# Using Explained Variance Allocation to analyse Importance of Predictors

Henri Wallard

CEDRIC, Centre d' Etude et de Recherche en Informatique et Communications.
CNAM CEDRIC, 292 rue Saint Martin ,75141 Paris , France.
Email :henriwallard@hotmail.com

**Abstract.** Using applications of linear regression, Market Research practitioners want to determine a ranking of predictors or a quantification of their respective importance for a desired outcome. As predictors are often correlated, regression coefficients can be difficult to use directly because they can be instable across samples and have negative values that are counterintuitive. To overcome these difficulties other methods have been proposed in the industry using squared semi partial correlation coefficients, squared zero order correlation coefficients or methods such as Shapley Value decomposition or decomposition via orthogonalisation in the space of predictors.

The proximity between the results obtained by different Variance Decomposition methods has led some authors to conclude that they are a fully valid approach. This paper will highlight theoretical reasons why these methods present similarities, offer a simple alternative new way to decompose variance but will also show the flaws and risks of relaying on Variance Decomposition for quantification of importance of predictors and why a Game Theory approach like Shapley Value can lead to misinterpretations. It will also present additional methods developed to compute β coefficients using Variance Decomposition as an intermediate step and propose recommendations for driver analysis.

**Keywords:** Variance Decomposition, Regression, Importance.

## 1 Introduction

In the field of Market Research, practitioners want to help their clients identify how to act on some factors such as quality of service or design of a product to achieve a desirable outcome such as purchase intent or satisfaction and loyalty of the customers. This is done with the desire to identify what are the best drivers of improvement, and quantify their respective impact. This is achieved through statistical modeling and simulation like in other fields such as Psychology, Social Sciences or Economics. The classic reference method used to do this Ordinary Least Square regression (OLS), but while this approach is recommended in the business literature it has some limitations. Because of sample size and design of the questions in Market research surveys, there may occur instability of coefficients. And also as questions can relate to similar topics there may be multi-collinearity and negative coefficients that are counter-intuitive. Causal assumptions and modeling options can lead to a variety of

*16rd ASMDA Conference Proceedings, 30 June – 4 July 2015, Piraeus, Greece*

results when one wants to quantify or simulate the impact of a given action on a predictor on the desired response. In Market Research many techniques are used: Regressions, Path Models, Bayesian Belief Networks, Random Forest to name a few. This paper focusses on relative importance in the context of Linear Regression.

## 2 The concept of relative importance

Johnson [1] points out that the terminology of relative importance is confusing because of the many different definitions used, and introduced the term “relative weights” to define the proportion of explained variance by the linear model allocated to each individual predictor. We propose first to define relative importance in a general way.
Let us consider p random variables as predictors and y a response. We can define a Relative Importance Function as a function that associates to each predictor (defined by its index j in the set P={1,…p}) a value of importance:

$$RI : \mathrm{P} \to \mathbb{R}$$

In practical applications, functions of relative importance are defined using matrix calculus and polynomial functions applied to the correlation matrix, explained variance of models and bivariate or multivariate correlation coefficients. These computations are applied using the estimates of these values. As a consequence we should like Grömping [2] clearly refer to estimators of relative importance. For instance we will see later that some relative importance functions are based on a full decomposition of the explained variance. As the estimator of explained variance in Linear Regression is biased, it is impossible that all estimators of relative importance for each predictor are unbiased as their sum is actually biased. This is why in all that follows we will only discuss the properties of estimators .The topic of relative importance has been discussed in many publications since at least 1936 and a history of the use of relative importance has been presented by Johnson and Lebreton [3]. This article will focus on relative importance evaluation based on allocation of shares of variance for linear regression. Grömping [2] gives an overview of Variance Decomposition methods. Some approaches allocate shares of Variance that can be negative and this has attracted criticism. Others propose the usage of values of relative importance that are all positive but do not add up to the total of explained variance. Lastly some methods fully decompose the explained variance across predictors. We will designate in this paper the methods that assign relative importance values to each predictor so that the sum of these relative importance adds up to the estimated R² as Variance Decomposition as opposed to Variance Allocation when the sum of the relative importance estimates is different from the estimated R².

