

Elizabeth A. Parillo

Department of Mathematical Sciences
Montana State University

December 16th, 2010

A writing project submitted in partial fulfillment
of the requirements of the degree

Master of Science in Ecological and Environmental Statistics

APPROVAL

of a writing project
submitted by Elizabeth Parillo

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographical style, and consistency, and is ready for submission to the Statistics faculty.

12/16/10
Date


Megan D. Higgs
Writing Project Advisor

Dec 16, 2010
Date

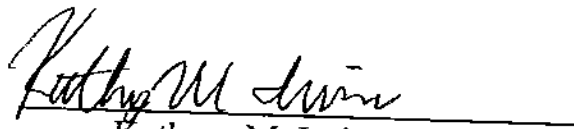

Kathryn M. Irvine
Writing Projects Coordinator

Table of Contents

| | |
|---|----|
| 1. Introduction | 4 |
| 2. Classification and Regression Tree (CART) Analysis Background | 6 |
| 2.1 Splitting Rules | 7 |
| 2.1.1 Classification Trees | 7 |
| 2.1.2 Regression Trees | 8 |
| 2.1.3 Tree Construction Summary | 9 |
| 3. Tree Popularity | 9 |
| 4. Examples | 11 |
| 4.1. Logistic Regression versus Classification Tree | 11 |
| 4.1.1 Models using age as only explanatory variable | 12 |
| 4.1.2 Models using age and gender as explanatory variables | 15 |
| 4.1.3 Interaction Models | 17 |
| 4.2 Multiple Linear Regression versus Regression Tree | 20 |
| 4.1.1. Use of the original response versus a log- transformed response..... | 20 |
| 4.1.2. Justifiable statistical inference | 24 |
| 5. Conclusion | 25 |
| 6. References | 27 |

1. Introduction

Leo Brieman published his book entitled *Classification and Regression Tree Analysis* in 1984. This book came about due to the advent of computer technology necessary to implement the tree methodology. "The use of trees was unthinkable before computers," says Brieman. This contrasts with previous statistical methods, which were first developed on paper, and moved over to computers as data requirements dictated. Brieman developed the CART © software, which he describes as a "flexible non-parametric tool [to add] to the data analyst's arsenal."

Classification and regression tree methods are increasingly seen in literature across many disciplines. For example, they have been found to be very useful in remote sensing land cover classification procedures, as well as other ecological and environmental applications. Medical studies of patient mortality, or indicators of presence or absence of disease, commonly employ tree-based analysis methods. CART is used in these applications to sort through large datasets with numerous variables and complex relationships in hopes of revealing the relationships between variables, highlighting important variables, delineating the data set into smaller homogeneous groups, and to accurately predict the response of a new subject.

Due to increased usage and praise of these models in various scientific communities, more insight is needed regarding how CART analysis fits into a statistician's tool arsenal. Further development of an appropriate statistical context of these methods is necessary to reconcile the researcher's use of these models. Their appropriate use, as with any analysis method, depends on the question of interest at hand, and the subsequent goal of a study. In this paper, we provide examples to help elucidate differences between Classification trees, logistic regression, and between regression trees and ordinary regression.

Sometimes the focus of statistical analysis is on trying to understand the relationship between large numbers of explanatory variables in their association with a certain phenomena. To accomplish this goal, traditional statistical tools are used to estimate model parameters, and calculate measures of uncertainty to quantify the degree of error in our estimates. Practically, standard errors and subsequent confidence intervals give researchers a way to quantify a range of plausible values for an estimated parameter.

On the other hand, sometimes the main goal of statistical analysis, rather than explanation-based parameter estimates, is accurate prediction of new observations. When the goal of a model is prediction, the primary focus is constructing a model that accurately classifies or predicts a new observation based on explanatory variable values. There are many ways to quantify predictive model accuracy, such as the common k-fold cross validation, or use of an independent validation data set. The most common form of k-fold cross validation is to leave one observation out ($k=1$), construct the predictive model, and then predict the observation that was left out. This process is repeated for all the observations. These methods quantify the misprediction rate of the model, or the mean prediction error. In this case, there is less of a desire to quantify uncertainty in the form of standard errors for estimates, as the focus is on the predictive accuracy of a model. Therefore, there is an important distinction between assessing accuracy and uncertainty in predictions and assessing accuracy and uncertainty in the estimation of means and probabilities.

Classification or Regression trees are useful when a predictive model is the goal. In other words, Classification and Regression trees often have high predictive accuracy in predicting class memberships or mean values for data. We can quantify this predictive accuracy using misprediction rates, as previously described. However, users should be aware that Classification

and Regression tree analyses do not quantify uncertainty in the estimated probabilities of class membership, and in the estimated means for each terminal node in a regression tree. There is a disconnect between the ability to quantify uncertainty in predictions made from trees, and the ability to quantify uncertainty in the tree itself. Other methods, such as linear or logistic regression, can be utilized as predictive models and we can also describe both how uncertain we are in the model itself.

This paper explains some of the background of Classification and Regression Tree analysis as well as describes the basics of tree construction. Some current applications of CART analysis and reasons for their current popularity will also be presented. Additionally, In order to examine some of the advantages and disadvantages of CART, two examples compare CART to both traditional logistic regression and multiple linear regression, followed by a discussion of the important differences.

