DOMINANCE ANALYSIS: A NECESSITY OF PAYING ATTENTION TO RELATIVE IMPORTANCE OF PREDICTORS IN DECISION MAKING ISSUES

S. Yasaman Amirkiaee

University of North Texas
College of Business, 1155 Union Circle #311160, Denton, TX 76203-5017
(786) 863-0478
SeyedeYasaman.Amirkiaee@unt.edu

ABSTRACT

Multiple regression analysis can be a very useful prediction tool to predict the value of an unknown dependent variable from the known values of a set of predictor variables. Having the regression equation, one can determine how much a unit increase in a given predictor value increases the dependent variable value, while other predictor values is assumed to be fixed. On the other hand, in most cases in the real world, the predictors cannot be considered isolated as they are inter-correlated. Knowing relative importance of each predictor provides decision makers a stronger tool to utilize for example in strategy development, organizational planning, policy making and so on. This paper recalls dominance analysis method as a complement to traditional multiple regression analysis in case of correlated predictors and the necessity of taking it into account in decision making issues.

INTRODUCTION

One of the most popular statistical analysis and mostly used one in business science is multiple regression which tries to predict the behavior of a dependent variable based on a set of identified independent variables that can be called predictors. However, one might be more interested in relative importance of each predictor that can represent the extent of the contribution that the predictor provides to the prediction of dependent variable by itself and relative to others in the model (Johnson and LeBreton, 2004). It is not an easy task to do in most cases in the presence of correlated predictors (Tonidandel & LeBreton, 2011). During the years, some methods have been developed as the supplement to regression analysis to provide more explanation that how much of the variance explained in dependent variable can be attributed to each predictor variable. The traditional estimates of importance like standardized regression coefficients and simple bivariate correlations have been documented to fail in case of inter-correlations among the predictors (Tonidandel & LeBreton, 2011). While, there are other methods like dominance analysis (Budescu, 1993) and relative weight analysis (Fabbries, 1980) that can overcome this issue and deserve to get more notices from researchers in business science. In this paper, we focus on dominance analysis method to show the importance of its utilization in business science especially in decision making problems. We, also show that how this simple method can be a strong competitor to complicated methods which are using now for ranking the predictors according to their importance.

DOMINANCE ANALYSIS PROCEDURE

Dominance Analysis can be differentiated from other methods for identifying relative importance by its unique features: 1) it considers a pairwise fashion for measuring relative importance, 2) all relative subset models (i.e. $2^{(p-2)}$ subset models, where p is the number of predictors) is considered when two predictors are comparing.

The original Dominance Analysis procedure can be followed by these steps, as they can be perceived from Azen and Budescu (2003), based on (Budescu, 1993):

1- The portion variance in Y explained by the predictors in the model X (ρ^2_{YX}) is provided.

2- The additional contributions of each predictor to each subset model is computed; that is measured by the increase in $\rho^2_{Y\hat{Y}}$ that results from adding that predictor to regression model.

3- Complete dominance should be checked for each pair of the predictors. That is, demonstrate if the predictor's additional contribution to each of the subset models is greater than that of the other predictor.

- If complete dominance cannot be established, and the predictor has greater additional contribution than some of the other predictors and not all, it is said to be undermined dominance. In this case, it should be continued with the following step.

4- Conditional dominance should be checked for each predictors which considers the average additional contribution to all subset models of a given model size, i.e. the number of predictors included in the subset model (k).

- If conditional dominance cannot be established, and the predictor has greater average additional contribution than some of the other predictors and not all, it should be continued with the following step.

5- General dominance should be checked for each predictors to all subset models by the calculation of the average of all the conditional values, i.e. averaging the average of the additional contribution values.

For special cases where the presence of a specific predictor (or a set of predictors) is theoretically or practically necessary for the model, a constrained dominance analysis can be established (Azen and Budescu, 2003) by considering this step at the beginning of the dominance analysis:

- Only subset models that includes the intended specific predictor (or a set of predictors) should be considered for the analysis and the relevant comparisons should be conducted among these desired sets.

This procedure can be illustrated by the example that has been provided in Azen and Budescu (2003) for four predictors that describe the overall measure of satisfaction with life (Y). They might be considered as self, housing, finances, and health (X_1, X_2, X_3, and X_4 respectively). With

four predictors, 2^4=16 models are available. In the following example from the Table1, the first two columns show the possible models and their corresponding ρ^2_{YX} values.