In terms of notation we will use the following in reference to the Linear Regression Model and focus this article on allocation of shares of variance of the response y into proportions due to the p predictors X’s ( and errors) :

$$y = X\beta + e = \sum \beta_i X_i + e \quad (1)$$

We will work in the case of p predictors that are linearly independent, and in the case of n observations n greater than p and the n x p matrix of predictors score is of full rank p. The variance explained by the p predictors (P is the set of the p predictors) is:

$$V(P) = \sum_{i=1}^{i=p} \beta_i^2 v_i + 2\sum_{i<j} \beta_i \beta_j \sqrt{v_i v_j} \quad (2)$$

with vi variance of Xi and $\rho_{ij}$ coefficient of corrélation between Xi et Xj.

We can assume a regression model without intercept and w.l.o.g that all X's are centered (i.e. have expectation 0). To simplify the notations in the rest of the article we will assume, unless specified, that y and the Xj's are centered and standardized.

We have identified 8 methods published and included in R packages. A detailed documentation on 6 of these methods is available on Pr. Grömping's website [4] dedicated to resources on the relaimpo R package. Another method proposed by both Genizi [9] and Johnson [1] called here Relative Weights (RW's) and finally Zuber and Strimmer [5] CAR scores (Correlation-Adjusted (marginal) correlation) are also available via R packages (relaimpo and yhat). We will first present 3 methods of allocation and then 5 methods of decomposition and then discuss some points of difference and convergence between these approaches.

.

## 3 Methods for Variance Allocation

3. 1 Allocation "first".

The measures are the squared correlations of the predictors with the response:

$$first(j) = \mathrm{cov}(y, Xj)$$

When the predictors are mutually de-correlated the sum of the measures "first" adds up to the overall R² of the model. When this is not the case, the sum of the first (j) over all p predictors is often higher that the overall R² of the model.cf. Grômping [4].

3.2 Allocation "last".

This measure attributes as Relative Importance for a predictor j the increase in R² when predictor j is included last in the model compared to the R² with only the other p-1 predictors. This measure is identical to the squared semi-partial correlation sr²(j), which is sometimes presented as the amount by which the R² is reduced when this predictor is deleted from the regression equation. See for instance Tabachnick and Fidell [6].

3.3 Allocation “betasquared”.

This relative importance measure consists in attributing as importance the square of the standardized regression coefficient. Like the measures 3.1 and 3.2 these are variance allocations as the sum of these measures for all p predictors do not in general add up to the R².

## 4 Methods of Variance Decomposition

4.1 Decomposition Hoffman-Pratt

This measure of relative importance noted pratt(j) attributes to a predictor j the product of the standardized multiple regression coefficient by the marginal correlation between the predictor j and the response When the predictors are standardized :

$$pratt(j) = \beta_j \rho_{yj}$$

From the properties of the OLS regression we can easily confirm that this measure is leads to a decomposition of the R².

$$R^2 = \mathrm{cov}(y, \sum_j \beta_j Xj) = \sum_j \beta_j r_{yj} \qquad (3)$$

4.2 Shapley Value or LMG or Average

This method has been assigned several names. See for instance Grömping [2] for an historical overview. We will call this measure here lmg(j) or SV(j). This measure is computed by averaging on all possible ordering of the p predictors the increase of the R² when the predictor j is added to the model based on the other predictors entered before j in the model. These values have been proposed by Lindeman, Merenda and Gold (1980), hence the name lmg. If we consider a game theory perspective where we assimilate the p predictors as players and define the game function of a coalition of k players as the R² achieved by the model based on these k predictors, it turns out that the application of Shapley Value to the game described above generates exactly the same values as lmg,, hence the possible notation SV(j).

Let us present below one notation and one of the ways to compute lmg. Let r be a permutation of P, this constitutes an ordering of the predictors. Each permutation r enables to define an order of entry of the predictors in the model.

Let Sj(r) be the set of predictors entered before j in the permutation r. We can compute:

$R^2(S_j(r))$ as the R² of the model including the predictors in Sj(r)

$R^2_{+j}\ (S_j(r))$ as the R² of the model including the predictors in Sj(r) U{j}

And define $\Delta_j(r) = R^2_{+j}\ (S_j(r)) - R^2(S_j(r))$

$\Delta_j(r)$ is the increase in R² when the predictor j is added to the predictors entered before j in the model with order resulting from the permutation r.

$$lmg(j) = \frac{1}{p!}\sum_r \Delta_j(r) \tag{4}$$

averaged on all $2^p$ permutations of the p predictors. This formula can be rewritten in different forms, combining the permutations that have the same sets Sj(r).

