Journal of the Korean
Statistical Society
Vol. 25, No. 1, 1996

Model Selection for Tree-Structured Regression[†]

Sung-Ho Kim[1]

Abstarct

In selecting a final tree, Breiman, Friedman, Olshen, and Stone (1984) compare the prediction risks of a pair of tree, where one contains the other, using the standard error of the prediction risk of the larger one. This paper proposes an approach to selection of a final tree by using the standard error of the difference of the prediction risks between a pair of trees rather than the standard error of the larger one. This approach is compared with CART's for simulated data from a simple regression model. Asymptotic results of the approaches are also derived and compared to each other. Both the asymptotic and the simulation results indicate that final trees by CART tend to be smaller than desired.

Key Words : CART; Prediction risk; Convergence in probability; Final tree.

[†] This research was supported by NON-DIRECTED RESEARCH FUND, Korea Research Foundation.

[1] School of Humanities and General Sciences, KAIST, Daejon, 305-701, Korea.

1. INTRODUCTION

Tree-structured regression (TSR) is a non-linear sequential regression scheme where regressor variables are involved one after another, each being selected, based on a part of data which is determined by the previously selected regressor variables, so that the prediction risk may be minimized. In a regular regression, we add a regressor variable when the whole data suggests that it may improve predictions most among the yet unselected regressor variables. On the other hand, once a regressor variable is selected in TSR, the subsequent selection is based on a part of the whole data set that corresponds to the outcome of the previously selected variables. The sequential selection scheme can be depicted in a tree-like graph. We call such a graph a tree.

The use of trees in regression analysis dates back to the Automatic Interaction Detection program (AID) developed by Morgan and Sonquist (1963), which was followed by the classification program THAID, developed by Morgan and Messenger (1973). Breiman, Friedman, Olshen, and Stone (1984) proposed an algorithm, which they called *Classification And Regression Trees* (CART), that is designed as a sequential decision aid for classification or regression problems. Loh and Vanichsetakul (1988) proposed an algorithm called *Fast Algorithm for Classification Trees* (FACT) which involves recursive application of linear discriminant analysis, with the regressor variables at each stage being appropriately chosen according to the data and the type of splits desired. These four tree algorithms are primarily for prediction problems.

AID, THAID and FACT build trees by adding nodes (variables) until a certain condition holds. CART builds trees in two steps. First, a tree is grown beyond the optimal tree-size, and then an optimal tree is chosen by pruning the grown-up tree. In general terms, CART uses a loss function in the growing process, which ends when the expected loss no longer decreases. It then estimates the prediction risks by the subtrees of the grown-up tree by the cross-validation or the test-set method and selects the subtree with the minimum prediction risk. In the estimation by cross-validation, CART uses as a criterion a liear combination of the loss function and the tree-size, to control the tree-size for the cross-validation subsets of data. Breiman et al. (1984) developed theories concerning the search of optimal trees using the criterion, and Kim (1994) extended the theory so that the observation cost may be considered in addition to the loss function and the tree-size.

One of the advantages of the tree-structured approach is that the tree procedure output gives easily understood and interpreted information regarding the predictive structure of the data. The tree procedure output, almost universally, provides an illuminating and natural way of understanding the structure of the problem (Breiman et al. (1984), p. 58). However, extensive exploration and careful interpretation are necessary to arrive at sound conclusion (Einhorn (1972), Doyle (1973), Breiman et al. (1984)).

Two key issues relevant to sound interpretation are instability and selection of the final tree. Instability issue is discussed in Breiman et al. (1984, pp. 156-160). The selection rule of the final tree by CART is described in subsection 3.4.3 and section 11.6 of Breiman et al. (1984). CART's selection rule of the final tree can be improved in the following sense. Suppose we have data from a simple linear regression model. We test significance of the regression model by testing whether the regression coefficient is zero or not, which is equivalent to testing whether the prediction risk or the mean squared error decreases by the simple regression model. We see an analogy in CART that it compares prediction risks between a pair of regression trees where one is a submodel of the other. But a distinction between the regular regression and the tree-structured regression by CART is that CART does not use a standardized form of the difference of the prediction risks between the optimal tree and any of its subtrees or an F-like statistic as we do in comparing a regular regression model and its reduced one. Actually, Loh and Vanichsetakul (1988) use F-statistics for splitting and stopping rules. CART uses the standard error (call it se_1) of the estimate of the prediction risk of the optimal tree in comparing prediction risks. It is desirable to use the standard error (call it se_2) of the difference between the two prediction risks. se_1 and se_2 are not equal to each other in general, and so the two standard errors may end up with different regression trees.

In this paper we will derive an asymptotic formula for the statistic that is proposed for selecting final trees and that for the statistic that is used in CART. We will then use simulated data to demonstrate differences between the proposed and CART's approaches. It is shown that CART's approach would often produce smaller trees than desired.

This paper consists of four sections and two appendices. Section 2 introduces notations, section 3 presents the main result of this paper, deriving the asymptotic property of the statistic which is proposed to be used for final tree selection, and section 4 concludes the paper. Appendix 1 proves the main theorem in section 3 and Appendix 2 contains a table that is referred to in section 3.

2. NOTATION

In this paper only one variable, if any, is used to split data at a node. We denote the decision rules before and after split by $d_0(\cdot)$ and $d_1(\cdot)$, respectively. To estimate the accuracy of the predictions by a decision rule, we will use a test set method. In a test set method, we divide the whole data set into two parts; one part being used for the construction of decision rules, the other for the test of the decision rules or the estimation of some parameters involved. The former part is called the *learning or training* (data) set, the latter the *test* (data) set. The size of the learning set is denoted by N_0, and by $N_1 = N - N_0$ for the test set.

We consider a data set, $\{(X_i, Y_i)\}_{i=1}^{N}$, from a simple regression model

$$Y = \beta X + \epsilon, \tag{2.1}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$, and ϵ is independent of X. This model will be assumed throughout this paper.

We will use a squared error loss as given by

$$L(Y_i, d(X_i)) = (Y_i - d(X_i))^2,$$

where $d(\cdot)$ stands for a decision rule. In a test set method, the expected loss of the prediction by $d(\cdot)$ is estimated by

$$R(d) = \sum_{i \in \eta} L(Y_i, d(X_i))/N_1, \tag{2.2}$$

where η is the index set of the elements of the test set.

3. MAIN RESULT AND DISCUSSION

The method to split data is described in detail in Breiman et al. (1984). We split data so that the prediction risk, as estimated based on the learning set, may decrease most among all the possible splits when we observe a selected regressor variable.

We denote the sample variance of $L(Y_i, d_0(X_i)) - L(Y_i, d_1(X_i))$ by $\hat{V}(L(d_0) - L(d_1))$ and the sample variance of $L(Y_i, d_1(X_i))$ by $\hat{V}(L(d_1))$. Let

$$IR_{dif} = \frac{R(d_0) - R(d_1)}{\sqrt{\hat{V}(L(d_0) - L(d_1))}}, \tag{3.1}$$

and

$$IR_1 = (R(d_0) - R(d_1))/\sqrt{\hat{V}(L(d_1))}. \tag{3.2}$$

The test set is split into two subsets according to the value of X. If the value of X is smaller than a certain value, the corresponding case is classified into one subset, otherwise into the other. We will call the former subset the *left* subset, and the latter the *right* subset. The terms left and right are from the left and the right child nodes in a tree. We will denote by π_L the probability that a case falls into the left subset and by π_R for the right subset. Let $\phi_X^{(a)}$ denote the ath central moment of X and X_L and X_R the random variables for X of the cases that fall into the left subset and the right subset, respectively. Let μ_X and σ_X^2 denote the mean and the variance of the subscript variable X, respectively.

