Enhancing Students' Understanding on the Method of Least Squares: an Interpretative Model Inspired By Historical and Epistemological Considerations

Michael KOURKOULOS, Constantinos TZANAKIS

Department of Education, University of Crete, 74100 Rethymnon, Greece

mkourk@edc.uoc.gr, tzanakis@edc.uoc.gr

Abstract

A didactical and epistemological analysis permits to identify models of situations, which can be used in teaching to enhance students' understanding of basic statistical methods and aggregates that involve sums of squared distances from a center (central point or central line, e.g the Method of Least Squares (MLS), variance, the Pearson coefficient). Here such a basic model, the model of springs in two dimensions, is analyzed with respect to its didactical virtues to facilitate the initial understanding of the MLS, taking into account elements from relevant individual interviews realised with prospective schoolteachers.

1 Introduction

Didactical point out that students encounter important difficulties to understand variation and its parameters concerning univariate distributions (e.g. Mevarech 1983, Shaughnessy 1992, Batanero et al. 1994, Watson et al. 2003, Reading & Shaughnessy 2004, delMas & Liu 2005) and that they encounter even more important ones in the case of bivariate distributions (e.g. Ross & Cousins1993, Batanero et al. 1996, Cobb et al. 2003, Moritz 2004, Scariano & Calzada 2004).

Our previous research concerning variation points out that: an important factor for the efficiency of introductory teaching approaches concerning the understanding of basic statistical methods and aggregates that involve sums of squared distances from a central point or a central line (e.g variance, Method of Least Squares (MLS), Pearson's coefficient) is the adequacy of the used body of situations' examples (Kourkoulos & Tzanakis 2003a, b, 2006a). The non-purely mathematical examples of situations employed in usual introductory statistics' courses are very often mainly (or almost exclusively) examples related to social phenomena (students' notes, peoples' weights, incoming etc), whereas, meaningful examples of situations from other domains, like physics, or geometry are absent.

The meaning of the aforementioned aggregates and methods is difficult to understand in the context of examples related to social phenomena, because: (i) in these cases the aggregates represent only data tendencies (often having a coherent meaning only at the purely numerical level); (ii) the sums of squares involved in the aggregates are quantities that have an unclear meaning in that context (squares of students' height, squares of distances of buses trips etc), or, even worse, they are dimensionally meaningless (squares of notes, weights, money etc), Kourkoulos & Tzanakis 2006a.

Restricting the body of examples used in introductory courses to this type of situations, is virtually a strong cause of important epistemological obstacles against students' understanding[1]. Moreover, the absence of adequate situations' examples, in which the aggregates have a clear meaning, deprive students of important interpretative elements that are essential to facilitate their comprehension (Tzanakis & Kourkoulos 2004).

2 Relevant historical elements

The MLS was conceived by Legendre at 1805 in connection with data treatment in problems of astronomy and geodesy. The method rapidly became the most important method of data treatment in astronomy and geodesy in the 19^{th} century, (Stigler 1986, ch. 1, Porter 1986, pp. 93–100). However, adequately transferring MLS, as well as other methods and tools developed for data treatment in these two fields, to the data treatment of social sciences demanded a laborious evolution for almost a century, and overcoming important conceptual barriers (Porter 1986, pp. 307–314). The conceptual framework of linear regression that Galton established working on heredity (from 1874 to 1889),[2] opened the way to the works of Edgeworth, Pearson and Yule, who elaborated adequate conceptual frameworks and the first efficient tools for statistical elaboration on problems of social sciences. It is characteristic of the importance of the conceptual difficulties encountered, that it is only as late as 1897, that on the basis of theoretical arguments, Yule proposed a generalised method of linear regression for problems in social sciences based on the use of least squares. (Stigler 1986, part 3, Porter 1986, pp. 286–296).

A main reason for these difficulties is the complexity of social phenomena, in which a very large number of factors interfere. In comparison, the phenomena examined in astronomy and geodesy were much simpler. A consequence of this complexity was that there were no theories of social phenomena that could incorporate coherently and efficiently all (or most of) the influencing factors. In contrast in astronomy and geodesy there was a solid theoretical background, Newtonian mechanics and its extensions, permitting to efficiently modelise and interpret the examined phenomena. This has several consequences: it provided meaning to the used statistical objects and methods, inspired and oriented their development and permitted to interpret their results. Furthermore, it provided reliable a priori expectations, a critical element for assessing the elaborated statistical methods.[3] On the contrary, in the treatment of social data, statistical objects were (and still are), in most of cases, only data tendencies, with a meaning much more difficult to construct.[4] Moreover, the absence of reliable a priori expectations made difficult to assess the statistical methods used, and of course the two previous aspects interacted increasing further the encountered difficulties. (Stigler, 1986, pp. 358–361).

[1] These obstacles are widely activated if the introductory course requires that students examine carefully and coherently the meaning of the newly introduced parameter (Kourkoulos & Tzanakis 2003a,b, 2006a).

