

The following material is copyrighted material, belonging to Bruce Ratner, as found in his book *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC Press, Boca Raton, 2003. Neither the above titled book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the author. Used with permission.

Variable Selection Methods in Regression: Ignorable Problem, Outing Notable Solution Bruce Ratner, Ph.D.

Variable selection in regression – identifying the best subset among many variables to include in a model – is arguably the hardest part of model building. Many variable selection methods exist. Many statisticians know them, but few know they produce poorly performing models. The wanting variable selection methods are a miscarriage of statistics because they are developed by debasing sound statistical theory into a misguided pseudo-theoretical foundation. The purpose of this article is two-fold: 1) To resurface the scope of literature on the weaknesses of variable selection methods. 2) To enliven anew a notable solution to defining a substantially performing regression model. To tactically achieve my goal, I divide the article into two objectives: 1) To review five widely used variable selection methods, itemize some of their weaknesses, and answer why they are used. 2) To present Tukey's EDA that is relevant to the titled topic: The Natural Seven-step Cycle of Statistical Modeling and Analysis – the outed notable solution to variable selection in regression. I feel that newcomers to Tukey's EDA need the Seven-step Cycle introduced within the narrative of Tukey's analytic philosophy. Accordingly, I enfold the solution with front and back matter – The Essence of EDA, and The EDA School of Thought, respectively. John W. Tukey (June 16, 1915 – July 26, 2000) was a mega-contributor to the field of statistics, and was a humble, unpretentious man, as he always considered himself as *a data analyst*. Tukey's seminal book, *Exploratory Data Analysis* is uniquely known by the book's initialed title, EDA.

I. Background

Classic statistics dictates that the statistician sets about dealing with a given problem with a pre-specified procedure designed for that problem. For example, the problem of predicting a continuous target variable (e.g., profit) is solved by using the ordinary least squares (OLS) regression model *along with checking* the well-known underlying OLS assumptions. [1] At hand there are *several* candidate predictor variables, allowing a workable-task for the statistician to check assumptions (e.g., predictor variables are linearly independent). Likewise, the dataset has a *practica-ble* number of observations, making it also a workable-task for the statistician to check assumptions (e.g., the errors are uncorrelated). As well, the statistician can perform the *regarded yet often-discarded* exploratory data analysis (aka EDA), such as examine and apply the appropriate remedies for individual records that contribute to sticky data-characteristics (e.g., gaps, clumps, and outliers). Quite important, EDA allows the statistician to assess whether or not a given variable,

say, X needs a transformation/re-expression (e.g., $\log(X)$, $\sin(X)$ or $1/X$). The traditional variable selection methods cannot do such transformations or a priori construction of new variables from the original variables. [1.1] *Inability* of construction of new variables is a serious weakness of the variable selection methodology. [2]

Nowadays, building an OLS regression model or a logistic regression model (LRM where the target variable is binary: yes-no/1-0) is problematic because of the size of the dataset. Modelers work on *big data* – consisting of a *teeming multitude* of variables, and an *army* of observations. The workable-tasks are no longer feasible. Modelers cannot sure-footedly use OLS regression and LRM on big data, as the two statistical regression models were conceived, testing and experimented within the *small-data setting* of the day over 50 and 205 years ago, for LRM and OLS regression, respectively. The regression theoretical foundation and the tool of significance testing employed on big data are without statistical binding force. Thus, fitting big data to a pre-specified *small-framed* model produces a *skewed* model with doubtful interpretability and questionable results.

Folklore has it that the knowledge and practice of variable selection methods were developed when small data slowly grew into the early size of big data circa late 1960s/early 1970s. With only a single bibliographic citation ascribing variable selection methods to unsupported notions, I believe a reasonable scenario of the genesis of the methods was as follows: [3] College statistics nerds (intelligent thinkers) and computer science geeks (intelligent doers) put together the variable selection methodology using a *trinity of selection-components*:

- 1) Statistical tests (e.g., F, chi-square, and t tests, and significance testing)
- 2) Statistical criteria (e.g., R-squared, adjusted R-squared, Mallows' C_p and MSE [3.1])
- 3) Statistical stopping rules (e.g., p-values *flags* for variable entry/deletion/staying in a model)

The created body of unconfirmed thinking about the newborn-developed variable selection methods was on bearing soil of expertness and adroitness in computer-automated, misguided statistics. The trinity distorts its components' original theoretical and inferential meanings when they are framed within the newborn methods. The statistician executing the computer-driven trinity of statistical apparatus in a seemingly intuitive, insightful way gave proof – *face validity* – that the problem of variable selection, aka subset selection, was solved (at least to the uninitiated statistician).

