Spring 2018

Probabilistic pairwise model comparisons based on discrepancy measures and a reconceptualization of the p-value

Benjamin N. Riedle
University of Iowa

PROBABILISTIC PAIRWISE MODEL COMPARISONS BASED ON DISCREPANCY MEASURES AND A RECONCEPTUALIZATION OF THE P-VALUE

by

Benjamin N. Riedle

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Biostatistics
in the Graduate College of
The University of Iowa

May 2018

Thesis Supervisors: Professor Joseph Cavanaugh
Professor Andrew Neath

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Benjamin N. Riedle

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Biostatistics at the May 2018 graduation.

Thesis committee: ______________________________
Joseph Cavanaugh, Thesis Supervisor

____
Andrew Neath, Thesis Supervisor

____
Jacob Oleson

____
Eric Foster

____
Linnea Polgreen

____
Philip Polgreen

ACKNOWLEDGEMENTS

As the English clergyman and poet John Donne said "No man is an island...every man is a piece of the continent." I believe no words could better describe how deeply indebted I am to the innumerable people who have been integral in my development as a student and as a person. While I will unfortunately be unable to list every individual who has had a significant impact on my life, I sincerely appreciate all the help and advice I have been provided. Such individuals include: teachers and professors, without whom completing a doctoral degree would have been impossible; fellow students, with whom I have spent countless hours doing homework and studying (and getting distracted by considerably more interesting topics of discussion!); and family and friends, who have provided me with truly invaluable guidance and camaraderie. Thank you for all your support throughout the years.

I would like to start off by thanking my dissertation co-advisor and friend, Dr. Joseph Cavanaugh. I believe that no accolade is inappropriate when describing Joe. Joe is an astoundingly thoughtful, friendly, hard-working and generous man who has inestimably contributed to my development as a statistician and researcher. I feel incredibly lucky to have met Joe and even more fortunate that he was willing to serve as my mentor and advisor. Throughout my tenure at the University of Iowa, Joe has been amazingly charitable with his time, with he and I having weekly meetings. While the academic topics of these meetings have changed throughout the years, ranging from reading and discussing others' research papers in the early years to, more recently, hashing out countless details of my dissertation, one thing has remained true: Joe has been an astonishingly helpful, insightful and caring mentor. Of course, Joe is much more than simply an advisor and mentor, he is also an awesome friend. While we eventually begin discussing statistical topics in our meetings, Joe and I both have a knack for distracting each other by starting to discuss unrelated (but typically more interesting!) topics, such as the Beatles, Led Zeppelin and Iowa football. As well as having a great deal of fun in our meetings, Joe is also a gracious and generous host, with him often treating me to burgers at Short's Restaurant. Joe, I would just like to say I am incredibly grateful for everything you have done for me, both as a mentor and as a friend. Thank you.

I would next like to thank my dissertation co-advisor, master's thesis advisor and friend, Dr. Andrew Neath. Andy is an amazingly intelligent, thoughtful, charitable and kind person. When I was an undergraduate who was beginning to work on my master's degree, Andy reached out to me to ask if I would like to work with him on my master's thesis. Though I considered this to be a very generous offer which I gladly accepted, I did not realize at the time how invaluable and rewarding my relationship with Andy would ultimately prove. As an advisor and mentor, Andy truly is second to none; he is very generous with his time, enthusiastically helpful, incredibly insightful, and a kindly explicator. Like Joe, my meetings with Andy were always quite academically beneficial, but we also allotted plenty of time to talk about St. Louis sports (especially the hapless Blues!), roller coasters, as well as many other fun topics. As a friend, Andy is kind and generous, having taken me to Blues and

Cardinals games as well as many coffee shops and restaurants. Andy has always been there for me as a mentor and friend, and I feel so lucky to have met him and worked so closely with him on two major research projects. Thanks, Andy.

I would also like to thank my research assistantship coordinator and mentor, Dr. Linnea Polgreen. Linnea and I have worked closely together for the entirety of my doctoral studies. Linnea is a very accomplished researcher and academician, all while remaining laid back and friendly. I have truly enjoyed working with Linnea and believe that my work with her has been invaluable to my development as an applied statistician.

I would next like to thank the remaining members of my committee, Dr. Philip Polgreen, Dr. Jacob Oleson and Dr. Eric Foster. I have had the good fortune of regularly working and meeting with Phil throughout my time at Iowa. Phil is a very accomplished and energetic researcher and physician who is also kind and funny. He has aided considerably in my development as an applied researcher, always devising new and interesting topics to research (I honestly do not know how he is able to come up with all his ideas!). I have been very lucky to have Jake as my professor for two classes. Jake is an incredibly friendly and approachable professor who truly cares about his students. While I have only had Eric as a professor for a few class periods, I have fortunately gotten to know him outside of the academic environment. Eric is truly one of the nicest and most thoughtful people I have ever met and is also a very skilled researcher and professor.

I also wish to express my sincere gratitude for the University of Iowa Biostatistics Department faculty, staff and student body. The rapport between and within the faculty and students of this department is remarkable. The faculty always have their doors open to help the students, and the students are always willing to aid one another. The Biostatistics Department is a sterling example of what academics should be all about: hard work, intellectual curiosity, important research, an eagerness and aptitude for teaching and learning, and a friendly and approachable environment. I consider it an honor to be a graduate from this impressive department.

While all my teachers were important to my academic development, I wish to individually thank a few teachers who were most influential. First, I wish to express gratitude for my first grade teacher, Mrs. Carolyn Henley. In first grade, I struggled mightily with basic reading and speech, but Mrs. Henley believed in me, which gave me confidence that I would eventually learn to read. I am also indebted to my third grade teacher, Ms. O'Donnell, who was a very energetic and passionate teacher, instilling in me a lifelong love of learning. I am also grateful for my fifth grade teacher, the late Mrs. Carol Downing, who was an incredibly talented educator, providing me with a very strong education as I entered middle school. Finally, I would like to thank my undergraduate math professor, Dr. Adam Weyhaupt, for being an astoundingly talented and helpful professor, which ultimately helped provide me with a very strong math background as I entered my doctoral work. Adam was always willing to help, often spending a great deal of time with his students during office hours due to the line out the door waiting for his highly useful guidance.

I would like to thank three of my fellow graduate students in particular, Dr.

Natalie Langenfeld, Dr. Janel Barnes, and soon-to-be-Dr. Ryan Peterson. I have known Natalie since my freshman year in college and have spent countless hours working on homework and studying with her as an undergraduate, master's student and doctoral student. Natalie is one of my best friends, and I sincerely thank her for all the help she has provided in my academic career. As doctoral students, Natalie, Janel and I spent a considerable amount of time working together on homework and studying for exams. In a semester in which we took an extra Ph.D.-level course in addition to our regular academic load, we spent upwards of 50 hours a week together. Pursuing my Ph.D. would not have been the same without them. Thanks for everything, Natalie and Janel. While Ryan and I have not spent much time working together on homework or studying, he has been an incredibly insightful collaborator and a very good friend. In our research meetings, Ryan always provides impressively useful and thought-provoking suggestions. Also, despite nearly always beating me now, I greatly look forward to when we get to play tennis together. Thanks, Ryan.

I want to thank some of my best friends, some of whom I have known since before high school. First, I want to acknowledge Adam Braundmeier, J.D., with whom I have been best friends since before third grade. Adam and I were next-door neighbors growing up and have been inseparable for much of our lives. I would also like to thank Adam for not killing me when he accidentally ran me over in the high school parking lot; I certainly would not be here completing this dissertation if he had been driving a little bit faster! To another of my best friends, Nick Griswold, I thank you for always keeping me up-to-date on the newest conspiracy theories and for providing “indisputable” evidence of the existence of aliens on Earth. While what Nick has to say may, at times, be of questionable veracity, it is undoubtedly never dull, and I always look forward to speaking with him. To Dan Turner, who I have known well since senior year in high school, thank you for being one of the nicest people I have ever met and also thank you for organizing “hot tub Tuesdays!” To Anthony Allsup, thank you for being awesome and always wanting to do something fun (including going on my poorly-planned and quite excruciating hiking and biking trips!). To Zach Bugger, thank you for being a loyal friend who I know will always be on my side (and thanks for having my favorite surname ever!). To James Tucker, first, thank you for your military service, and second, thank you for always wanting to do fun activities, including pool volleyball, bags, hockey and just about any other game under the sun. To Dr. Andy Rombach, one of the smartest, most well-rounded people I have ever met, thank you for being an incredibly loyal friend and for being a trivia master, securing victories for our team when the rest of us would not have had a chance. To Joe Moen, thank you for being an incredibly fun person who is always able to provide me with up-to-date results from the world of sports, and thank you for taking good care of my sister. Finally, thank you to Kayla Hoffman, for being an awesome neighbor, a loyal friend, and a person who is always fun to hang out with. Without good friends, life would be far less rich and meaningful, so I sincerely appreciate the friendship I have with each of you.

I would next like to thank my best friend, girlfriend, confidante and love of my life, Kristen Merkitch. Kristen and I have been dating for the vast majority of my time at Iowa. Meeting her was one of the most important and fortuitous moments of my life. Kristen is funny, kind, thoughtful, beautiful, intelligent, and just fun to be around. Her unique personality and wit bring an excitement to conversations that is irreplaceable. As a fellow Ph.D. student, Kristen has always been very understanding of the long nights which were sometimes necessary to study, do homework and complete my dissertation. Without Kristen by my side, graduating and completing this dissertation would have been much more difficult and far less enjoyable. Kristen, I love you sincerely and am so glad we are together.

I next wish to thank my loving family. My father and mother are truly the best of parents, being astoundingly loving, friendly, intelligent and wise. My dad is an amazingly hard-working, thoughtful and fun-loving man. My dad has, to the best of his ability, tried to instill his incredible work ethic in me. While I may not have always lived up to the standard he set, he has always been a very understanding person. I am very fortunate to have such a great father and am also lucky that we share many of the same hobbies, including going on several week-long hiking trips together. My mother is an astonishingly thoughtful, loving, patient and loyal person. Growing up, my mother would always be willing to go the extra mile for my sister and me; she would consistently sacrifice her time to take us to practices, games, and other events. My mother is quick to laugh and is very fun to be around, and I sincerely enjoy spending time with her. Whether I am visiting with my parents or am hours away, they always make it abundantly clear that both my sister and I are loved. I would also like to thank my sister, Danielle. Unlike many siblings, Danielle and I nearly always got along well when we were growing up. Despite being almost four years my junior, when we were kids, Danielle was a loyal companion who was always willing to participate in any games or activities my friends and I would devise. As we have gotten older, Danielle and I continue to stay close, and I greatly cherish our friendship and our annual camping trip. I would also like to express my gratitude to my parents' neighbors, Dr. Marcia Buono and the late Mr. Ralph Buono. Marcia and Ralph made an excellent team, both being incredibly kind, generous and very fun-loving people. I know I speak for everyone in my family when I say that we are so glad that Marcia is still our neighbor and that Ralph is sorely missed. I would like to further acknowledge Kristen's parents, Dr. Ken Merkitch and Mrs. Janine Merkitch, for being so kind and generous and welcoming me into the family as one of their own. I would also like to thank my wonderful grandparents, Lonnie and Rosalee Riedle and Frank and Isabelle Tolley. My Grandpa Lonnie and Grandma Rosalee have encouraged me as I have advanced my eduction and have provided me with love and support throughout my whole life. My Grandpa Lonnie's hardwork and perseverance is truly inspirational; while working a full-time job, he also helped raise five children and maintained approximately 20 rental houses at any given time (I unfortunately did not inherit his craftsmanship!). My Grandma Rosalee is very intelligent, earning the honor of valedictorian of her high school class, and a great caregiver, raising five wonderful children. While Grandpa Frank unfortunately passed away when I was quite young, I am very glad that I was able to meet him. He was a very good person and raised eight great children. My late Grandma Isabelle was one of the nicest people I have ever met. She was always willing to do anything for anybody. When my sister was very young, she loved playing the card game Uno. My grandma would play Uno with her for hours on end, not because she loved Uno, but because she loved Danielle and always wanted her to be happy. My grandma is truly missed. I would finally like to thank all my cousins, aunts and uncles. You mean the world to me, and I appreciate everything you have done for me.

ABSTRACT

Discrepancy measures are often employed in problems involving the selection and assessment of statistical models. A discrepancy gauges the separation between a fitted candidate model and the underlying generating model. In this work, we consider pairwise comparisons of fitted models based on a probabilistic evaluation of the ordering of the constituent discrepancies. An estimator of the probability is derived using the bootstrap.

In the framework of hypothesis testing, nested models are often compared on the basis of the p-value. Specifically, the simpler null model is favored unless the p-value is sufficiently small, in which case the null model is rejected and the more general alternative model is retained. Using suitably defined discrepancy measures, we mathematically show that, in general settings, the Wald, likelihood ratio and score test p-values are approximated by the bootstrapped discrepancy comparison probability (BDCP). We argue that the connection between the p-value and the BDCP leads to potentially new insights regarding the utility and limitations of the p-value. The BDCP framework also facilitates discrepancy-based inferences in settings beyond the limited confines of nested model hypothesis testing.

PUBLIC ABSTRACT

In recent years, the misunderstanding and misapplication of the p-value has led to skepticism in the general efficacy of hypothesis testing and the p-value in scientific research (Peng, 2015). This skepticism has led to some extreme decisions. For instance, in 2015, the editors of *Basic and Applied Social Psychology* decided to ban all p-values (Trafimow and Marks, 2015). The mounting controversy led the American Statistical Association (ASA) to take the unprecedented step of issuing a recent policy statement on p-values, hoping to reduce confusion about their proper interpretation and use (Wasserstein and Lazar, 2016). This dissertation will address one area of concern regarding the p-value: the difficulty in its interpretation. Statistics textbooks typically define the p-value as "the probability that if the null hypothesis is true, a test statistic will have a value as extreme or more extreme than the value we actually observe" (Gould and Ryan, 2013). We show that in general settings, the p-value can be interpreted in an arguably more intuitive manner. To see how, consider two competing statistical models, one of which corresponds to the null hypothesis and other to the alternative hypothesis. Deciding which of the two models is better can be done using a discrepancy function, which measures how well a statistical model adheres to the truth. In certain frameworks, we establish that the p-value can be interpreted as an estimator of the probability that the null model has a smaller discrepancy than the alternative (i.e. the null model should be preferred to the alternative). Thus, rather than having to assume the null is true and determining

a probability that reflects what was observed, as the standard p-value interpretation requires, the p-value can at times simply be interpreted as the probability the null model is better than the alternative.

TABLE OF CONTENTS

LIST OF TABLES xv

LIST OF FIGURES xvi

CHAPTER

1 INTRODUCTION 1

1.1 Hypothesis Testing and p-Values 1
1.2 p-Value Controversies 3
1.3 p-Value Alternatives 7
1.4 p-Value Alternative Interpretations 10
1.5 DCP/BDCP 14
1.6 Dissertation Outline 16

2 BACKGROUND 17

2.1 Discrepancy Functions 17
2.2 Using the Bootstrap to Estimate Discrepancy Functions 19
2.3 Discrepancy Comparison Probability (DCP) 21
2.4 Wald, Score and Likelihood Ratio (LR) Tests 28

3 INTRODUCTORY EXAMPLES 32

3.1 BDCP / p-Value Connection: One-Sample Example 32
3.2 BDCP / Wald p-Value Connection: Simplified GLM Setting . . . 37

4 BDCP / WALD P-VALUE CONNECTION: GENERAL SETTING . 40

5 BDCP / LR P-VALUE CONNECTION 45

5.1 Full Null Setting 45
5.2 Partial Null Setting 50

6 BDCP / SCORE P-VALUE CONNECTION 57

6.1 Full Null Setting 57
6.2 Partial Null Setting 60

7 SIMULATION STUDY 66

7.1 Design 66
7.2 Linear Regression 68
7.3 Logistic Regression 69
7.4 Interpretation of Results 70
7.5 Tables and Figures 72

8 BDCP / P-VALUE CONNECTION: FURTHER INVESTIGATION . 82

8.1 Undesirable Model Selection Properties of BDCP / p-Value . . . 82
8.2 Adjusted R^2 / Mean Squared Error as Model Selection Criteria . 83
8.3 AIC / DCP Comparison 84
8.4 Bias of Discrepancy Estimators Under KL Discrepancy 86
8.5 Biased Estimators and Connection to p-Value 88
8.6 Remedial Approaches for Bias 89

9 BENEFITS OF DCP / BDCP FRAMEWORK 92

9.1 Insights Gained from BDCP / p-Value Connection 92
9.2 Broader Utility of DCP / BDCP Framework 94

10 APPLICATIONS 98

10.1 ACE/ARB and Survival Study 98
10.1.1 Overview 98
10.1.2 Primary Results 101
10.1.3 Secondary Results 102
10.2 Phantom Antimicrobial Prescribing Study 105
10.2.1 Overview 105
10.2.2 Results 107

11 CONCLUDING REMARKS 109

11.1 Summary 109
11.2 Future Directions 111

12 REFERENCES 112

LIST OF TABLES

Table

7.1 $\hat{\rho}_c$ comparing $\hat{P}^*$ to its respective p-value in the linear regression setting. 72

7.2 $\hat{\rho}_c$ comparing $\hat{P}^*$ to its respective p-value in the logistic regression setting. 77

10.1 2×2 table showing the relationship between ACE/ARB use and one-year post-index discharge survival. A patient was considered an ACE/ARB user if and only if he or she filled a prescription for an ACE or ARB within 30 days of the index discharge date. 101

10.2 The Wald, score and LR p-values as well as their corresponding BDCPs determining the significance of ACE/ARB use on one-year survival. . . . 103

10.3 The BDCP under the WAD discrepancy for a variety of weighting schemes. The weighting schemes place probability w on the ACE/ARB parameter and distribute the remaining $1-w$ equally across the remaining parameters.105

10.4 The Wald, score and LR p-values as well as their corresponding BDCPs determining the significance of being white on whether a patient receives a phantom antimicrobial prescription. 108

LIST OF FIGURES

Figure

2.1 Visual representation of the resampling process that would, in principle, be used in the approximation of the DCP. 26

2.2 Visual representation of the resampling process used in the approximation of the BDCP. 27

7.1 Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an adequately specified null. Here, $\beta_0 = 0$, $\sigma^2 = 50$ and $n = 500$. 73

7.2 Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.175$, $\sigma^2 = 60$ and $n = 100$. 74

7.3 Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.10$, $\sigma^2 = 57$ and $n = 250$. 75

7.4 Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.05$, $\sigma^2 = 55$ and $n = 1000$. 76

7.5 Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an adequately specified null. Here, $\beta_0 = 0$ and $n = 500$. 78

7.6 Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.07$ and $n = 100$. 79

7.7 Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.04$ and $n = 250$. 80

7.8 Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.01$ and $n = 1000$. 81

CHAPTER 1
INTRODUCTION

In this dissertation, we introduce a pairwise model comparison probability based on discrepancy measures. We argue that this discrepancy comparison probability (DCP) has broader utility and necessitates fewer assumptions than what is required for conventional hypothesis testing and p-value interpretations. Despite the broader applicability of the DCP, this work attempts to establish a connection between the p-value and the DCP when hypothesis testing assumptions are met. Because the DCP cannot be directly calculated, bootstrap resampling techniques are used to derive an estimator. We show that in certain large-sample settings, the score, Wald and likelihood ratio (LR) test p-values are approximated by the bootstrapped discrepancy comparison probability (BDCP) using suitably defined discrepancy measures.

We begin the dissertation by motivating the study of the p-value. To do so, we first provide a brief introduction to hypothesis testing. We discuss some of the critiques of the p-value, as well as provide different interpretations of the p-value and alternatives to the p-value. We then informally introduce the DCP and BDCP. We end the introduction with an overview of the dissertation.

1.1 Hypothesis Testing and p-Values

Statistical hypothesis testing is often used in scientific research. Nevertheless, many scientific researchers lack a strong foundational understanding of hypothesis testing. We thus begin with a brief overview of hypothesis testing and the p-value.

A statistical model embodies a set of assumptions regarding how a set of data was generated. Ideally, a model will provide a good approximation to the data-generating mechanism. Using the observed data, a hypothesis regarding some aspect of the model can be evaluated via a hypothesis test. In hypothesis testing, two hypotheses, the null and alternative, are proposed. The null typically corresponds to an assumption of no effect or no difference. The model corresponding to the alternative is assumed to be adequately specified.

To test the null hypothesis, a test statistic is put forth that seeks to characterize the observed data in a manner that allows for discrimination between the null and alternative hypotheses. Typically, a test statistic is chosen because its distribution is known when the null hypothesis is true. The sampling distribution of the test statistic under the null is often referred to as the reference distribution. Intuitively, based on the observed data and reference distribution, we seek to determine whether the observed value of the test statistic is "extreme" under the assumption that the null is true. If the observed value of the test statistic is deemed "sufficiently extreme," then we consider this to be adequate evidence against the null hypothesis, and thus choose to reject the null. More formally, based on the reference distribution (which is derived under the null hypothesis), we seek to determine the probability of observing a value of the test statistic which is as extreme as or more extreme than the observed value. This probability is known as the p-value. If the p-value is sufficiently small, hence denoting an extreme value of the test statistic, then the null hypothesis is rejected in favor of the alternative. In the context of model selection and evaluation, hypothesis testing can be used to delineate between two competing nested models. To do so, the simpler null model is favored unless the p-value is sufficiently small, in which case the null model is rejected, and the larger alternative model is retained. A pre-specified level of significance α is often provided, and the null model is rejected if and only if the p-value is less than α.

1.2 p-Value Controversies

In recent years, misunderstanding and misapplication of the p-value has led to skepticism in the general efficacy of hypothesis testing in scientific research (Peng, 2015). This skepticism has led to some extreme decisions. For instance, in 2015, the editors of *Basic and Applied Social Psychology* decided to ban all p-values (Trafimow and Marks, 2015). The mounting controversy led the American Statistical Association (ASA) to take the unprecedented step of issuing a policy statement on p-values, hoping to reduce confusion about their proper interpretation and use (Wasserstein and Lazar, 2016).

Goodman (1999) argues that the use of hypothesis testing and the p-value in biomedical applications, without regard for biological plausibility, has led to numerous inaccurate scientific claims. Goodman (1999) contends that many scientific papers are written in a manner which suggests that the conclusions of a study are direct implications of the results of hypothesis testing. Such papers are written so that the primary concern is whether a p-value is below some threshold, such as $\alpha = 0.05$.

Whether the p-value meets this threshold is examined before any discussion of other relevant considerations, such as the proposed mechanism of action or the clinical significance of the result. The structure of such scientific papers has contributed to many readers being unable to discriminate between bona fide scientific conclusions and the results of hypothesis tests. Because the results of a single hypothesis test, even one which is highly significant, should rarely be considered overwhelming evidence against the null hypothesis, the language of "rejecting" the null based on a small p-value can be misleading.

That clinical researchers often misunderstand the p-value is not surprising when considering the standard p-value interpretation: "the probability that if the null hypothesis is true, a test statistic will have a value as extreme or more extreme than the value we actually observe" (Gould and Ryan, 2013). To interpret a p-value, instructors generally present the following counterintuitive line of reasoning. First, in many practical settings, a null hypothesis can *never* be precisely true. However, we will assume that it is *true.* Then, the p-value assesses the plausibility of the test statistic (*not the null hypothesis*) under the assumption that the null hypothesis is true (*even though it cannot be true*). Undoubtedly, this interpretation has confused numerous introductory statistics students. In a 2006 study of medical residents, 88% of those surveyed believed they had fair to complete confidence in their ability to interpret a p-value, but only 59% could successfully answer a basic question related to the interpretation of a result of $p > 0.05$ (Windish, Huot and Green, 2007). This finding illustrates that many medical practitioners, who need to have a basic comprehension of statistics in order to understand the medical literature, lack knowledge of important concepts taught in introductory statistics courses. According to Goodman (2008), the most common misinterpretation of the p-value is that $p = 0.05$ implies there is a 5% chance that the null hypothesis is true. Of course, the p-value is based on the assumption that the null is true and thus cannot also be a probability the null is false.

Arguments and criticisms against the p-value are numerous, yet use of the p-value in statistical inference has not slowed appreciably. Goodman (2001) believes one reason for the continued use, and often misuse, of the p-value is what he refers to as "naive inductism." Naive inductism refers to the notion that when two scientists look at the same data, they should come to the exact same conclusion, with no room for subjectivity. Naive inductism requires scientists to assume their statistical models accurately represent reality and that inferences drawn from these models make manifest some underlying truth about the world. The p-value can contribute to this problem because it is a probability, and therefore some individuals will mistakenly believe it is an objective measure of reality, without recognizing the underlying modeling assumptions.

In many real-world applications of the p-value, a result is considered "statistically significant" if and only if the corresponding p-value is less than 0.05. Similarly, when hypothesis testing is put in the model selection context, a cutoff of 0.05 is often used to delineate between two competing nested models. However, the use of such a pre-specified cutoff point is completely arbitrary; p-values of 0.0495 and 0.0505 represent nearly identical evidence against the null, but typically the former is deemed significant, while the latter is not significant.

In standard hypothesis testing, if the alternative model is not adequately specified, then the validity of the results of a hypothesis test, including the p-value, is brought into question. In many statistical modeling applications, the notion of *any* model being correct is difficult to defend, thus reminding one of the famous George Box (1976) quote: “All models are wrong; some are useful.” Unfortunately, hypothesis testing is often performed in settings where the alternative model is unlikely to provide an adequate characterization of the underlying phenomenon. Johnson (1995) provides ecological examples in which hypothesis testing was performed when the alternative model is almost certainly underspecified.

