AUTOMATIC LINEAR MODELING (LINEAR) in SPSS

Hongwei "Patrick" Yang, PhD

Assistant Professor
Department of Educational Policy Studies & Evaluation
College of Education
University of Kentucky
Lexington, KY 40506-0001
E-mail: patrick.yang@uky.edu

Presented at the 2014 Modern Modeling Methods (M^3) Conference
Neag School of Education, University of Connecticut
Storrs, CT 06269-3064

May 16, 2014

Background

- Traditionally, linear regression modeling in SPSS Statistics is carried out using the REGRESSION procedure which is capable of fitting linear models and computing a variety of model fit statistics
- Limitations of the REGRESSION procedure include
  - 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

Background

- Given the limitations of the traditional REGRESSION procedure, this presentation here introduces the new development in SPSS on linear modeling: The LINEAR procedure (available since version 19)
  - The LINEAR procedure accelerates the data analysis process through several automatic mechanisms
  - The LINEAR procedure improves over the traditional REGRESSION procedure in the limitations outlined above. Two of the major improvements are discussed here:
    - Automatic variable selection
    - Automatic data preparation

Two Major Improvements: Automatic Variable Selection

- In regression modeling, variable selection methods are used very often and they are also known as subset selection method, or feature/attribute selection method as in the field of data mining [2, 25, 30, 31, 49, 53]
- We typically want to choose, at least, one small subset from the pool of candidate predictors that gives adequate prediction accuracy for a reasonable cost of measurement [39]

Two Major Improvements: Automatic Variable Selection

- Among many variable selection methods, the stepwise method and the all-possible-subsets (a.k.a., best-subsets) method remain to be popular thanks to their availability in major statistics computer programs
  - This is changing because certain regularization methods like the *Least Absolute Shrinkage Selection Operator* or *LASSO* by Tibshirani [51] are gradually taking over as alternative variable selection methods
  - The regularization methods are also available in SPSS Statistics through its categorical regression procedure (CATREG)
    - Ridge regression
    - LASSO
    - Elastic net that combines ridge regression and LASSO

Two Major Improvements: Automatic Variable Selection

- 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
- 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 a much larger model space by considering all possible regression models from the pool of potential predictors
    - Given that the approach is computationally intensive, it works better when the number of potential predictors is not too large, say 20 or fewer [39, 54]

Two Major Improvements: Automatic Variable Selection

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

Two Major Improvements: Automatic Data Preparation

- Before any linear modeling is conducted, the data have to be 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

Two Major Improvements: Automatic Data Preparation

- In the LINEAR procedure, values of continuous predictors that lie beyond a cutoff value (three standard deviations from the mean) are treated as outliers
- Once the ADP option is selected, identified outliers are set to the cutoff value of three standard deviations from the mean [29]
- Given that many outliers, individually or collectively, have a strong influence on the fitted regression model, a.k.a., influential observations, the LINEAR procedure also provides a diagnostic statistic (Cook's Distance) that measures the impact of each of the identified outliers on the fitted model

Two Major Improvements: Automatic Data Preparation

- To measure the influence of outlying cases on the fitted model, the LINEAR procedure institutes the measure of Cook's Distance which takes into account the impact of both the predictor (leverage) and the DV (discrepancy) data on the estimates of model parameters
- The LINEAR procedure bases the determination of an influential case on a rule of thumb from Fox [16]. Once an outlying observation satisfies this rule, it is automatically displayed in the output as an influential case

Example One: Automatic Data Preparation

- The first example is based on a re-analysis of two benchmark data sets: One used in Belsley et al. [5] and the other from Chatterjee and Hadi [9]
- Both works provide a dedicated coverage on the issue of influential cases in linear regression. For the first study, Belsley et al. use the original "Boston Housing" data. Then, data for the present study were obtained from the UC Irvine (UCI) Machine Learning Repository [3]
- When analyzing each data set, we fit that same model as is presented in the original study

Example One: Automatic Data Preparation

- After running the LINEAR analysis with the Belsey et al., data, a total of 30 outlying cases are identified as influential and they are graphically 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 those 30 outlying cases
- By contrast, with the help of as many as four diagnostic measures, Belsley et al. (1980, p. 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)

