

Using JMP® to apply Decision Trees and Random Forests as Screening Tools for Limiting Candidate Predictors in Regression Models

Jason Brinkley, Ph.D. American Institutes for Research

Abstract

There are many techniques for evaluating candidate predictors in regression models. Some, such as stepwise regression, have been well studied and with known limitations. Others, such as shrinkage techniques (i.e. LASSO), have become increasingly popular and are showing true potential in helping to provide quality regression models for estimating effects in a multiple variable environment. Many of these techniques become difficult in the world of big data, especially if that data is long (many columns or variables) and short (fewer numbers of observations or rows). Tree based methods are a good alternative as a framework for data exploration and identification of variables to be used in building a quality regression model. Indeed tree methods provide an entirely different framework for model building that can oftentimes provide better predictions. However, when the main goal is still effect estimation, tree methods can be a very useful screening tool. This work examines the practice based on several examples using options very commonly found in both JMP® and JMP Pro software. The focus is on using EITHER single classification trees as well as so-called 'Random' or 'Bootstrap' Forests.

Introduction

Multiple regression has emerged as a stable and powerful technique for simultaneous exploration of many different potential predictor variables and how they may impact a target outcome of interest. The technique has been used across virtually every data driven discipline and has become a staple technique in forming statistical models that help to understand how different factors impact the world around us. In many modern scenarios, there is often a great wealth of available data and decisions need to be made to form statistical models that are more limited in scope. Sometimes those decisions are clinically based and there is a need to form parsimonious models that give as complete a view as possible in a way that lends itself well to interpretation. Other times there are statistical drivers for limiting the size and scope of statistical models, such as sample size or collinearity between candidate predictors. Despite the reasons for such limiting, much attention has been paid toward the so called 'variable selection' problem.

This paper briefly introduces the traditional and modern concepts for dealing with variable selection, then introduces tree based models as an alternative to linear regression models. Then through the use of a targeted example, tree methods integrate with multiple regression to help as an early screening tool for exploring or limiting candidates in regression models. Output with traditional model selection (stepwise), modern model selection (*lasso*), and combining trees/multiple regression results are shown for comparison. The interested reader should see Harrell (2015) and James (2013) for a very complete consideration of all things regression, variable selection, and tree based methods. This paper assumes the reader has a basic working knowledge of multiple regression methods and associated output in JMP®.

Model Selection in multiple linear regression

The traditional methodology for dealing with multiple candidate predictors is to use so called 'stepwise' regression techniques. These come in three types: Forward, Backward, and Stepwise. In Forward Selection, each candidate predictor is screened for possible inclusion in the regression model and one predictor chosen via some selection criteria (say p -values or measures like Akaike's Information Criterion, AIC). That is to say that all potential predictors are screened and the variable with the highest association to the outcome is selected for inclusion into the regression model. Next, the remaining variables are screened and the predictor with the highest association with the outcome (in a model that already contains the first selected variable) is included next. The process continues until no

candidate variables meet some inclusion threshold (say $p < 0.05$). In contrast, Backward Selection begins with a model containing all potential predictors and iteratively removes variables, one at a time, until all candidate predictors meet some inclusion threshold. Lastly, Stepwise Selection is a combination of Forward and Backward Selection methods.

Stepwise methods were the staple for multiple linear regression for many years, even though there are significant problems with the technique. These problems include multiple testing biases and biases in both regression parameters and measures of fit such as R^2 , coefficient of determination. Harrell suggests that if Stepwise methods are to be utilized that Backward methods would be preferred; however there is an additional problem here in that backward selection specifically fails if there are more candidate predictors than degrees of freedom allowed in the model. That is to say that if your number of candidate variables is larger than your sample size, Backward Selection can't start with a 'full model'.

Modern model selection techniques attempt to address these problems by creating different algorithms for determining number of variables in a regression model. The most popular of these emerging methods center on the concept of 'penalization'. The idea is that regression coefficients are given some sort of constraint to prevent bias and that constraint 'shrinks' the coefficients of less important variables to smaller values, eventually to zero. Many of these constraints are centered on the sample size and variables whose parameter estimates shrink to zero are not considered to be part of the model. The most popular of these penalized techniques is Tibshirani's *Lasso* (least absolute shrinkage and selection operator) method (1996). Shrinkage based variable selection techniques arrive at variable reduction or model selection in a very indirect way because no variables are actually 'removed' so to speak, only that their parameter estimates become zero so there is no need to continue to consider them as part of the statistical model. While *Lasso* has great potential for constructing regression models reduced for parsimony, it is not immediately designed to yield standard errors for parameter estimates or to do statistical inference on candidate variables. Tibshirani suggests the bootstrap and Harrell has a more thorough treatment of this in the previously cited text.

