VALID POST-SELECTION INFERENCE

Kai Zhang

A DISSERTATION
in
Statistics

For the Graduate Group in Managerial Science and Applied Economics
Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2012

Supervisor of Dissertation

____

Lawrence D. Brown, Miers Busch Professor, Statistics

Graduate Group Chairperson

____

Eric Bradlow, K.P. Chao Professor, Marketing, Statistics and Education

Dissertation Committee
Lawrence D. Brown, Miers Busch Professor, Statistics
Andreas Buja, The Liem Sioe Liong/First Pacific Company Professor, Statistics
Edward I. George, Universal Furniture Professor, Statistics
Dylan S. Small, Associate Professor of Statistics
Linda Zhao, Professor of Statistics

VALID POST-SELECTION INFERENCE

To my family.

Acknowledgement

My sincere gratitude goes to my family, all my teachers, and all my friends for their love and support over the past few years.

I am indebted to my thesis advisor Professor Lawrence D. Brown for his fatherly nurturing and guidance. His thinking is always extremely sharp and deep that inspires me to think further. His advice is always extremely insightful and helpful that stimulates me to work harder. He is also extremely patient and considerate to give me many warm encouragements that helps me to overcome difficulties. Although scientific research is never easy, Professor Brown is always my source of courage and power.

I thank my thesis committee members Professors Andreas Buja, Edward I. George, Dylan S. Small, and Linda Zhao for their kind and careful supervision. They provided me both deep understanding of the theories and fundamental tools in research. They also gave me very generous support in the past five years.

I am grateful to my collaborators Sathyanarayan Anand, Richard Berk, Dean Foster, Ruth Heller, Hongzhe Li, Scott Lorch, Zongming Ma, James Piette, Emil Pitkin, Alexander Rakhlin, Paul Rosenbaum, Jeffrey Silber, Sindhu Srinivas, and Mikhail Traskin, and instructors Tony Cai, Warren Ewens, Shane Jensen, Abba Krieger, Mark Low, Robin Pemantle, Paul Shaman, Lawrence Shepp, J. Michael

Steele, Robert Stine, Jessica Wachter, and Abraham Wyner. The inspiration and education I received from them will benefit my entire career.

Many thanks to the staff members Adam Greenberg, Daniel Huang, Catharina LeBlanc, Carol Reich, Andrew Romond, Sarin Sieng, Anand Srinivasan, Debbie Todd, and Tanya Winder for creating excellent environment for me to study here.

I would like to thank my colleague students. My years of graduate study here would not have been this colorful without them.

My final, and most heartfelt, acknowledgment must go to my family. This dissertation would not have been possible without their unconditional love and support. My debt to them is beyond measure.

ABSTRACT

VALID POST-SELECTION INFERENCE

Kai Zhang

Lawrence D. Brown

In the classical theory of statistical inference, data is assumed to be generated from a known model, and the properties of the parameters in the model are of interest. In applications, however, it is often the case that the model that generates the data is unknown, and as a consequence a model is often chosen based on the data. In my dissertation research, we study how to achieve valid inference when the model or hypotheses are data-driven. We study three scenarios, which are summarized in the three chapters.

In the first chapter, we study the common practice to perform data-driven variable selection and derive statistical inference from the resulting model. We find such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals when the model has been chosen a priori. We propose to produce valid "post-selection inference" by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing "simultaneity insurance" for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always more precise than full Scheffé protection. Importantly it does *not* depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

In the second chapter of this thesis, we propose a different approach to achieve valid post-selection inference which corresponds to the treatment of the design matrix predictors as random. Our methodology is based on two techniques, namely split samples and the bootstrap. Split-sample methodology generally involves dividing the observations randomly into two parts: one part for exploratory model building, a.k.a. the training set or planning sample, and the other part for confirmatory statistical inference, a.k.a. holdout set or analysis sample. We use a training sample only to seek a subset of predictors, and then perform the estimation and inference on the holdout set. As far as inference after selection in linear models is concerned, the main advantage of this technique is, roughly speaking, that it separates the data for exploratory analysis from the data for confirmatory analysis, thereby removing the contaminating effect of selection on inference. We show that the our procedure achieves valid inference asymptotically for any selection rule.

The third part of the thesis is an application of the split samples method to an observational study on the effect of obstetric unit closures in Philadelphia. The splitting was successful twice over: (i) it successfully identified an interesting and moderately insensitive conclusion, (ii) by comparison of the planning and analysis samples, it is clearly seen to have avoided an exaggerated claim of insensitivity to unmeasured bias that might have occurred by focusing on the least sensitive of many findings. Under the assumption of no unmeasured confounding, we found strong evidence that obstetric unit closures caused birth injuries. We also showed this conclusion to be insensitive to bias from a moderate amount of unmeasured confounding.

Contents

Dedication . . . . **iii**

Acknowledgement . . . . **iv**

Abstract . . . . **vi**

List of Tables . . . . **xiii**

List of Figures . . . . **xiv**

Preface . . . . **xv**

1 The PoSI Approach **1**

1.1 Introduction — The Problem with Statistical Inference after Model Selection 1

1.2 Model Selection Re-Interpreted 5

1.2.1 Post-Selection Inference for Full Model Parameters — a Dead End 5

1.2.2 The Meaning of Regression Coefficients 8

1.2.3 Assumptions, “Wrong Models”, and Error Estimates 12

1.3 Estimation and its Targets in Submodels 15

1.3.1 Multiplicity of Regression Coefficients 15

1.3.2 “Omitted Variables Bias” 18

1.3.3 Interpreting Regression Coefficients in First-Order Incorrect Models 19

1.4 Universally Valid Post-Selection Confidence Intervals 20

1.4.1 Test Statistics with One Error Estimate for All Submodels . . 20

1.4.2 Model Selection and Its Implications for Parameters 22

1.4.3 Valid Post-Selection Confidence Intervals 23

1.4.4 Universal Validity for all Selection Procedures 24
1.4.5 Restricted Model Selection 26
1.4.6 Reduction of Universally Valid Post-Selection Inference to Simultaneous Inference 28
1.4.7 Scheffé Protection 30
1.4.8 PoSI-Sharp Model Selection — "SPAR" and "SPAR1" 32
1.4.9 PoSI P-Value Adjustment for Model Selection 34
1.5 The Structure of the PoSI Problem 36
1.5.1 Canonical Coordinates 36
1.5.2 PoSI Coefficient Vectors in Canonical Coordinates 37
1.5.3 Orthogonalities of PoSI Coefficient Vectors 38
1.5.4 The PoSI Process 40
1.5.5 The PoSI Polytope 41
1.5.6 PoSI Optimality of Orthogonal Designs 42
1.5.7 A Duality Property of PoSI Vectors 45
1.6 Illustrative Examples and Asymptotic Results 47
1.6.1 Example 1: The PoSI Problem for Exchangeable Designs . . . 47
1.6.2 Example 2: Where $K(\mathbf{X})$ is close to the Scheffé Bound 50
1.6.3 Bounding Away from Scheffé 52

2 The Split Samples Approach 55
2.1 Introduction 55
2.1.1 Split Samples 56
2.1.2 Random Design View 57
2.1.3 Nonlinearity and Bootstrap 58
2.2 Population 60
2.3 Valid Inference in the Full Model 64
2.3.1 Least Squares Estimates in the Full Model 64
2.3.2 Bootstrap Inference for LS Estimates in the Full Model 66
2.4 Valid Post-Selection Inference via Split Samples and Bootstrap 69
2.4.1 Least Squares Estimates in A Submodel 69
2.4.2 Valid Bootstrap Inference under A Submodel 71
2.4.3 A Split Samples Procedure 72
2.5 Inference under First-Order Correctness and Homoscedasticity 74
2.5.1 The Properties of LS Estimates under First-Order Correctness and Homoscedasticity 75
2.5.2 Gaussian Case 76

2.5.3 Valid Bootstrap Inference in Submodels under the Gaussian Distribution 78

3 Using Split Samples in an Observational Study 79

3.1 Introduction: background; methodological outline 79

3.1.1 A wave of closures of hospital obstetrics units 79

3.1.2 Matching to build a control Philadelphia 81

3.1.3 Splitting 86

3.1.4 Evidence factors 89

3.2 Matching 91

3.2.1 Philadelphia and elsewhere, before and after matching 91

3.2.2 How the matching was done 95

3.3 Splitting 97

3.3.1 A 10%-90% random split for design and analysis 97

3.3.2 Birth injury in the planning sample: the largest difference, two nearly independent tests for effect and a test for unmeasured bias 99

3.4 Observational studies with binary outcome and difference-in-differences 104

3.4.1 Notation: base and intervention periods; exposed and unexposed regions 104

3.4.2 Model for sensitivity analysis 105

3.4.3 Sensitivity analysis with binary outcomes in difference-in-differences analysis 109

3.5 Confirmatory analysis using the 90% sample 113

4 Conclusion and Discussion 119

4.1 Conclusion and Discussion of the PoSI Approach 119

4.2 Conclusion and Discussion of the Split Samples Approach 121

4.3 Conclusion and Discussion of the Real Data Application of the Split Samples Method 123

4.4 Thoughts about Future Research 124

A Appendix 125

A.1 Proofs in Chapter 1 125

A.1.1 Proof of Theorem 1.4.3 125

A.1.2 Proof of Proposition 1.5.3 127

A.1.3 Proof of Duality: Lemma 1.5.1 and Theorem1.5.1 127

A.1.4 Proof of Theorem 1.6.1 130

A.1.5 Proof of Theorem 1.6.2 132
A.1.6 Proof of Theorem 1.6.3 135

Bibliography **139**

List of Tables

3.1 Covariate balance before and after matching 84
3.2 Results for birth injury in the planning component of the split sample 100
3.3 General form of the table under H_0 111
3.4 Results for birth injury in the analysis component of the split sample 114
3.5 Sensitivity analysis in the 90% analysis sample. 116

List of Figures

1.1 The PoSI polytope for $d = p = 2$ 43
1.2 The PoSI constant $K(\mathbf{X}, \alpha = 0.05)$ for exchangeable designs 48
1.3 The geometry of the limiting PoSI polytope for $p = 3$ for exchangeable designs 51

2.1 *The decomposition of* Y *when* $p = 1$. 62

3.1 Map of the City of Philadelphia showing hospitals that closed their obstetrics units 82
3.2 Covariate balance measured as the absolute difference in means . . . 94
3.3 Covariate imbalance for four continuous covariates 96

Preface

In a series of papers, John Tukey (1980a; 1980b) envisioned the future of statistics by distinguishing two important aspects: exploratory data analysis and confirmatory data analysis. Indeed, in many modern applications of statistics, data analyses are guided by both of the following ideas: (1) People want to understand the mechanism in the data with a parsimonious model. This model is to be found through some data-driven model selection; and (2) People want to make valid statistical inference from the model they select. In light of these ideas, I have been working on methodologies that achieve valid inference after model selection. My work is summarized in the three parts of this dissertation.

The first part of this thesis is based on the paper *Valid Post-Selection Inference (2012)* which is a joint work with Richard Berk, Lawrence Brown, Andreas Buja, and Linda Zhao. The second part of this thesis is based on a paper in preparation that will be a joint work with Richard Berk, Lawrence Brown, Andreas Buja, Edward George, Emil Pitkin, Mikhail Traskin, and Linda Zhao. The third part of this thesis is based on the paper *Using Split Samples and Evidence Factors in an Observational Study of Neonatal Outcomes (2011)* which is a joint work with Scott Lorch, Paul Rosenbaum, Dylan Small, and Sindhu Srinivas. I am very thankful for the inspiration and suggestions from my collaborators. In particular, I am grateful to Lawrence Brown, Andreas Buja, Paul Rosenbaum, and Dylan Small for their substantial contributions to the text of the manuscripts.

Chapter 1

The PoSI Approach

1.1 Introduction — The Problem with Statistical Inference after Model Selection

Classical statistical theory grants validity of statistical tests and confidence intervals assuming a wall of separation between the selection of a model and the analysis of the data being modeled. In practice, this separation rarely exists and more often a model is "found" by a data-driven selection process. As a consequence inferential guarantees derived from classical theory are invalidated. Among model selection methods that are problematic for classical inference, *variable selection* stands out because it is regularly taught, commonly practiced, and highly researched as a technology. Even though statisticians may have a general awareness that the data-driven selection of variables (predictors, covariates) must somehow affect subsequent classical inference from F- and t-based tests and confidence intervals, the practice is so pervasive that it appears in classical undergraduate textbooks on statistics such as Moore and McCabe (2003).

The reason for the invalidation of classical inference guarantees is that a data-driven variable selection process produces a model that is itself stochastic, and this stochastic aspect is not accounted for by classical theory. Models become stochastic when the stochastic component of the data is involved in the selection process. (In regression with fixed predictors the stochastic component is the response.) Models are stochastic in a well-defined way when they are the result of formal variable selection procedures such as stepwise or stagewise forward selection or backward elimination or all-subset searches driven by complexity penalties (such as C_p, AIC, BIC, risk-inflation, LASSO, ...) or prediction criteria such as cross-validation, or recent proposals such as LARS and the Dantzig selector (for an overview see, for example, Hastie, Tibshirani, and Friedman (2009)). Models are also stochastic but in an ill-defined way when they are informally selected through visual inspection of residual plots or normal quantile plots or generally through activities that may be characterized as "data snooping". Finally, models become stochastic in the most opaque way when their selection is affected by human intervention based on post hoc considerations such as "in retrospect only one of these two variables should be in the model" or "it turns out the predictive benefit of this variable is too weak to warrant the cost of collecting it." In practice, all three modes of variable selection may be exercised in the same data analysis: multiple runs of one or more formal search algorithms may be performed and compared, the parameters of the algorithms may be subjected to experimentation, and the results may be critiqued with graphical diagnostics; a round of fine-tuning based on substantive deliberations may finalize the analysis.

Posed so starkly, the problems with statistical inference after variable selection may well seem insurmountable. At a minimum, one would expect technical solutions to be possible only when a formal selection algorithm is (1) well-specified (1a) in advance and (1b) covering all eventualities, (2) strictly adhered to in the course of data analysis, and (3) not "improved" on by informal and post-hoc elements. It may, however, be unrealistic to expect this level of rigor in most data analysis contexts, with the exception of well-conducted clinical trials. The real challenge is therefore to devise statistical inference that is valid following any type of variable selection, be it formal, informal, post hoc, or a combination thereof. Meeting this challenge with a relatively simple proposal is the goal of this article. This proposal for valid ***Po***st-***S***election ***I***nference, or "***PoSI***" for short, consists of a large-scale family-wise error guarantee that can be shown to account for all types of variable selection, including those of the informal and post-hoc varieties. On the other hand, the proposal is no more conservative than necessary to account for selection, and in particular it can be shown to be less conservative than Scheffé's simultaneous inference.

The framework for our proposal is in outline as follows — details to be elaborated in subsequent sections: We consider linear regression with predictor variables whose values are considered fixed, and with a response variable that has normal and homoscedastic errors. The framework does not require that any of the eligible linear models is correct, not even the full model, as long as a valid error estimate is available. We assume that the selected model is the result of some procedure that makes use of the response, but the procedure does not need to be fully specified. A crucial aspect of the framework concerns the use and interpretation of the selected model: We assume that, after variable selection is completed, the selected predictor variables — and only they — will be relevant; all others will be eliminated from further consideration. This assumption, seemingly innocuous and natural, has critical consequences: It implies that statistical inference will be sought for the coefficients of the selected predictors only and in the context of the selected model only. Thus the appropriate targets of inference are the best linear coefficients within the selected model, where each coefficient is adjusted for the presence of all other included predictors but not those that were eliminated. Therefore the coefficient of an included predictor generally requires inference that is specific to the model in which it appears. Summarizing in a motto, a difference in adjustment implies a difference in parameters and hence in inference. The goal of the present proposal is therefore simultaneous inference for all coefficients in all submodels. Such inference can be shown to be valid following any variable selection procedure, be it formal, informal, post hoc, fully or only partly specified.

Problems associated with post-selection inference were recognized long ago, for example, by Buehler and Fedderson (1963), Brown (1967), Olshen (1973), Sen (1979), Sen and Saleh (1987), Dijkstra and Veldkamp (1988), Pötscher (1991), Kabaila (1998). More recently these problems have been the subject of incisive analyses by Leeb and Pötscher (2003; 2005; 2006a; 2006b; 2008a; 2008b), Kabaila and Leeb (2006), Leeb (2006), and Pötscher and Leeb (2009).

This article proceeds as follows: Section 1.2 starts by outlining some unsolvable difficulties of post-selection inference as they transpire from the work of Leeb and Pötscher cited above (Section 1.2.1); we then rethink the assumptions underlying their analyses and lay some groundwork by proposing new (or old) meanings for regression coefficients (Section 1.2.2); we conclude the section by discussing assumptions with a view towards valid inference in "wrong models" (Section 1.2.3). Section 1.3 is about estimation and its targets; Section 1.4 develops the methodology for PoSI confidence intervals (CIs) and tests. After some structural results for the PoSI problem in Section 1.5 , we show in Section 1.6 that with increasing number of predictors p the width of PoSI CIs can range between the asymptotic rates $O(\sqrt{\log p})$ and $O(\sqrt{p})$. We give examples for both rates and, inspired by problems in sphere packing and covering, we give upper bounds for the limiting constant in the $O(\sqrt{p})$ case. Some proofs are deferred to the appendix.

1.2 Model Selection Re-Interpreted

1.2.1 Post-Selection Inference for Full Model Parameters — a Dead End

It is a natural intuition that model selection distorts inference by distorting sampling distributions of parameter estimates: One expects that estimates in selected models tend to generate more Type I errors than conventional theory would suggest because the typical selection procedure favors models with strong, hence highly significant predictors. This intuition correctly points toward a multiplicity problem which would tend to become more severe as the number of predictors subject to selection increases. This problem will be addressed here with a simultaneous inference approach.

A second problem with inference after model selection is pointed out by Leeb and Pötscher in the above referenced series of articles. The problem exists even in a two-predictor situation, as illustrated by Leeb and Pötscher (2005): They analyze a case with a predictor that is protected from selection and a covariate that is subject to selection, and they provide an explicit finite-sample formula for the sampling distribution of the coefficient estimate of the protected predictor, as the covariate is randomly selected/deselected according to a BIC-equivalent test to grant consistent model selection (ibid., p. 29). The analysis reveals in graphic ways (ibid., Figure 2) that the sampling distribution depends critically on the unknown true coefficient of the covariate and the sample size, with egregious deviations from the fixed-model sampling distribution ranging from bi-modality to approximate normality with inflated variance. Because the true covariate slope is not known, there is no way of determining whether the sample size places the sampling distribution in this realm of deviation from classical theory.

Generalizing to arbitrary linear models Leeb and Pötscher (2003; 2005; 2006a; 2006b; 2008a; 2008b) show that sampling distributions cannot be estimated after model selection, not even asymptotically. Ironically, the asymptotics that describe the devious finite sample behavior of sampling distributions best are those based on consistent model selection. They show that asymptotic normality is riddled with extreme non-uniformity of convergence and that risk functions behave erratically when telescoping true slopes to zero so as to approach submodels. Leeb and Pötscher (2005, p. 27) arrive at the following conclusion: "the post-model-selection estimator ... is nothing else than a variant of Hodges' so-called superefficient estimator."

It is little comfort that these problems are non-existent for perfectly orthogonal regression designs (Leeb and Pötscher 2005, p. 43f). In the majority of practical contexts there is some degree of collinearity, and one of the purposes of model selection is to weed out predictor redundancies caused by partial collinearity. Leeb and Pötscher's analysis is compelling within their framework, but the intractable situation they expose suggests a need to renegotiate the assumptions that underlie their framework.

Leeb and Pötscher (ibid.), like many authors in this area, make the fundamental assumption that all estimation is in the full model. Thus, if a model selection procedure excludes a predictor, this is interpreted as forcing the estimate of its slope to zero. Consequently, a slope estimate $\hat{\beta}_j$ of a predictor is always defined, whether it is selected or not: $\hat{\beta}_j$ is the LS estimate in the selected submodel if the j^{th} predictor is included, and it is zero otherwise. Either way, the resulting value is interpreted as an estimate of β_j in the full model. A parallel consequence is that in this interpretation the coefficient of a predictor has always the same meaning as a full-model parameter, irrespective of which covariates are selected or excluded. It is under this framework that post-selection estimators can be interpreted as generalized superefficient Hodges estimators with the ensuing problems of non-uniformity (Leeb and Pötscher 2005). This problem can also be seen as an inferential analog of the problem of "omitted variables bias" well-known in econometrics (see, for example, Angrist and Pischke 2009).

1.2.2 The Meaning of Regression Coefficients

Our solution to the inferential problem associated with "omitted variables bias" is to assert that submodels have their own separate parameters, and it is these that are being estimated in the selected submodels. We start the discussion with the following questions:

(1) When we select submodels in practice, do we think of excluded predictors as having a zero slope?

(2) Does the full model necessarily have special status?

(3) Can a slope estimate be interpreted as estimating the same target parameter, regardless of what the other predictors are?

The short answers are:

(1) The slopes of excluded predictors are not zero; they are not defined and therefore don't exist.

(2) The full model has no special status.

(3) The meaning of a slope depends on which predictors are included in the selected model.

These statements call for elaboration:

As for (1), assigning a zero value to predictors that are not in the model is an elegant technical device, but it is not something that describes how we think or even how we *should* think about slopes and their estimates. The PoSI framework we describe in Section 1.3 will not require zero slope fill-ins.

As for (2), the full model cannot be argued to have generally special status because there is generally a question of predictor redundancy. It is a common experience that models are proposed on theoretical grounds but found on empirical grounds to have their predictors hopelessly entangled by collinearities that permit little meaningful statistical inference. This is best illustrated with a concrete example (inspired by Mosteller and Tukey (1977), p. 326f): Consider a study of performance of students in a large school system. Interested in socio-economic factors, investigators wish to pin down the predictors that are most strongly associated with children's success in school: father's and mother's highest education levels, their high school GPAs, their SAT scores, their frequencies of intensive reading, the perceived importance they each assign to education, and so on. There should be little surprise that, if all these predictors are included in the model, the overall test rejects but none of the individual predictors is statistically significant. Informal model selection, however, may show that each predictor is highly statistically significant if retained alone in the model. The obvious reason is that these predictors measure essentially the same trait in parents, hence are highly collinear with each other. As a consequence, the full model is not viable in the first place. This situation is not limited to the social sciences: in gene expression studies it may well occur that numerous sites have a tendency to be expressed concurrently, hence as predictors in disease studies they will be hopelessly confounded. The bias in favor of full models may be particularly strong in econometrics where there is a "notion that a longer regression ... has a causal interpretation, while a shorter regression does not" (Angrist and Pischke 2009, p. 59). Even in causal models, however, there is a possibility that included adjustor variables will "adjust away" some of the causal variables of interest. Generally, in any creative observational study involving novel predictors it will be difficult to exclude surprising collinearities that might force a rethinking of the role of predictors. In conclusion, whenever predictor redundancy is a potential issue, it cannot a priori be claimed that the full model provides the parameters of primary interest.

As for (3), we do teach that the meaning of a slope depends on what other predictors are included in the chosen model: "the slope is the average difference in the response for a unit difference in the predictor, *at fixed levels of all other predictors.*" This last condition is sometimes rendered as "*adjusted for all other predictors*" and called the "Ceteris Paribus" clause (see, for example, Angrist and Pischke 2009). It is an essential part of the meaning of a slope. That there is a difference in meaning when there is a difference in covariates is most drastically evident when there is a case of Simpson's paradox. This is again best illustrated with a concrete example: A company is introducing a new high-tech device and conducts a consumer survey that includes a response for self-reported purchase likelihood (PL), as well as two predictors, Age and $Income$. We consider a model with Age alone and one with both Age and $Income$. [Note that the smaller model cannot be disregarded as "wrong". If $Income$ is difficult to measure, it may be useful to rely on the equation with Age alone. Further, if the variables have a jointly normal distribution, every linear submodel is "correct".] Now, it is sensibly anticipated that younger respondents will rate themselves with higher PL, but a regression of PL on *Age* alone produces a significantly positive slope estimate, indicating that older respondents have higher PL. On the other hand, a regression of PL on both *Age* and $Income$ yields a significantly negative slope estimate for *Age*, indicating that, *comparing only respondents at the same* $Income$ *level*, younger respondents have indeed higher PL. This instance of Simpson's paradox is enabled by a positive collinearity between *Age* and $Income$ that turns *Age* into a proxy for $Income$. Must we use the full model? Not if the improvement in R^2 is practically irrelevant even though $Income$ is statistically significant (apart from the issue of availability of $Income$ measurements). Is the marginal slope of PL on *Age* an estimate of the *Income*-adjusted slope on *Age*? Certainly not — the two slopes answer very different questions, apart from having opposite signs. In conclusion, *differences in adjustment result in different parameters.*

From these considerations follows a framework in which the full model is no longer the sole provider of parameters, where rather each submodel defines its own. The consequences of this view will be elaborated in Section 1.3.

In the preceding discussions we assumed a focus on the interpretation of the selected submodel and hence on inference for its coefficients. When the focus is on prediction, on the other hand, the focus is on predicted values. Yet, even in prediction problems there is sometimes a question of which predictors matter most within a selected submodel, and here a suitable t-statistic of a coefficient estimate is a measure of the predictive power of a given predictor above and beyond the other predictors in the submodel. In this context the submodel-specific parameters are the appropriate ones to consider.

1.2.3 Assumptions, "Wrong Models", and Error Estimates

We state assumptions for estimation and for the construction of valid tests and CIs. A major goal is to prepare the ground for valid statistical inference after model selection in "first order wrong models".

We consider a quantitative response vector $\mathbf{Y} \in \mathbb{R}^n$, assumed random, and a full predictor matrix $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n \times p}$, assumed fixed. We allow $\mathbf{X}$ to be of non-full rank, and n and p to be arbitrary. In particular, we allow $n < p$. Throughout the article we let

$$d \;\triangleq\; \mathrm{rank}(\mathbf{X}) = \dim(\mathrm{span}(\mathbf{X})), \qquad \text{hence} \;\; d \leq \min(n, p). \tag{1.2.1}$$

Due to frequent reference we call $d = p$ $(\leq n)$ "**the classical case**".

It is common practice to assume the full model $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ to be correct. In the present framework, however, first-order correctness, $\mathbf{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$, will not be assumed. By implication, first-order correctness of any submodel will not be assumed either. Effectively,

$$\boldsymbol{\mu} \;\triangleq\; \mathbf{E}[\mathbf{Y}] \in \mathbb{R}^n \tag{1.2.2}$$

is allowed to be unconstrained and, in particular, need not reside in the column space of $\mathbf{X}$. In other words, the model given by $\mathbf{X}$ is allowed to be "first-order wrong", and hence we are in a well-defined sense serious about G.E.P. Box' famous dictum. What he calls “wrong models”, however, we prefer to call “approximations”: All predictor matrices $\mathbf{X}$ provide approximations to $\boldsymbol{\mu}$, some better than others, but the degree of approximation plays no role in the clarification of statistical inference. We will echo Box as follows: all models are mere approximations, yet some are useful. The main reason for elaborating this point is as follows: after model selection the case for “correct models” is clearly questionable, even for “consistent model selection procedures” (Leeb and Pötscher 2003, p. 101); but if correctness of submodels is not assumed, it is only natural to abandon this assumption for the full model also, in line with the idea that the full model has no special status. As we proceed with estimation and inference guarantees in the absence of first-order correctness we will rely on assumptions as follows:

- For estimation (Section 1.3), we will only need the existence of $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$.
- For testing and CI guarantees (Section 1.4), we will make conventional second order and distributional assumptions:

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}). \tag{1.2.3}$$

The assumptions (1.2.3) of homoscedasticity and normality are as questionable as first order correctness, and we will report elsewhere on approaches that avoid them. In the present work, we choose to follow the large model selection literature that relies on the technical advantages of assuming homoscedastic and normal errors.

Accepting the assumption (1.2.3), we address the issue of estimating the error variance σ^2, because the valid tests and CIs we construct require a valid estimate $\hat{\sigma}^2$

of σ^2 that is independent of LS estimates. In the classical case, the most common way to assert such an estimate is to assume that the full model is first order correct, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ in addition to (1.2.3), in which case the mean squared error (MSE), $\hat{\sigma}_F^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$, of the full model will do. However, other possibilities for producing a valid estimate $\hat{\sigma}^2$ exist, and they may allow relaxing the assumption of first order correctness:

- Exact replications of the response obtained under identical conditions might be available in sufficient numbers. An estimate $\hat{\sigma}^2$ can be obtained as the MSE of the one-way ANOVA of the groups of replicates.
- In general, a larger linear model than the full model might be considered as correct, hence $\hat{\sigma}^2$ could be the MSE from this larger model.
- A different possibility is to use another dataset, similar to the one currently being analyzed, to produce an independent estimate $\hat{\sigma}^2$ by whatever valid estimation method.
- A special case of the preceding is a random split-sample approach whereby one part of the data is reserved for producing $\hat{\sigma}^2$ and the other part for estimating coefficients, selecting models, and carrying out post-model selection inference.
- A very different type of estimates $\hat{\sigma}^2$ may be based on considerations borrowed from non-parametric function estimation (Hall and Carroll 1989).

The purpose of pointing out these possibilities is to separate at least in principle the issue of first-order model incorrectness from the issue of valid and independent error estimation under the assumption (1.2.3). This separation puts the case $n < p$

within our framework as the valid and independent estimation of σ^2 is a problem faced by all "$n < p$" approaches.

1.3 Estimation and its Targets in Submodels

Following Section 1.2.2, the meaning and numeric value of a regression coefficient depends on what the other predictors in the model are. This statement requires a qualification: it assumes that the predictors are non-orthogonal/partially collinear. If they are perfectly pairwise orthogonal, as in some designed experiments or in function fitting with orthogonal basis functions, a coefficient has the same identity across all submodels, both in meaning and in value, because adjustment of predictors for each other and the ceteris paribus clause become vacuous. This article is hence largely a story of (partial) collinearity.

1.3.1 Multiplicity of Regression Coefficients

We will give meaning to LS estimators and their targets in the absence of any assumptions other than the existence of $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$, which in turn is permitted to be entirely unconstrained in $\mathbb{R}^n$. Besides resolving the issue of estimation in "first order wrong models", the major point here is to follow up on the idea that the regression coefficient of a predictor generates different parameters in different submodels. As each predictor appears in 2^{p-1} submodels, the p regression coefficients of the full model generally proliferate into a plethora of as many as $p\,2^{p-1}$ distinct regression coefficients according to the submodels they appear in. To describe the situation we start with notation.

To denote a submodel we use the (non-empty) index set $\mathrm{M} = \{j_1, j_2, ..., j_m\} \subset \mathrm{M}_F = \{1, \ldots, p\}$ of the predictors $\mathbf{X}_{j_i}$ in the submodel; the size of the submodel is $m = |\mathrm{M}|$ and that of the full model is $p = |\mathrm{M}_F|$. Let $\mathbf{X}_\mathrm{M} = (\mathbf{X}_{j_1}, ..., \mathbf{X}_{j_m})$ denote the $n \times m$ submatrix of $\mathbf{X}$ with columns indexed by M. We will assume that only submodels M are considered for which $\mathbf{X}_\mathrm{M}$ is of full rank:

$$\mathrm{rank}(\mathbf{X}_\mathrm{M}) \;=\; m \;\leq\; d.$$

We let $\hat{\boldsymbol{\beta}}_\mathrm{M}$ be the unique least squares estimate in M:

$$\hat{\boldsymbol{\beta}}_\mathrm{M} = (\mathbf{X}_\mathrm{M}^T \mathbf{X}_\mathrm{M})^{-1} \mathbf{X}_\mathrm{M}^T \mathbf{Y}. \tag{1.3.1}$$

Now that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ is an estimate, what is it estimating? A conclusion from Section 1.2.1 is that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ does not estimate the coefficients in the full model. Because any larger model could have been the full model, we generalize by asserting that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ does not estimate parameters in any other model than M itself. In M, it is natural to ask that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ be an unbiased estimate of its target:

$$\begin{aligned} \boldsymbol{\beta}_\mathrm{M} \;\triangleq\; \mathbf{E}[\hat{\boldsymbol{\beta}}_\mathrm{M}] \;&=\; (\mathbf{X}_\mathrm{M}^T \mathbf{X}_\mathrm{M})^{-1} \mathbf{X}_\mathrm{M}^T \, \mathbf{E}[\mathbf{Y}] \\ &=\; \operatorname*{argmin}_{\boldsymbol{\beta}' \in \mathbb{R}^m} \|\boldsymbol{\mu} - \mathbf{X}_\mathrm{M} \boldsymbol{\beta}'\|^2 \end{aligned} \tag{1.3.2}$$

This definition requires only the existence of $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$ but no other assumptions. In particular there is no need to assume first order correctness of M or M_F. Nor does it matter to what degree M provides a good approximation to $\boldsymbol{\mu}$ in terms of approximation error $\|\boldsymbol{\mu} - \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}_{\mathrm{M}}\|^2$. Asserting that the model M is "correct" would mean $\boldsymbol{\mu} \in \mathrm{span}(\mathbf{X}_{\mathrm{M}})$ or equivalently the approximation error vanishes; in this case $\boldsymbol{\beta}_{\mathrm{M}}$ would be the "true" parameter.

In the classical case $d = p \leq n$, we can define the target of the full-model estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ as a special case of (1.3.2) with $\mathrm{M} = \mathrm{M}_F$:

$$\boldsymbol{\beta} \;\triangleq\; \mathbf{E}[\hat{\boldsymbol{\beta}}] \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}[\mathbf{Y}]. \tag{1.3.3}$$

In the general (non-classical) case, let $\boldsymbol{\beta}$ be any (possibly non-unique) minimizer of $\|\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}'\|^2$; the link between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{\mathrm{M}}$ is as follows:

$$\boldsymbol{\beta}_{\mathrm{M}} \;=\; (\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^T\mathbf{X}\boldsymbol{\beta}. \tag{1.3.4}$$

Thus the target $\boldsymbol{\beta}_{\mathrm{M}}$ is an estimable linear function of $\boldsymbol{\beta}$, without any first-order assumptions. Equation (1.3.4) follows from $\mathrm{span}(\mathbf{X}_{\mathrm{M}}) \subset \mathrm{span}(\mathbf{X})$.

