MULTIPLE REGRESSION AND HISTORICAL EXPLANATION

Frank W. Carlborg, Northern Illinois University

The problem which led to this paper was suggested to me by Stanley Parsons of the history department of the University of Missouri at Kansas City. His interest was in describing the Populists in Nebraska in the latter part of the last century. For this purpose, he analysed the elections of the time. Since the 88 counties of Nebraska varied widely in their support of the Populist movement, one might hope to explain the movement by comparing the county-by-county vote with other variables. The method of multiple regression suggests itself in such a situation. This paper is concerned with some of the questions that arise after a satisfactory function has been fitted to the data in the usual least-squares manner. What follows may be divided into three general parts:

A. The regression function for 1890 and the questions raised by it.
B. The general derivation of a procedure for answering those questions.
C. The application of the procedure to the election of 1890.

Part A on the handout gives some of the usual results from a multiple regression analysis. The dependent variable Y is the Populist vote in a county as a percent of the total vote. The six independent variables, the x's, were chosen on historical grounds. They include such things as the percent of farm income in each county paid in interest charges and such as the percent of the population in each county of Protestant cultural background. The quadratic regression function of display (A1) was fitted to the data with the results of display (A2). Before settling on any function in a problem such as this, one must decide when to stop looking for more independent variables; also, one must decide when to stop adding on terms in the equation. These are very important problems which to my knowledge are unsolved. This paper has nothing to say about either of these questions. Rather, I would like to assume that a satisfactory function has been found. Therefore, let me assume that the equation on the handout fits the election of 1890 to a degree that is acceptable to the historian and his critics.

The original means, variances, and standard deviations of all seven variables are given in Table 1. The correlation matrix for the six independent variables is below Table 1. In the regression function of display (A2), the independent variables have been transformed so that each has mean of zero and standard deviation of one; then, the correlation matrix is also the covariance matrix of the transformed variables. Under the assumption that the fit of the function is satisfactory, these figures now summarize the election of 1890. To the historian, however, this summary is probably not satisfactory; he requires further interpretation. One thing he may like is an objective measure of the importance of each independent variable. The remainder of this paper is concerned with developing a procedure for measuring the importance of independent variables. This measure will be a non-negative function of the betas and the marginal distribution. To motivate the desirability of such a measure, consider again the numerical example. If one were to interpret the results, he might look at the coefficients for the linear terms first. β_6 is 8.24, and this suggests that x_6 is important. β_1 is next largest here. Among the coefficients of the quadratic terms, β_{22}, β_{66}, and β_{44} are the large ones. Several of the coefficients of the cross-product terms appear large--particularly, those involving the variable x_5. With one or two exceptions, the pairwise correlations among the x_i's are near zero in this example, but how should these correlations be considered, if at all, in evaluating the importance of a variable? These observations suggest that variable x_6 is important. Other than that, it seems difficult to make a very definite statement. As mentioned before, this paper will suggest a procedure for this situation. Some rather arbitrary steps will be taken. I would like to point them out.

To begin the general derivation, consider the simplest possible situation which is given in display (B1)--that is, the linear regression function with only one independent variable. To the historian, the magnitude of β_0 is irrelevant because he is interested in explaining variations in Y. The absolute value of β_1 suggests more information. The larger this absolute value is; the more important the variable is. Of course, the value of β_1 may be changed at will by performing scale transformations on x_1. The measure of importance should be independent of such transformations. A reasonable way out of this is to multiply the absolute value of β_1 by the standard deviation of x_1. Henceforth, it will be assumed that all independent variables have been transformed to unit standard deviations. Also, the results are more simply stated if it is assumed that the independent variables have a mean of zero. The formulas will be derived for this standardized situation.

They are easily generalized to other situations. This argument has now led to the conclusion that the absolute value of β_1 is a reasonable way to measure the importance of x_1 in this very simple situation. Note that the absolute value of β_1 is also the absolute value of the slope of the regression function with respect to x_1. Thus, the importance of x_1 is measured by the amount of change in the dependent variable which one would expect with a unit change in the independent variable. This way of stating the result seems, to me at least, to be one possible way to evaluate an independent variable in this historical situation. It also provides a base for generalization. The intention here is to proceed from this base. It is certainly an arbitrary decision and is possibly unsatisfactory to some viewpoints. The remainder of this paper is concerned with extending this idea to more general regression functions and with its application to the Populist example.

