Optimal Variable Subset Selection Problem in Regression Analysis is NP-Complete

Paolo Victor T. Redondo
School of Statistics
University of the Philippines Diliman

Combinatorial and optimization problems are classified into different complexity classes, e.g., whether an algorithm that efficiently solve the problem exists or a hypothesized solution to the problem can be quickly verified. The optimal selection of subset variables in regression analysis is shown to belong to a complexity class called *NP*-hard (Welch, 1982) in which solutions to the problems in the same class may not be easily (in terms of computing speed) proven optimal. Variable selection in regression analysis based on correlations is shown to be *NP*-hard, i.e., a complexity class of problems with easily verifiable solutions.

Keywords: optimal variable selection, regression analysis, np-completeness

1. Introduction

Despite the technological advancements in computing power, most existing combinatorial and optimization problems have not yet been found an efficient solution. This gives rise to one of the seven Millennium Problems selected by the Clay Mathematics Institute in 2000; the *P* versus *NP* Problem (de Figueiredo, 2012). Suppose a decision problem, i.e., a question answerable by "yes" or "no", is easily solvable. Such problem belongs to the complexity class *P* (polynomial time) where the time it takes to find the solution is simply a polynomial function of the size of the inputs. Consider the basic multiplication of two -digit numbers. Regardless of how large n is, the algorithm to efficiently solve the problem only requires n^2 single-digit multiplications (a polynomial in). This is considerably "quick" compared to some exponential functions, say 3^n. Other examples are string matching, number sorting and finding the maximum and minimum value in an array. On the other hand, the complexity class *NP* (non-deterministic polynomial time) includes all problems that are solvable in nondeterministic polynomial time, i.e., the problem cannot be answered quickly. However, *NP* problems are verifiable in polynomial time. This means that given a certain “guess” or “solution”, one can quickly validate its correctness. An example is the classic sudoku puzzle. Given a standard *nxn* sudoku with several entries filled, an answer to the puzzle cannot be easily generated. Yet given an answered sudoku, verification if the entries are all correct can be done quickly. Moreover, increasing the number of grids *n* in the sudoku puzzle exponentially increase computing time for finding the solution, thus, solvable in nondeterministic polynomial time.

Intuitively, the class *P* is in the class *NP* since problems that are easily solved have solutions that are easily verifiable. The converse, on the other hand, is yet to be proven. That is, are there problems whose answer can be quickly checked but requires an impossibly long time to solve by any direct method? Proving that all problems in class are also in class *NP* will pave the way towards exploration of finding the most efficient one out of infinitely many possible solutions for any existing problem while providing evidence to its contrary blocks the opportunity for some problems to be efficiently solved. This concept is widely applied to statistical methods.

Since statistics deals with uncertainties, arriving at an optimal solution for a certain problem can also be viewed in the context of being in class *P* or *NP*. For example, Welch (1982) proved three problems from computational statistics belong to the class *NP*-Hard; cluster analysis, subset selection in regression and D-optimal exact design of experiments.

The paper focuses only on the optimal subset selection problem in regression analysis. Also, since most statisticians are not familiar with some computer science concepts, a brief review of terminologies is included in section II. In Section III, an overview of the optimal subset selection problem is presented while the statistical aspect to achieve *NP*-completeness is elaborated and is connected to real-life application strategies in Section IV. An illustration through simulation and actual data is in Section V. Lastly, implications of the problem being *NP*-complete to practice are discussed in Section VI.

2. Theory of NP-Completeness and NP-Hardness

What if $P = NP$? Then all problems we view before to be difficult become easy. Proofs to all theorems become trivial. There will always be an efficient way to look around things. Production for manufacturers can improve in speed with much less waste in resources. All forms of transportation can be scheduled optimally to move everyone and everything quicker than it used to be (Fortnow, 2009). What a great future to look forward to given the current traffic congestion in EDSA!

However, most computer scientists expect the complement is true, that $P \neq NP$, despite not having a formal proof, and believe that the millennium problem will still not be proven in the near future (Fortnow, 2009; D. Johnson, 2012).

However, not having an efficient solution to every known problem has its merits too. Had $P = NP$, public-key cryptography will become impossible. The ability of two parties to send secure messages to each other without exchanging private keys can never happen because anyone who can intercept the message can decode its contents (Fortnow, 2009).

