Advances in Model-agnostic Approaches to Statistical Inference

David Whitney

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Marco Carone, Chair

Noah Simon

Alex Luedtke

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Advances in Model-agnostic Approaches
to Statistical Inference

David Whitney

Chair of the Supervisory Committee:
PhD Marco Carone
Department of Biostatistics

This dissertation focuses broadly on contributing to understanding the impact of incorrect modeling assumptions on analyses and arguing for the use of methods for statistical inference that are valid under weaker conditions than required for traditional approaches. By statistical inference, we mean providing both a point estimate for the population value of a parameter as well as valid confidence intervals and hypothesis tests. In settings involving coarsened data or nuisance parameters that are difficult to estimate, model-based approaches to regression can provide misleading results when the model fails to hold. Through the examples of the partially linear additive model and Cox proportional hazards model, we provide guidance for evaluating the properties of model-based regressions and illustrate alternative model-agnostic approaches that avoid undesirable behaviors. We introduce a novel expansion of a remainder term to derive a framework to obtain doubly robust inference for a broad class of parameters. This work extends recent nonparametric methods to achieve doubly robust inference – rather than simply doubly robust estimation – for the average treatment effect specifically. While estimation of quantiles is not much more difficult than for means, construction of confidence intervals presents greater challenges. Hence, we study and evaluate several model-agnostic procedures to obtain confidence regions and hypothesis tests for quantiles.

TABLE OF CONTENTS

Page

List of Figures iii

Chapter 1: Introduction 1
1.1 Contributions 1
1.2 Parameters as mappings on statistical models 2
1.3 Asymptotic linearity, pathwise differentiability, and efficiency 3
1.4 Estimating equations, one-step estimators, and TMLE 6

Chapter 2: Models as deliberate approximations 11
2.1 Model-robust interpretation 12
2.2 Valid inference in the presence of irregular nuisances 17
2.3 Deliberate model-agnostic parameter extensions 23
2.4 Discussion 25

Appendices 28
Appendix 2.A Computing the OLS slopes for Figure 2.1 28
Appendix 2.B Solving the partial likelihood equations for Figure 2.2 30
Appendix 2.C Details of the PLAM nuisance contributions 30
Appendix 2.D Sampling distribution of the one-step estimator 31
Appendix 2.E Sample R code for computing PLAM estimators 37
Appendix 2.F Sample R code for computing the one-step estimator 40

Chapter 3: A General Approach to Doubly Robust Inference 46
3.1 Introduction 46
3.2 Doubly robustness 48
3.3 Approximating estimation remainders 52
3.4 Achieving doubly robust inference under missingness at random 57

3.5 Discussion 67

Appendices 69
Appendix 3.A Proof of Theorem 3.1 69
Appendix 3.B Proof of Lemma 3.1 69
Appendix 3.C Proof of Theorem 3.2 72

Chapter 4: Robust inference for quantiles 73
4.1 Efficiency theory for quantile estimators 74
4.2 Methods for obtaining pointwise confidence intervals for a fixed quantile . . . 79
4.3 Simulations 87
4.4 Discussion 89

Appendices 91
Appendix 4.A Sample R code for constructing confidence intervals 91

Bibliography 98

LIST OF FIGURES

Figure Number Page

2.1 Exposure-specific OLS regression functional value when the true regression model follows a partially linear additive model. The line $\theta_{\text{OLS}} = 1$ (black) denotes the exposure-specific regression coefficient for the true data-generating mechanism. The other lines represent the value of the OLS regression functional for different strengths of the exposure-confounder relationship. 14

2.2 The model-robust interpretation of the partial likelihood regression functional depends on the censoring distribution. For a binary covariate, the population hazard ratio at time t is assumed to be $\alpha t^{\alpha-1}$, with larger α values corresponding to greater departures from the PH model. The exposure group-specific censoring distributions are exponential with rates γ_0 and γ_1. 18

2.3 Empirical bias and standard error of model-based estimators $\theta_{0,n}$, $\theta_{1,n}$ and $\theta_{2,n}$ scaled by $n^{1/2}$ for sample sizes $n \in \{500, 1000, 2000, 3000, 5000\}$ computed using 5000 simulated datasets for each sample size under correct and incorrect model specifications. 22

2.4 Empirical bias and standard error of estimators $\theta_{\text{MPLE},n}$ and $\theta_{\text{OS},n}$ scaled by $n^{1/2}$ for sample sizes $n \in \{500, 1000, 2000, 3000, 5000\}$ computed using 5000 simulated datasets for each sample size under correct and incorrect model specifications. 27