Consider next the regression function of display (B2). This is a quadratic function with still only one independent variable. The problem is to develop a measure of importance of x_1 for this situation; the result should be consistent with the previous result when β_{11} is zero. Toward this end, consider the slope at any value of x_1 as in display (B3). The prior, linear case suggested the absolute value of this as a measure. However, this absolute value is itself a function of x_1. An overall measure might be taken as the average of the absolute slopes with respect to the marginal distribution of occurrences of x_1. In order to actually perform this calculation, one must return to each actual data-point and compute the slope. Consider instead the following slight variation: average the square of the slope with respect to the marginal distribution of x_1. Display (B4) gives this result. The square root of this is an approximation to the average absolute slope. Note that the calculation of this quantity does not require returning to the actual data-points. Working with the square of the slope is more convenient in several other ways, and this will be done in the future. Investigations have revealed that this is not an expensive convenience.

For the next case, take the regression function of display (A1) and let the problem be that of measuring the importance of any one independent variable, say x_k. At any data-point, the partial derivative of the regression function with respect to x_k is the appropriate slope, and it is given in (B5). This, then, gives the rate at which one expects changes in the dependent variable for small changes in x_k. The average of the square of this with respect to the distribution of data-points appears in (B6). Note that this is a quadratic form in the betas of the regression function. The square root of (B6) is an approximation to the average absolute slope, and it is the desired measure of importance. The quadratic regression function in (A1) is the most general to be considered here. The generalization of (B6) to higher degree polynomials raises no new problems.

The next step in the generalization is to consider the problem of measuring the historical importance of a pair of independent variables. It would be very convenient if the importance of a pair of independent variables were a simple function of their individual measures of importance. If this were not the case, then separate calculations would be required for each pair, each triple, etc. To start the derivation, display (B7) gives a linear regression function with two independent variables, x_1 and x_2. The measure of importance of this pair should be consistent with what has been done above, and the result should reduce to a previous result in degenerate cases. A generalized concept of the slope will be used to generate the needed measure. For a unit change in x_1 and x_2, what is the corresponding change in the expected value of Y? The answer depends upon the direction of the unit change. If Δx_1 and Δx_2 are the changes in x_1 and x_2, a unit change corresponds to the condition (B8). One possibility is to take the unit change in the direction leading to the greatest change in the expected value of the dependent variable. This change is given by (B9). This is the generalized slope previously mentioned. Note that the square of (B9) is the sum of the squares of the separate slopes for linear regression functions, and note that relation (B10) holds for this case. This derivation is really much more general than the linear case. For any regression function with any number of variables, the above steps may be retraced and suitably modified. The chief modification is that β_i of the linear model is replaced by the partial derivative of the regression function with respect to x_i. The details are omitted here. The major result is that these measures of importance combine like orthogonal vectors under addition. That is, relation (B10) is always true. Triples of independent variables are similarly handled.

Item C on the handout gives the numerical results of applying this measure of importance to the Populist example. Table 2 shows the six variables in the first column; the second column gives the individual importances; the third gives the squares of the individual importances. Note that x_2 and x_6 are the two most important variables. Although the measure of x_2 is about twice that for x_1, it should be remarked that since these measures add like vectors, it would take four independent variables like x_1 to equal the importance of x_2. Next look at the importance of the pair x_2 and x_6. This is found by adding 191 and 138 and taking the square root of the sum. This number is 18. Also look at the importance of all six taken together. This is found by taking the square root of the sum of the six numbers in the last column. The ratio of the former for x_2 and x_6 to the latter for all six is .80. This shows that x_2 and x_6 account for 80 percent of what is told by all six variables. This suggests to the historian that he should look at these two variables in further detail. This has turned out to be a good numerical example because so much is suggested about the election by looking at only these two variables. Table 3 shows the result of classifying the original data according to only these two variables. The high quadratic and cross-product betas for x_2 and x_6 suggest displaying the data at three levels of each variable. The nine numbers in parentheses are the actual frequencies. That is, the "7" in parentheses indicates that seven of the 88 counties ranked in the low third according to x_2 and in the high third according to x_6. The other nine numbers are the average vote of the counties in each category. That is, the "55.8" indicates that the seven counties had an average Populist vote of 55.8 percent. There were three major parties in this election. This table shows great variations in Populist vote among the various levels of x_2 and x_6. No historical interpretation of this will be attempted here. This table is probably the major result of the study. It is important to remark that Table 3 does not depend in any way upon the assumptions made in the analysis of Part B above. The analysis has only suggested where to look to find something interesting. The table was constructed directly from the raw data.

