Geometry of the Linear Model

V. Línek
Charles University Prague, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. The advantage of the geometric approach to linear model and its applications is known to many authors. In spite of that, it still remains to be rather unpopular in teaching statistics around the world and is almost missing in the Czech Republic. In this article, we use geometry to derive some well-known properties of the linear model and to explain some of the most familiar statistical methods to show usefulness of this attitude. Historical background is briefly discussed, too.

Normal linear model

In this article, we consider a normal linear model described in [Anděl, 1998], in which a random vector \( Y \) of \( n \) components satisfies a condition

\[ Y = X\beta + Z, \]

where \( X_{\text{con}} \) is a known matrix, \( \beta \) is an unknown vector of \( m \) parameters and \( Z \) follows a normal distribution \( N(0; \sigma^2 I_n) \), e.g. its components are mutually independent and have the same unknown variance and zero expected value. Having an actual value of \( Y \), the usual goals are to estimate \( \sigma^2 \), \( X\beta \) or even \( \beta \) and to find out whether it is possible to make the model more precise. We will show simple examples of how these goals can be achieved easily through the geometrical interpretation of the model.

Estimating \( X\beta \): an orthogonal projection

From a geometric point of view, the linear model says that \( X\beta \), the expected value of \( Y \) lies somewhere in the subspace spanned by the columns of \( X \); let’s denote this subspace by \( M \). It is natural to estimate \( X\beta \) by an orthogonal projection of \( Y \) into \( M^1 \), i.e., by a vector \( Y_M \) which lies in \( M \) with \( Y - Y_M \) being perpendicular to \( M \). Thus \( Y_M \) can definitely be found from the following conditions:

\[ Y_M = Xb, \]
\[ X^T(Y - Y_M) = 0. \]

When the columns of \( X \) are independent, we get

\[ b = (X^TX)^{-1}XY, \]
\[ Y_M = X(X^TX)^{-1}XY, \]

where \( b \) is the best unbiased estimate of \( \beta \) because it is a linear function of \( Y_M \):

\[ b = (X^TX)^{-1}X^TY_M. \]

Probability characteristics of orthogonal projections

Let’s suppose that the random vector \( Y \) of \( n \) components follows the normal distribution \( N(0; I_n) \) and that we project it orthogonally into two subspaces \( M \) and \( K \) of dimensions \( m \) and \( k \), which are orthogonal to each other. It is easy to prove that

\[ \|Y_M\|^2 \sim \chi^2_m \quad \text{and} \quad \frac{\|Y_M\|^2}{m} \sim F_{m,k}. \]

First, we introduce an alternative orthonormal base of \( \mathbb{R}^n \), formed by the vectors \( e_1, \ldots, e_n \) such that

\[1\text{ Recall that the Gauss-Markov theorem states that this projection is the best linear unbiased estimator of } X\beta. \]
Let's denote this projection with \( Y' \). Thus we get

\[
Y' = (Y_1', ..., Y_n') \quad \text{means} \quad Y = Y_1 e_1 + ... + Y_n e_n.
\]

Then \( Y' = \mathbf{B} Y \), where \( \mathbf{B} \) is an orthonormal matrix. Thus we get

\[
Y' \sim \mathcal{N}(\mathbf{B} \cdot 0; \mathbf{B} \cdot \mathbf{I} \cdot \mathbf{B}^T) = \mathcal{N}(0; \mathbf{I}_n).
\]

In other words, the components of \( Y' \) follow the \( \mathcal{N}(0; 1) \) distribution and they are independent, as zero covariance means independence in the case of normally distributed random values.

Using the alternative system of coordinates, it is obvious that \( Y_M' = (Y_1', ..., Y_m', 0, ..., 0) \) and \( Y_K' = (0, ..., 0, Y_{m+1}', ..., Y_{m+k}', 0, ..., 0) \); so we get

\[
\sum_{i=1}^{m} (Y_i')^2 \sim \chi^2_m \quad \text{and} \quad \sum_{i=m+1}^{m+k} (Y_i')^2 / k \sim F_{m,k}^2.
\]

But the sums in these formulas represent just the squared lengths of \( Y_M' \) and \( Y_K' \), as the squared length of a vector can be calculated by summing the squares of its coordinates related to any orthonormal base.

