PSYCHOMETRIKA—VOL. 69, NO. 3, 355–374
SEPTEMBER 2004

THE REGRESSION TRUNK APPROACH TO DISCOVER TREATMENT COVARIATE INTERACTION

ELISE DUSSELDORP
JACQUELINE J. MEULMAN

LEIDEN UNIVERSITY

The regression trunk approach (RTA) is an integration of regression trees and multiple linear regression analysis. In this paper RTA is used to discover treatment covariate interactions, in the regression of one continuous variable on a treatment variable with *multiple* covariates. The performance of RTA is compared to the classical method of forward stepwise regression. The results of two simulation studies, in which the true interactions are modeled as threshold interactions, show that RTA detects the interactions in a higher number of cases (82.3% in the first simulation study, and 52.3% in the second) than stepwise regression (56.5% and 20.5%). In a real data example the final RTA model has a higher cross-validated variance-accounted-for (29.8%) than the stepwise regression model (12.5%). All of these results indicate that RTA is a promising alternative method for demonstrating differential effectiveness of treatments.

Key words: classification and regression trees, multiple linear regression, treatment covariate interaction, differential effectiveness, aptitude treatment interaction.

1. Introduction

Treatment covariate interaction (TCI) is defined as the interaction between a treatment variable **t** (categorical) and a covariate **x** (continuous or categorical, also called predictor). Variable **t** has usually two categories, referring to two experimental conditions (e.g., a treatment condition and a control condition). The presence of a TCI indicates that the treatment "has one effect on one kind of person, and a different effect on another" (Cronbach & Snow, 1977, p. 3). In other words, the effect of variable **t** on an outcome variable **y** differs for persons belonging to different categories of (or having different values on) covariate **x**. We assume that the effectiveness of a treatment has been investigated by a randomized pre-test post-test design, and that the covariate has been measured at the pre-test. When the covariate has not been measured at the pre-test, changes on the covariate might be confounded with the treatment effect. The variable **y**, on which the treatment effectiveness is determined, is usually measured on the pre-test and on the post-test.

The presence of a TCI violates the assumption of homogeneity of the regression slopes in analysis of covariance (ANCOVA; see Stevens, 1992, p. 334). While the aim of ANCOVA is to *control* for individual differences within the experimental conditions, the aim of treatment covariate interaction research is to *discover* individual differences as important sources of differential effectiveness of treatments (Shomon-Salomon & Hannah, 1991). A TCI approach to treatment evaluation research makes it possible to abandon the concept of the average performer and to challenge the idea of a universally best treatment (Dance & Neufield, 1988; Van der Linden, 1981). In the field of education, TCI is known as *aptitude treatment interaction* (Cronbach &

Supported by The Netherlands Organization for Scientific Research (NWO), grant 030-56403 to Jacqueline Meulman for the Pioneer project "Subject-Oriented Multivariate Analysis," and grant 451-02-058 to Elise Dusseldorp for the Veni project "Modeling interaction effects as small trees in regression and classification." We thank Bram Bakker for making the data of his doctoral dissertation available, and David Hand, Jerome Friedman, Bart Jan van Os, Philip Spinhoven, and the anonymous reviewers for their helpful suggestions. Special thanks are due to Lawrence Hubert.

Requests for reprints should be sent to Elise Dusseldorp, Data Theory Group, Faculty of Social and Behavioral Sciences, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands (e-mail: dusseldorp@fsw.leidenuniv.nl)

0033-3123/2004-3/2001-0880-A $00.75/0

Snow, 1977; Snow, 1991) or *attribute treatment interaction*. We prefer the term TCI, which is used in medical research (e.g., Koziol & Wu, 1996), to refer to the more general situation of an interaction between **t** and **x**.

In the psychological literature of the last decade most applications of a TCI approach to evaluation research are found in education (Goska & Ackerman, 1996; McInerney, McInerney, & Marsh, 1997; Rueter, Conger, & Ramisetty-Mikler, 1999) and clinical psychology (Barber & Muenz, 1996; Digiusto & Bird, 1995; Glidden-Tracey & Wagner, 1995; Piper, McCallum, Joyce, Azim, & Ogrodniczuk, 1999). In all these applications the treatment variable **t** consisted of two categories. The TCI was represented in all but one study as a product between **t** and **x**, and the significance of the TCI was tested by a multiple linear or logistic regression analysis (for, respectively, a continuous or a dichotomous outcome variable). This is a commonly used analysis approach to TCI, recommended by Cronbach and Snow (1977). None of the studies included any quadratic terms of covariates. This implies that only linear main effects and linear-by-linear interaction effects were investigated. A majority of the studies included only one or two TCIs in the analysis. In two studies, multiple TCIs were included (6 and 12), and forward or backward stepwise selection was used to discover significant TCIs (Barber & Muenz, 1996; Digiusto & Bird, 1995). The majority of the studies reported significant TCIs. However, none of the studies corrected for multiple testing, or cross-validated the prediction results. The forward stepwise selection procedure (Digiusto & Bird, 1995) did not succeed in tracing significant TCIs. In the study with the backward stepwise selection procedure (Barber & Muenz, 1996), the final model contained 9 predictors, whereas the total sample size was only 84 subjects.

Earlier applications of TCI research in psychology (for an overview, see Dance & Neufield, 1988), commonly used a 2 × 2 analysis of variance (ANOVA) to detect possible TCIs. To apply this method, a continuous covariate was dichotomized into high and low (usually at the sample median). Although the interpretation of the interaction is easier than in multiple regression analysis, this method has been criticized in the literature because of its loss of statistical power (Cronbach & Snow, 1977; Humphreys & Fleishman, 1974).

The methodological literature on TCI research appears to be limited to situations with one or two TCIs, which are specified before the analysis (e.g., Cronbach & Snow, 1977). A problem that is often encountered in treatment effectiveness research, however, is how to discover one or more significant TCIs when the data contain multiple covariates. Snow (1991) noted in this respect, “As more variables are added, the conventional analysis quickly becomes weak, impracticable, and probably uninterpretable” (p. 213). Another problem that we wish to consider is the representation of the interaction effect. Is a linear-by-linear interaction (i.e., the product of **t** and **x**) the only appropriate representation of the TCI?

The present paper focuses on data containing a treatment variable and *multiple* covariates, and on situations where no *a priori* hypotheses (based on previous research) exist about possibly significant TCIs. An alternative analysis strategy, called the regression trunk approach (RTA), is proposed that integrates two existing analysis methods: multiple linear regression analysis and regression trees. Combining linear regression models and regression trees was done before in generalized regression trees (Chauduri, Lo, Loh, & Yang, 1995; Ciampi, 1991). Because many psychologists are probably unfamiliar with (generalized) regression trees, we first explain these methods before describing RTA. Furthermore, the difference between RTA and generalized regression trees will be clarified in the discussion section. The type of interaction discovered by RTA is not a linear-by-linear interaction but a *threshold* interaction. A threshold interaction occurs when a predictor variable has an effect on an outcome variable only above or below certain threshold value(s) on another predictor variable. In the case of a TCI, a threshold interaction means that an experimental condition (i.e., a category of **t**) has an effect on **y** only above or below certain threshold value(s) on **x**. Figures 1 and 2 show examples of a linear-by-linear and a threshold interaction. In Figure 2 we see that a threshold value on **x** divides a regression line for

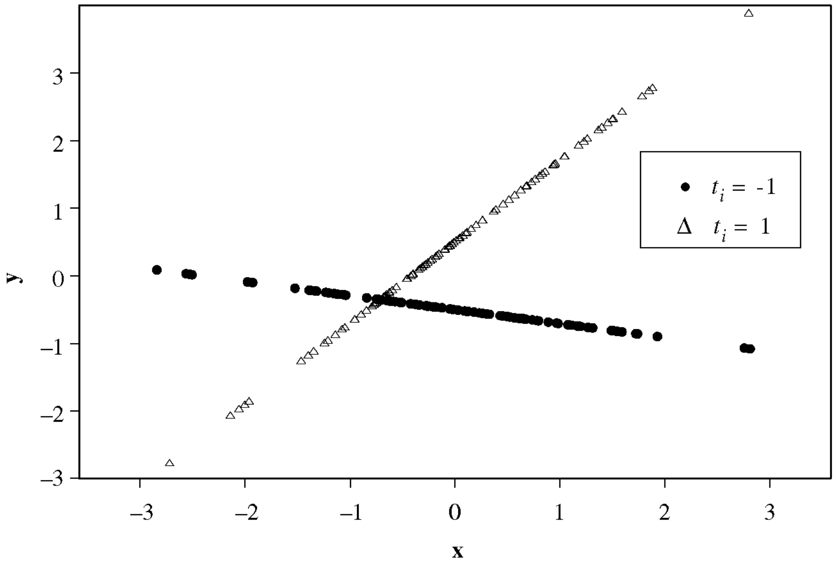

FIGURE 1.
Example of a linear-by-linear interaction between a dichotomous treatment variable **t** and covariate **x**.

