Application of CART, Neural Networks, and Generalized Additive Models: A Case Study

W. Holmes Finch **Mei Chang** **Andrew S. Davis**
Ball State University

Statistical prediction of an outcome variable using multiple independent variables is a common practice in the social and behavioral sciences. For example, neuropsychologists are sometimes called upon to provide predictions of pre-injury cognitive functioning for individuals who have suffered a traumatic brain injury. Typically these predictions are made using standard multiple linear regression models with several demographic variables (e.g., gender, ethnicity, education level) as predictors. Prior research has found conflicting evidence regarding the ability of such models to provide accurate predictions of outcome variables such as full-scale intelligence (FSIQ) test scores. The current study had two goals: 1) to demonstrate the utility of a set of alternative prediction methods that have been applied extensively in the natural sciences and business but which have not been frequently explored in the social sciences and 2) to develop models that can be used to predict premorbid cognitive functioning in preschool children. Prediction of Stanford Binet 5 FSIQ scores for preschool aged children is used to compare the performance of a multiple regression model with several of these alternative methods. Results demonstrate that classification and regression trees (CART) provided more accurate prediction of FSIQ scores than the more traditional regression approach. Implications of these results are discussed.

Clinical neuropsychologists are frequently required to determine if an individual has experienced changes in intellectual functioning resulting from a neurological insult such as a traumatic brain injury (TBI) or stroke. Accurate diagnosis and the determination of functional decline relies largely on a clinician's ability to compare current test performance to an estimate of premorbid (i.e., prior to injury) performance. It is common in clinical practice for neuropsychologists to use a discrepancy between a predicted and an obtained test score to assist in the determination of whether organic impairment or a progressive disease is present. Thus, an accurate estimation of premorbid intelligence is necessary to prevent errors such as under or overestimation of a patient's level of cognitive decline (Griffin, Mindt, Rankin, Ritchie, & Scott, 2002) and the availability of techniques demonstrating good validity and reliability for predicting premorbid intellectual functioning is a central concern of clinicians. When premorbid ability levels can be reasonably estimated, a diagnosis can be made with confidence and cognitive rehabilitation programs can be properly designed, monitored, and modified (Reynolds, 1997).

Traditional Methods of Prediction

A variety of approaches have been proposed and developed for the estimation of premorbid ability estimation (PAE), including (a) historical achievement-based and standardized group assessment data (e.g., Baade & Schoenberg, 2004; Schinka & Vanderploeg, 2000); (b) "hold/don't hold tests" estimates (Blair & Spreen, 1989; Lezak, Howieson, Loring, Hannay, & Fischer, 2004); (c) best current performance estimates (Lezak, 1995); (d) demographic-based regression formulas (e.g., Barona, Reynolds, & Chastain, 1984); (e) combinations of demographic and actual performance data (e.g., Schoenberg, Lange, & Saklofske, 2007a; Schoenberg, Lange, & Saklofske, 2007b; Schoenberg, Lange, Saklofske, Suarez, & Brickell, 2008; Schoenberg, Scott, Duff, & Adams, 2002; Vanderploeg, Schinka, & Axelrod, 1996); and (f) current word reading ability tests (e.g., Blair & Spreen; Wechsler, 2003). However, each approach has been shown to have some limitations in application.

Using Multiple Linear Regression to Predict Premorbid IQ

An alternative to these more ad hoc approaches to predicting premorbid IQ involves the use of multiple linear regression (MLR) to estimate IQ. Researchers in the field have developed models based on demographic variables in conjunction with performance on a task such as word reading or some comparable measure (Sellers, Burns, & Guyrke, 2002; Vanderploeg, Schinka, Baum, Tremont, & Mittenberg, 1998; Yeates & Taylor, 1997), while in other cases, only demographic variables were used. Crawford, Millar, and Milne (2001) found that for adults, the correlation between actual and predicted IQ, based on the demographic variables of education, socio-economic status and age, was 0.76, which was higher than that obtained through clinical judgment. A study focusing on predicting IQ for adolescents included variables such as gender, ethnicity, region of the U.S. in which the subject lived, age, and parental education level (Schoenberg et al., 2007a) and was found to provide predictions of FSIQ. Powell,

Brossart and Reynolds (2003) compared the performance of two regression models of the demographic information estimation formula index (DI) (Barona et al., 1984) and the Oklahoma Premorbid Intelligence Estimate (OPIE) (Krull, Sherer, & Adams, 1995) for estimating premorbid cognitive functioning in adults. Both models are based on linear equations that predict cognitive functioning using demographic variables (age, gender, race, education, occupation, urban/rural residence and current performance) on Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1983) Vocabulary and Picture Completion subtests. Their results demonstrated that the DI approach provided more accurate estimates of cognitive decline but were not as accurate when predicting FSIQ for individuals who did not suffer any brain injury. One issue with either of these approaches is that researchers must have access to all of the variables that serve as inputs to the standardized equations. Another issue is that regression-based estimates of premorbid IQ have been shown susceptible to error, particularly in outer ranges of intellectual function (Veiel & Koopman, 2001). In addition, the MLR model assumes a linear relationship exists between the outcome of interest and the predictors, unless the researcher explicitly includes a non-linear term. However, in many instances, it may be unclear whether a non-linear term should be included, and more importantly what type would be most appropriate.

Much of the focus in the prediction of premorbid IQ has been on adults with relatively little research devoted to predicting cognitive functioning in school-aged and younger children (Schoenberg et al., 2007a). However, some research has been conducted on the use of prediction equations based on only demographic variables with school-aged children (Pungello, Iruka, Dotterer, Koonce-Mills, & Reznick, 2009; Schoenberg et al., 2008; Schoenberg, Lange, Brickell, & Saklofske, 2007; Roberts, Bornstein, Slater, & Barrett, 1999; Sellers, Burns, & Guyrke, 1996). These studies included variables such as parental level of education, ethnicity, gender, age, region of the U.S. in which the child resided and parental occupation, and generally found that they could achieve an R^2 generally around 0.4 when predicting Full Scale IQ (FSIQ). Despite the publication of several studies, the problem of accurately predicting premorbid IQ, particularly in young children, has not been completely solved (Schoenberg et al., 2007b). Furthermore, prior work with adult populations has not definitively demonstrated linear models to be the universally most effective tool for predicting premorbid IQ scores, as was discussed previously. Therefore, alternative methods for prediction should be investigated in order to find the optimal tool(s) for the important task of obtaining reliable estimates of premorbid intellectual functioning for young children who have undergone a neurological insult and suffered from cognitive impairment.

