation for a geometric approach, if the matrix $\mathbf{X}$ is considered as a linear transformation from the parameter space to the data space. In this case, the subspace defined by vectors which are image by the transformation $\mathbf{X}$ of elements in the parameter space will be called the image of $\mathbf{X}$, denoted by $\mathrm{Im}(\mathbf{X}) = \{\mathbf{Xv}, \mathbf{v} \in \mathbb{R}^p\}$. Note that if the columns of $\mathbf{X}$ are represented by $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_n)$, then $\mathrm{Im}(\mathbf{X})$ is spanned by the vectors $\mathbf{X}_i \in \mathbb{R}^n$.

The subspace $\mathrm{Im}(\mathbf{X})$ is where belong all vectors that can represent the expected value of the random vector $\mathbf{Y}$. Geometrically, the linear model is describle in Figure 1 and the estimating process can be described simply as a rule that assigns

[1]Universidade Federal de Lavras – UFLA, Departamento de Ciências Exatas – DEX, CEP: 37200-000, Lavras, Minas Gerais, Brasil. E-mail: *leandrospereira@dex.ufla.br*; *lucas@dex.ufla.br*; *devanil@dex.ufla.br*; *carlosjreis17@yahoo.com.br*

to each observed vector $\mathbf{y}$ another vector in the space $\mathrm{Im}(\mathbf{X})$. If the estimator is linear, this rule is defined by a linear transformation. Since it is reasonable that, if the observed vector belongs to $\mathrm{Im}(\mathbf{X})$ then itself should be the estimative of $\mathbf{Y}$ expectation, then the linear estimation rules are defined by linear projections.

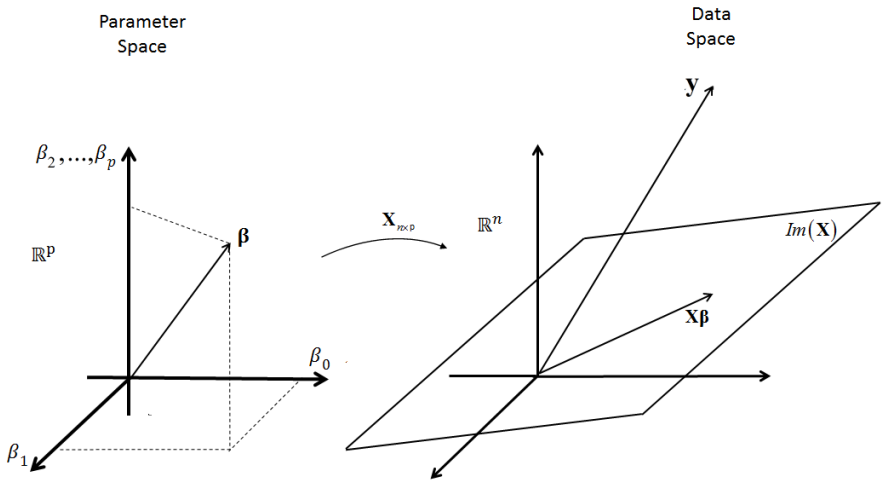

Figure 1 - Geometric model for multiple regression.

Definition 1.1. A linear transformation $\mathbf{A}$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ is a linear projector if $\mathbf{A}^2 = \mathbf{A}$. The linear projector $\mathbf{A}$ is said to be a orthogonal projector if, for all vector $\mathbf{w}$, $\mathbf{w} - \mathbf{A}\mathbf{w}$ is perpendicular to the subspace $\mathrm{Im}(\mathbf{X})$.

A useful concept, when dealing with projections, is the direct sum. A vector space V is orthogonal direct sum of V_1 and V_2, denoted by $\mathrm{V}_1 \oplus \mathrm{V}_2$, if all vector $\mathbf{v} \in \mathrm{V}$ is uniquely written as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, with $\mathbf{v}_1 \in \mathrm{V}_1$, $\mathbf{v}_2 \in \mathrm{V}_2$ and inner product is zero, $\mathbf{v}'\mathbf{w} = 0$. Note that if $\mathrm{P}_{\mathrm{V}_1}$ is the orthogonal projection onto the subspace V_1, then $\mathrm{P}_{\mathrm{V}_1}(\mathbf{v}) = \mathrm{P}_{\mathrm{V}_1}(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{v}_1$.

The orthogonal complement of a subspace $\mathrm{W} \subset \mathrm{V}$, denoted by $\mathrm{W}^{\perp}$, is defined as $\mathrm{W}^{\perp} = \{\mathbf{v} \in \mathrm{V}, \mathbf{v}'\mathbf{w} = 0 \,\forall \mathbf{w} \in \mathbf{W}\}$. Therefore we have a decomposition in terms of direct orthogonal sum, $\mathrm{W} \oplus \mathrm{W}^{\perp} = \mathrm{V}$. It is well known that a linear projector is orthogonal, if and only if, it is a symmetric linear transformation.

If $\mathrm{P}_{Im(\mathbf{X})}$ is the orthogonal projection onto $Im(\mathbf{X})$, then in terms of the $\mathbf{X}$ transformation $\mathrm{P}_{Im(\mathbf{X})} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$, where $\mathbf{X}'$ is the transpose of $\mathbf{X}$ defined by $(\mathbf{X}\mathbf{v})'\mathbf{w} = \mathbf{v}'(\mathbf{X}'\mathbf{w})$, $\forall \mathbf{v}, \mathbf{w}$.

The matrix $\frac{1}{\mathrm{n}}\mathbf{J}$, orthogonal projection onto the subspace spanned by the unit

vector $\mathbf{1} = (1, ..., 1)'$, is given by

$$\frac{1}{\mathrm{n}}\mathbf{J} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{bmatrix}_{n \times n}.$$

Note that if $\mathbf{y} = (y_1, \ldots, y_{\mathrm{n}})$, $\frac{1}{\mathrm{n}}\mathbf{J}\mathbf{y} = \bar{\mathbf{y}} = (\bar{y}, \ldots, \bar{y})$, wherein $\bar{y} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} y_i$.

The least squares estimator of the regression model can be geometrically described in a simple way. Once observed the data vector $\mathbf{y}$, it is one orthogonally projected onto $\mathrm{Im}(\mathbf{X})$ obtaining the vector $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. $\hat{\mathbf{y}}$ is denominated adjusted data vector. Since $\mathbf{X}$ is an injective linear transformation, there is only one vector $\hat{\boldsymbol{\beta}}$ belonging to the $\mathbb{R}^{\mathrm{p}}$ such that $\mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Again, by $\mathbf{X}$ injectivity, one achieves the expression of the least squares estimator (Gauss-Markov estimator) $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The Gauss-Markov estimator geometry can be described in terms of a fundamental triangle. Except for the C_p Mallow statistic, the subject and notation of this work are described as treated in Rencher and Schaalje (2008).

There are no new results in the present article, except a complete geometrical description of C_p statistic. However, the novelty relies in the geometric approach, which employees extensively the concepts of vectorial subspace, orthogonal projections and linear transformations. The use of geometric concepts in statistic dates back a long time, going back to Fisher on 1915 and involving great statisticians over the years (HERR, 1980). The most recent works, such as Margolis (1979), Bryant (1984), Saville (1986), Saville (1991), Bring (1996), even in applied statistics, as Bailey (2008), that uses geometric modeling concepts as factors, blocks and treatments, rely almost entirely on the concept of vectors. The same can be said about standart linear models books as Rao (1973), Draper and Smith (1998), Seber and Lee (2003) and Searle (1971). This type of approach, although geometric, can not be done without matrix algebra concepts. The big difference in relation to this work is that matrices are addressed as linear transformations, which effectively reduces to a minimum the use of algebraic properties of matrices. For the theory of matrix as linear transformation see Boldrini (1986). The article is restricted to the more theoretical aspects and application of the geometric method in practical situations will be subject in future work.