Notation: To distinguish regression coefficients as a function of the model they appear in, we write $\beta_{j\cdot\mathrm{M}} = \mathbf{E}[\hat{\beta}_{j\cdot\mathrm{M}}]$ for the components of $\boldsymbol{\beta}_{\mathrm{M}} = \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}]$ with $j \in \mathrm{M}$. An important convention we adopt throughout this article is that the index j of a coefficient refers to the coefficient's index in the original full model M_F: $\beta_{j\cdot\mathrm{M}}$ for $j \in \mathrm{M}$ refers not to the j'th coordinate of $\boldsymbol{\beta}_{\mathrm{M}}$, but to the coordinate of $\boldsymbol{\beta}_{\mathrm{M}}$ corresponding to the j'th predictor $\mathbf{X}_j$ in the full predictor matrix $\mathbf{X}$. We refer to this convention as "**full model indexing**".

1.3.2 “Omitted Variables Bias”

By allowing each $\hat{\beta}_{j\cdot\mathrm{M}}$ to estimate its own target $\beta_{j\cdot\mathrm{M}}$ and thereby relieving $\hat{\beta}_{j\cdot\mathrm{M}}$ of the burden of estimating the parameter β_j in the full model, we sidestep the problem of “omitted variables bias” and with it a major driver of the problems analyzed by Leeb and Pötscher (Section 1.2.1). In the present framework $\beta_j - \beta_{j\cdot\mathrm{M}}$ is not a bias as these are two different parameters that answer two different questions. Just the same, we consider briefly the difference between β_j and $\beta_{j\cdot\mathrm{M}}$ in the classical case $d{=}p{\leq}n$. Compare the following two definitions:

$$\boldsymbol{\beta}_{\mathrm{M}} \triangleq \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] \quad \text{and} \quad \boldsymbol{\beta}^{\mathrm{M}} \triangleq (\beta_j)_{j\in\mathrm{M}}, \tag{1.3.5}$$

the latter being the coefficients β_j from the full model M_F subsetted to the submodel M. While $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ estimates $\boldsymbol{\beta}_{\mathrm{M}}$, it does *not* generally estimate $\boldsymbol{\beta}^{\mathrm{M}}$. The difference $\boldsymbol{\beta}^{\mathrm{M}} - \boldsymbol{\beta}_{\mathrm{M}}$ is the vectorized “omitted variables bias”.

In general, the definition of $\boldsymbol{\beta}_{\mathrm{M}}$ involves $\mathbf{X}$ and all of $\boldsymbol{\beta}$, not just $\boldsymbol{\beta}^{\mathrm{M}}$, through (1.3.4). A little algebra shows that $\boldsymbol{\beta}_{\mathrm{M}} = \boldsymbol{\beta}^{\mathrm{M}}$ if and only if

$$\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}^c} \boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}, \tag{1.3.6}$$

where M^c denotes the complement of M in the full model M_F. Special cases of (1.3.6) include: (1) the column space of $\mathbf{X}_{\mathrm{M}}$ is orthogonal to that of $\mathbf{X}_{\mathrm{M}^c}$, and (2) $\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}$, meaning that the approximation to $\boldsymbol{\mu}$ in M_F is no better than in M, or if the full model M_F is first-order correct, so is the submodel M.

1.3.3 Interpreting Regression Coefficients in First-Order Incorrect Models

The regression coefficient $\beta_{j\cdot\mathrm{M}}$ is conventionally interpreted as the "average difference in the response for a unit difference in X_j, ceteris paribus in the model M". This interpretation no longer holds when the assumption of first order correctness is given up. Instead, the phrase "average difference in the response" should be replaced with the unwieldy but more correct phrase "average difference in the response approximated in the submodel M". The reason is that the fit in the submodel M is $\hat{\mathbf{Y}}_{\mathrm{M}} = \mathbf{H}_{\mathrm{M}}\mathbf{Y}$ ($\mathbf{H}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^T$) whose target is $\boldsymbol{\mu}_{\mathrm{M}} = \mathbf{E}[\hat{\mathbf{Y}}_{\mathrm{M}}] = \mathbf{H}_{\mathrm{M}}\mathbf{E}[\mathbf{Y}] = \mathbf{H}_{\mathrm{M}}\boldsymbol{\mu}$. Thus in the submodel M we estimate not the true $\boldsymbol{\mu}$ but the LS approximation $\boldsymbol{\mu}_{\mathrm{M}}$ to $\boldsymbol{\mu}$ using $\mathbf{X}_{\mathrm{M}}$: $\boldsymbol{\mu}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}_{\mathrm{M}}$, where $\boldsymbol{\beta}_{\mathrm{M}} = \mathrm{argmin}_{\boldsymbol{\beta}'}\|\boldsymbol{\mu} - \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}'\|^2$.

A second interpretation of regression coefficients is in terms of adjusted predictors: For $j \in \mathrm{M}$ define the M-adjusted predictor $\mathbf{X}_{j\cdot\mathrm{M}}$ as the residual vector of the regression of $\mathbf{X}_j$ on all other predictors in M. Multiple regression coefficients, both estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ and parameters $\beta_{j\cdot\mathrm{M}}$, can be expressed as simple regression coefficients with regard to the M-adjusted predictor:

$$\hat{\beta}_{j\cdot\mathrm{M}} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Y}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}, \qquad \beta_{j\cdot\mathrm{M}} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\boldsymbol{\mu}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}. \tag{1.3.7}$$

The left hand formula lends itself to an interpretation of $\hat{\beta}_{j\cdot\mathrm{M}}$ in terms of the well-known leverage plot which shows Y plotted against $\mathbf{X}_{j\cdot\mathrm{M}}$ and the line with slope $\hat{\beta}_{j\cdot\mathrm{M}}$. This plot is valid without first-order correctness assumption.

A third interpretation can be derived from the second: For notational reasons let $\mathbf{x} = (x_i)_{i=1\ldots n}$ be any adjusted predictor $\mathbf{X}_{j\cdot\mathrm{M}}$, so that $\hat{\beta} = \mathbf{x}^T\mathbf{Y}/\|\mathbf{x}\|^2$ and $\beta = \mathbf{x}^T\boldsymbol{\mu}/\|\mathbf{x}\|^2$ are the corresponding $\hat{\beta}_{j\cdot\mathrm{M}}$ and $\beta_{j\cdot\mathrm{M}}$. Introduce case-wise slopes through the origin, both as estimates $\hat{\beta}_{(i)} = Y_i/x_i$ and as parameters $\beta_{(i)} = \mu_i/x_i$, as well as case-wise weights $w_{(i)} = x_i^2/\sum_{i'=1\ldots n} x_{i'}^2$. Equations (1.3.7) are then equivalent to the following:

$$\hat{\beta} \;=\; \sum_i w_{(i)}\hat{\beta}_{(i)}, \qquad \beta \;=\; \sum_i w_{(i)}\beta_{(i)}.$$

Hence regression coefficients are weighted averages of case-wise slopes, and this interpretation holds without first-order assumptions.

1.4 Universally Valid Post-Selection Confidence Intervals

1.4.1 Test Statistics with One Error Estimate for All Submodels

After defining $\boldsymbol{\beta}_{\mathrm{M}}$ as the target of the estimate $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$, we consider inference for it in terms of test statistics. Following Section 1.2.3 we require a normal homoscedastic model for $\mathbf{Y}$, but we leave its mean $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$ unspecified: $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. We then have equivalently

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} \sim \mathcal{N}(\boldsymbol{\beta}_{\mathrm{M}}, \sigma^2(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}) \quad \text{and} \quad \hat{\beta}_{j\cdot\mathrm{M}} \sim \mathcal{N}(\beta_{j\cdot\mathrm{M}}, \sigma^2/\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2).$$

Again following Section 1.2.3 we assume the availability of a valid estimate $\hat{\sigma}^2$ of σ^2 that is independent of all estimates $\hat{\beta}_{j\cdot\mathrm{M}}$, and we further assume $\hat{\sigma}^2 \sim \sigma^2\chi_r^2/r$ for r degrees of freedom. If the full model is assumed correct, $n > p$ and $\hat{\sigma}^2 = \hat{\sigma}_F^2$, then $r = n-p$. In the limit $r \to \infty$ we obtain $\hat{\sigma} = \sigma$, the case of known σ, which will be used starting with Section 1.6.

Let $t_{j\cdot\mathrm{M}}$ denote a t-ratio for $\beta_{j\cdot\mathrm{M}}$ that uses $\hat{\sigma}$ irrespective of the submodel M:

$$t_{j\cdot\mathrm{M}} \triangleq \frac{\hat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}}}{((\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1})_{jj}^{\frac{1}{2}}\ \hat{\sigma}} = \frac{\hat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}}}{\hat{\sigma}/\|\mathbf{X}_{j\cdot\mathrm{M}}\|} = \frac{(\mathbf{Y} - \boldsymbol{\mu})^T\mathbf{X}_{j\cdot\mathrm{M}}}{\hat{\sigma}\|\mathbf{X}_{j\cdot\mathrm{M}}\|}. \tag{1.4.1}$$

[According to full model indexing, $(...)_{jj}$ refers to the diagonal element corresponding to $\mathbf{X}_j$.] The quantity $t_{j\cdot\mathrm{M}} = t_{j\cdot\mathrm{M}}(\mathbf{Y})$ has a central t-distribution with r degrees of freedom. Essential is that the standard error estimate in the denominator of (1.4.1) does *not* involve the MSE $\hat{\sigma}_{\mathrm{M}}$ from the submodel M, for two reasons:

- We do not assume that the submodel M is first-order correct; therefore each MSE $\hat{\sigma}_{\mathrm{M}}^2$ could have a distribution that is a multiple of a non-central χ^2 distribution with unknown non-centrality parameter.
- More disconcertingly, the MSE would be the result of selection: $\hat{\sigma}_{\hat{\mathrm{M}}}^2$. Not much of real use is known about its distribution (see, for example, Brown 1967 and Olshen 1973).

These problems are avoided by using one valid estimate $\hat{\sigma}^2$ that is independent of all submodels.

With this choice of $\hat{\sigma}$, a marginal $1-\alpha$ confidence interval for $\beta_{j\cdot \mathrm{M}}$ is

$$\begin{aligned} \mathrm{CI}_{j\cdot \mathrm{M}}(K) &\triangleq \left[\hat{\beta}_{j\cdot \mathrm{M}} \pm K \left[(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1}\right]_{jj}^{\frac{1}{2}} \hat{\sigma}\right] \\ &= \left[\hat{\beta}_{j\cdot \mathrm{M}} \pm K\,\hat{\sigma}/\|\mathbf{X}_{j\cdot \mathrm{M}}\|\right]. \end{aligned} \tag{1.4.2}$$

where $K = t_{r,1-\alpha/2}$ is the $1-\alpha/2$ quantile of a t-distribution with r degrees of freedom. This interval is valid, that is,

$$\mathbf{P}[\beta_{j\cdot \mathrm{M}} \in \mathrm{CI}_{j\cdot \mathrm{M}}(K)] \;\geq\; 1-\alpha,$$

under the assumption that the submodel M is chosen independently of the response $\mathbf{Y}$.

1.4.2 Model Selection and Its Implications for Parameters

In practice, the model M tends to be the result of some form of model selection that makes use of the stochastic component of the data, which is the response vector $\mathbf{Y}$ in the present context (Section 1.2.3). This model should therefore be expressed as $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$. In general we allow a model selection procedure to be any (measurable) map

$$\hat{\mathrm{M}}: \quad \mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y}), \qquad \mathbb{R}^n \to \mathcal{M}_{\mathrm{all}}, \tag{1.4.3}$$

where

$$\mathcal{M}_{\text{all}} \triangleq \{\mathrm{M} \,|\, \mathrm{M} \subset \{1, 2, ..., p\},\ \operatorname{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\,\} \qquad (1.4.4)$$

is the set of all full-rank submodels. Thus $\hat{\mathrm{M}}$ divides $\mathbb{R}^n$ into as many as $|\mathcal{M}|$ different regions with shared outcome of model selection.

Data dependence of the selected model $\hat{\mathrm{M}}$ has strong consequences:

- Most fundamentally, the selected model $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$ is now random. Whether the model has been selected by an algorithm or by human choice, if the response $\mathbf{Y}$ has been involved in the selection, the resulting model is a random object because it could have been different for a different realization of the random vector $\mathbf{Y}$.

- Associated with the random model $\hat{\mathrm{M}}(\mathbf{Y})$ is the parameter vector of coefficients $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, which is now randomly chosen also:

  (1) It has a random dimension, $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{m(\mathbf{Y})}$ for $m(\mathbf{Y}) = |\hat{\mathrm{M}}(\mathbf{Y})|$.

  (2) For any fixed j, it may or may not be the case that $j \in \hat{\mathrm{M}}(\mathbf{Y})$.

  (3) Conditional on $j \in \hat{\mathrm{M}}(\mathbf{Y})$, the parameter $\beta_{j\cdot\hat{\mathrm{M}}(\mathbf{Y})}$ changes randomly as the adjustor covariates in $\hat{\mathrm{M}}(\mathbf{Y})$ vary randomly.

Thus the set of parameters for which inference is sought is random also.

1.4.3 Valid Post-Selection Confidence Intervals

Unless a predictor is forced to be in the selected model, it is not meaningful to ask for marginal probability guarantees for $\mathrm{CI}_{j\cdot\hat{\mathrm{M}}}$ for a fixed j because the guarantee requires $j \in \hat{\mathrm{M}}(\mathbf{Y})$ whereas the probability $\mathbf{P}[j \in \hat{\mathrm{M}}(\mathbf{Y})]$ all by itself can be easily less than $1-\alpha$ even for a strong predictor. One may therefore be tempted to look for guarantees in terms of conditional probabilities given $j \in \hat{\mathrm{M}}$, but little is known about such events and the associated conditional distribution of $|t_{j\cdot\mathrm{M}}|$ for common selection methods. However, a solution in terms of marginal rather than conditional probability can be found by binding j with a quantifier and requiring a simultaneous guarantee in terms of $\mathbf{P}[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ \forall j \in \hat{\mathrm{M}}]$. For this mathematically well-defined probability there exists in principle a confidence guarantee through suitable choice of the constant K such that

$$\mathbf{P}\left[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ \forall j \in \hat{\mathrm{M}}\right] \;\geq\; 1-\alpha. \tag{1.4.5}$$

Thus the logical impossibility of a marginal guarantee for any particular $j \in \hat{\mathrm{M}}$ implies that only a simultaneous guarantee for all $j \in \hat{\mathrm{M}}$ can be given.

1.4.4 Universal Validity for all Selection Procedures

The difficulty with the guarantee (1.4.5) is that the constant would be specific to the model selection procedure $\hat{\mathrm{M}}$: $K = K(\hat{\mathrm{M}})$. Finding procedure-specific constants may be a challenge, and this is not what we attempt to do in this article. Rather, the "PoSI" procedure proposed here produces a constant K that provides universally valid post-selection inference *for all model selection procedures* $\hat{\mathrm{M}}$:

$$\mathbf{P}\left[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ \forall j \in \hat{\mathrm{M}}\right] \;\geq\; 1-\alpha \quad \forall\, \hat{\mathrm{M}}. \tag{1.4.6}$$

Universal validity irrespective of the model selection procedure $\hat{\mathrm{M}}$ is a strong property that raises questions of whether the approach is too conservative. There are, however, some arguments in its favor:

(1) Universal validity may be desirable or even essential for applications in which the model selection procedure is not specified in advance or for which the analysis involves some ad hoc elements that cannot be accurately pre-specified. Even so, we should think of the actually chosen model as part of a "procedure" $\mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y})$, and though the ad hoc steps are not specified for $\mathbf{Y}$ other than the observed one, this is not a problem because our protection is irrespective of what a specification might have been. This view also allows data analysts to change their minds, to improvise and informally decide in favor of a model other than that produced by a formal selection procedure, or to experiment with multiple selection procedures.

(2) There exists a model selection procedure that requires the full strength of universally valid PoSI, and this procedure may not be entirely unrealistic as an approximation to some types of data analytic activities: "significance hunting", that is, selecting that model which contains the statistically most significant coefficient; see Section 1.4.8.

(3) There is a general question about the wisdom of proposing ever tighter confidence and retention intervals for practical use when in fact these intervals are valid only under tightly controlled conditions. It might be reasonable to suppose that much applied work involves more data peeking than is reported in published articles. With inference that is universally valid after any model selection procedure we have a way to establish which rejections are safe, irrespective of unreported data

peeking as part of selecting a model.

1.4.5 Restricted Model Selection

The concerns over PoSI's conservative nature can be alleviated somewhat by introducing a degree of flexibility to the PoSI problem with regard to the universe of models being searched. Such flexibility is additionally called for from a practical point of view because it is not true that *all* submodels in $\mathcal{M}_{\text{all}}$ (1.4.4) are being searched all the time. Rather, in many applications the search is limited in a way that can be specified a priori, without involvement of $\mathbf{Y}$. For example, a predictor of interest may be forced into the submodels, or there may be a restriction on the size of the submodels. Indeed, if p is large, a restriction to a manageable set of submodels is a computational necessity. In much of what follows we can allow the universe $\mathcal{M}$ of submodels to be an (almost) arbitrary but pre-specified non-empty subset of $\mathcal{M}_{\text{all}}$; the only requirement is that every predictor is used in at least one model:

$$\bigcup_{\mathrm{M}\in\mathcal{M}} \mathrm{M} \;=\; \{1, 2, ..., p\}. \qquad (1.4.7)$$

Because we allow only non-singular submodels (see (1.4.4)) we have $|\mathrm{M}| \le d \; \forall \mathrm{M}\in\mathcal{M}$, where as always $d = \mathrm{rank}(\mathbf{X})$. — Selection procedures are now maps

$$\hat{\mathrm{M}} : \; \mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y}), \qquad \mathbb{R}^n \to \mathcal{M}. \qquad (1.4.8)$$

The following are examples of model universes with practical relevance (see also Leeb and Pötscher (2008a), Section 1.1, Example 1).

(1) Submodels that contain the first p' predictors ($1 \leq p' \leq p$):

$\mathcal{M}_1 = \{\mathrm{M} \in \mathcal{M}_{\text{all}} \,|\, \{1, 2, ..., p'\} \subset \mathrm{M}\}$.

Classical: $|\mathcal{M}_1| = 2^{p-p'}$. Example: forcing an intercept into all models.

(2) Submodels of size m' or less ("sparsity option"):

$\mathcal{M}_2 = \{\mathrm{M} \in \mathcal{M}_{\text{all}} \,|\, |\mathrm{M}| \leq m'\}$. Classical: $|\mathcal{M}_2| = \binom{p}{1} + ... + \binom{p}{m'}$.

(3) Submodels with fewer than m' predictors dropped from the full model:

$\mathcal{M}_3 = \{\mathrm{M} \in \mathcal{M}_{\text{all}} \,|\, |\mathrm{M}| > p - m'\}$. Classical: $|\mathcal{M}_3| = |\mathcal{M}_2|$.

(4) Nested models: $\mathcal{M}_4 = \{\{1, ..., j\} \,|\, j \in \{1, ..., p\}\}$. $|\mathcal{M}_4| = p$.

Example: selecting the degree up to $p-1$ in a polynomial regression.

(5) Models dictated by an ANOVA hierarchy of main effects and interactions in a factorial design.

This list is just an indication of possibilities. In general, the smaller the set $\tilde{\mathcal{M}} = \{(j, \mathrm{M}) \,|\, j \in \mathrm{M} \in \mathcal{M}\}$ is, the less conservative the PoSI approach is, and the more computationally manageable the problem becomes. With sufficiently strong restrictions, in particular using the sparsity option (2) and assuming the availability of an independent valid estimate $\hat{\sigma}$, it is possible to apply PoSI in certain non-classical $p > n$ situations.

Further reduction of the PoSI problem is possible by pre-screening adjusted predictors *without the response* $\mathbf{Y}$. In a fixed-design regression, any variable selection procedure that does *not* involve $\mathbf{Y}$ does *not* invalidate statistical inference. For example, one may decide not to seek inference for predictors in submodels that impart a "Variance Inflation Factor" (*VIF*) above a user-chosen threshold: $\mathit{VIF}_{j\cdot\mathrm{M}} = \|\mathbf{X}_j\|^2/\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2$ if $\mathbf{X}_j$ is centered, hence does not make use of $\mathbf{Y}$, and elimination according to $\mathit{VIF}_{j\cdot\mathrm{M}} > c$ does not invalidate inference.

1.4.6 Reduction of Universally Valid Post-Selection Inference to Simultaneous Inference

We show that universally valid post-selection inference (1.4.6) follows from simultaneous inference in the form of family-wise error control for all parameters in all submodels. The argument depends on the following lemma that may fall into the category of the "trivial but not immediately obvious".

Lemma 1.4.1. *("Significant Triviality Bound") For any model selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$*, the following inequality holds for all* $\mathbf{Y} \in \mathbb{R}^n$*:*

$$\max_{j \in \hat{\mathrm{M}}(\mathbf{Y})} |t_{j\cdot\hat{\mathrm{M}}(\mathbf{Y})}(\mathbf{Y})| \;\leq\; \max_{\mathrm{M}\in\mathcal{M}} \max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}(\mathbf{Y})|$$

Proof: This is a special case of the triviality $f(\hat{\mathrm{M}}(\mathbf{Y})) \leq \max_{\mathrm{M}} f(\mathrm{M})$, where $f(\mathrm{M}) = \max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}(\mathbf{Y})|$. □

For a model selection procedure $\hat{\mathrm{M}}$ that attains the right hand bound of the lemma, see Section 1.4.8.

Theorem 1.4.1. *Let K satisfy*

$$\mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}}\max_{j\in\mathrm{M}}|t_{j\cdot\mathrm{M}}|\leq K\right] \;\geq\; 1-\alpha. \tag{1.4.9}$$

Then the following holds for all model selection procedures $\hat{\mathrm{M}}:\mathbb{R}^n\to\mathcal{M}$:

$$\mathbf{P}\left[\max_{j\in\hat{\mathrm{M}}}|t_{j\cdot\hat{\mathrm{M}}}|\leq K\right] \;\geq\; 1-\alpha. \tag{1.4.10}$$

Proof: This follows immediately from Lemma 1.4.1. □

Although mathematically trivial we give the above the status of a theorem as it is the central statement of the reduction of universal post-selection inference to simultaneous inference.

Let K be the minimal constant satisfying (1.4.9). By definition K does not depend on the selection procedure $\hat{\mathrm{M}}$, but it does depend on the full predictor matrix $\mathbf{X}$, the set of submodels $\mathcal{M}$, the required coverage $1-\alpha$, and the degrees of freedom r in $\hat{\sigma}$. We will ignore the dependence on $\mathcal{M}$ if it is understood that $\mathcal{M}=\mathcal{M}_{\mathrm{all}}$ and we will variously write

$$K=K(\mathbf{X},\mathcal{M},\alpha,r),\quad K(\mathbf{X}),\quad K(\mathbf{X},p),\quad K(\mathbf{X},\alpha,p,r), \tag{1.4.11}$$

the last two being useful in the classical case ($d=p\leq n$) for asymptotics as $p\to\infty$. We call $K(\mathbf{X})$ the "PoSI constant", and for M and $j\in\mathrm{M}$ we call $\mathrm{CI}_{j\cdot\mathrm{M}}(K(\mathbf{X}))$ the "PoSI simultaneous confidence interval" or simply "PoSI CI". From (1.4.10) follows the desired coverage guarantee:

Theorem 1.4.2. *"Simultaneous Post-Selection Confidence Guarantees" hold for any model selection procedure* $\hat{\mathrm{M}}\colon \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}[\, \beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K(\mathbf{X},\alpha)) \;\forall j \in \hat{\mathrm{M}}\,] \;\geq\; 1-\alpha.$$

Simultaneous inference provides strong family-wise error control, which in turn translates to strong error control following model selection.

Theorem 1.4.3. *"Strong Post-Selection Error Control" holds for any model selection procedure* $\hat{\mathrm{M}}\colon \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}[\, \forall j \in \hat{\mathrm{M}} : |t^{(0)}_{j\cdot\hat{\mathrm{M}}}| > K(\mathbf{X},\alpha) \;\Rightarrow\; \beta_{j\cdot\hat{\mathrm{M}}} \neq 0\,] \;\geq\; 1-\alpha,$$

where $t^{(0)}_{j\cdot\mathrm{M}}$ *is the t-statistic for the null hypothesis* $\beta_{j\cdot\mathrm{M}} = 0$*.*

The proof is in the Appendix. The theorem states that, with probability $1-\alpha$, in a selected model *all* PoSI-significant rejections have detected true alternatives.

1.4.7 Scheffé Protection

Realizing the idea that the LS estimators in different submodels generally estimate different parameters, we generated a simultaneous inference problem involving up to $p\,2^{p-1}$ linear contrasts $\beta_{j\cdot\mathrm{M}}$. In view of the enormous number of linear combinations for which simultaneous inference is sought, one should wonder whether the problem is not best solved by Scheffé's method (1953; 1959) which provides simultaneous inference for *all* linear combinations. To accommodate rank-deficient $\mathbf{X}$, we cast Scheffé's result in terms of t-statistics for arbitrary non-zero $\mathbf{x} \in \text{span}(\mathbf{X})$:

$$t_{\mathbf{x}} \;\triangleq\; \frac{(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{x}}{\hat{\sigma} \|\mathbf{x}\|}. \tag{1.4.12}$$

The t-statistics in (1.4.1) are obtained for $\mathbf{x} = \mathbf{X}_{j \cdot \mathrm{M}}$. Scheffé's guarantee is

$$\mathbf{P}\left[\sup_{\mathbf{x} \in \text{span}(\mathbf{X})} |t_{\mathbf{x}}| \leq K_{\text{Sch}} \right] \;=\; 1 - \alpha, \tag{1.4.13}$$

where the Scheffé constant is

$$K_{\text{Sch}} \;=\; K_{\text{Sch}}(\alpha, d, r) \;=\; \sqrt{d \mathrm{F}_{d,r,1-\alpha}}. \tag{1.4.14}$$

It provides an upper bound for *all* PoSI constants:

Proposition 1.4.1$K(\mathbf{X}, \mathcal{M}, \alpha, r) \;\leq\; K_{\text{Sch}}(\alpha, d, r) \;\; \forall \mathbf{X},\, \mathcal{M},\, d{=}\text{rank}(\mathbf{X}).$

Thus parameter estimates $\hat{\beta}_{j \cdot \mathrm{M}}$ whose t-ratios exceed K_{Sch} in magnitude are universally safe from invalidation due to model selection. The universality of the Scheffé constant is a tip-off that it may be too loose for some predictor matrices $\mathbf{X}$, and obtaining the sharper constant $K(\mathbf{X})$ may be worth the effort. An indication is given by the following comparison as $r \to \infty$:

- For the Scheffé constant it holds $K_{\text{Sch}} \sim \sqrt{d}$.
- For orthogonal designs it holds $K_{\text{orth}} \sim \sqrt{2 \log d}$.

(For orthogonal designs see Section 1.5.6.) Thus the PoSI constant K_{orth} is much smaller than K_{Sch}. The big gap between the two suggests that the Scheffé constant may be too conservative at least in some cases. We will study the case of certain non-orthogonal designs for which the PoSI constant is $O(\sqrt{\log(d)})$ in Section 1.6.1. On the other hand, the PoSI constant can approach the order $O(\sqrt{d})$ of the Scheffé constant K_{Sch} as well, and we will study one such case in Section 1.6.2.

Even though in this article we will give asymptotic results for $d = p \to \infty$ and $r \to \infty$ only, we mention another kind of asymptotics whereby r is held constant while $d = p \to \infty$: In this case K_{Sch} is in the order of the product of $\sqrt{d}$ and the $1-\alpha$ quantile of the inverse-chi-square distribution with r degrees of freedom. In a similar way, the constant K_{orth} for orthogonal designs is in the order of the product of $\sqrt{2 \log d}$ and the 1–α quantile of the inverse-chi-square distribution with r degrees of freedom.

1.4.8 PoSI-Sharp Model Selection — "SPAR" and "SPAR1"

There exists a model selection procedure that requires the full protection of the simultaneous inference procedure (1.4.9). It is the "significance hunting" procedure that selects the model containing the most significant "effect":

$$\hat{\mathrm{M}}_{\text{SPAR}}(\mathbf{Y}) \triangleq \underset{\mathrm{M} \in \mathcal{M}}{\operatorname{argmax}} \max_{j \in \mathrm{M}} |t_{j \cdot \mathrm{M}}(\mathbf{Y})|.$$

We name this procedure "SPAR" for "*Single Predictor Adjusted Regression.*" It achieves equality with the "significant triviality bound" in Lemma 1.4.1 and is therefore the worst case procedure for the PoSI problem. In the selected submodel $\hat{\mathrm{M}}_{\mathrm{SPAR}}(\mathbf{Y})$ the less significant predictors matter only in so far as they boost the significance of the winning predictor by adjusting it accordingly. This procedure ignores the quality of the fit to $\mathbf{Y}$ provided by the model. While our present purpose is to point out the existence of a selection procedure that requires full PoSI protection, SPAR could be of practical interest when the analysis is centered on strength of "effects", not quality of model fit.

Practically of greater interest is a restricted version of SPAR whereby a predictor of interest is determined a priori and the search is for adjustment that optimizes this predictor's "effect". We name the resulting procedure "SPAR1". If the predictor of interest is $\mathbf{X}_p$, say, then the model universe is $\mathcal{M}_{\mathrm{SPAR1}} = \{\mathrm{M} \in \mathcal{M}_{\mathrm{all}} \,|\, p \in \mathrm{M}\,\}$ and the model selection procedure is

$$\hat{\mathrm{M}}_{\mathrm{SPAR1}}(\mathbf{Y}) \;\triangleq\; \underset{\mathrm{M}\in\mathcal{M}_{\mathrm{SPAR1}}}{\mathrm{argmax}}\; |t_{p\cdot\mathrm{M}}(\mathbf{Y})|.$$

Importantly, the SPAR1 guarantee that we seek is not for all coefficients in the models $\mathrm{M} \in \mathcal{M}_{\mathrm{SPAR1}}$ but only for the X_p-coefficient $\beta_{p\cdot\mathrm{M}}$:

$$\mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}_{\mathrm{SPAR1}}} |t_{p\cdot\mathrm{M}}| \le K_{\mathrm{SPAR1}}\right] \;\ge\; 1-\alpha,$$

where K_{SPAR1} is the minimal constant satisfying this condition. As $\mathcal{M}_{\mathrm{SPAR1}} \subset \mathcal{M}_{\mathrm{all}}$ and SPAR1 inference is for $j = p$ only, the unrestricted PoSI constant dominates the SPAR1 constant: $K(\mathbf{X}, \mathcal{M}_{\mathrm{all}}) \ge K_{\mathrm{SPAR1}}(\mathbf{X})$. Even so, we will construct in Section 1.6.2 an example where the SPAR1 constant increases at the Scheffé rate and is asymptotically more than 63% of K_{Sch}. This is the technical reason for introducing SPAR1.

1.4.9 PoSI P-Value Adjustment for Model Selection

Statistical inference for regression coefficients is more often carried out in terms of p-values than confidence intervals. The usual p-values are for null hypotheses $\beta_{j\cdot\mathrm{M}} = 0$, hence the test statistics are

$$t^{(0)}_{j\cdot\mathrm{M}} \;=\; \hat{\beta}_{j\cdot\mathrm{M}}/(\hat{\sigma}/\|\mathbf{X}_{j\cdot\mathrm{M}}\|), \qquad t^{(0)}_{\max} \;=\; \max_{\mathrm{M}\in\mathcal{M}} \max_{j\in\mathrm{M}} |t^{(0)}_{j\cdot\mathrm{M}}|.$$

To define marginal and adjusted p-values we introduce two c.d.f.s:

$$F_{j\cdot\mathrm{M}}(t) = \mathbf{P}[\,|t^{(0)}_{j\cdot\mathrm{M}}| < t\,], \qquad F_{\max}(t) \;=\; \mathbf{P}[t^{(0)}_{\max} < t]. \tag{1.4.15}$$

The former measures marginal null coverage of a two-sided retention interval $[-t, +t]$, while the latter measures simultaneous coverage of a retention cube $[-t, +t]^k$ where $k = |\{(j, \mathrm{M}) \,|\, j \in \mathrm{M} \in \mathcal{M}\}|$ is the number of tests performed, which can be as many as $p\,2^{p-1}$ in the classical case $d = p \leq n$ for $\mathcal{M} = \mathcal{M}_{\text{all}}$. Denoting by $t^{obs}_{j\cdot\mathrm{M}}$ and $t^{obs}_{\max}$ the observed values of $t^{(0)}_{j\cdot\mathrm{M}}$ and $t^{(0)}_{\max}$, respectively, the following p-values can be defined:

(1) Marginal: $\mathrm{pval}_{j\cdot\mathrm{M}} = 1 - F_{j\cdot\mathrm{M}}(\,|t^{obs}_{j\cdot\mathrm{M}}|\,)$

(2) Global adjusted: $\mathrm{pval}^{PoSI}_{j\cdot\mathrm{M}} = 1 - F_{\max}(t^{obs}_{\max})$

(3) Individual adjusted: $\mathrm{pval}^{PoSI}_{j\cdot\mathrm{M}} = 1 - F_{\max}(|t^{obs}_{j\cdot\mathrm{M}}|)$

Comments:

(1) The marginal p-value ignores the fact that k tests are being performed.

(2) The global adjusted p-value establishes whether at least the strongest "effect" is statistically significant, and it is therefore an overall test similar to, but more specific than, the overall F-test. Because the latter is derived from Scheffé protection, the global adjusted PoSI p-value is more powerful and still protects against any model selection in the model universe $\mathcal{M}$.