The newbie subset selection methods initially enjoyed wide acceptance with extensive use, and presently still do. Statisticians build *at-risk* accurate and stable models – either *unknowingly* using these unconfirmed methods or *knowingly* exercise these methods because *they know not what to do*. It was not long before these methods' weaknesses, some contradictory, generated many commentaries in the literature. I itemize nine ever-present weaknesses, below, for two of the traditional variable selection methods, *All-subset*, and *Stepwise*. I concisely describe the five frequently used variable selection methods in the next section.

1. For All-subset selection with more than 40 variables: [3]
 - a. The number of possible subsets can be huge.
 - b. Often, there are several good models, although some are unstable.
 - c. The best X variables may be no better than random variables, if size sample is relatively small to the number of all variables.

- d. The regression statistics and regression coefficients are biased.
2. All-subset selection regression can yield models that are *too small*. [4]
3. Why the number of candidate variables and not the number in the final model is the number of degrees of freedom to consider. [5]
4. The data analyst knows more than the computer ... and failure to use that knowledge produces inadequate data analysis. [6]
5. Stepwise selection yields confidence limits that are far too narrow. [7]
6. Regarding frequency of obtaining authentic and noise variables ... The degree of correlation among the predictor variables affected the frequency with which authentic predictor variables found their way into the final model. The number of candidate predictor variables affected the number of noise variables that gained entry to the model. [8]
7. Stepwise selection will not necessarily produce the best model if there are redundant predictors (common problem). [9]
8. There are two distinct questions here: (a) When is Stepwise selection appropriate? And (b) Why is it so popular? [10]
9. As to question (b) above ... there are two groups that are inclined to favor its usage. One consists of individuals, with little formal training in data analysis, which confuses knowledge of data analysis with knowledge of the syntax of SAS, SPSS, etc. They seem to figure that *if its there in a program, its gotta be good and better than actually thinking about what my data might look like*. They are fairly easy to spot and to condemn in a right-thinking group of well-trained data analysts. However, there is also a second group who is often well trained They believe in statistics ... given any properly obtained database, a suitable computer program can objectively make substantive inferences without active consideration of the underlying hypotheses. ... *Stepwise selection is the parent of this line blind data analysis* [11]

Currently, there is *burgeoning* research that continues the original efforts of subset selection by shoring up its pseudo-theoretical foundation. It follows a line of examination that adds assumptions and makes modifications for eliminating the weaknesses. As the traditional methods are being mended, there are innovative approaches with starting points far afield from their traditional counterparts. There are freshly minted methods, like the *enhanced variable selection method* built in the GenIQ Model, constantly being developed. [12] [13] [14] [15]

II. Introduction

Variable selection in regression – identifying the best subset among many variables to include in a model – is arguably hardest part of model building. Many variable selection methods exist because it provides a solution to one of the most important problems in statistics. [16] [17] Many

statisticians know them, but few know they produce poorly performing models. The wanting variable selection methods are a *miscarriage of statistics* because they are developed by debasing sound statistical theory into a misguided pseudo-theoretical foundation. They are executed with computer-intensive search heuristics guided by rules-of-thumb. Each method uses a unique trio of elements, one from each component of the trinity of selection-components. [18] Different sets of elements typically produce different subsets. The number of variables in common with the different subsets is small, and the sizes of the subsets can vary considerably.

An alternative view of the problem of variable selection is to examine certain subsets and select the best subset, which either maximizes or minimizes an appropriate criterion. Two subsets are obvious – the best single variable and the complete set of variables. The problem lies in selecting an intermediate subset that is better than both of these extremes. Therefore, the issue is how to find the *necessary variables* among the complete set of variables by deleting both *irrelevant variables* (variables not affecting the dependent variable), and *redundant variables* (variables not adding anything to the dependent variable). [19]

I review five frequently used variable selection methods. These *everyday* methods are found in major statistical software packages. [20] The test-statistic for the first three methods uses either the F statistic for a continuous dependent variable, or the G statistic for a binary dependent variable. The test-statistic for the fourth method is either R-squared for a continuous dependent variable, or the Score statistic for a binary dependent variable. The last method uses one of the criteria: R-squared, adjusted R-squared, Mallows' C_p .

