

SI-CHAID[®] 4.0

USER'S GUIDE

Jay Magidson

For more information about Statistical Innovations Inc. please visit our website at <http://www.statisticalinnovations.com>

or contact us at

Statistical Innovations Inc.
375 Concord Avenue, Suite 007
Belmont, MA 02478
e-mail: michael@statisticalinnovations.com

SI-CHAID® is a registered trademark of Statistical Innovations Inc.

Windows is a trademark of Microsoft Corporation.

SPSS is a trademark of SPSS, Inc.

Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

SI-CHAID® 4.0 User's Guide.

Copyright © 2005 by Statistical Innovations Inc.

All rights reserved.

No part of this publication may be reproduced or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission from Statistical Innovations Inc.

We strongly encourage any feedback on this manual or the program. Please send you comments directly to Michael Denisenko at michael@statisticalinnovations.com.

This document should be cited as " J. Magidson (2005) SI-CHAID 4.0 User's Guide. Belmont, Massachusetts: Statistical Innovations Inc."

Compatibility

SI-CHAID® is designed for computers running Windows 95, Windows 98, Windows 2000, Windows XP, Windows NT 4.0, or later

Customer Service

If you have any questions concerning your shipment or account, see Contacting Statistical Innovations. Please have your invoice number ready for identification when calling.

Training Seminars

We provide public and onsite training seminars on SI-CHAID. We also offer online courses. For information or to be placed on our mailing list, see Contacting Statistical Innovations or visit our website.

Tell Us Your Thoughts

Your comments are important to us. Please write or e-mail us about your experiences with SI-CHAID. We especially like to hear about new and interesting applications using SI-CHAID. Consider submitting examples and application ideas for inclusion on our website.

Contacting Statistical Innovations

To contact us or to be placed on our mailing list, visit our website at <http://www.statisticalinnovations.com> or write us at Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478. You can also e-mail us at michael@statisticalinnovations.com.

Preface

I am pleased to present SI-CHAID 4.0, the next generation of CHAID (CHI-squared Automatic Interaction Detection) analysis. SI-CHAID 4.0 features numerous improvements over our earlier programs, SPSS CHAID 6.0 for Windows and SI-CHAID 2.0, including the important extension to multiple dependent variables. That extension becomes possible in conjunction with either of our sister products Latent GOLD 4.0 and Latent GOLD Choice 4.0. In addition, the ability to save entire trees or tree branches allows additional applications such as the use of a holdout sample for validation (see Tutorial #3).

I hope that you find this manual as easy-to-use as the program. It begins with a brief overview of the program and new features, followed by four tutorials, which provide a step-by-step introduction to using the program. The Command References section contains the detailed descriptions of all features and aspects of the program. It is divided into the CHAID Define and the CHAID Explore sections, describing the Define and Explore modules of the program, respectively.

The first tutorial, "Beginning a CHAID Analysis", uses a traditional database marketing application to develop a response-based segmentation. It guides you through the major features of the program and is a good place to start for those who are new to CHAID. The second tutorial, "Using SI-CHAID to Identify Profitable Segments", shows how to develop a segmentation tree when the dependent variable is quantitative (measuring profitability). Tutorial #3, "Using SI-CHAID with a Hold-Out Sample", illustrates the use of the program with a hold-out sample. Tutorial #4, "Using CHAID with Multiple Correlated Dependent Variables", describes an extended CHAID analysis to develop a demographic segmentation that is predictive of 11 dependent variables. (See also Latent GOLD tutorial #4 for another application of this extended CHAID capability).

The Appendix contains my article, "The CHAID Approach to Segmentation Modeling: CHI-squared Automatic Interaction Detection", which provides technical details to supplement Tutorial #1. Reprints of 2 additional articles, which supplement Tutorials #2 and #4, are included with your program CD. Please visit the Statistical Innovations' website, <http://www.statisticalinnovations.com>, for up to date developments about SI-CHAID and our other programs.

I hope you enjoy using SI-CHAID to explore your data.

I wish to thank the Polk Company for making the magazine subscription data available. This data set accompanies the software and is used throughout this manual for purposes of illustration.

I also wish to thank J. Alexander Ahlstrom for his assistance in the design and development of the program and Michael Denisenko for his valuable contribution in the production of this manual.

Jay Magidson

Belmont, Massachusetts

April 2005

TABLE OF CONTENTS

| | |
|---|-----------|
| SI-CHAID Overview | 1 |
| New Features in SI-CHAID 4.0 | 2 |
| Tutorial 1: Beginning A CHAID Analysis | 3 |
| The Data | 3 |
| Setting up the Model | 5 |
| Opening the Data File | 5 |
| Assigning Variables | 6 |
| Scanning the Data | 7 |
| Setting Options | 7 |
| Growing a Tree | 8 |
| Growing a Tree in Automatic Mode | 9 |
| Gains Charts | 10 |
| Detailed Gains Charts | 10 |
| Summary Gains Chart | 12 |
| Scoring your file | 13 |
| Tables | 14 |
| After-Merge Table | 14 |
| Before-Merge Table | 15 |
| Comparing Tables Before and After Merging | 16 |
| Obtaining Frequency Counts | 16 |
| Growing a Tree in Interactive Mode | 17 |
| Rearranging Categories | 18 |

Tutorial 2: Using SI-CHAID to Identify Profitable Segments 19

The Data19

Modifying the Previous Analysis File20

Assigning Category Scores22

 Nominal Method22

 Ordinal Method27

Tutorial 3: Using SI-CHAID with a Hold-out Sample 31

Tutorial 4: Using CHAID with Multiple Correlated Dependent Variables
..... 38

The Data38

Steps Used to Obtain the CHAID Segments40

 Growing the CHAID Tree41

 Step 3: Show how the CHAID Segments Predict the
 11 Dependent Variables47

 Use of Correlated vs. Uncorrelated Dependent Variables55

SI-CHAID Define 56

Define Menus56

 File Menu57

 Edit Menu58

 View Menu58

 Model Menu58

 Help Menu60

 Menu Shortcuts60

Model Analysis Dialog Box60

 Variables Tab61

 Scan65

 Details65

 Groups65

Options Tab66

Technical Tab68

Predictor Options Tab70

SI-CHAID Explore 72

Tree Diagram View73

SI-CHAID® 4.0 USER'S GUIDE

| | |
|--|------------|
| Select Dialog | .74 |
| Rearrange Dialog | .75 |
| Delete | .75 |
| Hide | .76 |
| Node Items | .76 |
| Save | .77 |
| Restore | .77 |
| Tree Map View | .78 |
| Gains Chart View | .79 |
| Table View | .82 |
| Cell Format Options | .83 |
| Contents Options | .83 |
| Predictors Options: | .84 |
| Source Code View | .84 |
| SI-CHAID Explore Menu Reference | .85 |
| File Menu | .85 |
| Edit Menu | .85 |
| Tree Menu | .86 |
| View Menu | .87 |
| Window Menu | .87 |
| Help Menu | .88 |

The CHAID Approach to Segmentation Modeling:

CHI-Squared Automatic Interaction Detection 89

SI-CHAID Overview

SI-CHAID for Windows is a stand-alone program developed by Statistical Innovations Inc for performing CHAID (CHI-squared Automatic Interaction Detector) analyses. You can display your results simultaneously in the form of an intuitive tree diagram, crosstabulations, and a gains chart summary. Traditional CHAID analyses identify segments that are predictive of a single dependent variable which may be specified to be nominal or ordinal, and you can combine categories of a predictor variable in any way. For a detailed description of the nominal and ordinal CHAID algorithms, see Magidson (1994) and Magidson (1993) respectively.

The program accepts data directly from an ASCII data file. Alternatively, data, variable names and value labels may be imported from any .sav system file created by SPSS for Windows. SI-CHAID consists of two separate programs that work together - ChaidDefine and ChaidExplore. Either program may be launched from the Start Menu, or either can be used to execute the other.

The Define program is used to set up a CHAID Definition (.chd) file with the File → New command, or alter the specifications of an existing .chd file with File → Open. The typical setup includes the selection of the dependent variable, the predictor variables, the combine-type of the predictors, and various options for growing the tree (stopping rule, significance levels, etc.). Define may also be used to enter or modify scores for the categories of the dependent variable when the ordinal algorithm is specified. The model specifications, which are saved with a .chd extension, can be inspected with a text editor (Notepad, for example).

The Explore program allows you to grow or alter a SI-CHAID Tree, automatically or interactively, using the settings given in a previously saved (.chd) file. It can also be used to produce crosstabulations, gains charts, and if-then-else source code statements that can assist in scoring your data file.

SI-CHAID® 4.0 USER'S GUIDE

The application includes four tutorials. The first two tutorials introduce traditional uses of CHAID; the latter two illustrate new features in SI-CHAID 4.0. Specifically, Tutorial #1 illustrates the steps involved in setting up an analysis from scratch. Tutorial #2 builds on the analysis in Tutorial #1 and explores differences between the Nominal and Ordinal algorithms.

SI-CHAID is designed to be an exploratory analysis tool. The only limitation built into the program is that all variables are required to have at most 31 categories or levels. By default, continuous variables or other variables containing more than 31 levels will automatically be grouped into 16 levels. Alternatively, the grouping feature within SI-CHAID may be used to automatically reduce the number of categories to some specified number of levels.

Note that usage of (optional) numeric scores in SI-CHAID may serve different purposes:

- Category scores for an ordinal dependent variable provide a way to account for differential costs or gains associated with the categories of a dependent variable. For example, tutorial #2 illustrates the use of category scores to differentially weight the relative gains associated with paid responders, unpaid responders, and nonresponders in a direct marketing promotion. This example demonstrates the value of the ordinal algorithm in situations where the dependent variable contains more than 2 ordered categories and profitability (or other) scores are available.
- Scores are used in conjunction with the grouping feature to reduce the number of levels of a variable. Each reduced level is assigned a score equal to the mean score of the levels included in the new (grouped) level. If the variable being grouped has one or more values treated as missing, these missing variables are preserved in a separate last category of the grouped variable. In the case of a predictor variable, the resulting grouped variable may be included in an analysis using the FLOAT combine type.
- Scores may be used for the purpose of gains charts produced in a SI-CHAID analysis. A special SCORE option in the gains chart allows you to produce gains charts based on different sets of category scores without the need to create different .chd files.

NEW FEATURES IN SI-CHAID 4.0

The two major new features included in SI-CHAID 4.0 are the ability to produce segmentation trees that are predictive of *multiple* dependent variables (in conjunction with Latent GOLD 4.0 and/or Latent GOLD Choice 4.0), and the ability to save tree diagrams. For an example of the former, see Tutorial #4; for the latter, see Tutorial #3, which involves the use of a holdout sample.

Other new features include expanded Tables and Gains Chart options. Predictor by Dependent variable tables can now be obtained for *all* predictors (or all significant predictors) instead of just the current predictor) at any level of the tree. Gains Chart summaries now change interactively to reflect which tree node is specified as the active base. To obtain a gains chart summary for the *entire* tree, simply click on the root node of the tree to make it the active (current) node.

