Multiple Regression Analysis: A Model for Mass Communications Research

* Syed H. Pasha

* Ph.D. in Mass Communication, University of Wisconsin at Madison, United States. – Associate Professor of Mass communication, King Saud University.
Abstract

Multiple regression analysis is a highly sophisticated and powerful statistical and methodological technique for investigating interrelationships among three or more variables especially where the variables have been conceptually classified into dependent and independent and where causality is suspected. Sound research, however, is not based on elegant methodology alone but requires good grounding in theory just as it must satisfy the requirements of reliability and validity of the research instrument. Multiple regression procedures are based on certain descriptive and inferential assumptions which must be met in order for the results to be valid. However, regression techniques are fairly robust and minor violations of underlying assumptions do not affect their outcomes. The article also discusses the procedures for calculating R²'s and F tests. In the end a number of studies from mass communication literature are cited in relation to their use of multiple regression techniques either for exploring interrelationships among variables or for calculating path coefficients in causal models.
Introduction

This article is an attempt to outline some basic ideas in the use of multiple regression analysis in mass communications research. Multiple regression is a powerful statistical technique for analyzing interval level data. It is indicated in cases where variables have been conceptually classified into dependent and independent. In regression analysis, the dependent and independent variables are also known as criterion and predictor variables respectively. In the multiple regression situation, there is one dependent variable, Y, which is regressed upon two or more independent variables, X's. Multiple regression analyzes and measures the extent to which two or more X's contribute to variation in Y both individually, that is one X at a time, as well as jointly, that is two or more X's interacting with one another. Multiple regression could be used for descriptive as well as for inferential purposes.¹

The present article seeks to throw light on the following questions: what is the role of theory in research? What are reliability and validity notions? How is a sample selected? What are some of the assumptions underlying multiple regression analysis? and finally, how are hypotheses tested using multiple regression models?

The model presented in this article is expected to be of interest and use to students of social scientific research in general, and to students of mass communications research in particular.

THEORY

Kerlinger cites the case of the student who on being asked what it was that he wanted to study, replied that he planned on using a rather sophisticated statistical technique known as Analysis of Variance (Kerlinger, 1973). In other words, his research goals were centered around this methodological "hammer". So long as he could use the hammer freely for pounding purposes, it mattered little to him what it was that received the pounding and why. What the student needed was a little bit of theory which could have given meaning and direction to his pounding. It is well said that there is nothing more practical than a good theory.²

A theory is essentially a nomological network, a system of interrelationships between concepts and variables. In the present article, the variable of interest is political and public affairs knowledge. The task ahead is to identify some of the key factors that would explain or predict the acquisition of knowledge. Positing knowledge as the dependent or criterion variable, Y, the article seeks to explore a set of independent or predictor variables, X's, whose impact on knowledge could be demonstrated in the context of a multiple regression model.

One way to do this is to use a dragnet or shotgun approach - a methodological fishing expedition, as it were throwing all known or suspected covariates of knowledge including the kitchen sink and the garbage pail, as some would say into a grand regression equation, and wait, while the machine, through a stepwise procedure
churns out a list of variables which share the most variance with knowledge. The other way – and methodologically speaking a more elegant and perhaps a more useful way in the long run is to agonize intellectually over the research problem, and attempt to conceptualize it clearly at the theoretical level before submitting it to the machine for necessary computations. This is the way followed in the model presented in this article.

Given the wide availability of the mass media, and their frequent use by most people, Media Exposure suggests itself as an interesting concept to pursue in the theoretical quest to locate useful predictors of Knowledge. A quick literature review confirms this view. Media Exposure and Knowledge are indeed positively correlated in several studies. But what does this concept Media Exposure really mean? A closer look at the “dictionary” part of the theory seems called for at this stage. The mass media spectrum extends over a very wide range. Even when it is narrowed down to something more specific like Television, the issue is not still in sharp enough focus. It might be more profitable to narrow the concept down further to Television News, a program dealing more directly with political and public affairs information. Thus, the concept of Media exposure would be modified to TV News Exposure. Even though problems remain with the explication of Exposure, they would not be gone into here.