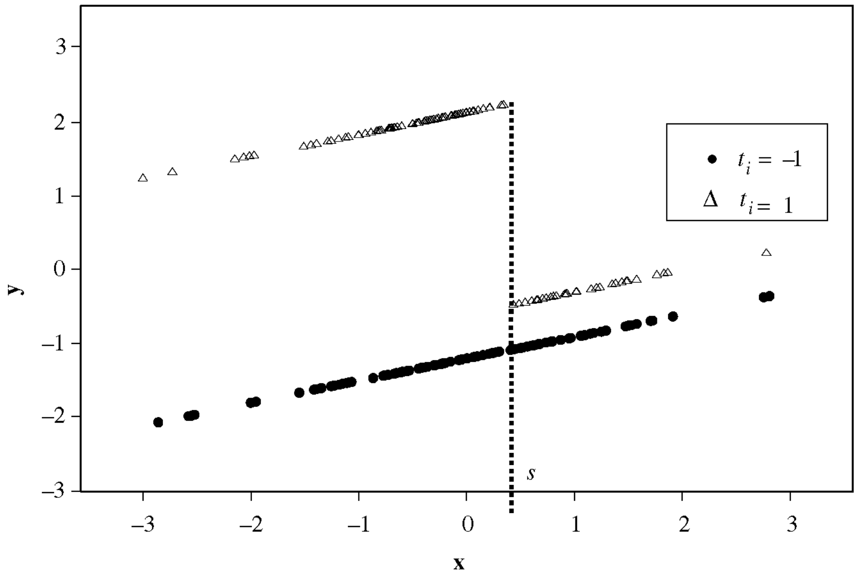

FIGURE 2.
Example of a threshold interaction between a dichotomous treatment variable **t** and covariate **x**. The symbol *s* denotes the threshold value on **x**.

a category of **t** into two parts having different intercepts. In contrast to Figure 1, the slopes of the regression lines remain the same. It will be shown in section 3 that a threshold interaction effect is modeled easily as a small regression tree.

The analyses are limited to first-order TCIs where the covariate **x** is continuous or ordinal, the treatment variable **t** is a dichotomous variable, and the outcome variable **y** is continuous.

Using two simulation studies and a real data example, we compare the performances of RTA and forward stepwise linear regression (STEP) in identifying significant TCIs. In the STEP approach the interactions are represented as linear-by-linear interactions (cross-products), because this is common practice in the TCI literature. In the simulation studies we examine prediction models with first-order threshold interactions. We hypothesize that RTA will be better in identifying this type of interaction than STEP. Besides comparing RTA with STEP, we are interested in the power of RTA; we therefore use different TCI effect sizes and vary the number of covariates in the true model. We also investigate the behavior of the approach when the true model has no significant TCIs. In the real data example we expect that the representation of an interaction effect as a threshold is more appropriate than as a cross-product. A special cross-validation procedure will be applied to estimate the predictive accuracy of the RTA model, and to determine the best size of the tree.

2. Regression Trees and Generalized Regression Trees

Morgan and Sonquist (1963) introduced one of the earliest implementations of regression trees, which they called the *automatic interaction detection* method (AID). AID was developed for the analysis of multiple predictors (covariates) and one continuous outcome variable (**y**). It looks for a partition of the subjects on the basis of a set of predictors into subgroups (called *nodes*) that are as homogeneous as possible with respect to **y**. The starting point of the partitioning is called the root node and contains the total sample of subjects. After the first split, two daughter nodes (i.e., subsets) emerge, each of which plays the role of parent node for the subsequent split. At the end of the partitioning process, the terminal nodes determine the final prediction of the subjects (often the mean of **y** of the subjects in a node). The partitioning criterion consists of a so-called impurity index (or deviance) combined with a stopping rule. For regression trees, the impurity index is defined as the average of the total within node sum of squares (see formula 8.11; Breiman, Friedman, Olshen, & Stone, 1984, p. 230). If all subjects in a node have the same score on **y**, the within node sum of squares is zero. The stopping rule determines at which node to stop splitting; for example, the minimal number of subjects required at a node. The partitioning process of a tree is binary and recursive, which means that the algorithm is repeated: At each split on the basis of the most suitable predictor variable, the subjects are divided into two disjoint and exhaustive subsets that maximize the decrease in impurity. In other words, the best split at a node is that split on a predictor variable that most successfully separates the high values on **y** from the low ones.

Although Sonquist was a sociologist, AID was seldom used in social science research. An important reason for this was the serious problem of overfitting, that is, “The trees grown are not the right size and the estimates are overly optimistic.” (Breiman et al., 1984, p. 232). In terms of the bias/variance trade-off, a small tree has low variance but high bias in the prediction estimates, whereas a large tree has high variance with low bias (Hand, 1997). AID was included by Breiman et al. in a much more extended approach that has become known as classification and regression trees (CART). This approach is suitable for the prediction of both a continuous outcome variable (i.e., regression trees) and a categorical outcome variable (i.e., classification trees). In CART, the problem of overfitting is reduced by a special pruning procedure. Pruning refers to the process where a relatively large tree (fitted with a liberal stopping rule) is reduced to a smaller tree by cutting off terminal nodes (leaf nodes) from the tree to decrease the variance. Pruning is usually done by using cross-validation, and can be seen as an analogous process to backward stepwise regression (Hand, 1997). CART uses a tenfold cross-validation procedure to determine the best pruning parameter, which indicates the best size of a tree. In the tenfold cross-validation procedure, the original data are divided into 10 mutually exclusive sets. One single tree is fitted for nine “training” sets that are pooled together, and the set held out is used as the "test" set. The deviance (e.g., the total within groups sum of squares) is computed from this test set for different values of the pruning parameter. This process is repeated 10 times with each time a different set playing the role of the test set. The deviances obtained for the 10 test sets are accumulated and then divided by the total sample size, resulting in one cross-validated deviance at each pruning parameter (see Breiman et al., p. 234). The minimum value of the cross-validated deviance plus one standard error indicates the best tree size (Breiman et al., p. 237 and p. 308). This cross-validation procedure does not indicate *which* variables to use in the pruned tree; it only determines the best size of the tree.

Independently from the CART framework, Quinlan (1993) introduced C4.5 in the area of machine learning, which is a classification tree method. Both CART and C4.5 are currently popular data mining techniques, included in many data mining software packages (see Gaul & Säuberlich, 1999, for an overview). Important disadvantages of regression trees are that they lack a formal procedure of statistical inference (Clark & Pregibon, 1993) and that they do not separate main effects of predictor variables from interaction effects.

These disadvantages were partly removed in generalized regression trees, independently proposed by Chaudhuri et al. (1995) and Ciampi (1991). The generalized regression trees of Chauduri et al. are constructed by recursively partitioning the data according to the signs of the residuals from a generalized linear model. First, a linear main effects model using all predictors is fitted. Then the residuals are computed and observations with a negative residual are classified into one node and the remainder in the other node. Subsequently, the predictor is selected to split the node, being the one with the largest absolute difference in means for the two nodes (determined by a t-test). This process is repeated, resulting in a tree with a separate linear regression model at each terminal node. In this approach, categorical predictors are converted into ordered variables before fitting a tree.

Ciampi's (1991) generalized regression trees look more like CART. The impurity indices are based on generalized linear models, and the choice of a particular index depends on the assumed distribution of the data. Furthermore, the model allows for adjusting the outcome variable for confounding variables. A Likelihood Ratio Statistic is used to evaluate the decrease in impurity by a split on a predictor variable. A greedy (exhaustive) search approach is used like in AID and CART.

Recently, Loh (2002) presented a new algorithm called GUIDE for generalized, unbiased interaction detection and estimation. It expands on the generalized regression trees of Chaudhuri et al. (1995). GUIDE also divides the residuals of a linear model into negative and positive signs, and uses a Chi-square test instead of a t-test to determine the best split. To perform this test, numerical variables are divided into four groups at the sample quartiles, and a two-way table is constructed with the groups as columns and the residual signs as rows. To test pair-wise interactions between predictors, numerical predictor variables are split into two halves at the sample median, and a two-way table is constructed with combinations of categories from two predictors as columns and the residual signs as rows. Advantages of GUIDE are that categorical predictors do not have to be quantified *a priori*, and that pair-wise interactions are investigated.

3. Regression Trunk Approach to TCI

Multiple regression analysis and regression trees are combined in the following way: Small regression trees are used to *discover* interaction effects, and multiple linear regression is subsequently used to *test* the interaction effects. We have labeled a small tree, a "regression trunk" (Dusseldorp & Meulman, 2001), and the combined method the regression trunk approach (RTA). The innovative element of the approach presented in this paper is that it has been especially designed to detect TCIs. As we mentioned before, a TCI represented by a regression trunk can be regarded as a *threshold* interaction. For example, when a regression trunk uses two predictor variables (**t** and **x**), a category of variable **t** has an effect on variable **y** only above or below (a) certain threshold value(s) on the continuous covariate **x**. The TCIs are not specified *before* the analysis but *during* the analysis. RTA to TCI consists of three subsequent phases of analysis.