Tree based methods for statistical models

As an alternative to multiple linear regression, tree-based methods provide an alternative approach with increasing popularity. James (2013) has a deep overview of the methods which encompass decision tree approaches such as regression and classification trees along with popular techniques such as bagging, boosting, and random forests. For this work, we will focus on the regression tree approach and the subsequent random forest. In general, tree-based methods fall into a set of computationally intensive methods where computer algorithms iteratively scan and split the data into categories based on potential predictors with the end goal of maximizing differences in a target outcome given some constraint or measure of interest. That is to say if the outcome is say, test scores, and predictors are person level characteristics such as age, race and gender then the idea is to split the data up into increasingly homogeneous subgroups in such a way as to maximize the differences in something like average test score across groups. The 'tree' terminology comes from the idea that the data is split multiple times and each subsequent splitting builds upon itself to create increasingly specialized subgroups. An example tree from James (2013) is shown below:

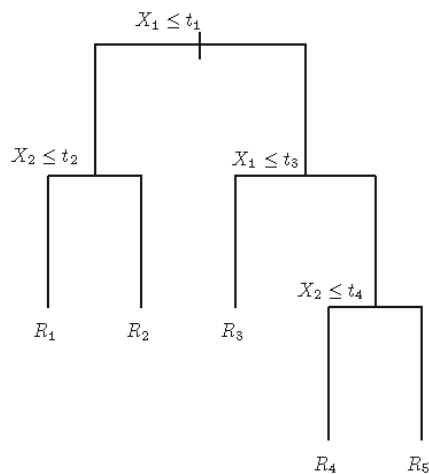


FIGURE 1. EXAMPLE DECISION TREE

Here we see five distinct subgroups based on splitting two variables at four points. Those splits are determined by some predefined criteria with a goal of creating disparate subgroups in relation to a specific outcome. Individual tree models have some well-known limitations such as overfitting and lack of exact reproducibility. Random forests have grown as the most common technique for harnessing the true potential of trees in using trees along with bootstrapping to create an ‘ensemble’ of trees from which more general information can be obtained. It is important to note that tree methods generally center on prediction and not effect estimation, that is to say that statistical inference is not a key component of these methods and in general there is not an analog to testing parameter estimates or regression beta coefficients here. Instead we explore how often different variables show up in tree splits either in individual trees or in a forest. We do have, however, the capacity to measure R^2 and AIC for these models and can use them for comparative purposes.

Example

Mann (2014) explored intravenous (IV) catheterizations from a sample of 592 aged 2 days to 18 years in a pediatric hospital setting. The data is available online [here](#) along with a full exploration of the number of attempts to successful IV start; this analysis will instead consider total active time to successful IV which was skewed and transformed using a natural log transformation. The predictors of interest are listed in the table below:

Variable Name	Label	Variable Name	Label
SHIFT	Shift (Day or Night)	COOPCH1	Nurse Assessed Cooperative Child First Attempt (Yes or No)
AGE	Age of Patient	OSBDA1	Patient Satisfaction Score (First Attempt)
GENDER	Gender of Patient	FAMPRSNT	Family is Present
WEIGHT	Weight of Patient	SUPPSTAFF	Support Staff is Present
PREVIV	Patient Ever Had an IV (Yes or No)	CHLDLFPREP	Child Specific Support Staff is Present
LOSTIV	Patient Lost and IV this Visit (Yes or No)	EXPRTN	First Attempt Nurse is an Expert
NUREXP1	First Attempt Nurse Experience (1-5)	OSBDM	Patient Satisfaction Score (Entire Process)
COMPY1	First Attempt Nurse Competence (1-5)	USEDASSTDV	Nurse Used an Assisted Device to Insert IV
DIFF1	Nurse Assessed Difficult First Attempt (Yes or No)	Dehydrated	Patient is Dehydrated (Yes or NO)

TABLE 1. POTENTIAL PREDICTORS FOR ACTIVE TIME TO SUCCESSFUL IV

The analysis below is only meant to demonstrate the potential for using forest models in addition to traditional regression methods and is not intended as a complete statistical analysis of this data. The goal is to find quality statistical models that look at time to successful IV as a multivariate function of the list in table 1.

