Open Journal of Statistics, 2015, 5, 754-767
Published Online December 2015 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2015.57075

Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis

Michael Olusegun Akinwande*, Hussaini Garba Dikko, Agboola Samson

Department of Mathematics, Ahmadu Bello University, Zaria, Nigeria
Email: *akinwandeolusegun@gmail.com, hgdikko@yahoo.com, abuagboola@gmail.com

Received 29 September 2015; accepted 21 December 2015; published 24 December 2015

Abstract

Suppression effect in multiple regression analysis may be more common in research than what is currently recognized. We have reviewed several literatures of interest which treats the concept and types of suppressor variables. Also, we have highlighted systematic ways to identify suppression effect in multiple regressions using statistics such as: R^2, sum of squares, regression weight and comparing zero-order correlations with Variance Inflation Factor (VIF) respectively. We also establish that suppression effect is a function of multicollinearity; however, a suppressor variable should only be allowed in a regression analysis if its VIF is less than five (5).

Keywords

Suppression Effect, Multicollinearity, Variance Inflation Factor (VIF), Regression and Correlation, Stepwise Selection

1. Introduction

When selecting a set of study variables, researchers frequently test correlations between the outcome variables (*i.e.* dependent variables) and theoretically relevant predictor variables (*i.e.* independent variables) [1]. In some instances, one or more of the predictor variables are uncorrelated with the outcome variable [2]. This situation poses the question of whether researchers' multiple regression analysis should exclude independent variables that are not significantly correlated with the dependent variable [3].

Questions such as this are most times not given the supposed credit. In multiple regression equations, suppressor variables increase the magnitude of regression coefficients associated with other independent variables

*Corresponding author.

How to cite this paper: Akinwande, M.O., Dikko, H.G. and Samson, A. (2015) Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open Journal of Statistics*, **5**, 754-767.
http://dx.doi.org/10.4236/ojs.2015.57075

or set of variables [4]. A suppressor variable correlates significantly with other independent variables, and accounts for or suppresses some outcome-irrelevant variations in such predictors as well as improving the overall predictive power of the model. Given this function, some prefer to call the suppressor variable an enhancer [5].

1.1. Stepwise Regression and Its Limitation

Stepwise regression is a common technique used to eliminate variables when the relationship of each predictor variable with an outcome variable is tested separately for statistical significance. Predictor variables that are not significantly related to outcome variables are often eliminated at the bi-variate level. Bi-variate results obtained from stepwise selection, provide only partial information about the relationship between a predictor and an outcome variable, and are an improper method for selecting variables for a multiple regression model. Some researchers have reported that when a multiple regression model is incorporated with a predictor variable that is uncorrelated with the outcome variable in a bi-variate model, the uncorrelated predictor variable sometimes significantly improved the explained variance [6]. Under such circumstances, the whole regression can be greater than the sum of parts [7]. Nevertheless, researchers often prematurely eliminate these variables during their variable selection process based on the variable's very low bi-variate correlation with the dependent (response) variable. However, eliminating these uncorrelated variables will cause the researcher to underestimate some of the parameters, and this will lead to undermining the predictive power of the model and may yield regression equations with less predictability because stepwise regression is a handicap in variable selection for variables with multicollinearity [8].

1.2. Multicolinearity

Collinearity is a linear association between two explanatory (predictor) variables. Two regressor variables are perfectly collinear if there is an exact linear relationship between the two.

Multicollinearity: Multicollinearity refers to a situation in which two or more explanatory (predictor) variables in a multiple regression model are related with each other and likewise related with the response variable. We have perfect multicollinearity if, for example as in the equation above, the correlation between two independent variables is equal to 1 or −1. In practice, we rarely face perfect multicollinearity in a data set. More commonly, the issue of multicollinearity arises when there is an approximate linear relationship among two or more independent variables.

In regression analysis, we look at the correlations between one or more input variables, or factors, and a response to visualize the strength and direction of association between them. But in practice, the number of potential factors you may include in a regression model is limited only by your imagination and your capacity to actually gather the desired data of interest.

1.3. Effect of Multicollinearity

Multicollinearity practically inflates unnecessarily the standard errors of the coefficients. Whereas, increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly far from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. Without multicollinearity (that is, with lower standard errors), those coefficients might be significant.

1.4. Handling Multicollinearity

A little bit of multicollinearity isn't necessarily a huge problem. But severe multicollinearity is a major problem, because it theoretically shoots up the variance of the regression coefficients, making them unstable. The more variance they have, the more difficult it is to interpret the coefficients. Some things to be concerned about when multicollinearity is a factor in multiple regression analysis are outlined as:

- A regression coefficient is not significant even though, in the real sense, that variable is highly correlated with Y.
- When you add or delete a predictor variable, the regression coefficients changes dramatically.
- Having a negative regression coefficient when the response should increase along with X.