(3) The individual adjusted p-value adjusts each $|t_{j\cdot \mathrm{M}}|$ as if it were a max statistic, hence results in an over-adjustment for all but $t_{\max}$. A sharper method than this "one-step adjustment" would be a simulation-based "step-down" method, but the computational expense may be prohibitive and the gain in statistical efficiency may be small.

The adjusted p-values are recommended because they account universally for any model selection in the model universe $\mathcal{M}$.
[Note on terminology: "adjustment of a p-value for simultaneity" and "adjustment of a predictor for other predictors" are two concepts that share nothing except the partial homonym.]

1.5 The Structure of the PoSI Problem

1.5.1 Canonical Coordinates

We can reduce the dimensionality of the PoSI problem from $n \times p$ to $d \times p$, where $d = \text{rank}(X) \leq \min(n, p)$, by introducing Scheffé's canonical coordinates. This reduction is important both geometrically and computationally because the PoSI coverage problem really takes place in the column space of $\mathbf{X}$.

DEFINITION: *Let* $\mathbf{Q} = (\mathbf{q}_1, ..., \mathbf{q}_d) \in \mathbb{R}^{n \times d}$ *be any orthonormal basis of the column space of* $\mathbf{X}$. *Note that* $\hat{\mathbf{Y}} = \mathbf{Q}\mathbf{Q}^T\mathbf{Y}$ *is the orthogonal projection of* $\mathbf{Y}$ *onto the column space of* $\mathbf{X}$ *even if* $\mathbf{X}$ *is not of full rank. We call* $\tilde{\mathbf{X}} = \mathbf{Q}^T\mathbf{X} \in \mathbb{R}^{d \times p}$ *and* $\tilde{\mathbf{Y}} = \mathbf{Q}^T\hat{\mathbf{Y}} \in \mathbb{R}^d$ *canonical coordinates of* $\mathbf{X}$ *and* $\hat{\mathbf{Y}}$.

We extend the notation $\mathbf{X}_{\mathrm{M}}$ for extraction of subsets of columns to canonical coordinates $\tilde{\mathbf{X}}_{\mathrm{M}}$. Accordingly slopes obtained from canonical coordinates will be denoted by $\hat{\boldsymbol{\beta}}_{\mathrm{M}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = (\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{X}}_{\mathrm{M}})^{-1}\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{Y}}$ to distinguish them from the slopes obtained from the original data $\hat{\boldsymbol{\beta}}_{\mathrm{M}}(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^T\mathbf{Y}$, if only to state in the following proposition that they are identical.

Proposition 1.5.1. *Properties of canonical coordinates:*

(1) $\tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y}$.

(2) $\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{X}}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}}$ *and* $\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{Y}} = \mathbf{X}_{\mathrm{M}}^T\mathbf{Y}$.

(3) $\hat{\boldsymbol{\beta}}_{\mathrm{M}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \hat{\boldsymbol{\beta}}_{\mathrm{M}}(\mathbf{X}, \mathbf{Y})$ *for all submodels* M.

(4) $\tilde{\mathbf{Y}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \sigma^2\mathbf{I}_d)$, *where* $\tilde{\boldsymbol{\mu}} = \mathbf{Q}^T\boldsymbol{\mu}$.

(5) $\tilde{\mathbf{X}}_{j\cdot\mathrm{M}} = \mathbf{Q}^T\mathbf{X}_{j\cdot\mathrm{M}}$*, where* $j \in \mathrm{M}$ *and* $\tilde{\mathbf{X}}_{j\cdot\mathrm{M}} \in \mathbb{R}^d$ *is the residual vector of the regression of* $\tilde{\mathbf{X}}_j$ *onto the other columns of* $\tilde{\mathbf{X}}_\mathrm{M}$*.*

(6) $t_{j\cdot M} = (\hat{\beta}_{j\cdot M}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) - \beta_{j\cdot M})/(\hat{\sigma}/\|\tilde{\mathbf{X}}_{j\cdot\mathrm{M}}\|)$*.*

(7) In the classical case $d = p$*,* $\tilde{\mathbf{X}}$ *can be chosen to be an upper triangular or a symmetric matrix.*

The proofs of *(1)-(6)* are elementary. As for *(7)*, an upper triangular $\tilde{\mathbf{X}}$ can be obtained from a QR-decomposition based on a Gram-Schmidt procedure: $\mathbf{X} = \mathbf{QR}$, $\tilde{\mathbf{X}} = \mathbf{R}$. A symmetric $\tilde{\mathbf{X}}$ is obtained from a singular value decomposition: $\mathbf{X} = \mathbf{UDV}^T$, $\mathbf{Q} = \mathbf{UV}^T$, $\tilde{\mathbf{X}} = \mathbf{VDV}^T$.

Canonical coordinates allow us to analyze the PoSI coverage problem in $\mathbb{R}^d$. In what follows we will freely assume that all objects are rendered in canonical coordinates and write $\mathbf{X}$ and $\mathbf{Y}$ for $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, implying that the predictor matrix is of size $d \times p$ and the response is of size $d \times 1$.

1.5.2 PoSI Coefficient Vectors in Canonical Coordinates

The PoSI coverage problem (1.4.9) can be simplified as follows: Due to pivotality of t-statistics, the problem is invariant under translation of $\boldsymbol{\beta}$ and rescaling of σ, and hence it suffices to solve coverage problems for $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma = 1$. In canonical coordinates this implies $\mathbf{E}[\tilde{\mathbf{Y}}] = \mathbf{0}_d$, hence $\tilde{\mathbf{Y}} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. For this reason we write $\mathbf{Z}$ instead of $\tilde{\mathbf{Y}}$, so that $\mathbf{Z}/\hat{\sigma}$ has a d-dimensional t-distribution with r degrees of freedom and any linear combination $\mathbf{u}^T\mathbf{Z}/\hat{\sigma}$ with a unit vector $\mathbf{u}$ has a 1-dimensional t-distribution. Letting $\mathbf{X}_{j\cdot\mathrm{M}}$ be the adjusted predictors in canonical coordinates, the estimates (1.3.7) and their t-statistics (1.4.1) simplify to

$$\hat{\beta}_{j\cdot\mathrm{M}} \;=\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Z}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2} \;=\; \boldsymbol{l}_{j\cdot\mathrm{M}}^T\mathbf{Z}, \qquad t_{j\cdot\mathrm{M}} \;=\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Z}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|\hat{\sigma}} \;=\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}^T\mathbf{Z}/\hat{\sigma}, \tag{1.5.1}$$

where we took advantage of the fact that these are linear functions of $\mathbf{Z}$ and $\mathbf{Z}/\hat{\sigma}$, respectively, with "PoSI coefficient vectors" $\boldsymbol{l}_{j\cdot\mathrm{M}}$ and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ that equal $\mathbf{X}_{j\cdot\mathrm{M}}$ up to scale:

$$\boldsymbol{l}_{j\cdot\mathrm{M}} \;\triangleq\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}, \qquad \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \;\triangleq\; \frac{\boldsymbol{l}_{j\cdot\mathrm{M}}}{\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|} \;=\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|}. \tag{1.5.2}$$

As we now operate in canonical coordinates we have $\boldsymbol{l}_{j\cdot\mathrm{M}} \in \mathbb{R}^d$ and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in S^{d-1}$, where S^{d-1} is the unit sphere in $\mathbb{R}^d$. To complete the structural description of the PoSI problem we let

$$\mathcal{L}(\mathbf{X},\mathcal{M}) \;\triangleq\; \{\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \,|\, j \in \mathrm{M} \in \mathcal{M}\} \;\subset\; S^{d-1}, \tag{1.5.3}$$

If $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$ we omit the second argument and write $\mathcal{L}(\mathbf{X})$.

Proposition 1.5.2. *The PoSI problem* (1.4.9) *is equivalent to a d-dimensional coverage problem for linear functions of the multivariate t-vector* $\mathbf{Z}/\hat{\sigma}$*:*

$$\mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}}\max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \le K\right] \;=\; \mathbf{P}\left[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}(\mathbf{X},\mathcal{M})} |\bar{\boldsymbol{l}}^T\mathbf{Z}/\hat{\sigma}| \le K\right] \;\overset{(\geq)}{=}\; 1-\alpha. \tag{1.5.4}$$

1.5.3 Orthogonalities of PoSI Coefficient Vectors

The set $\mathcal{L}(\mathbf{X},\mathcal{M})$ of unit vectors $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ has intrinsically interesting geometric structure, which is the subject of this and the following subsections. The next proposition (proof in Appendix A.1.2) elaborates in so many ways the fact that $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ is essentially the predictor vector $\mathbf{X}_j$ orthogonalized with regard to the other predictors in the model M. In what follows vectors are always assumed in canonical coordinates and hence d-dimensional.

Proposition 1.5.3. *Orthogonalities in $\mathcal{L}(\mathbf{X}, \mathcal{M})$: The following statements hold assuming that the models referred to are in $\mathcal{M}$ (hence are of full rank).*

1. *Adjustment properties:*

   $$\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in \mathrm{span}\{\, \mathbf{X}_j \,|\, j \in \mathrm{M}\} \quad \textit{and} \quad \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \perp \mathbf{X}_{j'} \ \ \textit{for } j \neq j' \textit{ both} \in \mathrm{M}.$$

2. *The following vectors form an orthonormal "Gram-Schmidt" series:*

   $$\{\bar{\boldsymbol{l}}_{1\cdot\{1\}},\ \bar{\boldsymbol{l}}_{2\cdot\{1,2\}},\ \bar{\boldsymbol{l}}_{3\cdot\{1,2,3\}},\ \ldots,\ \bar{\boldsymbol{l}}_{d\cdot\{1,2,\ldots,d\}}\}$$

   Other series are obtained using $(j_1, j_2, ..., j_d)$ in place of $(1, 2, ..., d)$.

3. *Vectors $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ and $\bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}$ are orthogonal if $\mathrm{M} \subset \mathrm{M}'$, $j \in \mathrm{M}$ and $j' \in \mathrm{M}' \setminus \mathrm{M}$.*

4. *Classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$: Each vector $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ is orthogonal to $(p–1)\, 2^{p-2}$ vectors $\bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}$ (not all of which may be distinct).*

The cardinality of orthogonalities in the classical case and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$ is as follows: If the predictor vectors $\mathbf{X}_j$ have no orthogonal pairs among them, then $|\mathcal{L}(\mathbf{X})| = p\, 2^{p-1}$. If there exist orthogonal pairs, then $|\mathcal{L}(\mathbf{X})|$ is less. For example, if there exists exactly one orthogonal pair, then $|\mathcal{L}(\mathbf{X})| = (p-1)\, 2^{p-1}$. When $\mathbf{X}$ is a fully orthogonal design, then $|\mathcal{L}(\mathbf{X})| = p$.

1.5.4 The PoSI Process

An alternative way of looking at the PoSI problem is in terms of a stochastic process indexed by (j, M) for $j \in \mathrm{M}$. We mention this view because it is the basis of some software implementations used to solve simultaneous inference and coverage problems, even though in this case it does not result in a practicable approach. In the PoSI problem the obvious process is $\mathbf{W} = (t_{j \cdot \mathrm{M}})_{j \in \mathrm{M} \in \mathcal{M}}$, which is a t-process for finite degrees of freedom r in $\hat{\sigma}$ and a Gaussian process in the limit $r \to \infty$.

The covariance structure of $\mathbf{W}$ exists for $r > 2$ and is proportional (by a factor $r/(r-2)$) to the correlation matrix

$$\mathbf{\Sigma} \; = \; (\Sigma_{j \cdot \mathrm{M};\, j' \cdot \mathrm{M}'}), \qquad \Sigma_{j \cdot \mathrm{M};\, j' \cdot \mathrm{M}'} \; \triangleq \; \bar{\boldsymbol{l}}_{j \cdot \mathrm{M}}^T \bar{\boldsymbol{l}}_{j' \cdot \mathrm{M}'}. \tag{1.5.5}$$

The coverage problem (1.5.4) can be written as $\mathbf{P}[\|\mathbf{W}\|_\infty \le K] = 1 - \alpha$. Software that computes such coverages (for example, Genz et al. (2010)) allows users to specify a structure such as $\mathbf{\Sigma}$, intervals such as $[-K, +K]$ for the components, and error degrees of freedom r. In our experiments this approach worked in the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$ for up to $p = 7$, the limiting factor being the space requirement $p\, 2^{p-1} \times p\, 2^{p-1}$ for the matrix $\mathbf{\Sigma}$. By comparison the approach described in Buja et al. (2012) works for up to $p \approx 20$.

Proposition 1.5.3 above implies that there exist certain necessary orthogonalities in $\mathcal{L}(\mathbf{X}, \mathcal{M})$. In terms of the correlation structure $\mathbf{\Sigma}$, orthogonalities in $\mathcal{L}(\mathbf{X}, \mathcal{M})$ correspond to zero correlations in $\mathbf{\Sigma}$. Part *4.* of the proposition states that in the classical case and $\mathcal{M}{=}\mathcal{M}_{\mathrm{all}}$ each "row" of $\mathbf{\Sigma}$ has $(p–1)\, 2^{p-2}$ zeros out of $p\, 2^{p-1}$ entries, amounting to a fraction $(p–1)/(2p) \to 0.5$, implying that the overall fraction of zeros in $\mathbf{\Sigma}$ approaches half for increasing p. Thus $\mathbf{\Sigma}$, though not sparse, is rich in zeros. It can be much sparser in the presence of exact orthogonalities among the predictors.

1.5.5 The PoSI Polytope

Coverage problems can be framed geometrically in terms of probability coverage of polytopes in $\mathbb{R}^d$. For the PoSI problem the polytope with half-width K is defined by

$$\mathbf{\Pi}_K \;=\; \mathbf{\Pi}_K(\mathbf{X}, \mathcal{M}) \;\triangleq\; \{\mathbf{z} \in \mathbb{R}^d \,|\; |\bar{\boldsymbol{l}}^T \mathbf{z}| \le K, \;\forall\, \bar{\boldsymbol{l}} \in \mathcal{L}(\mathbf{X}, \mathcal{M})\,\}, \qquad (1.5.6)$$

henceforth called the "PoSI polytope". The PoSI coverage problem (1.5.4) is equivalent to calibrating K such that

$$\mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{\Pi}_K] \;=\; 1 - \alpha.$$

The simplest case of a PoSI polytope, for $d{=}p{=}2$, is illustrated in Figure 1.1. More general polytopes are obtained for arbitrary sets $\mathcal{L}$ of unit vectors, that is, subsets $\mathcal{L} \subset S^{d-1}$ of the unit sphere in $\mathbb{R}^d$. For the special case $\mathcal{L} = S^{d-1}$ the "polytope" is the "Scheffé ball" with coverage $\sqrt{d\mathrm{F}_{d,r}} \to \sqrt{\chi_d^2}$ as $r \to \infty$:

$$\mathbf{B}_K \;\triangleq\; \{\mathbf{z} \in \mathbb{R}^d \,|\; \|\mathbf{z}\| \le K\,\}, \qquad \mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{B}_K] \;=\; F_{\mathrm{F}_{d,r}}(K^2/d).$$

Many properties of the polytopes $\mathbf{\Pi}_K$ are not specific to PoSI because they hold for polytopes (1.5.6) generated by simultaneous inference problems for linear functions with arbitrary sets $\mathcal{L}$ of unit vectors. These polytopes ...

1. ... form scale families of geometrically similar bodies: $\mathbf{\Pi}_K = K\mathbf{\Pi}_1$.

2. ... are point symmetric about the origin: $\mathbf{\Pi}_K = -\mathbf{\Pi}_K$.

3. ... contain the Scheffé ball: $\mathbf{B}_K \subset \mathbf{\Pi}_K$.

4. ... are intersections of "slabs" of width $2K$:

$$\mathbf{\Pi}_K \;=\; \bigcap_{\bar{\boldsymbol{l}}\in\mathcal{L}} \{\mathbf{z} \in \mathbb{R}^d |\ |\mathbf{z}^T\bar{\boldsymbol{l}}\,| \;\le K\,\}.$$

5. ... have $2\,|\mathcal{L}|$ faces (assuming $\mathcal{L} \cap -\mathcal{L} = \emptyset$), and each face is tangent to the Scheffé ball $\mathbf{B}_K$ with tangency points $\pm K\bar{\boldsymbol{l}}$ ($\bar{\boldsymbol{l}} \in \mathcal{L}$).

Specific to PoSI are the orthogonalities described in Proposition 1.5.3.

1.5.6 PoSI Optimality of Orthogonal Designs

In orthogonal designs, adjustment has no effect: $\mathbf{X}_{j\cdot\mathrm{M}} = \mathbf{X}_j$ for all $j \in \mathrm{M}$, hence $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} = \mathbf{X}_j/\|\mathbf{X}_j\|$ and $\mathcal{L}(\mathbf{X},\mathcal{M}) = \{\mathbf{X}_1/\|\mathbf{X}_1\|, ..., \mathbf{X}_p/\|\mathbf{X}_p\|\}$. The polytope $\mathbf{\Pi}_K$ is therefore a hypercube. This simple observation implies an optimality property of orthogonal designs if the submodel universes $\mathcal{M}$ are sufficiently rich to force $\mathcal{L}(\mathbf{X},\mathcal{M})$ to contain an orthonormal basis of $\mathbb{R}^d$: The polytope generated by an orthonormal basis is a hypercube, hence the polytope $\mathbf{\Pi}_K(\mathbf{X},\mathcal{M})$ is contained in this hypercube; thus $\mathbf{\Pi}_K(\mathbf{X},\mathcal{M})$ has maximal extent iff it is equal to this hypercube, which is the case iff $\mathcal{L}(\mathbf{X},\mathcal{M})$ is this orthonormal basis and nothing more, that is, $\mathbf{X}$ is an orthogonal design. — A simple sufficient condition for $\mathcal{M}$ to grant the existence of

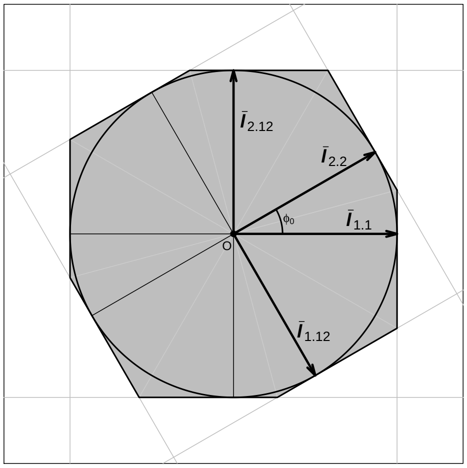

Figure 1.1: *The PoSI polytope* $\mathbf{\Pi}_{K=1}$ *tangent to the Scheffé disk (2-D ball)* $\mathbf{B}_{K=1}$ *for* $d = p = 2$*: The normalized raw predictor vectors are* $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$ *and* $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$*, and the normalized adjusted versions are* $\bar{\boldsymbol{l}}_{1\cdot\{1,2\}}$ *and* $\bar{\boldsymbol{l}}_{2\cdot\{1,2\}}$*. Shown in gray outline are the two squares (2-D cubes) generated by the o.n. bases* $(\bar{\boldsymbol{l}}_{1\cdot\{1\}}, \bar{\boldsymbol{l}}_{2\cdot\{1,2\}})$ *and* $(\bar{\boldsymbol{l}}_{2\cdot\{2\}}, \bar{\boldsymbol{l}}_{1\cdot\{1,2\}})$*, respectively. The PoSI polytope is the intersection of the two squares.*

an orthonormal basis in $\mathcal{L}(\mathbf{X}, \mathcal{M})$ is the existence of a maximal nested sequence of submodels such as $\{1\}$, $\{1,2\}$,...,$\{1,2,...,d\}$ in $\mathcal{M}$. It follows according to item 2. in Proposition 1.5.3 that there exists an orthonormal Gram-Schmidt basis in $\mathcal{L}(\mathbf{X}, \mathcal{M})$. We summarize:

Proposition 1.5.4. *Among predictor matrices with* rank($\mathbf{X}$)=d *and model universes $\mathcal{M}$ that contain at least one maximal nested sequence of submodels, orthogonal designs with $p = d$ columns yield*

- *the maximal coverage probability $\mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{\Pi}_K]$ for fixed K, and*
- *the minimal PoSI constant K satisfying $\mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{\Pi}_K] = 1 - \alpha$ for fixed α: $\inf_{\mathrm{rank}(X)=d} K(\mathbf{X}, \mathcal{M}, \alpha, r) = K_{\mathrm{orth}}(\alpha, d, r)$.*

The proposition holds not only for multivariate t-vectors and their Gaussian limits but for arbitrary spherically symmetric distributions. — Optimality of orthogonal designs translates to optimal asymptotic behavior of their constant $K(\mathbf{X}, \alpha)$ for large d:

Proposition 1.5.5. *Consider the Gaussian limit $r \to \infty$. For $\mathbf{X}$ and $\mathcal{M}$ as in Proposition 1.5.4, the asymptotic lower bound for the constant K as $d \to \infty$ is attained for orthogonal designs for which the asymptotic rate is*

$$\inf_{\mathrm{rank}(\mathbf{X})=d} K(\mathbf{X}, \mathcal{M}, \alpha) \;=\; K_{\mathrm{orth}}(d, \alpha) \;=\; \sqrt{2 \log d} + o(d).$$

The above facts show that the PoSI problem is bounded on one side by orthogonal designs: $\inf_{\mathrm{rank}(\mathbf{X})=d} K(\mathbf{X}, \alpha, r) = K_{\mathrm{orth}}(\alpha, d, r)$, for all α, d and r. On the other side, the Scheffé ball yields a loose upper bound: $\sup_{\text{rank}(\mathbf{X})=d,\mathcal{M}} K(\mathbf{X},\mathcal{M},\alpha,r) < K_{\text{Sch}}(\alpha,d,r)$. The question of how close to the Scheffé bound the PoSI upper bound $\sup_{\text{rank}(\mathbf{X})=d,\mathcal{M}} K(\mathbf{X},\mathcal{M},\alpha,r)$ can get for $r \to \infty$ will occupy us in Section 1.6.2. Unlike the infimum problem, the supremum problem does not appear to have a unique optimizing design $\mathbf{X}$ uniformly in α, d and r.

1.5.7 A Duality Property of PoSI Vectors

In the classical case $d\!=\!p$ and $\mathcal{M}\!=\!\mathcal{M}_{\text{all}}$ there exists a duality for PoSI vectors $\mathcal{L}(\mathbf{X})$ which we will use in Section 1.6.1 below but which is also of independent interest. We require some preliminaries: Letting $\mathrm{M}_F = \{1, 2, ..., p\}$ be the full model, we observe that the (unnormalized) PoSI vectors $\boldsymbol{l}_{j\cdot\mathrm{M}_F} = \mathbf{X}_{j\cdot\mathrm{M}_F}/\|\mathbf{X}_{j\cdot\mathrm{M}_F}\|^2$ form the rows of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ because $\hat{\boldsymbol{\beta}}_F = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. In a change of perspective, we interpret the transpose matrix

$$\mathbf{X}^* \;=\; \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

as a predictor matrix as well, to be called the “dual design” of $\mathbf{X}$. It is also of size $p \times p$ in canonical coordinates, and its columns are the PoSI vectors $\boldsymbol{l}_{j\cdot\mathrm{M}_F}$. It turns out that $\mathbf{X}^*$ and $\mathbf{X}$ pose identical PoSI problems if $\mathcal{M}\!=\!\mathcal{M}_{\text{all}}$:

Theorem 1.5.1. $\mathcal{L}(\mathbf{X}^*) = \mathcal{L}(\mathbf{X})$, $\;\boldsymbol{\Pi}_K(\mathbf{X}^*) = \boldsymbol{\Pi}_K(\mathbf{X})$, $\;K(\mathbf{X}^*) = K(\mathbf{X})$.

Recall that $\mathcal{L}(\mathbf{X})$ and $\mathcal{L}(\mathbf{X}^*)$ contain the normalized versions of the respective adjusted predictor vectors. The theorem follows from the following lemma which establishes the identities of vectors between $\mathcal{L}(\mathbf{X}^*)$ and $\mathcal{L}(\mathbf{X})$. We extend obvious notations from $\mathbf{X}$ to $\mathbf{X}^*$ as follows:

$$\mathbf{X}_j^* \;=\; \boldsymbol{l}_{j\cdot\{j\}}^* \;=\; \boldsymbol{l}_{j\cdot \mathrm{M}_F}\,.$$

Submodels for $\mathbf{X}^*$ will be denoted M^*, but they, too, will be given as subsets of $\{1, 2, ..., p\}$ which, however, refer to columns of $\mathbf{X}^*$. Finally, the normalized version of $\boldsymbol{l}_{j\cdot\mathrm{M}^*}^*$ will be written as $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}^*}^*$.

Lemma 1.5.1. *For two submodels* M *and* M^* *that satisfy* $\mathrm{M} \cap \mathrm{M}^* = \{j\}$ *and* $\mathrm{M} \cup \mathrm{M}^* = \mathrm{M}_F$*, we have*

$$\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}^*}^* \;=\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}\,, \qquad \|\boldsymbol{l}_{j\cdot\mathrm{M}^*}^*\|\,\|\boldsymbol{l}_{j\cdot\mathrm{M}}\| \;=\; 1$$

The proof is in Appendix A.1.3. The assertion about norms is really only needed to exclude collapse of $\boldsymbol{l}_{j\cdot\mathrm{M}^*}^*$ to zero.

A special case arises when the predictor matrix (in canonical coordinates) is chosen to be symmetric according to Proposition 1.5.1 (*7.*): if $\mathbf{X}^T = \mathbf{X}$, then $\mathbf{X}^* = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{X}^{-1}$, and hence:

Corollary 1.5.1. *If* $\mathbf{X}$ *is symmetric in canonical coordinates, then*

$$\mathcal{L}(\mathbf{X}^{-1}) = \mathcal{L}(\mathbf{X}), \quad \mathbf{\Pi}_K(\mathbf{X}^{-1}) = \mathbf{\Pi}_K(\mathbf{X}), \quad \textit{and} \quad K(\mathbf{X}^{-1}) = K(\mathbf{X})$$

1.6 Illustrative Examples and Asymptotic Results

In this section we consider examples in the classical case $d{=}p$ and $\mathcal{M}{=}\mathcal{M}_{\text{all}}$. Also, we work with the Gaussian limit $r \to \infty$ assuming $\sigma = 1$ is known.

1.6.1 Example 1: The PoSI Problem for Exchangeable Designs

In exchangeable designs $\mathbf{X}$ all pairs of predictor vectors enclose the same angle. In canonical coordinates a convenient parametrization of a family of symmetric exchangeable design is

$$\mathbf{X} \;=\; \mathbf{I}_p + a\mathbf{E}_{p\times p}, \tag{1.6.1}$$

where $-1/p < a < \infty$, and $\mathbf{E}$ is a matrix with all entries equal to 1. The range restriction on a assures that $\mathbf{X}$ is positive definite. Writing $\mathbf{X} = \mathbf{X}(a)$ when the parameter a matters, we will make use of the fact that

$$\mathbf{X}(a)^{-1} = \mathbf{X}(-a/(1+pa))$$

is also an exchangeable design. The function $c_p(a) = -a/(1+pa)$ maps the interval $(-1/p, \infty)$ onto itself, and it holds $c_p(0) = 0$, $c_p(a) \downarrow -1/p$ as $a \uparrow +\infty$, and vice versa. Exchangeable designs include orthogonal designs for $a = 0$, and they extend to two types of strict collinearities: for $a \uparrow \infty$ the predictor vectors collapse to a single dimension span($\mathbf{1}$), and for $a \downarrow -1/p$ they collapse to a subspace span$(\mathbf{1})^{\perp}$ of dimension $(p-1)$, where $\mathbf{1} = (1, 1, ..., 1)^T$.

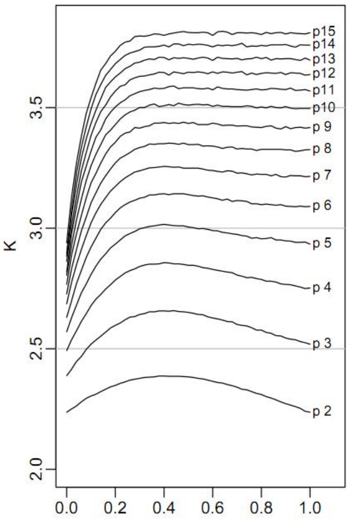

Figure 1.2: *The PoSI constant* $K(\mathbf{X}, \alpha = 0.05)$ *for exchangeable designs* $\mathbf{X} = \mathbf{I} + a\mathbf{E}$ *for* $a \in [0, \infty)$*. The horizontal axis shows* $a/(1+a)$*, hence the locations 0, 0.5 and 1.0 represent* $a = 0,\ 1,\ \infty$*, respectively. Surprisingly, the largest* $K(\mathbf{X})$ *is not attained at* $a = \infty$*, the point of perfect collinearity, at least not for dimensions up to* $p = 10$*. The graph is based on 10,000 random samples in p dimensions for* $p = 2, ..., 15$*.*

As non-orthogonality/collinearity drives the fracturing of the regression coefficients into model-dependent quantities $\beta_{j \cdot \mathrm{M}}$, it is of interest to analyze $K(\mathbf{X})$ as $\mathbf{X} = \mathbf{X}(a)$ moves from orthogonality at $a = 0$ toward either of the two types of collinearity. Here is what we find:

- Unguided intuition might suggest that the collapse to rank 1 calls for larger $K(\mathbf{X})$ than the collapse to rank $(p-1)$. This turns out to be entirely wrong: collapse to rank 1 or rank $p-1$ has identical effects on $K(\mathbf{X})$. The reason is duality (Section 1.5.7): for exchangeable designs, $\mathbf{X}(a)$ collapses to rank 1 iff

$\mathbf{X}(a)^* = \mathbf{X}(a)^{-1} = \mathbf{X}(-a/(1+pa))$ collapses to rank $p-1$, and vice versa, while $K(\mathbf{X}(a)^{-1}) = K(\mathbf{X}(a))$ according to Corollary 1.5.1.

- A more basic intuition would suggest that $K(\mathbf{X})$ increases as $\mathbf{X}$ moves away from orthogonality and approaches collinearity. Even this intuition is not fully born out: In Figure 1.2 we depict numerical approximations to $K(\mathbf{X}(a), \alpha = 0.05)$ for $a \in [0, \infty)$ ($a \in (-1/p, 0]$ being redundant due to duality). As the traces show, $K(\mathbf{X}(a))$ increases as $\mathbf{X}(a)$ moves away from orthogonality, up to a point, whereafter it descends as it approaches collinearity, at least for dimensions $p \leq 10$.

In summary, the dependence of $K(\mathbf{X})$ on the design $\mathbf{X}$ is not a simple matter. While duality provides some insights, there are no simple intuitions for inferring from $\mathbf{X}$ the geometry of the sets of unit vectors $\mathcal{L}(\mathbf{X})$, their polytopes $\mathbf{\Pi}_K$, their coverage probabilities and PoSI constants $K(\mathbf{X})$.

We next address the asymptotic behavior of $K = K(\mathbf{X}, \alpha, p)$ for increasing p. As noted in Section 1.4.7, there is a wide gap between orthogonal designs with $K_{\text{orth}} \sim \sqrt{2 \log p}$ and the full Scheffé protection with $K_{\text{Sch}} \sim \sqrt{p}$. The following theorem shows how exchangeable designs fall into this gap:

Theorem 1.6.1. *PoSI constants of exchangeable design matrices $\mathbf{X}(a)$ have the following limiting behavior:*

$$\lim_{p \to \infty} \sup_{a \in (-1/p, \infty)} \frac{K(\mathbf{X}(a), \alpha, p)}{\sqrt{2 \log p}} \; = \; 2.$$

The proof can be found in Appendix A.1.4. The theorem shows that for exchangeable designs the PoSI constant remains much closer to the orthogonal case than the Scheffé case. Thus, for this family of designs it is possible to improve on the Scheffé constant by a considerable margin.

The following detail of geometry for exchangeable designs has a bearing on the behavior of their PoSI constant: The angle between pairs of predictor vectors as a function of a is $\cos(\mathbf{X}_j(a), \mathbf{X}_{j'}(a)) = a(2+pa)/(pa^2+4a+2)$. In particular, as the vectors fall into the rank-$(p-1)$ collinearity at $a = -1/p$, the cosine becomes $-1/(2p-3)$, which converges to zero as $p \to \infty$. Thus, with increasing dimension, exchangeable designs approach orthogonal designs even at their most collinear extreme.

We finish with a geometric depiction of the limiting polytope $\mathbf{\Pi}_K$ as $\mathbf{X}(a)$ approaches either collinearity: For $a \uparrow \infty$, the predictor vectors fall into the 1-D subspace $\mathrm{span}(\mathbf{1})$, and for $a \downarrow -1/p$ they fall into $\mathrm{span}(\mathbf{1})^\perp$. With duality in mind and considering the permutation symmetry of exchangeable designs, it follows that the limiting polytope is a prismatic polytope with a p-simplex as its base in $\mathrm{span}(\mathbf{1})^\perp$. In Figure 1.3 we show this prism for $p = 3$. The unit vectors $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$, $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$ and $\bar{\boldsymbol{l}}_{3\cdot\{3\}} \sim \mathbf{X}_3$ form an equilateral triangle. The plane $\mathrm{span}(\mathbf{1})^\perp$ also contains the six once-adjusted vectors $\bar{\boldsymbol{l}}_{j\cdot\{j,j'\}}$ $(j' \neq j)$, while the three fully adjusted vectors $\bar{\boldsymbol{l}}_{j\cdot\{1,2,3\}}$ collapse to $\mathbf{1}/\sqrt{p}$, turning the polytope into a prism.

1.6.2 Example 2: Where $K(\mathbf{X})$ is close to the Scheffé Bound

We describe a situation in which the asymptotic upper bound for $K(\mathbf{X}, \alpha, p)$ is $O(\sqrt{p})$, hence close to the Scheffé constant K_{Sch} in terms of the asymptotic rate.

