ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/283695133

How a variable's partial correlation with other variable(s) can make a good predictor: the suppressor variable case

Article *in* International Journal of Advanced Statistics and Probability · November 2015

DOI: 10.14419/ijasp.v3i2.5400

CITATION
1

READS
175

5 authors, including:

Olusegun Akinwande
Nigerian Center for Disease Control
10 PUBLICATIONS **178** CITATIONS
SEE PROFILE

Umar Adamu Abubakar
Ahmadu Bello University
14 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project The Transmuted New Weighted Weibull Distribution (T-NWWD): It's Properties and Application View project

Project The Exponentiated New Weighted Weibull Distribution: Theory and Practice View project

All content following this page was uploaded by Olusegun Akinwande on 16 November 2015.

The user has requested enhancement of the downloaded file.

International Journal of Advanced Statistics and Probability, 3 (2) (2015) 210-214
www.sciencepubco.com/index.php/IJASP
©Science Publishing Corporation
doi: 10.14419/ijasp.v3i2.5400
Research Paper

How a variable's partial correlation with other variable(s) can make a good predictor: the suppressor variable case

Akinwande Michael Olusegun [1]*, Aminu Muktar [1], Kaile Nasiru Kabir [1], Ibrahim Abubakar Adamu [2], Umar Adamu Abubakar [1]

[1] *Department of Mathematics Ahmadu Bello University Zaria*
[2] *Department of Mathematics and Statistics Abubakar Tatari Ali Polytechnic Bauchi*
**Corresponding author E-mail: akinwandeolusegun@gmail.com*

Copyright © 2015 Akinwande Michael Olusegun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Suppression effect is one of the most elusive and difficult to understand dynamics in multiple regression analysis. Suppressor variable(s) and their dynamics in multiple regression analyses are important in reporting accurate research outcomes. However, quite a number of researchers are unfamiliar with the possible advantages and importance of these variables. Suppressor variables tend to appear useless as separate predictors, but have the potential to change the predictive ability of other variables and completely influence research outcomes. This research describes the role suppressor variables play in a multiple regression analysis and provides practical examples that further explain how suppressor effects can alter research outcomes. Finally, we employed mathematical set notation to demonstrate the concepts of suppressor effects.

Keywords: *Suppression Effect; Stepwise Process; Regression; Correlation; Set Notation.*

1. Introduction

When selecting input variables, most researchers often check for possible significant correlations between the dependent variables and potential relevant input variables [1]. In some instances, some of the input variables are uncorrelated or have a near zero correlation with the response variable [10]. This situation raises the question of whether a researchers' multiple regression analysis should exclude or incorporate input variables that are not significantly correlated with the response variable [12]. Such questions are mostly not given the requisite attention. In multiple regression equations, suppressor variables increase the weight of predictor variable's coefficients associated with other independent variables or set of variables [5] [9]. A suppressor variable correlates significantly with other predictor variables, and accounts for or suppresses some irrelevant noise in such predictor variables as well as improving the overall predictive power of the model [2]. Given this illustration, some researchers' prefer to refer to suppressor variable as an enhancer and not suppressor (McFatter, 1979). Suppressor variables are classified into four, namely:

- Classic suppression
- Negative suppression
- Reciprocal suppression
- Absolute/Relative suppression

1.1. Set notation illustration

Let A, B and C be any arbitrary variables where A is regarded as the response variable while B & C are the potential input variables ∃ A&B are related and B&C are also related but A&C are not directly related. But if A&B exist and B&C exists, → A&C exist by implication.

Also, if A∩B and B∩C → A∩C exists. This implies that the relationship between A&C is called a partial one and holds because of the relationship between B&C.

