Open Journal of Statistics, 2015, 5, 658-664
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57067

The Dual of the Least-Squares Method

Quirino Paris

Department of Agricultural and Resource Economics, University of California, Davis, CA, USA
Email: paris@primal.ucdavis.edu

Received 18 August 2015; accepted 8 December 2015; published 11 December 2015

Abstract

This paper presents the dual specification of the least-squares method. In other words, while the traditional (primal) formulation of the method minimizes the sum of squared residuals (noise), the dual specification maximizes a quadratic function that can be interpreted as the value of sample information. The two specifications are equivalent. Before developing the methodology that describes the dual of the least-squares method, the paper gives a historical perspective of its origin that sheds light on the thinking of Gauss, its inventor. The least-squares method is firmly established as a scientific approach by Gauss, Legendre and Laplace within the space of a decade, at the beginning of the nineteenth century. Legendre was the first author to name the approach, in 1805, as "méthode des moindres carrés", a "least-squares method". Gauss, however, used the method as early as 1795, when he was 18 years old. Again, he adopted it in 1801 to calculate the orbit of the newly discovered planet Ceres. Gauss published his way of looking at the least-squares approach in 1809 and gave several hints that the least-squares algorithm was a minimum variance linear estimator and that it was derivable from maximum likelihood considerations. Laplace wrote a very substantial chapter about the method in his fundamental treatise on probability theory published in 1812.

Keywords

Least Squares, Primal, Dual, Pythagoras Theorem, Noise, Value of Sample Information

1. Introduction

The least-squares method has primal and dual specifications. The primal specification is well known: Given a regression function (either linear or nonlinear) and a sample of observations, the goal is to minimize the sum of the squared deviations between the data and the regression relation, as discussed in Section 3. The dual specification is not known because it is not sought out over the past two hundred years. This paper presents such a dual specification in Section 4. First, however, the reader is offered a historical and illuminating perspective of the least-squares method in the words of its inventor.

How to cite this paper: Paris, Q. (2015) The Dual of the Least-Squares Method. *Open Journal of Statistics*, **5**, 658-664.
http://dx.doi.org/10.4236/ojs.2015.57067

2. Historical Perspective

Karl Friedrich Gauss, at the age of 18, conceived the least-squares (LS) method. However, he did not publish it until 1809, [1]. There, he states that "*Our principle, which we have used since the year* 1795, *has lately been published by Legendre in the work Nouvellesméthodes pour la détermination des orbites des comètes, Paris* 1805, *see* [2], *where several other properties of this principle have been explained, which, for the sake of brevity, we here omit*" (translation [3]). Furthermore, in the Preface to his book [1], Gauss gives an insightful and illuminating account of how the idea of the least-squares method came to him. Up to that time, "*… in every case in which it was necessary to deduce the orbits of heavenly bodies from observations, there existed advantages not to be despised, suggesting, or at any rate permitting, the application of special methods*; *of which advantages the chief one was, that by means of hypothetical assumptions an approximate knowledge of some elements could be obtained before the computation of the elliptic elements was commenced. Notwithstanding this, it seems somewhat strange that the general problem—To determine the orbit of a heavenly body, without any hypothetical assumption, from observations not embracing a great period of time, and not allowing the selection with a view to the application of special methods—was almost wholly neglected up to the beginning of the present century*; *or at least, not treated by any one in a manner worthy its importance*; *since it assuredly commended itself to mathematicians by its difficulty and elegance, even if its great utility in practice were not apparent. An opinion had universally prevailed that a complete determination from observations embracing a short interval of time was impossible—an ill-founded opinion—for it is now clearly shown that the orbit of a heavenly body may be determined quite nearly from good observations embracing only a few days*; *and this without any hypothetical assumption.*

Some idea occurred to me in the month of September of the year 1801, *engaged at the time on a very different subject, which seemed to point to the solution of the great problem of which I have spoken. Under such circumstances we not infrequently, for fear of being too much led away by an attractive investigation, suffer the associations of ideas, which more attentively considered, might have proved most fruitful in results, to be lost from neglect. And the same fate might have befallen these conceptions, had they not happily occurred at the most propitious moment for their preservation and encouragement that could have been selected. For just about this time the report of the new planet, discovered on the first day of January of that year with the telescope at Palermo, was the subject of universal conversation*; *and soon afterwards the observations made by the distinguished astronomer Piazzi from the above date to the eleventh of February were published. Nowhere in the annals of astronomy do we meet with so great an opportunity, and a greater one could hardly be imagined, for showing most strikingly, the value of this problem, than in this crisis and urgent necessity, when all hope of discovering in the heavens this planetary atom, among innumerable small stars after the lapse of nearly a year, rested solely upon a sufficiently approximate knowledge of its orbit to be based upon these very few observations. Could I ever have found a more seasonable opportunity to test the practical value of my conceptions, than now in employing them for the determination of the orbit of the planet Ceres, which during the forty-one days had described a geocentric arc of only three degrees, and after the lapse of a year must be looked for in a region of the heavens very remote from that in which it was last seen? This first application of the method was made in the month of October,* 1801, *and the first clear night, when the planet was sought for** (*by de Zach, December* 7, 1801) *as directed by the numbers deduced from it, restored the fugitive to observation. Three other new planets, subsequently discovered, furnished new opportunities for examining and verifying the efficiency and generality of the method.*