Further, it seems logical to argue that it is not just what people are exposed to, but also why they choose to be exposed to it, that may be significant. It is not only the physical act of exposure, but also the mental and emotional set accompanying or preceding exposure, the motivations behind exposure conceptualized in terms of gratifications sought from and the orientations brought to the exposure situation, that may play a key role in predicting Knowledge. Thus, Surveillance using the mass media with a view to find out what is going on in the world emerges as another explanatory variable in relation to knowledge.

As a result of this mini-theorizing (as contrasted with Parson’s grand theorizing and Merton’s middle range theorizing) two distinct conceptual sources of influence on knowledge have emerged: TV News Exposure and Surveillance motivation. The purpose of this theoretical note, however, is not to build a systematic theory of knowledge, but to demonstrate the importance of a theoretical attack on the research question before subjecting it to a methodological “pounding”.

**METHODOLOGY**

There are three variables in the theoretical model outlined in the previous section, namely, Knowledge, TV News Exposure and Surveillance. Each of these three variables could be operationalized and measured on a scale about which intervality would be presumed. It must be noted, however, that no matter how carefully the social scientific and behavioral variables are operationalized and measured, questions remain about important issues such as the nature and distance of intervals, meaning and location of zero, and so on. To be aware of these inherent limitations of social research is a sobering influence on the researcher.
RELIABILITY

One of the first tasks facing a researcher is to establish the reliability of his instrument. Reliability is the accuracy or precision of measurement resulting in stability or consistency of results over time, between measurements or tests, and between coders or raters. Theoretically speaking, reliability is the proportion of true to total variance in an obtained set of scores: (Kerlinger, 1973:446)

\[ r_t = \frac{V_{oo}}{V_t} \]

However, since true variance is never known, the following formula may be more useful for the computation of reliability coefficients:

\[ r_t = 1 - \frac{V_e}{V_t} = \frac{V_t - V_e}{V_t} \]

There are several ways in which reliability of the measuring instrument could be established. For example, the same test or questionnaire may be administered to a group of people at two different points of time (test-retest situation); or the sample, if large enough, could be divided into two halves and each segment tested separately (equivalence situation); or the same test could be repeated with a number of different samples.

VALIDITY

Reliability is a necessary but not a sufficient condition in the research process. A researcher must demonstrate that his instrument no matter how reliable it is-is measuring what it was really intended to measure, that is, his is a valid instrument. Kerlinger (1973:469) defines validity as:

\[ V_{vt} = \frac{V_{co}}{V_t} \]

that is, the proportion of common factor variance-variance shared between measures-to the total variance of a measure. Smith considers reliability to constitute the upper limit of validity, meaning an instrument cannot be more valid than it is reliable. (Smith, 1975:57-79)

Validity could be achieved in a number of ways. Since validity deals mainly with the domain of content and meaning, sometimes a mere assertion or statement could do it. This will be a kind of validity by fiat. But a more useful and systematic approach is to use convergent and discriminant validation procedures. These procedures are especially helpful in the area of construct validation. Using a multi-trait multi-method matrix is another useful approach. Certain multivariate statistical techniques such as Factor Analysis and MDS could also be used for validating certain category schemes.
THE SAMPLE

A sample plays a crucial role in any research, both in terms of its size and the manner of its selection. Random sampling is a key requirement of all inferential statistics. Significance tests (e.g. F-Test) are based on the assumption of random sampling.⁹

The size of the sample is closely related to the question of power. Power is defined as 1 − α, or the probability of successfully rejecting the null hypothesis when it is false. Power is a function of α (that is power increases as α increases), postulated effect size in the population and the sample size, N.⁷

Some other considerations which play an important role in sample selection include the time and resources available to the researcher.

SOME ASSUMPTIONS UNDERLYING MULTIPLE REGRESSION ANALYSIS

Multiple regression is based on certain basic assumptions. Some of these assumptions are descriptive. These include the assumptions of intervality, linearity, homoscedasticity, non-autoregressive errors, etc. Other assumptions are inferential such as eᵢ ∼ N(0, Q) and Cov eᵢ,eⱼ = 0, i ≠ j.⁷

In general, regression techniques are robust to violation of these assumptions. But it is desirable to be on guard, and check the data for signs of extreme and serious violations. Some assumptions are more important than others. Here an attempt will be made to examine some of the more important descriptive assumptions briefly.

