IDENTIFYING OUTLYING OBSERVATIONS IN REGRESSION TREES

NICHOLAS GRANERED AND SAMANTHA C. BATES PRINS

ABSTRACT. Regression trees are an alternative to classical linear regression models that seek to fit a piecewise linear model to data. The structure of regression trees makes them well-suited to the modeling of data containing outliers. We propose an algorithm that takes advantage of this feature in order to automatically detect outliers. This new algorithm performs well on the four test datasets [7] that are considered to be necessary for a valid outlier detection algorithm in a linear regression context, even though regression trees lack the global linearity assumption. We also show the practical use of this approach in detecting outliers in an ecological dataset collected in the Shenandoah Valley.

1. INTRODUCTION

While there are numerous well-understood outlier detection methods for use with classical linear regression models, e.g. [10, 12], fewer methods are available for regression trees [1, 2]. Regression trees are an alternative to classical linear regression that model the response-predictor relationship using a piecewise linear model. This is accomplished by making a series of binary splits in one or more of the predictors, such that these splits minimize the overall residual sum of squares (RSS) of the tree. Each binary split is based on a single predictor, e.g. $X_1 \leq 3.2$ and $X_1 > 3.2$, but subsequent binary splits may involve the same or a series of different predictors. The whole-tree RSS is then defined as the sum over the residual sum of squares in each of the T terminal nodes in the tree, with terminal nodes being defined as nodes that are not subdivided by subsequent binary splits. Thus, the whole-tree RSS is

$$\text{RSS} = \sum_{t=1}^{T} \text{RSS}_t = \sum_{t=1}^{T} \sum_{j=1}^{n_t} (y_j - \overline{y}_t)^2$$

where n_t is the number of observations in terminal node t such that $\sum_{t=1}^{T} n_t = N$ and N is the total number of observations used to build the tree, y_j is

the value of the response for the jth observation placed into terminal node t, and $\overline{y}_t = \sum_{j=1}^{n_t} y_j/n_t$ is the predicted value of the response at terminal node t.

Binary splits are added to the tree until either no splits reduce the whole-tree RSS enough to counter the added complexity of adding another split, or until it is no longer possible to divide any existing node into two sub-nodes that both satisfy $n_t \geq c$ where c is the minimum number of observations required in a terminal node. In the R [9] package TREE [11] used in this paper, the default value for c is 5. Using the tree for prediction of a new observation is accomplished by evaluating the decision rules at each binary split until a terminal node t is reached. The predicted value for the new observation is then the average response, $\overline{y}_t$, observed at that node.

Regression trees have several advantages over classical linear regression models: they are easy to interpret, make no assumption of normality of errors, do not assume a global linear relationship between the response and a predictor, do not have to discard observations with missing values, and their structure is not affected by monotonic transformations of the predictors. The structure of regression trees is also advantageous in its modeling of outliers since observations with unusual response values can be isolated in a small terminal node and these unusual observations are not used in the prediction of observations that do not fall into the same terminal node. Indeed a relatively large RSS_t for a terminal node can indicate the presence of one or more outliers in that node.

It is for these reasons that regression trees have been used to model response-predictor relationships in ecology [4, 5]. Zirkle [14] used data collected by the Friends of the Shenandoah River (FOSR) to build tree models for predicting metrics of water quality and based on these models, identified streams with atypical water quality. Such atypical streams may indicate the presence of karst geology. In Zirkle's approach, streams isolated in small nodes with large RSS_t whose removal from the dataset resulted in a structurally different tree and/or a tree with a substantially different R^2 value were identified as outliers. However, this approach ultimately focused on building the best tree for prediction with outliers being streams that changed the quality or value of the prediction. Our approach focuses more closely on detecting outliers rather than predicting the non-outlying observations well.

It is important to address two issues which affect all outlier detection methods. The first of these is masking, which occurs when the presence of one group of observations that are not outliers masks the presence of another group of observations which are truly outlying. Swamping is the

opposite problem, in which non-outlying observations are made to look like outliers due to the presence of another group. These problems typically arise in "forward-stepping" algorithms, which seek to identify outliers one at a time from the complete data set. "Backwards-stepping" methods identify a subset of the data as being the most likely to be outlying, and test to see if the entire subset exhibits significant outlier behavior. As such, backwards-stepping methods are less susceptible to masking and swamping.