Finally we can make our results more general. When the variation matrix of \( Y \) is \( \sigma I_n \) instead of \( I_n \), then the variation matrix of \( Y/\sigma \) is \( I_n \), so we only need to divide the left side of the first formula by \( \sigma^2 \) to make it correct; the second formula doesn’t change, as the \( \sigma \)'s cancel each other.

**Estimating \( \sigma^2 \)**

When we have \( \mathbf{X}\beta \) estimated with \( Y_M \), the orthogonal projection of \( Y \) into \( \mathbf{M} \), we can realise that \( Y - Y_M \) is another orthogonal projection of \( Y \) into \( \mathbf{M}^\perp \), the orthogonal complement of \( \mathbf{M} \) of dimension \( n - m \) (remember that we suppose that the columns of \( \mathbf{X} \) are independent, so the dimension of \( \mathbf{M} \) is \( m \)).

Further, because \( \mathbf{X}\beta \) lies in \( \mathbf{M} \), \( Y - Y_M \) is a projection of \( Y - \mathbf{X}\beta \) into \( \mathbf{M}^\perp \), too (see the Fig. 1). As the expected value of \( Y - \mathbf{X}\beta \) is \( 0 \), we can use the results from the previous paragraph and deduce that

\[
\frac{\|Y - Y_M\|^2}{\sigma^2} \sim \chi^2_{n-m},
\]

\[
\mathbb{E} \left( \frac{\|Y - Y_M\|^2}{\sigma^2} \right) = n - m,
\]

\[
\mathbb{E} \left( \frac{\|Y - Y_M\|^2}{n - m} \right) = \sigma^2.
\]

This means that the last expression in the brackets is an unbiased estimator of \( \sigma^2 \). It is usually denoted by \( s^2 \):

\[
s^2 = \frac{\|Y - Y_M\|^2}{n - m}.
\]

**Testing submodel**

As we will see later, any null hypothesis usually means the idea that the unknown expected value of \( Y \) lies in some particular subspace \( \mathbf{S} \) inside \( \mathbf{M} \) and is called the submodel. Let \( s \) be the dimension of \( \mathbf{S} \) (\( s < m \)). To verify the validity of the hypothesis, the first step is projecting \( Y \) orthogonally into \( \mathbf{S} \). Let’s denote this projection with \( Y_S \). Now we can write

\[
Y - \mathbf{X}\beta = (Y_S - \mathbf{X}\beta) + (Y_M - Y_S) + (Y - Y_M).
\]

\[2\]Remember possible definitions of these two important distributions, as they are presented for example in [Zvára & Štěpán, 2006]: when independent random values \( Y_1, ..., Y_m, Y_{m+1}, ..., Y_{m+k} \) follow the \( \mathcal{N}(0;1) \) distribution, then

\[
\sum_{i=1}^{m} Y_i^2 \sim \chi^2_m \quad \text{and} \quad \sum_{i=m+1}^{m+k} Y_i^2 / k \sim F_{m,k}
\]

and the expected value of the first expression is \( m \).
Figure 1. A scheme of geometrical relationships between the random vector $Y$, its orthogonal projection $Y_M$ into $M$ and its expected value $X\beta$.