It should be noted that while the focus of this research was on predicting IQ scores for young children using demographic variables, other recent examples using prediction in the social science literature include predicting school counselor evaluations of student performance (Granello, 2010), college student dropout (Nistor & Neubauer, 2010), impact of character education on social competence (Cheung & Lee, 2010) and parent training effectiveness (Lavigne, LeBailly, & Gouze, 2010) to name but a few. In the vast majority of this research some variant of linear regression was used to obtain predictions. However, because it is limited to linear or relatively simple non-linear forms, regression may not always be the optimal choice for this type of research (Berk, 2008).

The goals of this study were to describe some alternative methods of prediction that could be employed in the context of obtaining estimates of premorbid IQ in preschool-aged children. These methods, including Classification and Regression Trees, Neural Networks and Generalized Additive Models, have all been shown to be effective tools for prediction in simulation research, particularly in the presence of non-linear relationships between outcome and predictor variables (Chang, Finch, & Davis, 2011; Finch & Holden, 2010). Given their positive record of performance, and their relative novelty in the social science literature, it was hoped that a manuscript demonstrating how each could be used to solve a real world prediction problem would add to the quantitative methods literature. In addition, each of these alternative models presents the user with great flexibility in terms of model settings and the like, which can have tremendous impact on the final results of the analysis. Thus, in addition to demonstrating how these tools can be used in practice, a second goal of this manuscript is to discuss these various model settings and provide some guidance for their implementation. It should be noted that this paper is not intended to be a comprehensive review of these methods, but rather an introduction that should provide the interested researcher with the basic tools to conduct analyses with these modeling techniques. A number of more comprehensive works are referenced below for those who want to delve deeper (which we encourage wholeheartedly).

Alternative Methods of Prediction

The following is discussion of a number of alternative approaches for prediction that may prove useful when a relationship between variables is not strictly linear. Given some of the problems discussed earlier in the traditional approach of using MLR for predicting premorbid IQ and because very little work has been done examining the prediction of IQ in very young children, these alternative methods may prove to be interesting tools for this task. After a description of these approaches, the results of a study demonstrating how to use these techniques for the task of predicting IQ will be presented.

Classification and Regression Trees (CART)

CART (Breiman, Friedman, Olshen & Stone, 1984) arrives at predicted values for an outcome variable, *Y*, given a set of predictors by iteratively dividing individual members of the sample into ever more homogeneous groups, or nodes, based on values of the predictor variables. It can be thought of as a nonparametric approach because there are no assumptions regarding the underlying population from which the sample is drawn nor the form of the model linking the outcome and predictor variables. CART begins by placing all subjects into one node, or group, and then searches the set of predictors to find the value of one of those by which it can divide the observations into two new nodes, whose values on *Y* are as homogeneous as possible. For each of these new nodes, the predictors are once again searched for the optimal split by which the subjects can be further divided into ever more homogeneous nodes, where homogeneity is always based on the similarity of values of *Y*. This division of the data continues until a predetermined stopping point is reached, when further splits do not appreciably reduce the heterogeneity of the resulting nodes. At this point, the tree is complete and values of *Y* for new individuals can be obtained using the decision tree developed with this original training sample. The data for the new subject are fed into the tree, following the branches from node to node based on the values of the predictor variables until the individual is placed in one of the final, or terminal nodes. The predicted value for *Y* for each individual is then the mean for the training sample in this terminal node.

CART has a tendency to overfit the training data when developing the initial prediction tree (Berk, 2008), meaning that the final model may be too closely associated with the training sample to generalize well to other samples from the same population. In addition, trees produced by CART can sometimes contain terminal nodes with few individuals or terminal nodes that are very heterogeneous, which is characteristic of tree instability and an inability to generalize to the broader population (Hothorn, Hornik, & Zeileis, 2006). One commonly used method for ameliorating overfitting is the practice of pruning trees. This process, which is demonstrated in the results section below, involves the removal of terminal nodes that appear to provide little predictive power, and when included in the final model might lead to overfitting of the training sample. Pruning is not an automated process, and requires the direct involvement of the researcher, typically through an examination of results for multiple pruned trees, as is discussed below. In order to ascertain how much pruning is necessary, the researcher typically refers to a plot of the number of nodes by total model deviance. Total deviance for a tree corresponds to the sum of the sum of squared residuals within the terminal nodes (i.e., the sum of squared differences between the predicted and actual IQ scores in this example). The larger the deviance value, the greater the heterogeneity of scores within the terminal nodes, and the worse the CART solution. As terminal nodes are removed from the tree, the deviance will increase because more heterogeneous individuals are grouped together in the new, larger terminal node. The decision regarding the number of terminal nodes to retain in the pruned tree is based upon balancing this increased heterogeneity with a desire to have a more parsimonious tree, and one that generalizes better to other samples.

Neural Networks (NNET)

Another prediction method examined in this study is Neural Networks (NNET) (e.g., Marshall & English, 2000; see Garson, 1998 for a more technical description of the method). NNETs create a prediction model for *Y* by using a search algorithm that examines a large number of subsets of the predictors, as well as interactions among them. Interactions and powers of the predictors (referred to as hidden layers) are computed in conjunction with weights that are akin to regression slopes. Main effects and hidden layers to be included in the final model are selected by the algorithm so as to minimize the least squares criterion used in standard linear regression (i.e., minimizing the sum of squared differences between the observed and predicted values). The hidden layers are generally much more complex than the two and three way interactions common in regression, involving several predictors and higher order versions of the predictors in a single interaction (Schumacher, Robner, & Vach, 1996). In addition, they

are not specified *a priori* by the researcher, but instead are identified by the NNET algorithm based on their contribution to reducing the sum of squared residuals. In order to reduce the likelihood of finding locally optimal results that will not generalize beyond the training sample, random changes to the subset of predictors and interactions, not based on model fit, are also made. This method of obtaining optimal model fit is known as back-propagation, where the difference between actual and predicted outputs is used to find optimal weights for main effects and hidden layers. It is one of the most commonly used approaches in NNET applications (Garson, 1998).