It does seem difficult to explore the distributions of IR_1 and IR_{dif} theoretically. We will instead explore the stochastic limits of IR_1 and IR_{dif} and compare the performances IR_1 and IR_{dif} via simulation. The formulae of the stochastic limits are given in Theorems 3.1 and 3.2 below.

Theorem 3.1 Assume that the 4th moment of X is bounded. Suppose that $\delta_1 < \frac{N_0}{N_1} < \delta_2$, for some $0 < \delta_1 < \delta_2 < \infty$. Then, if V_{dif} in expression (3.3) is positive, IR_{dif} converges in probability at the rate of $N_1^{-1/2}$ to

$$K_{dif} = \frac{\pi_L \pi_R \beta^2 (\mu_{X_R} - \mu_{X_L})^2}{\sqrt{V_{dif}}}, \tag{3.3}$$

as $N_1 \rightarrow \infty$, where

$$\begin{aligned} V_{dif} &= \beta^4 \left\{ \phi_X^{(4)} - \overline{\phi_X^{(4)}} - (\sigma_X^4 - (\overline{\sigma_X^2})^2) - 4\pi_L\pi_R(\mu_{X_R} - \mu_{X_L})(\phi_{X_R}^{(3)} - \phi_{X_L}^{(3)}) \right\} \\ &\quad + 4\beta^2\sigma_\epsilon^2(\sigma_X^2 - \overline{\sigma_X^2}), \end{aligned}$$

where, for a non-negative integer a,

$$\overline{\phi_X^{(a)}} = \pi_L \phi_{X_L}^{(a)} + \pi_R \phi_{X_R}^{(a)} \tag{3.4}$$

and

$$\overline{\sigma_X^2} = \pi_L \sigma_{X_L}^2 + \pi_R \sigma_{X_R}^2.$$

Proof. See Appendix 1.

Both the numerator and the denominator in the right hand side of (3.3) are multiples of β. But we will keep the formula in (3.3) as it is, where V_{dif} is the stochastic limit of $\hat{V}(L(d_0) - L(d_1))$.

Theorem 3.2 Under the same condition as in Theorem 3.1, IR_1 converges at the rate of $N_1^{-1/2}$ in probability to

$$K_1 = \frac{\pi_L \pi_R \beta^2 (\mu_{X_R} - \mu_{X_L})^2}{\sqrt{\beta^4 \left(\phi_X^{(4)} - (\overline{\sigma_X^2})^2\right) + 4\beta^2 \sigma_\epsilon^2 \overline{\sigma_X^2} + \phi_\epsilon^{(4)} - \sigma_\epsilon^4}} \qquad (3.5)$$

as $n \longrightarrow \infty$.

Proof. Its proof is part of the proof of the preceding theorem. The denominator of the fraction in expression (3.5) is the square root of the limit of B in expression (3.24) in Appendix 1. The denominator can be derived straightforwardly from expressions (3.17) and (3.25) in Appendix 1. The numerator of the fraction in expression (3.5) is from (3.49) in Appendix 1. Property (3.33) in Appendix 1 is applied to obtain the limit (3.5).

CART uses a scalar multiple of $\sqrt{N_1} IR_1$ for selecting final trees, where $R(d_1)$ is the sample prediction risk from the optimal tree and $R(d_0)$ may be viewed as the sample prediction risk from any interested subtree of the optimal tree. Under the 1-se rule, CART selects the tree whose size is the largest among those whose corresponding $\sqrt{N_1} IR_1$ values are larger than or equal to 1. When data are from the simple regression model (2.1), the 1-se rule selects the X variable in (2.1) when $\sqrt{N_1} IR_1$ is not less than 1.

Before having a close look at the expressions in Theorems 3.1 and 3.2, we will compare, assuming that data are from the simple model (2.1), the selection frequencies of the regressor variable between $\sqrt{N_1} IR_1$ and $\sqrt{N_1} IR_{dif}$. Table A.1 in Appendix 2 is obtained from a simulation experiment where X in expression (2.1) is binary taking on 1 or 2 with equal probability, $N_0 = N_1 = 25, 50, 100, 200, 300$, $\beta = 0.5, 1, 2$, and $\sigma_\epsilon = 1, 3$. The last 4 columns of the table are the relative frequencies that $\sqrt{N_1} IR_1$ and $\sqrt{N_1} IR_{dif}$ each exceeds the corresponding thresholds out of 500 iterations. The thresholds, 1.65, 1.96, 2.33, are respectively 5, 2.5, 1 upper-percentiles of the standard normal distribution, and the threshold 1 refers to the *1-se* rule of CART as for $\sqrt{N_1}IR_1$.

The table indicates that $\sqrt{N_1}IR_1$ tends to recommend smaller final trees than $\sqrt{N_1}IR_{dif}$ unless σ_ϵ is relatively small compared with β. In the table, the frequency of variable-selection by $\sqrt{N_1}IR_1$ is not less than that by $\sqrt{N_1}IR_{dif}$ when $\beta = 2$, and $\sigma_\epsilon = 1$. Otherwise, $\sqrt{N_1}IR_{dif}$ recommends the variable-selection more often than $\sqrt{N_1}IR_1$. We will elaborate on this below.

Recall that, in the simulation above, X takes on 1 or 2 equally likely and the simple model (2.1) is assumed. So we can easily obtain that

$$\begin{aligned} V_1 &= 2\sigma_\epsilon^4 \\ V_{dif} &= \beta^2\sigma_\epsilon^2. \end{aligned}$$

From this result, we can have the values of $\sqrt{V_1}, \sqrt{V_{dif}}, K_1$, and K_{dif} as in Table 3.1 under the simulation setup.

The ranks of the K_1 values are of the same pattern as those of the K_{dif} values. The last column of Table 3.1 is the ranks of the K_{dif} values in the ascending order. The table shows that among the 6 combinations of β and σ_ϵ, the K_{dif} value is the smallest for the (0.5, 3) combination and the largest for the (2, 1) combination, and the same story for K_1. It is very interesting to see that the ranks of the relative frequencies of variable-selection as shown in Table A.1 match exactly to those of K_{dif} or K_1. In other words, the ranks of the relative frequencies are, asymptotically, subject to the values of IR_{dif} (or IR_1). Tables 3.1 and A.1 indicate that, as the values of IR_{dif} (or IR_1) get closer to 0, the regressor variable, X, will less likely be selected for observation.

Table 3.1. Values of $\sqrt{V_1}, \sqrt{V_{dif}}, K_1$, and K_{dif} under the simulation setting

β	σ	$\sqrt{V_1}$	$\sqrt{V_{dif}}$	K_1	K_{dif}	rank
0.5	1	1.41	0.5	0.044	0.125	3
	3	12.73	1.5	0.005	0.042	1
1	1	1.41	1	0.177	0.25	5
	3	12.73	3	0.020	0.084	2
2	1	1.41	2	0.707	0.5	6
	3	12.73	6	0.079	0.167	4

Given the learning data set, $L(Y_1, d(X_1)), \cdots, L(Y_{N_1}, d(X_{N_1}))$, are independent and identically distributed, and so are $L(Y_i, d_0(X_i)) - L(Y_i, d_1(X_i))$, for $i = 1, 2, \cdots, N_1$. Thus $R(d_0) - R(d_1)$ converges in law to a normal distribution. Furthermore, given the learning data set, IR_{dif} converges in law, under the condition of Theorem 3.1, to a normal distribution by Slutsky's Theorem. In particular, when $\beta = 0$, we have $K_{dif} = 0$. In other words, IR_{dif} converges in law to a normal distribution with mean equal to zero. But as indicated in Theorem 3.1 the convergence is very slow. Recall that $\hat{V}(L(d_0) - L(d_1))$ is the sample variance of $\{L(Y_i, d_0(X_i)) - L(Y_i, d_1(X_i))\}_{i=1}^{N_1}$. So the standardized form of $R(d_0) - R(d_1)$ is given by $\sqrt{N_1}IR_{dif}$ given the learning data set. When $\beta = 0$, $\sqrt{N_1}IR_{dif}$ converges in law to the standard normal distribution given the learning data set. $\sqrt{N_1}IR_1$ does not have this nice property.