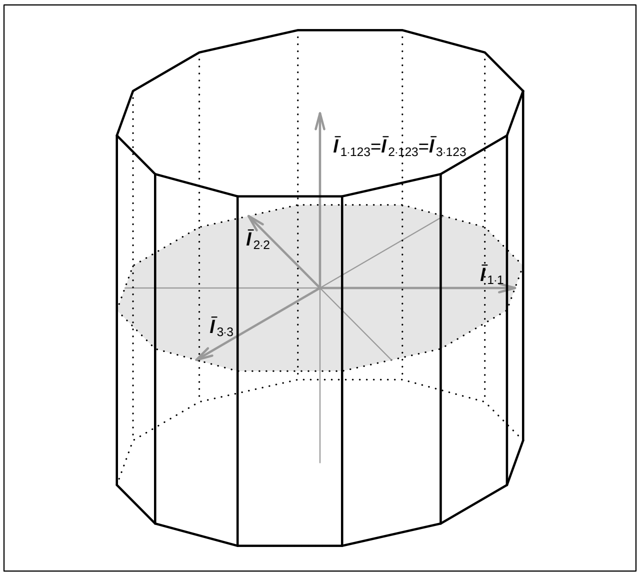

Figure 1.3: *Exchangeable Designs: The geometry of the limiting PoSI polytope for* $p = 3$ *as* $a \downarrow -1/p$ *or* $a \uparrow +\infty$ *in* (1.6.1).

We consider SPAR1 (Section 1.4.8) whereby a predictor of interest has been chosen, $\mathbf{X}_p$, say. The goal of model selection with SPAR1 is to "boost the effect" of $\mathbf{X}_p$ by adjusting it for optimally chosen predictors $\mathbf{X}_j$ $(j < p)$. The search is over the 2^{p-1} models that contain $\mathbf{X}_p$, but inference is sought only for the adjusted coefficient $\beta_{p\cdot\mathrm{M}}$.

The task is to construct a design for which simultaneous inference for all adjusted coefficients $\beta_{p\cdot\mathrm{M}}$ requires the constant $K_{\mathrm{SPAR1}}(\mathbf{X})$ to be in the order of $\sqrt{p}$. To this end consider the following upper triangular $p \times p$ design matrix in canonical coordinates:

$$\mathbf{X} = (\mathbf{e}_1, ..., \mathbf{e}_{p-1}, \mathbf{1}_p)\,, \tag{1.6.2}$$

where $\mathbf{e}_j$ are the canonical basis vectors, $(\mathbf{e}_j)_i = \delta_{ij}$, and $\mathbf{1}_p = (1, ..., 1)^T \in \mathbb{R}^p$. We have the following theorem:

Theorem 1.6.2. *The designs* (1.6.2) *have SPAR1 simultaneous* $1 - \alpha$ *confidence intervals for* $\mathbf{X}_p$ *of the form* $\left[\hat{\beta}_p \pm K_{\text{SPAR1}}(\mathbf{X})\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{pp}}\right]$ *where*

$$\lim_{p\to\infty} \frac{K_{\text{SPAR1}}(\mathbf{X})}{\sqrt{p}} \;=\; 0.6363....$$

A (partial) proof is in Appendix A.1.5 where we show the $\geq$ part. As always, we consider the case of "large r," that is, σ known; for small r the constant is larger. The theorem shows that even if we restrict consideration to a single predictor $\mathbf{X}_p$ and its adjustments, the constant K_{SPAR1} to reach valid simultaneous inference against all submodels containing that coefficient can be much greater than the $O(1)$ t-quantiles used in common practice. Also, since for the unrestricted PoSI constant $K(\mathbf{X})$ we have $K(\mathbf{X}) \geq K_{\text{SPAR1}}(\mathbf{X})$, the theorem shows that there exist predictor matrices for which the PoSI constants are of the asymptotic order of the Scheffé constants.

1.6.3 Bounding Away from Scheffé

We provide a rough asymptotic upper bound on all PoSI constants $K(\mathbf{X}, \mathcal{M}, \alpha, d)$. It is strictly smaller than the Scheffé constant but not by much. The bound, however, is loose because it is based on letting go of the rich structure of the sets $\mathcal{L}(\mathbf{X}, \mathcal{M})$ (Section 1.5.3) and only using their cardinality $|\mathcal{L}|$ $(= p\,2^{p-1}$ in the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\text{all}})$.

Theorem 1.6.3.*Denote by $\mathcal{L}_d$ arbitrary finite sets of d-dimensional unit vectors, $\mathcal{L}_d \subset S^{d-1}$, such that $|\mathcal{L}_d| \le a_d$ where $a_d^{1/d} \to a$ (> 0). Denote by $K(\mathcal{L}_d)$ the $(1-\alpha)$-quantile of $\sup_{\bar{\boldsymbol{l}} \in \mathcal{L}_d} |\bar{\boldsymbol{l}}^T \mathbf{Z}|$. Then the following describes an asymptotic worst-case bound for $K(\mathcal{L}_d)$ and its attainment:*

$$\lim_{d \to \infty} \sup_{|\mathcal{L}_d| \le a_d} \frac{K(\mathcal{L}_d)}{\sqrt{d}} \;=\; \left(1 - \frac{1}{a^2}\right)^{1/2}.$$

The proof of Theorem 1.6.3 (see the Appendix A.1.6) is an adaptation of Wyner's (1967) techniques for sphere packing and sphere covering. The worst-case bound ($\le$) is based on a surprisingly crude Bonferroni-style inequality for caps on spheres. Attainment of the bound ($\ge$) makes use of the artifice of picking the vectors $\bar{\boldsymbol{l}} \in \mathcal{L}$ randomly and independently. — Applying the theorem to PoSI sets $\mathcal{L} = \mathcal{L}(\mathbf{X}_{n \times p}, \mathcal{M}_{\text{all}})$ in the classical case $d = p$, we have $|\mathcal{L}| = p\,2^{p-1} = a_p$, hence $a_p^{1/p} \to 2$, so the theorem applies with $a = 2$:

Corollary 1.6.1.*In the classical case $d = p$ a universal asymptotic upper bound for the PoSI constant $K(\mathbf{X}_{n \times p}, \mathcal{M}_{\text{all}})$ is*

$$\lim_{p \to \infty} \sup_{\mathbf{X}_{n \times p}} \frac{K(\mathbf{X}_{n \times p}, \mathcal{M}_{\text{all}})}{\sqrt{p}} \;\le\; \frac{\sqrt{3}}{2} = 0.866\ldots.$$

The corollary shows that the asymptotic rate of the PoSI constant is strictly below that of the Scheffé constant, but possibly not by much. We do not know whether there exist designs for which the bound of the corollary is attained, but the theorem implies the bound is sharp for unstructured sets $\mathcal{L}$.

Chapter 2

The Split Samples Approach

2.1 Introduction

It is common practice to apply classical statistical inference to models that have been selected based on data. Typically, the same data is used for both selection and the inference after selection. Despite its prevalence, this practice is problematic because it ignores the fact that the inference is conditional on the model selection that is itself stochastic. The stochastic nature of the selection process affects and distorts sampling distributions of the post-selection parameter estimates, leading to invalid post-selection inference. The problems of post-selection inference have long been recognized and have been discussed recently by Leeb and Pötscher (2005; 2006b; 2008b) and Berk et al. (2010). Some conservative solutions to achieve valid post-selection inference are studied in Wang and Lagakos (2009) and Berk et al. (2012).

In this study, we propose a different approach to achieve valid post-selection inference for the problem of inference after variable selection in linear models. We suppose the response and the explanatory variables are generated from some general joint distribution where their relationship is not necessarily linear. Data is then gathered and analyzed, and a subset of the explanatory variables is chosen. These explanatory variables are then used to generate a linear submodel in approximation to the expected value of the response. We are interested in valid statistical inference after this process using the selected linear submodel. Our goal is to provide such valid inference that is universally valid for any variable selection procedure. Our methodology is based on two techniques, namely split samples and the bootstrap.

2.1.1 Split Samples

Split samples methodology generally involves dividing the observations randomly into two parts: one part for exploratory model building, a.k.a. the training set, and the other part for confirmatory statistical inference, a.k.a. holdout set. In a pioneering paper, Cox (1975) observed that split samples were more flexible, and perhaps more easily adapted to complex settings. Such data-splitting has been applied frequently and broadly in past literature on multiple testing problems. For example, in the statistical learning literature (e.g., Hastie et al. 2009, Chapter 1), the training sample is used to fit different models, and a holdout set is used for choosing the model with smallest prediction error; and in the causal inference literature (Heller et al. 2009; Zhang et al. 2011), a planning sample is used to design a study to be confirmed by the analysis sample.

Here, we utilize the split samples methodology in a slightly different way. We use a training sample only to seek a subset of predictors, and then perform both the estimation and inference on the holdout set.

As far as inference after selection in linear models is concerned, the main advantage of this data splitting technique is, roughly speaking, that it separates the data for exploratory analysis from the data for confirmatory analysis, thereby removing the contaminating effect of selection on inference. Such data-splitting is crucial for valid inference after model selection. Some qualitatively similar proposals are in Young and Karr (2011) who proposed the use of a holdout sample to test claims made from a modeling sample; in Hurvich and Tsai (1990) who studied coverage probabilities of post-selection confidence intervals via the split samples method; and in Wang and Lagakos (2009) who studied the potential of a version of this approach in linear models.

2.1.2 Random Design View

It is important to note that the bootstrap used in our split samples approach implicitly corresponds to the treatment of the design matrix predictors as random, where the training and holdout samples are distinct samples of predictors and responses from a larger population. In contrast, an approach for a fixed design would require identical designs in the two split samples.

As we will see, the random-design view changes the parameter space and the analysis for statistical inference from the conventional view. Further, there are other good reasons for taking the random-design point of view: (1) it is a proper view for observational data: each observation is an i.i.d. sample from a multivariate joint distribution; and (2) except for designed experiments, the fixed-design view for conventional inference is a theoretical artifice based on an ancillarity argument whose main benefit is facilitating inference calculation. With the bootstrap, the need for such fixed-design inference disappears. Furthermore, the fixed-design inference justification from ancillarity works only when the selected linear submodel is valid. We do not assume the response is linear. When nonlinearity is present with random predictors, the SE of slopes can be severely underestimated by the fixed design point of view (Will be explained in Section 2.1.3 and in further details in a later section).

We consider data that are n i.i.d. samples from random vectors $(X_1, \ldots, X_p, Y)$ with a non-degenerate $(p+1)$-dimensional joint distribution $\mathbf{P}_n\left(dx_1, \ldots, dx_p, dy\right)$, where Y is the response variable and X_1 through X_p are potential explanatory variables. It will be convenient to let $\vec{\mathbf{X}} = (1, X_1, \ldots, X_p)^T$, with a constant 1 appended. We also assume that the number of predictors, p, is fixed, while the joint distribution $\mathbf{P}_n$ can vary with n.

2.1.3 Nonlinearity and Bootstrap

In conventional inference for linear regression, it is assumed that (1) the relationship between response and explanatory variables is linear; and (2) the errors are homoscedastic; and (3) the underlying error distribution is independent Gaussian. Under these assumptions, the predictor values are conventionally treated as fixed preset constants even when they are random samples. The justification for such conditioning on the design matrix is that any predictor distribution is ancillary for the unknown parameters when the above assumptions are correct, hence conditioning on the design matrix produces valid frequentist inference for the desired parameters.

However, as will be discussed in more details later, when a model has been built based on the data, it is a fallacy to proceed as if the selected model were "true." Furthermore, the ancillarity argument should not be used, as the above assumptions of a linear model is never verifiable in practice. Indeed, a linear model should only be considered as an approximation instead. In fact, all linear models are approximations to generally nonlinear response surfaces, and the slopes are those of the best linear approximation. This view can be best described by a famous quote from G.E.P. Box: "all models are wrong, but some are useful." We take this view and consider regression models as linear approximations throughout this paper. In particular, in the context of inference after variable selection, we consider that each submodel has its own best linear approximation for the predictors it includes. This implies if a predictor is part of two different submodels, its slopes in the two submodels will be different in general. Bootstrap inference then frees us from being constrained by particular model assumptions such as linearity.

The rest of this paper is organized as follows. In Section 2.2 we introduce the population assumptions. In Section 2.3 we consider the bootstrap inference under the full linear model. We introduce our proposed split-sample procedure and states its asymptotic properties in Section 2.4. In Section 2.5 we consider a special case when the linear model is correct and the errors are homoscedastic.

2.2 Population

As stated before, we do not assume a linear relationship between the response and predictors in this paper. Instead, we rely only on our population assumption that the random vector $(X_1, \ldots, X_p, Y)$ has a non-degenerate joint distribution $\mathbf{P}_n$. Under $\mathbf{P}_n$, we can write out the conditional expectation of Y given $\vec{\mathbf{X}}$, i.e.,

$$\mathbf{E}\left[Y|\vec{\mathbf{X}}\right] = \mu_n\left(\vec{\mathbf{X}}\right) \tag{2.2.1}$$

for some $\mathbf{P}_n$ measurable function $\mu_n(\cdot)$.

In general, the "true response surface" or "true response function" $\mu_n(\vec{\mathbf{X}})$ need not to be linear in $\vec{\mathbf{X}}$ and can be any $\mathbf{P}_n$ measurable function. However, we will still be interested in inference based on a linear model approximation. Why? The linear approximation has long been considered an Occam's razor for its simplicity and effectiveness. Due to its advantages, it serves as the most common and fundamental method for modern data analysis. In our particular case, this linear approximation approach entails a linear combination of the components of $\vec{\mathbf{X}}$ which optimally approximates the general response function $\mu_n(\vec{\mathbf{X}})$. This set of coefficients for this linear combination are determined by the population distribution $\mathbf{P}_n\left(dx_1, \ldots, dx_p, dy\right)$.

Formally, suppose our loss is the mean squared error of prediction $\mathbf{E}\left[|Y - \mathbf{b}^T\vec{\mathbf{X}}|^2\right]$ for a $(p+1)$-dimensional vector $\mathbf{b}$. The best linear approximation of Y on $\vec{\mathbf{X}}$ under

$\mathbf{P}_n$ is a $(p+1)$-vector $\boldsymbol{\beta}_n(\mathbf{P}_n) = \boldsymbol{\beta}_n = (\beta_{n,0}, \beta_{n,1}, \ldots, \beta_{n,p})$ such that

$$\begin{aligned} \boldsymbol{\beta}_n(\mathbf{P}_n) &= \operatorname{argmin}_{\mathbf{b}} \mathbf{E}\big[|Y - \mathbf{b}^T\vec{\mathbf{X}}|^2\big] \\ &= \operatorname{argmin}_{\mathbf{b}} \mathbf{E}\left[|\mu_n(\vec{\mathbf{X}}) - (b_0 + b_1X_1 + \ldots + b_pX_p)|^2\right] \end{aligned} \tag{2.2.2}$$

Solving this optimization problem yields

$$\boldsymbol{\beta}_n(\mathbf{P}_n) = \mathbf{E}\big[\vec{\mathbf{X}}\vec{\mathbf{X}}^T\big]^{-1}\mathbf{E}\big[\vec{\mathbf{X}}Y\big]. \tag{2.2.3}$$

Equation (2.2.3) verifies that the vector $\boldsymbol{\beta}_n = \boldsymbol{\beta}_n(\mathbf{P}_n)$ is a population parameter. Furthermore, equation (2.2.3) provides the basis for the least squares (LS) estimator in linear models.

With the above notation, the nonlinearity of the linear approximation to $\mu_n(\vec{\mathbf{X}})$ is captured by

$$\eta_n = \eta_n(\vec{\mathbf{X}}) = \mu_n(\vec{\mathbf{X}}) - \boldsymbol{\beta}_n^T\vec{\mathbf{X}}. \tag{2.2.4}$$

We say the linear regression model is *first order correct* if $\eta_n(\vec{\mathbf{X}}) = 0$ $\mathbf{P}_n$-a.s., which in general may not be true. Such a nonzero $\eta_n(\vec{\mathbf{X}})$ can nullify the ancillarity of the predictor distribution, and distort the estimation of the variance of the LS estimators. This issue and its resolution by utilizing the bootstrap method are discussed in Buja's "Conspiracy Blurb" and also in a later section.

Let the "error" ϵ_n be the random variable

$$\epsilon_n = Y - \mu_n(\vec{\mathbf{X}}). \tag{2.2.5}$$

By construction $\mathbf{E}\left[\epsilon_n|\vec{\mathbf{X}}\right] = 0$ and $\mathbf{E}\left[\epsilon_n\eta_n(\vec{\mathbf{X}})\right] = 0$. The conditional variance of the error is then

$$\text{Var}\left[\epsilon_n|\vec{\mathbf{X}}\right] = \sigma_n^2(\vec{\mathbf{X}}). \tag{2.2.6}$$

If the errors are conditionally homoscedastic, then the above equation is simplified to

$$\text{Var}\left[\epsilon_n|\vec{\mathbf{X}}\right] = \sigma_n^2. \tag{2.2.7}$$

Under the above notation, the response Y can be decomposed as

$$Y = \boldsymbol{\beta}_n^T\vec{\mathbf{X}} + \eta_n(\vec{\mathbf{X}}) + \epsilon_n. \tag{2.2.8}$$

This decomposition when $p = 1$ is illustrated in Figure 1 below.

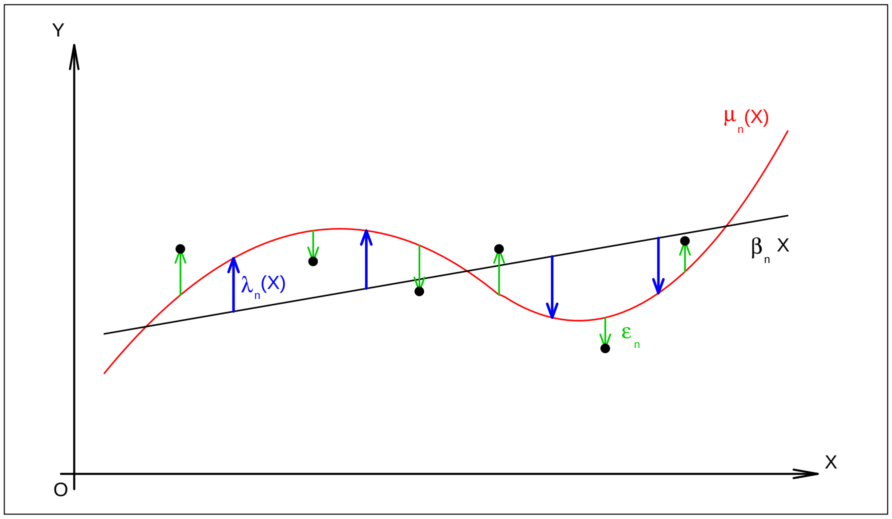

Figure 2.1: *The decomposition of Y when $p = 1$.*

We denote our n i.i.d. draws from $\mathbf{P}_n(dx_1, \dots, dx_p, dy)$ by $(X_{i,1}, \dots, X_{i,p}, Y_i)$ for $i = 1, 2, \dots, n$. From these, we form the responses $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, and column n-vectors $\mathbf{X}_j = (X_{1,j}, \dots, X_{n,j})^T$, which we collect into the $n \times (p+1)$ random predictor matrix $\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_p]$, appended by an intercept vector $\mathbf{1}$, i.e.,

$$\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \dots, \mathbf{X}_p] = \begin{pmatrix} \vec{\mathbf{X}}_1^T \\ \vdots \\ \vec{\mathbf{X}}_n^T \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{pmatrix} \tag{2.2.9}$$

where $\vec{\mathbf{X}}_i^T = (1, X_{i,1}, \dots, X_{i,p})$ for $i = 1, \dots, n$ are the transposed row vectors.

We finally collect the values $\mu_{n,i} = \mu_n(X_{i,1}, \dots, X_{i,p})$, $\eta_n(\vec{\mathbf{X}}_i) = \eta_n(X_{i,1}, \dots, X_{i,p})$, and $\epsilon_i = Y_i - \mu_{n,i}$ into random n-vectors denoted by $\boldsymbol{\mu}_n = (\mu_{n,1}, \dots, \mu_{n,n})^T$, $\boldsymbol{\eta}_n(\mathbf{X}) = (\eta_n(\vec{\mathbf{X}}_1), \dots, \eta_n(\vec{\mathbf{X}}_n))^T$, $\boldsymbol{\epsilon}_n = (\epsilon_1, \dots, \epsilon_n)^T$, and $\boldsymbol{\sigma}_n(\mathbf{X}) = \mathrm{diag}(\sigma_n^2(\vec{\mathbf{X}}_1), \dots, \sigma_n^2(\vec{\mathbf{X}}_n))$, respectively.

We shall make two remarks here:

1. If not otherwise stated, we consider a fixed p in this manuscript. However, some results which allow p to grow with n will be given.

2. In asymptotics, we consider $\boldsymbol{\beta}_n$ shrinking to 0 as $n \to \infty$ so that $\boldsymbol{\beta}_n$ is a function of n.

2.3 Valid Inference in the Full Model

2.3.1 Least Squares Estimates in the Full Model

The Least Squares estimate of $\boldsymbol{\beta}_n$ based on n observations, $\hat{\boldsymbol{\beta}}_n$, is given by

$$\hat{\boldsymbol{\beta}}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.3.1}$$

We shall first note that in general $\eta_n(\vec{\mathbf{X}})$ is not 0 a.s., and the LS estimate is biased:

$$\mathbf{E}\big[\hat{\boldsymbol{\beta}}_n\big] = \boldsymbol{\beta}_n + \mathbf{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}_n(\mathbf{X})\right]. \tag{2.3.2}$$

However, the LS estimate is asymptotically consistent, as seen in the following theorem which is a summary of the marginal properties of $\hat{\boldsymbol{\beta}}_n$.

Theorem 2.3.1. *1. $\hat{\boldsymbol{\beta}}_n$ is consistent, i.e.,*

$$\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n \xrightarrow{P} \mathbf{0}. \tag{2.3.3}$$

2. The variance of $\hat{\boldsymbol{\beta}}_n$ is given by

$$\mathrm{Var}\big[\hat{\boldsymbol{\beta}}_n\big] = \mathbf{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\sigma}_n^2(\mathbf{X})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right] + \mathrm{Var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}_n(\mathbf{X})\right]. \tag{2.3.4}$$

In particular, if the error is homoscedastic, then

$$\mathrm{Var}\big[\hat{\boldsymbol{\beta}}_n\big] = \sigma_n^2\mathbf{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right] + \mathrm{Var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}_n(\mathbf{X})\right]. \tag{2.3.5}$$

To compare the above actual properties of $\hat{\boldsymbol{\beta}}_n$ and the assumptions in the traditional inference, we also consider the conditional distribution of $\hat{\boldsymbol{\beta}}_n$ given $\mathbf{X}$. We shall illustrate their difference under the homoscedastic case. Note that the random projection or hat matrix generated by $\mathbf{W}$ is

$$\mathbf{H}_{\mathbf{W}} = \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T. \tag{2.3.6}$$

Let $\mathbf{X}_{\setminus j} = \left[\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_{j-1}, \mathbf{X}_{j+1}, \ldots, \mathbf{X}_p\right]$. Then $\mathbf{X}_j^{adj} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}_{\setminus j}})\mathbf{X}_j$ is the predictor $\mathbf{X}_j$ "adjusted for all other predictors". The estimate of the j-th coordinate of $\hat{\boldsymbol{\beta}}_n$, $\hat{\beta}_{n,j}$ can then be expressed as follows:

$$\hat{\beta}_{n,j} = \frac{\langle \mathbf{Y}, \mathbf{X}_j^{adj} \rangle}{\|\mathbf{X}_j^{adj}\|^2} \tag{2.3.7}$$

Note that also the coordinate-wise conditional expectation of $\hat{\boldsymbol{\beta}}_n$ is

$$\mathbf{E}\left[\hat{\beta}_{n,j}|\mathbf{X}\right] = \beta_{n,j} + \frac{\langle \mathbf{X}_j^{adj}, \boldsymbol{\eta}_n(\mathbf{X}) \rangle}{\|\mathbf{X}_j^{adj}\|^2} \tag{2.3.8}$$

and the coordinate-wise conditional variance of $\hat{\boldsymbol{\beta}}_n$ is

$$\mathrm{Var}\left[\hat{\beta}_{n,j}|\mathbf{X}\right] = \frac{\sigma_n^2}{\|\mathbf{X}_j^{adj}\|^2}. \tag{2.3.9}$$

Therefore, the marginal variance of $\hat{\beta}_{n,j}$ is given by

$$\mathrm{SE}_{marg}\left(\hat{\beta}_{n,j}\right)^2 = \mathrm{Var}\left[\hat{\beta}_{n,j}\right] = \mathbf{E}\left[\frac{\sigma_n^2}{\|\mathbf{X}_j^{adj}\|^2}\right] + \mathrm{Var}\left[\frac{\langle \mathbf{X}_j^{adj}, \boldsymbol{\eta}_n(\mathbf{X}) \rangle}{\|\mathbf{X}_j^{adj}\|^2}\right]. \tag{2.3.10}$$

This agrees with the diagonals in the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_n$ in (2.4.5).

Note that the conventional inference uses $\hat{\sigma_n}^2 = \frac{\mathbf{Y}^T(\mathbf{I}-\mathbf{H}_\mathbf{X})\mathbf{Y}}{n-p-1}$ as the estimate for σ_n^2 in (2.3.9), i.e., the conventional LS inference is based on

$$\begin{aligned}\mathrm{SE}_{conv}\left(\hat{\beta}_{n,j}|\mathbf{X}\right)^2 =& \frac{\mathbf{E}\left[\frac{1}{n-p-1}\left\|\mathbf{Y}^T\left(\mathbf{I}-\mathbf{H}_\mathbf{X}\right)\mathbf{Y}\right\|^2\middle|\mathbf{X}\right]}{\|\mathbf{X}_j^{adj}\|^2} \\ \simeq& \frac{\sigma_n^2 + \frac{n}{n-p-1}\mathbf{E}\left[\eta_n(\vec{\mathbf{X}})^2\right]}{\|\mathbf{X}_j^{adj}\|^2}\end{aligned} \tag{2.3.11}$$

which implies that

$$\mathbf{E}\left[\mathrm{SE}_{conv}\left(\hat{\beta}_{n,j}|\mathbf{X}\right)^2\right] = \mathbf{E}\left[\frac{\sigma_n^2}{\|\mathbf{X}_j^{adj}\|^2}\right] + \mathbf{E}\left[\frac{\left\|\left(\mathbf{I}-\mathbf{H}_\mathbf{X}^T\right)\boldsymbol{\eta}_n(\mathbf{X})\right\|^2}{\|\mathbf{X}_j^{adj}\|^2}\right]. \tag{2.3.12}$$

Therefore, if $\eta_n(\vec{\mathbf{X}})$ is not 0 a.s., then $\mathrm{SE}_{conv} \neq \mathrm{SE}_{marg}$. In practice, $\eta_n(\vec{\mathbf{X}}) = 0$ a.s. is often not an appropriate assumption to make (Boston Housing Data for example). Therefore, the conventional approach usually underestimate the true SE of $\hat{\beta}_{n,j}$, leading to less accurate inference. The problem can be severer if the errors are heteroscedastic.

2.3.2 Bootstrap Inference for LS Estimates in the Full Model

In this subsection we consider bootstrap confidence intervals under the full model. We consider the “pair-bootstrap” method, in which we generate a resample of size w, $\{(\vec{\mathbf{X}}_1^*, Y_1^*), (\vec{\mathbf{X}}_2^*, Y_2^*), \ldots, (\vec{\mathbf{X}}_w^*, Y_w^*)\}$, from the original data. The resample forms the bootstrap design matrix $\mathbf{X}^*$ and the bootstrap response $\mathbf{Y}^*$. The bootstrap LS estimate is then

$$\hat{\boldsymbol{\beta}}_n^* = \left(\mathbf{X}^{*T}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*T}\mathbf{Y}^*. \tag{2.3.13}$$

We use the bootstrap distribution $\mathcal{L}^*\left(\sqrt{w}\left(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n\right)\right)$ to approximate the law $\mathcal{L}\left(\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right)\right)$.

A strong result on bootstrap inference can be obtained as a corollary from the following theorem of Mammen (1993). This theorem considers the case when $w = n$ and convergence is under the Kolmogorov distance, d_∞, where $d_\infty(F, G) = \sup_x |F(x) - G(x)|$.

Theorem 2.3.2. *Consider the following data generation process:*

1. $(\vec{\mathbf{X}}_i, Y_i)$ *are i.i.d. with finite second moments* $\mathbf{E}[Y_i^2] < \infty$ *and* $\mathbf{E}[\|\vec{\mathbf{X}}_i\|^2] < \infty$.
2. *Let* $\boldsymbol{\beta}_n = \operatorname{argmin}_{\mathbf{b}}\mathbf{E}\left[(Y_i - \boldsymbol{\beta}_n^T\vec{\mathbf{X}}_i)^2\right]$ *and let* $\xi_{n,i} = Y_i - \boldsymbol{\beta}_n^T\vec{\mathbf{X}}_i$.

Assuming the following:

1. *The eigenvalues of* $\mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]$ *are bounded away from* 0 *and* ∞ *for all* n.
2. $0 < \inf_n \mathbf{E}[\xi_n^2] \le \sup_n \mathbf{E}[\xi_n^2] < \infty$.
3. *For some fixed constant* $b \ge 1/3$, $p^{1+b}/n \to 0$.
4. $$\sup_n \sup_{\|\mathbf{d}\|=1} \mathbf{E}\left[|\mathbf{d}^T\vec{\mathbf{X}}|^4(1 + \xi_n^2)\right] < \infty \tag{2.3.14}$$

   where B *is the smallest integer greater than or equal to* $2/b$.

5. *For* $\mathbf{c}_p \in \mathbb{R}^{p+1}$ *with* $\|\mathbf{c}_p\| = 1$,

$$\mathbf{E}\left[(\mathbf{c}_p^T\vec{\mathbf{X}})^2\xi_n^2 I\left[(\mathbf{c}_p^T\vec{\mathbf{X}})^2\xi_n^2 \geq \zeta n\right]\right] \to 0 \tag{2.3.15}$$

   for every fixed $\zeta > 0$.

Then

$$d_\infty\left(\mathcal{L}^*\left(\sqrt{n}\mathbf{c}_p^T\left(\hat{\boldsymbol{\beta}}_n^* - \hat{\boldsymbol{\beta}}_n\right)\right), \mathcal{L}\left(\sqrt{n}\mathbf{c}_p^T\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n\right)\right)\right) \xrightarrow{P} 0 \tag{2.3.16}$$

By (2.2.8), ξ_n in our case is

$$\xi_{n,i} = Y_i - \boldsymbol{\beta}_n^T\vec{\mathbf{X}}_i = \eta_n(\vec{\mathbf{X}}_i) + \epsilon_i. \tag{2.3.17}$$

Thus, we have the following corollary for fixed p.

Corollary 2.3.1. *Suppose the following conditions hold:*

1. *The number of predictors p is fixed.*
2. $(\vec{\mathbf{X}}_i, Y_i)$ *are i.i.d. with finite second moments* $\mathbf{E}[Y_i^2] < \infty$ *and* $\mathbf{E}[\|\vec{\mathbf{X}}_i\|^2] < \infty$.
3. *The eigenvalues of* $\mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]$ *are bounded away from* 0 *and* ∞ *for all* n.
4. $0 < \inf_n \mathbf{E}[(\eta_n(\vec{\mathbf{X}}) + \epsilon)^2] \leq \sup_n \mathbf{E}[(\eta_n(\vec{\mathbf{X}}) + \epsilon)^2] < \infty$.
5. $$\sup_n \mathbf{E}\left[\|\vec{\mathbf{X}}\|^4\big(1 + (\eta_n(\vec{\mathbf{X}}) + \epsilon)^4\big)\right] < \infty \tag{2.3.18}$$

Then

$$d_\infty\left(\mathcal{L}^*\left(\sqrt{n}\left(\hat{\beta}^*_{n,j}-\hat{\beta}_{n,j}\right)\right),\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}_{n,j}-\beta_{n,j}\right)\right)\right)\xrightarrow{P}0. \tag{2.3.19}$$

One then finds the appropriate quantiles of the bootstrap distribution to construct asymptotically valid confidence intervals.

We have a remark here: The convergence in distribution is for a particular coordinate. For a fixed p, coordinate-wise convergence implies convergence in the joint distribution. However, in general this is not true. On the other hand, the PoSI confidence interval (Berk et al. 2012) provides valid family wise error statements.

2.4 Valid Post-Selection Inference via Split Samples and Bootstrap

2.4.1 Least Squares Estimates in A Submodel

For a given submodel $\mathrm{M} \subset \{1,\ldots,p\}$ of cardinality m, we consider the data sampled from random vectors $(1, X_{j_1},\ldots,X_{j_m},Y)$, where $j_k \in \mathrm{M}$, $\forall k = 1,\ldots,m$. We denote $\vec{\mathbf{X}}_\mathrm{M} = \{1, X_{j_1},\ldots,X_{j_m}\}$, and define the slopes under this submodel, $\boldsymbol{\beta}_{n,\mathrm{M}}$, as the best linear approximation to Y from $\vec{\mathbf{X}}_\mathrm{M}$, i.e.,

$$\boldsymbol{\beta}_{n,\mathrm{M}} = \mathrm{argmin}_{\mathbf{b}}\mathbf{E}\left[\|Y-\mathbf{b}^T\vec{\mathbf{X}}_\mathrm{M}\|^2\right] = \mathbf{E}\left[\vec{\mathbf{X}}_\mathrm{M}\vec{\mathbf{X}}_\mathrm{M}^T\right]^{-1}\mathbf{E}\left[\vec{\mathbf{X}}_\mathrm{M}\mu(\vec{\mathbf{X}})\right]. \tag{2.4.1}$$

The least squares estimate under M is given by

$$\hat{\boldsymbol{\beta}}_{n,\mathrm{M}} = \left(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}}\right)^{-1} \mathbf{X}_{\mathrm{M}}^T \mathbf{Y} \tag{2.4.2}$$

If $\eta_n(\vec{\mathbf{X}})$ is not 0 a.s., then the LS estimate is biased:

$$\mathbf{E}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big] = \mathbf{E}\left[(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \boldsymbol{\mu}_n(\mathbf{X})\right]. \tag{2.4.3}$$

The consistency and variance of $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$ are summarized in the following theorem.