A primary strength of NNET models is that they can identify complex interactions among the predictor variables in the hidden layer that other approaches may ignore (Marshall & English, 2000). For example, whereas in regression it is common to express the interaction of two predictors as their product, or to square or cube a single variable if the relationship with the response is believed not to be linear, a NNET will create hidden layers as weighted products of perhaps several variables, thus allowing the model to be influenced by the predictors to varying degrees. The result is that fairly obscure relationships between the outcome and predictors will be automatically identified without the researcher having to explicitly include them in the model.

Conversely, this ability to identify extremely specific models to fit the data presents a potential problem in that NNETs can substantially overfit the training data used to estimate the model (Schumacher et al., 1996). In order to combat this problem, most NNET models apply what is called weight decay, which penalizes (i.e., reduces) the largest weights found in the original NNET analysis, in effect assuming that very large weights are at least partially driven by random variation unique to the training data. The researcher typically sets the value of the decay parameter, λ, with larger values shrinking the weights for non-linear terms to a greater degree, and thus reducing their impact on the final model, hopefully ameliorating problems of overfitting. Generally speaking, the value of λ for a given problem is selected by examining the ability of the model to correctly predict the outcome variable for a cross-validation sample (Hastie, Tibshirani, & Friedman, 2001). In other words, the decay parameter value associated with the most accurate prediction in the cross-validation sample is the one that is selected.

Another choice that the researcher must make when using NNETs is the number of hidden layers that will be allowed in the final model. The larger the number of hidden layers that are permitted in the model, the more complex the model could become by incorporating higher order non-linear terms. Including more hidden layers has both positive and negative aspects. On the one hand, such models are better able to identify complex relationships among the variables, but on the other, they may lead to overfitting of the training sample. Thus, the researcher is advised to try multiple settings for the number of hidden layers and decide on the optimal setting for this parameter based on the accuracy of predictions for a cross-validation sample (Garson, 1998).

Researchers using NNETs also have control over the range of random starting values for the hidden layer weights that will be used in the model. The algorithm selects initial weight values randomly within a predefined range. The weights are then updated based upon the minimization of the least squares criterion. When the weights are near 0, hidden layers are deemphasized and the model becomes essentially linear in form. As these weights increase, the hidden layers play a greater role in determining predicted values for the outcome variable. In the initial model setup, the randomly selected starting values are typically drawn from a fairly restricted range near 0; e.g. -0.5 to 0.5 in the case of R. However, the researcher can change the range of these starting values and attempt to find the optimal setting based upon prediction accuracy for a cross-validation sample, if (s)he believes that hidden layers will play a more (or less) important role in the final model. It is important to note that if starting values for the weights are too large, the final model performance may be compromised due to overfitting (Hastie et al., 2001). Examples of manipulation of each of these settings are provided in the results section.

Generalized Additive Models (GAM)

GAMs are a class of very flexible models that allow for the linking of *Y* with one or more predictor variables, using a wide variety of smoothing functions common in statistics. Each function is fit using a smoothing technique such as a thin plate spline (default in R), cubic spline or a P spline, with the goal of minimizing the penalized sum of squares criterion (Simonoff, 1996). For an excellent discussion of smoothing and splines, the reader is encouraged to refer to Keele (2008), and the aforementioned Simonoff . The penalized sum of squares (PSS) is based on the standard sum of squared residuals with a penalty applied for model complexity (i.e., the number of main effects and interactions included). The GAM algorithm works in an iterative fashion, beginning with the setting of the model intercept to the

mean of *Y*. Subsequently, the smoothing function of choice is applied to each of the independent variables in turn, selecting the smoothed predictor that minimizes the PSS. This iterative process continues until the smoothing functions stabilize (i.e., the PSS cannot be appreciably reduced further), at which point final model parameter estimates are obtained. The optimal model is typically selected so as to minimize the Generalized Cross Validation (GCV) score, which is based on an approximation of a jackknifed cross validation check of the training data. Essentially, the GCV score is a measure of prediction accuracy based on a sum of squares value when jackknifing (leave one out) is used. However, it is constructed such that actual jackknifing, which can be quite laborious for large datasets, is not necessary. Smaller values of GCV are associated with more accurate and generalizable models. In addition to the GCV, selection of optimal GAMs is also aided by the popular Akaike Information Criterion (AIC), for which smaller values indicate better model fit.

As was the case for CART and NNET, overfitting of the data can also be a problem with GAMs. In order to avoid overfitting, the researcher using GAM can change the smoothing parameter, γ, which appears in the equation for the GCV score. By default γ is set to 1, where larger values correspond to identifying a smoother model as optimal for the data. Kim and Gu (2004) found that γ of approximately 1.4 was effective at correcting the overfitting problem while not compromising model fit. The researcher also has control over the actual smoothing spline to be used in developing the GAM, a selection of which was mentioned above. Indeed, different smoothing splines could be used with different predictor variables, or combinations of these variables. Finally, the researcher has the option of selecting what is essentially the complexity of the smoothing function through the dimension of the function used by the smoothing algorithm. Larger values of this parameter, *k*, allow for more degrees of freedom in the smoother, which corresponds to a potentially more complex smoothing function. Typically, this value is not set extremely high in order to avoid the possibility of overfitting. It should also be noted that in practice, the value of *k* frequently has a minor impact on the final performance of the model in terms of prediction accuracy (Wood, 2006).