Turning back to Table A.1, suppose that we use $\sqrt{N_1}IR_{dif}$ in variable-selection. The table says, as is noted earlier, that the relative frequencies of variable-selection are the smallest when $\beta = 0.5$ and $\sigma_\epsilon = 3$. The relative frequencies appear near the upper-tail probabilities of the standard normal curve at the threshold points in the table when $N_1 = 25, 50$, while they go beyond the upper-tail probabilities for $N_1 = 100, 200, 300$. The farther K_{dif} stays away from 0 upwards, the farther the relative frequencies appear beyond the upper-tail probabilities upwards. Since $\sqrt{N_1}IR_{dif}$ follows a normal distribution asymptotically, we can pick a threshold, in a meaningful way, based on the standard normal curve. This feature looks more attractive than CART's rule for threshold-selection.

$\sqrt{N_1}IR_{dif}$ gives rise to larger relative frequencies of variable-selection than $\sqrt{N_1}IR_1$ when $\beta = 2$ and $\sigma_\epsilon = 1$, while it is not the case for the other combinations of (β, σ_ϵ) considered in Table A.1. We can see the reason of this in Table 3.1. Note that the numerators of $\sqrt{N_1}IR_{dif}$ and $\sqrt{N_1}IR_1$ are the same and that $\sqrt{V_1}$ is smaller than $\sqrt{V_{dif}}$ only for (2,1) slot out of the 6 (β, σ_ϵ) combinations. So, asymptotically speaking, IR_1 tends to select the regressor variable more often than IR_{dif} when $\beta = 2$ and $\sigma_\epsilon = 1$. This trend is displayed in the (2,1) slots of Table A.1.

Now we will have a closer look at Theorems 3.1 and 3.2. Example 3.1 below may help us better understand expression (3.3), and Example 3.2 illustrates the results of Theorems 3.1 and 3.2.

Example 3.1 Suppose that X is uniformly distributed over the interval [-a, a]. Under this distribution, we can see that $\pi_L = 0.5$. Actually, we split the data set so that

$$\sigma_Y^2 - (\pi_L \sigma_{Y_L}^2 + \pi_R \sigma_{Y_R}^2), \tag{3.6}$$

the variance of the conditional means of Y after split, is maximized. Under the simple (2.1), expression (3.6) can be rewritten as

$$\beta^2(\sigma_X^2 - \pi_L \sigma_{X_L}^2 - \pi_R \sigma_{X_R}^2),$$

which is simplified, from the uniform distribution of X, as in

$$\beta^2 a^2 (1 - \pi_L^3 - \pi_R^3)/3.$$

This expression is maximized when $\pi_L = 0.5$.

We know that

$$\mu_X = 0, \quad \mu_{X_L} = -a/2, \quad \sigma_X^2 = a^2/3, \quad \sigma_{X_L}^2 = a^2/12,$$

$$\phi_X^{(4)} = a^4/5, \quad \text{and} \quad \phi_{X_L}^{(4)} = a^4/(2^4 \cdot 5).$$

Since $\phi_{X_R}^{(3)} - \phi_{X_L}^{(3)} = 0$, we have

$$V_{dif} = \frac{\beta^4 a^4}{12} + \beta^2 a^2 \sigma_\epsilon^2.$$

Hence from (3.3) follows

$$K_{dif} = \frac{\sqrt{3}}{2\sqrt{1 + 12\frac{\sigma_\epsilon^2}{\beta^2 a^2}}},$$

and

$$K_1 = \frac{0.25}{\sqrt{\frac{1}{180} + \frac{\sigma_\epsilon^2}{3\beta^2 a^2} + \frac{\phi_\epsilon^{(4)} - \sigma_\epsilon^4}{\beta^4 a^4}}}$$

follows from (3.5). Under the normal assumption for ϵ, $\phi_\epsilon^{(4)} = 3\sigma_\epsilon^4$, which simplifies the previous expression into

$$K_1 = \frac{0.25}{\sqrt{\frac{1}{180} + \frac{\sigma_\epsilon^2}{3\beta^2 a^2} + \frac{2\sigma_\epsilon^4}{\beta^4 a^4}}}.$$

Both K_{dif} and K_1 decrease as σ_ϵ increases, and they increase as $|a\beta|$ increases. As σ_ϵ decreases or $a\beta$ increases, the data set is more likely to split.

Table 3.2 compares K_1 and K_{dif} when $\beta = 1$ and $a = 1$. From the table, we can safely say, assuming that data are from a linear regression model, that, for large N_1, IR_1 tends to select variables more often than IR_{dif} when σ_ϵ is relatively small compared with the coefficients of the regressor variables and less often than IR_{dif} when σ_ϵ is relatively large. In other words, CART tends to select smaller (or larger) final trees than desired when σ_ϵ is relatively larger (or smaller) compared with the regression coefficients.

Table 3.2. K_1 and K_{dif} values with $a = \beta = 1$

σ_ϵ	K_1	K_{dif}
0.20	1.682	0.712
0.40	0.753	0.507
0.60	0.403	0.375
0.80	0.245	0.294
1.00	0.163	0.240
1.20	0.116	0.203
1.40	0.087	0.175
1.60	0.067	0.154
1.80	0.053	0.137
2.00	0.043	0.124
2.20	0.036	0.113
2.40	0.030	0.103
2.60	0.026	0.096
2.80	0.022	0.089
3.00	0.019	0.083

Example 3.2 Suppose that X is binary taking on 1 and 2 equally likely, that β and σ_ϵ in expression (2.1) are both equal to 1. Then we can easily see that $K_1 = 0.177$ and $K_{dif} = 0.25$ from results (3.5) and (3.3), respectively. IR_1 and IR_{dif} appear near 0.18 and 0.25, respectively, in Table 3.3, confirming the results of Theorems 3.1 and 3.2 numerically. In the table, $N_1 = N_2$.

Table 3.3. Values of IR_{dif} under the condition of Example 3.2

N_1	20	60	100	200	300	400	500	600	700	800	900	1000
IR_1	0.193	0.187	0.174	0.183	0.180	0.180	0.180	0.180	0.175	0.185	0.179	0.181
IR_{dif}	0.252	0.269	0.243	0.254	0.255	0.255	0.250	0.253	0.244	0.261	0.250	0.255

4. CONCLUDING REMARKS

CART is one of the most popular computer program for tree-regression analysis. It is useful for analysing the data which involve both continuous and discrete or categorical variables, and useful when a relatively large number of variables are involved compared with the size of data. These, among other reasons, may have made CART used in a wide range of research fields including medicine, computer science, social and behavioral science, etc. Efforts to improve CART are as important as the popularity of the program.