Example One: Automatic Data Preparation

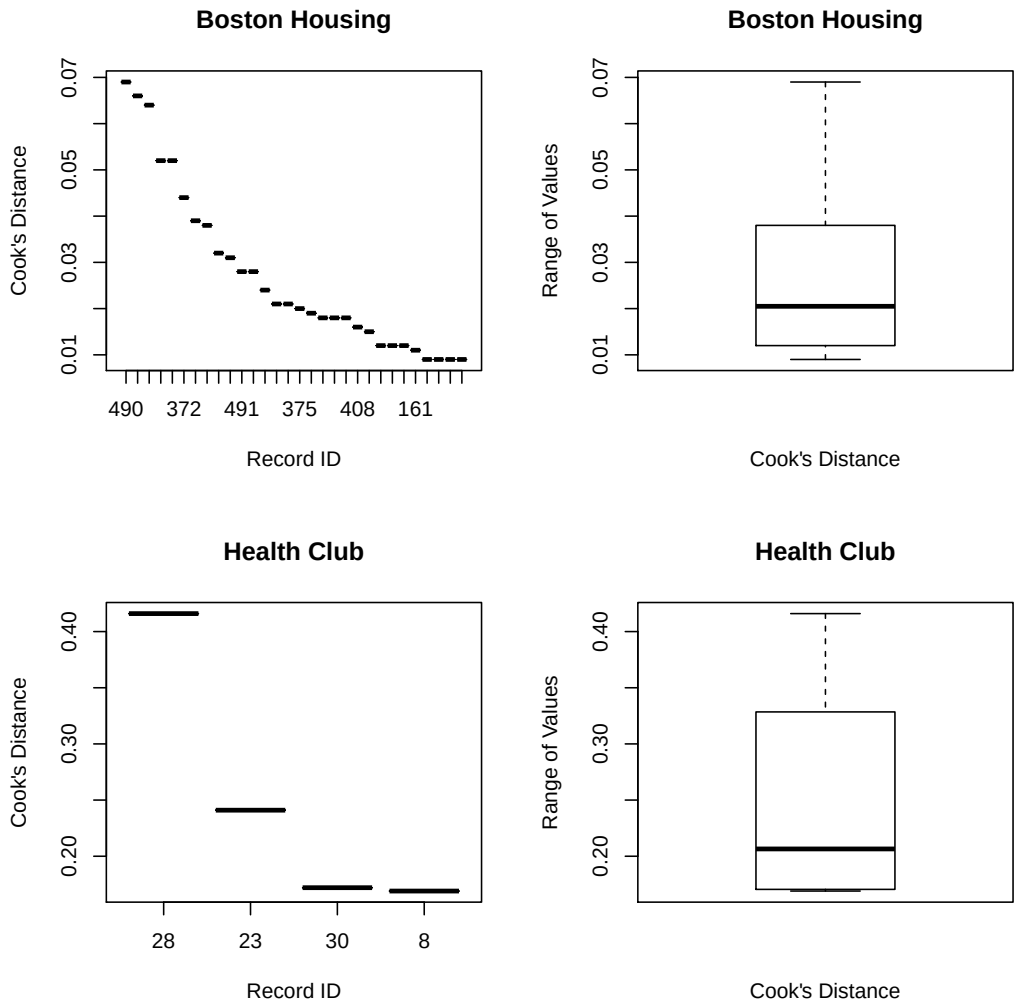
- Although the number of influential cases found by the LINEAR procedure is smaller, it should be noted that the procedure uses just a single diagnostic criterion to evaluate the impact of each observation that has already been identified as outlying whereas Belsley at al. use as many as four different measures on all observations in the original data regardless of whether any one is already identified as outlying
- So, it is not surprising that the latter comes up with more influential cases because of two things: 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, etc., and 2) influential observations identified by Belsley at al. include those that are not considered to be outlying (say, observations 124, 127, 143, 144, 148, and many more)

Example One: Automatic Data Preparation

- After running the LINEAR analysis with the Chatterjee and Hadi data, a total of four outlying cases are identified as having an influential impact on the parameter estimates of the model: Cases 28, 23, 30, and 8. This same set of cases is also selected by Chatterjee and Hadi (1988, p. 134) as influential when they use Cook's Distance, despite minor differences in 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