Current Study

As mentioned above, of particular interest in the current study is the investigation of how manipulating tuning parameters impacts each of the alternative methods (i.e., CART, NNET, and GAM). In much prior work, these methods have been studied using either default or generally recommended settings for these parameters. However, in actual practice analysis results can change with different values for these tuning parameters. Given that the general recommendation for these methods is to, in fact, try several values for these settings in order to find the optimal model (Hastie et al., 2001), the current study seeks to add to the literature regarding the most effective use of each approach by demonstrating how these settings can be manipulated in a common software package (R). It should be noted, however, that this work is not intended to represent a complete training in the use of these alternative prediction methods. Indeed, for each of them complete books are available to walk the researcher through planning, conducting and interpreting analyses. Rather, this study is intended to introduce interested readers to the basic sequence of using these methods for prediction, and to encourage further investigation of those methods that appear to be most appropriate for a given research scenario. There are many fine texts available for each approach, several of which are included in the references to this manuscript, and we encourage the interested reader to peruse these for a more complete discussion of the fine details of conducting each analysis. We are hopeful, however, that this paper will serve as a strong starting point from which a researcher interested in using one or more of these methods can begin their analysis with some confidence.

Methodology

Participants and Procedures

Participants for this study included 200 (n = 103 females; n = 97 males) preschool children. The sample was obtained from preschool facilities near a mid-sized city in the Midwest Demographic information for the total sample appears in Table 1. Only children who did not receive special education or related services, and whose parental consent was obtained, were included as participants. Once a signed parental permission form was obtained, the children were administered the Stanford-Binet Intelligence Scales – Fifth Edition (SB5; Roid, 2003) under standardized conditions by trained examiners. In addition, selected demographic data were also collected for all study participants.

Instrumentation

The SB5 (Roid, 2003) is an individually administered assessment of IQ appropriate for people between the ages of 2 and 85 years. It is theoretically grounded in the Cattell-Horn-Cattell (CHC) theory and intends to represent 5 CHC factors, including Fluid Intelligence (G*f*), Crystallized Knowledge (G*c*), Quantitative Knowledge (G*q*), Visual Processing (G*v*), and Short-Term Memory (G*sm*). The entire SB5 (5 verbal and 5 nonverbal subtests) was administered to the participants, and generated a Full Scale IQ (FSIQ). In relation to this study, the SB5 FSIQ score was used to indicate the children's comprehensive cognitive abilities. The SB5 was selected for use in this study because it is strongly grounded in CHC theory, has been normed for children as young as those used in this study, and has been shown to be a valid and reliable tool for such assessments.

Table 1. Descriptive Statistics for Total, Training and Cross-Validation Samples

Variable	Total Sample	Training	Cross-Validation
Gender			
Male	97 (48.5%)	73 (48.7%)	24 (48%)
Female	103 (51.5%)	77 (51.3%)	26 (52%)
Ethnicity			
Caucasian	124 (62%)	93 (62%)	31 (62%)
African-American	49 (24.5%)	38 (25.3%)	11 (22%)
Hispanic/Latino	2 (1%)	2 (1.3%)	0 (0%)
Bi-racial	20 (10%)	15 (10%)	5 (10%)
Other	3 (1.5%)	0 (0%)	3 (6%)
No report	2 (1%)	2 (1.3%)	0 (0%)
Father's education			
Less than High school	30 (15%)	24 (16%)	6 (12%)
High school/GED	78 (39%)	56 (37.3%)	22 (44%)
1-3 years of college	44 (22%)	35 (23.3%)	9 (18%)
4+ years of college	29 (14.5%)	21 (14%)	8 (16%)
No report	19 (9.5%)	14 (9.4%)	5 (10%)
Mother's education			
Less than High school	16 (8%)	13 (8.7%)	3 (6%)
High school/GED	48 (24%)	35 (23.3%)	13 (26%)
1-3 years of college	89 (49.5%)	65 (43.3%)	24 (48%)
4+ years of college	39 (19.5%)	31 (20.7%)	8 (16%)
No report	8 (4%)	6 (4%)	2 (4%)
Age (months) Mean	58.86 (5.38)	59.73 (5.50)	60.28 (5.04)
FSIQ (SD)	98.10 (11.81)	98.29 (11.17)	97.54 (13.67)

Prediction Models

The outcome variable of interest was the FSIQ from the SB5, while the predictors included years of education each for mother and father, and the child's age. These predictors were selected because they are typically available for any subject for who predicted IQ is required, and will not be impacted by a CNS injury. They have also been used in prior IQ prediction studies (Sellers et al., 1996). The models used to predict FSIQ with these demographic variables included MLR as well as CART, NNET, and GAM. All analyses were carried out using the R software package (R Development Core Team, 2007). These prediction methods were selected because they have been demonstrated in prior research to be effective tools in predicting continuous outcome variables (Chang et al., 2011; Finch &Holden, 2010).

In order to assess the predictive accuracy of the models, the original sample of 200 subjects was randomly divided into training (N=150) and cross-validation samples (N=50). For each method, the training sample was used to estimate a predictive model, which was in turn applied to the cross-validation sample to obtain predicted values for FSIQ. Prediction accuracy for the cross-validated sample was assessed through the bias of the predicted IQ: $\text{Bias} = \theta_{\text{Actual}} - \theta_{\text{Predicted}}$ and the Root Mean Square Error (RMSE) of the predictions for the cross-validation sample:

$$\text{RMSE} = \sqrt{\frac{\sum\left(\theta_{\text{Actual}} - \theta_{\text{Predicted}}\right)^2}{n}}.$$

Bias serves as a measure of the estimation accuracy, while RMSE reflects both accuracy and precision of the predicted values. In general, results with lower bias and lower RMSE can be viewed as better fitting.

Results

Following is a description of FSIQ prediction results for the cross-validation sample using CART, NNET, and GAM, along with ordinary least squares (OLS) regression, which will serve as the baseline for comparison with the alternative methods. The R commands necessary to run these analyses appear in italics in the text. Table 2 contains the bias and RMSE results for each model.

