METHODOLOGIES FOR IDENTIFYING SUBSETS OF THE POPULATION
WHERE TWO TREATMENTS DIFFER

by

Gregory A. Yothers

B.S. Mathematics, The Pennsylvania State University, 1985

M.A. Statistics, The University of Pittsburgh, 1998

Submitted to the Graduate Faculty of
Arts and Sciences in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2003
Three methodologies for the problem of identifying subsets of the population where two treatments differ are proposed here.

Our first methodology is applicable when the response is possibly right censored time-to-event data and we wish to compare two treatments using the stratified log-rank test on the overall sample as well as separately within the levels of a categorical co-factor used to stratify the overall test. We present a method of calculating the exact familywise type I error rate for this group of tests given the type I error for each of the component tests. We present a sensitivity analysis of power for the design of a clinical trial based on varying assumptions about the interaction between treatment and the co-factor and various allocations of the type I error to the hypothesis tests.

In our second methodology, we consider comparing two treatments using a given hypothesis test on the full sample and on all possible subsets, and we separately consider restricting the subsets considered to be those defined based on half-intervals of a covariate. Rather than treating this as a family of hypothesis tests, we instead choose the minimum p-value from the group of hypothesis tests as our test statistic. Simulation is employed to find an approximate critical value for our novel test statistic. We present a method for simulating a critical p-value which allows for hypothesis testing on all subsets of the sample while controlling type I error.

Our third methodology applies when we wish to compare the mean response between two treatments, but the mean response varies depending on the value of a continuous covariate in some nonlinear fashion. Our solution is to fit separate segmented linear models to each treatment to approximate the nonlinear relationship. We develop asymptotic simultaneous confidence bands for the difference of the expected value functions conditional on the covariate from two treatment groups. The technique applies to the case where the changepoints of the segmented linear models have to be estimated or the case of known changepoints. The method is used to find intervals of the covariate where the treatments differ significantly.
PREFACE

This work would not have been possible without the unending love and patience of my immediate family. My wife Robbin has had to bear more than her share of providing financially for our family and handling of routine household matters and she has not received the personal attention from me that she deserves over these last few years. My children Mitchell and Emily have not had as much of my time and attention as they deserve. I want to thank you all for your patience and perseverance; I will try to make it up to you.

I want to thank my extended family; my grandmother Eva, Mom and Dad Yothers, Mom and Dad Houston, Glenn and Lisa and family, Aaron and Jen, Terri and Paul and family, and Jessica for your support over these last years and for humoring me when I was absent from family functions.

I gratefully acknowledge my current employer The National Surgical Adjuvant Breast and Bowel Project (NSABP) as well as my former employer Respironics Incorporated for providing me the flexibility in my employment to pursue an advanced degree, paying the bulk of my tuition, and providing much of the computing resources necessary to accomplish this work.

My committee members Doctors John Bryant, Sam Wieand, and Leon Gleser have been invaluable to me through the process of completing my education and research. I thank them for their thoughtful mentoring and guidance, their outstanding instruction, and their insightful comments for improving this research.

Last but not least, I want to thank my advisor Doctor Allan Sampson for his tireless efforts in prodding me along to finally complete this research, his inspiring instruction, his editing of my work and coaching of my writing skills, and the wealth of knowledge he brought to bear in helping me tackle these topics. I will always be grateful.
TABLE OF CONTENTS

1 Introduction ................................................................. 1
   1.1 General problem statement ........................................... 1
   1.2 Standard solutions ....................................................... 1
   1.3 Controlling the Familywise Error for Multiple Log-Rank Tests ............. 2
   1.4 Inference Guided Data Exploration ................................... 3
   1.5 Asymptotic Simultaneous Confidence Band for the Difference of Segmented Linear Models .................................................. 6
   1.6 Comparison between Simultaneous Confidence Band for the Difference of Segmented Linear Models and Inference Guided Data Exploration by ANCOVA on Half Intervals of a Covariate................................. 6
   1.7 Conclusion ................................................................... 7

2 Controlling the Familywise Error Rate for Multiple Log-Rank Tests .......... 8
   2.1 Problem Statement ......................................................... 8
   2.2 Background ..................................................................... 8
   2.3 Approach to the problem ................................................ 9
   2.4 Allocation of Alpha .......................................................... 10
      2.4.1 Relationship between subgroup-level test statistics and overall statistic for 2 subgroups ......................................................... 10
      2.4.2 Definition of familywise power and FWE......................... 11
      2.4.3 Extension to many subgroups ......................................... 12
      2.4.4 Recursive definition of power function ............................. 13
      2.4.5 Controlling the FWE ..................................................... 13
   2.5 Motivating Example: NSABP Protocol B-29 ..................................... 14
   2.6 Choosing Type I Error Rates or How Best to Allocate Alpha .................. 15
      2.6.1 Relationship between FWE, overall $\alpha$ and subgroup-level $\alpha$ .......... 15
      2.6.2 Loss of power with no interaction between treatment and subgroup ...... 16
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.3</td>
<td>Partial Derivatives</td>
<td>70</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Equation for Simultaneous Confidence Band</td>
<td>71</td>
</tr>
<tr>
<td>4.4</td>
<td>Segmented Linear Models with Known Changepoints</td>
<td>72</td>
</tr>
<tr>
<td>4.5</td>
<td>Estimation and Practical Approach</td>
<td>73</td>
</tr>
<tr>
<td>4.6</td>
<td>Application</td>
<td>75</td>
</tr>
<tr>
<td>4.7</td>
<td>Simulation of Coverage Probabilities</td>
<td>79</td>
</tr>
<tr>
<td>4.8</td>
<td>Discussion</td>
<td>82</td>
</tr>
<tr>
<td>4.8.1</td>
<td>Comparison to Naïve segmented Working-Hotelling band: Unknown change points</td>
<td>82</td>
</tr>
<tr>
<td>4.8.2</td>
<td>Comparison to Naïve segmented Working-Hotelling band: Known change points</td>
<td>86</td>
</tr>
<tr>
<td>4.8.3</td>
<td>Difficulties Presented by Multiple Intervals of Significant Difference</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>Comparison between Simultaneous Confidence Band for the Difference of Segmented Linear Models and Inference Guided Data Exploration by ANCOVA on Half-Intervals of a Covariate</td>
<td>89</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>89</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Brief explanation of Simultaneous Confidence Band for the Difference of Segmented Linear Models</td>
<td>89</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Brief explanation of Inference Guided Data Exploration by ANCOVA on Half-Intervals of a Covariate</td>
<td>90</td>
</tr>
<tr>
<td>5.1.2.1</td>
<td>Stopping Criteria 1, Half-Intervals Minimum p-value</td>
<td>93</td>
</tr>
<tr>
<td>5.1.2.2</td>
<td>Stopping Criteria 2, Half-Intervals First Significant p-value</td>
<td>93</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Discussion of similarities and differences</td>
<td>94</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Simple models to illuminate the relative strengths of the methods</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>Measures of Performance</td>
<td>95</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Estimation of the True Threshold</td>
<td>95</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Coverage by the True Region were the Groups Differ</td>
<td>97</td>
</tr>
<tr>
<td>5.3</td>
<td>Simulations</td>
<td>98</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Simulation Model 1: Segmented lines with different changepoints, otherwise parallel</td>
<td>98</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Simulation Model 2: Treatment segmented, control single line</td>
<td>99</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Simulation Model 3: Step function for treatment, control single line</td>
<td>100</td>
</tr>
</tbody>
</table>
5.3.4 Minimum sample size and critical p-value used for half-intervals method

5.3.5 Simulation results

5.4 Discussion

5.4.1 Coverage by the True Region where the Groups Differ

5.4.2 Estimation of the True Threshold

5.4.3 Relationship between bias, coverage, and MSE

5.4.4 Areas of application for the methods (Exploratory vs. Regulatory)

6 Conclusion

6.1 Controlling the Familywise Error for Multiple Log-Rank Tests

6.2 Inference Guided Data Exploration

6.3 Asymptotic Simultaneous Confidence Band for the Difference of Segmented Linear Models

6.4 Comparison between Simultaneous Confidence Band for the Difference of Segmented Linear Models and Inference Guided Data Exploration by ANCOVA on Half-Intervals of a Covariate

6.5 Areas for future research

6.6 Acknowledgements

Appendix A

Recursive Definition of the Power Function for Multiple Log-rank Tests

Appendix B

Strong Control of FWE for Multiple Log-rank Tests

Appendix C

Equivalence of Protection Schemes for Multiple Log-rank Tests

Appendix D

Asymptotic Normality of our Parameterization of the Segmented Linear Model

Appendix E

Derivation of the Plug-in Estimator of the Information Matrix

Bibliography
**LIST OF TABLES**

Table 1: Possible alpha allocation for 2 subgroups and FWE = 0.05 ....................... 16
Table 2: Range of FWE for protected and unprotected schemes.......................... 27
Table 3: Numeric example of subsets with Z statistic........................................... 35
Table 4: Minimum subset size, two sample Z-test, alpha = 0.05, N = 10,000
iterations per simulation................................................................................. 37
Table 5: Minimum subset size, two-sample t-test, alpha = 0.05, N = 10,000
iterations per simulation................................................................................. 40
Table 6: Minimum subset size, two-sample t-test, upper bound, alpha = 0.05,
N = 10,000 iterations...................................................................................... 42
Table 7: Selected Z-scores from Table 4, Table 5, and Table 6............................ 43
Table 8: Half-intervals of covariate, two-sample t-test, alpha = 0.05, N = 10,000
iterations........................................................................................................ 48
Table 9: Half-intervals of covariate, Wilcoxon test, alpha = 0.05, N = 10,000
iterations....................................................................................................... 51
Table 10: Half-intervals of a covariate, ANCOVA t-test, alpha = 0.05, N = 10,000
iterations....................................................................................................... 53
Table 11: P values of t-test comparing all patient no older than tabulated age ......... 57
Table 12: Empirical Non-Coverage Probabilities with Covariates Uniform on (0, 10). 80
Table 13: Empirical Non-Coverage Probabilities with Covariates Fixed and Equally
Spaced.......................................................................................................... 81
Table 14: Data for method of half-intervals example ........................................... 92
Table 15: Performance of the methods under various simulation models and standard
deviations................................................................................................. 103
LIST OF FIGURES

Figure 1: NSABP Protocol B-29 Schema ................................................................. 15
Figure 2: Possible alpha allocation for 2 subgroups, FWE = 0.05, and \( \alpha_1 = \alpha_2 \) .... 16
Figure 3: Familywise power and power of overall stratified test, \( RR_1 = RR_2 = 0.75 \) .... 17
Figure 4: Familywise power and power of overall stratified test, \( RR_1 = 0.59 \) and \( RR_2 = 0.81 \) ................................................................. 18
Figure 5: Familywise power for various pairs of reduction in event rates ............... 19
Figure 6: Familywise power for various pairs of reduction in event rates, unequal allocation to subgroups (1:3) ................................................................. 20
Figure 7: Familywise Power for various pairs of reduction in event rates, no overall treatment effect ................................................................. 21
Figure 8: Expected familywise power for various allocations of events to the control arms of the 2 subgroups ................................................................. 22
Figure 9: FWE for protected and unprotected schemes ........................................ 25
Figure 10: 2-segment linear model fit to AC group .............................................. 77
Figure 11: 2-segment linear model fit to AC Tax group ........................................ 77
Figure 12: Difference in mean response of the 2 models with 95% simultaneous confidence band ................................................................. 78
Figure 13: 2-Segment Linear Fit to Treatment Group 1 Data .............................. 84
Figure 14: 2-Segment Linear Fit to Treatment Group 2 Data .............................. 84
Figure 15: Simultaneous Confidence Interval for the Difference (Group 2 minus Group 1) ................................................................. 85
Figure 16: Comparison between Naïve Working Hotelling and Our Simultaneous Confidence Bands ................................................................. 85
Figure 17: Comparison between Naïve Working Hotelling and Our Simultaneous Confidence Bands ................................................................. 87
Figure 18: Method of half-intervals, critical p-value = 0.0169, minimum sample size = 2, panels 1-5 ................................................................. 92
Figure 19: E(Y|X) under simulation models 1, 2, and 3 ....................................... 94
Figure 20: Example for SCB, identified region is covered by true region .......... 97
Figure 21: Mean response vs. covariate under simulation model 1 .................................... 99
Figure 22: Mean response vs. covariate under simulation model 2 .................................. 100
Figure 23: Mean response vs. covariate under simulation model 3 .................................. 100
Figure 24: Coverage by the true interval under simulation model 1 .................................. 104
Figure 25: MSE of threshold estimator under simulation model 1 .................................... 104
Figure 26: Coverage by the true interval under simulation model 2 .................................. 105
Figure 27: MSE of threshold estimator under simulation model 2 .................................... 105
Figure 28: Coverage by the true interval under simulation model 3 .................................. 106
Figure 29: MSE of threshold estimator under simulation model 3 .................................... 106
1 Introduction

1.1 General problem statement

Usual techniques for comparing two treatments assume that the differential response in patients receiving treatment versus control, which we refer to as the treatment effect, is constant over the population under study. In practice, this assumption is at best only approximately correct, and often there are important fluctuations in the effect of treatment due to known factors or latent factors unknown to the researchers. If we compare two treatments under the assumption that the treatment effect is constant over the population under study when in fact the effect of treatment varies due to factors, known or unknown, we are in effect testing whether the “average” treatment effect is different from zero. When such a comparison leads to a significant result, the results are generalized to the entire population and, as a result, if there are subsets of the population where the effect of treatment is negligible or detrimental we may be treating some patients unnecessarily or actually causing harm. Conversely, when such a comparison leads to a conclusion of accepting the null hypothesis of no treatment effect, we may deprive certain subsets of the population from access to a beneficial treatment.

The primary aim of this research is to leverage known factors which may be related to the size of the treatment effect in hopes of identifying subsets of the population where treatment is or is not beneficial. A portion of this research deals with examining all subsets of the population. A finding of significant difference between the treatments when all subsets are considered may imply there are factors unknown to the researcher which are related to the size of the treatment effect.

1.2 Standard solutions

It is common practice to attempt to model factors which are related (or possibly related) to the size of the response in order to reduce the variance of the parameter estimates related to the treatment comparison and provide for more powerful inference. Typically, some attempt is made to test for interaction between treatment and the other factors in the model. These interaction tests can identify situations where the treatment effect is not
constant over the population under study. However, studies are rarely planned to have adequate power for such tests (Pocock et al. (2002)).

When a factor is identified which interacts with treatment, the approach to analysis must be modified to address this. For categorical factors, this generally means that separate treatment comparisons are performed within each level of the factor and the issue of multiple comparisons arises. For continuous factors, the issues can be more complex and only a few standard approaches have been proposed. For example, the parallel slopes analysis of covariance model is often employed when the response is linear with respect to a covariate. Standard practice is to test the parallel slopes assumption prior to proceeding with the analysis. Violation of the parallel slopes assumption implies a treatment by covariate interaction. A solution for the case where the slopes are not parallel was proposed by Potthoff (1964) who constructs a simultaneous confidence band for the difference in mean response between the treatments with respect to the covariate. The treatments are found to be significantly different on regions of the covariate where zero difference is not contained in the simultaneous confidence band.

1.3 Controlling the Familywise Error for Multiple Log-Rank Tests

Our first methodology is described in Chapter 2 and concerns comparing two treatments when the response is possibly censored time-to-event data. More specifically, the problem is the design of an analysis plan for such data when there is some concern that the effect of treatment will be different on several levels of a cofactor. Our solution is to perform stratified log-rank tests within each level of the cofactor along with an overall log-rank test stratified over the levels of the cofactor as well as the other stratification factors. All of the tests are performed at reduced levels of type I error so that the type I error for the family of hypothesis tests is controlled to be level $\alpha$.

Generalized methods of dealing with the multiple comparison problem such as the Bonferroni procedure could be applied to this problem, but the conservative nature of such an approach is unappealing. Other more specialized multiple comparison procedures such as Tukey’s HSD, Fisher’s LSD, Dunnet’s MCP, Holm’s, or Sidak’s are applicable to the setting of ANOVA where one is interested in comparing multiple
treatments (Miller (1981) and Hsu (1996)). Since our aim is to test the same hypothesis for the overall sample as well as within each stratum, these methods do not apply.

Our major contribution is a method of calculating the exact familywise error rate (FWE) for this family of hypothesis tests given the type I error of each component test. Our method can be used to find a combination of type I error rates for the individual hypothesis tests which controls the familywise error rate at level $\alpha$. We present a sensitivity analysis based on the design of National Surgical Adjuvant Breast and Bowel Project (NSABP) Protocol B-29 (NSABP (2000)) to show how various allocations of the type I error to the constituent tests of hypothesis affect power under various alternatives. We find this approach to be efficient and effective when the sign of the treatment effect is the same for each subgroup and there may or may not be small to moderate differences in the magnitude of the effect between subgroups.

We show that ad hoc methods for dealing with the inflation of the type I error such as using a preliminary interaction test to decide whether to perform the overall test or the individual subgroup tests are largely ineffective. While such “protection” schemes do a better job of controlling the type I error than the “unprotected” approach of performing all tests at level $\alpha$, the approach of using a preliminary interaction test still results in roughly a doubling of the FWE.

1.4 Inference Guided Data Exploration

In Chapter 3, we consider comparing two treatments using a given hypothesis test on the full sample and on every subset where each treatment has at least $n_{\min}$ observations. Rather than treating this as a family of hypothesis tests, we instead choose the minimum p-value from the set of hypothesis tests as our test statistic. Simulation is employed to find an approximate critical value for our novel test statistic. Much as the Scheffé procedure provides a critical value for the test statistic which allows for comparing all contrasts in the analysis of variance, we derive a critical p-value which, while controlling the type I error, allows for hypothesis testing on every subset of the original sample where each treatment has at least $n_{\min}$ observations. This technique may prove useful in providing a rule of thumb for judging significant differences when an investigator has
“gone fishing” in his data, i.e., performed a large number of statistical tests of hypothesis on subsets of the sample.

There are problems similar to the all subsets problem in the field of spatial statistics. Worsley (1992) addressed the problem of finding regions of the brain which were active while the subjects performed some task. The volume of the brain is divided into hyper-rectangles called voxels. Brain activation was measured via Positron Emission Tomography (PET) for a number of subjects and a $t$-statistic is computed for the change in activation at each voxel. A typical experiment may include $10^5$ voxels, so the problem of determining which of the voxels experienced a significant change is non-trivial. He uses the concept of a Gaussian field to develop critical values for the $t$-statistics such that, under the null hypothesis of no change in activation, the probability that the $t$-statistic from at least one of the voxels will exceed the critical value is alpha. The difficulties with implementing an approach such as this on the current problem would involve finding an analogue for the “distance” between non overlapping subsets and dealing with the problem of overlapping subsets.

In our work, we consider two common tests of hypothesis, the two sample $Z$-test and two sample $t$-test, and we provide tables of the critical p-value for various sample sizes for the treatment and control groups and various values of the minimum sample size in each treatment group. The exact simulation solution is found to be computationally intractable for the two sample $t$-test for moderate to large sample sizes. This is not the case for the $Z$-test. A special property of the $Z$-test is exploited to find the critical p-value for this test even for large sample sizes. More specifically, if we consider all pairs of sub-samples of size $n_i$ from treatment and $n_j$ from control, then the most extreme $Z$-score for this combination of sub-sample sizes comes from the test comparing the largest $n_i$ order statistics from treatment with the smallest $n_j$ order statistics from control or from the test comparing the smallest $n_i$ order statistics from treatment with the largest $n_j$ order statistics from control. We refer to these two combinations as tests comparing maximal order statistics to minimal order statistics. The approach of only considering subsets where maximal order statistics are compared to minimal order statistics can also be applied to the setting of the $t$-test, and, while it provides an estimate of the critical p-value for the
t-test which is biased upwards, we show that this approximation is superior to using the critical p-value from the Z-test as the approximation.

We also consider a restriction on the subsets to be considered where the subsets are defined based on half-intervals of a covariate. For example, subsets are formed by including all observations except the $k$ largest/smallest order statistics of the covariate from the full sample with the treatments pooled so that, for example, $k = 0$ corresponds to the full sample comparison. Treatment comparisons are made for the subsets corresponding to $k = 0, 1, 2, \ldots$ until one of the treatments has less than $n_{\text{min}}$ observations in the subset. This restriction on the subsets considered leads to much less extreme critical p-values as opposed to considering all subsets.

In a related problem, Koziol and Wu (1996) consider the problem of determining a threshold value for baseline hemoglobin below which patients are less likely to require blood transfusion when treated with r-HuEPO, a recombinant form of human erythropoietin. Their method is applicable to the case where the response is binary and the probability of success is monotonically related to the value of the covariate.

For the approach based on half-intervals of a covariate, we provide tables of the critical p-value for three common tests of hypothesis, the two sample $t$-test, the normal approximation to the Wilcoxon rank sum test, and the analysis of covariance model (ANCOVA). We find that the critical p-value depends on the joint distribution of the response and the covariate for the $t$ and Wilcoxon tests, but for the ANCOVA model, where the dependence between the response and covariate is properly modeled, the critical p-value does not depend on the joint distribution.

In order to illustrate the method based on half-intervals of a covariate, we use tumor response data from NSABP Protocol B-27 (Mamounas (1997)). We identify a half-interval of the covariate (age) where the reduction in tumor size is greater in women receiving Taxotere than in women of the same age not receiving Taxotere.
1.5  Asymptotic Simultaneous Confidence Band for the Difference of Segmented Linear Models

Our final methodology is presented in Chapter 4. Here we are concerned with the comparison of two treatments when the expected value of the response for each treatment differentially depends on the value of a covariate. In particular, we are interested in situations where a linear model is inadequate to describe this relationship. We propose the use of segmented linear models as a flexible approach to describing such functional relationships. The approach that we take is related to that of Potthoff (1964), in that we construct an asymptotic simultaneous confidence band about the difference in mean response of two segmented linear models. In order to accomplish this, we extend results of Cox and Ma (1995) who give a general approach for constructing an asymptotic simultaneous confidence band for a single nonlinear regression function.

Simulations are performed to ascertain the true coverage of our simultaneous confidence bands for the difference of two segmented linear regressions. When the data are sufficient to identify the underlying model, we find that the coverage of our asymptotic simultaneous confidence band is slightly conservative for large sample sizes, i.e., the proportion of simulations where the simultaneous confidence band does not cover the true difference in mean response is less than \( \alpha \). When the covariates are drawn at random, so that there is positive probability of observing no data on one of the segments of the segmented linear model, the method still performs well for moderate to large sample sizes, however the small sample coverage is poor. The method is applied to breast cancer tumor size data from NSABP Protocol B-27 (Mamounas 1997) where we find an interval of the covariate (age) where the treatments are significantly different.

1.6  Comparison between Simultaneous Confidence Band for the Difference of Segmented Linear Models and Inference Guided Data Exploration by ANCOVA on Half Intervals of a Covariate

Chapter 5 compares the methods developed in Chapter 4 and Chapter 3. Each of these methods describes a region of the covariate where the treatments differ when a significant difference is found. We consider several simulation models where the treatments have the same mean response for covariate values up to a threshold value and different mean
responses for covariate values greater than the threshold. The methods are compared based on two performance measures, the first being the probability that the identified region (with respect to the covariate) where the method claims the treatments differ is covered by the true region where the treatments differ. The second performance measure is the mean squared error (MSE) of the threshold estimate relative to the true threshold, where we define the threshold estimate to be the smallest covariate value in the identified region of significant difference. In light of the results from these performance measures we discuss how one might choose between the methods based on the aims of the application.

Our simultaneous confidence band method performed best for coverage of the identified region of significant difference by the true region where the treatments differ. As a result, we recommend this method when the aim of the research is regulatory. The method of inference guided data exploration by ANCOVA on half-intervals performed best for the MSE of the estimate of the threshold value of the covariate defining the region where the treatments differ. Thus, we recommend this method when the aim of the research is exploratory.

1.7 Conclusion

In Chapter 6 we conclude with a brief review of Chapters 2 through 5, reflections on our major contributions and conclusions, and directions for future research.
2 Controlling the Familywise Error Rate
for Multiple Log-Rank Tests

2.1 Problem Statement

Our concern lies with the problem of designing a clinical trial for possibly right censored time-to-event data where we wish to compare two treatments with an overall stratified test along with individual treatment comparisons within subgroups of the data where the grouping factor(s) is one or more of the stratification factors. We wish to perform such a family of hypothesis tests while controlling the familywise type I error. This approach is appealing when the effect of treatment may not be the same in each of the subgroups (i.e., treatment by subgroup interaction). Multiple tests of hypotheses will inflate the type I error for a given clinical trial unless steps are taken to control for multiplicity. Common methods of controlling for multiple comparisons lead to a loss of power for the individual hypotheses so that subgroup analyses are often avoided. However, subgroup analyses often serve a legitimate scientific purpose when performed in a responsible manner.

The stratified log-rank test is often used as the primary analysis for time-to-event data. Log-rank or stratified log-rank tests can be used to determine the significance of the treatment effect separately within each subgroup. General methods for dealing with multiple comparisons, such as the Bonferroni procedure, are conservative and can lead to greatly reduced power. Rather than using such conservative tests, we propose for our goal the calculation of the exact familywise error (FWE) for the overall stratified log-rank test along with separate log-rank or stratified log-rank tests within each subgroup. This will allow for greater flexibility in the design of clinical trials and reduce the loss of power for the individual hypotheses.

2.2 Background

Generalized methods of dealing with the multiple comparison problem such as the Bonferroni procedure could be applied to this problem. The Bonferroni procedure is known to be conservative due to the nature of the underlying probability inequality. Our method exploits the fact that the overall stratified log rank statistic is a weighted sum of
the subgroup level statistics and the dependence among the relevant test statistics can be used to obtain an exact procedure which will be more powerful for each of the component hypotheses than would the Bonferroni procedure. Klein and Moeschberger (1997) give formulas for the stratified and ordinary log-rank test statistics. Both of the statistics have an asymptotic normal distribution.

Other more specialized multiple comparison procedures such as Tukey’s HSD, Fisher’s LSD, Dunnet’s MCP, Holm’s, or Sidak’s are applicable to the setting of ANOVA where one is interested in comparing multiple treatments (Miller (1981) and Hsu (1996)). Since our aim is to test the same hypothesis for the overall sample as well as within each subgroup, these methods do not apply.

2.3 Approach to the problem

Our concern is the design and analysis of clinical trials to compare two treatments where, in addition to the overall test, we desire to test the primary hypothesis on several subgroups which partition the sample. We propose a method whereby a pre-specified familywise alpha is allocated among the overall test and the constituent subgroup tests. We find the method to be efficient in terms of familywise power when the treatment effect in each subgroup is in the same direction and the magnitude of the range of treatment effects between subgroups is not too great. The procedure can be used to make the design of a clinical trial robust against the presence of a treatment by subgroup interaction when a significant interaction is not anticipated.

In Section 2.4, we derive expressions for the exact FWE and familywise power and adapt them to numerical solution. In Section 2.5, we present NSABP Protocol B-29 (NSABP (2000)) as a motivating example. Section 2.6 is a sensitivity analysis, loosely based on NSABP Protocol B-29, which illustrates how variation in design assumptions affects familywise power. We consider common approaches to dealing with a potential treatment by subgroup interaction in Section 2.7 and illustrate the deficiencies of such approaches. We close the chapter in Section 2.8 with a discussion of our conclusions.
2.4 Allocation of Alpha

In this section, we propose a multiple testing approach where one performs an overall test for treatment effect along with tests within each subgroup. All tests are carried out at reduced levels of significance so that the familywise level of significance is maintained at a specified rate. All tests are based on the stratified log-rank statistic, although the ordinary log-rank statistic may be used for a subgroup test when the subgroup consists of a single stratum.

2.4.1 Relationship between subgroup-level test statistics and overall statistic for 2 subgroups

In this subsection, we develop notation and describe the relationship between the subgroup level and overall log-rank test statistics for the case of $k = 2$ subgroups. For example, if we have one grouping factor with 2 levels and an additional stratification factor, then the two subgroup level tests will use log-rank statistics stratified for the additional stratification factor and the overall test is stratified both for the grouping factor as well as the additional stratification factor.

Let $L_1$ and $L_2$ be the stratified log-rank statistics from the 1st and 2nd subgroup tests comparing treatment to control, respectively, and let $V_1$ and $V_2$ be the respective variances. When a subgroup consists of a single stratum, the ordinary log-rank statistic is used. Formulas for the stratified and ordinary log-rank test statistic and its variance are given in Klein and Moeschberger (1997, pp. 192-193, 206). Let $L_0 = L_1 + L_2$ represent the overall stratified log-rank statistic comparing treatment to control aggregated over the subgroups and $V_0$ denote its variance. In the usual case, where failure and censoring times are independent between patients and censoring is non-informative, $L_1$ and $L_2$ are independent and $V_0 = V_1 + V_2$.

Now, the standardized log-rank statistics are represented by $Z_i = \frac{L_i}{\sqrt{V_i}}, \quad i = 0, 1, 2$. Since $L_0 = L_1 + L_2$ and $V_0 = V_1 + V_2$, it follows that:

$$Z_0 = Z_1 \sqrt{a} + Z_2 \sqrt{1-a}, \quad \text{where } a = \frac{V_1}{(V_1 + V_2)} \quad 0 < a < 1.$$ (2.4.1)

The ratio $a$ is approximately the proportion of events which occur in the first subgroup.
The standardized log rank statistics have an asymptotic standard normal distribution under the null hypothesis that the hazard rates, as a function of time, are equal up to the longest time on study (Klein and Moeschberger (1997)).

Let \( \alpha_0, \alpha_1, \) and \( \alpha_2 \) represent the nominal levels of significance of the two-sided tests based on \( Z_0, Z_1, \) and \( Z_2 \) and let \( c_0, c_1, \) and \( c_2 \) be the corresponding critical values (i.e., \( c_0 = \Phi^{-1}(1 - \alpha_0 / 2) \) where \( \Phi \) is the CDF of the standard normal distribution).

When the assumption of proportional hazards holds, it is convenient to describe the null and alternative hypotheses in terms of the Relative Risk (RR) or hazard ratio. Often the term “reduction in event rate” is used, this quantity is equal to one minus the RR. Let \( RR_i, i \in \{0, 1, 2\}, \) represent the relative risk in the overall sample, \( 1^{\text{st}} \) subgroup, and \( 2^{\text{nd}} \) subgroup respectively, so that \( RR_i = 1, i \in \{0, 1, 2\}, \) denotes the null hypothesis and \( RR_1 = RR_1^{\text{Alt}} \) and \( RR_2 = RR_2^{\text{Alt}} \) is a specific alternative hypothesis. The overall relative risk is a function of the relative risks in the subgroups, so that, in specifying a specific alternative, it suffices to specify the alternative for the subgroups only. When the alternative hypothesis is true, the standardized test statistic \( Z_i = L_i / \sqrt{V_i} \) is asymptotically distributed as \( N[\ln(RR_i^{\text{Alt}}) / \sqrt{V_i}, 1] \) (Schoenfeld (1981)). The expected value of the standardized log-rank statistic when the proportional hazards assumption does not hold is given by Lakatos (1988).

Let \( \theta_i \) be the expected value of \( Z_i \) under the alternative hypothesis. Then \( W_i = Z_i - \theta_i \) is distributed as standard normal under the alternative hypothesis. When the assumption of proportional hazards holds, \( \theta_i = \ln \left( RR_i^{\text{Alt}} \right) \sqrt{V_i} \). Under the null hypothesis of no difference between treatments, \( \theta_i = 0 \).

### 2.4.2 Definition of familywise power and FWE

We define familywise power to be the probability of detecting at least one significant difference for a member of our family of hypotheses when a specific alternative hypothesis holds. When the null hypothesis is true (\( \theta_i = 0 \) in each subgroup), we refer to the familywise power as the FWE. The power against a specific alternate hypothesis can
be written as:

$$\text{Power}(\alpha_0, \alpha_1, \alpha_2, a, \theta_1, \theta_2) = 1 - \int \int \phi(z_1 - \theta_1) \phi(z_2 - \theta_2) \, dz_1 dz_2,$$

(2.4.2)

where $\phi$ denotes the standard normal density, and the integral is taken over the acceptance region defined by $\{(z_1, z_2) : |z_1| < c_1, |z_2| < c_2, |z_2 - z_1| < c_0\}$.