Regression Model Selection

We start with a regression model of all predictors in the data, which has been fit using the 'Fit Model' feature in JMP® Pro 12. Note that the "full model" uses 531 of the 592 available data points and has an overall R-square value of 0.287. The adjusted R-square in combination with several predictors showing up as non-significant (i.e. p -values greater than 0.05) suggest that this model may be slightly overfit. Here it is important to note the intended purpose of regression modeling, this is a fine model from the perspective of wanting to know which predictors may have unique impact and whether certain variables predict the outcome of interest after adjusting for other variables in the statistical model.

Outcome - Natural Log of Total Active Minutes

Summary of Fit

RSquare	0.28706
RSquare Adj	0.254718
Root Mean Square Error	0.732243
Mean of Response	2.12134
Observations (or Sum Wgts)	531

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	23	109.45572	4.75894	8.8756
Error	507	271.84344	0.53618	Prob > F
C. Total	530	381.29916		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SHIFT	1	1	1.850364	3.4510	0.0638
AGE	1	1	0.000621	0.0012	0.9729
GENDER	1	1	0.299894	0.5593	0.4549
WEIGHT	1	1	0.012353	0.0230	0.8794
PREVIV	1	1	0.078560	0.1465	0.7020
LOSTIV	1	1	0.422014	0.7871	0.3754
NUREXP1	3	3	1.589687	0.9883	0.3980
COMPY1	4	4	5.598115	2.6102	0.0349*
DIFF1	1	1	5.243006	9.7784	0.0019*
COOPCH1	1	1	2.576671	4.8056	0.0288*
OSBDA1	1	1	5.455462	10.1747	0.0015*
FAMPRSNT	1	1	0.338629	0.6316	0.4272
SUPPSTAFF	1	1	2.365043	4.4109	0.0362*
CHLDLFPREP	1	1	3.670311	6.8453	0.0092*
EXPRTRN	1	1	12.556639	23.4187	<.0001*
Dehydrated	1	1	9.988779	18.6295	<.0001*
USEDASSTDV	1	1	12.647067	23.5873	<.0001*
OSBDM	1	1	5.589310	10.4243	0.0013*

FIGURE 2. FULL REGRESSION MODEL WITH ALL PREDICTORS OF INTEREST

In a first attempt to limit the number of potential predictors, the same variables are fit to a stepwise regression model using backward methods and p -value for inclusion of the JMP default 0.10. The results are shown below and show that a reduced model maintains a number of significant predictors with approximately the same predictive capacity with R-square measuring 0.274 (adjusted R-square is actually higher than the full model).

Fit Group

Outcome - Natural Log of Total Active Minutes

Summary of Fit

RSquare	0.273546
RSquare Adj	0.25716
Root Mean Square Error	0.731176
Mean of Response	2.13788
Observations (or Sum Wgts)	545

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	12	107.09694	8.92474	16.6937
Error	532	284.41673	0.53462	Prob > F
C. Total	544	391.51367		<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SHIFT	1	1	1.978045	3.6999	0.0549
NUREXP1(1-2&3&4)	1	1	2.602139	4.8673	0.0278*
COMPY1(1&2&3&4-5)	1	1	4.933458	9.2280	0.0025*
DIFF1	1	1	5.712913	10.6860	0.0011*
COOPCH1	1	1	2.985677	5.5847	0.0185*
OSBDA1	1	1	4.624028	8.6492	0.0034*
SUPPSTAFF	1	1	2.157356	4.0353	0.0451*
CHLDLFPREP	1	1	4.428752	8.2840	0.0042*
EXPRTRN	1	1	12.594264	23.5575	<.0001*
Dehydrated	1	1	10.235241	19.1450	<.0001*
USEDASSTDV	1	1	15.472970	28.9421	<.0001*
OSBDM	1	1	5.268286	9.8543	0.0018*

FIGURE 3. BACKWARD SELECTION MODEL OUTPUT

Next, contrast that with JMP output using Adaptive Lasso with the default AICc Validation. Note that in JMP the default criteria for using Backward Selection and Lasso are different. Even with that differential in mind, all three regression outputs agree that certain characteristics (age, gender, and weight for example) do not seem to have an association with time to start an IV after adjusting for health characteristics, setting and nurse characteristics.