Consider a decision problem is in class *NP* but not necessarily in class *P*. That decision problem may have no efficient solution unless $NP - P$ is empty. In preparation to this, another complexity class, *NP*-complete, is defined. Suppose a problem in class *NP* is "reducible" to another *NP* problem *Y*, i.e., the instance/input *x* of the problem *X* can be transformed (polynomially) to the instance/input *y* of the problem *Y* such that the answer to *x* is "yes" if and only if the answer to *y* is "yes." Then, the complexity class *NP*-complete represents the set of all problems *X* in class *NP* for which it is possible to quickly reduce any other *NP* problem *Y* to *X* (Wilf, 2002). This makes -complete problems to be the hardest problems in the class *NP*. The implication of *NP*-completeness is that once an *NP*-complete decision problem is efficiently solved, then all other problems in *NP* will automatically have an efficient solution. Imagine having a single solution to all your problems (in *NP*); from the very first *NP*-complete problem, the satisfiability problem (see Cook, 1971), the recently proven *NP*-complete problems such as deciding the closure of inconsistent rooted triples (see M. Johnson, 2018) up to all *NP* problems imaginable.

Meanwhile, the complexity class *NP*-hard is an extension of the class *NP*-complete. Any problem *X* is *NP*-hard if there is an *NP*-complete problem *Y*, such that *Y* is quickly reducible to *X*. Thus, the class -hard comprises of problems that are at least as hard as the *NP*-complete problems although not necessarily in (D. Johnson, 2012). Not being in means solutions to the problem are no longer easily verifiable. Therefore, saying a statistical problem is -hard suggests no efficient algorithm can find the optimal solution, and at the same time, a solution given may not be verified optimal. An example of such is the optimal subset selection problem in regression analysis, shown by Welch (1982) to be *NP*-hard. Thus, the paper exposes the statistical aspect where the optimal subset selection becomes *NP*-complete, in order to validate optimality of the common practice in choosing variables in regression analysis.

3. Optimal Subset Selection in Regression Analysis

Consider an $n x 1$ data vector $\boldsymbol{Y}$, an design matrix $\boldsymbol{X}$ where n and k are positive integers. The objective of the optimal subset selection is to find a $k x 1$ vector $\hat{\boldsymbol{\beta}}$ with only q nonzero elements such that $q \leq k$ where the residual sum of squares, denoted by $R(\hat{\boldsymbol{\beta}}) = (Y - X\hat{\boldsymbol{\beta}})'(Y - X\hat{\boldsymbol{\beta}})$, is minimized. That is, finding the optimal set of variables among all that will explain most of the variability in the response variable $\boldsymbol{Y}$.

A "yes/no" decision analogue of this is by introducing another input B and ask whether there exists a k x 1 vector $\hat{\boldsymbol{\beta}}$ with only q nonzero elements such that $R(\hat{\boldsymbol{\beta}}) \leq B$ (Welch, 1982). By polynomially transforming the Minimum Weight Solution to Linear Equations (MWSTLE) which is shown by Garey and Johnson (1979) to be *NP*-complete, Welch (1982) shown the subset selection problem to be *NP*-Hard.

Clearly, the problem of finding the optimal subset of variables that minimizes the residual sum of squares cannot be solved quickly as the number of choices k for the variables increases. Similarly, given a subset of q variables and a vector of $\hat{\boldsymbol{\beta}}$ estimates, verification of its correctness entails checking all possible combinations of variables as regressors. Mathematically, one must regress the response with sets comprised of single variables, of $\binom{k}{2}, \binom{k}{3}, \ldots, \binom{k}{k-1}$ combinations of variables and with all variables which is a total of $2^k - 1$ possible scenarios. Therefore, the computing time exponentially increases as the k increases which is why the problem is viewed to be *NP*-hard; a problem difficult to solve with solutions not easily verifiable.

4. Inducing NP-Completeness in Statistical Sense

Given the subset selection problem is *NP*-Hard, in order to find a scenario for the problem to be *NP*-complete, we need to impose conditions on which a provided solution is easily verifiable. Suppose a solution $\hat{\boldsymbol{\beta}}$ is optimal, which implies that $R(\hat{\boldsymbol{\beta}}) = (Y - X\hat{\boldsymbol{\beta}})'(Y - X\hat{\boldsymbol{\beta}})$ is minimized.