LINEARITY

It is the assumption that the highest term entering the equation specifying the relationship of X with Y is of the first degree. This assumption could be tested with the help of a bivariate scatterplot which would show the nature and extent of the relationship of the two variables, X and Y. If the linearity assumption holds, it should be possible to fit (draw) a straight line across the spread of the datapoints in the scatterplot. This means, Y is linearly related to X. If on the other hand, a curved rather than a straight line, would seem to fit the scatter of the datapoints better, that would mean that the two variables are related to one another in a curvilinear fashion. This would constitute a violation of the assumption of linearity.

For the three-variable model presented in this article, two rather than one scatterplots would be needed to check for linearity. The first of the two scatterplots would provide an indication of the relationship between X₁ and Y, that is between TV News Exposure and Knowledge, while the second would allow for an examination of the relationship between X₂ and Y, that is between Surveillance and Knowledge.
MULTICOLLINEARITY

It is the extent to which \( X_1 \) and \( X_2 \) overlap. Violation of this assumption would lead to serious problems in analysis. Role of non-orthogonal regressors in a regression equation may become ambiguous and at times even meaningless. Multi collinearity could be known by looking at the intercorrelations between X's, that is between a pair of any two given independent variables. For instance, \( r_{12} \), that is the correlation coefficient between \( X_1 \) and \( X_2 \) (TV News Exposure and Surveillance in the case of the present model), would provide an indication of the extent of overlap between these two variables.

As a rule of thumb, an \( r_{12} \) of .80 may be taken as the danger limit. But this does not mean that lesser correlations could be treated as totally safe. For example, an \( r_{12} \) of .50 may not constitute a serious violation of the assumption of the orthogonality of predictors in a regression equation, yet it would constitute a clear signal to the researcher that all is not well with the data. The researcher must take the presence of this correlation between the independent variables into account while interpreting his results, especially where it concerns the regression coefficients (beta weights) associated with the interaction term in the regression equation.

HOMOSCEDASTICITY

Homogeneity of variance is an important assumption in ANOVA. Discriminant Analysis assumes that the variance-covariance matrices for the different groups are equal. Similarly, multiple regression assumes that \( Y \) variance is uniform across all values of \( X \). Violation of this assumption could lead to serious substantive and methodological problems. This assumption could be tested by looking at the plot of residuals. In such a plot if the errors make a more or less gourd or square or some other similar shape, this would be taken as an indication of the violation of the assumption of homoscedasticity. Residuals spread out in a straight band would indicate that the assumption has not been seriously violated.\(^8\)

NON-AUTOREGRESSIVE ERRORS

One of the assumptions of Multiple Regression is that the error terms are not interdependent but randomly distributed. Positive autoregression will produce wider confidence intervals making it difficult to reject the false null hypothesis and increasing the likelihood of Type II error. Negative autocorrelation will have the reverse impact. However, adjacent residuals of the same sign which result from ordering values in increasing magnitude may provide an indication of nonlinearity, and suggest inclusion of a quadratic terms in the model. In general, it is possible to incorporate the autoregressive scheme into a regression framework by suitable data transformations.

Two methods could be used to test this assumption. First, the Durbin-Watson statistic could be of use. If the test yielded an observed \( d \) value which is greater than
the specified $d_2$ value, then in a test for negative autocorrelation, the null hypothesis ($H_0: \tau = 0$) would be rejected, and the alternative hypothesis ($H_1: \tau < 0$) would be sustained. As a double check a plot of the residuals could be visually inspected. If the disturbances by and large form a nice looking zigzag pattern, and are not grouped in bunches on either side, then this is an indication of independent errors.

**$R^2$'s AND F TESTS**

$R$ is a multiple correlation coefficient. When squared, it gives the amount of variance in the dependent variable, $Y$, explained by the independent variables, $X$'s. $R^2$ is obtained by dividing regression sum of squares by the total sum of squares. Thus, $R$ is a measure of explained variance as a proportion of total variance in $Y$. $F$ on the other hand, represents a measure of explained variance as a proportion of unexplained or error variance. $F$ is obtained by dividing regression mean square by residual mean square.