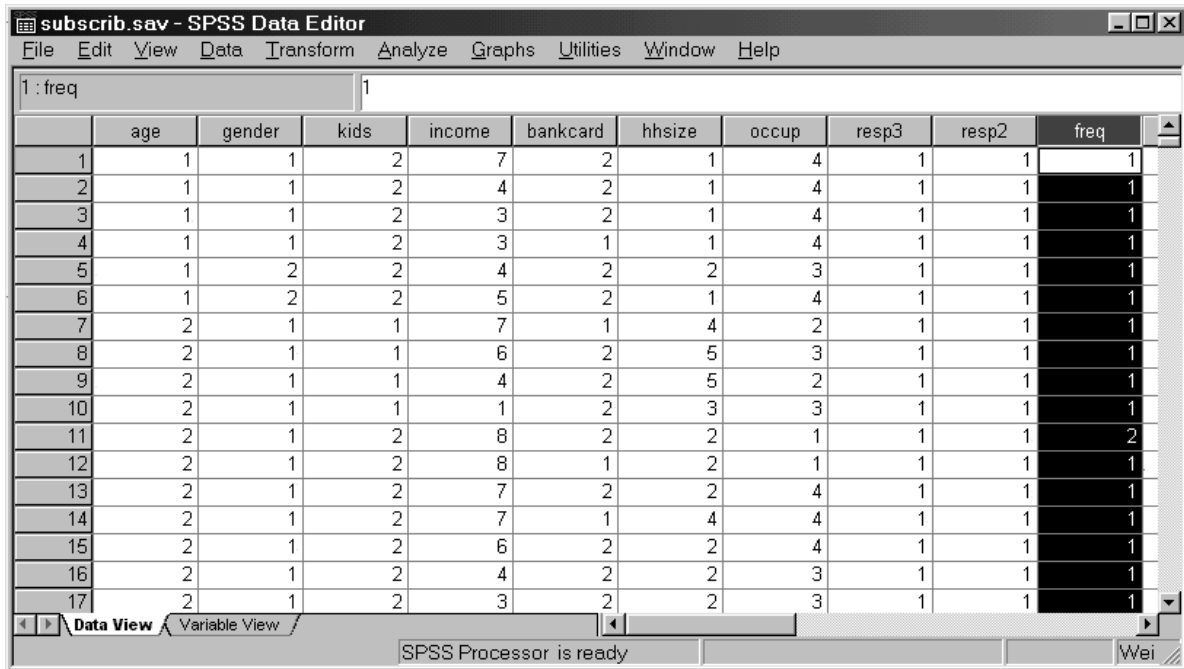
Tutorial 1: Beginning A CHAID Analysis

In this Tutorial we illustrate the basic functions and uses of SI-CHAID. We will show how to set up an analysis (.chd) file and grow a CHAID tree by using the standard CHAID algorithm, which is designed for a dichotomous or nominal dependent variable. In our example, we show how to determine CHAID segments that differ on response rates, and how gains charts can be used to predict the expected response from mailing/ targeting the most responsive segments. Tutorial #2 illustrates the use of the ordinal algorithm in SI-CHAID to identify segments best upon a profitability criterion. Both tutorials follow the analyses described in Magidson (1993).

The Data

In this tutorial, we will be using the SPSS file *subscrib.sav*, which contains information about a direct marketing promotion for a magazine subscription. Based on their response to this promotion, households were categorized as paid responders, unpaid responders, or nonresponders. Paid responders were households that returned a mail form, checked off the item that they would like to subscribe to the magazine, and later paid for the subscription. Unpaid responders were households that returned the form and checked off the item that they would like to subscribe to the magazine, but then cancelled their subscriptions prior to paying. Nonresponders includes all others (that is, households that did not request a subscription).

SI-CHAID® 4.0 USER'S GUIDE



| | age | gender | kids | income | bankcard | hhsiz | occup | resp3 | resp2 | freq |
|----|-----|--------|------|--------|----------|-------|-------|-------|-------|------|
| 1 | 1 | 1 | 2 | 7 | 2 | 1 | 4 | 1 | 1 | 1 |
| 2 | 1 | 1 | 2 | 4 | 2 | 1 | 4 | 1 | 1 | 1 |
| 3 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 1 | 1 | 1 |
| 4 | 1 | 1 | 2 | 3 | 1 | 1 | 4 | 1 | 1 | 1 |
| 5 | 1 | 2 | 2 | 4 | 2 | 2 | 3 | 1 | 1 | 1 |
| 6 | 1 | 2 | 2 | 5 | 2 | 1 | 4 | 1 | 1 | 1 |
| 7 | 2 | 1 | 1 | 7 | 1 | 4 | 2 | 1 | 1 | 1 |
| 8 | 2 | 1 | 1 | 6 | 2 | 5 | 3 | 1 | 1 | 1 |
| 9 | 2 | 1 | 1 | 4 | 2 | 5 | 2 | 1 | 1 | 1 |
| 10 | 2 | 1 | 1 | 1 | 2 | 3 | 3 | 1 | 1 | 1 |
| 11 | 2 | 1 | 2 | 8 | 2 | 2 | 1 | 1 | 1 | 2 |
| 12 | 2 | 1 | 2 | 8 | 1 | 2 | 1 | 1 | 1 | 1 |
| 13 | 2 | 1 | 2 | 7 | 2 | 2 | 4 | 1 | 1 | 1 |
| 14 | 2 | 1 | 2 | 7 | 1 | 4 | 4 | 1 | 1 | 1 |
| 15 | 2 | 1 | 2 | 6 | 2 | 2 | 4 | 1 | 1 | 1 |
| 16 | 2 | 1 | 2 | 4 | 2 | 2 | 3 | 1 | 1 | 1 |
| 17 | 2 | 1 | 2 | 3 | 2 | 2 | 3 | 1 | 1 | 1 |

Figure 1. Subscrib.sav file

The variables included in the file are:

- AGE age of head of household
- GENDER sex of head of household
- KIDS presence of children
- INCOME household income
- BANKCARD presence of bankcard
- HHSIZE household size
- OCCUP occupational status of head of household
- RESP3 coded 1 for paid, 2 for unpaid responders and 3 for nonresponders.
- RESP2 coded 1 for (paid and unpaid) responders, and 2 for nonresponders – to be used as the dependent variable in this tutorial
- FREQ number of cases (designated as a case weight in SPSS)

The purpose of our initial analysis is to identify household segments that are more likely to respond than other segments.

Setting up the Model

OPENING THE DATA FILE



To open the file,

- ▷ Open ChaidDefine.exe from the CHAID Directory
- ▷ Go to the File Menu and click New
- ▷ From the menu, select subscrib.sav

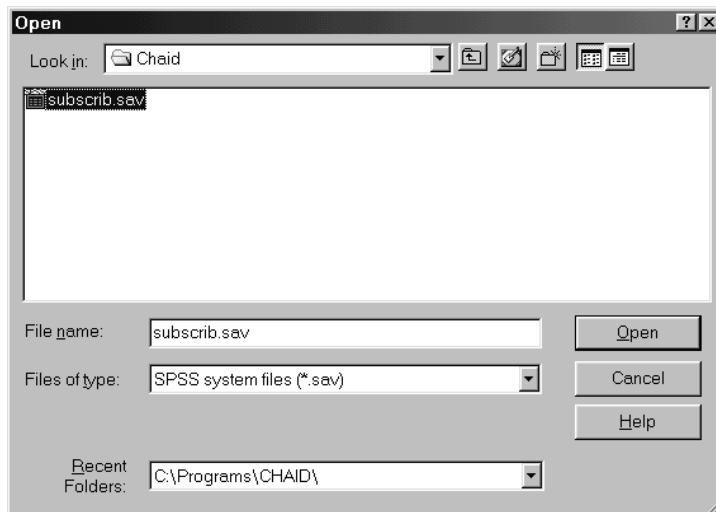


Figure 2. File New Dialog Box

Once you click on the file, the Model Analysis Dialog Box opens. It looks like this:

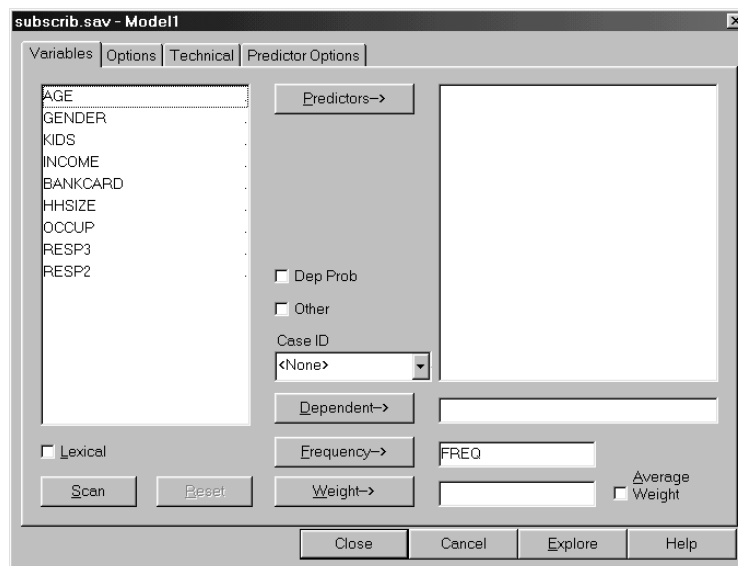


Figure 3. Model Analysis Dialog Box

SI-CHAID® 4.0 USER'S GUIDE

The variables in the data file subscrib.sav are included in the Variables List Box on the left, except for the variable `FREQ`. SI-CHAID automatically entered this variable in the frequency box because it was specified within SPSS to be used as a case weight when creating the SPSS save file.)

ASSIGNING VARIABLES

To begin a CHAID analysis, we need to select one (or more) dependent variables and at least one predictor. Optionally, one of two weight variables can be specified - a case weight (frequency) and a sampling weight (weight).

For this analysis, the dichotomous variable `RESP2` will be the single dependent variable. For an example of multiple dependent variables, see Tutorial #3 in this manual.



To select the dependent variable:

- ▷ Click on `RESP2` in the Variables Box.
- ▷ Click on "Dependent" to move `RESP2` to the Dependent Variable Box

Next, we will select the predictor variables. The predictor variables for this analysis will be `AGE`, `GENDER`, `KIDS`, `INCOME`, `BANKCARD`, `HHSIZE`, and `OCCUP`.

- ▷ Highlight `AGE`, `GENDER`, `KIDS`, `INCOME`, `BANKCARD`, `HHSIZE`, and `OCCUP`.
- ▷ Click on "Predictors" to move the above variables to the Predictor Variable Box.

The completed Model Analysis Dialog Box should look like this:

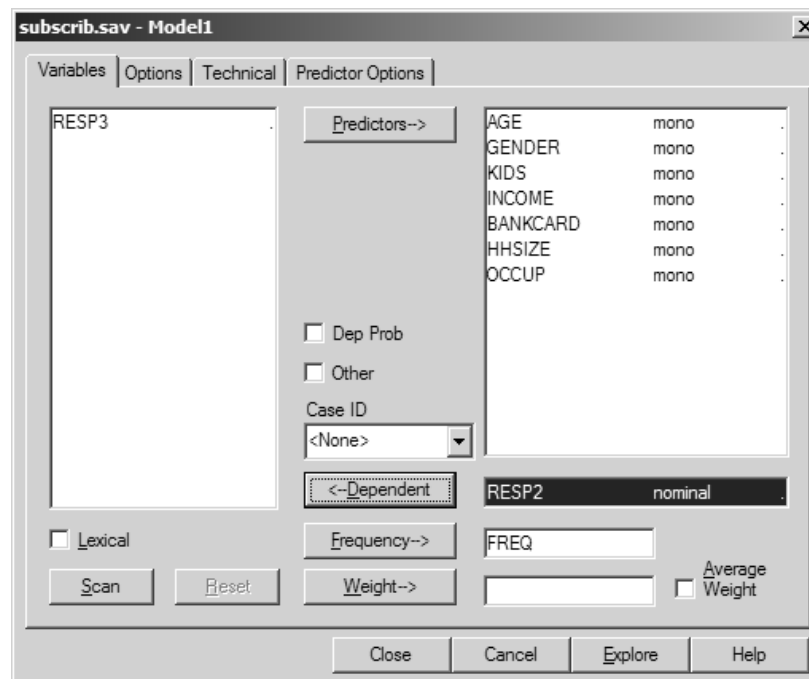


Figure 4. Model Analysis Dialog Box with variables in place

SCANNING THE DATA

Now that you have set your analysis options, you are ready to scan the data file.



