DRUG DEVELOPMENT AND INDUSTRIAL PHARMACY, 19(12), 1447-1459 (1993)

RELATIVE EFFICIENCY OF R^2 AND B^2 IN REGRESSION ANALYSIS FOR CALIBRATION AND FORMULATION

N. R. Bohidar
Philadelphia College of Pharmacy and Science
Villanova University

ABSTRACT

The assessment of the adequacy of a regression equation, as measured by the degree of closeness of the predicted values and their respective observed values is accomplished by the two contending statistics, R^2 and B^2. The derivation of the two statistics is presented and their relative performances are examined in the context of several pharmaceutics experiments involving, calibration, validation and formulation. The results strongly indicate that the B^2-statistic is much more sensitive and efficient than the R^2-statistic which has a tendency to inflate the magnitude irrespective of the data structure.

INTRODUCTION

Regression analysis plays a vital role in almost all pharmaceutics experiments, especially those associated with calibration (instruments, assay methods) and formulation development (optimization). Its greatest contribution is the formulation of a regression equation based on the estimated regression coefficients (intercept, slope), which are the functions of the independent variable (X) and the dependent variable (Y). The primary purpose of the regression equation is the prediction of one or more values of the dependent variable from one or more given values of the independent variable. Hence the success of a regression analysis is generally measured by the degree of overall closeness of the experimentally determined Y-values and their corresponding predicted values (Y^*). Therefore, the assessment of the adequacy of a regression function is directly linked to this composite measure of closeness. In the past, the quantity, r, the correlation coefficient, has been erroneously used for this purpose. Since r represents, the correlation between two dependent variables, the quantity is not defined in the context of a regression analysis which involves a dependent variable and one or more independent variables. Note that, one of the cardinal requirements of regression analysis is that, the X's are fixed and measured without error, and only the Y's are subjected to experimental error.

The two most appropriate statistics representing the composite measure of closeness between Y and Y^* are denoted by R^2 (the coefficient of determination) and B^2 (the coefficient of prediction). The primary purpose of this paper is to address the assessment of the adequacy of a regression function by exploring the merits and demerits of the two contending statistics noted above.

FORMULATION OF R^2 AND B^2 STATISTICS

Consider a pharmaceutics experiment (calibration or formulation optimization) involving K independent variables represented by X_1, X_2----X_k and let Y denote the dependent variable. Also let X, Y and B (without subscript) denote respectively, the matrix of the independent variables with (n x (k + 1)) elements, the vector of the dependent variable with n observations and the vector of regression coefficients with (k + 1) elements.

Regression Partition of Total Sum of Squares: (1,2,3)

The regression model involving the above three quantities can be expressed, in matrix notation, as, $Y = XB + E$, where E is the experimental error associated with the dependent variable. Now by applying the Gauss-Markoff least squares regression procedure for minimizing the error sum of squares ($E'E$), one has $Q = E'E = (Y - XB)'(Y - XB) = Y'Y - 2B'X'Y + B'X'XB$. Now, for minimization, $dQ/dB = -2X'Y + 2X'XB = 0$. Or, $X'XB = X'Y$, which are the normal equations associated with regression. Since in all pharmaceutics experiments, ($X'X$) has the property of non-singularity (full rank), an unique inverse exists and as such there is an unique solution of each coefficient in the equation, expressed as $B^* = (X'X)^{-1}X'Y$. Now one proceeds to partition the error sum of squares ($E'E$) into its constituent components, as follows:

$E^{*\prime}E^* = (Y - XB^*)'(Y - XB^*) = Y'Y - 2B^{*\prime}X'Y + B^{*\prime}X'XB^* = Y'Y - B^{*\prime}X'Y + B^{*\prime}[X'XB^* - X'Y]$. Now the expression in the brackets vanishes since $X'XB^* = X'Y$ as a consequence of the normal equations. Applying the correction for the mean and rearranging, one has the following,

$$(Y'Y - CF) = (B^{*\prime}X'Y - CF) + (Y - XB^*)'(Y - XB^*)$$

where, $CF = (\Sigma Y)^2/n$, $(Y'Y - CF)$ = Total sum of squares, $(B^{*\prime}X'Y - CF)$ = Regression sum of squares and the last term on the right hand side is the residual (error) sum of squares. The equation above constitutes the regression partition of the total sum of squares. Now dividing both sides of the equation by $(Y'Y - CF)$, one creates the following quantities,