The integration region can be re-expressed as:

$$\left\{(z_1, z_2) : -c_2 < z_2 < c_2, \max\left\{-c_1, \frac{c_0 - z_2 \sqrt{1-a}}{\sqrt{a}}\right\} < z_1 < \min\left\{c_1, \frac{c_0 - z_2 \sqrt{1-a}}{\sqrt{a}}\right\}\right\};$$

(2.4.3)

so that it follows:

$$\text{Power}(\alpha_0, \alpha_1, \alpha_2, a, \theta_1, \theta_2) =$$

$$1 - \int_{-c_2}^{c_2} \phi(z_2 - \theta_2) \left[ \Phi\left(\max\left\{-c_1, \min\left\{c_1, \frac{c_0 - z_2 \sqrt{1-a}}{\sqrt{a}}\right\}\right\} - \theta_1\right) - \Phi\left(\max\left\{-c_1, \min\left\{c_1, \frac{c_0 - z_2 \sqrt{1-a}}{\sqrt{a}}\right\}\right\} - \theta_1\right) \right] \, dz_2,$$

(2.4.4)

where $\Phi$ is the CDF of the standard normal distribution.

### 2.4.3 Extension to many subgroups

These results generalize to $k$ subgroups as follows:

$$Z_0 = \sum_{i=1}^{k} Z_i \sqrt{a_i}, \quad \text{where } a_i = V_i \sqrt{\sum_{j=1}^{k} V_j}, \quad 0 < a_i < 1, \quad \sum_{j=1}^{k} a_j = 1.$$  

(2.4.5)

The ratio $a_i$ is approximately equal to the proportion of events which occur in the $i$th subgroup.

The familywise power converges to:

$$\text{Power}(\alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k) =$$

$$1 - \prod_{i=1}^{k} \prod_{j=1}^{c_i} \phi(z_i - \theta_j) \left[ \Phi\left(\max\left\{-c_1, \min\left\{c_1, \frac{c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i}}{\sqrt{a_1}}\right\}\right\} - \theta_1\right) - \Phi\left(\max\left\{-c_1, \min\left\{c_1, \frac{c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i}}{\sqrt{a_1}}\right\}\right\} - \theta_1\right) \right] \, dz_2 \cdots dz_k.$$

(2.4.6)
2.4.4 Recursive definition of power function

The multiple integral in equation (2.4.6) can be difficult to evaluate numerically when the number of subgroups goes beyond about 3 or 4. Fortunately there is a recursive representation of the power function that facilitates computation when there are many subgroups.

Given \( \alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k \), define:

\[
\rho_r(z) = \Pr \left( \sum_{i=1}^{r} Z_i \leq z \mid \left| Z_j \right| \leq c_j \right). \tag{2.4.7}
\]

Then,

\[
\rho_1(z) = \Phi \left( \max \left[ -c_1, \min \left( c_1, z / \sqrt{a_1} \right) \right] - \theta_1 \right) - \Phi \left( -c_1 - \theta_1 \right), \tag{2.4.8}
\]

and,

\[
\rho_{r+1}(z) = \int_{-c_{r+1}}^{c_{r+1}} \phi(u - \theta_{r+1}) \rho_r \left( z - u / \sqrt{a_{r+1}} \right) du. \tag{2.4.9}
\]

For \( k \) subgroups,

\[
\text{Power} \left( \alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k \right) = 1 - \left[ \rho_k(c_0) - \rho_k(-c_0) \right]. \tag{2.4.10}
\]

See Appendix A for derivation and proof.

An S-Plus function implementing the recursive method of calculating power is available from the author.

2.4.5 Controlling the FWE

Given \( a_1, \ldots, a_k \), where \( a_i \) is the proportion of expected events in subgroup \( i \), controlling the familywise error to be level \( \alpha \) is a problem of choosing \( \alpha_0, \ldots, \alpha_k \) such that the power (2.4.10) is equal to \( \alpha \) when all null hypotheses are true (\( \theta_1 = \cdots = \theta_k = 0 \)). The solutions form a subspace of dimension \( k \), so that there is no unique solution. Trial and error can be used to find a solution reasonably quickly, and the process may be expedited by constraining the nominal type I errors for the within subgroup tests to all be equal. This approach provides strong control of the FWE in the sense that for any arrangement of
true and false null hypotheses, the FWE will be no more than $\alpha$. See Appendix B for a discussion of strong control.

2.5 Motivating Example: NSABP Protocol B-29

The primary aim of NSABP Protocol B-29 (NSABP (2000)) was to determine whether the addition of Octreotide to Tamoxifen alone or to Tamoxifen in combination with chemotherapy (AC) prolongs disease-free survival in women with axillary node-negative, estrogen-receptor (ER) positive primary invasive breast cancer. The protocol schema is shown in Figure 1.

An unusual feature of this clinical trial is that the patient has to decide along with her physician whether or not to use chemotherapy. The trial was designed this way because, at the time this trial was initiated, the standard of care for these patients was in a state of flux. Previously, the standard of care among node negative ER positive breast cancer patients had been Tamoxifen alone. The release of the results of NSABP Protocol B-20 (Fisher et al. (1997)) at about the same time as the initiation of Protocol B-29 caused a shift in the standard of care to chemotherapy.

The trial is designed to have 80% power to detect a 25% decrease in event rate (i.e., $RR = 0.75$) using a 0.05 level two-sided overall stratified log-rank test. The trial will accrue 3000 patients over 5 years, and the final analysis is scheduled to follow the 400th event, which is estimated to require an additional 3 years of follow-up.

Physicians involved in the design of the trial thought the effect of Octreotide would be unlikely to materially interact with chemotherapy status. However, in planning the trial it was felt to be important to provide for individual tests for the effect of Octreotide in the presence of chemotherapy as well as in its absence. It was considered unacceptable to treat these subgroup analyses as post-hoc, or exploratory, so it was necessary to design an analysis plan that controlled the familywise error rate. Among the many factors considered in making this decision, a primary consideration is that, if the trial shows positive results, the pharmaceutical company providing the test drug would likely use the trial as a basis for a new treatment indication from regulatory bodies.
The stratification factors are chemotherapy (yes, no), age (<50, 50+), and pathologic tumor size (0-2, 2.1-4, and 4.1+ cm). The subsets of interest are those defined by the grouping factor chemotherapy (yes, no) and the within subset tests for a treatment effect due to Octreotide are stratified for age and pathologic tumor size.

2.6 Choosing Type I Error Rates or How Best to Allocate Alpha

2.6.1 Relationship between FWE, overall $\alpha$, and subgroup-level $\alpha$

The question arises as to how the type I error should be divided between the overall and the subgroup-specific tests, or rather, how much alpha should be spent on the subgroup-specific tests. For $k = 2$ subgroups, Table 1 and Figure 2 show a variety of combinations of the nominal size of the overall test ($\alpha_0$) and the nominal size of the within subgroup tests ($\alpha_1$ & $\alpha_2$). For simplicity, we only consider the case where $\alpha_1 = \alpha_2$. The possibilities form a continuum between ($\alpha_0$, $\alpha_1 = \alpha_2$) = (0.05, 0) (no subgroup specific tests) to (0, 0.0253) (no overall test). Given $\alpha_{exp}$, $\alpha_0$, and the constraint $\alpha_1 = \alpha_2$, the common value of $\alpha_1$ & $\alpha_2$ is a function of $a$ (the proportion of events in the first subgroup), however, as may be seen in Figure 2, the effect of varying $a$ is weak.
Table 1: Possible alpha allocation for 2 subgroups and FWE = 0.05

<table>
<thead>
<tr>
<th></th>
<th>(\alpha_0)</th>
<th>(\alpha_1 = \alpha_2)</th>
<th>(\alpha_0)</th>
<th>(\alpha_1 = \alpha_2)</th>
<th>(\alpha_0)</th>
<th>(\alpha_1 = \alpha_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a = 0.50)</td>
<td>0.050</td>
<td>0.0000</td>
<td>0.050</td>
<td>0.0000</td>
<td>0.050</td>
<td>0.0000</td>
</tr>
<tr>
<td></td>
<td>0.045</td>
<td>0.0060</td>
<td>0.045</td>
<td>0.0060</td>
<td>0.045</td>
<td>0.0057</td>
</tr>
<tr>
<td></td>
<td>0.040</td>
<td>0.0099</td>
<td>0.040</td>
<td>0.0104</td>
<td>0.040</td>
<td>0.0108</td>
</tr>
<tr>
<td></td>
<td>0.030</td>
<td>0.0161</td>
<td>0.030</td>
<td>0.0168</td>
<td>0.030</td>
<td>0.0183</td>
</tr>
<tr>
<td></td>
<td>0.020</td>
<td>0.0207</td>
<td>0.020</td>
<td>0.0214</td>
<td>0.020</td>
<td>0.0229</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.0240</td>
<td>0.010</td>
<td>0.0244</td>
<td>0.010</td>
<td>0.0250</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.0253</td>
<td>0.000</td>
<td>0.0253</td>
<td>0.000</td>
<td>0.0253</td>
</tr>
</tbody>
</table>

Figure 2: Possible alpha allocation for 2 subgroups, FWE = 0.05, and \(\alpha_1 = \alpha_2\)

2.6.2 Loss of power with no interaction between treatment and subgroup

Here we consider the penalty in terms of power when some of our alpha is allocated to subgroup specific tests when the treatment effect is the same in all subgroups. Figure 3 shows the case where there is an overall 25% reduction in event rate, no interaction between treatment and subgroup \((RR_1 = RR_2 = 0.75)\), and the expected number of events
is equal in each of the $k = 2$ subgroups ($a = 0.5$). The overall stratified log-rank test at the 0.05 level has power 0.82. The alpha allocation of $\alpha_0 = 0.04$ and $\alpha_1 = \alpha_2 = 0.0099$ yields a power of 0.79 for the overall test and a familywise power of 0.80! Thus, spending 20% of total alpha (setting $\alpha_0 = 0.04$) on subgroup tests leads to a very small loss of power even in the case of no interaction. Setting $\alpha_0$ less than about 0.03, on the other hand, leads to a rather substantial loss in power.

We should point out that the small loss in power for the case of no treatment by subgroup interaction is not the only sacrifice when electing to spend some of our alpha on subgroup tests. When all of our alpha is allocated to the overall test, then rejecting the null hypothesis results in a conclusion generalizable to the entire population represented by the full sample. When some of our alpha is spent on subgroup tests, the trial may have a “successful” outcome (one of the null hypotheses is rejected), but the result may only be generalizable to the population represented by a subset. Thus the scope of application of a successful result may be reduced.

![Figure 3: Familywise power and power of overall stratified test, $RR_1 = RR_2 = 0.75$](image)
2.6.3 Power with interaction between treatment and subgroup

Next we consider how power is affected in the presence of treatment-subgroup interaction when some of our alpha is allocated to subgroup specific tests. Figure 4 shows the case of a 41% reduction in event rate in one subgroup and a 9% reduction in the other subgroup. We again assume that there are 400 total events and that the expected number of events on the control arm of each subgroup is equal. When the subgroups are pooled, we have an overall 25% reduction in event rate. The overall stratified log-rank test at the $\alpha_0 = 0.05$ level has power 0.829. Reducing $\alpha_0$ to 0.04 reduces the power of the overall test to 0.804, but increases the familywise power to 0.906! Spending any more than 20% of alpha on subgroup tests does not materially increase the familywise power even in the presence of this very substantial interaction.

![Figure 4: Familywise power and power of overall stratified test, $RR_1 = 0.59$ and $RR_2 = 0.81$](image)

In the previous subsection, we pointed out that allocating some of our alpha to subset tests has the potential of limiting the scope of application for a successful outcome when there is no interaction between treatment and subgroup. The gain in familywise power illustrated by Figure 4 is due to the increased probability of detecting the large treatment
difference in the 1st subgroup. Even if the scope of application is limited, it may be better to get something (a finding restricted to a subgroup) than nothing (a negative trial).

Now we examine how varying the magnitude of interaction affects the familywise power. Figure 5 shows a variety of pairs of reduction in event rates in the two subgroups such that when the subgroups are pooled we have a 25% reduction in event rate. We again assume that there are 400 total events and that the expected number of events on the control arm of each subgroup is equal. The overall stratified log-rank test at the $\alpha_0 = 0.05$ level has power in the range 0.820-0.844. Reducing $\alpha_0$ to 0.04 dramatically increases the familywise power in the presence of interaction. Spending any more than 20% of alpha on subgroup tests does not materially increase the familywise power even in the presence of substantial interaction. Note that for small interaction (the pair (35, 15)), the power is maximized near $\alpha_0 = 0.04$.

Figure 5: Familywise power for various pairs of reduction in event rates
2.6.4 Power with varying sample sizes between subgroups

Until now we have only considered what happens when the numbers of events (patients) in the two subgroups are roughly equal. We next consider the case where most of the patients are assigned to the subgroup with the smaller treatment effect. Figure 6 shows a variety of pairs of reduction in event rates for comparison with Figure 5. We again assume that there are 400 total events, but now the numbers of events on the control arms of the two subgroups are not equal. The expected number of events on the control arm of subgroup 2 is three times the expected number of events on the control arm of subgroup 1. When the subgroups are pooled, the reduction in event rate varies from 25% for the pair (25, 25) to 12.5% for the pair (50, 0). We see that the familywise power suffers when most of the events (patients) are in the subgroup with a small reduction in event rate. Note that a choice of size of overall test equal to 0.04 would give near maximum power given any of the states of nature explored in Figure 6.

Figure 6: Familywise power for various pairs of reduction in event rates, unequal allocation to subgroups (1:3)
2.6.5 Power with no overall effect but varying interaction between treatment and subgroup

Now consider the case of no overall treatment effect but varying degrees of treatment-subgroup interaction. Figure 7 shows a variety of pairs of reduction in event rates in the two subgroups such that when the subgroups are pooled we have no reduction in event rate. We again assume that there are 400 total events and that the number of events on the control arm of each subgroup is equal. Reducing $\alpha_0$ (the size of the overall test) increases the familywise power in the presence of qualitative interaction. When there is no overall effect and the treatment is beneficial in one subgroup and detrimental in the other, the alpha allocation approach described here is not very powerful. If, during the design phase, one anticipates substantial interaction, the trial should be powered for a test of treatment by subgroup interaction as the primary aim. Gail and Simon (1985) suggest a test for qualitative interaction which may be appropriate in this situation.

Figure 7: Familywise Power for various pairs of reduction in event rates, no overall treatment effect
2.6.6 Power when a prior probability distribution is placed on the difference in effect size between subgroups

In this subsection, we consider the average familywise power for various allocations of events to the control arms of the two subgroups when we place a prior probability distribution on the difference in percent reduction in event rate between the two subgroups. In each case, there is a 25% reduction in event rate when the subgroups are pooled. The prior is normal with mean zero and standard deviation such that there is a 5% probability of qualitative interaction (treatment is beneficial in one subgroup and detrimental in the other). Figure 8 shows the expected power given the prior distribution.

![Figure 8: Expected familywise power for various allocations of events to the control arms of the 2 subgroups](image)

We see that the alpha allocation procedure is most effective when the numbers of events on the control arms of the two subgroups are not too far out of balance. Note that taking the size of the overall test between 0.04 and 0.045 gives near optimal power when the balance of events is no worse than 1:3. If we relax the requirement that the size of the tests in the two subgroups be equal and instead spend more alpha on the subgroup with
the most events, we can improve the familywise power when the numbers of events in the
subgroups are out of balance.

2.7 Common Approaches to Controlling FWE

Various ad hoc methods for controlling the FWE are used in practice. For instance, one
method to reduce the inflation of the type I error for our problem might be to first
perform a test for a treatment by subgroup interaction and then follow with the overall
test if the test for interaction is negative or the individual subgroup specific tests if there
is interaction. Next, we will explore the consequences with respect to FWE of several
methods of dealing with the problem of multiple tests. A list of potential protection
schemes follows.

(1) Unprotected Subgroup Tests – Perform the overall stratified test at level $\alpha$;
follow-up with subgroup-specific $\alpha$ level tests.

(2) Protected Subgroup Tests – Perform the overall stratified test at level $\alpha$;
follow-up with subgroup-specific $\alpha$ level tests only if treatment-by-
subgroup interaction is significant.

(3) Protected Subgroup Tests – Test for treatment-by-subgroup interaction at
level $\alpha$. If interaction is significant, test for treatment effect individually
in each subgroup at level $\alpha$. If interaction is not significant, test for
overall treatment effect at level $\alpha$.

(4) Protected Subgroup Tests – Perform the overall stratified test at level $\alpha$;
follow-up with subgroup-specific level $\alpha$ tests only if the overall treatment
effect is significant.

Here the FWE is the probability of finding a significant difference between treatments on
either the overall stratified test or any of the subgroup-specific tests given that no
difference exists in any subgroup and, by implication, no overall treatment effect or
treatment by subgroup interaction.
Unprotected tests (1) perform an $\alpha$ level overall stratified test along with individual $\alpha$ level tests for each subgroup. An alternative to the unprotected test would be to “protect” the individual subgroup tests by allowing them only if a test for treatment by subgroup interaction is significant (2). A second alternative is to perform the interaction test first; if it is not significant, the overall stratified test is performed; otherwise, if the interaction is significant, the treatment effect is tested individually in each subgroup (3). These two alternatives for protecting the subgroup specific tests ((2) and (3)) are actually quite similar in operating characteristics, since if both interaction and main effect are significant, it is almost certain that at least one subgroup level test will also be significant. This is true with probability one in the case of $k = 2$ subgroups (see Appendix C). Thus, for the remainder of this section (2.7), we will consider only protection scheme (3).

The final protection method (4), where the subgroup specific tests are done only when the main effect test is significant, is much stronger in the sense that it maintains the FWE at level $\alpha$. However, this method is clearly not very powerful in the presence of substantial interaction, and will not be considered further.

The FWE for unprotected tests (protection scheme (1)) and tests protected by a preliminary interaction test (protection scheme (3)) for the case of $k = 2$ subgroups is shown in Figure 9. The FWE for the unprotected tests ranges from 0.0975 to 0.115, while the FWE for the protected tests ranges from 0.080 to 0.0975. The error rates vary as a function of $a = V_1/(V_1 + V_2)$, where $V_1$ and $V_2$ are the variances of the respective subgroup specific test statistics. This ratio is approximately equal to the proportion of events that occur in subgroup 1.
For any testing scheme where the subgroup-specific tests are independent and the overall test is independent of the interaction test (true in typical survival analyses), the probability of a type I error for the unprotected (1) and protected (3) schemes can be expressed as follows. We assume the null hypothesis is no treatment effect in any subgroup implying no effect in the overall test and no interaction. The two type I errors are:

\[
P_{\text{UNPROT}} \{ \text{Type I error} \} = P \{ \text{Type I error in overall test} \cup \text{Type I error in any subgroup test} \} \\
= P \{ \text{Type I error in any subgroup test} \} \\
+ P \{ \text{Type I error in overall test} \cap \text{No type I error in any subgroup test} \} \\
= 1 - (1 - \alpha)^k \\
+ P \{ \text{Type I error in overall test} \cap \text{No type I error in any subgroup test} \} \\
\geq 1 - (1 - \alpha)^k,
\]
and,

$$P_{\text{prot}} \{\text{Type I error}\} = \left\{ \begin{array}{l} \text{Type I error in interaction test} \\ \text{Type I error in any subgroup test} \end{array} \right\} + P \left\{ \begin{array}{l} \text{Type I error in interaction test} \\ \text{Type I error in overall test} \end{array} \right\}$$

$$\alpha - P \left\{ \begin{array}{l} \text{Type I error in interaction test} \\ \text{No type I error in any subgroup test} \end{array} \right\} + (1 - \alpha) \alpha \leq 2\alpha - \alpha^2 \leq 1 - (1 - \alpha)^2. \quad (2.7.2)$$

In Table 2, we illustrate the performance of these protection schemes with respect to FWE when they are extended to more than two subgroups. As in the case of $k = 2$ subgroups, the FWE varies as a function of $a = (a_1, \ldots, a_k)$, where $a_i = V_i / \sum_{j=1}^k V_j$. The ratio $a_i$ is approximately equal to the proportion of events that occur in subgroup $i$.

Equation (2.7.1) shows that the type I error of the unprotected test is bounded below by $1 - (1 - \alpha)^k$. This value is the limit of the type I error as any component $a_i$ of $a$ goes to 1. The type I error rate of the unprotected test appears to achieve its maximum at $a = (1/k, \ldots, 1/k)$. This conjecture is supported numerically, but has yet to be proven.

Equation (2.7.2) shows that the type I error of the protected test is bounded above by $1 - (1 - \alpha)^2$. This value is the limit of the type I error as any component $a_i$ of $a$ goes to 1. The type I error rate of the protected test appears to achieve its minimum at $a = (1/k, \ldots, 1/k)$. For the unprotected tests, the FWE goes up dramatically with the number of subgroups. The unprotected FWE is on the order of $k\alpha$, where $k$ is the number of subgroups. The upper bound for the FWE of the protected tests is $1 - (1 - \alpha)^2 (\approx 2\alpha$ for small alpha), about twice the nominal rate, and does not depend on the number of subgroups. Obviously the FWE goes up dramatically with the number of subgroups in the unprotected testing scheme, but even with the so called “protected” testing scheme the rate nearly doubles the nominal $\alpha$. One could use protection scheme (3) with all tests
performed at level $\alpha/2$ to maintain control of the type I error, but clearly this would result in a loss of power.

Table 2: Range of FWE for protected and unprotected schemes

<table>
<thead>
<tr>
<th>Number of Subgroups</th>
<th>Range of FWE$^a$</th>
<th>Unprotected Tests (1)</th>
<th>Protected Tests (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.098-.115</td>
<td>.080-.098</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.143-.161</td>
<td>.090-.098</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.185-.204</td>
<td>.094-.098</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.226-.245</td>
<td>.095-.098</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.302-.319</td>
<td>.096-.098</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>.401-.416</td>
<td>.097-.098</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ All tests performed at $\alpha = .05$.

2.8 Conclusion

The alpha allocation approach described here is very effective when the treatment effect is in the same direction in each subgroup and there may or may not be small to moderate differences in the size of the effect between subgroups. The method is sensitive to the balance of events (patients) between the subgroups. When the sizes of the subgroup level tests are equal, the approach seems to be quite effective when the balance is no worse than about 3 to 1. We suggest spending more alpha on the subgroup with the most patients (events) when the number of patients is out of balance between subgroups. Spending between 10 and 20 percent of alpha on the subgroup level tests (i.e., setting $\alpha_0$ equal to 0.045 to 0.04) would seem to be a prudent choice for $k = 2$ subgroups and the range of circumstances explored in this chapter when substantial interaction is thought to be unlikely a priori.

When there is no overall treatment effect but there may be offsetting effects between the subgroups, the alpha allocation approach is not very powerful. Designing the trial for a test for interaction would be much more effective in this situation. If one were to use the multiple testing procedure in this situation, most of the alpha should be spent on the within subgroup tests.
In the design of NSABP protocol B-29 (NSABP (2000)), we expected little or no interaction and nearly equal accrual to the two subgroups (chemotherapy and no chemotherapy). Given our design assumptions in B-29, we spent about 10% of alpha on subgroup level tests and set the size of the subgroup level tests equal. If we had anticipated unequal accrual to the subgroups or significant interaction, we likely would have altered our choices. Our choice of alpha allocation \((a_0 \approx 0.045, a_1 = a_2 = 0.006)\) proved to preserve power in the presence of mild perturbations of design assumptions.

Ad hoc methods for dealing with the inflation of the type I error such as using a preliminary interaction test to decide whether to perform the overall test or the individual subgroup tests are largely ineffective. While such “protection” schemes do a better job of controlling the type I error than the “unprotected” approach of performing all tests at level \(\alpha\), the approach of using a preliminary interaction test still results in roughly a doubling of the FWE.

The tools described in this chapter can be adapted to the design of other potential trials. Given prior beliefs regarding the likelihood of significant treatment-subgroup interaction, balance of accrual to the subgroup levels, and other factors, one can explore the sensitivity of power to design assumptions and parameters much as we have in Section 2.6 of this chapter.
3 Inference Guided Data Exploration

3.1 Problem Statement

Researchers are often tempted to look at subset analyses of their data; this is especially true when the primary hypothesis turns out to be non-significant. Formal control of the type I error becomes difficult or impossible when the same hypothesis is tested over and over on subsets of the original data. Statisticians sometimes refer to this scenario as a “fishing expedition”. Given the ad hoc nature of the hypotheses tested in this situation, formal inference is not really possible, but some sort of “rule of thumb” for deciding whether the most extreme p-value observed during a fishing expedition is likely to occur purely by chance would be quite useful in practice. A similar problem arises when the number of formal tests of hypothesis is small, but the subsets which are tested are selected after inspecting plots or summary measures of the data or the raw data itself.

Our technique is motivated by the following problem. Consider an experiment where a researcher wishes to compare the response of a group of individuals receiving some treatment to the response of another group of individuals who receive control. In addition to the overall test of hypothesis on these groups, the researcher is also interested in looking at the treatment effect in possibly overlapping subsets of the original sample. We first consider the problem where the subsets of interest are all subsets where a treatment comparison is possible and later consider restrictions on the subsets of interest.

The similarities to the first topic (multiple log-rank tests) are obvious. This work is a generalization where we do not restrict ourselves to using the log-rank statistics nor do we restrict the subsets under consideration to be disjoint. A major difference is that we do not consider the group of hypothesis tests as a family of tests, instead, we choose as our test statistic the minimum p-value from the group of tests. The aim of the research is to determine a critical value to control the type I error for our unusual test statistic.

Some notation is necessary to express the ideas that follow more precisely. Let $S_t$ and $S_c$ be the sets of individuals who received treatment and control respectively. Let $i$ and $j$ index the unique, unordered, and non-empty subsets of $S_t$ and $S_c$. If we have samples of
size \( n_t \) from treatment and \( n_c \) from control, then since \( \sum_{i=1}^{n} \binom{n}{i} = 2^n - 1 \), there are \( 2^n - 1 \) possible subsets of individuals who receive treatment and \( 2^n - 1 \) possible subsets of individuals who receive control.

The number of possible pairs of subsets is:

\[
\binom{2^n - 1}{2^n - 1}
\]  

(3.1.1)

Finally, let \( S_{ti} \) and \( S_{cj} \) denote the \( i \)th and \( j \)th subsets from treatment and control. We define \( P_{ij} \) to be the p-value from a test of hypothesis on the responses associated with the pair of subsets \( S_{ti} \) and \( S_{cj} \). Depending on the test statistic used, the cardinality of \( S_{ti} \) and \( S_{cj} \) must typically be at least 2 for \( P_{ij} \) to be defined. The notation \( |S_{ti}| \) means the cardinality of the set \( S_{ti} \). The most extreme p-value (\( P^* \)) obtained by testing all possible pairs of subsets would be:

\[
P^* = \inf_{\{i,j \mid |S_{ti}| \geq 2 \text{ and } |S_{cj}| \geq 2\}} P_{ij}.
\]  

(3.1.2)

Given the large number \((2^n - 1 - n_t)(2^n - 1 - n_c)\) of possible pairs of subsets which could be compared in evaluating (3.1.2), one may wish to consider some restriction criteria. If the universe of subsets which are considered for hypothesis testing is restricted, then under the null hypothesis, we would expect to see less extreme p-values than if all possible subsets are considered. A natural first candidate for a restriction criteria would be to increase the cardinality of the sets under consideration in equation (3.1.2) to a number larger than two.

Another appealing restriction criterion would be to choose subsets of the original sample based on half-intervals of a covariate. For instance, if the covariate is age, we could first consider the pair of subsets formed by including all subjects in treatment and control at most as old as the oldest person in the sample. Next, exclude the oldest person and compare the resulting subsets. Continue in this fashion until the cardinality criterion is violated in one of the treatment groups. This type of restriction criteria on the subsets considered has the advantage of allowing a generalization of a significant result in a subset of the sample to the corresponding subset of the population represented by the sample. Our first method, which considers all subsets, may find a significant result when
comparing 2 subsets, but there may be no way of defining the corresponding subsets of the population related to the sample.

To amplify on the method based on half-intervals, consider a situation where response is correlated with a covariate. For instance, wound healing rate may be correlated with initial wound area. It may be that the difference in mean response between treatment and control also depends on the covariate, i.e., larger wounds heal faster when treated, but treatment has little or no effect on smaller wounds. How does one choose inclusion/exclusion criteria in order to perform a study in this situation? Choosing eligibility criteria which are too liberal with respect to the covariate may include patients where the treatment effect is negligible making it difficult to prove the treatment superior to control at the end of the study. Likewise, choosing eligibility criteria which are too conservative with respect to the covariate may lead to denying therapy to patients who may benefit when the treatment is finally approved. A method is needed whereby the trial is designed to include all patients who could potentially benefit from treatment, but the analysis is designed to determine the largest subset with respect to the covariate where treatment is significantly better than control. Such a method would allow the target population for the treatment to be determined by the data rather than being chosen somewhat arbitrarily prior to the trial.

3.2 Background

The similarities of this problem to the multiple log-rank test procedure considered in Chapter 2 are such that the same arguments apply with respect to existing multiple comparison procedures. The Bonferroni procedure could be applied to this problem, but would prove to be extremely conservative and virtually unusable due to the large number of comparisons considered. More specialized multiple comparison procedures tend to be targeted at the problem of comparing multiple treatments on the same endpoint or the problem of making the same treatment comparison on multiple endpoints and hence do not apply here (Miller (1981) and Hsu (1996)).

Fleiss (1986) discusses a method of determining the region of a covariate where treatment is superior to control when the regression lines are not parallel; his presentation
uses Potthoff’s (1964) work in this regard. An expression for the difference in mean response between the treatments when the treatments are modeled independently via linear regression is derived, and a confidence interval for the difference of regressions is computed. The interval of the x-axis where the confidence interval lies totally above (below) zero defines the region where treatment is superior (inferior) to control. This approach can be applied to the type of problem we consider in Section 3.5 if one is willing to assume that the mean response is linear with respect to the covariate in both the treatment and control groups.

There are interesting parallels to this section’s problem in the field of spatial statistics. Worsley (1992) addressed the problem of finding regions of the brain which were active while the subjects performed some task. Cerebral blood flow was measured at baseline via Positron Emission Tomography (PET) and repeated while the subjects where performing the task of interest. The volume of the brain is divided into hyper-rectangles called voxels. The process is repeated for a number of subjects and a \( t \)-statistic is computed for the change in blood flow at each voxel. A typical experiment may include \( 10^5 \) voxels, so the problem of determining which of the voxels experienced a significant change is non-trivial. He uses the concept of a Gaussian field to develop critical values for the \( t \)-statistics such that, under the null hypothesis of no change in blood flow, the probability that the \( t \)-statistic from at least one of the voxels will exceed the critical value is alpha. The critical value depends on the volume of the brain and the number of voxels. The difficulties with implementing an approach such as this on the current problem would involve finding an analogue for the “distance” between non overlapping subsets and dealing with the problem of overlapping subsets.

Koziol and Wu (1996) consider the problem of determining a threshold value for baseline hemoglobin below which patients are less likely to require blood transfusion when treated with r-HuEPO, a recombinant form of human erythropoietin. Their method is applicable to the case where the response is binary and the probability of success is monotonically related to the value of the covariate. They test the null hypothesis that the probability of success (no transfusion) is the same increasing function with respect to the covariate
(baseline hemoglobin) versus the alternative that the probability of success differs for values of the covariate less than a threshold value.

3.3 Approach to the problem

We begin by developing the method of inference guided data exploration restricted by minimum subset sample size in Section 3.4. The method is then applied to find critical p-values for the two sample Z-test, the two sample \( t \)-test, and approximate critical values for the two sample \( t \)-test. In Section 3.5 we restrict our interest to subsets formed by considering only those observations where a covariate falls in a half interval (i.e., the value of the covariate exceeds some threshold). This additional restriction leads to less extreme critical p-values. The modified method is then applied to the two sample \( t \)-test, the two sample Wilcoxon test, and the analysis of covariance model. In Section 3.6, we apply the method developed in Section 3.5 to tumor response data from NSABP Protocol B-27 (Mamounas (1997)), and we conclude in Section 3.7 with a discussion of the results.

