

# 4

## ***Beginning a GOLDMineR Analysis***

This section contains two tutorials that will help you become familiar with the basic features of GOLDMineR.

- Tutorial #1 demonstrates how to obtain and interpret effects plots and related output listings.
- Tutorial #2 illustrates the Model Search procedure and the Specialized Charts feature.

### ***Tutorial #1***

In this tutorial we use sample dataset #4 (see Table 4-1) and show how to specify a model and obtain an effects plot. You will:

- open a previously saved data file
- select the dependent and predictor variable
- experiment with different scaling types
- generate plots and tables
- use various display options

Further illustrative analyses on these data can be found in Chapter 10.

Table 4-1  
Clinical Trial Data\*

TRTMNT	IMPROVE					Fixed X-scores
	Worse	Stationary	Slight	Moderate	Marked	
Test Drug	1	13	16	15	7	1
Placebo	5	21	14	9	3	0
Fixed Y-scores	-1	0	1	2	3	

\*(Source: DeJonge, H., 1983, 'Deficiencies In Clinical Reports For Registration Of Drugs', *Statistics in Medicine*, Vol. 2, 155-166.)

## Opening a previously created array file

For information on how the array was created and saved, see Chapter 5.

- ▶ To open the previously saved array file from the menus choose:

File

Open

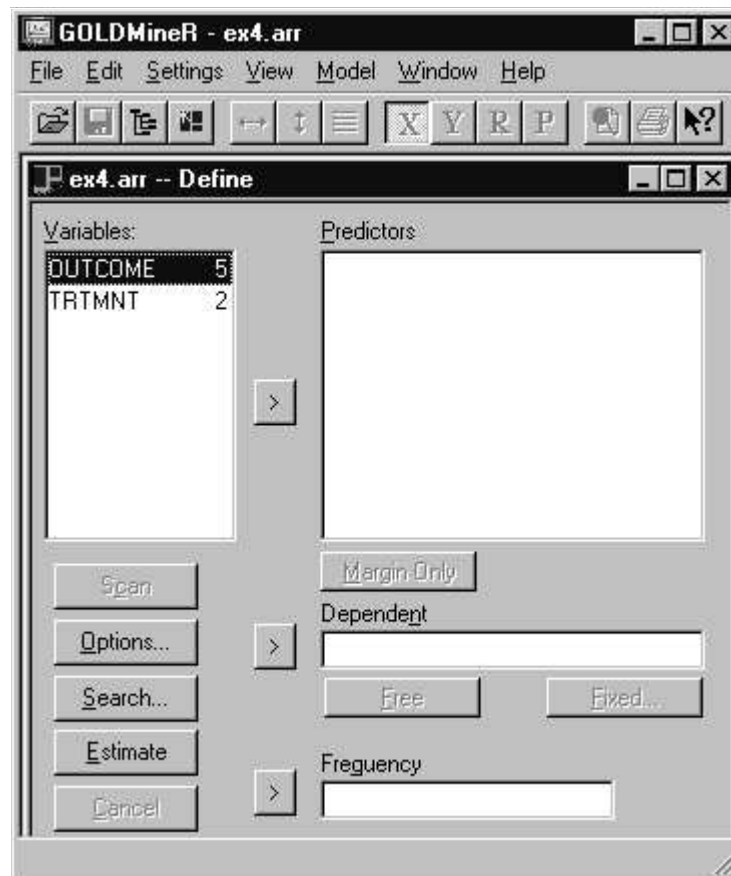
GOLDMineR places you in the Open dialog box. Click on the arrow next to Files of type and select array files (.ar\*) if this is not already the default listing. A list of all files with .arr extensions will now appear in the selection box (See Figure 4-1). If you copied the tutorial files into a directory other than the default directory, switch to that directory prior to retrieving the file.

Figure 4-1  
File Open dialog box



Highlight `ex4.arr` and click on `Open` to open the Define Model dialog box, shown in Figure 4-2.

Figure 4-2  
The Define Model Dialog Box



### Selecting the Dependent Variable

For this analysis, the 5-category variable OUTCOME will be the dependent variable. To select the dependent variable:

- ▶ Click on OUTCOME in the Variables box.

- ▶ Click on the > key next to the Dependent variable box
- ▶ or double click on OUTCOME (double clicking here only works if no other variable has already been designated as the dependent variable).

The designated dependent variable, OUTCOME, now appears in the dependent variable box. (If you mistakenly select a variable other than OUTCOME as the dependent variable, click on that variable in the Dependent variable box, and then click on the reverse arrow key to return it to the variables box.)

### **Selecting Predictor Variables**

To select the predictor variable:

- ▶ click on TRTMNT in the variables box
- ▶ Click on the > key to the left of the Predictors box
- ▶ or, after selecting the dependent variable, you may double click on TRTMNT and it will be moved to the predictor box.

(If you mistakenly select a variable that you do not wish to be a predictor, click on that variable in the Predictor variable box, and then click on the reverse arrow key to return it to the variables box.)

### **Assigning Category Scores**

Each dependent and predictor variable must be specified as **Free** or **Fixed** depending upon whether it is to be treated as qualitative, in which case category scores for that variable will be estimated, or treated as quantitative by assigning fixed scores. By default, GOLDMineR will select the **Fixed** scaling option for a given variable if quantitative values are used in the input data file or fixed category scores are assigned in an 'array file' for that variable. Fixed category scores (called X-scores and Y-scores) have been specified for the variables in ex4.arr using the 'Score' option (see Table 4-1 for scores or the detailed instructions on creating this array in Chapter 5).

We will first estimate a model using the fixed category scores that are specified in our input file.

Figure 4-3  
Model Define dialog box with options set



## Fitting a Model

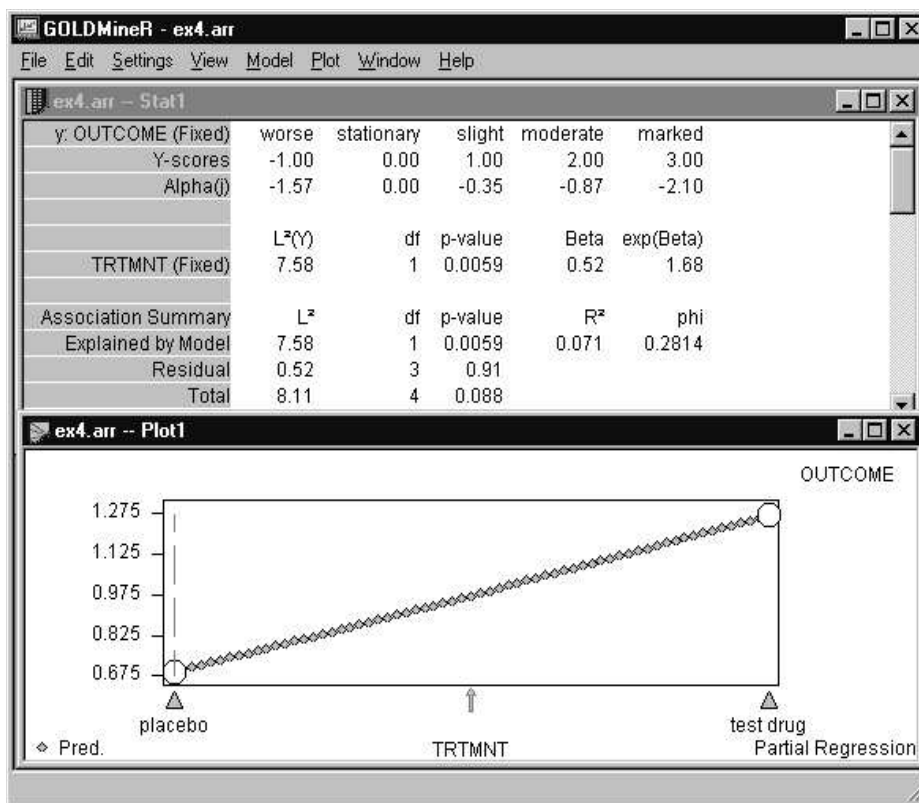
Now that we have defined our variables and have settled on our scaling options, we can estimate a model.

- To estimate a model from within the Model Define dialog box, click on Estimate .

GOLDMineR checks first that the parameters are identifiable and then begins to estimate them. When completed, two windows will appear on the screen as in Figure 4-4:

- a Statistics window (top)
- a Plot window which contains a partial regression plot (bottom).

Figure 4-4  
Initial Statistics and Plot Window for Model 1



### Interpreting the Information in the Statistics Window

The fixed scores ("Y-scores") assigned to the outcome categories for the model estimated are listed at the top of the Statistics window. Since the Y-scores are equidistant, the model specifies that the spacing between adjacent outcomes is equal.

The equidistant Y-score assumption may be formalized by the following three distinct restrictions:

$$y_5 - y_4 = y_4 - y_3$$

$$y_4 - y_3 = y_3 - y_2$$

$$y_3 - y_2 = y_2 - y_1$$

the validity of which is tested by the Residual  $L^2$  model fit statistic with 3 degrees of freedom (1 degree of freedom for each distinct restriction).

The Statistics window shows *Residual  $L^2$*  = 0.52 and the corresponding *p-value* is 0.91. The smaller the value of Residual  $L^2$ , the larger the corresponding p-value and the better the fit of the model assumptions to the data. The *p-value* of 0.91 suggests that the equidistant spacing assumption provides an excellent fit to these data. (Generally, a p-value greater than 0.05 is interpreted as providing a fit that is adequate.)

After verifying the model assumptions fit the data, we next examine the effect of the treatment. The statistic  $L^2(Y) = 7.58$  with 1 *degree of freedom* ( $p = .0059$ ) is used to test the hypothesis  $H_0: \text{Beta} = 0$  (or the equivalent  $H_0: \exp(\text{Beta}) = 1$ ). Since  $p < .05$ , we see that the effect of TRTMNT, estimated by  $\exp(\hat{\beta}) = 1.68$  is statistically significant at the .05 level (and even at the .01 level). Later, when we examine an Effects plot, we will see how the treatment effect of 1.68 can be interpreted in terms of an odds ratio.

## Regression Plot Features

The graphical display that appears by default in the Plot window is the Partial Regression view, which plots the regression predictions for Y as a function of X.

Within the Plot window, click on the circle above the *placebo* marker, which represents the observed average Y-value for patients receiving the *placebo*. The following message appears on the status bar in the lower left corner of your screen: Observed Average: TRTMNT[placebo] (avg=0.69, res=0.000). The message means that

- the average Y-value observed for persons given the placebo is 0.69 (falling between the scores for *stationary* and *slight* improvement), and
- the difference between this value and the corresponding predicted Y is 0 (adjusted residual = 0).

Click it again and a diamond symbol hidden behind the circle appears, which represents the predicted Y for persons given the placebo and the message in the status

bar changes to: Expected  $Y[x=0.00]=0.69$ . Thus, the predicted Y-value is identical to the observed average Y-value. Click on the circle and diamond above the *test drug* marker to also confirm that the predicted Y-value equals the corresponding observed average Y-value for persons receiving the *test drug*.

By default, the regression plot provides predictions for 63 X-values spaced equally between the lowest and highest X-scores. In Figure 4-4, the low point corresponds to the *placebo* ( $x=0$ ) and the high to the *test drug* ( $x=1$ ).

### Changing the default number of plotted points

- ▶ Click the right mouse button from inside the Plot window to bring up the plot control dialog box.

Figure 4-5  
Plot Control dialog box



- ▶ Highlight '63', type in a new number, '3'
- ▶ Click on Update.

The plot is now updated to reflect the reduced number of predicted Y diamond shaped points.

- ▶ Click on **Linear** and GOLDMineR overlays the linear regression line onto the plot.

For this simple example, Figure 4-4 (which plots the predicted OUTCOME score as a function of TRTMNT) is not very informative since TRTMNT contains only two categories and hence predictions between these two categories are generally not meaningful. A more useful display for this example is the Effects plot, which we will examine next.

### ***Obtaining an Effects Plot***

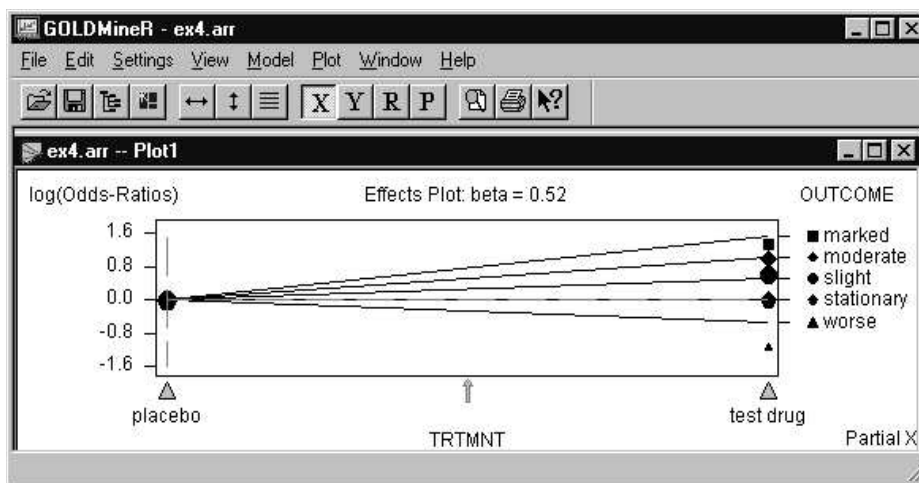
The Effects plots show the effect of the predictor(s) in the generalized logit model. There are two types of effects plots, partial and joint. For this example, because there is only one predictor variable, the partial and joint plots are identical. There are two views for each type of effects plot, X-view and Y-view. In an effects plot, the effects (estimated log-odds ratios) are plotted as a function of X (the X-view) or as a function of Y (the Y-view).

To obtain an Effects plot in the X-view:

- ▶ click anywhere within the Plot window to make it the active window (you'll know its active when the Plot Menu appears on the menu bar).
- ▶ Click on the 'X' speedbutton located on the toolbar.

An X-view effects plot replaces the regression plot in the Plot window. In this plot we see effect lines for each of the 5 outcomes plotted as a function of TRTMNT. The categories are represented by markers on the horizontal and vertical axes.

Figure 4-6  
Partial X Effects Plot with Log-Odds Ratios



### Changing the scale from Log-odds Ratios to Odds Ratios

The units displayed to the left of the vertical axis in Figure 4-6 correspond to log-odds ratios.

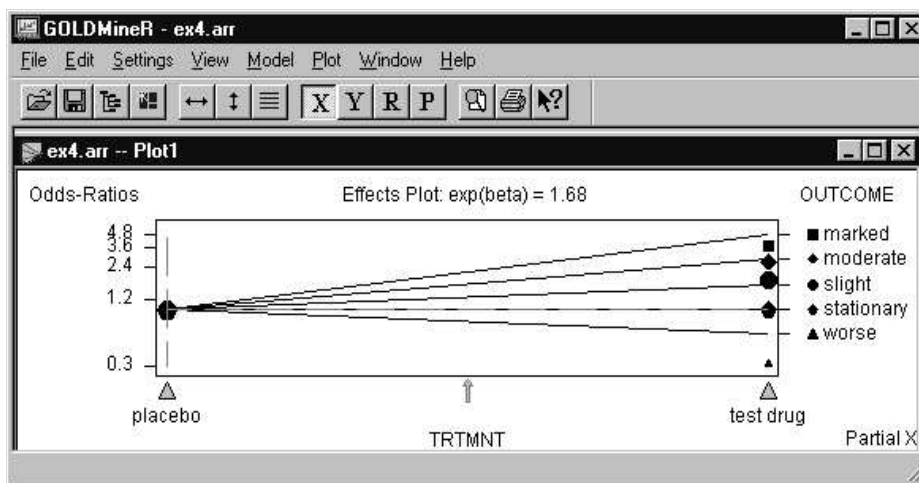
- To translate the log-odds ratios in the plot to the more easily interpretable odds ratios from the menu choose:

View

Standard

The plot itself is unchanged, but the units displayed to the left of the plot are now odds ratios. For the remainder of this tutorial, all plots and tables will now contain the standard (odds-ratio) units unless you change back to log-odds ratios by choosing View, Logarithmic. Your Plot window now looks like Figure 4-7.

Figure 4-7  
Partial X Effects Plot with Odds Ratios



We are now going to open up a Table window in order to help you obtain an accurate interpretation of the effects in terms of odds ratios.

## Generating Tabular Output

For each effects plot there is a corresponding table that can be opened to obtain increased precision in reading the desired effect estimate.

- ▶ To open a Table window from the menus choose:

Window

New Table

(If you have estimated more than one model during the current session, the Estimated Models dialog box will appear which allows you to select the table that corresponds to the appropriate model.) A Table window containing observed frequency counts (the same as Table 4-1) now appears. (The observed frequency counts can be checked against those given in Table 4-1 for accuracy.)

Figure 4-8  
Table containing Observed Frequency Counts

		TRTMNT			
		placebo	test drug		
OUTCOME	Y-scores	weights	1.00	0.00	X-ref (cell)
marked	3.00	0.00	3	7	3.00 f
moderate	2.00	0.00	9	15	9.00 f
slight	1.00	0.00	14	16	14.00 f
stationary	0.00	1.00	21	13	21.00 f
worse	-1.00	0.00	5	1	5.00 f
		Y-ref	21.00	13.00	21.00 f
average	score		0.69	1.27	104 .

A table can contain either observed or estimated expected frequency counts, probabilities, odds or odds ratios -- in standard or logarithmic units. For this example, we wish to replace the observed frequency counts with estimated expected odds ratios. To change the table so that it contains estimated expected odds ratios click anywhere in the Table window to make it the active window and then choose:

Table

Control

or click on the right mouse button to bring up the table control dialog box.

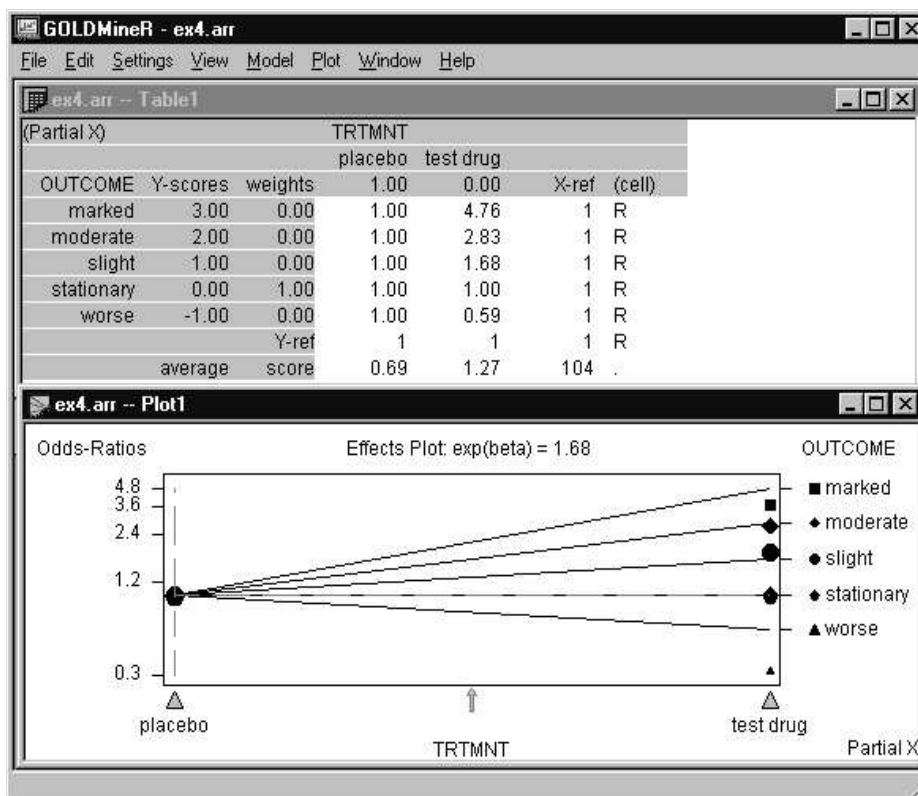
Figure 4-9  
Table Control dialog box



The default table contains observed frequency counts (Frequencies is marked with a check in the Observed options box). To view the expected odds ratios estimated by the model click on Frequencies in the Observed option box to remove the observed frequencies and then click on Ratios in the Expected option box to insert the estimated odds ratios.

Your Table window and Plot window should now look like Figure 4-10 (you may have to manually arrange the windows to see both at once).

Figure 4-10  
 Partial X Table (containing odds ratios) and corresponding Partial X Effects Plots



### Interpreting the X-View Effects Plot with Table Backup

The X-view of the effects plot contains an effects line associated with each OUTCOME category, the slope of which equals the estimated treatment effect with respect to attaining the  $j^{\text{th}}$  outcome,  $\beta^* y_j^*$ . Since the Y-score for OUTCOME = *stationary* is 0 (i.e.,  $y_2 = 0$ ), the effects line for *stationary* coincides with the horizontal axis (i.e., the slope of the *stationary* effects line equals 0) and serves as a reference for measuring the treatment effect. For example, the effect of the treatment on the odds of attaining a *marked* (vs. *stationary*) improvement can be read from the effects plot as follows:

- The odds of a *marked* improvement is about 4.8 times as likely for patients receiving the *test drug* rather than the *placebo*. More precisely, the estimate obtained using the table is 4.76.

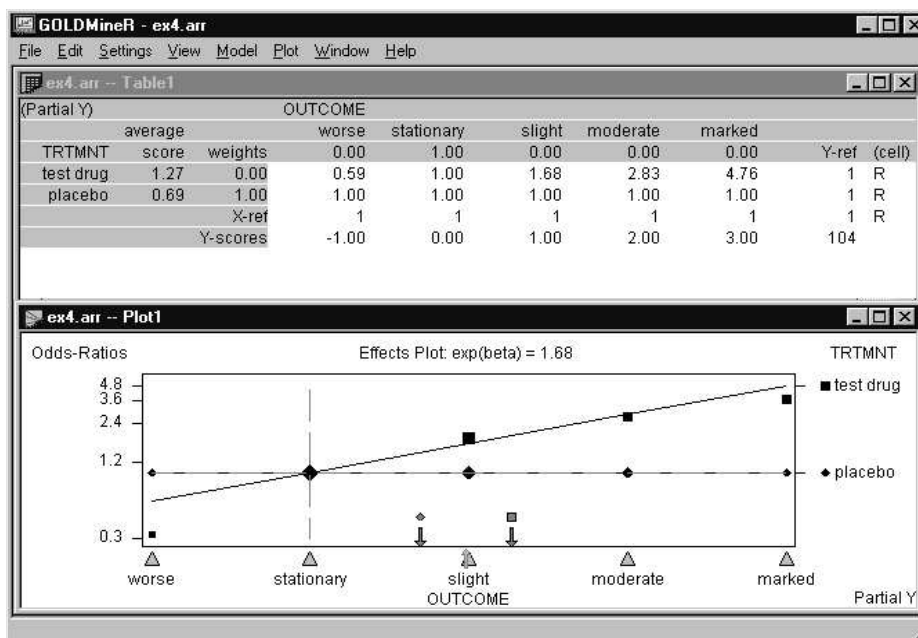
The OUTCOME categories are displayed along the side of the Effects plot with a symbol used to identify the corresponding observed odds ratio. The TRTMNT categories are labeled at the bottom of the effects plot below the associated marker.

To obtain an Effects plot in the Y-view click anywhere within the Plot window to make it the active window and click on the ‘Y’ speedbutton. A Y-view Effects plot will replace the X-view Effects plot in the Plot window. Also, you may wish to click in the Table window and click on the ‘Y’ speedbutton once again to transpose the rows and columns so that the Table changes from X-view to Y-view. Your Table window and Plot window should now look like Figure 4-11 (you may have to manually arrange them to see both at once).

### ***Interpreting the Y-view Effects Plot with Table Backup***

In the Y-view plot we see that an effects line is associated with each of the two types of TRTMNT, the slope of which equals  $\bar{\beta}_{x_i^*}$ . Since the  $X^*$ -score for the *placebo* category is zero (i.e.,  $x_2^* = 0$ ), the *placebo* effects line coincides with the horizontal axis and serves as a reference for measuring the treatment effect. Since the  $X^*$ -score for the *test drug* category is 1, the *test drug* effects line has slope equal to  $\exp(\text{Beta}) = 1.682$ . By examining the Table window we see that while we now have a different view of the effects, the odds ratios effect estimates themselves remain unchanged.

Figure 4-11  
 Partial Y Table (containing odds ratios) and corresponding Partial Y Effects Plot



The categories of OUTCOME are labeled at the bottom of the Effects plot beneath triangular shaped markers. The markers are equidistant from each other corresponding to the equidistant Y-scores assumed by the model. The good fit of the equidistant spacing assumption is supported by the fact that the observed odds ratio symbols appear close to the effects line which shows the expected odds ratios estimated under the model.

Since the Y\*-score for *marked* improvement is 3, the odds of having a *marked* improvement (vs. *stationary*) is about  $\exp(3\beta) = 4.76$  times as high for patients who received the *test drug* than for patients who received the *placebo*. As we saw earlier, the effect estimate of 4.76 can be approximated using the *test drug* effects line or obtained more accurately from the corresponding table.

#### **Predicted Y Arrows (available only in the Y-view)**

Along the bottom of the Partial-Y Effects plot are two downward-pointing arrows which represent the predicted outcomes for patients according to their TRTMNT

category. A symbol above the arrow identifies the TRTMNT category (see the symbols listed next to the category labels on the side of the plot). To obtain the specific values of the predictions click on a particular arrow and a message will appear on the status bar (the values can also be found using the Table window). Just by looking at the plot, you can see that the predicted outcome for *test drug* is located between *slight* and *moderate*, while the predicted outcome for placebo is between *stationary* (no change) and *slight*. In addition to the beta estimate, the distance between these two prediction arrows is another way to quantify the effect of the treatment.

The predicted values for patients receiving the *placebo* and *test drug* obtained by clicking on the arrows are the same as those we saw earlier when we examined the partial regression view (recall Figure 4-4).

### **Changing the Meaning of the Odds of Improvement by Changing the Reference Point**

The effect of the treatment on the odds of a *marked* improvement is defined in the current effects plot relative to the *stationary* outcome which serves as the reference for computing the odds. Click within the Table to make it the active window and then right click to open the Table control dialog box. Click on *expected odds* to add the expected odds to the table. Verify that the odds for *stationary* are equal to 1 which confirms that the *stationary* OUTCOME category is used as the Y-reference base for computing the odds.

We will now show how the odds, odds ratios and effects plot change when we change the Y-reference from *stationary* to some other outcome.

Note that the current origin of the plot in Figure 4-11 (the intersection of the axes) corresponds to the TRTMNT = *placebo*, OUTCOME = *stationary* reference point (each axis is represented in the effects plot by a series of straight dashed lines). The origin of the plot provides a visual representation of the baseline references used in defining the odds and odds ratios. The default setting for the reference point was established by assigning scores of '0' to the desired reference categories in the .arr file (see Table 4-1 for scores).

In the current plot, the horizontal axis corresponds to the X-reference and the vertical axis to the Y-reference. In general, each axis (reference) corresponds to a category of a variable, the mean of the variable, or some other weighted average of the categories. The weights assigned to each category of a variable may be customized by specifying category weights for that variable which consists of a set of numbers between 0 and 1 that sum to 1, one number for each category.

To illustrate the effect of changing the reference point, we are first going to use the OUTCOME category *worse* in place of *stationary* as the outcome reference. While changing the reference changes the perspective from which you view the effects plot (i.e., shifts the axis), it does not affect *Beta* or the *Predicted Y*.

To change the outcome reference choose:

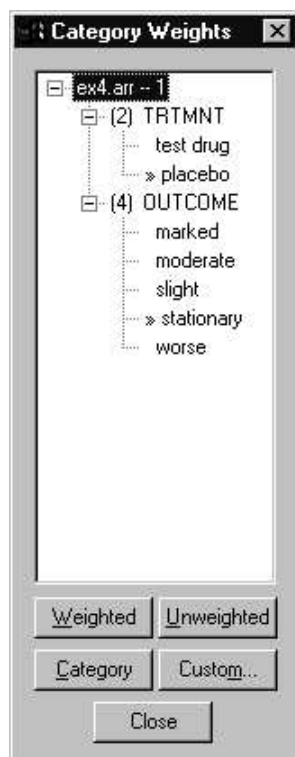
Model

Category Weights

or click on the Category Weights symbol ( ) on the toolbar.

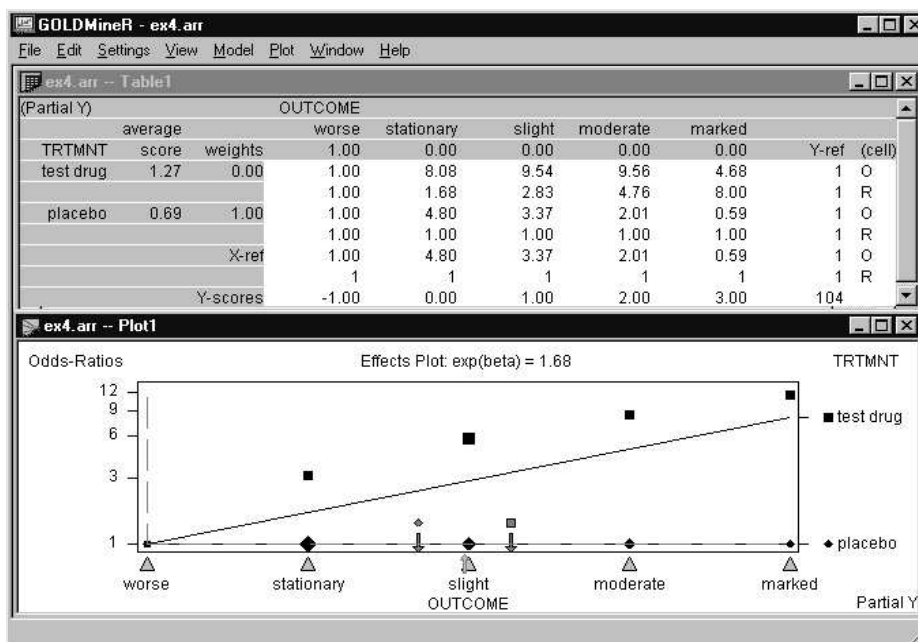
The Category Weights dialog box will appear with a list of each of the variables and its categories. A symbol identifying the current category weighting option is listed to the left of the variable name. The symbol “(k)” identifies the kth category as the reference. The k<sup>th</sup> category of the variable has a [>>] next to it.

Figure 4-12  
Category Weights dialog box



Select *worse* by highlighting it and clicking on *Category* or double click on *worse*. The [ $\gg$ ] will now be next to the *worse* category and “(5)” appears next to *OUTCOME* since *worse* is the 5th *OUTCOME* category. Alternatively, you may accomplish the same change to the Y-reference by double clicking on the triangular marker for *worse* in the Effects plot. Note that the odds in the table change so that the odds for *worse* now equals 1. Also note that in the Effects plot (Figure 4-13) the dashed vertical line (representing the Y- reference) is now aligned with the *worse* category marker.

Figure 4-13  
 OUTCOME reference changed to Worse



The odds of marked improvement (vs. worse) are about 8 times as high for patients who received the test drug than for those receiving the placebo. Note that the estimates for Beta and the predicted average score arrows are not affected by changes in the references. The odds and odds ratios have been redefined based on the new Y-reference.

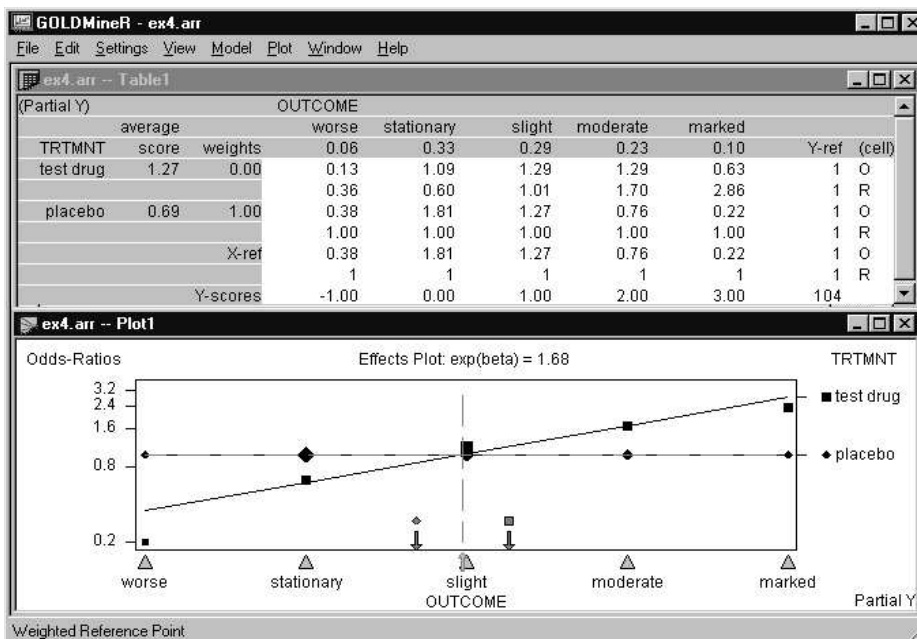
## Selecting Weighted Average References

Next, we're going to change the Y-reference to be based on the mean (i.e., the weighted average of the categories) using the "weighted average" option, which weights outcome  $j$  by the observed sample proportion for that outcome.

To change the Y-reference to be based on the mean, within the category weight dialog box, highlight the variable name OUTCOME, click on Weighted or double click on OUTCOME. A 'W' appears next to OUTCOME. Alternatively, you may double click on the upward pointing arrow at the bottom of the Y-view Effects plot which serves as the mean outcome marker.

The dashed vertical Y-reference line is now aligned with the mean outcome marker (Figure 4-14), which is just to the left of the *slight* marker. The odds and odds ratios as displayed in the graph change accordingly. The category weights used to define the Y-reference appear directly below the outcome labels in the Table window.

Figure 4-14  
OUTCOME reference changed to Weighted



The odds of marked improvement (vs. the “average” outcome) are 2.86 times as high for patients who received the test drug than for those receiving the placebo.

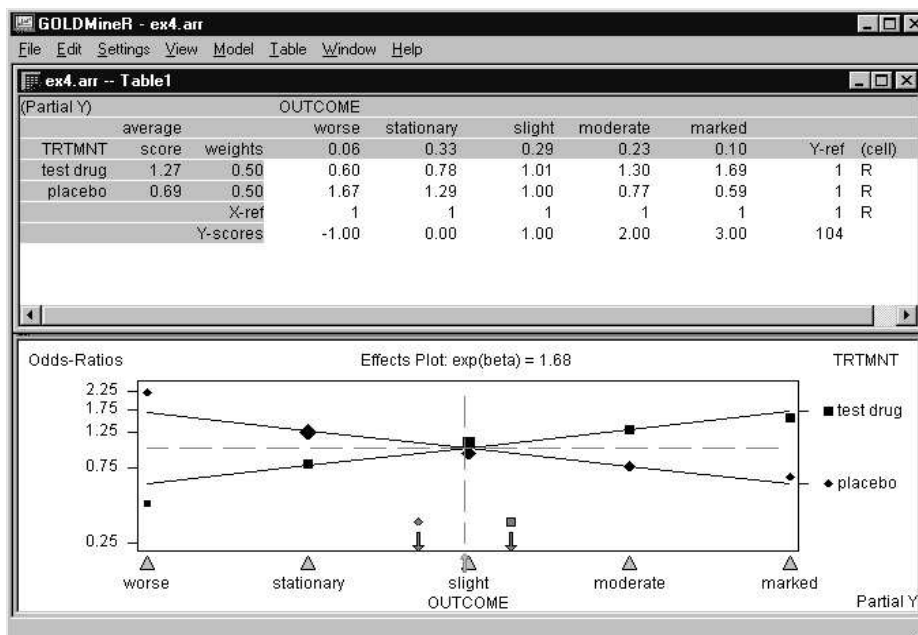
Next, we’re going to change the TRTMNT reference to be based on the mean. Since the cases in the sample are evenly divided among the treatment and placebo groups (52 in each) use of the (weighted) mean here to define the X-reference is equivalent to use of the *unweighted* mean (as in ANOVA). (For simplification, we will now remove the odds from the table. In the Table Control, click on Odds.)

To change the predictor reference for TRTMNT to be based on the mean, within the category weights dialog box, highlight the variable name TRTMNT, click on Weighted at the bottom of the box or double click on TRTMNT.

The dashed horizontal reference line is now positioned midway between the two treatment categories. The odds and odds ratios have been recalculated based on the new Y-reference.

Figure 4-15

*OUTCOME* reference = Weighted; *TRTMNT* reference = Weighted



The odds of marked improvement (vs. the “average” outcome) are 1.69 times as high for patients who received the test drug than for the average patient.

## Reversing Category ordering

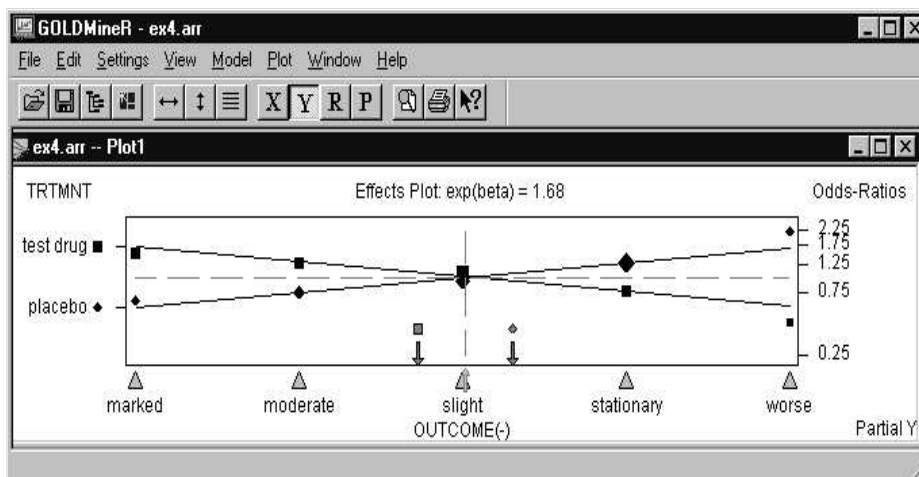
If at any time you wish to change the order of categories in the effects plot or tables (i.e., to make it easier to read), GOLDMineR allows you to reverse the order of categories in the effects plot or tables for either variable without changing the model or statistics. To reverse the order of categories choose:

View

Reverse Horizontal

to reverse the order of OUTCOME or click on the Reverse Horizontal symbol ( ) on the tool bar. The categories of OUTCOME will now be listed from *marked* to *worse* (Figure 4-16) instead of *worse* to *marked* (Figure 4-15).

Figure 4-16  
Reversing order of OUTCOME



### Obtaining Item Descriptions

Within any plot, at any time you may click on effects lines, observed points, the triangular markers and other items to receive a description of the item. Descriptions appear on the status bar in the lower left corner.

### Assessing the Treatment Effect Without Outcome Spacing Assumptions

Since our equidistant spacing assumption provided an excellent fit to these data, the current model is the final model which would be used to report results. However, for purposes of illustration, we will now change the model by setting OUTCOME to **Free** so that the relative spacing of the outcome categories will be freely estimated by GOLDMineR. (Later, we will also estimate a third model with different fixed Y-scores.)

To estimate category scores for OUTCOME choose

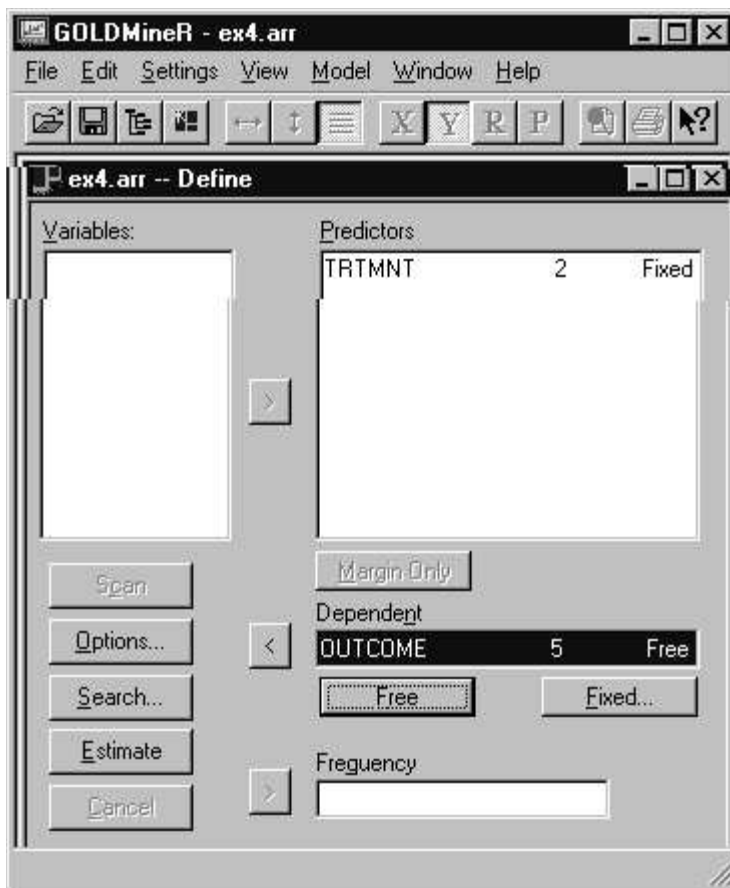
Model

Define

to bring up the Define dialog box. Highlight OUTCOME by clicking on it and then click on Free.

The word 'Free' should now appear next to OUTCOME in the dependent variable box.

Figure 4-17  
Model Define Setup for Model 2

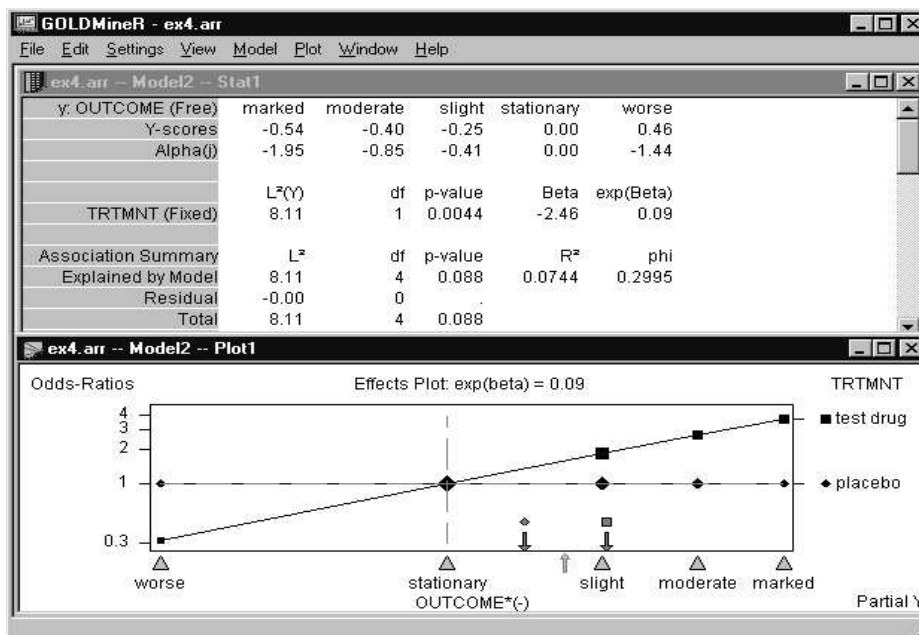


Click on Estimate and the default partial regression plot appears. To produce the graph in Figure 4-18:

- ▶ Make the plot window the active window
- ▶ Select the ‘Y’ button on the toolbar or from the Plot menu select Joint-Y
- ▶ Double click on the *stationary* marker to set the category weight for OUTCOME.
- ▶ Double click on *placebo* within the Category Weights dialog box to set the category weight for TRTMNT.
- ▶ Select ‘View, Reverse Horizontal’ to change the direction of OUTCOME

Note that the estimated Y-scores listed in the Statistics window are highest for *worse* and lowest for *marked* and the Beta estimate is negative. The interpretation of these results is identical to results obtained by multiplying each score and the Beta estimate by -1 (See Section A.5, Uniqueness of Y-score Parameters).

Figure 4-18  
Statistics and Plot Window for Model 2



The Statistics window shows that the treatment effect under the assumed model is significant and somewhat smaller in magnitude than that estimated previously. However, because the previous model fit extremely well (Residual  $L^2 = 0.524$ ), the new model does not provide a significant improvement and hence provides an overfit to the data. That is, the differential spacing between the outcomes in Figure 4-18 can be explained by chance.

An alternative approach to estimating the Y-scores that lets the user predetermine which of the two outcomes will have the lower (higher) score is to assign fixed scores to two outcome categories and the missing score code '.' to the others. For example, assigning  $y_1 = 0.54$  to *marked* and '0' to *stationary* results in the Y-scores in Figure 4-18 being multiplied by -1 and Beta changes from -2.46 to 2.46. (Only those Y-scores assigned the score code "." are estimated.)

### **Fitting a New Model Using other scores (for example, mid-rank**

**scores)**

GOLDMineR will also estimate models for other fixed Y-scores. Next, we will estimate the model that assumes a spacing of the outcomes dictated by Y-scores derived from midrank theory (see DeJonge, 1983). The Y-scores are 3, 2.29, 1.17, -0.17 and -1 listed in order from *marked* to *worse*.

To change the category scores for OUTCOME to be based on user defined scores choose:

Model

Define

to bring up the Define dialog box. Highlight OUTCOME by clicking on it and then click on Fixed .

The Fixed dialog box appears with the default scores for OUTCOME listed in category order from *marked* to *worse*.

Figure 4-19

Fixed Score dialog box



To input the midrank scores

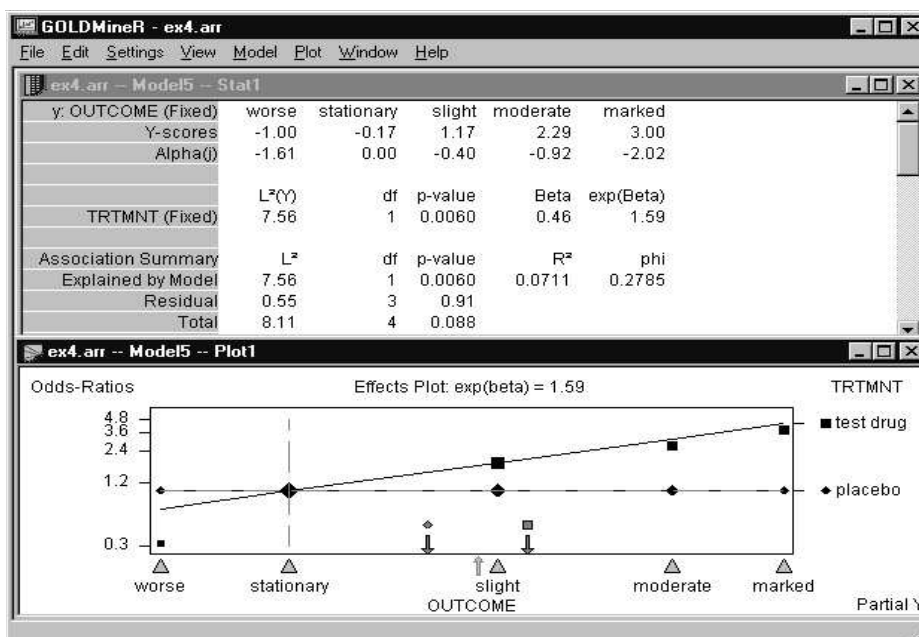
- ▶ double click on a previous score, type in the midrank score and click on Replace. Follow this step until all midrank scores are input.
- ▶ Click on OK.

- ▶ Click on Estimate to reestimate the model.

The default partial regression plot appears. To produce the graph in Figure 4-20:

- ▶ Make the plot window the active window.
- ▶ Select the 'Y' speedbutton on the toolbar or from the Plot menu select Joint Y.

Figure 4-20  
Statistics and Plot Window for Model 3



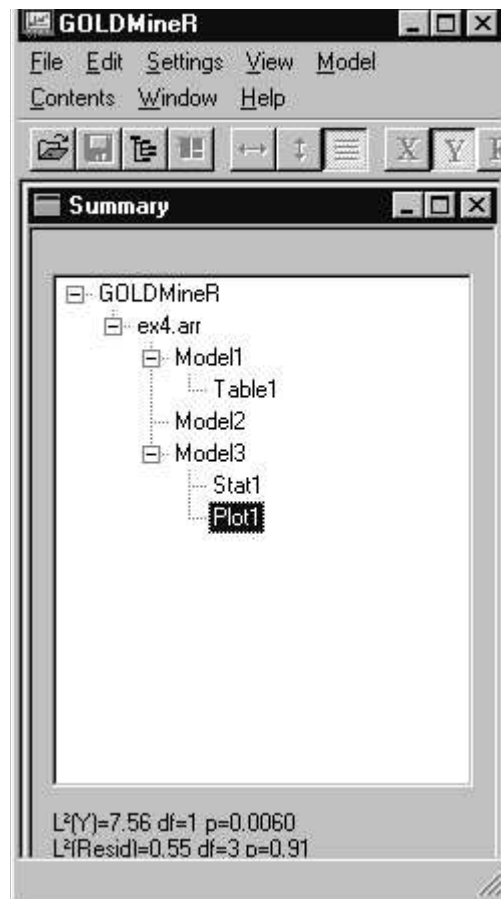
The Statistics window tells you that the current model fit (Residual  $L^2 = 0.55$ ) is slightly worse than that obtained from the first model which assumed equidistant spacing between the outcome categories. The estimated effect is now slightly less significant than before.

## Model Summary Window

To view a summary listing of the models estimated thus far, click on the Model Summary speed button or Ctrl-D or select Summary from the Windows menu to bring

up the Model Summary Window which shows that we have estimated 3 models on datafile ex4.arr and currently have 3 windows open.

Figure 4-21  
Model Summary Window



The window labeled Plot1 is highlighted indicating that it has been most recently active.

## **Tutorial #2**

This tutorial introduces two special features included in GOLDMineR – the model search feature and the quantile chart reporting feature. The model search feature is especially useful in exploratory situations when faced with many potential predictor variables. It can save you time by identifying important predictors to include in the model so that you do not have to build a model containing all the potential predictors when many are extraneous - - often a time consuming task.

The quantile chart provides simple tabular summaries that are especially useful in the case of many different X-profiles such as when a continuous predictor or many categorical predictors are included in the model . This tutorial uses sample dataset #1, the rheumatoid arthritis data (ex1.txt).

If you have just completed Tutorial #1, select File,Close to close that example and start with a clean slate.

### **The Search Procedure**

The Search Procedure in GOLDMineR is primarily useful for variable selection in the case of many predictors. For example, for a dataset that contains 100 candidate predictor variables, the search feature could be used to automatically identify 10 of the most important predictors to include in the model. For more detailed information regarding this procedure, see“SEARCH MENU” on page 151.

- ▶ To load the saved file choose:

File

Open

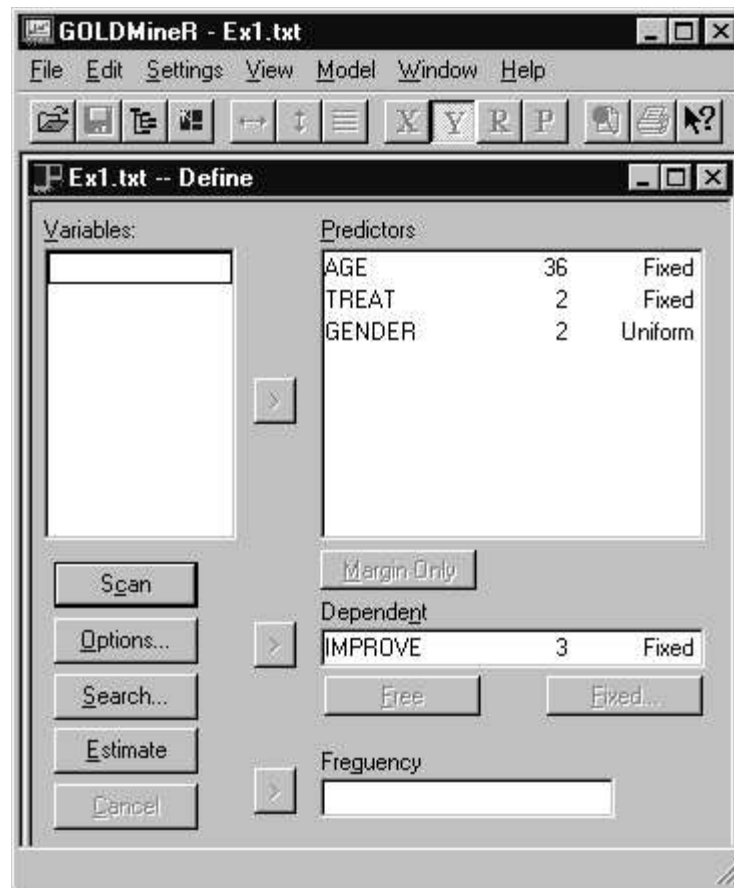
The File Open dialog box will appear. Click on the arrow next to Files of Type and select Data files (.dat or .txt) if this is not already the default listing. A list of all files with those extensions in the current directory will appear. If you copied the tutorial files into a directory other than the default directory, switch to that directory prior to retrieving the file. Highlight the filename “ex1.txt” and click on Open. The Define Model Dialog Box will appear.

Next, select the dependent variable and the set of predictor variables to be considered for inclusion in the model. To select the dependent variable, double click on IMPROVE and it will be moved into the dependent variable box. To select the

predictor variables, highlight all the predictor names (AGE, TREAT, GENDER) and click on the predictor arrow to move them to the predictor box.

Because this is an ASCII data file, click on Scan to scan the data. The Model Define Window should look like Figure 4-22 (we will use the default scaling for each variable which corresponds to Model C used in Chapter 7). (If the Search button is selected prior to scanning the file a scan will be automatically performed prior to bringing up the Search dialog box.)

Figure 4-22  
Model Define dialog box for Search example



- ▶ Click on Search to bring up the Search dialog box (depending on the size of your file, this may take a few moments).

The Search dialog box contains the set of predictors in the upper box.

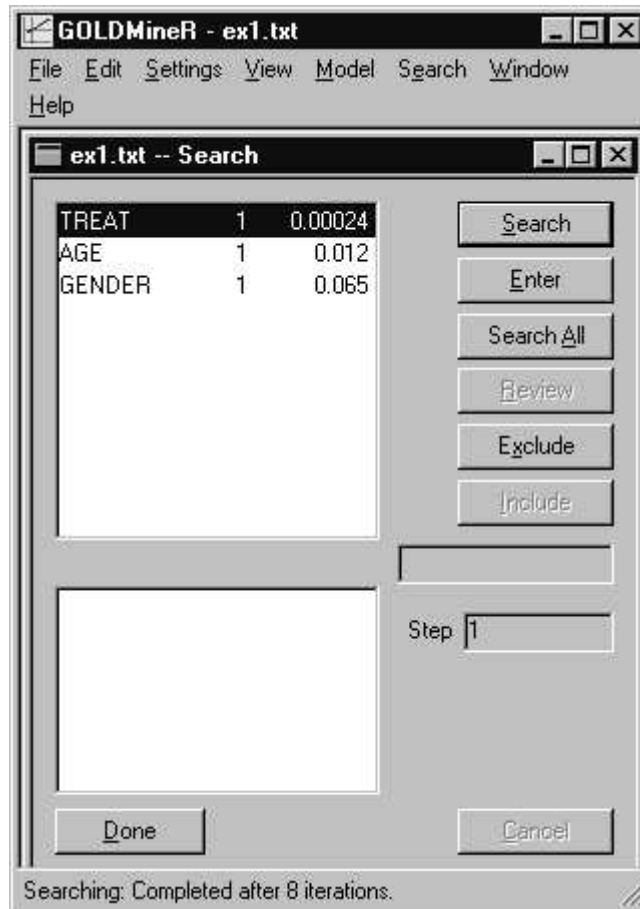
Figure 4-23  
The Search Dialog Box



- ▶ Click on the Search to manually search through all the predictors and rank them from most to least significant along with their associated p-values. The most significant predictor will be ranked first and highlighted (in our example TREAT is highlighted). The number “1” in the Step box (lower right corner) in Figure 4-24 indicates that we are searching for the 1<sup>st</sup> predictor to enter into the model. The number to the right of

each predictor represents the number of degrees of freedom associated with the p-value for that predictor.

Figure 4-24  
After Selecting Search



- Click on Enter to enter the current highlighted predictor into the model (you could alternatively highlight one of the other predictors and enter it). The selected predictor will now be moved into the lower box. If a candidate predictor variable is entered into the model in error, while still in the search dialog box, highlight the desired predictor for inclusion in the model and click on Enter. The desired predictor will be entered in place of the one entered in error.

Figure 4-25  
After Selecting TREAT to Enter into the Model



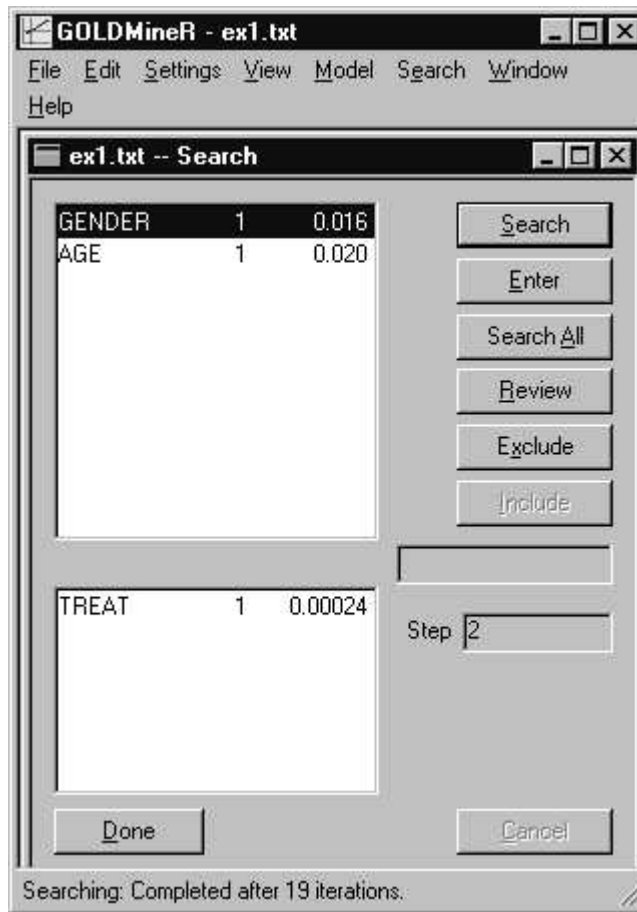
At this time you may examine the output from estimating the current model containing 1 predictor by clicking on Review. The Statistics window (absent the  $L^2(Y)$  statistic) and Plot window will appear on your screen.

- Open the Summary window (or the Window menu) and select Search to return to the Search dialog box. (If many candidate predictors were ranked in the upper window, you might want to highlight the ones that are not significant (have high p-values) and click on Exclude to omit them and save time during the search for the next best predictor.)

- ▶ Click on Search again to search for the 2<sup>nd</sup> predictor to enter into the model. Now the number “2” appears in the Step box to indicate that Search is now evaluating models with two predictors which contain TREAT as one of the two predictors. The significance level indicates the p-value that would be attained for each predictor if that predictor were included in the 2 predictor model.

Figure 4-26

Step 2



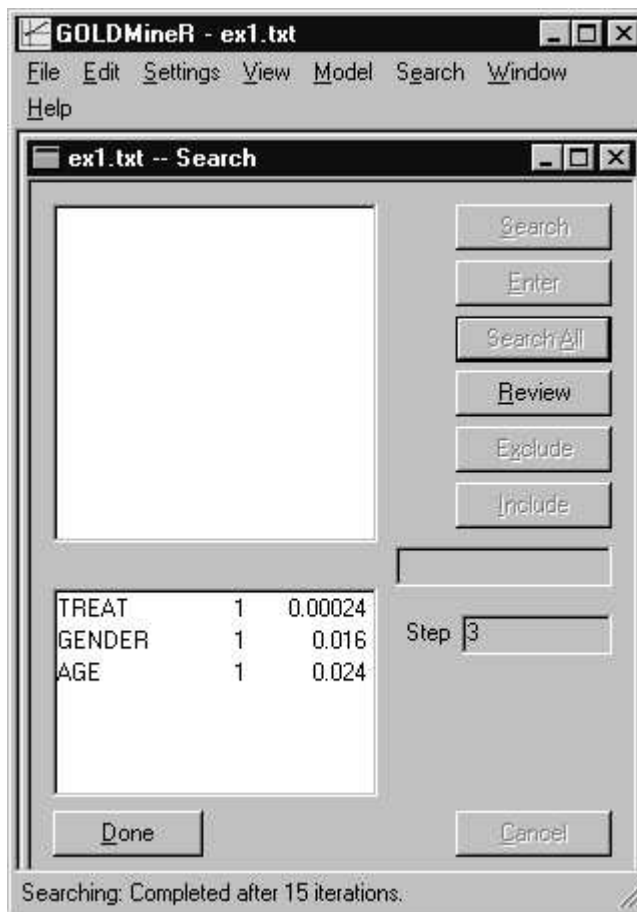
Note that even though GENDER would not be significant as the only predictor (see Figure 4-24), it would be significant ( $p=0.016$ ) in the two predictor model which

includes TREAT. Thus, it was wise not to have excluded GENDER during step 1 of the search procedure.

### ***Automatic Search***

To perform the remaining part of the search process automatically, click on the Search All (this option may be used at any time during the search process). The most significant predictor is automatically entered at each step and the process continues until no significant predictor remains in the upper box (i.e., all predictors have been entered into the model or the p-value associated with each predictor in the upper box exceeds .05). In this example, all three predictors are entered into the model. Depending upon your resources and the number of predictor variables being considered for inclusion, the automatic mode may take some time to complete. One strategy to expedite the search of many predictors is to exclude predictors that are not significant after the initial step (or after the second step) of the search.

Figure 4-27  
Results of Automatic Search



- ▶ When you have completed your Search, click on Done to examine the output windows for the model. You may use the WALD statistics to examine the significance level of each predictor in the current model. Alternatively, you may wish to return to the Model Define dialog box and reestimate the entire model in order to obtain the likelihood ratio difference statistic,  $L^2(Y)$ , to provide a better assessment of the contribution of each predictor.

## Generating Charts (Quantile and Profile)

### *Summary Tables for X-profiles*

The models estimated in “Tutorial #1” on p. 33 were based on a single dichotomous predictor, TRTMNT. Hence, the regression models provided only two outcome predictions, one for each X-profile – patients receiving the *placebo*, and those receiving the *test drug*. For our current model, although there are fewer cases in the sample, there are many more X-profiles (63). When the number of X-profiles is fairly large, the usual kinds of Table summaries may become unmanageable. In such cases specialized charts may be quite useful to summarize the results.

In this tutorial we illustrate various tabular summaries for our current application which contains 63 X-profiles.

To view the X-profiles in our current application, open the Window menu and select New Table. The Estimated Models dialog box appears, listing each model we have estimated. Select the last model estimated by highlighting it and clicking on Ok.

To obtain a joint table which lists all 63 X-profiles as rows of the table select Joint Y from the Table Menu.

By default, the X-profiles are ordered from high to low based on the predicted Y-values. Click on the reverse vertical speedbutton [ ] to display them from low to high or choose Table, Ordered to turn off score based ordering to return to the original ordering in the data file.

Figure 4-28  
Partial listing of Joint Y Table

(Joint Y)			IMPROVE				
	average		0	1	2		
X-profile	score	weights	1.00	0.00	0.00	Y-ref	(cell)
1,F,70	1.00	0.00	0	1	0	.	f
1,F,69	0.50	0.00	1	1	0	1.00	f
1,F,68	1.50	0.00	0	1	1	.	f
1,F,67	2.00	0.00	0	0	1	.	f
1,F,66	2.00	0.00	0	0	1	.	f
1,F,62	1.50	0.00	0	1	1	.	f
1,F,61	2.00	0.00	0	0	1	.	f
1,F,60	2.00	0.00	0	0	1	.	f
1,F,59	2.00	0.00	0	0	2	.	f

Various subsets of the 63 rows can be examined by constructing different Partial Tables, or the 63 rows can be reduced to a smaller number of X-profile groups using Quantile charts. Next, we illustrate the use of the Quantile chart to create a smaller, more manageable summary table.

### ***Generating a Quantile Chart***

A quantile chart contains percentile rankings of the X-profiles grouped by the predicted Y-score. The quantile chart contains an Id that specifies the quantile rank, plus individual and cumulative information containing the size of each X-profile group, percentage of total sample, average expected score and average observed score.

To create and display a quantile chart the Table window must be the active window. Open the Table Menu and select **Specialized Charts**. Your current table will be replaced by a quantile chart containing 10 groups (a decile chart).

Figure 4-29  
Quantile Chart

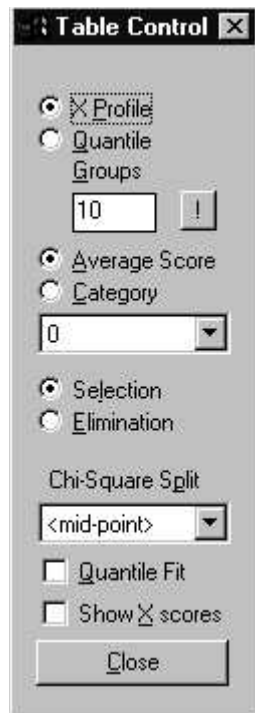
Selection							
	Predicted		Observed	Cumulative ...		Predicted	
Id	n	Score	Score	n	% of n	Score	Score
1	8.40	1.58	1.32	8.40	10.00	1.58	1.32
2	8.40	1.46	1.73	16.80	20.00	1.52	1.52
3	8.40	1.27	1.48	25.20	30.00	1.44	1.51
4	8.40	1.01	0.55	33.60	40.00	1.33	1.27
5	8.40	0.83	0.88	42.00	50.00	1.23	1.19
6	8.40	0.69	0.92	50.40	60.00	1.14	1.14
7	8.40	0.57	0.51	58.80	70.00	1.06	1.05
8	8.40	0.42	0.40	67.20	80.00	0.98	0.97
9	8.40	0.30	0.55	75.60	90.00	0.90	0.93
10	8.40	0.21	0.00	84.00	100.00	0.83	0.83

The decile chart in Figure 4-29 is created by rank ordering the 63 X-profiles from high to low on the predicted Y-value and then grouping adjacent X-profiles into deciles, each containing exactly 10% of the 84 total observations. To obtain exactly 8.4 observations in each decile of the quantile chart, a given X-profile may be proportionally allocated (as described next) between two (or more) decile groups.

### Generating a Profile Chart

The Profile chart lists each X-profile ranked in order of its predicted Y-value. To replace the quantile chart with a profile chart open the Table Menu and select Control or right click the mouse to bring up the Table control dialog box (which now contains chart options). Click on X Profile.

Figure 4-30  
Table Control dialog box



The profile chart in Figure 4-31 now replaces the quantile chart.

Figure 4-31  
Partial listing of the X-profile Chart

Id	n	Predicted Score	Observed Score	Cumulative n	% of n	Predicted Score	Observed Score
1,F,70	1	1.61	1.00	1	1.19	1.61	1.00
1,F,69	2	1.60	0.50	3	3.57	1.61	0.67
1,F,68	2	1.59	1.50	5	5.95	1.60	1.00
1,F,67	1	1.58	2.00	6	7.14	1.60	1.17
1,F,66	1	1.56	2.00	7	8.33	1.59	1.29
1,F,62	2	1.51	1.50	9	10.71	1.57	1.33
1,F,61	1	1.49	2.00	10	11.90	1.57	1.40
1,F,60	1	1.48	2.00	11	13.10	1.56	1.45
1,F,59	2	1.46	2.00	13	15.48	1.54	1.54
1,F,58	1	1.45	0.00	14	16.67	1.54	1.43

In Figure 4-31, the highlighted row, ('1, F, 62'), corresponds to the X-profile of 62 year old *females* ('F') who received the *test drug* ('1'). This X-profile consists of the 8<sup>th</sup> and 9<sup>th</sup> highest scoring observations, which are proportionally allocated between deciles 1 and 2 in Figure 4-29. Specifically, the two observations having this X-profile are counted as 1.4 observations in decile 1, and 0.6 observations in decile 2.

