The Dual of the Least-Squares Method

Quirino Paris
University of California, Davis

Abstract

The least-squares method was 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, is credited to have used it as early as 1795, when he was 18 years old. He, subsequently, 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. Surprisingly, there still remains an unexplored aspect of the least-squares method: since the traditional formulation is stated as minimizing the sum of squared deviations subject to the linear (or nonlinear) specification of a regression model, this mathematical programming problem must have a dual counterpart. This note fills this gap and shows that the least-squares estimates of unknown parameters and deviations can be obtained by maximizing the net value of sample information.

Key Words: 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 was 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.

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 (Gauss, p. 221). There, he states that "*Our principle, which we have used since the year 1795, has lately been published by Legendre in the work* Nouvelles méthodes pour la détermination des orbites des comètes, Paris

1806*, where several other properties of this principle have been explained, which, for the sake of brevity, we here omit.*" (translation by Charles Henry Davis, 1857, p. 270).

Furthermore, in the Preface to his book (pp. viii-x), 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 unfrequently, 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 propitius 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, Decembre 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. (Davis, pp. xiii-xvi).

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 (Davis, p. 249): "*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 (Davis, pp. 252-253): "*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 astonomy. 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 (Davis, p. 255): "*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 (Davis, pp. 260-261): "*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 (1805). In his preface, Legendre states (p. viii): "*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 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 (p. 348) – 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, 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 rather sizable gap. The objective of this note 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 x 1) vector of sample observations, X is an (n x k) matrix of predetermined values, β is a (k x 1) vector of unknown parameters to be estimated, and u is an (n x 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' disbelief that LS parameters and residuals may be estimated simultaneously.

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 x* 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:

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

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

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

$$\text{subject to} \qquad X'u = 0\,. \tag{12}$$

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 equations (13) and (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

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

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. This result was never indicated in any statistics or econometrics textbook.

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

$$\begin{aligned} y'y &= y'(X\beta + u) = (X\beta + u)'(X\beta + u)\\ &= \beta'X'X\beta + 2\beta'X'u + u'u\\ &= \beta'X'X\beta + u'u \end{aligned} \tag{15}$$

and also

$$
\begin{aligned}
y'y &= y'(X\beta + u) = y'X\beta + y'u \\
&= (X\beta + u)'X\beta + y'u \\
&= \beta'X'X\beta + y'u.
\end{aligned}
$$

Therefore,

$$u'u = y'u$$

that can be restated as

$$u'u/2 = y'u - u'u/2 \tag{16}$$

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 $u'u$ must be matched by the maximization of $\beta'X'X\beta$. 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 note 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 gives the name to the algorithm – is equivalent to the maximization of the net value of sample information.

References

Gauss, Carolo Friderico. (1809). *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*. Hamburgi Sumtibus Frid. Perthes et I. H. Besser.

Davis, Charles Henry. (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 Company, Boston.

Lagrange, Joseph-Louis. (1808). *Leçons sur le Calcul des Fonctions*, M.me V. Courcier, Imprimeur-Libraire pour la Mathématiques, quai des Augustins, n. 55, Paris.

Laplace, M. Le Comte. (1812). *Théorie Analytique des Probabilités*, M.me V. Courcier, Imprimeur-Libraire pour la Mathématiques, quai des Augustins, n. 57, Paris.

Legendre, Adrien Marie. (1805). *Nouvelles Méthodes pour la Détermination des Orbites* *des Comètes*, Chez Firmin Didot, Libraire pour la Mathématique, la Marine, l'Architecture, et les Éditions stéreotypes, rue de Thionville, n. 116, A Paris.

Afterword

I assert rather comfortably that very few present-day statisticians (if any) are aware of the history of the least-squares method and I am certain that none knows about its dual specification. In spite of these facts, the gatekeepers of statistics have desk rejected the present paper with justifications that – at times – are hopelessly disturbing. When a problem has primal and dual specifications, ignoring the dual model is akin to ignoring the other side of the story and, therefore, preferring ignorance to full knowledge.

I began this saga by submitting the paper to The American Statistician. I specifically indicated that it would have been suitable for its Teachers Corner. The editor, professor John Stufken, took more than 100 days to deliver the following nonsensical rejection: "*... your writing makes no distinction between errors (noise) and residuals. The model y = X beta + u, with u as the errors, is an identity. Here u is a random error term, and minimizing u'u makes no sense since the errors, while unobservable, are what they are. What does make sense is finding a beta_hat that minimizes (y - X beta)'(y - X beta) over all possible choices of beta. We would then define e = y - X beta_hat (this e is different from your e), and refer to this as the vector of residuals. But e and u are not the same. - thus the least square problem is normally written as an unconstrained minimization problem: min (y - X beta)' (y - X beta). What is gained by using your form (3) and (4), suggesting a constrained optimization problem? What insights are gained from any of the observations made in the paper?*"