In the first phase, a main effects model is estimated by the use of standard linear regression. The outcome variable **y** is regressed on all covariates and the dichotomous treatment variable. Instead of using standard multiple regression one might also use categorical regression (CATREG; Meulman, Heiser, & SPSS, 1999) to incorporate nonlinear relationships in the main effects model (see Dusseldorp & Meulman, 2001). In the second phase of the analysis, two regression trunks are fitted to the residuals from the first phase, one for the subjects in the first category of **t**, and one for the subjects in the second category. These small regression trees are combined into one regression trunk, of which the first split is made on the treatment variable **t**. Growing a regression trunk for each treatment group separately is similar to forcing the first split of the total regression trunk on the treatment variable. This forced split is necessary to discover TCIs, when multiple covariates are available. Why? We have two reasons for using a forced split. First, regression tree analysis uses a *sequential* partitioning algorithm, with each split maximizing the decrease in the impurity index. Because we use the residual ($\mathbf{y}_{res}$) as the outcome variable in the regression trunk analysis, the main effects of both the covariates and the treatment variable are already partialled out. This implies that there is no single best predictor for the first split of the regression trunk. Second, because the number of distinct values of the treatment variable is in general much lower than those of the covariate(s), it is much more likely that a split will be made on a covariate than on the treatment variable. This phenomenon is called variable (or attribute) selection bias: an unrestricted split selection process tends to select variables with more split points (Loh & Shih, 1997; Loh, 2002). To illustrate the impact of these two reasons, we note that for generated data from a model with a large TCI effect, the probability of detecting the TCI was 1 by RTA with a forced split, but only .553 without (see Table 1).

By forcing the first split on **t**, the trunk will find for each category of **t** the covariate (and the threshold value on that covariate, i.e., the split point) that divides the subjects of one treatment group into two subgroups that are as homogeneous as possible (i.e., the total within group sum of squares is as low as possible). The two subgroups will generally be more homogeneous with regard to $\mathbf{y}_{res}$ than the total treatment group, implying that the treatment effect differs for the two subgroups. The difference between the mean score of the subgroups on $\mathbf{y}_{res}$ is an indication of the strength of the differential treatment effectiveness. The approach described in Dusseldorp and Meulman (2001) does not use a forced split.

In the third phase, the total regression trunk is converted into a set of contrast variables. Subjects are assigned values on the contrast variables on the basis of their scores on the predictor variables that determine the splits of the branches leading to the terminal nodes of the trunk. For example, subjects whose scores on the predictor variables lead to terminal node 1 of the trunk, receive a value 1 on contrast variable 1, and a value 0 on the rest of the contrast variables. When a contrast variable consists of the values 1 and 0, we use the term dummy variable. However, other (and more) values may be chosen based on a particular scheme. The contrast variables are added to the main effects regression model as a second block. One has to exclude linearly dependent contrast variables. For example, the number of linearly independent dummy variables equals the total number of terminal nodes minus 2. Because **t** is a linear combination of the dummy variables, one more dummy variable has to be deleted than is commonly done in regression analysis. In other words, the consequence of forcing the first split on **t** is that the effects of the dummy variables are nested within the treatment categories. The choice of the category values of the contrast variables changes the parameter estimates for the individual variables but does not change the overall contribution of the converted regression trunk to the fit of the regression model. The difference in variance-accounted-for between the main effects model and the model including the main effects and the contrast variables, is used to estimate the effect size of the TCI(s) (see next section).

4. Simulation Studies

The performances of RTA and forward stepwise regression analysis (STEP) were compared with respect to (a) identifying one first-order TCI (investigated in Simulation Study 1), and (b) identifying two first-order TCIs (investigated in Simulation Study 2). The data for Study 1 were generated from the model

$$\mathbf{y}_{\text{obs}_1} = \sum_{j=1}^{J} w_j \mathbf{x}_j + .00\mathbf{t} + .49\mathbf{c}_1 + .71\mathbf{e}. \tag{1}$$

The data for Study 2 were generated from the model

$$\mathbf{y}_{\text{obs}_2} = \sum_{j=1}^{J} w_j \mathbf{x}_j + .00\mathbf{t} + .45\mathbf{c}_1 + .45\mathbf{c}_2 + .71\mathbf{e}. \tag{2}$$

The symbols used in (1) and (2) have the same meaning. Covariates are denoted by $\mathbf{x}_j$, with J the total number of covariates. The covariates were standard normally distributed variables. The value of J is a design factor in the simulation studies (see below). A population correlation between all covariates of .30 was chosen because most observed correlations between the covariates reported in the application studies (see the Introduction) were between .21 and .38.

The treatment variable $\mathbf{t}$ was defined as a dichotomous variable, coding two treatment conditions with $t_i = -1$ for the first condition, and $t_i = 1$ for the second; t_i denotes the ith element of $\mathbf{t}$, and $i = 1, \ldots, N$, with N the total number of subjects. The number of subjects in each condition was $.50N$. Subjects were randomly assigned to the two conditions (which is comparable to a randomized experiment). The resulting correlation between $\mathbf{t}$ and each covariate was approximately zero.

The contrast variable $\mathbf{c}_1$ was defined as the following Boolean combination of the covariate $\mathbf{x}_1$ and $\mathbf{t}$:

$$\begin{aligned} &\text{if } (t_i = -1 \text{ and } x_{i1} > s_1), \ \text{then } c_{i1} = 1, \\ &\text{if } (t_i = -1 \text{ and } x_{i1} \leq s_1), \ \text{then } c_{i1} = -1, \\ &\text{if } (t_i = 1), \ \text{then } c_{i1} = 0, \end{aligned} \tag{3}$$

where s_1 symbolizes the value of the covariate $\mathbf{x}_1$ at the split point. Figure 3 shows a graphical representation of this interaction. The contrast variable $\mathbf{c}_2$ in (2) was defined as the following Boolean combination of the covariate $\mathbf{x}_2$ and $\mathbf{t}$:

$$\begin{aligned} &\text{if } (t_i = 1 \text{ and } x_{i2} > s_2), \ \text{then } c_{i2} = 1, \\ &\text{if } (t_i = 1 \text{ and } x_{i2} \leq s_2), \ \text{then } c_{i2} = -1, \\ &\text{if } (t_i = -1), \ \text{then } c_{i2} = 0, \end{aligned} \tag{4}$$

where s_2 symbolizes the value of the covariate $\mathbf{x}_2$ at the split point. Figure 4 shows a graphical representation of this interaction. An advantage of this type of contrast variables to dummy variables is that their intercorrelation is very low.

The weights for the covariates used to generate $\mathbf{y}_{\text{obs}}$ are denoted by w_j. In models (1) and (2) the weights for covariate $\mathbf{x}_1$ and $\mathbf{x}_2$ (w_1 and w_2) and the weight for variable $\mathbf{t}$ were

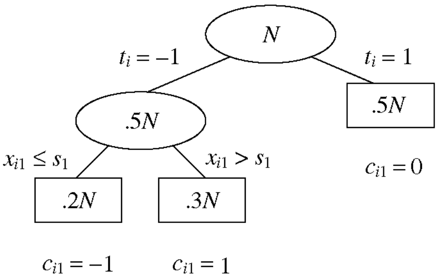

FIGURE 3.
Graphical representation of the true treatment covariate interaction in Simulation Study 1. Ellipses denote interior nodes and rectangles denote terminal nodes. The variables used for the splits and the terminal node sizes are fixed. Below the terminal nodes, the label of the corresponding contrast variable ($\mathbf{c}_1$) is displayed.

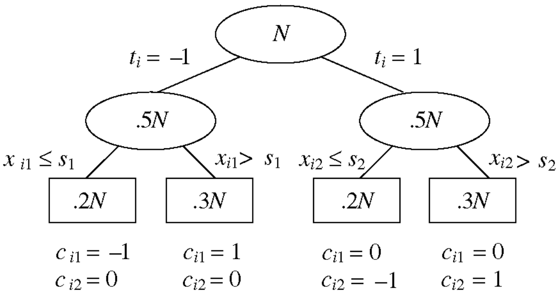

FIGURE 4.
Graphical representation of the true treatment covariate interactions in Simulation Study 2. The variables used for the splits and the terminal node sizes are fixed. Below the terminal nodes, the values of the corresponding contrast variables ($\mathbf{c}_1$ and $\mathbf{c}_2$) are displayed.

fixed at .00. The other weights varied. In both models the contrast variable $\mathbf{c}_1$ has a large weight, implying that the first treatment condition only has an effect for subjects whose score on $\mathbf{x}_1$ is greater than the value s_1. In addition, in model (2) the contrast variable $\mathbf{c}_2$ has a large weight, implying that the second treatment condition only has an effect for subjects whose score on $\mathbf{x}_2$ is greater than the value s_2. In other words, Study 1 contains one TCI (between $\mathbf{t}$ and $\mathbf{x}_1$) and Study 2 includes two TCIs (between $\mathbf{t}$ and $\mathbf{x}_1$, and between $\mathbf{t}$ and $\mathbf{x}_2$).

