Innovative Strategies for Strengthening
Interpretability of Covariance Analysis by Use of
Complementary Parametric and Nonparametric Methods

by

Catherine M. Tangen

A dissertation submitted to the faculty of The University of North Carolina in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill

1997

Approved by:

[Signatures]

Advisor
Reader
Reader
Reader
INNOVATIVE STRATEGIES FOR STRENGTHENING
INTERPRETABILITY OF COVARIANCE ANALYSIS BY USE OF
COMPLEMENTARY PARAMETRIC AND NONPARAMETRIC METHODS

by

Catherine M. Tangen

Department of Biostatistics
University of North Carolina

Institute of Statistics
Mimeo Series No. 2182T

May 1997
Catherine M. Tangen, Innovative Strategies for Strengthening Interpretability of Covariance Analysis by Use of Complementary Parametric and Nonparametric Methods. (Under the direction of Professor Gary G. Koch)

Abstract

Analysis of covariance serves two main purposes in the analysis of clinical trials. It provides for a more powerful statistical test through variance reduction for comparison of treatment groups; and it addresses comparisons between randomized groups with adjustment for random imbalance of covariates.

Certain methods of covariance analysis such as logistic regression for ordinal response data and proportional hazards (Cox) regression for survival data are not readily interpretable with regard to their properties of variance reduction and bias adjustment. Direct variance reduction is not observed when adjustment is made for various covariates. It is a commonly held misconception that adjusting for covariates with logistic or Cox models not only increases variance estimates, but also leads to unstable treatment parameter estimates. In addition, there may be concerns that technical assumptions regarding the structure of the log-linear model are not appropriate such as the proper form of the quantitative covariates, no interaction between treatment and covariates, and the proportional odds (or hazards) assumption.

This research provides non-parametric counterparts to clarify the components of bias adjustment and variance reduction. There are no formal assumptions required for how a response variable is related to the covariates. The rationale for nonparametric covariance analysis is based on the randomization in the study design. Computations for these methods are through the application of weighted least squares to fit linear models to the differences between treatment groups for the means of the response variable and the
covariates jointly with a specification that has null values for the differences in covariates. The primary response can be on the log scale (e.g. logit, log hazard), but covariate adjustment is performed on the linear scale. The resulting parameter is an unconditional population average estimate of treatment effect, adjusted for imbalance of covariates.

The proposed methodology is applied to several types of outcomes including proportions, ordinal scores, logits (both dichotomous and ordinal logistic), logrank and Wilcoxon scores, incidence densities, and survival rates. Methodologic differences for hypothesis testing versus confidence interval estimation, and extensions to the stratified case, multiple doses (> 2), and multivariate response are addressed.
Acknowledgements

My deepest gratitude goes to Gary Koch for his nurturing and generous support throughout my years at UNC. He took me in as a GRA with no computer experience, and he patiently showed me through his words and actions what it means to be a biostatistician. I cannot adequately express my appreciation to him.

I would like to thank my committee members Dana Quade, Harry Guess, Phil Sloane, and Ron Helms for their time and input.

My best memories of my time spent here at UNC are the friendships I have made. In particular, the comradery in the Biometric Consulting Lab provided for a very supportive learning environment and a lot of fun. Many thanks to my classmates, past and present, who really made my experience an enjoyable one.

I have been lucky to have such a supportive family behind me. Special thanks to my sisters Maribeth and Terry, my father Len, and the Hillard family for the long distance phone calls, newspaper clippings, and letters to report on the news (and ski conditions) in Seattle, and to provide encouragement to me.

Finally, thanks to Brent, my adventurous husband, who keeps me laughing even through the hard times. His love and unwavering support makes me feel like I can accomplish anything. I share this degree with him.
# Table of Contents

## Chapter

I. Introduction to Analysis Methods for Clinical Trials Including Analysis of Covariance ................................................................. 17

1.1 Introduction of Analysis of Covariance ................................................................. 17

1.1.1 Benefits of Analysis of Covariance ................................................................. 17

1.1.2 Role of Stratification in Randomized Clinical Trials ........................................ 18

1.2 Considerations for Analysis of Covariance .......................................................... 20

1.2.1 Data Structure ................................................................................................. 20

1.2.2 Sampling Framework ....................................................................................... 21

1.3 Analysis Approaches ......................................................................................... 23

1.3.1 Randomization Methods: Categorical Response Data ...................................... 23

1.3.2 Modeling: Categorical Response Data ............................................................ 26

1.3.3 Multiple Regression: Modeling Continuous Data ............................................ 29

1.3.4 Survival Analysis: Time-to-Event Data ......................................................... 31

1.4 Analysis of Covariance: Literature Review ....................................................... 35

1.4.1 Previous Nonparametric Analysis of Covariance Research ............................ 35

1.4.2 Issues Regarding Covariates ........................................................................... 39

1.4.3 Covariates in Logistic Regression ................................................................. 42

1.4.4 Treatment Parameter Estimation in Non-linear Models .................................. 45
1.4.5 Treatment Parameter Estimation in Survival Models ........................................ 45
1.4.6 Conditioning on Balanced Covariates................................................................. 46
1.4.7 Omitted Covariates in Cox Models ........................................................................ 47
1.4.8 Prespecification of Covariates .............................................................................. 48

1.5 Examples to be Used in This Thesis ........................................................................ 50
1.6 Thesis Proposal ......................................................................................................... 51

II. Application of Response Means Linear Model to a Randomized Clinical Trial. 52

2.1 Description of Clinical Trial Example ...................................................................... 52
2.2 Assessing Summary Response Across Four Visits ..................................................... 52
    2.2.1 Unadjusted and Covariate Adjusted Models ....................................................... 54
    2.2.2 Assessing Goodness of Fit (GOF) of Covariate Adjusted Linear Model .......... 55
    2.2.3 Interpreting GOF Test and Degree of Covariate Adjustment ......................... 55

2.3 Center as a Stratification Factor ............................................................................. 56
    2.3.1 Method 1 for Addressing Stratification ($\beta_w$) .................................................. 56
    2.3.2 Method 2 for Addressing Stratification ($\beta$) ...................................................... 56
    2.3.3 Results of Stratified Models ............................................................................. 57
    2.3.4 Assessing Center x Treatment Interaction ......................................................... 59

2.4 Assessing Covariate x Treatment Interaction .......................................................... 60

2.5 Multivariate Models ............................................................................................... 61
    2.5.1 Assessing Response for Each Visit Separately ................................................... 62
    2.5.2 Assessing Treatment x Visit Interaction ............................................................. 63
2.5.3 Assessing Ordinal Response at Each of the Four Visits..........................64

2.6 Difference in Means Parameterization.......................................................67

2.7 SAS Catmod Procedure Approximation....................................................68

2.8 Nonparametric Covariance Matrix............................................................68

III. Modeling Logit Response With Linear Covariate Adjustment.........................69

3.1 Traditional Logistic Regression...............................................................69

3.2 Logit Response With Linear Covariate Adjustment......................................73
  3.2.1 Parameter Interpretation........................................................................75

3.3 Stratified Logit Response With Covariate Adjustment..................................76

3.4 Traditional Proportional Odds Model - Cumulative Logits..........................77

3.5 Proportional Odds With Linear Covariate Adjustment..................................80

3.6 Stratified Proportional Odds With Covariate Adjustment..........................82

IV. Logrank and Wilcoxon Hypothesis Testing..................................................86

4.1 Introduction of Clinical Trial Example.......................................................86

4.2 Kaplan-Meier Estimates..............................................................................91

4.3 Stratified Logrank and Wilcoxon Tests.....................................................91

4.4 Proportional Hazards Regression...............................................................92

4.5 Logrank and Wilcoxon Testing With Linear Covariate Adjustment..............96

4.6 Simultaneous Modeling of Multiple Doses of Test Treatment.....................99

4.7 Bivariate Methods - Wilcoxon and Logrank Tests...................................101
  4.7.1 Bivariate Logrank Tests at 12 and 18 Months.................................103
4.7.2 Combining Logrank and Wilcoxon Tests ................................................. 104

V. Survival Estimation: Incidence Density and Survival Rates ........................................ 107

5.1 Introduction to Incidence Density Estimation .................................................... 107

5.1.1 Estimating Interval Specific Incidence Density ............................................. 107

5.1.2 Estimating Proportional Incidence Density .................................................. 109

5.1.3 Testing Proportional Incidence Density Assumption ...................................... 111

5.1.4 Simultaneous Modeling of Multiple Doses of Test Treatment .......................... 112

5.2 Piecewise Exponential Model ............................................................................. 114

5.3 Survival Rate Estimation ................................................................................... 117

5.3.1 Estimating Interval Specific Survival Rate Differences ................................. 117

5.3.2 Estimating Average Differences in Survival Rates ................................------- 119

5.3.3 Testing Homogeneity of Differences in Survival Rates Across Intervals .......... 120

5.3.4 Simultaneous Modeling of Multiple Doses of Test Treatment ...................... 122

5.4 Summary ......................................................................................................... 123

VI. Logrank Testing Revisited: Application to a Cancer Clinical Trial ......................... 126

6.1 Introduction of Clinical Trial Example ................................................................ 126

6.2 Kaplan-Meier Curves for Survival and Disease-Free Survival ......................... 130

6.3 Cox Regression Results .................................................................................... 130

6.4 Correlations of Covariates With Response and Treatment .............................. 131

6.5 Stratified Logrank Tests With Linear Covariate Adjustment ......................... 133

6.5.1 Assessing Treatment x Strata Interaction ..................................................... 135
6.6 Multivariate Assessment of Survival and Disease-Free Survival..................136

VII. Summary and Future Research.................................................................139

7.1 Summary.................................................................................................139

7.2 Proposed Future Research.........................................................................145

A. Appendix....................................................................................................150

A.1 Creating Response Vectors.................................................................150

A.1.1 Response Function for Linear Response........................................151

A.1.2 Response Function for Dichotomous Logit Response......................152

A.1.3 Response Function for Ordinal Logit Response.................................152

A.1.4 Response Function for Logrank and Wilcoxon Scores......................153

A.1.5 Response Function for Incidence Density.........................................156

A.1.6 Response Function for Survival Rates..............................................159

A.2 Covariance Matrices..............................................................................162

A.2.1 Estimation............................................................................................162

A.2.2 Hypothesis Testing..............................................................................167

A.3 Univariate Response Parameter Estimation........................................168

A.3.1 Unrestricted Covariate Adjustment..................................................168

A.3.2 Covariate Adjusted Estimation of Treatment Effect for Univariate Response..................................................170

A.3.3 Assessing Goodness of Fit...............................................................171

A.4 Modeling Multivariate Response.........................................................172

A.4.1 Multivisit Response..........................................................................172
A.4.2 Multiple Test Statistics - Logrank and Wilcoxon Multivariate Tests................................................................. 175

A.4.3 Proportional Treatment Parameter Estimation................................................................. 177

A.5 Mean Difference Parameterization................................................................................. 180

A.6 Calculating Variance Reduction..................................................................................... 182

A.7 Extent of Covariance Adjustment.................................................................................... 182

A.8 Stratification.................................................................................................................... 183
  A.8.1 Stratification Method 1 ($\beta_w$)............................................................................... 184
  A.8.2 Stratification Method 2 ($\bar{\beta}$).............................................................................. 185
  A.8.3 Assessing Treatment x Strata Interaction............................................................... 187
  A.8.4 Stratified Multivariate Case...................................................................................... 187

A.9 Treatment x Covariate Interaction.................................................................................. 188

A.10 Multiple Treatment Arms ($>2$) in the Stratified Case.................................................. 188
  A.10.1 Estimation Setting.................................................................................................... 188
  A.10.2 Hypothesis Testing Setting..................................................................................... 192

A.11 Computations............................................................................................................... 193

A.12 Equivalence of Various Model Estimation Methods...................................................... 195
  A.12.1 Treatment Parameter Equivalence: Univariate versus Multivariate Model............. 195
  A.12.2 Treatment Parameter Equivalence: Means versus Difference in Means Parameterization............................................................................................................. 197

References............................................................................................................................ 201
List of Tables

Table 2.1  Descriptive P-values for Association of Covariates with Treatment and Response Where Response is the Sum of Favorable Outcomes for the Four Visits.................................................. 53

Table 2.2  Difference in Number of Favorable Responses, Summed Over the Four Visits, Treatment Parameter is Based on a Linear Model............. 54

Table 2.3  Stratified Adjusted and Unadjusted Treatment Effect Estimation, Outcome of Interest is the Number of Favorable Responses Summed Across 4 Visits........................................................................................................ 58

Table 2.4  Test of Treatment x Center Interaction Based on the Linear Model Assessing Number of Favorable Responses.................................... 59

Table 2.5  Results for Treatment Parameter (Difference in Favorable Response Rates) From Linear Model for Dichotomous Response, By Visit........... 62

Table 2.6  Assessing Treatment x Visit Interaction for Four Different Models Where Favorable Response (Y/N) at Each Visit is the Outcome of Interest........................................................................................................ 64

Table 2.7  Illustrating Potential Dichotomies Among the Ordinal Response Categories........................................................................................................ 65

Table 2.8  Results for Treatment Parameter (Difference in Means) From Linear Model for Ordered Response Scores...................................................... 66

Table 2.9  Assessing Treatment x Visit Interaction for Four Different Models Where Ordinal Score at Each Visit is the Outcome of Interest............. 67

Table 3.1  Results of Maximum Likelihood Logistic Regression, Modeling Favorable Response (Yes/No) at Each Visit.................................................. 71

Table 3.2  Response Logit With Linear Covariate Adjustment, Modeling Favorable Response (Yes/No) at Each Visit, Using Weighted Least Squares Methodology............................................................................ 74

Table 3.3  Modeling Response Logit Using WLS With Stratification and Linear Covariate Adjustment, Modeling Favorable Response (Yes/No) at Each Visit................................................................. 77
Table 3.4  Results of ML Proportional Odds, Modeling Ordinal Response Score at Each Visit.............................................................................................................................................................. 78

Table 3.5  SAS Logistic Procedure, Test of Proportional Odds Assumption for Each Visit, Modeling Three Cumulative Logits........................................................................................................................................................................... 78

Table 3.6  Results of Proportional Odds With Linear Covariate Adjustment at Each Visit Fitting Three Cumulative Logits.............................................................................................................................................................................. 80

Table 3.7  Test of Proportional Odds Assumption For Each Visit, Fitting Three Cumulative Logits........................................................................................................................................................................................................ 81

Table 3.8  Results of Proportional Odds With Linear Covariate Adjustment at Each Visit, Stratifying on Center Using Two Different Methods............. 83

Table 3.9  Test of Proportional Odds Assumption for Each Visit Based on a Center Stratified Model Using $\bar{\beta}$ Method...................................................................................................................................................................................... 84

Table 4.1  Means of the Stratification Factor and 21 Covariates, and Wald Test P-values for Comparisons Between the Placebo Group and the Dose 100/200 Combined Group........................................................................................................................................................................................................... 88

Table 4.2  Assessment of Collinearity Among ALS Covariates and Stratification Factor........................................................................................................................................................................................................ 89

Table 4.3  Results of Logrank and Wilcoxon Tests - Each Pairwise Dose Comparison With Placebo, Stratifying on Limb/Bulbar Status............. 91

Table 4.4  Results of Cox Regression - Unadjusted and 21 Covariate Adjusted Test Treatment Effect, Stratifying on Limb/Bulbar Status............... 92

Table 4.5  Cox Regression Score Tests to Assess Residual GOF for Test Treatment With or Without 21 Covariates in the Model, Stratifying on Limb/Bulbar Status........................................................................................................................................................................................................... 94

Table 4.6  Results of Unadjusted and 21 Covariate Adjusted Logrank and Wilcoxon Tests Using Linear Models, Stratifying on Limb/Bulbar Status........................................................................................................................................................................................................ 98

Table 4.7  Test of Overall Treatment Effect for Wilcoxon and Logrank Tests With or Without 8 Covariate Adjustment, Stratified ($\beta_w$) on Limb/Bulbar Status........................................................................................................................................................................................................ 100
Table 4.8  Test of Linear Trend for Wilcoxon and Logrank Tests With or Without 8 Covariate Adjustment, Stratifying ($\beta_w$) on Limb/Bulbar Status........100

Table 4.9  Goodness of Fit Test: Appropriateness of Constraining 21 Covariates to be Equivalent for Pairwise Comparisons, Stratifying ($\beta_w$) on Limb/Bulbar Status......................................................101

Table 4.10 Results of Unadjusted and 21 Covariate Adjusted Bivariate Logrank and Wilcoxon Tests Using Linear Models, With Stratification.........102

Table 4.11 Results of Unadjusted and 21 Covariate Adjusted Bivariate Logrank Tests at 12 and 18 Months, Using Linear Models, Stratifying ($\beta_w$) on Limb/Bulbar Status........................................103

Table 4.12 Results of Unadjusted and 21 Covariate Adjusted Sums of Standardized (-) Logrank and Wilcoxon Tests at 18 Months of Follow-up, Stratifying ($\beta_w$) on Limb/Bulbar Status.................................105

Table 5.1 Differences in Log Incidence Density Estimates Between Respective Treatment Groups, Stratifying ($\beta_w$) on Limb/Bulbar Status.................108

Table 5.2 Linear Model for Proportional Incidence Density - Unadjusted and 21 Covariate Adjusted Treatment Effect, Stratifying on Limb/Bulbar ($\beta_w$) Status..............................................................110

Table 5.3 Test of Proportional Incidence Density Assumption Over 3 Intervals of Time, Using the Wald Chi-square.................................................................112

Table 5.4 Test of Dose Effect in Proportional ID With or Without Covariate Adjustment, Stratification ($\beta_w$) on Limb/Bulbar Status.........................113

Table 5.5 Test of Dose Linear Trend in Proportional ID With or Without Covariate Adjustment, Stratifying on Limb/Bulbar Status.........................114

Table 5.6 Comparison of Four Treatments for Grouped Event Data.........................115

Table 5.7 Parameter Estimates From Poisson Regression........................................116

Table 5.8 Difference in Cumulative Survival Estimates Between Test Treatment Group and Placebo, Stratifying on Limb/Bulbar Status..........................118

Table 5.9 Survival Estimates for the Combined Dose Group 100/200 and Placebo, Stratifying on Limb/Bulbar Status.................................................................119
Table 5.10 Difference in Survival Rate Estimates Between Test Treatment Group and Placebo, Stratifying ($\beta_w$) on Limb/Bulbar Status .............................................. 120

Table 5.11 Test of Homogeneous Differences in Survival Rates For 3 Intervals of Time .............................................................................................................. 0

Table 5.12 Global Test of Treatment Effect With or Without Covariate Adjustment, Stratified on Limb/Bulbar Status, Response is Difference in Survival Rates .................................................................................... 122

Table 5.13 Test of Dose Linear Trend in Survival Rates With or Without Covariate Adjustment, Stratifying on Limb/Bulbar Status .................................................... 123

Table 5.14 Summary of P-values From 21 Covariate Adjusted Analyses With Stratification on Limb/Bulbar Status .............................................................................. 124

Table 6.1 Descriptive Statistics for Covariates and Stratification Factors by Treatment Arm .............................................................................................................. 127

Table 6.2 Cox Regression Results, Stratifying ($\beta_w$) on Cooperative Group ...... 130

Table 6.3 Spearman Correlations and Corresponding P-values of Covariates and Prognostic Scores With Dichotomous Outcomes, Logrank Scores (LR), and Treatment for Both Survival (S) and Disease-Free Survival (DFS) .................................................. 132

Table 6.4 Logrank Test With Linear Covariate Adjustment, Stratifying ($\beta_w$) on Cooperative Group ........................................................................................................ 134

Table 6.5 Assessing Homogeneity of Treatment Effect Across Cooperative Groups in the Setting of No Covariance Adjustment ......................................................... 135

Table 6.6 Individual Cooperative Group Treatment Estimates for Survival and Disease-Free Survival (DFS) ................................................................................................. 135

Table 6.7 Joint Logrank Test for Survival and Disease-Free Survival at 3 Years of Follow-up, Stratifying on Cooperative Group ......................................................... 136

Table 6.8 Combining Logrank Tests for Survival and Disease-Free Survival at 3 Years of Follow-up Using the Sum of Standardized Z-Scores (1 df Test), With Stratification ................................................................. 137
**List of Figures**

<table>
<thead>
<tr>
<th>Figure 4.1</th>
<th>Survival Distribution by Dose of Drug</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 6.1</td>
<td>Overall Survival by Treatment Arm</td>
<td>130</td>
</tr>
<tr>
<td>Figure 6.2</td>
<td>Disease-Free Survival by Treatment Arm</td>
<td>128</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction to Analysis Methods for Clinical Trials,
Including Analysis of Covariance

1.1 Introduction to Analysis of Covariance

Analysis of covariance (ANCOVA) is an important technique for controlling for the effects of prognostic variables. It was introduced into statistical practice by Fisher (1932), and Cochran (1957) clearly documented its nature and uses. Fisher (1932) expressed it as follows: "The analysis of covariance combines the advantages and reconciles the requirements of the two very widely applicable procedures known as regression and analysis of variance."

1.1.1 Benefits of ANCOVA

The first illustration of analysis of covariance in the literature was a randomized experiment where adjustment for covariates was employed to increase precision (Fisher 32). The covariate $x$ is a measurement, taken on each experimental unit before the treatments are applied, and it is thought to predict to some degree the final response $y$ on that unit (Cochran 57). This is probably the most frequent application of analysis of covariance today. In this use the function of covariance is the same as that of local control (pairing and blocking). It removes the effects of an environmental source of variation that would otherwise inflate the experimental error. However, it differs from blocking and stratification in two important ways. First, its use is restricted to quantitatively expressed prognostic variables, whereas blocking and stratification may be used with either quantitative or categorical prognostic variables. Second, unlike the other
two techniques, the analysis of covariance is purely a statistical method of control. It does not require any special arrangement of the experimental units or any special randomization scheme.

1.1.2 Role of Stratification in Randomized Clinical Trials

In the randomized clinical trial setting, there seem to be two extreme views regarding the use of stratification. One view has been to use as few stratifying variables as possible, even to the point of no stratification at all, relying on the balance that will come about by randomization. People who hold this view would make use of baseline data for post hoc adjustment through some form of covariance analysis. Peto et al (76, 77) express this viewpoint. The view from the other end of the spectrum is to stratify on many variables, so as to achieve a balance among groups that is as close as possible (Brown 80). The difficulty here is that even a small number of variables used for cross-classification into cells for stratification yields a large number of cells with few subjects in each, even to the point of having zero observations in some cells (Brown 80).

Correctly taking into account stratification in the study design can be troublesome for many analysis methods when counts in cells are small. However, recent research has produced techniques which are baseline adaptive randomization methods (such as biased coin methods) to virtually guarantee balance between treatment groups with regard to several covariates (Frane in press). Although these methods balance covariates marginally, cross-classification of covariables may not be balanced across treatment groups.

The feature of added precision due to analysis of covariance is, on the one hand, a positive feature. A study design may call for the comparison of parallel groups, and the effects of one or more quantitative prognostic variables may be controlled using appropriate statistical procedures. However, some investigators implicitly distrust purely
statistical methods of control (Fleiss 86). Fleiss (86) suggests that it is often helpful to present analysis of covariance as an adaptation of and an improvement on a purely statistical method of control that all investigators are comfortable with, the analysis of change when the prognostic variable is the same one, measured before treatment, that is used to measure response to treatment. An example of this would be measuring systolic blood pressure as the response of interest after a patient receives the intervention, and placing baseline (prior to intervention) systolic blood pressure in the model as a covariate. Modeling post score as a response and pre score as a covariable provides a conditionally unbiased test of treatment effect. However, Overall and Magee (92) point out that modeling change as the response measure (post—pre scores) does not address baseline imbalance.

A second purpose ANCOVA serves is to remove the effects of confounding variables in observational studies. In areas where randomized experiments are not feasible, there may be two or more groups differing in some characteristic, and the goal is to discover whether there is an association between this characteristic and a response $y$ (Cochran 57). In observational studies it is widely realized that an observed association, even if statistically significant, may be due wholly or partly to other disturbing variables in which the groups differ. Where feasible, a common device, analogous to blocking in randomized experiments, is to match the groups for the disturbing variables thought to be most important. In the same way, a covariance adjustment may be applied for $x$-variables that have not been matched (Cochran 57). Observational studies are subject to difficulties of interpretation from which randomized experiments are more nearly free. Even if matching and covariance adjustment have been appropriately applied, we can never be sure that bias may not be present from some confounding variable that was overlooked.
A third application for analysis of covariance is to shed light on the nature of the treatment effects. This application is closely related to the previous one. Not only does analysis of covariance allow for clarification of the degree to which detected differences between groups are due to treatment rather than other factors associated with response, but it also provides some structure for evaluating homogeneity of treatment differences among subgroups (Cochran 57, Koch et al 82, Cox and McCullagh 82).

One final application of analysis of covariance is its use in the development of missing data algorithms (Hemmerle 82).

1.2 Considerations for Analysis of Covariance

1.2.1 Data Structure

The type of analysis of covariance which is performed is dependent on two considerations. The first consideration is the type of data which is available; that is, what is the scale of measurement for the outcome? The second consideration involves the actual or presumed sampling framework from which the data have arisen. Another way to consider this second issue is to assess what the target population is. The answer affects the type of analysis which can be appropriately performed, and the kind of inference which can be drawn.

1.2.1.a Categorical Data

Data can be thought of as continuous or categorical. Categorical data can be further grouped into dichotomous response, discrete counts, nominal or ordinal response, or grouped survival data. Dichotomous responses have two possible outcomes which often are "yes" or "no", for example, did the patient respond favorably to treatment? More than two responses which take on a natural ordering are considered ordinal. Responses which take the values low, medium or high may be considered ordinal. A
patient global rating scale of five categories (excellent (1), good (2), moderate (3), fair (4), poor (5)) can be thought of as an ordinal scale. When outcomes don't fall into labeled categories but instead are numbers themselves, they are considered discrete counts. The number of offspring that an individual has may be considered discrete counts, and these counts may not be normally distributed. Two or more response categories with no apparent ordering are considered a nominal measure.

The final type of categorical response, grouped survival times, occurs when the response variable represents the number of patients who have an event of interest (e.g. time to healing, time to recurrence) during a given time interval instead of time being measured on a continuous scale for each individual. This type of response is commonly recorded for an individual when the event of interest is assessed at clinical visits where a screening procedure is necessary to assess whether the event took place, e.g. an endoscopy for ulcer detection, so the exact time that the event occurred is not known, but the interval within which the event occurred is known. Subjects being assessed for an event with intervals 0-1 years, 1-2 years, and 2-3 years and categorization of individuals as no event, event, or lost to follow-up for a given interval provide an example of grouped survival time (Stokes et al 85).

1.2.1.b Continuous Data

Data are considered continuous if, between any two potentially observable values, there is always another potentially observable value (Kleinbaum et al 88). Examples of continuous variables are age, blood pressure, and weight. Discrete variables have gaps between observations, i.e., between any two potentially observable values, there is a value that is not possibly observable. Discrete variables can sometimes be treated as continuous variables, especially if the possible values of such a variable are not far apart and cover a wide range of values (Kleinbaum et al 88).
1.2.2 Sampling Framework

Data arise from different sampling frameworks. The nature of the sampling framework determines assumptions that can be made for the statistical analyses which in turn guides the type of analysis that can be applied (Stokes et al 95). The sampling framework also determines the type of inference that is possible. Study populations are limited to target populations, those populations to which inferences can be made, by assumptions justified by the sampling framework (Stokes et al 95).

In general, data come from one of three sampling frameworks: historical data, experimental data, and sample survey data. Historical data are observational data. This means the study population has a definition based on geography or the circumstances of the data collection process, and this collection process can be either retrospective or prospective (Stokes et al 95).

Experimental data come from studies that involve random allocation of subjects to different interventions of some kind. Clinical trials where patients are randomized to either treatment or placebo, and types of fertilizer applied to agricultural plots would be two examples of settings from which experimental data would arise (Stokes et al 95).

In sample survey studies, subjects are randomly chosen from a larger study population. Some sampling designs may also have an element of experimental design. Researchers may randomly select subjects from a sample population and then randomly assign treatment to the selected study subjects (Stokes et al 95).

The major difference in the three sampling frameworks is the use of randomization to obtain them. Observational data have no randomization, and so it is
often difficult to assume they are representative of a convenient population.
Experimental data have good coverage of the possibilities of alternative treatments for the
restricted protocol population, and sample survey data have very good coverage of some
larger population (Stokes et al 95).

Randomization can be applied to an individual or a cluster of subjects (e.g. clinic).
Randomization can also be applied to subsets, called strata or blocks, with either equal or
unequal probabilities. This can lead to more complex designs such as stratified random
samples, or multistage random samples. In the experimental design setting these issues
can lead to repeated measures applications (Stokes et al 95).

1.3 Analysis Approaches

There are basically two approaches to the analysis of data: hypothesis testing and
modeling (Stokes et al 95). Randomization methods are often used to address a specific
hypothesis regarding association. Most often the hypothesis of interest is whether the
association exists between an outcome measure and an explanatory variable. However, it
is often of interest to also describe the nature of the association in the data. Statistical
modeling techniques using maximum likelihood estimation or weighted least squares
estimation are employed to describe the variability in terms of a statistical model.

1.3.1 Randomization Methods: Categorical Data

1.3.1.a 2 x 2 Tables

One of the most widely used methods of assessing the null hypothesis of no
association between two dichotomous variables is the Fisher's exact test for 2 x 2 tables
or the chi-square approximation. These methods require that the margins be fixed under
the null hypothesis (Stokes et al 95). This requirement is readily satisfied under
randomized allocation of patients to treatment in the null case where response is not
affected by treatment. If data are considered a convenience sample, one can still condition on margin totals being fixed by specifying the null hypothesis as corresponding to a random distribution of favorable response across patients regardless of treatment (Stokes et al 95). In general the chi-square test requires an expected count $\geq 5$ in each cell of the table for the approximation to be appropriate.

1.3.1.b Sets of 2 x 2 Tables

A natural progression from the 2 x 2 table is the analysis of sets of 2 x 2 tables. The principle behind this is to assess the relationship between the row variable and the response variable, controlling for the effects of the stratification factors (Stokes et al 95). Sometimes the stratification arises from the design of the study (e.g., clinic), and other times it becomes necessary to control for the effect of certain explanatory variables (e.g., gender). This type of analysis addresses the same question as the analysis of a single table: Is there an association between row and column variables, and what is the magnitude of that association? Chi-square statistics and odds ratios are still employed, but now the overall association is being assessed, not just one table. The product hypergeometric distribution can be induced via conditional distribution arguments when there is post-randomization stratification or when there are independent binomial distributions from simple random sampling (Stokes et al 95, Koch et al 90).

The Mantel-Haenszel strategy ($Q_{MH}$) can potentially remove confounding effects of explanatory variables and so a gain of power is attained for detecting association by comparing like individuals with one another (Stokes et al 95). $Q_{MH}$ is effective when it is expected that the direction of the association between exposure and outcome is fairly homogeneous across the strata. $Q_{MH}$ may fail to pick up an association if the effects are in opposite directions (Koch et al 90, Stokes et al 95) so testing for homogeneity of odds
ratios across strata by a Breslow-Day test or another appropriate test is necessary. Mantel and Fleiss (80) proposed a criterion for determining whether the chi-square approximation is appropriate for the distribution of $Q_{MH}$ for $q$ strata. The criterion states that the across-strata sum of expected values for a particular cell have a difference of at least 5 from both the minimum possible sum and the maximum possible sum of observed values. One of the major limitations of the Mantel-Haenszel method is that when there are a large number of strata for which adjustment is needed, it is not unusual for the Mantel-Fleiss criteria to be violated.

1.3.1.c $S \times R$ Tables

Although $2 \times 2$ tables are encountered most frequently, Mantel-Haenszel methods can be extended to larger tables (Koch et al 90, Stokes et al 95). For $2 \times r$ tables, there is an interest in assessing the association between a variable which has multiple ordinal response categories and a two level factor such as treatment. For $s \times 2$ tables, one may be looking at a dichotomous response and a gradient dosage of a treatment. Ordinal data analysis involves the choosing of scores to apply to the multiple response categories. There are a variety of choices: integer, standardized midranks (modified ridit scores), and logrank scores are among the most frequently employed. Selection of the type of score is based on the presumed scale and spacing of the response levels (Koch et al 82, Koch et al 90, Stokes et al 95), relative to the alternative of interest for the detection of treatment differences.

The Mantel-Haenszel strategies produce tests for specific alternatives to no association: general association, location shifts, and linear trends (Koch et al 90, Stokes et al 95). These methods can also lead to confidence interval estimation (e.g., 95% confidence interval for odds ratio). However, these randomization based hypothesis
testing methods do not necessarily explain the nature of the association between various parameters and the outcome of interest.

### 1.3.2 Modeling: Categorical Response Data

Statistical modeling, on the other hand, allows one to address the relationship between various model parameters. Logistic regression is a form of statistical modeling that is often used for ordered categorical outcome variables. Its purpose is to describe the relationship between a categorical response variable and a set of explanatory variables. Observations are assumed to be independent of one another.

#### 1.3.2.a Logistic Regression

Logistic regression (or its proportional odds extension) is used when the response of interest is dichotomous or ordinal categorical (McCullagh 1980). Explanatory variables can be either categorical or continuous. For this type of modeling, subjects in each treatment group, for example, are presumed to be representative of a larger population of patients with similar covariate distributions, leading to a sampling framework defined by the product-multinomial distribution for response status across covariate x treatment sub-populations (Koch et al 80, Koch et al 90). This is equivalent to assuming a stratified simple random sample with replacement.

The relationship of the response distribution to covariates and treatment is assessed by parameters which are estimated by maximum likelihood methods. A clinical trial with stratified randomization in multiple centers has the strata managed as additional covariables for a logistic regression model. A logistic regression model (or its proportional odds extension) is used for the conditional distributions of the response variable in sub-populations of patients according to the cross-classification of the covariables and treatment. A result is a test for the null hypothesis of no treatment effect.
which is adjusted for the covariates in the model. A linear model is obtained by modeling the logit which is the log of an odds. One can obtain a model-predicted odds ratio by exponentiating model parameter estimates.

For the parameters of the multinomial distributions for an ordinal response variable with \( M \) categories, the proportional odds extension of the logistic regression model has the following specification (McCullagh 80, Koch et al 90, Koch et al in press):

\[
\sum_{m'=1}^{m} \theta_{hikm'} = \frac{\exp(a_m + \xi_h + \gamma(2-i) + \sum_{L=1}^{q} \beta_L x_{nkL})}{1 + \exp(a_m + \xi_h + \gamma(2-i) + \sum_{L=1}^{q} \beta_L x_{nkL})}
\text{for } m = 1, 2, ..., (M - 1) \tag{1.1}
\]

where \( \theta_{hikm'} \) denotes the probability that a randomly selected patient from the sub-population corresponding to the cross-classification of the \( h \)-th stratum, \( i \)-th treatment, and covariates \( x_{hik} = (x_{hik1}, ..., x_{hikt})' \) has the \( m' \)-th outcome for the response variable where \( h = 1, 2, ..., q; i = 1, 2 \text{ for test treatment and placebo respectively; and } m' = 1, 2, ..., M \); the \( \{ \xi_h \} \) are unknown parameters for the respective strata and have the restriction \( \xi_q = 0 \) for identifiability, \( \gamma \) is the unknown parameter for the effect of test treatment, and \( \beta = (\beta_1, \beta_2, ..., \beta_t)' \) is the vector of unknown coefficients for the covariates \( x_{hik} \). All of these unknown parameters are interpretable as logarithms of ratios (relative to the \( q \)-th center for the \( \xi_h \), relative to placebo for \( \gamma \), and per unit change in the \( x_{hikL} \) for the \( \beta_L \)) for all of the respective odds \( \{ \sum_{m'=1}^{m} \theta_{hikm'}/(1 - \sum_{m'=1}^{m} \theta_{hikm'}) \} \) where \( m = 1, 2, ..., (M - 1) \); i.e., they are logarithms of odds ratios, and their homogeneity across \( m \) is the proportional odds assumption in the specification (1.1) for the model. The \( \{ \alpha_m \} \) are unknown parameters that correspond to intercepts for the predicted log odds for patients from the \( q \)-th stratum with placebo and \( x_{hik} = 0_t \) (where \( 0_t \) denotes the \((t \times 1)\) vector of 0's).

When \( M = 2 \) for a dichotomous response variable, the left-hand side of (1.1) simplifies to predict \( \theta_{hik1} \) only, and the \( \{ \alpha_m \} \) on the right-hand side simplify to \( \alpha \). Also, the
proportional odds assumption is no longer necessary since \( \{ \theta_{h_1k_1}/\theta_{h_2k_2} \} \) is the only type of odds to which the model applies.

Maximum likelihood methods are typically used to estimate the parameters in the logistic model (1.1). For sufficiently large sample sizes (e.g., all \( n_h \geq 20 \), and \( \sum_{h=1}^{q} n_h \geq 30 \sqrt{(1 + t)} \), and each possible outcome of the response variable has at least 10 patients with not overly skewed distributions of the covariates), the maximum likelihood estimates approximately have a multivariate normal distribution for which a consistent estimator of the corresponding covariance matrix is available (Koch et al in press). One can test linear hypotheses concerning the parameters in the logistic model (1.1) with likelihood ratio chi-square statistics and Wald statistics and one can test hypotheses for the inclusion of additional covariables in the model with score statistics (Koch et al 90, Stokes et al 95).

Some of the benefits of the logistic and proportional odds log-linear model approach include the following:

(i) higher statistical power (relative to its unadjusted counterpart) for the comparison between the treatments

(ii) its estimate of the treatment parameter applies homogeneously within each subpopulation according to the cross-classification of the covariabls and the strata (as well as to the overall population average across them), and in this sense, is generalizable to populations which might not have the same distributions for the covariabls and the strata as the population for the patients in the clinical trial (Koch et al, in press).

(iii) estimates are available for the effects of the strata and the covariabls in addition to that for treatments

28
(iv) capability for the evaluation of homogeneity of treatment differences across subgroups based on the covariables and the strata through statistical tests for the addition of treatment x covariables interaction or treatment x strata interaction to the model (Koch et al, in press)

(v) its application with sufficiently comprehensive models does not explicitly require randomized assignment of patients to treatments and thereby is potentially useful for eliminating selection bias from comparison groups in observational studies (Anderson et al 80, Koch et al 80).

Covariance analyses with the logistic regression model have the following limitations:

(i) the stratified simple random sampling assumption for selection of patients to participate in the clinical trial may not be realistic

(ii) one or more of the technical assumptions for the model structure may not apply such as: linearity for the relationship with quantitative covariables, no interaction between treatments and covariables, i.e., parallelism, no interaction between treatments and strata, and the proportional odds assumption when the response variable has 3 or more ordered categories.

(iii) the large sample properties of maximum likelihood estimates may not apply when subgroups have small sample sizes (e.g., < 10) (McCullagh 80, Stokes et al 95).

1.3.3 Multiple Regression: Modeling Continuous Response

When the outcome of interest is continuous, another type of analysis of covariance (ANCOVA) regression is employed. It involves a multiple regression model in which the study factors (such as treatment) are all treated as nominal factors by being incorporated into the model as dummy variables (Kleinbaum 88). The covariates may be measurements on any scale. The general ANCOVA model usually contains a mix of
dummy variables and other types of variables, and the dependent variable is considered continuous (e.g., blood pressure). The following standard regression assumptions apply to this ANCOVA: (1) existence of finite mean and variance for Y, (2) independence of Y-values from one another, (3) linearity - the mean value of Y is a straight line function of X, (4) homoscedasticity - the variance of Y is the same for any X, and (5) and that the expected value of Y given X is normally distributed with variance $= \sigma^2$ (Kleinbaum 88). Although this last assumption is not needed for the least squares model fitting, it is needed for inference regarding the parameter estimates when sample sizes are small. The method is fairly robust to departures from normality (Kleinbaum 88) when sample sizes are large.

The ANCOVA model assumes that there is no interaction of covariates with study variables. A regression model is fit such that:

$$Y = B_0 + B_1X + B_2Z + E$$  \hspace{1cm} (1.2)

where $X$ is the covariate (e.g. age) and $Z$ is a dummy variable that indicates the two groups to be compared (e.g., treatment = 1, placebo = 0). This model assumes that the regression lines for placebo and treatment are parallel, i.e., there is no significant treatment x age interaction. Under this model, the adjusted mean scores for placebo and treatment are defined to be the predicted values one gets when assessing the model when $Z = 0$ and $Z = 1$ when $X$ is equal to the overall mean age for both groups (Kleinbaum 88). Given that the no treatment x age interaction assumption is appropriate, it is then possible to compare the mean score for the two treatment groups as if they had the same age distribution. This covariance approach statistically equates the age distributions by treating both groups as if they had the same mean age. If the interaction (lack of parallelism) is significant, then an analysis that allows two separate regression lines (one for each group) would be more appropriate, and the nature of the interaction needs to be described (Kleinbaum et al 88).
This type of ANCOVA modeling can be expanded to account for \( K \) groups (e.g. multi-dose trial), and \( P \) covariates. \( K - 1 \) dummy variables are needed to specify the groups (e.g. 3 variables for 4 treatment groups). In this expanded scenario, the assumption of no interaction implies parallel hypersurfaces in \((p + 1)\)-space. Linear regression ANCOVA adjusts for disparities in covariate distributions over groups by artificially assuming that all groups have the same set of mean covariate values (Kleinbaum et al 88). By using ANCOVA, validity is achieved by adjusting for confounding, giving an unbiased estimate of association. Although there may be no confounding in one's data, controlling for one or more covariates will still be beneficial through a gain in precision if the covariates are associated with the outcome (Beach and Meier 89, Canner 91, Cochran 57, Cox and McCullagh 82).