The rest of the table represents the additional contribution of a given predictor ($\rho^2_{Y\hat{Y}}$). For example, $\rho^2_{Yx_1x_2}$= 0.450 represents the variance in Y explained by the model consisting of x_1 and x_2; and the additional contribution of x_1 to the subset model $\{x_2\}$ is defined as $\rho^2_{Yx_1x_2} - \rho^2_{Yx_2}$ = 0.450 - 0.090 = 0.360. That is the difference between the proportion of variance in Y explained by both x_1 and x_2 and the proportion of variance in Y explained by x_2 alone. It can be easily calculated for each of the remaining predictors when x_1 can be added (i.e., the null subset {.}, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_2x_3\}$, $\{x_2x_4\}$, $\{x_3x_4\}$ and $\{x_2x_3x_4\}$) and for other predictors, as well.

Table1. Dominance Analysis in the Population for Hypothetical Example with Four Predictors (Azen and Budescu, 2003, Page 136)

		Additional Contribution of			
Subset Model (X)	ρ^2_{YX}	X_1	X_2	X_3	X_4
Null and k= 0 average	0	0.36	0.09	0.16	0.25
X_1	0.360		0.090	0.117	0.113
X_2	0.090	0.360		0.138	0.223
X_3	0.160	0.317	0.068		
X_4	0.250	0.223	0.063	0.030	
k =1 average		0.300	0.074	0.095	0.152
X_1X_2	0.450			0.124	0.073
X_1X_3	0.477		0.097		0.037
X_1X_4	0.473		0.051	0.041	
X_2X_3	0.228	0.346			
X_2X_4	0.313	0.210		0.025	
X_3X_4	0.280	0.233	0.058		
k = 2 average		0.246	0.069	0.063	0.073
$X_1X_2X_3$	0.574				0.010
$X_1X_2X_4$	0.523			0.061	
$X_1X_3X_4$	0.513		0.071		
$X_2X_3X_4$	0.338	0.246			
k = 3 average		0.246	0.071	0.061	0.010
$X_1X_2X_3X_4$	0.584				
Overall Average		0.292	0.076	0.095	0.121

In the pair comparisons, based on the complete dominance definition, x_1 is shown to be dominated all other predictors. To have a meaningful comparison, for example between x_1 and x_4, we consider all subset models to which both x_1 and x_4 can make additional contribution i.e. {.}, $\{x_2\}$, $\{x_3\}$, $\{x_2x_3\}$. For all these sets, the additional contribution of x_1 is larger than of x_4, 0.360 compare with 0.250, 0.360 compare with 0.223, 0.317 compare with 0.120, and 0.346 compare with 0.110, respectively. Continuing these pair comparison reveals that x_1 is completely dominate all other

predictors, too. It is not the case for other predictors, here and complete dominance cannot be established between x_2 and x_3, x_2 and x_4, or x_3 and x_4.

So, the comparison should be continued to weaker levels of dominance, first conditionally dominance. Comparing the average additional contribution within each model size (k= 0, 1, 2, and 3) which is shown in the specific rows of Table 1, represents that x_1 conditionally dominates x_2, x_3, and x_4 in the pair comparisons. But again it is not the case for other predictors and the additional analysis is needed.

To check the general dominance, the average of all the conditional values are needed which are calculated and shown in the last row of Table 1. It is easy and always possible to put the overall averages in order unless there are similar values for a pair of predictors. In this case, it is obvious that x_1 generally dominates all predictors, which is followed by x_4, x_3, and x_2. Thus, in our particular example for measuring overall satisfaction, self-satisfaction is more important factor for explaining variance in this criterion, following by health, finance, and housing.
Also, it is worth noting that the value of $\rho^2_{Y\hat{Y}}$ is calculated by summation of overall averages of all predictors i.e. 0.292 + 0.076 + 0.095 + 0.121 = 0.584.

If the first predictor, x_1 (self-satisfaction) has been considered as a mandatory component of the model, the constrained dominance analysis should be conducted (Azen and Budescu, 2003). As it is shown in Table 2, just the subset models that are included x1, now are desired and kept. Based on the Table 2, here x_3 completely dominates x_4 but this cannot be established between x_2 and x_3. While the general dominance order among the predictors is x_3, x_2, and x_4. However, $\rho^2_{Y\hat{Y}}$ of this model remains the same as the full model and equals 0.584.