2. Classification and Regression Tree Analysis Background

Whether a tree is a classification or a regression tree depends only on the type of response variable. If the response variable is quantitative, then the tree is termed a regression tree, and if the response variable is categorical, then a classification tree is grown.

CART uses classification rules to grow both types of trees. A classification rule is a systematic way of predicting what class, or homogeneous group within the data set, a specific observation belongs to (Brieman 1984). Classification and Regression trees are schematic tools to help one decipher how response variable values (categorical or quantitative) are grouped based on values of explanatory variables.

CART uses binary recursive partitioning to divide the data set into sequentially more homogenous subsets, or classes (Friedl and Brodley 1997). The binary recursive partitioning process is dictated by an “impurity” function that is based on a single explanatory variable value at a single split. This function acts as a splitting rule, which partitions the response variable into homogeneous groups. The data are a “learning sample” which “trains” the classifier in how to classify new response values. There are many different functions by which to split the data, and there are various “stopping rules”, which prevent the tree from growing, or splitting out more classes. These become the main tasks of classification and regression trees: knowing how to use the learning sample to determine the splits, and knowing when to stop growing a tree (Brieman 1984).

2.1. Split Rules

As stated, there are many different functions by which to split the data. Different statistical computing packages make these decisions automatically depending on different default values. These default values vary across packages, and therefore, a user of these models must be aware of what rules the tree is utilizing. In the statistical computing package, R, the default split rule is the gini index.

2.1.1. Classification Trees

There are three common splitting rules for growing classification trees (De'ath and Fabricus 2000): the entropy index, the gini index and the twoing index. Each one of the following rules split the data with a different approach. The entropy index (otherwise known as the information rule) is defined as

$$Entropy(t) = -\sum_i p_i \log p_i$$

Where p_i is the relative frequency of class i at node t . This rule works to minimize the heterogeneity within a group by identifying those splits which divide out as many groups as possible as precisely as possible. This rule has a higher predictive accuracy than the other rules when identifying rare classes (Zambon et al. 2006).

The gini index is defined as

$$Gini(t) = \sum_i p_i(1 - p_i)$$

This rule splits the data by first finding the largest homogeneous category within the data and separating it out from the rest of the dataset (Zambon et al. 2006).

The twoing index is

$$Twoing(t) = \frac{P_L P_R}{4} \left(\sum_i (|p(i|t_L) - p(i|t_R)|)^2 \right)$$

where $p(i|t)$ is the relative frequency of class i at node t and L and R denote the left and right sides of a given split respectively. Twoing tends to split the data up evenly in that it separates groups of the data by identifying groups that make up 50% of the remaining data at each node.

2.1.2. Regression Trees

If the dataset has a quantitative response variable, then a regression tree is grown. The most common impurity rules for regression trees include: the sums of squares about the group means, and the sums of absolute deviations from the group medians. Both of these rules split the data based on the largest reduction in the sum of squares about a group mean or median, akin to least squares linear models (De'ath and Fabricus 2000). The default splitting rule in the statistical computing package R is based on the group means. Each explanatory variable in the data set is

assessed for each split and the one explaining the most amount of variation in the response is chosen for that split (Crawly 2007).

2.1.3. Tree Construction Summary

To summarize the steps in growing a tree, all of the data, or the learning sample, start in a root node. The tree searches through all the explanatory variables available in the data set and picks the variable that best splits the data into the two most homogenous groups. This “best” split is one that minimizes the impurity in the data set, or maximizes the homogeneity (definition of “best split” dependent on the splitting rule applied). The tree does this until all the terminal nodes are either pure (have the same response variable value) or there are five or fewer observations in the node (the default “stopping rule” in the statistical computing package R). Additionally, single classification trees require “pruning” as they are susceptible to over-fitting the data. This means that the prediction error associated with a certain tree size can actually increase as the tree size increases. In other words, the tree model has an optimal size. Trees also require pruning because sometimes, as the size of the tree increases, there is no substantial additional amount of deviance explained. This means that, beyond a certain size, the classification produced by the tree will not improve (Brieman 1984).

3. Tree Popularity

Tree-based models have been increasingly popular in many sciences. The literature describes many advantages of tree-based models. For example, Zambon et al. (2006) describes decision trees as non-parametric methods independent from data distributional assumptions, and are useful for examining large, complex data sets. Crawly (2007) describes tree-based models are

a simple data exploration tool, producing a clear picture of the data structure, as well as providing an “intuitive insight into...the interactions between variables”. Because ecological and environmental data are often very complex, with many explanatory variables, tree-based models have become popular in this area. Scientists need a methodology that is flexible and robust in handling a wide range of data types, while also be easily interpretable, and has the capacity to handle missing data values (De'ath and Fabricius 2000).

Classification trees in Remote Sensing are often used to produce accurate land cover classifications needed for environmental monitoring practices (Chen and Paelinckx 2008). Other environmental monitoring applications include delineating wetland or riparian zones (Baker et al. 2006), documenting certain agricultural practices (Yang et al. 2002), or mapping invasive plant species (Lawrence et al. 2005). The search for the most capable classification tool in terms of its predictive accuracy has led remote sensing scientists through many different classification procedures.