Outcome - Natural Log of Total Active Minutes

Model Launch

Adaptive Lasso with AICc Validation

Model Summary		Estimation Details	
Response	Log[TOTACTVMIN]	Number of Grid Points	150
Distribution	Normal	Minimum Penalty Fraction	0
Estimation Method	Adaptive Lasso	Grid Scale	Linear
Validation Method	AICc		
Mean Model Link	Identity		
Scale Model Link	Identity		
Measure	Training		
Number of rows	592		
Sum of Frequencies	531		
-LogLikelihood	578.38191		
BIC	1263.4348		
AICc	1191.9568		
Generalized RSquare	0.2798072		

Solution Path

Parameter Estimates for Centered and Scaled Predictors

Parameter Estimates for Original Predictors

Effect Tests

Source	Nparm	DF	Wald ChiSquare	Prob > ChiSquare	
EXPRTRN	1	1	42.663303	<.0001*	
USEDASSTDV	1	1	24.657745	<.0001*	
Dehydrated	1	1	17.809757	<.0001*	
OSBDA1	1	1	9.1921157	0.0024*	
DIFF1	1	1	8.7021276	0.0032*	
OSBDM	1	1	7.8606875	0.0051*	
CHLDLFPREP	1	1	5.5439276	0.0185*	
SUPPSTAFF	1	1	2.669904	0.1023	
NUREXP1	3	1	2.6423073	0.1041	Levels removed: 2
COOPCH1	1	1	2.5345726	0.1114	
SHIFT	1	1	2.4076799	0.1207	
COMPY1	4	4	7.2387019	0.1238	
AGE	1	0	0	1.0000	Removed
GENDER	1	0	0	1.0000	Removed
WEIGHT	1	0	0	1.0000	Removed
PREVIV	1	0	0	1.0000	Removed
LOSTIV	1	0	0	1.0000	Removed
FAMPRSNT	1	0	0	1.0000	Removed

FIGURE 4. LASSO MODEL OUTPUT

Trees and Forests for candidate Reduction

Turning to tree based methods, start with a simple regression tree of the full data. Splits occur in order of importance, in terms of maximizing differences in means of the outcome. So for instance, the first split is between expert RN and non-experts. With a difference of natural logs of $3.24 - 2.10 = 1.14$ or a difference in minutes of $e^{(3.24)} - e^{(2.10)} = 17.4$ active minutes against experts. Note the finding is not intuitive but we see that expert nurses only stick 33 patients. Looking down the rest of the tree we see a mix of different predictors and the end result are a set of groupings that immediately interest across multiple variables. Tree-based methods demonstrate a capability for interactions in a way that are just not immediately possible in regression models.

