A Note on "Squared Versus Unsquared Deviations for Lines of Best Fit"

By David M. Bell

This note comments on an article by Harold B. Jones and Jack C. Thompson, "Squared Versus Unsquared Deviations for Lines of Best Fit," which appeared in the April 1968 issue of Agricultural Economics Research. The purpose of that article was "to compare and contrast the two approaches in the hope that more effective utilization of both techniques will result." The authors point out that alternatives to the least squares technique exist and that these alternatives should be considered for each problem so that the most appropriate procedure may be chosen. They suggest that the differences between the two concepts are frequently unrecognized or ignored except in mathematical theory studies (4, p. 64).[1]

Given their purpose, the paper falls short in two ways. First, the treatment of squared versus unsquared deviations is less than complete. And second, a significant part of the discussion is not closely related to the primary issue, and may confuse readers with limited statistical background—those to whom the article was primarily directed.

As an illustration of the latter, while discussing least squares the authors state that "... the attempt to substitute probability for logic or cause and effect relationships carries one beyond the realm of true scientific inquiry" (4, p. 65). This is true. Statistical techniques provide only probability statements that the researcher must then interpret. But this sheds little light on the basic issue—the choice of estimation technique.

Regression as an Estimation Procedure

Jones and Thompson assume that, usually, once a regression line is fitted to the data all statistical work is completed; the line is given—it is absolute. If this were the case, their argument would be more tenable, and if a regression line fitted by absolute deviations resulted in better estimators for the given purpose than those derived from squared deviations, then that technique would be preferable. But alternatives to squared deviations other than absolute deviations may also be considered.

Consider the general equation

$$\sum_{i=1}^{N} \left| (Y_i - \tilde{Y}_i)^{\alpha} \right|$$

where Y_i equals the observed value of Y corresponding to X_i and $\tilde{Y}_i$ is the "predicted" value of that Y_i. If $\alpha = 1$, we have the absolute deviations case; if $\alpha = 2$, the squared deviations case. But α could also be set equal to .5 or 1.5 or 4 or any other value. The basic question is which value of α should be chosen.

In fact, the choice of α should be determined by the loss function (8, p. 15).[2] Simply stated, the loss function is an approximation of the cost of making a wrong or bad decision. For example, suppose that the true value of some variable, which may range from 0 to m, is A (fig. 1). The loss may then be stated as some

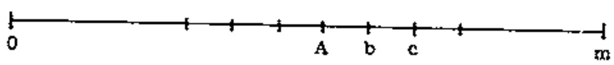

Figure 1

function of the difference between our estimate and the true value A. This function, which is determined by the characteristics of the situation, establishes weights on varying degrees of error. If deciding that c is the true value is twice as costly as deciding that b is the true value, the loss function would be linear; and in terms of

[1] Underscored numbers in parentheses indicate items in the References, page 79.

[2] We are under the assumption that no tests will be made on the line. Otherwise, other considerations to be discussed later would also be important.

regression, absolute deviations ($\alpha = 1$) would be most appropriate. If choosing c is four times as costly as choosing b, the loss function would be quadratic, and squared deviations ($\alpha = 2$) would be appropriate for the regression. α then, is determined by the nature of the loss function, and one could easily conceive a situation where $\alpha = .5$ or $\alpha = 4$ would be most appropriate.

Seldom does the researcher know precisely the true nature of the loss function. Consequently, he assumes a quadratic loss function in the belief that it is the best approximation. But if the researcher knew the loss function to be approximately linear, it would be wise to use absolute deviations. Even this does not remove all the possibility of error, however, since other values close to $\alpha = 1$ may be more appropriate.

Tests on Regression Estimators

But fitting the regression line is often only the beginning of the statistical tests. A well-developed theory exists for testing various aspects of the least squares regression line. Alternatively, the theory for testing various characteristics of the absolute deviations estimators is less developed. Some statisticians say no such test statistics exist. Others say they do, but are much too complicated to justify their use. Regardless of which is correct, one seldom sees them used.

