TOPICS IN ADAPTIVE INFERENCE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

William Fithian
June 2015

This dissertation is online at: http://purl.stanford.edu/qc856vk9328

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Trevor Hastie, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Emmanuel Candes

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jonathan Taylor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Robert Tibshirani

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

Abstract

The recent rapid expansion in data collection has presented the field of statistics with an inexhaustible array of exciting new applications and problem settings not directly covered by classical methods or theory. In more complex problems, we must choose an appropriate computational algorithm, statistical model, or scientific question that is well-tuned to the sampling distribution we are presented with. Thus the payoff from using methods that adapt to the data set at hand is high. This work details several such methods.

Chapter 2 — based on Fithian and Hastie (2014) — describes an adaptive subsampling strategy for computationally efficient inference on massive data sets. Using a modified case-control sampling algorithm, we "compress" a large data set into a much smaller subsample, enriching for the most informative observations. Our algorithm enables potentially much faster analysis, at a provably small cost to statistical performance.

Most statistical inference procedures are formally invalid if they are preceded by "data snooping." However, in most real data problems it is hard to specify a reliable model *a priori*. Chapter 3 — based on Fithian et al. (2014) — discusses methods for carrying out valid inference after adaptive model selection, correcting for selection by conditioning on the model selected. Chapter 4 is as-yet unpublished work and gives further examples of selective inference applied to specific problems. First, I discuss the problem of "rank verification" — testing, for example, whether the candidate who receives the most votes in a random survey is actually leading in the population from which the survey was taken. Second, I discuss an exact nonparametric selective inference procedure for use when a the data fail a goodness-of-fit test for some set of parametric assumptions.

Acknowledgments

I would like to thank a great many people inside and outside of the department who have made the last five years some of the happiest and most productive of my life.

First and foremost, I would like to thank my family. My wife Kari has been my emotional anchor throughout my time in graduate school. Her steady emotional support has made it possible for me to focus a great deal of energy on my work while mostly maintaining my sanity. She also played a large role in tempting me to leave my comfortable professional life as a New York financier for the riskier path of starting afresh in academia — a risk which has paid off better than I could have imagined. My parents have been extremely supportive throughout my life, and have ceaselessly pushed me to question assumptions, stretch myself intellectually, and pursue the things in life that inspire me. My wonderful siblings have also shown me by example how to follow my passions.

I would like to thank Trevor Hastie, who has been an ideal academic advisor, mentor, advocate, and friend. His sharply skeptical mind has helped me hone my intuitions about which projects are worth working on, and he has also consistently encouraged me to work on many of the varied projects that have caught my interest rather than prematurely optimizing my skills for one small subfield. His humility, intellectual honesty, generosity, good humor, and forthrightness have also been an inspiration to me. As a young student first learning to navigate the waters of academia, it has been extremely reassuring to have an advisor whose advice I can trust not only to further my professional interests, but also to reflect the values that I aspire to as an academic citizen.

In addition to Trevor I would also like to thank the other two authors of the *Elements of Statistical Learning*, Rob Tibshirani and Jerry Friedman, for writing the book that made me want to attend statistics graduate school in the first place. My life is immeasurably more meaningful as a direct result of their effort to create an exciting and accessible text for non-expert readers. I am uncommonly blessed to have gotten to collaborate with the authors whose words first lit the fire of my excitement about statistics.

Getting to work with Jonathan Taylor has been yet another stroke of great fortune. The first time I heard about his ideas on selective inference I was pretty sure that they made no sense, but after vacillating many times I have (for now) flipped firmly to believing that they are fundamental and essential for twenty-first century statistics. It has been a tremendous privilege to get to play a part in the development of these techniques. In addition, I am very pleased to call Jonathan a friend and look forward to fruitful collaborations in the future.

I am very thankful for my remaining thesis reader, Emmanuel Candès, as well. I believe that his excellent Stats 300C course (of which I was lucky to be an inaugural student) has dramatically impacted the research being done at Stanford. I learned a great deal in that class and also honed my intuition for multiple comparison problems (in fact, my role in my first paper on selective inference was based on my end-term project in that class). Yoav Benjamini's class the following year also played a large role in stimulating my understanding of and interest in multiple comparisons and selective inference, so I was quite lucky to make his acquaintance as well.

More generally, the students in the statistics department have made it a wonderful place to spend five years. The cohort of students who entered Stanford with me — Sam Gross, Jian Li, Linxi Liu, Leo Pekelis, Dennis Sun, Nike Sun, and Zhen Zhu — were a remarkable source of support during the long summer of 2011 when we all prepared for our quals. I will carry with me many fond memories of late-night problem set sessions. My lovely memories with friends in the statistics department include: sharing coffee, confidences, and neuroses with Nike; the excitement of discovering "brilliant ideas" with Stefan, most of which were cruelly snuffed out in their infancy by the harsh light of reason (at least we didn't have to write them up); impromptu late-night showtunes recitals with Jacob; losing track of time while walking through Cambridge arguing with Rahul about robust optimization; enjoying warm homemade scones in front of Alex and Max's fireplace on a snowy evening in Pittsburgh; losing to the "backdoor Bayesians" time and again at trivia night with Dennis; trading sarcastic remarks with Lucas about bootstrap-related squabbles; and intense and highly competitive arm-wrestling bouts with Noah (OK, that last one never happened).

Presenting my work in front of Rob, Trevor and Jonathan's lab group has been an indispensible part of my development as a researcher. I am especially indebted to the "golden oldies" of yesteryear — Ryan Tibshirani, Brad Klingenberg, Jacob Bien, Rahul Mazumder, Noah Simon, Max G'Sell, and Alex Chouldechova (nearly all of whom are actually younger than me) — who served as academic older brothers, sisters, and cousins and from whom I learned a great deal. I consider myself very fortunate to have been embedded in such a community of scholars.

I have also been blessed with many excellent collaborators: in addition to those previously mentioned, Dennis Sun, Rahul Mazumder, Lucas Janson, Julie Josse, Percy Liang, Sida Wang, Jane Elith, David Keith, and Asaf Weinstein.

Finally, all of the statistics department students, faculty, and staff have been immensely helpful. In particular, there is essentially zero chance that I would have graduated on time without the many, many reminders (and second reminders) from Susie Ementon about administrative deadlines!

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Adaptivity in Statistical Inference

With the rise of ever larger and more complex data sets, it is increasingly difficult to specify beforehand the most appropriate computational or statistical method of analysis. This thesis will cover two topics involving *adaptive inference*, in which the analyst defers certain choices about his or her analysis method until after observing the data.

One result of the trend toward collecting very large data sets is that computation becomes much more cumbersome, forcing the analyst either to use efficient model-fitting methods, or to discard much of the data, or both. To decide which examples to discard, it helps to look at (some of) the data and determine which examples are likely to carry a large amount of information about the parameters of interest. Traditional statistical sampling methods such as case-control sampling can reduce the sampling effort needed to attain a given statistical precision, by oversampling examples from a rare class relative to a more common class. Chapter 2 discusses the method of *local case-control sampling* proposed in Fithian and Hastie (2014), which adapts to the sampling distribution in order to better measure the importance of a data point, and subsamples only those examples which appear important based on the predictions of a *pilot model* trained on a small subsample of the data.

Another result of this trend is that increasingly many inquiries are data-driven rather than hypothesis-driven — that is, rather than beginning with a scientific question and conducting an experiment to answer it, scientists are more likely to begin with a large and complex data set and hunt for "interesting" signals in that data.

Model misspecification, a perpetual obstacle for statistical inference, can have especially pernicious effects when data sets are very large. The null hypothesis is nearly always false, and with enormous sample sizes a mild misspecification can lead to miniscule p-values even if the effects discovered are mere artifacts of model misspecification. It is ever more implausible that we can predict whether our model is "good enough" based on theoretical considerations alone. As a result, it is ever more vital to examine the data and ask whether we are using an appropriate method. Because this type of "data snooping" formally invalidates the error guarantees of classical frequentist tests, it is necessary to develop frameworks for inference that can preserve their guarantees in the face of adaptive model selection. Chapter 3 describes the methods of Fithian et al. (2014) for valid frequentist inference for a question that is adaptively selected after viewing the data.

In each of the above scenarios, there is an adaptive selection process (of data points, or or interesting questions) by the analyst, which induces a selection bias in the data that will subsequently be used for inference. By using the distribution of the data conditional on selection, we can make an appropriate adjustment for this bias.

1.2 Local Case-Control Sampling

Many scientific and industrial problems involve data sets so large that computational costs play a major role in constraining what statistical procedures we can use, or what fraction of the available data we can afford to analyze. This can lead to a direct tradeoff between computational efficiency and statistical performance.

In particular, many classification problems involve class imbalance, in which one class is marginally much rarer than the other. More generally, we may have *conditional class imbalance*, where for most values of the features X, the response $Y \in \{0, 1\}$ is quite predictable: either $\mathbb{P}(Y = 1 \mid X = x) \approx 0$ or $\mathbb{P}(Y = 1 \mid X = x) \approx 1$, but the majority label is not necessarily the same for all x. This occurs, for example, in the problem of email spam filtering: spam and legitimate email are both common, but well-trained classifiers make very few mistakes. When most examples are easy, the information in the data set is concentrated in the most surprising examples. Thus, if we subsample the data enriching for the surprises, we might hope to obtain more computationally efficient procedures that perform nearly as well as if we trained them on all the data.

Chapter 2 discusses one such scheme for statistically efficient subsampling, which we call *local case-control sampling*, that applies a simple accept-reject rule to condense the data set $\{(X_i, Y_i)\}_{i=1}^n$. Observation i is kept with probability

$$\mathbb{P}\left(\text{accept } (X_i, Y_i)\right) = a(X_i, Y_i) = \left| Y_i \; - \; \frac{e^{\tilde{\beta}' X_i}}{1 + e^{\tilde{\beta}' X_i}} \right|, \tag{1.1}$$

where $\tilde{\beta}' x$ in (1.1) is a pilot estimate of the log-odds. In an industrial setting, $\tilde{\beta}$ could represent yesterday's fit; otherwise, it could be an estimate computed on a smaller subsample. The size of the compressed data set we obtain using the rule (1.1) depends on the expectation of $a(X, Y)$, which measures how surprising Y is, given X. For example, suppose there are 1000 cases and 1,000,000 controls, and $\mathbb{P}(Y_i = 1 \mid X_i) \ll 1$ for all X_i. The rule (1.1) would keep nearly all 1000 of the cases and about 1000 controls, preferentially sampling controls that look the most like cases, for a subsample 500 times smaller than the original data set.

In effect, (1.1) exponentially tilts the conditional Bernoulli distribution of Y given $X = x$ for each x. If the conditional log-odds in the original population is $f(x)$, the log-odds in the subsample is $f(x) - \tilde{\beta}'x$, which is close to 0 if the pilot fit is good. Thus the subsample is conditionally balanced throughout the feature space. If we fit a logistic regression to the compressed data set, we can correct the sampling bias by adding back $\tilde{\beta}$ to the subsampled estimate of β.

When the logistic regression model is misspecified ($f(x)$ is not actually linear), let β^* denote the coefficients minimizing the risk under the logistic loss for the original population. If $\tilde{\beta}$ is consistent for β^*, standard case-control sampling — uniformly downsampling the majority class to balance the sample, estimating a logistic regression model on the subsample, and then adjusting the intercept — leads to biased estimates for β^*. As $n \to \infty$, the case-control estimate converges instead to another limit, $\beta^*_{\text{CC}} \neq \beta^*$. By contrast, the local case-control estimator $\hat{\beta}$ is consistent for β^*, as long as the pilot $\tilde{\beta}$ is.

If most examples are easy, the compressed data set can be orders of magnitude smaller than the original data set. Remarkably, however, the compressed sample contains exactly half the Fisher information in the original sample, so the asymptotic covariance matrix of $\hat{\beta}$ is twice that of the MLE computed on all n data points, assuming that $\tilde{\beta}$ is consistent. More generally, the factor of 2 improves to $1+1/c$ if we use acceptance probabilities $\max(c \cdot a(X, Y), 1)$ for $c > 1$ and assign weight $c \cdot a(X_i, Y_i)$ for samples with $c \cdot a(X_i, Y_i) > 1$. If we take $c = 4$ in our example, we obtain an 80%-efficient estimator by fitting a logistic regression on a subset of size roughly $n\,\mathbb{E}[\max(4a(X, Y), 1)] \approx 5000$, still 200 times smaller than $n = 1,001,000$. Thus $\hat{\beta}$ is a computationally efficient linear estimator that is provably almost as good as the MLE.

1.3 Selective Inference

Another strain of my research concerns post-selection inference. Some of my early research was on selective inference given independent test statistics $Y_i \overset{\text{i.i.d.}}{\sim} N(\mu_i, 1)$. Weinstein et al. (2013) inverts conditional tests to construct confidence intervals C_i for selected μ_i. These intervals are designed to control the *false coverage-statement rate* $\text{FCR} = \mathbb{E}\left[V/\max(1, R)\right]$, where R is the number of μ_i selected, and V is the number of noncovering intervals C_i for selected μ_i.

My recent work, discussed in Chapters 3–4, has dealt with the more general problem of inference after model selection. In Fithian et al. (2014), we view a statistical investigation as having two stages: first, a probabilistic model is chosen for the data, and some testing or other inference problem is posed in terms of unknown aspects of the sampling distribution — in short, a question is asked. Second, we use the model as a tool, along with the data, to perform an appropriate inference answering the question posed in stage 1. If the first stage is data-dependent, then a valid second-stage inference procedure must account for that dependence. Fithian et al. (2014) argues that a natural way to account for the randomness in stage 1 is by conditioning on the question asked. We propose to control the *selective type I error rate*: the chance of falsely rejecting a given true null hypothesis, conditional on that hypothesis being tested at all.

As a simple illustrative example, consider the problem of testing $H_0 : \mu = \mu_0$ in the model $Y \sim N(\mu, 1)$, where we only perform the test if $Y > 3$. Because the law $\mathcal{L}_\mu(Y \mid Y > 3)$ is an exponential family with natural parameter μ, there is a uniformly most powerful unbiased (UMPU) test of H_0. Inverting this test gives a confidence interval for μ. If Y barely passes the threshold, the confidence interval must be wide to account for the selection bias. By contrast, if Y is far above the threshold then $\mathcal{L}_\mu(Y \mid Y > 3) \approx N(\mu, 1)$ for all plausible μ, so the confidence interval is close to the nominal interval $Y \pm 1.96$.

More generally, we can derive for any exponential family model M, conditioned on any selection event A, finite-sample UMPU tests controlling the selective type I error at α for a variety of hypothesis testing and confidence interval problems. In particular, we obtain selective versions of the usual linear regression z-tests and t-tests, depending on whether the error variance σ^2 is known or unknown. In general, computing the rejection cutoffs can require computational finesse, but the computations are doable in various cases of interest.

Chapter 2

Local Case-Control Sampling

2.1 Introduction

In recent years statisticians, scientists, and engineers are increasingly analyzing enormous data sets. When data sets grow sufficiently large, computational costs may play a major role in the analysis, potentially constraining our choice of methodology or the number of data points we can afford to process. Computational savings can translate directly to statistical gains if they

1. open the door to using more sophisticated statistical techniques on a compressed data set,
2. allow us to refit our models more often to adapt to changing conditions,
3. allow for cross-validation, bagging, boosting, bootstrapping, or other computationally intensive statistical procedures, or
4. enable us to experiment with and prototype a variety of models, instead of trying only one or two.

Bottou and Bousquet (2008) discuss the tradeoffs arising when we adopt this point of view. One simple manifestation of these tradeoffs is that we may run out of computing resources before we run out of data, in effect making the sample size n a function of the efficiency of our fitting method.

2.1.1 Imbalanced Data Sets

Class imbalance is pervasive in modern classification problems and has received a great deal of attention in the machine learning literature (Chawla et al., 2004). It can come in two forms:

Marginal Imbalance One of the classes is quite rare; for instance, $\mathbb{P}(Y = 1) \approx 0$. Such imbalance typically occurs in data sets for predicting click-through rates in online advertising, detecting fraud, or diagnosing rare diseases.

Conditional Imbalance For most values of the features X, the response Y is very easy to predict; for instance, $\mathbb{P}(Y = 1 \mid X = 0) \approx 0$ but $\mathbb{P}(Y = 1 \mid X = 1) \approx 1$. For example, such imbalance might arise in the context of email spam filtering, where well-trained classifiers typically make very few mistakes.

Both or neither of the above may occur in any given data set. The machine learning literature on class imbalance usually focuses on the first type, but the second type is also common.

If, for example, our data set contains one thousand or one million negative examples for each positive example, then many of the negative data points are in some sense redundant. Typically in such problems, the statistical noise is primarily driven by the number of representatives of the rare class, whereas the total size of the sample determines the computational cost. If so, we might hope to finesse our computational constraints by subsampling the original data set in a way that enriches for the rare class. Such a strategy must be implemented with care if our ultimate inferences are to be valid for the full data set.

This chapter proposes one such data reduction scheme, local case-control sampling, for use in fitting logistic regression models. The method requires one parallelizable scan over the full data set and yields a potentially much smaller subsample containing roughly half of the information found in the original data set.

2.1.2 Subsampling

The simplest way to reduce the computational cost of a procedure is to subsample the data before doing anything else. However, uniform subsampling from an imbalanced data set is inefficient since it fails to exploit the unequal importance of the data points.

Case-control sampling — sampling uniformly from each class but adjusting the mixture of the classes — is a more promising approach. This procedure originated in epidemiology, where the positive examples (cases) are typically diseased patients and negative examples (controls) are disease-free (Mantel and Haenszel, 1959). Often, an equal number of cases and controls are sampled, resulting in a subsample with no marginal imbalance, and costly measurements of predictor variables are only made for selected patients (Breslow et al., 1980). This method is useful in our context as well, since a logistic regression model fitted on the subsample can be converted to a valid model for the original population via a simple adjustment to the intercept (Anderson, 1972; Prentice and Pyke, 1979).

However, standard case-control sampling still may not make most efficient use of the data. For instance, it does nothing to exploit conditional imbalance in a data set that is marginally balanced. Even with some marginal imbalance, a control that looks similar to the cases is often more useful for discrimination purposes than one that is obviously not a case.

We propose a method, local case-control sampling, which attempts to remedy imbalance *locally* throughout the feature space. Given a pilot estimate $(\tilde{\alpha}, \tilde{\beta})$ of the logistic regression parameters, local case-control sampling preferentially keeps data points for which Y is surprising given X. Specifically, if $\tilde{p}(x) = \frac{e^{\tilde{\alpha}+\tilde{\beta}'x}}{1+e^{\tilde{\alpha}+\tilde{\beta}'x}}$, we accept (x_i, y_i) with probability $|y_i - \tilde{p}(x_i)|$, the absolute residual of the pilot model. In the presence of extreme marginal or conditional imbalance, these errors will generally be quite small and the subsample can be many orders of magnitude smaller than the full data set.

Just as with case-control sampling, we can fit our model to the subsample and make an equally simple correction to obtain an estimate for the original data set. When the logistic regression model is correctly specified and the pilot is consistent and independent of the data, the asymptotic variance of the local case-control estimate is exactly twice the variance of a logistic regression fit on the (potentially much larger) full data set. This factor of two improves to $1 + \frac{1}{c}$ if we accept with probability $c|y_i - \tilde{p}(x_i)| \wedge 1$ and weight accepted points by a factor of $c|y_i - \tilde{p}(x_i)| \vee 1$. For example, if $c = 5$ then the variance of the subsampled estimate is only 20% greater than the variance of the full-sample MLE. The subsample we take with $c > 1$ is no more than c times larger than the subsample for $c = 1$, and for data sets with large imbalance is roughly $\frac{1+c}{2}$ times as large.

Local case-control sampling also improves on the bias of standard case-control sampling. When the logistic regression model is misspecified, case-control sampling is in general inconsistent for the risk minimizer in the original population. By contrast, local case-control sampling is always consistent given a consistent pilot, and is also asymptotically unbiased when the pilot is. Sections 2.5 and 2.6 present empirical results demonstrating the advantages of our approach in simulations and on the Yahoo! webspam data set.

2.1.3 Notation and Problem Setting

Our setting is that of predictive classification: we are given n observations, each consisting of a covariate vector $x_i \in \mathbb{R}^p$ and a binary response $y_i \in \{0, 1\}$, and our aim is to learn the function

$$p(x) = \mathbb{P}(Y = 1|X = x) \tag{2.1}$$

or equivalently to learn

$$f(x) = \operatorname{logit}(p(x)) = \log \frac{p(x)}{1 - p(x)} \tag{2.2}$$

which could be infinite for some x.

A linear logistic regression model assumes f is linear in x; that is,

$$f_\theta(x) = f_{\alpha,\beta}(x) = \alpha + \beta' x \tag{2.3}$$

where $\theta = (\alpha, \beta) \in \mathbb{R}^{p+1}$. This is less of a restriction than it might seem, since x may represent a very large basis expansion of some smaller set of “raw” features.

Although f is unlikely to satisfy our parametric model for any given basis x, under general conditions logistic regression in large samples will converge to the population maximizer of the expected log-likelihood

$$\theta^* = \arg\max_\theta \mathbb{E}(\ell(f_\theta(X); Y)) \tag{2.4}$$
$$= \arg\max_\theta \mathbb{E}\left[Y(\alpha + \beta' X) - \log\left(1 + e^{\alpha+\beta' X}\right)\right] \tag{2.5}$$

If $f = f_{\theta_0}$ for some θ_0, then $\theta^* = \theta_0$; otherwise f_{θ^*} is the best linear approximation to f in the sense of (2.4). Our goal here is to speed up computation while still obtaining a good estimate of θ^*.

2.1.4 Related Work

Recent years have seen substantial work on classification in imbalanced data sets. See Chawla et al. (2004) and He and Garcia (2009) for surveys of machine learning efforts on this problem. In particular, Owen (2007) discusses a limiting regime in which logistic regression converges to a sensible population estimate as marginal imbalance grows infinitely large. Many of the methods proposed for dealing with class imbalance involve some form of undersampling the majority class, oversampling the minority class, or both.

One recurring theme is to preferentially sampling negative examples that lie near positive examples in feature space. For example, Mani and Zhang (2003) proposes selecting majority-class examples whose average distance to its three nearest minority examples is smallest. Our method has a similar flavor since the probability of sampling a negative example $(x, 0)$ is $\tilde{p}(x)$, which is large when the features x are similar to those characteristic of positive examples.

Our proposal lies more in the tradition of the epidemiological case-control sampling literature. In particular, case-control sampling within several categorical strata has been studied by Fears and Brown (1986); Breslow and Cain (1988); Weinberg and Wacholder (1990); Scott and Wild (1991). Typically the strata are based on easy-to-measure screening variables available for a wide population, with more laborious-to-collect variables being measured on the sampled subjects.

Mineiro and Karampatziakis (2013) propose a somewhat similarly-inspired procedure as ours. For general empirical-risk minimization problems, they propose importance sampling examples with acceptance probability proportional to the loss $L(\theta; x, y)$ (capping probabilities above at 1 and below at a threshold $p_{\min}$), and adjusting for the biased sampling by computing a Horvitz–Thompson estimator. As a general recommendation, this proposal is somewhat odd in that it is not invariant to replacing $L(\cdot)$ with $L(\cdot) + c$. For logistic regression, however, their acceptance probabilities are quite correlated with ours, and so we include their method as a competitor in our simulations, using the value of $p_{\min}$ recommended in their paper.

2.2 Case-Control Subsampling

Although case-control sampling is commonly done by taking a fixed number of samples from each class, for our purposes it will be simpler to consider a nearly equivalent procedure based on accept-reject sampling.

Define some acceptance probability function $a(y)$ and let $b = \log \frac{a(1)}{a(0)}$, the log-selection bias. Consider the following algorithm:

1. Generate independent $z_i \sim \text{Bern}(a(y_i))$.
2. Fit a logistic regression to the subsample $S = \{(x_i, y_i) : z_i = 1\}$, obtaining unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$.
3. Assign $\hat{\alpha} \leftarrow \hat{\alpha}_S - b$ and $\hat{\beta} \leftarrow \hat{\beta}_S$.

This variant is convenient to analyze because the subsample thus obtained is an i.i.d. sample from a new population:

$$\mathbb{P}_S(X, Y) = \mathbb{P}(X, Y \mid Z = 1) = \frac{a(Y)\mathbb{P}(X, Y)}{\overline{a}} \tag{2.6}$$

with $\overline{a} = a(1)\mathbb{P}(Y = 1) + a(0)\mathbb{P}(Y = 0)$, the marginal probability of $Z = 1$.

The estimate $(\hat{\alpha}, \hat{\beta})$ is motivated by a simple application of Bayes' rule relating the odds of $Y = 1$ in $\mathbb{P}$ and $\mathbb{P}_S$. If $g(x)$ is the true conditional log-odds function for $\mathbb{P}_S$, we have

$$\begin{aligned} g(x) &= \log\left(\frac{\mathbb{P}(Y = 1 \mid X = x, Z = 1)}{\mathbb{P}(Y = 0 \mid X = x, Z = 1)}\right) && (2.7)\\ &= \log\left(\frac{\mathbb{P}(Y = 1 \mid X = x) \cdot \mathbb{P}(Z = 1 \mid Y = 1, X = x)}{\mathbb{P}(Y = 0 \mid X = x) \cdot \mathbb{P}(Z = 1 \mid Y = 0, X = x)}\right) && (2.8)\\ &= f(x) + b && (2.9) \end{aligned}$$

That is, the log-odds $g(x)$ in our biased population is simply a vertical shift by b of the log-odds $f(x)$ in the original population, so given an estimate of g we can subtract b to estimate f. If the model is correctly specified, logistic regression on the subsample yields a consistent estimate for the function $g(x)$, so the estimate for $f(x)$ is also consistent.

Note that the derivation (2.7-2.9) is equally valid if the sampling bias b depends on x, in which case we have $g(x) = f(x) + b(x)$. Local case-control sampling exploits this more general identity.

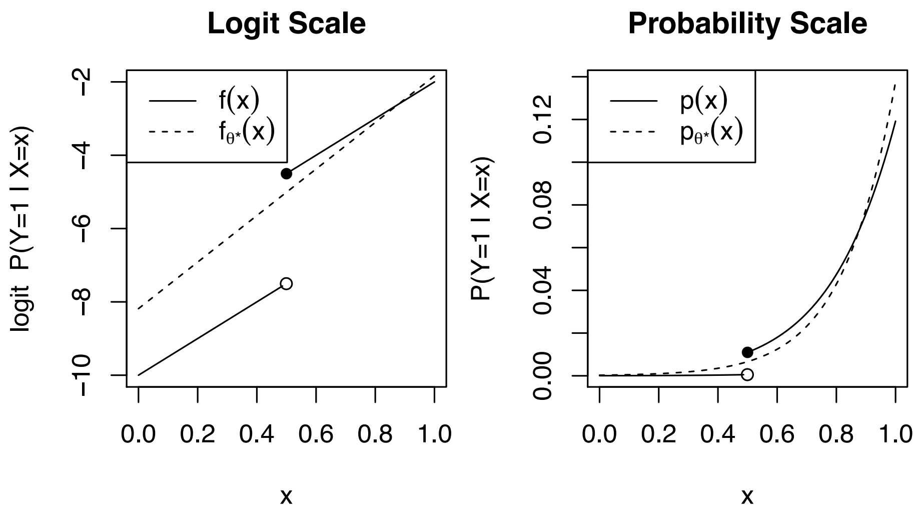


Figure 2.1: The best linear fit $f_{\theta^*}(x)$ approximates the true log-odds $f(x)$ in the sense of matching its implied conditional probabilities, not logits.

2.2.1 Conditional Probability and the Logit Loss

If X is integrable, then upon differentiating the population risk (2.4) with respect to θ we obtain the population score criterion:

$$0 = \mathbb{E}\left[\left(Y - \frac{e^{f_\theta(X)}}{1+e^{f_\theta(X)}}\right)\begin{pmatrix}1\\X\end{pmatrix}\right] \tag{2.10}$$

$$= \int (p(x) - p_\theta(x))\begin{pmatrix}1\\x\end{pmatrix} d\mathbb{P}(x) \tag{2.11}$$

Informally, the best linear predictor is the one that gets the conditional *probabilities* right on average. Note this is not the same as a predictor that gets the conditional log-odds right on average.

To illustrate the difference between approximating probabilities and approximating logits, suppose that $X \sim U(0,1)$ and $f(x) = -10 + 5x + 3 \cdot \mathbf{1}_{x>0.5}$. The left panel of Figure 2.1 shows $f(x)$ as a solid line and its best linear approximation as a dashed line. On the logit scale, the dashed line appears to be a very poor fit to the black curve. It fits reasonably well for large x, but it appears more or less to ignore the smaller values of x.

The right panel of Figure 2.1 shows why. When we transform both curves to the probability scale, the fit looks much more reasonable. $f_{\theta^*}(x)$ need not approximate $f(x)$ particularly well for small x, because in that range even a large change in the log-odds produces a negligible change in the conditional probability $p(x)$. By contrast, $f_{\theta^*}(x)$ needs to approximate $f(x)$ well for larger x, where $p(x)$ changes more rapidly.

In general, logistic regression places highest priority on fitting f where $\frac{d\,p(x)}{d\,f(x)}$ is largest: where

$f(x) \approx 0$ and $p(x) \approx 0.5$. In this example, with its strong marginal imbalance, the regions that matter most are those where $p(x)$ is largest. This often makes sense in applications such as medical screening or advertising click-through rate prediction, where accuracy is most important when the probability of disease or click-through is non-negligible. In Section 2.7 we consider how to modify the method to obtain classifiers that prioritize correctness near some other, user-defined level curve of $p(x)$.

Finally, note that Figure 2.1 suggests the case-control sampling estimate is unlikely to be consistent for θ^* in general. The nature of our linear approximation in the left panel is intimately related to the fact that $f(x) < 0$ everywhere in the sample space. If $f(x)$ were shifted upward by some constant, the response of the dashed curve would be more complicated than a simple constant shift, so estimating $f(x) + b$ and then subtracting b may not be a successful strategy.

2.2.2 Inconsistency of Case-Control Under Misspecification

If the linear model is misspecified, the case-control estimate is generically not consistent for the best linear predictor θ^* as $n \to \infty$ (Xie and Manski, 1989; Manski and Thompson, 1989). The unadjusted estimate will instead converge to the best linear predictor of g for the distribution $\mathbb{P}_S$, which solves the score criterion

$$0 = \int \left(\frac{e^{f(x)+b}}{1 + e^{f(x)+b}} - \frac{e^{f_\theta(x)}}{1 + e^{f_\theta(x)}} \right) \binom{1}{x} d\mathbb{P}_S(x) \tag{2.12}$$

Let $\theta^*_{\mathrm{CC}}(b)$ be the large-sample limit of the *adjusted* case-control sampling estimate with bias b. Then $\theta^*_{\mathrm{CC}}(b)$ solves the population score criterion

$$0 = \int \left(\frac{e^{f(x)+b}}{1 + e^{f(x)+b}} - \frac{e^{f_\theta(x)+b}}{1 + e^{f_\theta(x)+b}} \right) \binom{1}{x} d\mathbb{P}_S(x) \tag{2.13}$$

which differs from (2.10) in two ways. First, the integral is taken over a different distribution for X. Second, and more importantly, the integrand is different. We are now approximating $f(x)$ in a different sense than we were.

In general under misspecification, $\theta^*_{\mathrm{CC}}(b)$ is different for every b. If we sample cases and controls equally, in the limit we will get a different answer than if we sample twice as many controls; and in either case we will get a different answer than if we use the entire data set or subsample uniformly.

These differences can be quite consequential for our inferences about β or the predictive performance of our model, as we see next.

Example 1: Oatmeal and Disease Risk In this fictitious example we consider estimating the effect of exposure to oatmeal on a person's risk of developing some rare disease. Suppose that 10% of the population has a family history of the disease, half the population eats oatmeal (independently of family history), and that both exposure and family history are binary predictors. Suppose further

Original Population ($\mathbb{P}$)

Conditional Log-Odds (f)

	History -	History +
Oat -	-5	-4
Oat +	-10	-1

Conditional Probabilities

	History -	History +
Oat -	.007	.02
Oat +	5E-5	.37

Case-Control Population ($\mathbb{P}_S$)

Conditional Log-Odds (g)

	History -	History +
Oat -	-1.2	-.2
Oat +	-6.2	2.8

Conditional Probabilities

	History -	History +
Oat -	.24	.46
Oat +	.002	.94

Table 2.1: Disease risk in the full population, and in the population created by case-control sampling with equal numbers in each class.

that the true conditional log-odds function $f(x)$ is given by the top-left panel of Table 2.1.

The corresponding conditional probabilities $p(x)$ are given in the lower-left panel of Table 2.1. Notice that oatmeal increases the risk for people who are already at risk by virtue of their family history, but has a protective effect for everyone else. This interaction means that the additive logistic regression model is misspecified.

Because only the probabilities in the "History +" column are large enough to matter, the fitted model for $f(x)$ pays more attention to the at-risk population, for whom oatmeal elevates the risk of disease. A logistic regression on a large sample from this population estimates the coefficient for oatmeal as $\beta^*_{\text{Oatmeal}} = 1.4$, implying an odds ratio of about 4.0. This is close to the marginal odds ratio of roughly 4.3 that we would obtain if we did not control for family history.

Suppose however that we sampled an equal number of cases and controls. Then the conditional log-odds of disease in our sample would reflect the top-right panel of Table 2.1, with all cells increased by the same amount.

For large samples, the case-control estimate is $\beta^*_{\text{CC,Oatmeal}} = -0.83$, implying an odds ratio of about 0.44. Using case-control sampling has reversed our inference about the effect of oatmeal exposure, because after shifting the log-odds the left column becomes much more important.

Example 2: Two-Class Gaussian Model Suppose that $\mathbb{P}(Y = 1) = 1\%$, and that $X \mid Y \sim \mathrm{N}(\mu_Y, \Sigma_Y)$. Let

$$\mu_0 = (0, 0) \quad \Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{2.14}$$

$$\mu_1 = (1.5, 1.5) \quad \Sigma_1 = \begin{pmatrix} 0.3 & 0 \\ 0 & 5 \end{pmatrix} \tag{2.15}$$

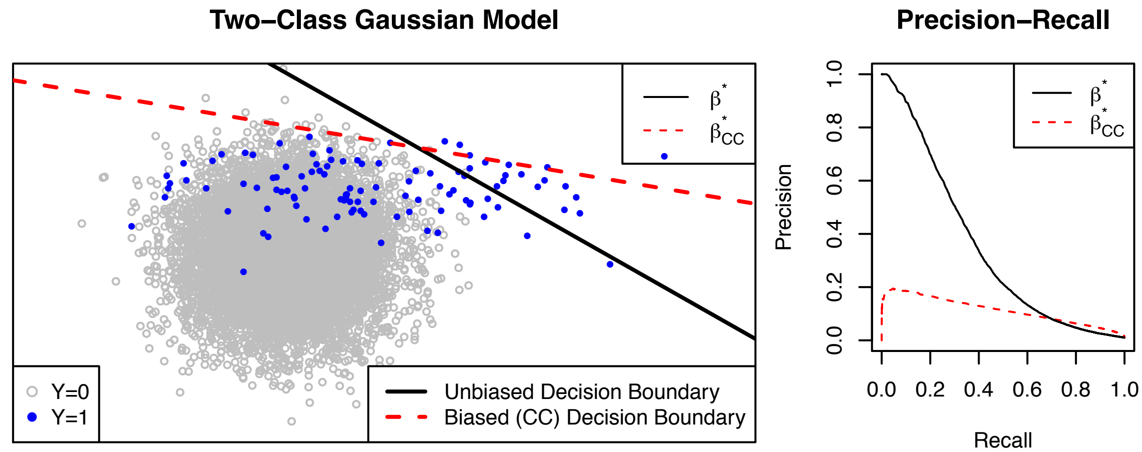


Figure 2.2: At left, biased (case-control) and unbiased decision boundaries for the bivariate Gaussian mixture model. At right, precision-recall curves for β^* and β^*_{CC}.