4.1 (A) The marginal distribution function F_0 and (B) marginal density f_0 for Y implied by the data-generating mechanism. The median of Y is plotted in each panel. 87

4.2 Empirical coverage and length of approximate 95% confidence intervals for sample sizes $n \in \{100, 200, 500, 1000, 2000\}$ computed using 1000 simulated datasets. 90

ACKNOWLEDGMENTS

Completing my graduate studies was not always assured. Many people helped me navigate the challenges of the last six years. First I would like to thank my dissertation advisor, Marco Carone, for his guidance, feedback and good humor throughout my time working with him as a student. I look forward to our continued collaborations. I thank my other reading committee members, Noah Simon and Alex Luedtke, for their dedication and for their thought-provoking questions in each of our interactions.

The Biostatistics Graduate Program provided pivotal support to me throughout my time as student. I thank Scott Emerson, Lurdes Inoue, and Gitana Garofalo for their mentorship and for helping me navigate difficult times during the program.

Outside of the university, I received support from my family and friends. My parents Paul and Julie Whitney, as ever, provided their unfailing love and support for my studies. My sister Kelly valiantly subjected herself to my presentations. My parents-in-law David and Alexis Krogh gave Jo and I a place to live while I wrote my dissertation. My good friend Corbin Muck bought me pizza while I wrote my applied qualifying exam. Thank you all.

This work was supported in part by the UW NIEHS sponsored Biostatistics, Epidemiologic and Bioinformatic Training in Environmental Health (BEBTEH) Training Grant, Grant #: NIEHS T32ES015459.

DEDICATION

To my wife, Johanna, who stood with me unfailingly through it all. I look forward to our next adventure together.

Chapter 1
INTRODUCTION

In public health, statistical analyses are often based on parametric or semiparametric regression models. As these models may fail to hold in practice, it is of interest to understand the behavior of model-based regressions in the context of potential misspecification. In many cases, model-agnostic alternatives that avoid the limitations of model-based regression approaches exist. For these model-agnostic approaches, we wish to provide valid statistical inference under the weakest conditions possible. To distinguish between model-based results for estimators, we will make reference to model-agnostic characteristics. By model-agnostic characteristics, we mean properties of a particular estimator (e.g., consistency, sampling distribution) when the model possibly fails to hold.

In section 1 of this chapter, we briefly outline the contributions of this dissertation. In the remaining sections we provide background for relevant statistical concepts and recur throughout the remainder. In section 2, we define the notions of model and parameter. In section 3, we review pathwise differentiability for parameter mappings and discuss the relevance of this property for obtaining inference. In section 4, we discuss three frameworks for deriving asymptotically linear estimators of pathwise differentiable parameters: M-estimators (and Z-estimators), one-step estimators, and targeted minimum loss-based estimators (TMLE).

1.1 Contributions

The remaining chapters focus broadly on understanding the impact of incorrect modeling assumptions on analyses and drawing statistical inference that is valid under weaker conditions than required for traditional approaches. By statistical inference, we mean providing both a point estimate for the population value of a parameter as well as valid confidence intervals

and hypothesis tests. In settings involving coarsened data or nuisance parameters that are difficult to estimate, model-based approaches to regression can provide misleading results when the model fails to hold. In Chapter 2, we provide guidance for evaluating the properties of model-based regressions and propose alternative model-agnostic approaches that avoid these undesirable behaviors. In Chapter 3, we introduce a novel expansion of the estimation remainder to derive a framework to obtain doubly robust inference for a broad class of parameters. This work extends recent nonparametric methods to achieve doubly robust inference – rather than simply doubly robust estimation – for the average treatment effect specifically. As an illustration of our general results, we provide doubly robust inference for quantile treatment effects. While estimation of quantiles is not much more difficult than for means, inference remains challenging. In Chapter 4, we propose and evaluate several procedures to obtain confidence regions and hypothesis tests for quantiles.

1.2 Parameters as mappings on statistical models

It is not uncommon for statements such as, "We fit a regression model of Y on X" to appear in descriptions of statistical analyses. Such statements can muddle the relationship between model and parameter. In our work, we will adopt broader definitions of *model* and *parameter*. The model represents the set of data-generating distributions that are compatible with the scientific knowledge about the experiment, whereas the parameter provides a summary of the data-generating mechanism and is typically of scientific interest.