It could be argued that statisticians should make an effort to develop these tests. But until they do, methods other than those of squared deviations will be less fruitful.

Other Considerations

Some contend that a study will be most useful when it is structured so that other researchers can interpret the findings. Following this philosophy, if tests do exist for absolute deviation estimators, but few understand them, the researcher might be well advised to use squared deviation techniques so that others could interpret the findings, and consequently use them. The authors agree with this, concluding their article with, "In the final analysis, it is only when research results are disseminated to others that anything worthwhile can be achieved . . . Any given method should be used but only where it is appropriate and preferably where the results are easily understood by those concerned with the problem. With this kind of philosophy, we can expect a wider acceptance of our research results" (4, p. 68). While I do not accept this argument entirely,[3] it does have some merit.

[3]If researchers continually used familiar and common techniques, progress would not be forthcoming.

The case of fitting a regression line is not unimportant. Simple calculating procedures have been developed for fitting the line by using squared deviations. For smaller problems, hand calculators can be used while larger problems can be solved by using standard regression programs on electronic computers. Fitting the regression by using absolute deviation is more difficult. In extremely simple problems involving only one independent variable, graphic methods can be used on a trial and error basis. Otherwise, a linear programming procedure is necessary. Electronic computers can handle both least squares regression and linear programming models with great speed, but the relative ease associated with least squares makes it less expensive and easier to manipulate for the average researcher.

Conclusions

In conclusion, the authors have two acceptable hypotheses in their article. The first may be interpreted as: Researchers use "popular" statistical techniques in their problem solving without fully understanding the techniques. Two aspects of this hypothesis are undoubtedly true. First, researchers often use a "popular" statistical technique without being certain it is most appropriate for their problem, although this does not, in itself, justify abandoning the "popular" technique. And second, many researchers do not fully understand the technique they utilize.

The second hypothesis is: In certain situations, the absolute deviation technique may be superior to squared deviations for fitting a regression line, and in such cases it should be used. I agree. There are situations where the objectives are best served by the absolute deviations method, and in those situations use of squared deviations would not be logical.

Separating these hypotheses would have enriched the usefulness of the article by eliminating the source of confusion to those with limited statistical background.

Had the authors more clearly set forth the merits and shortcomings of both techniques, more judicious use of the two would have been possible. Instead, they present a paper supporting and recommending increased use of absolute deviations—a technique many researchers consider sterile.

References

(1) Dixon, Wilfred J., and Frank J. Massey, Jr. Introduction to Statistical Analysis. McGraw-Hill Book Co., New York, 2d ed., 1957.

(2) Ezekiel, Mordecai, and Karl A. Fox. Methods of Correlation and Regression Analysis. John Wiley and Sons, New York, 13th ed. rev., 1958.

(3) Johnston, J. Econometric Methods. McGraw-Hill Book Co., New York, 1963.

(4) Jones, Harold B., and Jack C. Thompson. "Squared Versus Unsquared Deviations for Lines of Best Fit." Agr. Econ. Res. 20(2): 64-69, April 1968.

(5) Mood, Alexander M., and Franklin A. Graybill. Introduction to Theory of Statistics. McGraw-Hill Book Co., New York, 2d ed., 1963.

(6) Stapleton, J. H. Simple and Multiple Linear Regression. Mimeographed notes, Mich. State Univ. Bookstore, East Lansing, Mich.

(7) Steel, R. G. D., and J. H. Torrie. Principles and Procedures of Statistics. McGraw-Hill Book Co., New York, 1960.

(8) Theil, Henri. Applied Economic Forecasting. North-Holland Publishing Co., Amsterdam, 1966.

(9) Wagner, H. M. "Linear Programming Techniques for Regression Analysis." Jour. Amer. Statis. Assoc., Vol. 54, 1959.