[2] However, it is interesting to notice that Galton realised linear regression without using the MLS; in most of cases he found his regression coefficients by rough calculations based upon graphs. (Stigler 1986, ch 8)

[3] In this intellectual environment is not surprising that Legendre when initially presenting MLS (Legendre's appendix of 1805, pp.72-75; see Stigler 1986 pp. 11–15, 58) explained the meaning of the method and of the solution found by reference to equilibrium (directly, p. 73 and through an analogy to the center of gravity, p. 75). More precisely, in p. 73 he wrote for the MLS "Par ce moyen il s'établit entre les erreurs une sorte d'équilibre qui empêchant les extrêmes de prévaloir, est très-propre à faire connoître l'état du system le plus proche de la vérité.". The interpretative model that we analyze in section 3 could be considered as an operational realization of these Legendre's reference to equilibrium.

[4] In contrast to that, the aggregates of central tendency and of variation, in astronomy and geodesy, had the status of approximations to measures of "real objects" of central importance for the examined situation (e.g. a regression line can be an approximation to the trajectory of a celestial body, and square residuals can be a measure of the inaccuracy of observations). E.g. Consider the difficulty on understanding the meaning of a regression line of students' notes in mathematics and literature compare to that where the regression line is the approximation to the trajectory of a projectile or of a celestial body.

Conventional introductory statistics' courses do not take under consideration this imposing historical reality, and this omission allows for the existence of the important defect underlined in (1), concerning the characteristics of the set of the situations' examples used in these courses.

3 Physical models

3.1

Studying (i) students' difficulties to understand the discussed aggregates and methods (Kourkoulos & Tzanakis 2003a,b, 2006a), (ii) the historical development of these concepts in statistics (Stigler 1986, 1999, Porter 1986, Kourkoulos & Tzanakis, 2006a), and (iii) realizing a didactically oriented epistemological study of fundamental physical phenomena that are related to basic statistical concepts (Tzanakis & Kourkoulos, 2004), allows as to identify elementary physical situations that involve quantities conceptually close to the sums of squared distances from a center (central point or central line).

Further analysis led us to elaborate for didactical purposes two interpretative models (a model of moving particles and a model of springs)[5] for the variance. The models were used in two experimental courses on introductory statistics. Students' behavior was encouraging concerning the models' didactical potential to facilitate the understanding of variance and its properties. (Kourkoulos et al 2006b).

Here we present and comment didactically on an extension of the springs' models in two dimensions, elaborated to facilitate at an introductory level, the understanding of the MLS, the Least Squares Straight Line (LSSL) and its associated quantities (Pearson's coefficient, square residuals, ...). The presentation and the comments are enriched with results of the analysis of individuals interviews realized with 15 students.[6,7] Given their small number, these interviews constitute only a first tentative approach for the empirical investigation of the model. However, students' behavior and reactions appears often to be very insightful for further exploring the model didactically. The presentation is done, according to the elements and the order of presentation given to the students, albeit concisely, because of space limitations.

Initially the general problem of linear regression and the use of MLS were presented briefly to the students with data examples from everyday life situations and from constructions' and measurements' errors situations. Students posed interesting questions such as:

- Why to use squared distances and not "simple" (1^{st} degree) distances? What do these squared distances mean?
- Why to search a straight line and not another line of best fit? How to decide whether there is some straight line that fits well the data indeed?[8]

As we will see, to answer these types of questions, the use of the model can offer significant clarifications and insights.

[5] Though simple, these models are rooted in deep physical models that historically have been used as models of (i) thermal radiation, (ii) ideal gases, and (iii) a solid body (if one thinks of springs as microscopic oscillators) (*ibid* 2004)

[6] In the individual interviews one of the researchers presented the model to each one of the students and discussed the subject with him. The interviews lasted 4 to 6 hours (2 to 3 meetings) following students' background and questions. The discussion with them was registered and their written productions were collected.

[7] All students were volunteer students of the Department of Education; also the year before, they had followed one of the experimental courses mentioned in the previous paragraph.

[8] These questions were incited by (graph representation of) data examples, which seemed to fit better to other types of line or to be too scattered.

3.2 Initial state of the model

Consider that on a horizontal plane (e.g. a table) we have a set of fixed points, (figure 1)[9], and an attachment bar placed on Ox. Keeping the bar immovable, we attach springs to the points and to the bar, so that the springs' direction is parallel to Oy.

We consider the springs as ideal, obeying Hook's low, so: the force exerted to the bar by the spring attached to the point (x_i, y_i) is $F_i = ky_i$ and its potential energy is $E_i = \frac{1}{2}\, ky_i^2$.[10] Here, for simplicity we consider that the spring's constant is the same for all, $k = 1$ Nt/cm.[11] Therefore, the total initial potential energy of the system is

$$E_{\text{initial}} = E_1 + E_2 + \ldots + E_n = \frac{1}{2}ky_1^2 + \frac{1}{2}ky_2^2 + \ldots + \frac{1}{2}ky_n^2 \tag{1}$$

(n been the number of points).