- Having a positive regression coefficient when the response should decrease as *X* increases. One way to estimate multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases when predictors are correlated. If no factors are correlated, the VIFs will all be 1. If the variance inflation factor (VIF) is equal to 1 there is no multicollinearity among regressors, but if the VIF is greater than 1, the regressors may be moderately correlated. A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, it can be assumed that the regression coefficients are poorly estimated due to multicollinearity which should be handled accordingly. If multicollinearity is a problem in a multiple model, that is, the variance inflation factor (VIF) for a predictor is near or above 5. The solution may be simply to:
- **Remove highly correlated predictors from the model:** If there are one or more factors with a high VIF, one of the factors should be removed from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared. However, instead of tagging multicollinearity as a disadvantage in multiple linear regressions, we are viewing it as an advantage in the sense that predictors act as suppressor variables in regression analysis leveraging on presence of multicolinearity among independent variables because a predictor which shares zero order correlation with the response variable can only be retained in the model if and only if it is significantly correlated with one or more predictor variables under study. Having studied the concept and effect of multicollinearity we can theoretically say that a suppressor variable should be allowed in a regression model if and only if the variance inflation factor (VIF) is below 5, that is, if the strength of multicolinearity in the model does not account for rendering other predictors redundant (less significant when they should be practically significant) [9].

1.5. Categories of Suppressor Variables

Since the introduction of the concept of suppression, many authors have expanded the definition of these variables (see for example, [5] [9] [10] [11] opined that the name "suppressor variable" may have a pejorative connotation because "suppression" sounds like "repression". On the contrary, suppressor variables are actually advantageous because they improve the prediction of the criterion. In essence, these variables suppress irrelevant variance in the other predictor variable(s), thus indirectly allowing for a more concise estimate of the predictor-criterion relationship, even though the suppressor variable directly predicts none or almost none of the criterion variable's variance

There are four types of suppressor variables: the classic suppressor, the negative suppressor, the reciprocal suppressor, and the absolute and relative suppressor. We briefly introduce each type below.

1.6. Classic Suppression

Classic suppression in multiple regression analysis was originally introduced and was later demonstrated mathematically. Although, there exist a zero-order correlation between a suppressor and an outcome variable (zero correlation), the prediction in the outcome variable increases when a suppressor variable is added to the equation simply because the suppressor variable is correlated with another predictor (or set of predictors) that are correlated with the outcome (dependent) variable. In this case, the suppressor variable removes irrelevant predictive variance from the other predictor (or set of predictors) and increases the predictor's regression weight, thus increasing overall model predictability. Sometimes the suppressor variable may also receive nonzero regression weight with a negative sign. However, a variable is a suppressor only for those variables whose regression weights are increased. Thus, a suppressor is not defined by its own regression weight but rather by its effects on other variables in a regression system [4].

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

Here $r_{yx_1} = 0.707106$, $r_{yx_2} = 0.0$, and $r_{x_1x_2} = -0.707106$. For these data, the beta weight β_1 for the first predictor, X_1, will equal:

$$\begin{aligned}\hat{\beta}_1 &= \left[r_{yx_1} - \left(r_{yx_2}\right)\left(r_{x_1x_2}\right)\right] \Big/ 1 - r_{x_1x_2^2} = \left[0.707106 - (0.0)(-0.707106)\right] \Big/ 1 - (0.707106)^2 \\ &= \left[0.707106 - (0.0)(-0.707106)\right] \Big/ 1 - 0.5 = \left[0.707106 - (0.0)(-0.707106)\right] \Big/ 0.5 \\ &= \left[0.707106 - 0.0\right] \Big/ 0.5 = 0.707106 / 0.5 = 1.414213\end{aligned}$$

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

$$\begin{aligned}\hat{\beta}_2 &= \left[r_{yx_2} - \left(r_{yx_1}\right)\left(r_{x_1x_2}\right)\right] \Big/ 1 - r_{x_1x_2^2} \\ &= \left[0.0 - (0.707106)(-0.707106)\right] \Big/ 1 - (0.707106)^2 \\ &= \left[0.0 - (0.707106)(-0.707106)\right] \Big/ 1 - 0.5 \\ &= \left[0.0 - (0.707106)(-0.707106)\right] \Big/ 0.5 \\ &= \left[0.0 - (-0.5)\right] \Big/ 0.5 \\ &= 0.5/0.5 \\ &= 1.0\end{aligned}$$

The coefficient of determination R^2 for these equals:

$$\begin{aligned}R^2 &= \left(\hat{\beta}_1\right)\left(r_{yx_1}\right) + \left(\hat{\beta}_2\right)\left(r_{yx_2}\right) \\ &= (1.414213)(0.707106) + (1.0)(0.0) \\ &= 1.0 + 0.0 \\ &= 1.0\end{aligned}$$

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

1.7. Negative Suppressor

Negative suppression was introduced and later explained mathematically by [4]. A negative suppressor works in a manner similar to that of a classic suppressor by removing irrelevant variance from a predictor (or set of predictors), increasing the predictor's regression weight, and increasing overall predictability of the regression equation. The difference between these two types of suppressors is the negative suppressor's positive zero-order correlation with other predictor variable(s) and with the outcome variable; however, when entered in multiple regressions, the negative suppressor has a negative beta weight [4].