3.4 Inference Guided Data Exploration Restricted by Minimum Subset Sample Size

As noted in the problem statement and equation (3.1.2), there is typically some minimum subset sample size required for the test statistic to be defined. If, instead of restricting the subsets of interest to those where the test statistic is defined, we rather restrict our interest to subsets of at least size \( n_{\text{min}} \), then we would expect to see less extreme minimum p-values at the end of the procedure. Define \( P^*_{n_{\text{min}}} \), the most extreme p-value observed after testing all pairs of subsets with cardinality at least \( n_{\text{min}} \), as follows:

\[
P^*_{n_{\text{min}}} = \inf_{i,j} \{ i \neq j \mid \|S_i\| \geq n_{\text{min}} \text{ and } \|S_j\| \geq n_{\text{min}} \} P_{ij},
\]

(3.4.1)

The number of pairs of subsets is:

\[
N_{\text{pairs}} = \left[ \sum_{r=n_{\text{min}}}^{n} \binom{n}{r} \right] \times \left[ \sum_{r=n_{\text{min}}}^{n} \binom{n}{r} \right].
\]

(3.4.2)
A critical value for $P_{n_{\text{min}}}^*$, which controls the type I error at level $\alpha$, can be estimated via simulation as follows:

1) Sample responses for $S_t$ and $S_c$ from a null distribution (to be specified).
2) Calculate $P_{n_{\text{min}}}^*$ for the sample.
3) Repeat steps 1) and 2) $N$ times.
4) Estimate the critical value as the $(N \alpha)^{th}$ order statistic from the $N$ values of $P_{n_{\text{min}}}^*$.

This procedure, while computationally intensive, is feasible for small sample sizes on current PC’s. The estimated critical value depends on the test statistic selected, the sample sizes for treatment and control ($n_t$ and $n_c$), and the minimum subset size ($n_{\text{min}}$). The precision of the estimate depends on the number of simulations ($N$), and convergence is roughly proportional to $1/\sqrt{N}$.

3.4.1 Two sided Z-test assuming known variance

Here we apply the method of inference guided data exploration restricted by minimum subset sample size using the two sample Z-test as our statistic. The Z-test has the property that the mapping from the test statistic to the p-value does not depend on the underlying sample sizes used in the calculation. Therefore, one can directly determine a critical Z-score for our method. The $t$ statistic, for example, has a mapping to the p-value which depends on the degrees of freedom (a function of the constituent sample sizes). Hence, calculation of a critical $t$-score is not possible and one must work with the p-value.

The assumption of a known variance allows a shortcut for our method. The Z-statistic in this situation depends only on the difference in means and the size of each sub-sample. If we consider all pairs of sub-samples of size $n_i$ from treatment and $n_j$ from control, then the most extreme Z-score for this combination of sub-sample sizes will come from the pair having the largest difference in means. The largest difference in means will come from the test comparing the largest $n_i$ order statistics from treatment with the smallest $n_j$ order statistics from control or from the test comparing the smallest $n_i$ order statistics from treatment with the largest $n_j$ order statistics from control. When the sub-sample
sizes are not specified, we refer to these two combinations as tests comparing maximal order statistics to minimal order statistics. Because of this, we need only consider the $2(n_t - n_{\text{min}} + 1)(n_c - n_{\text{min}} + 1) - 1$ comparisons where maximal order statistics are compared to minimal order statistics as opposed to the large number found in (3.4.2).

For example if we have a treatment sample ($n_t = 3$) with responses $\{-1, 0, 1\}$, and control sample ($n_c = 3$) with responses $\{-2, 0, 1\}$, then the following pairs of sub-samples of at least size 2 ($n_{\text{min}} = 2$) can be compared (Table 3). The $Z$-statistic is computed assuming a known variance of 1.

**Table 3: Numeric example of subsets with $Z$ statistic**

| Sub-sample Sizes | Treatment | Control | |Z statistic| Comment |
|------------------|-----------|---------|-----------------|-----------|
| (3, 3)           | $\{-1, 0, 1\}$ | $\{-2, 0, 1\}$ | 0.408          | a, b      |
| (3, 2)           | $\{-1, 0, 1\}$ | $\{-2, 0\}$   | 1.095          | a, b      |
| (3, 2)           | $\{-1, 0, 1\}$ | $\{0, 1\}$    | 0.548          | a         |
| (3, 2)           | $\{-1, 0, 1\}$ | $\{-2, 1\}$   | 0.548          |           |
| (2, 3)           | $\{-1, 0\}$   | $\{-2, 0, 1\}$| 0.183          | a         |
| (2, 2)           | $\{-1, 0\}$   | $\{-2, 0\}$   | 0.5            |           |
| (2, 2)           | $\{-1, 0\}$   | $\{0, 1\}$    | 1.0            | a         |
| (2, 2)           | $\{-1, 0\}$   | $\{-2, 1\}$   | 0.0            |           |
| (2, 3)           | $\{0, 1\}$    | $\{-2, 0, 1\}$| 0.913          | a, b      |
| (2, 2)           | $\{0, 1\}$    | $\{-2, 0\}$   | 1.5            | a, b, c   |
| (2, 2)           | $\{0, 1\}$    | $\{0, 1\}$    | 0.0            |           |
| (2, 2)           | $\{0, 1\}$    | $\{-2, 1\}$   | 1.0            |           |
| (2, 3)           | $\{-1, 1\}$   | $\{-2, 0, 1\}$| 0.365          |           |
| (2, 2)           | $\{-1, 1\}$   | $\{-2, 0\}$   | 1.0            |           |
| (2, 2)           | $\{-1, 1\}$   | $\{0, 1\}$    | 0.5            |           |
| (2, 2)           | $\{-1, 1\}$   | $\{-2, 1\}$   | 0.5            |           |

- **a** Compares maximal order statistics with minimal order statistics
- **b** Most extreme statistic for treatment sample size and control sample size
- **c** Most extreme statistic overall

Note that for every pair of sub-sample sizes considered, with the exception of the full sample comparison (sizes (3, 3)), there are two entries labeled “a” where maximal order statistics are compared to minimal order statistics. For instance, when $n_t = 3$ and $n_j = 2$, one may have the three lowest order statistics from treatment and the two highest from control ($|Z| = 0.548$) or vice versa ($|Z| = 1.095$). As previously asserted, the most extreme
Z-score for a given pair of sample sizes (labeled “b”) is always one of the entries which compare maximal order statistics with minimal order statistics.

We performed simulations applying the inference guided data exploration restricted by minimum subset sample size method to normal independently and identically distributed samples and using the two-sample two-sided Z-test for the test of hypothesis on various combinations of sample sizes for the two treatment groups and the minimum sample size. The results are presented in Table 4. For these simulations both the treatment and control samples are drawn from the same null distribution $N(0, 1)$. The critical p-value is invariant under location and scale changes so that the tabulated values correspond to any normal distribution. The table lists the relevant parameters ($n_t$, $n_c$, and $n_{\text{min}}$), the critical value for the test statistic to control the type I error at level 0.05 ($|Z|$), and the p-value associated with the critical value ($p(|Z|)$). We performed $N = 10,000$ iterations for each simulation.

The table shows that critical values can be quite extreme when a large number of pairs of subsets are considered. For instance the entry for $n_t = 40$, $n_c = 40$, and $n_{\text{min}} = 20$, would consider (from equation (3.4.2)) \[\sum_{\delta=20}^{40} \binom{40}{\delta} \times 3.83 \times 10^{23}\] comparisons per iteration. Of course, our shortcut for this particular test statistic allows us to calculate the critical values with only $2(n_t - n_{\text{min}} + 1)(n_c - n_{\text{min}} + 1) - 1 = 881$ comparisons per iteration. The critical value for this entry is 6.5656 as compared to 1.96 for a critical value for a single comparison. Note that the critical values for entries where $n_t = n_c = n_{\text{min}}$ (and hence the number of pairs of subsets considered equals 1) cluster around 1.96 as would be expected.

Holding other parameters fixed, the critical p-value tends to decrease (become more extreme) as $n_t$ or $n_c$ increase and increase as $n_{\text{min}}$ increases.
Table 4: Minimum subset size, two sample Z-test, alpha = 0.05, \( N = 10,000 \) iterations per simulation

| \( n_t \) | \( n_c \) | \( n_{\min} \) | \(|Z|\) | \( p(|Z|)\) | \( n_t \) | \( n_c \) | \( n_{\min} \) | \(|Z|\) | \( p(|Z|)\) |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 2 | 2 | 1.9754 | 4.823E-02 | 7 | 4 | 2 | 3.3693 | 7.536E-04 |
| 3 | 2 | 2 | 2.2674 | 2.337E-02 | 7 | 5 | 2 | 3.5228 | 4.271E-04 |
| 3 | 3 | 2 | 2.6111 | 9.025E-03 | 7 | 6 | 2 | 3.6034 | 3.141E-04 |
| 3 | 3 | 3 | 1.9595 | 5.005E-02 | 7 | 7 | 2 | 3.7174 | 2.013E-04 |
| 4 | 2 | 2 | 2.4516 | 1.422E-02 | 7 | 3 | 3 | 2.7770 | 5.487E-03 |
| 4 | 3 | 2 | 2.8055 | 5.023E-03 | 7 | 4 | 3 | 3.1197 | 1.811E-03 |
| 4 | 4 | 2 | 2.9926 | 2.766E-03 | 7 | 5 | 3 | 3.3471 | 8.165E-04 |
| 4 | 3 | 3 | 2.2959 | 2.168E-02 | 7 | 6 | 3 | 3.4957 | 4.728E-04 |
| 4 | 4 | 3 | 2.6501 | 8.048E-03 | 7 | 7 | 3 | 3.6133 | 3.023E-04 |
| 4 | 4 | 4 | 1.9664 | 4.926E-02 | 7 | 4 | 4 | 2.6978 | 6.981E-03 |
| 5 | 2 | 2 | 2.6002 | 9.316E-03 | 7 | 5 | 4 | 3.0367 | 2.392E-03 |
| 5 | 3 | 2 | 2.9464 | 3.215E-03 | 7 | 6 | 4 | 3.2583 | 1.121E-03 |
| 5 | 4 | 2 | 3.1406 | 1.686E-03 | 7 | 7 | 4 | 3.4196 | 6.272E-04 |
| 5 | 5 | 2 | 3.2925 | 9.929E-04 | 7 | 5 | 5 | 2.5424 | 1.101E-02 |
| 5 | 3 | 3 | 2.4993 | 1.244E-02 | 7 | 6 | 5 | 2.8904 | 3.848E-03 |
| 5 | 4 | 3 | 2.8447 | 4.445E-03 | 7 | 7 | 5 | 3.0813 | 2.061E-03 |
| 5 | 5 | 3 | 3.0599 | 2.215E-03 | 7 | 6 | 6 | 2.3030 | 2.128E-02 |
| 5 | 4 | 4 | 2.2993 | 2.149E-02 | 7 | 7 | 6 | 2.6325 | 8.476E-03 |
| 5 | 5 | 4 | 2.6354 | 8.403E-03 | 7 | 7 | 7 | 1.9419 | 5.215E-02 |
| 5 | 5 | 5 | 1.9643 | 4.950E-02 | 8 | 8 | 2 | 3.9164 | 8.989E-05 |
| 6 | 2 | 2 | 2.7175 | 6.577E-03 | 8 | 8 | 3 | 3.8250 | 1.308E-04 |
| 6 | 3 | 2 | 3.0562 | 2.241E-03 | 8 | 8 | 4 | 3.7251 | 1.953E-04 |
| 6 | 4 | 2 | 3.2566 | 1.127E-03 | 8 | 8 | 5 | 3.4624 | 5.355E-04 |
| 6 | 5 | 2 | 3.4013 | 6.707E-04 | 8 | 8 | 6 | 3.0946 | 1.971E-03 |
| 6 | 6 | 2 | 3.5435 | 3.948E-04 | 8 | 8 | 7 | 2.6160 | 8.898E-03 |
| 6 | 3 | 3 | 2.6698 | 7.590E-03 | 8 | 8 | 8 | 1.9830 | 4.737E-02 |
| 6 | 4 | 3 | 2.9950 | 2.745E-03 | 9 | 9 | 2 | 4.0921 | 4.275E-05 |
| 6 | 5 | 3 | 3.2336 | 1.222E-03 | 10 | 10 | 2 | 4.2311 | 2.326E-05 |
| 6 | 6 | 3 | 3.3580 | 7.852E-04 | 11 | 11 | 2 | 4.3919 | 1.123E-05 |
| 6 | 4 | 4 | 2.5595 | 1.048E-02 | 12 | 12 | 2 | 4.5035 | 6.685E-06 |
| 6 | 5 | 4 | 2.8878 | 3.880E-03 | 13 | 13 | 2 | 4.6436 | 3.423E-06 |
| 6 | 6 | 4 | 3.0965 | 1.958E-03 | 14 | 14 | 2 | 4.7962 | 1.617E-06 |
| 6 | 5 | 5 | 2.2913 | 2.195E-02 | 16 | 16 | 14 | 3.0158 | 2.563E-03 |
| 6 | 6 | 5 | 2.6518 | 8.006E-03 | 20 | 20 | 10 | 5.1021 | 3.360E-07 |
| 6 | 6 | 6 | 1.9785 | 4.787E-02 | 30 | 30 | 15 | 5.8718 | 4.312E-07 |
| 7 | 2 | 2 | 2.7961 | 5.172E-03 | 40 | 40 | 20 | 6.5466 | 5.886E-11 |
| 7 | 3 | 2 | 3.1417 | 1.680E-03 | 50 | 50 | 35 | 4.8161 | 1.464E-06 |
To relate these results to the problem where a researcher has been “fishing” for a significant result in his/her data, consider the following hypothetical scenario. A statistician is approached by a researcher after an experiment has been performed and preliminary hypothesis tests suggest a significant result with $p = 0.001$. The statistician asks questions about the nature of the experiment and learns that 40 independent observations were made from each of two different treatment groups and the 2 sample $Z$-test was used since the researcher knows the value of the standard deviation from previous work. The original comparison on the full sample did not yield the expected result, i.e., $p = 0.07$, and the researcher was not able to reject the null hypothesis. After inspecting plots of the data against various covariates, the researcher discovered an interesting relationship with a covariate he/she had not previously considered important. After several tests of hypothesis using different threshold values for the covariate, the researcher comes to the statistician with a significant result on a subset of the data where each treatment has 24 observations. The statistician asks a few more questions and learns that no subgroups with fewer than 20 observations per treatment were considered and the researcher would not have been comfortable with results on a sample smaller than 20 per treatment.

Now clearly, in the preceding scenario, the researcher did not consider all of the $3.83 \times 10^{23}$ subsets where each treatment has at least 20 observations or even a significant fraction of this number. However, even if only a handful of formal comparisons were done, the fact that the researcher scanned plots of the data with respect to multiple covariates before selecting a hypothesis to test should not be ignored. The statistician refers to Table 4 and finds the entry for $n_r = 40$, $n_c = 40$, and $n_{\min} = 20$. He explains to the researcher that a conservative critical value, considering the way the subset was selected prior to the formal comparison, would be $Z = 6.5455$ or $p = 5.886 \times 10^{-11}$. He convinces the researcher to report the results as suggestive but not conclusive and requiring further confirmation.

### 3.4.2 Two sided $t$-test

Next we consider applying our method using the two-sample two-sided $t$-test for the test of hypothesis. The shortcut method of only considering subsets where maximal order
statistics are compared to minimal order statistics does not apply to this situation and so we must compare all pairs of subsets which meet the minimum sample size constraint. The shortcut method does not apply because the $t$-statistic, unlike the $Z$-statistic, estimates the variance based on the treatment and control samples under consideration so that the most extreme statistic for a given pair of sub-sample sizes may not necessarily come from the pair having the largest separation in means. Table 5 shows the results of simulations applying the inference guided data exploration restricted by minimum subset sample size method to normal independent identically distributed samples and using the two-sample two-sided $t$-test. For these simulations both the treatment and control samples are drawn from the same null distribution $N(0, 1)$. The table lists the relevant parameters ($n_t$, $n_c$, and $n_{\text{min}}$), the critical p-value for the test to control the type I error at level 0.05 ($p$), and the $Z$-score associated with the critical value ($Z(p)$). $N = 10,000$ iterations were performed for each simulation.

Due to the large amount of time required to run the simulations for larger sample sizes, this table is not as extensive as Table 4. For example, the entry for $n_t = 16$, $n_c = 16$, $n_{\text{min}} = 14$ required 48 hours on a 1.5 GHz Pentium 4 system.

The critical value must be determined based on the p-value for the $t$-test because the critical value will be used in evaluating subsets of varying sizes and the degrees of freedom for the $t$-statistic is of course based on the underlying sample size. The $Z$-score corresponding to the critical p-value is tabulated merely for the convenience of the reader as oftentimes statisticians are more comfortable understanding changes on this scale rather than very small changes to already small p-values.
Table 5: Minimum subset size, two-sample $t$-test, alpha = 0.05, $N = 10,000$ iterations per simulation

<table>
<thead>
<tr>
<th>$n_t$</th>
<th>$n_c$</th>
<th>$n_{\text{min}}$</th>
<th>$p$</th>
<th>$Z(p)$</th>
<th>$n_t$</th>
<th>$n_c$</th>
<th>$n_{\text{min}}$</th>
<th>$p$</th>
<th>$Z(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4.647E-02</td>
<td>1.9911</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4.701E-02</td>
<td>1.9862</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1.660E-02</td>
<td>2.3954</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2.272E-03</td>
<td>3.0522</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4.629E-03</td>
<td>2.8318</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>5.426E-04</td>
<td>3.4588</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5.322E-02</td>
<td>1.9331</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>2.289E-04</td>
<td>3.6848</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>8.010E-03</td>
<td>2.6517</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>9.885E-05</td>
<td>3.8934</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1.905E-03</td>
<td>3.1047</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>4.736E-05</td>
<td>4.0683</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>8.358E-04</td>
<td>3.3407</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>4.451E-03</td>
<td>2.8443</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1.718E-02</td>
<td>2.3827</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>1.014E-03</td>
<td>3.2865</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4.798E-03</td>
<td>2.8203</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>3.170E-04</td>
<td>3.6010</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5.133E-02</td>
<td>1.9487</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>1.598E-04</td>
<td>3.7753</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3.816E-03</td>
<td>2.8930</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>6.983E-03</td>
<td>2.6976</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>9.524E-04</td>
<td>3.3042</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>1.869E-03</td>
<td>3.1103</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>3.785E-04</td>
<td>3.5547</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>6.872E-04</td>
<td>3.3946</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1.798E-04</td>
<td>3.7458</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>1.671E-02</td>
<td>2.3930</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>7.194E-04</td>
<td>3.3821</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4.558E-03</td>
<td>2.8367</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2.016E-03</td>
<td>3.0879</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4.591E-02</td>
<td>1.9962</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>3</td>
<td>7.194E-04</td>
<td>3.3821</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>1.307E-03</td>
<td>3.2143</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>1.546E-02</td>
<td>2.4215</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>2.871E-04</td>
<td>3.6267</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4.945E-03</td>
<td>2.8106</td>
<td>16</td>
<td>16</td>
<td>14</td>
<td>8.143E-04</td>
<td>3.3479</td>
</tr>
</tbody>
</table>

The p-values (and associated $Z$-scores) are slightly more extreme in Table 5 than the corresponding values in Table 4 as may be expected since the $t$-test is not as powerful as the $Z$-test. For example, the entry corresponding to $n_t = 16$, $n_c = 16$, and $n_{\text{min}} = 14$ has a critical p-value of $8.143 \times 10^{-4}$ in Table 5, but a less extreme critical p-value of $2.563 \times 10^{-3}$ in Table 4. The large amount of computer time required to simulate the critical p-values for the relatively small sample sizes in Table 5 suggests that this approach is impractical if not impossible for moderate to large sample sizes. In the next section (3.4.3), we consider an approximation to the critical p-value for the two sample $t$-test.

Holding other parameters fixed, the critical p-value tends to decrease (become more extreme) as $n_t$ or $n_c$ increase and increase as $n_{\text{min}}$ increases. This result is consistent with Table 4.
3.4.3 Two sided $t$-test, upper bound

What are the consequences of applying the shortcut used for the Z-test to the $t$-test (i.e., only compare the most extreme $n_i$ order statistics from treatment with the most extreme $n_j$ order statistics from control for each pair of sub-sample sizes)? Our simulations show that this approach often misses the most extreme p-value obtained by considering all pairs of subsets where each treatment has at least $n_{\min}$ observations resulting in an estimated critical p-value which is upwardly biased (less extreme). However, if this approach comes close to estimating the most extreme p-value with a small upward bias, it may be good enough for practical purposes. In particular, the question arises as to whether or not this approach yields a better approximation to the critical p-value for the $t$-test than using the critical p-value from the Z-test tabulated in Table 4.

Table 6 shows an upper bound for the critical p-value based on the results of simulations applying the inference guided data exploration restricted by minimum subset sample size method to normal independent identically distributed samples and using the two-sample two-sided $t$-test for the test statistic. For these simulations both the treatment and control samples are drawn from the same null distribution $N(0, 1)$, and we only consider $t$-tests based on subsets where maximal order statistics are compared to minimal order statistics. We call this an upper-bound because the unbiased critical p-value could be smaller (more extreme) than value we list. The table lists the relevant parameters ($n_t$, $n_c$, and $n_{\min}$), the critical p-value for the test to control the type I error at level 0.05 ($p$), and the Z-score associated with the critical value ($Z(p)$). For each simulation, we performed $N = 10,000$ iterations.
Table 6: Minimum subset size, two-sample $t$-test, upper bound, $\alpha = 0.05$, $N = 10,000$ iterations

<table>
<thead>
<tr>
<th>$n_t$</th>
<th>$n_c$</th>
<th>$n_{min}$</th>
<th>$p$</th>
<th>$Z(p)$</th>
<th>$n_t$</th>
<th>$n_c$</th>
<th>$n_{min}$</th>
<th>$p$</th>
<th>$Z(p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4.64E-02</td>
<td>1.9911</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>2.495E-04</td>
<td>3.6627</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1.660E-02</td>
<td>2.3954</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>1.589E-04</td>
<td>3.7768</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>5.848E-03</td>
<td>2.7562</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>7.873E-05</td>
<td>3.9482</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>5.322E-02</td>
<td>1.9331</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>5.330E-05</td>
<td>4.0407</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>8.923E-03</td>
<td>2.6150</td>
<td>7</td>
<td>3</td>
<td>3</td>
<td>3.122E-03</td>
<td>2.9555</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2.807E-03</td>
<td>2.9882</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>7.401E-04</td>
<td>3.3743</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.472E-03</td>
<td>3.1802</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>3.205E-04</td>
<td>3.5982</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1.718E-02</td>
<td>2.3827</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>1.607E-04</td>
<td>3.7739</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5.476E-03</td>
<td>2.7776</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>8.856E-05</td>
<td>3.9200</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>5.087E-03</td>
<td>2.8015</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>1.073E-03</td>
<td>3.2707</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1.574E-03</td>
<td>3.1606</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>4.443E-04</td>
<td>3.5123</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>7.333E-04</td>
<td>3.3768</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>2.154E-04</td>
<td>3.7002</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>4.130E-04</td>
<td>3.5316</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>7.686E-03</td>
<td>2.6656</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>8.992E-03</td>
<td>2.6124</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>1.796E-03</td>
<td>3.1220</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2.589E-03</td>
<td>3.0128</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>7.733E-04</td>
<td>3.3622</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>3</td>
<td>1.024E-03</td>
<td>3.2839</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>1.684E-02</td>
<td>2.3903</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>1.546E-02</td>
<td>2.4215</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>4.469E-03</td>
<td>2.8430</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5.268E-03</td>
<td>2.7902</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>4.965E-02</td>
<td>1.9629</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4.701E-02</td>
<td>1.9862</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2.034E-05</td>
<td>4.2612</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3.488E-03</td>
<td>2.9211</td>
<td>8</td>
<td>8</td>
<td>3</td>
<td>2.857E-05</td>
<td>4.1845</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>1.083E-03</td>
<td>3.2679</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>4.823E-05</td>
<td>4.0640</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>4.727E-04</td>
<td>3.4958</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>1.629E-04</td>
<td>3.7705</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>2</td>
<td>2.280E-04</td>
<td>3.6858</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>6.736E-04</td>
<td>3.4001</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2</td>
<td>1.248E-04</td>
<td>3.8365</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>4.698E-03</td>
<td>2.8271</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>4.863E-03</td>
<td>2.8159</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>4.966E-02</td>
<td>1.9629</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>3</td>
<td>1.358E-03</td>
<td>3.2033</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>8.322E-06</td>
<td>4.4567</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>3</td>
<td>4.872E-04</td>
<td>3.4877</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>3.501E-06</td>
<td>4.6390</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3</td>
<td>2.923E-04</td>
<td>3.6220</td>
<td>11</td>
<td>11</td>
<td>2</td>
<td>1.671E-06</td>
<td>4.7897</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>7.337E-03</td>
<td>2.6812</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>6.249E-07</td>
<td>4.9834</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>2.093E-03</td>
<td>3.0767</td>
<td>13</td>
<td>13</td>
<td>2</td>
<td>3.196E-07</td>
<td>5.1115</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>4</td>
<td>8.794E-04</td>
<td>3.3265</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>1.510E-07</td>
<td>5.2513</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5</td>
<td>1.671E-02</td>
<td>2.3930</td>
<td>16</td>
<td>16</td>
<td>14</td>
<td>8.143E-04</td>
<td>3.3479</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4.704E-03</td>
<td>2.8266</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>1.051E-08</td>
<td>5.7224</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4.591E-02</td>
<td>1.9962</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.304E-11</td>
<td>6.7682</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2.139E-03</td>
<td>3.0703</td>
<td>40</td>
<td>40</td>
<td>20</td>
<td>2.309E-14</td>
<td>7.6321</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
<td>6.157E-04</td>
<td>3.4246</td>
<td>50</td>
<td>50</td>
<td>35</td>
<td>1.393E-08</td>
<td>5.6743</td>
</tr>
</tbody>
</table>
Holding other parameters fixed, the critical p-value tends to decrease (become more extreme) as $n_t$ or $n_c$ increase and increase as $n_{\text{min}}$ increases. This is consistent with Table 4 and Table 5.

To examine the applicability of this upper-bound, we compared the entries of Table 5, where an exhaustive search of all possible pairs of sub-samples at least as large as $n_{\text{min}}$ was considered, to the corresponding entries of Table 6. We see that the critical p-values are often less extreme in Table 6 than those in Table 5 so that these p-values are slightly anti-conservative. For example, consider the selected $Z$-scores displayed in Table 7 corresponding to the critical p-values from Table 4, Table 5, and Table 6. We have elected to work with the $Z$-score rather than the more relevant critical p-value because differences on this scale seem easier to comprehend. The difference between the $Z$-scores corresponding to the critical p-value in Table 6 minus the value from Table 5 (bias) is negative or zero meaning the estimated critical p-values listed in Table 6 are less extreme than the unbiased estimates in Table 5. The zero difference for the entry $n_t = 16$, $n_c = 16$, and $n_{\text{min}} = 14$ indicates that for the particular simulations carried out, the 5th percentile of the minimum p-value was the same to the number of decimal places recorded for both the shortcut method and the exhaustive search method. We consider this a coincidence and suspect additional simulations would show a small negative bias.

Table 7: Selected $Z$-scores from Table 4, Table 5, and Table 6

<table>
<thead>
<tr>
<th>$n_t$</th>
<th>$n_c$</th>
<th>$n_{\text{min}}$</th>
<th>$Z(p)$</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Table 4 ($Z$)</td>
<td>Table 5 ($t$)</td>
<td>Table 6 ($t$-upper)</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2.9926</td>
<td>3.3407</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3.2925</td>
<td>3.7458</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2.6354</td>
<td>2.8106</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>2</td>
<td>3.5435</td>
<td>4.0683</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3</td>
<td>3.3580</td>
<td>3.7753</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3.0965</td>
<td>3.3946</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>2</td>
<td>3.1417</td>
<td>3.6267</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>14</td>
<td>3.0158</td>
<td>3.3479</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>10</td>
<td>N/A</td>
<td>5.7224</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>20</td>
<td>6.5466</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Table 7 also allows comparison of the potential methods of approximating the critical p-value for the \( t \)-test, the \( Z \)-test from Table 4 and the upper bound for the \( t \)-test from Table 6. The difference between the \( Z \)-score corresponding to the critical p-value in Table 4 minus the value from Table 6 (additional bias) is always negative meaning the estimated critical p-values listed in Table 4 are less extreme than the upwardly biased estimates in Table 6. The approach of using the shortcut approximation for the critical p-value using the \( t \)-test, while being slightly biased upwards, would seem to be a better approximation than using the critical value from the \( Z \)-test. Also, since the direct approach employed for Table 5 is computationally intractable for even moderate sample sizes, we would recommend this approximation for the critical p-value of the \( t \)-test when using the method of inference guided data exploration restricted by minimum subset sample size.

3.5 Inference Guided Data Exploration Restricted by Half-Intervals of a Covariate

A further restriction on the subsets to be considered may be of interest in practice when a covariate is relevant. Consider the case where we observe a covariate along with the response in our experiment. One may have reason to believe that the size of the treatment effect could be related to the value of the covariate. If the eligibility criteria for study entry for the covariate is so wide as to allow inclusion of individuals for which the effect of treatment may be negligible, then the multiple subsets procedure we propose may be useful to find the range of the covariate where there is an effect.

Let \( k \) index the order statistics of the observed covariate \( X, X_{(1)}, \ldots, X_{(n + n^c)} \), where here we have combined the covariates of the treatment and control groups. Now, for convenience and without loss of generality, let the sets of individuals receiving treatment and control be denoted by:

\[
S_t = \{1, \ldots, n_t\}, \quad S_c = \{n_t + 1, \ldots, n_t + n_c\}.
\]  
(3.5.1)

Define the subsets of individuals where the covariate is no larger than the \( k^{th} \) order statistic of the covariates as:

\[
S^{(k)}_t = \left\{ i : i \leq n_t, x_i \leq X_{(k)} \right\}, \quad S^{(k)}_c = \left\{ i : n_t + 1 \leq i \leq n_t + n_c, x_i \leq X_{(k)} \right\},
\]  
(3.5.2)

where \( x_i \) is the value of the covariate associated with the \( i^{th} \) individual.
We define $P_k$ to be the p-value from a test of hypothesis on the responses associated with the pair of subsets $S_t^{(k)}$ and $S_c^{(k)}$. Define $P_{\text{max},X}^{*}$, the most extreme p-value observed after testing all pairs of subsets defined by half-intervals of the covariate and with cardinality at least $n_{\text{min}}$, as follows:

$$P_{\text{max},X}^{*} = \inf_{k : |c^{(i)}| \geq n_{\text{min}} \text{ and } |c^{(j)}| \geq n_{\text{min}}} P_k.$$ \hfill (3.5.3)

The number of pairs to be compared is at most:

$$\text{Max # of pairs} = (n_t + n_c - 2n_{\text{min}} + 2).$$ \hfill (3.5.4)

The simulation procedure from Section 3.4 can be used to estimate a critical value for $P_{\text{max},X}^{*}$ which controls the type I error at level $\alpha$, simply substitute $P_{\text{max},X}^{*}$ for $P_{\text{max}}^{*}$ as follows.

1) Sample responses and covariates for $S_t$ and $S_c$ from a null distribution (to be specified).
2) Calculate $P_{\text{max},X}^{*}$ for the sample.
3) Repeat steps 1) and 2) $N$ times.
4) Estimate the critical value as the $(N \alpha)^{th}$ order statistic from the $N$ values of $P_{\text{max},X}^{*}$.

The estimated critical value depends on the test statistic selected, the sample sizes for treatment and control ($n_t$ and $n_c$), the minimum subset size ($n_{\text{min}}$), and the dependence between the response and the covariate. For the case we consider, where the response and covariate are distributed as bivariate normal, the correlation ($\rho$) describes this dependence. The precision of the estimate depends on the number of simulations ($N$).

### 3.5.1 Two sided t-test

Here we apply the method of inference guided data exploration restricted by half-intervals of a covariate using the two sample t-test as our statistic. The data consist of a response and a covariate which are jointly distributed as bivariate normal with correlation $\rho$. If $Y$ and $X$ are the response and covariate, respectively, then both the treatment and
control samples are drawn from the following null distribution:

\[
\begin{pmatrix} Y \\ X \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right].
\] (3.5.5)

The critical p-value is invariant under location and scale changes so that the tabulated values correspond to any bivariate normal distribution with correlation \( \rho \). Independent identically distributed samples are generated and the two-sample two-sided \( t \)-test is used for the test statistic. Table 8 lists the relevant parameters (\( \rho, n_t, n_c, \) and \( n_{\text{min}} \)), the critical p-value for the test to control the type I error at level 0.05 (\( p \)), and the \( Z \)-score associated with the critical value (\( Z(p) \)). Each simulation consisted of 10,000 iterations.

