GOLD*MineR*™

User's Guide

Jay Magidson, Ph.D.
Statistical Innovations Inc.

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

GOLDMineR is a trademark of SI Software Limited Partnership and is based on U.S. patent #5,452,410.

Copyright © 1998 by Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478. All rights reserved.

Windows is a registered trademark of Microsoft Corporation.

GOLDMineR™ 2.0 User's Guide
Copyright © 1998 by Statistical Innovations Inc.
375 Concord Avenue
Belmont, MA 02478

All rights reserved.

Printed in the United States of America

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

To my Teachers

Leo A. Goodman

and the late Donald T. Campbell

Preface

The choice of which regression model to use to predict some dependent variable is made today based upon the metric of the dependent variable. For continuous dependent variables the linear regression model is generally employed. For dichotomous dependent variables (e.g., response vs. nonresponse) the logit model has become the accepted standard (logistic regression). Unfortunately, in practice, many dependent variables are neither continuous nor dichotomous.

Ordinal dependent variables lie between continuous and dichotomous variables. Such variables contain J>2 ordered outcomes. Common examples include a 5-point rating scale (e.g., very favorable, somewhat favorable, neither favorable or unfavorable, somewhat unfavorable, very unfavorable), outcomes from clinical trials (e.g., improvement = none, some, moderate, marked), and usage (none, light, heavy). Up to now, the rather limited availability of mainstream software for analyzing ordinal variables has encouraged researchers to either 1) dichotomize the ordinal variable and use logistic regression or 2) assign quantitative scores (e.g., a Likert scale assigns the integers 1,2,..,J) and use the linear regression model. However, these alternatives involve untested assumptions and may provide misleading results.

In recent years, major methodological advances (see e.g. Goodman, 1979, 1981, 1983, 1985, 1991, 1996) have occurred which suggest the existence of a *single* regression model (i.e., a supermodel) that can accommodate dependent variables of any metric -- whether dichotomous, ordinal or continuous! Even under the bivariate normal distributional assumption that yields a linear regression, the supermodel holds true (see Appendix A), but unlike the linear model it also holds true under many other important exponential family bivariate distributions as well. The supermodel subsumes both the linear and logistic regression models as special cases.

Conceptually, if a positive correlation exists between X and Y, the supermodel provides a generalization from a linear function of X to an always increasing (monotonic) function of X. GOLDMineR (Graphical Ordinal Logit Displays based on Monotonic Regression) is intended for persons who desire to go beyond the limitations of the linear regression model. It is the first mainstream program that is designed to analyze dichotomous, and ordinal, as well as grouped continuous dependent variables.

In addition to traditional regression output, GOLDMineR includes several new interactive graphs (see Magidson 1995, 1996a, 1996b) which ease the process of fitting models and interpreting results. Since the graphs present the essence of the supermodel in a visual fashion, they often reveal new insights into the data which can not generally be ascertained from a mere scanning of traditional tabular statistical output.

GOLDMineR can also generate code for scoring external files and includes a fast Search algorithm that selects important predictors from a large number of potential predictors in a fashion similar to my earlier software product, SPSS CHAID (now called ANSWER TREE).

Acknowledgments

From the time the development began with SBIR grant # 1R43 CA61507-01 from the National Institutes of Health in September 1993 up until now, I have benefited from conversations with many individuals. In particular, I wish to thank Alan Agresti, Tony Babinec, Graham Barlow, Larry Bernstein, Hans Cheng, Tom Chiu, Leo Goodman, Shelby Haberman, Michael Hack, Gary Hoffmann, Paul Holland, David Hosmer, Gary Koch, Jeffrey Lehman, Betsy Magidson, Jeremy Magland, Donald Olivier, Keith Potts, Giuliana Quiroz, Rajesh Selukar, Jing Shyr, Lina Stephan, Bruce Sunstein and Jeroen Vermunt.

The work itself was largely a three person effort. I am very grateful to J. Alexander Ahlstrom for superb programming and to Frank Derrick for his exceptional efforts in the organization and development of this manual and excellent advice throughout. Finally, I wish to express gratitude to Michael Mayer at Beneficial Management Corp. and Ed Hamburg at SPSS for their continuing support.

Jay Magidson
Belmont, Massachusetts
May 1998

Compatibility

GOLDMineR is designed to operate on computer systems running under Windows 95, 98 or NT 4.0.

1

Overview

GOLDMineR (Graphical Ordinal Logit Displays based on Monotonic Regression) is a generalized regression program for predicting a dichotomous, ordinal or grouped continuous dependent variable Y containing J≥ 2 outcome categories, as a function of M quantitative and/or qualitative predictor variables $X = (X_1,X_2,...,X_M)$. When the dependent variable contains more than 2 outcome categories, some or all of the Y-scores that scale the outcomes may be treated as unknown parameters and estimated simultaneously with the regression coefficients.

Certain aspects of monotonic regression modeling, as implemented in GOLDMineR, are new (e.g., the monotonic regression form of the model which differs from the generalized logit form of the model, the way contrast coding for qualitative predictors is specified). In addition, important concepts from log-linear and logit modeling may be unfamiliar to persons experienced with traditional regression approaches (e.g., interpretation of effects in terms of odds ratios, partitioning chi-square, assessing model fit in a manner distinctly different from the use of R^2). Therefore, in addition to the usual kind of "how to use the program" information contained in *all* computer manuals, Chapters 7, 8, 9, 10, 11 and the Appendix contain material of an explanatory nature which address "how to interpret the output" kinds of questions.

GOLDMineR can be used to develop the following kinds of monotonic regression models:

- **Type 1**: logit/logistic regression models (analysis of a binary dependent variable)
- **Type 2**: ordinal logit models with quantitative scores assigned to each category of the dependent variable
- **Type 3**: ordinal logit models with unknown scores for one or more categories of the dependent variable
- **Type 4**: models for grouped continuous dependent variables

Since most of us were taught initially to think in terms of linear relationships, at first glance, the XY scatterplot in Figure 1-1 may give the appearance of linearity. However, the monotonic curve estimated by GOLDMineR provides a better fit to the data. In

practice, the restrictive assumption of linearity may often be replaced by the less restrictive assumption of monotonicity to provide improved prediction without having to choose from among different transformations.

Figure 1-1
Graph of Montonic Regression Curve and Linear Regression Line

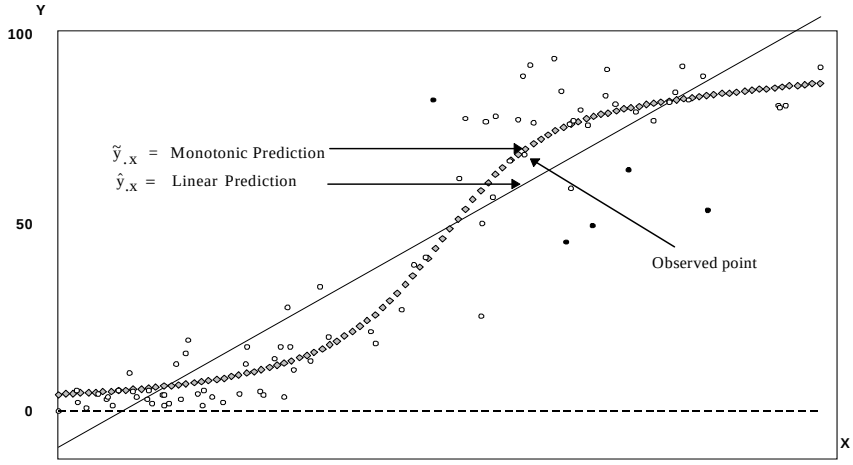


Figure 1-1 shows improved prediction under the monotonic regression model estimated by GOLDMineR (R^2=.90) over the traditional linear regression model (R^2=.85). N = 90 observations.

Sample Datasets

Four sample datasets are used throughout the manual to illustrate various features of the program, and each is analyzed in detail in separate chapters (Chapters 7, 8, 9 and 10). Copies of each dataset are distributed with the program in all three GOLDMineR input formats - - array (.arr), text (.dat or .txt) and SPSS saved files (.sav).

- #1: Rheumatoid Arthritis Data (ex1.*); Source: Koch, G. and Edwards, S. (1988)
- #2: Hypothetical Mail Survey Data (ex2.*)

- #3: Mail Survey Data (ex3.*); Source:Magidson, J. (1994b)
- #4: Clinical Trial Data (ex4.*); Source:DeJonge, H. (1983)

About This Manual

Chapter 2 introduces the Summary and Model History windows which help you to keep track of your data sets, models estimated and interactive output windows.

Chapter 3 provides a detailed description of the three kinds of interactive output windows and describes the various model parameters and statistics obtainable through model estimation.

Chapter 4 provides two hands-on tutorials to help you get the program up and running quickly.

Chapter 5 describes the various input data options that are supported by GOLDMineR.

Chapter 6 details all GOLDMineR menus and commands.

Chapters 7- 10 present analyses for each sample data set which serve to introduce the different model types and illustrate major features of the program. Chapter 7 compares the use of model types 1, 2 and 3 to assess the effect of a drug treatment for rheumatoid arthritis (sample dataset #1). In a marketing research application (sample dataset #2), Chapter 8 illustrates how the use of seemingly reasonable scores for a predictor variable may yield a statistically significant *negative* effect estimate despite the true effect being *positive*. This chapter also illustrates how the model fit statistic can signal a scaling problem (or other kind of model misspecification) and how the model is modified to overcome the problem. Chapter 9 uses sample dataset #3 to illustrate the use of various contrast coding options for estimating the effects of a qualitative predictor. Chapter 10 uses sample dataset #4 to illustrate the relationship between odds, odds ratios and the model parameter estimates, and provides guidance in the interpretation and use of the Association Summary Section of the Statistics window.

Chapter 11 illustrates the use of model Type 3 in a classical multinomial logit/ multiple group discriminant analysis/cluster analysis type of application.

The technical Appendix documents the models and statistics which form the basis of GOLDMineR. Beyond the Appendix and the articles listed as references, for further information and related material users are referred to books by Agresti (1990), Ishii-Kuntz, (1994), and Clogg and Shihadeh (1994).

2

Tracking Datasets, Estimated Models and Interactive Output Windows

In any session of GOLDMineR, you may estimate many models on each of several different data files, and for each model, you may open one or more Statistics, Plot or Table windows. There are several ways to help you keep track of the various windows that have been opened:

- The **Summary window** provides a hierarchical listing of all windows that have been opened and can be used to move between windows.
- The **Model History** dialog box lists all models estimated.
- The name of the data file, a model number and window designation appear at the top of each window.
- The name of the data file and model number appear at the bottom of each page of printed output.

Summary Window

In a given GOLDMineR session, users may attend to different datasets, models and the associated interactive output windows by using the Summary window. The Summary window provides

- a hierarchical listing of each open data file.
- each model estimated for that data file.
- the associated Statistics, Plot and Table windows opened for each model. If you are running consecutive models on the same data file, GOLDMineR will replace the default Plot and Statistics windows with each new model. To view the windows for an earlier model, highlight the model number, right click to bring up the Contents Menu and select the type of window (Statistics, Plot, Table).

E To access the Summary window, from the menus choose:

Window

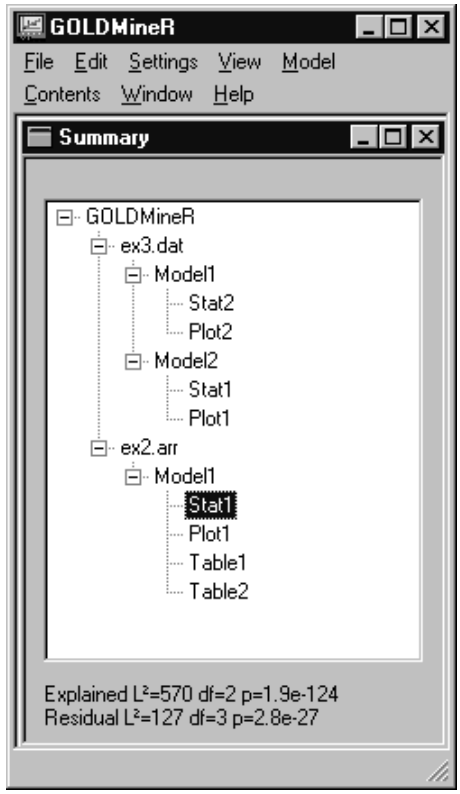
Summary Window

E or type <ctrl-D>.

If the Summary window has been opened, it will be listed at the bottom of the Window menu with all other currently open windows. To display more or less information in the Summary window, use the mouse to select the +- keys. When the Summary window is the active window, the Contents Menu will appear on the menu bar. Special features include:

- Right clicking on any item in the Summary window will open the Contents Menu (depending on where you right click, only certain menu items may be available). See “CONTENTS MENU” on page 150 for more details.
- Double clicking on ‘GOLDMineR’ opens the Open dialog box.
- After one or more data files have been opened, double clicking on the data file name opens the Define dialog box which contains the settings for the most recently estimated model.
- After one or more models have been estimated, double clicking on the Model number opens the Contents Menu. See “CONTENTS MENU” on page 150 for more details.
- Double clicking on a specific window name (for example, Stat1, Plot2) will make that window the active window.

Figure 2-1
Summary Window



For example, Figure 2-1 tells you that:

- For the data set named ex3.dat, there are two models that have been estimated. Each model has one Statistics and one Plot window open.
- For the dataset named ex2.arr, one model has been estimated and one Statistics, one Plot and two Table windows are open.

The summary statistics listed at the bottom of the window relate to whichever model is highlighted in the window (the statistics listed above relate to the ex2.arr data set, Model 1).

Model History

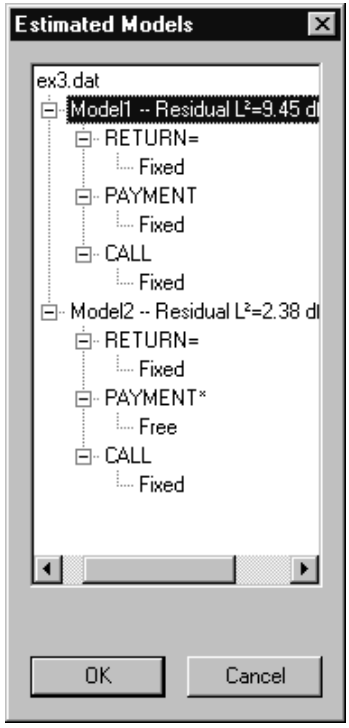
The Model History dialog box provides a more detailed listing than the Summary window for each model estimated (including Residual L^2, df, p-value, variable names and scaling) for the *current* dataset only. Double clicking any item within a model listing will change the current window to that model.

E To open the Model History dialog box from the menus choose:

Model

History

Figure 2-2
Model History dialog box



For example, Figure 2-2 tells you that for the dataset named ex3.dat, two models have been estimated for the dependent variable named RETURN. One difference between

the models is that Model 1 utilized **Fixed** scaling for the predictor variable named PAYMENT while Model 2 utilized **Free** scaling.

3

Interactive Output Windows

Three interactive output windows are available to examine the results of model estimation from different perspectives:

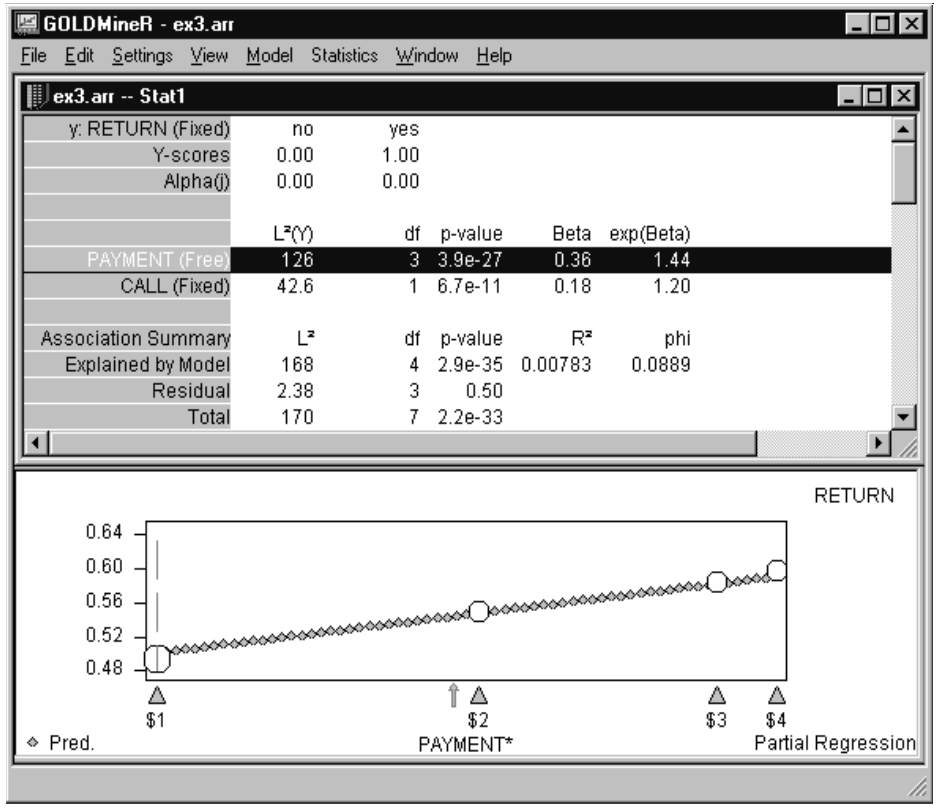
- The **Statistics Window** contains traditional statistical output (parameter estimates, standard errors, significance tests and assessment of model fit).
- The **Plot Window** contains graphical displays (regression plots and effects plots).
- The **Table Window** displays various predictor by dependent variable relationships.

To customize views for each output window, from within the corresponding window, right click to bring up the control dialog box. In addition, users may interact with the output windows by changing the contrast coding for a qualitative predictor in a given model which creates instantaneous changes in the relevant sections of each of these windows.

Upon completion of the model estimation procedure GOLDMineR automatically:

- Opens a Statistics Window in the upper part of your screen highlighting the contribution of the first predictor in the model.
- Opens a Plot Window in the lower part of the screen visually *displaying* the contribution of that predictor in the model (the Partial Regression plot), as in Figure 3-1. The plot for the first predictor may be replaced by a plot for a different predictor by simply clicking on a different predictor variable name in the Statistics Window.

Figure 3-1
Initial GOLDMineR display



Alternatively, the Partial Regression plot can be changed to a joint Regression plot which shows the contribution of all predictors simultaneously or to a corresponding *Effects* plot which displays the effect of the predictor in the generalized logit form of the model (see"Plot Types" on page 133 for further information).

A Table window may also be opened to obtain precise numerical values that correspond to points on the plot as well as other information on the predictor(s) by dependent variable relationships. The Table window (as well as the Plot window) can display information associated with a selected predictor ("partial" displays) or information based on all predictors simultaneously ("joint" displays).

The three different kinds of output windows provide alternative, but complementary ways to view results of model estimation. Specifically, from different perspectives, each provides information about

- the magnitude of the effect of any selected predictor
- the predictive power of the model as a whole
- how well the model fits the observed data.

In this chapter we describe the components of each window and illustrate how various statistics, plots and tables can be effectively used together to help interpret the results of your analysis.

All output in this section is based on an analysis of sample data set #3. For further interpretation and discussion of model fitting strategies for these data see Chapter 9.

The Statistics Window

Following is an annotation of the first three sections of the Statistics window (Model Summary, Category-Specific Parameter Estimates, and Scores) for those components displayed by default. Optional information may also be displayed in these sections such as correlations or covariances for the alpha and beta parameter estimates and the design matrix. In addition, the name of the data file and the model number appears at the bottom of each printed page of output.

The Statistics window features quick transfer between the Model Summary and the Category-Specific Parameter Estimates section by double clicking on a variable name in either section. Upon double clicking, immediate transfer is made to the other section where information on the selected variable will be highlighted. The mouse pointer changes from the default logo (an arrow) to a pointing hand when the mouse pointer is moved to variable names and other transfer points in the statistics output. In addition, if the most recently active Plot window contains either a partial regression or partial effects plot for some predictor, highlighting any predictor variable name in the Statistics window will automatically change the partial plot to one associated with the highlighted predictor.

Model Summary Statistics

The first section of the Statistics window contains Model Summary Statistics.

Figure 3-2
Example of Model Summary Statistics

y: RETURN (Fixed)	no	yes				
Y-scores	0.00	1.00				
Alpha	0.00	0.00				
		$L^2(Y)$	df	p-value	Beta	exp(Beta)
PAYMENT (Free)		126	3	3.9e-27	0.36	1.44
CALL (Fixed)		42.6	1	6.7e-11	0.18	1.20
Association Summary		L^2	df	p-value	R^2	phi
Explained by Model		168	4	2.9e-35	0.00783	0.0889
Residual		2.38	3	0.50		
Total		170	7	2.2e-33		

The Model Summary Statistics section contains:

A. For each dependent variable outcome category, the **Y-scores** used in the current model and the estimated **Alpha** parameter for that category. The Y-score may be either standardized (a V-score) or unstandardized. While the unstandardized Y-scores are not affected by changes to the E-weights, standardized Y-scores (V-scores) are updated immediately whenever the E-weights are changed. The alpha parameters, representing baseline logits, are updated immediately when either the E-contrast weights or the C-weights (for any predictor) are changed. For J > 8, only the first 8 alpha parameters are displayed. For further details, including definitions of the E and C-weights, see "A.3 Estimating X-Scores for Qualitative Predictor Variables" on page 214.

B. For each predictor,

- **Beta.** The effect estimate for that predictor in the additive form of the model. Beta is interpretable as a log-odds ratio.
- **exp(Beta).** The effect estimate for that predictor in multiplicative form of the model. exp(beta) is interpretable as an odds ratio.
- **L^2(Y)**. The likelihood ratio chi-squared difference statistic used to assess the statistical significance of Beta (and exp(Beta)).
- **p-value**. The level of statistical significance associated with the chi-squared statistic L^2 or $L^2(Y)$ with **df** degrees of freedom

C. The **Association Summary** contains the analysis of association which partitions the total association between the predictors and the dependent variable, **Total L^2**, into the **Explained by Model L^2,** the amount of association which is explained by the model

plus the residual association, **Residual L^2,** the amount of association that remains unexplained, which tests the validity of the model assumptions.

- The **p-value** corresponding to the Explained L^2 is used to test H_0:$R^2 = 0$ (and H_0': phi = 0).
- $0 \leq \mathbf{R}^2 \leq 1$, the monotonic R^2, measures the percentage of variance of Y that is explained by the monotonic regression model, computed as the ratio of the variance of the predicted Y to the variance of the actual Y (see "A.14 Analysis of a Continuous Dependent Variable" on p. 235).

  It is important to note that R^2 (as well as the usual linear R^2, displayable as an option) may be quite low when the dependent variable is not continuous and/or highly skewed (R^2 = .007 in Figure 3-2) even with a model that contains highly significant predictors like the one summarized in Figure 3-2. Hence, R^2 is not as useful as it is when the conventional linear regression assumptions are met.
- **Phi** ≥ 0 measures the degree of association between Y and the predicted Y (phi is the absolute value of the association parameter in Goodman's weighted RC model – see "A.5 Model Identification Issues" on p. 218.)

The following additional information may be obtained using the Statistics Menu or Statistics Control dialog box (accessed while right clicking within a Statistics window).

D. The Detail option provides the **Pearson Total X^2** and **Residual X^2**. Alternative measures of model fit that are asymptotically equivalent to Total L^2 and Residual L^2 when the model is valid and the data is nonsparse (See "A.8 Partitioning Chi-Square" on p. 226).

E. Design Matrix. The Design Matrix option provides the design matrix used to compute the parameter estimates (see "A.11 Design Matrix" on p. 231).

F. Correlation/Covariances of Parameter Estimates. The Correlations option provides the correlations between the beta estimates, between the alpha estimates or between alpha and beta estimates. Optionally, covariances may be listed in the output instead of correlations**.**

G. Convergence Information. Information as to whether the estimation algorithm converges, and the **number of iterations** is reported. Optionally, users can obtain further information by selecting Detail which includes 1) the number of digits to which the estimated frequency counts reproduce the observed **linear R^2** (a condition of convergence – see "A.10 Estimation Algorithm" on p. 229), 2) a comparison of the

linear R^2 based on the observed ("**Obs**") and the estimated expected ("**Exp**") frequencies, and 3) the value used for the technical parameter **Delta**.

(Note: The number of iterations reported following estimation may differ depending upon whether the LR Contribution option is designated because the estimation algorithm utilizes different start values under the LR Contribution setting.)

H. Decile Fit. When requested in the Model Options dialog box before a model has been estimated, the following new statistic is suggested as an alternative to Residual L^2 and Residual X^2 when the data is sparse (See "A.15: Assessing Model Fit when Analyzing Sparse Data" on p. 244).

I. Delta. When requested in the Model Options dialog box before a model has been estimated, the user-specified positive quantity is distributed among the observed cell counts prior to estimation to resolve identification problems caused by zero cells. By default, Delta = 0

Category-Specific Parameter Estimates

The second section of the Statistics window contains the parameter estimates Alpha(j) for each category j = 1, 2, ..., J, of the dependent variable, Beta(k), for each category k = 1,2,...,K_m of the mth predictor treated as Free and the corresponding standard errors. Each of these statistics is updated interactively as the associated category weights (E-weights for the dependent variable and C-weights for each predictor) are changed. (see "Interpreting the Effect Estimates" on page 195 and sections A.1-A.3 of the Appendix.)

Figure 3-3
Example of Category-Specific Parameter Estimates

Category-Specific Parameter Estimates

RETURN (Fixed)	Y-score	E-weights	Alpha(j)	Std. Err.	Wald	p-value	exp(Alpha)		
yes	1.00	0.00	0.00	0.03	0.02	0.88	1.00		
no	0.00	1.00	0.00	0.00	.	.	1.00		
PAYMENT (Free)	X-score	C-weights	Beta(k)	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
$1	0.00	1.00	0.00	0.00	.	.	1.00	.	.
$2	0.52	0.00	0.19	0.04	25.02	5.7e-7	1.21	1.12	1.30
$3	0.91	0.00	0.33	0.04	75.34	4.0e-18	1.39	1.29	1.50
$4	1.00	0.00	0.36	0.04	89.40	3.2e-21	1.44	1.33	1.55
CALL (Fixed)	x0	C-weights	Beta	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
	0.00	(2)	0.18	0.03	42.56	6.9e-11	1.20	1.13	1.26

For the *dependent* variable, the Statistics window output lists:

- **Y-score.** Category scores for the dependent variable are reported in unstandardized (Figure 3-3) or standardized units depending on the current setting of the View, Unstandardized scores option. For an interpretation of the standardized Y-scores (called V-scores), see "A.5 Model Identification Issues" on p. 218. For a Free scale dependent variable, the estimated Y^*-scores are shown.
- **E-weights**. Dependent variable category weights used to define the baseline logits, Alpha(j), j = 1, 2, ..., J.
- Point estimates for the baseline logit parameters, **Alpha(j),** and the corresponding baseline odds, **exp(Alpha)**. (For J > 8, only the first 8 parameter estimates are displayed unless the option to display covariances of the Alpha parameters is requested.)
- **Std. Err.** The estimated standard errors for each Alpha(j).
- **Wald**. The Wald statistic used to test H_0:Alpha(j) = 0.
- **p-value**. The p-value assessing the significance of the Wald statistic.

For each *predictor* treated as *qualitative* (**Free** scale), such as PAYMENT in Figure 3-3, the Statistics window output lists:

- **X-score.** Category scores for a Free scale predictor variable are reported in unstandardized (see Figure 3-3) or standardized units depending on the setting of the View, Unstandardized score option. For the interpretation of the scores, see section A.3 of the Appendix for the unstandardized X-scores and section A.5 of the Appendix for the standardized X-scores (called U-scores).
- **C-weights**. Predictor variable category weights. For a Free scale predictor such as PAYMENT in Figure 3-3 the C-weights specify the contrast used to define the associated Beta(k) parameters. For example, the C-weights for PAYMENT specify dummy coding with category 1 (PAYMENT = $1) serving as the reference (see A.3).
- Point estimates for the category-specific parameters **Beta(k)** and **exp[Beta(k)]**.
- **Std. Err**. The standard error for each Beta(k).
- **Wald**. The Wald statistic used to test H_0:Beta(k) = 0.
- **p-value**. The p-value assessing the significance of the Wald statistic.
- **Lower/Upper**. The lower and upper 95% confidence limits for each exp(β(k)).

For each *predictor* variable treated as *quantitative* (**Fixed** scale), such as CALL in Figure 3-3, the Statistics window output lists:

- **x_0**. The X-reference score. (In the partial plots and partial tables, predictors other than the selected predictor are held constant at the X-reference score (see sections A.1, A.3, and A.12 of the Appendix).
- **C-weights**. A symbolic representation for the category weights used to define the X-reference score. C-weights define the reference score for both Fixed and Free predictors. For Free predictors, C-weights also are used to specify the contrast coding. The table below shows the relationship between the symbolic representations (used for Fixed predictors) and the contrast coding used for Free predictors:

	Contrast Type
(k)	dummy coding with category k as the reference category
UnWtd	Unweighted Avg. (effect coding)
Wtd	Weighted Avg. (weights used are marginal proportions)
(C)	Custom (contrast weights input by user)

- **Beta**. Same as Beta in "Model Summary Statistics" on page 13.
- **Std. Err.** The estimated standard error for Beta.
- **Wald**. The Wald statistic used to test H_0:Beta = 0 (H_0:exp(Beta) = 1).
- **p-value**. The p-value assessing the significance of the Wald statistic.
- **exp(Beta)**. Same as exp(Beta) in "Model Summary Statistics" on page 13.
- **Lower/Upper**. The lower and upper 95% confidence limits for exp(beta).

Scores

The third section of the Statistics Window contains information on the scores used for the variables.

Figure 3-4
Example of Scores Section

Scores	Stnd	UnStnd
RETURN (Fixed)		
yes	2.02	1.00
no	0.00	0.00
y0	0.00	0.00
mn	1.14	0.57
sd	1.00	0.50
PAYMENT (Free)		
$1	0.00	0.00
$2	1.05	0.52
$3	1.83	0.91
$4	2.02	1.00
mn	0.97	0.48
sd	0.87	0.49
CALL (Fixed)		
yes	1.01	1.00
no	0.00	0.00
x0	0.00	0.00
mn	0.51	0.50
sd	0.50	0.50

The Scores section contains:

- **Category scores** (both standardized and unstandardized) for each variable that has fewer than 8 categories.
- For the dependent variable and predictors treated as quantitative (Fixed), the reference score $\mathbf{y_0}$ and $\mathbf{x_0}$.
- **mn**. The mean score.
- **sd**. The standard deviation.

For variables designated as Free scale, the reference score is always zero and is not displayed.

The Plot Window

The plot window can display either:

- Regression Plots or
- Effects Plots

which can be specified as either:

- partial or
- joint

Effects plots can also be viewed in either:

- X-view or
- Y-view.

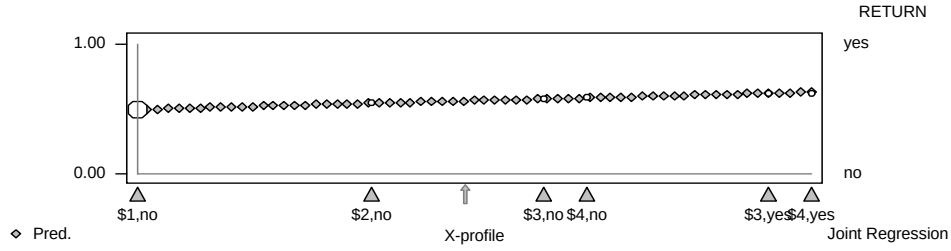
See “Plot Types” on page 133. for specific details on each type of plot.

The default range used for the vertical axis of the joint regression plots differs from that used for partial plots.

Joint Regression Plots

Figure 3-5 is an example of a joint regression plot.

Figure 3-5
Joint Regression Plot with selected X-profiles



The **joint** regression plot contains:

A. The **monotonic regression curve** (represented by diamond shaped symbols) plots the predicted y-value as a function of the joint X-variable which attaches the score $X_i = \Sigma \hat{\beta}_m x_{mi}$ to the ith X-profile (X_{1i}, X_{2i}, …, X_{Mi}). In Figure 3-5, selected X-profiles are plotted on the horizontal axis (using triangular markers) according to their score on the joint X-variable. Options:

- You can choose the X-profiles to be displayed in the graph - - the two extreme X-profiles are always displayed by default. See “Label Display” on p. 131.
- You can control the density of the monotonic regression curve by specifying the number of diamond shape symbols to be plotted.
- The horizontal axis can be labeled using equidistant quantitative values of the joint X-variable rather than each joint category/value, by selecting the continuous score display option.

B. The **observed average y-score** (represented as a circle) is plotted for each selected X-profile. The diameter of the circle is proportional to the number of observations having the given X-profile. Clicking on a circle causes the following message (specific to the circle clicked) to be displayed on the status bar: Observed Average: X-Profile [yes, $1] (obs y avg.=.552, res=1.45). **Res** denotes the adjusted residuals which assess the difference between the predicted Y and the average Y observed in the sample for that X-profile. Circles corresponding to profiles where the predicted Y is significantly different from the observed average (at the .05 level), as identified by |res| > 1.96, are differentiated from other circles by the shading. See Screen “Colors” on page 110 for exact colors of residuals. (For the partial Regression plot associated with the m^{th} predictor, observed average Y-score circles only appear when category weights for all other predictors are consistent with dummy coding.)

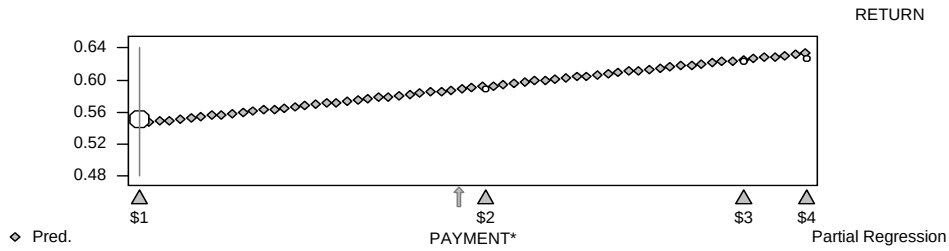
By default, the vertical scale ranges from the minimum to the maximum Y-score observed in the sample, but any alternative range may be specified by the user.

Partial Regression Plots

Figure 3-6 and Figure 3-7 illustrate two partial regression plots for PAYMENT-- 1) given CALL = 'No' and 2) given CALL = 'Yes'. These figures are a subset of Figure 3-5.

Figure 3-6
Partial Regression Plot for PAYMENT given CALL = no

Figure 3-7
Partial Regression Plot for PAYMENT given CALL = yes



The **partial** regression plot associated with predictor $X_{m'}$ contains:

A. The **partial regression curve**. Identical to the joint plot, but predicted values of Y are plotted as a funtion of $X_{m'}$ for specified X-reference levels X_{m0} of the other predictors, X_m, m=1,2,..,M, $m \neq m'$.

B. Observed Average Y-score. Identical to the joint effects plot, but only values for the selected subset of X-profiles are plotted. That is, for a selected predictor $X_{m'}$ (PAYMENT in Figure 3-6) only X-profiles are plotted that are included in the joint X-variable range X=($X_{m'}$, $X_m = x_{mo}$).

The following may optionally be appended to the plot:

- The linear regression line
- The derivative of the predicted Y value with respect to the selected predictor

By default the range of the vertical axis is determined so that the plot contains the average Y values predicted and observed for all X-profiles represented in the plot.

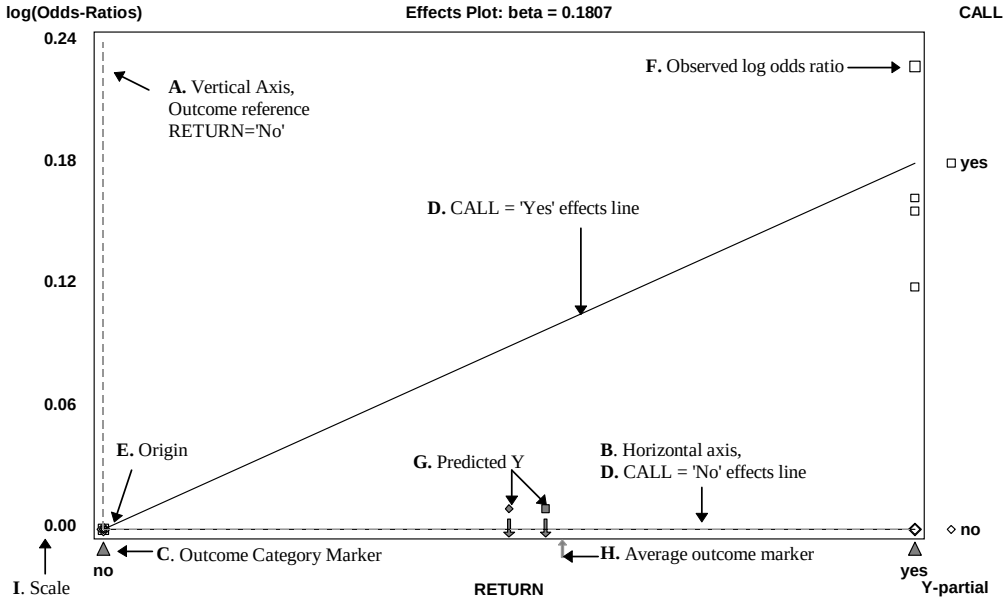
Effects Plots

Partial effects and joint effects estimated by a model are odds ratios. When displayed in logarithmic units, the estimated expected log-odds ratios may be plotted as an exact linear function of X (the X-view) or Y (the Y-view). In X-view, the partial or joint X variable is represented on the horizontal axis while in the Y-view, Y is represented by the horizontal axis. The slope of the Y-view partial effects line associated with category k of a predictor is equal to the category specific parameter estimate for that predictor, Beta(k).

Annotated Partial Y Effects Plot

Following is an annotated partial Y effects plot with details of its components.

Figure 3-8
Partial Y-View Effects Plot for the selected predictor CALL



Each Y-view of the effects plot is a semi-log plot which contains the following components:

A. The **vertical axis** corresponds to the selected Y-reference which serves as the base in defining the generalized odds. In Figure 3-8 the RETURN = *no* category was selected as the Y-reference and coincides with the vertical axis. (The Y-reference is selected by the user to be any category of the dependent variable or an 'average' response category by specifying the appropriate E-contrast weights.)

B. The **horizontal axis** corresponds to a category (or an 'average' category) of the selected predictor which serves as the X-reference or baseline used in defining the odds ratios. In Figure 3-8 we see that the CALL = *no* category was selected as the X-reference.

C. The **outcome category markers**, spaced according to the Y-scores that are assigned to or estimated for the response categories, are signified by triangular markers at the bottom of the plot. For example, Figure 3-8 displays triangular markers associated with the two levels of RETURN, using Y-scores 0 and 1. Double clicking on an outcome marker changes the E-weights so that the Y-reference corresponds to that outcome category.

D. The **effects line** associated with a category of a selected predictor plots expected log-odds ratios associated with that category as a linear function of Y. The slope of the effects line shown equals *Beta*, the effect estimate for that predictor.

E. The **origin**, the point of intersection between the vertical and horizontal axes, represents the reference point (or base) for interpreting the odds ratios. In Figure 3-8 the origin is associated with the (RETURN = *No*, CALL = *No*) cell.

F. The observed log-odds ratio associated with each outcome category is displayed for each selected X-profile observed in the sample. For example, in Figure 3-8 four observed nonzero log-odds ratios associated with RETURN = *Yes* are plotted using square shaped symbols (for the X-profiles (CALL, PAYMENT = (*Yes, $1*), (*Yes, $2*), (*Yes, $3*) and (*Yes, $4*)). Larger symbols reflect cells that are based on larger sample sizes. The lack of fit of the model to the data is ascertained by the Residual L^2 statistic which measures how distant these symbols are from the corresponding effects line.

G. Each **predicted Y,** designated by an arrow pointing downwards onto the horizontal axis, corresponds to the predictions given for each category of the selected predictor with the other predictors held constant at their X-reference levels. In Figure 3-8 the arrows represent predicted return rates associated with CALL = *Yes* and CALL = *No* for a given payment of $1 (PAYMENT = *$1*). (To change the PAYMENT level, use Model, Category Weights.)

H. An **Average Outcome marker** identifies $\bar{y}$, the mean Y-score. Double clicking this marker changes the E-weights so that the Y-reference corresponds to $\bar{y}$ (i.e., $y_0 = \bar{y}$) and the plot is updated immediately to reflect the revised definitions for odds and odds ratios.

I. The scale that appears next to a vertical border at the left or right of the plot represents logarithmic (default) or standard **odds ratio units.** The estimated odds ratio is given above the partial effects plot in default logarithmic units (Beta = .1807) or in standard units (exp(Beta) = 1.198).

The Table Window

GOLDMineR creates various predictor by dependent variable tables that can be displayed in either the X-view or the Y-view. In Y-view the outcome categories comprise the columns of the table and the i^{th} row of the table corresponds to either a particular X-profile formed by observed combinations of the predictor variables (a joint table) or a category of a single predictor variable (partial table). For a partial table, the cell entries in the i^{th} row are associated with the i^{th} category of the selected predictor for given reference levels of the other predictors.

A reversal of a Y-view table created by transposing rows and columns of the table yields the X-view of that table. For simplicity, for the remainder of this section when we speak of a table we will be referring to the Y-view of the table.

Cell (i,j) of a table may contain:

- Observed or expected frequency counts
- Probabilities/proportions, odds or odds ratios calculated from either observed or expected frequency counts.
- Standard or logarithmic units.

Figure 3-9 is the joint table containing observed frequency counts for sample dataset #3 analyzed in Chapter 9.

Figure 3-9
Joint Table with Observed Frequency Counts

(Joint Y)			RETURN			
	average		yes	no		
X-profile	score	weights	0.00	1.00	Y-ref	(cell)
$1,yes	0.55	0.00	2407	1954	1954	f
$1,no	0.50	1.00	2133	2176	2176	f
$2,yes	0.59	0.00	1265	881	881	f
$2,no	0.55	0.00	1156	942	942	f
$3,yes	0.62	0.00	1340	809	809	f
$3,no	0.58	0.00	1262	897	897	f
$4,yes	0.63	0.00	1306	779	779	f
$4,no	0.60	0.00	1248	839	839	f
	X-ref		2133	2176	2176	f
	Y-scores		1.00	0.00	21394	

Figure 3-10 is the partial table (a subset of Figure 3-9) associated with the predictor PAYMENT for the subset of observations in Figure 3-9 for which CALL = 'Yes'.

Figure 3-10
Partial Table for PAYMENT with Observed Frequency Counts Given CALL = 'Yes'

(Partial Y)			RETURN			
	average		yes	no		
PAYMENT	score	weights	0.00	1.00	Y-ref	(cell)
$1	0.55	1.00	2407	1954	1954.00	f
$2	0.59	0.00	1265	881	881.00	f
$3	0.62	0.00	1340	809	809.00	f
$4	0.63	0.00	1306	779	779.00	f
		X-ref	2407.00	1954.00	1954.00	f
		Y-scores	1.00	0.00	10741	

In addition to the particular (i,j) cell entry (frequency count, probability, odds or odds ratio) each GOLDminer table also contains information identifying the row and column references that are used to compute odds ratios, the Y-scores associated with the current model and the average scores for each row of the table. More specifically, each table contains the following components:

- The **cell entry identifier** provides a one letter character that identifies the content of the cell. One or more of the following cell entries can be selected:

f	observed frequencies
F	expected frequencies
p	observed proportions
P	expected probabilities
o	observed odds
O	expected odds
r	observed odds ratios
R	expected odds ratios

The single **cell entry** selected for Figure 3-9 and Figure 3-10 is "f", the observed frequency count.

- Reference scores (**Y-reference** and **X-reference**). As illustrated in Chapter 10, the observed odds of outcome j may be calculated for the i^{th} row of the table as $f_{ij}/f_{i0,}$ where f_{ij} is the observed frequency count for cell (i,j) and f_{i0} is the associated observed Y-reference count for row i used as the base in computing odds. The Y-reference count will equal the observed frequency associated with a particular outcome or an average of the frequency counts in that row depending upon how the category weights E_0 are chosen. In Figure 3-9 and Figure 3-10 the Y-reference count corresponds to the RETURN = *No* outcome.

  The *baseline* odds in favor of outcome j, f_{0j}/f_{00}, is the odds calculated for the designated baseline reference. The baseline reference or X-reference, f_{0j}, may correspond to a particular row, or it may correspond to some "average" row. In Figure 3-9, the X-reference corresponds to row i=2 associated with CALL = *No* and PAYMENT = *$1*. In Figure 3-10, the X-reference corresponds to row i=1, PAYMENT = *$1* for observations where CALL = *Yes*.

- Category **weights** associated with the rows and the columns define the Y-reference and X-reference used in the calculation of odds and odds ratios. The weights consist of a set of numbers between 0 and 1 that sum to 1. For example, in Figure 3-10, the row contrast is C = (1,0,0,0,), which allocates 100% of the weight to the first category of PAYMENT. (For qualitative predictors, the C-weights define the contrasts used for the category-specific beta parameters.) The column weights in Figure 3-10 is E = (0,1), which allocates 100% of the weight to the second category of RETURN.

- **Y-scores** are used by the model to quantify the ordinal outcome categories. The scores displayed in Figure 3-9 and Figure 3-10 are the original unstandardized scores (*Yes* = 1, *No* = 0) used in the analysis conducted in Chapter 9. Unstandardized scores are not effected by changing the E-contrast.

Alternatively, standardized Y-scores ('V-scores - - see "A.5 Model Identification Issues" on p. 218) may be displayed in the table. See "Unstandardized Scores" on page 114 for details on selecting the type of Y-score displayed in the Table.

- The **average score** is computed for each row of the table based on the Y-score (unstandardized or standardized) displayed. In Figure 3-9 and Figure 3-10, since the Y-scores are 1 and 0 and the *observed* frequency counts are selected, the average score represents the observed return rate for each X-profile (observed conditional proportion). If *expected* frequency counts were selected instead of the *observed* for display in the table, the average score displayed would be the *expected* return rate as predicted by the model.
- The cell reference count $\mathbf{f_{00}}$ is located in the table cell where the X-reference column and Y-reference row intersect .
- The total sample size **N** is provided in the last cell in the lower right corner of the table (21,394 in Figure 3-9). Estimated expected frequency counts always sum to the total observed sample size. $N = \sum_i \sum_j f_{ij} = \sum_i \sum_j \hat{F}_{ij}$.
- **Adjusted Residuals** (optional). For each row of the table the adjusted residuals assess model fit by computing the standardized difference between the predicted Y-score and the observed average Y. (An adjusted residual having magnitude of 1.96 or greater indicates that the difference is significant at the .05 level.)

Figure 3-11 is the partial Y-view table containing the estimated expected odds ratios associated with PAYMENT given CALL = *No*. This table corresponds to Figure 3-6 . (The predicted Y values displayed in the table correspond to some of the points plotted in the regression curve.)

Figure 3-11
Partial Table for PAYMENT with Expected Odds Ratios Given CALL = 'No'

(Partial Y)		RETURN				
	average		yes	no		
PAYMENT	score	weights	0.00	1.00	Y-ref	(cell)
$1	0.50	1.00	1.00	1.00	1	R
$2	0.55	0.00	1.21	1.00	1	R
$3	0.58	0.00	1.39	1.00	1	R
$4	0.59	0.00	1.44	1.00	1	R
	X-ref		1	1	1	R
	Y-scores		1.00	0.00	10653	

Figure 3-12 is the partial Y-view table containing expected odds-ratios for CALL given PAYMENT = *$1*. In Figure 3-8, the estimated expected odds ratios are plotted (in logarithmic units) to create the effects lines. Since the average scores in Figure 3-12 are based on estimated expected frequency counts, they correspond to the predicted score arrows displayed in Figure 3-8.

Figure 3-12
Partial Table for CALL with Expected Odds Ratios Given PAYMENT = '$1'

(Partial Y)			RETURN			
	average		yes	no		
CALL	score	weights	0.00	1.00	Y-ref	(cell)
yes	0.55	0.00	1.20	1.00	1	R
no	0.50	1.00	1.00	1.00	1	R
		X-ref	1	1	1	R
		Y-scores	1.00	0.00	8670	

Specialized Charts

The Table Window also displays charts which group X-profiles into quantiles and rank them by the predicted Y-value. Such charts can be created for each X-profile observed in the sample (profile chart) or for summary groups containing exactly the same number of observations in each group (quantile chart).

Profile Chart

The first type of specialized chart is known as the Profile chart, which can be shown in either “average score” or “category” format.

Figure 3-13
Example of a Profile Chart

Selection

Id	n	Predicted Score	Observed Score	Cumulative ... n	% of n	Predicted Score	Observed Score
$4,yes	2085	0.63	0.63	2085	9.75	0.63	0.63
$3,yes	2149	0.63	0.62	4234	19.79	0.63	0.62
$2,yes	2146	0.59	0.59	6380	29.82	0.62	0.61
$4,no	2087	0.59	0.60	8467	39.58	0.61	0.61
$3,no	2159	0.58	0.58	10626	49.67	0.60	0.60
$2,no	2098	0.55	0.55	12724	59.47	0.60	0.60
$1,yes	4361	0.55	0.55	17085	79.86	0.58	0.58
$1,no	4309	0.50	0.50	21394	100.00	0.57	0.57

The "average score" format of the profile chart displayed in Figure 3-13 has the following components:

- Profile **Id** defines the profile using labels assigned to the predictor variables. There will be one row for each combination of the predictor variables.
- **n** is the number of sample observations for the given X-profile.
- **Predicted Score** ($\tilde{y}_{.x}$) is the predicted Y-value for the given X-profile.
- **Observed Score** ($\bar{y}_{.x}$) is the observed mean Y-score for observations having the given X-profile.
- **Cumulative n** is the cumulative number of observations that have a predicted Y-value $\geq \tilde{y}_{.x}$.
- **Cumulative % of n** is the cumulative percentage of observations that have a predicted Y-value $\geq \tilde{y}_{.x}$.
- **Cumulative Predicted Score** is the predicted Y-value for observations that have a predicted Y-value $\geq \tilde{y}_{.x}$.
- **Cumulative Observed Score** is the observed mean value for observations that have a predicted Y-value $\geq \tilde{y}_{.x}$.

Quantile Chart

The second type of specialized chart is known as the Quantile chart.

Figure 3-14
Example of a Quantile Chart

Selection

Id	n	Predicted Score	Observed Score	Cumulative ... n	% of n	Predicted Score	Observed Score
1	2139.40	0.63	0.63	2139.40	10.00	0.63	0.63
2	2139.40	0.62	0.62	4278.80	20.00	0.63	0.62
3	2139.40	0.59	0.59	6418.20	30.00	0.62	0.61
4	2139.40	0.59	0.60	8557.60	40.00	0.61	0.61
5	2139.40	0.58	0.58	10697.00	50.00	0.60	0.60
6	2139.40	0.55	0.55	12836.40	60.00	0.60	0.60
7	2139.40	0.55	0.55	14975.80	70.00	0.59	0.59
8	2139.40	0.55	0.55	17115.20	80.00	0.58	0.58
9	2139.40	0.50	0.50	19254.60	90.00	0.57	0.57
10	2139.40	0.50	0.50	21394.00	100.00	0.57	0.57

The quantile chart with 10 groups (a decile chart) in Figure 3-14 has 21,394/10 = 2,139.40 observations per group. Fractional observations are created by allocating individual observations that have identical X-profiles proportionally into the relevant quantile groups. All other components are the same as in the Profile Chart.

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.

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

Table 4-1
Clinical Trial Data[*]

*(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.

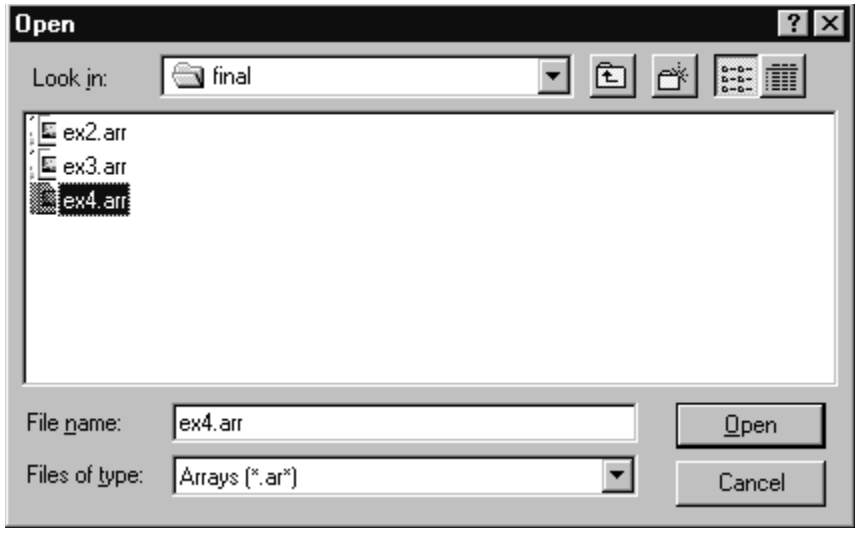
E To open the previously saved array file from the menus choose:

File

Open

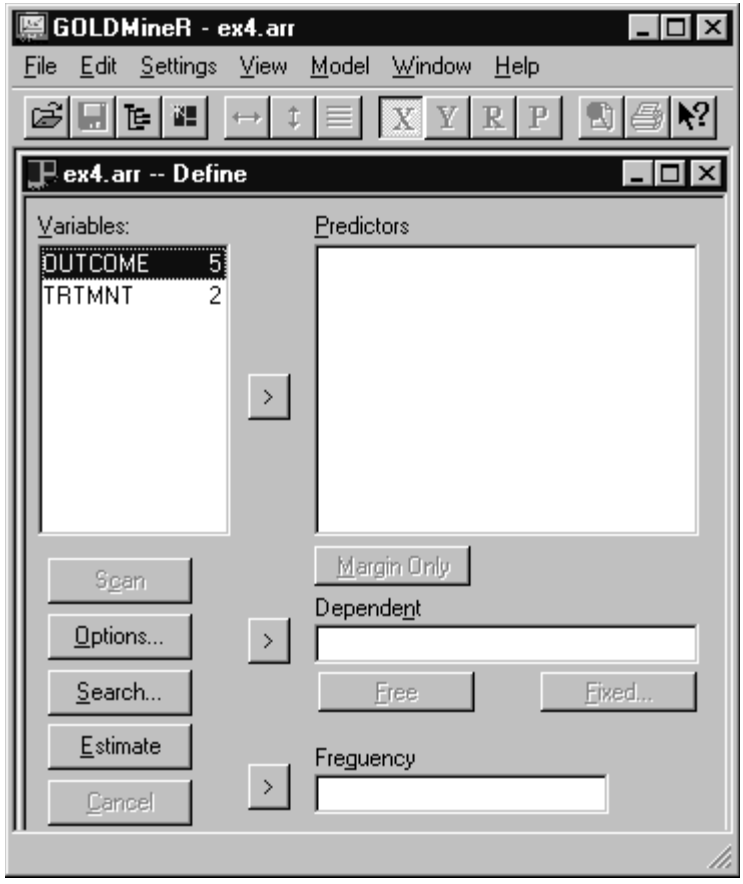
GOLDMineR places you in the Open dialog box. Click on the arrow next to Files of type and select array files (.ar*) if this is not already the default listing. A list of all files with .arr extensions will now appear in the selection box (See Figure 4-1). If you copied the tutorial files into a directory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

E Click on OUTCOME in the Variables box.

E Click on the > key next to the Dependent variable box

E 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:

E click on TRTMNT in the variables box

E Click on the > key to the left of the Predictors box

E or, after selecting the dependent variable, you may double click on TRTMNT and it will be moved to the predictor box.

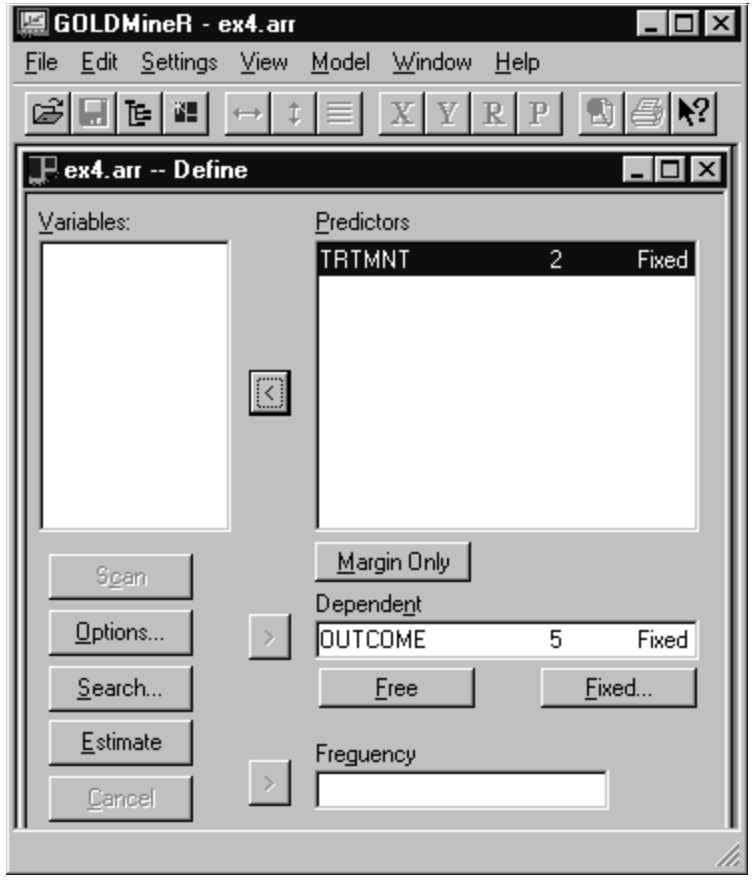
(If you mistakenly select a variable that you do not wish to be a predictor, click on that variable in the Predictor variable box, and then click on the reverse arrow key to return it to the variables box.)

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

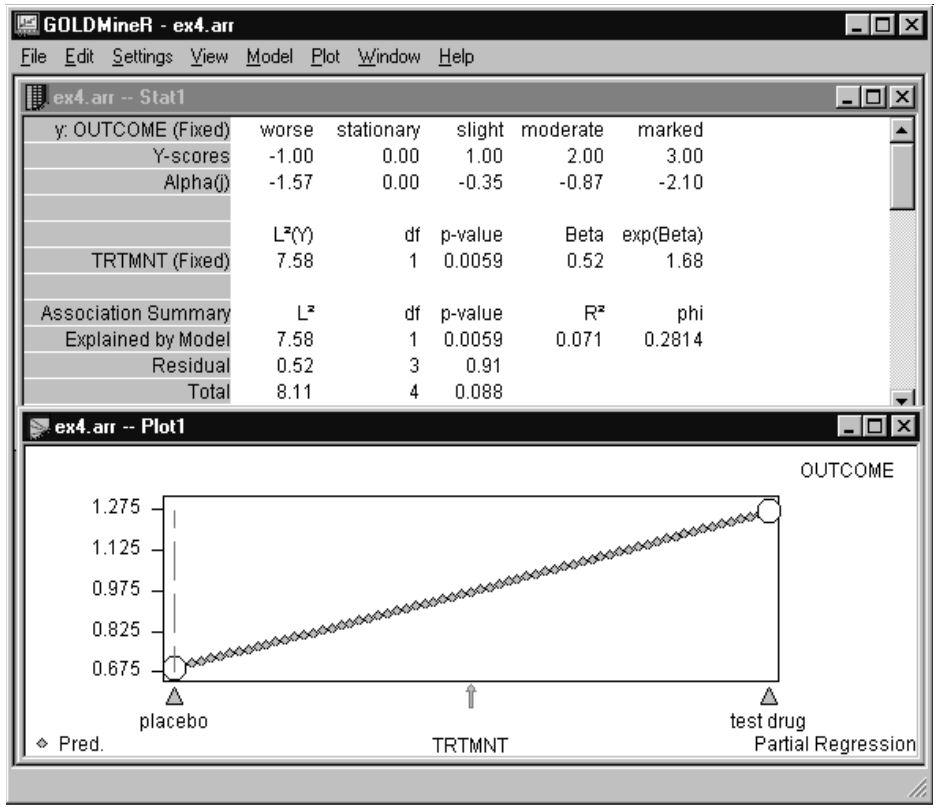
Now that we have defined our variables and have settled on our scaling options, we can estimate a model.

E 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:Beta = 0 (or the equivalent H_0: exp(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

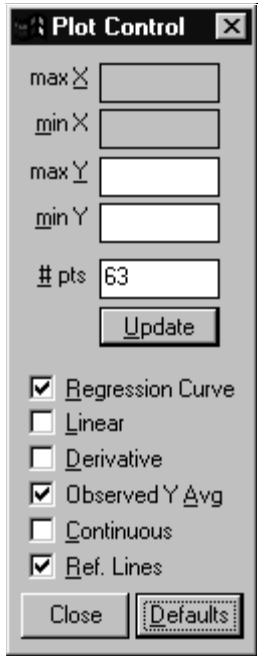
bar changes to: Expected Y[x=0.00]=0.69. Thus, the predicted Y-value is identical to the observed average Y-value. Click on the circle and diamond above the *test drug* marker to also confirm that the predicted Y-value equals the corresponding observed average Y-value for persons receiving the *test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

E 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



E Highlight '63', type in a new number, '3'

E Click on Update.

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

E 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).

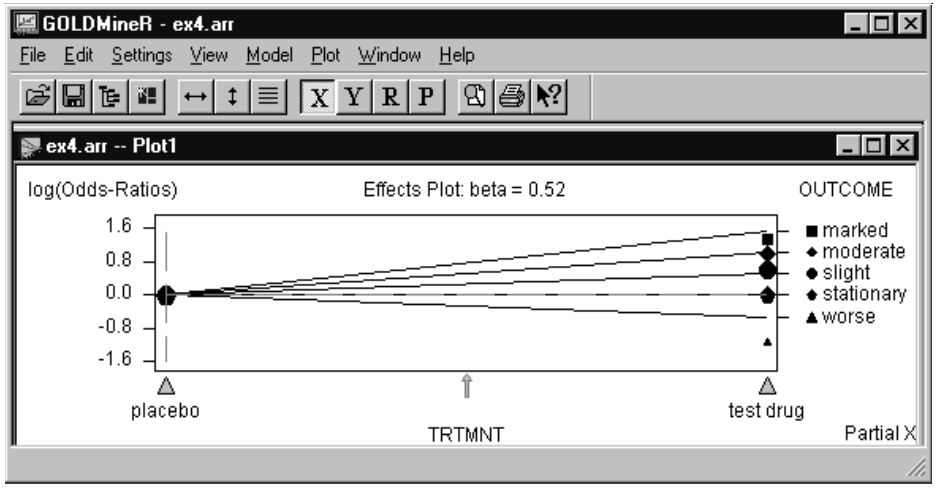
To obtain an Effects plot in the X-view:

E 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).

E 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.

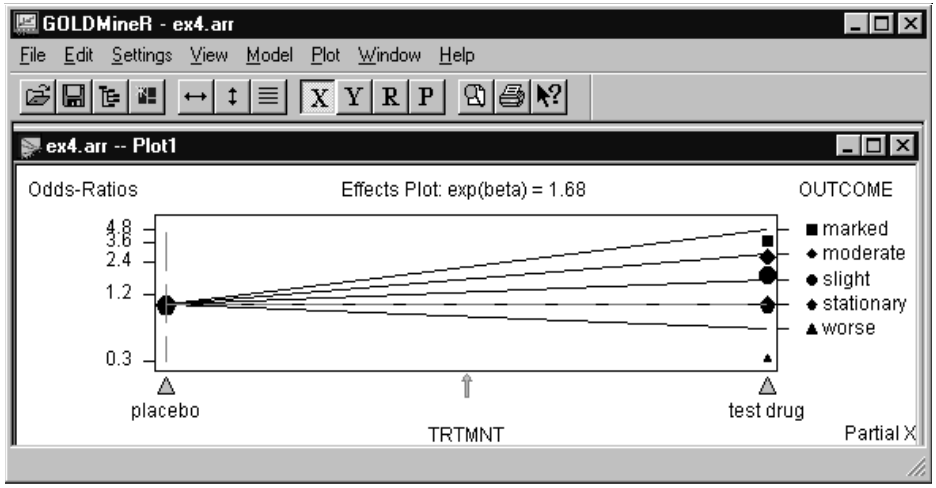
E To translate the log-odds ratios in the plot to the more easily interpretable odds ratios from the menus choose:

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.

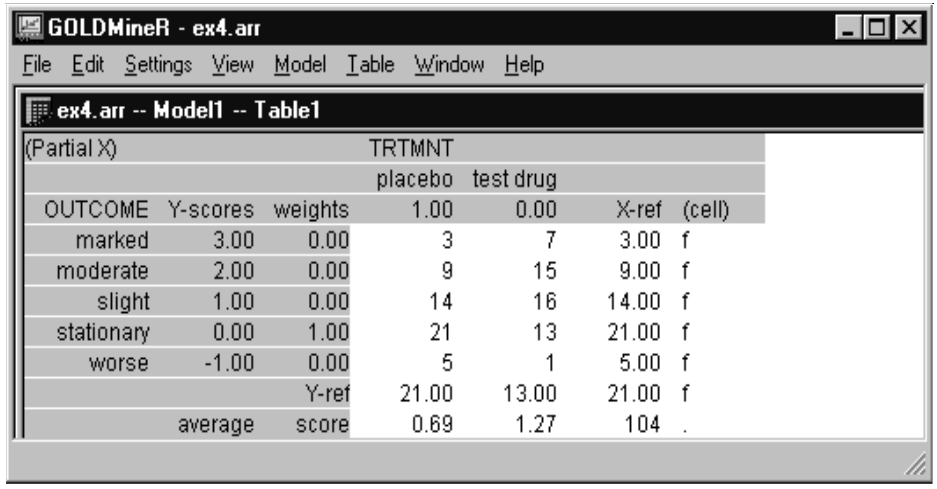
E To open a Table window from the menus choose:

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



(Partial X)			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	.

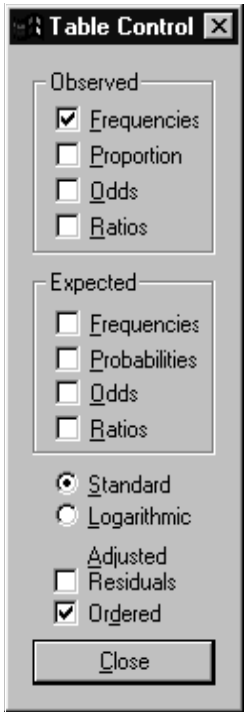
A table can contain either observed or estimated expected frequency counts, probabilities, odds or odds ratios - - in standard or logarithmic units. For this example, we wish to replace the observed frequency counts with estimated expected odds ratios. To change the table so that it contains estimated 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.

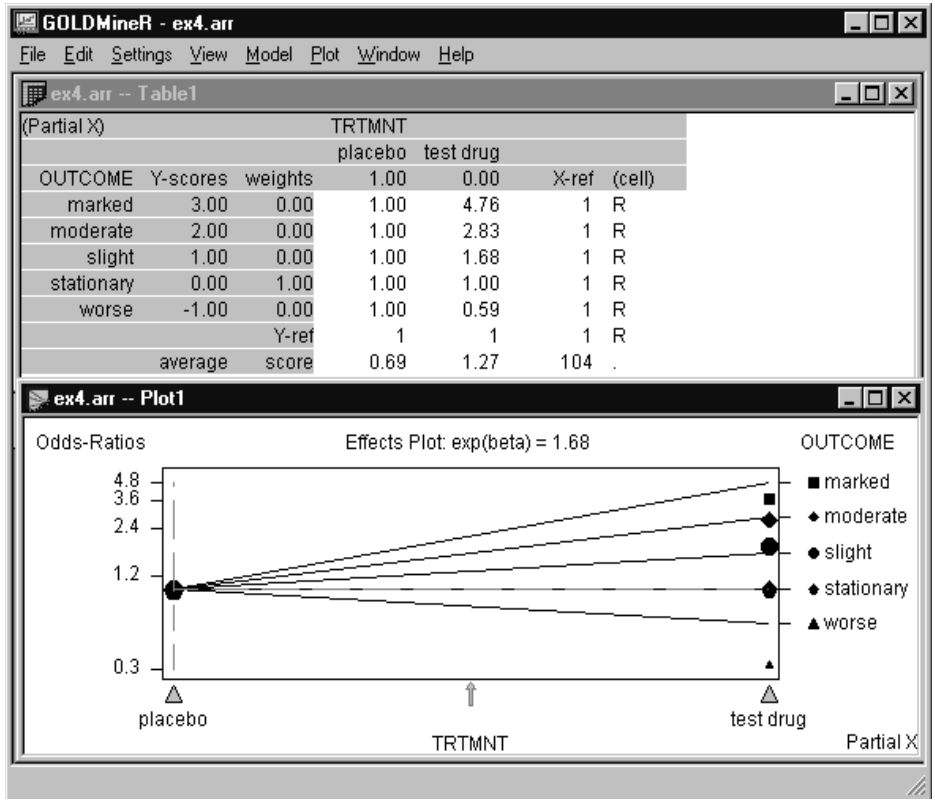
Figure 4-9
Table Control dialog box



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 jth outcome, βy_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.

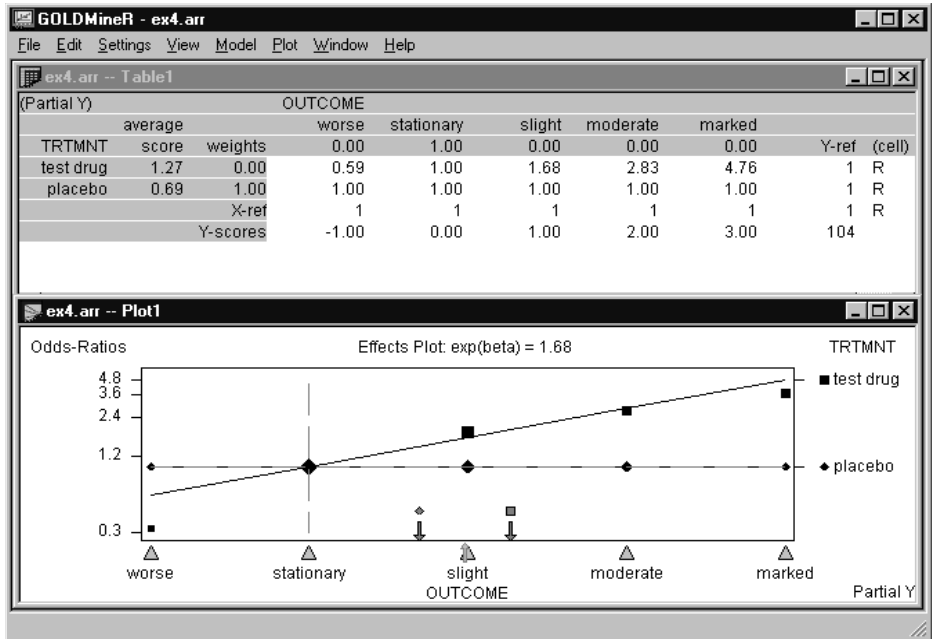
The OUTCOME categories are displayed along the side of the Effects plot with a symbol used to identify the corresponding observed odds ratio. The TRTMNT categories are labeled at the bottom of the effects plot below the associated marker.

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 $\hat{\beta}_{x_i^*}$. 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(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 $\hat{\beta}$) = 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:

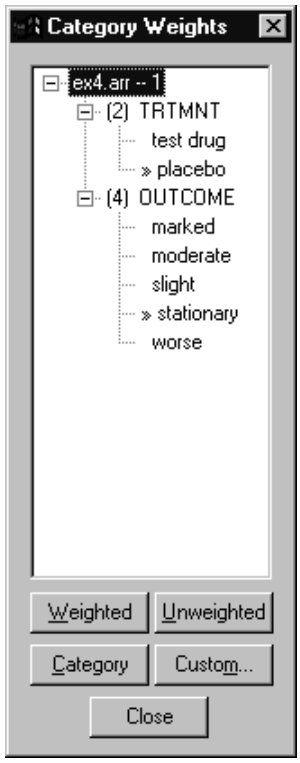
Model

Category Weights

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

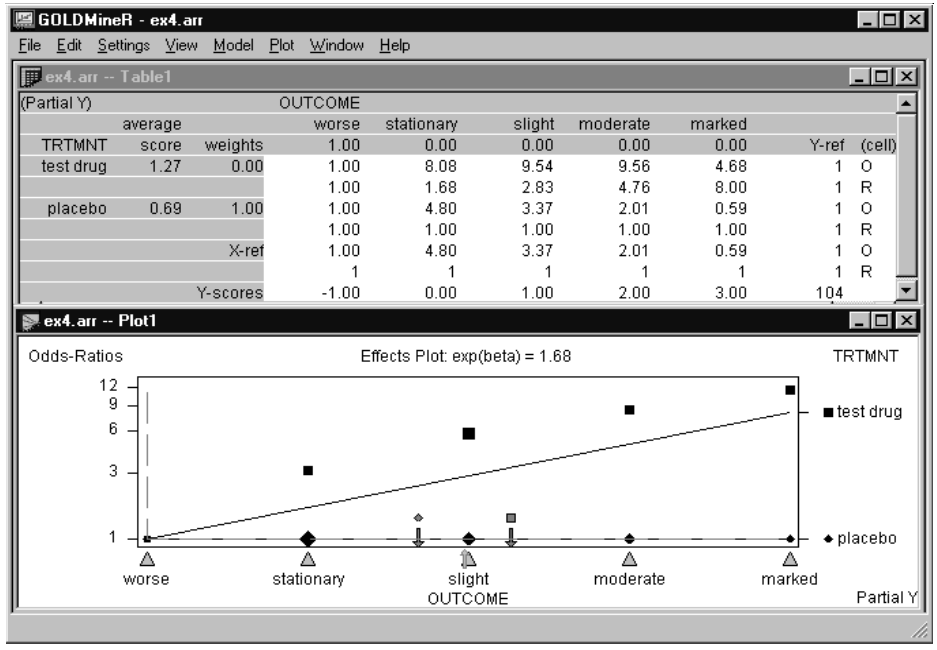
The Category Weights dialog box will appear with a list of each of the variables and its categories. A symbol identifying the current category weighting option is listed to the left of the variable name. The symbol "(k)" idendifies the kth category as the reference. The k^{th} category of the variable 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 [>>] 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.

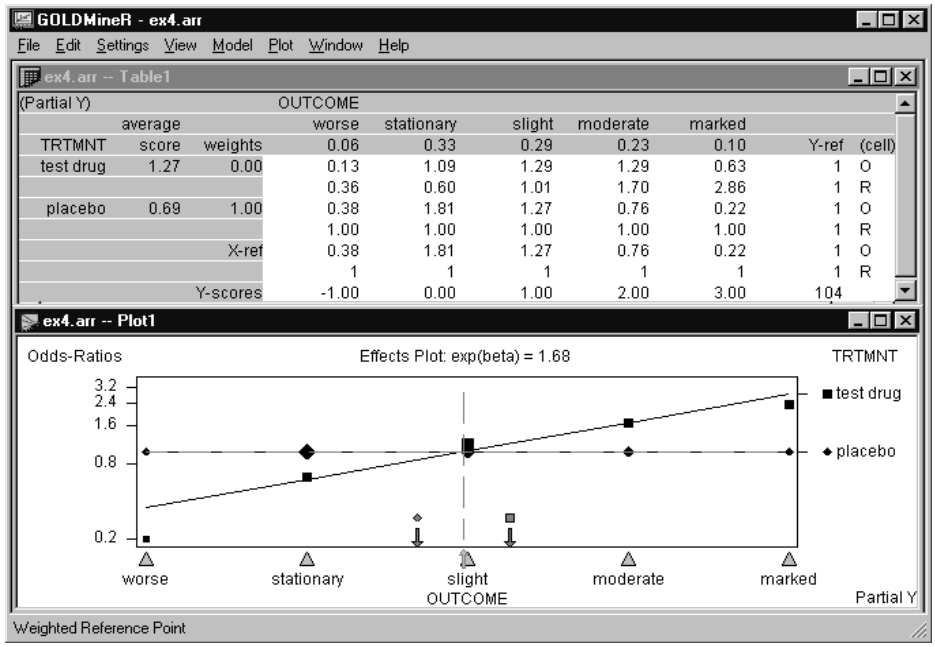
Selecting Weighted Average References

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



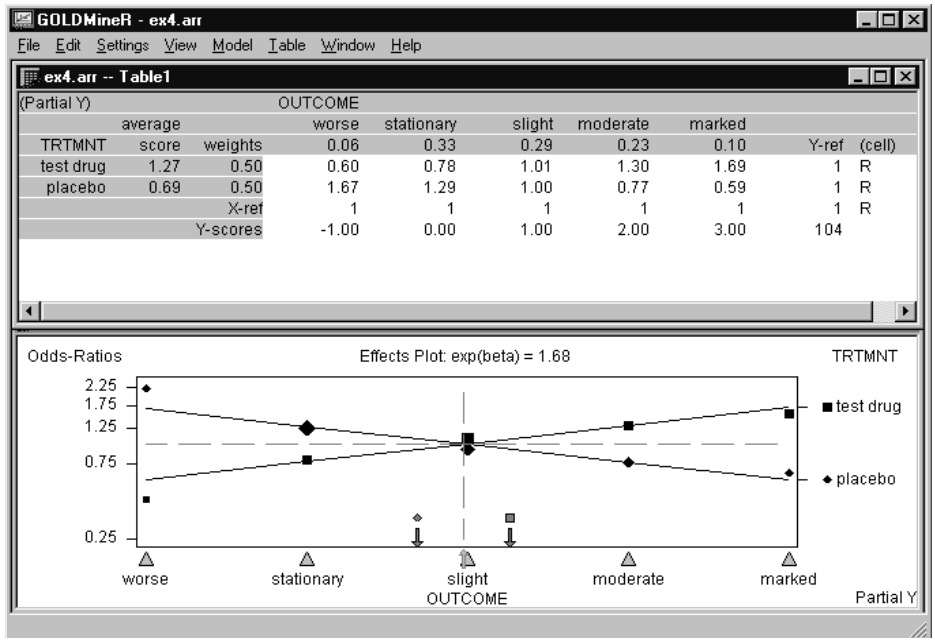
The odds of marked improvement (vs. the "average" outcome) are 2.86 times as high for patients who received the test drug than for those receiving the placebo.

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 wieghts 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

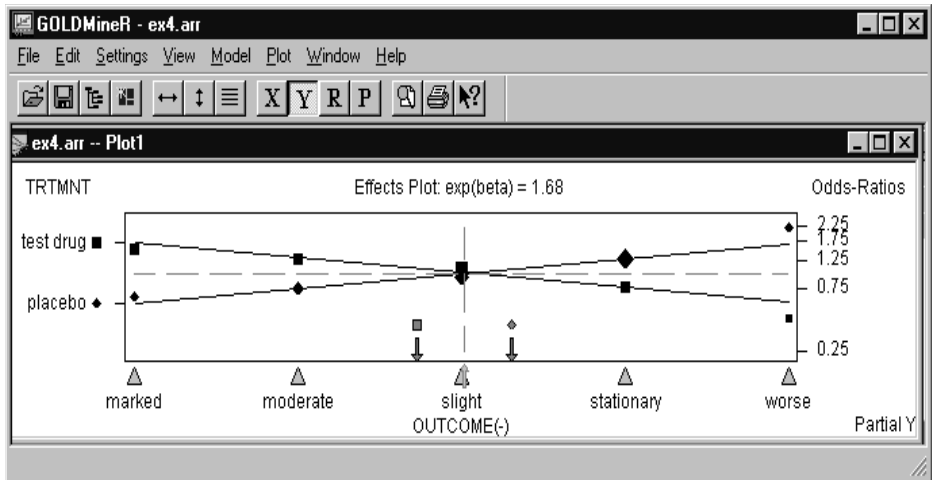
If at any time you wish to change the order of categories in the effects plot or tables (i.e., to make it easier to read), GOLDMineR allows you to reverse the order of categories in the effects plot or tables for either variable without changing the model or statistics. To reverse the order of categories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

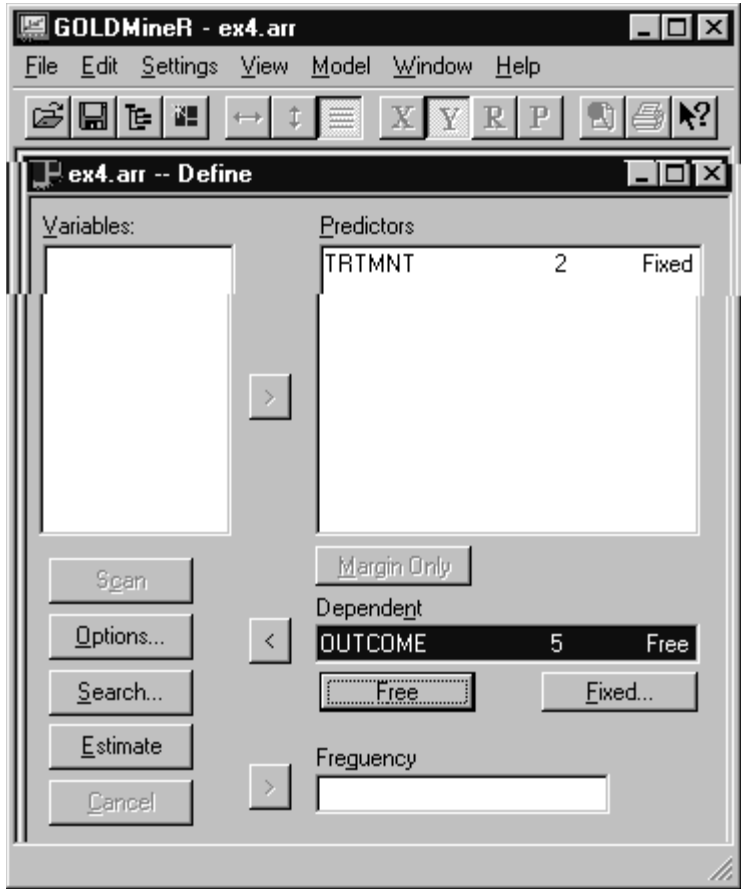
Model

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:

E Make the plot window the active window

E Select the ‘Y’ button on the toolbar or from the Plot menu select Joint-Y

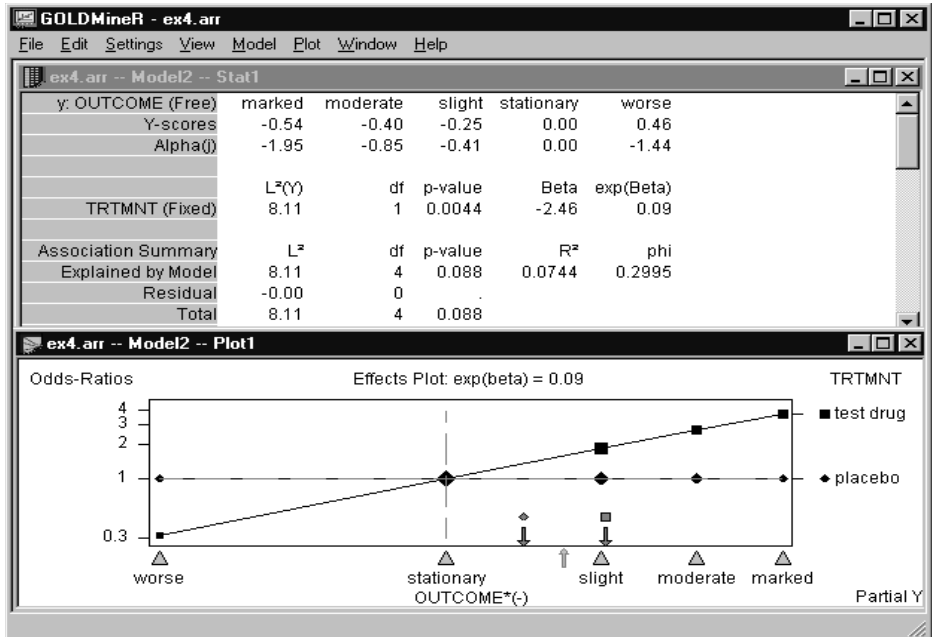
E Double click on the *stationary* marker to set the category weight for OUTCOME.

E Double click on *placebo* within the Category Weights dialog box to set the category weight for TRTMNT.

E 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:

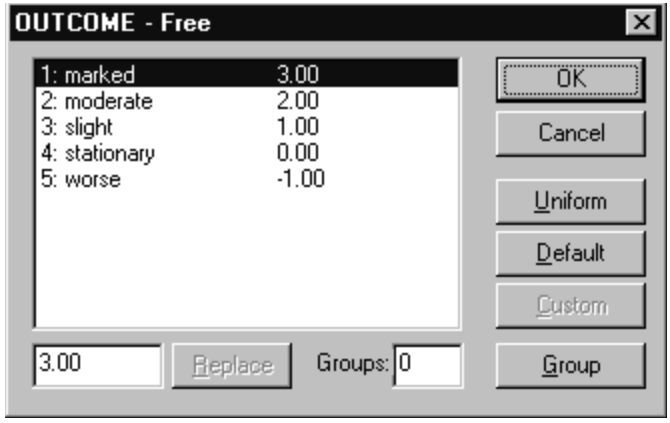
Model

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

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

E Click on OK.

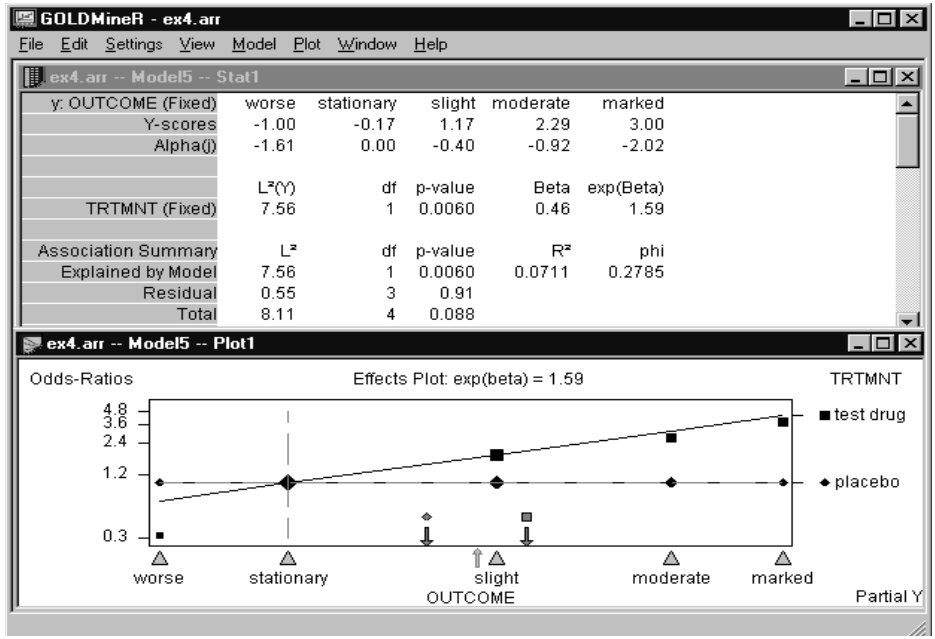
E Click on Estimate to reestimate the model.

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

E Make the plot window the active window.

E 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



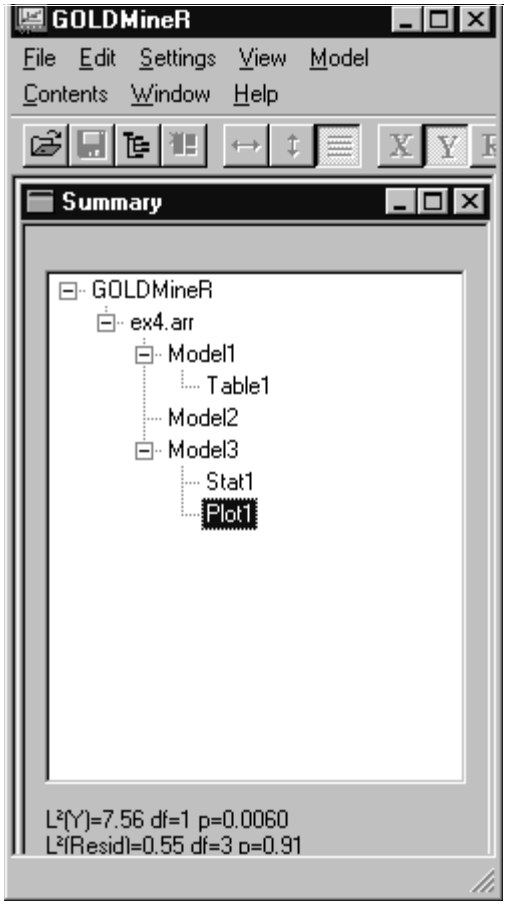
The Statistics window tells you that the current model fit (Residual $L^2 = 0.55$) is slightly worse than that obtained from the first model which assumed equidistant spacing between the outcome categories. The estimated effect is now slightly less significant than before.

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.

E To load the saved file choose:

File

Open

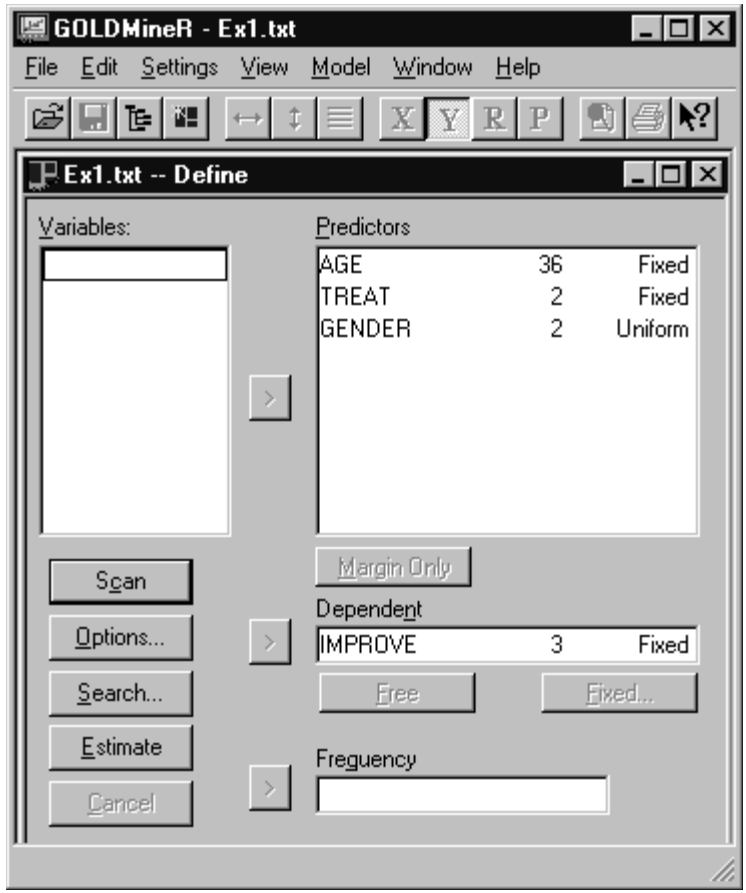
The File Open dialog box will appear. Click on the arrow next to Files of Type and select Data files (.dat or .txt) if this is not already the default listing. A list of all files with those extensions in the current directory will appear. If you copied the tutorial files into a directory other than 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



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

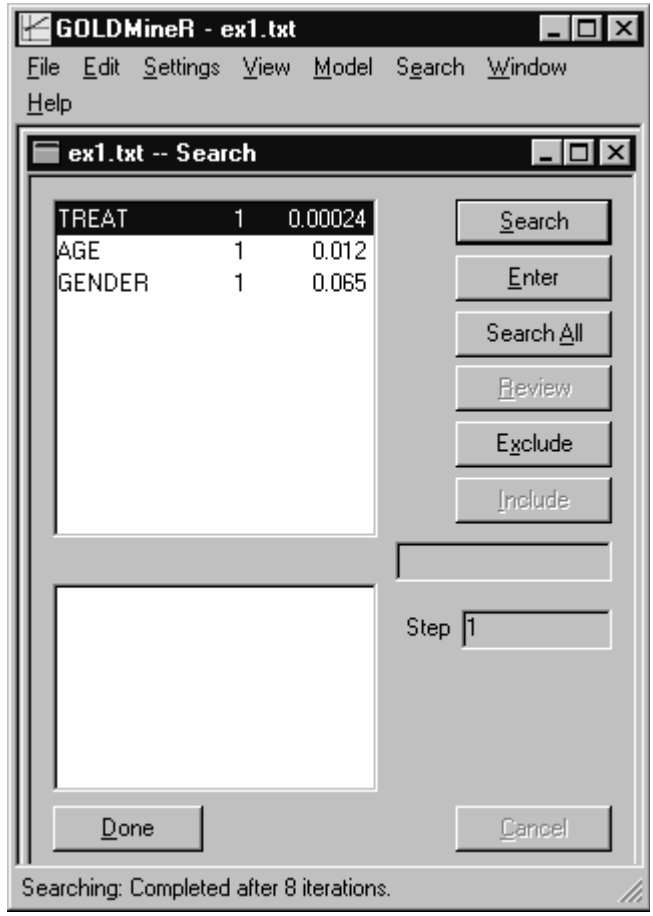
The Search dialog box contains the set of predictors in the upper box.

Figure 4-23
The Search Dialog Box

E 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 1st predictor to enter into the model. The number to the right of

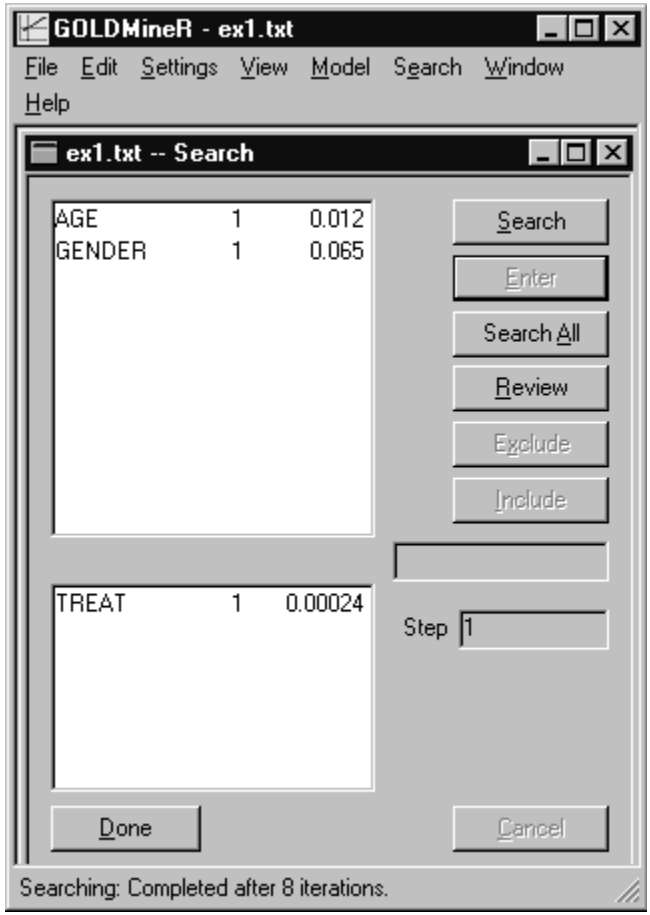
each predictor represents the number of degrees of freedom associated with the p-value for that predictor.

Figure 4-24
After Selecting Search



E 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

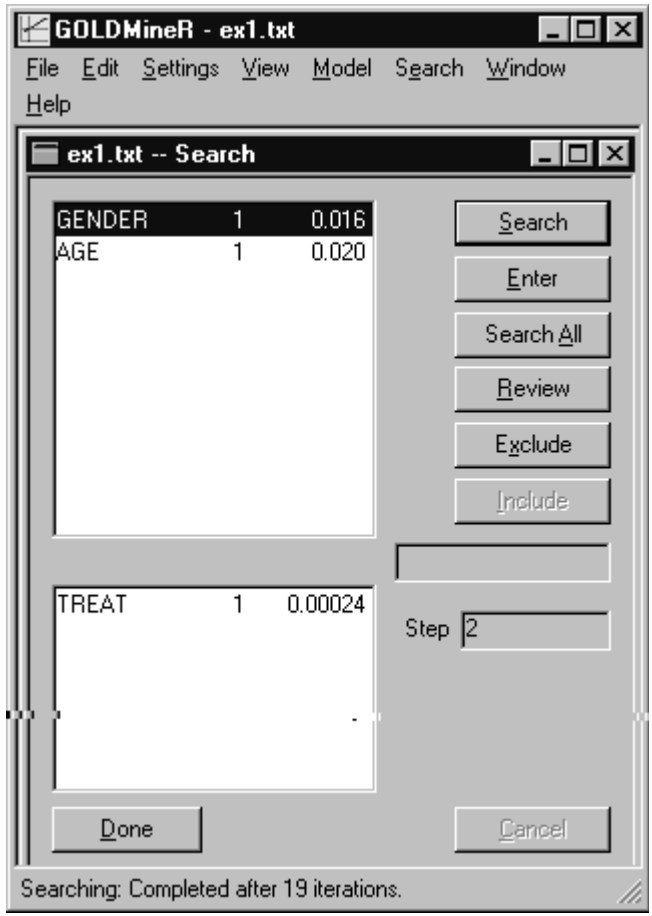


At this time you may examine the output from estimating the current model containing 1 predictor by clicking on Review. The Statistics window (absent the $L^2(Y)$ statistic) and Plot window will appear on your screen.

E 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.)

E Click on Search again to search for the 2nd 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



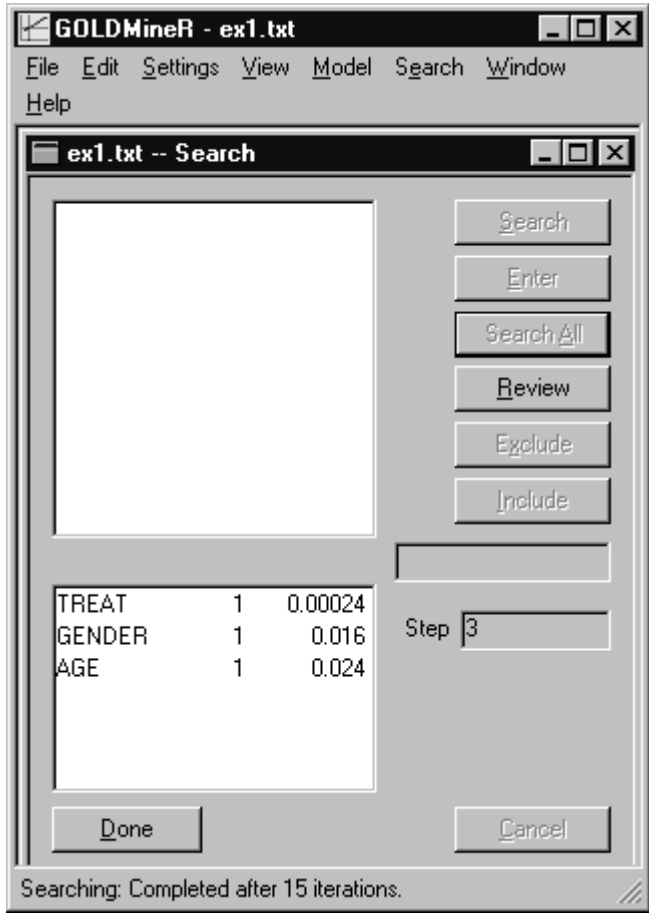
Note that even though GENDER would not be significant as the only predictor (see Figure 4-24), it would be significant (p=0.016) in the two predictor model which

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



E 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 signficance 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 unmangeable. In such cases specialized charts may be quite useful to summarize the results.

In this tutorial we illustrate various tabular summaries for our current applicaton 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

ex1.txt -- Table1

(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

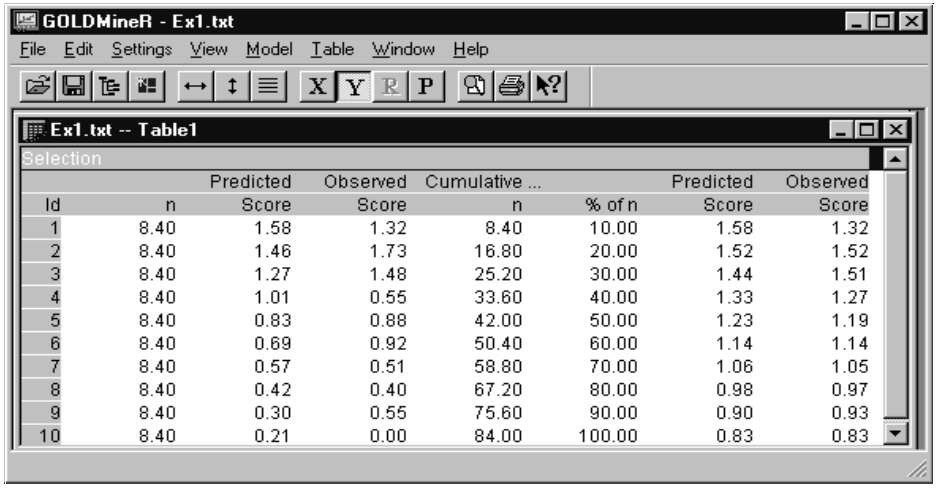
Various subsets of the 63 rows can be examined by constructing different Partial Tables, or the 63 rows can be reduced to a smaller number of X-profile groups using Quantile charts. Next, we illustrate the use of the Quantile chart to create a smaller, more manageable summary table.

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



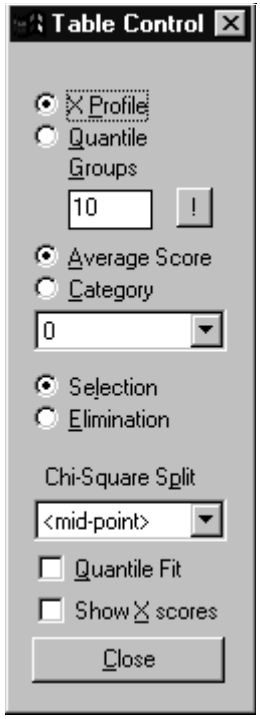
Id	n	Predicted Score	Observed Score	Cumulative ... n	% of n	Predicted Score	Observed Score
1	8.40	1.58	1.32	8.40	10.00	1.58	1.32
2	8.40	1.46	1.73	16.80	20.00	1.52	1.52
3	8.40	1.27	1.48	25.20	30.00	1.44	1.51
4	8.40	1.01	0.55	33.60	40.00	1.33	1.27
5	8.40	0.83	0.88	42.00	50.00	1.23	1.19
6	8.40	0.69	0.92	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

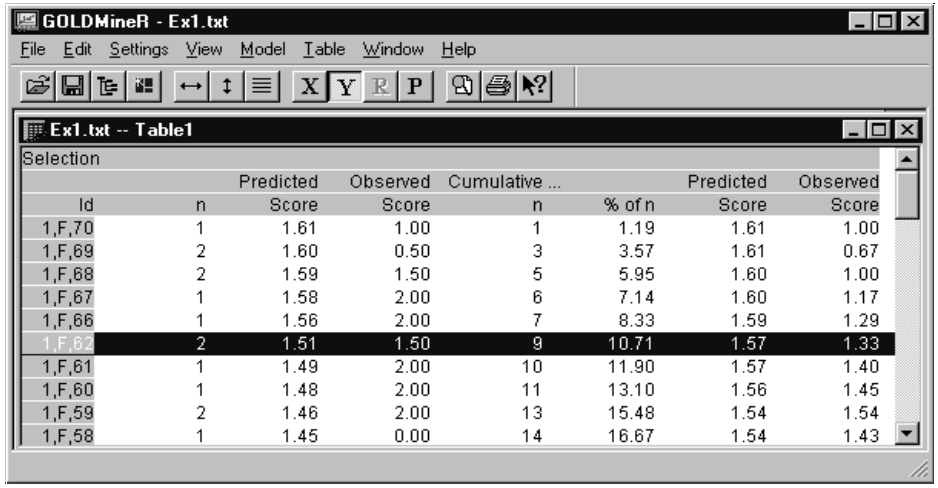
The Profile chart lists each X-profile ranked in order of its predicted Y-value. To replace the quantile chart with a profile chart open the Table Menu and select Control or right click the mouse to bring up the Table control dialog box (which now contains chart options). Click on X Profile.

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^{th} and 9^{th} 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.

5

Formatting your data for a GOLDMineR analysis

GOLDMineR can read input files in any of the following data formats:

- SPSS saved file (*.sav)
- rectangular (cases by variables) ASCII data file (.txt or .dat)
- ASCII array file (*.arr), ("array format")

The two latter formats are described further in this chapter. Regarding SPSS saved files, GOLDMineR can read SPSS variable names and *category* (value) labels (less than 40 characters), but does not support SPSS *variable* labels or date variables. In addition, if a SPSS saved file contains multiple user missing designations for a numeric variable, GOLDMineR will combine them into one 'missing' category. GOLDMineR does not recognize user missing specifications for non-numeric variables.

The advantage of using a SPSS saved file or the array format is that you can assign category labels. If you have a large number of variables or variables with many categories, a SPSS file (or a rectangular file) is preferable to creating a large array file, which can be difficult. For information on creating an SPSS save file, see your SPSS manuals (programs such as DBMS Copy are available to convert SAS and other data formats to the form of a SPSS saved file). Note that for SPSS save files and rectangular files, string variables with lower and upper case letters, such as 'f' and 'F', are distinguished as separate categories.

Rectangular files

A rectangular file can be created in any spreadsheet or wordprocessor using the 'Save as.." command to save the data as a text file with either a .dat or .txt extension. The first line of the file needs to contain the variable names (including a name for the optional frequency variable if a frequency variable is present in the file), followed by the data. For any variable, data may be either numeric or alphanumeric (such as 'Female', 'Male'). The data should be separated by either a space or a tab (no other

delimiters, such as "," are allowed). Each row should contain just one record. For numeric codes, do not use commas (such as 1,092). Below is a partial listing of file ex1.dat (sample dataset #1). This data contains one record per case. Note that records 6 and 7 are identical (0 30 0 F).

IMPROVE	AGE	TREAT	GENDER
0	23	1	F
0	23	0	F
1	27	1	M
0	29	1	M
0	30	1	M
0	30	0	F
0	30	0	F
1	31	0	F
2	32	1	M
0	32	1	F
0	32	0	F
2	33	0	F
1	37	1	F
0	37	0	M
...			

For data sets consisting of many observations that contain identical values on all variables, the inclusion of an optional frequency variable will reduce the number of physical records in the file. Below is a partial listing of the file ex3.dat (sample dataset #3). FREQ contains the total number of observations associated with that combination of categories for the variables. For example, the value of FREQ for the first record, defined by the combination of categories (1, 1, 1) is 2,407.

- The values taken on by a frequency variable will generally be an *integer* representing the number of observations associated with the particular combination of categories.
- GOLDMineR allows a frequency variable to take on fractional values to support certain advanced applications, discussion of which are beyond the scope of this manual. (**Warning!** The use of fractional sampling weights as a frequency variable will generally cause all statistical tests and p-values reported in the statistics output to be incorrect.)

CALL	PAYMENT	RETURN	FREQ
1	1	1	2407
1	2	1	1265
1	3	1	1340
1	4	1	1306
0	1	1	2133
0	2	1	1156
0	3	1	1262
0	4	1	1248
1	1	0	1954
1	2	0	881
1	3	0	809
1	4	0	779
0	1	0	2176
0	2	0	942
0	3	0	897
0	4	0	839

Arrays

An array formatted file can be created in any word processor or editor program. It contains frequency counts for each cell in the multiway table. The general ASCII format is:

```
ARRAY
<d1> <d2> <d3> ...
LABELS
<var>/<cat1>/<cat2>/<cat3>/...
...
SCORES
<var>/<scr1>/<scr2>/<scr3>/...
...
DATA
f1...11 f1...12 f1...13 ...
f1...21 f1...22 f1...23 ...
... fn...n
```

The array below (sample file ex3.arr) reexpresses ex3.dat (shown above in rectangular format) in array format.

```
ARRAY
2 4 2
LABELS
CALL/yes/no
PAYMENT/$1/$2/$3/$4
RETURN/yes/no
SCORES
CALL/1/0
PAYMENT/1/2/3/4
RETURN/1/0
DATA
2407 1954
1265 881
1340 809
1306 779
2133 2176
1156 942
1262 897
1248 839
```

The key words “ARRAY”, “LABELS”, “SCORES” and “DATA” must begin in column 1 and be written in upper case. Following the “LABELS” key word, exactly one label record must appear for each variable. The “SCORES” section of the file, which may be used to assign scores to the categories, is an optional section. If the keyword “SCORES” appears in the file, at least one score record must follow.

The frequency counts are entered on lines following the DATA keyword in free format, and must be entered in the order such that the first dimension varies least rapidly..., and the last dimension most rapidly. (The use of scientific notation for frequency counts is *not* supported.)

Step by Step Instructions for Creating an 'Array Format' File

To create the array file for the following table (based on sample data set #4) follow each step below:

	TRTMNT	
OUTCOME	**Test Drug**	**Placebo**
Marked Improvement	7	3
Moderate Improvement	15	9
Slight Improvement	16	14
Stationary	13	21
Worse	1	5

E **Line 1**: Type ‘ARRAY’ in all caps

E **Line 2**: <dn> are integers which describe the number of categories in each dimension. The number of integers in the list sets the dimensionality of the array. For example, for the Clinical Trial data there are 2 dimensions or variables (var1:OUTCOME var2:TRTMNT,) with five and two categories respectively (var1: marked improvement, moderate improvement, some improvement, stationary, worse, var2: test drug/placebo). Type ‘5 2’.

E **Line 3**: Type ‘LABELS’ in all caps.

E **Lines 4-5**. The LABELS section supplies dimension (i.e., variable names) and category labels. The LABELS sections is optional. If the section is present, there must

be one entry for each dimension; the category labels may be omitted. Each label on an individual line is separated from each other with a forward slash '/'. The label information must be entered as a single "line" of 4096 characters or less for each variable. Continuation of a physical line can be effected by placing a backslash '\' at the end of a line.

For our example, type the following on lines 4-5:

```
OUTCOME/Marked/Moderate/Slight/Stationary/Worse
TRTMNT/Test Drug/Placebo
```

E **Line 6:** Type 'SCORES' in all caps.

E **Lines 7-8:** The SCORES section supplies dimension and category scores and is optional. Unlike the LABELS section, it is not required to supply scores for all dimensions. Each score on an individual line is separated from each other with a forward slash '/'. If no LABELS section is present, the <var> item should be a dimension number. The score information must be entered as a single "line" of 4096 characters or less for each variable. Continuation of a physical line can be effected by placing a backslash '\' at the end of a line. For our example, type the following on lines 7-8:

```
OUTCOME/3/2/1/0/-1
TRTMNT/1/0
```

E **Line 9**: Type 'DATA' in all caps.

E **Line 10 - ?:** The DATA section is required and must supply the frequency counts for the table. The counts are in "free format", separated by spaces, tabs or newlines. Do not use commas (such as 1,092).

Below is the completed array in its entirety, saved as file ex4.arr on your diskette.

```
ARRAY
5 2
LABELS
OUTCOME/marked/moderate/slight/stationary/worse
TRTMNT/test drug/placebo
SCORES
OUTCOME/3/2/1/0/-1
TRTMNT/1/0
DATA
 7  3
15  9
16 14
13 21
 1  5
```

Once created in the form of an SPSS saved file, a rectangular file or an array file, the data can be opened for a GOLDMineR analysis using File, Open (See “Opening a Data File” on page 85). For convenience, once opened the data may be saved in the format of a GOLDMineR saved file (.gsv) along with current or new settings (scores, etc.) using the File Save menu option.

6

Reference

FILE MENU

Use the File menu to open, close, save and print your data file.

Opening a Data File

Use File Open to load your data files and to define your model. You may load an SPSS saved file, a rectangular file, an array file or a previously saved GOLDMineR file.

E From the menus choose:

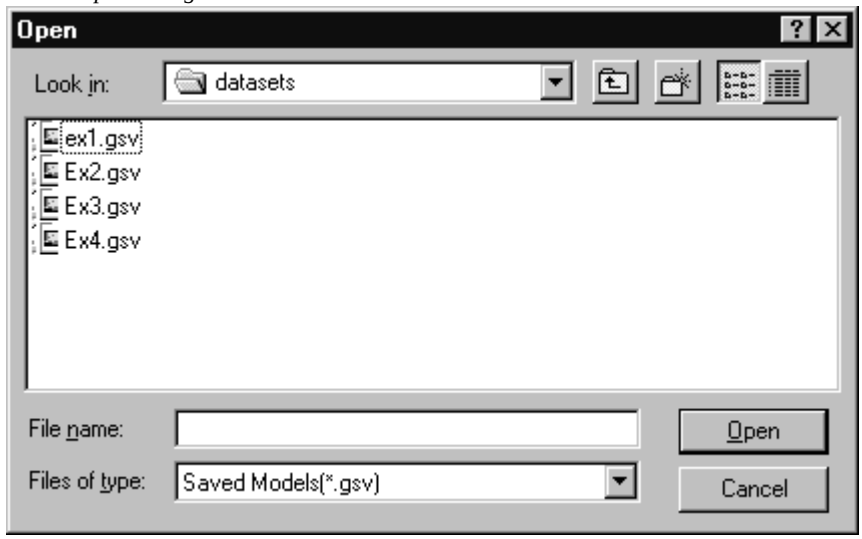
File

Open...

E In the Open File dialog box, select the file you want to open.

E Click Open.

Figure 6-1
The File Open dialog box



File Open Options

Look in: To change drives and directories.

File Name. You can type in a filename, a path and filename or a wildcard search. By default, GOLDMineR looks for all files in the current directory with the same extension as the last file that was opened and displays them on the files list. All files in the directory that match the wildcard search appear on the Files list.

List Files of Type. To specify the type of data file that will be loaded, you can choose one of the following alternatives:

- **GOLDMineR save files**. Loads previously saved analyses. The default extension is .gsv.
- **SPSS system files**. Loads SPSS saved files (through version 8.0). The default extension is .sav. GOLDMineR can read SPSS variable names and *category* (value) labels (less than 40 characters), but does not support SPSS *variable* labels or date variables. In addition, if a SPSS saved file contains multiple user missing designations for a numeric variable, GOLDMineR will combine them into one 'missing' category.

  Since non-numeric variables will generally be treated as nominal using the Free scaling option, users may wish to include SPSS user missing categories as

additional valid categories in an analysis. (Users wishing to omit such cases from an analysis may use GOLDMineR's Missing Value Options.) Hence, GOLDMineR ignores SPSS user missing specifications for non-numeric variables.

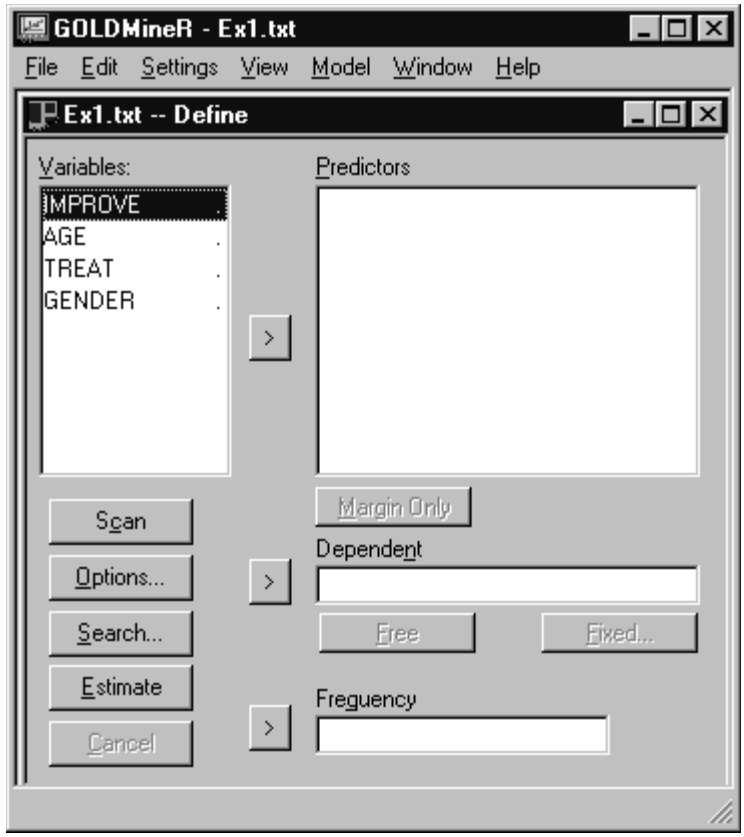
- **Rectangular files**. Loads ASCII rectangular 'text' files. The default extension is .dat or .txt.
- **Array files**. Loads ASCII array files. The default extension is .arr.

For information on creating rectangular and array files, see Chapter 5.

Model Setup

Once you have opened a file, the **Model Define** dialog box will appear. Use this to designate variables, assign category scores, and set options for your analysis as well as for scanning SPSS and text files. (Note that the Model Define dialog box can be viewed and modified at any later time by choosing Model, Define from the menus. Also note that there is only one Model Define dialog box that can be opened for each dataset.) For previously saved GOLDMineR files (*.gsv) the Model Define dialog box will contain the saved settings.

Figure 6-2
Model Define dialog box prior to specifying setup



All eligible variables that may be included in an analysis are listed under **Variables**. Variables may be designated as one of three types: Dependent, Predictor or Frequency.

Defining Variables

You must specify one dependent and at least one predictor variable to begin an analysis.

E Highlight a variable name in the Variables list box.

E Click on the corresponding arrow button to specify the variable as one of the following:

Dependent Variable. Assign one variable to be used as the dependent variable

Predictors. Assign one or more variables to be used as predictors.

Frequency Variable (optional). Assign one variable to be used as a frequency variable. See Chapter 5 for an example that uses a frequency variable.

Note that:

- Double clicking a variable name in the variables list box will send that variable to the **Dependent** box unless the box already contains another variable. Subsequent double clicks send variables to the **Predictors** box.
- A variable designated to weight cases in an SPSS saved file according to the SPSS data dictionary will automatically appear in the **Frequency** box.
- To remove a variable from the **Dependent**, **Predictors** or **Frequency** box, highlight the variable name and click on the corresponding reverse arrow.
- To remove a predictor variable from the model, but retain it for use with the other predictors for assessing the overall model fit, highlight the variable name and click Margin Only. This is an advanced option not generally used. See "Model Define: Margin Only" on page 95 for further details.

Scanning SPSS and Rectangular Files

This option Scans the data file and identifies each distinct category or value for variables in the **Dependent** and **Predictors** boxes (this option is not executed for files in array format). Prior to your data file being scanned, only variable names will be listed in the variable box. Once you have defined your variables (as dependent, predictor or frequency), click on Scan to scan the datafile. The number of distinct categories (or values) now appears along with the scaling option (Free, Fixed) to the right of the assigned predictors. If a Scan has not been selected before clicking on Estimate, a scan will automatically be performed by GOLDMineR prior to beginning the estimation algorithm and default settings will be used to scale the variables in the model.

Variable Scaling Options

The dependent and each predictor variable can be specified as **Free** or **Fixed** depending upon whether category scores for that variable are free to be estimated by the model or are fixed at specified values respectively. The scaling option may be specified by the user for any predictor or dependent variable before or after the file has been scanned, or if not specified, default specifications will be assigned by GOLDMineR following a scan of the file.

- For any variable having K categories, its category scores may be fixed automatically at the equidistant values (k-1)/(K-1), k = 1, 2, …, K by assigning the **Uniform** scaling option (a special case of Fixed) to the variable.
- For variables that are other than dichotomous, GOLDMineR selects **Free** as the default scaling option for both the dependent variable and the predictors unless quantitative values for those variables are found in the input file in which case the default is **Fixed**. If scores for a variable are included in an array file or quantitative values are found by a scan of the data file, and the variable is specified as **Fixed**, the values in the file are assigned to the corresponding categories as scale scores and used in the model.
- For dichotomous variables, all statistical results obtained from model estimation are identical regardless of the scaling option used. By default, dichotomous variables are treated as **Fixed** if scores are found in the file. If scores are not found in the file, the variable is treated as **Uniform** (a special case of Fixed), which assigns the scores 0 and 1.
- For predictor variables specified as **Free**, category-specific parameter estimates, β_k, k = 1, 2, …K, will be estimated in addition to the single parameter β associated with the predictor variable as a whole. In addition, the score $x_k^* = \beta_k / \beta$ will be assigned to the k^{th} predictor category for use in plots.
- For a dependent variable specified as **Free**, a score for each category, y_k^*, is estimated during the analysis and used to establish an ordering of the categories in the Model Summary section of the Statistics window.
- (see sections A.3, A.5, and A.6 in the Appendix for a discussion of identification issues regarding the **Free** scaling option.)

To specify or change the scaling option for a particular variable in the **Dependent** or **Predictors** box, highlight the variable name, then click on the corresponding button to select one of the following two options:

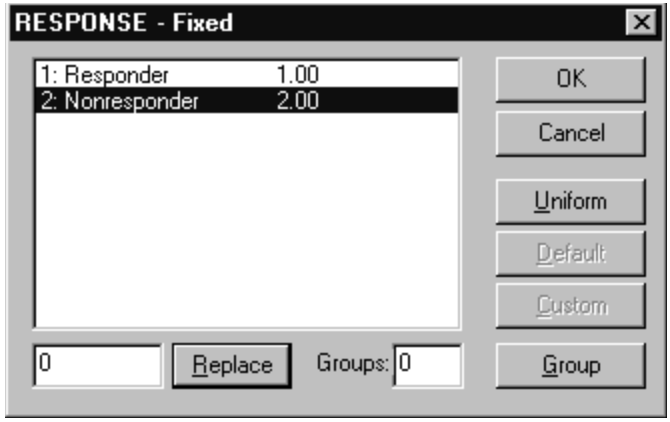
Free. Category scores remain free to be estimated by the model. (Free variables are marked by '*' in the plot window.) In order that the category scores estimated for any Free scale variable be identifiable, two restrictions must be imposed to uniquely define a zero point and a unit. By default, GOLDMineR restricts the range of the estimated scores to be 1, and the weighted average of the scores (based on current category weights) to be 0. If no category weights have been specified, the lowest score is restricted to be 0.

Fixed. Allows users to assign their own custom scores to the categories of a variable. The current scores are listed for each category in the score dialog box. To change a score:

E Highlight a variable

E Select Fixed or double click on the variable name to open the **Fixed Scaling** dialog box. The current category score for a Fixed variable or the Uniform score for Free variables appears to the right of the category label.

Figure 6-3
Fixed scaling dialog box



E Double click on any score you wish to change

E Type in a value for the new score (in Figure 6-3, we are changing '2' to '0')

E Click on Replace.

In Figure 6-3 we change the score for the 'Nonresponder' category of the dependent variable from '2' to '0' in order that the predictions be interpretable as predicted response rates (between 0 and 1) rather than numbers between 1 and 2.

E Click OK to accept the new scores.

The following options are special cases of Fixed scaling:

Uniform. The variable is assigned Fixed integer scores that are equidistant and have a low score of 0 and a high score of 1. For example, for a 3 category variable, the Fixed scores assigned are 0, 0.5, and 1.

Default. If scores have been changed from their original default value (either the original values contained in the data file, scores specified in the array file or if no scores were specified, the default GOLDMineR values), clicking on Default will reset the scores to their default value.

Custom (or user scores). If the user inputs scores (such as we did above), and then later resets them by clicking on Uniform or Default, selecting Custom will reset the scores to their most recent custom setting.

Group. The Group option may be used to reduce the number of original categories for a variable by combining adjacent categories. The score assigned to a group of adjacent categories is the average of the original (default) scores assigned to the original categories. GOLDMineR groups adjacent categories that have fewer observations before grouping categories having larger number of observations towards the goal of having approximately equal sized groups.

The algorithm used is identical to the Rank Cases Tranformation used in SPSS with the number of groups equal to NTILES. The number of groups is specified by the user by replacing the current number in the group box with the number of groups desired and clicking on Group. The actual number of groups will be approximately equal to the number specified. Specifying '0' or a number greater than or equal to the number of original categories restores the variable to its original number of categories.

To use the group option:

E Highlight a variable in the Model Define dialog box that has been selected as the dependent or a predictor variable.

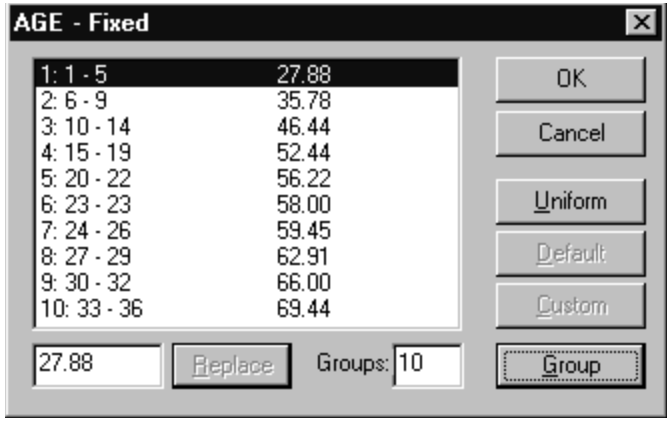
E Select Fixed as the scaling option. The Fixed dialog box will appear

E Type the number of desired groups into the group box

E Click on Group.

The Fixed dialog box changes to show the new groups.

Figure 6-4
Fixed Scaling dialog box with 36 category AGE variable regrouped into 10 categories



In Figure 6-4, the 36-category AGE variable has been merged into 10 groups. New group #1 consists of the original categories 1-5 (abbrieviated as "1:1-5").

E You may click OK to accept the new category grouping and new scores or replace one or more scores with new scores.

The new number of categories will now be listed next to the variable name in the model define window. After grouping, a group variable may be treated as Free by highlighting it in the model define window and clicking on Free.

To ungroup a grouped variable and reset the original category scores:

E Highlight the grouped variable name in the model define window

E Select Fixed to bring up the Fixed dialog box

E Type '0' or a number greater than or equal to the number of original categories in the groups box

E Click on OK

Warning! Use of the Group option will cause any previously set Custom scores to be ignored. Custom scores for the new categories of a grouped variable should be applied *after* the Group option is used.

Missing Values

The special score code '.' may be assigned to one or more categories of a variable to designate those categories as "missing". Missing categories receive special treatment during the estimation process depending upon the option settings (See “Model Define:Options” on page 96.). The '.' score code is assigned in one of the following five ways:

- After a data file has been opened, the missing code '.' may be assigned to any categories or levels of any Fixed variable containing three or more categories/levels in the same manner that scores are assigned.
- For SPSS saved files, values assigned the system missing specification for any variable are identified during the scan process and are recoded to the GOLDMineR missing value code '.'.
- For a rectangular file (.txt, .dat), during the scanning process, if the first code for a given variable is found to be numeric, GOLDMineR assumes the variable to be quantitative and the **Fixed** scaling option is assigned by default. Any non-numeric codes for that variable found subsequently in the input file are assumed to represent missing values and assigned the missing values code '.'. Alternatively, if the first code is found to be non-numeric, the variable is treated as qualitative and the **Free** scaling option is assigned by default. Subsequently, any codes, whether numeric or non-numeric, are treated as valid category labels.
- For an array file, the SCORES option may be used to assign the missing values code '.' to one or more categories of any variable.

Imputing Scores for some (but not all) categories of the Dependent Variable

For any dependent variable with three or more categories, GOLDMineR will replace the special score code '.' for each missing category with an imputed score if different numeric scores have been assigned to two or more categories of the variable and the technical missing value option Y is set to impute. (At least two fixed scores are

required because any scale is meaningless without a zero point and a unit for the purpose of scale identification.)

To have GOLDMineR impute one or more scores for a dependent variable:

E Click on Options. Check that the Missing Value Option for Y is set to impute (the default). Click OK.

E Highlight the variable name in the dependent variable box

E Select Fixed or double click on the variable name.

E Double click on each score(s) you wish to estimate

E Type '.' for each score you wish to estimate

E Click on Replace.

E After assigning the desired scores and '.' to the categories, Click OK.

If fewer than 2 distinct scores have been assigned GOLDMineR will open a message window alerting you that "there must be at least 2 distinct scores for score estimation" .

E Click Estimate to estimate the imputed score(s) (which may be viewed in the output windows).

Model Define: Margin Only

Margin Only (only for predictors). Uses the selected predictor variable in defining the cells used to assess the model fit, but restricts the Beta effect parameter for the predictor to be zero (an advanced feature generally not used). To turn this option off once it has been selected for a particular predictor, highlight the predictor name and assign a scaling option (Free or Fixed).

Use of this feature causes the Total L^2, Residual L^2 and the associated degrees of freedom (but not the explained L^2) to be different than what it would be if the selected predictor variable(s) were removed from the predictor box.

Model Define:Options

E To change selected technical options for your analysis, from the Model define dialog box:

E Click on Options to bring up the Technical Options dialog box

E or when the Model Define dialog box is the current window, from the menus choose:

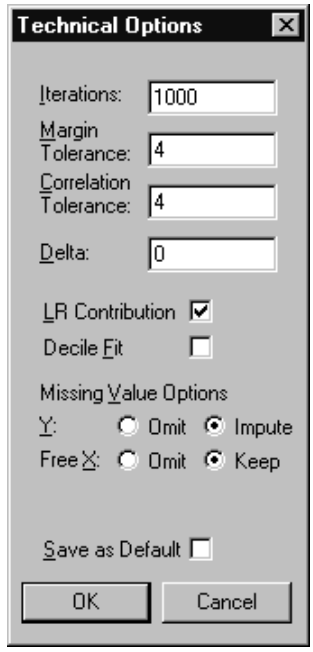
Model

Options

E Change selected options.

E Click on OK.

Figure 6-5
Technical Options dialog box



Changes in any of these options will affect new output obtained from estimating the current model, but will not affect current output windows. It is beneficial to check your technical options settings before estimating a model.

The iterative estimation procedure converges when both the margin tolerance and the correlation tolerance attain the specified limits. The iterative procedure continues until either 1)it achieves convergence or 2) the maximum number of iterations is exceeded. For specific formula, see the maximum likelihood equations in Section A.10.

Iterations. Iterations specifies the maximum number of iterations allowed when estimating a model. The default is 1000. In general, a lower value for iterations may speed up processing by stopping the estimation prior to achieving convergence. However, if the algorithm does not converge, the model parameter estimates and associated statistics may not be accurate. To see if the model has converged after estimating, check the convergence information listed at the bottom of the Model Summary section of the Statistics window.

Margin Tolerance. The maximum likelihood equations for the current model requires that certain margins (the ML margins) be fit. For example, the estimated expected frequency count for each observed X-profile (the estimated expected X-margin) must equal the corresponding observed frequency (observed X-margin), and the estimated expected count for each dependent variable category (estimated expected Y-margin) must equal the corresponding observed frequency (observed Y-margin). The Margin Tolerance option requires that the iterative estimation algorithm continue until either each observed and expected "ML margin" agrees to the number of decimal places specified by the margin tolerance or the maximum number of iterations is exceeded. The maximum value is 13 and the default value is 4. In general, a lower value for margin tolerance speeds up processing. To see if the model has converged after estimating, check the convergence information listed at the bottom of the Model Summary section of the Statistics window. If the algorithm does not converge the model parameter estimates and related statistics may not be accurate.

Correlation Tolerance. The maximum likelihood equations require that the observed correlation between Y and the joint X variable ($X = \sum\beta_m X_m$) be identical to the correlation calculated using the estimated expected frequency counts in place of the observed frequency counts. In the program, convergence is achieved if the observed and expected correlations agree to the number of decimal places specified by the correlation tolerance. (The correlation tolerance criterion is used in the convergence for all models except the independence model for which the expected correlation is always 0). The maximum value is 13 and the default value is 4. In general, a lower value for correlation tolerance speeds up processing. To see if the model has converged after estimating, check the convergence information listed at the bottom of the Model Summary section of the Statistics window. If the algorithm does not converge to the desired tolerance, the model results may not be accurate.

Delta. Set Delta by specifying a non-negative value “d” or by entering a non-negative value followed by a percent sign, “d%”, to express Delta as a percentage of the total sample size. In the former case, Delta = d; in the latter, Delta = [(d/100)N], where N is the sample size. The default value for Delta is 0. A nonzero Delta value is used to solve identification problems that may occur from empty cells.

Warning! Use of a nonzero value for Delta alters the original data in a way that reduces the significance of any effects. For technical details, refer to the “Delta Option” on page 223).

LR Contribution ("L^2(Y)"). When this option is selected (default), GOLDMineR estimates one additional model for each predictor included in the current model to assess the unique contribution of that predictor to the model chi-square (“Explained by

Model L^2”). $L^2(Y)$ is calculated as the difference between the model chi-square under the current model and under the comparison model where the predictor is treated as "Margin Only and is reported under the "$L^2(Y)$" column heading for each predictor in the Statistics window. If the dependent variable is designated as Free, scores for the dependent variable in the comparison model are fixed to be identical to the estimated scores obtained for the current model. (Selection of this option may increase the time to perform an analysis by a substantial amount, especially where there are many predictors. When this option is not selected, users can rely upon the Wald statistic to assess the contribution of each predictor.) For specific formulas, See “A.8 Partitioning Chi-Square” on page 226.

$L^2(Y)$ is not reported in the Statistics window when the Search procedure is used to define a model.

Decile Fit. When this option is selected, the decile fit statistic is computed and displayed in the statistics output. The decile fit may be used as an alternative to the Residual L^2 and Residual X^2 model fit statistics when analyzing sparse data**.** For technical details, See “A.15: Assessing Model Fit when Analyzing Sparse Data” on page 244.

Missing Value Options

During model estimation, cases assigned the missing value code ‘.’ in place of a quantitative value (Y-score or X-score) on a variable are treated in accordance with the missing value options that are in effect. Different options are available for dependent variables, Free predictor variables and Fixed predictors.

Y-score options (Y) for the Dependent Variable

Omit. Exclude from the analysis cases coded missing on the dependent variable and reduce the number of dependent variable categories accordingly. For example, if 2 of 5 dependent variable categories are assigned the missing value code, any cases included in those 2 categories will be omitted from the analysis and all output will reflect the resulting 3-category dependent variable.

Impute (default): Retain all categories of the dependent variable in the analysis and estimate a Y-score for each category assigned the missing value code.

Missing Value Options for Predictors

For Fixed predictors, cases assigned the missing value code on one or more Fixed predictors are always excluded from the analysis.

For Free predictors, cases assigned the missing value code on one or more Free predictors are treated based on the Free X option in effect:

Free X options

Omit. Cases assigned the missing value code on one or more Free predictors are excluded from the analysis.

Keep (default): Retain all categories of any Free predictors in the analysis and estimate an X-score (and associated category-specific Beta) for each such category.

Save as Default

Save current technical settings as defaults.

The Search Procedure

For information on the Search Procedure, see “SEARCH MENU” on page 151.

Estimating Your Model

After your model has been defined in the Model Define dialog box, click on Estimate (or select Model, Estimate when the Model Define dialog box is the current window) to have GOLDMineR estimate the model, the results of which are displayed in the Statistics, Plot and Table windows.

During the estimation procedure, the status bar desplays two measures of the progress towards convergence:

- the current iteration number
- the ratio of 1) the linear R^2 based on the current estimates for the expected frequencies to 2) the linear R^2 based on the observed frequencies, expressed as a percentage. As the estimation procedure approaches convergence, the ratio approaches 100%

Cancelling Model Estimation

To terminate the model estimation process once it has begun, select Cancel. Cancel is available from the time estimation is begun (by clicking on Estimate) until the estimation is completed. If Cancel is selected during the estimation of the full model, the algorithm will complete the current iteration (h) and produce output. The message '*Failed to converge after h iterations'* will appear in the Statistics window. If cancel is selected prior to estimation of the full model, the program will terminate immediately without producing output.

Close

E To close a previously opened GOLDMineR data file, click on any output window based on the data file and from the menus choose:

File

Close

GOLDMineR will close all windows (Plot, Table, Statistics, Define) pertaining to the current file listed at the top of the active window.

Save

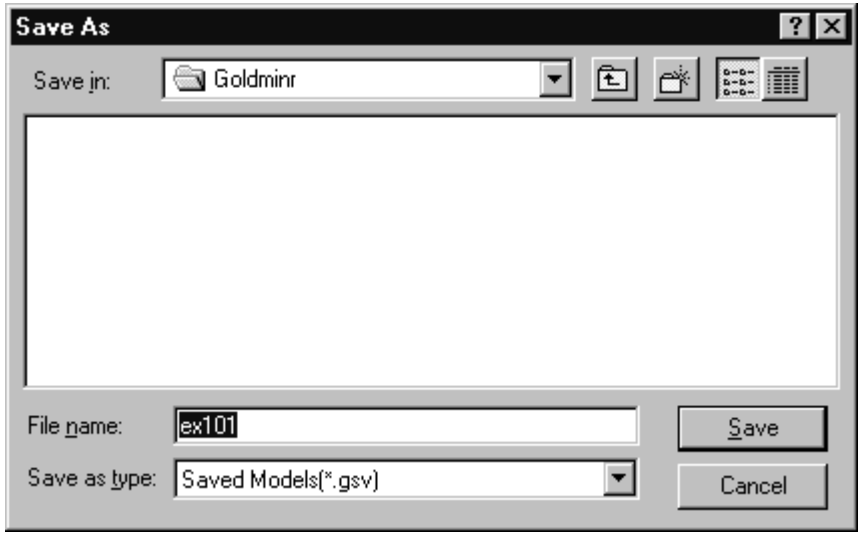
E To save the current GOLDMineR analysis settings (i.e., the analysis represented by the active window when save is implemented) from the menus choose:

File

Save...

E Enter the name of the document and click Save (GOLDMineR will prompt you with a default name and a .gsv extension).

Figure 6-6
Save As dialog box



The following information is saved:

- Current data file associated with the active window
- Dependent variable selected and scaling options for the dependent variable
- Predictors selected and scaling options selected for each predictor
- Technical options specified for this analysis

Note that specific Plot, Table and Statistics windows are not saved.

Upon opening a saved file, the Define Model dialog box will appear with the saved options implemented.

Print

Use this option to print the output from any Statistics, Plot or Table window. The name of the data file and model number will appear at the bottom of each page of printed output.

E Select either Plot, Statistics or Table as your active window.

E From the menus choose:

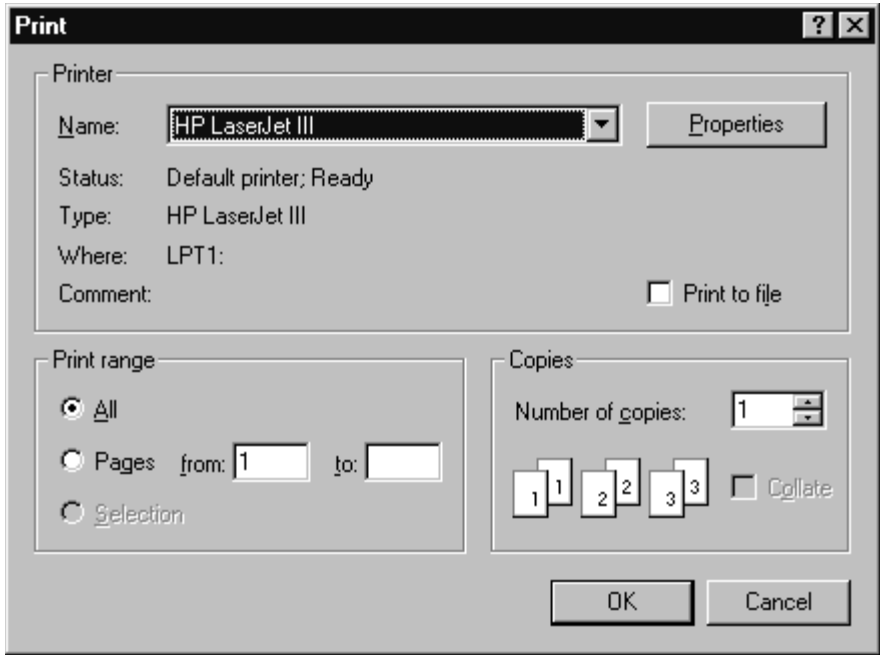
File

Print...

E Select the Print settings you want.

E Click OK to print.

Figure 6-7
Print dialog box



The following options allow you to specify how the document should be printed:

Printer. This is the active printer and printer connection.

Print Range. Currently, all GOLDMineR output is printed on as many pages as needed. Plots are always 1 page.

Copies. Specify the number of copies you want to print for the above page range.

Collate Copies. Prints copies in page number order, instead of separated multiple copies of each page.

Properties. Select Properties to choose from such option as orientation, paper size and source.

Print Preview

Use this command to display the active view as it would appear when printed. When you choose this command, the main window will be replaced with a print preview window in which one or two pages will be displayed in their printed format. The print preview toolbar offers you options to view either one or two pages at a time; move back and forth through the document; zoom in and out of pages; and initiate a print job.

The **print preview toolbar** offers you the following options:

Print. Bring up the print dialog box, to start a print job.

Next Page. Preview the next printed page.

Prev Page. Preview the previous printed page.

One Page / Two Page. Preview one or two printed pages at a time.

Zoom In. Take a closer look at the printed page.

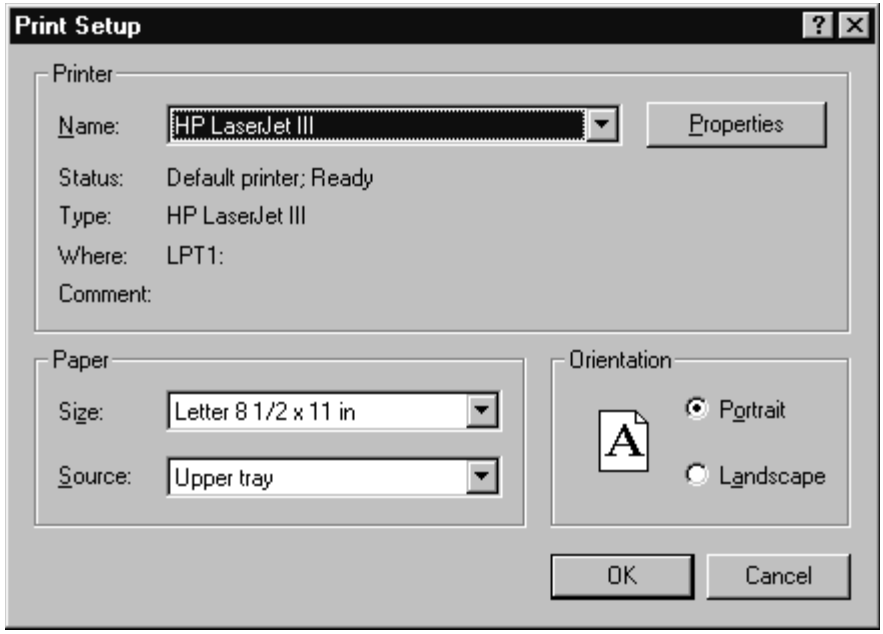
Zoom Out. Take a larger look at the printed page.

Close. Return from print preview to the editing window.

Print Setup

Used to set various options regarding printing, such as printer, orientation, paper size and other options.

Figure 6-8
Print Setup dialog box



The following options allow you to select the destination printer and its connection.

Name. Select the printer you want to use. Choose the Default Printer; or choose the Specific Printer option and select one of the current installed printers shown in the box. You install printers and configure ports using the Windows Control Panel.

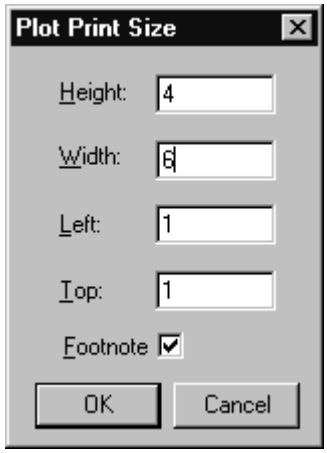
Orientation. Choose Portrait or Landscape.

Properties. Displays a dialog box where you can make additional choices about printing, specific to the type of printer you have selected.

Page Setup

Sets the size and location for printed plots and adds a footnote to all printed pages of output.

Figure 6-9
Plot Print Size dialog box



The following options are available (default starting values in inches are shown in "()"):

Height of plot (4).

Width of plot (6).

Left margin (1).

Top margin (1).

The **Footnote** option lists the data file and model at the bottom of the printed page for Statistics, Plot and Table windows.

Numbered Files

The numbered entries record the most recently opened files. Selection of one of these opens the indicated file.

Exit

E To leave the GOLDMineR program, from the menus choose:

File

Exit

EDIT MENU

Use the Edit Menu to cut and paste plots and tables into other programs or to use the stand-alone tools for calculations of chi-square and predicted Y-score.

Copy

Enabled when a model is active. The format of data sent to the Clipboard depends on the window:

- for the Plot window, a Windows Metafile picture of the plot is placed in the Clipboard
- for the Table and Statistics windows, tab-delimited text is made available.

Prob-Chi (Chi-Square Calculator) stand-alone utility

This tool allows you to obtain a p-value for a given chi-square (or vice-versa) for a specified number of degrees of freedom, DF. To use this tool:

E From the menus choose

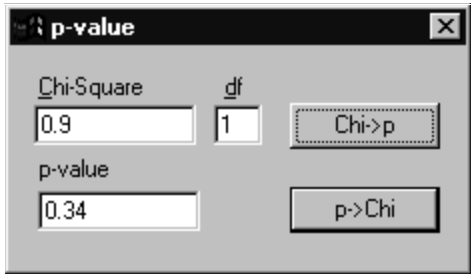
Edit

ProbChi...

E Enter a Chi-square value (or p-value) and number of degrees of freedom.

E Click on Chi->p (or p->Chi).

Figure 6-10
ProbChi dialog box



- Given two models, one a restricted form of another model, this tool can be used to test the significance of those restrictions by using the difference of the likelihood ratio fit statistics for the two models, Residual L^2(B) - Residual L^2(A) with DF = DF(B) - DF(A) degrees of freedom, and selecting chi->p to see if the difference in the chi-squares is significant.

Example: Using results obtained from estimating Models A and B for sample dataset #1 (see Table 7-2), we obtain a likelihood ratio difference of 25.3 - 24.4 = 0.9, DF = 4 - 3= 1. When input, we obtain a p-value of 0.34, a nonsignificant chi-squared difference.

Expected Y stand-alone utility

For a given value for the standardized U-score associated with the joint X variable (**Joint U**), this tool calculates:

- the monotonic regression prediction for Y (**Y**).
- the linear regression prediction for Y (**OLS Y**).
- the derivative of the monotonic regression curve (**deriv.**) in either unstandardized (default) or standardized units.

This utility calculates the monotonic prediction for Y for each X-profile, entered one at a time. For the definition of U See “A.5 Model Identification Issues” on page 218.) Values for U can be obtained for any observed X-profile from the Profile chart (see “Specialized Charts” on page 145) or by clicking on the triangular markers beneath the joint X-view effects or joint regression plots. To use this tool:

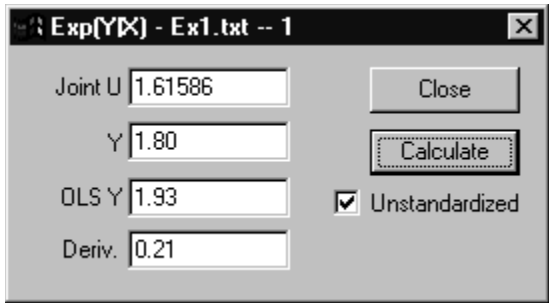
E From the menus choose

Edit

Expected Y...

E Enter a joint U-score in the **Joint U** box and click on Calculate.

Figure 6-11
Expected Y dialog box



The following option is available:

Unstandardized Scores. The units for Y are set to unstandardized when this option is selected. Otherwise, the scores will be in standardized units. By default, this option will be set to coincide with the Unstandardized score setting in the View Menu.

To score a file by saving the predicted Y for each X-profile corresponding to cases in an external file, see “SPSS Code” on page 123.

SETTINGS MENU

Use the Settings Menu to change fonts, colors, and formats.

Font

Use the Select Font dialog box to pick a font and its characteristics for the indicated device. Font selections are maintained separately for the screen and printer devices. For plot views, the size of the plot components is proportional to the size of the

selected font. Any changes made to screen fonts will only be in effect for new windows that are opened after the change has been made.

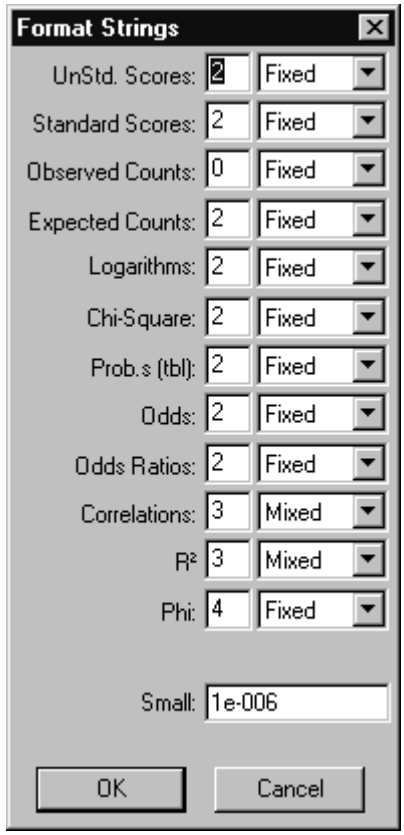
Colors

- The plot components box selects the plot component; the space to the right of the component name shows the current effective color.
- The color assignments box selects colors by reference to the user's Windows system color settings (set with the Control Panel) and also shows the actual color. The entry <custom> at the top of the list allows the user to select a color from the palette.
- When <custom> is selected for a plot component, the user must click in a box in the palette to the right. For such a selection to take effect, the Set button must be clicked. This button is enabled only when this operation is required.
- **Default** resets all assignments to internal defaults.
- Colors are set independently for the printer and screen devices. The device is indicated in the dialog caption.

Formats

The format specification contains fields that allow you to set the precision and type for twelve different kinds of output quantities. For example, the fields identified by the label "UnStd Scores" control the format specification for the unstandardized scores and the optional design matrix listed in the Statistics window, the values that appear along the vertical axis in the joint regression plot, beneath the markers on the horizontal axis of the plots (when the continous display option is selected from the plot control dialog box), and the values that appear on the status bar when clicking on either the markers (when the continuous display option is not selected) or the prediction arrows.

Figure 6-12
Format Strings dialog box



In the Statistics window formats for the:

- *Alpha*, *Beta* and *Std. Err.* are set according to the Logarithms format.
- *Wald* statistic is set according to the Chi-Square format.
- *exp(Beta)*, *Lower* and *Upper* are set according to the Odds format. Note: Since odds ratios and exp(β) may become very large, users may wish to change the default type from Fixed to Mixed for the Odds Ratios format.

In the Table window:

- *weights* are set according to the Unstd. Scores format.
- *Adjusted Residuals* are set according to the Odds Ratios format.

The format specification is controlled by the following two fields:

precision. Valid range of numbers to choose from is 0-20.

- For Fixed and Scientific types, this number specifies the number of digits after the decimal point.
- For Mixed types this number specifies the maximum number of significant digits with trailing zeroes truncated.

type. Choose from the following options:

- **Fixed** (fixed decimal place format)
- **Scientific** (exponential format)
- **Mixed** (treated as Fixed if value is in range 10**-4 ... 10**<precision>, otherwise treated as Scientific)

An additional option at the bottom of the Format dialog box controls when to display zero in the output:

Small. Values less than or equal to this value are displayed as zero in the output. The default value is $1e^{-.006}$.

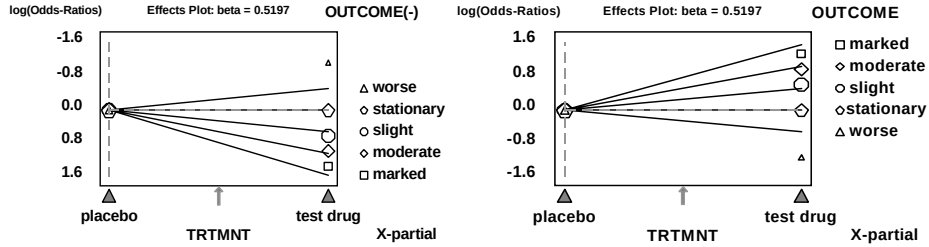
VIEW MENU

Use the View Menu to select from various plot options.

Reverse Vertical

This command is used to reverse the order of categories on the vertical axis (Y-axis) in the effects plot and/or the rows in a Table window. For example, if OUTCOME is on the vertical axis and listed from top to bottom as *worse*.…*marked*, selecting Reverse Vertical will change the order to be *marked*…..*worse*.

Figure 6-13
Effects of Using Reverse Vertical

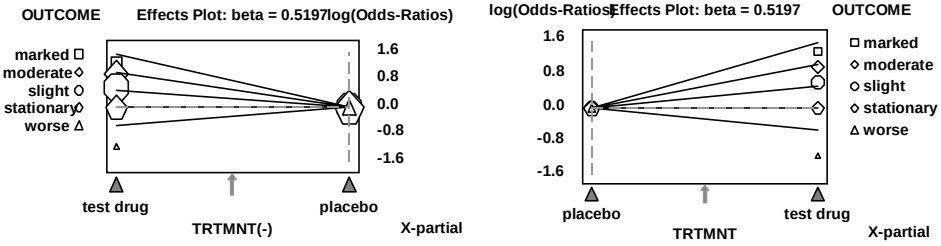


Note: Using Reverse Vertical does not alter any parameter estimates, scores or other statistical output.

Reverse Horizontal

This command is used to reverse the order of categories on the horizontal axis (X-axis) in the effects plot and/or the columns in a Table window. For example, if TRTMNT is on the horizontal axis and listed from left to right as *test drug* then *placebo*, selecting Reverse Horizontal will change the order to be *placebo* then *test drug*.

Figure 6-14
Effects of Using Reverse Horizontal



Note: Using Reverse Horizontal does not alter any parameter estimates, scores or other statistical output.

Standard vs. Logarithmic

Used to select the scale that appears next to a vertical border at the left or right of an Effects plot and for the cell entries in the Table window. To change the unit type for Tables only, See “Control” on page 139. Select one of the following:

Standard (or “multiplicative”). For an effects plot, this option causes the scale for the odds ratios along the vertical axis to be displayed in standard odds ratios units, without altering the effects plot itself. For tables, current cell entries will be displayed in standard units.

Logarithmic (default). For an effects plot, this options causes the scale for odds ratios along the vertical axis to be displayed in logarithmic units. For Tables, current cell entries will be displayed in logarithmic units.

Unstandardized Scores

When selected (the default), scores for individual variables (X-scores, Y-scores and avg. Y-scores), are displayed in unstandardized units. When not selected, scores are displayed in standardized units (U-scores, V-scores and avg. V-scores). This option does not affect the Statistics output where both standardized and unstandardized scores are displayed nor does it affect scores for X-profiles that are always displayed in *standardized* units in joint tables, joint plots and specialized charts.

Tool Bar

The tool bar provides quick, easy access to common tasks. ToolTips provide a brief description of each tool when you put the mouse pointer on the tool. All options, including those on the toolbar, relate to the file that is listed at the top of the active window.

Figure 6-15
Tool bar

File Open

File Save

Show Summary Window

Define New Model

Reverse Horizontal Axis

Reverse Vertical Axis

Show Category Weights

Change to X-View Effects Plot/Table

Change to Y-View Effects Plot/Table

Change to Regression Plot

Change to Partial View

Print Preview

Print

Get Help using mouse cursor

To Show or Hide the Tool Bar

E From the menus choose:

View

Toolbar

To Move the Tool Bar

E click anywhere in the Tool Bar outside the Tool Bar buttons.

E Drag the Tool Bar where you want it.

- Dragging the Tool Bar to the left or right side of the window attaches the Tool Bar vertically to that side.
- Dragging the Tool Bar to the top or bottom of the window attaches the Tool Bar horizontally.
- Dragging the Tool Bar anywhere other than the window borders creates a detached, floating Tool Bar.

Status Bar

The status bar at the bottom of the GOLDMineR application window provides you with:

- current information about various menu options and the status of a current model estimation.
- basic descriptive information regarding variable labels and values associated with various objects highlighted on an active plot. Objects that can be highlighted include effects lines, observed points, diamond symbols marking regression predictions and the triangular markers at the bottom of the effects and regression plot. To highlight a particular object, either click on it or use the arrow + Tab keys to cycle through the plot points.

To Show or Hide the Status Bar

E from the menus choose:

View

Status Bar

Synch Views

When this option is checked (the default), all currently opened Plot and Table windows associated with the current model will be synchronized. Synchronized Plot and Table windows are all updated simultaneously as any of the following options are changed:

View Menu

- Reverse Horizontal
- Reverse Vertical
- Standard vs. Logarithmic
- Unstandardized

In addition, all currently opened Plot windows associated with the current model will be updated simultaneously if the following option is changed:

Plot Menu

- Label Display

When this option is not checked, changes to the above menu items affect only the active Plot or Table window.

MODEL MENU

Use the model menu to define a model, view results of previously estimated models, set category weights or to output to a file a design matrix or SPSS code to score an external data file.

Define

See “Model Setup” on p. 87 for details.

The following Model Menu options are only active when the Model Define dialog box is open. For details on these options, see“Model Setup” on p. 87.

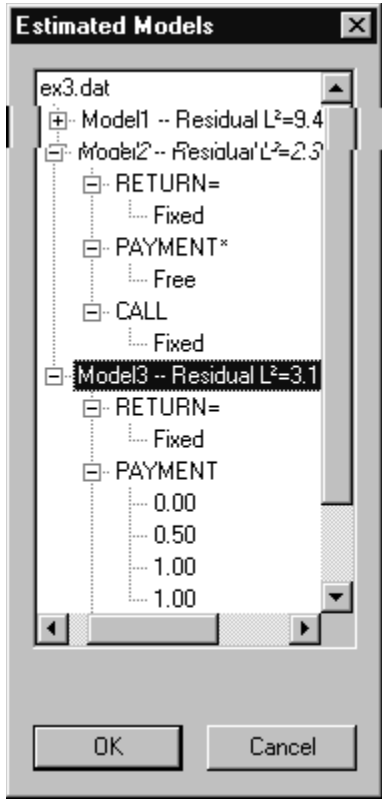
- Scan

- Options
- Estimate
- Search
- Cancel

History

Select History to view the Estimated Models dialog box which contains a listing of each individual model that has been estimated on the current data set along with the Residual L^2 model fit statistic, degrees of freedom, and p-value.

Figure 6-16
Model History dialog box



- To help distinguish between models estimated, Click on the +- key which expands the model listing to display the dependent variable and predictors included in a particular model and their scaling (Fixed or Free). If custom scores were used in the model for one or more variables, they will be listed also.
- The contents of any active output window can be replaced with the corresponding output associated with a different model by highlighting that model in the Estimated Models dialog box and double clicking on it.

Category Weights

A vector of category weights is defined for the dependent variable, denoted E_0, and for each predictor, denoted C_0. Specifically, the vector associated with a variable having K categories is an ordered K-tuple of nonzero positive category weights which sum to one. The category weights define the baseline references for calculating the *odds* and *odds ratios*, the *Alpha* parameters, the contrast coding for qualitative predictors, the reference scores y_0 and x_0 used to define the relative scores and identify the level of each predictor variable being held constant in the partial plots and partial tables. The weights applied to each category of a variable are displayed in the table window under the heading marked 'weights'. Reference scores y_0 and x_0 are used to define the axes in the plots (displayed using dashed vertical and horizontal lines).

For a qualitative (**Free**) predictor variable, C_0 determines the contrast coding used for the category-specific beta parameter estimates -- dummy coding, effect coding or other coding. For example, dummy coding for a Free dichotomous predictor with the second category as the reference is defined by $C_0 = (0,1)$. Alternatively, the contrast $C_0 = (1,0)$ represents the dummy coding with the first category as the reference. ANOVA-like effect coding is represented by equal category weights (**Unweighted** option). For example, for a dichotomous predictor, the ANOVA contrast is defined by $C_0 = (0.5, 0.5)$. For further details see chapter 9 and sections A.1-A.3, A.11 and A.13 in the Appendix.

Figure 6-17
Nine Alternative Reference Perspectives for Displaying the Effect in a 2 x 2 Table TRT (t,p) x RESP (IMP, NC)

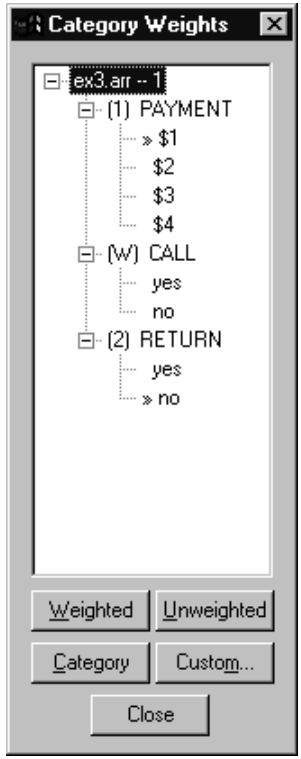
[have SPSS insert]

Assigning Category Weights

When you select Model, Category Weights, the Category Weights dialog box appears with a list of all the variables and their categories. A symbol identifying the weighting selections currently in effect is listed to the left of the variable name. The symbol (“k”) identifies the kth category of the variable as the reference category (that category will be marked by a ‘>>’). By default, if scores are associated with a variable either from values in the input data file or by assigning Fixed scores, GOLDMineR will use the weighting (0,…,1….,0) for that variable where the category/level of that variable assigned the score "0" is taken as the reference and assigned a weight of 1. If no category has been assigned a score of "0", by default, the first category of the variable is used as the base and assigned a weight of 1.

Changing the weights does not effect the model statistics (L^2, R^2, ϕ, *p-values*) or the *Beta* effect estimates, only the category-specific estimates (i.e., the *Beta(k)*), the baseline logits [*Alpha(j)*], the *log-odds ratios* and the *origin* (reference point) for the effects plots. To view the current weights in use for any variable, see the column and row labeled “weights” in the Table window.

Figure 6-18
Category Weights dialog box



E To change category weights, from the menus choose:

Model

Category Weights...

The Category Weights dialog box will appear.

E Set the category weighting option:

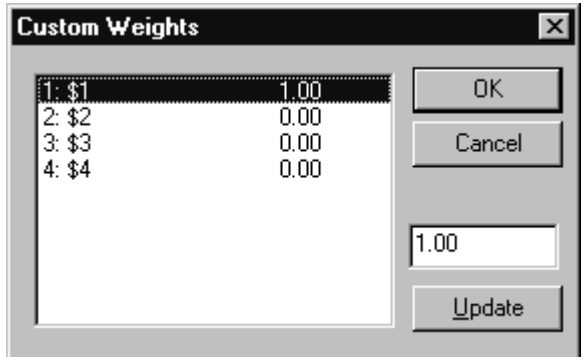
- To assign all the weight to one category of the variable, either 1) highlight the category you wish to select and click on Category, 2) double click the category label or 3) in the effects plot double click the label or marker associated with that category (this last option works only for the variable on the horizontal axis).
- To weight each category by the observed marginal proportion in that category, either 1) click on a variable name and then click on Weighted , 2) double click on

the variable name (to cycle through the Weighted and Unweighted options) or 3) in an Effects plot double click on the average outcome marker (see page 23 for its general location on the plot).

- To weight each category equally, either 1) click on a variable name and then click on Unweighted or 2) double click on the variable name twice (to cycle through the Weighted and Unweighted options).
- To use some other weights:

EHighlight the variable name and click on Custom to bring up the Custom dialog box which contains the current category weights.

Figure 6-19
Custom Weights dialog box



EDouble click on each weight you wish to change

EType in a new weight

EClick on Update

The following options are available:

Weighted (W). Uses sample marginal proportions to weight each variable category. For example, for a dichotomous variable with 20 observations in the first category and 80 observations in the second, C_0 (or E_0) = (0.2,0.8).

Unweighted (U). Assigns equal weights to each variable category. For example, for a dichotomous variable, C_0(or E_0) = (0.5, 0.5). When used with a **Free** predictor, the

unweighted option yields ANOVA-like effect coding interpretations of the category-specific Beta parameter estimates.

Category (k). When used with a **Free** predictor, this option is equivalent to using dummy coding. Select a specific category of the current variable as the base reference.

For example, if you selected category 3 of a 5 category variable as your reference, the contrast would be listed as [3] and would be equal to specifying C_0(or E_0) = (0,0,1,0,0).

Custom (C). Allows direct input of the weights. Users may assign any positive weights to each variable category so long as these weights sum to 1.

For a **Free** predictor variable, the weights define the Beta(k) effect estimates listed in the Category-Specific Parameter Estimates section of the Statistics window. For example, for a **Free** dichotomous predictor, if you wanted $\hat{\beta}_1$ to represent the effect of category 1 relative to an "average" base constructed to be 25% category 1 and 75% category 2, set C_0 = (0.25,0.75).

GOLDMineR will reject any negative weights as well as weights that do not sum to 1. To insure that your weights sum to 1 you may input the special score code '.' for one category and GOLDMineR will automatically replace it with the score that yields a sum of 1. (Custom is an advanced option which is generally not used.)

Design Matrix

Writes to a file the design matrix A used to specify the generalized logit model Ψ = Ab where Ψ is the vector of expected logits under the fitted model and b is the parameter vector consisting of all *Betas* (for **Fixed** predictors), *Beta(k)* (for **Free** predictors) and *Alpha*. See “A.11 Design Matrix” on p. 231 for further description.

SPSS Code

Writes to a file SPSS syntax code for the selected model that can be used to score an external file. The code computes the predicted dependent variable YPREDCTD for cases in the external file based on the values of the predictor variables included in the selected model using the parameter estimates obtained from that model. The code also computes a standardized version of the predicted Y variable (YSTNDRD) which equals YPREDCTD divided by its standard deviation. Predictor variables are recoded according to specifications for any grouped variables if the grouping feature has been used.

In order for YPREDCTD to be computed correctly on the external file it is necessary that the dependent variable and all predictors included in the model have the same variable names in the external file as in the file analyzed by GOLDMineR and that the variables have been coded in the same manner on both files. For example, any recodes used to create the analysis file should be reproduced on the external file prior to executing the syntax code.

Scores for the joint X-variable associated with the model predictors (XSTNDRD) are also computed by the syntax code based on the Weighted Average category weighting for each predictor. XSTNDRD is a measure of the joint X-variable formed by the model predictors. For any X-profile, XSTNDRD equals the sum of the standardized X-scores associated with that X-profile when the Weighted Average weighting option is specified for all predictors in the model. To equate the unit of measurement used for the joint X-variable to be consistent with the current weight settings for the predictors, add the sum of the means of the standardized X-scores to XSTNDRD. These means are displayed in the 'Scores' section of the Statistics window directly beneath each predictor variable. (The mean of any standardized X-score variable will equal zero whenever the Weighted Average weighting option is specified for that predictor.)

The "Weighted Average" category weighting is also used to scale YPREDCTD and YSTNDRD when the dependent variable is scaled as a **Free** variable. The mean score for any Free dependent variable is assumed to be zero for the purpose of computing the predicted value for Y. Since the actual mean score for a Free dependent variable is determined by the choice of category weights, the predictions given by YPREDCTD will differ from those displayed in the Table window by the mean of the unstandardized Y-score. To equate the unit of measurement used for the predicted dependent variable score to be consistent with the current category weighting, add the mean value of the unstandardized Y-score to YPREDCTD. The mean of the unstandardized Y-score is displayed in the SCORES section of the Statistics window directly above the standard deviation. (The mean of the unstandardized and the standardized Y-score variables will always equal zero whenever the Weighted Average weighting option is used).

All computed variables are invariant of the category weight settings specified in GOLDMineR.

Figure 6-20:Sample of SPSS Code with Grouped Variable (“Age”)

```
compute #xu_1 = $SYSMIS .
do   if ( AGE < 44 ) .
compute #xu_1 = -0.7506353 . /*cat.no 1*/
else if ( AGE < 54 ) .
compute #xu_1 = -0.1700097 . /*cat.no 2*/
else if ( AGE < 59 ) .
compute #xu_1 = 0.0978437 . /*cat.no 3*/
else if ( AGE < 64 ) .
compute #xu_1 = 0.2600892 . /*cat.no 4*/
else  .
compute #xu_1 = 0.4945331 . /*cat.no 5*/
end if .
compute #xu_2 =   (TREAT - 0.4880952) * 1.521967 . /* MT_FIXED */
compute #xu_3 = $SYSMIS .
do   if ( GENDER = 'F       ' ) . /*cat.no 1*/
compute #xu_3 = 0.3165111 .
else if ( GENDER = 'M       ' ) . /*cat.no 2*/
compute #xu_3 = -0.7469662 .
end if .
compute xstndrd = 0.0 . /* standardized x-score */
compute xstndrd = xstndrd + #xu_1 .
compute xstndrd = xstndrd + #xu_2 .
compute xstndrd = xstndrd + #xu_3 .
******* #yn_N = F0j[j]/F00 * exp( phi * y * x ) .
compute #yn_1 = 1.409154 * exp( -0.5897802 * xstndrd ) .
compute #yn_2 = 0.5372483 * exp( 0.117956 * xstndrd ) .
compute #yn_3 = 0.8155958 * exp( 0.8256923 * xstndrd ) .
compute #yd_ = 0.0 .
compute #yd_ = #yd_ + #yn_1 .
compute #yd_ = #yd_ + #yn_2 .
compute #yd_ = #yd_ + #yn_3 .
compute ystndrd = 0.0 . /* standardized expected y */
compute ystndrd = ystndrd + #yn_1 * -0.9284767 / #yd_ .
compute ystndrd = ystndrd + #yn_2 * 0.1856953 / #yd_ .
compute ystndrd = ystndrd + #yn_3 * 1.299867 / #yd_ .
******* ypredctd is unstandardized score .
compute ypredctd = ystndrd * 0.8975275 + 0.8333333 .
execute .
```

Upon selecting this option, the Save as dialog box will appear with a suggested name and a default extension of .sps.

For further details regarding the standardized scores see manual Appendix "A.5 Model Identification Issues".

STATISTICS MENU

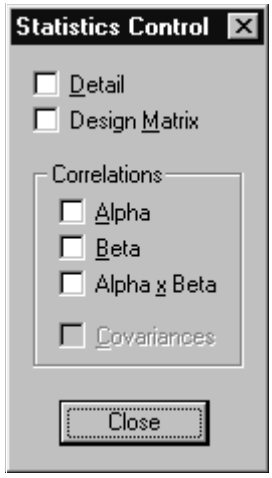
The Statistics Menu is only available when the Statistics window is the active window. For details on the contents of the first three sections of the Statistics window which are

displayed by default, see Chapter 3. Statistics options can be accessed either by the control dialog box or the Statistics Menu.

Control

Selecting Control brings up the Statistics Control Dialog Box. This box may also be accessed by clicking on the right mouse button when the Statistics window is the active window. Statistics Control allows the user to include various technical statistics in the Statistics window.

Figure 6-21
Statistics Control dialog box



The following options are available (and are also listed separately on the menu):

Detail

When this option is checked:

- The Pearson chi-squared statistics, Total X^2 and Residual X^2, are displayed in the Model Summary section of the Statistics window (and the decile fit statistic if 'decile fit' was selected under Options). Residual X^2 may differ substantially from the corresponding likelihood ratio statistic, Residual L^2, in the case that the model does not hold true or in the case of sparse data. In the case of sparse data, the decile

fit statistic may be used to assess model fit. See “A.8 Partitioning Chi-Square” on page 226 for formulae.

- Messages appear at the bottom of the Model Summary section that contain the degree to which the estimated frequency counts reproduce the observed linear correlations(a condition of convergence). A second message is a comparison of the squares of the linear correlation, R^2, calculated from the observed ("Obs") and estimated expected ("Exp") frequencies.

Design Matrix

When this option is checked the design matrix A is displayed. The design matrix is used to estimate the *Alpha* and *Beta* parameters. See “A.11 Design Matrix” on page 231 for details.

Correlations

The following correlations options are available to choose from:

Alpha. When this option is checked the correlation (or covariance) matrix for the *Alpha* parameter estimates is displayed. In addition, if J>8, selection of this option also causes the standard errors for the *Alpha* estimates α_j, j=9,10,…,J to be included in the category-specific parameter estimates section.

Beta. When this option is checked the correlation (or covariance) matrix for the *Beta* parameter estimates is displayed.

Alpha x Beta. When this option is checked the correlations (or covariances) between the *Alpha* and *Beta* parameter estimates is displayed.

Covariances. When this option is checked covariances rather than correlations are displayed when the *Alpha*, *Beta*, or *Alpha* by *Beta* options are checked.

PLOT MENU

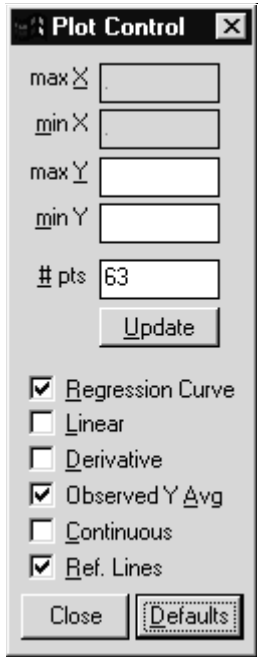
The Plot menu is only available when a GOLDMineR plot is the active window. For details on the contents of the specific plots, see Chapter 3.

Control

Selecting Control brings up the Plot Control Dialog Box appropriate for that plot. This box may also be accessed by clicking the right mouse button when a plot is the active window. Plot Control allows users to customize the appearance of the Plot view. The plot control options for Effects plots and Regression plots are somewhat different.

Plot Control for Regression Plots

Figure 6-22
Plot Control for Regression Plots



The following Plot Control options are available for partial and joint Regression plots unless specified otherwise:

Setting the Scale

max X:, min X (Joint Regression Plot only). Determines the domain of the X-profiles to be plotted along the horizontal dimension of the regression plot by specifying the minimum and maximum U-scores for the X-profiles. Changes take effect only after Update is clicked.

max Y:, min Y. Determines the range for the Y-values plotted along the vertical dimension of the Regression plot. Changes take effect only after Update is clicked.

#pnts. Controls the number of points on the monotonic regression curve to be plotted between the minimum and maximum points. Changes take effect only after Update is clicked.

Update. You need to click on Update whenever you change a min, max or #pnts. value. (You do not need to click on Update for changes to the check boxes).

Specifying Which Features to Include in Plot

Regression Curve. Plots the predicted Y-score according to the monotonic function of the predictors given by the fitted model.

Linear. Overlays the Ordinary Least Squares Regression line. Plots the linear regression predictions for Y as a function of the U-score.

Derivative. Plots the derivative of the predicted Y-score as a function of X which equals ϕVAR(Y|X). (The derivative curve displayed is scaled by subtracting the mean Y-score.)

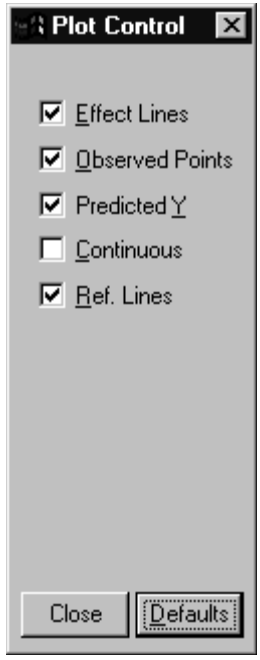
Observed Y Avg. The average Y-score is calculated for each X-profile observed in the sample and displayed as a circle with diameter proportional to the number of observations with that X-profile. Shaded circles correspond to observed average Y-values that differ significantly (at the .05 level) from the monotonic prediction. See Screen “Colors” on page 110 for exact color of residuals.

Continuous. Changes the horizontal scale labels to 5 equally spaced quantitative values in standardized X-score units. For the joint regression plot, this option also changes the vertical scale labels from the observed Y values/categories to 5 equally spaced quantitative values.

Ref. Lines. When this option is selected, horizontal and vertical dashed lines are displayed corresponding to the X and Y reference scores x_0 and y_0 as determined by the current weights set under Model, Category Weights.

Plot Control for Effects Plots

Figure 6-23
Plot Control for Effects Plots



The following Plot Control options are available for Effects plots:

Effect Lines. Estimated effects lines are included in the plot when this option is checked (default).

Observed Points. Symbols representing observed odds ratios (or log-odds ratios) are included in the plot when this option is checked (default). The diameter of the symbol is proportional to the number of observations in the associated cell.

Predicted Y (Y-view only). Arrows pointing downward onto the horizontal line representing the predicted outcome for each predictor category are included in the plot when this option is checked (default).

Continuous. Relabels the horizontal scale from the actual categories to 5 equally spaced quantitative values. For additional labeling options, see “Label Display” on page 131.

Ref. Lines. When this option is selected (default), dashed lines will appear for the selected contrasts set under Model, Category Weights.

To Set Control back to Defaults

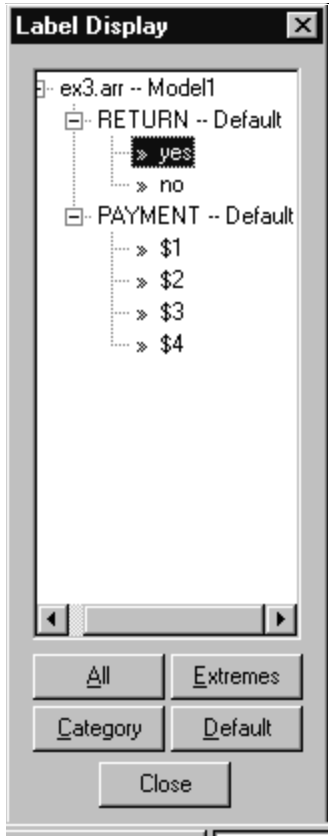
Click on Default to reset all the options to the original defaults.

Label Display

Selecting Label Display brings up the Label Display dialog box that can be used to control the number of labels to plot along the axes to improve readability. The dialog box contains the name of each axis (a variable name or "X-profile") and labels for each associated category. Labels currently displayed in the plot are preceded by the >> symbol.

Category markers together with their associated labels are displayed along the horizontal and vertical axes of the plots. By default, when more than eight categories exist, only five labels are displayed - - two labels associated with the minimum and maximum category scores, the label associated with the median score, and the two labels associated with the category scores at the 1^{st} and 3^{rd} quartile.

Figure 6-24
Label Display dialog box



To change the number of labels displayed in the plot, from the Label Display box select either the horizontal or vertical axis by clicking the axis name (or a label listed under the name) and select from the following activated options:

All. Selecting All causes one marker and associated label to be displayed for every observed category (value) of the variable. Several labels in the same location may be distinguished by clicking in the location to make one of the labels visible and using the tab key to display the next label.

Extremes. Selecting Extremes causes only the two labels associated with the minimum and maximum category scores to be displayed.

Category. Highlight any category label and select Category to alternate between displaying and removing that label from the plot. Highlighting a label preceded by the >> symbol and selecting category causes that label (currently displayed in the plot) to be removed. Highlighting a label absent the >> symbol and selecting Category causes that label to appear in the plot.

Default. Selecting Default restores the labels in the plot for the selected axis to their original default levels.

Plot Types

Output from any model may be examined from different perspectives using different plot types. Following are samples of the six possible plot types that may be constructed from sample dataset #3 for results from Model A in Chapter 9. GOLDMineR displays a message in the lower right corner of the plot that identifies the plot type that is displayed. For details of the components of these plots, see Chapter 3. For information on options, see "Control" on p. 128. Note: A table may be opened and customized to provide supplementary information for any particular plot.

How to Select Plot Types

E If no Plot windows are open or if you wish to open a new Plot window, from the menus choose:

Window

New Plot

E otherwise within an active Plot window from the menus choose:

Plot

E select the plot type.

- For Partial plots, the Select Effect dialog box will appear with a list of the predictor variables preceded by a symbol in parenthesis regarding their current category weight setting (See page 122 for the definition of the symbols). Select the predictor effect by highlighting the name of the predictor and either double clicking or clicking on OK. Selection of a predictor will produce a plot for that predictor where all other predictors are held constant at their current category weight setting.

Figure 6-25
Select Effect dialog box for Partial Plots

Click on the +- symbol to list the category labels for each predictor.

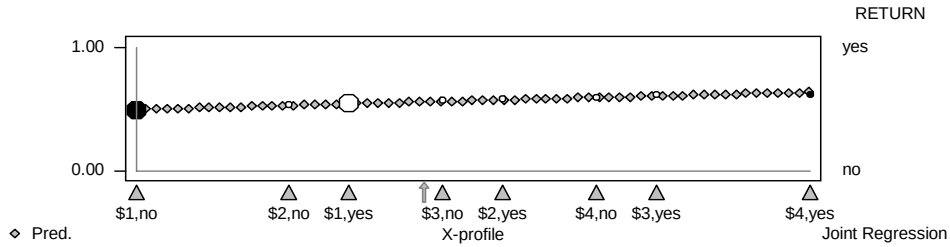
Following are the six plot types (two types of Regression plots and four types of Effects plots):

Regression Plots

In the Regression plot the predicted value of the dependent variable is plotted as a function of either 1) the joint (composite) X variable (**joint** view), where X = (X_1, X_2, …, X_M) or 2) a selected predictor variable (**partial** view), X_m, controlling for set levels (the X-reference levels x_{m0}) of the other predictors. The U-score is used to quantify the joint X variable on the horizontal axis of the joint Regression plot. For the definition of U, see "A.5 Model Identification Issues" on page 218 of the Appendix. The Regression plot is generally more useful with quantitative rather than qualitative predictors. In the case of a model with only one predictor variable, except for the range and labeling of the vertical axis, the partial and joint plots are identical.

The Joint Regression Plot

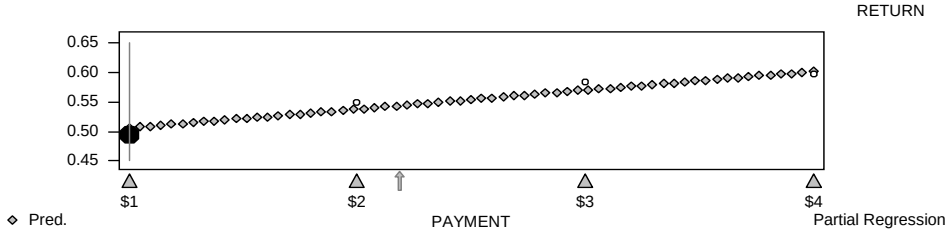
Figure 6-26
Joint Regression plot for all predictors



Standardized U-scores and predicted Y-values may be viewed by clicking on plot symbols. The predicted Y-value may be viewed in unstandardized or standardized units depending upon whether Unstandardized is selected in the View menu.

The Partial Regression Plot

Figure 6-27
Partial Regression plot for PAYMENT given CALL = No



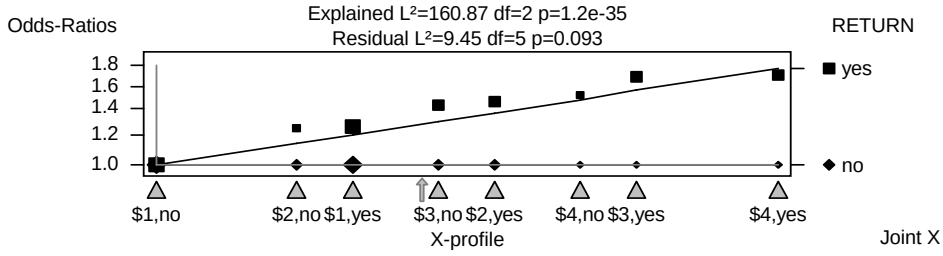
Effects Plots

Partial effects and joint effects estimated by a model are odds ratios. When displayed in logarithmic units, the log-odds ratios may be plotted as an exact linear function of X (the X-view) or as a linear function of Y (the Y-view). In the X-view, the partial or joint X variable is represented on the horizontal axis while in the Y-view, Y is represented on the horizontal axis. In the joint effects plot, selected X-profiles (i.e., selected categories of $X = (x_1, x_2, \ldots, x_M)$ that are observed in the sample) appear in

the plot. In the partial effects plot, selected individual categories of the selected predictor appear in the plot controlling for set levels (the X-reference levels) of other predictors. In the case of a model with only one predictor variable, the partial and joint plots are identical. The particular categories that are displayed are controlled by the Label Display option in the Plot menu.

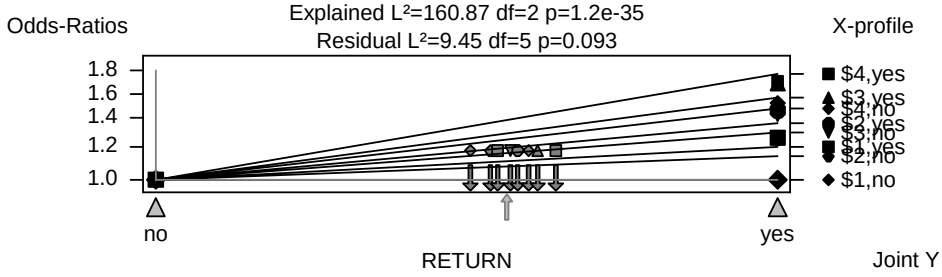
The Joint X-view Effects Plot

Figure 6-28
Joint X plot



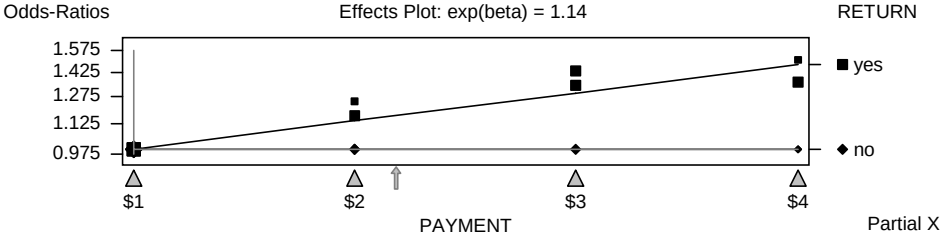
The Joint Y-view Effects Plot

Figure 6-29
Joint Y plot



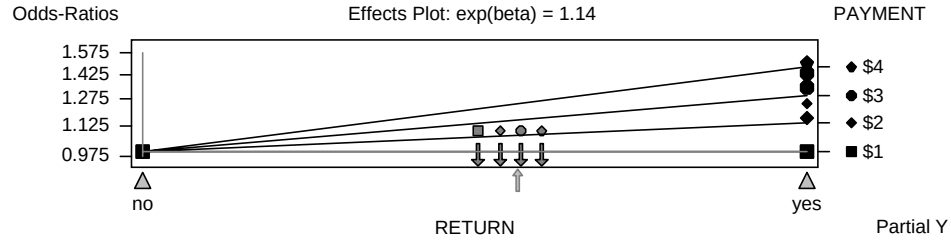
The Partial X-view Effects Plot

Figure 6-30
Partial X plot for PAYMENT given CALL = No



The Partial Y-view Effects Plot

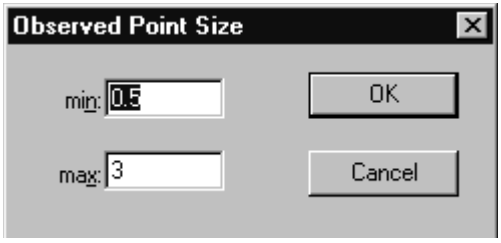
Figure 6-31
Partial Y plot for PAYMENT given CALL = No



Observed (Point Size)

Symbols representing observed odds-ratios in Effects plots and observed average Y-values in Regression plots are drawn in the Plot window with diameter that varies according to the cell frequency count associated with that symbol. This dialog box allows users to set the minimum and maximum diameter size for the symbols associated with the minimum and maximum cell frequency count. The unit of measurement is the size of the 'x' character of the current font.

Figure 6-32
Observed Point Size dialog box



Save…

The graphical plot view may be saved in Windows Metafile Format (.wmf) with the Plot Save command. This saved file is created for editing purposes using software external to GOLDMineR and cannot be reopened using GOLDMineR.

TABLE MENU

The **Table Menu** is only available when a GOLDMineR table is the active window. After the New Tables option has been used to open at least one table from the Windows menu, the **Table Menu** may be used to select the table type and the cell contents. For the active Table window, the Table Menu identifies the various table options that are in effect for that table (each option in effect is marked with a checkmark), and may also be used to open or close the Table control dialog box, order the table rows or columns according to the predicted value of Y, or produce specialized charts. For details on the contents of the Table window, see Chapter 3.

Control

Selecting Control opens the Table Control dialog box. This box may also be opened (and closed) by clicking on the right mouse button from within an active Table window. The Table Control allows you to choose the cell contents of the table. Multiple cell options may be selected at once (for example, a table can contain both observed counts and expected odds ratios).

Figure 6-33
Table Control dialog box

The following options are available using Table Control:

Observed vs. Expected

Observed (default). Cell contents are based on the sample observations (observed frequency counts).

Expected. Cell contents are based on expected frequency counts estimated by the model.

Standard vs. Logarithmic

This option controls the units to be displayed in the tables. (The units may also be set in the View menu.)

Standard. (default) Select Standard to have the current table entries displayed in standard units.

Logarithmic. Select Logarithmic to have the current table entries displayed in logarithmic units.

Cell Contents

Frequencies (default). In standard units, cell (i,j) of the table contains the observed frequency counts, f_{ii}, for that cell or the estimated expected count, $\hat{F}_{ij}$. In logarithmic units, $\ln(f_{ij})$ and $\ln(\hat{F}_{ij})$ are displayed. For cells in which there are no sample observations, display of the observed frequency count will show a 0 (or in partial tables where there are no sample observations for an entire X-profile, a '.'). When logarithmic units are displayed for cells in which $f_{ij} = 0$ or '.', since ln(0) is undefined, '.' is displayed as the cell entry.

Proportions/Probabilities. Cell (i,j) may contain the observed conditional proportion $p_{j.i} = f_{ij} \Big/ \sum_{i=1}^{I} f_{ij}$ and/or the estimated conditional probability $P_{j.i} = \hat{F}_{ij} \Big/ \sum_{i=1}^{I} \hat{F}_{ij}$ depending upon whether *observed proportions* or *expected probabilities* is checked. Summing the proportions/probabilities over any given row i (in Y-view) or column i (in X-view) will total 100%.

Odds. Cell (i,j) contains the odds in favor of the jth outcome category relative to the Y-reference identified by the current E-weights. For each predictor category, the odds is calculated as the ratio of the expected (or observed) frequency count for the jth outcome category to the corresponding reference.

Ratios. Cell (i,j) contains the odds ratio effect defined by the reference point (x_0, y_0) identified by the current E-weights and C-weights. The *expected* odds ratios define the effects lines in Effects plots (in logarithmic or standard units). The *observed* odds ratios (in logarithmic or standard units) correspond to the points that appear as symbols in Effects plots.

Adjusted Residuals

When selected, an additional column is appended to the table containing adjusted residuals associated with each given X-profile (each row in Y-View). Each adjusted residual assesses the difference between the observed average Y-score and the predicted Y-value for that X-profile. Adjusted residuals have a standard normal distribution under the assumption that the fitted model is true.

Ordered

When selected (the default), rows and columns are ordered based on the associated X-score and Y-score. Such ordering serves to align the table with the Effects plot axes. When not selected, rows (columns) associated with a nonnumeric variable are ordered alphabetically and rows (columns) associated with the joint variable are sorted by each component variable. For related options see “Reverse Vertical” on page 112, “Reverse Horizontal” on page 113 and “Synch Views” on page 117.

Table Types

Following are samples of the four possible table types (Joint X, Joint Y, Partial X or Partial Y) constructed for results from Model A in Chapter 9. GOLDMineR displays a message in the upper left corner of the table that identifies the table type. The type for the active table is marked by a checkmark in the menu. For details of the components of a table, see Chapter 3. For information on options, see “Control” on p. 139.

In the X-view, the partial or joint X variable makes up the rows of the table and the Y variable the columns, while in the Y-view, the Y variable comprises the rows of the table and the partial or joint X variable the columns. In joint tables each X-profile (i.e., each joint category of $X = (x_1, x_2, \ldots, x_M)$ that is observed in the sample) makes up the rows or columns. In partial tables (which are a subset of the joint tables), each individual category of the selected predictor comprises the rows or columns controlling for set levels (the X-reference levels) of other predictors. In the case of a model with only one predictor variable, the partial and joint tables are identical.

How to Select Table Types

E If no table windows are open or if you wish to open a new Table window, from the menus choose:

Window

New Table

E otherwise within an active Table window from the menus choose:

Table

E select the table type.

- For partial tables, the Select Effect dialog box will appear with a list of the predictor variables preceded by a symbol in parenthesis regarding their current category weight setting (See page 122 for the definition of the symbols). Select the predictor effect by highlighting the name of the predictor and either double clicking or clicking on OK. Selection of a predictor will produce a table for that predictor where all other predictors are held constant at their current category weight setting.

Figure 6-34
Select Effect dialog box for Partial Tables

Click on the +- symbol to list the category labels for each predictor.

The following four table types are available to choose from:

Joint X

The expected odds ratios (shown in Figure 6-35) are used to create the effects lines in the Joint X plot (see Figure 6-28).

Figure 6-35
Joint X table for all predictors – Expected Odds Ratios

(Joint X)			X-profile									
			$1,no	$2,no	$1,yes	$3,no	$2,yes	$4,no	$3,yes	$4,yes		
RETURN	Y-scores	weights	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	X-ref	(cell)
yes	1.00	0.00	1.00	1.14	1.20	1.30	1.36	1.48	1.55	1.77	1	R
no	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1	R
		Y-ref	1	1	1	1	1	1	1	1	1	R
	average	score	0.51	0.54	0.55	0.57	0.58	0.60	0.61	0.64	21394	F

Joint Y

The expected odds ratios (shown in Figure 6-36) are used to create the effects lines in the Joint Y plot (see Figure 6-29).

Figure 6-36
Joint Y table for all predictors – Expected Odds Ratios

(Joint Y)			RETURN			
	average		no	yes		
X-profile	score	weights	1.00	0.00	Y-ref	(cell)
$4,yes	0.64	0.00	1.00	1.77	1	R
$3,yes	0.61	0.00	1.00	1.55	1	R
$4,no	0.60	0.00	1.00	1.48	1	R
$2,yes	0.58	0.00	1.00	1.36	1	R
$3,no	0.57	0.00	1.00	1.30	1	R
$1,yes	0.55	0.00	1.00	1.20	1	R
$2,no	0.54	0.00	1.00	1.14	1	R
$1,no	0.51	1.00	1.00	1.00	1	R
		X-ref	1	1	1	R
		Y-scores	0.00	1.00	21394	

Partial X

The expected odds ratios (shown Figure 6-37) are used to create the effects lines in the Partial X plot (see Figure 6-30).

Figure 6-37
Partial X table for PAYMENT given CALL = 'No' – Expected Odds Ratios

(Partial X)			PAYMENT					
			$1	$2	$3	$4		
RETURN	Y-scores	weights	1.00	0.00	0.00	0.00	X-ref	(cell)
yes	1.00	0.00	1.00	1.14	1.30	1.48	1	R
no	0.00	1.00	1.00	1.00	1.00	1.00	1	R
		Y-ref	1	1	1	1	1	R
	average	score	0.51	0.54	0.57	0.60	10653	.

Partial Y

The expected odds ratios (shown in Figure 6-38) are used to create the effects lines in the Partial Y plot (see Figure 6-31).

Figure 6-38
Partial Y table for PAYMENT given CALL = 'No' – Expected Odds Ratios

(Partial Y)			RETURN			
	average		no	yes		
PAYMENT	score	weights	1.00	0.00	Y-ref	(cell)
$4	0.60	0.00	1.00	1.48	1	R
$3	0.57	0.00	1.00	1.30	1	R
$2	0.54	0.00	1.00	1.14	1	R
$1	0.51	1.00	1.00	1.00	1	R
		X-ref	1	1	1	R
		Y-scores	0.00	1.00	10653	

Specialized Charts

Selecting Specialized Charts from the Table menu replaces the active table with a specialized chart based on your current model. There are two types of specialized charts, profile or quantile. There are also two ways to format each chart type, either using the average Y-score or a selected category of the dependent variable.

Obtaining a Specialized Chart

E Within an open Table window, from the menus choose:

Table

Specialized Charts

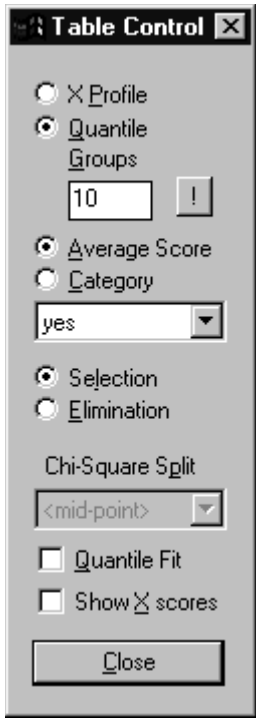
E Select Options from the Table Control dialog box for Specialized Charts.

E Click on Close.

Control

If a chart is the active window, either right click from within the window or select Table, Control to bring up the control dialog box containing chart options.

Figure 6-39
Table Control dialog box for Specialized Charts



You may select from the following options:

Type of Chart

Choose one of the following alternatives to produce either a profile or quantile chart.

X Profile. An X-profile chart contains a row for each category of the joint X-variable (i.e., each X-profile), and rank orders all of these profiles from high to low (or low to high) based on the predicted Y-score (average score format type) or the percentage in a selected category of the dependent variable (category format type). The profile chart can become quite large if many predictor variables or predictors with many categories (or both) are included in the model.

Figure 6-40
Profile Chart based on Average Score Type

Selection

Id	n	Predicted Score	Observed Score	Cumulative ... n	% of n	Predicted Score	Observed Score
$4,yes	2085	0.64	0.63	2085	9.75	0.64	0.63
$3,yes	2149	0.61	0.62	4234	19.79	0.63	0.62
$4,no	2087	0.60	0.60	6321	29.55	0.62	0.62
$2,yes	2146	0.58	0.59	8467	39.58	0.61	0.61
$3,no	2159	0.57	0.58	10626	49.67	0.60	0.60
$1,yes	4361	0.55	0.55	14987	70.05	0.59	0.59
$2,no	2098	0.54	0.55	17085	79.86	0.58	0.58
$1,no	4309	0.51	0.50	21394	100.00	0.57	0.57

The Profile chart based on the predicted Y (average score format type) (see Figure 6-40) provides an identifying label for each X-profile. Individual and cumulative information is provided for the number of sample observations having the X-profile, the predicted Y and the corresponding observed average score. For further details on these components, see Chapter 3.

Figure 6-41
Profile Chart based on Category Format

Selection

		Predicted		Observed		Cumulative ...		Predicted		Observed			
Id	n	resps	%	resps	%	n	% of n	resps	% resps	% of resps	resps	% resps	% of resps
$4,yes	2085	1343.96	64.46	1306	62.64	2085	9.75	1343.96	64.46	11.09	1306	62.64	10.78
$3,yes	2149	1319.99	61.42	1340	62.35	4234	19.79	2663.95	62.92	21.99	2646	62.49	21.84
$4,no	2087	1256.78	60.22	1248	59.80	6321	29.55	3920.73	62.03	32.36	3894	61.60	32.14
$2,yes	2146	1251.05	58.30	1265	58.95	8467	39.58	5171.78	61.08	42.68	5159	60.93	42.58
$3,no	2159	1232.00	57.06	1262	58.45	10626	49.67	6403.78	60.27	52.85	6421	60.43	52.99
$1,yes	4361	2403.00	55.10	2407	55.19	14987	70.05	8806.78	58.76	72.68	8828	58.90	72.86
$2,no	2098	1129.76	53.85	1156	55.10	17085	79.86	9936.54	58.16	82.00	9984	58.44	82.40
$1,no	4309	2180.46	50.60	2133	49.50	21394	100.00	12117.00	56.64	100.00	12117	56.64	100.00

The profile chart based on the category format (see Figure 6-41) ranks the X-profiles, i = 1, 2, …, I by the predicted percentage "Response", where "Response" corresponds to the selected outcome category j. In addition to providing the ID and number of sample observations, this chart also contains the predicted number of responders ("Predicted resps"), the predicted response rate ('%') and the corresponding observed percentage that respond in that outcome category. Similar to the average score format, the rightmost columns provide the corresponding cumulative information. Note: when printing a category formatted chart, select 'landscape'.

Quantile. A Quantile chart contains percentile rankings of joint X-variable groupings by predicted Y-score.

Figure 6-42
Quantile Chart

Selection

Id	n	Predicted Score	Observed Score	Cumulative ... n	% of n	Predicted Score	Observed Score
1	2139.40	0.64	0.63	2139.40	10.00	0.64	0.63
2	2139.40	0.61	0.62	4278.80	20.00	0.63	0.62
3	2139.40	0.60	0.60	6418.20	30.00	0.62	0.62
4	2139.40	0.58	0.59	8557.60	40.00	0.61	0.61
5	2139.40	0.57	0.58	10697.00	50.00	0.60	0.60
6	2139.40	0.55	0.55	12836.40	60.00	0.59	0.60
7	2139.40	0.55	0.55	14975.80	70.00	0.59	0.59
8	2139.40	0.54	0.55	17115.20	80.00	0.58	0.58
9	2139.40	0.51	0.50	19254.60	90.00	0.57	0.57
10	2139.40	0.51	0.50	21394.00	100.00	0.57	0.57

The Quantile chart (see Figure 6-42) contains an Id that specifies the number of quantiles selected (the figure above represents the default decile chart which contains 10 quantiles). Individual and cumulative information is provided for the segment size, percentage of total sample, average expected score and average observed score. For further details on these components, see Chapter 3.

Groups. By default, segments are grouped into 10 equal intervals called deciles. You may specify an alternative number of groups by inputting a number of quantile groups between 2 and the total sample size. For example, inputting '4' will produce a chart with 4 quartiles (25% intervals). To change this setting, highlight the number currently listed, type in a new number, and click on !.

Format Type

Two different report formats for summarizing the model predictions and comparing them with the corresponding observed quantities are available. The first is based on the predicted Y-values and compares these values with the corresponding observed average scores (obtained by selecting "Average Score" in the Table Control). The second focuses on a designated outcome category. Click "Category" and select the designated outcome category to obtain the second format.

Choose one of the two format types to base your chart rankings on:

Average Score (default). Based on the predicted Y-value and summarizes the observed average Y-scores for comparison.

Category. Focuses on the percentage responding in a designated outcome category.

Order

Choose one of the following alternatives from the Order group:

Selection.(default) A selection report ranks groupings from high to low. The predicted Y-value (or predicted response rate) is sorted in descending order, and the cumulative statistics reflect the successive addition of each new grouping.

Elimination. An elimination report ranks groupings from low to high. The predicted Y-value (or predicted response rate) is sorted in ascending order, and the cumulative statistics reflect the successive elimination of groupings.

Other specialized chart options include:

Chi-square Split (experimental). Useful for J>2 outcome categories only. Select option A or B to dichotomize the dependent variable for the computation of the experimental quantile fit statistic. The dichotomization is only used for the computation of this statistic and does not affect any other computation.

- Option A: Select "midpoint" to dichotomize the dependent variable so that category 1 and 2 have exactly the same number of observations (Observations in one or more original outcome categories may be proportionately allocated between the two categories. For example, if J=3, the number of observations in outcomes 1 and 3 are equal, and 5 observations are in outcome 2, these 5 observations will be divided equally, 2.5 in new category 1 and 2.5 in new category 2.)
- Option B: Select outcome category j' as the cutpoint. This selection combines adjacent outcome categories so that new category 1 contains categories 1,2,…,j' and new category 2 contains categories j'+1,j'+2,…,J.

Quantile Fit. (experimental). When this option is selected, GOLDMineR will compute the quantile fit based on the number of quantiles selected under **Quantiles**. This experimental fit statistic may be used as an alternative to the Residual L^2 and Residual X^2 model fit statistics when there is sparse data (such as with continuous variables). If J>2, the dependent variable will be dichotomized prior to the computation. The output appears above the quantile chart. For technical information, see “A.15: Assessing Model Fit when Analyzing Sparse Data” on page 244.

Show X-scores. An additional column appears in the specialized charts which contains the standardized X-scores associated with each X-profile or each quantile group.

CONTENTS MENU

The Contents Menu is only available when the Summary window is the active window. This menu can also be accessed by right clicking from within the Summary window (see Chapter 2 for more details). The following options are available:

Open New Data. Allows the user to open a new data file to begin analysis on. The Open dialog box will appear (this option works the same as File, Open).

Define New Model. Allows the user to define a new model for the current data file. The Define dialog box will appear (this option works the same as Model, Define).

Note: There is one Define dialog box per data file.

- If you select Define New Model and the Define dialog box for the current data file is not already open, the Define dialog box will contain the settings for the current model.
- If you select Define New Model and the Define dialog box for the current data file is already open, the Define dialog box will contain the settings for the most recently estimated model. In this case, to view the settings for a particular model you must first close the open Define dialog box.

Activate View. In the Summary window, highlight a type of window (Stat, Plot or Table) and select Activate View to make that window the active window.

Statistics, Plot, Table. In the Summary window, highlight either a model name or type of window and select Statistics, Plot or Table to open a new Statistics, Plot or Table window for the current model. Similar to the Window, New Statistics, New Plot and New Table commands.

SEARCH MENU

About Search

- The Search feature is a heuristic procedure primarily useful in automating variable selection prior to performing a model estimation in applications where many predictors exist.
- The Search feature is most useful with datasets containing many predictor variables as a fast way of identifying a smaller number of significant predictors. For example, for a dataset that contains 100 candidate predictor variables, the

search feature could be used first to automatically identify 8 of the most significant predictors to include in the model.

- The Search feature is activated by clicking on Search in the Model Define dialog box (or selecting Model, Search when the Model Define dialog box is the current window).
- Search performs a stepwise inclusion process which evaluates each predictor variable among a selected set, one at a time, regarding its unique contribution to the model fit chi-square.
- During Step 1, Search ranks each predictor from most to least significant based on the p-value associated with the model L^2 statistic for the model that contains only that predictor.
- During subsequent steps, Search ranks each predictor from most to least significant based on the p-value associated with the likelihood ratio difference statistic $L^2(Y)$ for that predictor. (Any predictors treated as Margin Only are ranked at the bottom of the list.)
- Optionally, users can examine the ranked set and select any predictor (not necessarily the most significant) to enter into the model at that step.
- $L^2(Y)$ for the m^{th} predictor represents the increase in Residual L^2 resulting from imposing the restriction $\beta_m = 0$ (i.e., treating the m^{th} predictor as **Margin Only**) in the model that includes the predictors that have already entered the model plus the m^{th} predictor.
- Cautionary Notes: Predictors are not removable from the model once they have been entered. The category scores for a Free dependent variable are reestimated at each step of the Search procedure. There is no guarantee that the final set of predictors is optimal in any sense.

Using Search

E Click on Search in the Model Define dialog box. The Search dialog box appears with an upper and lower box. The upper box initially contains a list of all predictors to be considered for possible inclusion in the model.

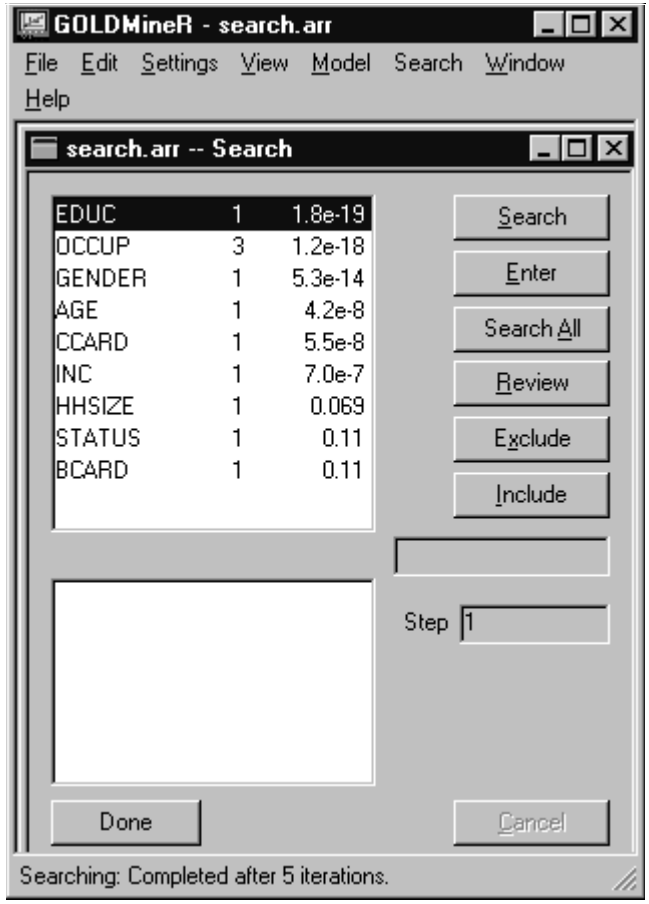
E Select Search to rank the predictors from most to least significant by p-value with associated degrees of freedom.

E Highlight a variable and click on Enter to enter it into the model. Generally, the most significant is selected.

E Repeat the previous two steps until you no longer want to enter any more variables into the model.

E Click on Done.

Figure 6-43
Search dialog box



Search Options

- After any given step, a subset of candidate predictors in the upper box may be excluded from future consideration using the Exclude option. Previously excluded predictors can be included in future steps using the Include option.
- After any step of the search process, output from the current model may be examined by using the Review option and then the Search process may be continued (or ended by clicking on Done).
- Search may be performed in automatic mode using the Search All option. In this case the most significant predictor is entered at each step automatically and the search process continues until no candidate predictors are significant at the .05 level or the user manually terminates the procedure by clicking on Cancel.

Following are the options listed in detail:

Search. At each step, all eligible predictors in the upper box are rank ordered by significance (p-value, listed to the right of degrees of freedom). The most significant predictor, (i.e., the one with the lowest p-value) is listed first and highlighted.

Enter. At the conclusion of each step of the search process, click on Enter to select the most significant predictor (which is highlighted) for inclusion in the model and the predictor will move to the lower box. Optionally, any other eligible predictor displayed in the upper box may be selected and entered instead. Only one variable can be entered into the model at each step. If you make a mistake and enter the wrong predictor, highlight the correct predictor and click on Enter again. The correct predictor will change places with the other predictor. To continue the selection process, click on Search again or click on Search All.

Search All. Search All performs the search process, enters the most significant predictor into the model, and continues the search procedure until none of the p-values associated with the eligible predictors in the upper box is statistically significant at the .05 level. (Any predictors treated as Margin Only are not included in the model.)

Review. Used to view GOLDMineR output associated with the current model at any particular step and causes the current model to be listed in the Estimated Model dialog box. To return to the Search box after a Review, select 'filename<search> ' from the listing of active windows at the bottom of the Windows Menu. During a Review of the Statistics output, a '.' (which indicates not available) is displayed below the column titled 'L^2(Y)' because to expedite the computations the L^2(Y) statistics associated with the current model are not calculated by the Search procedure except for the predictor

most recently entered into the model. (The WALD statistic may be examined in the absence of $L^2(Y)$.)

Exclude. To exclude a predictor in the upper box from being considered for inclusion into the model. In the case that many predictors are not significant after the first step, they may be excluded from future steps to speed up the search process. (However, beware that in some cases a variable that is not significant at one step may become significant at a later step - - See "Tutorial #2" on page 62.)

Include. To include a predictor that has been previously excluded in the upper box.

Step. The step number S indicates that the Search is currently evaluating a model containing S predictors (initially, S = 0).

Cancel. To terminate the current step of the model estimation procedure once it has begun, select Cancel. Cancel is available from the time Search (or Search All) is begun until Search (or Search All) is completed.

Done. When you have completed your Search (or no eligible predictors remain) select Done (instead of Review) to close the Search dialog box and view the output.

WINDOW MENU

The Window menu contains three sections:

- The first section may be used to open the summary window or a new output window.
- The second section may be used to alter the arrangement of the current windows.
- The third section provides a list of all current Statistics, Plot, Table and Summary windows that are open and marks the active window with a checkmark.

After a model is estimated, a Statistics window and a Plot window are opened automatically. In order to open a Table window the Window menu must be used. If a Table window is opened, the table type for this window will be consistent with the category ordering and type of the most recently opened Plot window that corresponds to the current model.

Each current window is listed in the Windows menu and is identified by 1) the dataset name, 2) the model number, and 3) the type of window (Statistics, Plot, or Table) and a sequence number for windows of that type opened (1,2,3,…) for the current model. Once opened, each window can be customized in different ways and

hence multiple Plot windows or multiple Table windows may be open at any given time. See Chapter 2 for information about keeping track of open windows.

Summary Window

E To open the Summary window, from the menus choose:

Window

Summary Window

The Summary window provides a hierarchical listing of all windows that have been opened and can be used to move from one window to another. See Chapter 2 for more detail.

New Plot

E To open a new Plot window, from the menus choose:

Window

New Plot

E If multiple models have been estimated, The Estimated Models dialog box will appear with a list of the current models run.

E Highlight the model you wish to open a new Plot window for and select Ok.

New Statistics

E To open a new Statistics window, from the menus choose:

Window

New Statistics

E If multiple models have been estimated, The Estimated Models dialog box will appear with a list of the current models run.

E Highlight the model you wish to open a new Statistics window for and select Ok.

New Table

E To open a Table window, from the menus choose:

Window

New Table

E If multiple models have been estimated, The Estimated Models dialog box will appear with a list of the current models run.

E Highlight the model you wish to open a new Table window for and select Ok.

Cascade

E To cascade the display of open windows, from the menus choose:

Window

Cascade

The active window appears in front of the others. To activate another window and bring it forward, click anywhere on the visible portion.

Tile Vertical

E To tile the display of open windows vertically (two windows displayed side by side), from the menus choose:

Window

Tile Vertical

This rearranges and resizes the open windows with no overlap so that you can see all of them at once. The active window is placed in the upper left hand corner. To activate another tiled window, click anywhere within it. If you activate a window and then click on its Maximize button, it expands to fill the application window. Clicking the Restore button sends the window back to its place in the tiled set. You can also resize tiled windows by clicking and dragging the sides or corners.

Tile Horizontal

E To tile the display of open windows horizontally (two windows displayed one above the other), from the menus choose:

Window

Tile Horizontal

This rearranges and resizes the open windows with no overlap so that you can see all of them at once. The active window is placed in the upper left hand corner. To activate another tiled window, click anywhere within it. If you activate a window and then click on its Maximize button, it expands to fill the application window. Clicking the Restore button sends the window back to its place in the tiled set. You can also resize tiled windows by clicking and dragging the sides or corners.

Arrange Icons

If you have iconized any windows, this option arranges the icons at the bottom of the GOLDMineR main window.

Selecting a Window from the Menu

At the bottom of the Window menu is a list of open windows, whether iconized or not. The active window has a checkmark beside it. You can activate and display one of the other windows by selecting it from the menu.

7

Example 1 Assessing the Effects of a New Treatment for Rheumatoid Arthritis

Table 7-1 contains clinical trial data on a new drug treatment for persons with rheumatoid arthritis (included in the file 'ex1.*'). As is often the case in biomedical outcomes, the dependent variable, a rating of improvement, is measured on an ordered categorical scale. The outcome categories for Y: IMPROVE – *None*, *Some* or *Marked*, are coded 0, 1 and 2 in the data file. The predictor variables are X_1:AGE, X_2:TREAT and X_3:GENDER.

IMPROVE	AGE	TREAT	GENDER	IMPROVE	AGE	TREAT	GENDER	IMPROVE	AGE	TREAT	GENDER
0	23	1	F	0	52	0	M	1	62	1	F
0	23	0	F	0	53	0	M	2	62	1	F
1	27	1	M	0	53	0	F	0	62	0	M
0	29	1	M	0	54	0	F	0	62	0	M
0	30	1	M	0	54	0	F	0	63	1	M
0	30	0	F	2	54	0	F	0	63	1	M
0	30	0	F	2	55	1	F	2	63	0	M
1	31	0	F	2	55	1	F	1	63	0	F
2	32	1	M	2	55	0	F	0	64	1	M
0	32	1	F	2	56	1	F	1	64	1	M
0	32	0	F	2	57	1	F	0	64	0	F
2	33	0	F	2	57	1	F	2	65	0	F
1	37	1	F	2	57	1	F	2	66	1	F
0	37	0	M	0	57	0	F	0	66	0	F
0	37	0	F	1	57	0	F	0	66	0	F

Table 7-1
Rheumatoid Arthritis Data (Source: Koch, G. and Edwards, S. (1988), chapter 9 in K. Pearce, ed., Biopharmaceutical Statistics for Drug Development, Marcel Dekker, Inc., p. 406-407. Reprinted and distributed by permission of publisher.

IMPROVE	AGE	TREAT	GENDER	IMPROVE	AGE	TREAT	GENDER	IMPROVE	AGE	TREAT	GENDER
0	23	1	F	0	52	0	M	1	62	1	F
0	41	1	F	2	58	1	M	1	66	0	F
2	41	1	F	0	58	1	F	2	67	1	F
0	44	0	M	1	58	0	F	1	68	1	F
0	44	0	F	0	59	1	M	2	68	1	F
0	45	0	F	2	59	1	M	1	68	0	F
2	46	1	M	2	59	1	F	0	69	1	M
0	46	0	F	2	59	1	F	0	69	1	F
0	48	1	F	0	59	0	M	1	69	1	F
2	48	1	F	0	59	0	M	2	70	1	M
0	48	0	F	1	59	0	F	1	70	1	F
0	49	0	F	2	59	0	F	2	74	0	F
0	50	0	M	2	60	1	F				
0	51	0	M	2	61	1	F				
0	51	0	F	0	61	0	F				

Table 7-1
Rheumatoid Arthritis Data (Source: Koch, G. and Edwards, S. (1988), chapter 9 in K. Pearce, ed., Biopharmaceutical Statistics for Drug Development, Marcel Dekker, Inc., p. 406-407. Reprinted and distributed by permission of publisher.

Our goal is to assess the effect of the *test drug* (TREAT = 1) vs. the *placebo* (TREAT = 0) under each of three models that differ with respect to the coding used to scale the dependent variable:

Model A: Dichotomous Scale - - IMPROVE = *None* (0), *Some* (1), *Marked* (1). Equivalent to the simple logistic regression model that defines improvement as a dichotomous variable (*Some* or *Marked* improvement vs. *None*). Model A is specified in GOLDMineR by assigning exactly two scores to the outcome categories - - specified here by maintaining the original scores of 0 and 1 for *None* and *Some* improvement respectively and recoding the score for *Marked* improvement from 2 to 1.

Model B: An Unrestricted Ordinal Scale - - IMPROVE = *None* (0), *Some* (y_2^*), *Marked* (1). The ordinal logit model where no assumption is made on the relative spacing of the improvement categories. Since any scale is meaningless without 1) a zero point and 2) a unit point, exactly two scores need to be assigned to ordinal outcome categories for the purpose of scale identification. In order to make meaningful comparisons of results between Models A and B, we assign the same scores to the outer categories as used in model A (0 and 1), set the score associated with

the middle improvement category, y_2^*, to missing (“.”), and use GOLDMineR's imputation option to estimate the unknown Y-score.

Model C: An Equidistant Scale - - IMPROVE = *None* (0), *Some* (0.5), *Marked* (1). The ordinal logit model which assumes that *Some* improvement is equidistant between *None* and *Marked* is called the “equal adjacent category odds ratio” model and is specified by assigning equidistant scores such as 0, 0.5 and 1 to the outcome categories. (Without loss of generality, for scale identification, scores of 0 and 1 are again assigned to the outer categories. Thus, each of the three models can be distinguished from each other by the Y-score used for the *Some* improvement category.)

Interpretation of Model Results

For each of the three models, Table 7-2 lists the R^2 statistic, which assesses the percentage of the variance of Y explainable by a model, and the results of the test H_0:$R^2 = 0$, as well as the score estimate for y_2^* obtained under Model B. The results indicate that the R^2 statistic is 1) significantly different from zero and 2) similar in magnitude under each model. (The likelihood ratio chi-squared statistic, Explained L^2, tests the null hypothesis H_0:R^2=0 under each model.)

Model B is not as parsimonious as Models A and C because it requires an additional parameter to be estimated, y_2^*. Hence, we would only accept Model B as our final model if it provided a significantly better fit to the data than both Models A and C. The estimate for y_2^* is 0.74, which positions the *Some* improvement category almost ¾ of the way towards *Marked* improvement (see Figure 7-2). However, the 95% confidence interval for y_2^* is [0.26, 1.37], which includes both 0.5 (the score for y_2 assumed under Model C), and 1.0 (the score for y_2 assumed under Model A). Hence, the relative positioning of the *Some* improvement category given under Model B is not significantly different from the positioning estimated under parsimonious Models A and C (see Figure 7-1 and Figure 7-3).

	Y-scores				Test of Null Hypothesis H_0:R^2=0		
	y_1	y_2	y_3	R^2	Explained L^2	d.f.	Significance (p-value)
Model A	0	1	1	0.26	24.4	3	p = 2.1x10-5
Model B	0	0.74*	1	0.27	25.3	4	p = 4.4x10-5
Model C	0	0.50	1	0.26	24.3	3	p = 2.2x10-5

Table 7-2
Model Summary Statistics for Models A, B and C

The treatment effect estimated under each model is displayed in Table 7-3 in units of log-odds ratios (see the column titled β_2) as well as odds ratios ($\exp(\beta_2)$). The interpretation of these estimates may be visualized using the partial Effects plots in Figure 7-1, Figure 7-2, and Figure 7-3, where the treatment effect corresponds to the slope of the effects line.

	log-odds ratio					odds ratio	
	β_2	Standard Error	Wald Chi-square	d.f.	Significance (p-value)	$\exp(\beta_2)$	95% Confidence Interval
Model A	1.76	0.54	10.76	1	.001	5.8	[2.0, 16.6]
Model B	2.03	0.58	12.03	1	.0005	7.6	[2.4, 23.9]
Model C	2.13	0.60	12.42	1	.0004	8.4	[2.6, 27.3]

Table 7-3
Estimates of the Treatment Effect and Related Statistics under Models A, B and C

Figure 7-1
Effects Plot for Treatment under Model A

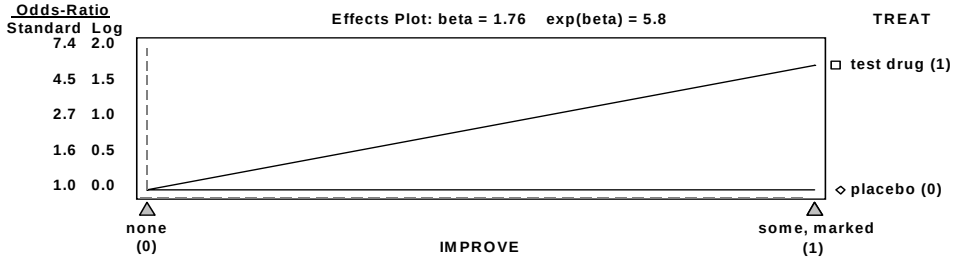


Figure 7-1 shows that patients receiving the test drug are estimated under Model A to be exp($\hat{\beta}_2$) = 5.8 times as likely to improve (some or marked improvement) than patients of the same age and gender who received the placebo. (The slope of the effects line is $\hat{\beta}_2$ = 1.76, the log of the estimated expected odds ratio 5.8).

Figure 7-2
Effects Plot for TREAT under Model B

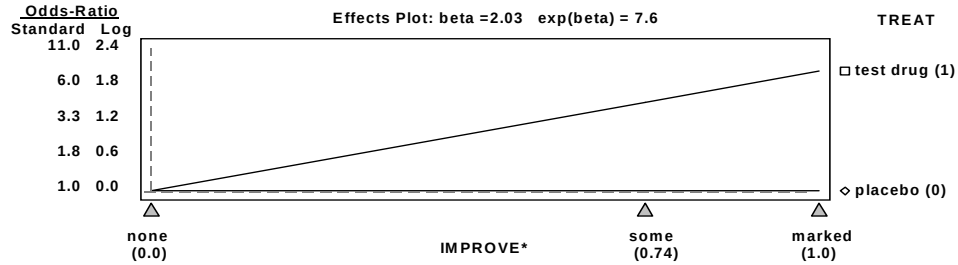


Figure 7-2 shows that patients receiving the test drug are estimated under Model B to be exp($\hat{\beta}_2$) = 7.6 times as likely to achieve a marked improvement (relative to no improvement) and exp(.74 $\hat{\beta}_2$) = 4.5 times as likely to achieve some improvement (relative to no improvement) than patients of the same age and gender who received the placebo.

Figure 7-3
Effects Plot for TREAT under Model C

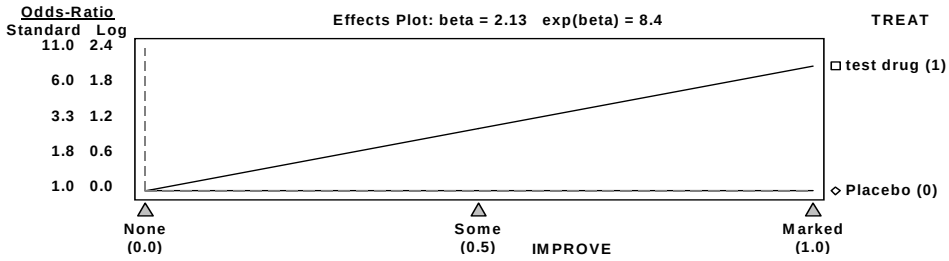


Figure 7-3 shows that patients receiving the test drug are estimated under Model C to be exp($\hat{\beta}_2$) = 8.4 times as likely to achieve a marked improvement (relative to no improvement) and exp(0.5 $\hat{\beta}_2$) = 2.9 times as likely to achieve some improvement (relative to no improvement) than patients of the same age and gender who received the placebo.

Examination of either Table 7-3 or Figure 7-1, Figure 7-2, and Figure 7-3 shows that the effect estimate is positive and statistically significant ($p < .001$) under each of the three models. However, the precise interpretation of the effect depends upon the scale used to measure improvement. In particular, when interpreting the effect estimates obtained under the monotonic regression models it is important to know that β and $\exp(\beta)$ represent the effect of the predictor variable per unit of the dependent variable.

While it is tempting to compare effect estimates obtained from Models A, B and C with each other, the different Y-scores used in the models make comparisons somewhat difficult. In Model A (Figure 7-1), a unit is defined as the distance between none (coded '0') and *some* or *marked* improvement (coded '1'). In Models B and C, a unit is defined as the distance between *none* (0) and *marked* (1) improvement, with *some* improvement being three-quarters of the way towards *marked* (Figure 7-2) or half way (Figure 7-3). Hence, proper interpretation of effect estimates must take into account the Y-scores used to scale the dependent variable. See Chapter 10 for an introduction to the interpretation of odds and odds ratios, log-odds and log-odds ratios (See Rudas, 1998 and Ishii-Kuntz, 1994, for a more thorough discussion of using odds ratios to measure relationships in data).

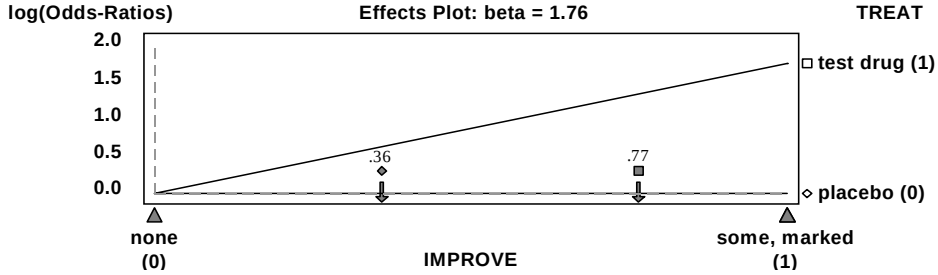
Examination of the 95% confidence intervals in Table 7-3 finds that the odds ratio of 2.0 is the smallest number included in each of the confidence intervals for the three models. Hence, we may conclude that the test drug is likely to at least double the odds of a *marked* improvement regardless of whether improvement is defined according to

Model A, B or C. The use of GOLDMineR output to select one of the three models over the others is discussed later (see "Model Selection" on p. 172).

Monotonic Regression Predictions

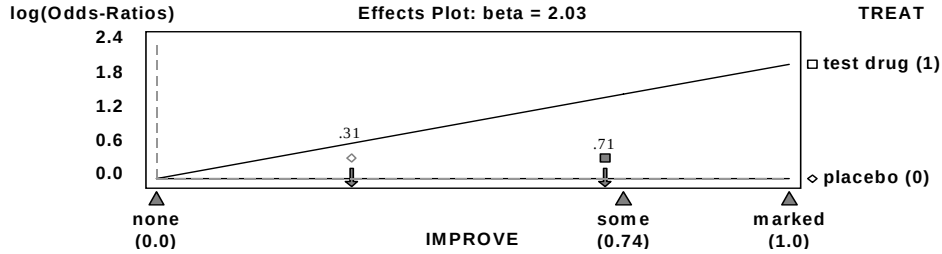
In addition to defining the treatment effect in terms of the log-odds ratio, β, or the odds ratio, $\exp(\beta)$, another way to measure the effect is to compute the difference between the predicted improvement score given the *test drug* vs. the *placebo*. Predictions of the amount of improvement, $\tilde{y}_{.x}$, can be visualized in GOLDMineR's partial Effects plot by examining the distance between the 'arrows' associated with any given covariate-based profile formed by different AGE x GENDER combinations. In Figure 7-4, Figure 7-5, and Figure 7-6, for example, prediction arrows have been included in the graph for 50 year old females. In Figure 7-4 we see that under Model A the estimated effect of the *test drug* for 50 year old females, is an increase of 77%-36% = 41% in the chance of improvement.

Figure 7-4
Partial effects plot for Model A including predictions under the 'test drug' and 'placebo' conditions for GENDER = 'Female', AGE = '50'



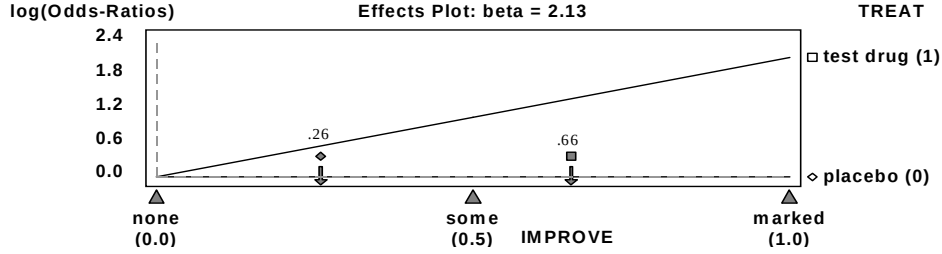
Under Model A, 50 year old females who received the test drug are estimated to have a 77% chance of some or marked improvement compared to 36% of those who received the placebo, reflecting an expected gain due to the test drug of 77% - 36% = 41%.

Figure 7-5
Partial effects plot for Model B including predictions under the 'test drug' and 'placebo' conditions for GENDER = 'Female', AGE = '50'



Under Model B, 50 year old females who received the test drug are estimated to have an improvement score of .71 compared to .31 for those who received the placebo, reflecting an expected gain due to the test drug of .71-.31 =.40 units on the improvement scale.

Figure 7-6
Partial effects plot for Model C including predictions under the 'test drug' and 'placebo' conditions for GENDER = 'Female', AGE = '50'



Under Model C, 50 year old females who received the test drug are estimated to have an improvement score of .66 compared to .26 for those who received the placebo, reflecting an expected gain due to the test drug of .66-.26 =.40 units on the improvement scale.

A test of significance for the distance between the test drug and placebo arrows for any covariate-based profile is equivalent to the significance test of the earlier measure of the treatment effect estimate, $\hat{\beta}_2$ (i.e., the test of the null hypothesis H_0:β_2=0),

reported earlier in Table 7-3. In general, both measures of the treatment effect may be useful in reporting results.

Regression Curves for Multiple Predictor Models

The monotonic regression curve given in Figure 7-7 and Figure 7-8 displays the predicted value for Y, as a function of the single predictor variable X. When there are *multiple* predictors as in the current example, two different displays of the regression curve may be obtained in a GOLDMineR session – a partial view for a selected predictor (illustrated in Figure 7-7) and a joint view of all predictors simultaneously . Figure 7-7 displays the partial view under Model A for AGE, controlling for *females* who received the *placebo*. For comparison, Figure 7-8 displays the corresponding (lower) predictions for *males* who received the *placebo*. Partial displays can also be obtained for TREAT (controlling for AGE and GENDER) and for GENDER (controlling for AGE and TREAT).

Note that the predicted improvement, represented by one of the arrows in Figure 7-4, for 50 year old *females* who received the *placebo*, is shown in Figure 7-7 as a point on the partial regression curve as a function of AGE.

Since the Y-scores used in Model A are 1 and 0, the expected value of Y for a given X as predicted in Figure 7-7 and Figure 7-8 by $\tilde{y}_{.X}$ represents the probability of *some* or *marked* improvement $P_{Y.x}$ as a function of AGE, where X = (AGE, GENDER = *Female*(in Figure 7-7) or *Male* (in Figure 7-8), and TREAT = *Placebo*):

$$E(Y|X) = 1(P_{Y.x}) + 0(1-P_{Y.x}) = P_{Y.x}$$

Figure 7-7
Plot of the likelihood of improvement predicted under Model A as a function of AGE given GENDER = Female and TREAT = Placebo

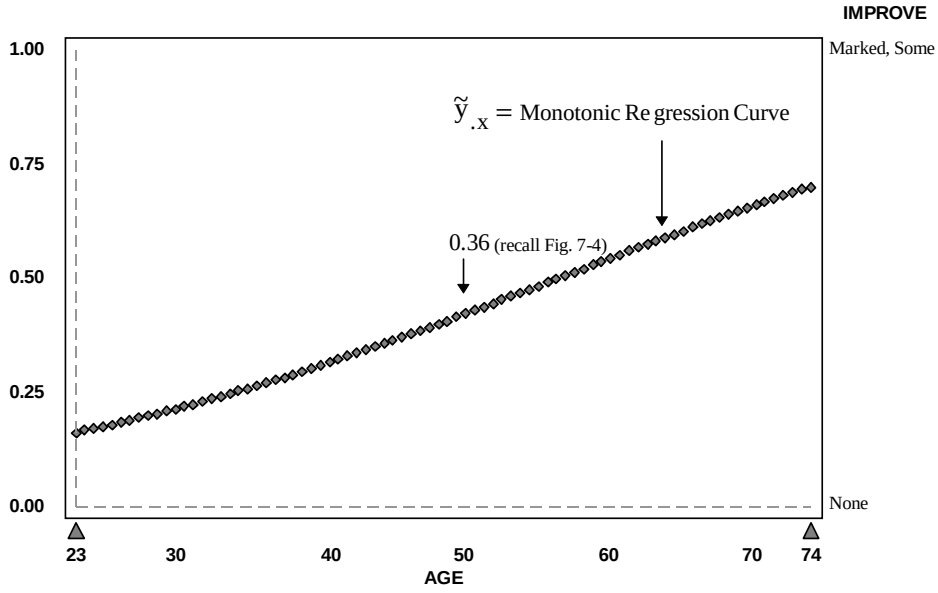
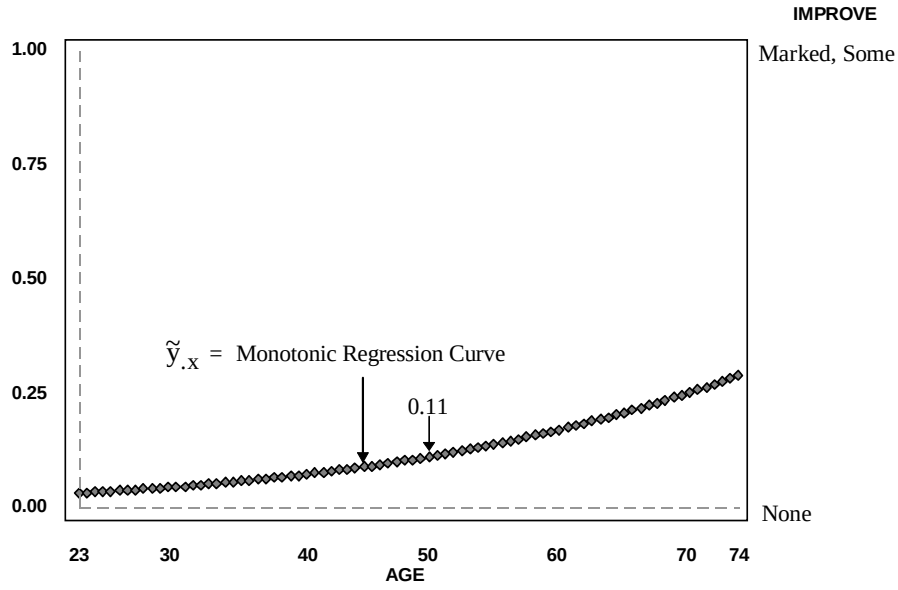


Figure 7-8 shows how the predicted probability of improvement, $\tilde{y}_{.x}$, increases with AGE at a continuous accelerating rate. Specifically, the curve shows the increase expected for *male* patients who received the *placebo*, as AGE varies from *23* (the minimum age of patients in the sample) to *74* (the maximum age).

Figure 7-8
Plot of the likelihood of improvement predicted under Model A as a function of AGE given GENDER = Male and TREAT = Placebo

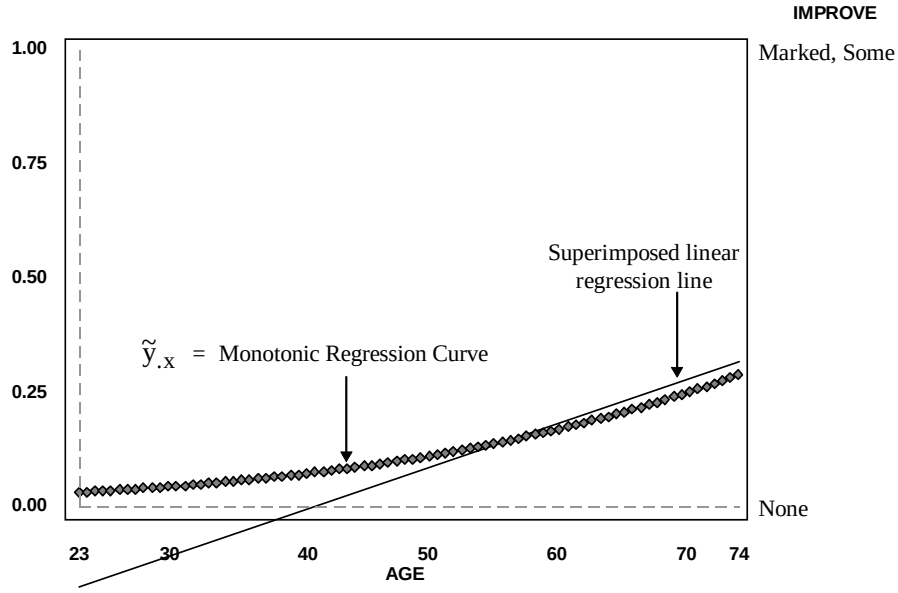


One of the properties of the monotonic predictor $\tilde{y}_{.x}$ is that predictions for X = x (whether x is observed in the sample or not) always fall between the minimum and maximum values for Y that are observed in the sample:

$y_{min} \leq \tilde{y}_{.x} \leq y_{max}$ for any x

Hence, meaningless out-of-bound predictions that occur frequently with the linear predictor $\hat{y}_{.x}$ (such as predicting probabilities or other positive expectations to be negative) never occur with the predictor $\tilde{y}_{.x}$. Figure 7-9 shows that the linear predictor yields meaningless predictions for the probability of improvement for control *males* under 40 years of age.

Figure 7-9
Plot of the likelihood of improvement predicted under Model A as a function of AGE given GENDER = Male and TREAT = Placebo including superimposed OLS regression line



GOLDMineR will optionally estimate a least squares regression line for the variables used in the monotonic regression model and overlay it on the plot as in Figure 7-9. Figure 7-10 and Figure 7-11 show that the linear model yields similar negative predictions under Models B and C as well. As in Model A where negative predictions took on the interpretation of meaningless probabilities, the negative predictions under Models B and C are also meaningless because for these data it is not possible to have an outcome less than the worst outcome (*None*), which has a Y-score of 0.

(Note: The optional regression line overlay is provided only to highlight by contrast the property of the monotonic regression curve that it never results in meaningless predictions.)

Figure 7-10
Plot of the likelihood of improvement predicted under Model B as a function of AGE given GENDER = Male and TREAT = Placebo including superimposed OLS regression line

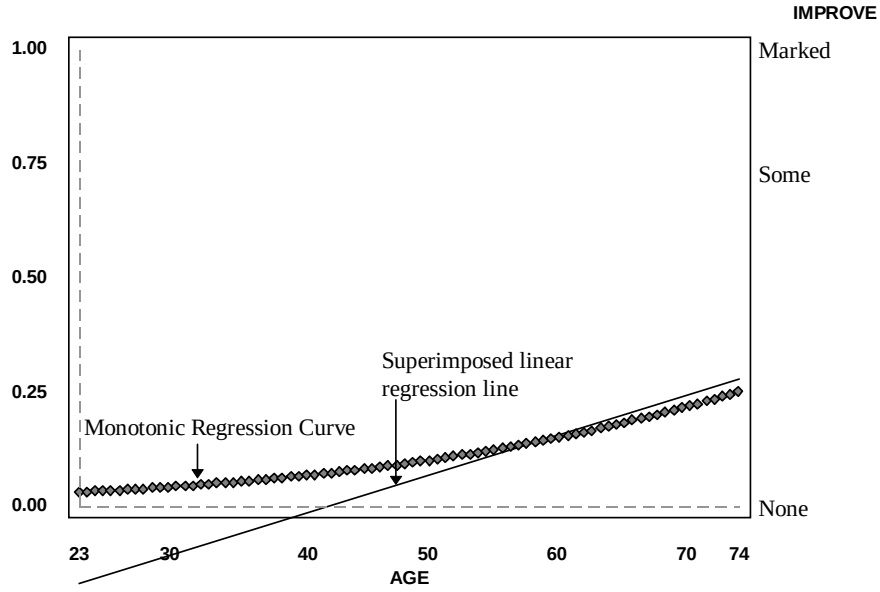
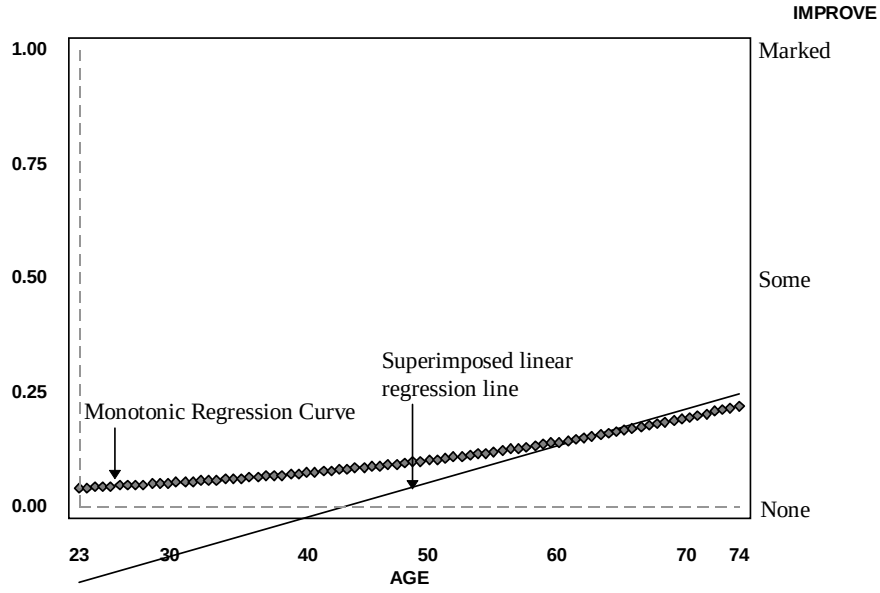


Figure 7-11
Plot of the likelihood of improvement predicted under Model C as a function of AGE given GENDER = Male and TREAT = Placebo including superimposed OLS regression line



Model Selection

Table 7-4 reports the estimates for β_1, β_2, and β_3 obtained under Models A and C together with the associated likelihood ratio chi-squared tests of significance given by $L^2(Y)$. All estimates are statistically significant. (The optional likelihood ratio difference statistic $L^2(Y)$ is generally preferable to the Wald chi-square given in Table 7-3 as a test of significance – both statistics may be displayed in GOLDMineR statistical output. Both yield comparable significance levels for these data.)

	m		Beta	L^2(Y)	d.f.	Significance (p-value)
	1	TREAT	1.76	12.2	1	.0005
Model A	2	GENDER	-1.49	6.98	1	.0083
	3	AGE	0.05	6.16	1	.013
	1	TREAT	2.13	14.4	1	.0002
Model C	2	GENDER	-1.48	5.4	1	.02
	3	AGE	0.05	5.1	1	.024

Table 7-4
Beta Estimates for Models A and C

After estimating several models, a single model is often selected that is the simplest (i.e., most parsimonious in terms of the number of parameters) among models that provide an adequate fit to the data. Models A and C are both simpler than Model B because only three parameters are estimated (β_1, β_2, β_3) rather than four (β_1, β_2, β_3 and y_2), as counted by the number of degrees of freedom ("d.f") in Table 7-2. Since the resulting increase in R^2 from about 0.26 (under Models A and C) to about 0.27 (under Model B) is not statistically significant*, one would generally use either Model A or C (not Model B) in reporting results in the analysis of the rheumatoid arthritis data.

The important topic of how to assess model fit using the conventional chi-square fit statistics, Residual L^2 or Residual X^2, when all predictors are categorical is discussed in Chapters 8 and 9. In the case that one or more predictors is continuous, as in the current example, we must resort to less conventional approaches to assess fit (see "A.15: Assessing Model Fit when Analyzing Sparse Data" on page 242).

* For example, from Table 7-2 we obtain a nonsignificant chi-squared difference: L^2(B) - L^2(C) = 1.0 with 4-3 = 1 degree of freedom. A similar chi-squared difference is obtained when comparing Model B with Model A.

8

Example 2 Assessing the Effects of a Reminder Call

Imagine that you design an experiment to estimate the return rate of your mail survey as a function of 1) the dollar amount mailed out with the survey (X_1: PAYMENT - *$1*, *$2*, or *$10*) and 2) a reminder call (X_2: CALL - - *Yes* or *No*). Table 8-1 contains frequency counts for a hypothetical mail survey (included in the file ex2.*) where the dichotomous dependent variable RETURN indicates whether an individual completes and returns the survey or not: *Yes*(1) or *No*(0), for the joint X variable formed by the different possible combinations of the two categorical predictor variables CALL and PAYMENT.

		Y:RETURN			
Joint X		**yes(1)**	**no(0)**	**Total**	**Percent Response**
	$1, yes(1)	266	734	1,000	26.6%
	$1, no(0)	19	81	100	19.0%
PAYMENT, CALL	**$2, yes(1)**	286	214	500	57.2%
	$2, no(0)	263	237	500	52.6%
	$10, yes(1)	82	18	100	82.0%
	$10, no(0)	801	199	1,000	80.1%

Table 8-1
Mail Survey Experiment: Hypothetical Data and Simple Descriptive Statistics

We will assess the effects of CALL and PAYMENT under each of two models that differ with respect to the category scores assigned to the three PAYMENT levels:

Model A: Fixed Dollar PAYMENT scale: scores = 1, 2, 10

Model B: Fixed Equidistant PAYMENT scale: scores = 1, 2, 3

Model A uses the original dollar payment amount to scale PAYMENT while Model B scales PAYMENT using the equidistant scores 1, 2, 3 in which the *$2* option is centered midway between the lower and higher payment amounts.

This example illustrates

- the use of the goodness of fit statistic to examine the validity of the scaling and other assumptions made by a model
- the poor prediction and biased parameter estimates that can result from model misspecification (i.e., use of improper scores).

Interpretation of Model Results

Table 8-2 shows the surprising result that very different effect estimates are obtained depending upon whether PAYMENT is scaled according to Model A or B. Specifically, we see that under Model B, the estimated effect of **CALL** is **positive** (0.22) and significant ($p < .05$) but under Model A, the estimated effect of **CALL** is **negative** (-0.19) and significant ($p < .05$).

While the results obtained under each of these models find the respective estimate to be statistically significant, the test of significance relies on the assumption that the model is true. Given one or more models, how can one tell which (if any) model is true? Fortunately, a chi-squared goodness of fit statistic is available which assesses the truth (validity) of assumptions made by a model, whenever the sample is sufficiently large and can be expressed as a table of nonsparse multiway frequency counts as in Table 8-1.

The model fit statistic, Residual L^2, for each model is reported in Table 8-3. The corresponding p-values show that Model B provides an excellent fit to the data in Table 8-1 ($p = 0.81$) while the fit under Model A is so poor ($p = 2.8 \times 10^{-27}$) that the fixed dollar scaling assumption made under that model must be rejected at any reasonable significance level. (The addition of a CALL x PAYMENT interaction effect in Model A does not improve the fit and the estimated effect of CALL becomes even more negative.)

		$\hat{\beta}$	**Significance (p-value)**
Model A	**PAYMENT**	0.21	5.1×10^{-70}
	CALL	-0.19	.042
Model B	**PAYMENT**	1.32	22×10^{-97}
	CALL	0.22	.037

Table 8-2
Effect Estimates

	Model Fit Residual L^2	**d.f.**	**Significance (p-value)**
Model A	126.7	3	2.8×10^{-27}
Model B	0.94	3	0.81

Table 8-3
Model Fit Statistics

Comparing the actual sample return rate in Table 8-1 for persons receiving a reminder call with the rate observed for persons receiving no reminder call but the same payment, suggests that the effect of the reminder call **is** positive -- an increase from 19.0% to 26.6% for those receiving a *$1* payment, 52.6% to 57.2% for those receiving a *$2* payment and 80.1% to 82.0% for those receiving a *$10* payment (see Table 8-4). However, Table 8-4 shows that under Model A, the predicted return rate is not only poor (in comparison with the observed rate), but persons who received a reminder call are predicted to have a lower return rate than the corresponding group receiving no reminder call (a consequence of the negative effect estimate for CALL).

		Response Rate		
	CALL	**$1**	**$2**	**$10**
Observed	**Yes**	26.6%	57.2%	82.0%
	No	19.0%	52.6%	80.1%
Model A	**Yes**	35.4%	40.3%	78.7%
	No	39.9%	45.1%	81.8%
Model B	**Yes**	26.4%	57.3%	83.4%
	No	22.4%	51.9%	80.1%

Table 8-4
Response Rates

Various plots available in GOLDMineR can provide insight into understanding why the biasing effect for CALL occurs under Model A. Figure 8-1 and Figure 8-2 show the partial Regression plot for PAYMENT under the CALL = *yes* and CALL = *no* situation respectively. The observed return rates given in Table 8-4 are plotted as circles in Figure 8-1 and Figure 8-2, the diameter of which is proportional to the corresponding sample size (shading means that the estimated expected rate differs significantly from the corresponding observed rate).

While the observed rates are higher for persons receiving the reminder call, the monotonic regression curve (as well as the linear regression line) is affected by a large correlation between PAYMENT and CALL (the sample sizes are largest for “no call” persons who received higher payment and “called” persons who received lower payments). This correlation serves to inflate the predictions of the return rates for persons receiving no reminder call and deflate the predictions for those called (i.e., the regression curve is “pulled down” in Figure 8-1 and “pulled up” in Figure 8-2 by the larger circles which carry greater weight because of the larger sample sizes).

Figure 8-1
Predicted return rate under Model A as a function of PAYMENT given CALL = 'Yes'

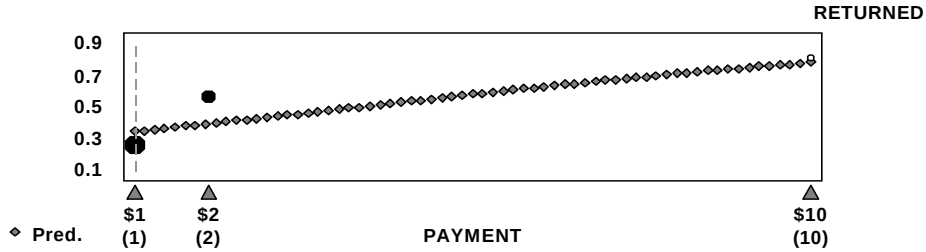


Figure 8-2
Predicted return rate under Model A as a function of PAYMENT given CALL = 'No'

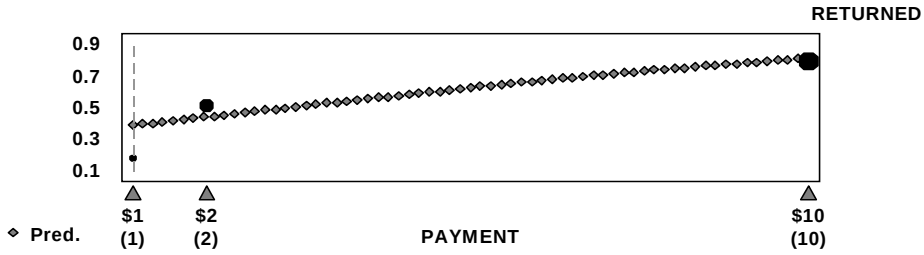


Figure 8-3 and Figure 8-4 show the corresponding plots under Model B where equidistant scores for PAYMENT of 1, 2 and 3 are used in place of 1,2, and 10. The latter model provides an excellent fit to the data and shows a significant *positive* effect for CALL.

Figure 8-3
Predicted return rate under Model B as a function of PAYMENT given CALL = 'Yes'

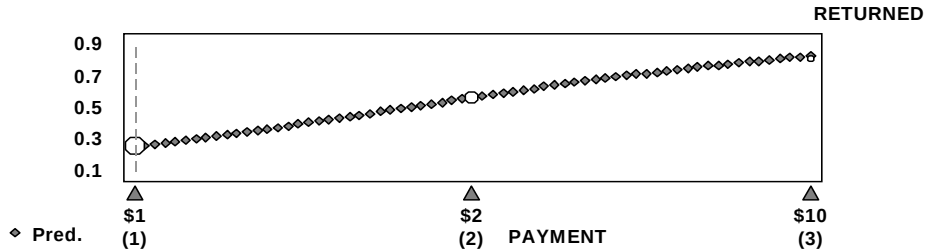
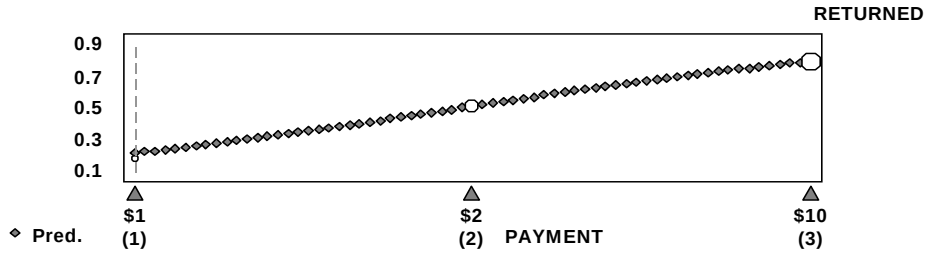


Figure 8-4
Predicted return rate under Model B as a function of PAYMENT given CALL = 'No'



Hence, the large goodness of fit statistic (Residual L^2=126.7) alerts us to problems with Model A and causes us to accept Model B (Residual L^2 = 0.94) over Model A for these data and to conclude that the effect of CALL is positive, not negative.

9

Example 3 Analysis of a Qualitative Predictor Variable

In Chapter 8 hypothetical mail survey data were used to demonstrate the importance of assessing the model fit to test the validity of the assumptions made by a model. In the current example we will analyze real data from the Mail Survey Experiment (Magidson, 1994b) where the actual dollar payments tested were *$1* (the control), *$2, $3* and *$4*, the experiment being designed so that there is no correlation between X_1:PAYMENT and X_2:CALL. We will again use the model fit statistic to choose between two alternative models, one that treats PAYMENT as a quantitative predictor with equidistant category scores (Model A) and one that treats PAYMENT as a qualitative predictor (Model B).

The data from the Mail Survey Experiment is summarized in Table 9-1 (included in file ex3.*). The observed rate of returning the survey is given for each joint category formed by the joint predictor X=CALL x PAYMENT, in percentages, odds and logit units. For example, among persons who received both a $1 payment and a reminder call, 55.2% responded. The observed odds in favor of response for this group is 55.2% / (100% - 55.2%) = 1.23 and the logit = ln(1.23) = 0.21.

		RETURN		Likelihood of Response		
CALL	**PAYMENT**	**Yes**	**No**	**Percentage**	**Odds**	**Logit**
Yes (1)	**$1 (1)**	2,407	1,954	55.2 %	1.23	0.21
	$2 (2)	1,265	881	58.9 %	1.44	0.36
	$3 (3)	1,340	809	62.4 %	1.66	0.50
	$4 (4)	1,306	779	62.6 %	1.68	0.52
No (0)	**$1 (1)**	2,133	2,176	49.5 %	0.98	-0.02
	$2 (2)	1,156	942	55.1 %	1.23	0.20
	$3 (3)	1,262	897	58.5 %	1.41	0.34
	$4 (4)	1,248	839	59.8 %	1.49	0.40

Table 9-1
Mail Survey Experiment with Descriptive Statistics

Model A, which treats PAYMENT as a *quantitative* predictor with equidistant category scores, was estimated using the **Uniform** scoring option. **Uniform** scores are equidistant scores that have a low score of 0 and a high score of 1.

The overall results for Model A are summarized in Figure 9-1 and the partial Effects plot for PAYMENT is displayed in Figure 9-2.

Figure 9-1
Model Summary Results of Model A which Specifies Equidistant Payment Scores

y: RETURN (Fixed)	0	1				
Y-scores	0.00	1.00				
Alpha	0.00	0.02				
		$L^2(Y)$	df	p-value	Beta	exp(Beta)
PAYMENT (Fixed)		119	1	1.1e-27	0.39	1.48
CALL (Fixed)		42.6	1	6.7e-11	0.18	1.20
Association Summary		L^2	df	p-value	R^2	
Explained by Model		161	2	1.2e-35	0.00749	
Residual		9.45	5	0.093		

Figure 9-2
Model A: PAYMENT is treated as Fixed with $1 as the reference

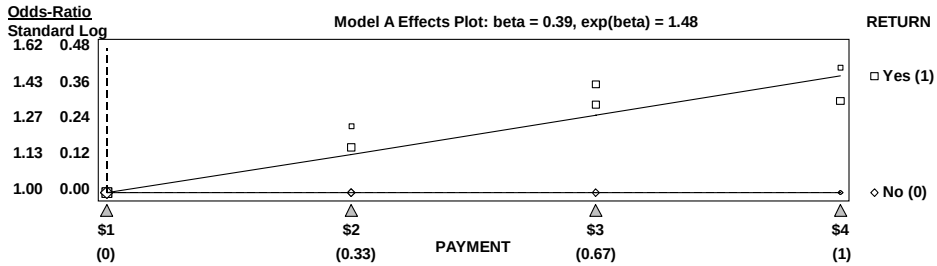


Figure 9-2 shows that persons receiving a $2 payment and for whom CALL = Yes(No) are estimated to be exp(0.33 $\hat{\beta}_1$) = 1.14 times as likely to return the survey than persons receiving a $1 payment and for whom CALL = Yes(No). Similarly, given CALL = Yes(No), the effect of a $3 payment is estimated to be exp(0.67 $\hat{\beta}_1$) = 1.30 and a $4 payment is exp($\hat{\beta}_1$) = 1.48 times as likely to return the survey than persons receiving a $1 payment (who were exposed to the same calling option). Uniform scores are displayed in parenthesis.

Model B, which treats PAYMENT as a *qualitative* predictor, was estimated using the **Free** scaling option for the PAYMENT variable. In order that the category scores estimated for any **Free** scale variable such as PAYMENT be identifiable, two standardizing restrictions must be imposed to uniquely define a zero point and a unit. By default, like **Uniform** scores, GOLDMineR restricts the scores by fixing the lowest at 0 and highest at 1. The overall model results for Model B are summarized in Figure 9-3 and the corresponding Effects plot is displayed in Figure 9-4. The estimated PAYMENT scores, labeled X-score in Figure 9-3, are used to position the payment levels on the horizontal axis of the partial Effects plot.

Figure 9-3
Model Summary Results from Model B which specifies Free PAYMENT Scores (with $1 payment serving as the reference category)

y: RETURN (Fixed)	0	1
Y-scores	0.00	1.00
Alpha	0.00	0.00

	$L^2(Y)$	df	p-value	Beta	exp(Beta)
PAYMENT (Free)	126	3	3.9e-27	0.36	1.44
CALL (Fixed)	42.6	1	6.7e-11	0.18	1.20

Association Summary	L^2	df	p-value	R^2
Explained by Model	168	4	2.9e-35	0.00783
Residual	2.38	3	0.50	
Total	170	7	2.2e-33	

Category-Specific Parameter Estimates

PAYMENT (Free)	X-score	C-weights	Beta(k)	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
1	0.00	1.00	0.00	.	.	.	1.00	.	.
2	0.52	0.00	0.19	0.04	25.02	5.7e-7	1.21	1.12	1.30
3	0.91	0.00	0.33	0.04	75.34	4.0e-18	1.39	1.29	1.50
4	1.00	0.00	0.36	0.04	89.40	3.2e-21	1.44	1.33	1.55

Figure 9-4
Model B: PAYMENT is treated as a FREE predictor with $1 used as the reference

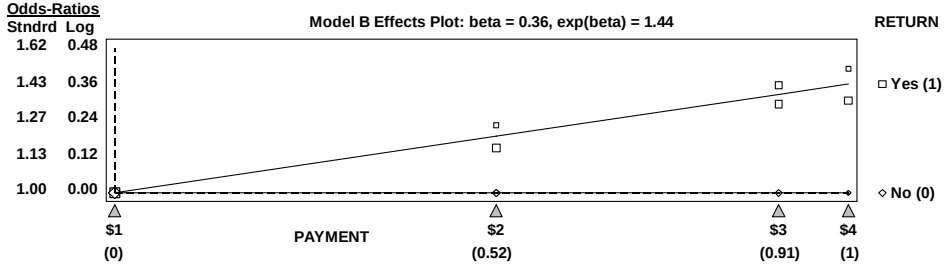


Figure 9-4 shows that persons receiving a $2 payment and for whom CALL = Yes(No) are estimated to be exp(0.52$\hat{\beta}_1$) = 1.21 times as likely to return the survey than persons receiving a $1 payment and for whom CALL = Yes(No). Similarly, given CALL = Yes(No), the effect of a $3 payment is estimated to be exp(0.91$\hat{\beta}_1$) = 1.39 and a $4 payment is exp($\hat{\beta}_1$) = 1.44 times as likely to return the survey than persons receiving a $1 payment (who were exposed to the same calling option).

Since PAYMENT is treated as a **Free** scale (qualitative) variable, GOLDMineR estimates category-specific effects $\beta_k = \beta x^*_{1k}$, for each payment level k = 1, 2, 3,4 and lists them in a separate section of the output. By default, dummy contrast coding is used with the category allocated the 0 score (the *$1* PAYMENT category) being the one omitted from the regression (see "Category Weights" on page 119). The estimate for a *$2* payment, shown in Figure 9-3 as *Beta*(2) = 0.19, represents the increased likelihood of returning the survey associated with a *$2* payment vs. a *$1* payment. Examining the column titled *p-value* in Figure 9-3, we see that this effect is statistically significant ($p=5.7x10^{-7}$), and that $\hat{\beta}_3$ and $\hat{\beta}_4$ are also significantly different from zero.

To test the adequacy of the equidistant scoring assumption we compare the model fit statistics from the two models (see "A.10 Estimation Algorithm" on p. 228). Examining the output in Figure 9-1 and Figure 9-3 we see that Residual L^2 has been reduced from 9.45 with 5 d.f. under Model A to 2.38 with 3 d.f. under Model B, a reduction of 7.07 (with 5-3 = 2 d.f.) which is statistically significant at the p = .05 level. Hence, we reject the equidistant scaling assumption made by Model A in favor of Model B.

Suppose that we estimated only Model A. Would we still reject the equidistant scoring assumption? While Residual $L^2 = 9.45$ (with 5 d.f.) is an adequate fit (p = .09), a comparison of the observed and expected log-odds ratio in Figure 9-2 might lead us to reexamine the equidistant scoring assumption. Specifically, the observed log-odds ratios for persons who received a reminder call, as well as the ratios for those who did not receive a call, both appear above the corresponding expected ratios (i.e., above the effects line) for each of the two middle payment categories (*$2* and *$3* payments). Note the improvement in fit attained in Figure 9-4 when the PAYMENT scores are no longer restricted to be equidistant.

Figure 9-5 and Figure 9-6 illustrate partial Regression plots for PAYMENT given CALL under Model A. A careful examination of the pattern of residuals in these figures would also raise some questions about the correctness of the equidistant PAYMENT scores assumption. In particular, the two shaded circles show that the predicted return rates for the '*$1*, no call' (see Figure 9-5) and '*$4*, call' (see Figure 9-6) cells differ significantly from the observed rates. Figure 9-7 and Figure 9-8 show the closer fit to the observed return rates obtained under Model B where PAYMENT is treated as a qualitative variable (**Free** scale).

Figure 9-5
Model A: Predicted Probability of RETURN = 'Yes' by PAYMENT given CALL = 'No'

o = observed return rate (shaded means that observed rate is significantly different from expected)

Figure 9-6
Model A: Predicted Probability of RETURN = 'Yes' by PAYMENT given CALL = 'Yes'

Figure 9-7
Model B:Predicted Probability of RETURN = 'Yes' by PAYMENT given CALL = 'No'

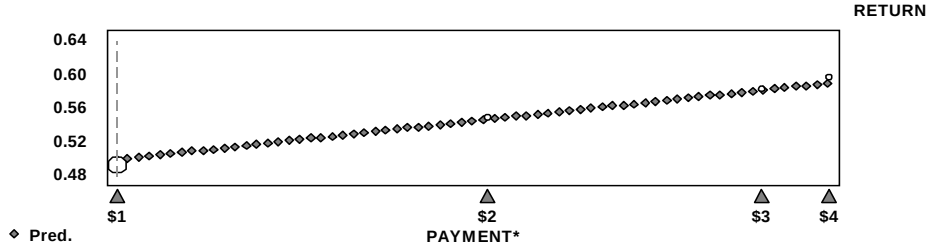
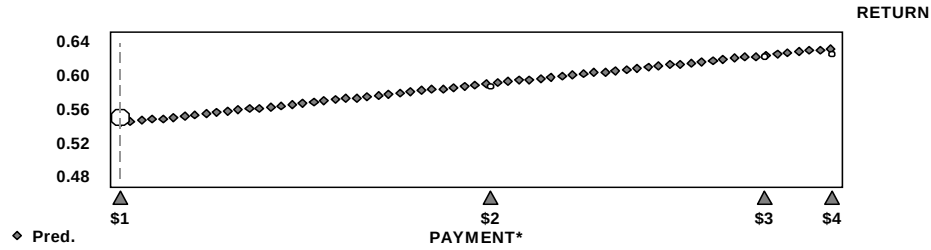


Figure 9-8
Model B: Predicted Probability of RETURN = 'Yes' by PAYMENT given CALL = 'Yes'



Should we accept Model B as our final model?

The category-specific parameter estimate section of Figure 9-3 confirmed that a *$2* payment (as well as *$3* and *$4* payments) are significantly better than a *$1* payment ($p = 5.7 \times 10^{-7}$). That is, all other factors being equal, persons receiving a *$2* payment are estimated to be 1.21 times as likely as those receiving a *$1* payment to return the survey (see the caption under Figure 9-4). However, it is not clear from Figure 9-3 whether the estimate of $\exp(\beta_4) = 1.44$ is significantly higher than the estimate of $\exp(\beta_3) = 1.39$. Similarly, from an examination of the Effects plot in Figure 9-4 we might ask the question of whether a *$4* payment is significantly better than a *$3* payment.

To test the effect of a *$4* payment relative to a *$3* payment we must change the reference category from the *$1* level to the *$3* level by the appropriate change in contrast coding. Once a model has been estimated, alternative contrast coding options may be selected to obtain different graphical views of the model and different category-specific parameter estimates, without reestimating the model (i.e., only the category-specific parameter estimate section of the output changes). Figure 9-9 and Figure 9-10 provide results for Model B when the reference is changed to the *$3* level. Note that the X-scores in Figure 9-9 are lower than those in Figure 9-3 by 0.91, the value of the original X-score shown in Figure 9-3 for the *$3* payment level).

Figure 9-9
Category-specific parameter estimates for PAYMENT under Model B where $3 is the reference

Category-Specific Parameter Estimates

PAYMENT (Free)	X-score	C-weights	Beta(k)	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
$1	-0.91	0.00	-0.33	0.04	75.34	4.0e-18	0.72	0.67	0.78
$2	-0.39	0.00	-0.14	0.04	10.12	0.0015	0.87	0.80	0.95
$3	0.00	1.00	0.00	.	.	.	1.00	.	.
$4	0.09	0.00	0.03	0.04	0.59	0.44	1.03	0.95	1.13

Figure 9-10
*Model B: PAYMENT is treated as a **Free** predictor with $3 used as the reference to assess the significance of $4 vs. $3*

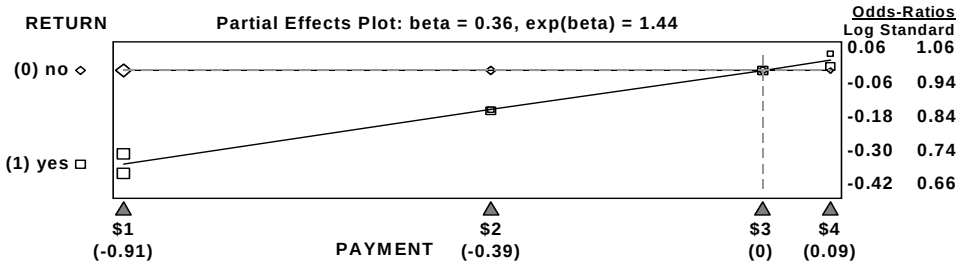


Figure 9-10 shows that persons receiving a $4 payment and for whom CALL = Yes(No) are estimated to be exp(0.09$\hat{\beta}_1$) = 1.04 times as likely to return the survey than persons receiving a $3 payment and for whom CALL = Yes(No).

Figure 9-9 shows that the effect of *$4* relative to *$3*, $\hat{\beta}_4$ = .034, is not significant at the .05 level (p = 0.44). This result suggests that we assign the same score to a *$3* and a *$4* payment and estimate a final model. One possibility for a final model is to recode the original datafile to combine payment levels *$3* and *$4* and estimate a model with PAYMENT treated as either Free or Fixed using equidistant scores. Recoding could also be implemented by using the Group feature in GOLDMineR which will automatically combine the *$3* and *$4* payment levels if 3 groups are specified (for more information on the Grouping feature, see "Variable Scaling Options" on page 90).

For the remainder of this section, we maintain separate categories for the *$3* and *$4* payment levels and assign **Fixed** PAYMENT scores of 0, 0.5, 1 and 1. That is, model C positions the *$2* payment equidistant between *$1* and *$3*, and positions the *$4*

payment at the same place on the plot as the *$3* payment. Figure 9-11 summarizes the results and Figure 9-12 displays the effects plot under Model C.

Figure 9-11
*Results of model specifying **Fixed** PAYMENT scores: 0, 0.5, 1, 1*

y: RETURN (Fixed)	0	1				
Y-scores	0.00	1.00				
Alpha	0.00	0.01				
		$L^2(Y)$	df	p-value	Beta	exp(Beta)
PAYMENT (Fixed)		125	1	4.5e-29	0.35	1.41
CALL (Fixed)		42.6	1	6.6e-11	0.18	1.20
Association Summary		L^2	df	p-value	R^2	
Explained by Model		167	2	5.1e-37	0.00779	

Figure 9-12
*Effects Plot for Model C specifying **Fixed** PAYMENT scores: 0, 0.5, 1, 1*

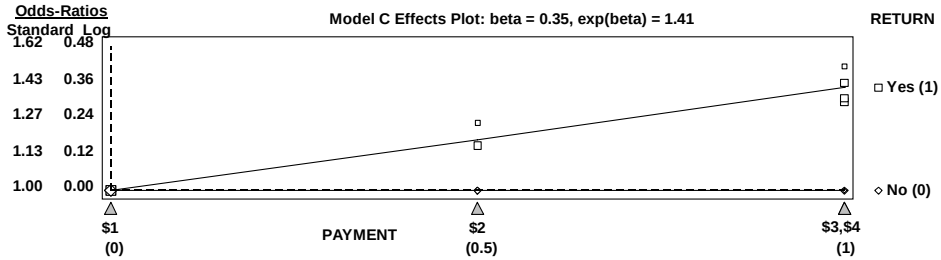


Figure 9-12 shows that persons receiving a $3 or $4 payment and for whom CALL = Yes(No) are estimated to be exp($\hat{\beta}_1$) = 1.41 times as likely to return the survey than persons receiving a $1 payment and for whom CALL = Yes(No).

Under Model C, each additional *$1* PAYMENT up to *$3* is expected to increase the likelihood of RETURN. A payment of *$4* is not expected to lift response above that of a *$3* payment.

10

Example 4 Measuring and Analyzing the Association Explained by a Model

This chapter introduces some advanced topics that are important to understand when interpreting the results of estimating a simple logit or logistic regression model or a more complex model. These topics are not typically included in an introductory course on regression. Specifically, the section titled "Understanding the Analysis of Association Summary" on page 191 uses sample dataset #4 to show how the independence chi-squared statistic can be partitioned into two different chi-squared statistics, one that tests for association related to an ordinal dependent variable and a second that tests the validity of assumptions made by the fitted model. The section titled "Interpreting the Effect Estimates" on page 195 defines odds and odds ratios which are important to understand when interpreting the parameter estimates, and uses graphical displays to help visualize the meaning of the odds ratios and estimated parameters as well as issues of model fit.

Understanding the Analysis of Association Summary

Included in the Statistics output is an analysis of association (ANOAS) summary which partitions the total amount of qualitative association between the rows and columns of a 2-way table into an ordinal portion explainable by the assumed model in which Y-scores are used to quantify the ordinal structure of the dependent variable, and a residual portion representing unexplained association. The total qualitative association is measured by the likelihood ratio chi-squared statistic, Total L^2, which tests for independence between the rows and columns of the table:

$$\text{Total } L^2 = \text{Explained } L^2 + \text{Residual } L^2$$

In this section, we use a simple 2 x 5 crosstabulation of clinical trial data included in the file ex4.* consisting of a single dichotomous predictor variable X:TRTMNT (*test*

drug, placebo), and a five category ordinal dependent variable Y:OUTCOME (*marked improvement, moderate improvement, slight improvement, stationary, worse*) to illustrate an ANOAS partitioning. Since there are only 2 rows in the table – 1) patients who receive the *test drug*, and 2) patients who receive the *placebo*, the model fit statistic assesses the validity of the assumed Y-scores in the model (See "A.8 Partitioning Chi-Square" on p. 225).

			(Y) OUTCOME					
		j =	**1**	**2**	**3**	**4**	**5**	**Avg.**
i	**x**	**X:TRTMNT**	**Worse**	**Stationary**	**Slight**	**Moderate**	**Marked**	**Y-Score**
1	**1**	**Test Drug**	1	13	16	15	7	1.3
2	**0**	**Placebo**	5	21	14	9	3	0.7
		Y-scores =	-1	0	1	2	3	

Table 10-1
Observed Counts (f_{ij}) for the 2x5 Table of Treatment by Outcome (Source: DeJonge, 1983)

For our analysis of the data in Table 10-1, we will use the equidistant Y-scores = -1, 0, 1, 2, and 3, to structure the ordinal nature of OUTCOME. We will see that the use of equidistant scores results in the treatment effect β being interpretable as the adjacent category log odds ratio. The ANOAS Summary in Table 10-2 shows that 1) the Explained L^2 is large enough to reject the hypothesis that $\beta = 0$ so we conclude that the treatment effect is statistically significant (at the .01 level), and 2) the Residual L^2 is small so we accept the model hypothesizing equidistant Y-scores since it provides a good fit to the data (p=.91):

Analysis of Association	**Hypothesis Tested**	**L^2**	**d.f.**	**p-value**	**Result**
Explained by Model	$\beta = 0$ in model	7.6	1	0.01	REJECT
Residual	Equidistant Y-scores	0.5	3	0.91	ACCEPT
Total Association	Independence	8.1	4	0.09	ACCEPT

Table 10-2
Analysis of Association for the Equidistant Y-Score Model

While the analysis leads us to conclude that the treatment effect is statistically significant (Explained L^2 =7.6, df = 1, p = .01), the test for independence would suggest that the treatment effect is *not* significant at the .05 level (Total L^2 = 8.1, df =

4, p = .09). Reconciliation of the results obtained from the different tests may be obtained by noting that the test based on Explained L^2 utilizes only 1 degree of freedom and hence is more powerful than the test for independence. By treating OUTCOME as a qualitative (nominal) variable, the test for independence lacks power to detect significant relationship between TRTMNT and Y. However, the more focused test, base on Explained L^2 reveals a highly significant relationship between TRTMNT and Y. The focused test has higher power than the test for independence by making use of the assumed Y-score structure to preserve all of the Y-Association in the data but filter out any residual association that is not relevant to the assumed Y-scores (see "A.8 Partitioning Chi-Square" on p. 225).

In summary, as an alternative to a GOLDMineR analysis, if we tested only for independence in Table 10-1 we would not have detected a significant treatment effect. On the other hand, by taking advantage of the fact that the hypothesis of independence (with 4 degrees of freedom) is equivalent to the hypotheses that 1) our assumed model is true (with 3 degrees of freedom) and 2) $\beta = 0$ in our model (with 1 degree of freedom), we are able to partition the test for independence into two portions. The first portion provides a test of the truth of the model. We conclude that the model with the assumed Y-scores provides a good fit to the data. The second portion provides a much more powerful test for a treatment effect by reducing the number of degrees of freedom from 4 to 1. We conclude that the treatment effect is highly significant (p = .01).

Table 10-3 provides estimated expected counts obtained under the assumed model. Notice that the outcome score predicted for each treatment condition matches exactly the corresponding observed quantities shown in Table 10-1 (a predicted outcome of $\tilde{y}_{\cdot 1} = 1.3$ for patients who received the *test drug* and $\tilde{y}_{\cdot 2} = 0.7$ for those who received the *placebo*). The exact match is due to the fact that maximum likelihood estimates for the expected frequencies obtained under the assumed model preserve all of the Y-Association in the data (see "A.10 Estimation Algorithm" on p. 228). The test of independence applied to the estimated expected counts in Table 10-3 yields and Explained L^2 of 7.6, a measure of the amount of Y-Association in the data.

			(Y) OUTCOME					E(Y|X=x)
		j =	**1**	**2**	**3**	**4**	**5**	**Avg.**
i	**x**	**X:TRTMNT**	**Worse**	**Stationary**	**Slight**	**Moderate**	**Marked**	**Y-Value**
1	**1**	**Test Drug**	1.58	12.78	15.1	15.12	7.41	1.3
2	**0**	**Placebo**	4.42	21.22	14.90	8.88	2.59	0.7
		Y-scores =	-1	0	1	2	3	

Table 10-3
Expected Frequency Counts (F_{ij}) Estimated Under the Assumed Model

Note: The predicted Y-values in Table 10-3 are obtained using the estimated expected frequencies to compute a weighted average of the Y-scores. For example, the average Y-value for persons taking the 'test drug' is calculated as (1.58-1 + 12.78*0 + 15.10*1 + 15.12*2 + 7.41*3) / (1.58+12.78+15.10+15.12+7.41) = 1.3.*

Table 10-4 provides expected counts estimated under the independence model (obtained in GOLDMineR by specifying TRTMNT to be a Margin Only predictor). Notice that the outcome score predicted for each treatment condition is identical ($\tilde{y}_{\cdot 1} = \tilde{y}_{\cdot 2} = 0.9$), a consequence of the fact that under the independence model all association (including Y-Association) is assumed to be null. (A test of independence applied to Table 10-4 yields $L^2 = 0$.)

			(Y) OUTCOME					Predicted E(Y|X=x)
		j =	**1**	**2**	**3**	**4**	**5**	**Avg.**
i	**x**	**X:TRTMNT**	**Worse**	**Stationary**	**Slight**	**Moderate**	**Marked**	**Y-Value**
1	**1**	**Test Drug**	3	17	15	12	5	0.9
2	**0**	**Placebo**	3	17	15	12	5	0.9

Table 10-4
Expected Frequency Counts (F_{ij}) Estimated Under the Model of Independence

Interpreting the Effect Estimates

In general, for any 2-way table, all of the qualitative association between the I rows and J columns of the table can be expressed in terms of (I-1)x(J-1) distinct effects which correspond to odds ratios. Table 10-5 shows the (2-1)x(5-1) = 4 distinct odds ratios (in logarithmic units) obtained from Table 10-1 for the *test drug* using the *placebo* treatment and *stationary* outcome as bases for the calculations. Notice that the log odds ratios increase as the outcome goes from *worse* to *marked* improvement, a pattern that indicates positive association in the table. For comparison, Table 10-5 also shows the corresponding log odds ratios obtained from the model that posited equidistant Y-scores, and those obtained from the independence model that assumes no association in the table.

In this section we show how log-odds ratios are calculated and how they relate to the β-effect parameters that are estimated by GOLDMineR for a given model.

Model:	**Worse**	**Stationary**	**Slight**	**Moderate**	**Marked**	**Total Association**
No Model (Table 10-1)	-1.13	0.00	0.61	0.99	1.33	8.1
Y-Association (Table 10-3)	-0.52	0.00	0.52	1.04	1.56	7.6
Independence (Table 10-4)	0.00	0.00	0.00	0.00	0.00	0.0

Table 10-5
Comparison of Log Odds Ratios Obtained Under 3 Models

Proper interpretation of the α and β-effect estimates, Effects plots, and related L^2-statistics that are presented as output from a GOLDMineR analysis requires an understanding of generalized odds and odds ratios (in both standard and logarithmic units) since these quantities are the "building blocks" that define the effects. In this section we illustrate the computation of those generalized odds ratios that are obtained using the default “baseline contrasts” in a GOLDMineR analysis. We also show how to obtain different odds ratios by changing the E-weights for the dependent variable and/or C-weights for the predictor(s) and how to interpret the resulting effect estimates.

Baseline Odds

By default, a particular OUTCOME category is used as the base for computing baseline odds. In our example, we implicitly designated the *stationary* (j=2) category as the base by assigning it a Y-score of 0 (i.e., $y_2 = 0$).

The baseline odds predicted under the assumed model for a given row in Table 10-3 are calculated by dividing each estimated expected count in that row by the corresponding estimated expected count associated with the OUTCOME base or reference category, *stationary*. Table 10-6 provides the results of calculating the baseline odds. For example, for patients who receive the *test drug*, the expected odds in favor of *marked* (vs. *stationary*) improvement is 7.41/12.78 = 0.58 (see Table 10-6). Note that by definition, the expected odds associated with the base category is always equal to 1 (here, 12.78/12.78 = 1).

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	0.12	1.00	1.18	1.18	0.58
Placebo	0.21	1.00	0.70	0.42	0.12

Table 10-6
Expected (Baseline) Odds for Clinical Trial Data; Reference Point: (OUTCOME = 'Stationary')

Next, the predicted baseline odds *ratios* are obtained by dividing each expected odds (from Table 10-6) by the corresponding *baseline* odds associated with the predictor reference category *placebo*, which was assigned an X-score of 0 (see Table 10-7 for the results of the calculation). For example, as given previously, the estimated expected odds ratio associated with the *test drug* and *marked* improvement cell is computed by dividing the estimated expected odds in favor of *marked* (vs. *stationary*) improvement given the *test drug* by the corresponding odds in favor of a *marked* improvement given the *placebo* -- i.e., 0.58/0.12 = 4.75.

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	0.59	1.00	1.68	2.83	4.75
Placebo	1.00	1.00	1.00	1.00	1.00

Table 10-7
Expected Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = 'Stationary', TRTMNT = 'Placebo')

From Table 10-7 we see that the odds of a 1-unit improvement (from y=0 to y=1) is 1.68, a 2-unit improvement (from y=0 to y=2) is 2.83 and a 3-unit improvement (from y=0 to y=3) is 4.75 times as high for patients who received the *test drug* than for those who received the *placebo*. The corresponding odds ratio for a 4-unit improvement (from y=-1 to y=3) is obtained by dividing the odds ratio for a 3-unit improvement from the stationary category (4.75) by the odds ratio for a 1-unit change in the negative direction (0.59), which yields 8.05.

Conversion to Baseline Logits

Odds are often expressed in logarithmic units (called logits) in order to translate the underlying multiplicative model into an additive form (see Equation 10-1 and Equation 10-2). Taking the natural log of both the baseline odds and odds ratios yields the baseline logits and log-odds ratios given in Table 10-8 and Table 10-9 respectively.

The difference between the log-odds ratios between adjacent categories in Table 10-8 is constant and is called the adjacent category log-odds ratio. To verify that it is constant, compute the difference in the log-odds ratios between *slight* and *stationary* (0.52 - 0 = 0.52), between *moderate* and *slight* (1.04 - 0.52 = 0.52), and between *marked* and *moderate* (1.56 - 1.04 = 0.52).

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	-2.09	0.00	0.17	0.17	-0.54
Placebo	-1.57	0.00	-0.35	-0.87	-2.10

Table 10-8
Expected Baseline Logits for Clinical Trial Data; Base: (OUTCOME = 'Stationary')

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	-0.52	0.00	0.52	1.04	1.56
Placebo	0.00	0.00	0.00	0.00	0.00

Table 10-9
Expected Log-odds Ratios for Clinical Trial Data; Reference Point: (OUTCOME = 'Stationary', TRTMNT = 'Placebo')

Formally, the model for this example expresses the conditional baseline logits as the sum of two terms:

$$\Psi_{j.i} = \alpha_j + \beta\, x_i\, y_j\,; \qquad j = 1,2,3,4,5 \qquad \text{Equation 10-1}$$

$$\text{where } x_i = \begin{cases} 1 \text{ if test drug } (i = 1) \\ 0 \text{ if placebo } (i = 2) \end{cases}$$

and the OUTCOME categories are assumed to be equidistant by assigning the equidistant scores -1, 0, 1, 2 and 3.

The intercept, α_j, represents the jth logit under the *placebo* condition and the second term, β, represents the effect of the treatment defined as the expected improvement associated with the *test drug*. The expected improvement is calculated in logarithmic units for the j^{th} outcome as the difference between the expected logit under the drug treatment ($\Psi_{j.1}$) and the expected logit under the placebo condition ($\Psi_{j.2}$).

From Equation 10-1, the expected logit under the drug treatment is $\Psi_{j.1} = \alpha_j + \beta*1*y_j = \alpha_j + \beta y_j$, and the expected logit under the placebo condition is $\Psi_{j.2} = \alpha_j + \beta*0*y_j = \alpha_j$. Hence, the effect of the treatment in Equation 10-1 is $\Psi_{j.1} - \Psi_{j.2} = \alpha_j + \beta y_j - \alpha_j = \beta y_j$. Since $y_j = 1$ for the outcome of *slight* improvement (j = 2), we see that the effect parameter β measures the effect of a 1-unit change from *stationary* (y = 0) to *slight* (y = 1). For this example, $\hat{\beta} = 0.52$, the estimated adjacent category log-odds ratio.

The multiplicative form of the model specified in Equation 10-1 is obtained through exponentiation:

$$\exp(\Psi_{j.i}) = \exp(\alpha_j)\,[\exp(\beta X)]^{yj} \qquad \text{Equation 10-2}$$

Since the additive effect $\beta\, x_i\, y_j$ is a log-odds ratio, the multiplicative effect, $\exp(\beta\, x_i\, y_j)$ is an odds ratio. Estimates for α and β in both additive and multiplicative form are given in Table 10-10. (Further technical information on Equation 10-1 and

Equation 10-2, including the identifying restrictions on the parameters are given in sections A.1, A.2, A.3 and A.5 of the Appendix.)

j	**Y**	**OUTCOME**	α_j	**exp(α_j)**
1	3	marked	-2.10	0.12
2	2	moderate	-0.87	0.42
3	1	slight	-0.35	0.70
4	0	stationary	0.00	1.00
5	-1	worse	-1.57	0.21
			β	**exp(β)**
	X	**TRTMNT**	0.52	1.68

Table 10-10
Parameter Estimates

It is important to note that since the log-odds ratios are measures of the association in the data, under the model specified in Equation 10-1, only 1 parameter, β, is needed to describe all of the association in the table. Specifically, the expected associations are given by Table 10-11:

	Worse	**Stationary**	**Slight**	**Moderate**	**Marked**
Test Drug	-β	0	β	2β	3β
Placebo	0	0	0	0	0

Table 10-11
Expected Associatons

The fact that only 1 parameter is required to describe the association is the reason that there is only 1 degree of freedom in the ANOAS results corresponding to Y-Association. Next, we will calculate the corresponding log-odds ratios in the observed data and show how to read an effects plot to determine how well the assumed model fits the data.

Calculating and Plotting Observed Odds Ratios

The cell entries in Table 10-6, Table 10-7, Table 10-8 and Table 10-9 were derived using expected counts given in Table 10-4, estimated by GOLDMineR under the assumed model. Table 10-12, Table 10-13, Table 10-14 and Table 10-15 provide the corresponding *observed* odds, odds ratios, log-odds and log-odds ratios computed using the original sample data in Table 10-1.

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	0.08	1.00	1.23	1.15	0.54
Placebo	0.24	1.00	0.67	0.43	0.14

Table 10-12
Observed Odds for Clinical Trial Data; Reference Point: (OUTCOME = 'Stationary')

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	0.32	1.00	1.85	2.69	3.77
Placebo	1.00	1.00	1.00	1.00	1.00

Table 10-13
Observed Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = 'Stationary', TRTMNT = 'Placebo')

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	-2.56	0.00	0.21	0.14	-0.62
Placebo	-1.44	0.00	-0.41	-0.85	-1.95

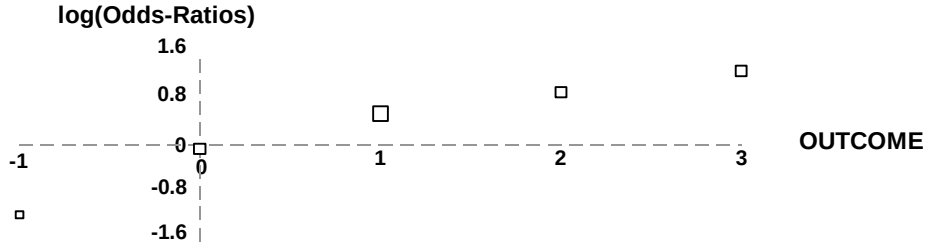
Table 10-14
Observed Log-Odds for Clinical Trial Data; Reference Point: (OUTCOME = 'Stationary')

	Worse	Stationary	Slight	Moderate	Marked
Test Drug	-1.13	0.00	0.61	0.99	1.33
Placebo	0.00	0.00	0.00	0.00	0.00

Table 10-15
Observed Log-Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = 'Stationary', TRTMNT = 'Placebo')

Figure 10-1 plots the observed log-odds ratios for the *test drug* from Table 10-15 as a function of the assumed equidistant Y-scores. Notice that the points lie almost on a straight line as expected under the assumed model. The approximate linearity provides evidence that the equidistant Y-score assumption is valid. Note that larger symbols reflect cells that are based on larger sample sizes.

Figure 10-1
Plot of Observed Log-Odds

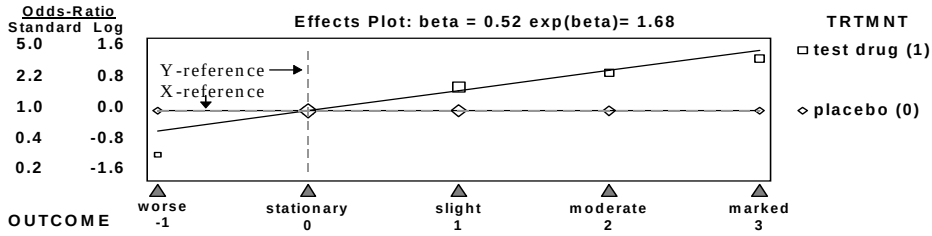


Reading the Effects Plot

Figure 10-2 superimposes "effects lines" onto the plot in Figure 10-1 and adds the 0 log-odds ratios corresponding to the placebo condition. Effects lines for each treatment condition are created by plotting the expected log-odds ratios for that treatment (from Table 10-9) as a function of Y and connecting the points. (An alternative view of the effects lines, called the X-view, is obtained by plotting the expected log-odds ratios for each outcome as a function of X.) Formally, the effects lines are obtained by re-expressing Equation 10-1 as $\Psi_{j.i} - \alpha_j = \beta x_i y_j$, where $\Psi_{j.i} - \alpha_j$

represents the expected log-odds ratio. Hence, β represents the slope of the effects line for the *test drug*.

Figure 10-2
Effects Plot with Observed Log Odds Ratio Points and Effects Lines



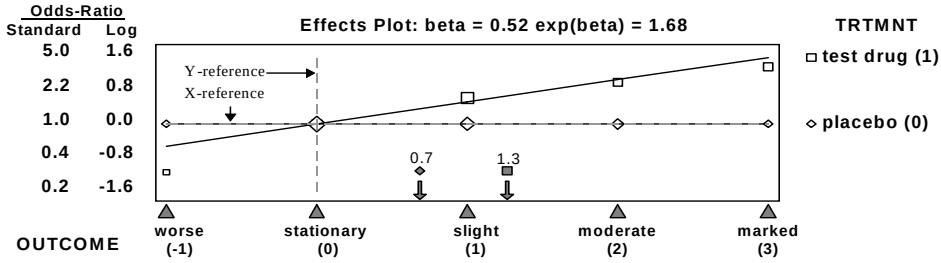
It can be seen from the test drug effect line that the odds of having a 'marked' improvement (vs. 'stationary') is exp(3 beta) = 4.75 times as high for patients who received the test drug than for patients who received the placebo. See Table 10-7 for corresponding expected odds ratios. The good fit is supported by the fact that the observed log-odds ratio symbols appear close to the effects line representing the expected log-odds ratios under the model.

Outcome Predictions

The predicted value of Y for a given value of X (i.e., the estimate for $E[Y|X = x_i]$) is shown in the rightmost column of Table 10-3 and is denoted in Figure 10-3 by vertical arrows pointing downwards onto the horizontal axis.

Note that the predicted Y-value for *placebo* falls between *stationary* and *slight* improvement while for *test drug* it falls between *slight* and *moderate* improvement. The distance between the arrows associated with the *test drug* and *placebo*, representing the difference in the predicted outcome under the *test drug* vs. the *placebo*, is a measure of the amount of Y-Association in the data and also is an alternative to β as a measure of the treatment effect. The test for $\beta = 0$ given in Table 10-2, is equivalent to the test for $E(Y|X=1) = E(Y|X=0)$.

Figure 10-3
Effects Plot with Observed Log Odds Ratio Points, Effects Lines and Prediction Arrows



Under the model, the outcome prediction arrows tell us that patients who received the test drug are estimated to have an improvement score of 1.3 compared to 0.7 for those who received the placebo, reflecting an expected gain due to the test drug of 1.3-0.7=0.60 units on the improvement scale.

The slope of the *test drug* effects line, being a log-odds ratio, provides a visual measure of the *relative* effect analogous to the correlation or regression coefficient. The distance between the arrows provides a visual measure of the *absolute* effect, an improvement from less than *slight* (the predicted score expected under the *placebo* condition) to somewhat more than slight (the predicted score expected under the *test drug* condition). Hence, the effects plot in Figure 10-3 contains both relative and absolute measures of effect, which are assessed for statistical significance using the same test statistic.

Changing the Base for Calculating Odds

The treatment effect displayed as a slope in Figure 10-3 is defined relative to the (0,0) reference point (*stationary, placebo*). The corresponding odds ratio is defined as the ratio of the odds of improvement for patients who received the *test drug* to the odds of improvement for patients who received the *placebo*, where the odds of improvement is defined relative to the *stationary* outcome. We will now show how the graphical display changes based on a change in the reference point. First, we will change the Y-reference, then the X-reference.

In general, the Y-reference is defined by $Y = y_0$, where y_0 can be any number between the observed minimum (y_{min}) and observed maximum (y_{max}) outcome scores categories regardless of whether it coincides with a particular outcome category. For

example, taking $y_0 = \bar{y}$, the mean Y-score causes the "weighted average" to be used as the Y-reference which represents the weighted average of all the outcome categories.

$$\bar{y} = \sum_{j=1}^{5} p_j y_j = 0.98$$

where p_j = observed proportion of the patients having outcome j

Table 10-16, Table 10-17, Table 10-18 ,Table 10-19 and Table 10-20 provide estimated expected frequencies, odds, odds ratios, log-odds, log-odds ratios and Y*-scores computed when the Y-reference is changed to be the weighted average reference. Figure 10-4 contains the corresponding Effects plot. The modified Y-scores, $y^*_j = y_j - y_0$ are displayed below the corresponding category labels.

				(Y) OUTCOME				
		j =	**1**	**2**	**3**	**4**	**5**	**Y-ref.**
i	**x**	**X:TRTMNT**	**Worse**	**Stationary**	**Slight**	**Moderate**	**Marked**	$\hat{F}_{i0}$
1	**1**	**Test Drug**	1.58	12.78	15.10	15.12	7.41	11.73
2	**0**	**Placebo**	4.42	21.22	14.90	8.88	2.59	11.69
		Y^*-scores =	-1.98	-0.98	0.02	1.04	2.02	

Table 10-16
Expected Frequency Counts (F_{ij}) estimated under the equal adjacent category odds ratio model

Note: For each row i, $\hat{F}_{i0}$ is the geometric mean of the estimated expected frequencies in that row. $\hat{F}_{i0} = \prod_{j=1}^{J} \left(\hat{F}_{ij}\right)^{p_j}$

	Worse	**Stationary**	**Slight**	**Moderate**	**Marked**	**Y-reference**
Test Drug	0.13	1.09	1.29	1.29	0.63	1.0
Placebo	0.38	1.81	1.27	0.76	0.22	1.0

Table 10-17
Expected Odds for Clinical Trial Data; Reference Point: (OUTCOME = Weighted Average)

	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	0.36	0.60	1.01	1.70	2.86	1.0
Placebo	1.00	1.00	1.00	1.00	1.00	1.0

Table 10-18
Expected Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = Weighted Average, TRTMNT = 'Placebo')

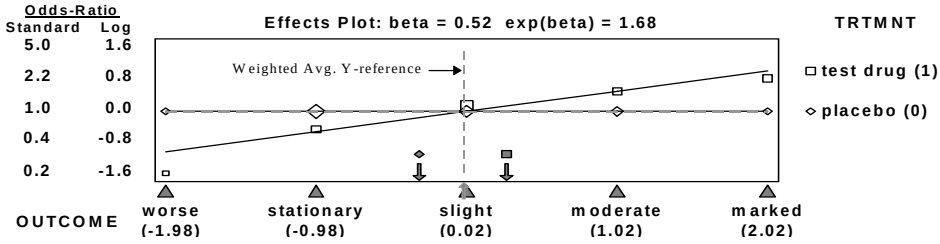
	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	-2.00	0.09	0.25	0.25	-0.46	0.0
Placebo	-0.97	0.59	0.24	-0.27	-1.51	0.0

Table 10-19
Expected Log-Odds for Clinical Trial Data; Reference Point: (OUTCOME = Weighted Average)

	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	-1.03	-0.51	0.01	0.53	1.05	0.0
Placebo	0.00	0.00	0.00	0.00	0.00	0.0

Table 10-20
Expected Log-Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = Weighted Average, TRTMNT = 'Placebo')

Figure 10-4
Effects Plot with Reference Points: (OUTCOME = Weighted Average, TRTMNT = 'Placebo')



The dashed vertical line representing the Y-reference now extends from the weighted average arrow on the plot. The odds of marked improvement (vs. the "average" outcome) are exp(2.02 beta) = 2.86 times as high for patients who received the test drug than for those receiving the placebo. See Table 10-17 for corresponding odds ratios.

Changing the Baseline for Calculating Odds Ratios

Similar to use of y_0 to define the Y-reference, the X-reference is defined by some number $X = x_0$, between the minimum and maximum observed X-score regardless of whether the value $X = x_0$ is observed in the sample.

We will now change the X-reference to represent the weighted average of the TRTMNT categories (i.e., $x_0 = \bar{x} = 0.5$). Table 10-21, Table 10-22, Table 10-23, Table 10-24 and Table 10-25 provide estimated expected frequencies, odds, odds ratios, log-odds and log-odds ratios computed when the X-reference and Y-reference are defined as weighted averages of the row and column categories. Figure 10-5 displays the corresponding Effects plot. Note that modified X-scores, $x^*_j = x_j - x_0$ replace the original X-scores.

				(Y) OUTCOME				
			j =	**1**	**2**	**3**	**4**	**5**	**Y-ref.**
i	**x^*-scores**	**X:TRTMNT**		**Worse**	**Stationary**	**Slight**	**Moderate**	**Marked**	$\hat{F}_{0j}$
1	**1**	**Test Drug**		1.58	12.78	15.10	15.12	7.41	11.73
2	**0**	**Placebo**		4.42	21.22	14.90	8.88	2.59	11.69
		Y^*-scores =		-1.98	-0.98	0.02	1.04	2.02	
		X-reference		2.64	16.47	15.0	11.59	4.38	

Table 10-21
Expected Frequency Counts (F_{ij}) estimated under the equal adjacent category odds ratio model

Note: For each column j, $\hat{F}_{0i}$ is the geometric mean of the estimated expected frequencies for that column $\hat{F}_{0j} = \prod_{i=1}^{I} \left(\hat{F}_{ij}\right)^{p_i}$

	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	0.13	1.09	1.29	1.29	0.63	1.0
Placebo	0.38	1.81	1.27	0.76	0.22	1.0
X-Reference	0.23	1.41	1.28	0.99	0.37	

Table 10-22
Expected Odds for Clinical Trial Data; Reference Point: (OUTCOME = Weighted Average)

	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	0.60	0.78	1.00	1.30	1.69	1.0
Placebo	1.67	1.29	1.00	0.77	0.59	1.0
X-Reference	1.0	1.0	1.0	1.0	1.0	

Table 10-23
Expected Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = Weighted Average, TRTMNT = Weighted Average)

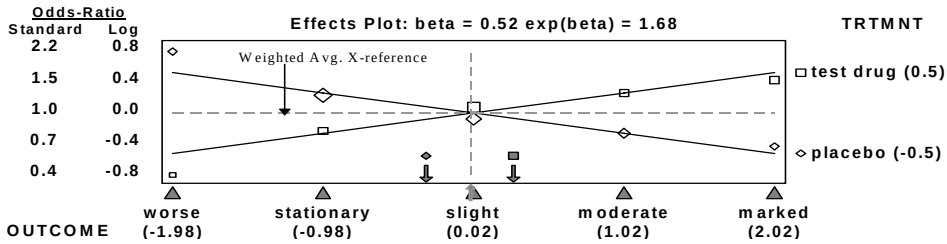
	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	-2.00	0.09	0.25	0.25	-0.46	0.0
Placebo	-0.97	0.59	0.24	-0.27	-1.51	0.0
X-Reference	-1.49	0.34	0.25	-0.01	-0.98	

Table 10-24
Expected Log-Odds for Clinical Trial Data; Reference Point: (OUTCOME = Weighted Average)

	Worse	Stationary	Slight	Moderate	Marked	Y-reference
Test Drug	-0.51	-0.25	0.01	0.26	0.52	0.0
Placebo	0.51	0.25	-0.01	-0.26	-0.52	0.0
X-Reference	0.0	0.0	0.0	0.0	0.0	

Table 10-25
Expected Log-Odds Ratios for Clinical Trial Data; Reference Points: (OUTCOME = Weighted Average, TRTMNT = Weighted Average)

Figure 10-5
Y-View Effects Plot - - Weighted Average Reference Points



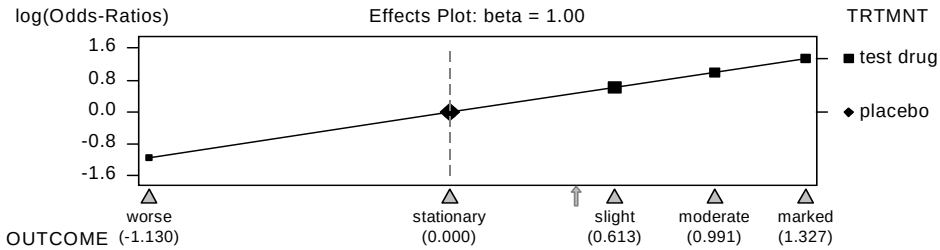
The dashed horizontal line representing the X-reference is now positioned at the "average". The odds of marked improvement (vs. the "average" outcome) are exp[(0.5)(2.02)beta] = 1.69 times as high for patients who received the test drug than for the average patient. See Table 10-23 for corresponding odds ratios.

It is important to note that changing the reference point does not change the model (the Beta and predicted Y-scores remain the same – see "A.13 Monotonic Regression Model is Invariant of the Reference Point (x0, y0)" on page 233). Changing the reference only alters the perspective from which you view the Effects plot (i.e., the origin (x_0, y_0)), which serves as the base for the calculation of the odds ratios. See Chapter 9 for a different change in reference and the resulting change in the category-specific beta parameter estimates.

The Estimated Y-Score Model

In the case where there are only two X-profiles, Y-scores can always be estimated to yield a perfect fit to the data. However, the increase in fit comes at the expense of the estimation of J-2 = 3 distinct Y-score parameters. For example, using the observed log-odds ratios for *worse* and *slight* in Table 10-5 to fix Y-scores for the first two outcome categories, estimation of the others will match perfectly the other observed log-odds ratios and yields Figure 10-6 with *Beta* = 1. (Two identifying restrictions need to be imposed on the Y-scores -- recall the discussion for Model B in Chapter 7).

Figure 10-6
The Estimated Y-Score Model



However, since the additional amount of association that is explained is small (L^2[Residual] = 0.5 in Table 10-2) and costs 3 degrees of freedom, the gain is not significant with the current data. Hence, we conclude that the estimated Y-score model overfits these data. A model that overfits data is less likely to validate on other data sets and causes the estimate of β to be overstated. Hence, a more parsimonious model should always be selected over one that provides an overfit. For the current example, the equidistant Y-score model provides a reasonable choice for Y-scores. See Chapter 4 for an example of alternative Y-scores that are justified on theoretical grounds.

Analysis of Association	**Hypothesis Tested**	**L^2**	**d.f.**	**p-value**
Y-Association	Independence	8.1	4	0.09
Residual Association	None	0.0	0	---
Total Association	Independence	8.1	4	0.09

Table 10-26
Analysis of Association for the Estimated Y-Score Model

Users are referred to the other examples in this manual and to the references for proper interpretation of results from models estimated on data containing more than two X-profiles.

1

Example 5
An Example from Discriminant Analysis

When the goal is to classify individuals or objects into one of three or more groups, the model used in GOLDMineR offers advantages over the traditional approach used in discriminant analysis if the grouping variable is ordinal. As an example, we reanalyze the classic iris data set of Fisher (1936) which is frequently used to introduce the topic of discriminant analysis.

In this example, four centimeter measurements – petal length (PETLEN), petal width (PETW), sepal length (SEPLEN), and sepal width (SEPW) are used to classify 150 iris flowers into one of three SPECIES – Setosa, Virginica and Versicolor. Fisher (1936, p.185) alludes to his belief that SPECIES is an ordinal variable by noting that Versicolor is a hybrid of the other two species:

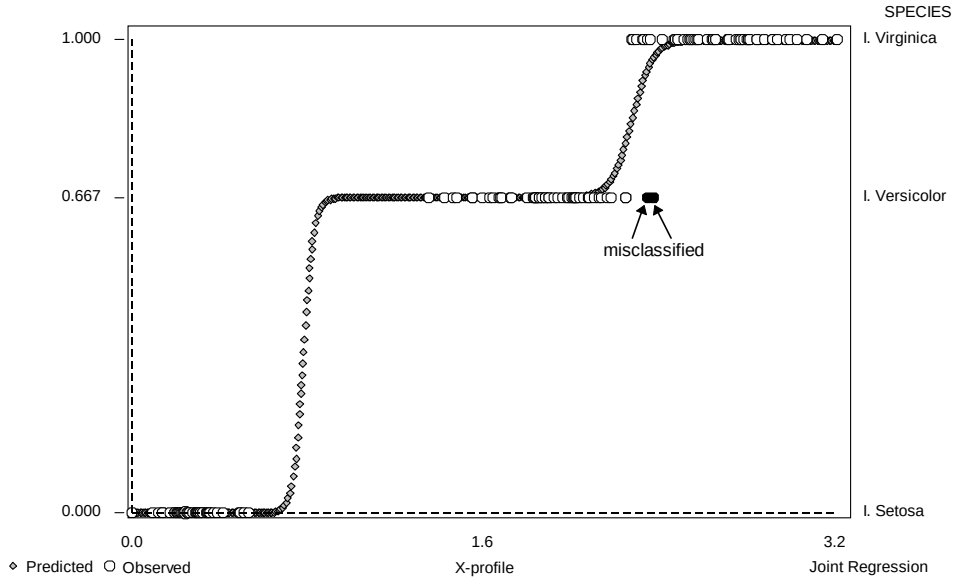
> We shall, therefore, consider whether, when we use the linear compound of the four measurements most appropriate for discriminating three such species, the mean value for I. Versicolor takes an intermediate value, and, if so, whether it differs twice as much from I. Setosa as from I. Virginica, as might be expected, if the effects of genes are simply additive, in a hybrid between a diploid and a tetraploid species.

For our analysis, we assigned scores to order the specimens in a manner consistent with the assumed hybrid theory – '0' for Setosa, '1' for Virginica and '2/3' for Versicolor. Since the results from model estimation indicated that neither predictor SEPW nor SEPLEN was significant when both were included in the model, SEPW was omitted from our final model. (Fisher noted that SEPW differed from the other variables in that its mean among the hybrid (Versicolor) specimens was *not* intermediate between the means among the other 2 species.)

Figure 1-1 compares the monotonic regression predictions (plotted using the diamond symbols), with the actual species scores of 0, 0.667 and 1 (plotted using circles). For these data, the monotonic regression curve looks very much like a step function with almost perfect prediction ($R^2 = 0.99$ as shown in Figure 1-2). An

excellent fit to the data is evidenced by the fact that only 2 of the 150 specimens – the 2 represented by shaded circles -- differ significantly from the model predictions.

Figure 1-1:Joint Regression Plot for Fisher Data



Results from discriminant analysis are often assessed by the number of misclassifications. Since the model predicts a score of 0.000 for all 50 Setosa specimens, all of these specimen are classified correctly by our model. Discrimination between Versicolor and Virginica is somewhat more difficult. Use of predicted score 0.8 as a cutoff point to distinguish between these two species results in the 2 specimens represented by shaded circles being the *only* ones misclassified. However, use of 0.8 can be criticized on grounds that it was selected after looking at the data.

A more appropriate way to classify the specimens that does not capitalize on chance is to base the classifications on the predicted probabilities obtained from the Table window (Expected Probability) for the estimated model. Table 1-1 contains predicted probabilities for the 13 specimens most difficult to classify correctly; those having a predicted score between 0.70 and 0.99. We classify each specimen into that species having the highest predicted probability of membership. The result shown in Table 1-1 is that only 3 specimens are misclassified -- an improvement over discriminant analysis where 4 specimens would be misclassified if a comparable analysis were undertaken.

PETLEN	PETW	SEPLEN	predicted score	Predicted Probability for: Setosa	Versicolor	Virginica	Actual Species	Classified Correctly?
5.1	2.0	6.5	0.99	0.00	0.02	0.98	Virginica	Yes
5,1	1.9	6.3	0.98	0.00	0.05	0.95	Virginica	Yes
5.1	1.6	6,6	0.96	0.00	0.11	0.89	Versicolor	No
4.9	1.8	6.1	0.95	0.00	0.15	0.85	Virginica	Yes
4.8	1.8	5.9	0.94	0.00	0.18	0.82	Versicolor	No
4.5	1.7	4.9	0.94	0.00	0.19	0.81	Virginica	Yes
4.8	1.8	6.0	0.92	0.00	0.24	0.76	Virginica	Yes
4.9	1.8	6.3	0.91	0.00	0.28	0.72	Virginica	Yes
4.8	1.8	6.2	0.86	0.00	0.41	0.59	Virginica	Yes
5.0	1.5	6.0	0.84	0.00	0.48	0.52	Virginica	Yes
5.1	1.5	6.3	0.82	0.00	0.53	0.47	Virginica	No
5.0	1.7	6.7	0.78	0.00	0.65	0.35	Versicolor	Yes
4.9	1.5	6.3	0.70	0.00	0.89	0.11	Versicolor	Yes

Table 1-1:An Illustration of Classification Based on the Highest Predicted Probability

Figure 1-2 provides the parameter estimates and model summary statistics for our final model. The fact that the beta estimates for PETLEN and PETW are approximately equal (28.8 vs. 29.9) suggests that these variables can be added together so that only a single petal measurement, the sum of Petal length and Petal width (PETLENW = PETLEN + PETW), is needed along with SEPLEN to provide good classification. Figure 1-3 shows how the three species are differentiated by plotting PETLENW by SEPLEN.

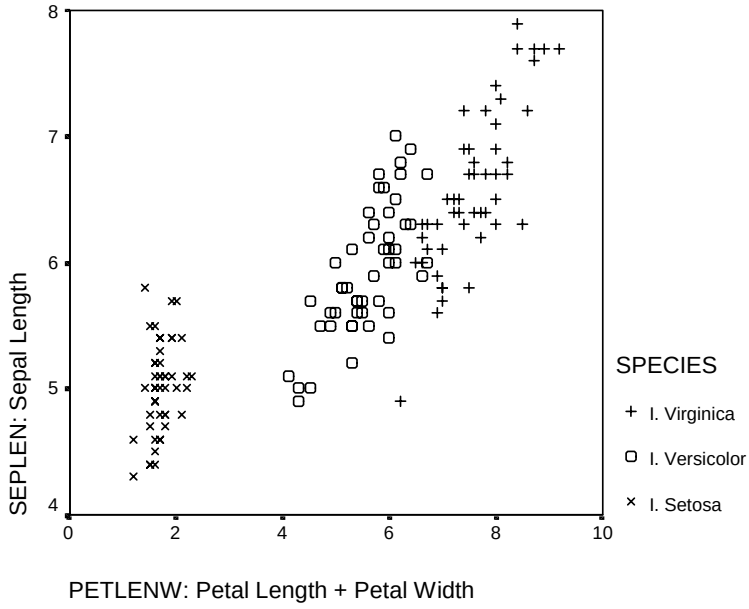
Figure 1-2:Partial listing of Statistics window

y: SPECIES (Fixed)	I. Setosa	I. Versicolor	I. Virginica					
Y-scores	0	0.667	1					
	$L^2(Y)$	df	p-value	Beta				
PETLEN (Fixed)	17.85	1	2.40E-05	28.8				
PETW (Fixed)	8.36	1	0.0039	29.9				
SEPLEN (Fixed)	5.07	1	0.024	-11.4				
Association Summary	L^2	df	p-value	R^2	phi	Pearson X^2	p-value	
Explained by Model	314.10	3	8.80E-68	0.99	27.00			
Residual	15.49	281	1			17.03	1	
Total	329.60	284	0.032			300	0.25	

Use of an alternative model where the scores for SPECIES are estimated rather than fixed provides similar results as those reported in this chapter with the exception that

the estimated scores turns out to be 0.00, 0.50, 1.00 for Setosa, Versicolor, and Virginica respectively.

Figure 1-3: SPSS Scatterplot



The number of iterations required to achieve convergence is high because it is very easy to discriminate perfectly between Setosa and the other species, thus resulting in an R^2 which approaches 1. The number of iterations can be reduced substantially by omitting the Setosa species from the analysis. For all practical purposes, the resulting model is identical.

.

K

Appendix

A.1 The Monotonic Regression Model

Given quantitative values y_1, y_2, …, y_J (Y-scores) assigned to the J different outcomes associated with an ordinal dependent variable Y, and given M qualitative and/or quantitative predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_M)$, following Magidson (1995) the monotonic regression of Y on **X** is:

$$E(Y|\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{J} P_{j.x}\, y_j \quad \text{Monotonic Regression Model} \qquad \text{Equation K-1}$$

where $P_{j.\mathbf{x}}$, the conditional probability of the occurrence of $Y=y_j$ given $\mathbf{X}=\mathbf{x}$, is estimated from a sample of N observations using the following 2-step process:

1) Conditional logits $\Psi_{j.\mathbf{x}}$ (defined below in Equation K-8) are predicted using Equation K-3

2) the predicted logits are transformed to predicted probabilities using the identity:

$$P_{j.x} \equiv \exp\left(\Psi_{j.x}\right) \Big/ \sum_{j=1}^{J} \exp\left(\Psi_{j.x}\right) \qquad \text{Equation K-2}$$

The Y-scores, which determine the ordering and relative spacing of the J outcomes, may be specified by the user. In the case that one or more of the Y-scores are unknown or otherwise unspecified prior to model estimation, they are treated as model parameters and estimated simultaneously with the other parameters. For a traditional logit analysis where J=2, users will typically set $y_1=1$ and $y_2=0$.

The model used to predict the generalized logits, $\Psi_{j.x^*}$, is:

$$\Psi_{j.x} = \alpha_j + \left(\beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_M x_M^*\right) y_j^* \quad j = 1, 2, \ldots, J \quad \text{Generalized Logit Model} \qquad \text{Equation K-3}$$

where the "*relative* Y-score" y_j^*, is the difference between y_j and some Y-reference score y_0 defined as some weighted average of the original Y-scores:

$$y_j^* = y_j - y_0; \qquad \text{where } y_0 = \sum_{j=1}^{J} e_j y_j \text{ for } e_j \geq 0, \quad j = 1,2,...,J \text{ and } \sum_{j=1}^{J} e_j = 1 \qquad \text{Equation K-4}$$

$$\text{Note that Equation K-4 implies } \sum_{j=1}^{J} e_j y_j^* = 0 \qquad \text{Equation K-5}$$

and for the m^{th} predictor variable, the relative X-score, x_m^*, is the difference between x_m and some X_m-reference score x_{m0} defined in Equation K-16:

$$x_m^* = x_m - x_{m0} \qquad m = 1, 2, \ldots, M \qquad \text{Equation K-6}$$

If the m^{th} predictor is quantitative, values for X_m are either read from the data file or input by the user as **Fixed** scores. If the m^{th} predictor is qualitative, scores for each category of the predictor X_{mk}, k = 1, 2, …, K_m are treated as **Free** parameters and estimated simultaneously with the other model parameters. Also, if the m^{th} predictor is qualitative, in addition to the estimation of β_m in Equation K-3, category-specific beta parameters, β_{mk}, k = 1, 2, …, K_m are estimated according to the contrast coding option selected.

The parameters in the generalized logit model (Equation K-3) require standardizing restrictions in order to be uniquely identified. Sections “A.3 Estimating X-Scores for Qualitative Predictor Variables” on p. 214, “A.5 Model Identification Issues” on p. 218 and “A.6 Nonuniqueness of Monotonic Regression” on p. 221 describe the standardizing restrictions used in GOLDMineR.

A.2 The Generalized Logit Model

A logit, also known as a "log-odds" is the natural logarithm of an odds. If Y is dichotomous, the simple odds in favor of the first outcome (j = 1) is P_1/P_2 or equivalently $P_1/(1-P_1)$ where P_j is the probability of the jth outcome occurring. For J ≥ 2, the generalized odds in favor of the jth outcome is defined as P_j / P_0 where P_j is the probability of occurrence of the jth outcome and P_0 is a reference probability. P_0 is determined by the user-defined category weights (E-weights) also used to define the reference score y_0:

$$P_0 = \prod_{j=1}^{J} (P_j)^{e_j} \qquad \text{Equation K-7}$$

For example, when J = 2, and $E_0 \equiv (e_1,e_2) = (0,1)$, the generalized odds in favor of the first outcome is the simple odds P_1/P_2 and the corresponding logit is $\ln(P_1/P_2)$.

Alternatively, for $(E_0) = (p_1,p_2)$ where p_j is the proportion of observations for which the jth outcome occurs, P_0 is the geometric mean of the outcome probabilities.

For a given **X**=**x**, the generalized logit in Equation K-3 is defined as

$$\Psi_{j.\mathbf{x}} \equiv \ln(P_{j\cdot\mathbf{x}} / P_{0\cdot\mathbf{x}})$$ Equation K-8

where $P_{j\cdot\mathbf{x}}$ is the conditional probability of the jth outcome occurring when **X**=**x**

and $P_{0.x} = \prod_{j=1}^{J}(P_{j.x})^{e_j}$ Equation K-9

Formulation of Equation K-3 in terms of the relative scores x* and y* rather than x and y allow the α-parameters to take on the interpretations of meaningful baseline logits:

$$\alpha_j = \Psi_{j.x_0}$$ Equation K-10

where $\mathbf{x}_0 = (x_{10}, x_{20},\ldots,x_{M0})$

The Generalized Logit Model (Equation K-3) can also be expressed as

$$\Psi_{j.x[i]} = \alpha_j + \left(\beta_1 x^*_{1[i]} + \beta_2 x^*_{2[i]} + \cdots + \beta_M x^*_{M[i]}\right) y^*_j \qquad j = 1, 2, \ldots, J$$ Equation K-11

where i indexes the observed X-profiles **X**[i] = **x**[i] i = 1, 2, …, I

In the case that Y and/or at least one X variable is continuous, I will equal the number of observations in the sample. In this case, the definition of the generalized logit given in Equation K-8 is replaced by

$$\Psi_{j.\mathbf{x}} \equiv \ln[f_{Y,\mathbf{X}}(y,\mathbf{x})/f_{Y,X}(y_0,\mathbf{x})]$$ Equation K-12

where $f_{Y,\mathbf{X}}$ denotes the joint density function. Under the usual distributional assumptions that justify the linear regression, it can be shown that the β_m -parameters in Equation K-11 are proportional to the regression coefficients β^*_m in the usual linear regression model:

$$Y = \alpha^* + \beta^*_1 X_1 + \beta^*_2 X_2 + \ldots \beta^*_M X_M + \varepsilon.$$ Equation K-13

The exact relationship between the parameters in this case is $\beta^*_m = \sigma^2\beta_m$, where σ^2 is the variance of the error term ε. (See the section “A.14 Analysis of a Continuous Dependent Variable” on p. 235 for further discussion of the relationship to the linear regression model.)

A.3 Estimating X-Scores for Qualitative Predictor Variables

Estimation of X-Scores and selection of contrast coding options for a qualitative predictor are accomplished in GOLDMineR by using the **Free** scaling option for that predictor and specifying a set of category weights (C-weights).

Similar to the E-weights in Equation K-4 and Equation K-5, for the mth predictor, the C-weights, $C_{m0} = (c_{m1}, c_{m2}, \ldots, c_{mKm})$ are nonnegative numbers that sum to one:

$$\sum_{k=1}^{K_m} c_{mk} = 1 \qquad c_{mk} > 0 \qquad k = 1, 2, \ldots, K_m \qquad \text{Equation K-14}$$

Category specific betas satisfy the following identifying restriction:

$$\sum_{k=1}^{K_m} c_{mk} \hat{\beta}_{mk} = 0 \qquad \text{Equation K-15}$$

Hence, there are K_m-1 degrees of freedom used for the test of significance for the contribution of the m^{th} predictor if qualitative, and K_m-1 columns will be included in the design matrix (see section "A.11 Design Matrix" on p. 231).

By default, dummy contrast codes are used by GOLDMineR for qualitative predictors, with category 1 serving as the base reference. For example, when the predictor variable PAYMENT in sample dataset #3 is treated as qualitative in Figure 9-3 and Figure 9-4, the default dummy contrast codes were specified through use of the C-weights $C_1 = (1,0,0,0)$. Hence, the parameter estimate for each category k = 1, 2, 3, 4 is interpretable as the effect of that category relative to the base category k=1. A parameter estimate for the base category (zero, by definition) is also provided in the output in Figure 9-3 (i.e., "Beta(1) = 0").

Figure 9-9 and Figure 9-10 showed how the results changed when the dummy contrast $C_1 = (0,0,1,0)$ is used. In this case, the revised category-specific parameter estimate for category k takes on the interpretation as the effect of that category vs. the new base category k=3. Hence, Figure 9-9 now lists Beta(3) = 0.

ANOVA-type contrast coding based on unweighted means (i.e., "effect coding") is accomplished in GOLDMineR by changing the dummy contrast to an unweighted contrast. Use of the **unweighted** option for the qualitative predictor PAYMENT in example 3 changes the C-weights to $C_1 = (.25, .25, .25, .25)$.

In general, GOLDMineR uses Equation K-15 to compute the estimate of Beta(k) for the qualitative predictor category omitted from the design matrix. For effect coding, Equation K-15 implies that the (unweighted) sum of the Beta(k), k =1,2,3,4 equals 0 so that the Beta(k) associated with the omitted category is computed by GOLDMineR

as the negative of the sum of the Beta(k) estimates for the other categories. For more detail and sample design matrices, see section "A.11 Design Matrix" on p. 231. For more information on the contrast coding features available in GOLDMineR see "Category Weights" on page 119.

The X-scores associated with a qualitative predictor with K categories are estimated subject to the identifying restrictions that the lowest estimated score is 0 and the highest 1. (Hence, if a dichotomous predictor is designated as qualitative, the X-scores 0 and 1 will be used.) In addition, the reference score x_0 in Equation K-6 is determined for a GOLDMineR analysis based on the current setting for $C = (c_1, c_2, \ldots, c_K)$:

$$x_0 = \sum_{k=1}^{K} c_k x_k \qquad \text{Equation K-16}$$

For examples of the X-scores that are obtained for different contrast coding selections, see Figure 9-3, Figure 9-7 and Figure 9-9.

Note that Equation K-11 remains true if β_m is replaced by $-\beta_m$ and estimated X-scores are mulitplied by -1. Alternatively, instead of being mulitplied by -1, negative X-scores could be replaced by nonnegative scores using the two step transformation $x'_{mk} = 1 - x_{mk}$, k = 1, 2, …, K as illustrated by the table below.

k	**x*k**	**-x*k**	**x'*k**
1	0.0	0.0	1.0
2	0.2	-0.2	0.8
3	0.6	-0.6	0.4
4	1.0	-1.0	0.0

In the first step Equation K-11 remains true because each category score x_k is multiplied by -1 (to obtain the scores in column 2 above) and the associated β is also multiplied by -1. During the first step the range of X-scores changes from [0,1] to [-1,0]. In the second step the dummy contrast is changed from C = (1,0,0,0) to C' = (0,0,0,1) so that the minimum score -1 is subtracted from each score to yield the scores in column 3 above. The end result is that the first category which originally is assigned the lowest score (0), becomes the category with the highest score (1) and vice versa.

To establish uniqueness for the estimated X-scores, an identifying restriction is required to eliminate the choice between the scores in columns 1 and 2 above. GOLDMineR resolves this identification problem by selecting those scores (x_m or $-x_m$) that yield a nonnegative $\hat{\beta}_m$.

In addition to the X^* scores for quantitative and qualitative predictor variables, GOLDMineR also displays a standardized version of the X-scores, called U-scores, which are related to the scores for the row variable used in Goodman's column effects association model. For further information on U-scores, see the section "Standardized Form of Model" on p. 219.

A.4 Monotonicity

The term "monotonic regression" comes from the monotonicity properties (Equation K-17 and Equation K-18) possessed by the regression model (Equation K-1, Equation K-2 and Equation K-3):

$$\frac{\partial}{\partial X_m} E(Y|\mathbf{X} = \mathbf{x}) = \beta_m \, VAR(Y|\mathbf{X}=\mathbf{x}) \qquad \text{Equation K-17}$$

$$\text{and } \frac{\partial}{\partial \sum_{m=1}^{M} \beta_m X_m^*} E(Y|\mathbf{X} = \mathbf{x}) = VAR(Y|\mathbf{X}=\mathbf{x}) > 0 \qquad \text{Equation K-18}$$

Since $VAR(Y|\mathbf{X}=\mathbf{x}) > 0$ for any $\mathbf{x}$, property (Equation K-17) states that as the m^{th} predictor is increased (while the other predictors are held constant), the predicted value for Y increases, decreases or remains constant depending upon whether $\hat{\beta}_m > 0$, $\hat{\beta}_m < 0$, or $\hat{\beta}_m = 0$, respectively. That is, the prediction of Y as a function of predictor X_m is monotonic for each m = 1, 2, …, M.

Proof of Equation K-17:

Let $f(\mathbf{x},y_j) = \exp(\alpha_j + \sum_{m=1}^{M} \beta_m X_m^* y_j)$

$$\frac{\partial}{\partial X_m} P_{j.\mathbf{x}} = \frac{\partial}{\partial X_m^*} P_{j.\mathbf{x}} = \frac{\partial}{\partial X_m^*} f\left(x, y_j\right) \Big/ \sum_{j=1}^{J} f\left(x, y_j\right)$$

$$= \beta_m y_j P_{j.\mathbf{x}} - \beta_m \, f\left(x,y_j\right) \sum_{j=1}^{J} y_j \, f\left(x,y_j\right) / \left[\sum_{j=1}^{J} f\left(x,y_j\right)\right]^2$$

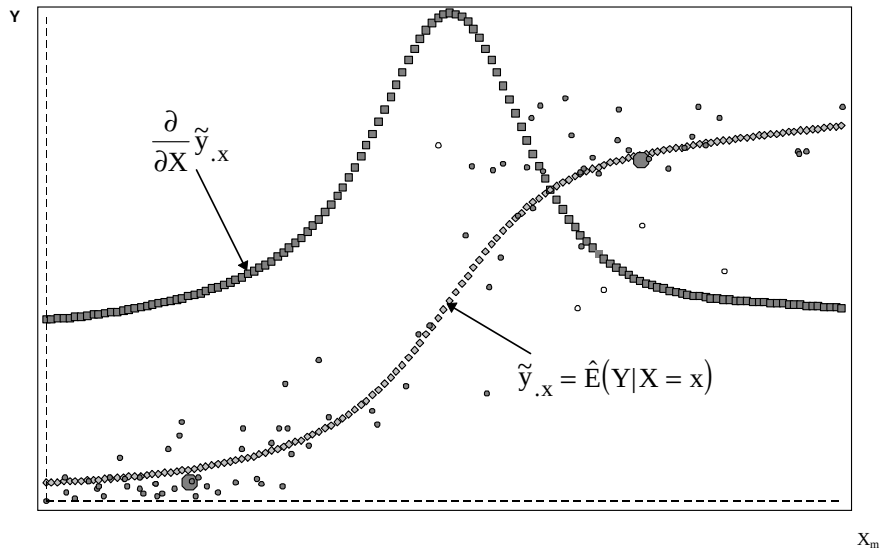
$$= \beta_m y_j P_{j.\mathbf{x}} - \beta_m P_{j.\mathbf{x}} \sum_{j=1}^{J} y_j P_{j.x}$$

$$\frac{\partial}{\partial X_m} E(Y|\mathbf{X}) = \sum_{j=1}^{J} \frac{\partial}{\partial X_m} P_{j.\mathbf{x}}\, y_j = \beta_m \left[\sum_{j=1}^{J} y_j^2 P_{j.x} - \left(\sum_{j=1}^{J} y_j P_{j.x} \right)^2 \right]$$

$$= \beta_m \, VAR(Y|\mathbf{X}) \qquad \text{Equation K-19}$$

Figure K-1 is an example of a partial regression curve for a single predictor with $\hat{\beta}_m > 0$ and the always positive derivative as displayed in GOLDMineR. (The derivative curve that is displayed in the regression plots is scaled in Figure K-1 by subtracting the mean Y-score so as to fit in the same Plot window as the regression curve.)

Figure K-1
Partial Regression Curve and Derivative



One of the usual assumptions made with the linear regression model (Equation K-13) is homoscedasticity, which states that $VAR(Y|\mathbf{X}=\mathbf{x}) = \sigma^2$, where σ^2 represents the unknown constant error variance that is assumed. When the usual linear regression assumptions are all true, the β–parameters in the monotonic model turn out to be exactly proportional to the corresponding β* parameters in the linear model. The constant of proportionality is σ^2 (see Equation K-48). The monotonic model however is more general than the linear model and neither requires an assumption of

homoscedasticity (constant error variance) nor any specific assumption about the form that heteroscedasticity may take.

A.5 Model Identification Issues

Uniqueness of Y-score parameters

In the case that all J of the Y-scores are treated as parameters and estimated, note that y_j^* in Equation K-3 may be replaced by any $y_j' = b\,y_j^*$ for some $b \neq 0$ and Equation K-3 will still hold true if the β parameters in this case are replaced by $\beta_m' = \beta_m / b$.

In addition, note that y_j^* may be replaced by $y_j' = y_j^* + a$ and Equation K-3 will still hold true if α_j is replaced by $\alpha_j' = \alpha_j + a \sum_{m=1}^{M} \beta_m x_m^*$. Hence, in order that the estimated Y-scores be unique, two identifying restrictions must be imposed on the Y-score estimates.

By default, and without loss of generality we identify the scale of the estimated Y-scores using the same restrictions used to identify the X-scores for a qualitative predictor:

- the minimum estimated Y-score is 0
- the maximum estimated Y-score is 1

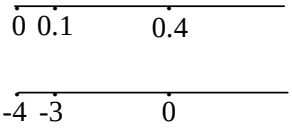
In addition, GOLDMineR also computes a standardized form of Y-scores called V-scores, which are related to the scores for the column variable used in Goodman's row effects association model. For further information on V-scores, see “Standardized Form of Model” on p. 219.

As a more general alternative to the two identifying restrictions given above, users may select any two real numbers that differ in value, assign them as scores for any two categories of Y and estimate the remaining scores. (For Model B of Example 1, the scores 0 and 1 were assigned to the outer categories and the middle score was estimated.) Assigning different scores to exactly two categories and estimating the others or estimating all scores subject to the restrictions that the low score is 0 and the high score is 1 will yield Y-equivalent scores y_j' and y_j'', j = 1, 2, …, J. Y-equivalent scores yield identical estimates for all the model parameters and associated statistics.

Formally, any two sets of Y-scores, y' and y'' are said to be Y-equivalent if there exists constants a and $b \neq 0$ such that $y_j'' = a + b\,y_j'$ j = 1, 2, …, J. Y-equivalent scores preserve the relative distances between each of the outcome categories. For example,

it is easy to check that the relative distance between outcome categories 1, 2 and 3 is preserved by the Y-equivalent scores shown below:

$$\frac{y_2' - y_1'}{y_3' - y_1'} = \frac{y_2'' - y_1''}{y_3'' - y_1''} = 0.25$$



Uniqueness of Alpha parameters

Without loss of generality, the following restriction is used to identify the α-parameters:

$$\sum_{j=1}^{J} e_j \alpha_j = 0 \qquad \text{Equation K-20}$$

Standardized Form of Model

The Generalized Logit Model (Equation K-3) can be expressed in the form of a weighted RC association model (Goodman, 1991) using u_m to denote the standardized version of predictor variable X_m and v_j to denote the standardized score for the j^{th} outcome category:

$$\Phi_{j.\mathbf{x}} = \phi(u_1 + u_2 + \ldots + u_M)\, v_j \qquad \text{Equation K-21}$$

where

$\Phi_{j.\mathbf{x}} = \Psi_{j.\mathbf{x}} - \alpha_j$ represents the generalized log-odds ratio Equation K-22

$v_j = y_j^* / S_Y$, where S_Y = standard deviation of the Y-scores Equation K-23

$u_m = \hat{\beta}_m x_m^* / S_{\mathbf{X}\beta}$, where $S_{\mathbf{X}\beta}$ = standard deviation of $\sum_{m=1}^{M} \hat{\beta}_m X_m$ Equation K-24

and $|\phi| = S_Y S_{\mathbf{X}\beta}$ Equation K-25

While the particular value of the log-odds ratio $\Phi_{j.\mathbf{x}}$ depends upon the reference point $(y_0, \mathbf{x}_0)$, ϕ is invariant of the reference point and is invariant of Y-equivalent scores. For example, if $y_j'' = 2y_j'$, it follows that $S_{y''} = 2S_{y'}$ and $\beta'' = \frac{1}{2}\beta'$, so that $S_{x\beta''} = \frac{1}{2} S_{x\beta'}$ and from Equation K-25 it follows that ϕ is unchanged. For insight into the meaning of ϕ, readers are referred to the discussion of the weighted association model in Goodman (1991), Goodman (1981) and section"A.14 Analysis of a Continuous Dependent Variable" on p. 235.

Note that in the case where M=1 (a single predictor variable) and $\hat{\beta} \neq 0$, defining $u \equiv u_1$, and $x^* \equiv x_1^*$, we have $u = x^*/S_X$ so that u (and v) represent generalizations of the usual standardized forms for x and y respectively, i.e., VAR(u) = VAR(v) = 1. When the C- and E-weights equal the corresponding marginal proportions, u and v are z-scores. When $\hat{\beta}_m = 0$, $u_m = 0$.

$U = u_1 + u_2 + \ldots + u_M$ is used to position the joint X-variable along the horizontal axis of the joint regression plot and may be obtained by selecting the "Show X-scores" option in the specialized charts table control box. In the degenerate case that each $\hat{\beta}_m = 0$, m = 1, 2, …, M, $S_{x\beta} = 0$. In this case, GOLDMineR sets each $u_m = 0$.

Identification of ϕ

ϕ in Equation K-21 is a measure of the strength of association between Y and **X** (i.e., between v and $u = (u_1, u_2, \ldots, u_M)$. Without loss of generality, ϕ is taken to always be positive to eliminate a basic indeterminacy in the model. To illustrate the indeterminacy, consider the simple case where M=1. From Equation K-21, Equation K-23 and Equation K-24 we have

$\Phi_{j.x} = \phi \hat{\beta} x^* y_j^* / S_Y S_{\mathbf{X}\beta}$ Equation K-26

For given values of x^* and y^*, ϕ and $\hat{\beta}$ are involved in an indeterminacy since they can be replaced by $-\phi$ and $-\hat{\beta}$ respectively without altering the equation. In the case of 1 predictor, in terms of population parameters, we have

$$\beta = \frac{\phi}{\sigma_Y \sigma_X}$$ Equation K-27

An alternate expression for Equation K-21 in terms of the **X**-profiles is the score for the joint X-variable (joint X-score)

$$\Phi_{j.\mathbf{x}(i)} = \phi U_i V_j$$ Equation K-28

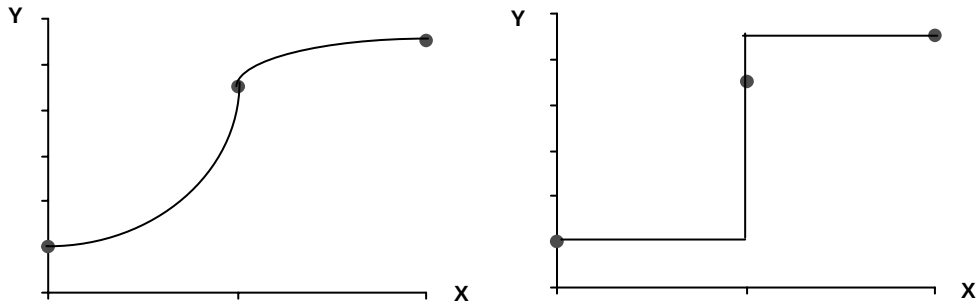
$$\text{where } U_i = \sum_{i=1}^{I} u_{mi}$$ Equation K-29

A.6 Nonuniqueness of Monotonic Regression

A second indeterminacy involving $\phi\beta$ occurs in cases where a monotonic regression of Y on **X** provides perfect predictions for Y. In this case $\phi\beta$ is not estimable, $\hat{\phi}$ approaches infinity, and hence the estimation algorithm fails to converge.

For purposes of illustration, consider a simple example where N=3 and fixed scores are given for X and Y. In Figure K-2 the graph on the left illustrates the situation where a monotonic curve can provide perfect prediction of Y. The graph on the right provides another monotonic curve that also provides perfect prediction. It is clear that if it is possible to find one monotonic regression curve that provides perfect prediction of Y, it is possible to find an infinite number of different monotonic regression curves that also provide perfect predictions for Y.

Figure K-2



The indeterminacy associated with a perfect monotonic relationship between Y and **X** occurs if Y is perfectly predicted from any X_m. Hence, prior to beginning the estimation, GOLDMineR initiates a check on each predictor by dependent variable table to determine whether a perfect monotonic relationship exists. If such is found (see Case A below), GOLDMineR displays the message "predictor X_m is monotonic" (with respect to Y) and fails to begin to estimate the parameters. In this case, the users should investigate the reason that the predictor can provide perfect prediction of the dependent variable and either omit the predictor variable or take some other appropriate action.

After defining a model and clicking on Estimate, GOLDminer also performs additional identification checks on the model parameters which depend upon the scaling options in place (see Cases B, C and D below).

Case A

Both the dependent variable and predictor are treated as quantitative variables (using **Uniform** or other **Fixed** scores). Cases are ordered from low to high on the X-score and examined to determine whether there is a perfect ordering on the Y-score. If a perfect ordering is found, the message "Predictor <predictor name> is monotonic" is displayed.

Case B

Predictor is treated as qualitative (**Free**) or is dichotomous, but dependent variable has more than 2 categories and is treated as having **Fixed** scores (**Uniform** or custom **Fixed** scores)

For the kth predictor category, k = 1,2,...,K, check row k to see if the number of nonempty cells is at least 2, a necessary condition for β_k to be identified. For example, in the following frequency table, row 2 fails the row check:

X/Y	**1**	**2**	**3**
1	0	13	56
2	100	0	0
3	0	22	1
4	18	5	3

Case C

Predictor has more than 2 categories and is treated as having **Fixed** scores but dependent variable is **Free** or dichotomous.

Same as Case B but check columns instead of rows.

Case D

Predictor is **Free** or dichotomous and dependent variable is **Free** or dichotomous. Check rows as in Case C and columns as in Case D.

In the case that one or more identification problems are found, GOLDMineR displays a summary description of the problem(s) and suggests alternatives to resolve the problem.

Delta Option

GOLDMineR offers an option to add a positive quantity to each observed cell count to resolve identification problems caused by empty cells. To use this option, users specify some non-negative number Delta to be proportionately allocated to the cell counts in a manner that preserves the observed distributions of both the X-profiles and the dependent variable. The quantity added to the observed (i,j) cell frequency count is:

$$p_i \, p_j \, \text{Delta}$$

where

p_i = proportion of the sample in the i^{th} X-profile

p_j = proportion of the sample in the j^{th} outcome category

Delta represents the total quantity added to the cells:

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_i \, p_j \, \text{Delta} = \text{Delta}$$

The quantities added to the cells do not affect the expected value of the dependent variable but do serve to dilute any effect estimates and significance tests by lowering the Total L^2. When used, *Delta* should be set at a low value such as 1 or 0.1% of the total sample size.

A.7 Multicollinearity

The final area of indeterminacy occurs as a result of multicollinearity among the X_m in Equation K-3. As in linear regression, in the case that any X_m can be expressed as an exact linear function of the remaining X-variables, β_m is not identifiable. GOLDMineR checks for multicollinearity by calculating the rank of the design matrix A (See "A.11 Design Matrix" on p. 231). If found not to be of full column rank, the Statistics window will display a message to that effect, which is a signal that one or more *alpha* or *beta* estimates are not identifiable, and adjust the number of degrees of freedom as described below.

In the case of multicollinearity, some of the *alpha* and *beta* estimates displayed in the Statistics window are nonunique estimates based on single value decomposition. To help determine which parameter estimates have been affected by the indeterminancy, GOLDMineR calculates the correlations among the parameter estimates (an optional statistic). In the case of a collinearity between two parameters, the correlation between the two parameter estimates will equal one. Also, parameter estimates involved in a multicollinearity may tend to have inflated standard errors and the unique contribution to the model chi-squared of a predictor involved in a multicollinearity, $L^2(Y)$, may be zero.

Following is an example of a multicollinearity associated with 3 dichotomous predictors, Z1, Z2, and Z3 due to the identity Z1 + Z2 + Z3 = 1 (see Table K-1).

Table K-1: Data for Multicollinearity Example

Z1	**Z2**	**Z3**	**Y**	**frq**
1	0	0	1	500
1	0	0	0	40
0	1	0	1	100
0	1	0	0	100
0	0	1	1	23
0	0	1	0	345

The design matrix in Figure K-3 was written to file 'design.txt' using the Design Matrix option in the Model menu after estimating the model that specifies that *Y* is the dependent variable, *Z1*, *Z2* and *Z3* are predictors and *frq* is a Frequency variable. Rank 4 is the full column rank because there are 4 columns in A corresponding to 1) the effect of *Z1*, 2) the effect of *Z2*, 3) the effect of *Z3*, and 4) the constant term *alpha*. The

message below the design matrix obtained from SVD, refers to the check for multicollinearity. The message states that the matrix is of "rank = 3", one less than full column rank. The second number ("0") in the message "sv = 2,0" represents the formal check for multicollinearity used in GOLDMineR. In the case of multicollinearity, the second number will be 0 (GOLDMineR considers any number smaller than 10^{-10} to be zero for this check).

Figure K-3: Design Matrix for Multicollinearity Example

```
Design Matrix:
     0     0     0     0
     0     0     0     0
     0     0     0     0
     0     0     1     1
     0     1     0     1
     1     0     0     1
SVD info=0 rank=3 wneed=268 wsize=268 sv=2,0
```

When multicollinearity is detected, the difference between the full rank and the actual rank is subtracted from the model chi-squared df and added to the residual chi-squared df prior to being reported in the Statistics window. After subtraction, the degrees of freedom from the model chi-squared will always be less than the sum of the df's associated with the chi-squared contributions, $L^2(Y)$.

Figure K-4:Partial Listing of Statistics window for Multicollinearity Example

y: y (Fixed)	1	0			
Y-scores	0.000	1.000			
Alpha(j)	0.00	-0.02			
	$L^2(Y)$	df	p-value	Beta	exp(Beta)
Z1 (Fixed)	-0.000	1	1.00	2.54	12.7
Z2 (Fixed)	-0.000	1	1.00	0.02	1.02
Z3 (Fixed)	-0.000	1	1.00	-2.69	0.0679

Association Summary	L^2	df	p-value	R^2	phi
Explained by Model	784.276	2	5.0e-171	0.6018	1.16
Residual	0.000	0	.		
Total	784.276	2	5.0e-171		

Converged after 90 iterations
Model is not full rank -- expected 4, obtained 3: df may not be correct.
(1 df subtracted from Model and added to Residual above).
Check standard errors, correlation of parameter estimates,
and $L^2(Y)$ for evidence of collinearities.

In Figure K-4 the degrees of freedom associated with the model chi-squared is 2, one fewer than the sum of the three 1- degrees of freedom associated with the contribution

chi-squared. Note also that the $L^2(Y)$ contribution calculated for each predictor is 0 yet the corresponding *beta* estimates are nonzero. This is an indication that the *beta* estimates are not identifiable. Whenever $L^2(Y) = 0$ for a given predictor, the predictor may be safely omitted from the regression without affecting any of the other statistics. In this example, any one of the predictors may be omitted without consequence.

A.8 Partitioning Chi-Square

The Association Summary in the statistics output partitions the independence chi-squared statistic (labeled "Total L^2") into a portion explained by the fitted model plus a residual:

$$\text{Total } L^2 = \text{Explained } L^2 + \text{Residual } L^2 \qquad \text{Equation K-30}$$

Total L^2 counts up all of the (qualitative) association in the data. Explained L^2 assesses the amount of association that is explained by the fitted model. Residual L^2 measures the extent to which the model explains less than all the association in the data.

Let the observed X-profiles form rows i =1,2,…,I, and the outcome categories form columns j =1,2,…,J in the following tables of frequency counts:

- the table of observed frequencies, $\{f_{ij}\}$
- the table of estimated expected frequencies under the fitted model, $\{\hat{F}_{ij}\}$
- the table of estimated expected frequencies obtained under the model of independence between the rows and columns, $\{\ddot{F}_{ij}\}$

When the table of observed frequencies is nonsparse, under large sample theory each L^2 component has a chi-squared distribution. For any fitted model containing d distinct model parameters, the L^2 likelihood ratio statistics are computed as follows:

Chi-squared Component	**L^2**	**degrees of freedom**
Explained	$2\sum\sum f_{ij}\ln(\hat{F}_{ij} / \ddot{F}_{ij})$	d
Residual	$2\sum\sum f_{ij}\ln(f_{ij}/ \hat{F}_{ij})$	(I-1)x(J-1) - d
Total	$2\sum\sum f_{ij}\ln(f_{ij}/ \ddot{F}_{ij})$	(I-1)x(J-1)

Chapters 8-10 illustrate the use of Residual L^2 to assess model fit (or lack of fit). In Chapters 8 and 9, lack of fit related to the omission of 1) category-specific parameters (for predictors with fixed scores) and/or 2) predictor interaction parameters. In

Chapter 10, lack of fit was related to the extent to which the association in the table required changes to the assumed Y-scores.

In general, for a given model H, the validity of *selected* model assumptions can be assessed by comparing it to a more general expanded model H^+ which includes an additional g distinct parameters excluded from model H. The significance of those g parameters can be assessed by examining the following likelihood ratio difference statistic with g degrees of freedom:

$$\text{Explained } L^2(H^+) - \text{Explained } L^2(H) = \text{Residual } L^2(H) - \text{Residual } L^2(H^+)$$

$$= 2\sum\sum f_{ij} \ln[\hat{F}_{ij}(H^+)/[\hat{F}_{ij}(H)] \qquad \text{Equation K-31}$$

where $\hat{F}_{ij}(H)$ and $\hat{F}_{ij}(H^+)$ denote the estimated expected frequencies under models H and H^+ respectively

In Chapter 10, the expanded model was a *saturated* model so that the expected frequencies estimated under the expanded model were equal to the observed frequencies. Substitution of f_{ij} for $\hat{F}_{ij}(H^+)$ in Equation K-31 shows that when model H^+ is saturated, the likelihood ratio difference statistic in Equation K-31 is identical to the Residual L^2 associated with model H. In example 1, the expanded model B maintained the same main-effects-only structure as the original models A and C but utilized an additional parameter to estimate scores for Y. The goal was to use the likelihood ratio difference between each original model and the expanded model B to test the theories that suggested use of the particular Y-scores in models A and C.

Depending upon the application, it may be appropriate to choose the expanded model to be the saturated model or it may be more appropriate to choose an unsaturated model to assess *selected* model assumptions. Magidson 1994b considers the case where Y-scores measure profitability and argues that in this context it is more useful to assess the extent to which the model explains not all of the *qualitative* association, but the quantitative association that is relevant to the Y(profitability) scores (called "Y-Association" in Magidson, 1993, 1994a, 1994b).

Similarly, tests of *significance* for parameters in model H may be performed by estimating a reduced model H^- that excludes g parameters that are included in model H, and computing the following likelihood ratio difference statistic:

$$\text{Explained } L^2(H) - \text{Explained } L^2(H^-) = \text{Residual } L^2(H^-) - \text{Residual } L^2(H)$$

$$= 2\sum\sum f_{ij} \ln[\hat{F}_{ij}(H)/[\hat{F}_{ij}(H^-)] \quad \text{Equation K-32}$$

In Chapter 7, the treatment effect was assessed using the optional LR contribution statistic $L^2(Y)$, the uses a likelihood ratio difference to assess the unique contribution of the treatment effect, with 1 degree of freedom. GOLDMineR computes $L^2(Y)$ for each predictor by estimating (behind the scenes) the reduced model that excludes the g distinct effect parameters associated with that predictor while retaining the Y-scores from the full model. The number of distinct effect parameters for any predictor corresponds to the number of columns in the design matrix -- g = K-1 for a qualitative predictor with K categories, g = 1 if the predictor is treated as quantitative.

The independence model is a reduced model that excludes all d distinct parameters. Substitution of $\ddot{F}_{ij}$ for $\hat{F}_{ij}(H^-)$ in Equation K-32 shows that when model H^- is the null model of independence, the likelihood ratio difference statistic in Equation K-32 is identical to the Explained L^2 associated with model H.

In general, the Pearson chi-squared statistic, denoted X^2, can not be partitioned into components. The corresponding formula for X^2 are

$$\text{Total } X^2 = \sum\sum\left(f_{ij} - \ddot{F}_{ij}\right)^2 \Big/ \ddot{F}_{ij} \quad \text{Equation K-33}$$

$$\text{Residual } X^2 = \sum\sum\left(f_{ij} - \hat{F}_{ij}\right)^2 \Big/ \hat{F}_{ij} \quad \text{Equation K-34}$$

A.9 Estimation of Standard Errors and Adjusted Residuals

As mentioned in section “A.8 Partitioning Chi-Square” on p. 226, the LR contribution “$L^2(Y)$“ assesses the contribution of each predictor in the model conditional on the Y-scores used in that model. Correlations, covariances and standard errors for the parameter estimates and adjusted residuals are also estimated conditional on the Y-scores. Conditional on the Y-scores, Equation K-3 is a log-linear model and hence the traditional approaches may be used to obtain these statistics and related p-values. (See e.g., Haberman, 1979, section 5.4 for relevant formulae.)

The p-value associated with the “Explained L^2“ assesses the significance of all model parameter estimates simultaneously. The number of degrees of freedom used to compute the p-value equals the number of distinct parameters that are estimated, including any estimated Y-scores. Hence, the difference between the associated Explained L^2 statistics between a model A that is equivalent to a model B except that some Y-scores estimated under model B are fixed under model A, is a chi-squared test

of those restrictions with degrees of freedom equal to the difference in the degrees of freedom used with each model.

Estimates for the Y-scores depend upon the predictors that are included in the model and the scaling used for those predictors.

A.10 Estimation Algorithm

GOLDMineR employs Goodman's (1979) unidimensional Newton algorithm successively to each standardized predictor $u_1, u_2, \ldots, u_M$ to obtain estimates $\hat{F}_{ij}$ that satisfy the maximum likelihood Equation K-35, Equation K-36 and Equation K-37, where n_i denotes the number of observations with profile $\mathbf{X}[i] = \mathbf{x}_i$ and n_j denotes the number of observations for which the jth outcome is obtained.

$$\sum_{j=1}^{J} \hat{F}_{ij} = \sum_{j=1}^{J} f_{ij} = n_i \qquad \text{Equation K-35}$$

$$\sum_{i=1}^{I} \hat{F}_{ij} = \sum_{i=1}^{I} f_{ij} = n_j \qquad \text{Equation K-36}$$

and $$\sum_i \sum_j \hat{F}_{ij} u_{mi} v_j = \sum_i \sum_j f_{ij} u_{mi} v_j \qquad \text{for } m = 1, 2, \ldots, M \qquad \text{Equation K-37}$$

It is easy to show that Equation K-37 still holds with the unstandardized Y-score y_j replacing the standardized score v_j, j=1,2,…,J. From Equation K-37 it follows that Equation K-38 also holds which states that the correlation between Y and the joint **X** variable is reproduced perfectly when the estimated expected frequency counts are used in place of the observed frequencies.

$$\sum_i \sum_j \hat{F}_{ij}(u_{1i} + u_{2i} + \ldots + u_{Mi}) v_j = \sum_i \sum_j f_{ij}(u_{1i} + u_{2i} + \ldots + u_{Mi}) v_j \qquad \text{Equation K-38}$$

For any **Free** predictor, say X_m, having K_m categories, let f_{kj} = observed frequency count in cell (k,j), of the two way table X_m x Y and let $\hat{F}_{kj}$ be the corresponding estimated expected frequency count k = 1, 2, …, K_m, j = 1, 2, …, J. Then the maximum likelihood estimates for $\hat{F}_{kj}$ also satisfy:

$$\sum_{j=1}^{J} f_{kj}\, y_j = \sum_{j=1}^{J} \hat{F}_{kj}\, y_j \quad \text{for } k = 1, 2, \ldots, K_m \qquad \text{Equation K-39}$$

Equation K-39 implies that if a predictor is treated as **Free** and scores x_k, $k = 1, 2, \ldots, K$ are estimated, the predicted (Y-score) for each category of X is equal to the observed conditional mean (Y-score) given $X = x_k$.

In GOLDMineR, the algorithm proceeds as follows:

The observed X-profiles are formed using all I combinations of the predictor variables that are observed in the sample. First, the independence model is estimated to obtain Total L^2 reported in the Association Summary section of the Statistics window. If the Y-scores are Fixed, the algorithm proceeds according to Step A below. Otherwise, the algorithm proceeds according to step B below.

If the LR contribution technical option is selected, M separate models are estimated, one at a time to assess the contribution of each predictor m = 1,2,…,M. As mentioned in section "A.8 Partitioning Chi-Square" on p. 226, GOLDMineR computes $L^2(Y)$ for each predictor by estimating a reduced model that excludes the distinct effect parameters associated with that predictor while retaining the Y-scores from the full model. Since estimates for the Y-scores depend upon the predictors included in the model, in the case that one or more Y-scores are estimated, all predictors must be included in the model first to obtain the Y-score estimates prior to obtaining $L^2(Y)$ for each predictor. Hence, in the case that one or more Y-scores are estimated, step B is applied first to obtain the estimated Y-scores prior to computing $L^2(Y)$.

If the decile fit technical option or the quantile fit specialized chart option is selected, additional models are estimated for the purpose of assessing model fit (see section "About The Decile Fit Statistic Interpretation in GOLDMineR" on p. 247 for details).

Computation of $L^2(Y)$.

For the mth reduced model, predictor m is treated as "margin only" by setting $u_{mi} = 0$ for each X-profile, i = 1,2,…,I prior to beginning the algorithm. Since Equation K-37 is satisfied, once set to zero, the u_m scores remain at 0.

Step A. Fixed Y-scores

For the mth predictor treated as qualitative:

Estimation proceeds according to Goodman's (1979) algorithm for the R model, modified according to (Goodman, 1985) when other than the **Uniform** scaling option is used for Y to take into account unequal spacing.

For the mth predictor treated as quantitative (with known x-scores):

In the case of **Fixed** Y-scores, estimation proceeds according to Goodman's algorithm for the U model, modified when other than the **Uniform** scaling option is used for X_m and/or Y to take into account unequal spacing.

Step B. Estimated Y-scores

For the mth predictor treated as qualitative:

Estimation proceeds according to Goodman's (1979) RC algorithm. In the case that at least one but not all Y-scores are estimated, estimation proceeds according to a mixture of the R and RC algorithm (see Magidson and SPSS, 1993).

For the mth predictor treated as quantitative (with known x-scores):

Estimation proceeds according to Goodman's (1979) algorithm for the C model modified when other than the **Uniform** scaling option is used for X_m to take into account unequal spacing.

A.11 Design Matrix

Let the i^{th} X-profile (profile **X**[i]) be defined by the M-tuple of observed values of the joint predictor $\mathbf{X} = \mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{Mi})$, let I be the number of **X**-profiles observed in the sample, J the number of categories for the dependent variable Y, and P the total number of distinct Alpha and Beta parameters to be estimated. Then, the generalized logit model (Equation K-3) can be expressed in matrix notation as $\Psi = Ab$ where Ψ is the IJx1 vector of logits $(\Psi_{1.1}, \Psi_{1.2,}\ldots,\Psi_{J.I})'$. The rows of A are sorted first by the categories of Y (the primary sort) and then by **X** so that each set of I rows, which correspond to the observed **X**-profile are repeated J times. The first I rows are

associated with category 1 of Y (i.e., associated with $\Psi_{1.i}$, i = 1, 2, ...I), the next with category 2 of Y and so on.

The design matrix A contains IxJ rows and P columns. The right-most J-1 columns of A represent contrasts that define alpha(1) - alpha(J-1). Each remaining column is associated with either a single beta coefficient if the associated predictor has only 2 categories or has more than 2 categories but is treated as quantitative (a FIXED predictor), or with one of the K_m-1 distinct components of the beta-vector representing contrasts associated with the K_m categories of the m^{th} predictor which is treated as qualitative (a FREE predictor).

The columns of the design matrix are constructed as follows:

The first P - J +1 columns correspond to the distinct *Beta* parameters that are estimated – 1 column for each FIXED predictor and K_m-1 columns for each free predictor X_m having K_m categories. For each FIXED predictor, the elements in the column associated with dependent variable category j are computed as $x^*_{mk} \times y^*_j$ where x^*_{mk} is the score associated with the kth category of that predictor (after subtracting the reference score x_{m0}) and y^*_j is the score associated with the jth category of the dependent variable (after subtracting the reference score y_0).

If the mth predictor variable is FREE, define K_m -1 indicator variables $Z(1_m)$, $Z(2_m)$, ...,$Z(K_m\text{-}1)$, where the mth indicator variable $Z(k_m)$ is '1' for those rows corresponding to the k_mth non-omitted category of that predictor, $-c_k/c_{k'}$ for those rows corresponding to the omitted category (category k' of that predictor), and 0 for the remaining rows. The elements in column K_m associated with dependent variable category j are computed as $Z(k_m) \times y^*_j$.

For each j = 1,2,...,J-1, the column corresponding to Alpha(j) is coded ‘1’ for those I rows corresponding to the jth dependent variable category, ‘-e(j)/e(0)’ for those I rows corresponding to the omitted "0" dependent variable category and ‘0’ for the remaining I x (J-2) rows, where $e = (e_1, e_2, ..., e_J)$ are the E-weights.

Table K-2 below illustrates various design matrices for the one predictor variable PAYMENT from example 3 where $E_0 = (0\ 1)$.

PAYMENT = Fixed (X=1,2,3,4; X_0 = 1) C_1 = (1,0,0,0)

b1(1)	alpha
0	1
1	1
2	1
3	1
0	0
0	0
0	0
0	0

PAYMENT = Fixed (X=1,2,3,3; X_0 = 1) C_1 = (1,0,0,0)

b1(1)	alpha
0	1
1	1
2	1
2	1
0	0
0	0
0	0

Grouped PAYMENT

PAYMENT = Fixed (X=1,2,3; X_0 = 1) C_1 = (1,0,0,0)

b1(1)	alpha
0	1
1	1
2	1
0	0
0	0
0	0

PAYMENT = Free (dummy coding, x-ref = (1)) C_1 = (1,0,0,0)

b1(1)	b1(2)	b1(3)	alpha
0	0	0	1
1	0	0	1
0	1	0	1
0	0	1	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

PAYMENT = Free (dummy coding, x-ref = (3)) C_1 = (0,0,1,0)

b1(1)	b1(2)	b1(3)	alpha
1	0	0	1
0	1	0	1
0	0	0	1
0	0	1	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

PAYMENT = Margin Only

alpha
1
1
1
1
0
0
0
0

PAYMENT = Free
(wtd. avg. coding - marg. proport.)
C_1 = (0.41,0.20,0.20,0.20)

b1(1)	b1(2)	b1(3)	alpha
-0.4895	-0.4969	-0.4812	1
1	0	0	1
0	1	0	1
0	0	1	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

PAYMENT = Free
(unwtd. avg. - effect coding)
C_1 = (0.25,0.25,0.25,0.25)

b1(1)	b1(2)	b1(3)	alpha
-1	-1	-1	1
1	0	0	1
0	1	0	1
0	0	1	1
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Table K-2
Various Design Matrices for predictor variable PAYMENT

A.12 Partial Effects Plots

Equation K-21can be rewritten as

$$\Phi_{j.\mathbf{x}} = \sum_{m=1}^{M} \Phi_{j.x}(m) \qquad \text{Equation K-40}$$

where $\Phi_{j.x}(m) = \phi\, u_m\, v_j$ Equation K-41

is the partial log-odds ratio associated with the m^{th} predictor, m = 1, 2, …, M. Hence, for any m, Equation K-41 provides the partial Effects plot displayed in GOLDMineR.

Similarly, we can express the generalized logit as a function of $X_{m'}$ by fixing the values of the other predictors, say at the X_m-reference level $X_m = x_{m0}$, m = 1, 2, …, M $m \neq m'$:

$$\Psi_{j.x}(m') = \alpha_j + \phi u_{m'} v_j \qquad \text{Equation K-42}$$

In the case that all predictors other than $X_{m'}$ are fixed at the X_m-reference value, Equation K-3 reduces to Equation K-29 in which case Equation K-1 becomes the partial regression curve associated with predictor m′.

A.13 Monotonic Regression Model is Invariant of the Refer-

ence Point (x_0, y_0)

Once a model is estimated in GOLDMineR, and the results are viewed from a given reference point, additional insights can often be obtained by viewing the results from a different reference point by changing the contrast weights associated with one or more variables. For example, in Chapter 9, the reference associated with the **Free** predictor PAYMENT was changed from the $1 level to the $3 level. We were then able to examine the estimate for β_4 under the revised contrast coding to assess the significance of the effect of a $4 payment vs. a $3 payment.

In Chapter 10 we saw that a change to the C-weights for some predictor does not affect the definition of the generalized logits ($\Psi_{j.\mathbf{x}}$ in Equation K-3), and hence does not alter the predicted probabilities ($P_{j.\mathbf{x}}$ in Equation K-2). Given a set of Y-scores (or Y-equivalent scores), estimates for the probabilities $P_{j.\mathbf{x}}$ in Equation K-2 are determined uniquely from estimates of the β_m parameters, which are not affected by the contrasts E_0 or C_m, m = 1, 2, …, M. Any model estimated by GOLDMineR is not only invariant of the reference point ($\mathbf{x}_0$, y_0), but also invariant with respect to Y-equivalent scores.

In the case that the mth predictor is treated as qualitative, estimates for X_m, and the category-specific beta parameters, β_{mk} k = 1, 2, …, K will change to reflect the change in coding. Similarly, changes to the dependent variable contrast, E_0, result in different definitions for the generalized logit, $\Psi_{j.\mathbf{x}}$, and therefore changes the estimates for the α_j, j = 1, 2, …, J. However, neither of these changes alter the $P_{j.\mathbf{x}}$, and hence the predictions for Y obtained from the monotonic regression model (Equation K-1) are not affected by changes to the category weights.

A.14 Analysis of a Continuous Dependent Variable

GOLDMineR is designed for the analysis of a dependent variable that is dichotomous or ordinal. However, a theoretical justification for the use of the monotonic regression model with continuous dependent variables is given in this section. From a practical perspective, however, memory requirements limit the analysis of a **Fixed** dependent variable to approximately 100 distinct values. If the memory limit is exceeded, GOLDMineR displays the message "Memory limit exceeded: ___ kb required for design matrix". (In such cases one may attempt to reduce the memory requirements using the Group option for one or more variables, reduce the number of predictors (especially those scaled as Free) or try the Search Procedure.) Typically, grouping the dependent variable into 20 categories causes minimal loss in power.

For continuous variables, we define the generalized log-odds ratios in terms of an assumed bivariate density function $f_{Y,X}$. For example, for reference points ($x_0 = \mu_X$, $y_0 = \mu_Y$) we have:

$$\Phi_{Y,X}(y,x) \equiv \ln \frac{f_{Y,X}(y,x) / f_{Y,X}(\mu_Y,x)}{f_{Y,X}(y,\mu_X) / f_{Y,X}(\mu_Y,\mu_X)} \qquad \text{Equation K-43}$$

We will use Equation K-43 to show that our monotonic regression model holds true in the case of bivariate normality.

For simplicity but without loss of generality we take $\sigma_Y = \sigma_X = 1$ and use the standardized from of the monotonic regression model as given in eq. (A12). The bivariate normal density is

$$f_{Y,X}(y,x) = A_1 \exp(-\tfrac{1}{2}\,[1/(1-\rho^2)][(x-\mu_x)^2 + (y-\mu_Y)^2 - 2\rho(x-\mu_x)(y-\mu_Y)])$$

$$= A_1 \exp[(-\tfrac{1}{2}\phi/\rho)((x-\mu_x)^2 + (y-\mu_Y)^2 - 2\rho(x-\mu_x)(y-\mu_Y))]$$

$$= A_1 \exp[\Phi_{Y,X}(y,x) - (\tfrac{1}{2}\phi/\rho)(x-\mu_X)^2 - (\tfrac{1}{2}\phi/\rho)(y-\mu_Y)^2] \qquad \text{Equation K-44}$$

where $A_1 = \dfrac{1}{2\pi\sqrt{\rho/\phi}}$; μ_Y and μ_X are population means.

$$= f_{Y,X}(\mu_Y,\mu_X)$$

and $\qquad \phi = \rho/(1-\rho^2) \qquad$ Equation K-45

Substitution of Equation K-44 into Equation K-43 for each of the four terms of $f_{Y,X}$ and simplifying yields:

$$\Phi_{Y,X}(y,x) = \phi(x-\mu_X)(y-\mu_Y) \qquad \text{Equation K-46}$$

Hence, Equation K-46 states that monotonic regression model (Equation K-21) holds true under bivariate normality (set M=1, $y_0 = \mu_y$ and $x_0 = \mu_x$). However, unlike the linear regression model which also holds true under bivariate normality, our monotonic model also holds true under many other important bivariate exponential family distributions as well (see Goodman, 1981, 1985 and Holland and Wang, 1987). Further generalization to the multivariate normal and multivariate exponential family distributions is given in the following section.

If the joint (X,Y) distribution is known to be bivariate normal, the regression of Y on X is linear and traditional linear regression programs should be employed to obtain increased efficiency. However, under most other exponential family bivariate distributions, the regression remains monotonic but is not linear in which case the linear regression provides only a linear approximation to the true monotonic curve. Since one rarely knows for certain that bivariate normality holds in practice, use of the monotonic regression model in GOLDMineR instead of linear regression is preferred on theoretical grounds.

But what if bivariate normality does hold? How might results from a monotonic regression be expected to compare to a linear regression on the same data? To begin to address this question we will use a small simulated sample of 30 cases from a bivariate normal distribution. More specifically, the assumed model is

$$Y = \alpha^* + \beta^* X + \varepsilon \quad \text{where } X \sim N(0,1), \text{ and } \varepsilon \sim N(0,1) \text{ independent of } X \qquad \text{Equation K-47}$$

For simplicity we fix $\alpha^* = 0$, and take $\beta^* = 1$ so that the population correlation is 0.7071 (i.e., $\rho^2 = .5$). Hence, $\sigma_Y^2 = 2$.

Since $\beta^* = \rho(\sigma_Y/\sigma_X)$, substituting Equation K-45 into Equation K-27 shows that under bivariate normality:

$$\beta^* = \sigma_\varepsilon^2 \beta \qquad \text{Equation K-48}$$

Since $\varepsilon \sim N(0,1)$, $\sigma_\varepsilon^2 = 1$, so that $\beta = \beta^*$. Hence, estimation of β in generalized logit model (Equation K-3) also provides an estimate for the linear regression coefficient β^* in this case. However, the additional term in the right hand side of Equation K-48 suggests that the estimate will not be as efficient as the usual least squares estimator under the linear model (Equation K-47).

Table K-3 and Figure K-5 compare predictions (and residuals) obtained from GOLDMineR with those obtained from linear regression for the simulated sample. The predictions coincide almost perfectly and both identify the same observation (case # 27) for which the residual is statistically significant at the .05 level.

			Linear Regression Results			Monotonic Regresson Results		
Case #	**X**	**Y**	$\hat{y}_{.x}$	**Residual**	**Std. Residual**	$\tilde{y}_{.x}$	**Residual**	**Adj. Residual**
1	-0.04	-0.14	0.08	-0.22	-0.19	0.08	-0.23	-0.21
2	-0.24	-0.47	-0.15	-0.31	-0.28	-0.15	-0.32	-0.29
3	0.07	0.57	0.21	0.36	0.32	0.21	0.36	0.34
4	-1.52	-0.73	-1.62	0.89	0.78	-1.59	0.86	0.90
5	0.57	0.75	0.78	-0.04	-0.03	0.77	-0.02	-0.02
6	-1.44	-0.99	-1.53	0.53	0.47	-1.52	0.52	0.53
7	-0.64	1.05	-0.61	1.67	1.46	-0.62	1.67	1.52
8	0.03	1.00	0.16	0.83	0.73	0.17	0.83	0.77
9	0.60	0.91	0.82	0.09	0.08	0.80	0.11	0.10
10	0.85	1.58	1.10	0.48	0.42	1.09	0.49	0.47
11	0.18	0.80	0.34	0.47	0.41	0.33	0.47	0.44
12	-1.03	-0.54	-1.05	0.51	0.45	-1.07	0.53	0.50
13	1.47	3.32	1.82	1.50	1.32	1.90	1.42	1.44
14	-1.13	-2.69	-1.17	-1.52	-1.33	-1.18	-1.51	-1.44
15	-1.38	-2.49	-1.45	-1.04	-0.91	-1.45	-1.04	-1.04
16	0.37	-0.75	0.55	-1.30	-1.14	0.54	-1.29	-1.22
17	0.51	0.47	0.71	-0.24	-0.21	0.69	-0.22	-0.21
18	-0.45	-1.23	-0.39	-0.83	-0.73	-0.40	-0.83	-0.76
19	-0.79	-1.30	-0.78	-0.53	-0.46	-0.79	-0.52	-0.47
20	-0.52	-1.07	-0.47	-0.60	-0.53	-0.48	-0.60	-0.54
21	-1.04	-0.46	-1.06	0.60	0.53	-1.08	0.62	0.58
22	-0.78	0.84	-0.76	1.60	1.41	-0.78	1.61	1.48
23	0.40	-1.03	0.59	-1.62	-1.42	0.58	-1.61	-1.53
24	-0.36	1.69	-0.29	1.97	1.73	-0.28	1.97	1.79
25	0.30	-1.21	0.47	-1.68	-1.47	0.46	-1.67	-1.58
26	-1.69	-1.08	-1.81	0.73	0.64	-1.75	0.67	0.73
27	-0.48	-2.91	-0.42	-2.49	-2.19	-0.43	-2.49	-2.26
28	0.31	0.50	0.48	0.02	0.02	0.47	0.03	0.03
29	-1.03	-2.21	-1.05	-1.16	-1.02	-1.07	-1.14	-1.07
30	-0.19	1.23	-0.09	1.32	1.16	-0.08	1.32	1.21

Table K-3
Results from a simulated sample of 30 bivariate normal observations (x,y) (observed correlation = 0.626)

Figure K-5

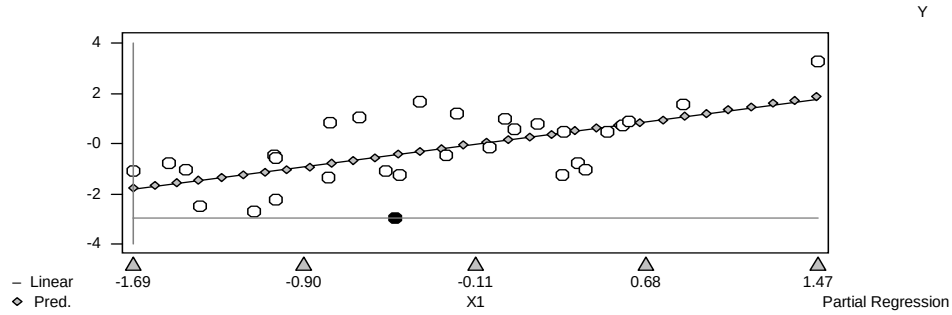

Y
4
2
-0
-2
-4
– Linear
◇ Pred.
-1.69
-0.90
-0.11
0.68
1.47
X1
Partial Regression

Figure K-6:Results from GOLDMineR

	$L^2(Y)$	df	p-value	Beta	exp(Beta)		
X (Fixed)	14.81	1	0.00012	0.932	2.54		
Association Summary	L^2	df	p-value	R^2	phi	Pearson X^2	p-value
Explained by Model	14.81	1	0.00012	0.392	1.013		
Residual	189.3	840	1.00			787.4	0.90
Total	204.1	841	1.00			870	0.24
Decile Fit	9.854	8	0.28			8.129	0.42

Converged after 46 iterations -- correlation reproduced to 4 digits.

Linear R^2	0.39203	(Obs)
	0.39203	(Exp)

Category-Specific Parameter Estimates

Y (Fixed)	y0	E-weights
	-2.91	(1)

X (Fixed)	x0	C-weights	Beta	Std. Err.	Wald	p-value	exp(Beta)	Lower	Upper
	-1.69	(1)	0.932	0.320	8.475	0.0036	2.54	1.36	4.76

Scores	Stnd	UnStnd
Y (Fixed)		
y0	0.000	-2.91
mn	1.908	-0.22
sd	1.000	1.41
X (Fixed)		
x0	0.000	-1.69
mn	1.799	-0.30
sd	1.000	0.77

Table K-4
Results from Linear Regression

R^2		**Beta**	**Std. Error**	**t**	**Sig.**
0.39203		1.146	.270	4.249	.00002

The Beta estimates turn out to be similar (.932 from GOLDMineR vs. 1.146 from linear regression) and the significance test is somewhat more powerful under the linear regression approach because the information that the regression is linear (which follows from the bivariate normal assumption) is taken into account. Under both approaches, the estimate for β is statistically significant (from Table K-4 we have $p = .00002$ compared to $p = .0001$ from Figure K-6 based on the likelihood ratio statistic). While both statistics are theoretically justified for these data, the likelihood ratio statistic does not assume bivariate normality and is also justified under more general conditions.

Note that the measure of model *fit* in Figure K-6 based on the likelihood ratio statistic and the Pearson statistic differ by a considerable amount for these data (Residual $L^2 = 189$, Residual $X^2 = 759$) indicating that the data is sparse and that these fit statistics should not be used in this application. However, the Decile Fit L^2 and X^2 statistics both yield the correct conclusion that the model assumptions fit the data. (See section"About The Decile Fit Statistic Interpretation in GOLDMineR" on p. 247 for further information on these statistics.)

The R^2 statistic used in GOLDMineR for the monotonic regression is a new statistic which is somewhat analogous to the traditional R^2 statistic in linear regression which is interpretable as the percentage of variance of the dependent variable Y that is explained by the linear regression model. We define the **monotonic R^2** statistic (to be distinguished from the usual ***linear* R^2**) as the ratio of the variance of the predicted values $\tilde{y}_{.x}$ to the variance of Y. In the case of a dichotomous dependent variable it can be shown that the **monotonic R^2** equals the R^2 analog proposed earlier by several others. See e.g., Light and Margolin (1971), Magidson (1981), and Haberman (1982).

Goodman (1981) points out that the estimate for ϕ can be used to provide an alternative to the usual approach to estimating the correlation coefficient ρ when the bivariate normal assumption holds true.

Solving Equation K-41 for ρ yields

$$\rho = (1 + \eta^2)^{1/2} - \eta, \qquad \text{where } \eta = 1/(2\phi) \qquad \text{Equation K-49}$$

Substituting the estimate 1.013325 for ϕ obtained from Figure K-6 into Equation K-49 yields and estimate $\tilde{\rho} = .622$ which is similar to the usual estimate r = .626.

Extension to Multiple Predictors

The results in Equation K-43 - Equation K-46 are extended here to the situation where there are multiple predictor variables and the predictors together with the dependent variable jointly follow the multivariate normal distribution.

Suppose there are M predictor variables, $\mathbf{X} = (X_1, X_2, ..., X_M)$. For notational simplicity we shall assume that each of the X variables and Y is standardized to have zero mean and unit variance. If Y and **X** follow a multivariate normal distribution, we have

$$f_{Y,\mathbf{X}}(y,\mathbf{x}) = A_4 \exp[-\tfrac{1}{2}(y\ \mathbf{x})\mathbf{V}^{-1}(y\ \mathbf{x})'] \qquad \text{Equation K-50}$$

$$\text{where } A_4 = \frac{1}{(2\pi)^{(M+1)/2}\sqrt{|V|}}, \qquad \text{Equation K-51}$$

V is the variance-covariance matrix, and |V| denotes the determinant of V.

For example, for the 1 predictor case examined earlier, we have:

$$f_{Y,X}(y,x) = A_1 \exp\left[-\frac{1}{2}(y\ x)\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}^{-1}\begin{pmatrix}y\\ x\end{pmatrix}\right] \qquad \text{Equation K-52}$$

$$= A_1 \exp\left[-\frac{1}{2}(y\ x)\begin{pmatrix}\phi/\rho & -\phi\\ -\phi & \phi/\rho\end{pmatrix}\begin{pmatrix}y\\ x\end{pmatrix}\right] \qquad \text{Equation K-53}$$

It will be instructive to partition $\mathbf{X} = (X_1,\mathbf{X}_2)$, where X_1 represents any single predictor variable and the remaining predictor variables are denoted by $\mathbf{X}_2 = (X_2,...,X_K)$. (Since the predictors can always be reordered in any fashion, this requirement involves no loss of generality.)

Now, we relax the assumption of zero means and take $\mu_k \equiv E(X_k)$ and $\mu_2 = (\mu_2,\mu_3, ... , \mu_K)$ and as a generalization of Equation K-43, we define the generalized log-odds ratio between Y and X:

$$\Phi_{Y,X1,\mathbf{X}2}(y,x_1,\mathbf{x}_2) = \ln \frac{f_{Y,X_1,X_2}(y,x_1x_2)\Big/ f_{Y,X_1,X_2}(\mu_y,x_1x_2)}{f_{Y,X_1,X_2}(y,\mu_1\mu_2)\Big/ f_{Y,X_1,X_2}(\mu_y,\mu_1\mu_2)} \qquad \text{Equation K-54}$$

The generalized log-odds ratio can be partitioned into separate portions associated with each X_k, or associated with any subsets such as X_1 and $\mathbf{X}_2$. For example, it is easy to verify that multivariate normality implies:

$$\Phi_{Y,X1,\mathbf{X}2}(y,x_1,\mathbf{x}_2) = \Phi_{Y,X1\cdot\mathbf{X}2} + \Phi_{Y,\mathbf{X}2\cdot X1} \qquad \text{Equation K-55}$$

where $\Phi_{Y,X1\cdot\mathbf{X}2}$ is the partial log-odds ratio between Y and X_1 with the remaining predictors $\mathbf{X}_2$ held constant:

$$\Phi_{Y,X1\cdot X2} = \ln \frac{f_{Y,X_1,X_2}(y, x_1 \mathbf{x}_2) \big/ f_{Y,X_1,X_2}(\mu_y, x_1 \mathbf{x}_2)}{f_{Y,X_1,X_2}(y, \mu_1 \mathbf{x}_2) \big/ f_{Y,X_1,X_2}(\mu_y, \mu_1 \mathbf{x}_2)} \qquad \text{Equation K-56}$$

$$\text{and } \Phi_{Y,X2\cdot X1} = \ln \frac{f_{Y,X_1,X_2}(y, x_1 \mathbf{x}_2) \big/ f_{Y,X_1,X_2}(\mu_y, x_1 \mathbf{x}_2)}{f_{Y,X_1,X_2}(y, x_1 \mu_2) \big/ f_{Y,X_1,X_2}(\mu_y, x_1 \mu_2)} \qquad \text{Equation K-57}$$

$$\Phi_{Y,X_1,X_2}(y, x_1, x_2) = \ln \frac{f_{Y,X_1,X_2}(y, x_1, x_2) / f_{Y,X_1,X_2}(\mu_Y, x_1, x_2)}{f_{Y,X_1,X_2}(y, \mu_1, \mu_2) / f_{Y,X_1,X_2}(\mu_Y, \mu_1, \mu_2)}$$

It follows that $\Phi_{Y,X1,\mathbf{X}2}$ may be expressed as the summation of terms associated with each predictor variable:

$$\Phi_{Y,X1,\mathbf{X}2}(y,x_1,\mathbf{x}_2) = \phi_1(x_1-\mu_1)(y-\mu_Y) + \sum_{m=2}^{M} \phi_m (x_m-\mu_m)(y-\mu_Y) \qquad \text{Equation K-58}$$

For example, for K = 2 we have:

$$V^{-1} = \frac{1}{D}\begin{pmatrix} 1-\rho_{12}^2 & -(\rho_{Y1}-\rho_{Y2}\rho_{12}) & -(\rho_{Y2}-\rho_{Y1}\rho_{12}) \\ -(\rho_{Y1}-\rho_{Y2}\rho_{12}) & 1-\rho_{Y2}^2 & -(\rho_{12}-\rho_{Y1}\rho_{Y2}) \\ -(\rho_{Y2}-\rho_{Y1}\rho_{12}) & -(\rho_{12}-\rho_{Y1}\rho_{Y2}) & 1-\rho_{Y1}^2 \end{pmatrix} \qquad \text{Equation K-59}$$

where $$D = 1-\rho_{12}^2-\rho_{Y1}^2-\rho_{Y2}^2+2\rho_{Y1}\rho_{Y2}\rho_{12} \qquad \text{Equation K-60}$$

$$= (1-\rho_{12}^2)(1-\rho_{Y.12}^2)$$

$$\phi_1 = \frac{\rho_{Y1}-\rho_{Y2}\rho_{12}}{D} \quad \text{and} \quad \phi_2 = \frac{\rho_{Y2}-\rho_{Y1}\rho_{12}}{D} \qquad \text{Equation K-61}$$

Now express the linear regression model in the usual form:

$$y^*_i = \mu_Y + (\mathbf{x}_i - \mu_{\mathbf{X}})' \beta^* + \varepsilon_i \qquad \text{Equation K-62}$$

where $\mathbf{x}'_i = (x_{i1}, x_{i2}, ..., x_{iM})$ Equation K-63

$\beta^* = (\beta^*_1, \beta^*_2, ..., \beta^*_M)'$ Equation K-64

and $\mu_{\mathbf{X}} = (\mu_1, \mu_2, ..., \mu_M)'$ Equation K-65

It is well known that $\beta^*_m = (\sigma_Y/\sigma_m)\rho^*_m$, where ρ^*_m is a multivariate generalization of the (bivariate) correlation called the path coefficient for variable X_m. Since $(1 - \rho^2_{Y\cdot\mathbf{X}})$ is the error variance σ_ε^2, establishing $\phi_m = \rho_m/(1- \rho^2_{Y\cdot\mathbf{X}})$ as the generalization to identity (Equation K-45), yields the following exact relationship between the parameters of the two models:

$$\beta^*_m = \sigma_\varepsilon^2 \beta_m \qquad \text{Equation K-66}$$

A.15: Assessing Model Fit when Analyzing Sparse Data

For the analysis of sample datasets 2, 3 and 4 in Chapters 8, 9 and 10 we used the likelihood ratio chi-squared statistic, Residual L^2, to assess and compare the fit of various models. The implicit assumption that Residual L^2 follows a chi-squared distribution was justified in those examples because the data arrays are comprised of nonsparse frequency counts. On the other hand, for our analysis of sample dataset #1 in Chapter 7, we did not use Residual L^2 because the chi-squared assumption is violated due to sparse data.

Sparse data occurs when the total sample size is small relative to the number of cells in the frequency table being analyzed. The number of cells is computed by multiplying the number of X-profiles observed in the sample by the number of outcome categories. Table K-5 below shows that the frequency table for sample dataset #1 consists of 189 cells but only 84 observations. The other datasets all contain many more observations than cells. When data is sparse, large sample theory can not be relied upon to justify the chi-squared assumption, and use of the Residual L^2 statistic should be avoided.

	Observed in Sample			
Example #	**Number of X-Profiles**	**Outcome Categories**	**Number of Cells**	**Sample Size**
1	63	3	189	84
2	6	2	12	3,200
3	8	2	16	10,653
4	2	5	10	104

Table K-5
Number of X-profiles in Example datasets

GOLDMineR output provides an optional alternative statistic for sparse data, called the "decile L^2",which utilizes a grouping of the joint X variable into deciles based on the predicted Y-value (and a dichotomous grouping of the dependent variable) to obtain nonsparse counts. A similar decile chi-squared statistic has been implemented in several logistic regression programs for dichotomous dependent variables (see “About The Decile Fit Statistic Interpretation in GOLDMineR” on p. 247).

How can you tell if your data is sufficiently sparse so that you should use decile L^2 instead of Residual L^2 to assess the fit of your model? GOLDMineR offers you a check on the validity of Residual L^2 by comparing it to the optional Pearson fit statistic, Residual X^2. When your data is nonsparse and your model is true, Residual L^2 should indicate that the model fits the data and Residual X^2 should be close in value. Hence, if you are unsure if your data is sufficiently sparse to invalidate the use of the Residual L^2 statistic for some model, you may request GOLDMineR to display Residual X^2 along with Residual L^2 . If Residual L^2 and Residual X^2 are <u>not</u> close in value, use decile L^2. If Residual L^2 indicates that the model fits the data, and Residual X^2 <u>is</u> close in value, then you have additional reason to accept the truth of your model. For *nonsparse* data you should always use Residual L^2 since decile L^2 has a somewhat conservative bias and the p-value will generally be somewhat inflated.

For the models estimated on nonsparse data in examples 2 and 3, Table K-6 shows that the Residual L^2 and Residual X^2 are close in value. Thus, we need not consider using decile L^2. (Note that the p-value for decile L^2 is only slightly higher than the corresponding p-value associated with Residual L^2 in each model that provides an adequate fit to the data. However, even for the model that does not fit the data (Model A in Chapter 8), despite the conservative bias, decile L^2 agrees with Residual L^2 in that

it provides strong evidence for rejection of Model A (decile L^2 = 81.51 as compared with Residual L^2 = 126.7).

For models estimated on sparse data in Example 1, Table K-6 shows some divergence in value between Residual L^2 and Residual X^2 under each of the models estimated. Therefore, we use the decile fit statistic here. We find that the low values for decile L^2 provide no evidence for rejecting Models A, B or C.

	Model	Residual L^2	p-value	Residual X^2	p-value	decile L^2	p-value
Example #1	**A**	112.9	1.00	128.3	1.00	5.62	0.69
	B	112.0	1.00	122.4	1.00	6.39	0.60
	C	113.0	1.00	121.9	1.00	7.50	0.48
Example #2	**A**	126.7	$28x10^{-27}$	124.7	$7.4x10^{-27}$	81.51	$1.5x10^{-17}$
	B	0.94	0.81	0.92	0.82	0.64	0.89
Example #3	**A**	9.45	0.09	9.46	0.09	9.06	0.11
	B	2.38	0.50	2.38	0.50	2.32	0.51
	C	3.18	0.67	3.18	0.67	3.05	0.69

Table K-6
Comparison of Three Fit statistics and Associated p-values for assessing fit of models

Table K-7 illustrates the use of a decile chart to compare average predictions from the monotonic and linear models by decile of X. The chart, like the graph in Figure 1-1, suggests that the predictions from the monotonic model tend to be better than the corresponding linear prediction for these data. Also, note that the average linear prediction in decile 1 is negative (-4.8), which reflects a structural problem with applying the linear regression model to the data displayed in Figure 1-1. This problem is circumvented by use of the monotonic model which always provides predictions between the observed minimum and maximum values of the dependent variable.

Decile Group g	$y_{\cdot g}$	$\hat{y}_{\cdot g}$	$\tilde{y}_{\cdot g}$
1	3.2	-4.8	5.2
2	4.6	4.0	6.4
3	7.8	10.5	7.7
4	8.1	18.8	10.4
5	19.3	29.5	16.4
6	50.0	50.4	47.4
7	80.9	62.5	71.2
8	76.1	72.2	81.5
9	85.2	81.3	85.9
10	87.4	98.3	90.4

Table K-7
Chart comparing grouped means for Y, $\hat{y}$ and $\tilde{y}$ by decile of X.

The decile chart in Table K-7 compares the grouped means observed for Y ($y_{\cdot g}$), with average predictions given by linear regression, ($\hat{y}_{\cdot g}$) , and monotonic regression, ($\tilde{y}_{\cdot g}$), within each decile group g. The observed means exhibit an increasing trend, which is clearly captured by both predictive models. However, predictions from the monotonic model capture more of the rapid increase between the 4th and 7th deciles.

About The Decile Fit Statistic Interpretation in GOLDMineR

Hosmer and Lemeshow (1980) proposed a goodness of fit statistic using decile groups based on predicted probabilities for the standard logistic regression situation (J=2) where at least one of the predictors is continuous.

Hosmer, Le Cessie and Lemeshow (1997) compare results from several computer packages that have implemented the Hosmer-Lemeshow statistic ("H-L"). While all packages produced the same predictions, the p-value assessing model fit based on their statistic varies from .02 computed from BMDPLR to .16 from SAS. The wide range in p-values results from different approaches used for selecting the cutpoints that define the deciles.

GOLDMineR contains an implementation of H-L statistic that avoids the ambiguity associated with defining the deciles by requiring every decile group to be based on exactly the same number of observations. Fractional "quasi-frequency" counts are constructed in cases where the sample size is not a multiple of 10. The approach to forming deciles is illustrated at the end of Tutorial #2. Given the resulting 10 x 2 table

of quasi-frequencies, the implementation requires estimation of an additional model with fixed X-scores assigned to each decile.

GOLDMineR also includes a quantile version of the H-L statistic (the quantile fit statistic) where any number of groups (not necessarily 10) can be specified. The GOLDMineR implementation of the H-L statistic has the property that for J = 2, if the number of groups is set to the sample size, the resulting quantile fit statistic (and associated p-value) are equivalent to the Residual L^2 (and associated p-value) obtained for the original model.

In cases where J>2, the H-L decile and quantile statistics are computed by estimating an additional model using a dichotomized form of the dependent variable. The dependent variable is dichotomized at the midpoint for the purpose of the H-L decile fit statistic reported in the Statistics window. For the more general H-L quantile statistic, two options are provided for dichotomizing the dependent variable. However, the dichotomization is performed exclusively for the purpose of computing the H-L statistic. All other output will be based on the total number of dependent variable categories J

Our implementation of H-L differs in two additional respects –

- the degrees of freedom are reduced in situations where the number of **X**-profiles is less than 10 so that the statistic when displayed with nonsparse data applications contains the correct number of degrees of freedom. (It can be shown that as the number of equally sized quantile groups approaches the sample size, H-L approaches the traditional chi-squared fit statistic.)
- H-L is further generalized for J>2 by dichotomizing the dependent variable. GOLDMineR contains two options for dichotomizing (see “Specialized Charts” on page 145).

Because the H-L statistic is still undergoing development, GOLDMineR includes an option to change the number of quantile groups from 10 (the default used for the decile fit statistic reported in the Statistics output) to any number between 2 and the total sample size N and obtain the corresponding “quantile fit” statistic. (See “Specialized Charts” on page 145) If a dichotomous dependent variable is being analyzed and the number of quantile groups is increased to N, the resulting quantile fit statistic equals the usual fit statistic.

Bibliography

Agresti, A. 1990. *Categorical Data Analysis*, New York: John Wiley and Sons.

Anderson, J.A.,1984. Regression and Ordered Categorical Variables (with discussion), *Journal of the Royal Statistical Society*, Series B. 46:1-40.

Clogg, C. and E. Shihadeh, 1994. *Statistical Models for Ordinal Variables*, Series: Advanced Quantitative Techniques in the Social Sciences, Volume 4, Thousand Oaks, CA.Sage Publications.

DeJonge, H., 1983. Deficiencies In Clinical Reports for Registration of Drugs, *Statistics in Medicine*, Vol. 2, 155-166.

Fisher, R.A., 1936. The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7, 179-188.

Goodman, Leo 1979. Simple Models for the Analysis of Associations in Cross-Classifications Having Ordered Categories, *Journal of the American Statistical Association*, 74:537-552. Reprinted in *The Analysis of Cross-Classified Data Having Ordered Categories*, 1984, Harvard University Press.

_____ 1981. Association Models and the Bivariate Normal for Contingency Tables with Ordered Categories, *Biometrika*, 68: 347-355.

_____ (1991a). Measures, Models and Graphic Displays in the Analysis of Cross-Classified Data, *Journal of the American Statistical Association*, 86:1085-1111.

_____ (1991b). New Methods in the Age of Progress in Contingency Table Analysis Using Measures, Models and Graphical Displays, Rejoinder, *Journal of the American Statistical Association*, 86: 1124-1138.

_____ (1996). A Single General Method for the Analysis of Cross-Classified Data: Reconciliation and Synthesis of Some Methods of Pearson, Yule, and Fischer, and also some Methods of Correspondence Analysis and Association Analysis, *Journal of the American Statistical Association*, 91: 408-428.

Haberman, S. 1979. *Analysis of Qualitative Data. Volume 1: Introductory Topics*, New York:Academic Press.

_____ 1982. The Analysis of Dispersion of Multinomial Responses, *Journal of the American Statistical Association*, 77:568-580.

Holland, P.W. and Y.J. Wang, 1987. Dependence Functions for Continuous Bivariate Densities, *Communications in Statistics – Theory Methods*, 16(3):863-876.

Hosmer and Lemeshow 1980. A goodness of fit test for the multiple logistic regression model, *Communications in Statistics*, A10:1043-1069.

Hosmer, S. Le Cessie and Lemeshow 1997. A comparison of goodness of fit tests for the logistic regression model, *Statistics in Medicine*, 16:965-980.

Ishii-Kuntz, Masako 1994. *Ordinal Log-linear Models*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-097, Thousand Oaks, CA:Sage.

Koch, G. and Edwards, S. 1988. Chapter 9 in K. Pearce, ed., *Biopharmaceutical Statistics for Drug Development*, Marcel Dekker, Inc., p. 406-407.

Light, R.J., and Margolin, B.H. 1971. An analysis of variance for categorical data, *Journal of the American Statistical Association* 63:1-28.

Magidson, Jay 1981. Qualitatitive Variance, Entropy and Correlation Ratios for Nominal Dependent Variables, *Social Science Research*, 10:177-194.

_____1993a. *SPSS for Windows CHAID Release 6.0*, Chicago: SPSS Inc.

_____1993b. The Use of the New Ordinal Algorithm in CHAID to Target Profitable Segments, *Journal of Database Marketing*, London: Henry Stewart Publications.

_____1994a. The CHAID Approach to Segmentation Modeling: CHi-squared Automatic Interaction Detection, chapter 5 in Bagozzi, Richard, *Advanced Marketing Research*, Cambridge MA.:Blackwell.

_____1994b. Multivariate Statistical Models for Categorical Data, chapter 4 in Bagozzi, Richard *Advanced Marketing Research*, Cambridge MA.:Blackwell.

_____ 1995. Introducing a New Graphical Model for the Analysis of an Ordered Categorical Response - Part I, *Journal of Targeting, Measurement and Analysis for Marketing (UK)*, Vol. IV, No. 2: 133-148.

_____ 1996a. Maximum Likelihood Assessment Of Clinical Trials Based On An Ordered Categorical Response, *Drug Information Journal*, Volume 30: 143-170.

_____1996b. Introducing a New Graphical Model for the Analysis of an Ordered Categorical Response - Part II, *Journal of Targeting, Measurement and Analysis for Marketing (UK)*, Vol. IV, No. 3: 214-227.

_____1998. Using New General Ordinal Logit Displays to Visualize the Effects in Categorical Outcome Data, chapter 34 in Blasius, Jorg and Greenacre, Michael, eds., *Visualization of Categorical Data*, Boston, MA: Academic Press.

Rudas, Tamás, 1998. *Odds Ratios in the Analysis of Contingency Tables*, Series:Quantitative Appliations in the Social Sciences, 119, Thousand Oaks: Sage Publications.