In this paper we pointed out that the final tree selection by CART may have to be modified so that the differences of prediction risks between a pair of trees, where one is nested in the other, may be standardized. Compared with the standardized version (i.e., $\sqrt{N_1}IR_{dif}$), it is indicated in the simulation result of section 3 and in Examples 3.1 and 3.2 that CART tends to recommend smaller trees than desired. When using the standardized version, we may use thresholds based on the standard normal curve, since the standardized version converges in law to the standard normal distribution although its convergence is very slow. The standardized version can be applied without difficulty when using the test set method in estimating the prediction risks and the variance of the differences of the prediction risks between a pair of trees.

In Table 3.2 we have seen that K_1 is smaller than K_{dif} for $\sigma_\epsilon \leq 0.8$. This implies that the estimate of the variance of $R(d_1)$ is larger than that of $R(d_0) - R(d_1)$ for $\sigma_\epsilon \leq 0.8$ under the setup of Example 3.1. Actually the ratio of the latte to the former gets larger asymptotically as σ_ϵ increases, indicating that the trees by CART may become far smaller than the trees by the standardized version as σ_ϵ increases. Another possible drawback in CART is, as indicated in Table 3.1, that when the regressor variable is binary, the regression coefficient β is not reflected in V_1 while it is in V_{dif}. This is because V_1 involves $L(Y, d_1(X))$ only.

We may safely conclude that CART tends to give smaller trees than the desired which may be obtained by applying the standardized version in selecting the final tree starting back from the optimal tree (see section 1). The process from the optimal tree to the final tree is analogous to the backward elimination process of regression modelling.

APPENDIX 1

In this section we prove Theorem 3.1. We will use the superscript, asterisk (*), to indicate that the superscribed refer to the learning set, and variables, sets, or numbers concerning the test set are not superscribed. Recall that N_0 is the size of the learning set and N_1 the size of the test set.

$$\begin{aligned} R(d_0) &= \frac{1}{N_1}\sum_{i\in\eta} L(Y_i, d_0(X_i)) \\ &= \frac{1}{N_1}\sum_{i\in\eta}(Y_i - d_0(X_i))^2, \end{aligned} \tag{3.7}$$

where η is the index set of X from the test set. Since the rule $d_0(\cdot)$ is based on the learning data set before splitting,

$$d_0 = \overline{Y}^*. \tag{3.8}$$

Denote by $\hat{\beta}^*$ the least squares estimate of β based on the learning set. Then, we may write

$$Y_i = \hat{\beta}^* X_i + e_i, \tag{3.9}$$

and

$$\overline{Y}^* = \hat{\beta}^* \overline{X}^*. \tag{3.10}$$

From equations (3.7) through (3.10), we have

$$\begin{aligned} R(d_0) &= \frac{1}{N_1}\sum_{i\in\eta}(\hat{\beta}^* X_i + e_i - \hat{\beta}^* \overline{X}^*)^2 \\ &= \hat{\beta}^{*2}\tau_{2,0} + 2\hat{\beta}^*\tau_{1,1} + \tau_{0,2}, \end{aligned} \tag{3.11}$$

where for $a, b \geq 0$,

$$\tau_{a,b} = \frac{1}{N_1}\sum_{i\in\eta}(X_i - \overline{X}^*)^a e_i^b. \tag{3.12}$$

If we consider the left and the right subsets, we can express $R(d_1)$ by

$$\begin{aligned} R(d_1) &= \frac{1}{N_1}\sum_{i\in\eta_L} L(Y_{Li}, d_L) + \frac{1}{N_1}\sum_{i\in\eta_R} L(Y_{Ri}, d_R) \\ &= \frac{1}{N_1}\left[\sum_{\eta_L}(Y_{Li} - d_L(X_i))^2 + \sum_{\eta_R}(Y_{Ri} - d_R(X_i))^2\right]. \end{aligned} \tag{3.13}$$

where η_L and η_R are index sets for the test set elements that belong to the left and the right subsets, respectively; Y_{Li} and Y_{Ri} are the Y-values of the ith case in the left and the right subset, respectively; and d_L and d_R are the prediction rules for the left and the right subsets, respectively.

Since X and Y are from a regression model in (2.1), we may write

$$Y_{Li} = \hat{\beta}^* X_{Li} + e_{Li} \ \text{ and } \ Y_{Ri} = \hat{\beta}^* X_{Ri} + e_{Ri}. \tag{3.14}$$

$$d_L(X_i) = \overline{Y}^*_L = \hat{\beta}^*\overline{X}^*_L + \overline{e}^*_L \ \text{ and } \ d_R(X_i) = \overline{Y}^*_R = \hat{\beta}^*\overline{X}^*_R + \overline{e}^*_R, \tag{3.15}$$

where $\overline{Y}^*_L$ ($\overline{Y}^*_R$) is the sample mean of the Y's that belong to the left (right) learning subset after split.

$$\overline{X}^*_L = \sum_{\eta^*_L} X^*_i / N_{0L} \ \text{ and } \ \overline{X}^*_R = \sum_{\eta^*_R} X^*_i / N_{0R}, \tag{3.16}$$

where η^*_L (η^*_R) is the index set of X's that belong to the left (right) learning subset after split, and N_{0L} (N_{0R}) is the cardinality of η^*_L (η^*_R).

Using the above 3 expressions, we can write

$$\begin{aligned} &R(d_1) \\ &= \frac{1}{N_1}\left[\sum_{i\in\eta_L}(\hat{\beta}^* X_{Li} + e_{Li} - \overline{Y}^*_L)^2 + \sum_{i\in\eta_R}(\hat{\beta}^* X_{Ri} + e_{Ri} - \overline{Y}^*_R)^2\right], \\ &= \frac{1}{N_1}\left[\sum_{i\in\eta_L}(\hat{\beta}^*(X_{Li} - \overline{X}^*_L) + e_{Li} - \overline{e}^*_L)^2\right] + \frac{1}{N_1}\left[\sum_{i\in\eta_R}(\hat{\beta}^*(X_{Ri} - \overline{X}^*_R) + e_{Ri} - \overline{e}^*_R)^2\right] \\ &= \hat{P}_L\left[\hat{\beta}^{*2}\lambda_{2,0} + 2\hat{\beta}^*\lambda_{1,1} + \lambda_{0,2}\right] + \hat{P}_R\left[\hat{\beta}^{*2}\rho_{2,0} + 2\hat{\beta}^*\rho_{1,1} + \rho_{0,2}\right], \end{aligned} \tag{3.17}$$

where for $a, b \geq 0$,

$$\lambda_{a,b} = \frac{1}{N_{1L}} \sum_{i \in \eta_L} (X_{Li} - \overline{X_L^*})^a (e_{Li} - \overline{e_L^*})^b, \tag{3.18}$$

$$\rho_{a,b} = \frac{1}{N_{1R}} \sum_{i \in \eta_R} (X_{Ri} - \overline{X_R^*})^a (e_{Ri} - \overline{e_R^*})^b, \tag{3.19}$$

$$\hat{P}_L = \frac{N_{1L}}{N_1}, \text{ and } \hat{P}_R = 1 - \hat{P}_L.$$

From expressions (3.11) and (3.17) follows

$$\begin{aligned} R(d_0) - R(d_1) &= \hat{\beta^*}^2 (\tau_{2,0} - \hat{P}_L \lambda_{2,0} - \hat{P}_R \rho_{2,0}) + 2\hat{\beta}^* (\tau_{1,1} - \hat{P}_L \lambda_{1,1} - \hat{P}_R \rho_{1,1}) \\ &\quad + (\tau_{0,2} - \hat{P}_L \lambda_{0,2} - \hat{P}_R \rho_{0,2}). \end{aligned} \tag{3.20}$$