In contrast to Table 4, Table 5, and Table 6, the comparable entries for critical p-values are not nearly as extreme in Table 8. There are two reasons for this. First of all, we have restricted our attention to pairs of subsets formed by half-intervals of a covariate which leads to fewer pairs of subsets to be compared per iteration. For instance the entry for \( n_t = 10, n_c = 10, \) and \( n_{\text{min}} = 2 \), would consider \( 2(n_t - n_{\text{min}} + 1)(n_c - n_{\text{min}} + 1) - 1 = 161 \) comparisons per iteration in Table 4 and Table 6 (using the shortcut for the \( Z \) and \( t \) statistics respectively), while in this table we would only consider at most \( (n_t + n_c - 2n_{\text{min}} + 2) = 18 \) comparisons per iteration. Secondly, Table 4 and Table 6 only consider pairs of subsets where maximal order statistics are compared to minimal order statistics, while in this table the subsets considered are more “typical”. By typical, we mean the following:

Consider the case of \( \rho = 0 \), this implies the response and covariate are independent under the assumption of a bivariate normal distribution. Now, selecting a subset of responses based on the (independent) covariate is like selecting a subset at random. In the case of \( \rho = 1 \), we would compare minimal order statistics to minimal order statistics in (3.5.2) or maximal order statistics to maximal order statistics (when \( x_i \geq X_{(k)} \) in (3.5.2)). This is because restricting \( x_i \leq X_{(k)} \) with \( \rho = 1 \) causes us to examine the minimal order statistics of the response. For either extreme and for intermediate values of the correlation, the pairs of subsets considered will tend to produce less extreme p-values than would be
expected by comparing maximal order statistics to minimal order statistics as we do in Table 4 and Table 6.

Holding other parameters fixed, the critical p-value tends to decrease (become more extreme) as \( n_r, n_c, \) or \( \rho \) increase and increase as \( n_{\text{min}} \) increases. We chose not to tabulate negative correlation as the entries are symmetric about zero.

The critical p-value depends on the joint distribution of the response and the covariate. Further simulations with a fixed set of covariates and the responses generated from the conditional distribution of the response given the covariate revealed that the critical p-value also depends on the set of observed covariates so that the results tabulated in Table 8 do not apply to an analysis conditional on the observed covariates, but rather apply to the case of the response and covariate being random with a joint normal distribution. When approaching a real problem where the covariates are considered fixed, we would recommend that simulations be carried out using the fixed covariates and generating the responses from the assumed null conditional distribution.

The entries in Table 8 where \( \rho = 0 \) correspond to the case where the response and covariate are independent. While we fail to see why one would want to choose subsets based on a covariate assumed to be independent of the response, these critical values may be used whenever the marginal distribution of the response is approximately normal and one is willing to assume that the response and covariate are independent regardless of the distribution of the covariate. The critical p-values would even be appropriate to an analysis conditional on the observed covariates provided the assumption of independence holds.
Table 8: Half-intervals of covariate, two-sample \( t \)-test, alpha = 0.05, N = 10,000 iterations

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>( n_t )</th>
<th>( n_c )</th>
<th>( n_{\min} )</th>
<th>( p )</th>
<th>( Z(p) )</th>
<th>( \rho )</th>
<th>( n_t )</th>
<th>( n_c )</th>
<th>( n_{\min} )</th>
<th>( p )</th>
<th>( Z(p) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.945E-02</td>
<td>2.3369</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.307E-02</td>
<td>2.4818</td>
</tr>
<tr>
<td>0.25</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.768E-02</td>
<td>2.3722</td>
<td>0.25</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.175E-02</td>
<td>2.5194</td>
</tr>
<tr>
<td>0.5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.998E-02</td>
<td>2.3267</td>
<td>0.5</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.223E-02</td>
<td>2.5053</td>
</tr>
<tr>
<td>0.75</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.828E-02</td>
<td>2.3599</td>
<td>0.75</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.022E-02</td>
<td>2.5683</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.570E-02</td>
<td>2.4158</td>
<td>0.95</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>6.896E-03</td>
<td>2.7018</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.658E-02</td>
<td>2.3959</td>
<td>1</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>7.782E-03</td>
<td>2.6614</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.834E-03</td>
<td>2.5816</td>
<td>0</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.565E-02</td>
<td>2.4169</td>
</tr>
<tr>
<td>0.05</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.892E-03</td>
<td>2.5796</td>
<td>0.05</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.376E-02</td>
<td>2.4635</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.386E-03</td>
<td>2.5977</td>
<td>0.1</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.346E-02</td>
<td>2.4714</td>
</tr>
<tr>
<td>0.3</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.352E-03</td>
<td>2.5989</td>
<td>0.3</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.249E-02</td>
<td>2.4981</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.176E-03</td>
<td>2.6054</td>
<td>0.5</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>8.058E-03</td>
<td>2.6496</td>
</tr>
<tr>
<td>0.7</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>8.539E-03</td>
<td>2.6300</td>
<td>0.7</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>8.632E-03</td>
<td>2.6263</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>7.126E-03</td>
<td>2.6909</td>
<td>0.9</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.227E-02</td>
<td>2.5041</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>6.666E-03</td>
<td>2.7131</td>
<td>0.95</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.065E-02</td>
<td>2.5540</td>
</tr>
<tr>
<td>0.99</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>6.164E-03</td>
<td>2.7390</td>
<td>0.99</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>9.637E-03</td>
<td>2.5886</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>5.172E-03</td>
<td>2.7961</td>
<td>1</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>8.894E-03</td>
<td>2.6123</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.897E-02</td>
<td>2.3461</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>6.814E-03</td>
<td>2.7058</td>
</tr>
<tr>
<td>0.25</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.776E-02</td>
<td>2.3707</td>
<td>0.25</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>6.775E-03</td>
<td>2.7077</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.656E-02</td>
<td>2.3963</td>
<td>0.5</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.084E-02</td>
<td>2.5480</td>
</tr>
<tr>
<td>0.75</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.489E-02</td>
<td>2.4351</td>
<td>0.75</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.108E-02</td>
<td>2.5401</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.271E-02</td>
<td>2.4918</td>
<td>0.95</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>9.396E-03</td>
<td>2.5973</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.136E-02</td>
<td>2.5313</td>
<td>1</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>8.049E-03</td>
<td>2.6500</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.082E-02</td>
<td>2.1594</td>
<td>0</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>5.653E-03</td>
<td>2.7672</td>
</tr>
<tr>
<td>0.25</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.394E-02</td>
<td>2.1208</td>
<td>0.25</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>5.562E-03</td>
<td>2.7725</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.154E-02</td>
<td>2.1501</td>
<td>0.5</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.912E-02</td>
<td>2.3432</td>
</tr>
<tr>
<td>0.75</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2.563E-02</td>
<td>2.2318</td>
<td>0.75</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.832E-02</td>
<td>2.3591</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2.407E-02</td>
<td>2.2560</td>
<td>0.95</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.684E-02</td>
<td>2.3902</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2.163E-02</td>
<td>2.2968</td>
<td>1</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.400E-02</td>
<td>2.4572</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
<td>60</td>
</tr>
</tbody>
</table>
When $\rho = 1$, the value of the response is determined by the value of the covariate. These entries are identical to what would be obtained by selecting subsets based on the order statistics of the response rather than the covariate. When the other parameters are fixed, the entries for $\rho = 1$ tend to produce the most extreme critical p-values when compared to other values of $\rho$, the few exceptions likely due to sampling variation in estimating the critical p-value. For $\rho = 1$, outliers from one tail of the distribution of the response for each treatment will be included with certainty in the sub-samples. These outliers will be more and more influential as the sub-sample size decreases. To the extent that one treatment group has more outliers than the other; this may explain some of the most extreme p-values. We suspect the critical p-values for $\rho = 1$ are the most extreme because, for values of $\rho$ less than one, outliers have some probability of being included in sub-samples, but are not included with certainty. This heuristic argument leads us to conjecture that the most extreme critical p-value for any joint distribution between the response and the covariate where the response has a marginal normal distribution corresponds to the case of $\rho = 1$.

We would recommend that the critical p-value for $\rho = 1$ be used as a conservative estimate of the critical p-value whenever the joint distribution of the response and the covariate is unknown and the marginal distribution of the response can be assumed to be approximately normal. The empirical evidence clearly supports this recommendation when the response and covariate are assumed to have a bivariate normal distribution, and our conjecture leads to this recommendation whenever the response has a marginal normal distribution.

### 3.5.2 Two sided Wilcoxon rank sum test

In this section we consider using a Wilcoxon rank sum test and apply the inference guided data exploration restricted by half-intervals of a covariate method. The data consist of a response and a covariate which are jointly distributed as bivariate normal with correlation $\rho$. If $Y$ and $X$ are the response and covariate, respectively, then both the
treatment and control samples are drawn from the following null distribution:

\[
\begin{pmatrix} Y \\ X \end{pmatrix} \sim MVN \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).
\] (3.5.6)

Independent identically distributed samples were generated and the normal approximation to the Wilcoxon rank sum test was used for the test statistic. The exact form of the Wilcoxon rank sum test would be preferable, but was not used because it is a permutation test and, as such, was not amenable to large scale simulations. Table 9 lists the relevant parameters \((\rho, n_t, n_c, \text{ and } n_{\text{min}})\), the critical p-value for the test to control the type I error at level 0.05 \((p)\), and the Z-score associated with the critical p-value \((Z(p))\). \(N = 10,000\) iterations were performed for each simulation.

The critical p-values in this table seem to be slightly less extreme than those in Table 8 (\(t\)-test). The differences are largest for the smallest sample sizes considered and become smaller for increasing sample sizes. We suspect the difference is explained by our use of the normal approximation to the Wilcoxon test and the relatively small sample sizes employed. The small difference that persists into the larger sample sizes is likely explained by the fact that the \(t\)-test is more powerful than the Wilcoxon test for normal data. Some of the smallest sample sizes have identical critical p-values for different values of \(\rho\). The Wilcoxon test is rank based, and, as such, there are only a discrete number of possible outcomes. This explains the duplicity of critical p-values for the smallest sample sizes.

Holding other parameters fixed, the critical p-value tends to decrease (become more extreme) as \(n_t, n_c, \text{ or } \rho\) increase and increase as \(n_{\text{min}}\) increases. This is consistent with Table 8.

When the joint distribution of the response and the covariate is unknown, we recommend, as discussed earlier at the end of Section 3.5.1, that the critical p-value for \(\rho = 1\) be used as a conservative estimate of the critical p-value.
Table 9: Half-intervals of covariate, Wilcoxon test, alpha = 0.05, N = 10,000 iterations

<table>
<thead>
<tr>
<th>ρ</th>
<th>n_t</th>
<th>n_c</th>
<th>n_min</th>
<th>p</th>
<th>Z(p)</th>
<th>ρ</th>
<th>n_t</th>
<th>n_c</th>
<th>n_min</th>
<th>p</th>
<th>Z(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5.183E-02</td>
<td>1.9445</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.677E-02</td>
<td>2.3917</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5.183E-02</td>
<td>1.9445</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.456E-02</td>
<td>2.4431</td>
</tr>
<tr>
<td>0.5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5.183E-02</td>
<td>1.9445</td>
<td>0.5</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.500E-02</td>
<td>2.4325</td>
</tr>
<tr>
<td>0.75</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5.183E-02</td>
<td>1.9445</td>
<td>0.75</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.286E-02</td>
<td>2.4875</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5.183E-02</td>
<td>1.9445</td>
<td>0.95</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.085E-02</td>
<td>2.5476</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>5.183E-02</td>
<td>1.9445</td>
<td>1</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>9.108E-03</td>
<td>2.6080</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>2.156E-02</td>
<td>2.2981</td>
<td>0</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.443E-02</td>
<td>2.4464</td>
</tr>
<tr>
<td>0.05</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>2.186E-02</td>
<td>2.2929</td>
<td>0.05</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.649E-02</td>
<td>2.3980</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>2.168E-02</td>
<td>2.2959</td>
<td>0.1</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.495E-02</td>
<td>2.4335</td>
</tr>
<tr>
<td>0.3</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>2.157E-02</td>
<td>2.2978</td>
<td>0.3</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.258E-02</td>
<td>2.4954</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.996E-02</td>
<td>2.3270</td>
<td>0.5</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>9.837E-03</td>
<td>2.5815</td>
</tr>
<tr>
<td>0.7</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.996E-02</td>
<td>2.3270</td>
<td>0.7</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>9.490E-03</td>
<td>2.5939</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.842E-02</td>
<td>2.3571</td>
<td>0.9</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.326E-02</td>
<td>2.4767</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.680E-02</td>
<td>2.3910</td>
<td>0.95</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.256E-02</td>
<td>2.4959</td>
</tr>
<tr>
<td>0.99</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.491E-02</td>
<td>2.4346</td>
<td>0.99</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.152E-02</td>
<td>2.5264</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>1.601E-02</td>
<td>2.4087</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.034E-02</td>
<td>2.5641</td>
</tr>
<tr>
<td>0.25</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2.387E-02</td>
<td>2.2592</td>
<td>0.25</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>7.458E-03</td>
<td>2.6757</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2.327E-02</td>
<td>2.2690</td>
<td>0.5</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>6.348E-03</td>
<td>2.7293</td>
</tr>
<tr>
<td>0.75</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2.299E-02</td>
<td>2.2736</td>
<td>0.75</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.147E-02</td>
<td>2.5281</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>2.113E-02</td>
<td>2.3056</td>
<td>0.95</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.099E-02</td>
<td>2.5431</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.996E-02</td>
<td>2.3270</td>
<td>1</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.111E-02</td>
<td>2.5393</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.407E-02</td>
<td>2.1193</td>
<td>0</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>6.230E-03</td>
<td>2.7354</td>
</tr>
<tr>
<td>0.25</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.734E-02</td>
<td>2.0821</td>
<td>0.25</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>5.417E-03</td>
<td>2.7811</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.680E-02</td>
<td>2.0880</td>
<td>0.5</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.995E-02</td>
<td>2.3272</td>
</tr>
<tr>
<td>0.75</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.121E-02</td>
<td>2.1544</td>
<td>0.75</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.833E-02</td>
<td>2.3588</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.049E-02</td>
<td>2.1637</td>
<td>0.95</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.648E-02</td>
<td>2.3980</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2.949E-02</td>
<td>2.1769</td>
<td>1</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.511E-02</td>
<td>2.4297</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2.949E-02</td>
<td>2.1769</td>
<td>0.95</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.179E-02</td>
<td>2.5183</td>
</tr>
<tr>
<td>1</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.067E-02</td>
<td>2.5535</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.5.3 **ANCOVA t-test**

If there is a relationship between the expected value of the response and the value of the covariate, a proper model can exploit this relationship to reduce the variance of the estimator of the mean response and provide a more powerful test. Here, we consider a situation where the expected value of the response is linearly related to the value of a covariate in both treatment groups, the slopes of the expected value of the response are assumed equal for the two groups, and the intercepts may differ. The appropriate model is the analysis of covariance. Our half-intervals methodology is applicable when the ANCOVA model is assumed to hold for at least a portion of the range of the covariate defined by a half-interval.

Table 10 shows the results of simulations applying the method of inference guided data exploration restricted by half-intervals of a covariate. The data consist of a response and a covariate which are jointly distributed as bivariate normal with correlation \( \rho \). If \( Y \) and \( X \) are the response and covariate, respectively, then both the treatment and control samples are drawn from the following null distribution:

\[
\begin{pmatrix}
Y \\
X
\end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\
0 \\
1 \rho \\
\rho 1 \end{pmatrix} \right).
\]  

(3.5.7)

Independent identically distributed samples were generated and the \( t \)-test that the coefficient of the treatment indicator (\( \beta_1 \)) is equal to zero in the ANCOVA model, \( Y = \beta_0 + \beta_1 I_{[\text{treatment}]} + \beta_2 X + \epsilon \), was used for the test statistic. The alternative again is two-sided. This model apparently controls for the correlation between response and covariate so that it may be seen that the critical p-value no longer depends on the correlation. The table lists the relevant parameters (\( \rho, n_t, n_c, n_{\text{min}} \)), the critical p-value for the test to control the type I error at level 0.05 (p), and the Z-score associated with the critical value (\( Z(p) \)). For each simulation, we performed \( N = 10,000 \) iterations.
Table 10: Half-intervals of a covariate, ANCOVA t-test, alpha = 0.05, N = 10,000 iterations

<table>
<thead>
<tr>
<th>ρ</th>
<th>n_t</th>
<th>n_c</th>
<th>n_min</th>
<th>p</th>
<th>Z(p)</th>
<th>ρ</th>
<th>n_t</th>
<th>n_c</th>
<th>n_min</th>
<th>p</th>
<th>Z(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.677E-02</td>
<td>2.3916</td>
<td>0</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.381E-02</td>
<td>2.4622</td>
</tr>
<tr>
<td>0.25</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.736E-02</td>
<td>2.3790</td>
<td>0.25</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.330E-02</td>
<td>2.4756</td>
</tr>
<tr>
<td>0.5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.677E-02</td>
<td>2.3917</td>
<td>0.5</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.135E-02</td>
<td>2.5316</td>
</tr>
<tr>
<td>0.75</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.754E-02</td>
<td>2.3752</td>
<td>0.75</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.349E-02</td>
<td>2.4706</td>
</tr>
<tr>
<td>0.95</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.734E-02</td>
<td>2.3794</td>
<td>0.95</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.284E-02</td>
<td>2.4881</td>
</tr>
<tr>
<td>0.99</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1.590E-02</td>
<td>2.4111</td>
<td>0.99</td>
<td>20</td>
<td>20</td>
<td>8</td>
<td>1.238E-02</td>
<td>2.5011</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.037E-03</td>
<td>2.6107</td>
<td>0</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.535E-02</td>
<td>2.4240</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.068E-03</td>
<td>2.6095</td>
<td>0.25</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.564E-02</td>
<td>2.4172</td>
</tr>
<tr>
<td>0.3</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>8.617E-03</td>
<td>2.6269</td>
<td>0.5</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.491E-02</td>
<td>2.4346</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>8.448E-03</td>
<td>2.6336</td>
<td>0.75</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.475E-02</td>
<td>2.4386</td>
</tr>
<tr>
<td>0.7</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>8.398E-03</td>
<td>2.6356</td>
<td>0.95</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.487E-02</td>
<td>2.4354</td>
</tr>
<tr>
<td>0.9</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>8.387E-03</td>
<td>2.6361</td>
<td>0.99</td>
<td>30</td>
<td>30</td>
<td>15</td>
<td>1.574E-02</td>
<td>2.4150</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.052E-03</td>
<td>2.6101</td>
<td>0</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.201E-02</td>
<td>2.5119</td>
</tr>
<tr>
<td>0.99</td>
<td>10</td>
<td>10</td>
<td>2</td>
<td>9.300E-03</td>
<td>2.6008</td>
<td>0.25</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.129E-02</td>
<td>2.5336</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.874E-02</td>
<td>2.3507</td>
<td>0.5</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.171E-02</td>
<td>2.5207</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.945E-02</td>
<td>2.3368</td>
<td>0.75</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.163E-02</td>
<td>2.5230</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.699E-02</td>
<td>2.3870</td>
<td>0.95</td>
<td>50</td>
<td>50</td>
<td>20</td>
<td>1.285E-02</td>
<td>2.4880</td>
</tr>
<tr>
<td>0.75</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.803E-02</td>
<td>2.3651</td>
<td>0</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.123E-02</td>
<td>2.5355</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.924E-02</td>
<td>2.3409</td>
<td>0.25</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.100E-02</td>
<td>2.5426</td>
</tr>
<tr>
<td>0.99</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>1.665E-02</td>
<td>2.3944</td>
<td>0.5</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.090E-02</td>
<td>2.5460</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>2.802E-02</td>
<td>2.1970</td>
<td>0.75</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>9.993E-03</td>
<td>2.5761</td>
</tr>
<tr>
<td>0.25</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.078E-02</td>
<td>2.1598</td>
<td>0.95</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.089E-02</td>
<td>2.5464</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.167E-02</td>
<td>2.1486</td>
<td>0.99</td>
<td>80</td>
<td>80</td>
<td>30</td>
<td>1.025E-02</td>
<td>2.5672</td>
</tr>
<tr>
<td>0.75</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.289E-02</td>
<td>2.1334</td>
<td>0</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.933E-02</td>
<td>2.3391</td>
</tr>
<tr>
<td>0.95</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.219E-02</td>
<td>2.1420</td>
<td>0.25</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.849E-02</td>
<td>2.3557</td>
</tr>
<tr>
<td>0.99</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>3.032E-02</td>
<td>2.1659</td>
<td>0.5</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>1.885E-02</td>
<td>2.3484</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Although the critical p-value no longer depends on the parameter $\rho$, the small variation seen in the critical p-values when the other parameters are fixed is due to sampling variation from the simulations. By employing a more complicated model which controls for the correlation, we can eliminate the dependence between the value of the critical p-value and the parameter $\rho$. This is true when the model properly reflects the underlying joint distribution.

In Section 3.5.1, we pointed out that the method restricted by half-intervals of a covariate produced critical p-values appropriate for an analysis where the response and covariate are jointly random and that, conditional on the observed set of covariates, the critical p-value depends on the covariates. In the current section (3.5.3), we considered a model which controls for the dependence between the response and the covariate so that a single critical p-value can be computed for the family of joint distributions where the model holds. We ran further simulations to explore whether the critical p-value for a fixed set of covariates would still depend on the covariates. Twenty-five simulations were run where, for each simulation, the set of covariates was fixed at a single realization of a sample from $X_i \sim iid \ N(0, 1)$ and the $Y_j$'s were then generated from the model $Y_j = \beta X_j + \epsilon_j$ where $\epsilon_j \sim iid \ N(0, 1)$. We used parameters $n_r = 50$, $n_c = 50$, $n_{\min} = 20$, $\beta = .5$, and $N = 10,000$ iterations for each fixed realization. The twenty-five simulated critical p-values were all consistent with the entries in Table 10 corresponding to parameters $n_r = 50$, $n_c = 50$, and $n_{\min} = 20$. We thus conclude that the critical p-values listed in Table 10 are appropriate for an analysis conditional on the observed covariate when the dependence between the response and the covariate is properly modeled.

With the other parameters fixed, the critical p-value tends to decrease (become more extreme) as $n_r$, or $n_c$, increase and increase as $n_{\min}$ increases. Unlike Table 8 and Table 9 where we used the $t$ and Wilcoxon rank sum tests to compare the treatments, the parameter $\rho$ has no effect on the critical p-value when the model controls for the correlation.
To illustrate the application of our method, we use data from the National Surgical Adjuvant Breast and Bowel Project (NSABP) Protocol B-27 (Mamounas (1997)). The NSABP is a National Cancer Institute funded cooperative group which performs clinical trials to study breast and colorectal cancer. Patients in Protocol B-27 are randomized into three groups. The control group (arm 1) receives 4 cycles of Adriamycin Cyclophosphamide (AC) chemotherapy given preoperatively along with 5 years of Tamoxifen beginning concurrently with AC. The two experimental groups receive the same AC chemotherapy and Tamoxifen followed by 4 cycles of Taxotere (T) given preoperatively (arm 2) or postoperatively (arm 3). Women with palpable operable invasive breast cancer confined to a single breast and possibly the ipsilateral (same side) axillary lymph nodes are eligible for the study, additional inclusion/exclusion criteria are described in Mamounas (1997).

In the present analysis, we are concerned with the change in tumor size between baseline (randomization) and surgery. We compare arms 1 and 3 combined, both of which received 4 cycles of AC prior to surgery with arm 2 where patients received 4 cycles of AC followed by 4 cycles of T prior to surgery. Our response variable is proportional reduction in tumor size defined as (baseline tumor size – tumor size at surgery) / baseline tumor size. Our analysis is restricted to patients whose baseline tumor size is at least 1.0 centimeter and for whom the tumor size at surgery is non-missing. A tumor size of zero at surgery is allowable and corresponds to a complete preoperative response to therapy; such patients will have a proportional reduction in tumor size value of 1.0. Our interest concerns the effect of a patient’s age on her tumor response measured by its proportional reduction, and whether or not this effect is comparable between arms 1/3 and arm 2.

Our previous experience with this dataset (further described in Section 4.6) suggests the effect of Taxotere in helping to reduce tumor size is more pronounced in younger women. We apply the method of Inference Guided Data Exploration by Half-intervals of a Covariate to this dataset using proportional reduction in tumor size as the response, age as the covariate with subsets formed by including all women no older than a series of decreasing age thresholds, and the two sample t-test as the test of hypothesis. The control
group (arms 1/3) has 1,117 observations, and the group treated with Taxotere preoperatively (arm 2) has 512 observations. We do not wish to consider any subsets where either treatment arm has less than 50 observations. A conservative estimate of the critical p-value was obtained by simulation using bivariate normal data and parameters $\alpha = 0.05$, $\rho = 1$, $n_t = 512$, $n_c = 1,117$, $n_{\text{min}} = 50$, and $N = 10,000$ iterations by the same methodology used in Section 3.5.1. Since we do not know the true relationship between age and proportional reduction in tumor size, we used a correlation of 1 as this produced the most extreme critical p-values in Section 3.5.1. The simulation yielded an estimated critical p-value of 0.00267. There were no more than $(n_t + n_c - 2n_{\text{min}} + 2) = 1,531$ comparisons per iteration of the simulation, and the simulation required approximately 2 hours on a 700 MHz Pentium III PC.

We performed a series of two sample $t$-tests where we compared the control and treatment groups on subsets of the data beginning with the full sample comparison, then including women no older than the second highest age, and continuing in this fashion until the final comparison which included women no older than 36 years of age. The next comparison would have included women no older than 35, but was not performed since the treatment group would have had less than $n_{\text{min}} = 50$ observations. The results of these hypothesis tests are displayed in Table 11 where we observe that the minimum p-value obtained from this series of tests was $p = 0.0000785$ and corresponded to a subset defined as all women no older than 53 years of age. Since the observed minimum p-value (0.0000785) was less than the critical p-value (0.00267), we reject the null hypothesis of no difference in proportional reduction in tumor size between the treatments and conclude that among women no older than 53 the reduction is greater in women receiving Taxotere than in women not receiving Taxotere at the $\alpha = 0.05$ level.
Table 11: P values of \( t \)-test comparing all patient no older than tabulated age

<table>
<thead>
<tr>
<th>Maximum Age (threshold defining subset)</th>
<th>Number of control patients in subset</th>
<th>Number of treatment patients in subset</th>
<th>P-value of ( t )-test on this subset</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>1,117</td>
<td>512</td>
<td>0.0003732</td>
</tr>
<tr>
<td>77</td>
<td>1,116</td>
<td>512</td>
<td>0.0003625</td>
</tr>
<tr>
<td>76</td>
<td>1,114</td>
<td>512</td>
<td>0.0003719</td>
</tr>
<tr>
<td>74</td>
<td>1,114</td>
<td>511</td>
<td>0.0003946</td>
</tr>
<tr>
<td>73</td>
<td>1,114</td>
<td>509</td>
<td>0.0004917</td>
</tr>
<tr>
<td>72</td>
<td>1,110</td>
<td>507</td>
<td>0.0004783</td>
</tr>
<tr>
<td>71</td>
<td>1,106</td>
<td>506</td>
<td>0.0005853</td>
</tr>
<tr>
<td>70</td>
<td>1,101</td>
<td>504</td>
<td>0.0004348</td>
</tr>
<tr>
<td>69</td>
<td>1,091</td>
<td>501</td>
<td>0.0004135</td>
</tr>
<tr>
<td>68</td>
<td>1,083</td>
<td>498</td>
<td>0.0004609</td>
</tr>
<tr>
<td>67</td>
<td>1,079</td>
<td>495</td>
<td>0.0004810</td>
</tr>
<tr>
<td>66</td>
<td>1,076</td>
<td>495</td>
<td>0.0004105</td>
</tr>
<tr>
<td>65</td>
<td>1,072</td>
<td>488</td>
<td>0.0004817</td>
</tr>
<tr>
<td>64</td>
<td>1,054</td>
<td>483</td>
<td>0.0004134</td>
</tr>
<tr>
<td>63</td>
<td>1,040</td>
<td>473</td>
<td>0.0001582</td>
</tr>
<tr>
<td>62</td>
<td>1,026</td>
<td>465</td>
<td>0.0003010</td>
</tr>
<tr>
<td>61</td>
<td>1,009</td>
<td>460</td>
<td>0.0001957</td>
</tr>
<tr>
<td>60</td>
<td>988</td>
<td>454</td>
<td>0.0001452</td>
</tr>
<tr>
<td>59</td>
<td>969</td>
<td>447</td>
<td>0.0001578</td>
</tr>
<tr>
<td>58</td>
<td>941</td>
<td>435</td>
<td>0.0002719</td>
</tr>
<tr>
<td>57</td>
<td>912</td>
<td>422</td>
<td>0.0004558</td>
</tr>
<tr>
<td>56</td>
<td>887</td>
<td>411</td>
<td>0.0009168</td>
</tr>
<tr>
<td>55</td>
<td>856</td>
<td>399</td>
<td>0.0008892</td>
</tr>
<tr>
<td>54</td>
<td>829</td>
<td>377</td>
<td>0.0015703</td>
</tr>
<tr>
<td>53</td>
<td>801</td>
<td>361</td>
<td>0.0000785</td>
</tr>
<tr>
<td>52</td>
<td>773</td>
<td>344</td>
<td>0.0000807</td>
</tr>
<tr>
<td>51</td>
<td>739</td>
<td>333</td>
<td>0.0001388</td>
</tr>
<tr>
<td>50</td>
<td>684</td>
<td>307</td>
<td>0.0001041</td>
</tr>
<tr>
<td>49</td>
<td>627</td>
<td>292</td>
<td>0.0003315</td>
</tr>
<tr>
<td>48</td>
<td>588</td>
<td>273</td>
<td>0.0008473</td>
</tr>
<tr>
<td>47</td>
<td>540</td>
<td>245</td>
<td>0.0020021</td>
</tr>
<tr>
<td>46</td>
<td>486</td>
<td>225</td>
<td>0.0024372</td>
</tr>
<tr>
<td>45</td>
<td>440</td>
<td>199</td>
<td>0.0032663</td>
</tr>
<tr>
<td>44</td>
<td>396</td>
<td>179</td>
<td>0.0026331</td>
</tr>
<tr>
<td>43</td>
<td>365</td>
<td>162</td>
<td>0.0028408</td>
</tr>
<tr>
<td>42</td>
<td>327</td>
<td>139</td>
<td>0.0052053</td>
</tr>
<tr>
<td>41</td>
<td>297</td>
<td>122</td>
<td>0.0102753</td>
</tr>
<tr>
<td>40</td>
<td>260</td>
<td>102</td>
<td>0.0024475</td>
</tr>
<tr>
<td>39</td>
<td>225</td>
<td>91</td>
<td>0.0050965</td>
</tr>
<tr>
<td>38</td>
<td>189</td>
<td>75</td>
<td>0.0059949</td>
</tr>
<tr>
<td>37</td>
<td>155</td>
<td>65</td>
<td>0.0102552</td>
</tr>
<tr>
<td>36</td>
<td>130</td>
<td>58</td>
<td>0.0330878</td>
</tr>
</tbody>
</table>
3.7 Discussion

The method of inference guided data exploration restricted by minimum subset sample size, which we considered in Section 3.4, provides a critical p-value for the minimum p-value after testing all possible pairs of subsets at least as large as the minimum sample size criteria. Direct application of the method can be computationally prohibitive for all but the smallest sample sizes as we observed in Section 3.4.2 where we applied the method using the two sample $t$-test as our test of hypothesis. The shortcut method of only considering pairs of sub-samples where maximal order statistics are compared to minimal order statistics makes the method computationally feasible, but this approach only yields exact critical p-values for certain hypothesis tests such as the two sample $Z$-test considered in Section 3.4.1. The method produces exact results asymptotically. Our work with finite simulations of $N = 10,000$ iterations seemed to reproduce the first two significant digits of our critical p-values consistently. Another hypothesis test which lends itself to the shortcut method is the Wilcoxon rank sum test. For any given pair of sub-sample sizes $n_i$ and $n_j$, the p-value for this test depends only on the sum of the pooled ranks from one of the treatments. The most extreme rank sum for a given pair of sample sizes will always come from a pair of sub-samples which compare maximal order statistics to minimal order statistics. The shortcut method is also useful for finding an approximate critical p-value for other tests of hypothesis such as the two sample $t$-test considered in Section 3.4.3. We found this approximate critical p-value to be a better approximation for the two sample $t$-test than using the critical p-value from the two sample $Z$-test.