4.3 PMVD (Proportional Marginal Variance Decomposition).

This measure is also a variance decomposition and is computed similarly as for lmg but with weights attached to each single permutation:

$$pmvd(j) = \frac{1}{p!}\,p(r)\sum_r \Delta_j(r) \tag{5}$$

For more details about PMVD see Feldman [7] and also Grömping [2].

4.4 Relative Weights

Fabbris [8] has proposed a way to decompose the explained variance using the Singular Value Decomposition of the matrix X. Later Genizi [9] and Johnson [1] used this approach in a different way. This decomposition is a particular case of a more general approach consisting of using a set of mutually uncorrelated variable to decompose the explained variance. We will formalize the orthogonal decomposition in general and then present the Relative Weights computation.

Let $z_i, i = 1,...p$ as set of p orthogonal standardized predictors:

Let us note $\lambda_{ji} = \mathrm{cov}(z_j, X_i)$ and $\beta_i = \mathrm{cov}(y, z_i)$

We compute the Orthogonal Decomposition RW with the zi's as follows

$$RW(j) = \sum_{i=1}^{i=p} \lambda_{ji}^2 \beta_i^2 \tag{6}$$

The Relative Weights generate a full variance decomposition because:

$$\sum_{j=1}^{j=p} RW(j) = \sum_{j=1}^{j=p}\sum_{i=1}^{i=p} \lambda_{ji}^2 \beta_i^2 = \sum_{i=1}^{i=p} \beta_i^2 \sum_{j=1}^{j=p} \lambda_{ji}^2 \quad (7)$$

$\lambda_{lm} = \mathrm{cov}(z_l, x_m)$, and the $z_i$ being a set of standardized orthogonal vectors and as the $x_j$ are also standardized finally:

$$\sum_{i=1}^{i=p} RW(\mathrm{i}) = V(y) = 1 \quad (8)$$

So the Relative Weights computed with any set of $z_i$ enables the computation of a full decomposition of the $R^2$ using the $RW_j$. The decomposition proposed by Genizi and Johnson consists in computing the Relative Weights using a specific set of orthogonal predictors that minimize the sum of the squares between each $X_j$ and $z_j$.

So in terms of variables minimizing: $\Psi = E[(z - X)'(\mathrm{z} - X)]$

In the case of a specific dataset with n observations and p predictors this leads to consider a specific matrix Z of the zj is as follows:

Let X be the n x p matrix of standardized centered observations. Let $X = P\Delta Q'$ the singular value decomposition of X. The set of zi minimizing the abovementioned sum of squares is $Z = PQ'$.The orthogonal decomposition using this specific set of orthogonal vectors are the Relative Weights. We will use the notation RW(j) from now on for this specific decomposition and Vo(j) in case we use another set of zi’s.

## 4.5 CAR scores

The CAR scores are the squared correlations between the response and the vectors Z as defined in 4.4.So: CAR(j)=$\lambda j^2$

This is a recent Variance Decomposition proposed by Zuber and Strimmer [10]. They use the term CAR standing for Correlation-Adjusted (marginal) coRelation.

We have limited the presentations of these methods to the strict minimum detail, but the documents in reference offer additional perspective on the Game Theory approach and axiomatic definitions of desirable properties in variance decomposition.

## 5 Results on Variance Decomposition

There are important difference between the usage of the Linear Model and the interpretation of Variance Decomposition values. In the case of lmg for instance, it is important not to use in a simplistic way the variance decomposition as if they were equivalent to the coefficients generated by Linear Regression Model. First because they are terms of variance that are actually homogeneous to squared values of the β's. If we consider an ideal case with mutually decorrelated predictors the lmg value would be distorted compared to the relative values generated by the Linear Model.
Also another way to write lmg in the case of two predictors is as follows:.

$$\mathrm{lmg}(1) = \frac{V(y) + (\beta_1{}^2 * \mathrm{v}_1 - \beta_2{}^2 * \mathrm{v}_2) * (1 - \rho_{12}{}^2)}{2} \qquad \mathrm{lmg}(2) = \frac{V(y) + (\beta_2{}^2 * \mathrm{v}_2 - \beta_1{}^2 * \mathrm{v}_1) * (1 - \rho_{12}{}^2)}{2}$$

If we consider a case with two predictors a sufficiently high correlation and a third predictor uncorrelated with the first two the reallocation of importance between the two lmg values can lead ultimately to different rankings between the importance measures if we apply the beta squared versus lmg.