Table 2. RMSE and Bias Values for Cross-Validation Sample: OLS, CART, GAM and NNET Models

Model	RMSE	Bias
OLS	15.0567	-6.13
CART, 13 nodes	15.2883	-6.24
CART, 10 nodes	15.46269	-6.43
CART, 8 nodes	15.19276	-6.51
NNET, 2 hidden layers	15.05665	-6.13
NNET, 5 hidden layers	15.05665	-6.13
NNET, 10 hidden layers	16.16655	-6.11
NNET, 20 hidden layers	14.71749	-5.2
NNET, 2 hidden layers, decay=0.5	16.43546	-6.08
NNET, 5 hidden layers, decay=0.5	15.48161	-6.35
NNET, 10 hidden layers, decay=0.5	15.08878	-5.18
NNET, 20 hidden layers, decay=0.5	15.31007	-7.06
NNET, 2 hidden layers, decay=0.75	16.16945	-6.34
NNET, 5 hidden layers, decay=0.75	15.27232	-4.96
NNET, 10 hidden layers, decay=0.75	15.67306	-5.88
NNET, 20 hidden layers, decay=0.75	15.67511	-6.67
NNET, 2 hidden layers, range=-1 to 1	15.05665	-6.13
NNET, 5 hidden layers, range=-1 to 1	15.05665	-6.13
NNET, 10 hidden layers, range=-1 to 1	15.27986	-6.59
NNET, 20 hidden layers, range=-1 to 1	15.05665	-6.13
GAM, thin plate	15.06799	-5.49
GAM, cubic	15.87559	-5.95
GAM, thin plate, g=1.4	15.06799	-5.49
GAM, cubic, g=1.4	15.56384	-6.53

CART

CART found a tree with 13 terminal nodes, and underestimated FSIQ in the model was 15.29. In terms of conducting analysis in R, the *library* command loads the tree library (which we would have previously installed in our version of R), and the *iq.cart<-tree(IQ~age+fathered+mothered)* command creates the prediction tree and saves it in the R object *iq.cart*. The descriptive output, produced by the *summary* command, appears below. This output shows us the deviance value for the tree (Residual mean deviance), where larger values indicate a greater difference in observed and predicted FSIQ for the training sample.

```
library(tree)
iq.cart<-tree(IQ~age+fathered+mothered)
summary(iq.cart)

Regression tree:
tree(formula = IQ ~ age + fathered + mothered)
Number of terminal nodes:  13
Residual mean deviance:  83.15 = 11390 / 137
Distribution of residuals:

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-34.330  -6.375   1.255   0.000   6.176  20.860
```

In order to determine how much pruning should be done, we used the following set of commands to create the graph in Figure 1, showing the relationship between the deviance and the number of terminal nodes.

```
iq.cart.prune<-prune.tree(fsiqtrain.cart)
plot(iq.cart.prune)
```

Moving right to left on the x-axis, we can see that the first big increase in deviance occurs between 10 and 9 terminal nodes. Thus, we may elect to fit a tree with only 10 terminal nodes rather than the original 13, using the following commands in R. Note that the subcommand *best=10* requests that the 10 terminal node tree with the smallest deviance be selected. The deviance for this tree

was 85.08, which is not much larger than the 83.15 for the original 13 node tree, suggesting that losing the three weakest terminal nodes did not substantially damage model fit for the training sample. In addition, the mean bias and RMSE values in Table 2 for the 10 node tree were very similar to those for the full tree.

```
iq.cart.prune10<-prune.tree(iq.cart,best=10)
summary(iq.cart.prune10)

Regression tree:
snip.tree(tree = iq.cart, nodes = c(4, 22, 15))
Number of terminal nodes:  10
Residual mean deviance:  85.08 = 11910/140
Distribution of residuals:

   Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
-34.330  -6.000   1.160   0.000   5.784  24.150
```

Likewise, we produced a tree with 8 terminal nodes, which also had very similar bias and RMSE values to the other two trees.

```
iq.cart.prune8<-prune.tree(iq.cart,best=8)
summary(iq.cart.prune8)

Regression tree:
snip.tree(tree = iq.cart, nodes = c(4, 15, 11))
Number of terminal nodes:  8
Residual mean deviance:  88.59 = 12580/142
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-34.330  -6.281   1.465   0.000   5.899  20.380
```

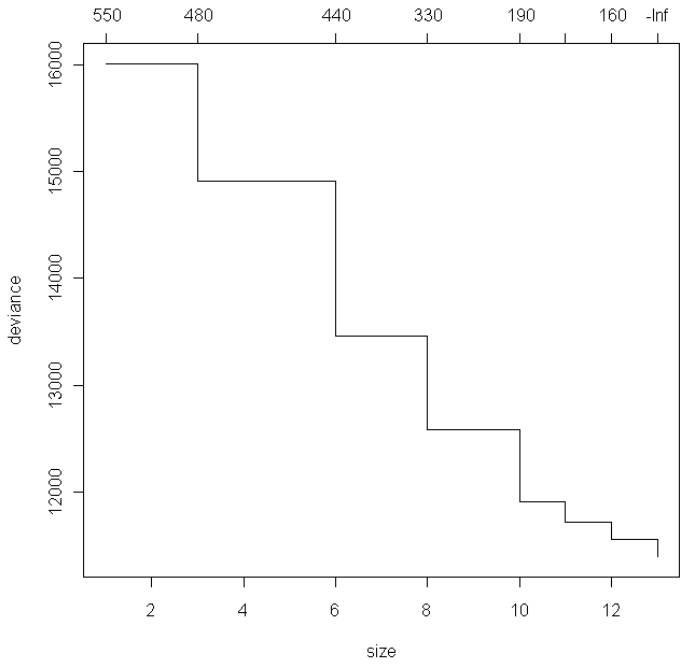

Figure 1.Total tree deviance by number of terminal nodes

Taken together, these results would suggest that for the purposes of predicting FSIQ for the cross-validation sample, the 8 terminal node tree was just as effective as the full 13 node tree. Furthermore, all of the tree models produced generally comparable results to those for the other alternative models included in this study.