In applying the method restricted by minimum subset sample size, one tests the null hypothesis of no difference between treatments in any pair of subsets considered versus the alternative that the treatments differ in at least one of the pairs of subsets considered. Concluding the alternative does not lead to a generalizable result, but may lead to further research to identify the factors related to the variation in the effect of treatment. When considering only the sample under study and not attempting to generalize the results, any pair of subsets tested which produced a p-value less than the critical p-value is identified as a pair of subsets where the treatments differ. This methodology will be useful to
provide a “rule of thumb” for a statistician to judge potential significance when encountering a researcher who has already been on a fishing expedition in his or her data. This approach may also prove useful as a planned methodology when the aim of the research is primarily exploratory in nature.

The method of inference guided data exploration restricted by half-intervals of a covariate, considered in Section 3.5, provides a critical p-value for the minimum p-value after testing multiple pairs of subsets defined by including all observations where the covariate is at least as large as (not larger than) a series of threshold values. The threshold values are taken from the order statistics of the pooled covariates beginning with the first (last), and the cardinality of the subsets considered must also satisfy a minimum sample size criteria. The methodology using the half-intervals restriction provides critical p-values which are much less extreme than those provided when the minimum sample size restriction is used alone. The critical p-values are less extreme because the number of pairs of subsets compared is smaller and some of the more extreme comparisons (i.e., comparing maximal order statistics with minimal order statistics) are avoided. For this methodology, the critical p-value depends on the relationship between the response and the covariate (correlation for a linear relationship), but, as seen in Section 3.5.3, the dependence can be removed when the relationship is properly modeled (ANCOVA model for a linear relationship). In practice, the method could be employed by using the critical p-value from the worst case correlation ($\rho = 1$) which would make the method slightly conservative. The results in Sections 3.5.1 (t-test) and 3.5.2 (Wilcoxon) are only applicable to the case where the response and covariate are jointly random, i.e., we do not consider the covariates to be fixed. Conditional on a fixed set of covariates, the critical p-value depends on the covariates observed. In Section 3.5.3 (ANCOVA), we showed that, when the model properly controls for the dependence between the response and covariate, the critical p-value no longer depends on the observed fixed set of covariates.

The methodology employing the half-intervals restriction begins with a comparison of the full sample from each treatment group and then compares successively smaller subsets until the minimum sample size criteria is reached. When the critical p-value is
available prior to using the method on a set of data, one may consider an alternative stopping criteria where the method stops with the first (largest) subset which produces a p-value less than or equal to the critical p-value. Under the null hypothesis, the $\alpha^{th}$ percentile of the minimum p-value with this additional stopping criteria is identical to the $\alpha^{th}$ percentile of the minimum p-value when this stopping criteria is not employed so that the critical p-values already tabulated apply to a method using this stopping criteria. We explore the relative merits of this stopping criteria as compared to the methodology not including this stopping criteria in Chapter 5.

We feel the methodology employing the half-intervals restriction considers the types of subsets a researcher would more commonly be interested in considering. For example, if the researcher feels that the effect of treatment may not be as strong in patients with very small values of a specific covariate, then this method could be employed to guard against inclusion criteria which were too liberal and possibly included patients where the effect of treatment was negligible. We again test the null hypothesis of no difference between treatments in any pair of subsets considered versus the alternative that the treatments differ in at least one of the pairs of subsets considered, however we are now able to generalize our results beyond the sample under study because the significant subsets are now identifiable based on the value of a covariate exceeding some threshold. As with the minimum sample size restriction, we believe the method employing the half-intervals restriction will be used to provide a “rule of thumb” for judging significance in situations where a researcher has already examined subsets with respect to a covariate and will also be used as a planned methodology when the aim of the research is primarily exploratory in nature.

If there are multiple covariates possibly related to the response, a logical extension of this work might be to restrict the subsets of interest to simultaneous half-intervals or quadrants based on two covariates or orthants when more than two covariates are considered. The solution would depend on the joint distribution of the response and the covariates. Our experience with a single covariate suggests that the most extreme critical p-values arise in the case of perfect correlation where we are essentially choosing subsets based on the order statistics of the response. We speculate that if the most extreme
critical p-values from the case of orthants for multiple covariates also corresponds to
choosing subsets based on the order statistics of the response, then the tabulated values in
Section 3.5 corresponding to $\rho = 1$ would be applicable as a conservative estimate of the
critical p-value.
4 Asymptotic Simultaneous Confidence Band for the Difference of Segmented Linear Models

4.1 Introduction and Background

A problem that has received substantial statistical attention is the comparison of two treatments when the expected value of the response depends on the value of a covariate. When the relationship between the response and the covariate is linear and the slopes of the treatments are parallel, analysis of covariance can be used to determine if the treatments differ. The problem becomes more difficult for the case of non-parallel slopes because the difference between the treatments depends on the slopes. In this case, one wants to identify regions of covariate values where the treatments are significantly different. Potthoff (1964) addressed this problem by constructing a simultaneous confidence band about the difference of the regression lines for the two treatments. The treatments are declared to be significantly different on intervals of the covariate where the simultaneous confidence band for the difference does not contain zero.

When a linear regression model does not adequately describe the effect of the covariate on the expected response, continuous segmented linear functions have been used for more general modeling. The segmented linear model has theoretical justification for a number of physical and biological processes (Julious (2001)) and can be used as an approximation for continuous functional relationships. For our purposes, we assume the number of segments to be modeled is known, and that the location(s) of the changepoint(s) are to be estimated from the data. Other authors have addressed the problem of determining the appropriate number of segments from the data, for example, Hinkley (1988). Comparing two treatments in the setting of segmented linear models is difficult because the treatments can differ in their intercepts, the slopes of each segment, and the locations of the changepoints.

In our setting, we are again interested in identifying regions of the covariate values where the two treatments differ. The approach we take is analogous to that of Potthoff (1964) where we construct a simultaneous confidence band about the difference of two
segmented linear models. In order to accomplish this, we extend results of Cox and Ma (1995) who give a general approach to construct a simultaneous confidence band for a single nonlinear regression function.

Among other areas of application, our simultaneous confidence interval approach is useful for determining the target population for a medical treatment based on the value of a covariate. For example, the effectiveness of a treatment for high blood pressure may depend on the patient’s pre-treatment blood pressure. Our technique could then be used to determine the threshold pre-treatment blood pressure value where the treatment becomes beneficial compared to its control.

Segmented linear models with sufficient numbers of segments are quite flexible as approximations to arbitrary functions. When the locations of the change points are known, parameter estimation for the segmented linear model is relatively easily accomplished through use of techniques from linear models. If the change points are to be estimated from the data, then non-linear estimation techniques must be used (Hudson, 1966).

To identify the region(s) of covariate values where two treatments with individual segmented regression functions differ, we first recognize that the difference of two segmented regression functions is also a segmented regression function. Thus, the basic issue in this setting is to find suitable simultaneous confidence bands for a single segmented regression function. Cox and Ma’s (1995) techniques, suitably modified, are used to construct these bands.

In Section 4.2, we introduce our model for the problem of comparing two treatments. Section 4.3 provides Cox and Ma’s (1995) method for computing simultaneous confidence bands for the expected value function in the setting of nonlinear regression and extends their approach to the difference of segmented linear models. In Section 4.4, we show how our method applies to the case of known changepoints in the segmented linear model. Our approach to estimation for the segmented regression problem is presented in Section 4.5, along with advice on the practical application of the method. In Section 4.6 our method is applied to breast cancer tumor size data from NSABP Protocol
B-27 (Mamounas (1997)). Simulations to describe the coverage probabilities for our simultaneous confidence bands are provided in Section 4.7 and Section 4.8 gives a discussion of the results.

4.2 Segmented Linear Model with Unknown Change Points for Two Treatments

Let \( f \) define an \( m \)-segment continuous piecewise linear function with parameter vector \( \mathbf{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{m1}, \theta_{m2})' \):

\[
f(x, \mathbf{\theta}) = \theta_{11} + \theta_{12} x + \sum_{k=2}^{m} \theta_{k2}(x - \theta_{k1}) I_{[x > \theta_{k1}]}.
\]

(4.2.1)

The parameters in (4.2.1) have the following interpretations: \( \theta_{11} \) is the \( y \) intercept of the first segment; \( \theta_{12} \) is the slope of the first segment; for \( k = 2, \ldots, m \); \( \theta_{k1} \) is the \( x \) coordinate of the point of change of the \( k^{th} \) segment; \( \theta_{k2} \) is the change in slope of the \( k^{th} \) segment relative to the \((k - 1)^{st}\); where we assume, without loss of generality, that the changepoints are ordered \( \theta_{21} < \cdots < \theta_{m1} \). The function \( f \) is continuous in \( x \), but has discontinuities of the first derivative with respect to \( x \) at the \((m - 1)\) changepoints \( \theta_{21}, \ldots, \theta_{m1} \).

For treatment group \( i \), consider the \( m_i \)-segment linear model for an observation \( Y \) with a scalar covariate \( x \) as:

\[
Y = f(x, \mathbf{\theta}^{(i)}) + \varepsilon_i
\]

\[
= \theta_{11}^{(i)} + \theta_{12}^{(i)} x + \sum_{k=2}^{m_i} \theta_{k2}^{(i)}(x - \theta_{k1}^{(i)}) I_{[x > \theta_{k1}^{(i)}]} + \varepsilon_i
\]

(4.2.2)

where \( i = 1, 2; \varepsilon_i \sim N(0, \sigma_i^2) \); and \( \sigma_1^2 \) and \( \sigma_2^2 \) are parameters. Define the joint parameter vector for the mean function of two treatments as \( \mathbf{\theta} = (\mathbf{\theta}^{(1)}', \mathbf{\theta}^{(2)}')' = (\theta_{11}^{(1)}, \theta_{12}^{(1)}, \ldots, \theta_{m_1}^{(1)}, \theta_{m_2}^{(1)},\theta_{11}^{(2)}, \theta_{12}^{(2)}, \ldots, \theta_{m_1}^{(2)}, \theta_{m_2}^{(2)})' \). Clearly \( E_{\mathbf{\theta}^{(i)}}(Y \mid x) = f(x, \mathbf{\theta}^{(i)}) \). Now suppose that we have independent random samples of size \( n_i \) for treatment group \( i \), where these observations are denoted by \( (Y_{ij}, x_{ij}); i = 1, 2; j = 1, \ldots, n_i \).

We can compare the two treatment groups using the difference of the mean functions in terms of the covariate. The difference is expressed as:

\[
d(x, \mathbf{\theta}) = E_{\mathbf{\theta}^{(1)}}(Y \mid x) - E_{\mathbf{\theta}^{(2)}}(Y \mid x) = f(x, \mathbf{\theta}^{(1)}) - f(x, \mathbf{\theta}^{(2)}).
\]

(4.2.3)
For a given \( x \), positive values of \( d \) indicate the expected value of treatment group 2 is
greater than that of group 1 at that value of \( x \).

As the difference of two segmented lines with \( (m_1 - 1) \) and \( (m_2 - 1) \) changepoints,
respectively, \( d \) must be a segmented line with \( (m_1 + m_2 - 2) \) changepoints assuming all
changepoints are distinct. Since for every segmented line with \( k \) changepoints there are
in general \( k + 1 \) slopes and 1 intercept, it follows that one needs \( (m_1 + m_2 - 2) + (m_1 + m_2
- 2 + 1 + 1) = 2(m_1 + m_2 - 1) \) parameters to describe \( d \). Although \( d \) is defined in terms of
the \( 2(m_1 + m_2) \) parameters in \( \theta \), it is clear that \( 2(m_1 + m_2 - 1) \) parameters would be
sufficient to describe the \( m_1 + m_2 - 1 \) segment linear function. While a
reparameterization is possible, it is not necessary for our purposes.

Since the errors are normally distributed and independent between treatment groups and
the subsets of the parameter vector relevant to the 2 groups are distinct, we can obtain the
maximum likelihood estimate (MLE) of \( \theta \) from the individual MLE’s of \( \theta^{(1)} \) and \( \theta^{(2)} \),
\( \hat{\theta}^{(1)} \) and \( \hat{\theta}^{(2)} \), using the individual models (4.2.2), with \( i = 1, 2 \). The difference function
\( d(x, \theta) \) given by (4.2.3) is then estimated by:

\[
d(x, \hat{\theta}) = f(x, \hat{\theta}^{(2)}) - f(x, \hat{\theta}^{(1)}).
\]

4.3 Construction of Simultaneous Confidence Band

Cox and Ma (1995) give a method for constructing a simultaneous confidence band about
the mean function for generalized nonlinear regression models. Let \( Y_1, \ldots, Y_n \) be \( n \)
independent random variables where the cumulative distribution functions (cdf’s) of
\( Y_1, \ldots, Y_n \) respectively, are \( H_1(y|\theta), \ldots, H_n(y|\theta) \), \( \theta = (\theta_1, \ldots, \theta_p) \in \Theta \). Further suppose
there are covariates \( x_1, \ldots, x_n \), associated with \( Y_1, \ldots, Y_n \), respectively, and a cdf \( H(y|x, \theta) \)
which depends on \( \theta \in \Theta \) and \( x \), such that finally \( H_j(y|x, \theta) = H(y|x_j, \theta) \) for \( j = 1, \ldots, n \). For
this setting and the case of vector valued covariates, Cox and Ma (1995) provide a
simultaneous confidence band for the mean \( E_\theta(Y) = g(x, \theta) \) viewed as a function of \( x \).

They require that the parameter estimates \( \hat{\theta}_n \) of \( \theta \), in their model based upon a sample of
size \( n \), have an asymptotic joint normal distribution, that is \( \sqrt{n}(\hat{\theta}_n - \theta) \overset{L}{\to} N(\theta, \Sigma). \)
Further, they require that there is a consistent estimator $J_n(\hat{\theta}_n)$ of the covariance matrix of $\hat{\theta}_n$, so that $n J_n(\hat{\theta}_n) \overset{p}{\to} \Sigma$. When there is no risk of ambiguity, we suppress the notation showing the dependence of the estimators of the parameters and covariance matrix on the sample size $n$.

The form of Cox and Ma’s (1995) asymptotic $100(1 - \alpha)\%$ simultaneous confidence band for $g(x, \theta)$ is:

$$g(x, \hat{\theta}) \pm \left( \frac{\chi^2_{p,1-\alpha}}{\sqrt{n}} \right)^{1/2} \left[ J(\hat{\theta}) \left( \frac{\partial g(x, \theta)}{\partial \theta} \bigg| \theta = \hat{\theta} \right) \right]^{1/2},$$

(4.3.1)

where $\partial g(x, \theta)/\partial \theta = \left( \frac{\partial g(x, \theta)}{\partial \theta_1}, \ldots, \frac{\partial g(x, \theta)}{\partial \theta_p} \right)'$ and $\chi^2_{p,1-\alpha}$ denotes the $(1 - \alpha)^{th}$ quantile of the chi-squared distribution on $p$ degrees of freedom.

The simultaneous confidence interval is valid over the range of the observed values of the covariate. The band can be extended beyond the range of the observed covariates if we assume the functional form of $g$ is correct for covariate values outside the observed range. Extrapolating far beyond the original data becomes impractical because the width of the simultaneous confidence bands tends to increase substantially as the covariate value moves further from the range of the observed data.

Cox and Ma (1995) only consider the confidence band in the case where the function $g$ is the expected value function from a single nonlinear regression. For our context, we need to use their technique to find a simultaneous confidence band for the difference of the expected value functions from two nonlinear regressions.

To construct their confidence band, they first use the Cauchy-Schwarz inequality to derive simultaneous confidence bands for all linear combinations of the parameters. Next, they use the delta method to extend the band to $g(x, \theta)$. However, this step is applicable to any function $g$ of the parameters where the delta method is applicable (i.e., the partial derivatives with respect to the parameters exist). Hence, the method can be applied to the difference function in (4.2.3) provided that: (a) we can obtain an estimator
of the joint parameter vector, \( \boldsymbol{\theta} = (\theta^{(1)'}, \theta^{(2)'})' \), which has an asymptotic normal distribution; (b) we can obtain an estimator of the covariance matrix of the estimators of the joint parameters which is consistent; and (c) the partial derivatives of the function (4.2.3) with respect to the parameters in \( \boldsymbol{\theta} \) exist.

### 4.3.1 Asymptotic Normality Requirement

The asymptotic distribution theory appropriate to linear segmented regression was obtained by Feder (1975, Corollary 4.5, p. 71) and can be applied to the case of normally distributed errors. He shows that the maximum likelihood estimators asymptotically have a joint normal distribution. This asymptotic result is shown to hold whenever the model is identifiable which can be summarized by the following three conditions: (i) the changepoints (given by \( \theta^{(i)}_{k1}, i = 1, 2, k = 2, \ldots, m_i \)) are distinct; (ii) the change in slope between adjacent segments (given by \( \theta^{(i)}_{k2}, i = 1, 2, k = 2, \ldots, m_i \)) are not zero; and (iii) a sufficient number of observations (in the available sample) have covariate values on each segment of the model and the covariate values are arranged in such a way that the parameters associated with each segment can be identified. For our purposes in the setting of linear segmented regression, condition (iii) is satisfied if, for each segment, there are at least two observations having distinct covariate values on the segment. A further condition for the validity of the asymptotic result relates to the distribution of the covariates which can be summarized by the following three conditions: (iv) the covariates in the sample are selected in such a way that the empirical distribution formed from the covariates converges in distribution to some distribution function \( F \); (v) the observed covariate values represent points of increase in the distribution \( F \) (i.e., \( dF/dx \) exists at this point and \( dF/dx > 0 \) or \( F \) has a step at this point); and (vi) \( F \) is continuous at each of the changepoints and has support on \((A, B)\), where all of the changepoints are strictly greater than \( A \) and strictly less than \( B \). Condition (v) assures that if the model is identifiable based on the sample, then it will remain identifiable in the limit. Conditions (iv) and (v) are clearly satisfied when the covariates are drawn at random from a distribution \( F \) satisfying (vi). Certain nonrandom schemes for selecting the covariates may also satisfy these conditions, for example, if the covariates are selected equally.
spaced on \((A, B)\), then the limiting distribution \(F\) is Uniform\((A, B)\), and conditions \((iv)\) and \((v)\) are satisfied.

Feder (1975) used a different parameterization for the segmented linear model than we have chosen; however, we show in Appendix D that from his results and by use of the multivariate delta method, the parameter estimates for our parameterization also have an asymptotic joint normal distribution. For the \(i^{th}\) group, the asymptotic distribution of the MLE \(\hat{\theta}^{(i)}\) is given by:

\[
\sqrt{n} \left( \hat{\theta}^{(i)} - \theta^{(i)} \right) \xrightarrow{L} N(0, \Sigma_i),
\]

(4.3.2)

where, based upon the Fisher’s information \(I(\cdot)\) about \(\theta^{(i)}\) contained in a single observation, which depends on the value of the covariate, we have that \(\Sigma_i = \left[ \int I(\theta^{(i)}, x) \, dF(x) \right]^{-1}\) (i.e., the inverse of the expected value with respect to the distribution of the covariate of the information matrix).

We use the inverse of the plug-in estimator of the information matrix, i.e \(\hat{\Sigma}_i(\hat{\theta}^{(i)}) = I(\hat{\theta}^{(i)}, x_1, \ldots, x_m)^{-1} = [\sum_{j=1}^n I(\hat{\theta}^{(i)}, x_j)]^{-1}\), as our estimator of the covariance matrix of the parameter estimates \(\hat{\theta}^{(i)}\). The estimator is consistent since, as shown in Appendix E, \(n \hat{\Sigma}_i(\hat{\theta}^{(i)}) \xrightarrow{p} \Sigma_i\). The convergence of the covariance estimator requires two further assumptions. Let \(X_n\) be a random variable with distribution given by the empirical distribution formed from the covariates for a sample of size \(n\) and \(X\) be a random variable with the limiting distribution \(F\), then for some positive \(\delta\); \((i)\) \(E(|X_n|^{2+\delta}) < \infty\) for all \(n\) and \((ii)\) \(E(|X|^{2+\delta}) < \infty\). The first assumption implies that the random variables \(X_n\) and \(X_n^2\) are uniformly integrable and the existence of \(\Sigma_i\) is guaranteed by the second assumption.

It follows by independence that \(\hat{\theta}\) has an asymptotic normal distribution with covariance:

\[
\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix},
\]

(4.3.3)

where \(\Sigma_1\) and \(\Sigma_2\) are the asymptotic covariance matrices of \(\hat{\theta}^{(1)}\) and \(\hat{\theta}^{(2)}\), respectively.
4.3.2 The Estimated Covariance Matrix of the Parameter Estimates

From Section 4.3.1, it follows that a consistent estimator \( \hat{\Sigma}(\hat{\theta}) \), of \( \Sigma \), is given by:

\[
\hat{\Sigma}(\hat{\theta}) = \begin{pmatrix}
\hat{\Sigma}_1(\hat{\theta}^{(1)}) & 0 \\
0 & \hat{\Sigma}_2(\hat{\theta}^{(2)})
\end{pmatrix} = \begin{pmatrix}
[I(\hat{\theta}^{(1)}), x_{11}, \ldots, x_{1n_1}]^{-1} & 0 \\
0 & [I(\hat{\theta}^{(2)}), x_{21}, \ldots, x_{2n_2}]^{-1}
\end{pmatrix}, \tag{4.3.4}
\]

where

\[
I(\hat{\theta}^{(i)}, x_{i1}, \ldots, x_{im_i}) = \frac{1}{\sigma_i^2} \sum_{j=1}^{n} \left( \frac{\partial f(x_j, \theta^{(i)})}{\partial \theta^{(i)}} \bigg|_{\theta^{(i)} = \hat{\theta}^{(i)}} \right) \left( \frac{\partial f(x_j, \theta^{(i)})}{\partial \theta^{(i)}} \bigg|_{\theta^{(i)} = \hat{\theta}^{(i)}} \right) ', \tag{4.3.5}
\]

and

\[
\frac{\partial f(x, \theta^{(i)})}{\partial \theta^{(i)}} = \begin{pmatrix} 1, x, -\theta_{21}^{(i)} I_{x \geq \theta_{21}^{(i)}} \end{pmatrix}' \begin{pmatrix} x - \theta_{21}^{(i)} I_{x \geq \theta_{21}^{(i)}} \end{pmatrix}', \ldots, \begin{pmatrix} -\theta_{m_2}^{(i)} I_{x \geq \theta_{m_1}^{(i)}} \end{pmatrix}' \begin{pmatrix} x - \theta_{m_2}^{(i)} I_{x \geq \theta_{m_1}^{(i)}} \end{pmatrix}', \tag{4.3.6}
\]

Equations (4.3.5) and (4.3.6) are derived in Appendix E. See Appendix E for discussion of the nonexistence of some of the partial derivatives at the \((m_1 + m_2 - 2)\) covariate values coinciding with the values of the changepoints \( \theta_{k1}^{(i)}, k \geq 2 \). Equation (4.3.6) substitutes the value zero for these partials when they do not exist.

Notice that, unlike Potthoff’s (1964) method for the difference of linear models, we have not assumed a common error variance. Cox and Ma’s (1995) simultaneous confidence bands are asymptotic, and they essentially assume the covariance is known allowing the use of a critical value from the chi-squared distribution. Potthoff’s (1964) method is not asymptotic; however, the use of a critical value from the \( F \) distribution forces the assumption of a common error variance to remove dependence on the unknown true variance(s). There appears to be no compelling reason from the mechanics of estimation or inference for this problem to assume a common variance, nevertheless, we describe how to estimate the common variance should this assumption be appropriate.

Our rationale follows from a straightforward argument for a model where the two treatment groups are assumed to have a common variance:

\[
Y_{ij} = f(x_j, \theta^{(i)}) + \varepsilon_j, \quad i = 1, 2, \quad j = 1, \ldots, n_i, \quad \varepsilon_j \sim \text{iid } N(0, \sigma^2). \tag{4.3.7}
\]
Imagine fitting the model (4.3.7) simultaneously for the two treatment groups in order to estimate the common variance. The MLE, \( \hat{\theta} = (\hat{\theta}^{(1)}, \hat{\theta}^{(2)})' \), minimizes the sum of squared errors (SSE), and the usual MLE based estimate of the common variance, \( \hat{\sigma}^2 \), is the mean squared error \( MSE = \frac{SSE_{\text{min}}}{n_1 + n_2 - 2m_1 - 2m_2} \). Let \( SSE_1 \) and \( SSE_2 \) be the sum of squared errors due to treatment groups 1 and 2 respectively, so that \( SSE = SSE_1 + SSE_2 \).

\( SSE_1 \) depends only on \( \hat{\theta}^{(1)} \) and \( SSE_2 \) depends only on \( \hat{\theta}^{(2)} \). Minimizing \( SSE \) with respect to \( \hat{\theta} \) is equivalent to minimizing \( SSE_1 \) with respect to \( \hat{\theta}^{(1)} \) and minimizing \( SSE_2 \) with respect to \( \hat{\theta}^{(2)} \). But, this is equivalent to fitting separate models for each of the 2 groups using the model (4.2.2) except that we would now produce separate estimates \( \hat{\sigma}_i^2 = \frac{MSE_i}{n_i - 2m_i} \) of the variance for each group. The pooled variance estimate of the \( MSE, \hat{\sigma}_p^2 \) (4.3.8), is a function of the individual group variance estimates obtained by fitting separate models for each group and is equivalent to the variance estimate obtained by fitting the model (4.3.7).

\[ \hat{\sigma}_p^2 = \left[ \frac{(n_1 - 2m_1) \hat{\sigma}_1^2 + (n_2 - 2m_2) \hat{\sigma}_2^2}{n_1 + n_2 - 2m_1 - 2m_2} \right] \] (4.3.8)

To assume a common error variance for the treatment groups, substitute \( \hat{\sigma}_p^2 \) (4.3.8) for \( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_2^2 \) in the formula for the estimated information (4.3.5).

While we have illustrated this argument using our nonlinear regression function \( f(4.2.1) \), the argument is general and applies to any regression function, linear or nonlinear, where the parameters associated with the two groups are distinct.

### 4.3.3 Partial Derivatives

In order to implement the method of Cox and Ma (1995), we must compute the partial derivatives of \( d(x, \theta) \) given by (4.2.3) with respect to \( \theta \). For fixed \( x \):

\[ \frac{\partial d(x, \theta)}{\partial \theta} = \left( -\frac{\partial f(x, \theta^{(1)})}{\partial \theta^{(1)}} \right)' \left( \frac{\partial f(x, \theta^{(2)})}{\partial \theta^{(2)}} \right)' \right)' \quad \theta_{k_1}^{(i)} \neq x, \ k \geq 2, \] (4.3.9)

where \( \frac{\partial f(x, \theta^{(i)})}{\partial \theta^{(i)}} \) is given by (4.3.6). See Appendix E for discussion of the
nonexistence of the derivative at a finite number of covariate values coinciding with the
values of the changepoints $\theta_{k1}^{(i)}$, $k \geq 2$.

4.3.4 Equation for Simultaneous Confidence Band

Using our asymptotic normal results, our consistent covariance estimator, and the
computed partial derivatives of $d(x, \theta)$, we use the results in (4.3.1) of Cox and Ma
(1995) to obtain a $100(1 - \alpha)\%$ asymptotic simultaneous confidence band for $d(x, \theta)$:

$$d(x, \hat{\theta}) \pm \sqrt{\chi^2_{2(m_1 + m_2 - 1), 1-\alpha} \left( \frac{\partial d(x, \theta)}{\partial \theta} \bigg| \theta = \hat{\theta} \right) \hat{\Sigma}(\hat{\theta}) \left( \frac{\partial d(x, \theta)}{\partial \theta} \bigg| \theta = \hat{\theta} \right)}.$$  (4.3.10)

The band is undefined for $\hat{\theta}_{k1}^{(i)} = x$ where $k \in \{2, \ldots, m_1\}$ and $i \in \{1, 2\}$. This will occur at
$m_1 + m_2 - 2$ covariate values coinciding with the estimated changepoints. Since these are
the natural points of discontinuity in the confidence intervals, we omit these points from
our confidence band. Confidence bands which are simultaneous for all but a finite
number of points will still serve as simultaneous confidence bands. The band is valid
over the range of the observed covariate values and can be extrapolated if we assume the
functional form of $d$ holds beyond the range of the observed data.

As noted in Section 4.2, only $2(m_1 + m_2 - 1)$ parameters are necessary to describe the
segmented line $d$ as opposed to the $2(m_1 + m_2)$ parameters in $\theta$. If they chose, one could
reparameterize the model, but we note that the quantity under the radical sign in (4.3.10)
would be identical to the value derived from the reparameterized model up to the value of
the $\chi^2$ multiplier. It is necessary, however, to use the degrees of freedom for the $\chi^2$
multiplier from the reduced model which explains why the degrees of freedom are not
equal to the number of elements in $\theta$.

We expect this asymptotic simultaneous confidence band to be somewhat conservative
for large samples. We note that the underlying probability inequality used by Cox and
Ma (1995) in deriving their bands is a sharp inequality for the space of all linear
combinations of the parameters. However, the set of linear combinations of the
parameters necessary to approximate the segmented linear function $d$ is a subset of all
linear combinations. It appears that, in most cases, the resulting approximate set will not
be large enough to make the underlying inequality sharp. As a result, this will cause conservativeness in the coverage probability.

For actual use, another contribution to the conservative nature of the bands is that they would be typically only used for the observed range of the covariate rather than the entire line.

4.4 Segmented Linear Models with Known Changepoints

While the case of known changepoints is not a focus of this chapter, we note that a few minor changes allow the preceding technique to be applied to the case where the changepoints are known. We reparameterize the model (4.2.2) as:

\[ Y = f(x, \theta^{(i)}) + \varepsilon_i = \theta^{(i)}_1 + \theta^{(i)}_2 x + \sum_{k=2}^{m} \theta^{(i)}_k (x - \gamma^{(i)}_k) I_{[\gamma^{(i)}_{k-1}, \gamma^{(i)}_k)} + \varepsilon_i \] (4.4.1)

where \(i = 1, 2; \gamma^{(i)}_2 < \ldots < \gamma^{(i)}_m\) are the known changepoints; and \(\varepsilon_i \sim N(0, \sigma^2_i)\). Define the joint parameter vector for the two treatments as \(\theta = (\theta^{(1)}, \theta^{(2)})' = (\theta^{(i)}_1, \theta^{(i)}_2, \theta^{(i)}_2, \ldots, \theta^{(i)}_m, \theta^{(2)}_1, \theta^{(2)}_2, \theta^{(2)}_2, \ldots, \theta^{(2)}_m)'\).

The vector of partial derivatives of \(f(4.3.6)\) becomes:

\[ \frac{\partial f(x, \theta^{(i)})}{\partial \theta^{(i)}} = (1, x, (x - \gamma^{(i)}_2) I_{[\gamma^{(i)}_2, \gamma^{(i)}_3)}, \ldots, (x - \gamma^{(i)}_m) I_{[\gamma^{(i)}_{m-1}, \gamma^{(i)}_m)})'. \] (4.4.2)

The degrees of freedom change for the simultaneous confidence band for the difference \(d\) so that equation (4.3.10) becomes:

\[ d(x, \hat{\theta}) \pm \sqrt{\chi^2_{(m^2 + m)} \cdot \left( \left. \frac{\partial d(x, \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}} \right)' \hat{\Sigma}(\hat{\theta}) \left( \left. \frac{\partial d(x, \theta)}{\partial \theta} \right|_{\theta = \hat{\theta}} \right)} \] (4.4.3)

Again, the degrees of freedom are adjusted to reflect the number of free parameters necessary to describe the segmented line \(d\).

All of the arguments of the previous section go through unchanged except that we no longer have the problem of the non-existence of the confidence band at a finite number of
points. The bands produced when the changepoints are known are continuous but may have a discontinuity of their first derivative at the changepoints.

4.5 Estimation and Practical Approach

To handle the numerical difficulties arising when fitting segmented linear models with unknown changepoints to data we have chosen an estimation procedure proposed by Larson (1992). His method has the advantage of always finding the global maximum likelihood estimates. For a two-segment linear model, the method proceeds by fitting the model to the data assuming the changepoint is known. A model is fit using each observed covariate value as the known changepoint and the sum of squared errors (SSE) is computed for each model along with the left and right side derivatives of the SSE with respect to the changepoint. Larson (1992) shows that a necessary and sufficient condition for the SSE to achieve a local minimum between two adjacent covariate observations is that the right side derivative of the SSE is negative at the covariate on the left and the left side derivative is positive at the covariate on the right. For each pair of adjacent covariates where a local minimum exists, the minima is solved and another 2-segment linear model is fit using this location as the changepoint. Finally, the model providing the smallest sum of squared errors is chosen from among all the models fit.

