Regression Analysis in Medical Research

GUY B. FAGUET, MD, and HARRY C. DAVIS, MS, Augusta, Ga

ABSTRACT: Even the most respected medical journals continue to publish articles containing unwarranted conclusions, which thus appear validated. This often results from the unfamiliarity of medical investigators with statistics leading to improper study design, data collection, analysis, and presentation. The increased use of multivariate analysis adds to the perplexity of medical readers not adequately prepared to judge the statistical method. This article attempts to acquaint readers with the terminology of regression analysis and how to use regression formulas.

THE UNFAMILIARITY of medical investigators with statistics often results in improper study design or inappropriate collection, analysis, and presentation of data.[1-3] Such studies may adversely affect patient management and subsequent related research,[4] needlessly place subjects at risk, and misuse resources. Unfortunately, misleading information still finds its way into the medical literature, bestowing legitimacy upon erroneous conclusions.[5-8] This fact is underscored by the recognition that the majority of readers of medical journals do not themselves have the necessary expertise to judge the statistical method. The increased use of multivariate analysis in clinical research has added to the bewilderment of many physicians. A better understanding of these regression techniques has thus become necessary.

The most commonly used regression analysis technique is multiple linear regression. However, other techniques such as logistic regression, nonlinear regression, and discriminant analysis are also encompassed by the term.[9-11] Regression is usually used (1) to test hypotheses, (2) to select variables for prediction models, or (3) to generate prediction formulas.

Because they are particularly suited to assess the correlation among variables and to establish the dependence of one variable upon others, regression equations can be used effectively by the clinician to make a diagnosis or assess prognosis. Calculation of the value of predictor variables singly or in combination might be of greater interest to the medical investigator more concerned with the assessment of the pathophysiology of disease or to the therapist attempting to improve survival by modulating risk factors.[12] This report briefly examines the interpretation and use of regression formulas, particularly with regard to estimating and evaluating the dependence relationships. The selection of variables for a prediction model is dealt with only as it pertains to the use of regression formulas. The use of regression for hypothesis testing is ignored.

Emphasis will be placed on types of regression that are linear with respect to the terms in the model (predictor variables within a regression formula), though the model terms themselves may be nonlinear transformations of the original variables. In some instances, a linear model may not fit the data sufficiently well. One method of determining whether there may be some pattern in the data that is not well explained by a linear model is through an examination of the residuals (a residual is the difference between the actual value of the dependent variable and the predicted value based upon a model). Several statistics have been developed to test for various patterns in residuals and there are a variety of nonlinear regression techniques which might prove to be best in some instances. Because these techniques are complex, it is advisable to consult a statistician to aid in the analyses of residuals and in the use of nonlinear regression techniques.

REGRESSION FORMULAS

Just as a clinician uses pieces of information which are combined to form a single diagnosis, regression analysis can be used to make predictions based upon a formula derived from empirically relating a set of one or more predictor (independent) variables to a single predicted (dependent) variable. Such a formula might look as follows:

$$Y = a + b_1X_1 + b_2X_2 + \ldots . b_nX_n$$

(Formula I) where Y is the value of the predicted variable; a is the value of a constant, commonly called the intercept; X_1, X_2, X_n are the values of predictor variables; and b_1, b_2, and b_n are the constants by which the values of the respective predictor variables are multiplied. The b's are often referred to as regression weights (slopes).

From the Departments of Medicine and Cell and Molecular Biology, Medical College of Georgia, and the Medical and Research Services, Veterans Administration Medical Center, Augusta. Mr. Davis is consulting statistician, Medical College of Georgia, Augusta.

Reprint requests to Guy B. Faguet, MD, Department of Medicine, Medical College of Georgia, Augusta, GA 30910.