Hadi and Simonoff [7] proposed two approaches for detecting outliers in classical linear regression, each of which involves two stages. We extend their second method, which uses single-linkage clustering and backwards-stepping for use in regression trees. Our approach works well on all four examples that Hadi and Simonoff [7] state as necessary in order to have a viable outlier detection method despite the fact that regression trees do not make use of the global linearity assumption and thus may not be expected to perform well on these datasets.

2. Method

In the first step of Hadi and Simonoff's [7] Method 2 algorithm, a subset of K observations is selected as an initial set of proposed outliers. The remaining $N - K$ observations are placed in a "clean" set, M, of proposed non-outlying observations. After first scaling each predictor and the response to have mean zero and standard deviation one, we used single-linkage clustering to identify the initial set of K outliers by selecting the last K observations connected to the cluster, under the assumption that these are the most outlying. As per [7], in choosing between two clusters at a particular link, we selected the smaller of the two as more outlying. In some cases, this approach identified $K+1$ candidate outliers due to the link connecting two clusters of size one. We believe that increasing the size of K in this case is superior to choosing one of the two observations at random or excluding both.

In practical situations, the true number of outliers, K, is almost always unknown but the choice of K can have a great effect on how well the procedure is able to identify outliers [7]. Selecting a K that is too large can result in swamping non-outliers. Selecting a K that is too small can result in masking of outliers or even failure to detect any known outliers at all. We recommend initially using $\lceil 0.1N \rceil \leq K \leq \lceil 0.2N \rceil$, trending towards the upper end when there are a greater number of suspected outliers, and the lower end when it is suspected that there are fewer.

The remaining two steps of Hadi and Simonoff's [7] Method 2 algorithm seek to identify outliers:

(2) Fit a regression model using the clean set M and calculate the absolute size of the internally studentized residuals or scaled prediction error, calling the value for the jth observation, d_j, where $j = 1, \ldots, N$.
(3) Use d_j to place all N observations in ascending order and then:
   (A) if $d_{<|M|+1>} \geq G$ where $d_{<j>}$ is the jth order statistic of $d_1, \ldots, d_N$, $|M|$ is the number of observations in the clean set M, and G is an appropriate cutoff, immediately identify all observations with $d_j \geq G$ as outliers and halt the algorithm, or
   (B) (i) if $K \geq 1$, let the new M consist of $d_{<1>}, \ldots, d_{<|M|+1>}$, decrement K by 1 and return to Step 2.
   (ii) If $K = 1$, decide there are no outliers in the dataset and stop.

We note that the calculation of d_j and the appropriate cutoff, G, were selected by Hadi and Simonoff [7] to be appropriate in a linear regression context, and that Step 3(A) will identify at least K outliers.

To adapt Hadi and Simonoff's [7] algorithm for regression tree models, we chose a new error measure, $d_j = |(y_j - \overline{y}_t)/s_t|$ where y_j is the response value for observation $j = 1, \ldots, N$, $\overline{y}_t$ is the predicted response value at the terminal node t to which observation j is assigned, and s_t is the standard deviation of response values at this terminal node. Note that $\overline{y}_t$ and s_t are based on only M, and that observation j may be in M or in the proposed outlier set.

This measure does well at detecting outlying observations with predictor value(s) that are within the range of those observed in M. If an observation has predictor value(s) outside those observed in M, and these predictors are used in the decision rules in the tree, then the observation will be predicted by the most extreme terminal node with respect to those predictors. This can result in poor prediction. To illustrate this, Figure 1 shows a scatterplot of the 20 observations in the First Word-Gesell Adaptive Score dataset. It is desired to predict the Gesell score using a child's age when they speak their first word. Also shown is the regression tree built using a clean set M that contained all but $K = 3$ potential outliers, namely observations 2, 18, and 19. Two binary splits at ages 9 and 11 created a $T = 3$ terminal node tree and the predicted values $\overline{y}_1, \overline{y}_2, \overline{y}_3$ at these nodes are indicated by the horizontal line segments. In a classical linear regression setting, observation 19 is considered an outlier but observation 18 is a leverage point since both observations 2 and 18 follow or influence the linear trend [12]. However, in a regression tree context observations 2, 18, and 19 all have large error measure values if they are excluded from M and may all be identified as outliers. This mischaracterization of observations 2 and 18

is due to all observations with child's age above 11 being modeled with the constant predicted value of $\overline{y}_3$. A linear regression model can make use of the linearity assumption to improve this prediction.