The full regression model for the two independent variables, TV News Exposure and Surveillance, will be as follows:

$$Y = a + b_1 x_1 + b_2 x_2 + b_3 (x_1 x_2) + e$$

where $Y$ is the dependent variable (Knowledge), $b_i$ is the regression coefficient for the predictor variables, $x_i$ is the value of the independent variable, $(x_1 x_2)$ is the interaction term and $e$ is the error. Assuming this model yielded an $R^2$ of certain magnitude, this model could then be tested for significance against the ultimate null hypothesis, $H_0: Y = a + e$. The $F$ value for this model would be calculated as follows:

$$F = \frac{R^2/k}{1 - R^2 / (N - K - 1)}$$

This $F$ value would then be checked in the table of $F$ for significance.$^9$

At this stage, one of the important decisions facing the researcher would be what specific model to choose to enter the variables into the regression equation. There are two choices before him: a hierarchical model or a stepwise model. A hierarchical model is chosen when there is theory guiding analysis. In the absence of theory, stepwise procedures may be employed.

In a hierarchical model, the researcher enters the independent variables ($X$'s) in the equation one at a time based on theoretical considerations. Proceeding from a theoretical model, the researcher would normally have a fairly clear idea as to which independent variables would be expected to emerge as strong or weak covariates and predictors of the dependent variable, $Y$. In the case of a stepwise model on the other hand, the machine and not the man, that is the computer and not the researcher, determines the order of the entry of variables in the equation. The computer does so on the basis of strength of the impact of each $X$ on $Y$ as indicated by the amount of variance shared by the two.
If the researcher chooses a hierarchical model, then he is faced with another important decision. He must now decide whether to choose a "step up" or "step down" hierarchical model. The former model is indicated in cases where 'theory building' rather than 'theory testing' is involved. At the stage of theory building, the researcher is not sure of his theory. As a result, he enters the variables in the regression equation tentatively to see how each one fares. For theory testing purposes, a step down model is indicated, for at this stage, the researcher knows the probable covariates of his dependent variable, and from among them enters those into the equation first which he thinks would be most useful. At this stage, he is interested in parsimony, that is trying to see how he could explain most variance with the smallest number of independent variables. Since the theory underlying the model presented in this article is clear but not too strong, a step up hierarchical model would probably be most appropriate.

In such a model, first of all, the contribution of TV News Exposure to variance in Knowledge would be tested all by itself against the ultimate null hypothesis. For such a model, the null and alternative hypotheses would be as follows:

\[ H_0: Y = a + e \]
\[ H_1: Y = a + b_1x_1 + e \]

The F value for this model would be obtained by the following procedure:

\[ F = \frac{R^2_{Y,X_1}}{K} \]
\[ \frac{(1 - R^2_{Y,X_1})}{N - K - 1} \]

Where \( R^2_{Y,X_1} \) is the variance explained by \( X_1 \) in \( Y \), that is, by TV News Exposure in Knowledge, \( K \) is the number of independent variables in the equation and \( N \) is the size of the sample. The obtained value of \( F \) will then be checked in the \( F \) table for the probability level associated with it.

At the second step up the ladder, the role of Surveillance would be checked to see how much variance this new variable contributed after accounting for the variance contributed by the first variable, TV News Exposure. The regression equation at this stage would be as follows:

\[ H_0: Y = a + b_1x_1 + b_2x_2 + e \]
\[ H_1: Y = a + b_1x_1 + b_2x_2 + e \]

The F value for this equation will be calculated as follows:

\[ F = \frac{R^2_{Y,X_1,X_2} - R^2_{Y,X_1}K_1 - K_2}{(1 - R^2_{Y,X_1,X_2})N - K_1 - 1} \]
At the third step, the role of TV News Exposure Surveillance interaction would be tested to see if it will make any significant contribution to Knowledge in an equation that already contains $X_1$ and $X_2$ as additive variables. In other words, the interactive term would be entered into the equation to see if it would explain variance in $Y$ (Knowledge) over and above what the two $X$'s have additively explained. The $H_0$, $H_1$ and the $F$ value for such an equation will be as follows:

$$H_0: Y = a + b_1x_1 + b_2x_2 + e$$
$$H_1: Y = a + b_1x_1 + b_2x_2 + b_3(x_1x_2) + e$$

$$F = \frac{R^2_{Y x_1 x_2} (x_1 x_2) - R^2_{Y x_1 x_2}}{K_1 - K_2}$$

$$\frac{(1 - R^2_{Y x_1 x_2})}{N - K_1 - 1}$$

**APPLICATION**

Multiple regression analysis has been widely used in mass communications research, especially in empirical studies investigating causal relationships among variables and those using simultaneous equations for the estimation of path coefficients. Grunig used multiple regression analysis along with canonical correlation analysis to predict the nature of relationship between media use patterns and certain antecedent variables such as family and work background and time-budgets and level of involvement (Grunig, 1979:248). Prisuta employed multiple regression procedures to ascertain the inter-relationship between certain background variables such as age, race, peer interest in public affairs, etc. and television news viewing preferences among adolescents (Prisuta, 1979:277).