To exemplify the theoretical results obtained by geometric constructions, we will follow an didactic example (RENCHER; SCHAALJE, 2008, p.140).

2 The Fundamental Triangle

The geometry of multiple linear regression models is defined by a right-angled triangle according to Figure 2.

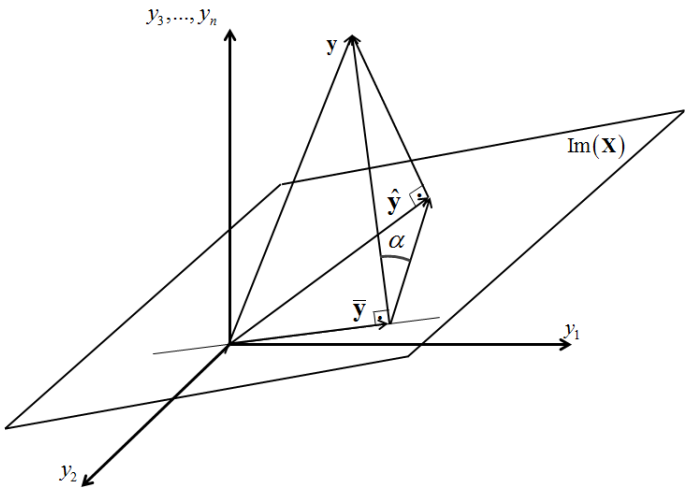

Figure 2 - The fundamental triangle.

$$(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \bar{\mathbf{y}}) = \mathbf{y} - \bar{\mathbf{y}}$$

$$(\mathbf{y} - \hat{\mathbf{y}})' (\hat{\mathbf{y}} - \bar{\mathbf{y}}) = 0$$

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = \|\mathbf{y} - \bar{\mathbf{y}}\|^2$$

$$SST = \|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\text{hypotenuse})^2$$

$$SSR = \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\text{adjacent side})^2$$

$$SSE = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\text{opposite side})^2$$

$$\text{SST} = \text{SSR} + \text{SSE}.$$

Example (RENCHER; SCHAALJE, 2008): Consider the data in Table 7.1, page 140 (TABLE 1).

Table 1 - Data for the example

Observation Number	$\mathbf{y}$	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ \vdots \\ 14 \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ \vdots & \vdots & \vdots \\ 1 & 8 & 13 \end{pmatrix}, \mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix},$$

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 2.8044 \\ 3.6861 \\ 2.4006 \\ 4.9716 \\ 5.8533 \\ 7.1388 \\ 8.4242 \\ 10.5914 \\ 9.3059 \\ 11.8769 \\ 10.1876 \\ 12.7586 \end{pmatrix}, \quad \bar{\mathbf{y}} = \frac{1}{12}\mathbf{J}\mathbf{y} = \begin{pmatrix} 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \\ 7.5 \end{pmatrix}.$$

■

The sums of squares can be explained in terms of projections in their respective subspaces.

$$
\begin{aligned}
\text{SSR} &= \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 \\
&= (\hat{\mathbf{y}} - \bar{\mathbf{y}})' (\hat{\mathbf{y}} - \bar{\mathbf{y}}) \\
&= \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} - \frac{1}{\mathrm{n}}\mathbf{J}\mathbf{y}\right)' \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} - \frac{1}{\mathrm{n}}\mathbf{J}\mathbf{y}\right) \\
&= \mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathbf{J}\right)' \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathbf{J}\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathbf{J}\right)^2 \mathbf{y} \\
&= \mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathbf{J} - \frac{1}{\mathrm{n}}\mathbf{J}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' + \frac{1}{\mathrm{n}}\mathbf{J}\right)\mathbf{y}. \quad (1)
\end{aligned}
$$

Here, let us assume that $\overrightarrow{\mathbf{1}} \in \mathrm{Im}(\mathbf{X})$. This is the hypothesis of the linear model with intercept, which is the most used. In this case, since $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ is the projection on $\mathrm{Im}(\mathbf{X})$ and $\frac{1}{\mathrm{n}}\mathbf{J}$ the projection on the subspace spanned by $\overrightarrow{\mathbf{1}}$, so $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathbf{J}\mathbf{y} = \frac{1}{\mathrm{n}}\mathbf{J}\mathbf{y}$ for all $\mathbf{y}$, i.e. , $\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathbf{J} = \frac{1}{\mathrm{n}}\mathbf{J}$.

On the other hand, $\frac{1}{\mathrm{n}}\mathbf{J}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} = \frac{1}{\mathrm{n}}\mathbf{J}\hat{\mathbf{y}} = \bar{\hat{\mathbf{y}}}$. With this insight we are able to prove the following proposition.

Proposition 2.1. *If* $\overrightarrow{\mathbf{1}} = (1, 1, \cdots, 1)' \in \mathrm{Im}(\mathbf{X})$ *then the mean of adjusted vector is equal to the mean of observed vector, i.e.,* $\bar{\hat{\mathrm{y}}} = \bar{\mathrm{y}}$.

Proof: (first proof) Consider the following decomposition into orthogonal direct sum. If $\mathrm{V}_{\overrightarrow{\mathbf{1}}}$ is the subspace spanned by the vector $\overrightarrow{\mathbf{1}}$, $\mathrm{V}_{\overrightarrow{\mathbf{1}}} = \{(\mathrm{a}, \mathrm{a}, \cdots, \mathrm{a}),\ \mathrm{a} \in \mathbb{R}\}$, then (Figure 3)

$$
\mathbb{R}^n = \mathrm{V}_{\overrightarrow{\mathbf{1}}} \oplus \left\{\mathrm{V}_{\overrightarrow{\mathbf{1}}}{}^{\perp} \cap \mathrm{Im}\,(\mathrm{X})\right\} \oplus (\mathrm{Im}\,(\mathrm{X}))^{\perp}.
$$

With this direct sum we have $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3$ with $\mathbf{v}_1 \in \mathrm{V}_{\overrightarrow{\mathbf{1}}}$, $\mathbf{v}_2 \in \left(\mathrm{V}_{\overrightarrow{\mathbf{1}}}{}^{\perp} \cap \mathrm{Im}\,(\mathrm{X})\right)$ and $\mathbf{v}_3 \in \mathrm{Im}(\mathrm{X})^{\perp}$, $\mathrm{P}_{\mathrm{V}_1}(\mathbf{v}) = \mathbf{v}_1$ and $\mathrm{P}_{\mathrm{Im}(\mathrm{X})} = \mathbf{v}_1 + \mathbf{v}_2$. Therefore,

$$
\mathrm{P}_{\mathrm{V}_1}\left(\mathrm{P}_{\mathrm{Im}(\mathrm{X})}(\mathbf{v})\right) = \mathrm{P}_{\mathrm{V}_1}(\mathbf{v}).
$$

If $\mathbf{y}$ is the observed vector, $\mathrm{P}_{\mathrm{Im}(\mathrm{X})}(\mathbf{y}) = \hat{\mathbf{y}}$ e $\mathrm{P}_{\mathrm{V}_1}(\mathbf{y}) = \bar{\mathbf{y}}$. Therefore,

$$
\bar{\mathbf{y}} = \mathrm{P}_{\mathrm{V}_1}(\mathbf{y}) = \mathrm{P}_{\mathrm{V}_1}\left(\mathrm{P}_{\mathrm{Im}(\mathrm{X})}(\mathbf{y})\right) = \mathrm{P}_{\mathrm{V}_1}(\hat{\mathbf{y}}) = \widehat{\bar{\mathbf{y}}}.
$$