Several astronomers wished me to publish the methods employed in these calculations immediately after the second discovery of Ceres; *but many things—other occupations, the desire of treating the subject more fully at some subsequent period, and, especially, the hope that a further prosecution of this investigation would raise various parts of the solution to a greater degree of generality, simplicity, and elegance—prevented my complying at the time with these friendly solicitations. I was not disappointed in this expectation, and I have no cause to regret the delay. For the methods first employed have undergone so many and such great changes that scarcely any trace of resemblance remain between the method in which the orbit of Ceres was first computed, and the form given in this work. Although it would be foreign to my purpose, to narrate in detail all the steps by which these investigations have been gradually perfected, still, in several instances, particularly when the problem was one of more importance than usual, I have thought that the earlier methods ought not to be wholly suppressed. But in this work, besides the solution of the principal problems, I have given many things which, during*

the long time I have been engaged upon the motions of the heavenly bodies in conic sections, struck me as worthy of attention, either on account of their analytical elegance, or more especially on account of their practical utility, see [3].

This lengthy quotation points to several aspects of discovery of which scientists were aware more than two hundred years ago: elegance as a crucial scientific criterion, serendipity, and the importance of long periods of reflection in order to better understand the properties of new methods. This last aspect perfectly fits the spirit of the present note that is devoted to the presentation of the dual specification of the least-squares method, a property that was neglected for over two hundred years.

Another striking feature of Gauss' thinking process about measuring the orbit of heavenly bodies consists in his clearly stated desire to achieve the highest possible accuracy, see [3]: "*If the astronomical observations and other quantities, on which the computation of orbits is based, were absolutely correct, the elements also, whether deduced from three or four observations, would be strictly accurate* (*so far indeed as the motion is supposed to take place exactly according to the laws of Kepler*), *and, therefore, if other observations were used, they might be confirmed, but not corrected. But since all our measurements and observations are nothing more than approximations to the truth, the same must be true of all calculations resting upon them, and the highest aim of all computations made concerning concrete phenomena must be to approximate, as nearly as practicable, to the truth. But this can be accomplished in no other way than by a suitable combination of more observations than the number absolutely requisite for the determination of the unknown quantities. This problem can only be properly undertaken when an approximate knowledge of the orbit has been already attained, which is afterwards to be corrected so as to satisfy all the observations in the most accurate manner possible.*

It can only be worth while to aim at the highest accuracy, when the final correction is to be given to the orbit to be determined. But as long as it appears probable that new observations will give rise to new corrections, it will be convenient to relax more or less, as the case may be, from extreme precision, if in this way the length of the computations can be considerably diminished. We will endeavor to meet both cases".

Here, Gauss seems to be totally aware of the problem connected to out-of-sample prediction and the necessity or, at least, convenience of a recursive algorithm to account for the information carried by new observations.

