The Case for Being Automatic: Introducing the Automatic Linear Modeling (LINEAR) Procedure in SPSS Statistics

Hongwei Yang
University of Kentucky

This article reviews the new Automatic Linear Modeling (LINEAR) procedure in SPSS 21.0 and introduces it as a new analytical tool for researchers who regularly use linear regression. To that end, the article uses benchmark applications to examine two of its main features: 1) automatic data preparation and 2) automatic subset selection. Additionally, the article discusses limitations of the LINEAR procedure, ways to improve it, and the model ensemble capability of the procedure as a possible topic for future research. The article concludes with a summary of discussions.

Traditionally, linear regression modeling in the program Statistical Package for the Social Sciences (SPSS) is carried out using the REGRESSION procedure, which is capable of fitting linear models and computing a variety of model fit statistics. However, this aspect of the program also has limitations:

- Limited to the stepwise method only with no capability of conducting all-possible-subsets regression
- Limited in terms of optimality statistics for variable selection and existing criteria are in the form of significance tests prone to Type I/II errors
- Unable to automatically identify and handle outlying cases
- Unable to conduct model ensemble to improve predictions
- Unable to interact with the SPSS Server program to work with very large data

Given the limitations of the traditional REGRESSION procedure, this article introduces the new development in SPSS pertaining to linear modeling: The LINEAR procedure (note: available since v. 19.0), which accelerates the data analysis process through several automatic mechanisms. The new procedure is an improvement over the traditional technique in the limitations outlined above. Particularly, two major areas of improvement are in automatic variable selection and automatic data preparation.

Automatic Variable Selection

Researchers often collect a data set with a large number of independent variables and each of them is a potential predictor of the dependent variable (DV). The problem of deciding which subset(s) of the large pool of potential predictors to include in a linear regression model is; therefore, very common and arguably the hardest part of regression modeling (Ratner, 2012). One wants to choose, at least, one small subset from the pool of candidate predictors that gives adequate prediction accuracy for a reasonable cost of measurement (Miller, 2002). To that end, often used are variable selection methods (i.e., the subset selection method or the feature/attribute selection method as found in the field of data mining (Ao, 2008; Hastie, Tibshirani, & Friedman, 2011; Kantardzic, 2011; Karagiannopoulos, Anyfantis, Kotsiantis, & Pintelas, 2007; Stine, 2005; Witten, Frank, & Hall, 2011). With relevant predictors identified by such methods, estimates and predictions tend to be more precise (Tamhane & Dunlop, 2000; Weisberg, 2005).

Among many variable selection methods, the stepwise method and the all-possible-subsets (i.e., best-subsets) methods remain as popular techniques thanks to their availability in major statistical computer programs; although, certain regularization methods like the Least Absolute Shrinkage Selection Operator or LASSO by Tibshirani (1996) are gradually taking over. The two popular approaches differ primarily in the extent to which the model space is searched. The stepwise approach searches only a subset of the entire model space, whereas the all-possible-subsets approach explores the space in its entirety. Also, because selecting variable subsets requires a numerical criterion that measures the quality of each subset, both approaches count on many of the same optimality criteria. After a criterion is specified, each approach iteratively searches for the best predictor subset(s) that optimize the criterion. Next, the two automatic search methods are discussed in greater detail:

- Stepwise method: This approach enters or removes predictors one at a time after taking into account the marginal contribution of a predictor controlling for other variables already in the model. Multiple criteria exist for evaluating this marginal contribution: the Partial F-test, the

Akaike's Information Criterion Corrected (AICC) (Hurvich & Tsai, 1989), etc. The stepwise method has three variants: 1) forward selection, 2) backward elimination, and 3) forward stepwise (Efroymson, 1960). The forward stepwise variant is probably the most popular technique, where after each variable (other than the first) is added to the set of selected predictors, the algorithm examines to see if any of the previously selected variables could be dropped without appreciably increasing the residual sum of squares. This process continues until the stopping criteria are met.

- All-possible-subsets method: Compared with the stepwise approach that economizes on computational efforts by exploring only a certain part of the model space, the all-possible-subsets approach conducts a computationally intensive search of the entire model space by considering all possible regression models from the pool of potential predictors. Given p predictors, there is a total of 2^p regression models (including the intercept-only model) to be estimated. Clearly, the number of models under the all-possible-subsets approach increases rapidly as the number of candidate predictors increases. Given that the approach is computationally intensive, it works better when the number of potential predictors is not too large, for example 20 or fewer (Miller, 2002; Yan & Su, 2009).

As for popularity, the all-possible-subsets method seems to be preferred by many researchers over the stepwise method. Perhaps, this is because researchers are no longer content with the selection of a single, final model based solely on an automatic variable selection method in statistics. Instead, the recommendation is to evaluate multiple, more promising subsets that are best according to the optimality criterion of choice (Kutner, Nachtsheim, & Neter, 2004; Miller, 2002; Tamhane & Dunlop, 2000). The all-possible-subsets method is able to provide best subsets after evaluating all possible regression models and the researcher can then choose an appropriate final model from the most promising subsets after factoring in additional considerations beyond solely statistical ones.