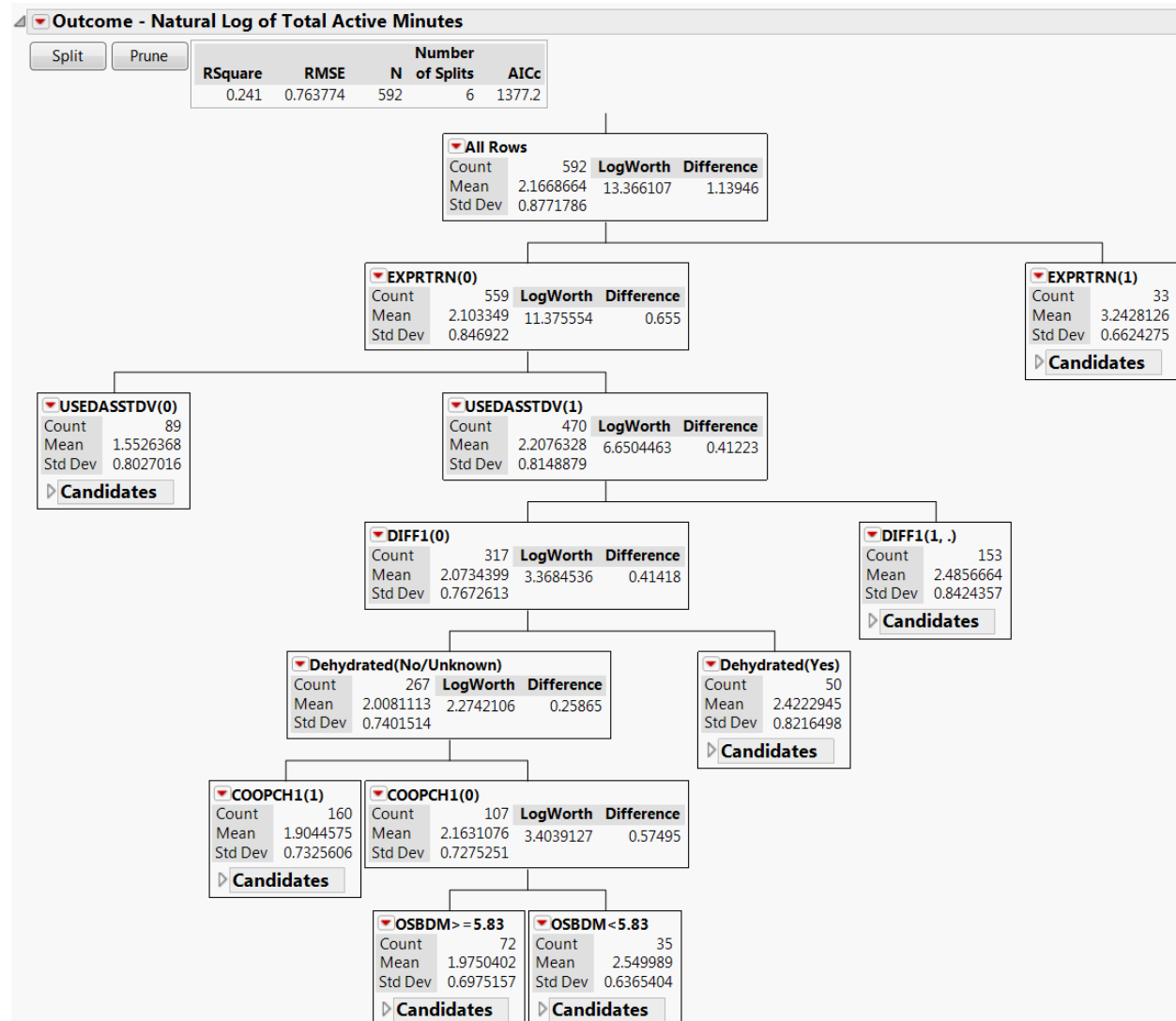


FIGURE 5. REGRESSION TREE FIT TO FULL DATA

Do note that the R-square for the current model is approximately the same as our regression models, indicating that neither modeling framework shows superior predictive capacity. Likewise note that the simple tree model does not utilize patient characteristics age, gender or weight; agreeing with previous output.

Turning toward the JMP® Pro feature 'bootstrap forest' which fits random forest models, we see the beginnings of a different story. Readers familiar with best practices in predictive modeling will note that the full data was fit to the forest model, instead of the usual training and validation samples. Also note that individual trees are not shown and the measure of variable impact is the 'Column Contributions' feature which measures the number of splits and model contribution that each variable has to the ensemble forest.

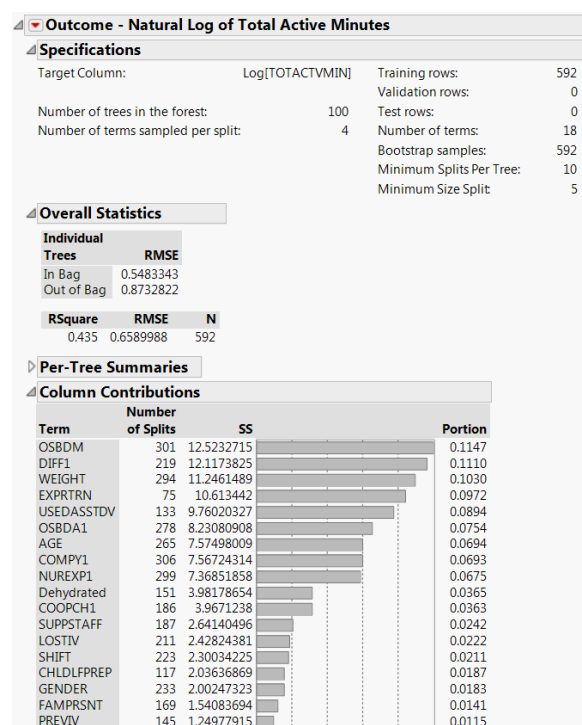


FIGURE 6. JMP PRO BOOTSTRAP FOREST OUTPUT

In direct contrast to previous output, note that 'Weight' is listed as a predictor with a lot of impact in the ensemble forest model. Likewise, this output seems to rely more on OSBDM patient satisfaction and nurse perceived difficulty rather than the expertise of the nurse or the use of assisted devices which show up as important variables in all of the other statistical models. Simply put, the variables that this model deems 'important' are very different than what is seen in all the other models.