Without loss-of-generality, let the first q variables be associated with the q nonzero entities in $\hat{\boldsymbol{\beta}}$. That is, for

$$\boldsymbol{Y} = [Y_1, \ldots, Y_n]',$$

$$\boldsymbol{X} = [\boldsymbol{X_1} \ldots \boldsymbol{X_q}\, \boldsymbol{X_{q+1}} \ldots \boldsymbol{X_k}] \text{ where } \boldsymbol{X_j} = [X_{i,j}] \text{ for } i = 1, \ldots, n,$$

$$\widehat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ldots \hat{\beta}_q \hat{\beta}_{q+1} \ldots \hat{\beta}_k]' \text{ where}$$

$$\hat{\beta} > 0, \forall j = 1, \ldots, q \text{ and } \hat{\beta}_j = 0, \forall j = q+1, \ldots, k$$

Thus, we are interested in minimizing

$$R(\hat{\boldsymbol{\beta}}) = (Y - X\hat{\boldsymbol{\beta}})'(Y - X\hat{\boldsymbol{\beta}})$$

$$= \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} \hat{\beta}_j X_{i,j} \right)^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} \hat{\beta}_j X_{i,j} \right)^2$$

Note that $R(\hat{\boldsymbol{\beta}})$, expressed as sum of squared values, can only be minimized if each squared term is minimized. Hence, the minimum value of $R(\hat{\boldsymbol{\beta}})$ is attained when $\left(Y_i - \sum_{j=1}^{q} \hat{\beta}_j X_{i,j}\right)^2 = B_i$ where B_i are some constant for all i = 1,..., n.

Now, the minimum value of $R(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} B_i = \sum_{i=1}^{n}\left(\sqrt{B_i}\right)^2 = \sum_{i=1}^{n}\left(\varepsilon_i\right)^2$ where $\varepsilon_i = \sqrt{B_i}$. Recall the property of the sample mean; the term $\sum_{i-1}^{n}\left(X_i - c\right)^2$ is minimized when c is the sample mean. Since $R(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}\left(\varepsilon_i - 0\right)^2$ is minimized, this forces the mean of ε_i to be zero. Thus, for $R(\hat{\boldsymbol{\beta}})$ to be minimized, $Y_i = \sum_{j=1}^{q} \beta_j X_{i,j} + \varepsilon_i$ where $E(\varepsilon_i) = 0$. In practice, we refer ε_i to as the random error term which by assumption in the linear regression model has three properties; i) independent ii) has mean zero and constant variance and iii) normally distributed (Montgomery, Peck and Vining, 2012). On the other hand, additional assumptions on the model considered by practitioners is that these error terms are also independent from any covariate $\boldsymbol{X}_j$ to ensure "validity" of the model and that any covariate $\boldsymbol{X}_j$ is independent from another covariate $X_{j'}$, where $j \neq j'$ so that information contributed by one variable is nonredundant upon considering any other variables. These assumptions comprise the statistical aspect where the optimal subset regression problem becomes *NP*-complete, i.e., a solution is easily verifiable.

For the problem to be *NP*-complete, since it has been proven by Welch (1982) that the problem is *NP*-Hard, it suffices to show that the problem belongs in class *NP*. This means we need to show that a provided solution is easily verifiable. However, finding out whether a "guess" attains the minimum value for $R(\hat{\boldsymbol{\beta}})$ cannot be done in polynomial time. Therefore, we need to visualize another cost function for the "yes/no" decision analogue of the problem based on some relationship measure of $\boldsymbol{Y}$ and $\boldsymbol{X}_j$**s**.

One of the frequently used method for searching the most appropriate covariates to be included in a linear regression model is the Exploratory Data Analysis (EDA) (Ratner, 2010). This method tries to visually find variables exhibiting linear relationship with the outcome variable Y. Thus, most statisticians incorporate variables having high correlation with their dependent variable Y in their model since correlation measures the strength of linear relationship between two continuous variables.

We now investigate the correlation of each independent variable $\boldsymbol{X}_j$ with the outcome variable $\boldsymbol{Y}$ through their respective covariances. Suppose $\boldsymbol{Y}$ truly has linear relationship with some covariates $\boldsymbol{X}_j$s plus some random error term with mean zero such that all X_js are independent from one another and the random error term ε is independent of any covariate X_j.