In SPSS Statistics, the LINEAR procedure provides both the all-possible-subsets and the stepwise capability (i.e., forward stepwise only). Both approaches are guided by multiple optimality statistics. Specifically, the two variable selection platforms share three optimality criteria (i.e., AICC, adjusted R-square, and overfit prevention criterion) and the stepwise approach employs an additional criterion in the form of an F statistic.

Automatic Data Preparation

Before any linear modeling is conducted, the data should be cleaned and ready for use: 1) missing values replaced, 2) date/month/hour data converted to duration data, 3) categorical predictors specified, 4) outliers identified and handled properly, etc. To that end, the LINEAR procedure provides an automatic data preparation (ADP) platform to perform many of the above tasks. Here, we examine its ability to identify and handle outliers. In the LINEAR procedure, values of continuous predictors that lie beyond a cutoff value (three standard deviations from the mean) are treated as outliers (IBM SPSS Inc., 2012). Once the ADP option is selected, identified outliers are set to the cutoff value of three standard deviations from the mean. Given that many outliers, individually or collectively, have a strong influence on the fitted regression model (i.e., influential observations) (Chatterjee & Hadi, 1988), the LINEAR procedure also provides a diagnostic statistic, Cook's (1977) Distance, that measures the impact of each of the identified outliers on the fitted model. Though, it should be noted that outlying cases and influential cases are not necessarily the same because each is defined to measure a specific feature of the data given the structure of the model (Belsley, Kuh & Welsch, 1980; Kutner et al., 2004; Mendenhall & Sincich, 2003; Tamhane & Dunlop, 2000). According to Belsley et al., an influential observation has a demonstrably larger impact on the calculated values of various estimates (e.g., coefficients, standard errors, *t* values) than is the case for most of the other observations. The fact that a small subset of the data can have a disproportionate influence on the estimated parameters and/or predictions is of concern for if this is the case, it is quite possible that the model estimates are based primarily on this small data subset rather than on the majority of the data. Therefore, it is important for the researcher to be able to assess the impact of identified outliers on the model and decide whether or not to remove those outlying cases that are also influential at the same time so that they do not jeopardize the main analysis.

As has been described, to detect any influential outlying cases, the LINEAR procedure institutes the measure of Cook's Distance that takes into account the impact of both the predictor (i.e., leverage) and the DV (i.e., discrepancy) data on the estimates of model parameters. According to IBM SPSS Inc. (2012, p. 528), the LINEAR procedure bases the determination of an influential case on a rule of thumb from Fox (1997). Once an outlying observation satisfies this rule, it is automatically displayed in the output as an influential case. Many times, plenty of the outliers identified are indeed also influential at the same time, so the researcher can decide whether to eliminate them in any subsequent analysis. With some or all of identified influential outlying cases removed, it is hoped that the parameter estimates are decided primarily by the majority of the data, instead of by just a small subset that are unusual or inordinately influential. Unfortunately, at this point, the LINEAR procedure does not have the capability to measure the presence and intensity of collinear relations among the regression data, which is another common diagnostic problem in regression (Belsley, et al., 1980). Hopefully, this issue can be addressed in a future release of the program.

Goal of Study

Given the new features of the LINEAR procedure, it is important for researchers who use regression analysis regularly to take advantage of them. However, a review of the most recent regression literature indicates very few texts ever mention its use. This review is limited to texts that have a publication date that is about one year after the procedure was initially made available in August of 2010. Additionally, some researchers argue that the learning curve for this procedure is rather steep (StackExchange, 2011). Thus, given that scarcity of coverage in the literature may prevent the features of the new and challenging procedure from being fully utilized; this article provides an overview of two main features of the program: 1) automatic data preparation, and 2) building a standard model under the two variable selection methods. To that end, two numerical examples are provided next with each example discussing one feature.

Two Numerical Examples

Example 1: Automatic Data Preparation

The first example is based on a re-analysis of two benchmark data sets: one used in Belsley et al. (1980) and the other from Chatterjee and Hadi (1988). Both works provide dedicated coverage concerning the issue of influential cases in linear regression. For the first study, Belsley et al. use the original "Boston Housing" data. For the present study, data were obtained from the UC Irvine (UCI) Machine Learning Repository (Bache & Lichman, 2013). The data set has 506 cases with one DV and 13 predictor variables. Belsley et al. applied various transformations to the data set before working on identifying cases that had an influential impact on a first order linear regression model containing the main effects of all 13 predictors. This article follows their practice and performs the same set of transformations before having the LINEAR procedure automatically prepare the data for the same model. Twelve of the 13 predictor variables were continuous and one was treated as continuous (i.e., a binary, categorical predictor coded into a dummy indicator). To learn more about the model and the data, refer to Belsley et al. (p. 231), and Harrison and Rubinfeld (1978). For the Chatterjee and Hadi application, the "Health Club" data consist of responses from 30 employees with each individual measured on one DV and four predictor variables. Chatterjee and Hadi use the data to fit a first order linear regression model containing the main effects of all four predictors. For comparison purposes, the article fits that same model. All four predictor variables used in the model are continuous. To learn more about the model and the data, refer to Chatterjee and Hadi (pp. 128-134).