Theorem 2.4.1. *1. $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$ is consistent*

$$\hat{\boldsymbol{\beta}}_{n,\mathrm{M}} - \boldsymbol{\beta}_{n,\mathrm{M}} \overset{P}{\to} \mathbf{0}. \tag{2.4.4}$$

2. The variance of $\hat{\boldsymbol{\beta}}_n$ is given by

$$\begin{aligned}&\mathrm{Var}\big[\hat{\boldsymbol{\beta}}_n\big]\\ =&\mathbf{E}\left[(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \boldsymbol{\sigma}_n^2(\mathbf{X}) \mathbf{X}_{\mathrm{M}} (\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1}\right] + \mathrm{Var}\left[(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \boldsymbol{\mu}_n(\mathbf{X})\right].\end{aligned} \tag{2.4.5}$$

In particular, if the error is homoscedastic, then

$$\mathrm{Var}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big] = \sigma_n^2 \mathbf{E}\left[(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1}\right] + \mathrm{Var}\left[(\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1} \mathbf{X}_{\mathrm{M}}^T \boldsymbol{\mu}_n(\mathbf{X})\right]. \tag{2.4.6}$$

2.4.2 Valid Bootstrap Inference under A Submodel

To achieve valid post-selection inference, we first consider the valid inference under a given submodel M. Mammen's theorem (Theorem 2.3.2) shows that this valid inference can be obtained through bootstrap. In fact, similarly as in (2.3.17), if we define

$$\xi_{n,i,\mathrm{M}} = Y_i - \boldsymbol{\beta}_{n,\mathrm{M}}^T \vec{\mathbf{X}}_{i,\mathrm{M}} = \boldsymbol{\beta}_n^T \vec{\mathbf{X}}_i + \eta_n(\vec{\mathbf{X}}_i) + \epsilon_i - \boldsymbol{\beta}_{n,\mathrm{M}}^T \vec{\mathbf{X}}_{i,\mathrm{M}} \tag{2.4.7}$$

where $\vec{\mathbf{X}}_{i,\mathrm{M}}$ is the collection of the variables in M in the i-th observation, or

$$\xi_{n,\mathrm{M}} = Y - \boldsymbol{\beta}_{n,\mathrm{M}}^T \vec{\mathbf{X}}_{\mathrm{M}} = \boldsymbol{\beta}_n^T \vec{\mathbf{X}} + \eta_n(\vec{\mathbf{X}}) + \epsilon - \boldsymbol{\beta}_{n,\mathrm{M}}^T \vec{\mathbf{X}}_{\mathrm{M}}, \tag{2.4.8}$$

then a sufficient condition for the asymptotically valid bootstrap inference for fixed p is as follows:

Corollary 2.4.1. *Suppose the following conditions hold:*

1. *The number of predictors p is fixed.*

2. $(\vec{\mathbf{X}}_{i,\mathrm{M}}, Y_i)$ *are i.i.d. with finite second moments* $\mathbf{E}[Y_i^2] < \infty$ *and* $\mathbf{E}[\|\vec{\mathbf{X}}_{i,\mathrm{M}}\|^2] < \infty$.

3. *The eigenvalues of* $\mathbf{E}[\vec{\mathbf{X}}_{\mathrm{M}} \vec{\mathbf{X}}_{\mathrm{M}}^T]$ *are bounded away from* 0 *and* ∞ *for all* n.

4. $0 < \inf_n \mathbf{E}[\xi_{n,\mathrm{M}}^2] \le \sup_n \mathbf{E}[\xi_{n,\mathrm{M}}^2] < \infty$.

5. $$\sup_n \mathbf{E}\left[\|\vec{\mathbf{X}}_{\mathrm{M}}\|^4 (1 + \xi_{n,\mathrm{M}}^4)\right] < \infty. \tag{2.4.9}$$

Then for any $j \in \mathrm{M}$,

$$d_\infty \left(\mathcal{L}^* \left(\sqrt{n} \left(\hat\beta^*_{n,j\cdot\mathrm{M}} - \hat\beta_{n,j\cdot\mathrm{M}} \right) \right), \mathcal{L} \left(\sqrt{n} \left(\hat\beta_{n,j\cdot\mathrm{M}} - \beta_{n,j\cdot\mathrm{M}} \right) \right) \right) \xrightarrow{P} 0. \tag{2.4.10}$$

2.4.3 A Split Samples Procedure

Based on the results in previous sections, we propose a split samples procedure for valid post-selection inference. The procedure is done by the following three steps.

1. Randomly split the data $\{(\vec{\mathbf{X}}_i, Y_i)\}_{i=1}^n$ into a model selection sample of size n_S, $\{(\vec{\mathbf{X}}_i^S, Y_i^S)_{i=1}^{n_S}\}$, and an inference sample of size n_I, $\{(\vec{\mathbf{X}}_i^I, Y_i^I)\}_{i=1}^{n_I}$.

2. In the model selection sample, apply a model selection rule $\mathcal{M}(\cdot)$ to choose submodel $\hat{\mathrm{M}} = \mathcal{M}(\mathbf{X}^S, \mathbf{Y}^S)$.

3. In the inference sample, estimate $\boldsymbol{\beta}_{n,\hat{\mathrm{M}}}$ by the LS estimate

$$\hat{\boldsymbol{\beta}}^I_{n,\hat{\mathrm{M}}} = \left((\mathbf{X}^I_{\hat{\mathrm{M}}})^T \mathbf{X}^I_{\hat{\mathrm{M}}} \right)^{-1} \left(\mathbf{X}^I_{\hat{\mathrm{M}}} \right)^T \mathbf{Y}^I. \tag{2.4.11}$$

Also, for $j \in \hat{\mathrm{M}}$, use the bootstrap distribution to obtain valid $(1-\alpha)$ confidence intervals $\mathrm{CI}^*_{n,j\cdot\hat{\mathrm{M}}}(1-\alpha)$. Denote the bootstrap LS estimate by

$$\hat{\boldsymbol{\beta}}^*_{n,\hat{\mathrm{M}}} = \left((\mathbf{X}^{I*}_{\hat{\mathrm{M}}})^T \ \mathbf{X}^{I*}_{\hat{\mathrm{M}}} \right)^{-1} \ \left(\mathbf{X}^{I*}_{\hat{\mathrm{M}}} \right)^T \ \mathbf{Y}^{I*}, \tag{2.4.12}$$

and denote the bootstrap distribution of $\left(\hat\beta^*_{n,j\cdot\hat{\mathrm{M}}} - \hat\beta_{n,j\cdot\hat{\mathrm{M}}} \right)$ by $F^*_{n,j\cdot\hat{\mathrm{M}}}$. A $1-\alpha$ confidence interval can be formed from appropriate quantiles of the bootstrap distribution:

$$\mathrm{CI}^*_{n,j\cdot\hat{\mathrm{M}}}(1-\alpha) = \left[2\hat{\beta}^I_{n,j\cdot\hat{\mathrm{M}}} - F^*_{n,j\cdot\hat{\mathrm{M}},1-\alpha/2}, 2\hat{\beta}^I_{n,j\cdot\hat{\mathrm{M}}} - F^*_{n,j\cdot\hat{\mathrm{M}},\alpha/2}\right]. \tag{2.4.13}$$

We have the following asymptotic result on this procedure for fixed p. It states that confidence intervals for LS estimates based on the split samples bootstrap procedure have correct asymptotic coverage probability.

Corollary 2.4.2. *Suppose the following conditions hold:*

1. *The number of predictors p is fixed.*

2. *$(\vec{\mathbf{X}}_i, Y_i)$ are i.i.d. with finite second moments $\mathbf{E}[Y_i^2] < \infty$ and $\mathbf{E}[\|\vec{\mathbf{X}}_i\|^2] < \infty$.*

3. *The eigenvalues of $\mathbf{E}[\vec{\mathbf{X}}\vec{\mathbf{X}}^T]$ are bounded away from 0 and ∞ for all n.*

4. *For $\xi_{n,\mathrm{M}}$ defined in (2.4.8), $0 < \inf_n \inf_{\mathrm{M}} \mathbf{E}[\xi^2_{n,\mathrm{M}}] \le \sup_n \sup_{\mathrm{M}} \mathbf{E}[\xi^2_{n,\mathrm{M}}] < \infty$.*

5. $$\sup_n \sup_{\mathrm{M}} \mathbf{E}\left[\|\vec{\mathbf{X}}_{\mathrm{M}}\|^4(1+\xi^4_{n,\mathrm{M}})\right] < \infty. \tag{2.4.14}$$

6. *The inference sample size n_I satisfies $\liminf(n_I/n) > 0$ as $n \to \infty$.*

Then for any $j \in \hat{\mathrm{M}} = \mathcal{M}(\mathbf{X}^S, \mathbf{Y}^S)$, as $n \to \infty$,

$$d_\infty\left(\mathcal{L}^*\left(\sqrt{n_I}\left(\hat{\beta}^*_{n,j\cdot\hat{\mathrm{M}}} - \hat{\beta}^I_{n,j\cdot\hat{\mathrm{M}}}\right)\right), \mathcal{L}\left(\sqrt{n_I}\left(\hat{\beta}^I_{n,j\cdot\hat{\mathrm{M}}} - \beta_{n,j\cdot\hat{\mathrm{M}}}\right)\right)\right) \xrightarrow{P} 0. \tag{2.4.15}$$

Moreover, for $\mathrm{CI}^*_{n,j\cdot\hat{\mathrm{M}}}(1-\alpha)$ *defined as in* (2.4.13)*, as* $n \to \infty$*,*

$$\mathbf{P}\left(\beta_{n,j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}^*_{n,j\cdot\hat{\mathrm{M}}}(1-\alpha)\right) \to 1-\alpha. \tag{2.4.16}$$

Here are some remarks:

1. There is no requirement on sample size on the model selection sample; however, it is desirable that the size of the inference sample should be large.
2. This result does not depend on the model selection procedure as it requires the regularity conditions to hold uniformly for every submodel.

2.5 Inference under First-Order Correctness and Homoscedasticity

This section is organized as follows: The first subsection introduces the general properties of LS estimates under the first-order correctness and homoscedasticity. The second subsection focuses on the properties of the nonlinearity term in a fixed submodel $\xi_{n,\mathrm{M}}$. The third subsection describes some properties of the LS estimates if we further assume the distribution of $\vec{\mathbf{X}}$ is Gaussian. With the results from those three subsections, we derive the theorem that the split samples bootstrap procedure gives valid post-selection inference.

2.5.1 The Properties of LS Estimates under First-Order Correctness and Homoscedasticity

If $\eta_n(\vec{\mathbf{X}}) = 0$ a.s., then by (2.3.2),

$$\mathbf{E}\big[\hat{\boldsymbol{\beta}}_n\big] = \boldsymbol{\beta}_n \tag{2.5.1}$$

So the LS estimate in the full model is unbiased.

To see if the LS estimate in a submodel is unbiased, we first note that under a submodel M, by (2.4.1)

$$\boldsymbol{\beta}_{n,\mathrm{M}} = \left(\mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}}^T\right]\right)^{-1} \mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}^T\right] \boldsymbol{\beta}_n. \tag{2.5.2}$$

With this notation, we have the following theorem on the expectation of the LS estimate under M, $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$.

Theorem 2.5.1. *Suppose the conditional expectation of $\vec{\mathbf{X}}_{\mathrm{M}^c}$ be linear given $\vec{\mathbf{X}}_{\mathrm{M}}$:*

$$\mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}^c}\middle|\vec{\mathbf{X}}_{\mathrm{M}}\right] = \mathbf{L}\vec{\mathbf{X}}_{\mathrm{M}}, \tag{2.5.3}$$

where $\mathbf{L}$ is some matrix of size $|\mathrm{M}^c| \times |\mathrm{M}|$. Then $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$ is an unbiased estimate of $\boldsymbol{\beta}_{n,\mathrm{M}}$.

Proof:

Let $\boldsymbol{\beta}_{n,[\mathrm{M}]}$ denote the vector consisting of those components of vector $\boldsymbol{\beta}_n$ with indices corresponding to variables in model M, let $\boldsymbol{\Sigma}_{[\mathrm{M},\mathrm{M}]} = \mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}}^T\right]$, and let

$\mathbf{\Sigma}_{\mathrm{M}^c|\mathrm{M}} = \mathbf{\Sigma}_{[\mathrm{M}^c,\mathrm{M}^c]} - \mathbf{\Sigma}_{[\mathrm{M}^c,\mathrm{M}]}\mathbf{\Sigma}^{-1}_{[\mathrm{M},\mathrm{M}]}\mathbf{\Sigma}_{[\mathrm{M},\mathrm{M}^c]}$. Then we have the following proposition

With the above notation, we have the following decomposition

$$\boldsymbol{\beta}_{n,\mathrm{M}} = \boldsymbol{\beta}_{n,[\mathrm{M}]} + \left(\mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}}^T\right]\right)^{-1}\mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}^c}^T\right]\boldsymbol{\beta}_{n,[\mathrm{M}^c]}. \tag{2.5.4}$$

Note that by (2.5.4),

$$\boldsymbol{\beta}_{n,\mathrm{M}} = \boldsymbol{\beta}_{n,[\mathrm{M}]} + \left(\mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}}^T\right]\right)^{-1}\mathbf{E}\left[\mathbf{E}\left[\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}^c}^T|\vec{\mathbf{X}}\right]\right]\boldsymbol{\beta}_{n,[\mathrm{M}^c]} = \boldsymbol{\beta}_{n,[\mathrm{M}]} + \mathbf{L}\boldsymbol{\beta}_{n,[\mathrm{M}^c]}. \tag{2.5.5}$$

Note also that by (2.4.3),

$$\mathbf{E}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big] = \boldsymbol{\beta}_{n,[\mathrm{M}]} + \mathbf{E}\left[\mathbf{E}\left[\big(\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}}^T\big)^{-1}\vec{\mathbf{X}}_{\mathrm{M}}\vec{\mathbf{X}}_{\mathrm{M}^c}^T\boldsymbol{\beta}_{n,[\mathrm{M}^c]}|\vec{\mathbf{X}}\right]\right] = \boldsymbol{\beta}_{n,[\mathrm{M}]} + \mathbf{L}\boldsymbol{\beta}_{n,[\mathrm{M}^c]}. \tag{2.5.6}$$

Hence, $\mathbf{E}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big] = \boldsymbol{\beta}_{n,\mathrm{M}}$. □

For condition (2.5.3) to hold we need either

1. $\mathbf{E}[\vec{\mathbf{X}}] = \mathbf{0}$, or

2. Full model and submodel M both include an intercept.

2.5.2 Gaussian Case

Suppose further that $(X_1, \dots, X_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Then we can have the following result on the conditional distribution of $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$.

Theorem 2.5.2. *Conditional on* $\vec{\mathbf{X}}_{\mathrm{M}}$,

$$\begin{aligned} \mathbf{E}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big|\mathbf{X}_{\mathrm{M}}\big] &= \boldsymbol{\beta}_{n,\mathrm{M}} & (2.5.7) \\ \mathrm{Var}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big|\mathbf{X}_{\mathrm{M}}\big] &= \kappa_{\mathrm{M}}^2(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1} & (2.5.8) \end{aligned}$$

where $\kappa_{\mathrm{M}}^2 = (\boldsymbol{\beta}_{n,[\mathrm{M}^c]}^T\boldsymbol{\Sigma}_{\mathrm{M}^c|\mathrm{M}}\boldsymbol{\beta}_{n,[\mathrm{M}^c]} + \sigma_n^2)$.

The unconditional distribution of $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$ is obtained as follows.

Theorem 2.5.3.

$$\hat{\boldsymbol{\beta}}_{n,\mathrm{M}} = \boldsymbol{\beta}_{n,\mathrm{M}} + \kappa_{\mathrm{M}}T_{m,n-m+1,\boldsymbol{\Sigma}_{[\mathrm{M},\mathrm{M}]}^{-1}/(n-m+1)} \qquad (2.5.9)$$

where $T_{m,n-m+1,\boldsymbol{\Sigma}_{[\mathrm{M},\mathrm{M}]}^{-1}/(n-m+1)}$ *is an m-dimensional random vector having multivariate t-distribution with* $n-m+1$ *degrees of freedom, location parameter* 0 *and scale matrix* $\boldsymbol{\Sigma}_{[\mathrm{M},\mathrm{M}]}^{-1}/(n-m+1)$.

In particular, the unconditional mean and variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}$ *are given by*

$$\begin{aligned} \mathbf{E}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big] &= \boldsymbol{\beta}_{n,\mathrm{M}} & (2.5.10) \\ \mathrm{Var}\big[\hat{\boldsymbol{\beta}}_{n,\mathrm{M}}\big] &= \frac{\kappa_{\mathrm{M}}^2}{n-m-1}\big(\boldsymbol{\Sigma}_{[\mathrm{M},\mathrm{M}]}\big)^{-1} & (2.5.11) \end{aligned}$$

2.5.3 Valid Bootstrap Inference in Submodels under the Gaussian Distribution

Under Gaussian distribution, we can further simplify the conditions in Corollary 2.4.2. This is seen by noting that if $(X_1, \ldots, X_p) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_n)$, then $\vec{\mathbf{X}}$ and $\xi_{n,\mathrm{M}} = \boldsymbol{\beta}_n^T \vec{\mathbf{X}} - \boldsymbol{\beta}_{n,\mathrm{M}}^T \vec{\mathbf{X}}_{\mathrm{M}} + \epsilon_n$ have bound fourth moments. Thus, we have the following corollary of Corollary 2.4.2, which shows that the split samples procedure gives valid post-selection confidence intervals asymptotically under the normal distribution.

Corollary 2.5.1. *Suppose* $\{(\vec{\mathbf{X}}_i, Y_i)\}_{i=1}^n$ *are i.i.d. observations from a Gaussian population* $\mathbf{P}_n(dx_1, \ldots, dx_p, dy)$ *such that the following conditions hold:*

1. $\mathbf{E}\left[\vec{\mathbf{X}}\right] = \mathbf{0}$*,* $\mathbf{E}\left[Y|\vec{\mathbf{X}}\right] = \boldsymbol{\beta}_n^T \vec{\mathbf{X}}$ *for some* $\boldsymbol{\beta}_n$*, and* $0 < \mathbf{E}\left[(Y - \boldsymbol{\beta}_n^T \vec{\mathbf{X}})^2 | \vec{\mathbf{X}}\right] = \sigma_n^2 < \infty$.
2. $\sup_n \boldsymbol{\beta}_n^T \mathbf{E}\left[\vec{\mathbf{X}}\vec{\mathbf{X}}^T\right] \boldsymbol{\beta}_n < \infty$.
3. *The inference sample size* n_I *satisfies* $\liminf(n_I/n) > 0$ *as* $n \to \infty$.

Then for any $j \in \hat{\mathrm{M}} = \mathcal{M}(\mathbf{X}^S, \mathbf{Y}^S)$*, as* $n \to \infty$*,*

$$d_\infty \left(\mathcal{L}^* \left(\sqrt{n} \left(\hat{\beta}^*_{n,j\cdot\hat{\mathrm{M}}} - \hat{\beta}^I_{n,j\cdot\hat{\mathrm{M}}} \right) \right), \mathcal{L} \left(\sqrt{n} \left(\hat{\beta}^I_{n,j\cdot\hat{\mathrm{M}}} - \beta_{n,j\cdot\hat{\mathrm{M}}} \right) \right) \right) \xrightarrow{P} 0. \tag{2.5.12}$$

Chapter 3

Using Split Samples in an Observational Study

3.1 Introduction: background; methodological outline

3.1.1 A wave of closures of hospital obstetrics units

Beginning in 1997, a series of community hospitals in Philadelphia closed their obstetrics units, so mothers who would normally have delivered at these hospitals had to seek care at the city's large regional hospitals whose obstetrics units remained open. Between 1997 and 2007, 12 of 19 hospitals in the city closed their obstetrics units. Nothing similar happened at this time in other major cities, which experienced only sporadic changes in the availability of obstetrics units. For instance, in Pittsburgh, Los Angeles, San Diego and San Francisco less than 5% of the deliveries in 1995 and 1996 were in obstetric units that subsequently closed between 1997-2005. Babies born in these and other cities will serve as controls. By contrast, in Philadelphia, over 30% of the deliveries in 1995 and 1996 occurred at obstetrics units that subsequently closed between 1997 and 2005. It is not entirely surprising that a hospital facing competitive or financial pressures would consider closing its obstetrics and neonatal units: these fields have unusually high costs associated with malpractice litigation and malpractice insurance (Kirby et al. 2006). Why closures should have concentrated in Philadelphia is less clear. In its densely urban center, Philadelphia is home to several large hospitals associated with major medical schools, but beyond its urban center, Philadelphia sprawls at considerable distance into a variety of diverse neighborhoods served by smaller community hospitals; the closures occurred here.

Of 19 Philadelphia hospitals with obstetrics units in 1995, 12 closed their obstetrics units between 1997 and 2007; see Figure 3.1. In part based on a split sample analysis described below, the analysis presented here focuses on five hospitals that abruptly closed in 1997-1999, before the City of Philadelphia intervened in 2000 to organize and slow the pace of subsequent closures and to offer strategies to allow for the remaining hospitals to accommodate the increased obstetric volume. It is interesting to note that four of the five closures during 1997-1999 were geographically close, suggesting a cascade in which each successive closure increased the stress on near-by units that remained open, perhaps leading to their closure. Conceivably, the geography of Philadelphia's closures explain why there was a wave of closures in Philadelphia with no similar pattern in other cities.

What was the effect of the 1997-1999 hospital closures on the health of mothers and their newborn babies? Stories were told — perhaps some were even true — of women in labor being delivered by ambulance to a hospital that had closed its obstetrics unit the previous week. Other stories were told — more likely true — of women in labor, some of them poor, traveling longer distances, perhaps in rush hour, to reach an open obstetrics unit, of overcrowding and inadequate staffing at the units that remained open. A closure in one neighborhood may force a mother who lives in that neighborhood to travel a long distance to a hospital in another neighborhood, but it may also cause overcrowding in a hospital remote from the closure, and so it may affect mothers who live near the hospital that remained open. It is easy to imagine a long trip to an overcrowded obstetrics unit is not beneficial. Then again, many of the hospitals that remained open have excellent reputations, better perhaps than the reputations of the hospitals that closed their obstetrics units. Then again, teaching hospitals are home to the most and least experienced doctors, professors of medicine and medical residents, who usually work in tandem, but who found themselves short of staff. Then again, the human race has managed to reproduce in circumstances considerably more dire than traffic and overcrowding. It is hard to know what, if anything, to expect from the five closures in 1997-1999.

3.1.2 Matching to build a control Philadelphia

For each birth in Philadelphia in 1995-2003, we used multivariate techniques and an optimal assignment algorithm to match a control birth from elsewhere in Pennsyl-

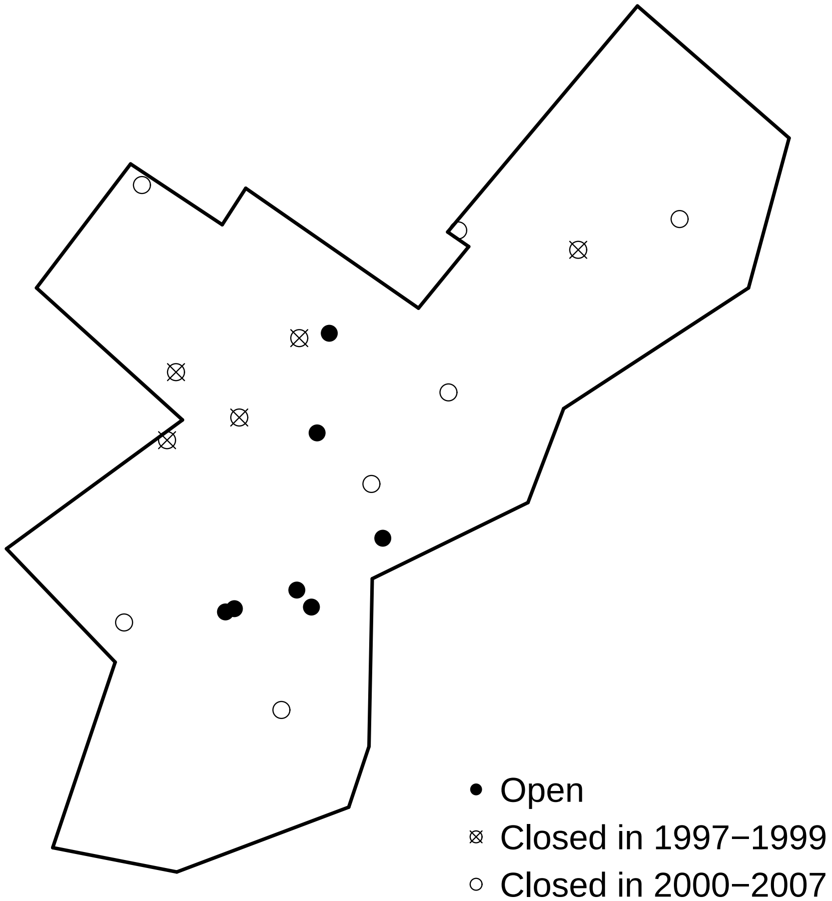

Figure 3.1: *Map of the City of Philadelphia showing hospitals that closed their obstetrics units. The analysis in the current paper focuses on closures in 1997-1999, before the City intervened to pace and organize the process of closure.*

vania or California or Missouri, the three states for which we had the needed data. Because there were 132,786 births in Philadelphia and 5,998,111 potential control births elsewhere, the matching was on an unusually large scale. The matching was done year-by-year, so a Philadelphia birth in 1995 was matched to a control birth in 1995, and it controlled not only characteristics of the mother and baby, but also characteristics of the mother's neighborhood, such as typical income, the frequency of poverty, and the level of education in the neighborhood. During this time period, Philadelphia mothers were quite different from the unmatched potential control group: they came from neighborhoods with lower income, more poverty, and fewer high school graduates; however, the mothers themselves (as opposed to their neighborhoods) were more likely than potential controls to have graduated high school. Philadelphia mothers were somewhat younger with less prenatal care, but their babies were, on average, slightly smaller. All of these measured differences and many other measured differences were removed year by year using matching techniques; see Section 3.2. The control mothers and infants are not only similar as individuals: as a group, they have similar temporal and measured neighborhood characteristics to births in Philadelphia in 1995-2003. Here, neighborhood characteristics are measured at the zip-code level and are indicated in Table 3.1.

Why build a control Philadelphia? Because of the geography of Philadelphia, the closures might be expected to affect certain neighborhoods more than others, and each neighborhood has its own demographics, income, social and health problems. A control Philadelphia permits straightforward questions about how mothers and neighborhoods in Philadelphia changed in comparison with similar mothers and

Table 3.1: *Covariate balance before and after matching. For Zip Code data, zip-fr means the fraction of the Zip Code with this attribute. An absolute standardized difference in mean of 0.2 or greater is in **bold**.*

Sample Size	5,998,111	132,786	132,786	Absolute	
	Potential	Philadelphia	Matched	Standardized	
	Controls	Births	Controls	Difference	
Covariate	Covariate Mean or Proportion			Before	After
	Mom's Neighborhood (Zip code)				
Income (K$)	46	30	30	**1.16**	0.04
Income Missing	0.00	0.00	0.00	0.06	0.00
Poverty (zip-fr)	0.15	0.25	0.23	**0.91**	0.13
Poverty Missing	0.00	0.00	0.00	0.06	0.00
High School (zip-fr)	0.74	0.68	0.69	**0.37**	0.07
HS Missing	0.00	0.00	0.00	0.06	0.00
College (zip-fr)	0.22	0.15	0.15	**0.51**	0.01
College Missing	0.00	0.00	0.00	0.06	0.00
	Mom				
Mom's Age	28	26	26	**0.21**	0.01
Parity	2.10	2.20	2.20	0.07	0.03
Parity Missing	0.00	0.01	0.01	0.09	0.04
Prenatal Care (Month Started)	2.40	2.70	2.60	**0.22**	0.04
PC Missing	0.02	0.11	0.08	**0.37**	0.11
	Mom's Education				
Below 8th Grade	0.10	0.02	0.02	**0.32**	0.02
Some High School	0.17	0.21	0.20	0.11	0.04
HS Graduate	0.30	0.38	0.40	0.17	0.05
Some College	0.20	0.19	0.19	0.02	0.01
College Graduate	0.13	0.09	0.10	0.11	0.01
More than College	0.09	0.06	0.06	0.11	0.00
Missing	0.01	0.04	0.04	0.17	0.04
	Mom's Race				
White	0.71	0.31	0.32	**0.87**	0.03
Black	0.07	0.42	0.46	**0.88**	0.11
Asian	0.07	0.03	0.03	0.18	0.03
Other	0.12	0.06	0.05	**0.20**	0.05
Missing	0.02	0.17	0.14	**0.52**	0.13
	Mom's Health Insurance				
Government	0.40	0.40	0.39	0.01	0.02
Other Insurance	0.57	0.58	0.60	0.02	0.04
Uninsured	0.03	0.01	0.01	0.11	0.04
Missing	0.00	0.01	0.00	0.11	0.06
	Baby				
Birth Weight, (grams)	3345	3179	3189	**0.26**	0.02
Birth Weight Missing	0.00	0.00	0.00	0.04	0.03
Gestational Age (Weeks)	39	38	38	0.14	0.01
Gestational Age Missing	0.05	0.01	0.01	**0.22**	0.02
Small at Gestational Age	0.09	0.14	0.12	0.16	0.05

neighborhoods elsewhere.

Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010) developed an innovative approach to using aggregate data to synthesize a control for a region that was subjected to an intervention. Their synthetic control is a weighted combination of actual regions that were not subjected to the intervention. For example, in their study of the economic impact of terrorism in the Basque Country, Abadie and Gardeazabal (2003) use a weighted combination of two Spanish regions to approximate the economic growth that the Basque Country would have experienced in the absence of terrorism. The weighted combination is chosen to match the region subjected to the intervention in its covariates and trajectory of outcomes prior to the intervention. Abadie, Diamond and Hainmueller (2010) developed an inferential approach when using synthetic controls that is akin to permutation inference. They use placebo tests to examine whether or not the estimated effect of the actual intervention is large relative to the distribution of the effects estimated for the regions not exposed to the intervention, where the synthetic control method is also used to estimate effects for regions not exposed to the intervention. A valuable feature of Abadie et al.'s synthetic control approach is that it only requires aggregate data on regions, which are often the only type of data available. For our study of the effect of the obstetric unit closures in Philadelphia, we are fortunate to have individual data on mothers and babies, which permit, for example, comparisons of parts of Philadelphia with its control.

3.1.3 Splitting

Philadelphia mothers and infants may have differed from controls in ways that were not measured and hence not controlled by matching for observed covariates. After adjustment for observed covariates, the key source of uncertainty in an observational study is the possibility that differences in outcomes between treated and control subjects are not effects of the treatment but rather biases from some unmeasured way in which treated and control subjects were not comparable. Our analysis is largely directed at this possibility.

A sensitivity analysis asks how failure to control some unmeasured covariate might alter the conclusions of a study. Many issues affect the sensitivity of conclusions to unmeasured biases (Rosenbaum 2004; 2010a, Part III; 2010b), but most of these issues are difficult to appraise in the absence of data. Heller et al. (2009) made a formal argument for splitting the sample at random into a small planning sample of perhaps 10% and a large analysis sample of perhaps 90%. The planning sample is used to design the study — to frame questions and guide the analytical plan — whereupon the planning sample is discarded; then, all conclusions are based on the untouched, unexamined, untainted analysis sample. If one were to perform several or many analyses of a single data set, noting that a particular conclusion was insensitive to unmeasured biases, then one would not know whether this judgement about sensitivity to bias was distorted by capitalizing on chance in picking the most favorable of these analyses. In contrast, the use of a split sample permits exploration of unlimited scope in a planning sample, and an independent, untainted, highly focused analysis of the analysis sample. Cox (1975) evaluated splitting to control for multiple testing in randomized experiments, but Heller et al. (2009) find that splitting is even more useful in sensitivity analyses in observational studies because the biases from unmeasured covariates do not diminish as the sample size increases. If one could make decisions that would make the study less sensitive to unmeasured biases by sacrificing a small portion of the sample, then that sacrifice might be well worth making. The formal argument in Heller et al. (2009) evaluates power and design sensitivity in split samples.

As Cox (ibid.) emphasized, splitting has an important advantage over most methods that address multiple testing, namely it permits human judgement to play an informed role between exploratory analysis of the planning and focused confirmatory analysis of the analysis sample. Formal or algorithmic procedures that address multiple testing, such as the Bonferroni inequality, do not leave a role for judgement; rather, their form must be prespecified. In the current study, this meant that an extensive analysis of the planning sample was discussed at a meeting of the clinicians and statisticians, and the analysis plan that emerged from that meeting reflected results from the planning sample combined with clinical and statistical judgement. For instance, before looking at any data, we thought that overcrowding in an obstetrics ward might result in an increase in Caesarean sections and birth injuries of various kinds, but the planning sample strongly suggested a focus on serious birth injuries (ICD-9 767-3), and not a focus on Caesarean sections. In part, our focus on serious birth injuries reflects what we saw in the planning sample, but in part it reflects a judgement about an effect that seems both plausible and clinically interesting. The planning split also revealed that several outcomes were simply too rare to study even with the much larger analysis sample; here, it is not the P-value but the event rate that provides information relevant to power computations for the as yet unexamined analysis sample. Although one can mechanize the evaluation of many P-values, one cannot mechanize an evaluation of many P-values that incorporates human judgement about what is plausible and interesting. Because human judgement cannot be mechanized, it is not typically possible to perform the same analysis on many repeated splits of the sample, as one might do in cross-validation.