Another useful tool commonly used for image classification is binomial or multinomial Logistic regression (Panatelson et al. 2009). In the case of remote sensing data, a binary response could be an indicator of land cover class (e.g. water versus not water). The explanatory variables for the model are the spectral responses, or scaled radiance values (i.e. digital numbers), in each band of remote sensing data. Logistic regression can be used to model the probability of a specific pixel belonging to a certain land cover class given a specific set of spectral characteristics. Logistic regression also provides an estimate of the uncertainty associated with that probability. When a probability cut-off rule is applied to the estimated probability, logistic regression becomes a classification tool.

Classification trees are also used to classify imagery (Lawrence and Wright 2001). Many studies have found that classification tree analysis has yielded higher predictive accuracies when compared to traditional classification methods such as minimum distance to means or parallelepiped (Lawrence et al. 2004). One can use classification tree analysis to estimate the probability of a pixel belonging to a certain land cover class by using the relative proportions of that class in that node (Breiman 1984). However, when one does this there is no estimate of the uncertainty assigned to that probability as there is in logistic regression. This fact is one of the main catalysts for this discussion.

4. Examples

The following two relatively simple examples illustrate the differences in inference between classification trees and logistic regression, and between regression trees and multiple linear regression.

4.1. Logistic Regression versus Classification Tree

The Donner Party was a pioneer group that attempted a route from Wyoming to the Sacramento Valley in 1846. The party was stranded in the eastern Sierra Nevada Mountains in a late October snow, and when the last survivor was rescued in April of 1847, 40 of the 87-member party had perished. The Donner Party data set consists of the party members sex, age and whether or not they survived (Ramsey and Schafer 2002).

In this case, the response is survival (0 if the party member died, 1 if they lived). Logistic regression is a statistical modeling tool appropriate for data in which the response variable is binary. From this model, one could estimate the probability of surviving given a certain age or

gender, the odds of survival for females relative to males for a given age, or the odds of survival for females (or males) at different ages. These estimates could be used to study whether there is evidence that the females were more capable of withstanding harsh conditions than the males were. Additionally, confidence intervals for parameters of interest can be a way to describe a set of “plausible values”. This is a way to quantify the degree of uncertainty in our estimates, which is traditionally one of the main goals of statistics. The mathematical details of logistic regression are beyond the scope of this paper, but for a detailed discussion of logistic regression see page 70 of Alan Agresti’s *An Introduction to Categorical Data Analysis*, 2007.

A classification tree could also be fit to these data since the response is categorical (lived vs. died). This tree is used to determine what values of age and/or gender best predict whether or not a person lived or died. The branches of the tree are rules that decide which survival category a person belongs to based on their age and sex. It is possible to obtain predicted probabilities from a tree model through a Class Probability Tree. A Class Probability Tree is very similar to Classification Tree, but instead of only getting the predicted group assignment for a given observation, we can also get the probability of belonging to that class. The probabilities of belonging to category j given that the observation is in terminal node t , are the relative proportions of class j cases in that node t (Brieman 1984). For this example, all the tree models were grown using the gini splitting rule.

4.1.1. Models using age as the only explanatory variable

We first compare a logistic regression model including only age to a tree based only on age. The estimated logistic regression equation is

$$\text{logit}(\hat{\pi}_i) = 1.81152 - 0.06647\text{age}_i$$

From this logistic regression equation, we can calculate the probability of survival for a given age, or the odds of survival of a specific age relative to another age. We would also be able to give a range of plausible values for the true probability of survival based on the age of an individual.

Figure 1 displays the resulting classification tree, which splits the data into two homogenous groups: those who survived and those who died. If an individual has an age greater than or equal to 22.5 years, then they would be classified as survived. If they are less than this age, they would be classified as dead. The numbers at the split and at each terminal node represent the number of people that truly lived or died. For example, in the survival node (which is labeled as 1), there are six individuals who actually survived, and one who died. Therefore, there are seven observations in that node total. Similarly, in the "died" node, there are 14 actual survivors, and 24 actual deaths, for a total of 38 observations in the node. The predicted probabilities of survival from CART are calculated from the full tree, not the pruned tree. This means the predicted probabilities from CART are the relative proportions of class j cases in node t in each terminal node constructed, not necessarily shown. It should be noted that this is not in the software documentation, nor explicitly stated in the 1984 Brieman text.

Figure 2 compares the predicted probabilities of survival from the logistic regression model to the predicted probabilities of survival from CART using only age. In order to classify individuals into survival categories based on their predicted probability of survival from logistic regression, a cutoff probability of 0.5 was used. The red dots indicate individuals of the Donner Party who actually survived, and the blue dots indicate those that actually died. Because there are only four distinct values for the probabilities from CART, these probabilities were "jittered" along the y-axis. This means the response values were scattered a small random amount in order

to display points that were overlaid on top of one another. The misprediction rate for the logistic regression model was 42%. The misprediction rate from CART was only 26%. Therefore, the predictive accuracy of the CART model is higher than the logistic regression method for this example. However, inferences from the logistic regression model may be more useful to the researcher's question of interest as we could make statements about the odds or probability of survival for an individual with a measure of uncertainty associated with this estimate.