After running the LINEAR analysis with the Belsey et al., (1980) data, a total of 30 outlying cases were identified as influential and they are presented in the top part of Figure 1: 1) the plot (left) is a plot of Cook's Distances on record ID as provided by the LINEAR procedure and 2) the plot (right) is a boxplot of values of Cook's Distance of the 30 outlying cases. By contrast, with the help of as many as four diagnostic measures, Belsley et al. (pp. 238-239) identify 67 cases as abnormalities: 1) outlying only (3 cases), 2) influential only (39 cases), and 3) outlying plus influential (25 cases). Out of those 30 influential cases detected by the LINEAR procedure, 26 of them are endorsed by at least one diagnostic measure in Belsley et al. as influential (short of cases 469, 375, 470, and 480). Although the number of influential cases found by the LINEAR procedure is smaller, it should be noted that the procedure uses just a single diagnostic criterion to evaluate the impact of each observation that has already been identified as outlying; whereas Belsley at al. use as many as four different measures on all observations in

the original data regardless of whether any one is already identified as outlying. Therefore, it is not surprising that the latter comes up with more influential cases because of two issues: 1) each diagnostic measure evaluates the data from its own perspective under the context of the fitted model: Leverage only, leverage plus discrepancy without row deletion, leverage plus discrepancy with row deletion and 2) influential observations identified by Belsley at al. include those that are not considered to be outlying (e.g., observations 124, 127, 143, 144, 148, and many more).

After running the LINEAR analysis with the Chatterjee and Hadi (1988) data, a total of four outlying cases were identified as having an influential impact on the parameter estimates of the model: Cases 28, 23, 30, and 8. This same set of cases was also selected by Chatterjee and Hadi (p. 134) as influential when they used Cook's Distance despite minor differences in the rank ordering of the impact. Further, both studies agree that observation 28 is the most influential case out of the four. In the bottom part of Figure 1, these four influential cases are presented graphically in two plots in a manner similar to those for the first data set.

Given that, with just a single criterion (i.e., Cook's Distance), the LINEAR procedure is able to identify about 40% of the influential cases found by Belsley et al. (1980) using as many as four criteria and detect 100% of the influential cases found by Chatterjee and Hadi (1988) employing the same criterion. That is, there was certain supportive evidence found on the effectiveness of the procedure in finding abnormalities in the data. Because the procedure is able to automatically detect influential cases, it should help researchers with initial data preparation, particularly when the dimension of the data is so large that manually identifying abnormal cases is too time-consuming to be accomplished in a reasonable amount of time. With the data preparation feature of the LINEAR procedure illustrated, we next proceed to its variable selection feature.

Example 2: Variable Selection under Two Methods

For the second example, we selected a total of 10 benchmark applications for assessing the subset selection capability of the LINEAR procedure. The 10 data sets were retrieved separately from three sources: 1) the UCI Machine Learning Repository by Bache and Lichman (2013), 2) the R package mlbench by Leisch and Dimitriadou (2013), and 3) the R package ElemStatLearn by Halvorsen (2013). A brief summary of the 10 benchmark applications is found in Table 1. It should be noted that the last two applications (WineQuality-RedWine and WineQuality-WhiteWine) each have a DV (i.e., graded quality of wine) that is in fact ordinal in measurement ranging from 0 (Very Bad) to 10 (Very Excellent). Here, we still analyze each DV in a continuous scale following the practice in the original paper that presents the two applications by Cortez, Cerdeira, Almeida, Matos, and Reis (2009).

Given multiple predictors from each application, we subjected them separately to the forward stepwise and the all-possible-subsets methods. As for the entry/removal criterion, we selected the AICC statistic due to three reasons: 1) it is not prone to Type I/II errors, 2) it seems to work well for small as well as large samples, and 3) it is moderately recommended by Miller (2002). When the optimal model from each variable selection method was finally identified and estimated, we evaluated the degree of model fit using the correlation coefficient between the observed and the predicted values of the DV; an approach adopted by similar studies on subset selection/feature selection such as Karagiannopoulos, et al. (2007). A summary of the model search/evaluation process is also found in Table 1.