NNET

Of the three alternative modeling techniques examined here, NNET has the largest number of potential settings that a researcher can change. In this study, we estimated a large number of NNET models, results of which are presented in Table 2 with regard to prediction accuracy for the cross-validation sample. In addition, compared to the other two approaches featured here, the output from a NNET analysis is not particularly informative regarding either model fit or the actual nature of the model itself. Indeed, the most useful information regarding the fit of a NNET model comes from its ability to accurately predict the outcome variable for the cross-validation sample. As an initial example, the following are the commands for running a basic NNET with 2 hidden layers, and default weight decay of 0 and range of random weight starting values from -0.5 to 0.5.

```
library(nnet)
iq.nnet2<-
nnet(IQ~age+fathered+mothered,size=2,
linout=T,skip=T)

# weights:  14
initial  value 3088737.481193
iter  10 value 28190.551478
iter  20 value 21395.190892
iter  30 value 21354.815727
iter  40 value 21298.496237
iter  50 value 21296.878978
iter  50 value 21296.878961
iter  50 value 21296.878961
final  value 21296.878961
converged
```

The library to be used in this case is nnet, which was previously installed in our version of R. The *linout=T* command is necessary when the outcome variable is continuous, as is the case for FSIQ. The default setting for NNET in R is a categorical outcome, leading to a model corresponding to logistic regression. The *skip=T* subcommand allows for the inclusion of main effects as well as hidden layers in the final model. If this command is set to *F*, there will be no weights directly linking each of the main effects to the outcome variable. We can view the weights by typing the command to the right.

```
summary(iq.nnet2)
a 3-2-1 network with 14 weights
options were - skip-layer connections  linear output units
 b->h1 i1->h1 i2->h1 i3->h1
  0.55  -0.52   0.55  -0.26
 b->h2 i1->h2 i2->h2 i3->h2
 -0.69  -0.57  -0.61  -0.29
 b->o h1->o h2->o i1->o i2->o i3->o
93.04  0.37 -0.21 -0.01  0.12  1.69
```

In this table, b is the intercept, i1, i2, and i3 are the three independent variables, h1 and h2 are the hidden layers, and o is the outcome variable. Thus, we can see that the weight for variable i1, age, to the first hidden layer is -0.52, while the weight of father's education (i2) to this hidden layer is 0.55, and so on. Finally, looking at the last line, we can see that the variable with the largest direct weight to the outcome variable is mother's education (i3), with a value of 1.69. At the same time, neither age nor father's education had a strong direct relationship with FSIQ. In addition, hidden layer 1 (h1), which was dominated primarily by an interaction of age and father's education, had a somewhat stronger impact on FSIQ than did hidden layer 2. It should be noted that from this output, we do not have an indication of which of these weights could be statistically significant in the classic hypothesis testing sense, nor do we know how well this particular model fits the training data. We can, however, examine its fit to the cross-validation data in Table 2, and see that it performs similarly to the CART models described earlier. Finally, we can set the number of hidden layers with the *size* subcommand, as below for a NNET with 10 hidden layers.

iq.nnet10<-nnet(IQ~age+fathered+mothered,size=10,linout=T,skip=T)

In order to change the value of the weight decay (λ) parameter to 0.5, we would use the following command in R. Remember that larger values of λ tend to shrink the size of the weights for the hidden layers. To select the optimal λ value various values would be tried and their relative impact determined through an examination of the accuracy of results for the cross-validation sample (see Table 2).

iq.nnet2.decay<-nnet(IQ~age+fathered+mothered,size=2,decay=.5,linout=T,skip=T)

In addition to manipulating the number of hidden nodes and the weight decay parameter, the researcher also has the option of changing the range of random starting values for the weights. By default, R draws these weights randomly from between -0.5 and 0.5. However, as discussed above, the range of starting values can be changed in order to reflect *a priori* beliefs regarding the importance of the hidden layers. A larger range of starting values for the weights allows for the possibility that the hidden layers are more important than if the range of starting values is tightly clustered near 0. As an example, the following R commands include 2 hidden layers, a λvalue of 0 and the range of starting values between -1 and 1.

iq.nnet2.range1<-nnet(IQ~age+fathered+mothered,size=2,linout=T,skip=T,rang=1)

```
# weights:  14
initial  value 2588587.290328
iter  10 value 15631.624853
final  value 15597.617140
converged
```

summary(iq.nnet2.range1)

```
a 3-2-1 network with 14 weights
options were - skip-layer connections  linear output units
 b->h1  i1->h1  i2->h1  i3->h1
 -0.36   0.40   0.68   0.81
 b->h2  i1->h2  i2->h2  i3->h2
 -2.38 -142.77  -6.38  -11.60
  b->o   h1->o   h2->o   i1->o   i2->o   i3->o
 47.02  46.02 -86.92  -0.01   0.12   1.69
```

We can see that the weights for the hidden layers in this model are generally larger those of the 2 hidden node with starting value range from -0.5 to 0.5 above. Bias and RMSE results for a number of NNET models with the cross-validation data appear in Table 2. It appears that the NNET model with 20 hidden layers provided the best fit to the cross-validation sample, across all of the models examined in this study.

GAM

In order to build a GAM with the thin plate smoothing spline (the default in R), we would use the following command sequence. GAM is included in the *mgcv* library of R functions that we would have previously installed in our version of R. The actual *gam* command used here sets the dimensions of the basis function, *k*, equal to 6 for both mother's and father's education. The reason for this is that the number of observed values for these variables was only 6, and there cannot be more dimensions to the basis function than there are values of the variable. The default value in R is 10. We will note the GCV score of 111.6503 and compare it with the GCV scores for alternative GAMs below.

```
library(mgcv)
iq.gam.tp<-gam(IQ~s(age, bs="tp")+s(fathered, k=6, bs="tp")
+s(mothered, k=6, bs="tp"),family=gaussian)
iq.gam.tp
```

```
Family: gaussian
Link function: identity

Formula:
IQ ~ s(age, bs = "tp") + s(fathered, k = 6, bs = "tp")
+ s(mothered,     k = 6, bs = "tp")

Estimated degrees of freedom:
1 1 1  total = 6

GCV score: 111.6503
```

In order to use an alternative smoother, such as the cubic spline, we would change the previous commands as follows, replacing *tp* with *cs*.

```
iq.gam.cs<-gam(IQ~s(age, bs="cs")+s(fathered, k=6, bs="cs")+s(mothered,
k=6, bs="cs"),family=gaussian)
iq.gam.cs
```

```
Family: gaussian
Link function: identity

Formula:
IQ ~ s(age, bs = "cs") + s(fathered, k = 6, bs = "cs") + s(mothered,
   k = 6, bs = "cs")

Estimated degrees of freedom:
4.4562e+00 3.0283e+00 6.7655e-10  total = 10.48453

GCV score: 105.3774
```

Note that the GCV score for the cubic spline GAM is somewhat lower than that of the thin plate spline model, indicating that it provides a better fit to the data.