The sample variance of $L(Y, d_0(X)) - L(Y, d_1(X))$ is

$$\hat{V}(L(d_0) - L(d_1)) = \frac{1}{N_1 - 1} \sum_{i \in \eta} \{L(Y_i, d_0(X_i)) - L(Y_i, d_1(X_i)) - (R(d_0) - R(d_1))\}^2 .$$

Now, we will represent $\hat{V}(L(d_0) - L(d_1))$ in terms of X, $\hat{\beta}^*$, λ and ρ.

$$\begin{aligned} \hat{V}(L(d_0) - L(d_1)) &= \frac{1}{N_1 - 1} \left\{ \sum_{i \in \eta} (L(Y_i, d_0) - R(d_0))^2 + \sum_{i \in \eta} (L(Y_i, d_1) - R(d_1))^2 \right. \\ &\quad \left. -2 \sum_{i \in \eta} L(Y_i, d_0) L(Y_i, d_1) + 2N_1 R(d_0) R(d_1) \right\}. \end{aligned} \tag{3.21}$$

For the first part in expression (3.21), let

$$A = \frac{1}{N_1} \sum_{i \in \eta} (L(Y_i, d_0) - R(d_0))^2 .$$

Recall that $L(Y_i, d_0(X_i)) = (Y_i - d_0(X_i))^2$. From equations (3.8), (3.9), and (3.10) follows

$$Y_i - d_0(X_i) = \hat{\beta}^* (X_i - \overline{X}^*) + e_i.$$

Thus, we have

$$\begin{aligned} A &= \frac{1}{N_1} \sum_{i \in \eta} L(Y_i, d_0)^2 - R(d_0)^2 \\ &= \hat{\beta^*}^4 \tau_{4,0} + 4\hat{\beta^*}^3 \tau_{3,1} + 6\hat{\beta^*}^2 \tau_{2,2} + 4\hat{\beta}^* \tau_{1,3} + \tau_{0,4} - R(d_0)^2. \end{aligned} \tag{3.22}$$

For the second part in expression (3.21), let

$$B = \frac{1}{N_1} \sum_{i \in \eta} \left(L(Y_i, d_1) - R(d_1) \right)^2 . \tag{3.23}$$

After splitting, Y's are predicted by d_1. The prediction values for the left and the right subsets will be denoted by d_L and d_R, respectively.

$$\begin{aligned} B &= \frac{1}{N_1} \sum_{i \in \eta} L(Y_i, d_1)^2 - R(d_1)^2 \\ &= \frac{1}{N_1} \left\{ \sum_{i \in \eta_L} L(Y_{Li}, d_L)^2 + \sum_{i \in \eta_R} L(Y_{Ri}, d_R)^2 \right\} - R(d_1)^2. \end{aligned} \tag{3.24}$$

Using the notation in expressions (3.14) and (3.15), we can rewrite the first summation part of (3.24) as

$$\hat{P}_L \left(\hat{\beta^*}^4 \lambda_{4,0} + 4\hat{\beta^*}^3 \lambda_{3,1} + 6\hat{\beta^*}^2 \lambda_{2,2} + 4\hat{\beta}^* \lambda_{1,3} + \lambda_{0,4} \right) .$$

By applying the same algebra to the second summation part of (3.24), we have

$$\begin{aligned} B \;=\; & \hat{\beta^*}^4 (\hat{P}_L \lambda_{4,0} + \hat{P}_R \rho_{4,0}) + 4\hat{\beta^*}^3 (\hat{P}_L \lambda_{3,1} + \hat{P}_R \rho_{3,1}) + 6\hat{\beta^*}^2 (\hat{P}_L \lambda_{2,2} + \hat{P}_R \rho_{2,2}) \\ & + 4\hat{\beta}^* (\hat{P}_L \lambda_{1,3} + \hat{P}_R \rho_{1,3}) + (\hat{P}_L \lambda_{0,4} + \hat{P}_R \rho_{0,4}) - R(d_1)^2. \end{aligned} \tag{3.25}$$

For the last part of equation (3.21), we obtain, after an algebra,

$$\begin{aligned} & \frac{1}{N_1} \sum_{i \in \eta} L(Y_i, d_0) L(Y_i, d_1) \\ =& \hat{\beta^*}^4 (\hat{P}_L \lambda_{4,0} + \hat{P}_R \rho_{4,0}) + 4\hat{\beta^*}^3 (\hat{P}_L \lambda_{3,1} + \hat{P}_R \rho_{3,1}) + 6\hat{\beta^*}^2 (\hat{P}_L \lambda_{2,2} + \hat{P}_R \rho_{2,2}) \\ & + 4\hat{\beta}^* (\hat{P}_L \lambda_{1,3} + \hat{P}_R \rho_{1,3}) + (\hat{P}_L \lambda_{0,4} + \hat{P}_R \rho_{0,4}) + \\ & \Bigg\{ \hat{P}_L (\hat{\beta}^* (\overline{X^*_L} - \overline{X^*}) + \overline{e^*_L})^2 \frac{\sum_{\eta_L} L(Y_{Li}, d_L)}{n_L} + \hat{P}_R (\hat{\beta}^* (\overline{X^*_R} - \overline{X^*}) + \overline{e^*_R})^2 \\ & \frac{\sum_{\eta_R} L(Y_{Ri}, d_R)}{n_R} \Bigg\} + 2 \Bigg\{ \hat{P}_L (\hat{\beta}^* (\overline{X^*_L} - \overline{X^*}) + \overline{e^*_L}) \frac{\sum_{\eta_L} L(Y_{Li}, d_L)^{\frac{3}{2}}}{n_L} \\ & + \hat{P}_R (\hat{\beta}^* (\overline{X^*_R} - \overline{X^*}) + \overline{e^*_R}) \frac{\sum_{\eta_R} L(Y_{Ri}, d_R)^{\frac{3}{2}}}{n_R} \Bigg\} . \end{aligned} \tag{3.26}$$

In the above lengthy expression,

$$\frac{\sum_{\eta_L} L(Y_{Li}, d_L)}{N_{1L}} = \hat{\beta^*}^2 \lambda_{2,0} + 2\hat{\beta}^* \lambda_{1,1} + \lambda_{0,2},$$

and

$$\frac{\sum_{\eta_L} L(Y_{Li}, d_L)^{\frac{3}{2}}}{N_{1L}} = \hat{\beta^*}^3 \lambda_{3,0} + 3\hat{\beta^*}^2 \lambda_{2,1} + 3\hat{\beta}^* \lambda_{1,2} + \lambda_{0,3}.$$