To scan the file,

- ▷ Click on Scan

After the data scans, the default combine types appear next to each predictor. The combine type specifies how the categories of the predictor are allowed to merge. You can change the combine type for a predictor from the Predictor Options tab or by right clicking on the variable and selecting the desired combine type name from the pop-up menu.

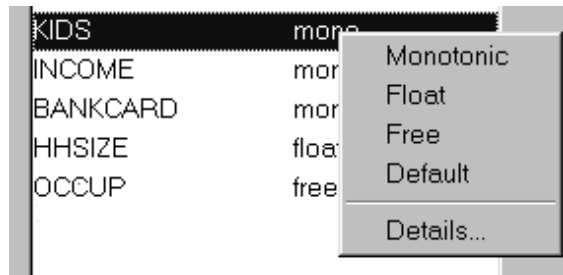


Figure 5. Predictor Options pop-up menu

- ▷ Right-click on OCCUP and select "Free" to define OCCUP as a free variable

You may view category labels by selecting Details... from this menu or by double-clicking on a predictor or the dependent variable name. This action brings up the category-labels window.

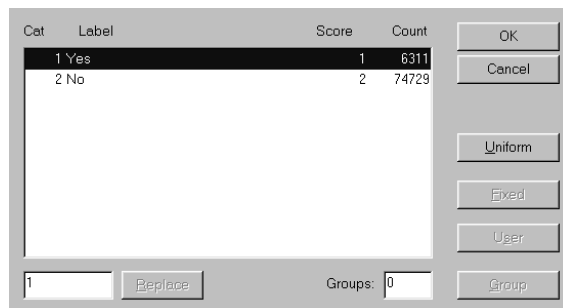


Figure 6. Category Labels Window

SETTING OPTIONS

The Options Tab controls the operation of the CHAID segmentation algorithm, including the stopping rule and the minimum segment size.

SI-CHAID® 4.0 USER'S GUIDE

- ▷ Click on the Options Tab to open the Options Dialog Box
- ▷ Double-click on the Depth Limit text box and enter 2 to set the analysis depth limit at 2. That tells SI-CHAID that the tree should expand to no more than two levels deep.
- ▷ Leave the other options, Merge Level and Eligibility Level, at their default levels.
- ▷ Select Auto in the Startup Mode Menu on the right. This tells SI-CHAID to run the analysis automatically.

Your Options Tab should now look like this:

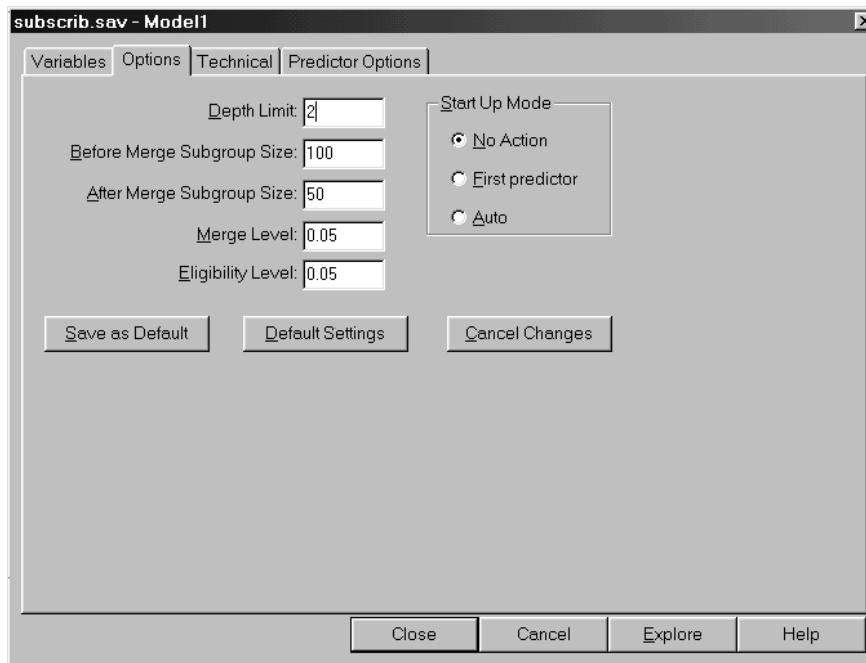


Figure 7. Options Tab

Growing a Tree

After you have set all the options, you are now ready to grow a segmentation tree.

- ▷ Click Explore

SI-CHAID automatically prompts you to save the new model with a Save As dialog box.

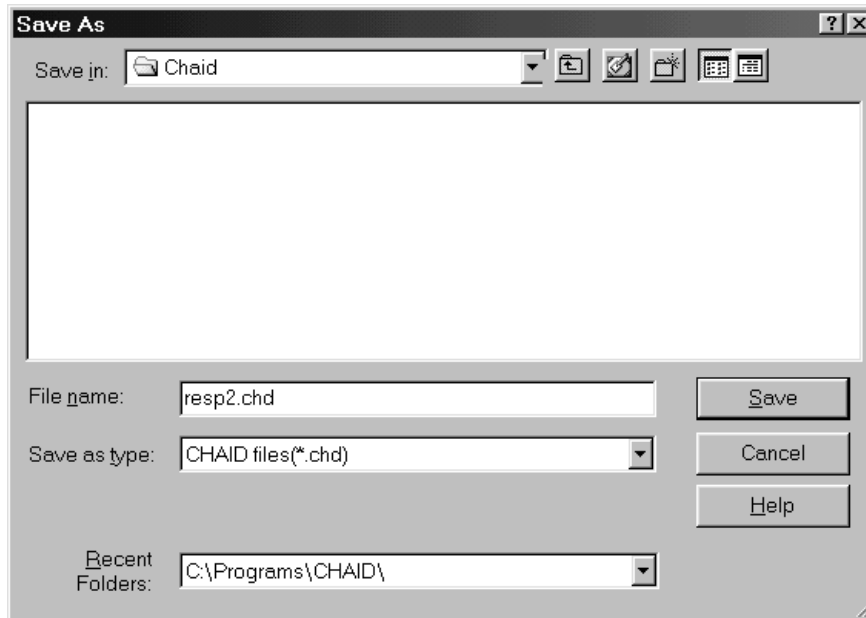


Figure 8. Save As Dialog Box

In the File Name box, type resp2 to override the suggested filename and click on Save. That tells SI-CHAID to save your analysis settings to an analysis file with the name resp2.chd. All printed and saved output will be prefixed by the name resp2.

GROWING A TREE IN AUTOMATIC MODE

After you click Save, SI-CHAID automatically opens the ChaidExplore program and grows the tree.

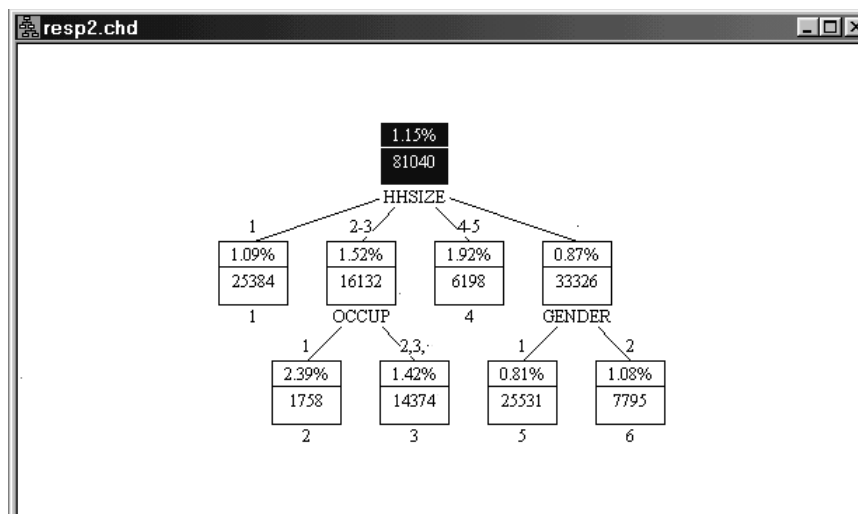


Figure 9. Tree Diagram

By default, SI-CHAID displays the tree diagram in local mode. The local mode displays detailed results within each node, and numbers each terminal node. The results of the CHAID tree shows 6 segments, details for which are displayed in each of the 6 terminal nodes. The highest response rate is obtained from segment 2, defined as households of size 2 or 3 (HHSIZE = 2-3) and occupation = 'white collar' (OCCUP = 1). Terminal node #2 shows

that there are a total of 1,758 cases in this segment and the response rate is 2.39%. The next best segment is obtained from households containing 4 or more persons (terminal node #4), and the response rate for this segment is 1.92%.

For large trees, all terminal nodes may not be visible at once. In this case, a global 'Tree Map' view is useful to get a better feel for the entire tree. To switch to global mode,

- ▷ Click on Window
- ▷ Select New Tree Map

The Global Tree Window then appears

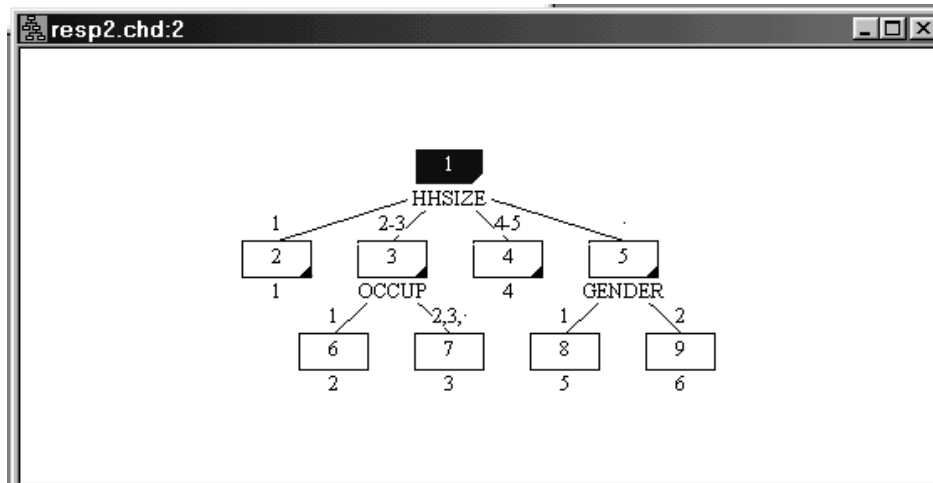


Figure 10. Global Tree Window

Gains Charts

The results of a CHAID analysis can also be displayed in the form of Gains Charts, which sort all or a subset of the segments from best to worst and also provides cumulative results expected based on the best K of these segments (or best quantile). In our current analysis, best is defined based on the percentage of cases in the first category of the dependent variable (response rate).

If the root node is the current node, the gains charts include all segments. If some other node is current, the gains charts are based on segments derived from the current node.

DETAILED GAINS CHARTS



To produce a detailed gains chart corresponding to the entire CHAID tree:

- ▷ Click on the root node of the tree diagram to make it the current node
- ▷ Click on Window to display the Window options
- ▷ Select New Gains

SI-CHAID displays a detailed gains chart, where the segments are listed from best to worst.

| Id | size | % of all | resp | %resp | score | index | Cum: size | % of all | resp | %resp | score | index |
|----|-------|----------|------|-------|-------|-------|-----------|----------|------|-------|-------|-------|
| 2 | 1758 | 2.2 | 42 | 4.5 | 2.39 | 208 | 1758 | 2.2 | 42 | 4.5 | 2.39 | 208 |
| 4 | 6198 | 7.6 | 119 | 12.8 | 1.92 | 167 | 7956 | 9.8 | 161 | 17.3 | 2.02 | 176 |
| 3 | 14374 | 17.7 | 204 | 21.9 | 1.42 | 124 | 22330 | 27.6 | 365 | 39.2 | 1.63 | 142 |
| 1 | 25384 | 31.3 | 276 | 29.6 | 1.09 | 95 | 47714 | 58.9 | 641 | 68.9 | 1.34 | 117 |
| 6 | 7795 | 9.6 | 84 | 9.0 | 1.08 | 94 | 55509 | 68.5 | 725 | 77.9 | 1.31 | 114 |
| 5 | 25531 | 31.5 | 206 | 22.1 | 0.81 | 70 | 81040 | 100.0 | 931 | 100.0 | 1.15 | 100 |

Figure 11. Gains Chart

The column labeled Id contains segment numbers. The next column (size) contains the number of cases in this segment, followed by a re-expression of segment size in terms of a percentage (% of all). The 4th column (resp) contains the number of responders in the segment, followed by a re-expression of this quantity in terms of percentage. Thus, we see that segment 2 represents 2.2% of all cases, but accounts for 4.5% of all respondents.