We can change the degree of smoothing itself by setting γ=1.4, for example here with the cubic spline smoother. In this instance, the GCV actually increased from the cubic spline model with the default γ=1, suggesting that this latter model does not provide as good a fit to the data.

```
iq.gam.cs.gamma14<-gam(IQ~s(age, bs="cs")+s(fathered, k=6, bs="cs")
    +s(mothered, k=6, bs="cs"),family=gaussian,gamma=1.4)
iq.gam.cs.gamma14

        Family: gaussian
        Link function: identity

        Formula:
        IQ ~ s(age, bs = "cs") + s(fathered, k = 6, bs = "cs")
          + s(mothered,     k = 6, bs = "cs")

        Estimated degrees of freedom:
        4.8614e-03 6.6852e-03 7.6595e-06  total = 3.011554

        GCV score: 107.4748
```

```
 AIC(iq.gam.tp)
[1] 1129.699
AIC(iq.gam.cs)
[1] 1125.545
AIC(iq.gam.cs.gamma14)
[1] 1126.74
```

In addition to the GCV, it is also possible to compare the fit of GAMs using the AIC, which can be obtained with the R commands to the right. Remember that using this criterion, the optimal model is the one with the smallest AIC value, which in this case is the cubic spline model with γ =1, which was also the best fitting model based on the GCV score.

We estimated models for GAMs with cubic and thin plate splines and for both γ=1 and γ=1.4. Results for these in terms of prediction of the cross-validation sample appear in Table 2. The GAMs with a thin plate spline and γ=1 or γ=1.4 provided the lowest bias and RMSE values of the GAMs, despite the fact that based on the GCV and AIC values, the cubic splines appeared to be slightly better. In addition, the GAMs had slightly better RMSE and bias values than both OLS and the CART models, and performed comparably to the NNETs, though as noted above the NNET model with 20 hidden layers provided the best fit for the cross-validation sample. While not at all dramatic, the small discrepancy in terms of which model appears to be best fitting for the training and cross-validation samples does suggest the need for extra care with regard to the problem of overfitting when using these complex modeling techniques. In short, it appears that the cubic spline models may have overfit the training data somewhat, when compared with the thin plate splines.

Discussion

Prediction is an important aspect of statistical practice in psychology and the other social sciences, which had traditionally been done using standard MLR. For example, prior studies in the area of premorbid IQ prediction have generally been based on MLR models with adolescent and older populations. However, it has been argued that the regression based approach may not always be optimal (Veiel & Koopman, 2001), nor has it been shown that such predictions can be accurately made for preschool age children. Problems with relying too completely on strictly linear model forms are not limited to the prediction of premorbid IQ. In recent years, a number of alternative prediction modeling methods have become more widely available in popular software packages such as R. While offering the promise of greater prediction accuracy, however, these more complex models also present the researcher with a sometimes bewildering array of tuning parameters that must be set in order for them to perform optimally. Thus, a primary goal of this study was to demonstrate how one might use these modern methods of prediction in practice with a real prediction problem. The methods featured here were selected because they have been shown to be effective in both simulation and applied research, as noted above.

To briefly summarize the results of this study, it appears that in terms of the outcome variables included here, bias and RMSE all of the methods provided generally comparable predictions of FSIQ, with the NNET model with 20 hidden nodes being somewhat more accurate than the others. This approach demonstrated both the least bias and the lowest RMSE value. MLR, CART and GAM performed very similarly to one another. Given that prior Monte Carlo simulation work has shown that linear models such as MLR perform poorly for prediction when there are a number of interactions among predictor variables in the population (Garson, 1998), we may be able to infer from the relative success of MLR in this case that the relationships among these predictors and FSIQ are largely linear.

In conclusion, we hope that this manuscript contributes to research practice by demonstrating three proven and effective methods of prediction that are available in situations where it is known or believed that the relationships between predictor and outcome variables is not linear in nature. Furthermore, we have attempted to demonstrate how one can optimize these models through the judicious use of tuning parameters and/or pruning, in the case of CART. In the final analysis, the selection of optimal models and settings should be based on their accuracy with respect to a cross-validation sample. In all three cases, there is a distinct risk of overfitting the training data, which results in models that are not generalizable to the broader population. Therefore, simply assessing model fit for the training sample will likely leave the researcher with a less than optimal model for practice. However, systematically altering the tuning parameters and examining their impact on prediction for the cross-validation sample can result in selection of a model that provides the most accurate predictions possible for samples from across the population.

References

Baade, L. E., & Schoenberg, M. R. (2004). A proposed method to estimate premorbid intelligence utilizing group achievement measures from school records. *Archives of Clinical Neuropsychology*, *19*, 227-243.

Barona, A., Reynolds, C.R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52,* 885-887.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. New York: Springer.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3,* 129-136.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chang, M., Finch, W. H., & Davis. A. S. (2011, April). *The prediction of intelligence in preschool children using alternative models to regression*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Cheung, C-K, & Lee, T-Y. (2010). Improving social competence through character education. *Evaluation and Program Planning, 33,* 255-263.