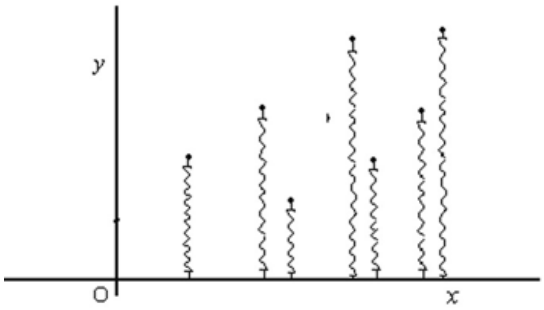


Figure 1

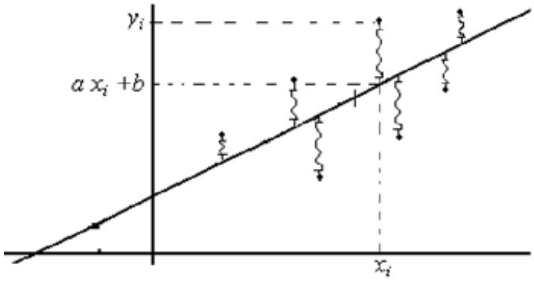


Figure 2

3.3 Leaving the bar free

3.3.1

We consider that when we leave the bar free, the end of each spring attached to the bar can move only parallely to Oy (the other edge remains fixed).[12] We suppose also that when the bar and the springs move there is (small but non-negligible) friction.

Once liberated, the bar is attracted by the springs towards the set of points and because of friction, it finally stops somewhere between the points, after oscillating some time around its final equilibrium position, until totally loosing, because of frictions, its kinetic energy (figure 2). All students found very natural that the bar finally stops at some position and that this position is somewhere between the points. [13]

Then the researcher told the students that the bar at rest is on some straight line (e_{final}) of the form $y = ax + b$ and asked them to "calculate" (express) the force exerted on the bar by the spring attached to the point (x_i, y_i) and its potential energy. Twelve students

[9] Such illustrations were given to the students on paper and they could work on them.

[10] The researcher reminded to the students the two properties of a spring obeying Hook's law, but students had no particular difficulties on this subject, since, they had been taught Hook's law in high school and in compulsory physics course at the University. Moreover, in their recent course of introductory statistics, they had already used models with such springs (see note 7, Kourkoulos end al 2006b).

[11] However, if we consider the springs' constants as different then they can represent frequencies associate to the attachment points.

[12] To the two students who asked, we gave examples of different technical realizations permitting this motion of the springs when they are connected to the bar.

[13] Alternatively we could consider that: there are no frictions but we apply adequate external resistance to the bar (e.g. we hold it adequately) so that it follows smoothly the attraction of the springs until it attains an equilibration position. In that case the liberated dynamic energy will be consumed by the external resistance.

achieved to apply Hook's law without help of the researcher (but having figure 2 at their disposal) and found: $F_i' = k(y_i - (ax_i + b))$, $E_i' = \frac{1}{2}k(y_i - (ax_i + b))^2$.

The remaining three, obtained the same result after the researcher helped him to express the length of the spring, $y_i - (ax_i + b)$ (their main difficulty was to decode the graphical representation).

After that, the researcher asked them to express the total potential energy of the system when the bar is at rest. Thirteen of them succeeded to do so without help and gave answers of the type:

$$\begin{aligned} E_{\text{fin}} &= \frac{1}{2}k(y_1 - (ax_1 + b))^2 + \frac{1}{2}k(y_2 - (ax_2 + b))^2 + \ldots + \frac{1}{2}k(y_n - (ax_n + b))^2, \\ E_{\text{fin}} &= \frac{1}{2}k[(y_1 - (ax_1 + b))^2 + (y_2 - (ax_2 + b))^2 + \ldots + (y_n - (ax_n + b))^2] \end{aligned} \tag{2}$$

To the two others the subject was explained by the researcher.

Then the researcher remarked that during the motion of the bar from its initial to its final position there was loss of energy because of friction. This energy was "taken" from the energy stored in the springs, since it was the only energy existing in the system and there was no external energy supply. Therefore $E_{\text{initial}} > E_{\text{fin}}$. All students easily accepted this assertion as correct and no objections or difficulties to understand it appeared.

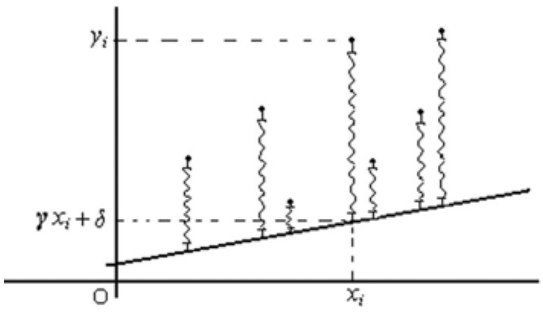


Figure 3

3.3.2

After that, the researcher asked students to consider what will happen to the bar and the total energy of the springs if we hold it fixed on another straight line ($y = \gamma x + \delta$), figure 3, and then we liberate it. He also told them that: the equilibrium position seen previously (see §3.3.1) we will prove later on that it is the only equilibrium position of the bar.[14] All students considered that obviously the bar will move and finally will rest at the unique equilibrium position and succeeded to write the potential energy of the system at the new position.

$$E = \frac{1}{2}k[(y_1 - (\gamma x_1 + \delta))^2 + (y_2 - (\gamma x_2 + \delta))^2 + \ldots + (y_n - (\gamma x_n + \delta))^2] \tag{3}$$

Furthermore, all but one, answered easily that in this case $E > E_{\text{fin}}$ as well (because there are frictions during the movement and so there is loss of energy).