1.3.4 Survival Analysis: Time-to-event Data

It is often of interest to know whether two or more groups differ with regard to time-to-event data, usually referred to generically as survival data. This can include time until disease recurs, time until healing, and time until death. For samples with progressively censored observations, the Kaplan-Meier product limit (P-L) method is appropriate for estimating the survival function (Fleming and Harrington 91). The next step is to assess whether the survival functions differ between groups. This is only appropriate if it is assumed that the groups are homogeneous with regard to background characteristics. A randomized clinical trial would be an appropriate data structure to apply P-L methods for comparing treatment effects.

1.3.4.1 Logrank and Wilcoxon Tests: Hypothesis Testing in Survival Setting

There are two different types of hypothesis testing for comparing groups in a manner which is applicable to data with progressive censoring: Generalization of the
Wilcoxon test (Gehan's, and Peto's & Peto's) and the non-Wilcoxon test (Cox-Mantel and logrank test) (Lee 92). The choice of test is based on the fact that one type of test is more powerful than another for a given circumstance. There is little difference between the Cox-Mantel and logrank test and between the two generalized Wilcoxon tests (Lee 92). However, the logrank test is more powerful than the Wilcoxon tests in detecting departures when the two hazard functions are parallel (proportional hazards), or there is random but equal censoring and when there is no censoring in the sample. The generalized Wilcoxon tests appear to be more powerful than the logrank test for detecting other types of differences, such as when the hazard functions are not parallel. The generalized Wilcoxon tests give more weight to early failures than later failures whereas the logrank test gives equal weight to all failures.

1.3.4.b Cox Regression: Modeling in the Survival Setting

As in the case with linear and logistic modeling, we often wish to adjust for covariates in our assessment of the effects of study variable on survival. It is often of interest to assess the relationship between possible prognostic variables and survival time. In the multiple regression approach, the survival time of the $i$-th patient, $t_i$, is the outcome of interest. We want to define a relationship of $t_i$ or a function of $t_i$, and covariates $(x_{1i}, x_{2i}, \ldots, x_{pi})$ that can be expressed in a regression function. Regression models for multivariate time-to-event analysis generally involve the assumption of proportional hazard functions (Lee 92). A proportional hazards model assumes that different groups of individuals have hazard functions that are proportional to one another, i.e., the ratio of the hazard functions for two subjects with covariates $x_1 = (x_{11}, x_{21}, \ldots, x_{pi})$ and $x_2 = (x_{12}, x_{22}, \ldots, x_{p2})$ does not vary with time. This implies that the hazard function given a set of covariables $x = (x_1, x_2, \ldots, x_p)$ can be written as $h(t \mid x) = h_0(t)g(x)$ where $g(x)$ is a function of $x$ and $h_0(t)$ can be thought of as the baseline
hazard function of an individual with \( g(x) = 1 \) (reference level of all covariates) (Lee 92).

Cox (Fleming and Harrington 91) introduced a general nonparametric model which is appropriate for survival analysis with and without censoring. His model uses the hazard function as the dependent variable. One needs to assume that survival time is continuous (i.e., probability of ties can be ignored). The model takes the following form:

\[
h(t|x) = h_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p} \quad (1.3)
\]

where \( h_0(t) \) is the hazard function of the underlying survival distribution when \( x = 0 \), and the \( \beta \)'s are regression coefficients. Cox's model assumes that the hazard of the study group is proportional to \( h_0(t) \). It is also assumed that the individual observations are independent, and censoring is uniform and non-informative.

To estimate the coefficients Cox suggests a maximum likelihood procedure where the likelihood function is based on the conditional probability of failure (Fleming and Harrington 91). Maximum likelihood estimates of \( \beta \)'s are obtained by solving simultaneously the \( p \) equations by the Newton-Raphson method (Lee 92). The proportional hazards model assumes that the hazard rate of an individual with prognostic variables, \( x \), is a constant multiple of the baseline hazard rate at all times. Although this assumption may be met in many situations, it is not reasonable for other cases. To account for this nonproportional case, the Cox model can be generalized using stratification. In this case the underlying hazard function, \( h_{0i}(t) \), may be different for each stratum; however, the regression coefficients are the same for all strata. With stratification the proportional hazards assumption applies within each stratum.
1.3.4.c Piecewise Exponential Models: Modeling Grouped Survival Data

Statistical models can be used to analyze grouped survival data by providing a description of the pattern of event rates. These models can describe this pattern over time, and they can also describe the variation in patterns due to the influence of treatment and other explanatory variables. Several assumptions must be made to use this model. They include: (1) assume withdrawals are uniformly distributed during the time intervals in which they occur and are unrelated to treatment failures, and (2) the within-interval probabilities of the treatment failures are small (Stokes et al 95). The time-to-event failures have independent exponential distributions (Stokes et al 95).

The piecewise exponential likelihood is written as follows:

$$
\phi_{PE} = \prod_{i=1}^{g} \prod_{k=1}^{t} \lambda_{ik}^{n_{ik}} \{\exp[-\lambda_{ik}N_{ik}]\}
$$  \hspace{1cm} (1.4)

where $n_{ik}$ is the number of failures for the $i$-th group during the $k$-th interval, $N_{ik}$ is the total person-months of exposure, and $\lambda_{ik}$ is the hazard parameter (Stokes et al 95). The piecewise exponential model assumes that there are independent exponential distributions with hazard parameters $\lambda_{ik}$ for the respective time periods. The $N_{ik}$, the amount of person time for individuals from population $i$ at risk for a given interval $k$, are computed as the length of the $k$-th interval multiplied by the sum from population $i$ of those completing the interval, half of the number who failed during the interval, and half of the number lost to follow-up during the interval.

If one thinks of the numbers of events $n_{ik}$, conditional on their amounts of exposure $N_{ik}$, as having independent Poisson distributions then it is possible to write a Poisson likelihood for the data (Stokes et al 95). Both the Poisson and piecewise exponential likelihood are proportional, and because of this result, one can assume the piecewise exponential model, but obtain estimates from Poisson regression computations.
which are more accessible (Stokes et al 95). It is also possible to impose a proportional hazards structure on the piecewise exponential model. The parameter vector $\beta$ relates the hazard function for the $i$-th population to the explanatory variables $x_i$ regardless of the interval (Stokes et al 95). Unlike the Cox model, one can test the piecewise exponential model for the proportional hazards assumption among the $t$ intervals.

1.4 Analysis of Covariance: Literature Review

1.4.1 Previous Nonparametric Analysis of Covariance Research

Initial encouragement of the methods to be introduced in this thesis came from research by Quade (67). The context of his research was mainly in terms of ranks. Data are thought to be representative of a random sample from each of $k$ populations. The multivariate covariates are perceived as random variables instead of constants. The key assumption is that the $X$ marginal distribution is the same for all $k$ populations. Prior to this research, nonparametric analysis of covariance was focused on $k = 2$ populations and only one covariate. Results of rank ANCOVA may be interpreted in terms of the probability that a response chosen at random from one population will exceed one from the "average" population less the probability of the opposite ordering.

Each variate is ordinal, but it is not necessary for it to be continuous. $Z_{ij} = a$ score which is the residual from a regression of ranks. A one-way analysis of variance is then performed on the scores, and a variance ratio test is used. The null hypothesis is that the conditional distribution of $Y$ given $X$ is the same for each population. The alternative hypothesis is that some populations have greater values of $Y$ than others for all fixed values of $X$.

Further amplification of these concepts was shown in Koch et al (82) and Amara and Koch (80). These nonparametric methods were revisited in the context of finite
population random sample models. For this setting subjects under study are viewed as a finite population which has been strictly randomized into two groups prior to treatment. Hypergeometric models are used as the basis for tests which take the form of a Mantel-Haenszel type test. The null hypothesis is that active and placebo treatment groups have equivalent bivariate distributions of covariate and response. In the single stratum case, the statistic \( Q_{MR}(Y, X) = Q_{MR}(X) + Q_{MR}(Y|X) \), (where \( MR \) stands for multivariate randomization), can be partitioned into two independent components under the null hypothesis. \( Q_{MR}(Y|X) \) can be interpreted as a covariance-adjusted test statistic. Because the criterion \( Q_{MR}(Y|X) \) is a multivariate randomization statistic with respect to the residuals of a combined sample set of multiple regressions of the \( Y_{ij} \) on the \( X_{ij} \) (where \( i \) indicates individual and \( j \) indicates treatment group), it is directly analogous to the rank analysis of covariance methods discussed by Quade (67). Koch et al (82) and Amara et al (80) focused on the chi-square approximation while Quade (67) emphasized the t-test or ANOVA.

In the latter portions of the Koch et al paper (82) and Amara and Koch paper (80), they provide extensions of their methods to the stratified case although no concrete illustration was shown. Methods for hypothesis testing and confidence interval estimation were provided. The covariance matrix which is used for confidence interval estimation = \( V_F \), a block diagonal matrix of within-treatment group covariance matrices. In contrast, \( V_o \), is used for hypothesis testing, and it is based on a covariance matrix estimated from the pooled treatment groups.

One additional contribution made by Koch et al (82) and Amara et al (80) is the somewhat crude introduction of the Mann-Whitney statistic where the null hypothesis is that the probability that active treatment response is at least as favorable as placebo response = 0.50. The analysis reflects covariance adjustment for a nonparametric
ranking function that takes into account the ordinal nature of the data but involves no scaling assumptions.

Carr, Hafner and Koch (89) refined the Mann-Whitney statistic with covariance adjustment, confidence intervals, multivariate settings, and analysis of data with stratification and missing values.

By use of an example, Koch et al (90) illustrated more explicitly the stratification methods for nonparametric analysis of covariance which were discussed in Koch et al (82). In addition, extensions to multivariate response were specified.

1.4.1.a Nonparametric Methods Using the Matching Principle

Quade (82) points out the distinction between "adjustment for \(X\)" which is performed in the classical ANCOVA setting and "holding \(X\) constant" which underlies procedures based on matching or blocking. In his paper, Quade (82) introduces two topics: (1) nonparametric analysis of covariance by matching, and (2) analysis of matched differences. His nonparametric analysis of covariance by matching can be regarded as a compromise between the standard analysis of covariance and the standard analysis of independent matched pairs. Quade (82) points out that there is no need to restrict attention to independent matched pairs, but rather all matched pairs can be incorporated.

When \(X\) is concomitant (i.e., its distribution is the same regardless of treatment) then methods based on randomization may be used. When \(X\) is not concomitant, methods related to partial correlation (between \(Y\) and treatment, given \(X\)) are applicable. Quade points out that either the actual values of \(Y\) or ranks of \(Y\) can be used with these methods (82). Throughout his paper, Quade (82) assumes that the observations form
independent random samples or that they come from a completely randomized design. The concept of "tolerance" was also introduced as the maximum distance between any two values of $X$ in which the two values are still considered matched. When $X$ is discrete, tolerance is often allowed to equal zero.

The null hypothesis for the analysis of covariance by matching method is that the conditional distribution is the same in all populations. A mean response ($Y_{ij}$) is calculated for all observations which are caliper matched on $X$. The corresponding score is the difference between the observed response and the mean response of the matched observations ($Y_{ij} - \bar{Y}_{ij}$). Under $H_0$, these residual scores are interchangeable and therefore can be used in an ANOVA. Both an ANOVA approximation and exact randomization method for small samples are possible for assessing statistical significance.

The concept of matched differences was also introduced in this paper (Quade 82), particularly for the case with one treatment group and one observation group. The assumption of concomitance is not required for this method so it can be applied in non-randomized situations. The interpretation of the population matched difference in mean is the conditional expected difference between the response of a treated observation and the response of a control observation, given that they are matched on $X$ (within the tolerance limit). A matched difference defined as an average conditional difference is analogous to a partial correlation defined as an average conditional correlation (Quade 82). Quade also noted that matched differences may be recognized as ratios of two-sample U-statistics (Quade 82).
1.4.2 Issues Regarding Covariates

1.4.2.a Selection of Covariates

Cox and McCullagh (82) state that in clinical trials there are two major benefits from analysis of covariance: (1) Adjustment for precision in designed experiments (i.e., increase precision of treatment contrasts). (2) Adjustment for bias in studies (comparability of groups) where the crude treatment effect can be partitioned into three categories: (i) the real group or treatment effect; (ii) an effect attributable to incidental variables, z; and (iii) unexplained or random variation. On average over all allocations, randomization ensures balance. However, a given treatment allocation is never balanced. Analysis of covariance is often employed as a means of controlling for imbalance which occurs despite randomization. The question that follows this course of action is which covariates to include in the analysis.

1.4.2.b Testing Balance of Covariates

Checking baseline comparability among treatment groups has been recommended by various authorities and experts (The U.S. FDA appears to recommend tables and graphs (Senn 94)). Stephen Senn (89, 94), among others (Altman 85, Altman and Dore 90, Tukey 93), has been a vocal opponent of statistical testing for imbalance. The alternative hypothesis of imbalance can only be true if randomization has not taken place, implying that deception was involved with treatment allocation or that some incompetence occurred. Senn argues that this is not the usual reason why such tests are carried out, but instead one is interested in making a statement about the observed allocation itself. Analysts who test for baseline balance can appear to be responsibly meticulous in inferential matters by claiming that if they find significant imbalance they will use the added protection of analysis of covariance, but Senn (94) points out that they can just as well be criticized as being insufficiently thorough because when these analysts do not find significant imbalance they take a risk and do not use analysis of covariance.
Balance has nothing to do with validity of statistical inference. Balance concerns efficiency of statistical inference, and valid inference depends on correct conditioning (Senn 94). An unconditional test of the effect of treatment in the context of a clinical trial is one that takes no account of the actual distribution of covariates which results from a randomization. A conditional test is one which does take account of an actual observed distribution of covariates. The result of an unconditional test can be regarded as corresponding to the average over all possible randomizations of the various conditional test results that would be obtained given the actual outcomes observed and the various baseline distributions corresponding to the possible results of the randomization (Senn 94). A conditional analysis of an unbalanced experiment produces a valid inference (Senn 94). Balance is useful in that, having decided to condition, we minimize the standard errors of our treatment effects if our covariate is orthogonal to treatment (Senn 94). Either one favors unconditional analysis, in which case there is no need to take account of balance, or conditional inference is favored, in which case one needs to take account of what is considered to be important whether or not it is unbalanced.

1.4.2.c Proposed Criteria for Covariate Selection

Covariates should not be fitted according to whether or not they are 'unbalanced' but according to whether or not they are important. (Altman 85, Altman and Dore 91, Begg 90, Senn 89, Senn 90) Senn (91) provides a guideline for choosing covariates to use in an analysis of covariance. (1) Identify on the basis of previous studies covariates of prognostic value. Cox and McCullagh (82) suggest that a covariate of prognostic value is one for which the partial correlation with the outcome variable given other important covariates is expected to be at least 0.30, but in many cases no great harm to efficiency will be done in using a lower threshold value. Specify these covariates a priori in the protocol or analysis plan. (2) Fit these covariates in an analysis of covariance
whatever the degree of imbalance. (3) Do not look at other covariates except possibly for the purpose of obtaining hints for future studies. An exception might be where evidence has accrued from other sources during the period between planning the trial and beginning the analysis and suggests that other covariates not originally considered to be important may now indicate prognostic value.

Enas, Enas et al (90) suggest identifying possible covariates at the design stage and categorizing these variables into one of three groups: (i) variables known to be important predictors of outcome, (ii) those for which historical evidence suggests the possibility of a relationship with outcome, and (iii) those for which there is no evidence of a relationship with outcome so they simply describe the population sampled. Enas et al (90) recommend including all from the first group, important factors from the second group (e.g. make a rule that the correlation should be $\geq 0.5$ in the current trial between the historically suggestive covariate and outcome for both treatment groups in order to be included in an analysis of covariance), and none from the last group. Enas et al (90) suggest that because imbalance on covariates used in ANCOVA may make it difficult to verify ANCOVA assumptions, it may be prudent to prevent imbalance on the most important covariates. However, they point out that stratification can make the study execution more difficult and is often impractical when the number of such baseline variables is large or the variables have many levels. Enas et al (90) suggest that tables displaying baseline patient information should be shown for the pooled patient sample. The exception would be covariates used in the analysis, generally specified in the protocol, for which summary statistics should be shown for each treatment group.

Overall and Magee (92) illustrate that analyses of simple pre - post difference scores are highly dependent on the direction of chance baseline deviations in relation to the directional treatment effect that would be inferred from rejection of the null
hypothesis. Since all three methods (raw scores, differences, and analysis of covariance) give unbiased estimates when averaged over all randomizations, the relative values of the three estimates in a given trial don't have anything to do with clinical relevance and everything to do with chance. Only analysis of covariance gives a conditionally unbiased estimate.

1.4.3 Covariates in Logistic Regression

1.4.3.a Effect of Covariates in Logistic Regression

Hauck, Neuhaus, et al (91) have specified two principal criteria for confounding in their paper. A covariate is considered a "classic" confounder if it is associated with the exposure and causally related to the outcome. The "operational" criterion has a broader definition. It states that a covariate is a confounder if the estimate of exposure effect is changed by inclusion of the covariate. A covariate which satisfies the operational criterion but not the classic criterion can be considered to be a "maverick". The authors point out that mavericks are a problem in logistic regression. They make a distinction between the bias considered due to omitting a maverick and that due to classical confounding. If the exposure actually has no effect, omitting a maverick cannot artificially introduce an effect. On the other hand, omitting a classical confounder can introduce an effect. Because of this, Hauck, Neuhaus et al (91) suggest that mavericks and classical confounders be considered as separate problems. Five results are specified in their paper, although prior researchers have also illustrated many of these points. (1) The effect of omitting a maverick is to bias the odds ratio towards no effect which is a generalization of the result that ignoring matching biases results towards no effect. (2) The magnitude of the bias caused by omitting a maverick increases with the variance of the omitted mavericks and with the magnitude of the effect of the maverick on the outcome. (3) A covariate that is related to the outcome can be omitted without introducing bias into the estimation of the odds ratio if: (i) the outcome is independent of
the exposure conditional on the covariate; or (ii) conditional on the outcome, the
covariate is independent of the exposure. (4) When outcomes are rare, the effect of
omitted mavericks becomes negligible. (5) Tests of the hypothesis of a no exposure-
outcome association remain valid when a maverick is omitted. This follows because
omitting a maverick cannot artificially introduce an effect. Omitting a maverick can be
thought of as mismodeling the explanatory variables. This has been shown to lead to a
loss in asymptotic efficiency, and so we would expect this type of omission to lead to a
loss in power (Hauck et al 91).

Odds ratios and logistic models are with reference to the proportion of individuals
with the outcome in a population that is defined by the exposure and covariates that are
incorporated in the model. Therefore, we are considering models for the proportion of
individuals with the outcome that have common values of these variables, but this
proportion is averaged across varying levels of all covariates or factors not included in the
model. Estimates of odds ratios give a comparison, with reference to the proportion
without exposure that have the event conditional on covariates in the model. When a new
covariate is added to the model or a covariate is deleted, we implicitly change the
structure of the reference population. Estimates of odds ratios then give a comparison
with reference to this new population structure. Omission of a maverick will lead to a
different odds ratio that is relevant and correct for a population structure in which the
maverick is allowed to vary; the bias refers to the fact that this is not the same odds ratio
that one obtains when the maverick is held fixed. Whenever mavericks are omitted, odds
ratios are closer to unity (in magnitude) of what they would be with all mavericks
included. Hauck, Neuhaus et al (91) conclude that both classical confounders and
maverick covariates which are strongly correlated with outcome should be considered
when estimating odds ratios.
1.4.3.b Asymptotic Relative Precision in Logistic Models

In Robinson and Jewell's paper (91) they look at the asymptotic relative precision (ARP) of an estimator in the adjusted relative to the unadjusted case. Their paper assumes both treatment and covariate are dichotomous. They show that adjustment for covariates always leads to a loss (or at best no gain) of precision with respect to logistic regression models. The variance of the pooled estimate is always less than or equal to the variance of the adjusted estimate (Robinson and Jewell 91). The conventional wisdom of variance reduction with adjustment does not apply in the logistic case. However, they contend that it is always as or more efficient to adjust for covariates when testing for the presence of a treatment effect in randomized studies, in the context of a logistic regression model, despite the associated loss of precision which is demonstrated in this paper.

The ARP of an estimator is the inverse of its asymptotic variance (Var unadj./Var adj.) relative to its unadjusted counterpart. In the linear regression case, when treatment and covariate are uncorrelated, ARP > 1. When the correlation between covariate and outcome is zero then ARP < 1 (Robinson and Jewell 91). A strong association between outcome and covariate has a beneficial effect upon precision of treatment effect, whereas a strong association between covariate and treatment has a detrimental effect, and therefore the precision of the treatment parameter reflects the competing effects of these outcome/covariate and treatment/covariate relationships. For the logistic case, ARP ≤ 1 with equality occurring if and only if the covariate is independent of outcome and treatment (Robinson and Jewell 91). Classic linear ARP can be <, >, or = 1 but logistic ARP is always ≤ 1 (Robinson and Jewell 91).
1.4.4 Treatment Parameter Estimation in Non-linear Models

Gail et al (84) showed in the non-linear regression setting that when treatment and covariate are independent, the value of the unadjusted treatment parameter falls between 0 and the adjusted treatment parameter, and so we see that the point estimate for the unadjusted treatment effect is smaller in absolute value (perfect randomization would ensure independence but in practice this rarely occurs). It can be concluded that the asymptotic relative efficiency (adjusted to unadjusted under the null hypothesis of no treatment effect) is \( \geq 1 \) with equality occurring only if the covariate is independent of outcome and treatment (Gail et al 84). To test the null hypothesis of no treatment effect in a randomized study, it is always as or more efficient (greater power) to adjust for the covariate when logistic models are used. In this regard the logistic regression model behaves similarly to the classic linear regression model. It was also noted in Robinson and Jewell's paper (91) that it would be straightforward to extend these results to the case where the covariate is discrete and multivariate.

1.4.5 Treatment Parameter Estimation in Survival Models

Chastang, Byar, and Piantadosi (88) explored aspects of bias in estimating treatment effect in survival models when there is failure to adjust on balanced prognostic variables. No such bias would occur in the more familiar linear regression model, although some efficiency would be lost in estimating treatment effect. The essential problem giving rise to the bias in survival models is that when important balanced covariates are omitted, the hazards in the treatment groups are no longer proportional, that is, the model is misspecified. Proportional hazards can hold either when the covariate is included or omitted, but not both (Chastang et al 88). Randomly mixing different hazards by ignoring strata does not give estimates equal to those within strata except in special cases. The extent of the asymptotic bias depends on the prognostic strength of the omitted covariates (Chastang et al 88).
1.4.6 Conditioning on Balanced Covariates

The results of the Chastang et al paper (88) suggest that it may also be important to adjust on balanced covariates as well, particularly if they have substantial prognostic value. If important differences are noted between unadjusted and adjusted analyses, it is probably preferable to base inference on the latter. This position is argued to be most defensible when careful specification of adjustment variables was made in advance because otherwise one might be criticized for exploratory data analyses directed toward trying to produce some particular result. Chastang et al (88) argue that preference for the adjusted model can be justified in two distinct ways: First, if the true model involves the covariates then one avoids bias in estimating the treatment effect. We never know what the "true" model is, but if there is a large difference in the adjusted and unadjusted treatment effect, it might be possible to show that the adjusted model fits the data better than the model with just treatment alone. In this case, the better fitting model can be considered closer to the "true" model. The second justification for preferring the adjusted model is that with it a different and possibly more important question is being answered which is specific to individual patients as characterized by their covariate values $z$ rather than averaging the treatment effect over all values of $z$ in the population. A proper view is that the adjusted and unadjusted treatment effect estimates are providing answers to different questions, the effect of treatment given $z$ versus the effect of treatment averaging over $z$ (Chastang et al 88).

Forsythe (77) examined the consequences of six strategies for choosing between analysis of variance and covariance where there is one potential covariate. The six different strategies were: (1) never use the covariate, (2) always use the covariate, (3) use the covariate if a preliminary test of the common slope yields significance at $\alpha \leq 0.05$ for that data (i.e., significant association between outcome and covariate in pooled sample),
(4) use the covariate if the test of treatment effect is more significant with the analysis of covariance than with analysis of variance, (5) use the covariate if a preliminary test of the equality of covariate means between groups is significant at $\alpha \leq 0.05$ (i.e., if there is significant imbalance of covariate between treatments), and (6) use the covariate if the slope is significantly different from zero and the covariate means differ significantly between groups. The results of the simulations showed that using criterion (4) severely distorts the actual level of the test, leading to too many false positive conclusions, and so it is not recommended. The suggestion made by the author is to always use the covariate in the analysis, and if a post hoc decision must be made whether to use the covariate, criterion (3) is the safest method under the conditions of the study that they assessed (Forsythe 77).

1.4.7 Omitted Covariates in Cox Models

Chastang, Byar, and Piantadosi (88) show through simulations that estimates of treatment effect in the Cox model are biased if an important balanced covariate is omitted even when the data are uncensored. This same result was confirmed by Gail, Wieand and Piantadosi (84). The relative bias is most affected by the prognostic importance of the covariate, the percentage censored, and the distribution of the omitted covariate. Gail, Wieand and Piantadosi (84) showed that the estimates of treatment effect are less biased by omission of needed covariates when the analysis is based on the exponential model than when the analysis is based on Cox's model, provided censoring is moderate. With heavy censoring, the two analyses yield similarly biased estimates of treatment effect.

The term "bias" in non-linear models may be more a matter of parameter interpretation than true bias. As Hauck et al (91) point out, adding covariates to a logistic model changes the interpretation of the treatment parameter. The treatment parameter applies to those individuals with the same covariate values, and if one changes the list of
covariates in the model then one also is also changing the interpretation of the treatment parameter. The same analogy holds true for the Cox model. The treatment parameter in the unconditional model is the population average estimate. On the other hand, treatment parameter estimates with covariates in the model are conditional on those covariate values. Each type of parameter has its own interpretation, but perhaps the term "bias" is misleading. However, there is concern about proportional hazards holding for the model, and bias may be an issue regarding model fit.

1.4.8 Prespecification of Covariates

Beach and Meier's paper (89) shows that unless tight control over the analysis plan is established in advance, covariate adjustment can lead to seriously misleading inferences. Illustrations from the clinical literature were provided. Typically there are too many covariates to include all of them in the analysis for large clinical trials. Beach and Meier (89) examined a method for assessing the impact of covariate adjustment based on choosing covariates which are both disparate (imbalance as measured by two sample t-test) and influential (as measured by \( R^2 \) or \( Z \), the test statistic for the associated coefficient in the model). The combination was measured by the product of influence and disparity. This was done using both formal analysis and simulation. Their opinion is that in the Cox model variance remains unchanged or increases with covariate adjustment so the justification for adjustment based on precision is a weak one, and they point out that the relevant measure of sample size with time-to-event data is the number of events, not the number at risk (Beach and Meier 89). {We take exception to this observation that precision is not obtained with covariance adjustment. It will be illustrated in the following chapters that log-linear models provide precision by increasing the size of the parameter estimate to a greater extent than increasing the standard error}. Beach and Meier note that covariate selection based on disparity outperforms selection based on influence. The criterion for "performance" was defined as a decrease in the mean square
error (MSE). This observation provides an additional argument for prespecifying covariates prior to looking at the data. Choosing covariates based on imbalance can lead to a more biased model. Canner (91) demonstrated that it is possible for baseline covariates that have only moderate disparity between two treatment groups or are only modestly associated with the outcome variable (p-value as high as 0.30 - 0.40) to cause a substantial adjustment of a treatment effect. Use of these results allow an assessment of sensitivity of findings to a range of covariate adjusted estimates based on various rules for selecting covariates.

The following common criticisms of analysis of covariance are addressed by Stephen Senn (94). (a) It often is pointed out that the relationship between a given covariate and outcome depends on the treatment, i.e. there is a lack of parallelism. Senn's response is that such a lack of parallelism is a problem, but the point to understand is that this covariate by treatment interaction exists whether or not the covariate is fitted in an analysis of covariance. He argues that it would seem inconsistent to want to remain ignorant regarding covariates but nevertheless look at their baseline distribution. (b) Sometimes the relationship between covariate and outcome might be non-linear, e.g. a quadratic relationship with age. Senn replies that if age is strongly related to prognosis then fitting the linear component alone is better than fitting none. Only the orthogonal component of age squared is not adjusted for. (c) It is often true that covariates tend to be heavily confounded with one another, and there is concern that this will have an adverse effect on the efficiency of estimates. Senn states that the point to remember is that it is the effect on the efficiency of the treatment estimate which is important. That covariates are confounded with one another does not matter so long as their effect on outcome is not of itself of primary importance. (d) Some suggest trying to fit various models, using different covariates, to check for robustness of the conclusions to the assumptions of the model. Senn points out that this has consequences for the sample size.
of studies. They will need to be larger for a given chosen power if all possible analyses are required to be significant before a success is claimed.

1.5 Examples to be Used in This Thesis

Three examples will be used in this thesis to illustrate different methods of covariance analysis. The first example involves a randomized clinical trial at two centers to compare two treatments (test, placebo) for 111 patients with a respiratory disorder (Koch et al 90). Global ratings of patient status (0 if terrible, 1 if poor, 2 if fair, 3 if good, 4 if excellent) at each of four visits during the treatment period are the response criteria, and baseline rating, age, and gender are the covariates.

The second example involves a randomized clinical trial which compares 3 doses of test treatment and placebo (placebo, 50 mg, 100 mg, 200 mg) for 959 patients with a life-threatening disorder. With 18 months of follow-up, the primary response criterion is overall survival. There were 22 covariates specified as a priori candidates for adjustment.

The third example involves 618 individuals with surgically resected colon cancer who were entered on a large national clinical trial (Moertel et al 90). These subjects were deemed eligible if they were free of disease at the time of registration and they were equally randomized to either observation or a chemotherapy regimen. The primary outcomes of interest were overall survival and disease-free survival, where time to disease recurrence or death is the event of interest. Several baseline clinical and pathologic covariates were collected on each subject.

The first and last examples show a strong association between treatment and favorable outcome when no adjustment for covariates is performed. The second example suggests a treatment effect, but the unconditional test is not statistically significant. The
first and second examples showed indications of random imbalance favoring the placebo group while the last example gave indications of random imbalance favoring treatment groups.

1.6 Thesis Proposal

Some types of modeling such as logistic regression and its extensions, and proportional hazards regression show no obvious benefit of variance reduction when analysis of covariance is performed. By applying non-parametric methods to log-linear response with adjustment for covariates on the linear scale, we expect to gain insight into the components of variance reduction and bias adjustment for a number of different settings. Methods for confidence interval estimation versus statistical testing will be explored in various scenarios.
Chapter 2
Application of Response Means Linear Model
To a Randomized Clinical Trial

2.1 Description of Clinical Trial Example

The data for this chapter come from a randomized clinical trial that was designed to assess the effect of treatment for 111 patients with a respiratory disorder. There are two centers within which patients were randomly assigned to two treatments (active, placebo) in successive blocks of six. The status of each patient was classified relative to a set of five ordinal categories (0 = terrible, 1 = poor, 2 = fair, 3 = good, 4 = excellent) at baseline and at each of four visits during the time period when treatment was being received. Age and gender were assessed for all patients at the time of randomization to treatment.

The two primary goals of this study are to estimate the treatment effect, and to assess whether there is any treatment x covariate interaction. For this study covariates of interest are Gender, Age, Baseline Status, and Center. Center is actually a stratification factor because of its role in the randomization process. Its effect will be assessed in three different ways: (1) ignoring stratification, (2) adjusting for Center as a covariate, and (3) managing Center as a stratification factor.
2.2 Assessing Summary Response Across Four Visits

A dichotomy can be defined for each of the four visits such that if a patient had a score of 3 or 4 at a given visit then favorable response = 1, otherwise favorable response = 0 for the corresponding visit. Each subject would then have assessments for four dichotomies corresponding to the four visits. An initial outcome of interest is the sum of these four dichotomies. The interpretation of this outcome would be the number of visits in which a favorable response was documented. An individual could have a response ranging from 0 (no favorable response for the four visits) to 4 (all visits had favorable response). A question of interest would be whether the mean number of favorable responses across the four visits is different for the two treatment groups (See A.1.1 for description of response function).

It is often interesting to obtain univariate correlations of covariates with response in order to assess which predictors are most strongly associated with the outcome of interest. In addition, correlations of covariates with treatment give an indication of how random the imbalance is between treatment groups for a given covariate.

Table 2.1 Descriptive P-values for Association of Covariables with Treatment and Response Where Response is the Sum of Favorable Outcomes for the Four Visits

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender</th>
<th>Baseline</th>
<th>Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>0.12</td>
<td>0.46</td>
<td>0.0001</td>
<td>0.004</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.59</td>
<td>0.015</td>
<td>0.98</td>
<td>0.93</td>
</tr>
</tbody>
</table>

P-values from Table 2.1 are based on Spearman correlations which were obtained by using SAS Proc Corr (SAS 90). Baseline status and Center are most strongly associated with response. There is also a suggestion of an atypical imbalance of gender for the two treatment groups.
2.2.1 Unadjusted and Covariate Adjusted Models

There are several approaches to comparing treatment effects for this outcome. One can estimate the unadjusted treatment difference by a method which involves no covariance adjustment, or the treatment difference can be estimated after adjusting for one or more of the covariates of interest. For the method of analysis that will be used in this chapter, covariate adjustment corresponds to restricting the difference between the two treatment groups in mean covariate values to be zero (see A.3.1, A.3.2 for model specifications).

Table 2.2 Difference in Number of Favorable Responses, Summed Over the Four Visits Treatment Parameter is Based on a Linear Model

<table>
<thead>
<tr>
<th>Model</th>
<th>Treatment Parameter</th>
<th>Standard Error</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.950</td>
<td>0.291</td>
<td>0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.980</td>
<td>0.235</td>
<td>0.00003</td>
</tr>
</tbody>
</table>

Model 1 is based on a linear model which estimates the treatment difference using an unstructured design matrix for response means and allows the four covariate differences to vary randomly for each treatment group. Model 2 estimates the treatment difference for the response after constraining the difference between treatment groups for average covariate values to be equal to zero. This second model manages Center (Center 1 = 1, Center 2 = 0), Age, Gender (Males = 1, Females = 0), and Baseline as covariates, and a restricted design matrix is employed (see A.2.1.a for structure of covariance matrix and A.3.1, A.3.2 for estimation of treatment effect). The unadjusted test indicates that active treatment has an average of 0.95 more favorable visits (among a total of four visits) than the placebo group, and this favorable treatment effect is significant (p=0.001). With adjustment for the four covariates, there is a small increase in the treatment parameter, suggesting that random imbalance at baseline favors the placebo group. Once this covariate imbalance has been adjusted, active treatment has a somewhat
larger effect. Moreover, improvement in precision from covariate adjustment leads to a smaller variance estimate (35% reduction). A sample size that is 65% of what would otherwise be required is sufficient with this variance reduction (see A.6 for formula). The combination of a larger estimated treatment effect and smaller variance produces a more significant test of treatment in the adjusted case (p=0.00003).

2.2.2 Assessing Goodness of Fit of Covariate Adjusted Linear Model

The goodness of fit (GOF) test compares the unadjusted model with the model that adjusts for four covariates (Center, Age, Gender, Baseline). The GOF test evaluates the compatibility of differences in mean covariate values with constraints to be zero (See A.3.3). The Wald chi-square = 7.00 with 4 degrees of freedom, and p=0.14. The results of this test confirm that constraining the four covariate mean differences to be zero is compatible with the estimated values for these quantities.

2.2.3 Interpreting GOF Test and Degree of Covariate Adjustment

Obtaining a small p-value from a GOF test indicates that there is substantial random imbalance among the treatment groups. There also is a corresponding suggestion of incompatibility with the constraint that differences in covariate means between treatment groups is zero. One might be cautious with the interpretation of ANCOVA results, although it should be recognized that ANCOVA is still justifiable. Randomization implies that the covariate adjusted model has a priori appropriateness, and correspondingly, it can be recognized that the unadjusted analysis should be viewed cautiously due to the atypical imbalance in covariates. This presents a dilemma for interpretation. The adjusted and unadjusted methods are equally problematic.

One possible solution is to consider a weighted combination of the adjusted and unadjusted methods. One can explore the stability of conclusions across a range of
believable weights that support the alternative hypothesis (see A.7 for a discussion of this issue). It should be noted that a similar approach has been introduced in the setting of cross-over trials (Tudor and Koch, 1994). In that setting, various weights were used to assess the effect of treatment in the presence of a carryover effect.

2.3 Center as a Stratification Factor

In the previous analysis, Center was managed as a covariate, but it can also be thought of as a stratification factor. Whether one considers Center as a covariate or a stratification factor affects the type of analysis performed. With Center as a stratification factor, mean response vectors and covariance matrices for each treatment group are calculated separately within each center. Mean values for the outcome (sum of favorable responses) and covariates (Age, Gender, Baseline) are calculated for treatment and placebo for Center 1 and Center 2 separately for a total of four hypothetical populations. The next step in a stratified analysis is dependent on the sample size of each stratum x treatment group classification.

2.3.1 Method 1 for Addressing Stratification ($\beta_w$)

If there are 25-30 individuals in each stratum x treatment group and more than one or two covariates to adjust for, one probably needs to combine the mean response vectors and covariance matrices across the two strata with weights which are proportional to stratum sample sizes or some other appropriate specification (see A.8.1). The result is one weighted response vector and one weighted covariance matrix for each treatment group based on contributions from both Centers. It is then possible to estimate treatment effects with or without covariance adjustment as was illustrated for the unstratified case. The resulting treatment effect is considered to be stratified by Center, yet concerns about asymptotics are addressed by combining across strata prior to estimation.
2.3.2 Method 2 for Addressing Stratification ($\bar{\beta}$)

If there are 50 or more subjects in each stratum x treatment group, then it is possible to calculate adjusted and unadjusted treatment effects within each stratum separately, and then a weighted treatment effect parameter can be estimated for the combined strata with weights that correspond to sample size or to reciprocals of estimated variances. It should be noted that weights based on the reciprocals of estimated variances require more stability so larger sample sizes are required for this type of weighting. For estimation purposes this second strategy for stratified analysis involves inverting covariance matrices for four populations (2 treatments x 2 centers) instead of two populations (2 treatments) as in method 1; therefore, stability of the covariance matrices is a concern. An additional consideration is the number of covariates one wishes to adjust for in the analysis. The larger the number of covariates, the greater the dimensions of the covariance matrix, and a larger sample size is required to obtain a stable, non-singular covariance matrix. However, if one does have an adequate sample size, this method of calculating adjusted treatment effect parameters within each center initially and then pooling across strata is more efficient than the method of proportionally combining means from the strata and then estimating the treatment effect (see A.8.2).

2.3.3 Results of Stratified Models

In this particular example there are 27 - 29 subjects in each treatment group within each center. We want to adjust for three covariates (Age, Gender, Baseline). This will involve estimation of a 4 x 4 covariance matrix (1 outcome, 3 covariates) for each stratum x treatment group. Sample size is marginal for applying the latter method of stratified analysis (i.e., Section 2.3.2, $\bar{\beta}$). However, both methods were applied to these data to compare results.
For the following table, Center is treated as a stratification factor instead of a covariate as was previously done in Section 2.2.1. Both stratification methods and a variety of weights are used to estimate the difference in response for the two treatments with or without covariate adjustment. For these models, there are three covariates (Age, Gender, Baseline).