2. Methodology

2.1. Coefficient of correlation

Correlation analysis comes into play in this study, in the sense that the test for association within input and response variable will be done bi-variately and separately which can be referred to as the limitation in determining suppression effect, this is done in this manner so as to be able to get a clearer picture of the kind of relationship the predictors share within themselves and also, the nature and kind of relationship between the input and response variables arbitrarily say x and y usually denoted by $r_{(x,y)}$ or simply $r_{x,y}$. This is a numerical measure of the linear relationship between the two random variables, and it is defined mathematically as:

$$r = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{\bar{x}})^2}} \tag{1}$$

2.2. Stepwise process

Empirically, the stepwise process employs the F-statistic in the partial F-test for its selection process. The test statistic for the stepwise process is denoted by F^* which compares the Means Square of the Regressors (MSR) with the Mean Square of the Error (MSE) for selecting relevant variables.

$$F^* = \frac{MSR}{MSE} \tag{2}$$

The stepwise process begins by fitting a simple regression model for each of the $p-1$ potential X variables:

$$F^* = \frac{MSR(X_k)}{MSE(X_k)} \tag{3}$$

$$F_1^* = \frac{MSR\left(X_1/X_2,\ldots,X_{p-1}\right)}{MSE\left(X_1,\ldots,X_{p-1}\right)} \tag{4}$$

$$F_k^* = \frac{MSR\ (X_k/X_i)}{MSE\ (X_i,X_k)} = \left[\frac{b_k}{S(b_k)}\right]^2 \tag{5}$$

Assuming X_2 is the variable entered in step 1, the stepwise process will fit all regression models with all variables where X_2 is one of the pair. Therefore for such regression model, the partial F test statistic will be:

$$F^* = \frac{SSR\ (X_2/X_1,X_3\ldots,X_{k-1},X_{k+1},\ldots,X_{p-1})}{MSE} \tag{6}$$

If H_0 holds, then $F^* \sim F_{(1,n-p)}$. Large values of F^* leads to the conclusion of H_a. Recall that MSR (X_k) = SSR (X_k) measures the reduction in the total variation of Y associated with the use of variable X_k. The variable X with the largest F^* values is selected as the candidate variable for addition if F^* value exceeds a predetermined level. Thus, the variable X is added otherwise the program terminates with no X variable is considered sufficiently helpful to enter into the regression model (John, William, & Michael, 1983).

2.3. Multiple regression analysis

Multiple Linear Regression analysis is an improvement on Simple Linear Regression analysis so as to be able to incorporate more than one predictor variable, the statistics used to assess the association between two or more input variables and a single response variable [7] [13]. The general form of a multiple linear regression equation is as follows:

$$Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_kX_k + \varepsilon_i \tag{7}$$

$$Y_i = \beta_0 + \sum_{j=1}^{n} X_{ij}\beta_{ij} + \varepsilon_i \tag{8}$$

2.4. Parameter estimation

To estimate the parameters in a Multiple Regression Analysis, we use the extension of the least square estimation procedure. Given the general case of the multiple regression model:
The least square function is written as:

$$S=(\beta_0, \beta_1, \beta_2, .., \beta_k) = \sum_{i=1}^{k} \varepsilon_i^2 = \sum_{i=1}^{k}\left(y_i - \beta_0 \sum_{i=1}^{n} \beta_j x_{ij}\right)^2 \quad (9)$$

The function S is been minimized with respect to $\beta_0, \beta_1, \beta_2, .., \beta_k$ which must satisfy the partial derivativeδS.
Therefore;

$$\frac{\delta S}{\delta \beta_0}(\beta_0, \beta_1, \beta_2, .., \beta_k) = -2\sum_{i=1}^{n}\left(y_i - \beta_0 \sum_{i=1}^{n} \hat{\beta}_j X_{ij}\right) = 0 \quad (10)$$

And

$$\frac{\delta S}{\delta \beta_1}(\beta_0, \beta_1, \beta_2, .., \beta_k) = -2X_1\sum_{i=1}^{n}\left(y_i - \beta_0 \sum_{i=1}^{n} \hat{\beta}_j X_{ij}\right) = 0 \quad (11)$$

We solve the above equations and obtain the least square normal equations as follows:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{12} + \cdots + \hat{\beta}_j \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \quad (12)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_j \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} y_i x_{i1} \quad (13)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}\, x_{i2} + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} y_i x_{i1} \quad (14)$$

$\vdots$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{ik} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}x_{ik} + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} y_i x_{ik} \quad (15)$$

Consider an illustration involving two predictor variables, X_1 and X_2.