For $\forall\, j = 1,2,\ldots, q$,

$$Cov(Y, X_j) = Cov\,(\beta_1 X_1 + \cdots + \beta_j X_j + \cdots + \beta_q X_q + \varepsilon, X_j)$$

$$= \beta_1\, Cov(X_1, X_j) + \cdots + \beta_j Cov(X_j, X_j) + \cdots + \beta_q Cov(X_q, X_j) + Cov(\varepsilon, X_j)$$

$$= \beta_j Var(X_j) \neq 0 \quad \text{since } Cov(X_j, X_j) = Var(X_j),$$

$$Cov(X_{j'}, X_j) = 0, \forall\, j' \neq j \text{ and } Cov(\varepsilon, X_j) = 0 \text{ by assumption.}$$

On the other hand, when $j = q+1,\ldots,k$,

$$Cov(Y, X_j) = Cov\,(\beta_1 X_1 + \cdots + \beta_q X_q + \varepsilon, X_j)$$

$$= \beta_1 Cov(X_1, X_j) + \cdots + \beta_q Cov(X_q, X_j) + Cov(\varepsilon, X_j) = 0$$

since $Cov(X_{j'}, X_j) = 0$, $\forall\, j' \neq j$ and $Cov(\varepsilon, X_j) = 0$ by assumption.

But since the correlation $\rho(Y, X) = \dfrac{Cov(Y, X)}{\sqrt{Var((Y)Var(X)}}$ where the variances in the denominator are nonnegative, we have shown that $\rho(Y, X_j) \neq 0$, $\forall j = 1,\ldots, q$ while $\rho(Y, X_j) = 0$, $\forall j = q + 1, \ldots, k$.

Hence, minimizing ($\hat{\boldsymbol{\beta}}$) is similar to finding the estimate for the q nonzero $\hat{\beta}_j s$ via those covariates X_1, X_2, and X_q whose correlation with the dependent variable is nonzero. Thus, given a "guess" on $\hat{\boldsymbol{\beta}}$, we can verify easily if this is the optimal solution via checking if 1) the associated covariate s has a nonzero correlation with the dependent variable and 2) if the "guess" on $\hat{\boldsymbol{\beta}}$ is equal to the estimate under the ordinary least squares estimation which is known to minimize the sum of squared residuals. Hence, under the perfect scenario that Y can be truly expressed as $Y = \beta_1 X_1 + \cdots + \beta_q X_q + \varepsilon$ where $E(\varepsilon) = 0$ and the X_is and ε are independent, we have shown that the optimal subset selection problem is in class *NP* and hence, is *NP*-complete.

5. Illustrations

To illustrate the statistical aspect discussed above, two scenarios are simulated; $k = 3$ with $q = 1$ and $k = 5$ with $q = 2$. For simplicity, 500 samples are simulated for all explanatory variables X_js and the error term ε independently from a standard normal distribution.

Table 1. Correlation of all Explanatory Variables with the Response Variable

Variables	Correlation with Y	P-value
X_1	0.8944	3.61e-176
X_2	0.0000	1.00
X_3	0.0000	1.00

Table 2. Residual Sum of Squares for All Possible Combinations of Regressors

Regressing Y on	$R(\hat{\beta})$
X_1	499
X_2	2495
X_3	2495
X_1 and X_2	499
X_1 and X_3	499
X_2 and X_3	2495
X_1, X_2 and X_3	499

Table 3. OLS Coefficient Estimates for All Possible Combinations of Regressors

Regressing Y on	Estimated $(\hat{\beta})$
X_1	2.00
X_2	-5.96e-16
X_3	9.94e-16
X_1 and X_2	2.00 , -2.09e-16
X_1 and X_3	2.00 , 3.68e-16
X_2 and X_3	-5.96e-16, 1.33e-15
X_1, X_2 and X_3	2.00 , -2.09e-16 , 3.45e-16

For the first case, suppose $\boldsymbol{Y} = 2\boldsymbol{X}_1 + \boldsymbol{\varepsilon}$. Note that the correlation of $\boldsymbol{Y}$ with $\boldsymbol{X}_1$ is nonzero while correlation of $\boldsymbol{Y}$ with $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are both zero (see Table 1). Also, the minimum residual sum of squares is attained when regressing $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ (see Table 2) where $\boldsymbol{X}_1$ having the only nonzero coefficient estimate as given by the ordinary least squares (see Table 3).