The vector $\mathbf{e}$ denotes the error, randomly generated with $N(0, 1)$. The weights in (1) and (2) were chosen in such a way that the variance-accounted-for (VAF) by $\mathbf{e}$ is about .50, and that the TCI effect size was within a certain range (as noted below this range was a design factor in the simulation studies). As measure of the TCI effect size we used (also, see Cohen, 1988, p. 410):

$$f^2 = \frac{R^2_{y.A,B} - R^2_{y.A}}{1 - R^2_{y.A,B}}, \tag{5}$$

where $R^2_{y.A}$ denotes the VAF by the main effects only model (i.e., the model including all covariates and $\mathbf{t}$). In Study 1, $R^2_{y.A,B}$ denotes the VAF by the model including main effects and the contrast variable $\mathbf{c}_1$; in Study 2, $R^2_{y.A,B}$ denotes the VAF by the model including main effects and the contrast variables $\mathbf{c}_1$ and $\mathbf{c}_2$. Note that the denominator in (5) is approximately .50, because the error variance proportion is fixed at about .50. The value .50 was chosen because most squared multiple correlation coefficients in the application studies (see the Introduction) didn't exceed this value.

In both studies, we systematically varied two factors in a complete factorial design: the range of the effect size of the TCI (f^2), and the ratio between the number of subjects (N) and the number of predictor variables in the model (M). We chose the following ranges for f^2 in Study 1: .00–.01, .02–.08, .09–.15, .16–.22, .23–.29, and .30–.36; and in Study 2: .00–.02, .03–.12, .13–.22, .23–.32, .33–.42, and .43–.52. In both studies, we chose three levels for the ratio between N and M: 20, 15, and 10. A minimum N/M of 15 is often used as guideline for a reliable regression equation in social science research. The three values of N/M were realized by varying the number of covariates (J) in (1) and (2). The value of N was fixed at 120. Thus, the values of M were 6, 8, and 12. For Study 1, M equals the number of covariates plus 2 (i.e., $\mathbf{t}$ and $\mathbf{c}_1$). This implies that the values of J in (1) were 4, 6, and 10. For Study 2, M equals the number of covariates plus 3 (i.e., $\mathbf{t}$, $\mathbf{c}_1$, and $\mathbf{c}_2$). This implies that the values of J in (2) were 3, 5, and 9. Although the number of covariates varied in the models, the TCI for Study 1 was always defined as an interaction between $\mathbf{t}$ and $\mathbf{x}_1$ (see (3)). Similarly, the TCIs for Study 2 were always defined as an interaction between $\mathbf{t}$ and $\mathbf{x}_1$ and an interaction between $\mathbf{t}$ and $\mathbf{x}_2$ (see (3) and (4)). The population correlations between the covariates were fixed at .30.

One additional design factor was varied in Study 1: the number of subjects with $c_{i1} = 1$ in (3). The split point s_1 in (3) was varied in each analysis in such a way that the size of the group with $c_{i1} = 1$ was 30% or 15% of N (36 or 18, respectively). Because this factor appeared to have no influence (see section 4.1), the size of the groups with $c_{i1} = 1$ and $c_{i2} = 1$ in (4) was fixed at $.15N$ for Study 2.

In Study 1, for each of the $6 \times 3 \times 2 = 36$ combinations, 1000 random samples were drawn, resulting in 36,000 analyses (for RTA and STEP separately). In Study 2 and for each of the $6 \times 3 = 18$ combinations, 1000 random samples were also drawn, resulting in 18,000 analyses. In each random sample we investigated whether RTA or STEP identified the true TCI(s). Following the RTA procedure, TCIs were defined through the use of contrast variables that represented the categorical trunk variable. In the first phase, a main effects model was fitted on $\mathbf{y}_{\text{obs}}$ using all the covariates and the treatment variable $\mathbf{t}$. Second, a tree with two terminal nodes was fitted for each treatment condition separately (because we limit our analysis to first-order interactions). As a stopping rule, we used a minimum terminal node size of $.10N$. If the fitted tree was larger than two nodes, the tree was pruned back to two terminal nodes because we limited the analyses to first-order interactions. The two small trees were combined into one regression trunk, with a first split on $\mathbf{t}$. Figure 5 shows a graphical representation of this trunk. Third, the regression trunk was

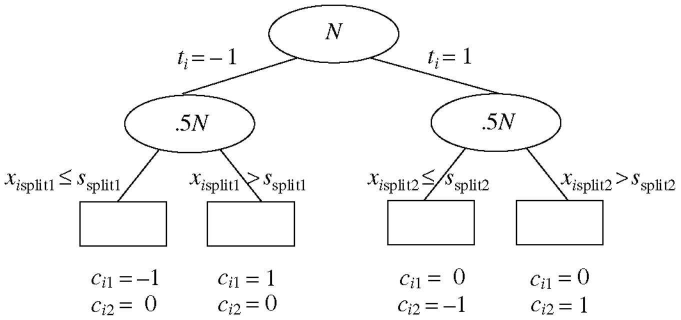
N
$t_i = -1$
$t_i = 1$
.5N
.5N
$x_{i\text{split1}} \le s_{\text{split1}}$
$x_{i\text{split1}} > s_{\text{split1}}$
$x_{i\text{split2}} \le s_{\text{split2}}$
$x_{i\text{split2}} > s_{\text{split2}}$
$c_{i1} = -1$
$c_{i2} = 0$
$c_{i1} = 1$
$c_{i2} = 0$
$c_{i1} = 0$
$c_{i2} = -1$
$c_{i1} = 0$
$c_{i2} = 1$

FIGURE 5.
The regression trunk fitted by RTA in both simulation studies. The covariates used for the splits ($\mathbf{x}_{\text{split1}}$ and $\mathbf{x}_{\text{split2}}$) and the split points ($s_{\text{split1}}$ and s_{split2}) may vary. Below the terminal nodes, the values of the corresponding contrast variables ($\mathbf{c}_1$, and $\mathbf{c}_2$) are displayed.

converted into a set of contrast variables, defined in the following way:

$$\begin{aligned}
&\text{if } (t_i = -1 \text{ and } x_{i\,\text{split1}} \le s_{\text{split1}}), \text{ then } c_{i1} = -1,\\
&\text{if } (t_i = -1 \text{ and } x_{i\,\text{split1}} > s_{\text{split1}}), \text{ then } c_{i1} = 1, \text{ else } c_{i1} = 0;\\
&\text{if } (t_i = 1 \text{ and } x_{i\,\text{split2}} \le s_{\text{split2}}), \text{ then } c_{i2} = -1,\\
&\text{if } (t_i = 1 \text{ and } x_{i\,\text{split2}} > s_{\text{split2}}), \text{ then } c_{i2} = 1, \text{ else } c_{i2} = 0,
\end{aligned} \tag{6}$$

where $x_{i\,\text{split1}}$ denotes the covariate used to split the subjects in the first treatment condition, and s_{split1} denotes the split point on this covariate. Consequently, $x_{i\,\text{split2}}$ denotes the covariate used to split the subjects in the second treatment condition, and s_{split2} denotes the split point on this covariate.

To test the TCIs both contrast variables were added to the main effects regression model of the first phase. RTA was said to detect the true TCI in Study 1 (see (1)), if covariate $\mathbf{x}_1$ was used to split the subjects in treatment condition $k = 1$, and the regression weight of the contrast variable $\mathbf{c}_1$ was significant (with $p \le .01$). RTA was said to detect the true TCIs in Study 2 (see (2)), if covariate $\mathbf{x}_1$ was used to split the subjects in the first treatment condition, covariate $\mathbf{x}_2$ was used to split the subjects in the second treatment condition, and the regression weights of variables $\mathbf{c}_1$ and $\mathbf{c}_2$ were both significant (with $p \le .01$).

Following the procedure of STEP, all possible first-order TCIs were specified before the analysis (by computing the cross-product between **t** and each covariate). For example, when four covariates were used in the true model (see (1)), four TCIs were computed. The forward stepwise selection procedure was used to select significant TCIs, in addition to the main effects model. The STEP procedure was said to detect the true TCI in Study 1, if the product between **t** and $\mathbf{x}_1$ was selected in the final model and its regression weight was significant (with $p \le .01$). The STEP procedure was said to detect the true TCIs in Study 2, if both the products between **t** and $\mathbf{x}_1$, and between **t** and $\mathbf{x}_2$ were selected in the final model, and the regression weights of these products were significant (with $p \le .01$).