So all in all lmg produces distortion and can potentially change the ranking between the importance of predictors versus the results derived from the Linear Model. This is why we need to be careful not to consider them as a full alternative to standard models. Conklin and Lipovetsky [11] have considered adjusting regression coefficients using Shapley Value as an intermediate step of calculation and computing coefficients in resolving a quadratic equation equalizing the Hoffman values and lmg for each predictor. Grömping and Landau [12] have criticized this approach.

Regarding similarities, Johnson and Lebreton [3] have observed the proximity between the results of relative weights and lmg and state :« *Despite being based on entirely different mathematical models, Johnson's epsilon and Budescu's dominance measures ( Note : Budescu's dominance is one of the denomination of Shapley Value /lmg ) provide nearly identical results when applied to the same data these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors. The convergence between these two mathematically different approaches suggests that substantial progress has been made toward furnishing meaningful estimates of relative importance among correlated predictors".*

We will first analyze and formulate results in the case of two predictors. Starting with two uncorrelated standardized variables E1 and E2, we will construct X1, X2, and y:

$$X_1 = \cos(\varphi)E_1 - \sin(\varphi)E_2 \qquad X_2 = \cos(\varphi)E_1 + \sin(\varphi)E_2$$
$$y = \cos(\psi)E_1 + \sin(\psi)E_2$$

From this we can actually compute:

$$\beta_1 = \frac{\sin(\varphi-\psi)}{\sin(2\varphi)} \qquad r_{y1} = \cos(\varphi+\psi)$$
$$r_{y2} = \cos(\varphi-\psi)$$
$$\beta_2 = \frac{\sin(\varphi+\psi)}{\sin(2\varphi)} \qquad \rho_{12} = \cos(2\varphi)$$

$$\text{last}(1) = \sin^2(\varphi-\psi);\text{first}(1) = \cos^2(\psi+\varphi)$$
$$\text{last}(2) = \sin^2(\psi+\varphi);\text{first}(2) = \cos^2(\psi-\varphi)$$

$$\text{SV}(1) = \frac{(1-\sin(2\varphi)\sin(2\psi))}{2}$$
$$\text{SV}(2) = \frac{(1+\sin(2\varphi)\sin(2\psi))}{2}$$

It is also possible to compute the result of orthogonal decomposition using any orthogonal base of the considered plane let us consider z1 and z2 such as:

$$z_1 = \cos(\omega)E_1 + \sin(\omega)E_2$$
$$z_2 = -\sin(\omega)E_1 + \cos(\omega)E_2$$

The results of an orthogonal decomposition process using the zi defined by the choice of a specific value of ω are Vo(1) and Vo(2) as computed below :

$$Vo(1) = \cos^2(\psi-\omega)\cos^2(\omega+\varphi) + \sin^2(\psi-\omega)\cos^2(\omega-\varphi)$$
$$Vo(2) = \cos^2(\psi-\omega)\sin^2(\omega+\varphi) + \sin^2(\psi-\omega)\sin^2(\omega-\varphi)$$

It is easy to demonstrate (w.l.o.g with φ≤π/2),that the specific $z_i$ considered earlier to implement the Relative Weights of the variance decomposition proposed by Johnson and Genizi, corresponds to the case when ω=-π/4. Taking ω=-π/4 in the computations of Vo(1) and Vo(2) and simplifying we get:

$$\text{RW}(1) = \frac{(1-\sin(2\varphi)\sin(2\psi))}{2};\text{RW}(2) = \frac{(1+\sin(2\varphi)\sin(2\psi))}{2}$$

We recognize here the formula for $SV(1)$ and $SV(2)$. So we have demonstrated through a trigonometric approach that in the case of two predictors the relative weights and the lmg (or Shapley Values) are identical (cf. also Thomas and al. [14]). This result is just the application of simple trigonometric equivalences and should not in our view lead to conclude that because the two methods converge this is in itself a justification of their validity. The demonstration proposed here enables easy visualization of the impact of the choice of orthogonalisation if we let ω vary. In the case of two predictors it also enables to demonstrate that the CAR scores may remain constant even when the correlation between predictors vary. We can also notice some other links between orthogonalisation procedures and lmg if we use some particular sets of orthogonal vectors in the space generated by the Xj's. As Relative Weights is a particular case of decomposition by orthogonalisation , these links help understand the proximity between lmg which is an averaging of last values over submodels and relative weights , which is a decomposition by orthogonalisation.