1.8. Reciprocal Suppressor

Reciprocal suppression was introduced by [4]. Some authors have also called this concept suppressing confounders [1]. Here, both the predictor and the suppressor variable have a positive zero-order correlation with the outcome (response) variable but have a negative zero-order correlation with each other this part the regressors share is actually irrelevant to Y_i, that is; X_1 and X_2 having a negative zero-order correlation with each other. When Y_i is regressed on these two variables, X_1 and X_2 will suppress some of their irrelevant information, increase the regression weight of each other making their regressor coefficients positive respectively, and thus improve model R^2.

1.9. Absolute/Relative Suppressor

Absolute and relative suppression was originally introduced by [4] and further clarified by [12]. According to [12], absolute suppression is defined by the relationship between the predictor's weight in bi-variate regression equation and its weight in multivariate equations. It exists whenever adding predictors increases the weight of the variable relative to its weight in the bi-variate equation. On the other hand, if the regression weight of a predictor variable increases when a new variable is added to a regression equation, but the increase is not beyond the respective weight of the predictor in the bi-variate mode, then the new variable is a relative suppressor [12]. Therefore, relative suppression is tested hierarchically, and the researcher must compare the standardized beta weights of the predictors (regressos) in the equation before and after the inclusion of the variable that may be a potential relative suppressor. Hence, relative suppression should be tested only when there are three or more predictors [12].

1.10. How Common Are Suppressor Variable(s) in Multiple Regression?

The use of suppressor variables in multiple regressions is more common than currently recognized [13] [14]. This lack of recognition may be as a result of the fact that suppressor variables are not necessarily a special category of predictor (independent) variables; they can be any predictor (or independent) variable in a multiple regression model, including variables for race/ethnicity, income, education, and self-worth [14]. Using a multiple regression model to predict the salary of administrators at educational institutions, [11] found that the variable for level of education attained acted as a suppressor variable. The variable for level of education had a close to zero (but positive) zero-order correlation with administrators' salaries (dependent (response) variable) at both public and private institutions (r = 0.010 and 0.014 for public and private institutions respectively). However, the model's regression coefficient associated with the level of education was not only statistically significant but was also negative. This finding prompted [11] to test the level of education variable for its suppression effect. He noted that at the bi-variate level, the level of education variable was weakly correlated with the dependent variable (salary) but was significantly correlated with other independent variables, including respondent's age. To determine if level of education was a suppressor variable, [11] ran the regression model with and without the level of education variable included in the regression models predicting salary (dependent variable). In the model for public institutions, the addition of the level of education variable increased the (coefficient of multiple determination) R^2 from 0.26 to 0.28. In the model for private institutions that excluded the level of education variable, the R^2 was 0.22; the inclusion of the level of education variable increased the R^2 to 0.36. [11] concluded that the level of education was a suppressor variable in predicting salary of administrators for both public and private educational institutions [15].

Similarly [16], the suppressor effect of a variable for cognitive ability was demonstrated by [17] in a study examining outcomes of medical rehabilitation among older adults. Specifically, the study examined the probability of a patient's returning to independent living (*i.e.* living alone) versus living with others. [17] and colleagues noted that demographic variables for age and education became significant predictors of return to independent living only when the model included the variable for cognitive ability. Although the authors concluded the cognitive ability variable produced a suppressive effect, they did not analyze the nature of suppression.

Having reviewed relevant literatures as to the nature, implication, behavior and identification of suppressor variable(s) and its effect in multiple regression analysis validation and reporting of results, we can say to a reasonable extent that the concept of suppression effect in multiple regression has for long been in existence but has not been in lime light due to the fact that suppressor variables are not necessarily a special category of predictor (independent) variable in regression analysis. However they can simply be referred to as any predictor or independent variable that are not necessarily correlated with the outcome or response variable but linearly correlated with some or all the other predictors so to say.

1.11. Identifying Suppressor Variable(s)

As a result of the fact that researchers are in a perpetual search for substantive relationships between variables, they usually try to use predictors that they believe will be highly correlated with the response variable. For this reason, suppressor variables are usually not consciously sought out to be included in a regression equation. Fortunately, suppressor variables can be incorporated into a study unknown to the researcher. In these situations, even variables that would not be considered theoretically reasonable as direct predictors have possibilities for suppressor effect [4].

Another complication in detecting suppressor variables is that they may simply be overlooked because of their low zero-order correlations or non-correlation with the response variable [10]. The definitions above pay particular attention to two indicators of a suppressor effect: beta (β) weights and correlations between the predictors. However, many researchers neglect either one or the other [18]. The emphasis here is that interpretation of either beta (β) weights alone or correlation coefficients (r) alone may lead to major oversights in data analysis which should be stated that "the thoughtful researcher should always interpret either (a) both the beta weights and the structure coefficients or (b) both the beta weights and the bi-variate correlations of the predictors with Y (response)".