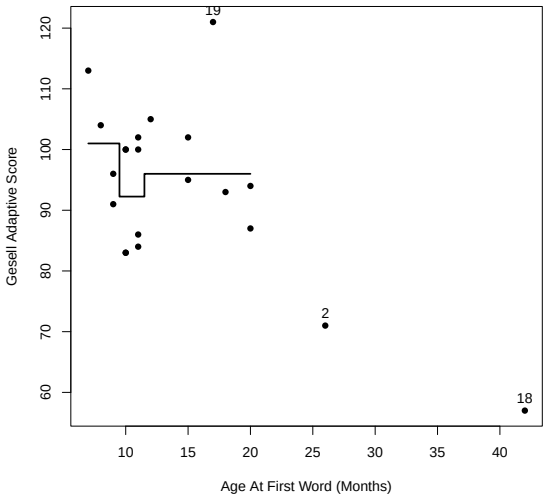

FIGURE . Scatterplot of First Word-Gesell Adaptive Score data with regression tree superimposed. The tree was built using a clean set M that excluded $K = 3$ observations, namely, observations 2, 18 and 19.

For the cutoff, G, we use $t_f(1-\alpha/K, n_t-1)$, the value on a folded-t distribution with n_t-1 degrees of freedom and probability $1-\alpha/K$ below it. Notice that K will change with each iteration through Step 3(Bi), and that n_t (the number of observations in M that were assigned to node t) will change depending on the node that an observation is placed in or predicted by. If one is willing to assume that response values at each terminal node of the tree come from a normal distribution with node-specific mean and standard deviation then it seems appropriate to assume that $(y_j-\overline{y}_t)/s_t$ is a value from a t-distribution with n_t-1 degrees of freedom. In this case, d_j would have a folded-t distribution [8]. While normality within a terminal node is dubious, especially in cases where M may indeed include outliers, it seems the most appropriate assumption to make.

As discussed in [13] one can control either the individual or experiment-wise error rate, although approximate rather than precise control is needed, and further that the Bonferroni correction may be overly conservative for

the context. Choosing to control the individual error rate so that, for example, we would nominally expect at the $\alpha = 0.05$ level, 5% of observations in a truly clean data set to be incorrectly identified as outliers, seemed appropriate in our motivating context of outlier detection in ecology due to the exploratory nature of the problem; we are interested in detecting potential outliers so that they can be investigated further. Thus, we conducted a simulation study to determine the appropriate divisor of α in order to achieve an individual error rate of α. Each of the 10,000 datasets in our simulation consisted of $N = 60$ observations on a single predictor and response. The predictor values consisted of 20 observations from each of three non-overlapping uniform distributions in order to create clear separation between the tree nodes. The response values were generated from normal distributions with means and standard deviations specific to each of the three nodes. In simulation 1, we held out $K = 3$ of the 60 observations (one per node) so that $|M|$ was initially 57, and in simulation 2, K was increased to 6 with two per node held-out of M. If at any stage of the algorithm, a tree failed to assign the N observations to their correct nodes, that dataset was discarded and replaced.

Table 1 shows the average percentage of the $N = 60$ observations identified as outliers across the 10,000 clean datasets using each of four corrections:

(I) no adjustment to α,
(II) division by the number, K, of comparisons implied by $d_j \geq G$ in Step 3A, recognizing that K changes each time through Step 3(Bi),
(III) same as adjustment (II) but with K fixed at the initial value in all iterations through Step 3(Bi), and
(IV) dividing by $|M| + 1$ as done in [7].