Then the researcher remarked that, for the same reasons all the positions ($\neq e_{\text{final}}$) have a corresponding potential energy that is greater than the potential energy of the equilibrium position.

[14] The researcher anticipated the result of a proof that followed (see page 8) in order to avoid considerations such as: what will happen if there are more than one equilibrium positions? What will happen if there is a whole domain of such positions? And so on, given that they don't concern our model

The purpose of the previous discussion was to present a conceptually simple explanation[15] that students could understand on the relation between the position of minimal dynamic energy and the equilibrium position of the system, given that they were not taught the corresponding general principle in physics. In this respect, as we have described, their reaction was encouraging.

Then the researcher remarked that this consideration is in agreement with a principle of Physics saying that the positions of minimal potential energy of a system are equilibrium positions of the system and that in case there is only one position of minimum potential energy this is the only position of stable static equilibrium of the system.

Remarks (1): As we have seen, already when we introduced the model (§§2.1, 2.2) basic quantities related to the LSSL have a clear interpretation:

- The sum of points' squared deviations from any straight line corresponds to the potential energy of the system (total potential energy of the springs) when the attachment bar is on this line (eq (3)). (Thus, square residuals obtain also a clear meaning; they correspond to the minimum potential energy of the system.)

- The LSSL is interpreted in two ways: (a) the position of the attachment bar for which the system has its minimal potential energy, (b) the equilibrium position of the bar.

  That LSSL is the equilibrium position is one of its principal characteristics; however, it is a characteristic difficult to be seen in the usual purely mathematical elaboration (here equilibrium is static in the sense that the bar does not move when it is at the equilibrium position; this aspect cannot appear and be understood within the usual mathematical elaboration since movement is absent there).[16] Therefore the model is particularly useful for understanding this characteristic.

- The characteristics (a) and (b) are connected in a clear way with a simple argumentation. The simplicity and clarity of this argumentation is due to the characteristics of the model.

  (Moreover because of this connection students obtained some interesting introductive insights on the corresponding general physical principle.)

4 Approaches for finding the LSSL

(A) Typical approach in statistics

Initially, the researcher reminded to the students how to differentiate 2nd degree polynomials and to use them to find the extremum of such functions (since 10 students claimed "not to remember anything" on this from high school).

Then he tried to explain the concept of partial differentiation in this case. Students have not been taught previously partial differentiation and 8 of them had important difficulties on understand it.

Finally he presented the typical approach in statistics' courses for finding the LSSL, by partial differentiation of the sum of squared deviations, $\frac{2}{k}E = (y_1-(\gamma x_1+\delta))^2+(y_2-(\gamma x_2+\delta))^2+\ldots+(y_n-(\gamma x_n+\delta))^2$, with respect to γ, δ.

All students understood the new elements that the solution found added to the meaning of LSSL already presented in §(3.3): it passes through the point $(\overline{y}, \overline{x})$[17] and its inclination

[15] Even though somewhat simplified

[16] See also footnote 3.

[17] The researcher remarked to the students that this point is a center of the set of points, also called their mathematical center of gravity.

relative to Ox is:

$$a = \frac{\frac{\sum_{i=1}^{n} y_i x_i}{n - \overline{y}\,\overline{x}}}{\sigma_x^2} \tag{4}$$

All students were able to apply the two conditions and find the LSSL in specific examples.

Moreover, the researcher remarked that: the solution process constitutes also a proof that LSSL is unique for a set of points, (if $\sigma_x^2 \neq 0$). In the context of the model, this means that LSSL is the only position of minimum potential energy of the system and, following the corresponding physical principal, it is the only position of stable equilibrium of the bar.

Nine students faced important difficulties to understand the solution process, mainly because of the concept of partial differentiation.

Only six students presented evidences[18] that they have satisfactorily understood the solution process.