The next column displays the response rate for the associated segment (score). Thus, we see that segment 2 has the highest response rate (2.39%). The next highest response rate is 1.92% (segment 4).

The score represents the mean category score. By default, the category scores are '1' for the first category, and '0' for all others, so that the mean score corresponds to the % in the first category (responders in this example).



To change the category scores,

- ▷ right click on the gains chart to bring up the gains chart control panel.

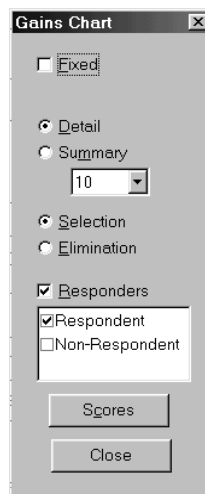


Figure 12. Gains Chart Control Panel

Note that a check mark appears next to Responders to indicate that the default gains chart is presented.

- ▷ Click the Scores button, to bring up the gains chart category scores window.
- ▷ Double click the score you wish to change, enter the replacement score and click the Replace button.
- ▷ Click OK after all the new scores have been entered.



To view the new gains chart based on the revised scores,

- ▷ click Responders in the Gains Chart control to remove the check mark for the default gains chart.
- ▷ Now click Responders once again in the Gains Chart control panel to restore the default gains chart.

The *index* column for a given segment measures the average response score for that segment relative to the average score for the total sample. The index score for segment 2 is 208, which is computed as $(2.39\% / 1.15\%) \times 100$. This means that the response rate for this segment is 108% higher than average.

Columns 8 through 13 in the gains chart present cumulative statistics. From the columns labeled *Cum: size, % of all*, and *score*, you can see that the three highest responding segments constitute 27.6% of the sample and have a combined response rate of 1.63%. The final column, *Cum: index*, measures the cumulative average response score for these segments relative to the average score for the total sample. For example, the index for the three best segments is 142 ($1.63\% / 1.15\%$). Thus, the three best segments, taken together, responded at a rate 42% higher than average.

If you know the break-even response rate (or if the category scores reflect profitability), you can use gains charts to determine the segments to which you should mail future promotions. For example, suppose that when you take into account the cost of mailing and the gain from responders, you need a response rate of 1.45% to break even. Looking at the Gains chart above, (and assuming that this is your final segmentation), you would expect to make a profit if you mailed only the top two segments, since the score for the remaining households falls below the break-even level. Large savings could be gained by mailing only to segments with the highest response rates.

SUMMARY GAINS CHART

The summary gains chart summarizes the predicted response rate at various depths of the file. That is, the summary gains chart tells you the results that would be attained by targeting the best Q-percent of the file. This form of the gains chart is especially useful for comparing the results of 2 or more different CHAID trees. By default, the results are displayed in deciles.



To obtain a summary gains chart,

- ▷ click Summary on the (top) of the gains chart control panel.

The gains chart changes to the following:

| tile | size | resp | %resp | score | index |
|------|-------|------|-------|-------|-------|
| 10 | 8104 | 163 | 17.5 | 2.01 | 175 |
| 20 | 16208 | 278 | 29.9 | 1.72 | 149 |
| 30 | 24312 | 387 | 41.5 | 1.59 | 138 |
| 40 | 32416 | 475 | 51.0 | 1.46 | 127 |
| 50 | 40520 | 563 | 60.4 | 1.39 | 121 |
| 60 | 48624 | 651 | 69.9 | 1.34 | 117 |
| 70 | 56728 | 735 | 78.9 | 1.30 | 113 |
| 80 | 64832 | 800 | 86.0 | 1.23 | 107 |
| 90 | 72936 | 866 | 93.0 | 1.19 | 103 |
| 100 | 81040 | 931 | 100.0 | 1.15 | 100 |

Figure 13. Summary Gains Chart

The score column shows that, the predicted response rate would be 2.01% if the best decile were mailed.

Scoring your file

You can obtain source code, which will allow you to score your file with segment definitions.

- ▷ Select New Source from the Windows menu

A window appears containing SPSS if-then-else statements which compute the variable chdsegmt containing the CHAID segment number.

```

resp2.chd:4
compute chdsegmt = $sysmis .
compute chderror = $sysmis .
do if (missing(HHSIZE)=0 & ((HHSIZE=1))) .
- compute chdsegmt = 1 .
else if (missing(HHSIZE)=0 & ((2<=HHSIZE & HHSIZE<=3))) .
- do if (missing(OCCUP)=0 & ((OCCUP=1))) .
- compute chdsegmt = 2 .
- else if ((2<=OCCUP & OCCUP<=3)
| missing(OCCUP)=1) .
- compute chdsegmt = 3 .
- else .
- compute chderror = 1 .
- end if .
else if (missing(HHSIZE)=0 & ((4<=HHSIZE & HHSIZE<=5))) .
- compute chdsegmt = 4 .
else if (missing(HHSIZE)=1) .
- compute chderror = 1 .

```

Figure 14. Source File

Tables

The *New Table Window* option displays a table of the dependent variable (columns) by the current predictor variable (rows). You can control whether the table displays row percentages, column percentages, total percentages, or cell frequencies, and whether the table shows merged or unmerged categories of the predictor.

AFTER-MERGE TABLE



To view a table showing row percentages for merged categories of HHSIZE at the top of the tree:

- ▷ Click the top (root) node of the tree diagram
- ▷ Select Window
- ▷ Click on New Table

Values in the Respondent column match the values displayed in each of the four HHSIZE nodes:

| HHSIZE (after) | Respondent | Non-Respondent | row % Total |
|-------------------|------------|----------------|----------------|
| 1 | 1.09 | 98.91 | 25384 |
| 2-3 | 1.52 | 98.48 | 16132 |
| 4-5 | 1.92 | 98.08 | 6198 |
| . | 0.87 | 99.13 | 33326 |
| Total | 1.15 | 98.85 | 81040 |

LR chi-square=70.96 df=3 prob=5.8e-14(adj)

Figure 15. After Merge Table

Notice that SI-CHAID merged categories 2 and 3, as well as categories 4 and 5.

The probability displayed in the bottom of the after-merge table, 2.7×10^{-15} , is adjusted for the fact that categories have been merged. The probability used by CHAID to rank predictors is the smaller of this adjusted probability and the probability associated with the table computed before category merging.

BEFORE-MERGE TABLE



To view a row percentage table of HHSIZE by RESP2 for unmerged HHSIZE categories:

- ▷ Right-click on the Table to bring up Table Display.
- ▷ In the pop-up menu, click on *Before Merge*

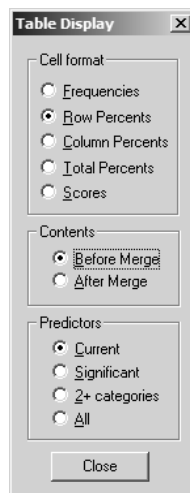


Figure 16. Table Display Menu

SI-CHAID automatically produces a table of row percentages before HHSIZE categories are merged, as shown below:

| HHSIZE (before) | Respondent | Non-Respondent | row % Total |
|--------------------|------------|----------------|----------------|
| 1 | 1.09 | 98.91 | 25384 |
| 2 | 1.49 | 98.51 | 11240 |
| 3 | 1.59 | 98.41 | 4892 |
| 4 | 1.79 | 98.21 | 3187 |
| Five or more | 2.06 | 97.94 | 3011 |
| . | 0.87 | 99.13 | 33326 |
| Total | 1.15 | 98.85 | 81040 |

LR chi-square=71.79 df=5 prob=4.4e-14

Figure 17. Before Merge Table

The table shows you the percentage of households in each HHSIZE category that responded to the promotion. For example, 1.09% of one-person households responded. Note that the total count in the lower right corner of the table (81,040) corresponds to the size of the highlighted node.

The table also displays the probability value (p value), a measure of statistical significance. The smaller the p value, the more statistically significant the predictor. The p value for HHSIZE before categories are merged is 4.4e- 14 (shorthand for 4.4 x 10-14, a highly significant result). In fact, HHSIZE is the most significant of all the predictors. That is why the first split in the tree is based on household size categories.

COMPARING TABLES BEFORE AND AFTER MERGING

To see why some of the categories of HHSIZE have been merged, compare the Before- and After- Merge tables. SI-CHAID merged two-person and three-person households because their before-merge response rates (1.49% and 1.59%) are not significantly different. The combined response rate for the merged categories is 1.52%. Similarly, SI-CHAID merges four- and five-person households, since the response rates for these subgroups (1.79% and 2.06%) are statistically indistinguishable. The combined response rate for the joint category is 1.92%.

OBTAINING FREQUENCY COUNTS



To obtain frequency counts before HHSIZE categories are merged

- ▷ Right-click on the Table to bring up *Table Display*.
- ▷ In the pop-up menu, click on *Frequencies*.

SI-CHAID automatically produces the table of frequency counts shown below:

| HHSIZE (before) | Respondent | Non-Respondent | n Total |
|--------------------|------------|----------------|------------|
| 1 | 276 | 25108 | 25384 |
| 2 | 168 | 11072 | 11240 |
| 3 | 78 | 4814 | 4892 |
| 4 | 57 | 3130 | 3187 |
| Five or more | 62 | 2949 | 3011 |
| . | 290 | 33036 | 33326 |
| Total | 931 | 80109 | 81040 |

LR chi-square=71.79 df=5 prob=4.4e-14

Figure 18. Frequency Count Table

The first row of the table indicated that 276 one-person households responded. The response rate displayed on the tree diagram (1.09%) is obtained by dividing the frequency by the total number of one-person households (25,384).

Growing a Tree in Interactive Mode



To explore your data in interactive mode, simply select any node of the tree you wish to analyze:

- ▷ Using the mouse or arrow keys, move to the HHSIZE = 23 node
- ▷ Right-click on the 23 node and select Select from the pop-up menu

The Select Predictors dialog box will come up. Three predictors show up as offering significant splits of this subgroup. They are ranked from most to least significant. At this point you may a) split the subgroup using the best predictor (OCCUP), b) select one of the other predictors to split on, or c) change the Detail level display selection to include variables that are not significant in the list of predictors.