To use a regression formula, it is necessary only to substitute the values of the X's into the formula. To exemplify, let us assume that from studying a group of subjects we have developed a formula to predict systolic blood pressure based on stress and daily sodium intake as follows:

Systolic Blood Pressure =
50 + .825x Stress + (−.25)x Daily Sodium Intake

(Formula II) where mean systolic blood pressure = 130, SD = 15 when mean stress (arbitrary units) = 100, SD = 37 and mean daily sodium intake (arbitrary units) = 10, SD = 1. This formula would yield a prediction for systolic blood pressure of 160 for a patient with a stress score of 137 and daily sodium intake of 12. For such a prediction to be accurate, however, the formula must be appropriate to the new sample. This means that the new and the original sample from which the formula derives must be comparable with respect to characteristics that might influence the dependent or independent variables under study. To the extent that the original sample from which the formula derives represents the parent population, the formula will be appropriate and generalizable to that population. In our example, for formula II to have predictive value, the scores for stress and daily sodium intake of the patient must not exceed the range of the respective scores of the original sample. In addition, if the above formula were developed using only middle-aged men, for example, its use to predict systolic blood pressure for an adolescent white woman or an elderly white man would probably be inappropriate.

STANDARDIZING THE VARIABLES

In our example, the regression weight for stress being greater than the regression weight for daily sodium intake does not mean that stress is more important than daily sodium intake in predicting systolic blood pressure. The relative importance of the variables within a regression formula can be ascertained by their conversion to z-scores. To convert to z-scores, the mean is subtracted from each score and the difference is then divided by the standard deviation. Thus, a z-score of −1.2 indicates a score that is below the mean by 1.2 standard deviations.[13] In this case, the regression formula no longer has an intercept term, and the regression weights are called beta (β) weights.[14,15] Using the above example to illustrate, after converting to z-scores, a recomputation of the regression formula II would now generate the following equation:

Systolic Blood Pressure =
.50 Stress + .75 Daily Sodium Intake

(Formula III). Thus, the patient with a stress score of 137 and daily sodium intake of 12, has z-scores for those variables of [(137 − 100)/37] and [(12 − 10)/1], that is, 1 and 2, respectively. The predicted systolic blood pressure is then (.50 × 1 + .75 × 2) or 2 on a z-score scale, which indicates that the predicted blood pressure score is 2 standard deviations above the mean [130 + (2 × 15)] or 160. From the relative weights in formula III it can be seen that daily sodium intake is more important than stress (β weights .75 and .50, respectively) for predicting systolic blood pressure within this formula. It must be remembered, however, that because weights reflect the relative contribution of their respective independent variables to the value of the formula in predicting the dependent variable, adding, subtracting, or replacing variables within a formula will alter the β weights of the variables remaining in the equation. As a corollary, β weights do not necessarily reflect the correlation that each independent variable might have with the dependent variable individually, outside regression formulas. This is true because the predictor variables usually have some intercorrelation, or overlap in prediction. The information as to the predictive value of each independent variable outside regression formulas can be ascertained by calculating its individual correlation with the dependent variable.

STABILITY AND VALIDITY OF REGRESSION WEIGHTS

Regression weights are notoriously unstable, that is, the weights often change drastically when new regression formulas using the same variables are derived from new or altered samples. Normally, the higher the ratio of the number of subjects to the number of variables, the more stable the regression weights tend to be. Though the stability of regression weights might appear desirable for the clinician applying a regression formula, to the investigator developing a regression formula the stability of the regression weights is usually a minor consideration, since he is usually much more concerned with the validity of a regression formula and the proper selection of the variables to be included in the formula. A formula is valid for a sample in the measure that it predicts the dependent variable. The level of prediction is often called "predictive accuracy." To the extent that the sample is representative of the parent population, the formula will also be valid for the population (generalizable). Thus, the validity of a regression formula is usually the most important consideration for both investigator and clinician. Because large samples are not usually required for validity (and often involve an unnecessary waste of resources), some methods (eg, ridge regression,[16,17] Tukey's jackknife technique,[18] and equal weighing of variables[19-22] have been developed to stabilize the regression weights from relatively small samples.

MEASURE OF VALIDITY OF REGRESSION FORMULAS

The most common measure of validity in regression is the multiple correlation coefficient (R) or its square (R^2). This coefficient can range from 0 to 1, with a 1 indicating perfect prediction. The squared multiple correlation coefficient is equal to the proportion by

which error variance can be reduced through the use of the regression formula as opposed to predicting the mean of the dependent variable for all subjects. For example, if we did not know that systolic blood pressure is related to stress and daily sodium intake, we would probably predict the mean for the sample under consideration in order to minimize error variance. Thus, given three values of systolic blood pressure of 136, 130, and 124 (mean = 130), the minimal error variance (the squared difference between the predicted and the actual values) would be $[(136 - 130)^2 + (130 \pm 130)^2 + (124 - 130)^2]/3$, or 24. However, if we are given information about stress and daily sodium intake scores for each individual, this error variance can be reduced considerably through the use of a regression formula that takes such information into account. Thus, if we apply formula II (assumed to have an R^2 of .83) to the above example, the error variance would be reduced from 24 to 4.09 as shown in Table 1.