Here r_{yx_1} =0.707106, r_{yx_2} =.0, and $r_{x_1x_2}$ = -0.707106. For these data, the beta weight $\beta_1$for the first predictor, X_1, will equal:

$\hat{\beta}_1$= [r_{yx_1}- (r_{yx_2}) ($r_{x_1x_2}$)] / 1-$r_{x_1x_2^2}$

= [0.707106 - (.0) (-0.707106)] / 1-(0.707106)^2

= [0.707106 - (.0) (-0.707106)] / 1- 0.5

= [0.707106 - (.0) (-0.707106)] / 0.5

= [0.707106 - .0] / 0.5

= 0.707106 / 0.5

= 1.414213.

The beta weight β_2 for the second predictor, X_2, will equal:

$\hat{\beta}_2$ = [r_{yx_2}- (r_{yx_1}) ($r_{x_1x_2}$)] / 1-$r_{x_1x_2^2}$

= [.0 - (0.707106) (-0.707106)] / 1 - (0.707106)^2

= [.0 - (0.707106) (-0.707106)] / 1 - 0 .5

= [.0 - (0.707106) (-0.707106)] / 0.5

= [.0 - (-0.5)] / 0.5

= 0.5 / 0.5

= 1.0.

The coefficient of determination R^2 for these equals:

$R^2 = (\hat{\beta}_1)\,(r_{yx_1}) + (\hat{\beta}_2)\,(r_{yx_2})$

= (1.414213) (0.707106) + (1.0) (.0)
= 1.0 + 0.0

= 1.0.

Thus, in this example, even though X_2 has a zero correlation with Y_i, the use of X_2 aspart of prediction along with X_1 doubles the predictive efficacy of the predictors, yielding perfect prediction.

3. Discussion

Some of the advantages accrue for accurately identifying suppression effects in multiple regression analysis. Incorporating suppressor variables in multiple regressions will amount to three positive outcomes: determining more accurate regression coefficients associated with predictor variables; improving overall predictability of the regression model; and enhancing accuracy of theory and model building.

Firstly, the risks inherent in excluding a relevant variable far outweigh the risks inherent in including an irrelevant variable [3] [4]. The regression weight of a predictor variable may change depending on its association with other predictor variables in the model. If a suppressor variable that should have been in the model is missing, that omission may substantially alter the research results, including an underestimated regression coefficient of the suppressed variable, higher model error sum of squares, and lower predictive power of the model as it has been shown in the analysis [6] [8]. An incomplete set of predictor variables may not only underestimate regression coefficients, but in some instances, will increase the probability of making a Type II error by failing to reject the null hypothesis when it is false [17]. In contrast, although including irrelevant variables in a model can contribute to multi-collinearity and loss of degrees of freedom. Hence, the risk of excluding a relevant variable outweighs the risk of including an irrelevant variable. To avoid underestimating the regression coefficient of a particular predictor variable, it is important to understand the nature of its relationship with other predictor variables [14] [16]. The concept of suppression provokes researchers to think about the presence of outcome-irrelevant variation in an independent variable that may mask that variable’s genuine relationship with the outcome variable.

However, in most research, predictor variables are inter-correlated, and regression coefficients are calculated after adjusting for all the bi-variate correlations between independent variables [16]. When a multiple regression model is altered by adding a variable that is uncorrelated with other predictor variables, the usual outcome is that the uncorrelated variable reduces the regression weight of the other predictor variable(s) [11]. The impact will be different if the added variable (or set of variables) is a suppressor variable. The suppressor variable will account for irrelevant predictive variance in some predictors and therefore will yield an increase in the regression weight of those predictors. Moreover, the regressor weight of the suppressor may improve, thus improving the overall predictive power of the model [6]. Suppression implies that the relationship between some independent variables of interest and the outcome variables are blurred because of outcome-irrelevant variance; the addition of suppressor variables clears or, purifies the outcome-irrelevant variation from the independent variables, thus revealing the true relationship between the independent and outcome variables [19].