One final problem in detecting suppressor variables is the type of statistical analysis employed. The only analysis that has been discussed to this point is that of linear regression where the predictors are inter-correlated [16]. Knowledgeable researchers understand that all least squares analyses are in fact forms of the General Li-

near Model. For example, [18] demonstrated that multiple linear regression subsumes all uni-variate parametric methods as special cases and that a uni-variate general linear model can be used for all uni-variate analyses. Ten years later, [18] demonstrated that canonical correlation analysis subsumes all parametric analyses, both uni-variate and multivariate, as special cases. Thus, it is not surprising that there is the possibility to obtain a suppressor effect in other forms of analysis.

2. Methodology

We undertook a review of science literatures and various databases to understand the concept of multicollinearity and suppressor variables in regression analysis, again we went ahead to further examine the linkage between multicollinearity and suppression effect keeping in mind the supposed implication of multicollinearity in over or underestimating regression inferences . Next, we designed a sample study for the purpose of illustrating the setbacks of refusing to allow a suppressor variable in a regression analysis without obtaining it variance inflation factor (VIF).

Data Source and Type

Solely for the purpose of illustration, in our investigation we employed the use of a simulated data from MINITAB (14) and Microsoft Excel (2007). These data is a 5 variable data, we have also assigned arbitrary names to the variables which include: Grain Yield, Plant Heading, Plant Height, Tiller Count and Panicle Length. A limitation of this analysis is that we have as a result of the fact that it is sometime nearly hard to have a set of data which has a zero order correlation between them, but having our objective in mind, that is, to show the limitation of stepwise selection in been able to select a variable with zero or nearly zero order correlation with the response variable and to show that we cannot talk about suppression effect in analysis without talking about multicollinearity. Therefore we require a set of predictor variables that exhibit the basic nature of effect we intend to show that is, independent variables that have near zero or very weak correlation with the outcome (dependent) variable and other predictor variables that has a non-zero correlation with the response variable. We have ignored limitations that are inherent in the use of such data. Readers should ignore all implications of our findings, taking away from this exercise only the discussion that pertains to the limitation of stepwise selection and the advantage of multicollinearity as regards suppression effect.

The statistical packages used for this study are **R-Package 3.2.2**, **Stat-Graphics (version 17), Minitab (version 14) and Microsoft Excel 2010**. The choice of these packages is due to preference.

3. Analysis and Results

Quite a number of authors have proposed the understanding suppressor variables by evaluating regression weights [4] [12] [18] [19]. Instead of the regression weights, some researchers have preferred squared semipartial correlation of the suppressor variable in evaluating suppressor effect of a variable [18] [20] [10]. In the current study, we intend to show the limitation of stepwise selection and the advantage of multicollinearity in regression analysis by evaluating the regressor weights and the general predictability of the regression model with VIF as a constraint.

3.1. Hypothesis

From the simulated data, we hypothesized that the Grain Yield of wheat if solely dependent on Plant Heading, Plant Height, Tiller Count and Panicle Length. Specifically, we examined the following hypothesis:

- The grain yield of wheat depends on the plant heading;
- The grain yield of wheat depends on plant height;
- The grain yield of wheat depends on tiller count;
- The grain yield of wheat depends on panicle length,

3.2. Measures

We picked five (5) variables from the simulated wheat grain yield data: 1) Grain yield; 2) Plant Heading; 3) Plant Height; 4) Tiller Count; 5) panicle Length. We treated plant heading, plant height, tiller count and panicle length as predictor (independent) variables while grain yield as response (dependent) variable.

4. Results

The first step of analysis involves a Pearson zero order correlation of the five variables that is, Grain Yield, Plant Heading, Plant Height, Tiller Count and Panicle Length. From **Table 1** below, we can clearly see that Grain Yield is remotely/weakly correlated with Tiller Count ($r = 0.001$) and Plant Height ($r = 0.039$), but it is significantly related with Plant Heading ($r = 0.591$) and Panicle Length ($r = 0.767$) respectively. From the zero order correlation result, we are able to see that just two out the four predictor variables are significantly correlated with the outcome (response) variable (that is, Plant Heading and Panicle Length). Therefore, we may just conclude that the variables to be selected should be Plant Heading and Panicle Length leaving out Plant Height and Tiller Count.

4.1. Correlation

The second analytic step is to clearly outline the correlated predictors. To this end, we check for multicollinearity among these four independent (predictor) variables. Therefore, from **Table 1**, the zero order correlation values between the four independent variables are:

- Plant Heading and Plant Height, Tiller Count, Panicle Length ($r = 0.093, -0.326, 0.261$);
- Plant Height and Tiller Count and Panicle Length ($r = 0.007, 0.174$);
- Tiller Count and Panicle Length ($r = 0.179$).