The p-value combines two important pieces of information, namely the estimated effect size and the precision of the effect estimate. A small effect that can be very accurately estimated will yield a small p-value, but may be of no practical significance. For instance, Berkson (1938) illustrates this phenomenon with a hypothetical example in which 200,000 observations are drawn from some population. If a test of normality is administered on this data, then it is quite likely that the p-value will be very small, indicating significant departures from normality. This will occur even when the distribution of the data is approximately bell-shaped. In another example of accurate estimation leading to a small p-value, Riedle et al. (2017) show that the odds of a patient over 85 years of age receiving an antimicrobial prescription without a corresponding visit to a clinician is only 1.055 $(95\% \; CI : (1.023, 1.088))$ times the odds for a patient under 85, yet the p-value is 0.0007. Because this odds ratio is accurately estimated, then a fairly modest effect is able to reach a high level of statistical significance. Fortunately, the authors present the effect estimate and the 95% confidence interval, so the p-value can be placed in proper context.

A final point of confusion regarding the p-value is what Goodman (1999) refers to as the p-value fallacy. The p-value fallacy is based on the mistaken idea that an observation can be viewed from both a short-term and long-term perspective. The p-value is incorrectly assumed to quantify both the behavior of an experiment over repeated trials and the implications of a single observation. Neyman (1937) first described a similar notion regarding confidence intervals. Specifically, once a $100(1-\alpha)\%$ confidence interval is calculated using observed data, then it is incorrect to claim that the probability the confidence interval covers the true parameter value is $100(1-\alpha)\%$. This observed confidence interval either contains the true value or not, and thus the probability is either one or zero. Instead, confidence intervals should be interpreted in terms of long-run probabilities: over many repeated samples, a $100(1-\alpha)\%$ confidence interval should include the true value over $100(1-\alpha)\%$ of the samples.

1.3 p-Value Alternatives

As Efron and Tibshirani (1994) showed, bootstrap resampling techniques can be used to define a bootstrap analogue to hypothesis testing. An analogue to the p-value, known as the achieved significance level (ASL), can be derived from the bootstrap-based hypothesis testing procedure. The ASL uses the bootstrap to construct a reference distribution for a desired test statistic. The observed value of the test statistic is then compared to this bootstrap-based reference distribution. Conceptually, the bootstrap test modifies the original data so as to make the null hypothesis true, while preserving all other aspects of the distribution of the original data. By preserving all other aspects of the data, the bootstrap test is robust to model misspecifications. The estimator of the ASL is the proportion of bootstrap samples drawn from the modified data (defined so that the null is true) in which the bootstrapped test statistic is as extreme as or more extreme than the observed test statistic (which arises from the data generating mechanism in which the null may or may not be true).

While the standard p-value is derived under a classical paradigm, Bayesian analogues to the p-value have been developed. A standard p-value is derived by defining a test statistic that is calculated from observed data. The test statistic must be a pivotal quantity and thus must have a null distribution that does not depend on unknown parameters. The reference distribution for the test statistic is derived. The p-value is then the probability, based on the reference distribution, that a future observation of the test statistic will be as extreme as or more extreme than that which was actually observed.

Bayesian p-values are similar in spirit to frequentist p-values in that they also compare the observed value of a test statistic to a reference distribution. We briefly introduce two Bayesian p-values, the posterior predictive p-value and the prior predictive p-value. When the null hypothesis pre-specifies all parameter values, then the standard p-value is equivalent to both the prior and posterior predictive p-values. However, in the presence of nuisance parameters, Bayesian p-values differ from their frequentist counterparts in how the reference distribution is constructed.

The posterior predictive p-value uses the posterior predictive distribution to derive the reference distribution of the test statistic. Guttman (1967) first introduced the notion of posterior predictive assessment. Rubin (1984) showed that the posterior predictive p-value can be interpreted as the posterior mean of the standard p-value, where averaging is done according to the posterior distribution of the nuisance parameters under the null hypothesis. In contrast to the posterior predictive p-value, Box (1980) proposed using the prior predictive distribution in the calculation of the reference distribution, leading to the prior predictive p-value.

Another Bayesian alternative to the p-value is the Bayes factor. In a hypothesis testing setting, the Bayes factor is the ratio of the posterior odds of the null being true to the prior odds on the null. For a thorough treatment of Bayes factors, see Kass and Raftery (1995). Because scientists may rightly have strong beliefs regarding the relative likelihood of proposed hypotheses, Goodman (1999) argues that the Bayesian paradigm is more aligned with how science should proceed. Namely, if a hypothesis is viewed as very likely (or unlikely), then less (or more) data-based evidence should be required to conclude in favor of that hypothesis. Thus, Goodman (1999) favors using the Bayes factor instead of the p-value in most medical research. Another advantage of the Bayes factor is that it can be viewed both objectively and subjectively. The Bayes factor can be viewed objectively in a manner similar to a likelihood ratio in the sense that if the Bayes factor equals B, then the observed data are B times as likely under the null hypothesis as under the alternative. The Bayes factor can also be viewed subjectively in the sense that it is the ratio of the posterior odds in favor of the null to the prior odds in favor of the null. Thus, the Bayes factor allows a practitioner to understand how the observed data alters the prior belief of the relative merits of two hypotheses.

1.4 p-Value Alternative Interpretations

Ideally, a statistical testing problem would be summarized through a measure of belief and/or a measure of evidence. Specifically, one may wish for the p-value to provide a probability of the null hypothesis being true. The Bayesian framework allows for such a probability in the form of the posterior probability on the null. The p-value, on the other hand, does not typically provide such an interpretation. For instance, Lindley (1957) showed that for a point null hypothesis, the p-value and the posterior probability can vary considerably. According to Shafer (1982), for a point null, the p-value tends to be smaller than the Bayesian posterior probability. Berger and Sellke (1987) provide a more thorough exploration of the relationship between the p-value and the posterior probability. Specifically, the authors consider the p-value derived under a two-sided hypothesis test of a point null in comparison to the posterior probability in favor of the null. The authors consider a variety of classes of priors in the calculation of the posterior probability. For many classes of priors, Berger and Sellke (1987) show that the infimum, over the given class of priors, of the posterior probability on the null is considerably greater than the p-value. Thus, in point null two-sided testing, the p-value will typically overstate the evidence against the null hypothesis in comparison to the Bayesian posterior probability. However, there are situations in which the p-value provides a more desirable interpretation. For instance, Casella and Berger (1987) show that in a one-sided testing problem, an equivalence exists between the p-value and the posterior probability on the null when using a symmetric, noninformative prior. Thus, for certain settings and priors, the p-value can be thought of as a reasonable measure of the probability that the null is true in a Bayesian sense.

Goodman (2001) provides another interpretation of the p-value by transforming it into a minimum Bayes factor (MBF). An advantage of the MBF is that it requires no additional information beyond what is necessary for the p-value. Further, specifying a prior distribution for the alternative model is unnecessary because the MBF is the minimum over all such priors. In a certain sense, the MBF can therefore be thought of as an "objective" measure in that it does not require stipulating a subjective prior distribution for the alternative. While transforming the p-value to the MBF is a straightforward process, Goodman (2001) argues that this conversion represents more than a simple translation of statistical quantities, but instead allows for the information contained in the p-value to be viewed in a different paradigm. To see how, consider that the MBF is the lower bound on the ratio of the null model posterior odds to the null model prior odds. In other words, the MBF gives the ratio of the null model posterior odds to the prior odds, where the alternative model is taken to be the model which best fits the data at hand. Thus, transforming the p-value into the MBF illustrates that the information contained in the p-value can be used to compare the relative merits of the null and alternative hypotheses. In an example provided by Goodman (2001), a p-value of 0.03 is observed. The corresponding MBF is 0.10, indicating that the observed data reduce the odds in favor of the null by, at most, a multiplicative factor of 10. Converting the p-value to the MBF allows the practitioner to consider how much the observed data increases or decreases the odds in favor of the null hypothesis, without having to invoke subjective Bayesian concepts, such as prior distributions.

A further criticism of the p-value is that it may not adhere to the likelihood principle. Intuitively, one feature of this principle is that statistical evidence should only depend on the experiment which was performed and the data observed from the experiment, not on any known or unknown intentions of the researcher performing the experiment. Neath (2017) provides a useful and accessible explanation of the likelihood principle and why the p-value may violate this principle; for a more rigorous treatment of the likelihood principle, see Pawitan (2001) and Berger et al. (1988).

For an intuitive understanding of why, in some instances, the p-value violates the likelihood principle, consider that the p-value is a tail probability based on the probability of getting a result as extreme as or more extreme than that which was observed. The p-value depends not only on the observed value of the test statistic, but also on the sampling distribution of the test statistic under the null hypothesis. Therefore, the p-value depends on the assumed distribution of the data, which can depend on the intentions of the experimenter. Neath (2017) explores a canonical example of this violation in which an experimenter flips a coin 12 times and observes 3 heads. If the intention of the researcher was to flip the coin 12 times and record the number of heads, then this random variable should follow a binomial distribution. On the other hand, if the coin were to be flipped repeatedly until 3 heads were observed, then the total number of flips should follow a negative binomial distribution. Suppose the null hypothesis that the probability of heads is 0.50 is being tested against the lower-sided alternative. Then, if the former experimental design is applied, the p-value is 0.0730, whereas if the latter design is applied, the p-value is 0.0327. Using the same hypotheses and the same data, the p-value can vary depending on the intentions of the experimenter, and thus the p-value is in clear violation of the likelihood principle.

Despite this violation, Neath (2017) argues that when testing a point null, a small modification of such a p-value can lead to a quantity which complies with the likelihood principle. Consider the likelihood ratio (LR) statistic, which adheres to the likelihood principle. Suppose that the testing problem represents the *regular case*, where the log-likelihood can be approximated by a quadratic function. Using a one-to-one transformation, the LR statistic can be translated into a function of the Wald test statistic. We can then solve for the corresponding Wald statistic based on the LR statistic. In settings where a normal approximation is appropriate, this Wald statistic can then be translated into a p-value. Because this p-value ultimately depends only on the LR statistic (which adheres to the likelihood principle), then these two transformations unambiguously lead to a p-value which adheres to the likelihood principle. In the regular case, Wilks's theorem (1938) can also be applied to derive a p-value which complies with the likelihood principle.

1.5 DCP/BDCP

A discrepancy function gauges the separation between a fitted candidate model and the underlying generating model. Discrepancies can be used to delineate between models, with the notion that a smaller discrepancy signifies a model which more closely adheres to the truth. When using discrepancies to select an appropriate model, it is typically unnecessary to assume one of the candidate models represents the truth. Rather, most model selection techniques employing discrepancies seek to determine which model most closely corresponds to the truth, without requiring any candidate model to be precisely true. However, because a discrepancy depends on the unknown data generating process, calculating the discrepancy is impossible. Instead, under suitable conditions, model selection criteria may be derived as asymptotically unbiased estimators of expected discrepancies.

By utilizing discrepancy measures, this dissertation introduces a novel interpretation of the p-value. To derive this alternative interpretation, our work introduces the discrepancy comparison probability (DCP), which is a pairwise model comparison probability based on discrepancy measures. The DCP has broader utility and requires fewer assumptions than hypothesis testing and the p-value. Despite the broader applicability of the DCP, we attempt to establish a connection between the p-value and the DCP when hypothesis testing assumptions are met. Because the constituent discrepancies of the DCP cannot be calculated, an estimator of the DCP is derived using the bootstrap. We show that in certain large-sample settings, the p-value and the bootstrapped discrepancy comparison probability (BDCP) are approximately equal.

To understand the importance of the p-value / BDCP connection, consider the standard interpretation of the p-value, as a probability conditioned on the null being true. The validity of discrepancy functions, and thus the BDCP, does not depend on whether the model is true. Connecting the p-value to the BDCP requires the assumption that the alternative model is adequately specified, but the null may be underspecified. Therefore, by drawing a connection between the BDCP and the p-value, we show that the p-value can at times simply be interpreted as a bootstrap-based probability that the null model is better than the alternative. Realize that the null model may be better without conforming precisely to the truth; if the bias of the null model is negligible compared to the additional estimation error of the alternative model, then the null will be preferred.

While the BDCP can be connected to the p-value, the BDCP can also be employed in settings that will not lead to an approximation of the p-value. We discuss the benefits that can be gleaned from the connection between the BDCP and the p-value, including alternative interpretations of the p-value, which lead to potentially new insights regarding its behavior. The utility of the DCP framework in settings that do not necessarily lead to a natural connection to the p-value will also be considered.

1.6 Dissertation Outline

In Chapter 2, we formally introduce discrepancy functions and the Wald, score and LR p-values. We also provide a thorough introduction of the DCP and demonstrate how to use the bootstrap to derive the BDCP. In Chapter 3, we explore an introductory example in which we connect the BDCP to the p-value in a one-sample test of means setting. We also connect the BDCP to the Wald test p-value in a specific generalized linear modeling (GLM) framework. We mathematically connect the Wald test p-value to the BDCP in a general setting in Chapter 4. In Chapter 5, we mathematically show that the LR test p-value can be connected to the BDCP and examine some of the complications in making this connection. In Chapter 6, we show how the BDCP approximates the score test p-value. Chapter 7 provides a thorough simulation study in which we verify the mathematical results of Chapters 4 - 6 for finite sample sizes. In Chapter 8, we argue that, in general, the bootstrap-based discrepancy estimators are biased. We discuss the ramifications of this bias and suggest methods of correcting for it. In Chapter 9, we discuss the practical implications of connecting the p-value to the BDCP, and argue that the BDCP is valid in many settings in which the p-value is not. We also consider the development of the BDCP for a general discrepancy. In Chapter 10, we illustrate our methodology with two biomedical applications. Finally, in Chapter 11, we provide some concluding remarks.

CHAPTER 2
BACKGROUND

In this chapter, we begin with a brief introduction to discrepancy functions and discuss how the bootstrap can be applied to estimate the distribution of the overall discrepancy. In Section 2.3, we formally introduce the discrepancy comparison probability (DCP), and its bootstrap-based estimator, the BDCP. In Section 2.4, we introduce the Wald, score and LR test p-values in two important settings. In the first setting, the null hypothesis pre-specifies all parameter values. In the other setting, the null only pre-specifies a subset of the parameters, leaving the remaining to be treated as nuisance parameters.

2.1 Discrepancy Functions

Model selection problems often employ discrepancy functions to aid in the choice between competing models. Suppose we have a vector of independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, which is generated from an unknown, true distribution $g(\boldsymbol{y})$, which is not necessarily parametric. Further, suppose a parametric candidate model $f(\boldsymbol{y}|\boldsymbol{\theta})$ is put forth to approximate the observed data $\boldsymbol{y}$. Specifically, assume the candidate model belongs to a parametric class of densities

$$\mathscr{F} = \{f(\boldsymbol{y}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\},$$

where $\boldsymbol{\Theta}$ is the parameter space for $\boldsymbol{\theta}$. A discrepancy function $d(g, f)$ provides a measure of the disparity between the true density $g(\boldsymbol{y})$ and a parametric model $f(\boldsymbol{y}|\boldsymbol{\theta})$

that satisfies

$$d(g, f) \geq d(g, g).$$

A discrepancy function need not be a formal distance metric. However, a discrepancy should still behave in a manner similar to a distance. Namely, as the dissimilarity between $g(\boldsymbol{y})$ and $f(\boldsymbol{y}|\boldsymbol{\theta})$ increases, the discrepancy $d(g, f)$ should increase accordingly. For notational simplicity, we assume candidate parametric models can be characterized by their parameter vector $\boldsymbol{\theta}$, and will thus denote $d(g, f)$ by $d(g, \boldsymbol{\theta})$.

Let $\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log f(\boldsymbol{y}|\boldsymbol{\theta})$ be the natural logarithm of the likelihood function for the candidate model. Accordingly, let $\ell_i(\boldsymbol{\theta}|y_i)$ represent the contribution of the i^{th} observation to the log-likelihood. The Kullback-Leibler (KL) discrepancy between the true model $g(\boldsymbol{y})$ and candidate model $f(\boldsymbol{y}|\boldsymbol{\theta})$ is defined as

$$d_{KL}(g, \boldsymbol{\theta}) = E_g \left\{-2\ell(\boldsymbol{\theta}|\boldsymbol{z})\right\},$$

where E_g denotes expectation with respect to the true distribution g, and $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ is a sample of independent observations drawn from the true distribution g, generated independently of $\boldsymbol{y}$. The KL discrepancy was introduced by Solomon Kullback and Richard Leibler (1951). The KL discrepancy assesses how well the candidate model predicts future data arising from the true distribution.

Quantifying how well the *fitted* candidate model approximates the true distribution is often of interest in a model selection problem. Let $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}}\, \ell(\boldsymbol{\theta}|\boldsymbol{y})$ denote the maximum likelihood estimator of $\boldsymbol{\theta}$. Similarly, let $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ denote the corresponding fitted model, and let $\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) = \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$. The discrepancy between the true model g and the fitted candidate model $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is referred to as the *overall* discrepancy, and is denoted $d(g, \hat{\boldsymbol{\theta}})$. The overall KL discrepancy for the fitted candidate model $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is then

$$d_{KL}(g, \hat{\boldsymbol{\theta}}) = E_g \left\{-2\ell(\boldsymbol{\theta}|\boldsymbol{z})\right\} |_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Note that the overall discrepancy $d(g, \hat{\boldsymbol{\theta}})$ is a random variable, as it is a function of the estimated parameter vector $\hat{\boldsymbol{\theta}}$, and thus depends on $\boldsymbol{y}$. Therefore, it is useful to think of the *distribution* of $d(g, \hat{\boldsymbol{\theta}})$. Ideally, we would like to approximate the distribution of the overall discrepancy $d(g, \hat{\boldsymbol{\theta}})$ by evaluating it over many repeated samples $\boldsymbol{y}$ arising from the true distribution g. However, even if we could draw repeated samples from g, evaluating the overall discrepancy is still not possible because the overall discrepancy depends on the unknown g. Thus, model selection criteria have been developed which seek to estimate the distribution of $d(g, \hat{\boldsymbol{\theta}})$, or some characteristic of its distribution. For instance, under certain regularity conditions, the Akaike information criterion (AIC; Akaike, 1973, 1974) serves as an asymptotically unbiased estimator of the *expected value* of the overall KL discrepancy. For an overview of model selection criteria which focus on the expected value of overall discrepancies, see McQuarrie and Tsai (1998) and Burnham and Anderson (2003).

2.2 Using the Bootstrap to Estimate Discrepancy Functions

Instead of relying on asymptotics to estimate the expected value of the overall discrepancy $d(g, \hat{\boldsymbol{\theta}})$, this paper uses bootstrap resampling techniques to approximate the distribution of $d(g, \hat{\boldsymbol{\theta}})$. Efron (1983, 1986) was the first to develop the idea of using the bootstrap in the model selection context. The simulation results presented later in this paper employ the non-parametric bootstrap, but parametric or semi-parametric techniques can also be used. The non-parametric bootstrap samples are of size n, drawn with replacement from $\boldsymbol{y}$. Note that for most applications, for $i = 1, \ldots, n$, each observation y_i will have corresponding covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})^T$, where for $j = 1, \ldots, J$, x_{ij} denotes the i^{th} observation on the j^{th} covariate. For each observation y_i included in a bootstrap sample, its corresponding covariate vector $\boldsymbol{x}_i$ is also included, and thus the selection of y_i implies the selection of $(y_i, \boldsymbol{x}_i^T)^T$. Following a common convention of modeling notation, we will often let the covariates $\boldsymbol{x}_i$ and the outcome y_i be represented simply by y_i.

We can use the bootstrap to estimate the distribution of the overall discrepancy $d(g, \hat{\boldsymbol{\theta}})$, by applying what Efron and Tibshirani (1994) refer to as the "plug-in principle." The plug-in principle dictates that each element of the overall discrepancy is replaced by its bootstrap analogue. For instance, applying the plug-in principle to the overall KL discrepancy can be summarized by the following replacements:

$$\begin{gathered} g \to \hat{g} \\ \boldsymbol{y} \to \boldsymbol{y}^* \\ E_g \to E_{\hat{g}} \\ \hat{\boldsymbol{\theta}} \to \hat{\boldsymbol{\theta}}^*. \end{gathered}$$

Here, $\hat{g}$ is the empirical distribution; $\boldsymbol{y}^*$ is a bootstrap sample drawn from $\hat{g}$, and $\hat{\boldsymbol{\theta}}^*$ is the MLE of $\boldsymbol{\theta}$ derived under the bootstrap sample $\boldsymbol{y}^*$. Because the observations $y_1, \ldots, y_n$ are independent, the bootstrap analogue to the overall KL discrepancy is then given by

$$\begin{aligned} d_{KL}\left(\hat{g}, \hat{\boldsymbol{\theta}}^*\right) &= E_{\hat{g}}\left\{-2\ell(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^*} \\ &= \sum_{i=1}^{n}\left\{-2\ell_i(\hat{\boldsymbol{\theta}}^*|y_i)\right\} \\ &= -2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}). \end{aligned}$$

Rather than just thinking in terms of the overall KL discrepancy, consider a generic discrepancy d. To derive a bootstrap-based estimator of the distribution of the overall discrepancy, we first draw $b = 1, \ldots, B$ bootstrap samples from $\boldsymbol{y}$. For $b = 1, \ldots, B$, let the MLE of $\boldsymbol{\theta}$ based on the b^{th} bootstrap sample be denoted $\hat{\boldsymbol{\theta}}^*(b)$. Then, for $b = 1, \ldots, B$, calculate the bootstrap analogue to the overall discrepancy $d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right)$. The set

$$\left\{d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right) : b = 1, \ldots, B\right\}$$

serves as a bootstrap-based approximation to the distribution of the overall discrepancy.

2.3 Discrepancy Comparison Probability (DCP)

Suppose that two nested models are put forth to approximate observed data $\boldsymbol{y}$. Delineating between these two models is often done using hypothesis testing, where we choose in favor of the null model unless the p-value is sufficiently small, in which case we reject the null and decide in favor of the alternative. However, deciding between these two competing models could also be done using discrepancy functions. Let the MLE of $\boldsymbol{\theta}$ for the null and alternative models be denoted $\hat{\boldsymbol{\theta}}_0$ and $\hat{\boldsymbol{\theta}}$, respectively, with corresponding overall discrepancies $d(g, \hat{\boldsymbol{\theta}}_0)$ and $d(g, \hat{\boldsymbol{\theta}})$. There exist many ways in which a model can be evaluated in the discrepancy function framework. For instance, suppose one is interested in which of the two models has a smaller expected overall KL discrepancy. Then, choosing the model with the smaller AIC would be an appropriate way of proceeding. However, in this paper, we do not seek to delineate between two models using their expected overall discrepancies, but instead evaluate the models using the probability

$$P = Pr\left[d(g, \hat{\boldsymbol{\theta}}_0) < d(g, \hat{\boldsymbol{\theta}})\right],$$

which we refer to as the **discrepancy comparison probability** (DCP). The DCP is the probability that the fitted null model will be more congruous with the true model than the fitted alternative, as measured by the discrepancy function d. To help better understand the DCP, suppose $P = 0.80$. Then, the fitted null model will have a smaller overall discrepancy than the fitted alternative in 80% of samples of size n drawn from the generating distribution. The null model may be *better* without conforming precisely to the truth; if the bias of the null model is negligible compared to the additional estimation error of the alternative model, then the null will be preferred. Importantly, the DCP is a pairwise comparison of the two competing models because it compares the null and alternative overall discrepancies *derived under the same samples.* Therefore, the DCP is actually a measure on the joint distribution of $d(g, \hat{\boldsymbol{\theta}}_0)$ and $d(g, \hat{\boldsymbol{\theta}})$.

To help better understand how the DCP could be approximated in principle, refer to Figure 2.1. For $k = 1, \ldots, K$, let the k^{th} sample of n independent observations drawn from the generating distribution g be denoted $\boldsymbol{y}_k$. Also, for $k = 1, \ldots, K$, let the null and alternative model MLE of $\boldsymbol{\theta}$ based on the k^{th} bootstrap sample be denoted $\hat{\boldsymbol{\theta}}_0(k)$ and $\hat{\boldsymbol{\theta}}(k)$, respectively. Then, as displayed in this figure, begin approximating the DCP by drawing a first sample $\boldsymbol{y}_1$ from the data generating mechanism g. Based on this first sample, fit the null and alternative models to derive $\hat{\boldsymbol{\theta}}_0(1)$ and $\hat{\boldsymbol{\theta}}(1)$, respectively. Based on the two fitted models, calculate the null and alternative discrepancies, $d(g, \hat{\boldsymbol{\theta}}_0(1))$ and $d(g, \hat{\boldsymbol{\theta}}(1))$, respectively. Determine if $d(g, \hat{\boldsymbol{\theta}}_0(1)) < d(g, \hat{\boldsymbol{\theta}}(1))$. Repeat this process by drawing $k = 1, \ldots, K$ independent samples from g, and determine the proportion of those K samples in which $d(g, \hat{\boldsymbol{\theta}}_0(k)) < d(g, \hat{\boldsymbol{\theta}}(k))$. This proportion would serve as an approximation of the DCP.

Of course, because the true distribution g is unknown, we cannot calculate either $d(g, \hat{\boldsymbol{\theta}}_0)$ or $d(g, \hat{\boldsymbol{\theta}})$. Instead, as outlined in the previous section, we employ bootstrap resampling to approximate their joint distribution. For the null and alternative models, let the MLE of $\boldsymbol{\theta}$ derived using the bootstrap sample be denoted as $\hat{\boldsymbol{\theta}}_0^*$ and $\hat{\boldsymbol{\theta}}^*$, respectively. Also, let the null and alternative model bootstrap-based estimator of the overall discrepancy be denoted $d\left(\hat{g}, \hat{\boldsymbol{\theta}}_0^*\right)$ and $d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*\right)$, respectively. Finally, for $b = 1, \ldots, B$, let the null and alternative model MLE of $\boldsymbol{\theta}$ based on the b^{th} bootstrap sample be denoted $\hat{\boldsymbol{\theta}}_0^*(b)$ and $\hat{\boldsymbol{\theta}}^*(b)$, respectively. Then, for $b = 1, \ldots, B$, we apply the plug-in principle to derive the following empirical approximation of the joint distribution of $d(g, \hat{\boldsymbol{\theta}}_0)$ and $d(g, \hat{\boldsymbol{\theta}})$:

$$\left\{ \left(d\left(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(b)\right), d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right) \right) : b = 1, \ldots, B \right\}.$$

Because the DCP P is of particular interest in this paper, we use the bootstrap to derive an estimator. Let Pr^* denote probability with respect to the joint bootstrap distribution of $d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*)$ and $d(\hat{g}, \hat{\boldsymbol{\theta}}^*)$. Following the plug-in principle, the **bootstrap-based discrepancy comparison probability** P^*, or the BDCP, is then

$$P^* = Pr^* \left[d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*) < d(\hat{g}, \hat{\boldsymbol{\theta}}^*) \right].$$

The BDCP P^* is the probability the bootstrap-based estimator of the overall discrepancy is smaller under the null than alternative. Let $\mathbb{1}(\cdot)$ denote the indicator function. We can approximate P^* by drawing $b = 1, \ldots, B$ bootstrap samples, and calculating

$$\hat{P}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1} \left\{ d\left(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(b)\right) < d\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right) \right\},$$

which is simply the proportion of the B bootstrap samples in which the bootstrap-based overall discrepancy estimator is smaller under the null model than under the alternative.