■

Example (continuation):

$$
\bar{\mathbf{y}} = \frac{1}{12}\mathbf{J}\mathbf{y} = \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \end{bmatrix} \begin{pmatrix} 2 \\ 3 \\ \vdots \\ 14 \end{pmatrix} = \begin{pmatrix} 7.5 \\ 7.5 \\ \vdots \\ 7.5 \end{pmatrix},
$$

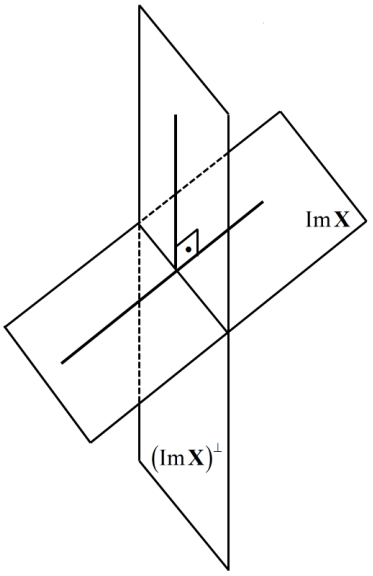

Figure 3 - $\mathbb{R}^n$ as a orthogonal direct sum.

$$\begin{aligned}
\widehat{\bar{\mathbf{y}}} &= \frac{1}{12}\mathbf{J}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \end{bmatrix} \begin{bmatrix} 0.3943 & 0.1892 & \cdots & -0.1182 \\ 0.1892 & 0.1908 & \cdots & -0.0567 \\ \vdots & \vdots & \ddots & \vdots \\ -0.1182 & -0.0567 & \cdots & 0.2854 \end{bmatrix} \begin{pmatrix} 2 \\ 3 \\ \vdots \\ 14 \end{pmatrix} \\
&= \begin{pmatrix} 7.5 \\ 7.5 \\ \vdots \\ 7.5 \end{pmatrix}.
\end{aligned}$$

■

Another more geometrical demonstration can be based on the two triangles in Figure 4.

Proof: (second proof) Suppose that the orthogonal projection of $\hat{\mathbf{y}}$ in the direction of $\overrightarrow{\mathbf{1}}$ is not $\bar{\mathbf{y}}$. Then, we have two right-angled triangles. $c^2 = b^2 + a^2$ and $e^2{=}d^2 + a^2$. As $c < e$ and $d < b$ since the projections are orthogonal, we have $c^2 < e^2$, $d^2 < b^2$. Then $b^2 + a^2 < e^2{=}a^2 + d^2 \Rightarrow b^2 < d^2$, which is a contradiction.

■

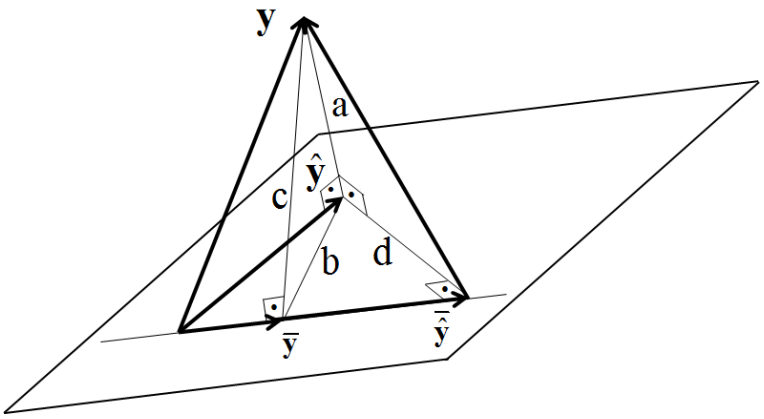

Figure 4 - Geometrical proof of Proposition 2.1.

An example that shows that hypothesis $\overrightarrow{1} \in \mathrm{Im}(\mathrm{X})$ is essential is for $\mathrm{X} = \begin{pmatrix} 1 & 1 \\ 1 & -2 \\ -2 & 1 \end{pmatrix}$, $\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, we have that $\overrightarrow{1} \notin \mathrm{Im}(\mathbf{X})$. $\hat{\mathbf{y}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$, and thus $\overline{\hat{\mathrm{y}}} = 0 \neq 2 = \bar{\mathrm{y}}$.

By Proposition 2.1, $\frac{1}{\mathrm{n}}\mathbf{J}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y} = \frac{1}{\mathrm{n}}\mathbf{J}\hat{\mathbf{y}} = \overline{\hat{\mathbf{y}}} = \bar{\mathbf{y}} = \frac{1}{\mathrm{n}}\mathbf{J}\mathbf{y}$. Using this fact the expression of SSR can be simplified.

$$\mathrm{SSR} = \mathbf{y}' \left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' - \frac{1}{\mathrm{n}}\mathbf{J} \right) \mathbf{y}.$$

For those readers who prefer the algebraic approach, a good challenge is to prove algebraically that $\frac{1}{\mathrm{n}}\mathbf{J}\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\frac{1}{\mathrm{n}}\mathbf{J} = \frac{1}{\mathrm{n}}\mathbf{J}$.

For the sum of squares SSE we have

$$
\begin{aligned}
\text{SSE} &= \|\mathbf{y}-\hat{\mathbf{y}}\|^2 \\
&= (\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}}) \\
&= \left(\mathbf{y}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right)'\left(\mathbf{y}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathbf{y}\right) \\
&= \left(\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathbf{y}\right)'\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)'\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)^2\mathbf{y} \\
&= \mathbf{y}'\left(\mathbf{I}-\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)^2\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathbf{y}.
\end{aligned}
$$

Again, the fundamental identity can be demonstrated:

$$
\begin{aligned}
\text{SSE}+\text{SSR} &= \mathbf{y}'\left(\mathbf{I}-\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\right)\mathbf{y}+\mathbf{y}'\left(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'-\frac{1}{\mathrm{n}}\mathbf{J}\right)\mathbf{y} \\
&= \mathbf{y}'\left(\mathbf{I}-\frac{1}{\mathrm{n}}\mathbf{J}\right)\mathbf{y} \\
&= \mathbf{y}'(\mathbf{y}-\bar{\mathbf{y}}) \\
&= (\mathbf{y}'-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}}) \\
&= \text{SST}.
\end{aligned}
$$

Example (continuation): $\text{SSR}=\|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|^2$=139.541, $\text{SST}=\|\mathbf{y}-\bar{\mathbf{y}}\|^2$=165, $\text{SSE}=\|\mathbf{y}-\hat{\mathbf{y}}\|^2$ = 25.459 and $\text{SST}=\text{SSR}+\text{SSE}$ =139.541 + 25.459=165. ■