Under each variable selection method, the correlations between the observed and predicted values of the DV are generally high (above .80) with only two exceptions (about .51 and .54), which suggests that the patterns of the two sets of values are relatively consistent with each other in almost all 10 benchmark applications. This provides support for the final, fitted model from each variable selection method. One thing to note is that for almost all data sets, the final models from the two variable selection methods tend to be identical with the exception of the BodyFat data where the two optimal models differ. This finding is worth further investigation because the all-possible-subsets method supposedly searches the entire model space; whereas the forward stepwise method only does a small portion. Although the possibility does exist that they end up finding the same optimal model, the number of times they agree with each other here in the LINEAR procedure may still be too high. Given that multiple algorithms exist in implementing the stepwise and the all-possible-subsets method (Furnival & Wilson, 1974; Miller, 2002), it could have been a characteristic of the particular algorithms adopted by the procedure. However, further

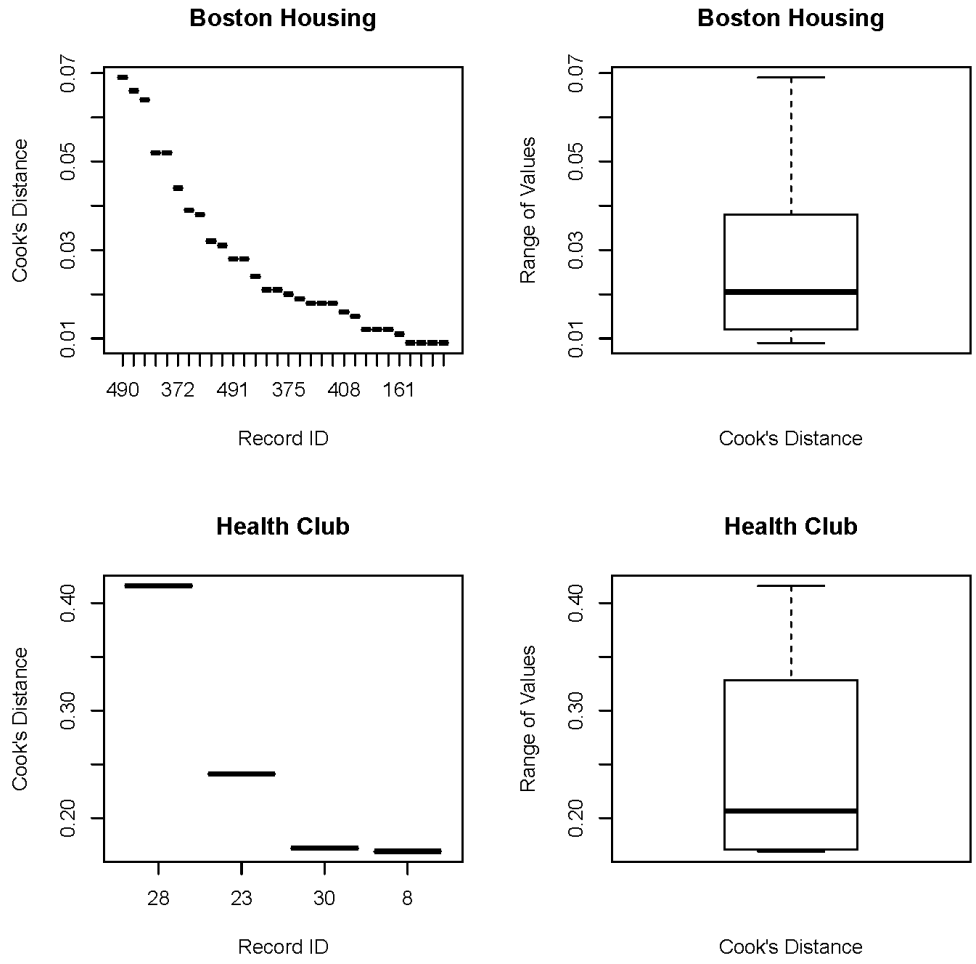

Figure 1. Automatic data preparation: Cook's Distance values from two analyses.

Table 1. Automatic Variable Selection: Data Sets and Correlations Between Observed and Predicted

Dataset	Sample size	Number of predictors	Correlations: Forward stepwise	Correlations: All possible subsets	Final model identical (Y/N)	Source
BodyFat	252	14	.988	.988	N	UCI
Bone	485	02	.510	.510	Y	ElemStatLearn
BostonHousing Corrected	506	13	.868	.868	Y	mlbench
CPU Performance	209	06	.847	.847	Y	UCI
Galaxy	323	04	.948	.948	Y	ElemStatLearn
Ozone	366	09	.811	.811	Y	mlbench
Prostate	097	08	.816	.816	Y	ElemStatLearn
Servo	167	04	.847	.847	Y	mlbench
WineQuality-RedWine	1,599	11	.608	.608	Y	UCI
WineQuality-WhiteWine	4,898	11	.537	.537	Y	UCI

Note. The Boston housing data set analyzed here is the corrected version, not the same as the one used in the other numerical example.

study is needed to fully investigate if this tentative explanation is reasonable. Finally, the all-possible-subsets method provides us with a set of (up to) 10 best models that we can treat as more promising subsets and examine more closely by factoring in additional considerations beyond solely statistical ones; a recommended approach in the literature (Kutner, et al., 2004; Miller, 2002; Tamhane & Dunlop, 2000).