- ▷ Highlight AGE and click OK to select it as the next predictor

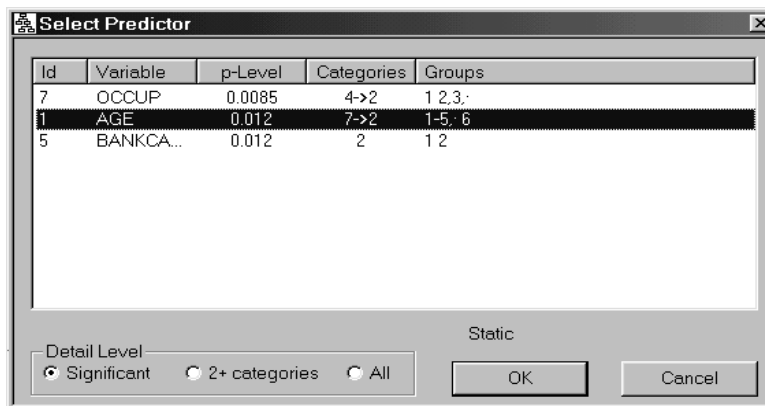


Figure 19. Selecting Predictor AGE

The tree now looks as follows:

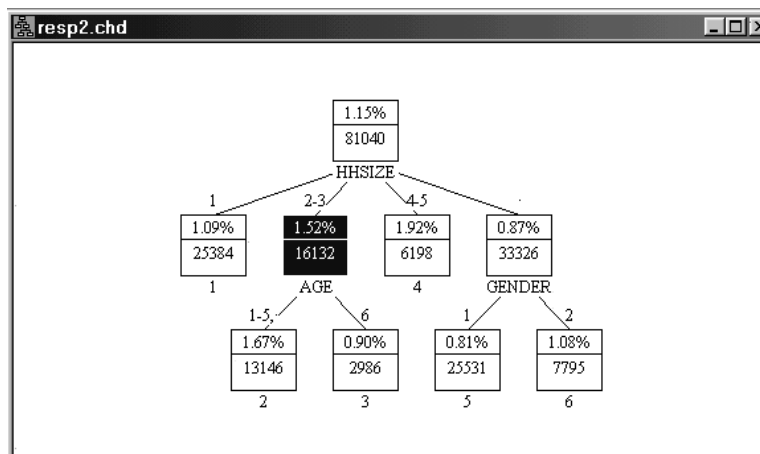


Figure 20. Tree Diagram with AGE used to Split the HHSIZE = 2-3 Parent Node

REARRANGING CATEGORIES

- ▷ Right click and select Rearrange
- ▷ Select the 5 age range categories between 18-64 as the 1st re-arranged category
- ▷ click the right arrow to move them to the right-most window

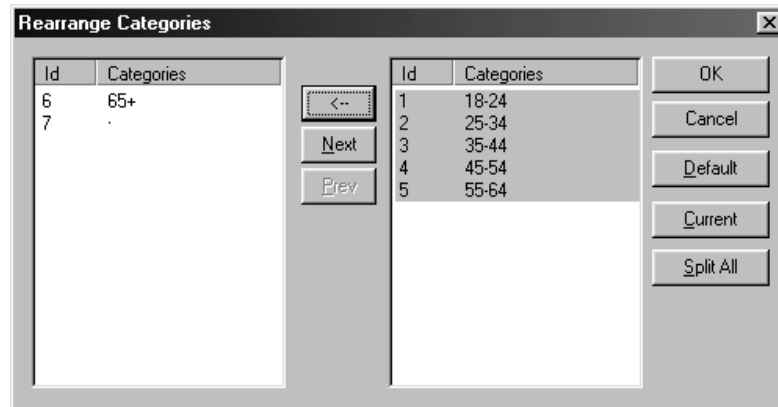


Figure 21. Rearranging Categories

- ▷ Click Next
- ▷ Select age 65+ as the 2nd re-arranged category
- ▷ click the right arrow
- ▷ click next
- ▷ Select the missing age group
- ▷ Click the right arrow
- ▷ Click OK

The rearranged tree will now look as follows:

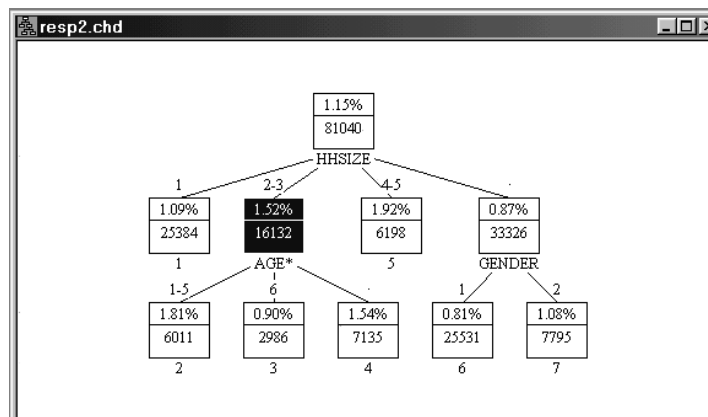


Figure 22. Rearranged Tree Diagram

SI-CHAID is designed as a useful tool to explore your data. There are no right or wrong trees. Feel free to explore your data as you wish.

Tutorial 2: Using SI-CHAID to Identify Profitable Segments

This tutorial shows how to use the CHAID ordinal algorithm to segment based on profitability scores. We will again use the magazine subscription data set, *subscribe.sav*, used previously in Tutorial 1. However, our dependent variable will now be *RESP3*, coded 1 (paid responder), 2 (unpaid responder) and 3 (nonresponder). We'll compare a default nominal CHAID segmentation of *RESP3* to the ordinal CHAID analysis that takes into account the gain (or loss) associated with each response group. For simplicity, we utilize the SI-CHAID option settings used in Magidson (1993).

The Data

For this Tutorial, we will be using the same data file as for Tutorial 1: Beginning a CHAID Analysis. The file *subscribe.sav* contains information about a direct marketing promotion used to encourage people to subscribe to a magazine. Households that were sent the promotion were categorized as paid responders, unpaid responders, or nonresponders. The data and analyses are described in more detail in Magidson (1993).

Modifying the Previous Analysis File



If your analysis file from tutorial #1 is not still open, re-open it:

- ▷ Open the Define program
- ▷ Select Open from the File Menu
- ▷ From the files listed select 'resp2.chd' and click the Open button

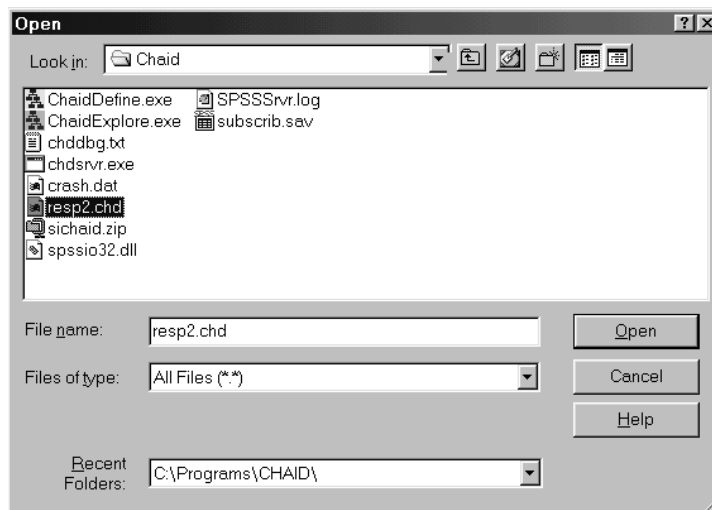


Figure 23. File Open Dialog Box

Your earlier analysis file is retrieved:

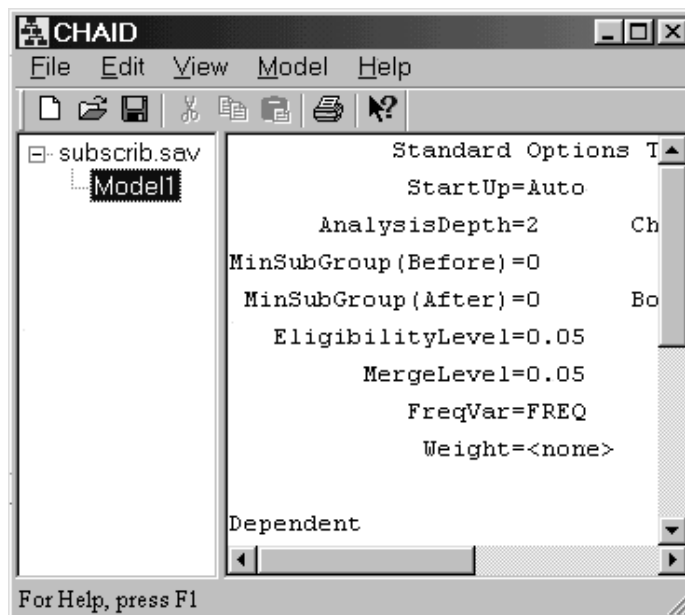


Figure 24. Analysis File for Model1

USING SI-CHAID TO IDENTIFY PROFITABLE SEGMENTS



To enter the Variables tab of the Model Analysis Dialog Box:

- ▷ Right-click on 'Model1' and select 'Edit'

Or alternatively,

- ▷ double-click on 'Model1'

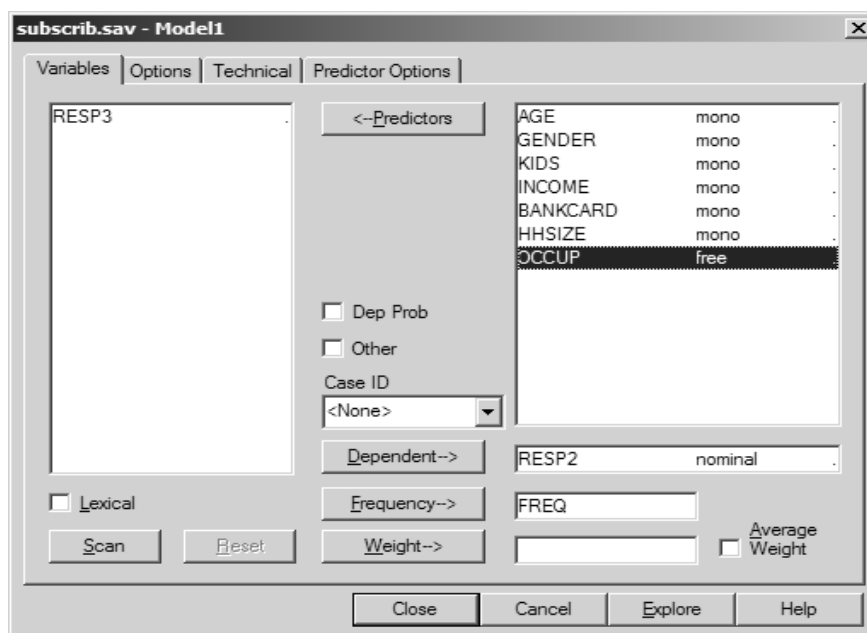


Figure 25. Model Analysis Dialog Box



To change the dependent variable from Resp2 to Resp3 and re-scan the data file:

- ▷ Click on Resp2
- ▷ Click the Dependent button
- ▷ Select Resp3 from the Variables box
- ▷ Click the Dependent button
- ▷ Click Scan

The Model Analysis Dialog Box should now look like this:

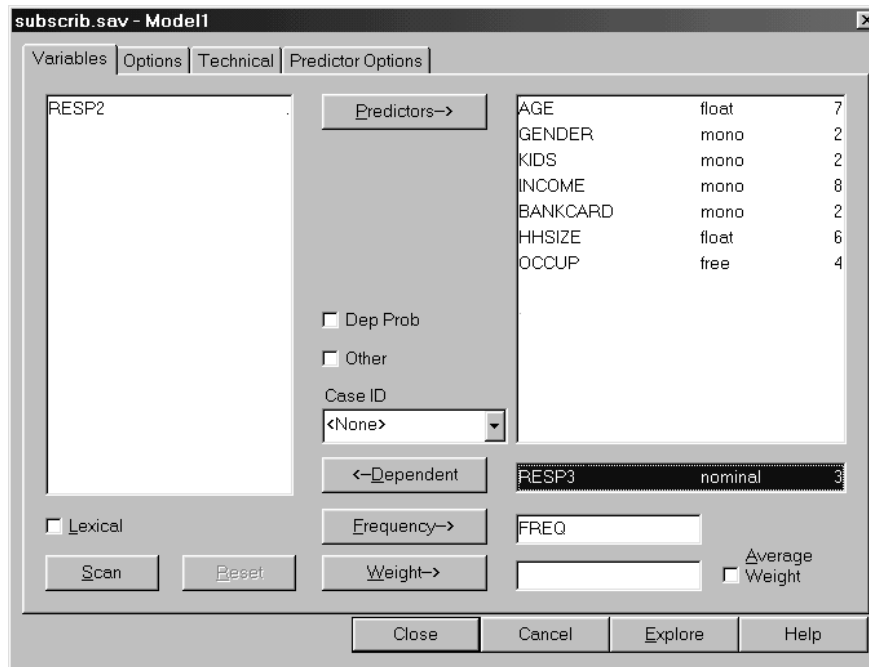


Figure 26. Model Analysis Dialog Box after editing

Assigning Category Scores

NOMINAL METHOD

Before growing the new tree, we will assign profitability scores to the categories of the dependent variable for future use. Although the standard CHAID algorithm (the 'nominal' algorithm) does not utilize these scores to grow the tree, the scores may still be used by the gains chart to identify which of the resulting segments are most profitable. Later we will compare results from the nominal segmentation to the segmentation obtained from the ordinal algorithm.