The third step involves assessment of Tiller Count as possible suppressor variable. Since Tiller Count is not significantly related with the outcome variable Grain Yield but the Tiller Count variable is significantly associated with the other predictor variables (that is, Plant Heading, Plant Height and Panicle Length) and therefore this suggests Tiller Count as a potential suppressor variable.

But before we go ahead to investigate the presence of suppression among the predictor variables, it is expedient to employ the already existing methods of variable selection in regression analysis to get a clear picture of the potentially relevant variable(s) that will be suggested by the various methods of variable selection as it are so as to further buttress our point.

4.2. Forward Selection

Stepwise Regression: Grain Yield versus Plant Heading, Plant Height, Tiller Count and Panicle Length. Response is Grain Yield on 4 predictors, with $N = 50$. From **Table 2** above, the forward selection process selects the plant heading and panicle length variable at 0.05 (α) as the significant variables to be included in the model as suggested by the correlation result in **Table 1** above with their corresponding p-values.

4.3. Backward Elimination

Stepwise Regression: Grain Yield versus Plant Heading, Plant Height, Tiller Count and Panicle Length.

Response is Grain Yield on 4 predictors, with $N = 50$.

From **Table 3** above, the forward selection process selects the plant heading and panicle length variable at

Table 1. Bi-variate zero order correlation.

	Grain Yield	Plant Heading	Plant Height	Tiller Count	Panicle Length
Grain Yield	1				
Plant Heading	0.5917	1			
***P*-Value**	0.000				
Plant Height	0.0393	0.0936	1		
***P*-Value**	0.786	0.518			
Tiller Count	**0.0016**	−0.3264	0.0070	1	
***P*-Value**	0.991	0.021	0.961		
Panicle Length	0.7675	0.2618	0.1745	0.1792	1
***P*-Value**	0.000	0.066	0.225	0.213	

0.05 (α) as the significant variables to be included in the model as suggested by the correlation result in **Table 1** above with their corresponding *p*-values.

4.4. Stepwise Selection

Stepwise Regression: Grain Yield versus Plant Heading, Plant Height, Tiller Count and Panicle Length.

Response is Grain Yield on 4 predictors, with $N = 50$.

From the three methods of variable selection (**Tables 2-4**) (that is, forward selection, backward elimination

Table 2. Forward selection alpha-to-enter: 0.05(α).

Step	1	2
Constant	253.38	−95.54
Panicle Length	0.807	0.691
T-Value	8.30	8.76
P-Value	0.000	0.000
Plant Heading		0.406
T-Value		5.59
P-Value		0.000
R-Sq	58.91	75.30
R-Sq (Adj)	58.06	74.25

Table 3. Backward elimination. alpha-to-remove: 0.05(α).

Step	1	2	3
Constant	144.30	155.43	−95.54
Plant Heading	0.421	0.412	0.406
T-Value	5.37	5.76	5.59
P-Value	0.000	0.000	0.000
Plant Height	−1.11	−1.11	
T-Value	−1.62	−1.64	
P-Value	0.112	0.108	
Tiller Count	0.4		
T-Value	0.31		
P-Value	0.759		
Panicle Length	0.704	0.711	0.691
T-Value	8.50	9.06	8.76
P-Value	0.000	0.000	0.000
R-Sq	76.71	76.67	75.30
R-Sq (Adj)	74.64	75.14	74.25

Table 4. Alpha to enter and remove: 0.05(α).

Step	1	2
Constant	253.38	−95.54
Panicle Length	0.807	0.691
T-Value	8.30	8.76
P-Value	0.000	0.000
Plant Heading		0.406
T-Value		5.59
P-Value		0.000
R-Sq	58.91	75.30
R-Sq (Adj)	58.06	74.25

and stepwise selection) above, we are able to deduce that Plant Heading and Panicle Length are the potentially relevant variables to be included in the model as suggested by the three variable selection methods. But it is against this backdrop that we suggest the presence of a suppressor variable from the zero order correlation of the four predictor (independent) variables. We analyze having identified the existence of multicollinearity among the said predictor (independent) variables. To this end, we, therefore, identify Tiller Count as a potential suppressor variable because of its significant correlation with other predictor (Plant Heading) which is said to be the presence of multicollinearity within the said variables.

The fourth analytic step is to run a regression of the variables both in the bi-variate and multiple variable cases to explicitly reveal the suppression effect of the Tiller Count variable as the potential classic suppressor in the regression model.