On the other hand, for the second case, assuming $\boldsymbol{Y} = 2\boldsymbol{X}_1 - 1.5\boldsymbol{X}_2 + \boldsymbol{\varepsilon}$, the only correlation of $\boldsymbol{Y}$ with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are nonzero while correlation of $\boldsymbol{Y}$ with the rest are all zeroes (see Table 4). Again, the minimum residual sum of squares is attained when regressing $\boldsymbol{Y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (see Table 5) where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ having the only nonzero coefficient estimates as given by the ordinary least squares (see Table 6).

Table 4. Correlation of all Explanatory Variables with the Response Variable

Variables	Correlation with Y	P-value
X_1	0.7428	8.24e-89
X_2	-0.5571	4.21e-42
X_3	0.0000	1.00
X_4	0.0000	1.00
X_5	0.0000	1.00

Table 5. Residual Sum of Squares for All Possible Combinations of Regressors

Regressing Y on	$R(\hat{\beta})$	Regressing Y on	$R(\hat{\beta})$	Regressing Y on	$R(\hat{\beta})$	Regressing Y on	$R(\hat{\beta})$
X_1	1621.75	X_1, X_5	1621.75	X_1, X_2, X_4	499	X_3, X_4, X_5	3617.75
X_2	2495	X_2, X_3	2495	X_1, X_2, X_5	499	X_1, X_2, X_3, X_4	499
X_3	3617.75	X_2, X_4	2495	X_1, X_3, X_4	1621.75	X_1, X_2, X_3, X_5	499
X_4	3617.75	X_2, X_5	2495	X_1, X_3, X_5	1621.75	X_1, X_2, X_4, X_5	499
X_5	3617.75	X_3, X_4	3617.75	X_1, X_4, X_5	1621.75	X_1, X_3, X_4, X_5	1621.75
X_1, X_2	499	X_3, X_5	3617.75	X_2, X_3, X_4	2495	X_2, X_3, X_4, X_5	2495
X_1, X_3	1621.75	X_4, X_5	3617.75	X_2, X_3, X_5	2495	X_1, X_2, X_3, X_4, X_5	499
X_1, X_4	1621.75	X_1, X_2, X_3	499	X_2, X_4, X_5	2495		

Table 6. OLS Coefficient Estimates for All Possible Combinations of Regressors

Regressing Y on	Estimated ($\hat{\beta}$)	Regressing Y on	Estimated ($\hat{\beta}$)
X_1	2.00	X_1, X_2, X_4	2.00 , -1.50 , 3.65e-16
X_2	-1.50	X_1, X_2, X_5	2.00 , -1.50 , 1.24e-16
X_3	1.03e-15	X_1, X_3, X_4	2.00 , -2.58e-16 , -5.37e-16
X_4	1.36e-15	X_1, X_3, X_5	2.00 , -2.58e-16 , 9.94e-16
X_5	9.19e-16	X_1, X_4, X_5	2.00 , -8.75e-16 , 1.11e-16
X_1, X_2	2.00 , -1.50	X_2, X_3, X_4	-1.50 , 1.31e-15 , 2.17e-15

X_1, X_3	2.00 , -2.58e-16	X_2, X_3, X_5	-1.50 , 1.31e-15 , -5.37e-16
X_1, X_4	2.00 , -8.75e-16	X_2, X_4, X_5	-1.50 , 2.08e-15 , -7.95e-17
X_1, X_5	2.00 , 1.07e-15	X_3, X_4, X_5	1.03e-15 , 1.67e-15 , 8.35e-16
X_2, X_3	-1.50 , 1.31e-15	X_1, X_2, X_3, X_4	2.00 , -1.50 , 2.49e-16 , 3.38e-16
X_2, X_4	-1.50 , 2.09e-15	X_1, X_2, X_3, X_5	2.00 , -1.50 , 2.49e-16 , 1.14e-16
X_2, X_5	-1.50 , -2.98e-16	X_1, X_2, X_4, X_5	2.00 , -1.50 , 3.65e-16 , 1.34e-16
X_3, X_4	1.03e-15 , 1.67e-15	X_1, X_3, X_4, X_5	2.00 , -2.58e-16 , -5.367e-16 , 1.07e-15
X_3, X_5	1.03e-15 , 5.17e-16	X_2, X_3, X_4, X_5	-1.50 , 1.31e-15 , 2.17e-15 , -2.09e-16
X_4, X_5	1.36e-15 , 1.07e-15	X_1, X_2, X_3, X_4, X_5	2.00 , -1.50 , 2.49e-16 , 3.38e-16 , 1.39e-16
X_1, X_2, X_3	2.00 , -1.50 , 2.49e-16		