Gauss' reading becomes even more exciting, see [3]: "*But when we have a longer series of observations, embracing several years, more normal positions can be derived from them*; *on which account, we should not insure the greatest accuracy, if we were to select three or four positions only for the determination of the orbit, and neglect all the rest. But in such a case, if it is proposed to aim at the greatest precision, we shall take care to collect and employ the greatest possible number of accurate places. Then, of course, more data will exist that are required for the determination of the unknown quantities*: *but all these data will be liable to errors, however small, so that it will generally be impossible to satisfy all perfectly. Now as no reason exists, why, from among those data, we should consider any six as absolutely exact, but since we must assume, rather, upon the principles of probability, that greater or less errors are equally possible in all, promiscuously*; *since, moreover, generally speaking, small errors oftener occur than large ones*; *it is evident, that an orbit which, while it satisfies precisely the six data, deviates more or less from the others, must be regarded as less consistent with the principles of the calculus of probabilities, than one which, at the same time that it differs a little from those six data, presents so much the better an agreement with the rest. The investigation of an orbit having, strictly speaking, the maximum probability, will depend upon a knowledge of the law according to which the probability of errors decreases as the errors increase in magnitude*: *but that depends upon so many vague and doubtful considerations—physiological included—which cannot be subjected to calculation, that it is scarcely, and indeed less than scarcely, possible to assign properly a law of this kind in any case of practical astronomy. Nevertheless, an investigation of the connection between this law and the most probable orbit, which we will undertake in its utmost generality, is not to be regarded as by any means a barren speculation*". This quotation suggests the seed of a maximum likelihood approach. Which takes on a clear statement in the following quote, see [3]: "*Now in the same manner as, when any determinate values whatever of the unknown quantities being taken, a determinate probability corresponds, previous to observation, to any system of values of the functions* (*of the unknown parameters*); *so, inversely, after determinate values of the functions have resulted from observation, a determinate probability will belong to every system of values of the unknown quantities, from which the value of the functions could possibly have resulted*: *for, evidently, those systems will be regarded as the more probable in which the greater expectation had existed of the event which actually occurred. The estimation of this probability rests upon the following theorem—If, any hypothesis* H *being made, the probability of any determinate event*

E *is* h, *and if, another hypothesis* H' *being made excluding the former and equally probable in itself, the probability of the same event is* h': *then I say, when the event* E *has actually occurred, that the probability that* H *was the true hypothesis, is to the probability that* H' *was the true hypothesis, as* h *to* h'".

Gauss proceeds to state, analytically, the function that represents the probability of an event composed of many observations and to derive from such a statement the least-squares principle, see [3]: "*Therefore, that will be the most probable system of values of the unknown quantities* (*parameters*) *in which the sum of the squares of the differences between the observed and computed values of the functions* (*of the unknown parameters*) *is a minimum, if the same degree of accuracy is to be presumed in all the observations… The principle explained in the preceding* (*paragraph*) *derives value also from this, that the numerical determination of the unknown quantities is reduced to a very expeditious algorithm, when the functions* (*of the unknown parameters*) *are linear*". This quotation contains a clear statement of the LS approach as the minimum variance linear estimator.

Gauss did not name his approach as the least-squares method. This name was suggested first by Adrien Marie Legendre in 1805. In his preface, Legendre states, see [2]: "*After all the problem's conditions have been appropriately specified, it is necessary to calculate the coefficients in such a manner as to make the errors as small as possible. To this goal, the method which seems to me the simplest and most general one consists in minimizing the sum of the squared errors. In this way, one obtains as many equations as unknown coefficients; a way to calculate all the orbit's elements. The method that I will present, and that I call the least-squares method, may be very useful in all problems of physics and astronomy where one needs to obtain the most precise results possible from observations*". Surprisingly, Legendre does not mention Gauss' success in predicting Ceres' orbit that was obtained in 1801 and was—apparently, according to Gauss—very acclaimed among the world's astronomers. Also Legendre derives his LS method directly by stating the problem as a linear function of the unknown parameters, without the more elaborate construct of maximizing the likelihood function formulated by Gauss.

There remains to mention Laplace. In 1812, he published a fundamental textbook about probability theory, see [4], and devoted chapter 4 of book 2 to a probability treatment of the LS methodology. The book was dedicated to Napoleon the Great who, in that year, undertook the ill-fated invasion of Russia. The chapter in question is titled: *The probability of the errors of the average results based upon a large number of observations, and the most advantageous average results.* In this chapter one finds a theoretical foundation of the least-squares method (for linear systems) which results as a consequence of the analysis that the mean observational error will fall within certain given limits. The analysis—says Laplace [4]—leads directly to the results associated with the least-squares method.

When all the properties and features of the LS method were thought to be well known, and when all the possible ways of obtaining the least-squares estimates of a linear system's parameters were thought to have been discovered, there surfaced an intriguing question: What is the dual specification of the least-squares method? It is difficult or, better, impossible to conjecture whether such a question could have occurred to either Gauss, or Legendre, or Laplace. The Lagrangean method [5], that is crucial for answering this question, was published by Lagrange in 1804, with revisions in 1806 and 1808. Perhaps, the greatest obstacle to the idea of the dual LS specification has been the particular way in which the LS problem is formulated and presented to students. To date, the traditionally and universally used approach to the LS estimator has hidden away the analytical path to the dual problem. By now one can say that, at least from the viewpoint of fully understanding its structure, the neglect of the dual of the LS method has left a surprising gap. The objective of this paper is to fill this gap.