Table2. Constrained (x1 Included in All Models) Dominance Analysis in the population for Hypothetical Example with Four Predictors (Azen and Budescu, 2003, Page 139)

		Additional Contribution of		
Subset Model (X)	ρ^2_{YX}	X_2	X_3	X_4
(X_1) and k =1 average	0.36	0.09	0.117	0.113
(X_1)X_2	0.45		0.124	0.073
(X_1)X_3	0.477	0.097		0.037
(X_1)X_4	0.473	0.051	0.041	
k = 2 average		0.074	0.082	0.055
(X_1)X_2X_3	0.574			0.01
(X_1)X_2X_4	0.523		0.061	
(X_1)X_3X_4	0.513	0.071		
k = 3 average		0.071	0.061	0.01
(X_1)$X_2X_3X_4$	0.584			
Overall Average		0.078	0.087	0.059

The steps of Dominance Analysis procedure can be followed easily via Diagram 1.

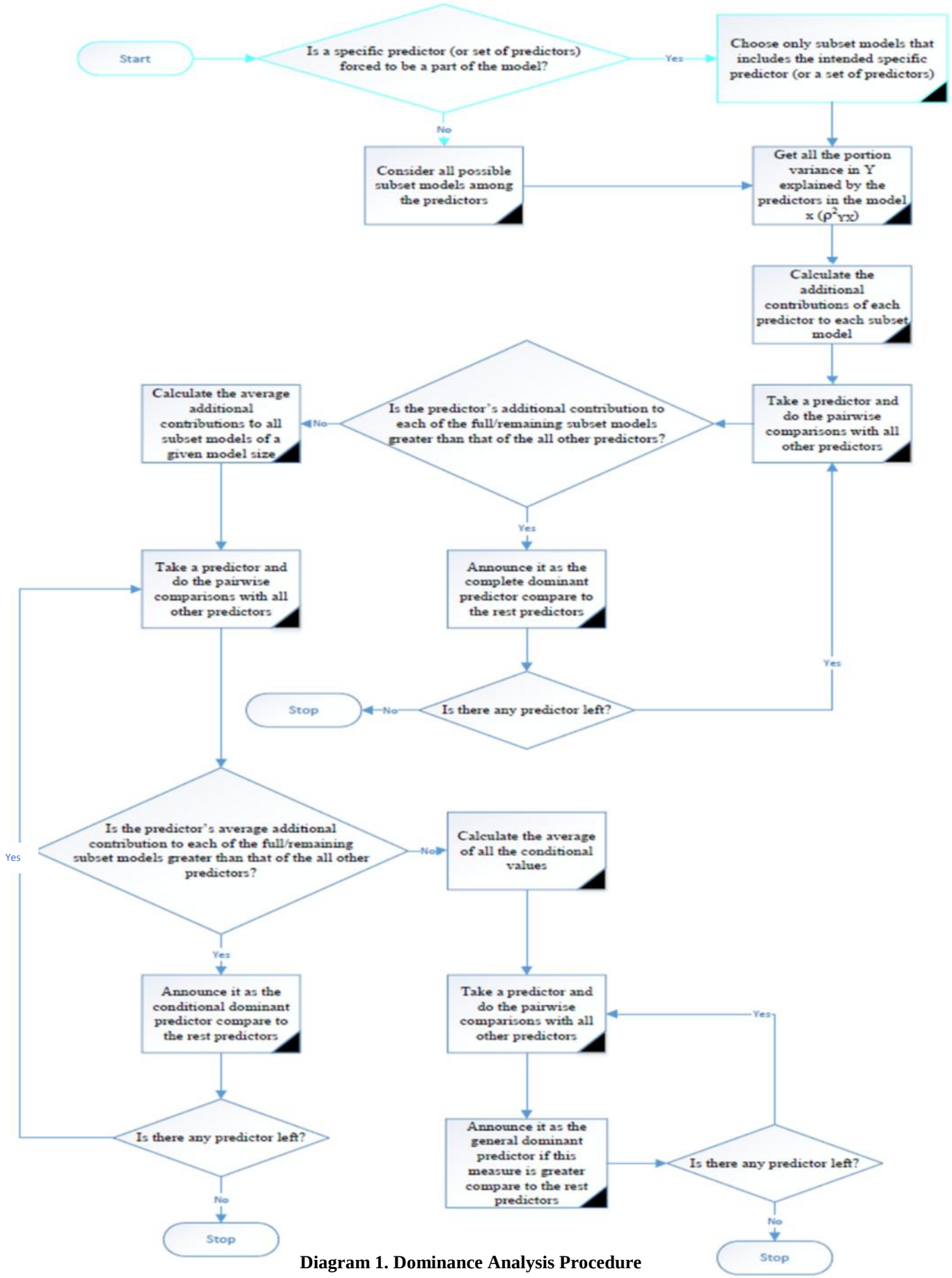

Diagram 1. Dominance Analysis Procedure

CASE STUDY

Here the results of running the Dominance Analysis on a data set is presented and discussed. The data set has been recently collected and denoted to the UCI Machine Learning repository by Fernandes, et al. (2015) for the initial purpose of the online news popularity prediction. Online popularity in the Web and social networks is often measured by number of interactions that can be exposure via shares, likes, comments. To this end, they analyzed the content of 39000 articles from the Mashable website, which is one of the largest news website. They provided an extensive list of features that could explain different aspects of the articles and were supposed to influence the number of interactions.

As a part of their results, they ranked the features according to their importance via Random Forest method (Liaw and Wiener, 2002), the first ten ranked features is provided in Table 3.

Table3. Ranking of Features According to Their Importance in the Random Forest Model (Fernandes, et al. 2015, page 543)

Feature*	Rank (#)
Avg. keyword (avg. shares)	0.0430 (1)
Avg. keyword (max. shares)	0.0398 (2)
Closeness to top 3 LDA topic	0.0320 (3)
Min number of shares of the referenced articles	0.0299 (4)
Ave number of shares of the referenced articles	0.0289 (5)
Closeness to top 2 LDA topic	0.0285 (6)
Closeness to top 5 LDA topic	0.0282 (7)
Best keyword (avg. shares)	0.0277(8)
Worst keyword (avg. shares)	0.0276 (9)
Closeness to top 1 LDA topic	0.0276 (10)

* Note: Name of features in the paper were inconsistent with the given dataset. So, this table has been revised based on the latest results that the first author of (Fernandes, et al. 2015) provided.