(B) AN ALTERNATIVE WAY INDUCED BY THE SPRINGS' MODEL

The researcher presented the subject in the following manner:

By an equilibrium position of the bar we mean that, if originally we hold it fixed there, it remains at rest even if we liberate it afterwards. For this to happen, it must neither be displaced nor be rotating. Hence, it must satisfy two equilibrium conditions: (i) the total force exerted on it must be 0; (ii) the total moment around some point A of the plane must be 0 (figure 4).[19]

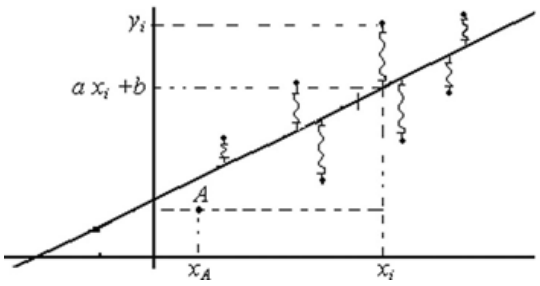


Figure 4

Most of the students easily accepted and understood the two equilibrium conditions:

Students had been taught the 1st condition in their physics courses as a condition holding for a solid body at rest (but also it appears to them as intuitively clear). They also had been taught that if a plane solid body is attached to a point A of its plane[20] then: if it stays at rest the total moment of the forces exerted on the body around A is zero. The researcher reminded them this property (focusing to the case of a bar). After that reminding, only three students claimed not to understand the property.

Then the researcher explained that if the body, here the bar, is not attached to A and remains at rest, then we can attach it to A without disturbing its equilibrium (and without exerting any additional force on it). Thus we can apply the previous property and obtain that the total moment around A of the forces exerted on the body is zero. Obviously, this held also when the body was not attached because no additional force was exerted on it

[18] They were able to reproduce the general solution process (with others letters instead of γ and δ) with only minor corrections and instructions from the researcher.

[19] Condition (i) and (ii) together, are also sufficient conditions of static equilibrium. However, since students knew that an equilibrium position of the bar exists (§3.3.1), discussing this aspect was not necessary for the treatment of the problem and to keep the discussion shorter we had not discussed it. Nevertheless, it is interesting to consider it with the students in a further didactical investigation of the subject.

[20] So that it can only turn around the point, in the plane.

because of the attachment. Therefore, this leads to condition (ii). Only two students, among the three above, found difficult to understand these explanations.[21]

Given the work done in §§3.2, 3.3, students had no significant difficulties to express the 1st condition:

$$F_{\text{total}} = F_1+F_2+\ldots+F_n = k(y_1-(ax_1+b))+k(y_2-(ax_2+b))+\ldots+k(y_n-(ax_n+b)) = 0 \quad (5)$$

For the 2nd condition, five students initially needed help to express the moment of a spring around A: $M_{iA} = k(y_i - (ax_i + b))(x_i - x_A)$, but managed to do so by themselves for the others springs. Ten students managed by themselves to express algebraically the 2nd condition:

$$\begin{aligned} M_{\text{total}\,A} = \quad & k(y_1 - (ax_1 + b))(x_1 - x_A) + k(y_2 - (ax_2 + b))(x_2 - x_A)+ \\ & \ldots + k(y_n - (ax_n + b))(x_n - x_A) = 0 \end{aligned} \quad (6)$$

All students understood without serious difficulties, the necessary algebraic transformations presented by the researcher to find the solution: a **unique**[22] equilibrium straight line that satisfies the same conditions (also expressed in the same form) as the LSSL found previously, in §4(A).

Moreover, the researcher showed them that with somewhat different transformations of (5) and (6) we obtain the inclination a in a different form:

$$a = \frac{\frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{n}}{\sigma_x^2} \quad (7)$$

Then, the researcher remarked to the students that since the equilibrium straight line is unique, as explained previously (see §3.3.2) it is also the position of minimum potential energy of the system of springs and thus the LSSL of the set of points.

Remarks (2): Comparison of the solution processes A & B

- The process B is mathematically easier than A, since, it doesn't involve partial derivations, or some other rather complicated mathematical procedure to minimize the sum of squared deviations (SSD).[23] Moreover the two equations obtained are of first-degree in the unknowns. However for understanding process B, it is necessary that students have some rudimentary knowledge of elementary physics.

- Process A focuses on minimizing the SSD and, thus, in the context of the model, on minimizing the potential energy of the system. Process B focuses on the characteristic of LSSL as an equilibrium position, and, allows to clarify further this characteristic (in addition to the immobility aspect, see Remark 1): It clarifies with respect to which quantities LSSL is an equilibrium position (what quantities equilibrate at this position): the springs' forces (and equivalently the deviations from the LSSL) and the momentum exerted to the bar (so that the bar don't turn).[24]

[21] If someone work with students knowing more physics than ours, these explanations will be unnecessary, since the two conditions are typical conditions of static equilibrium.

[22] The researcher also remarked to the students: that since conditions (i) and (ii) are necessary equilibrium conditions, the solution process is also a proof that there is at most one equilibrium straight line (when $\sigma_x^2 \neq 0$); given that there is some equilibrium straight line (see §3.31.), we are sure that there is one and only one equilibrium straight line.

[23] For such procedures that do not use partial differentiations see Darlington 1969, Stanley & Glass 1969, Gordon & Gordon 2004, Scariano & Calzada 2004.

[24] The 2nd condition is difficult to be explained as an equilibrium condition within a purely algebraic and/or geometrical elaboration.

- Process B cannot be extended beyond 3 dimensions in an elementary way, since the model cannot; process A has not this important restriction.
- Concerning introductory statistics, it is interesting to present to students both processes since they enlighten different aspects of the subject. Moreover the understanding of one process can interfere constructively with the understanding of the other.