The performance of both RTA and STEP was expressed as the number of times the method detected the true TCI(s). Analyses of variance (ANOVA) were performed for the simulation studies separately, with the detection of the true TCI (1 = yes, 0 = no) as the dependent variable, and the corresponding design factors plus a dichotomous variable indicating the method applied (RTA or STEP) as independent variables. We were interested in which factors influenced the probability of detection of a TCI. Therefore, we omitted the lowest effect size ranges (.00–.01 for Study 1, and .00–.02 for Study 2) because these ranges implied that the simulated model did not contain a TCI. ANOVA assumes a normally distributed dependent variable. Although in this study the dependent variable was binomially distributed, we assumed that this distribution was approximated by a normal distribution (using the guideline that $np > 5$ and $n(1-p) > 5$, where n denotes the total number of simulations and p denotes the overall detection probability of the true TCI). Given the large value of n (60,000 in Study 1 and 30,000 in Study 2) the approximation was reasonable (Rice, 1995, p. 172).

4.1. Results of Simulation Study 1

The mean proportion of detection of the TCI (between **t** and $\mathbf{x}_1$) in Study 1 is displayed in Table 1, for RTA and STEP separately. RTA was superior to STEP for each cell in the design. On average, RTA detected 82.3% of the TCIs (omitting the lowest effect size range), and STEP detected 56.5%. We interpreted the probabilities given in the column of the lowest effect size range as Type I errors. The Type I errors of RTA were higher than those of STEP, but they were reasonable: between .023 and .060. The probabilities in the other columns in Table 1 were

TABLE 1.
Results of Simulation Study 1: Probability of detection of a treatment covariate interaction between variable **t** and $\mathbf{x}_1$ by the regression trunk approach (RTA) or by forward stepwise regression (STEP).

	Effect Size Range					
Group Size	.00–.01	.02–.08	.09–.15	.16–.22	.23–.29	.30–.36
			RTA			
.30N						
$N/M = 20$	.040	.288	.859	.997	1.000	1.000[a]
$N/M = 15$	.029	.278	.836	.994	.998	1.000
$N/M = 10$	.025	.247	.799	.980	.997	1.000
.15N						
$N/M = 20$	.060	.324	.887	.995	1.000	1.000
$N/M = 15$	.044	.303	.872	.996	.999	1.000
$N/M = 10$	.023	.230	.818	.992	.997	1.000
			STEP			
.30N						
$N/M = 20$	.006	.125	.472	.711	.840	.918
$N/M = 15$	.010	.137	.436	.679	.809	.905
$N/M = 10$	.005	.132	.408	.622	.781	.879
.15N						
$N/M = 20$	.010	.119	.394	.616	.764	.864
$N/M = 15$	.004	.122	.384	.599	.754	.860
$N/M = 10$	.017	.137	.356	.583	.721	.813

Note. Number of simulations per cell equals 1000. N/M is the ratio between sample size N and number of predictor variables M in the model.
[a] The probability in this cell was .553 when we used RTA without a forced split on **t**.

considered as indications of the power of the method. The power of RTA was reasonable (i.e., $\geq$ .80) for effect sizes greater than or equal to .09. The power of STEP was reasonable for effect sizes greater than or equal to .30, and for effect sizes between .23–.29 if the group size was .30N and the N/M ratio was 15 or higher. In the case of one predictor, Cohen (1988) gives the operational definitions of "small," "medium," and "large" values for f^2 of .01, .10, and .33, respectively. Taking these values into account, the results indicate that only RTA is able to detect a TCI (modeled as a threshold interaction) with a medium effect size. Both methods detect a TCI with a large effect size.

Because the sample size was very large (60,000), most effects were significant (with $p <$ .001) in the ANOVA analysis. Therefore we inspected the profile plots and the effect size (η^2) to determine which effects were important. The main effects of Method and Effect Size Range were important, and the first-order interaction effect of Method × Effect Size Range. None of the second- and third-order interactions were important. The small main effect of Group Size (size of the group of subjects with $c_{i1} = 1$, in (3)) indicated that this factor had no influence on the probability of detection. The main effect of Method confirmed that RTA was better in detecting the true TCI than STEP. The interpretation of the main effect of Effect Size Range was straightforward: the smaller the effect size of the TCI, the smaller the probability of detection. The first-order effect Method × Effect Size Range revealed that the detection rate of STEP was decreasing regularly with the decrease of the effect size range, while the detection rate of RTA showed a large drop from effect size ranges above .09 to ranges below .09.

We also compared the split point s_{split1} found by RTA with the simulated split point (s_1), see Figure 5 and Figure 3. Remember that s_1 was chosen in such a way that the group size (i.e., the terminal node size of the subjects with $c_{i1} = 1$) was fixed at .30N or .15N. As a consequence, s_1 was different in each sample. We looked at the results of the RTA analyses for the effect size range .30–.36, both group sizes and all three N/M-ratios (6,000 samples in total). RTA detected the true TCI in all of these samples (see the last column in Table 1). In 61% of the samples, the group size found by RTA was equal to the true group size (36 or 18), indicating that s_{split1} was very close (or equal) to s_1. In 97%, the group size was equal to the true group size plus or minus 3. In 94%, s_{split1} was within one standard deviation of s_1.

4.2. Results of Simulation Study 2

The mean proportion of detection of the TCIs (between **t** and $\mathbf{x}_1$, and between **t** and $\mathbf{x}_2$) in Study 2 is displayed in Table 2, for RTA and STEP separately. On average, RTA detected 52.3% of the true TCIs (omitting the lowest effect size range), and STEP detected 20.5%. These percentages are much lower than those in Study 1. Although the model increased only a bit in complexity (two TCIs instead of one TCI), the probability of detection of the TCIs decreased a lot.

The probabilities in the column of the lowest effect size range indicated that the Type I errors were very low for both methods. Inspecting the other columns in Table 2, we concluded that the power of RTA was reasonable for effect sizes higher or equal to .33, given that the N/M-ratio equals 20. The power of STEP was low. None of the probabilities were greater or equal than .80. In the case of more than one predictor, the operational definitions of "small," "medium," and "large" values of f^2 are .02, .15, and .35, respectively (Cohen, 1988, pp. 413–414). Taking these values into account, the results indicate that RTA is able to detect two TCIs when the common effect size is large and the N/M-ratio is 20.

Again, the main effects of Method and Effect Size Range were important in the ANOVA analysis. None of the first- or second-order interactions were important. The interpretations of the main effects are similar to those in Study 2.

TABLE 2.

Results of Simulation Study 2: Probability of detection of two treatment covariate interactions, between variable **t** and $\mathbf{x}_1$ and between variable **t** and $\mathbf{x}_2$, by the regression trunk approach (RTA) or by forward stepwise regression (STEP).

	Effect Size Range					
Group Size	.00–.02	.03–.12	.13–.22	.23–.32	.33–.42	.43–.52
			RTA			
.15N						
$N/M = 20$	.005	.109	.476	.680	.795	.847
$N/M = 15$	.003	.089	.365	.585	.765	.843
$N/M = 10$	.000	.033	.255	.520	.696	.786
			STEP			
.15N						
$N/M = 20$	.000	.011	.070	.215	.356	.500
$N/M = 15$	.000	.016	.072	.180	.330	.415
$N/M = 10$	.000	.013	.054	.164	.269	.408

Note. Number of simulations per cell equals 1000. N/M is the ratio between sample size N and number of predictor variables M in the model.

5. Real Data Example: The Treatment of Panic Disorder Study

In a twelve-week, pre-test post-test placebo-controlled study, 131 patients with panic disorder were randomly assigned to four conditions: paroxetine ($n = 32$), clomipramine ($n = 32$), cognitive therapy ($n = 35$), or placebo ($n = 32$) (Bakker, van Dyck, Spinhoven, & van Balkom, 1999). Both paroxetine and clomipramine are antidepressants. In addition, 31 patients who had refused randomized treatment received cognitive therapy (CT) by preference (Bakker, Spinhoven, van Balkom, Vleugel, & van Dyck, 2000).

Several outcome measures were used to estimate the efficacy of the treatment conditions. We limited our analysis to one outcome measure: the Hamilton Rating Scale for Anxiety (HAMA). Patients treated with an antidepressant (either paroxetine or clomipramine) showed a significantly higher decrease in anxiety than patients receiving the placebo (Bakker et al., 1999). Furthermore, patients treated with an antidepressant showed a significantly higher decrease in anxiety than those treated with cognitive therapy (in the intention-to-treat sample). No significant difference in anxiety reduction was found between the two groups of patients treated with a different antidepressant, either paroxetine or clomipramine. In addition, no significant difference in anxiety reduction was found between the two groups of patients treated with CT, either by chance allocation or by their own preference (Bakker et al., 2000). Because of these findings we could merge the patients treated with antidepressants (AD) into one group ($n = 64$) and the patients treated with CT into another group ($n = 66$).

In the present study the following research question was investigated: Which patients benefit from AD and which from CT? We were especially interested in the possible influence of locus of control orientation (a multidimensional concept indicating a patient's beliefs about the controllability of the panic disorder) on differential effectiveness of the two treatment groups. As Bakker, Spinhoven, van der Does, van Balkom, and van Dyck (2002) concluded, "another interesting question that needs further research is the possibility that locus of control orientation constitutes a prescriptive indicator, that is, a variable predicting differential response to one versus another treatment" (p. 61).