Data simulated from this model are shown in the left panel of Figure 2.2. In this example, the true log-odds $f(x)$ is an additive quadratic function of the two coordinates X_1 and X_2.

In this example as in the previous one, the population-optimal case-control parameters differ substantially from the optimal parameters in the original population, with dramatic effects for the predictive performance of the model. The decision boundaries for the two estimates are overlayed on the left panel of Figure 2.2. In the right panel, we plot the precision-recall curves resulting from each set of parameters.

2.2.3 Weighted Case-Control Sampling

A simple alternative to standard case-control sampling is to weight the subsampled data points by the inverse of their probability of being sampled. We include weighted case-control sampling as a competitor in our simulation studies in Section 2.5. This sampling scheme succeeds in removing the bias induced by the case-control sampling, but at a cost of increasing the variance since the effective sample size is reduced (Scott and Wild, 1986, 2002).

2.3 Local Case-Control Subsampling

In this section, we describe local case-control subsampling, a generalization of standard case-control sampling that both improves on its efficiency and resolves its problem of inconsistency. To achieve these benefits we require a pilot estimate, i.e. a good guess $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta})$ for the population-optimal θ^*.

2.3.1 The Local Case-Control Sampling Algorithm

Local case-control sampling differs from case-control sampling only in that the acceptance probability a is allowed to depend on x as well as y. Our criterion for selection will be the degree of "surprise" we experience upon observing y_i given x_i:

$$a(x, y) = |y - \tilde{p}(x)| = \begin{cases} 1 - \tilde{p}(x) & y = 1 \\ \tilde{p}(x) & y = 0 \end{cases} \tag{2.16}$$

where $\tilde{p}(x) = \frac{e^{\tilde{\alpha}+\tilde{\beta}'x}}{1+e^{\tilde{\alpha}+\tilde{\beta}'x}}$ is the pilot estimate of $\mathbb{P}(Y = 1 \,|\, X = x)$. The algorithm is:

1. Generate independent $z_i \sim \text{Bern}(a(x_i, y_i))$
2. Fit a logistic regression to the sample $S = \{(x_i, y_i) : z_i = 1\}$ to obtain unadjusted estimates $\hat{\theta}_S = (\hat{\alpha}_S, \hat{\beta}_S)$.
3. Assign $\hat{\alpha} \leftarrow \hat{\alpha}_S + \tilde{\alpha}$ and $\hat{\beta} \leftarrow \hat{\beta}_S + \tilde{\beta}$

As in Section 2.2, the adjustment is justified by (2.7-2.9), only now with the constant b replaced by

$$b(x) = \log\left(\frac{a(x, 1)}{a(x, 0)}\right) = -\tilde{\alpha} - \tilde{\beta}'x \tag{2.17}$$

In other words, the subsample is drawn from a measure with

$$g(x) = f(x) - \tilde{\alpha} - \tilde{\beta}'x \tag{2.18}$$

If $f(x)$ is well approximated by the pilot estimate, then $g(x) \approx 0$ throughout feature space. That is, conditional on selection into S, y_i given x_i is nearly a fair coin toss.

Recall that the Fisher information for a Bernoulli random variable with natural parameter η and mean $p_\eta = \frac{e^\eta}{1+e^\eta}$ is $p_\eta(1 - p_\eta)$. Since this quantity is maximized when $\eta = 0$ and $p_\eta = 0.5$, fair coin tosses are more informative than heavily biased ones. In effect, local case-control sampling tilts the conditional distribution $\mathcal{L}(Y \,|\, X = x)$ by the amount $-\tilde{\alpha} - \tilde{\beta}'x$, making each y_i in the subsample more informative. The algorithm can be described as follows:

1. Tilt all the $\mathbb{P}(Y|X = x)$ toward the most favorable sampling measure.
2. Fit a logistic regression to the sample from the tilted measure.
3. Tilt back to obtain an estimate for the original population.

In marginally imbalanced data sets where $\mathbb{P}(Y = 1|X = x)$ is small everywhere in the predictor space, a good pilot has $\tilde{p}(x) \approx 0$ for all x, and the number of cases discarded by this algorithm will be quite small. If we wish to avoid discarding any cases, we can always modify the algorithm so that instead of keeping $(x, 1)$ with probability $a(x, 1)$, we keep it with probability 1 and assign weight $a(x, 1)$.

2.3.2 Choosing the Pilot Fit

In many applications, there may be a natural choice of pilot fit $\tilde{\theta}$; for instance, if we are re-fitting a classification model every week to adapt to a changing world, then last week's fit is a natural choice for this week's pilot.

If no pilot fit is available from such a source, we recommend an initial pass of weighted case-control sampling (described in Section 2.2.2) to obtain the pilot. Because weighted case-control sampling is itself consistent and asymptotically unbiased for the true parameters, the entire procedure would then enjoy consistency and asymptotic unbiasedness per the results in Section 2.4.

Our simulations suggest that mild inaccuracy in the pilot estimate does not unduly derade the performance of the local case-control algorithm. The main role of the pilot fit is to guide us in discarding most of the data points for which y_i is obvious given x_i while keeping those for which y_i is conditionally surprising.

Because standard case-control sampling amounts to local case-control sampling with a constant-only pilot fit, we might expect that the pilot fit need not be perfect to improve upon case-control sampling. Our experiments in Sections 2.5 and 2.6 support this intuition.

2.3.3 Taking a Larger or Smaller Sample

As we will see in Section 2.4.3, under correct model specification the baseline procedure outlined above has exactly twice the asymptotic variance as a logistic regression estimated with the full sample, despite using a potentially very small subset of the data. We can improve upon this factor of two by increasing the size of the subsample.

One simple way to achieve this is to multiply all acceptance probabilities by some constant c, e.g. $c = 5$. When deciding whether to sample the point (x_i, y_i), we would then generate $z_i \sim \text{Bern}(ca(x_i, y_i) \wedge 1)$ and assign weight $w_i = ca(x_i, y_i) \vee 1$ to each sampled point. This amounts to a larger, weighted subsample from $\mathbb{P}_S$, and we can make the same correction to the estimates from the subsample. We see in Section 2.4.4 that for $c > 1$ the factor of two is replaced by a factor of $1 + \frac{1}{c}$.

In the case of large imbalance, most of the $\tilde{p}(x_i)$ are near 0 or 1. For $c > 1$ the marginal acceptance probability at x_i becomes

$$\mathbb{P}(z_i = 1 \,|\, x_i = x) = p(x)(c(1 - \tilde{p}(x)) \wedge 1) + (1 - p(x))(c\tilde{p}(x) \wedge 1) \tag{2.19}$$

$$\approx (1 + c)p(x)(1 - p(x)) \tag{2.20}$$

where the approximation holds for $p(x) \approx \tilde{p}(x) \approx 0$ or 1. For $c = 1$, the marginal acceptance probability is $p(x)(1-\tilde{p}(x)) + (1-p(x))\tilde{p}(x) \approx 2p(x)(1-p(x))$, so for $c > 1$ we take roughly $\frac{1+c}{2}$ times as many data points as for $c = 1$. For example, if $c = 5$, the subsample accepted is roughly 3 times as large, and the relative efficiency improves from 1/2 to 5/6.

Alternatively, if n is extremely large, even a small fraction of the full data set may still be too large a sample. In that case we can proceed as above with $c < 1$, or simply sample any desired number n_s of data points uniformly from the local case-control subsample.

2.4 Asymptotics of the Local Case-Control Estimate

We now turn to examining the asymptotic behavior of the local case-control estimate. We first establish consistency, assuming a consistent pilot estimate $\tilde{\theta}$. We expressly do not assume that the pilot estimate is independent of the data, since in some cases we may recycle into the subsample some of the data we used to calculate the pilot.

Assuming independence of $\tilde{\theta}$ and the data gives finer results about the asymptotic distribution of $\hat{\theta}$. It is asymptotically unbiased when $\tilde{\theta}$ is, and derive the asymptotic variance of the estimate. When the logistic regression model is correctly specified, the local case-control estimate has exactly twice the asymptotic variance of the MLE for the full data set.

2.4.1 Preliminaries

For better clarity of notation in this section, we will use the letter λ in place of $\tilde{\theta}$ to denote pilot estimate. Additionally, we drop the notation $\binom{1}{x}$ and assume x possibly includes a constant term, so that $f_\theta(x) = \theta' x$.

The local case-control subsampling scheme for pilot λ effectively samples from the probability measure $\mathbb{P}_\lambda$, where

$$d\mathbb{P}_\lambda(x, y) = \frac{a_\lambda(x, y) d\mathbb{P}(x, y)}{\overline{a}(\lambda)}, \tag{2.21}$$

and $\overline{a}(\lambda) = \int a_\lambda(x, y)\, d\mathbb{P}(x, y)$ is the marginal probability of acceptance.

The population risk of the logistic regression parameters θ with respect to sampling measure $\mathbb{P}_\lambda$ is

$$R_\lambda(\theta) = -\int \left[\frac{e^{f(x)-\lambda' x}}{1 + e^{f(x)-\lambda' x}} \theta' x - \log\left(1 + e^{\theta' x}\right) \right] d\mathbb{P}_\lambda(x) \tag{2.22}$$

$$= \frac{-1}{\overline{a}(\lambda)} \int \left[\frac{e^{f(x)-\lambda' x}}{1 + e^{f(x)-\lambda' x}} \theta' x - \log\left(1 + e^{\theta' x}\right) \right] \left[\frac{e^{\lambda' x} + e^{f(x)}}{(1 + e^{\lambda' x})(1 + e^{f(x)})} \right] d\mathbb{P}(x) \tag{2.23}$$

We obtain (2.22) by conditioning on X, since $\operatorname{logit} \mathbb{P}_\lambda(Y = 1 \mid X = x) = f(x) - \lambda' x$. The second factor in (2.23) arises from the fact that the marginal acceptance probability given x is

$$\tilde{p}(x)(1-p(x)) + (1-\tilde{p}(x))p(x) = \frac{e^{\lambda' x} + e^{f(x)}}{(1+e^{\lambda' x})(1+e^{f(x)})} \tag{2.24}$$

Note that the function $h(\eta) = -A\eta + \log(1+e^{\eta})$ for $A \in [0,1]$, which plays a major role, is strictly convex, its magnitude is bounded by $1+|\eta|$, and it has Lipschitz constant ≤ 1. As a function of λ, $a_\lambda(x,y) = \left|y - \frac{e^{\lambda' x}}{1+e^{\lambda' x}}\right|$ is strictly positive, bounded by 1, and has Lipschitz constant $\leq \|x\|$. By the same token, $\overline{a}(\lambda) = \mathbb{E}a_\lambda(X,Y) \in (0,1)$ with Lipschitz constant $\leq \mathbb{E}\|X\|$.

By Cauchy-Schwarz, the integrand in (2.23) is bounded by $2(1+\|\theta\|\|x\|)$. If $\mathbb{E}\|X\| < \infty$, then, we may appeal to dominated convergence and take limits with respect to θ and λ inside the integral.

Writing $R_\lambda(\theta) = \mathbb{E}(r_\lambda(\theta' X))$, note that r_λ is a strictly convex function. It follows that $R_\lambda(\theta)$ is strictly convex as well, provided there is no v for which $\mathbb{E}|v'X| = 0$. Additionally, assume that $\mathbb{P}(Y=1 \mid X=x) \in (0,1)$ on some neighborhood of x with positive measure, so that $R_\lambda(\theta) \to \infty$ as $\theta \to \infty$ in any given direction. Consequently, $R_\lambda(\theta)$ attains a unique minimum.

Denote by $\widehat{R}_\lambda^{(0)}(\theta)$ the empirical risk on a local case-control subsample taken using the pilot estimate λ. Assume that, given the data and the pilot, the acceptance decisions z_i are independent with success probability $a_\lambda(x_i, y_i)$. Then

$$\widehat{R}_\lambda^{(0)}(\theta) = -\left(\sum_{i=1}^n z_i\right)^{-1} \sum_{i=1}^n z_i \left[y_i \theta' x_i - \log\left(1+e^{\theta' x_i}\right)\right] \tag{2.25}$$

It will be somewhat simpler to replace the random subsample size $\sum_{i=1}^n z_i$ with its expectation $n\overline{a}(\lambda)$. Define

$$\widehat{R}_\lambda(\theta) = -\frac{1}{n\overline{a}(\lambda)} \sum_{i=1}^n z_i \left[y_i \theta' x_i - \log\left(1+e^{\theta' x_i}\right)\right] \tag{2.26}$$

Since minimizing (2.25) with respect to θ is equivalent to minimizing (2.26), the two are equivalent for our purposes.

If the unadjusted parameters $\hat{\theta}_S$ minimize $\widehat{R}_\lambda$, the local case-control estimate $\hat{\theta} = \hat{\theta}_S + \lambda$ is an M-estimator minimizing $\widehat{Q}_\lambda(\theta) = \widehat{R}_\lambda(\theta - \lambda)$. We use analogous notation for the population version:

$$Q_\lambda(\theta) = R_\lambda(\theta - \lambda) \tag{2.27}$$

For any given pilot estimate λ and large n we expect

$$\hat{\theta} \approx \arg\min_\theta Q_\lambda(\theta) \tag{2.28}$$

Define the right-hand side of (2.28) to be $\overline{\theta}(\lambda)$, the large-sample limit of local case-control sampling with pilot estimate fixed at λ. The best linear predictor for the original population corresponds to the case $\lambda = 0$ (uniform subsampling), i.e. $\theta^* = \overline{\theta}(0)$. Consistency means that for large n, $\hat{\theta} \xrightarrow{p} \theta^*$.

Recall that if the model is correctly specified with true parameters θ_0, then $\overline{\theta}(\lambda) = \theta_0$ for *any* fixed pilot estimate λ. Minimizing $\widehat{Q}_\lambda$ therefore yields a consistent estimate. Unfortunately, in the misspecified case $\overline{\theta}(\lambda) \neq \overline{\theta}(0) = \theta^*$. In this sense, local case-control sampling with the pilot λ held fixed is in general *not* consistent for θ^*. However, we see below that it is consistent if $\lambda = \theta^*$.

Proposition 1. *Suppose X is integrable and that $\theta^* = \overline{\theta}(0)$ is the best linear predictor for the untilted population $\mathbb{P}_0$. Then*

$$\theta^* = \arg\min_\theta Q_{\theta^*}(\theta) = \overline{\theta}(\theta^*) \tag{2.29}$$

In other words, if we could only choose our pilot perfectly, then the local case-control estimate would converge to θ^* as $n \to \infty$.

Proof. The population optimality criterion is

$$\begin{aligned} 0 &= -\overline{a}(\lambda)\nabla_\theta Q_\lambda(\theta) && (2.30) \\ &= \int \left[\frac{e^{f(x)-\lambda' x}}{1+e^{f(x)-\lambda' x}} - \frac{e^{(\theta-\lambda)' x}}{1+e^{(\theta-\lambda)' x}} \right] \left[\frac{e^{\lambda' x} + e^{f(x)}}{(1+e^{\lambda' x})(1+e^{f(x)})} \right] x \, d\mathbb{P}(x) && (2.31) \end{aligned}$$

If we evaluate the above at $\lambda = \theta = \theta^*$, after some simplifications we obtain

$$\begin{aligned} -\overline{a}(\theta^*)\nabla_\theta Q_{\theta^*}(\theta^*) &= \frac{1}{2} \int \frac{e^{f(x)} - e^{\theta^{*\prime} x}}{\left(1+e^{f(x)}\right)\left(1+e^{\theta^{*\prime} x}\right)} x \, d\mathbb{P}(x) && (2.32) \\ &= \frac{1}{2} \int \left(\frac{e^{f(x)}}{1+e^{f(x)}} - \frac{e^{\theta^{*\prime} x}}{1+e^{\theta^{*\prime} x}} \right) x \, d\mathbb{P}(x) && (2.33) \\ & && (2.34) \end{aligned}$$

which is exactly the population score (2.10) for the original population. Since θ^* optimizes the risk for the original population, this value is 0.

□

Of course, in practice we will never have a perfect pilot — if we did, we wouldn't need to estimate θ^* — but Proposition 1 suggests that if λ is near θ^*, minimizing $\widehat{Q}_\lambda$ yields a good estimate. In fact we will see that if $\lambda \xrightarrow{p} \theta^*$ then $\hat{\theta} \xrightarrow{p} \theta^*$ as well.

2.4.2 Consistency

Consider a sequence of data sets with size n tending to infinity. The main result of this section is that if our pilot estimate λ_n is consistent for θ^*, then so is the local case-control estimate $\hat{\theta}_n$. $\mathbb{E}\|X\| < \infty$ is assumed throughout.

First, we establish pointwise convergence of the function we actually minimize to the function we would prefer to minimize:

Proposition 2. *If $\lambda_n \xrightarrow{p} \lambda_\infty$, then for each $\theta \in \mathbb{R}^{p+1}$,*

$$\widehat{Q}_{\lambda_n}(\theta) \xrightarrow{p} Q_{\lambda_\infty}(\theta) \tag{2.35}$$

Because we avoid assuming independence between the pilot λ_n and the data (x_i, y_i), the proof is somewhat technical and is deferred to the Appendix. The proof relies on coupling the acceptance decisions z_i for different pilot estimates through a shared uniform random variable. With this coupling two nearby pilot estimates will differ on very few accept-reject decisions.

Second, we control the oscillation of the functions in (2.35) to obtain uniform convergence on compact sets Θ:

Proposition 3. *If $\lambda_n \xrightarrow{p} \lambda_\infty$, then for compact $\Theta \subseteq \mathbb{R}^{p+1}$,*

$$\sup_{\theta \in \Theta} |\widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta)| \xrightarrow{p} 0 \tag{2.36}$$

Proof. Define

$$F_n(\theta) = \widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta) \tag{2.37}$$

Proposition 2 implies $F_n(\theta) \xrightarrow{p} 0$ pointwise. Next we establish that F_n is Lipschitz, handling each term separately.

Note that the integrand in (2.31) is x times two factors that each are bounded by ± 1, hence

$$\|\overline{a}(\lambda_\infty)\nabla_\theta Q_{\lambda_\infty}\| \leq \int \|x\| \, d\mathbb{P}(x) = \mathbb{E}\|X\| \tag{2.38}$$

Similarly for $\widehat{Q}_{\lambda_n}$, we have

$$\nabla_\theta \widehat{Q}_{\lambda_n} = -\frac{1}{n\overline{a}(\lambda_n)} \sum_{i=1}^n z_i \left(y_i - \frac{e^{(\theta-\lambda_n)' x_i}}{1 + e^{(\theta-\lambda_n)' x_i}} \right) x_i \tag{2.39}$$

so that

$$\sup_\theta \|\nabla_\theta \widehat{Q}_{\lambda_n}\| \leq \overline{a}(\lambda_n)^{-1} \frac{1}{n} \sum_{i=1}^n \|x_i\| \tag{2.40}$$

$$\xrightarrow{p} \overline{a}(\lambda_\infty)^{-1} \mathbb{E}\|X\| \tag{2.41}$$

It follows that, with probability tending to 1, $F_n(\theta)$ has Lipschitz constant less than $c = 3\overline{a}(\lambda_\infty)^{-1}\mathbb{E}\|X\|$.

Now, for any $\varepsilon > 0$, we can cover Θ with finitely many Euclidean balls of radius $\delta = \varepsilon / c$, centered at $\theta_1, \ldots, \theta_{M(\varepsilon)}$. Let $A_n(\varepsilon)$ be the event that F_n has Lipschitz constant less than c and

$$\sup_{1 \le j \le M(\varepsilon)} |F_n(\theta_j)| < \varepsilon \tag{2.42}$$

On $A_n(\varepsilon)$, we have $\sup_{\theta \in \Theta} |F_n(\theta)| < 2\varepsilon$, and $\mathbb{P}(A_n(\varepsilon)) \to 1$ as $n \to \infty$. $\square$

Finally we come to the main result of the section, in which we prove that the local case-control estimate is consistent when the pilot is.

Theorem 4. *If $\lambda_n \xrightarrow{p} \theta^*$ then the local case-control estimate $\hat{\theta}_n \xrightarrow{p} \theta^*$ as well.*

Proof. Let $\Theta \in \mathbb{R}^{p+1}$ be any compact set with θ^* in its interior, and let

$$\varepsilon = \inf_{\theta \in \partial\Theta} Q_{\theta^*}(\theta) - Q_{\theta^*}(\theta^*) > 0 \tag{2.43}$$

where the strict inequality follows from strict convexity. Uniform convergence implies that with probability tending to 1,

$$\sup_{\theta \in \Theta} |\widehat{Q}_{\lambda_n}(\theta) - Q_{\theta^*}(\theta)| < \varepsilon/2 \tag{2.44}$$

which implies in turn that

$$\inf_{\theta \in \partial\Theta} \widehat{Q}_{\lambda_n}(\theta) > \widehat{Q}_{\lambda_n}(\theta^*) \tag{2.45}$$

Whenever this is the case, the strictly convex function $\widehat{Q}_{\lambda_n}$ has a unique minimizer in the interior of Θ. Since Θ was arbitrary, we can take its diameter to be less than any $\delta > 0$. Hence, $\hat{\theta}_n \xrightarrow{p} \theta^*$. $\square$

2.4.3 Asymptotic Distribution

In this section we derive the asymptotic distribution of the local case-control logistic regression estimate. To simplify matters we assume the pilot estimate λ is independent of our data set. This would not be the case if our pilot were based on a subsample of the data (the procedure we use for all our simulations), but it could hold if the pilot came from a model fitted to data from an earlier time period.

The main result of this section is that if the logistic regression model is correctly specified and the pilot is consistent, the asymptotic covariance matrix of the local case-control estimate for θ is exactly twice the asymptotic covariance matrix of a logistic regression performed on the entire data set.

We assume throughout this section that $\mathbb{E}\|X\|^2 < \infty$. It will be convenient to give names to some recurring quantities. First, we have seen that if $\mathbb{E}\|X\| < \infty$ we can differentiate $Q_\lambda(\theta)$ inside the integral to obtain the gradient of the population risk:

$$
\begin{aligned}
G(\theta, \lambda) &\triangleq -\overline{a}(\lambda)\nabla_\theta Q_\lambda(\theta) && (2.46)\\
&= \int \left(y - \frac{e^{(\theta-\lambda)'x}}{1+e^{(\theta-\lambda)'x}} \right) x \, d\mathbb{P}_\lambda(x, y) && (2.47)\\
&= \int \left(\frac{e^{f(x)-\lambda'x}}{1+e^{f(x)-\lambda'x}} - \frac{e^{(\theta-\lambda)'x}}{1+e^{(\theta-\lambda)'x}} \right) \left(\frac{e^{\lambda'x}+e^{f(x)}}{(1+e^{\lambda'x})(1+e^{f(x)})} \right) x \, d\mathbb{P}(x) && (2.48)
\end{aligned}
$$

Whereas (2.47) is the expectation of the logistic regression score with respect to $\mathbb{P}_\lambda$, we can also define its covariance matrix, which is finite since $\mathbb{E}_\lambda \|X\|^2$ is:

$$
J(\theta, \lambda) \triangleq \mathrm{Var}_\lambda \left[\left(Y - \frac{e^{(\theta-\lambda)'X}}{1+e^{(\theta-\lambda)'X}} \right) X \right] \quad (2.49)
$$

The above is continuous in θ and λ by dominated convergence.

Since the derivatives of the integrand in (2.48) are uniformly bounded by $2\|x\|^2$, dominated convergence implies we can again differentiate inside the integral. Differentiating with respect to θ we obtain

$$
\begin{aligned}
H(\theta, \lambda) &\triangleq -\overline{a}(\lambda)\nabla^2_\theta Q_\lambda(\theta) && (2.50)\\
&= \int \frac{e^{(\theta-\lambda)'x}}{(1+e^{(\theta-\lambda)'x})^2} \left(\frac{e^{\lambda'x}+e^{f(x)}}{(1+e^{\lambda'x})(1+e^{f(x)})} \right) xx' \, d\mathbb{P}(x) && (2.51)
\end{aligned}
$$

Here the integrand is dominated by xx', so dominated convergence again applies and thus we see that H is continuous in θ and λ. $H(\theta, \lambda) \succ 0$ for any θ, λ provided there is no nonzero v for which $\mathbb{E}|v'X| = 0$. Finally, define the matrix of cross partials:

$$
C(\theta, \lambda) \triangleq \nabla_\lambda G(\theta, \lambda) \quad (2.52)
$$

To be concrete, $C_{i,j} = \frac{\partial^2}{\partial\theta_i \, \partial\lambda_j} Q_\lambda(\theta)$. Continuity of C again follows from noting the derivative of the integrand in (2.48) with respect to λ is dominated by $8\|x\|^2$.

To begin we consider the behavior of $\overline{\theta}(\lambda)$ for λ near θ^*. By Proposition 1 we have $G(\theta^*, \theta^*) = 0$. Since $H(\theta, \lambda) \succ 0$, we can apply the Implicit Function Theorem to the relation $G(\overline{\theta}(\lambda), \lambda) = 0$ to obtain

$$
\overline{\theta}(\lambda) = \theta^* + H(\theta^*, \theta^*)^{-1} C(\theta^*, \theta^*)(\lambda - \theta^*) + o(\|\lambda - \theta^*\|) \quad (2.53)
$$

By standard M-estimator theory, if we fix λ and send $m \to \infty$ the coefficients of a logistic regression performed on a sample of size $|S|$ from $\mathbb{P}_\lambda$ would be asymptotically normal with covariance matrix $\frac{1}{|S|} H(\overline{\theta}(\lambda), \lambda)^{-1} J(\overline{\theta}(\lambda), \lambda) H(\overline{\theta}(\lambda), \lambda)^{-1}$. In light of this and the fact that $|S| \approx \overline{a}(\lambda)n$, we might predict the following:

Theorem 5. *Assume* $\mathbb{E}\|X\|^2 < \infty$*. If* $\lambda \xrightarrow{p} \theta^*$ *independently of the data, then*

$$\sqrt{n}\left(\hat{\theta} - \overline{\theta}(\lambda)\right) \xrightarrow{\mathcal{D}} N(0, \overline{a}(\theta^*)^{-1}\Sigma) \tag{2.54}$$

with $\Sigma = H(\theta^*, \theta^*)^{-1} J(\theta^*, \theta^*) H(\theta^*, \theta^*)^{-1}$*.*

Again, we defer the proof to the Appendix. We can combine (2.54) with (2.53) to immediately obtain the following reassuring facts:

Corollary 6. *Assume* $\mathbb{E}\|X\|^2 < \infty$ *and* λ_n *is a sequence of pilot estimators given independently of the data. Then*

(a) If λ_n *is* $\sqrt{n}$*-consistent, so is* $\hat{\theta}_n$*.*

(b) If λ_n *is asymptotically unbiased, so is* $\hat{\theta}_n$*.*

(c) If $\sqrt{n}\,(\lambda_n - \theta^*) \xrightarrow{\mathcal{D}} N(0, V)$ *then* $\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{\mathcal{D}} N\left(0, \Sigma\right)$ *with*

$$\Sigma = H^{-1}\left(CVC' + \overline{a}^{-1}J\right)H^{-1} \tag{2.55}$$

In (2.55) *we have suppressed the arguments of* θ^* *in* $H, C, \overline{a}$*, and* J*.*

In the special case where logistic regression model is correctly specified, we have the following:

Theorem 7. *Assume the logistic regression model is correct and let* $\frac{1}{n}\Sigma_{\text{full}}$ *be the asymptotic variance of the MLE for the full sample. Then if* $\mathbb{E}\|X\|^2 < \infty$ *and* $\lambda \xrightarrow{p} \theta_0$ *independently of the data, we have*

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{D}} N\left(0, a(\theta_0)^{-1}\Sigma\right) = N\left(0, 2\Sigma_{\text{full}}\right) \tag{2.56}$$

Hence, although the size of a local case-control subsample is roughly $n\overline{a}(\lambda)$, the variance of $\hat{\theta}$ is the same as if we took a uniform sample of size $n/2$ from the full data set. In other words, each point sampled is worth about $\frac{1}{2\overline{a}(\tilde{\theta})}$ points sampled uniformly.

Proof. If logistic regression is correctly specified for $\mathbb{P}$, it is also for $\mathbb{P}_\lambda$, regardless of λ, so $\overline{\theta}(\lambda) \equiv \theta_0$. Furthermore, by standard maximum likelihood theory $J(\theta_0, \lambda) = H(\theta_0, \lambda)^{-1}$ for each λ. Therefore, (2.54) specializes to

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{\mathcal{D}} N(0, \overline{a}(\theta_0)^{-1} H(\theta_0, \theta_0)^{-1}) \tag{2.57}$$

But

$$H(\theta, \lambda) = \overline{a}(\lambda)^{-1} \int \left[\frac{e^{(\theta-\lambda)'x}}{(1 + e^{(\theta-\lambda)'x})^2}\right]\left[\frac{e^{\lambda'x} + e^{f(x)}}{(1 + e^{\lambda'x})(1 + e^{f(x)})}\right] xx'\, d\mathbb{P}(x) \tag{2.58}$$

If $f(x) = \theta_0' x$ and $\lambda = \theta_0$, then (2.58) simplifies to

$$H(\theta_0, \theta_0) = \overline{a}(\theta_0)^{-1} \frac{1}{2} \int \frac{e^{\theta_0' x}}{(1 + e^{\theta_0' x})^2} x x' \, d\mathbb{P}(x) \tag{2.59}$$

$$= \overline{a}(\theta_0)^{-1} \frac{1}{2} H(\theta_0, 0) \tag{2.60}$$

$$= (2\overline{a}(\theta_0)\Sigma_{\text{full}})^{-1} \tag{2.61}$$

$\square$

Notice that before subsampling, the acceptance probability given X is 1 for every value x, but the second derivative of the log-likelihood for an observation with $X = x$ is $-\frac{e^{\theta_0' x}}{(1+e^{\theta_0' x})^2}$. After subsampling with pilot $\lambda = \theta_0$, the acceptance probability is $2\frac{e^{\theta_0' x}}{(1+e^{\theta_0' x})^2}$, but the second derivative becomes $-\frac{1}{4}$. Heuristically, in each neighborhood of x space, we may be discarding most of the data points, but each accepted one has proportionately higher variance-reduction value.

The practical meaning of Theorem 7 is that local case-control sampling is most advantageous when $\overline{a}(\theta_0) = \mathbb{E}(|Y - \tilde{p}(X)|)$ is small; i.e. when Y is easy to predict throughout much of the covariate space. This can happen as a result of marginal or conditional imbalance, or both. Standard case-control sampling can also improve our efficiency in the presence of marginal imbalance, but unlike local case-control sampling, it does not exploit conditional imbalance. Hence we would expect local case-control to outperform standard case-control in cases where the marginal imbalance is very high, as in the simulation of Section 2.5.2.

2.4.4 Variance for a Larger Sample

In Section 2.3.3 we proposed increasing the size of the local case-control subsample by multiplying all the acceptance probabilities $a(x, y)$ by a constant $c > 1$ and assigning weight $w = ca(x, y)$ when $ca(x, y) > 1$. We analyze the asymptotic variance here as a function of c. To simplify matters suppose the model is correctly specified and λ is fixed at θ_0.

The weighted log-likelihood for the subsample and its derivatives are then

$$\ell_w(\theta) = \sum_{i=1}^{n} z_i w_i (y_i \theta' x_i - \log(1 + e^{\theta' x_i})) \tag{2.62}$$

$$\nabla_\theta \ell_w(\theta) = \sum_{i=1}^{n} z_i w_i (y_i - p_\theta(x_i)) x_i \tag{2.63}$$

$$\nabla_\theta^2 \ell_w(\theta) = \sum_{i=1}^{n} z_i w_i p_\theta(x_i)(1 - p_\theta(x_i)) x_i x_i' \tag{2.64}$$

Conditionally on x, there is a $p(x) \cdot (c(1 - p(x)) \wedge 1)$ chance $y = z = 1$ and $w = c(1 - p(x)) \vee 1$, where $p(x) = p_{\theta_0}(x)$. Similarly there is a $(1 - p(x)) \cdot (cp(x) \wedge 1)$ chance $y = 0, z = 1$, and $w = cp(x) \vee 1$.

We immediately obtain

$$\mathbb{E}(yzw \mid x) = cp(1-p), \quad \mathbb{E}(zw \mid x) = 2cp(1-p), \quad \mathbb{E}(zw^2 \mid x) \le c(c+1)p(1-p) \tag{2.65}$$

The expectation and variance of the score evaluated at 0 are

$$\mathbb{E}\nabla_\theta \ell_w(0) = n \int \mathbb{E}(zw(y-1/2) \mid x) x \, d\mathbb{P}(x) = 0 \tag{2.66}$$

$$J = \mathrm{Var}\left(\nabla_\theta \ell_w(0)\right) = n \int \mathbb{E}(z^2 w^2 (y-1/2)^2 \mid x) x x' \, d\mathbb{P}(x) \tag{2.67}$$

$$= \frac{n}{4} \int \mathbb{E}(zw^2 \mid x) x x' \, d\mathbb{P}(x) \preceq \frac{c(c+1)}{4} \Sigma_{\text{full}} \tag{2.68}$$

and the expected Hessian is

$$H = \mathbb{E}\nabla_\theta^2 \ell_w(0) = \frac{n}{4} \int \mathbb{E}(zw \mid x) x x' \, d\mathbb{P}(x) = \frac{c}{2} \Sigma_{\text{full}}^{-1} \tag{2.69}$$

We have derived

$$H^{-1} J H^{-1} \preceq \left(1 + \frac{1}{c}\right) \Sigma_{\text{full}} \tag{2.70}$$

For $c = 1$, we recover the factor of two from (2.56), but for e.g. $c = 5$ we only pay 20% increased variance relative to the full sample.

2.5 Simulations

Here we compare our method to standard weighted and unweighted case-control sampling for two-class Gaussian models like the one considered in Section 2.2.2. The standard case-control estimates use a 50-50 split between the two classes.

2.5.1 Simulation 1: Two-Class Gaussian, Different Variances

We begin with a five-dimensional two-class Gaussian simulation where the classes have different covariance matrices. If $X \mid Y = y \sim N(\mu_y, \Sigma_y)$, then

$$\log \frac{\mathbb{P}(x \mid Y=1)}{\mathbb{P}(x \mid Y=0)} = -\frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)'\Sigma_0^{-1}(x-\mu_0) + \text{const} \tag{2.71}$$

(2.71) is linear if $\Sigma_1 = \Sigma_0$, and quadratic otherwise, so if the two covariance matrices were the same the linear logistic model would be correctly specified. In this case the model is incorrectly specified, letting us compare the behavior of the different methods under model misspecification.

Take $\mathbb{P}(Y = 1) = 1\%$, $\mu_0 = 0$, and $\mu_1 = (1, 1, 1, 1, 4)'$. The covariance matrices are $\Sigma_0 = \mathrm{diag}(1, 1, 1, 1, 9)$ and $\Sigma_1 = I_5$. Hence $f(x)$ is additive, but with a nonzero quadratic term in x_5.

Simulation 1 ($\Sigma_0 \neq \Sigma_1 \;\Rightarrow$ model misspecified)

	$\widehat{\text{Bias}^2}$	(s.e.)	$\widehat{\text{Var}}$	(s.e.)
C-C	0.15	(0.0011)	0.04	(0.00063)
Weighted C-C	0.022	(0.0016)	0.15	(0.0026)
Local C-C	0.0028	(0.00016)	0.023	(0.00034)
LP Sampling	0.0037	(0.00032)	0.055	(0.00099)

Table 2.2: Estimated bias and variance of $\hat{\beta}$ for each sampling method. For $\hat{\beta} \in \mathbb{R}^p$, we define $\text{Bias}^2 = \|\mathbb{E}\hat{\beta} - \beta\|^2$ and $\text{Var} = \sum_{j=1}^p \text{Var}(\hat{\beta}_j)$.

Simulation 2 ($\Sigma_0 = \Sigma_1 \;\Rightarrow$ model correct)

	$\widehat{\text{Bias}^2}$	(s.e.)	$\widehat{\text{Var}}$	(s.e.)
C-C	0.047	(0.0038)	0.86	(0.0079)
Weighted C-C	0.52	(0.02)	1.6	(0.016)
Local C-C	0.0038	(0.00027)	0.038	(0.00043)
LP Sampling	0.043	(0.0011)	0.13	(0.0017)

Table 2.3: Estimated bias and variance of $\hat{\beta}$ for each sampling method. For $\hat{\beta} \in \mathbb{R}^p$, we define $\text{Bias}^2 = \|\mathbb{E}\hat{\beta} - \beta\|^2$ and $\text{Var} = \sum_{j=1}^p \text{Var}(\hat{\beta}_j)$.

For our simulation, we first generate a large ($n = 10^6$) sample from the population described above. Second, we obtain a pilot model using the weighted case-control method on $n_s = 1000$ data points. Next, we take a local case-control sample of size 1000 using that pilot model.

For comparison, we obtain standard case-control (CC) and weighted case-control (WCC) estimates. For the comparison estimators we do not use a sample of size 1000 again but rather use the total number of observations seen by the LCC model or the pilot model, roughly 2000, so the LCC estimate must pay for its pilot sample. We repeat this entire procedure 1000 times.

Table 2.2 shows the squared bias and variance of $\hat{\beta}$ over the 1000 realizations for each of the three methods. As expected, we face a bias-variance tradeoff in choosing between the WCC and CC methods, whereas the LCC method improves substantially on the bias of CC and the variance of WCC. Standard errors for both bias and variance are computed via bootstrapping the 1000 realizations.

More surprising is the fact that LCC enjoys smaller bias than WCC and smaller variance than CC, dominating the other two methods on both measures. The improvement in variance over the CC estimate is likely due to the conditional imbalance present in the sample, while the improvement in bias over the WCC estimate may come from the fact that the methods are only unbiased asymptotically and the LCC estimate is closer to its asymptotic limiting behavior. Also included for comparison is the proposal of Mineiro and Karampatziakis (2013).

2.5.2 Simulation 2: Two-Class Gaussian, Same Variance

Next we simulate a two-class Gaussian model with each class having the same variance, so that the true log-odds function f is linear. We also increase the dimension to 50 for this simulation.

Since the model is now correctly specified, all three methods are asymptotically unbiased. However, in this case we introduce more substantial conditional imbalance, to demonstrate the variance-reduction advantages of local case-control sampling in that setting.

For this example, $\mathbb{P}(Y = 1) = 10\%$, $\mu_1 = \binom{1_{25}}{0_{25}}$, $\mu_0 = 0_{50}$, and $\Sigma_0 = \Sigma_1 = I_{50}$. We repeat the procedure from Section 2.5.1, now with $n_s = 10^4$. Instead of generating a full sample, the full data set is implicit and we sample directly from $\mathbb{P}_S$.

In this example, the difference between the methods is more dramatic. Table 2.3 shows the squared bias and variance of the three methods. Here local case-control enjoys substantially better bias than the other two methods, improving on CC more than twenty-fold. For the correct pilot model, $\overline{a}(\theta_0)$ is roughly 0.005, so the local case-control subsample size is around $n/200$. Since the model is correctly specified, the variance is roughly twice that of logistic regression on the full sample of size n. In other words, local case-control subsampling is roughly 100 times more efficient than uniform subsampling.

Asymptotically, all three methods are unbiased but it appears that LCC again enjoys a smaller bias in finite sample.

2.6 Web Spam Data Set

Relative to standard case-control sampling, local case-control sampling is especially well-suited for data sets with significant conditional imbalance; that is, data sets in which y_i is easy to predict for most x_i.

One such application is spam filtering. To demonstrate the advantages of local case-control sampling and compare asymptotic predictions to actual performance, we test our method on the Web Spam data available on the LIBSVM website [1] and originally from Webb et al. (2006). The data set contains 350,000 web pages, of which about 60% are labeled as "web spam," i.e. web pages designed to manipulate search engines rather than display legitimate content. This data set is marginally balanced, though as we will see the conditional imbalance is considerable.

As features we use frequency of the 99 unigrams that appeared in at least 200 documents, log-transformed with an offset so as to reduce skew in the features. In this data set the downsampling ratio $\overline{a}$ is around 10%; that is, when using a good pilot we will retain about 10% of the observations.

Since we only have a single data set, we use subsampling as a method to assess the sampling distribution of our estimators. In each of 100 replications, we begin by taking a uniform subsample

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

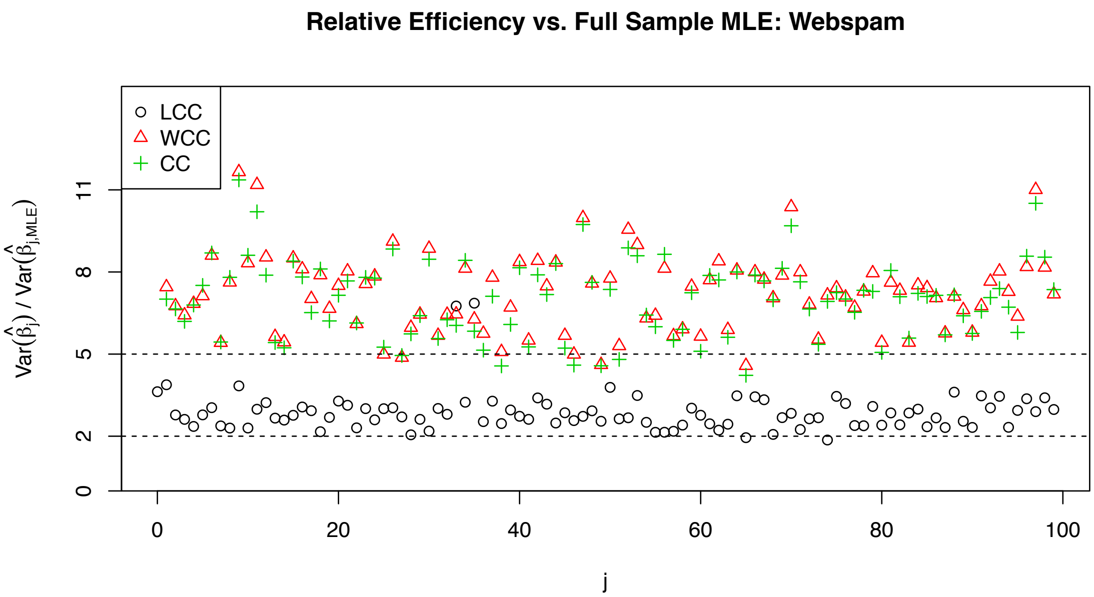


Figure 2.3: Relative variance of coefficients for different subsampling methods. The theoretical predictions ($2\times$ variance for local case-control, $5\times$ variance for standard) are reasonably close to the mark, though a bit optimistic.

of size $n = 100,000$ from the population of 350,000 documents. After obtaining 100 data sets of size $n = 100,000$, we use the same procedure as we used in our two simulations with $n_S = 10,000$.

Our asymptotic theory predicts that the variance of the local case-control sampling estimate of θ should be a little more than twice the variance using the full sample (more because the model is misspecified and our pilot has some variance). Because the full sample is close to marginally balanced, the standard case-control sampling methods should do about as well as a uniform subsample of size 20,000 — that is, they should have variance roughly 5 times that of the full sample.

Note that 20,000 is roughly twice the size of the local case-control sample, since we are counting the pilot sample against the local case-control method. If we had a readily available pilot model, as we would in many applications, it would be more relevant to give the CC and WCC methods access to only 10,000 data points, doubling their variance relative to the observed variance in this experiment.

The theoretical predictions come reasonably close in this experiment, as shown in Figure 2.3. The horizontal axis indexes each of the 100 coefficients to be fit (there are 99 covariates and an intercept), and the vertical axis gives the variance of each estimated coefficient, relative to the variance of the same coefficient in a model fitted to the full sample.

The magnitude of our improvement over standard case-control sampling is substantial here, but could be much larger in a data set with an even stronger signal. The key point is that standard case-control methods have no way to exploit conditional imbalance, so the more there is, the more local case-control dominates the other methods.

2.7 Discussion

In imbalanced logistic regression, we can speed up computation by subsampling the data in a biased fashion and making post-hoc correction to the coefficients estimated in the subsample. Standard case-control sampling is one such scheme, but it has two main flaws: it has no way to exploit conditional imbalance, and when the model is misspecified it is inconsistent for the population risk minimizer.

Local case-control sampling generalizes standard case-control sampling to address both flaws, subsampling with a bias that is allowed to depend on both x and y. When the pilot is consistent, our estimate is consistent even under misspecification, and if the model is correct then local case-control sampling has exactly twice the asymptotic variance of logistic regression on the full data set. Our simulations suggest that local case-control performs favorably in practice.

2.7.1 Extensions

This work suggests extensions in several directions, described below:

Indifference Point Other Than 50% In some applications (e.g. diagnostic medical screening), a false negative may be more costly than a false positive, or vice-versa. One of the implications of the discussion in Section 2.2.2 is that the Bernoulli log-likelihood implicitly places most emphasis on approximating the log-odds well near the 0 (50% probability) level curve, which may not be appropriate if the decision boundary relevant to our application is at 10%. In general, we would expect to obtain a better model in the large-n limit if we target the decision boundary we care most about.

In a sense, the reason that standard case-control sampling performed so badly in Example 2 of Section 2.2.2 is that it targeted a level curve of $\mathbb{P}(Y = 1 \mid X = x)$ other than 50%. Specifically, it targeted the level curve corresponding to 50% in the subsampling population for equal-sampled case-control sampling, which corresponds to the marginal $\mathbb{P}(Y = 1)$ level curve in the original population.

What happened by accident in Example 2 need not always be one, and it would be interesting to generalize our procedure so as to target any chosen decision threshold. More generally still, our indifference point could depend on our features x — in online advertising, for instance, some advertisers may be willing to pay more per click than others.

Boosting There is no reason in principle why the pilot model must be linear, or belong to the same model space as the model we fit to the local case-control sample, since we can use any $\tilde{p}(x) = \frac{e^{\tilde{f}(x)}}{1+e^{\tilde{f}(x)}}$

in the algorithm. Moreover, we could model the log-odds in the subsample however we choose, e.g. as arising from a regression tree. Whatever log-odds function $f_s(x)$ we fit to the local case-control sample can simply be added to $\tilde{f}(x)$ to obtain an estimate for $f(x)$ in the original population.

This observation suggests the possibility of iteratively fitting a "base model" to the subsample, then adding it to $\tilde{f}(x)$ to obtain a new pilot for the next iteration. Indeed, that iterative algorithm is closely related to the AdaBoost algorithm of Freund and Schapire (1997). Even more similarly to AdaBoost, we could weight each point by $|y_i - \tilde{p}(x_i)|$ instead of sampling it with that probability.

Friedman et al. (2000) show that the AdaBoost algorithm can be thought of as fitting a logistic regression model additive in base learners. In AdaBoost, the function $F_M(x) = \sum_{m=1}^{M} f_m(x)$ simply records the number of classifiers f_m classifying x as belonging to class $+1$ minus the number classifying it as class -1, and Friedman et al. show that $\frac{1}{2}F_M(x)$ can be thought of as approximating the log-odds of $Y = +1$ given $X = x$.

The difference is that while AdaBoost weights the point (x_i, y_i) by $e^{(2y_i-1)F_m(x_i)}$, the local case-control version would use weights

$$|y_i - p_M(x_i)| = \frac{e^{y_i F_m(x_i)}}{1 + e^{F_m(x_i)}} = \frac{e^{(2y_i-1)F_m(x_i)}}{1 + e^{(2y_i-1)F_m(x_i)}} \tag{2.72}$$

Operationally, this alternative weighting scheme limits the influence of "outliers," i.e. hard-to-classify points that can unduly drive the AdaBoost fit.

Logistic Regression with Regularization In high-dimensional settings, lasso- or ridge-penalized logistic regressions are often preferable to standard logistic regression, the model considered here. One could use local case-control sampling with a regularized version of logistic regression, but our asymptotic results might need revisiting in such a case — especially in a high-dimensional asymptotic regime ($p \gg n$ or $p/n \to \gamma \in (0, \infty)$). Since the high-dimensional setting is important in modern statistics and machine learning, this bears further investigation.

Other Generalized Linear Models One way of viewing the method is as a way of "tilting" the conditional distribution of Y by a linear function of X in the natural parameter space so as to enrich our subsample for more informative observations. We could use similar tricks on other GLMs.

For instance, suppose we are given a Poisson variable with natural parameter $\eta = \log \mathbb{E}Y$. By sampling with acceptance probability proportional to $e^{\xi Y}$, we obtain (conditional on acceptance) a Poisson with natural parameter $\eta + \xi$. Since Poisson variables with larger means carry more information, this could yield a substantial improvement over uniform subsampling.

If our data arise from a Poisson GLM with $\eta(x) \approx \alpha + \beta' x$, we could generalize the local case-control scheme by sampling (x_i, y_i) with probability proportional to $\exp\{(\xi_0 - \alpha - \beta' x_i) y_i\}$, where the extra parameter ξ_0 guarantees that we always tilt the conditional mean of y_i upward. Similar generalizations may apply for multinomial logit and survival models.

Chapter 3

Optimal Inference After Model Selection

3.1 Introduction

A typical statistical investigation can be thought of as consisting of two stages:

1. **Selection:** The analyst chooses a probabilistic model for the data at hand, and formulates testing, estimation, or other problems in terms of unknown aspects of that model.

2. **Inference:** The analyst attempts the chosen problems using the data and the selected model.

Informally, the selection stage determines what questions to ask, and the inference stage answers those questions. Most statistical methods carry an implicit assumption that selection is *non-adaptive* — that is, choices about which model to use, hypothesis to test, or parameter to estimate, are made before seeing the data. *Adaptive selection* (also known colloquially as "data snooping") violates this assumption, formally invalidating any subsequent inference.

In some cases, it is possible to specify the question prior to collecting the data—for instance, if the data are governed by some known physical law. However, in most applications, the choice of question is at least partially guided by the data. For example, we often perform exploratory analyses to decide which predictors or interactions to include in a regression model or to check whether the assumptions of a test are satisfied. In this work, we define a framework for valid inference after adaptive selection.

If we do not account properly for adaptive model selection, the resulting inferences can have troubling frequency properties, as we now illustrate with an example.

Example 1 (File Drawer Effect)**.** Suppose a scientist observes n independent observations $Y_i \sim N(\mu_i, 1)$. He focuses only on the apparently large effects, selecting only the indices i for which

$|Y_i| > 1$, i.e.

$$\widehat{I} = \{i : |Y_i| > 1\}.$$

He wishes to test $H_{0,i} : \mu_i = 0$ for each $i \in \widehat{I}$ at the $\alpha = 0.05$ significance level. Most scientists intuitively recognize that the nominal test that rejects $H_{0,i}$ when $|Y_i| > 1.96$ is invalidated by the selection.

What exactly is "invalid" about this test? After all, the probability of falsely rejecting a given $H_{0,i}$ is still α, since most of the time, $H_{0,i}$ is simply not tested at all. Rather, the troubling feature is that the error rate among the hypotheses *selected* for testing is possibly much higher than α. To be precise, let n_0 be the number of true null effects and suppose $n_0 \to \infty$ as $n \to \infty$. Then, in the long run, the fraction of errors among the true nulls we test is

$$\begin{aligned}
\frac{\text{\# false rejections}}{\text{\# true nulls selected}} &= \frac{\dfrac{1}{n_0} \displaystyle\sum_{i:\, H_{0,i} \text{ true}} 1\{i \in \widehat{I},\ \text{reject } H_{0,i}\}}{\dfrac{1}{n_0} \displaystyle\sum_{i:\, H_{0,i} \text{ true}} 1\{i \in \widehat{I}\}} \\
&\to \frac{\mathbb{P}_{H_{0,i}}(i \in \widehat{I},\ \text{reject } H_{0,i})}{\mathbb{P}(i \in \widehat{I})} \\
&= \mathbb{P}_{H_{0,i}}(\text{reject } H_{0,i} \mid i \in \widehat{I}), \qquad (3.1)
\end{aligned}$$

which for the nominal test is $\Phi(-1.96)/\Phi(-1) \approx .16$.

Thus, we see that (3.1), the probability of a false rejection conditional on selection, is a natural error criterion to control in the presence of selection. In this example, we can directly control (3.1) at level $\alpha = 0.05$ simply by finding the critical value c solving

$$\mathbb{P}_{H_{0,i}}\left(|Y_i| > c \;\middle|\; |Y_i| > 1\right) = 0.05.$$

In this case $c = 2.41$, which is more stringent than the nominal 1.96 cutoff.

This chapter will develop a theory for inference after selection based on controlling the *selective type I error rate* (3.1). Our guiding principle is:

The answer must be valid, given that the question was asked.

For all its disarming simplicity, Example 1 can be regarded as a stylized model of science. Imagine that each Y_i represents an estimated effect size from a scientific study. However, only the large estimates are ever published—a caricature which may not be too far from the truth, as recently demonstrated by Franco et al. (2014). To compound the problem, there may be many reasonable methodologies to choose from, even once the analyst has decided roughly what scientific question to address (Gelman and Loken, 2013). Because of the resulting selection bias, the error rate among published claims may be very high, leading even to speculation that "most published research findings are false" (Ioannidis, 2005). Thus, selection effects may be a partial explanation for the replicability crisis reported in the scientific community (Yong, 2012) and the popular media (Johnson, 2014).

The setting of Example 1 has been studied extensively in the literature of simultaneous and selective inference, and several authors have proposed adjusting for selection by means of conditional inference. Zöllner and Pritchard (2007) and Zhong and Prentice (2008) construct selection-adjusted estimators and intervals for genome-wide association studies for genes that pass a fixed initial significance threshold, based on a conditional Gaussian likelihood. Cohen and Sackrowitz (1989) obtain unbiased estimates for the mean of the population whose sample mean is largest by conditioning on the ordering of the observed sample means, and Sampson and Sill (2005) and Sill and Sampson (2009) apply the same idea to obtain estimates for the best-performing drug in an adaptive clinical trial design. Hedges (1984) and Hedges (1992) propose methods to adjust for the file drawer effect in meta-analysis when scientists only publish significant results.

Another framework for selection adjustment is proposed by Benjamini and Yekutieli (2005), who consider the problem of constructing intervals for a number R of parameters selected after viewing the data. Letting V denote the number of non-covering intervals among those constructed, they define the *false coverage-statement rate* (FCR) as the expected fraction $V/\max(R,1)$ of non-covering intervals. Controlling the FCR at level α thus amounts to "coverage on the average, among selected intervals." As we will see further in Section 3.8, FCR control is closely related to the selective error control criterion we propose. In fact, Weinstein et al. (2013) employ conditional inference to construct FCR-controlling intervals in the context of Example 1. Rosenblatt and Benjamini (2014) propose a similar method for finding correlated regions of the brain, also with a view toward FCR control.

3.1.1 Conditioning on Selection

In classical statistical inference, the notion of "inference after selection" does not exist. The analyst must specify the model, as well as the hypothesis to be tested, in advance of looking at the data. A classical level-α test for a hypothesis H_0 under model M must control the usual or *nominal type I error rate*:

$$\mathbb{P}_{M,H_0}(\text{reject } H_0) \leq \alpha. \tag{3.2}$$

The subscript in (3.2) reminds us that the probability is computed under the assumption that the data Y are generated from model M, and H_0 is true; if M is misspecified, there are no guarantees on the rejection probability.

While ruling out model selection avoids the selection problem altogether, it does not realistically describe most statistical practice: statisticians are trained to check their models and to tweak them if they diagnose a problem. Under the purist view, model checking is technically forbidden, since it leaves open the possibility that the model will change after we see the data. We will argue that if the model and hypothesis are selected randomly, we should instead control the selective type I error rate

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) \leq \alpha. \tag{3.3}$$

One can argue that models and hypotheses are practically never truly fixed but are chosen randomly, since they are based on the outcomes of previous experiments in the (random) scientific process. Typically, we ignore the random selection and use classical tests that control (3.2), implicitly assuming that the randomness in selecting M and H_0 is independent of the data used for inference. In that case,

$$\mathbb{P}_{M,H_0}(\text{reject } H_0 \mid (M, H_0) \text{ selected}) = \mathbb{P}_{M,H_0}(\text{reject } H_0). \tag{3.4}$$

It may seem pedantic to point out that model selection is random if based on previous experiments. Even so, this viewpoint gives us a prescription for what to do when science does not dictate a model. If it is possible to split the data $Y = (Y_1, Y_2)$ with Y_1 independent of Y_2, then we can imitate the scientific process by setting aside Y_1 for selection and Y_2 for inference. If selection depends on Y_1 only, then any nominal level-α test based on the value of Y_2 will satisfy (3.4), so the nominal test based on Y_2 also controls the selective error (3.3).

This meta-algorithm for generating selective procedures from nominal ones is called *data splitting* or *sample splitting*. The idea dates back at least as far as Cox (1975), and, despite the paucity of literature on the topic, is common wisdom among practitioners. For example, it is customary in genetics to use one cohort to identify loci of interest and a separate cohort to confirm them (Sladek et al., 2007). Wasserman and Roeder (2009) and Meinshausen et al. (2009) discuss data-splitting approaches to high-dimensional inference.

The popularity of data splitting owes in no small part to its transparent justification, which non-experts can easily appreciate: if we imagine that Y_1 is observed "first," then we can proceed to analyze Y_2 as though model selection took place "ahead of time." Equation (3.4) guarantees that this temporal metaphor will not lead us astray even if it does not describe how Y_1 and Y_2 were actually collected.

Data splitting elegantly solves the problem of controlling selective error, but at a cost. It not only reduces the amount of data available for inference, but also reduces the amount of data available for selection. Furthermore, it is not always possible; for example, spatial and time series data often exhibit autocorrelation that rules out splitting the data into independent parts.

In this chapter, we propose directly controlling the selective error rate (3.3) by conditioning on the event that (M, H_0) is selected. As with data splitting, we treat the data as though it were revealed in stages: in the first stage, we "observe" just enough data to resolve the decision of whether to test (M, H_0), after which we can treat the data $(Y \mid (M, H_0) \text{ selected})$ as "not yet observed" when stage two commences.

The intuition of the above paragraph can be expressed formally in terms of the filtration

$$\mathscr{F}_0 \underbrace{\subseteq}_{\text{used for selection}} \mathscr{F}(\mathbf{1}_A(Y)) \underbrace{\subseteq}_{\text{used for inference}} \mathscr{F}(Y), \tag{3.5}$$

where $\mathscr{F}(Z)$ denotes the σ-algebra generated by random variable Z (informally, everything we know about the data after observing Z), $\mathscr{F}_0$ is the trivial σ-algebra (representing complete ignorance), and A is the *selection event* $\{(M, H_0) \text{ selected}\}$. We can think of "time" as progressing from left to right in (3.5). In stage one, we learn just enough to decide whether to test (M, H_0), and no more, advancing our state of knowledge from $\mathscr{F}_0$ to $\mathscr{F}(\mathbf{1}_A(Y))$. We then begin stage two, in which we discover the actual value of Y, advancing our knowledge to $\mathscr{F}(Y)$. Because our selection decision is made at the end of stage one, everything revealed during stage two is fair game for inference.

In effect, controlling the type I error conditional on A prevents us from appealing to the fact that $Y \in A$ as evidence against H_0. Even if $Y \in A$ is extremely surprising under H_0, we still will not reject unless we are surprised anew in the second stage. In this sense, conditioning on a random variable discards the information it carries about any parameter or hypothesis of interest. By contrast with data splitting, which can be viewed as conditioning on Y_1 instead of $\mathbf{1}_A(Y_1)$, we advocate discarding as little data as possible and reserving the rest for stage two. This frugality results in a more efficient division of the information carried by Y — *data carving*, to introduce an evocative metaphor.

3.1.2 Outline

In Section 3.2 we formalize the problem of selective inference, discuss general properties of selective error control, and address key conceptual questions. Conditioning on the selection event effectively discards the information used for selection, but some information is left over for second-stage inference. We will also see that a major advantage of selective error control is that it allows us to consider only one model at a time when designing tests and intervals, even if *a priori* there are many models under consideration.

If $\mathcal{L}(Y)$, the law of random variable Y, follows an exponential family model, then for any event A, $\mathcal{L}(Y \mid A)$ follows a closely related exponential family model. As a result, selective inference dovetails naturally with the classical optimality theory of Lehmann and Scheffé (1955); Section 3.3 briefly reviews this theory and derives most powerful unbiased selective tests in arbitrary exponential family models after arbitrary model selection procedures. Because conditioning on more data than is necessary saps the power of second-stage tests, data splitting yields inadmissible selective tests under general conditions.

Section 3.5 gives some general strategies for computing rejection cutoffs for the tests prescribed in Section 3.3, while Sections 3.4–3.6 derive selective tests in specific examples. Section 3.4 focuses on the case of linear regression, generalizing the recent proposals of Taylor et al. (2014), Lee et al. (2013), and others. We derive new, more powerful selective z-tests, as well as t-tests that do not require knowledge of the error variance σ^2.

Several simulations in Section 3.7 compare the post-lasso selective z-test with data splitting, and illustrate a *selection–inference tradeoff*, between using more data in the initial stage and reserving more information for the second stage. Section 3.8 compares and contrasts selective inference with multiple inference, and Section 3.9 concludes.

3.2 The Problem of Selective Inference

3.2.1 Example: Regression and the Lasso

In the previous section, we motivated the idea of conditioning on selection. Arguably, the most familiar example of this "selection" is variable selection in linear regression. In regression, the observed data $Y \in \mathbb{R}^n$ is assumed to be generated from a multivariate normal distribution

$$Y \sim N_n(\mu,\ \sigma^2 I_n). \tag{3.6}$$

The goal is to model the mean μ as a linear function of predictors X_j, $j = 1, \ldots, p$. To obtain a more parsimonious model (or simply an identifiable model when $p > n$), researchers will often use only a subset $M \subseteq \{1, \ldots, p\}$ of the predictors. Each subset M leads to a different probabilistic model corresponding to the assumption $\mu = X_M \beta^M$, where X_M denotes the matrix consisting of columns X_j for $j \in M$. Then, it is customary to report tests of $H_{0,j}^M : \beta_j^M = 0$ for each coefficient in the model. If M was chosen in a data-dependent way, then to control selective error we must condition on having selected $(M, H_{0,j}^M)$, which in this case is the same as conditioning on having selected model M.

There are many data-driven methods for variable selection in linear regression, ranging from AIC minimization to forward stepwise selection, cf. Hastie et al. (2009). We will consider one procedure in particular, based on the lasso, mostly because selective inference in the context of the lasso (Lee et al., 2013) was a main motivation for the present work. The lasso (Tibshirani, 1996) provides an estimate of $\beta \in \mathbb{R}^p$ that solves

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ ||Y - X\beta||_2^2 + \lambda||\beta||_1, \tag{3.7}$$

where X is the "full" matrix consisting of all p predictors. The first term is the usual least-squares objective, while the second term encourages many of the coefficients to be exactly zero. Because of this property, it makes sense to define the model "selected" by the lasso to be the set of variables

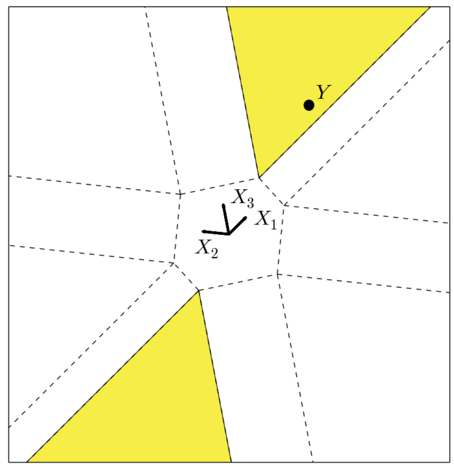


Figure 3.1: An example of the lasso with $n = 2$ observations and $p = 3$ variables. We base tests on the distribution of Y, conditional on its landing in the highlighted region.

with non-zero coefficients, i.e.,

$$\widehat{M}(Y) = \{j : \hat{\beta}_j \neq 0\}.$$

Notice that $\widehat{M}(Y)$ can take on up to 2^p possible values, one for each subset of $\{1, ..., p\}$. The regions $A_M = \{y : \widehat{M}(y) = M\}$ form a partition of $\mathbb{R}^n$ into regions that correspond to each model. To control the selective error after selecting a particular M, we must condition on the event that Y landed in A_M. The partition for a lasso problem with $p = 3$ variables in $n = 2$ dimensions is shown in Figure 3.1. An explicit characterization of the lasso partition can be found in Lee et al. (2013); see also Harris (2014) for more plots visualizing the way the lasso partitions the sample space. If we used a different selection procedure, we would obtain a different partition. Characterizations of the partitions in forward stepwise selection and marginal screening can be found in Loftus and Taylor (2014) and Lee and Taylor (2014), respectively.

Let us imagine that in stage one, we loaded the data into a software package and computed $\widehat{M}(Y)$, but we remain otherwise ignorant of the value Y — that is, we have observed *which* of the regions Y falls into but not *where* Y is in that region. Now that we have chosen the model, we will construct tests of $H_{0,j}^M : \beta_j^M = 0$ for each of the selected variables. In the example shown in Figure 3.1, we selected variables 1 and 3 and thus test the two hypotheses

$$H_{0,1}^{\{1,3\}} : \beta_1^{\{1,3\}} = 0$$
$$H_{0,3}^{\{1,3\}} : \beta_3^{\{1,3\}} = 0.$$

Notice that we have to be careful to always specify the model along with the coefficient, since the coefficient for variable j does not necessarily have a consistent interpretation across different models.

Each regression coefficient summarizes the effect of that variable, adjusting for the other variables in the model. For example, "What is the effect of IQ on salary?" is a genuinely different question from "What is the effect of IQ on salary, after adjusting for years of education?" Both questions are meaningful, but they are fundamentally different.[1]

Having chosen the model M and conditioned on the selection, we will base our tests on the precise location of Y, which we do not know yet. Conditionally, Y is not Gaussian, but it does follow an exponential family. As a result, we can appeal to the classical theory of Lehmann and Scheffé (1955) to construct tests or confidence intervals for its natural parameters, which are β^M if σ^2 is known, and otherwise are $(\beta^M/\sigma^2, 1/\sigma^2)$.

With this concrete example in mind, we will now develop a general framework of selective inference that is much more broadly applicable. Because we are explicitly allowing models and hypotheses to be random, it is necessary to carefully define our inferential goals. We first discuss selective inference in the context of hypothesis testing. The closely related developments for confidence intervals will follow in Section 3.2.3.

3.2.2 Selective Hypothesis Tests

We now introduce notation that we will use for the remainder of the chapter. Assume that our data Y lies in some measurable space $(\mathcal{Y}, \mathscr{F})$, with unknown sampling distribution $Y \sim F$. The analyst's task is to pose a reasonable probability model M — i.e., a family of distributions which she believes contains F — and then carry out inference based on the observation Y.

Let $\mathcal{Q}$ denote the *question space* of inference problems q we might tackle. A hypothesis testing problem is a pair $q = (M, H_0)$ of a model M and null hypothesis H_0, by which we mean a submodel $H_0 \subseteq M$.[2] We write $M(q)$ and $H_0(q)$ for the model and hypothesis corresponding to q. Without loss of generality, we assume $H_0(q)$ is tested against the alternative hypothesis $H_1(q) = M(q) \setminus H_0(q)$. To avoid measurability issues, we will assume throughout that $\mathcal{Q}$ is countable, although our framework can be extended to other question spaces with additional care.

In Section 3.2.1 where we test each variable in a selected regression model, the question space is

$$\mathcal{Q} = \{(M, H_{0,j}^M) : M \subseteq \{1, \dots, p\}, j \in M\}.$$

Note our slight abuse of notation in using M interchangeably to refer both to a subset of variable indices and to the corresponding probability model $\left\{N_n(X_M\beta_M, \sigma^2 I_n) : \beta \in \mathbb{R}^{|M|}\right\}$.

We model selective inference as a process with two distinct stages:

[1] We use the word "effect" here informally to refer to a regression coefficient, recognizing that regression cannot establish causal claims on its own.

[2] We identify a "null hypothesis" like $H_0 : \mu(F) = 0$ with the corresponding subfamily or "null model" $\{F \in M : \mu(F) = 0\}$. This should remind us that the error guarantees of a test do not necessarily extend beyond the model it was designed for.

1. **Selection:** From the collection $\mathcal{Q}$ of possible questions, the analyst selects a subset $\widehat{\mathcal{Q}}(Y) \subseteq \mathcal{Q}$ to test, based on the data.

2. **Inference:** The analyst performs a hypothesis test of $H_0(q)$ against $M(q) \setminus H_0(q)$ for each $q \in \widehat{\mathcal{Q}}$.

In the case of the simple regression example shown in Figure 3.1, where we selected variables 1 and 3, $\widehat{\mathcal{Q}}$ would consist of the hypotheses for each of the two variables in the model. To be completely explicit,

$$\widehat{\mathcal{Q}}(Y) = \left\{ \left(\{1,3\}, H_{0,1}^{\{1,3\}} \right), \left(\{1,3\}, H_{0,3}^{\{1,3\}} \right) \right\}.$$

A correctly specified model M is one that contains the true sampling distribution F. Importantly, we expressly do not assume that all — or any — of the candidate models are correctly specified. Because the analyst must choose M without knowing F, she could choose poorly, in which case there may be no formal guarantees on the behavior of the test she performs in stage two. Some degree of misspecification is the rule rather than the exception in most real statistical applications, whether models are specified adaptively or non-adaptively. Our analyst would be in the same position if she were to select a (probably wrong) model using Y, then use that model to perform a test on new data Y^* collected in a confirmatory experiment. See Section 3.2.5 for further discussion of this issue.

For our purposes, a *hypothesis test* is a function $\phi(y)$ taking values in $[0,1]$, representing the probability of rejecting H_0 if $Y = y$. In most cases, the value of the function will be either 0 or 1, but with discrete variables, randomization may be necessary to achieve exact level α.

To adjust for selection in testing q, we condition on the event that the question was asked, which we describe by the selection event

$$A_q = \{q \in \widehat{\mathcal{Q}}(Y)\}, \tag{3.8}$$

i.e., the event that q is among the questions asked. In general, the selection events for different questions are not disjoint. In the regression example, we only ever test $H_{0,j}^M$ when model M is selected, so conditioning on A_q is equivalent to simply conditioning on $\widehat{M}$.

In selective inference, we are mainly interested in the properties of a test ϕ_q for a question q, conditional on A_q. We say that ϕ_q controls *selective type I error* at level α if

$$\mathbb{E}_F \left[\phi_q(Y) \mid A_q \right] \leq \alpha, \quad \text{for all } F \in H_0. \tag{3.9}$$

and define its *selective power function* as

$$\text{Pow}_{\phi_q}(F \mid A_q) = \mathbb{E}_F[\phi_q(Y) \mid A_q]. \tag{3.10}$$

Because $\mathcal{Q}$ is countable, the only relevant q are those for which $\mathbb{P}(A_q) > 0$.

Notice that only the model M and hypothesis H_0 are relevant for defining the selective level of a test. This means that in designing valid ϕ_q, we can concentrate on one q at a time, even if there are many mutually incompatible candidate models in $\mathcal{Q}$. As long as each ϕ_q controls the selective error at level α given its selection event A_q, then a global error is also controlled:

$$\frac{\mathbb{E}\,[\#\text{ false rejections}]}{\mathbb{E}\,[\#\text{ true nulls selected}]} \leq \alpha, \tag{3.11}$$

provided that the denominator is finite. Equation (3.11) holds for countable $\mathcal{Q}$ regardless of the dependence structure across different q. The fact that we can design tests one q at a time makes it much easier to devise selective tests in concrete examples, which we take up in Sections 3.3–3.6.

Suppose that each scientist in a discipline controls the selective error rate for each of his or her own experiments. Then the discipline as a whole will achieve long-run control of the type I error rate among true *selected* null hypotheses, in the same sense as they would if there were no selection. No coordination is required between different research groups.

Proposition 8 (Discipline-Wide Error Control)**.** *Suppose there are n independently operating research groups in a scientific discipline with a shared, countable question space $\mathcal{Q}$. Research group i collects data $Y_i \sim F_i$, applies selection rule $\widehat{\mathcal{Q}}_i(Y_i) \subseteq \mathcal{Q}$, and carries out selective level-α tests $(\phi_{q,i}(y_i), q \in \widehat{\mathcal{Q}}_i)$. Assume each research group has probability at least $\delta > 0$ of carrying out at least one test of a true null, and for some common $B < \infty$,*

$$\mathbb{E}_{F_i}\left[|\widehat{\mathcal{Q}}_i(Y_i)|^2\right] \leq B, \quad \text{for all } i.$$

Then as n grows, the discipline as a whole achieves long-run control over the frequentist error rate

$$\limsup_{n\to\infty} \frac{\#\text{ false rejections}}{\#\text{ true nulls selected}} \overset{a.s.}{\leq} \alpha. \tag{3.12}$$

The proof is deferred to Appendix B. Though the assumption of independence across research groups can be weakened without necessarily affecting the conclusion, we do not pursue such generalizations. Note that there is no counterpart to Proposition 8 for multiple-inference error rates such as the false discovery rate (FDR) (Benjamini and Hochberg, 1995) or familywise error rate (FWER); even if every research group controls its own FWER or FDR at level α, there is no guarantee we will control FWER or FDR after aggregating across the different groups.

3.2.3 Selective Confidence Intervals

If the goal is instead to form confidence intervals for a parameter $\theta(F)$, it is more convenient to think of $\mathcal{Q}$ as containing pairs $q = (M, \theta(\cdot))$ of a model and a parameter. By analogy to (3.9), we will call a set $C(Y)$ a $(1-\alpha)$ *selective confidence set* if

$$\mathbb{P}_F(\theta(F) \in C(Y) \mid A_q) \geq 1-\alpha, \quad \text{for all } F \in M. \tag{3.13}$$

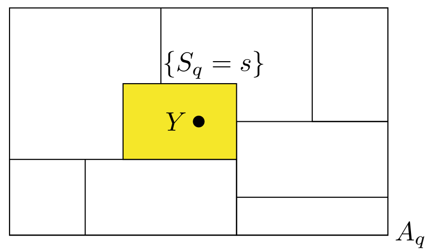


Figure 3.2: Instead of conditioning on the selection event A_q that question q is asked, we can condition on a finer event, the value of the random variable S_q. We call S_q the *selection variable*.

The next result establishes that selective confidence sets can be obtained by inverting selective tests, as one would expect by analogy to the classical case.

Proposition 9 (Duality of Selective Tests and Confidence Sets)**.** *Suppose we form a confidence interval for $\theta(F)$ on the event A_q. Suppose also that on this event, we form a test ϕ_t of $H_{0,t} = \{F : \theta(F) = t\}$ for all t. Let $C(Y)$ be the set of t for which ϕ_t does not (always) reject:*

$$C(Y) = \{t : \phi_t(Y) < 1\}. \tag{3.14}$$

If each ϕ_t is a selective level-α test, then $C(Y)$ is a selective $(1-\alpha)$ confidence set.

Proof. The selective non-coverage probability is

$$\mathbb{P}_F(\theta(F) \notin C(Y) \mid A_q) \;=\; \mathbb{P}_F(\phi_{\theta(F)}(Y) = 1 \mid A_q) \;\leq\; \mathbb{E}_F\left[\phi_{\theta(F)}(Y) \mid A_q\right] \;\leq\; \alpha.$$

□

3.2.4 Conditioning Discards Information

Because performing inference conditional on a random variable effectively disqualifies that variable as evidence against a hypothesis, we will typically want to condition on as little data as possible in stage two. Even so, some selective inference procedures condition on more than A_q. For example, data splitting can be viewed as inference conditional on Y_1, the part of the data used for selection. More generally, we say a *selection variable* is any variable $S_q(Y)$ whose level sets partition the sample space more finely than A_q; i.e., $A_q \in \mathscr{F}(S_q)$. Informally, we can think of conditioning on a finer partition of A_q, as shown in Figure 3.2.

We say ϕ controls the *selective type I error with respect to S_q* at level α if the error rate is less than α given $S_q = s$ for $\{S_q = s\} \subseteq A_q$. More formally,

$$\mathbb{E}_F\left[\phi(Y)\mathbf{1}_{A_q}(Y) \mid S_q\right] \overset{\text{a.s.}}{\leq} \alpha, \quad \text{for all } F \in H_0 \tag{3.15}$$

Taking $S_q(y) = \mathbf{1}_{A_q}(y)$, the coarsest possible selection variable, recovers the baseline selective type I error in (3.9). The definition of a selective confidence set may be generalized in the same way.

Generalizing (3.5) to finer selection variables gives

$$\mathscr{F}_0 \underbrace{\quad\subseteq\quad}_{\text{used for selection}} \mathscr{F}(S(Y)) \underbrace{\quad\subseteq\quad}_{\text{used for inference}} \mathscr{F}(Y), \tag{3.16}$$

suggesting that the more we refine $S(Y)$, the less data we have left for second-stage inference. Indeed, the finer S is, the more stringent is the requirement (3.15):

Proposition 10 (Monotonicity of Selective Error)**.** *Suppose $\mathscr{F}(S_1) \subseteq \mathscr{F}(S_2)$. If ϕ controls the type I error rate at level α for $q = (M, H_0)$ w.r.t. the finer selection variable S_2, then it also controls the type I error rate at level α w.r.t. the coarser S_1.*

Proof. If $F \in H_0$, then

$$\mathbb{E}_F\left[\phi(Y)\mathbf{1}_A(Y) \,|\, S_1\right] \;=\; \mathbb{E}_F\left[\,\mathbb{E}_F\big[\phi(Y)\mathbf{1}_A(Y)\,|\,S_2\big]\,|\,S_1\right] \;\overset{\text{a.s.}}{\leq}\; \alpha.$$

□

Because $S(y) = \mathbf{1}_A(y)$ is the coarsest possible choice, a test controlling the type I error w.r.t. any other selection variable also controls the selective error in (3.9). At the other extreme, if $S(y) = y$, then we cannot improve on the trivial "coin-flip" test $\phi(y) \equiv \alpha$. Proposition 10 suggests that we will typically sacrifice power as we move from coarser to finer selection variables. Even so, refining the selection variable can be useful for computational reasons. For example, in the case of the lasso, by conditioning additionally on the signs of the nonzero $\hat{\beta}_j$, the selection event becomes a convex region instead of the union of up to $2^{|\widehat{M}|}$ disjoint convex regions (Lee et al., 2013). Another valid reason to refine S_q beyond $\mathbf{1}_{A_q}$ is to strengthen our inferential guarantees in a meaningful way; for example, we can achieve achieve false coverage-statement rate (FCR) control by choosing $S_q = (\mathbf{1}_{A_q}(Y), |\widehat{\mathcal{Q}}(Y)|)$ (see Section 3.8, Proposition 18).

Data splitting corresponds to setting every selection variable equal to $S = Y_1$. As a result, data splitting does not use all the information that remains after conditioning on A, as we see informally in the three-stage filtration

$$\mathscr{F}_0 \underbrace{\quad\subseteq\quad}_{\text{used for selection}} \mathscr{F}(\mathbf{1}_A(Y_1)) \underbrace{\quad\subseteq\quad}_{\text{wasted}} \mathscr{F}(Y_1) \underbrace{\quad\subseteq\quad}_{\text{used for inference}} \mathscr{F}(Y_1, Y_2). \tag{3.17}$$

As we will see in Section 3.3.2, this waste of information means that data splitting is inadmissible under fairly general conditions.

We can quantify the amount of leftover information in terms of the Fisher information that remains in the conditional law of Y given S. In a smooth parametric model, we can decompose the

Hessian of the log-likelihood as

$$\nabla^2 \ell(\theta;\, Y) = \nabla^2 \ell(\theta;\, S) + \nabla^2 \ell(\theta;\, Y \mid S) \tag{3.18}$$

The conditional expectation

$$\mathcal{I}_{Y \mid S}(\theta; S) = -\mathbb{E}\left[\nabla^2 \ell(\theta;\, Y \mid S) \mid S\right] \tag{3.19}$$

is the *leftover Fisher information* after selection at $S(Y)$. Taking expectations in (3.18), we obtain

$$\mathbb{E}\left[\mathcal{I}_{Y \mid S}(\theta; S)\right] \;=\; \mathcal{I}_Y(\theta) - \mathcal{I}_S(\theta) \;\preceq\; \mathcal{I}_Y(\theta). \tag{3.20}$$

Thus, on average, the price of conditioning on S — the price of selection — is the information S carries about θ.[3] In some cases this loss may be quite small, which a simple example elucidates.

Example 2. Consider selective inference under the univariate Gaussian model

$$Y \sim N(\mu, 1), \tag{3.21}$$

after conditioning on the selection event $A = \{Y > 3\}$.

Figure 3.3a plots the leftover information as a function of μ. If $\mu \ll 3$, there is very little information in the conditional distribution: whether $\mu = -10$ or $\mu = -11$, Y is conditionally highly concentrated on 3. By contrast, if $\mu \gg 3$, then $\mathbb{P}_\mu(A) \approx 1$, the conditional law is practically no different from the marginal law, and virtually no information is lost in the conditioning.

Figure 3.3b shows the confidence intervals that result from inverting the tests described in Section 3.3. When $Y \gg 3$, the interval essentially coincides with the nominal interval $Y \pm 1.96$ because there is hardly any selection bias and no real adjustment is necessary. By contrast, when Y is close to 3 it is potentially subject to severe selection bias. This fact is reflected by the confidence interval, which is both longer than the nominal interval and centered at a value significantly less than Y.

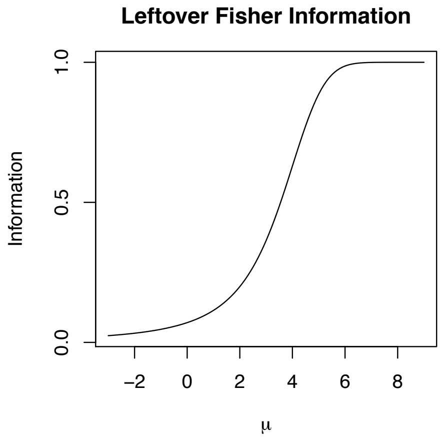
3.2.5 Conceptual Questions

We now pause to address conceptual objections that may have occurred to a skeptical reader.

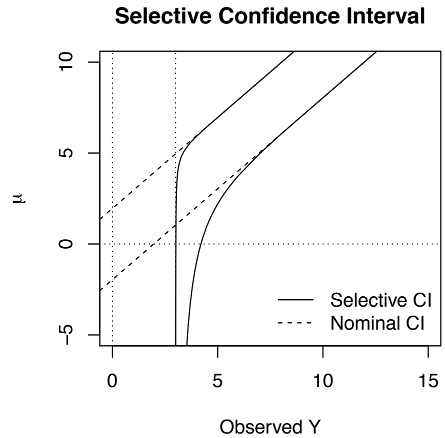
How Can the Model Be Random?

In our framework, inference is based on a probabilistic model that is allowed to be chosen randomly, based on the data Y; the reader may wonder whether that randomness muddies the interpretation of whatever inference is carried out in the second stage.

[3] Note that we do not necessarily have $\mathcal{I}_{Y \mid S}(\theta; S) \preceq \mathcal{I}_Y(\theta)$ for every S. In fact there are interesting counterexamples where $\mathcal{I}_{Y \mid A}(\theta) \gg \mathcal{I}_Y(\theta)$ for certain θ, but we will not take them up here.



(a) Leftover Fisher information as a function of μ. For $\mu \ll 3$, then there is very little information in the conditional distribution, since Y is conditionally highly concentrated on 3. For $\mu \gg 3$, then $\mathbb{P}_\mu(A) \approx 1$ and virtually no information is lost.



(b) Confidence intervals from inverting the UMPU tests of Section 3.3. For $Y \gg 3$, the interval essentially coincides with the nominal interval $Y \pm 1.96$. For Y close to 3, the wide interval reflects potentially severe selection bias.

Figure 3.3: Univariate Gaussian. $Y \sim N(\mu, 1)$ with selection event $A = \{Y > 3\}$.

First, note that in our framework the *true* sampling distribution F is not "selected" in any sense; it is entirely outside the analyst's control. The only thing selected is the *working model*, a tool the analyst uses to carry out inference, which may or may not include the true F.

Thoughtful skeptics may find reasons for concern about any approach — data carving, data splitting, or selecting a model based on a previous experiment — in which a random model is selected, believing that statistical testing is only appropriate when a probabilistic model can be based purely on convincing theoretical considerations. We answer only that this point of view would rule out most scientific inquiries for which statistics is ever used. However, if one is comfortable with choosing a random model based on data splitting or a previous experiment, we see no special reason to be any more concerned about choosing a random model based on data carving.

In any case, it is by no means required that the model M be random. For example, in our clinical trial example of Section 3.6.1, the probabilistic model is always the same but we choose which null hypotheses to test after inspecting the data. The same is true of the saturated-model selective z-test described in 3.4.2.

What if the Selected Model is Wrong?

If we were not writing about model selection, we might have begun by stating a formal mathematical assumption that the sampling distribution F belongs to a known model M, and then devised a test ϕ that behaves well when $F \in M$. The same ϕ might not work well at all for $F \notin M$: for example, if we choose to apply the one-sample t-test of $\mu = 0$ to a sample $Y_1, \ldots, Y_n$ whose observations are highly correlated, then the probability of rejection may be a great deal larger than the nominal α, even if $\mathbb{E}[Y_i] = 0$. This is not a mistake in the formal theory, nor does it make the t-test an inherently invalid test; rather, the validity or invalidity of a test is defined with respect to its behavior when $F \in H_0 \subseteq M$.

In any given application, the analyst must choose from among many statistical methods knowing that each one is designed to work under a particular set of assumptions about F — i.e., under a particular model M. Because our theory encompasses both the choice and the subsequent analysis, it would not be sensible to assume that the analyst is infallible and always selects a correct model. Typically some candidate models M are correctly specified (i.e., $F \in M$), others are not ($F \notin M$), and the analyst can never know for sure which are which.

Of course, the possibility of misspecification is not restricted to adaptive procedures like data carving and data splitting: selecting an inappropriate model *after* seeing the data leaves us no better or worse off than if we had chosen the same inappropriate model *before* seeing the data. Each ϕ_q or C_q is designed with respect to a particular model $M(q)$, and properties like selective type I error control or selective coverage only constrain its behavior for $F \in M(q)$.

There is a separate question of robustness: if $F \notin M$ but is "close" in some sense, we may still want our procedure to behave predictably. However, even if some model gives a reasonable approximation to $\mathcal{L}(Y)$, there is no guarantee that the induced model for $\mathcal{L}(Y \mid A)$ is reasonable, since conditioning can introduce new robustness problems. For example, suppose that a test statistic $Z_n(Y)$ tends in distribution to $N(0,1)$ under H_0 as $n \to \infty$. In a non-selective setting, we might be comfortable modeling it as Gaussian as a basis for hypothesis testing. In this case it is also true that $\mathcal{L}(Z_n \mid Z_n > c)$ converges to a truncated Gaussian law for any fixed $c \in \mathbb{R}$, but the approximation may be much poorer for intermediate values of n. Worse, if we use increasing thresholds $c_n \to \infty$ with n, the truncated Gaussian approximation may never become reasonable.

3.2.6 Prior Work on Selective Inference

This chapter takes its main inspiration from a recent ferment of work on the problem of inference in linear regression models after model selection. Lockhart et al. (2014) derive an asymptotic test for whether the nonzero fitted coefficients at a given knot in the lasso path contain all of the true nonzero coefficients. Taylor et al. (2014) provided an exact (finite-sample) version of this result and extended it to the LARS path, while Lee et al. (2013), Loftus and Taylor (2014), and Lee and Taylor (2014) used similar approaches to derive exact tests for the lasso with a fixed value of regularization parameter λ, forward stepwise regression, and regression after marginal screening, respectively. All of the above approaches are derived assuming that the error variance σ^2 is known or an independent estimate is available.

The present work attempts to unify the above approaches under a common theoretical framework generalizing the classical optimality theory of Lehmann and Scheffé (1955), and elucidate previously unexplored questions of power. It also lets us generalize the results to the case of unknown σ^2, and to arbitrary exponential families after arbitrary selection events.

Other works have viewed selective inference as a multiple inference problem. Recent work in this vein can be found in Berk et al. (2013) and Barber and Candes (2014). Section 3.8 argues that inference after model selection and multiple inference are distinct problems with different scientific goals; see Benjamini (2010) for more discussion of this distinction. An empirical Bayes approach for selection-adjusted estimation can be found in Efron (2011).

There has also recently been work on inference in high-dimensional linear regression models, notably Belloni et al. (2011), Belloni et al. (2014), Zhang and Zhang (2014), Javanmard and Montanari (2013), and van de Geer et al. (2013); see Dezeure et al. (2014) for a review. These works focus on approximate asymptotic inference for a fixed model with many variables, while we consider finite-sample inference after selecting a smaller submodel to focus our inferential goals.

Leeb and Pötscher (2005, 2006, 2008) prove certain impossibility results regarding estimating the distribution of post-selection estimators. These results do not apply to our framework; under the statistical models we use, the post-selection distributions of our test statistics are known and thus do not require estimation.

The foregoing works are frequentist, as is this work. Because Bayesian inference conditions on the entire data set, conditioning first on a selection event typically has no operative effect on the posterior: if p and π are respectively the marginal likelihood and prior, then $p(Y \mid A, \theta) \cdot \pi(\theta \mid A) \propto p(Y \mid \theta) \cdot \pi(\theta)$ for $Y \in A$ (Dawid, 1994). Yekutieli (2012) argues that in certain cases it is more appropriate to use the likelihood conditional on selection condition the likelihood on selection without changing the prior to reflect that conditioning, resulting in a posterior proportional to $p(Y \mid A, \theta) \cdot \pi(\theta)$. The credible intervals discussed in Yekutieli (2012) resemble the confidence intervals proposed in this chapter, and the discussion therein presents a somewhat different perspective on how and why conditioning can adjust for selection.

Though our goals are very different, our theoretical framework is in some respects similar to the conditional confidence framework of Kiefer (1976), in which inference is made conditional on some estimate of the confidence with which a decision can be made. See also Kiefer (1977); Brownie and Kiefer (1977); Brown (1978); Berger et al. (1994).

Olshen (1973) discussed error control given selection in a two-stage multiple comparison procedure, in which an F-test is first performed, then Scheffé's S-method applied if the F-test rejects. For large enough rejection thresholds, simultaneous coverage in the second stage is less than $1 - \alpha$ conditional on rejection in stage one.

3.3 Selective Inference in Exponential Families

As discussed in Section 3.2.2, we can construct selective tests "one at a time" for each model–hypothesis pair (M, H_0), conditional on the corresponding selection event A_q and ignoring any other models that were previously under consideration. This is because the other candidate models and hypotheses are irrelevant to satisfying (3.9). For that reason, we suppress the explicit dependence on $q = (M, H_0)$ except where it is necessary to resolve ambiguity.

Our framework for selective inference is especially convenient when M corresponds to a multiparameter exponential family

$$Y \sim f_\theta(y) = \exp\{\theta' T(y) - \psi(\theta)\}\, f_0(y) \tag{3.22}$$

with respect to some dominating measure. Then, the conditional distribution given $Y \in A$ for any measurable A is another exponential family with the same natural parameters and sufficient statistics but different carrier distribution and normalizing constant:

$$(Y \mid Y \in A) \sim \exp\{\theta' T(y) - \psi_A(\theta)\}\, f_0(y)\, \mathbf{1}_A(y) \tag{3.23}$$

This fact lets us draw upon the rich theory of inference in multiparameter exponential families.

3.3.1 Conditional Inference and Nuisance Parameters

Classically, conditional inference in exponential families arises as a means for inference in the presence of nuisance parameters, as in Model 11 below.

Model 11 (Exponential Family with Nuisance Parameters)**.** *Y follows a p-parameter exponential family with sufficient statistics $T(y)$ and $U(y)$, of dimension k and $p-k$ respectively:*

$$Y \sim f_{\theta,\zeta}(y) = \exp\{\theta' T(y) + \zeta' U(y) - \psi(\theta, \zeta)\}\, f_0(y), \tag{3.24}$$

with $(\theta, \zeta) \in \Theta \subseteq \mathbb{R}^p$ open.

Assume θ corresponds to a parameter of interest and ζ to an unknown nuisance parameter. The conditional law $\mathcal{L}(T(Y) \mid U(Y))$ depends only on θ:

$$(T \mid U = u) \sim g_\theta(t \mid u) = \exp\{\theta' t - \psi_g(\theta \mid u)\}\, g_0(t \mid u), \tag{3.25}$$

letting us eliminate ζ from the problem by conditioning on U. For $k = 1$ (i.e., for $\theta \in \mathbb{R}$), we obtain a single-parameter family for T.

Consider testing the null hypothesis $H_0 :\ \theta \in \Theta_0 \subseteq \Theta$ against the alternative $H_1 :\ \theta \in \Theta_1 =$

$\Theta \setminus \Theta_0$. We say a level-$\alpha$ selective test $\phi(y)$ is *selectively unbiased* if

$$\text{Pow}_\phi(\theta \mid A) = \mathbb{E}_\theta[\phi(Y) \mid A] \geq \alpha, \quad \text{for all } \theta \in \Theta_1. \tag{3.26}$$

The condition (3.26) specializes to the usual definition of an unbiased test when there is no selection (when $A = \mathcal{Y}$). Unbiasedness rules out tests that privilege some alternatives to the detriment of others, such as one-sided tests when the alternative is two-sided.

A *uniformly most powerful unbiased* (UMPU) selective level-α test is one whose selective power is uniformly highest among all level-α tests satisfying (3.26). A selectively unbiased confidence region is one that inverts a selectively unbiased test, and confidence regions inverting UMPU selective tests are called uniformly most accurate unbiased (UMAU). All of the above specialize to the usual definitions when $A = \mathcal{Y}$.

See Lehmann and Romano (2005) or Brown (1986) for thorough reviews of the rich literature on testing in exponential family models. In particular, the following classic result of Lehmann and Scheffé (1955) gives a simple construction of UMPU tests in exponential family models.

Theorem 12 (Lehmann and Scheffé (1955))**.** *Under Model 11 with $k = 1$, consider testing the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{against} \quad H_1 : \theta \neq \theta_0 \tag{3.27}$$

at level α. There is a UMPU test of the form $\phi(Y) = f(T(Y), U(Y))$ with

$$f(t, u) = \begin{cases} 1 & t < c_1(u) \text{ or } t > c_2(u) \\ \gamma_i & t = c_i(u) \\ 0 & c_1(u) < t < c_2(u) \end{cases} \tag{3.28}$$

where c_i and γ_i are chosen to satisfy

$$\mathbb{E}_{\theta_0}\left[f(T, U) \mid U = u\right] = \alpha \tag{3.29}$$

$$\mathbb{E}_{\theta_0}\left[T f(T, U) \mid U = u\right] = \alpha \, \mathbb{E}_{\theta_0}\left[T \mid U = u\right]. \tag{3.30}$$

The condition (3.29) constrains the power to be α at $\theta = \theta_0$, and (3.30) is obtained by differentiating the power function and setting its derivative to 0 at $\theta = \theta_0$.

Because $\mathcal{L}(Y \mid A)$ is an exponential family, we can simply apply Theorem 12 to the conditional law $\mathcal{L}(Y \mid A)$ to obtain an analogous construction in the selective setting.

Corollary 13 (UMPU Selective Tests)**.** *Under Model 11 with $k = 1$, consider testing the hypothesis*

$$H_0 : \theta = \theta_0 \quad \textit{against} \quad H_1 : \theta \neq \theta_0 \tag{3.31}$$

at selective level α on selection event A. There is a UMPU selective test of the form $\phi(Y) = f(T(Y), U(Y))$ with

$$f(t,u) = \begin{cases} 1 & t < c_1(u) \text{ or } t > c_2(u) \\ \gamma_i & t = c_i(u) \\ 0 & c_1(u) < t < c_2(u) \end{cases} \tag{3.32}$$

for which c_i and γ_i solve

$$\mathbb{E}_{\theta_0}\left[f(T,U) \,|\, U = u,\ Y \in A\right] = \alpha \tag{3.33}$$

$$\mathbb{E}_{\theta_0}\left[T f(T,U) \,|\, U = u,\ Y \in A\right] = \alpha\, \mathbb{E}_{\theta_0}\left[T \,|\, U = u,\ Y \in A\right]. \tag{3.34}$$

It is worth keeping in mind that unbiasedness is only one way to choose a test when there is no completely UMP one. For example, another simple choice is to use the equal-tailed test from the same conditional law (3.25). The equal-tailed level-α rejection region is simply the union of the one-sided level-$\alpha/2$ rejection regions. While the equal-tailed and UMPU tests choose c_i and γ_i in different ways, both tests take the form (3.28). In fact, as we will see next, *all* admissible tests are of this form, which implies that data splitting tests are usually inadmissible.

3.3.2 Conditioning, Admissibility, and Data Splitting

A selective level-α test ϕ is *inadmissible* on selection event A if there exists another selective level-α test ϕ^* for which

$$\mathbb{E}_{\theta,\zeta}[\phi^*(Y)|A] \geq \mathbb{E}_{\theta,\zeta}[\phi(Y)|A], \quad \text{for all } (\theta,\zeta) \in \Theta_1, \tag{3.35}$$

with the inequality strict for at least one (θ,ζ). In the main result of this section, we will show that tests based on data splitting are nearly always inadmissible.

Let Y be an observation from Model 11, and suppose we wish to test

$$H_0:\ \theta = \theta_0 \quad \text{against} \quad H_1:\ \theta \neq \theta_0. \tag{3.36}$$

We will assume all tests are functions of the sufficient statistic and write (with some abuse of notation) $\phi(T,U)$ for $\phi(Y)$. We can do this without loss of generality because any test $\phi(Y)$ can be Rao-Blackwellized, i.e.,

$$\phi(T,U) \equiv \mathbb{E}[\phi(Y)|T,U],$$

to obtain a new test that is a function of (T,U), with the same power function as the original. Therefore, if $\phi(T,U)$ is inadmissible, then so is the original test $\phi(Y)$.

Now we can apply the following result of Matthes and Truax (1967).

Theorem 14 (Matthes and Truax, Theorem 3.1)**.** *Let Y be an observation from Model 11, and* *suppose we wish to test*

$$H_0:\ \theta = \theta_0 \quad \textit{against} \quad H_1:\ \theta \neq \theta_0. \tag{3.37}$$

Let $\mathscr{C}$ denote the class of all level-α tests $\phi(T,U)$ of the form

$$\phi(t,u) = \begin{cases} 0 & t \in \mathit{int}\ C(u) \\ \gamma(t,u) & t \in \partial C(u) \\ 1 & t \notin C(u) \end{cases}, \tag{3.38}$$

and $C(u)$ is a convex set for every u. Then, for any $\phi \notin \mathscr{C}$, there exists $\phi^* \in \mathscr{C}$ such that

$$\mathbb{E}_{\theta,\zeta}[\phi^*(T,U)] \geq \mathbb{E}_{\theta,\zeta}[\phi(T,U)], \quad \textit{for all } (\theta,\zeta) \in \Theta_1. \tag{3.39}$$

Notice that, if (3.39) holds with equality for all (θ,ζ), then by the completeness of (T,U) we have $\phi \overset{\text{a.s.}}{=} \phi^*$. Hence, every admissible test is in $\mathscr{C}$ or almost surely equal to a test in $\mathscr{C}$.

In order to apply this result to data splitting, we first introduce a generic exponential family composed of two independent data sets governed by the same parameters:

Model 15 (Exponential Family with Data Splitting)**.** *Model independent random variables $(Y_1, Y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ as*

$$Y_i \sim \exp\left\{\theta T_i(y) + \zeta' U_i(y) - \psi_i(\theta,\zeta)\right\}\ f_{0,i}(y), \quad i = 1, 2, \tag{3.40}$$

with $\theta \in \mathbb{R}$ and with the models for Y_i both satisfying Model 11.

Model 15 would, for example, cover the case where Y_1 and Y_2 are the responses for two linear regressions with different design matrices but the same regression coefficients.

For a selection event $A = A_1 \times \mathcal{Y}_2$, we say ϕ is a *data-splitting test* if $\phi(Y) = \phi_2(Y_2)$; that is, the selection stage uses only Y_1 and the inference stage uses only Y_2. Again, by Rao-Blackwellization, we can assume w.l.o.g. that the test is of the form $\phi(T_2, U_2)$.

Next, define the *cutoff gap* $g^*(\phi)$ as the largest $g \geq 0$ for which the acceptance and rejection regions are separated by a "cushion" of width g. If T_2^* is a conditionally independent copy of T_2 given U_2, then

$$g^*(\phi) = \sup\left\{g:\ \mathbb{P}_{\theta,\zeta}(|T_2 - T_2^*| < g,\ \phi(T_2,U_2) > 0,\ \phi(T_2^*,U_2) < 1) = 0\right\}. \tag{3.41}$$

Note that the support of (T_2, T_2^*, U_2) does not depend on θ or ζ; thus, neither does g^*. For most tests, $g^*(\phi) = 0$. For example, $g^* = 0$ if either cutoff is in the interior of $\operatorname{supp}(T_2 \mid U_2)$ with positive probability, or if ϕ is a randomized test for discrete (T_2, U_2).

Next we prove the main technical result of this section: ϕ is inadmissible unless T_1 is determined by U_1 on A_1, within an amount g^* of variability.

Theorem 16. *Let T_1^* denote a copy of T_1 that is conditionally independent given U_1 and $Y_1 \in A$, and let ϕ be a data-splitting test of (3.37) in Model 15. If*

$$\mathbb{P}_{\theta,\zeta}(|T_1 - T_1^*| > g^*(\phi) \,|\, Y_1 \in A) > 0$$

then ϕ is inadmissible.

Proof. Construct conditionally independent copies T_i^* with $(T_1, T_1^*, U_1) \perp\!\!\!\perp (T_2, T_2^*, U_2)$, and assume that ϕ is of the form $\phi(T, U)$ with $T = T_1 + T_2$ and $U = U_1 + U_2$ (otherwise we could Rao-Blackwellize it). If ϕ is admissible, then by Matthes and Truax (1967), it must be a.s. equivalent to a test of the form (3.38). That is, there exist $c_i(U)$ for which

$$\mathbb{P}_{\theta,\zeta}(\phi(T,U) < 1,\ T \notin [c_1(U), c_2(U)] \,|\, A) = \mathbb{P}_{\theta,\zeta}(\phi(T,U) > 0,\ c_1(U) < T < c_2(U) \,|\, A) = 0. \quad (3.42)$$

Now, by assumption, there exists $\delta > g^*(\phi)$ for which

$$B_1 \triangleq \{|T_1 - T_1^*| > \delta\}$$

occurs with positive probability. By the definition of $g^*(\phi)$ in (3.41), the event

$$B_2 \triangleq \{|T_2 - T_2^*| > \delta,\ \phi(T_2, U_2) > 0,\ \phi(T_2^*, U_2) < 1\}$$

also occurs with positive probability. Since the two events are independent, $B = B_1 \cap B_2$ occurs with positive probability.

Next, assume w.l.o.g. that the event in (3.41) can occur with $T_2^* > T_2$ (otherwise we could reparameterize with natural parameter $\xi = -\theta$, for which $-T_i$ would be the sufficient statistics for Y_i). Then for some $\delta > g^*(\phi)$, the event

$$B = \{T_1 + \delta < T_1^*,\ \ T_2 < T_2^* < T_2 + \delta,\ \ \phi(T_2) > 0,\ \ \text{and } \phi(T_2^*) < 1\}$$

occurs with positive probability for all θ, ζ. On B,

$$T_1 + T_2 < T_1 + T_2^* < T_1^* + T_2 < T_1 + T_2^*,$$

but $\phi(T, U) > 0$ for $T = T_1 + T_2$ and $T = T_1^* + T_2$ and $\phi(T, U) < 1$ for the other two, ruling out the possibility of (3.42). ☐

In the typical case $g^* = 0$ and we have

Corollary 17. *Suppose ϕ is a data-splitting test of (3.37) in Model 15 with $g^*(\phi) = 0$. Then ϕ is inadmissible unless T_1 is a function of U_1 (that is, unless $T_1 \in m\mathscr{F}(U_1)$).*

Example 3. To illustrate Theorem 16, consider a bivariate version of Example 2:

$$Y_i \sim N(\mu, 1), \quad i = 1, 2, \quad \text{with } Y_1 \perp\!\!\!\perp Y_2, \tag{3.43}$$

in which we condition on the selection event $A = \{Y_1 > 3\}$.

With data splitting, we could construct a 95% confidence interval using only Y_2; namely, $Y_2 \pm 1.96$. This interval is valid but does not use all the information available. A more powerful alternative is to construct an interval based on the law

$$\mathcal{L}_\mu \left(Y_1 + Y_2 \;\middle|\; Y_1 > 3 \right), \tag{3.44}$$

which uses the leftover information in Y_1.

Figure 3.4a shows the Fisher information that is available to each test as a function of μ. The Fisher information of data splitting is exactly 1 no matter what μ is, whereas the optimal selective test has information approaching 2 as μ increases. Figure 3.4b shows the expected confidence interval length of the equal tailed interval as a function of μ. For $\mu \gg 3$, the data splitting interval is roughly 41% longer than it needs to be (in the limit, the factor is $\sqrt{2} - 1$).

Together, the plots tell a consistent story: when the selection event is not too unlikely, discarding the first data set exacts an unnecessary toll on the power of our second-stage procedure.

3.4 Selective Inference for Linear Regression

For a concrete example of the exponential family framework discussed in Section 3.3, we now turn to linear regression, which is one of the most important applications of selective inference. In linear regression, the data arise from a multivariate normal distribution

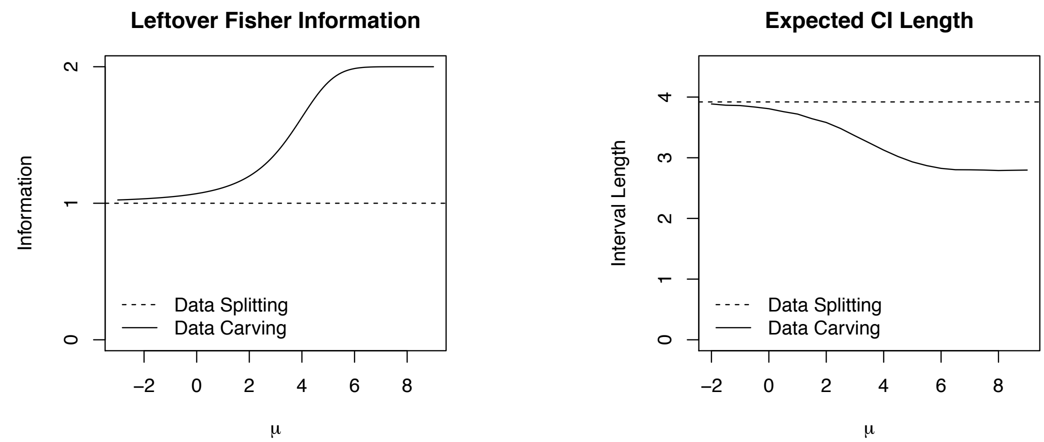
$$Y \sim N_n(\mu, \, \sigma^2 I_n),$$

where μ is modeled as

$$\mu = X_M \beta^M. \tag{3.45}$$

To avoid trivialities, we will assume that X_M has full column rank for all M under consideration, so that β^M is well-defined.

Depending on whether σ^2 is assumed known or unknown, hypothesis tests for coordinates β_j^M generalize either the z-test or the t-test. In the non-selective case, z- and t-tests are based on



(a) Fisher information available for second-stage inference.

(b) Expected confidence interval length.

Figure 3.4: Contrast between data splitting and data carving in Example 3, in which $Y_i \sim N(\mu, 1)$ independently for $i = 1, 2$. Data splitting discards Y_1 entirely, while data carving uses the leftover information in Y_1 for the second-stage inference. When $\mu \ll 3$, data carving also uses about one data point for inference since there is no information left over in Y_1. But when $\mu \gg 3$, conditioning barely effects the law of Y_1 and data carving has nearly two data points left over.

coordinates of the ordinary least squares (OLS) estimator $\hat{\beta} = X_M^\dagger Y$, where $X_M^\dagger$ is the Moore-Penrose pseudoinverse. For a particular j and M, it will be convenient to write $\hat{\beta}_j^M = \eta' Y$ with

$$\eta_j^M = \frac{X_{j \cdot M}}{\|X_{j \cdot M}\|^2}, \quad \text{where } X_{j \cdot M} = \mathcal{P}_{X_{M \setminus j}}^{\perp} X_j \tag{3.46}$$

is the remainder after adjusting X_j for the other columns of X_M, and $\mathcal{P}_{X_{M \setminus j}}$ denotes projection onto the column space of $X_{M \setminus j}$. Letting $\hat{\sigma}^2 = \|\mathcal{P}_{X_M}^{\perp} Y\|^2 / (n - |M|)$, the test statistics

$$Z = \frac{\eta_j^{M\prime} Y}{\sigma \|\eta_j^M\|} \quad \text{and} \quad \widetilde{T} = \frac{\eta_j^{M\prime} Y}{\hat{\sigma} \|\eta_j^M\|} \tag{3.47}$$

are respectively distributed as $N(0, 1)$ and $t_{n-|M|}$ under $H_0 : \beta_j^M = 0$. Henceforth, we will suppress the subscript and superscript for η_j^M, simply writing η when there is no ambiguity.

We will see that the optimal selective versions of these tests are based on the same test statistics, but compared against different null distributions.

3.4.1 Inference Under the Selected Model

Suppressing the superscript M in β^M, the selected model has the form

$$Y \sim \exp\left\{\frac{1}{\sigma^2}\beta' {X_M}' y - \frac{1}{2\sigma^2}\|y\|^2 - \psi(X_M\beta, \sigma^2)\right\} \tag{3.48}$$

If σ^2 is known, the sufficient statistics are $X_k'Y$ for $k \in M$, and inference for β_j is based on

$$\mathcal{L}_{\beta_j}\left(X_j'Y \;\middle|\; {X_{M\setminus j}}'Y,\ A\right). \tag{3.49}$$

Otherwise, $\|Y\|^2$ represents another sufficient statistic and inference is based on

$$\mathcal{L}_{\beta_j/\sigma^2}\left(X_j'Y \;\middle|\; {X_{M\setminus j}}'Y,\ \|Y\|,\ A\right). \tag{3.50}$$

Decomposing

$$X_j'y = X_j'\mathcal{P}_{X_{M\setminus j}}y + X_j'\mathcal{P}^{\perp}_{X_{M\setminus j}}y \tag{3.51}$$

$$= X_j'\mathcal{P}_{X_{M\setminus j}}y + \|X_{j.M}\|^2\eta' y, \tag{3.52}$$

we see that $Z = \eta' Y/\sigma\|\eta\|$ is a fixed affine transformation of $X_j'Y$ once we condition on $X_{M\setminus j}'Y$. If σ^2 is known, then, we can equivalently base our selective test on

$$\mathcal{L}_{\beta_j}\left(Z \;\middle|\; {X_{M\setminus j}}'Y,\ A\right). \tag{3.53}$$

While Z is marginally independent of $X_{M\setminus j}'Y$, it is generically not conditionally independent given A, so that the null distribution of Z generically depends on $X_{M\setminus j}'Y$.

If σ^2 is unknown, we may observe further that

$$\hat{\sigma}^2 = \frac{\|\mathcal{P}^{\perp}_{X_M}Y\|^2}{n-|M|} = \frac{\|Y\|^2 - \|\mathcal{P}_{X_{M\setminus j}}Y\|^2 - (\eta' Y)^2/\|\eta\|^2}{n-|M|}. \tag{3.54}$$

Writing $Z_0(Y) = \eta' Y/\|\eta\|$, we have $\widetilde{T}(Y) = (n-|M|)\, Z_0/(\|Y\|^2 - \|\mathcal{P}_{X_{M\setminus j}}Y\|^2 - Z_0^2)$, which is a monotone function of $\eta' Y$ after fixing $\|Y\|^2$ and $X_{M\setminus j}'Y$. Thus, our test is based on the law of

$$\mathcal{L}_{\beta_j/\sigma^2}\left(\widetilde{T} \;\middle|\; {X_{M\setminus j}}'Y,\ \|Y\|,\ A\right). \tag{3.55}$$

Note that, given A, $\hat{\sigma}^2$ in (3.54) is neither unbiased for σ^2 nor χ^2-distributed. We recommend against viewing it as a serious estimate of σ^2 in the selective setting.

Constructing a selective t-interval is not as straightforward as the general case described in Section 3.5.2 because β_j is not a natural parameter of the selected model; rather, β_j/σ^2 is. Testing

$\beta_j = 0$ is equivalent to testing $\beta_j/\sigma^2 = 0$, but testing $\beta_j = c$ for $c \neq 0$ does not correspond to any point null hypothesis about β_j/σ^2. However, we can define

$$\widetilde{Y} = Y - bX_j \sim N(X\beta - bX_j, \sigma^2 I). \tag{3.56}$$

Because $(\beta_j - b)/\sigma^2$ is a natural parameter for $\widetilde{Y}$, we can carry out a UMPU selective t-test for $H_0 : \beta_j = b \iff (\beta_j - b)/\sigma^2 = 0$ based on the appropriate conditional law of $\widetilde{Y}$.

3.4.2 Inference Under the Saturated Model

Even if we do not take the linear model (3.45) seriously, there is still a well-defined best linear predictor in the population for design matrix X_M:

$$\theta^M = \arg\min_\theta \mathbb{E}_\mu \left[\|Y - X_M\theta\|^2\right] = X_M^\dagger \mu, \tag{3.57}$$

We call θ^M the *least squares coefficients* for M. According to this point of view, each θ_j^M corresponds to the linear functional $\eta_j^{M\prime}\mu$.

This point of view is convenient because the least-squares parameters are well-defined under the more general saturated model (3.6), leading to meaningful inference even if we do a poor job of selecting predictors. In particular, Berk et al. (2013) adopt this perspective as a way of avoiding the need to consider multiple candidate probabilistic models.

Several recent papers have tackled the problem of exact selective inference in linear regression after specific selection procedures (Lee et al., 2013; Loftus and Taylor, 2014; Lee and Taylor, 2014). These works, as well as Berk et al. (2013), assume the error variance is known, or that an estimate may be obtained from independent data, and target least-squares parameters in the saturated model.

Under the selected model, $\beta_j^M = \theta_j^M = \eta'\mu$, whereas under the saturated model β^M may not exist (i.e., there is no β^M such that $\mu = X_M\beta^M$). Compared to the selected model, the saturated model has $n - |M|$ additional nuisance parameters corresponding to $\mathcal{P}_{X_M}^\perp \mu$.

We can write the saturated model in exponential family form as

$$Y \sim \exp\left\{\frac{1}{\sigma^2}\mu' y - \frac{1}{2\sigma^2}\|y\|^2 - \psi(\mu, \sigma^2)\right\}, \tag{3.58}$$

which has $n+1$ natural paramaters if σ^2 is unknown and n otherwise. To perform inference on some least-squares coefficient $\theta_j^M = \eta'\mu$, we can rewrite (3.58) as

$$Y \sim \exp\left\{\frac{1}{\sigma^2\|\eta\|^2}\mu'\eta\,\eta' y + \frac{1}{\sigma^2}(\mathcal{P}_\eta^\perp \mu)'\,(\mathcal{P}_\eta^\perp y) - \frac{1}{2\sigma^2}\|y\|^2 - \psi(\mu, \sigma^2)\right\}. \tag{3.59}$$

If σ^2 is known, inference for θ_j^M after selection event A is based on the conditional law $\mathcal{L}_{\theta_j^M}\left(\eta' Y \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$, or equivalently $\mathcal{L}_{\theta_j^M}\left(Z \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$.

If σ^2 is unknown, we must instead base inference on

$$\mathcal{L}_{\theta_j^M/\sigma^2}\left(\eta' Y \;\middle|\; \mathcal{P}_\eta^\perp Y,\ \|Y\|,\ A\right). \tag{3.60}$$

Unfortunately, the conditioning in (3.60) is too restrictive. The set

$$\left\{y:\ \mathcal{P}_\eta^\perp y = \mathcal{P}_\eta^\perp Y,\ \ \|y\| = \|Y\|\right\} \tag{3.61}$$

is a line intersected with the sphere $\|Y\| S^{n-1}$, and consists only of the two points $\{Y,\ Y - 2\eta' Y\}$, which are equally likely under the hypothesis $\theta_j^M = 0$. Thus, under the saturated model, conditioning on $\|Y\|$ leaves insufficient information about θ_j^M to carry out a meaningful test.

3.4.3 Saturated Model or Selected Model?

When σ^2 is known, we have a choice whether to carry out the z-test with test statistic $Z = \eta' Y/\sigma\|\eta\|$ in the saturated or the selected model. In other words, we must choose either to assume that $\mathcal{P}_{X_M}^\perp \mu = 0$ or to treat it as an unknown nuisance parameter. Writing

$$U = X_{M\setminus j}{}' Y, \quad \text{and} \quad V = \mathcal{P}_{X_M}^\perp Y, \tag{3.62}$$

we must choose whether to condition on U and V (saturated model) or only U (selected model). Conditioning on both U and V can never increase our power relative to conditioning only on U, and will in most cases lead to an inadmissible test per Theorem 16.

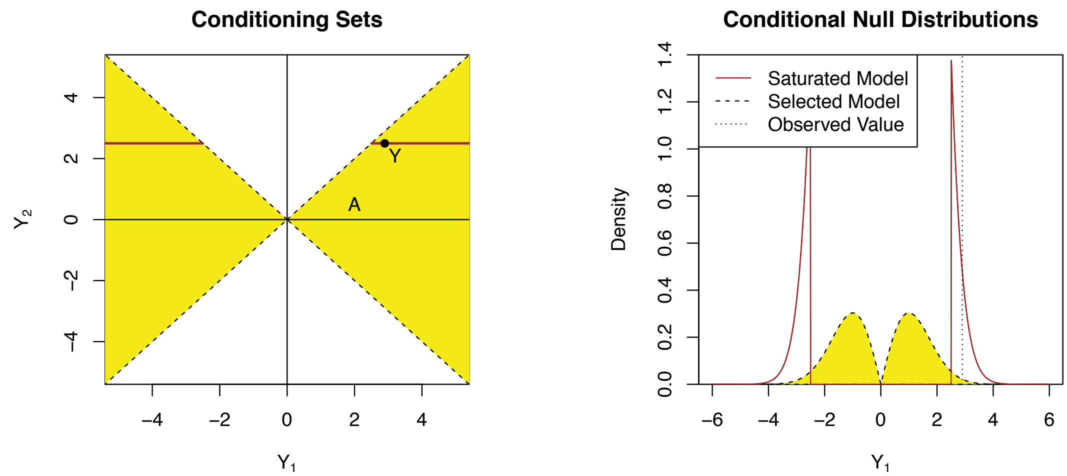
In the non-selective case, this choice makes no difference at all since T, U, and V are mutually independent. In the selective case, however, the choice may be of major consequence as it can lead to very different tests. In general, T, U, and V are not conditionally independent given A, and $\mathcal{P}_{X_M}^\perp \mu$ may play an important role in determining the conditional distribution of T. If we needlessly condition on V, we may lose a great deal of power, whereas failing to condition on V could lead us astray if $\mathcal{P}_{X_M}^\perp \mu$ is large. A simple example can elucidate this contrast.

Example 4. Suppose that $y \sim N_2(\mu, I_2)$, with design matrix $X = I_2$, and we choose the best one-sparse model. Our selection procedure chooses $M = \{1\}$ if $|Y_1| > |Y_2|$, and $M = \{2\}$ otherwise.

Figure 3.5 shows one realization of this process with $Y = (2.9, 2.5)$. $|Y_1|$ is a little larger than $|Y_2|$, so we choose $M = \{1\}$. The yellow highlighted region $A = \{|Y_1| > |Y_2|\}$ is the chosen selection event, and the selected model is

$$Y \sim N_2\left((\mu_1, 0), I_2\right). \tag{3.63}$$

In this case, $T = Y_1$, $V = Y_2$, and there is no U since X_M has only one column. The selected-model test is based on $\mathcal{L}(Y_1 \,|\, A)$, whereas the saturated-model test is based on $\mathcal{L}(Y_1 \,|\, Y_2, A)$. The



(a) For $Y = (2.9, 2.5)$, the selected-model conditioning set is $A = \{y : |y_1| > |y_2|\}$, a union of quadrants, plotted in yellow. The saturated-model conditioning set is $\{y : y_2 = 2.5\} \cap A = \{y : y_2 = 2.5, |y_1| > 2.5\}$, a union of rays, plotted in brown.

(b) Conditional distributions of Y_1 under $H_0 : \mu_1 = 0$. Under the hypothesis $\mu = 0$, the realized $|Y_1|$ is quite large given A, giving p-value 0.015. By contrast, $|Y_1|$ is not too large given $A \cap \{y : y_2 = Y_2\}$, giving p-value 0.3.

Figure 3.5: Contrast between the saturated-model and selected-model tests in Example 4, in which we fit a one-sparse model with design matrix $X = I_2$. The selected-model test is based on $\mathcal{L}_0(Y_1 \mid A)$, whereas the saturated-model test is based on $\mathcal{L}_0(Y_1 \mid Y_2, A)$.

second conditioning set, a union of two rays, is plotted in brown. Under the hypothesis $\mu = 0$, the realized $|Y_1|$ is quite large given A, giving p-value 0.015. By contrast, $|Y_1|$ is not terribly large given $\{Y_2 = 2.5\} \cap A = \{Y_2 = 2.5, |Y_1| > 2.5\}$, leading to p-value 0.30.

The difference between the saturated and selected models is especially important in early steps of sequential model-selection procedures that use a form of the saturated-model z-test. It has been observed in several cases that if there are two strong variables with similar effect sizes, the p-value in the first step may not be very small (Taylor et al., 2014; G'Sell et al., 2013). We will explore this subtle issue further in a forthcoming companion chapter dealing with sequential model selection.

3.5 Computations

We saw in Section 3.3 that inference in the one-parameter exponential family requires knowing the conditional law $\mathcal{L}_\theta(T \mid U, A)$. In a few cases, such as in the saturated model viewpoint, this conditional law can be determined fairly explicitly. In other cases, we will need to resort to Monte Carlo sampling. In this section, we suggest some general strategies.

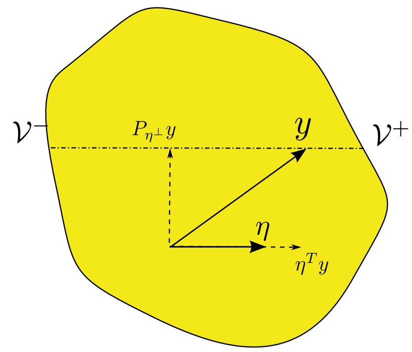


Figure 3.6: Saturated-model inference for a generic convex selection set for $Y \sim N(\mu, I_n)$. After conditioning on the yellow set A, $\mathcal{V}^+$ is the largest $\eta' Y$ can get while $\mathcal{V}^-$ is the smallest it can get. Under $H_0 : \ \eta'\mu = 0$, the test statistic $\eta' Y$ takes on the distribution of a standard Gaussian random variable truncated to the interval $[\mathcal{V}^-, \mathcal{V}^+]$. As a result, $W(Y) = \frac{\Phi(\eta' Y) - \Phi(\mathcal{V}^-)}{\Phi(\mathcal{V}^+) - \Phi(\mathcal{V}^-)}$ is uniformly distributed.

3.5.1 Gaussians Under the Saturated Model

As we discussed in Section 3.4.2, the previous chapters by Lee et al. (2013); Loftus and Taylor (2014); Lee and Taylor (2014) adopted the saturated model viewpoint with known σ^2. In this case, $\mathcal{L}_\theta(T \,|\, U, A) = \mathcal{L}_\theta\left(\eta' Y \;\middle|\; \mathcal{P}_\eta^\perp Y,\ A\right)$ is a truncated univariate Gaussian, since $\eta' Y$ is a Gaussian random variable and $\mathcal{P}_\eta^\perp Y$ is independent of $\eta' Y$. If A is convex, then the truncation is to an interval $[\mathcal{V}^-(Y), \mathcal{V}^+(Y)]$, where the endpoints represent the maximal extent one can move in the η direction at a "height" of $\mathcal{P}_\eta^\perp Y$, while still remaining inside A, i.e.,

$$\mathcal{V}^+(Y) = \sup_{\{t:\, Y + t\eta \,\in\, A\}} \eta'(Y + t\eta) \tag{3.64}$$

$$\mathcal{V}^-(Y) = \inf_{\{t:\, Y + t\eta \,\in\, A\}} \eta'(Y + t\eta). \tag{3.65}$$

The geometric intuition is illustrated in Figure 3.6.

When A is specifically a polytope, we can obtain closed-form expressions for $\mathcal{V}^-$ and $\mathcal{V}^+$. The generalization to regions A that are non-convex is straightforward (i.e., instead of truncating to a single interval, we truncate to a union of intervals). For further discussion of these points, see Lee et al. (2013).

3.5.2 Monte Carlo Tests and Intervals

More generally, if we can obtain a stream of samples from $\mathcal{L}_\theta(T \mid U, A)$ for any value of θ, then we can carry out hypothesis tests and construct intervals. This can be done efficiently via rejection sampling if, for example, we can sample efficiently from $\mathcal{L}_\theta(Y \mid U)$ and $\mathbb{P}_\theta(Y \in A \mid U)$ is not too small. Otherwise, more specialized sampling approaches may be required.

More generally, we consider how to construct a test based on the statistic Z, which is distributed according to a one-parameter exponential family

$$Z \sim g_\theta(z) = e^{\theta z - \psi(\theta)}\, g_0(z). \tag{3.66}$$

If in addition to Z we are given an independent sequence from the reference distribution

$$Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} g_0(z), \tag{3.67}$$

then an exact Monte Carlo one-sided test of $H_0 : \theta \leq 0$ rejects if the observed value Z is among the $(n+1)\,\alpha$ largest of $Z, Z_1, \ldots, Z_n$ (Barnard, 1963).

By reweighting the samples, we can use the same sequence to test $H_0 : \theta \leq \theta_0$ for any other θ_0. Denote the importance-weighted empirical expectation as

$$\widehat{\mathbb{E}}_\theta\, h(Z) = \frac{\sum_{i=1}^n h(Z_i) e^{\theta Z_i}}{\sum_{i=1}^n e^{\theta Z_i}} \tag{3.68}$$

$$\xrightarrow{a.s.} \mathbb{E}_\theta\, h(Z) \quad \text{as } n \to \infty \text{ for integrable } h. \tag{3.69}$$

In effect, we have put an exponential family "through" the empirical distribution of the Z_i in the manner of Efron et al. (1996); see also Besag (2001). The Monte Carlo one-sided cutoff for a test of $H_0 : \theta \leq \theta_0$ is the smallest c_2 for which

$$\widehat{\mathbb{P}}_{\theta_0}(Z > c_2) \leq \alpha. \tag{3.70}$$

The test rejects for $Z > c_2$ and randomizes appropriately at $Z = c_2$.

The Monte Carlo UMPU two-sided test of $H_0 : \theta = \theta_0$ is a bit more involved, but similar in principle. We can solve for $c_1, \gamma_1, c_2, \gamma_2$ for which

$$\widehat{\mathbb{E}}_{\theta_0} \phi(Z) = \alpha \tag{3.71}$$

$$\widehat{\mathbb{E}}_{\theta_0} \left[Z\phi(Z)\right] = \alpha\, \widehat{\mathbb{E}}_{\theta_0} Z. \tag{3.72}$$

In Appendix B we discuss how (3.71–3.72) can be solved efficiently for fixed θ_0 and inverted to obtain a confidence interval. Monte Carlo inference as described above is computationally straightforward once $Z_1, \ldots, Z_n$ are obtained.

More generally, the Z_i could represent importance samples with weights W_i, or steps in a Markov chain with stationary distribution $g_{\theta_0}(z)$. The same methods apply as long as we still have

$$\widehat{\mathbb{E}}_\theta\, h(Z) = \frac{\sum_{i=1}^n W_i\, h(Z_i)\, e^{\theta Z_i}}{\sum_{i=1}^n W_i\, e^{\theta Z_i}} \tag{3.73}$$

$$\xrightarrow{a.s.} \mathbb{E}_\theta\, h(Z), \qquad \text{for integrable } h. \tag{3.74}$$

Numerical problems may arise in solving (3.71–3.72) for θ_0 far away from the reference parameter used for sampling. Combining appropriately weighted samples from several reference values θ can help to keep the effective sample size from getting too small for any θ_0. For further references on Monte Carlo inference see Jockel (1986); Besag and Clifford (1989); Forster et al. (1996); Mehta et al. (2000).

3.5.3 Sampling Gaussians with Affine and Quadratic Constraints

In the case where Y is Gaussian, several simplifications are possible. For one, there are many ways to sample from a truncated multivariate Gaussian distribution. In this chapter, we use hit-and-run Gibbs sampling algorithms, while Pakman and Paninski (2014) suggest another approach based on Hamiltonian Monte Carlo.

Efficient sampling from multivariate Gaussian distributions under such constraints is the main algorithmic challenge for most of the Gaussian selective tests proposed in this chapter. The works cited above use the saturated model exclusively which means they do not require any sampling.

In many cases, the sampling problem may be greatly facilitated by refining the selection variable that we use. For example, Lee et al. (2013) propose conditioning on the variables selected by the lasso as well as the signs of the fitted $\hat{\beta}_j$, leading to a selection event consisting of a single polytope in $\mathbb{R}^n$. If we condition only on the selected variables and not on the signs, the selection event is a union of up to 2^s polytopes, where s is the number of variables in the selected model (though most of the polytopes might be excluded after conditioning on U).

Refining the selection variable never impairs the selective validity of the procedure, but it typically leads to a loss in power. However, this loss of power may be quite small if, for example, the conditional law puts nearly all of its mass on the realized polytope. This price in power is acceptable if it is the only way to obtain a tractable test. Quantifying the tradeoff between computation and power is an interesting topic for further work.

When carrying out selective t-tests, it is necessary to condition further on the realized vector length $\|Y\|$, adding a quadratic equality constraint to the support. To deal with this, we sample instead from a ball and project the samples onto the sphere using an importance sampling scheme. Appendix B gives details.

3.6 Selective Inference in Non-Gaussian Settings

In this section we describe tests in two simple non-Gaussian settings, selective inference in a binomial problem, and tests involving a scan statistic in Poisson process models. More generally, we address the question of selective inference in generalized linear models.

3.6.1 Selective Clinical Trial

To illustrate the application of our approach in a simple non-Gaussian setting we discuss a selective clinical trial with binomial data. The experiment discussed here is similar to an adaptive design proposed by Sill and Sampson (2009).

Consider a clinical trial with m candidate treatments for heart disease. We give treatment j to n_j patients for $0 \leq j \leq m$, with $j = 0$ corresponding to the placebo. The number of patients on treatment j to suffer a heart attack during the trial is

$$Y_j \overset{\text{ind.}}{\sim} \text{Binom}(p_j, n_j), \qquad \text{with } \log \frac{p_j}{1-p_j} = \begin{cases} \theta & j = 0 \\ \theta - \beta_j & j > 0 \end{cases}, \tag{3.75}$$

so β_j measures the efficacy of treatment j. The likelihood for Y is

$$Y \sim \exp\left\{ \theta \sum_{j=0}^{m} y_j - \sum_{j=1}^{m} \beta_j y_j - \psi(\theta, \beta) \right\} \prod_{j=0}^{m} \binom{Y_j}{n_j}, \tag{3.76}$$

an exponential family with $m+1$ sufficient statistics. Define $\hat{p}_j = Y_j / n_j$, and let $\hat{p}_{(j)}$ denote the jth smallest order statistic.

After observing the data, we select the best $k < m$ treatments in-sample, then construct a confidence interval for each one's odds ratio relative to placebo. If there are ties, we select all treatments for which $\hat{p}_j \leq \hat{p}_{(k)}$ (so that we could possibly select more than k treatments).

For simplicity, assume that treatments $1, \ldots, k$ are the ones selected. Inference for β_1 is then based on the conditional law

$$\mathcal{L}_{\beta_1}\left(Y_1 \;\middle|\; \sum_{j=0}^{m} Y_j,\ Y_2, \ldots, Y_J,\ \{j = 1 \text{ selected}\} \right) \tag{3.77}$$

Under this law, $Y_2, \ldots, Y_m$ are fixed, as is $Y_0 + Y_1$, with Y_0 and Y_1 the only remaining unknowns. Before conditioning on selection, we have the two-by-two multinomial table

	Control	Treatment
Heart attack	Y_0	Y_1
No heart attack	$n_0 - Y_0$	$n_1 - Y_1$

The margins are fixed, and conditioning on selection gives an additional constraint that $Y_1 \leq n_1\hat{p}_{(k)}$, where the right-hand side is known after conditioning on the other Y_j. Rejecting for conditionally extreme Y_1 amounts to a selective Fisher's exact test. Aside from the constraint on its support, the distribution of Y_1 is hypergeometric if $\beta_1 = 0$ and otherwise noncentral hypergeometric with noncentrality parameter β_1. We can use this family to construct an interval for β_1.

3.6.2 Poisson Scan Statistic

As a second simple example, consider observing a Poisson process $Y = \{Y_1, \ldots, Y_{N(Y)}\}$ on the interval $[0, 1]$ with piecewise-constant intensity, possibly elevated in some unknown window $[a, b]$. That is, $Y \sim \text{Poisson}(\lambda(t))$ with

$$\lambda(t) = \begin{cases} e^{\alpha+\beta} & t \in [a, b] \\ e^{\alpha} & \text{otherwise.} \end{cases} \tag{3.78}$$

Our goal is to locate $[a, b]$ by maximizing some scan statistic, then test whether $\beta > 0$ or construct a confidence interval for it. Assume we always have $[\hat{a}, \hat{b}] = [Y_i, Y_j]$ for some i, j; this is true, for example, if we use the multi-scale-adjusted likelihood ratio statistic proposed in Rivera and Walther (2013).

The density of Y can be written in exponential family form as

$$Y \sim \exp\left\{ \sum_{i=1}^{N(y)} \log \lambda(y_i) - \int_0^1 \lambda(s)\, ds \right\} \tag{3.79}$$

$$= \exp\big\{\alpha\, N(Y) + \beta\, T(y) - \psi(\alpha, \beta)\big\}, \tag{3.80}$$

where

$$T(y) = \sum_{i=1}^{N(y)} \mathbf{1}\{y_i \in [a, b]\} \quad \text{and} \quad \psi(\alpha, \beta) = e^{\alpha}(1 - b + a) + e^{\alpha+\beta}(b - a). \tag{3.81}$$

If A is the event that $[a, b]$ is chosen, we carry out inference with respect to $\mathcal{L}_\beta\,(T \mid N, A)$. Note that under $\beta = 0$ and conditional on N, Y is an i.i.d. uniform random sample on $[0, 1]$.

Once we condition on the event $\{a, b \in Y\}$, the other $N - 2$ values are uniform. Thus, we can sample from $\mathcal{L}_\beta(T \mid N, A)$ with $\beta = 0$ by taking Y to include a, b, and $N - 2$ uniformly random points, then rejecting samples for which $[a, b]$ is not the selected window.

3.6.3 Generalized Linear Models

Our framework extends to logistic regression, Poisson regression, or other generalized linear model (GLM) with response Y and design matrix X, since the GLM model may be represented as an exponential family of the form

$$Y \sim \exp\{\beta' X'y - \psi(X\beta)\}\ f_0(y). \tag{3.82}$$

As a result, we can proceed just as we did in the case of linear regression in the reduced model, conditioning on $U = X_{M\setminus j}{}'Y$ and basing inference on $\mathcal{L}_{\beta_j^M}(X_j'Y \mid U, A)$.

A difficulty may arise for logistic or Poisson regression due to the discreteness of the response distribution Y. If some control variable X_1 is continuous, then for almost every realization of X, all configurations of Y yield unique values of $U = X_1'Y$. In that case, conditioning on $X_1'Y$ means conditioning on Y itself. No information is left over for inference, so that the best (and only) exact level-α selective test is the trivial one $\phi(Y) \equiv \alpha$. By contrast, if all of the control variables are discrete variables like gender or ethnicity, then conditioning on U may not constrain Y too much.

Because $X'Y$ is approximately a multivariate Gaussian random variable, a more promising approach may to base inference on the asymptotic Gaussian approximation, though we will not pursue that here. Tian and Taylor (2015) discuss selective inference in certain non-Gaussian problems.

3.7 Simulation: High-Dimensional Regression

As a simple illustration, we compare selective inference in linear regression after the lasso for $n = 100, p = 200$. Here, the rows of the design matrix X are drawn from an equicorrelated multivariate Gaussian distribution with pairwise correlation $\rho = 0.3$ between the variables. The columns are normalized to have length 1.

We simulate from the model

$$Y \sim N(X\beta, I_n), \tag{3.83}$$

with β 7-sparse and its non-zero entries set to 7. The signal to noise ratio (SNR) (magnitude of β) was chosen so that data splitting with half the data yielded a superset of the true variables on roughly 20% of instances. For data splitting and carving, Y is partitioned into selection and inference data sets Y_1 and Y_2, containing n_1 and $n_2 = n - n_1$ data points respectively.

We compare two procedures:

Data Splitting after Lasso on Y_1 (Split$_{n_1}$): Use the lasso on Y_1 to select the model, and use Y_2

for inference.

Data Carving after Lasso on Y_1 (Carve$_{n_1}$): Use the lasso on Y_1 to select the model, and use Y_2 and whatever is left over of Y_1 for inference.

Procedure Carve$_{100}$ is inference after the using the lasso on the full data set Y.

For the data carving procedures, we use the selected-model z-test of Section 3.4.1 after lasso variable selection using Lagrange parameter

$$\lambda = 2\mathbb{E}(\|X^T \epsilon\|_\infty), \quad \epsilon \sim N(0, \sigma^2 I)$$

as described in (Negahban et al., 2012). In addition, we condition on the signs of the active lasso coefficients, so that procedure Carve$_{100}$ is the inference-after-lasso test considered in Lee et al. (2013).[4]

We know from Theorem 16 that procedure Carve$_{n_1}$ strictly dominates procedure Split$_{n_1}$ for any n_1, but there is a selection–inference tradeoff between data-carving procedures Carve$_n$ and Carve$_{n_1}$ for $n_1 < n$. Carve$_n$ uses all of the data for selection, and is therefore likely to select a superior model, whereas procedure Carve$_{n_1}$ reserves more power for the second stage.

Let R be the size of the model selected and V the number of noise variables included. We compare the procedures with respect to aspects of their selection performance:

- chance of screening, i.e. obtaining a correct model ($\mathbb{P}(R - V = 7)$ or p_{screen}).
- expected number of true variables selected ($\mathbb{E}[R - V]$),
- false discovery rate of true variables selected ($\mathbb{E}[V/\max(R, 1)]$ or FDR),

Conditional on having obtained a correct model, we also compare them on aspects of their second stage performance:

- probability of correctly rejecting the null for one of the true variables (Power),
- probability of incorrectly rejecting the null for a noise variable (Level).

The results, shown in Table 3.1, bear out the intuition of Section 3.3.2. Because procedure Carve$_{100}$ uses the most information in the first stage, it performs best in terms of model selection, but pays a price in lower second-stage power relative to Split$_{50}$ or Carve$_{50}$. The procedure Carve$_{50}$ clearly dominates Split$_{50}$, as expected. Increasing n_1 from 50 to 75 improves p_{screen} for Split$_{75}$, but Split$_{75}$ suffers a drop in power. Procedure Carve$_{75}$ seems to strike a better compromise.

Figure 3.7 shows the tradeoff curve of model selection success (as measured by the probability of successful screening) against second-stage power conditional on successful screening. As n_1 increases,

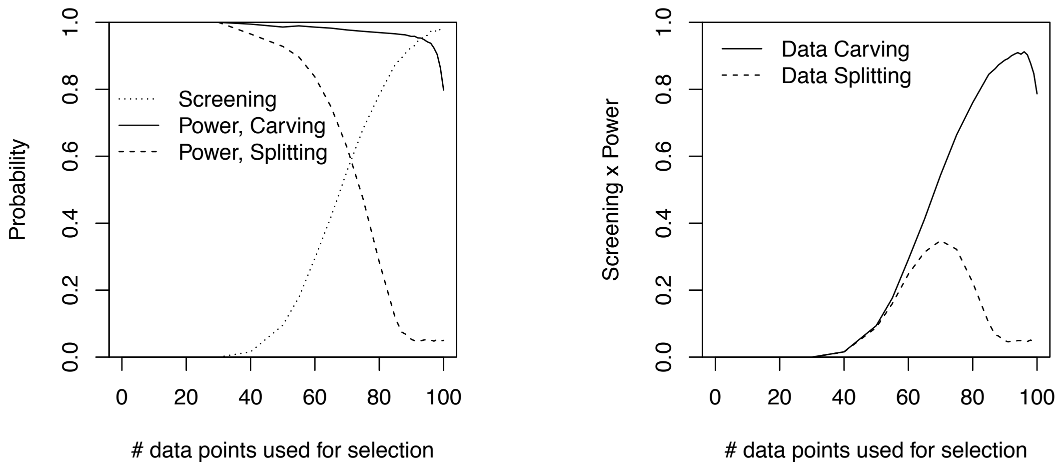
[4] Because of the form of the selection event when we use the lasso after n data points, the test statistic is conditionally independent of $\mathcal{P}_{X_M}^{\perp} Y$. Thus, there is no distinction between the saturated- and selected-model z-tests after the lasso on all n data points.

Algorithm	p_{screen}	$\mathbb{E}[R-V]$	$\mathbb{E}[R]$	FDR	Power	Level
Carve_{100}	0.98	6.98	15.28	0.52	0.79	0.05
Split_{50}	0.17	5.21	14.19	0.62	0.93	0.05
Carve_{50}	0.17	5.21	14.19	0.62	0.99	0.05
Split_{75}	0.76	6.70	15.63	0.55	0.48	0.05
Carve_{75}	0.76	6.70	15.63	0.55	0.97	0.06

Table 3.1: Simulation results. p_{screen} is the probability of successfully selecting all 7 true variables, and Power is the power, conditional on successful screening, of tests on the true variables. The more data we use for selection, the better the selected model's quality is, but there is a cost in second-stage power. Carve_{75} appears to be finding a good tradeoff between these competing goals. Carve_{n_1} always outperforms Split_{n_1}, as predicted by Theorem 16.

Algorithm	p_{screen}	$\mathbb{E}[R-V]$	$\mathbb{E}[R]$	FDR	Power	Level
Carve_{100}	0.97	6.97	15.11	0.51	0.80	0.05
Split_{50}	0.18	5.26	14.30	0.62	0.93	0.05
Carve_{50}	0.18	5.26	14.30	0.62	0.98	0.06

Table 3.2: Simulation results under misspecification. Here, errors ϵ are drawn independently from Student's t_5. Our conclusions are identical to Table 3.1.



(a) Probability of successful screening, and power conditional on screening, for Split_{n_1} and Carve_{n_1}.

(b) Probability of successful screening times power conditional on screening, for Split_{n_1} and Carve_{n_1}.

Figure 3.7: Tradeoff between power and model selection. As n_1 increases and more data is used in the first stage, we have a better chance of successful screening (picking all the true nonzero variables). However, increasing n_1 also leads to reduced power in the second stage. Data splitting suffers much more than data carving, though both are affected.

stage-one performance improves while stage-two performance declines, but the decline is much slower for data carving. Surprisingly, Carve_{98} and Carve_{99} have much higher power than Carve_{100}: 90%, 86%, and 79% respectively. We cannot explain why holding out just one or two data points in the first stage improves power so dramatically. Better understanding this tradeoff is an interesting topic of further work.

Finally, to check the robustness of data carving, we replace the Gaussian errors with independent errors drawn from Student's t distribution with five degrees of freedom. The numbers barely change at all; see Table 3.2. Tian and Taylor (2015) rigorously analyze the case of non-Gaussian errors.

3.8 Selective Inference and Multiple Inference

A common approach to the problem of inference after selection is to replace the nominal error rate with an alternative joint error rate that is deemed appropriate to the scientific setting.

For example, suppose that $\theta_q,\ q = 1, \ldots, m$ correspond to parameters of a common model M. We designate a small number $R(Y) = |\widehat{\mathcal{Q}}(Y)|$ of them as interesting and construct a confidence interval $C_q(Y)$ for each $q \in \widehat{\mathcal{Q}}$. Benjamini and Yekutieli (2005) propose controlling the *false coverage-statement rate* (FCR)

$$\mathbb{E}\left[\frac{V}{\max(R, 1)}\right], \quad \text{where} \quad V(Y) = \left|\left\{q : q \in \widehat{\mathcal{Q}},\ \theta_q(F) \notin C_q(Y)\right\}\right| \tag{3.84}$$

is the number of non-covering intervals constructed.

FCR control is closely related to selective coverage. By choosing appropriate selection variables S_q, we can adapt selective coverage to control the FCR.

Proposition 18 (FCR Control via Selective Error Control)**.** *Assume $\mathcal{Q}$ is countable with each $q \in \mathcal{Q}$ corresponding to a different parameter θ_q for the same model M. Let $R(Y) = |\widehat{\mathcal{Q}}(Y)|$ with $R(Y) < \infty$ a.s., and define $V(Y)$ as in (3.84).*

If each C_q enjoys coverage at level $1-\alpha$ given $S_q = \big(\mathbf{1}_{A_q}(Y), R(Y)\big)$, then the collection of intervals $(C_q, q \in \widehat{\mathcal{Q}})$ controls the FCR at level α:

$$\mathbb{E}\left[\frac{V}{\max(R,1)}\right] \;\le\; \mathbb{E}\left[\frac{V}{R} \;\Big|\; R \ge 1\right] \;\le\; \alpha. \tag{3.85}$$

Proof. Let $V_q(Y) = \mathbf{1}\left\{q \in \widehat{\mathcal{Q}}(Y),\, \theta_q(F) \notin C_q(Y)\right\}$, so that $V = \sum_{q\in\mathcal{Q}} V_q$. For $R \ge 1$, and for any $F \in M$,

$$\mathbb{E}_F\left[V \mid R\right] \;=\; \sum_{q\in\mathcal{Q}} \mathbb{E}_F\left[V_q \mid R\right] \;\le\; \sum_{q\in\mathcal{Q}} \alpha\, \mathbb{E}_F\left[\mathbf{1}_{A_q}(Y) \mid R\right] \;=\; \alpha R, \tag{3.86}$$

hence $\mathbb{E}\left[V/R \mid R\right] = \alpha$ for each $R \ge 1$. $\square$

Other authors have addressed inference after selection by proposing to control the FWER, the chance that any selected test incorrectly rejects the null or any constructed confidence interval fails to cover its parameter. For example, the "post-selection inference" (PoSI) method of Berk et al. (2013) constructs simultaneous $(1-\alpha)$ confidence intervals for the least-squares parameters of all linear regression models that were ever under consideration. As a result, no matter how we choose the model, the overall probability of constructing any non-covering interval is controlled at α.

Selective error control can be adapted to control the FWER as well. If our selection rule always chooses a single hypothesis to test, then we have overall FWER control.

Proposition 19 (FWER Control for Singleton $\widehat{\mathcal{Q}}$)**.** *Assume that $|\widehat{\mathcal{Q}}(Y)| \overset{a.s.}{=} 1$, and let $Q(Y) = (M(Y), H_0(Y))$ denote the single (random) selected model–hypothesis pair. If each ϕ_q controls the selective error at level α, then the test $\phi(y) = \phi_{Q(y)}(y)$ controls the FWER at level α:*

$$\mathbb{E}_F\left[\phi(Y);\, F \in H_0(Y)\right] \le \alpha \tag{3.87}$$

Proof. Condition on Q:

$$\mathbb{E}_F\left[\phi(Y)\mathbf{1}\{F \in H_0(Y)\}\right] = \mathbb{E}_F\left[\, \mathbb{E}_F\left[\phi(Y) \mid Q(Y)\right] \mathbf{1}\{F \in H_0(Y)\}\,\right] \tag{3.88}$$

$$\le \alpha\, \mathbb{P}_F\left[F \in H_0(Y)\right] \tag{3.89}$$

$\square$

More generally, it is clear that if we construct $(\phi_q, q \in \mathcal{Q})$ to control any joint error rate conditional on the entire selected set $\widehat{\mathcal{Q}}$, we will also have marginal control of the same joint error rate.

However, the converse of Proposition 19 is not true: FWER control does *not* in general guarantee control of relevant selective error rates. For example, suppose $Q(Y) = 1$ with probability 0.9 and $Q(Y) = 2$ otherwise. If ϕ_1 and ϕ_2 have selective error rates $\alpha_1 = 0.02$ and $\alpha_2 = 0.3$ respectively, the overall FWER is still controlled at $\alpha = 0.05$.

Does our conservatism when asking question 1 compensate for our anti-conservatism when asking question 2? To answer this question we must consider not only mathematics but also the relevant scientific context. If the different questions represent a bag of anonymous, *a priori* undifferentiated hypotheses, then a joint error rate like the FWER or FDR may be a good proxy for our scientific goals. For example, when performing a large-scale genome-wide "fishing expedition" for loci associated with type II diabetes, the fraction of null genes among all purported discoveries is a very relevant quantity: it measures what fraction of our time and money will be wasted following up on false leads.

In other scientific applications, however, different hypotheses have quite distinct identities and may vary greatly in their importance and interpretation. For example, a confidence interval for the effect of gender on salary after controlling for one's job title may be much more socially consequential than an interval for the effect of job title after controlling for gender. As such, averaging our error rates across the two questions is inappropriate.

3.9 Discussion

Selective inference concerns the properties of inference carried out after using a data-dependent procedure to select which questions to ask. We can recover the same long-run frequency properties among answers to *selected* questions that we would obtain in the classical non-adaptive setting, if we follow the guiding principle of selective error control:

The answer must be valid, given that the question was asked.

Happily, living up to this principle can be a simple matter in exponential family models including linear regression, due to the rich classical theory of optimal testing in exponential family models. Even if we are possibly selecting from a large menu of diverse and incompatible models, we can still design tests one model at a time and control the selective error using the test designed for the selected model. We generally pay a price for conditioning, so it is desirable to condition on as little as possible. Data carving can dramatically improve on data splitting by using the leftover information in Y_1, the data set initially designated for selection.

Many challenges remain. Deriving the cutoffs for sample carving tests can be computationally difficult in general. In addition, the entire development of this chapter takes the model selection procedure $\widehat{\mathcal{Q}}$ as given, when in reality we can choose $\widehat{\mathcal{Q}}$. More work is needed to learn what model selection procedures lead to favorable second-stage properties.

As data sets and research questions become more and more complex, we have less and less hope of specifying adequate probabilistic models ahead of time. As such, a key challenge of complex research is to balance the goal of choosing a realistic model against the goal of inference once we have chosen it. We hope that the ideas in this chapter represent a step in the right direction.

Chapter 4

Selective Inference: More Examples

This remaining chapter details several further forays into specific problems in selective inference that I find illuminating or otherwise interesting.

4.1 Rank Verification

Our next example concerns the question of how to statistically verify empirical rankings based on simple exponential family models. As motivation, consider a May 4, 2015 Quinnipiac University poll of 667 Iowa Republicans Poll (2015b). The poll found that Scott Walker was in the lead with 21% of the vote, Rand Paul and Marco Rubio shared second place with 13%, and twelve other candidates (including "Don't Know") trailed the leading candidates.

Table 4.1 shows the results of the poll. Note that we will make a simplifying assumption that poll represents a random sample of Iowa Republicans — i.e., that the data are multinomial. In reality, Quinnipiac has applied proprietary algorithms for post-processing the data to make the sample more representative of the population of likely voters.

Having observed that Scott Walker received the most votes in this poll, we may be motivated to ask whether he is actually in the lead in the population of all Iowa Republicans. In Section 4.1.2 we construct a test for the hypothesis that the winning candidate in a poll is actually leading in the population.

Before discussing the multinomial case, we address a similar problem in the Gaussian case, of testing whether the largest sample mean in an ANOVA actually came from the group with the largest population mean.

Rank	Candidate	Result	Votes
1.	Scott Walker	21%	140
2.	Rand Paul	13%	87
3.	Marco Rubio	13%	87
4.	Ted Cruz	12%	80
⋮		⋮	
14.	Bobby Jindal	1%	7
15.	Lindsey Graham	0%	0

Table 4.1: Results from a Quinnipiac University poll of 667 Iowa Republicans. To compute the last column (Votes), we make the simplifying assumption that the reported percentages in the third column (Results) correspond to raw vote shares among survey respondents.

4.1.1 Gaussian Case: Ranking Group Means

Consider a one-way layout in which we observe samples of size n from m Gaussian populations with the same variance. The model is

$$M: \; Y_{i,j} \overset{\text{ind.}}{\sim} N(\mu_i, \sigma^2), \quad i = 1, \ldots, m, \;\; j = 1, \ldots, n. \tag{4.1}$$

For example, we could imagine a multi-arm trial with m competing treatments, where larger values of Y correspond to better outcomes.

After observing all the $Y_{i,j}$, we compute a mean $\overline{Y}_i$ for each sample and denote $i^* = \arg\max_i \overline{Y}_i$. Having observed which group has the largest sample mean, we wish to test whether the corresponding μ_i is in fact the largest one — i.e., whether treatment i is in fact the best treatment. An issue of selection bias arises because we only test H_i on the occasions when $\overline{Y}_i$ is large; thus, we might expect "naïve" procedures that do not account for selection to be anti-conservative.

Informally, if $\overline{Y}_i$ is the largest, then we test the null hypothesis that μ_i is *not* the largest. Note that in this case, the statistical model is not selected but is the same for every question; only the null hypothesis differs from one question to the next. Formally, in the notation of Chapter 3, the possible questions are $q_i = (H_i, M)$ where

$$H_i : \; \mu_i \leq \max_{j \neq i} \mu_j \quad = \quad \bigcup_{j \neq i} H_{i,j} : \; \mu_i \leq \mu_j. \tag{4.2}$$

The selection event for q_i is $A_i = \left\{ \overline{Y}_i > \max_{j \neq i} \overline{Y}_j \right\}$. Because the selection events partition the sample space, we always ask exactly one question, and so controlling the selective error $\mathbb{P}_{H_i}(\text{ reject } H_i \mid A_i)$ at level α implies control of the family-wise error rate $\mathbb{P}(\text{reject any true } H_i)$ at α.

Pairwise Comparisons Approach

As stated, this problem is amenable to classical methods for post-hoc comparisons in ANOVA. Let

$$\varepsilon_i = \overline{Y}_i - \mu_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2/n), \tag{4.3}$$

and

$$S^2 = \frac{1}{m(n-1)} \sum_{i,j} (Y_{i,j} - \overline{Y}_i)^2 \sim \frac{\sigma^2}{m(n-1)} \; \chi^2_{m(n-1)}. \tag{4.4}$$

Then Tukey's correction for all pairwise comparisons (Tukey, 1951) would reject $H_{i,j}$ if $\overline{Y}_i - \overline{Y}_j > S\sqrt{2/n}\, r_{m,n,\alpha}$, where r_α is the $1 - \alpha$ quantile of the *studentized range distribution*

$$R \; = \sup_{1 \leq i, j \leq m} \frac{|\varepsilon_i - \varepsilon_j|}{S\sqrt{2/n}}, \tag{4.5}$$

If $m = 2$, R is the absolute value of a student's t-distribution with $2(n-1)$ degrees of freedom and $r_{2,n,\alpha} = t_{2(n-1),\alpha/2}$, the upper $\alpha/2$ quantile of a t-distribution. Otherwise, R is a maximum of $\binom{m}{2}$ correlated t-distributions with $m(n-1)$ degrees of freedom.

If σ^2 is known, we replace S with σ in (4.5). The upper quantile of the corresponding distribution is denoted by $r_{m,\infty,\alpha}$, and $r_{2,\infty,\alpha} = z_{\alpha/2}$, the upper $\alpha/2$ quantile of a $N(0,1)$ distribution.

By construction, Tukey's procedure makes no errors unless R is larger than its upper quantile. Consequently, it controls the familywise error rate at level α. In the end, then, we declare group i to be the largest if and only if

$$\overline{Y}_i \quad \geq \quad \sqrt{2/n}\; S\, r_{m,n,\alpha} + \max_{j \neq i} \overline{Y}_j, \tag{4.6}$$

or if σ^2 is known, we replace $S\, r_{m,n,\alpha}$ with $\sigma\, r_{m,\infty,\alpha}$ in (4.6).

Selective Approach

By applying the techniques of Chapter 3 we can obtain a substantially more powerful test. We begin by constructing valid $p_{i,j}$ for the hypothesis $H_{i,j}$ on event A_i. To simplify notation, assume without loss of generality that $i = 1$ and $j = 2$.

If we reparametrize the problem as

$$\theta = \frac{\mu_1 - \mu_2}{2\sigma^2/n}, \;\; \zeta = \frac{\mu_1 + \mu_2}{2\sigma^2/n}, \tag{4.7}$$

then we obtain the exponential family model

$$Y \sim \exp\left\{\theta(\overline{Y}_1 - \overline{Y}_2) + \zeta(\overline{Y}_1 + \overline{Y}_2) + \sum_{i>2} \frac{n\mu_i}{\sigma^2}\overline{Y}_i - \frac{1}{2\sigma^2}\|Y\|^2 - \psi(\theta, \zeta, \mu, \sigma^2)\right\}, \tag{4.8}$$

with $H_{1,2}$ corresponding to the hypothesis $\theta \leq 0$.

If σ^2 is known, we can take $p_{i,j}$ to be the survival function of

$$\mathcal{L}_{\theta=0}\left(\overline{Y}_1 - \overline{Y}_2 \;\middle|\; \overline{Y}_1 + \overline{Y}_2,\, Y_3, \ldots, Y_m,\, A_1\right), \tag{4.9}$$

and reject if $p_{i,j} \leq \alpha$, which occurs with probability less than α under the null.

If σ^2 is unknown, then we modify (4.9) by conditioning on $\|Y\|$ as well. In that case the distribution of $\overline{Y}_1 - \overline{Y}_2$ is conditionally t-distributed.

Having constructed $p_{i,j}$ in this way, we can take

$$p_i = \max_{j \neq i} p_{i,j}, \tag{4.10}$$

and reject H_i if $p_i \leq \alpha$; i.e., if every $p_{i,j}$ is smaller than α. Then, we have

$$\mathbb{P}(p_i \leq \alpha \mid A_i) \leq \min_{j \neq i} \mathbb{P}(p_{i,j} \leq \alpha \mid A_i). \tag{4.11}$$

If H_i is true, then at least one $H_{i,j}$ is true, so the right-hand side of (4.11) is smaller than α.

Perhaps surprisingly, in the case of known σ^2, the selective approach described in the preceding paragraphs reduces to performing an *unadjusted* z-test comparing the largest sample mean to the second-largest one. That is, the entire procedure amounts to declaring μ_i to be the largest if and only if

$$\overline{Y}_i \;\; \geq \;\; \sqrt{2/n}\,\sigma\, z_{\alpha/2} + \max_{j \neq i} \overline{Y}_j, \tag{4.12}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of a $N(0, 1)$ distribution. An interesting question for further study is whether a similar result holds in the case of unknown σ^2, with σ^2, with $\sigma z_{\alpha/2}$ replaced by $St_{m(n-1),\alpha/2}$.

Theorem 20. *If $\overline{Y}_i > \overline{Y}_j > \max_{k \notin \{i,j\}} \overline{Y}_k$, then the selective p-value for H_i is*

$$p_i = p_{i,j} = 2\left\{1 - \Phi\left(\frac{\overline{Y}_i - \overline{Y}_j}{\sigma\sqrt{2/n}}\right)\right\}. \tag{4.13}$$

Thus, the test rejects H_i if and only if (4.12) holds.

The proof is deferred to the Appendix.

Power Comparison: Tukey vs. Selective Test

The selective procedure and Tukey's procedure coincide exactly if there are exactly two means. For $m > 2$, the two procedures differ in the required gap between the largest and second-largest sample means. Tukey's procedure requires the gap to be a factor $r_{m,\infty,\alpha}/z_{\alpha/2}$ times larger; for large m this ratio is on the order of $\sqrt{\log m}$.

Figure 4.1 displays power curves for a "one-sparse" model in which $\mu_j = 0$ for all $j > 1$ and $\mu_1 > 0$. Here $\sigma^2 = 1$ is known, and $n = 1$. If there are 30 groups and $\mu_1 = 6$, the selective procedure enjoys 86% power, compared to 11% power for Tukey's procedure.

It may appear quite surprising that "verifying a winner" as we propose in this section requires no adjustment for multiplicity, in the sense that we only need to compare the winner to the runner-up, just as we would have done if there were only two candidates. After all, we know that the winner suffer's from a "winner's curse": given that treatment 5 wins the trial, it likely that Y_5 is an overestimate of μ_5, and the more groups there are, the more severe this bias is likely to be. However, if treatment 9 is the trial's runner-up, then Y_9 is likely biased upward nearly as much as Y_5. Informally, we might say that Y_9 suffer's from a "runner-up's curse" that is nearly as bad as the winner's curse; as a result, $Y_5 - Y_9$ is actually not all that biased upward. Performing a two-tailed instead of a one-tailed test takes care of what little bias there is. Indeed, performing a two-tailed test can be viewed as a sort of "correction for multiplicity," albeit one that we would have to make even if there were only two treatments in the trial.

4.1.2 Multinomial Case: Ranking Candidates Using Polling Data

We return now to the question of whether Scott Walker was really leading among Iowa Republicans. We model the number of votes received by the candidates as multinomial:

$$M:\ (Y_1, \ldots, Y_m) \sim \text{Multinom}(n, \pi) \tag{4.14}$$

We set up our decision problem in direct analogy to Section 4.1.1. We will test

$$H_i:\ \pi_i \leq \max_{j \neq i} \pi_j \quad = \quad \bigcup_{j \neq i} H_{i,j}:\ \pi_i \leq \pi_j. \tag{4.15}$$

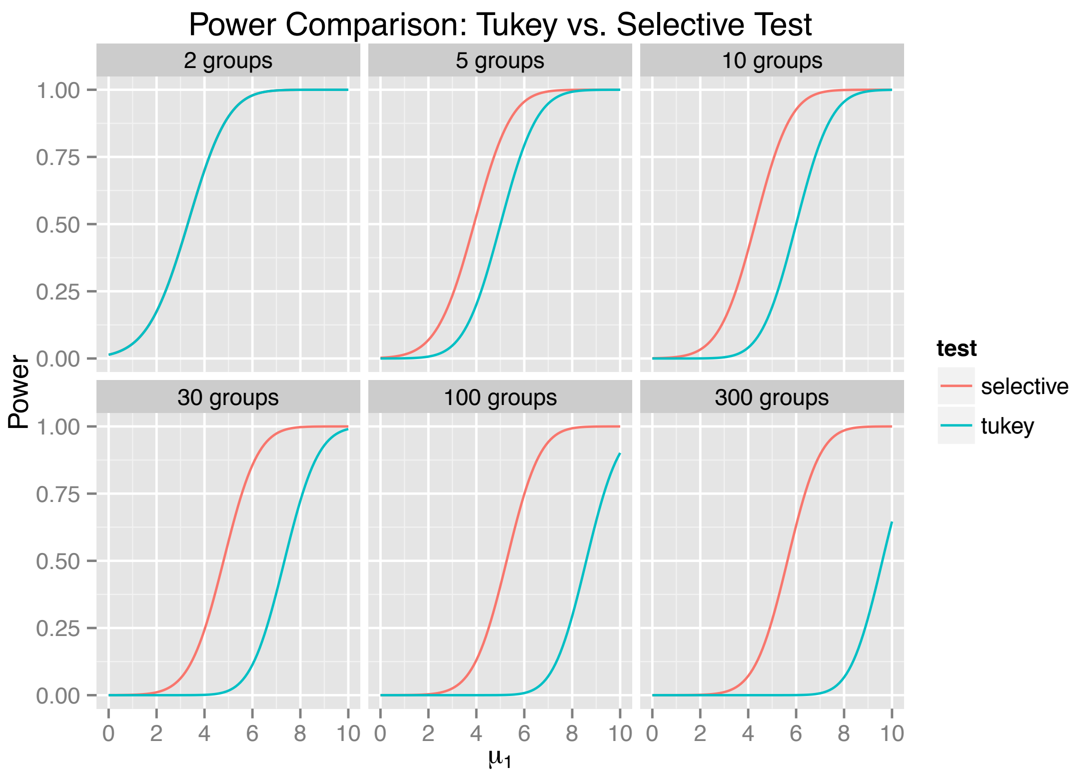


Figure 4.1: Power curves as a function of μ_1 for the simulation comparing the power of Tukey's procedure to the selective test.

when Y_i is larger than the other Y_j. To break ties, we introduce independent auxiliary random variables $U_i \sim U[0,1]$ and take

$$A_i = \left\{ Y_i + U_i > \max_{j \neq i} Y_j + U_j \right\} \tag{4.16}$$

As before, we will construct the one-sided p-value $p_{i,j}$ for testing $H_{i,j}$ on A_i, and take $p_i = \max_{j \neq i} p_{i,j}$.

Designing a selective pairwise test is a simple matter once we reparameterize our model in exponential family form:

$$\exp\left\{ \sum_i Y_i \beta_i - n \log \sum_i e^{\beta_i} \right\} \frac{n!}{\prod_i Y_i!},$$

with $\pi_i = e^{\beta_i} \big/ \sum_j e^{\beta_j}$.

Without loss of generality, assume we are testing $H_{1,2}$ on the event A_1. The parameter of interest is $\theta = \beta_1 - \beta_2$, and $H_{1,2}$ is equivalent to the hypothesis that $\theta \leq 0$.

To eliminate the nuisance parameter $(\beta_1 + \beta_2, \beta_3, \ldots, \beta_m)$, we condition on $(Y_1 + Y_2, Y_3, \ldots, Y_m)$. The test statistic is $Y_1 - Y_2$, or equivalently Y_1, whose distribution given the above is

$$Y_1 \mid (Y_1 + Y_2, Y_3, \ldots, Y_m) \sim \text{Binom}\left(Y_1 + Y_2,\ \frac{e^\theta}{1 + e^\theta} \right). \tag{4.17}$$

After conditioning further on A_1, the distribution of Y_1 is a truncated binomial distribution.

As in the Gaussian case, the entire procedure reduces to a single two-sided binomial test involving only the largest two counts.

Theorem 21. *Let $Y_i \geq Y_j \geq \max_{k \notin \{i,j\}} Y_k$, with i selected. The selective p-value p_i for H_i reduces to the unadjusted two-sided binomial p-value for comparing Y_i to Y_j.*

That is, if $Y_i = Y_j$ then $p_i = p_{i,j} = 1$, and otherwise

$$p_i = p_{i,j} = 2\left\{1 - B_{Y_i + Y_j, 1/2}(Y_i - 1)\right\} \tag{4.18}$$

$$= 2 \sum_{x = Y_i}^{Y_i + Y_j} \binom{Y_i + Y_j}{x} \Big/ 2^{Y_i + Y_j}, \tag{4.19}$$

where $B_{n,p}$ denotes the cumulative distribution function of $\text{Binom}(n,p)$.

As before, we defer the proof to the appendix. Applying Theorem 21 to the Quinnipiac poll is as simple as typing `binom.test(140, 140 + 87)` into R. We obtain the p-value 0.00053 and lower confidence bound 0.20 for $\beta_{\text{SW}} - \max_{j \neq \text{SW}} \beta_j$: that is, we can state with 95% confidence that Scott Walker's support is at least 22% higher than all other candidates.

Contrasting the Selective and Pairwise Comparisons Approaches

There is an asymptotic analog of Tukey's procedure for multinomial data. If the counts are relatively large, then we can model $n \sim \text{Poisson}(\lambda)$, in which case

$$Y_i \overset{\text{ind.}}{\sim} \text{Poisson}(\lambda\pi_i) \approx N(\lambda\pi_i, \lambda\pi_i) \tag{4.20}$$

We can apply this normal approximation along with the Games–Howell correction for unequal variances to obtain an asymptotic post-hoc p-value of 0.033 for H_i.

While 0.033 is still significant by the usual standards, this test is substantially less powerful than the selective multinomial approach outlined above. If we resample $Y^* \sim \text{Multinom}(667, \hat{\pi})$, we can compute the probability of declaring with 95% confidence that Scott Walker the true leader, conditional on Walker receiving the most votes. Under this resampling distribution, the Tukey–Games–Howell procedure has 35% power, while the selective multinomial test has 88% power, a substantial difference.

4.1.3 A Stepdown Procedure for Rank Verification

Table 4.2 shows the results of a companion poll of Iowa Democrats taken by Quinnipiac University at the same time as the poll of Table 4.1 (Poll, 2015a). There is little doubt that Hillary Clinton is currently the leading Democratic candidate in Iowa, but can we also state confidently who is in second place, third place, and so on?

Assume that $Y_1 > Y_2 > Y_3 \geq \max_{j>3} Y_j$. If we reject $H_1 = \bigcup_{j>1} H_{1,j}$, then we may wish to test

$$H_2^{\{2,\ldots,m\}} = \bigcup_{j>2} H_{2,j}, \tag{4.21}$$

on the event

$$A_2^{1,\{2,\ldots,m\}} = \left\{ Y_1 + U_1 > Y_2 + U_2 > \max_{j>2} Y_j + U_j \right\} \cap \{\text{reject at step 1}\} \tag{4.22}$$

Conditional on Y_1, the event that the first test rejected at level α is simply the event that $\max_{j>1} Y_j < C_1(Y_1)$, for some cutoff value $C_1(Y_1) \leq Y_1$. The pairwise p-value $p_{2,j}$ at step 2 is based on upper quantiles of

$$\mathcal{L}\left(Y_2 \mid Y_2 + Y_j,\ (Y_i)_{i \notin \{2,j\}},\ A_2^{1,\{2,\ldots,m\}} \right) \tag{4.23}$$

The p-value $p_{2,j}$ is thus nearly the same as it would have been if candidate 1 were removed from the data set entirely. The only difference is that we condition on $Y_2 < C_1(Y_1)$, which makes the test a bit more powerful than it would have been without that constraint. Again, we will reject if $\max_{j>2} p_{2,j} \leq \alpha$. Note that the hypothesis $H_{2,1}: \pi_2 \leq \pi_1$ is not relevant to $H_2^{\{2,\ldots,m\}}$, so we do not test it and we never construct a p-value $p_{2,1}$.

Rank	Candidate	Result	Votes
1.*	Hillary Clinton	60%	415
2.*	Bernie Sanders	15%	104
3.*	Joe Biden	11%	76
4.*	Don't Know	7%	48
5.	Jim Webb	3%	21
6.	Mark O'Malley	3%	21
7.	Lincoln Chafee	0%	0

Table 4.2: Results from a Quinnipiac University poll of 692 Iowa Democrats. The starred ranks represent ranks that can be verified by the stepdown procedure with confidence of 95%. That is, we can confidently declare Clinton, Sanders, Biden, and "Don't Know" to be the four most popular responses in the population from which the respondents were sampled.

With only a minor modification to the proof of Theorem 21, we can see that rejecting at step 2 reduces to checking whether the third largest value $\max_{j>2} Y_j$ is smaller than some cutoff $C_2(Y_1, Y_2)$. Applying this logic inductively, we see that at each step k, we merely apply the test of Theorem 21, with upper bound $\min_{j<k} C_j(Y_1, \dots, Y_j)$ on the remaining counts.

We can combine these single-step p-values to obtain a stepdown procedure in which we continue rejecting until one of the p-values is larger than some pre-specified α. This procedure controls the familywise error rate.

Proposition 22. *The stepdown procedure controls the familywise error rate at level α.*

Applying this stepdown procedure to Table 4.2, we find that we can declare, with 95% confidence, the most popular responses to be Hillary Clinton, Bernie Sanders, Joe Biden, and "Don't Know."

Proof. Either the empirical ranks of all candidates are identical to their true ranks, or there is some smallest rank k_0 for which the k_0th leading candidate in the poll is *not* the k_0th leading candidate in the population. Let $k_0 = \infty$ if the empirical ordering is perfect.

The probability of making an error is

$$\text{FWER} = \sum_{k=1}^{\infty} \mathbb{P}(k_0 = k)\, \mathbb{P}(\text{reject at step } k_0 \mid k_0 = k), \tag{4.24}$$

so it suffices to show, for each k, that the error probability at step k is less than α, given that k presents the first opportunity to make a mistake.

Assume without loss of generality that we are on the event

$$A_k^{1,\ldots,k-1,\{k,\ldots,m\}} = \left\{ Y_1 > \cdots > Y_{k-1} > Y_k + U_k > \max_{j>k} Y_j + U_j \right\} \cap \{\text{reject } 1, \ldots, k-1\}, \quad (4.25)$$

since we certainly would not have rejected $k-1$ times if there were any ties before step k.

Then $k_0 = k$ means that $\pi_1 > \cdots > \pi_{k-1} > \max_{j>k-1} \pi_j$, but the null hypothesis

$$H_k^{\{k,\ldots,m\}} : \; \pi_k \leq \max_{j>k} \pi_j \quad (4.26)$$

is true.

By construction the tests described in this section are valid tests of $H_k^{\{k,\ldots,m\}}$ on $A_k^{1,\ldots,k-1,\{k,\ldots,m\}}$; as a result, $\mathbb{P}(\text{reject at step } k_0 \mid k_0 = k)$ is controlled at level α. $\square$

4.2 Selective Permutation Tests: Exact Nonparametric Selective Inference

All of the examples of selective inference that we have considered thus far are parametric, but there is nothing inherently parametric about the approach outlined in Chapter 3. It is instructive to consider an example where nonparametric selective inference is still conceptually straightforward.

In certain nonparametric models, we can obtain exact tests by conditioning on sufficient statistics for the null model. The best known example of a test of this type is the permutation test. Given two samples $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} G$, consider testing the null hypothesis $H_0 : \; F = G$ against the alternative that the two distributions are not identical.

Let $Z = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ denote the vector of all $m+n$ data points, with the X values coming first. Under H_0, the distribution of Z — and, therefore, of any test statistic $T(Z)$ — is invariant to randomly permuting the $m+n$ observations, thereby assigning a random m of them to be the X sample and the other n to be the Y sample. Each permutation π results in a different value $T(\pi Z) = T(Z_{\pi(1)}, \ldots, Z_{\pi(m+n)})$. If the observed $T(Z)$ is larger than most values of $T(\pi Z)$, then we reject H_0. Permutation tests were introduced by Fisher (1935).

This procedure amounts to a conditional test that is directly analogous to the framework proposed by Lehmann and Scheffé (1955). Define $U = \{Z_i\}_{i=1}^{n+m}$, the set of observed values with the sample labels "forgotten." If H_0 is true, then U is sufficient for Z and

$$\mathcal{L}(T \mid U) = \frac{1}{|\Sigma(m+n)|} \sum_{\pi \in \Sigma(m+n)} \delta_{T(\pi Z)} \quad (4.27)$$

where $\Sigma(m+n)$ is the permutation group on $m+n$ elements. We then reject H_0 if the observed $T(Z)$ is too large to sustain the hypothesis that it was sampled from (4.28). Typically $\Sigma(m+n)$ is too large to compute the full permutation distribution, but we can always obtain an exact Monte Carlo test by sampling B independent values from the null law in (4.28) and rejecting if $T(Z)$ is one of the largest $\lfloor (1+B)\alpha \rfloor$ values.

Once we view the permutation test as a conditional test, we see that we can extend it to a selective test simply by conditioning further on some selection event A. For any event A, let $\Sigma_A = \{\pi \in \Sigma(m+n) : \pi Z \in A\}$. Then

$$\mathcal{L}(T \mid U, A) = \frac{1}{|\Sigma_A|} \sum_{\pi \in \Sigma_A} \delta_{T(\pi Z)} \tag{4.28}$$

Note that if $Z \in A$ then $|\Sigma_A| \geq 1$. We can always sample from $\mathcal{L}(T \mid U, A)$ simply by sampling random permutations and then rejecting samples for which $\pi Z \notin A$. If $\pi Z \in A$ is very unlikely, the computational cost of this approach may be prohibitive, in which case we would need to explore alternative sampling methods such as MCMC methods.

4.2.1 A Selective Two-Sample Test

Consider a randomized experiment in which m mice are given a particular treatment, n other mice receive a control treatment, and a univariate response is measured for each mouse. In such settings, it is often unclear *a priori* whether it is appropriate to use a test that makes Gaussian assumptions, or whether a nonparametric test should be used instead. Denote the treated mice's responses as $X_i \overset{\text{i.i.d.}}{\sim} F$ and the controls as $Y_i \overset{\text{i.i.d.}}{\sim} G$, and consider adaptively testing the null hypothesis $H_0 : F = G$ (no treatment effect) against the alternative that F is stochastically larger than G.

A prevailing recommendation in such cases is to apply some test $\tilde{\phi}(Z)$ of the parametric (Gaussian) model, and then use the t-test if the parametric model is not rejected, and some nonparametric permutation-based method such as Wilcoxon's rank-sum test if the test fails. This adaptive approach is a form of data snooping and requires some correction, but intuitively this sort of snooping feels harmless and so we might hope the correction will be relatively small.

To fix ideas, suppose that we employ the following two-stage procedure: let S_X^2 and S_Y^2 denote the sample variances of X and Y respectively; that is, $S_X^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \overline{X})^2$. First, we carry out a test $\tilde{\phi}$ based on the normalized residuals

$$R = \left(\frac{X_1 - \overline{X}}{S_X}, \ldots, \frac{X_m - \overline{X}}{S_X}, \; \frac{Y_1 - \overline{Y}}{S_Y}, \ldots, \frac{Y_n - \overline{Y}}{S_Y} \right).$$

If $\tilde{\phi}(R) = 0$ (if the data pass the test for Gaussianity), we perform a two-sample t-test. Otherwise, we fall back on the Wilcoxon rank-sum test.

First, notice that we require no correction if $\tilde{\phi}(R) = 0$ and we proceed with the t-test. Under the Gaussian model underlying that test,

$$X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2), \quad \text{and} \quad Y_i \overset{\text{i.i.d.}}{\sim} N(\nu, \tau^2).$$

The two-sample t-statistic is a function of $\overline{X}, \overline{Y}, S_X^2$, and S_Y^2, which are jointly independent of R. As a result there is no need to make any correction whatsoever for any selection rule based on R.

On the other hand, we will in general need a correction if $\tilde{\phi}(R) = 1$ and we perform a rank-sum test instead. We can carry this test out by trying random permutations of $Z = (X, Y)$ until we obtain B permutations for which $\tilde{\phi}(R(\pi Z)) = 1$, and then compare the Wilcoxon statistic for the observed sample to the statistic for the B permutations.

If $R(Z)$ do not appear Gaussian, then typically we expect $R(\pi Z)$ will not appear Gaussian either for most π, especially if $\tilde{\phi}$ is powerful against bimodal alternatives. Thus, accept-reject should be a reasonably computationally efficient algorithm in most cases.

4.3 Selective UMVU Estimators

As a final vignette, I consider uniform minimum-variance unbiased estimation. The risk of an estimator $\hat{\theta}(Y)$ for some parameter $\theta(F)$ is

$$R\left(\hat{\theta}(\cdot); F\right) = \mathbb{E}_F\left[L\left(\hat{\theta}(Y), \theta(F)\right)\right], \tag{4.29}$$

where $L(t, \theta)$ is the loss associated with guessing t when the truth is θ. If L is convex in t then there is often an unbiased estimator that minimizes $R\left(\hat{\theta}(\cdot); F\right)$ uniformly (i.e., for all F belonging to the model M in question), subject to the constraint that $\hat{\theta}$ be unbiased, i.e. $\mathbb{E}_F(\hat{\theta}(Y)) = \theta(F)$ for all $F \in M$. In particular, if there is a complete sufficient statistic T for the model, then

$$\hat{\theta}(T) = \mathbb{E}[\tilde{\theta}(Y) \mid T] \tag{4.30}$$

is UMVU (Rao, 1947; Lehmann and Scheffé, 1950). This device is commonly called *Rao–Blackwellization*, See Lehmann and Romano (2005) for a discussion.

By analogy to the selective Type I error rate, we can define the *selective risk* of an estimator $\hat{\theta}$ as

$$R\left(\hat{\theta}(\cdot); F\right) = \mathbb{E}_F\left[L\left(\hat{\theta}(Y), \theta(F)\right) \mid Y \in A\right], \tag{4.31}$$

where A is the event that we actually report an estimate of θ. If T is a complete sufficient statistic for the model conditional on A, and $\tilde{\theta}(Y)$ is any *selectively unbiased* estimator of θ ($\mathbb{E}_F(\tilde{\theta}(Y)) = \theta(F)$ for all $F \in M$), then

$$\hat{\theta}(T) = \mathbb{E}[\tilde{\theta}(Y) \mid T,\ Y \in A] \tag{4.32}$$

has minimum selective risk among all selectively unbiased estimators. In selective inference, this device was used by Sampson and Sill (2005) and Sill and Sampson (2009) to obtain a UMVU estimator of a treatment effect in a two-stage selective "drop-the-losers" clinical trial design.

In general, Rao–Blackwellization is a very useful tool for obtaining selective UMVU estimators. Consider the data splitting model introduced in Chapter 3, and reproduced below

Model 23 (Exponential Family with Data Splitting)**.** $(Y_1, Y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ *are independent with*

$$Y_i \sim \exp\left\{\theta T_i(y) + \zeta' U_i(y) - \psi_i(\theta, \zeta)\right\} \; f_{0,i}(y), \quad i = 1, 2, \tag{4.33}$$

with $\theta \in \mathbb{R}$.

Assume that $\tilde{\theta}$ is any unbiased estimator that is measurable with respect to Y_2. Assume without loss of generality that $\tilde{\theta}$ is a function of (T_2, U_2). Then we can obtain the UMVU estimator by Rao–Blackwellizing $\tilde{\theta}$:

$$\begin{aligned}
\hat{\theta}(T, U) &= \mathbb{E}\left[\tilde{\theta}(T_2, U_2) \;\mid\; T, U, \, Y_1 \in A_1\right] \\
&= \mathbb{E}\left[\tilde{\theta}(T - T_1(Y_1), U - U_1(Y_1)) \;\mid\; T, U, \, Y_1 \in A_1\right],
\end{aligned}$$

so computing $\hat{\theta}$ is thereby reduced to sampling from $\mathcal{L}(Y_1 \mid T, U, \; Y_1 \in A)$.

In the case of data carving with repeated univariate Gaussian samples, we can compute $\hat{\theta}$ explicitly. Assume that the data follow the distribution

$$Y_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2), \qquad i = 1, \ldots, n, \tag{4.34}$$

and that we will report an estimate of μ if and only if $\overline{Y}_1 \in A_1$, where

$$\overline{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_i, \quad \overline{Y}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n} Y_i, \quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{4.35}$$

We will Rao–Blackwellize $\tilde{\mu} = \overline{Y}_2$, an unbiased estimator with variance σ^2/n_2.

To reduce the number of parameters in the problem, assume $\sigma^2 = n$, so that there is exactly one unit of information, and let $\gamma = n_1/n$, the fraction of data used for selection. By a sufficiency reduction, we have simply

$$\overline{Y}_1 \sim N(\mu, \, \gamma^{-1}), \qquad \overline{Y}_2 \sim N\left(\mu, \, (1-\gamma)^{-1}\right), \qquad \overline{Y} = \gamma \overline{Y}_1 + (1-\gamma)\overline{Y}_2 \sim N(\mu, \, 1). \tag{4.36}$$

We will also use the identity

$$\overline{Y}_1 \mid \overline{Y} \sim N\left(\overline{Y}, \; \frac{1-\gamma}{\gamma}\right). \tag{4.37}$$

The selective UMVU estimator for μ is

$$\begin{aligned}
\hat{\mu}(\overline{Y}) &= \mathbb{E}\left[\overline{Y}_2 \mid \overline{Y},\ \overline{Y}_1 \in A_1\right] && (4.38)\\
&= \mathbb{E}\left[\frac{1}{1-\gamma}\left(\overline{Y} - \gamma\overline{Y}_1\right) \mid \overline{Y},\ \overline{Y}_1 \in A_1\right] && (4.39)\\
&= \overline{Y} - \frac{\gamma}{1-\gamma}\mathbb{E}\left[\overline{Y}_1 - \overline{Y} \mid \overline{Y},\ \overline{Y}_1 \in A_1\right] && (4.40)\\
&= \overline{Y} - \frac{\gamma}{1-\gamma}\mathbb{E}\left[\overline{Y}_1 - \overline{Y} \mid \overline{Y},\ \overline{Y}_1 - \overline{Y} \in A_1 - \overline{Y}\right] && (4.41)\\
&= \overline{Y} - \nu\,\mathbb{E}\left[Z \mid \overline{Y},\ Z \in \nu(A_1 - \overline{Y})\right], && (4.42)
\end{aligned}$$

where $Z \sim N(0,1)$ independent of $\overline{Y}$ and $\nu = \sqrt{\gamma/(1-\gamma)}$.

If $A_1 = [a,b]$ then (4.42) simplifies further, to

$$\hat{\mu}(\overline{Y}) = \overline{Y} - \nu\,\frac{\phi\left(\nu(a-\overline{Y})\right) - \phi\left(\nu(b-\overline{Y})\right)}{\Phi\left(\nu(b-\overline{Y})\right) - \Phi\left(\nu(a-\overline{Y})\right)}, \tag{4.43}$$

and similarly if $A_1 = (-\infty, a] \cup [b, +\infty)$, we have

$$\hat{\mu}(\overline{Y}) = \overline{Y} - \nu\,\frac{\phi\left(\nu(b-\overline{Y})\right) - \phi\left(\nu(a-\overline{Y})\right)}{1 - \Phi\left(\nu(b-\overline{Y})\right) + \Phi\left(\nu(a-\overline{Y})\right)}, \tag{4.44}$$

Because $\overline{Y}$ is the usual (non-selective) estimator of μ, the second term in (4.42) is an additive bias correction for selection. Taking $\gamma \to 0$, we have $\nu \to 0$ and the correction disappears. This corresponds to the case where we use very little data for selection: selection is nearly independent of the sufficient statistic $\overline{Y}$ and no correction is necessary. Taking $\gamma \to 1$ (holding out very little data) we have $\nu \to \infty$. Then, the correction tends to 0 for $\overline{Y} \in A_1$ but diverges for $\overline{Y} \notin A_1$. The interesting behavior is near the boundary, as we discuss in Section 4.3.1.

Figure 4.2 shows $\hat{\mu}(\overline{Y})$ for several values of γ, with $A_1 = (-\infty, -3] \cup [3, \infty)$. For small γ, $\hat{\mu} \approx \overline{Y}$, while for large γ the behavior is more erratic.

While it may appear from Figure 4.2 that $\hat{\mu}$ will behave very badly for $\gamma = 0.9$, we must keep in mind that $\overline{Y} \approx 0$ is a very rare event: $\overline{Y}$ is highly correlated with $\overline{Y}_1$, and $\overline{Y}_1$ is constrained to be larger than 3 in absolute value. Figure 4.3 shows $\mathrm{Var}_\mu(\hat{\mu})$ for the same four values of μ. By the Rao–Blackwell Theorem, $\mathrm{Var}(\hat{\mu}) < \mathrm{Var}(\tilde{\mu})$ (the inequality is strict because $\hat{\mu} \neq \tilde{\mu}$ almost surely).

Interestingly, the variance is smaller than 1 in a neighborhood of $\mu = 0$ for $\gamma = 0.1$ and $\gamma = 0.25$. This may seem impossible because $\overline{Y}$, the marginal UMVUE, has variance 1: thus, there is an interesting super-efficiency phenomenon at $\mu = 0$. Note that, because the selective UMVUE $\hat{\mu}$ is undefined when $Y_1 \notin A_1$, our $\hat{\mu}$ does not actually represent a "better" estimator than $\overline{Y}$, in the sense of marginal risk.

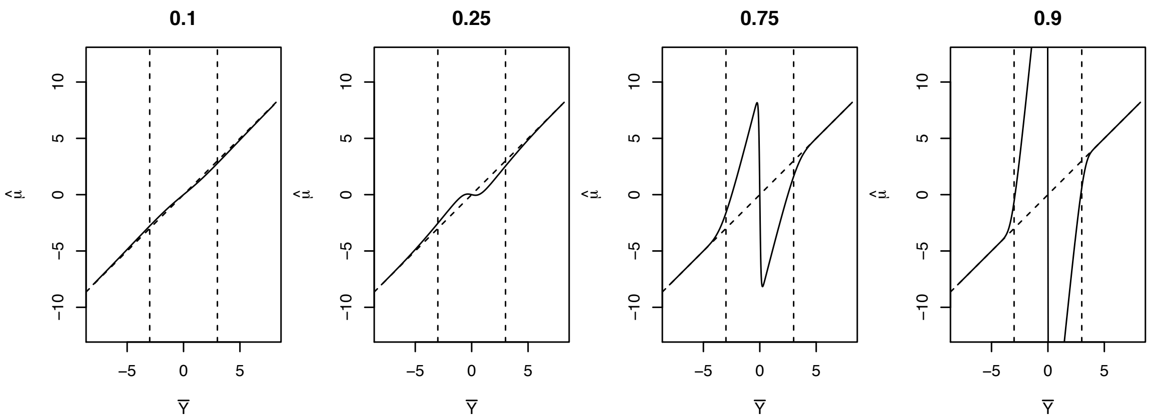


Figure 4.2: The UMVU estimator $\hat{\mu}(\overline{Y})$ as a function of $\overline{Y}$, for several values of γ. We estimate μ when $\overline{Y}_1 \in A_1 = (-\infty, -3] \cup [3, \infty)$. Outside of the threshold, $\hat{\mu} \approx \overline{Y}$. The vertical dashed lines show the threshold for $\overline{Y}_1$, while the diagonal dashed lines show the unadjusted estimator $\overline{Y}$.

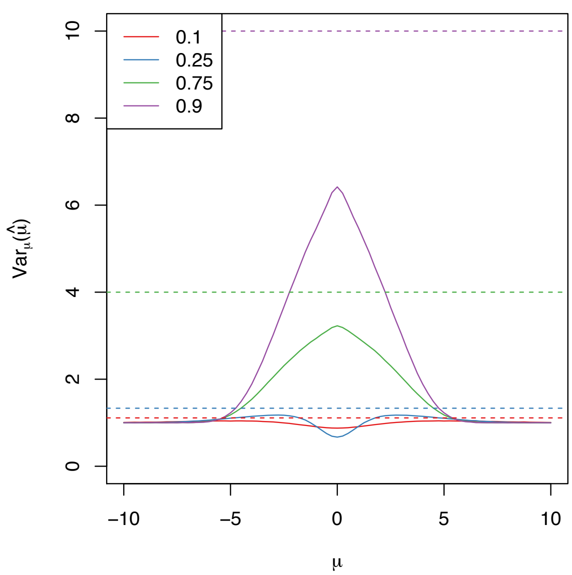


Figure 4.3: Variance of the UMVU estimator $\hat{\mu}(\overline{Y})$ for several values of γ. The dashed lines, provided for comparison, are the variances of the corresponding data-splitting estimators $\tilde{\mu} = \overline{Y}_2$, which we have Rao–Blackwellized to obtain our $\hat{\mu}$.

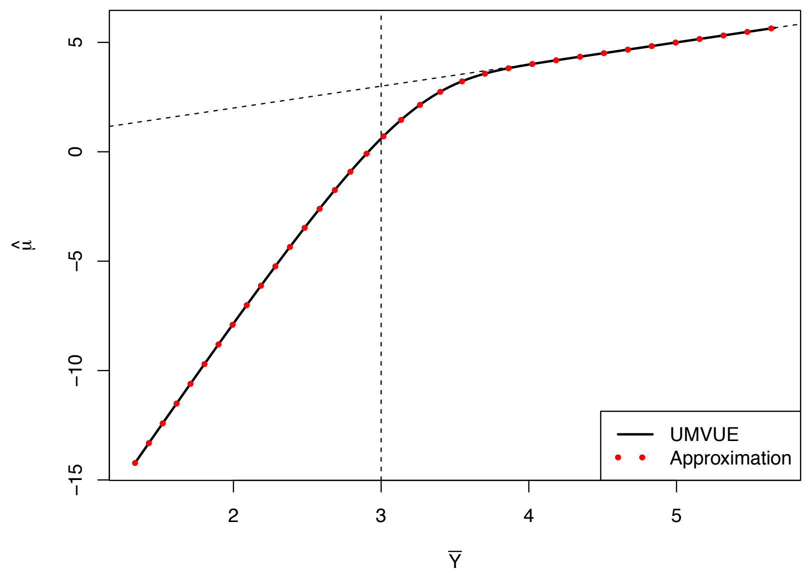


Figure 4.4: The "zoomed-in" UMVU estimator $\hat{\mu}(\overline{Y})$ as a function of $\overline{Y}$ near the threshold 3, for $\gamma = 0.9$. The approximation $\hat{\mu}(3 + \delta/\nu) \approx \overline{Y} - \nu\phi(\delta)/\Phi(\delta)$ is extremely accurate.

4.3.1 Limiting Behavior of $\hat{\mu}$ For $\gamma \to 1$

Assume that a is a boundary point of A_1 with $(a - \varepsilon, a) \subseteq A_1^C$ and $(a, a + \varepsilon) \subseteq A_1$ for some $\varepsilon > 0$. Then, taking $\nu \to \infty$,

$$\frac{\hat{\mu}(a + \delta/\nu) - \overline{Y}}{\nu} = -\mathbb{E}\left[Z \mid Z \in \nu(A_1 - a - \delta/\nu)\right] \tag{4.45}$$

$$\to -\mathbb{E}\left[Z \mid Z \geq -\delta\right] \tag{4.46}$$

$$= \frac{-\phi(\delta)}{\Phi(\delta)} \tag{4.47}$$

Figure 4.4 shows a more zoomed-in picture of $\hat{\mu}(\overline{Y})$ near 3, for $\gamma = 0.9$. The approximation is very good in this case because the two lobes of A_1 are well separated.

4.3.2 Data Carving for the Saturated Model

We can repeat a similar derivation for data carving in the saturated model with known $\sigma^2 = 1$. Similarly as before, assume that

$$Y_1 \sim N_n(\mu, \gamma^{-1} I_n), \qquad Y_2 \sim N_n(\mu, (1-\gamma)^{-1} I_n), \qquad Y = \gamma Y_1 + (1-\gamma) Y_2 \sim N_n(\mu, I_n). \tag{4.48}$$

If such a split does not exist in the data *a priori*, we can always construct one synthetically by building an n-variate Brownian bridge B_t with $B_0 = 0$ and $B_1 = Y$ and taking $Y_1 = \gamma^{-1} B_\gamma$.

For some η, assume that we report an estimate for $\theta = \eta'\mu$ whenever $Y_1 \in A_1$. Then we can Rao–Blackwellize $\tilde{\theta} = \eta' Y_2$, obtaining

$$\hat{\theta}(Y) = \mathbb{E}\left[\eta' Y_2 \mid Y,\ Y_1 \in A_1\right] \tag{4.49}$$

$$= \eta' Y - \frac{\gamma}{1-\gamma}\mathbb{E}\left[\eta'(Y_1 - Y) \mid Y,\ Y_1 - Y \in A_1 - Y\right] \tag{4.50}$$

$$= \eta' Y - \nu \|\eta\|\, \mathbb{E}\left[\eta' Z \mid Y,\ Z \in \frac{\nu}{\|\eta\|}(A_1 - Y)\right], \tag{4.51}$$

where $Z \sim N_n(0, I_n)$ independent of Y. There is a similar bias correction in (4.51) as in (4.42).

Chapter 5

Discussion

This dissertation has presented several examples of statistical problems in which it is advantageous to defer our choice of inference procedure until after some partial observation of the data.

Adapting to data typically creates some inherent bias in the subsequent procedure, and as such we must take care when designing adaptive inference procedures to adjust correctly for the bias thereby introduced.

The traditional, conservative recommendation for bias correction is to separate our analysis into two stages. The first stage is called *exploratory* analysis, in which "all bets are off:" we can do whatever we want and throw all caution — and inferential tools — to the wind. Order is restored in the second *confirmatory* stage, in which we collect fresh data and rigorously analyze it, having made all model-selection choices in the first stage. In light of recent discoveries, this approach appears a little too conservative, leading in many cases to inefficient use of data. This inefficiency is a main theme of Chapter 3, and Chapter 4 gives two examples — the nonparametric test and the rank-verification procedure — in which no adjustment, or very little adjustment is necessary. If a poll makes us wonder whether Scott Walker was truly in the lead, it would be terribly wasteful to throw away that poll and collect an entirely new data set.

Even if we are to adopt the conservative approach in certain problems and reserve a confirmatory data set for the purposes of a maximally convincing end-stage analysis, the methods discussed in Chapter 3 can still play a valuable role in improving the quality of the first-stage inference. The goal of the exploratory stage is to *learn from data* — so why should it not benefit from a century of intellectual heritage in statistics? Things we might do in the exploratory stage include examining the data set to learn what variables are important, what interactions are necessary, what models are "reasonable," and so on. But the field of statistics was invented to help analysts perform precisely these sorts of tasks! How wasteful it would be to insist on disregarding the tools of statistics in our exploratory analyses — or to apply these tools in a way which we know to be woefully invalid. Statistical tests and intervals are not just ritual procedures that sanctify scientific results once the scientist knows what she is looking for; they belong in the hurly-burly of open-ended discovery.

While much important science is still hypothesis-driven, every year brings new opportunities to carry out more data-driven investigations using enormous data sets that can support aggressive data snooping and still have a great deal of information left over to produce rigorous inference for a final analysis. Statistical methods of the future will scan through a large and complex spaces of potentially plausible models or potentially interesting findings and report back to scientists with answers to questions that the scientists never could have known they were interested in. Designing such algorithms will require a great many insights from researchers in diverse fields, but I hope that the work in my dissertation has contributed in some small part to advancing this research agenda.

Appendix A

Appendix For Chapter 2

Proof of Proposition 2 (Pointwise convergence)

Proof. Fix θ and begin by writing

$$\ell_i^\lambda = y_i(\theta - \lambda)' x_i - \log\left(1 + e^{(\theta-\lambda)' x_i}\right) \tag{A.1}$$

Let z_i^λ be the Bernoulli selection decisions, generated by comparing mutually independent $u_i \sim U(0,1)$ to the threshold $a_\lambda(x_i, y_i)$. The z_i^λ are independent conditional on λ and the data. Also, write $q_i^\lambda = z_i^\lambda \ell_i^\lambda$, so that $\widehat{Q}_\lambda(\theta) = \frac{-1}{n\overline{a}(\lambda)} \sum_{i=1}^n q_i^\lambda$.

By the Cauchy-Schwarz inequality, we have

$$|\ell_i^\lambda| \leq 1 + |\|\theta - \lambda\|\|x_i \tag{A.2}$$

Now, for $\delta > 0$ define $\Lambda_\delta = \{\lambda \,:\, \|\lambda - \lambda_\infty\| < \delta\}$. For $\lambda \in \Lambda_1$, we have

$$|q_i^\lambda| \leq m_i \triangleq 1 + (\|\theta - \lambda_\infty\| + 1)\|x_i\| \tag{A.3}$$

which is integrable by assumption. Finally let $\mathbb{E}_n$ denote an average taken over indices $i = 1, \ldots, n$, i.e. $\mathbb{E}_n f = \frac{1}{n}\sum_{i=1}^n f_i$. Then

$$\widehat{Q}_{\lambda_n}(\theta) - Q_{\lambda_\infty}(\theta) = \overline{a}(\lambda_n)^{-1}\mathbb{E}_n q^{\lambda_n} - \overline{a}(\lambda_\infty)^{-1}\mathbb{E} q^{\lambda_\infty} \tag{A.4}$$

By continuity, $\overline{a}(\lambda_n) \xrightarrow{p} \overline{a}(\lambda_\infty) > 0$. Therefore it suffices to show $\mathbb{E}_n q^{\lambda_n} \xrightarrow{p} \mathbb{E} q^{\lambda_\infty}$. Because $\mathbb{E}_n q^{\lambda_\infty} \xrightarrow{a.s.} \mathbb{E} q^{\lambda_\infty}$ by the law of large numbers, it suffices equally well to show that $\mathbb{E}_n q^{\lambda_n} - \mathbb{E}_n q^{\lambda_\infty} \xrightarrow{p} 0$.

Now fix $\varepsilon > 0$ and take K large enough that $\mathbb{E}(m\mathbf{1}_{m>K}) < \varepsilon$. For $\lambda_n \in \Lambda_1$ we have

$$\left|\mathbb{E}_n q^{\lambda_n} - \mathbb{E}_n q^{\lambda_\infty}\right| \leq \left|\mathbb{E}_n(q^{\lambda_n} - q^{\lambda_\infty})\mathbf{1}_{m\leq K}\right| + 2\mathbb{E}_n m\mathbf{1}_{m>K} \tag{A.5}$$

With probability one the second term is eventually less than 2ε. Further, for $\lambda_n \in \Lambda_\delta$, we have

$$|q_i^{\lambda_n} - q_i^{\lambda_\infty}| = \frac{1}{2}\left|(z_i^{\lambda_n} - z_i^{\lambda_\infty})(\ell_i^{\lambda_n} + \ell_i^{\lambda_\infty}) + (z_i^{\lambda_n} + z_i^{\lambda_\infty})(\ell_i^{\lambda_n} - \ell_i^{\lambda_\infty})\right| \tag{A.6}$$

$$\leq |z_i^{\lambda_n} - z_i^{\lambda_\infty}|m_i + \delta\|x_i\| \tag{A.7}$$

Now, write

$$d_i = |z_i^{\lambda_n} - z_i^{\lambda_\infty}|m_i\mathbf{1}_{m_i\leq K} \tag{A.8}$$

$z_i^{\lambda_n} \neq z_i^{\lambda_\infty}$ iff u_i lies between $a_{\lambda_n}(x_i, y_i)$ and $a_{\lambda_\infty}(x_i, y_i)$. Hence conditionally on λ_n and the data, the d_i are mutually independent non-negative random variables bounded by K with means

$$\mu_i = |a_{\lambda_n}(x_i, y_i) - a_{\lambda_\infty}(x_i, y_i)|m_i\mathbf{1}_{m_i\leq K} < \delta K^2 \tag{A.9}$$

since $\nabla_\lambda a_\lambda(x_i, y_i) \leq \|x_i\| < m_i$.

Continuing, we have

$$\left|\mathbb{E}_n(q^{\lambda_n} - q^{\lambda_\infty})\mathbf{1}_{m\leq K}\right| \leq \mathbb{E}_n(d - \mu) + \mathbb{E}_n\mu + \delta\mathbb{E}_n\|x\|\mathbf{1}_{m\leq K} \tag{A.10}$$

$$\leq \mathbb{E}_n(d - \mu) + \delta K^2 + \delta K \tag{A.11}$$

Conditioning on λ and $\{(x_i, y_i)\}$, the first term is a sum of independent zero-mean random variables that are bounded in absolute value by K. By Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} d_i - \mu_i\right| \geq \varepsilon \;\middle|\; \lambda_n, \{(x_i, y_i)\}\right) \leq 2\exp\left[-n\varepsilon^2/(2K^2)\right] \tag{A.12}$$

Since this bound is deterministic, the same applies to the unconditional probability that $\mathbb{E}_n(d-\mu)$ is large. Take $\delta = \varepsilon/(K + K^2)$. With probability tending to 1, $\lambda_n \in \Lambda_\delta$ and the event in (A.12) holds, in which case

$$|\mathbb{E}_n(q^{\lambda_n} - q^{\lambda_\infty})| \leq 4\varepsilon \tag{A.13}$$

Since ε was arbitrary, the proof is complete. □

Proof of Theorem 5 (Distribution of $\hat{\theta} - \overline{\theta}(\lambda)$)

Proof. By the Mean Value Theorem, we have for each n

$$\nabla_\theta \widehat{Q}_{\lambda_n}(\hat{\theta}_n) = \nabla_\theta \widehat{Q}_{\lambda_n}(\overline{\theta}(\lambda_n)) + \nabla^2_\theta \widehat{Q}_{\lambda_n}(\phi_n)\left(\hat{\theta}_n - \overline{\theta}(\lambda_n)\right) \tag{A.14}$$

where ϕ_n is some convex combination of $\hat{\theta}_n$ and $\overline{\theta}(\lambda_n)$. Noting that the LHS is by definition 0 and rearranging, we obtain

$$\sqrt{n}\left(\hat{\theta}_n - \overline{\theta}(\lambda_n)\right) = \nabla^2_\theta \widehat{Q}_{\lambda_n}(\phi_n)^{-1} \cdot \sqrt{n}\,\nabla_\theta \widehat{Q}_{\lambda_n}(\overline{\theta}(\lambda_n)) \tag{A.15}$$

If we can show the first factor tends in probability to $\nabla^2_\theta Q_{\theta^*}(\theta^*)^{-1}$ and the second tends in distribution to $N\left(0, \overline{a}(\theta^*)^{-1} J(\theta^*, \theta^*)\right)$, then by Slutsky's Theorem we have the desired result.

Using the Skorokhod construction define a joint probability space for λ_n such that $\lambda_n \xrightarrow{a.s.} \theta^*$. We will condition on the sequence λ_n and use a triangular array central limit theorem for the random variables

$$\begin{aligned} g_{ni} &= \frac{z_{ni}}{\overline{a}(\lambda_n)}\left(y_i - \frac{e^{(\overline{\theta}(\lambda_n)-\lambda_n)' x_i}}{1 + e^{(\overline{\theta}(\lambda_n)-\lambda_n)' x_i}}\right) x_i & \text{(A.16)} \\ &= \frac{z_{ni}}{\overline{a}(\lambda_n)} \nabla_\theta \ell(\theta - \lambda_n; x_i, y_i)\Big|_{\theta = \overline{\theta}(\lambda_n)} & \text{(A.17)} \end{aligned}$$

Because λ_n is independent of the data, $\mathbb{E}(f(g_{ni}) \,|\, \lambda_n, z_{ni} = 1) = \mathbb{E}_{\lambda_n}(f(g_{ni}))$ for any f. The triangular array CLT applies since

$$\begin{aligned} \mathbb{E}(g_{ni} \,|\, \lambda_n) &= 0 & \text{(A.18)} \\ \mathrm{Var}(g_{ni} \,|\, \lambda_n) &= \mathbb{E}\left[\mathrm{Var}(g_{ni} \,|\, \lambda_n, z_{ni}) \,|\, \lambda_n\right] & \text{(A.19)} \\ &= \mathbb{P}(z_{ni} = 1 \,|\, \lambda_n) \overline{a}(\lambda_n)^{-2} \mathrm{Var}_{\lambda_n}(\nabla_\theta \ell(\overline{\theta}(\lambda_n) - \lambda_n; x_{ni}, y_{ni})) & \text{(A.20)} \\ &= \overline{a}(\lambda_n)^{-1} J(\overline{\theta}(\lambda_n), \lambda_n) & \text{(A.21)} \\ &\xrightarrow{a.s.} \overline{a}(\theta^*)^{-1} J(\theta^*, \theta^*) & \text{(A.22)} \end{aligned}$$

Therefore, defining $S_n = n^{-1/2} \sum_{i=1}^n g_{ni}$ and $Z = N(0, a(\theta^*)^{-1} J(\theta^*, \theta^*))$, the CLT tells us $\mathbb{P}(S_n \in A \,|\, \lambda_n) \to \mathbb{P}(Z \in A)$ whenever $\lambda_n \to \theta^*$, which we assumed occurs with probability 1. By dominated convergence, we also have $\mathbb{P}(S_n \in A) \to \mathbb{P}(Z \in A)$.

Next we turn to the Hessian. We have $\hat{\theta}_n \xrightarrow{p} \theta^*$ by Theorem 4, so $\phi_n \xrightarrow{p} \theta^*$ as well. Writing

$$h_i^{\theta,\lambda} = \frac{e^{(\theta-\lambda)' x_i}}{(1 + e^{(\theta-\lambda)' x_i})^2} x_i x_i' z_i^\lambda \tag{A.23}$$

we need to show that

$$\overline{a}(\lambda_n)^{-1}\left(\mathbb{E}_n h^{\phi_n,\lambda_n}\right)^{-1} \xrightarrow{p} \overline{a}(\theta^*)^{-1}\left(\mathbb{E}h^{\theta^*,\theta^*}\right)^{-1} \tag{A.24}$$

Note that $\|h_i^{\theta,\lambda}\|_F \le \|x_i\|^2$, which is integrable; hence $\mathbb{E}_n h^{\theta^*,\theta^*} \xrightarrow{p} \mathbb{E}h^{\theta^*,\theta^*} = H(\theta^*,\theta^*) \succ 0$. Since $\overline{a}$ is continuous and strictly positive, and $\lambda_n \xrightarrow{p} \theta^*$, it suffices to show that

$$\left\|\mathbb{E}_n h^{\phi_n,\lambda_n} - \mathbb{E}_n h^{\theta^*,\theta^*}\right\|_F \xrightarrow{p} 0 \tag{A.25}$$

Note that $h_i^{\theta^*,\theta^*} = \frac{1}{4}x_i x_i'$, and define $w_{ni} = \frac{e^{(\phi_n-\lambda_n)'x_i}}{(1+e^{(\phi_n-\lambda_n)'x_i})^2}$.

Following the structure of the proof of Proposition 2, take K large enough that $\mathbb{E}\|x\|^2\mathbf{1}_{\|x\|>K} < \varepsilon$ and truncate the h_i.

$$\begin{aligned}\left\|\mathbb{E}_n h^{\phi_n,\lambda_n} - \mathbb{E}_n h^{\theta^*,\theta^*}\right\|_F &\le \left\|\mathbb{E}_n\left(h^{\phi_n,\lambda_n} - h^{\theta^*,\theta^*}\right)\mathbf{1}_{\|x\|\le K}\right\|_F \\ &\quad + \left\|\mathbb{E}_n\left(h^{\phi_n,\lambda_n} - h^{\theta^*,\theta^*}\right)\mathbf{1}_{\|x\|>K}\right\|_F \qquad \text{(A.26)} \\ &\le K^2\mathbb{E}_n\left|w_n z_n^{\lambda_n} - \frac{1}{4}z_n^{\theta^*}\right|\mathbf{1}_{\|x\|\le K} + 2\mathbb{E}_n\|x\|^2\mathbf{1}_{\|x\|>K} \qquad \text{(A.27)}\end{aligned}$$

The second term is eventually less than 2ε. Now, $w_{ni} - \frac{1}{4}$ is small, because

$$\left|\frac{d}{d\eta}\left(\frac{e^\eta}{(1+e^\eta)^2}\right)\right| = \left|\frac{e^\eta(e^\eta-1)}{(1+e^\eta)^3}\right| \le \frac{e^\eta}{(1+e^\eta)^2} \le \frac{1}{4} \tag{A.28}$$

Hence by Cauchy-Schwarz

$$\left|w_{ni} - \frac{1}{4}\right| \le \frac{1}{4}\|\phi_n - \lambda_n\|\|x_i\| \tag{A.29}$$

So on the event $\{\max\|\lambda_n - \theta^*\|, \|\phi_n - \theta^*\| < \delta\}$, we have

$$\begin{aligned}&\mathbb{E}_n\left|w_n z_n^{\lambda_n} - \frac{1}{4}z_n^{\theta^*}\right|\mathbf{1}_{\|x\|\le K} \qquad \text{(A.30)} \\ &= \frac{1}{2}\mathbb{E}_n\left|(z^{\lambda_n} - z^{\theta^*})\left(w_n + \frac{1}{4}\right) + (z^{\lambda_n} + z^{\theta^*})\left(w_n - \frac{1}{4}\right)\right|\mathbf{1}_{\|x\|\le K} \qquad \text{(A.31)} \\ &\le \mathbb{E}_n|z^{\lambda_n} - z^{\theta^*}|\mathbf{1}_{\|x\|\le K} + \delta K \qquad \text{(A.32)}\end{aligned}$$

Finally, we can bound the first term exactly as we did in the proof of Proposition 2, defining $d_i = |z_i^{\lambda_n} - z_i^{\theta^*}|K^2\mathbf{1}_{\|x_i\|\le K}$ and $\mu_i = \mathbb{E}(d_i \mid x_i, y_i, \lambda_n) \le \delta K^3$. The same argument implies $\mathbb{P}(\mathbb{E}_n(d-\mu) \ge \varepsilon) \le 2\exp[-n\varepsilon^2/(2K^4)]$, so as $n \to \infty$ we have with probability approaching 1,

$$\begin{aligned}\left\|\mathbb{E}_n\left(h^{\phi_n,\lambda_n} - h^{\theta^*,\theta^*}\right)\right\|_F &\le \mathbb{E}_n(d-\mu) + \mathbb{E}_n\mu + \delta K^3 + 2\mathbb{E}_n\|x\|^2\mathbf{1}_{\|x\|>K} \qquad \text{(A.33)} \\ &\le 3\varepsilon + 2\delta K^3 \qquad \text{(A.34)}\end{aligned}$$

so taking $\delta < \varepsilon/K^3$, the right-hand side is less than 5ε. $\square$

Appendix B

Appendix For Chapter 3

Proof of Proposition 8

Proof. For group i, let R_i be the number of true nulls selected, i.e.,

$$R_i = \left|\left\{(M, H_0) : (M, H_0) \in \widehat{\mathcal{Q}}_i(Y_i),\ F_i \in H_0 \subseteq M\right\}\right|,$$

and let V_i denote the number of false rejections. If $Z_n^V = \sum_{i=1}^n V_i$ and $Z_n^R = \sum_{i=1}^n R_i$, then we need to show $\limsup_{n\to\infty} Z_n^V / Z_n^R \leq \alpha$.

By design, $0 \leq V_i \leq R_i$ and $\mathbb{E}(V_i) \leq \alpha\, \mathbb{E}(R_i)$. As a result, $\mathbb{E}[Z_n^V]/\mathbb{E}[Z_n^R] \leq \alpha$ for every n, so we just need to show that the two sums are not far from their expectations. Because

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(R_i)}{i^2} \leq B \sum_{i=1}^{\infty} \frac{1}{i^2} < \infty,$$

we can apply Kolmogorov's strong law of large numbers to the independent but non-identical sequence $R_1, R_2, \ldots$ to obtain

$$\frac{1}{n}(Z_n^R - \mathbb{E}Z_n^R) \xrightarrow{a.s.} 0, \quad \text{so} \quad \left|\frac{Z_n^R}{\mathbb{E}Z_n^R} - 1\right| \leq \left|\frac{\delta}{n}(Z_n^R - \mathbb{E}Z_n^R)\right| \xrightarrow{a.s.} 0.$$

As for Z_n^V, we have

$$\frac{1}{n}(Z_n^V - \mathbb{E}Z_n^V) \xrightarrow{a.s.} 0, \quad \text{so} \quad \frac{Z_n^V}{\mathbb{E}Z_n^R} - \alpha \leq \frac{\delta}{n}(Z_n^V - \mathbb{E}Z_n^V) \xrightarrow{a.s.} 0;$$

in other words, $Z_n^R / \mathbb{E}Z_n^R \xrightarrow{a.s.} 1$ and $\limsup_n Z_n^V / \mathbb{E}Z_n^R \overset{\text{a.s.}}{\leq} \alpha$. $\square$

*Proof of Theorem 16

Proof. Let $T = T_1 + T_2$, $U = U_1 + U_2$. If ϕ is admissible then

$$\phi(Y) \in m\mathscr{F}(T, U, A) \subseteq m\mathscr{F}(T, U_1, U_2, A). \tag{B.1}$$

Given $(A, U_1 = u_1, U_2 = u_2)$, the density of (Y_1, Y_2) simplifies to

$$Y \sim \exp\left\{\theta\left(T_1(y_1) + T_2(y_2)\right) - \psi(\theta)\right\} \mathbf{1}_A(y_1) \prod_{i=1}^{2} f_{0,i}(y_i \mid u_i). \tag{B.2}$$

T_1 and T_2 are conditionally independent given (U_1, U_2, A) because Y_1 and Y_2 are.

If T_1 is not a function of U_1 on A then there exist $\delta > 0$ and real-valued τ for which $\min \mathbb{P}(A^-), \mathbb{P}(A^+) > 0$ where

$$A^- = \{T_1 < \tau(U_1) - \delta\} \cap A, \quad \text{and} \tag{B.3}$$
$$A^+ = \{T_1 > \tau(U_1) + \delta\} \cap A. \tag{B.4}$$

This probability depends on (θ, ζ) but if it is positive for one (θ, ζ) it is positive for all.

If ϕ is admissible among level α tests given Y_1, it must have acceptance regions of the form

$$1 - \phi(Y) = \mathbf{1}\left\{c_1(U_2) \leq T_2 \leq c_2(U_2)\right\} \tag{B.5}$$

for some (possibly infinite) cutoffs c_1, c_2. Because $\alpha > 0$, we must have at least one of the $c_i \in (a, b)$ with positive probability. Note we can always replace T_i with $-T_i$; thus, without loss of generality assume

$$\mathbb{P}\left(c_2(U_2) < b\right) > 0 \tag{B.6}$$

Define the event

$$B = \{c_2(U_2) < b\} \bigcap \{c_2(U_2) - c_1(U_2) > \varepsilon\} \tag{B.7}$$

which has positive probability for some $\varepsilon \in (0, \delta)$.

Moreover, T_2 must place some mass near the cutoff on each side, so that $\min \mathbb{P}(B^-), \mathbb{P}(B^+) > 0$ where

$$B^- = \{c_2(U_2) - \varepsilon < T_2 < c_2(U_2)\} \cap B, \quad \text{and} \tag{B.8}$$
$$B^+ = \{c_2(U_2) < T_2 < c_2(U_2) + \varepsilon\} \cap B. \tag{B.9}$$

Notice that

$$A^- \cap B^- \Rightarrow T - c_2 - \tau \;\in\; (-\delta - \varepsilon,\;\; -\delta), \tag{B.10}$$

$$A^- \cap B^+ \Rightarrow T - c_2 - \tau \;\in\; (-\delta,\;\; -\delta + \varepsilon), \tag{B.11}$$

$$A^+ \cap B^- \Rightarrow T - c_2 - \tau \;\in\; (\delta - \varepsilon,\;\; \delta), \tag{B.12}$$

$$A^+ \cap B^+ \Rightarrow T - c_2 - \tau \;\in\; (\delta,\;\; \delta + \varepsilon), \tag{B.13}$$

corresponding to four disjoint intervals of increasing value of T given (U_1, U_2). Furthermore, note that $\phi(Y)$ takes values $0, 1, 0, 1$ on the four respective events. This fact rules out the possibility that the acceptance region of ϕ has convex (U_1, U_2)-sections. $\square$

Monte Carlo Tests and Confidence Intervals: Details

Assume Z arises from a one-parameter exponential family

$$Z \sim g_\theta(z) = e^{\theta z - \psi(\theta)}\, g_0(z). \tag{B.14}$$

We wish to compute (by Monte Carlo) the UMPU two-sided rejection region for the hypothesis $H_0 : \theta = \theta_0$. Let $U \sim \text{Unif}[0, 1]$ be an auxiliary randomization variable.

Define the dictionary ordering on $[0, 1]$:

$$(z_1, u_1) \prec (z_2, u_2) \iff z_1 < z_2 \text{ or } (z_1 = z_2 \text{ and } u_1 < u_2). \tag{B.15}$$

If $\Gamma_1 = (c_1, \gamma_1)$ and $\Gamma_2 = (c_2, 1 - \gamma_2)$, then the region

$$R_{\Gamma_1,\Gamma_2} = \{(z, u) : (z, u) \prec \Gamma_1 \text{ or } (z, u) \succ \Gamma_2\} \tag{B.16}$$

implements the rejection region for the test with cutoffs c_1, c_2 and boundary randomization parameters γ_1, γ_2.

For $\Gamma_1 \prec \Gamma_2$, write

$$K_1(\Gamma_1, \Gamma_2; \theta) = \mathbb{P}_\theta(R_{\Gamma_1,\Gamma_2}) - \alpha \tag{B.17}$$

$$K_2(\Gamma_1, \Gamma_2; \theta) = \mathbb{E}_\theta(Z \,|\, (Z, U) \in R^C_{\Gamma_1,\Gamma_2}) - \mathbb{E}_\theta(Z), \tag{B.18}$$

so that the correct cutoffs Γ_i are those for which $K_1(\Gamma_1, \Gamma_2; \theta) = K_2(\Gamma_1, \Gamma_2; \theta) = 0$. For fixed θ, K_1 is decreasing in Γ_1 and increasing in Γ_2, while K_2 is increasing in both Γ_1 and Γ_2.

Let $(Z_1, W_1), (Z_2, W_2), \ldots$ be a sequence of random variables for which

$$\widehat{\mathbb{E}}_\theta^n h(Z) = \frac{\sum_{i=1}^n W_i h(Z_i) e^{\theta Z_i}}{\sum_{i=1}^n W_i e^{\theta Z_i}} \tag{B.19}$$

$$\xrightarrow{a.s.} \mathbb{E}_\theta h(Z). \tag{B.20}$$

for all integrable h. This would be true if (Z_i, W_i) are a valid i.i.d. sample or i.i.d. importance sample from g_0, or if they come from a valid Markov Chain Monte Carlo algorithm.

If $\widehat{K}_i^n$ are defined analogously to K_i for $i = 1, 2$, with $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ replaced with their importance-weighted empirical versions $\widehat{\mathbb{E}}_\theta^n$ and $\widehat{\mathbb{P}}_\theta^n$, then $\widehat{K}_i^n \xrightarrow{a.s.} K$ pointwise as $n \to \infty$, and $\widehat{K}_i^n$ satisfy the same monotonicity properties almost surely for each n. As a result, we have almost sure convergence on compacta for $(\widehat{K}_1^n, \widehat{K}_2^n)$:

$$\sup_{(\Gamma_1,\Gamma_2)\in G} \max_i \left\| \widehat{K}_i^n(\Gamma_1, \Gamma_2; \theta) - K_i(\Gamma_1, \Gamma_2; \theta) \right\| \tag{B.21}$$

for each θ, for compact $G \in (\mathbb{R} \times [0, 1])^2$.

We carry out our tests by solving for Γ_1 and Γ_2 which solve $\widehat{K}_1^n$ and $\widehat{K}_2^n$, in effect defining the UMPU tests for a one-parameter exponential family through the approximating empirical measure. Specifically, we can define

$$\widehat{\Gamma}_2(\Gamma_1; \theta) = \inf \left\{ \Gamma_2 : \widehat{K}_1^n(\Gamma_1, \Gamma_2; \theta) = 0 \right\}, \tag{B.22}$$

with $\widehat{\Gamma}_2 = \infty$ if the set is empty. That is, for a given lower cutoff we define the upper cutoff to obtain a level-α acceptance region if that is possible. Then, $\widehat{K}_2^n \left(\Gamma_1, \widehat{\Gamma}_2(\Gamma_1; \theta); \theta \right)$ is an increasing function and we can solve it using binary search. Let $\widehat{R}_\theta$ denote the rejection region so obtained.

Note that (z, u) is in the left-tail of $\widehat{R}_\theta$ if and only if $\widehat{K}_2^n \left((z, u), \widehat{\Gamma}_2((z, u)); \theta \right) < 0$. This fact, paired with an analogous test for whether (z, u) is in the right tail, gives us a quick way to carry out the test. It also allows us to quickly find the upper and lower confidence bounds for the approximating empirical family, via binary search.

Sampling for the Selective t-Test: Details

Let $C \subseteq \mathbb{R}^k$ denote a set with nonempty interior and consider the problem of integrating some integrable function $h(y)$ against the uniform probability measure on $C \cap S^{k-1}$, where S^{k-1} is the unit sphere of dimension $k - 1$, assuming the intersection is non-empty. Assume we are given an i.i.d. sequence of uniform samples $Y_1, Y_2, \ldots$ from $C \cap B^k$, where B^k is the unit ball.

Let $R \sim \frac{r^{k-1}}{k}$, so that if $Z \sim \text{Unif}(S^{k-1})$, then $Y = RZ \sim \text{Unif}(B^k)$. Let

$$W(Z) = \left(\int_0^1 \mathbf{1}\{rZ \in C\} \frac{r^{k-1}}{k} \, dr \right)^{-1} \tag{B.23}$$

We can use the Y_i for which $Z_i = Y_i / \|Y_i\| \in C$ as a sequence of importance samples with weights $W(Z_i)$, since

$$\mathbb{E}(h\,(Z)\mathbf{1}\,\{Y, Z \in C\}\, W(Z)) \tag{B.24}$$

$$= \int_{S^{k-1}} \int_0^1 h(z)\mathbf{1}\,\{z, rz \in C\}\, W(z) \frac{r^{k-1}}{k} \, dr \, dz \tag{B.25}$$

$$= \int_{S^{k-1}} h(z)\mathbf{1}\,\{z \in C\} \, dz \tag{B.26}$$

$$= \mathbb{E}\,(h(Z)\mathbf{1}\,\{Z \in C\})\,. \tag{B.27}$$

To carry out the selective t-test of $H_0 :\ \beta_j = 0$, we need to sample from

$$\mathcal{L}\left(\eta' Y \;\middle|\; \mathcal{P}_{X_{M \setminus j}} Y,\ \ \|Y\|,\ \ A\right). \tag{B.28}$$

Let $U = \mathcal{P}_{X_{M \setminus j}} Y$, and let $Q \in \mathbb{R}^{n \times (n - |M| - 1)}$ be such that $QQ' = \mathcal{P}^{\perp}_{X_{M \setminus j}}$. Then $L^2 \triangleq \|Q'Y\|^2 = \|Y\|^2 - \|U\|^2$ is fixed under the selection event. Let

$$C = \{v : \, U + Qv \in A\}, \tag{B.29}$$

so that $A_U = U + QC$, an $(n - |M| - 1)$-dimensional hyperplane intersected with A, is the event we would sample from for the selective z-test.

Under H_0, Y is uniformly distributed on

$$(U + QC) \cap \|Y\| S^{n-1} = U + Q\left(C \cap LS^{n-|M|-2}\right). \tag{B.30}$$

Assume we can resample Y^* uniformly from $A_U \cap (U + LB^{n-|M|-1})$, which is just sampling from A_U with an additional quadratic constraint. Then $V^* = Q'(Y^* - U)$ is a sample from the ball of radius L, intersected with C. We can turn V^* into an importance-weighted sample from the sphere via the scheme outlined above; then, the same importance weight suffices to turn Y^* into a sample from the selective t-test conditioning set.

Appendix C

Appendix For Chapter 4

Proof of Theorem 1 (Verifying the Winner, Gaussian Case)

Proof. Assume without loss of generality that $\overline{Y}_1 > \overline{Y}_2 > \max_{j>2} \overline{Y}_j$, with $i = 1$. First, I will show that

$$p_{1,2} = 2\left\{1 - \Phi\left(\frac{\overline{Y}_1 - \overline{Y}_2}{\sigma\sqrt{2/n}}\right)\right\}, \tag{C.1}$$

and second, I will show that $p_{1,2}$ is larger than any other $p_{1,j}$, verifying that $p_1 = p_{1,2}$.

For $j \neq 1$, write

$$Z_j = \frac{\overline{Y}_1 - \overline{Y}_j}{\sigma\sqrt{2/n}}, \tag{C.2}$$

and let

$$\mathscr{F}_{1,j} = \mathscr{F}\left(\overline{Y}_1 + \overline{Y}_j,\, (\overline{Y}_k)_{k\notin\{1,j\}}\right). \tag{C.3}$$

$Z_j \sim N(0,1)$ independently of $\mathscr{F}_{1,j}$, and $p_{1,j}$ is the upper one-sided p-value for $\mathcal{L}(Z_j \mid \mathscr{F}_{1,j}, A_1)$ under $\theta = 0$.

Define

$$V_j = \frac{\overline{Y}_1 + \overline{Y}_j}{2}, \qquad \text{and} \tag{C.4}$$

$$W_j = V_j \;\vee\; \max_{k\notin\{1,j\}} \overline{Y}_k, \tag{C.5}$$

both of which are measurable with respect to $\mathscr{F}_{1,j}$. Noting that $\overline{Y}_1 > \overline{Y}_j \iff \overline{Y}_1 > V_j$, we can rewrite A_1 as

$$\begin{aligned}
Y \in A_1 &\iff \overline{Y}_1 > \max\{\overline{Y}_2, \ldots, \overline{Y}_m\} && \text{(C.6)}\\
&\iff \overline{Y}_1 > W_j && \text{(C.7)}\\
&\iff \sigma Z_j/\sqrt{2n} + V_j = \overline{Y}_1 > W + j && \text{(C.8)}\\
&\iff Z_j > \sqrt{2n}(W_j - V_j)/\sigma. && \text{(C.9)}
\end{aligned}$$

Thus, the one-sided p-value $p_{1,j}$ is

$$p_{1,j} = \frac{1 - \Phi(Z_j)}{1 - \Phi\left(\frac{W_j - V_j}{\sigma/\sqrt{2n}}\right)} \tag{C.10}$$

For $j = 2$ we have $W_2 = V_2$ and the denominator in (C.10) is $1 - \Phi(0) = 1/2$, establishing (C.1). It remains to show that $p_{1,j} \leq p_{1,2}$ for all $j > 2$. To see why, begin by writing

$$p_{i,j} = \frac{1 - \Phi\left\{\Delta_j + (Z_j - \Delta_j)\right\}}{1 - \Phi\left(\Delta_j\right)}, \tag{C.11}$$

with

$$\Delta_j = \frac{W_j - V_j}{\sigma/\sqrt{2n}} \geq 0 = \Delta_2 \tag{C.12}$$

Moreover, because $W_j \leq W_2 = (\overline{Y}_1 + \overline{Y}_2)/2$, we have

$$\begin{aligned}
Z_j - \Delta_j &= \frac{\sqrt{2n}}{\sigma}\left(\frac{\overline{Y}_1 - \overline{Y}_j}{2} - (W_j - V_j)\right) && \text{(C.13)}\\
&= \frac{\sqrt{2n}}{\sigma}\left(\overline{Y}_1 - W_j\right) && \text{(C.14)}\\
&\geq \frac{\sqrt{2n}}{\sigma}\left(\overline{Y}_1 - W_2\right) && \text{(C.15)}\\
&= Z_2 \quad = Z_2 - \Delta_2 && \text{(C.16)}
\end{aligned}$$

Next, define the function

$$f(a, b) = \frac{1 - \Phi(a + b)}{1 - \Phi(a)}, \qquad \text{so that} \quad p_{1,j} = f(\Delta_j, Z_j - \Delta_j). \tag{C.17}$$

Then combining $Z_j - \Delta_j \geq Z_2 - \Delta_2$ and $\Delta_j \geq \Delta_2$, we have the result if we can show that $f(a, b)$ is decreasing in a and b.

But

$$\frac{\partial f(a, b)}{\partial b} = \frac{-\phi(a + b)}{1 - \Phi(a)} < 0, \tag{C.18}$$

and

$$\frac{\partial f(a,b)}{\partial a} = \frac{1}{(1-\Phi(a))^2}\{\phi(a)(1-\Phi(a+b)) - \phi(a+b)\Phi(a)\} \tag{C.19}$$

$$= \frac{\phi(a)\phi(a+b)}{(1-\Phi(a))^2}\left\{\frac{1-\Phi(a+b)}{\phi(a+b)} - \frac{1-\Phi(a)}{\phi(a)}\right\} \tag{C.20}$$

$$< 0 \tag{C.21}$$

The last inequality follows from the fact that the Mills ratio $(1-\Phi(x))/\phi(x)$ (Mills, 1926) is strictly decreasing. See Baricz (2008) for a in-depth discussion of properties of the Mills ratio. □

Proof of Theorem 2 (Verifying the Winner, Multinomial Case)

Lemma 24. *The function*

$$f(y,z,w) = \sum_{x=y}^{y+z}\binom{y+z}{x} \Big/ \sum_{x=w}^{y-1}\binom{y+z}{x}, \tag{C.22}$$

is increasing in the arguments z and w.

Proof. Increasing w only makes the denominator smaller in (C.22), so it is clear that f is increasing in w. Next using the fact that $\binom{n-1}{x} = \binom{n}{x}\frac{n-x-1}{n}$, we see that

$$f(y,z-1,w) = \sum_{x=y}^{y+z-1}\binom{y+z-1}{x} \Big/ \sum_{x=w}^{y-1}\binom{y+z-1}{x} \tag{C.23}$$

$$= \sum_{x=y}^{y+z}\binom{y+z}{x}\frac{y+z-x}{y+z} \Big/ \sum_{x=w}^{y-1}\binom{y+z}{x}\frac{y+z-x}{y+z} \tag{C.24}$$

$$< \sum_{x=y}^{y+z}\binom{y+z}{x} \Big/ \sum_{x=w}^{y-1}\binom{y+z}{x} \tag{C.25}$$

$$= f(y,z,w) \tag{C.26}$$

The inequality in (C.25) comes from the fact that $(y+z-x)/(y+z)$ is smaller than $z/(y+z)$ for every term in numerator, and larger than $z/(y+z)$ for every term in the denominator. □

Next, we prove the main result.

Proof. This proof will largely parallel the proof of Theorem 1 above. Assume without loss of generality that $Y_1 \geq Y_2 \geq \max_{j>2} Y_j$, so that $i = 1$.

Let $\mathscr{F}_{1,j} = \mathscr{F}(Y_1 + Y_j,\, (Y_k)_{k\notin\{1,j\}})$ denote the σ-algebra to condition on. Conditionally on $\mathscr{F}_{1,j}$, $Y_1 \sim \text{Binom}(Y_1+Y_j, 1/2)$ under $\theta = 0$, and $p_{1,j}$ will be the upper one-sided binomial p-value for

$\mathcal{L}(Y_1 \mid \mathscr{F}_{1,j}, A_1)$.

Suppose that $Y_1 = Y_j$. Then, given $Y_1 + Y_j$, Y_1 could not be any smaller or else we would have $Y_1 < Y_j$ and A_1 would be impossible. Thus, the realized value of Y_1 is as small as possible under $\mathcal{L}(Y_1 \mid \mathscr{F}_{1,j}, A_1)$, so the upper one-sided p-value is exactly $p_{1,j} = 1$. It follows that $p_1 = 1$ if $Y_1 = Y_2$.

Next, consider the case $Y_1 > Y_2 \geq Y_j$. Define

$$V_j = \frac{Y_1 + Y_j}{2}, \qquad \text{and} \tag{C.27}$$

$$W_j = V_j \ \vee \max_{k \notin \{1,j\}} Y_k, \tag{C.28}$$

both of which are measurable with respect to $\mathscr{F}_{1,j}$.

We can rewrite A_1 as

$$A_1 = \{Y_1 > W_j\} \ \cup \ \{Y_1 = W_j, \ \text{tie broken for } 1\}, \tag{C.29}$$

Now, consider sampling a new value Y_1 from the law $\mathcal{L}(Y_1 \mid \mathscr{F}_{1,j}, A_1)$. Let d_j denote how many Y_k would be tied with Y_1 in the case where $Y_1 = W_j$. That is,

$$d_j = \#\{k > 1 : Y_k = Y_1 \text{ if } Y_1 = W_j\}, \tag{C.30}$$

which is measurable with respect to $\mathscr{F}_{1,j}$. If W_j is not an integer, then $d_j = 0$, otherwise $d_j \geq 1$.

The probability of a tie being broken for 1 is $1/(d_j + 1)$, and the survival function of the law $\mathcal{L}(Y_1 \mid \mathscr{F}_{1,j}, A_1)$ is

$$G_{1,j}(y_1 \mid \mathscr{F}_{1,j}) \ = \ \frac{\mathbb{P}_0(A_1, Y_1 \geq y_1 \mid \mathscr{F}_{1,j})}{\mathbb{P}_0(A_1 \mid \mathscr{F}_{1,j})} \tag{C.31}$$

$$= \sum_{x \geq y_1} \binom{2V_j}{x} \Bigg/ \left\{ \binom{2V_j}{W_j} \frac{1_{d_j > 0}}{d_j + 1} + \sum_{x > W_j} \binom{2V_j}{x} \right\} \tag{C.32}$$

and $p_{1,j} = G_{1,j}(Y_1 \mid \mathscr{F}_{1,j})$.

Now consider $j = 2$. We have assumed $Y_1 > Y_2 \geq \max_{k>2} Y_k$; therefore,

$$W_2 \ = \ V_2 \ > \ Y_2 \ \geq \ \max_{k>2} Y_k \tag{C.33}$$

and consequently $d_2 = 1$ if $Y_1 + Y_2$ is even and otherwise $d_2 = 0$. In either case, the denominator in (C.32) is exactly $2^{1-(Y_1+Y_2)}$, giving

$$p_{1,2} = 2 \sum_{x=Y_i}^{Y_1+Y_2} \binom{Y_1 + Y_2}{x} / 2^{Y_1+Y_2}, \tag{C.34}$$

as claimed. It remains only to show that the other $p_{1,j}$ are no larger than $p_{1,2}$.

For all $j > 1$,

$$\frac{p_{1,j}}{1-p_{1,j}} = \sum_{x=Y_1}^{Y_1+Y_j} \binom{Y_1+Y_j}{x} \Bigg/ \left\{ \binom{Y_1+Y_j}{W_j} \frac{1_{d_j>0}}{d_j+1} + \sum_{x=\lceil W_j+1/2 \rceil}^{Y_1-1} \binom{Y_1+Y_j}{x} \right\} \tag{C.35}$$

$$\leq \sum_{x=Y_1}^{Y_1+Y_j} \binom{Y_1+Y_j}{x} \Bigg/ \sum_{x=\lceil W_j+1/2 \rceil}^{Y_1-1} \binom{Y_1+Y_j}{x} \tag{C.36}$$

$$= f(Y_1, Y_j, \lceil W_j + 1/2 \rceil). \tag{C.37}$$

If we perform a similar manipulation replacing the $1_{d_j>0}/(d_j+1)$ with 1 instead of 0, we obtain

$$f(Y_1, Y_j, \lceil W_j \rceil) \leq \frac{p_{1,j}}{1-p_{1,j}} \leq f(Y_1, Y_j, \lceil W_j + 1/2 \rceil) \tag{C.38}$$

If $Y_{1,j} = Y_{1,2}$, then by the symmetry of the problem we have $p_{1,j} = p_{1,2}$. Otherwise $Y_{1,j} < Y_{1,2}$ and $W_j \leq W_2 - 1/2$. Thus,

$$\frac{p_{1,j}}{1-p_{1,j}} \leq f(Y_1, Y_j, \lceil W_j + 1/2 \rceil) \leq f(Y_1, Y_2, \lceil W_2 \rceil) \leq \frac{p_{1,2}}{1-p_{1,2}}. \tag{C.39}$$

Because $x \mapsto x/(1-x)$ is monotone for $x \in [0, 1)$, it follows that $p_{1,j} \leq p_{1,2}$, completing the proof. $\square$

Bibliography

James A Anderson. Separate sample logistic discrimination. *Biometrika*, 59(1):19–35, 1972.

Rina Foygel Barber and Emmanuel Candes. Controlling the false discovery rate via knockoffs. *arXiv preprint arXiv:1404.5609*, 2014.

Árpád Baricz. Mills' ratio: monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340(2):1362–1370, 2008.

GA Barnard. Discussion of professor bartlett's paper. *Journal of the Royal Statistical Society*, 1963.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*, 2011.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Yoav Benjamini. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010.

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

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

James O Berger, Lawrence D Brown, and Robert L Wolpert. A unified conditional frequentist and bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, pages 1787–1807, 1994.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Julian Besag. Markov chain monte carlo for statistical inference. *Center for Statistics and the Social Sciences*, 2001.

Julian Besag and Peter Clifford. Generalized monte carlo significance tests. *Biometrika*, 76(4): 633–642, 1989.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20:161–168, 2008.

NE Breslow and KC Cain. Logistic regression for two-stage case-control data. *Biometrika*, 75(1): 11–20, 1988.

Norman E Breslow, Nicholas E Day, et al. *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.*, volume 1. Distributed for IARC by WHO, Geneva, Switzerland, 1980.

Lawrence D Brown. A contribution to kiefer's theory of conditional confidence procedures. *The Annals of Statistics*, pages 59–71, 1978.

Lawrence D Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-monograph series*, pages i–279, 1986.

C Brownie and J Kiefer. The ideas of conditional confidence in the simplest setting. *Communications in Statistics-Theory and Methods*, 6(8):691–751, 1977.

Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

Arthur Cohen and Harold B Sackrowitz. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters*, 8(3):273–278, 1989.

DR Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

AP Dawid. Selection paradoxes of bayesian inference. *Lecture Notes-Monograph Series*, pages 211–220, 1994.

Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *arXiv preprint arXiv:1408.4026*, 2014.

Bradley Efron. Tweedies formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Bradley Efron, Robert Tibshirani, et al. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24(6):2431–2461, 1996.

Thomas R Fears and Charles C Brown. Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics*, pages 955–960, 1986.

Ronald Aylmer Fisher. The design of experiments. 1935.

William Fithian and Trevor Hastie. Local case-control sampling: Efficient subsampling in imbalanced data sets. *The Annals of Statistics*, 42(5):1693–1724, 10 2014.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Jonathan J Forster, John W McDonald, and Peter WF Smith. Monte carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 445–453, 1996.

Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: unlocking the file drawer. *Science*, 2014.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2): 337–407, 2000.

Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Downloaded January*, 30:2014, 2013.

Max Grazier G'Sell, Jonathan Taylor, and Robert Tibshirani. Adaptive testing for the graphical lasso. *arXiv preprint arXiv:1307.4765*, 2013.

Naftali Harris. Visualizing lasso polytope geometry, June 2014. URL `http://www.naftaliharris.com/blog/lasso-polytope-geometry/`.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

Larry V Hedges. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, 9(1):61–85, 1984.

Larry V Hedges. Modeling publication selection effects in meta-analysis. *Statistical Science*, pages 246–255, 1992.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Adel Javanmard and Andrea Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *arXiv preprint arXiv:1301.4240*, 2013.

Karl-Heinz Jockel. Finite sample properties and asymptotic efficiency of monte carlo tests. *The annals of Statistics*, pages 336–347, 1986.

George Johnson. New truths that only one can see. *The New York Times*, 2014.

Jack Kiefer. Admissibility of conditional confidence procedures. *The Annals of Statistics*, pages 836–865, 1976.

Jack Kiefer. Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, 72(360a):789–808, 1977.

Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. *arXiv preprint arXiv:1402.5596*, 2014.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference with the lasso. *arXiv preprint arXiv:1311.6238*, 2013.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005.

Hannes Leeb and Benedikt M Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, pages 2554–2591, 2006.

Hannes Leeb and Benedikt M Pötscher. Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(02):338–376, 2008.

EL Lehmann and Joseph P Romano. *Testing statistical hypotheses*. New York:. Springer, 2005.

EL Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(3):219–236, 1955.

El L Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: the Indian Journal of Statistics*, pages 305–340, 1950.

Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso (with discussion). *The Annals of Statistics*, 42(2):413–468, 2014.

Joshua R Loftus and Jonathan E Taylor. A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*, 2014.

Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*, 2003.

Charles F Manski and T Scott Thompson. Estimation of best predictors of binary response. *Journal of Econometrics*, 40(1):97–123, 1989.

Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

Ted K Matthes and Donald R Truax. Tests of composite hypotheses for the multivariate exponential family. *The Annals of Mathematical Statistics*, pages 681–697, 1967.

Cyrus R Mehta, Nitin R Patel, and Pralay Senchaudhuri. Efficient monte carlo methods for conditional logistic regression. *Journal of The American Statistical Association*, 95(449):99–108, 2000.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488), 2009.

John P Mills. Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika*, pages 395–400, 1926.

Paul Mineiro and Nikos Karampatziakis. Loss-proportional subsampling for subsequent erm. *arXiv preprint arXiv:1306.1840*, 2013.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of MM-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, November 2012. ISSN 0883-4237. doi: 10.1214/12-STS400. URL `http://projecteuclid.org/euclid.ss/1356098555`.

Richard A Olshen. The conditional level of the ftest. *Journal of the American Statistical Association*, 68(343):692–698, 1973.

A.B. Owen. Infinitely imbalanced logistic regression. *The Journal of Machine Learning Research*, 8:761–773, 2007.

Ari Pakman and Liam Paninski. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.

Quinnipiac University Poll. May 7, 2015 - what trouble? clinton has early lock on iowa caucus, quinnipiac university poll finds; sanders, biden are only dems over 3%, May 2015a. URL `http://www.quinnipiac.edu/news-and-events/quinnipiac-university-poll/iowa/release-detail?Releas`

Quinnipiac University Poll. May 6, 2015 - walker in front of pack in iowa gop caucus, quinnipiac university poll finds; at 5%, bush is a distant seventh, May 2015b. URL `http://www.quinnipiac.edu/news-and-events/quinnipiac-university-poll/iowa/release-detail?Releas`

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

C Radhakrishna Rao. Minimum variance and the estimation of several parameters. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 43, pages 280–283. Cambridge Univ Press, 1947.

Camilo Rivera and Guenther Walther. Optimal detection of a jump in the intensity of a poisson process or in a density with likelihood ratio statistics. *Scandinavian Journal of Statistics*, 40(4): 752–769, 2013.

JD Rosenblatt and Yoav Benjamini. Selective correlations; not voodoo. *NeuroImage*, 103:401–410, 2014.

Allan R Sampson and Michael W Sill. Drop-the-losers design: Normal case. *Biometrical Journal*, 47(3):257–268, 2005.

AJ Scott and CJ Wild. Fitting logistic regression models in stratified case-control studies. *Biometrics*, pages 497–510, 1991.

Alastair Scott and Chris Wild. On the robustness of weighted methods for fitting models to case–control data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2): 207–219, 2002.

Alastair J Scott and CJ Wild. Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 170–182, 1986.

Michael W Sill and Allan R Sampson. Drop-the-losers design: Binomial case. *Computational statistics & data analysis*, 53(3):586–595, 2009.

Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, 2014.

Xiaoying Tian and Jonathan Taylor. Asymptotics of selective inference. *arXiv preprint arXiv:1501.03588*, 2015.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

John W Tukey. Quick and dirty methods in statistics. part ii. simple analyses for standard designs. In *Proceedings of the 5th Annual Convention. American Society for Quality Control*, pages 189–197, 1951.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Steve Webb, James Caverlee, and Calton Pu. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *Proceedings of the Third Conference on Email and Anti-Spam (CEAS)*, 2006.

Clarice R Weinberg and Sholom Wacholder. The design and analysis of case-control studies with biased sampling. *Biometrics*, pages 963–975, 1990.

Asaf Weinstein, William Fithian, and Yoav Benjamini. Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108(501): 165–176, 2013.

Gary M Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004.

Yu Xie and Charles F Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302, 1989.

Daniel Yekutieli. Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):515–541, 2012.

Ed Yong. Replication studies: Bad copy. *Nature*, 485(7398):298–300, 2012.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Hua Zhong and Ross L Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9(4):621–634, 2008.

Sebastian Zöllner and Jonathan K Pritchard. Overcoming the winners curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615, 2007.