3 The centered model

In certain situations, it is convenient to change the $\mathbf{X}$ matrix by $\mathbf{X}^*=\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{X}$. In this way, the new matrix columns are obtained by the orthogonal projection of the columns of $\mathbf{X}$ in the subspace $\text{Im}(\mathbf{X})$ perpendicular to the vector $\overrightarrow{\mathbf{1}}$. The columns of $\mathbf{X}$ are subtracted from the mean of the correspondent column. So, $\mathbf{X}^*$ has columns with zero sum, and it is said that the model is in centered form, i.e., centered on the mean. Geometrically, each column vector of $\mathbf{X}_i$ is subtracted from its orthogonal projection onto $\overrightarrow{\mathbf{1}}$. If $\mathbf{X}$ defines a model with intercept, $\mathbf{X}_1$ is a column containing only $1's$ and therefore $\mathbf{X}_1^*$ will be a column formed by zeros. Substituting this column by one with $1's$ we then construct a new model with intercept, $\tilde{\mathbf{X}}^*=\left[\mathbf{1},\mathbf{X}_1^*,\ldots,\mathbf{X}_p^*\right]$. Since $\text{Im}\,(\mathbf{X})=\text{Im}\left(\widetilde{\mathbf{X}}^*\right)$, the regression models, that is, the subspace where the expectation of the response vector $\mathbf{Y}$ is supposed to

be, are the same. Thus, $\mathbf{Y} = \widetilde{\mathbf{X}}^*\beta + \boldsymbol{\varepsilon}$ is just a reparametrization of $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. If $\hat{\boldsymbol{\beta}}^* = \left(\hat{\boldsymbol{\beta}}_0^*, \hat{\boldsymbol{\beta}}_1^*, \ldots, \hat{\boldsymbol{\beta}}_p^*\right)$ is the least squares estimator of the regression $\mathbf{Y} = \tilde{\mathbf{X}}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with respect to the original regression $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ follows that:

Proposition 3.1. $\hat{\boldsymbol{\beta}}_0^* = \bar{\mathrm{y}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{i}}^* = \hat{\boldsymbol{\beta}}_{\mathrm{i}}$, $\mathrm{i} = 1, \cdots, \mathrm{p}$

Proof:

$$\begin{aligned}
\hat{\mathbf{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\
&= \sum_{\mathrm{i}=0}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}} \\
&= \hat{\boldsymbol{\beta}}_0 \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}} \\
&= \hat{\boldsymbol{\beta}}_0 \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \left(\mathbf{X}_{\mathrm{i}} - \frac{1}{\mathrm{n}} \mathbf{J}\mathbf{X}_{\mathrm{i}} + \frac{1}{\mathrm{n}} \mathbf{J}\mathbf{X}_{\mathrm{i}} \right) \\
&= \hat{\boldsymbol{\beta}}_0 \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \frac{1}{\mathrm{n}} \mathbf{J}\mathbf{X}_{\mathrm{i}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \left(\mathbf{X}_{\mathrm{i}} - \frac{1}{\mathrm{n}} \mathbf{J}\mathbf{X}_{\mathrm{i}} \right) \\
&= \hat{\boldsymbol{\beta}}_0 \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \frac{1}{\mathrm{n}} \mathbf{J}\mathrm{X}_{\mathrm{i}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}}^* \\
&= \hat{\boldsymbol{\beta}}_0 \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \bar{x}_{\mathrm{i}} \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}}^* \\
&= \left(\hat{\boldsymbol{\beta}}_0 + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \bar{x}_{\mathrm{i}} \right) \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}}^*.
\end{aligned}$$

Therefore,

$$\frac{1}{\mathrm{n}} \mathbf{J}\hat{\mathbf{y}} = \frac{1}{\mathrm{n}} \mathbf{J} \left[\left(\hat{\boldsymbol{\beta}}_0 + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \bar{x}_{\mathrm{i}} \right) \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}}^* \right]$$

$$\overline{\hat{\mathbf{y}}} = \left(\hat{\boldsymbol{\beta}}_0 + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \bar{x}_{\mathrm{i}} \right) \overrightarrow{\mathbf{1}}$$

However, by proposition 2.1, $\overline{\hat{\mathbf{y}}} = \bar{\mathbf{y}} \Rightarrow \hat{\boldsymbol{\beta}}_0 + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \bar{x}_{\mathrm{i}} = \bar{\mathrm{y}}$. So,

$$\hat{\mathbf{y}} = \bar{\mathrm{y}} \overrightarrow{\mathbf{1}} + \sum_{\mathrm{i}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{i}} \mathbf{X}_{\mathrm{i}}^* = \widetilde{\mathbf{X}}^* \begin{pmatrix} \bar{\mathrm{y}} \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{\mathrm{p}} \end{pmatrix}$$

■

Example (continuation):

$$\mathbf{X}^* = \left(\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{12} & \frac{1}{12} & \cdots & \frac{1}{12} \end{bmatrix} \right) \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ \vdots & \vdots & \vdots \\ 1 & 8 & 13 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & -4.3333 & -6.5 \\ 0 & -2.3333 & -2.5 \\ \vdots & \vdots & \vdots \\ 0 & 3.6666 & 4.5 \end{pmatrix},$$

$$\tilde{\mathbf{X}}^* = \begin{pmatrix} 1 & -4.3333 & -6.5 \\ 1 & -2.3333 & -2.5 \\ \vdots & \vdots & \vdots \\ 1 & 3.6666 & 4.5 \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}}^* = \left(\tilde{\mathbf{X}}^{*'} \tilde{\mathbf{X}}^* \right)^{-1} \tilde{\mathbf{X}}^{*'} \mathbf{y} = \begin{pmatrix} 7.500 \\ 3.011 \\ -1.285 \end{pmatrix}.$$

■

The prediction equation $\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}'\mathrm{x}$ also allows an interesting geometric interpretation. While the columns of the matrix $\mathbf{X}$ are vectors in the data space $\mathbb{R}^{\mathrm{n}}$ and span the subspace $\mathrm{Im}\,(\mathrm{X})$, the rows of $\mathbf{X} = \begin{bmatrix} \mathbf{L}_1 \\ \vdots \\ \mathbf{L}_n \end{bmatrix}$ can be viewed as vectors in the parametric space $\mathbb{R}^{\mathrm{p}}$.

Since $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$, then

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} \mathbf{L}_1 \\ \mathbf{L}_2 \\ \vdots \\ \mathbf{L}_{\mathrm{n}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_{\mathrm{p}} \end{bmatrix} = \begin{bmatrix} \hat{\mathrm{y}}_1 \\ \hat{\mathrm{y}}_2 \\ \vdots \\ \hat{\mathrm{y}}_{\mathrm{p}} \end{bmatrix},$$

and thus

$$\mathbf{L}_1\hat{\boldsymbol{\beta}} = \hat{\mathrm{y}}_1, \ \ldots, \mathbf{L}_{\mathrm{n}}\hat{\boldsymbol{\beta}} = \hat{\mathrm{y}}_{\mathrm{n}}.$$

$\mathbf{L}_{\mathrm{i}} = (\mathrm{x}_{\mathrm{i}1}, \mathrm{x}_{\mathrm{i}2}, \ \ldots \ , \mathrm{x}_{\mathrm{ip}})$ is a vector of covariates values, i.e.,

$$\hat{\mathrm{y}}_{\mathrm{i}} = \mathbf{L}_{\mathrm{i}}\hat{\boldsymbol{\beta}} = \sum_{\mathrm{j}=1}^{\mathrm{p}} \hat{\boldsymbol{\beta}}_{\mathrm{j}} \mathrm{x}_{\mathrm{ij}} = \hat{\boldsymbol{\beta}}'\mathbf{x}_{\mathrm{i}}.$$

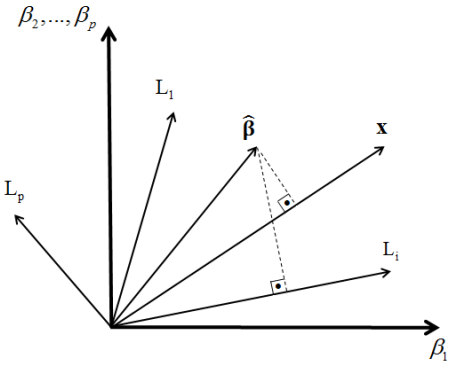

Figure 5 - Geometric interpretation of prediction equation.