Similarly for

$$\frac{\sum_{\eta_R} L(Y_{Ri}, d_R)}{N_{1R}} \text{ and } \frac{\sum_{\eta_R} L(Y_{Ri}, d_R)^{\frac{3}{2}}}{N_{1R}}$$

by replacing λ with ρ. Then we can express $\frac{1}{N_1} \sum_{i \in \eta} L(Y_i, d_0) L(Y_i, d_1)$ in terms of $\hat{\beta}^*$, X, λ and ρ as below:

$$\begin{aligned}
&\frac{1}{N_1} \sum_{i \in \eta} L(Y_i, d_0) L(Y_i, d_1) \\
=&\hat{\beta^*}^4 (\hat{P}_L \lambda_{4,0} + \hat{P}_R \rho_{4,0}) + 4\hat{\beta^*}^3 (\hat{P}_L \lambda_{3,1} + \hat{P}_R \rho_{3,1}) + 6\hat{\beta^*}^2 (\hat{P}_L \lambda_{2,2} + \hat{P}_R \rho_{2,2}) \\
&+ 4\hat{\beta}^* (\hat{P}_L \lambda_{1,3} + \hat{P}_R \rho_{1,3}) + (\hat{P}_L \lambda_{0,4} + \hat{P}_R \rho_{0,4}) + \\
&\Big\{ \hat{P}_L (\hat{\beta}^* (\overline{X}^*_L - \overline{X}^*) + \overline{e^*_L})^2 (\hat{\beta^*}^2 \lambda_{2,0} + 2\hat{\beta}^* \lambda_{1,1} + \lambda_{0,2}) \\
&+ \hat{P}_R (\hat{\beta}^* (\overline{X}^*_R - \overline{X}^*) + \overline{e^*_R})^2 (\hat{\beta^*}^2 \rho_{2,0} + 2\hat{\beta}^* \rho_{1,1} + \rho_{0,2}) \Big\} \\
&+ 2 \Big\{ \hat{P}_L (\hat{\beta}^* (\overline{X}^*_L - \overline{X}^*) + \overline{e^*_L}) (\hat{\beta^*}^3 \lambda_{3,0} + 3\hat{\beta}^*_2 \lambda_{2,1} + 3\hat{\beta}^* \lambda_{1,2} + \lambda_{0,3}) \\
&+ \hat{P}_R (\hat{\beta}^* (\overline{X}^*_R - \overline{X}^*) + \overline{e^*_R}) (\hat{\beta^*}^3 \lambda_{3,0} + 3\hat{\beta}^*_2 \lambda_{2,1} + 3\hat{\beta}^* \lambda_{1,2} + \lambda_{0,3}) \Big\}. \qquad (3.27)
\end{aligned}$$

By combining the three results (3.22), (3.25), and (3.27), we have

$$\begin{aligned}
&\frac{N_1 - 1}{N_1} \hat{V}(L(d_0) - L(d_1)) \\
=\;& A + B - 2 \left\{ \frac{1}{N_1} \sum_{i \in \eta} L(Y_i, d_0) L(Y_i, d_1) - R(d_0) R(d_1) \right\} \\
=\;& \hat{\beta^*}^4 (\tau_{4,0} - \hat{P}_L \lambda_{4,0} - \hat{P}_R \rho_{4,0}) + 4\hat{\beta^*}^3 (\tau_{3,1} - \hat{P}_L \lambda_{3,1} - \hat{P}_R \rho_{3,1}) \\
&+ 6\hat{\beta^*}^2 (\tau_{2,2} - \hat{P}_L \lambda_{2,2} - \hat{P}_R \rho_{2,2}) + 4\hat{\beta}^* (\tau_{1,3} - \hat{P}_L \lambda_{1,3} - \hat{P}_R \rho_{1,3}) \\
&+ (\tau_{0,4} - \hat{P}_L \lambda_{0,4} - \hat{P}_R \rho_{0,4})
\end{aligned}$$

$$
\begin{aligned}
&-2\Bigg[\Big\{\hat{P}_L(\hat{\beta}^*(\overline{X^*_L}-\overline{X^*})+\overline{e^*_L})^2(\hat{\beta^*}^2\lambda_{2,0}+2\hat{\beta}^*\lambda_{1,1}+\lambda_{0,2})\\
&+\hat{P}_R(\hat{\beta}^*(\overline{X^*_R}-\overline{X^*})+\overline{e^*_R})^2(\hat{\beta^*}^2\rho_{2,0}+2\hat{\beta}^*\rho_{1,1}+\rho_{0,2})\Big\}\\
&+2\Big\{\hat{P}_L(\hat{\beta}^*(\overline{X^*_L}-\overline{X^*})+\overline{e^*_L})(\hat{\beta^*}^3\lambda_{3,0}+3\hat{\beta}^*_2\lambda_{2,1}+3\hat{\beta}^*\lambda_{1,2}+\lambda_{0,3})\\
&+\hat{P}_R(\hat{\beta}^*(\overline{X^*_R}-\overline{X^*})+\overline{e^*_R})(\hat{\beta^*}^3\rho_{3,0}+3\hat{\beta}^*_2\rho_{2,1}+3\hat{\beta}^*\rho_{1,2}+\rho_{0,3})\Big\}\Bigg]\\
&-(R(d_0)-R(d_1))^2. \qquad (3.28)
\end{aligned}
$$

Now we will see how

$$IR_{dif}=\frac{R(d_0)-R(d_1)}{\sqrt{\hat{V}(L(d_0)-L(d_1))}}$$

converges as

$$\min\{N_0,N_1\}\to\infty,$$

by examining how $R(d_0)-R(d_1)$ and $\hat{V}(L(d_0)-L(d_1))$ converge.

Some of the results concerning the O_p, o_p definitions are stated below without proof (see Bishop et al. (1975), p. 484 for results (3.29) and (3.31).); for real constants, c and d,

$$
\begin{aligned}
O_p(c)o_p(d) &= o_p(cd). && (3.29)\\
O_p(c)O_p(d) &= O_p(cd). && (3.30)\\
O_p(1)+O_p(1) &= O_p(1). && (3.31)
\end{aligned}
$$

$$\text{If } U_i=O_p(1), \text{ then } cU_i=O_p(1). \qquad (3.32)$$

Suppose $U_i=O_p(1)$ and $V_i=d+o_p(1)$, for $d>0$, then we have (see Rao (1973, p. 124))

$$U_i/V_i=O_p(1). \qquad (3.33)$$

We will see how $\overline{e^*_L}$ (expression (3.15)) converges.

Lemma 3.1 Suppose that the variance of X is bounded. Then, under the condition that $0<\delta_1<\frac{N_0}{N_1}<\delta_2<\infty$,

$$
\begin{aligned}
\overline{e^*_L} &= O_p(N^{-\frac{1}{2}}), \quad \text{and} && (3.34)\\
\overline{e^*_R} &= O_p(N^{-\frac{1}{2}}). && (3.35)
\end{aligned}
$$

Proof.

$$\begin{aligned} \overline{e_L^*} &= \frac{1}{N_{0L}^*}\sum_{\eta_L^*}\left(Y_{Li}^* - \hat{\beta}^* X_{Li}^*\right) \\ &= \frac{1}{N_{0L}}\sum\left((\beta - \hat{\beta}^*)X_{Li}^* + \epsilon_{Li}^*\right) \\ &= \overline{\epsilon_L^*} - (\hat{\beta}^* - \beta)\overline{X}_L^*. \end{aligned} \tag{3.36}$$

Since N_0/N_1 is bounded, we can see, under the condition of the lemma, that

$$\overline{X}_L^* - \mu_{X_L} = O_p(N^{-\frac{1}{2}}), \tag{3.37}$$

$$\overline{X}_R^* - \mu_{X_R} = O_p(N^{-\frac{1}{2}}), \tag{3.38}$$

$$\hat{\beta}^* - \beta = O_p(N^{-\frac{1}{2}}). \tag{3.39}$$

In the same context, we will use $O_p(N^{-1/2})$ instead of $O_p(N_0^{-1/2})$ or $O_p(N_1^{-1/2})$ throughout the rest of the paper.

Since $\overline{\epsilon_L^*} = O_p(N^{-\frac{1}{2}})$, by equations (3.37) and (3.39) and results (3.31) and (3.32), expression (3.34) follows. Expression (3.35) is proved in the same way.