Subsequently, the researcher considered with the students some important quantities related to LSSL and their interpretation in the context of the model.

Because of space limitations, we report briefly on this point.

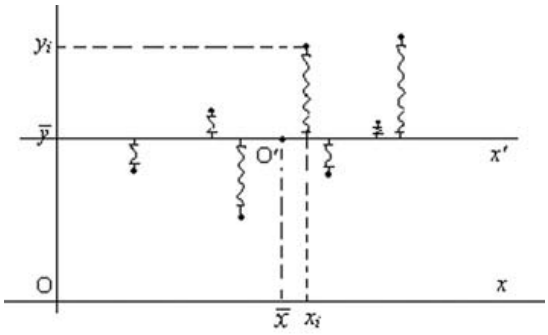


Figure 5

- The sum $\dfrac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{n}$ that appears in (7), is the **covariance** of the statistical variables X, Y.

When the bar passes from the point O, with coordinates $(\overline{y},\overline{x})$, and it is parallel to Ox (position Ox of the bar in figure 5) its total moment around $(\overline{y},\overline{x})$ is: $k\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})$, so by dividing with n we obtain **the average moment per spring** around O. Thus we have a clear interpretation of the covariance as proportional to this quantity. As our students had not been taught the covariance previously, this interpretation was used for introducing this concept. The subject was only touched upon and, given its importance, it merits a systematic didactical study. However it is interesting to remark that once the model is established, it leads naturally to the introduction of covariance, which appears as an important and conceptually clear quantity in this context.

Pearson's correlation coefficient

Consider a parallel displacement of the initial coordinate system $O(x,y)$ to $O'(x',y')$ with the origin at the centre of gravity $(\overline{x},\overline{y})$: $x'_i=x_i-\overline{x}, y'_i=y_i-\overline{y}$.

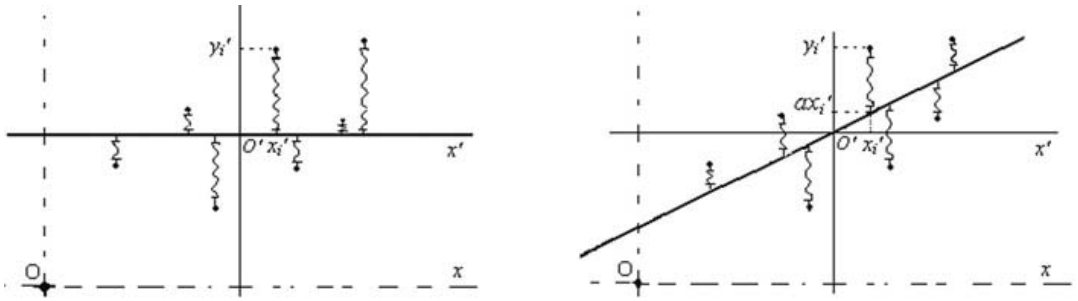
Consider that the initial position of the bar is Ox (figure 6a). The total energy of the system is:

$$E_{\text{initial}}=\frac{1}{2}k\sum_{i=1}^{n}{y'}_i^2 \quad {}^{25}$$

At the equilibrium position (figure 6b), the remaining potential energy of the system is:

$$E_{\text{remainingMin}}=\frac{1}{2}k\sum_{i=1}^{n}(y'_i-ax'_i)^2$$

[25]This quantity permits also to interpret the Variance of the statistical variable Y. The subject was not discussed analytically with students since a detailed work on interpreting Variance, was done in their previous introductory statistics' course (footnote 8, Kourkoulos et al 2006b).



a) b)

Figure 6

The liberated potential energy of the system is:

$$E_{\text{liberatedMax}} = E_{\text{initial}} - E_{\text{remainingMin}} = \frac{1}{2}k\sum_{i=1}^{n} y'^2_i - \frac{1}{2}k\sum_{i=1}^{n}(y'_i - ax'_i)^2$$

This is the maximum amount of potential energy that the system can liberate since the remaining potential energy is the minimum.

Let us consider the ratio $E_{\text{liberatedMax}}/E_{\text{initial}}$; this coefficient gives the maximum percentage of the initial potential energy that the springs' system can liberate. So, it is a coefficient of efficiency of the system, if the system is considered as an energy reservoir.

It is easy to prove that this simple proportion is the square of Pearson's correlation coefficient. Thus, in the context of the model the Pearson coefficient gets a clear meaning.

Moreover, it is easy to see that when the minimum remaining potential energy (the non-exploitable energy) is small compared to the total potential energy, P^2 is large (and inversely)

$$P^2 = \frac{E_{\text{liberatedMax}}}{E_{\text{initial}}} = 1 - \frac{E_{\text{remainingMin}}}{E_{\text{initial}}}$$

This also concerns the corresponding squared deviations:

$$P^2 = 1 - \frac{\sum_{i=1}^{n}(y'_i - ax'_i)^2}{\sum_{i=1}^{n} y'^2_i}$$

Qualitatively, it is clear that:

When the deviations of the attachment points from the LSSL are small **compared** to their distances from the axis $O'x'$, then $|P|$ is large (close to 1), and if the attachment points are on the least squares' straight line then $|P| = 1$.