Example One: Automatic Data Preparation

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



Example One: Automatic Data Preparation

- Given that, with just a single criterion (Cook's Distance), the LINEAR procedure is able to identify about 40% of the influential cases found by Belsley et al. [5] using as many as four criteria, and detect 100% of the influential cases found by Chatterjee and Hadi [9] using the same criterion, we may say that certain supportive evidence has been found on the effectiveness of the procedure in finding abnormalities in the data
- This feature is very useful in 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

Example Two: Automatic Data Preparation

- For the second example, we select a total of 10 benchmark applications for assessing the subset selection capability of the LINEAR procedure
- The 10 data sets are retrieved separately from three sources: 1) the UCI Machine Learning Repository by Bache and Lichman [3], 2) the R package *mlbench* by Leisch and Dimitriadou [34], and 3) the R package *ElemStatLearn* by Halvorsen [21]
- A brief summary of the 10 benchmark applications is found in Table 1

Example Two: Automatic Data Preparation

Table 1: Data Sets and Correlations Between Observed and Predicted

Dataset	Sample size	Number of predictors	Correlation: Stepwise	Correlation: All possible subsets	Final model identical	Source
Bodyfat	252	14	.988	.988	N	UCI
Bone	485	02	.510	.510	Y	ElemStatLearn
BostonHousing	506	13	.868	.868	Y	mlbench
CPUPerformance	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
RedWine	1,599	11	.608	.608	Y	UCI
WhiteWine	4,898	11	.537	.537	Y	UCI

Example Two: Automatic Data Preparation

- Given multiple predictors from each application, we subject them separately to the forward stepwise and the all-possible-subsets (best-subsets) method: AICC statistic as the entry/removal criterion
- When the optimal model from each variable selection method is finally identified and estimated, we evaluate the degree of model fit using the correlation coefficient between the observed and predicted values of the DV, an approach adopted by similar studies on subset selection/feature selection, such as Karagiannopoulos, et al. [31]
- A summary of the model search/evaluation process is also found in Table 1

Example Two: Automatic Data Preparation

- 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 finally fitted model from each variable selection method provided in the LINEAR procedure
- 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 [32, 39, 50]

Discussions

- Through the two examples, we may say that the LINEAR procedure provides an effective, new solution to linear regression modeling in SPSS. Compared with the traditional REGRESSION procedure, the LINEAR procedure functions well as its substitute: It provides almost everything found in the traditional procedure, but it also offers additional, typically more advanced features not available in the traditional procedure
- Here, we would like to discuss several additional issues regarding the procedure with the hope to address some philosophical concerns, suggest several possible improvements, and advance new research

Discussions: Automation versus Human Input

- There are mixed feelings about the LINEAR procedure which 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: Some people hate the procedure, like Field [15], whereas others disagree and argue the procedure is a great idea [36]
  - People who hate it:
    `http://www.youtube.com/watch?v=pltr74IlxOg`
  - People who like it:
    `http://www.youtube.com/watch?v=JIs4sMxAFg0`

Discussions: Automation versus Human Input

- We argue that at least the point of being automatic is well justified in the literature (say, in data mining), particularly when the data are huge:
  - Big data are an outgrowth of today's digital environment which 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 us in transforming the vast amounts of data into useful information
    - Automatically analyze the data, automatically classify it, automatically summarize it, automatically discover and characterize trends in it, automatically flag anomalies, among other things [22, 42, 53]

Discussions: Automation versus Human Input

- Given huge 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 [30, 55]
- 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

Discussions: Automation versus Human Input

- On the other hand, despite the need for automation in data mining, we also argue that automation is no substitute for human input [1, 33]
- Under the LINEAR procedure, human participation is also needed:
  - Before the analysis, the researcher needs to review the literature extensively to decide how to phrase the question(s) correctly and what variables should be measured to collect data that could be used to answer the question(s)
  - 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
  - After the analysis, given multiply selected models from subset selection, the researcher needs to factor in theoretical considerations before making the decision on the choice of the very final model