In addition, the following methodological question was studied: Which approach is more appropriate to discover important treatment covariate interactions, the regression trunk approach (RTA) or forward stepwise regression (STEP)? From a clinical psychological point of view, we did not expect that the treatment groups had a reverse linear relationship with locus of control orientation (as would be in a disordinal product interaction when the regression lines of each group intersect). We expected, however, that patients' internal control orientation influenced the effect of cognitive therapy. As Thompson and Wierson (2000) stated, "given the pervasive importance of personal judgments of control, it is not surprising that they can play a central role in the process and outcomes of psychotherapy." In the literature, however, no hypotheses appear to be available about the nature of the interaction. To assess the interaction by RTA (and to represent it as a small tree) seems natural, because RTA can identify the change point in baseline internal control orientation, above which patients appear to be more responsive to cognitive therapy.

The following variables were used as covariates in the analyses, all measured at the pre-test: Age, Gender, Duration of Panic Attack, Severity of Panic Attack, Agoraphobia, Depression, Internal Locus of Control, Chance Locus of Control, Medication Locus of Control, and Therapist Locus of Control. All locus of control scales were measured by the Multidimensional Anxiety Locus of Control scale (MALC; Bakker et al., 2002). We recoded the four locus of control scales described by Bakker et al. (2002) in such a way that a higher score on a scale indicated a higher specific locus of control. In addition, Pre-test Anxiety was used as a covariate, and Post-test Anxiety was the outcome variable ($\mathbf{y}$; both variables were measured by the HAMA). Only subjects with no missing values on these variables were included in the analyses, resulting in a sample size of $N = 102$ (AD: $n = 55$; CT: $n = 47$).

In the first step of the RTA analysis, a main effects model was estimated. In the second step, a regression trunk was fitted to the residuals of the main effects model for the CT and AD patients separately, using a stopping rule of 10% of the number of subjects (N). These regression trunks were merged into one regression trunk. In the third step, the regression trunk was converted into a set of dummy variables, which were added in the last step to the main effects model. To perform the STEP analysis, all possible treatment covariate interactions were specified *a priori* (11 in total). The STEP procedure was used to select significant TCIs, in addition to the main effects model. Tenfold cross-validation (see the Introduction) was used to estimate the predictive accuracy of the models (expressed as cross-validated VAF). For both RTA and STEP, and in each training set the *complete* procedure was performed to estimate the parameters of the training model (including the estimation of the main effects model and the selection of important TCIs). This training model was then used to predict $\mathbf{y}$ for the test set.

5.1. Results of the Real Data Example

The fitted regression trunk by RTA for the total sample consisted of six terminal nodes (Figure 6). From this regression trunk several dummy variables were extracted (see Figure 6). We computed dummy variables using the nodes of the second layer of the regression trunk ($\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, and $\mathbf{d}_4$), and dummy variables using the nodes of the third layer ($\mathbf{d}_5$, $\mathbf{d}_6$, $\mathbf{d}_7$, and $\mathbf{d}_8$). For example, patients were assigned a value 1 to $\mathbf{d}_1$ if they were in the CT condition and their score on the internal locus of control scale was lower than 3.25. The cross-validated VAF of several RTA models was estimated; these models included different sets of dummy variables representing the TCI(s). In this way, we could determine the best size of the regression trunk, and also the best representation of the TCI(s).

Table 3 displays the different RTA models and their cross-validated VAF. This table shows that the RTA model including $\mathbf{d}_5$, and $\mathbf{d}_6$ had the highest cross-validated VAF (29.8%). This

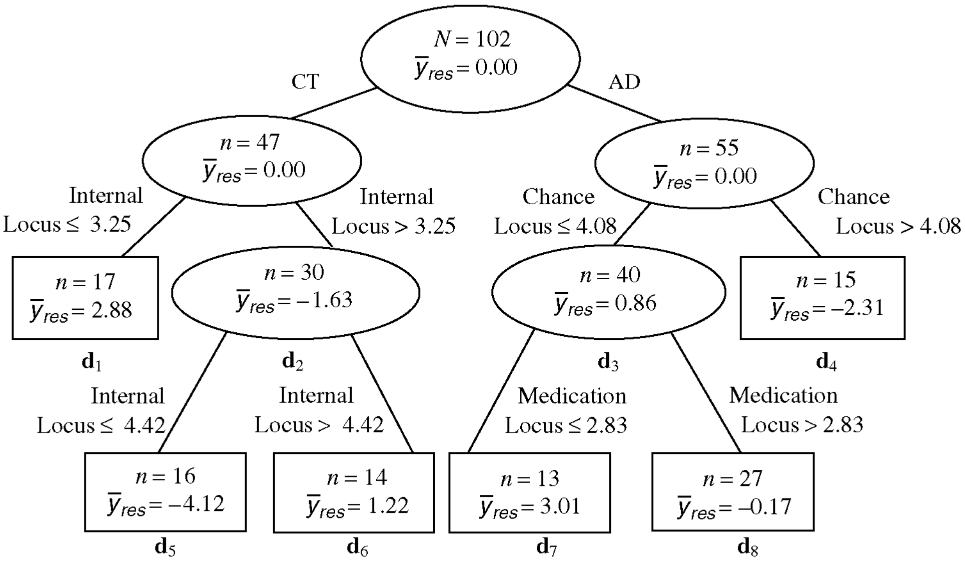

FIGURE 6.

Regression trunk for the total sample ($N = 102$) of panic disorder patients. The outcome variable is the residual of the linear main effects model ($\mathbf{y}_{res}$). The number of patients and their mean value on the outcome variable are given in each node. CT = cognitive therapy; AD = antidepressants. Below the nodes, the labels of the corresponding dummy variables ($\mathbf{d}_1, \ldots, \mathbf{d}_8$) are displayed.

TABLE 3.
Predictive accuracy of the models estimated on the panic disorder data, with Post-test Anxiety as outcome variable. Results of the tenfold cross-validation procedure for the main effects model, and for the models with different representation of the treatment covariate interaction(s). The meaning of the dummy variables used in the RTA models are shown in Figure 6.

Method	Model	VAF	VAF-CV
Linear Regression	Main effects (Pre-test Anxiety, Treatment, and 10 covariates)	36.5%	20.9%
Regression Trunk	Main effects, $\mathbf{d}_2$ and $\mathbf{d}_4$	46.2%	25.1%
Approach (RTA)	Main effects, $\mathbf{d}_5$ and $\mathbf{d}_6$	46.3%	29.8%
	Main effects, $\mathbf{d}_4$, $\mathbf{d}_5$, and $\mathbf{d}_6$	49.1%	26.7%
	Main effects, $\mathbf{d}_5$, $\mathbf{d}_6$, $\mathbf{d}_7$, and $\mathbf{d}_8$	51.3%	28.2%
Forward Stepwise Regression	Main effects, Treatment × Chance Locus of Control	38.1%	12.5%

Notes. VAF = variance-accounted-for in original sample ($N = 102$); VAF-CV = variance-accounted-for determined by tenfold cross-validation. This value was computed as the square of the correlation between the original outcome variable and the predicted value.

implies that the best size of the regression trunk is four terminal nodes (see Figure 7). In the tenfold cross-validation, one part of the regression trunk (the left part in Figure 6) always came out the same not only with regard to the covariate used for the splits (Internal Locus of Control) but also with regard to the split points. In other words, the dummy variables $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_5$, and $\mathbf{d}_6$, always referred to the same combination of the treatment variable and Internal Locus of Control in all training models. The other part of the ten regression trunks (the right part in Figure 6) varied over the tenfold cross-validation. Only four times was Chance Locus of Control used in the second split and Medication Locus of Control in the third. Consequently, the dummy variables $\mathbf{d}_3$, $\mathbf{d}_4$, $\mathbf{d}_7$, and $\mathbf{d}_8$ referred to different combinations of the treatment variable and the covariates in the ten training models.

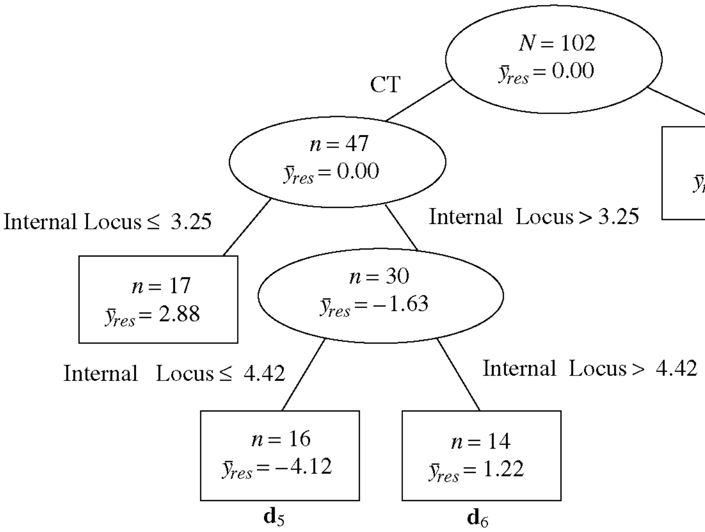

FIGURE 7.
Best size of the regression trunk for the total sample ($N = 102$) of panic disorder patients, determined by tenfold cross-validation. CT = cognitive therapy; AD = antidepressants.