1. Forward Selection (FS) - This method adds variables to the model until no remaining variable (outside the model) can add anything significant to the dependent variable. FS begins with no variable in the model. For each variable, the test-statistic (TS), a measure of the variable's contribution to the model, is calculated. The variable with the largest TS value that is greater than a preset value C is added to the model. Then the test-statistics is calculated again for the variables still remaining, and the evaluation process is repeated. Thus, variables are added to the model one by one until no remaining variable produces a TS value that is greater than C . Once a variable is in the model, it remains there.
2. Backward Elimination (BE) - This method deletes variables one by one from the model until all remaining variables contribute something significant to the dependent variable. BE begins with a model which includes all variables. Variables are then deleted from the model one by one until all the variables remaining in the model have TS values greater than C . At each step, the variable showing the smallest contribution to the model (i.e., with the smallest TS value that is less than C) is deleted.
3. Stepwise (SW) - This method is a modification of the forward selection approach and differs in that variables already in the model do not necessarily stay. As in Forward Selection, SW adds variables to the model one at a time. Variables that have a TS value greater than C are added to the model. After a variable is added, however, SW looks at all the variables already included to delete any variable that does not have a TS value greater than C .

4. R-squared (R-sq) - This method finds several subsets of different sizes that best predict the dependent variable. R-sq finds subsets of variables that best predict the dependent variable based on the appropriate TS. The best subset of size k has the largest TS value. For a continuous dependent variable, TS is the popular measure R-squared, the coefficient of multiple determination, which measures the proportion of the *explained* variance in the dependent variable by the multiple regression. For a binary dependent variable, TS is the theoretically correct but less-known Score statistic [21]. R-sq finds the best one-variable model, the best two-variable model, and so forth. However, it is unlikely that one subset will stand out as clearly being the best, as TS values are often bunched together. For example, they are equal in value when rounded at the, say, third place after the decimal point. [22] R-sq generates a number of subsets of each size, which allows the user to select a subset, possibly using nonstatistical conditions.
5. All-possible Subsets – This method builds all one-variable models, all two-variable models, and so on, until the last all-variable model is generated. The method requires a powerful computer (because a lot of models are produced), and selection of any one of the criteria: R-squared, adjusted R-squared, Mallows' C_p .

III. Weakness in the Stepwise

An ideal variable selection method for regression models would find one or more subsets of variables that produce an *optimal* model. [22.1] Its objectives are that the resultant models include: accuracy, stability, parsimony, interpretability, and avoid bias in drawing inferences. Needless to say, the above methods do not satisfy most of these goals. Each method has at least one drawback specific to its selection criterion. In addition to the nine weaknesses mentioned above, I itemize a compiled list of weaknesses of the *most popular Stepwise* method. [23]

1. It yields R-squared values that are badly biased high.
2. The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
3. The method yields confidence intervals for effects and predicted values that are falsely narrow.
4. It yields p-values that do not have the proper meaning and the proper correction for them is a very difficult problem.
5. It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large).
6. It has severe problems in the presence of collinearity.
7. It is based on methods (e.g., F tests) that were intended to be used to test pre-specified hypotheses.
8. Increasing the sample size doesn't help very much.
9. It allows us to not think about the problem.
10. It uses a lot of paper.

11. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.

I add to the tally of weaknesses by stating *common* weaknesses in regression models, as well as those specifically related to OLS regression model and LRM:

The everyday variable selection methods in regression model typically results in models having too many variables, an indicator of overfitted. The prediction errors, which are inflated by outliers, are not stable. Thus, model implementation results in unsatisfactory performance. For ordinary least squares regression, it is well known in the absence of normality or absence of linearity assumption or outlier(s) presence in the data, variable selection methods poorly perform. For logistic regression, the reproducibility of the computer-automated variable-selection models is unstable and not reproducible. The variables selected as predictor variables in the models are sensitive to unaccounted for sample variation in the data.