Notice that I specifically avoided the use of terms such as "errors" and "residuals" (I called them "deviations") in order to concentrate on the essential aspects of the least-squares method. A highly surprising statement of professor Stufken, then, deals with his writing that "*The model y = X beta + u, with u as the errors, is an identity.*" This opinion is certifiably wrong since, otherwise, every equation would be an identity and the error specification analysis would be useless. The next statement is even more appalling: "*…minimizing u'u makes no sense… [but] What does make sense is finding a beta_hat that minimizes (y - X beta)'(y - X beta)…*" To help editor Stufken in his understanding of the least-squares approach, I suggest that he ought to reflect, for a moment, on how the latter sum of squares *(y - X beta)'(y - X beta)* is obtained. In the end, Stufken's questions "*What is gained by using your form (3) and (4), suggesting a constrained optimization problem? What insights are gained from any of the observations made in the paper?*" reveal his total inability to appreciate the process of "**understanding**" since all his efforts are concentrated on the process of "**how to do things**." Unfortunately, this is a rather common example of the scientific prejudice that must be confronted when dealing with gatekeepers of science of professor Stufken's type.

I lowered my expectations and the second submission was to the online journal Economics Bulletin whose mission statement includes "*…free and extremely rapid scientific communication across the entire community of research economists*." In fact, when I submitted my paper, the average editorial lag for refereed contents over the last 12 months was of 53.88 days. Unfortunately, after more than 120 days, the editor, professor John P. Conley, notified me that "*I have not yet succeeded in getting a response from the referees.*" I withdrew my submission.

Daringly, the third submission was to the Journal of the American Statistical Association. In a moment of scientific enlightenment, I thought that full knowledge and understanding of both the history and the structure of the least-squares methodology ought to be disseminated in a prominent way among statisticians who, probably, use this approach on a daily basis. I was wrong. Professor Xuming He, Co-Editor of JASA wrote: "*I regret to inform you that we have now considered your paper but unfortunately feel it unsuitable for publication in Journal of the American Statistical Association (JASA). Your paper discussed an interesting duality for the least squares method, but its level of novelty and significance falls short of what we normally expect from JASA publications*." According to editor Xuming He, although nobody knows the duality of least squares, "*its level of novelty and significance falls short of what we normally expect from JASA publications.*"

What about the Journal of the Royal Statistical Society, Series B? Another desk rejection. The Joint Editor, professor Gareth Roberts, wrote: "*I enjoyed reading through your account. Certainly much of the historical information about the discovery of the least squares method was new to me, and I found it of some interest. However this is not material which is suitable for JRSS B. The other contribution of your article is to consider the dual to the least squares minimisation problem. Unfortunately, although this perspective is quite unusual, I did not find it a sufficiently substantial observation to consider that it would fare well in a full review*."

Down from the Olympus Mountain of statistics, the next submission was to the American Journal of Agricultural Economics, whose contributors and readers are among the most active users of the least-squares approach. No luck. The editor, professor David Hennessy, wrote: "*In the case of your submission, I have determined that it is not appropriate for this journal and it would not be a good use of the referees time or your time to send this manuscript to outside reviewers. In this case, I am very confident that if I were to choose outside reviewers, they would concur with my decision. In any event, as one of the ultimate gatekeepers of what articles are published in this journal, I do not feel that this article would be appropriate for publication in American Journal of Agricultural Economics* ***even if your paper did happen to receive a positive review***." Prejudice or Catch 22?

The last submission was to the Journal of Statistics Education. I thought that knowing two ways for explaining the LS method to statistics students ought to be better than knowing only one. The editor, professor John Gabrosek, does not agree. He wrote: "*The* *words primal and dual need a definition. How should the reader (who is likely a teacher of statistics and not a researcher) interpret these words?*" This is discouraging: I thought that, even before reading my paper, a teacher of statistics ought to know the meaning of the words **least squares**. Gabrosek's comment is equivalent to say that these words need a definition. Of course, my paper defines the primal and dual **problems** (see the Introduction) and in so doing it defines also the primal and dual words. The editor also wrote: "*If the practical application of the least-squares method is the same whether using the primal result or the dual result then what is gained by this foray into history and the dual result? The introductory statistics class … is already jammed full of material. What practical benefit is there to adding this to the curriculum?*" For professor Gabrosek, crunching numbers seems to be more important than understanding and knowledge.

Gatekeepers of science create obstacles to scientific progress and delay further discoveries. Fortunately, the internet has arrived to deflate their censorial and arbitrary power. It means only that my paper will not receive brownie points by some administrator…