Regression Revisited

Returning to regression analysis, the goal is to somehow incorporate the effect of weight into a statistical model that gives similar R-square values but could explain why neither Backward Selection nor Lasso selected it to be a part of the model. Many different models were explored (results not shown) with the focus of exploring interactions that traditional model selection methods tend to ignore without specific user input. In the end, the best solution seems to come from the output shown in figure 7. Note that when a combination of effects for support staff presence, child weight, and their interaction are included in the model we see significant effects. However, those effects all seem to become spurious when also adjusting for the use of an assisted device. This result may be clinically important in that the use of assisted devices may depend on the size of the child (say infants versus others) or the absence of additional staff. If model interpretation is important, then the regression output in figure 7 offer something of competing views of the data. Parsimony may suggest a smaller number of factors, however the implication from such models would be that maybe assisted devices should be used in all patients which may not be supported by the scenarios from which the data are observed.

Outcome - Natural Log of Total Active Minutes						Outcome - Natural Log of Total Active Minutes					
Whole Model						Whole Model					
Summary of Fit						Summary of Fit					
RSquare			0.250182			RSquare			0.268944		
RSquare Adj			0.240462			RSquare Adj			0.258094		
Root Mean Square Error			9.212265			Root Mean Square Error			9.10471		
Mean of Response			12.00821			Mean of Response			12.00821		
Observations (or Sum Wgts)			548			Observations (or Sum Wgts)			548		
Analysis of Variance						Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio		Source	DF	Sum of Squares	Mean Square	F Ratio	
Model	7	15290.668	2184.38	25.7392		Model	8	16437.403	2054.68	24.7863	
Error	540	45827.546	84.87		Prob > F	Error	539	44680.810	82.90		Prob > F
C. Total	547	61118.213			<.0001*	C. Total	547	61118.213			<.0001*
Effect Tests						Effect Tests					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
OSBDM	1	1	1352.4715	15.9366	<.0001*	OSBDM	1	1	1332.6344	16.0760	<.0001*
DIFF1	1	1	1297.4413	15.2881	0.0001*	DIFF1	1	1	1227.4923	14.8077	0.0001*
EXPRTRN	1	1	6181.5793	72.8394	<.0001*	EXPRTRN	1	1	6053.0275	73.0198	<.0001*
OSBDA1	1	1	1478.4729	17.4213	<.0001*	OSBDA1	1	1	1359.5317	16.4005	<.0001*
SUPPSTAFF	1	1	62.4414	0.7358	0.3914	SUPPSTAFF	1	1	208.7517	2.5182	0.1131
WEIGHT	1	1	838.9698	9.8858	0.0018*	WEIGHT	1	1	318.1003	3.8374	0.0506
SUPPSTAFF*WEIGHT	1	1	425.7829	5.0171	0.0255*	SUPPSTAFF*WEIGHT	1	1	172.2757	2.0782	0.1500
						USEDASSTDV	1	1	1146.7352	13.8335	0.0002*

FIGURE 7. REGRESSION OUTPUT TO EXPLORE WEIGHT

Discussion

Multiple linear regression is a well-tested and often used mechanism for which one can gain a larger sense of the interplay between multiple variables and how they associate or potentially influence an outcome of interest. There is often a need to reduce a model with too many terms or to fit models to data for which there are more candidate predictors than are allowed in a fully saturated model. Model selection techniques have arose to help create more useful models by either trimming away less important variables or developing small models with a set number of very important key indicators. Tree-based methods have emerged as an interesting alternative to explore these same associations or potential influence in different manner. Random forests can offer unique insights which can be considered on their own or in combination with regression models to help provide a more complete picture. The true potential in random forests come from the notion that they do not rely on statistical inference and in some ways provide information regarding associations free from issues of type 1 error or other statistical quandaries. Although caution is also recommended, the true downfall of model selection techniques is that they allow an analyst to 'not think' about the data. Augmenting tree-based methods to regression analysis is not a substitute for this clinical and analytic insight.

References

- Harrell, F. (2015). Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: springer.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.
- Mann, J., Larsen, P., & Brinkley, J. (2014). Exploring the use of negative binomial regression modeling for pediatric peripheral intravenous catheterization. Journal of Medical Statistics and Informatics, 2(1), 6.

Acknowledgments

Special thanks to Professor Richard De Veaux from Williams College whose workshop on Practical Data Mining: Challenges and Solutions provided the original insights and inspiration for this document.

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Name: Jason Brinkley
Enterprise: American Institutes for Research
Address: 100 Europa Drive Suite 315
City, State ZIP: Chapel Hill, NC 27517
E-mail: jbrinkley@air.org

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.