We denote a typical observed data unit by $Z \in \mathcal{Z} \subset \mathbf{R}^d$ and denote the (statistical) model by $\mathcal{M}$. By the model for Z, we mean a collection (set) of probability measures, P, defined on a common sigma field $\mathcal{A}$ of $\mathcal{Z}$. Hence, for each $P \in \mathcal{M}$, the triple $(\mathcal{Z}, \mathcal{A}, P)$ is a probability space. We consider parameters as mappings $\Psi : \mathcal{M} \to \mathbf{K}$, where $\mathbf{K}$ is often of finite dimension (e.g. vector-valued parameters in $\mathbf{R}^d$) but may be infinite dimensional (e.g. function-valued parameters). Unless otherwise stated, we suppose that the population value $\psi_0 \equiv \Psi(P_0)$ is the target of statistical inference. Regression functions are a particular class of parameters. For $Z = (Y, X) \sim P$, the conditional expectation of Y given $X = x$,

denoted by $\bar{Q}_P(x) = E_P(Y \mid X = x)$, is an example of a regression function. Depending on the model $\mathcal{M}$, regression functions may be completely unspecified, or perhaps (partially) known except for a finite-dimensional index.

1.2.1 Example: generalized linear model

We illustrate how these terms apply in a popular example. For $Z = (Y, X)$, with $Y \in \mathbf{R}$ and $X \in \mathbf{R}^d$, the generalized linear model (GLM) with link function $g : \mathbf{R} \to \mathbf{R}$ and conditional variance $\text{var}_P(Y|X = x) = \sigma_P^2(x)$ consists of all probability measures P such that

$$g(\bar{Q}_P(x)) = \alpha_P + x^\top \beta_P,$$

for $\alpha_P \in \mathbf{R}$ and $\beta_P = (\beta_{P,1}, \ldots, \beta_{P,d})^\top \in \mathbf{R}^d$. Here, the regression function $\bar{Q}_P(x)$ has a form that is known up to the value of $\Psi(P) = (\alpha_P, \beta_{P,1}, \ldots, \beta_{P,d})^\top \in \mathbf{R}^{d+1}$. Essentially, this means that estimating the function $\bar{Q}_P$ is no more difficult than estimating the vector $\Psi(P)$. In a GLM, the β_P component of $\Psi(P)$ may be of further interest as a measure of association between variables in X with the outcome Y. In particular, for g the identity function, the GLM reduces to the linear model and $\beta_{P,j}$ is interpretable as the difference in the conditional mean of Y at $X = x$ and $X = x'$ where $x = (x_1, \ldots, x_j, \ldots, x_d)$ and $x' = (x_1, \ldots, x_j + 1, \ldots, x_d)$:

$$\beta_{P,j} = Q_P(x_1, \ldots, x_j + 1, \ldots, x_d) - Q_P(x_1, \ldots, x_j, \ldots, x_d)$$

where $x_1, \ldots, x_d$ are arbitrary. Model-based regression estimators have been ubiquitous across disciplines because of the interpretability of their parameters as well as the ease with which statistical inference may be performed. In Chapter 2, we consider the impact on inference when $P_0 \notin \mathcal{M}_0$, but model-based estimators that assume $P_0 \in \mathcal{M}_0$ have been implemented.

1.3 Asymptotic linearity, pathwise differentiability, and efficiency

Even in cases in which P_0 is known to be in a regular parametric model indexed by ψ_0, and the maximum likelihood estimator (MLE) ψ_n is used, the exact sampling distribution

of ψ_n is commonly intractable at finite sample sizes. When P_0 belongs to a more complex model, this difficulty remains. Hence, inference for ψ_0 is often based on approximations to the asymptotic sampling distribution of ψ_n.

1.3.1 Asymptotic linearity and Wald-based inference

For independent observations $Z_1, \ldots, Z_n$ identically distributed according to P_0, an estimator ψ_n of population parameter ψ_0 is asymptotically linear with influence function ϕ_{P_0} if it admits the expansion

$$\psi_n - \psi_0 = \frac{1}{n} \sum_{i=1}^{n} \phi_{P_0}(Z_i) + o_p(n^{-1/2}),$$

for a function $\phi_{P_0} : \mathcal{Z} \to \mathbf{K}$ in $L_2(P_0)$, the space of mean-zero, square-integrable functions with respect to the probability measure P_0. This expansion is desirable as it tells us that ψ_n behaves like a sample average and that confidence intervals and hypothesis tests based on a normal approximation to the sampling distribution of ψ_n are asymptotically valid.