Lemma 3.2 Let a and b be non-negative integers with $a + b > 0$. Suppose that the $(a + b)$th moment of X and the bth moment of ϵ are bounded, and that $\delta_1 < \frac{N_0}{N_1} < \delta_2$, for some $0 < \delta_1 < \delta_2 < \infty$. Then we have,

$$\lambda_{a,b} = \phi_{X_L}^{(a)}\phi_\epsilon^{(b)} + O_p(N^{-\frac{1}{2}}), \quad \text{and} \tag{3.40}$$

$$\rho_{a,b} = \phi_{X_R}^{(a)}\phi_\epsilon^{(b)} + O_p(N^{-\frac{1}{2}}), \tag{3.41}$$

where $\phi_{X_L}^{(a)}$ and $\phi_{X_R}^{(a)}$ are the ath central moments of X_L and X_R, respectively, and $\phi_\epsilon^{(b)}$ is the bth central moment of ϵ.

Proof. $\lambda_{a,b}$ is given in expression (3.18). By expressions (3.14) and (3.36),

$$\begin{aligned} e_{Li} - \overline{e_L^*} &= (\beta - \hat{\beta}^*)X_{Li} + \epsilon_{Li} - \overline{\epsilon_L^*} + (\hat{\beta}^* - \beta)\overline{X}_L^* \\ &= (\beta - \hat{\beta}^*)(X_{Li} - \overline{X}_L^*) + (\epsilon_{Li} - \overline{\epsilon_L^*}). \end{aligned} \tag{3.42}$$

Hence,

$$\begin{aligned}
&\lambda_{a,b} \\
&= \frac{1}{N_{1L}} \sum_{\eta_L} \left(X_{Li} - \overline{X_L^*} \right)^a (e_{Li} - \overline{e_L^*})^b \\
&= \frac{1}{N_{1L}} \sum_{\eta_L} \left(X_{Li} - \overline{X_L^*} \right)^a \left((\beta - \hat{\beta}^*)(X_{Li} - \overline{X_L^*}) + (\epsilon_{Li} - \overline{\epsilon_L^*}) \right)^b \qquad \text{by (3.42)} \\
&= \frac{1}{N_{1L}} \sum_{t_1=0}^{b} \begin{pmatrix} b \\ t_1 \end{pmatrix} \left(\beta - \hat{\beta}^* \right)^{t_1} \sum_{\eta_L} \left(X_{Li} - \overline{X_L^*} \right)^{a+t_1} (\epsilon_{Li} - \overline{\epsilon_L^*})^{b-t_1}. \qquad (3.43)
\end{aligned}$$

Expanding the terms $\left(X_{Li} - \overline{X_L^*} \right)^{a+t_1}$ and $(\epsilon_{Li} - \overline{\epsilon_L^*})^{b-t_1}$ in expression (3.43) yields

$$\begin{aligned}
&\lambda_{a,b} \\
&= \sum_{t_1=0}^{b} \sum_{t_2=0}^{a+t_1} \sum_{t_3=0}^{b-t_1} \begin{pmatrix} b \\ t_1 \end{pmatrix} \begin{pmatrix} a+t_1 \\ t_2 \end{pmatrix} \begin{pmatrix} b-t_1 \\ t_3 \end{pmatrix} (\beta - \hat{\beta}^*)^{t_1} (\mu_{X_L} - \overline{X_L^*})^{a+t_1-t_2} \\
&\quad (-\overline{\epsilon_L^*})^{b-t_1-t_3} \left\{ \frac{1}{N_{1L}} \sum_{\eta_L} (X_{Li} - \mu_{X_L})^{t_2} \epsilon_{Li}^{t_3} \right\}. \qquad (3.44)
\end{aligned}$$

If we consider the convergence rate, we may write, for $t_2 \geq 0, t_3 \geq 0$, and $t_2 + t_3 > 0$,

$$\frac{1}{N_{1L}} \sum_{\eta_L} (X_{Li} - \mu_{X_L})^{t_2} \, \epsilon_{Li}^{t_3} = \phi_{X_L}^{(t_2)} \cdot \phi_{\epsilon}^{(t_3)} + O_p(N^{-\frac{1}{2}}). \qquad (3.45)$$

By expressions (3.37), (3.39), (3.44), and (3.45), we have

$$\begin{aligned}
&\lambda_{a,b} \\
&= \sum_{t_1=0}^{b} \sum_{t_2=0}^{a+t_1} \sum_{t_3=0}^{b-t_1} \begin{pmatrix} b \\ t_1 \end{pmatrix} \begin{pmatrix} a+t_1 \\ t_2 \end{pmatrix} \begin{pmatrix} b-t_1 \\ t_3 \end{pmatrix} \left(O_p(N^{-\frac{1}{2}}) \right)^{t_1} \cdot \left(O_p(N^{-\frac{1}{2}}) \right)^{a+t_1-t_2} \cdot \\
&\quad \left(O_p(N^{-\frac{1}{2}}) \right)^{b-t_1-t_3} \cdot \left(\phi_{X_L}^{(t_2)} \cdot \phi_{\epsilon}^{(t_3)} + O_p(N^{-\frac{1}{2}}) \right) \\
&= \sum_{t_1=0}^{b} \sum_{(t_2,t_3) \neq (a,b)} \begin{pmatrix} b \\ t_1 \end{pmatrix} \begin{pmatrix} a+t_1 \\ t_2 \end{pmatrix} \begin{pmatrix} b-t_1 \\ t_3 \end{pmatrix} \left(O_p(N^{-\frac{1}{2}}) \right)^{t_1} \cdot \left(O_p(N^{-\frac{1}{2}}) \right)^{a+t_1-t_2} \cdot \\
&\quad \left(O_p(N^{-\frac{1}{2}}) \right)^{b-t_1-t_3} \cdot \left(\phi_{X_L}^{(t_2)} \cdot \phi_{\epsilon}^{(t_3)} + O_p(N^{-\frac{1}{2}}) \right) + \phi_{X_L}^{(a)} \phi_{\epsilon}^{(b)} + O_p(N^{-\frac{1}{2}}). \\
&\qquad (3.46)
\end{aligned}$$

By applying the properties of O_p as expressed in (3.31) and (3.32), we can see that the summation part in expression (3.46) may be replaced by $O_p(N^{-\frac{1}{2}})$ under the condition of the lemma. Hence, we have the result (3.40). The result (3.41) is proved in the same way.

Lemma 3.3 Under the same condition of Lemma 3.2,

$$\tau_{a,b} = \phi_X^{(a)}\phi_\epsilon^{(b)} + O_p(N^{-\frac{1}{2}}),$$

where $\tau_{a,b}$ is in expression (3.12).

Proof. Its proof is in the similar way as for Lemma 3.2, and is thus omitted.

Now we will show weak convergence of $R(d_0)-R(d_1)$ and $\hat{V}(L(d_0)-L(d_1))$.

For $R(d_0) - R(d_1)$:

Using the O_p notation, we can write

$$\hat{P}_L = \frac{N_{1L}}{N_1} = \pi_L + O_p(N^{-\frac{1}{2}}) \text{ and } \hat{P}_R = ((n-1)/N_1)\frac{N_{1L}}{N_1} = \pi_R + O_p(N^{-\frac{1}{2}}). \tag{3.47}$$

By applying the results (3.39) and (3.47) and Lemmas 3.2 and 3.3, we have, from expression (3.20),

$$R(d_0) - R(d_1) = \beta^2\left(\phi_X^{(2)} - \overline{\phi_X^{(2)}}\right) + O_p(N^{-\frac{1}{2}}). \tag{3.48}$$

See expression (3.4) for $\overline{\phi_X^{(2)}}$.