Here, we took a small random sample of the matched pairs, 10% or 13,278 pairs in this study, and used it to plan the main analysis, which concerned the complementary 90% of pairs or 119,508 pairs. Among many outcomes examined using the planning split sample, we were led to focus on birth injuries, specifically ICD code 767.3, and on the years 1997-1999 when five hospitals abruptly closed their obstetrics units. Beginning in 2000, the City of Philadelphia intervened to slow down and organize closures. Before looking at the planning sample, it was not obvious to us whether the City's intervention had been more than a symbolic gesture, but the planning sample suggested that most of the action occurred in 1997-1999, that is, after the City's intervention there was no discernable effect of hospital closures. If this analytic focus had come about after examining many outcomes and various comparisons for those outcomes using the complete data, then there would naturally be reason for concern that the focus was distorted by capitalizing on chance events that only appear to be systematic patterns. However, this analytic focus came about by examining a random sample of 10% of the pairs, and 90% of the pairs remain to put this carefully chosen, very specific focus to a proper test. One might imagine two investigators, one who early on published a small, informal, exploratory, highly speculative and not particularly convincing study involving many comparisons, with the second investigator taking the one promising result from the first study and confirming it in a much larger independent sample. From an inferential point of view, it makes no difference whether there were two investigators or only one, that is, no difference between, on the one hand, replicating a promising but speculative finding by someone else and, on the other hand, generating both the speculative finding and the confirmation using split samples.

3.1.4 Evidence factors

If we are looking at a treatment effect, not a bias from unmeasured covariates, then we anticipate several patterns. First, when compared to similar births in other states, an effect of the closures should be absent in 1995-1996 and present in 1997-1999. For birth injuries, a binary outcome, this leads to a difference-in-difference analysis along the lines suggested by Gart (1969) for randomized crossover studies; see Section 3.4 where discordant pairs become the counts in a 2 × 2 table that is subjected to a sensitivity analysis. Second, we identified thirteen zip codes in northern Philadelphia as close to the hospitals with closures (specifically, 19115, 19119, 19121, 19127, 19128, 19129, 19131, 19132, 19135, 19136, 19144, 19149, 19152). Of course, overcrowding occurred in the obstetrics units that remained open, and many of these were at some distance from the closures; nonetheless, it is reasonable to contrast zip codes with closures to zip codes without closures in 1997-1999, anticipating a larger effect on zip codes with closures. Finally, if the difference between the Philadelphia-versus control difference in the zip codes with closures and in zip codes without closures was already apparent in 1995-1996, before the closures, then that cannot plausibly be an effect of the closures; rather, it must indicate that our matching and difference-in-differences have failed to compare comparable mothers under different treatments. The first two comparisons are an example of evidence factors, that is, of (nearly) independent tests of the hypothesis of no treatment effect that are susceptible to different kinds of unmeasured biases (Rosenbaum 2010c), whereas the third comparison is a test for unmeasured bias (Rosenbaum 1984).

The method of difference-in-differences has a long history; see, for instance, Campbell (1957, 1969), Meyer (1995), Angrist and Krueger (2000), Shadish, Cook and Campbell (2002) and Athey and Imbens (2006). A conventional description of difference-in-differences follows, although Proposition 3.4.1 departs from this description by studying sensitivity to biases that can affect difference-in-difference studies. In a nonrandomized treatment-versus-control comparison the treatment effect is aliased with stable but unmeasured baseline differences between treated and control groups, whereas in a before-versus-after comparison, the treatment effect is aliased with trends over time. In contrast, in a difference-in-differences study, the treatment effect is aliased neither with stable unmeasured baseline differences between treated and control groups nor with trends over time that affect all groups in the same way, but it is aliased with the interaction of those two sources of bias. Proposition 3.4.1 examines sensitivity of inferences about effects to biases from such interactions. Although difference-in-differences is conventionally defined in terms of the passage of time, it is more generally relevant to situations in which a treatment effect is aliased with the interaction of two sources of bias, and this generality is exploited here in the second evidence factor, where time is replaced by Philadelphia zip codes near closures.

For a recent review of matching techniques, see Stuart (2010). For discussion of the importance of anticipated patterns in observational studies, see Campbell (1957), Trochim (1985), Shadish, Cook and Campbell (2002) and West et al. (2008). Various methods of sensitivity analysis in observational studies are discussed by Cornfield et al. (1959), Rosenbaum and Rubin (1983), Yanagawa (1984), Gastwirth(1992), Gastwirth, Krieger, and Rosenbaum (1998), Rosenbaum (1995; 2002, Section 4), Marcus (1997), Lin et al. (1998), Robins et al. (1999), Copas and Eguchi (2001), Imbens (2003) and DiPrete and Gangl (2004).

3.2 Matching

3.2.1 Philadelphia and elsewhere, before and after matching

We obtained birth certificates from all deliveries occurring in Pennsylvania, California and Missouri between 1/1/1995 and 6/30/2005. Each state's department of health linked these birth certificates to death certificates using name and date of birth, and then de-identified the records. We then linked over 98% of birth certificates to maternal and newborn hospital records. Over 80% of the remaining unlinked birth certificate records failed to identify a hospital, suggesting a birth at home or a birthing center. The unlinked records had similar gestational age and racial/ethnic distributions to the linked records. For the maternal and newborn hospital records, California, Missouri, and Pennsylvania routinely collect information on all hospital admissions within each state. Each patient record contains the UB-92 form submitted by each hospital to the state, with 15 to 25 fields for principal diagnoses and procedures occurring during the hospital stay. Birth certificates contain information on birth weight, gestational age, and patient-level demographic variables and obstetric risk factors. Sociodemographic information on the mother's zip code is obtained from the Bureau of the Census.

Each baby born in Philadelphia was matched with a baby born in other regions of Pennsylvania or California or Missouri. In each year, the match balanced 59 observed covariates. Of these, 34 covariates are listed in Table 3.1, which gives their means among potential controls outside Philadelphia, in Philadelphia, and in the matched controls. These covariates describe the socioeconomic status of mom's neighborhood, mom's own age, parity, prenatal care, education, race, and health insurance, and baby's birth weight and gestational age, two key measures of a newborn's health status. Because we are interested in the effects of the hospitals at the time of delivery, we adjust for quantities such as gestational age and birth weight that are essentially determined prior to admission to the hospital. These factors are associated with different risks of many neonatal outcomes (Stoll et al. 2010). A study of prenatal care, as opposed to care around the time of delivery, would not adjust for gestational age and birth weight, although in fact there is little compelling evidence that prenatal medical care has much effect on preterm delivery (American College of Gynecology 2003, Hollowell et al. 2011). Babies were also matched exactly for year of birth.

For each of the 34 covariates, Table 3.1 also gives the standardized absolute difference in means before and after matching, that is, Philadelphia-versus-potential controls and Philadelphia-versus-matched controls. The pooled standard deviation used in this measure is calculated as the square root of the equally weighted average of the sample variances inside and outside Philadelphia before matching, so matching changes the numerator, that is the difference in means, but it does not change the denominator, the pooled standard deviation. See Rosenbaum and Rubin (1985) for discussion of this conventional measure of covariate imbalance. In addition to the covariates in Table 3.1, there are 25 other covariates, $59 = 34 + 25$, which describe rare congenital anomalies or problems in the pregnancy that existed long before the start of labor.

Before matching, compared to potential controls, Philadelphia mothers were, on average, more likely to live in a low income neighborhood in which fewer people had college degrees, slightly younger with a little less prenatal care, more likely to have completed 8th grade, more often black, and gave birth to somewhat smaller babies.

Figure 3.2 displays all 59 absolute standardized differences in means in each of five years, 1995-1999. Before matching several covariates differed by more than 0.8 standard deviations. After matching, all $295 = 5 \times 59$ standardized differences in means after matching are less than 0.2 standard deviations. Before matching, the maximum and upper quartile of the 295 absolute standardized differences were 1.19 and 0.18, whereas after matching they were 0.19 and 0.06, respectively. For comparison, a Normal distribution has 95% of its probability on an interval that is

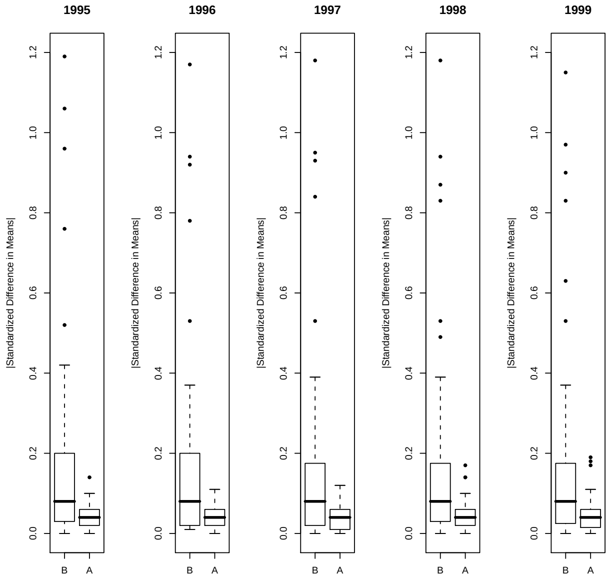

Figure 3.2: *Covariate balance before (B) and after (A) matching for 59 covariates in each of five years, measured as the absolute difference in means in units of a pooled standard deviation.*

approximately four standard deviations in length, so 0.19 and 0.06 of a standard deviation are approximately 5% and 2% of such an interval. In brief, Figure 3.2 shows that after matching, all of the 59 covariate means were in reasonable balance in every year; that is, Philadelphia and control-Philadelphia were similar in terms of these covariates year by year. Figure 3.3 displays balance for four continuous covariates.

3.2.2 How the matching was done

There were 132,786 births in Philadelphia and 5,998,111 potential control births to choose from in building the matched comparison. In matching, a large sample size should be a luxury, but if inappropriate methods are used, it can appear to be a hindrance. A 132786×5998111 distance matrix would contain approximately 7.96×10^{11} numbers, and this is well beyond what can be handled with current combinatorial optimization techniques on current computers. There is a simple solution, however: match exactly for some important covariates, thereby reducing one large problem to a series of smaller problems; see Rosenbaum (2010a, Section 9.3).

We ordered the covariates by priority, year of birth being first because of the structure of the study, followed by gestational age in weeks $(0, 33]$, $(33, 36]$, $(36, 38]$, $(38, 40]$ and $(40, \infty)$, categories based on an estimated propensity score for the propensity to be born in Philadelphia, mother's age in years $(0, 18]$, $(18, 34]$, $(34, \infty)$, mother's education in four groups by degree. The algorithm first looked at the size of the distance matrix within a given year; if that was too large, it looked at the size

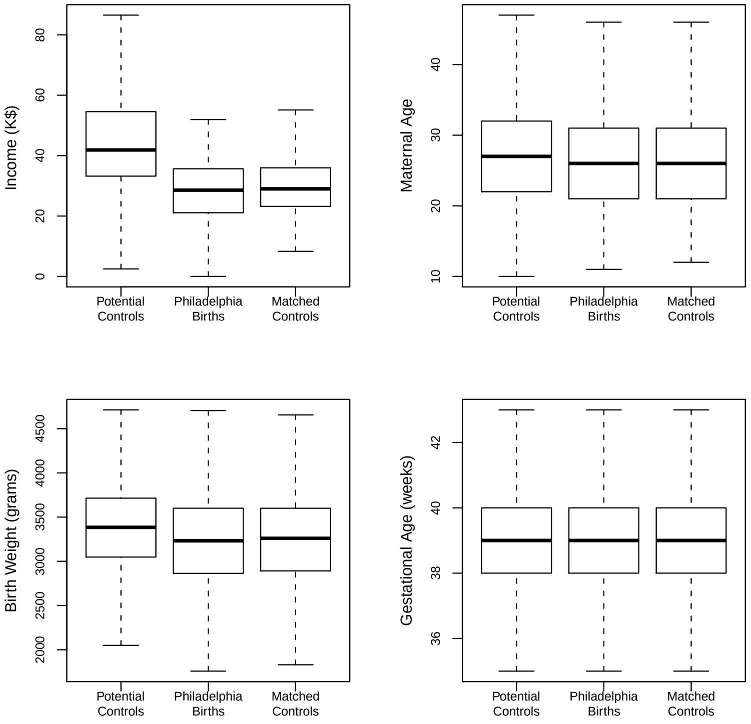

Figure 3.3: *Covariate imbalance before and after matching for four continuous covariates, namely income, maternal age, birth weight, and gestational age.*

of the distance matrix within a given year and gestational age; if that was too large, it looked within a given year, gestational age and propensity score group, and so on. Once the size of the distance matrix was manageable, the distance matrix was computed using a rank-based Mahalanobis distance within calipers for an estimated propensity score (Rosenbaum and Rubin 1985; Rosenbaum 2010a, Section 8), and an optimal match was determined to minimize the total distance within matched pairs (Rosenbaum 1989; 2010a, Section 8). Calipers on the propensity score ensure a close match on a unidimensional summary sufficient to remove bias from imbalances in observed covariates; see Rosenbaum and Rubin (1985) and Abadie and Imbens (2011) for discussion of calipers and unidimensionality in matching. The computations used Hansen's (2007) `optmatch` package in `R`; see also Hansen and Klopfer (2006).

3.3 Splitting

3.3.1 A 10%-90% random split for design and analysis

As proposed by Heller et al. (2009), within each year, the planning sample was a 10% sample of pairs drawn at random without replacement. The analysis sample was the complementary 90% of pairs. As noted in Section 3.1.2, the base period, 1995-1996 had no closures of obstetrics units, 1997-1999 had five abrupt closures, whereas beginning in 2000 the City of Philadelphia intervened to prevent abrupt closures so that closures followed some delay and reorganization among open hospitals. The planning sample looked at 38 outcomes in each of two time periods defined by the

City's intervention in the process of closure, 1997-1999 and 2000-2003, for all zip codes, for zip codes close to closures and for zip codes remote from closures, so a total of $38 \times 2 \times 3 = 228$ significance levels were computed. Consistent with the discussion by Cox (1975) and Heller et al. (2009), sample splitting served as a substitute for a correction for multiple testing.

The planning sample suggested several interesting hypotheses, and here we focus on one of these, namely birth injury ICD-9 767.3. Unlike some of the other 767 codes, code 767.3 is a serious injury, such as fracture of long bones or the skull, not a routine abrasion of a normal birth. The planning sample suggested an increase in such birth injuries in Philadelphia in 1997-1999 with a return to normal in 2000-2003, with some indication that the increase was more pronounced for mothers who lived in zip codes affected by closures.

The planning sample is used informally to suggest interesting hypotheses and appropriate analyses. To motivate and clarify the theoretical discussion in Section 3.4, we present an analysis of birth injury for the 10% planning sample in the same form that will be used in the final analysis of the complementary 90% sample. Actually, we did quite a bit of analysis of the planning sample before settling upon this form. Having selected this form, the analysis of the complementary 90% sample simply used this one form on this outcome. The analysis of the 90% sample incorporates a sensitivity analysis developed in Section 3.4.

3.3.2 Birth injury in the planning sample: the largest difference, two nearly independent tests for effect and a test for unmeasured bias

Table 3.2 is the analysis of birth injury for the 10% planning sample. It has four panels labeled " a comparison focused on the most affected groups," " factor 1," " factor 2," " bias test."Factor 1 is the simplest comparison, so it is described first; then the other parallel comparisons are described briefly. Table 3.2 counts Philadelphia-control pairs discordant for birth injury, that is, pairs in which exactly one baby experienced a birth injury. Factor 1 compares Philadelphia to control in 1997-1999 versus 1995-1996. In 1995-1996, there were 85 pairs containing one birth injury, and in 43 pairs it was the Philadelphia baby who was injured and in 42 pairs it was the control baby who was injured. In contrast, during the period of closures, 1997-1999, there were 184 pairs with birth injuries, and in 141 of the 184 pairs it was the Philadelphia baby who experienced the injury. The odds ratio in this 2×2 table is 3.19, so it looks as if there was an increase in the risk of birth injury in Philadelphia during the period of hospital closures. Because of this observation in the planning sample, the analysis in the complementary 90% sample will look for an increase in risk for this same outcome. Our data do not locate the birth injury as occurring either in the hospital or prior to reaching the hospital, say in an ambulance. The most affected group contrasts Philadelphia zip codes near closures to matched controls in 1995-1996 and in 1997-1999; both a priori and as indicated in this planning split sample, it seems reasonable to think that if a strong effect is

Table 3.2: *Results for birth injury in the planning component of the split sample. The table counts discordant pairs in which exactly one baby in the pair was injured. Factor 1 contrasts the affected years (1997-1999) with hospital closures in Philadelphia to the base years (1995-1996) without closures. Factor 2 looks within the affected years (1997-1999) and contrasts zip codes with (W) closures to zip codes without (W/O) closures. The bias test contrasts the same zip codes, but in the years (1995-1996) prior to closures, so a difference there cannot be an effect caused by hospital closures, and would instead indicate a failure to control some unmeasured bias. The P-values and odds ratios are from Gart's (1969) procedure.*

A comparison focused on the most affected groups				
Birth Outcomes in Discordant Pairs		Zip Codes With Closures 1995-1999		
Philadelphia Baby	Control Baby	Affected 1997-1999	Base 1995-1996	Total (+)
Injured	Not Injured	52	8	60
Not Injured	Injured	12	11	23
	Total (+)	64	19	83
Odds Ratio		5.80		
Alternative		1-sided		
P-value		0.0016		
95% Interval		$[2.03, \infty)$		

		Factor 1			Factor 2			Bias Test		
Birth Outcome		1995-1999			1997-1999			1995-1996		
Discordant Pairs		Time Period			Zip Code Closures			Zip Code Closures		
Philadelphia Baby	Control Baby	Affected 97-99	Base 95-96	+	W	W/O	+	W	W/O	+
Injured	Not Injured	141	43	184	52	89	141	8	35	43
Not Injured	Injured	43	42	85	12	31	43	11	31	42
	Total (+)	184	85	269	64	120	184	19	66	85
Odds Ratio		3.19			1.51			0.65		
Alternative		1-sided			1-sided			2-sided		
P-value		0.000023			0.19			0.44		
95% Interval		$[1.95, \infty)$			$[0.76, \infty)$			[0.20, 2.03]		

to be found, it will be found here.

Gart (1969) proposed an analysis for a randomized, two-period cross-over experiment with a binary outcome which we generalize for use here. His analysis is suggested by a logit model with additive pair and time effects plus a treatment effect. In such a model, the nuisance parameters are eliminated by conditioning on sufficient statistics, so that the treatment effect is tested by comparing two sets of discordant matched pairs to the hypergeometric distribution in a 2×2 table analysis. In Table 3.2, we perform this analysis several times, and in Section 3.4 we examine the analysis in the context of a non-randomized observational study and generalize it to permit a sensitivity analysis. Happily, after a few steps, the sensitivity analysis for binary difference-in-differences turns out to be an almost standard sensitivity analysis for a 2×2 table, so the situation in observational studies develops in parallel with Gart's (1969) analysis for a randomized cross-over study. There is, however, a curious transformation of the magnitude of the sensitivity parameter; see Proposition 3.4.1.

Judged by Gart's test, the increase in risk of birth injury in " factor 1"in the planning sample is significantly different from an odds ratio of 1, with one-sided significance level 0.000023 and one-sided 95% confidence interval $[1.95, \infty)$. In the planning sample alone, if one did a Bonferroni correction for 228 two-sided tests, the significance level would be approximately 0.01.

In Table 3.2, factor 2 looks just at the years of closures, 1997-1999, and contrasts zip codes near closures in 1997-1999 to zip codes remote from closures. As mentioned in Section 3.1.2, the overcrowding did not occur at the closed obstetrics units but at the ones that remained open, so mothers in zip codes remote from closures may have been affected by sharing an overcrowded obstetrics unit with mothers who came from zip codes with closures. On the other hand, mothers in zip codes with closures faced a newly lengthened trip to the obstetrics unit and may have been unexpected there. In any event, factor 2 is another difference-in-difference analysis in the manner of Gart (1969) but now contrasting Philadelphia-control pairs for zip codes near closures to pairs for zip codes remote from closures. The odds ratio is 1.51, consistent with increased risk, but it does not differ significantly from 1 in this 10% planning sample. The panel labeled " bias test" in Table 3.2 is the same comparison but done in the years before closures: any systematic difference here could not be an effect of the closures and must reflect some uncontrolled bias. The odds ratio is 0.65 and is not significantly different from 1 in this 10% planning sample.

The analysis for the most affected group in Table 3.2 looks just at zip codes near closures, comparing 1997-1999 to 1995-1996. It is in this comparison that we might anticipate the largest effect. The odds ratio is 5.8 with a one-sided 95% confidence interval of $[2.03, \infty)$. Because this is one of the largest of hundreds of estimated odds ratios in the 10% planning sample, we have reason to suspect that it is biased upwards; nonetheless, this seems like a promising comparison to make in the independent 90% analysis sample which will be examined in Section 3.5.

There is an important difference between, on the one hand, factors 1 and 2 and, on the other hand, the analysis of the most affected groups. Factors 1 and 2 are not redundant; indeed, they are nearly independent tests when the hypothesis of no treatment effect is true, that is, they are approximate evidence factors. If the null hypothesis of no effect were true, then exact evidence factors would be statistically independent (Rosenbaum 2010c) and, strictly speaking, factors 1 and 2 in Table 3.2 do not qualify; however, they are nearly independent and so are approximate evidence factors (Rosenbaum 2011, Lemma 4 and Section 7). Moreover, the unmeasured biases that affect these two comparisons are different — in factor 1, unmeasured ways Philadelphia changed over time differently than control-Philadelphia, in factor 2 unmeasured ways that the difference between Philadelphia moms and controls in zip codes with closures in 1997-1999 differed from the pairs for zip codes without closures. In this sense, the two factors are providing separate, not redundant, information about birth injuries possibly caused by abrupt hospital closures. In contrast, the most affected analysis in Table 3.2 is heavily redundant with the other two analyses; it expresses the same evidence in a different way.

What does it mean to say that two evidence factors are " nearly independent" ?It means that under the null hypothesis, the two P-values for the two factors are stochastically larger than the uniform distribution on the unit square, so viewing them as independent P-values would not lead to inflation of the type-1 error rate. For example, in a 2×3 contingency table, the null hypothesis of independence may be tested by computing a chi-square for independence with one degree of freedom comparing column one to the total of columns two and three, and another chi-square for independence comparing columns two and three (Lancaster 1949, expression 18). These two P-values are not independent, because the second column of the first table is the marginal row total of the second table; however, the pair of resulting P-values are stochastically larger than uniform under the null hypothesis of independence. For detailed discussion of approximate evidence factors together with associated sensitivity analyses, see Rosenbaum (2011).

It was a given that we would look at infant mortality, so that decision was made without reference to the planning sample, and the entire data set was used. Although we do not present that analysis here, it is worth mentioning that for death there were no significant differences in the four analyses that parallel Table 3.2 and the point estimates suggest that nothing dramatic had occurred.

3.4 Observational studies with binary outcome and difference-in-differences

3.4.1 Notation: base and intervention periods; exposed and unexposed regions

There are I pairs, $i = 1, \ldots, I$, of two mothers, $k = 1, 2$, who gave birth in the same year, one giving birth in Philadelphia, denoted $Z_{ik} = 1$, the other giving birth elsewhere, denoted $Z_{ik} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. The mothers have been matched for an observed covariate $\mathbf{x}_{ik}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but there is concern also about an unobserved covariate u_{ik} that was not matched, so possibly $u_{i1} \neq u_{i2}$. Because we match for year of birth, year is included in $\mathbf{x}_{ik}$.

In using mothers outside Philadelphia as controls for mothers inside Philadelphia, we are contemplating what would have happened to paired mothers had they interchanged roles, the Philadelphia mother living and delivering in Pittsburgh, say, and the Pittsburgh mother with whom she is paired delivering in Philadelphia. That is to say, each mother (or her newborn baby) has two potential binary responses, r_{Tik} if mother ik delivered in Philadelphia or r_{Cik} if mother ik delivered elsewhere; see Neyman (1923) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : r_{Tik} = r_{Cik}$ for $i = 1, \ldots, I$,

$k = 1, 2$. In Table 3.2, (r_{Tik}, r_{Cik}) refers to birth injury of type ICD-9 767.3, and $(r_{Tik}, r_{Cik}) = (1, 0)$ indicates that baby ik would have experienced a birth injury in Philadelphia but not in, say, Pittsburgh. Under Fisher's H_0, $(r_{Tik}, r_{Cik}) = (0, 0)$ or $(r_{Tik}, r_{Cik}) = (1, 1)$, so some babies had birth injuries and others did not, but changing where mother ik delivered would not change whether a birth injury occurred. Write $R_{ik} = Z_{ik}\, r_{Tik} + (1 - Z_{ik})\; r_{Cik}$ for the observed response of mother ik. Also, write $\mathcal{F} = \{(r_{Tik}, r_{Cik}, \mathbf{x}_{ik}, u_{ik})\,,\, i = 1, \ldots, I,\, k = 1, 2\}$.

3.4.2 Model for sensitivity analysis

Even if Fisher's null hypothesis H_0 were true, birth outcomes might be different in Philadelphia and elsewhere because mothers in Philadelphia differ from mothers elsewhere. This may be expressed in terms of a model that relates delivery in Philadelphia to characteristics of mothers and their neighborhoods in $\mathcal{F}$. This model begins by describing the situation prior to matching. The model says that prior to matching, the Z_{ik} were conditionally independent given $\mathcal{F}$with

$$\Pr\left(Z_{ik} = 1 |\, \mathcal{F}\right) = \frac{\exp\left\{\kappa\left(\mathbf{x}_{ik}\right) + \gamma u_{ik} + \varrho r_{Cik}\right\}}{1 + \exp\left\{\kappa\left(\mathbf{x}_{ik}\right) + \gamma u_{ik} + \varrho r_{Cik}\right\}},\ 0 \leq u_{ik} \leq 1 \qquad (3.4.1)$$

where $\kappa\left(\cdot\right)$ is an unknown function. In (3.4.1), by Bayes theorem, the term $\kappa\left(\mathbf{x}_{ik}\right)$ permits the distribution of observed covariates $\mathbf{x}_{ik}$ in Philadelphia to differ from the distribution among potential controls before matching, as indeed is seen to be the case in Table 3.1; moreover, because year is in $\mathbf{x}_{ik}$, (3.4.1) permits this difference in observed covariates to be different in different years.

In (3.4.1), if $\varrho \neq 0$ then the response r_{Cik} the mother or baby would exhibit outside Philadelphia is related to whether the mother delivers in Philadelphia; that is, by Bayes theorem under (3.4.1), birth injuries may be more or less common in Philadelphia than elsewhere. A bias of the form $\varrho \neq 0$ would be the worst type of bias if one were comparing Philadelphia to matched control, but the study compares Philadelphia in two time periods to control in two time periods, and for this comparison $\varrho \neq 0$ is less of a problem. Of course, we cannot estimate ϱ because we observe R_{ik} not r_{Cik}; in particular, we never observe r_{Cik} when $Z_{ik} = 1$, so we could not fit (3.4.1) even if we somehow knew that $\gamma = 0$.

If $\gamma \neq 0$ in (3.4.1), then the unobserved (and hence unmatched) covariate u_{ik} is related to whether a mother delivers in Philadelphia. Because $0 \leq u_{ik} \leq 1$ in (3.4.1), two mothers ik and ik' with $(\mathbf{x}_{ik}, r_{Cik}) = (\mathbf{x}_{ik'}, r_{Cik'})$ may differ in their odds of delivering in Philadelphia by a factor of at most $\Gamma = \exp(\gamma)$ because u_{ik} and $u_{ik'}$ differ. Because u_{ij} is otherwise unconstrained, it may be different in Philadelphia and control in a different way before and after hospital closures. The term γu_{ik} with $0 \leq u_{ik} \leq 1$ introduces a bias of entirely unspecified form but of a magnitude determined by the magnitude of the sensitivity parameter Γ.

To aid interpretation, it is sometimes convenient to unpack the single parameter Γ into two parameters (Δ, Λ) as $\Gamma = (1 + \Delta\Lambda) / (\Delta + \Lambda)$ where Λ controls the relationship between $u_{i1} - u_{i2}$ and $Z_{i1} - Z_{i2}$ and Δ controls the relationship between $u_{i1} - u_{i2}$ and $r_{Ci1} - r_{Ci2}$. Here, $Y_{Ci} = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$ is 1 if the Philadelphia baby would have had a birth injury if delivery had occurred outside Philadelphia but the control would not, $Y_i = -1$ if the situation were reversed, and $Y_i = 0$ if both babies would have had the same outcome outside Philadelphia. If $\varrho = 0$ so that McNemar's test may be used in a sensitivity analysis comparing Philadelphia babies to controls, a value of $\Gamma = 1.25$ unpacks into the curve $1.25 = (1 + \Delta\Lambda) / (\Delta + \Lambda)$, which includes, for example, $(\Delta, \Lambda) = (2, 2)$ for a u_{ik} that doubles the odds of delivering in Philadelphia and doubles the odds of a birth injury, but it also includes $(\Delta, \Lambda) = (1.4, 5)$ and $(\Delta, \Lambda) = (5, 1.4)$. Analogously, $\Gamma = 2$ unpacks into $(\Delta, \Lambda) = (3, 5)$ and $(\Delta, \Lambda) = (5, 3)$ and other values on the curve $\Gamma = (1 + \Delta\Lambda) / (\Delta + \Lambda)$. For discussion of various aspects of this interpretation of the magnitude of Γ, see Gastwirth, Krieger and Rosenbaum (1998, Section 2) and Rosenbaum and Silber (2009a).

Our analysis eliminates ϱ in (3.4.1) as a nuisance parameter; see Proposition 3.4.1. In one sense the value of ϱ does matter because it affects the patterns of data we see, but in another sense it does not matter because no matter what value ϱ takes on, the difference-in-differences analysis will fully account for it. Because of this and because (3.4.1) is linear in u_{ik} and r_{Cik} on the logit scale, we may assume without loss of generality that the unobserved covariate, u_{ik}, is uncorrelated with birth injuries in the absence of closures, r_{Cik}, because if this were not the case, we could replace u_{ik} by its least squares residual $\breve{u}_{ik} = u_{ik} - (\vartheta + \eta r_{Cik})$, so $\breve{u}_{ik}$ and r_{Cik} are uncorrelated, and $\kappa(\mathbf{x}_{ik}) + \gamma u_{ik} + \varrho r_{Cik}$ in (3.4.1) equals $\{\kappa(\mathbf{x}_{ik}) + \vartheta\} + \gamma \breve{u}_{ik} + (\varrho + \eta) r_{Cik}$. In other words, an unobserved covariate u_{ik} cannot bias the analysis by virtue of being related to birth injuries; it must instead in Factor 1 be related to birth injuries in a different way in different years, or in Factor 2 it must be related to birth injuries in a different way in different zip codes. Although this appears to be an attractive feature of the difference-in-differences analysis, there is a nontrivial price to be paid for it. If ϱ were known to be zero, then Philadelphia and control-Philadelphia could be compared directly, say using McNemar's test for binary responses in matched pairs, and the bias from u_{ik} would be of magnitude γ on the logit scale or $\Gamma = \exp(\gamma)$ in terms of odds; see Rosenbaum (2002, Section 4.3.2). In contrast, although the difference-in-differences analysis may take u_{ik} to be uncorrelated with r_{Cik}, the analysis faces a bias from u_{ik} of magnitude 2γ on the logit scale or $\Theta = \Gamma^2 = \exp(2\gamma)$ in terms of odds; again, see Proposition 3.4.1. In brief, the difference-in-difference analysis is completely unaffected by certain unmeasured biases perfectly correlated with r_{Cik}, but is twice as sensitive to certain other unmeasured biases uncorrelated with r_{Cik}. A mathematically distinct yet conceptually related phenomenon has been noted previously, with difference-in-differences studies being more severely affected by errors-of-measurement (Freeman 1984, Griliches and Hausman1986).

After matching for $\mathbf{x}_{ik}$, so that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ and $Z_{i1} + Z_{i2} = 1$, the model (3.4.1) implies

$$\Pr\left(Z_{i1} = 1 \middle| \mathcal{F}, Z_{i1} + Z_{i2} = 1\right) = \frac{\exp(\gamma u_{i1} + \varrho r_{Ci1})}{\exp(\gamma u_{i1} + \varrho r_{Ci1}) + \exp(\gamma u_{i2} + \varrho r_{Ci2})}. \quad (3.4.2)$$

In particular, (3.4.2) is $\frac{1}{2}$ if $\gamma = \varrho = 0$, but otherwise treatment assignment is biased.

An alternative but nearly equivalent formulation of the model would omit reference to the population prior to matching — that is, omit reference to (3.4.1) — and take (3.4.2) as the starting point, that is, take (3.4.2) as a model for treatment assignment Z_{ik} within a given matched pair i. Our sense is that the step from

(3.4.1) to (3.4.2) is useful in making it clear what matching for $\mathbf{x}_{ik}$ does and what it fails to do. There is, however, one advantage in beginning with (3.4.2). Once a matched pair is formed, there is one Philadelphia zip code attached to that pair, and by including that zip code in $\mathcal{F}$ as an attribute of the pair i (not the mother k), we may understand (3.4.2) as a model for the identity k of the Philadelphia mother in pair i. That is, in this formulation, (3.4.2) asks: Given that pair i contains two mothers, one from Philadelphia zip-code xxxxx and the other from a zip code with similar attributes elsewhere in Pennsylvania, California or Missouri, and given specific values of (u_{i1}, r_{Ci1}) and (u_{i2}, r_{Ci2}) for these two mothers, what is the chance that mother $i1$ is the Philadelphia mother and $i2$ is the mother from elsewhere? This distinction between starting with (3.4.1) and starting with (3.4.2) is relevant only to comparisons of pairs with a zip code near a hospital closure versus pairs with a zip code remote from closures — in such comparisons, zip code is treated as a fixed attribute of the pair, as year is treated as a fixed attribute of the pair in temporal comparisons.