We denote weak convergence of a sequence X_n to a tight limit X_0 by $X_n \rightsquigarrow X_0$. A central limit theorem-based argument leads to the conclusion that if ψ_n is asymptotically linear at ψ_0, then

$$n^{1/2}(\psi_n - \psi_0) \rightsquigarrow \mathbb{G}_{P_0}\phi_{P_0}$$

where $\mathbb{G}_{P_0}\phi_{P_0} = \int \phi_{P_0}(z) d\mathbb{G}_{P_0}(z)$ is a Brownian bridge process applied to the influence function ϕ_{P_0}. Considering convergence of $n^{1/2}(\psi_n - \psi_0)$ to a stochastic process is useful when ψ_0 is function-valued, such as the survival function or quantile function. If the parameter values are vectors in $\mathbf{K} = \mathbf{R}^d$, then we find that

$$n^{1/2}(\psi_n - \psi_0) \rightsquigarrow L$$

where L is a d-dimensional multivariate normal random vector with mean zero and covariance matrix given by $\Sigma = E_{P_0}\{\phi_{P_0}(Z)\phi_{P_0}(Z)^\top\}$. In this case, we write $L \sim \mathcal{N}_d(0, \Sigma)$. Note that in this case, we can also say that the sequence $n^{1/2}(\psi_n - \psi_0)$ converges in distribution to L.

The convergence in distribution for asymptotically linear estimators of $\psi \in \mathbf{R}^d$ suggests the use of Wald-based inference, which approximates the sampling distribution of ψ_n at fixed n by $\mathcal{N}_d(\psi_0, n^{-1}\Sigma_n)$ with $\Sigma_n = \frac{1}{n}\sum_{i=1}^n \phi_{P_n}(Z)\phi_{P_n}(Z)^\top$. Letting $a \in \mathbf{R}^d$, the two-sided $(1-\alpha)\times 100\%$ Wald confidence interval for $a^\top\psi_0$ has limits given by the familiar formula

$$a^\top\psi_n \pm z_{1-\alpha/2}\frac{1}{n^{1/2}}\left(a^\top\Sigma_n a\right)^{1/2},$$

where $z_{1-\alpha/2}$ is defined as the $1-\alpha/2$ quantile of a $\mathcal{N}(0,1)$ random variable.

In a nonparametric (unrestricted) model, any two regular asymptotically linear estimators ψ_n and ψ_n' for ψ_0 have the same influence function ϕ_{P_0}. Hence ψ_n and ψ_n' have the same asymptotic sampling distribution. For general semiparametric and parametric models, there exists an infinity of influence functions. This entails potentially different covariances for ψ_n and ψ_n' and motivates comparison of asymptotic relative efficiencies for these estimators. To characterize the collection of influence functions requires a notion of differentiability for the parameter mapping $\Psi : \mathcal{M} \to \mathbf{K}$ over the model $\mathcal{M}$.

1.3.2 Tangent space, pathwise differentiability and the efficient influence function

We consider the set of maps $\epsilon \mapsto P_\epsilon$ that are differentiable in quadratic mean (van der Vaart 1998, Chapter 25), that is, which take $\epsilon \in \mathbf{R}$ to an element of $\mathcal{M}$ such that there exists a function $s : \mathcal{Z} \to \mathbf{K}$ for which

$$\lim_{\epsilon\to 0}\int\left[\frac{p_\epsilon(z)^{1/2} - p_0(z)^{1/2}}{\epsilon} - \frac{1}{2}s(z)p_0(z)^{1/2}\right]^2 d\mu(z) = 0,$$

where p_ϵ and p_0 are densities with respect to some common dominating measure μ. The collection of all functions s arising in this way is called the tangent set at P_0 in the model $\mathcal{M}$. We call the closed linear span of the tangent set the tangent space at P_0 in $\mathcal{M}$, denoted by $\mathcal{T}_\mathcal{M}(P_0)$. The tangent space can be interpreted as all possible score functions at $\epsilon = 0$ for parametric submodels $\{P_\epsilon : \epsilon\} \subset \mathcal{M}$ for which $P_\epsilon = P_0$ when $\epsilon = 0$. As our interest pertains to neighborhoods of $\epsilon = 0$, we index submodels by their score at zero, writing $P_{\epsilon,s}$ for a fixed $s \in \mathcal{T}_\mathcal{M}(P_0)$.

The parameter mapping $\Psi : \mathcal{M} \to \mathbf{K}$ is pathwise differentiable in the direction s if

$$\frac{d}{d\epsilon}\Psi(P_{\epsilon,s})\bigg|_{\epsilon=0} = \int s(z)h_0(z)dP_0(z) \tag{1.1}$$

for $s \in \mathcal{T}_{\mathcal{M}}(P_0)$ and some $h_0 \in L_2(P_0)$. The Riesz representation guarantees (1.1) holds provided the mapping of s to the derivative on the left-hand side is a bounded linear functional. The function h_0 is called a gradient and is not unique, in general. Equality holds for any $h' = h_0 + s^{\perp}$, where $s^{\perp}$ is orthogonal to all $s \in \mathcal{T}_{\mathcal{M}}(P_0)$ in $L_2(P_0)$ so that $\int s(z)s^{\perp}(z)dP_0(z) = 0$. The canonical gradient $\phi^*_{P_0}$ is defined as the unique gradient which belongs to $\mathcal{T}_{\mathcal{M}}(P_0)$.

