

THE UNIVERSITY OF TEXAS AT AUSTIN

SCHOOL OF INFORMATION

MATHEMATICAL NOTES FOR LIS 397.1

INTRODUCTION TO RESEARCH IN

LIBRARY AND INFORMATION SCIENCE

Ronald E. Wyllys
Last revised: 2003 Jan 15

USING REGRESSION TO ESTIMATE AND PREDICT

What Is Regression?

The word "regression" has a technical meaning in statistics that is rather far removed from its everyday meaning. Though the connection could be traced historically, we shall not do so here. It is easiest simply to think of "regression" in statistics as a name for careful, informed prediction and estimation.

Our discussion of regression will concentrate on *linear* regression. Its central idea is to find the straight line that "best" fits a set of observed points, i.e., observed pairs of values of two variables. With observed data we often have problems of minor errors of measurement and with the effects of chance deviations from some underlying pattern. For example, height and weight data are subject to errors of measurement (and of reporting!) and are also affected by the fatness or thinness of the persons measured. Nevertheless, there is an underlying pattern of greater weight with greater height and vice versa.

One way of describing regression is to say that its purpose is to try to see through the obscuring haze of variation to the underlying pattern or trend exhibited by two variables. Once the existence of such a pattern for a pair of variables has been established and its nature determined, the pattern can be used to estimate, or predict, the probable value for one of the variables that corresponds to a particular value for the other variable. The ability to make this kind of prediction is especially useful when the *predicted* variable is difficult to observe or will take on its value in the future, whereas the *predictor* variable is easy to observe or can be observed with little or no delay.

Another way of describing regression is to say that it concerns the fitting of smoothed curves--for our purposes in LIS 397.1, the kind of smoothed curve known as the straight line--to points, i.e., to sets of data that consist of pairs of values. Smoothed curves have the virtue of filtering out minor fluctuations in the observed behavior of a pair of variables, yielding a clearer picture of the real behavior of the variables. You will often find regression referred to as "curve fitting" in textbooks and in calculator and computer manuals.

In general terms, regression starts with a set of observed pairs of values of two variables, measured on the elements in a sample drawn from some population having two variables in which you are interested. The variables can be labelled X and Y; the observed values of the variables can be called $X_1, X_2, X_3, \text{etc.}$, and $Y_1, Y_2, Y_3, \text{etc.}$; and the pairs can be denoted as $(X_1, Y_1), (X_2, Y_2), \text{etc.}$, or, in general, as (X_i, Y_i) , where i is an arbitrary index used simply to distinguish the various pairs. Such pairs can be plotted as points on a chart, usually with the values of the X_i s measured along the horizontal axis and the values of the Y_i s along the vertical axis.

For example, you could (theoretically, at least) measure the height and the weight of each woman graduate student at UT-Austin. Then, to the resulting scattered pairs of data, you could fit a straight line by the regression technique. This line would represent the average change in weight corresponding to a given change in height. Knowing a woman graduate student's height would then enable you to predict her weight. (Note: This statement holds if height was being used as the independent variable, in a sense that is discussed later. If weight was being used as the independent variable, then the line would represent the average change in height corresponding to a given change in weight.)

In a similar but more practical way, one can predict a student's graduate GPA from his or her GRE score, by using the regression-line equation derived from a sample of observed pairs of GRE scores and GPAs.

What linear regression does is to find the straight line that "best" fits any set of pairs of observed values. What "best fit" means is well defined: viz., given any set of pairs of observed values (i.e., points), the line of "best fit" to the points is that line whose position and orientation are such that the sum of the squares of the vertical distances of the points from that particular line is less than the analogous sum for any other possible line. (Note: Strictly speaking, these distances are measured parallel to the coordinate axis of the dependent variable. Since we are assuming throughout this discussion that the dependent variable is Y, whose axis is the vertical, we can speak of "vertical distances" here.)

The line with the property just described is known as the "least-squares fitted" line for the particular set of points, or the "regression line". Roughly speaking, the line of best fit is the line that comes *as close as possible simultaneously to all the actually observed points*.