If $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \ \ldots \ , \mathrm{x}_\mathrm{p})$ is a vector defined by arbitrary values of the covariates, the predicted value $\hat{\mathbf{y}}$ is then acquired by the prediction equation (Figure 5) with $\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}'\mathbf{x}$.

Proposition 3.2. *If $\bar{y}$ is the mean of the n observed values and $\bar{\mathrm{x}} = (\bar{\mathrm{x}}_1, \cdots, \bar{\mathrm{x}}_\mathrm{p})$ the vector defined by the mean of covariates used values, then $\bar{y} = \hat{\beta}'\,\bar{\mathrm{x}}$.*

Proof:

$$\hat{\mathrm{y}}_\mathrm{i} = \mathbf{L}_\mathrm{i}\hat{\boldsymbol{\beta}}$$

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\hat{\mathrm{y}}_\mathrm{i} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\mathbf{L}_\mathrm{i}\hat{\boldsymbol{\beta}}$$

$$\begin{aligned}
\overline{\hat{\mathbf{y}}} &= \left(\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\mathbf{L}_\mathrm{i}\right)\hat{\boldsymbol{\beta}} \\
&= \left(\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\mathrm{x}_{\mathrm{i}1}, \mathrm{x}_{\mathrm{i}2}, \ \ldots \ , \mathrm{x}_{\mathrm{ip}})\right)\hat{\boldsymbol{\beta}} \\
&= (\bar{\mathrm{x}}_1, \bar{\mathrm{x}}_2, \ \ldots \ , \bar{\mathrm{x}}_\mathrm{p})\,\hat{\boldsymbol{\beta}}.
\end{aligned}$$

Since $\overline{\hat{\mathbf{y}}} = \bar{\mathbf{y}}$, then $\bar{\mathbf{y}} = \hat{\boldsymbol{\beta}}'\,\bar{\mathbf{x}}$.

■

4 Geometric interpretation of determination coefficient

The coefficient of determination in a multiple regression is expressed in terms of the fundamental triangle as the quotient :

$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \left(\frac{\text{adjacent side}}{\text{hypotenuse}}\right)^2 = \cos^2(\alpha)\,.$$

Such R^2 is defined as the squared cosine of the angle α in the fundamental triangle, its interpretation becomes much more evident. If α is approximately 0° $\left(R^2 \approx 1\right)$ this indicates that SSE ≈ 0, i.e., the triangle opposite side is small relative to the measure of the hypotenuse and therefore it can be said that the regression (the adjusted vector $\hat{\mathbf{y}}$) is close enough to the data represented by the vector $\mathbf{y}$, and hence explains well the data. In the other hand, α close to 90° $\left(R^2 \approx 0\right)$ has the opposite explanation.

Also follows in the fundamental triangle, by the vector inner product concept, the characterization of R^2 as

$$R^2 = \cos^2(\alpha) = \left(\frac{(\mathbf{y} - \bar{\mathbf{y}})' (\hat{\mathbf{y}} - \bar{\mathbf{y}})}{\|\mathbf{y} - \bar{\mathbf{y}}\| \ \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|} \right)^2.$$

By Proposition 2.1, $\overline{\hat{\mathbf{y}}} = \bar{\mathbf{y}}$ and therefore,

$$R^2 = \left(\frac{(\mathbf{y} - \bar{\mathbf{y}})' (\hat{\mathbf{y}} - \overline{\hat{\mathbf{y}}})}{\|\mathbf{y} - \bar{\mathbf{y}}\| \ \|\hat{\mathbf{y}} - \overline{\hat{\mathbf{y}}}\|} \right)^2$$

which is the sampling correlation between the random variables $\mathbf{Y}$ and $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

In the simple regression case, another interesting interpretation is also possible. $\hat{\mathbf{y}} - \overline{\hat{\mathbf{y}}}$ is a vector belonging to $\mathrm{Im}(\mathbf{X})$ and is perpendicular to the unidimensional subspace spanned by $\overrightarrow{\mathbf{1}}$. The column vector $\mathbf{x}$ of $\mathbf{X} = (\mathrm{j},\mathbf{x})$ subtracted from its means $\mathbf{x} - \bar{\mathbf{x}}$ is also perpendicular to this same subspace. As $\dim(\mathrm{Im}(\mathbf{X})) = 2$, then the vectors $\hat{\mathbf{y}} - \overline{\hat{\mathbf{y}}}$ and $\mathbf{x} - \bar{\mathbf{x}}$ are parallels each other and therefore has the same angle with $\mathbf{y} - \bar{\mathbf{y}}$. Therefore,

$$R = \cos(\alpha) = \frac{(\mathbf{y} - \bar{\mathbf{y}})' (\hat{\mathbf{y}} - \overline{\hat{\mathbf{y}}})}{\|\mathbf{y} - \bar{\mathbf{y}}\| \ \|\hat{\mathbf{y}} - \overline{\hat{\mathbf{y}}}\|} = \frac{(\mathbf{y} - \bar{\mathbf{y}})' (\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{y} - \bar{\mathbf{y}}\| \ \|\mathbf{x} - \bar{\mathbf{x}}\|}$$

and R can be viewed as the sample correlation coefficient between $\mathbf{Y}$ and the $\mathbf{x}$ covariate.

Another property that admits an immediate demonstration, much simpler than the usual algebraic proof is

Proposition 4.1. *For a regression linear model* $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$*, if a covariate is added then the model with this new covariate,*

$$\mathbf{Y}_{n\times 1} = \tilde{\mathbf{X}}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

$\tilde{\mathbf{X}}_{n\times(p+1)} = \begin{pmatrix} & & * \\ & \mathrm{X} & * \\ & & \vdots \\ & & * \end{pmatrix}$*, has the* R^2 *value increased.*

Proof: The subspace $\mathrm{Im}(\mathbf{X})$ is contained in $\mathrm{Im}\left(\tilde{\mathbf{X}}\right)$. In the original model, $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathrm{Im}(\mathbf{X})$. Also, in the augmented model, $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ in the subspace $\mathrm{Im}\left(\tilde{\mathbf{X}}\right)$. Thus, the norm $\|\mathbf{y}-\hat{\mathbf{y}}\|$ decreases and the norm $\|\mathbf{y}-\bar{\mathbf{y}}\|$ remains the same. Then $\hat{\mathbf{y}}-\bar{\mathbf{y}}$, the adjacent side, increases and, consequently, α becomes lower. In this way, R^2 increases.

■

Example (continuation): Adding a new covariate given in Table 2.

Table 2 - Data for the example

Observation Number	$\mathbf{y}$	x_1	x_2	x_3
1	2	0	2	4
2	3	2	6	8
3	2	2	7	8.5
4	7	2	5	7
5	6	4	9	11
6	8	4	8	6
7	10	4	7	11
8	7	6	10	7
9	8	6	11	12
10	12	6	9	12
11	11	8	15	16
12	14	8	13	10

For the original data with x_1 and x_2 we have $R^2 = 0.8457$ for the augmented model. The sample correlaction between x_1 and x_3 is 0.7186 and the sample correlaction between x_2 and x_3 is 0.7742. Then,the new coefficient of determination is

$$\mathrm{R}^2 = \left(\frac{(\mathbf{y}-\bar{\mathbf{y}})'(\hat{\mathbf{y}}-\bar{\mathbf{y}})}{\|\mathbf{y}-\bar{\mathbf{y}}\|\,\|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|}\right)^2 = 0.8653.$$

■

A common mistake in using the concept of the coefficient of determination R^2 is , despite being a measure of how the model fits the data, it can't be used as a criterion of the quality to the predictive model, which is how much the model explains the unobserved data. In other words we can't use R^2 as a criterion for model selection. A suitable technique for this is the use of the C_p Mallows's statistic, that also admits a geometric description in terms of the fundamental triangle.

