A General Overview of Regression Analysis

George H Olson

Appalachian State University

March 2014

Today we move into a study of what is agreeably the most widely used statistical procedure in the behavioral sciences: regression analysis, more precisely, Multiple Linear Regression (MLR) analysis. Actually, the more generic term is General Linear Model analysis (GLM). Virtually all the parametric statistical procedures we have examined so far in this course can be subsumed under the GLM. This includes the *t* test and the Analysis of Variance. It includes, also, Analysis of Covariance, which we will examine later in this course, and a wide variety of more advanced statistical procedures including multiple correlation, principal component analysis, factor analysis, discriminant analysis, hierarchal linear regression, and covariance structure analysis. Gaining a firm grasp of MLR (or the GLM) will facilitate your understanding of those more complex statistical procedures.

We begin with MLR, of which simple regression is a subset. In simple regression there is one *independent* variable and one *dependent* variable. More typically, we will have multiple *independent* variables, but still only one *dependent* variable (when there is more than one dependent variable the procedure is classified as *multivariate* multiple regression analysis, which would be the subject of a much more advanced course in statistics.)

**Simple Linear Regression**

In simple linear regression, we examine the functional relationship,

 (1)

that is, the relationship between the measured dependent variable, *y*, for individual *i* and individual’s score on the independent variable, *x*. In simple *linear* regression, the form of this function is assumed to be linear. Typically, we will write the function as,

 (2)

which you should all recognize as an equation for a straight line. The *coefficient*, *α*, is the intercept, and the coefficient, *β*, is the slope of the line.

The goal in regression analysis is to find reliable *estimates* of *α* and *β,*  *a* and *b*, say, which when substituted in (2) yields,

 (3)

where an additional term, *ei*, has been added to account for the fact that it is not likely (in fact it is highly unlikely) all the *y*i sit on the straight line. Using the estimates of the coefficients, we compute *estimated* (or *predicted*) values for *yi*which we identify as  (read as -hat sub *i*), where

 (4)

The technique is to find estimates *a* and *b* such that the sum of the squared differences,  is minimized. This is known as the *least squares criterion* in regression analysis. Note that the difference, , so that the least squares criterion is satisfied when the sum of the *ei*2 is minimized.

**Prediction and Explanation**

Regression analysis typically is used for two different purposes. It is often used to *predict* outcomes. This is how it is used in the EVASS system developed by the SAS Institute and used in North Carolina schools. Used this way, the *yi* are scores on some particular achievement test, e.g., End-of-Year science, and the *xi* (of which there usually are several for each individual) are scores (and other measures) taken on the individual students in previous years. In the simplest case, where only one *x* is used, the prediction equation,



is used to *predict* or *forecast* individuals’ score on an achievement test at the end of a given year. There is no intention of *explaining* or *accounting for* end of year’s test performance. If shoe size provided an adequate minimization of  then shoe size would be used in EVASS. The point is, when used for purposes of prediction, regression analysis can use any set of *predictor* variables. Of course, this does not mean that the prediction will necessary be accurate. A measure of the accuracy of prediction is given by a statistic called *multiple correlation squared* or *R*2, which gives the proportion of variance in the *criterion* variable (i.e., the *y*) predicted by the *predictor* variable(s) (i.e., the *x*s).

A more sophisticated use of regression analysis is its use in explaining how variance in variables of interest (e.g., achievement) can be accounted for by variance in independent variables (e.g., motivation, parental involvement, birth order, and so on.) Over the next few weeks we will investigate this use of regression analysis in some detail.

The *parametric* model for the regression of *Y* on *X* is given earlier by (2)



The model for the regression of *Y* on *X* in a *sample* was given in (3), earlier



Calculation of the constants in the model:

the *slope* (**) is given by

**, (4)

where **

,

and the intercept by

**. (5)

Using the coefficients, *a* and *b*, leads us to equation (4) for estimating (or *predicting*) an individual’s score on *Y*:



If we take a closer look at the regression equation, (5), using (3) and (4) leads us to

 (6)

**Partitioning the *Sum of Squares***, **

First, consider the following identity

. (7)

If we subtract from each side of the equation, we obtain

 (8)

After squaring and summing, we have

 (9)

or, after simplifying[[1]](#footnote-1),

 (10)

Where *SSreg* = regression sum of squares (the sum of squared deviations in *Y* *accounted for* or *explained by* the regression of *Y* on the *X*) , and *SSres* = residual sum of squares (or the sum of squares left *unexplained* by or *un-accounted for*).

Dividing (10) through by the total sum of squares, *SStot* (= ) gives

 (11)

or

 (12)

The first ratio in (12), gives the proportion of sum of squares in *Y* accounted for, or explained, by regression; the second ratio gives the proportion of sum of squares left unexplained. The proportion of sum of squares explained is denoted as the *coefficient of determination*, or 

**A computational example**

It is often useful to devise simple computational examples, such as the following:

|  |  |
| --- | --- |
| *Y*  3  1  0  4  5 | *X*  1  0  1  -1  2 |

The means of the two variables, *Y* and *X*, are



Having computed the means, we now compute the deviations (*x* and *y*), squares of deviations (*x2* and *y*), and cross-products of deviations (*xy*):

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Deviations, squares, and cross-products | | | | | | |
| *Y*  3  1  0  4  5 | *y*  .4  -1.6  -2.6  1.4  2.4 | *y2*  .16  2.56  6.76  1.96  5.76 | *X*  1  0  1  -1  2 | *x*  .4  -.6  -1.6  .4  1.4 | *x2*  .16  .36  2.56  .16  1.96 | *xy*  .16  .96  4.16  .56  3.36 |

The *sums* of squares and cross-products are computed as



and the regression coefficients as

*b = *1.769, and

*=* 2.6 – 1.769(.6) = 2.6 – 1.061 = 1.539.

The *regression equation* can now be written as

.

From an earlier equation (10) we obtain

****  (13)

Note that we could also have computed

 (14)

An alternative calculation is given by

 (15)

Hence,

 (16)

Recall that the equation for the Pearson correlation is

 (17)

Therefore, *SSres* can also be computed as

 (18)

Appendix

Showing the Simplification of the Partitioning of SS*Y*

Beginning with,



we need only show that



Recalling (4, 5 & 6) we can write,



1. See the appendix to see how the equation simplifies. [↑](#footnote-ref-1)