If one of the variables is known to depend on the other, we can speak of the "independent" variable and of the "dependent" variable; on the other hand, if there is no particular reason to view one variable as "more independent" than the other, we can choose arbitrarily to consider one of them as the independent variable and the other as the dependent. The "dependence" *does not need to be*, and often *is not*, a real, causal relationship; we simply call "dependent" the variable whose values we want to *predict* on the basis of specified values for the other variable, the one we call the "independent". In these terms, we can call the regression line the "line of regression of the dependent variable on the independent variable". For convenience, we shall herein call X the independent variable and Y the dependent.

The primary goal in regression is to find out how to express the regression line in algebraic form. That is, we want to use the information in the sample to derive the algebraic equation of the line of best fit to the observed sample pairs (X_i, Y_i) . Once we have that equation, we can use it to give us the predicted value of the dependent variable that corresponds to any particular value of interest for the independent variable.

The Linear Regression Equation

The definition of the line of best fit in terms of minimizing a sum of squared distances may strike you as a very difficult definition to satisfy. Actually, with the aid of calculus it is easy to determine which line satisfies the definition, not only with respect to any particular set of pairs of observations, but also with respect to sets of data in general. It turns out that the general solution--for the case in which X is being considered the independent, and Y the dependent, variable--can be written in the form of the following equation:

Equation 1

$$\hat{Y} = B_0 + B_1 X$$

In writing this algebraic equation for the regression line, we use a circumflex accent, "^", over the symbol for the dependent variable, Y, to indicate that the value of Y is estimated, i.e., predicted. The combination is usually pronounced "Y hat." The regression-line equation is calculated by using the observed sample pairs to provide numerical values for B_0 and B_1 via the formulas that follow. (You would use these formulas if you were using a simple calculator--one lacking a built-in regression or trend-line function] to do regression; ordinarily nowadays, one does regression with the aid of a computer, and these formulas are used internally by the statistical program package.)

We begin by calculating the sample Pearson correlation coefficient, using data obtained from the sample of pairs of observations (X_i, Y_i)

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)s_X s_Y}$$

Then we easily calculate

$$B_0 = \bar{Y} - B_1 \bar{X}$$

and

$$B_1 = r_{XY} \left(\frac{s_Y}{s_X} \right)$$

These equations satisfy the definition, and they constitute one way in which you can calculate the regression line for a set of pairs of observations. Ordinarily, however, nowadays one finds a regression line by using either a computer program or an electronic calculator that incorporates a regression-analysis, or trend-line, function. The particular equation that results from inserting the particular numerical values B_0 and B_1 that have been determined from a particular set of pairs of observations will be the regression equation for those data; it will represent the regression line for those pairs of data when viewed as points.

When used in the regression equation, the numerical values of B_0 and B_1 will yield, for each of the originally observed values of the independent variable (viz., X_1, X_2, X_3 , etc.), the corresponding "predicted", or estimated, value of the dependent variable, Y . That the predicted values are labeled with a circumflex accent, " $\hat{}$ ", helps to distinguish them from the originally observed values of the dependent variable. Thus

$$\hat{Y}_1$$

is the *predicted* value of Y corresponding to X_1 , whereas Y_1 is the *observed* value of Y that corresponds to X_1 . The predicted values

$$\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \dots$$

all lie on the regression line.

Equation 1 can be put into several different forms. Let us look at one of them. Using the equations above for B_0 and B_1 , we can write

$$B_0 = \bar{Y} - B_1 \bar{X} = \bar{Y} - r_{XY} \left(\frac{s_Y}{s_X} \right) \bar{X}$$

Thus Equation 1 can be written as

$$\hat{Y}_i = \bar{Y} - r_{XY} \left(\frac{s_Y}{s_X} \right) \bar{X} + r_{XY} \left(\frac{s_Y}{s_X} \right) X_i$$

from which, rearranging terms, we obtain

Equation 2

$$\hat{Y} = \bar{Y} + r_{XY} \left(\frac{s_Y}{s_X} \right) (X_i - \bar{X})$$

Equation 2 shows that the predicted value of Y corresponding to any particular value of X that happens to be of interest can be computed as follows: To the mean value,

$$\bar{Y}$$

of the observed Y s we add a certain quantity. This added quantity starts with the correlation coefficient multiplied by the ratio of the *predicted* variable's SD to the *predictor* variable's SD. This product, in turn, is multiplied by the difference between the interesting particular value of X and the mean value of all the observed X s.