5 The Mallows's C_p statistic

This section follows the ideas of Sheden (2014). Consider a linear model denoted by M, with $\mathbf{Y}_{\mathrm{n}} = \mathbf{X}_{\mathrm{n}\times\mathrm{p}}\boldsymbol{\beta}_{\mathrm{p}\times 1} + \boldsymbol{\varepsilon}_{\mathrm{n}\times 1}$, $\mathrm{E}[\mathbf{Y}] = \boldsymbol{X\beta}$ and $\mathrm{cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$. The model assumes that the parameter mean of $\mathbf{Y}$ belongs to the linear subspace $\mathrm{Im}(\mathbf{X})$. If the model is not correct then $\mathrm{E}[\mathbf{Y}] = \boldsymbol{\theta} \neq \boldsymbol{X\beta}$. In the incorrect model M, the closest vector to $\boldsymbol{\theta}$ will be denoted by $\theta_{\mathbf{X}}$ and is determined by the orthogonal decomposition $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathbf{X}} + \boldsymbol{\theta}_{\mathbf{X}^{\perp}}$ with respect to the subspace $\mathrm{Im}(\mathbf{X})$. $\theta_{\mathbf{X}}$ is the parametric vector which will be estimated by the model M. In the correct model $\mathbf{Y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon}$, the error vector can also be decomposed as $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_{\mathrm{X}} + \boldsymbol{\varepsilon}_{\mathrm{X}^{\perp}}$. The covariance matrix of the random vector $\hat{y}$ is given by

$$\begin{aligned}
\mathrm{E}\left[(\hat{\mathbf{y}}-\boldsymbol{\theta})(\hat{\mathbf{y}}-\boldsymbol{\theta})'\right] &= \mathrm{E}\left[(\boldsymbol{\theta}_{\mathbf{X}}+\boldsymbol{\varepsilon}_{\mathbf{X}}-(\boldsymbol{\theta}_{\mathbf{X}}+\boldsymbol{\theta}_{\mathbf{X}^{\perp}}))(\boldsymbol{\theta}_{\mathbf{X}}+\boldsymbol{\varepsilon}_{\mathbf{X}}-(\boldsymbol{\theta}_{\mathbf{X}}+\boldsymbol{\theta}_{\mathbf{X}^{\perp}}))'\right] \\
&= \mathrm{E}\left[(\boldsymbol{\varepsilon}_{\mathbf{X}}-\boldsymbol{\theta}_{\mathbf{X}^{\perp}})(\boldsymbol{\varepsilon}_{\mathbf{X}}-\boldsymbol{\theta}_{\mathbf{X}^{\perp}})'\right] \\
&= \mathrm{E}\left[\boldsymbol{\varepsilon}_{\mathbf{X}}\boldsymbol{\varepsilon}'_{\mathbf{X}}+\boldsymbol{\theta}_{\mathbf{X}^{\perp}}\boldsymbol{\theta}'_{\mathbf{X}^{\perp}}\right] \\
&= \mathrm{E}\left[\boldsymbol{\varepsilon}_{\mathbf{X}}\boldsymbol{\varepsilon}'_{\mathbf{X}}\right]+\boldsymbol{\theta}_{\mathbf{X}^{\perp}}\boldsymbol{\theta}'_{\mathbf{X}^{\perp}} \\
&= \sigma^2\mathbf{I}_{p\times p}+\boldsymbol{\theta}_{\mathbf{X}^{\perp}}\boldsymbol{\theta}'_{\mathbf{X}^{\perp}},
\end{aligned}$$

and mean square error

$$\begin{aligned}
EQM(\hat{\mathbf{y}}) &= \mathrm{E}\left[(\hat{\mathbf{y}}-\theta)'(\hat{\mathbf{y}}-\theta)\right] \\
&= tr\left(\mathrm{E}\left[(\hat{\mathbf{y}}-\theta)'(\hat{\mathbf{y}}-\theta)\right]\right) \\
&= tr\left(\mathrm{E}\left[(\hat{\mathbf{y}}-\theta)(\hat{\mathbf{y}}-\theta)'\right]\right) \\
&= tr\left[\sigma^2\boldsymbol{I}+\boldsymbol{\theta}_{\mathbf{X}^{\perp}}\boldsymbol{\theta}'_{\mathbf{X}^{\perp}}\right] \\
&= p\sigma^2+\|\boldsymbol{\theta}_{\mathbf{X}^{\perp}}\|^2.
\end{aligned}$$

When choosing among various models, we generally chose one whose estimator has smaller mean square error. As $EQM(\hat{\mathbf{y}})$ depends on the population parameters, we seek then a unbiased estimator to this value. First of all, is necessary to calculate the expectation of $\hat{\sigma}^2 = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}{n-p}$.

$$
\begin{aligned}
\mathrm{E}\left[\hat{\sigma}^2\right] &= \mathrm{E}\left[\frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}{n-p}\right] \\
&= \mathrm{E}\left[\frac{\mathbf{y}'\left(\mathbf{I}-\mathrm{P}_{\mathbf{X}}\right)\mathbf{y}}{n-p}\right] \\
&= \frac{1}{n-p}\mathrm{E}\left[(\boldsymbol{\theta}+\boldsymbol{\varepsilon})'\left(\boldsymbol{I}-\mathrm{P}_{\mathrm{X}}\right)(\boldsymbol{\theta}+\boldsymbol{\varepsilon})\right] \\
&= \frac{1}{n-p}\mathrm{E}\left[(\boldsymbol{\theta}+\boldsymbol{\varepsilon})'\left(\boldsymbol{\theta}_{\mathrm{X}^\perp}+\boldsymbol{\varepsilon}_{\mathrm{X}^\perp}\right)\right] \\
&= \frac{1}{n-p}\mathrm{E}\left[(\boldsymbol{\theta}_{\mathbf{X}^\perp}+\boldsymbol{\varepsilon}_{X^\perp})'(\boldsymbol{\theta}_{\mathbf{X}^\perp}+\boldsymbol{\varepsilon}_{X^\perp})\right] \\
&= \frac{1}{n-p}\mathrm{E}\left[(\boldsymbol{\theta}_{\mathbf{X}^\perp})'\theta_{\mathbf{X}^\perp}+2(\boldsymbol{\theta}_{\mathbf{X}^\perp})'\boldsymbol{\varepsilon}_{X^\perp}+(\boldsymbol{\varepsilon}_{X^\perp})'\boldsymbol{\varepsilon}_{X^\perp}\right] \\
&= \frac{1}{n-p}\mathrm{E}\left[(\boldsymbol{\theta}_{\mathbf{X}^\perp})'\boldsymbol{\theta}_{\mathbf{X}^\perp}+(\boldsymbol{\varepsilon}_{X^\perp})'\boldsymbol{\varepsilon}_{X^\perp}\right] \\
&= \frac{1}{n-p}\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2+\frac{1}{n-p}\mathrm{E}\left[\|\boldsymbol{\varepsilon}_{X^\perp}\|^2\right] \\
&= \frac{1}{n-p}\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2+\frac{1}{n-p}(n-p)\sigma^2 \\
&= \frac{1}{n-p}\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2+\sigma^2.
\end{aligned}
$$