Given the litany of weaknesses cited, the lingering question is: Why do statisticians use variable selection methods to build regression models? To paraphrase Mark Twain: “Get your [data] first, and then you can distort them as you please.” [23.1] The author’s answer is: “Modelers use variable selection methods every day *because they can.*” As a counterpoint to the absurdity of “because they can,” I enliven anew Tukey’s solution of Natural Seven-step Cycle of Statistical Modeling and Analysis to defining a substantially performing regression model. I feel that newcomers to Tukey’s EDA need the Seven-step Cycle introduced within the narrative of Tukey’s analytic philosophy. Accordingly, I enfold the solution with front and back matter – The Essence of EDA, and The EDA School of Thought, respectively. I delve into the trinity of Tukey’s masterwork; but first I discuss, below, an enhanced variable selection method, for which I might be the only exponent for appending this method to the current baseless arsenal of variable selection.

IV. Enhanced Variable Selection Method

In lay terms, the variable-selection problem in regression can be stated:

Find the best combination of *the original variables* to include in a model. The variable selection method *neither states nor implies* that it has an *attribute to concoct new variables stirred up by mixtures of the original variables.*

The attribute – data mining – is either overlooked, perhaps, because it is reflective of the simple-mindedness of the problem-solution at the onset, or is currently sidestepped as the problem is too difficult to solve. A variable selection method without a data mining attribute obviously hits a *wall*, which beyond it would otherwise increase the predictiveness of the technique. In today’s terms, the variable selection methods are *without* data mining capability. They cannot dig the data for the mining of potentially important new variables. (This attribute, which has never surfaced during my literature search, is a partial mystery to me.) Accordingly, I put forth a definition of an enhanced variable selection method:

An enhanced variable selection method is one that identifies a subset that consists of the original variables *and* data-mined variables, whereby *the latter are a result of the data-mining attribute of the method itself*.

The following five discussion-points clarify the attribute-weakness, and illustrate the concept of an enhanced variable selection method.

1. Consider the complete set of variables, X_1, X_2, \dots, X_{10} . Any of the current variable selection in use finds the best combination of the original variables (say X_1, X_3, X_7, X_{10}); but, it can never automatically transform a variable (say transform X_1 to $\log X_1$) if it were needed to increase the *information content (predictive power)* of that variable. Furthermore, none of the methods can generate a re-expression of the original variables (perhaps X_3/X_7) if the constructed variable, structure, were to offer more predictive power than the original component variables combined. In other words, current variable selection methods cannot find an *enhanced subset*, which needs, say, to include transformed and re-expressed variables (possibly $X_1, X_3, X_7, X_{10}, \log X_1, X_3/X_7$). A subset of variables without the potential of new structure offering more predictive power clearly limits the modeler in building the best model.
2. Specifically, the current variable selection methods fail to identify structure of the types discussed here. *Transformed variables* with a *preferred* shape. A variable selection procedure should have the ability to transform an individual variable, if necessary, to induce symmetric distribution. Symmetry is the preferred shape of an individual variable. For example, the workhorse of statistical measures – the mean and variance – is based on symmetric distribution. Skewed distribution produces inaccurate estimates for means, variances, and related statistics, such as the correlation coefficient. Symmetry facilitates the interpretation of the variable's effect in an analysis. Skewed distribution are difficult to examine because most of the observations are bunched together at one end of the distribution. Modeling and analyses based on skewed distributions typically provide a model with doubtful interpretability and questionable results.
3. The current variable selection method also should have the ability to *straighten* nonlinear relationships. A linear or straight-line relationship is the *preferred* shape when considering two variables. A straight-line relationship between independent and dependent variables is an assumption of the popular statistical linear regression models (e.,g., OLS regression and LRM). (Remember that, a linear model is defined as a sum of weighted variables, such as $Y = b_0 + b_1 * X_1 + b_2 * X_2 + b_3 * X_3$.) [24] Moreover, straight-line relationships among all the independent variables is a *desirable property*. [25] In brief, straight-line relationships are easy to interpret: A unit of increase in one variable produces an expected constant increase in a second variable.
4. *Constructed variables* from the original variables using simple arithmetic *functions*. A variable selection method should have the ability to construct simple re-expressions of the the original variables. Sum, difference, ratio, or product variables potentially offer more information than the original variables themselves. For example, when analyzing the efficiency of an automobile engine, two important variables are miles traveled and fuel used

(gallons). However, we know the ratio variable of miles per gallon is the best variable for assessing the engine's performance.