- ▷ Right-click on RESP3 in the dependent box of the Model Analysis Dialog Box
- ▷ In the pop-menu, select Details

USING SI-CHAID TO IDENTIFY PROFITABLE SEGMENTS

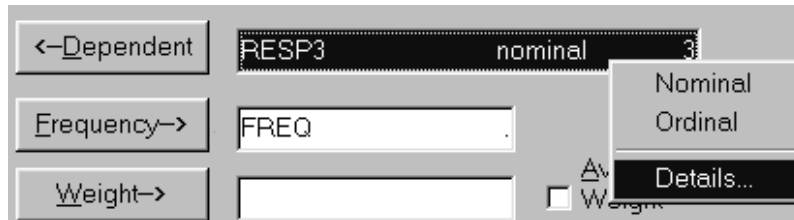


Figure 27. Options pop-up menu

Clicking Details will bring up the Edit Scores Box

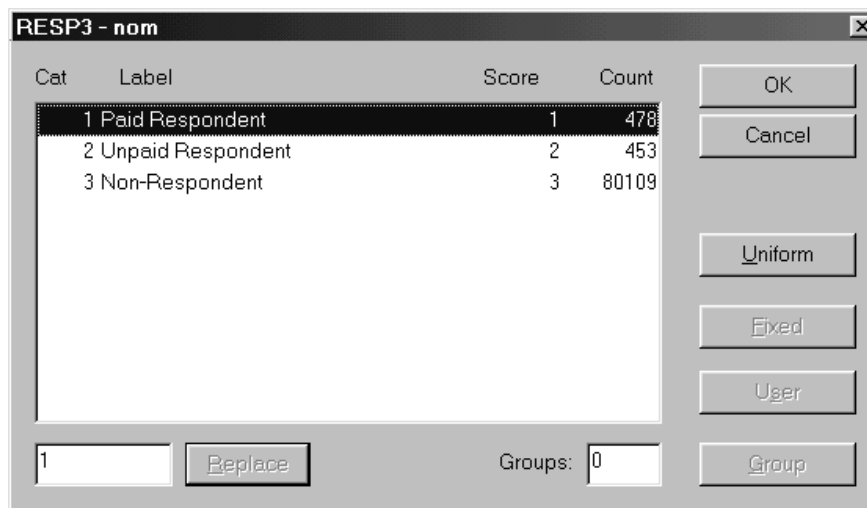


Figure 28. Edit Scores Box

(Alternatively, double-clicking on Resp3 would also get us to this screen)

The first category (Paid Respondent) is highlighted. The default scores correspond to the integer codes used in the SPSS file – 1,2 and 3. To change the score for Paid Respondents,

- ▷ Double-click on the 'Paid Respondent' label

The score '1' is highlighted in the Edit Scores box

- ▷ Replace the score '1' with the score '35' and click the Replace button

Now repeat these steps for the other categories:

- ▷ Double-click on the second category ('Unpaid Respondent').
- ▷ Replace the score '2' with the score '-7' and click the Replace button.
- ▷ Double-click on the third category ('Nonresponder').

- ▷ Replace the score '3' with the score '-0.15' and click the Replace button.

Your screen should now look like this:

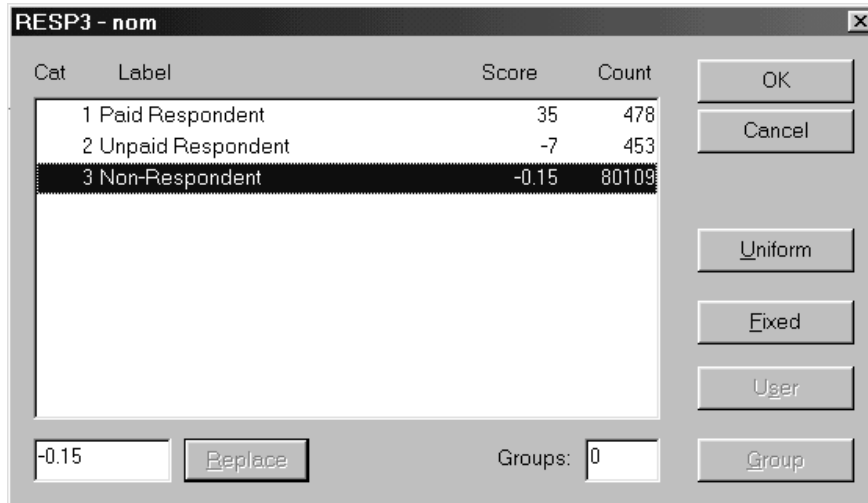


Figure 29. Edit Scores Box showing New Category Scores

- ▷ Click OK to return to the Model Analysis Dialog Box
- ▷ Now, go to the Options Tab
- ▷ Change the "Before Merge Subgroup Size" to '4500' and the "After Merge Subgroup Size" to '1500'. These were the settings used in the Magidson (1994) article.

The Options Tab should now look like this:

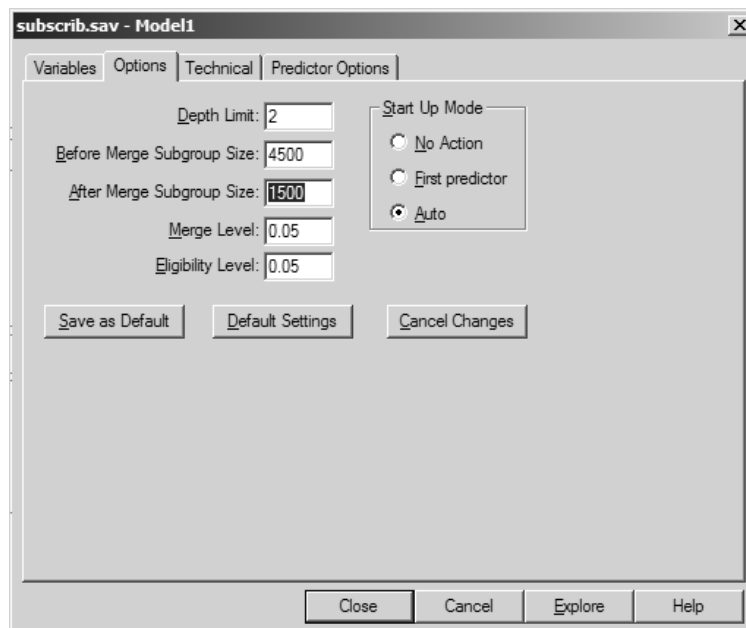


Figure 30. Options Tab after Editing

USING SI-CHAID TO IDENTIFY PROFITABLE SEGMENTS



To save the new analysis file and grow the tree:

- ▷ Click Explore
- ▷ In the File name box type *RESP3nom.chd* to override the suggested filename
- ▷ Click the Save button

This tells SI-CHAID to save your analysis settings to an analysis file with the name *RESP3nom.chd*. All printed and saved output will be prefixed by the name *RESP3nom*. Later, we will create another analysis file with named *RESP3ord.chd* corresponding to the ordinal algorithm.

After you click Save, SI-CHAID automatically opens ChaidExplore and generates the following 7-segment tree:

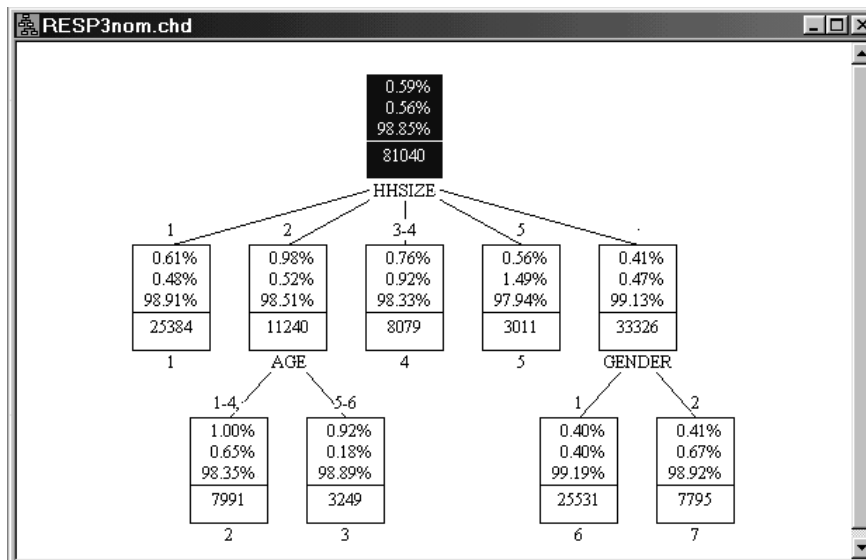


Figure 31. Tree Diagram showing 7 Segments

Notice that this *RESP3nom* solution differs from our earlier 6-segment *RESP2* solution (recall *Tutorial 1: Beginning a CHAID Analysis*). For example, while *HHSIZE* is still used for the first split, it is now merged into five categories instead of four. In our earlier analysis, *HHSIZE* categories 2 and 3 were merged. Now category 2 is a separate category and categories 3 and 4 are merged.

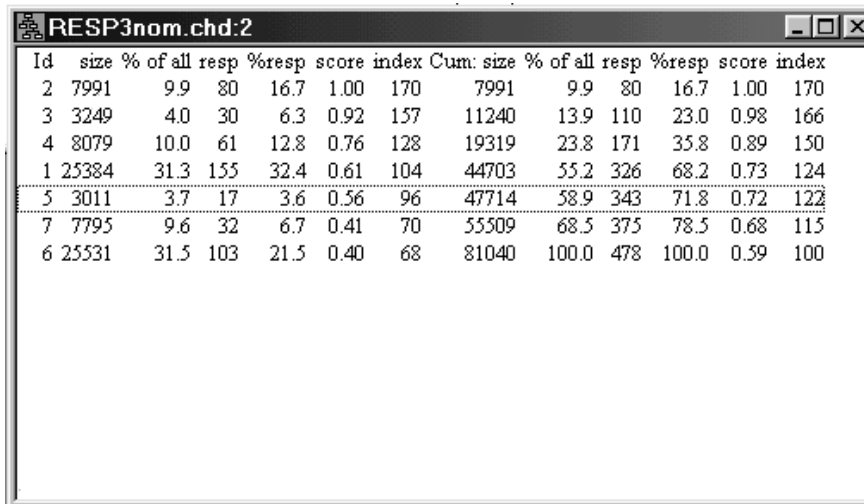


To obtain a gains chart for this segmentation,

- ▷ Select 'New Gains' from the Windows menu.