In the final model estimated by STEP on the original sample, the first-order interaction Treatment × Chance Locus of Control was selected. The cross-validated VAF of this model is lower than the main effects model (12.5% versus 20.9%, see Table 3). The models estimated in the ten training sets were very different. Four times the first-order interaction Treatment × Chance Locus of Control was selected, and four times none of the first-order interactions.

The cross-validated VAF for each RTA model in Table 3 is higher than for the STEP model. Overall, the RTA model with $\mathbf{d}_5$ and $\mathbf{d}_6$ has the best cross-validated fit. The estimated regression coefficients for this model are shown in Table 4. Four main effects and one interaction effect ($\mathbf{d}_5$) are significant. Figure 8 depicts the threshold interaction between Treatment and Internal Locus of Control.

The significant main effect of Treatment indicates that, on average, patients treated by antidepressants experience a lower level of Post-test Anxiety than patients treated by cognitive therapy. However, the treatment covariate interaction effect reveals that patients treated by cognitive therapy with a medium Internal Locus of Control (between 3.25 and 4.42) also experience a lower level of Post-test Anxiety. This means that cognitive therapy was especially effective for patients with a *medium* Internal Locus of Control. The change in VAF from the main effects only model to the main effects model plus dummy variable $\mathbf{d}_5$ was 9%; the corresponding effect size of the addition of $\mathbf{d}_5$ (f^2) was 0.17.

6. Discussion

This study proposes a new tool, called the regression trunk approach (RTA), for modeling a specific type of interaction in prediction problems, namely threshold interactions. An important virtue of RTA is that regression trees are fitted to the *residuals* of a linear main effects model. A regression tree is fitted separately for the subsamples of a treatment variable to trace treatment covariate interaction effects. By first removing the linear main effects, RTA efficiently finds these interaction effects. A disadvantage of ordinary regression trees is "not utilizing continuous vari-

TABLE 4.
Parameter estimates of the best model on the panic disorder data, fitted by the regression trunk approach. Lower and upper bound of the 95% confidence interval are displayed.

Predictor Variable	*b*	Lower	Upper
Intercept	2.95	−8.27	14.17
Pre-test Anxiety	0.03	−0.17	0.23
Treatment (CT versus AD)	3.79**	2.12	5.46
Age	0.08	−0.08	0.24
Gender	0.13	−2.36	2.61
Duration of Panic Attack	0.22*	0.00	0.45
Severity of Panic Attack	0.65	−1.90	3.21
Agoraphobia	0.21**	0.06	0.35
Depression	−0.02	−0.22	0.19
Internal Locus of Control	−0.52	−2.17	1.14
Chance Locus of Control	0.70	−0.68	2.09
Medication Locus of Control	1.57**	0.61	2.53
Therapist Locus of Control	−0.99	−2.24	0.25
$\mathbf{d}_5$ (Treatment = CT, and 3.25 < Internal locus of control < 4.42)	−8.13**	−12.40	−3.85
$\mathbf{d}_6$ (Treatment = CT, and Internal locus of control > 4.42)	−3.04	−8.51	2.42

Note. *b* = regression coefficient. CT = cognitive therapy, AD = antidepressants; $^*p < .05$, $^{**}p < .01$

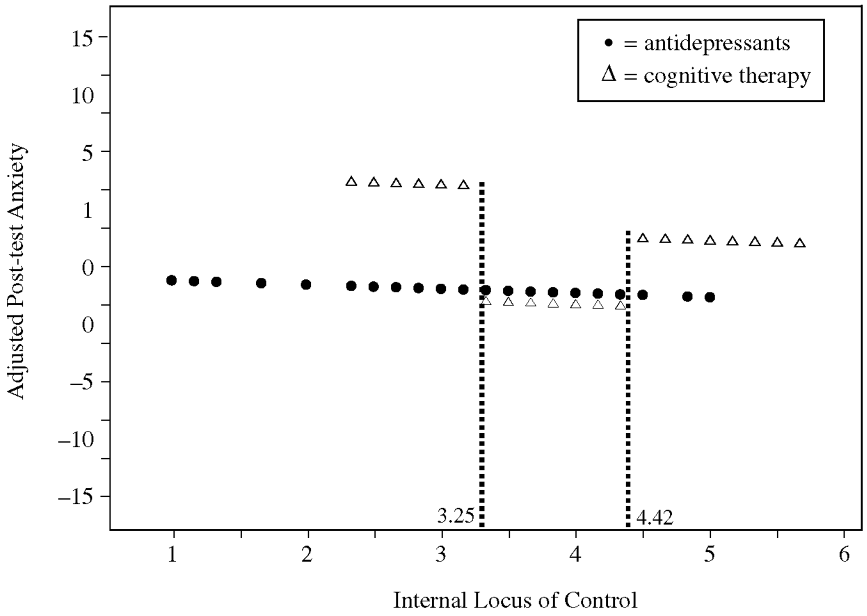

FIGURE 8.
Regression lines of Adjusted Post-test Anxiety on Internal Locus of Control for the antidepressants condition and for the cognitive therapy condition. The regression line for the cognitive therapy condition is divided into three parts, determined by the threshold values on Internal Locus of Control. Post-test Anxiety is adjusted for the main effects of the other covariates (i.e., Pre-test Anxiety, Age, Gender, Duration of Panic Attack, Severity of Panic Attack, Agoraphobia, Depression, Chance Locus of Control, Medication Locus of Control and Therapist Locus of Control).

ables effectively" (Harrell, 2001, p. 27); for example, multiple splits on the same continuous variable are needed to model a linear main effect. RTA overcomes this shortcoming by integrating regression trees into a linear regression model. A second virtue of RTA is that, after growing the trees for the subsamples, the trees are converted into dummy variables and added to the linear main effects model. In this way, we can compare the cross-validated fit of the regression trunk model with the cross-validated fit of the main effects model, and determine what the additive variance-accounted-for is by the dummy variables. We correct for the risk of overfitting by the use of cross-validation on the whole process.

Both the simulation studies and the real data example show results in favor of RTA, compared to forward stepwise regression analysis (STEP), a commonly suggested approach to select the most important (interaction) predictors in a linear regression analysis. We want to draw attention to two points concerning the simulation studies. First, the *true* treatment covariate interactions were modeled as threshold interactions. In this case, STEP does not perform very well, and the results show that RTA is more appropriate. If the true interaction effects were modeled as products, STEP would probably have performed better. Second, the STEP approach included only linear-by-linear interactions. In future research we want to compare RTA with polynomial regression analysis that also models, for example, curvilinear-by-linear interactions (using a quadratic term of the covariate).

In real-life situations it depends on the true type of interaction if modeling the interaction as a threshold has to be preferred over modeling it as a product. For an interaction between two categorical variables (especially when the variables have more than two categories), however, modeling the interaction as a threshold is much more efficient. If one cell of the full cross tabulation (of the two categorical variables) is responsible for the interaction, RTA will usually separate that cell from the other cells in the regression trunk (in two splits only). In contrast, for linear regression all possible combinations of dummy variables (representing the categories of the variables) have to be entered in the analysis as products.

The threshold representation of interaction in RTA is comparable with the representation of a TCI in the 2×2 ANOVA approach (with a split at the median), because in both RTA and ANOVA the continuous covariate is dichotomized. An important difference, however, is that in RTA the covariate is dichotomized optimally *during* the analysis, while in ANOVA the dichotomization is specified *beforehand.* Because RTA finds the optimal cut-off point, loss of power is less an issue than in the ANOVA approach. (Note that in RTA a covariate can also be discretized into more than two categories, as is shown in the real data example.) The simulation studies showed that the price paid by RTA for increased power was small: Type I errors remained below or equal to .06. It appeared to be necessary, however, to use $p \leq .01$ to determine the significance of the TCI-effect in the final RTA model. If we used $p \leq .05$, the Type I errors were larger.