5. *Constructed variables* from the original variables using *a set of functions* (e.g., arithmetic, trigonometric, and/or Boolean functions). A variable selection method should have the ability to construct complex, re-expressions with mathematical functions that capture the complex relationships in the data, thusly, potentially offer more information than the original variables themselves. In an era of data warehouses and the internet, big data consisting of hundreds of thousands-to-millions of individual records and hundreds-to-thousands of variables are commonplace. Relationships among many variables produced by so many individuals are sure to be complex, beyond the simple straight-line pattern. Discovering the mathematical expressions of these relationships, although difficult although practical guidance exist, should be the hallmark of a high-performance variable selection method. For example, consider the well-known relationship among three variables: The lengths of the three sides of a right triangle. A powerful variable selection procedure would identify the relationship among the sides, even in the presence of measurement error: The longer side (diagonal) is the square root of the sum of squares of the two shorter sides.

In sum, the attribute-weakness implies: *A variable selection method should have the ability of generating an enhanced subset of candidate predictor variables.*

V. EDA

I present the trinity of Tukey's EDA that is relevant to the proposed topic: a) The Essence of EDA, b) The Natural Seven-step Cycle of Statistical Modeling and Analysis, serving as the outed notable solution to variable selection in regression, and c) The EDA School of Thought.

a) The Essence of EDA is best described in Tukey's own words: "Exploratory data analysis is detective work – numerical detective work –or counting detective work – or graphical detective work... [It is] about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights." EDA includes the following characteristics:

1. Flexibility - techniques with greater flexibility to delve into the data
2. Practicality - advice for procedures of analyzing data
3. Innovation - techniques for interpreting results
4. Universality - use all of statistics that apply to analyzing data
5. Simplicity – above all, the belief that simplicity is the golden rule

The professional statistician has also been empowered by the personal computer (PC) computational strength, which without the natural seven-step cycle of statistical modeling and analysis would not be possible. The PC and the analytical cycle comprise the perfect pairing as long as the steps are followed in order and the information obtained from one-step is used in the

next step. Unfortunately, statisticians are human and succumb to taking shortcuts through the seven-step cycle. They ignore the cycle and focus solely on the sixth step listed below. However, careful statistical endeavor requires additional procedures, as described in the originally outlined seven-step cycle that follows [25.1]:

b) The Natural Seven-step Cycle of Statistical Modeling and Analysis

1. Definition of the problem.

Determining the best way to tackle the problem is not always obvious. Management objectives are often expressed qualitatively, in which case the selection of the outcome or target (dependent) variable is subjectively biased. When the objectives are clearly stated, the appropriate dependent variable is often not available, in which case a surrogate must be used.

2. Determining technique.

The technique first selected is often the one with which the data analyst is most comfortable; it is not necessarily the best technique for solving the problem.

3. Use of competing techniques.

Applying alternative techniques increases the odds that a thorough analysis is conducted.

4. Rough comparisons of efficacy.

Comparing variability of results across techniques can suggest additional techniques or the deletion of alternative techniques.

5. Comparison in terms of a precise (and thereby inadequate) criterion.

Explicit criterion is difficult to define; therefore, precise surrogate are often used.

6. Optimization in terms of a precise and similarly inadequate criterion.

Explicit criterion is difficult to define; therefore, precise surrogates are often used.

7. Comparison in terms of several optimization criteria.

This constitutes the final step in determining the best solution.

The founding fathers of classical statistics – Karl Pearson and Sir Ronald Fisher would have delighted in the PCs ability to free them from time-consuming empirical validations of their concepts. Pearson, whose contributions include regression analysis, the correlation coefficient, the standard deviation (a term he coined), and the chi-square test of statistical significance, would have likely developed even more concepts with the free time afforded by the PC. One can further speculate that the PCs functionality would have allowed Fisher's methods of maximum likelihood estimation, hypothesis testing, and analysis of variance to have immediate, practical applications.

c) The EDA School of Thought

Tukey's book is more than a collection of new and creative rules and operations; it defines EDA as a discipline that holds that data analysts fail only if they fail to try many things. It further espouses the belief that data analysts are especially successful if their detective work forces them to notice the unexpected. In other words, the philosophy of EDA is a trinity of *attitude* and *flexibility* to do whatever it takes to refine the analysis, and *sharp-sightedness* to observe the unexpected when it does appear. EDA is thus a self-propagating theory; each data analyst adds his or her own contribution, thereby contributing to the discipline, as I hope to accomplish with this book.