Notice that $Y_M - Y_S$ is perpendicular to $S$ and so it lies inevitably in $M - S$, the orthogonal complement of $S$ inside $M$, the dimension of which is $m - s$. Provided the null hypothesis is correct, $Y_S - X\beta$ lies in $S$ and all three vectors in brackets in the formula above are orthogonal to each other. This means that we have projected the error vector $Z = Y - X\beta$ orthogonally into three subspaces, which are all orthogonal to each other. Thus we can make use of the second formula from the previous paragraph to claim that

$$\frac{\|Y_M - Y_S\|^2}{(m - s)} \sim F_{m-s, n-m}.$$

Because it is natural to reject the null hypothesis when $Y_M - X_S$ is too long compared to $Y - Y_M$, we have the proper statistics which we can use to make our decision – we reject the null hypothesis when the actual value is higher than the critical one.

The table below shows the usual form in which the results of this process are presented; we replaced the usual notation with the labels of the relevant geometrical objects. For better orientation and further thinking, the picture below hopefully can help.

**An example: Linear regression**

Let’s suppose that $Y$ is composed of seven components $Y_i$, the expected values of which depend linearly on three more variables. In addition, there is a basic level common to all components and the random components $Z_i$:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_7 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ \vdots \\ 1 & x_{71} & x_{72} & x_{73} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{pmatrix} + \begin{pmatrix} Z_1 \\ \vdots \\ Z_7 \end{pmatrix}.$$

Here, the last three columns of the matrix represent values of the independent quantities, each of them being determined seven times. The values $\beta_1, \beta_2, \beta_3$ are the unknown coefficients of the dependence and $\beta_0$ is the basic level.

When we get the actual value of $Y$, we project it into the subspace spanned by the columns of $X$; the dimension of this subspace is four. (The estimate of $\beta$ is usually determined during this process.) Having this projection $Y_M$, $Y_M - Y$ is an orthogonal projection into a three-dimensional complement of $M$. Thus we can estimate $\sigma^2$ with $s^2$:

$$s^2 = \frac{\|Y - Y_M\|^2}{3}.$$
Table 1. The table of analysis of variance interpreted geometrically ($m$, $s$ denote the dimensions of model and submodel; $n$ stands for the dimension of the whole space, i.e., the number of components of $Y$).

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>Degrees of freedom</th>
<th>Mean sum of squares</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>$|Y_M - Y_s|^2$</td>
<td>$m - s$</td>
<td>$|Y_M - Y_s|^2 / (m - s)$</td>
<td>$F_{M, S}$</td>
</tr>
<tr>
<td>Error</td>
<td>$|Y - Y_M|^2$</td>
<td>$n - m$</td>
<td>$|Y - Y_M|^2 / (n - m)$</td>
<td>$F_{M, S}$</td>
</tr>
<tr>
<td>Total$^3$</td>
<td>$|Y - Y_s|^2$</td>
<td>$n - s$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. A scheme of geometrical representation of the model, its submodel and vectors involved.

Further, we can consider the possibility that some of the three independent variables do not have any influence on the expected value of $Y$. This idea can be formulated as the null hypothesis that the relevant $\beta_i$ are equal to zero. That means that we might leave out the relevant columns of $X$ and replace the original model with a simpler one – its submodel. Typically we omit one column; than the dimension of the subspace given by the submodel is three and we can verify the null hypothesis by means of this statistics:

$$\frac{\|Y_M - Y_s\|^2/1}{\|Y - Y_M\|^2/1} \sim F_{1, 3},$$

because the dimension of the orthogonal complement of $S$ (three-dimensional) inside $M$ (four-dimensional) is one.$^4$

In another usual case we suggest omitting all columns representing the variables on which $Y$ depends. Then there is only the first column of $X$ left, representing the basic level of components of $Y$. This time the null hypothesis can be verified using statistics

$$\frac{\|Y_M - Y_s\|^2/3}{\|Y - Y_M\|^2/3} \sim F_{3, 3},$$

because the dimension of the orthogonal complement of $S$ (one-dimensional) inside $M$ (four-dimensional) is three.$^5$

$^3$ Notice the Pythagorean theorem in the picture below.

$^4$ A standard method of verifying this null hypothesis is based on another statistics, which follows the $t$ distribution. We do not deal with the $t$ statistics here as their geometrical interpretation is not so much illustrative. Anyway, these two methods are equivalent (at least in the case of two-tailed alternative hypothesis), because if $X \sim t_n$, then $X^2 \sim F_{1, n}$. 

$^5$
A brief historical survey

The history of the geometrical view of statistics is closely related to the history of mathematical statistics itself. There are many reasons to think that sir Ronald Aylmer Fisher (1890 – 1962), a British biologist and brilliant mathematician who is respected as the founder of the mathematical statistics, could reach most of his fundamental results just by the means of his strong geometrical sense. His daughter claims in [Box, 1978] that he had to develop it as a child when he suffered from his poor eyesight and that it influenced his manner of attacking mathematical problems throughout his life.

The story began when another great person, William Sealy Gosset (1876 – 1937), a British chemist and mathematician as well, published his famous article ["Student", 1908] dealing with the distribution of 

\[ z = \frac{\bar{x}}{s}, \]

i.e., the sample average of zero expected value divided by the sample standard deviation. This problem had been fully ignored so far and the article published under a pseudonym “Student” did not initially attract much attention – except for Fisher. He grasped the problem in his own way and during the correspondence with Gosset, he conceived the notion of the geometrical representation of the configuration of the sample in \( n \)-dimensional space. This representation immediately gave him the concept of degrees of freedom, yielded the mathematical proof of “Student’s” distribution and finally led to wider applications [Box, 1981].

Unfortunately, there is not much geometry in Fisher’s works and he uses it explicitly very rarely (see [Fisher, 1915] or [Fisher, 1923] for examples). Most of evidence of his geometrical thinking comes from his correspondence with Gosset and from his daughter’s works ([Box, 1978, 1981]). The reason for that is probably that he did not succeed in explaining his thoughts by the means of geometry because not everyone could follow it. That was why he expressed himself in algebraic terms when publishing came to pass. Consequently, the algebraic approach became standard instead of the original one.

In spite of that, there are still some authors who devote themselves to this attitude. Saville & Wood [1991] express an aim “to retrieve Fisher’s lost insight”. Box [1978] keeps calling attention to the advantages of the geometrical approach “by which the results can be immediately seen rather than laboriously derived”. Wichura [2006] writes that the mathematical theory underlying the analysis of variance and regression became clear to him after he had read a draft of William Kruskal’s excellent (but unpublished) treatise on the geometric approach to these subjects. In [Herr, 1988], four more authors publishing throughout the twentieth century are mentioned, who deal with the linear model from more or less geometrical point of view. A question why the geometrical approach is unpopular at present is also analysed here. The author finds out that the reasons for this are above all of historical nature.

There are almost no mentions about this possibility in the Czech literature except of [Zvára, 2008] and [Pázman, 1988], but these sources are aimed at engaged readers and may appear inaccessible for those who are unfamiliar with the topic.

Conclusion

We presented a rough demonstration of how geometry can be used in understanding some of the most important statistical concepts together with a fleeting glance on the history of this attitude. Although the geometric approach to the linear model still remains more or less unpopular, we are strictly convinced that it deserves an extraordinary attention because of its didactical potential and

---

5 We can similarly verify any null hypothesis which is represented by a system of linear conditions concerning the values of \( \beta \) (for example \( \beta_1 = 5 \) or \( 2\beta = \beta_1 \)). The idea of the test remains the same, as these conditions (in the case of independence of the columns of \( \mathbf{X} \)) always represent limiting the original subspace \( \mathbf{M} \) of possible expected values of \( \mathbf{Y} \) to its inner subspace \( \mathbf{S} \).

6 Another possible reason that easy computing formulae were needed in these days is suggested in [Saville & Wood, 1991].
close relation to historical roots of mathematical statistics. A deeper analysis of this topic placing emphasis on solid theoretical and historical background will follow as soon as possible.

References

“Student”: The Probable Error of a Mean. Biometrika 6, 1–25, 1908.