Refer to Figure 2.2 to better understand how the BDCP is approximated. For $b = 1, \ldots, B$, let the b^{th} bootstrap sample of size n drawn from the empirical distribution $\hat{g}$ be denoted $\boldsymbol{y}_k^*$. Begin approximating the BDCP by drawing a first bootstrap sample $\boldsymbol{y}_1^*$ from $\hat{g}$. Using $\boldsymbol{y}_1^*$, obtain $\hat{\boldsymbol{\theta}}_0^*(1)$ and $\hat{\boldsymbol{\theta}}^*(1)$ by fitting the null and alternative models. Calculate the null and alternative discrepancy estimators, $d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(1))$ and $d(\hat{g}, \hat{\boldsymbol{\theta}}^*(1))$, respectively. Determine if $d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(1)) < d(\hat{g}, \hat{\boldsymbol{\theta}}^*(1))$. Repeat this process by drawing $b = 1, \ldots, B$ samples from $\hat{g}$, and determine the proportion of those B samples in which $d(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(b)) < d(\hat{g}, \hat{\boldsymbol{\theta}}^*(b))$. This proportion serves as an approximation to the BDCP. Conceptually, the BDCP mimics the DCP in that the DCP is a probability based on repeated samples drawn from the generating distribution, whereas the BDCP is a probability based on drawing repeated bootstrap samples from the sample $\boldsymbol{y}$. The bootstrap is, in some sense, assuming the data $\boldsymbol{y}$ is the "truth," and then taking repeated bootstrap samples from this "truth." The BDCP is then the proportion of the samples in which the fitted null model is more congruous with the "truth" $\boldsymbol{y}$ than the fitted alternative.

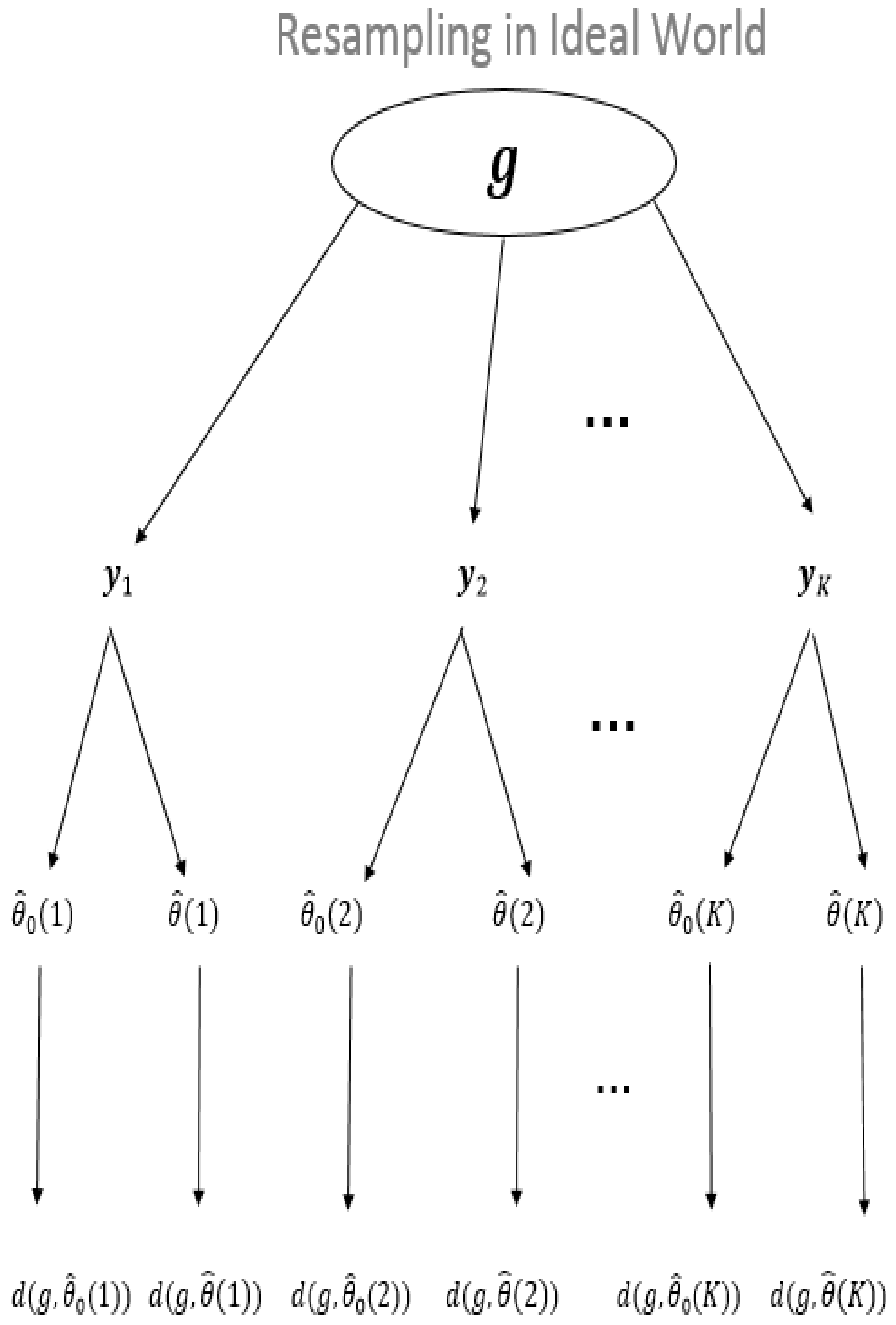


Figure 2.1: Visual representation of the resampling process that would, in principle, be used in the approximation of the DCP.

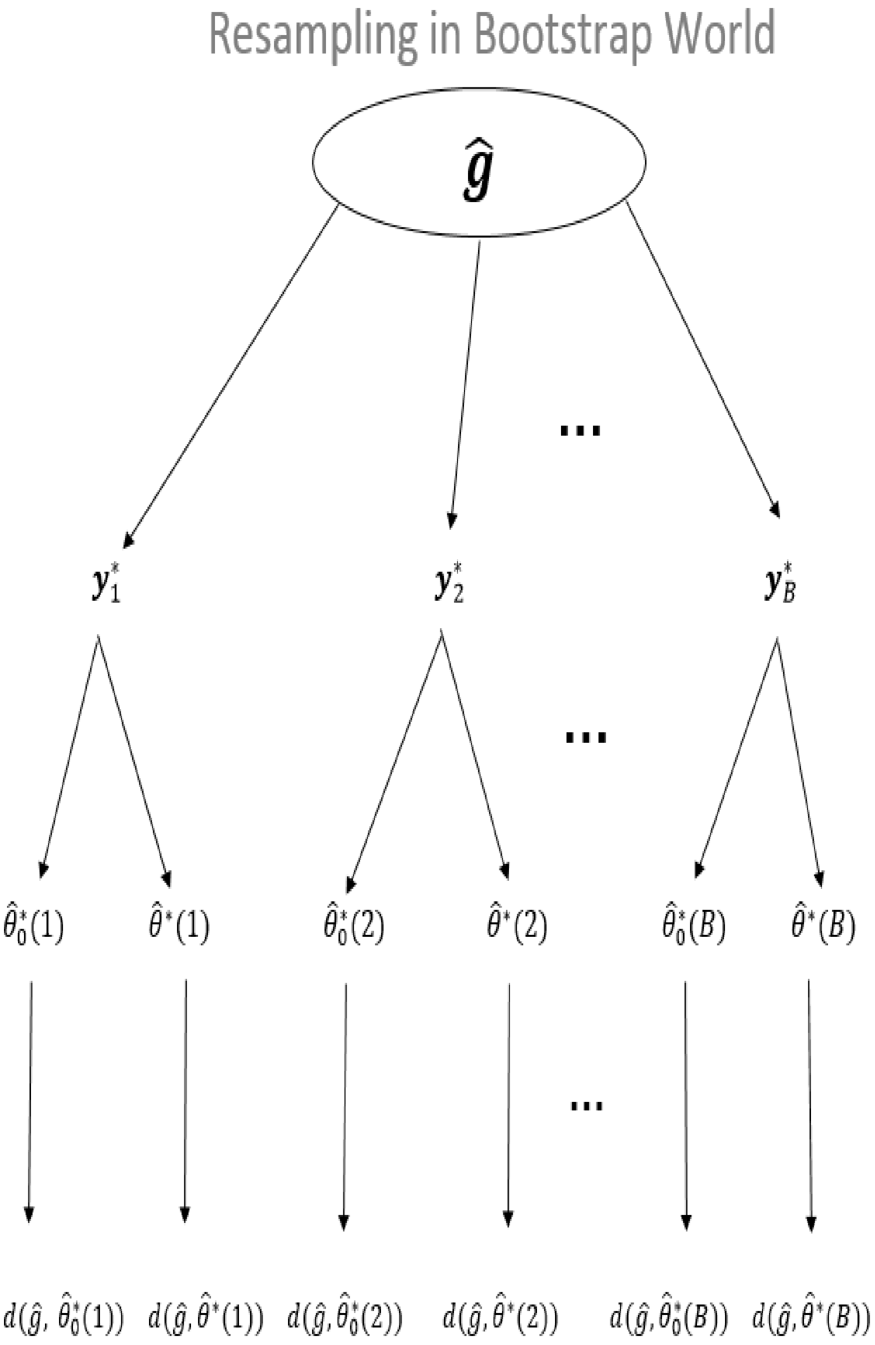

Resampling in Bootstrap World
$\hat{g}$
$\boldsymbol{y}_1^*$
$\boldsymbol{y}_2^*$
$\boldsymbol{y}_B^*$
...
$\hat{\theta}_0^*(1)$
$\hat{\theta}^*(1)$
$\hat{\theta}_0^*(2)$
$\hat{\theta}^*(2)$
$\hat{\theta}_0^*(B)$
$\hat{\theta}^*(B)$
$d(\hat{g}, \hat{\theta}_0^*(1))$
$d(\hat{g}, \hat{\theta}^*(1))$
$d(\hat{g}, \hat{\theta}_0^*(2))$
$d(\hat{g}, \hat{\theta}^*(2))$
$d(\hat{g}, \hat{\theta}_0^*(B))$
$d(\hat{g}, \hat{\theta}^*(B))$

Figure 2.2: Visual representation of the resampling process used in the approximation of the BDCP.

2.4 Wald, Score and Likelihood Ratio (LR) Tests

In Chapters 4, 5 and 6, we show that the Wald, LR and score p-values are approximated by the BDCP using a suitably chosen discrepancy function. This connection will be made in two important settings. First, the connection is made in settings where the null hypothesis pre-specifies all model parameter values, a setting which we call a full null. Second, the approximation is made when the null hypothesis pre-specifies only a subset of the parameters, which we call a partial null.

In this section, we provide the form of the Wald, LR and score p-values in both settings. We first introduce some relevant notation. Let the score vector be denoted

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}|y_i).$$

Let the observed Fisher information be denoted

$$I(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{-\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\sum_{i=1}^{n} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell_i(\boldsymbol{\theta}|y_i) \right).$$

Let the expected Fisher information be denoted

$$\mathscr{I}(\boldsymbol{\theta}) = E\left[I(\boldsymbol{\theta}|\boldsymbol{y})\right] = E\left[U(\boldsymbol{\theta}|\boldsymbol{y})U^T(\boldsymbol{\theta}|\boldsymbol{y})\right].$$

Finally, let $I^{-1}(\boldsymbol{\theta}|\boldsymbol{y})$ and $\mathscr{I}^{-1}(\boldsymbol{\theta})$ denote the inverse of the observed and expected informations, respectively.

In the full null setting, the null hypothesis pre-specifies all parameter values. Thus, the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is tested against the general alternative $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. In the full null setting, the Wald test statistic, is

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathscr{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

The LR statistic is

$$L = 2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_0|\boldsymbol{y})\right),$$

and the score test statistic is

$$S = \boldsymbol{U}^T(\boldsymbol{\theta}_0|\boldsymbol{y})I^{-1}(\boldsymbol{\theta}_0|\boldsymbol{y})\boldsymbol{U}(\boldsymbol{\theta}_0|\boldsymbol{y}).$$

If we let $dim(\boldsymbol{\theta}) = p_A$, then the p-value for each of these tests is

$$p = Pr\left[\chi^2_{p_A} > T\right],$$

where $\chi^2_{p_A}$ is a central chi squared random variable with p_A degrees of freedom, and T is each test's respective test statistic.

In the partial null setting, suppose the null hypothesis pre-specifies k of the p_A parameter values. Let the parameter vector $\boldsymbol{\theta}$ be partitioned into the vector of parameters which the null pre-specifies, denoted by $\boldsymbol{\theta}_{(1)}$, and the parameters which are not pre-specified, denoted $\boldsymbol{\theta}_{(2)}$. We refer to $\boldsymbol{\theta}_{(1)}$ as the parameters of interest and to $\boldsymbol{\theta}_{(2)}$ as the nuisance parameters. The null hypothesis of the form $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ is tested against the alternative $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$. Let $\hat{\boldsymbol{\theta}}_{0(2)} = \text{argmax}_{\boldsymbol{\theta}_{(2)}} \ell(\boldsymbol{\theta}_{0(1)}, \boldsymbol{\theta}_{(2)}|\boldsymbol{y})$ be the MLEs of the nuisance parameters derived under the null hypothesis. Then, let the MLEs derived under the null hypothesis and the sample $\boldsymbol{y}$ be denoted $\hat{\boldsymbol{\theta}}_0 = (\boldsymbol{\theta}^T_{0(1)}, \hat{\boldsymbol{\theta}}^T_{0(2)})^T$. Let the unrestricted MLEs of $\boldsymbol{\theta}$ derived from the sample $\boldsymbol{y}$ also be partitioned so that $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}^T_{(1)}, \hat{\boldsymbol{\theta}}^T_{(2)})^T$.

Let the score vector $\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y})$ be partitioned into the derivatives of the log-likelihood with respect to the parameters of interest and the nuisance parameters.

Specifically, let

$$\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}|\boldsymbol{y})}{\partial \boldsymbol{\theta}_{(1)}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_{(1)}} \ell_i(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}|y_i),$$

be the parameter of interest portion of the score vector, and

$$\boldsymbol{U}_{(2)}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\partial \ell(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}|\boldsymbol{y})}{\partial \boldsymbol{\theta}_{(2)}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}_{(2)}} \ell_i(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}|y_i)$$

be the nuisance parameter portion of the score vector. We will also find it useful to partition the observed and expected information matrices into components consisting of the parameters of interest and nuisance parameters. Write

$$I(\boldsymbol{\theta}|\boldsymbol{y}) = \begin{pmatrix} I_{11}(\boldsymbol{\theta}|\boldsymbol{y}) & I_{12}(\boldsymbol{\theta}|\boldsymbol{y}) \\ I_{21}(\boldsymbol{\theta}|\boldsymbol{y}) & I_{22}(\boldsymbol{\theta}|\boldsymbol{y}) \end{pmatrix},$$

and

$$\mathscr{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathscr{I}_{11}(\boldsymbol{\theta}) & \mathscr{I}_{12}(\boldsymbol{\theta}) \\ \mathscr{I}_{21}(\boldsymbol{\theta}) & \mathscr{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

Similarly, we partition the observed and expected inverse information matrices. Let

$$I^{-1}(\boldsymbol{\theta}|\boldsymbol{y}) = \begin{pmatrix} I^{11}(\boldsymbol{\theta}|\boldsymbol{y}) & I^{12}(\boldsymbol{\theta}|\boldsymbol{y}) \\ I^{21}(\boldsymbol{\theta}|\boldsymbol{y}) & I^{22}(\boldsymbol{\theta}|\boldsymbol{y}) \end{pmatrix},$$

and

$$\mathscr{I}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathscr{I}^{11}(\boldsymbol{\theta}) & \mathscr{I}^{12}(\boldsymbol{\theta}) \\ \mathscr{I}^{21}(\boldsymbol{\theta}) & \mathscr{I}^{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

In the nuisance parameter setting, the Wald test statistic is

$$W = (\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{0(1)})^T \left[\mathscr{I}^{11}(\hat{\boldsymbol{\theta}})\right]^{-1} (\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{0(1)}).$$

The LR statistic is

$$\begin{aligned} L &= 2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y})\right) \\ &= 2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y})\right), \end{aligned}$$

and the score test statistic is

$$\begin{aligned} S &= \left[\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y})\right]^T I^{11}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) \left[\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})|\boldsymbol{y}\right] \\ &= \left[\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y})\right]^T I^{11}(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y}) \left[\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y})\right]. \end{aligned}$$

The p-value for these three tests is

$$p = Pr[\chi^2_k > T],$$

where χ^2_k is a central chi squared random variable with k degrees of freedom, and T is the observed test statistic.

The Wald test was first developed by Abraham Wald (1943). Wilks (1938) first provided the asymptotic distribution of the likelihood ratio statistic. Rao (1948) developed the score test in the full null setting, and Neyman (1979) broadened the applicability of the score test to include settings with nuisance parameters.

CHAPTER 3
INTRODUCTORY EXAMPLES

3.1 BDCP / p-Value Connection: One-Sample Example

To illustrate the concepts presented in the previous sections, we focus on a simple problem. By first considering a simple setting, we lay the groundwork for understanding more general settings, which will be explored later in the dissertation.

Suppose $y_1, \ldots, y_n, y \sim g$ are independent and identically distributed random variables with an unknown mean μ and known variance σ^2. The problem is to predict y after observing $y_1, \ldots, y_n$. Let $\hat{y}$ define the predicted value of y. We will require a judgement on the accuracy of the prediction. Define

$$\begin{aligned}\Delta(\hat{y}) &= E_g(y - \hat{y})^2 \\ &= \sigma^2 + (\hat{y} - \mu)^2\end{aligned}$$

as the mean squared error of prediction. Aside from the distribution variance, prediction accuracy under the mean squared error criterion depends only on the accuracy of $\hat{y}$ as an estimator of the distribution mean μ. So for this problem, a prediction $\hat{y}$ is synonymous with an estimator of μ. The choice of an estimator will depend on the selection of a model. The following models are under consideration:

$$M_0 : y_1, \ldots, y_n, y \sim N(\mu_0, \sigma^2)$$

and

$$M_A : y_1, \ldots, y_n, y \sim N(\mu, \sigma^2).$$

Let $\tilde{\mu}$ represent a model estimator for μ. Under model M_0, we put forth the "estimator" $\tilde{\mu} = \mu_0$ regardless of the observed data. Under model M_A, we use the data information in creating the estimator $\tilde{\mu} = \hat{\mu} = \bar{y}$. The decision between models is to be based on which model puts forth the more accurate estimator.

The problem as stated is reminiscent of a hypothesis testing problem. Data $y_1, \ldots, y_n$ is to be used in testing the null hypothesis $H_0 : \mu = \mu_0$ against a general alternative $H_A : \mu \neq \mu_0$. A decision in a null hypothesis significance testing problem may proceed through the use of a p-value. Define

$$\begin{aligned} p &= 2(1 - \Phi(|z|)) \\ &= Pr(\chi_1^2 > z^2) \end{aligned}$$

where $z = \sqrt{n}(\bar{y} - \mu_0)/\sigma$ is the standardized test statistic. The problem is also reminiscent of a model selection problem within a discrepancy function framework. Write

$$d(\mu, \tilde{\mu}) = (\tilde{\mu} - \mu)^2.$$

The squared error of estimation d may be treated as a discrepancy function, where the fit of a model is judged by a comparison between the estimated mean and the true mean. The discrepancy d is often the focus of interest in a model selection problem. However, d is typically a random variable since the fitted value $\tilde{\mu}$ is often a function of the sample $y_1, \ldots, y_n$. Thus, we instead should say the *distribution* of the discrepancy d is the quantity of interest.

Under general model M_A, the distribution of the discrepancy $d(\mu, \bar{y}) = (\bar{y} - \mu)^2$

is induced from the distribution on the sample mean. However, under the null model M_0, the discrepancy $d(\mu, \mu_0)$ does not involve the observed sample. Its distribution is simply a point mass at $(\mu_0 - \mu)^2$. As outlined in Section 2.1, model selection criteria are often developed by focusing on the expected value of an overall discrepancy. Instead of summarizing the distributions via an expectation, we will base our model evaluation on the discrepancy comparison probability (DCP), introduced in Section 2.3. In this example the DCP is

$$P = Pr\left[d(\mu, \mu_0) < d(\mu, \bar{y})\right] = Pr\left[(\mu_0 - \mu)^2 < (\bar{y} - \mu)^2\right].$$

Thus, the preferred model is that which is most likely to provide the more accurate estimator of the true distribution mean. We may conceptualize the null model discrepancy $(\mu_0 - \mu)^2$ as bias due to model misspecification. Alternatively, we may conceptualize the general model discrepancy $(\bar{y} - \mu)^2$ as an error due to parameter estimation. When model bias is negligible in comparison to estimation error, the null model is preferred. This may be so without the null conforming precisely to the truth. A fundamental aspiration of the discrepancy function approach to model selection is to achieve a balance between goodness-of-fit and parsimony.

We will use the bootstrap to estimate the distributions of the respective discrepancies. Let $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^T$ denote a bootstrap sample of size n drawn from the empirical distribution. The sample mean $\bar{y}$ serves as the empirical distribution mean. Let $\bar{y}^*$ denote the mean of the bootstrap sample. A bootstrap realization from the distribution of $d(\mu, \mu_0)$ is written as

$$\hat{d}(\mu, \mu_0) = d(\bar{y}, \mu_0) = (\mu_0 - \bar{y})^2.$$

Since the distribution of $d(\mu, \mu_0)$ consists of only a point mass, so does its estimator. Denote a bootstrap realization from the distribution of $d(\mu, \bar{y})$ as

$$\hat{d}(\mu, \bar{y}) = d(\bar{y}, \bar{y}^*) = (\bar{y}^* - \bar{y})^2.$$

Repeat for $b = 1, \ldots, B$ to create a bootstrap estimator of the distribution of the alternative model discrepancy. Let Pr^* denote probability with respect to this bootstrap distribution. The bootstrap sampling scheme leads to the following BDCP:

$$P^* = Pr^* \left[(\mu_0 - \bar{y})^2 < (\bar{y}^* - \bar{y})^2 \right]. \tag{3.1}$$

For $b = 1, \ldots, B$, let $\bar{y}^*(b)$ denote the sample mean of the b^{th} bootstrap sample. Then, using the B bootstrap samples, the BDCP can be approximated by

$$\hat{P}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1} \left\{ (\mu_0 - \bar{y})^2 < (\bar{y}^*(b) - \bar{y})^2 \right\}. \tag{3.2}$$

We see in (3.2) the features which define the problem of deciding between two models. Support for the alternative model is strongest when the distance between the null mean and the sample mean is large compared to the sampling variability. These same features appear when we take a hypothesis testing approach to model selection.

We are now in a position to connect the BDCP to the z-test p-value. If the model assumptions on the true distribution are nearly correct, then the sampling distribution of the sample mean is approximately normal,

$$\bar{y} \mathrel{\dot\sim} N\left(\mu, \frac{\sigma^2}{n}\right).$$

Bootstrap resampling is ideal in the case when the empirical distribution captures the features of the true distribution. If the true distribution is approximated by the empirical distribution, then the bootstrap distribution of the sample mean is also approximately normal,

$$\bar{y}^* \dot{\sim} N\left(\bar{y}, \frac{\sigma^2}{n}\right).$$

The bootstrap estimator of the distribution of the overall discrepancy is induced to become

$$(\bar{y}^* - \bar{y})^2 \dot{\sim} \frac{\sigma^2}{n}\chi_1^2,$$

where χ_1^2 denotes a central chi squared random variable with one degree of freedom. That is, the estimated distribution of the sampling error under the general model is a scaled chi squared distribution.

Now, again consider the BDCP in (3.1). Write

$$\begin{aligned} P^* &= Pr^*\left[(\mu_0 - \bar{y})^2 < (\bar{y}^* - \bar{y})^2\right] \\ &\approx Pr\left[\frac{\sigma^2}{n}\chi_1^2 > (\mu_0 - \bar{y})^2\right] \\ &= Pr\left[\chi_1^2 > z^2\right] \\ &= p. \end{aligned}$$

Thus, we have the approximation

$$P^* \approx p. \tag{3.3}$$

Expression (3.3) establishes the p-value as a bootstrap-based estimator of the probability that the null model provides a more accurate estimator than the general model.

3.2 BDCP / Wald p-Value Connection: Simplified GLM Setting

In the previous section, we established that, when the hypothesis test assumptions are met, the one-sample z-test p-value is approximated by the BDCP under the specified discrepancy. In this section, we show that the development applies in a broader setting.