Discussion

Through the two examples, the LINEAR procedure provides an effective, new solution to linear regression modeling in SPSS in comparison to the traditional REGRESSION procedure; where the LINEAR procedure functions well as the latter's substitute. That is, LINEAR provides almost everything found in the traditional procedure, but it also offers additional, typically more advanced, features not available in REGRESSION. Subsequently, we would like to discuss several additional issues regarding the LINEAR procedure with the hope of addressing some philosophical concerns, suggesting several possible improvements, and advancing new research.

Automation versus Human Input

First, there are mixed feelings about the LINEAR procedure that automates many aspects of linear regression modeling; a fundamental predictive data mining (DM) tool that serves as the building block of more complex DM algorithms (Cerrito, 2006; Cios, Pedrycz, Swiniarski & Kurgan 2007; Hill, Malone, & Trocine, 2004; Larose, 2005; Ratner, 2012; Witten, et al., 2011). Some researchers dislike the procedure, such as Field (2013); whereas others disagree and argue that the procedure is very useful such as Lynda Pod Cast (2011). After inspecting those differing views, we argue that at least the point of being automatic is well justified in the literature (i.e., in data mining), particularly when the data are very large. As is known, big data are an outgrowth of today's digital environment that generates data flowing continuously at unprecedented speed and volume. This explosive growth in data has generated an urgent need for automated tools that can intelligently assist in transforming the vast amounts of data into useful information. Specifically, because we are "drowning" in data, but" starving" for knowledge, we must find ways to automatically analyze the data, to automatically classify it, to automatically summarize it, to automatically discover and characterize trends in it, and to automatically flag anomalies (Han & Kamber, 2006; Rubenking, 2001; Witten et al.). Given large data, knowledge discovery has to be a cooperative effort of humans and computers through automatic as well as manual methods. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers (Kantardzic, 2011; Yao, Zhong, & Zhao, 2008). Clearly, the LINEAR procedure represents an important move in the direction of automating the data analysis process, and this move is further enhanced by the procedure's ability to communicate with the SPSS Server program designed to improve the efficiency and productivity of analyzing very large data sets.

On the other hand, despite the need for automation in data mining, we also argue that automation is no substitute for human input (Ankerst, 2003; Larose, 2005). Data mining is not a black box that can automatically answer any question that is asked. Humans need to be actively involved at every phase of the data mining process via use of theory, among other things. In the case of linear regression modeling under the LINEAR procedure, human participation is needed in each step of the analysis: 1) before the analysis, the researcher needs to review the literature extensively to decide how to phrase the question(s) correctly and which variables should be measured to collect data that could be used to answer the question(s) and 2) during the analysis, human input is needed to examine the context of any problematic cases identified by the LINEAR procedure to figure out why a particular case is abnormal. To accomplish this, the researcher must examine responses to each variable for the cases in question and interpret them in light of what is being measured (Belsley et al., 1980, pp. 237-241). Lastly, 3) after the analysis, given multiply selected models from subset selection, the researcher needs to factor in theoretical considerations before making the decision related to the choice of the final model. In addition, several other aspects of data mining, where human participation plays an important role, includes the selection of the right model given the distribution of the data, the use of the right algorithm under the context of the data so that the analysis can be completed in a reasonable amount of time, and the presentation of the results in a manner that is easy to understand for the targeted audience. In the end, we believe that the very power of the formidable data mining algorithms embedded in many black-box data mining software programs makes

their misuse proportionally more dangerous. Therefore, any inappropriate use of the LINEAR procedure could also give us a highly distorted picture of reality.

Grounds for Improvements

Despite the new features that the LINEAR procedure offers, it still falls short of expectations in numerous ways so there is plenty of room for improvement. For instance, predicated on the two features that we examine here, the procedure can be enhanced in many ways.