**Case 1:** Let us consider y* as the projection of y on the space of the predictors and let us choose one given predictor and an orthogonal set of zi's with the condition that:

$$zj = \frac{y^*}{\| y^* \|}$$

We have

$$\forall i \neq j, y.zi = 0 \qquad and \quad y^*.zj = 1$$

Let us now use the RW calculation formula:

$$\mathrm{Vo(j)} = \sum_{i=1}^{i=p} \lambda^2 \,\mathrm{ji}^* \beta^2 \,\mathrm{i} \ \text{ as:}$$

$$\lambda ji = \mathrm{cov(zj}, Xi) \quad \beta i = \mathrm{cov(y, zi)} \quad \beta j \neq 0; \beta j = 1$$

we have :

$$\mathrm{Vo(j)} = \mathrm{cov}^2(\mathrm{y, xj}) = \mathrm{first(j)}$$

This means that for any j there is always at least one choice of orthogonal decomposition that will allocate first (j) to that predictor.

**Case 2**: This time we will consider an orthogonal set so that:

$$zj = \frac{uj}{\| uj \|}$$

$u_j$ being the residual of the regression of $X_j$ on the other variables.

$$if \quad i \neq j \quad \lambda ji = \mathrm{cov}(\mathrm{zj}, Xi) = 0, \quad and \quad \lambda^2 jj = \mathrm{cov}^2(\mathrm{z}\, j, Xj) = 1 - \mathrm{R}_j^2$$

$$As \quad \beta^2 j = \mathrm{cov}^2(\mathrm{y}, \mathrm{zj}), \quad \mathrm{Vo(j)} = last(j)$$

These examples show that orthogonalisation methods do enable with specific sets of orthogonal vectors to allocate either first(j) or last(j) for one given predictor. As Johnson is a particular case of orthogonalisation and lmg (j) is an average of last (j) across submodels, it confirms why there can be a proximity between variance decomposition via orthogonalisation and lmg in the case of more than 2 predictors. Both lmg and Relative Weights are computer intensive methods. We introduce an alternative variance decomposition method that is much more computationally efficient and offers very similar results to lmg and relative weights. The method will allocate to each predictor j a share of variance that is a weighted average between first(j) and last (j) , hence the name weifila for weighted first last.

Here are the computation steps: let L and F be the sum of first and last for all predictors:

$$L = \sum_j last(j) \qquad F = \sum_j \mathrm{first}(j)$$

We will consider two cases in the usual situation where $L \neq F$

$$If \quad L < R^2 < F \quad \mathrm{W}(j) = last(j)\left(\frac{F - R^2}{F - L}\right) + first(j)\left(\frac{R^2 - L}{F - L}\right) \tag{9}$$

$$If \quad F < R^2 < L \quad \mathrm{W}(j) = last(j)\left(\frac{R^2 - F}{L - F}\right) + first(j)\left(\frac{L - R^2}{L - F}\right) \tag{10}$$

The case where R² would be outside of the interval between F and L is not encountered in practice. By construction in both cases above: $\sum w(j) = R^2$

We will note also that in the case of two predictors the weifila values equate to the lmg and relative weights values as shown below:

$$F = first(1) + first(2) = 1 + \cos(2\psi)\cos(2\varphi)$$

$$L = last(1) + last(2) = 1 - \cos(2\psi)\cos(2\varphi)$$

$$F - L = 2\cos(2\varphi)\cos(2\psi) \qquad w(j) = \frac{first(\mathrm{j}) + \mathrm{last(j)}}{2}$$

So we recognize the formula above for lmg and this confirms that:

$$w\left(j\right)=lmg\left(j\right)=SV\left(j\right)=RW\left(j\right) \qquad (11)$$