Consider regression coefficient testing within a generalized linear modeling (GLM) framework. Define the competing models

$$M_0 : \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1},$$

and

$$M_A : \eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \beta_j x_j,$$

where η is a function of the mean response. The model M_0 is nested within M_A, varying only in whether β_j is included in the model. Deciding between the null model M_0 and the alternative model M_A is analogous to testing the null hypothesis $H_0 : \beta_j = 0$ versus $H_A : \beta_j \neq 0$, which can be done using a Wald test. As described in Section 2.4, the Wald p-value is

$$p_W = Pr(\chi^2_1 > W),$$

where

$$W = \left(\frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \right)^2,$$

and $\widehat{SE}(\hat{\beta}_j) = \sqrt{\imath^{jj}(\hat{\boldsymbol{\beta}})}$, where $\imath^{jj}(\hat{\boldsymbol{\beta}})$ is the $(j,j)^{th}$ element of $\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$.

Delineating between these models can also be done in the discrepancy function framework. Define the squared error of estimation (SEE) discrepancy as

$$d_{SEE}(\beta_j, \tilde{\beta}_j) = (\tilde{\beta}_j - \beta_j)^2,$$

where $\tilde{\beta}_j$ is the model estimator of β_j. Model M_0 puts forth the estimator $\tilde{\beta}_j = 0$, while M_A estimates β_j using the MLE $\hat{\beta}_j$. The DCP under the SEE discrepancy is

$$P_{SEE} = Pr\left[(0 - \beta_j)^2 < (\hat{\beta}_j - \beta_j)^2\right].$$

The true value β_j is unknown, so we employ the bootstrap to estimate P_{SEE}. Applying the plug-in principle, β_j is replaced by $\hat{\beta}_j$. The alternative model estimator $\hat{\beta}_j$ is replaced by $\hat{\beta}^*_j$, the MLE of β_j under the bootstrap sample. The null model estimator of β_j is 0, regardless of the bootstrap sample. The BDCP is then

$$P^*_{SEE} = Pr^*\left[(0 - \hat{\beta}_j)^2 < (\hat{\beta}^*_j - \hat{\beta}_j)^2\right]. \tag{3.4}$$

Based on bootstrap samples $b = 1, \ldots, B$, the BDCP can be approximated. For, $b = 1, \ldots, B$, let $\hat{\beta}^*_j(b)$ denote the MLE of β_j based on the b^{th} bootstrap sample. Then, write

$$\hat{P}^*_{SEE} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{(0 - \hat{\beta}_j)^2 < (\hat{\beta}^*_j(b) - \hat{\beta}_j)^2\right\}.$$

Assuming the alternative model M_A is adequately specified (i.e. $g(\boldsymbol{y}) \in \mathscr{F}$), then $\hat{\beta}_j$ should follow an approximate normal distribution, so that

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)} \dot{\sim} N(0,1).$$

Applying this result in the bootstrap context, $\hat{\beta}^*_j$ should also follow an approximate normal distribution:

$$\frac{\hat{\beta}^*_j - \hat{\beta}_j}{\widehat{SE}(\hat{\beta}^*_j)} \dot{\sim} N(0,1),$$

where $\widehat{SE}(\hat{\beta}_j^*) = \imath^{jj}(\hat{\boldsymbol{\beta}}^*) \approx \imath^{jj}(\hat{\boldsymbol{\beta}})$. Thus, the approximate distribution of the alternative model bootstrap-based discrepancy estimator is

$$(\hat{\beta}_j^* - \hat{\beta}_j)^2 \mathrel{\dot{\sim}} \imath^{jj}(\hat{\boldsymbol{\beta}})\chi_1^2, \tag{3.5}$$

Applying (3.5) to the inequality involving P_{SEE}^*, displayed in (3.4), allows us to assert

$$\begin{aligned} P_{SEE}^* &= Pr^* \left[(0 - \hat{\beta}_j)^2 < (\hat{\beta}_j^* - \hat{\beta}_j)^2 \right] \\ &\approx Pr \left[(0 - \hat{\beta}_j)^2 < \imath^{jj}(\hat{\boldsymbol{\beta}})\chi_1^2 \right] \\ &= Pr \left[\chi_1^2 > \frac{\hat{\beta}_j^2}{\imath^{jj}(\hat{\boldsymbol{\beta}})} \right] \\ &= Pr[\chi_1^2 > W] \\ &= p_W. \end{aligned} \tag{3.6}$$

When hypothesis testing assumptions are met, (3.6) states that the BDCP under the SEE discrepancy approximates the Wald test p-value in a GLM setting in which the null model has one fewer mean structure parameters than the alternative. In the next chapter, we draw a connection between the Wald test p-value and the BDCP in yet a more general setting.

CHAPTER 4
BDCP / WALD P-VALUE CONNECTION: GENERAL SETTING

In Sections 3.1 and 3.2, we showed that a one-sample z-test p-value for a test on a mean, and the Wald test p-value in a specific GLM setting, can be interpreted as a bootstrap-based estimator of the probability that the overall discrepancy is smaller under the null model than the alternative. In these settings, the discrepancy can be interpreted as a squared error of estimation. We now expand on these results by showing that, when hypothesis testing assumptions are met, the Wald test p-value can be approximated by the BDCP in a general setting.

To understand how the Wald test p-value can be interpreted in such a way, consider a partial null hypothesis which pre-specifies k of a set of p_A parameters. Using the notation introduced in Section 2.4, suppose the null hypothesis of the form $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ is tested against the alternative $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$. As described in Section 2.4, the Wald statistic in this setting is

$$W = (\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{0(1)})^T \left[\mathscr{I}^{11}(\hat{\boldsymbol{\theta}})\right]^{-1} (\hat{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{0(1)}),$$

leading to the Wald test p-value

$$p_W = Pr[\chi^2_k > W].$$

Deciding between the null and alternative models can also be done in a discrepancy function framework. If we were to follow the same procedure as in Section 3.2, then a squared error of estimation discrepancy would be put forth. Perhaps the most obvious analogue to the previous setting's overall SEE discrepancy is to sum the squared differences between the model-based parameter estimators $\tilde{\boldsymbol{\theta}}_{(1)}$ and their true values $\boldsymbol{\theta}_{(1)}$:

$$d_{SEE}(\boldsymbol{\theta}_{(1)}, \tilde{\boldsymbol{\theta}}_{(1)}) = (\tilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)})^T (\tilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}).$$

Let the unrestricted MLEs of $\boldsymbol{\theta}$ derived from the bootstrap sample $\boldsymbol{y}^*$ be partitioned so that $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\theta}}^{*T}_{(1)}, \hat{\boldsymbol{\theta}}^{*T}_{(2)})^T$. Applying the plug-in principle to estimate the null and alternative discrepancies justifies the following replacements: (1) the true parameter of interest vector $\boldsymbol{\theta}_{(1)}$ is replaced by the unrestricted MLE $\hat{\boldsymbol{\theta}}_{(1)}$; (2) the unrestricted MLE under the original sample $\hat{\boldsymbol{\theta}}_{(1)}$ is replaced by the unrestricted MLE under the bootstrap sample $\hat{\boldsymbol{\theta}}^*_{(1)}$, and (3) the null model puts forth the "estimator" $\boldsymbol{\theta}_{0(1)}$, regardless of the sample, so no replacement is necessary.

We will be unable to connect the BDCP under this discrepancy to the Wald p-value. To see why, consider the BDCP under this discrepancy:

$$\begin{aligned} P^*_{SEE} &= Pr^* \left[d_{SEE}(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{0(1)}) < d_{SEE}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}^*_{(1)}) \right] \\ &= Pr^* \left[(\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)}) < (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \right]. \end{aligned}$$

Assuming an adequately specified alternative model, for $j = 1, \ldots, k$,

$$\frac{(\hat{\theta}^*_j - \hat{\theta}_j)^2}{\imath^{jj}(\hat{\boldsymbol{\theta}})} \dot{\sim} \chi^2_1.$$

This result may appear promising in the goal of connecting P^*_{SEE} and the Wald p-value. However, for $m \neq n = 1, \ldots, k$, $cov(\hat{\theta}^*_m, \hat{\theta}^*_n) \neq 0$, in general. Because the parameter estimates may be correlated, then

$$\sum_{j=1}^{k} \frac{(\hat{\theta}_j^* - \hat{\theta}_j)^2}{\imath^{jj}(\hat{\boldsymbol{\theta}})} \; \dot{\not\sim} \; \chi^2_k.$$

Therefore, the alternative model SEE discrepancy estimator does not follow an approximate χ^2 distribution, and we will thus be unable to connect the BDCP to the Wald p-value using this discrepancy.

Instead, we define a discrepancy which is similar in spirit to the SEE discrepancy, but also yields a BDCP which approximates the Wald p-value. Define the **overall Wald discrepancy** as

$$d_W(\boldsymbol{\theta}_{(1)}, \tilde{\boldsymbol{\theta}}_{(1)}) = (\tilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)})^T \left[\mathscr{I}^{11}(\boldsymbol{\theta})\right]^{-1} (\tilde{\boldsymbol{\theta}}_{(1)} - \boldsymbol{\theta}_{(1)}).$$

Applying the plug-in principle, the alternative discrepancy estimator is

$$d_W(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}^*_{(1)}) = (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)}).$$

Note that the bootstrapped Wald discrepancy estimator contains the observed information rather than the expected information as a consequence of the plug-in principle. Similarly, the bootstrap estimator of the null model discrepancy is

$$d_W(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{0(1)}) = (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)}). \tag{4.1}$$

The BDCP under the Wald discrepancy is then

$$\begin{aligned} P_W^* &= Pr^*\left[d_W(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{0(1)}) < d_W(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}^*_{(1)})\right] \\ &= Pr^*\Bigg[(\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \\ &\qquad\qquad < (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})\Bigg]. \end{aligned} \tag{4.2}$$

Proposition 4.1. *Assuming that the large-sample properties of the MLEs hold and that the alternative model is adequately specified, then for testing a partial null hypothesis of* $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ *versus the alternative of* $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$,

$$p_W \approx P_W^*.$$

Proof. We begin by stating a well-known result pertaining to maximum likelihood estimators which will be applied later in the proof. Under certain regularity conditions and an adequately specified alternative model, for large n,

$$\hat{\boldsymbol{\theta}}_{(1)} \dot{\sim} N_k \left(\boldsymbol{\theta}_{(1)}, \mathscr{I}^{11}(\boldsymbol{\theta})\right). \tag{4.3}$$

Conditioning on the observed data $\boldsymbol{y}$, the null model discrepancy estimator $d_W(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{0(1)})$, displayed in (4.1), is fixed. For large n, under certain regularity conditions, we have that $I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \approx \mathscr{I}^{11}(\hat{\boldsymbol{\theta}})$, and thus the null discrepancy estimator should approximate the Wald test statistic:

$$\begin{aligned} d_W(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{0(1)}) &= (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \\ &\approx (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[\mathscr{I}^{11}(\hat{\boldsymbol{\theta}})\right]^{-1} (\boldsymbol{\theta}_{0(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \\ &= W. \end{aligned} \tag{4.4}$$

By applying (4.4) to (4.2), we can write

$$P_W^* \approx Pr^* \left[(\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)}) > W\right]. \tag{4.5}$$

Thus, in order to complete the proof, we need to show that under Pr^*, the term on the left-hand side of the inequality in (4.5) follows an approximate χ^2_k distribution:

$$(\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \dot{\sim} \chi^2_k.$$

Recall that for large n, under certain regularity conditions, $I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \approx \mathscr{I}^{11}(\hat{\boldsymbol{\theta}})$. Thus, assuming the data at hand adequately characterizes the sampling distribution of $\hat{\boldsymbol{\theta}}_{(1)}$ via the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*_{(1)}$, then applying the large-sample result (4.3) to the bootstrapping context yields

$$\begin{aligned} \hat{\boldsymbol{\theta}}^*_{(1)} &\mathrel{\dot\sim} N_k\left(\hat{\boldsymbol{\theta}}_{(1)}, \mathscr{I}^{11}(\hat{\boldsymbol{\theta}})\right) \\ &\mathrel{\dot\sim} N_k\left(\hat{\boldsymbol{\theta}}_{(1)}, I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right) \end{aligned}$$

Thus, we write

$$(\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \mathrel{\dot\sim} \chi^2_k. \tag{4.6}$$

Applying (4.6) to the inequality involving P^*_W displayed in (4.2) allows us to assert

$$\begin{aligned} P^*_W &\approx Pr^*\left[(\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)})^T \left[I^{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right]^{-1} (\hat{\boldsymbol{\theta}}^*_{(1)} - \hat{\boldsymbol{\theta}}_{(1)}) > W\right] \\ &\approx Pr\left[\chi^2_k > W\right] \\ &= p. \end{aligned}$$

This completes the proof. □

This proof holds regardless of whether the null hypothesis pre-specifies all parameters. We are thus able to interpret the Wald p-value as a bootstrap-based estimator that the overall Wald discrepancy will be smaller under the null model than under the alternative.

CHAPTER 5
BDCP / LR P-VALUE CONNECTION

In this chapter, we establish a connection between the LR test p-value and the BDCP. In Chapter 4, regardless of the presence of nuisance parameters, the null model bootstrap-estimated Wald discrepancy $d_W(\hat{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{0(1)})$ was fixed under Pr^*, allowing for a connection between the Wald test p-value and the BDCP which did not necessitate distinguishing between the full and partial null settings. Unlike the Wald p-value, connecting the LR p-value to the BDCP will require differentiating between the full and partial null settings. Specifically, to connect the BDCP to the LR p-value, different discrepancies will be necessary in the full and partial null settings. In Section 5.1 we show that in the full null setting, the LR p-value is approximated by the BDCP under the KL discrepancy. However, in the partial null setting, establishing this connection between the LR p-value and the KL discrepancy is not always possible. Instead, in Section 5.2, we introduce the "parameters of interest" Kullback-Leibler (PIKL) discrepancy, and show that the BDCP under the PIKL discrepancy approximates the LR test p-value in the partial null setting.

5.1 Full Null Setting

In the full null setting, the null hypothesis pre-specifies all parameter values. Using the notation from Section 2.4, the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is tested against the general alternative $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. If we let $dim(\boldsymbol{\theta}) = p_A$, with the likelihood ratio statistic denoted by

$$L = 2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_0|\boldsymbol{y})\right), \tag{5.1}$$

then the likelihood ratio test p-value for these hypotheses is

$$p_{LR} = Pr[\chi^2_{p_A} > L]. \tag{5.2}$$

Also, recall from Section 2.1 that the overall KL discrepancy of the model corresponding to the alternative hypothesis is

$$d_{KL}(g, \hat{\boldsymbol{\theta}}) = E_g\left\{-2\ell(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

whose bootstrap-based estimator is

$$d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*) = -2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}). \tag{5.3}$$

Similarly, the null overall KL discrepancy is

$$d_{KL}(g, \boldsymbol{\theta}_0) = E_g\left\{-2\ell(\boldsymbol{\theta}_0|\boldsymbol{z})\right\}.$$

Unlike the alternative model whose parameter vector is maximized over its parameter space, the null parameter vector is fixed at $\boldsymbol{\theta}_0$ in the full null setting. Therefore, the bootstrap-based “estimator” of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_0$ for all bootstrap samples. The bootstrap-based estimator of the null overall KL discrepancy is then

$$d_{KL}(\hat{g}, \boldsymbol{\theta}_0) = -2\ell(\boldsymbol{\theta}_0|\boldsymbol{y}). \tag{5.4}$$

Note that conditioned upon the observed data $\boldsymbol{y}$, the bootstrap-based estimator of the null overall KL discrepancy $d_{KL}(\hat{g}, \boldsymbol{\theta}_0)$ is actually fixed, and thus does not vary from one bootstrap sample to the next. Applying the null and alternative bootstrap-based KL discrepancy estimators in (5.3) and (5.4), respectively, yields the following BDCP:

$$P^*_{KL} = Pr^* \left[-2\ell(\boldsymbol{\theta}_0|\boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \right]. \tag{5.5}$$

Proposition 5.1.1. *Assuming that the large-sample properties of the MLEs hold and that the alternative model is adequately specified, then for testing a full null hypothesis of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus the alternative of $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,*

$$p_{LR} \approx P^*_{KL}.$$

Proof. We begin the proof by stating a well-known result from maximum likelihood estimation which will be applied later in the proof. Recall that for large n, under certain regularity conditions with the alternative model being adequately specified,

$$\hat{\boldsymbol{\theta}} \dot{\sim} N_{p_A}\left(\boldsymbol{\theta}, \mathscr{I}^{-1}(\boldsymbol{\theta})\right).$$

It follows that

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathscr{I}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi^2_{p_A}. \tag{5.6}$$

Recall from (5.5) that the BDCP in this setting is

$$P^*_{KL} = Pr^* \left[-2\ell(\boldsymbol{\theta}_0|\boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \right].$$

Because Pr^* denotes probability with respect to the joint distribution of $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ and $d_{KL}(\hat{g}, \boldsymbol{\theta}_0)$, we can conceptualize the observed data $\boldsymbol{y}$ as being fixed under Pr^*. The bootstrap sample, and thus $\hat{\boldsymbol{\theta}}^*$, are random under Pr^*. We re-arrange P^*_{KL} so as to introduce the likelihood ratio statistic L from (5.1):

$$\begin{aligned} P^*_{KL} &= Pr^* \left[-2\ell(\boldsymbol{\theta}_0|\boldsymbol{y}) < -2\ell\left(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}\right)\right] \\ &= Pr^* \left[2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_0|\boldsymbol{y})\right) < 2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right)\right] \\ &= Pr^* \left[2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right) > L\right]. \end{aligned} \tag{5.7}$$

Under fixed observed data $\boldsymbol{y}$, the likelihood ratio statistic L is fixed. In order to show that the LR p-value in (5.2) is approximated by P^*_{KL} in (5.7), we need to show that under Pr^*,

$$2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right) \dot{\sim} \chi^2_{p_A}.$$

Consider taking a second-order Taylor series expansion of $\ell\left(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}\right)$ about $\hat{\boldsymbol{\theta}}$, which yields:

$$\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \approx \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T I(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

For large n, the observed information is approximated by the expected information, and thus we can write

$$2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right) \approx (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T \mathscr{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}). \tag{5.8}$$

Assuming the data at hand adequately characterizes the sampling distribution of $\hat{\boldsymbol{\theta}}$ via the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*$, then applying the large-sample result (5.6) to the bootstrapping context yields

$$(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T \mathscr{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) \dot{\sim} \chi^2_{p_A}. \tag{5.9}$$

Applying (5.8) and (5.9), we have

$$2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right) \dot{\sim} \chi^2_{p_A}.$$

Referring back to (5.7), we can establish our desired result:

$$\begin{aligned} P^*_{KL} &= Pr^* \left[2 \left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \right) > L \right] \\ &\approx Pr \left[\chi^2_{p_A} > L \right] \\ &= p_{LR}. \end{aligned} \tag{5.10}$$

We have thus shown that, assuming an adequately specified alternative model, $P^*_{KL} \approx p_{LR}$ when the null hypothesis pre-specifies all parameter values. This completes the proof. □

Result (5.10) establishes the LR p-value as a bootstrap-based estimator of the probability that the null is "better" than the alternative, as measured by the bootstrapped overall KL discrepancy. Clearly, the p-value is not the probability that the null is *true*, but it is not a requirement for the null model to be true for it to better than the alternative. If $\boldsymbol{\theta}_0$ provides a reasonable characterization of $\boldsymbol{\theta}$, then due to the sampling variability incurred under the alternative model, the null model may be more accurate than the alternative model.

Because the null hypothesis pre-specifies all parameter values in the full null setting, the bootstrap-based parameter vector estimator for the null model is fixed under Pr^*, leading to the null bootstrap-based KL discrepancy estimator $d(\hat{g}, \boldsymbol{\theta}_0)$ also being fixed. In the following section, we address the partial null setting, in which the null hypothesis does not pre-specify all parameter values.

5.2 Partial Null Setting

In the partial null setting, we again wish to show that the LR p-value is approximated by the BDCP. As described in Section 2.4, if we let the LR statistic be denoted

$$\begin{aligned} L &= 2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y})\right) \\ &= 2\left(\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y})\right), \end{aligned}$$

then the likelihood ratio test p-value in the partial null setting is

$$p_{LR} = Pr\left[\chi^2_k > L\right].$$

Unlike in the full null setting, we will be unable to connect the LR p-value to the BDCP under the conventional KL discrepancy. That the BDCP under the conventional KL discrepancy does not approximate the p-value in this setting does not preclude one from using the KL discrepancy; the BDCP under the KL discrepancy is still a valid tool for deciding between two competing models, it simply does not have the connection with the LR p-value which we seek.

To draw a connection between the p-value and the BDCP, we instead use a modified version of the KL discrepancy, which we refer to as the **parameter of interest Kullback-Leibler** (PIKL) **discrepancy**. The PIKL discrepancy constitutes only a small modification of the conventional KL discrepancy and is a sensible tool for evaluating models which contain nuisance parameters. Of particular importance to this dissertation, the BDCP under the PIKL discrepancy approximates the LR p-value in the partial null setting, as we will soon establish. To understand the

PIKL discrepancy, we first introduce the notion of the pseudo-true parameter $\bar{\boldsymbol{\theta}}$. The pseudo-true parameter is defined as the parameter value which minimizes the KL discrepancy. Write

$$\bar{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, d_{KL}(g, \boldsymbol{\theta}).$$

If the model is adequately specified (i.e. $g(\boldsymbol{y}) \in \mathscr{F}$), then $\bar{\boldsymbol{\theta}}$ is the true value of $\boldsymbol{\theta}$. However, $\bar{\boldsymbol{\theta}}$ is well-defined, regardless of whether the model is adequately specified. Let the pseudo-true nuisance parameter vector for the null model be denoted $\bar{\boldsymbol{\theta}}_{0(2)} = \operatorname{argmin}_{\boldsymbol{\theta}_{(2)}} d_{KL}\left(g, (\boldsymbol{\theta}_{0(1)}, \boldsymbol{\theta}_{(2)})\right)$. For the alternative model, let the pseudo-true parameter vector be denoted by $\left(\bar{\boldsymbol{\theta}}_{(1)}^T, \bar{\boldsymbol{\theta}}_{(2)}^T\right)^T = \operatorname{argmin}_{\left(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}\right)} d_{KL}\left(g, \left(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}\right)\right)$. Also, let the MLEs of the parameters of interest, derived under the restriction that $\boldsymbol{\theta}_{(2)} = \bar{\boldsymbol{\theta}}_{(2)}$, be denoted $\hat{\boldsymbol{\theta}}_{C(1)} = \operatorname{argmax}_{\boldsymbol{\theta}_{(1)}} \ell\left(\left(\boldsymbol{\theta}_{(1)}, \bar{\boldsymbol{\theta}}_{(2)}\right) | \boldsymbol{y}\right)$. We will refer to $\hat{\boldsymbol{\theta}}_{C(1)}$ as the conditional MLEs, to emphasize their dependence on the condition $\boldsymbol{\theta}_{(2)} = \bar{\boldsymbol{\theta}}_{(2)}$.

The overall PIKL discrepancy evaluates the KL discrepancy at the pseudo-true values of the nuisance parameters and at the conditional MLEs of the parameters of interest. Specifically, the overall PIKL discrepancy for the null model is

$$d_{PIKL}\left(g, (\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)})\right) = E_g\left\{-2\ell(\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)} | \boldsymbol{z})\right\},$$

and for the alternative, it is

$$d_{PIKL}\left(g, (\hat{\boldsymbol{\theta}}_{C(1)}, \bar{\boldsymbol{\theta}}_{(2)})\right) = E_g\left\{-2\ell(\boldsymbol{\theta}_{(1)}, \bar{\boldsymbol{\theta}}_{(2)} | \boldsymbol{z})\right\}|_{\boldsymbol{\theta}_{(1)} = \hat{\boldsymbol{\theta}}_{C(1)}}.$$

Note that the null model overall PIKL discrepancy $d_{PIKL}\left(g, (\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)})\right)$ is fixed, as is $d_{KL}(g, \boldsymbol{\theta}_0)$ in the full null setting.

To use the bootstrap to estimate the null and alternative discrepancies, we must apply the plug-in principle to the pseudo-true parameter vector. The pseudo-true parameter vector $\bar{\boldsymbol{\theta}}$ minimizes the KL discrepancy under the true distribution g, so the bootstrap-based version of $\bar{\boldsymbol{\theta}}$ should minimize the bootstrap-based KL discrepancy estimator $-2\ell(\boldsymbol{\theta}|\boldsymbol{y})$. Therefore, we use $\hat{\boldsymbol{\theta}}$ as the plug-in for $\bar{\boldsymbol{\theta}}$. Accordingly, the bootstrap-based pseudo-true nuisance parameters for the null and alternative models are $\hat{\boldsymbol{\theta}}_{0(2)}$ and $\hat{\boldsymbol{\theta}}_{(2)}$, respectively. Let $\hat{\boldsymbol{\theta}}^*_{C(1)} = \text{argmax}_{\boldsymbol{\theta}_{(1)}}\, \ell\left(\left(\boldsymbol{\theta}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}\right)|\boldsymbol{y}^*\right)$. Then, the null model bootstrap-based PIKL discrepancy estimator is

$$d_{PIKL}\left(\hat{g}, (\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})\right) = -2\ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}),$$

and the alternative model bootstrap-based estimator is

$$d_{PIKL}\left(\hat{g}, (\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right) = -2\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}).$$

The BDCP under the PIKL discrepancy is then

$$\begin{aligned} P^*_{PIKL} &= Pr^*\left[d_{PIKL}\left(\hat{g}, (\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})\right) < d_{PIKL}\left(\hat{g}, (\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right)\right] \\ &= Pr^*\left[-2\ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right]. \end{aligned}$$

Proposition 5.2.1. *Assuming that the large-sample properties of the MLEs hold and that the alternative model is adequately specified, then for testing a partial null hypothesis of* $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ *versus the alternative of* $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$,

$$p_{LR} \approx P^*_{PIKL}.$$

Proof. We begin by stating well-known results pertaining to score statistics which will be applied later in the proof. First, recall that for large n, under certain regularity conditions with the alternative model being adequately specified,

$$\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{y}) \dot{\sim} N_{p_A}\left(\mathbf{0}, \mathscr{I}(\boldsymbol{\theta})\right).$$

Thus, the score vector for the parameters of interest also follows an approximate normal distribution:

$$\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{y}) \dot{\sim} N_k\left(\mathbf{0}, \mathscr{I}_{11}(\boldsymbol{\theta})\right).$$

The preceding result leads to

$$\boldsymbol{U}_{(1)}^T(\boldsymbol{\theta}|\boldsymbol{y})\mathscr{I}_{11}^{-1}(\boldsymbol{\theta})\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{y}) \dot{\sim} \chi_k^2. \tag{5.11}$$

We now start by adding $2\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})$ to each side of the inequality that defines P_{PIKL}^*, yielding:

$$\begin{aligned}
P_{PIKL}^* &= Pr^*[-2\ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) < -2\ell(\hat{\boldsymbol{\theta}}_{C(1)}^*, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})] \\
&= Pr^*\left[2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y})\right)\right. \\
&\qquad\quad \left.< 2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_{C(1)}^*, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right)\right] \\
&= Pr^*\left[2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_{C(1)}^*, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right) > L\right].
\end{aligned} \tag{5.12}$$

Thus, in order to complete the proof, we need to show that under Pr^*, the term on the left-hand side of the inequality in (5.12) follows an approximate χ_k^2 distribution:

$$2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_{C(1)}^*, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right) \dot{\sim} \chi_k^2.$$

To establish this result, consider taking a second-order Taylor series expansion of $\ell(\hat{\boldsymbol{\theta}}_{C(1)}^*, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})$ around $\left(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}\right)$. Write

$$\ell(\hat{\boldsymbol{\theta}}_{C(1)}^*, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \approx \ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \frac{1}{2}\left(\hat{\boldsymbol{\theta}}_{C(1)}^* - \hat{\boldsymbol{\theta}}_{(1)}\right)^T I_{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\left(\hat{\boldsymbol{\theta}}_{C(1)}^* - \hat{\boldsymbol{\theta}}_{(1)}\right). \tag{5.13}$$

Replacing the observed information with the expected information, approximation (5.13) implies that

$$2\left[\ell(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right] \approx \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^T \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right). \quad (5.14)$$

Applying a first-order Taylor series expansion of $\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*)$ about $(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)})$ leads to

$$\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*) \approx \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*) - I_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*)\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right).$$

Based on the definition of $\hat{\boldsymbol{\theta}}^*_{C(1)}$, we have that $\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*) = \boldsymbol{0}$. Thus, we write

$$\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*) \approx I_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*)\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right). \quad (5.15)$$

For large n, the bootstrap distribution of $\boldsymbol{y}^*$ should mimic the sampling distribution of $\boldsymbol{y}$, thus leading to $I_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*) \approx I_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})$. Also for large n, under certain regularity conditions, we have that $I_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \approx \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)})$. Therefore, under these conditions, $I_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}^*) \approx \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}_{(1)},\hat{\boldsymbol{\theta}}_{(2)})$. This approximation in combination with approximation (5.15) leads to:

$$\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)} \approx \mathscr{I}_{11}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*). \quad (5.16)$$

Result (5.16) implies that

$$\begin{aligned}\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^T \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right) &\approx \boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*)\mathscr{I}_{11}^{-1}(\hat{\boldsymbol{\theta}})\mathscr{I}_{11}(\hat{\boldsymbol{\theta}})\mathscr{I}_{11}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*) \\ &= \boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*)\mathscr{I}_{11}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*). \quad (5.17)\end{aligned}$$

Applying (5.11) to the bootstrapping context, we have

$$\boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*)\mathscr{I}_{11}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}^*) \mathrel{\dot\sim} \chi^2_k. \quad (5.18)$$

Combining (5.17) and (5.18), we see that

$$\left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right)^T \mathscr{I}_{11}(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}\right) \dot{\sim} \chi^2_k. \tag{5.19}$$

In conjunction with (5.14), the preceding distributional result yields the desired distributional result:

$$2\left[\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right] \dot{\sim} \chi^2_k. \tag{5.20}$$

Finally, applying (5.20) to the inequality involving P^*_{PIKL} displayed in (5.12) allows us to assert

$$\begin{aligned} P^*_{PIKL} &= Pr^*\left[2\left(\ell(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right) > L\right] \\ &\approx Pr[\chi^2_k > L] \\ &= p_{LR}. \end{aligned}$$

This completes the proof. □

We have thus shown that the LR p-value can be approximated by a BDCP under an appropriately chosen discrepancy, regardless of whether the null hypothesis pre-specifies all parameter values. In this section, we focus on the PIKL discrepancy because its BDCP provides an approximation to the LR p-value in the partial null setting. However, we again note that the BDCP under the conventional KL discrepancy is also a valid tool for choosing between two competing models in the partial null setting.

At first glance, the overall PIKL discrepancy may seem convoluted because it depends on the pseudo-true nuisance parameters, which are unknown. Thus, we cannot evaluate the overall PIKL. However, we are also unable to evaluate the conventional overall KL discrepancy, and thus the inability to evaluate the PIKL discrepancy causes no additional duress. Instead, both the PIKL and KL discrepancies are easily estimated using the bootstrap. The PIKL may also be appealing from a practical standpoint; if one is concerned with only the parameters of interest, then setting the nuisance parameters to their best possible values, as both the PIKL and its bootstrap-based estimator do, is reasonable. The PIKL discrepancy and its estimator are also akin to a plug-in likelihood in which the nuisance parameter vector is evaluated at its global MLE.

CHAPTER 6
BDCP / SCORE P-VALUE CONNECTION

Similar to the previous chapter, connecting the score test p-value to the BDCP again requires distinguishing between the full and partial null settings. In Section 6.1, we introduce the score discrepancy and show that the BDCP under this discrepancy approximates the score test p-value in the full null setting. We introduce the parameter of interest piecewise score (PIPS) discrepancy in Section 6.2, and show an equivalence between the score test p-value and the BDCP under the PIPS discrepancy in the partial null setting.

6.1 Full Null Setting

Again using the notation of Section 2.4, in the full null setting, the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ is tested against the general alternative $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. If we let $dim(\boldsymbol{\theta}) = p_A$, with the score test statistic denoted by

$$S = \boldsymbol{U}^T(\boldsymbol{\theta}_0|\boldsymbol{y})I^{-1}(\boldsymbol{\theta}_0|\boldsymbol{y})\boldsymbol{U}(\boldsymbol{\theta}_0|\boldsymbol{y}),$$

then the score test p-value for these hypotheses is

$$p_S = Pr[\chi^2_{p_A} > S], \tag{6.1}$$

where $\chi^2_{p_A}$ is a central chi-square random variable with p_A degrees of freedom.

We seek to define a discrepancy under which the BDCP approximates the score test p-value. Let $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}^*$ denote model-based parameter estimators of $\boldsymbol{\theta}$ using the original sample and bootstrap sample, respectively. Then, let the **overall score**

discrepancy be defined as

$$d_S(g, \tilde{\boldsymbol{\theta}}) = \left[E_g \left\{\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right]^T \left[\mathscr{I}(\tilde{\boldsymbol{\theta}})\right]^{-1} \left[E_g \left\{\boldsymbol{U}(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right].$$

The bootstrap-based estimator of the overall score discrepancy is

$$\begin{aligned} d_S(\hat{g}, \tilde{\boldsymbol{\theta}}^*) &= \left[\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_i(\tilde{\boldsymbol{\theta}}^*|y_i)\right]^T \left[\sum_{i=1}^{n} -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \ell_i(\tilde{\boldsymbol{\theta}}^*|y_i)\right]^{-1} \left[\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_i(\tilde{\boldsymbol{\theta}}^*|y_i)\right] \\ &= \boldsymbol{U}^T(\tilde{\boldsymbol{\theta}}^*|\boldsymbol{y}) I^{-1}(\tilde{\boldsymbol{\theta}}^*|\boldsymbol{y}) \boldsymbol{U}(\tilde{\boldsymbol{\theta}}^*|\boldsymbol{y}) \end{aligned}$$

Thus, the BDCP under the score discrepancy is

$$\begin{aligned} P_S^* &= Pr^* \left[d_S(\hat{g}, \boldsymbol{\theta}_0) < d_S(\hat{g}, \hat{\boldsymbol{\theta}}^*)\right] \\ &= Pr^* \left[\boldsymbol{U}^T(\boldsymbol{\theta}_0|\boldsymbol{y}) I^{-1}(\boldsymbol{\theta}_0|\boldsymbol{y}) \boldsymbol{U}(\boldsymbol{\theta}_0|\boldsymbol{y}) < \boldsymbol{U}^T(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) I^{-1}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right] \\ &= Pr^* \left[\boldsymbol{U}^T(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) I^{-1}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) > S\right]. \end{aligned} \tag{6.2}$$

Proposition 6.1.1. *Assuming that the large-sample properties of the MLEs hold and that the alternative model is adequately specified, then for testing a full null hypothesis of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus the alternative of $H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$,*

$$p_S \approx P_S^*.$$

Proof. Under Pr^*, we can think of the observed data $\boldsymbol{y}$ as being fixed. The bootstrap sample is random and thus $\hat{\boldsymbol{\theta}}^*$ and $\boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})$ are also random. In order to show that the score p-value in (6.1) is approximated by P_S^* in (6.2), we must argue that under Pr^*,

$$\boldsymbol{U}^T(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) I^{-1}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \dot{\sim} \chi^2_{p_A}.$$

Let $L(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) = f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}^*)$, and let $E_*(\cdot)$ denote expectation taken with respect to the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*$. Then, consider the expectation of $\boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})$ under Pr^* by writing:

$$\begin{aligned}
E_*\left\{\boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\} &= E_*\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\} \\
&= \int L(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\frac{\partial L(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})/\partial\boldsymbol{\theta}}{L(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})}d\hat{\boldsymbol{\theta}}^* \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\int L(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})d\hat{\boldsymbol{\theta}}^* \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}(1) \\
&= \mathbf{0}.
\end{aligned}$$

Because $E_*\left\{\boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\} = \mathbf{0}$, the multivariate central limit theorem can be invoked to arrive at the following approximate distribution:

$$\boldsymbol{U}\left(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}\right) \dot{\sim} N_{p_A}\left(\mathbf{0}, \mathscr{I}\left(\hat{\boldsymbol{\theta}}^*\right)\right).$$

For large n, the expected information is approximated by the observed information, and thus we write

$$\boldsymbol{U}\left(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}\right) \dot{\sim} N_{p_A}\left(\mathbf{0}, I\left(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}\right)\right). \tag{6.3}$$

Applying (6.3) leads to the following distribution on the alternative model discrepancy estimator:

$$\boldsymbol{U}^T(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})I^{-1}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\boldsymbol{U}(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \dot{\sim} \chi^2_{p_A}. \tag{6.4}$$

Applying (6.4) to (6.2) allows us to establish the desired result:

$$\begin{aligned} P_S^* &= Pr^* \left[\boldsymbol{U}^T(\hat{\boldsymbol{\theta}}^* | \boldsymbol{y}) I^{-1}(\hat{\boldsymbol{\theta}}^* | \boldsymbol{y}) \boldsymbol{U}(\hat{\boldsymbol{\theta}}^* | \boldsymbol{y}) > S \right] \\ &\approx Pr \left[\chi^2_{p_A} > S \right] \\ &= p_S. \end{aligned} \tag{6.5}$$

This completes the proof. □

Result (6.5) establishes that, under an adequately specified full null and the assumptions of hypothesis testing, the score test p-value can be alternatively interpreted as a bootstrap-based estimator of the probability that the null model is better than the alternative, as measured by the score discrepancy. This proof linking the score test p-value and the BDCP hinges upon the fact the null model has no nuisance parameters. Rather than use the score discrepancy to connect the BDCP to the score p-value in the partial null setting, we will make the connection using the parameter of interest piecewise score (PIPS) discrepancy.

6.2 Partial Null Setting

In the partial null setting, we again wish to establish that the score p-value is approximated by the BDCP. Using the notation of Section 2.4, the null hypothesis of the form $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ is tested against the alternative $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$. The score test statistic is

$$\begin{aligned} S &= \left[\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)} | \boldsymbol{y}) \right]^T I^{11}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)} | \boldsymbol{y}) \left[\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)} | \boldsymbol{y}) \right] \\ &= \left[\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_0 | \boldsymbol{y}) \right]^T I^{11}(\hat{\boldsymbol{\theta}}_0 | \boldsymbol{y}) \left[\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}_0 | \boldsymbol{y}) \right], \end{aligned} \tag{6.6}$$

yielding the score test p-value of

$$p = Pr[\chi_k^2 > S].$$

We wish to define a discrepancy under which the BDCP approximates the score test p-value in the partial null setting. Define the indicator function $\mathbb{1}_e(\cdot)$ as follows:

$$\mathbb{1}_e(\boldsymbol{\theta}) = \begin{cases} 1, & \text{if any element in } \boldsymbol{\theta} \text{ is to be estimated for the given model.} \\ 0, & \text{if all elements in } \boldsymbol{\theta} \text{ are fixed for the given model.} \end{cases}$$

Then, let the **overall piecewise score discrepancy** be defined as follows:

$$\begin{aligned} d_{PS}\left(g, \tilde{\boldsymbol{\theta}}\right) = & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right]^T \left[\mathscr{I}_{11}(\tilde{\boldsymbol{\theta}})\right]^{-1} \\ & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right] \mathbb{1}_e(\tilde{\boldsymbol{\theta}}_{(1)}) \\ & + \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right]^T \left[\mathscr{I}^{11}(\tilde{\boldsymbol{\theta}})\right] \\ & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}\right] \left(1 - \mathbb{1}_e(\tilde{\boldsymbol{\theta}}_{(1)})\right). \end{aligned}$$

Note that the indicator specifies the appropriate variance/covariance structure for $\boldsymbol{U}_{(1)}(\boldsymbol{\theta}|\cdot)$ depending on whether the parameters of interest $\boldsymbol{\theta}_{(1)}$ are estimated for a given model.

Let the model-based parameter estimators $\tilde{\boldsymbol{\theta}}$ be partitioned into the parameters of interest and nuisance parameters: $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_{(1)}^T, \tilde{\boldsymbol{\theta}}_{(2)}^T)^T$. Similar to Section 5.2, we wish to define a variant of the overall piecewise score discrepancy, which we call the **overall parameter of interest piecewise score** (PIPS) **discrepancy**, that evaluates the nuisance parameters at their pseudo-true values $\bar{\boldsymbol{\theta}}_{(2)}$ and the parameters of interest at their MLEs derived under the restriction that $\boldsymbol{\theta}_{(2)} = \bar{\boldsymbol{\theta}}_{(2)}$. Under the null model, $\boldsymbol{\theta}_{(1)}$ is fixed at $\boldsymbol{\theta}_{0(1)}$, and thus the null model overall PIPS discrepancy is

$$\begin{aligned} d_{PIPS}\left(g, (\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)})\right) = & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{z})\right\}\right]^T \left[\mathscr{I}^{11}(\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)})\right] \\ & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \bar{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{z})\right\}\right]. \end{aligned}$$

The $\boldsymbol{\theta}_{(1)}$ component is estimated under the alternative model, and thus the alternative model overall PIPS discrepancy is

$$\begin{aligned} d_{PIPS}\left(g, (\hat{\boldsymbol{\theta}}_{C(1)}, \bar{\boldsymbol{\theta}}_{(2)})\right) = & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{(1)}, \bar{\boldsymbol{\theta}}_{(2)}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}_{(1)}=\hat{\boldsymbol{\theta}}_{C(1)}}\right]^T \left[\mathscr{I}_{11}(\hat{\boldsymbol{\theta}}_{C(1)}, \bar{\boldsymbol{\theta}}_{(2)})\right]^{-1} \\ & \left[E_g\left\{\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{(1)}, \bar{\boldsymbol{\theta}}_{(2)}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}_{(1)}=\hat{\boldsymbol{\theta}}_{C(1)}}\right]. \end{aligned}$$

As described in Section 5.2, in applying the plug-in principle to estimate the null and alternative PIPS discrepancy, $\hat{\boldsymbol{\theta}}$ is used as the plug-in for $\bar{\boldsymbol{\theta}}$. The null model bootstrap-based PIPS discrepancy is then

$$\begin{aligned} d_{PIPS}\left(\hat{g}, (\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})\right) = & \left[\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y})\right]^T I^{11}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) \\ & \left[\boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y})\right], \end{aligned} \tag{6.7}$$

and the alternative model bootstrap-based estimator is

$$\begin{aligned} d_{PIPS}\left(\hat{g}, (\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right) = & \left[\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right]^T \left[I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right]^{-1} \\ & \left[\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right]. \end{aligned} \tag{6.8}$$

Applying the null and alternative discrepancy estimators in (6.7) and (6.8), respectively, as well as the definition of the score statistic, displayed in (6.6), allows us to write the BDCP under the PIPS discrepancy as follows:

$$\begin{aligned} P^*_{PIPS} &= Pr^* \left[d_{PIPS}\left(\hat{g}, (\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)})\right) < d_{PIPS}\left(\hat{g}, (\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right) \right] \\ &= Pr^* \Big[\boldsymbol{U}^T_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) I^{11}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) \boldsymbol{U}_{(1)}(\boldsymbol{\theta}_{0(1)}, \hat{\boldsymbol{\theta}}_{0(2)}|\boldsymbol{y}) \\ &\qquad < \boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \left[I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \right]^{-1} \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \Big] \\ &= Pr^* \left[\boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \left[I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \right]^{-1} \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) > S \right]. \end{aligned} \tag{6.9}$$

Proposition 6.2.1. *Assuming that the large-sample properties of the MLEs hold and that the alternative model is adequately specified, then for testing a partial null hypothesis of* $H_0 : \boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{0(1)}$ *versus the alternative of* $H_A : \boldsymbol{\theta}_{(1)} \neq \boldsymbol{\theta}_{0(1)}$,

$$p_S \approx P^*_{PIPS}.$$

Proof. To complete this proof, we need to show that under Pr^*, the term on the left-hand side of the inequality in (6.9) follows an approximate χ^2_k distribution:

$$\boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \left[I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \right]^{-1} \boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \dot{\sim} \chi^2_k.$$

To establish this result, consider taking a first-order Taylor series expansion of $\boldsymbol{U}_{(1)}\left(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}\right)$ around $\left(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}\right)$. Write

$$\begin{aligned} \boldsymbol{U}_{(1)}\left(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}\right) &\approx \boldsymbol{U}_{(1)}\left(\hat{\boldsymbol{\theta}}|\boldsymbol{y}\right) - \left(I_{11}(\hat{\boldsymbol{\theta}}|\boldsymbol{y}), I_{12}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right) \begin{pmatrix} \hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)} \\ \hat{\boldsymbol{\theta}}_{(2)} - \hat{\boldsymbol{\theta}}_{(2)} \end{pmatrix} \\ &= -I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}). \end{aligned} \tag{6.10}$$

Using a demonstration similar to that which produced (5.19), we argue that

$$(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \dot{\sim} N_k\left(\mathbf{0}, \mathscr{I}^{-1}_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)})\right).$$

For large n, under certain regularity conditions, $\mathscr{I}_{11}^{-1}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}) \approx I_{11}^{-1}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})$. Thus, we write

$$(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \dot{\sim} N_k\left(\mathbf{0}, I_{11}^{-1}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right). \tag{6.11}$$

Combining (6.10) and (6.11), we see that

$$\begin{aligned}\boldsymbol{U}_{(1)}\left(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}\right) &\approx -I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^*_{C(1)} - \hat{\boldsymbol{\theta}}_{(1)}) \\ &\dot{\sim} -I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})N_k\left(\mathbf{0}, I_{11}^{-1}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right) \\ &\sim N_k\left(\mathbf{0}, I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right). \end{aligned} \tag{6.12}$$

Because the distribution of $\hat{\boldsymbol{\theta}}^*_{C(1)}$ is centered around $\hat{\boldsymbol{\theta}}_{(1)}$, it follows that $I_{11}(\hat{\boldsymbol{\theta}}_{(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) \approx I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})$. This approximation in conjunction with (6.12) establishes that

$$\boldsymbol{U}_{(1)}\left(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}\right) \dot{\sim} N_k\left(\mathbf{0}, I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right) \tag{6.13}$$

Result (6.13) implies that

$$\boldsymbol{U}^T_{(1)}\left(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}\right)\left[I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right]^{-1}\boldsymbol{U}_{(1)}\left(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}\right) \dot{\sim} \chi^2_k. \tag{6.14}$$

Finally, applying (6.14) to the inequality involving P^*_{PIPS} displayed in (6.9) allows us to assert

$$\begin{aligned}P^*_{PIPS} &= Pr^*\left[\boldsymbol{U}^T_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\left[I_{11}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y})\right]^{-1}\boldsymbol{U}_{(1)}(\hat{\boldsymbol{\theta}}^*_{C(1)}, \hat{\boldsymbol{\theta}}_{(2)}|\boldsymbol{y}) > S\right] \\ &\approx Pr[\chi^2_k > S] \\ &= p_S.\end{aligned}$$

This completes the proof. □

We have thus established that the score p-value can be approximated by a BDCP under an appropriately chosen discrepancy, regardless of whether the null hypothesis pre-specifies all parameter values.

CHAPTER 7
SIMULATION STUDY

7.1 Design

To further support the mathematical results showing the connection between the Wald, score and LR test p-values, we have also performed a simulation study. The primary goal of this simulation study is to compare how well each p-value is approximated by its respective BDCP for which the approximation has been mathematically shown in previous chapters. The simulation study employs a factorial design composed of three factors. First, we consider both a linear and a logistic regression modeling framework. Within both frameworks, we consider full and partial null hypotheses. Finally, within each combination of modeling framework and type of null, we examine a setting in which the null is adequately specified and another setting in which the alternative is true and null is underspecified. Based on compiled results not shown, the sample size n affects the quality of the approximations more in the true alternative, false null setting than for an adequately specified null. Therefore, in the underspecified null settings, we present results for 3 sample sizes, namely $n = 100$, $n = 250$ and $n = 1,000$, and when the null is adequately specified, we present results for one sample size, $n = 500$.

In both the linear and logistic regression modeling frameworks, we draw independent samples of size n of an outcome variable y, as well as corresponding covariates x_1, x_2 and x_3. In the linear regression setting, for $i = 1, \ldots, n$, the generating model is of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

where $\epsilon_i \backsim N(0, \sigma^2)$. Similarly, in the logistic regression setting, for $i = 1, \ldots, n$, the observed data is generated as $y_i \backsim bin(1, \pi_i)$, where

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

In both settings the distribution of the covariates is

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \backsim N_3\left(\begin{pmatrix} 2 \\ 2 \\ -2 \end{pmatrix}, \begin{pmatrix} 100 & 64 & 64 \\ 64 & 100 & 64 \\ 64 & 64 & 100 \end{pmatrix}\right).$$

In the linear regression setting, the alternative model always corresponds to the model in which the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ is unrestricted. In the linear regression setting, the alternative model estimator of σ^2 is unrestricted on $[0, \infty)$. In each linear regression simulation, the true variance is $\sigma^2 = 50$. In the full null setting, the null model sets $\boldsymbol{\beta} = \mathbf{0}$. While not typically done in practice, in the full null setting, the null model must provide a pre-specified value of σ^2, denoted by σ_0^2. The pre-specified σ_0^2 varies across simulation sets. Thus, we can write the hypotheses for the full null linear regression setting as $H_0 : \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \sigma_0^2 \end{pmatrix}$ versus $H_A : \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} \neq \begin{pmatrix} \mathbf{0} \\ \sigma_0^2 \end{pmatrix}$. The partial null in the linear regression framework tests $H_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ versus the general alternative. Simulations sets differ according to the values of the true parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$.

Like the linear regression simulations, in the logistic regression setting, the alternative model parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ is unrestricted. In the logistic regression setting, the full null tests $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus the general alternative. Also similar to the linear regression setting, the partial null in the logistic regression setting corresponds to a test of $H_0 : \binom{\beta_2}{\beta_3} = \binom{0}{0}$ versus the general alternative. Simulation sets vary according to the values of the true parameter vector $\boldsymbol{\beta}$.

Each set of simulation results is based on drawing 50 original samples. From each original sample, we calculate the full and partial null Wald, LR and score p-values, and based on $B = 10,000$ bootstrap samples, we approximate the BDCP under each of the constituent discrepancies.

For each set of simulation results, we present a scatterplot with the corresponding p-value on the x-axis and the estimated BDCP on the y-axis. A 45 degree line running through the origin is placed on each graph to aid in determining how close the estimated BDCP is to its corresponding p-value. Each graph will contain 50 ordered pairs, one for each of the original samples. For each of the simulation results, we also present an estimate of the concordance correlation coefficient (CCC), labeled $\hat{\rho}_c$, which is a numerical measure of how close a set of ordered pairs falls to the line $y = x$ (Lawrence and Lin, 1989). The CCC is a measure that lies between -1 and 1, inclusive, with $\rho_c = 1$ indicating exact agreement.

All simulations, calculations and scatterplots were performed and created using R (R Core Team, 2017) and R studio (RStudio Team, 2017).

7.2 Linear Regression

Figure 7.1 presents the scatterplots comparing the estimated BDCPs to their respective p-values in the adequately specified null setting. The first 6 entries of

Table 7.1 present the CCC values comparing the estimated BDCPs to their respective p-values.

In the underspecified null setting, we wish to avoid a scenario in which most p-values are very close to zero. To achieve this goal, as we increase the sample size, the absolute value of the elements of the true parameter vector $\boldsymbol{\beta}$ must get smaller. With this in mind, for each of the three sample sizes, we vary $\boldsymbol{\beta}$ as well as the prespecified null variance estimator σ_0^2. For each sample size, samples are drawn from the generating distribution in which we set $\beta_0 = -\beta_1 = \beta_2 = -\beta_3$, and $\sigma^2 = 50$. For $n = 100$, we set $\beta_0 = 0.175$ and set the null model variance "estimator" to $\sigma_0^2 = 60$; for $n = 250$, set $\beta_0 = 0.10$ and $\sigma_0^2 = 57$; and for $n = 1,000$, set $\beta_0 = 0.05$ and $\sigma_0^2 = 55$.

Figure 7.2 presents scatterplots for the $n = 100$ setting; Figure 7.3 presents the results for $n = 250$; and the $n = 1,000$ setting is presented in Figure 7.4. The final three sections of Table 7.1 present the CCC values in each of these three settings, respectively.

7.3 Logistic Regression

The logistic regression results are presented in a fashion similar to the linear regression setting. In the adequately specified null setting $\boldsymbol{\beta} = \mathbf{0}$. In the underspecified null setting, we again set $\beta_0 = -\beta_1 = \beta_2 = -\beta_3$. When $n = 100$, we set $\beta_0 = 0.07$; for $n = 250$, set $\beta_0 = 0.04$; and for $n = 1,000$, $\beta_0 = 0.01$. The scatterplots for the adequately specified null, where we set $n = 500$, are presented in Figure 7.5. Figure 7.6 presents the scatterplots for the underspecified null with $n = 100$; results for when

$n = 250$ are given in Figure 7.7; and Figure 7.8 presents results for $n = 1,000$. Table 7.2 gives the CCC values for each of these combinations.

7.4 Interpretation of Results

For both the linear and logistic regression modeling frameworks, Figures 7.1 and 7.5 support that when the null model is adequately specified, the p-values are closely approximated by their respective BDCPs. This finding is especially strong in the logistic regression setting, where most points on the scatterplot fall very close to the 45 degree line.

In the $n = 100$ underspecified null setting, as illustrated by Figures 7.2 and 7.6, the BDCPs exhibit a considerable amount of positive bias for their respective p-values in both the linear and logistic regression settings. For the LR p-value, the bias is more pronounced in the full null setting. In the linear regression framework, the Wald- and score-based BDCPs exhibit less positive bias in the partial null setting than in the full null. For the partial null setting, the Wald- and score-based BDCPs also exhibit less bias in the linear regression setting than in the logistic.

In this underspecified null setting, when we increase the sample size from $n = 100$ to $n = 250$, as displayed in Figures 7.3 and 7.7, a great deal of the positive bias is attenuated for both the linear and logistic regression frameworks. Specifically, for linear regression, the BDCPs under the Wald, PIKL, score and PIPS discrepancies display very little bias, while the BDCP under the KL discrepancy still exhibits noticeable bias (albeit far less bias than when $n = 100$). For the logistic regression setting, increasing the sample size from $n = 100$ to $n = 250$ drastically reduced the positive bias, but most graphs still indicate some bias, with the BDCP under the PIKL discrepancy being the only setting where the bias is not particularly noticeable.

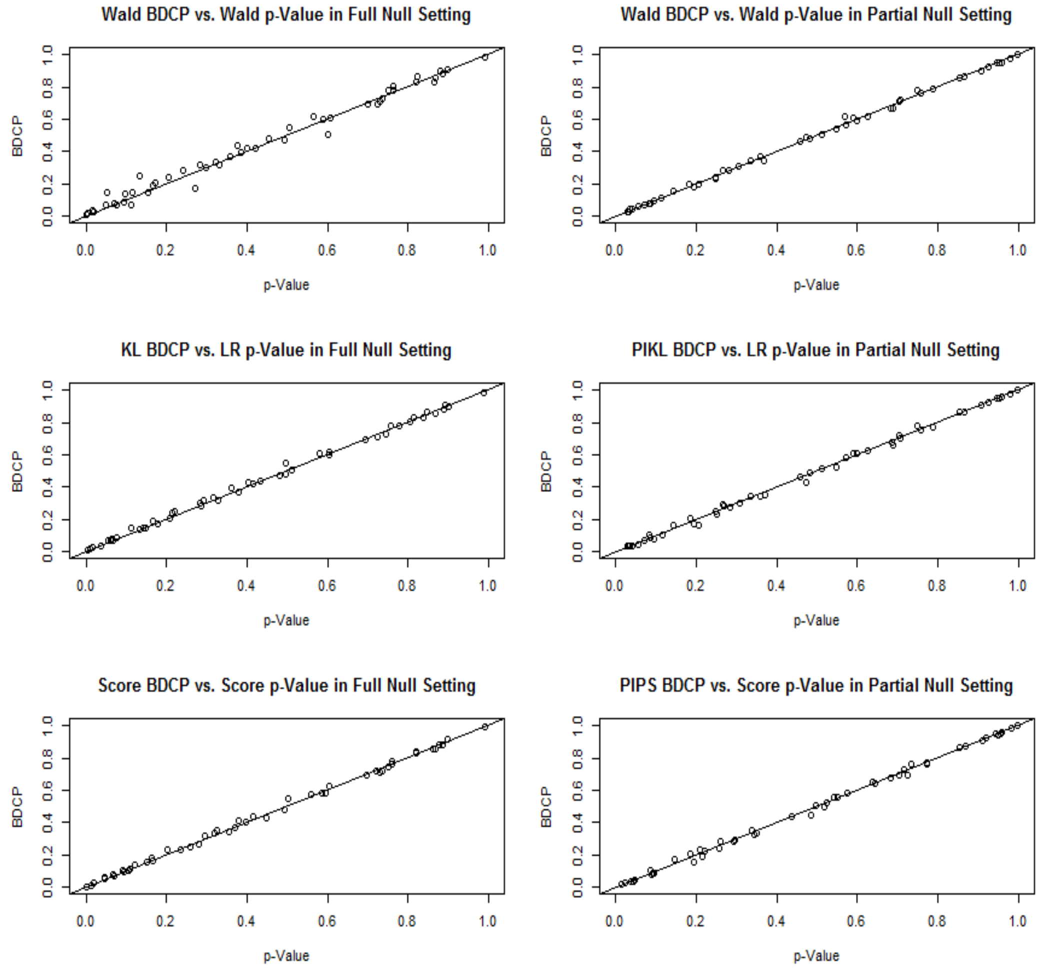
Finally, when we increase the sample size from $n = 250$ to $n = 1,000$, almost all bias disappears. Specifically, in the linear regression setting, Figure 7.4 shows that nearly all ordered pairs fall very close to the 45 degree line, indicating that almost all positive bias is removed. According to Table 7.1, in the linear regression setting with $n = 1,000$, the BDCP under the KL discrepancy is the worst approximation, but is still quite good with $\hat{\rho}_c = 0.99590$. Similarly, Figure 7.8 indicates that nearly all positive bias is removed in the logistic regression framework. Nevertheless, in this setting, there is still a very small, but noticeable, positive bias in the full null settings, with any bias in the partial settings being far less noticeable.

We find that the simulation results strongly support the mathematical findings which connect the BDCPs to their respective p-values. For adequately specified null hypotheses, the approximations hold quite well for moderate sample sizes. For underspecified null hypotheses, the approximations improve as the sample size increases. This, however, is to be expected because the mathematical proofs connecting the BDCPs to their p-values rely on large sample theory.

7.5 Tables and Figures

Table 7.1: $\hat{\rho}_c$ comparing $\hat{P}^*$ to its respective p-value in the linear regression setting.

Null Specification	n	β_0	σ_0^2	Null Type	Test	Disc.	$\hat{\rho}_c$
adequate	500	0	50	full	Wald	Wald	0.99897
adequate	500	0		partial	Wald	Wald	0.99923
adequate	500	0	50	full	LR	KL	0.99878
adequate	500	0		partial	LR	PIKL	0.99884
adequate	500	0	50	full	score	score	0.99892
adequate	500	0		partial	score	PIPS	0.99886
under	100	0.175	60	full	Wald	Wald	0.97849
under	100	0.175		partial	Wald	Wald	0.98870
under	100	0.175	60	full	LR	KL	0.95098
under	100	0.175		partial	LR	PIKL	0.99235
under	100	0.175	60	full	score	score	0.95999
under	100	0.175		partial	score	PIPS	0.99113
under	250	0.10	57	full	Wald	Wald	0.99509
under	250	0.10		partial	Wald	Wald	0.99324
under	250	0.10	57	full	LR	KL	0.98953
under	250	0.10		partial	LR	PIKL	0.99509
under	250	0.10	57	full	score	score	0.99168
under	250	0.10		partial	score	PIPS	0.99468
under	1000	0.05	55	full	Wald	Wald	0.99662
under	1000	0.05		partial	Wald	Wald	0.99863
under	1000	0.05	55	full	LR	KL	0.99590
under	1000	0.05		partial	LR	PIKL	0.99910
under	1000	0.05	55	full	score	score	0.99600
under	1000	0.05		partial	score	PIPS	0.99916



Wald BDCP vs. Wald p-Value in Full Null Setting

Wald BDCP vs. Wald p-Value in Partial Null Setting

KL BDCP vs. LR p-Value in Full Null Setting

PIKL BDCP vs. LR p-Value in Partial Null Setting

Score BDCP vs. Score p-Value in Full Null Setting

PIPS BDCP vs. Score p-Value in Partial Null Setting

Figure 7.1: Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an adequately specified null. Here, $\beta_0 = 0$, $\sigma^2 = 50$ and $n = 500$.

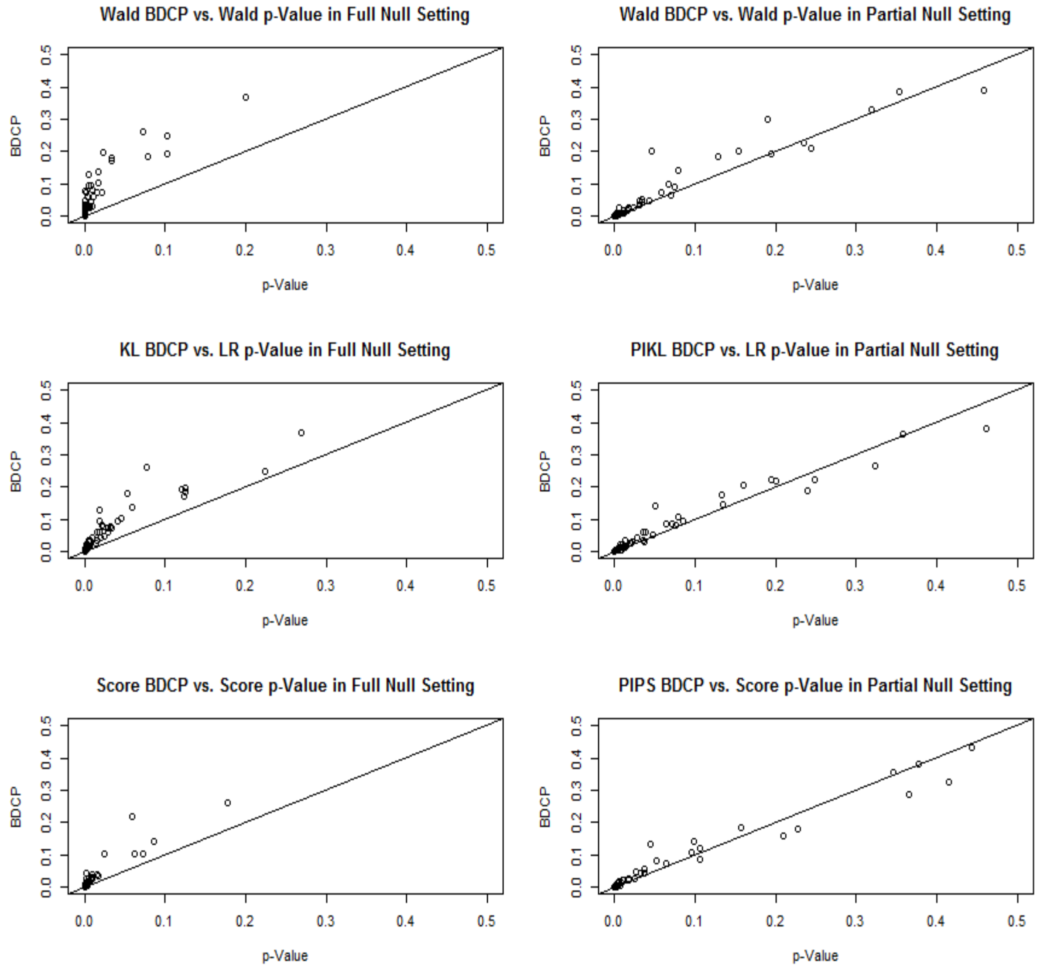


Figure 7.2: Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.175$, $\sigma^2 = 60$ and $n = 100$.

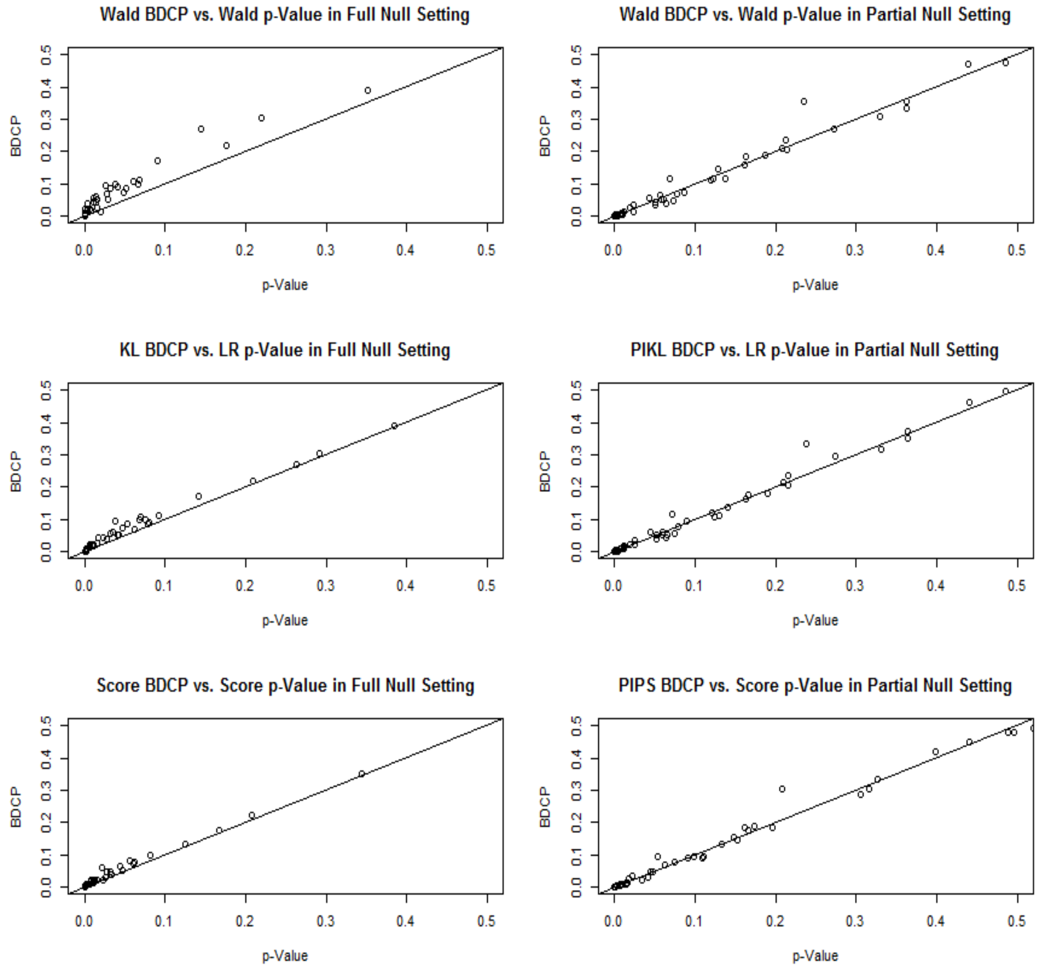


Figure 7.3: Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.10$, $\sigma^2 = 57$ and $n = 250$.

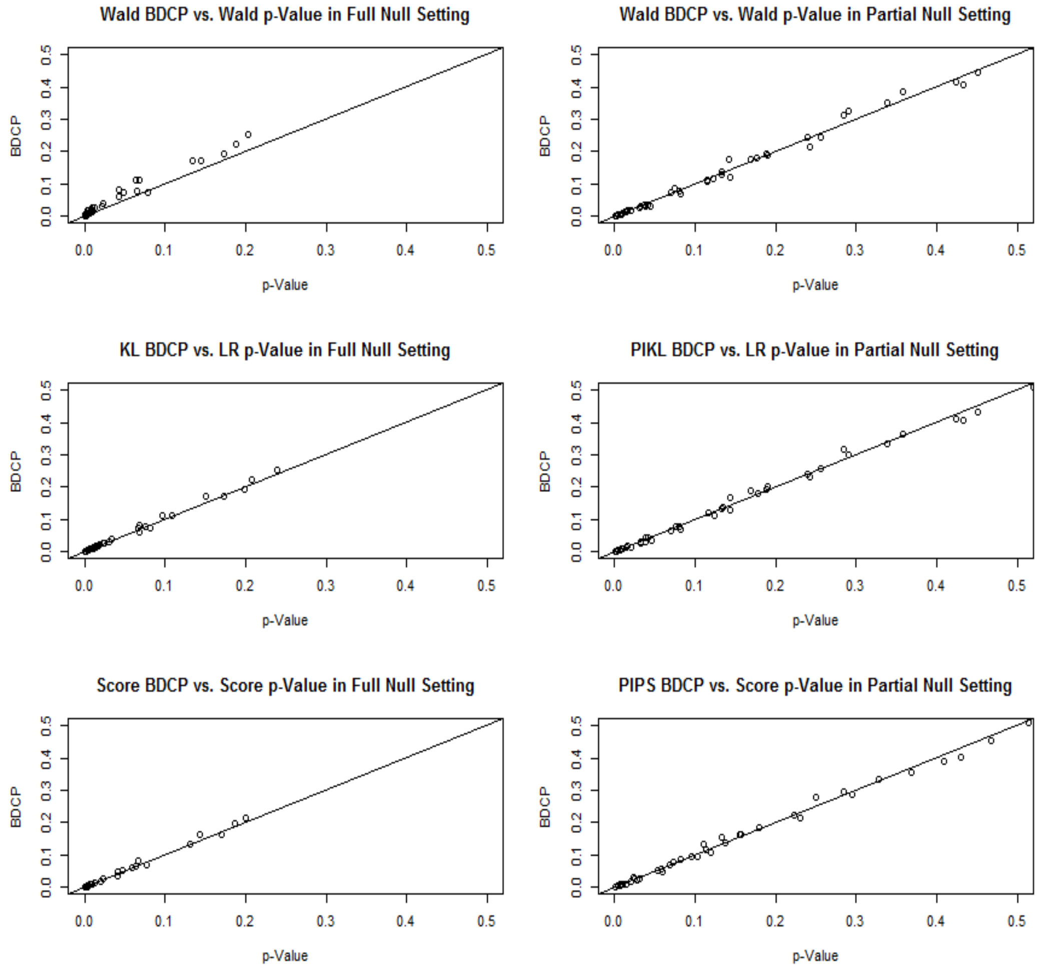
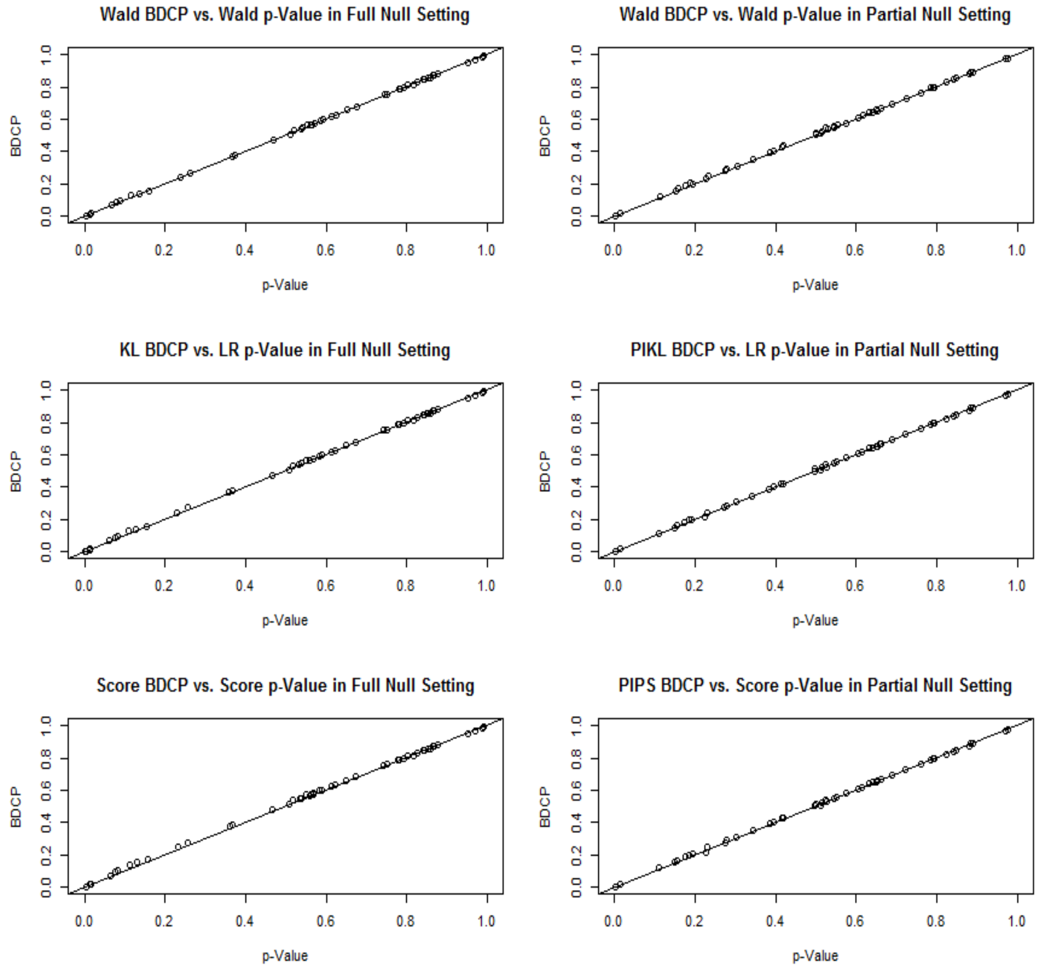


Figure 7.4: Scatterplots of estimated BDCPs vs. their respective p-values in the linear regression setting with an underspecified null. Here, $\beta_0 = 0.05$, $\sigma^2 = 55$ and $n = 1000$.

Table 7.2: $\hat{\rho}_c$ comparing $\hat{P}^*$ to its respective p-value in the logistic regression setting.

Null Specification	n	β_0	Null Type	Test	Disc.	$\hat{\rho}_c$
adequate	500	0	full	Wald	Wald	0.99982
adequate	500	0	partial	Wald	Wald	0.99981
adequate	500	0	full	LR	KL	0.99986
adequate	500	0	partial	LR	PIKL	0.99980
adequate	500	0	full	score	score	0.99974
adequate	500	0	partial	score	PIPS	0.99976
under	100	0.07	full	Wald	Wald	0.85941
under	100	0.07	partial	Wald	Wald	0.98605
under	100	0.07	full	LR	KL	0.83085
under	100	0.07	partial	LR	PIKL	0.99352
under	100	0.07	full	score	score	0.73545
under	100	0.07	partial	score	PIPS	0.99013
under	250	0.04	full	Wald	Wald	0.97823
under	250	0.04	partial	Wald	Wald	0.99879
under	250	0.04	full	LR	KL	0.98201
under	250	0.04	partial	LR	PIKL	0.99926
under	250	0.04	full	score	score	0.95987
under	250	0.04	partial	score	PIPS	0.99889
under	1000	0.01	full	Wald	Wald	0.99989
under	1000	0.01	partial	Wald	Wald	0.99986
under	1000	0.01	full	LR	KL	0.99990
under	1000	0.01	partial	LR	PIKL	0.99993
under	1000	0.01	full	score	score	0.99985
under	1000	0.01	partial	score	PIPS	0.99992


Wald BDCP vs. Wald p-Value in Full Null Setting

Wald BDCP vs. Wald p-Value in Partial Null Setting

KL BDCP vs. LR p-Value in Full Null Setting

PIKL BDCP vs. LR p-Value in Partial Null Setting

Score BDCP vs. Score p-Value in Full Null Setting

PIPS BDCP vs. Score p-Value in Partial Null Setting

Figure 7.5: Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an adequately specified null. Here, $\beta_0 = 0$ and $n = 500$.

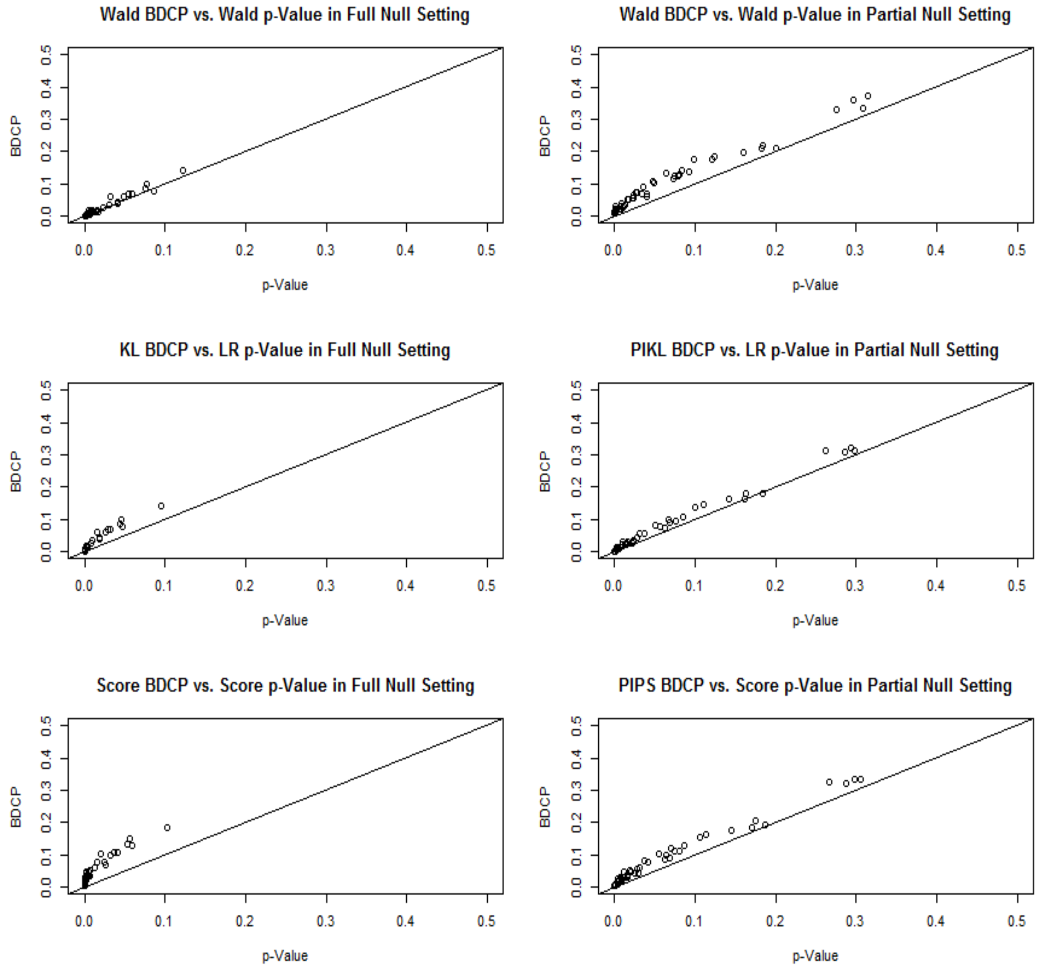


Figure 7.6: Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.07$ and $n = 100$.

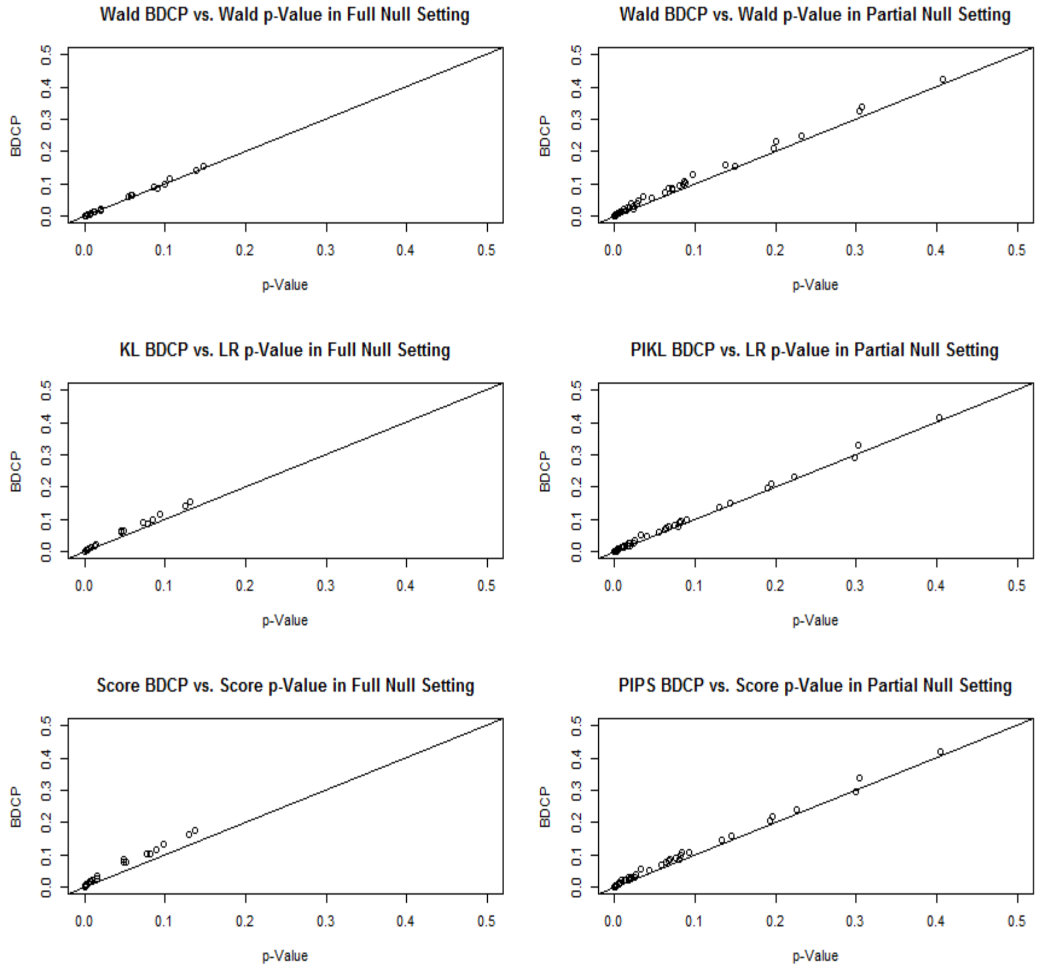


Figure 7.7: Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.04$ and $n = 250$.

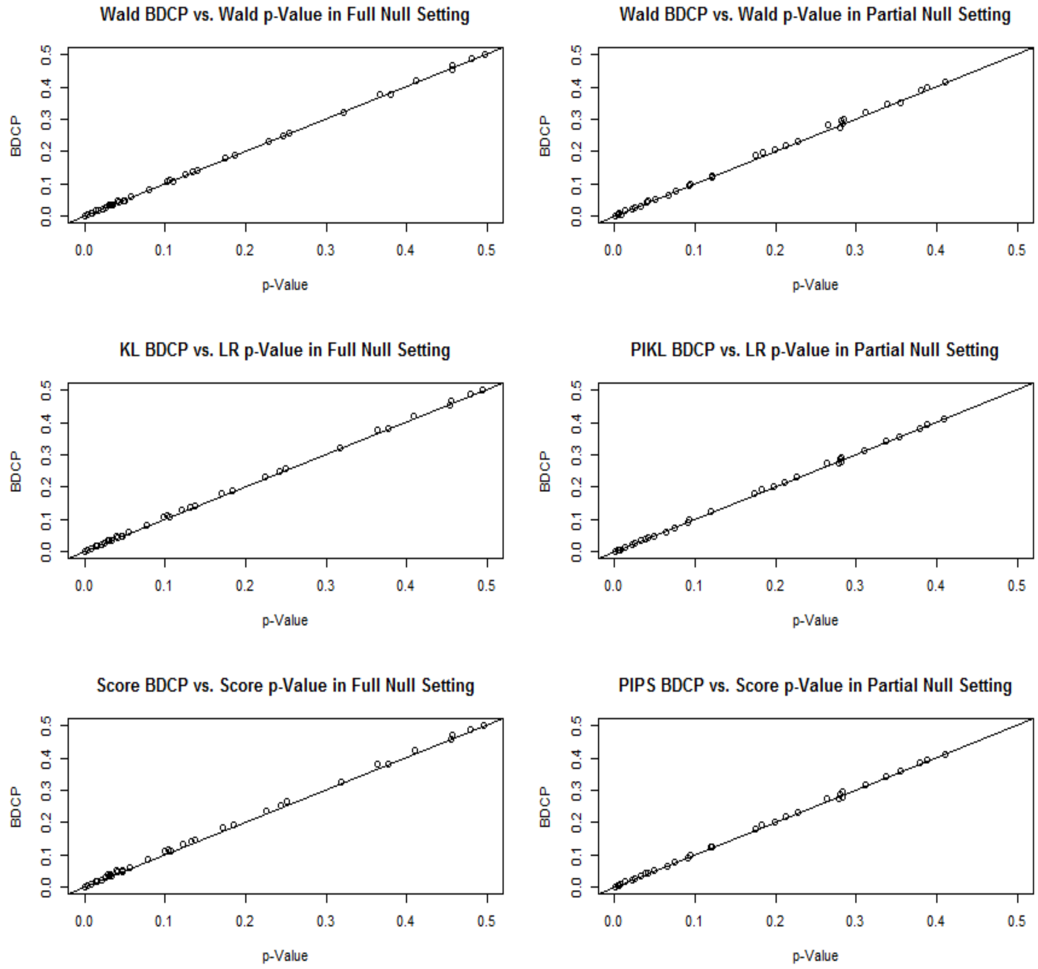


Figure 7.8: Scatterplots of estimated BDCPs vs. their respective p-values in the logistic regression setting with an underspecified null. Here, $\beta_0 = 0.01$ and $n = 1000$.

CHAPTER 8
BDCP / P-VALUE CONNECTION: FURTHER INVESTIGATION

We have shown mathematically that, using suitably defined discrepancy measures, the BDCP approximates the Wald, score and LR test p-values. Because the hypothesis testing and discrepancy function approaches to model selection are notably different, this connection is perhaps surprising, and we thus further investigate this approximation.

8.1 Undesirable Model Selection Properties of BDCP / p-Value

In settings in which the null model is badly underspecified compared to the alternative, both the p-value and the BDCP under suitably chosen discrepancies should be very small, indicating strong support for the alternative. On the other hand, the p-value and the BDCP will struggle to delineate between two adequately specified models. To see why, consider a setting in which the null model is correct, and the alternative model is overspecified. In this case, it is desirable for a model selection quantity to indicate support for the null. For instance, AIC will often register a preference for the smaller model when it is properly specified. On the other hand, this is not how the p-value behaves. Under a true null, the p-value follows an approximate $Uniform(0,1)$ distribution. For the BDCP to approximate the p-value, it too must approximately follow this distribution. Thus, under a true null, this BDCP will not typically register strong support for the null model, with about half of all such cases yielding a BDCP less than 0.50.

Ostensibly, a BDCP less than 0.50 should indicate at least modest support for the alternative model, because if $P^* < 0.50$, then across repeated bootstrap samples, the fitted alternative model has a greater than 50% chance of being more congruous with $\hat{g}$ than the fitted null. Based on this interpretation of an observation of $P^* < 0.50$, when the null is precisely true, there is about a 50% chance that the BDCP will indicate at least modest support for the alternative.

8.2 Adjusted R^2 / Mean Squared Error as Model Selection Criteria

That P^* under certain discrepancies will not tend to strongly prefer a properly specified null model may be considered an undesirable attribute. This behavior is, nevertheless, very similar to the behavior of certain model selection criteria, including adjusted R^2 and the mean squared error (MSE). For instance, suppose we must choose between the following two nested linear regression models:

$$M_0 : E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1},$$

and

$$M_A : E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta_k x_k.$$

Assuming the R^2 value of the larger model is less than one, then the adjusted R^2 of M_0 is less than that of M_A if and only if $|t_k| < 1$, where $t_k = \hat{\beta}_k \;/\; \widehat{SE}(\hat{\beta}_k)$ is the Student t statistic for testing $\beta_k = 0$ (Dufour, 2011). What then is the probability of the alternative model being favored by adjusted R^2 even if M_0 is properly specified? Assuming the null is true, t_k follows a t distribution. Depending on the degrees of freedom, $Pr\left\{|t_k| < 1\right\}$ is bounded between approximately 0.50 and 0.68. Thus, even in the case which is mostly likely to favor the null (i.e. when we assume t_k is distributed normally), the probability that the alternative model adjusted R^2 will be greater than that of the null model is approximately $1 - 0.68 = 0.32$. In other words, when the additional parameter β_k is precisely equal to 0, there is at least a 32% chance that adjusted R^2 will favor the larger model. This behavior is quite similar to that of the p-value and the BDCP under suitably defined discrepancies.

For another example of a model selection tool struggling to delineate between adequately specified models, suppose we use MSE to choose between two nested linear models. Note that $MSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \;/\; (n-(k+1))$, where $\hat{y}_i$ is the fitted value of the i^{th} observation, and $k+1$ is the rank of the model's design matrix. If we impose the assumption that the smaller model is adequately specified, then the MSE for both the smaller and larger model is an unbiased estimator of the error variance σ^2. Because both models are assumed to contain all necessary parameters, then σ^2 is the same for both models. Therefore, choosing between the two models using MSE will not tend to strongly favor the null model, even when it is properly specified. This behavior is also qualitatively similar to that of the p-value and the BDCP under suitable discrepancies.

8.3 AIC / DCP Comparison

When two competing models contain all true predictors, then many traditional model selection criteria endeavor to choose the more parsimonious model. For instance, consider AIC, which is defined as

$$AIC = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2k,$$

where k is the dimension of the model's parameter vector. AIC is designed to punish models which have unnecessary complexity. Specifically, $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ is the goodness-of-fit term, which will always improve with additional model complexity. However, the $2k$ penalty term, which of course increases with model complexity, can often prevent models with extraneous parameters from being favored simply because their goodness-of-fit term is better than a more parsimonious model.

While we argue that the DCP and AIC are both useful model evaluation tools, they will not necessarily produce similar model evaluations. To understand why, consider a setting where the true model is contained in the class of models being considered and that the vector of MLEs $\hat{\boldsymbol{\theta}}$ satisfies conventional large-sample properties. AIC then serves as an asymptotically unbiased estimator of the *expected* overall KL discrepancy, $E_g\left\{d_{KL}(g, \hat{\boldsymbol{\theta}})\right\}$, where

$$E_g\left\{d_{KL}(g, \hat{\boldsymbol{\theta}})\right\} = E_g\left\{E_g\left\{-2\ell(\boldsymbol{\theta}|\boldsymbol{z})\right\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\}.$$

In other words, AIC is estimating the average of the overall KL discrepancy, where the averaging is done over the distribution of $\hat{\boldsymbol{\theta}}$. Therefore, if one uses AIC to delineate between two (or more) models, then one is implicitly concerned with the average of each model's overall KL discrepancy. On the other hand, if one uses the DCP under the KL discrepancy to decide between two models, then focus is instead placed on the joint distribution of the overall discrepancies for the two models. We thus should not necessarily assume AIC and the DCP will yield similar model evaluations. For instance, suppose the joint distribution of the overall KL discrepancies of two competing models is such that for 40% of samples, the overall discrepancy for Model 1 is *much* smaller than that of Model 2, and in 60% of samples, Model 2 is *slightly* better. In such an instance, AIC may favor Model 1 because it will have a smaller expected overall KL discrepancy, but the DCP will favor Model 2 because it has a smaller overall KL discrepancy in a majority (60%) of samples.

8.4 Bias of Discrepancy Estimators Under KL Discrepancy

Despite AIC and the DCP under the KL discrepancy trying to characterize different aspects of the overall KL discrepancy, one may still desire for the bootstrap-based estimator of the overall KL discrepancy to be asymptotically unbiased for the expected overall KL discrepancy, as AIC is. However, over repeated samples from the true distribution and repeated bootstrap samples drawn from each of these samples, the bootstrap-based estimator of the overall KL discrepancy $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*) = -2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})$ is negatively biased for the expected overall KL discrepancy $E_g\left\{d_{KL}(g, \hat{\boldsymbol{\theta}})\right\}$. Formally, this result can be characterized as follows.

Proposition 8.4.1. *For large* n, *if the assumptions under which AIC is derived are met, then*

$$E_g\left\{E_*\left\{-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\}\right\} \approx E_g\left\{d_{KL}(g, \hat{\boldsymbol{\theta}})\right\} - k,$$

where $E_*(\cdot)$ *denotes expectation taken with respect to the bootstrap distribution of* $\hat{\boldsymbol{\theta}}^*$, *and* k *is the dimension of the parameter vector.*

Proof. We begin by stating a well-known result from maximum likelihood estimation that will be used later in the proof. Recall that for large n, under certain regularity conditions and an adequately specified model,

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T I(\boldsymbol{\theta}|\boldsymbol{y})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \mathrel{\dot\sim} \chi^2_k. \tag{8.1}$$

Then, consider taking a second-order expansion of $-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})$ about $\hat{\boldsymbol{\theta}}$ to obtain

$$-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y}) \approx -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + (\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T I(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}).$$

We then take the expectation of both sides with respect to the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*$, which yields

$$E_* \left\{-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\} \approx -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + E_* \left\{(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T I(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})\right\}. \tag{8.2}$$

Assuming the data at hand adequately characterizes the sampling distribution of $\hat{\boldsymbol{\theta}}$ via the bootstrap distribution of $\hat{\boldsymbol{\theta}}^*$, then applying (8.1) to the boostrapping context yields

$$(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T I(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) \mathrel{\dot\sim} \chi^2_k.$$

Therefore,

$$E_* \left\{(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})^T I(\hat{\boldsymbol{\theta}}|\boldsymbol{y})(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})\right\} \approx k. \tag{8.3}$$

Combining expressions (8.2) and (8.3) yields

$$\begin{aligned} E_* \left\{-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\} &\approx -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + k \\ &= AIC - k. \end{aligned}$$

Recall that if the true model is contained in the class of models being considered and the large-sample properties of MLEs hold, then AIC is an asymptotically unbiased estimator of the overall KL discrepancy. Therefore,

$$\begin{aligned} E_g\left\{E_*\left\{-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\}\right\} &\approx E_g\left\{AIC - k\right\} \\ &\approx E_g\left\{d_{KL}(g,\hat{\boldsymbol{\theta}})\right\} - k. \end{aligned}$$

This completes the proof. □

8.5 Biased Estimators and Connection to p-Value

Another way to conceptualize Proposition 8.4.1 is that

$$E_g\left\{E_*\left\{-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})\right\}\right\} \approx E_g\left\{-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\right\} + k.$$

Whereas AIC includes the $2k$ penalty term for model complexity, the bootstrap-based estimator of the overall KL discrepancy essentially contains a "$1k$ penalty term." This $1k$ penalty term provides an explanation for why P^*_{KL} is unable to strongly delineate between the null and alternative model when the null is true. To see why, consider that, under appropriate regularity conditions and a true null,

$$-2\ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y}) - (-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})) \,\dot{\sim}\, \chi^2_m,$$

where $m = dim(\boldsymbol{\theta}) - dim(\boldsymbol{\theta}_0)$. Therefore,

$$E_g\left\{-2\ell(\hat{\boldsymbol{\theta}}_0|\boldsymbol{y}) - (-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}))\right\} \approx m.$$

Thus, for each extraneous parameter added to the alternative model, the alternative model goodness-of-fit term $-2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ is expected to improve by one unit. Consequently, the $1k$ penalty term of $-2\ell(\hat{\boldsymbol{\theta}}^*|\boldsymbol{y})$ essentially "cancels" the improvement in the goodness-of-fit term. A penalty term smaller than $1k$ implies that adding unnecessary complexity would generally be favored, whereas a penalty term greater than $1k$ would not typically favor unnecessary complexity. The $1k$ penalty term of the bootstrap-based estimator of the KL discrepancy is such that unnecessary complexity is neither systematically favored nor disfavored. We see this characteristic in how P^*_{KL} does not typically strongly favor a true null, regardless of how many extraneous parameters the alternative model contains. Note that because P^*_{KL} is based on pairwise comparisons of the joint distribution of the overall discrepancy estimators rather than comparisons of the *expected value* of the overall discrepancy estimators, then this result only offers some intuition behind why $P^*_{KL} \dot{\sim} Uniform(0,1)$ under a true null.

8.6 Remedial Approaches for Bias

If one desired to attenuate the effect of the bias of the bootstrap-based estimator of the KL discrepancy $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ on the bootstrap-based DCP P^*_{KL}, then what could be done? We should first note that attempts to correct for the bias of $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ will break the connection between the BDCP and the p-value. If the bias is corrected, then the BDCP should tend to favor the null model when it is properly specified, which is unlike how the p-value behaves. Nevertheless, suppose a practitioner did not care about the connection to the p-value. Perhaps then the most obvious way of correcting for the bias is by shifting the joint distribution of $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ and $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*_0)$ by the bias of the discrepancy estimators. Recall that the bias is approximately the dimension of the parameter vector, which we denote p_A and p_0 for the alternative and null models, respectively. Such an adjustment would ensure that the means of the null and alternative bootstrap-based estimators $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ and $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}_0^*)$ are approximately equal to the means of $d_{KL}(g, \hat{\boldsymbol{\theta}})$ and $d_{KL}(g, \hat{\boldsymbol{\theta}}_0)$. Then, based on B bootstrap samples, the approximation of this bias-corrected BDCP would be

$$\hat{P}_{KL}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1} \left\{ d_{KL}\left(\hat{g}, \hat{\boldsymbol{\theta}}_0^*(b)\right) + p_0 < d_{KL}\left(\hat{g}, \hat{\boldsymbol{\theta}}^*(b)\right) + p_A \right\}. \tag{8.4}$$

Note that shifting the joint distribution of $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}^*)$ and $d_{KL}(\hat{g}, \hat{\boldsymbol{\theta}}_0^*)$ only ensures that the mean of the joint distribution aligns with that of $d_{KL}(g, \hat{\boldsymbol{\theta}})$ and $d_{KL}(g, \hat{\boldsymbol{\theta}}_0)$, but does not necessarily imply the actual distributions will be similar. Moreover, this shift does not imply that the bias-corrected P_{KL}^* is unbiased for P_{KL}.

The bias-corrected BDCP in (8.4) utilizes Proposition 8.4.1 to establish the bias of the null and alternative bootstrap-based discrepancy estimators. However, the result of Proposition 8.4.1 is only valid under adequately specified models, and thus the bias corrections employed in (8.4) are only accurate when the null and alternative models are adequately specified. While this fact may seem to diminish the impact of the bias-corrected BDCP, note that the bias-corrected BDCP is still a useful model evaluation tool when the alternative model is adequately specified and the null is underspecified. To see why, consider that in such a setting, the BDCP (regardless of whether the bias correction is used) should be small, indicating support for the alternative. In other words, when the null is badly underspecified compared to the alternative, the alternative is favored by the BDCP to such an extent that the veracity of the bias corrections is immaterial from a practical standpoint.

Recall that when the null is true, the p-value will be less than 0.50 in approximately 50% of samples. Thus, using a cutoff of $p < 0.50$ to choose in favor of the alternative model would not typically be a wise threshold. Similarly, because the uncorrected BDCP under a suitably defined discrepancy approximates the p-value, then using a threshold of $\hat{P}^* < 0.50$ to choose in favor of the alternative model will also typically be errant. Instead, a smaller cutoff may be a more appropriate threshold for choosing in favor of the alternative model. On the other hand, the bias-corrected BDCP should provide a more balanced comparison of the null and alternative models. Thus, in order to favor the alternative model, the bias-corrected BDCP should not be required to reach the same threshold as the p-value or uncorrected BDCP. Let P_C^* denote the bias-corrected BDCP. Then, depending on the context, if P_C^* indeed provides an equitable comparison of the models, then $\hat{P}_C^* < 0.50$ could be considered a reasonable threshold for deciding in favor of the alternative.

With that in mind, we briefly look forward to the application presented in detail in Section 10.1. In this study, the unadjusted BDCP under the PIKL discrepancy is $\hat{P}_{PIKL}^* = 0.048$, and $p_{LR} = 0.044$. When we apply the proposed k bias correction to the null and alternative discrepancy estimators, then the approximation of this bias-corrected BDCP under the PIKL discrepancy is $\hat{P}_C^* = 0.093$, nearly twice that of the unadjusted BDCP. However, if we apply a threshold of $\hat{P}_C^* < 0.50$ for choosing in favor of the alternative model, then this observed value of $\hat{P}_C^*$ still lends fairly strong support in favor of the alternative model.

CHAPTER 9
BENEFITS OF DCP / BDCP FRAMEWORK

In this chapter we address two important contributions of this work. First, by drawing a connection between the p-value and the BDCP, we can provide an alternative interpretation of the p-value, from which we gain new insights regarding the behavior of the p-value. Second, while we have shown a connection between the p-value and the BDCP when hypothesis testing assumptions are met, the BDCP framework can be applied to a considerably broader collection of settings than those in which the p-value is valid.

9.1 Insights Gained from BDCP / p-Value Connection

The standard interpretation of the p-value is arguably confusing and counter-intuitive, especially to students or researchers who must use statistics in their work but who may not specialize in the field. By drawing a connection between the p-value and the BDCP when hypothesis testing assumptions are met, we allow for a perhaps more intuitively pleasing interpretation of the p-value. Instead of interpreting the p-value in the usual manner, we can instead interpret it as a reflection of the probability, based on the sample at hand, that the fitted null model is closer to the "truth" than the fitted alternative, where proximity is based on a suitably chosen discrepancy. In other words, rather than assuming the null is true and calculating a quantity which reflects the probability of what was observed under this assumption, we can instead think of the p-value as a bootstrap-based probability that the null is, in a certain sense, “better” than the alternative, without the null having to be strictly true. This interpretation offers a frequentist interpretation of the p-value which is better aligned with assessing the probability of a hypothesis.

Providing a non-standard interpretation of the p-value also allows us to assess its behavior in a different light. For instance, consider a setting in which a subtle effect yields a small p-value due to a very large sample size. The standard interpretation of the p-value indicates that the result is statistically significant even though it may not be of practical importance. Viewing this phenomenon in the BDCP framework may be beneficial. In choosing between competing models based on the BDCP, concepts pertinent to statistical modeling naturally arise. For instance, there is a bias-variability tradeoff that is inherent in statistical modeling; a larger model should be less biased, but at the cost of increased variability. A small BDCP indicates that, for most bootstrap samples, the discrepancy tends to prefer the more complex fitted alternative model to the simpler fitted null model, since the subtle effect can be estimated with sufficient accuracy to justify its inclusion. Stated another way, the adverse impact of the increased variability of the alternative model is outweighed by the impact of the null model bias due to omitting a nonzero effect. This interpretation of the BDCP may provide a clearer way of understanding the bias-variability tradeoff than the standard interpretation of the p-value.

9.2 Broader Utility of DCP / BDCP Framework

Beyond the advantage of providing a new interpretation of the p-value, the BDCP also has the strength that it can be applied in a broader collection of settings than hypothesis testing. For instance, hypothesis testing requires the alternative model to be adequately specified or else the corresponding p-value may be invalid. The BDCP, on the other hand, provides a valid comparison of competing models, regardless of the veracity of the alternative model (although the BDCP may not approximate the p-value in this setting). Because the notion of either the null or alternative being true is hard to defend in many practical settings, this advantage of the BDCP greatly enhances its utility.

Hypothesis testing typically requires the null model to be nested within the larger alternative, but the BDCP under certain discrepancies, such as the KL discrepancy, does not require nested models in order to be valid. There are many settings in which we would like to compare nonnested models. For instance, suppose we wish to compare a model which enters an effect linearly and another which enters the effect as a categorical variable. Standard hypothesis testing cannot be used to distinguish between these models, while the BDCP under the KL or PIKL discrepancies can easily be used.

Formal hypothesis testing requires pre-planned hypotheses in order to control Type I and Type II error rates. However, in practice, hypothesis testing is often applied in instances in which the data is used to make decisions regarding the selection of model and which hypotheses to test. While standard hypothesis testing techniques will typically no longer control for Type I and Type II error rates when applied in this manner, such hypothesis tests can still provide useful information regarding the implausibility of the null hypothesis. Nevertheless, in instances in which the hypotheses are not strictly pre-planned, a model evaluation tool that is not associated with the formality of controlling error rates may be preferable, because the use of such a tool could reduce the risk of incorrect interpretations. The BDCP simply seeks to quantify which of two models is closer to the truth and is therefore unconcerned with these long-term error rates. Accordingly, the BDCP is well-suited for use in settings without pre-planned hypotheses because it may produce a model evaluation which is less likely to be misconstrued than the evaluation provided by the p-value.

The BDCP can be approximated for any discrepancy in which the plug-in principle can be applied. For instance, suppose we were interested in assessing the absolute deviations between a fitted model's J-dimensional parameter vector $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_J)^T$ and the true parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)^T$. Then, for $j = 1, \ldots, J$, we could put forth a discrepancy based on the sum of scaled absolute deviations (SAD), such as

$$d_{SAD}(g, \tilde{\boldsymbol{\theta}}) = \sum_{j=1}^{J} \frac{|\tilde{\theta}_j - \theta_j|}{\sqrt{\imath^{jj}(\boldsymbol{\theta})}},$$

where $\imath^{jj}(\cdot)$ represents the $(j,j)^{th}$ element of the inverse expected information matrix. Let the estimated parameter vector using the bootstrap sample be denoted $\tilde{\boldsymbol{\theta}}^* = (\tilde{\theta}_1^*, \tilde{\theta}_2^*, \ldots, \tilde{\theta}_J^*)^T$, and let the estimated parameter vector under the general model be denoted $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_J)^T$. Then, applying the plug-in principle, the bootstrap-based estimator of the SAD discrepancy is

$$d_{SAD}(\hat{g}, \tilde{\boldsymbol{\theta}}^*) = \sum_{j=1}^{J} \frac{|\tilde{\theta}_j^* - \hat{\theta}_j|}{\sqrt{i^{jj}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}},$$

where $i^{jj}()$ represents the $(j,j)^{th}$ element of the inverse observed information matrix.

To assess whether the null or alternative model has a smaller sum of scaled absolute deviations in a hypothesis testing framework would require developing distributional theory. However, distributional theory for absolute values is notoriously difficult, and thus deriving a p-value to compare models on the basis of absolute deviations would be challenging. Instead, a practitioner may begrudgingly use a Wald test in such a setting because it is based on a measure that reflects the weighted squared distance between the MLEs and the hypothesized values. On other hand, by applying the plug-in principle and drawing repeated bootstrap samples, one can easily approximate the BDCP for this discrepancy. Importantly, the BDCP under the SAD discrepancy need not approximate a known p-value.

We next introduce a generalization of the SAD discrepancy that may be of practical use and for which the BDCP can be easily approximated. Again, consider a setting with two competing nested models, where the larger model consists of J parameters, and the smaller model pre-specifies some subset of these parameters. Suppose that accurate estimation of some parameters is more important than for others. Such a setting could arise in applications in which there are a handful of parameters of interest, which the user will want to estimate very accurately, and a collection of nuisance parameters, which may not require as accurate of estimation. Then, one could put forth a weighted sum of absolute deviations (WAD) discrepancy, where the weights are user-defined. We define the WAD discrepancy as

$$d_{WAD}(g, \tilde{\boldsymbol{\theta}}) = \sum_{j=1}^{J} \frac{w_j|\tilde{\theta}_j - \theta_j|}{\sqrt{\imath^{jj}(\boldsymbol{\theta})}},$$

where $w_1, \ldots, w_J$ are the user-defined weights. The WAD discrepancy is equivalent to the SAD discrepancy when $w_1 = \ldots = w_J = 1$. Similar to the bootstrap-based estimator of the SAD discrepancy, the estimator of the WAD discrepancy is

$$d_{WAD}(\hat{g}, \tilde{\boldsymbol{\theta}}^*) = \sum_{j=1}^{J} \frac{w_j|\tilde{\theta}_j^* - \hat{\theta}_j|}{\sqrt{i^{jj}(\hat{\boldsymbol{\theta}}|\boldsymbol{y})}}.$$

Comparing two models on the basis of a sum of weighted absolute deviations would be very difficult in the hypothesis testing framework, but can easily be achieved in the BDCP framework. In Subsection 10.1.3, for one of our applications, we calculate the BDCP under the WAD discrepancy for a variety of weighting schemes.

In this dissertation we primarily focused on discrepancies whose BDCP approximates the p-value, but practitioners do not need to confine their choice of discrepancy to this small class. We believe that, regardless of the connection to the p-value, the BDCP can be a valuable piece of information in choosing between two competing statistical models.

CHAPTER 10
APPLICATIONS

In this chapter, we apply our methodology to two biomedical applications. The primary purpose of these applications is to show that the Wald, score and LR p-values are approximated by their respective BDCPs. The first application investigates the effects of certain blood pressure medications on the probability of one-year survival in a high-risk Medicare cohort. For this application, we provide secondary results which illustrate that the BDCP can be applied in settings that do not provide a connection with the p-value. Specifically, we first illustrate that the BDCP can employ discrepancies which do not yield an approximation to the p-value. We then use the BDCP to compare nonnested models, which standard hypothesis testing is unable to do. The second application explores whether white patients are more likely than their non-white counterparts to receive antimicrobial prescriptions without a physician visit. We hypothesized that white patients may receive preferential treatment from some physicians, and thus be more likely to have prescriptions without a visit.

10.1 ACE/ARB and Survival Study

10.1.1 Overview

Our first study seeks to determine the effects of angiotensin converting enzyme inhibitors (ACEs) and angiotensin II receptor blockers (ARBs) on one-year survival for a high-risk Medicare population, all of whom have suffered an acute myocardial infarction (AMI). There exists some evidence supporting that ACEs and ARBs may be beneficial to patients who have suffered an AMI. For instance, Setoguchi et al. (2008) found that the use of ACE/ARBs helped explain a reduction in post-AMI patient mortality from 1995 to 2004. Also, in a large clinical trial, known as The Survival and Ventricular Enlargement (SAVE) study, Pfeffer et al. (1992) found treating patients who recently suffered an AMI with captopril, an ACE, led to a significantly decreased risk of mortality when compared to patients receiving a placebo. However, the mean age for patients in that clinical trial is 59.4 years ($sd = 10.6$), whereas the youngest a member of the present study's cohort can be is 66, and the mean age is 78.3 years ($sd = 7.9$). Thus, while the SAVE study presents strong evidence that captopril, and perhaps other ACEs, decrease the risk of mortality among a cohort of patients considerably younger than the present study's cohort, it is unable to definitively confirm that using an ACE is beneficial to a more elderly, and perhaps sicker, patient cohort.

The cohort consists of 8,682 Medicare beneficiaries, all of whom suffered an AMI (an inpatient stay with an ICD-9 diagnosis code of 410.x1) in 2007 or 2008. All patients were also discharged alive from the hospital stay in which the AMI was diagnosed and survived for at least 30 days post-discharge.

Unless a patient has a drug contraindication, medical practice dictates that patients suffering an AMI should typically be placed on either an ACE or an ARB (Smith et al., 2011; O'Gara et al., 2013). Despite the medical recommendation, only 4,327 (49.8%) members of the cohort filled a prescription for an ACE or ARB in the month following their discharge. A patient was considered an ACE/ARB user if and only if he or she filled a prescription for an ACE/ARB within 30 days of discharge. Note that in this study we do not differentiate between ACEs and ARBs; we simply create an indicator of whether the patient filled a prescription for either of the drugs in the 30 days post-discharge.

To help better understand the relationship between ACE/ARB use and survival, Table 10.1 presents a 2×2 table of ACE/ARB use and one-year survival. From Table 10.1, we determine that the unadjusted odds ratio comparing ACE/ARB use and one-year survival is 1.686 ($95\% \, CI : (1.496, 1.900)$; $p < 0.0001$). Thus, this perhaps naive analysis suggests quite strongly that ACE/ARB use increases the probability of one-year survival. However, this analysis is unable to account for the fact that patients who fill a prescription for an ACE/ARB may be considerably different than patients who do not. The result may simply be indicating that patients who receive an ACE/ARB are healthier on average and are thus less likely to die. Therefore, rather than rely on unadjusted analyses, we instead employ a multivariable logistic regression model to assess the relationship between ACE/ARB use and one-year survival. The model will control for a variety of covariates, including measures of patient demographics and socioeconomic status, measures of patient severity, comorbidities, drugs taken before the AMI, procedures before and during the AMI stay, drug contraindications, etc. To determine the importance of ACE/ARB use, we test the null hypothesis $H_0 : \beta = 0$ versus the general alternative $H_A : \beta \neq 0$, where β is the parameter corresponding to ACE/ARB use. The null model includes the control variables, and the alternative model includes the same control variables and the ACE/ARB indicator.

Table 10.1: 2 × 2 table showing the relationship between ACE/ARB use and one-year post-index discharge survival. A patient was considered an ACE/ARB user if and only if he or she filled a prescription for an ACE or ARB within 30 days of the index discharge date.

		One-Year Survival	
		yes	no
ACE/ARB use	yes	3814	513
	no	3550	805

10.1.2 Primary Results

To further understand this application, consider that the adjusted odds ratio estimate comparing ACE/ARB use and one-year survival is 1.151 with a Wald-based 95% confidence interval of $(1.002, 1.322)$. Thus, at the 0.05 significance level, we find ACE/ARB use to increase the probability of one-year survival, holding all other covariates constant. However, the adjusted result is considerably less significant ($p = 0.0440$) than the unadjusted results ($p < 0.0001$). Also, the estimated effect size is smaller in the adjusted results, with an estimated odds ratio of 1.151, whereas the unadjusted odds ratio estimate is 1.686. This illustrates the general concept that, when using observational data, the effect of a treatment may not be adequately characterized if we do not control for important covariates related to the probability of receiving treatment.

The Wald, LR and score test p-values as well their estimated BDCPs based on $B = 1,000$ bootstrap samples are presented in Table 10.2. Table 10.2 shows that the Wald, score and LR test p-values are very closely approximated by their bootstrap-based DCPs. For instance, we find that the LR p-value is 0.0440 and that the corresponding BDCP under the PIKL discrepancy is 0.048. These results suggest that we can then interpret the LR p-value as a bootstrap-based estimator of the probability that the fitted null model will have smaller overall PIKL discrepancy than the fitted alternative. Similar interpretations apply to the Wald and score test p-values. The idea of ACE/ARB use having no effect is not a scientifically valid hypothesis, so rejection of the null adds little to the underlying science. Instead, we may interpret the BDCP, and by its approximate equivalence the p-value, as a low probability that the model which does not account for ACE/ARB use is better than the alternative model which includes this predictor, without having to assume either candidate model precisely matches the truth. We can then conclude that the information from the sample is enough to estimate the effect of ACE/ARB use on survival with sufficient accuracy.

10.1.3 Secondary Results

As previously mentioned, use of the BDCP need not be limited to discrepancies which yield a connection with the p-value. To illustrate that the BDCP can be defined

Table 10.2: The Wald, score and LR p-values as well as their corresponding BDCPs determining the significance of ACE/ARB use on one-year survival.

Test	p-value	$\hat{P}^*$
Wald	0.0442	0.0460
Score	0.0441	0.0480
LR	0.0440	0.0480

and estimated for arbitrary discrepancies, we estimate the BDCP under the WAD discrepancy, which is described in Section 9.2. The BDCP will again be comparing the null model, which sets the ACE/ARB parameter to zero, and the alternative model, which estimates the ACE/ARB parameter. The WAD discrepancy requires a user-specified weighting scheme, so we apply a variety of weighting schemes to compare the two models. In each scheme, the ACE/ARB parameter receives a certain weight w, with the remaining $1 - w$ weight being distributed equally among the remaining 101 model parameters. We will let the weight on the ACE/ARB parameter be $w = 1, 0.50, 0.10, 0.05, 0.02$ and 0.0098. The weight of $w = 0.0098$ constitutes equal weighting across all model parameters. If a practitioner is interested only in the ACE/ARB indicator, then assessing the null and alternative models with $w = 1$ is a reasonable choice. On the other hand, if one is interested in an overall assessment of the model, without singling out any particular parameter, then calculating the BDCP with $w = 0.0098$ is justified.

Table 10.3 displays the BDCPs under the WAD discrepancy with the given weighting schemes. From Table 10.3, first note that each BDCP is less than 0.50, indicating that the alternative model is preferred in a majority of bootstrap samples for each studied weighting scheme. However, as the weight placed on the ACE/ARB indicator decreases, the corresponding BDCP increases. In other words, the null model fares better when it is compared across all parameters than when more weight is placed on its "estimator" of the ACE/ARB parameter. For instance, if we compare the models only on the basis of the ACE/ARB parameter estimator (i.e. using $w = 1$), then the alternative model is preferred in a large majority (94.96%) of bootstrap samples. On the other hand, when each parameter receives equal weighting, then the null model has a smaller discrepancy estimate in more than a quarter (27.21%) of bootstrap samples. Thus, depending on how we compare models, considerably different model assessments are possible. Further, note that using $w = 1$ produces a BDCP which approximates the p-value, but the weighting schemes which place less weight on the ACE/ARB indicator produce BDCPs that vary considerably from the p-value. This result illustrates that the BDCP can be approximated for discrepancies which do not necessarily have a connection with the p-value. All that is required for use of the BDCP is successful application of the plug-in principle, so users have wide latitude in choosing appropriate context-specific discrepancies. While the connection between the BDCP and the p-value is useful, practitioners may in certain instances prefer to use a discrepancy which does not provide such an approximation.

Suppose now that rather than determining the effect of ACE/ARBs, we are instead interested in determining whether age should be entered linearly or categori-

Table 10.3: The BDCP under the WAD discrepancy for a variety of weighting schemes. The weighting schemes place probability w on the ACE/ARB parameter and distribute the remaining $1 - w$ equally across the remaining parameters.

w	$\hat{P}^*$
1	0.0504
0.50	0.0568
0.10	0.0824
0.05	0.1121
0.02	0.2630
0.0098	0.2721

cally. Using the same modeling framework as before, we compare two models which contain the same set of predictors, except that the "null" model enters age linearly, and the "alternative" model enters ages categorically, with categories of 66-70, 71-75, 76-80, 81-85 and over 85. Here, the BDCP under the KL discrepancy is approximately 0.519, indicating no strong preference between the competing models. This analysis illustrates that the BDCP can easily compare nonnested models, a comparison for which standard hypothesis testing cannot be used.

10.2 Phantom Antimicrobial Prescribing Study

10.2.1 Overview

Our second application seeks to determine whether white patients who are prescribed antimicrobials are more likely to receive the prescription without a corresponding physician visit. This application is based on results from Riedle et al. (2017).

Antimicrobial resistance is a major public health concern. Decreasing the number of improper antibiotic prescriptions constitutes one important strategy in combatting antimicrobial resistance. Many important stewardship programs have been implemented to attempt to reduce inappropriate prescriptions. Nearly all such programs are to be applied during office visits. However, a considerable number of antimicrobial prescriptions are filled without an in-person visit to a physician. Thus, for these prescriptions without a corresponding visit (i.e. phantom prescriptions) traditional approaches designed to ensure proper prescribing may not be effective. Therefore, one important aspect of the Riedle et al. (2017) study is to determine important predictors of whether a prescription was filled without a corresponding visit. For this application, we specifically seek to determine whether white patients are more likely to receive at least one phantom antimicrobial prescription.

Our cohort consists of 3,262 Medicare recipients who suffered an AMI in 2007 or 2008. Further inclusion criteria include filling a prescription for at least one antimicrobial in the year following index discharge. A prescription was considered to be phantom if and only if the patient did not have any inpatient, outpatient or carrier claims in the 7 days, inclusive, prior to the date the prescription was filled. The outcome of interest is an indicator denoting whether the patient had at least one phantom antimicrobial prescription in the year following index discharge. The predictor is an indicator of whether the patient was white. Of the 3,262 patients, 2,717 (83.3%) were white, and 887 (27.2%) received at least one phantom antimicrobial prescription in the year following discharge.

In this application, we employ a logistic regression model. To determine whether being white is an important predictor of whether a patient receives a phantom antimicrobial, we test the null hypothesis $H_0 : \beta = 0$ versus the general alternative $H_A : \beta \neq 0$, where β is the parameter corresponding to the white indicator. The null model includes an intercept term, while the alternative model includes an intercept term and the white indicator.

10.2.2 Results

Table 10.4 presents the Wald, score and LR p-values as well as their corresponding BDCPs. The BDCPs are approximated using $B = 5,000$ bootstrap samples. From Table 10.4, we see that the Wald, score and LR p-values are each less than 0.02, with the LR p-value being slightly smaller than the other two. Similarly, the corresponding BDCPs are each approximately 0.02, with the LR-based BDCP being slightly smaller than the others. Importantly, we again see a strong level of agreement between the p-values and their respective BDCPs.

The odds ratio comparing the white indicator and whether the patient receives one or more phantom antimicrobial prescriptions was 1.294, with a Wald-based 95% confidence interval of $(1.042, 1.606)$. Thus, in the hypothesis testing context, of the patients who received at least one antimicrobial in the year after discharge, white patients were significantly more likely to have at least one of these prescriptions without a corresponding visit to a clinician. In the BDCP framework, we find that, across most bootstrap samples, the model which estimates the parameter corresponding to

Table 10.4: The Wald, score and LR p-values as well as their corresponding BDCPs determining the significance of being white on whether a patient receives a phantom antimicrobial prescription.

Test	p-value	$\hat{P}^*$
Wald	0.0194	0.0218
Score	0.0192	0.0206
LR	0.0176	0.0190

being white is preferred over the model which sets the white effect to zero. This result may be due to subtle biases in prescribing patterns or a whole host of other potential factors. Regardless of causality, our analysis suggests that white patients are more likely to receive antimicrobials without a visit and thus circumvent most antimicrobial stewardship programs.

CHAPTER 11
CONCLUDING REMARKS

11.1 Summary

When evaluating models on the basis of discrepancy functions, we merely wish to know which model is most congruous with the truth, without having to assume one of the candidate models is true. On the other hand, the paradigm for hypothesis testing typically assumes one of two competing models is precisely true. Despite these differences in underlying philosophy, when the assumptions of hypothesis testing are met, we have shown that a bootstrap-based discrepancy comparison probability estimator can approximate the Wald, likelihood ratio (LR) and score p-value.

The primary purpose of this dissertation was to introduce the discrepancy comparison probability (DCP) and show that, under specifically formulated discrepancies, its bootstrap-based estimator approximates the three aforementioned p-values. Because the bootstrap-based DCP approximates the p-value, our work does not alleviate many of the problems with or abuses of the p-value. Instead, this methodology allows us to conceptualize the p-value in a different way. The alternative interpretation of the p-value that our work provides is better aligned with assessing a probability on the null hypothesis, without having to assume the null hypothesis matches the truth.

This dissertation also addresses the bias of the bootstrap-based discrepancy estimators. Using suitably defined discrepancies, the p-value is approximated by the BDCP, which relies on these biased estimators of the overall discrepancies. Because the p-value is approximated by a quantity based on biased discrepancy estimators, a quantity employing unbiased discrepancy estimators may provide a more attractive comparison of the competing models than the p-value. For the Kullback-Leibler discrepancy, we quantify the bias of the discrepancy estimator and introduce a modified version of the BDCP which accounts for the bias of the null and alternative discrepancy estimators. While this modified version of the BDCP no longer approximates the p-value, it may be more useful to statistical practitioners because it should better identify when the null model is adequately specified.

While the principal goal of this thesis is to connect the BDCP with the p-value, the BDCP is more broadly applicable than hypothesis testing and the p-value. For instance, to establish a connection between the BDCP and the p-value, we needed to assume that the larger model was adequately specified. However, this assumption does not have to be met in order for the BDCP to be valid. Also, unlike the p-value derived from standard hypothesis testing, the BDCP under certain discrepancies can compare nonnested models. Further, while we have only considered the BDCP under a small number of discrepancies, the methodology presented here can be implemented for any discrepancy in which the plug-in principle can be applied. Finally, while the BDCP as it is formulated in this dissertation can only compare two models, using bootstrap-based estimators of overall discrepancies to delineate between models need not be limited to comparisons of only two, as illustrated by the methodology presented in Neath, Cavanaugh and Riedle (2012).

11.2 Future Directions

Because we believe the BDCP can be a practical model selection tool, future work will explore the potential uses of the BDCP, without regard for a connection to the p-value. Our methodology is quite flexible in the choice of discrepancy, only requiring successful application of the plug-in principle to derive an approximation of the BDCP. As shown in Subsection 10.1.3, the BDCP under the WAD discrepancy can produce considerably different model evaluations depending on the user-specified weighting scheme. Thus, future work will examine the behavior of the BDCP under a variety of discrepancies. Also, recall from Chapter 8 that the bootstrap-based estimator of the KL discrepancy is biased. We propose an alternative BDCP which accounts for the bias of the constituent discrepancy estimators. Future work will explore the bias of the bootstrap-based estimators for a variety of other discrepancies. We will also consider methods of correcting for this bias beyond those presented in this dissertation.

CHAPTER 12
REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, In *2nd International Symposium on Information Theory* (Eds., Petrov B. N. and Csáki, F.), pp. 267–281, Akadémia Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association* **82**, 112–122.

Berger, J.O., Wolpert, R.L., Bayarri, M.J., DeGroot, M.H., Hill, B.M., Lane, D.A. and LeCam, L. (1988). The likelihood principle. *Institute of Mathematical Statistics, Lecture Notes-Monograph Series* **6**, iii–199.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association* **33**, 526–536.

Box, G.E. (1976). Science and statistics. *Journal of the American Statistical Association* **71**, 791–799.

Box, G.E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A* **143**, 383–430.

Burnham, K. P. and Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Springer.

Casella, G. and Berger, R.L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association* **82**, 106-111.

Dufour, J.M. (2011). Coefficients of Determination, Technical Report. *McGill University, Center for Interuniversity Research and Analysis on Organizations, and Center for Interuniversity Quantitative Economics Research.*

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**, 461–470.

Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap.* CRC Press.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals of Internal Medicine* **130**, 995–1004.

Goodman, S.N. (2001). Of p-values and Bayes: A modest proposal. *Epidemiology* **12**, 295–297.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology* **45**, 135–140.

Gould, R. and Ryan, C. (2013). *Introductory Statistics: Exploring the World through Data.* Pearson Education.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B* **29**, 83–100.

Johnson, D.H. (1995). Statistical sirens: The allure of nonparametrics. *Ecology* **76**, 1998–2000.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Lawrence, I and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268.

Lindley, D.V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

McQuarrie, A.D. and Tsai, C. (1998). *Regression and Time Series Model Selection.* World Scientific.

Neath, A.A. (2017). A note on the connection between likelihood inference, Bayes factors, and p-values. *Open Access Library Journal* **4**, 1–11.

Neath, A.A., Cavanaugh, J.E. and Riedle, B. (2012). A bootstrap method for assessing uncertainty in Kullback-Leibler discrepancy model selection problems. *Mathematics in Engineering, Science & Aerospace* **3**, 381–391.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **236**, 333–380.

Neyman, J. (1979). $C(\alpha)$ tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A* **41**, 1–21.

O'Gara, P.T., Kushner, F.G., Ascheim, D.D., Casey, D.E., Chung, M.K., De Lemos, J.A., ... and Granger C.B. (2013). 2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction. *Journal of the American College of Cardiology* **61**, 78–140.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford University Press.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance* **12**, 30–32.

Pfeffer, M.A., Braunwald, E., Moyé L.A., Basta, L., Brown Jr, E.J., Cuddy, T.E., ... and Klein, M. (1992). Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction: Results of the Survival and Ventricular Enlargement Trial. *New England Journal of Medicine* **327**, 669–677.

Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

R Core Team. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. URL http://www.R-project.org/.

Riedle, B.N., Polgreen, L.A., Cavanaugh, J.E., Schroeder, M.C. and Polgreen, P.M. (2017). Phantom prescribing: Examining the frequency of antimicrobial prescriptions without a patient visit. *Infection Control & Hospital Epidemiology* **38**, 273–280.

RStudio Team. (2017). RStudio: Integrated development environment for R. *RStudio, Inc.* URL http://www.rstudio.com/.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**, 1151–1172.

Setoguchi, S., Glynn, R.J., Avorn, J., Mittleman, M.A., Levin, R. and Winkelmayer, W.C. (2008). Improvements in long-term mortality after myocardial infarction and increased use of cardiovascular drugs after discharge: a 10-year trend analysis. *Journal of the American College of Cardiology* **51**, 1247–1254.

Shafer, G. (1982). Lindley's Paradox. *Journal of the American Statistical Association* **77**, 325–334.

Smith, S.C., Benjamin, E.J., Bonow, R.O., Braun, L.T., Creager, M.A., Franklin, B.A., ... and Lloyd-Jones, D.M. (2011). AHA/ACCF secondary prevention and risk reduction therapy for patients with coronary and other atherosclerotic vascular disease: 2011 update. *Circulation* **124**, 2458–2473.

Trafimow, D. and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology* **37**, 1–2.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society* **54**, 426–482.

Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician* **70**, 129–133.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**, 60–62.

Windish, D.M., Huot, S.J. and Green, M.L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *Journal of the American Medical Association* **298**, 1010–1022.