- First, there is currently only a single diagnostic statistic designed for identifying and flagging influential cases, which is insufficient because researchers tend to use multiple criteria for that purpose (Belsley et al., 1980). Hopefully, the LINEAR procedure would provide at least one measure from each of the groups of diagnostic statistics as recommended by Chatterjee and Hadi (1988, pp. 183-184). Particularly, given that some predictors may be of more theoretical interest than others, it would be helpful for the procedure to provide diagnostic statistics (e.g., DFBETAS) that are able to investigate each predictor variable separately to evaluate individual coefficient sensitivity. Such statistics would be a useful addition to those that assume all regression coefficients are of equal interest and evaluate the overall/global sensitivity of all predictors combined.
- Second, the procedure currently does not make available many of the details of the linear modeling process, which is acknowledged by the SPSS support team (D. Matheson, personal communication, May 20, 2013). It does not even allow the researcher to save regression residuals; although, those values are used internally to compute many of the displayed statistics. Because the researcher may have the need to explore the data/model(s) further beyond what the program provides, it would be helpful for the procedure to make accessible many of the currently hidden statistics. As will be seen later, the article would have assessed the model ensemble features of the procedure to see if they are truly able to improve predictions, had certain statistics (i.e., analysis weights, model weight, and residuals at each base model) been available.
- Finally, two other features that we believe should be added to the LINEAR procedure are data partitioning/splitting and forced entry of variables (i.e., terms). With the partitioning of data capability, already available in other data mining software programs such as SAS Enterprise Miner, the researcher could easily split the data into two sets with one set used for model development (i.e., training) and the other set for model validation, which is a common strategy in data mining to better assess the quality of model generalization. Specifically, predictions would be made for the independent data in the validation data set from the regression model developed from the model-building data set to calibrate the predictive ability of this model for the new data; a process known as cross-validation. At this point, the LINEAR procedure does not have the ability to split the data within the procedure so that the entire sample data are used for both model development and model validation, which typically over-estimates the predictive ability of the fitted model (Kutner, et al., 2004; SAS Institute, Inc., 2011a; Tamhane, & Dunlop, 2000). As for the feature of forced entry of variables (i.e., terms), it is about always including in the model some predictor variables of choice, regardless of their significance levels, and at the same time applying subset selection to all other predictors. This feature is already available in the NOMREG procedure in SPSS Statistics and also in the SPECIFICATION SEARCH platform in SPSS AMOS, but it is yet to be offered in the LINEAR procedure. From the perspective of modeling, this feature would require theoretically interesting variables whose effects need to be controlled (e.g., pre-test scores) and should always be part of the model without being subjected to the process of subset selection. Then, the modeling process would focus on the relationship between the DV and the remaining predictors when simultaneously adjusting for the effects from all forced entry predictors (i.e., terms).

Model Ensemble for Future Research

Finally, the article suggests a topic for future research that relates to the ensemble capability offered in LINEAR procedure. Basically, the procedure offers two ensemble methods: 1) Bootstrap aggregating or bagging by Breiman (1996) and 2) adaptive boosting or AdaBoost by Drucker (1997) and Freund and Schapire (1996; 1997). In addition, the procedure also provides several special ensemble methods for

large data that are able to communicate with the SPSS Server program (IBM SPSS Inc., 2012, pp. 299-301). The ensemble techniques are designed to create a model ensemble containing multiple component/base models that are averaged in a certain manner to improve the stability/accuracy of predictions, and the techniques can be incorporated into either variable selection method.

Here, we take the boosting approach as an example to briefly demonstrate how it is incorporated in the LINEAR procedure in SPSS. It should be noted that both the bagging and the boosting approaches can be used for either the prediction of a continuous outcome or the classification of a categorical outcome. Like bagging, the boosting approach is one of many ensemble methods that are rooted in machine learning (Freund, 1995; Schapire, 1990; Schapire, 2002). They both leverage the power of multiple models to achieve better prediction or classification accuracy than any of the individual models could perform on their own (Ao, 2008; Oza, 2005). Both of these techniques rely on a usually user-specified number of base models with each one estimated under a perturbed version of the original data set. Whereas the bagging method needs multiple parallel bootstrap samples independently generated from the original sample to estimate the base models, the boosting method uses the base models in a sequential collaboration manner where each new model concentrates more on the observations for which the previous models had larger residuals or larger discrepancies (Barutcuoglu & Alpaydin, 2003; Mevik, Segtnan, & Næs, 2005). In other words, the boosting technique works by sequentially applying the linear model as the learning machine to different, perturbed versions of the original data (i.e., generated by applying the analysis weights that are updated from one base model to the next) to obtain the outputs of each base model. At each base model, a single model weight statistic is computed from the corresponding, updated analysis weights for the data. Finally, the core ensemble output in the form of predicted values of the DV is a weighted combination of the base model outputs.

There are many different implementations of the boosting approach (Barutcuoglu & Alpaydin, 2003; Buhlmann & Hothorn, 2007; Hothorn, Buhlmann, Kneib, Schmid, & Hofner, 2010; Schapire, 2002; Schonlau, 2005). The LINEAR procedure implements the AdaBoost method by Drucker (1997). The AdaBoost algorithm is an efficient and popular implementation of the boosting principle. Although it was originally designed for classification problems, it has been adapted to handle the prediction of a continuous outcome with the linear regression model used as the learning machine. A description of several versions of AdaBoost is found in Barutcuoglu and Alpaydin.

Within the implementation of AdaBoost, the LINEAR procedure uses the weighted median method to score the ensemble model that consists of a user-specified number of base models that are regarded as one combined model (IBM SPSS Inc., 2012). The weighted median for an observation is computed by first sorting all base model predicted values of this observation in ascending order of magnitude, and then summing their model weights until the sum is no less than half of the total model weights. Then, the computed median can be used to index, from within all base model predicted values of this observation, the final predicted value of this observation for the ensemble model. To learn more about the ensemble algorithms implemented in the LINEAR procedure, refer to IBM SPSS Inc. (pp. 294-305).