To summarize, this paper has presented a procedure to measure the importance of independent variables in a multiple regression. This procedure is intended to be useful when the investigator's purpose is to explain a phenomenon, such as a historical event. The basic idea is to weight an independent variable according to the expected change in the dependent variable resulting from a change in the independent variable. The procedure would not be appropriate in many other regression situations--such as when the investigator's purpose is prediction or control of the dependent variable.

HANDOUT

Part A: Regression function for the election of 1890

(A1) $$E(Y) = f(x_1,\ldots,x_n) = \beta_0 + \Sigma\beta_i x_i + \Sigma\beta_{ij}x_i x_j + \Sigma\beta_{ii}x_i^2$$

	$\beta_0 = 42.70$	$\beta_{11} = -.42$	$\beta_{24} = -1.69$	$\beta_{44} = 1.68$
		$\beta_{12} = 2.20$	$\beta_{25} = -4.62$	$\beta_{45} = -4.64$
	$\beta_1 = 4.10$	$\beta_{13} = 1.08$	$\beta_{26} = -2.69$	$\beta_{46} = -1.99$
	$\beta_2 = 3.32$	$\beta_{14} = -4.70$		
(A2)	$\beta_3 = 3.00$	$\beta_{15} = 3.83$	$\beta_{33} = -.34$	$\beta_{55} = -1.08$
	$\beta_4 = -1.85$	$\beta_{16} = 1.27$	$\beta_{34} = 2.71$	$\beta_{56} = 5.66$
	$\beta_5 = 2.17$		$\beta_{35} = -1.42$	
	$\beta_6 = 8.24$	$\beta_{22} = -7.02$	$\beta_{36} = 2.86$	$\beta_{66} = 2.73$
		$\beta_{23} = -4.45$		

Table 1: Original Variables

Variable	Mean	Variance	Std. dev.
x_1	77.6	277.02	16.6
x_2	17.6	40.79	6.4
x_3	5.3	18.92	4.3
x_4	82.8	50.31	7.1
x_5	34.6	135.67	11.6
x_6	91.0	26.10	5.1
Y	37.3	285.94	16.9

The correlation matrix $[\rho_{ij}]$ of the marginal distribution of the x_i's is

$$\begin{bmatrix} 1.00 & .32 & .00 & .30 & -.03 & .22 \\ & 1.00 & .01 & .14 & -.47 & .10 \\ & & 1.00 & -.11 & -.14 & -.03 \\ & & & 1.00 & .04 & .66 \\ & & & & 1.00 & -.07 \\ & & & & & 1.00 \end{bmatrix}$$

The multiple correlation coefficient is .81

Part B: General derivation

Importance of one variable

(B1) Linear function: $E(Y) = f(x_1) = \beta_0 + \beta_1 x_1$

(B2) Quadratic function: $E(Y) = f(x_1) = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2$.

(B3) $f_1 = \frac{df}{dx_1} = \beta_1 + 2\beta_{11} x_1$

(B4) Average $(f_1^2) = \beta_1^2 + 4\beta_{11}^2$ with Average $(x_1) = 0$, Average $(x_1^2) = 1$

(B5) $f_k = \frac{\partial f}{\partial x_k} = \beta_k + 2\beta_{kk} x_k + \sum_{i \neq k} \beta_{ik} x_i$

(B6) General result for quadratic function:

Average$(f_k^2) = \beta_k^2 +$ Average$(2\beta_{kk} x_k + \Sigma\, \beta_{ik} x_i)^2$

where Average$(x_i) = 0$, Average$(x_i^2) = 1$, and Average$(x_i x_j) = \rho_{ij}$

Importance of more than one independent variable

(B7) $E(Y) = f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

(B8) $(\Delta x_1)^2 + (\Delta x_2)^2 - 1 = 0$

(B9) Importance of x_1 and x_2 = $\frac{\text{Maximal } \Delta E(Y)}{\text{subject to (B8)}} = (\beta_1^2 + \beta_2^2)^{1/2}$

(B10) [Importance of x_1 and $x_2]^2$ = [Importance of $x_1]^2$ + [Importance of $x_2]^2$

Part C: Application

Table 2: Measures of importance

Variable	Importance	[Importance]^2
x_1	6.5	43
x_2	13.8	191
x_3	6.9	47
x_4	7.3	53
x_5	6.8	47
x_6	11,8	139

Table 3: Vote as a function of x_2 and x_6

	low x_2	middle x_2	high x_2
high x_6	55.8 (7)	51.8 (10)	37.6 (12)
middle x_6	35.3 (9)	38.0 (11)	39.6 (9)
low x_6	25.1 (14)	26.5 (8)	33.3 (8)