In all cases, the target individual error rate was 5%.

TABLE 1. Average percentage of the 60 observations in 10,000 different no-outlier datasets that are identified as outliers based on four different corrections made to the Type 1 error rate, α. K_{initial} indicates that K was fixed at its initial value throughout and K_{current} indicates that K changed with each Step 3(Bi).

α **Correction Used**	**Simulation 1** ($K = 3$)	**Simulation 2** ($K = 6$)
I: None	4.50%	7.96%
II: $\alpha/K_{\text{initial}}$	1.99%	1.80%
III: $\alpha/K_{\text{current}}$	3.02%	3.72%
IV: $\alpha/\lvert M\rvert + 1$	0.10%	0.17%

We chose to adjust α by the current value of K rather than using no adjustment, as this resulted in values that were closest to and lower than 5%. Use of $\alpha/K_{\text{current}}$ rather than α would make it somewhat more difficult to declare an observation an outlier. As an aside, if one wished to control the experiment-wise error rate, as we assume Hadi and Simonoff [7] wished to, then $|M|+1$ would be best.

Another consideration for tree models is setting the minimum number of observations for each terminal node of the tree, c. We have found that c should never be set lower than two so as to avoid $s_t = 0$, and no higher than five for small data sets in order to avoid masking outliers.

Our proposed algorithm for detecting outliers in regression trees is then:

(1) Use single-linkage clustering to identify an initial set of K (or $K+1$) outliers after scaling the predictor(s) and response to each have mean 0 and standard deviation 1. If $K+1$ outliers are identified, then this is the new K.
(2) Construct a regression tree with terminal nodes $t = 1, \ldots, T$ using the clean data set M consisting of $N-K$ observations and minimum node size of c. Assign each observation $j = 1, \ldots, N$ to the appropriate node t in the tree according to the tree's decision rules and then:
    (A) calculate the error measure $d_j = |(y_j - \overline{y}_t)/s_t|$ using the node specific mean and standard deviation.
    (B) Calculate $p_j = d_j/t_f(1-\alpha/K, n_t - 1)$ where the denominator is the cutoff, G, discussed earlier, noting that p_j is node specific and dependent on the current K.
(3) Sort the $p_1, \ldots, p_N$ into ascending order.
    (A) if $p_{<|M|+1>} \geq 1$ where $p_{<j>}$ is the jth order statistic of $p_1, \ldots, p_N$ and $|M|$ is the number of observations in the clean set M, immediately identify all observations with $p_j \geq 1$ as outliers, or
    (B) (i) if $K \geq 1$, let the new M consist of the $|M|+1$ observations with lowest p_j values i.e. $p_{<1>}, \ldots, p_{<|M|+1>}$, decrement K by 1 and return to Step 2.
        (ii) If $K = 1$, decide there are no outliers in the dataset and end algorithm.

3. Results

In order to verify the validity of the algorithm, we observed how it performed on several known data sets, including the quartet of benchmark data sets identified by [7] as necessary for a valid algorithm in multiple outlier detection problems, as well as examples which highlight the importance of

the more subjective variables in our algorithm. Hadi and Simonoff's [7]'s quartet are examples 1–4 below. We present results for minimum node sizes of $c \in \{2, 3, 4, 5\}$ and unless otherwise stated we used an individual error rate of $\alpha = 0.05$ with correction III applied. Although our recommendation is that $\lceil 0.1N \rceil \leq K \leq \lceil 0.2N \rceil$ be used in practice, we include results for a broader range of K values for completeness (the upper bound being the largest value for which a tree could be built).

Examples

(1) **Telephone Data:** This data set contains the number of international telephone calls made in Belgium for each of 24 years. The response is the number of calls and the predictor is year. The data is known to contain six outliers (years 1964-1969) due to changes in the measurement unit for the response. The years 1963 and 1970 were also partially affected, but are not considered significant outliers. Our algorithm correctly identifies as outliers only the six years with the different recording system, for $K \geq 4$ for all c.