$1.0 = (B^{*\prime}X'Y - CF)/(Y'Y - CF) + [(Y - XB^*)'(Y-XB^*)]/(Y'Y - CF)$. The first term on the right hand side is the quantity widely known as, R^2, where, $R^2 = (B^{*\prime}X'Y) - CF)/Y'Y - CF)$. This ratio represents the proportion of the total variation which is attributable to regression. The range of R^2 is from zero to one. Note that the fitted regression hyperplane is denoted by $Y^* = XB^*$ and the residuals from regression are denoted by $E^* = (Y - Y^*)$, in vector notation. As R^2 approaches the value of 1.0, E^* approaches the value of 0.0.

Regression Partition of Individual Observation: (1,2,3)

The above development accomplishes the partition of the total sum of squares into its constituent components. Now the interest is focused on accomplishing the partition of an individual observation into its appropriate segments measured by distances. Consider that a regression analysis of Y on X_1, X_2---X_k has been performed and a multidimensional regression diagram (graph) consisting of (a) fitted regression hyperplane, (b) $\bar{Y}$-hyperplane, (c) $\bar{X}$-hyperplane and (d) the K coordinates has been accomplished. (Note that the fitted regression hyperplane passes through the intersection of the $\bar{Y}$ and $\bar{X}$ hyperplanes). Let an observation Y_i with its K coordinates (X_1, X_2, ----X_k) be located in the K-dimensional space. If a perpendicular is drawn from that point parallel to the Y-axis on to the X-coordinate axis, that single perpendicular (vertical) line will intersect the following hyperplanes, (a) the fitted regression hyperplane, (b) the $\bar{Y}$-hyperplane and (c) naturally the X-coordinate axis (at the bottom of the diagram (graph)). The interest here is to express the distances between the above intersection points as a function of the Y-variable. Consider the following line-diagram which shows the total distance from Y_i and X_i and the two points of intersection denoted by Y^* and $\bar{Y}$. (This should be a vertical line, however, it is presented as a horizontal line for space considerations). Note that, all distances are measured from the point X_i.

Y_i Y^* $\bar{Y}$ X_i

Now the following distances are of interest (a) the distance between Y_i and X_i which is Y_i (the numerical value of the Y_i observation), (b) the distance between Y_i and Y^* which is $(Y_i - Y^*)$, deviation from regression, (c) the distance between Y^* and Y which is $(Y^* - \bar{Y})$, measuring the distance created by the inclination of the regression slope and (d) the distance between X_i and $\bar{Y}$ which is $\bar{Y}$. So the total distance now can be algebraically expressed as a sum of the constituent segments, of a single observation Y_i, as follows:

$$Y_i = \bar{Y} + (Y_i^* - \bar{Y}) + (Y_i - Y_i^*)$$

For statistical purposes, one has the following rearranged expression,

$$(Y_i - \bar{Y}) = (Y_i^* - \bar{Y}) + (Y_i - Y_i^*)$$

This shows that the total distance between the observation and the mean is segmented into two parts, (i) the distance between the fitted line and the mean, a separation caused by regression, and (ii) the distance between the observation and the fitted line, a separation caused by residual from regression. The above two expressions are true identities. Since the component segments and the total length are distances, the absolute value of the total must be equal to the sum of the absolute values of the two component segments, as follows:

$$(Y_i - \bar{Y})^{abs} = (Y_i^* - \bar{Y})^{abs} + (Y_i - Y_i^*)^{abs}$$

where $(\quad)^{abs}$ represents the absolute value of the expression (difference) inside the parentheses. Now summing over all the n observations (i = 1,2, ---n), one has, $\Sigma(Y_i - \bar{Y})^{abs} = \Sigma(Y_i^* - \bar{Y})^{abs} + \Sigma(Y_i - Y_i^*)^{abs}$

Since each term is positive, one divides both sides of the identity by $\Sigma(Y_i - \bar{Y})^{abs}$ with the following result,

$$1.0 = \Sigma(Y_i^*-\bar{Y})^{abs}/\Sigma(Y_i-\bar{Y})^{abs} + \Sigma(Y_i-Y_i^*)^{abs}/\Sigma(Y_i-\bar{Y})^{abs}$$