Two consequences of these facts are worth noting here. First, if the correlation coefficient, r_{XY} , is zero, then the predicted value of Y will be

$$\hat{Y} = \bar{Y}$$

no matter what value of X is used. In other words, if there is no correlation between X and Y , then knowing the value of X does not help you improve your estimate of the probable value of Y ; your best guess for Y is simply the mean value of all the Y s. Second, when

$$X = \bar{X}$$

the last parenthesis in Equation 2 will contain zero; hence, once again, the predicted value of Y will be

$$\hat{Y} = \bar{Y}$$

That is, the predicted value of Y corresponding to the mean value of X is just the mean value of Y . Another way of putting this is that the point whose coordinates are the mean values of X and Y , viz.,

$$(\bar{X}, \bar{Y})$$

lies on the regression line.

The Regression Coefficients

The coefficients B_0 and B_1 are known jointly as the "regression coefficients", since they define the regression line. Unfortunately, the singular form, "regression coefficient", is sometimes used to refer to B_1 alone. A better name for B_1 is "slope", since the value of B_1 equals the slope of the regression line. The slope of any line is the ratio of (i) the change in vertical position along a piece of the line to (ii) the change in horizontal position along that piece. That is,

$$\text{slope} = \frac{\text{vertical change}}{\text{horizontal change}}$$

The other coefficient, B_0 , is also known as the "Y-intercept", or just "intercept", since its value is the height at which the regression line intercepts (i.e., crosses) the vertical axis of the graph, the Y-axis. (Note that the Y-axis we are speaking of here is the basic vertical reference axis, i.e., the line that goes vertically through the origin of coordinates, the point (0,0). This axis is not necessarily the left margin of a particular graph, for the left margin is usually chosen to suit the particular data being displayed, and therefore may not include the point $X = 0$.)

To understand why B_0 is known as the "Y-intercept", observe from the regression line's equation

$$\hat{Y} = B_0 + B_1X$$

that if you put in zero as the value of X , then you get B_0 as the corresponding predicted value of Y . Since $X = 0$ for all the points of the vertical axis of any graph, what you just did was to show that the point on the regression line where it crosses the Y-axis is the point for which $X = 0$ and $Y = B_0$, i.e., the point (0, B_0).

Confidence Intervals for Predicted Values and for the Regression Coefficients

As with estimates of population means, it is possible to develop confidence intervals for the predicted values \hat{Y} and for the regression coefficients B_0 and B_1 . However, since such confidence intervals involve uncertainties associated with not just one but two variables, they are somewhat more complicated to develop and to state than are confidence intervals for population means. Since the goal of LIS 397.1 is merely to provide an introductory level of understanding of statistical techniques, in this note we ignore the development of confidence intervals for predicted values and for the regression coefficients.¹

When Is the Use of Regression Justified?

Strictly speaking, regression ought to be used only under the conditions that: (i) both X and Y are interval variables; (ii) both X and Y are Gaussianly distributed; and (iii) X and Y are known to be correlated (i.e., the population Pearson correlation coefficient of X and Y is known to be different from zero).

However, it must be admitted that regression is frequently used when these conditions have not been satisfied. Subjective judgment is often involved, especially when the investigator feels confident that X and Y really are correlated, even though the sample that he or she used was too small to prove the correlation beyond a reasonable doubt.

Similarly, people often use regression with time as the independent variable, even though time is clearly a non-Gaussian variable, simply because in their opinion the convenience of doing so outweighs the strictures.