To extend the method to an $m$-segment linear model, Larson (1992) proposes first finding the best 2-segment model; then, fixing this changepoint, use the algorithm to find the best second changepoint. The second changepoint is then fixed and the location of the first changepoint is re-estimated using the same algorithm. The process continues alternately fixing each of the 2 changepoints until convergence of the estimated changepoint locations suggests the best 3-segment model is found. Analogously, the method could then be used fixing 2 change points and finding the best 3rd for a four-segment model. Clearly, this could become computationally prohibitive if more than a few change points are required.

SAS/IML® code is available from the author. The code implements Larson’s (1992) estimation procedure for two-segment linear models. The user specifies for each of two treatment groups whether a 1 or 2-segment linear model is to be fit, the program then fits
the specified model to each group and plots the estimated difference function between the
groups with a simultaneous confidence band.

For very large segmented linear models where computation may be prohibitive using
Larson’s method, we recommend Tishler and Zang’s (1981) method. Gallant and Fuller
(1973) provide a specialized method which is applicable to segmented polynomial
models with continuous first derivatives. Julious (2001) presents a method which
appears to be equivalent to Larson’s for a two-segment linear model and discusses a
bootstrap test attributed to Hinkley (1988) for the presence of a change point.

The following practical suggestions may be helpful when approaching a problem where
our method is applicable. Our method assumes the number of segments to be modeled
for each treatment is known. In most practical applications, the number of segments will
have to be estimated from the data. We suggest that one first estimate the number of
segments required for each treatment by plotting the data with a non-parametric curve fit
and visually inspecting the plots. Next, rule out a simpler model by using Hinkley’s
(1988) bootstrap likelihood ratio test with the null model having one less segment than
visually estimated in the first step.

If both treatments have the same number of segments, then we want to rule out the
possibility that a simpler parallel segmented model applies. In a parallel segmented
model all parameters for the two treatments would be equal except for the intercepts, i.e.,
the treatment groups have the same segmented linear representation except one is
vertically shifted with respect to the other. Let \( \hat{\theta}^{(i)*} \) be the estimated parameter vector for
the \( i \)th group with the first parameter (corresponding to the intercept term) omitted.
Similarly, let \( \hat{\Sigma}^{*} \) be the estimated covariance matrix for the \( i \)th group where the first row
and column are omitted. We use the following test statistic to determine whether the
parallel segmented model applies:

\[
C = (\hat{\theta}^{(1)*} - \hat{\theta}^{(2)*})' \left( \hat{\Sigma}_{1}^{*} + \hat{\Sigma}_{2}^{*} \right)^{-1} (\hat{\theta}^{(1)*} - \hat{\theta}^{(2)*}) .
\]  

(4.5.1)

Under the null hypothesis that the parameters are equal and under reasonable regularity
conditions, \( C \) is distributed as \( \chi^2 \) with degrees of freedom equal to the number of elements
in \( \hat{\theta}^{(i)} \). If the null cannot be rejected, then the parallel segmented model should be fit with the two treatments having separate intercepts. The treatment comparison is then a test for equivalence of the two intercept terms. Depending on the application, it may also be appropriate to consider more complex similarities between the models for the two treatments. For example, an approach similar to the preceding could be used to test whether the treatments only differ in the slope of the final segment.

Given a reasonable basis for the number of segments to be modeled for each treatment and ruling out the possibility of a parallel segmented model, it would then be appropriate to apply our simultaneous confidence band approach for comparing two treatments.

### 4.6 Application

To illustrate the application of our method, we again use the data discussed in Section 3.6 from NSABP Protocol B-27 (Mamounas (1997)). Patients in Protocol B-27 are randomized into three groups. The control group (arm 1) receives 4 cycles of Adriamycin Cyclophosphamide (AC) chemotherapy given preoperatively along with 5 years of Tamoxifen beginning concurrently with AC. The two experimental groups receive the same AC chemotherapy and Tamoxifen followed by 4 cycles of Taxotere (T) given preoperatively (arm 2) or postoperatively (arm 3). Women with palpable operable invasive breast cancer confined to a single breast and possibly the ipsilateral (same side) axillary lymph nodes are eligible for the study, additional inclusion/exclusion criteria are described in Mamounas (1997).

In the present analysis, we are concerned with the change in tumor size between baseline (randomization) and surgery. We compare arms 1 and 3 combined, both of which received 4 cycles of AC prior to surgery with arm 2 where patients received 4 cycles of AC followed by 4 cycles of T prior to surgery. Our response variable is proportional reduction in tumor size defined as \((\text{baseline tumor size} - \text{tumor size at surgery}) / \text{baseline tumor size}\). Our analysis is restricted to patients whose baseline tumor size is at least 1.0 centimeter and for whom the tumor size at surgery is non-missing. A tumor size of zero at surgery is allowable and corresponds to a complete preoperative response to therapy; such patients will have a proportional reduction in tumor size value of 1.0. Our interest
concerns the effect of a patient’s age on her tumor response measured by its proportional reduction, and whether or not this effect is comparable between arms 1/3 and arm 2. Initial exploration with non-parametric curves suggests that 2-segment linear models are appropriate to describe the functional relationship between reduction in tumor size and patient age.

Figure 10 shows the fit of a 2-segment linear model to 1,117 observations from arms 1/3 (AC group). The estimated parameters are $\hat{\theta}^{(1)} = (0.5589, -0.0006514, 64.84, 0.01103)'$, so that $f(x, \hat{\theta}^{(1)}) = 0.5589 - 0.0006514 x + 0.01103 (x - 64.84) I_{[x > 64.84]}$. There is some issue of lack of normality of the errors since 14% of the observations have a proportional reduction in tumor size of 1.0 (complete response).

Figure 11 shows the fit of a 2-segment linear model to 512 observations from arm 2 (AC→Tax group). The estimated parameters are $\hat{\theta}^{(2)} = (1.127, -0.01078, 50.99, 0.01035)'$, so that $f(x, \hat{\theta}^{(2)}) = 1.127 - 0.01078 x + 0.01035 (x - 50.99) I_{[x > 50.99]}$. There is some issue of lack of normality of the errors since 28% of the observations have a proportional reduction in tumor size of 1.0 (complete response).
Figure 10: 2-segment linear model fit to AC group

Figure 11: 2-segment linear model fit to AC → Tax group
To compare the groups, we first calculate the difference of the estimated expected value functions depicted by the heavy 3-segmented line in Figure 12. The estimate of the variance-covariance matrix of the parameter estimates is:

\[
\hat{\Sigma} = \begin{bmatrix}
4.270E-3 & -8.785E-4 & -0.01293 & 8.785E-4 & 0 & 0 & 0 & 0 \\
-8.785E-4 & 1.868E-6 & 3.015E-3 & -1.868E-6 & 0 & 0 & 0 & 0 \\
-0.01293 & 3.015E-3 & 50.36 & 7.000E-2 & 0 & 0 & 0 & 0 \\
8.785E-4 & -1.868E-6 & 7.000E-2 & 2.114E-4 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 3.994E-2 & -9.306E-4 & -0.7262 & 9.306E-4 \\
0 & 0 & 0 & 0 & -9.306E-4 & 2.208E-5 & 1.887E-2 & -2.208E-5 \\
0 & 0 & 0 & 0 & -0.7262 & 1.887E-2 & 46.87 & 2.215E-3 \\
0 & 0 & 0 & 0 & 9.306E-4 & -2.208E-5 & 2.215E-3 & 5.378E-5 \\
\end{bmatrix}, (4.6.1)
\]

from which (by (4.3.10)) we can compute a 95% simultaneous confidence band for the population difference. This band is depicted by the light solid lines in Figure 12.

We can conclude the treatment groups are significantly different in tumor response for any interval of age where zero is not contained in the 95% simultaneous confidence band for the difference. Specifically, for patients between the ages of 26.2 and 44.8 years, the reduction in tumor size for patients treated with AC followed by preoperative Taxotere is
significantly greater (at the 0.05 level) than for patients treated with AC alone. The trend from the model would suggest that Taxotere may also be beneficial for patients less than 26.2 years of age, but the data were insufficient to demonstrate a significant difference for this age range.

4.7 Simulation of Coverage Probabilities

We simulated the coverage probabilities for a particular pair of 2-segment linear models with varying common error variances, sample sizes, and sets of observed covariates. For the first treatment group, the parameters are $\mathbf{\theta}^{(1)} = (0, 0, 6, 1)'$ so that $f(x, \mathbf{\theta}^{(1)}) = (x - 6) I_{[x>6]}$ and for the second treatment group $\mathbf{\theta}^{(2)} = (0, 0.25, 4, 1)'$ so that $f(x, \mathbf{\theta}^{(2)}) = 0.25 x + (x - 4) I_{[x>4]}$. Within each treatment group and for each sample size and error variance combination, we simulated 40 to 100 sets of covariate values (denoted in Table 12 as $k$: number of cases) from the model $x_{i1}, \ldots, x_{i_n} \sim \text{iid Uniform}(0, 10)$. Within each case, 500 to 1,000 sets of errors (denoted in Table 12 as $n$: number of iterations per case) were simulated and responses calculated from the model.

The estimated non-coverage probability varies depending on the particular set of covariates observed which implies that a simple Binomial model would underestimate the variance of the estimated non-coverage proportion for the case of random covariates. To accommodate the variation of the underlying true proportion, we assume the number of true mean functions not covered by the simultaneous confidence band out of $n$ iterations within a case follows a Beta-Binomial distribution. Here, the Binomial parameters are $n$ and $p$, but the parameter $p$ varies from case to case depending on the particular set of covariates observed. The Binomial parameter $p$ is distributed as $\text{Beta}[\mu(r-1)/r, (\mu-1)(r-1)/r]$ where $\mu$ and $r$ are unknown, so that $E(p) = \mu$ and $\text{Var}(p) = r \mu(1 - \mu)$. Kleinman’s (1973) method of moments estimators are used to estimate the average non-coverage probability $\mu$ as well as the parameter $r$ when the covariates follow the specified model along with an estimate of the standard error.
Let $\hat{p}_1, \ldots, \hat{p}_k$ denote the estimated non-coverage probabilities from each of the $k$ cases, where $\hat{p}_i$ is based upon $n$ iterations within the case. Then:

$$\hat{\mu} = \frac{1}{k} \sum_{i=1}^{k} \hat{p}_i \quad \text{and} \quad \widehat{\text{Var}}(\hat{\mu}) = \frac{\hat{\mu}(1 - \hat{\mu})(1 + (n - 1) \hat{r})}{nk},$$

(4.7.1)

where $n$ is the number of iterations per case and

$$\hat{r} = \frac{1}{n - 1} \left( \frac{n \sum_{i=1}^{k} (\hat{p}_i - \hat{\mu})^2}{k \hat{\mu}(1 - \hat{\mu})} - 1 \right).$$

The choices for the common sample size were $n_1 = n_2 = 10, 20, 50, 80, 150, 250, \text{ and } 500$ and for the common error variance $\sigma_1 = \sigma_2 = 0.1, 0.5, \text{ and } 1.0$. Limited combinations of these sample sizes and variances were employed. The nominal asymptotic coverage probability of our bands was 95%, and we assumed a common error variance. When the average non-coverage probability to be estimated is small, say less than 0.05, a larger number of iterations per case ($n = 1,000$) is required to reliably estimate the proportion. When the number of iterations per case is smaller ($n = 500$ or 800), we increase the number of cases ($k$) to reduce the standard error of the estimated average non-coverage probability. The detailed results are given in Table 12.

**Table 12: Empirical Non-Coverage Probabilities with Covariates Uniform on (0, 10)**

<table>
<thead>
<tr>
<th>$n_1 = n_2$</th>
<th>$\sigma_1 = \sigma_2$</th>
<th>Number of Cases ($k$)</th>
<th>Number of Iterations per Case ($n$)</th>
<th>Estimated Non-Coverage Probability ($\hat{\mu}$)</th>
<th>Standard Error of the Estimate $\sqrt{\text{Var}(\hat{\mu})}$</th>
<th>($\hat{r}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.1</td>
<td>100</td>
<td>500</td>
<td>0.1933</td>
<td>0.0243</td>
<td>3.771E-1</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>70</td>
<td>800</td>
<td>0.0886</td>
<td>0.0046</td>
<td>1.725E-2</td>
</tr>
<tr>
<td>50</td>
<td>1.0</td>
<td>40</td>
<td>1,000</td>
<td>0.0742</td>
<td>0.0024</td>
<td>2.384E-3</td>
</tr>
<tr>
<td>80</td>
<td>1.0</td>
<td>40</td>
<td>1,000</td>
<td>0.0507</td>
<td>0.0020</td>
<td>2.393E-3</td>
</tr>
<tr>
<td>150</td>
<td>1.0</td>
<td>40</td>
<td>1,000</td>
<td>0.0343</td>
<td>0.0013</td>
<td>1.063E-3</td>
</tr>
<tr>
<td>250</td>
<td>1.0</td>
<td>40</td>
<td>1,000</td>
<td>0.0228</td>
<td>0.0013</td>
<td>1.846E-3</td>
</tr>
<tr>
<td>500</td>
<td>1.0</td>
<td>40</td>
<td>1,000</td>
<td>0.0150</td>
<td>0.0006</td>
<td>1.065E-4</td>
</tr>
</tbody>
</table>
The poor coverage and high standard error observed for the sample size of 10 and, to a lesser extent, 20 prompted further investigation. Some cases, i.e., sets of covariates, produced non-coverage estimates \( (\hat{p}) \) of 0.9 to 1.0 while the majority of cases produced non-coverage estimates more in line with those observed for larger samples. Inspection of the covariates of the extreme cases revealed that all or all but one of the covariates fell on one segment of the underlying 2-segment model. This explains the poor coverage since we had essentially no data to estimate the location of the mean function on one of the segments. It is not surprising that our method breaks down when the data are insufficient to estimate the underlying model parameters.

We ran additional simulations with fixed covariates equally spaced on (0, 10), here \( x_{ij} = 10j / (n_i + 1), j = 1, \ldots, n_i \). The parameters of the underlying models, variances, and sample sizes were the same as the previous simulations. With a fixed set of covariates, the true non-coverage probability is also fixed so that a Binomial model is appropriate for estimation. We performed 10,000 iterations for each sample size/variance combination. The detailed results are given in Table 13.

**Table 13: Empirical Non-Coverage Probabilities with Covariates Fixed and Equally Spaced**

<table>
<thead>
<tr>
<th>( n_1 = n_2 )</th>
<th>( \sigma_1 = \sigma_2 )</th>
<th>Number of Iterations (n)</th>
<th>Estimated Non-Coverage Probability ( (\hat{p}) )</th>
<th>Standard Error of the Estimate ( \sqrt{\hat{p}(1-\hat{p})/n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.1</td>
<td>10,000</td>
<td>0.0376</td>
<td>0.0019</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>10,000</td>
<td>0.0611</td>
<td>0.0024</td>
</tr>
<tr>
<td>50</td>
<td>1.0</td>
<td>10,000</td>
<td>0.0635</td>
<td>0.0024</td>
</tr>
<tr>
<td>80</td>
<td>1.0</td>
<td>10,000</td>
<td>0.0495</td>
<td>0.0021</td>
</tr>
<tr>
<td>150</td>
<td>1.0</td>
<td>10,000</td>
<td>0.0313</td>
<td>0.0017</td>
</tr>
<tr>
<td>250</td>
<td>1.0</td>
<td>10,000</td>
<td>0.0221</td>
<td>0.0015</td>
</tr>
<tr>
<td>500</td>
<td>1.0</td>
<td>10,000</td>
<td>0.0149</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

The estimated non-coverage probabilities in Table 13 are close to the nominal non-coverage probability of 0.05 for small sample sizes and seem to decrease for larger
sample sizes. Even though our procedure is based on asymptotic distribution theory, it seems to behave well for small to moderate sample sizes.

For the particular pair of models used in the simulation, the simultaneous confidence bands become increasingly conservative as the sample size increases. There are two interrelated factors which may help to explain this behavior. First, the inequality used to derive the simultaneous bounds is not sharp, and, in the simulation, non-coverage is checked only for the range of covariates in the observed sample. Second, the \( \chi^2 \) multiplier used to achieve simultaneous coverage is appropriate if we assume the covariance matrix is known. A critical value from the \( F \) distribution might be more appropriate for an unknown covariance matrix which has to be estimated from the data. Thus, if the critical value were appropriate to the sample size, we would expect to see conservative coverage, something that is apparent for the larger sample sizes. However, for smaller sample sizes, with a liberal critical value, the two factors seem to negate each other and coverage is reasonably close to the true level.

### 4.8 Discussion

Another approach to the problem of constructing simultaneous confidence bands about the difference of segmented linear models might be to naively apply the Working-Hotelling bands (for example, Neter et al. (1996, p. 234) for a single regression line or Potthoff (1964) for the difference of two regression lines) to each segment of the difference with a Bonferroni correction for the confidence level. We compare our technique with naïve segmented Working-Hotelling bands for the case of unknown change points in Section 4.8.1 and for the case of known changepoints in Section 4.8.2. In Section 4.8.3, we consider the difficulties presented when our method identifies multiple disjoint intervals of significant difference between the treatments.

#### 4.8.1 Comparison to Naïve segmented Working-Hotelling band: Unknown change points

Using simulated data, we illustrate the difference between the Working-Hotelling approach and our simultaneous confidence band. The data are simulated for a particular pair of 2-segment linear models as follows: for the first treatment group, \( \theta^{(1)} = (0, 0.1, \)
6.5, 0.75)’ and for the second treatment group \( \theta^{(2)} = (-0.125, 0.15, 3.5, 0.75)' \). Within each treatment group, we simulated the covariate values from the model \( x_1, \ldots, x_{n_i} \sim \text{Uniform}(0, 10) \). Each treatment group used a common sample size of \( n_1 = n_2 = 80 \) and common error variance \( \sigma_1 = \sigma_2 = 1.0 \). The pooled estimate of the common variance was used. The nominal asymptotic coverage probability of our bands is 95%.

Noting that the estimated difference function is a three-segment linear function, we construct the naïve Working-Hotelling bands by allocating \( \alpha \) among the three segments. Since we expect roughly equal numbers of observations on each of the three segments, we have used the Bonferroni procedure to allocate \( \alpha \) so that, for each segment, the Working-Hotelling bands will have coverage probability \( (1 - \alpha / 3) \).

To construct the naïve Working-Hotelling bands, we first fit the segmented linear models to estimate the parameters. Next, we use the estimated changepoints from the two models to define the 3 segments of the difference function. From Potthoff (1964), the Working-Hotelling simultaneous confidence band for the difference when the covariate \( x \) is in segment \( j \) is:

\[
d(x, \hat{\theta}) \pm \sqrt{2F_{2, n_{ij} + n_{kj} - 4, 1 - \alpha/3} \hat{\sigma}^2 \sum_{i=1}^{2} \left( \frac{1}{n_{ij}} + \frac{(x - \bar{x}_{ij})^2}{\sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij})^2} \right)},
\]

where \( F_{n_1, n_2, p} \) is the \( p^{th} \) quantile of the F distribution on \( n_1 \) and \( n_2 \) degrees of freedom, \( x_{ijk} \) is the value of the \( k^{th} \) covariate in segment \( j \) from treatment \( i \), \( n_{ij} \) is the number of covariates in segment \( j \) from treatment \( i \), and \( \bar{x}_{ij} = \frac{\sum_{k=1}^{n_{ij}} x_{ijk}}{n_{ij}} \).

The construction of the naïve Working-Hotelling bands outlined in the previous paragraph takes the estimated changepoints and treats them as known. For this reason, it is not readily apparent that the naïve Working-Hotelling bands actually constitute a 95% simultaneous confidence band.

Figure 13 and Figure 14 show the 2-segment models fit to the data. Figure 15 shows our simultaneous confidence band for the data, and Figure 16 shows the comparison between
our simultaneous confidence band and a naïve application of the Working Hotelling band.

Figure 13: 2-Segment Linear Fit to Treatment Group 1 Data
Solid line – estimated mean function

Figure 14: 2-Segment Linear Fit to Treatment Group 2 Data
Solid line – estimated mean function
Figure 15: Simultaneous Confidence Interval for the Difference (Group 2 minus Group 1)

Center solid line – estimated mean function for the difference
Center dashed line – true mean function for the difference
Outer solid lines – our simultaneous 95% confidence band for the difference

Figure 16: Comparison between Naïve Working Hotelling and Our Simultaneous Confidence Bands

Center solid line – estimated mean function for the difference
Outer solid lines – our simultaneous 95% confidence band for the difference
Outer dashed lines – naïve simultaneous 95% Working-Hotelling bands for the difference
While the naïve Working-Hotelling bands appear to be narrower in certain regions than our simultaneous band, our band seems to be narrower near the change points. The explanation for this latter phenomenon would be that, for our bands, the continuity between adjacent segments is taken into account, but the naïve Working-Hotelling bands are fit individually to the segments without accounting for continuity.

Since we cannot establish that the naïve Working-Hotelling bands actually constitute a simultaneous band for the case of unknown changepoints, we cannot at this time recommend their use.

4.8.2 Comparison to Naïve segmented Working-Hotelling band: Known change points

We construct an example, using simulated data, to illustrate the difference between the Working-Hotelling approach and our simultaneous confidence bands for the case of known changepoints. The data are simulated for a particular pair of 2-segment linear models as follows: for the first treatment group; \( \mathbf{\theta}^{(1)} = (0, 0.1, 0.75)' \), \( \gamma_2^{(1)} = 6.5 \) and for the second treatment group \( \mathbf{\theta}^{(2)} = (-0.125, 0.15, 0.75)' \), \( \gamma_2^{(2)} = 3.5 \). Within each treatment group, we simulated the covariate values from the model \( x_{i1}, \ldots, x_{in} \sim \text{Uniform}(0, 10) \).

Each treatment group used a common sample size of \( n_1 = n_2 = 80 \) and common error variance \( \sigma_1 = \sigma_2 = 1.0 \). The pooled estimate of the common variance was used. The nominal asymptotic coverage probability of our bands is 95%.

Construction of the naïve Working-Hotelling bands is the same as the previous section. For the case of known changepoints considered here, the naïve Working-Hotelling bands should actually constitute a 95% simultaneous confidence band. The band may be more conservative than necessary since the construction procedure assumes 4 parameters were estimated for each segment (for a total of 12) when, in fact, only 6 parameters are necessary to describe the 3-segment difference function.

Figure 17 shows the comparison between our simultaneous confidence band and a naïve application of the Working-Hotelling band.
Figure 17: Comparison between Naïve Working Hotelling and Our Simultaneous Confidence Bands

Center solid line – estimated mean function for the difference
Outer solid lines – our simultaneous 95% confidence band for the difference
Outer dashed lines – naïve simultaneous 95% Working-Hotelling bands for the difference

For the case of known change points, we see that our simultaneous band appears to be quite a lot narrower on average than the naïve Working-Hotelling bands. Our band is much narrower near the change points. The explanation for this would be that, for our band, the continuity between adjacent segments is taken into account, but the Working-Hotelling bands are fit individually to the segments without accounting for continuity.

Our procedure produces continuous confidence bands when the change points are known. The decreased width of our bands when the change points are known is primarily due to the reduction (from 6 in Section 4.8.1 to 4 in Section 4.8.2) in the degrees of freedom for the chi-squared multiplier.

While the naïve Working-Hotelling bands do actually constitute a simultaneous band for the case of known changepoints, their inability to properly account for the continuity between adjacent segments and their greater width in general would suggest that, in general, the naïve Working-Hotelling bands are inferior to the simultaneous confidence bands produced by our method.
4.8.3 Difficulties Presented by Multiple Intervals of Significant Difference

For intervals of the covariate where zero is contained in the simultaneous confidence bands for the difference of the segmented regressions we conclude there is no difference between the two groups. Likewise, for intervals of the covariate where the simultaneous confidence bands lie entirely above or below zero we conclude there is a significant difference between the groups.

Practical difficulties may arise since it is possible to find multiple disjoint intervals of the covariate where the groups are significantly different using our method. When such issues arise, the details of the application should probably dictate interpretation.
5 Comparison between Simultaneous Confidence Band for the Difference of Segmented Linear Models and Inference Guided Data Exploration by ANCOVA on Half-Intervals of a Covariate

5.1 Introduction

We consider here two previously introduced methods of approaching the problem of comparing two treatments when the mean value of the response depends on the value of the covariate and the treatments differ in mean response for only part of the range of the covariates studied. The methods applicable to this problem are our simultaneous confidence bands developed for the difference of segmented linear models and our inference guided data exploration using ANCOVA on half-intervals of a covariate. Our interest lies in comparing the methods on several measures of performance including; (i) coverage, which is the probability that the identified region(s) of significant difference is contained in the true region where the groups differ; (ii) the mean squared error (MSE) and (iii) bias of the estimated threshold value of the covariate defining the true region of significant difference.

5.1.1 Brief explanation of Simultaneous Confidence Band for the Difference of Segmented Linear Models

The two treatment groups are modeled separately by \( m \)-segment linear models in this method; we restrict our attention to 2-segment models in this chapter. A simultaneous confidence band for the difference of the group means as a function of the covariate is constructed. Treatments are found to be significantly different for values of the covariate where zero is not contained in the simultaneous confidence band. When the difference is defined as \( E(\text{Treatment} \mid x) – E(\text{Control} \mid x) \), regions of the covariate where the upper confidence band is negative imply that the control group is superior and regions where the lower confidence band is positive imply the treatment group is superior. This method has the potential of identifying multiple, possibly disjoint, regions of significant difference between the treatments. A more detailed development of this method can be found in Chapter 4. For the sake of brevity we may later refer to this method as simultaneous confidence bands or simply SCB.
5.1.2 Brief explanation of Inference Guided Data Exploration by ANCOVA on Half-Intervals of a Covariate

In this method, parallel slopes ANCOVA models are fit to subsets of the data defined by choosing all observations where the covariate is at least as large as (no smaller than) a series of increasing (decreasing) threshold values (in this chapter we restrict our attention to the case of increasing threshold values). Order statistics from the covariates of the two groups combined are used as successive threshold values beginning with the first order statistic and ending when one of the treatments reaches a minimum sample size criteria or when some other stopping criteria is triggered. We use the term “obtained p-value” to refer to the minimum p-value observed when applying the method to a set of data. Type I error is controlled by comparing the obtained p-value to a critical p-value generated by simulation from the null distribution using the same sample sizes for treatment and control and the same minimum sample size criteria. When the obtained p-value is less than or equal to the critical p-value, then either the treatment or control group is declared superior based on the sign of the coefficient of the treatment indicator in the ANCOVA model. When the obtained p-value is greater than the critical p-value we conclude the null hypothesis is true and there is no difference between treatments. The obtained p-value is not a true p-value, but rather a test statistic. The critical p-value is the $a^{th}$ quantile of the empirical distribution of the obtained p-value under the null hypothesis. See Sections 3.5 and 3.5.3 for a more detailed description of the method.

For each threshold value, the ANCOVA model is fit and a p-value is calculated for the difference between groups. We need a rule to choose one obtained p-value for the method from among all the p-values associated with the various threshold values. We consider two such rules in Sections 5.1.2.1 and 5.1.2.2 which we refer to as stopping criteria. The term half-intervals or HI may be used to refer to this method generically when the stopping rule is not important.

A visual example of the method of half-intervals is presented in Figure 18 based on the data in Table 14. Both treatment groups have samples of size 4 and we use a minimum sample size criteria of 2 so that the critical p-value from Table 10 is 0.0169 (since the
critical p-values do not depend on $\rho$ for this model, this is the average over all Table 10 values of $\rho$ for the entries with sample sizes $n_t = 4$, $n_c = 4$, and $n_{\text{min}} = 2$).
Table 14: Data for method of half-intervals example

<table>
<thead>
<tr>
<th></th>
<th>Control Group</th>
<th>Treatment Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>2 4 6 8</td>
<td>1 3 5 7</td>
</tr>
<tr>
<td>Y</td>
<td>-0.05 0.2 0.4</td>
<td>-0.4 0.6 1.3 2.3</td>
</tr>
</tbody>
</table>

Figure 18: Method of half-intervals, critical p-value = 0.0169, minimum sample size = 2, panels 1-5
5.1.2.1 Stopping Criteria 1, Half-Intervals Minimum p-value

Under this stopping criteria, every subset is tested until the minimum sample size criteria is reached. We then choose the threshold value which produced the minimum p-value observed for any subset and evaluate significance based on the critical p-value. We refer to this p-value as the obtained p-value. The motivation for this stopping criteria is that this threshold value in some sense defines the most “likely” region for a difference in treatments since this threshold produced the most extreme p-value which implies that it was least “likely” to come from the null hypothesis of no treatment difference. Under this stopping rule we may later refer to this method as half-intervals minimum p-value or simply HIM.

Using this stopping criteria, all tests depicted in Figure 18 would be performed and we would choose the test depicted in panel 3 which produced the minimum p-value (0.015). Since the obtained p-value (0.015) is less than the critical p-value (0.0169), we find a significant difference between treatments. The associated threshold value of the covariate is \( x = 3 \) (Table 14).

5.1.2.2 Stopping Criteria 2, Half-Intervals First Significant p-value

For this criteria, we stop the first time the observed p-value is less than the critical p-value. The threshold value associated with this p-value will be used as our threshold estimate, and we refer to this p-value as the obtained p-value. The motivation for this stopping criteria is that, since our threshold values begin with the first order statistic (smallest) of the combined covariates and move successively through the order statistics, this criteria will produce the largest possible region of the covariate where the treatments are found to be significantly different. If the minimum sample size criteria is reached without stopping, then we conclude there is no difference between the treatments, and the obtained p-value is the minimum observed p-value. We may later refer to this method as half-intervals first significance or HIF.

In the example described by Table 14 and Figure 18, this stopping criteria would lead us to stop in panel 2 since the obtained p-value (0.0161) is less than the critical p-value (0.0169). The associated threshold value is \( x = 2 \).
5.1.3 Discussion of similarities and differences

The method of half-intervals assumes the underlying model is linear with parallel slopes for the treatments on the region of the covariate where the treatments differ. The method of simultaneous confidence bands can accommodate parallel slopes but is also well suited to the case where the slopes are not parallel. We employ two simulation models where the assumption of parallel slopes for the two treatments holds and one simulation model where this assumption does not hold in order to contrast this difference between the methods (Figure 19). The simultaneous confidence bands method assumes the expected value for each treatment group is continuous over the range of the observed covariates in the underlying model. The method of half-intervals can accommodate discontinuity by considering only a subset of the data; however continuity is necessary within the subset under consideration. Two simulation models where continuity does hold and one where it does not hold are employed to contrast this difference between the methods.

Figure 19: $E(Y|X)$ under simulation models 1, 2, and 3
5.1.4 Simple models to illuminate the relative strengths of the methods

The first simulation model considered has each treatment following a 2-segment linear model. The models for each treatment will share common slopes for both segments and a common intercept for the first segment; the models will differ only in the location of the changepoint. For simulation model 1, continuity of the mean response is satisfied and the parallel slopes ANCOVA model roughly holds on the region of the covariate where the treatments differ in mean response. The second simulation model has the treatment group following a 2-segment linear model and the control group following a simple linear model. The first segment of the treatment model shares a common slope and intercept with the linear model of the control group. Under simulation model 2, continuity of the mean response is satisfied and the assumptions of the parallel slopes ANCOVA model are violated on the region of the covariate where the treatments differ in mean response. The third simulation model we consider has the control group following a simple linear model and the treatment group having a constant slope but with a step in the response. For covariate values less than the location of the step, both models share a common slope and intercept. To the right of the step, both models share a common slope. For simulation model 3, continuity of the mean response is violated, but the assumptions of the parallel slopes ANCOVA model hold on the region of the covariate where the treatments differ in mean response. These three simulation models are defined more rigorously in Section 5.3 where we describe our simulations.

5.2 Measures of Performance

We will compare the SCB and HI methods using two criteria; coverage of the identified region(s) of significant difference by the true region of significant difference and the mean squared error (MSE) and bias of the estimated threshold value of the covariate where the groups begin to differ. The threshold estimate is defined first to facilitate the definition of coverage.