The sharp-sightedness of EDA warrants more attention, as it is a very important feature of the EDA approach. The data analyst should be a keen observer of those indicators that are capable of being dealt with successfully, and use them to paint an analytical picture of the data. In addition to the ever-ready visual graphical displays as an *indicator* of what the data reveal, there are numerical indicators, such as counts, percentages, averages and the other classical descriptive statistics (e.g., standard deviation, minimum, maximum and missing values). The data analyst's personal judgment and interpretation of indicators are not considered a bad thing, as the goal is to draw informal inferences, rather than those statistically significance inferences that are the hallmark of statistical formality.

In addition to visual and numerical indicators, there are the *indirect messages* in the data that force the data analyst to take notice, prompting responses such as "the data look like..." or, "it appears to be" Indirect messages may be vague; but their importance is to help the data analyst draw informal inferences. Thus, indicators do not include any of the hard statistical apparatus, such as confidence limits, significance test, or standard errors.

With EDA, a new trend in statistics was born. Tukey and Mosteller quickly followed up in 1977 with the second EDA book (commonly referred to EDA II), *Data Analysis and Regression*, which recasts the basics of classical inferential procedures of data analysis and regression as an assumption-free, nonparametric approach guided by "(a) a sequence of philosophical attitudes ... for effective data analysis, and (b) a flow of useful and adaptable techniques that make it possible to put these attitudes to work." [26]

Hoaglin, Mosteller and Tukey in 1983 succeeded in advancing EDA with *Understanding Robust and Exploratory Analysis*, which provides an understanding of how badly the classical methods behave when their restrictive assumptions do not hold, and offers alternative robust and exploratory methods to broaden the effectiveness of statistical analysis. [27] It includes a collection of methods to cope with data in an informal way, guiding the identification of data structures relatively quickly and easily, and trading off optimization of objective for stability of results.

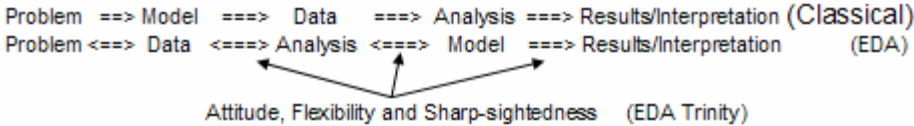
Hoaglin, Mosteller and Tukey in 1991 continued their fruitful EDA efforts with *Fundamentals of Exploratory Analysis of Variance*. [28] They recast the basics of the analysis of variance with the classical statistical apparatus (e.g., degrees of freedom, F ratios and p-values) in a host of numerical and graphical displays, which often give insight into the

structure of the data, such as sizes effects, patterns and interaction, and behavior of residuals.

EDA set off a burst of activity in the visual portrayal of data. Published in 1983, *Graphical Methods for Data Analysis* (Chambers et al) presents new and old methods - some which require a computer, while others only paper and pencil - but all are powerful data analysis tools to learn more about data structure. [29] In 1986 du Toit et al came out with *Graphical Exploratory Data Analysis*, providing a comprehensive, yet simple presentation of the topic. [30] Jacoby with *Statistical Graphics for Visualizing Univariate and Bivariate Data* (1997), and *Statistical Graphics for Visualizing Multivariate Data* (1998) carries out his objective to obtain pictorial representations of quantitative information by elucidating histograms, one-dimensional and enhanced scatterplots and non-parametric smoothing. [31, 32] In addition, he successfully transfers graphical displays of multivariate data on a single sheet of paper, a two-dimensional space.

EDA presents a major paradigm shift in the ways models are built. With the mantra “Let your data be your guide,” EDA offers a view that is a complete reversal of the classical principles that govern the usual steps of model building. The EDA declares the model must always follow the data, not the other way around, as in the classical approach.