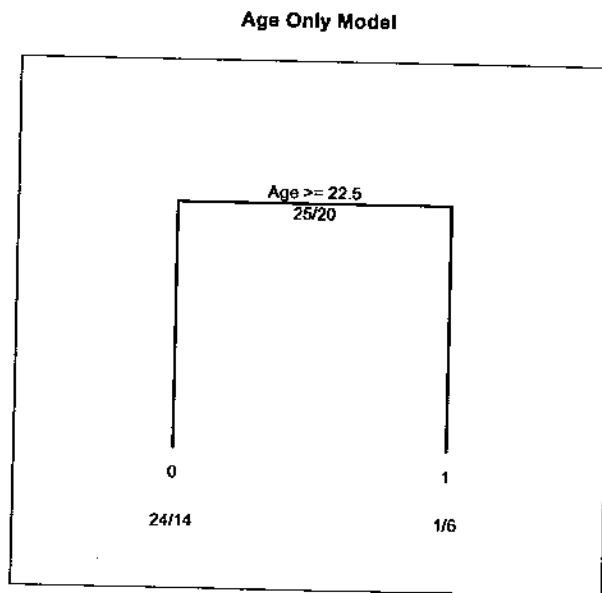


Figure 1: This is the survival classification tree from R using age as the only explanatory variable.

4.1.2. Models using age and gender as explanatory variables

The second pair of models included both a party member's gender and age as explanatory variables. *Female* is an indicator variable for a party member's gender such that if an observation is from a female it equals 1, and 0 if from a male. The estimated logistic regression equation is

$$\text{logit}(\hat{\pi}_i) = 1.63312 + 1.59729 \text{female}_i + 0.0782 \text{age}_i$$

If one were using logistic regression to model the probability of survival for the Donner party members, there is evidence that gender is an important variable in explaining survival after accounting for age ($t = 2.114$, two-sided p -value = 0.0345). We could make inferences about the odds of female to male survival given a certain age, while also providing a range of plausible values for the parameter. For example, the odds of survival for females are estimated to be 4.94 times that of males of the same age. A 95% confidence interval ranges from 1.12 to 21.54 times.

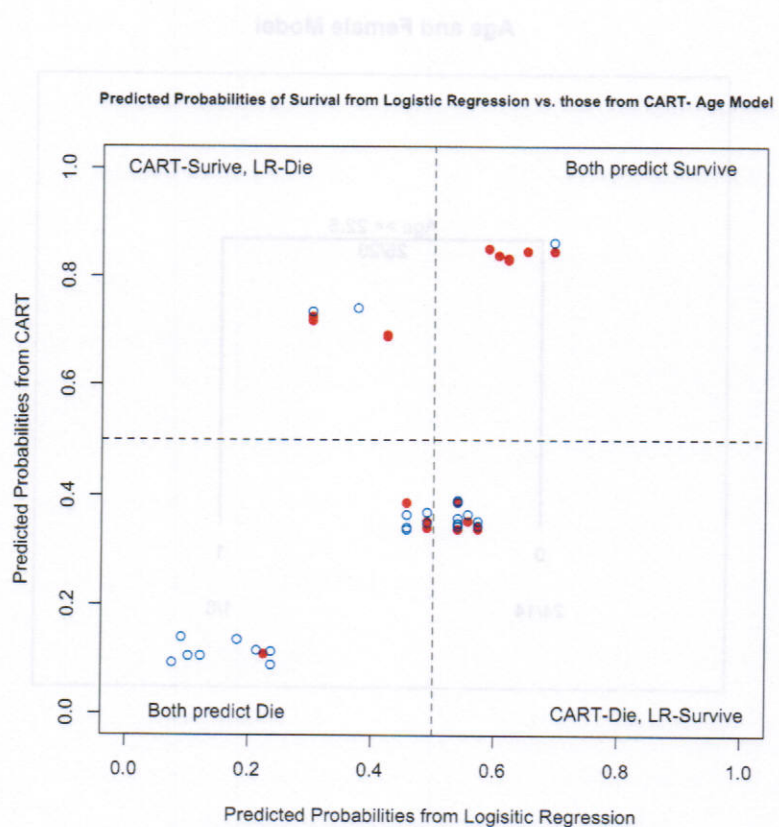


Figure 2: Predicted Probabilities from Logistic Regression vs. those from CART using age as the only explanatory variable. The red dots indicate actually survival, while blue indicates actual death.

The classification tree constructed with age and gender as explanatory variables appears identical to the previous tree. The tree does not use gender as an important variable in predicting the probability of survival and therefore, one would not suspect the relationship of gender with

survival, unless one examined the whole tree grown. The predicted probabilities of survival change in comparison to the previous model, but the classification rule to assign party members to groups did not change with the additional explanatory variable. This is an interesting point that users should be aware of. We can see from the logistic regression model that there is evidence that gender is associated with the odds or probability of survival, but one does not see the relationship of gender in the pruned the classification tree shown in Figure 3.

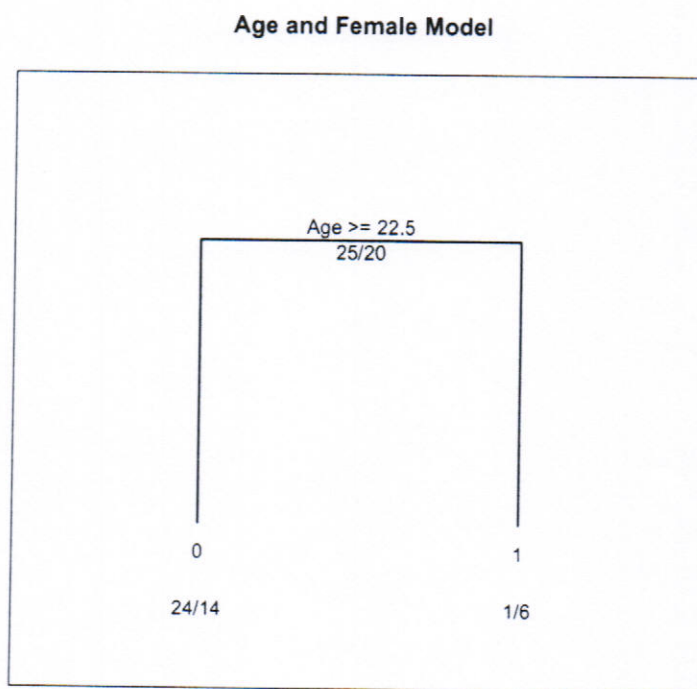


Figure 3: This is the classification tree constructed in R using gender and age as explanatory variables.