3.4.3 Sensitivity analysis with binary outcomes in difference-in-differences analysis

We wish to focus on a set $\mathcal{S} \subseteq \{1, \ldots, I\}$ of the pairs, and to contrast two subsets of the pairs in $\mathcal{S}$, denoted by $v_i = 1$ and $v_i = 0$. In the first evidence factor in Table 3.2, all pairs are used, $\mathcal{S} = \{1, \ldots, I\}$, and $v_i = 1$ for birth pairs in years 1997-1999 and $v_i = 0$ for pairs in 1995-1996. In the second evidence factor in Table 3.2,

$\mathcal{S} \subset \{1, \ldots, I\}$ are the pairs in 1997-1999, and $v_i = 1$ for pairs with a Philadelphia mother in a zip code near a closure and $v_i = 0$ for pairs with a Philadelphia mother not near a closure.

Consider testing Fisher's null hypothesis $H_0 : r_{Tik} = r_{Cik}$ using the conditional distribution of $T' = \sum_{i \in \mathcal{S}} \sum_{k=1}^{2} v_i \, Z_{ik} \, R_{ik}$ given $W' = \sum_{i \in \mathcal{S}} \sum_{k=1}^{2} Z_{ik} \, R_{ik}$. In the first evidence factor in Table 3.2, this is the conditional distribution of T', the number of birth injuries in Philadelphia during the years 1997-1999 of abrupt closures, given the total W' of birth injuries in Philadelphia in all years 1995-1999. If H_0 is true, then $r_{Tik} = r_{Cik} = R_{ik}$, and T' and W' receive only constant contributions from concordant pairs with $0 = R_{i1} - R_{i2} = r_{Ci1} - r_{Ci2}$. Renumber the pairs so that pairs $j = 1, \ldots, J$ are both in $\mathcal{S}$ and are discordant pairs in the sense that $R_{j1} \neq R_{j2}$, and pairs $j+1, \ldots, I$ are either not in $\mathcal{S}$ or are concordant pairs with $R_{j1} = R_{j2}$. Let $T = \sum_{j=1}^{J} \sum_{k=1}^{2} v_j \, Z_{jk} \, R_{jk}$ and $W = \sum_{j=1}^{J} \sum_{k=1}^{2} Z_{jk} \, R_{jk}$ and notice that, given $\mathcal{F}$ and $Z_{i1} + Z_{i2} = 1$, $i = 1, \ldots, I$, they differ from T' and W' by a constant when H_0 is true. Write $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{J2})^T$ and $\mathbf{r}_C = (r_{C11}, r_{C12}, \ldots, r_{CJ2})^T$ for the $2J$-dimensional vectors, and write $\mathcal{Z}$ for the set containing the 2^J vectors $\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{J2})^T$ with each $z_{jk} = 0$ or $z_{jk} = 1$ and $z_{j1} + z_{j2} = 1$. With a slight abuse of notation, conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ will be abbreviated to conditioning on $\mathcal{Z}$. Write $v_+ = \sum_{j=1}^{J} v_j$.

In Proposition 3.4.1, the case (3.4.5) of $\Gamma = 1$ is essentially due to Gart (1969). In (3.4.4) conditioning on W has eliminated the potential bias in (3.4.2) from ϱr_{Ci1}, leaving only the potential bias from γu_{i1}.

Table 3.3: *General form of the table under H_0 after renumbering within the J discordant pairs so that $r_{Cj1} = 1$ and $r_{Cj2} = 0$ for each j.*

	$v_j = 1$	$v_j = 0$	Total
$z_{j1} = 1$	$\sum_{j=1}^{J} v_j z_{j1}$	$\sum_{j=1}^{J} (1 - v_j) z_{j1}$	w
$z_{j1} = 0$	$\sum_{j=1}^{J} v_j (1 - z_{j1})$	$\sum_{j=1}^{J} (1 - v_j) (1 - z_{j1})$	$J - w$
Total	v_+	$J - v_+$	J

Proposition 3.4.1. *Let $\Theta = \Gamma^2$. Under H_0 and the sensitivity model (3.4.1),*

$$\Upsilon\left(J,\, w,\, v_+,\, t,\, \frac{1}{\Theta}\right) \leq \Pr\left(T \geq t \middle|\, \mathcal{F},\, \mathcal{Z},\, W = w\right) \leq \Upsilon\left(J,\, w,\, v_+,\, t,\, \Theta\right) \qquad (3.4.3)$$

where

$$\Upsilon\left(J,\, w,\, v_+,\, t,\, \Theta\right) = \frac{\sum_{k=\max(t,w+v_+-J)}^{\min(w,v_+)} \binom{v_+}{k}\binom{J-v_+}{w-k}\Theta^k}{\sum_{k=\max(0,w+v_+-J)}^{\min(w,v_+)} \binom{v_+}{k}\binom{J-v_+}{w-k}\Theta^k} \qquad (3.4.4)$$

is the extended hypergeometric distribution. In particular, if $\gamma = 0$ in (3.4.1), so that $\Gamma = 1$, then

$$\Pr\left(T \geq t \middle|\, \mathcal{F},\, \mathcal{Z},\, W = w\right) = \sum_{k=\max(t,w+v_+-J)}^{\min(w,v_+)} \frac{\binom{v_+}{k}\binom{J-v_+}{w-k}}{\binom{J}{w}} \qquad (3.4.5)$$

is the hypergeometric distribution.

Proof. The proof consists in transforming a sensitivity analysis for 2 × 2 tables counting discordant pairs, such as the 2 × 2 tables in Table 3.2, into a sensitivity analysis for unrelated events in 2 × 2 tables, and then applying standard methods for the latter situation. Throughout the proof, assume H_0 is true for the purpose of testing it, so $r_{Tik} = r_{Cik} = R_{ik}$. Using (3.4.2), we have

$$\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathcal{F},\, \mathbf{Z} \in \mathcal{Z}\right) = \frac{\exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} u_{jk} + \varrho \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} r_{Cjk}\right)}{\prod_{j=1}^{J} \left\{\exp\left(\gamma u_{j1} + \varrho r_{Cj1}\right) + \exp\left(\gamma u_{j2} + \varrho r_{Cj2}\right)\right\}}. \tag{3.4.6}$$

Let $\mathcal{Z}_w = \left\{\mathbf{z} \in \mathcal{Z} : w = \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} r_{Cjk}\right\}$. Then $|\mathcal{Z}_w| = \binom{J}{w}$. Conditioning on $W = w$ or equivalently on $\mathbf{Z} \in \mathcal{Z}_w$ yields

$$\Pr\left(\mathbf{Z} = \mathbf{z} \middle| \mathcal{F},\, \mathbf{Z} \in \mathcal{Z}_w\right) = \frac{\exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} u_{jk}\right)}{\sum_{\mathbf{b} \in \mathcal{Z}_w} \exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} b_{jk} u_{jk}\right)}$$

which no longer depends upon ϱ. Because the J pairs are discordant, $1 = |r_{Cj1} - r_{Cj2}|$ for every j, we may without loss of generality renumber the two subjects in each pair j so that $r_{Cj1} = 1$ and $r_{Cj2} = 0$; then $v_j \sum_{k=1}^{2} z_{jk} r_{Cjk} = v_j z_{j1}$ and $T = \sum_{j=1}^{J} v_j z_{j1}$ and $W = \sum_{j=1}^{J} z_{j1}$; see Table 3.3. Also, write $\widetilde{u}_j = u_{j1} - u_{j2}$, so that $-1 \leq \widetilde{u}_j \leq 1$. Define the J-dimensional vectors $\widetilde{\mathbf{u}} = \left(\widetilde{u}_1, \ldots, \widetilde{u}_J\right)^T$, $\mathbf{v} = \left(v_1, \ldots, v_J\right)^T$ and $\mathbf{1} = \left(1, \ldots, 1\right)^T$. Let $\chi(A) = 1$ if event A occurs and $\chi(A) = 0$ otherwise. Then using $\sum_{k=1}^{2} z_{jk} u_{jk} = u_{j2} + z_{j1}\left(u_{j1} - u_{j2}\right)$ and simplifying

$$\begin{aligned} &\Pr\left(T \geq t \middle| \mathcal{F},\, \mathbf{Z} \in \mathcal{Z}_w\right) \\ =&\frac{\sum_{\mathbf{z} \in \mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j \sum_{k=1}^{2} z_{jk} r_{Cjk} \geq t\right) \exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} z_{jk} u_{jk}\right)}{\sum_{\mathbf{b} \in \mathcal{Z}_w} \exp\left(\gamma \sum_{j=1}^{J} \sum_{k=1}^{2} b_{jk} u_{jk}\right)} \\ =&\frac{\sum_{\mathbf{z} \in \mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j z_{j1} \geq t\right) \exp\left(\gamma \sum_{j=1}^{J} z_{j1} \widetilde{u}_j\right)}{\sum_{\mathbf{b} \in \mathcal{Z}_w} \exp\left(\gamma \sum_{j=1}^{J} b_{j1} \widetilde{u}_j\right)} = \lambda_t\left(\widetilde{\mathbf{u}}\right), \text{ say.} \end{aligned} \tag{3.4.7}$$

Then to prove (3.4.3) it suffices to show

$$\lambda_t\left(\mathbf{1}-2\mathbf{v}\right) \leq \lambda_t\left(\widetilde{\mathbf{u}}\right) \leq \lambda_t\left(2\mathbf{v}-\mathbf{1}\right), \tag{3.4.8}$$

because $w = \sum_{j=1}^{J} z_{j1}$ is fixed for $\mathbf{z} \in \mathcal{Z}_w$, so that, for example,

$$\begin{aligned} \lambda_t\left(2\mathbf{v}-\mathbf{1}\right) &= \frac{\sum_{\mathbf{z}\in\mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j z_{j1} \geq t\right) \exp\left\{\gamma \sum_{j=1}^{J} z_{j1}\left(2v_j-1\right)\right\}}{\sum_{\mathbf{b}\in\mathcal{Z}_w} \exp\left\{\gamma \sum_{j=1}^{J} b_{j1}\left(2v_j-1\right)\right\}} \\ =& \frac{\sum_{\mathbf{z}\in\mathcal{Z}_w} \chi\left(\sum_{j=1}^{J} v_j z_{j1} \geq t\right) \exp\left(2\gamma \sum_{j=1}^{J} z_{j1} v_j\right)}{\sum_{\mathbf{b}\in\mathcal{Z}_w} \exp\left(2\gamma \sum_{j=1}^{J} b_{j1} v_j\right)} = \Upsilon\left(J,\, w,\, v_+,\, t,\, \Gamma^2\right). \end{aligned} \tag{3.4.9}$$

The proof of (3.4.8) is identical to the proof of Proposition 1 in Rosenbaum (1995), except in that proof, $0 \leq u_j \leq 1$ whereas here $-1 \leq \widetilde{u}_j \leq 1$, so the upper bound in (3.4.3) is attained with $\widetilde{u}_j = 2v_j - 1$ rather than with $u_j = v_j$ (or with $u_j = r_j$ in the notation of that proof). □

3.5 Confirmatory analysis using the 90% sample

Table 3.4 is for the analysis sample of 90% of pairs but is otherwise parallel to Table 3.2 for the 10% planning sample. The initial impression of Table 3.4 is that it exhibits many of the same patterns as Table 3.2, albeit sometimes in a more muted form. For instance, in Table 3.2, the odds ratio for the most affected groups was 5.80, whereas in Table 3.4 it is 2.19. This is not surprising given that Table 3.2 was selected as the most promising of many possible analyses, while Table 3.4

Table 3.4: *Results for birth injury in the analysis component of the split sample. This table, which is the basis for conclusions rather than hypothesis generation, has the same structure as Table 3.2 but is based on an independent sample of pairs that is approximately nine times larger.*

A comparison focused on the most affected groups				
Birth Outcomes in Discordant Pairs		Zip Codes With Closures 1995-1999		
Philadelphia Baby	Control Baby	Affected 1997-1999	Base 1995-1996	Total (+)
Injured	Not Injured	475	131	606
Not Injured	Injured	137	83	220
	Total (+)	612	214	826
Odds Ratio		2.19		
Alternative		1-sided		
P-value		3.71×10^{-6}		
95% Interval		[1.63, ∞)		

		Factor 1			Factor 2			Bias Test		
Birth Outcome		1995-1999			1997-1999			1995-1996		
Discordant Pairs		Time Period			Zip Code			Zip Code		
Philadelphia	Control	Affected	Base		Closures			Closures		
Baby	Baby	97-99	95-96	+	W	W/O	+	W	W/O	+
Injured	Not Injured	1231	505	1736	475	756	1231	131	374	505
Not Injured	Injured	514	339	853	137	377	514	83	256	339
	Total (+)	1745	844	2589	612	1133	1745	214	630	844
Odds Ratio		1.61			1.73			1.08		
Alternative		1-sided			1-sided			2-sided		
P-value		4.37×10^{-8}			9.33×10^{-7}			0.69		
95% Interval		[1.39, ∞)			[1.42, ∞)			[0.78, 1.51]		

is an independent replication of that one most promising analysis. As in Table 3.2, Table 3.4 provides several pieces of information consistent with an increase in birth injuries caused by abrupt hospital closures. First, in Factor 1, there is an increase from 1995-1996 to 1997-1999 in the relative frequency of birth injuries in Philadelphia when contrasted with control-Philadelphia. Second, in Factor 2, in the years 1997-1999, there is a greater excess of birth injuries in zip codes near hospital closures than in zip codes remote from hospital closures when contrasted with matched pairs in control-Philadelphia. The test for bias looks at these same zip code groups but in the years before closures, yielding an odds ratio of 1.08 which does not differ significantly from 1. That is to say, zip codes with closures look different after the closures but did not look different before the closures. These pieces of information are not greatly redundant with each other; that is, the first two pieces are approximate evidence factors. The most affected group contrasts zip codes near closures in 1995-1996 to 1997-1999 to matched controls in control-Philadelphia; this yields the largest estimated odds ratio of 2.19. In the absence of bias from unmeasured covariates, this would suggest roughly a doubling of the odds of birth injuries in the affected regions of Philadelphia during the period of abrupt closures.

Unlike Factor 1 in Table 3.2, in Table 3.4 there is strong evidence that birth injuries were more common in Philadelphia than in control-Philadelphia in 1995-1996 when there were no closures. Specifically, if McNemar's test is applied to the $844 = 505 + 339$ pairs discordant for birth injury in 1995-1996, the two-sided P-value is 1.2×10^{-8}. Expressed in terms of (3.4.1), it appears that $\varrho \neq 0$, so the elimination of ϱ by conditioning is essential. We could not reasonably apply McNemar's test to the $1745 = 1231 + 514$ discordant pairs in 1997-1999, because the comparison in 1995-1996 suggests that at least part of the difference in birth injuries in 1997-1999 was already present in 1995-1996 when there were no closures.

Table 3.5 is the sensitivity analysis based on Table 3.4 using Proposition 3.4.1. Table 3.5 eliminates ϱ by conditioning and worries about an unobserved covariate u_{ik} uncorrelated with birth injuries in the absence of closures, r_{Cik}, but possibly related to changes or differences in the frequencies of birth injuries. In Table 3.5, the analysis is reported in terms of Γ, but from Proposition 3.4.1 the sensitivity bound is calculated using the extended hypergeometric distribution with parameter $\Theta = \Gamma^2$.

Table 3.5: *Sensitivity analysis in the 90% analysis sample. The table gives the upper bound on the one-sided P-value testing the null hypothesis of no effect of closures on birth injuries for the three effect comparisons in Table 3.4 for departures from random assignment of various magnitudes* Γ.

Γ	Upper bound on 1-sided P-value		
	Most Affected	Factor 1	Factor 2
1.0	0.0000	0.0000	0.0000
1.1	0.0003	0.0007	0.0012
1.15	0.0019	0.0145	0.0118
1.2	0.0083	0.1126	0.0636
1.25	0.0277	0.3892	0.2066
1.3	0.0730	0.7301	0.4445

Birth injuries were more common in Philadelphia than among matched controls even before Philadelphia hospitals began to close their obstetrics units; however, there was a substantial increase in the relative frequency of birth injuries during the years 1997-1999 of abrupt closures, and this increase was substantially more pronounced in zip codes served by hospitals that closed. Moreover, zip codes served by hospitals that closed did not exhibit any relative excess of birth injuries in the years 1995-1996 prior to closures. A moderate bias from an unobserved covariate u_{ik} of magnitude $\Gamma = 1.3$ (or $\Lambda = 2$ and $\Delta = 2.3$ in Section 3.4.2) could produce any one of these associations, but this u_{ik} would need to be somewhat unusual: it would need to be uncorrelated with birth injuries r_{Cik} (see Section 3.4.2) yet strongly correlated with the change in birth injuries over time and with the post-closure difference in zip codes with closures. Such unobserved covariate is logically possible, but is rendered somewhat less plausible by the need to explain the results in factor 1, factor 2 and the bias test, no one of which is redundant with another.

Table 3.4 permits two other informative analyses. Although one expects an effect of closures in zip codes with closures, as discussed earlier it is less clear what one should expect for mothers living in zip codes without closures. Comparing pairs discordant for birth injuries in zip codes without closures in 1997-1999 and 1995-1996, the point estimate of the odds ratio is 1.35 with 95% confidence interval [1.11, 1.63], suggesting a small increase in birth injuries for mothers in zip codes without closures. In addition, in the $2 \times 2 \times 2$ table in Table 3.4 recording pairs discordant for birth injuries, time interval, and with or without closures, the three factor interaction in a log-linear model is not plausibly zero, with likelihood ratio chi-square of 6.27 on 1 degree of freedom, P-value $= 0.012$, so the increase in birth injuries appears to have been larger in zip codes with closures than in zip codes without closures. This pattern of results is not inconsistent with overcrowding at the hospitals that remained open, with mothers remote from the closures being nonetheless affected by the influx of mothers from zip codes with closures.

Chapter 4

Conclusion and Discussion

4.1 Conclusion and Discussion of the PoSI Approach

In Chapter 1, we investigated the Post-Selection Inference or "PoSI" problem for linear models whereby valid statistical tests and confidence intervals are sought after variable selection, that is, after selecting a subset of the predictors in a data-driven way. We adopted a framework that does *not* assume any of the linear models under consideration to be correct. We allowed the response vector to be centered at an arbitrary mean vector but with homoscedastic and Gaussian errors. We further allowed the full predictor matrix $\mathbf{X}_{n \times p}$ to be rank-deficient, $d = \text{rank}(\mathbf{X}) < p$, and we also allowed the set $\mathcal{M}$ of models M under consideration to be largely arbitrary. In this framework we showed that valid post-selection inference is possible via simultaneous inference. An important enabling factor is the principle that the regression coefficient of a given predictor as distinct when it occurs in different submodels: $\beta_{j\cdot\mathrm{M}}$ and $\beta_{j\cdot\mathrm{M}'}$ are generally different parameters if $\mathrm{M}\neq\mathrm{M}'$. We showed that simultaneity protection for all parameters $\beta_{j\cdot\mathrm{M}}$ provides valid post-selection inference. In practice this means enlarging the constant $t_{1-\alpha/2,r}$ used in conventional inference to a constant $K(\mathbf{X}_{n\times p}, \alpha, r)$ that provides simultaneity protection for up to $p\,2^{p-1}$ parameters $\beta_{j\cdot\mathrm{M}}$. We showed that the constant depends strongly on the predictor matrix $\mathbf{X}$ as the asymptotic bound for $K(\mathbf{X}, \alpha, r)$ with $d = \mathrm{rank}(\mathbf{X})$ ranges between the minimum of $\sqrt{2\log d}$ achieved for orthogonal designs on the one hand, and a large fraction of the Scheffé bound $\sqrt{d}$ on the other hand. This wide asymptotic range suggests that computation is critical for problems with large numbers of predictors. In the classical case $d=p$ our current computational methods are feasible up to about $p\approx 20$.

We carried out post-selection inference in a limited framework. Several problems remain open, and many natural extensions are desirable:

- Among open problems is the quest for the largest fraction of the asymptotic Scheffé rate $\sqrt{d}$ attained by PoSI constants. So far we know this fraction to be at least 0.6363 but no more than 0.8660... in the classical case $d=p$.

- Computations for $p > 20$ are a challenge. Straight enumeration of the set of up to $p\,2^{p-1}$ linear combinations should be replaced with heuristic shortcuts that yield practically useful upper bounds on $K(\mathbf{X}_{n\times p},\ \mathcal{M},\ \alpha, r)$ that are specific to $\mathbf{X}$ and the set of submodels $\mathcal{M}$, unlike the 0.8660 fraction of the Scheffé bound which is universal.

- The methodology is easily adapted to practically useful variations by suitable choice of the set of models $\mathcal{M}$: (1) Data analysts might be interested only in small submodels, $|\mathrm{M}| \leq 5$, say, when p is large. (2) We introduced SPAR ("Single Predictor Adjusted Regression", Section 1.4.8) defined as "significance hunting" or the search for the strongest adjusted "effect" in any predictor. In practice one might be more interested in SPAR1 or the search for strong adjusted effects in one predetermined focal predictor. — Any limitation to a lesser number of submodels or regression coefficients to be searched increases the computationally accessible number of predictors.
- Among models to which the PoSI framework should be extended next are generalized linear models and mixed effects models.

R code for computing the PoSI constant for up to $p = 20$ can be obtained from the authors' webpages (manuscript describing the computations is available from the authors).

4.2 Conclusion and Discussion of the Split Samples Approach

In Chapter 2, we studied the problem of inference after model selection for random-design matrices. In this study, we neither assumed linearity nor homoscedasticity. Instead, we only assumed that the observations are i.i.d. from a fixed dimensional multivariate joint distribution. We showed that under mild conditions on the moments of the joint distribution, we can achieve valid post-selection inference via split samples and bootstrap. The proposed procedure suggested randomly split the data into a model selection sample and an inference sample. The model selection sample was used to choose a submodel by some selection criteria, whereupon the sample was discarded; then, bootstrap inference was produced for the selected set of variables based on the inference sample. We showed that for any explanatory variable in the selected model, its bootstrap confidence interval has proper coverage probability asymptotically. Furthermore, this coverage probability is universally valid for any the model selection rule.

The split-sample approach has been shown to be very effective in protecting valid confirmatory inference from exploratory model building by ruling out conditional inference. However, several fundamental questions are to be answered: (1) What is the “sweet spot” of the proportion of the exploratory sample in maximizing the power of the inference afterwards? (2) We can literally build the model using the exploratory sample by any procedure. What would be a suitable criterion to compare the procedures, and what is the optimal procedure among all of them under such criterion? (3) A potential extension of the split-sample method is to allow sequential splitting, in which we can keep splitting more observations into the exploratory sample, using the new splits for model assessment and model revision, until reaching a satisfactory model. Will there be any general guidance for this method? (4) How does the split-sample method relate to Bayesian methods by viewing the exploratory sample as the a priori information?

4.3 Conclusion and Discussion of the Real Data Application of the Split Samples Method

In Chapter 3, our study built a control Philadelphia with some of the temporal and sociodemographic structure of Philadelphia thereby framing and simplifying questions about how Philadelphia might have changed in the absence of widespread closures of obstetrics units.

Because this series of hospital closures is a unique event, it will never be possible to replicate this study using a new independent sample. Motivated by considerations of improved design sensitivity (Heller et al. 2009), we created an internal replication, a small planning sample of about 13,000 pairs of mothers, and an independent confirmatory analysis sample of about 120,000 pairs. The planning sample suggested a focus on serious birth injuries (ICD-9 767.3), with a relative increase in injuries in the years 1997-1999 of abrupt closures, especially in zip codes served by obstetrics units that abruptly closed. This led to two evidence factors, one test for bias from unmeasured covariates, and a sensitivity analysis.

In a scientific report, what is the appropriate way to report a split sample analysis? In our methodological discussion here, we have focused on one confirmatory analysis. Our sense is that both exploratory and confirmatory analyses should be presented (Tukey 1980b), but that these two types of analyses should be distinguished based on their different histories. That is, a table might present parallel analyses for many interesting outcomes with a bright red line separating confirmatory from exploratory analyses. Above the red line are a few analyses suggested by the planning sample, with independent confirmation or not from the much larger analysis sample. Below the line are exploratory analyses of many outcomes, perhaps aided by some interpretive guidance from multiple testing procedures, such as the Bonferroni inequality, and their associated sensitivity analyses (e.g., Heller et al. 2009, Section 3.3; Rosenbaum and Silber 2009b, Section 4.5). Though perhaps interesting and worthy of further study, hypotheses that are first suggested by the analysis sample or the complete data would inevitably be regarded as speculative unless confirmed by multiple testing procedures.

4.4 Thoughts about Future Research

Statistics is about learning, and statisticians strive to learn from both exploratory and confirmatory analysis of the historical data. My Ph.D. research on valid post-selection inference is aimed at providing a bridge between exploratory data analysis and confirmatory data analysis, and I would like to devote myself to the pursuit of more general and more efficient inference procedures under this framework.

Appendix A

Appendix

A.1 Proofs in Chapter 1

A.1.1 Proof of Theorem 1.4.3

We start with the statement of strong family-wise error control by defining the true null hypotheses and true alternatives for the true $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu}$, as well as the sets of insignificant and significant tests for the observed $\mathbf{Y}$:

$$
\begin{aligned}
H_0 &\triangleq \{\, (j, \mathrm{M}) \mid \beta_{j\cdot\mathrm{M}} = 0,\ j \in \mathrm{M} \in \mathcal{M} \,\}, \\
H_1 &\triangleq \{\, (j, \mathrm{M}) \mid \beta_{j\cdot\mathrm{M}} \neq 0,\ j \in \mathrm{M} \in \mathcal{M} \,\}, \\
\hat{H}_0 &\triangleq \{\, (j, \mathrm{M}) \mid |t^{(0)}_{j\cdot\mathrm{M}}| \leq K(\mathbf{X}, \alpha),\ j \in \mathrm{M} \in \mathcal{M} \,\}, \\
\hat{H}_1 &\triangleq \{\, (j, \mathrm{M}) \mid |t^{(0)}_{j\cdot\mathrm{M}}| > K(\mathbf{X}, \alpha),\ j \in \mathrm{M} \in \mathcal{M} \,\}.
\end{aligned}
$$

where $t^{(0)}_{j\cdot\mathrm{M}} \triangleq \hat{\beta}_{j\cdot\mathrm{M}} / (\hat{\sigma} / \|\mathbf{X}_{j\cdot\mathbf{M}}\|)$ has the parameter set to $\beta_{j\cdot\mathrm{M}} = 0$.

Lemma A.1.1. *"Strong Family-Wise Error Control" holds for* $K(\mathbf{X},\alpha)$*:*

$$\mathbf{P}[H_0 \subset \hat{H}_0] \; = \; \mathbf{P}[H_1 \supset \hat{H}_1] \; \geq \; 1-\alpha.$$

Proof: Standard; just the same: $H_0 \subset \hat{H}_0 \; \Leftrightarrow \; H_1 \supset \hat{H}_1$ implies the equality of the two probabilities. Further, using $t^{(0)}_{j\cdot\mathrm{M}} = t_{j\cdot\mathrm{M}} \Leftrightarrow (j,\mathrm{M}) \in H_0$,

$$\begin{aligned} \mathbf{P}[\, H_0 \subset \hat{H}_0 \,] \; &= \; \mathbf{P}[\max_{(j,\mathrm{M})\in H_0} |t_{j\cdot\mathrm{M}}| \leq K(\mathbf{X},\alpha)\,] \\ &\geq \; \mathbf{P}[\max_{\mathrm{M}\in\mathcal{M}} \max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \leq K(\mathbf{X},\alpha)\,] \; \geq \; 1-\alpha \end{aligned}$$

by the definition of $K(\mathbf{X},\alpha)$ (1.4.11). □

THEOREM 1.4.3. *"Strong Post-Selection Error Control" holds for any model selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}[\forall j \in \hat{\mathrm{M}} : |t^{(0)}_{j\cdot\hat{\mathrm{M}}}| > K(\mathbf{X},\alpha) \; \Rightarrow \; \beta_{j\cdot\hat{\mathrm{M}}} \neq 0\,] \; \geq \; 1-\alpha.$$

Proof: Define $\hat{\mathrm{M}}' \triangleq \{(j,\hat{\mathrm{M}}) \,|\, j \in \hat{\mathrm{M}}\}$. The event $H_1 \supset \hat{H}_1$ implies the event $H_1 \cap \hat{\mathrm{M}}' \supset \hat{H}_1 \cap \hat{\mathrm{M}}'$, hence, using Lemma A.1.1:

$$1-\alpha \; \leq \; \mathbf{P}[\, H_1 \supset \hat{H}_1 \,] \; \leq \; \mathbf{P}[\, H_1 \cup \hat{\mathrm{M}}' \supset \hat{H}_1 \cup \hat{\mathrm{M}}' \,]. \qquad \square$$

A.1.2 Proof of Proposition 1.5.3

1. The matrix $\mathbf{X}^*_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}(\mathbf{X}^T_{\mathrm{M}}\mathbf{X}_{\mathrm{M}})^{-1}$ has the vectors $\boldsymbol{l}_{j\cdot\mathrm{M}}$ as its columns. Thus $\boldsymbol{l}_{j\cdot\mathrm{M}} \in \mathrm{span}(\mathbf{X}_j : j \in \mathrm{M})$. Orthogonality $\boldsymbol{l}_{j\cdot\mathrm{M}} \perp \mathbf{X}_{j'}$ for $j' \neq j$ follows from $\mathbf{X}^T_{\mathrm{M}}\mathbf{X}^*_{\mathrm{M}} = \mathbf{I}_p$. The same properties hold for the normalized vectors $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$.

2. The vectors $\{\bar{\boldsymbol{l}}_{1\cdot\{1\}},\ \bar{\boldsymbol{l}}_{2\cdot\{1,2\}},\ \bar{\boldsymbol{l}}_{3\cdot\{1,2,3\}},\ \ldots,\ \bar{\boldsymbol{l}}_{p\cdot\{1,2,\ldots,p\}}\}$ form a Gram-Schmidt series with normalization, hence they are an o.n. basis of $\mathbb{R}^p$.

3. For $\mathrm{M} \subset \mathrm{M}'$, $j \in \mathrm{M}$, $j' \in \mathrm{M}' \setminus \mathrm{M}$, we have $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \perp \bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}}$ because they can be embedded in an o.n. basis by first enumerating M and subsequently $\mathrm{M}' \setminus \mathrm{M}$, with j being last in the enumeration of M and j' last in the enumeration of $\mathrm{M}' \setminus \mathrm{M}$.

4. For any (j_0, M_0), $j_0 \in \mathrm{M}_0$, there are $(p-1)\,2^{p-2}$ ways to choose a partner (j_1, M_1) such that either $j_1 \in \mathrm{M}_1 \subset \mathrm{M}_0 \setminus j_0$ or $\mathrm{M}_0 \subset \mathrm{M}_1 \setminus j_1$, both of which result in $\bar{\boldsymbol{l}}_{j_0\cdot\mathrm{M}_0} \perp \bar{\boldsymbol{l}}_{j_1\cdot\mathrm{M}_1}$ by the previous part.