Figure 1.1 EDA Paradigm



In the classical approach, the problem is stated and formulated in terms of an outcome variable Y . It is assumed that the *true* model explaining all the variation in Y is known. Specifically, it is assumed that all the structures (predictor variables, X_i s) affecting Y and their forms are known and present in the model. For example, if Age affects Y , but the log of Age reflects the true relationship with Y , then log of Age must be present in the model. Once the model is specified, the data are taken through the model-specific analysis, which provides the results in terms of numerical values associated with the structures, or estimates of the true predictor variables’ coefficients. Then, interpretation is made for declaring X_i an important predictor, assessing how X_i affects the prediction of Y , and ranking X_i in order of predictive importance.

Of course, the data analyst never knows the true model. So, familiarity with the content domain of the problem is used to explicitly put forth the true *surrogate* model, from which good predictions of Y can be made. According to Box, “all models are wrong, but some are useful.” [33] In this case, the model selected provides serviceable predictions of Y . Regardless of the model used, the assumption of knowing the truth about Y sets the statistical logic in motion to cause likely bias in the analysis, results and interpretation.

In the EDA approach, not much is assumed beyond having some prior experience with content domain of the problem. The right attitude, flexibility and sharp-sightedness are the forces behind the data analyst, who assesses the problem and lets the data guide the analysis, which then suggests the structures and their forms of the model. If the model passes the validity check, then it is considered final and ready for results and interpretation to be made. If not, with the force still behind the data analyst, the analysis and/or data are revisited until new structures produce a sound and validated model, after which final results and interpretation are made. See Figure 1.1. Without exposure to assumption violations, the EDA paradigm offers a degree of confidence that its prescribed exploratory efforts are not biased, at least in the manner of classical approach. Of course, no analysis is bias-free, as all analysts admit their own bias into the equation.

With all its strengths and determination, EDA as originally developed had two minor weaknesses that could have hindered its wide acceptance and great success. One is of a subjective or psychological nature, and the other is a misconceived notion. Data analysts know that failure to look into a multitude of possibilities can result in a flawed analysis, thus finding themselves in a competitive struggle against the data itself. Thus, EDA can foster in data analysts an insecurity that their work is never done. The PC can assist the data analysts in being thorough with their analytical due diligence, but bears no responsibility for the arrogance EDA engenders.

The belief that EDA, which was originally developed for the small-data setting, does not work as well with large samples is a misconception. Indeed, some of the graphical methods, such as the stem-and-leaf plots, and some of the numerical and counting methods, such as folding, and binning, do breakdown with large samples. However, the majority of the EDA methodology is unaffected by data size. Neither the manner in which the methods are carried out, nor the reliability of the results is changed. In fact, some of the most power EDA techniques scale up quite nicely, but do require the PC to do the serious number crunching of the big data. [34] For example, techniques such as ladder of powers, re-expressing and smoothing are very valuable tools for large sample or big data applications.

VI. Conclusion

Finding the best possible subset of variables to put in a model has been a frustrating exercise. Many variable selection methods exist. Many statisticians know them, but few know they produce poorly performing models. The wanting variable selection methods are a miscarriage of statistics because they are developed by debasing sound statistical theory into a misguided pseudo-theoretical foundation. I review the five widely used variable selection methods, itemize some of their weaknesses, and answer why they are used. Then, I present the outed solution to variable selection in regression: The Natural Seven-step Cycle of Statistical Modeling and Analysis. I feel that newcomers to Tukey's EDA need the Seven-step Cycle introduced within the narrative of Tukey's analytic philosophy. Accordingly, I enfold the solution with front and back matter – The Essence of EDA, and The EDA School of Thought, respectively.