Further, the ensemble capability of the LINEAR procedure can be used in combination with each of its two subset selection methods to improve the performance of individual models. This combination of subset selection and machine learning algorithms has recently become popular. For example, Liu, Cui, Jiang, and Ma (2004) apply the ensemble neural networks with combinational feature selection to the microarray experiments for tumor classification and have obtained remarkably improved results (Ao, 2008). As far as the LINEAR procedure is concerned, the combinational work of feature selection and model ensemble is for the prediction of a continuous outcome, instead of the classification of a categorical outcome. The learning machine used in the LINEAR procedure is the linear model, instead of the neural networks. This modeling process can be conducted under each of the two subset selection methods to create and estimate each base model for the ensemble method of choice (i.e., bagging or boosting). The final ensemble model consists of base models; each one of which is constructed through the subset selection method of choice.

Certainly, SPSS LINEAR is not the only statistics program that provides the model ensemble capability. There are other programs that also offer similar features either as an official routine or through an un-official plug-in. Here, we use SAS and STATA as an example to see what ensemble features each of the two programs offers.

The ensemble capability in SAS is provided mainly through its Enterprise Miner (EM) program (SAS Institute, Inc., 2011a; SAS Institute, Inc., 2011b). We use EM 7.1 in SAS 9.3 as an example. In this version of the EM program, there are several nodes that can be used to conduct model ensemble. First, the Gradient Boosting node that implements the boosting approach, with the learning machine being the decision tree for both categorical and continuous outcomes instead of the linear regression model as implemented in the LINEAR procedure that is only for the continuous outcome, a major limitation of the LINEAR procedure. Second, the Ensemble node that is able to create a model ensemble from multiple preceding nodes for modeling such as the Regression node, the Decision Tree node, and the AutoNeural node. Again, the Ensemble node is able to handle categorical as well as continuous outcomes. Finally, there is a third option that uses the Start Groups node in conjunction with the End Groups node. The combined use of the two nodes allows data that are segmented into different groups to be processed in different ways. When analyzing each group of data, typical ensemble techniques such as Boosting or Bagging can be specified as the group processing function within the Start Groups node.

As for the STATA program, it is currently limited in the ensemble capability. As of version 13, no official command for model ensemble has been provided within the STATA program. According to Schonlau (2005), one option that was published in the official journal of STATA is to take advantage of a plug-in program that provides a new, but still unofficial, STATA command known as boost, which implements the boosting algorithm described in Hastie, Tibshirani, and Friedman (2011, p. 322). The boost command offers the boosting option and it accommodates continuous (i.e., linear model), categorical (i.e., logistic regression), and count/frequency (i.e., Poisson regression) outcomes, and the command can also be used in combination with common variable selection methods such as stepwise regression.

Finally, the comparison of ensemble methods in terms of improving the accuracy/stability of individual models, with/without simultaneous subset selection, is an active area of research (Barutcuoglu & Alpaydin, 2003; Breiman, 1996; Cortez, et al., 2009; Drucker, 1997; Freund & Schapire, 1997; Liu, et al., 2004; Schapire, 2002). Therefore, it would be interesting to assess the performance of the ensemble capability of the LINEAR procedure with/without also conducting subset selection of variables. Such a study is well beyond the scope of this article that aims to provide an overall introduction of various features of this new LINEAR procedure and not just its ensemble capability. A separate, dedicated study would be needed that takes advantage of both simulated and benchmark data sets to compare ensemble methods with each other and with regular methods that do not conduct model ensemble. The comparisons can be made under separate conditions when it comes to the construction of each base model: no application of any variable selection method versus application of each variable selection method of choice. However, it is recommended that such a study should wait until future releases of this LINEAR procedure become available which, hopefully, would allow more details of the ensemble/subset selection process to be available to the user as outputs. At this point, the entire ensemble process conducted in the LINEAR procedure is almost completely invisible to the user, and plenty of relevant information is missing from the outputs of the LINEAR procedure. For example, neither the analysis weights of the data at a base model nor the model weight statistic of the base model can be retrieved from the analysis. In addition, no residuals of any sort can be saved from the LINEAR procedure, which makes it difficult to assess model performance. Given the limited amount of outputs in the existing LINEAR procedure, the amount of work in assessing the performance of its ensemble capability would be excessive, if not impossible.

Conclusion

This article demonstrates the effectiveness of two features of the LINEAR procedure in SPSS Statistics using benchmark applications. Through the demonstration, the article argues for the use of the LINEAR program as a substitute for the traditional REGRESSION procedure; particularly with very large data where manual handling of the data becomes too time-consuming/work-demanding to be practically accomplished. The article also discusses philosophical issues related to the new procedure, aspects of this procedure where improvements can be made, and its ensemble capability as a topic for future research.

References

Ankerst, M. (2003). The perfect data mining tool: Interactive or automation – Report on the SIGKDD-2002 Panel. *SIGKDD Explorations, 5*, 110-111.

Ao, S. (2008). *Data mining and applications in Genomics*. Berlin, Germany: Springer Science+Business Media.