Perhaps a reasonable middle ground is to compute the sample Pearson correlation coefficient, r_{XY} , and check to see whether it is large enough to permit the rejection of the null hypothesis that the population correlation coefficient is zero. If r_{XY} is large enough for rejection, you can proceed; if not, then before proceeding, you should consider carefully whether you feel confident that there really is a correlation between the variables because of other knowledge you have about the situation.

A frequently encountered pitfall of regression occurs when predictions are made outside the interval containing the observed values of X and Y . A moment's reflection should assure you that the phenomenon you are investigating could, and sometimes will, behave quite differently outside the interval of your observations from the way it behaves inside that interval. Hence, predictions outside the observed interval are dangerous. A common situation in which one wants to make such predictions is when the independent variable is time, and one wants to predict future values of the dependent variable. For practical reasons, such predictions are often made; just be wary of relying on them too heavily.

Other Kinds of Regression

In conclusion, we mention two extensions of linear regression that you should be aware of. First, we have been talking about fitting sets of observed data pairs with a straight line, the technique of *linear* regression. In the linear regression equation, the exponent of the independent variable X is one (since, by convention, the value of an exponent is understood to be one when the exponent is not explicitly written). By using terms containing the (single) independent variable with higher *exponents*, e.g., as in a regression equation like

$$\hat{Y} = B_0 + B_1 X + B_2 X^2 + B_3 X^3$$

it is possible to fit sets of observed points with smoothed approximations in the form of various types of curved lines. Not surprisingly, this is known as "curvilinear regression".

Second, by using *more than one independent variable*, e.g., as in a regression equation like

$$\hat{Y} = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3$$

(where the subscripts on the X s are used to distinguish different *variables* rather than different observed *values of a single* variable, as in curvilinear regression), it is possible to quantify the relative importance of several variables in producing the overall effect represented by the dependent variable. This is known as "multiple regression".

For example, an investigation into the management of academic libraries might make use of an equation like the one above, but with more terms. For example, an investigator might gather data from a sample of university libraries and represent the data for each library as follows: Y , size of the library's collection; X_1 , number of students at the university; X_2 , number of faculty members; X_3 , number of academic majors available at the undergraduate level; X_4 , number of fields in which master's degrees are offered by the university; and X_5 , number of fields in which doctoral degrees are offered. By observing the values of these six variables in a sample of universities and carrying out the multiple-regression process, the investigator could derive an improved understanding of the typical relations among the variables.

From the mathematical viewpoint, essentially what happens is that if some variable X_i is important in determining the value of the dependent variable Y , then its coefficient B_i will be a large (positive or negative) number, whereas if some other variable X_j is not important in determining Y , then its coefficient B_j will be close to zero.

From the practical standpoint, what can happen is this. Suppose the director of a university library carries out such a study on the libraries of several dozen universities of similar size to her own, and obtains a regression equation for Y ; i.e., she obtains the values of the coefficients B_i that result from the data in her sample. Suppose, further, that she then takes the values for the variables X_i for her own university and puts them into the regression equation; i.e., she multiplies the X_i values for her institution by the coefficients B_i yielded by her sample. Suppose, finally, that the result is a predicted collection size for her library that is substantially larger than its actual size. That would mean

that, in comparison with the libraries of other similar universities, her library's collection is badly undersized. Such an outcome would be an excellent basis for a campaign by her to win a substantial increase in her acquisitions budget over the next 5 years or so.

Both curvilinear and multiple regression are widely used. In part because computers have made it easy--so far as humans are concerned--to do the often horrendous arithmetic of multiple regression, this technique has become one of the favorite tools of researchers and managers who want to investigate which factors are important, and which are unimportant, in complicated situations.

Endnote

1. A source that provides a detailed treatment of confidence intervals for the predicted values in regression and for the regression coefficients is: Hays, William L. *Statistics*. 4th edition. Ft. Worth, TX: Holt, Rinehart and Winston; 1988. ISBN:0-03-002464-1. See especially pp. 568-573.