References

1. Classical underlying assumptions, http://en.wikipedia.org/wiki/Regression_analysis.
- 1.1 The variable selection methods do not include the new breed of methods that have data mining capability.
2. Tukey, J.W., *The Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.
3. Miller, A., J., *Subset Selection in Regression*, Chapman and Hall, NY, 1990, pp. iii-x.
- 3.1 Statistica-Criteria-Supported-by-SAS.pdf (<http://www.geniq.net/res/Statistical-Criteria-Supported-by-SAS.pdf>)
4. Roecker, Ellen B. 1991. Prediction error and its estimation for subset-selected models. *Technometrics* 33, 459-468.
5. Copas, J. B. 1983. Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B* 45, 311-354.
6. Henderson, H. V., Velleman, P. F., 1981. Building multiple regression models interactively. *Biometrics* 37, 391-411.
7. Altman, D. G., Andersen, P. K. 1989. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 771-783.
8. Derksen, S., Keselman, H. J., 1992. Backward, forward and stepwise automated subset selection algorithms. *British Journal of Mathematical and Statistical Psychology*, 265-282.
9. Judd, C. M., McClelland, G. H. 1989. *Data analysis: A model comparison approach*. Harcourt Brace Jovanovich, New York.
10. Bernstein, I., H., 1988. *Applied Multivariate Analysis*, Springer -Verlag, New York.
11. Comment without an attributed citation: Frank Harrell, Vanderbilt University School of Medicine, Department of Biostatistics, Professor of Biostatistics, and Department Chair.
12. Kashid, D. N., Kulkarni, S. R. 2002. A More General Criterion for Subset Selection in Multiple Linear Regression. *Communication in Statistics-Theory & Method*, 31(5), 795-811.
13. Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *J. Royal Statistic. Society B.*, Vol. 58, No. 1, 267-288.
14. Ratner, B., 2003. *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC Press, Boca Raton, Chapter 15, which presents the GenIQ Model (www.GenIQModel.com).
15. Chen, Shyi-Ming, Shie, Jen-Da. 1995. *A New Method for Feature Subset Selection for Handling Classification Problems*, ISBN: 0-7803-9159-4.
16. SAS Proc Reg Variable Selection Methods.pdf
17. Comment without an attributed citation: In 1996, Tim C. Hesterberg, Research Scientist at Insightful Corporation, asked Brad Efron for the most important problems in statistics, fully expecting the answer to involve the bootstrap, given Efron's status as inventor. Instead, Efron named a single problem, variable selection in regression. This entails selecting variables from among a set of candidate variables, estimating parameters for those variables, and inference – hypotheses tests, standard errors, and confidence intervals.
18. Other criteria are based on information theory, and bayesian rules.
19. Dash, M., and Liu, H. 1997. Feature Selection for Classification, *Intelligent Data Analysis*, Elsevier Science Inc.
20. SAS/STAT Manual. See PROC REG, and PROC LOGISTIC
21. R-squared theoretically is not the appropriate measure for a binary dependent variable. However, many analysts use it with varying degrees of success.

- 21.1 Mark Twain quotation: “Get your facts first, then you can distort them as you please.”
http://thinkexist.com/quotes/mark_twain/
22. For example, consider two TS values: 1.934056 and 1.934069. These values are equal when rounding occurs at the third place after the decimal point: 1.934.
- 22.1 Even if there were perfect variable selection method, it is unrelastic to believe there is a unique best subset of variables.
23. Comment without an attributed citation: Frank Harrell, Vanderbilt University School of Medicine, Department of Biostatistics, Professor of Biostatistics, and Department Chair.
24. The weights or coefficients (b_0 , b_1 , b_2 and b_3) are derived to satisfy some criterion, such as minimize the mean squared error used in ordinary least-square regression, or minimize the joint probability function used in logistic regression.
25. Fox, J., 1997. Applied Regression Analysis, Linear Models, and Related Methods, Sage Publications, California.
- 25.1. The seven steps are Tukey’s. The annotations are the author’s.
26. Mosteller, F and Tukey. J. W., *Data Analysis and Regression*, Addison-Wesley, Reading, MA, 1977.
27. Hoaglin, D.C., Mosteller, F and Tukey. J. W., *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, Inc., NY, 1983.
28. Hoaglin, D.C., Mosteller, F and Tukey. J. W., *Fundamentals of Exploratory Analysis of Variance*, John Wiley & Sons, Inc., NY, 1991.
29. Chambers, M. J., Cleveland, W. S., Kleiner, B and Tukey, P.A., *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole Publishing Company, CA, 1983
30. du Toit, Steyn, A. G. W., and Stumpf, R. H., *Graphical Exploratory Data Analysis*, Springer-Verlag, NY, 1986.
31. Jacoby, W. G., *Statistical Graphics for Visualizing Univariate and Bivariate Data*, Sage Publication, CA., 1997.
32. Jacoby, W. G., *Statistical Graphics for Visualizing Multivariate Data*, Sage Publication, CA, 1998.
33. Box, G. E. P., Science and statistics. *Journal of the American Statistical Association* 71, 1976, 791-799.
34. Weiss, S. M., and Indurkha, N., *Predictive Data Mining*, Morgan Kaufman Publishers Inc. San Francisco, CA., 1998.