

A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results

by

Beau Coker

Department of Statistical Science
Duke University

Date: _____

Approved:

Sayan Mukherjee, Supervisor

Cynthia Rudin

Alexander Volfovsky

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2018

ABSTRACT

A Theory of Statistical Inference for Ensuring the Robustness
of Scientific Results

by

Beau Coker

Department of Statistical Science
Duke University

Date: _____

Approved:

Sayan Mukherjee, Supervisor

Cynthia Rudin

Alexander Volfovsky

An abstract of a thesis submitted in partial fulfillment of the requirements for
the degree of Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2018

Copyright © 2018 by Beau Coker
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Inference is the process of using facts we know to learn about facts we do not know. A theory of inference gives assumptions necessary to get from the former to the latter, along with a definition for and summary of the resulting uncertainty. Any one theory of inference is neither right nor wrong, but merely an axiom that may or may not be useful. Each of the many diverse theories of inference can be valuable for certain applications. However, no existing theory of inference addresses the tendency to choose, from the range of plausible data analysis specifications consistent with prior evidence, those that inadvertently favor one's own hypotheses. Since the biases from these choices are a growing concern across scientific fields, and in a sense the reason the scientific community was invented in the first place, we introduce a new theory of inference designed to address this critical problem. From this theory, we derive "hacking intervals," which are the range of summary statistic one may obtain given a class of possible endogenous manipulations of the data. They make no appeal to hypothetical data sets drawn from imaginary superpopulations. A scientific result with a small hacking interval is more robust to researcher manipulation than one with a larger interval, and is often easier to interpret than a classic confidence interval. Hacking intervals turn out to be equivalent to classical confidence intervals under the linear regression model, and are equivalent to profile likelihood confidence intervals under certain other conditions, which means they may sometimes provide a more intuitive and potentially more useful interpretation of classical intervals.

Contents

Abstract	iv
List of Tables	vii
List of Figures	viii
List of Abbreviations and Symbols	ix
Acknowledgements	xii
1 Introduction	1
1.1 A New Theory of Inference	4
2 Prescriptively Constrained Hacking Intervals	10
2.1 Examples	12
2.1.1 k -NN	12
2.1.2 Matching	14
2.1.3 Adding a New Feature	16
3 Tethered Hacking Intervals	23
3.1 Examples	25
3.1.1 SVM	25
3.1.2 Kernel Regression	28
3.1.3 PCA Feature Selection	29
3.2 Connection to Profile Likelihood	32
3.3 Generalization Bound	38

4	Tethered Hacking Intervals for Linear Regression	42
4.1	Scenario 1: Average Treatment Effect	43
4.1.1	Relationship to classical confidence intervals.	45
4.1.2	Non-classical Choices for θ	46
4.1.3	Combining with data variance	47
4.1.4	Illustration	48
4.2	Scenario 2: Individual Treatment Effect	49
4.2.1	Illustration	53
5	Application: Recidivism Prediction	55
5.1	Prescriptively Constrained Example: Adding a New Feature	58
5.2	Tethered Example: SVM	59
6	Related works and Discussion	67
6.1	Conclusion	74
A	Proofs	75
	Bibliography	87

List of Tables

3.1	Illustration of generalization bound.	41
4.1	Hacking interval for linear regression.	49

List of Figures

2.1	Hacking intervals for k -NN.	14
3.1	Hacking interval for SVM.	27
3.2	Level sets of the Mahalanobis distance that defines the kernel regression function	29
3.3	PCA feature selection hacking.	32
4.1	Average treatment effect prediction curves.	50
4.2	Individual treatment effect prediction curves.	54
5.1	Hacking interval for odds ratio.	59
5.2	SVM hacking intervals by COMPAS score.	63
5.3	SVM hacking intervals for all individuals with a COMPAS score of 3.	64
5.4	SVM hacking intervals for all individuals with a COMPAS score of 8.	64
5.5	SVM hacking intervals by race.	66

List of Abbreviations and Symbols

Symbols

n	number of observations.
p	number of covariates (sometimes there are more than p , for example if we separate out a treatment variable from p confounders).
$i \in \{1, 2, \dots, n\}$	indexes observations.
$k \in \{1, 2, \dots, n\}$	also indexes observations if necessary.
$j \in \{1, 2, \dots, p\}$	indexes features.
x	scalar covariate.
x_i	i th observation x .
\mathbf{x}	$p \times 1$ vector of covariates.
\mathbf{x}_i	i th observation of \mathbf{x} .
$\mathbf{x}_{\cdot j}$	j th component of \mathbf{x} .
\mathbf{x}_{ij}	j th component of i th observation of \mathbf{x} .
$\mathbf{X} = (\mathbf{x}_1^T, \dots, \mathbf{x}_n^T)^T$	$n \times p$ matrix of covariates.
$x^{(\text{new})}$	Scalar covariate not included in the training data.
$x_i^{(\text{new})}$	i th observation of $x^{(\text{new})}$.
$\mathbf{x}^{(\text{new})}$	$p \times 1$ vector of covariates not include in the training data.
$\mathbf{x}_i^{(\text{new})}$	i th observation of $\mathbf{x}^{(\text{new})}$.
$\mathbf{x}_{\cdot j}^{(\text{new})}$	j th component of $\mathbf{x}^{(\text{new})}$.
$\mathbf{x}_{ij}^{(\text{new})}$	j th component of i th observation of $\mathbf{x}^{(\text{new})}$.

$\mathbf{X}^{(\text{new})} = (\mathbf{x}_1^{(\text{new})T}, \dots, \mathbf{x}_m^{(\text{new})T})^T$ $m \times p$ matrix of covariates.

y scalar outcome.

y_i i th observation of a scalar outcome.

$\mathbf{Y} = (y_1, \dots, y_n)^T$ $n \times 1$ vector of scalar outcomes.

\mathcal{Y} Space of \mathbf{Y} .

$z = [x, y]$ tuple.

$z_i = [x_i, y_i]$ tuple.

$\mathbf{z} = [\mathbf{x}, y]$ tuple.

$\mathbf{z}_i = [\mathbf{x}_i, y_i]$ tuple.

$\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ tuple.

\mathcal{Z} tuple.

$\phi : \mathcal{Z} \rightarrow \mathcal{Z}$ Data adjustment function. We require that ϕ can be separated into two functions $\phi_{\mathbf{X}}$ and $\phi_{\mathbf{Y}}$ such that for any $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ we have $\phi(\mathbf{Z}) = [\phi_{\mathbf{X}}(\mathbf{X}), \phi_{\mathbf{Y}}(\mathbf{Y})]$.

Φ space of ϕ .

ψ hyperparameter(s). Can be a vector or scalar. We can also write ψ_d to denote “default” hyperparameter(s).

Ψ space of ψ .

$f : \mathcal{X} \rightarrow \mathcal{Y}$ prediction function.

f_{β} prediction function parameterized by β (β is a placeholder. Any letter will do).

\mathcal{F} space of f .

\mathcal{F}_{ψ} space of f parameterized by hyperparameters ψ .

$L : \mathcal{Z} \times \mathcal{F} \times \Psi \rightarrow \mathbb{R}$ Loss function. If there are no hyperparameters, we omit their notation by letting $L : \mathcal{Z} \times \mathcal{F} \rightarrow \mathbb{R}$

$t(\cdot)$ real-valued summary statistic. Depending on the context, we may have t map from any collection of \mathcal{Z} , \mathcal{X} , and \mathcal{F} (in that order), where \mathcal{X} is for predictions on new points $\mathbf{X}^{(\text{new})}$ (or $\mathbf{x}^{(\text{new})}$).

Generally, bold/unbold and lowercase/uppercase mean the following:

- | | |
|-------------------|---|
| Bold, lowercase | $p \times 1$ vector (<i>e.g.</i> , \mathbf{x}). |
| Bold, uppercase | $n \times p$ matrix (<i>e.g.</i> , \mathbf{X}). |
| Unbold, lowercase | scalar (<i>e.g.</i> , x). |
| Unbold, uppercase | random variable (<i>e.g.</i> , X). |

Acknowledgements

I am grateful for my advisor, Cynthia Rudin, for support and kindness; our collaborator, Gary King, for insightful debates; and Aaron Fisher, for detailed comments, discussions, and assistance with the proofs. I also want to thank Faustine Li, Gilad Amitai, and Lisa Lebovici for listening to me talk about this project far too much and kindly providing their thoughts anyway.

Introduction

The numerous choices in even “best practice” data analysis procedures lead to high levels of unmeasured and unreported uncertainty in research publications. These choices include, among others, variable selection and transformations, data subsetting, identification and elimination of outliers, functional forms, distributional assumptions, priors, estimators, nonparametric preprocessing (such as matching), and procedures to control for unmeasured confounders (such as difference in differences or instrumental variables). (See Wicherts et al. 2016 for an attempt to enumerate a complete list.) Classical statistical inference conditions on whichever choices the analyst makes and focuses on uncertainty induced by observing only a sample of data. This is an uncertainty *between* datasets, one observed and one imagined from an assumed superpopulation. However, *within* a single dataset, the often considerable variation *across potential “plausible” analyses* leads to an often wide range of empirical estimates that can overwhelm uncertainty induced by sampling. Neglecting the sensitivity to analyst choices may be part of the reason why scientific studies may fail to replicate at expected rates. We thus propose that researchers (and readers) ask a simple question that gets to the heart of whether or not a quantitative conclusion

can be trusted: “Could a different reasonable researcher analyzing the same data come to a different conclusion?” This paper provides a quantitative framework for answering this question.

We use the term *hacking* to describe an earnest researcher working hard to choose appropriately among many data analysis choices. Although the term is sometimes used to describe dishonest manipulation of results, we use it solely (in the positive sense of a “hackathon”) to refer to honest scientists genuinely trying to get the right answer. For a given model class and loss function, a *hacking interval* is the smallest and largest value of a summary statistic (*e.g.*, a coefficient in a regression) that can be achieved over a set of constraints for which the researcher, readers, and the scientific community would like robustness. It quantifies the extent to which a different, also reasonable, analyst could come to different conclusions. Researchers that report hacking intervals are more transparent about the evidence available to support their hypotheses. Hacking intervals are designed to reveal the information any research publication should provide to make it more difficult for researchers, and readers of their work, to be fooled. A major benefit of hacking intervals is that they are easy to understand, interpret, and teach. They can be taught at the same time, before, or even without reference to classical confidence intervals or any other theory of inference. They require no knowledge of probability.

The hacking intervals we propose come in two varieties. *Prescriptively constrained* hacking intervals allow for an explicit definition the analysis choices that reasonable researchers will make and find the range of a summary statistic over these choices. They are useful when one wishes to limit which analysis choices are valid. The second type, *tethered hacking intervals*, avoid the explicit enumeration of analysis choices and require only that the predictive model chosen by the researcher has a small enough loss on the observed data. Each type of hacking interval is a consequence of the defined set of researcher constraints. In a maximum likelihood scenario, tethered

hacking intervals are mathematically equivalent to profile likelihood confidence intervals (as shown in Section 3.2). Our work therefore provides a new interpretation of profile likelihood confidence intervals that requires no understanding of probability.

Quantifying the potential impact of hacking is especially — but not only — important if researchers are biased towards a conclusion. Standard data analysis procedures leave researchers in a situation that meets all the conditions social-psychologists have shown lead to biased choices: In the presence of high levels of discretion, many analysis choices, little objective way to know which is best, and access to the estimates each choice results in, even honest, hard working, earnest researchers may wind up inadvertently biasing results toward their favorite hypotheses (Gilbert, 1998; Banaji and Greenwald, 2013; Kahneman, 2011). If a researcher (or reader) is concerned that analysis choices were only chosen because they yielded results consistent with the bias of the researcher, a hacking interval informs them of the degree to which this can matter. A small hacking interval says that *any* researcher within defined constraints, but biased towards a conclusion or not, could have only a small impact on the result. While conducting and aggregating many studies, where each is based on different researchers making their own analysis choices, is a natural solution for understanding the impact of analysis choices, this process can be costly, time consuming, and, if the biases are concentrated towards one conclusion, the space of researcher hacking may be left under-explored. Hacking intervals explore this space systematically with computational and mathematical tools. This exploration can also provide insight into the state of researcher bias: if the hacking interval is large, and the range of conclusions from many published studies is small, then this suggests researchers may be biased towards a specific conclusion. There are existing formalized procedures that aim to mitigate the impact of bias, for example pre-registration, “best practices,” ignorance-based approaches (*e.g.*, double blind experiments or journal reviews), or requiring replication datasets (King, 1995), but the problem of a different reason-

able researcher being able to reach a different conclusion would still exist even if researchers were completely unbiased, so these procedures address only part of the problem. The sheer number of possible analysis choices leaves unchecked uncertainty in scientific results unless the space of choices is rigorously defined and explored.

Throughout this work, we offer examples and illustrations of hacking intervals, in the context of k -nearest neighbors (kNN), matching, variable selection, support vector machines, and, in more detail, linear regression. We present an analysis of recidivism prediction. We find evidence that the COMPAS score, which is a commonly used risk scoring system used in bail and parole decisions, may often be calculated incorrectly. This can lead in practice to dangerous criminals being released, as well as low risk individuals being unfairly sentenced or denied bail or parole. We find cases where the COMPAS score is not within the hacking interval, meaning that no reasonable model (by our definition of reasonable, and according to our dataset) would agree with the risk assessment provided by COMPAS.

1.1 A New Theory of Inference

Each of the diverse collection of theories of inference is united by a common goal — to understand if an observed effect is robust over counterfactual worlds imagined to have occurred — but can be distinguished by which counterfactual worlds are of interest. For example, p-values consider if an effect is robust to counterfactual *data* from a superpopulation. Fisher’s exact p-values fix the data and consider if an effect is robust to counterfactual *treatment assignments* from every possible assignment. Causal sensitivity analysis considers if an effect is robust to counterfactual *unmeasured confounders* from a defined set (Ding and VanderWeele, 2016; Liu et al., 2013). There are many other theories of inference, and they can be probabilistic or non-

probabilistic,¹ and deductive or inductive.² In part because the sum of uncertainties from different forms of inference is usually too large to be able to conclude almost anything at all, current practice is to present, in every applied publication, intervals or another summary from *only one* chosen form of uncertainty stemming from a single theory of inference, and to temporarily assume away other forms of uncertainty. Another reason is that one theory of inference may be of significantly more use than another depending on the context. For example, despite studies showing a strong correlation between smoking and lung cancer, the question of whether or not smoking caused lung cancer was unsettled in the 1950s because of the possibility of an unmeasured confounding genetic variable. The Cornfield Conditions assumed that the causal effect were zero and deduced properties of the unmeasured confounding genetic variable, properties that were deemed biologically infeasible (Cornfield et al., 1959). This theory of inference was instrumental in taking the scientific community from facts that were known (smoking correlates with lung cancer and there is an approximate biological limit on how much a genetic variable and smoking could be related) to a fact that was unknown (smoking causes lung cancer).

Motivated by a growing mistrust of scientific results on the basis of researcher degrees of freedom, our theory of inference considers if an effect is robust to counterfactual *researchers* making analysis choices from a defined set. Like many other theory of inferences, it is completely deductive. We hypothesize which choices a reasonable researcher will make, either by explicitly constraining their choices or by allowing a tolerance in the loss function, and deduce the range of effects — the

¹ For example, a p-value gives the probability of an equal or more extreme result under the null hypothesis, while causal sensitivity analysis gives a non-probabilistic bound on an attribute of the unmeasured confounder.

² For example, Bayesian inference can be thought of as inductive (in the sense that newly observed data updates prior beliefs) or deductive (in the sense that posterior predictive checks can be used to falsify a model). Fisher’s p-values, Neyman-Pearson hypothesis tests, and Neyman’s confidence intervals are typically thought of as deductive (Gelman and Shalizi, 2013).

hacking interval — that they can achieve subject to these constraints. A hacking interval can be used to judge whether or not the observed effect is robust to researcher choices. It can be probabilistic or, in a practical sense, non-probabilistic. If one hypothesizes that reasonable researchers make choices within the specified constraints with probability $1 - \delta$, for any $\delta \in [0, 1]$, then the effect they report will be in the hacking interval with probability $1 - \delta$. In the special case that $\delta = 0$, this is practically equivalent to saying the interval is non-probabilistic (although technically it holds with probability 1). While a hacking interval is designed to estimate the range of conclusions that *reasonable* researchers could report, *any* researcher acting within the constraints will report results within the hacking interval.

An alternative to our approach is a greatly expanded Bayesian model (perhaps via robust Bayes combined with Bayesian model averaging) that formally specifies all possible modeling decisions, enables the choice of priors or classes of priors and the many associated hyperprior values over this large set, and computes classes of posteriors as a result. We do not recommend this approach because it adds numerous researcher choices for which prior information is rarely available, and thus may exacerbate the very problem of hacking we seek to address. Our preferred theory of inference explicitly gives up the goal of full posterior distributions or classes of posterior distributions. In their place, it seeks the more limited goal of an interval as a summary of uncertainty. What we get in return for limiting our goal to intervals is clearer ways of specifying assumptions and researcher discretion reflected in easy-to-interpret results.

Hacking intervals and classical confidence intervals (which we take to mean either a frequentist confidence interval or a Bayesian credible interval) each convey important but different components of the strength of evidence in observed data, but hacking intervals are a more natural starting point. When a researcher observes a numerical result of scientific interest, they want to quantify how strongly

the observed data supports the result. As with p-values, classical confidence intervals quantify the robustness of the result to sampling variability. If the result could be reversed under different, likely to occur datasets, this implies the result is not robust.³ Hacking intervals quantify the robustness of the result to researcher choices. If the result could be reversed under different but also reasonable analysis choices, then the result is not robust. A large interval of either type degrades the robustness of a result. However, as we argue now, hacking intervals are a more natural starting point. Consider the following advantages of hacking intervals as compared to classical confidence intervals:

1. *They are easier to explain.* Hacking intervals are the solutions to an optimization problem that requires no understanding of probability. In contrast, despite repeated clarifications of their interpretation (Wasserstein and Lazar, 2016), frequentist confidence intervals are routinely misinterpreted and misexplained, to the point where they have been banned (Trafimow and Marks, 2015). We discuss problems with classical confidence intervals and teaching hacking intervals in Chapter 6.
2. *They do not require estimation of a probability distribution.*
3. *They are more natural when the superpopulation imagined in classical inference does not exist.* We discuss this point in the next paragraph.
4. *They could be wider than classical confidence intervals.*

Given these relative advantages of hacking intervals, and that the analyst simply wants to find patterns in the data that are robust, we advise the analyst to calculate

³ Frequentist confidence intervals and Bayesian credible intervals have precise definitions. We mean this statement only in regards to their spirit. In a Bayesian context, the observed data is viewed in the context of prior information.

a hacking interval first and then decide if calculating a classical interval is worth the trouble or is even possible in the context of the problem.

In regards to the third point in the above list, consider problems from the political science fields of comparative politics and international relations, where country level or time-series cross-sectional data are available. The cause of (for instance) civil wars is incredibly important for understanding the past, and we may like to determine patterns that characterize political situations that have led to civil wars. One might hypothesize that countries with many people in poverty, having many young men (ages, say 15-25 years), with neighboring countries in civil war, and with no strong government could be prone to have civil wars. The data are observational, randomization is impossible for events that happened in the past, and no more relevant data may ever be collected. In these cases, researchers often use some type of regression to estimate causal relationships. If the researcher learns that a variable has a large coefficient in a regression for predicting aspects relating to a civil war, then she may use confidence intervals to determine whether this result is robust — robust across possible model specifications. She may use traditional inference notation (confidence intervals, null hypotheses), but since the idea of a superpopulation may not even make sense, the null hypothesis does not exist, and she may be better off computing a hacking interval. The researcher is not interested in constructing an imaginary superpopulation of world systems with different countries; we really only care about the current countries. The question is whether the researcher can claim a robust empirical relationship, or whether she demonstrated only that it was merely possible to find one of a million model specifications that was consistent with her causal hypothesis. In this case, the researcher may wish to call the measurement uncertainty a hacking interval, rather than a confidence interval. However, to do this requires a specific mathematical framework for this interval, which we provide here.

This work is done in collaboration with my advisor, Professor Cynthia Rudin (Duke University), and Professor Gary King (Harvard University).

Prescriptively Constrained Hacking Intervals

Denote $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{n \times p}$ as covariates, $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^{n \times 1}$ as outcomes, $\mathbf{Z} \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as datasets, and $f : \mathcal{X} \rightarrow \mathcal{Y}$ as prediction functions from a class \mathcal{F}_ψ , where $\psi \in \Psi$ denotes a vector of hyperparameters. For example, \mathcal{F}_ψ could be the space of all binary decision trees of maximum depth ψ . Let $L : \mathcal{Z} \times \mathcal{F}_\psi \times \Psi \rightarrow \mathbb{R}$ be a loss or regularized loss function, and $t : \mathcal{Z} \times \mathcal{X} \times \mathcal{F}_\psi \rightarrow \mathbb{R}$ be a summary statistic of interest. The loss function may or may not depend on the hyperparameters ψ , so if not we omit writing ψ . Similarly, while the summary statistic must depend on f , it may or may not depend on Z , which is the observed training data, or $\mathbf{X}^{(\text{new})}$, which are covariates for observations the model is not trained on. Depending on the context we may omit writing \mathbf{Z} and/or $\mathbf{X}^{(\text{new})}$ in the definition of t . For hyperparameters ψ , training data $\mathbf{Z} \in \mathcal{Z}$, and, optionally, test data $\mathbf{X}^{(\text{new})}$, we assume the user finds f^* that minimizes the loss $L(\mathbf{Z}, f^*, \psi)$ and then computes the summary statistic $t^* := t(\mathbf{Z}, \mathbf{X}^{(\text{new})}, f^*)$ based on this result. For instance, in linear regression, the user finds the linear function $f^*(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}^*$ that minimizes the quadratic loss on the dataset $\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$. Possible summary statistics include an estimate of a single

regression coefficient $t(f^*) = \beta_j^*$, a goodness-of-fit measurement of f^* on \mathbf{Z} , or a prediction $t(\mathbf{x}^{(\text{new})}, f^*) = f^*(\mathbf{x}^{(\text{new})}) = \mathbf{x}^{(\text{new})T} \boldsymbol{\beta}^*$ on a single test observation $\mathbf{x}^{(\text{new})} \in \mathcal{X}$. Our interest is in the range of summary statistics t^* that could be achieved if the researcher were permitted to adjust the dataset \mathbf{Z} and hyperparameters ψ .

The approach to this problem is to explicitly constrain data adjustments $\phi : \mathcal{Z} \rightarrow \mathcal{Z}$ to a set Φ and hyperparameters to a set Ψ . We assume that ϕ can be separated into two functions $\phi_{\mathbf{X}}$ and $\phi_{\mathbf{Y}}$ such that for any $\mathbf{Z} = [\mathbf{X}, \mathbf{Y}]$ we have $\phi(\mathbf{Z}) = [\phi_{\mathbf{X}}(\mathbf{X}), \phi_{\mathbf{Y}}(\mathbf{Y})]$. We then wish to calculate the minimum and maximum summary statistics over these two sets, Ψ and Φ , which constrain researcher choices:

$$a_{\min} := \min_{\psi \in \Psi, \phi \in \Phi} t \left(\phi(\mathbf{Z}), \phi_{\mathbf{X}}(\mathbf{X}^{(\text{new})}), \underset{f \in \mathcal{F}_{\psi}}{\operatorname{argmin}} L(\phi(\mathbf{Z}), f, \psi) \right), \quad (2.1)$$

$$a_{\max} := \max_{\psi \in \Psi, \phi \in \Phi} t \left(\phi(\mathbf{Z}), \phi_{\mathbf{X}}(\mathbf{X}^{(\text{new})}), \underset{f \in \mathcal{F}_{\psi}}{\operatorname{argmin}} L(\phi(\mathbf{Z}), f, \psi) \right). \quad (2.2)$$

Notice hyperparameters ψ impact a_{\min} and a_{\max} through \mathcal{F}_{ψ} (*e.g.*, by controlling the max depth of decision trees) as well as through the loss directly (*e.g.*, by controlling the regularization). In other words, ψ is assumed to contain all relevant hyperparameters, both to determine hard constraints on the function class, as well as soft constraints through regularization. We define the interval $[a_{\min}, a_{\max}]$ as the *prescriptively constrained hacking interval*. For example, if the summary statistic t is a prediction of f on a new point $\mathbf{x}^{(\text{new})}$, then Equations (2.1) and (2.2) can be written as:

$$a_{\min} = \min_{\psi \in \Psi, \phi \in \Phi} f(\phi_{\mathbf{X}}(\mathbf{x}^{(\text{new})})) \quad \text{s.t.} \quad f \in \underset{\mathcal{F}_{\psi}}{\operatorname{argmin}} L(\phi(\mathbf{Z}), f, \psi)$$

$$a_{\max} = \max_{\psi \in \Psi, \phi \in \Phi} f(\phi_{\mathbf{X}}(\mathbf{x}^{(\text{new})})) \quad \text{s.t.} \quad f \in \underset{\mathcal{F}_{\psi}}{\operatorname{argmin}} L(\phi(\mathbf{Z}), f, \psi).$$

While a prescriptively constrained hacking interval is designed for a single loss function, one could include in ψ a hyperparameter that switches between more than

loss function, allowing for specification of the loss function to be among the researcher choices.

2.1 Examples

We present examples of prescriptively constrained hacking intervals for k-nearest neighbor (where the researcher chooses k within a reasonable range), matching for causal inference (where the researcher chooses a matching algorithm), and adding a new feature (where the researcher adds a new feature constrained by its relationship to existing data).

2.1.1 k -NN

This is a simple example. Suppose we have observed data $\mathbf{Z} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ and we wish to predict on a new point $\mathbf{x}^{(\text{new})}$ by averaging nearby observations. In this example we will keep the data \mathbf{Z} fixed but allow the researcher to choose the hyperparameter k , the number of nearest neighbors over which to average. To construct a simple prescriptively constrained hacking interval, we define a subset of reasonable hyperparameter choices Ψ , which in this case we can write as a range $[k_{\min}, k_{\max}]$, and find the range of predictions on a new point $\mathbf{x}^{(\text{new})}$ subject to the constraint that $k \in [k_{\min}, k_{\max}]$:

$$\max/\min_{k \in [k_{\min}, k_{\max}]} \frac{1}{k} \sum_j \eta_{i^{(\text{new})}j} y_j$$

where η_{ij} tells us if \mathbf{x}_j is within the k th nearest neighbor of \mathbf{x}_i . This range of predictions is the prescriptively constrained hacking interval. Notice that there is no loss function. The hyperparameter k allows for only one function in the function space \mathcal{F}_k (namely, the one that averages over the k nearest neighbors). To solve this problem, we evaluate the nearest neighbor average for each k within the range $\Psi = [k_{\min}, k_{\max}]$.

Prescriptively constrained hacking intervals require that the researcher justify to readers their choice of Ψ , and we recommend that this discussion be briefly included in every paper. This approach therefore does not remove all research discretion, and arguably not all hacking, but it changes the nature of scholarly papers from a justification of a single specification to one where they justify a definition for the range of reasonable specifications.

One possibility for this choice is to center $\Psi = [k_{min}, k_{max}]$ around a fixed value and calculate the hacking interval over $[k_{min}, k_{max}]$ constraints of increasing width. For example, find $k^* \in [1, n - 1]$ that minimizes the training error and then find the hacking interval over $\Psi(m) := [k^* - m, k^* + m]$ for each $m = 1, 2, 3, \dots, m_{max}$, where m_{max} handles the edge cases so that $\Psi \subset [0, n]$. Figure (2.1.1) shows the results of such a procedure for a dataset in two dimensions and $\mathbf{x}^{(new)} = (0.5, 0.5)$. We find that $k^* = 5$ minimizes the training error and the resulting prediction on $\mathbf{x}^{(new)}$ is 0.6. However, if the researcher is allowed to pick any k in $[k^* - 2, k^* + 2] = [3, 5]$, for example, then the prediction ranges from .57 to .70. This is the hacking interval for $m = 2$. Displaying the hacking interval as a function of m illustrates the sensitivity of the hacking interval to the freedom given to the researcher.

Another choice for the range of k could be to use prior information of acceptable past researcher choices. We might choose the range of k large enough to include the smallest and largest k 's used in k -nearest neighbor in any article in the last 5 years in that field. In practice, that interval may actually be the smallest and largest values that would not be objected to by reviewers.

Other researcher choices for k -NN that we did not consider in this example could include the distance function or the weighting of the k nearest neighbors. The use of k -NN as the predictive function class could also be considered a researcher degree of freedom. We could use a binary hyperparameter to switch between k -NN and any other regression algorithm.

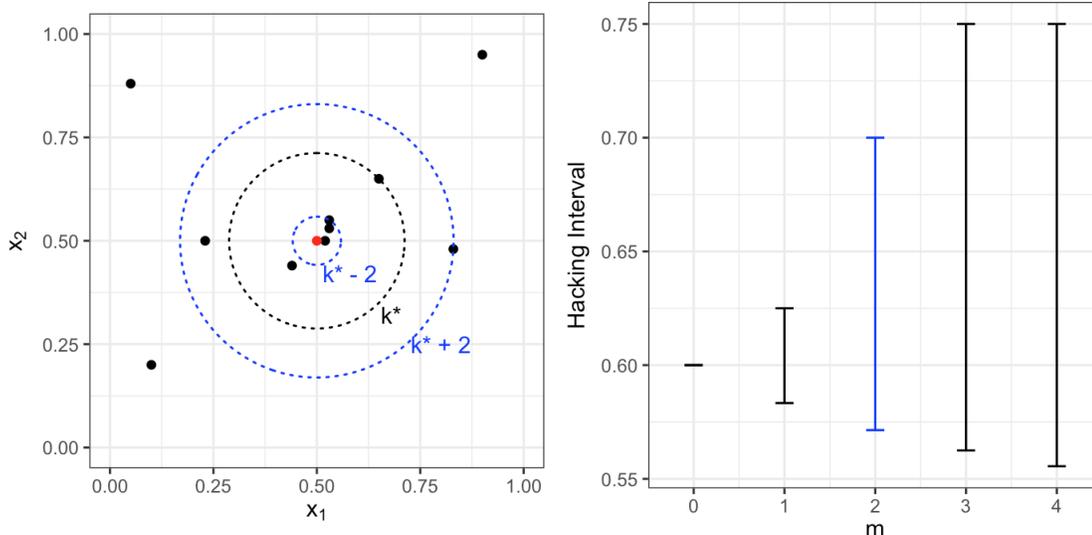


FIGURE 2.1: *Left:* Observed data with distance to $k = 3, 5$, and 7 nearest neighbors highlighted, where $k^* = 5$. *Right:* Hacking intervals as function of the hyperparameters space width m . $m = 2$ corresponds to a hacking interval over researcher choice $[k^* - 2, k^* + 2] = [3, 7]$.

2.1.2 Matching

Matching methods in causal inference are a key example of where a hacking interval can be useful in strengthening or weakening conclusions made from data. An analyst’s choice of matching algorithm may have a huge impact on the conclusion, and this impact could be much larger than the uncertainty due to randomness in the data. It is entirely possible that two well-meaning analysts, given the same data, would choose two different matching algorithms for treatment and control units and reach two entirely different conclusions. It does not make sense to model the set of choices that analysts would make when choosing between possible matching algorithms. The analyst’s choice could be arbitrary; it depends on the algorithms available at the time, the popularity of these algorithms among scientists, the order of the data in the database, and other choices that are totally separate from the ground truth treatment effect. The analyst does not choose matches from a uniform distribution among reasonable matching possibilities (simple examples where one

good match assignment is clearly better than another can show why this would be an unreasonable assumption). In the work of Noor-E-Alam and Rudin (2015), the authors take a hacking interval approach by specifying that an *unreasonable* match assignment would have least one matched treatment/control pair whose covariates are far away from each other. The converse of this set consists of *reasonable* match assignments. Noor-E-Alam and Rudin (2015) compute a tethered hacking interval, in that they compute the range of treatment effect estimates corresponding to *all* reasonable match assignments. For some datasets, they find that all reasonable match assignments yield the same conclusion. In other cases, the range of treatment effects corresponding to reasonable matches is very large. Their technique uses mixed-integer programming, so that they can determine the maximum and minimum test statistics over all match assignments without having to enumerate them.

If we find that *any* reasonable match assignment yields the same conclusion, then it is a much stronger result than saying that *one* reasonable match assignment (as is typically considered) yields a particular conclusion. Consider the experiment done by Noor-E-Alam and Rudin (2015) on the GLOW (Global Longitudinal study of Osteoporosis in Women) data Hosmer et al. (2013). The goal was to determine whether smoking causes bone fractures using McNemar’s test. Their experiments showed that no matter which reasonable analyst creates the matched pairs, the conclusion is the same: smoking causes bone fractures. Figure ??, reproduced from Noor-E-Alam and Rudin (2015), shows the tethered hacking intervals of the P -value of McNemar’s test for different numbers of matched groups. The figure shows that if any reasonable analyst constructs match assignments with 30 or more “unmatched pairs (where treatment and control outcomes differ), the worst (highest) P -value possible they can achieve is 0.003, meaning the result will be significant at the 0.05 level no matter which matching method the analyst uses. This example demonstrates how hacking intervals can sometimes strengthen scientific conclusions. (In other

cases, hacking intervals can weaken conclusions).

2.1.3 Adding a New Feature

The addition of a new feature to a given collection of features, possibly from existing features (such as an interaction term) or from new data, is a data adjustment that can impact the conclusions about prediction models. A prescriptively constrained hacking interval in this context is the range of a summary statistic that can be achieved over all of the possible choices made by the researcher about new features, subject to explicit constraints on those choices. If the researcher is given the freedom to choose each of n feature values (one for each observation), then solving this problem requires optimizing over a potentially large space, since there are n choices made by the researcher. Fortunately, it may only be necessary to specify a small number of attributes about the new feature to calculate its impact on the summary statistic. The prescriptively constrained hacking interval would then be an optimization problem over a smaller space of attributes, subject to explicit constraints on the attributes.

In a causal inference setting, where the researcher observes a treatment feature among other possibly confounding features, sensitivity analysis deals with this exact problem. The goal is to find the impact on a causal effect (the summary statistic) of an unmeasured confounder u (the new feature).¹ To do this one needs to choose a value for several attributes about the unmeasured confounder. There are a number of approaches to this problem that require different attributes of u to be chosen (see Liu et al., 2013, for a review), but generally only a few attributes are required: its distribution, its relationship to the outcome, and its relationship to the treatment. In applications of causal sensitivity analysis, a researcher will often display the adjusted

¹ Alternatively, the goal may be to assume the unmeasured confounder reduced the causal effect to zero and see what this would imply about the unmeasured confounder.

causal effect for each of a few choices of these attributes. If we explicitly define a range of choices for each attribute, then the maximum and minimum causal effect over these ranges is a prescriptively constrained hacking interval.

The motivation of causal sensitivity analysis and prescriptively constrained hacking are different. In causal sensitivity analysis, u exists but is unmeasured by the researcher. Constraints on the values of the attributes of u are based on what we believe is scientifically reasonable. In prescriptively constrained hacking, u is created by the researcher. Constraints on the values of the attributes of u are based on what we believe is a reasonable amount of researcher freedom.

We now define our approach in more detail. Let $\mathbf{Y} = (y_1, \dots, y_n)^T \in \{-1, 1\}^n$ be a $n \times 1$ matrix of observed binary outcomes, $\mathbf{X} = (\mathbf{x}_1^T, \dots, \mathbf{x}_n^T)^T$ an $n \times p$ matrix of observed covariates, and $\mathbf{W} = (w_1, \dots, w_n)^T \in \{0, 1\}^n$ an $n \times 1$ matrix of observed binary covariates. In a causal inference setting, \mathbf{W} is the treatment. The researcher degrees of freedom constitute the choice of an additional binary covariate $\mathbf{U} = (u_1, \dots, u_n)^T \in \{0, 1\}^n$. This is equivalent to the choice of a data adjustment function $\phi : [\mathbf{Y}, \mathbf{X}, \mathbf{W}] \mapsto [\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}]$. Once ϕ has been chosen we assume the researcher finds a model f from a set of linear functions \mathcal{F} of the form:

$$f([\mathbf{x}, w, u]) = \beta_0 + \mathbf{x}^T \boldsymbol{\beta}_{\mathbf{x}} + \beta_w w + \beta_u u$$

by minimizing the logistic loss function:

$$L([\mathbf{X}, \mathbf{W}, \mathbf{U}], f) = \sum_{i=1}^n \log(1 + e^{-y_i f([\mathbf{x}_i, w_i, u_i])}),$$

Notice this is equivalent to maximizing the likelihood under the model:

$$\text{logit } \Pr(Y = 1 \mid \mathbf{x}, w, u) = f([\mathbf{x}, w, u]),$$

where Y is the random variable corresponding to the observed y . In other words, the researcher performs logistic regression. (For simplicity, the objective is fixed and

there are no user choices except to add the extra feature.) We further assume the researcher is interested in the odds ratio of y and w controlling for covariates \mathbf{x} and u :

$$OR_{yw|\mathbf{x}u} := \frac{\Pr(Y = 1 \mid \mathbf{x}, w = 1, u)}{\Pr(Y = 1 \mid \mathbf{x}, w = 0, u)},$$

so we set the test statistic to be $t(f) := e^{\beta w}$. The steps followed by the researcher can be summarized as follows:

- *1a*: Choose a data adjustment $\phi \in \Phi$ (we discuss Φ below).
- *1b*: Find $\hat{f}([\mathbf{x}, w, u]) = \hat{\beta}_0 + \mathbf{x}^T \hat{\beta}_{\mathbf{x}} + \hat{\beta}_w w + \hat{\beta}_u u$ that minimizes the logistic loss on the adjusted data, $(\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}) = \phi(\mathbf{Y}, \mathbf{X}, \mathbf{W})$.
- *1c*: Calculate the summary statistic $\widehat{OR}_{yw|\mathbf{x}u} = t(\hat{f}) = e^{\hat{\beta} w}$.

The prescriptively constrained hacking interval is the maximum and minimum values of $t(\hat{f})$ that can be achieved over all of the possible researcher choices of $\phi \in \Phi$. There are no hyperparameters in this example.

Interestingly, we can calculate $\widehat{OR}_{yw|\mathbf{x}u}$ without knowing the researcher-created covariate u exactly. We need only know the relationship of u to the binary covariate w , specified by $p_0 := p(U \mid w = 0)$ and $p_1 := \Pr(U \mid w = 1)$ (where U is the random variable corresponding to u), and the relationship of u to the binary outcome y , specified by $OR_{yu} := \Pr(Y = 1 \mid u = 1)/\Pr(Y = 1 \mid u = 0)$. When p_0 , p_1 , and OR_{yu} are known, Lin et al. (1998) show² that we can derive the odds ratio adjusting for \mathbf{x} and u , $\widehat{OR}_{yw|\mathbf{x}u} = t(\hat{f})$, from the odds ratio that only adjusts for \mathbf{x} , $\widehat{OR}_{yw|\mathbf{x}}$, by the following formula:

$$\widehat{OR}_{yw|\mathbf{x}u} = \frac{1}{AF} \widehat{OR}_{yw|\mathbf{x}} \tag{2.3}$$

² Lin et al. (1998) show this result exactly for log-linear regression, but they argue it should hold approximately for logistic regression.

where

$$AF = \frac{(OR_{yu} - 1)p_1 + 1}{(OR_{yu} - 1)p_0 + 1}. \quad (2.4)$$

We write OR_{yu} rather than \widehat{OR}_{yu} because the former quantity is the true odds ratio, not one estimated from the data.

Since $\widehat{OR}_{yw|x}$ can be estimated from the observed data, Equations (2.3) and (2.4) imply the impact of the researcher choice of u is completely summarized by p_1 , p_0 , and OR_{yu} , since they determine AF . Conversely, if we knew the data adjustment ϕ we could estimate p_1 , p_0 , and OR_{yu} , calling the estimates \hat{p}_1 , \hat{p}_0 , and \widehat{OR}_{yu} , respectively, from the adjusted data. Steps 1a-1c are therefore equivalent to Steps 2a-2d defined by:

- 2a: Calculate $\widehat{OR}_{yw|x}$.
- 2b: Choose a data adjustment $\phi \in \Phi$.
- 2c: Calculate \widehat{OR}_{yu} , \hat{p}_1 , and \hat{p}_0 using the adjusted data $[\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}] = \phi([\mathbf{Y}, \mathbf{X}, \mathbf{W}])$.
- 2d: Calculate the summary statistic $\widehat{OR}_{yw|x} = \frac{1}{\widehat{AF}} \widehat{OR}_{yw|x}$, where \widehat{AF} analogous to Equation (2.4) but depends on the estimated quantities \widehat{OR}_{yu} , \hat{p}_1 , and \hat{p}_0 :

$$\widehat{AF} = \frac{(OR_{yu} - 1)\hat{p}_1 + 1}{(\widehat{OR}_{yu} - 1)\hat{p}_0 + 1}. \quad (2.5)$$

Notice that the researcher's choice of a data adjustment ϕ implies a value for u and the three attributes about u — \widehat{OR}_{yu} , \hat{p}_1 , and \hat{p}_0 — but it is through these three attributes only that ϕ impacts the summary statistic. If we instead allow the researcher to choose only the three attributes, we can find the impact on the summary

statistic without ever knowing u . We just need to define the space of allowable data adjustments Φ in terms of its impact on these three attributes:

$$\Phi := \left\{ \phi : (\mathbf{Y}, \mathbf{X}, \mathbf{W}) \mapsto (\mathbf{Y}, \mathbf{X}, \mathbf{W}, \mathbf{U}) \mid \widehat{OR}_{yu} \in [a, b], |\hat{p}_1 - \hat{p}_0| \leq c, \hat{p}_0 > d \right\},$$

for constants a , b , and $c < d$ (the reason for these exact constraints will be clear later). Then, Steps 2a-2d can be replaced with Steps 3a-3c defined by:

- 3a: Calculate $\widehat{OR}_{yw|\mathbf{x}}$.
- 3b: Choose OR_{yu} , p_0 , and p_1 such that $OR_{yu} \in [a, b]$, $|p_1 - p_0| \leq c$, and $p_0 \geq d$.
- 3c: Calculate the summary statistic $\widehat{OR}_{yw|\mathbf{x},u} = \frac{1}{AF} \widehat{OR}_{yw|\mathbf{x}}$, where AF depends on OR_{yu} , p_0 , and p_1 .

For equivalent choice of constraints, the maximum and minimum values of $\widehat{OR}_{yw|\mathbf{x},u}$ that could be achieved by any of the three sequences of Steps (1a-1c, 2a-2d, and 3a-3c) are all equal. We can think of finding the maximum and minimum values of $\widehat{OR}_{yw|\mathbf{x},u}$ for each of the three sequences as the three following optimization problems (each solved for the maximum and minimum):

$$\text{Steps 1a-1c: } \max/\min_{\phi \in \Phi} \{OR_{yw|\mathbf{x}u}\} \tag{2.6}$$

$$\text{Steps 2a-2d: } \begin{cases} \max/\min & \left\{ \frac{1}{AF} \widehat{OR}_{yw|\mathbf{x}} \right\} \\ \phi \text{ s.t. } & \begin{cases} \widehat{OR}_{yu} \in [a, b] \\ |\hat{p}_1 - \hat{p}_0| \leq c \\ \hat{p}_0 > d \end{cases} \end{cases} \tag{2.7}$$

$$\text{Steps 3a-3c: } \max/\min_{OR_{yu} \in [a, b]} \left\{ \frac{1}{AF} \widehat{OR}_{yw|\mathbf{x}} \right\}. \tag{2.8}$$

$$|p_1 - p_0| \leq c$$

$$p_0 \geq d$$

Optimization Problem (2.8) will prove the most useful as it does not require knowledge of u . Since $\widehat{OR}_{yw|x}$ is estimated from the observed data, solving Optimization Problem (2.8) is equivalent to solving for the maximum and minimum values of AF subject to the same constraints and dividing $\widehat{OR}_{yw|x}$ by each value. Using Equation (2.4) for AF , we find the maximum and minimum values of AF by solving the following optimization problem:

$$\max_{OR_{yw}, p_1, p_0} / \min \frac{(OR_{ty} - 1)p_1 + 1}{(OR_{yu} - 1)p_0 + 1} \quad \text{s.t.} \quad \begin{cases} OR_{yu} \in [a, b] \\ |p_1 - p_0| \leq c \\ p_0 \geq d \end{cases}, \quad (2.9)$$

Dividing $\widehat{OR}_{yw|x}$ by the maximum and minimum values given by Optimization Problem (2.9) gives the minimum and maximum values, respectively, of $OR_{yw|x}$, which define the hacking interval in this case.

We can solve Equation (2.9) for the case where OR_{yu} is fixed greater than one (implying $\Pr(Y = 1 | u = 1) > \Pr(Y = 1 | u = 0)$). In this case, the maximization problem in Equation (2.9) (*i.e.*, the hacking interval upper bound) becomes:

$$\begin{aligned} \max_{|p_1 - p_0| \leq c} \frac{(OR_{yu} - 1)p_1 + 1}{(OR_{yu} - 1)p_0 + 1} &= \max_{p_0 \geq d} \frac{(OR_{yu} - 1)(p_0 + c) + 1}{(OR_{yu} - 1)p_0 + 1} \\ p_0 &\geq d \\ &= \max_{p_0 \geq d} 1 + \frac{(OR_{yu} - 1)c}{(OR_{yu} - 1)p_0 + 1}, \end{aligned}$$

while the minimization problem (*i.e.*, the hacking interval lower bound) becomes:

$$\begin{aligned} \min_{|p_1 - p_0| \leq c} \frac{(OR_{yu} - 1)p_1 + 1}{(OR_{yu} - 1)p_0 + 1} &= \min_{p_0 \geq d} \frac{(OR_{yu} - 1)(p_0 - c) + 1}{(OR_{yu} - 1)p_0 + 1} \\ p_0 &\geq d \\ &= \min_{p_0 \geq d} 1 - \frac{(OR_{yu} - 1)c}{(OR_{yu} - 1)p_0 + 1}. \end{aligned}$$

In each case, the optimum occurs at $p_0 = d$. Therefore, Equation (2.9) can be solved when OR_{yu} is fixed greater than one. We apply this result in Section 5.1.

This section shows how results from causal sensitivity analysis can be leveraged to solve problems where the researcher is permitted to hack a new feature. Here, we have been in a non-causal inference setting of logistic regression modeling. In Section 5.1 we apply these results to a recidivism dataset.

Tethered Hacking Intervals

In prescriptively constrained hacking intervals, discussed in Chapter 2, we optimize over a data adjustment function ϕ and hyperparameters ψ constrained to be in sets Φ and Ψ , respectively. An advantage of this approach is that we can clearly define acceptable researcher adjustments. A disadvantage is that the possible adjustments may be difficult to enumerate or optimize over efficiently. One way to circumvent this requirement is to allow *any* choice of ψ and ϕ so long as the loss using the unadjusted data Z and a set of default hyperparameters ψ_d is not too large. The *tethered hacking interval* is the minimum and maximum summary statistic under this constraint. In other words, it is given by the interval $[b_{\min}, b_{\max}]$,

$$b_{\min} := \min_{f \in \mathcal{F}_{\psi_d}} t(\mathbf{Z}, \mathbf{X}^{(\text{new})}, f) \quad \text{s.t.} \quad L(\mathbf{Z}, f, \psi_d) \leq \theta, \quad (3.1)$$

$$b_{\max} := \max_{f \in \mathcal{F}_{\psi_d}} t(\mathbf{Z}, \mathbf{X}^{(\text{new})}, f) \quad \text{s.t.} \quad L(\mathbf{Z}, f, \psi_d) \leq \theta. \quad (3.2)$$

given a fixed, chosen value of θ .

For example, suppose \mathcal{F} is the set of constant functions $f(x) = \lambda$, $t(\mathbf{Z}, \mathbf{X}^{(\text{new})}, f) = \lambda$ is the parameter λ that defines f , and L is the quadratic loss for each of n observations in dataset \mathbf{Z} . There are no hyperparameters ψ_d so we suppress their notation.

Then Equations (3.1) and (3.2) become:

$$b_{\min} = \min_{\lambda} \lambda \quad \text{s.t.} \quad \sum_{i=1}^n (\lambda - y_i)^2 \leq \theta$$

$$b_{\max} = \max_{\lambda} \lambda \quad \text{s.t.} \quad \sum_{i=1}^n (\lambda - y_i)^2 \leq \theta.$$

For another example, if \mathcal{F} is the set of linear functions $f(x) = \lambda_0 + \lambda_1 x$, $t(Z, f) = \lambda_0 + \lambda_1 \mathbf{x}^{(\text{new})}$ is a prediction of f on a new point $\mathbf{x}^{(\text{new})}$, and L is the same quadratic loss, then Equations (3.1) and (3.2) become:

$$b_{\min} = \min_{\lambda_0, \lambda_1} \lambda_0 + \lambda_1 \mathbf{x}^{(\text{new})} \quad \text{s.t.} \quad \sum_{i=1}^n (\lambda_0 + \lambda_1 x_i - y_i)^2 \leq \theta$$

$$b_{\max} = \max_{\lambda_0, \lambda_1} \lambda_0 + \lambda_1 \mathbf{x}^{(\text{new})} \quad \text{s.t.} \quad \sum_{i=1}^n (\lambda_0 + \lambda_1 x_i - y_i)^2 \leq \theta.$$

In general, when the summary statistic is a prediction on a new point $\mathbf{x}^{(\text{new})}$, Equations (3.1) and (3.2) become:

$$b_{\min} = \min_{\mathcal{F}_{\psi_d}} f(\mathbf{x}^{(\text{new})}) \quad \text{s.t.} \quad L(\mathbf{Z}, f, \psi_d) \leq \theta$$

$$b_{\max} = \max_{\mathcal{F}_{\psi_d}} f(\mathbf{x}^{(\text{new})}) \quad \text{s.t.} \quad L(\mathbf{Z}, f, \psi_d) \leq \theta.$$

The interpretation of a tethered hacking interval is that a researcher could have hacked the data or adjusted the hyperparameters to obtain any value of the test statistic in the interval. In other words, for each point $b' \in [b_{\min}, b_{\max}]$ there could exist a data adjustment function ϕ' and a set of hyperparameters ψ' such that b' is the output of the summary statistic when applied to the minimum loss predictive model f using ϕ' and ψ' . That is,

$$b' = t \left(\phi'(\mathbf{Z}), \phi'_{\mathbf{X}}(\mathbf{X}^{(\text{new})}), \underset{f \in \mathcal{F}_{\psi'}}{\operatorname{argmin}} L(\phi'(\mathbf{Z}), f, \psi') \right).$$

This interpretation describes how results are hacked in practice. First, a researcher chooses how to adjust a dataset and which hyperparameters are appropriate, and then summarizes the resulting best function in a class. The purpose of a tethered hacking interval is to bound the results of this procedure by specifying a single constraint on the loss function.

The work of Fisher et al. (2017) introduces a measure of variable importance for a class of prediction functions based on the Rashomon set. While their “empirical model class reliance” measure of variable importance could be viewed as a type of hacking interval, their goal is to study the population version of this quantity, in order to judge variable importance for the population.

Notice two things about tethered hacking intervals. First, when the loss function corresponds to a likelihood function, tethered hacking intervals are equivalent to profile likelihood confidence intervals for an appropriate choice of the loss threshold θ . We explore this in Section 3.2. Second, as with prescriptively constrained hacking intervals, a tethered hacking interval is a statement about the degree to which summaries of a single observed dataset could be hacked by a researcher. It does not require an assumption about a true data generating procedure. In Section 3.3 we make such an assumption about the true data generating procedure and derive an appropriate generalization bound in order to unite traditional inference with our new inference paradigm. Before that, we discuss examples of tethered hacking intervals for SVM, kernel regression, and variable selection using PCA.

3.1 Examples

3.1.1 SVM

In this section we demonstrate how hacking intervals can be calculated in the context of support vector machines (SVM). Recall that SVM is trained by minimizing the

following loss function:

$$L(\mathbf{Z}, f, \psi_d) = \frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 + \psi_d \sum_{i=1}^n (1 - y_i f(\mathbf{x}_i))_+,$$

where $f(\mathbf{x}) := \boldsymbol{\lambda}^T \mathbf{x} + \lambda_0$ is the distance of \mathbf{x} to the separating hyperplane and $\psi_d \in \mathbb{R}^+$ is a hyperparameter that controls the degree of regularization. Here, we define the summary statistic as the distance of a new point $\mathbf{x}^{(\text{new})}$ to the separating hyperplane. The hacking interval is then given by:

$$\max_{\boldsymbol{\lambda}, \lambda_0} / \min_{\boldsymbol{\lambda}, \lambda_0} \boldsymbol{\lambda}^T \mathbf{x}^{(\text{new})} + \lambda_0 \quad \text{s.t.} \quad \frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 + \psi_d \sum_{i=1}^n (1 - y_i (\boldsymbol{\lambda}^T \mathbf{x}_i + \lambda_0))_+ \leq \theta, \quad (3.3)$$

where θ controls the loss tolerance. Figure 3.4 illustrates this problem.

For simplicity we can write both the min and max problems from Equation (3.3) as a single minimization problem that depends on the choice of a binary variable $s \in \{-1, +1\}$ ($s = 1$ for min, $s = -1$ for max). If we also write the loss constraint in terms of slack variables $\boldsymbol{\xi}$ then Equation (3.3) becomes:

$$\min_{\boldsymbol{\lambda}, \lambda_0, \boldsymbol{\xi}} s \boldsymbol{\lambda}^T \mathbf{x}^{(\text{new})} + s \lambda_0 \quad \text{s.t.} \quad \begin{cases} y_i (\boldsymbol{\lambda}^T \mathbf{x}_i + \lambda_0) \geq 1 - \xi_i, \quad \forall i \\ \xi_i \geq 0, \quad \forall i \\ \frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 + \psi_d \sum_{i=1}^n \xi_i \leq \theta. \end{cases} \quad (3.4)$$

This is a convex optimization problem. The objective is linear. The first two constraints are the same as in non-separable SVM and are linear. The last constraint is the sum of a norm (always convex) and a linear function in $\boldsymbol{\xi}$, so it is convex; also, it is the objective function for non-separable SVM. Therefore, we can apply the KKT conditions to obtain the dual problem. The following proposition shows the result.

Proposition 1 (Hacking Intervals for SVM). *The solution to Optimization Problem (3.4) is given by*

$$\boldsymbol{\lambda}^* = \frac{1}{\beta^*} \left(-s \mathbf{x}^{(\text{new})} + \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \right)$$

and

$$\lambda_0^* = y_{i_{sv}} - \boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}},$$

where i_{sv} is such that $0 < \alpha_{i_{sv}}^* < \beta^* \psi_d$ and the optimal dual variables $(\boldsymbol{\alpha}^*, \beta^*)$ are the solutions to the following dual problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \beta} & -\frac{1}{2\beta} \left[\mathbf{x}^{(new)T} \mathbf{x}^{(new)} - 2s \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x}^{(new)} + \sum_i \sum_k \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k \right] + \sum \alpha_i - \beta \theta \\ \text{s.t.} & \begin{cases} 0 \leq \alpha_i \leq \beta \psi_d, \forall i \\ \sum_{i=1}^n \alpha_i y_i = s \\ \beta \geq 0 \end{cases} \end{aligned} \quad (3.5)$$

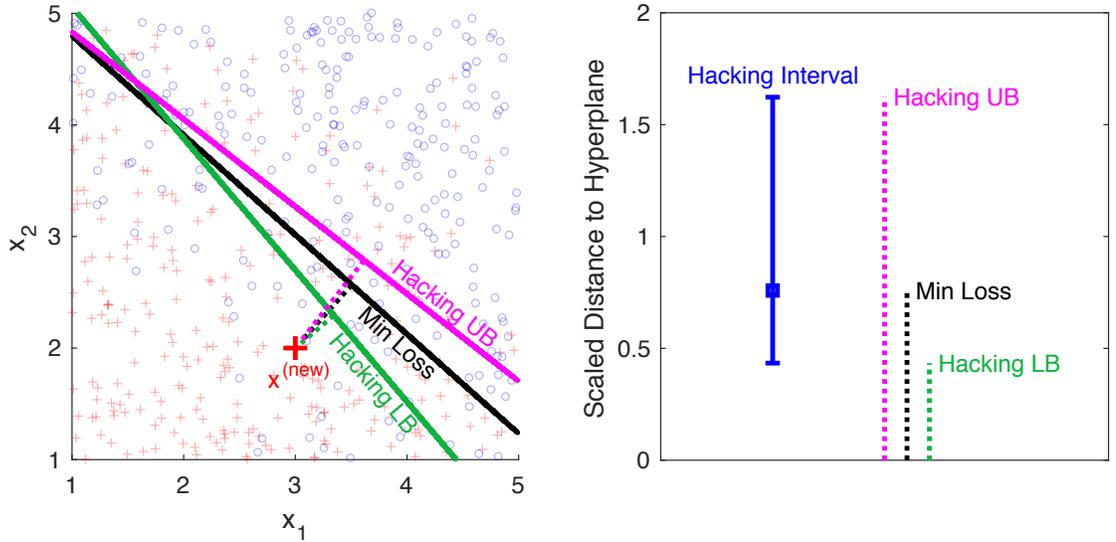


FIGURE 3.1: Hacking interval for an SVM prediction. The summary statistic being hacked is the distance from the separating hyperplane to a new observation, $\mathbf{x}^{(new)}$. For a default regularization tradeoff of $\psi_d = 1$ and a 5% tolerance on the loss relative to the minimum loss solution, SVM will always predict a +1 label, but the distance to the hyperplane can range from about 0.4 to about 1.6.

In Section 5.2 we apply SVM hacking intervals to a recidivism dataset.

3.1.2 Kernel Regression

Consider the form of kernel regression where prediction models $f \in \mathcal{F}$ evaluated at a point $\mathbf{x}_i \in \mathbb{R}^p$ are weighted averages of the labels of the other points, $\{y_j\}_{j \neq i}$, and where the weight is determined by a kernel function $k_{\psi_d}(d_A(\mathbf{x}_i, \mathbf{x}_j))$ that depends inversely on hyperparameters ψ_d and the distance $d_A(\mathbf{x}_i, \mathbf{x}_j)$ between points \mathbf{x}_i and \mathbf{x}_j for parameters A . That is, f is of the form:

$$f_A(\mathbf{x}_i) = \frac{\sum_{j \neq i} y_j k_{\psi_d}(d_A(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{j \neq i} k_{\psi_d}(d_A(\mathbf{x}_i, \mathbf{x}_j))}.$$

We suppose a quadratic loss function $L(\mathbf{Z}, f_A, \psi_d) = \sum_i (f_A(\mathbf{x}_i) - y_i)^2$, a Gaussian kernel $k_{\psi_d}(d(\mathbf{x}_i, \mathbf{x}_j)) = 1/\sqrt{2\pi\psi_d^2} \exp(-d(\mathbf{x}_i, \mathbf{x}_j)/\psi_d^2)$, and a Mahalanobis distance $d_A(\mathbf{x}_i, \mathbf{x}_j) = (x_i - x_j)^T A^T A (x_i - x_j)$, where A is a $p \times p$ matrix of parameters. Methods for learning A are considered by Weinberger and Tesauro (2017). A tethered hacking interval $[b_{\min}, b_{\max}]$ for prediction on a new point $\mathbf{x}^{(\text{new})}$ is given by:

$$b_{\min} = \min_A \frac{\sum_j y_j k(d_A(\mathbf{x}^{(\text{new})}, \mathbf{x}_j))}{\sum_j k(d_A(\mathbf{x}^{(\text{new})}, \mathbf{x}_j))} \quad \text{s.t.} \quad \sum_i \left(y_i - \frac{\sum_{j \neq i} y_j k(d_A(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{j \neq i} k(d_A(\mathbf{x}_i, \mathbf{x}_j))} \right)^2 \leq \theta$$

$$b_{\max} = \max_A \frac{\sum_j y_j k(d_A(\mathbf{x}^{(\text{new})}, \mathbf{x}_j))}{\sum_j k(d_A(\mathbf{x}^{(\text{new})}, \mathbf{x}_j))} \quad \text{s.t.} \quad \sum_i \left(y_i - \frac{\sum_{j \neq i} y_j k(d_A(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{j \neq i} k(d_A(\mathbf{x}_i, \mathbf{x}_j))} \right)^2 \leq \theta.$$

Figure 3.1.2 shows an example. Covariates are uniformly distributed on $[0, 10] \times [0, 10]$ and any points on a line of slope one have the same mean outcome. Therefore, the prediction function $f_A(\mathbf{x})$ should assign higher weight to the points to the upper right and to the lower left of \mathbf{x} since these points should have outcomes similar to the outcome of \mathbf{x} . Figure 3.1.2 shows this result in the level sets of the Mahalanobis distance metric that defines f_A . The middle panel corresponds to the minimum loss prediction function, while the left and right panels correspond to the prediction functions that minimize and maximize prediction on $\mathbf{x}^{(\text{new})} = (5, 5)^T$, respectively, within a loss constraint $\theta = 2000$ and for $\psi_d = 1$. These minimum and

maximum predictions, which define a tethered hacking interval, are -0.94 and 1.17 . The minimum loss prediction is -0.32 .

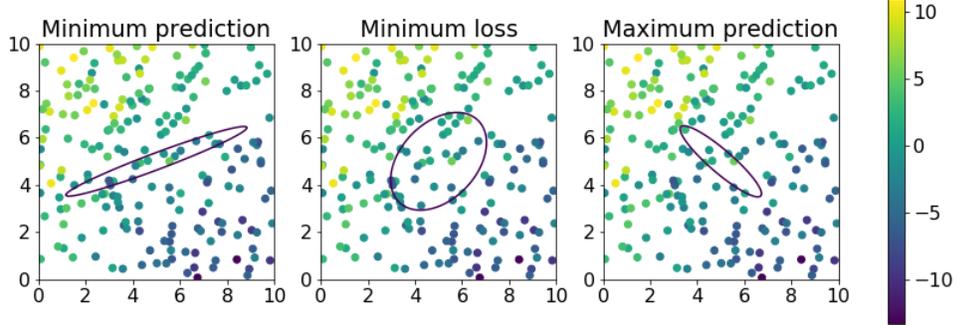


FIGURE 3.2: Level sets of the Mahalanobis distance that defines the kernel regression function f_A . The middle panel corresponds to the minimum loss prediction function, while the left and right panels correspond to the prediction functions that minimize and maximize prediction on $\mathbf{x}^{(\text{new})} = (5, 5)^T$, respectively, within a loss constraint $\theta = 2000$ and for $\psi_d = 1$.

3.1.3 PCA Feature Selection

Hacking intervals can also be used in the context of feature selection. We consider the example where principal components analysis is employed. There are a number of proposed methods (McCabe, 1984; Jolliffe, 1972; Xu et al., 2008), but we focus on one proposed by Guo et al. (2002), which is similar to Krzanowski (1987).

In this method, we start with a matrix \mathbf{X} of n observations and p features and we wish to find the matrix \mathbf{X}_q of n observations and $q < p$ of these features that gives to closest approximation to \mathbf{X} when we compare the principal component scores of \mathbf{X} and \mathbf{X}_q . That is, we do a principal component decomposition of each matrix, writing $\mathbf{S} = \mathbf{X}\mathbf{W}$ and $\mathbf{S}_q = \mathbf{X}_q\mathbf{W}_q$, where \mathbf{S} and \mathbf{S}_q are the matrices of principal component scores of \mathbf{X} and \mathbf{X}_q , respectively, and \mathbf{W} and \mathbf{W}_q the matrices of eigenvectors of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}_q^T\mathbf{X}_q$, respectively. We assume the columns of \mathbf{W} and \mathbf{W}_q are ordered

from largest to smallest eigenvalue, so that the columns of \mathbf{S} and \mathbf{S}_q are ordered from largest to smallest variance. We then pick an integer $k \leq q$ and compare the first k columns (*i.e.*, principal component scores) of \mathbf{S} and \mathbf{S}_q by “superimposing” one matrix on the other. That is, we find their Procrustes distance, which compares the matrices after optimal translation, scaling, and rotation. If we define the function space \mathcal{F} as the set of all selector functions that map the complete set $\{1, \dots, p\}$ of p feature indices to the selected set of $q < p$ feature indices, then the f (*i.e.*, the choice of q features) that maximizes this Procrustes distance for a given k will minimize the following loss function:

$$L(\mathbf{X}, f, k) = \text{trace}(\mathbf{S}^{(k)T} \mathbf{S}^{(k)} + \mathbf{S}_q^{(k)T} \mathbf{S}_q^{(k)} - 2\mathbf{\Sigma}) \quad (3.6)$$

where $\mathbf{S}^{(k)}$ and $\mathbf{S}_q^{(k)}$ are the first k columns of \mathbf{S} and \mathbf{S}_q , respectively, and $\mathbf{\Sigma}$ is the diagonal matrix of singular values of $\mathbf{S}^{(k)T} \mathbf{S}_q^{(k)}$. Equation (3.6) represents the loss of structural information in a candidate subset. In practice we scale the score matrices so that the loss is between 0 and 100. The number of component scores k is a hyperparameter. Note that while the researcher must choose the number of selected variables q , this number actually defines the problem, so we do not consider it a hyperparameter.

Among all of the q feature subsets that result in a loss of less than a small threshold θ , there are three questions we wish to ask. (i) Is a particular subset $j \in \{1, 2, \dots, \binom{p}{q}\}$ one of these subsets? That is, does subset j of q features yield a small loss of information? (ii) Is feature $i \in \{1, 2, \dots, p\}$ included or not included in any of these subsets? That is, to achieve a small loss of information using $q < p$ features, can we determine if a particular feature i must or must not be used? (iii) What is the maximum Hamming distance of these subsets to the optimal subset? That is, how different could a subset of q features that yields a small loss of information be from the subset of q features that yields the least loss of information? Each of these

questions corresponds to a different summary statistic. In (i) it is a binary indicator equal to 1 for subset j and 0 otherwise. In (ii) it is a binary indicator equal to 1 if variable i is included in a subset of q features and 0 if not. In (iii) it is the Hamming distance between two subsets (if we represent the subsets as binary indicators for each feature). Notice that in the first two cases, the hacking interval is either $[0, 0]$, $[1, 1]$, or $[0, 1]$, while in the last case the hacking interval is between 0 and, at most, $2q$, since at most q variables can differ.

Several PCA variable selection papers (Jolliffe, 1972; Guo et al., 2002; Krzanowski, 1987) have used a dataset on alate alleges (winged aphids), so we will do the same for comparison. This dataset consists of 40 observations of 19 variables. See Jeffers (1967) for a full description. Keeping with common practice on this dataset (Guo et al., 2002), we will restrict our analysis to selecting $q = 4$ and set a default $k = 4$ for the number of principal component scores. Reading the three panels from left to right in Figure 3.3 we see answers to our three questions for this dataset. Note that in each case, θ' is a number added to the minimum loss (which is out of 100).

Of the fourteen feature selection methods analyzed on this dataset by Guo et al. (2002), the selected features have losses given by Equation (3.6) that differ by as much as 7.65 (although, note that only 3 of the methods seek to minimize this particular loss) and are concentrated around a group of fifteen of the features (*i.e.*, four of the features are not selected by any of the fourteen methods). Our analysis shows that this greatly underestimates the diversity of features that could be selected, where “selected” means a feature subset yields a loss within a given tolerance of the minimum loss. Examining Figure 3.3, we find that within a loss tolerance of 7.65, more than 92% of the 3876 possible 4-feature subsets could be selected (left panel); within a loss tolerance of only 1.66, each of the 19 features is contained in at least one 4-feature subset that could be selected (middle panel); and within loss tolerance of only 0.59, a four-feature subset could be selected that is disjoint from

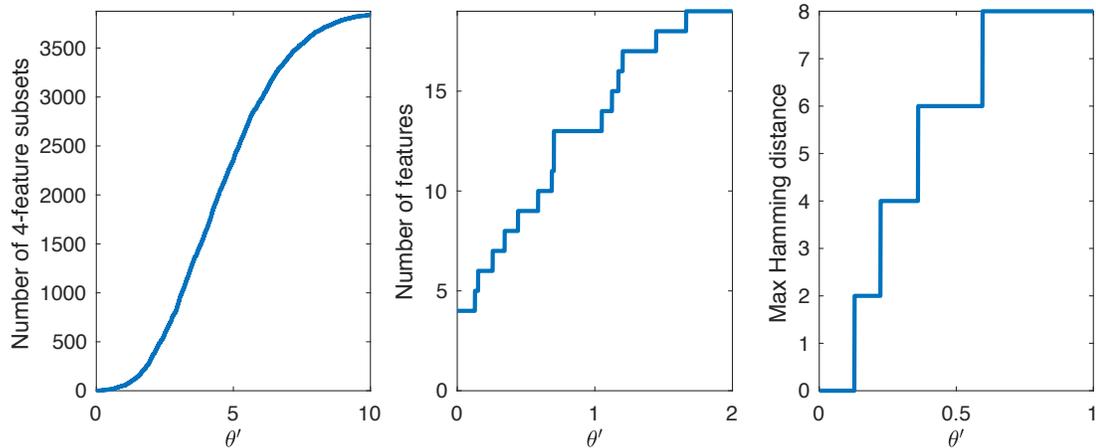


FIGURE 3.3: *Left:* Number of q -feature subsets that yield a loss within the tolerance *Middle:* Number of unique features that are included in at least one subset that yields a loss within the tolerance. *Right:* Maximum Hamming distance between the optimal q -feature solution any q -feature solution that yields a loss within the tolerance. θ' is a number added to the minimum loss (which is out of 100)

the optimal 4-feature subset (right panel). This illustrates the advantage of our systematic approach over the approach of aggregating past studies.

3.2 Connection to Profile Likelihood

Tethered hacking intervals are closely related to profile likelihood confidence intervals. When the loss function L corresponds to a likelihood function \mathcal{L} and the test statistic t is a single parameter of the learned prediction function f , then a tethered hacking interval is mathematically equivalent to a profile likelihood confidence interval for an appropriate choice of the loss threshold θ . In this section we quantify this equivalence.

Suppose a likelihood function $\mathcal{L}(\lambda, \xi)$ depends on a low-dimensional parameter of interest λ and a higher-dimensional nuisance parameter ξ . The *profile likelihood* \mathcal{L}_p focuses attention on λ by “profiling out” ξ :

$$\begin{aligned} \mathcal{L}_p(\lambda) &:= \sup_{\xi} \mathcal{L}(\lambda, \xi) \\ &= \mathcal{L}(\lambda, \hat{\xi}_{\lambda}), \end{aligned}$$

where $\hat{\xi}_\lambda$ maximizes the likelihood when λ is fixed. The profile likelihood \mathcal{L}_p is a lower dimensional function that for many purposes can be used instead of the full, higher-dimensional likelihood \mathcal{L} . Notably, $[\hat{\lambda}, \hat{\xi}]$ maximizes the full likelihood if and only if $\hat{\lambda}$ maximizes the profile likelihood. Interesting for our purposes is the property that the ratio of profile likelihoods $\mathcal{L}_p(\lambda_0)/\mathcal{L}_p(\hat{\lambda})$ equals the ratio of likelihoods Λ_{λ_0} for testing the null hypothesis $H_0 : \lambda = \lambda_0$:

$$\frac{\mathcal{L}_p(\lambda_0)}{\mathcal{L}_p(\hat{\lambda})} = \frac{\sup_{\xi} \mathcal{L}(\lambda_0, \xi)}{\sup_{\xi} \mathcal{L}(\hat{\lambda}, \xi)} = \frac{\sup_{\xi} \mathcal{L}(\lambda_0, \xi)}{\sup_{\lambda, \xi} \mathcal{L}(\lambda, \xi)} := \Lambda_{\lambda_0}.$$

By Wilks' Theorem, if H_0 is true and a few regularity conditions are met, then as the sample size $n \rightarrow \infty$,

$$-2 \log \Lambda_{\lambda_0} \xrightarrow{d} \chi_m^2,$$

where m is the difference in the dimensions of λ and ξ (Wilks, 1938). A hypothesis test for H_0 would reject H_0 if $-2 \log \Lambda_{\lambda_0}$ is large, which happens when the maximum likelihood under H_0 , $\sup_{\xi} \mathcal{L}(\lambda_0, \xi)$, is small. If λ is a scalar, the set of λ_0 for which $H_0 : \lambda = \lambda_0$ is not rejected provides a confidence interval for λ . That is, a $1 - \alpha$ *profile likelihood likelihood confidence interval* for λ is given by $[p_{\min}, p_{\max}]$, where:

$$p_{\min} := \min \lambda \quad \text{s.t.} \quad -2 \log(\Lambda_\lambda) \leq \chi_{m, 1-\alpha}^2,$$

$$p_{\max} := \max \lambda \quad \text{s.t.} \quad -2 \log(\Lambda_\lambda) \leq \chi_{m, 1-\alpha}^2,$$

and $\chi_{m, 1-\alpha}^2$ is the $1 - \alpha$ quantile of a χ^2 distribution with m degrees of freedom.

Equivalently, in terms of the profile likelihood we have:

$$p_{\min} = \min \lambda \quad \text{s.t.} \quad \log \mathcal{L}_p(\lambda) \geq \log \mathcal{L}_p(\hat{\lambda}) - \frac{1}{2} \chi_{m, 1-\alpha}^2, \quad (3.7)$$

$$p_{\max} = \max \lambda \quad \text{s.t.} \quad \log \mathcal{L}_p(\lambda) \geq \log \mathcal{L}_p(\hat{\lambda}) - \frac{1}{2} \chi_{m, 1-\alpha}^2. \quad (3.8)$$

If we define $\theta_p(\alpha) := \log \mathcal{L}_p(\hat{\lambda}) - \frac{1}{2} \chi_{m, 1-\alpha}^2$ then we have:

$$p_{\min} = \min \lambda \quad \text{s.t.} \quad \log \mathcal{L}_p(\lambda) \geq \theta_p(\alpha), \quad (3.9)$$

$$p_{\max} = \max \lambda \quad \text{s.t.} \quad \log \mathcal{L}_p(\lambda) \geq \theta_p(\alpha). \quad (3.10)$$

Equations (3.9) and (3.10) are similar to Equations (3.1) and (3.2) that define tethered hacking intervals if the summary statistic t is a single parameter λ of the prediction function $f_{\lambda,\xi}$; that is, if $t(f_{\lambda,\xi}) = \lambda$. In this case, both the profile likelihood confidence interval and the tethered hacking interval are the minimum and maximum of the summary statistic that can be achieved subject to a constraint on how well the prediction function fits the observed data. In the case of a profile likelihood confidence interval, the fit constraint is a lower bound $\theta_p(\alpha) := \log \mathcal{L}_p(\hat{\lambda}) - \frac{1}{2}\chi_{m,1-\alpha}^2$ on the profile likelihood. In the case of a tethered hacking interval, the fit constraint is an upper bound θ on the loss. To summarize, if $t(f_{\lambda,\xi}) = \lambda$, then:

$$\text{Profile likelihood interval: } \max_{\lambda} / \min_{\lambda} \lambda \text{ s.t. } \log \mathcal{L}_p(\lambda) \geq \theta_p(\alpha),$$

$$\text{Tethered hacking interval: } \max_{\lambda,\xi} / \min_{\lambda,\xi} \lambda \text{ s.t. } L(Z, f_{\lambda,\xi}, \psi_d) \leq \theta.$$

Notice that the profile likelihood confidence interval is an optimization over λ only, while the tethered hacking interval is an optimization over λ and ξ . However, since the objective function of the tethered hacking interval does not depend on ξ , and the loss is constrained by an upper bound, we can do no better in the optimization than by plugging in the ξ that minimizes the loss for a fixed λ , $\hat{\xi}_\lambda := \operatorname{argmin}_\xi L(Z, f_{\lambda,\xi}, \psi_d)$, into the tethered hacking interval constraint. In other words, we can “profile out” the nuisance parameter ξ from the loss function as we did with the likelihood function. Therefore, we have:

$$\text{Profile likelihood interval: } \max_{\lambda} / \min_{\lambda} \lambda \text{ s.t. } \log \mathcal{L}_p(\lambda) \geq \theta_p(\alpha) \quad (3.11)$$

$$\text{Tethered hacking interval: } \max_{\lambda} / \min_{\lambda} \lambda \text{ s.t. } L(Z, f_{\lambda,\hat{\xi}_\lambda}, \psi_d) \leq \theta. \quad (3.12)$$

This shows the equivalence of a profile likelihood confidence interval and a tethered hacking interval when the summary statistic is a single parameter λ of the prediction function. Notice the profile likelihood confidence interval requires the existence of

a likelihood model, whereas the tethered hacking interval requires the existence of only a loss function.

If we are given a likelihood function $\mathcal{L}(\lambda, \xi)$ and threshold $\theta_p(\alpha)$ that define a profile likelihood confidence interval for the parameter λ at confidence level α , we can construct a loss function $L(\mathbf{Z}, f_{\lambda, \xi}, \psi_d)$ and loss threshold θ that define an equivalent tethered hacking interval. We do this by defining the loss function as the negative log likelihood and the loss threshold as the negative likelihood threshold:

$$L(\mathbf{Z}, f_{\lambda, \xi}, \psi_d) := -\log \mathcal{L}(\lambda, \xi), \quad (3.13)$$

$$\theta := -\theta_p(\alpha). \quad (3.14)$$

We assume the function class \mathcal{F} is clear from the definition of the likelihood. Taking the infimum over ξ of Equation (3.13) we have:

$$\begin{aligned} \inf_{\xi} L(\mathbf{Z}, f_{\lambda, \xi}, \psi_d) &= \inf_{\xi} -\log \mathcal{L}(\lambda, \xi) \\ \iff L(\mathbf{Z}, f_{\lambda, \hat{\xi}_{\lambda}}, \psi_d) &= -\sup_{\xi} \log \mathcal{L}(\lambda, \xi) \\ &= -\log \sup_{\xi} \mathcal{L}(\lambda, \xi) \\ &= -\log \mathcal{L}_p(\lambda). \end{aligned}$$

This means:

$$\begin{aligned} \{\lambda \mid \log \mathcal{L}_p(\lambda) \geq \theta_p(\alpha)\} &= \{\lambda \mid -\log \mathcal{L}_p(\lambda) \leq -\theta_p(\alpha)\} \\ &= \left\{ \lambda \mid L(\mathbf{Z}, f_{\lambda, \hat{\xi}_{\lambda}}, \psi_d) \leq \theta \right\}. \end{aligned}$$

The profile likelihood confidence interval and tethered hacking interval defined in Equations (3.11) and (3.12), respectively, will therefore be the same, since each is defined by the minimum and maximum value of the same objective function over the same set.

We illustrate the construction of a tethered hacking interval from a profile likelihood confidence interval with the example of linear regression. Suppose the outcomes

$\mathbf{y} = (y_1, \dots, y_n)^T$ are generated by the following linear model with independent Gaussian noise of known variance σ^2 :

$$Y \sim N(\mathbf{X}\boldsymbol{\xi} + \mathbf{W}\lambda, \sigma^2 I),$$

where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ and $\mathbf{W} = (w_1, \dots, w_n)^T$ are, respectively, $n \times p$ and $n \times 1$ matrices of covariates, $\boldsymbol{\xi}$ is a $p \times 1$ vector of nuisance parameters, and λ is the scalar parameter of interest. The log likelihood for this model is:

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\xi}, \lambda_0) &= \log \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{x}_i^T \boldsymbol{\xi} + w_i \lambda - y_i)^2 \right\} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \boldsymbol{\xi} + w_i \lambda - y_i)^2. \end{aligned} \quad (3.15)$$

By Equation (3.11), the profile likelihood confidence interval at confidence level α is defined by the minimum and maximum values of λ for which the following inequality holds:

$$\begin{aligned} \log \mathcal{L}_p(\lambda) &\geq \theta_p(\alpha) \\ \log \mathcal{L}_p(\lambda) &\geq \log \mathcal{L}_p(\hat{\lambda}) - \frac{1}{2} \chi_{1,1-\alpha}^2 \\ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \hat{\boldsymbol{\xi}}_\lambda + w_i \lambda - y_i)^2 &\geq -\frac{n}{2} \log(2\pi\sigma^2) \\ &\quad - \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \hat{\boldsymbol{\xi}} + w_i \hat{\lambda} - y_i)^2 - \frac{1}{2} \chi_{1,1-\alpha}^2 \\ \sum_{i=1}^n (\mathbf{x}_i^T \hat{\boldsymbol{\xi}}_\lambda + w_i \lambda - y_i)^2 &\leq \sum_{i=1}^n (\mathbf{x}_i^T \hat{\boldsymbol{\xi}} + w_i \hat{\lambda} - y_i)^2 + \sigma^2 \chi_{1,1-\alpha}^2. \end{aligned} \quad (3.16)$$

To construct an equivalent tethered hacking interval, we define the loss by Equation

(3.13):

$$\begin{aligned}
L(\mathbf{Z}, f_{\xi, \lambda}, \psi_d) &:= -\log \mathcal{L}(\xi, \lambda) \\
&= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \xi + w_i \lambda - y_i)^2
\end{aligned}$$

and the loss threshold by Equation (3.14):

$$\begin{aligned}
\theta &:= -\theta_p(\alpha) \\
&= -\log \mathcal{L}_p(\hat{\lambda}) + \frac{1}{2} \chi_{m, 1-\alpha}^2 \\
&= \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \hat{\xi} + w_i \hat{\lambda} - y_i)^2 + \frac{1}{2} \chi_{1, 1-\alpha}^2.
\end{aligned}$$

By Equation (3.12), the tethered hacking interval is defined by the minimum and maximum values of λ for which the following inequality holds:

$$\begin{aligned}
L(Z, f_{\lambda, \hat{\xi}_\lambda}, \psi_d) &\leq \theta \\
\frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \hat{\xi} + w_i \lambda - y_i)^2 &\leq \frac{n}{2} \log(2\pi\sigma^2) \\
&\quad + \frac{1}{2\sigma^2} \sum_{i=1}^n (\mathbf{x}_i^T \hat{\xi} + w_i \hat{\lambda} - y_i)^2 + \frac{1}{2} \chi_{1, 1-\alpha}^2 \\
\sum_{i=1}^n (\mathbf{x}_i^T \hat{\xi}_\lambda + w_i \lambda - y_i)^2 &\leq \sum_{i=1}^n (\mathbf{x}_i^T \hat{\xi} + w_i \hat{\lambda} - y_i)^2 + \sigma^2 \chi_{1, 1-\alpha}^2.
\end{aligned} \tag{3.17}$$

Since Equations (3.16) and (3.17) are the same, the profile likelihood confidence interval and tethered hacking intervals are the same, since each is defined by the minimum and maximum value of λ over the same set of λ . Notice any monotonic function applied to the loss function and loss threshold will yield the same set of λ for which the loss function exceeds the loss threshold.

3.3 Generalization Bound

Tethered hacking intervals do not require the assumption of a true data generating process; however, if we do assume a true data generating process μ , we can extend the interpretation of tethered hacking intervals to this setting and, in the case of learning classification functions, derive a generalization bound that incorporates both the uncertainty due to observing a random draw from the true data generating process *and* the uncertainty due to researcher hacking. The interpretation of a hacking interval in this setting begins with the following sequence of events:

1. The true distribution μ generates independently and identically a “pristine” dataset \mathbf{Z}_p of n observations.
2. $\tilde{\phi} : \mathcal{Z} \rightarrow \mathcal{Z}$ transforms the pristine data, \mathbf{Z}_p , into what the researcher actually observes, \mathbf{Z}_o .
3. $\phi : \mathcal{Z} \rightarrow \mathcal{Z}$ transforms the observed data, \mathbf{Z}_o , into the hacked data the researcher uses in their analysis, \mathbf{Z}_h . This is the same ϕ as in the tethered hacking interval definition.

In other words, in order to do inference or prediction on the true data generating process μ , a researcher would like to perform their analysis on a dataset generated by μ , which we call the pristine data \mathbf{Z}_p . However, we assume the researcher only observes \mathbf{Z}_o , which is the pristine data after applying an unknown data adjustment function $\tilde{\phi}$. In an attempt to undo $\tilde{\phi}$, the researcher chooses a data adjustment function of their own, ϕ . The result of applying ϕ to \mathbf{Z}_o is the hacked data \mathbf{Z}_h . Schematically, we can write this procedure as:

$$\underbrace{\mathbf{Z}_p}_{\text{Pristine data}} \xrightarrow{\tilde{\phi}} \underbrace{\mathbf{Z}_o}_{\text{Observed data}} \xrightarrow{\phi} \underbrace{\mathbf{Z}_h}_{\text{Hacked data}} .$$

On each of the three datasets — \mathbf{Z}_p , \mathbf{Z}_o , and \mathbf{Z}_h — there is at least one function from a function class \mathcal{F} that minimizes the empirical risk on a loss function $L(\mathbf{Z}, f)$.¹ We call these functions f_p , f_o , and f_h , respectively. That is, we define the following:

- $f_p \in \arg \min_{f \in \mathcal{F}} R_{\mathbf{Z}_p}^{\text{emp}}(f) := \arg \min_{f \in \mathcal{F}} L(\mathbf{Z}_p, f)$ minimizes the empirical risk of the pristine dataset, \mathbf{Z}_p . Since we do not observe \mathbf{Z}_p we cannot learn it.
- $f_o \in \arg \min_{f \in \mathcal{F}} R_{\mathbf{Z}_o}^{\text{emp}}(f) := \arg \min_{f \in \mathcal{F}} L(\mathbf{Z}_o, f)$ minimizes the empirical risk of the observed dataset, \mathbf{Z}_o . This is the function that would be learned if we did not allow for hacking.
- $f_h \in \arg \min_{f \in \mathcal{F}} R_{\mathbf{Z}_h}^{\text{emp}}(f) := \arg \min_{f \in \mathcal{F}} L(\mathbf{Z}_h, f)$ minimizes the empirical risk of the hacked dataset, \mathbf{Z}_h . It is learned by the researcher.

In a classical statistical setting — where the researcher observes \mathbf{Z}_p and computes f_p — a question from statistical learning theory is how the true risk of f_p , $R_\mu^{\text{true}}(f_p) := \mathbb{E}_{Z \sim \mu} L(Z, f_p)$, differs from the empirical risk of f_p on \mathbf{Z}_p , $R_{\mathbf{Z}_p}^{\text{emp}}(f_p) := L(\mathbf{Z}_p, f_p)$. If \mathcal{F} consists of classification functions, a bound on this difference can be found using statistical learning theory. In a hacking setting — where the researcher observes \mathbf{Z}_o , adjusts the data to obtain \mathbf{Z}_h , and computes f_h — there is additional uncertainty due to the data adjustments. In order to derive an analogous bound, this time to understand how the true risk of f_p differs from the empirical risk of f_h on \mathbf{Z}_h , we will need to make a few assumptions about the impact of the data adjustments:

- $\left| R_{\mathbf{Z}_p}^{\text{emp}}(f_p) - R_{\mathbf{Z}_o}^{\text{emp}}(f_p) \right| \leq \theta_1$, which we call “reverse tethering (part 1).” It means that the function f_p that minimizes the loss on the pristine data does not yield a loss too different on the observed data.

¹ Both the function class \mathcal{F} and the loss function $L(\mathbf{Z}, f)$ can depend on hyperparameters ψ_d , but hyperparameters are not important to this section so we suppress their notation.

- $|R_{\mathbf{Z}_o}^{\text{emp}}(f_p) - R_{\mathbf{Z}_o}^{\text{emp}}(f_o)| \leq \theta_2$, which we call “reverse tethering (part 2).” It means that the functions learned from the pristine data and the observed data do not have losses too different from each other on the observed data.
- $|R_{\mathbf{Z}_o}^{\text{emp}}(f_o) - R_{\mathbf{Z}_o}^{\text{emp}}(f_h)| \leq \theta_3$, which is our standard tethering constraint used in Equations (3.1) and (3.2). It means that the functions learned from the observed data and the hacked data do not have losses too different from each other on the observed data.
- Let $|R_{\mathbf{Z}_o}^{\text{emp}}(f_h) - R_{\mathbf{Z}_h}^{\text{emp}}(f_h)| = \theta_4$. This is not an assumption since it can be calculated by the researcher. For the function f_h that minimizes the loss on the hacked data, it is the difference between the loss on the observed data and the hacked data.

We can bound the difference between the true risk of f_h and the empirical risk of f_h on \mathbf{Z}_h by applying the triangle inequality several times and bounding each intermediate difference. The assumptions about θ_1 , θ_2 , and θ_3 , and the calculated θ_4 , provide bounds on all but one of these intermediate differences. The final intermediate difference is between the true risk of f_h and the empirical risk of f_h on f_p . This can be bounded by the same learning theory bound we would derive in a classical statistical setting, since it holds uniformly for all functions in \mathcal{F} . Table 3.3 summarizes the relationships and Proposition 2 gives the result.

Proposition 2 (Generalization Bound for Hacked Data). *If \mathcal{F} is a set of classification functions with Vapnik-Chervonenkis dimension h , then, for all $\delta > 0$, with probability of at least $1 - \delta$ with respect to data \mathbf{Z}_p drawn i.i.d. from an unknown distribution μ on $\mathbb{R}^{n \times p} \times \{-1, 1\}^n$:*

$$|R_{\mu}^{\text{true}}(f_p) - R_{\mathbf{Z}_h}^{\text{emp}}(f_h)| \leq 2\sqrt{2\frac{h \log \frac{2eh}{n} + \log \frac{4}{\delta}}{n}} + \sum_{i=1}^4 \theta_i.$$

	μ		\mathbf{Z}_p		\mathbf{Z}_o		\mathbf{Z}_h
f_p	$R_{Z_\mu}^{\text{true}}(f_p)$	\xrightarrow{VC}	$R_{Z_p}^{\text{emp}}(f_p)$	$\xrightarrow{\theta_1}$	$R_{Z_o}^{\text{emp}}(f_p)$		$R_{Z_h}^{\text{emp}}(f_p)$
f_o	$R_{Z_\mu}^{\text{true}}(f_o)$		$R_{Z_p}^{\text{emp}}(f_o)$		$R_{Z_o}^{\text{emp}}(f_o)$	$\downarrow \theta_2$	$R_{Z_h}^{\text{emp}}(f_o)$
f_h	$R_{Z_\mu}^{\text{true}}(f_h)$		$R_{Z_p}^{\text{emp}}(f_h)$		$R_{Z_o}^{\text{emp}}(f_h)$	$\downarrow \theta_3$	$R_{Z_h}^{\text{emp}}(f_h)$
						$\xrightarrow{\theta_4}$	$R_{Z_h}^{\text{emp}}(f_h)$

Table 3.1: We can relate the true risk of f_h , $R_{Z_\mu}^{\text{true}}(f_p)$, to the empirical risk of f_h on \mathbf{Z}_h , $R_{Z_h}^{\text{emp}}(f_h)$, by applying a VC bound and the triangle inequality four times.

Tethered Hacking Intervals for Linear Regression

We develop hacking intervals in detail for two linear regression scenarios:

- *Scenario 1: average treatment effect.* We assume a class of linear functions \mathcal{F} with p confounders and an indicator covariate for the treatment (1 if treatment, 0 if control). We write $f \in \mathcal{F}$ as:

$$f(\mathbf{x}, \text{treated or control}) = \beta_1 x_{.1} + \beta_2 x_{.2} + \dots + \beta_p x_{.p} + \beta_0 \mathbf{1}_{\text{treated}}.$$

The goal is to construct a tethered hacking interval for β_0 , the coefficient of the treatment indicator. In other words, the test statistic is $t(\mathbf{Z}, f) = \beta_0$. The coefficient represents β_0 represents the average treatment effect. Section 4.1 develops this in detail.

- *Scenario 2: individual treatment effect.* We assume a class of linear functions \mathcal{F} with p confounders for both the treatment and control groups. We write

$f \in \mathcal{F}$ as:

$$f(\mathbf{x}, \text{treated or control}) = 1_{\text{control}}[\beta_1^c x_{.1} + \beta_2^c x_{.2} + \dots \beta_p^c x_{.p}] + \\ 1_{\text{treated}}[\beta_1^t x_{.1} + \beta_2^t x_{.2} + \dots \beta_p^t x_{.p}],$$

where 1_{control} is 1 only for the control group and 1_{treated} is 1 only for the treatment group. The goal is to construct a tethered hacking interval for a prediction of f on a new point $[\mathbf{x}^{(\text{new})}, \text{treated or control}]$. In other words, the test statistic is $t(\mathbf{Z}, [\mathbf{x}^{(\text{new})}, \text{treated or control}], f) = f(\mathbf{x}^{(\text{new})}, \text{treated or control})$. The value $f(\mathbf{x}^{(\text{new})}, \text{treated or control})$ represents the prediction for a person with covariates $\mathbf{x}^{(\text{new})}$. Section 4.2 develops this in detail.

In both scenarios we assume ignorability, we assume no unmeasured confounding, and we use a quadratic loss function $L(\mathbf{Z}, f) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, 1_{[i \text{ treated}]}))^2$, where $\mathbf{Z} = \{[\mathbf{x}_i, 1_{[i \text{ treated}]}, y_i]\}_{i=1}^n$ is the observed data. There are no hyperparameters so we suppress their notation in the loss function. Throughout this section, we assume the means have been subtracted from each feature so that they are normalized.

4.1 Scenario 1: Average Treatment Effect

The goal is to find the range of treatment effects, β_0 , corresponding to all possible ways the analyst can hack the observed data subject to a constraint in the loss. Thus our goal is to solve:

$$\max_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \beta_0 \quad \text{s.t.} \quad \sum_{i=1}^n (y_i - \beta \mathbf{x}_i - \beta_0 1_{[i \text{ treated}]})^2 \leq \theta, \quad (4.1)$$

and

$$\min_{\beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \beta_0 \quad \text{s.t.} \quad \sum_{i=1}^n (y_i - \beta \mathbf{x}_i - \beta_0 1_{[i \text{ treated}]})^2 \leq \theta. \quad (4.2)$$

This is a convex quadratically constrained linear program. Since there are inequality constraints we require the full KKT conditions (the method of Lagrange multipliers does not handle inequality constraints). As it turns out, answers to these problems can be found analytically. This is one of the rare problems for which a subset of the KKT conditions can be used to find a closed form solution.

Theorem 3 (Hacking Interval for Least-Squares ATE). *Define the following:*

- $\beta_{LS}^* := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, the optimal least square solution from regressing \mathbf{Y} on \mathbf{X} . The coefficient within this vector for the treatment variable is denoted $\tilde{\beta}_{0,LS}^*$.
- $\tilde{\beta}_{LS}^* := (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y}$, the optimal least square solution from regressing \mathbf{Y} on $\tilde{\mathbf{X}} := [\mathbf{X}, \mathbf{1}_{[treated]}]$.
- $\gamma_{LS}^* := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{[treated]}$, the optimal least square solution from regressing $\mathbf{1}_{[treated]}$ on \mathbf{X} .
- $V_{tt} := \left(\left[\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \right]^{-1} \right)_{tt}$, the diagonal entry corresponding to the treatment variable of $\left[\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \right]^{-1}$.

Then, the solutions of the optimization problem (4.1) are:

$$\beta_{0,\max}^* = \tilde{\beta}_{0,LS}^* + \sqrt{V_{tt}} \sqrt{\theta - SSE}, \quad (4.3)$$

$$\beta_{\max}^* = \beta_{LS}^* - \beta_{0,\max}^* \gamma_{LS}^*. \quad (4.4)$$

and the solutions of the optimization problem (4.2) are:

$$\beta_{0,\min}^* = \tilde{\beta}_{0,LS}^* - \sqrt{V_{tt}} \sqrt{\theta - SSE}, \quad (4.5)$$

$$\beta_{\min}^* = \beta_{LS}^* - \beta_{0,\min}^* \gamma_{LS}^*. \quad (4.6)$$

From this theorem, one can see that the range $\beta_{0,\max}^* - \beta_{0,\min}^*$ scales as the square root of the permitted level of optimality θ . The solution is not difficult to find if the relevant KKT conditions are substituted into each other in a particular order.

Next, we relate the new confidence intervals to the standard ones and then produce new interpretations for confidence intervals, based on in-sample error increases. In the process, we will discuss possible meanings for the user-defined parameter θ .

4.1.1 Relationship to classical confidence intervals.

We have just produced a confidence interval for β_0 . How does that compare with a typical confidence interval produced using the standard approach where we assume a null distribution? The confidence interval is symmetric in both cases around the least squares solution, so we must be able to equate them. We next equate traditional confidence intervals with our confidence intervals, which relates α for a significance test with θ for our robust confidence interval.

Theorem 4 (Relating Least-Squares ATE Hacking Intervals to Standard Confidence Intervals). *Start with a standard confidence interval for β_0 under usual assumptions (normality of errors given a linear model), which is given by:*

$$\left[\tilde{\beta}_{LS} - t_{(1-\alpha/2),(n-p-1)} \sqrt{\frac{SSE}{n-p-1}} \sqrt{V_{tt}}, \tilde{\beta}_{LS} + t_{(1-\alpha/2),(n-p-1)} \sqrt{\frac{SSE}{n-p-1}} \sqrt{V_{tt}} \right]$$

where $t_{(1-\alpha/2),(n-p-1)}$ is the $1 - \alpha/2$ quantile of a t distribution with $n - p - 1$ degrees of freedom (we estimate p coefficients plus the treatment variable). Then, in order to keep the new confidence interval from Theorem 3 the same as the standard one, we would take the following value for θ :

$$\theta = SSE \left(1 + \frac{t_{(1-\alpha/2),(n-p-1)}^2}{n-p-1} \right).$$

Thus, for teaching purposes, rather than explaining the t distribution or the meaning of α to a student unfamiliar with these topics, we can explain θ first and later convert to α for those who want this interpretation.

4.1.2 Non-classical Choices for θ .

In classical hypothesis testing, one would choose the significance level α and say that if the data were drawn repeatedly from the true model, the probability that an estimated value of β_0 would be within the confidence interval with probability at least $1 - \alpha$. We propose *in-sample* alternatives based on the meaning of θ . Here are some natural choices:

- *Choose θ as a percentage of the SSE:* Assume the user would not allow a model that would achieve more than 10% higher error than the SSE. Then we set $\theta = .1 \cdot \text{SSE}$. Generally if we do not tolerate more than $r\%$ error higher than the SSE, we would choose $\theta = r \cdot \text{SSE}$.

To use this, we would ask questions like: "If we were to tolerate any type of change to the data or model that would incur an additional error of 10%, what are the largest and smallest treatment effect one could estimate?" If the answer is that the treatment effect estimate is robust to 10% error due to user hacking, then the estimate is reliable.

- *Choose θ as the minimum loss suffered to allow the treatment effect coefficient to be 0.* Let us say without loss of generality that the estimated treatment effect coefficient is negative. Then the upper confidence interval is (using Theorem 3):

$$\beta_{0,\max}^* = \beta_{0,LS}^* + \sqrt{V_{tt}}\sqrt{\theta - \text{SSE}},$$

We can set this value to 0, which would provide the minimum sacrifice in least square error necessary for that coefficient to become 0. We thus need to solve

for θ in the following:

$$\begin{aligned} 0 &= \beta_{0,LS}^* + \sqrt{V_{tt}}\sqrt{\theta_0 - \text{SSE}}, \\ \theta_0 &= \frac{(\beta_{0,LS}^*)^2}{V_{tt}} + \text{SSE}. \end{aligned}$$

In other words, we would need to sacrifice a least square error of at least $\frac{(\beta_{0,LS}^*)^2}{V_{tt}}$ beyond that of the optimal solution in order that the regression coefficient could be 0.

To use this, we would ask questions like: "How much loss would need to be sacrificed in order for the treatment to have the opposite estimated effect?"

4.1.3 Combining with data variance

The bounds of a the hacking interval, $\beta_{0,\max}^*$ and $\beta_{0,\min}^*$, are deterministic functions of a fixed dataset $[\tilde{\mathbf{X}}, \mathbf{Y}]$. If we assume the outcomes \mathbf{Y} are one possible realization of a ground truth linear process given by:

$$\mathbf{Y} \sim N(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 I), \quad (4.7)$$

then the bounds of the hacking interval are random variables. The following theorem gives their variance.

Theorem 5 (Variance of Least-Squares ATE Hacking Interval Bounds). *If outcomes \mathbf{Y} are generated by equation 4.7 and the threshold θ is set to $(1+r)\text{SSE}$ for any $r > 0$, then the variance of both hacking interval bounds $\beta_{0,\min}^*$ and $\beta_{0,\max}^*$ given by Equations (4.5) and (4.3), respectively, is:*

$$\mathbb{V} \left[\beta_{0,\min}^* \mid \tilde{\mathbf{X}} \right] = \mathbb{V} \left[\beta_{0,\max}^* \mid \tilde{\mathbf{X}} \right] = \sigma^2 V_{tt} (1 + r(n - p - 1 - \mu^2)), \quad (4.8)$$

where

$$\mu = \left(\frac{\sqrt{2}\Gamma((n-p)/2)}{\Gamma((n-p-1)/2)} \right). \quad (4.9)$$

4.1.4 Illustration

Let us consider an illustrative example. We suppose a ground truth with two covariates called $v_{.1}$ and $v_{.2}$, chosen uniformly and independently over the interval $[1,5]$, a $1/2$ probability of treatment assignment for each observation, and outcomes generated by the following process:

$$y_i = 2 \times 1_{[i \text{ treated}]} + v_{i1} + v_{i2} + \epsilon_i, \quad (4.10)$$

where $\epsilon_i \sim N(0, 1)$. In this illustration, the researcher observes more than v_{i1} , v_{i2} , and the treatment indicator, $1_{[i \text{ treated}]}$. We assume they observe monomials $\mathbf{x}_i = (v_{i1}, v_{i2}, v_{i1}^2, v_{i2}^2, v_{i1}v_{i2}, v_{i1}v_{i2}^2, v_{i1}^2v_{i2}, v_{i1}^2v_{i2}^2)$ and $1_{[i \text{ treated}]}$. In the language of Section 3.3, $\{[v_{i1}, v_{i2}, 1_{[i \text{ treated}]}], y_i\}_{i=1}^n$ is the pristine data and $\{[\mathbf{x}_i, 1_{[i \text{ treated}]}], y_i\}_{i=1}^n$ is the observed data. This puts the researcher at risk of overfitting the observed covariates in \mathbf{x}_i that are not part of the ground truth.

We simulated a dataset of $n = 500$ observations and used Theorem 3 to find the values of β_{\max}^* , $\beta_{0,\max}^*$, β_{\min}^* , and $\beta_{0,\min}^*$, where θ was set to 10% higher than the least squares loss of β_{LS}^* . Table 4.1.4 gives the results for the coefficient on treatment indicator, β_0 . To illustrate these results, on a grid of $v_{.1}^{\text{new}}$ and $v_{.2}^{\text{new}}$, we found the vector of monomials that would be observed by the researcher, $\mathbf{x}^{(\text{new})}$, and evaluated the four possible outcome predictions (max and min, treatment and control):

$$\hat{y}_{\max,\text{treated}} = \mathbf{x}^{(\text{new})} \beta_{\max}^* + 1 \times \beta_{0,\max}^* \quad (4.11)$$

$$\hat{y}_{\min,\text{treated}} = \mathbf{x}^{(\text{new})} \beta_{\min}^* + 1 \times \beta_{0,\min}^* \quad (4.12)$$

$$\hat{y}_{\min,\text{untreated}} = \mathbf{x}^{(\text{new})} \beta_{\min}^* + 0 \times \beta_{0,\min}^* \quad (4.13)$$

$$\hat{y}_{\max,\text{untreated}} = \mathbf{x}^{(\text{new})} \beta_{\max}^* + 0 \times \beta_{0,\max}^* \quad (4.14)$$

Equations (4.11) through (4.14) are ordered by value, from largest to smallest. This gives four surface plots, shown in Figure 4.1 from different rotations. Asymptotically, or if we had a larger number of points, the curves would be hyperplanes since the

$\beta_{0,\min}^*$	β_{LS}^*	$\beta_{0,\max}^*$
1.52	2.16	2.80

Table 4.1: Minimum, least-squares, and maximum coefficient on the treatment indicator. $[\beta_{0,\min}^*, \beta_{0,\max}^*]$ is the tethered hacking interval. The ground truth is $\beta_0 = 2$.

ground truth in Equation (4.10) depends linearly on $v_{.1}$ and $v_{.2}$. As it stands, the curves are very close to the optimal hyperplanes, overfitting only slightly.

We would like to consider *individual* treatment effects, where the treatment effects can differ between units. The simple regression setting above will predict a constant treatment effect for all units, so we need to have a more flexible modeling approach.

4.2 Scenario 2: Individual Treatment Effect

For the second regression scenario we consider, the regression model is more flexible, including separate terms for treatment and control. Our goal is to find the range of treatment effects for a particular point $\mathbf{x}^{(\text{new})}$.

To explain the motivation for this problem, let us consider a new patient receiving a prediction of the expected treatment effect for a drug. Before taking the drug, the patient might want to know whether there are other reasonable models that give different predictions. That is, the patient might want to know the answer to the following: *Considering all reasonable models for predicting treatment effects, what are the largest and smallest possible predicted treatment effects for this drug on me?*

To determine the range, we solve:

$$\max_{\beta, \beta_0} f(\mathbf{x}^{(\text{new})}) \quad \text{s.t.} \quad \sum_{i=1}^n \left(f(\mathbf{x}_i^{(\text{new})}) - y_i^{(\text{new})} \right)^2 \leq \theta.$$

and

$$\min_{\beta, \beta_0} f(\mathbf{x}^{(\text{new})}) \quad \text{s.t.} \quad \sum_{i=1}^n \left(f(\mathbf{x}_i^{(\text{new})}) - y_i^{(\text{new})} \right)^2 \leq \theta.$$

All four treatment prediction curves.

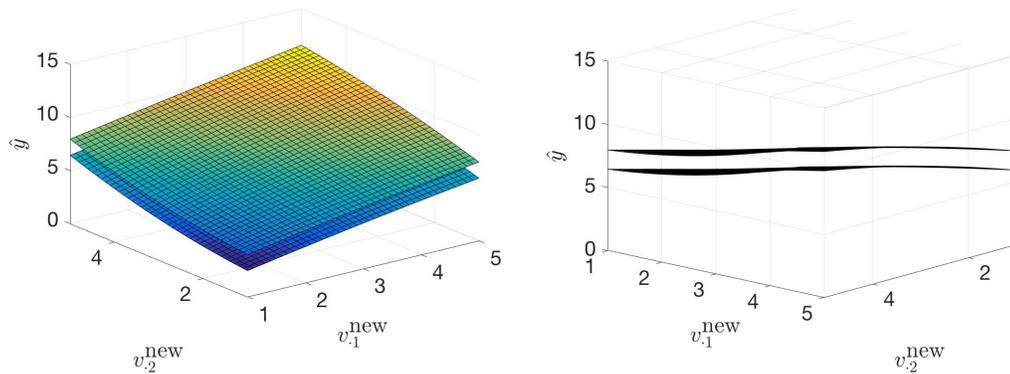
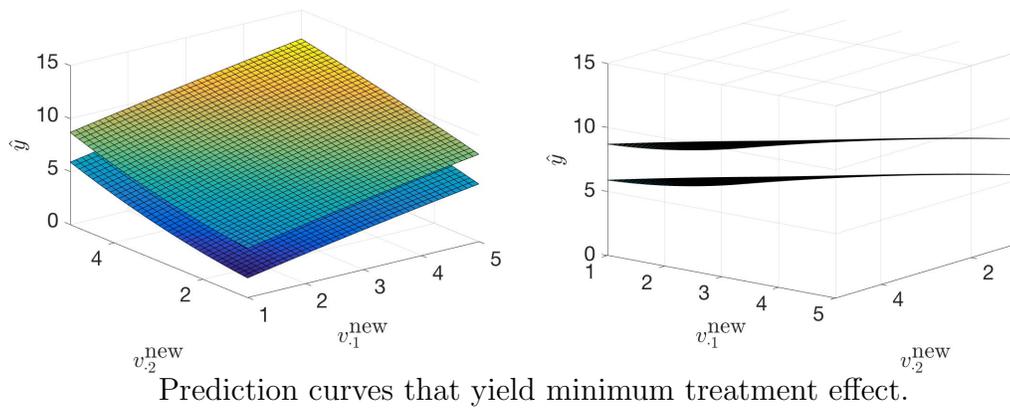
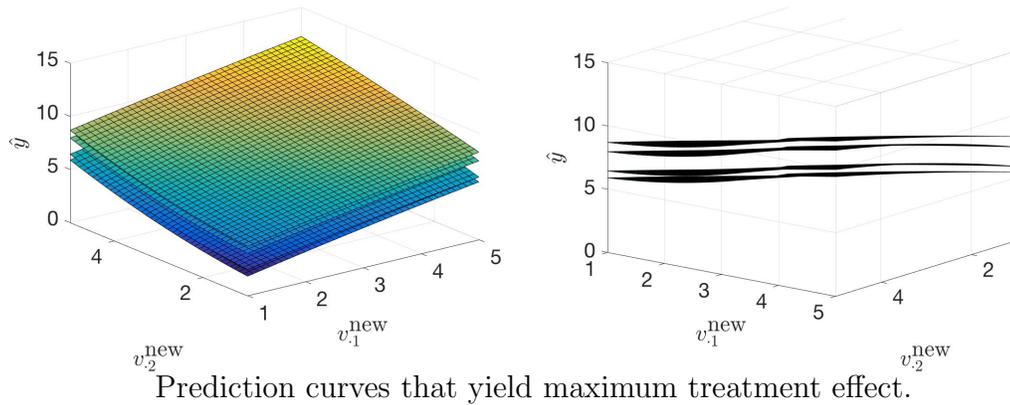


FIGURE 4.1: For all rows, the left and right figures are two different vantage points of the same figure. Since the true data generation process in Equation (4.10) depends linearly on only v_1 and v_1 , the optimal prediction curve as a function of v_1 and v_1 is a hyperplane. The addition of monomials to the observed \mathbf{x} causes some overfitting. (Top Row) All four prediction curves (max/min, treatment/control). (Second Row) Prediction curves that yield the maximum treatment effect. The upper curve shows $\hat{y}_{\max, \text{treated}}$ and the lower curve shows $\hat{y}_{\max, \text{untreated}}$. The difference between the curves is the maximum treatment effect, $\beta_{0, \max}^*$. These curves correspond to the top and bottom curves in the top row of plots. (Third Row) Prediction curves that yield the minimum treatment effect. The upper curve shows $\hat{y}_{\min, \text{treated}}$ and the lower curve shows $\hat{y}_{\min, \text{untreated}}$. The difference between the curves is the minimum treatment effect, $\beta_{0, \min}^*$. These curves correspond to the middle two curves in the top row of plots. Code written by Cynthia Rudin.

The model is:

$$f(\mathbf{x}, \text{treated or control}) = 1_{\text{control}}[\beta_1^c x_{.1} + \beta_2^c x_{.2} + \dots + \beta_p^c x_{.p}] + 1_{\text{treated}}[\beta_1^t x_{.1} + \beta_2^t x_{.2} + \dots + \beta_p^t x_{.p}].$$

Using notation $w_i = 1$ for treatment points, and $w_i = 0$ for control points, the least squares loss thus decouples, leading to separate regression problems for the treatment and control points:

$$\begin{aligned} & \sum_{i=1}^n (f(\mathbf{x}_i, w_i) - y_i)^2 \\ &= \sum_{i:w_i=1} (f(\mathbf{x}_i, 1) - y_i)^2 + \sum_{i:w_i=0} (f(\mathbf{x}_i, 0) - y_i)^2 \\ &= \sum_{i:w_i=1} ([\beta_1^c x_{i1} + \beta_2^c x_{i2} + \dots + \beta_p^c x_{ip}] - y_i)^2 + \sum_{i:w_i=0} ([\beta_1^t x_{i1} + \beta_2^t x_{i2} + \dots + \beta_p^t x_{ip}] - y_i)^2. \end{aligned}$$

Because the first sum involves only the control observations and control coefficients, and the second sum involves only treatment observations and treatment coefficients, this decouples as two separate regressions, one for the control group, and one for the treatment group. We will assume that the user wants neither of the regressions to be too suboptimal, so we will have separate constraints θ on the quality of each regression. We will find the maximum and minimum values for the control regression and the treatment regressions (four values). All of these optimization problems are very similar, so for simplicity, we solve the optimization problem on a generic regression problem, for point $\mathbf{x}^{(\text{new})}$. Here $\mathbf{x}^{(\text{new})}$ does not need to be one of the training observations.

Theorem 6 (Hacking Intervals for Least-Squares Individual TE). *Consider the hack-*

ing interval optimization problems:

$$\begin{aligned} \max_{\boldsymbol{\beta}}(\mathbf{x}^{(\text{new})}\boldsymbol{\beta}) \text{ such that } \sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 \leq \theta, \\ \min_{\boldsymbol{\beta}}(\mathbf{x}^{(\text{new})}\boldsymbol{\beta}) \text{ such that } \sum_{i=1}^n (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 \leq \theta. \end{aligned}$$

Define $\boldsymbol{\beta}_{LS}^* := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, define $\Upsilon = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{(\text{new})T}$, which is a vector of size p , $SSE = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|^2$, and

$$\tilde{\mu} = \frac{\sqrt{\theta - SSE}}{\|\mathbf{X}\Upsilon\|}.$$

The solutions to the optimization problems above are:

$$\boldsymbol{\beta}_+^* = \boldsymbol{\beta}_{LS}^* - \tilde{\mu}\Upsilon, \quad \boldsymbol{\beta}_-^* = \boldsymbol{\beta}_{LS}^* + \tilde{\mu}\Upsilon.$$

Theorem 7. (Relating hacking intervals of individual treatment effects to standard confidence intervals) Start with a standard confidence interval for $\mathbf{x}^{(\text{new})}\boldsymbol{\beta}$ under usual assumptions (normality of errors given a linear model), which is given by the boundary points:

$$\boldsymbol{\beta}_{LS}^* \pm t_{(1-\alpha/2), (n-p-1)} \sqrt{\frac{SSE}{n-p-1}} \sqrt{\mathbf{x}^{(\text{new})}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{(\text{new})T}}$$

where $t_{(1-\alpha/2), (n-p-1)}$ is the $1 - \alpha/2$ quantile of a t distribution with $n - p - 1$ degrees of freedom. Then, in order to keep the hacking interval from Theorem 6 the same as the standard one, we would take the following value for θ :

$$\theta = SSE \left(1 + \frac{t_{(1-\alpha/2), (n-p-1)}^2}{n-p-1} \right).$$

We can use the result of Theorem 6 to determine the hacking interval, which in this case is the range of causal effect estimates for $\mathbf{x}^{(\text{new})}$. Let us apply Theorem 6 to

the treatment regression and the control regression separately. We thus obtain β_+^{t*} , β_-^{t*} , β_+^{c*} , and β_-^{c*} . To find the maximum of the causal effect estimate, use:

$$\max(\mathbf{x}^{(\text{new})}\beta_+^{t*}, \mathbf{x}^{(\text{new})}\beta_-^{t*}) - \min(\mathbf{x}^{(\text{new})}\beta_+^{c*}, \mathbf{x}^{(\text{new})}\beta_-^{c*}).$$

To find the minimum of the causal effect estimate, use:

$$\min(\mathbf{x}^{(\text{new})}\beta_+^{t*}, \mathbf{x}^{(\text{new})}\beta_-^{t*}) - \max(\mathbf{x}^{(\text{new})}\beta_+^{c*}, \mathbf{x}^{(\text{new})}\beta_-^{c*}).$$

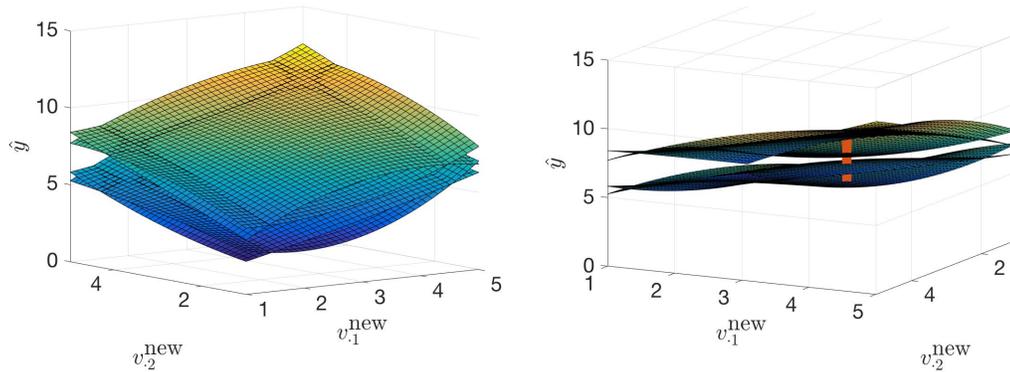
4.2.1 Illustration

We continue with the same data generation process we used in Section 4.1.4, where the ground truth outcomes are created as follows:

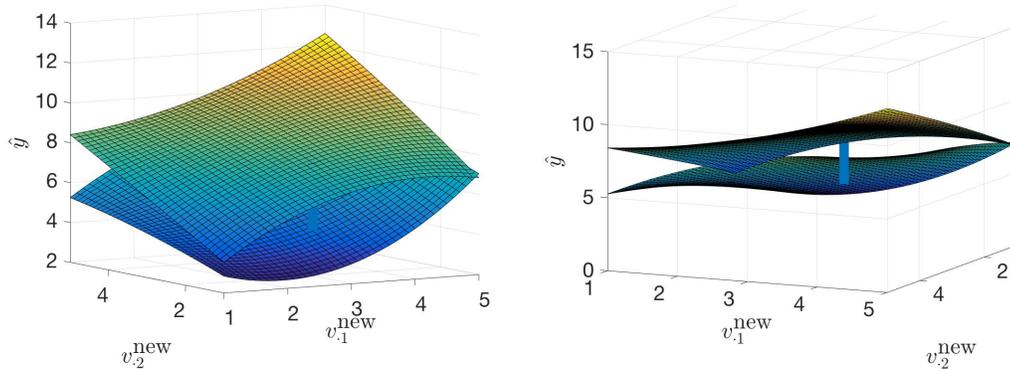
$$y_i = 2 \times 1_{[\text{treated}]} + v_{i1} + v_{i2} + \epsilon$$

. We chose $\mathbf{x}^{(\text{new})}$ to be created from the point $v_1^{\text{new}} = 3$, $v_2^{\text{new}} = 2$. Here we created four separate regressions. One regression maximizes the expected outcome at $\mathbf{x}^{(\text{new})}$ for the treatment observations. Another regression minimizes the expected outcome at $\mathbf{x}^{(\text{new})}$ on the treatment observations. Analogous regressions are created for the control observations. Figure 4.2 shows these regressions explicitly for $\mathbf{x}^{(\text{new})} = \{3, 2\}$. One can see the regressions starting to bend away from each other at $\mathbf{x}^{(\text{new})}$ for the maximization problem, and bend towards each other for the minimization problem. We placed a blue line between the curves at the point $\mathbf{x}^{(\text{new})}$.

All four models: max and min at $\mathbf{x}^{(\text{new})}$ of regressions for treatment and control



Max of treatment and min of control at $\mathbf{x}^{(\text{new})}$. Vertical line drawn at $\mathbf{x}^{(\text{new})}$



Min of treatment and max of control at $\mathbf{x}^{(\text{new})}$. Vertical line drawn at $\mathbf{x}^{(\text{new})}$

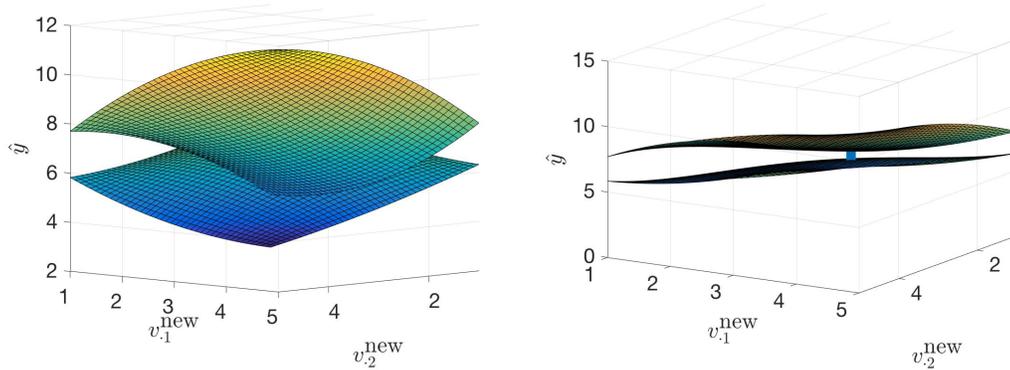


FIGURE 4.2: For all rows, the left and right figures are two different vantage points of the same figure. (Top Row) All four regressions (max/min, treatment/control). (Second Row) Maximizing the gap between treatment and control at $\mathbf{x}^{(\text{new})}$. The upper curve is the regression for maximizing expected outcomes on the treated at $\mathbf{x}^{(\text{new})}$. The lower curve is the regression for minimizing expected outcomes on the control units at $\mathbf{x}^{(\text{new})}$. One can see how the curves pull away from each other at $\mathbf{x}^{(\text{new})}$ to make the differences between treatment and control as large as possible. (Third Row) Minimizing the gap between treatment and control at $\mathbf{x}^{(\text{new})}$. The upper curve is the regression for minimizing outcomes on the treatment units at $\mathbf{x}^{(\text{new})}$. The lower curve is the regression for maximizing the control outcomes at $\mathbf{x}^{(\text{new})}$. Here the curves pull towards each other to minimize the estimated treatment effect. Code written by Cynthia Rudin.

Application: Recidivism Prediction

Understanding the potential impact of researcher choices on machine learning methods becomes especially important when issues of fairness are involved. Although there does not exist a widely accepted mathematical definition of fairness when assessing risk with machine learning (Berk, 2016), if a machine learning method could reach opposing conclusions about a person or group of persons were small adjustments to a dataset or hyperparameters made, then this could potentially undermine any definition of fairness (one could simply argue a negative decision to be unfair because an equally good model exists that predicts the opposite). A hacking interval quantifies the degree to which this can happen.

In the criminal justice system, algorithms are increasingly being used to make risk assessments about defendants, for example their risk of failing to appear in court or reoffending. Clearly, issues of fairness are involved. One such algorithm is COMPAS (or, Correctional Offender Management Profiling for Alternative Sanctions), created by Northpointe Inc. (now Equivant). COMPAS produces three decile scores that indicate the risk that a defendant will fail to reappear in court, reoffend, or violently reoffend. As of October 2017 it was used by 4 of 58 counties in California (Back

et al., 2017). It is a proprietary algorithm that bases its assessment on a questionnaire that is either pulled from criminal records or answered by the defendant. The data gathered by the questionnaire is not publicly available. ProPublica assembled COMPAS scores and other data — including criminal history and demographic information — on more than 7,000 defendants in Broward County, Florida, from 2013 through 2014 with the help of the Broward County Sheriff’s Office . Using the same metric used by Northpointe — whether or not a defendant was charged with a crime within two years of the COMPAS score calculation — ProPublica concluded that COMPAS was biased against African Americans. For example, they found that of African American defendants who did not reoffend, 45 percent were misclassified as higher risk, while of Caucasian defendants who did not reoffend, only 23 percent were misclassified as higher risk (Angwin et al., 2016). Northpointe has issued a rebuttal that argues a definition of fairness based on a false-positive rate is not appropriate in this case (Dieterich et al., 2016). Angelino et al. (2017) and Fisher et al. (2017) argued that since African Americans tend to have longer criminal histories, then as long as criminal history is permitted, COMPAS may depend only on criminal history and not race, which agrees with the sentiment of other work on interpretable models for recidivism (Zeng et al., 2017).

In our analysis, we use the data collected by ProPublica, but our interest is not in comparing a risk assessment score like COMPAS against a given definition of fairness. Rather, we are interested in the impact that researcher choices could have on conclusions made about this dataset. In Section 5.1, we use the methods of Section 2.1.3 to assess the impact that a new feature created by the researcher could have on inferences about the population, in this case the odds ratio of reoffending and gender. This is an example of a prescriptively constrained hacking interval since we explicitly constraint researcher choices about the new feature. In Section 5.2, we use the methods of Section 3.1.1 to assess the impact of researcher choices on

the predictions of a support vector machine about individual defendants. This is an example of a tethered hacking interval since we constrain researcher choices only through their impact on the loss function. For both applications we use the following set of features:

- *c_charge_degree_F*: Binary indicator if the most recent charge prior to the COMPAS score calculation is a felony.
- *sex_Male*: Binary indicator if the defendant is male.
- *age_screening*: Age in years at the time of the COMPAS score calculation.
- *age_18_20*, *age_21_22*, *age_23_25*, *age_26_45*, and *age_45*: Binary indicators based on *age_screening* for age groups 18-20, 21-22, 23-25, 26-45, and greater than or equal to 45, respectively.
- *juvenile_felonies_0*, *juvenile_misdemeanors_0*, and *juvenile_crimes_0*: Binary indicators of whether there is more than one juvenile felony, misdemeanor, or crime, respectively. We use binary indicators because the counts of each are highly right-skewed.
- *priors_0*, *priors_1*, *priors_2_3*, and *priors_3*: Binary indicators of whether the number of priors is 0, 1, 2-3, or more than 3, respectively.

We filtered the dataset to include only defendants whose most recent charge prior to the COMPAS score calculation was a felony or misdemeanor and occurred at most 30 days prior to the COMPAS score calculation (otherwise we assume this charge did not trigger the COMPAS score calculation, so it seems that data about this defendant are missing). The binary indicator variables for age and number of priors were added to the dataset because, in general, recidivism is highly nonlinear with respect to these features.

5.1 Prescriptively Constrained Example: Adding a New Feature

We suppose a researcher is interested in the odds ratio between gender and recidivism but is allowed to create a new binary feature u , perhaps as a function of the existing features or by introducing new data. Notice this is not a valid causal question since gender is not assignable, but we only use the mathematical tools of causal sensitivity analysis. A benefit of this approach is that we do not need to understand exactly what the new feature is, only its relationship to the outcome y (whether or not a defendant reoffends) and “treatment” w (gender). In the setup described in Section 2.1.3, this means the researcher specifies constraints $OR_{yu} \in [a, b]$, $|p_1 - p_0| \leq c$, and $p_0 \geq d$ (by specifying a , b , c and d), where $p_0 := p(U | w = 0)$, $p_1 := p(U | w = 1)$. We will use a simple version where OR_{yu} is fixed (or, equivalently, $a = b = OR_{yu}$). As shown in Section 2.1.3, the hacking interval can be calculated as a function of c .

Figure 5.1 shows hacking intervals for $OR_{yw|xu}$ — the odds ratio between recidivism and gender adjusted for the observed covariates \mathbf{x} and the new feature u — for each combination of $c \in (0.1, 0.15, 0.2, 0.25, 0.3)$ and $OR_{yu} \in (1.5, 1.75)$. These constraints are picked arbitrarily for illustration. In practice, the choice of these constraints describes the degree of freedom given to the researcher. For example, if the researcher were permitted to pick any new binary feature u such that the odds ratio between the outcome and the new feature were $OR_{yu} = 1.5$ and the difference between p_1 and p_0 (the probability of the new feature when the treatment w is present or not present, respectively) were constrained to be less than or equal to $c = 0.3$, then the value of $OR_{yw|xu}$ they could get would necessarily be in the hacking interval $[1.03, 1.37]$. For the same restriction of $c = 0.3$, if the researcher were permitted to pick u such that $OR_{yu} = 1.75$, indicating a stronger relationship between the new feature and the outcome, then they could obtain a value of $OR_{yw|xu}$ above or below one, since Figure 5.1 shows the hacking interval in this case overlaps with one. In

other words, with this freedom given to the researcher, they could conclude that the odds ratio between recidivism and gender, after controlling for measured covariates and the new covariate they created, could be above or below one.

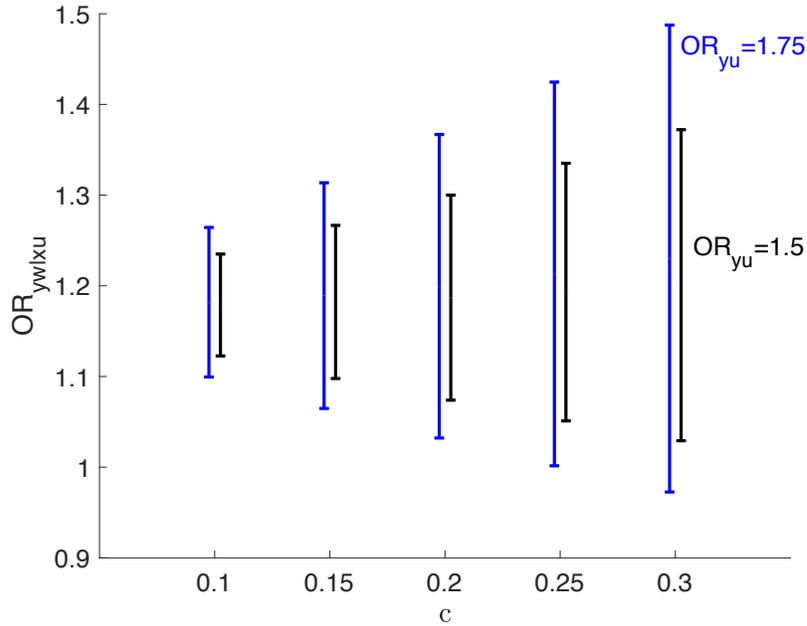


FIGURE 5.1: Hacking intervals for $OR_{yw|xu}$ for different values of constraints c and $a = b = OR_{yu}$.

5.2 Tethered Example: SVM

We now consider the impact of researcher hacking on predictions of two year recidivism for individual defendants. We use a support vector machine (SVM) as our predictive model. For prediction on a new defendant represented by $\mathbf{x}^{(new)}$, SVM calculates the distance of $\mathbf{x}^{(new)}$ to the hyperplane that minimizes the hinge loss. If the distance is positive, the model predicts the defendant will reoffend within two years. If the distance is negative, the model predicts the defendant will not reoffend within two years. By adjusting the hyperplane, the tethered hacking interval is the range of distances of $\mathbf{x}^{(new)}$ to the hyperplane that can be achieved within a constraint on the

loss. As discussed in Section 3.1.1, we can find this range of values by solving the dual problem in Equation (3.5) for $s = -1$ and $s = 1$. We do this using the `fmincon` function in MATLAB. We thus solved two optimization problems for each defendant.

Figure 5.2 shows the hacking intervals for 10 selected defendants from each group of COMPAS scores. We included a few individuals highlighted in an article by ProPublica (Angwin et al., 2016) and randomly selected the rest. The loss is constrained to be within 5% of the minimum loss on a group of 1000 defendants randomly selected from the remaining defendants (so, each prediction in Figure 5.2 is out of sample).

Consider three possible cases: (i) The hacking interval is entirely below zero, (ii) the hacking interval is entirely above zero, and (iii) the hacking interval overlaps with zero. In case (i), this means there does not exist an SVM model such that the loss on the 1000 training observations is within 5% of the minimum loss and the model predicts the defendant will reoffend; all “reasonable” models (*i.e.*, within this loss constraint) predict the defendant *will not* reoffend. In case (ii), when the hacking interval is entirely above zero, the interpretation is the same except all reasonable models predict the defendant *will* reoffend. In (iii), when the hacking interval overlaps with zero, then reasonable SVM models exist that make either prediction. Although this is only a sample of the data, notice that of the ten defendants shown here with COMPAS scores of ten — the riskiest possible COMPAS score — eight of them have hacking intervals that overlap with zero. On the other hand, of the ten people shown here with COMPAS scores of one — the least risky COMPAS score — six of them have hacking intervals entirely above zero.

We can also examine the individuals highlighted in the ProPublica article. In the article, several pairs of defendants are highlighted. In each case one defendant received a low COMPAS score despite a significant criminal history, while the other received a high COMPAS score despite a limited criminal history. For example,

James Rivelli and Robert Cannon were both charged with theft, but Rivelli was charged with felony grand theft and possession of heroin, while Cannon was charged with misdemeanor petit theft. In addition, Rivelli had three prior arrests, including for felony aggravated assault and felony grand theft, while Cannon had none. Despite this, Rivelli — who is white — received a low risk COMPAS score of three, while Cannon — who is black — received a medium risk COMPAS score of six. Rivelli later reoffended in Broward County with grand theft again while Cannon did not. Interestingly, the hacking intervals for both defendants overlapped with zero, indicating justifiable SVM models (on our limited feature set) could have made either prediction. The hacking intervals also overlap with zero for the similarly contrasting pair of Bernard Parker and Dylan Fugett, both arrested on drug charges. For the pair of Vernon Prater and Brisha Borden, both arrested on petty theft charges, the more experienced criminal Prater also has a hacking interval that overlaps with zero but we do not have data on Borden. The exception is Mallory Williams, who received a medium risk COMPAS score of six after a DUI arrest and only two prior misdemeanors. Her hacking interval is entirely below zero, meaning no justifiable SVM model would predict she would reoffend in this experiment. She did not reoffend. In general, we see a high degree of uncertainty from SVM models for the individuals discussed in this article.

The counterpart to Mallory Williams in the ProPublica article, Gregory Lugo, illustrates how offense data can be easily misinterpreted. Gregory Lugo was charged with a DUI but had zero priors according to the data we used in our analysis. Not surprisingly, his COMPAS score was low and his hacking interval was entirely below zero. However, ProPublica claimed he had four priors, including three DUIs, and used this as an example of a poorly calibrated COMPAS score. This appears to be a misinterpretation of the data: all of his supposed prior offenses have the same offense date as the offense related to his COMPAS score calculation, so the supposed prior

offenses appear to be re-recordings (perhaps for ordinary bureaucratic reasons) of the same offense.

There are other interesting examples in Figure 5.2. Claudio Tamarez, a 30 year old Caucasian male, received a COMPAS score of 4, which means low risk, following a charge for possession of phentermine and despite 9 priors that included battery on officer. In contrast, his hacking interval was entirely above zero. He ended up not recidivating, though. Daniel Chiswell, a 41-year old (at the time of the COMPAS score calculation) Caucasian male, was assigned a COMPAS score of only one despite being charged with felony possession of heroin and having previously been charged with felony battery on an officer. His hacking interval overlapped with zero, meaning there exists a reasonable SVM model that would have predicted he would reoffend. He was charged again with felony possession of heroin later that year. Valentina Parrish, a 21 year old Caucasian female, was charged with driving under the influence and possession of less than 20 grams of cannabis. She was given a COMPAS score of ten. In contrast, her hacking interval, $[-2.16, 0.50]$, also overlapped with zero but it was skewed towards the negative end. She did not reoffend. There are also examples that illustrate limitations of our limited feature set. Victor Moreno, a 31 year old African American male, received a COMPAS score of 10 despite zero priors. However, the arrest related to his COMPAS score calculation included felony charges of battery, tampering with a victim, tampering with physical evidence, and delivering cocaine. Our SVM model, without access to the content of these charges, not surprisingly gave him a low hacking interval given his lack of prior offenses.

Figures 5.2 and 5.2 show the hacking intervals for every defendant in our dataset with COMPAS scores of three and eight, respectively. The loss constraint is the same as above (within 5% of the minimum loss on the same 1000 defendants). Of the 663 people in our dataset with COMPAS scores of three — a “low risk” score — 75 of them had hacking intervals entirely above zero. Again, this means that,

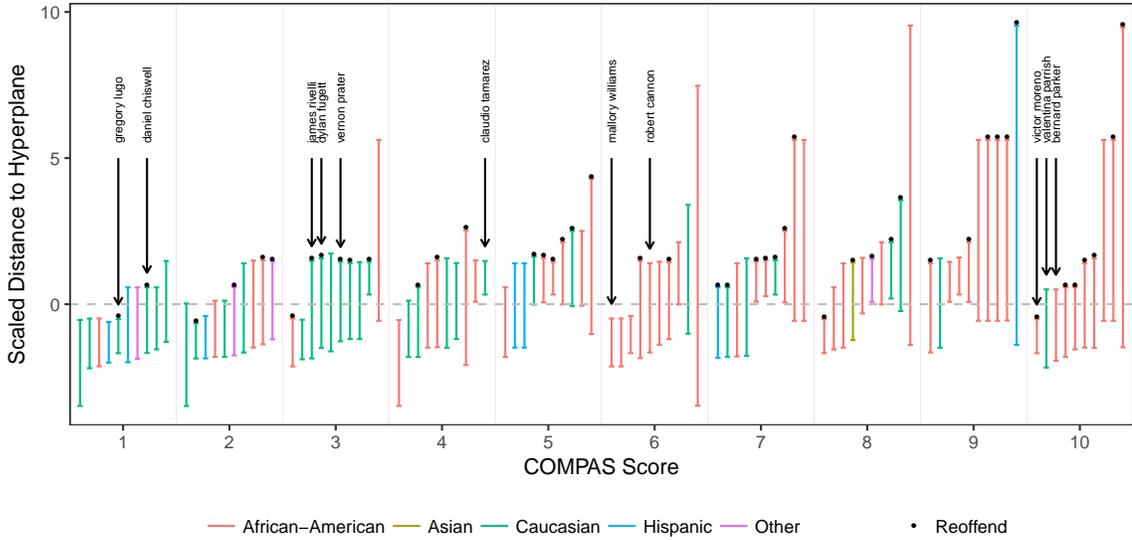


FIGURE 5.2: SVM hacking intervals for 10 defendants for each COMPAS score. Loss is constrained to be within 5% of the minimum loss on a random sample of 1000 defendants.

had SVM been used for prediction, any reasonable model would have predicted that they would reoffend. These 75 people had an average of about 6.3 priors and 35 of them reoffended. Conversely, of the 428 people in our dataset with COMPAS scores of eight — a “high risk” score — 121 of them had hacking intervals entirely below zero, meaning any reasonable SVM model would predict that they would not reoffend. These 121 people had an average of about 8.75 priors and 94 of them reoffended. This potentially means we may be missing data on their past criminal history that is not in the dataset we use for our analysis. While it is possible that missing information can explain COMPAS scores that are high, it cannot explain COMPAS scores that are too low.

We also show hacking intervals grouped by race in Figure 5.2. As before, we allow for a 5% tolerance on the loss on a sample of 1000 defendants, but for this figure we use a different sample of defendants. Each hacking interval in Figure 5.2 is out-of-sample (*i.e.*, the defendant corresponding to the hacking interval was not

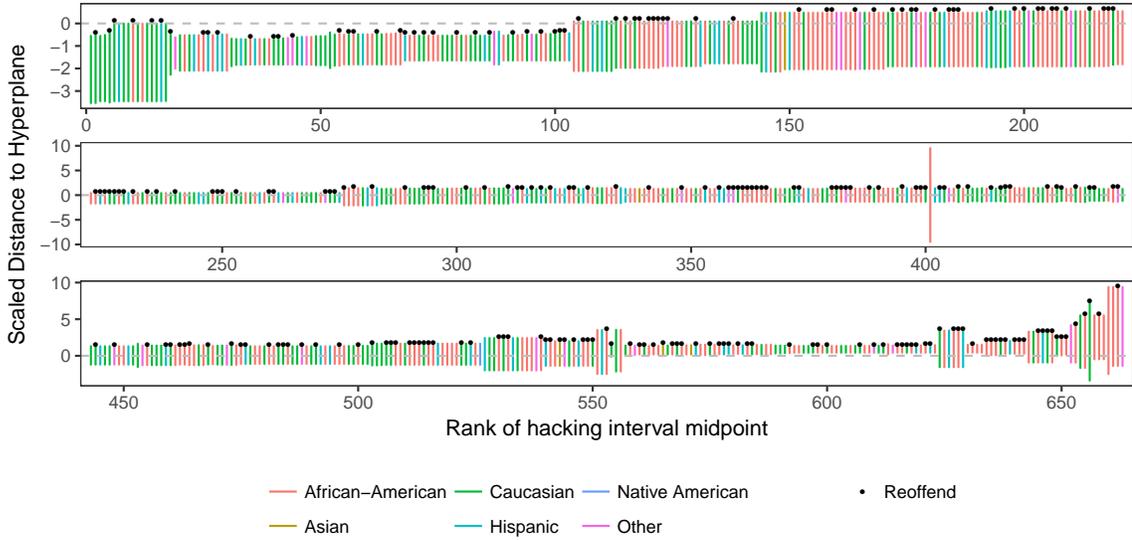


FIGURE 5.3: SVM hacking intervals for all defendants with a COMPAS score of 3. Loss is constrained to be within 5% of the minimum loss on a random sample of 1000 defendants.

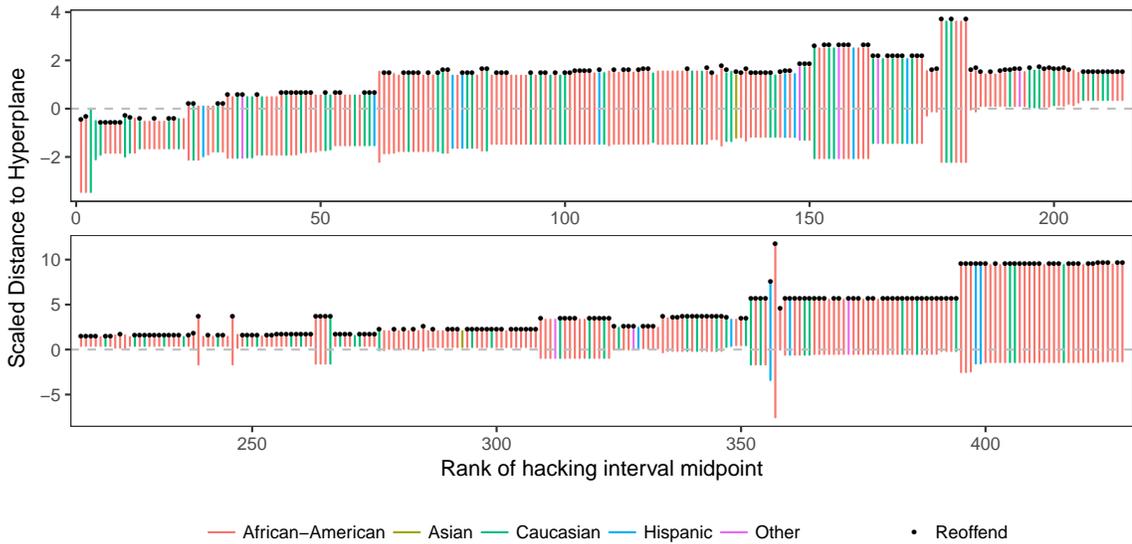


FIGURE 5.4: SVM hacking intervals for all defendants with a COMPAS score of 8. Loss is constrained to be within 5% of the minimum loss on a random sample of 1000 defendants.

included in the 1000 defendant training sample used for the loss constraint). Some of the COMPAS scores again do not align with the hacking intervals. Consider Edwin Chaj, a 27 year-old Hispanic male with only one prior related to trespassing, received a COMPAS score of nine following a charge of disorderly intoxication. In contrast to the high-risk COMPAS score, his hacking interval was low ($[-1.59, 0.24]$), although not entirely below zero. He did not reoffend. Similarly, Cuong Do, a 32 year old Asian male with no priors, received a COMPAS score 8 following charges with felony burglary and misdemeanor petit theft. In contrast to the high-risk COMPAS score, his hacking interval was entirely below zero, meaning no SVM model with less than a 5% loss relative to the minimum loss would predict he would recidivate within two years. He did not recidivate. On the other hand, consider Mories Abdo, a 27 year old Asian male with 6 priors, received a COMPAS score of 3 following a battery charge. In contrast to the low-risk COMPAS, his hacking interval was entirely above zero. He did not recidivate. Figure 5.2 also indicates the individuals discussed in the ProPublica article. Since the 1000 defendant training sample is different from Figure 5.2, the hacking intervals are slightly different, but they are each in the same category (below zero, overlapping with zero, or above zero).

We summarize Figures 5.2 through 5.2 with a couple observations:

- *If we had used SVM on our limited data set rather than the COMPAS score to predict re-offense, then for most people there is enough uncertainty in predictions that we could justifiably predict either reoffend or not reoffend. This can be seen in Figures 5.2 and 5.2, where 75% and 67% of defendants with COMPAS scores of three and eight, respectively, have hacking intervals that overlap with zero, meaning justifiable SVM models exist that could make either prediction. Even for the extreme cases discussed in the ProPublica article, the hacking intervals often overlapped with zero.*

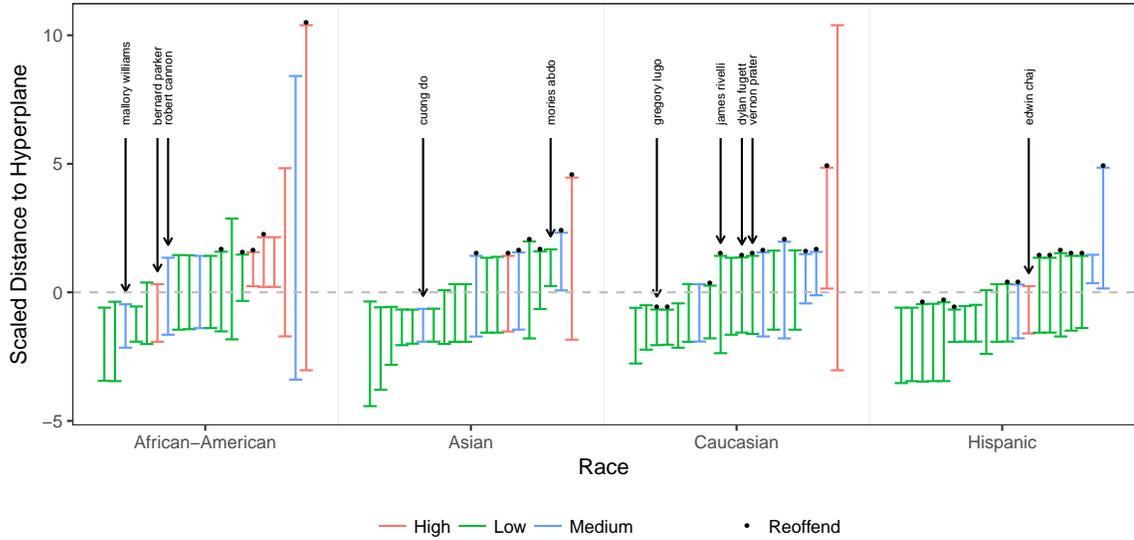


FIGURE 5.5: SVM hacking intervals for 10 randomly selected defendants for each race in the dataset (except Native American as there are only 11 in the dataset). Loss is constrained to be within 5% of the minimum loss on a random sample of 1000 defendants.

- *There are many individuals for which no justifiable SVM model would agree with the COMPAS score using our feature set.* In the case of an individual with a low COMPAS score, this means their hacking interval is entirely above zero, while in the case of an individual with a high COMPAS score, this means their hacking interval is entirely below zero. Figure 5.2 shows 75 examples of the former case and Figure 5.2 shows 121 examples of the latter case.

Related works and Discussion

Hacking intervals are designed to quantify a form of uncertainty that is usually ignored in statistical inference. This could have implications for scientific research; let us discuss this first.

Problems with replication of scientific studies and proposed solutions

The evidence for p -hacking primarily comes from two types of meta-analyses: replication studies, and the distribution of p -values for a set of independent findings (or “ p -curve” Simonsohn et al., 2014). To make matters worse, both findings are contested. For an example of the former approach, a major 2015 study attempted to replicate 100 studies and found very few findings could be reproduced (Open Science Collaboration, 2015), but a replication of this replication found that the percent of studies that were replicated was not statistically different from the fraction that would be expected to replicate due to chance alone (Gilbert et al., 2016). Camerer et al. (2016) found a higher initial percentage being replicable in 18 economic studies, but still reflecting a problem in the field. In commercial applications, large corpo-

rations are keenly aware of this problem: based on their own comparisons, Bayer HealthCare found that only about 20-25% of preclinical studies were completely in line with their in-house findings (Prinz et al., 2011). Amgen replicated 11% of 53 scientific findings (Begley and Ellis, 2012).

For the p -curve approach, a uniform distribution of p -values across articles indicates a lack of significant results, a right-skew indicates a general existence of significant results, and a left-skew, especially near the 0.05 threshold, supposedly indicates p -hacking. Head et al. (2015) concluded that the evidence indicates the existence of “widespread” evidence for p -hacking after searching all Open Access papers in the PubMed database (~100,000 papers). This type of analysis has also been contested (Bishop et al., 2016), and not all meta-analyses have found evidence of p -hacking (Jager and Leek, 2014).

Several types of solutions to p -hacking have been proposed:

- We could require researchers to “pre-register” the details of their study, so that they cannot selectively make choices to achieve significant results, but this rules out learning from the data in any other way.
- Another proposal is to reduce the significance threshold (Monogan, 2015; Humphreys et al., 2012; Simmons et al., 2011; Gelman and Loken, 2013), because when explicitly considering multiple comparisons, decreasing the threshold for significance is sensible (e.g., the Bonferroni correction). Recently, a group of 72 scientists advocated reducing it to 0.005 (J. Benjamin et al., 2018), which might lessen false positives but would also invalidate the quantitative meaning of the p -value in the first place. This is also a drastic measure, leading to a higher true negative rate, and thus many important results being dismissed as insignificant.
- A third proposal is to create Bayesian confidence intervals or Bayesian hi-

erarchical models. In comparison to frequentist hypothesis testing, Bayesian hypothesis testing provides a more comfortable interpretation of the conclusion (the probability that the alternative hypothesis is true), but it is still subject to hacking: the introduction of a prior gives the researcher even more discretion, which may lead to more user choices (see Gelman et al., 2012, for examples of complicated priors leading to bias). If we place a prior on analysts' decisions, it is easy to argue that any given prior is wrong. An example of this, discussed shortly, is the choice of matching algorithm for treatment and control units in a matched pairs experiment. This is a case where uniform priors do not make sense, but any other choice of prior is not defensible either.

- In the case where the researcher does variable selection, *post selection inference* can be used to adjust classical confidence intervals in order to account for the variables being chosen after examining the data. In the case of linear regression, Tibshirani et al. (2016) present a framework for *specific* variable selection procedures (forward stepwise regression, least angle regression, and the lasso regularization path) and Berk et al. (2013) present a framework that holds for *all* variable selection procedures that is more conservative than Scheffé protection (Scheffé, 1959). Hacking intervals differ from post-selection inference in at least two ways: (1) Hacking intervals are more general, as they could include uncertainty to many choices made by the analyst for *any* prediction problem (not just regression), and do not necessarily require *i.i.d.* Gaussian errors; (2) post selection is useful when you already have a model selected and you want to do regular inference, whereas hacking intervals consider robustness to other models that *could* have been selected. Post-selection confidence intervals can be combined with hacking intervals to account for other researcher choices.

These solutions are obviously sometimes useful, but often unfulfilling, highlighting

the importance, inherent difficulty, and urgency of the problem.

Problems with classical inference that are easy to overlook

Here we highlight some drawbacks to classical inference in the way it is used in practice and how hacking intervals can help to fix these issues. Note that while we generally have frequentist or Bayesian inference in mind, there are alternative forms of inference, including fiducial inference (see Hannig et al. (2016) for a review of a modern version).

In cases where a superpopulation exists, the null hypothesis for data analysis is not the correct null hypothesis. The entire confidence interval (CI) calculation for an observed dataset is conditional on statistical assumptions about measurement, distributions, asymptotics, and modeling, among others. Changes in any of these can greatly impact the resulting substantive conclusions, a problem known as *model dependence* (King and Zeng, 2006; Iacus et al., 2011). The null hypothesis used for the analysis depends on the processed data and thus is subject to model dependence. Let us say we want to know whether a pharmaceutical drug causes a side effect. We might process data by choosing covariates, choosing a match assignment, perform regression with a choice of regularization, and so on. The “true” null hypothesis is that the drug does not have any side effect. Instead, the null hypothesis that is actually tested is that the drug has no effect after the researcher’s pre-processing is done to future instances of raw data. It is not clear which pre-processing steps will make the researcher’s null hypothesis close to the true null hypothesis on the correct superpopulation. If the researcher’s results are robust to a range of possible data processing options, then this range may include processing that brings the data closer to a sample drawn from the true superpopulation. To analyze the data in this case, we would want a combination of a hacking interval (for the data processing choices)

and a regular confidence interval (for the processed data) to ensure robustness both to user manipulation and to randomness in the sample of data. We discuss such combinations in Section 4.1.3 for regression. To summarize, hacking intervals help to ensure that the conclusions about the true null hypothesis with respect to the true superpopulation are valid.

It does not make sense to explicitly model analyst choices. In the case of Bayesian model averaging or other decision-making frameworks, one might try to model the way the analyst might treat the data and average over realistic choices an analyst might make. However, this makes little sense. The hypothesis is about the ground truth, not about researcher choices. We would like the result to be robust to *any* choices made by a reasonable researcher.

One could argue that a meaningful prior could be placed on regularization parameters, but there are other analyst choices where there is no meaningful prior. Consider the case of matching in causal inference. When we match treatment and control units and afterwards perform regression to estimate treatment effects, we assume the matched pairs are approximately equivalent to draws from a standardized controlled trial, where the treatment and control units are exactly matched. We might then ignore the uncertainty in the choice of matching procedure, proceed to regression analysis, and report estimated treatment effects. However, an equivalently good match assignment may also exist that yields the opposite results: estimated treatment effects with the opposite sign. (This uncertainty in the matching assignment, unfortunately, is typically ignored in the analysis, despite its key role in the conclusion.) It makes little sense, however, to then model the analysts' choice of matching algorithm. Whether users choose nearest neighbor matching, genetic matching, optimal matching, fast-large-almost-matching-exactly, coarsened exact matching, or matching algorithms that do not exist yet is not relevant to the ground truth. Even if it were, no prior over match assignments (definitely not a uniform distribution)

would reasonably represent match assignments chosen by researchers. A result that *any* reasonable researcher choosing a reasonable match assignment would find the same conclusion is much stronger than saying *one*, arbitrarily chosen match assignment by one researcher yields a particular conclusion. A robust result is also stronger than saying “if users choose match assignments according to an unrealistic prior then the results holds.”

Mathematical equivalence of hacking intervals to other problems, but with different meaning

In some contexts, hacking intervals bear mathematical equivalence to other problems, which means we can leverage existing methods in some cases. Prescriptively-constrained hacking intervals often fall under a form of sensitivity analysis (Leamer, 2010). If we consider uncertainty in the inputs to a mathematical model (usually in an applied-math context), they fall under the field of Uncertainty Quantification. If we consider uncertainty in prior specification, they fall under Robust Bayesian Analysis. If we consider uncertainty in assumptions for causal inference, they fall under (causal) Sensitivity Analysis. See Ghanem et al. (2017), Berger (1994), and Liu et al. (2013) for overviews of these fields, respectively. Uncertainty Quantification provides useful computational tools, like Monte-Carlo simulation and surrogate models (Sudret et al., 2017). In the latter two methods, theoretical bounds on effect estimates have been proven. Berger (1990) determines the range of a posterior quantity priors contained in a certain class. In causal inference we can find the range of effect estimates subject to an unmeasured confounder being within specified bounds on its relationship to both the treatment and the outcome (Lin et al., 1998; Vanderweele and Arah, 2011). If we think of an unmeasured confounder as an additional feature created by a researcher, we can use these results to find the prescriptively-constrained hacking interval under this researcher degree of freedom. We applied

this idea in Section 2.1.3. Tethered hacking intervals are equivalent to profile likelihood confidence intervals (Bjornstad, 1990) when the loss function corresponds to a likelihood. We discussed this in more detail in Section 3.2.

Finding hacking intervals can be viewed as a form of robust optimization. Robust optimization serves as a worst case analysis in decision theory. Uncertainty sets are the primitives for hacking intervals, namely the ranges of user choices we are willing to consider. In prescriptively-constrained hacking intervals, the uncertainty set is the range of prescriptive choices the researcher is allowed to make. In tethered hacking intervals, the uncertainty set is determined by the set of functions achieving low loss (This set is also called the Rashomon Set, see Fisher et al., 2017).

If we cannot easily determine the uncertainty set in advance, we may be able to learn the uncertainty sets from related problems if data (from other sources) are available. This would be similar to Tulabandhula and Rudin (2014), who use machine learning to determine uncertainty sets for decision making.

Teaching of hacking intervals

A major benefit of hacking intervals is that they are easy to explain. Confidence intervals and p -values are difficult to teach and interpret, and are regularly misinterpreted. In response, the American Statistical Association recently issued a document explaining hypothesis testing to users (Wasserstein and Lazar, 2016), and the field of Basic and Applied Social Psychology banned p -values altogether (Trafimow and Marks, 2015), but as the authors of these proposals recognize this does not fully solve the problem.

Hacking intervals are easy to explain, do not require knowledge of probability to understand, and sometimes capture as much, if not more, uncertainty as regular confidence intervals. Teaching hacking intervals first may give a gentle introduction to the effect of uncertainty on conclusions.

6.1 Conclusion

In this work, we presented an alternative theory of inference. It complements existing theories in that it handles a form of uncertainty that arises from analyst choices, rather than from randomness in the data. We presented several examples of hacking intervals stemming from regression and classification, dimension reduction and feature selection. We showed in a real example how hacking intervals can be helpful – in particular, our results indicate that a commonly used model for pre-trial risk analysis may often be incorrectly computed, potentially leading to suboptimal judicial decision making throughout the U.S. Our examples indicate that is possible that these incorrectly computed risk scores could lead (or have led) to dangerous criminals being released prior to trial.

A

Proofs

Proof. (**Proposition 1: Hacking Intervals for SVM**)

Notice that dual problem given by Equation (3.5) is convex. The last two terms are linear, $1/\beta$ is convex, and the coefficient on this term is positive since it is a square. We will assume Slater's condition is satisfied. That is, that there exists a primal solution for which all inequality constraints are strictly satisfied. In this case, the KKT conditions provide necessary and sufficient conditions for optimality. We start by writing down the Lagrangian:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}, \lambda_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \mathbf{r}, \beta) = & s\boldsymbol{\lambda}^T \mathbf{x}^{(\text{new})} + s\lambda_0 - \sum_{i=1}^n \alpha_i [y_i(\boldsymbol{\lambda}^T \mathbf{x}_i + \lambda_0) - 1 + \xi_i] \\ & - \sum_{i=1}^n r_i \xi_i + \beta \left[\frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 + \psi_d \sum_{i=1}^n \xi_i - \theta \right], \end{aligned}$$

where we have introduced dual variables $\{\alpha_i\}_{i=1}^n$, $\{r_i\}_{i=1}^n$, and β .

Next, we check the KKT conditions:

- *Lagrangian stationarity*: $\nabla \mathcal{L} = 0$. So, we have:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = s \mathbf{x}^{(\text{new})} - \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i + \beta^* \boldsymbol{\lambda}^* = 0$$

$$\boldsymbol{\lambda}^* = \frac{1}{\beta^*} \left(-s \mathbf{x}^{(\text{new})} + \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \right), \quad (\text{A.1})$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_0} = s - \sum_{i=1}^n \alpha_i^* y_i = 0 \quad (\text{A.2})$$

$$\sum_{i=1}^n \alpha_i^* y_i = s, \quad (\text{A.3})$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = -\alpha_i^* - r_i^* + \beta^* \psi_d = 0 \quad (\text{A.4})$$

$$r_i^* = \beta^* \psi_d - \alpha_i^*. \quad (\text{A.5})$$

- *Complementary slackness*: There are three conditions:

$$\alpha_i^* [y_i (\boldsymbol{\lambda}^{*T} \mathbf{x}_i + \lambda_0^*) - 1 + \xi_i^*] = 0, \quad \forall i \quad (\text{A.6})$$

$$r_i^* \xi_i^* = 0, \quad \forall i \quad (\text{A.7})$$

$$\beta^* \left[\frac{1}{2} \|\boldsymbol{\lambda}^*\|_2^2 + \psi_d \sum_{i=1}^n \xi_i^* - \theta \right] = 0.$$

- *Dual feasibility*: There are three conditions:

$$\alpha_i^* \geq 0, \quad \forall i \quad (\text{A.8})$$

$$r_i^* \geq 0, \quad \forall i$$

$$\beta^* \geq 0. \quad (\text{A.9})$$

Notice if we plug in the stationarity condition given by Equation (A.5), we have:

$$r_i^* \geq 0 \iff \beta^* \psi_d - \alpha_i^* \geq 0 \iff \alpha_i^* \leq \beta^* \psi_d. \quad (\text{A.10})$$

- *Primal feasibility*: There are three conditions:

$$y_i(\boldsymbol{\lambda}^{*T}x + \lambda_0^*) - 1 + \xi_i^* \geq 0, \quad \forall i$$

$$\xi_i^* \geq 0, \quad \forall i$$

$$\frac{1}{2} \|\boldsymbol{\lambda}^*\|_2^2 + \psi_d \sum_{i=1}^n \xi_i^* - \theta \leq 0.$$

Now, let us simplify the Lagrangian:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}, \lambda_0, \boldsymbol{\xi}, \alpha, r, \beta) &= s\boldsymbol{\lambda}^T \mathbf{x}^{(\text{new})} + s\lambda_0 - \boldsymbol{\lambda} \sum \alpha_i y_i \mathbf{x}_i - \lambda_0 \sum \alpha_i y_i \\ &\quad + \sum \alpha_i - \sum \alpha_i \xi_i - \sum r_i \xi_i + \beta \left[\frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 + \psi_d \sum_{i=1}^n \xi_i - \theta \right] \\ &= \boldsymbol{\lambda}^T \left(s\mathbf{x}^{(\text{new})} - \sum \alpha_i y_i \mathbf{x}_i \right) + \lambda_0 \left\{ s - \sum \alpha_i y_i \right\} \\ &\quad + \sum \alpha_i + \sum \{-\alpha_i - r_i + \beta \psi_d\} \xi_i + \beta \frac{1}{2} \|\boldsymbol{\lambda}\|_2^2 - \beta \theta \\ &= -\beta \boldsymbol{\lambda}^T \left\{ \frac{1}{\beta} \left(-s\mathbf{x}^{(\text{new})} + \sum \alpha_i y_i \mathbf{x}_i \right) \right\} + \lambda_0 \left\{ s - \sum \alpha_i y_i \right\} \\ &\quad + \sum \alpha_i + \sum \{-\alpha_i - r_i + \beta \psi_d\} \xi_i + \beta \frac{1}{2} \boldsymbol{\lambda}^T \boldsymbol{\lambda} - \beta \theta \end{aligned}$$

Next we plug the KKT conditions (A.1), (A.2), and (A.4) into the Lagrangian:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\lambda}^*, \lambda_0^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}^*, \mathbf{r}^*, \beta^*) &= -\beta^* \boldsymbol{\lambda}^{*T} \{\boldsymbol{\lambda}^*\} + \lambda_0^* \{0\} + \sum \alpha_i^* + \sum \{0\} \xi_i^* \\ &\quad + \beta^* \frac{1}{2} \boldsymbol{\lambda}^{*T} \boldsymbol{\lambda}^* - \beta^* \theta \\ &= -\beta^* \frac{1}{2} \boldsymbol{\lambda}^{*T} \boldsymbol{\lambda}^* + \sum \alpha_i^* - \beta^* \theta. \end{aligned}$$

Plugging in Equation (A.1) we have:

$$\begin{aligned}
& \mathcal{L}(\boldsymbol{\lambda}^*, \lambda_0^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}^*, \mathbf{r}^*, \beta^*) \\
&= -\beta^* \frac{1}{2} \left\{ \frac{1}{\beta^*} \left(-s\mathbf{x}^{(\text{new})} + \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \right)^T \frac{1}{\beta^*} \left(-s\mathbf{x}^{(\text{new})} + \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \right) \right\} \\
&\quad + \sum \alpha_i^* - \beta^* \theta \\
&= -\frac{1}{2\beta^*} \left[s^2 \mathbf{x}^{(\text{new})T} \mathbf{x}^{(\text{new})} - 2s \sum \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x}^{(\text{new})} + s^2 \sum_i \sum_k \alpha_i^* \alpha_k^* y_i y_k \mathbf{x}_i^T \mathbf{x}_k \right] \\
&\quad + \sum \alpha_i^* - \beta^* \theta \\
&= -\frac{1}{2\beta^*} \left[\mathbf{x}^{(\text{new})T} \mathbf{x}^{(\text{new})} - 2s \sum \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x}^{(\text{new})} + \sum_i \sum_k \alpha_i^* \alpha_k^* y_i y_k \mathbf{x}_i^T \mathbf{x}_k \right] \\
&\quad + \sum \alpha_i^* - \beta^* \theta.
\end{aligned}$$

Notice $s^2 = 1$, so we can eliminate it. Since the Lagrangian only depends on the dual variables $\boldsymbol{\alpha}$ and β , the dual problem is:

$$\begin{aligned}
& \max_{\boldsymbol{\alpha}, \beta} -\frac{1}{2\beta} \left[\mathbf{x}^{(\text{new})T} \mathbf{x}^{(\text{new})} - 2s \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x}^{(\text{new})} + \sum_i \sum_k \alpha_i \alpha_k y_i y_k \mathbf{x}_i^T \mathbf{x}_k \right] + \sum \alpha_i - \beta \theta \\
& \quad \text{s.t.} \quad \begin{cases} 0 \leq \alpha_i \leq \beta \psi_d, \quad \forall i \\ \sum_{i=1}^n \alpha_i y_i = s \\ \beta \geq 0, \end{cases}
\end{aligned}$$

where the constraints come from Equations (A.8), (A.3), (A.9), and (A.10). Once the optimal $(\boldsymbol{\alpha}^*, \beta^*)$ have been found, we find the optimal $\boldsymbol{\lambda}^*$ from Equation (A.1).

To find λ_0^* we can use the complementary slackness conditions. For some i_{sv} such that $r_{i_{sv}}^* > 0$ and $\alpha_{i_{sv}}^* > 0$, we have $\xi_{i_{sv}}^* = 0$ and $y_{i_{sv}} (\boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}} + \lambda_0^*) - 1 + \xi_{i_{sv}}^* = 0$, by Equations (A.6) and (A.7), respectively. Then, $y_{i_{sv}} (\boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}} + \lambda_0^*) = 1$, so $\lambda_0^* = y_{i_{sv}} - \boldsymbol{\lambda}^{*T} \mathbf{x}_{i_{sv}}$. \square

Proof. (**Proposition 2: Generalization Bound for Hacked Data**)

As a result of Vapnik and Chervonenkis (1981) and the Vapnik-Chervonenkis-Sauer-Shelah lemma (proved independently by Vapnik and Chervonenkis (1971), Sauer (1972), and Shelah (1972)), we have that with probability $1 - \delta$:

$$\left| R_\mu^{\text{true}}(f_p) - R_{\mathbf{Z}_p}^{\text{emp}}(f_p) \right| \leq 2\sqrt{2\frac{h \log \frac{2eh}{n} + \log \frac{4}{\delta}}{n}}$$

Applying the triangle inequality, Equation (A), the assumptions about θ_1 , θ_2 , and θ_3 , and the calculated θ_4 gives:

$$\begin{aligned} \left| R_\mu^{\text{true}}(f_p) - R_{\mathbf{Z}_h}^{\text{emp}}(f_h) \right| &\leq \left| R_\mu^{\text{true}}(f_p) - R_{\mathbf{Z}_p}^{\text{emp}}(f_p) \right| \\ &\quad + \left| R_{\mathbf{Z}_p}^{\text{emp}}(f_p) - R_{\mathbf{Z}_o}^{\text{emp}}(f_p) \right| \\ &\quad + \left| R_{\mathbf{Z}_o}^{\text{emp}}(f_p) - R_{\mathbf{Z}_o}^{\text{emp}}(f_o) \right| \\ &\quad + \left| R_{\mathbf{Z}_o}^{\text{emp}}(f_o) - R_{\mathbf{Z}_o}^{\text{emp}}(f_h) \right| \\ &\quad + \left| R_{\mathbf{Z}_o}^{\text{emp}}(f_h) - R_{\mathbf{Z}_h}^{\text{emp}}(f_h) \right| \\ &\leq 2\sqrt{2\frac{h \log \frac{2eh}{n} + \log \frac{4}{\delta}}{n}} + \sum_{i=1}^4 \theta_i. \end{aligned}$$

□

Proof. (Theorem 3: Hacking Interval for Least-Squares ATE)

Let us rewrite the problems as follows:

$$\max / \min_{\boldsymbol{\beta} \in \mathbb{R}^p, \beta_0 \in \mathbb{R}} \beta_0 \quad \text{such that} \quad g(\boldsymbol{\beta}, \beta_0) - \theta \leq 0,$$

where $g(\boldsymbol{\beta}, \beta_0) = \sum_{i=1}^n (y_i - \boldsymbol{\beta} \mathbf{x}_i - \beta_0 \mathbf{1}_{[i \text{ treated}]})^2$.

Let us write some of the KKT stationarity conditions:

$$\text{for } j = 1 \dots p, \quad \left. \frac{\partial \beta_0}{\partial \beta_j} \right|_{\beta_0^*, \boldsymbol{\beta}^*} = 0,$$

$$\left. \frac{\partial g(\boldsymbol{\beta}, \beta_0)}{\partial \beta_j} \right|_{\beta_0^*, \boldsymbol{\beta}^*} = -2 \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^* - \beta_0^* \mathbf{1}_{[i \text{ treated}]}) x_{ij},$$

that is:

$$\mathbf{0} = \mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}^* - \beta_0^* \mathbf{X}^T \mathbf{1}_{[\text{treated}]}$$

and solving for $\boldsymbol{\beta}^*$:

$$\boldsymbol{\beta}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} - \beta_0^* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{[\text{treated}]}. \quad (\text{A.11})$$

Changing notation in (A.11),

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}_{LS}^* - \beta_0^* (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{[\text{treated}]}, \quad (\text{A.12})$$

where $\boldsymbol{\beta}_{LS}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ is the optimal least square solution. We do not yet know β_0^* , and for that, we will use complementary slackness.

$$\mu g(\boldsymbol{\beta}^* \beta_0^*) = 0 \rightarrow g(\boldsymbol{\beta}^* \beta_0^*) = 0, \text{ that is, } \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta}^* - \beta_0^* \mathbf{1}_{[i \text{ treated}]})^2 - \theta = 0. \quad (\text{A.13})$$

Equations (A.12) and (A.13) will suffice to find all solutions. Substituting (A.12) into (A.13),

$$\begin{aligned} 0 &= \sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^* - \beta_0^* [\mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{[\text{treated}]} - \mathbf{1}_{[i \text{ treated}]}])^2 - \theta, \\ 0 &= \sum_i (d_i + \beta_0^* h_i)^2 - \theta, \end{aligned}$$

where we have defined differences $d_i := y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*$ and $h_i := \mathbf{x}_i (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{[\text{treated}]} - \mathbf{1}_{[i \text{ treated}]}$. Continuing,

$$0 = \left(\sum_i d_i^2 - \theta \right) + 2\beta_0^* \left(\sum_i h_i d_i \right) + \beta_0^{*2} \sum_i h_i^2.$$

Thus, using the quadratic formula:

$$\begin{aligned} \beta_0^* &= \frac{-2 \sum_i h_i d_i \pm \sqrt{(2 \sum_i h_i d_i)^2 - 4(\sum_i h_i^2)((\sum_i d_i^2) - \theta)}}{2 \sum_i h_i^2} \\ &= -\frac{\mathbf{h}^T \mathbf{d}}{\|\mathbf{h}\|^2} \pm \frac{1}{\|\mathbf{h}\|} \sqrt{\frac{(\mathbf{h}^T \mathbf{d})^2}{\|\mathbf{h}\|^2} - \|\mathbf{d}\|^2 + \theta}. \end{aligned}$$

Now, suppose θ were set to the SSE. Then the contents of the square root must become 0, since if this were not true, the solution to the robust optimization problem would disagree with the solution to the least squares minimization problem in the case where θ is the SSE, which would be a contradiction. In that case, we are back to the least square solution, which must be both $\tilde{\beta}_{0,LS}^*$ and $-\frac{\mathbf{h}^T \mathbf{d}}{\|\mathbf{h}\|^2}$. Next, setting the contents of the square root to 0 and solving for θ (which was set to the SSE), we find $\theta = \|\mathbf{d}\|^2 - \frac{(\mathbf{h}^T \mathbf{d})^2}{\|\mathbf{h}\|^2}$. Putting this together we have:

$$\beta_0^* = \beta_{LS}^* \pm \frac{1}{\|\mathbf{h}\|} \sqrt{\theta - \text{SSE}}. \quad (\text{A.14})$$

Notice we can write V_{tt} , the diagonal entry corresponding to the treatment variable of $[\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}]^{-1}$, as follows:

$$\begin{aligned} V_{tt} &:= \left([\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}]^{-1} \right)_{tt} \\ &= \left(\left[\begin{bmatrix} \mathbf{X}^T \\ \mathbf{1}_{\text{treated}}^T \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{1}_{\text{treated}} \end{bmatrix} \right]^{-1} \right)_{tt} \\ &= \left(\left[\begin{array}{cc} \mathbf{X}^T \mathbf{X} & \mathbf{X}^T \mathbf{1}_{\text{treated}} \\ (\mathbf{X}^T \mathbf{1}_{\text{treated}})^T & \mathbf{1}_{\text{treated}}^T \mathbf{1}_{\text{treated}} \end{array} \right]^{-1} \right)_{tt} \\ &= \left[\begin{array}{cc} (\mathbf{X}^T \mathbf{X})^{-1} + \frac{1}{k} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{\text{treated}} \mathbf{1}_{\text{treated}}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} & \frac{1}{k} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{\text{treated}} \\ \frac{1}{k} \mathbf{1}_{\text{treated}}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} & \frac{1}{k} \end{array} \right]_{tt} \end{aligned} \quad (\text{A.15})$$

$$= \frac{1}{k}, \quad (\text{A.16})$$

where $k = \mathbf{1}_{\text{treated}}^T \mathbf{1}_{\text{treated}} - \mathbf{1}_{\text{treated}}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{\text{treated}} = \mathbf{1}_{\text{treated}}^T (\mathbf{I}_n - \mathbf{H}) \mathbf{1}_{\text{treated}}$, $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is the hat matrix, and \mathbf{I}_n is an identity matrix of size n . Equation (A.15) follows from a common block matrix inversion formula (see Proposition

2.8.7 in Bernstein (2005), for example). Next, notice that $\|\mathbf{h}\|^2$ simplifies to:

$$\begin{aligned}
\mathbf{h}^T \mathbf{h} &= -\mathbf{1}_{\text{treated}} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{\text{treated}} + \mathbf{1}_{\text{treated}}^T \mathbf{1}_{\text{treated}} \\
&= \mathbf{1}_{\text{treated}}^T (\mathbf{I}_n - \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{1}_{\text{treated}} \\
&= \mathbf{1}_{\text{treated}}^T (\mathbf{I}_n - \mathbf{H}) \mathbf{1}_{\text{treated}} \\
, &= k
\end{aligned} \tag{A.17}$$

Therefore, from Equations (A.16) and (A.17), we have $V_{tt} = 1/\|\mathbf{h}\|^2$. Plugging this result into Equation (A.14) we have the desired result for $\beta_{0,\max}$ and $\beta_{0,\min}$:

$$\begin{aligned}
\beta_{0,\max}^* &= \tilde{\beta}_{0,LS}^* + \sqrt{V_{tt}} \sqrt{\theta - \text{SSE}} \\
\beta_{0,\min}^* &= \tilde{\beta}_{0,LS}^* - \sqrt{V_{tt}} \sqrt{\theta - \text{SSE}}.
\end{aligned}$$

Finally, defining $\boldsymbol{\gamma}_{LS}^* := (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{1}_{[\text{treated}]}$ as the optimal least square solution from regressing $\mathbf{1}_{[\text{treated}]}$ on \mathbf{X} , from Equation (A.12) we have:

$$\begin{aligned}
\beta_{\max}^* &= \beta_{LS}^* - \beta_{0,\max}^* \boldsymbol{\gamma}_{LS}^*, \\
\beta_{\min}^* &= \beta_{LS}^* - \beta_{0,\min}^* \boldsymbol{\gamma}_{LS}^*.
\end{aligned}$$

□

***Proof.* (Theorem 4: Relating Least-Squares ATE Hacking Intervals to Standard Confidence Intervals)**

The boundary points of a standard confidence interval are:

$$\begin{aligned}
\mathbf{x}^{(\text{new})} \boldsymbol{\beta}_{LS}^* &\pm t_{(1-\alpha/2), (n-p-1)} \sqrt{\text{MSE}(\mathbf{x}^{(\text{new})} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^{(\text{new})T})} \\
\mathbf{x}^{(\text{new})} \boldsymbol{\beta}_{LS}^* &\pm t_{(1-\alpha/2), (n-p-1)} \sqrt{\text{SSE}/(n-p-1)} \sqrt{\mathbf{x}^{(\text{new})} \Upsilon}
\end{aligned}$$

By Theorem 6, the boundary points of the robust interval are:

$$\begin{aligned}
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \tilde{\mu}\mathbf{x}^{(\text{new})}\Upsilon \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \frac{\sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|}}{\|\mathbf{X}\Upsilon\|}\mathbf{x}^{(\text{new})}\Upsilon \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \frac{\sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|}}{\sqrt{(\mathbf{X}\Upsilon)^T(\mathbf{X}\Upsilon)}}\mathbf{x}^{(\text{new})}\Upsilon \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \frac{\sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|}}{\sqrt{(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{(\text{new})T})^T(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{(\text{new})T})}}\mathbf{x}^{(\text{new})}\Upsilon \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \frac{\sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|}}{\sqrt{\mathbf{x}^{(\text{new})}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^{(\text{new})T}}}\mathbf{x}^{(\text{new})}\Upsilon \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \frac{\sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|}}{\sqrt{\mathbf{x}^{(\text{new})}\Upsilon}}\mathbf{x}^{(\text{new})}\Upsilon \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|}\sqrt{\mathbf{x}^{(\text{new})}\Upsilon} \\
\mathbf{x}^{(\text{new})}\boldsymbol{\beta}_{LS}^* &\pm \sqrt{\theta - \text{SSE}}\sqrt{\mathbf{x}^{(\text{new})}\Upsilon}.
\end{aligned}$$

Comparing the standard and robust confidence intervals, we have:

$$\begin{aligned}
t_{(1-\alpha/2), (n-p-1)}\sqrt{\text{SSE}/(n-p-1)} &= \sqrt{\theta - \|Y - \mathbf{X}\boldsymbol{\beta}_{LS}^*\|} \\
\theta &= \text{SSE} \left(1 + \frac{t_{(1-\alpha/2), (n-p-1)}^2}{n-p-1} \right).
\end{aligned}$$

□

Proof. **(Theorem 5: Variance of Least-Squares ATE Hacking Interval Bounds)**

Let $\beta_{0,s}^* = \tilde{\beta}_{0,LS}^* + s\sqrt{V_{tt}}\sqrt{\theta - \text{SSE}}$, where $s = 1$ gives $\beta_{0,\max}^*$ and $s = -1$ gives $\beta_{0,\min}^*$ as defined by Equations (4.5) and (4.3), respectively. It is well-known that for linear regression, the maximum likelihood estimates $\tilde{\boldsymbol{\beta}}_{LS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\mathbf{Y}$ and $\tilde{\sigma}_{LS}^2 = \frac{1}{n}(\mathbf{Y} - \tilde{X}\tilde{\boldsymbol{\beta}}_{LS})^T(\mathbf{Y} - \tilde{X}\tilde{\boldsymbol{\beta}}_{LS})$ for $\boldsymbol{\beta}$ and σ^2 , respectively, have the following properties:

- *Property 1:* $\tilde{\beta}_{LS} \sim N(\boldsymbol{\beta}, (\tilde{X}^T \tilde{X})^{-1} \sigma^2)$.
- *Property 2:* $\frac{n\tilde{\sigma}_{LS}^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} \sim \chi_{n-p-1}^2$. Consequently, $\sqrt{\text{SSE}/\sigma}$ has a chi distribution with mean μ given by Equation (4.9) for $n - p - 1$ degrees of freedom.
- *Property 3:* $\tilde{\beta}_{LS}$ and $\tilde{\sigma}_{LS}^2$ are independent conditional on $\tilde{\mathbf{X}}$. Consequently, $\tilde{\beta}_{0,LS}^*$ and $s\sqrt{rV_{tt}}\sqrt{n\tilde{\sigma}_{LS}^2} = s\sqrt{rV_{tt}}\sqrt{\text{SSE}}$ are also independent for fixed $\tilde{\mathbf{X}}$.

Therefore:

$$\begin{aligned}
\mathbb{V} \left[\beta_{0,s}^* \mid \tilde{\mathbf{X}} \right] &= \mathbb{V} \left[\tilde{\beta}_{0,LS}^* + s\sqrt{V_{tt}}\sqrt{\theta - \text{SSE}} \mid \tilde{\mathbf{X}} \right] \\
&= \mathbb{V} \left[\tilde{\beta}_{0,LS}^* + s\sqrt{V_{tt}}\sqrt{(1+r)\text{SSE} - \text{SSE}} \mid \tilde{\mathbf{X}} \right] \\
&= \mathbb{V} \left[\tilde{\beta}_{0,LS}^* + s\sqrt{rV_{tt}}\sqrt{\text{SSE}} \mid \tilde{\mathbf{X}} \right] \\
&= \mathbb{V} \left[\tilde{\beta}_{0,LS}^* \right] + \mathbb{V} \left[s\sqrt{rV_{tt}}\sqrt{\text{SSE}} \mid \tilde{\mathbf{X}} \right] \tag{A.18}
\end{aligned}$$

$$= \sigma^2 V_{tt} + s^2 r V_{tt} \mathbb{V} \left[\sqrt{\text{SSE}} \mid \tilde{\mathbf{X}} \right] \tag{A.19}$$

$$\begin{aligned}
&= \sigma^2 V_{tt} + r \sigma^2 V_{tt} \mathbb{V} \left[\sqrt{\text{SSE}/\sigma^2} \mid \tilde{\mathbf{X}} \right] \\
&= \sigma^2 V_{tt} + r V_{tt} (n - p - 1 - \mu^2) \tag{A.20}
\end{aligned}$$

$$= \sigma^2 V_{tt} (1 + r(n - p - 1 - \mu^2)),$$

where μ is given by Equation (4.9). Equation (A.18) follows from Property 3, the first term in Equation (A.19) follows from Property 1, and Equation (A.20) follows from the variance formula for a chi distribution. Notice the final result does not depend on s , so it holds for both $\beta_{0,\max}^*$ and $\beta_{0,\min}^*$.

□

Proof. **(Theorem 6: Hacking Intervals for Least-Squares Individual TE)**

Starting again with the stationarity conditions:

$$\nabla(\mathbf{x}^{(\text{new})}\boldsymbol{\beta}) = \mu \nabla \left[\sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta})^2 - \theta \right],$$

evaluating the gradients with respect to β_j , we know that the optimal solution $\boldsymbol{\beta}^*$ obeys

$$x_j^{(\text{new})} = 2\mu \left[\sum_i (y_i - \mathbf{x}_i \boldsymbol{\beta}^*) (-x_{ij}) \right]$$

and the full gradients in vector form are:

$$\mathbf{x}^{(\text{new})} = 2\mu(\mathbf{X}^T \mathbf{X} \boldsymbol{\beta}^* - \mathbf{X}^T \mathbf{Y}),$$

Solving for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}^* = \frac{1}{2\mu} [(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^{(\text{new})T}] + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \tilde{\mu} \Upsilon + \boldsymbol{\beta}_{LS}^*$$

where $\boldsymbol{\beta}_{LS}^*$ is the optimal least squares solution, $(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, we defined $\tilde{\mu} = 1/(2\mu)$, and $\Upsilon = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}^{(\text{new})}$. Again using complementary slackness,

$$\mu \left(\sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta}^*)^2 - \theta \right) = 0 \rightarrow \left(\sum_{i=1}^n (y_i - \mathbf{x}_i \boldsymbol{\beta}^*)^2 - \theta \right) = 0.$$

Substituting from (A.21),

$$\sum_{i=1}^n (y_i - \tilde{\mu} \mathbf{x}_i \Upsilon - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*)^2 - \theta = 0, \text{ and simplifying yields}$$

,

$$0 = \left[\sum_i (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*)^2 - \theta \right] - 2\tilde{\mu} \sum_i (\mathbf{x}_i \Upsilon) (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*) + \tilde{\mu}^2 \sum_i (\mathbf{x}_i \Upsilon)^2$$

$$0 = \left[\sum_i (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*)^2 - \theta \right] - 2\tilde{\mu} V + \tilde{\mu}^2 \sum_i (\mathbf{x}_i \Upsilon)^2,$$

where we let $V = \sum_i (\mathbf{x}_i \Upsilon) (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*)$. Notice it is equal to zero:

$$\begin{aligned}
V &= \sum_i (\mathbf{x}_i \Upsilon) (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*) \\
&= (X \Upsilon)^T (Y - X \boldsymbol{\beta}_{LS}^*) \\
&= (X (X^T X)^{-1} \mathbf{x}^{(\text{new})T})^T (Y - X \boldsymbol{\beta}_{LS}^*) \\
&= (\mathbf{x}^{(\text{new})} (X^T X)^{-1} X^T) (Y - X \boldsymbol{\beta}_{LS}^*) \\
&= \mathbf{x}^{(\text{new})} (X^T X)^{-1} X^T Y - \mathbf{x}^{(\text{new})} (X^T X)^{-1} (X^T X) \boldsymbol{\beta}_{LS}^* \\
&= \mathbf{x}^{(\text{new})} \boldsymbol{\beta}_{LS}^* - \mathbf{x}^{(\text{new})} \boldsymbol{\beta}_{LS}^* \\
&= 0.
\end{aligned}$$

Therefore, the quadratic formula yields:

$$\tilde{\mu} = \frac{\pm \sqrt{-4 (\sum_i (\mathbf{x}_i \Upsilon)^2) [\sum_i (y_i - \mathbf{x}_i \boldsymbol{\beta}_{LS}^*)^2 - \theta]}}{2 \sum_i (\mathbf{x}_i \Upsilon)^2}.$$

Simplifying, abusing notation by letting $\tilde{\mu}$ be the positive solution, and plugging these solutions back into (A.21) yields the result:

$$\boldsymbol{\beta}_+^* = \boldsymbol{\beta}_{LS}^* - \tilde{\mu} \Upsilon, \quad \boldsymbol{\beta}_-^* = \boldsymbol{\beta}_{LS}^* + \tilde{\mu} \Upsilon.$$

□

Bibliography

- Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. (2017), “Learning Certifiably Optimal Rule Lists for Categorical Data.” *arXiv:1704.01701v2*.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016), “Machine Bias,” .
- Back, B. J., Rodriguez, L. R., Boessenecker, M., Calvo, A., Castro, A., Chittick, H. A., Eskin, G. C., Gordon, S. M., Jackson, T. L., McCabe, B. L., Murillo, S. R., and Pichon, R. J. (2017), “Pretrial Detention Reform — Recommendations to the Chief Justice,” Tech. rep., Judicial Branch of California.
- Banaji, M. R. and Greenwald, A. G. (2013), *Blindspot: Hidden biases of good people*, Random House LLC.
- Begley, C. G. and Ellis, L. M. (2012), “Raise standards for preclinical cancer research,” *Nature*, 483, 531–533.
- Berger, J. O. (1990), “Robust Bayesian Analysis: Sensitivity to the Prior,” *Journal of Statistical Planning and Inference*, 25, 303–328.
- Berger, J. O. (1994), “An Overview of Robust Bayesian Analysis,” *Test*, 3, 5–124.
- Berk, R. (2016), “A Primer on Fairness in Criminal Justice Risk Assessments,” *Working Paper No. 2016-5.0*.
- Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), “Valid post-selection inference,” *The Annals of Statistics*, 41, 802–837.
- Bernstein, D. S. (2005), *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*, Princeton University Press.
- Bishop, D. V. M., Chen, J., and Thompson, P. A. (2016), “Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value,” *PeerJ*.
- Bjornstad, J. F. (1990), “Predictive Likelihood: A Review,” *Statistical Science*, 5, 242–254.

- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016), “Evaluating replicability of laboratory experiments in economics,” *Science*, 351, 6050–6051.
- Cornfield, J., Haenszel, W., and Hammond, E. C. (1959), “Smoking and lung cancer: recent evidence and a discussion of some questions.” *Journal of the National Cancer Institute*, 22, 173–203.
- Dieterich, W., Mendoza, C., and Brennan, T. (2016), “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity,” Tech. rep., Northpointe.
- Ding, P. and VanderWeele, T. J. (2016), “Sensitivity Analysis Without Assumptions,” *Epidemiology*, 27, 368–377.
- Fisher, A., Rudin, C., and Dominici, F. (2017), “Model Class Reliance: Variable Importance Measures for any Machine Learning Model Class, from the “Rashomon” Perspective,” *arXiv:1801.01489v2*.
- Gelman, A. and Loken, E. (2013), “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,” .
- Gelman, A. and Shalizi, C. R. (2013), “Philosophy and the practice of Bayesian statistics,” *Br J Math Stat Psychol.*, 66, 8–38.
- Gelman, A., Hill, J., and Yajima, M. (2012), “Why We (Usually) Dont Have to Worry about Multiple Comparisons,” *Journal of Research on Educational Effectiveness*, 5, 189–211.
- Ghanem, R., Higdon, D., and Owhadi, H. (2017), *Handbook of Uncertainty Quantification*, Springer International Publishing.
- Gilbert, D., King, G., Pettigrew, S., and Wilson, T. (2016), “Comment on ‘Estimating the Reproducibility of Psychological Science’,” *Science*, 351, 1037a–1038a.
- Gilbert, D. T. (1998), “Ordinary Psychology,” in *The Handbook of Social Psychology*, eds. D. T. Gilbert, S. T. Fiske, and G. Lindzey, vol. 2, pp. 89–150, McGraw Hill, New York.
- Guo, Q., Wu, W., Massart, D. L., Boucon, C., and De Jong, S. (2002), “Feature selection in principal component analysis of analytical data,” *Chemometrics and Intelligent Laboratory Systems*, 61, 123–132.

- Hannig, J., Iyer, H., Lai, R. C., and Lee, T. C. M. (2016), “Generalized fiducial inference: A review and new results,” *Journal of the American Statistical Association*, 111, 1346–1361.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015), “The Extent and Consequences of P-Hacking in Science,” *PLoS Biology*, 13.
- Hosmer, D., Lemeshow, S., and Sturdivant, R. (2013), *Applied Logistic Regression: Third Edition.*, John Wiley and Sons Inc.
- Humphreys, M., Sanchez De La Sierra, R., and Van Der Windt, P. (2012), “Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration,” *Political Analysis*, pp. 1–20.
- Iacus, S. M., King, G., and Porro, G. (2011), “Multivariate Matching Methods that are Monotonic Imbalance Bounding,” *Journal of the American Statistical Association*, 106, 345–361.
- J. Benjamin, D., O. Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E.-J., Berk, R., Bollen, K., Brembs, B., Brown, L., Camerer, C., Cesarini, D., D. Chambers, C., Clyde, M., D. Cook, T., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., and Johnson, V. (2018), “Redefine Statistical Significance,” *Nature Human Behaviour*, pp. 6–10.
- Jager, L. R. and Leek, J. T. (2014), “An estimate of the science-wise false discovery rate and application to the top medical literature,” *Biostatistics*, 15, 1–12.
- Jeffers, J. N. R. (1967), “Two Case Studies in the Application of Principal Component Analysis,” *Journal of the Royal Statistical Society*, 16, 225–236.
- Jolliffe, I. T. (1972), “Discarding Variables in a Principal Component,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 21, 160–173.
- Kahneman, D. (2011), *Thinking, fast and slow*, Macmillan.
- King, G. (1995), “Replication, Replication,” *PS: Political Science and Politics*, 28, 443–499.
- King, G. and Zeng, L. (2006), “The Dangers of Extreme Counterfactuals,” *Political Analysis*, 14, 131–159.
- Krzanowski, W. J. (1987), “Selection of Variables to Preserve Multivariate Data Structure, Using Principal Components,” *Source Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36, 22–33.
- Leamer, E. E. (2010), “Extreme bounds analysis,” in *Microeconometrics*, pp. 49–52, Springer.

- Lin, D. Y., Psaty, B. M., and Kronmal, R. a. (1998), “Assessing the sensitivity of regression results to unmeasured confounders in observational studies.” *Biometrics*, 54, 948–963.
- Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013), “An Introduction to Sensitivity Analysis for Unobserved Confounding in Non-Experimental Prevention Research,” *Prev Sci*, 14, 570–580.
- Mccabe, G. P. (1984), “American Society for Quality Principal Variables,” *Technometrics*, 26, 137–144.
- Monogan, J. E. (2015), “Research preregistration in political science: The case, counterarguments, and a response to critiques,” *PS: Political Science & Politics*, 48, 425–429.
- Noor-E-Alam, M. and Rudin, C. (2015), “Robust Testing for Causal Inference in Observational Studies,” Working paper.
- Open Science Collaboration (2015), “Estimating the reproducibility of psychological science,” *Science*, 349.
- Prinz, F., Schlange, T., and Asadullah, K. (2011), “Believe it or not: how much can we rely on published data on potential drug targets?” *Nature Reviews*.
- Sauer, N. (1972), “On the Density of Families of Sets,” *Journal Of Combinatorial Theory (A)*, 13, 145–147.
- Scheffé, H. (1959), *The Analysis of Variance*, Wiley, New York.
- Shelah, S. (1972), “A Combinatorial Problem; Stability And Order For Models And Theories In Infinitary Languages,” *Pacific Journal Of Mathematics*, 41.
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011), “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant,” *Psychological Methods*, 22, 1359–1366.
- Simonsohn, U., Nelson, L. D., and Simmons, J. P. (2014), “P-Curve: A Key to the File-Drawer,” *Journal of Experimental Psychology: General*, 143, 534–547.
- Sudret, B., Marelli, S., and Wiart, J. (2017), “Surrogate models for uncertainty quantification: An overview,” *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pp. 793–797.
- Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), “Exact post-selection inference for sequential regression procedures,” *Journal of the American Statistical Association*, 111, 600–620.

- Trafimow, D. and Marks, M. (2015), “Editorial,” *Basic and Applied Social Psychology*, 37, 1–2.
- Tulabandhula, T. and Rudin, C. (2014), “Robust Optimization using Machine Learning for Uncertainty Sets,” in *Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM)*.
- Vanderweele, T. J. and Arah, O. A. (2011), “Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Bias Formulas for Sensitivity Analysis,” *Epidemiology*, 22, 42–52.
- Vapnik, V. N. and Chervonenkis, A. Y. (1971), “On The Uniform Convergence Of Relative Frequencies Of Events To Their Probabilities,” *Theory of Probability and Its Applications*, 16, 264–281.
- Vapnik, V. N. and Chervonenkis, A. Y. (1981), “Necessary And Sufficient Conditions For The Uniform Convergence Of Means To Their Expectations,” *Theory of Probability and Its Applications*, 26, 532–554.
- Wasserstein, R. L. and Lazar, N. A. (2016), “The ASA’s Statement on p-Values: Context, Process, and Purpose,” *The American Statistician*, 70, 129–133.
- Weinberger, K. Q. and Tesauro, G. (2017), “Metric Learning for Kernel Regression,” *Proceedings of Machine Learning Research*, 2, 612–619.
- Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., and van Assen, M. A. (2016), “Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking,” *Frontiers in Psychology*, 7, 1–12.
- Wilks, S. S. (1938), “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses,” *The Annals of Mathematical Statistics*, 9, 60–62.
- Xu, J.-l., Xu, B.-w., Zhang, W.-f., and Cui, Z.-f. (2008), “Principal Component Analysis Based Feature Selection For Clustering Clustering,” *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*, pp. 12–15.
- Zeng, J., Ustun, B., and Rudin, C. (2017), “Interpretable classification models for recidivism prediction.” *Journal of the Royal Statistical Society*, 180, 689–722.