Remarks (3):

(i) As we have seen for the variance (Kourkoulos et al 2006b, Tzanakis, Kourkoulos 2004) and for the LSSL (previously), when an adequate physical model is established, not only the examined elements get a clear initial meaning, but also properties and aspects otherwise difficult to understand can be easily clarified; the same holds for P in the context of this model. For example, a common misunderstanding concerning P is that if $P = 0$ then the statistical variables X, Y are independent. From our interpretation, we have that $P = 0$ when $E_{\text{liberatedMax}} = 0$ and $E_{\text{initial}} \neq 0$. For having $E_{\text{liberatedMax}} = 0$ (no liberated energy at all), the attachment bar must not move from the initial position $O'x'$. Thus, any distribution of attachment points such that springs annihilate mutually their influences (forces and moments) leaving the bar immobile at $O'x'$, gives $P = 0$. On the basis of this

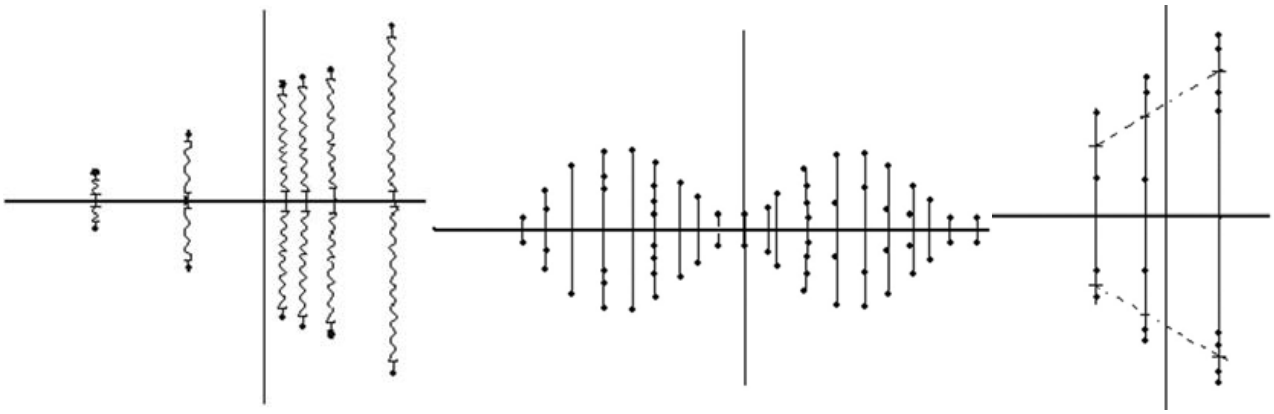


Figure 7

remark it is easy to construct as many examples as one wishes (in fact one can construct whole categories of them) where P=0 but obviously X, Y are dependent.

The three examples above belong to the large category of examples for which the average ordinate of the points having the same abscissa is 0 (so the springs with the same abscissa annihilate their forces and moments).

(ii) Although we didn't discuss this with our students, the model offers important interpretative possibilities for elaborating on other interesting questions (of open type). This permits a more thorough understanding of the involved statistical objects:

a) What kind of changes in a set of points leave unchanged the LSSL and/or P?

b) If some points of the set change, how do their changes influence covariance, the variance of the variables, the LSSL and P? Conversely, how can we change the position of some points of the set in order to obtain a given change of the aforementioned quantities?

c) Are LSSL and P internal characteristics of the set of points?

For a given set of points in the plane, if we rotate the axes Ox, Oy, do LSSL and/or P change? If yes, in which way?

Final remarks

- Using models as the examined one in the introductory teaching of statistics allows students to meaningfully interpret the purely mathematical version of statistical methods; in this case MLS and their associate aggregates (LSSL, Pearson coefficient, squares residuals, ...). This interpretation clarifies important aspects of the subject and ameliorates students' understanding of the mathematical version of statistical methods and aggregates. This amelioration, as well as the fact that the students dispose interpretative models, constitute important assets in the effort to understand the meaning of the methods and aggregates in more difficult contexts (such as those referring to social phenomena) where aggregates express only data tendencies. On the contrary, as remarked in section 1, confining the body of used examples in situations related to social phenomena constitute an important defect of introductory teaching approaches.

- The behavior of our students furnish initial indications, given their small number, that introducing the examined model in introductory teaching approaches of statistics will be feasible and fruitful, on the condition that the students dispose some rudiments of knowledge in elementary physics. However, further investigation is needed, especially concerning its use in whole class course.

- Here we studied the didactical virtues of the model concerning the introduction of the discussed statistical concepts. However, the model offers important such possibilities, which concern more thorough aspects of these concepts as well (e.g. see remarks 3 (ii)); their didactical investigation is an appealing possibility.