We need now an unbiased estimator for σ^2. This can be obtained if we have a much higher dimension model that contains the model M as a submodel. Such higher dimension model, ideally contains more variables and much more observatios. With this model we can obtain a good estimative of σ^2. This estimative will be denoted by $\hat{\sigma}^{*2}$. We need too an unbiased estimator of $\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2$. As $\mathrm{E}\left[\hat{\sigma}^2\right] = \frac{1}{n-p}\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2+\sigma^2$ so $\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2 = (n-p)\,\mathrm{E}\left[\hat{\sigma}^2\right]-(n-p)\,\sigma^2$ an unbiased estimator is

$$
\widehat{\|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2} = (n-p)\,\hat{\sigma}^2-(n-p)\,\sigma^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|^2-(n-p)\,\sigma^2.
$$

Therefore an unbiased estimator for $EQM\left(\hat{\mathbf{y}}\right)$ is

$$
\widehat{EQM}\left(\hat{\mathbf{y}}\right) = \|\mathbf{y}-\hat{\mathbf{y}}\|^2-(n-p)\,\hat{\sigma}^{*2}+p\hat{\sigma}^{*2} = \|\mathbf{y}-\hat{\mathbf{y}}\|^2+2p\hat{\sigma}^{*2}-n\hat{\sigma}^{*2}
$$

because

$$
\begin{aligned}
\mathrm{E}\left[\widehat{EQM}\left(\hat{\mathbf{y}}\right)\right] &= \mathrm{E}\left[\|\mathbf{y}-\hat{\mathbf{y}}\|^2-(n-p)\,\hat{\sigma}^{*2}+p\hat{\sigma}^{*2}\right] \\
&= (n-p)\,\mathrm{E}\left[\hat{\sigma}^2\right]-(n-p)\,\sigma^2+p\sigma^2 \\
&= \|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2+(n-p)\,\sigma^2-(n-p)\,\sigma^2+p\sigma^2 \\
&= \|\boldsymbol{\theta}_{\mathbf{X}^\perp}\|^2+p\sigma^2 \\
&= EQM\left(\hat{\mathbf{y}}\right).
\end{aligned}
$$

This justify the definition of the Mallows's C_p statistic as

$$\begin{aligned}
\mathrm{C}_p &= \frac{E\widehat{Q}M\left(\hat{\mathbf{y}}\right)}{\hat{\sigma}^{*2}} \\
&= \frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2 + 2p\hat{\sigma}^{*2} - n\hat{\sigma}^{*2}}{\hat{\sigma}^{*2}} \\
&= \frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2}{\hat{\sigma}^{*2}} + 2p - n \\
&= \frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2}{\hat{\sigma}^{*2}} - (n-2p)\,.
\end{aligned}$$

The correct model is $Y = \theta + \epsilon$, that is, we are use all the data and without any supposition about the covariates. In this way we have the unbiased estimator of σ^2 by $\hat{\sigma}^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$. Observe that $\hat{\sigma}^{*2}$ is taken regardless of the models in study.

As we must prefer models with lowest mean square error the idea is to calculate the value of C_p statistic and the correct model. In a p-dimensional correct model situation, $\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2 \approx (n-p)\,\sigma^2$, and $\hat{\sigma}^{*2} \approx \sigma^2$. So,

$$\mathrm{C}_p = \frac{\left\|\mathbf{y}-\hat{\mathbf{y}}\right\|^2}{\hat{\sigma}^{*2}} - (n-2p) \approx \frac{(n-p)\,\sigma^2}{\sigma^2} - (n-2p) = p.$$

So, when choosing between models we must choose one with $\mathrm{C}_p \approx p$.

Originally, C_p statistic was defined for covariate selection models (MALLOWS, 1973), that is, if $\mathrm{x}_1, \ldots, \mathrm{x}_k$ are k covariates, we want to select p covariates that best describe the response variable. Is possible in this context to have nested models. In this case, it is advantageous to introduce an alternative form of C_p statistic, as follows.

Let $\mathrm{r} = \mathrm{k} - \mathrm{p}$ be the number of removed covariates. Considering the orthogonal projections $\hat{\mathbf{y}}_\mathrm{k}$ of $\mathbf{y}$ data vector on the k-dimensional subspace defined by all covariates and $\hat{\mathbf{y}}_\mathrm{p}$ on the p-dimensional subspace defined by chosen covariates we have again the fundamental triangle, as in Figure 6.

Considering as correct model the model with the k covariates then we have

$$\mathrm{C}_k = \frac{\left\|\mathbf{y}-\hat{\mathbf{y}}_\mathrm{k}\right\|^2}{\hat{\sigma}^{*2}} - (n-2k) \approx \mathrm{k}$$

and

$$\frac{\left\|\mathbf{y}-\hat{\mathbf{y}}_\mathrm{k}\right\|^2}{\hat{\sigma}^{*2}} \approx \mathrm{n} - \mathrm{k}.$$

The C_p statistic for the model with p selected covariates can be expressed by

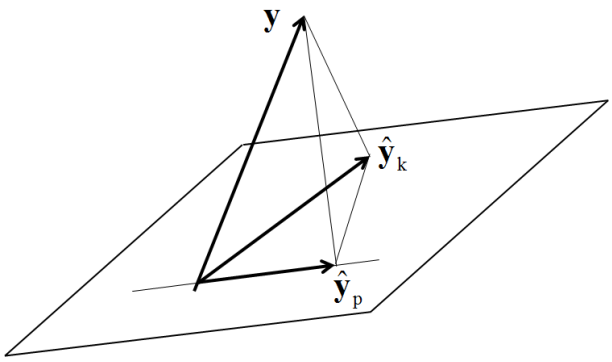

Figure 6 - Adjusted vector for nested models.

$$
\begin{aligned}
\mathrm{C}_p &= \frac{\|\mathbf{y}-\hat{\mathbf{y}}_{\mathrm{p}}\|^2}{\hat{\sigma}^{*2}} - (\mathrm{n}-2\mathrm{p}) \\
&= \frac{\|\mathbf{y}-\hat{\mathbf{y}}_{\mathrm{k}}\|^2 + \|\hat{\mathbf{y}}_{\mathrm{k}}-\hat{\mathbf{y}}_{\mathrm{p}}\|^2}{\hat{\sigma}^{*2}} - (\mathrm{n}-2\mathrm{p}) \\
&= \frac{\|\hat{\mathbf{y}}_{\mathrm{k}}-\hat{\mathbf{y}}_{\mathrm{p}}\|^2}{\hat{\sigma}^{*2}} + \frac{\|\mathbf{y}-\hat{\mathbf{y}}_{\mathrm{k}}\|^2}{\hat{\sigma}^{*2}} - (\mathrm{n}-2\mathrm{p}) \\
&\approx \frac{\|\hat{\mathbf{y}}_{\mathrm{k}}-\hat{\mathbf{y}}_{\mathrm{p}}\|^2}{\hat{\sigma}^{*2}} + (\mathrm{n}-\mathrm{k}) - (\mathrm{n}-2\mathrm{p}) \\
&= \frac{\|\hat{\mathbf{y}}_{\mathrm{k}}-\hat{\mathbf{y}}_{\mathrm{p}}\|^2}{\hat{\sigma}^{*2}} + (2\mathrm{p}-\mathrm{k})\,.
\end{aligned}
$$

6 Geometrical interpretation of C_p statistic

The geometry of construction of C_p statistic can be summarized in the Figure 7.