(2) **Hertzsprung-Russell Star Data:** This data set contains the logarithms of the effective surface temperature and the light intensity for the 47 stars in the star cluster CYG OB1, of which four stars are red giants and hence are different from the rest. Temperature is the predictor of light intensity. Our algorithm correctly identifies only the four giant stars as outlying for $3 \leq K \leq 25$ for all c, and for at least one c for $K \geq 26$.

(3) **Hawkins-Bradu-Kass Data:** This artificial data set contains 75 observations and has 3 predictors. Observations 1-10 are outliers and there are good leverage points at observations 11-14. Our algorithm successfully identifies the 10 outlying observations as outlying for $9 \leq K \leq 16$ for all c. For $17 \leq K \leq 26$ the algorithm works correctly for at least one of these c; however, for other c values, the algorithm identified at least one observation in addition to the 10 outliers.

(4) **Modified Wood Gravity Data:** The modified wood gravity data set contains information on five anatomical factors of wood and how these affect the wood's specific gravity for each of 20 samples. Observations 4, 6, 8, and 19 are outliers. Given the smaller number of observations in this problem, we increased our individual error rate to 0.1. With this α, our algorithm correctly identifies only the four outliers for $K \geq 5$ for at least one c.

(5) **The single outlier case:** The First Word-Gesell Adaptive Score data discussed earlier serves as an example for the single outlier case. The algorithm correctly identifies only point 19 as an outlier

for $K \in \{2, 5, 6, 7, 8\}$ for at least one c. For $K \in \{3, 4\}$ the algorithm identified the leverage points, observations 2 and 18, in addition to observation 19.

We note that with a few exceptions the algorithm correctly detected only the known outliers for all K within the recommended ranges. The four exceptions ($K = 3$ in Example 1, $K = 8$ in Example 3, and $K \in \{2, 3\}$ in Example 4) were cases where the true number of outliers was at least one more than the K used. Given the relatively good performance of the algorithm for K that were initially too large, this suggests that it is better to use larger rather than smaller values of K initially.

Application to the Friends of the Shenandoah River Monitoring Data

The Friends of the Shenandoah River (FOSR) data set [6] includes measurements collected from 2001-2011 on 222 river and stream sites in the Shenandoah Valley, including six water metrics: nitrate, orthophosphate, ammonia, dissolved oxygen, pH, and turbidity. For each site, we also have data on 53 geographical predictors and these are used to predict the six water metrics. The hypothesis is that sites whose water metrics differ significantly from other sites with similar geography may be affected by karst geology. Features of karst landscape may include caves and underground drainage systems. The karst and outlier status of the observations in the FOSR dataset is unknown. In order to maintain the assumption of independence of sites, we limited our study to only those sites that are upstream, defined as being sites whose entire flow of water has never flowed through any other site in the data set. This left us with 52 observations.

We ran the algorithm with $c \in \{2, 3, 4, 5\}$ and $2 \leq K \leq 10$. One site in particular (JR10) was identified as an outlier for three of the six response metrics for most K and c values, and would be the prime candidate to explore further for the presence of karst geology. The algorithm also identified multiple outlying sites for each individual metric, and three sites (JR01, JR06, and NR05) that are potential outliers with respect to at least two metrics. The three JR prefixed sites mentioned are all within 10 miles of one another which may be promising. While it is possible that some of the identified sites are not outliers, in this context incorrectly identifying a site as an outlier is preferable to failing to identify an outlying site due to the ability to do more physical exploration of the site.

4. Discussion

We proposed an automatic method for detecting outliers using regression tree models and applied this method to four example data sets on

which good performance is considered necessary for a viable outlier detection algorithm in the context of linear regression. Our algorithm performed effectively on these benchmark examples even without the use of the global linearity assumption available in linear regression. A simulation study suggested an appropriate correction to use in controlling the individual error rate.

Example 5 (First Word-Gesell Adaptive Score data) suggests that our algorithm could be susceptible to falsely identifying observations that are "good" leverage points in the context of classical linear regression models as being outliers. However, we believe this is due to the data in this example following a linear trend that is better suited to a linear regression model than a regression tree. In practice, regression trees are not always the best choice of model.