Blankenberg used a series of multiple regression analysis to assess the relative influence of structural and discretionary factors on newspaper circulation (Blankenberg, 1981:543). In another study, Roberts used regression techniques to investigate the effect of time spent by parents and children with television on their perception of violence in television programs (Roberts, 1981:556).

Pasha used a computer-based stepwise multiple regression model to assess the validity of an earlier theory-based hierarchical regression model. Results of the new analysis confirmed earlier findings as the machine selected and included in the model predictor variables in the same order as the previously tested hierarchical model, and the beta values remained unaffected. (Pasha, 1983:1)

In an attempt to estimate path coefficients to predict the mass media behavior of Korean immigrants in Hawaii, Yum employed multiple regression procedures through a series of simultaneous equations (Yum, 1982:154). In another interesting study Reeves and Garramone used regression techniques to investigate the difference identification with television characters made on children's evaluation of their peers on the one hand, and the variance in children's rating of their peers as a result of television viewing (Reeves & Garramone, 1982:317)
Wakshlag and Greenberg made use of regression procedures to determine the extent to which children were adventurous in their choice of programs, and to investigate the success of programming strategies by American television networks (Wakshlag & Greenberg, 1979:58). Some other examples of research employing multiple regression analysis as their primary methodological tool include studies by Berman and Stookey (1980:330) Austin (1984,74) and Singer et al (1984:73)

Berman and Stookey were interested in looking at the interrelationship between television viewing habits of adolescents and their affective orientation toward government. Austin was trying to explain attendance at art film shows (as contrasted with the general Hollywood fare of a more popular genre) using a set of variables such as brochure write-ups, having a series ticket, newspaper advertisements, color photography, etc. He used a stepwise model with forward inclusion. Singer et al. were estimating the extent to which a combined set of family and television variables contributed to children’s perception of their neighborhoods and the broader world as “scary”.

In a recent textbook on mass communication research methods, Wade has reviewed a number of statistical designs and procedures (Wade, 1981). In this review, she has attempted to explain and illustrate some of the key concepts of multiple regression analysis. The treatment, however, is not detailed and leaves some basic questions unanswered.

**SUMMARY AND CONCLUSION**

This article presented an outline of a multiple regression model for use in mass communications research. Starting with an explanation of the role of theory in research, the article postulated a theoretical model containing three concepts, namely, Knowledge, TV News Exposure and Surveillance motivation for television viewing. Notions of reliability, validity and sample selection were then explicated. Following this, some of the more important descriptive assumptions underlying multiple regression analysis were discussed. Finally, the procedures for hypothesis-testing in multiple regression analysis were outlined, and the use of R² and F-ratio in multiple regression models was illustrated. The article is expected to be of use to those interested in social research in general and in mass communications research in particular.

**NOTES**


**Bibliography**

Austin, B. "Portrait of an Art Film Audience". *Journal of Communication*, Winter 1984, 74.


Franklin, B.J. and Osborne, H.W. (eds.) "Time Budgets, Level of Involvement and Use of the Mass Media", *Journalism Quarterly*, Summer 1979, 248


Nie, N.H., Hull, C.H.
Jenkins, J.G.,
Steinbrenner, K. And
Bent, D.H.
Pasha, S.H.

Prisuta, R.

Reeves, B and
Garramone, G.
Roberts, C.

Singer, J. Singer, D.
and Repaczynski, W.
Smith H.W.

Wade, S.

Wakshlaq and
Greenberg, B.
Yum, J.


"Family Patterns and Television Viewing as Predictors of Children's Beliefs and Agression", Journal of Communication, Spring 1984, 73.


"Programming Strategies and the Popularity of Television for Children", Human Communication Research, Fall 1979, 58.