- The examined model provides an example on the clear meaning statistical concepts, which are considered to be obscure and difficult for the students, can get in the context of adequate physical situations. An important relevant issue is the elaboration of other adequate interpretative models for these concepts, since the use of more than one such model in the teaching activities creates interactions that are positive for students' understanding.

References

– Batanero, C., Godino, J. D., Vallecillos, A., Green, D. E., Holmes, P., 1994, "Errors and difficulties in understanding elementary statistical concepts", *I.J.M.E.S.T.*, **25**(4), pp. 527–547.

– Batanero, C., Estepa, A., Godino, J. D., Green, D. R., 1996, "Intuitive strategies and preconceptions about association in contingency tables", *Journal for Research in Mathematics Education, 27*, pp. 151–169.

– Cobb, P., McClain, K., Gravemeijer, K., 2003, "Learning About Statistical Covariation", *Cognition and Instruction*, **21**(1), pp. 1–78.

– Darlington, R., 1969, "Deriving Least-Squares Weights Without Calculus", *The American Statistician*, **23**(5), pp. 41–42.

– del Mas, R., Liu, Y., 2005, "Exploring students' conceptions of the standard deviation", *Statistics Education Research Journal (SERJ)*, **4**(1), pp. 55–81.
http://www.stat.auckland.ac.nz/serj

– Forester, P. A., 2005, "Introducing the Least Squares Regression Principle with Computer Technologies" in A. Rogerson (ed), *Proceeding of the 8^{th} International Conference of the "Mathematics Education into the 21st Century Project"*, Universiti Teknologi Malaysia, pp. 87–91.
http://math.unipa.it/∼grim/21_project/21_malasya_Forster87-91_05.pdf

– Gordon, Sh., Gordon, F., 2004, "Deriving the Regression Equations without Calculus", *Mathematics and Computer Education*, **38**(1), pp. 64–68.

– Kourkoulos, M., Tzanakis, C., 2003a, "Graphic representations of data and their role in understanding elementary statistical concepts: An experimental teaching based on guided research work in groups" (in Greek), *Proceeding of the 3^{rd} Colloquium on Didactics of Mathematics*, University of Crete, pp. 209–228.

– Kourkoulos, M., Tzanakis, C., 2003b, "Introductory Statistics with problem-solving activities and guided research work, assisted by the use of EXCEL" in Triandafyllidis, T. & Hadjikyriakou, C., (eds) *Proceedings of ICTMT6*, Athens: New Technologies Publications, pp. 109–117.

– Kourkoulos, Tzanakis, 2006a, "An epistemological and didactical analysis concerning statistical variance supported by experimental teaching work", preprint, University of Crete.

– Kourkoulos, M., Mandadakis, V., Tzanakis, C., 2006b, "Didactical models enhancing students understanding of the concept of Variance in Statistics", *Proccedings of the 3rd ICTM, Istanbul*, Pub. N.Y. John Wiley & Sons, CD-ROM, Paper-151.pdf.

– Mevarech, Z., 1983, "A deep structure model of students' statistical misconceptions", *Educational Studies in Mathematics* **14**, pp. 415–429.

– Moritz, J., 2004, "Reasoning about covariation", in D. Ben-Zvi & G. Garfield (eds), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, Dordrecht : Kluwer, pp. 227–255.

– Porter, Th. M., 1986, *The rise of statistical thinking: 1820–1900*, Princeton : Princeton University Press.

– Reading, C., Shaughnessy, J. M., 2004, "Reasoning about variation", in D. Ben-Zvi, G. Garfield (eds), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, Dordrecht : Kluwer, pp. 201–226.

– Ross, J. A., Cousins, J. B., 1993, "Patterns of student growth in reasoning about correlational problems", *Journal of Educational Psychology*, **85**(1), pp. 49–65.

– Scariano, S. M., Calzada, M., 2004, "Three Perspectives on Teaching Least Squares", *Mathematics and Computer Education* 38(3), pp. 255–264.

– Shaughnessy, J. M., 1992, "Research in probability and statistics: reflections and directions", in D. A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning*, N.Y. : Macmillan, pp. 465–494.

– Stanley, J., Glass, G., 1969, "An Algebraic Proof that the Sum of the Squared Errors in Estimating Y from X via b, and bo is Minimal", *The American Statistician*, 23(1), pp. 25–26.

– Stigler, S. M.,1986, *The History of Statistics: The measurement of uncertainty before 1900*, Cambridge, MA : Harvard University Press.

– Stigler, S. M., 1999, *Statistics on the table: The history of statistical concepts and methods*, Cambridge, MA : Harvard University Press.

– Tzanakis, C., Kourkoulos, M., 2004, "May history and physics provide a useful aid for introducing basic statistical concepts?", *Proceedings of the HPM Satellite Meeting of ICME-10*, Upsalla University, pp. 425–437.

– Watson, J. M., Kelly, B. A., Callingham, A., Shaughnessy, J. M, 2003, "The measurement of school students' understanding of statistical variation", *I.J.M.E.S.T.*, 34(1), pp. 1–29.