SI-CHAID® 4.0 USER'S GUIDE

The gains chart appears as follows:



| Id | size | % of all resp | %resp | score | index | Cum: size | % of all resp | %resp | score | index | | |
|----|-------|---------------|-------|-------|-------|-----------|---------------|-------|-------|-------|------|-----|
| 2 | 7991 | 9.9 | 80 | 16.7 | 1.00 | 170 | 7991 | 9.9 | 80 | 16.7 | 1.00 | 170 |
| 3 | 3249 | 4.0 | 30 | 6.3 | 0.92 | 157 | 11240 | 13.9 | 110 | 23.0 | 0.98 | 166 |
| 4 | 8079 | 10.0 | 61 | 12.8 | 0.76 | 128 | 19319 | 23.8 | 171 | 35.8 | 0.89 | 150 |
| 1 | 25384 | 31.3 | 155 | 32.4 | 0.61 | 104 | 44703 | 55.2 | 326 | 68.2 | 0.73 | 124 |
| 5 | 3011 | 3.7 | 17 | 3.6 | 0.56 | 96 | 47714 | 58.9 | 343 | 71.8 | 0.72 | 122 |
| 7 | 7795 | 9.6 | 32 | 6.7 | 0.41 | 70 | 55509 | 68.5 | 375 | 78.5 | 0.68 | 115 |
| 6 | 25531 | 31.5 | 103 | 21.5 | 0.40 | 68 | 81040 | 100.0 | 478 | 100.0 | 0.59 | 100 |

Figure 32. Gains Chart

The most profitable of these 7 segments (at the top of the list) is segment #3. The expected profit of \$.16 from mailing each household in this segment is computed by SI-CHAID as follows:

$$.0092 \times (\$35) + .0018 \times (-\$7) + .9889 \times (-\$0.15) = \$0.16$$

- ▷ Click the X in the upper right of the gain-chart to close it



To display the expected profit in each node of the tree rather than the percentages for paid, unpaid and non-responders:

- ▷ Right click in any node of the tree diagram
- ▷ Select 'node items' from the pop-up menu
- ▷ Click the box to the left of 'Score'

A check-mark appears in this box.

To remove the percentages from each node of the tree:

- ▷ Click the box to the left of 'Percents'

The check-mark disappears from this box.

- ▷ Click 'Close'

USING SI-CHAID TO IDENTIFY PROFITABLE SEGMENTS

The revised tree display is as follows:

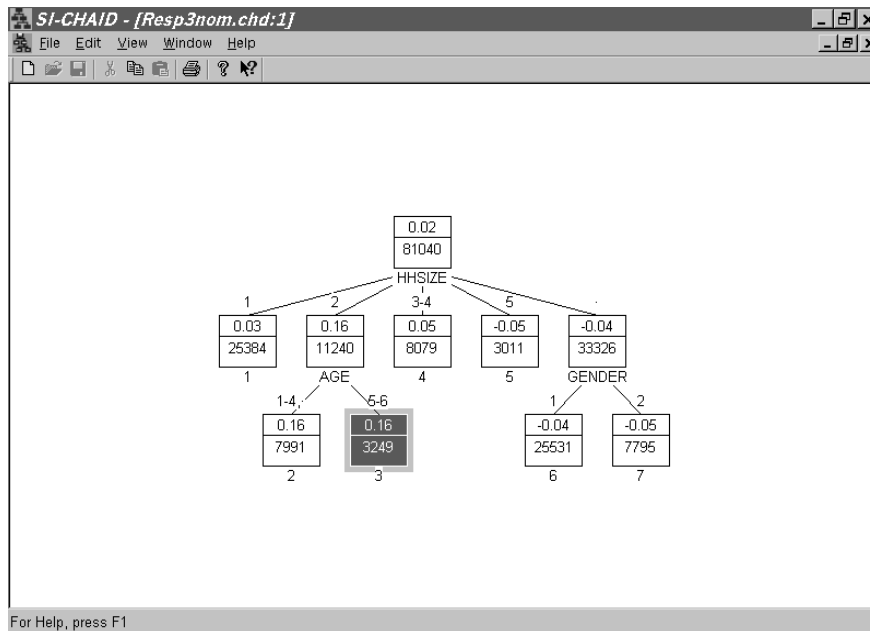


Figure 33. Tree Diagram showing Average Scores

ORDINAL METHOD

We will now reanalyze these data using the same category scores but we will use the ordinal method, which treats the dependent variable as ordinal.

- ▷ Return to ChaidDefine and double-click on "Model 1" in the left pane.

The Model Analysis Dialog Box pops up

- ▷ Right-click on RESP3 in the Dependent variable box and select Ordinal from the pop-up menu
- ▷ Click Explore
- ▷ Enter the filename RESP3ord.chd so as to not replace our earlier analysis file RESP3nom.chd
- ▷ Click Save

The following tree diagram is displayed:

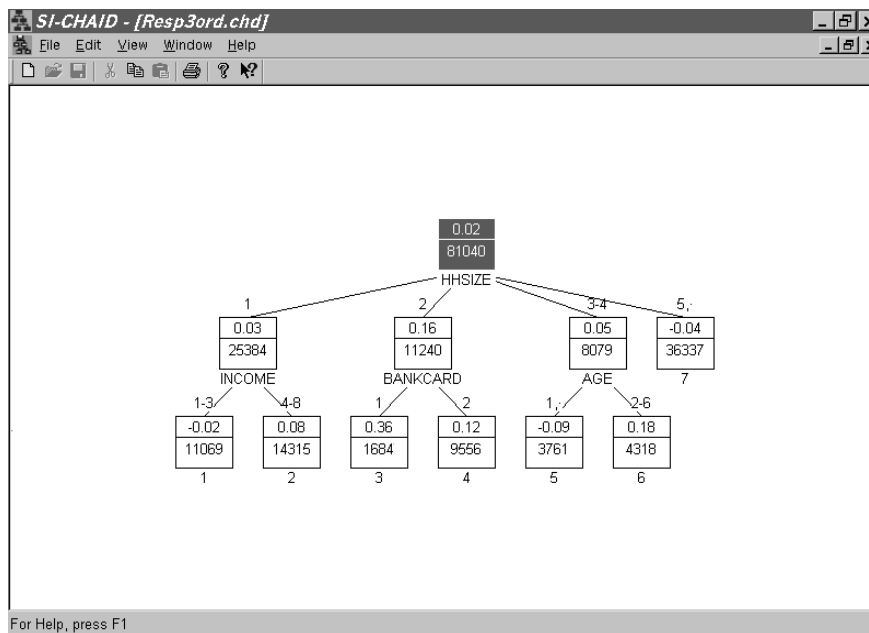


Figure 34. Tree Diagram obtained using Ordinal Algorithm

To display the Nominal and Ordinal segmentation trees side-by-side:

- ▷ Select 'Tile Vertical' from the Windows menu

Note that two-person households are now split based on whether they own a bankcard rather than based on Age, and that the expected gain for two-person households that own a bankcard (0.36) is three times greater than the expected gain for two-person households that do not own a bankcard (0.12).

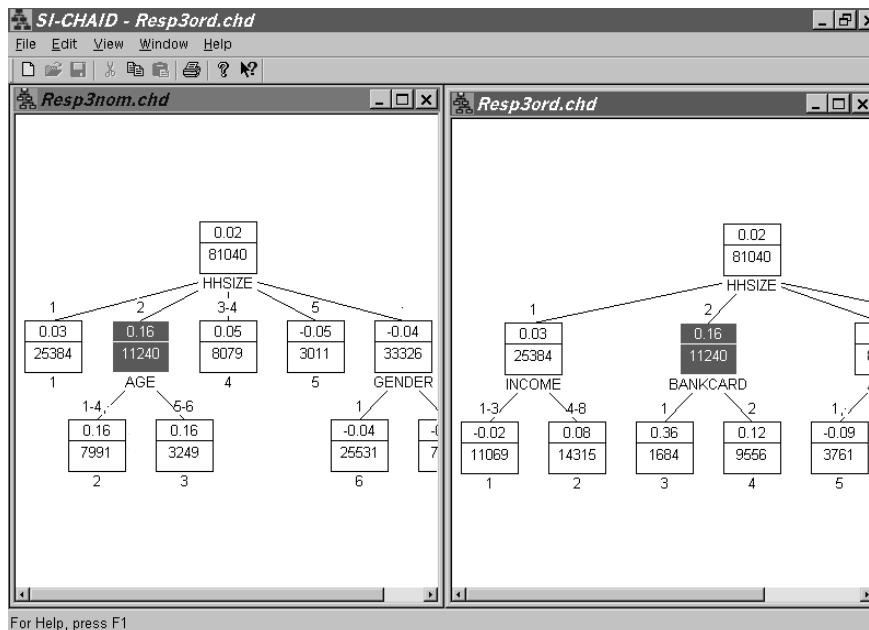


Figure 35. Tree Diagrams for Nominal vs. Ordinal Algorithms side-by-side

USING SI-CHAID TO IDENTIFY PROFITABLE SEGMENTS

Return to the nominal segmentation and click on the node corresponding to HHSIZE =2

- ▷ Right-click and choose 'Select'

Notice that only a single predictor, AGE, is listed as a candidate for splitting this subgroup using the nominal method. The nominal test of significance is not powerful enough to identify the important BANKCARD effect. By taking into account the profitability scores, the ordinal test of significance utilizes only a single degree of freedom. Thus, it provides a more powerful test of significance and a better segmentation model than the nominal method (For further details, see Magidson, 1994).



To compare gains charts from the different segmentations:

- ▷ Click in the Window of the nominal segmentation tree to make it active
- ▷ Click on the root node to make it the current node
- ▷ Select New Gains from the Windows menu
- ▷ Right-click on this gains chart and select Gains Items from the pop-up menu
- ▷ Select Summary to display the quantile format and change the default to 5 percentile units
- ▷ Click Close to close this Window

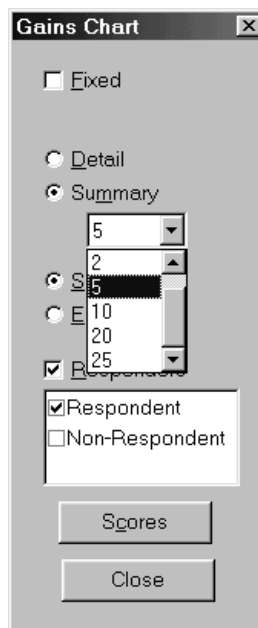


Figure 36. Gains Chart Control Panel



Repeat these steps to obtain a corresponding gains chart for the ordinal segmentation tree:

- ▷ Click in the Window of the ordinal segmentation tree to make it active
- ▷ Click on the root node to make it the current node
- ▷ Select New Gains from the Windows menu
- ▷ Right-click on this gains chart and select Gains Items from the pop-up menu
- ▷ Select Summary to display the quantile format and change the default to 5 percentile units
- ▷ Click Close to close this Window.
- ▷ Rearrange the gains Windows to present them side-by-side:

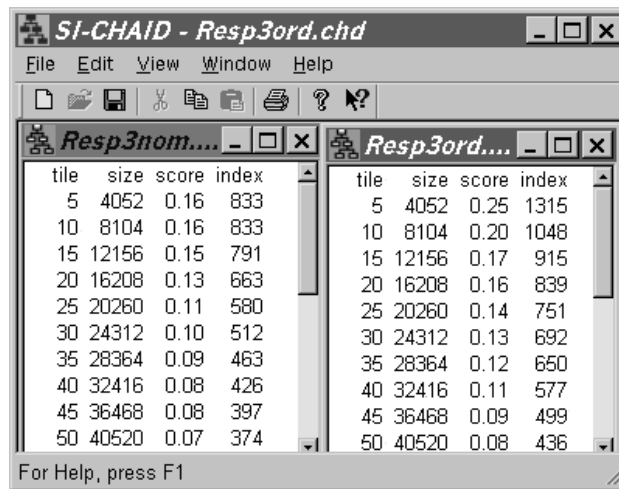


Figure 37. Two Gains Charts side-by-side

Comparison of these gains charts show that the ordinal segmentation would be expected to outperform the nominal segmentation for mailings involving profitable segments (less than 50% of all cases). Hence, by taking into account the profitability scores, the ordinal algorithm provides a more profitable segmentation.

Note: If the node corresponding to HHSIZE=2 is the current node for each tree as in Figure 35, the gains charts comparison will be based on the parent node.

Tutorial 3: Using SI-CHAID with a Hold-out Sample

Sometimes cases on the analysis file are randomly assigned to a 'hold-out' sample and not used in the development of the segmentation tree. Instead, such cases are reserved for the purpose of 'validating' the tree. In this tutorial we utilize the data file holdout.sav to illustrate the use of SI-CHAID in this way.