Bache, K., & Lichman, M. (2013). *UCI machine learning repository* retrieved from http://archive.ics.uci.edu/ml

Barutcuoglu, Z., & Alpaydin, E. (2003). A comparison of model aggregation methods for regression. In O. Kaynak, E. Alpaydin, E. Oja, & L. Xu. (Eds.), *Artificial neural networks and neural information processing - ICANN/ICONIP 2003* (pp. 76–83). New York: Springer.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons, Inc.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Buhlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science, 22*, 477-505.

Cerrito, P. B. (2006). *Introduction to data mining: Using SAS Enterprise Miner*. Cary, NC: SAS Institute Inc.

Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity analysis: Linear regression*. New York: John Wiley & Sons, Inc.

Cios, K., Pedrycz, W., Swiniarski, R. W. & Kurgan, L. A. (2007). *Data mining: A knowledge discovery approach*. New York: Springer.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*, 15–18.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*, 547-553.

Drucker, H. (1997). Improving regressor using boosting techniques. *Proceedings of the 14th International Conferences on Machine Learning*, 107-115.

Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical methods for digital computers* (pp. 191–203). New York: Wiley.

Field, A. (2013). *SPSS Automatic Linear Modeling*. Retrieved from http://www.youtube.com/watch?v=pltr74IlxOg

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage Publications, Inc.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation, 121*, 256-285.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, 148-156.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119-139.

Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics, 16*, 499-511.

Halvorsen, K. (2013). *ElemStatLearn: Data sets, functions and examples.* Retrieved from http://cran.r-project.org/web/packages/ElemStatLearn/index.html

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publisher.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management, 5*, 81-102.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Hill, C. M., Malone, L. C., & Trocine, L. (2004). Data mining and traditional regression. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 233–249). Boca Raton, FL: CRC Press LLC.

Hothorn, T., Buhlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research, 11*, 2109-2113.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297-307.

IBM SPSS Inc. (2012). *IBM SPSS Statistics 21 Algorithms*. Chicago: Author.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B., & Pintelas, P. E. (2007). *Feature selection for regression problems*. Proceedings of the 8th Hellenic European Research on Computer Mathematics and its Applications (HERCMA), Athens, Greece.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models* (4th ed.). New York: McGraw-Hill/Irwin.

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: John Wiley & Sons, Inc.

Leisch, F., & Dimitriadou, E. (2013). *mlbench: Machine learning benchmark problems*. Retrieved from http://cran.r-project.org/web/packages/mlbench/index.html

Liu, B., Cui, Q., Jiang, T., & Ma, S. (2004). A combinational feature selection and ensemble neural network method for classification of gene expression data. *BMC Bioinformatics, 5*,136.

Lynda Pod Cast. (2011). *How to use automatic linear modeling*. Retrieved from http://www.youtube.com/watch?v=JIs4sMxAFg0

Mendenhall, W., & Sincich, T. (2003). *A second course in statistics: Regression analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Mevik, B., Segtnan, V. H., & Næs, T. (2005). Ensemble methods and partial least squares regression. *Journal of Chemometrics, 18*, 498-507.

Miller, A. J. (2002). *Subset selection in regression* (2nd ed.). New York: CRC.

Oza, N. C. (2005). Ensemble data mining methods. In J. Wang (Ed.), *Encyclopedia of Data Warehousing and Mining* (pp. 448-453). Hershey, PA: Information Science Reference.

Ratner, B. (2012). *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data* (2nd ed.). Boca Raton, FL: Taylor & Francis Group.

Rubenking, N. J. (2001). *Hidden messages*. Retrieved from http://www.pcmag.com/article2/0,2817,8637,00.asp

SAS Institute, Inc. (2011a). *Getting started with SAS Enterprise Miner 7.1*. Cary, NC: Author.

SAS Institute, Inc. (2011b). *SAS Enterprise Miner 7.1 extension nodes: Developer's guide*. Cary, NC: Author.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*, 197-227.

Schapire, R. E. (2002). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *MSRI workshop on nonlinear estimation and classification*. New York: Springer.

Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a STATA plugin. *The STATA Journal, 5*, 330-354.

StackExchange. (2011). *Is automatic linear modeling in SPSS a good or bad thing?* Retrieved from http://stats.stackexchange.com/questions/7432/is-automatic-linear-modelling-in-spss-a-good-or-bad-thing

Stine, R. A. (2005, January). *Feature selection for models in data mining*. Paper presented at the data mining conference at The College of New Jersey, Ewing, NJ.

Tamhane, A. C., & Dunlop, D. D. (2000). *Statistics and data analysis: From elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267-288.

Weisberg, S. (2005). *Applied linear regression* (3rd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Elsevier Inc.

Yan, X., & Su, X. G. (2009). *Linear regression analysis: Theory and computing*. Singapore: World Scientific.

Yao, Y., Zhong, N., & Zhao, Y. (2008). A conceptual framework of data mining. *Studies in Computational Intelligence, 118*, 501–515.

Send correspondence to:
Hongwei Yang
University of Kentucky
Email: hya222@uky.edu