Crawford, J. R., Millar, J., & Milne, A. B. (2001). Estimating premorbid IQ from demographic variables: A comparison of regression equation vs. clinical judgment. *British Journal of Clinical Psychology, 40,* 97-105.

Finch, W. H., & Holden, J. E. (2010). Prediction accuracy: A Monte Carlo comparison of several methods in the continuous variable case. *Multiple Linear Regression Viewpoints, 36*, 13-28.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. London: SAGE Publications.

Granello, D. H. (2010). Cognitive complexity among practicing counselors: How thinking changes with experience. *Journal of Counseling & Development, 88,* 92-100.

Griffin, S. L., Mindt, M. R., Rankin, E. J., Ritchie A. J., & Scott, J. G. (2002). Estimating premorbid intelligence: Comparison of traditional and contemporary methods across the intelligence continuum. *Archives of Clinical Neuropsychology, 17*, 497–507.

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

Hothorn, T., Hornick, K., Zeilleis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of computational and graphical statistics, 15,* 651-674.

Keele, L. (2008). *Semiparametric regression in the social sciences*. Hoboken, NJ: John Wiley & Sons.

Kim, Y. J. & Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B, 66,* 337-356.

Krull, K. R., Sherer, M., & Adams, R. L. (1995). A comparison of indices of premorbid intelligence in clinical populations. *Applied Neruopsychology, 2,* 35-38.

Lavigne, J. V., LeBailly, S. A., & Gouze, K. R. (2010). Predictors and correlates of completing behavioral parent training for the treatment of oppositional defiant disorder in pediatric primary care. *Behavior Therapy, 41,* 198-211.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford UniversityPress.

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Marshall, D. B., & English, D. J. (2000). Neural Network modeling of risk assessment in child protective services. *Psychological Methods*, *5,* 102-124.

Nistor, N., & Neubauer, K. (2010). From participation to dropout: Quantitative participation patterns in online university courses. *Computers & Education, 55,* 663-672.

Powell, B. D., Brossart, D. F., & Reynolds, C. R. (2003). Evaluation of the accuracy of two regression-based methods for estimating premorbid IQ. *Archives of clinical neuropsychology, 18,* 277-292.

Pungello, E. P., Iruka, I. U., Dotterer, A. M., Koonce-Mills, R., & Reznick, J. S. (2009). The effects of socioeconomic status, race, and parenting on language development in early childhood. *Developmental Psychology, 45,* 544-557.

R Development Core Team. (2007). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reynolds, C. R. (1997). Postscripts on premorbid ability estimation: Conceptual addenda and a few words on alternative and conditional approaches. *Archives of Clinical Neuropsychology, 12,* 769-778.

Roberts, E., Bornstein, M. H., Slater, A. M., & Barrett, J. (1999). Early cognitive development and parental education. *Infant and Child Development, 8*, 49-62.

Roid, G. H. (2003). *Stanford-Binet intelligence scales (5th ed.)*: Technical Manual. Itasca, IL: Riverside Publishing.

Schinka, J. A., & Vanderploeg, R. D. (2000). Estimating premorbid level of functioning. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Schoenberg, M. R., Lange R. T., Brickell T. A., & Saklofske D. H. (2007). Estimating premorbid general cognitive functioning for the American WISC-IV: Demographic and current performance approaches. *Journal of Child Neurology, 22*, 379-388.

Schoenberg M. R., Lange R. T., & Saklofske D. H. (2007a). Estimating premorbid FSIQ scores for the Canadian WISC-IV: Demographic and combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology, 29,* 867-878.

Schoenberg, M. R., Lange, R. T., & Saklofske, D. H. (2007b). A proposed method to estimate full scale intelligence quotient (FSIQ) for the Canadian Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV) using demographic and combined estimation procedures. *Journal of Clinical and Experimental Neuropsychology*, *29,* 867-878.

Schoenberg, M. R., Lange, R. T., Saklofske, D. H., & Suarez, M. (2008). Validation of the child premorbid intelligence estimate method to predict premorbid Wechsler Intelligence Scale for Children—Fourth Edition Fall Scale IQ among children with brain injury. *Psychological Assessment, 20,* 377-384.

Schoenberg, M. R., Scott, J. G., Duff, K., & Adams, R. L. (2002). Estimation of WAIS-III intelligence from combined performance and demographic variables: Development of the OPIE-3. *The Clinical Neuropsychologist, 16*, 426-438.

Schumacher, M., Robner, R. & Vach, W. (1996). Neural networks and logistic regression: Part I. *Computational Statistics and Data Analysis*, *21*, 661-682.

Sellers, A. H., Burns, W. J., & Guyrke, L. (2002). Differences in young children's IQs on the Wechsler Preschool and Primary Scale of Intelligence-Revised as a function of stratification variables. *Applied Neuropsychology, 9,* 65-73.

Sellers, A. H., Burns, W. J., & Guyrke, L. (1996). Prediction of premorbid intellectual functioning of young children using demographic information. *Applied Neuropsychology, 9,* 65-73.

Simonoff, J. S. (1996). *Smoothing methods in statistics.* New York: Springer.

Vanderploeg, R. D., Schinka, J. A., & Axelrod, B. N. (1996). Estimation of WAIS-R premorbid intelligence: current ability and demographic data used in a best-performance fashion. *Psychological Assessment, 8,* 404-411.

Vanderploeg, R. D., Schinka, J. A., Baum, K. M., Tremont, G., & Mittenberg, W. (1998). WISC-III premorbid prediction strategies: Demographic and best performance approaches. *Psychological Assessment, 10(3),* 277-284.

Veiel, H. O. F., & Koopman, R. F. (2001). The bias in regression-based indices of premorbid IQ. *Psychological Assessment, 13,* 356-368.

Wechsler, D. (1983). *Wechsler Adult Intelligence Scale-Revised (WAIS-R)*. Technical Manual. New York: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wood, S.N. (2006). *Generalized Additive Models: An introduction with R.* Boca Raton, FL: Chapman & Hall.

Yeates, K. O., & Taylor, H. G. (1997). Predicting premorbid neuropsychological functioning following pediatric traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 19,* 825-.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu