2018-01-01

Hierarchical Multiplicity Control Methods For Linear Models

Dimuthu Dilshan Fernando
University of Texas at El Paso, ddfernando@miners.utep.edu

HIERARCHICAL MULTIPLICITY CONTROL METHODS FOR LINEAR MODELS

DIMUTHU FERNANDO

Master's Program in Mathematical Sciences

APPROVED:

Amy Wagler, Ph.D., Chair

Xiaogang Su, Ph.D.

Jeffrey T. Olimpo, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

HIERARCHICAL MULTIPLICITY CONTROL METHODS FOR LINEAR MODELS

by

DIMUTHU FERNANDO, B.Sc.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2018

Acknowledgements

In writing this thesis the inspiration and motivation come from different directions and sources. Many people have contributed in socially, academically, and spiritually to the successful accomplishment of this research. Though it might be difficult to address all the individuals by name, they all have my heartfelt gratitude.

Most especially, my deep appreciation goes to my thesis adviser Dr.Amy Wagler for the useful advices and necessary suggestions given when writing this research study. I thank you for working with me patiently throughout the time and for supporting my writing of this research. Your critical comments, timely feedback, stimulating suggestions, and commitment to my writing made this thesis what it is.

I also wish to thank the other members of my committee, Dr. Xioagang Su of Department of Mathematical Sciences and Dr. Jeffrey Olimpo of Department of Biological Sciences, both at The University of Texas at El Paso. Their suggestions, comments and additional guidance were invaluable to the completion of this work.

I must extend my gratitude to Dr. Ori Rosen, Director of Graduate Studies in Statistics, all the Professors and members of staff, at the Department of Mathematical Sciences to encouraging and helping me in various ways to fulfill this task.

Finally I will express my gratitude to my wife, beloved parents for their unfailing love and courage given to make this task completed, my friends for their corporation for the completion of the thesis.

NOTE: This thesis was submitted to my Supervising Committee on the May 31, 2018.

Abstract

Hypothesis testing is a commonly used statistical inference technique on which a statement of the population is investigated through the evidence from a representative sample of the population. With simultaneous testing of more than one null hypotheses need for an appropriate multiple comparison method is essential. With motivation from the study of Bogomolov et al. (2017) we have modified a multiple comparison tree structure to build the required comparisons and focus on controlling the FWER (Family Wise Error Rate) using the Bonferroni procedure. The proposed method has advantages such as controlling the global error rates separately at each level, families of hypotheses at high resolution are tested only when their parent hypotheses are rejected.

In this study, a level restricted method is used to control the FWER at each level and a simulation study is performed to justify the proposed method. Additionally, the proposed method was applied to two real data sets in an educational setting to make multiple comparisons.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In modern scientific research the analysis of large data sets has become very common with the use of advanced methods for statistical inference. Hypothesis testing is a commonly used statistical inference technique in which a statement of the population was investigated through the evidence from a representative sample of the population. The goal of hypothesis testing is to determine whether there is sufficient evidence to reject the null hypothesis (H_0) a claim regarding a population parameter in favor of an alternative hypothesis (H_a).

1.1 Multiple Comparisons in Hypothesis Testing

When we consider educational or genetic data, this generates a multiplicity of data leading to perform many multiple statistical hypothesis tests. In most research settings, more than one (H_0) is tested at a time. With simultaneous testing of more than one null hypotheses (H_0)'s the need for appropriate multiple comparisons correction method is essential. There are two types of errors that researchers make during hypothesis testing. One is the **type I error** which occurs when the researcher rejects (H_0) when its true. This is also referred as a false positive and its probability is controlled by α. The other error is **type II error** which occurs when (H_0) is false but the researcher fails to reject it. This is also referred as a false negative.

Multiple comparisons methods are used to control the overall type I error rate. Each of the individual tests or confidence intervals has a type I error rate α_i that can be controlled by the experimenter. If we conduct tests together as a family then a combined type I error rate can be computed for the family of tests. The family wise error rate (FWER) is the

probability of rejecting at least one true null hypotheses in a series of hypothesis tests. When a family of tests contains more true null hypotheses, the many true null hypothesis will be rejected and increase the type I error rate of the family. Another class of approaches mainly focus on not reducing the family wise error rate but on controlling the expected proportion of false positives which can be referred to as False Discovery Rate (FDR). Hence it is very important to use procedures to control type I error rates or false positives when proceeding with multiple comparisons. When considering the the adjustment methods for multiple testing there isn't a universally accepted way to control the problem of multiple testing, but researchers determine which approach best suits the setting.

1.2 Methods of Adjustments for Multiple Testings

The first step in controlling for multiple comparisons is to quantify the probability of obtaining false positives. Many multiple testing adjustment methods exist for controlling error rates.

- The control of the FWER is important when a conclusion from various individual hypothesis tests are likely to be erroneous. We can identify Bonferroni, Holm and Tukey methods which focus on controlling FWER (the probability of atleast one type I error). There are two general types of family wise error rate corrections.
  - **Single step**- Equivalent adjustments made to each p-value
  - **Sequential**- Adaptive adjustment made to each p-value
- In some occasions we are more concerned about false positives and its essential to monitor FDR. Benjamini and Hochberg (1995) suggest a different point of view on error controlling for multiple testing. FDR is defined as the erroneous rejections among all rejections. If all tested null hypotheses are true, controlling the FDR is the same as controlling FWER.

Table 1.1: Notations representing the number of decisions when testing a set of null hypotheses (adapted from Benjamini and Hochberg (1995))

Truth/Decision	**Declared non-significant**	**Declared Significant**	**Total**
True Null Hypothesis	U	V	N
Non-true Null Hypothesis	T	S	M-N
Total	Ac	Re	M

1.3 Strategies for Controlling FWER using Tree Structures

Different strategies to control FWER in a tree structure are possible. The goal is to perform meaningful comparisons at increasing levels of resolution. Below are some methods for controlling FWER in a tree structure. For the tree structure in Figure 1.1 threshold value for FWER calculated from each of these methods are mentioned in the brackets.

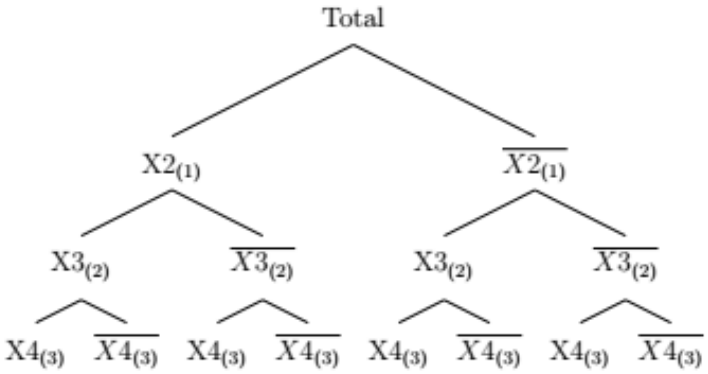


Figure 1.1: Tree structure with 3 levels

- **Outer node**- The interest lies in the discoveries that are not parent to any other families. This can be also referred to as controlling at the highest resolution level. (Compare the p-values of the level 3 hypotheses with level of significance $\alpha/4$).

- **Full tree**- We are interested in the entire set of discoveries and controlling by using traditional method of dividing the level of significance by the number of tested hypotheses. (Compare the p-values of the 7 hypotheses with level of significance $\alpha/7$).

- **Level restricted**- The researcher is interested only in the discoveries on a specific level of the tree. The level restricted FWER is also a model for the FWER of the entire study when Bonferroni procedure was applied separately to several families of hypotheses. (Compare the p-values of the level 1 hypotheses with level of significance α, Compare the p-values of the level 2 hypotheses with level of significance $\alpha/2$ and Compare the p-values of the level 3 hypotheses with level of significance $\alpha/4$).

The **research question** for this study is: Does the level restricted method still control the global FWER?

Chapter 2

Literature Review

2.1 FWER Correction Methods for Multiplicity Adjustment

Many procedures have been developed to control the FWER. These techniques are usually used when it's important not to make any type I errors at all.

2.1.1 Bonferroni Procedure

The Bonferroni technique is the most simple and widely used method among multiple comparison procedures. It uses a single step FWER correction method on which equivalent adjustments made to each p-value. Let K be the number of null hypothesis to be tested and P_i be the p-value for testing H_{0i}. Using this procedure, we can obtain simultaneous 1-α confidence intervals by constructing individual confidence intervals with a coverage probability of $1 - \alpha/k$ and will reject H_{0i} if $P_i < \alpha/k$. This procedure run each test with a level of α/k and controls the strong family wise error rate.

2.1.2 Holm Procedure

The Holm procedure can be viewed as a Sequential FWER correction method in which adaptive adjustments are made to each p-value. Suppose we are conducting m tests α is the type I error.

- Order the unadjusted p-values such that $p_1 \leqslant p_2 \leqslant ... \leqslant p_m$

- Then reject $H_{0(i)}$ if

$$p_{(j)} \leqslant \alpha/(m-j+1) \; for \; all \; j = 1, 2, ...i.$$

2.2 Controlling the False Discovery Rate

Benjamini and Hochberg (1995) advocate a different approach for multiple testing which controls the expected proportion of falsely rejected hypotheses (FDR) among all rejected tests. In many occasions, we can live with a certain number of false positives and in these situations maintaining FDR control is essential.

Suppose there are a total of R tests claimed as significant and from these S tests are truly significant and remaining V are non significant (i.e. $R = S + V$). Since we never know the exact count of true positives and false positives both S and V are considered as unobserved random variables. False discovery rate can be defined as

$$FDR = E(V/R)$$

which shows the expected value of the proportion of false positives from the total count of null hypotheses rejected. Let m be the number of tests and δ is level of controlling FDR

- Order the unadjusted p-values such that $p_1 \leqslant p_2 \leqslant ... \leqslant p_m$
- Find the test with the highest rank j for which the p-value (p_j) is less than or equal to $(j/m) * \delta$.
- Declare the tests of rank 1, 2, ..., j as significant

$$p(j) \leq \delta * (j/m)$$

2.3 Multiple Comparisons in Hierarchical Modeling

Bogomolov et al. (2017) proposed a new multiple testing procedure which can control error rates at multiple levels of the data. The required hypotheses which needs to be tested are organized hierarchically in a tree structure with the required error rates specified at each level. Below will be a graphical representation of how the layout is set to test the hypotheses extracted.

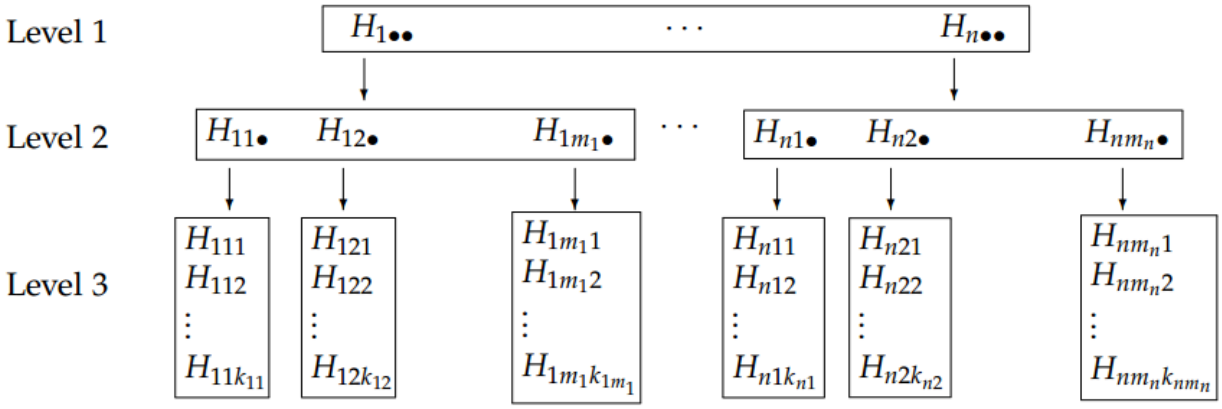


Figure 2.1: Hierarchical structure of the hypotheses

In Figure 2.1, a rectangular box identifies a set of related hypotheses and are always tested simultaneously. Furthermore in the above graph, each hypothesis at level ι is parent to a family of hypotheses at level $\iota + 1$. A motivating example for this hierarchical tree structure comes from eQTL (expression Quantitative Trait Loci) studies where,

- Level 1 hypothesis $H_{i..}$ corresponds to that SNP (Single Nucleotide Polymorphism) i is not involved in the regulation of the expression of any gene in any tissue.
- Level 2 hypothesis $H_{ij.}$ states that SNP i does not affect the expression of gene j in any tissue.
- Level 3 hypothesis H_{ijt} specifies that SNP i is not involved in the regulation of gene j

in tissue t.

Discoveries in Level 1 correspond to the identification of important SNPs, in Level 2 of the affected genes, and in Level 3 of the tissues where this regulation is active. This testing procedure has many positive outcomes such as:

- It can accommodate any number of levels hierarchically
- It can control the errors within the families at the given level
- It can target different error rates varying across different families

The work of Yekutieli (2008) applied hierarchical testing into micro array experiments that tested expression levels of genes in brain regions of mice. In the hierarchical approach, log-linear models are fitted for contingency tables containing two rat types. The loglinear models were built hierarchically in which the Benjamini and Hochberg procedure was applied to the main effects model, the two way interaction model and the three way interaction model.

This approach uses increasing levels of precision to increase power in data settings where the hypotheses can be meaningfully grouped. They have used FDR as the method for controlling the error rate and defined 3 levels of FDR correction for the tree.

Gelman et al. (2012) present a Bayesian multilevel modeling framework on which they are able to find a more reliable point estimate. Fitting a multilevel model shifts point estimates and their corresponding intervals toward each other, this process is often referred to as partial pooling. Whereas classical procedures typically keep the point estimates stationary, adjusting for multiple comparisons by making the intervals wider.
When considering multilevel estimate's, they make comparisons appropriately more conservative, so that the resulting intervals for comparisons are more likely to include zero. At the same time this adjustment does not decrease the power of detecting true differences.

The work of Yi et al. (2014) presented a hierarchical modeling approach in a genetic framework. The proposed hierarchical models simultaneously fit as many variables as pos-

sible and shrinks unimportant effects towards zero by using the hierarchical model. A hierarchical model yields to more accurate parameter estimates while addressing the multiple comparison problem by reducing the effective number of genetic effects and the number of statistically significant effects. In their study, they used a traditional Bonferroni correction. Bonferroni is appropriate since hierarchical modeling influences the dependence among parameters and thus decreases the number of independent hypothesis tests.

Lindquist and Mejia (2015) discussed approaches which can be used for correcting for multiple tests. They used multiple comparisons to analyze functional magnetic resonance imagining (fMRI) data and discussed the in adequateness of using the same method to threshold a test statistic when dealing with families consisting of many tests. When the chosen threshold is too conservative, this results with losing power to detect meaningful results. On the other hand, if the chosen cutoff is too liberal, this results in an excessive numbers of false positives.

Lindquist and Mejia (2015) illustrated the difference of using FWER and FDR using a fMRI study of 100,000 brain voxels tested using a threshold of α=0.001. They explained that this will lead to a significant percentage of false positives and no way of identifying those voxels. They have used FDR and controlled it at q=0.05 to overcome this issue which helps to increase the reliability of the results.

Chapter 3

Methodology

In the following, a modification of the Yekutieli and Bogomolov hierarchical methodology for multiple comparisons is developed with a focus on controlling FWER. The method will be extended to an educational data research setting, demonstrating the broad applicability of this multiple comparison method.

3.1 FWER Control using a Tree Structure

In this study we have used a new notion of level specific FWER that reflects a hierarchical order where the families of hypotheses at the highest resolution level are tested only if their parent hypotheses are rejected. The level 1 family (F^1)is always tested and from level 2 onwards, a family of hypotheses are tested only if its parent hypothesis is rejected which is similar to the work of Yekutieli (2008).

Let H_{0i} be the hypothesis that there is no effect from outcome interested i , where i =1,2,...,m. here m represents the total number of hypotheses tested simultaneously. The family wise null hypothesis (H_0) denote that there is no effect of the outcome interested for all m hypotheses. Mathematical notation=

$$H_0 = \cap_i H_{0i}$$

Thus if we reject a single H_{0i} will reject the family wise null hypothesis and a false positive from any test would increase family wise error (FWE). According to the work of Lindquist and Mejia (2015) FWER is defined in the following manner. Assuming (H_0) is true they needed to control the probability of falsely rejecting (H_0) at controlled by some fixed value

α. The mathematical expression is represented as follows.

$$FWER = P(\cup_i \{T_i \geq u_\alpha\} | H_0) \leq \alpha$$

In other words, the FWER represents the probability under H_0 that any of the m test statistics (T_i) take the value above the threshold (u_α).

In this study we have defined the average FWER and FWEP as below. Here w_i^l are random weights for $(FWEP)_l$ that sum to 1. The weights can be simply equal to the inverse of the total number of families tested at level l.

$$\overline{FWEP}^l = \sum_{F_i^l \ is \ tested} w_i^l (FWEP)_{F_i^l} \ for \ l = 1, 2, 3$$

$$FWER^l = E(\overline{FWEP}^l)$$

where FWEP (family wise error proportions) defined as below when the weights are equal to the inverse of the total number of families tested at level l.

$$\overline{FWEP}^l = \frac{\sum_{F_i^l \ is \ tested} (FWEP)_{F_i^l}}{\mid \{i \ : \ F_i^l \ is \ tested\} \mid}$$

where F_i^l = number of families tested at l^{th} level

Corrollary 1. For hierarchical testing procedure for the hypotheses in Figure 2.1:

- For the l^{th} family tested $\mid k_l \mid$ denotes the number of hypotheses tested in the l^{th} level. p_{k_l} denotes the p-value for the l^{th} family and k^{th} test.

$$I(p_{k_l} < \alpha_l) = 1 \ if \ p_{k_l} < \alpha_l \ otherwise = 0$$

- The Family wise error rate for each level can be calculated by getting the expectation of below formula

$$FWER^l = E[\frac{I(p_{k_l} < \alpha_l)}{\mid k_l \mid}]$$

3.2 Proposed Methodology

3.2.1 Building a Linear Model

Initially, we build a linear model using the variables. The variables which were used to partitioning of the data are sometimes known before hand (as in the study of Bogomolov et al. (2017)) or we can use data driven methods to identify the splits. At the first stage, significance of the terms of the main effects models was checked using an unadjusted α level. At the second stage, two-way interactions were tested with each of the significant main effects from the first stage. In the third stage, three-way interactions with each of the significant two-way interaction were tested and so on.

3.2.2 Formulating the Tree Structure

After identifying the splitting variables and building the required linear models, we can start creating the tree structure for hypothesis testing. The hypotheses that are required to be tested are organized in a tree structure and the FWER is controlled separately at each level.

3.2.3 Defining the Level of Significance for Hypothesis Testing

The level restricted method is used to control the FWER at each level. The families of hypotheses at high resolution are tested only when their parent hypotheses are rejected. This is an advantage since the researcher will not investigate higher resolution tests when splitting variables are non-significant. From this proposed method we focus on FWER control for complex structures.

Given the proof in section 3.1 we know that the level restricted procedure will control the global FWER. It remains to be shown whether the level restricted method is conservative in these settings. If it is too conservative, it will result in hiding true findings and will lose power.

Chapter 4

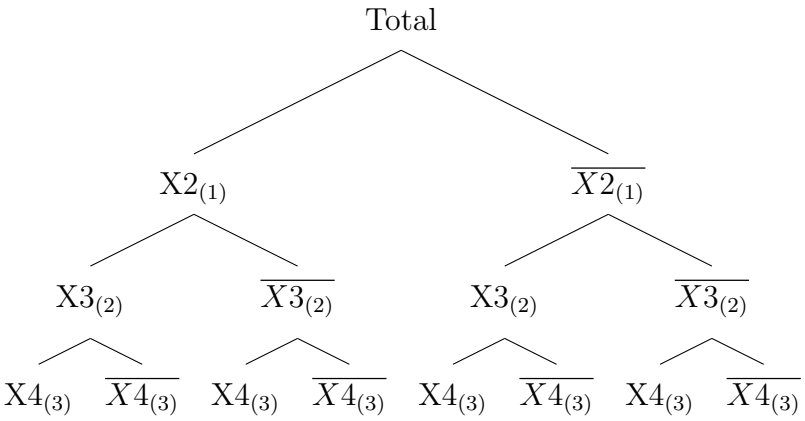
Simulations

4.1 Creating Variables and Tree Structure

The response variable and the control variate were generated using standard normal distribution. X2, X3 and X4 are binary variables generated using binomial distribution. The simulations have been carried out for different sample sizes and on different levels of significance (α).

4.1.1 Created Tree Structure for Simulations

Below is the tree structure for checking the effect of each variable variable and the effect of the each contrast was tested using any α level.



4.2 Hypotheses to be Tested in Tree Structure

There are three levels in this set-up, which correspond to the comparisons outlined below.

- Level 1 hypothesis

$$H^1(0) : \alpha_{X2} = 0$$

- Level 2 hypothesis

$$H_1^2(0) : \alpha_{(\overline{X3},\ \overline{X2})-(X3,\ \overline{X2})} = 0 \quad -C1$$

$$H_2^2(0) : \alpha_{(\overline{X3},\ X2)-(X3,\ X2)} = 0 \quad -C2$$

- Level 3 hypotheses

$$H_1^3(0) : \alpha_{(\overline{X3},\ \overline{X2},\ X4)-(\overline{X3},\ \overline{X2},\ \overline{X4})} = 0 \quad -C3$$

$$H_2^3(0) : \alpha_{(\overline{X3},\ X2,\ X4)-(\overline{X3},\ X2,\ \overline{X4})} = 0 \quad -C4$$

$$H_3^3(0) : \alpha_{(X3,\ \overline{X2},\ X4)-(X3,\ \overline{X2},\ \overline{X4})} = 0 \quad -C5$$

$$H_4^3(0) : \alpha_{(X3,\ X2,\ X4)-(X3,\ X2,\ \overline{X4})} = 0 \quad -C6$$

The proposed tree structure was used at each level of the tree and the significance of the contrasts were tested using the defined level restricted α level.

4.3 Simulation Results

Table 4.1 to Table 4.7 represents the estimated power of the test for each contrast. The bold highlighted cells represent the values obtained using the full tree based method. Since we have 7 contrasts, in this scenario the full tree based method means comparing each contrast with the level of $(\alpha/7)$. With the created structure, the expectation is to obtain contrast 2 is more significant than contrast 1 and within contrast 2, contrast 6 to be more significant than contrast 4.

4.3.1 Simulation Results for Initial Split Variable X2

Table 4.1 represents the simulation results for the level 1 hypothesis and the full tree based results are highlighted in bold. The level 2 hypothesis are only tested for significant level 1 hypothesis based on the tree based cut off level.

Table 4.1: Simulation Results for intial split variable

nsim	n	alpha	tree based %	**full tree %**
1000	100	0.05	0.529	**0.254**
1000	200	0.05	0.849	**0.589**
5000	100	0.05	0.534	**0.246**
10000	200	0.05	0.825	**0.578**
1000	200	0.1	0.893	**0.671**
10000	200	0.1	0.898	**0.685**
1000	200	0.2	0.95	**0.749**
10000	200	0.2	0.951	**0.773**

4.3.2 Simulation Results for α=0.05

For α=0.05 the simulations were done for n=100, n=200 while increasing the number of simulations. For any sample size, contrast 2 has been more significant than contrast 1 and within contrast 2 contrast 6 has been significant more times than contrast 4. There is less estimated power in the bold highlighted cells since it compares each contrast with ($\alpha/7$) level of significance. When the sample size increases the estimated power of the test for each contrast also increase.

Table 4.2: Estimated power of the test for α=0.05, n=100, nsim=5000

Contrast 1				Contrast 2			
0.0122		**0.0026**		0.811		**0.6784**	
Contrast 3		Contrast 5		Contrast 4		Contrast 6	
0	**0**	0	**0**	0.0048	**0.0008**	0.2946	**0.1646**

Table 4.3: Estimated power of the test for α=0.05, n=200, nsim=10000

Contrast 1				Contrast 2			
0.0142		**0.003**		0.99		**0.9735**	
Contrast 3		Contrast 5		Contrast 4		Contrast 6	
0	**0**	0	**0**	0.0062	**0.0017**	0.7466	**0.6098**

4.3.3 Simulation Results for α=0.1

Table 4.4 and Table 4.5 show that when α is increased up to 0.1 we can see an increment of estimated power of the test for each contrast. The same pattern of contrast 2 is more significant and within contrast 2, contrast 6 is more significant which can be observed in results with α=0.1.

Table 4.4: Estimated power of the test for α=0.1, n=200, nsim=1000

Contrast 1				Contrast 2			
0.031		**0.005**		0.993		**0.982**	
Contrast 3		Contrast 5		Contrast 4		Contrast 6	
0	**0**	0	**0**	0.016	**0.004**	0.814	**0.694**

Table 4.5: Estimated power of the test for α=0.1, n=200, nsim=10000

Contrast 1				Contrast 2			
0.026		**0.0075**		0.9954		**0.985**	
Contrast 3		Contrast 5		Contrast 4		Contrast 6	
0.0002	**0**	0.0003	**0**	0.013	**0.0042**	0.8187	**0.6923**

4.3.4 Simulation Results for α=0.2

Table 4.6 and Table 4.7 shows that when α increased up to 0.2 we can see a increment of estimated power of the test for each contrast. As found earlier the same pattern of contrast 2 is more significant and within contrast 2 contrast 6 is more significant can be observed in results with α=0.2.

Table 4.6: Estimated power of the test for α=0.2, n=200, nsim=1000

Contrast 1				Contrast 2			
0.055		**0.013**		0.996		**0.989**	
Contrast 3		Contrast 5		Contrast 4		Contrast 6	
0.001	**0**	0.001	**0**	0.023	**0.008**	0.882	**0.772**

Table 4.7: Estimated power of the test for α=0.2, n=200, nsim=10000

Contrast 1				Contrast 2			
0.0507		**0.016**		0.9983		**0.9925**	
Contrast 3		Contrast 5		Contrast 4		Contrast 6	
0.0008	**0.0003**	0.0005	**0.0001**	0.0288	**0.0068**	0.8828	**0.7676**

Chapter 5

Discussion and Applications

5.1 Discussion

When a family of tests contains high proportion of true null hypotheses, many true null hypothesis will be rejected and increase the type I error rate of the family. When considering the adjustment methods for multiple testing there isn't a universally accepted way to control the problem of multiple testing. However in practice, researchers determine which approach best suits the setting. In this study, a hierarchical multiplicity correction method is applied to control FWER in complex data structures. Simulation results indicate that power of tests are increased by employing the proposed hierarchical tree structure method. Power of the tests are significantly higher than the results of full tree based method applied at all levels of the tree structure. After the simulation study, the proposed method was applied to two real data sets.

5.2 Limitations and Future Work

In this study, use of recursive partition models may induce over-optimism for the set of mean comparisons. Use of theory based partition methods are recommended. The use of Bonferroni adjustment for controlling FWER assumes the resulting set of mean comparisons are independent. Using this methodology and due to the hierarchical structure, the comparisons are most certainly dependent. Future studies should incorporate a multiplicity control procedure that assumes dependencies of the test end points.

5.3 Applications

5.3.1 Research Driven Course Effect on Student Performance

Student Behaviors and Instructional Practices on Laboratory Courses

Laboratory practices play a major role in undergraduate student education in fields of Science, Technology, Engineering and Mathematics (STEM) and for non-STEM. The amount of the student learning in any educational field mainly depends on the effectiveness of the instructors. By identifying important instructor behaviors, which improves the student performance or focus of the student towards the course, will be beneficial to increase the retention of the student in the subject field. Various methods have been used to identify instructional behaviors in the laboratory such as student surveys, audio/video recordings of students working in the labs, one -to-one meetings with the students. Students previous educational experiences and background knowledge about the content area impacts the student behaviors as well as their feedback on instructional activities.

Student's experiences in their undergraduate science courses plays a major role when it comes to student retention in STEM majors. Hence, by optimizing the instructor effectiveness and by identifying the key factors which has a effect on the student outcomes, it will help to improve the student preference on selecting a STEM major.

Instructional practices play a major role in student retention in any field of education. The work of Kendall and Schussler (2013) undergraduate experiences in lower division science courses are key factors for a student when choosing a science major. The courses normally includes lecture portions which is taught by a faculty member and some smaller sections such as discussions and laboratory sessions taught by graduate teaching assistants (GTAs).

In studies about student perception of instructors, GTAs have been identified with positive indicators such as being engaging, approachable, informal, relaxed, interactive and understanding. Meanwhile, they have also been identified with negative indicators, including a lack of confidence, knowledge, experience and authority.

The work of Bangera and Brownell (2014) emphasizes the current approaches which improve diversity in scientific research targets on graduating from STEM majors, but graduation with only a STEM undergraduate degree is not enough for graduate school entry. Further, they discussed undergraduate independent research experiences, which is becoming a key factor when students are seeking graduate school admission. They suggest CUREs should be added as required introductory courses, as it will give students the chance to engage in authentic research in the undergraduate stage.

Olimpo et al. (2016) developed a mixed-methods approach to examine the student engagement in scientific research following implementation of a novel CURE in an introductory cell/molecular biology course. Findings indicate that CURE students exhibited more expert-like outcomes on these constructs relative to their non- CURE participants, including those in areas related to self-efcacy, self-determination and problem-solving strategies.

When considering students enrolled in introductory cell/molecular and organismal biology, CUREs indicates a stronger interest in scientific research as compared with their colleagues completing traditional lab coursework.

Past Studies on Factors Effecting Student Performance

Over the past decades, researchers have examined the key factors of students' performance on standardized achievement tests. The work of Emilio et al. (2004), income inequality is strongly related to differences in years of schooling among individuals.

The substance of Guimarães and Sampaio (2013) article is that they analyzed the effect of variables like family income, gender, race, religion and high school attended on entering a university in Brazil. They studied how determinants such as parents education level and family income will affect the average scores of students in university entrance tests. Furthermore, they investigated the interaction between income and the probability of attending public schools and private tutoring classes.

The gender of the student may also be a factor in determining student performance. Childhood training and experience, gender differences in attitudes, parental and teacher

expectations and behaviors, differential course taking and biological differences between the sexes may all be instrumental in giving rise to gender differences in achievement Feingold (1988). Gender disparity in various spheres of public life and the patriarchal social structure in Turkey may also lead to poor academic performance among female university students.

At the primarylevel, female students are generally found to get better course grades but perform worse than males in achievement tests like Dayioğlu and Türüt-Aşik (2007). This pattern is explained by the better work habits and better language abilities of females. The work of Young and Fisler (2000) explain that better performance of males in SAT-M by referring to the different socio-economic background of students. They note that males generally come from households where the parents' socio-economic status is higher. Others have argued that the content of the test or of its administration favors males Bridgeman and Wendler (1991).

Data Processing for Student Performance data set

The first data set consists of 146 student records of the undergraduate students who are enrolled in either Cure or Non- Cure lab course type. It consists of data collected over 5 fields and prior to the study, there is a necessity of possible modifications to the data set.

Missing values are identified in each variable and removed the rows which consist of those values. After this step, 111 data values remained in the data set. The data set consists of student related and instructor related variables. Since each of the variable type consists of a large number of variables, a dimension reduction of the variables were needed to included them in model. Table 5.1 represents the variable types and count of variables in each section.

Table 5.1: Number of Variables in each type

Class	**Type**	**No Of VAR**
Student	Demographic	10
Student	Performance	23
Student	Behaviors	52
Instructor	Demographic	7
Instructor	Behaviors	44

Table 5.2 summarizes the base categories of the Binary variables which has been used in the model.

Table 5.2: Reference Categories for binary variables

Variable	**Notation**	**Levels**	**Code**
Ethnicity	URM	Caucasian	1
		Other	0
First Language	ESL	Native English Speaker	1
		Other	0
Highest Biology Course	HS	Advanced Bio	1
		General Bio	0
Gender	Gender	Male	1
		Female	0
SCI 1301 Enrollment	SCI	Yes	1
		NO	0
Academic Program	STEM	Stem	1
		Non-Stem	0
Over Confidence	OC	Over Confident	1
		Non-Over Confident	0
Lab Course Type	Lab Course Type	Cure	1
		Non-Cure	0

Dimension reduction of variables

The principal Component Analysis (PCA) method was applied for Student performance, Student behaviors and Instructor behaviors sections and selected principal components are used as predictors when model fitting is done.

Principal Component Analysis for Student Performance

Student performance is the response variable which was used in model fitting and it is created by performing principal component analysis on the student performance variables. Only the shift variables were used to create the response variable and the shifts were taken as post-pre.

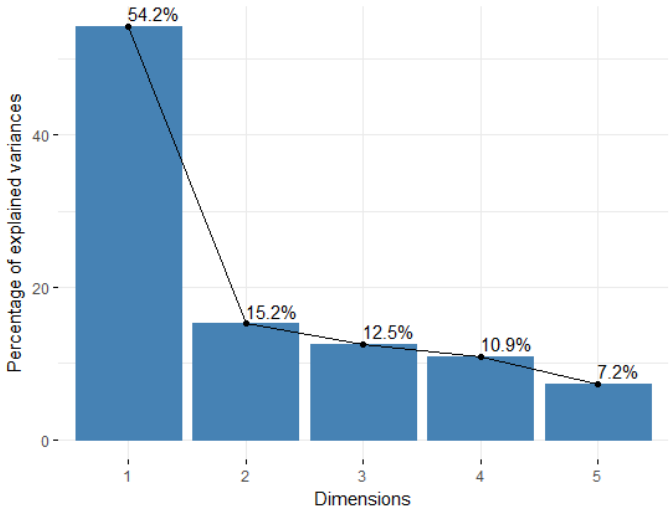


Figure 5.1: Scree plot for the Student Performance variables

Table 5.3: Percentage of Variance explained by each Dimension for Student Performance

Dimension	**eigenvalue**	**variance %**	**cumulative variance %**
Dim.1	2.708	54.157	54.157
Dim.2	0.76	15.197	69.354
Dim.3	0.625	12.505	81.859
Dim.4	0.545	10.898	92.757
Dim.5	0.362	7.243	100

Since first principal component explains around 54% of the variance it was selected to represent student performance variables.

- This principal component is positively loaded with all five original variables and can be described as a summary of the motivation factors of the student. These variables are shift (measured as post-previous) variables and they are, Intrinsic Motivation, Career Motivation, Self Determination, Self Efficacy and Science Identity.

Principal Component Analysis for Student Behaviors

There were originally 52 variables in the student behaviors section and Principal Component analysis was performed to reduce the dimension of the variables.

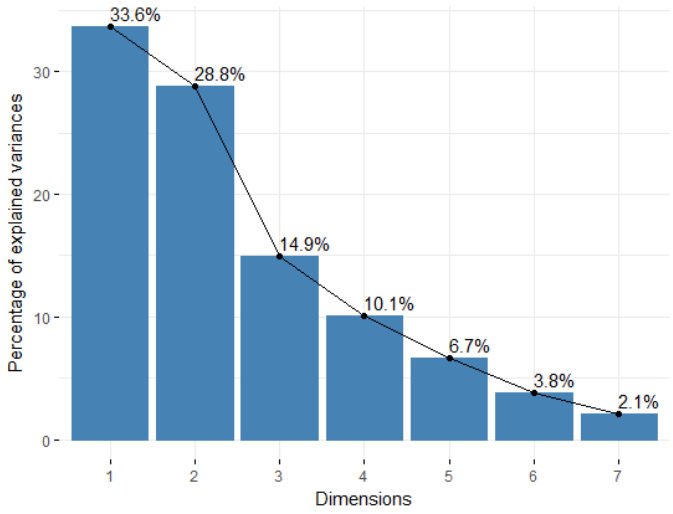


Figure 5.2: Scree plot for the Student Behaviors variables

Table 5.4: Percentage of Variance explained by each Dimension for Student Behaviors

Dimension	**eigenvalue**	**variance %**	**cumulative variance %**
Dim.1	17.492	33.638	33.638
Dim.2	14.98	28.809	62.447
Dim.3	7.765	14.932	77.379
Dim.4	5.237	10.072	87.45
Dim.5	3.468	6.668	94.119
Dim.6	1.979	3.807	97.926
Dim.7	1.079	2.074	100

Since the first two principal components explains around 63% of the variance, the first two principal components were selected to represent student behavior variables.

- PC 1 for the student behaviors was mainly classified as behaviors with high level student interactions vs behaviors with low level student interactions. (positively loaded on high level interaction student behaviors and negatively loaded on low level interaction student behaviors). Examples for behaviors with high level student interactions are engaged in lab, student asking questions in class and students engaging in groups. Examples for behaviors with low level student interactions are listening, taking a test or quiz and students waiting.
- PC 2 for the student behaviors was mainly classified with the student behaviors in the classroom vs student behaviors in the lab (positively loaded with student behaviors in class and negatively loaded with student behaviors in the lab).

Principal Component Analysis for Instructor Behaviors

There were originally 44 variables in the Instructor behaviors section and Principal Component analysis was performed to reduce the dimension of the variables.

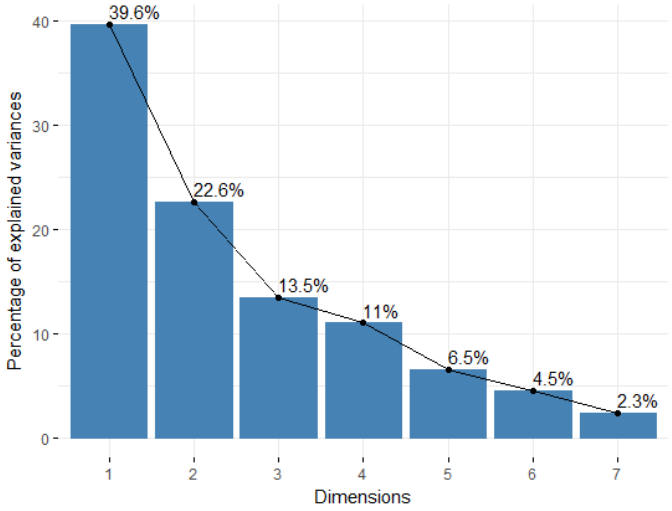


Figure 5.3: Scree plot for the Instructor Behaviors variables

Table 5.5: Percentage of Variance explained by each Dimension for Instructor Behaviors

Dimension	**eigenvalue**	**variance %**	**cumulative variance %**
Dim.1	17.411	39.571	39.571
Dim.2	9.952	22.617	62.188
Dim.3	5.923	13.461	75.65
Dim.4	4.833	10.984	86.634
Dim.5	2.861	6.502	93.136
Dim.6	1.993	4.53	97.665
Dim.7	1.027	2.335	100

Since the first two principal components explains around 62% of the variance, the first two principal components were selected to represent Instructor behavior variables.

- PC 1 for the instructor behaviors was mainly classified as teacher centered vs student centered. (positively loaded on teacher centered variables and negatively loaded on student centered variables). Examples for teacher centered variables are lecturing, real time writing on the board, conducting demo. Examples for student centered variables are monitoring class groups and performing administrative duties.
- PC 2 for the instructor behaviors was mainly classified as high level vs low level of one-on-one conversations. (positively loaded on high level student one-on-one conversation variables and negatively loaded low level student one-on-one conversation variables). Examples for high level one-on-one conversation variables are providing feedback on activity, talking to students individually. Examples for low level one-on-one conversation variables are lecturing, real time writing on board and posting lab related questions.

Model Building Process for Student Performance Data

Step wise selection procedure is used for model building and the initial model included all the explanatory variables taking student performance as the response. P-value of the explanatory variables and AIC value was considered as model selection tools in each step.

Model 1 with all Covariates taking Student Performance as the Response

Initially starts with the model containing all the variables and this model was named as model 1. The predictor variable with highest P-value was removed from the model also considering the AIC value at each step.

Table 5.6: Parameter estimates for model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6873	0.5586	1.23	0.2216
URM	0.1751	0.2571	0.68	0.4976
FirstGen	0.1241	0.1587	0.78	0.4359
ESL1	-0.3022	0.1447	-2.09	0.0394
High_Biol	0.0692	0.1632	0.42	0.6724
Gendermale	0.1957	0.1441	1.36	0.1777
GPA	-0.0756	0.1223	-0.62	0.5378
factor(STEM)STEM	-0.2115	0.1694	-1.25	0.2151
factor(Academic_Year)Junior	-0.3053	0.5453	-0.56	0.5769
factor(Academic_Year)Sophomore	0.0990	0.1675	0.59	0.5560
scale(Stu_bepc1)	0.2540	0.1838	1.38	0.1702
scale(Stu_bepc2)	-0.4455	0.5226	-0.85	0.3961
factor(Lab_Course_Type)NON CURE	1.1535	0.7053	1.64	0.1052
scale(Ins_beha_PC1)	0.0744	0.3026	0.25	0.8064
scale(Ins_beha_PC2)	-0.1769	0.1310	-1.35	0.1801
factor(OverConf)1	-1.5556	0.1478	-10.52	0.0000

Finalized Linear Model

Table 5.7: Finalized linear model

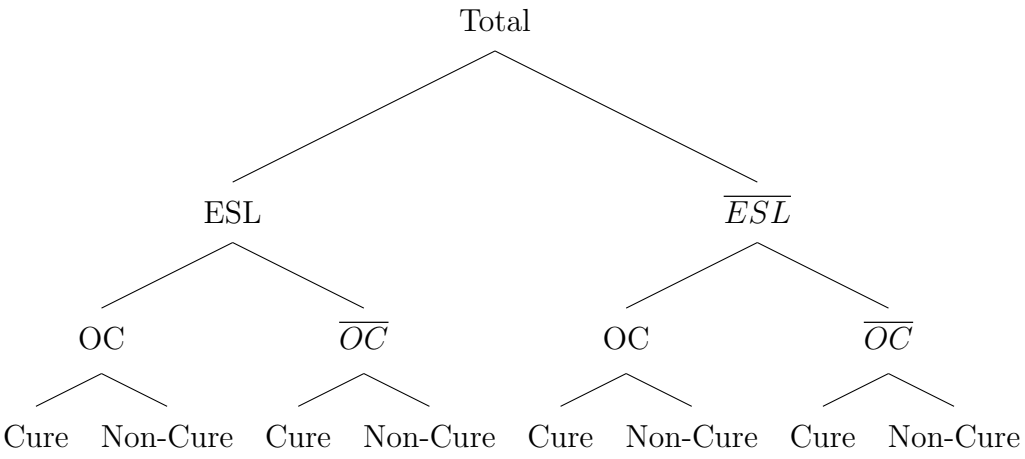
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5305	0.5335	0.99	0.3224
URM	0.2603	0.2327	1.12	0.2658
ESL1	-0.2849	0.1418	-2.01	0.0473
Gendermale	0.1967	0.1405	1.40	0.1646
GPA	-0.0465	0.1147	-0.41	0.6859
scale(Stu_bepc1)	0.2605	0.1801	1.45	0.1511
scale(Stu_bepc2)	-0.4165	0.5074	-0.82	0.4136
(Lab_Course_Type)NON CURE	1.1547	0.6836	1.69	0.0943
scale(Ins_beha_PC1)	0.0636	0.2934	0.22	0.8289
scale(Ins_beha_PC2)	-0.1533	0.1259	-1.22	0.2263
(OverConf)1	-1.5331	0.1438	-10.66	0.0000

After proceeding with several iterations linear model displayed by Table 5.7 is selected with URM, ESL, Gender, GPA, Lab Course Type, Student behavior PC1, Student behavior PC2, Instructor behavior PC1, Instructor behavior PC2 and Over confidence.

Tree Structure for Student Performance Data

Here the level 1 hypothesis is tested by checking the significance of the ESL variable and the level 2 hypotheses (between OC and Non-OC) are tested only if the level 1 hypotheses is significant. The level 3 hypotheses (between Cure and Non-Cure courses) are tested only if level 2 hypotheses were significant.

The tree diagram for above hypotheses testing is displayed below and all hypothesis testings is carried out using α =0.1.



Checking the Significance of Level 1 Hypotheses

The level 1 hypotheses are tested by fitting a main effects model for the data .

Parameter Estimates for the Fitted Main Effects Model

Table 5.8: Parameter estimates for the main effects model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5305	0.5335	0.99	0.3224
factor(URM)1	0.2603	0.2327	1.12	0.2658
factor(ESL)1	-0.2849	0.1418	-2.01	0.0473
Gendermale	0.1967	0.1405	1.40	0.1646
GPA	-0.0465	0.1147	-0.41	0.6859
scale(Stu_bepc1)	0.2605	0.1801	1.45	0.1511
scale(Stu_bepc2)	-0.4165	0.5074	-0.82	0.4136
scale(Ins_beha_PC1)	0.0636	0.2934	0.22	0.8289
scale(Ins_beha_PC2)	-0.1533	0.1259	-1.22	0.2263
Lab_Course_TypeNON CURE	1.1547	0.6836	1.69	0.0943
OC1	-1.5331	0.1438	-10.66	0.0000

The effect of the ESL variable is tested as the level 1 hypothesis.

$$H^1(0) : \alpha_{ESL} = 0$$

From the parameter estimate of fitted main effects model there is a negative effect from the ESL variable on the response (student performance) of interest. Constructed 90% simultaneous confidence interval for ESL is (-0.520,-0.049). The result indicates that there is a significant effect of ESL variable. Since the level 1 hypothesis is significant we can proceed testing on level 2 hypotheses.

Checking the Significance of Level 2 Hypotheses

The level 2 hypotheses are tested by fitting a two way interaction model with ESL and Over confidence interaction term.

Parameter Estimates for the Model with OC*ESL Interaction Term

Table 5.9: Parameter estimates for the two way interaction model

	Estimate	Std. Error	t value	Pr(>|t|)
ESL_OC0.0	1.0216	0.4203	2.43	0.0169
ESL_OC1.0	0.9099	0.4319	2.11	0.0376
ESL_OC0.1	-0.3721	0.3920	-0.95	0.3448
ESL_OC1.1	-0.7863	0.4006	-1.96	0.0524
factor(URM)1	0.2685	0.2346	1.14	0.2552
Gendermale	0.1981	0.1417	1.40	0.1654
GPA	-0.0185	0.1144	-0.16	0.8718
scale(Stu_bepc2)	0.3462	0.2464	1.41	0.1631
scale(Stu_bepc1)	0.0240	0.1097	0.22	0.8270
scale(Ins_beha_PC1)	0.3385	0.2473	1.37	0.1742
scale(Ins_beha_PC2)	-0.0354	0.1083	-0.33	0.7446

Parameter estimates for the ESL*OC interaction term indicates that not being a native english speaker and not being overconfident would have the largest positive effect on student performance. Below will be the results for the hypothesis testing using the proposed method where the effect of the OC*ESL variable is tested as the level 2 hypothesis.

Table 5.10: Confidence intervals for level 2 contrasts

Contrast	**Estimate**	**L-bound**	**U-bound**
$H_1^2(0) : \alpha_{(\overline{OC},\overline{ESL})-(OC,\overline{ESL})} = 0$	1.3937	0.9415	1.846
$H_2^2(0) : \alpha_{(\overline{OC},ESL)-(OC,ESL)} = 0$	1.6962	1.2232	2.1691

Since both of the simultaneous confidence intervals does not include 0 there is a significant difference in student performance for over confident and Non-Over confident students within the ESL and Non-ESL.

Checking the Significance of Level 3 Hypotheses

The level 3 hypotheses are tested by fitting model with ESL, Over confidence and course type three way interaction term. Each of the confidence intervals are created with 100*(1-α /4)% level of significance.

Parameter Estimates for the Model with OC*ESL*Course type Three Way Interaction Term

Table 5.11: Parameter estimates for three way interaction model

	Estimate	Std. Error	t value	Pr(>|t|)
all0.0.CURE	0.1798	0.6175	0.29	0.7716
all1.0.CURE	0.1752	0.5608	0.31	0.7554
all0.1.CURE	-1.0250	0.5298	-1.93	0.0560
all1.1.CURE	-1.4218	0.5373	-2.65	0.0095
all0.0.NON CURE	1.7125	0.5536	3.09	0.0026
all1.0.NON CURE	1.7643	0.5949	2.97	0.0038
all0.1.NON CURE	0.3177	0.5421	0.59	0.5592
all1.1.NON CURE	-0.1192	0.5567	-0.21	0.8309
factor(URM)1	0.2493	0.2386	1.04	0.2988
Gendermale	0.1956	0.1461	1.34	0.1840
GPA	-0.0531	0.1158	-0.46	0.6477
scale(Stu_bepc2)	-0.5559	0.5226	-1.06	0.2901
scale(Stu_bepc1)	0.3225	0.1877	1.72	0.0890
scale(Ins_beha_PC1)	0.0120	0.2985	0.04	0.9680
scale(Ins_beha_PC2)	-0.1723	0.1279	-1.35	0.1811

Required hypotheses to be tested and their results are listed below.

Table 5.12: Confidence intervals for level 2 contrasts

Contrast	**Estimate**	**L-bound**	**U-bound**
$H_1^3(0) : \alpha_{(\overline{OC},\overline{ESL},Cure)-(\overline{OC},\overline{ESL},Non-Cure)} = 0$	-1.5327	-3.6417	0.5763
$H_2^3(0) : \alpha_{(\overline{OC},ESL,Cure)-(\overline{OC},ESL,Non-Cure)} = 0$	-1.5891	-3.5679	0.3898
$H_3^3(0) : \alpha_{(OC,\overline{ESL},Cure)-(OC,\overline{ESL},Non-Cure)} = 0$	-1.3428	-3.2443	0.5588
$H_4^3(0) : \alpha_{(OC,ESL,Cure)-(OC,ESL,Non-Cure)} = 0$	-1.3026	-3.2455	0.6403

When considering the simultaneous confidence intervals for the contrasts all the confidence intervals includes zero. This indicates that there is no significant difference in student performance between Cure and Non-Cure course studying students, even we observed a difference being over confident.

5.3.2 Gender Concordance in Mentoring Relationships

Data Processing for Mentor and Mentee Relationship data set

The second data set consists of 80 student records on 14 variables regarding the mentor and mentee gender concordance. The student responses are recorded based on their gender, sums scores measuring their skills in different attributes and some demographic variables. The response variable is Thinking and working as Scientist and the explanatory variables are Category relationship, Parental education, Research time, Nativity and Hispanic. The Category relationship and parental education has been identified as the most important variables from the recursive partitioning and were used to create the tree structures. Figure 5.4 displays the structure F- Faculty member, P- Postgraduate mentor , U- Under graduate student, Solid line represents gender is matched and Dash Line represents gender is unmatched.

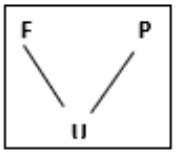
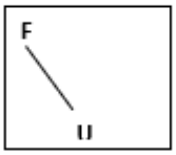
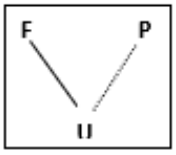
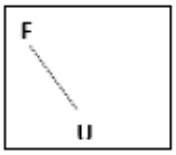
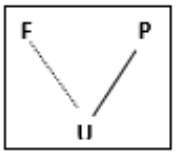
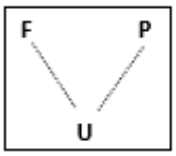
Categories	Pictographic depiction	Description in relation to the student mentee
1		Gender-Matched Faculty & Gender-Matched Postgraduate
2		Gender-Matched Faculty
3		Gender-Matched Faculty & Gender-Unmatched Postgraduate
4		Gender-Unmatched Faculty
5		Gender-Unmatched Faculty & Gender-Matched Postgraduate
6		Gender-Unmatched Faculty & Gender-Unmatched Postgraduate

Figure 5.4: Mentoring relationships based on gender concordance of the mentee, faculty mentor and postgraduate mentor

Model Building Process for Mentor Mentee Gender Concordance Data

The response variable is Thinking and working as a scientist which was a measure of the students ability to think and work as a scientist. The decision tree algorithm selects a model that incorporates ideal splits in order to use a minimum number of explanatory variables. The splitting criterion is universal across different methodologies and includes finding splits that improve the fit criterion (Mallows Cp) for the model (Therneau, Atkinson, & Ripley, 2015). Recursive partitioning is a fundamental tool in data mining, which helps us to explore the structure of a set of data. With the help of recursive partitioning parental education level and category relationship is identified as splitting variables to construct the tree to proceed the multiple comparisons.

Finalized Main Effects Model

Table 5.13: Finalized main effects model

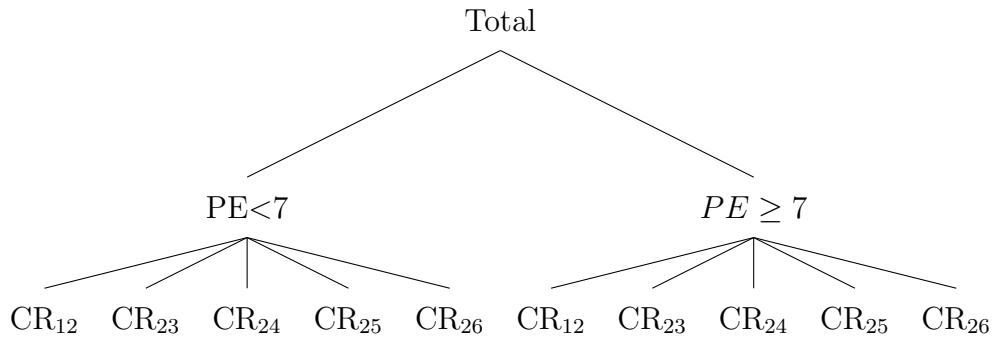
	Estimate	Std. Error	t value	Pr(>|t|)
factor(Category_relationship)1	30.4804	2.9201	10.44	0.0000
factor(Category_relationship)2	32.7756	3.3928	9.66	0.0000
factor(Category_relationship)3	29.4202	3.1811	9.25	0.0000
factor(Category_relationship)4	30.8854	3.2147	9.61	0.0000
factor(Category_relationship)5	29.6015	3.2468	9.12	0.0000
factor(Category_relationship)6	20.2689	4.3383	4.67	0.0000
ResearchTime	0.0818	0.5742	0.14	0.8872
factor(Hispanic)1	-2.2454	1.6993	-1.32	0.1920
factor(Nativity)1	2.6121	2.1056	1.24	0.2201
PEdu	2.6784	1.5228	1.76	0.0843

Catergory relationship and parental education level variables are significant at 0.1 level of significance.

Comparing the Effect when Faculty and Student and Gender were Matched With all Other Categories

Here the level 1 hypothesis is tested by checking the significance of the Parental education level variable and the level 2 hypotheses (comparing 2^{nd} category of Category relationship with other categories) are tested only if the level 1 hypotheses was significant. The tree diagram for above hypothesis testing is displayed as below and all hypothesis testings are carried out using α =0.1 (un-adjusted).

Here CR_{ij} denotes the difference of the effcect between i^{th} and j^{th} category.



Checking the Significance of Level 1 Hypotheses

The level 1 hypothesis is tested by fitting a main effects model (Table 5.13) for the data and the construted 90% confidence interval (0.130,5.226) does not include zero implies that the effect of Parental education level is significant.

$$H^1(0) : \alpha_{PE} = 0$$

Since the level one hypotheses is significant we can test the effect of category relationship within the parental education.

Checking the Significance of Level 2 Hypotheses

The level 2 hypotheses are tested by fitting a two way interactions model with the Parental education (PE) and Category relationship (CR) interaction term.

Parameter Estimates for Model with PE*CR Interaction Term

Table 5.14: Parameter estimates for two way interaction model

	Estimate	Std. Error	t value	Pr(>|t|)
CR_PE1.0	35.4549	4.0985	8.65	0.0000
CR_PE2.0	35.3645	2.9315	12.06	0.0000
CR_PE3.0	29.1124	4.2233	6.89	0.0000
CR_PE4.0	28.5970	3.0904	9.25	0.0000
CR_PE5.0	24.5747	3.0991	7.93	0.0000
CR_PE6.0	13.0357	5.2545	2.48	0.0166
CR_PE1.1	31.8664	2.3704	13.44	0.0000
CR_PE2.1	30.8827	3.5037	8.81	0.0000
CR_PE3.1	31.8983	2.7349	11.66	0.0000
CR_PE4.1	35.2553	3.0147	11.69	0.0000
CR_PE5.1	37.9280	3.6364	10.43	0.0000
CR_PE6.1	25.7932	4.2499	6.07	0.0000
ResearchTime	0.1616	0.5143	0.31	0.7546
factor(Hispanic)1	-1.1589	1.6309	-0.71	0.4807
factor(Nativity)1	1.7999	1.9195	0.94	0.3530

Table 5.15 represents required hypotheses to be tested in level 2 and their corresponding confidence intervals. For identification purpose $PE < 7 = \overline{PE}$ and $PE \geq 7 = PE$.

Table 5.15: Confidence intervals for level 2 contrasts

Contrast	**Estimate**	**L-bound**	**U-bound**
$H_1^2(0) : \alpha_{(\overline{PE},CR_1)-(\overline{PE},CR_2)} = 0$	0.09	-10.27	10.45
$H_2^2(0) : \alpha_{(\overline{PE},CR_3)-(\overline{PE},CR_2)} = 0$	-6.25	-16.64	4.14
$H_3^2(0) : \alpha_{(\overline{PE},CR_4)-(\overline{PE},CR_2)} = 0$	**-6.76**	**-13.02**	**-0.52**
$H_4^2(0) : \alpha_{(\overline{PE},CR_5)-(\overline{PE},CR_2)} = 0$	**-10.79**	**-18.68**	**-2.90**
$H_5^2(0) : \alpha_{(\overline{PE},CR_6)-(\overline{PE},CR_2)} = 0$	**-22.33**	**-8.47**	**-36.18**
$H_6^2(0) : \alpha_{(PE,CR_1)-(PE,CR_2)} = 0$	0.98	-6.14	8.10
$H_7^2(0) : \alpha_{(PE,CR_3)-(PE,CR_2)} = 0$	1.01	-6.76	8.79
$H_8^2(0) : \alpha_{(PE,CR_4)-(PE,CR_2)} = 0$	4.37	-2.73	11.48
$H_9^2(0) : \alpha_{(PE,CR_5)-(PE,CR_2)} = 0$	7.04	-2.17	16.26
$H_{10}^2(0) : \alpha_{(PE,CR_6)-(PE,CR_2)} = 0$	5.09	-5.23	15.41

When we refer constructed 99% simultaneous confidence intervals (Table 5.15) for contrasts, we can observe significant effects from below 3 contrasts (highlighted in bold in Table 5.15).

$$H_3^2(0) : \alpha_{(\overline{PE},CR_4)-(\overline{PE},CR_2)} = 0$$

$$H_4^2(0) : \alpha_{(\overline{PE},CR_5)-(\overline{PE},CR_2)} = 0$$

$$H_5^2(0) : \alpha_{(\overline{PE},CR_6)-(\overline{PE},CR_2)} = 0$$

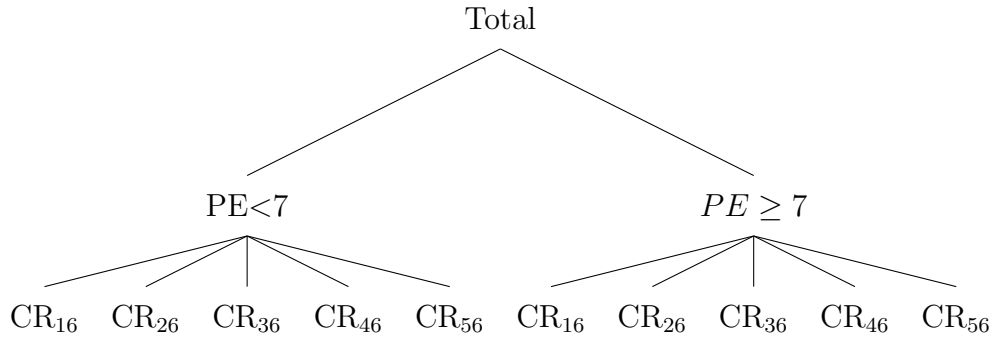
This indicates that for **parent education level less than 7**,

- There is a significant difference in Thinking and working like a scientist when faculty and student gender were matched vs unmatched. There is a significant difference in Thinking and working like a scientist when faculty and student gender were matched vs student and faculty gender were unmatched, matched postgraduate. There is a

significant difference in Thinking and working like a scientist when faculty and student gender were matched vs faculty, student and postgraduate gender unmatched.