4.5. Regression Analysis

The Bi-variate Case

Regression Analysis: Grain Yield versus Plant Heading

The regression equation is

$$\text{Grain Yield} = 554 + 0.573 \text{ Plant Heading} \quad (1)$$

$$RSq = 35.0\% \quad RSq(adj) = 33.7\%$$

Regression Analysis: Grain Yield versus Plant Height

The regression equation is

$$\text{Grain Yield} = 1153 + 0.37 \text{ Plant Height} \quad (2)$$

$$RSq = 0.2\% \quad RSq(adj) = 0.0\%$$

Regression Analysis: Grain Yield versus Tiller Count

The regression equation is

$$\text{Grain Yield} = 1246 + 0.03 \text{ Tiller Count} \quad (3)$$

$$RSq = 0.0\% \quad RSq(adj) = 0.0\%$$

Regression Analysis: Grain Yield versus Panicle Length

The regression equation is

$$\text{Grain Yield} = 253 + 0.807 \text{ Panicle Length} \quad (4)$$

$$RSq = 58.9\% \quad RSq(adj) = 58.1\%$$

Results obtained from **Tables 5-12** that is; the regression analysis in the bi-variate cases shows that the significant predictors among the four predictor variables are plant heading and panicle length. This is in agreement with the correlation result obtained in **Table 1** and also the suggestions by the stepwise selection process. The next step is to carry out the regression analysis in the multiple variable cases.

Table 5. Summary of regression coefficients.

Predictor	Coef	SE Coef	*T*-Value	*P*-Value
Constant	553.8	136.2	4.06	0.000
Plant Heading	0.5727	0.1126	5.09	0.000

Table 6. Analysis of variance.

Source	Df	Sum of Squares	Mean Square	*F*-Ratio	*P*-Value
Regression	1	13,980	13,980	25.86	0.000
Residual Error	48	25,948	541		
Total	49	4,939,928			

Table 7. Summary of regression coefficients.

Predictor	Coef	SE Coef	*T*-Value	*P*-Value
Constant	1152.5	343.9	3.35	0.002
Plant Height	0.368	1.346	0.27	0.786

Table 8. Analysis of variance.

Source	Df	Sum of Squares	Mean Square	*F*-Ratio	*P*-Value
Regression	1	61.9	61.9	0.07	0.786
Residual Error	48	39,866.3	830.5		
Total	49	39,928.2			

Table 9. Summary of regression coefficients.

Predictor	Coef	SE Coef	*T*-Value	*P*-Value
Constant	1245.85	49.30	25.27	0.000
Tiller Count	0.025	2.174	0.01	0.991

Table 10. Analysis of variance.

Source	Df	Sum of Squares	Mean Square	*F*-Ratio	*P*-Value
Regression	1	0.01	0.01	0.00	0.991
Residual Error	48	39,928.1	831.8		
Total	49	39,928.2			

Table 11. Summary of regression coefficients.

Predictor	Coef	SE Coef	*T*-Value	*P*-Value
Constant	253.4	119.7	2.12	0.040
Panicle Length	0.80689	0.09728	8.30	0.000

Table 12. Analysis of variance.

Source	Df	Sum of Squares	Mean Square	*F*-Ratio	*P*-Value
Regression	1	23,523	23,523	68.83	0.000
Residual Error	48	16,405	342		
Total	49	39,928			

4.6. Multiple Variable Cases

Regression Analysis: Grain Yield versus Plant Heading, Panicle Length

The regression equation is

$$\text{Grain Yield} = -96 + 0.406\ \text{Plant Heading} + 0.691\ \text{Panicle Length} \quad (5)$$

$$Sq = 75.3\% \quad RSq(adj) = 74.3\%$$

Regression Analysis: Grain Yield versus Plant Heading, Tiller Count and Panicle Length

The regression equation is

$$\text{Grain Yield} = 95 + 0.416\ \text{Plant Heading} + 0.38\ \text{Tiller Count} + 0.699\ \text{Panicle Length} \quad (6)$$

$$RSq = 75.4\% \quad RSq(adj) = 74.7\%$$

From the four regression analyses in the bi-variate case, in model 1, we regressed our outcome variable Grain

Yield on the predictor variable Plant Heading was significant and accounted 33.7% of the variance in the outcome variable. Plant Heading was positively associated with grain yield in the bi-variate correlation ($\beta_1 = 0.573, t = 5.09, p < 0.05$). This implies that as Plant Heading increases by one unit Grain Yield increases by 57.3%.

In model 2, Grain yield versus Plant Height which was insignificant as suggested by the correlation result in **Table 1** and the stepwise process respectively. Plant Height accounts for only 0.0% of the variance in the outcome variable ($\beta_1 = 0.37, t = 0.27, p > 0.05$). This suggests that the relationship between grain yield and plant height is negligible.

In model 3, Grain Yield versus Tiller Count was insignificant as suggested by the correlation result in **Table 1** and the stepwise process respectively. Tiller Count and Grain Yield were not associated this does not account for any variability in the outcome variable ($\beta_1 = 0.03\%, t = 0.01, p > 0.05$). This theoretically implies that there is no relationship at all between Grain Yield and Tiller Count.

In model 4, Grain Yield versus Panicle Length was significant as expected and accounted for about 58.1% of the variance in the outcome variable. Panicle Length which was positively associated with Grain yield has ($\beta_1 = 0.807, t = 8.30, p < 0.05$). This implies that Panicle Length increases by one unit Grain Yield increases by 80.7%.