The simulation illustrates how a "solution" $\hat{\boldsymbol{\beta}}$ with only q nonzero entries can be easily verified to be the optimal solution. We have shown, mathematically, that the optimal set of q variables which accounts for the most variability present in the response $\boldsymbol{Y}$ are only those variables having nonzero correlation with $\boldsymbol{Y}$. Thus, instead of checking all $2^k - 1$ possible scenarios (exponential) for the regression analysis, we only need to consider finding all k correlations with $\boldsymbol{Y}$ which is still considerably fast even for large values of k. The implication of this to practitioners is simple yet important. The common practice of examining correlations before incorporating variables in a linear regression model becomes computationally optimal provided that the conditions mentioned are satisfied.

6. Conclusions

We have shown that the optimal subset selection problem in regression analysis is *NP*-complete given that $\boldsymbol{Y}$ truly has linear relationship with covariates $\boldsymbol{X}_j$s and on the random error term with mean zero assuming that the following conditions are satisfied: i) all $\boldsymbol{X}_j$s are independent from one another, and; ii) the random error term is independent on any covariate $\boldsymbol{X}_j$. If any of these two assumptions is violated, verification of the optimality of a certain "guess" will again require infinite amount of resources; an unwanted scenario in modelling. For example, if the assumption about the independence of the covariates is violated, multicollinearity in regression analysis arises where two variables share the same contribution on explaining variability in the dependent variable. This could cast doubt on appropriateness of the estimated regression model given by the ordinary least squares estimation due to the ill-conditioned $\hat{\boldsymbol{\beta}}$ estimates (unstable). Also, because of the dependencies among some of the $\boldsymbol{X}_j$s, it is possible to obtain more than q variables having nonzero correlation with $\boldsymbol{Y}$ which again makes the proposed "solution" to be nonverifiable (in polynomial time). This is commonly true in practice because the true relationship of the dependent variable with the covariates are seldom known. Nonetheless, by assuming such logical and practical relationship between the response and the covariates exists, practitioners are given more confidence in pursuing a parsimonious variable selection without the need to computationally exhaust all possibilities. Because when the problem becomes theoretically hard, it does not mean we cannot devise a technique to come up with a relatively good solution (Johnson, 2012).

Acknowlegement

I would like to thank Dr. Erniel Barrios and Dr. Henry Adorna for guiding me in the direction of this paper and for giving insightful comments to improve illustrations presented above.

REFERENCES

COOK, S., 1971, The complexity of theorem-proving procedures, *In Proceedings of the 3rd ACM Symposium on the Theory of Computing,* pp. 151-158. NY: ACM.

DE FIGUEIREDO, C., 2012, The P versus NP-complete dichotomy of some challenging problems in graph theory, *Discrete Applied Mathematics, 160(18)*:2681-2693.

FORTNOW, L. ,2009, The status of the P versus NP problem. *Communications of the ACM, 52(9)*:78-86.

GAREY, M., & JOHNSON, D., 1979, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* San Francisco: Freeman.

JOHNSON, D., 2012, A brief history of NP-completeness, 1954-2012, *Documenta Mathematica Extra Volume ISMP*, 359-376.

JOHNSON, M., 2018, Deciding the closure of inconsistent rooted triples is NP-complete. Available at: https://arxiv.org/pdf/1807.00030.pdf.

MONTGOMERY, D., PECK, E., & VINING, G., 2012, *Introduction to Linear Regression Analysis, 5th Ed.* New Jersey: John Wiley & Sons, Inc.

RATNER, B., 2010, Variable selection methods in regression: Ignorable problem, outing notable solution, *Journal of Targeting, Measurement and Analysis for Marketing, 18(1)*:65-75.

WELCH, W., 1982, Algorithmic Complexity: Three NP-Hard Problems in Computational Statistics, *Journal of Statistical Computation and Simulation*, 15(1): 17-25.

WILF, H., 2002, *Algorithms and Complexity, 2nd Ed.*. New York: Taylor and Francis Group.