5.2.1 Estimation of the True Threshold

All of the simulation models which we will consider are constructed so that, for all covariate values less than the true threshold value \(x_0\), the mean response for both groups
is the same and, for all covariate values greater than or equal to the threshold, the mean responses differ. The methods will be compared on their ability to estimate the true threshold value by the MSE and bias of the threshold estimators when the estimator exists. The estimator exists whenever the method finds a significant difference between the treatments.

The natural estimators of the true threshold for the method of half-intervals are as follow. For the HIM stopping criteria, the threshold estimate is the covariate value used as the threshold for the comparison which yielded the minimum p-value, provided it is less than or equal to the critical p-value. For the HIF stopping criteria, it is the covariate value used as the threshold for the comparison which yielded the first sub-critical p-value. If the obtained p-value for either stopping criteria exceeds the critical p-value, then no significant difference is found between the treatments and, consequently, there is no threshold estimate. These estimators are consistent with the true threshold in the sense that the method of half-intervals has found a significant difference between the treatments while considering the subset of observations where the covariate is greater than or equal to the threshold estimate.

The method of simultaneous confidence bands does not provide a natural estimator for the threshold since there may be multiple disjoint intervals of significant difference identified. In constructing a threshold estimate for this method we proceed as follows, where we provide in Figure 20 a visual example. This method may produce multiple disjoint intervals of the covariate where the treatments are found to be significantly different. All such intervals have either the upper boundary of the simultaneous confidence band less than zero or the lower boundary greater than zero. Our strategy is to trace the upper and lower boundaries of the simultaneous confidence band to find the smallest covariate value where one of these conditions is satisfied and declare this covariate value our threshold estimate. Following this strategy, all regions of significant difference will be contained in the region to the right of the threshold estimate, inclusive. When no region of significant difference between the treatments is identified (zero difference lies between the confidence bands for all values of the covariate), there is no threshold estimate. In practice, the trace of the boundaries of the simultaneous
confidence band requires a grid search, so we will restrict the search to the range of values of the observed covariates. Let $u(x)$ and $l(x)$ define the upper and lower boundaries of the simultaneous confidence bands respectively, then the threshold estimate is:

$$\hat{x}_0 = \inf_{x \in (A,B)} \left\{ x : u(x) < 0 \text{ or } l(x) > 0 \right\},$$

(5.2.1)

where $(A, B)$ is the range of the observed covariates.

Figure 20: Example for SCB, identified region is covered by true region

Since the width of the simultaneous confidence bands increases dramatically outside the range of the observed covariates, it would be quite rare, but not impossible, for a region of significance to occur entirely outside the range of our search. When a region of significant difference overlaps the lower boundary of the range of our search, we introduce upward bias in our threshold estimate but coverage is unaffected. When a region of significant difference overlaps the upper boundary of the range of our search, neither coverage nor the threshold estimate is affected.

5.2.2 Coverage by the True Region were the Groups Differ

Our simulation models are constructed so that the true region where the groups differ is $(x_0, \infty)$ and the treatment group is superior to control on this region. The methods are evaluated based on the proportion of identified regions of significant difference found by
the method which are contained in the true region. If no region of significant difference is identified by the method, then appropriately we use the convention that the identified region is covered by the true region. If the true region where the treatments differ does not cover the region(s) of significant difference identified by the method, then there are values of the covariate where the method claims the treatments differ significantly when in fact there is no difference.

The definition of coverage for both the simultaneous confidence bands and half-intervals methods depends on the associated threshold estimates as follows. The existence of a threshold estimate, as defined in Section 5.2.1, implies there exist one or more regions of significant difference and that all covariate values in these regions are at least as large as the threshold estimate. If the threshold estimate exists, then, provided the threshold estimate is greater than or equal to the true threshold value, the identified region(s) is covered. If no threshold estimate was identified, then no region of significant difference is identified and, by convention, the region is covered.

5.3 Simulations

We consider three different simulation models to evaluate the relative performance of our methods and stopping criteria. Rather than fixing the standard deviation and varying the sample sizes to illustrate the relative performance of the methods, we have decided to fix the sample sizes and vary the standard deviation. Since increasing sample size should be roughly equivalent to decreasing standard deviation, we feel this is a reasonable compromise as simulations for small standard deviation are much more practical than simulations for large sample size. We use a sample size of 50 for each treatment and consider various values of the standard deviation ranging from 0.1 to 10.0.

5.3.1 Simulation Model 1: Segmented lines with different changepoints, otherwise parallel

Both treatment groups follow a 2-segment linear model with all parameters identical except for the locations of the changepoints. This simulation model is perfectly suited to the simultaneous confidence band (SCB) method. For covariate values greater than the second changepoint the parallel slopes ANCOVA model is appropriate so that the half-
intervals (HI) method is also relatively well suited. Simulation model 1 is given by:

\[\begin{align*}
Y_{\text{Control}} &= (X - \alpha_1) \beta I_{[X > \alpha_1]} + \epsilon \\
Y_{\text{Treatment}} &= (X - \alpha_2) \beta I_{[X > \alpha_2]} + \epsilon
\end{align*}\]  

(5.3.1)

where \(\alpha_1 = 6\), \(\alpha_2 = x_0 = 4\), \(\beta = 1\), and \(\epsilon\) is distributed as \(N(0, \sigma^2)\). The model is illustrated in Figure 21.

Both SCB and HI appear to be applicable to the setting of model 1.

5.3.2 Simulation Model 2: Treatment segmented, control single line

The control group follows a simple linear model and the treatment group follows a 2-segment linear model. The 2-segment linear models of the SCB method will fit the treatment group well, but the control group will be over fit since we will be fitting a 2-segment line to data generated from a simple linear model. This simulation model is ill suited to the half-intervals method as the parallel slopes assumption does not hold on the region of the covariate where the treatments differ. Simulation model 2 is given by:

\[\begin{align*}
Y_{\text{Control}} &= \epsilon \\
Y_{\text{Treatment}} &= (X - \alpha) \beta I_{[X > \alpha]} + \epsilon
\end{align*}\]  

(5.3.2)

where \(\alpha = x_0 = 5\), \(\beta = 1\), and \(\epsilon\) is distributed as \(N(0, \sigma^2)\). Figure 22 illustrates the model.
For model 2, SCB seems applicable while HI seems less so.

5.3.3 Simulation Model 3: Step function for treatment, control single line

Here, the control group follows a simple linear model and the treatment group has a constant slope but is discontinuous at the point of the step. This simulation model is not well suited to modeling with the simultaneous confidence band method since the assumption of continuity is violated for the treatment group. Like simulation model 1, the parallel slopes ANCOVA model will fit the data well for covariate values greater than the location of the step in the treatment group. Simulation model 3 is given by:

\[
\begin{align*}
Y_{\text{Control}} &= \varepsilon \\
Y_{\text{Treatment}} &= \beta I_{|X > x_0|} + \varepsilon
\end{align*}
\]

(5.3.3)

where \(\alpha = x_0 = 5\), \(\beta = 1\), and \(\varepsilon\) is distributed as \(N(0, \sigma^2)\). Figure 23 illustrates the model.

For model 3, HI seems applicable while SCB seems less so.
5.3.4 Minimum sample size and critical p-value used for half-intervals method

In order to employ the method of half-intervals, we need a minimum sample size criteria and a critical p-value based on simulation from the null hypothesis of no difference between treatments. A minimum sample size criteria of 20 for each treatment was used in all applications of the half-intervals method (i.e., in estimating the critical p-value as well as later simulations to compare methods). The critical p-value is the 0.05 quantile of the empirical distribution of the obtained p-value based on 30,000 simulated iterations. In Section 3.5.3, we found that the critical p-value for the method of half-intervals by ANCOVA did not depend on the correlation so we will use independent responses and covariates. For each iteration, sets of 50 independent covariates are generated for each of treatment and control from Uniform(0, 10). Independent error terms are generated from $N(0, \sigma^2)$ and the responses are set equal to the errors for both treatment and control so that the model is linear with zero slope and intercept. Finally, the HIM stopping criteria is applied to the data. The 0.05 quantile of the obtained p-values from these 30,000 simulations is 0.0123 and is used as the critical p-value in all later applications of the method of half-intervals for either stopping criteria. This value is consistent, up to the sampling variation of the estimated critical p-value, with the estimates in Table 10 corresponding to sample sizes of 50 for each treatment and a minimum sample size criteria of 20. In fact, even the distribution of the covariate apparently has no influence on the critical p-value since the ANCOVA analysis is done conditional on the observed covariates (Neter et al. (1996), p. 1017).

5.3.5 Simulation results

A total of 10,000 iterations were performed for each combination of simulation model and standard deviation. For each iteration, sets of 50 independent covariates were generated for each of treatment and control from Uniform(0, 10). The expected value of the response was calculated based on the simulation model for the specified alternative. Independent error terms were generated from $N(0, \sigma^2)$, for selected values of the standard deviation, and added to the expected values to give the sets of responses. Finally, each of the methods, SCB, HIM, and HIF, were applied to the data. Both the half-intervals and
simultaneous confidence band methods assumed a common standard deviation for both treatment groups.

Table 15 gives the detailed results of the simulations including coverage by the true interval, MSE of the threshold estimate, and bias of the threshold estimate. Figure 24 through Figure 29 use the data in Table 15 to compare the methods on coverage by the true interval and MSE of the threshold estimate with separate comparisons for each simulation model. Figure 24, for example, plots the coverage (labeled “Cover” in Table 15) versus the standard deviation for each of the three methods (SCB, HIM, and HIF) based on simulation model 1.
Table 15: Performance of the methods under various simulation models and standard deviations

<table>
<thead>
<tr>
<th>Simulation model</th>
<th>Method</th>
<th>Performance Measure</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cover</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>0.149</td>
</tr>
<tr>
<td></td>
<td>SCB</td>
<td>Cover</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>1.901</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>-1.226</td>
</tr>
<tr>
<td></td>
<td>HIM</td>
<td>Cover</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>HIF</td>
<td>Cover</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>SCB</td>
<td>Cover</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>-2.060</td>
</tr>
<tr>
<td></td>
<td>HIM</td>
<td>Cover</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td>5.165</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>-1.791</td>
</tr>
<tr>
<td></td>
<td>HIF</td>
<td>Cover</td>
<td>0.000</td>
</tr>
</tbody>
</table>

103
Figure 24: Coverage by the true interval under simulation model 1

Figure 25: MSE of threshold estimator under simulation model 1
Figure 26: Coverage by the true interval under simulation model 2

Figure 27: MSE of threshold estimator under simulation model 2
Figure 28: Coverage by the true interval under simulation model 3

Figure 29: MSE of threshold estimator under simulation model 3
5.4 Discussion

Within each simulation model, the performance measures for each method were estimated for various values of the standard deviation with all other parameters (including sample sizes) fixed. The effect of decreasing standard deviation should be roughly equivalent to the effect of increasing sample sizes. Each performance measure will be examined in detail focusing on the differences between the methods. In the final section (5.4.4), we discuss the relative merits of the methods considered and describe how the priorities of the application influence the choice of a method.

5.4.1 Coverage by the True Region where the Groups Differ

The most striking result from examining the coverage performance of the methods under the various simulation models is the almost universal superiority of the method of simultaneous confidence bands. The results for coverage are given in Figure 24, Figure 26, and Figure 28 for simulation models 1, 2, and 3, respectively. The method of half-intervals with the minimum p-value stopping criteria outperformed the SCB method only on simulation model 3 and only for a single value of standard deviation ($\sigma = 0.3$).

Comparing the two stopping criteria for the method of Half Intervals, we see that the minimum p-value criteria universally outperforms the first significant p-value criteria.

The method of simultaneous confidence bands produced coverage in excess of $0.95 (= 1 - \alpha)$ for all values of standard deviation under simulation models 1 and 2 where the underlying model was consistent with the assumptions of the method. For simulation model 3 (step function for treatment), where the assumption of continuity in the underlying model was violated, we see that the coverage dropped off dramatically for the smallest values of standard deviation. With small values of the standard deviation (or large sample sizes), the width of the simultaneous confidence band becomes narrow and, since the underlying model is inconsistent with the model assumed by the method, the coverage tends towards zero. The design of the method of SCB is such that it should provide coverage in excess of $1 - \alpha$ when the assumptions of the method hold, but we see that this level of coverage may not be achieved when the underlying model is inconsistent with the assumptions of the method.
5.4.2 Estimation of the True Threshold

The clearest difference observed between methods based upon the mean squared error performance measure was the difference between the two stopping criteria for the half-intervals method. The MSE of the threshold estimate is plotted against standard deviation for each method/stopping criteria in Figure 25, Figure 27, and Figure 29 for simulation models 1, 2, and 3, respectively. The most significant p-value stopping criteria is superior to the first significant p-value stopping criteria for all combinations of simulation models and standard deviations examined.

For all three simulation models and for all but the smallest values of standard deviation, the method of half-intervals with the most significant p-value stopping criteria proved to be a superior estimator (as measured by MSE) of the true threshold value when compared to the method of simultaneous confidence bands. The SCB method is superior to the method of HIM for the smallest values of standard deviation considered for simulation models 1 and 2 and, by extension, would appear to be superior for very large sample sizes. Under simulation models 1 and 2 where the underlying model is consistent with the assumptions for the method of SCB, it seems that the limiting value of MSE for the method of SCB is zero as the standard deviation approaches zero. This is consistent with the fact that the threshold estimator for SCB should be asymptotically approximately unbiased when the assumptions of the model are satisfied.

5.4.3 Relationship between bias, coverage, and MSE

There is a relationship between the bias of the threshold estimate (Table 15) and the coverage which helps to explain the coverage results in Section 5.4.1. The identified region of significant difference can only be covered by the true region if the threshold estimate is greater than or equal to the true threshold. If the threshold estimate is positively biased, then, on average, more of the identified regions will tend to be covered by the true region than if it is unbiased or negatively biased. This explains the coverage difference for the stopping criteria for the half-intervals method since, for HIF, we stop at the first threshold value providing a significant p-value while, for HIM, we continue examining larger threshold values in hopes of finding a smaller p-value. The HIF method
inherently has smaller bias than the HIM method and hence, poorer coverage. This relationship also explains the superior coverage performance observed for the SCB method, since the SCB method generally produced a strong positive bias of the threshold estimate on the simulation models considered.

The MSE of the threshold estimate is equal to the variance of the estimate plus the square of the bias. This helps to explain the performance difference observed for the two half-intervals stopping criteria with respect to MSE since, in general, the HIM criteria was negatively biased or nearly unbiased while the HIF criteria had a strong negative bias for all cases considered. If the two stopping criteria produce similar variances for the estimated threshold, then the difference in bias explains the observed differences in MSE.

5.4.4 Areas of application for the methods (Exploratory vs. Regulatory)

We now consider the factors which may influence the choice of method when faced with the problem of comparing two treatments where the mean value of the response depends on the value of the covariate and the treatments differ in mean response for only part of the range of the covariates studied. The choice of a method should be dictated by the aims of the application. If the primary aim is regulatory, for example to get an indication from a regulatory body for a new treatment on a subset of the covariate values under study, then coverage would seem to be the most important performance measure, and simultaneous confidence bands then would be the clear choice of method. If, on the other hand, the aim of the application is exploratory in nature, then the important performance measure is estimation of the true threshold, and the method of half-intervals with the minimum p-value stopping criteria would apparently be the method of choice. The method of Half Intervals with the first significant p-value did not prove to be a superior method for any combination of simulation models or performance measures we considered; hence, we currently would not recommend its further use.
6 Conclusion

We have considered in this dissertation three methodologies related to the general problem of finding subsets of the population where two treatments differ. In this closing chapter, we revisit each methodology focusing on the specific areas of application, our major contribution in developing the methodology, and a description of the types of subsets which may be identified. We also discuss the conclusions from the comparison between the methods of inference guided data exploration by ANCOVA on half-intervals of a covariate and simultaneous confidence bands for the difference of segmented linear models. In the final section, we discuss directions for future research.

6.1 Controlling the Familywise Error for Multiple Log-Rank Tests

The methodology of controlling the familywise error for multiple log-rank tests, considered in Chapter 2, is applicable when the response is possibly right censored time-to-event data and we have one or more categorical factors possibly related to the size of the treatment effect. The method controls the familywise error rate for a group of hypothesis tests where we test for the difference between treatments with an overall stratified log-rank test along with individual tests within each subgroup. The subgroups here are defined by the levels of one or more related categorical factors. Our contribution here is a way of calculating the exact familywise error for the family of tests given the type I error rates of the constituent tests. The method will identify subsets of the population where two treatments differ defined by the subgroups involved in any tests which prove to be significant.

Ad hoc methods for dealing with the inflation of the type I error such as using a preliminary interaction test to decide whether to perform the overall stratified test or the individual subgroup tests are largely ineffective. While such “protection” schemes do a better job of controlling the type I error than the “unprotected” approach of performing all tests at level $\alpha$, the approach of using a preliminary interaction test still results in roughly a doubling of the FWE.
6.2 Inference Guided Data Exploration

For the method of inference guided data exploration, described in Chapter 3, we have two related methodologies. The first method, given in Section 3.4, considers all subsets of the population with the only restriction being the minimum sample size allowed for each treatment. This method is applicable to any type of response and any method of inference which produces a p-value. Instead of treating the large number of tests of hypothesis on these subsets as a family of hypothesis tests, we rather choose the minimum p-value over all subsets as our test statistic. Our contribution to the development of this methodology was in recognizing the potential of this novel test statistic and developing a technique for finding a critical value for our statistic which controls the type I error. In applying the method, one tests the null hypothesis of no difference between treatments in any of the subsets considered versus the alternative that the treatments differ in at least one of the subsets considered. Concluding the alternative does not lead to a generalizable result, but may lead to further research to identify the factors related to the variation in the effect of treatment. When considering only the sample under study and not attempting to generalize the results, any subset test which produced a p-value less than the critical p-value is identified as a subset where the treatments differ. The critical p-value from this methodology can be used as guidance to judge whether the result of a “fishing expedition” may be significant. Provided the minimum sample size is selected in advance, the critical p-value from this methodology is valid even when the subsets are selected by inspecting the data.

In Section 3.5, we describe the second method where we introduce greater restrictions on the subsets under consideration. Here we consider only the subsets of the population defined by the value of a covariate exceeding some threshold value with an additional restriction that the sample size for each treatment is at least as large as the minimum sample size. The method is again applicable to any type of response and any method of inference which produces a p-value. We again choose the minimum p-value over all subsets considered as our test statistic. Our additional contribution for this methodology was to find a reasonable restriction on the subsets considered so that the critical p-values are not as extreme as for the method only restricted by minimum sample size. The less
extreme critical p-values make this methodology more attractive for use as a planned methodology rather than as a reaction to a fishing expedition. In applying the method, we again test the null hypothesis of no difference between treatments in any of the subsets considered versus the alternative that the treatments differ in at least one of the subsets considered. Under this method, we are able to identify the subsets whose p-value is at least as extreme as the critical p-value based on the threshold value of the covariate used in defining the subset so that the results are now generalizable beyond the particular sample under study.

6.3 Asymptotic Simultaneous Confidence Band for the Difference of Segmented Linear Models

Our final methodology of asymptotic simultaneous confidence band for the difference of segmented linear models is presented in Chapter 4. Here we are concerned with comparing two treatments when the value of the continuous response depends on the value of some continuous covariate and the relationship between the response and covariate is not adequately described by a standard linear model. We model the treatments separately by segmented linear models and construct a simultaneous confidence band for the difference in mean response between the two models. Our contribution here was in recognizing the potential of the general approach employed by Potthoff (1964) for the case of non-parallel slopes in ANCOVA and extending his approach to the setting of segmented linear models. The potential scope of this method is extended since the segmented linear model provides a reasonable approximation to many nonlinear models. Our method identifies regions with respect to the covariate where the two treatments differ in mean response.

6.4 Comparison between Simultaneous Confidence Band for the Difference of Segmented Linear Models and Inference Guided Data Exploration by ANCOVA on Half-Intervals of a Covariate

We compared the method of simultaneous confidence band for the difference of segmented linear models in Chapter 5, to the method of inference guided data exploration by ANCOVA on half-intervals of a covariate. Two stopping rules, minimum p-value and first significant p-value, were employed for the half-intervals method. The methods and
stopping criteria were compared on the proportion of identified regions of significant
difference which were “covered” by the true region where the treatments differed in
mean response (coverage) and the mean squared error of the estimated threshold value of
the covariate (MSE). The threshold value defines the boundary between the regions
where the treatments do and do not differ. Several simulation models were employed to
explore the robustness of the methods to violations of underlying assumptions. The first
significant p-value stopping criteria for inference guided data exploration by ANCOVA
on half-intervals of a covariate was not found to be superior for coverage or MSE on any
of the simulation models we explored, hence we would not recommend its further use.
The minimum p-value stopping criteria for inference guided data exploration by
ANCOVA on half-intervals of a covariate was generally superior with respect to MSE.
The method of simultaneous confidence band for the difference of segmented linear
models was generally superior with respect to coverage. The choice of a method should
be dictated by the aims of the application. If the primary aim is regulatory, then coverage
would be the most important performance measure, and simultaneous confidence band
seems to be the clear choice of method. If, on the other hand, the aim of the application
is exploratory in nature, then the important performance measure is the MSE of the
threshold estimator, and the method of half-intervals with the minimum p-value stopping
criteria would apparently be the method of choice.

6.5 Areas for future research

The general approach used in controlling the familywise error rate for multiple log-rank
tests (Chapter 2) should be easily transferable to any situation where the overall test of
hypothesis is based on a normally distributed statistic expressed as a weighted sum of
normally distributed test statistics from disjoint subsets whose union is the original
sample. The approach would be more widely applicable if it can be extended to statistics
which follow the Student’s $t$ distribution. While we recognize the difficulty of the latter
problem, each of these avenues could be pursued in future research.

In order to employ the method of inference guided data exploration (Chapter 3), one must
first have an estimate of the critical p-value based on simulation from the null
distribution. This critical p-value is specific to the parameters involved including the
sample sizes for treatment and control and the minimum subset sample size. The tables of critical values we have provided in Chapter 3 are not very extensive, due to the computational demands of performing these simulations, so that the first step in a practical application of this method would be to perform a simulation for the particular parameters involved. An alternative to simulating a critical p-value for a given test procedure and set of underlying distributional assumptions would be to perform a randomization test for an observed pair of samples for treatment and control. The computational effort involved in a randomization test would be similar to that required to simulate a critical p-value. It would be interesting to compare these randomization tests to the methods of inference guided data exploration already developed.

Under the null hypothesis of no difference between treatments, the test statistic for inference guided data exploration by half-intervals of a covariate (Section 3.5) should behave as if it were on a random walk for successive values of the covariate threshold. If the theory of sequential sampling could be adapted to this problem, it may be possible to derive a closed form solution for the critical value of certain test statistics so that simulation is unnecessary.

The approach of constructing a simultaneous confidence band about the difference of two regression models, which we applied to segmented linear models in Chapter 4, may have extension to other problems. For example, the effect of a continuous covariate may be different for treatment and control in a logistic regression or a proportional hazards model. If this is the case, one can imagine fitting separate models for treatment and control and constructing a simultaneous confidence band about the difference of the two models in order to determine regions, with respect to the covariate, where the treatments differ significantly.

6.6 Acknowledgements

We wish to thank Professor Stewart Anderson Ph.D. of the University of Pittsburgh’s Department of Biostatistics and the NSABP for providing the tumor response data used in our application.
Appendix A

Recursive Definition of the Power Function for Multiple Log-rank Tests

We derive the recursive power function given in Section 2.4.4 in this appendix.

Given \( \alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k \), define:

\[
\rho_r(z) \equiv \Pr \left\{ \left( \sum_{i=1}^{r} Z_i \sqrt{a_i} \leq z \right) \cap \left( \bigcap_{j=1}^{r} [|Z_j| \leq c_j] \right) \right\}.
\]  
(A.1)

Then,

\[
\rho_{r+1}(z) = \Pr \left\{ \left( \sum_{i=1}^{r+1} Z_i \sqrt{a_i} \leq z \right) \cap \left( \bigcap_{j=1}^{r+1} [|Z_j| \leq c_j] \right) \right\}
= \Pr \left\{ \left( \sum_{i=1}^{r} Z_i \sqrt{a_i} \leq z - Z_{r+1} \sqrt{a_{r+1}} \right) \cap \left( \bigcap_{j=1}^{r+1} [|Z_j| \leq c_j] \right) \right\}
= \int_{c_{r+1}}^{z} \phi(u - \theta_{r+1}) \Pr \left\{ \left( \sum_{i=1}^{r} Z_i \sqrt{a_i} \leq z - u \sqrt{a_{r+1}} \right) \cap \left( \bigcap_{j=1}^{r+1} [|Z_j| \leq c_j] \right) \right\} du
= \int_{c_{r+1}}^{z} \phi(u - \theta_{r+1}) \rho_r(z - u \sqrt{a_{r+1}}) du,
\]  
(A.2)

and,

\[
\rho_1(z) = \Pr \left\{ \left( Z_1 \sqrt{a_1} \leq z \right) \cap \left( |Z_1| \leq c_1 \right) \right\}
= \Pr \left\{ -c_1 \leq Z_1 \leq \min \left\{ c_1, z/\sqrt{a_1} \right\} \right\}
= \int_{-c_1}^{\max \left\{ -c_1, \min \left\{ c_1, z/\sqrt{a_1} \right\} \right\}} \phi(u - \theta_1) du
= \Phi \left\{ \max \left[ -c_1, \min \left\{ c_1, z/\sqrt{a_1} \right\} \right] - \theta_1 \right\} - \Phi (-c_1 - \theta_1).
\]  
(A.3)

For \( k \) subgroups, we claim that:

\[
\text{Power} \left( \alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k \right) = 1 - \left[ \rho_k(c_0) - \rho_k(-c_0) \right].
\]  
(A.4)
Assume (A.4), then:

\[
\text{Power} (\alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k) = 1 - \left[ \rho_k (c_0) - \rho_k (-c_0) \right]
\]

\[
= 1 - \int_{-c_k}^{c_k} \phi(z_k - \theta_k) \rho_{k-1} \left( c_0 - z_k \sqrt{a_k} \right) dz_k + \int_{-c_k}^{c_k} \phi(z_k - \theta_k) \rho_{k-1} \left( -c_0 - z_k \sqrt{a_k} \right) dz_k
\]

\[
+ \int_{-c_k}^{c_k} \cdots \int_{-c_k}^{c_k} \phi(z_2 - \theta_2) \cdots \phi(z_k - \theta_k) \rho_1 \left( c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i} \right) dz_2 \cdots dz_k
\]

\[
= 1 - \int_{-c_k}^{c_k} \cdots \int_{-c_k}^{c_k} \prod_{i=2}^{k} \phi(z_i - \theta_i) \left[ \rho_1 \left( c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i} \right) - \rho_1 \left( -c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i} \right) \right] dz_2 \cdots dz_k
\]

\[
= 1 - \int_{-c_k}^{c_k} \cdots \int_{-c_k}^{c_k} \prod_{i=2}^{k} \phi(z_i - \theta_i) \left[ \Phi \left\{ \max \left[ -c_1, \min \left( c_1, \frac{c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i}}{\sqrt{a_1}} \right) \right] \right\} - \Phi(-c_1 - \theta_1) \right]
\]

\[
- \Phi \left\{ \max \left[ -c_1, \min \left( c_1, \frac{c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i}}{\sqrt{a_1}} \right) \right] \right\} - \theta_1 \right\} + \Phi(-c_1 - \theta_1) \right\} \right] ) dz_2 \cdots dz_k
\]

\[
= 1 - \int_{-c_k}^{c_k} \cdots \int_{-c_k}^{c_k} \prod_{i=2}^{k} \phi(z_i - \theta_i) \left[ \Phi \left\{ \max \left[ -c_1, \min \left( c_1, \frac{c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i}}{\sqrt{a_1}} \right) \right] \right\} - \theta_1 \right\}
\]

\[
- \Phi \left\{ \max \left[ -c_1, \min \left( c_1, \frac{c_0 - \sum_{i=2}^{k} z_i \sqrt{a_i}}{\sqrt{a_1}} \right) \right] \right\} - \theta_1 \right\} dz_2 \cdots dz_k.
\]

\[(A.5)\]

This agrees with equation (2.4.6) in Section 2.4.3.
Appendix B

Strong Control of FWE for Multiple Log-rank Tests

Here we show that strong control of the FWE is achieved in the sense that for any arrangement of true and false null hypotheses, the FWE will be no more than $\alpha$. The point is best illustrated using a slight modification of the non-recursive representation of the power function (2.4.6) as follows:

$$
\text{Power}(\alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k) =
1 - \int_{-c_1}^{c_1} \cdots \int_{-c_2}^{c_2} \max\left\{ -c_i, \min\left( c_i, \left\{ c_i - \sum_{t=2}^{k} z_i \sqrt{a_i} \right\} / \sqrt{a_i} \right) \right\}_{i=1}^{k} \prod_{i=1}^{k} \phi(z_i - \theta) \, dz_1 \cdots dz_k.
$$

(B.1)

We control the FWE at level $\alpha$ by choosing $\alpha_0, \ldots, \alpha_k$ such that the power (B.1) is $\alpha$ when all null hypotheses are true ($\theta_1 = \cdots = \theta_k = 0$), given $a_1, \ldots, a_k$, where $a_i$ is the proportion of expected events in subgroup $i$. Examination of this equation reveals that the FWE is simply 1 minus the probability of making no type I error. In a situation where some of the null hypotheses are true and some are false ($\theta_i \neq 0$ for some $i$), the FWE is 1 minus the probability of rejecting no true null hypotheses. An expression for the FWE can be derived by substituting $\infty$ for the critical value $c_i$ of any false null hypothesis ($\theta_i \neq 0, i = 0, \ldots, k$) in (B.1). This substitution allows the relevant random variables where the null hypothesis is false to assume all possible values since by definition they cannot contribute to a type I error. When the random variables with the false nulls are allowed to range over the full support of the normal density rather than being restricted to range within the critical values, the value of the multiple integral must increase and one minus the multiple integral (the FWE) must decrease. Thus strong control of the FWE is achieved.
For the case where the overall null hypothesis is false ($\theta_0 \neq 0$), the expression for the FWE can be simplified to give:

$$FWE(\alpha_0, \alpha_1, \ldots, \alpha_k, a_1, \ldots, a_k, \theta_1, \ldots, \theta_k) = 1 - \prod_{i, \theta_i = 0} (1 - \alpha_i)$$

(B.2)
Appendix C

Equivalence of Protection Schemes for Multiple Log-rank Tests

Protection schemes (2) and (3) are equivalent if our conjecture that, given that there is a type I error on the overall test as well as the test of interaction between treatment and subgroup, then, there is a type I error in at least one subgroup with probability one. We prove the conjecture for the case of $k = 2$ subgroups in this appendix.

Let $L_1$ and $L_2$ be the stratified log-rank statistics for subgroups 1 and 2 respectively, and let $V_1$ and $V_2$ be the corresponding variances of the statistics.

The standardized statistics are $Z_j = L_j / V_j$, $j = 1, 2$, $Z_j \sim N(0, 1)$.

$L_0 = L_1 + L_2$ is the overall stratified log-rank statistic.

$V_0 = V_1 + V_2$ is the variance of the overall statistic.

Then, $Z_0 \sim N(0, 1)$ where $Z_0 = L_0 / V_0 = Z_1 \sqrt{a + Z_2 \sqrt{1-a}}$, and $a = \frac{V_1}{V_1 + V_2}$, $0 < a < 1$.

Similarly, $Z_{\text{inter}} = Z_1 \sqrt{1-a} - Z_2 \sqrt{a}$.