The threshold representation of interaction in RTA is highly comparable with the representation of interaction in generalized regression trees. Several differences between these methods have to be mentioned, however. In RTA, the residuals of the linear main effects model are computed once and a tree is fitted for these residuals, resulting in one final linear model instead of several at the terminal nodes of a tree (as in Chaudhuri et al., 1995). In RTA, categorical predictors do not have to be quantified before using them (as in Chaudhuri et al.). Furthermore, RTA always uses the *same* predictors in the linear main effects model as in the tree-modeling process, and in this way, we disentangle main effects from interaction effects. This implies that in RTA the interaction effects of predictors (modeled as a tree) are estimated *above* their linear main effects. In Ciampi's trees (1991), on the other hand, some predictors may act as confounders (for which $\mathbf{y}$ is linearly adjusted), and *other* predictors may be used in the tree. An important difference between GUIDE (Loh, 2002) and RTA is that the latter finds the optimal threshold(s) on a continuous predictor variable that interacts with a treatment variable, whereas GUIDE divides a continuous predictor variable at the sample median (a priori) to detect an interaction with, for example, a treatment variable. An *a priori* dichotomization induces loss of power (like in the 2×2 ANOVA approach). Finally, RTA uses a forced first split on the treatment variable, which is not done in generalized regression trees and GUIDE. In this way, RTA reduces a general search for interactions between predictors to a search for TCIs only.

In the simulation studies we pruned the regression trunks back to two terminal nodes. In this way, we assumed that the true model included first-order interaction effects only, and no effects for higher-order interactions. The same assumption was used in the STEP procedure. We only computed cross-products of the treatment variable with one covariate (not with two covariates or more). Future research has to be done to develop an efficient pruning procedure for RTA. We applied the pruning procedure of CART to determine the best size of the regression trunks in the simulation studies but this procedure appeared to be too conservative.

In the real data example, all RTA models show a higher cross-validated fit than the STEP model. Furthermore, RTA discovered a threshold interaction that is easy to interpret. In this way, RTA maintains the easy interpretability of the 2×2 ANOVA interactions. The threshold interaction found by RTA incorporates a nonlinear effect of the covariate Internal Locus of Control (by selecting only the subjects with a medium score). As suggested by a reviewer, we investigated whether this type of interaction effect could have been discovered when modeling it as a product between the treatment variable and a quadratic term of Internal Locus of Control (LOC), that is, as a curvilinear-by-linear interaction. We fitted a regression model on the panic disorder data including all linear main effects, plus a quadratic term of LOC (LOC^2), plus the product of Treatment and LOC, plus the product of Treatment and LOC^2. The VAF by this model was 38.2% for the total sample; this is only 1.7% more than the VAF by the main effects model. The regression coefficients of the two product terms were not significant, implying that the treatment covariate interaction effect could not be represented as a curvilinear-by-linear interaction.

As we expected (see section 5), internal locus of control influenced the effect of cognitive therapy. The finding that especially patients with a medium internal control orientation profit from cognitive therapy has a plausible clinical psychological explanation. Patients having a *low* internal control orientation are not convinced that they can decrease their anxiety level by altering their cognitions (one of the goals of cognitive therapy). Patients having a *high* control orientation, on the other hand, are convinced that they can handle a panic attack themselves, but they have a risk of overestimating their own capability or they blame themselves too much for having panic attacks. Patients with a *medium* internal control orientation apparently have a realistic view of their own capability of managing their panic attacks, which led in combination with cognitive therapy (appealing to someone's self management skills) to a lower level of post-test anxiety. It should be noted, however, that this is an *a posteriori* explanation that needs to be confirmed in future clinical research.

The results of the present study of the panic disorder data indicate that some patients profit from cognitive therapy, whereas other patients profit from antidepressants. In this study, we mainly wish to optimize the allocation of patients into different treatment categories (and not to predict Post-test Anxiety). In this case, it may be better to use the bootstrap (Efron & Tibshirani, 1993) in establishing the stability of the splitting variables and the split points that allocate the patients into groups.

We limited our analyses to a treatment variable with two categories (representing two experimental conditions). The described analysis strategy by RTA can be generalized, however, to treatment variables with more than two categories. Only the second phase of the RTA has to be modified, with regression trunks fitted to each treatment group separately. These regression trunks can be merged into one regression trunk with multiple splits forced on the treatment variable.

7. Computational Note

The ANOVA analyses were performed with SPSS, version 10.0. All other analyses were performed with Splus 2000. The default random generator of Splus was used. The regression tree and forward stepwise regression analyses were carried out with the Splus-functions *tree* and *menuStep*, respectively. The "forward" option in the function *menuSTEP* adds predictor variables in a linear regression model on the basis of Mallows's C_p statistic (1973). An author prepared function was written in Splus to perform the regression trunk approach.

References

Bakker, A., van Dyck, R., Spinhoven, P., & van Balkom, A.J.L.M. (1999). Paroxetine, clomipramine and cognitive therapy in the treatment of panic disorder. *Journal of Clinical Psychiatry*, *60*, 831–838.

Bakker, A., Spinhoven, P., van Balkom, A.J.L.M., Vleugel, L., & van Dyck, R. (2000). Cognitive therapy by allocation versus cognitive therapy by preference in the treatment of panic disorder. *Psychotherapy and Psychosomatics*, *69*, 240–243.

Bakker, A., Spinhoven, P., van der Does, A.J.W., van Balkom, A.J.L.M., & van Dyck, R. (2002). Locus of control orientation in panic disorder and the differential effects of treatment. *Psychotherapy and Psychosomatics*, *71*, 85–89.

Barber, J.P., & Muenz, L.R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the treatment for depression collaborative research program. *Journal of Consulting and Clinical Psychology*, *64*, 951–958.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., & Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica*, *5*, 641–666.

Ciampi, A. (1991). Generalized regression trees. *Computational Statistics and Data Analysis*, *12*, 57–78.

Clark, L.A., & Pregibon, D. (1993). Tree-based models. In J.M. Chambers & T.J. Hastie (Eds.), Statistical models in S (pp. 377–419). London: Chapman & Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J., & Snow, R.E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Dance, K.A., & Neufield, R.W.J. (1988). Aptitude-treatment interaction in the clinical setting: A review of attempts to dispel the "patient uniformity" myth. *Psychological Bulletin, 104*, 192–213.

Digiusto, E., & Bird, K.D. (1995). Matching smokers to treatment: Self-control versus social support. *Journal of Consulting and Clinical Psychology, 63*, 290–295.

Dusseldorp, E., & Meulman, J.J. (2001). Prediction in medicine by integrating regression trees into regression analysis with optimal scaling. *Methods of Information in Medicine, 40*, 403–409.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Gaul, W., & Säuberlich, F. (1999). Classification and positioning of data mining tools. In W. Gaul & H. Locarek-Junge (Eds.), *Classification in the information age. Studies in classification, data analysis, and knowledge organization* (pp. 145–154). Heidelberg: Springer.

Glidden-Tracey, C.E., & Wagner, L. (1995). Gender salient attribute × treatment interaction effects on ratings of two analogue counselors. *Journal of Counseling Psychology, 42*, 223–231.

Goska, R.E., & Ackerman, Ph. L. (1996). An aptitude-treatment interaction approach to transfer within training. *Journal of Educational Psychology, 88*, 249–259.

Hand, D.J. (1997). *Construction and assessment of classification rules.* Chichester: Wiley.

Harrell, F.E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* New York: Springer.

Humphreys, L.G., & Fleishman, A. (1974). Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology, 66*, 464–472.

Koziol, J.A., & Wu, S.C.H. (1996). Changepoint statistics for assessing a treatment-covariate interaction. *Biometrics, 52*, 1147–1152.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica, 12*, 361–386.

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica, 7*, 815–840.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics, 15*, 661–675.

McInerney, V., McInerney, D.M., & Marsh, H.W. (1997). Effects of metacognitive strategy training within a cooperative group learning context on computer achievement and anxiety: An aptitude-treatment study. *Journal of Educational Psychology, 89*, 686–695.

Meulman, J.J., Heiser, W.J., & SPSS (1999). *SPSS Categories 10.0.* Chicago, IL: SPSS Inc.

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

Piper, W.E., McCallum, M., Joyce, A.S., Azim, H.F., & Ogrodniczuk, J.S. (1999). Follow-up findings for interpretive and supportive forms of psychotherapy and patient personality variables. *Journal of Consulting and Clinical Psychology, 67*, 267–273.

Quinlan, J.R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Rice, J.A. (1995).*Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury Press.

Rueter, M.A., Conger, R.D., & Ramisetty-Mikler, S. (1999). Assessing the benefits of a parenting skills training program: A theoretical approach to predicting direct and moderating effects. *Family Relations, 48*, 67–77.

Shomon-Salomon, V., & Hannah, M.T. (1991). Client-treatment interaction in the study of differential change processes. *Journal of Consulting and Clinical Psychology, 59*, 217–225.

Snow, R.E. (1991). Aptitude-treatment-interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59*, 205–216.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Thompson, S.C., & Wierson, M. (2000). Enhancing perceived control in psychotherapy. In C.R. Snyder & R.E. Ingram (Eds.), *Handbook of psychological change: Psychotherapy processes & practices for the 21st Century* (pp. 177–197). New York: Wiley.

Van der Linden, W.J. (1981). Using aptitude measurements for the optimal assignment of subjects to treatments with and without mastery scores. *Psychometrika, 46*, 257–274.

Manuscript received 20 FEB 2001
Final version received 20 JUN 2003