For the purpose of this paper, first these 10 features (predictors) has been examined in terms of correlation, which the results show that they are pretty correlated. Next, these 10 predictors are picked for running the dominance analysis in SAS to reveal the relative importance of each predictors. Table 4 shows the results of the dominance analysis (general dominance) on this data set which is a very close match to Random Forest method results.

Table 4. Dominance Analysis in the Population for the Popularity of Online News Data Set with Ten Predictors

Variable	Rank (#)
Avg. keyword (avg. shares)	.0082 (1)
Avg. keyword (max. shares)	.0023 (2)
Closeness to top 3 LDA topic	.0023 (3)
Min number of shares of the referenced articles	.0015 (4)
Ave number of shares of the referenced articles	.0013 (5)
Best keyword (avg. shares)	.0009 (6)
Closeness to top 2 LDA topic	.0005 (7)
Closeness to top 5 LDA topic	.0004 (8)
Worst keyword (avg. shares)	.0004 (9)
Closeness to top 1 LDA topic	.0003 (10)

CONCLUSION

The results of dominance analysis provides decisions makers with a powerful tool by representing the relative importance of each predictor in a multiple regression problem. It is especially important in the presence of big data set which includes inter-correlation components. The purpose of this paper was to emphasize the importance of this useful method in the context of decision science and provide the information and examples to facilitate utilizing of this method. The comparison of the dominance analysis results and the random forest method that is much more complicated method for this purpose, shows that dominance analysis can be a reliable simple method that gives similar results. So, it seems that it is time to recognize this simple method as a useful tool to improve decision making.

REFERENCES

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, *8*(2), 129.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*,*114*(3), 542.

Fabbris, L. (1980). Measures of predictor variable importance in multiple regression: An additional suggestion. *Quality & Quantity*, *14*(6), 787-792.

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In *Progress in Artificial Intelligence* (pp. 535-546). Springer International Publishing.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, *7*(3), 238-257.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest.*R news*, *2*(3), 18-22.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis: A useful supplement to regression analysis. *Journal of Business and Psychology*,*26*(1), 1-9.