3. The Primal of the Least-Squares Method

We abstract from any statistical distribution of the error terms and hypothesis-testing consideration. The traditional (primal) LS approach consists of minimizing the squared deviations from an average relation of, say, a linear model that consists of three parts:

$$y = X\beta + u \tag{1}$$

where y is an $(n\times 1)$ vector of sample observations, X is an $(n\times k)$ matrix of predetermined values, β is a $(k\times 1)$ vector of unknown parameters to be estimated, and u is an $(n\times 1)$ vector of deviations from the quantity $X\beta$.

In the terminology of information theory, relation (1) may be regarded as representing the decomposition of a message into signal and noise, that is

$$\text{message} = \text{signal} + \text{noise} \tag{2}$$

with the obvious correspondence: y = message, $X\beta$ = signal, and u = noise. The quantity y is more generally known as the sample information.

The least-squares methodology, then, minimizes the squared deviations (noise) subject to the model's specification given in Equation (1). Symbolically,

Primal: $$\min SSD = u'u/2 \tag{3}$$

subject to $$y = X\beta + u \tag{4}$$

where *SSD* stands for sum of squared deviations. An intuitive interpretation of the objective function (3) is the minimization of a cost function of noise. We call model (3) and (4) the Primal LS model. The solution of model (3) and (4) by any appropriate mathematical programming routine gives the LS estimates of parameters β and deviations (noise) u.

Traditionally, however, the LS method is presented as the minimization of the sum of squared deviations defined as $SSD = (y - X\beta)'(y - X\beta)$ with the necessity of deriving, first, an estimate of the β parameters and then using their least-squares estimates $\hat{\beta}$ to obtain the LS residuals: $\hat{u} = y - X\hat{\beta}$. This way of presenting the LS method obscures the derivation of the dual specification and is the source of some readers' surprise that LS parameters and residuals may be estimated simultaneously by means of a nonlinear programming solver.

4. The Dual of the Least-Squares Method

The Lagrange approach is eminently suitable for deriving the dual of the least-squares method. Hence, choosing the ($n \times 1$) vector variable e to indicate n Lagrange multipliers (or dual variables) of constraints (4), the relevant Lagrangean function is stated as:

$$L(u, \beta, e) = u'u/2 + e'(y - X\beta - u) \tag{5}$$

with first order necessary conditions (FONC)

$$\frac{\partial L}{\partial u} = u - e = 0 \tag{6}$$

$$\frac{\partial L}{\partial \beta} = -X'e = 0 \tag{7}$$

$$\frac{\partial L}{\partial e} = y - X\beta - u = 0\,. \tag{8}$$

A first remarkable insight is that, from FONC (6), the Lagrange multipliers (dual variables), e, of the LS method are identically equal to the deviations (primal variables, noise), u. Each observation in model (4), then, is associated with its specific Lagrange multiplier that turns out to be identically equal to the corresponding deviation. A Lagrange multiplier measures the amount of change in the objective function due to a change in one unit of the associated observation. If a Lagrange multiplier is too large, the corresponding observation may be an outlier. Secondly, FONC (6) and (7), combined into $X'u = 0$, represent the orthogonality condition between the vector of deviations and the space of predetermined values of the linear model (1) that characterizes the LS approach. The equations $X'u = 0$ constitute the constraints of the dual model. In general, the dual objective function is given by the maximization of the Lagrangean function with respect to dual variables, keeping in mind that $e = u$. And since we are dealing with a quadratic specification, the Lagrangean function can be simplified substantially by means of relation (6), restated as:

$$u = e \quad \text{and} \quad u'u = u'e\,. \tag{9}$$

Therefore, the Lagrangean function can be streamlined as:

$$L(u, \beta, e) = u'u/2 + e'(y - X\beta - u) = u'y - u'u/2 \tag{10}$$

using relations (7) and (9).

The Dual of the LS model can now be assembled as:

Dual: $$\max NVSI = u'y - u'u/2 \tag{11}$$

subject to $$X'u = 0 \tag{12}$$

Constraints (12) constitute the orthogonality conditions of the LS approach, already mentioned above. An intuitive interpretation of the dual objective function can be formulated within the context of information theory. Hence, the dual problem seeks to maximize the net value of the sample information (*NVSI*). Typically, dual variables (Lagrange multipliers) are regarded as marginal sacrifices or implicit (shadow) prices of the corresponding constraints. We have already seen that dual variables *e* are identically equal to primal variables *u*. Thus, in the LS specification, the variables *u* have a double role: as deviations in the primal model (noise) and as "implicit prices" in the dual model. The quantity $u'y$, therefore, is interpreted as the gross value of sample information. This quantity is netted out of the "cost of noise", $u'u/2$, to provide the highest possible level of the *NVSI* objective function.