Figure 4 is a plot of CART probabilities versus logistic regression probabilities when both age and gender are included. When compared to Figure 2, it is clear that the CART probabilities from this model differ from the age-only model probabilities of survival. Again, we would not have recognized how the probabilities of survival changed when included female in

the model if we were to only examine the pruned tree. The misprediction rates of the Logistic Regression model including age and gender is equal to the misprediction rate of the corresponding CART model at 22%.

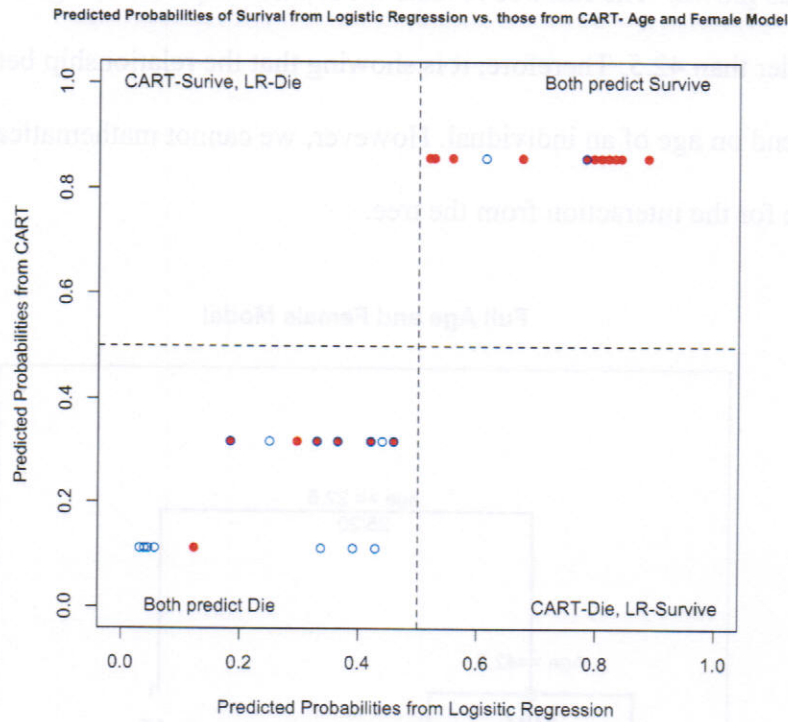


Figure 4: These are the predicted probabilities from the logistic regression model versus those from CART using age and gender as explanatory variables. Red indicates actual survival while blue indicates actual death.

4.1.3. Interaction Models

Lastly, an interaction model was fit, which would allow the relationship between gender and survival to be dependent on age. The estimated logistic regression equation is

$$\text{logit}(\hat{\pi}_i) = 0.31834 - 0.03248age_i + 6.92805Female_i - 0.1616age_i * Female_i$$

There is suggestive evidence that the relationship between gender and probability of survival does vary across ages as indicated by evidence for the interaction term in the model (t-stat = -1.74, two sided p-value = 0.0865).

Figure 5 is a schematic of the full grown tree using the age and gender explanatory variables. Since the user of these models cannot include an interaction term in the model construction, in order to view potential interactions between age and gender, the full tree using age and gender was grown. The full tree reveals the relationship between gender and survival for individuals older than 42.5. Therefore, it is showing that the relationship between gender and survival may depend on age of an individual. However, we cannot mathematically describe or obtain an estimate for the interaction from the tree.

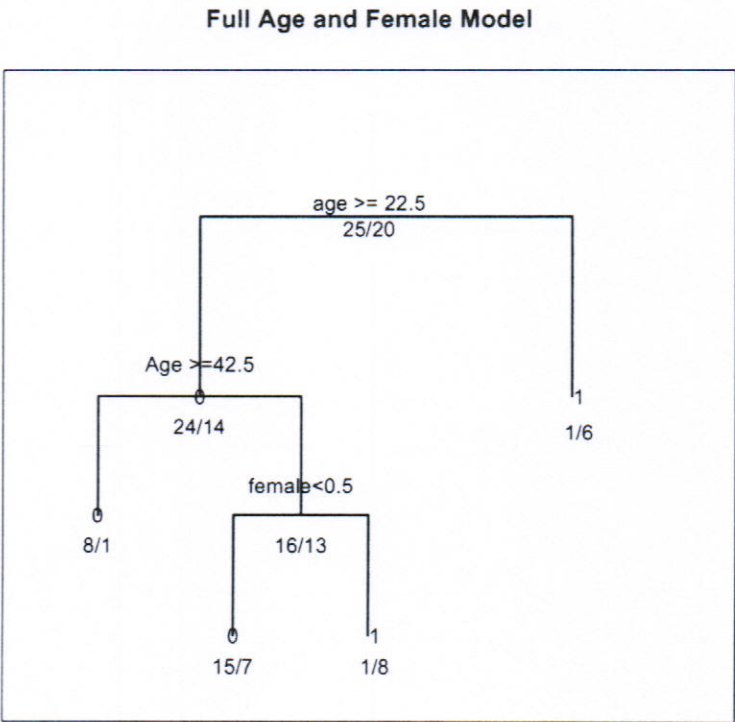


Figure 5: This is the full classification tree for the model constructed using age and gender as explanatory variables.