The statistical information contained in the two ratios on the right hand side of the above equation is of paramount importance in regression. The second term portrays the proportion of the total absolute mean deviations primarily attributable to the total absolute deviations from regression. As such, the proportion directly attributable to regression can be expressed as,

$$B^2 = 1.0 - [\Sigma(Y_i-Y_i^*)^{abs}/\Sigma(Y_i-\bar{Y})^{abs}] \text{ or,}$$

$B^2 = \Sigma(Y_i^*-\bar{Y})^{abs}/\Sigma(Y_i-\bar{Y})^{abs}$, depicting the relative contribution due to regression. This is the composite measure of closeness between Y and Y* without involving any sum of squares. The range of B^2 is always between zero and one, and as B^2 approaches the value of one, $E^*(= Y - Y^*)$ approaches the value of zero. (Note the distinction between B(vector of coefficients), B^*(vector of estimated coefficients) and B^2(coefficient of prediction)).

RELATIVE PERFORMANCE OF R^2 AND B^2 IN PHARMACEUTICS EXPERIMENTS: RESULTS AND DISCUSSION

The purpose of this section is to examine the merits and demerits of the two statistics, R^2 (coefficient of determination) and B^2(coefficient of prediction), estimated from the same experimental data. The results of the regression analysis are presented in self-explanatory tables, and because of the nature of the topic, the interpretation will be confined only to the computed values of the R^2 and B^2 statistics.

Consider an experiment involving the calibration of a new assay method (here, gravimetric) for determining calcium in the presence of large amount of magnesium (4,5). In this study 10 different samples containing, TABLE-I-A

#	X	Y	Y^*	$(Y-Y^*)$	$(Y-\bar{Y})$
1	20.0	19.8	19.838	-0.038	-11.21
2	22.5	22.8	22.354	0.446	- 8.21
3	25.0	24.5	24.870	-0.370	- 6.51
4	28.5	27.3	28.393	-1.093	- 3.71
5	31.0	31.0	30.909	0.091	- 0.01
6	33.5	35.0	33.426	1.574	3.99
7	35.5	35.1	35.439	-0.339	4.09
8	37.0	37.1	36.948	0.152	6.09
9	38.0	38.5	37.955	0.545	7.49
10	40.0	39.0	39.968	-0.968	7.99

TABLE-I-B

1	20.0	19.8	19.828	-0.028	-10.028
2	22.5	22.8	22.365	0.435	- 7.028
3	25.0	24.5	24.901	-0.401	- 5.328
5	31.0	31.0	30.988	0.012	1.171
7	35.5	35.1	35.553	-0.453	5.271
8	37.0	37.1	37.075	0.025	7.271
9	38.0	38.5	38.090	0.410	8.671

TABLE-I-C

1	20.0	19.8	19.801	-0.0013	-9.500
5	31.0	31.0	30.996	0.0038	1.700
8	37.0	37.1	37.102	-0.0025	7.800

by design, known amounts of CaO are analyzed by the new method. The laboratory assay value is considered as the dependent Y-variable and the true composition of the sample is considered as the independent X-variable. A linear regression analysis is conducted and the results are presented in TABLE-I-A,-B and -C. The intent of the analysis is to determine the degree of closeness of Y^*-values estimated by the regression equation ($Y^* = -0.2927 + 1.0065X$) and their respective observed Y-values, based on the R^2 and B^2 statistics. The computation of the two statistics consists of: (i) $R^2 = (B^{*\prime}X'Y - CF)/(Y'Y - CF) = B^{*2}\Sigma(X-\bar{X})^2/\Sigma(Y-\bar{Y})^2 = 433.4967/438.8887 = 0.9877$ and (ii) $B^2 = 1.0-[\Sigma(Y-Y^*)^{abs}/\Sigma(Y - \bar{Y})^{abs}] = 1.0 - (5.6153/59.30) = 0.9053$. The magnitude of R^2 above conveys the impression that the regression equation fits the observed data perfectly ($R^2 = 0.99$, rounded). Whereas, the B^2-value, which is lower than the R^2-value indicates clearly that the fit is not perfect because there are some data points for which the $(Y - Y^*)$ values are much higher than that for the other data points (See TABLE-I-A). A cursory examination of the $(Y - Y^*)$ column shows that, for each of the data points #4, #6 and #10, there is approximately one mg. discrepancy in each of their predicted values. This is exactly what is reflected in the magnitude of B^2. For further comparison between the two statistics, a linear regression analysis is conducted without these three data points, and the results are presented in TABLE-I-B. Now, R^2 is equal to 0.9978, giving again the impression that the regression equation fits the data perfectly ($R^2 = 1.0$, rounded). However, B^2 (=0.9606) clearly indicates that it is not necessarily so. Just to examine the consequences of eliminating those data points whose $(Y-Y^*)$ values are either equal to or above 0.4 mg. (#2,3,7 and 9), TABLE-I-C shows the regression analysis of the remaining three data points, #1, #5 and #8. The R^2-value is 1.000 and the B^2-value is 0.9996. It is clearly demonstrated in these tables that B^2 is extremely sensitive to the true nature of the data structure and to the magnitude of the difference between Y and Y^*. This is evidenced by the fact that, it changed from 0.9053 to 0.9606 and to 0.9996, as the difference between Y and Y^* decreased. However, R^2 tends to inflate the magnitude and remains unchanged. These analysis are conducted primarily for examining the relative efficiency of R^2 and B^2. This is not recommended for regular routine regression analysis.