Table 2.3 Stratified Adjusted and Unadjusted Treatment Effect Estimation
Outcome of Interest is the Number of Favorable Responses Summed Across 4 Visits

<table>
<thead>
<tr>
<th>Stratified Model</th>
<th>Weight</th>
<th>Tx Estimate</th>
<th>SE</th>
<th>Pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\beta}$ Method, Unadjusted</td>
<td>MH</td>
<td>0.943</td>
<td>0.281</td>
<td>0.0008</td>
</tr>
<tr>
<td>$\bar{\beta}$ Method, Adjust for 3 Cov.</td>
<td>MH</td>
<td>0.983</td>
<td>0.238</td>
<td>0.00003</td>
</tr>
<tr>
<td>$\beta_w$ Method, Unadjusted</td>
<td>MH</td>
<td>0.943</td>
<td>0.281</td>
<td>0.0008</td>
</tr>
<tr>
<td>$\beta_w$ Method, Adjust for 3 Cov.</td>
<td>MH</td>
<td>1.021</td>
<td>0.229</td>
<td>$8.0 \times 10^{-6}$</td>
</tr>
<tr>
<td>$\beta_w$ Method, Unadjusted</td>
<td>Variance</td>
<td>0.968</td>
<td>0.280</td>
<td>0.0005</td>
</tr>
<tr>
<td>$\beta_w$ Method, Adjust for 3 Cov</td>
<td>Variance</td>
<td>0.994</td>
<td>0.224</td>
<td>$8.9 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

For the particular outcome of mean number of favorable responses, the unadjusted treatment effect, treatment effect adjusted for four covariates, treatment effect stratifying on center, and treatment effect stratifying on center and adjusting for three covariates all lead to the same conclusion. Active treatment had significantly more favorable responses than did placebo. Adjusting for covariates and/or stratifying by center provided a more significant test of treatment than the unadjusted test. The results of various stratified analyses can be seen in Table 2.3 (see A.8.1, A.8.2 for discussion of weighting schemes). On average, across the different stratification methods and weight schemes, there was a 33% variance reduction between the stratified unadjusted and stratified 3 covariate adjusted models. This translates into a 49% savings in sample size (see A.6 for formula).
When one compares Table 2.3 to Table 2.2 (stratified versus unstratified), we see that for the stratified case using Mantel-Haenszel weights, the unadjusted treatment parameter estimate is slightly smaller than the unstratified unadjusted treatment parameter (0.943 versus 0.950, respectively), and the standard error for the stratified unadjusted test is smaller than the unstratified unadjusted test (0.281 versus 0.291, respectively). This leads to a net result of a more significant test of treatment for the stratified case. Both Method 1 ($\beta_w$) and Method 2 ($\tilde{\beta}$) stratification methods using MH weights and no covariate adjustment provided the exact same test of treatment effect in this example. Stratified estimates using inverse variance weighting were larger than their MH weight stratified counterparts (e.g. unadjusted 0.968 versus 0.943, respectively).

2.3.4 Assessing Center x Treatment Interaction

When one uses the second method for stratification adjustment with treatment parameter estimates and covariance matrices being calculated separately for each Center, it is straightforward to assess a potential strata (Center) x treatment interaction. This is equivalent to testing the null hypothesis that the treatment parameter for Center 1 is equivalent to the treatment parameter estimate for Center 2 (See A.8.3 for details of hypothesis testing). Results of this test can be seen in Table 2.4.

<table>
<thead>
<tr>
<th>Interaction/Factor</th>
<th>Chi-square</th>
<th>DF</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment x Center, Unadjusted</td>
<td>0.75 (0.236/0.316)</td>
<td>1</td>
<td>0.39</td>
</tr>
<tr>
<td>Treatment x Center, 3 Cov. Adjust</td>
<td>0.30 (0.064/0.211)</td>
<td>1</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The first row of Table 4 assesses the treatment x center interaction when there is no covariate adjustment. The second row displays the result of the test of interaction when there is adjustment for the three covariates (Age, Gender, Baseline) within each
Center. The numerator of the test is the squared difference in treatment parameter estimates for the two centers, and the denominator is the sum of the variance of the treatment parameter estimate for each of the centers. For both the unadjusted and covariate adjusted models, Center 2 had a larger treatment effect than Center 1 (Center 1: 0.70, 0.90, Center 2: 1.19, 1.15, respectively). One can see that the variance of the test of interaction is reduced for the covariate adjusted test. It would seem that random imbalance at baseline tended to favor the placebo group in Center 1, and tended to favor the treatment group in Center 2 because the treatment parameter estimate increased for Center 1 with covariate adjustment and decreased slightly for Center 2. Both of these tests of interaction are based on the $\bar{\beta}$ method of stratification (A.8.2). In both the adjusted and unadjusted models, results are compatible with homogeneous treatment effects across Centers.

2.4 Assessing Covariate x Treatment Interaction

It is possible to assess covariate x treatment interactions with the response means linear model. For this particular clinical trial, this assessment of interaction will be illustrated for gender x treatment. Initially, estimates of treatment are obtained for males and females separately within each of the two centers. This would involve four treatment parameter estimates. The next step would be to combine the gender specific treatment estimates across centers using weights which are proportional to sample size. The result is a center stratified treatment estimate for males, and a center stratified treatment estimate for females. With these gender specific estimates of treatment effect, three different objectives may be addressed.

1. One may assess treatment effect separately for each of the sexes,
2. One can compare the two treatment estimates (test of interaction), and
3. One can estimate covariate adjusted or unadjusted overall treatment effect by a weighted combination of the two gender specific estimates.
For (3), when there is concern about sample size, weights which are proportional to sample size are usually preferred. Precision weights may also be used when sample size is not a problem, and there is not a concern about a gender x treatment interaction being present (see A.9 for details).

The center stratified treatment x gender test of interaction which combines estimates across strata using Mantel-Haenszel weights yields the following result: (1) no covariate adjustment, Wald chi-square = 3.50 with 1 df, p-value = 0.061, and (2) with covariate adjustment (age, baseline), Wald chi-square = 5.19 with 1 df, p-value = 0.023. The unadjusted difference in estimated treatment parameters for (females - males) = 0.247 (s.e. = 0.619) and the covariate adjusted difference in estimated treatment parameters for (female - male) = 1.22 (s.e. = 0.53). These results indicate that females had a significantly more favorable response to treatment relative to the males and that this difference between the genders is more pronounced with covariate adjustment due to a larger difference in treatment parameter estimates and a smaller variance.

The method in the previous paragraph which describes the assessment of covariate x treatment interaction is considered the more standard method. However, if sample size is quite large, it is possible to treat gender as a post-stratification factor. One could fully stratify on Center x Gender x Treatment. A covariate adjusted (Age, Baseline) treatment effect would then be estimated for each of the four strata. One would then proceed to weight adjusted estimates of treatment across strata. This is in contrast to the standard method which performs covariate adjustment after a weighted pooling across strata. It is apparent that a large sample size is needed for this second method of assessing covariate x treatment interactions because each stratum x covariate x treatment subgroup must be fairly large in order to obtain stable, non-singular estimates of covariance matrices and parameters for each subgroup.
2.5 Multivariate Models

2.5.1 Assessing Response for Each Visit Separately

The next step in the analysis of this clinical trial is to assess, for each treatment group, the proportion of patients with favorable response at each visit separately. This involves four dichotomous outcomes for each individual. For each visit, the treatment effect can be estimated in the unadjusted case, adjusting for covariates, stratifying by Center, or a combination of the latter two.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>19.4%</td>
<td>31.8%</td>
<td>26.6%</td>
<td>17.3%</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>9.2%</td>
<td>9.0%</td>
<td>9.1%</td>
<td>9.4%</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.036</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>0.067</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>19.5%</td>
<td>32.7%</td>
<td>27.1%</td>
<td>18.7%</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>7.5%</td>
<td>8.4%</td>
<td>8.0%</td>
<td>8.2%</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.010</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.023</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>19.2%</td>
<td>31.6%</td>
<td>26.5%</td>
<td>17.0%</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>9.0%</td>
<td>9.0%</td>
<td>9.0%</td>
<td>9.1%</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.033</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>0.061</td>
</tr>
<tr>
<td>4</td>
<td>Treatment</td>
<td>21.1%</td>
<td>33.4%</td>
<td>28.2%</td>
<td>19.4%</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>7.2%</td>
<td>8.3%</td>
<td>7.9%</td>
<td>8.3%</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.003</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.019</td>
</tr>
</tbody>
</table>

Model Identification:
1) Unadjusted treatment effect - no stratification or covariate adjustment
2) Covariate adjustment (Age, Gender, Center, Baseline)
3) Stratify on Center, No Covariate Adjustment, \( \beta \) stratification using MH weights
4) Stratify on Center, Adjust for 3 cov.(Gender, Age, Baseline), \( \beta \) strat, MH weights

62
The results in Table 2.5 were obtained using mean response coding in a multivariate setting, i.e., all four visits were modeled simultaneously (see Appendix 1 for detailed specification of the model). Comparison of results using a univariate approach (one visit per model) and multivariate approach (4 visits simultaneously) yield the same treatment parameters and standard errors for each visit. This was true for both unadjusted and covariate adjusted models (see A.12.a for a proof of these results). In the stratified, multivariate setting it is more common to use Mantel-Haenszel weights when combining strata because the multivariate covariance matrix has fairly large dimensions, and there is more of a concern about sample size in this setting.

The results of Table 2.5 indicate that there are significantly more favorable responses in the treatment group than the placebo group for each of the four visits. However, Visit 4 showed only a marginally significant treatment effect for the unadjusted case (p = 0.067 unstratified, p = 0.061 stratified), but after covariate adjustment the effect of treatment was more clearly significant (p = 0.023 unstratified, p = 0.019 stratified). This was due to both an increase in the magnitude of the treatment effect and a decrease in the variance. Stratification had little effect on either treatment parameter estimation or variance reduction.

2.5.2 Assessing Treatment x Visit Interaction

In this multivariate response setting with four outcomes, it is of interest to know whether treatment effects are homogeneous across the four visits. The Wald Chi-square is used to test the null hypothesis of no treatment x visit interaction. Table 2.6 displays the results of testing for interaction for each of the four models specified in Table 2.5.

The results of these tests of treatment x visit interaction in Table 2.6 indicate that the treatment effect is fairly homogeneous across the four visits with respect to the
dichotomous outcome of favorable response: Yes vs No. This non-significant treatment x visit interaction supports the decision that was made earlier in this chapter to assess response as the sum of favorable responses across visits. One could also apply a model that imposes no treatment x visit interaction for the four models for the four visits. This would imply modifying the design matrix so that treatment effect estimation for the four visits would now have averaging across the visits (see A.4.1.c for details of assessing treatment x visit interaction).

Table 2.6 Assessing Treatment x Visit Interaction for Four Different Models
Where Favorable Response (Y/N) at Each Visit is the Outcome of Interest

<table>
<thead>
<tr>
<th>Model</th>
<th>Wald Chisquare</th>
<th>degrees of freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.1</td>
<td>3</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>3.2</td>
<td>3</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>3.2</td>
<td>3</td>
<td>0.36</td>
</tr>
<tr>
<td>4</td>
<td>3.1</td>
<td>3</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Model Identification:
1) Unadjusted treatment effect - no stratification or covariate adjustment
2) Covariate adjustment (Age, Gender, Center, Baseline)
3) Stratify on Center, No Covariate Adjustment, \( \bar{\beta} \) Method, MH weights
4) Stratify on Center, Adjust for 3 cov.(Gender, Age, Baseline), \( \bar{\beta} \) Method, MH weights

For the center stratified model (\( \bar{\beta} \) method, MH weights) with no covariate adjustment, the average treatment parameter estimate across the four visits is 0.261 ± 0.068 \(( p = 0.0001)\), and the covariate adjusted average treatment parameter estimate is 0.275 ± 0.053 \(( p < 0.0001)\). The same pattern of variance reduction with covariate adjustment can be seen when average treatment effect is estimated.
2.5.3 Assessing Ordinal Response at Each of Four Visits

It can be noted that there are four possible dichotomous cutoff points among the five adjacent ordinal response values at each of four visits. These dichotomies are represented in Table 2.7 where the last row of the table is the sum of the columns.

<table>
<thead>
<tr>
<th></th>
<th>Terrible</th>
<th>Poor</th>
<th>Fair</th>
<th>Good</th>
<th>Excellent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Because the sum of these dichotomies results in an ordered response variable with five levels, it is appropriate to consider a response mean outcome which takes discrete values 0 to 4. Using this summary measure reduces the 16 possible responses (4 dichotomies x 4 visits) to 4 ordinal scores (one for each visit). In this way one avoids concerns about multiplicity. At each of the four visits, differences in ordinal response means can be assessed for each treatment group in an unadjusted, adjusted, stratified with no covariate adjustment, or stratified and covariate adjusted model.

The results in Table 2.8 are based on mean response coding using a multivariate model which simultaneously applies to all four visits (see A.4.1, A.8.4 for multivariate response). The treatment parameter estimate for each visit can be interpreted as the difference in mean response between treatment and placebo from the linear model for ordered response scores. In the multivariate setting where the dimensions of the covariance matrix will be larger, there is more concern about appropriate sample size so MH weights are typically used.
In Table 2.8, stratifying on Center had virtually no effect on the estimated treatment parameter or reducing the variance relative to the unstratified model. However, adjusting for Age, Gender, and Baseline in both the stratified and unstratified models led to an increase in the magnitude of the treatment parameter and a decrease in the variance estimate. For Visit 1, this covariate adjustment changed a marginally significant test of treatment \((p = 0.056\) unstratified, \(p = 0.049\) stratified) to a clearly significant test of treatment \((p = 0.010\) unstratified, \(p = 0.003\) stratified).

Table 2.8 Results for Treatment Parameter (Difference in Means) From Linear Model for Ordered Response Scores

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.400</td>
<td>0.989</td>
<td>0.827</td>
<td>0.625</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.209</td>
<td>0.222</td>
<td>0.244</td>
<td>0.254</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.056</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.014</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.426</td>
<td>1.013</td>
<td>0.873</td>
<td>0.661</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.165</td>
<td>0.199</td>
<td>0.223</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.010</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>0.394</td>
<td>0.986</td>
<td>0.826</td>
<td>0.619</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.200</td>
<td>0.220</td>
<td>0.244</td>
<td>0.249</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.049</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.013</td>
</tr>
<tr>
<td>4</td>
<td>Treatment</td>
<td>0.460</td>
<td>1.062</td>
<td>0.903</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.157</td>
<td>0.193</td>
<td>0.220</td>
<td>0.229</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.003</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Model Identification:
1) Unadjusted treatment effect - no stratification or covariate adjustment
2) Covariate adjustment (Age, Gender, Center, Baseline)
3) Stratify on Center, No Covariate Adjustment, Method 2 using MH weights
4) Stratify on Center, Adjust for 3 cov. (Gender, Age, Baseline), Method 2, MH weights
Although the treatment x visit interaction is not significant for differences in proportion of favorable response (see Table 2.6), the treatment effect does not appear to be homogeneous across visits when the ordinal score is the outcome of interest in Table 2.9. It would be possible to identify which of the response dichotomies is the source for this interaction.

Table 2.9. Assessing Treatment x Visit Interaction for Four Different Models Where Ordinal Score at Each Visit is the Outcome of Interest

<table>
<thead>
<tr>
<th>Model</th>
<th>Wald Chisquare</th>
<th>degrees of freedom</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.8</td>
<td>3</td>
<td>0.008</td>
</tr>
<tr>
<td>2</td>
<td>12.7</td>
<td>3</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>12.2</td>
<td>3</td>
<td>0.007</td>
</tr>
<tr>
<td>4</td>
<td>13.4</td>
<td>3</td>
<td>0.004</td>
</tr>
</tbody>
</table>

"Model" corresponds to the specified models found in Table 2.8

It would also be possible to assess covariate x treatment interaction for response means as was done for differences in proportion of favorable response (Section 2.4 of this chapter).

2.6 Difference in Means Parameterization

An alternative to modeling response means is to model the average difference between treatment groups outright. This method will be discussed in the setting of a treatment comparison when there are two treatment groups. One creates a difference vector between the two groups which is equivalent to subtracting the mean response and mean covariate values for the placebo group from the mean response and mean covariate values for the treatment group. Because only differences between means are estimated, (and not placebo means and treatment group differences, as in the response means parameterization), fewer parameters need to be estimated. In cases with smaller sample sizes and/or a large number of covariates, this difference between means parameterization
may be more stable. See A.5 for a complete explanation of this difference between means parameterization.

2.7 SAS Catmod Procedure Approximation

All parameter estimates and standard errors in this chapter have been obtained by programming in SAS IML (SAS 90). However, the Catmod procedure in SAS can be used to closely approximate the results in this chapter. Details of model specification and a comparison of the two methods can be seen in A.11.a.

2.8 Nonparametric Covariance Matrix

Although the approach in this chapter is primarily nonparametric, it could be more purely nonparametric if a randomization covariance matrix \( V_o \) were used. A randomization covariance matrix is calculated for the pooled sample instead of for each treatment group separately. It is called the randomization covariance matrix because, under randomization, the means and variances of the two treatment groups are considered to be equal. The randomization covariance matrix is to some extent considered a fixed quantity under \( H_o \) (see A.2.2 for a description of its structure). In addition, see Koch et al 1990, for a more in depth discussion of randomization covariance matrices with regard to this particular clinical example.

One of the benefits of the randomization covariance matrix is that one can work with a smaller sample size because the covariance matrix is based on the pooled sample. A drawback to using \( V_o \) is that only test statistics are obtained and not true confidence interval estimates of treatment effects. Further illustrations of the use of \( V_o \) will be explored in subsequent chapters.
Chapter 3
Modeling Logit Response With Linear Covariate Adjustment

In this chapter we will continue to evaluate the clinical trial that was used as an example in Chapter 2. In the previous chapter we explored differences in proportion of favorable outcome for the two treatment groups at each of four visits. Favorable outcome has been defined as a dichotomous response. If score equals 3 or 4 then favorable response equals 1 for that respective visit; otherwise favorable response equals 0 for that visit. The previous chapter emphasized response means (proportions or ordinal scores) for the two groups in the unadjusted and various stratified and covariate adjusted cases.

Another obvious criterion to consider for dichotomous response is the logit of the response. If \( \bar{y}_j \) is the proportion of favorable response for the \( j^{th} \) treatment group for a given visit, then the corresponding logit is defined as: \( \text{logit}(\bar{y}_j) = \log \left\{ \frac{\bar{y}_j}{1 - \bar{y}_j} \right\} \). The logistic regression model can be used to assess the effect of treatment on the response logit, either by itself or by adjusting for various covariates. We will evaluate three different models which assess treatment effect at each visit. (1) treatment alone - no covariable adjustment, (2) treatment adjusting for center as a covariable, and (3) treatment adjusting for center, gender, age, and baseline status in the model.

3.1 Traditional Logistic Regression

When considering Model 1 (treatment only, no covariate adjustment), one must make the assumption that patients from each treatment group represent a simple random
sample from an infinite population. The resulting odds ratio for treatment can be thought of as a population average odds ratio, and it can be interpreted as the odds of favorable response for the treatment group relative to the placebo group at a specific visit. This odds ratio can be thought of as being applicable to the entire population that is representative of those randomized to treatment or placebo.

However, when one chooses to adjust for covariates in a logistic regression model (see Models 2, 3 in Table 3.1), additional assumptions are required. Patients must now be considered representative of a stratified simple random sample where stratification can be thought of as a cross-classification of subjects based on their explanatory values. This stratified random sample assumption is not based on stratification as a design component of the study but instead one conditions on an individual's covariate values. The resulting treatment odds ratio from a covariate adjusted model is a subpopulation specific odds ratio which applies only to those who share the same values of explanatory variables.

In addition to assuming a stratified random sample, other assumptions must be made for a covariate adjusted logistic model. It is assumed that there is no interaction of covariates with treatment or with other covariates. In addition, it is also assumed that covariates in the model are being specified in their proper form. For example, if age is in the logistic model, we assume that the linear form of this covariate is appropriate, and no higher order covariate is necessary. Another way to express these additional two assumptions is to say that the covariate adjusted model is correctly specified. These additional assumptions can cause concern about the appropriateness of the logistic model. For this particular example, treating Center as a covariate might not be entirely appropriate. By doing so, we are conditioning on Center in the logistic model when in fact it is a stratification factor. However, traditional logistic regression only allows for treating stratification factors and covariates in the same manner as explanatory variables.
The logistic model which is applied to dichotomous responses has the following form:

\[ \theta_{hi} = \frac{\exp\{\alpha + \sum_{k=1}^{t} \beta_k x_{hik}\}}{1 + \exp\{\alpha + \sum_{k=1}^{t} \beta_k x_{hik}\}} \]

where the quantity \( \alpha \) is the intercept parameter, the \( \{x_{hik}\} \) are the \( t \) explanatory variables for the \( h \)th stratum and \( i \)th treatment; \( k = 1, 2, ..., t \); and the \( \{\beta_k\} \) are the \( t \) regression parameters. The odds of an event for the \( h \)th group is:

\[ \frac{\theta_h}{1-\theta_h} = \exp\{\alpha + \sum_{k=1}^{t} \beta_k X_{hik}\} \]

By taking the natural log of both sides, a linear model is obtained for the logit:

\[ \log\left\{\frac{\theta_h}{1-\theta_h}\right\} = \alpha + \sum_{k=1}^{t} \beta_k X_{hik} \]

The logit is the log of an odds. One can obtain a model-predicted odds ratio by exponentiating model parameter estimates.

The results of Table 3.1 were obtained by using the logistic procedure in SAS (SAS 92). Estimates of treatment effect are obtained by maximum likelihood (ML) methods. The treatment parameter estimate for Model 1, Visit 1 is 0.813. One can exponentiate the parameter estimate to obtain an odds ratio. The population average odds ratio for treatment at Visit 1 is \( 2.25 = \exp(0.813) \). Those receiving treatment have 2.25 times higher odds of favorable response than the placebo group at Visit 1. This treatment effect is marginally significant (\( p = 0.040 \)). In general, Visits 2 and 3 seemed to show the strongest treatment effect. Comparing Model 2 to Model 1 for each visit we see that conditioning on Center yields a larger treatment parameter estimate. In addition, the standard errors also increased when adjusting for Center. Model 3, which adjusts for four explanatory variables, provides even larger parameter and standard error estimates relative to Models 1 and 2.
Table 3.1 Results of Maximum Likelihood Logistic Regression, Modeling Favorable Response (Yes/No) at Each Visit

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter*</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.813</td>
<td>1.329</td>
<td>1.131</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.395</td>
<td>0.403</td>
<td>0.403</td>
<td>0.386</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.040</td>
<td>0.001</td>
<td>0.001</td>
<td>0.070</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.859</td>
<td>1.361</td>
<td>1.150</td>
<td>0.758</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.410</td>
<td>0.411</td>
<td>0.409</td>
<td>0.407</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.036</td>
<td>0.001</td>
<td>0.005</td>
<td>0.063</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>1.311</td>
<td>1.582</td>
<td>1.464</td>
<td>0.987</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.530</td>
<td>0.461</td>
<td>0.489</td>
<td>0.469</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.013</td>
<td>0.001</td>
<td>0.003</td>
<td>0.035</td>
</tr>
</tbody>
</table>

* Parameter is the log-odds for treatment

Models: (1) Treatment alone
        (2) Treatment adjusting for Center
        (3) Treatment adjusting for Age, Gender, Center, Baseline

It is interesting to recall that in the previous chapter, we saw that in evaluating differences in mean response for this same clinical trial example, adjusting for covariates caused the treatment parameter to increase slightly and the variance estimate to decrease. Variance reduction, which provides an increase in precision, is generally expected as a benefit of analysis of covariance. However, this benefit is not directly obtained for logistic models which adjust for covariates as in Table 3.1. It should also be noted for Table 3.1 that, although covariate adjustment does not lead directly to variance reduction, the parameter estimate increase is larger than the increase in the standard error. This leads to a net result of a more significant test of treatment; that is, the p-values are smaller for the covariate adjusted models.
3.2 Logit Response With Linear Covariate Adjustment

One can consider a possible logistic-like analysis where the emphasis is on addressing the population average treatment parameter for the response logit. This analysis can be performed for the unadjusted estimate of treatment effect or the covariate adjusted treatment estimate. The methods employed are similar to those described in the previous chapter. A response mean vector is created for each treatment group where covariates to be used in the analysis are also considered random variables along with the response of interest. The outcome of interest is the mean response logit instead of the mean proportion (or score) as in the previous chapter (see A.1.2 for response function structure). Using a Taylor's series approximation, an estimated covariance matrix can be obtained for the response logit and covariates (see A.2.1.b). Design matrices are used which can either allow the covariates to vary randomly between the two groups or one can restrict the difference in covariate means to be equal to zero (See A.3.1, A.3.2 for specific modeling and estimation details).

The treatment parameter in Table 3.2, Model 1 corresponds to the population average treatment effect for each visit and is estimated with weighted least squares methodology. This unadjusted parameter estimate (Model 1) for each visit is the exact same parameter estimate obtained from the corresponding maximum likelihood estimation (MLE) from the logistic model in SAS (Table 3.1, Model 1). However, the corresponding standard error estimate in Table 3.2 using weighted least squares is slightly larger relative to its MLE counterpart from Proc Logistic (± 0.004) in Table 3.1, Model 1. Table 3.2, Model 2, which adjusts for random imbalance between Centers, results in a slightly reduced estimate of treatment effect. This may indicate that the random imbalance of centers favored treatment slightly so adjusting for center effect reduces the treatment parameter estimate.
Table 3.2 Response Logit With Linear Covariate Adjustment, Modeling Favorable Response (Yes/No) at Each Visit, Using Weighted Least Squares Methodology

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter*</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.813</td>
<td>1.329</td>
<td>1.131</td>
<td>0.699</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.399</td>
<td>0.407</td>
<td>0.407</td>
<td>0.390</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.041</td>
<td>0.001</td>
<td>0.005</td>
<td>0.073</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.803</td>
<td>1.323</td>
<td>1.126</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.384</td>
<td>0.401</td>
<td>0.403</td>
<td>0.371</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.037</td>
<td>0.001</td>
<td>0.005</td>
<td>0.064</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>0.820</td>
<td>1.375</td>
<td>1.175</td>
<td>0.758</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.322</td>
<td>0.380</td>
<td>0.359</td>
<td>0.339</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.011</td>
<td>0.0002</td>
<td>0.001</td>
<td>0.026</td>
</tr>
</tbody>
</table>

* Parameter is the log-odds for treatment relative to placebo

Models: (1) Treatment alone
(2) Treatment adjusting for center
(3) Treatment adjusting for age, gender, center, baseline

It is not surprising that adjusting for center has little effect on the treatment parameter estimate because treatment randomization was allocated in blocks within each center so random imbalance between centers should be minimal. What is interesting to note is that there is a slight amount of variance reduction in Model 2 relative to Model 1. On average, across the four visits, variance was reduced by 3%. Model 3 provides a larger estimate of treatment effect relative to the unadjusted estimate. There is also greater variance reduction relative to Models 1 and 2 after adjusting for random imbalance for Age, Gender, Center, and Baseline. This is an expected benefit of covariance adjustment. On average, the standard error was reduced by 13% in Model 3 (4 covariate adjustment) relative to Model 1 which has no covariate adjustment (See A.6 for variance reduction and sample size increase formulas). This added precision due to variance reduction is comparable to a gain in sample size of 38%. An increased treatment parameter estimate and a decrease in the standard error leads to a more significant test of
treatment. This is the same trend that was seen in the previous chapter where it was shown that random imbalance of baseline covariates favors placebo in this particular clinical trial so adjustment for covariates leads to a larger estimate of treatment effect.

3.2.1 Parameter Interpretation

The pattern of comparisons for treatment significance is similar for Tables 3.1 and 3.2. For both tables treatment was found to be more significant for visits 2 and 3 relative to visits 1 and 4, and adjustment for covariates leads to smaller p-values for tests of treatment relative to the unadjusted tests of treatment. However, the mechanisms for the two tables are different. In Table 3.1, covariance adjustment provides an increase in the treatment parameter estimate and a small increase in the corresponding standard error estimate relative to the unadjusted estimate. On the other hand, Table 3.2 shows that covariate adjustment provides somewhat larger (unconditional) treatment parameter estimates and smaller standard error estimates relative to their unadjusted counterparts. The impact is the same for both types of models (nearly identical p-values), but the mechanism is different. Additionally, the treatment parameters for the two models have different interpretations. For the logistic model with a linear covariance adjusted treatment parameter estimate, the parameter is interpreted as a population average treatment estimate. The traditional logistic regression covariate adjusted treatment parameter is a subpopulation average parameter, and this structure for subpopulations changes depending on which covariates are in the model. The subpopulation average odds ratio applies only to patients through subsets which share the same values of the covariates as conditioned upon in the model. In general, the significance level of both types of methods agree with one another. This may be an indication that the ML logistic regression model fits the data adequately, and the logistic model with linear covariance adjustment provides an assessment of random imbalance and variance reduction.
3.3 Stratified Logit Response With Covariate Adjustment

In Table 3.1, Center was managed as a covariate because traditional logistic regression treats both covariates and stratification factors as explanatory variables. In Table 3.2 Center was treated as a covariate and the adjusted estimate of treatment involved restricting differences in centers between treatment groups to be equivalent. Handling Center as a covariate may be simpler computationally, but it is not really compatible with the study design. Treating Center as a covariate assumes that the centers are random when actually they are known, controlled factors. It is more appropriate, if sample size allows, to treat Center as a stratification factor. Stratified analyses can be performed by either (1) combining weighted covariate and response means data across strata and then performing an adjusted analysis on the combined strata data ($\beta_w$), or (2) conduct a covariate adjusted analysis within each stratum and then use a weighted combination of the resulting estimated parameters ($\bar{\beta}$). This latter method is used when sample size per stratum is fairly large. (See A.8.1, A.8.2 for a more detailed description of these stratification methods).

Table 3.3 shows the results of analyses when Center is treated as a stratification factor and there is no covariate adjustment (Model 1), or Center is treated as a stratification factor and the covariates Age, Gender, and Baseline status are adjusted to equivalence between treatment groups (Model 2). For the results in Table 3.3, covariance adjustment is performed within each of the two centers separately and then estimates are combined using Mantel-Haenszel weights ($\bar{\beta}$ method).

The stratified results in Table 3.3 are similar to those found in Table 3.2. Model 2, which stratifies on Center and adjusts for 3 covariates, provides larger treatment parameter estimates, and on average, the standard error estimates are reduced by 26% relative to the stratified model with no covariate adjustment (Model 1).
Table 3.3 Modeling Response Logit Using WLS With Stratification and Linear Covariate Adjustment, Modeling Favorable Response (Yes/No) at Each Visit

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter*</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.836</td>
<td>1.347</td>
<td>1.142</td>
<td>0.756</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.426</td>
<td>0.422</td>
<td>0.419</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.050</td>
<td>0.001</td>
<td>0.006</td>
<td>0.069</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.852</td>
<td>1.420</td>
<td>1.229</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.316</td>
<td>0.385</td>
<td>0.363</td>
<td>0.376</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.007</td>
<td>0.0002</td>
<td>0.0007</td>
<td>0.020</td>
</tr>
</tbody>
</table>

* Parameter is difference in mean response logits between treatment groups (log-odds)
Models: (1) Treatment, stratifying on center (\(\bar{\beta}\) method)
(2) Treatment, stratifying on center (\(\bar{\beta}\)) and adjusting for age, gender, baseline status

3.4 Traditional Proportional Odds Model - Cumulative Logits

In addition to modeling dichotomous response with one logit, it is possible to fit a proportional odds model to the ordinal score for each visit. Because the score can take five possible values (0, 1, 2, 3, or 4), the proportional odds model can be fit to four response logits which is equivalent to the number of possible dichotomous splits. However, due to sparse numbers of patients reporting a score of 0 or 1 (poor response), especially for the latter visits, these two categories were combined (0-1, 2, 3, 4) and three logits were evaluated for the data.

The assumptions for the proportional odds model are the same as for the logistic regression model (see section 3.1) with one additional assumption required. It is assumed that the regression parameters (\(\beta_k\)) for each of the \(k\) logits are equivalent (\(\beta_1 = \beta_2 = \ldots = \beta_k\)) with only an intercept \(\alpha_k\) differentiating the relationships across types of logits. In other words, the regression lines for the different cumulative log odds are parallel to each other, for a single continuous explanatory variable, differing only by the difference between the values of the intercept parameter.
Table 3.4 Results of ML Proportional Odds, Modeling Ordinal Response Score at Each Visit

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter*</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.602</td>
<td>1.414</td>
<td>1.071</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.347</td>
<td>0.365</td>
<td>0.355</td>
<td>0.348</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.082</td>
<td>0.0001</td>
<td>0.003</td>
<td>0.052</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.685</td>
<td>1.486</td>
<td>1.072</td>
<td>0.727</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.351</td>
<td>0.369</td>
<td>0.355</td>
<td>0.354</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.051</td>
<td>0.0001</td>
<td>0.003</td>
<td>0.040</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>0.970</td>
<td>1.726</td>
<td>1.269</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.386</td>
<td>0.400</td>
<td>0.380</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.012</td>
<td>0.0001</td>
<td>0.0008</td>
<td>0.024</td>
</tr>
</tbody>
</table>

* Parameter is the difference in mean response cumulative logits between treatment and placebo, scores 0-1 are combined into one group.

Models: (1) Treatment alone
(2) Treatment adjusting for center
(3) Treatment adjusting for age, gender, center, baseline

The results for the proportional odds model are obtained by using the logistic procedure in SAS (SAS 92). These estimates, like those for the logistic model, are obtained by maximum likelihood methods. The treatment parameter estimate for Model 1, Visit 1 is 0.602. This can be interpreted as the log-odds for treatment versus placebo in the cumulative logit for categories above a score (more favorable) relative to the scores directly below it (less favorable). The corresponding population average treatment odds ratio is $1.83 = \exp(0.602)$. The p-value = 0.082 is suggestive of a treatment benefit, but it is not statistically significant. Like the maximum likelihood logistic models (Table 3.1), adjustment for covariates causes the estimated treatment parameter to increase, and so does its standard error.
Table 3.5  SAS Logistic Procedure, Test of Proportional Odds Assumption for Each Visit, Modeling Three Cumulative Logits

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Score Chi-square</td>
<td>2.445</td>
<td>0.273</td>
<td>0.969</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td>df</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>pvalue</td>
<td>0.295</td>
<td>0.872</td>
<td>0.616</td>
<td>0.956</td>
</tr>
<tr>
<td>2</td>
<td>Score Chi-square</td>
<td>3.712</td>
<td>0.599</td>
<td>2.357</td>
<td>9.730</td>
</tr>
<tr>
<td></td>
<td>df</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>pvalue</td>
<td>0.446</td>
<td>0.963</td>
<td>0.670</td>
<td>0.045</td>
</tr>
<tr>
<td>3</td>
<td>Score Chi-square</td>
<td>12.178</td>
<td>7.935</td>
<td>13.289</td>
<td>25.578</td>
</tr>
<tr>
<td></td>
<td>df</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>pvalue</td>
<td>0.273</td>
<td>0.635</td>
<td>0.208</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Models: 
1. P. odds assumption with only treatment (no covariates) in the model
2. P. odds assumption with center and treatment in the model
3. P. odds assumption with treatment, age, gender, center, baseline in model

The SAS proportional odds goodness of fit test checks whether there is a common parameter vector $\beta$ instead of distinct $\beta_k$ for each cumulative logit. The hypothesis can be stated as $\beta_k = \beta$ for all $k$. Rejecting the null hypothesis rejects the assumption of proportional odds and a different type of model needs to be considered. If the null hypothesis is not rejected, then the assumption of proportional odds is compatible with the data. The test is comparing $t$ parameters for the explanatory variables across $(r - 1)$ logits, where $r$ is the number of response levels, it has $t*(r - 2)$ df (Stokes et al 95).

For Model 3 which adjusts for 4 covariates plus treatment ($t = 5$) and has 4 response levels 0-1, 2, 3, 4 ($r - 2 = 2$), there are 10 degrees of freedom for the proportional odds test. The results of this test can be found in Table 3.5. The sample size requirements for this proportional odds test are moderately demanding; one needs at least 5 observations at each outcome at each level of each main effect (Stokes et al 95).
is some indication that, for Visit 4, the covariate adjusted proportional odds models (Models 2, 3) do not have a particularly good fit.

3.5 Proportional Odds With Linear Covariate Adjustment

The results of the proportional odds models in Table 3.6 are obtained by weighted least squares methodology (see A.1.3 for response function structure and A.2.1.c for estimated covariance structure). As in the logistic model with linear covariate adjustment (Table 3.2), the difference in mean covariate values between treatment groups is constrained to equal zero for the respective covariates in Models 2 and 3 (see A.4.1.a, A.4.1.b). The parameter estimates and standard errors for Model 1 in Table 3.6 are comparable to those for Model 1 using maximum likelihood estimation methods in Table 3.4. However, the covariate adjusted Models 2 and 3 of Table 3.6 provide fairly different parameter estimates than the same models in Table 3.4. With linear covariate adjustment the variance is reduced and the treatment parameter estimate increases by only small amounts.

Table 3.6 Results of Proportional Odds With Linear Covariate Adjustment at Each Visit Fitting Three Cumulative Logits

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter*</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.588</td>
<td>1.410</td>
<td>1.060</td>
<td>0.676</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.351</td>
<td>0.368</td>
<td>0.359</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.094</td>
<td>0.0001</td>
<td>0.003</td>
<td>0.054</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.581</td>
<td>1.402</td>
<td>1.058</td>
<td>0.674</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.335</td>
<td>0.362</td>
<td>0.356</td>
<td>0.335</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.083</td>
<td>0.0001</td>
<td>0.003</td>
<td>0.044</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>0.657</td>
<td>1.473</td>
<td>1.152</td>
<td>0.758</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.278</td>
<td>0.327</td>
<td>0.317</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.018</td>
<td>&lt;0.0001</td>
<td>0.0003</td>
<td>0.013</td>
</tr>
</tbody>
</table>

* Parameter is difference in mean cumulative response logits between treatment groups

Models: (1) Treatment alone
(2) Treatment adjusting for center
(3) Treatment adjusting for age, gender, center, baseline

80
Table 3.6 provides insight into the mechanism of adjustment for random imbalance and variance reduction which is not directly observable in the ML estimates in Table 3.4. The important thing to note is that both methods, despite differences in parameter estimates and standard errors, show agreement in level of statistical significance. The corresponding p-values for Tables 3.4 and 3.6 are very similar.

The results of Table 3.7 indicate that a proportional odds model fits the data adequately with or without covariate adjustment for each of the four visits (each p-value $> 0.25$). It is interesting to remember that this was not the case for the ML proportional odds model tests in Table 3.5. They indicated that the models which conditioned on covariates (Models 2, 3) did not have an appropriate fit for proportional odds for the data at Visit 4. (See A.4.3.b for an explanation of the test of the proportional odds assumption for these models with linear covariate adjustment).

**Table 3.7 Test of Proportional Odds Assumption for Each Visit, Fitting Three Cumulative Logits**

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wald Chi-square</td>
<td>2.322</td>
<td>0.267</td>
<td>0.940</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>df</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>pvalue</td>
<td>0.313</td>
<td>0.875</td>
<td>0.625</td>
<td>0.957</td>
</tr>
<tr>
<td>2</td>
<td>Wald Chi-square</td>
<td>2.438</td>
<td>0.421</td>
<td>0.634</td>
<td>0.144</td>
</tr>
<tr>
<td></td>
<td>df</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>pvalue</td>
<td>0.296</td>
<td>0.810</td>
<td>0.729</td>
<td>0.930</td>
</tr>
</tbody>
</table>

Models: (1) P. odds assumption with no linear covariate adjustment
(2) P. odds assumption adjusting for linear age, gender, center, baseline

We saw from the previous chapter where mean response is the outcome of interest that constraining the difference in mean covariate values to be equal to zero is appropriate for this clinical trial example (Chapter 2.2.2), i.e., the goodness of fit test is adequate.
This test of covariate imbalance is equivalent regardless of whether the response of interest is on the linear or logit scale. The covariates remain on the linear scale regardless of the scale of the response.

The full goodness-of-fit test for the linear models with proportional odds is comprised of two parts: (1) equality of covariates between treatment groups, and (2) the proportional odds assumption. It is possible to simultaneously test both of these assumptions in one GOF test (see A.4.3.c for further specifications).

3.6 Stratified Proportional Odds With Covariate Adjustment

As previously discussed in the earlier part of this chapter, the traditional logistic and proportional odds models handle true stratification factors as explanatory variables in the model. In a case where centers are directly related to the randomization process, the factor should be treated as a stratification factor. For the results shown in Table 3.8, two methods of stratification are employed. The first method involves combining the response and covariate mean vectors for treatment groups across the strata by use of weights (in this case Mantel-Haenszel weights (MH)). Covariate adjusted analysis is then performed on the combined strata adjusted mean vector (\( \beta_w \) method). The other method (\( \bar{\beta} \)) involves performing covariate adjustment within each stratum separately and then combining parameter estimates with appropriately selected weights (MH weights). For a discussion of these two stratification methods and the choice of various weighting schemes, see A.8.1, A.8.2, A.8.4.

Both stratification methods provide essentially the same conclusions; however, the second method (\( \bar{\beta} \)) tends to provide a more statistically significant test of treatment effect relative to its first method counterpart (\( \beta_w \)) for this particular trial. For example, the unadjusted test of treatment effect for Visit 1 with the \( \beta_w \) method provides a p-
value = 0.095 where the same unadjusted test for $\bar{\beta}$ method provides a p-value = 0.041.

The covariate adjusted counterparts of these same models provide $p = 0.016$ and $p = 0.001$, respectively.