Weifila is a way to select an intermediate point w(j) inside the interval between first(j) and last(j) for each j. We have compared these 3 measures on two different datasets. The results are presented below:

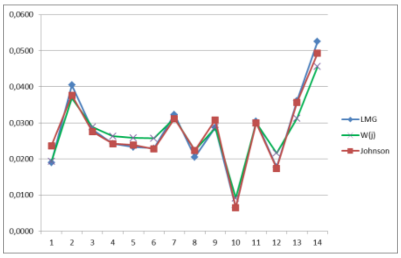

Fig 1: 1499 obsevations.14 predictors

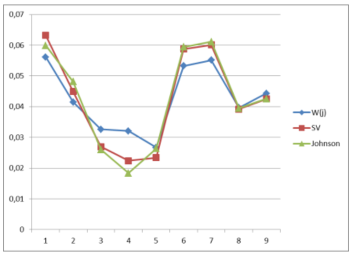

Fig 2: 499 observations 9 predictors

The weighted first last average "weifila" is much simpler to compute and delivers very similar results at least in the typical size of datasets and number of drivers used in practical applications for marketing and social research.

## 6 Conclusions

Among several methods to allocate variance among predictors, the proximity between lmg and Relative Weights has been noted (Johnson and Lebreton [3]), and seen as a justification of their validity. This proximity is actually a complete equality in the case of a model with two predictors and results from simple geometric properties. Also there exists variance decomposition via orthogonalisation that allocate exactly last(j) or first(j) for any of the predictors. So this proximity should not be seen in itself as a justification of validity.

The new method of variance decomposition proposed in this paper via a weighted average between first(j) and last(j) for each predictor provides very consistent results with lmg and Relative Weights but is simpler and less computer intensive. This method has been successfully tested with datasets typical of situations encountered in marketing research applications.

As underlined by Johnson [1] and Grömping [2], variance decomposition should not be seen as a substitute for linear regression models or path analytical models and models based on theory driven explanations can be more relevant than using directly variance decomposition. However when a model based on theory is not available variance decomposition can help identify important variables. Lastly the usage of modern machine learning techniques can also be considered.

## References

1. J. W.Johnson. A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression. Multivariate Behavioral Research,35(1) 1-19. 2000.
2. U.Grömping. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. The American Statistician, Vol 61 , No2 p139 2007.
3. J.W Johnson and J.M. Lebreton. History and Use of Relative Importance Indices in Organizational Research. *Organizational Research Methods* **7**, 238 - 257. . 2004.
4. http://prof.beuth-hochschule.de/groemping/relaimpo/
5. V.Zuber, and K.Strimmer. Variable importance and model selection by decorrelation. Preprint. http://arxiv.org/abs/1007.5516 (2010).
6. B. Tabachnick L. Fidell , Using Multivariate Statistics , Fifth Edition , 5.6.1.1 Pearson 2006.
7. B Feldman . "Relative Importance and Value." Manuscript version 1.1, 2005 http://www.prismanalytics.com/docs/RelativeImportance.pdf

8. L. Fabbris. Measures of regressor importance in multiple regression: an additional suggestion. *Qual Quant* 1980, 4:787–792

9. A. Genizi. Decomposition of $R^2$ in Multiple Regression with correlated regressors. Statiscica Sinica 4 , 407-420. 1993.

10. V. Zuber, K Strimmer. High-Dimensional Regression and Variable Selection Using CAR Scores. Statistical Applications in Genetics and Molecular Biology **10**: 34. 2011.

11. S**.**Lipovetsky, M. Conklin. Analysis of Regression in Game Theory Approach. *Applied Stochastic Models in Business and Industry* **17**, 319-330.2001

12. U Grömping , S.Landau.Do not adjust coefficients in Shapley Value Regression *Applied Stochastic Models in Business and Industry* **17**, 319-330.2009

13. U Grömping. Variable importance in regression models *WIRE's Comput Stat 2015,7:137-152. Doi:10.1002/wics.1346*

14. R D Thomas , Bruno D Zumbo, Ernest Kwan, Linda Schweitzer. On Johnson's (2000) Relative Weights Method for Assessing Variable Importance A Reanalysis : *Multivariate Behavioral Research July 18 2014*