4. Summary and conclusion

Our primary goal in this work as initially stated in our objectives is to highlight the dynamics of suppression effect and the limitation of stepwise selection in Multiple Regression Analysis research as well as to draw the attention of researchers to the fact that suppressor variables in multiple regression analysis are more prevalent than previously recognized [18]. The idea that a variable, which is unrelated to the dependent variable, should be retained not only for theoretical purposes but also to improve overall predictability of a model. Horst (1941) has recommended that researchers should retain a variable, even if it has negligible/weak correlation with the dependent (response) variable but has a significant correlation with other predictor (independent) variables. Furthermore, other benefits accrue from including suppressor variables in multiple regression models. Including a suppressor variable will eliminate the danger of rejecting a true hypothesis as false [15].

References

[1] Akinwande, M. O., Dikko H. G., & Gulumbe S. U (2015) Identifying the Limitation of Stepwise Selection for VariableSelection in Regression Analysis. *American Journal of Theoretical and Applied Statistics*. 414-419. http://dx.doi.org/10.11648/j.ajtas.20150405.22
[2] Bertrand, P. V. (1988). A quirk in multiple regression: The whole regression can be greater than the sum of its parts. *The Statistician*, 371-374. http://dx.doi.org/10.2307/2348761.
[3] Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (Revised ed.). New York: Routledge.
[4] Cohen, J. C. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.
[5] Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 35-46. http://dx.doi.org/10.1177/001316447403400105.
[6] Courville, T. &. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational & Psychological Measurement*, 229-248.
[7] Fox, J. (1991). Regression diagnostics. *Beverly Hills, CA: Sage*.
[8] Henard, D. (1998). Suppressor variable effects: Toward understandingan elusive data dynamic. *SouthwestEducational Research Association, Houston*.
[9] Liebscher, G. (2012). A Universal Selection Method in Linear Regression Models. *Open Journal of Statistics* , 153-162. http://dx.doi.org/10.4236/ojs.2012.22017.
[10] Lutz, J. G. (1983). A method for constructing data which illustrate three types of suppressor variables. *Educational and Psychological Measurement*, 373-377. http://dx.doi.org/10.1177/001316448304300206.
[11] McFatter, R. M. (1979). The use of structural equation models in interpreting regression equations including suppressor and enhancer variables. *Applied Psychological Measurement*, 123-135. http://dx.doi.org/10.1177/014662167900300113.
[12] Morrow-Howell, N. (1994). The M word: Multicollinearity in multiple regression. *Social Work Research*, 247-251.
[13] Nathans, L. L. (2012). Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. *Practical Assessment, Research & Evaluation , 17*, 123-136.
[14] Paulhus, D. L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research*, 303-328. http://dx.doi.org/10.1207/s15327906mbr3902_7.
[15] Rosenberg, M. (1973). The logical status of suppressor variables. *Public Opinion Quarterly, 37*, 359-372. http://dx.doi.org/10.1086/268098.
[16] Shanta, P., & Williams, E. (2010). Suppressor Variables in Social Work Research: Ways to Identify in Multiple Regression Models. *Journal of the Society for Social Work and Research* , 28-40.
[17] Smith, R. L. (1992). Suppressor variables in multiple regression/correlation. *Educational and Psychological Measurement*, 17-29. http://dx.doi.org/10.1177/001316449205200102.
[18] Tzelgov, J. &. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin*, 524-536. http://dx.doi.org/10.1037/0033-2909.109.3.524.
[19] Walker, D. A. (2003). Suppressor variable(s) importance within a regression model: An example of salary compression from career services. *Journal of College Student Development*, 127-133. http://dx.doi.org/10.1353/csd.2003.0010.

View publication stats