In the dual model, the vector of parameters β is obtained as a vector of Lagrange multipliers of constraints (12). In fact, from the Lagrangean function of the dual problem stated as:

$$L^*(u,\mu) = y'u - u'u/2 - \mu'[X'u]$$

where μ is a $(k \times 1)$ vector of Lagrange multipliers associated with constraints (12), the corresponding FONCs are

$$\frac{\partial L^*}{\partial u} = y - u - X\mu = 0 \tag{13}$$

$$\frac{\partial L^*}{\partial \mu} = -X'u = 0\,. \tag{14}$$

Hence, from Equation (13) and Equation (14), we can write

$$X'y - X'u - X'X\mu = 0 = X'y - X'X\mu$$

that results (assuming the nonsingularity of the $(X'X)$ matrix) in the formula of the well known LS estimator

$$\hat{\mu} = (X'X)^{-1}X'y = \hat{\beta}\,.$$

All the information of the traditional LS primal problem is contained in the LS dual model, and vice versa. Hence, the pair of dual problems—the primal [(3)-(4)] and the dual [(11)-(12)]—provides identical LS solutions for separating signal from noise.

At optimal solutions, $\hat{u}$, of both the primal and the dual LS models, the two objective functions are equal and can be written as

$$\text{Primal} = \text{Dual}$$
$$\hat{u}'\hat{u}/2 = \hat{u}'y - \hat{u}'\hat{u}/2.$$

It follows that

$$\frac{\partial\, \hat{u}'\hat{u}/2}{\partial y} = \hat{u}$$

which demonstrates a previous assertion, namely that the change in the primal objective function corresponding to a marginal change in each sample observation is equal to its associated Lagrange multiplier that is identically equal to the corresponding deviation. The two primal and dual objective functions can also be rewritten as:

$$\frac{\hat{u}'\hat{u}}{n} = \frac{\hat{u}'y}{n}\,.$$

Hence, the quantity $\hat{u}'y/n$ represents an equivalent way to estimate the variance of the sample deviations.

5. The Dual of the LS Method and Pythagoras Theorem

An interpretation of the dual pair of LS problems, without reference to any empirical context, can be formulated

using the Pythagorean theorem. With the knowledge that a solution to the LS problem requires the fulfillment of the orthogonality conditions $X'u = 0$, given in (12), Pythagoras theorem allows for the statement

$$y'y = y'(X\beta + u) = (X\beta + u)'(X\beta + u) = \beta' X'X\beta + 2\beta' X'u + u'u = \beta' X'X\beta + u'u \quad (15)$$

and also

$$y'y = y'(X\beta + u) = y'X\beta + y'u = (X\beta + u)' X\beta + y'u = \beta' X'X\beta + y'u.$$

Therefore,

$$u'u = y'u$$

that can be restated as:

$$u'u/2 = y'u - u'u/2 \quad (16)$$

which corresponds to the two objective functions of the primal (3) and the dual (11): the left-hand-side of equation (16) is the primal objective function to be minimized and the right-hand-side of the same equation is the dual objective function to be maximized. By the Pythagoras theorem (expressed by Equation (15)), for any given vector of observations *y*, the minimization of the noise function $u'u$ must be matched by the maximization of $\beta' X'X\beta$ which corresponds to the maximization of the signal. Equivalently, minimizing the length of the deviation vector *u* corresponds to maximizing the length of the vector $X\beta$, which is the projection of the observation vector *y* onto the space of predetermined variables *X*.

6. Conclusion

This paper has retraced the history of the least-squares method and has developed the dual specification of it which is a novel way of looking at the LS approach. It has shown that the traditional minimization of the sum of squared deviations that give the name to the algorithm is equivalent to the maximization of the net value of sample information.

References

[1] Carolo Friderico, G. (1809) Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium. Hamburgi Sumtibus Frid, Pertheset I. H. Besser.

[2] Adrien Marie, L. (1805) Nouvelles Méthodes pour la Détermination des Orbites des Comètes, Chez Firmin Didot, Libraire pour la Mathématique, la Marine,l'Architecture, et les Éditionsstéreotypes, rue de Thionville. A Paris.

[3] Charles Henry, D. (1857) Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections: A Translation of Gauss's "Theoria Motus" with an Appendix. Little Brown and Com pany, Boston.

[4] Le Comte, L.M. (1812) Théorie Analytique des Probabilités, M.me V. Courcier, Imprimeur-Libraire pour la Mathématiques, quai des Augustins. Paris.

[5] Joseph-Louis, L. (1808) Leçonssur le Calcul des Fonctions, M.me V. Courcier, Imprieur-Librairepour la Mathématiques, quai des Augustins. Paris.