A.1.3 Proof of Duality: Lemma 1.5.1 and Theorem1.5.1

The proof relies on a careful analysis of orthogonalities as described in Proposition 1.5.3, part *3.* In what follows we write $[\mathbf{A}]$ for the column space of a matrix $\mathbf{A}$, and $[\mathbf{A}]^\perp$ for its orthogonal complement. We show first that, for $\mathrm{M} \cap \mathrm{M}^* = \{j\}$, $\mathrm{M} \cup \mathrm{M}^* = \mathrm{M}_F$, the vectors $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}$ and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ are in the same one-dimensional subspace, hence are a multiple of each other. To this end we observe:

$$\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in [\mathbf{X}_{\mathrm{M}}]\,, \qquad \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in [\mathbf{X}_{\mathrm{M}\setminus j}]^{\perp}, \tag{A.1.1}$$

$$\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \in [\mathbf{X}^*_{\mathrm{M}^*}]\,, \qquad \bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \in [\mathbf{X}^*_{\mathrm{M}^*\setminus j}]^{\perp}, \tag{A.1.2}$$

$$[\mathbf{X}^*_{\mathrm{M}^*}] = [\mathbf{X}_{\mathrm{M}\setminus j}]^{\perp}, \qquad [\mathbf{X}^*_{\mathrm{M}^*\setminus j}]^{\perp} = [\mathbf{X}_{\mathrm{M}}]\,. \tag{A.1.3}$$

The first two lines state that $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ and $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}$ are in the respective column spaces of their models, but orthogonalized with regard to all other predictors in these models. The last line, which can also be obtained from the orthogonalities implied by $\mathbf{X}^T\mathbf{X}^* = \mathbf{I}_p$, establishes that the two vectors fall in the same one-dimensional subspace:

$$\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in [\mathbf{X}_{\mathrm{M}}] \cap [\mathbf{X}_{\mathrm{M}\setminus j}]^{\perp} \;=\; [\mathbf{X}^*_{\mathrm{M}^*}] \cap [\mathbf{X}^*_{\mathrm{M}^*\setminus j}]^{\perp} \ni \bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}.$$

Since they are normalized, it follows $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} = \pm\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$. This result is sufficient to imply all of Theorem 1.5.1. The lemma, however, makes a slightly stronger statement involving lengths which we now prove. In order to express $\boldsymbol{l}_{j\cdot\mathrm{M}}$ and $\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}$ according to (1.5.2), we use $\mathbf{P}_{\mathrm{M}\setminus j}$ as before and we write $\mathbf{P}^*_{\mathrm{M}^*\setminus j}$ for the analogous projection onto the space spanned by the columns $\mathrm{M}^* \setminus j$ of $\mathbf{X}^*$. The method of proof is to evaluate $\boldsymbol{l}^T_{j\cdot\mathrm{M}}\,\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}$. The main argument is based on

$$\mathbf{X}_j^T(\mathbf{I} - \mathbf{P}_{\mathrm{M}\setminus j})(\mathbf{I} - \mathbf{P}^*_{\mathrm{M}^*\setminus j})\mathbf{X}^*_j \;=\; 1, \tag{A.1.4}$$

which follows from these facts:

$$\mathbf{P}_{\mathrm{M}\backslash j}\mathbf{P}^*_{\mathrm{M}^*\backslash j} = \mathbf{0}, \quad \mathbf{P}_{\mathrm{M}\backslash j}\mathbf{X}^*_j = \mathbf{0}, \quad \mathbf{P}^*_{\mathrm{M}^*\backslash j}\mathbf{X}_j = \mathbf{0}, \quad \mathbf{X}_j^T\mathbf{X}^*_j = 1,$$

which in turn are consequences of (A.1.3) and $\mathbf{X}^T\mathbf{X}^* = \mathbf{I}_p$. We also know from (1.5.2) that

$$\|\boldsymbol{l}_{j\cdot\mathrm{M}}\| \;=\; 1/\|(\mathbf{I}-\mathbf{P}_{\mathrm{M}\backslash j})\mathbf{X}_j\|\,, \qquad \|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\| \;=\; 1/\|(\mathbf{I}-\mathbf{P}^*_{\mathrm{M}^*\backslash j})\mathbf{X}^*_j\|\,. \tag{A.1.5}$$

Putting together (A.1.4), (A.1.5), and (1.5.2), we obtain

$$\boldsymbol{l}_{j\cdot\mathrm{M}}^T\,\boldsymbol{l}^*_{j\cdot\mathrm{M}^*} \;=\; \|\boldsymbol{l}_{j\cdot\mathrm{M}}\|^2\,\|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\|^2 \;>\; 0. \tag{A.1.6}$$

Because the two vectors are scalar multiples of each other, we also know that

$$\boldsymbol{l}_{j\cdot\mathrm{M}}^T\,\boldsymbol{l}^*_{j\cdot\mathrm{M}^*} \;=\; \pm\,\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|\,\|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\|. \tag{A.1.7}$$

Putting together (A.1.6) and (A.1.7) we conclude

$$\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|\,\|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\| \;=\; 1, \qquad \bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \;=\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}},$$

This proves the lemma and the theorem. □

A.1.4 Proof of Theorem 1.6.1

The parameter a can range from $-1/p$ to ∞, but because of duality there is no loss of generality in considering only the case in which $a \geq 0$, and we do so in the following. Let $\mathrm{M} \subset \{1, \ldots, p\}$ and $j \in \mathrm{M}$. If $\mathrm{M} = \{j\}$ then $\boldsymbol{l}_{j\cdot\mathrm{M}} = \mathbf{X}_j$, the j-th column of $\mathbf{X}$, and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} = \boldsymbol{l}_{j\cdot\mathrm{M}}/\sqrt{pa^2 + 2a + 1}$. It follows that for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$,

$$|\bar{\boldsymbol{l}}_{j\cdot\mathbf{M}}^T \mathbf{Z}| \leq |\sum_{k\neq j} Z_k|/\sqrt{p} + |Z_j| \leq \sqrt{2\log p}(1 + o_p(1)) \qquad \text{(A.1.8)}$$

because $\|\mathbf{Z}\|_\infty = (1 + o_p(1))\sqrt{2\log p}$.

Because of (A.1.8) we now need only consider model selection sets, M, that contain at least two indices. For notational convenience, consider the case that $j = 1$ and $\mathrm{M} = \{1, \ldots, m\}$ with $2 \leq m \leq p$. The following results can then be applied to arbitrary j and M by permuting coordinates.

When $m \geq 2$ the projection of $\mathbf{X}_1$ on the space spanned by $\mathbf{X}_2, \ldots, \mathbf{X}_m$ must be of the form

$$\mathrm{Proj} \;=\; \frac{c}{m-1}\sum_{k=2}^{m} \mathbf{X}_k \;=\; \left(ca, \underbrace{ca + \frac{c}{m-1}, \ldots, ca + \frac{c}{m-1}}_{m-1}, \underbrace{ca, \ldots, ca}_{p-m} \right)$$

where the constant c satisfies $\boldsymbol{l}_{1\cdot\mathrm{M}} = (\mathbf{X}_1 - \mathrm{Proj}) \perp \mathrm{Proj}$. This follows from symmetry;

no calculation of projection matrices is needed to verify this. Let $d = 1 - c$. Then

$$(\boldsymbol{l}_{1\cdot\mathrm{M}})_k = \begin{cases} 1 + da & k = 1 \\ da - \frac{1-d}{m-1} & 2 \leq k \leq m \\ da & k \geq m+1 \end{cases} .$$

Some algebra starting from $\boldsymbol{l}_{1\cdot\mathrm{M}}^T \mathbf{X}_2 = 0$ yields

$$d \ = \ \frac{1/(m-1)}{pa^2 + 2a + 1/(m-1)}.$$

The term $d = d(a)$ is a simple rational function of a, and it is easy to check when $m \geq 2$ that $0 \leq da < 1/(2\sqrt{p})$.

Note also that $\|\boldsymbol{l}_{1\cdot\mathrm{M}}\| \geq 1$. Hence $\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}} = \boldsymbol{l}_{1\cdot\mathrm{M}}/\|\boldsymbol{l}_{1\cdot\mathrm{M}}\|$ satisfies

$$|\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}}^T \mathbf{Z}| \leq |Z_1| + |\frac{1}{m-1}\sum_{j=2}^{m} Z_j| + |\frac{1}{2\sqrt{p}}\sum_{j=1}^{p} Z_j| \leq 2\sqrt{2\log p}(1 + o_p(1)) + O_p(1).$$

This verifies that

$$\limsup_{p\to\infty} \frac{\sup_{a\in(-1/p,\infty)} K(\mathbf{X}(a))}{\sqrt{2\log p}} \leq 2 \quad \text{in probability.} \tag{A.1.9}$$

It remains to prove that equality holds in (A.1.9). Let $Z_{(1)} < Z_{(2)} < \ldots < Z_{(p)}$ denote the order statistics of $\mathbf{Z}$. Fix m. It is well-known that, in probability,

$$\lim_{p\to\infty} \frac{Z_{(1)}}{\sqrt{2\log p}} = -1 \quad \text{and} \quad \lim_{p\to\infty} \frac{Z_{(j)}}{\sqrt{2\log p}} = 1 \quad \forall j:\ p - m \ \leq \ j \ \leq \ p.$$

Note that

$$\lim_{a\to\infty} da = 0 \quad \text{and} \quad \lim_{a\to\infty} \|\boldsymbol{l}_{1\cdot\mathrm{M}}\|^2 = 1 + (m-1)^{-1}.$$

For any given $\mathbf{Z}$ one may choose to look at $\boldsymbol{l}_{j^*\cdot\mathrm{M}^*}$, with j^* being the index of $Z_{(1)}$ and $\mathrm{M}^* = \{j^*\} \cup \{j \,|\, Z_j = Z_{(k)}, p-m+2 \leq k \leq p\}$. The above then yields, in probability,

$$\lim_{p\to\infty, a\to\infty} \frac{|\bar{\boldsymbol{l}}^T_{j^*\cdot\mathrm{M}^*}\mathbf{Z}|}{\sqrt{2\log p}} \geq \frac{2}{\sqrt{1+(m-1)^{-1}}}.$$

Choosing m arbitrarily large and combining this with (A.1.9) yields the desired conclusion.

A.1.5 Proof of Theorem 1.6.2

Consider a primary predictor and controls (PP&C) design matrix

$$\mathbf{X} = \begin{pmatrix} \sqrt{1-(p-1)c^2} & \mathbf{0}^T_{p-1} \\ c\mathbf{1}_{p-1} & \mathbf{I}_{p-1} \end{pmatrix} \tag{A.1.10}$$

where c is the correlation between the primary predictor and the control predictors.

For a model $\mathrm{M} \supset \{1\}$ with $|\mathrm{M}| = m$, $\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}}$ is of the following form:

$$\bar{\boldsymbol{l}}_{1\cdot\mathrm{M},j} = \begin{cases} \sqrt{\frac{1-(p-1)c^2}{1-(m-1)c^2}} & j = 1 \\ 0 & j \in \mathrm{M}\backslash\{1\} \\ \frac{c}{\sqrt{1-(m-1)c^2}} & j \in \mathrm{M}^c \end{cases} \tag{A.1.11}$$

Therefore,

$$z_{1\cdot\mathrm{M}} = \bar{\boldsymbol{l}}_{1\cdot\mathrm{M}}^T \mathbf{Z} = \sqrt{\frac{1-(p-1)c^2}{1-(m-1)c^2}} Z_1 + \frac{c}{\sqrt{1-(m-1)c^2}} \sum_{j \in \mathrm{M}^c} Z_j. \tag{A.1.12}$$

Note that for each fixed m,

$$\begin{aligned} &\max_{\mathrm{M}} \left| \frac{c}{\sqrt{1-(m-1)c^2}} \sum_{j \in \mathrm{M}^c} Z_j \right| \\ =&\frac{c}{\sqrt{1-(m-1)c^2}} \max \left(\sum_{j=1}^{p-m-1} Z_{(p-j+1)}, -\sum_{j=1}^{p-m-1} Z_{(j)} \right) \end{aligned} \tag{A.1.13}$$

where $Z_{(j)}$ is the j-th order statistic. Note also that

$$\sup_c \frac{c}{\sqrt{1-(m-1)c^2}} = \frac{1}{\sqrt{p-m}} \tag{A.1.14}$$

and the equality is attained as $c \to 1/\sqrt{p-1}$. Therefore, for each fixed m,

$$\begin{aligned} &\sup_c \frac{1}{\sqrt{p}} \max_{|\mathrm{M}|=m} |z_{1\cdot\mathrm{M}}| \\ =&\sqrt{\frac{p}{p-m}} \max \left(\sum_{j=1}^{p-m-1} \frac{1}{p} Z_{(p-j+1)}, -\sum_{j=1}^{p-m-1} \frac{1}{p} Z_{(j)} \right) + O_p\left(\sqrt{\frac{\log p}{p}} \right). \end{aligned} \tag{A.1.15}$$

Suppose $m = rp$. Note that as $p \to \infty$, by Bahadur (1966),

$$Z_{(p-j+1)} = \Phi^{-1}\left(\frac{p-j+1}{p+1} \right) + O_p(p^{-1/2}). \tag{A.1.16}$$

Note also that $\sum_{j=1}^{p-m-1} \frac{1}{p} \Phi^{-1}\left(\frac{p-j+1}{p+1}\right)$ is a good approximation of the Riemann integral $\int_r^1 \Phi^{-1}(x)dx$:

$$\int_r^1 \Phi^{-1}(x)dx = \sum_{j=1}^{p-m-1} \frac{1}{p}\Phi^{-1}\left(\frac{p-j+1}{p+1}\right) + O(p^{-2}). \tag{A.1.17}$$

Therefore,

$$\sum_{j=1}^{p-m-1} \frac{1}{p} Z_{(p-j+1)} = \int_r^1 \Phi^{-1}(x)dx + O_p(p^{-1/2}). \tag{A.1.18}$$

Similarly,

$$-\sum_{j=1}^{m-1} \frac{1}{p} Z_{(j)} = \int_r^1 \Phi^{-1}(x)dx + O_p(p^{-1/2}). \tag{A.1.19}$$

Summarizingly,

$$\begin{aligned}
&\sup_c \frac{1}{\sqrt{p}} \max_{|\mathrm{M}|=m} |z_{1\cdot\mathrm{M}}| \\
=&\sqrt{\frac{p}{p-m}} \max\left(\sum_{j=1}^{p-m-1} \frac{1}{p} Z_{(p-j+1)}, -\sum_{j=1}^{p-m-1} \frac{1}{p} Z_{(j)}\right) + O_p(\sqrt{\log p/p}) \\
=&\frac{1}{\sqrt{1-r}} \int_r^1 \Phi^{-1}(x)dx + O_p(p^{-1/2}) + O_p(\sqrt{\log p/p}) \\
=&\frac{1}{\sqrt{1-r}} \phi(\Phi^{-1}(r)) + O_p(\sqrt{\log p/p}).
\end{aligned} \tag{A.1.20}$$

The function $f(r) = \frac{1}{\sqrt{1-r}}\phi(\Phi^{-1}(r))$ is maximized at $r^* = 0.73$ with $f(r^*) = 0.636$. Therefore,

$$\lim_{p\to\infty} \sup \sup_c \frac{1}{\sqrt{p}} \max_{\mathrm{M}} |z_{1\cdot\mathrm{M}}| = 0.636. \tag{A.1.21}$$

This sharpness of this bound is seen by considering the model with the first or last $m^* = r^*p$ order statistics of $\mathbf{Z}$ when $p \to \infty$ and $c \to \frac{1}{\sqrt{p-1}}$.

With (A.1.21), we conclude that $K_1(\mathbf{X}) \sim 0.636\sqrt{p}$. $\square$

A.1.6 Proof of Theorem 1.6.3

We will show that if $a_p^{1/p} \to a$ (> 0), we have

- a uniform asymptotic worst-case bound:
  $\lim_{p\to\infty} \sup_{|\mathcal{L}_p|\le a_p} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/\sqrt{p} \overset{\mathbf{P}}{\le} \sqrt{1-1/a^2}$;

- attainment of the bound when $|\mathcal{L}_p| = a_p$ and $\bar{\boldsymbol{l}} \in \mathcal{L}_p$ are i.i.d. Unif(S^{p-1}) independent of $\mathbf{Z}$:
  $\lim_{p\to\infty} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/\sqrt{p} \overset{\mathbf{P}}{\ge} \sqrt{1-1/a^2}$.

These facts imply the assertions about $(1-\alpha)$-quantiles $K(\mathcal{L}_p)$ of $\max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{Z}|$ in Theorem 1.6.3. We decompose $\mathbf{Z} = R\mathbf{U}$ where $R^2 = \|\mathbf{Z}\|^2 \sim \chi_p^2$ and $\mathbf{U} = \mathbf{Z}/\|\mathbf{Z}\| \sim$ Unif(S^{p-1}) are independent. Due to $R/\sqrt{p} \xrightarrow{\mathbf{P}} 1$ it is sufficient to show the following:

- uniform asymptotic worst-case bound:

$$\lim_{p\to\infty} \sup_{|\mathcal{L}_p|\le a_p} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{U}| \quad \overset{\mathbf{P}}{\le} \quad \sqrt{1-1/a^2}\,; \tag{A.1.22}$$

- attainment of the bound when $|\mathcal{L}_p| = a_p$ and $\bar{\boldsymbol{l}} \in \mathcal{L}_p$ are i.i.d. Unif(S^{p-1}) independent of $\mathbf{U}$:

$$\lim_{p\to\infty} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{U}| \quad \overset{\mathbf{P}}{\ge} \quad \sqrt{1-1/a^2}\,. \tag{A.1.23}$$

To show (A.1.22), we upper-bound the non-coverage probability and show that it converges to zero for $K' > \sqrt{1 - 1/a^2}$. To this end we start with a Bonferroni-style bound, as in Wyner (1967):

$$\begin{aligned}
\mathbf{P}[\max_{\bar{\boldsymbol{l}} \in \mathcal{L}} |\bar{\boldsymbol{l}}^T \mathbf{U}| > K'] &= \mathbf{P} \bigcup_{\bar{\boldsymbol{l}} \in \mathcal{L}} [\, |\bar{\boldsymbol{l}}^T \mathbf{U}| > K'] && \text{(A.1.24)} \\
&\leq \sum_{\bar{\boldsymbol{l}} \in \mathcal{L}} \mathbf{P}[\, |\bar{\boldsymbol{l}}^T \mathbf{U}| > K'] && \text{(A.1.25)} \\
&= |\mathcal{L}_p| \, \mathbf{P}[\, |U| > K'], && \text{(A.1.26)}
\end{aligned}$$

where U is any coordinate of $\mathbf{U}$ or projection of $\mathbf{U}$ onto a unit vector. We will show that the bound (A.1.26) converges to zero. We use the fact that $U^2 \sim \mathrm{Beta}(1/2, (p-1)/2)$, hence

$$\mathbf{P}[\, |U| > K'] \;=\; \frac{1}{\mathrm{B}(1/2, (p-1)/2)} \int_{K'^2}^{1} x^{-1/2} (1-x)^{(p-3)/2} dx \qquad \text{(A.1.27)}$$

We bound the Beta function and the integral separately:

$$\frac{1}{\mathrm{B}(1/2, (p-1)/2)} \;=\; \frac{\Gamma(p/2)}{\Gamma(1/2)\Gamma((p-1)/2)} \;<\; \sqrt{\frac{(p-1)/2}{\pi}},$$

where we used $\Gamma(x + 1/2)/\Gamma(x) < \sqrt{x}$ (a good approximation, really) and $\Gamma(1/2) = \sqrt{\pi}$.

$$\int_{K'^2}^{1} x^{-1/2} (1-x)^{(p-3)/2} dx \;\leq\; \frac{1}{K'} \frac{1}{(p-1)/2} (1 - K'^2)^{(p-1)/2},$$

where we used $x^{-1/2} \leq 1/K'$ on the integration interval. Continuing with the chain of bounds from (A.1.26) we have:

$$|\mathcal{L}_p|\mathbf{P}[\,|U| > K'] \;\leq\; \frac{1}{K'}\left(\frac{2}{(p-1)\pi}\right)^{1/2}\left(|\mathcal{L}_p|^{1/(p-1)}\sqrt{1-K'^2}\right)^{p-1}.$$

Since $|\mathcal{L}_p|^{1/(p-1)} \to a$ (> 0) by assumption, the right hand side converges to zero at geometric speed if $a\sqrt{1-K'^2} < 1$, that is, if $K' > \sqrt{1-1/a^2}$. This proves (A.1.22).

To show (A.1.23), we upper-bound the coverage probability and show that it converges to zero for $K' < \sqrt{1-1/a^2}$. We make use of independence of $\bar{\boldsymbol{l}} \in \mathcal{L}_p$, as in Wyner (1967):

$$\begin{aligned}
\mathbf{P}[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p}|\bar{\boldsymbol{l}}^T\mathbf{U}| \leq K'] &= \prod_{\bar{\boldsymbol{l}}\in\mathcal{L}_p}\mathbf{P}[\,|\bar{\boldsymbol{l}}^T\mathbf{U}| \leq K'] \;=\; \mathbf{P}[\,|U| \leq K']^{|\mathcal{L}_p|} && \text{(A.1.28)}\\
&= (1-\mathbf{P}[\,|U| > K'])^{|\mathcal{L}_p|} && \text{(A.1.29)}\\
&\leq \exp\left(-|\mathcal{L}_p|\,\mathbf{P}[\,|U| > K']\right). && \text{(A.1.30)}
\end{aligned}$$

We will lower-bound the probability $\mathbf{P}[\,|U| > K']$ recalling (A.1.27) and again deal with the Beta function and the integral separately:

$$\frac{1}{\mathrm{B}(1/2,(p-1)/2)} \;=\; \frac{\Gamma(p/2)}{\Gamma(1/2)\Gamma((p-1)/2)} \;>\; \sqrt{\frac{p/2-3/4}{\pi}},$$

where we used $\Gamma(x+1)/\Gamma(x+1/2) > \sqrt{x+1/4}$ (again, a good approximation really).

$$\int_{K'^2}^{1} x^{-1/2}(1-x)^{(p-3)/2}dx \;\geq\; \frac{1}{(p-1)/2}(1-K'^2)^{(p-1)/2},$$

where we used $x^{-1/2} \geq 1$. Putting it all together we bound the exponent in (A.1.30):

$$|\mathcal{L}_p| \, \mathbf{P}[\, |U| > K'] \ \geq \ \frac{\sqrt{p/2 - 3/4}}{\sqrt{\pi}\,(p-1)/2} \left(|\mathcal{L}_p|^{1/(p-1)} \sqrt{1 - K'^2} \right)^{p-1} .$$

Since $|\mathcal{L}_p|^{1/(p-1)} \to a \ (> 0)$ by assumption, the right hand side converges to $+\infty$ at nearly geometric speed if $a\sqrt{1 - K'^2} > 1$, that is, if $K' < \sqrt{1 - 1/a^2}$. This proves (A.1.23).

Bibliography

[1] Abadie, A., Diamond, A. and Hainmueller, J. (2010). Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* **105**, 493–505.

[2] Abadie, A. and Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review* **93**, 112–132.

[3] Abadie, A. and Imbens, G. W. (2011). Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business and Economic Statistics* **29**, 1–11.

[4] American College of Gynecology (2003). Practice Bulletin: Management of Preterm Labor. *Obstetrics and Gynecology* **101**, 1039–1047.

[5] Angrist, J. and Krueger, A. (2000). Empirical Strategies in Labor Economics, in *Handbook of Labor Economics* (O. Ashenfelter and D. Card eds), 1277–1366, Amsterdam: Elsevier.

[6] Angrist, X. and Pischke, X. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

[7] Athey, S. and Imbens, G.W. (2006). Identification and Inference in Nonlinear Difference-in-Differences Models. *Econometrica* **74**, 431–497.

[8] Bahadur, R. (1966). A Note on Quantiles in Large Samples. *The Annals of Mathematical Statistics* **37**, 577–580.

[9] Berk, R., Brown, L. D. and Zhao, L. (2010). Statistical Inference after Model Selection. *Journal of Quantitative Criminology* **26**, 217–236.

[10] Berk, R., Brown, L. D., Buja, A., Zhang, K. and Zhao, L. (2012). Valid Post-Selection Inferece. Submitted.

[11] Brown, L. D. (1967). The Conditional Level of Student's t-Test. *The Annals of Mathematical Statistics* **38**, 1068–1071.

[12] Brown, L. D. (1990). An Ancillarity Paradox which Appears in Multiple Linear Regression. *The Annals of Statistics* **18**, 471–493.

[13] Buehler, R. J. and Fedderson, A. P. (1963). Note on a Conditional Property of Student's t. *The Annals of Mathematical Statistics* **34**, 1098–1100.

[14] Buja, A., Zhang, K., Berk, R., Brown, L. D. and Zhao, L. (2012). Computational Methods for Post-Selection Inference. (Forthcoming; partly contained in an older version of the present article and available at `http://stat.wharton.upenn.edu/ buja/PAPERS/PoSI.pdf`).

[15] Campbell, D. T. (1957). Factors Relevant to the Validity of Experiments in Social Settings. *Psychological Bulletin* **54**, 297–312.

[16] Campbell, D. T. (1969). Reforms as Experiments. *American Psychologist* **24**, 409–429.

[17] Copas, J. and Eguchi, S. (2001). Local Sensitivity Approximations for Selectivity Bias. *Journal of the Royal Statistical Society* B **63**, 871–896.

[18] Cornfield, J., Haenszel, W., Hammond, E. et al. (1959). Smoking and Lung Cancer. *Journal of the National Cancer Institute* **22**, 173–203.

[19] Cox, D. (1975). A Note on Data-Splitting for the Evaluation of Significance Levels. *Biometrika* **62**, 441–444.

[20] Dijkstra, T. K. and Veldkamp, J. H. (1988). Data-driven Selection of Regressors and the Bootstrap, in *On Model Uncertainty and Its Statistical Implications* (T. K. Dijkstra, ed.), 17–38, Berlin: Springer.

[21] DiPrete, T. A. and Gangl, M. (2004). Assessing Bias in the Estimation of Causal Effects. *Sociological Methodology* **34**, 271–310.

[22] Fisher, R.A. (1935). *The Design of Experiments*, Edinburgh: Oliver & Boyd.

[23] Freeman, R.B. (1984). Longitudinal Analyses of the Effect of Trade Unions. *Journal of Labor Economics* **2**, 1–26.

[24] Gart, J. J. (1969). An Exact Test for Comparing Matched Proportions in Crossover Designs. *Biometrika* **56**, 75–80.

[25] Gastwirth, J. L. (1992). Methods for Assessing the Sensitivity of Comparisons in Title VII Cases to Omitted Variables. *Jurimetrics Journal* **33**, 19–34.

[26] Gastwirth, J. L., Krieger, A. M. and Rosenbaum, P. R. (1998). Dual and Simultaneous Sensitivity Analysis for Matched Pairs. *Biometrika* **85**, 907–920.

[27] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B. and Hothorn, T. (2010). *mvtnorm: Multivariate Normal and t Distributions*, `http://cran.r-project.org/web/packages/mvtnorm/`.

[28] Griliches, Z. and Hausman, J.A. (1986). Errors in Variables in Panel Data. *Journal of Econometrics* **31**, 93–118.

[29] Hall, P. and Carroll, R. (1989). Variance Function Estimation in Regression: The Effect of Estimating the Mean. *Journal of the Royal Statistical Society* B **51**, 3–14.

[30] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning — Data Mining, Inference, and Prediction*, 2nd ed. Corr. 3rd printing. Springer Series in Statistics, New York: Springer.

[31] Hansen, B. B. (2007). Optmatch: Flexible, Optimal Matching for Observational Studies. *R News* **7**, 18-24.

[32] Hansen, B. B. and Klopfer, S. O. (2006). Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics* **15**, 609–627.

[33] Heller, R., Rosenbaum, P. and Small, D. (2009). Split Samples and Design Sensitivity in Observational Studies. *Journal of the American Statistical Association* **104**, 1090–1101.

[34] Hollowell, J., Oakley, L., Kurinczuk, J. J., Brocklehurst, P. and Gray, R. (2011). The Effectiveness of Antenatal Care Programmes to Reduce Infant Mortality and Preterm Birth in Socially Disadvantaged and Vulnerable Women in High-Income Countries: A Systematic Review. *BMC Pregnancy Childbirth* **11**, 13.

[35] Hurvich, C. and Tsai, C. (1990). The Impact of Model Selection on Inference in Linear Regression. *The American Statistician* **44**, 214–217.

[36] Imbens, G. W. (2003). Sensitivity to Exogeneity Assumptions in Program Evaluation. *American Economic Review 93*, 126–132.

[37] Kabaila, P. (1998). Valid Confidence Intervals in Regression after Variable Selection. *Econometric Theory* **14**, 463–482.

[38] Kabaila, P. and Leeb, H. (2006). On the Large-Sample Minimal Coverage Probability of Confidence Intervals after Model Selection. *Journal of the American Statistical Association* **101** (474), 619–629.

[39] Kirby, P.B., Spetz, J., Maiuro, L. and Scheffler, R. M. (2006). Changes in Service Availability in California Hospitals, 1995 to 2002. *Journal of Healthcare Management* **51**, 26–38.

[40] LANCASTER, H. O. (1949). The Derivation and Partition of χ^2 in Certain Discrete Distributions. *Biometrika* **36**, 117–129.

[41] LEEB, H. (2006). The Distribution of a Linear Predictor after Model Selection: Unconditional Finaite-Sample Distributions and Asymptotic Approximations. *IMS Lecture Notes - Monograph Series* **49**, 291–311.

[42] LEEB, H. and PÖTSCHER, B. M. (2003). The Finite-Sample Distributions of Post-Model-Selection Estimators and Uniform versus Nonuniform Approximations. *Econometric Theory* **19**, 100–142.

[43] LEEB, H. and PÖTSCHER, B. M. (2005). Model Selection and Inference: Facts and Fiction, *Econometric Theory* **21**, 21–59.

[44] LEEB, H. and PÖTSCHER, B. M. (2006). Performance Limits for Estimators of the Risk or Distribution of Shrinkage-Type Estimators, and Some General Lower Risk-Bound Results. *Econometric Theory* **22**, 69–97.

[45] LEEB, H. and PÖTSCHER, B. M. (2006). Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators? *The Annals of Statistics* **34**, 2554–2591.

[46] LEEB, H. and PÖTSCHER, B. M. (2008a). Model Selection, in *The Handbook of Financial Time Series* (T. G. Anderson, R. A. Davis, J. -P. Kreib, and T. Mikosch, eds), 785–821, New York: Springer.

[47] LEEB, H. and PÖTSCHER, B. M. (2008b). Can One Estimate the Uncondi-

tional Distribution of Post-Model-Selection Estimators? *Econometric Theory* **24** (2), 338–376.

[48] Leeb, H. and Pötscher, B. M. (2008c). Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator. *Journal of Econometrics* **142**, 201–211.

[49] Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998). Assessing Sensitivity of Regression to Unmeasured Confounders in Observational Studies. *Biometrics* **54**, 948–963.

[50] Mammen, E. (1993). Bootstrap and Wild Bootstrap for High-Dimensional Linear Models. *The Annals of Statistics* **21**, pp. 255–285.

[51] Marcus, S. M. (1997). Using Omitted Variable Bias to Assess Uncertainty in the Estimation of an AIDS Education Treatment Effect. *Journal of Educational Statistics* **22**, 193–201.

[52] Meyer, B.D. (1995). Natural and Quasi-Natural Experiments in Economics. *Journal of Business and Economic Statistics* **13**, 151–161.

[53] Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics*, 4th ed., New York: W. H. Freeman and Company.

[54] Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, New York: Addison-Wesley.

[55] Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Reprinted in *Statistical Science* **5**, 463–480.

[56] Olshen, R. A. (1973). The Conditional Level of the F-Test. *Journal of the American Statistical Association* **68**, 692–698.

[57] Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False Discovery Control for Random Fields. *Journal of the American S-tatistical Association*, **99** (468), 1002–1014.

[58] Pötscher, B. M. (1991). Effects of Model Selection on Inference. *Econometric Theory* **7**, 163–185.

[59] Pötscher, B. M. and Leeb, H. (2009). On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding. *Journal of Multivariate Analysis* **100**, 2065–2082.

[60] Robins, J. M., Rotnitzky, A. and Scharfstein, D. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models, in *Statistical Models in Epidemiology* (E. Halloran and D. Berry eds), 1–94, New York: Springer.

[61] Rosenbaum, P. R. (1984). From Association to Causation in Observational Studies: the Role of Tests of Strongly Ignorable Treatment Assignment. *Journal of the American Statistical Association* **79**, 41–48.

[62] Rosenbaum, P.R. (1989). Optimal Matching in Observational Studies. *Journal of the American Statistical Association* **84**, 1024–1032.

[63] Rosenbaum, P. R. (1995). Quantiles in Nonrandom Samples and Observational Studies. *Journal of the American Statistical Association* **90**, 1424–1431.

[64] Rosenbaum, P. R. (2002). *Observational Studies*, New York: Springer.

[65] Rosenbaum, P. R. (2004). Design Sensitivity in Observational Studies. *Biometrika* **91**, 153–64.

[66] Rosenbaum, P. R. (2010a). *Design of Observational Studies*, New York: Springer.

[67] Rosenbaum, P. R. (2010b). Design Sensitivity and Efficiency in Observational Studies. *Journal of the American Statistical Association* **105**, 692–702.

[68] Rosenbaum, P. R. (2010c). Evidence Factors in Observational Studies. *Biometrika* **97**, 333–345.

[69] Rosenbaum, P. R. (2011). Some Approximate Evidence Factors in Observational Studies. *Journal of the American Statistical Association* **106**, 285–295.

[70] Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. *Journal of the Royal Statistical Society* B **45**, 212–218.

[71] Rosenbaum, P.R. and Rubin, D.B. (1985). Constructing a Control Group by Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *American Statistician* **39**, 33–38.

[72] Rosenbaum, P. R. and Silber, J. H. (2009a). Amplification of Sensitivity Analysis in Observational Studies. *Journal of the American Statistical Association* **104**, 1398–1405.

[73] Rosenbaum, P. R. and Silber, J. H. (2009b). Sensitivity Analysis for Equivalence and Difference in an Observational Study of Neonatal Intensive Care Units. *Journal of the American Statistical Association* **104**, 501–511.

[74] Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **66**, 688–701.

[75] Shadish, W. R., Cook, T. D. and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inferenc*, Boston: Houghton-Mifflin.

[76] Scheffé, H. (1953). A Method for Judging All Contrasts in the Analysis of Variance. *Biometrika* **40**, 87–104.

[77] Scheffé, H. (1959). *The Analysis of Variance*, New York: John Wiley & Sons.

[78] Sen, P. K. (1979). Asymptotic Properties of Maximum Likelihood Estimators Based on Conditional Specification. *The Annals of Statistics*, **7**, 742–755.

[79] Sen, P. K. and Saleh, A. K. M. E. (1987). On Preliminary Test and Shrinkage M-Estimation in Linear Models. *The Annals of Statistics*, **15**, 1580–1592.

[80] Stoll, B. J., Hansen, N. I., Bell, E. F., Shankaran, S., Laptook, A. R., Walsh, M. C., Hale, E. C. et al. (2010). Neonatal Outcomes of Ex-

tremely Preterm Infants from the NICHD Neonatal Research Network. *Pediatrics* **126**, 443–456.

[81] Stuart, E.A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* **25**, 1–21.

[82] Trochim, W. M. K. (1985). Pattern Matching, Validity and Conceptualization in Program Evaluation. *Evaluation Review* **9**, 575–604.

[83] Tukey, J. (1980a). Methodological Comments Focused on Opportunities, in *Multivariate Techniques in Human Communication Research* (P. R. Monge & J. N. Cappella, eds), 489–528, New York: Academic Press.

[84] Tukey, J. (1980b). We Need Both Exploratory and Confirmatory. *The American Statistician* **34**, 23–25.

[85] Wang, R. and Lagakos, S. W. (2009). Inference after Variable Selection Using Restricted Permutation Methods. *The Canadian Journal of Statistics* **37** (4), 625–644.

[86] West, S.G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D.C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M. and Mullen, P.D.(2008). Alternatives to the Randomized Controlled Trial. *American Journal of Public Health* **98**, 1359–1366.

[87] Wyner, A. D. (1967). Random Packings and Coverings of the Unit n-Sphere. *Bell System Technical Journal*, **46**, 2111–2118.

[88] Yanagawa, T. (1984). Case-Control Studies: Assessing the Effect of a Confounding Factor. *Biometrika* **71**, 191–194.

[89] Young, A. and Karr, S. (2011). Deming, data and observational studies. *Significance* **8** (3), 116–120.

[90] Zhang, K., Small, D., Lorch, S., Srinivas, S. and Rosenbaum, P. (2011). Using Split Samples and Evidence Factors in an Observational Study of Neonatal Outcomes. *Journal of the American Statistical Association* **106**, 511-524.