Grounds for Improvements

- There is currently only a single diagnostic statistic built in for identifying and flagging influential cases, which is insufficient because researchers tend to use multiple criteria for that purpose [5]
  - Consider at least one measure from each of the groups of diagnostic statistics as recommended by Chatterjee and Hadi [9]
- The procedure currently does not make available many of the details of the linear modeling process. 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
  - We would have assessed the model ensemble features of the procedure to see if they are truly able to improve predictions, had certain statistics (analysis weights, model weight, and residuals at each base model, etc.) been available

Grounds for Improvements

- Two other features that we believe should be added to the LINEAR procedure are data partitioning/splitting and forced entry of variables (terms). These two procedures are readily available in other statistics/data mining programs
  - The partitioning of data capability splits the data into two or three sets for model training, model validation, and sometimes, model testing
  - The forced entry capability is useful for a hierarchical approach to regression modeling

Grounds for Improvements

- The procedure primarily offers two ensemble methods: 1) Bootstrap aggregating or bagging by Breiman [6], and 2) adaptive boosting or AdaBoost by Drucker [13], and Freund and Schapire [18, 19]
- The ensemble techniques are designed to create a model ensemble/committee containing multiple component/base models which are *averaged* in a certain manner to improve the stability/accuracy of predictions; and the techniques can be incorporated into a common variable selection method (genetic algorithm, stepwise method, all-possible-subsets method, etc.)

Grounds for Improvements

- 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 component/base models
- This combination of subset selection and machine learning algorithms has recently become popular. For example, Liu, Cui, Jiang, and Ma [35] apply the ensemble neural networks with combinational feature selection to the microarray experiments for tumor classification and they have obtained remarkably improved results [2]

Grounds for Improvements

- 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 [4, 6, 12, 13, 19, 35, 46]
- Therefore, it would also be interesting to assess the performance of the ensemble capability of the LINEAR procedure with/without also conducting subset selection of variables in terms of improving the predictions

Conclusions

- The study demonstrates the effectiveness of two features of the LINEAR procedure in SPSS Statistics using benchmark applications
- Through the demonstration, the study argues for the use of the LINEAR program as a substitute for the traditional REGRESSION procedure
- In the end, the study 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 I


Ankerst, M. (2003). The perfect data mining tool: Interactive or automation - Report on the SIGKDD-2002 Panel. *SIGKDD Explorations, 5*(1), 110-111.


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


Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science.


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). NYC, NY: 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*. NYC, NY: Springer.

References II


Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 19*(1), 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). NYC, NY: Wiley.


Field, A. (2013, March 12). SPSS Automatic Linear Modeling. Retrieved May 12, 2013, 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*(4), 499-511.

References III


Halvorsen, K. (2013). ElemStatLearn: Data sets, functions and examples from the book: "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman [Computer software manual]. Retrieved from `http://cran.r-project.org/web/packages/ElemStatLearn/index.html` (R package version 2012.04-0)


Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). San Francisco, CA: 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*. NYC, NY: Springer.


Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). NYC, NY: 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, IL: SPSS, Inc.

References IV


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 [Computer software manual]. Retrieved from `http://cran.r-project.org/web/packages/mlbench/index.html` (R package version 2.1-1)


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, August 22). *How to use automatic linear modeling*. Retrieved May 12, 2013, 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., & Nes, T. (2005). Ensemble methods and partial least squares regression. *Journal of Chemometrics, 18*(11), 498-507.

References V


Miller, A. J. (2002). *Subset selection in regression* (2nd ed.). NYC, NY: 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 June 08, 2013, 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: SAS Institute, Inc.


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


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*. NYC, NY: Springer.


Schonlau, M. (2005). Boosted regression (boosting): An introductory tutorial and a STATA plugin. *The STATA Journal, 5*(3), 330-354.


StackExchange. (2011). *Is automatic linear modelling in SPSS a good or bad thing?* Retrieved May 12, 2013, from `http://stats.stackexchange.com/questions/7432/is-automatic-linear-modelling-in-spss-a-good-or-bad-thing`

References VI


Stine, R. A. (2005). *Feature selection for models in data mining*. Paper presented at the data mining conference at The College of New Jersey (TCNJ), 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*(1), 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 Scienti?c.


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

Questions & Answers

THANK YOU!

Questions?