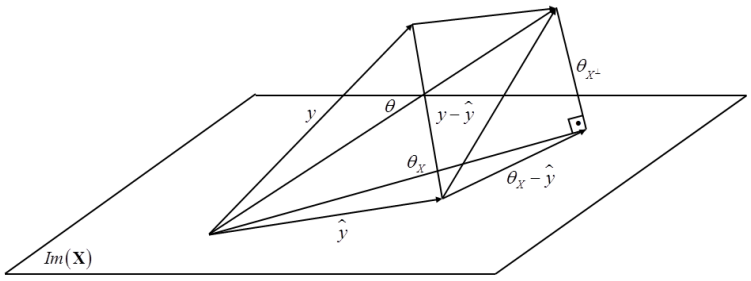

Figure 7 - Geometry of C_p.

The fundamental triangle is given by Figure 8.

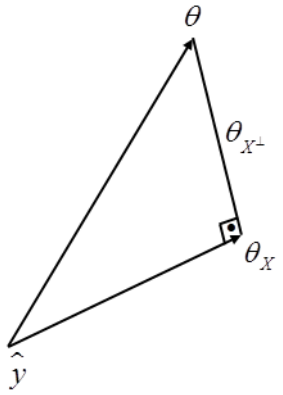

Figure 8 - Fundamental triangle.

$$\|\hat{\mathbf{y}} - \theta\|^2 = \widehat{\|\theta_{\mathbf{X}^{\perp}}\|^2} + \|\hat{\mathbf{y}} - \theta_X\|^2,$$

As $\hat{y}$ is an unbiased estimator of θ_X $\|\hat{\mathbf{y}} - \theta_X\|^2$ can be estimate by $p\sigma^{*2}$. In the subspace orthogonal to the plane define by the model we have

$$\mathbf{y} - \hat{\mathbf{y}} = \theta_{\mathbf{X}^{\perp}} + \varepsilon_{\perp}.$$

Therefore,

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \left(\theta_{\mathbf{X}^{\perp}} + \varepsilon_{\perp}\right)' \left(\theta_{\mathbf{X}^{\perp}} + \varepsilon_{\perp}\right) = \|\theta_{\mathbf{X}^{\perp}}\|^2 + 2\left(\theta_{\mathbf{X}^{\perp}}\right)' \varepsilon_{\perp} + \|\varepsilon_{\perp}\|^2.$$

So, as $E\left[\left(\theta_{\mathbf{X}^{\perp}}\right)' \varepsilon_{\perp}\right] = 0$ and $E\left[\|\varepsilon_{\perp}\|^2\right] = (n-p)\,\sigma^2$ then

$$E\left[\|\mathbf{y} - \hat{\mathbf{y}}\|^2\right] = \|\theta_{\mathbf{X}^{\perp}}\|^2 + (n-p)\,\sigma^2.$$

In this case we can estimate $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ by $\|\theta_{\mathbf{X}^{\perp}}\|^2 + (n-p)\,\sigma^2$. So,

$$\|\theta_{\mathbf{X}^{\perp}}\|^2 \approx \|\mathbf{y} - \hat{\mathbf{y}}\|^2 - (n-p)\,\widehat{\sigma}^{*2}.$$

Finally, the hipotenuse of the fundamental triangle that is a mesure of the mean square error of the model is

$$\|\hat{\mathbf{y}} - \theta\|^2 \approx \|\mathbf{y} - \hat{\mathbf{y}}\|^2 - (n-p)\,\widehat{\sigma}^{*2} + p\widehat{\sigma}^{*2} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + (2p-n)\,\widehat{\sigma}^{*2}.$$

Example (continued): Consider the models M_1 with only the covariate x_1, M_2 with the covariate x_2, M_3 with the covariate x_3, M_4 with the covariates x_1 and x_2,

M_5 with the covariates x_1 and x_3, M_6 with the covariates x_2 and x_3 and the model M_7 with the covariates x_1, x_2, and x_3. The C_p values for these models are given in the Table 3.

Table 3 - Results for selection of different models with C_p statistic

Model	C_p	$p - C_p$
M_1	-4.9849	6.9849
M_2	-2.5713	4.5713
M_3	-0.9301	2.9301
M_4	-4.3027	7.3027
M_5	-2.9897	5.9897
M_6	-0.6314	3.6314
M_7	-2.5187	6.5187

The choosen model by C_p statistic is the model M_3.

Conclusions

The geometric method is a natural approach to the various aspects of linear model theory. It has the advantage of emphasizing the statistical concepts involved and also has generality, since the various problems can be solved by essentially the same procedures.

PEREIRA, L. S.; CHAVES, L. M.; SOUZA, D. J. Geometria de resultados básicos em regressão linear e da estatística C_p de Mallows.. *Rev. Bras. Biom.,* São Paulo, v.33, n.3, p.357-377, 2015.

- ABSTRACT: Uma abordagem completamente geométrica, empregando somente subespaços vetoriais e projeções ortogonais, é aplicada para a teoria dos modelos lineares. Resultados básicos, que são usualmente provados em textos didáticos utilizam álgebra matricial não trivial, são demonstrados usando somente geometria. Como uma aplicação não usual, o método é aplicado à construção da estatística C_p de Mallows.

- KEYWORDS: Método geométrico; C_p de Mallows; Modelos Lineares; Projeções ortogonais.

References

BAILEY, R. A. *Design of Comparative Experiments*. Cambridge: Cambridge University Press, 2008. 330p.

BOLDRINI, J. L. et al. *Álgebra Linear*. 2.ed. São Paulo: Harbra, 1986. 411p.

BRING, J. A Geometric Approach to Compare Variables in a Regression Model. *The American Statistician*, Alexandria VA USA, v.50, p.57-62, 1996.

BRYANT, P. Geometry, Statistics, Probability: Variations on a Common Theme. *The American Statistician*, Alexandria VA USA, v.38, n.1, p.38-48, 1984.

DRAPER, N. R.; SMITH, H. *Applied Regression Analysis*. New Jersey: John Wiley & Sons, 1998. 736p.

HERR, D. G. On the History of the Use of Geometry in the General Linear Model. *The American Statistician*, Alexandria VA USA, v.34, n.1, p.43-47, 1980.

MALLOWS, C. L. Some Comments on C P. *Technometrics*, Alexandria VA USA, v.15, n.4, p.661-675, 1973.

MARGOLIS, M. S. Perpendicular Projections and Elementary Statistics. *The American Statistician*, Alexandria VA USA, v.33, n.3, p.131-135, 1979.

RAO, C. R. *Linear Statistical Inference and its Applications*. New Jersey: John Wiley & Sons, 1973.

RENCHER, A. C; SCHAALJE, G. B. *Linear Models in statistics*. New Jersey: John Wiley & Sons, 2008. 672p.

SAVILLE, D. J; WOOD, G. L. A Method for Teaching Statistics Using N-Dimensional Geometry. *The American Statistician*, Alexandria VA USA, v.40, p.205-214, 1986.

SAVILLE, D. J; WOOD, G. L. *Statistical Methods*: *The Geometric Aproach*. New York: Springer-Verlag, 1991. 560p.

SEARLE, S. R. *Linear Models*. New Jersey: John Wiley & Sons, 1971.

SEBER, G. A. F.; LEE, A. J. *Linear Regression Analysis*. New Jersey: John Wiley & Sons, 2003.

SHEDEN, K. *Model selection*. University of Michigan, Department of Statistics, 2014. 25p. `http://dept.stat.lsa.umich.edu/~kshedden/Courses/Stat600/Notes/model_selection.pdf` acessed in 06-24-2015.

Recebido em 27.11.2014.

Aprovado após revisão em 30.07.2015.