This experiment involves the validation of an HPLC assay method for Product-D with nine selected concentrations considered as the independent variable (X) and the measured area under the chromatographic peak for each concentration considered as the dependent variable (Y). A linear regression analysis is undertaken to examine the relative efficiency of the R^2 and B^2 statistics. It should be noted here that, in this laboratory, one of the strict requirements for an assay method to be considered valid, is that the R^2-value must not be less than 0.9999. The interest here is to demonstrate that a regression equation with a R^2-value of as high as 0.9999 may not necessarily provide a perfect fit of the observed data and may not necessarily provide appreciable closeness between Y and Y^* for all data points. The results of the analysis are presented in TABLE-II. Here the R^2-value is equal to 0.9999 and the B^2-value is equal to 0.9889, based on the regression equation, $Y^* = 60.922 + 33910.87X$.

TABLE-II

#	X	Y	Y^*	$(Y-Y^*)$	$(Y-\bar{Y})$
1	0.02	707.05	739.14	-32.090	-2820.32
2	0.04	1404.35	1417.36	-13.007	-2123.02
3	0.06	2107.58	2095.57	12.006	-1419.79
4	0.08	2786.77	2773.79	12.978	- 740.60
5	0.10	3469.55	3452.01	17.541	- 57.82
6	0.12	4166.49	4130.23	36.263	639.13
7	0.14	4807.49	4808.44	- 0.954	1280.13
8	0.16	5487.19	5486.66	0.529	1959.83
9	0.20	6809.83	6843.10	-33.266	3282.46

Since the percent difference between Y and Y^* for data point #1 is 4.5% and for the others it is less than 1%, the next regression analysis is conducted without the data point #1. Now the R^2-value is 0.9999 and the B^2-value is 0.9908, based on the regression equation, $Y^* = 81.67 + 33762.08X$. Since the percent difference between Y and Y^* for data point #2 is 2% and for the others it is less than 1%, the next regression analysis is conducted without the data point #2. Now the R^2-value is 0.9999 and the B^2-value is 0.9912, based on the regression equation, $Y^* = 105.94 + 33596.84X$. Since the percent difference between Y and Y^* for data point #6 is 0.7% and for the others it is less than 0.3%, the next regression analysis is accomplished without the data point #6. Now the R^2-value is 1.0000 and the B^2-value is 0.9950, based on the regression equation, $Y^* = 100.27 + 33603.77X$. Without resorting to further analysis, it is clearly demonstrated in these several results that (a) a R^2-value of 0.9999 does not necessarily guarantee that the regression equation fits the observed data perfectly (See column $(Y-Y^*)$), (b) the R^2-value merely imparts an impression that the equation fits perfectly, which is not necessarily true, (c) R^2-value, indeed has a tendency to inflate the true magnitude, (d) B^2-value is extremely sensitive to even the modest changes in the $(Y-Y^*)$ values, and, (d) in these analyses, the B^2-value attained the magnitudes of 0.9889, 0.9908, 0.9912 and 0.9950, reflecting appropriately the changes in the data structure, whereas, R^2-value remained unchanged at 0.9999, indicating insensitivity to structural changes in the data.