This increased predictability is due to the relationship between systolic blood pressure (the dependent variable) and stress and daily sodium intake (the two independent variables), a relationship not previously considered. Although the multiple correlation (R) is the most common measure of validity in regression,[14] certain regression techniques use other indices as measures of validity. For example, discriminant analysis usually uses Wilks' lambda,[23] whereas logistic regression usually uses a χ^2 test of model fit.[24] The validity of these techniques is sometimes ascertained via the percentage of subjects correctly classified by the derived equation(s). For example, a study might be done to determine if stress and daily sodium intake can be used to discriminate between persons with high systolic blood pressure and those with normal systolic blood pressure. Once the discriminant formula is obtained, predicted and actual group membership for each individual in the sample and the percentage of grouped cases correctly classified are determined as illustrated in Table 2. The equation used in this example correctly classified 92% of persons with high systolic blood pressure and 80% of those with normal systolic blood pressure, for an overall correct classification of 86%. It should be noted that using the percentage of subjects correctly classified as a measure of validity can be misleading because regression formulas are seldom developed to maximize this criterion.

TABLE 1.

Stress Scores	*Daily Sodium Intake Scores*	*Predicted Systolic Blood Pressure*	*Actual Systolic Blood Pressure*	*Squared Difference Between Actual and Predicted Systolic Blood Pressure*
111	12	138.575	136	6.63
100	10	130	130	0
95	8	126.375	124	5.64
				12.27 ÷ 3 = 4.09

TABLE 2.

		Actual	
		High	*Normal*
Predicted	High	23 (92%)	5
	Normal	2	20 (80%)
		25 (100%)	25 (100%)

GENERALIZABILITY OF REGRESSION FORMULAS

A regression equation can show substantial prediction of the dependent variable in the sample on which it was developed, yet have little or no predictive value for the parent population from which the sample was drawn. The phenomenon is usually called "capitalizing on chance" and because of it, the sample multiple correlation coefficient (R) is often a misleading indicator of the validity of a regression equation for the parent population. Formulas have been developed that estimate the validity of a regression equation for the parent population (generalizability of the formula) by weight correction of R according to the number of subjects, and of predictor variables in the sample generating the equation.[25-28] One such formula[27] is:

$$\hat{R}^2 = \frac{\rho^4 (N - k - 3) + \rho^2}{\rho^2 (N - 2k - 2) + k}$$

(Formula IV) where R^2 = the squared value of the cross-validated multiple R for the population from which the sample was drawn, N = number of subjects, k = number of predictors, and ρ^2 and ρ^4 = multiple correlation coefficient squared and to the fourth power, respectively, corrected for shrinkage according to the standard formula developed by Wheary. The cross-validated multiple R is usually lower than the sample multiple R, the magnitude of the difference (or shrinkage) depending mostly on the predictive power of the variables and the ratio of sample size to number of variables. Thus, it is generally advisable that an appropriate sample size and number of predictor variables be determined before beginning a study.[5] Unfortunately, this is often not done.[6-8] In general, if the ratio of subjects to predictors is low, the population estimate of validity ($\hat{R}$) will be markedly lower than that for the original sample (R). However, because of the greater predictor power of its independent variables, a regression equation developed within a small study with a low subject-predictor ratio might be more valid and generalizable than one developed within a larger study. To demonstrate, let us assume two regression equations obtained in order to predict systolic blood pressure. The first equation developed on a study using 25 subjects and two predictor variables yielded a multiple R^2 of .35. The second equation yielded a multiple R^2 of .30 from a study of 200 subjects and 15 predictor

variables. Using formula IV, the cross-validated multiple $\hat{R}^2$s for the equations derived from the large and the small studies would be .26 and .20, respectively.

STATISTICAL SIGNIFICANCE AND CLINICAL IMPORTANCE

In the last example, the equation developed from the smaller study would be clearly preferable to that developed from the larger study because (a) it is more predictive within the sample from which it derives, and (b) it is calculated to yield more accurate predictions in new samples from the parent population. Such preference would be appropriate despite the statistical significance of the regression equation obtained from the large study being much greater ($P = .0000001$) than that obtained from the smaller study ($P = .01$). This difference in P values reflects more the size of the two samples than the validity of the equations or the power of their respective predictor variables. The level of statistical significance is often mistakenly used to judge the relative merit of a study,[8,29] with the more statistically significant studies judged to be better. Once a study is found to be statistically significant, judgment of its merit should be based on the clinical significance of its findings. In the previous example, the clinical significance of the two equations can be ascertained by their validity for the original sample (R or R^2 values), and their generalizability to the parent population ($\hat{R}$ or $\hat{R}^2$ values).

DECLINING VALIDITY OF REGRESSION FORMULAS

A regression equation with good prediction of the dependent variable in the original sample retains its predictive value for new samples derived from the parent population so long as the dependent and independent variables remain unchanged. However, this is often not the case. Under such circumstances, optimizing existing regression equations may require adjustments ranging from simple recomputation of the formula to the generation of new regression formulas, depending upon the degree to which the dependent and independent variables have changed since the formula was generated. A new formula would be indicated, for example, to predict prognosis for patients whose survival has so markedly improved over a period of time that factors formerly known to influence survival show declining prognostic significance as previously unsuspected risk factors emerge.[30] Reevaluation of predictive formulas, like reassessment of medical knowledge, is dictated by the large and constant influx of new data bearing on our current understanding and treatment of disease processes.

CONCLUSIONS

This paper has attempted to acquaint the reader with some of the terminology of regression analysis and to show how regression formulas can be used. Such formulas and computer-assisted programs are increasingly being proposed in the practice of clinical medicine to aid in diagnosis,[31] to prescribe proper treatment,[32] and to predict probable outcome.[30,33] As the field of medicine becomes more complex, and with the increasing use of computer programs that aid the clinician in patient management, regression analysis will be found to be extremely useful for generating the necessary formulas for such computer aids. These formulas, if properly generated and applied, will prove to be a boon to scientific medical practice. Although numerous studies have suggested that such actuarial predictions are nearly always superior to intuitive predication,[34] the physician will still need to be the final judge of the utility and applicability of such formulas. There is still no substitute for useful experience and sound judgment.[35]