In particular, from each dependent category ('paid respondents', 'unpaid respondents' and 'non-responders') we randomly assigned each case in the 'subscrib.sav' file to one of two equally likely groups by generating the variable SAMPLE (1=test, 2 = holdout).

| | sample | age | gender | kids | income | bankcard | hhsz | occup | resp3 | resp2 |
|----|--------|-----|--------|------|--------|----------|------|-------|-------|-------|
| 1 | 1 | 1 | 1 | 2 | 7 | 2 | 1 | 4 | 1 | 1 |
| 2 | 2 | 1 | 1 | 2 | 4 | 2 | 1 | 4 | 1 | 1 |
| 3 | 2 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 1 | 1 |
| 4 | 1 | 1 | 1 | 2 | 3 | 1 | 1 | 4 | 1 | 1 |
| 5 | 2 | 1 | 2 | 2 | 4 | 2 | 2 | 3 | 1 | 1 |
| 6 | 1 | 1 | 2 | 2 | 5 | 2 | 1 | 4 | 1 | 1 |
| 7 | 1 | 2 | 1 | 1 | 7 | 1 | 4 | 2 | 1 | 1 |
| 8 | 1 | 2 | 1 | 1 | 6 | 2 | 5 | 3 | 1 | 1 |
| 9 | 1 | 2 | 1 | 1 | 4 | 2 | 5 | 2 | 1 | 1 |
| 10 | 1 | 2 | 1 | 1 | 1 | 2 | 3 | 3 | 1 | 1 |
| 11 | 2 | 2 | 1 | 2 | 8 | 2 | 2 | 1 | 1 | 1 |
| 12 | 2 | 2 | 1 | 2 | 8 | 2 | 2 | 1 | 1 | 1 |
| 13 | 1 | 2 | 1 | 2 | 8 | 1 | 2 | 1 | 1 | 1 |
| 14 | 2 | 2 | 1 | 2 | 7 | 2 | 2 | 4 | 1 | 1 |
| 15 | 1 | 2 | 1 | 2 | 7 | 1 | 4 | 4 | 1 | 1 |
| 16 | 2 | 2 | 1 | 2 | 6 | 2 | 2 | 4 | 1 | 1 |
| 17 | 1 | 2 | 1 | 2 | 4 | 2 | 2 | 3 | 1 | 1 |
| 18 | 2 | 2 | 1 | 2 | 3 | 2 | 2 | 3 | 1 | 1 |
| 19 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 4 | 1 | 1 |
| 20 | 1 | 2 | 1 | 2 | 2 | 2 | 3 | 3 | 1 | 1 |

Figure 38. Holdout.sav file

In this tutorial we will use this data file to grow a segmentation tree on the test file and see how well it validates on the holdout sample. This will be accomplished using the following steps:

- Use the 'First predictor' option to force the variable SAMPLE (test vs. holdout) to yield the first split
- Use the 'auto' option to grow the tree only on the SAMPLE = test group
- Save the resulting tree
- Apply the saved tree to the SAMPLE = 'holdout' group
- Compare gains-charts for the test and holdout samples

▷ From the Define program, select File Open 'holdout.chd'

Your display should now look like Figure 39. Note that the options shown in the Contents Pane indicate that the tree will be grown using the file 'holdout.sav' with the First Predictor option and the Ordinal method.

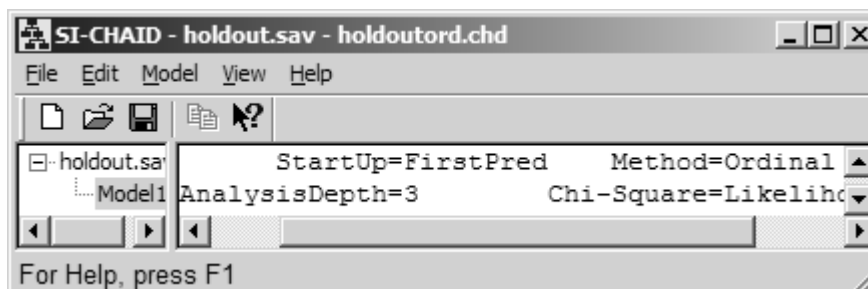


Figure 39. Holdout.sav in Chaid Define



To open the analysis dialog box:

- ▷ From the Model menu select 'Edit' (or double click on 'Model1')
- ▷ Click Scan

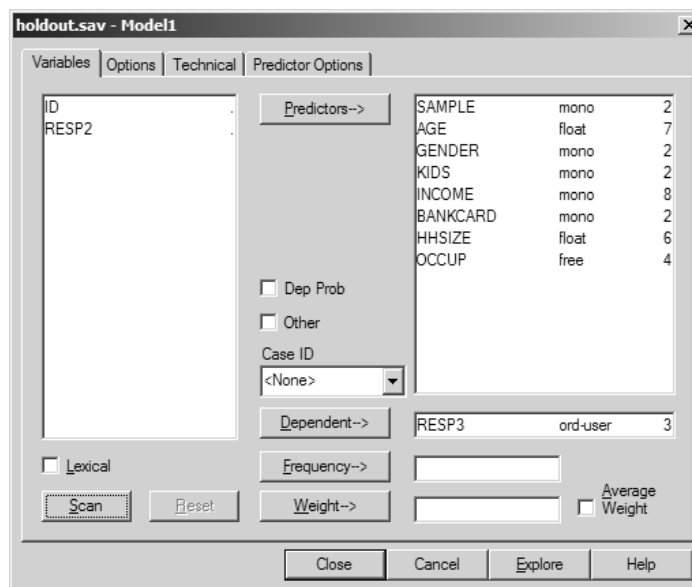


Figure 40. Analysis Dialog Box for Holdout.sav

Note that the dependent, predictor variables and scale types are identical to that used in the ordinal model developed in Tutorial #2, except that the new variable SAMPLE is used as the first predictor.

- ▷ Click 'Options' to open the Options tab

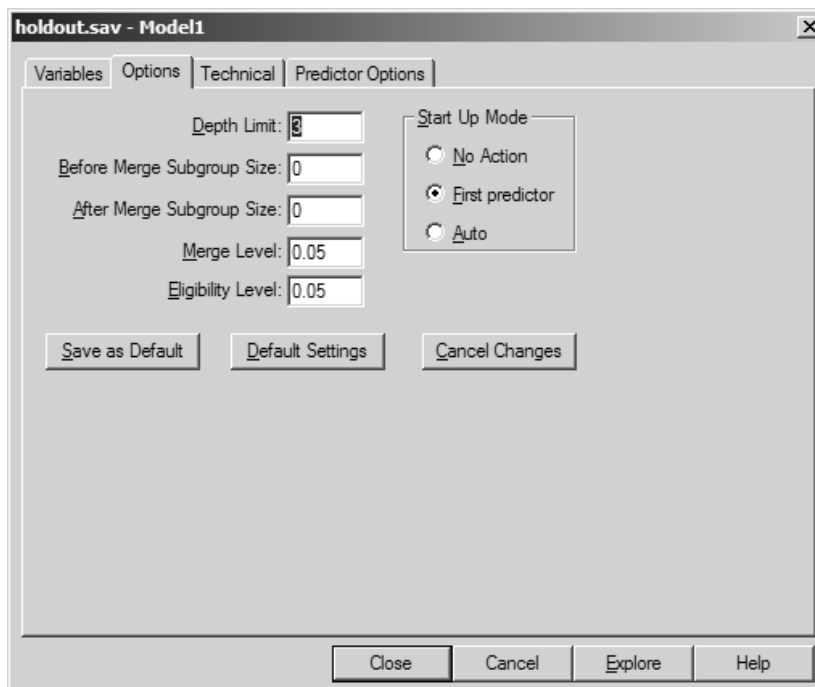


Figure 41. Options Tab for Holdout.sav

The 'First Predictor' option means that the categories of the first predictor variable SAMPLE will be used to define the initial CHAID split. This is indicated in the Start-Up Mode box.

- ▷ Click Explore
- ▷ When prompted, enter the file name 'holdout.chd'
- ▷ Select Yes, to replace the current file of the same name

The Explore program opens and grows the tree to one level, using the 2 categories of SAMPLE as shown below.

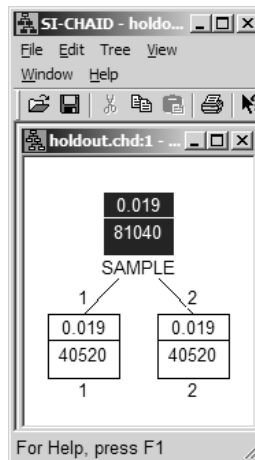


Figure 42. Tree Diagram for SAMPLE

The contents of the nodes shows that both the SAMPLE = 1 (test group) and SAMPLE = 2 (holdout group) consist of exactly half of the cases (N=40,520), each having an average profit of \$.019 per case.



To grow the tree within the test sample,

- ▷ Click on node 1
- ▷ From the Tree menu, select auto

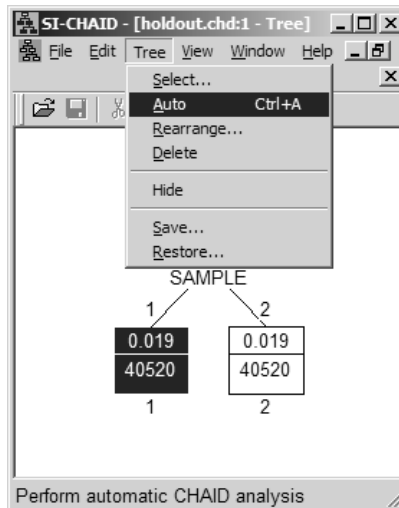


Figure 43. Selecting Auto from the Tree menu

USING SI-CHAID WITH A HOLD-OUT SAMPLE

The resulting tree consists of 5 segments, numbered 1-5. Segment #2 shows the highest profit (\$.467), followed by segment # 4 (\$.237), segment #3 (\$.102), segment #1 (\$.043) and segment #5 (-\$.061).

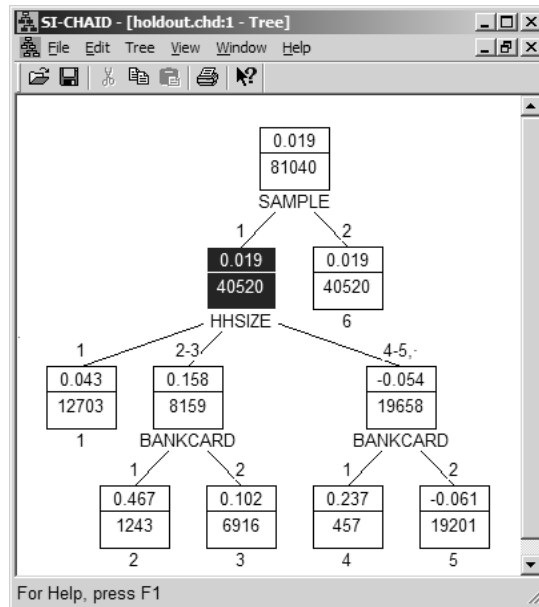


Figure 44. 5 segment Tree Diagram



One way to apply this tree to the holdout sample is to

- ▷ Select Edit → Copy
- ▷ Click on node #6
- ▷ Select Edit → Paste

An alternative approach is to save the tree to a file and then restore it to the holdout sample



To save the tree in Figure 44 corresponding to `SAMPLE=1`,

- ▷ from the Tree menu, select Save
- ▷ when prompted for a file name, enter '5segments.ctf'
- ▷ Click Save

The CHAID tree file '5segments.ctf' is saved

To apply this tree to the holdout sample,

- ▷ click on node #6
- ▷ from the Tree menu, select Restore
- ▷ When prompted for a file, select '5segments.ctf'