A problem arises when the standard deviation of response values within a terminal node (s_t) is zero. This is more likely to occur (although still rare) when the minimum node size is small, but we have observed it as high as $c = 4$. With no within-node variability, the error measure, d_j, for observations within M that lie in that node and outlier candidates that are predicted by that node, will be undefined. Our current recommendation is that if d_j is undefined and observation j is within the current clean set M, then d_j should be set to zero. This follows because j is joined by at least one other observation with similar predictor values and the same response value, within the currently assumed non-outlier observations. Setting d_j to zero is equivalent to assuming that j is not an outlier. If an observation outside M is predicted by a node with zero-variance, the recommendation is less certain. If one is not weary of false-positives, we would recommend setting d_j equal to a value that ensures that this observation remains in the outlier set.

More work is needed in examining the performance of the algorithm on datasets with multiple predictors in which the commonly accepted outliers are not extreme in any single predictor. For example, the dataset studied by [3] contained five predictors which sought to explain the verbal mean test score for all sixth grade students from 20 schools in New England. The three outliers do not differ significantly in any one predictor and the algorithm fails to identify them.

The choice of error measure, d_j, will be examined further in future research, with hopes of removing the issues discussed previously. In particular, there may be promise in the use of the median absolute deviation. However, the current measure is simple and intuitive, and still provides an effective automatic approach for identifying outliers in the context of regression tree models that can be used when one is unable to justify the assumptions necessary for use of linear regression.

Acknowledgments

The authors are grateful to the Jeffrey E. Tickle Scholarship for funding this work conducted during a summer research experience for undergraduates and to Dr Bruce Wiggins for providing the FOSR application.

References

[1] R. Chambers, A. Hentges, and X. Zhao, *Robust automatic methods for outlier and error detection*, Journal of the Royal Statistical Society A, **167.2** (2006), 323–339.
[2] N. Cheze and J.-M. Poggi, *Iterated Boosting for Outlier Detection*, Data Science and Classification, Springer, Berlin, 2006, pp. 213–220.
[3] J. Coleman et al., *Equality of Educational Opportunity*, Office of Education, U.S. Department of Health: Washington, D.C., 1966.
[4] G. De'ath and K. Fabricius, *Classification and regression trees: A powerful yet simple technique for ecological data analysis*, Ecology, **81.11** (2000), 3178–3192.
[5] J. Elith, J. Leathwick, and T. Hastie, *A working guide to boosted regression trees*, Journal of Animal Ecology, **77** (2008), 802–813.
[6] Friends of the Shenandoah River, Monitoring Data Files - FOSR, 2013, `http://fosr.org/state-of-the-river/monitoring-data-files/`.
[7] A. Hadi and J. Simonoff, *Procedures for the identification of multiple outliers in linear models*, Journal of the American Statistical Association, **88.424** (1993), 1264–1272.
[8] S. Psarakis and J. Panaretos, *The folded t distribution*, Communications in Statistics-Theory and Methods, **19.7** (1990), 2717–2734.
[9] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, `http://www.R-project.org/`.
[10] A. Rahmatulla Imon, *Identifying multiple influential observations in linear regression*, Journal of Applied Statistics, **32.9** (2006), 929–946.
[11] B. Ripley, *TREE: Classification and Regression Trees*, R package version 1.0-35, 2014.
[12] P. Rousseuw and A. Leroy, *Robust Regression and Outlier Detection*, John Wiley, New York, 1987.
[13] J. Simonoff, *General approaches to stepwise identification of unusual values in data analysis*, Directions in Robust Statistics and Diagnostics, Springer, New York, 1991, pp. 223–242.
[14] K. Zirkle, *Predicting water quality in the Shenandoah Valley*, Senior Honor's Thesis, James Madison University, Harrisonburg, VA, 2013.

MSC2010: 62G08

Key words and phrases: Outlier detection, influential observations, backward-stepping, robust models, outliers, CART

Department of Mathematics and Statistics, James Madison University, MSC 1911, Harrisonburg, VA 22807
E-mail address: granerne@dukes.jmu.edu

Department of Mathematics and Statistics, James Madison University, MSC 1911, Harrisonburg, VA 22807
E-mail address: prinssc@jmu.edu