Table 3.8 Results of Proportional Odds With Linear Covariate Adjustment at Each Visit, Stratifying on Center Using Two Different Methods

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter*</th>
<th>Visit 1</th>
<th>Visit 2</th>
<th>Visit 3</th>
<th>Visit 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Treatment</td>
<td>0.612</td>
<td>1.442</td>
<td>1.138</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.366</td>
<td>0.379</td>
<td>0.360</td>
<td>0.369</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.095</td>
<td>0.0001</td>
<td>0.002</td>
<td>0.061</td>
</tr>
<tr>
<td>2</td>
<td>Treatment</td>
<td>0.681</td>
<td>1.589</td>
<td>1.499</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.282</td>
<td>0.342</td>
<td>0.306</td>
<td>0.331</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.016</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>3</td>
<td>Treatment</td>
<td>0.725</td>
<td>1.498</td>
<td>1.113</td>
<td>0.752</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.355</td>
<td>0.377</td>
<td>0.356</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.041</td>
<td>&lt;0.0001</td>
<td>0.002</td>
<td>0.038</td>
</tr>
<tr>
<td>4</td>
<td>Treatment</td>
<td>0.846</td>
<td>1.669</td>
<td>1.488</td>
<td>1.084</td>
</tr>
<tr>
<td></td>
<td>S.E.</td>
<td>0.262</td>
<td>0.334</td>
<td>0.294</td>
<td>0.320</td>
</tr>
<tr>
<td></td>
<td>Pvalue</td>
<td>0.001</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

* Parameter is difference in mean response cumulative logits between treatment groups

Models: 1: Treatment, stratify on Center, $\beta_w$ method

2: Treatment adjusting for Age, Gender, Baseline ($\beta_w$ method)

3: Treatment, stratify on Center, $\bar{\beta}$ method

4: Treatment adjusting for Age, Gender, Baseline ($\bar{\beta}$ method)

With both methods one can see the point estimate increase with covariate adjustment and the variance decreases providing an overall more significant test of treatment effect after adjusting for covariates. Stratification with $\beta_w$ method, on average, provides a reduction in variance of 27% while the $\bar{\beta}$ stratification method reduces variance 30% relative to the counterpart model with no covariate adjustment.
For stratified logit response results in this chapter (Tables 3.3, 3.8), logits were formed within each stratum (Center) separately and then combined across strata. Because functions of means are taken within each stratum, a larger sample size is needed for each stratum in a manner similar to what is required for stratification method 2 ($\beta$). The treatment odds ratio is no longer interpreted as a population average estimate but instead the odds ratio is conditional on Center. The odds ratio only applies to the population average which corresponds to a given Center, i.e., a stratum specific population average odds ratio. A more stable method of estimating logits would be to form means within strata, combine the means across strata, and then create logits based on Center stratified means. The resulting treatment parameter estimates are not conditional on the stratum; they are the population average odds ratios.

Table 3.9  Test of Proportional Odds Assumption for Each Visit Based on a Center

<table>
<thead>
<tr>
<th>Stratified Model Using $\beta$ Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Models: (1) P. odds assumption with center stratification and no covariate adjustment
(2) P. odds assumption with center stratification and adjustment for age, gender, and baseline status

Like the unstratified case, the center stratified proportional odds model for each visit seems to fit the data adequately. This is true for the unadjusted and covariate adjusted models. See Table 3.9 for results.
Linear covariate adjustment for assessing treatment effect with logit response provides insight into the robustness of assumptions required for logistic regression and proportional odds models. Because the significance level of both methods tended to agree with one another, there is an indication that the traditional log-linear model assumptions are probably compatible with these data. By using weighted least squares, it is easier to identify the components of bias adjustment and variance reduction due to covariate adjustment. It is also possible to perform a stratified analysis which is consistent with the design of the clinical trial. Confirmation of significance level (similar magnitude of p-values across both methods) provides some indication of comfort with the traditional logistic model assumptions.
Chapter 4
Logrank and Wilcoxon Hypothesis Testing

In clinical trials where time-to-event is the outcome of interest, it is traditional to employ methods such as Kaplan-Meier estimation of survival curves, (stratified) logrank or Wilcoxon tests to assess differences in survival distributions, and Cox modeling to assess treatment effect while adjusting for the effects of covariates. However, these methods require certain assumptions which can have uncertain applicability. In addition, covariate adjusted analysis using the Cox model can provide results which are unexpected and difficult to interpret relative to the model with no covariance adjustment.

This chapter intends to show, through the use of a randomized clinical trial example, the results of traditional time-to-event analyses. Additional complementary nonparametric methods will then be proposed. The results of these latter analyses should help shed light on the behavior for methods of traditional analysis of covariance in the survival setting.

4.1 Introduction of Clinical Trial Example

Amyotrophic lateral sclerosis (ALS), also commonly known as Lou Gehrig's disease, is a progressive and fatal motor neuron disease for which there is no adequate treatment (Bensiman 94). In this example 959 patients with ALS were included in a double-masked clinical trial involving the drug riluzole. There were 4 treatment arms:
placebo, 50 mg, 100 mg, and 200 mg. Time until death or tracheostomy, which will generically be referred to as "survival", was the response of interest. Patient follow-up was truncated at 18 months. Pairwise comparisons of the three doses with placebo for survival were of interest. Because a previous pilot study used a 100 mg. dose which showed promising results relative to placebo, the comparison of 100/200 mg. combined versus placebo was of particular interest for this larger scale trial. Assessment of linear trend among the doses was also identified as a major objective.

Twenty-one covariates were specified a priori in the protocol or analysis plan for this study. These covariates included measures of baseline vital status, baseline measures of disease status, age, gender, disease onset information, neurologic function, and measures of family history for neurologic disorders. There was also one stratification factor which had two levels that corresponded to disease status at randomization: limb versus bulbar.

The pre-specified analysis plan called for the use of logrank statistics to compare survival distributions for the treatments. After adjustment for the effect of covariates, the treatment effects were estimated using a Cox model. Table 4.1 shows the average values of the 21 covariates and stratification factor for the treatment comparison of the 100/200 dose combination versus placebo. Individuals receiving Dose 50 are excluded from this table. As one would expect by chance with randomization, 2 p-values were less than 0.05, and 4 p-values were less than 0.10.
Table 4.1 Means of the Stratification Factor and 21 Covariates and Wald Test p-values for Comparisons Between the Placebo Group and The Dose 100/200 Combined Group

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Placebo Mean</th>
<th>Dose 100/200 Mean</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulbar Form at Randomization (strat factor)</td>
<td>0.31</td>
<td>0.30</td>
<td>0.85</td>
</tr>
<tr>
<td>(1=Bulbar, 0=Limb)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Muscle Testing Total Score</td>
<td>87.9</td>
<td>89.0</td>
<td>0.44</td>
</tr>
<tr>
<td>Vital Capacity Ratio</td>
<td>87.6</td>
<td>88.5</td>
<td>0.64</td>
</tr>
<tr>
<td>Age at Randomization (in years)</td>
<td>56.0</td>
<td>56.9</td>
<td>0.32</td>
</tr>
<tr>
<td>Disease Duration (in years)</td>
<td>1.83</td>
<td>1.68</td>
<td>0.33</td>
</tr>
<tr>
<td>Visual Analogue Scale Fatigue</td>
<td>43.3</td>
<td>44.5</td>
<td>0.19</td>
</tr>
<tr>
<td>Visual Analogue Scale Stiffness</td>
<td>27.6</td>
<td>26.8</td>
<td>0.45</td>
</tr>
<tr>
<td>Visual Analogue Scale - Cramping</td>
<td>29.4</td>
<td>26.1</td>
<td>0.09</td>
</tr>
<tr>
<td>Visual Analogue Scale - Fasciculation</td>
<td>39.5</td>
<td>42.0</td>
<td>0.57</td>
</tr>
<tr>
<td>Bulbar and Limb status at onset</td>
<td>0.38</td>
<td>0.31</td>
<td>0.03</td>
</tr>
<tr>
<td>Severity of Illness</td>
<td>3.98</td>
<td>3.98</td>
<td>1.00</td>
</tr>
<tr>
<td>Neurologic Diagnosis: definite vs probable</td>
<td>0.50</td>
<td>0.46</td>
<td>0.22</td>
</tr>
<tr>
<td>Diastolic Blood Pressure</td>
<td>80.2</td>
<td>81.4</td>
<td>0.16</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>132.1</td>
<td>134.3</td>
<td>0.05</td>
</tr>
<tr>
<td>FEV Ratio</td>
<td>85.1</td>
<td>86.4</td>
<td>0.57</td>
</tr>
<tr>
<td>Family Neurologic History of ALS</td>
<td>0.045</td>
<td>0.042</td>
<td>0.62</td>
</tr>
<tr>
<td>Height</td>
<td>168.6</td>
<td>168.4</td>
<td>0.45</td>
</tr>
<tr>
<td>Weight</td>
<td>68.1</td>
<td>68.1</td>
<td>0.59</td>
</tr>
<tr>
<td>Total Score Norris Bulbar Scale</td>
<td>30.3</td>
<td>31.0</td>
<td>0.37</td>
</tr>
<tr>
<td>Total Score Norris Limb Scale</td>
<td>44.1</td>
<td>44.7</td>
<td>0.30</td>
</tr>
<tr>
<td>Gender: 1=Male, 0=Female</td>
<td>0.63</td>
<td>0.61</td>
<td>0.22</td>
</tr>
<tr>
<td>Sitting Pulse</td>
<td>76.4</td>
<td>74.8</td>
<td>0.09</td>
</tr>
</tbody>
</table>

With so many covariates, there may be some concern about collinearity. Collinearity concerns relationships among predictor variables and does not directly involve the response variable. One informative way to examine collinearity is to consider what happens if each predictor variable is the response variable in a multiple regression model in which the independent variables are all the remaining predictors. The variance inflation factor is often used to measure collinearity in a multiple regression analysis. It
may be computed as \( VIF_c = \frac{1}{1-R_c^2} \) where \( c = 1, 2, \ldots m \) covariates and \( R_c^2 \) is the squared multiple correlation based on regressing \( x_c \) on the remaining \( m - 1 \) predictors. A rule of thumb for evaluating VIFs is to be concerned with any value larger than 10.0 which corresponds to \( R_c^2 > 0.90 \). Table 4.2 contains the VIF for each covariate based on the entire sample of patients. The results indicate that there is no problem with collinearity in this particular data set.

Table 4.2 Assessment of Collinearity Among ALS Covariates and Stratification Factor

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Variance Inflation Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulbar Form at Randomization (strat factor) (1=Bulbar, 0=Limb)</td>
<td>3.87</td>
</tr>
<tr>
<td>Muscle Testing Total Score</td>
<td>3.30</td>
</tr>
<tr>
<td>Vital Capacity Ratio</td>
<td>2.80</td>
</tr>
<tr>
<td>Age at Randomization (in years)</td>
<td>1.31</td>
</tr>
<tr>
<td>Disease Duration (in years)</td>
<td>1.16</td>
</tr>
<tr>
<td>Visual Analogue Scale Fatigue</td>
<td>1.31</td>
</tr>
<tr>
<td>Visual Analogue Scale Stiffness</td>
<td>1.35</td>
</tr>
<tr>
<td>Visual Analogue Scale - Cramping</td>
<td>1.21</td>
</tr>
<tr>
<td>Visual Analogue Scale - Fasciculation</td>
<td>1.24</td>
</tr>
<tr>
<td>Bulbar and Limb status at onset</td>
<td>1.70</td>
</tr>
<tr>
<td>Severity of Illness</td>
<td>1.76</td>
</tr>
<tr>
<td>Neurologic Diagnosis: definite vs probable</td>
<td>1.46</td>
</tr>
<tr>
<td>Diastolic Blood Pressure</td>
<td>1.66</td>
</tr>
<tr>
<td>Systolic Blood Pressure</td>
<td>1.72</td>
</tr>
<tr>
<td>FEV Ratio</td>
<td>2.54</td>
</tr>
<tr>
<td>Family Neurologic History of ALS</td>
<td>1.03</td>
</tr>
<tr>
<td>Height</td>
<td>2.49</td>
</tr>
<tr>
<td>Weight</td>
<td>1.81</td>
</tr>
<tr>
<td>Total Score Norris Bulbar Scale</td>
<td>3.01</td>
</tr>
<tr>
<td>Total Score Norris Limb Scale</td>
<td>3.81</td>
</tr>
<tr>
<td>Gender: 1=Male, 0=Female</td>
<td>1.99</td>
</tr>
<tr>
<td>Sitting Pulse</td>
<td>1.06</td>
</tr>
</tbody>
</table>
Figure 4.1 Survival Distribution By Dose of Drug

Follow-up Time in Days

DOSE
- 100 mg
- - - 200 mg
- - - 50 mg
- - - placebo
4.2 Kaplan-Meier Survival Estimates

Figure 4.1 displays the Kaplan-Meier curves for each of the four treatment groups. In addition, the 12 and 18 month estimated survival rates for each treatment group are displayed. Placebo appears to have a poorer survival rate than the active treatment groups, and this difference seems to be most pronounced at 12 months than at 18 months. The 100 mg. dose tends to have the best survival rates of the three test treatment groups.

4.3 Stratified Logrank and Wilcoxon Tests

Table 4.3 displays the results of logrank and Wilcoxon tests stratified on Limb/Bulbar status. The LIFETEST procedure in SAS (SAS 92) was used to obtain these test results, specifying STRATA as Limb/Bulbar status and TEST as the appropriate treatment indicator for the pairwise comparison of interest. There is a trend of some treatment effect in both the 100 mg and 200 mg doses, and combinations of doses relative to placebo, but the significance level is not particularly compelling. The Wilcoxon test tends to be slightly more significant than the protocol prespecified logrank test for each pairwise comparison. For each comparison in Table 4.3, only the subset of subjects randomized to the respective treatment groups of interest are included in each analysis.

<table>
<thead>
<tr>
<th>Dose Comparison</th>
<th>Wilcoxon p-value</th>
<th>Logrank p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg vs. Placebo</td>
<td>0.234</td>
<td>0.251</td>
</tr>
<tr>
<td>100 mg vs Placebo</td>
<td>0.050</td>
<td>0.076</td>
</tr>
<tr>
<td>200 mg vs. Placebo</td>
<td>0.061</td>
<td>0.075</td>
</tr>
<tr>
<td>100/200 vs. Placebo</td>
<td>0.026</td>
<td>0.036</td>
</tr>
<tr>
<td>Active Drug vs. Placebo</td>
<td>0.038</td>
<td>0.048</td>
</tr>
</tbody>
</table>
4.4 Proportional Hazards Regression

For adjustment for the effects of the 21 covariates, a proportional hazards model was fit stratifying on Limb/Bulbar status. This was done by implementing the option "STRATA" for Limb/Bulbar status in the PHREG Procedure instead of treating the stratification factor as an explanatory variable.

Table 4.4. Results of Cox Regression - Unadjusted and 21 Covariate Adjusted Test Treatment Effect, Stratifying on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Dose Comparison</th>
<th>Unadjusted Model</th>
<th>21 Covariate Adjusted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.159</td>
<td>-0.259</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.133</td>
<td>0.137</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.233</td>
<td>0.058</td>
</tr>
<tr>
<td>100 mg. vs Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.241</td>
<td>-0.413</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.135</td>
<td>0.139</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.073</td>
<td>0.003</td>
</tr>
<tr>
<td>200 mg. vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.237</td>
<td>-0.463</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.134</td>
<td>0.139</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.078</td>
<td>0.001</td>
</tr>
<tr>
<td>100/200 vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.239</td>
<td>-0.438</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.115</td>
<td>0.119</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.038</td>
<td>0.0002</td>
</tr>
<tr>
<td>Active Drug vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.212</td>
<td>-0.379</td>
</tr>
<tr>
<td>Standard Estimate</td>
<td>0.108</td>
<td>0.111</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.048</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Results from the various Cox models can be seen in Table 4.4. The parameter estimate can be interpreted as the log of the hazard ratio for those receiving the corresponding treatment relative to placebo. A negative parameter indicates less risk of "event" (in this example death or tracheostomy). The Active Drug indicator equals 1 if treatment is dose 50, 100, or 200, and it equals 0 if the treatment is placebo. Three
different models were fit in order to assess treatment effect: Either, (1) 3 indicators of treatment (50, 100, 200) were placed in the model, or (2) 2 indicators were in the model (50 mg, 100/200 mg combined), or (3) 1 indicator for treatment (active drug) was placed in the model. Each treatment comparison was fit twice; once with only treatment indicators in the model, and then again with the 21 covariates also in the model.

It is interesting to note that the treatment parameter estimate for the 21 covariate adjusted model is larger in absolute magnitude than its unadjusted counterpart (further from zero). In addition, the standard error for the adjusted model increases slightly which appears to contradict what is normally expected from analysis of covariance; namely, better precision due to variance reduction. The covariate adjusted model provides much more significant results for tests of treatment effect. Because the increase in the estimate of treatment effect (larger negative parameter) for the adjusted model and the slight increase in its standard error are somewhat unexpected results of analysis of covariance, the parameter estimates of treatment effects for the covariate adjusted model can be misinterpreted as potentially unstable.

One complication of the Cox model involves the interpretation of treatment parameter estimates. In the unadjusted model, the treatment parameter can be exponentiated and interpreted as the unconditional population average hazard ratio for test treatment relative to placebo with regard to death or tracheostomy. On the other hand, the exponentiated parameter from the 21 covariate adjusted model has a conditional interpretation. This treatment hazard ratio applies to individuals who have the same values for all 21 covariates included in the model. As one adds or removes covariates from the model, the conditional interpretation also changes. One can not readily compare the unconditional and conditional treatment parameters since they have different
interpretations. This is the same scenario one also encounters with logistic models where the parameter interpretation is dependent on covariates in the model.

Table 4.5 provide results of tests of goodness-of-fit (GOF). The PHREG procedure in SAS (SAS 92) was used to fit these models. A stepwise selection method was chosen for model building. The first column includes no variables in the model, and the corresponding treatment parameters were specified outside of the model. The selection entry p-value (SLE) was specified to be so small that no variable could enter the model. The option DETAILS was specified in order to produce a residual chi-square statistic for testing treatment parameters; the degrees of freedom were equal to the number of variables outside of the model. To obtain results for the second column of Table 4.5, one forces the 21 covariates into the model (INCLUDE=21), sets the entry criteria p-value so small that no treatment parameters enter the model, and specifies the DETAILS option again.

Table 4.5 Cox Regression Score Tests to Assess Residual GOF for Test Treatment With or Without 21 Covariates in the Model, Stratifying on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Model</th>
<th>No Covariates in Model</th>
<th>21 Covariates in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>50, 100, 200 (3 treatment indicators)</td>
<td>4.34</td>
<td>13.78</td>
</tr>
<tr>
<td>Chi-square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>degrees of freedom</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.23</td>
<td>0.003</td>
</tr>
<tr>
<td>50, 100/200 (2 treatment indicators)</td>
<td>4.34</td>
<td>13.67</td>
</tr>
<tr>
<td>Chi-square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>degrees of freedom</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p-value</td>
<td>0.11</td>
<td>0.001</td>
</tr>
<tr>
<td>Active Treatment (1 treatment indicator)</td>
<td>3.92</td>
<td>11.69</td>
</tr>
<tr>
<td>Chi-square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>degrees of freedom</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.05</td>
<td>0.0006</td>
</tr>
<tr>
<td>Linear Trend (1 linear dose variable)</td>
<td>2.84</td>
<td>10.25</td>
</tr>
<tr>
<td>Chi-square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>degrees of freedom</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.092</td>
<td>0.001</td>
</tr>
</tbody>
</table>
The residual score statistic is directed at the extent to which the residuals from the model are linearly associated with the other potential (treatment) explanatory variables. If there is an association, this is an indication that these variables should be included in the model. That is why the residual score test can be thought of as a GOF test. The results of the first column of Table 4.5 indicate that when there is no adjustment for covariates, the score goodness-of-fit statistic is marginal in each case (p-values range from 0.23 to 0.05) indicating that the treatment explanatory variables should not be included in the model. However, after conditioning on the 21 covariates, the goodness-of-fit assumption is clearly violated (p-values range from 0.003 to 0.0006). The treatment parameters are needed to explain the variability in the covariate adjusted model, and they should be included.

The strong difference in conclusions between the Cox unadjusted and covariate adjusted models in this example (Tables 4.4, 4.5) can also raise concerns about the assumptions required for fitting a proportional hazards model, and the interpretability of these models. Cox models assume that the hazards are proportional, that there are no covariate x treatment interactions, and that the covariates are specified correctly. Often diagnostic procedures concerning these assumptions for the Cox model are not straightforward to implement.

Cox (1972) introduced the proportional hazards model, a general semi-parametric model, which is appropriate for survival analysis with and without censoring. His model uses the hazard function as the response variable. When survival time can be assumed to be continuous (i.e., probability of ties can be ignored), the model for the hazard is:

\[ h(t|X) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \]
\[ = h_0(t)exp(\sum_{j=1}^{p} \beta_j x_j) \]
where $X = (x_1, x_2, \ldots, x_p)$ and $h_0(t)$ is the hazard function of the underlying survival distribution when $X = 0$, and the $\beta$'s are regression coefficients. This model assumes that the hazard of any group with covariates $X$ is proportional to $h_0(t)$. It is also assumed that the individual observations of survival are independent, and censoring is uniform and non-informative.

4.5 Logrank and Wilcoxon Testing With Linear Covariate Adjustment

It is possible to conduct hypothesis testing by using a stratified logrank or Wilcoxon test with linear covariate adjustment. The only assumption required for this methodology is randomization. Initially, logrank (or Wilcoxon) scores are assigned to individuals within each stratum based on their survival information (see A.1.4.a for logrank score formula and A.1.4.b for Wilcoxon score formula). Mean response vectors are then formed which are comprised of the average response score (logrank or Wilcoxon) and average value of each covariate for each treatment group within each stratum separately. For this example, each mean response vector would have length 22, 1 response and 21 covariates (see A.1.4 for specifications of the form of the response vector).

With this hypothesis testing setting, the covariance matrix, $V_{h0}$, is estimated by the pooled covariance matrix for each stratum because under the null hypothesis, the expected value of the response for a given stratum is equal for all treatment groups within that given stratum, and under randomization the expected values of the covariates are equal for the treatment groups (see A.2.2 for covariance specification). If one initially considers the case when there are two treatments, a difference of mean responses vector can be created for each stratum. This vector is obtained by subtracting the mean response vector for treatment group 1 from the mean response vector for treatment group 2 (see A.5 for details on difference in means parameterization). Difference vectors and
covariance matrices are then combined across strata using an appropriate weight (see A.8.1, \( \beta_w \) method).

For this example, a Mantel-Haenszel weight was used to combine response vectors and covariance matrices across the two strata. The \( \beta_w \) stratification method was used because we wish to adjust for 21 covariates which results in a 22 x 22 covariance matrix for each stratum, and there may be some concern about matrix non-singularity if covariance adjustment was performed within each of the two strata separately (\( \bar{\beta} \) method) and then combined.

One can specify a design matrix which either allows the difference in average covariates to vary randomly between the treatment groups, or one can have the vector of differences in covariate values between treatment groups restricted to equal zero (see A.5 for design matrix specifications). Using the estimating equations for weighted least squares regression specified in A.5, unadjusted logrank (or Wilcoxon) or covariate adjusted logrank (or Wilcoxon) tests can be obtained. Only those individuals randomized to a treatment group that is used in the pairwise comparison are included in the analysis.

The unadjusted tests in Table 4.6 agree very closely with the results provided by the SAS LIFETEST procedure (SAS 93) in Table 4.3. The stratified unadjusted logrank and Wilcoxon tests indicate a trend between treatment and length of survival, although the statistical significance is not as strong as one would like. However, after adjusting for the random imbalance among the 21 covariates for the respective treatment comparisons, the test of treatment in every case except dose 50 versus placebo is much more compelling. For the 100 mg dose versus placebo, the unadjusted logrank test p-
value = 0.076 which is not particularly strong. After adjusting for the 21 covariates, the p-value = 0.018, a much stronger basis for the treatment effect.

Table 4.6 Results of Unadjusted and 21 Covariate Adjusted Logrank and Wilcoxon Tests Using Linear Models, Stratifying on Limb/Bulbar Status

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg vs. Pl.</td>
<td>Parameter</td>
<td>0.058</td>
<td>0.051</td>
<td>-0.072</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.047</td>
<td>0.037</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.222</td>
<td>0.169</td>
<td>0.250</td>
</tr>
<tr>
<td>100 mg vs. Pl.</td>
<td>Parameter</td>
<td>0.097</td>
<td>0.105</td>
<td>-0.111</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.047</td>
<td>0.037</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.038</td>
<td>0.005</td>
<td>0.076</td>
</tr>
<tr>
<td>200 mg vs. Pl.</td>
<td>Parameter</td>
<td>0.086</td>
<td>0.094</td>
<td>-0.109</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.046</td>
<td>0.037</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.065</td>
<td>0.012</td>
<td>0.075</td>
</tr>
<tr>
<td>100/200 vs. Pl.</td>
<td>Parameter</td>
<td>0.091</td>
<td>0.100</td>
<td>-0.109</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.040</td>
<td>0.033</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.023</td>
<td>0.002</td>
<td>0.039</td>
</tr>
<tr>
<td>Active Drug vs. Pl.</td>
<td>Parameter</td>
<td>0.080</td>
<td>0.081</td>
<td>-0.097</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.038</td>
<td>0.031</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>p-value</td>
<td>0.034</td>
<td>0.009</td>
<td>0.052</td>
</tr>
</tbody>
</table>

Parameter is the difference in average logrank (or Wilcoxon) scores

The adjusted difference in average logrank scores between Dose 100 and placebo becomes larger in absolute value with covariate adjustment, perhaps indicating a random covariate imbalance which favors placebo.

Larger negative logrank treatment parameters and larger positive Wilcoxon treatment parameters indicate a benefit for test treatment. The other observed consequence of covariate adjustment is a reduced standard error from 0.062 to 0.051 for
the Dose 100 logrank test which is equivalent to a variance reduction of 32 percent. Parameter adjustment for random covariate imbalance and variance reduction are more expected results of analysis of covariance, and in this particular case the combined effect is a more significant test of treatment effect relative to the unadjusted test.

This general trend of a more significant test treatment effect with stratified logrank and Wilcoxon tests with covariate adjustment provides support for the results of the Cox regression models where a similar pattern was seen (Tables 4.4, 4.5). For most of the treatment comparisons, the p-values from the covariate adjusted logrank or Wilcoxon test and the corresponding Cox model were of the same magnitude. For example, the 100 mg. vs. placebo comparison had a covariate adjusted Wilcoxon p-value = 0.005, and a Cox regression p-value = 0.003. One notable exception is the 50 mg. dose versus placebo comparison. The covariate adjusted logrank test $p = 0.254$, but the Cox model $p = 0.058$. The overall agreement of the two types of covariate adjusted tests (Cox or Wilcoxon/logrank) leads one to have confidence that the proportional hazards model fits the data fairly well.

4.6 Simultaneous Modeling of Multiple Doses of Test Treatment

For the results in Tables 4.7 and 4.8, a compound difference vector was created which was a concatenation of the (Dose 50 - placebo), (Dose 100 - placebo), and (Dose 200 - placebo) difference in means vectors. All patients are used in this analysis. The randomization covariance matrix, which is estimated from the pooled treatment groups, is adjusted to account for the fact that the four treatment groups are dependent within a given stratum (see A.10.2 for specific details of the covariance structure). Because we are comparing more than two treatment groups within a stratum, Mantel-Haenszel weights are not straightforward to use to combine strata.
For these results in the two following tables, either no stratification was performed (see A.10.2.a for details), or an extended Mantel-Haenszel method which is comparable to $\beta$ stratification method is employed (see A.10.2.b for computation details).

Table 4.7. Test of Overall Treatment Effect for Wilcoxon and Logrank Tests With or Without 21 Covariate Adjustment, With or Without Stratification

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Chi-square</td>
<td>5.013</td>
<td>8.408</td>
<td>3.850</td>
<td>6.279</td>
</tr>
<tr>
<td>degrees of freedom</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.171</td>
<td>0.038</td>
<td>0.278</td>
<td>0.099</td>
</tr>
</tbody>
</table>

Stratifying ($\beta$) on Limb/Bulbar Status:

| Wald Chi-square         | 5.297               | 7.157                 | 4.218             | 5.200               |
| degrees of freedom      | 3                   | 3                     | 3                 | 3                   |
| p-value                 | 0.151               | 0.067                 | 0.239             | 0.158               |

The null hypothesis that is being tested in Table 4.7 is that there is no treatment effect, i.e., there is no effect of Dose 50, 100, or 200 relative to placebo. The alternative hypothesis is that at least one of the doses is significantly different than placebo with regard to survival. This test is somewhat analogous to the ANOVA main effect global test of treatment with placebo as the reference level.

Table 4.8. Test of Linear Trend for Wilcoxon and Logrank Tests With or Without 21 Covariate Adjustment, With or Without Stratification on Limb/Bulbar Status

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Chi-square</td>
<td>3.622</td>
<td>6.696</td>
<td>3.150</td>
<td>5.598</td>
</tr>
<tr>
<td>degrees of freedom</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.057</td>
<td>0.010</td>
<td>0.076</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Stratification ($\beta$ method) on Limb/Bulbar Status:

| Wald Chi-square    | 3.986               | 5.893                 | 3.443             | 4.694               |
| degrees of freedom | 1                   | 1                     | 1                 | 1                   |
| p-value            | 0.046               | 0.015                 | 0.064             | 0.030               |
In Table 4.8 we are testing the hypothesis that there is a linear relationship between dose and survival as assessed by logrank (or Wilcoxon) scores. We would expect in a linear relationship that survival should improve as the dose increases. Linearity is tested by specifying a contrast [-1 1 3] where -1 corresponds to Dose 50 vs. placebo, 1 to Dose 100 vs. placebo, and 3 to Dose 200 vs. placebo. If one plotted the contrast values (-1, 1, and 3) on the x-axis and the doses (50, 100, and 200) on the y-axis, the three points corresponding to dose and contrast would fall on a line, hence the test of linearity of dose. The small p-values (p < 0.05) tend to support the hypothesis of a linear dose relationship.

Table 4.9 Goodness of Fit Test: Appropriateness of Constraining 21 Covariates to be Equivalent for Pairwise Comparisons, Stratifying (β_w) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th></th>
<th>50 vs. Pl</th>
<th>100 vs. Pl</th>
<th>200 vs. Pl</th>
<th>100/200 vs. Pl</th>
<th>Tx vs. Pl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square*</td>
<td>20.3</td>
<td>21.5</td>
<td>27.6</td>
<td>26.6</td>
<td>26.2</td>
</tr>
<tr>
<td>df</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>p-value</td>
<td>0.50</td>
<td>0.43</td>
<td>0.15</td>
<td>0.18</td>
<td>0.20</td>
</tr>
</tbody>
</table>

* Wald Chi-square

The results of Table 4.9 indicate that no extreme randomization imbalance was observed among these comparisons with placebo, and that constraining the difference in mean covariates to be equal to zero between test treatments and placebo is appropriate.

4.7 Bivariate Methods - Wilcoxon and Logrank Tests

An interesting trend that can be seen in Tables 4.3 and 4.6 is that, in general, the Wilcoxon test provides a more significant test concerning treatment effects than its logrank counterpart for this clinical trial. For example, the stratified unadjusted logrank test for Dose 100 versus placebo provides a p-value = 0.076, whereas the Wilcoxon test for the same comparison yields a p-value = 0.038. Because the protocol for this study
explicitly stated that the logrank test would be used, there are limitations on the role of the Wilcoxon results in the primary analysis.

One possible solution that could have been implemented prior to the unmasking of the study (and preferably at the protocol development stage) would be to specify that a bivariate test would be used as the primary analysis for the logrank and Wilcoxon tests simultaneously after 18 months of follow-up. Because this would provide a Wald chi-square with 2 degrees of freedom, a larger test statistic would be required to address the type I error rate. This bivariate test could be performed with or without covariate adjustment (see A.4.2, A.4.2.a for detailed specifications).

Table 4.10 Results of Unadjusted and 21 Covariate Adjusted Bivariate Logrank and Wilcoxon Tests Using Linear Models, Stratifying on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Dose Comparison</th>
<th>Unadjusted Wilcoxon and Logrank</th>
<th>21 Covariate Adjusted Wilcoxon and Logrank</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg vs. Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>1.51</td>
<td>2.21</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p-value</td>
<td>0.469</td>
<td>0.331</td>
</tr>
<tr>
<td>100 mg vs Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>4.97</td>
<td>8.87</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p-value</td>
<td>0.083</td>
<td>0.012</td>
</tr>
<tr>
<td>200 mg vs Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>3.42</td>
<td>6.58</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p-value</td>
<td>0.181</td>
<td>0.037</td>
</tr>
<tr>
<td>100/200 vs Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>5.44</td>
<td>9.87</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p-value</td>
<td>0.066</td>
<td>0.007</td>
</tr>
<tr>
<td>Active Drug vs. Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>4.69</td>
<td>7.12</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>p-value</td>
<td>0.096</td>
<td>0.029</td>
</tr>
</tbody>
</table>

102
If the two degrees of freedom bivariate test meets the significance criterion as specified in the analysis plan, then either the logrank or the Wilcoxon test can be chosen as the preferred univariate test. This procedure should only be used if one can assume that both tests will provide a fairly strong test of treatment; otherwise combining the two tests can be counter-productive.

4.7.1 Bivariate Logrank Test at 12 and 18 Months

As was seen in the Kaplan-Meier curves in Figure 4.1, the survival benefit for active treatment seems to be more pronounced at 12 months than at 18 months. A bivariate test could also have been prespecified to address this possible pattern if it was assumed, a priori, that the combined survival evidence at 12 and 18 months would be stronger than either timepoint individually. As before, an adjustment of the critical value needs to be made to account for the two degrees of freedom in the test.

For this method of bivariate logrank hypothesis testing, one set of logrank scores is assigned to individuals for the interval 0 – 12 months where individuals surviving after 12 months are censored at 12 months. Another set of logrank scores is assigned to the interval of follow-up from 0 – 18 months. Each of these two logrank scores is considered a separate response and a multivariate model is fit to these two responses (see A.4.2 for modeling details). The results in Table 4.11 indicate that the joint logrank test for both 12 months and 18 months is quite significant after adjustment for random imbalance of covariates.
Table 4.11 Results of Unadjusted and 21 Covariate Adjusted Bivariate Logrank Tests at 12 and 18 Months, Using Linear Models, Stratifying (βw) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Dose Comparison</th>
<th>Unadjusted 12 and 18 Mo. Logrank</th>
<th>21 Covariate Adjusted 12 and 18 Mo. Logrank</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg vs. Pl.</td>
<td>Chi-square* 2.65</td>
<td>3.80</td>
</tr>
<tr>
<td></td>
<td>df 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>p-value 0.266</td>
<td>0.150</td>
</tr>
<tr>
<td>100 mg. vs Pl.</td>
<td>Chi-square* 3.57</td>
<td>5.81</td>
</tr>
<tr>
<td></td>
<td>df 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>p-value 0.168</td>
<td>0.055</td>
</tr>
<tr>
<td>200 mg. vs. Pl.</td>
<td>Chi-square* 7.43</td>
<td>11.42</td>
</tr>
<tr>
<td></td>
<td>df 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>p-value 0.024</td>
<td>0.003</td>
</tr>
<tr>
<td>100/200 vs. Pl.</td>
<td>Chi-square* 6.28</td>
<td>10.05</td>
</tr>
<tr>
<td></td>
<td>df 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>p-value 0.043</td>
<td>0.007</td>
</tr>
<tr>
<td>Active Drug vs. Pl.</td>
<td>Chi-square* 5.80</td>
<td>8.39</td>
</tr>
<tr>
<td></td>
<td>df 2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>p-value 0.055</td>
<td>0.015</td>
</tr>
</tbody>
</table>

*Wald chi-square

4.7.2 Combining Logrank and Wilcoxon Tests

When both the logrank and Wilcoxon tests are significant (as specified in the analysis plan), it is possible to combine these two test statistics into a one degree of freedom test. This particular strategy is used when one wishes to combine outcomes which are considered correlated with one another. This method allows for responses with different scales to be combined.
Table 4.12. Results of Unadjusted and 21 Covariate Adjusted Summing of Standardized(-) Logrank and Wilcoxon Tests at 18 Months of Follow-up, Stratifying ($\beta_w$) on Limb/Bulbar Status, Using a Wald Chi-square

<table>
<thead>
<tr>
<th>Dose Comparison</th>
<th>Unadjusted LR and Wilcoxon Sum Test</th>
<th>21 Covariate Adjusted LR and Wilcoxon Sum Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg vs. Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>1.48</td>
<td>1.67</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.223</td>
<td>0.196</td>
</tr>
<tr>
<td>100 mg. vs Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>3.53</td>
<td>6.31</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.060</td>
<td>0.012</td>
</tr>
<tr>
<td>200 mg. vs. Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>2.95</td>
<td>5.88</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.086</td>
<td>0.015</td>
</tr>
<tr>
<td>100/200 vs. Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>4.37</td>
<td>8.15</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.037</td>
<td>0.004</td>
</tr>
<tr>
<td>Active Drug vs. Pl.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>3.96</td>
<td>6.04</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.047</td>
<td>0.014</td>
</tr>
</tbody>
</table>

First, we note that large negative logrank scores indicate better survival, and large positive Wilcoxon scores indicate better survival so the opposite sign of the logrank score is used for this analysis. $Z$ scores ($mean = 0$, $std$ dev $= 1$) are created for the Wilcoxon and ($-$) logrank scores separately using the Standard procedure (SAS 92) in SAS. The sum of these two resulting $Z$ scores is then used as the outcome of interest. Means of covariates and $z$ score sums are calculated for each of the two treatment groups within each stratum separately, and the difference vectors between means and the corresponding $V_0$ matrix are used to perform tests of significance (see A.2.2) where in this case the response of interest is no longer an individual logrank or Wilcoxon score but instead is the sum of the two standardized scores.
The results of Table 4.12 indicate that when the response of interest is the combined logrank and Wilcoxon scores through a 1 df test, a pattern of more powerful tests of treatment differences is evident when covariate adjustment is performed.

Nonparametric hypothesis testing was performed using logrank scores, Wilcoxon scores, logrank and Wilcoxon scores jointly in a multi-degree of freedom test, and the sum of standardized logrank and Wilcoxon scores in a 1 degree of freedom test. A similar pattern was seen across all of these strategies. The unadjusted test of treatment effect showed a trend that there was some treatment effect, but the significance level was not particularly compelling. However, adjusting for the 21 prespecified covariates provided for a much stronger test of treatment effect. Parameter estimates increased and standard errors decreased with linear covariate adjustment. These resulting p-values were of the same magnitude as the covariate adjusted Cox model p-values for test treatment.

This nonparametric evidence would lead one to believe that the dramatic results that were seen in the covariate adjusted Cox model relative to the unadjusted model (see Tables 4.4, 4.5) were not the result of unstable parameter estimates or major violations of proportional hazards model assumptions. The result of these analyses provide a greater comfort level with the results of the covariate adjusted Cox regression models.
Chapter 5
Survival Estimation: Incidence Densities and Survival Rates

For this chapter we will continue to consider the ALS clinical trial that was introduced in the previous chapter. In Chapter 4 we performed hypothesis testing using logrank or Wilcoxon scores. Hypothesis testing addresses whether there is a significant difference in mean response between treatment groups. However, hypothesis testing tells nothing about the magnitude of the test treatment effect. In this chapter, we will propose two different types of estimators for assessing treatment effect in the survival setting; incidence density and survival rates.

5.1 Introduction of Incidence Density Estimation

The first proposed estimator, incidence density (ID), can be thought of as a ratio of the number of events that occur in an interval of follow-up time divided by the number of units of person-time at risk for the same interval. The amount of follow-up time for the clinical trial is divided into \( v \) intervals and incidence density is calculated for each interval separately. For linear model fitting, the log of incidence density (ID) is the response of interest. The response is estimated for each treatment group within each interval separately, and the treatment parameter of interest is the difference in average log ID between treatment groups.

5.1.1 Estimating Interval Specific Incidence Density

Table 5.1 shows the difference in log ID among various treatment groups (see A.1.5 for structure of the response vector), for three different intervals of time (0 – 6
months, 6 – 12 months, and 12 – 18 months). An estimated covariance matrix, in contrast to a randomization covariance matrix, is used to estimate standard errors (see A.2.1.d) for these tests in Table 5.1. The $\beta_w$ method of stratification is used (see A.8.1 for details). Either no covariance adjustment is performed (unadj. column) or difference in mean covariate values between treatment groups is restricted to equal zero (21 cov adj. column). These are the same 21 covariates and the same stratification factor that are described in Table 4.1.