Again, the predicted probabilities from the CART versus the predicted probabilities from the logistic regression interaction model are plotted in Figure 6.

Predicted Probabilities of Survival from Logistic Regression vs. those from CART- Interaction Model

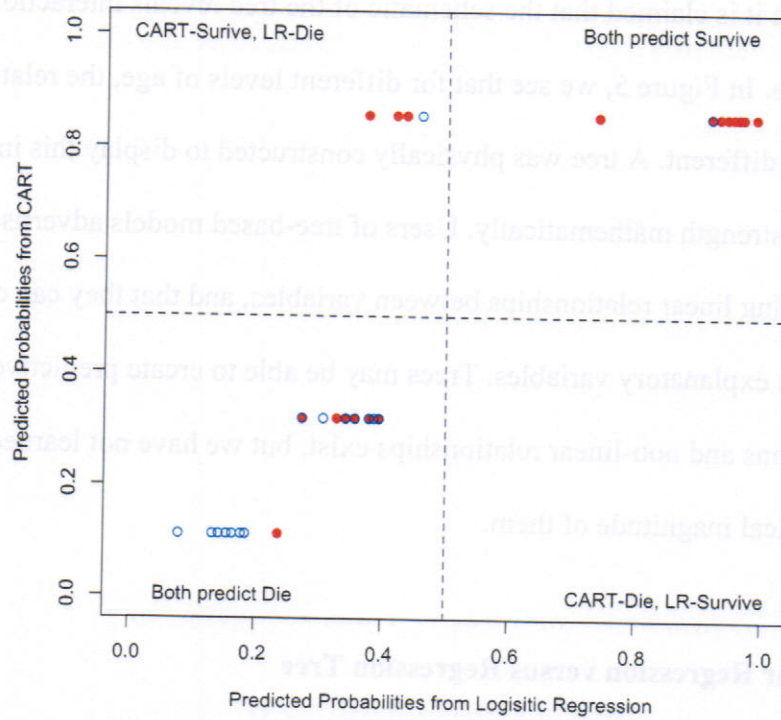


Figure 6: Predicted probabilities from the logistic regression interaction model versus those from CART using only age and gender as explanatory variables. The red indicates actually survival while blue indicates actual death.

The interaction logistic regression model predicted three individuals to die who actually survived, while these three individuals had been predicted to survive by the no-interaction model. This caused the misprediction rate to increase to 26%, which is a slight increase from the age and female only model. While the predictive accuracy may have gone down in the interaction model, estimating the interaction effect is a useful part of logistic regression. With logistic regression, we can mathematically estimate the interaction while also calculating a range of plausible values for the true change in the relationship of gender for varying ages. In other words, we can quantify our uncertainty in the estimate of the interaction effect. We cannot however, estimate the interaction effect from the tree model, nor can we describe how uncertain

we are in the interaction. Tree models do not allow a programmer to include an interaction term in the tree-model, but it is claimed that the schematic of the tree reveals interactions between explanatory variables. In Figure 5, we see that for different levels of age, the relationship of gender to survival is different. A tree was physically constructed to display this interaction, but we do not know the strength mathematically. Users of tree-based models advertise that trees are not limited to modeling linear relationships between variables, and that they can deal with interactions between explanatory variables. Trees may be able to create predictive models for data where interactions and non-linear relationships exist, but we have not learned anything about the mathematical magnitude of them.

4.2. Multiple Linear Regression versus Regression Tree

The previous example focused on the comparison of classification trees to logistic regression as a classification tool. The next example shows a comparison of a regression tree and classical multiple linear regression analysis.

4.2.1. The use of the original versus a transformed response

Brain size is an interesting variable with regard to how it may be related to evolution. One might expect that larger brained animals would be better suited to survive evolutionarily. Yet, there may be some downsides related to having a larger brain, such as the need for longer pregnancies, and/or a smaller number of offspring. The dataset used in the following example consists of average brain weight (*Brain.size*) values, body weights (*Body.weight*), gestation lengths (*Gest.time*), and litter (*Litter*) sizes for 96 mammal species (Ramsey and Schafer, 2002). As brain size and body weight are related, multiple linear regression could be used here to

determine which variables, if any, are associated with brain size after accounting for body weight. In order to model these data using multiple linear regression, the response variable (brain weight) was transformed using the natural log function because the use of the raw data violated the necessary homogeneity of variance and normality assumptions of the residuals for linear regression. For further discussion of the use of multiple linear regression see *The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, page 42.

The estimated regression equation is

$$\hat{\mu}\{\log(\text{brain.size})\} = 2.13 + 0.015\text{Gest.time} - 0.005\text{Body.weight} - 0.2149\text{Litter}$$

A regression tree was also fit to the same set of data to determine how it would be split the data into homogenous brain size groups based on explanatory variables. The variables chosen by the tree are often considered by the researchers to be the most important variables in explaining the response, which in this case, is brain weight. Because trees are advertised as not having data distributional assumptions, the first tree was constructed on the untransformed data set, as often users do not transform the response variable when using trees. The resulting tree is shown in Figure 7.