We also call $\phi^*_{P_0}$ the efficient influence function for estimating $\Psi(P_0)$ in the model $\mathcal{M}$. This is justified as:

1. the function $\phi^*_{P_0}$ defines the score at $\epsilon = 0$ in a least favorable parametric submodel of $\mathcal{M}$ through P_0, hence defining the semiparametric efficiency bound (van der Vaart 1998, Lemma 25.19);

2. an estimator ψ_n is asymptotically efficient for ψ_0 at P_0 relative to $\mathcal{M}$ if and only if $\psi_n - \psi_0 = \frac{1}{n}\sum_{i=1}^n \phi^*_{P_0}(Z_i) + o_p(n^{-1/2})$ (van der Vaart 1998, Lemma 25.23).

In the remainder of this chapter, we discuss several common frameworks for constructing asymptotically linear estimators. A fundamental result of Klaassen (1987) tells us that, for a given gradient h_0 of Ψ at P_0 in $\mathcal{M}$, a consistent estimator of h_0 exists if and only if there exists an asymptotically linear estimator of ψ_0 with influence function h_0. Thus, we do not find it surprising that the frameworks we consider typically involve estimation of the gradient h_0 in the construction of an estimator ψ_n of ψ_0.

1.4 Estimating equations, one-step estimators, and TMLE

We make use of the following strategies for constructing estimators in later chapters.

1.4.1 Estimating equations-based M- and Z-estimators

An estimating function $\tilde{\phi}_{\psi,\eta}$ for $\psi_0 \in \mathbf{R}^d$ is a function $z \mapsto \tilde{\phi}_{\psi,\eta}(z) \in \mathbf{R}^d$ such that $\int \tilde{\phi}_{\psi,\eta_{P_0}}(z)dP_0(z) = 0$ has unique solution ψ_0, where $\eta_0 = \eta(P_0)$ for some nuisance parameter η that is typically unknown. A Z-estimator ψ_n of ψ_0 is defined as a solution in ψ to

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}_{\psi,\eta_n}(Z_i) = 0,$$

where η_n is an estimator of η_0. For arbitrary probability measure P and function $f \in L_2(P)$, we will write $Pf = \int f(z)dP(z)$. By $\mathbb{P}_n$ we denote the probability measure placing mass n^{-1} on each observation and mass zero otherwise, so that $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$. Following the notation of van der Vaart (1998) Theorem 5.31, we define $V_{\psi_0,\eta_{P_0}}$ to be the matrix of partial derivatives of $\psi \mapsto P_0\tilde{\phi}_{\psi_0,\eta_0}$ evaluated at (ψ_0, η_{P_0}). Under regularity conditions, it can be shown that

$$\psi_n - \psi_0 = -V^{-1}_{\psi_0,\eta_{P_0}}(\mathbb{P}_n\tilde{\phi}_{\psi_0,\eta_0} + P_0\tilde{\phi}_{\psi_0,\eta_n}) + o_p(n^{-1/2} + \|P_0\tilde{\phi}_{\psi_0,\eta_n}\|) \tag{1.2}$$

From this expansion, we find that ψ_n is an asymptotically linear estimator of ψ_0 with influence function $\tilde{\phi}_{\psi_0,\eta_0}$ provided the drift term $P_0\tilde{\phi}_{\psi_0,\eta_n}$ is $o_p(n^{-1/2})$.

1.4.2 One-step estimators

An alternate characterization of pathwise differentiability to that in (1.1) is described, for example, by Pfanzagl (1982). We say that $\Psi : \mathcal{M} \to \mathbf{K}$ is pathwise differentiable with strong gradient $h_{P'}$ if we have

$$\Psi(P') - \Psi(P_0) = -P_0 h_{P'} + R(P', P_0) \tag{1.3}$$

where $h_{P'} \in L_2(P')$ and $R(P_0, P') = o(\rho(P_0, P'))$ for a metric ρ defining distances between elements of $\mathcal{M}$. This formulation implies that the representation in (1.1) holds uniformly over all possible scores at $\epsilon = 0$ for submodels through P_0. We find (1.3) useful as the basis for proofs that an estimator ψ_n is asymptotically linear.

If P_n is an estimator of P_0, then we will call $\psi_n = \Psi(P_n)$ a plug-in or substitution estimator for $\psi_0 = \Psi(P_0)$. For pathwise differentiable Ψ with strong gradient ϕ_{P_0} relative to $\mathcal{M}$, we have the following expansion for the substitution estimator ψ_n of ψ_0:

$$\psi_n - \psi_0 = \underbrace{\mathbb{P}_n\phi_{P_0}}_{\text{linear}} - \underbrace{\mathbb{P}_n\phi_{P_n}}_{\text{bias}} + \underbrace{(\mathbb{P}_n - P_0)(\phi_{P_n} - \phi_{P_0})}_{\text{empirical process}} + \underbrace{R(P_n, P_0)}_{\text{remainder}}. \tag{1.4}$$

For ψ_n to be asymptotically linear at ψ_0 with influence function ϕ_{P_0}, we must verify that the sum of the bias, empirical process, and remainder terms in (1.4) is $o_p(n^{-1/2})$. Under regularity conditions, if P_n converges in an appropriate sense to P_0 at a fast enough rate, then it can be argued that the remainder term $R(P_n, P_0)$ is $o_p(n^{-1/2})$. For the empirical process term $(\mathbb{P}_n - P_0)(\phi_{P_n} - \phi_{P_0})$ to be $o_p(n^{-1/2})$ it suffices that ϕ_{P_n} belongs to a P_0-Donsker class of functions with probability tending to one as n increases and that $P_0(\phi_{P_n} - \phi_{P_0})^2 = o_p(1)$ (van der Vaart and Wellner 1996).

If the bias term $\mathbb{P}_n\phi_{P_n}$ is $o_p(n^{-1/2})$, we say that ψ_n has the small-bias property. However, $\mathbb{P}_n\phi_{P_n}$ is not typically $o_p(n^{-1/2})$. One approach to account for this is to define an estimator $\psi_n^1 = \psi_n + \mathbb{P}_n\phi_{P_n}$. Estimators constructed in this manner are called one-step or generalized Newton-Raphson estimators for ψ_0 and have been discussed, for example, by Pfanzagl (1982). Provided the remainder and empirical process terms are $o_p(n^{-1/2})$, ψ_n^1 is asymptotically linear at ψ_0 with influence function ϕ_{P_0}.

We note, however, that ψ_n^1 is not itself a substitution estimator. For fixed n, this means that ψ_n^1 may take values outside of the parameter space defined by the mapping Ψ. In Chapter 3, we will discuss an additional limitation of the one-step estimation framework in the context of robustifying asymptotic linearity to inconsistent estimation of nuisance parameters in certain missing data problems.

1.4.3 The TMLE framework

For a pathwise differentiable parameter $\Psi : \mathcal{M} \to \mathbf{K}$, the targeted maximum likelihood estimation and more general targeted minimum loss-based estimation (both referred to as

TMLE) frameworks proposed by van der Laan and Rubin (2006) result in efficient substitution estimators of ψ_0 that possess the small-bias property. The TMLE algorithm maps an initial estimator P_n of P_0 to a targeted estimator P_n^*. The resulting substitution estimator $\psi_n^* = \Psi(P_n^*)$ is also called the TMLE. By construction of P_n^*, the bias term $\mathbb{P}_n \phi_{P_n^*}^*$ arising in (1.4) is $o_p(n^{-1/2})$. Given appropriate regularity conditions hold, the remainder and empirical process term are both $o_p(n^{-1/2})$, as well. We then conclude that ψ_n is an asymptotically linear estimator of ψ_0 with influence function given by the efficient influence function $\phi_{P_0}^*$.

Suppose that p_n^1 is an initial density estimator for the density p_0 of P_0 with respect to common dominating measure μ. The TMLE algorithm for estimating ψ_0 is as follows:

1. Set $k = 1$.

2. Define a parametric submodel $p_{n,\epsilon}^k$ such that $p_{n,0}^k = p_n^k$ and $\frac{d}{d\epsilon} \log p_{n,\epsilon}^k \big|_{\epsilon=0} = \phi_{P_n^k}^*$.

3. Set $\epsilon_{n,k} = \arg\max_\epsilon \frac{1}{n} \sum_{i=1}^n \log p_{n,\epsilon}^k(Z_i)$.

4. Define $p_n^{k+1} = p_{n,\epsilon_{n,k}}^k$ and set $k = k + 1$.

5. If $\frac{1}{n} \sum_{i=1}^n \phi_{P_n^k}^*(Z_i) \approx 0$, set $p_n^* = p_n^k$ and $\psi_n^* = \Psi(P_n^*)$, where P_n^* is the probability measure corresponding to the density p_n^*.

6. Otherwise, continue to step 2 and repeat until convergence.

In most semiparametric models, estimating the entire density function is undesirable. This leads us to consider loss functions other than the negative log-likelihood in the TMLE procedure.

We suppose that $\Psi(P)$ depends on P only through a summary feature $P \mapsto \eta_P$ defined on the model $\mathcal{M}$. In this case, we write $\Psi(P) = \Psi(\eta_P)$ and call η_P a nuisance parameter for the estimation of $\Psi(P)$. This suggests construction of a substitution estimator for ψ_0 based on an initial nuisance estimator η_n, so that $\psi_n = \Psi(\eta_n)$. We suppose also that ϕ_P^* depends

on P_0 through η_P and some additional nuisance parameter π_P of $P \in \mathcal{M}$, so that we may write $\phi^*_{\eta,\pi}$ for given η and π.

In this case, for $\epsilon \in \mathbf{R}$ we define η_ϵ to be a fluctuation of η such that $\eta_0 = \eta$ and $z \mapsto \ell_{\eta,\pi}(z)$ a loss function such that

$$\arg\min_{\eta} P_0 \ell_{\eta,\pi_0} = \eta_0$$

If η_n^1 is an initial estimator for the nuisance parameter value η_0 and π_n is an estimator of π_0, then the TMLE algorithm for estimating ψ_0 is as follows:

1. Set $k = 1$.
2. Define a fluctuation $\eta^k_{n,\epsilon}$ such that $\eta^k_{n,0} = \eta^k_n$ and $\frac{d}{d\epsilon}\ell_{\eta^k_{n,\epsilon},\pi_n}\Big|_{\epsilon=0} = \phi^*_{\eta^k_n,\pi_n}$.
3. Set $\epsilon_{n,k} = \arg\min_\epsilon \frac{1}{n}\sum_{i=1}^n \ell_{\eta^k_{n,\epsilon},\pi_n}(Z_i)$.
4. Define $\eta_n^{k+1} = \eta^k_{n,\epsilon_{n,k}}$ and set $k = k + 1$.
5. If $\frac{1}{n}\sum_{i=1}^n \phi^*_{\eta^k_n,\pi_n}(Z_i) \approx 0$, set $\eta^*_n = \eta^k_n$ and $\psi^*_n = \Psi(\eta^*_n)$.
6. Otherwise, continue to step 2 and repeat until convergence.

Chapter 2

MODELS AS DELIBERATE APPROXIMATIONS

A version of this chapter is being published as Whitney et al. (in press). The concerns raised herein motivate the challenges addressed in later chapters.

In public health, statistical analyses are often based on parametric or semiparametric regression models. However, in many applications, assumptions such as linearity or proportional hazards are not expected to hold. This inherent model misspecification motivates us to consider the agnostic characteristics of popular model-based estimators. By agnostic characteristics, we mean properties of the estimator (e.g. consistency, sampling distribution) when the model possibly fails to hold. For maximum likelihood estimators (MLEs) in parametric models, the story is clear. The work of Huber (1967) shows that, under misspecification, MLEs converge in probability to a projection in terms of Kullback-Leibler divergence and that they are asymptotically linear under certain regularity conditions. As more complex statistical methods enter widespread use, it is important to understand the model-agnostic properties of these methods so that scientific conclusions may be drawn appropriately.

Recent work by Buja et al. (2019a) and Buja et al. (2019b) draws further attention to the problem of model misspecification in regression and to the study of its ramifications. In their work, the authors advocate for viewing model-based regression coefficients as non-parametric functionals of the data-generating mechanism. This viewpoint has the advantage of clarifying the definition of the estimand and formalizing how to perform model-robust inference based upon influence functions. While this previous work covered settings of parametric models in complete-data, the authors did not treat more general semi-parametric models. Here, we continue the conversation along these lines. We wish to highlight additional considerations that arise in the context of model misspecification in a broader range of scenarios.

The structure of the chapter is as follows. In Section 2.1, we argue that the model-robust interpretation of model-based estimands may not always be appealing, particularly when there is significant model misspecification or the sampling scheme includes some form of coarsening. In Section 2.2, we show that when the model fitting procedure involves data-adaptive estimation of nuisances, model-robust valid inference may be much more difficult to perform. We note in Section 2.3 that these difficulties can be preempted by defining deliberate projection parameters and using suitable non- or semi-parametric techniques for inference. Section 2.4 closes the chapter with a discussion of outstanding challenges in defining and estimating these (and other) parameters using non- or semi-parametric techniques.

