ON LINEAR REGRESSION IN
THREE-DIMENSIONAL EUCLIDEAN SPACE.

O. V. Ageev, R. A. Sharipov

Abstract. The three-dimensional linear regression problem is a problem of finding a spatial straight line best fitting a group of points in three-dimensional Euclidean space. This problem is considered in the present paper and a solution to it is given in a coordinate-free form.

1. Introduction.

The linear regression problem in two-dimensional case (i.e. on a plane) typically arises when approximating experimental data with a linear function (see [1]). Its solution using least squares method was first published by Legendre in 1805 (see [2]). In an unpublished form the least squares method is attributed to Carl Friedrich Gauss 1795. His work was published only in 1809 (see [3]).

There are various fitting problems in three-dimensional Euclidean space (see plane, circle and ellipse fitting problems in [4] and [5], see ellipsoid fitting problem in [6] and [7]). The linear regression problem in three-dimensional case is the problem of best fitting some straight line to a group of points in three-dimensional Euclidean space. A solution of this problem is given by Jean Jacquelin in [8]. His method is substantially based on direct calculations using coordinates. Our goal in the present paper is to give a coordinate-free solution to the problem.

2. Parametric and non-parametric vectorial equations of a straight line.

Let’s consider the straight line $AX$ in Fig 2.1. The point $A$ is a fixed point of this line, its radius-vector is $r_0$. The point $X$ is a variable point, its radius-vector is $r$. These two radius-vectors are related to each other by means of the equation

$$r = r_0 + a t,$$

where $a$ is some non-zero vector on the line and $t$ is a scalar parameter. The equality (2.1) is called the vectorial parametric equation of the line in the space (see [9]).

The choice of the point $A$ on the line is not unique. Therefore the equation (2.1) has some extent of ambiguity. In order to avoid this ambiguity non-parametric
equations are used. Let’s multiply both sides of the equality (2.1) by the vector \( \mathbf{a} \) using the vector product\(^1\) operation. As a result we get

\[
[r, a] = [r_0, a]. \tag{2.2}
\]

The product of two constant vectors in the right hand side of (2.2) is a constant vector. If we denote it through \( \mathbf{b} \), we get the equality

\[
[r, a] = \mathbf{b}, \quad \text{where} \quad \mathbf{b} \perp \mathbf{a}. \tag{2.3}
\]

The equality (2.3) is known as the non-parametric vectorial equation of the line in the space (see [9]). Note that the vector \( \mathbf{b} = [r_0, a] \) has no ambiguity arising from the uncertainty in choosing the initial point \( A \) on the line. Indeed, it is easy to see that \( \mathbf{b} \) is invariant with respect to the transformation \( r_0 \to r_0 + \mathbf{a} t \).

3. The statement of the problem.

Let \( X_1, \ldots, X_n \) be a group of points in the space given by their radius-vectors \( r_1, \ldots, r_n \). The linear regression problem consists in finding a line given by the equation (2.2) such that the root mean square of the distances \( d_1, \ldots, d_n \) from the points \( X_1, \ldots, X_n \) to the line (2.2) takes its minimal value:

\[
d^2 = \frac{1}{n} \sum_{i=1}^{n} d_i^2. \tag{3.1}
\]

4. The solution of the problem.

The distance from the point \( X_i \) to the line (2.1) is given by the formula

\[
d_i = \frac{|[r_i - r_0, a]|}{|a|}. \tag{4.1}
\]

Without loss of generality we can assume that

\[
|a| = 1. \tag{4.2}
\]

Then, taking into account \( \mathbf{b} = [r_0, a] \) and (4.2), from (4.1) we derive

\[
d_i = |[r_i, a] - \mathbf{b}|. \tag{4.3}
\]

Now we substitute (4.3) into (3.1). As a result we obtain

\[
d^2 = |\mathbf{b}|^2 - \frac{2}{n} \sum_{i=1}^{n} ([r_i, a], \mathbf{b}) + \frac{1}{n} \sum_{i=1}^{n} |[r_i, a]|^2. \tag{4.4}
\]

The formula (4.4) is an analog of the formula (2.3) in [4]. The round brackets in (4.4) denote the scalar product\(^2\) operation.

**Definition 4.1.** A line given by the equation (2.3) with \( |a| = 1 \) is called an optimal root mean square line if the quantity (4.4) takes its minimal value.

---

1 It is also called the cross product, i.e. \([x, y] = x \times y\).

2 It is also called the dot product, i.e. \((x, y) = x \cdot y\).
The right hand side of (4.4) is a quadratic polynomial with respect to the components of the vector \( \mathbf{b} \). It takes its minimal value if \( \mathbf{b} \) is given by the formula

\[
\mathbf{b} = \frac{1}{n} \sum_{i=1}^{n} [r_i, \mathbf{a}]. 
\]

Substituting (4.5) back into (4.3), we derive

\[
\bar{d}^2 = \frac{1}{n} \sum_{i=1}^{n} |[r_i, \mathbf{a}]|^2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} ([r_i, \mathbf{a}], [r_j, \mathbf{a}]).
\]

The formula (4.6) is an analog of the formula (2.5) in [4]. Its right hand side is a quadratic form with respect to the vector \( \mathbf{a} \). We denote it through \( Q(\mathbf{a}, \mathbf{a}) \):

\[
Q(\mathbf{a}, \mathbf{a}) = \frac{1}{n} \sum_{i=1}^{n} |[r_i, \mathbf{a}]|^2 - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} ([r_i, \mathbf{a}], [r_j, \mathbf{a}])
\]

and call the non-linearity form for a group of points in three-dimensional Euclidean space. Like the non-flatness form (2.14) in [4], the non-linearity form (4.7) is positive in the sense of the following inequality:

\[
Q(\mathbf{a}, \mathbf{a}) \geq 0 \quad \text{for } \mathbf{a} \neq 0.
\]

Like in [4] one can draw some analogy to mechanics using the inertia tensor. However, we shall not do it now. We just note that like any quadratic form \( Q(\mathbf{a}, \mathbf{a}) \) diagonalizes in some orthonormal basis associated with its primary axes.

Let’s introduce the following notation analogous to (2.6) in [4]:

\[
r_{\text{cm}} = \frac{1}{n} \sum_{i=1}^{n} r_i.
\]

The vector \( r_{\text{cm}} \) in (4.8) is the radius-vector of the center of mass of a group of points \( X_1, \ldots, X_n \) if assume that unit masses are placed at each of these points. In terms of (4.8) the formula (4.5) is written as

\[
\mathbf{b} = [r_{\text{cm}}, \mathbf{a}].
\]

Comparing (4.9) with \( \mathbf{b} = [r_0, \mathbf{a}] \), we conclude that the optimal line should pass through the center of mass of a group of points. Its direction is determined by the non-linearity form \( Q(\mathbf{a}, \mathbf{a}) \) according to the following theorem.

**Theorem 4.1.** A line is an optimal root mean square line for a group of points if and only if it passes through the center of mass of these points and if its direction vector \( \mathbf{a} \) is directed along the primary axis of the non-linearity form \( Q \) of these points corresponding to its minimal eigenvalue.
5. Conclusion.

Theorem 4.1 solves the linear regression problem formulated in Section 3. Its proof is obvious from the consideration preceding it. Practically this theorem means that in order to find a line best fitting a group of points in three-dimensional Euclidean space one should find their center of mass and diagonalize the symmetric matrix associated with their non-linearity form (4.7). In some cases this matrix can have two minimal eigenvalues $\lambda_1 = \lambda_2 < \lambda_3$. In these cases the shape of the group of points resembles a disc and hence there is no preferable direction for the optimal line within the plane of this disc.

If $\lambda_1 = \lambda_2 = \lambda_3$, the shape of the group of points resembles a ball. In this case we have no preferable direction for the optimal line at all.

References

1. *Linear least squares*, Wikipedia, Wikimedia Foundation Inc., San Francisco, USA.
2. Legendre A. M., *Nouvelles méthodes pour la détermination des orbites des comètes*, F. Didot, Paris, 1805.
3. Gauss C. F., *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, Perthes & Besser, Hamburg, 1809.
4. Sharipov R. A., *Algorithms for laying points optimally on a plane and a circle*, e-print arXiv:0705.0350.
5. Gander W., Golub G. H., Strebel R., Least-squares fitting of circles and ellipses, BIT Numerical Mathematics 34 (1994), no. 4, 558–578.
6. Reza A., Sengupta A. S., *Least square ellipsoid fitting using iterative orthogonal transformations*, e-print arXiv:1704.04877.
7. Anwar R., Hamilton M., Nadolsky P. M., *Fast ellipsoidal fitting of discrete multidimensional data*, e-print arXiv:1901.05511.
8. Jacquelin J., *Regressions et trajectoires en 3D*, Online resource doc/31477970 at scribd.com, 2002, 2011.
9. Sharipov R. A, *Course of analytical geometry*, Bashkir State University, Ufa, 2010; see also arXiv:1111.6521.

Self-employed individual, Ufa, Russia

E-mail address: ageev-ufa@yandex.ru

Bashkir State University, 32 Zaki Validi street, 450074 Ufa, Russia

E-mail address: r-sharipov@mail.ru

URL: http://ruslan-sharipov.ucoz.com
http://freetextbooks.narod.ru