One might interpret Figure 7 to mean that gestation time is the most important variable in predicting brain size group. It split the data such that a group of 7 observations have a mean brain size of 1169 units, while 89 observations have a brain size of 144.3 units. A regression tree is based on splitting the data for the largest reduction in the sum of squares about group means or medians, akin to least squares linear models (De'ath and Fabricus 2000). Therefore, in the context of this data set, one needs to consider the variances of brain weights. There are animals in this data set that range from a field mouse to an elephant and this is the reason why we transform the response variable for the multiple linear regression model. If one of the split rules

for regression trees involves minimizing deviations from a group mean, and we are growing a tree in the context of this data set, we must consider that the squared deviations from a group mean are likely to be very large and very variable. Means are not robust measures of center in the presence of significant skewness. In other words, means are heavily influenced by extremely large or small values, which is the case for our dataset. Therefore, it is unclear how to proceed with regression tree analysis in a situation with such extreme data values. Perhaps the split rule involving medians would be a better rule to choose for this data set, but this was not investigated in this paper. Instead, to further compare regression trees and multiple linear regression with this data set, the logged response variable was used to construct another regression tree.

Regression Tree with Regular Response

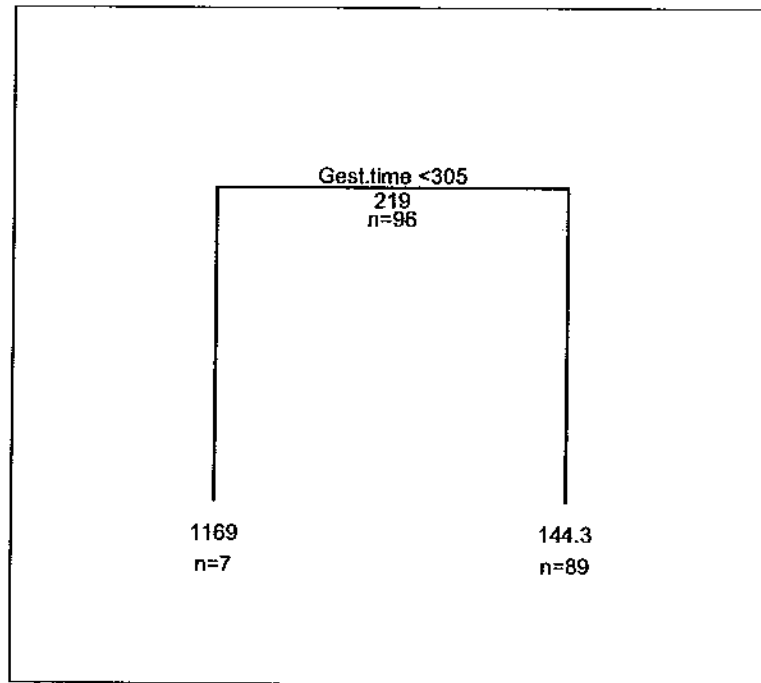


Figure 7: This is the regression tree for brain size using body weight, gestation length, and litter size as explanatory variables.

Regression tree using the transformed response

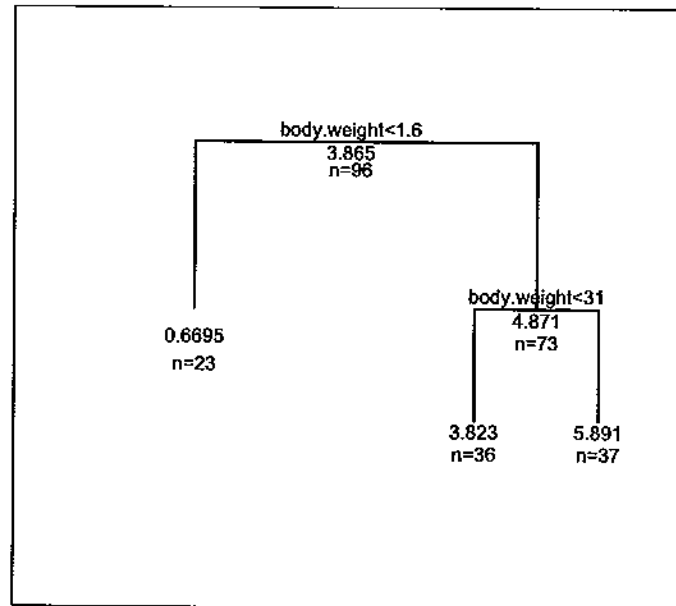


Figure 8: This is the regression tree using the log transformed brain size and all three explanatory variables

This tree chose body weight as the “first split”, or best explanatory variable to predict brain size. However, the question of interest was to assess how gestation time and litter size explain brain size after accounting for body size, as body and brain sizes are likely to be already related. While we could have concluded from the output of the multiple linear regression that gestation time and litter size are associated with brain size even after accounting for body weight, one is not capable of assessing this question of interest from a regression tree because it does not bring gestation time and litter size in as splitting variables. Therefore, it is unclear how to interpret the tree when there are several collinear explanatory variables in an observational study setting.