References

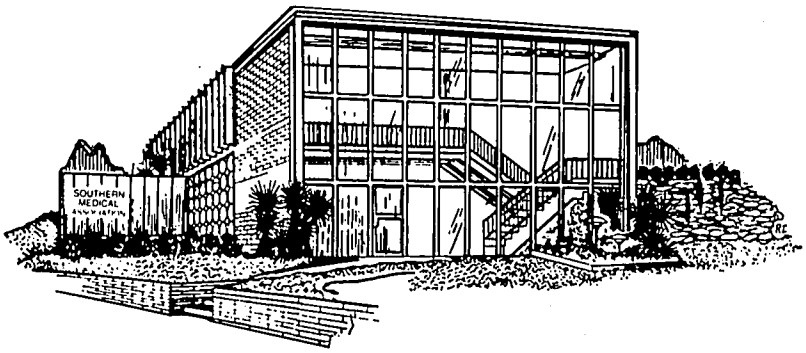
1. Altman DG: Statistics and ethics in medical research. collecting and screening data. *Br Med J* 281:1399-1401, 1980
2. Altman DG: Statistics and ethics in medical research. V. Analysing data. *Br Med J* 281:1473-1475, 1980
3. Altman DG: Statistics and ethics in medical research. VI. Presentation of results. *Br Med J* 281:1542-1544, 1980
4. Altman DG: Statistics and ethics in medical research. misuse of statistics is unethical. *Br Med J* 281:1182-1184, 1980
5. Schor S: Statistical proof in inconclusive "negative" trials. *Arch Intern Med* 141:1263-1264, 1981
6. Sibor S, Karten I: Statistical evaluation of medical journal manuscripts. *JAMA* 195:1123-1264, 1981
7. Gore S, Jones I, Rytter E: Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *Br Med J* 1:85-87, 1977
8. Reed JR, Slaichert W: Statistical proof in inconclusive "negative" trials. *Arch Intern Med* 141:1307-1310, 1981
9. Draper NR, Smith H: *Applied Regression Analysis.* New York, John Wiley & Sons Inc, 1966
10. Bock RD: *Multivariate Statistical Methods in Behavioral Research.* New York, McGraw-Hill Book Co, 1971
11. Press SJ, Wilson S: Choosing between logistic regression and discriminant analysis. *J Am Statist Assoc* 364:699-705, 1978
12. McNeil BJ, Hanley JA: Statistical approaches to clinical predictions. *N Engl J Med* 304:1292-1294, 1981
13. Hays WL, Winkler RL: *Statistics: Probability, Inference and Decision.* New York, Holt Rinehart & Winston, 1970
14. Kerlinger EN, Pedhazur ED: *Multiple Regression in Behavioral Research.* New York, Holt, Rinehart & Winston, 1973
15. Bobko P, Schemmer FM: Note on standardized regression estimators. *Psychol Bull* 88:233-236, 1980
16. Price B: Ridge regression: application to nonexperimental data. *Psychol Bull* 84:759-760, 1977
17. Rozeboom WW: Ridge regression: bonanza or beguilement. *Psychol Bull* 86:242-249, 1979
18. Gray HL, Schucany WR: *The Generalized Jack-knife Statistics.* New York, M. Dekker Inc, 1972
19. Wainer H: Estimating coefficients in linear models: it don't make no nevermind. *Psychol Bull* 83:213-217, 1976
20. Laughlin JE: Comment on "Estimating coefficients in linear models: it don't make no nevermind." *Psychol Bull* 85:247-253, 1978
21. Pruzek RM, Frederick BC: Weighting predictors in linear models: alternatives to least squares and limitation of equal weights. *Psychol Bull* 85:254-266, 1978
22. Wainer H: On the sensitivity of regression and regressors. *Psychol Bull* 85:267-273, 1978
23. Klecka W: *Discriminant Analysis.* Beverly Hills, Sage Publ, 1980
24. Cox DR: *The Analysis of Binary Data.* London, Methuen & Co, 1970
25. Cattin P: Note on the estimation of the squared cross-validated multiple correlation of a regression model. *Psychol Bull* 87:63-65, 1980
26. Rozeboom WW: Estimation of cross-validation multiple correlation: a clarification. *Psychol Bull* 88:1348-1351, 1978
27. Browne MW: Predictive validity of a linear regression equation. *Br J Math Stat Psychol* 28:79-87, 1975
28. Lucke JF, Whitely SE: Assessing squared validity: The estimation of multiple determination and cross-validity in multivariate normal random predictor regression. Unpublished technical report, 1981
29. Castelli WP, Gordon T, Hjortland M, et al: Alcohol and blood lipids: the cooperative lipoprotein phenotyping study. *Lancet* 2:153-155, 1977

Continued on page 729

8. Loda FA: Day care. *Pediatr Rev* 1:277, 1980
9. Scurletis TD, Peters AD, Robie WA: Attitudes of pediatricians toward day care. *Pediatrics* 38:44, 1966
10. Chang A, O'Neill R: Pediatricians and day care programs. *Pediatrics* 64:389, 1979
11. *Physician characteristics and distribution in the US, 1981*. Chicago: American Medical Association, Department of Data Release Services, 1982
12. *AMA Physician Masterfile for 1979*. Chicago, American Medical Association Department of Data Release Services, 1982
13. The Gallup Report. (Report 203.) Princeton, NJ, August 1982

Continued from page 725

30. Faguet GB, Davis HC: Survival in Hodgkin's disease. the role of immunocompetence and of conventional risk factors. *Blood* 59:938-945, 1982
31. Miller RA, People HE Jr, Myers JD: Internist-I, and experimental computer-based diagnostic consultant for general internal medicine. *N Engl J Med* 307:468-476, 1982
32. Yu VL, Fagan LM, Wraith SM, et al: Antimicrobial selection by computer. *JAMA* 242:1279-1282, 1979
33. Smith TL, Geham EA, Keating MJ, et al: Prediction of remission in adult acute leukemia. *Cancer* 50:466-472, 1982
34. Sawyer J: Measurement and prediction, clinical and statistical. *Psychol Bull* 66:178-200, 1977
35. Meehl PE: Clinical versus statistical prediction: theoretical analysis and a review of the evidence. Minneapolis, University of Minnesota Press, 1954



"The exclusive purpose of this Association shall be to develop and foster scientific medicine."