Table 5.1 Differences in Log Incidence Density Estimates Between Respective Treatment Groups, Stratifying ($\beta_w$) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Treatments</th>
<th>At 6 Mo.</th>
<th>At 12 Mo.</th>
<th>At 18 Mo.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unadj. 21 cov. adj.</td>
<td>unadj. 21 cov.adj.</td>
<td>unadj. 21 cov. adj.</td>
</tr>
<tr>
<td><strong>50 vs. Pl</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>-0.335 -0.402</td>
<td>-0.229 -0.139</td>
<td>0.129 0.182</td>
</tr>
<tr>
<td>std. error</td>
<td>0.263 0.230</td>
<td>0.205 0.184</td>
<td>0.249 0.237</td>
</tr>
<tr>
<td>p-value</td>
<td>0.202 0.081</td>
<td>0.266 0.452</td>
<td>0.606 0.444</td>
</tr>
<tr>
<td><strong>100 vs. Pl</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>-0.284 -0.442</td>
<td>-0.464 -0.428</td>
<td>0.078 0.139</td>
</tr>
<tr>
<td>std. error</td>
<td>0.257 0.228</td>
<td>0.215 0.199</td>
<td>0.247 0.234</td>
</tr>
<tr>
<td>p-value</td>
<td>0.270 0.053</td>
<td>0.031 0.031</td>
<td>0.751 0.553</td>
</tr>
<tr>
<td><strong>200 vs. Pl</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>-0.298 -0.433</td>
<td>-0.400 -0.406</td>
<td>-0.018 0.049</td>
</tr>
<tr>
<td>std. error</td>
<td>0.257 0.230</td>
<td>0.209 0.189</td>
<td>0.261 0.250</td>
</tr>
<tr>
<td>p-value</td>
<td>0.247 0.060</td>
<td>0.056 0.032</td>
<td>0.946 0.846</td>
</tr>
<tr>
<td><strong>Dose 100/200</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>-0.281 -0.416</td>
<td>-0.431 -0.447</td>
<td>0.036 0.123</td>
</tr>
<tr>
<td>std. error</td>
<td>0.216 0.192</td>
<td>0.178 0.161</td>
<td>0.221 0.210</td>
</tr>
<tr>
<td>p-value</td>
<td>0.194 0.030</td>
<td>0.015 0.005</td>
<td>0.870 0.557</td>
</tr>
<tr>
<td><strong>Active vs. Pl</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>-0.288 -0.393</td>
<td>-0.358 -0.358</td>
<td>0.074 0.167</td>
</tr>
<tr>
<td>std. error</td>
<td>0.202 0.178</td>
<td>0.164 0.147</td>
<td>0.208 0.196</td>
</tr>
<tr>
<td>p-value</td>
<td>0.153 0.027</td>
<td>0.029 0.015</td>
<td>0.721 0.393</td>
</tr>
</tbody>
</table>

parameter is the difference in log incidence density (treatment - placebo)
As in previous applications of this linear model methodology (response means, logits, scores), covariance adjustment provides for a slight adjustment in the parameter estimate due to random imbalance favoring the placebo groups, and the standard error is reduced leading to increased precision with covariance adjustment. It is interesting to note that adjustment for random imbalance can favor either test treatment or placebo. Both are illustrated in Table 5.1. For example, at 15 months the Dose 50 versus placebo parameter estimate actually gets smaller with covariate adjustment, and despite a reduction in the standard error, the net result is a less significant test of test treatment (Dose 50) with covariate adjustment for that particular interval.

There is a general trend among the treatment comparisons for a survival benefit from 0 – 6 months for treatment relative to placebo, although the test is not always statistically significant. This survival benefit for treatment becomes more pronounced for the 6 – 12 month interval, but the benefit seems to disappear for the 12 – 18 month interval. The treatment benefit is more compelling for the higher doses of treatment, whereas dose 50 shows minimal benefit relative to placebo. These trends agree with the Kaplan-Meier curves found in Figure 4.1. For those curves, survival benefit for treatment seems to peak somewhere around 12 months (i.e., distance between treatment and placebo curves is at a maximum), and the curves tend to become closer together again around 18 months, indicating less treatment benefit at this point in time.

5.1.2 Estimating Proportional Incidence Density

In Table 5.1 there were three different estimates of treatment effect for each pairwise treatment comparison corresponding to the three intervals of time. It would be desirable to have a summary measure of test treatment effect across intervals such as a mean. In order to produce the mean response over the intervals of time, we need to assume that the test treatment effect is homogeneous over the intervals; that is, the
treatment effect is proportional over time. In some ways, this approach is analogous to the assumption of proportional hazards where the ID (or hazard) does not need to be consistent over time, but the difference in log ID (or ratio of hazards) between the two respective treatment groups needs to be a constant. For Table 5.2, an average difference in log incidence density is estimated (see A.4.3 for estimation methods), and it is the resulting treatment parameter.

Table 5.2. Linear Model for Proportional Incidence Density - Unadjusted and 21 Covariate Adjusted Treatment Effect, Stratifying ($\beta_w$) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Dose Comparison</th>
<th>Unadjusted Model</th>
<th>21 Covariate Adjusted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>50 mg vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.151</td>
<td>-0.134</td>
</tr>
<tr>
<td>ID Ratio</td>
<td>0.860</td>
<td>0.875</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.136</td>
<td>0.108</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.265</td>
<td>0.215</td>
</tr>
<tr>
<td><strong>100 mg vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.245</td>
<td>-0.273</td>
</tr>
<tr>
<td>ID Ratio</td>
<td>0.783</td>
<td>0.761</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.137</td>
<td>0.113</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.074</td>
<td>0.015</td>
</tr>
<tr>
<td><strong>200 mg vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.264</td>
<td>-0.303</td>
</tr>
<tr>
<td>ID Ratio</td>
<td>0.768</td>
<td>0.739</td>
</tr>
<tr>
<td>Standard Estimate</td>
<td>0.138</td>
<td>0.111</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.055</td>
<td>0.007</td>
</tr>
<tr>
<td><strong>100/200 vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.257</td>
<td>-0.295</td>
</tr>
<tr>
<td>ID Ratio</td>
<td>0.773</td>
<td>0.745</td>
</tr>
<tr>
<td>Standard Estimate</td>
<td>0.117</td>
<td>0.094</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.028</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Active Drug vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>-0.220</td>
<td>-0.242</td>
</tr>
<tr>
<td>ID Ratio</td>
<td>0.803</td>
<td>0.785</td>
</tr>
<tr>
<td>Standard Estimate</td>
<td>0.109</td>
<td>0.086</td>
</tr>
<tr>
<td>Pvalue</td>
<td>0.042</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Parameter estimate is the log(incidence density) for treatment
One can exponentiate the treatment parameter estimate to obtain the ratio of incidence density of test treatment relative to the incidence density of placebo, somewhat analogous to a hazard ratio in a Cox regression model. A ratio which is less than 1 is an indication of a survival benefit for test treatment. Corresponding 95% confidence intervals for this ratio can be estimated as \( \exp[(\beta_{trt}) \pm 1.96(s.e.)] \). For the covariate adjusted treatment comparison of Dose 100 versus placebo, the ID ratio is 0.761 and the corresponding 95% confidence interval for the ratio is (0.610, 0.949). Because this interval does not contain the value 1, the test of treatment effect is significant at the \( \alpha=0.05 \) level.

In each comparison of test treatment versus placebo (except Dose 50), the parameter estimate became more negative indicating a larger treatment effect. In addition, variance was reduced for each comparison. The covariate adjusted estimate of test treatment incidence density ratio ranged from 0.74 to 0.88 depending on the dose comparison. This can be interpreted as there being 12% to 26% fewer deaths on the test treatment arm relative to placebo, and this estimated treatment effect was significant for all pairwise comparisons except Dose 50 versus placebo. The resulting p-values are in the same ballpark as the covariate adjusted Cox model p-values found in Table 4.4.

5.1.3 Testing Proportional Incidence Density Assumption

In order to be comfortable with the results in Table 5.2, it is necessary to test the appropriateness of averaging the treatment effect across \( v \) intervals. A non-significant test in Table 5.3 would support compatibility of the results with the assumption of homogeneity of treatment effects. This can also be thought of as applying to "proportional incidence density". See A.4.3.b for a discussion of how this assumption is tested. In general, the results of Table 5.3 support the assumption of homogeneity,
although dose 100/200 versus placebo and any active drug versus placebo show some lack of homogeneity for the covariate adjusted test.

Table 5.3 Test of Proportional Incidence Density Assumption Over 3 Intervals of Time, Using the Wald Chi-square

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unadjusted</th>
<th>21 Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>50 mg. vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>1.892</td>
<td>2.835</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.388</td>
<td>0.242</td>
</tr>
<tr>
<td><strong>100 mg. vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>2.765</td>
<td>4.017</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.251</td>
<td>0.134</td>
</tr>
<tr>
<td><strong>200 mg. vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>1.339</td>
<td>2.475</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.512</td>
<td>0.290</td>
</tr>
<tr>
<td><strong>100/200 vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>2.726</td>
<td>4.983</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.256</td>
<td>0.083</td>
</tr>
<tr>
<td><strong>Active Drug vs. Placebo</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>2.818</td>
<td>5.404</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.244</td>
<td>0.067</td>
</tr>
</tbody>
</table>

In Table 5.1 we see that for 100/200 versus placebo and active treatment versus placebo, both of the first two intervals show a significant test treatment effect with covariate adjustment. However, this treatment effect is not seen for the last interval of time (12 – 18 months). In fact the sign of the treatment parameter estimate for the last interval is positive, indicating an incidence ratio greater than 1 (i.e., not favoring test treatment). It is this heterogeneity which is underlying the nearly significant proportional incidence density test for these treatment comparisons.
5.1.4 Simultaneous Modeling of Multiple Doses of Test Treatment

If one continues to assume proportional incidence density, it is possible to jointly model the test treatment effect of Dose 50 vs. placebo, Dose 100 vs. placebo, and Dose 200 vs. placebo in the same model. By doing this, all observations are used in the model. For the results in Tables 5.4 and 5.5, a difference vector is created which stacks each of the three individual (test dose - placebo) difference in means vectors on top of each other vertically. A estimation matrix is used whose 3 diagonal components are: Cov(Dose 50 + Cov(placebo), Cov(Dose 100) + Cov(placebo), and Cov(Dose 200) + Cov(placebo). The covariance in each of the off-diagonal positions is Cov(placebo). Because we are stratifying on limb/bulbar status, the concatenated difference in means vector, and the block diagonal estimation matrix are estimated for each stratum separately (See A.10.1 for specifics). In order to combine response vectors from the two strata (βw stratification method), weights are based on the ratio of the stratum sample size relative to the total sample size (as was explained in Chapter 4.6). In addition, the subset of 8 covariates that was identified in the stepwise Cox model with no treatment in the model is used due to concern about invertibility of a 66 x 66 covariance matrix instead of the stepwise reduced 27 x 27 covariance matrix.

Table 5.4 Test of Dose Effect in Proportional ID With or Without Covariate Adjustment, Stratifying (βw) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unadjusted</th>
<th>8 Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dose Effect</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>5.08</td>
<td>10.18</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.166</td>
<td>0.017</td>
</tr>
</tbody>
</table>
Table 5.4 is testing the null hypothesis of no treatment effect across each of the three doses relative to placebo. A significant test indicates that at least one of the three doses is significantly different than placebo with regard to average incidence density. The unadjusted test provides little evidence of a significant effect, but the covariate adjusted test is more compelling. Covariate adjustment is performed within each pairwise dose comparison with placebo, i.e., differences in means are constrained to be equal to zero for the following groups: Dose 50 and placebo, Dose 100 and placebo, and Dose 200 and placebo. Pairwise covariate differences among the various doses of active treatment are left to be imbalanced.

Table 5.5 Test of Dose Linear Trend in Proportional ID With or Without Covariate Adjustment, Stratifying on Limb/Bulbar Status, Using a Wald Chi-square

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unadjusted</th>
<th>8 Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Trend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>4.33</td>
<td>8.30</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p-value</td>
<td>0.038</td>
<td>0.004</td>
</tr>
</tbody>
</table>

In Table 5.5 we are testing whether there is a linear dose trend. By specifying the contrast [-1 1 3], we are assessing whether survival benefit (as estimated with average incidence density), increases in a linear fashion with dose. The significant tests for both the unadjusted and covariate adjusted linear trend test indicate rejection of the null hypothesis of a linear trend.

5.2 Piecewise Exponential Model

Statistical models can extend the analysis of grouped survival data by providing a description of the pattern of event rates. This pattern can be described over time as well as describe the variation due to the influence of treatment and other group explanatory variables. One particularly useful model is the piecewise exponential model.
Table 5.6 Comparison of Four Treatments for Grouped Event Information

<table>
<thead>
<tr>
<th>Dose</th>
<th>Time (months)</th>
<th>Death or Tracheostomy</th>
<th>Lost to Follow-up</th>
<th>Survived</th>
<th>Exposure (in days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PI</td>
<td>0-6</td>
<td>37</td>
<td>1</td>
<td>204</td>
<td>41204</td>
</tr>
<tr>
<td></td>
<td>6-12</td>
<td>53</td>
<td>0</td>
<td>151</td>
<td>33174</td>
</tr>
<tr>
<td></td>
<td>12-18</td>
<td>30</td>
<td>59</td>
<td>62</td>
<td>212999</td>
</tr>
<tr>
<td>50 mg</td>
<td>0-6</td>
<td>25</td>
<td>1</td>
<td>211</td>
<td>40539</td>
</tr>
<tr>
<td></td>
<td>6-12</td>
<td>45</td>
<td>2</td>
<td>164</td>
<td>34312</td>
</tr>
<tr>
<td></td>
<td>12-18</td>
<td>35</td>
<td>62</td>
<td>67</td>
<td>22473</td>
</tr>
<tr>
<td>100</td>
<td>0-6</td>
<td>26</td>
<td>1</td>
<td>209</td>
<td>40457</td>
</tr>
<tr>
<td></td>
<td>6-12</td>
<td>36</td>
<td>0</td>
<td>173</td>
<td>35529</td>
</tr>
<tr>
<td></td>
<td>12-18</td>
<td>37</td>
<td>63</td>
<td>73</td>
<td>24114</td>
</tr>
<tr>
<td>200</td>
<td>0-6</td>
<td>28</td>
<td>1</td>
<td>215</td>
<td>41826</td>
</tr>
<tr>
<td></td>
<td>6-12</td>
<td>39</td>
<td>0</td>
<td>176</td>
<td>36294</td>
</tr>
<tr>
<td></td>
<td>12-18</td>
<td>35</td>
<td>64</td>
<td>77</td>
<td>23267</td>
</tr>
</tbody>
</table>

If the following assumptions can be made, then a piecewise exponential model can be fit to these data (Stokes et al. 95):

1. The withdrawals are uniformly distributed during the time intervals in which they occur and are unrelated to treatment failures.
2. The within-interval probabilities of the treatment failures are small. The time-to-failure events have independent exponential distributions.

One can assume the piecewise exponential model but obtain estimates from Poisson regression computations (see Chapter 1.3.4.1 for details). Poisson regression is readily accessible using the SAS Genmod procedure (SAS 97).

The results in Table 5.7 indicate that log incidence density increases significantly with the 6-12 month interval relative to 0-6 months, and it continues to increase for the 12-18 month interval but the increase is not significantly different than the 6-12 month
interval. The negative parameters for treatment indicate a decreased log incidence density for the test treatment groups relative to the placebo group, and this difference is borderline significant for both the 100 mg. and 200 mg. doses.

Table 5.7 Parameter Estimates from Poisson Regression

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Chi-square</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-7.098</td>
<td>0.120</td>
<td>3496.1</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>0-6 months</td>
<td>0.000</td>
<td>0.000</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6-12 months</td>
<td>0.566</td>
<td>0.120</td>
<td>0.505</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>12-18 months</td>
<td>0.759</td>
<td>0.126</td>
<td>36.1</td>
<td>0.0001</td>
</tr>
<tr>
<td>Placebo</td>
<td>0.000</td>
<td>0.000</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>50 mg. dose</td>
<td>-0.160</td>
<td>0.134</td>
<td>1.428</td>
<td>0.232</td>
</tr>
<tr>
<td>100 mg. dose</td>
<td>-0.255</td>
<td>0.136</td>
<td>3.532</td>
<td>0.060</td>
</tr>
<tr>
<td>200 mg. dose</td>
<td>-0.231</td>
<td>0.135</td>
<td>2.947</td>
<td>0.086</td>
</tr>
</tbody>
</table>

The score statistic (Residual Chi-Square) for the contribution of the pairwise interactions (dose of treatment x interval of time) which are not in the model is $Q_s = 4.17$, df = 6 (# of interaction terms), and p-value = 0.654. This result was obtained by using the logistic procedure in SAS (see Chapter 15.5.2 of Stokes et al, 1995 for details). The results of this score statistic indicate that the proportional hazards assumption is reasonable, and the main effects model (with treatment and time main effects) fits adequately.

The unconditional tests of treatment for this Poisson regression method agree quite well with the unadjusted tests of treatment for the nonparametric proportional incidence density method shown in Table 5.2. Both the log incidence density estimates and corresponding p-values are very similar. However, one shortcoming of the Poisson model is that one must have grouped survival data, and individual continuous covariates cannot be incorporated into the model. Like other log-linear models, the treatment
parameters that one obtains when any group covariates are in the model (e.g., age group, gender), are conditional on these covariates.

5.3 Survival Rate Estimation

A second type of parameter for estimating test treatment effect is a survival rate. A survival rate can be interpreted as the proportion of individuals alive at the end of the \( v \)-th interval. For each interval an individual is categorized as being at risk during the interval (yes or no) and whether the event (death or tracheostomy) occurred during that interval (yes or no). A survival rate is estimated as the number of individuals alive at the end of the \( v \)-th interval divided by the number of individuals at risk through that same interval (see A.1.6 for a more detailed algorithm) where individuals who withdraw during an interval are regarded as contributing follow-up for half of that interval.

5.3.1 Estimating Interval Specific Survival Rate Differences

Table 5.8 shows the difference in survival rates between the various test treatment groups and placebo for both the unadjusted and 21 covariate adjusted models for each of the three intervals of time (0 – 12 months, 12 – 15 months, and 15 – 18 months). The interval from 0 – 12 months shows the strongest survival benefit for test treatment over placebo. This trend becomes less evident for the 12 – 15 month and 15 – 18 month interval. The point estimate supports the conclusion of a trend in treatment benefit, and after adjusting for covariates the overall result from the test of treatment is stronger (see A.1.6 and A.2.1.e for specification of response vector and covariance matrix structure, respectively).

In previous examples of covariate adjustment in this thesis, the random imbalance tended to favor placebo so adjustment for covariates usually increased the estimate of treatment effect (i.e., the estimate increased in absolute magnitude). However, in Table
5.8 we see that the effect of covariate adjustment works in both directions. Dose 100 versus placebo is an example where covariate adjustment increased the treatment effect for each interval. For Dose 50 versus placebo, covariate adjustment actually decreased the estimate of treatment effect (closer to 0) indicating a baseline covariate random imbalance favoring Dose 50 with respect to survival rates.

Table 5.8 Differences in Cumulative Survival Estimates Between Test Treatment Group and Placebo, Stratifying \( \beta_w \) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Treatments</th>
<th>At 12 Mo.</th>
<th></th>
<th>At 15 Mo.</th>
<th></th>
<th>At 18 Mo.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unadj. 21 cov. adj.</td>
<td></td>
<td>unadj. 21 cov. adj.</td>
<td></td>
<td>unadj. 21 cov. adj.</td>
<td></td>
</tr>
<tr>
<td>50 vs. Pl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>7.5% 6.5%</td>
<td></td>
<td>4.3% 3.3%</td>
<td></td>
<td>5.0% 3.5%</td>
<td></td>
</tr>
<tr>
<td>std. error</td>
<td>4.3% 3.6%</td>
<td></td>
<td>4.5% 3.7%</td>
<td></td>
<td>4.9% 4.1%</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.081 0.067</td>
<td></td>
<td>0.341 0.368</td>
<td></td>
<td>0.300 0.398</td>
<td></td>
</tr>
<tr>
<td>100 vs. Pl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>10.9% 12.1%</td>
<td></td>
<td>9.8% 10.6%</td>
<td></td>
<td>5.3% 5.8%</td>
<td></td>
</tr>
<tr>
<td>std. error</td>
<td>4.2% 3.6%</td>
<td></td>
<td>4.4% 3.6%</td>
<td></td>
<td>4.9% 4.2%</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.010 0.0007</td>
<td></td>
<td>0.027 0.003</td>
<td></td>
<td>0.274 0.166</td>
<td></td>
</tr>
<tr>
<td>200 vs. Pl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>10.0% 11.4%</td>
<td></td>
<td>5.2% 5.9%</td>
<td></td>
<td>8.5% 8.6%</td>
<td></td>
</tr>
<tr>
<td>std. error</td>
<td>4.2% 3.5%</td>
<td></td>
<td>4.4% 3.7%</td>
<td></td>
<td>4.8% 4.1%</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.018 0.107</td>
<td></td>
<td>0.243 0.107</td>
<td></td>
<td>0.074 0.035</td>
<td></td>
</tr>
<tr>
<td>Dose 100/200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>10.4% 12.2%</td>
<td></td>
<td>7.4% 8.5%</td>
<td></td>
<td>6.9% 7.0%</td>
<td></td>
</tr>
<tr>
<td>std. error</td>
<td>3.7% 3.0%</td>
<td></td>
<td>3.9% 3.1%</td>
<td></td>
<td>4.2% 3.5%</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.005 &lt;0.0001</td>
<td></td>
<td>0.055 0.007</td>
<td></td>
<td>0.098 0.048</td>
<td></td>
</tr>
<tr>
<td>Active vs. Pl</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimate</td>
<td>9.5% 10.7%</td>
<td></td>
<td>6.4% 7.0%</td>
<td></td>
<td>6.3% 5.8%</td>
<td></td>
</tr>
<tr>
<td>std. error</td>
<td>3.5% 2.9%</td>
<td></td>
<td>3.7% 2.9%</td>
<td></td>
<td>3.9% 3.3%</td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.007 0.0002</td>
<td></td>
<td>0.081 0.017</td>
<td></td>
<td>0.110 0.079</td>
<td></td>
</tr>
</tbody>
</table>

The parameter is the difference in survival rates between groups (Treatment - Placebo).

Regardless of which direction the random imbalance lies, covariance adjustment provides for a reduction in the standard error. If the decrease in the magnitude of the treatment parameter (due to random imbalance) is larger than the decrease in the standard error.
error, the net result is a less significant test of treatment effect relative to its unadjusted counterpart.

Table 5.9 shows the survival rate estimates (analogous to Kaplan-Meier estimates) for the Dose 100/200 group and the placebo group at three points in time. The difference in survival rates is what is shown in Table 5.8. For example, Table 5.9 reports a covariate adjusted Dose 100/200 survival rate at 12 months = 73.5%, and the placebo group has a corresponding 12 month survival rate = 61.3%. The difference in these two estimated rates, 12.2%, is the reported value in Table 5.8 for covariate adjusted estimates at 12 months, and the p-value corresponding to the null hypothesis that the difference in survival rates between these two treatment groups is zero highly significant, \( p < 0.0001 \).

<table>
<thead>
<tr>
<th></th>
<th>Dose 100/200</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>12 mo. 15 mo. 18 mo.</td>
<td>12 mo. 15 mo. 18 mo.</td>
</tr>
<tr>
<td>No Cov. Adj.</td>
<td>73.1% 64.4% 53.1%</td>
<td>62.7% 56.0% 46.2%</td>
</tr>
<tr>
<td>Cov. Adj.</td>
<td>73.5% 64.6% 53.5%</td>
<td>61.3% 56.1% 46.5%</td>
</tr>
</tbody>
</table>

5.3.2 Estimating Average Differences in Survival Rates

Like the incidence density setting, it is desirable to summarize the treatment effect across the \( v \) intervals of time in order to obtain one treatment parameter estimate. Average difference in survival rates across the three intervals would be of interest. In Table 5.10, the last column provides the estimated difference in survival rates between the respective test treatment group and placebo averaged across intervals. The corresponding standard error and p-value corresponding to the null hypothesis that the average difference in survival rates equals zero are provided. These estimates are adjusted for the effect of the 21 covariates, and the test is stratified (\( \beta_w \)) by limb/bulbar
status. In each case, except Dose 50, the test treatment arm has significantly better overall survival than the placebo group. The interpretation of the Dose 100 versus placebo covariate adjusted parameter estimate is that, on average, 10.0% more subjects are alive in the Dose 100 group than the placebo group over the 18 month duration after adjusting for the effect of covariate random imbalance. This survival difference is significant ($p = 0.002$).

Another test of interest is whether the null hypothesis that the difference in survival rates between treatment group and placebo is equal to zero for each interval jointly. This involves a $v$ (in this case $v = 3$) degree of freedom test corresponding to the number of intervals being jointly tested. Rejection of this null hypothesis implies that at least one of the intervals has a difference in survival rates which is not equal to zero.

The results of the left hand side of Table 5.10 indicate that, after covariate adjustment, there is at least one interval for each of the treatment comparisons, except Dose 50, where the respective test treatment has different survival rates than placebo. P-values ranged from $0.0007 - 0.236$. If the overall test of treatment effect is significant, then it would be appropriate to look at individual survival estimates for each interval being sure to specify a procedure for dealing with multiple comparisons. These specifications should be made a priori in the study's analysis plan.
Table 5.10  Difference in Survival Rate Estimates Between Test Treatment Group and Placebo, Stratifying ($\beta_w$) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Treatments</th>
<th>Simultaneous Test of 3 Intervals</th>
<th>Average Difference Over 3 Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unadjusted</td>
<td>21 covariate adj</td>
</tr>
<tr>
<td><strong>Dose 50 vs. Pl</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald Chi-square</td>
<td>4.25</td>
<td>4.25</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.236</td>
<td>0.236</td>
</tr>
<tr>
<td><strong>Dose 100 vs. Pl</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald Chi-square</td>
<td>7.11</td>
<td>11.88</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.068</td>
<td>0.008</td>
</tr>
<tr>
<td><strong>Dose 200 vs. Pl</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald Chi-square</td>
<td>9.56</td>
<td>14.71</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.023</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Dose 100/200 vs. Pl</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald Chi-square</td>
<td>8.93</td>
<td>16.89</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.030</td>
<td>0.0007</td>
</tr>
<tr>
<td><strong>Active vs. Pl</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald Chi-square</td>
<td>8.68</td>
<td>14.83</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>p-value</td>
<td>0.034</td>
<td>0.002</td>
</tr>
</tbody>
</table>

5.3.3 Testing Homogeneity of Differences in Survival Rates Across Intervals

Table 5.11 tests the appropriateness of averaging treatment effect over the three intervals of time in both the unadjusted and 21 covariate adjusted model. In the covariate adjusted case for the Dose 200 versus placebo comparison, there is a slight indication that test treatment effect is not homogeneous across the three intervals for the higher doses of treatment, but in general the assumption of homogeneity seems to be well supported. This is not surprising because we see the estimate of treatment effect favoring active treatment for each dose within each of the three intervals when one looks at the individual interval results in Table 5.8.
Table 5.11 Test of Homogeneous Differences in Survival Rates for 3 Intervals of Time

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unadjusted</th>
<th>21 Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 mg. vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>1.76</td>
<td>1.84</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.416</td>
<td>0.398</td>
</tr>
<tr>
<td>100 mg. vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>1.95</td>
<td>2.59</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.376</td>
<td>0.274</td>
</tr>
<tr>
<td>200 mg. vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>4.42</td>
<td>5.42</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.110</td>
<td>0.067</td>
</tr>
<tr>
<td>100/200 vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>2.32</td>
<td>4.19</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.313</td>
<td>0.123</td>
</tr>
<tr>
<td>Active Drug vs. Placebo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>2.74</td>
<td>4.52</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.254</td>
<td>0.104</td>
</tr>
</tbody>
</table>

5.3.4 Simultaneous Modeling of Multiple Doses of Test Treatment

For the results in Tables 5.12 and 5.13, a compound difference vector was created which was a concatenation of the Dose 50 - placebo, Dose 100 - placebo, and Dose 200 - placebo difference in means vectors. All patients are used in this analysis. An estimation covariance matrix is used (see A.10.1). For the results in the following two tables, a weight proportional to the overall stratum sample size was employed. The method of $\beta_w$ strata combination (see A.8.1) was used, and a subset of the covariates were used for adjustment.
Table 5.12 Global Test of Treatment Effect With or Without Covariate Adjustment, Stratifying ($\beta_w$) on Limb/Bulbar Status, Response is Difference in Survival Rates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unadjusted</th>
<th>8 Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>6.59</td>
<td>13.56</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.086</td>
<td>0.004</td>
</tr>
</tbody>
</table>

The null hypothesis that is being tested in Table 5.10 is that there is no treatment effect, i.e., there is no effect of Dose 50, 100, or 200 relative to placebo. The alternative hypothesis is that at least one of the doses has significantly different survival rates than placebo. This global test is somewhat analogous to the ANOVA main effect test of treatment with placebo as the reference level.

Table 5.13 Test of Dose Linear Trend in Survival Rates With or Without Covariate Adjustment, Stratifying ($\beta_w$) on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Unadjusted</th>
<th>8 Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>5.39</td>
<td>11.71</td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>pvalue</td>
<td>0.020</td>
<td>0.001</td>
</tr>
</tbody>
</table>

In Table 5.13 we are testing the null hypothesis that there is a linear relationship between dose and survival as assessed by survival rates. We would expect in a linear relationship that survival should improve as the dose increases. Linearity is tested by specifying a contrast [-1 1 3] where -1 corresponds to Dose 50 vs. placebo, 1 to Dose 100 vs. placebo, and 3 to Dose 200 vs. placebo. Because the p-value is so small for both the unadjusted and covariate adjusted linear trend test, we reject the null hypothesis that there is a linear dose trend.
5.4 Summary

Table 5.13 provides a summary of pairwise p-values from the various covariate adjusted analyses that were used to assess test treatment in Chapters 4 and 5. The first column shows the p-values from the covariate adjusted Cox model, the second and third columns provide p-values from hypothesis testing with logrank or Wilcoxon scores respectively, the fourth and fifth columns provide p-values from estimation methods with average differences in log incidence densities and average differences in survival rates.

Table 5.14 Summary of P-values From 21 Covariate Adjusted Analyses With Stratification on Limb/Bulbar Status

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Cox Regression</th>
<th>Logrank</th>
<th>Wilcoxon</th>
<th>Incidence Density</th>
<th>Survival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 vs. Pl</td>
<td>0.058</td>
<td>0.254</td>
<td>0.169</td>
<td>0.215</td>
<td>0.121</td>
</tr>
<tr>
<td>100 vs. Pl</td>
<td>0.003</td>
<td>0.018</td>
<td>0.005</td>
<td>0.015</td>
<td>0.002</td>
</tr>
<tr>
<td>200 vs. Pl</td>
<td>0.001</td>
<td>0.023</td>
<td>0.012</td>
<td>0.007</td>
<td>0.002</td>
</tr>
<tr>
<td>100/200 vs. Pl</td>
<td>0.0002</td>
<td>0.007</td>
<td>0.002</td>
<td>0.002</td>
<td>0.0004</td>
</tr>
<tr>
<td>Any Tx vs. Pl</td>
<td>0.001</td>
<td>0.021</td>
<td>0.009</td>
<td>0.005</td>
<td>0.001</td>
</tr>
</tbody>
</table>

* p-value corresponds to parameter estimate for average across 3 intervals of time

The Cox regression model requires one to make several assumptions such as proportional hazards, that the model is correctly specified, and that there are no interactions of covariates with treatment or other covariates. When the covariate adjusted treatment parameter estimate is substantially different from the unadjusted one, there are concerns about these Cox model assumptions. By applying nonparametric hypothesis testing and treatment estimation, which only require an assumption of randomization, we can gain insight into the robustness of the results of the Cox model (incidence density and survival rate estimation also require the assumption of homogeneity of treatment effect across intervals of time).
Because there is substantial agreement across the various analysis methods in terms of strength of statistical significance of treatment effect, we can conclude that there is strong evidence to indicate a significant treatment effect. This effect seems to be more pronounced in the higher doses (100, 200), but even the estimates of Dose 50 versus placebo followed the same trend as the higher doses in terms of direction of treatment effect, although it did not attain compelling statistical significance.
Chapter 6
Logrank Testing Revisited:
Application to a Cancer Clinical Trial

6.1 Introduction of Clinical Trial Example

This large national randomized cancer clinical trial involved 618 individuals with Stage III colon cancer. Patients were considered to have Stage III disease if they have at least one local lymph node invaded with cancer in addition to their primary tumor. Disease had to be contained locally. No distant metastasis was allowed. After initial diagnosis, individuals underwent a curative resection to remove all cancer from the bowel area. Surgical and pathologic review was employed to be reasonably sure that no disease remained after the surgery. Patients were required to be disease-free at the time of randomization to the clinical trial.

This study is known as an "intergroup" trial because several different national cooperative cancer research groups joined together to conduct a trial in order to get a larger accrual of patients than would be achieved by any of the cooperative groups individually. The three cooperative groups which contributed patients to this study were SWOG, ECOG, and NCCTG. For the analyses in this chapter, stratification will be performed by cooperative group.

Patients were randomized to one of two treatment arms. The first arm, Observation, involved following the individual after surgery with no further treatment. The other arm involved chemotherapy with a combination regimen of fluorouracil (5-FU)
and levamisole. The objective of the study was to assess whether post-surgical treatment with this chemotherapy regimen was beneficial in the reduction of relapses and deaths in Stage III colon cancer patients. There were two primary endpoints of interest; overall survival and disease-free survival after three years of follow-up. Time until death due to any cause is assessed in overall survival, and time until death or relapse (minimum time of two) is evaluated in disease-free survival.

There are three covariates of interest in this study. The covariate NODES takes the value 1 if the patient had 5 or more positive lymph nodes at the time of surgical resection and it takes the value 0 if 1-4 lymph nodes were involved. The second covariate is SEROSA which is an indicator of how deeply the tumor invaded the colon wall. The value Serosa = 1 is an indicator of deeper tumor penetration than Serosa = 0. The final covariate is OBSTRUCT and it takes the value 1 if the primary tumor obstructed the colon, and it equals 0 if there was no obstruction. A value of 1 for any of these three covariates (Nodes, Serosa, Obstruct) is an indication of poor prognosis.

Table 6.1 Descriptive Statistics for Covariates and Stratification Factors by Treatment Arm

<table>
<thead>
<tr>
<th>Factor</th>
<th>Observation (n = 303)</th>
<th>5-FU + Levamisole (n = 315)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes (&gt; 4)</td>
<td>27.3%</td>
<td>26.1%</td>
</tr>
<tr>
<td>Obstruction</td>
<td>20.0%</td>
<td>17.5%</td>
</tr>
<tr>
<td>Serosal Involvement</td>
<td>84.1%</td>
<td>84.8%</td>
</tr>
<tr>
<td>ECOG participants</td>
<td>32.1%</td>
<td>32.3%</td>
</tr>
<tr>
<td>SWOG participants</td>
<td>39.7%</td>
<td>38.6%</td>
</tr>
<tr>
<td>NCCTG participants</td>
<td>28.3%</td>
<td>29.0%</td>
</tr>
</tbody>
</table>

127
Figure 6.1 Overall Survival By Treatment Arm

Survival Distribution Function

Follow-up Time in Days

TX → 5FU + Lev  ○○ Observation
Figure 6.2 Disease—Free Survival By Treatment Arm

Follow-up Time in Days

TX  \( \rightarrow \) 5FU + Lev  \(-\rightarrow\) Observation
6.2 Kaplan-Meier Curves for Survival and Disease-Free Survival

One can see from Table 6.1 that the two treatment groups are very well balanced on covariates and stratification factors. Because the main objectives involve time-to-event outcomes, Kaplan-Meier curves for survival and disease-free survival by treatment arm are shown in Figures 6.1 and 6.2, respectively.

The difference between treatment groups is more pronounced in the disease-free survival setting relative to overall survival, indicating that there are more relapses on the observation arm than the chemotherapy arm.

6.3 Cox Regression Results

In order to assess the effect of treatment, Cox regression models with stratification on cooperative group (STRATA option) are evaluated with and without covariate adjustment using the PHREG procedure in SAS (SAS 92). Both overall survival and disease-free survival are assessed as the outcome of interest.

<table>
<thead>
<tr>
<th>Table 6.2 Cox Regression Results Stratifying on Cooperative Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor</strong></td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Survival</td>
</tr>
<tr>
<td>Tx Parameter</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>P-value</td>
</tr>
<tr>
<td>Disease-Free Survival</td>
</tr>
<tr>
<td>Tx Parameter</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>P-value</td>
</tr>
</tbody>
</table>

Parameter is the log hazard ratio of treatment relative to placebo.
The results of Table 6.2 indicate that the chemotherapy regimen is significantly better than the observation arm with regard to overall survival and disease-free survival. The unconditional survival hazard ratio of treatment relative to observation is equal to \( \exp(-0.335) = 0.715 \), (95% CI: 0.535, 0.956), and the disease-free survival hazard ratio = \( \exp(-0.459) = 0.632 \), (95% CI: 0.496, 0.806). Because both of these confidence intervals for the hazard ratio do not include the value 1 we can conclude that 5-FU + levamisole significant reduces the risk of relapse and death relative to no further treatment in this clinical trial. It is interesting to note that for overall survival, the covariate adjusted parameter estimate actually decreases slightly in magnitude ( -0.335 unadjusted, -0.326 adjusted). This would appear to be a result of adjusting for random imbalance by moving the parameter estimate towards the null, and the counterdynamic of a small amount of variance reduction which gets manifest through an increase in the parameter estimate. The net result is a slight decrease in the parameter estimate and virtually no change to the standard error estimate. For the disease-free survival outcome, there is no noticable change in the parameter estimate and the standard error with covariance adjustment. This would seem to indicate that the increase in the parameter estimate due to covariance adjustment and the decrease in the parameter estimate due to adjustment for random imbalance cancel each other out, leading to a net result of no change.

6.4 Correlations of Covariates With Response and Treatment

It is possible to create a prognostic score for each subject based on their covariate information. A Cox regression model is fit to only the three covariates (Nodes, Serosa, Obstruct) with no treatment indicator in the model. The \( X\beta \) score (where \( X \) is the vector of covariate values for an individual, and \( \beta \) are the corresponding parameters from the Cox model) is then output into a dataset. This model fitting and prognostic score creation is performed separately for prediction of survival and then again for disease-free survival.
These two scores are then correlated with an indicator for survival (1 = died within 3 year interval of trial, 0 = alive), and an indicator for disease-free survival (1 = died or relapsed within the 3 year interval, 0 = alive and disease-free within the interval). In addition, the prognostic scores are also correlated with the survival and disease-free survival logrank scores, and an indicator for treatment (1 = chemotherapy, 0 = observation).

The correlations in Table 6.3 indicate that the three covariates combined as a prognostic score are fairly weak predictors of both survival and disease-free survival as assessed by either their correlations with the dichotomous responses (Y/N) ($r = 0.28 - 0.31$) or their corresponding logrank scores ($r = 0.27 - 0.30$). The correlations of individual covariates with response are weaker still ($r = 0.07 - 0.29$).

"Nodes" has the strongest correlation of the covariates with the responses. Because the covariates are not strong predictors, we would expect only a small amount of variance reduction from covariate adjustment. We also note that higher prognostic scores are correlated with risk of death and relapse. Because prognostic scores are inversely correlated with treatment, we conclude that random baseline covariate imbalance slightly favors treatment, i.e., the observation group received a group of patients that had a slightly higher risk profile than the chemotherapy group. We would expect that adjustment for random imbalance of covariates would cause the estimate of chemotherapy effect to be reduced. This result was seen in the Cox regression results of Table 6.2 for disease-free survival.
Table 6.3 Spearman Correlations and Corresponding P-values of Covariates and Prognostic Scores With Dichotomous Outcome Measures, Logrank Scores (LR), and Treatment for Both Survival (S) and Disease-Free Survival (DFS)

<table>
<thead>
<tr>
<th>DFS Prog. Score</th>
<th>Survival (Y/N)</th>
<th>DFS (Y/N)</th>
<th>Treatment</th>
<th>Survival LR</th>
<th>DFS LR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.296</td>
<td>0.285</td>
<td>-0.027</td>
<td>0.271</td>
<td>0.300</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.506</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>S. Prog. Score</td>
<td>0.312</td>
<td>0.276</td>
<td>-0.027</td>
<td>0.271</td>
<td>0.300</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.506</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Obstruct</td>
<td>0.107</td>
<td>0.054</td>
<td>-0.032</td>
<td>0.077</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>0.008</td>
<td>0.177</td>
<td>0.426</td>
<td>0.057</td>
<td>0.091</td>
</tr>
<tr>
<td>Nodes</td>
<td>0.287</td>
<td>0.245</td>
<td>-0.014</td>
<td>0.271</td>
<td>0.276</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.730</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Serosa</td>
<td>0.098</td>
<td>0.160</td>
<td>0.010</td>
<td>0.070</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td>0.015</td>
<td>0.0001</td>
<td>0.813</td>
<td>0.082</td>
<td>0.001</td>
</tr>
</tbody>
</table>

6.5 Stratified Logrank Tests With Linear Covariate Adjustment

Table 6.4 provides the results of logrank tests, stratifying on cooperative group for the outcomes of overall survival and disease-free survival with follow-up truncated at 3 years (see A.1.4.a for specifics regarding assigning logrank scores). Because we are in the hypothesis testing setting, the estimated null covariance matrix, $V_0$, is used (see A.2.2 for specific structure of $V_0$). The first model (Unadj.) provides the results of the logrank test with no covariate adjustment while the second model (Cov. Adj.) is the logrank test adjusting for the random imbalance of the covariates Serosa, Obstruct, and Nodes. This is done by constraining the difference in mean covariate values between treatment groups to be equal to zero (see A.3.1.a for unadjusted hypothesis testing and A.3.2 for covariate adjusted modeling specifications). The stratification method used for these models is $\beta_w$, which involves combining mean vectors across strata by Mantel-Haenszel weighting and then covariate adjustment is performed on the resulting stratification adjusted difference in means vector (see A.8.1 for details of stratification method).
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Survival</th>
<th>Disease-free Survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadj.</td>
<td>Treatment*</td>
<td>-0.100</td>
<td>-0.198</td>
</tr>
<tr>
<td></td>
<td>Standard Error</td>
<td>0.044</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.023</td>
<td>0.0002</td>
</tr>
<tr>
<td>Cov. Adj.</td>
<td>Treatment*</td>
<td>-0.089</td>
<td>-0.185</td>
</tr>
<tr>
<td></td>
<td>Standard Error</td>
<td>0.042</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>P-value</td>
<td>0.034</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

* Treatment parameter = difference in average logrank scores (Tx - Obs), $\beta_w$ strat method

In both the survival and disease-free survival outcomes, the covariate adjusted parameter in Table 6.4 is smaller (closer to 0) than its unadjusted counterpart, indicating a random baseline covariate imbalance which favors the treatment group. This supports what we observed with the Spearman correlations in Table 6.3. In addition, there is a slight reduction in the standard error due to increased precision from covariate adjustment. It is interesting to note that the significance level for the unadjusted and covariate adjusted test of treatment remains virtually unchanged. This is due to the fact that any gain in precision due to variance reduction is cancelled out by the reduction in magnitude of the treatment parameter. This same lack of change in significance level for the conditional model was also seen in the Cox regression results in Table 6.2. The results in Table 6.4, which show both the components of variance reduction and adjustment for random imbalance, give us insight into what is happening in the Cox models in Table 6.2.

For the models in Table 6.4, it is necessary to test the appropriateness of constraining the difference in mean covariate values to be equal to zero between the treatment groups. The goodness-of-fit test provides a Wald Chi-square = 1.056 with 3
degrees of freedom and $p = 0.788$ indicating that the model has an adequate fit (see A.3.3 for GOF details). This is not surprising since the covariates are so well balanced between the two groups as we saw in Table 6.1.

6.5.1 Assessing Treatment x Strata Interaction

Because we are stratifying on cooperative group, it is important to assess whether the treatment effect is homogeneous across these three groups. Table 6.5 provides the results of testing for homogeneity of treatment effect for both the outcomes of survival and disease-free survival (see A.8.3 for treatment x strata interaction details).

Table 6.5 Assessing Homogeneity of Treatment Effect Across Cooperative Groups in the Setting of No Covariance Adjustment

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Survival</th>
<th>Disease-Free Survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Chi-square</td>
<td>2.25</td>
<td>7.70</td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P-value</td>
<td>0.324</td>
<td>0.021</td>
</tr>
</tbody>
</table>

The results in Table 6.5 indicate that the assumption of homogeneity is compatible with this clinical trial for the outcome of overall survival, but there is some indication of heterogeneity of treatment effect in the disease-free survival setting. To get a better sense of what is going on, the unconditional treatment effect is estimated for each cooperative group separately. Because covariance adjustment provides little change in the parameter estimates, the within stratum estimates are only examined in the case with no covariate adjustment.

Table 6.6 Unadjusted Individual Cooperative Group Treatment Estimates for Survival and Disease-Free Survival (DFS)

<table>
<thead>
<tr>
<th>Cooperative Group</th>
<th>Survival Tx Estimate</th>
<th>DFS Tx Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCCTG</td>
<td>-0.041</td>
<td>-0.089</td>
</tr>
<tr>
<td>SWOG</td>
<td>-0.166</td>
<td>-0.238</td>
</tr>
<tr>
<td>ECOG</td>
<td>-0.073</td>
<td>-0.245</td>
</tr>
</tbody>
</table>
The results in Table 6.6 indicate that NCCTG had a much smaller unadjusted treatment effect with respect to disease-free survival (-0.089) relative to the other two cooperative groups, and this is probably the reason for the significant test which indicates lack of homogeneity. However, all estimates in the table (survival and disease-free survival for each group) have a negative sign indicating a trend in treatment benefit.

6.6 Multivariate Assessment of Survival and Disease-Free Survival

Because there are really two primary objectives for this clinical trial, it would have been possible to specify in the analysis plan of the protocol that the two endpoints would be assessed jointly. This can be done by assessing survival and disease-free survival simultaneously in a 2 degrees of freedom test. The results of these tests can be seen in Table 6.7 (see A.4.2 for multivariate logrank testing specifications).

Table 6.7 Joint Logrank Test for Survival and Disease-free Survival at 3 Years of Follow-up, Stratifying on Cooperative Group

<table>
<thead>
<tr>
<th>Parameter</th>
<th>No Covariate Adjustment</th>
<th>Covariate Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Chi-square</td>
<td>15.32</td>
<td>14.92</td>
</tr>
<tr>
<td>DF</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P-value</td>
<td>0.0005</td>
<td>0.0006</td>
</tr>
<tr>
<td>$\beta_w$</td>
<td>stratification method</td>
<td></td>
</tr>
</tbody>
</table>

It is not surprising that the results of jointly testing survival and disease-free survival treatment effect are very significant. We know from previous results (Tables 6.2, 6.4) that treatment is quite significant for each outcome individually so the joint test should be significant as well. However, if there had been concern about multiple outcomes and prioritizing which of the two outcomes was primary a priori, this joint test would have been a reasonable proposal. One would specify that if the global 2 df test was significant, then it is allowable to look at the two individual components.
Another alternative to dealing with multiplicity in endpoints is to create a summary measure which is a combination of the two outcomes. This can be done by creating a standardized z-score (mean = 0, std dev = 1) for each logrank score, one each for survival and disease-free survival. These z-scores are created using the Standard procedure in SAS (SAS 92). Then the sum of the two z-scores is considered the single outcome of interest. The results of analyzing this summary score can be found in Table 6.8. Because we are stratifying on cooperative group, standardized z-scores are calculated within each stratum separately by using the BY option in the Standard procedure in SAS. The linear model is then fit using the \( \beta_{\omega} \) method of handling stratification (see A.8.1 for details).

Table 6.8 Combining Logrank Tests For Survival and DF Survival at 3 Years of Follow-up Using the Sum of Standardized Z-Scores (1 DF Test), With Stratification

<table>
<thead>
<tr>
<th>Tx Parameter</th>
<th>Unadjusted</th>
<th>Covariate Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.475</td>
<td>-0.454</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.153</td>
<td>0.144</td>
</tr>
<tr>
<td>P-value</td>
<td>0.002</td>
<td>0.002</td>
</tr>
</tbody>
</table>

The results in Table 6.8 confirm what was observed in Table 6.4. There is a slight reduction in variance due to covariate adjustment and the magnitude of the treatment parameter estimate is reduced slightly due to adjustment for random imbalance which favors treatment. The net result is the same level of significance for the unadjusted and covariate adjusted model. The chemotherapy regimen significantly improves survival and disease-free survival relative to the observation arm. Both methods dealing with multiplicity of endpoints, the joint 2 df test and the 1 df z-score sum test, provide the same conclusion that the chemotherapy regimen significantly improves survival and disease-free survival.
This colon cancer clinical trial provides an illustration of random baseline covariate imbalance favoring treatment. It is often the case that Cox regression models that condition on covariates appear to have little or no impact on the estimation of the treatment parameter. However, by applying nonparametric hypothesis testing with logrank scores one can gain insight into what the components of random imbalance and variance reduction are really doing. Because the p-values agree in the corresponding logrank testing and Cox modeling, and we see that the components of adjustment for random imbalance and variance reduction cancel each other, we gain a level of comfort in the results of the covariate adjusted Cox model.
Chapter 7

Summary of Thesis and Proposal for Future Research

7.1 Summary

Analysis of covariance serves two main purposes in the analysis of clinical trials. It provides for a more powerful statistical test through variance reduction for comparison of treatment groups (narrower confidence intervals); and it addresses comparisons between randomized groups with adjustment for random imbalance, thereby allowing for clarification of the degree to which detected differences between randomized groups are due to treatment rather than other factors which are associated with response.

Certain methods of covariance analysis such as logistic regression for ordinal response data and proportional hazards (Cox) regression for survival data are not readily interpretable with regards to the components of variance reduction and bias adjustment. Direct variance reduction is not observed when adjustment is made for various covariates. It is a commonly held misconception that logistic and Cox modeling not only increase variance estimates, but also lead to unstable treatment parameter estimates. In addition, there may be concerns that technical assumptions regarding the structure of the log-linear model are not appropriate such as the following: linearity for the relationship with quantitative covariables, no interaction between treatment and covariates, no interaction between treatments and strata, and the proportional odds (or hazards) assumption.

Our proposal has been to apply non-parametric counterparts to clarify the components of bias adjustment and variance reduction. For nonparametric methods,
there are no formal assumptions for how a response variable is related to the covariates, but strong correlation between response and covariates is necessary to have substantial variance reduction. The rationale for nonparametric covariance analysis is based on the randomization in the study design. Computations for these methods are through the application of weighted least squares to fit linear models to the differences between treatment groups for the means of the response variable and the covariates jointly with a specification that has null values for the differences that correspond to the covariates. By using this method, the primary response can be on the log scale (e.g. logit, log hazard, log incidence density), but covariate adjustment is performed on the linear scale. The resulting parameter will be an unconditional population average estimate of treatment effect, adjusted for imbalance of covariates. Appropriateness of the application of this methodology is based on sampling arguments.

This proposed methodology was applied to several types of outcomes including differences between treatment groups for: proportions, ordinal scores, logits (both dichotomous and ordinal logistic), logrank and Wilcoxon scores, incidence densities, and survival rates. Methodologic differences for hypothesis testing versus estimation were illustrated. Extensions of this methodology were also made to the stratified case. Two different methods for handling stratification were explored: (1) mean response vectors were combined across strata and then covariate adjustment was performed, or (2) covariate adjustment was performed within each stratum and then adjusted estimates were combined across strata. Appropriate application of each stratification method and various weighting schemes were illustrated. Often clinical trials have more than two treatment groups. With this methodology it is possible to specify contrasts to test various linear hypotheses regarding treatment effects.
Often patients are assessed for response at several clinical visits, or estimates of survival are of interest at several times (e.g., 2 yr., 3 yr.). Sometimes there are several different types of endpoints (e.g., survival, disease-free survival) that need to be assessed. Each of these examples falls into the category of multivariate response, and several methods were introduced to analyze this type of data structure.

The methods of this thesis were illustrated for three clinical trial examples. The first example involved a randomized clinical trial at two centers (strata) to compare two treatments (test, placebo) for 111 patients with a respiratory disorder. Global ratings of patient status (0 - 4) at each of four visits during the treatment period were the response criteria, and there were three covariables.

In Chapter 2 we examined the difference in means between two treatment groups. The respiratory disorder clinical trial was used for this purpose. The initial outcome of interest was the sum of favorable responses across the four visits. Covariate adjusted analyses showed that after adjusting for random imbalance the estimate of test treatment became larger, indicating that random imbalance favored placebo. There was also variance reduction with covariate adjustment, and the combination of the two dynamics resulted in a more significant test of treatment. Center was initially treated as a covariate, but later in Chapter 2 these nonparametric linear modeling methods were extended to the stratified case. Center stratified analyses were performed using two different methods of stratification, and various weighting schemes for combining information across strata. A method for assessing covariate x treatment interaction was also introduced. In the second half of Chapter 2, the concept of multivariate response was introduced where each response from the four visits could be modeled jointly. Average treatment response across visits was of interest, and a test for homogeneity across visits was obtained. These
multivariate methods were shown for average differences in proportions for treatment groups as well as differences in ordinal scores.

In Chapter 3 we continued to use the respiratory disorder clinical trial. For each visit, response was dichotomized as favorable (yes/no). A logistic regression model was fit for each visit, predicting odds of favorable response for treatment with or without covariate adjustment. We observed that the standard error increased slightly and the point estimate for treatment increased to a greater extent leading to a more significant test of treatment for each visit. We then introduced an extension to the linear models methodology where logit response was modeled with linear covariate adjustment. These results showed a slight increase in the parameter estimate and a reduction in the standard error estimate leading to a more significant test of treatment relative to the unadjusted model. The p-values from the linear model with logit response agreed very closely with the p-values from the maximum likelihood (ML) traditional logistic regression. We then applied the logit response model with Center stratification. In the latter portion of Chapter 3 we assessed the traditional proportional odds model results for each visit. We then modeled proportional odds logits with linear covariate adjustment, and compared results with the ML methods. Goodness of fit tests were assessed for both types of models. A pattern similar to what was seen with logistic regression was observed.

The second example used in this thesis was a randomized clinical trial which compared 3 doses of test treatment and placebo (placebo, 50 mg, 100 mg, 200 mg) for 959 patients with ALS. With 18 months of follow-up, the primary response criterion was survival. There were 21 covariables specified as a priori candidates for adjustment, and there were two strata. In Chapter 4, we introduce the concept of hypothesis testing. In contrast to an estimation covariance matrix, a randomization covariance matrix was used in hypothesis testing, and its form was described in this chapter. Hypothesis testing was
assessed by using logrank and Wilcoxon scores with stratification. The results of traditional survival analysis were shown (Kaplan-Meier curves, Cox regression models). The covariate adjusted Cox regression model provided substantially more significant tests of test treatment versus placebo relative to its unadjusted counterpart. This result caused some concern about whether the Cox regression model fit the data from this trial. A covariate adjusted logrank (or Wilcoxon) hypothesis test was introduced where, like the previous chapters, covariate adjustment was performed by constraining that the difference in mean covariates be equal to zero between test treatment and placebo. With covariate adjustment, the parameter estimate for test treatment tended to became larger indicating that random imbalance favored the placebo group. In addition, the standard error of the parameter was reduced with the net result being a more significant test of treatment. The p-values from this hypothesis testing where of the same magnitude as the corresponding test of treatment in the Cox model.

Also in Chapter 4 we introduced bivariate testing methods where one could test the logrank and Wilcoxon tests jointly, test the logrank scores at more than one point in time, or one could combine multiple tests by standardizing the scores and creating a single summary measure. This summary measure could then be assessed as a univariate response. In addition, testing for trends across multiple doses was also illustrated.

In Chapter 5 we continued to examine the ALS clinical trial. Two estimation parameters were introduced in this chapter in order to estimate confidence intervals for treatment effect in the survival setting. Estimation methods for incidence density measures were described, and results were provided for each of three intervals of time. Average incidence density was then obtained by averaging across intervals, and a test of homogeneity of treatment effect across intervals was shown. These methods were initially provided for pairwise comparisons (test treatment versus placebo), but extensions
were also made to the multiple dose setting where trends in treatment were of interest. In the latter half of Chapter 5, survival rate estimation was introduced. Results of interval-specific and average treatment effect estimation were provided. Like the incidence density parameter, homogeneity of treatment effect across intervals of time for survival rates was assessed. Whether there was a linear trend in doses was also examined.

The last example used in this thesis involved a randomized clinical trial to compare a chemotherapy regimen with observation for 618 patients with surgically resected Stage III colon cancer. Patients were free of disease at the time of randomization. The outcomes of interest were survival and disease-free survival. There were three covariates based on disease characteristics, and there was one stratification factor with three levels. This last example was used in order to provide further evidence that the results found in previous chapters (especially Chapter 4) were not just isolated results from one study, but instead provided patterns that could be seen if applied to any randomized trial.

Kaplan-Meier curves and Cox regression models were initially assessed for this cancer clinical trial. There was little change in the treatment parameter estimate or the corresponding standard error in the Cox regression models when there was adjustment for covariates. The covariate adjusted logrank test, which was introduced in Chapter 4, was applied to this example. Adjustment for random imbalance provided a smaller estimate of treatment effect, indicating that random imbalance favored treatment. In addition, the standard error estimate decreased relative to the unadjusted logrank test. The net result of a reduced parameter estimate and a decrease in the standard error was approximately the same p-value for the unadjusted and covariate adjusted tests. The p-values from the covariate adjusted logrank tests agreed with the p-values from the covariate adjusted Cox regression models. This clinical trial provided an interesting example of how the
adjustment for random imbalance and the accompanying variance reduction can cancel each other out in terms of significance level if random imbalance favors treatment. Because both survival and disease-free survival were of interest in this study, a bivariate test of both outcomes using logrank scores was performed with and without covariate adjustment.

Nonparametric analysis of covariance with linear models has essentially no assumptions and provides increased statistical power through variance reduction. Such variance reduction is evident in the smaller standard errors that covariance adjustment with the nonparametric methods provides for the estimates of treatment differences for the three previously mentioned clinical trial examples. Counteraction of random imbalances in the distributions of covariables for the treatment groups is another product of covariance adjustment. For nonparametric methods with linear models, such counteraction can either make the adjusted estimate for the difference between treatments larger (or smaller) than its unadjusted counterpart in a random manner that corresponds to whether the random imbalance of the covariables provides less favorable or more favorable prognosis for better response to the patients with test treatment versus those with placebo. More specifically, the adjusted estimate for such counteraction is smaller for the cancer clinical trial, and larger for the respiratory disorder and ALS trials.

The p-values from covariance adjustment with logistic regression models or Cox regression models tend to agree well with those corresponding to nonparametric methods with linear models. This would seem to indicate that the adjusted treatment parameter in the log-linear model is usually as reasonable for its population counterpart as the adjusted estimate from the linear model.
7.2 Proposed Future Study

7.2.1 Missing Data Methods

In addition to the research methods described in this thesis, other topics remain which could be identified as future research. Methods described in this thesis depended on complete data vectors. However, it is often unrealistic to assume no missing data. Future research could be focused on methods to deal with missing data issues. We have already done some preliminary work in this area.

Preliminary methods for missing data have been applied to a randomized clinical trial for Alzheimer’s disease. There are two treatment groups, test treatment and placebo, there is one covariate, baseline Alzheimer function score, and patients are stratified by Center. Patients were assessed at 4 time points after randomization, and an Alzheimer function response score was assessed at each visit. Many patients have missing response scores at one or more visits.

Residuals were obtained for each individual at each visit based on a linear regression model which fit Center and baseline score as predictors of Alzheimer score, and this model was fit separately for each visit. Individuals with missing scores were not included in the regression models. For patients with missing scores, a residual score equal to zero was substituted for each missing response. By using this imputation feature we are assuming that those individuals who receive a residual equal to zero are predicted perfectly by their baseline and center information, and treatment is considered uninformative. This has the effect of drawing treatment closer to the null.

In addition to the residuals, an indicator of missing response (Yes/No) was created for each individual at each visit. For each treatment group within each stratum, a response vector was created which took the following form:
\( f_{hj} = \{ \overline{y}_{hj1}, \overline{y}_{hj2}, \overline{y}_{hj3}, \overline{y}_{hj4}, \bar{m}_{hj1}, \bar{m}_{hj2}, \bar{m}_{hj3}, \bar{m}_{hj4}, \bar{b}_{hj} \} \)

where \( h \) indexes the \( q \) centers, and \( j \) indexes the 2 treatment groups. The response \( \overline{y}_{hju} \) is the average Alzheimer score, \( \bar{m}_{hju} \) is the average prevalence of missing values, and \( \bar{b}_{hj} \) is the average baseline score for the \( j \)-th treatment group from the \( h \)-th stratum at the \( u \)-th visit. It is then possible to fit a stratified linear model in which a multivariate residual response vector is fit. Covariate adjustment for baseline status could be performed. Also, it would be possible to constrain the difference in proportion of missing values to be equal between the two treatment groups. It would also be of interest to fit this model to the ranks of the residuals.

The method described above is of particular interest for intent-to-treat analyses where it is desirable for all randomized individuals to be included in the analysis regardless of level of compliance. By constraining the difference in missing indicators to be equal to zero between treatment groups, we would be evaluating treatment effect under the comparability of missing prevalence. In this case, the goodness-of-fit (GOF) test would be informative because missingness is not randomized so we cannot assume the expected value of the difference is equal to zero. A significant GOF test would indicate that there is substantial imbalance with regard to a missing data pattern, and this missingness may not be ignorable.

7.2.2 Stepwise Covariate Selection

There are various methods for selecting covariates to be used in an analysis of covariance. There is often concern surrounding the stepwise method of covariate selection. It could be that adjusting for covariates selected by a stepwise method may yield treatment effect confidence intervals which are narrower than one is entitled. Coverage probability may be less than thought.
It would be possible to explore, through simulations, various aspects of this issue. The first step would be to create a super population based on a distributional structure. One would then proceed in one of two directions: hypothesis testing or estimation. For hypothesis testing one would, under the null hypothesis of no treatment effect, draw a sample from the superpopulation and randomize that sample to treatment. A stepwise covariate selected model would be identified, and treatment would be tested with and without covariate adjustment. The next step would be to re-randomize treatment to the same sample, identify a stepwise model and test treatment. Re-randomization and hypothesis testing would be performed numerous times. In addition, it would be of interest to draw several different samples and perform re-randomization and hypothesis testing on these additional samples to demonstrate replication of findings across samples.

The other aspect, estimation, would be performed under the alternative hypothesis with some treatment group differences. One would draw a simple random sample stratified by treatment group with some known correlation structure. One would draw a sample, identify a stepwise model, and calculate a confidence interval for treatment. The next step would be to draw a different sample, identify a stepwise model, and calculate a confidence interval. Repeated sampling would be performed to get multiple estimates of treatment effect.

7.2.3 Extensions to U-Statistics

Another possible area of future study would be variations on U-statistic formulation. Some preliminary illustrations and discussions of applications of U-statistics to nonparametric analysis of covariance can be found in Koch et al 82, particularly for contingency table data. Additional discussion in the setting of case record data can be found in Carr et al 89. For ordered data, Mann-Whitney or Kendall's tau type U-statistics are appealing because it is not necessary to assign scores or assume
proportional odds. One can obtain a response vector which includes the U-statistic and covariates. Aspects of the response vector and covariance matrix components have been discussed in previous publications (Koch et al, 82; Carr et al, 89). Improving the "friendliness" of these computations and extending these methods to more than two treatment groups would be of interest.

7.2.4 Clustered Data

Some clinical trials randomize treatment to groups of individuals (e.g., clinic). Individuals within a cluster can be thought of as being correlated. The methods in this thesis could be extended to cases where there is a cluster design.

7.2.5 Stratification

In Chapters 4 and 5, it was pointed out that it was not straightforward to apply Mantel-Haenszel weights to combine strata when there are more than two treatment groups. Alternative ways to combine data, especially when there is an interest in trend or multi-group comparisons could be explored further.

7.2.6 Extent of Covariate Adjustment

Section A.7 provides an expression for describing the extent of covariate adjustment in a model. For this thesis, we only presented results when there was no covariate adjustment and when there was complete covariate adjustment. It would be possible to assess robustness of results under various extents of covariate adjustment.
APPENDIX

For the notation and equations in this appendix, subscripts index individuals, treatment
groups, and covariates as follows:

\[ i = 1 \text{ to } n_j \text{ where } i \text{ identifies an individual in the } j^{th} \text{ treatment group} \]

\[ j = 1 \text{ to } s \text{ where } j \text{ identifies the treatment group} \]

\[ c = 1 \text{ to } m \text{ where } m \text{ identifies the number of covariates in the model} \]

A.1 Creating Response Vectors

A.1.a General Methodology

Averages of covariates and responses are obtained for each treatment group.

\[ y_{ij} \text{ is a response for the } i^{th} \text{ individual in the } j^{th} \text{ treatment group} \]

\[ y_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \text{ the average response value for the } j^{th} \text{ treatment group} \]

\[ \bar{x}_j \text{ is a } m \times 1 \text{ column vector which contains the average for each of the covariates that} \]

\[ \text{will be used in the analysis of covariance, } \bar{x}_{cj} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{icj} \]

where \( c = 1, 2, \ldots, m \) covariates.

\( f_{ij} \) is a vector of observed values for response and covariates from the \( i^{th} \) individual in the \( j^{th} \) treatment group

\[ f_{ij} = [y_{ij}, x_{1ij}, x_{2ij}, \ldots, x_{mij}]' \]

The key assumption for this linear models methodology is that the \( f_{ij} \) arise from a simple
random sample from an infinite population. If not specifically arising from SRS, vectors
from individuals need to represent a relevant conceptual population in such a sense.
\( f_j \) is a vector of average response functions for the \( j^{th} \) treatment group with dimensions for univariate response of \((1 + m) \times 1\) and it takes the general form:

\[
f_j = [y_j \bar{x}_{1j} \bar{x}_{2j} \ldots \bar{x}_{mj}]' \quad (A.1)
\]

A vector \( f \) is then formed which stacks the \( f_j \) vectors (one from each treatment group) on top of each other vertically. For univariate response, \( f \) has dimensions \((s)(1 + m) \times 1\) where \( s \) is the total number of treatment groups being modeled.

For the case with two treatment groups, \( f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \)

The response variable can take many forms. In some cases we are interested in the average response for each treatment group, and in other cases we are more interested in the difference in average response between treatment groups. In many cases this involves two treatment groups, but sometimes there are more than two groups and contrasts among these groups are of interest. In the following sections each type of response that is illustrated in this dissertation will be defined.

**A.1.1 Response Function for Linear Response (Chapter 2)**

For the clinical trial example in Chapter 2, the response \( (y_{ij}) \) can take several forms: (1) it can take the values 0, 1, 2, 3, or 4 corresponding to the number of favorable responses across the respective visits for a given individual, (2) it can be the response score for a given visit (values 0 - 4), or (3) it can take values 0 or 1, indicating favorable response Yes/No at a given visit. \( \bar{x}_j \) is a \( m \times 1 \) column vector which contains the average for each of the covariates. For the example in Chapter 2 there are four covariates: Age (in years), Gender (male=1, female=0), an indicator for Center 1, and Baseline Score for the \( f^{th} \) treatment group so, \( \bar{x}_{cj} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{icj} \) where \( c = 1, 2, 3, 4 \). For some illustrations in Chapters 2 and 3, Center is treated as a covariate and sometimes it is treated as a stratification factor.
A.1.2 Response Function for Dichotomous Logit Response (Chapter 3)

The same clinical trial example that was used in Chapter 2 is also used in Chapter 3. Averages of covariates and outcomes are obtained for each of the 2 treatment groups as is specified in the general methodology. \( y_{ij} \) is a response for the \( i^{th} \) individual in the \( j^{th} \) treatment group. In the first half of this chapter we are assessing the dichotomous response at a given visit (out of 4 possible visits) where 1 = favorable response, and 0 = otherwise for a given individual.

\[ y_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \]

the proportion of favorable response for the \( j^{th} \) treatment group at a given visit. The outcome of interest for this example is the logit of \( \bar{y}_j \), i.e. \( \log \left[ \frac{\bar{y}_j}{1-\bar{y}_j} \right] \). This logit response takes the place of the linear response in the response vector, \( \mathbf{f}_j \) and \( \mathbf{f} \).

\( \bar{x}_j \) is the same \( m \times 1 \) column vector which contains the average for each of the covariates as in the previous linear response section because we are using the same covariates as in Chapter 2: Age, Gender (male=1, female=0), an indicator for Center 1, and Baseline Score.

\[ \mathbf{f}_j = \begin{bmatrix} \text{logit}(\bar{y}_j) & \bar{x}_{1j} & \bar{x}_{2j} & \ldots & \bar{x}_{mj} \end{bmatrix} \quad (A.2) \]

A.1.3 Response Function for Ordinal Logit Response (Chapter 3)

The first half of Chapter 3 concerns a single logit for each visit. Now we are interested in modeling multiple logits based on the ordinal response score. It is necessary to define adjacent dichotomies between each of the ordinal scores where the score can take values 0, 1, 2, 3, or 4. With five values, there are four possible dichotomies which can be formed at each visit: (a) 0 vs. 1-4, (b) 0-1 vs. 2-4, (c) 0-2 vs. 3-4, and (d) 0-3 vs. 4. For proportional odds modeling, it is necessary to have 5-10 individuals categorized for each possible outcome. However, for the latter visits there were no individuals in the treatment group who had poor response, and so the ordinal score categories 0 and 1 were combined into one group. From this collapsing, there now are three resulting dichotomies: (a) 0-1 vs. 2-4, (b) 0-2 vs. 3-4, and (c) 0-3 vs. 4.
Three indicators $d_{ijl}$ were created for each individual for each visit where $i$ indexes individuals from 1 to $n_j$, where $j$ indexes treatment group, and $l = 1$ to 3 indexes the particular dichotomy.

Coding is specified as follows:

if score = 0 or 1 then $d_{1ij} = 1; d_{2ij} = 1; d_{3ij} = 1;
if score = 2 then $d_{1ij} = 0; d_{2ij} = 1; d_{3ij} = 1;
if score = 3 then $d_{1ij} = 0; d_{2ij} = 0; d_{3ij} = 1;
if score = 4 then $d_{1ij} = 0; d_{2ij} = 0; d_{3ij} = 0;

For each treatment group, the mean of each of the dichotomy indicators is estimated as:

$$\bar{d}_{ij} = \frac{1}{n_j} \sum_{i=1}^{n_j} d_{ij}$$

A logit is then created for each of these dichotomy averages, $\text{logit}(\bar{d}_{ij}) = \log\left(\frac{\bar{d}_{ij}}{1-\bar{d}_{ij}}\right)$

For this example with 3 cumulative logits, the mean response vector for a given treatment group takes the following form:

$$f_j = [\text{logit}(\bar{d}_{1j}), \text{logit}(\bar{d}_{2j}), \text{logit}(\bar{d}_{3j}), \bar{x}_{1j}, ..., \bar{x}_{m_j}]' \quad (A.3)$$

A.1.4 Response Function for Logrank and Wilcoxon Scores (Chapter 4)

Because the primary interest with Wilcoxon and Logrank Scores is hypothesis testing, true parameter estimates are not really obtained from these linear models. The parameters are differences in average logrank or Wilcoxon scores between treatment groups with or without covariate adjustment. These scores are used when time-to-event is the outcome of interest.
A.1.4.a Logrank Scores

Consider the case when there are two treatments \((s = 2)\).

A logrank score is assigned to each individual based on their survival data, i.e. event and censoring information. Logrank scores are approximately functions of the logarithms of the survival function.

Let there be \(r\) intervals (times of failures and censoring) which are indexed by \(p\).

\(n_{p0} = \) the number of censored observations during the \(p\)-th interval

\(n_{p1} = \) the number of events during the \(p\)-th interval

\(n_{(r+1),0} = \) the number of censored individuals after the \(r\)-th interval

\(n_{(r+1),1} = \) the number of events after the \(r\)-th interval

\(n_p = (n_{p0} + n_{p1})\)

Logrank scores are assigned to individuals as follows:

(a) \(w_{p,0} = -\sum_{k=1}^{p} \frac{n_{k+1}}{\sum_{i=1}^{n_i}}\), a person censored at interval \(p\) has a logrank score which is equivalent to the summing of the ratio (number of events/number at risk) for intervals 1 through \(p\).

(b) \(w_{p,1} = (1 + w_{p,0})\), a person with the event at interval \(p\) has a logrank score which is equal to 1 plus the logrank score for a person who is censored at the same interval \(p\).

(c) \(w_{(r+1),1} = w_{(r+1),0} = w_{r,0}\), those individuals who have either an event or are censored after the \(r\)-th interval receive the same logrank score as an individual censored at the \(r\)-th interval (see (a) for formula).

The larger the uncensored observation, the smaller its score. Censored observations receive negative scores. The \(\sum_{i=1}^{n} w_i = 0\), i.e. scores sum to zero for the two respective treatment groups together (where \(n_1 + n_2 = n\)).
A.1.4.b Wilcoxon Scores

Gehan's generalized Wilcoxon scores are used for Wilcoxon tests in this example. Each observation in group 1 has a survival value \( x_i \) or \( x_i^+ \), and every individual in group 2 has a value \( y_j \) or \( y_j^+ \) where a (+) sign indicates that the observation is censored. To create these scores, every observation is compared with every other observation \((n_1 + n_2 - 1\) comparisons) for each observation.

For the following expressions, \( x_i \neq x_i' \) (Equations A.5)

Define: 
\[
U_{ij} = +1 \text{ if } x_i > y_j \text{ or } x_i^+ \geq y_j \text{ (group 1 obs compared with group 2 obs)} \\
U_{ij} = +1 \text{ if } x_i > x_i' \text{ or } x_i^+ \geq x_i' \text{ (group 1 obs compared with group 1 obs)}
\]

\[
U_{ij} = 0 \text{ if } x_i = y_j \text{ or } x_i^+ < y_j \text{ or } y_j^+ < x_i \text{ (compare grp 1 and 2)}
\]

\[
U_{ij} = 0 \text{ if } x_i = x_i' \text{ or } x_i^+ < x_i' \text{ or } x_i'^+ < x_i \text{ (compare grp 1 w/ grp 1)}
\]

\[
U_{ij} = -1 \text{ if } x_i < y_j \text{ or } x_i \leq y_j^+ \text{ (compare grp 1 and 2)}
\]

\[
U_{ij} = -1 \text{ if } x_i < x_i' \text{ or } x_i \leq x_i'^+ \text{ (compare grp 1 with grp 1)}
\]

A \( U_{ij} \) score (-1, 0, or 1) is given to the result of every pairwise comparison and \( U_i \) is the sum of these comparisons for a given observation. Let \( U_i, i = 1, 2, \ldots, (n_1 + n_2) \) be the number of the remaining \((n_1 + n_2 - 1)\) observations in which the \(i\)-th observation is definitely greater minus the number which it is definitely less. The \((n_1 + n_2)\) \( U_i \)'s define a finite population with mean 0. For the stratified case, a type of midrank standardization (range from -1 to 1) is performed within each stratum \( \{u_i/(n_h + 1)\} = u_{im} \) so that larger strata do not carry more weight due to larger Wilcoxon scores when they are combined with smaller strata.

Mantel's simplified method for calculating the Gehan Wilcoxon score is used for examples in this dissertation (Mantel 67). Assigning logrank or Wilcoxon scores is based on (a) what the respective treatment comparison is, and (b) whether there is stratification present. For the clinical trial example in Chapter 4, there are 4 treatment arms. Logrank
or Wilcoxon scoring was performed for each of the relevant dose comparisons based on the subset of individuals being used in the direct comparison. For the comparison of any active dose of treatment versus placebo and linear trend analyses across all doses, all individuals in the clinical trial were pooled together and scored. In the non-stratified case, those individuals who received Dose 50 would have a logrank score for the 50 versus placebo comparison and another logrank score for the active treatment versus placebo comparison. Other comparison logrank scores (e.g. 100 mg. vs. placebo) would be set equal to missing for persons receiving the 50 mg dose. The same method is used for Wilcoxon scoring. For the stratified case, scores are assigned to individuals within each stratum separately, i.e., the sum of scores within a given stratum equal zero.

For both the logrank and Wilcoxon tests, the response of interest is \( \bar{w}_j \), the average logrank (or Wilcoxon) score for the \( j \)-th treatment group. This response is included in \( f_j \) as the mean vector is created for each treatment group. Scores are reassigned for each pairwise treatment comparison. The mean response vector takes the following form:

\[
f_j = [\bar{w}_j, \bar{x}_{1j}, ..., \bar{x}_{mj}]' \quad (A.6)
\]

A.1.5 Response Function for Incidence Density (Chapter 5)

Based on the length of follow-up and number of events in a randomized clinical trial, the range of follow-up time is divided into several intervals. Ideally, there should be at least 10 events in each interval for each treatment group when there is no covariate adjustment, and 15 – 20 events per interval for each treatment group when there is covariate adjustment.
For the example in Chapter 5 which has 18 months of follow-up, we chose 3 intervals: 0 – 6 months (0 – 180 days), 6 – 12 months (180 – 365 days), and 12 – 18 months (365 – 540 days).

Death or tracheostomy is considered the event of interest (event = 1, otherwise event = 0). Futime = follow-up time in days for an individual from randomization to event or censoring. Patients who have an event or are censored in a prior interval contribute nothing to event indicators or follow-up time in subsequent intervals.

Indicators and follow-up time variables were created for each interval for each individual. 

(A) person had event or is censored during the first interval

if follow-up time \( \leq 180 \) days then time1 = futime; time2 = 0; time3 = 0;

\quad if event = 0 (no event) then event1 = 0; event2 = 0; event3 = 0;

\quad if event = 1 (event) then event1 = 1; event2 = 0; event3 = 0;

(B) person had event or is censored during the second interval

if 180 days < follow-up time \( \leq 365 \) days then

time1 = 180; time2 = futime – 180; time3 = 0;

\quad if event = 0 (no event) then event1 = 0; event2 = 0; event3 = 0;

\quad if event = 1 (event) then event1 = 0; event2 = 1; event3 = 0;

(C) person had event or is censored during the third interval

if 365 days < follow-up time \( \leq 540 \) days then

time1 = 180; time2 = (365 – 180); time3 = futime – 365;

\quad if event = 0 (no event) then event1 = 0; event2 = 0; event3 = 0;

\quad if event = 1 (event) then event1 = 0; event2 = 0; event3 = 1;
(D) person had no event during the 3 intervals (follow-up time \( > 540 \) days) then
time1 = 180; time2 = (365 – 180); time3 = (540 - 365); (full followup at each
interval) event1 = 0; event2 = 0; event3 = 0; (no event throughout the 3 intervals)

Consider \( k = 1 \) to \( t \) where \( k \) identifies the interval of time
\[ T_{ijk} = \text{the followup time for the } i^{th} \text{ individual in the } j^{th} \text{treatment group for the } k^{th} \text{ interval.} \]
\[ y_{ijk} = \text{indicator of event (1 = event, 0 = no event) for the } i^{th} \text{ individual in the } j^{th} \text{treatment group for the } k^{th} \text{ interval.} \]

For the example in Chapter 5, there are three intervals of time \((t = 3)\). Each individual
has a data vector comprised of event indicators, follow-up measures, and 21 covariates.
\[ f_{ijk} = [y_{i1j}, T_{ij1}, y_{i2j}, T_{ij2}, y_{i3j}, T_{ij3}, x_{ij1}, \ldots, x_{ij21}]' \]
which has length = 27 for this particular example.

For each of \( j \) treatment groups, a mean response vector is formed:
\[ f_j = [\log(\text{ID})_{j1}, \log(\text{ID})_{j2}, \log(\text{ID})_{j3}, \bar{x}_{j1}, \ldots, \bar{x}_{j21}]' \tag{A.7} \]
where
\[ \log(\text{ID})_{jk} = \log \left\{ \frac{\frac{1}{n_j} \sum_{i=1}^{n_j} y_{ijk}}{\frac{1}{n_j} \sum_{i=1}^{n_j} T_{ijk}} \right\} = \log \left\{ \frac{\bar{y}_{jk}}{\bar{T}_{jk}} \right\} \tag{A.8} \]

In the case where \( j = 2 \), \[ f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \]

The log incidence density (log ID) for a given treatment group \( j \) during interval \( k \), is the
log of the ratio of the average number of events for the \( j^{th} \) treatment group during the \( k^{th} \)
interval relative to the average number of days at risk for the \( j^{th} \) treatment group during
that same interval \( k \).
\[ \bar{x}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \] has length \( m \) or in this particular case 21 (number of covariates), and
includes the average covariate values for the \( j^{th} \) treatment group. None of these
covariates are time dependent. For this example there are $t$ responses of interest, the $t$ average log(incidence density) measures for each treatment group.

A.1.6 Response Function for Survival Rates (Chapter 5)

Based on the length of follow-up of a clinical trial, define $k = 1$ to $t$ intervals in which to estimate survival rates for treatment groups. In this ALS clinical trial, there were 18 months of follow-up. Those followed throughout the trial with no event are censored at 18 months. Intervals were divided as follows: 0 – 12 months, 12 – 15 months, and 15 – 18 months; however, smaller intervals could also have been defined. Differences in survival rates between treatment groups at 12, 15, and 18 months are the outcomes of interest. Survival rates are estimated as follows:

For each interval $k$, define two variables for each individual $i$ from treatment group $j$ in the study: $S_{ktij}$ which identifies whether an individual $i$ from group $j$ survives through the $k$-th interval (no tracheostomy or death), and $R_{ktij}$ which identifies whether the individual from group $j$ contributed follow-up during the $k$-th interval. Individuals fall into one of three categories for each interval:

(A) Individuals followed throughout the entire $k$-th interval without an event are

given the following values: $S_{ktij} = 1$ and $R_{ktij} = 1$. All prior intervals each have $S_{1ij}$

through $S_{(k-1)ij}$ and $R_{1ij}$ through $R_{(k-1)ij}$ equal to 1. For subsequent intervals without complete followup, see (B).

(B) Those with no event observed during the $k$-th interval, but who also do not

have complete follow-up for the entire interval (i.e., censored) receive the following

values: $S_{ktij} = 0.5$, $R_{ktij} = 0.5$. All prior intervals (1, 2, ..., $k - 1$) have $S$ and $R$

indicators set equal to 1 because prior intervals have complete follow-up. All subsequent $S$ and $R$ indicators ($k + 1$ through $t$) are set equal to zero because no additional follow-up or event is contributed by this individual. By assigning a value of 0.5 for both at risk ($R$)
and survival \((S)\) for the \(k\)-th interval, we are assuming that censoring is uniform over the duration of the \(k\)-th interval. Other conventions are possible for handling censoring with the two extremes being (i) assuming censored individuals are events, and (ii) assuming censored individuals are survivors.

(C) Those with an event during the \(k\)-th interval receive the following values:
\(S_{kij} = 0, \ R_{kij} = 1\). As in the censored case, all subsequent \(S\) and \(R\) indicators (intervals \(k + 1\) through \(t\)) are set equal to zero because this patient is no longer at risk for the event. All prior \((k - 1)\) intervals have complete followup so \(S\) and \(R\) indicators for intervals 1, 2, ..., \((k - 1)\) are equal to 1.

When this variable assignment is completed, each individual will have \(t\) survival indicator variables, \(S\), and \(t\) at risk indicators \(R\). For each treatment group \(j\), the mean of each of the \(S_{kij}\) and \(R_{kij}\) indicators are calculated as follows: \(\bar{S}_{kj} = \frac{1}{n_j} \sum_{i=1}^{n_j} S_{kij}\), and
\(\bar{R}_{kj} = \frac{1}{n_j} \sum_{i=1}^{n_j} R_{kij}\).

Survival rates are then estimated as follows:
\[
P_{kj} = \frac{\bar{S}_{kj}}{\bar{R}_{kj}}
\]  \hspace{1cm} (A.9)

\(P_{kj}\) is the conditional probability for a member of the \(j\)-th group to survive the \(k\)-th interval given that the individual survived the \(k - 1\) previous intervals, so the overall cumulative survival rate (\(SR_{kj}\)) for the \(k\)-th interval is the product of the probabilities of surviving through all prior intervals including the \(k\)-th interval \((1, 2, ..., k)\),

\[
SR_{kj} = \prod_{i=1}^{k} P_{ij}
\]  \hspace{1cm} (A.10)

A mean response vector \(f_j\) is created for each of the \(j\) treatment groups. The vector will have length \(2t + m\), which for this particular example corresponds to the 3 survival indicators, the 3 at risk indicators, and 21 covariates which will be used in the analysis of covariance.
\( f_j \) takes the following form for each treatment group:

\[
\left[ \overline{S}_{1j} \quad \overline{R}_{1j} \quad \overline{S}_{2j} \quad \overline{R}_{2j} \quad \overline{S}_{3j} \quad \overline{R}_{3j} \quad \overline{x}_{1j} \quad \overline{x}_{2j} \ldots \overline{x}_{mj} \right]'^T \quad (A.11)
\]

For \( s = 2 \) treatment groups, the mean response vectors are stacked vertically, \( f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \)

A survival rate response vector is then formed by taking the following steps.

1. \( \log f = \log(f) \), which has length \( = (s)(2t + m) \), in this case \( (2)(6 + 21) = 54 \).

2. Define a matrix \( A_{ij} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0_{1x21} \\ 1 & -1 & 1 & -1 & 0 & 0 & 0_{1x21} \\ 1 & -1 & 1 & -1 & 1 & -1 & 0_{1x21} \\ 0_{21x1} & 0_{21x1} & 0_{21x1} & 0_{21x1} & 0_{21x1} & 0_{21x1} & I_{21} \end{bmatrix} \quad (A.12) \)

\( 24 \times 27 \)

Because there are 2 treatment groups, \( A_{1j} \) is duplicated for each treatment group, \( I_2 \otimes A_{1j} = A_1 \) which has dimensions \( 48 \times 54 \).

\( A_1 \log f = A_1 \cdot \log f \), where \( (A_1 \log f) \) has dimensions \( 48 \times 1 \).

This matrix multiplication involves taking \( \{ \log(\overline{S}_{kj}) - \log(\overline{R}_{kj}) \} \) for each interval, and summing the cumulative intervals for each treatment group separately. Summing the difference in \( \{ \log(\overline{S}_{kj}) - \log(\overline{R}_{kj}) \} \) across intervals is comparable to multiplying the \( P_{kj} \) across intervals on the linear scale.

It is important that no covariate mean values are \( \leq 0 \) for any treatment group because we are taking the natural log of these mean values. However, it is easy to add a constant value to each treatment covariate mean so that the new value is positive. This does not cause problems because we are testing differences between treatment group means, and adding the same constant to all groups does not affect the magnitude of difference between groups.

3. \( \exp[A_1 \log f] = da \), which transforms values from the log scale back to the linear scale
(4) define $A2 = [I_{24} \ - I_{24}]$ where the dimensions of $I$ are based on $(t + m)$. In this case, 3 intervals of time and 21 covariates.

$d = A2 * da$, which is our goal, the difference in mean responses (survival rates and covariates) for the two treatment groups with length $(t + m)$.

A.2 Covariance Matrices

There are two strategies available for addressing the covariance structure of the compound vector $d = (d_y \ d_x)'$ of differences for the response variable and the covariates jointly. One is solely for nonparametric hypothesis testing with no statistical assumptions concerning the method of selection for the patients in the clinical trial; and the other is for situations where both confidence intervals are of interest and the patients in the clinical trial represent some very large conceptual population in a sense comparable to a simple random sample. The former uses the covariance matrix $V_0$ that exactly applies from the randomization distribution and it applies under the null hypothesis that each patient has the same response to each of the two treatments. The other strategy for addressing the covariance matrix of $d$ uses the estimator $V_*$ Under the assumption that the patients in the clinical trial represent a very large population in a sense which is comparable to a simple random sample, the random matrix $V_*$ is an unbiased estimator (Koch et al 82) for the unconditional covariance matrix of $d$. The following sections specify in detail what the covariance structures are for each type of response of interest.

A.2.1. Estimation

The estimated covariance matrix of $f$ is $V_f$.

$V_f = \text{block diagonal of } (V_{f_1}, V_{f_2}, \ldots V_{f_s})$  \hspace{1cm} (A.13)
where $V_{fj}$ is the covariance matrix for the $j^{th}$ treatment group and there are $s$ treatment groups. $V_f$ has null matrices in the off-diagonal positions because the treatment groups are independent of one another.

$V_{fj}$ takes the form:

$$
\frac{1}{(n_j)(n_j-1)} \sum_{i=1}^{n_j} \begin{bmatrix}
(y_{ij} - \bar{y}_j)(y_{ij} - \bar{y}_j)' \\
(x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)'
\end{bmatrix}
\begin{bmatrix}
(y_{ij} - \bar{y}_j)(y_{ij} - \bar{y}_j)' \\
(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)'
\end{bmatrix}
$$

(A.14)

$V_{fj}$ can also be expressed as:

$$
\begin{bmatrix}
V_{YY,j} & V_{XY,j} \\
V_{XY,j}' & V_{XX,j}
\end{bmatrix}
$$

In the univariate case, $V_{fj}$ has dimensions $(1 + m) \times (1 + m)$, so it follows that the dimensions of $V_f$ are $(s)(1 + m) \times (s)(1 + m)$. For the example in Chapter 2 (linear response) and the first half of Chapter 3 (single logit response) there is 1 response and 4 covariates, $V_{fj}$ is $5 \times 5$, and because there are two treatment groups, $V_f$ is $10 \times 10$ and is comprised of two block matrices, $V_{f1}$ and $V_{f2}$.

A.2.1.a Linear Covariance (Estimation) - Chapter 2

Because the response of interest is just the mean score (ordinal or proportion), and not any function of $\bar{y}$, the estimated covariance matrix used for estimation is just $V_f$, i.e., $V = V_f$.

A.2.1.b Logistic Covariance (Estimation) - Chapter 3

We need to take into account the fact that we are using a function of $\bar{y}$, i.e., logit of $\bar{y}$.

Using the Taylor's series expansion to estimate the variance of the function, it has been shown that the first derivative of the function with respect to the random variable provides the pre- and post-multiplier which leads to a consistent estimator. More specifically,

$$
\frac{d}{d\bar{y}_j} [\log(\frac{\bar{y}_j}{1-\bar{y}_j})] = \frac{1}{\bar{y}_j(1-\bar{y}_j)}, \quad \frac{d}{d\bar{x}_{cj}}(\bar{x}_{cj}) = 1
$$
where \( c \) indexes covariates and takes values from 1 to \( m \) and \( j \) indexes treatment.

To incorporate the derivative of the function of each of the random variables, a diagonal matrix, \( H_j \), is created for each treatment group. Its diagonals are the first derivatives of the functions of means and takes the following form:

\[
H_j = \text{diagonal}[\frac{1}{\bar{y}_{ij}(1-\bar{y}_{ij})}, 1, \ldots, 1_{[m]}]
\]  \hspace{1cm} (A.15)

where there are \( m \) values of 1 on the diagonal to correspond with the \( m \) covariates. The dimensions of \( H_j \) are \((1 + m) \times (1 + m)\)

\( H \) is then formed as a diagonal matrix with \([H_1, H_2, \ldots, H_s]\) as the block diagonals. There are zero's in the off-diagonal positions. The dimensions are then \( s(1 + m) \times s(1 + m) \). The overall estimated covariance matrix used for parameter estimation with logit response is then:

\[
V = HV_fH' \hspace{1cm} (A.16)
\]

which also has dimensions \( s(1 + m) \times s(1 + m) \).

A.2.1.c Proportional Odds Covariance (Estimation) - Chapter 3

The covariance matrix structure for the proportional odds model found in the latter portion of Chapter 3 takes a similar form to that for the logistic regression model found in the earlier portion of the same chapter. Instead of just one logit there are \( l = 3 \) cumulative logits to model. The dimensions of \( V_f \) are now \( s(l + m) \times s(l + m) \) instead of the single logit dimensions of \( s(1 + m) \times s(1 + m) \). In addition, there are now \( l = 3 \) logits for which to take the derivative of the function of means for each treatment group.

\( H_j \) takes the following form:

\[
H_j = \text{diagonal}[\frac{1}{\bar{y}_{ij}(1-\bar{y}_{ij})}, \frac{1}{\bar{y}_{2j}(1-\bar{y}_{2j})}, \ldots, \frac{1}{\bar{y}_{lj}(1-\bar{y}_{lj})}, 1, \ldots, 1_m] \hspace{1cm} (A.17)
\]

As with the logistic regression model, \( H = \text{block}(H_1, H_2, \ldots, H_s) \)  \hspace{1cm} (A.18)

The resulting covariance matrix \( V = HV_fH' \) and has dimensions \( s(l + m) \times s(l + m) \)
A.2.1.d Incidence Density Covariance (Estimation) - Chapter 5

define \( z_{ij} = [d_{ij1}, y_{ij1}, d_{ij2}, y_{ij2}, d_{ij3}, y_{ij3}]' \) (A.19)

and \( \bar{z}_j = [\bar{d}_{j1}, \bar{y}_{j1}, \bar{d}_{j2}, \bar{y}_{j2}, \bar{d}_{j3}, \bar{y}_{j3}]' \) (A.20)

\( V_{fj} \) takes the form :

\[
\left( \frac{1}{\eta_j(n_j-1)} \right) \sum_{i=1}^{n_j} \begin{bmatrix}
    (z_{ij} - \bar{z}_j)(z_{ij} - \bar{z}_j)' & (z_{ij} - \bar{z}_j)(x_{ij} - \bar{x}_j)'
    
    (x_{ij} - \bar{x}_j)(z_{ij} - \bar{z}_j)' & (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)'
\end{bmatrix}
\] (A.21)

where \( k = 1 \) to \( t \) and indexes intervals of time.

The outcome of interest is comprised of a ratio of two random variables, average number of outcomes for interval \( k \) for treatment group \( j \), \( (\bar{d}_{jk}) \), and average number of days of follow-up time for interval \( k \) for treatment group \( j \), \( (\bar{y}_{jk}) \). The response is not linear, but instead is on the log scale (log incidence density). For the example in Chapter 5, \( t = 3 \) intervals, and there are 21 covariates to be used in the ANCOVA. To account for this log-linear function, we take the derivative of the log response function with respect to each random variable.

\[
\log(\frac{\bar{d}_{jk}}{\bar{y}_{jk}}) = \log(\bar{d}_{jk}) - \log(\bar{y}_{jk})
\]

\[
\frac{d(\log(\bar{d}_{jk}))}{d(\bar{d}_{jk})} = \frac{1}{\bar{d}_{jk}} , \text{ and } \frac{d(-\log(\bar{y}_{jk}))}{d(\bar{y}_{jk})} = -\frac{1}{\bar{y}_{jk}} , \text{ and } \frac{d}{d(x_{mj})} = 1
\]

\( H_j \) = block diagonal\( [(\frac{1}{d_{j1}} - \frac{1}{y_{j1}}), (\frac{1}{d_{j2}} - \frac{1}{y_{j2}}), ... , (\frac{1}{d_{jt}} - \frac{1}{y_{jt}}), 1_1, ..., 1_{21}] \) (A.22)

For this particular example, \( H_j \) has dimensions 24 x 27, \( (t + m) \times (2t + m) \)

\( V_j = H_j V_{fj} H_j' \) with dimensions 24 x 24 \((3 \log(\text{ID}) + 21 \text{ cov}) \times (3 \log(\text{ID}) + 21 \text{ cov})\)

If \( j = 2 \) treatment groups. \( f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \) with dimensions 48 x 1.

\( V = \text{block diagonal}[V_1, V_2] \) with dimensions 48 x 48

With a mean response vector \( (f) \), and a corresponding covariance matrix \( (V) \) one can specify design matrices to estimate treatment parameters and form confidence intervals.
A.2.1.e Survival Rates Covariance Matrix - Chapter 5

The covariance matrix of the response means of \( f \) is \( V_f \).

\[
V_f = \text{block diagonal of } (V_{f_1}, V_{f_2}, \ldots V_{f_s})
\]

where \( V_{f_j} \) is the variance for the \( j^{th} \) treatment group. \( V_f \) has null matrices in the off-diagonal positions because the treatment groups are independent of one another.

\( V_{f_j} \) takes the form:

\[
\left( \frac{1}{(n_i)(n_j-1)} \right) \sum_{i=1}^{n_i} \begin{bmatrix}
(z_{ij} - \bar{z}_j)(x_{ij} - \bar{x}_j)' \\
(z_{ij} - \bar{z}_j)(x_{ij} - \bar{x}_j)'
\end{bmatrix}
\]

(A.23)

where for this particular case with \( t = 3 \) time intervals,

\[
z_{ij} = [S_{1ij} \ R_{1ij} \ S_{2ij} \ R_{2ij} \ S_{3ij} \ R_{3ij}]'
\]

(A.24)

\[
\bar{z}_j = [\bar{S}_{1j} \ \bar{R}_{1j} \ \bar{S}_{2j} \ \bar{R}_{2j} \ \bar{S}_{3j} \ \bar{R}_{3j}]'
\]

(A.25)

where \( S_{ki} \) is an indicator of survival, and \( R_{ki} \) is an indicator of being at risk for the \( i \)-th individual from the \( j \)-th treatment group during the \( k \)-th interval.

\( V_{f_j} \) can also be expressed as:

\[
\begin{bmatrix}
V_{ZZ,j} & V_{XZ,j} \\
V_{ZX,j} & V_{XX,j}
\end{bmatrix}
\]

\( V_{f_j} \) has dimensions \((2t + m)x(2t + m)\), so it follows that the dimensions of

\( V_f \) are \((s)(2t + m)x(s)(2t + m)\). In this particular example, \( V_{f_j} \) is \((6 + 21)x(6 + 21)\),

and because there are two treatment groups, \( V_f \) is \( 56 \times 56 \) and is comprised of two block matrices, \( V_{f_1} \) and \( V_{f_2} \).

In the survival rate estimation section (A.1.6) we specify that the \( \log(\frac{S_{zi}}{R_{ki}}) \) is initially taken. Because we are taking a function of the response means for each treatment group at each interval, we must perform the corresponding adjustment to the covariance structure. The following steps are taken:
(1) A diagonal vector is created, $D1$, where the diagonal elements of the matrix are the elements of $f$, where $f$ is the vertically stacked vector of mean responses defined in A.1.6. $\text{diag}(f) = D1$.

(2) $\text{inv}(D1) = INV \times D1$, is the inverse of the diagonal matrix, $D1$. This is done to account for the fact that each of the response means have been converted to the log scale, and $\frac{d}{d(\log f)} = \text{inv}(f)$.

(3) A diagonal matrix is created whose elements are $da$ as was specified in A.1.5 ($= A1 \times \text{log} f$). This new diagonal matrix is called $D2$.

(4) Pre and Post multiply $V_f$ in the same order as the functions are being applied to the response means vector, $f$. $A1$ and $A2$ are the same matrices specified in the response means section (see A.1.6).

$$V = A2 \times D2 \times A1 \times INV \times D1 \times V_f \times INV \times D1 \times A1' \times D2 \times A2'$$ (A.26)

For this particular example where we are interested in the difference between two treatment groups over 3 time intervals, and 21 covariates to adjust for, $V$ has dimensions 24 x 24.

This matrix $V$ is the resulting covariance matrix for the corresponding difference in mean vectors including a difference in average survival rates between treatment groups, $d = f_1 - f_2$ (see Section A.5 for specifications on difference parameterization).

A.2.2. Hypothesis Testing - Chapters 4 and 6

Because we are hypothesis testing and not estimating confidence intervals, a randomization covariance matrix is used. This type of covariance matrix is used in Chapters 4 and 6. Under $H_0$, we assume that the logrank (or Wilcoxon) scores have
expected values which are the same for all treatment groups, and under the principle of randomization the covariate expected values are the same for all treatment groups.

\[
V_o = \left( \frac{n}{(n_1)(n_2)(n-1)} \right) \sum_{i=1}^{n} \left[ (w_{ij} - \bar{w})(w_{ij} - \bar{w})' \right. \\
\left. (w_{ij} - \bar{w})(x_{ij} - \bar{x})' \right]
\]

(A.27)

where \( n = n_1 + n_2 \) and \( \bar{w} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij}, \) and \( \bar{x} = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} x_{ij} \)

A.3 Univariate Response Parameter Estimation

A.3.1 Unrestricted Covariate Adjustment

The methodology in this section assumes that response data are in the form \( f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \)

where there are two treatment groups.

A design matrix, \( X_{unadj} \), can be specified which allows means for covariates to vary randomly among the treatment groups. Treatment effect which is unadjusted for random covariate imbalance is estimated in the form of 1 treatment parameter when one uses the following reference cell coding for two treatment groups.

The unadjusted design matrix for the univariate response case in Chapters 2 and 3 takes the form:

\[
X_{unadj} = \begin{bmatrix} I_{1+m} & 0_{(1+m)\times(1+m)} \\ I_{1+m} & I_{1+m} \end{bmatrix} = \begin{bmatrix} 1 & 0_{4x4} & 0_{4x1} & 0_{4x4} \\ 0_{4x1} & I_{4x4} & 0_{4x1} & 0_{4x4} \\ 1 & 0_{4x4} & 1 & 0_{4x4} \\ 0_{4x1} & I_{4x4} & 0_{4x1} & I_{4x4} \end{bmatrix}
\]

(A.28)

\[ s(1 + m) \times s(1 + m) \quad 2(5) \times 2(5) \]

This design matrix would be used in the setting of Chapters 2 and 3 where univariate response is being modeled (e.g. differences between two treatment groups with regard to proportions, ordinal scores, one logit (logistic) at one particular visit).
This model provides estimates for the treatment effect parameter relative to the reference level (placebo). Each of the covariate means (Age, Gender, Center, Baseline) is allowed to vary randomly for each of the two treatment groups. $X_{\text{unadj}}$ has dimensions $10 \times 10$. Given the design matrix $X_{\text{unadj}}$, the Beta vector is estimated by:

$$\hat{\beta}_{\text{unadj}} = (X'_{\text{unadj}} V^{-1} X_{\text{unadj}})^{-1} X'_{\text{unadj}} V^{-1} f$$

(A.29)

$$= X_{\text{unadj}}^{-1} f$$

and the variance of $\hat{\beta}_{\text{unadj}}$, $V_{\beta} = (X'_{\text{unadj}} V^{-1} X_{\text{unadj}})$

(A.30)

where $f$ and $V$ are the appropriate response mean vector and covariance matrix ($V$ for estimation, $V_o$ for hypothesis testing), respectively, as specified in sections A.1 and A.2 for the outcome of interest.

For example, here is a description of Beta parameters resulting from the $X_{\text{unadj}}$ design matrix for the clinical trial example in Chapter 2 and Chapter 3 for univariate response:

<table>
<thead>
<tr>
<th>Element of $\beta$ vector</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>average response (proportion, ordinal score, logit) for placebo</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>average of covariate 1 for the placebo group</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>average of covariate 2 for the placebo group</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>average of covariate 3 for the placebo group</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>average of covariate 4 for the placebo group</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>difference for response average (proportion, ordinal score, logit) between treatment and placebo</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>difference for covariate 1 average between treatment and placebo</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>difference for covariate 2 average between treatment and placebo</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>difference for covariate 3 average between treatment and placebo</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>difference for covariate 4 average between treatment and placebo</td>
</tr>
</tbody>
</table>

$\beta_6$ is the parameter of interest, the unadjusted treatment effect.
A.3.1.a Unadjusted Hypothesis Testing (Univariate Response)

The same methodology specified in A.3.1 can be used to test whether there is a significant difference in average logrank (or Wilcoxon) scores between treatment groups. In the example in Chapter 4 there are 21 covariates specified in the analysis plan so \( X_{unadj} \) has dimensions \( (1 + 21) \times (1 + 21) \) and takes the general structure as the matrix in equation A.28. Applying the estimation equations, \( V = V_\beta \), the randomization covariance matrix, is used for hypothesis testing.

\[
\hat{\beta}_{unadj} = (X'_{unadj} V^{-1}_o X_{unadj})^{-1} X'_{unadj} V^{-1}_o f \quad (A.31)
\]

\[
V_\beta = (X'_{unadj} V^{-1}_o X_{unadj}) \quad (A.32)
\]

A.3.2 Covariate Adjusted Estimation of Treatment Effect for Univariate Response

The unadjusted design matrix can usually be simplified so that the difference in average covariate values among treatment groups is restricted to be zero since randomization in study design supports equivalence of groups for averages of covariates. The resulting design matrix for the univariate response modeled in Chapters 2 and the first half of Chapter 3 (single logit) is:

\[
X_{adj} = \begin{bmatrix}
I_{m+1} & 0_{(m+1)\times 1} \\
1 & \end{bmatrix} = \begin{bmatrix}
1 & 0_{1\times 4} & 0 \\
0_{4\times 1} & I_{4\times 4} & 0 \\
1 & 0_{1\times 4} & 1 \\
0_{4\times 1} & I_{4\times 4} & 0 \\
\end{bmatrix}
\]

\[
s(m + 1) \times (m + 2) = 10 \times 6
\]

After applying the estimating equations of A.29, A.30 using the \( X_{adj} \) design matrix, the resulting \( \beta_{adj} \) vector has 6 parameters for the example in Chapter 3 when \( m = 4 \) covariates.
The first element ($\beta_1$) represents the average of the response (proportion, ordinal score, or logit) for the placebo group.

The next four elements ($\beta_2 - \beta_5$) are the means of the four covariates for the pooled treatment groups, i.e. restricting that the means are equivalent between the two groups.

The last parameter ($\beta_6$) represents the adjusted difference in average response (proportion, ordinal score, or logit) between test treatment and placebo.

This is the covariate adjusted test of treatment, the parameter of interest.

**A.3.3 Assessing Goodness of Fit**

A test of goodness of fit (GOF) can be performed on the constrained model relative to the unrestricted model where the covariates are allowed to vary among treatments. The test takes the form:

$$ (f - \hat{f})'V^{-1}(f - \hat{f}) = Q $$  \hspace{1cm} (A.34)

where $Q$ is distributed as a chi-square with

$$ \text{degrees of freedom (df)} = \dim(\beta_{unadj}) - \dim(\beta_{adj}) $$

$\hat{f}$ is estimated by $(X_{adj})\hat{\beta}_{adj}$

$f$ is the appropriate column vector of average response and covariate values for the $s$ treatment groups (see A.1) from the unconstrained model which is also equal to $(X_{unadj})\hat{\beta}_{unadj}$, and $V$ is the corresponding estimation ($V$) or hypothesis testing ($V_0$) covariance matrix for the response of interest (see A.2).

In the comparison of the models specified by $X_{adj}$ and $X_{unadj}$ for univariate response in Chapters 2 and 3, the degrees of freedom for the GOF test would be: $DF = (10 - 6) = 4$.

This is equal to the number of covariate constraints that were imposed between the two treatment groups (i.e., age, gender, center, and baseline).
A.4 Modeling Multivariate Response

A.4.1 Multivisit Response

Consider the subscript \( k = 1 \) to \( v \) where \( v \) indicates the number of response (or visit) when there is more than one response. For the example in Chapters 2 and 3, \( v = 4 \) visits. Sometimes it is of interest to model the 4 visits simultaneously in a multivariate model. The methods previously described for univariate response (A.3.1, A.3.2) can be expanded to encompass the case when there are \( v \) responses (visits).

\( f_j \) is now a \( (v + m) \) vector reflecting \( v \) average response measures (one response for each visit) and \( m \) average covariate values for the \( j^{th} \) treatment group. For the clinical trial example in Chapters 2 and 3 where there are 4 visits and 4 covariates, the multivariate \( f_j \) has dimensions \( 8 \times 1 \).

\[
f_j = [y_{1j}, \ldots, y_{vj}, x_{1j}, \ldots, x_{mj}]'
\]  
(A.35)

and \( V_{f_j} \) is now a \( (v + m) \times (v + m) \) covariance matrix (whereas univariate \( V_{f_j} \) has dimensions \( (1 + m) \times (1 + m) \)). Estimation equations for the multivariate case are the same as for the univariate case where the corresponding multivariate response vectors and covariance matrices are used (see equations A.29, A.30 for estimation and A.31, A.32 for hypothesis testing).

A.4.1.a Unadjusted Treatment Test for Multivariate Case

The unadjusted design matrix for the multivariate case where there are four visits, four covariates, and two treatment groups takes the form:

\[
X_{\text{unadj, MV}} = \begin{bmatrix}
I_{(v+m)} & 0_{(v+m) \times (v+m)} \\
I_{(v+m)} & I_{(v+m)}
\end{bmatrix} = \begin{bmatrix}
I_4 & 0_{4 \times 4} & 0_{4 \times 4} & 0_{4 \times 4} \\
0_{4 \times 4} & I_4 & 0_{4 \times 4} & 0_{4 \times 4} \\
I_4 & 0_{4 \times 4} & I_4 & 0_{4 \times 4} \\
0_{4 \times 4} & I_4 & 0_{4 \times 4} & I_4
\end{bmatrix}
\]  
(A.36)

\[
s(v + m) \times s(v + m) \\
2(4 + 4) \times 2(4 + 4)
\]

The \( \beta_1 - \beta_4 \) parameters are the average responses (proportions, ordinal scores, logit) for the placebo group for visits 1 - 4, respectively.
\( \beta_6 - \beta_8 \) are the average of the four covariate values (age, gender, center, baseline) respectively, for placebo.

\( \beta_9 - \beta_{12} \) are the estimated treatment parameters for visits 1 - 4 respectively (unadjusted difference in mean response between the two treatment groups)

\( \beta_{13} - \beta_{16} \) are the differences in average covariate values between treatment and placebo.

Each treatment parameter can be interpreted as the difference in average response (proportion, ordinal score, logit) between active and placebo for each visit without covariate adjustment. When response is measured at multiple times, it is often of interest to test whether treatment effect is homogenous across visits (treatment x visit interaction).

For the unadjusted, multivariate model this question can be addressed by specifying the following contrast matrix:

\[
C_{\text{unadj}, t \times v} = \begin{bmatrix}
0_{1 \times 8} & 1 & -1 & 0 & 0 & 0_{1 \times 4} \\
0_{1 \times 8} & 1 & 0 & -1 & 0 & 0_{1 \times 4} \\
0_{1 \times 8} & 1 & 0 & 0 & -1 & 0_{1 \times 4}
\end{bmatrix} (v - 1) \times s(v + m)
\]  

(A.37)

Each test statistic is formed as follows:

\[
\hat{C}\beta_{\text{unadj}} = \hat{\theta}, \ Var(\hat{\theta}) = \hat{C}V_{\hat{\beta}}C'
\]

(Wald Chi-square test = \( (C\beta_{\text{unadj}})'(CV_{\hat{\beta}}C')^{-1}(C\beta_{\text{unadj}}) \))  

(A.38)

(A.39)

df = number of rows of the corresponding \( C \) contrast matrix, in this case \((v - 1)\) df test.

A.4.1.1 Covariate Adjusted Multivariate Analysis

The covariate adjusted multivariate model has a similar form as the univariate model:

\[
X_{\text{adj}, MV} = \begin{bmatrix}
I_{(v+m)} & 0_{(v+m) \times v} \\
I_{(v+m)} & 0_{(m \times v)}
\end{bmatrix} = \begin{bmatrix}
I_{4} & 0_{4 \times 4} \\
0_{4 \times 4} & I_{4} \\
I_{4} & 0_{4 \times 4} \\
0_{4 \times 4} & I_{4}
\end{bmatrix} (A.40)
\]

\[ s(v + m) \times (2v + m) \quad 16 \times 12 \]
By applying estimating equations A.29-30 to the multivariate response vector and covariance matrix and specifying the restricted design matrix, there are \(2v + m\) resulting parameters. The \(\beta_1 - \beta_4\) parameters are the average responses (proportions, ordinal scores, logit) for the placebo group for visits 1 - 4, respectively. \(\beta_5 - \beta_8\) are the average of the four covariate values (age, gender, center, baseline) for the two treatment groups. i.e., restricting the covariate means to be equal across treatment groups.

\(\beta_9 - \beta_{12}\) are the estimated adjusted treatment parameters for visits 1 - 4 respectively (difference in mean response between the two treatment groups, adjusting for random imbalance among the 4 covariates).

The goodness of fit (GOF) test comparing the covariate adjusted and unadjusted model is the same as in the univariate case (A.3.3). It has degrees of freedom equivalent to the number of covariate constraints applied to the adjusted model regardless of whether the response criteria is univariate or multivariate.

**A.4.1.c Assessing Treatment x Visit Interaction**

The treatment x visit interaction can also be assessed for the covariate adjusted model with a Wald test. The corresponding contrast matrix is:

\[
\mathbf{C}_{adj(t x v)} = \begin{bmatrix}
\mathbf{0}_{1x8} & 1 & -1 & 0 & 0 \\
\mathbf{0}_{1x8} & 1 & 0 & -1 & 0 \\
\mathbf{0}_{1x8} & 1 & 0 & 0 & -1 \\
\end{bmatrix}
\]

(A.41)

\((v - 1) \times (2v + m)\)

If the treatment x visit interaction \((v - 1\) df) is not significant for the covariate adjusted model (implying homogeneous treatment effect), it is possible to simplify the estimation of the treatment parameter by assessing average treatment effect across visits using a column vector instead of four separate treatment parameters.
This can be seen in the following design matrix:

\[
X_{adj, MV, no\, interaction} = \begin{bmatrix}
I_4 & 0_{4x4} & 0_{4x1} \\
0_{4x4} & I_4 & 0_{4x1} \\
I_4 & 0_{4x4} & I_{1x1} \\
0_{4x4} & I_4 & 0_{4x1}
\end{bmatrix}
\] (A.42)

16 x 9

The ninth $\beta$ parameter would be interpreted as the covariate adjusted average treatment effect across the four visits (average difference between treatment group and placebo in response of interest).

**A.4.2 Multiple Test Statistics - Logrank and Wilcoxon Multivariate Tests**

Some of the illustrations in Chapter 4 show the assessment of two logrank tests (two different truncation times), or a logrank and Wilcoxon test at the same time in a 2 degree of freedom global test. This situation also falls under the category of multivariate response. One can think of each scoring (type of score or truncation point) as a different response. For these examples $k = 2$ responses, and the clinical trial used in Chapter 4 has 21 covariates specified for adjustment, therefore $f_j$ has length $(2 + 21)$ and $V_o$ (hypothesis testing covariance matrix) has dimension $23 \times 23$.

To evaluate the logrank test at two different truncation points, one must assign logrank scores using two different sets of follow-up information. For the earlier time point, all events and censoring that occur after the first timepoint are censored at the first timepoint. Logrank (or Wilcoxon) scores are then assigned based on this truncated information. For assigning scores to the longer followup interval, additional survival information that was truncated previously between timepoints 1 and 2 is allowed to be used. A new set of scores is assigned for information up to and including the latter timepoint.
A.4.2.a No Covariate Adjustment - Logrank or Wilcoxon Multivariate (MV) Tests

Because we are in the hypothesis testing mode, we are not interested in mean response for each treatment group. A mean logrank score for a treatment group has no meaningful interpretation. What is of interest is the difference in mean response between treatment groups. Difference in mean parameterization (see A.5) is used for this purpose. The model with no covariate adjustment can be specified with the following design matrix:

\[
X_{\text{unadj}, \text{mv}} = [I_{k+m}] = [I_{23}]
\]  
(A.43)

There are 23 resulting parameters when this unadjusted design matrix is used with the estimating equations (A.31, A.32). These parameters represent the difference in average logrank or Wilcoxon score \((\beta_1, \beta_2)\) for each time point, and difference in average 21 covariate values \((\beta_3, ..., \beta_{23})\) between the two treatment groups, respectively.

A covariate adjusted design matrix can be specified which takes the following form when two treatment groups are being compared:

\[
X_{\text{adj}, \text{mv}} = \begin{bmatrix} I_{(k)} \\ 0_{(m+k)} \end{bmatrix} = \begin{bmatrix} I_2 \\ 0_{21 \times 2} \end{bmatrix}
\]  
\((k + m) \times k 
\quad 23 \times 2
\)  
(A.44)

When one applies the estimation equations (A.31, A.32), there are 2 resulting parameters. These parameters \((\beta_1, \beta_2)\) are the covariate adjusted difference in average scores (logrank, Wilcoxon, or both) between the treatment groups. The difference in average covariate values between treatment groups are constrained to be equal to zero for each of the 21 covariates.

The next step is to specify a contrast matrix, either unadjusted or covariate adjusted, which selects both of the treatment parameters simultaneously.

\[
C_{\text{unadj}, \text{mv}} = \begin{bmatrix} 1 & 0 & 0_{1 \times 21} \\ 0 & 1 & 0_{1 \times 21} \end{bmatrix}
\]  
(A.45)

\[
C_{\text{adj}, \text{mv}} = [I_2]
\]  
(A.46)
A Wald test is then performed which tests whether the difference in adjusted (or unadjusted) mean scores (logrank, Wilcoxon, or both) for the two tests are jointly equal to zero. This test takes the following form (where $C_{mv}$ corresponds to either the adjusted or unadjusted contrast matrix depending on the covariate setting, and $\beta$ can be either the adjusted or unadjusted estimated parameter counterpart):

$$C_{mv}^\wedge \beta = \theta, \ Var(\theta) = C_{mv}^\wedge V_\beta C_{mv}^{\prime \wedge}$$ (A.47)

Wald Chi-square test $= (C_{mv}^\wedge \beta)'(C_{mv}^\wedge V_\beta C_{mv}^{\prime \wedge})^{-1}(C_{mv}^\wedge \beta)$ (A.48)

The degrees of freedom for the Wald test are equal to the number of parameters being jointly tested (number of rows of $C_{mv}$) which in this case equals 2 for both the adjusted and unadjusted case for the examples in Chapter 4.

A.4.3 Proportional Treatment Parameter Estimation

A.4.3.a Response Vector and Estimation

In the previous two sections we addressed the situation where "multivariate" implied multi-visit or multiple logrank (or Wilcoxon) testing. However, several estimation methods in this thesis involve estimating multiple parameters which are not thought of as being multiple visits. Proportional odds models involve estimating multiple ($l$) cumulative logits, and both incidence densities and survival rates estimate treatment effect over $t$ intervals. These situations can also be thought of as multivariate because more than one treatment parameter is being estimated.

We can think of proportional odds, incidence density, and survival rates in the same generic notation, i.e., there are $k$ estimates of treatment effect that correspond to cumulative logits or intervals of time. These can be modeled in the same manner as the previous section with or without covariate adjustment. The length of the mean response vector for each treatment group, $f_j$, has length $k + m$ where $m$ is the number of
covariates to be adjusted. Either cell mean parameterization or difference in mean parameterization may be used with specification of the corresponding design matrix (adjusted or unadjusted) and application of equations A.29, A.30.

A.4.3.b Testing Proportionality Assumption

For incidence density, proportional odds, and survival rates, one of the more compelling questions of interest is whether these \( k \) estimates of treatment are proportional to one another. In other words, (a) is the proportional odds assumption appropriate? (b) is the treatment effect (magnitude of difference in response between the two treatments) constant across intervals of time? This can be tested in the unadjusted and covariate adjusted setting, and the form of the contrast is analogous to testing whether there is a treatment x visit interaction as was explored in section A.4.1.c. These methods of testing proportionality will be illustrated for the difference parameterization (see A.5).

\[
X_{unadj, MV} = \begin{bmatrix} I_{k+m} \end{bmatrix} \tag{A.49}
\]

\[
X_{adj, MV} = \begin{bmatrix} I(k) \\ 0_{(m+k)} \end{bmatrix} \tag{A.50}
\]

where \( k \) is the number of treatment parameters estimated (cumulative logits, incidence densities, survival rates) and \( m \) is the number of covariates to be adjusted for.

Proportionality is tested by specifying the following contrast matrix for the unadjusted or adjusted case as appropriate (for this example we will assume \( k = 3 \)):

\[
C_{unadj, prop} = \begin{bmatrix} 1 & -1 & 0 & 0_{1 \times m} \\ 1 & 0 & -1 & 0_{1 \times m} \end{bmatrix} \tag{A.51}
\]

\[
C_{adj, prop} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \tag{A.52}
\]

A Wald statistic can be specified which takes the following form (where \( C_{prop} \) and \( \beta \) can be either the adjusted or unadjusted components):

\[
\text{Wald Chi-square test} = (C_{prop} \beta)^\intercal (C_{prop} V_\beta C_{prop}^\intercal)^{-1} (C_{prop} \beta) \tag{A.53}
\]
where the degrees of freedom of the test are equal to \# of rows of $C_{prop} = (k - 1)$

If there is compatibility with the assumption of proportional treatment effect (across cumulative logits or intervals of time) then it is appropriate to average the $k$ estimates of treatment effect to come up with one summary treatment effect measure.

The following unadjusted and adjusted design matrices average the $k$ treatment parameters when difference parameterization (see A.5) is used:

$$X_{unadj, prop} = \begin{bmatrix} 1_k & 0_{k \times m} \\ 0_{m \times 1} & I_m \end{bmatrix} \quad (A.54)$$

$(k + m) \times (1 + m)$

$$X_{adj, prop} = [1_k] \quad (A.55)$$

$k \times 1$

For the unadjusted covariate proportional model, there are $(1 + m)$ resulting parameters. The first parameter ($\beta_1$) can be interpreted as the average treatment parameter over $k$ cumulative logits or intervals of time. In the Chapter 5 proportional incidence density example, the parameter of interest is the difference in mean response (log ID) between test treatment and placebo averaged across the $t$ intervals of time without covariate adjustment. The additional $m$ parameters ($\beta_{k+1}$ through $\beta_{k+m}$) are the differences in average covariate values between the two treatment groups for the $m$ covariates, where $m = 21$, respectively.

For the adjusted covariate model there is only one resulting parameter. It has the interpretation of being the covariate adjusted average treatment effect (e.g., log odds ratio, log ID) over $k$ cumulative logits or intervals of time. This parameter is the adjusted difference in mean response between test treatment and placebo averaged across $k$ values.
The appropriateness of this estimate is based on the previously specified test of homogeneity of treatment effects.

A.4.3.c GOF for Covariate Adjustment and Proportional Odds

Section A.3.3 discusses how the goodness-of-fit test (GOF) can be performed to assess the compatibility of constraining the difference in mean covariate values to be equal to zero between treatment groups. In Section A.4.3.b a strategy for testing the assumption of homogeneity of treatment effect is discussed. It is also possible to test both of these hypotheses simultaneously. This GOF test can be split into two non-overlapping parts:

1. homogeneity assumption, \( df = (s - 1)(k - 1) \)
2. restricting covariates to be equal, \( df = (s - 1)(m) \)

where \( s \) is the number of treatment groups, \( k \) is the number of logits (or intervals) in which treatment is estimated, and \( m \) are the number of covariates that are being adjusted.

If two of the test results are known, the third unknown test and degrees of freedom can be deducted from the following two simple equations:

Combined GOF test = (test of homogeneity assumption) + (test of covariate assumption)

GOF df = (homogeneity df) + (covariate restriction df) = \( (s - 1)(k + m - 1) \)

A.5 Mean Difference Parameterization

This method of modeling will be illustrated for the case when there are two treatment groups. After calculating \( f_1, f_2, \) and \( V_{F1}, V_{F2} \) (as previously specified for the response means parameterization in Sections A.1, A.2), a difference vector is created:

\[
d = f_1 - f_2anumber{A.56}

which has dimensions \((k + m) \times 1\). This method can be used in the univariate \((k = 1)\) or multivariate \((k > 1)\) response setting, and the mean difference vector has length equal to the number of responses + the number of covariates in the model. In the estimation setting (see A.2.1), the covariance matrix for \( d \) is equal to:
\[ V_d = \text{Var}(f_1) + \text{Var}(f_2) = V_{f_1} + V_{f_2} \quad (A.57) \]

There is no covariance component between the two treatment groups because the covariance matrix is calculated separately for each respective treatment group, and the two groups are independent.

In the hypothesis testing mode (see A.2.2), the covariance matrix is estimated from the pooled sample, and not each treatment group individually. Therefore, there is a covariance between the treatment groups that must be accounted for.

\[
V_{do} = \text{Var}(f_1 - f_2) = \text{Var}(f_1) + \text{Var}(f_2) - 2(\text{cov} f_1, f_2)
\]

\[
= \{ \frac{n}{(n_1)(n_2)(n-1)} \sum_{i=1}^{n} \begin{bmatrix} (y_{ij} - \bar{y})(y_{ij} - \bar{y})' & (y_{ij} - \bar{y})(x_{ij} - \bar{x})' \\ (x_{ij} - \bar{x})(y_{ij} - \bar{y})' & (x_{ij} - \bar{x})(x_{ij} - \bar{x})' \end{bmatrix} \}
\]

The estimating equations for the mean difference method take a similar form to the response means method:

\[
\hat{\beta}_{diff} = (X'V_{do}^{-1}X)^{-1}X'V_{do}^{-1}d \quad (A.58)
\]

and the variance of \( \hat{\beta}_{diff} \),

\[
V_{\hat{\beta}} = X^{-1}V_{do}^{-1}X \quad (A.59)
\]

The design matrix, \( X_{unadj, diff} \), for the unadjusted case takes the simple form, \( I_{(k+m)} \), an identity matrix with dimensions \((k + m) \times (k + m)\). This allows the covariate differences between treatment groups to vary randomly while an unadjusted treatment effect is estimated for each of the \( k \) responses.

For the constrained case, \( X_{adj, diff} = \begin{bmatrix} I_k \\ 0_{m \times k} \end{bmatrix} \) \( (k + m) \times k \) \( (A.60) \)

This design matrix allows for the estimation of treatment effect for the \( k \) covariate adjusted responses while restricting the differences in the \( m \) average covariate values to be zero.
For the univariate constrained case, 

\[ X_{adj, diff} = \begin{bmatrix} 1 \\ 0_{m \times 1} \end{bmatrix} \]  

(A.61)

which is a column vector with dimensions \((1 + m) \times 1\).

### A.6 Calculating Variance Reduction

Percent variance reduction is calculated by:

\[ 1 - \left\{ \frac{s.e.{}^2(Tx\ parameter)_{unadj} - s.e.{}^2(Tx\ parameter)_{cov\ adj}}{s.e.{}^2(Tx\ parameter)_{unadj}} \right\} \times 100\% \]  

(A.62)

For the example in Table 2.2 of Chapter 2:

\[ \frac{(0.291^2) - (0.235^2)}{0.291^2} = 0.35, \]

\(.35 \times 100 = \text{percent variance reduction with covariance adjustment} = 35\%\)

There is a 35% reduction in the treatment parameter variance when the model adjusts for the effect of Center, Age, Gender, and Baseline relative to the model with no covariate adjustment.

How does this translate into savings in sample size?

\[ V = Var(Treatment)_{unadj}, V^* = Var(Treatment)_{adj} \]

Variance Ratio \(= \frac{V^*}{V} \)

\[ \left( \frac{V^*}{V} - 1 \right) = \left( \frac{n^*}{n} - 1 \right) \]

(A.63)

sample size increase for comparable power = \(\left( \frac{V^*}{V} - 1 \right) \times 100\%\).

\[ \frac{(0.291^2)}{(0.235^2)} - 1 = 0.55 \]  

so there is a sample size savings of 55% that accompanies the variance reduction of 35%.

### A.7 Extent of Covariate Adjustment

For the following expressions, a univariate response and a single covariate will be assumed in order to illustrate the concept of extent of covariate adjustment.
Using difference in mean response parameterization (see A.5), we have:

\[
\begin{align*}
g &= \begin{bmatrix} d \\ u \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} b \\
V_g &= \begin{bmatrix} V_{dd} & V_{du} \\
              V_{ud} & V_{uu} \end{bmatrix} \\
d &= \overline{y}_1 - \overline{y}_2
\end{align*}
\]  

(A.64)  (A.65)

where \(d\) is the unadjusted difference in mean response between the two groups

\[
b = \overline{y}_1 - \overline{y}_2 - \{\frac{V_{du}}{V_{uu}}(\overline{x}_1 - \overline{x}_2)\}  \tag{A.66}
\]

where \(b\) is the covariate adjusted difference in mean response between the two treatment groups.

\[
b_\lambda = \lambda b + (1 - \lambda)d, \text{ where } 0 \leq \lambda \leq 1
\]

(A.67)

\[
b_\lambda = (\overline{y}_1 - \overline{y}_2) - \lambda\{\frac{V_{du}}{V_{uu}}(\overline{x}_1 - \overline{x}_2)\}
\]

\[
b_\lambda = d \text{ when } \lambda = 0, \text{ and } b_\lambda = b \text{ when } \lambda = 1
\]

\[
\text{Var}(b_\lambda) = V_{dd} + \lambda^2(\frac{V_{du}^2}{V_{uu}})(V_{uu}) - 2\lambda(\frac{V_{du}}{V_{uu}})(V_{du})
\]

\[
= V_{dd} + \lambda^2(\frac{V_{du}^2}{V_{uu}V_{dd}})(V_{dd}) - 2\lambda(\frac{V_{du}^2}{V_{uu}V_{dd}})V_{dd}
\]

\[
= V_{dd}\{1 + (\lambda^2 - 2\lambda)R_{du}^2\} \tag{A.68}
\]

when \(\lambda = 0\), \(\text{Var}(b_\lambda) = V_{dd}\), and

when \(\lambda = 1\), \(\text{Var}(b_\lambda) = V_{dd}(1 - R_{du}^2) = V_{dd} - \frac{V_{du}^2}{V_{uu}} \tag{A.69}
\]

One can assess various values of \(\lambda\) which support the alternative hypothesis of a treatment effect. Covariate adjustment would be maximized when \(\lambda = 1\), and no covariate adjustment would be performed when \(\lambda = 0\).

**A.8 Stratification**

For the stratified case, subjects are thought of as being representative for an infinite population through a stratified simple random sample from an infinite population. One can define \(h = 1\) to \(q\) strata. In the example in Chapters 2 and 3 there are two strata based
on the number of centers in the study design, and in Chapters 4 and 5 there are two strata based on disease onset (limb vs. bulbar).

**A.8.1 Stratification Method 1 ($\beta_w$)**

When $n_{hj}$ are small (e.g., $n_{hj} \leq 20$) and more than two covariates are being used in the ANCOVA, this first method of handling stratification should be implemented. Because the covariance matrix is inverted to estimate model parameters, there is concern about estimating a stable, non-singular covariance matrix for each stratum $x$ treatment group when sample sizes are small. For this first method, one calculates $f_{11}, f_{12}, \ldots, f_{1s}, \ldots, f_{q1}, \ldots f_{qs}$ which are the mean response vectors for each stratum $x$ treatment group (for the example with two treatments and two strata: $f_{11}, f_{12}, f_{21}, f_{22}$). In addition, each $V_{hj}$ covariance matrix is estimated for each stratum $x$ treatment group. The next step is to combine treatment groups across strata so there is one resulting response vector, $f_{wj}$, and one resulting covariance matrix, $V_{wj}$ for each treatment group in a form similar to $V_j$ (see A.2). This can be accomplished in the following way:

$$
    f_{wj} = \sum_{h=1}^{2} w_h f_{hj} \quad \text{and} \quad V_{wj} = \sum_{h=1}^{q} (w_h^2) V_{hj}, \quad \text{where} \quad \sum_{h=1}^{2} w_h = 1 \quad \text{(A.71)}
$$

Some examples of $w_h$ can be: (1) $\frac{1}{q} = w_h$ where each of the $q$ strata contributes equally to the overall estimate. This would be equal to $\frac{1}{2}$ for both Centers in the example in Chapters 2 and 3, or (2) $w_h = \frac{n_h}{\sum n_h}$, i.e., weight is proportional to the sample size of the stratum relative to the total sample size. (3) Another weight to consider when combining strata is proportional to:

$$
    w_h = ((n_{h1})(n_{h2})/(n_{h1} + n_{h2}))^c \quad \text{(A.72)}
$$

where $0 \leq c \leq 1$ and weights are standardized to add to 1, and $s = 2$ treatment groups. This last class of weights fall into the category of "Mantel-Haenszel" type weights.
These weighted treatment response vectors and covariance matrices can then be used with a design matrix, as in the unstratified case, to calculate stratification adjusted model parameters using this first method for smaller sample sizes.

$$\hat{\beta}_w = (X'V_w^{-1}X)^{-1}X'V_w^{-1}f_w$$ \hspace{1cm} (A.73)

where $f_w = \begin{bmatrix} f_{w1} \\ f_{w2} \end{bmatrix}$ and $V_w = \begin{bmatrix} V_{w1} & 0 \\ 0 & V_{w2} \end{bmatrix}$ \hspace{1cm} (A.74)

where 1, 2 subscripts indicate treatment group, and $X$ can be any appropriate design matrix which may or may not adjust for covariates. Resulting parameter estimates are considered to be adjusted for the stratification factor.

The assessment of $\beta_w$ when the sample sizes for the combined strata are not large is possible through an exact p-value from the randomization distribution for its specification with $V_{wo} = \sum_{h=1}^{s} w_h^2 V_{ho}$ where $V_{ho}$ is the counterpart to $V_o$ in section A.2.2 for the $h$-th stratum. Another consideration for confidence intervals based on $\hat{\beta}_w$ is that sample sizes may not be large enough for all diagonal elements of all $V_h$ to be positive. In these situations, such diagonal elements can be replaced by their $V_{ho}$ counterparts (Koch et al, in press).

A.8.2 Stratification Method 2 ($\tilde{\beta}$)

When $n_{hj}$ are larger (e.g. $n_{hj} \geq 15\sqrt{(1 + m)}$) there is less concern about the stability of the corresponding covariance matrices, $V_{hj}$. With this method, parameters are estimated separately for each stratum ($\beta_h$) and then stratum-specific estimates are pooled across the strata. There are several choices of weights for this method. In general, there are two categories of weights: (1) explicit weights which are proportional to sample size, and (2) precision weights based on the inverse of covariance matrix weights. Weights in the first class are applicable even when there is heterogeneity among the strata. This first class of weights is preferred when there are borderline conditions for asymptotics. Examples of these weights are equal weights for each stratum, Mantel-Haenszel weights, and weights
based on sample size. These explicit weights tend to be used for the \( \beta_w \) Method of stratification assessment because one is already in a setting where there is concern about asymptotics, hence the reason why strata are combined prior to estimation. The second class of weights presumes homogeneity across strata. A nonsignificant treatment \( \times \) strata interaction would be reasonably supportive of using precision weights. Larger treatment \( \times \) strata sample sizes are usually required because an assumption of relative stability of the covariance matrix is needed. If homogeneity is reasonable to assume and sample sizes are large then precision weights are fully efficient, and the treatment estimation is more optimal.

Means for covariates and responses are estimated for each treatment group within each stratum separately. The \((m \times 1)\) vector \( \bar{x}_{hj} = \left\{ \sum_{i=1}^{n_{hj}} x_{hij} / n_{hj} \right\} \) corresponds to the means for the \( m \) covariates for treatment \( j \) at stratum level \( h \). The scaler \( \bar{y}_{hj} = \left\{ \sum_{i=1}^{n_{hj}} y_{hij} / n_{hj} \right\} \) corresponds to the mean response for treatment \( j \), stratum \( h \). A mean response vector can be created for both the response and covariables for treatment \( j \) at center \( h \).

\[
\mathbf{f}_{hj} = (\bar{y}_{hj}', \bar{x}_{hj}')'
\]

and has dimensions \((1 + m) \times 1\) for this univariate response example.

The resulting vector, \( \mathbf{f}_h \), involves stacking the \( \mathbf{f}_{hj} \) vectors into a column vector for each stratum separately.

The estimation covariance matrix \( \mathbf{V}_h \) has block diagonal form (with null matrices on the off-diagonal) where \( h \) indexes the stratum and \( j = 1 \) to \( s \) indexes treatment:

\[
\mathbf{V}_h = (\mathbf{V}_{h1}, \mathbf{V}_{h2}, ..., \mathbf{V}_{hs})
\]

(\ref{eq:75})

\( \beta_h \) are the estimated parameters for the \( h \)-th stratum which are calculated as follows:

\[
\hat{\beta}_h = (X'_h \mathbf{V}^{-1}_h X_h)^{-1} X'_h \mathbf{V}^{-1}_h \mathbf{f}_h \quad \text{and} \quad \mathbf{V}_{\hat{\beta}_h} = (X'_h \mathbf{V}^{-1}_h X_h)^{-1}
\]

(\ref{eq:76})

\( X \) is a specified design matrix which can allow for unadjusted or covariate adjusted parameter estimation. The resulting estimated parameters, \( \hat{\beta}_h \) from each stratum are then
combined across strata using an appropriate weighting scheme as described in this section.

\[
\hat{\beta} = \sum_{h=1}^{q} \hat{w}_h \beta_h \quad \text{and} \quad V_\beta = \sum_{h=1}^{q} (\hat{w}_h^2) V_h, \quad (A.77)
\]

where \( \sum_{h=1}^{q} \hat{w}_h = 1 \).

Wald tests can then be performed which take the following form:

\[
\hat{\beta} V_\beta^{-1} \hat{\beta} = Q_\beta \quad (A.78)
\]

### A.8.3 Assessing Treatment x Strata Interaction

In some cases, assessment of the homogeneity of the \( \{\beta_h\} \) across the strata may be of interest. A test statistic for the setting of scalar response is

\[
Q_{H\beta} = \sum_{h=1}^{q} \hat{w}_h (\beta_h - \tilde{\beta})^2 \quad (A.79)
\]

where \( \hat{w}_h = (1/V_{\beta_h}) \) and \( \beta = \{\sum_{h=1}^{q} \hat{w}_h \beta_h / \sum_{h=1}^{q} \hat{w}_h\} \). It has an approximately chi-square distribution with \( df = (q - 1) \) when homogeneity applies.

### A.8.4 Stratified Multivariate Case

The same methods that are described in A.4 for modeling multivariate response can also be applied to the stratified case. Either method of stratification (A.8.1, A.8.2) can be applied to the multivariate case. However, because the dimensions of \( V_{hj} \) increase \( ((k + m) \times (k + m) \) where \( k \) = number of responses) the guideline for assessing appropriate sample size for stratification methods is slightly altered. Each \( n_{hj} \) should be \( \geq 15 \sqrt{(k + m)} \) (instead of univariate criterion \( \geq 15 \sqrt{(1 + m)} \) ) in order to account for concern about the stability of the larger corresponding covariance matrix using \( \hat{\beta} \) stratification method (Section A.8.2). If sample sizes are smaller, the \( \beta_w \) method should be used.
A.9 Treatment x Covariate Interaction

The stratified approach could also be used for other categorical covariates to test
covariate x treatment interaction. One would think of each level of the covariate as a
stratum. Treatment effect parameters would be estimated within each level of the
covariate. It is then straightforward to compare the various estimated treatment
parameters across discrete covariate values by specifying a contrast matrix. This Wald
test would have degrees of freedom = (s - 1)(p - 1) where s is the number of treatments
and p is the number of levels of the covariate (e.g., for gender p = 2 levels). One specific
element of this would involve estimating a treatment parameter for males and females
separately and then comparing the treatment parameters between the two strata (gender).
This test comparing treatment vs. placebo for gender would be a 1 df chi-square test.

A.10 Multiple Treatment Arms (> 2) In The Stratified Case

A.10.1 Estimation Setting

For estimating effect across multiple treatment arms, the difference in means
parameterization is used. The response vector for the h-th stratum takes the following
form:
\[
d_h = \begin{bmatrix}
  f_{h,50} - f_{h,0} \\
  f_{h,100} - f_{h,0} \\
  f_{h,200} - f_{h,0}
\end{bmatrix} = \begin{bmatrix}
  d_{h,50} \\
  d_{h,100} \\
  d_{h,200}
\end{bmatrix}
\]

\[(s - 1)(k + m)\]

where the subscripts indicate the stratum (h) and dose of treatment (0 = placebo).
The placebo means are subtracted from the means of each of the test treatment doses in a
pairwise fashion. In the example in Chapter 5 there is one response (average difference
in log incidence density, or average difference in survival rates), 21 covariates, and 2
strata (limb/bulbar status). The vector \(d_h\) would have dimensions \((3)(1+21) \times 1 = 66 \times 1\).
Because this is a very large covariance matrix and there is some concern about
invertability of such a large matrix, we chose to adjust for a subset of the covariates when assessing each of the doses in a global fashion.

This subset of covariates was selected by performing a stepwise selection model building using Cox regression. The analysis was stratified by limb/bulbar status. No covariables were forced into the model, and treatment was not allowed as a candidate. Only the 21 potential covariates were included in the stepwise model. A criteria of $\alpha \leq 0.05$ for entry and to remain in the model was employed. There were 8 resulting covariates from this stepwise model, and they were adjusted for when global treatment and linear trend testing was performed. The resulting covariance matrix, $V$, now has dimensions $27 \times 27$ instead of $66 \times 66$.

Because we are in the estimation setting, we estimate a covariance matrix for each treatment group within each stratum separately. In this way, treatment groups are independent of one another. We know from Section A.2.1 that the covariance matrix for the difference in Dose 50 and placebo means is:

$$V(d_{h,50}) = V_h(f_{50} - f_0) = V_{h,50} + V_{h,0}$$

To expand to the multiple dose setting, we see that:

$$V_{dh} = \begin{bmatrix} V_{h,50} + V_{h,0} & V_{h,0} & V_{h,0} \\ 0 & V_{h,100} + V_{h,0} & V_{h,0} \\ V_{h,0} & V_{h,0} & V_{h,200} + V_{h,0} \end{bmatrix}$$

where the pairwise covariances are summed on each of the corresponding diagonals, and the covariance for the placebo group is located in each of the off diagonal positions.

We then combine the response vectors ($d_h$) and covariance matrices ($V_{dh}$) from the two strata. This is done by choosing a weight which is proportional to the stratum sample
size. \( w_{th} = \frac{n_h}{\sum_{h=1}^{5} n_h} \) where \( n_h = (n_{h0} + n_{h50} + n_{h100} + n_{h200}) \), which is equivalent to the stratum \( h \) sample size. The denominator of the weight is the total sample size.

The resulting stratified response vector and covariance matrix are:

\[
d = (wt_1) \ast (d_1) + (wt_2) \ast (d_2) \quad \text{and} \quad V_d = (wt_1^2 \ast V_{d1}) + (wt_2^2 \ast V_{d2})
\]

It is then straightforward to apply the standard estimating equation and design matrices which are specified in equation A.58 for differences in means parameterization.

Covariance adjustment is then performed on the stratification adjusted difference in response means vector (\( \beta_w \) method of stratification).

\[
\hat{\beta}_{diff} = (X'V_d^{-1}X)^{-1}X'V_d^{-1}d, \quad \text{and} \quad V_{\hat{\beta}} = X^{-1}V_d^{-1}X
\]

It is possible to perform a global (\( g \)) test of treatment which is equivalent to testing the null hypothesis that the difference in response for each dose of treatment with placebo is equal to zero, (50 - placebo = 0, 100 - placebo = 0, and 200 - placebo = 0). This is a 3 degree of freedom (\( s-1 \)) test which can be assessed in the unadjusted or covariate adjusted setting.

The specific contrasts in the case where there are three doses being compared to placebo, and 8 covariates identified by the stepwise model are:

\[
C_{g,\text{unadj}} = \begin{bmatrix}
1 & 0_{1 \times 8} & 0_{1 \times 9} & 0_{1 \times 9} \\
0_{1 \times 9} & 1 & 0_{1 \times 8} & 0_{1 \times 9} \\
0_{1 \times 9} & 0_{1 \times 9} & 1 & 0_{1 \times 8}
\end{bmatrix} \quad \text{and} \quad C_{g,\text{adj}} = I_3
\]

A.82

It is also possible to test whether there is a linear dose trend (\( t \)) among these groups. The contrast matrices for this test would take the following form:

\[
C_{t,\text{unadj}} = \begin{bmatrix}
-1 & 0_{1 \times 8} & 1 & 0_{1 \times 8} & 3 & 0_{1 \times 8}
\end{bmatrix} \quad \text{and} \quad C_{t,\text{adj}} = \begin{bmatrix}
-1 & 1 & 3
\end{bmatrix}
\]

A Wald chi-square test is obtained by using the corresponding unadjusted (\( \beta_{\text{unadj}} \)) or adjusted (\( \beta_{\text{adj}} \)) parameter estimates in conjunction with the appropriate contrast matrix in
the following formula, where test degrees of freedom are equal to the number of rows of \( C \):

\[
C: \quad \text{Wald Chi-square test} = (C\beta)'(C\Sigma_{\beta}C')^{-1}(C\beta)
\]

### A.10.2.a Unstratified Hypothesis Testing Setting

The same response vector that is shown in the previous section is also employed in this setting. The difference is the covariance matrix which is used for hypothesis testing. The treatment groups are not independent so the resulting covariance matrix is not block diagonal.

If we define \( V_{of} = \frac{1}{(n-1)} \sum_{j=1}^{S} \sum_{i=1}^{n_i} \begin{bmatrix} (w_{ij} - \bar{w})(w_{ij} - \bar{w})' & (w_{ij} - \bar{w})(x_{ij} - \bar{x})' \\ (x_{ij} - \bar{x})(w_{ij} - \bar{w})' & (x_{ij} - \bar{x})(x_{ij} - \bar{x})' \end{bmatrix} \)

where \( w_{ij} \) is the logrank or Wilcoxon score and \( x_{ij} \) is the vector of covariate values for the \( i \)-th individual from the \( j \)-th treatment group, and \( \bar{w} \) is the mean score and \( \bar{x} \) is the mean covariate vector for the pooled sample. For the example in Chapter 4 where there is one response measure and 8 covariates identified by the stepwise model described in the previous section, \( V_{of} \) has dimensions 9 x 9.

For the response vector: \( \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} \), the covariance matrix takes the following form where subscripts indicate treatment groups:

\[
V_o = \begin{bmatrix}
\frac{1}{n_1} (1 - \frac{n_1}{n}) & -\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} \\
-\frac{1}{n} & \frac{1}{n_2} (1 - \frac{n_2}{n}) & -\frac{1}{n} & -\frac{1}{n} \\
-\frac{1}{n} & -\frac{1}{n} & \frac{1}{n_3} (1 - \frac{n_3}{n}) & -\frac{1}{n} \\
-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \frac{1}{n_4} (1 - \frac{n_4}{n})
\end{bmatrix} \otimes (V_{of})
\]
In order to obtain the resulting response vector, \[
\begin{bmatrix}
\bar{f}_2 - \bar{f}_1 \\
\bar{f}_3 - \bar{f}_1 \\
\bar{f}_4 - \bar{f}_1
\end{bmatrix},
\]
it is necessary to apply a contrast matrix, \( C \), to the column vector of mean responses such that
\[
\begin{bmatrix}
-1 & 1 & 0 & 0 \\
-1 & 0 & 1 & 0 \\
-1 & 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\bar{f}_1 \\
\bar{f}_2 \\
\bar{f}_3 \\
\bar{f}_4
\end{bmatrix} = \begin{bmatrix}
\bar{f}_2 - \bar{f}_1 \\
\bar{f}_3 - \bar{f}_1 \\
\bar{f}_4 - \bar{f}_1
\end{bmatrix}.
\]

To obtain the resulting covariance matrix, we pre and post multiply the covariance matrix for the means by the contrast matrix. \( CV_o C' = V_{od} \), the randomization covariance matrix for the difference in means.
\[
V_{od} = \begin{bmatrix}
\frac{1}{n_1} + \frac{1}{n_2} & \frac{1}{n_1} & \frac{1}{n_1} \\
\frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_3} & \frac{1}{n_1} \\
\frac{1}{n_1} & \frac{1}{n_1} & \frac{1}{n_1} + \frac{1}{n_4}
\end{bmatrix} \otimes V_{of}
\]
which results in a 27 x 27 covariance matrix for the example in Chapter 4.

Parameters are then estimated by using the following equation with the randomization covariance matrix:
\[
\hat{\beta}_{diff} = (X'V-o^{-1}X)^{-1}X'V-o^{-1}d, \quad \text{and} \quad V_\beta = X^{-1}V-o^{-1}X
\]
It is also possible to specify the same contrast matrices for global tests and linear trends as were specified in A.10.1, with the exception that the randomization covariance matrix is used instead of the estimation covariance matrix. The same type of Wald chi-square test is the result.

A.10.2.b Stratified Hypothesis Testing Setting

For hypothesis testing relative to \( V_o \) or its stratified extensions, one can determine the test statistic \( Q_\beta \) in section A.8.2 in the following steps:

1. For each stratum, residuals are computed from the fit of the linear regression
model which only includes covariates to the data for the pooled treatment groups (i.e., treatment is not a component of the model);

2. The extended Mantel-Haenszel statistic is applied to compare the residuals from (1) for the two treatments with adjustment for the strata.

In the SAS System, the GLM Procedure is used for (1) with the specification, "response = covariates," in a "by stratum" manner; and residuals are obtained as output. For (2), the FREQ Procedure is used with the specification, "strata \times treatments \times residuals," and "table scores" as default.

A similar two-stage computational procedure can be used to obtain a test statistic which is comparable to $Q_{\beta_w}$. For this method, (1) is modified to (1W) as follows:

1W. For the combined strata, residuals are computed from the fit of the linear regression model which only includes strata and the covariates to the data for the pooled treatment groups; i.e., the model with the specification, "response = strata + covariates."

The test statistic with respect to these residuals is then determined through (2).

Analysis involving confidence intervals through $V_x$ in (A.3) or its stratified extensions or the methods outlined in Sections A.3, A.4, A.5 are straightforward to apply through computing procedures for matrix specifications; e.g., SAS IML. Their implementation is also possible through a combination of computing procedures for means and variances, correlation coefficients, and weighted least squares (e.g., the MEANS, CORR, and CATMOD Procedures in the SAS System).

A.11 Computations

A.11.a Estimation of Linear Response Using Catmod (Chapter 2)

Although all of these calculations were programmed using SAS IML, SAS Proc Catmod can be used to provide very similar results. For the unstratified case, the treatment groups
would be specified in the POPULATION statement; treatment and strata would be specified as populations for the stratified case. On the left hand side of the equal sign would be the response variable(s) and covariates. To the right of the equal sign one would specify a user defined $X$ matrix. RESPONSE MEANS would be invoked. The same function vector as is used in the IML method is used in Catmod. The only difference between the two methods involves the estimated covariance matrix for each population.

For the unstratified case: \[
\frac{(n_j - 1)}{n_j} V_{j, IML} = V_{j, Catmod}
\]

and for the stratified case: \[
\frac{(n_{hj} - 1)}{n_{hj}} V_{hj, IML} = V_{hj, Catmod}
\]

This method is applicable for the results in Chapter 2 where the response of interest is a difference in proportions or ordinal scores between treatment groups.

A.11.b Approximations with Unweighted Least Squares Computations

One can see from the expressions in equations A.29 and A.30 that the results from the methods there would seem similar to those that would come from ordinary least squares (OLS) computations for multiple linear regression for a response variable. However, to use the OLS method for formal inferential purposes, one needs to be able to argue that estimates and test statistics have support from the appropriateness of the usual assumptions for multiple linear regression; e.g., independent normal distributions with homogeneous variances, for measurements of response variables. These assumptions are much different from those needed for the methods described in A.1.a. Nevertheless, ad hoc computations with OLS can produce results that roughly approximate those from the univariate methods in A.3. In particular, this method can be applied to mean response, or logrank or Wilcoxon scores, i.e. cases where Taylor's series expansion is not required to estimate the variance and the response is univariate. In this spirit, one can view the OLS method as useful for informal or exploratory applications where the computations for the
methods in Section A.3 are not conveniently feasible; and one can view the methods in Section A.1 as supporting the robustness of OLS in situations where the assumptions for multiple linear regression do not strictly apply. These comments do not apply to formal inferential situations since one should address their needs for the methods in Section A.3 with accordance to the specifications there.

A.12 Equivalence of Various Model Estimation Methods

A.12.1 Univariate Estimation Versus Multivariate Estimation

Assume two responses, $y_{1j}$ and $y_{2j}$ for each of two treatment groups $(j = 1, 2)$ and one covariate $x$. We want to show that modeling the difference in response for each of these outcomes jointly provides the same parameter estimates as one would get if each parameter was estimated separately in a univariate model.

Define $d = \begin{bmatrix} \overline{y}_{11} - \overline{y}_{12} \\ \overline{y}_{21} - \overline{y}_{22} \\ \overline{x} - \overline{x}_2 \end{bmatrix} = \begin{bmatrix} y_{1 \text{diff}} \\ y_{2 \text{diff}} \\ x_{\text{diff}} \end{bmatrix}$ and

$$V_d = \begin{bmatrix} V(Y_1) & Cov(Y_1Y_2) & Cov(Y_1X) \\ Cov(Y_1Y_2) & V(Y_2) & Cov(Y_2X) \\ Cov(Y_1X) & Cov(Y_2X) & V(X) \end{bmatrix} = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{bmatrix}$$

where $V(Y_1) = Var(\overline{y}_{11}) + Var(\overline{y}_{12})$, $V(Y_2) = Var(\overline{y}_{21}) + Var(\overline{y}_{22})$, $V(X) = Var(\overline{x}) + Var(\overline{x}_2)$, $Cov(Y_1Y_2) = Cov(\overline{y}_{11}, \overline{y}_{21}) + Cov(\overline{y}_{12}, \overline{y}_{22})$, $Cov(Y_1X) = Cov(\overline{y}_{11}, \overline{x}) + Cov(\overline{y}_{12}, \overline{x}_2)$, $Cov(Y_2X) = Cov(\overline{y}_{21}, \overline{x}) + Cov(\overline{y}_{22}, \overline{x}_2)$.

We will generically define $V_d^{-1} = \begin{bmatrix} c_{11} & c_{12} \\ c'_{12} & c_{22} \end{bmatrix}$

We know from the definition of an inverse that:

$$\begin{bmatrix} c_{11} & c_{12} \\ c'_{12} & c_{22} \end{bmatrix} \cdot \begin{bmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$
We can produce expressions for the various elements of the inverse:
\[ c_{11}V_{11} + c_{12}V'_{12} = I, \quad c_{11}V_{12} + c_{12}V_{22} = 0 \]
\[ c_{11} = [V_{11} - V_{12}(V_{22}^{-1})V'_{12}]^{-1} \]
and \[ c_{12} = -(c_{11}V_{12}V_{22}^{-1}) \]

The equation for estimating covariate adjusted differences in mean response between treatment groups takes the following form: \( \beta = (X'V_d^{-1}X)^{-1}X'V_d^{-1}d \)

For the case with bivariate response and 1 covariate, the covariate adjusted design matrix would take the following form:
\[ \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = X = \begin{bmatrix} I \\ 0 \end{bmatrix} \]

\[ \text{Var}(\beta) = X'V_d^{-1}X = \begin{bmatrix} I & 0 \\ c_{11} & c_{12} \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} = c_{11} \]

\[ \beta = (c_{11})^{-1}X'V_d^{-1}d \]
\[ = (c_{11})^{-1}[I \ 0] \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix} \begin{bmatrix} y_{1\text{diff}} - y_{12} \\ y_{2\text{diff}} - y_{22} \\ x_{\text{diff}} - x_{2} \end{bmatrix} \]
\[ = (c_{11})^{-1}[c_{11} \ c_{12}] \begin{bmatrix} y_{1\text{diff}} \\ y_{2\text{diff}} \\ x_{\text{diff}} \end{bmatrix} \]

\[ = [I \ (c_{11})^{-1}(c_{12})] \begin{bmatrix} y_{1\text{diff}} \\ y_{2\text{diff}} \\ x_{\text{diff}} \end{bmatrix}, \]

\[ (c_{11})^{-1}(c_{12}) = (c_{11}^{-1})(c_{12}) - (V_{12}V_{22}^{-1}) = -(V_{12}V_{22}^{-1}) \]
\[ = \begin{bmatrix} 1 & 0 & -\text{Cov}(XY_1)V(X)^{-1} \\ 0 & 1 & -\text{Cov}(XY_2)V(X)^{-1} \end{bmatrix} \begin{bmatrix} y_{1\text{diff}} \\ y_{2\text{diff}} \\ x_{\text{diff}} \end{bmatrix} \]

\[ \beta = \begin{bmatrix} y_{1\text{diff}} - \text{Cov}(Y_1X)V(X)^{-1} \\ y_{2\text{diff}} - \text{Cov}(Y_2X)V(X)^{-1} \end{bmatrix} \]

Because the covariate adjusted treatment parameter estimate for the first response does not involve any information from the second response, and the covariate adjusted estimate for the second response does not involve the first response, we conclude that
results for the univariate response are equivalent to those for the individual elements in the joint multivariate response.

**A.12.2 Proving That Treatment Effect Estimation is Equivalent With Mean Response Parameterization and Difference in Mean Parameterization**

Define two components, \( F = \begin{pmatrix} \bar{y}_1 \\ \bar{x}_1 \end{pmatrix} \), \( G = \begin{pmatrix} \bar{y}_2 - \bar{y}_1 \\ \bar{x}_2 - \bar{x}_1 \end{pmatrix} \) where response \( y \) and covariates \( x \), can be either univariate or multivariate.

A corresponding design matrix can be specified,

\[
\begin{bmatrix}
  I & 0 & | & 0 \\
  0 & I & | & 0 \\
  - & - & | & - \\
  0 & 0 & | & I \\
  0 & 0 & | & 0
\end{bmatrix} = X
\]

which allows for the mean response and mean covariates in sample 1 (\( F \)) to be estimated without any restrictions. Also, the estimation of differences in means (\( G \)) can take several forms. In this particular design matrix, difference in response between treatment groups is estimated with the constraint that the difference in covariates equals zero.

\[ f = \begin{bmatrix} F \\ G \end{bmatrix} \], with design matrix \[ \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} \], and covariance matrix \[ \begin{bmatrix} V_F & V_{FG} \\ V_{FG} & V_G \end{bmatrix} \]

This design matrix has a component \( D \), which corresponds to the \( G \) portion of the response vector. This component can either take the form of an identity matrix in which case there is no covariate adjustment, or it can restrict differences in mean covariates \( (\bar{x}_2 - \bar{x}_1) \) to be equal to zero. We will leave this component of the design in its generic form so that results will be generalizable to both situations.

\[ E\{f\} = X\beta, \text{ where } \beta = \begin{bmatrix} \hat{\beta}_F \\ \hat{\alpha} \end{bmatrix}, \text{ and } \beta = (X'V^{-1}X)^{-1}X'V^{-1}f \]

We want to show that the lower half of the beta parameters, \( \hat{\beta}_G \), are estimated without any influence from the \( F \) portion of the response vector. Our goal is to find out that \( \hat{\beta}_G = (D'V_G^{-1}D)^{-1}D'V_G^{-1}G \) is the same as \( \beta \).
Estimation of parameters for the entire response vector \( f \), takes the following form:

\[
\hat{\beta} = \left( I 0 \right) \left[ V_F V_{FG}^{-1} \right]^{-1} \left( I 0 \right) \left( V_F V_{FG}^{-1} \right)^{-1} \left( F^T \right)
\]

We now want to define a square matrix, \( A \), that when pre and post multiplying this matrix with the covariance matrix for \((F, G)\) the result is null matrices in the off-diagonal position, indicating independence between the two transformed components of \( F \) and \( G \).

In other words: \( A^* \left[ V_F V_{FG} \right] A' = \left[ V_L 0 \right] \left[ 0 V_G \right] \)

We choose \( A = \left[ I - V_{FG}V_{G}^{-1} \right] \), to satisfy this condition.

We know that \( E(f) = X\beta \) so \( E(Af) = AX\beta \) and \( V(Af) = A*Var(f)*A' \)

We will now call \( X_A \) our new design matrix = \( AX \)

\[
X_2 = \left[ I - V_{FG}V_{G}^{-1} \right] \left[ I 0 \right] = \left[ I - V_{FG}V_{G}^{-1} D \right] = \left[ I K \right]
\]

where \( K = - V_{FG}V_{G}^{-1} D \)

\[
Af = \left[ I - V_{FG}V_{G}^{-1} \right] \left[ F \right] = \left[ F - V_{FG}V_{G}^{-1} G \right] = \left[ L \right] = g
\]

so the resulting transformed mean vector \( g = Af \) is composed of a new top component which we will call \( L \), and the same unchanged component \( G \), the difference in means between treatment groups.

\[
Var(Af) = A*Var(f)*A' = \left[ I - V_{FG}V_{G}^{-1} \right] \left[ V_F V_{FG} \right] \left[ I \right] = \left[ I - V_{FG}V_{FG}^{-1} \right] \left[ I \right]
\]

\[
Var(Af) = \left[ V_F - V_{FG}V_{G}^{-1}V_{FG} \right] \left[ 0 \right] = \left[ V_L 0 \right] = V_g
\]

We now use the transformed response vector \( \left[ L \right] \), new design matrix \( X_A \), \( \left[ I K \right] \), and the corresponding covariance matrix for \( L \) and \( G \), \( \left[ V_L 0 \right] \), to estimate parameters using the following equation:

\[
\hat{\beta} = (X_A^T V_g^{-1} X_A)^{-1} X_A^T V_g^{-1} g
\]
Using the specific corresponding matrices, it follows:

\[
(X'_g V^{-1} g X_A)^{-1} = \left\{ \begin{bmatrix} I & 0 \\ K' & D' \end{bmatrix} V^{-1}_L \begin{bmatrix} 0 & 0 \\ 0 & V^{-1}_G L \end{bmatrix} \begin{bmatrix} I & K \\ 0 & D \end{bmatrix} \right\}^{-1} = \left\{ \begin{bmatrix} V^{-1}_L & 0 \\ K' V^{-1}_L & D' V^{-1}_G \end{bmatrix} \begin{bmatrix} I & K \\ 0 & D \end{bmatrix} \right\}^{-1}
\]

\[
= \begin{bmatrix} V^{-1}_L & V^{-1}_L K \\ K' V^{-1}_L & K' V^{-1}_L (K + D' V^{-1}_G D) \end{bmatrix}
\]

We need to calculate the inverse of the above matrix:

We know by definition of inverse that:

\[
\begin{bmatrix} C_{11} & C_{12} \\ C_{12}' & C_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}
\]

\[
C_{22} = (V_{22} - V_{12} V^{-1}_{11} V_{12})^{-1}
\]

\[
C_{12}' = -(V_{22} - V_{12} V^{-1}_{11} V_{12})^{-1} V_{12} V^{-1}_{11}
\]

using this information, we obtain the inverse,

\[
= \begin{bmatrix} * & -K(D' V^{-1}_G D) \cr -(D' V^{-1}_G D) K' & (K' V^{-1}_L (K + D' V^{-1}_G D - K' V^{-1}_L V_{11} V^{-1}_L K)) \end{bmatrix}
\]

we place a * in the upper lefthand corner because this is related to \(V(L)\) which is not used in the covariate adjusted estimation of \(G\). For this, we only need the components for covariances between \(L\) and \(G\) and the covariance matrix.

\[
(X'_g V^{-1} g X_A)^{-1} X'_g V^{-1} g =
\]

\[
= \begin{bmatrix} * & K(D' V^{-1}_G D) \cr (D' V^{-1}_G D) K & (K' V^{-1}_L (K + D' V^{-1}_G D - K' V^{-1}_L V_{11} V^{-1}_L K)) \end{bmatrix} \begin{bmatrix} L \\ G \end{bmatrix}
\]

\[
= \begin{bmatrix} * & -K(D' V^{-1}_G D) \cr -(D' V^{-1}_G D) K' & (D' V^{-1}_G D) \end{bmatrix} \begin{bmatrix} V^{-1}_L L \\ K' V^{-1}_L L + D' V^{-1}_G G \end{bmatrix}
\]

\[
= \begin{bmatrix} * \\ -(D' V^{-1}_G D) K' V^{-1}_L L \cr (D' V^{-1}_G D) \end{bmatrix}
\]

\[
= \begin{bmatrix} (D' V^{-1}_G D) \end{bmatrix}
\]

\[
= 199
\]
Again, we use the convention that if the component (response or covariance) is related only to functions of \( L \) and not \( G \), then an asterisk (*) or (**) is shown instead of listing the full expression.

\[ ^{\hat{\beta}}_G = \left( D'V_G^{-1}D \right)^{-1} D'V_G^{-1}G \]

Therefore difference in means parameterization provides the same estimates as cell mean parameterization, regardless of whether there is covariate adjustment or not.
References


Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. Controlled Clinical Trials 1989, 10:161S-175S.


Conover WI, Iman RL. Rank transformations as a bridge between parametric and non-parametric statistics, American Statistician 1981, 35:124-129.