It is possible that if the main goal was simply to predict brain size for some other mammal based on the body weight, this regression tree could be an accurate and functional tool. However, It remains unclear as to what useful conclusions can be drawn from this regression tree. As it stands, it says that different body weights explain these relatively homogenous groupings of logged brain size data. This is not as useful as a multiple linear regression model on a logged response, from which you can make estimates of gestation time and litter size, which include measure of uncertainty of those estimates. Users of these models should take care when trying to attach importance to variables based on the early splits in the tree. Multicollinearity still must be considered in the context of interpretation.

4.2.2. Justifiable Statistical Inference

As a reminder, conventional statistical inference dictates that inferences to a larger population can only be made when random sampling is employed, and that cause and effect statements between variables can only be made under the condition of random assignment to groups. Therefore, the tree-based models, as with any model, do not allow for population inference or causal statements when these two conditions are not met. It is surprising therefore, to consistently read inference statements that violate the conventional statistical rules of inference. For example:

- "...the aim is to understand **causes** of variation in the [response]..." (Crawley 2007)
- "...[explanatory variable] is a key **determinant** in the [response]..." (Crawley 2007)

- "...the strongest effects were **due to** [explanatory variable]..." (De'ath and Fabricius 2000)
- "...the [explanatory variables] are possibly major **determinants** of [response variable]..." (De'ath and Fabricius 2000).

Causal inference is not justified as the design did not involve random assignment.

Unjustifiable statistical inferences are clearly not restricted to tree-based models, but one must apply the same critical thinking to these models as with all other models in regards to justifiable statistical conclusions.

5. Conclusion

Classification trees are claimed advantageous because they can handle large data sets with numerous variables and complex structures by outputting an easily understood schematic representation of that complex structure. Classification and Regression trees can be useful schematic tools to decipher how response variables are grouped based on values of explanatory variables. However, a user of CART must be aware that different statistical packages use various splitting rules in tree construction, and as a result, different trees can be constructed using the same set of data. CART can also be a useful classification or prediction tool as it often has higher predictive accuracy when compared to logistic regression as a classification tool. However, CART analysis does lack parameter inferences and the ability to quantify the degree of uncertainty in our estimates. Additionally, trees allow users to visually see interactions between explanatory variables, but we cannot mathematically quantify this interaction relationship using CART the way we have using traditional regression estimation techniques. It also remains unclear as to how to interpret regression trees in the presence of collinear explanatory variables,

or what useful conclusions can be drawn from regression trees when the goal is not group mean prediction.

As a statistical scholar, one is educated to think critically about the methods applied for data analysis, the associated assumptions, and what the allowable inference of a certain result is. Classification and Regression tree analysis can be a useful data exploratory tool, or an accurate prediction tool. Therefore, if data exploration or new observation prediction is the research goal, then CART may be an appropriate analysis choice. However, when the question of interest involves interpretation of variable relationships, CART lacks the ability to quantify uncertainty of the estimates of those relationships, and the ability to mathematically quantify interactions between explanatory variables. If Classification and Regression trees are to be used as a statistical model, then they are not exempt from the model considerations traditionally necessitated. As usual, the analyst needs to consider the question of interest, and what methods of analysis may be most appropriate for the task at hand.

6. References

- Baker, C., R. Lawrence, C. Montagne, D. Patten. 2006. Mapping wetlands and riparian areas using Landsat ETM+ imager and decision-tree-based models. *The Society of Wetland Scientists*. 26: 465-474.
- Brickleyer, R.S., R.L. Lawrence, P.R. Miller. 2002. Documenting no-till and conventional till practices using Landsat ETM+ imagery and logistic regression. *Journal of Soil and Water Conservation*. 57:267-271.
- Brieman, L., J.H. Friedman, R.A. Olshen, C.J. Stone. 1984. *Classification and Regression Trees*.
- Chan, J.C., D. Paelinckx. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*. 112: 2999-3011.
- Crawley, M. 2007. Tree Models. *The R Book*. 685-699.
- De'ath, G., K.E. Fabricius. 2000. Classification and Regression Trees: A Powerful yet simple technique for Ecological Data Analysis. *Ecology*. 81:3178-3192.
- Friedl, M.A., C.E. Brodley. 1997. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*. 61:399-409.
- Lawrence, R. L., A. Bunn, S. Powell, M. Zambon. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*. 90:331-336.
- Lawrence, R.L., A. Wright. 2001. Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering & Remote Sensing*. 67:1137-1142.
- Lawrence, R.L., S. Wood, R. Sheley. 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). *Remote Sensing of Environment*. 100:356-362.
- Ramsey, F.L., D.W. Schafer. 2002. *The Statistical Sleuth*. Chapters 8 and 20.
- Pantaleoni, E. R.H. Wynne, J.M. Galbraith, J.B. Campbell. 2009. Mapping Wetlands using ASTER data: a comparison between classification trees and logistic regression. *International Journal of Remote Sensing*. 30:2423-3440.
- Yang, C.C., S.O. Prasher, P. Enright, C. Madramootoo, M. Burgess, P.K. Goel, I. Callum. 2003. Application of decision tree technology for image classification using remote sensing data. *Agricultural Systems*. 76:1101-1117.
- Zambon, M., R.L. Lawrence, A. Bunn, and S. Powell. 2006. Effect of alternative splitting rules on image processing using classification tree analysis. *Photogrammetric Engineering & Remote Sensing*. 72:25-30.