4.6.1. Multicollinearity

Multicollinearity in regression is viewed as more of disadvantage, as it practically inflates unnecessarily the standard errors of coefficients in regression. Having studied Variance Inflation Factor (VIF) we know that a VIF of 5 and above is not good for regression model because it might render other significant variables redundant. Therefore, from our equation in **Table 13**, we can see that Plant Heading and Panicle Length share the same VIF 1.1 and 1.1 respectively as well as in **Table 14**, the analysis is very significant from the p-value column also in **Table 15** containing the three variables Plant Heading, Tiller Count and Panicle Length the VIFs are 1.3, 1.2, and 1.2 respectively and was also significant from the *p*-value column in **Table 16**. The Tiller Count and Panicle

Table 13. Summary of regression coefficients.

Predictor	Coef	SE Coef	*T*-Value	*P*-Value	VIF
Constant	−95.5	112.7	−0.85	0.401	
Plant Heading	0.40602	0.07270	5.59	0.000	1.1
Panicle Length	0.69141	0.07896	8.76	0.000	1.1

Table 14. Analysis of variance.

Source	Df	Sum of Squares	Mean Square	*F*-Ratio	*P*-Value
Regression	2	30,068	15,034	71.66	0.000
Residual Error	47	9860	210		
Total	49	39,928			

Table 15. Analysis of variance.

Source	Df	Sum of Squares	Mean Square	*F*-Ratio	*P*-Value
Regression	3	30,089	10,030	46.89	0.000
Residual Error	46	9840	214		
Total	49	39,928			

Table 16. Summary of regression coefficients.

Predictor	Coef	SE Coef	*T*-Value	*P*-Value	VIF
Constant	94.6	119.3	0.89	0.376	
Plant Heading	0.41584	0.07984	5.21	0.000	1.3
Tiller Count	0.381	1.219	0.31	0.756	1.2
Panicle Length	0.69910	0.08331	8.21	0.000	1.2

Length variable have the same VIF that is to say Tiller Count in the model serves as a classic suppressor Panicle Length. However, having studied this effect we say that instead of viewing multicollinearity as a disadvantage we are viewing it as an advantage since suppressor variables leverage on multicollinaerity among variables to act. That is to say, suppression effect is a function of multicollinearity. Therefore to this end we say that a suppressor variable should be allowed a place in a multiple regression model if its VIF is less than five (5).

4.6.2. Classic Suppression

Having identified Tiller Count as a suppressor variable, that is, Classic suppressor, from the correlation result in **Table 1** we can now infer from the inclusion of Tiller Count Variable in the multiple regression model for Grain Yield versus Plant Heading, Tiller Count and Panicle Length was significant as argued and the suppressor variable has improved the Beta Weight (coefficient) of the predictor variable Plant Heading from (0.406 to 0.416, $p < 0.05$) with that of Panicle Length from (0.691 to 0.699, $p < 0.05$) and has also improved the general predictability of the model as a whole.

Therefore from the above illustration we have been able to show a clear case of classic suppression in the regression model for Grain Yield versus Plant Heading, Tiller Count and Panicle Length.

4.6.3. Reciprocal Suppression

The final analytic step is to check for reciprocal suppression effect in the overall model side by side classic suppression. From the definition of reciprocal suppression; here, both the predictor variables (Tiller Count and Plant Heading) have a positive correlation with the outcome (response) variable but have a negative zero-order correlation with each other. When the response variable is regressed on these two variables, they will suppress some of their irrelevant information, increase the regression weight of each other, and thus improve model R^2. From our correlation result in **Table 1** above, we can clearly see that Plant Heading and Tiller Count are negatively correlated which suggests the presence of a reciprocal suppression effect. Therefore haven satisfied the condition for a reciprocal suppression effect, the regression analysis for Grain Yield versus Plant Heading, Plant Height Tiller Count and Panicle Length shows the result as expected. That is, the Beta (regressor) weights for Plant Heading, Tiller Count and Panicle Length are positive, the weights were also improved accordingly by clearing out the outcome irrelevant variances for each other and improving the overall predictability of the model.

4.7. Discussion

In this section, we discuss some of the advantages of accurately identifying suppression effects and the benefits of using suppressor variables in multiple regression analysis. Using suppressor variables in multiple regressions will yield three positive outcomes: determining more accurate regression coefficients associated with independent variables; improving overall predictive power of the model; and enhancing accuracy of theory building.

First, the risks associated with excluding a relevant variable are much greater than the risks associated with including an irrelevant variable. The regression weight of an independent variable may change depending upon its correlation with other independent variables in the model. If a suppressor variable that should have been in the model is missing, that omission may substantially alter the 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 above . An incomplete set of independent 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. In contrast, although including irrelevant variables in a model can contribute to multi-collinearity and loss of degrees of freedom, those variables will not affect the predictive power of the model. 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 independent variable, it is important to understand the nature of its relationship with other independent variables. 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.

Only when a predictor variable that is uncorrelated with other predictors is included in a multiple regression, will the regression weight of other predictor variables remain stable and not change. However, in most research, explanatory variables are inter-correlated, and regression coefficients are calculated after adjusting for all the

bi-variate correlations between independent variables. 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). 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.

Our example using the simulated Wheat Grain Yield data illustrates that the regression weight may change substantially when potential suppressor variables are included in models. If the regression weights of included variables improve dramatically due to the presence of a variable that was insignificant at the bi-variate level, then one or more of the independent variables may be acting as a suppressor. In our example, the presence of Tiller Count improved the regressor weights of Plant Heading and Panicle Length. Also, Plant Heading and Tiller Count served as Reciprocal suppressors in the overall model thereby clearing out the outcome irrelevant variance in each other thus improving the weights of each other.

References

[1] Cohen, J., Cohen, P., West, S.G. and Aiken, L.S. (2013) Applied Multiple Regression/Correlation Analysis for the Behavioral Science. Routledge, New York.

[2] Henard, D. (1998) Suppressor variable effects: Toward understandingan elusive data dynamic. *SouthwestEducational Research Association*, *Houston*.

[3] Morrow-Howell, N. (1994) The M Word: Multicollinearity in Multiple Regression. *Social Work Research*, **18**, 247-251.

[4] Conger, A.J. (1974) A Revised Definition for Suppressor Variables: A Guide to Their Identification and Interpretation. *Educational and Psychological Measurement*, **34**, 35-46. http://dx.doi.org/10.1177/001316447403400105

[5] McFatter, R.M. (1979) The Use of Structural Equation Models in Interpreting Regression Equations Including Suppressor and Enhancer Variables. *Applied Psychological Measurement*, **3**, 123-135. http://dx.doi.org/10.1177/014662167900300113

[6] Courville, T. (2001) Use of Structure Coefficients in Published Multiple Regression Articles: β Is Not Enough. *Educational & Psychological Measurement*, **61**, 229-248.

[7] Bertrand, P.V. (1988) A Quirk in Multiple Regression: The Whole Regression Can Be Greater than the Sum of Its Parts. *The Statistician*, **37**, 371-374. http://dx.doi.org/10.2307/2348761

[8] Lutz, J.G. (1983) A Method for Constructing Data Which Illustrate Three Types of Suppressor Variables. *Educational and Psychological Measurement*, **43**, 373-377. http://dx.doi.org/10.1177/001316448304300206

[9] Akinwande, M. O., Dikko H. G., &Gulumbe S. U (2015) Identifying the Limitation of Stepwise Selection for Variable Selection in Regression Analysis. *American Journal of Theoretical and Applied Statistics*, **4**, 414-419

[10] Velicer, W.F. (1978) Suppressor Variables and the Semipartial Correlation Coefficient. *Educational and Psychological Measurement*, **38**, 953-958. http://dx.doi.org/10.1177/001316447803800415

[11] 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*, **44**, 127-133. http://dx.doi.org/10.1353/csd.2003.0010

[12] Tzelgov, J. and Henik, A. (1991) Suppression Situations in Psychological Research: Definitions, Implications, and Applications. *Psychological Bulletin*, **109**, 524-536. http://dx.doi.org/10.1037/0033-2909.109.3.524

[13] Rosenberg, M. (1973) The Logical Status of Suppressor Variables. *Public Opinion Quarterly*, **37**, 359-372. http://dx.doi.org/10.1086/268098

[14] Nathans, L.L., Oswald, F.L. and Nimon, K. (2012) Interpreting Multiple Linear Regression: A Guidebook of Variable Importance. *Practical Assessment*, *Research & Evaluation*, **17**, 123-136.

[15] Paulhus, D.L., Robins, R.W., Trzesniewski, K.H. and Tracy, J.L. (2004) Two Replicable Suppressor Situations in Personality Research. *Multivariate Behavioral Research*, **39**, 303-328. http://dx.doi.org/10.1207/s15327906mbr3902_7

[16] Fox, J. (1991) Regression Diagnostics. Sage, Beverly Hills.

[17] MacNeill, S.E., Lichtenberg, P.A. and LaBuda, J. (2000) Factors Affecting Return to Living Alone after Medical Rehabilitation: A Cross-Validation Study. *Rehabilitation Psychology*, **45**, 356-364. http://dx.doi.org/10.1037/0090-5550.45.4.356

[18] Cohen, J., Cohen, P., West, S.G. and Aiken, L.S. (2003) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Lawrence Erlbaum, Mahwah.

[19] Pedhazur, E.J. (1997) Multiple Regression in Behavioral Research: An Explanation and Prediction. Holt, Rinehart & Winston, New York.

[20] Smith, R.L., Ager Jr., J.W. and Williams, D.L. (1992) Suppressor Variables in Multiple Regression/Correlation. *Educational and Psychological Measurement*, **52**, 17-29. http://dx.doi.org/10.1177/001316449205200102