In a formulation experiment, it is proposed to predict the disintegration times in minutes based on six selected physical as well as chemical factors using a multiple regression analysis with six independent variables. However, there are only 9 observations (9 rows and 6 columns) available for the analysis, leaving a residual degrees of freedom (DF) of (n-k-1) 2DF (n = no. of observations and k = no. of independent variables) with not enough DF left for a reliable estimate of the standard deviation. This is known as the saturation model, in which the R^2-value invariably, automatically gets inflated irrespective of the data structure and of the degree of relationships among the variables. The interest here is to show that even in this case, the B^2-value shows sensitivity to the structure of the data and does not get inflated automatically. The observed Y-values are 6, 2, 1, 5, 4, 9, 3, 7 and 8, and their corresponding Y^*-values are, 6.247, 2.171, 0.571, 5.940, 3.649, 8.717, 3.510, 6.325 and 7.869, based on the multiple regression equation, $Y^* = -1.367 + 0.356X_1 + 0.295X_2 + 0.988X_3 + 0.363X_4 - 0.103X_5 - 0.166X_6$. The R^2-value is 0.9651 and the B^2-value is 0.8132 with at least 3 data points showing appreciable discrepancies.

In a content uniformity experiment, it is proposed to conduct a regression analysis of content uniformity of tablets (Y) and their respective weights (X) with 12 available data points. Since each variable is required to be as close to its specification value as experimentally possible, the attainment of a high R^2-value is not in the realm of possibility. The interest here is to demonstrate that if a high or a low data point is inserted (as an artifice) into a set of data points which appears more or less as a circular cluster on a graph, the R^2-value of such a configuration gets automatically inflated without regard to the structure of the data. However, under the similar condition, the B^2-value remains at a reasonable level depending upon the data structure and the degree of relationships among the variables. In this study, R^2-value is 0.0236 and the B^2-value is 0.006, indicating strongly that there is no trend in the data, and that the structure of the data is a circular cluster with no perceptible relationship between the two variables. To study the effect of inserting (artificially) a single high point on the regression analysis, a simulated data point is created by adding 6.5 units to the highest weight value and 4.3 units to the highest content uniformity value. The regression analysis with the simulated data point included, produced a R^2-value of 0.8846 (0.9, rounded) and a B^2-value of 0.5209 (0.5, rounded). The R^2-value soars up automatically and the B^2-value strongly indicates that the prediction efficiency of the regression equation, here, is highly questionable.

Each of the four experiments mentioned above depicts a specific pharmaceutic activity, such as, calibration, validation, formulation and confirmation, and demonstrates the vital role played by regression analysis. The analysis primarily focuses on the assessment of the adequacy of the regression equation by comparing and contrasting the two contenders, R^2 and B^2. The results strongly show that the B^2-statistic is sensitive to the distribution of the data structure as well as the inherent relationships among the variables considered. R^2-statistic, on the other hand, has a tendency to inflate the magnitude irrespective of the data structure and the existing relationships among the variables. Presently, both statistics should be computed for the purpose of comparison and for appropriate pharmaceutics decisions.

ACKNOWLEDGEMENT

Grateful thanks are due to Mrs. Barbara J. Tomlinson for her meticulous efforts in accomplishing the word-processing task with utmost rapidity.

REFERENCES

1. N. R. Bohidar and K. E. Peace, Pharmaceutical Formulation Development. Chapter IV. "Biopharmaceutical Statistics in Drug Development." Marcel Dekker, Inc. New York, N.Y. 149-229 (1988)
2. N. R. Bohidar, Pharmaceutical Formulation Optimization Using SAS. Drug Development and Industrial Pharmacy, Vol. 17, No.3, 421-441 (1991).
3. N. R. Bohidar, Application of Optimization Techniques in Pharmaceutical Formulation-An Overview. Proceedings of the American Statistical Association. Biopharmaceutical Section. 6-13 (1984).
4. A. H. Bowker and G. J. Lieberman, "Engineering Statistics" Prentice-Hall, Inc. N.J. 241 (1961).
5. N. R. Bohidar, Statistical Aspects of Chemical Assay Validation. Proceedings of the American Statistical Association. Biopharmaceutical Section. 57-62 (1983).