Now, given that $(|Z_0| > z_{a/2}$ and $|Z_{\text{inter}}| > z_{a/2})$, there are four possibilities:

(i) $(Z_0 > z_{a/2}$ and $Z_{\text{inter}} > z_{a/2})$
(ii) $(Z_0 < -z_{a/2}$ and $Z_{\text{inter}} > z_{a/2})$
(iii) $(Z_0 > z_{a/2}$ and $Z_{\text{inter}} < -z_{a/2})$
(iv) $(Z_0 < -z_{a/2}$ and $Z_{\text{inter}} < -z_{a/2})$
Assume \((i) (Z_0 > z_{a/2} \text{ and } Z_{\text{inter}} > z_{a/2})\), then:

\[
\left( Z_1 \sqrt{a} + Z_2 \sqrt{1-a} > z_{a/2} \text{ and } Z_1 \sqrt{1-a} - Z_2 \sqrt{a} > z_{a/2} \right)
\]

\[
\Rightarrow \left( Z_1 > \frac{z_{a/2} - Z_2 \sqrt{1-a}}{\sqrt{a}} \text{ and } Z_1 > \frac{z_{a/2} + Z_2 \sqrt{a}}{\sqrt{1-a}} \right). \tag{C.1}
\]

\(Z_1\) is minimized where the right hand sides of the above inequalities coincide.

\[
\Rightarrow z_{a/2} - Z_2 \sqrt{1-a} = z_{a/2} + Z_2 \sqrt{a}
\]

\[
\Rightarrow Z_2 = z_{a/2} \left( \sqrt{1-a} - \sqrt{a} \right)
\]

\[
\Rightarrow Z_1 > \frac{z_{a/2} + z_{a/2} \left( \sqrt{1-a} - \sqrt{a} \right) \sqrt{a}}{\sqrt{1-a}}
\]

\[
= \frac{1 + (\sqrt{1-a} - \sqrt{a}) \sqrt{a}}{\sqrt{1-a}}
\]

\[
= \frac{(1-a) + \sqrt{a} \sqrt{1-a}}{\sqrt{1-a}}
\]

\[
= z_{a/2} \left( \sqrt{1-a} + \sqrt{a} \right)
\]

\[
> z_{a/2}, \text{ since } 0 < a < 1. \tag{C.2}
\]

By similar argument, \((ii) \Rightarrow Z_2 < -z_{a/2}, (iii) \Rightarrow Z_2 > z_{a/2}, \text{ and } (iv) \Rightarrow Z_1 < -z_{a/2} \).

Hence, \(\Pr \{|Z_1| > z_{a/2} \text{ or } |Z_2| > z_{a/2}\} = 1 \text{ if } (|Z_0| > z_{a/2} \text{ and } |Z_{\text{inter}}| > z_{a/2})\).
Appendix D

Asymptotic Normality of our Parameterization of the Segmented Linear Model

Feder (1975) shows that, under his parameterization of the segmented linear model, the parameters have an asymptotic joint normal distribution. We show here that, under our parameterization of this model, the parameters are also asymptotically distributed as joint normal. We treat the case of a single treatment and drop notation dealing with the treatment group.

Feder’s (1975) parameterization for an \( m \)-segment broken line is:

\[
f(x_j, \tau) = \begin{cases} 
\tau_{11} + \tau_{12}x_j, & x_j \in S_1 \\
\vdots & \\
\tau_{m1} + \tau_{m2}x_j, & x_j \in S_m 
\end{cases}
\]

\[
= \sum_{k=1}^{m} (\tau_{k1} + \tau_{k2}x_j) I_{[s_j, a_j]},
\]

where \( j = 1, \ldots, n; \quad k = 1, \ldots, m; \quad \tau = (\tau_{11}, \tau_{12}, \ldots, \tau_{m1}, \tau_{m2})' \);

\[
S_k = \left[ \begin{array}{cc}
\tau_{(k-1)1} - \tau_{k1} & \tau_{(k-1)2} - \tau_{k2} \\
\tau_{k1} - \tau_{(k+1)1} & \tau_{k2} - \tau_{(k+1)2}
\end{array} \right];
\]

\[
\frac{\tau_{01} - \tau_{11}}{\tau_{12} - \tau_{02}} \equiv -\infty; \quad \text{and} \quad \frac{\tau_{m1} - \tau_{(m+1)1}}{\tau_{(m+1)2} - \tau_{m2}} \equiv \infty.
\]

Feder shows that under reasonable regularity conditions:

\[
\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{L} N(0, [\int I(\tau, x) dF(x)]^{-1}),
\]

where \( \hat{\tau} \) is the MLE, \( I(\tau, x) \) is the Fisher’s information about \( \tau \) contained in a single observation with covariate value \( x \), and \( F \) is the distribution of the covariate.
For a single treatment group, we solve for the parameters of our parameterization in terms of Feder’s parameters:

\[
\begin{align*}
\theta_1 &= h_1(\tau) = \tau_{11} \\
\theta_2 &= h_2(\tau) = \tau_{12} \\
\theta_{21} &= h_{21}(\tau) = (\tau_{11} - \tau_{21}) / (\tau_{22} - \tau_{12}) \\
\theta_{22} &= h_{22}(\tau) = \tau_{22} - \tau_{12} \\
\vdots \\
\theta_{m1} &= h_{m1}(\tau) = (\tau_{(m-1)1} - \tau_{11}) / (\tau_{m2} - \tau_{(m-1)2}) \\
\theta_{m2} &= h_{m2}(\tau) = \tau_{m2} - \sum_{k=1}^{m-1} \tau_{k2}.
\end{align*}
\]  

(D.3)

The Jacobian matrix \( H \) is equal to:

\[
H = \begin{pmatrix}
\frac{\partial h_1(x, \tau)}{\partial \tau_{11}} & \frac{\partial h_1(x, \tau)}{\partial \tau_{12}} & \cdots & \frac{\partial h_1(x, \tau)}{\partial \tau_{m2}} \\
\frac{\partial h_2(x, \tau)}{\partial \tau_{11}} & \frac{\partial h_2(x, \tau)}{\partial \tau_{12}} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial h_{m2}(x, \tau)}{\partial \tau_{11}} & \cdots & \frac{\partial h_{m2}(x, \tau)}{\partial \tau_{m2}}
\end{pmatrix}.
\]  

(D.4)

\[
H = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
0 & \frac{\tau_{11} - \tau_{21}}{\tau_{22} - \tau_{12}} & \frac{\tau_{11} - \tau_{21}}{\tau_{22} - \tau_{12}} & \frac{-1}{\tau_{22} - \tau_{12}} \\
0 & \frac{-1}{\tau_{22} - \tau_{12}} & \frac{-1}{\tau_{22} - \tau_{12}} & \frac{0}{\tau_{22} - \tau_{12}} \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 \\
0 & 0 & \frac{\tau_{(m-1)1} - \tau_{11}}{\tau_{m2} - \tau_{(m-1)2}} & \frac{\tau_{(m-1)1} - \tau_{11}}{\tau_{m2} - \tau_{(m-1)2}} \\
0 & 0 & \cdots & 0 \\
0 & 0 & \frac{-1}{\tau_{m2} - \tau_{(m-1)2}} & \frac{-1}{\tau_{m2} - \tau_{(m-1)2}} \\
0 & -1 & \cdots & 1
\end{pmatrix}.
\]  

(D.5)

\( H \) is non-singular since the determinant \( \prod_{k=2}^{m} \left( \tau_{k2} - \tau_{(k-1)2} \right) \) is non-zero when the model is identifiable (slopes differ between adjacent segments). Hence, by the multivariate delta method, our parameter estimates obtained via least squares/maximum likelihood have an asymptotic joint normal distribution (Lehmann (1999, p. 315)). This distribution is given by \( \sqrt{n} (\hat{\theta} - \theta) \xrightarrow{L} N\left(0, H \left[ \int I(\tau, x) dF(x) \right]^{-1} H' \right) \).
But, since by Lehmann (1999, p. 500, eq. 7.5.13) as corrected by Hunter (2001):

\[
H \left[ \int I(\tau, x) dF(x) \right]^{-1} H' = \left( (H')^{-1} \left[ \int I(\tau, x) dF(x) \right] H^{-1} \right)^{-1}
\]

\[
= \left[ \int (H')^{-1} I(\tau, x) H^{-1} dF(x) \right]^{-1}
\]

\[
= \left[ \int I(\theta, x) dF(x) \right]^{-1},
\]

(D.6)

it follows that \( \sqrt{n} (\hat{\theta} - \theta) \rightarrow^L N (0, \left[ \int I(\theta, x) dF(x) \right]^{-1}) \).
Appendix E

Derivation of the Plug-in Estimator of the Information Matrix

We derive here the plug-in estimator of the information matrix of the parameters given in equation (4.3.5) and show its consistency.

For a single treatment we have (again suppressing notation dealing with treatment group) a set of responses $Y_1, \ldots, Y_n$ and associated covariates $x_1, \ldots, x_n$ such that:

$$Y_i = f(x_i, \theta) + \varepsilon_i, \quad (E.1)$$

where $\varepsilon_i$ is distributed as iid $N(0, \sigma^2)$.

It follows that $Y_i$ is distributed as independent $N(f(x_i, \theta), \sigma^2)$.

Now let $\phi(y)$ denote the Normal density. Then the log likelihood for a single observation with covariate $x$ is:

$$\log \phi(y) = -\frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - f(x, \theta))^2 - \log \sqrt{2\pi}. \quad (E.2)$$

From Lehmann (1999, Definition 7.5.1, p.497) the information matrix for a single observation with covariate $x$ has elements:

$$I_{jk}(\theta, x) = E \left[ \frac{\partial}{\partial \theta_j} \log \phi(y) \frac{\partial}{\partial \theta_k} \log \phi(y) \right]. \quad (E.3)$$

Now,

$$\frac{\partial}{\partial \theta_j} \log \phi(y) = \frac{1}{\sigma^2} (y - f(x, \theta)) \frac{\partial}{\partial \theta_j} f(x, \theta), \quad (E.4)$$
so that,

\[
I_{jk}(\theta, x) = E \left[ \frac{1}{\sigma^2} (y - f(x, \theta)) \frac{\partial}{\partial \theta_j} f(x, \theta) \frac{1}{\sigma^2} (y - f(x, \theta)) \frac{\partial}{\partial \theta_k} f(x, \theta) \right]
\]

\[
= \frac{\sigma^2}{\sigma^4} \frac{\partial}{\partial \theta_j} f(x, \theta) \frac{\partial}{\partial \theta_k} f(x, \theta) \tag{E.5}
\]

\[
= \frac{1}{\sigma^2} \frac{\partial}{\partial \theta_j} f(x, \theta) \frac{\partial}{\partial \theta_k} f(x, \theta).
\]

Which can be written in matrix notation as:

\[
I(\theta, x) = \frac{1}{\sigma^2} \left( \frac{\partial}{\partial \theta} f(x, \theta) \right) \left( \frac{\partial}{\partial \theta} f(x, \theta) \right)^\prime, \tag{E.6}
\]

where \( \frac{\partial}{\partial \theta} f(x, \theta) = \left( \frac{\partial}{\partial \theta_1} f(x, \theta), \frac{\partial}{\partial \theta_2} f(x, \theta), \ldots, \frac{\partial}{\partial \theta_m} f(x, \theta), \frac{\partial}{\partial \theta_n} f(x, \theta) \right)^\prime \), so that \( I(\theta, x) \) has dimension \( 2m \times 2m \).

Now for a sample of size \( n \), by the additivity of the information (Lehmann, (1999), Thm. 7.2.2, p. 465):

\[
I(\theta, x_1, \ldots, x_n) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta} f(x_i, \theta) \right) \left( \frac{\partial}{\partial \theta} f(x_i, \theta) \right)^\prime, \tag{E.7}
\]

and the plug-in estimator of the information is given by:

\[
I(\hat{\theta}, x_1, \ldots, x_n) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta} f(x_i, \theta) \bigg|_{\theta = \hat{\theta}} \right) \left( \frac{\partial}{\partial \theta} f(x_i, \theta) \bigg|_{\theta = \hat{\theta}} \right)^\prime. \tag{E.8}
\]

This agrees with (4.3.5).

Equation (4.2.1) defines the segmented linear function \( f \). We repeat the definition here, again suppressing treatment group notation.

\[
f(x, \theta) = \theta_{11} + \theta_{12} x + \sum_{k=2}^{m} \theta_{k2} (x - \theta_{k1}) I_{[x > \theta_{k1}]} \tag{E.9}
\]

Fixing \( x \) and any parameters not being differentiated, the partial derivatives are:

\[
\frac{\partial f}{\partial \theta_{11}} = 1, \quad \frac{\partial f}{\partial \theta_{12}} = x \tag{E.10}
\]

126
and for \( k = 2, \ldots, m \),

\[
\frac{\partial f}{\partial \theta_{ki}} = \begin{cases} \theta_{ki} I_{[x>\theta_{ki}]}, & \theta_{ki} \neq x \\ \text{Not Differentiable}, & \theta_{ki} = x \end{cases}
\]  \hspace{1cm} (E.11)

\[
\frac{\partial f}{\partial \theta_{k2}} = (x-\theta_{k1})I_{[x>\theta_{k1}]},
\]  \hspace{1cm} (E.12)

hence, the vector of partial derivatives \( \frac{\partial f(x, \theta)}{\partial \theta} \) does not exist for \( \theta_{k1} = x \) where 
\( k \in \{2, \ldots, m\} \) since equation (E.11) is undefined.

If \( \frac{\partial f(x, \theta)}{\partial \theta} \) fails to exist for any observed covariate \( x_t \) in the summation (E.8), then the estimator of the information does not exist. When working with data, it is not unusual to find that one or more of the estimated changepoints \((\hat{\theta}_{k1}, k \in \{2, \ldots, m\})\) coincide with an observed covariate value, so that this problem cannot be ignored.

We propose an alternate estimator where we substitute the value zero for \( \frac{\partial f}{\partial \theta_{ki}} \) whenever it is undefined or equivalently let \( \frac{\partial f}{\partial \theta_{ki}} \equiv -\theta_{ki} I_{[x>\theta_{ki}]} \) for all \( x \). Then,

\[
\frac{\partial f(x, \theta)}{\partial \theta} = (1, x, -\theta_{z1} I_{[x>\theta_{z1}]}, (x-\theta_{z1}) I_{[x>\theta_{z1}]}, \ldots, -\theta_{z1} I_{[x>\theta_{z1}]}, \ldots, (x-\theta_{m}) I_{[x>\theta_{m}]})' \].  \hspace{1cm} (E.13)

We now have a proxy for the vector of partial derivatives which exists for all \( x \), and hence, the proposed estimator of the information (E.8) will always exist for any observed set of covariates. Any subsequent reference to the estimated information (E.8) will imply the substitution of the proxy (E.13) for the vector of partial derivatives.

In order to use our results related to Cox and Ma (1995), we need to show that

\[
n \hat{\Sigma}(\hat{\theta}) \overset{p}{\to} \Sigma, \text{ or equivalently that } n[I(\hat{\theta}, x_1, \ldots, x_n)]^{-1} \overset{p}{\to} \left[\int I(\theta, x) \ dF(x)\right]^{-1}
\]

or

\[
I(\hat{\theta}, x_1, \ldots, x_n)/n \overset{p}{\to} \int I(\theta, x) \ dF(x). \quad \text{We will prove the latter by showing that each}
\]

element of the matrix on the left hand side converges in probability to the corresponding element on the right hand side, without loss of generality, we treat the case of \( m = 2 \) segments.
We can write \( I(\hat{\theta}, x_1, \ldots, x_n) / n \) for the case of \( m = 2 \) segments as (omitting the symmetric elements below the diagonal):

\[
\begin{align*}
\frac{I(\hat{\theta}, x_1, \ldots, x_n)}{n} &= \frac{1}{n \hat{\sigma}^2} \sum_{i=1}^{n} \begin{pmatrix}
1 & x_i & (x_i - \hat{\theta}_{21})I_{[x_i > \hat{\theta}_{21}]} \\
& x_i^2 & \hat{\theta}_{22}I_{[x_i > \hat{\theta}_{21}]} \\
& & -\hat{\theta}_{22}I_{[x_i > \hat{\theta}_{21}]} \\
& & \hat{\theta}_{22}^2I_{[x_i > \hat{\theta}_{21}]} \\
& & x_i(x_i - \hat{\theta}_{21})I_{[x_i > \hat{\theta}_{21}]} \\
& & \hat{\theta}_{22}(x_i - \hat{\theta}_{21})I_{[x_i > \hat{\theta}_{21}]} \\
& & \hat{\theta}_{22}^2(x_i - \hat{\theta}_{21})I_{[x_i > \hat{\theta}_{21}]} \\
& & (x_i - \hat{\theta}_{21})^2I_{[x_i > \hat{\theta}_{21}]} \\
\end{pmatrix}
\end{align*}
\]

(E.14)

From Feder (1975, Thm. 3.6, p. 57) we have that \( \hat{\theta} \xrightarrow{p} \theta, \hat{\sigma}^2 \xrightarrow{p} \sigma^2 \), and, by assumption, the \( x_i \) are selected in such a way that the empirical distribution of the covariates \( F_n \) formed from the sample, \( x_1, \ldots, x_n \), converges in distribution to that of a random variable \( X \) having distribution \( F \). Let \( X_n \) be a random variable with distribution given by the empirical distribution \( F_n \) of \( x_1, \ldots, x_n \). It is clear that \( X_n \xrightarrow{L} X \), where \( X \) has the limiting distribution \( F \). By the definition of the empirical distribution function, we can write (E.14) as:

\[
\frac{I(\hat{\theta}, x_1, \ldots, x_n)}{n} = \frac{1}{\hat{\sigma}^2} \begin{pmatrix}
1 & E_F(X_n) & -\hat{\theta}_{22}E_F(I_{[x_i > \hat{\theta}_{21}]}) & E_F([X_n - \hat{\theta}_{21}]I_{[x_i > \hat{\theta}_{21}]}) \\
& E_F(X_n^2) & -\hat{\theta}_{22}E_F(X_n I_{[x_i > \hat{\theta}_{21}]}) & E_F([X_n - \hat{\theta}_{21}]I_{[x_i > \hat{\theta}_{21}]}) \\
& & \hat{\theta}_{22}^2E_F(I_{[x_i > \hat{\theta}_{21}]}) & -\hat{\theta}_{22}E_F([X_n - \hat{\theta}_{21}]I_{[x_i > \hat{\theta}_{21}]}) \\
& & & E_F([X_n - \hat{\theta}_{21}]^2I_{[x_i > \hat{\theta}_{21}]}) \\
\end{pmatrix}
\]

(E.15)

The corresponding elements (omitting those below the diagonal) for which we wish to show convergence of (E.15) are given by:

\[
\int I(\theta, x) dF(x) = \frac{1}{\sigma^2} \begin{pmatrix}
1 & E_F(X) & -\theta_{22}E_F(I_{[X > \theta_{21}]}) & E_F([X - \theta_{21}]I_{[X > \theta_{21}]}) \\
& E_F(X^2) & -\theta_{22}E_F(X I_{[X > \theta_{21}]}) & E_F([X - \theta_{21}]I_{[X > \theta_{21}]}) \\
& & \theta_{22}^2E_F(I_{[X > \theta_{21}]}) & -\theta_{22}E_F([X - \theta_{21}]I_{[X > \theta_{21}]}) \\
& & & E_F([X - \theta_{21}]^2I_{[X > \theta_{21}]}) \\
\end{pmatrix}
\]

(E.16)

From Lehmann (1999, Thm. 2.4.1, p. 51), we have that \( 1/\hat{\sigma}^2 \xrightarrow{p} 1/\sigma^2 \) which is sufficient to prove the convergence of cell (1, 1).

In order to proceed, we require two further assumptions; for some positive \( \delta \);

(i) \( E(|X_n|^2 + \delta) < \infty \) for all \( n \) and (ii) \( E(|X|^2 + \delta) < \infty \). The first assumption implies that the random variables \( X_n \) and \( X_n^2 \) are uniformly integrable and the existence of the
expectations in (E.16) is guaranteed by the second assumption. Since \( E(|X_n|^2 + \delta) < \infty \) and \( X_n \xrightarrow{P} X \), it follows that \( E(X_n) \rightarrow E(X) \) and \( E(X_n^2) \rightarrow E(X^2) \) (Billingsley, 1995, corollary to Thm. 25.12, p. 338). The convergence of cells (1, 2) and (2, 2) then follows from \( 1/\hat{\sigma}^2 \xrightarrow{P} 1/\sigma^2 \) by Slutsky’s Theorem.

**Lemma D.1:** Let \( G_n \) be a sequence of distributions and \( G \) be a distribution such that \( G_n(y) \rightarrow G(y) \) at all continuity points of \( G \). For every continuity point \( a \) of \( G \) there exists a neighborhood \( R \) of \( a \) such that \( G_n(y) \rightarrow G(y) \) uniformly for \( y \in R \). That is, for every \( \varepsilon > 0 \) there exists \( N(\varepsilon) > 0 \) such that \( n \geq N(\varepsilon) \) implies that \( |G_n(y) - G(y)| < \varepsilon \) for every \( y \in R \).

**Proof:** The proof follows the proof of Polya’s Lemma given in Roussas (1997, pp. 206-207).

Since \( G \) is continuous at \( a \), there exists for some positive \( \gamma \) a neighborhood of \( a \), say \( (a - \gamma, a + \gamma) \), where \( G \) is continuous. Let \( \delta = \gamma/2 \), then \( 0 \leq G(a - \delta) \leq G(a + \delta) \leq 1 \), and the continuity of \( G \) on \( (a - \gamma, a + \gamma) \) implies its uniform continuity on \( [a - \delta, a + \delta] \). It follows that we can construct a finite partition \( a - \delta = y_1 < y_2 < \cdots < y_r = a + \delta \) of \( [a - \delta, a + \delta] \) such that:

\[
G(y_{j+1}) - G(y_j) < \varepsilon / 2, \quad j = 1, \ldots, r - 1. \tag{E.17}
\]

Next, \( G_n(y_j) \rightarrow G(y_j) \) implies there exists \( N_j(\varepsilon) > 0 \) such that for all \( n \geq N_j(\varepsilon) \):

\[
|G_n(y_j) - G(y_j)| < \varepsilon / 2, \quad j = 1, \ldots, r. \tag{E.18}
\]

By taking \( n \geq N(\varepsilon) = \max(N_1(\varepsilon), \ldots, N_r(\varepsilon)) \), we then have that:

\[
|G_n(y_j) - G(y_j)| < \varepsilon / 2, \quad j = 1, \ldots, r. \tag{E.19}
\]

Next let \( y \in [a - \delta, a + \delta] \), then \( y_j < y < y_{j+1} \) for some \( j = 1, \ldots, r - 1 \). By (E.17) and (E.19) and for \( n \geq N(\varepsilon) \), we have the following string of inequalities:

\[
G(y_j) - \varepsilon / 2 < G_n(y_j) \leq G_n(y) \leq G_n(y_{j+1}) < G(y_j) + \varepsilon \leq G(y) + \varepsilon \leq G(y_{j+1}) + \varepsilon. \tag{E.20}
\]

Hence \( 0 \leq G(y) + \varepsilon - G_n(y) \leq G(y_{j+1}) + \varepsilon - G(y_j) + \varepsilon / 2 < 2\varepsilon \), and therefore \( |G_n(y) - G(y)| < \varepsilon \) for every \( y \in R \), where \( R = (a - \delta, a + \delta) \). ■
Lemma D.2: Assume $A_n$ is a sequence of random variables such that $A_n \overset{p}{\to} a$, $G_n$ is a sequence of distributions and $G$ is a distribution such that $G_n(y) \to G(y)$ at all continuity points of $G$, and $a$ is a continuity point of $G$, then $G_n(A_n) \overset{p}{\to} G(a)$.

Proof: It suffices to show that for arbitrary $\delta > 0$ and $\varepsilon > 0$, we can choose $N$, depending on $\delta$ and $\varepsilon$, such that $n > N$ implies $P\{|G_n(A_n) - G(a)| > \delta\} < \varepsilon$. We have:

\[
P\{|G_n(A_n) - G(a)| > \delta\} = P\{|G_n(A_n) - G(A_n) + G(A_n) - G(a)| > \delta\}
\leq P\{|G_n(A_n) - G(A_n)| > \delta / 2\} \cup \{|G_n(A_n) - G(a)| > \delta / 2\},
\]

(E.21)

so that it remains to show that $n > N$ implies $P\{|G_n(A_n) - G(A_n)| > \delta / 2\} < \varepsilon / 2$ and $P\{|G(A_n) - G(a)| > \delta / 2\} < \varepsilon / 2$.

By Lehmann (1999, Thm. 2.1.4, p. 51), since $A_n \overset{p}{\to} a$, and $G$ is continuous at $a$, $G(A_n) \overset{p}{\to} G(a)$ and we can choose $N_1$ such that $n > N_1$ implies $P\{|G(A_n) - G(a)| > \delta / 2\} < \varepsilon / 2$.

By Lemma D.1, there exists a neighborhood $R$ of $a$, say $(a - \gamma, a + \gamma)$ where $G_n \to G$ uniformly. Since $A_n \overset{p}{\to} a$, we can choose $N_2$ such that $n > N_2$ implies $P\{|A_n - a| > \gamma\} = P\{A_n \not\in R\} < \varepsilon / 2$. We can choose $N_3$, by Lemma D.1, such that $n > N_3$ implies $|G_n(y) - G(y)| < \delta / 2$ for $y \in R$. Now for $n > \max(N_2, N_3)$:

\[
P\{|G_n(A_n) - G(A_n)| > \delta / 2\} = \left. P\{|G_n(A_n) - G(A_n)| > \delta / 2\} \right\} \cup \left\{ A_n \not\in R \right\}
\leq P\{A_n \not\in R\} < \varepsilon / 2
\]

(E.22)

The result follows if we choose $N = \max(N_1, N_2, N_3)$.

Returning to our argument regarding the convergence of the elements of the information, we can write the expectations in cells (1, 3) and (3, 3) of (E.15) and (E.16) as:

\[
E_{F_n}(I_{\{X_n > \hat{\theta}_{21}\}}) = P(X_n > \hat{\theta}_{21}) = 1 - F_n(\hat{\theta}_{21}),
\]

(E.23)

\[
E_{F}(I_{\{X > \theta_{21}\}}) = P(X > \theta_{21}) = 1 - F(\theta_{21}).
\]

(E.24)
By assumption (Section 4.3.1), the changepoints \((\theta_{k1}, k \geq 2)\) occur at continuity points of \(F\) and \(F_n \xrightarrow{P} F\) at all continuity points of \(F\). Since \(\hat{\theta}_{21} \xrightarrow{p} \theta_{21}\) and \(F\) is continuous at \(\theta_{21}\), it follows from Lemma D.2 that \(E_{F_n}(I_{\{X > \hat{\theta}_{21}\}}) \xrightarrow{p} E_F(I_{\{X > \theta_{21}\}})\). The convergence of cells (1, 3) and (3, 3) then follows from \(\frac{1}{\sigma^2} \xrightarrow{P} 1/\sigma^2\) and \(\hat{\theta}_{22} \xrightarrow{p} \theta_{21}\) by Slutsky’s Theorem.

We now focus on the remaining cells of (E.15) and (E.16). Let \(Y_n = X_n I_{\{X \leq \theta_{21}\}}\) and \(Y = X I_{\{X \leq \theta_{21}\}}\), then the corresponding distributions of \(Y_n\) and \(Y\), respectively are:

\[
G_n(y) = \frac{F_n(y)}{F_n(\theta_{21})} I_{\{y \leq \theta_{21}\}} + I_{\{y > \theta_{21}\}},
\]

(E.25)

\[
G(y) = \frac{F(y)}{F(\theta_{21})} I_{\{y \leq \theta_{21}\}} + I_{\{y > \theta_{21}\}}.
\]

(E.26)

Since \(F_n \xrightarrow{P} F\) at all continuity points of \(F\) and in particular at \(\theta_{21}\), it follows that \(G_n \xrightarrow{P} G\) at all continuity points of \(G\) and hence that \(Y_n \xrightarrow{L} Y\). Furthermore, since \(X_n\) and \(X_n^2\) are uniformly integrable it follows that \(Y_n\) and \(Y_n^2\) are uniformly integrable.

For cell (2, 3), we can write the expectations as:

\[
E_F(X_n I_{\{X > \theta_{21}\}}) = E_{F_n}(X_n) - \int_{-\infty}^{\hat{\theta}_{21}} x dF_n(x)
\]

\[
= E_{F_n}(X_n) - \int_{-\infty}^{\theta_{21}} x dF_n(x) + \int_{\theta_{21}}^{\hat{\theta}_{21}} x dF_n(x)
\]

\[
= E_{F_n}(X_n) - F_n(\theta_{21}) \int_{-\infty}^{\theta_{21}} \left[ x / F_n(\theta_{21}) \right] dF_n(x) + \int_{\theta_{21}}^{\hat{\theta}_{21}} x dF_n(x)\]

(E.27)

\[
= E_{F_n}(X_n) - F_n(\theta_{21}) G_n(Y_n) + r_n(\hat{\theta}_{21}, F_n),
\]

where \(r_n(\hat{\theta}_{21}, F_n) = \int_{\theta_{21}}^{\hat{\theta}_{21}} x dF_n(x),\)

\[
E_F(X I_{\{X > \theta_{21}\}}) = E_F(X) - \int_{-\infty}^{\hat{\theta}_{21}} x dF(x)
\]

\[
= E_F(X) - F(\theta_{21}) \int_{-\infty}^{\theta_{21}} \left[ x / F(\theta_{21}) \right] dF(x)
\]

(E.28)

\[
= E_F(X) - F(\theta_{21}) G(Y).
\]

As previously shown, \(E(X_n) \xrightarrow{D} E(X)\) and \(F_n(\theta_{21}) \xrightarrow{P} F(\theta_{21})\), and by the uniform integrability of \(Y\) we have \(E(Y_n) \xrightarrow{D} E(Y)\) by (Billingsley, 1995, corollary to Thm. 25.12, p. 338). It remains to show that \(r_n(\hat{\theta}_{21}, F_n) \xrightarrow{p} 0\).
Now,
\[ |r_n(\hat{\theta}_{21}, F_n)| = \left| \int_{\hat{\theta}_{21}}^{\theta_{21}} x \, dF_n(x) \right| \leq \int_{\hat{\theta}_{21}}^{\theta_{21}} |x| \, dx, \]  
(E.29)
and \( \hat{\theta}_{21} \to \theta_{21} \) implies the right hand side of (E.29) converges in probability to zero. It follows from Slutsky’s theorem that \( E_{F_n}(X_n I_{\{X_n > \hat{\theta}_{21}\}}) \to E_F(X I_{\{X > \theta_{21}\}}) \). The convergence of cell (2, 3) again follows from the convergence in probability of the parameter estimates by Slutsky’s theorem.

We finish by considering cell (2, 4), the remaining cells can be proven by similar argument. Write the expectations as:
\[ E_{F_n}(X_n(X_n - \hat{\theta}_{21})) = E_{F_n}(X_n^2) + \hat{\theta}_{21} E_{F_n}(X_n) - \hat{\theta}_{21} \int_{-\infty}^{\theta_{21}} x \, dF_n(x) \]
\[ = E_{F_n}(X_n^2) - \hat{\theta}_{21} \int_{-\infty}^{\theta_{21}} x \, dF_n(x) + \hat{\theta}_{21} \int_{\hat{\theta}_{21}}^{\theta_{21}} x \, dF_n(x) \]
(E.30)
where \( r_{n1}(\hat{\theta}_{21}, F_n) = \int_{\hat{\theta}_{21}}^{\theta_{21}} x^2 \, dF_n(x) \) and \( r_{n2}(\hat{\theta}_{21}, F_n) = \int_{\hat{\theta}_{21}}^{\theta_{21}} x \, dF_n(x) \).

\[ E_F(X(X - \theta_{21}) I_{\{X > \theta_{21}\}}) = E_F(X^2 I_{\{X > \theta_{21}\}}) + \theta_{21} E_F(X) I_{\{X > \theta_{21}\}} \]
\[ = E_F(X^2) - F(\theta_{21}) \int_{-\infty}^{\theta_{21}} x^2 / F(\theta_{21}) \, dF(x) \]
\[ = E_F(X^2) - F(\theta_{21}) F(\theta_{21}) \int_{-\infty}^{\theta_{21}} x / F(\theta_{21}) \, dF(x) \]
\[ = E_F(X^2) - F(\theta_{21}) E_G(Y^2) + \theta_{21} E_F(X) - \theta_{21} F(\theta_{21}) E_G(Y). \]  
(E.31)

By similar argument as used in proving convergence of cell (2, 3), \( E(X_n) \to E(X) \), \( E(X_n^2) \to E(X^2) \), \( E(Y_n) \to E(Y) \), \( E(Y_n^2) \to E(Y^2) \), \( F_n(\theta_{21}) \to F(\theta_{21}) \), \( \hat{\theta}_{21} \to \theta_{21} \), \( r_{n1}(\hat{\theta}_{21}, F_n) \) \( \to 0 \), and \( r_{n2}(\hat{\theta}_{21}, F_n) \) \( \to 0 \). It follows that \( E_{F_n}(X_n(X_n - \hat{\theta}_{21}) I_{\{X_n > \hat{\theta}_{21}\}}) \to E_F(X(X - \theta_{21}) I_{\{X > \theta_{21}\}}) \), and hence we have the converge of the entries in cell (2, 4).
BIBLIOGRAPHY
Bibliography