A simple algebra leads us to another expression for $R(d_0) - R(d_1)$ as in

$$R(d_0) - R(d_1) = \pi_L\pi_R\beta^2\left(\mu_{X_R} - \mu_{X_L}\right)^2 + O_p(N^{-\frac{1}{2}}). \tag{3.49}$$

For $\hat{V}(L(d_0) - L(d_1))$:

Under the condition of Theorem 3.1, we know that

$$\begin{aligned} \overline{X}^* - \mu_X &= O_p(N^{-\frac{1}{2}}), && (3.50)\\ \overline{X}^*_R - \mu_{X_R} &= O_p(N^{-\frac{1}{2}}). && (3.51) \end{aligned}$$

In expression (3.28), we apply Lemma 3.3 to the τ terms, Lemma 3.2 to the λ and the ρ terms, Lemma 3.1 to $\overline{e^*}$ terms, equation (3.48) to $(R(d_0)-R(d_1))^2$, the results (3.37), (3.39), (3.47), (3.50) and (3.51) to the other terms in (3.28), and the properties (3.31) and (3.32) to have

$$\begin{aligned}
&\frac{N_1-1}{N_1}\hat{V}(L(d_0)-L(d_1)) \\
&= \beta^4\left\{\phi_X^{(4)} - \overline{\phi_X^{(4)}} - (\sigma_X^4 - (\overline{\sigma_X^2})^2) - 4\pi_L\pi_R(\mu_{X_R}-\mu_{X_L})(\phi_{X_R}^{(3)} - \phi_{X_L}^{(3)})\right\} \\
&\quad +4\beta^2\sigma_\epsilon^2(\sigma_X^2 - \overline{\sigma_X^2}) + O_p(N^{-\frac{1}{2}}) \\
&= \gamma + O_p(N^{-1/2}), \text{ say.}
\end{aligned}$$

Thus we have

$$\hat{V}(N(d_0)-L(d_1)) = \gamma + O_p(N^{-1/2}). \tag{3.52}$$

By applying property (3.33) to the results (3.49) and (3.52), we have the desired result of Theorem 3.1.

APPENDIX 2

Table A.1 Table of the Frequencies of Variable-Selection

N_1	β	σ_ϵ	Using	Thresholds			
				1	1.65	1.96	2.33
25	0.5	1.0	a^*	0.054	0.014	0.010	0.006
			b^*	0.322	0.160	0.118	0.076
		3.0	a	0.004	0.002	0.002	0.002
			b	0.126	0.036	0.024	0.008
	1.0	1.0	a	0.414	0.188	0.128	0.066
			b	0.578	0.354	0.252	0.172
		3.0	a	0.030	0.006	0.002	0.002
			b	0.200	0.084	0.048	0.026
	2.0	1.0	a	0.952	0.902	0.846	0.790
			b	0.936	0.788	0.728	0.586
		3.0	a	0.122	0.036	0.016	0.006
			b	0.416	0.228	0.150	0.098

(Note) *: $a = \sqrt{N_1}IR_1$.
$b = \sqrt{N_1}IR_{dif}$.

(Continued)

N_1	β	σ_ϵ	Using	Thresholds			
				1	1.65	1.96	2.33
50	0.5	1.0	*a*	0.052	0.004	0.002	0.000
			b	0.440	0.234	0.172	0.104
		3.0	*a*	0.002	0.000	0.000	0.000
			b	0.136	0.040	0.018	0.006
	1.0	1.0	*a*	0.620	0.320	0.202	0.124
			b	0.750	0.550	0.446	0.332
		3.0	*a*	0.014	0.000	0.000	0.000
			b	0.294	0.132	0.076	0.030
	2.0	1.0	*a*	0.982	0.966	0.958	0.944
			b	0.976	0.952	0.924	0.860
		3.0	*a*	0.182	0.034	0.016	0.006
			b	0.536	0.330	0.224	0.140
100	0.5	1.0	*a*	0.080	0.004	0.000	0.000
			b	0.548	0.364	0.266	0.172
		3.0	*a*	0.000	0.000	0.000	0.000
			b	0.184	0.078	0.046	0.022
	1.0	1.0	*a*	0.830	0.508	0.368	0.240
			b	0.898	0.762	0.672	0.534
		3.0	*a*	0.008	0.000	0.000	0.000
			b	0.362	0.170	0.110	0.076
	2.0	1.0	*a*	1.000	1.000	1.000	1.000
			b	1.000	1.000	0.998	0.992
		3.0	*a*	0.336	0.070	0.034	0.006
			b	0.750	0.518	0.420	0.306
200	0.5	1.0	*a*	0.164	0.018	0.004	0.000
			b	0.760	0.572	0.452	0.318
		3.0	*a*	0.000	0.000	0.000	0.000
			b	0.256	0.124	0.082	0.044
	1.0	1.0	*a*	0.978	0.878	0.778	0.604
			b	0.990	0.956	0.926	0.884
		3.0	*a*	0.010	0.000	0.000	0.000
			b	0.598	0.354	0.262	0.164
	2.0	1.0	*a*	1.000	1.000	1.000	1.000
			b	1.000	1.000	1.000	1.000
		3.0	*a*	0.602	0.138	0.050	0.012
			b	0.918	0.780	0.692	0.572

(Continued)

N_1	β	σ_ϵ	Using	Thresholds			
				1	1.65	1.96	2.33
300	0.5	1.0	*a*	0.248	0.012	0.002	0.000
			b	0.852	0.674	0.580	0.426
		3.0	*a*	0.000	0.000	0.000	0.000
			b	0.328	0.162	0.096	0.044
	1.0	1.0	*a*	0.998	0.964	0.932	0.862
			b	1.000	0.988	0.984	0.966
		3.0	*a*	0.014	0.000	0.000	0.000
			b	0.664	0.448	0.330	0.230
	2.0	1.0	*a*	1.000	1.000	1.000	1.000
			b	1.000	1.000	1.000	1.000
		3.0	*a*	0.714	0.286	0.132	0.044
			b	0.950	0.836	0.756	0.646

REFERENCES

(1) Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: The MIT Press.

(2) Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees,* Wadsworth International Group, Belmont, California.

(3) Doyle, R. M. (1973). The use of automatic interaction detector and similar search procedures. *Operational Res. Quart.*, **24**, 465-467.

(4) Einhorn, H. (1972). Alchemy in the behavioral sciences. *Pub. Op. Quart.*, **36**, 367-378.

(5) Kim, S. H. (1994). A general property among nested, pruned subtrees of a decision-support tree. *Communications in Statistics Theory and Methods*, **23**, 4, 1227-1238.

(6) Loh, W. Y. and Vanichesetakul, N. (1988). "Tree-structured classification via generalized discriminant analysis," *Journal of American Statistical Association,* **83**, 715-728.

(7) Morgan, J. N. and Messenger, R. C. (1973). *THAID: a sequential search program for the analysis of nominal scale dependent variables.* Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

(8) Morgan, J. N. and Sonquist, J. A. (1963) "Problems in the analysis of survey data, and a proposal," *Journal of American Statistical Association*, **58**, 415-434.

(9) Rao, C. R. (1973). *Linear Statistical Inference and Its Applications.* 2nd ed. New York, New York: John Wiley & Sons.