Comparing the Effect when Faculty, Student and Postgraduate Gender were Unmatched With all Other Categories

Here the level 1 hypothesis is tested by checking the significance of the Parental education level variable and the level 2 hypotheses (comparing 6^{th} category of Category relationship with other categories) are tested only if the level 1 hypotheses was significant. The tree diagram for above hypotheses testing is displayed as below and all hypothesis testings was carried out using α =0.1. (adjustment was done level wise)



Checking the Significance of Level 1 Hypotheses

The level 1 hypothesis is tested by fitting a linear model (model 1) for the data and the construted 90% confidence interval (0.130,5.226) does not include zero implies that the effect of Parental education level is significant.

$$H^1(0) : \alpha_{PE} = 0$$

Since the level one hypothesis is significant we can test the effect of category relationship within the parental education.

Checking the Significance of Level 2 Hypotheses

The level 2 hypotheses are tested by fitting a linear model with the Parental education (PE) and Category relationship (CR) interaction term for the data. Table 5.16 represents required hypotheses to be tested in level 2 and their corresponding confidence intervals. For identification purpose $PE < 7 = \overline{PE}$ and $PE \geq 7 = PE$.

Table 5.16: Confidence intervals for level 2 contrasts

Contrast	**Estimate**	**L-bound**	**U-bound**
$H_1^2(0) : \alpha_{(\overline{PE},CR_1)-(\overline{PE},CR_6)} = 0$	**22.42**	**6.84**	**38.00**
$H_2^2(0) : \alpha_{(\overline{PE},CR_2)-(\overline{PE},CR_6)} = 0$	**22.33**	**8.47**	**36.18**
$H_3^2(0) : \alpha_{(\overline{PE},CR_3)-(\overline{PE},CR_6)} = 0$	16.07	-0.04	32.20
$H_4^2(0) : \alpha_{(\overline{PE},CR_4)-(\overline{PE},CR_6)} = 0$	**15.56**	**1.87**	**29.24**
$H_5^2(0) : \alpha_{(\overline{PE},CR_5)-(\overline{PE},CR_6)} = 0$	11.54	-2.8	25.87
$H_6^2(0) : \alpha_{(PE,CR_1)-(PE,CR_6)} = 0$	6.07	-4.18	16.33
$H_7^2(0) : \alpha_{(PE,CR_2)-(PE,CR_6)} = 0$	5.09	-5.23	15.41
$H_8^2(0) : \alpha_{(PE,CR_3)-(PE,CR_6)} = 0$	6.10	-4.38	16.59
$H_9^2(0) : \alpha_{(PE,CR_4)-(PE,CR_6)} = 0$	9.46	-1.01	19.93
$H_{10}^2(0) : \alpha_{(PE,CR_5)-(PE,CR_6)} = 0$	**12.13**	**0.47**	**23.79**

Table 5.16 shows constructed 99% simultaneous confidence intervals for the contrasts and we can observe significant effects from below 4 contrasts (highlighted in bold in Table 5.16).

$$H_1^2(0) : \alpha_{(\overline{PE},CR_1)-(\overline{PE},CR_6)} = 0$$

$$H_2^2(0) : \alpha_{(\overline{PE},CR_2)-(\overline{PE},CR_6)} = 0$$

$$H_4^2(0) : \alpha_{(\overline{PE},CR_4)-(\overline{PE},CR_6)} = 0$$

$$H_{10}^2(0) : \alpha_{(PE,CR_5)-(PE,CR_6)} = 0$$

This indicates that for **parental education level less than 7**,

- There is a significant difference in Thinking and working like a scientist when faculty, student and postgraduate gender were matched vs unmatched. There is a significant difference in Thinking and working like a scientist when faculty and student gender were matched vs faculty, student and postgraduate gender were unmatched. There is a significant difference in Thinking and working like a scientist when faculty and student gender were unmatched vs faculty, student and post graduate gender unmatched.

Furthermore, for **parental education level greater than 7** we can state that there is a significant difference in Thinking and working like a scientist when student and postgraduate gender were matched vs unmatched. This provides evidence that the post graduate gender concordance has an effect.

References

Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE-Life Sciences Education*, *13*(4), 602–606.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bogomolov, M., Peterson, C. B., Benjamini, Y., & Sabatti, C. (2017). Testing hypotheses on a tree: new error rates and controlling strategies. *arXiv preprint arXiv:1705.07529*.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology*, *83*(2), 275.

Dayioğlu, M., & Türüt-Aşik, S. (2007). Gender differences in academic performance in a large public university in turkey. *Higher Education*, *53*(2), 255–277.

Emilio, D. R., Belluzzo Jr, W., & Alves, D. C. (2004). Uma análise econométrica dos determinantes do acesso à universidade de são paulo.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, *43*(2), 95.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211.

Guimarães, J., & Sampaio, B. (2013). Family background and students achievement on a university entrance exam in brazil. *Education Economics*, *21*(1), 38–59.

Kendall, K. D., & Schussler, E. E. (2013). Evolving impressions: undergraduate perceptions of graduate teaching assistants and faculty members over a semester. *CBE-Life Sciences Education*, *12*(1), 92–105.

Lindquist, M. A., & Mejia, A. (2015). Zen and the art of multiple comparisons. *Psychoso-*

matic Medicine, *77*(2), 114.

Olimpo, J. T., Fisher, G. R., & DeChenne-Peters, S. E. (2016). Development and evaluation of the tigriopus course-based undergraduate research experience: Impacts on students content knowledge, attitudes, and motivation in a majors introductory biology course. *CBE-Life Sciences Education*, *15*(4), ar72.

Therneau, T., Atkinson, B., & Ripley, B. (2015). *rpart: Recursive partitioning and regression trees. r package version 4.1–10.*

Yekutieli, D. (2008). Hierarchical false discovery rate–controlling methodology. *Journal of the American Statistical Association*, *103*(481), 309–316.

Yi, N., Xu, S., Lou, X.-Y., & Mallick, H. (2014). Multiple comparisons in genetic association studies: a hierarchical modeling approach. *Statistical Applications in Genetics and Molecular Biology*, *13*(1), 35–48.

Young, J. W., & Fisler, J. L. (2000). Sex differences on the sat: An analysis of demographic and educational variables. *Research in Higher Education*, *41*(3), 401–416.

Curriculum Vitae

Dimuthu Fernando was born in January 22, 1988. His initial foundation for science was laid at S. Thomas College, Bandarawela, Sri Lanka. Where he sat for the G.C.E Advanced Level Examination (2007), Following Combined Mathematics, Chemistry, and Physics. Having excelled at this highly competitive examination, he was fortunate to gain entrance to the most prestigious university in Sri Lanka, University of Colombo, to follow an undergraduate degree in Industrial Statistics and Mathematical Finance. He graduated as a Bachelor of Science (Special) in Industrial Statistics in 2013 with a Second Class Honours (Upper Division). The completion of his undergraduate research project titled : A Multilevel Study to Determine the Factors Affecting the Stream Selected by University Entering Students enhanced his interest in Applied Statistics and research through Statistical applications.

After completion of B.Sc degree he worked as an Assistant Manager at MAS Active which is a globally recognized multinational company in apparel manufacturing from March 2013 to August 2016. In the fall of 2016, he entered the Graduate School of The University of Texas at El Paso. While pursuing a masters degree in Statistics, he worked as a teaching assistant and as a tutor in MARCS.

Permanent address: A9/2/1 Manning Town Housing Scheme
Matha road, Colombo 08, Sri Lanka.