2.1 Model-robust interpretation

Framing regression coefficients as indices for the 'projection' of the true regression function onto the specified model is intuitively appealing. In our experience, most practitioners are aware that this is implicitly what they are doing when fitting regression models. However, it must be stressed that not all projections are useful projections. Below, we highlight that model-based regression coefficients may have a poor interpretation when (a) the model used is overly parsimonious, or (b) when the data are subject to some form of coarsening.

2.1.1 Targeted versus indiscriminate parsimony

A key reason for the popularity of regression models is their ability to summarize parsimoniously key relationships. However, parsimony can have several impacts on the interpretation of regression coefficients. For example, it can mask effect modification — this occurs if the portion of the model pertaining to the exposure of interest is parsimonious. This may be desirable if the goal is to succinctly summarize population-averaged relationships. This targeted form of parsimony is what renders regression models attractive. However, parsimony could also result in poor confounding control — this occurs when the portion of the model that involves potential confounders is too inflexible to allow sufficient deconfounding. This is an example of indiscriminate parsimony, which is both unnecessary — it can often be mi-

tigated by the use of regression models with parsimonious exposure involvement but flexible confounding adjustment — and possibly harmful.

2.1.2 Example: ordinary least squares can result in indiscriminate parsimony

As an illustration, we expand upon a simple example stemming from the discussion found in Section 10 of Buja et al. (2019a). There, the authors note that when the underlying associations exhibit symmetry, there may be little to no linear trend. To be concrete, suppose that the data unit consists of the triple (W, X, Y), including a continuous outcome Y, exposure of interest X, and confounder W, generated from data-generating distribution P. Ordinary least-squares (OLS) regression may often be used in this context, with exposure and confounder both included as main terms, and reported upon with the appropriate caveat that the model coefficients represent indices of the least-squares projection. We show with a numerical example that the resulting estimand may not be particularly useful. Specifically, we consider P to be specified by $W \sim U(-2, 2)$,

$$X|W \sim \mathcal{N}(b_X W^4, 1) \text{ and } Y|(X, W) \sim \mathcal{N}(X + b_Y W^2, 1) \ .$$

Coefficients b_X and b_Y control the strength of the exposure-confounder and (nonlinear) outcome-confounder relationships, respectively. In this example, the deconfounded linear relationship between Y and X is unambiguous: the regression slope equals one. However, the OLS estimand has explicit form

$$\theta_{\text{OLS}}(b_X, b_Y) = 1 + \frac{1}{7}\left(\frac{7680 b_X b_Y}{225 + 4096 b_X^2}\right);$$

a range of numerical values are displayed in Figure 2.1 for various b_X and b_Y values. Depending on the strength of the underlying associations, the resulting estimand can be stronger or weaker, and of possibly the opposite sign as the true slope. This emphasizes that not all projections are useful — in fact, when the postulated regression model is strongly misspecified, there is a risk of inadequate deconfounding, and the regression functional may not be reflective of the underlying association of interest.

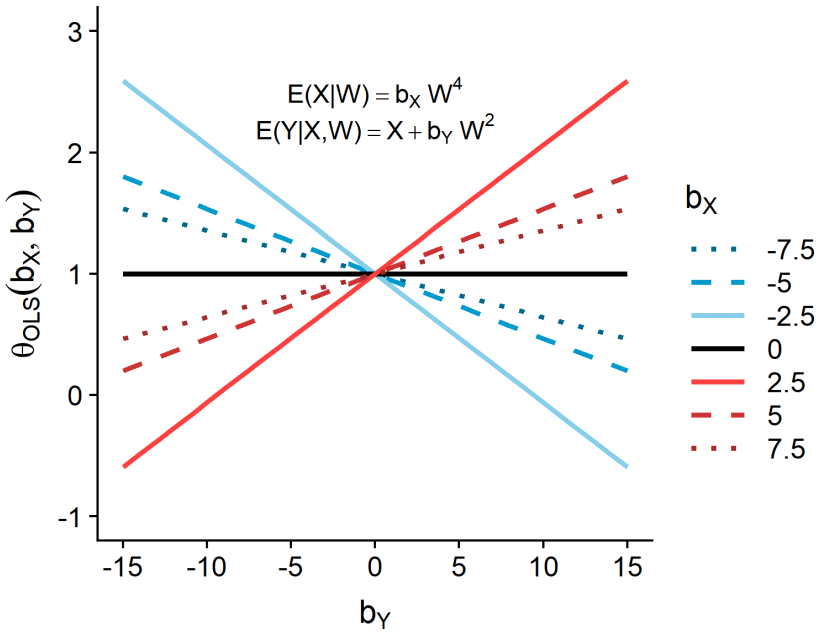

Figure 2.1: Exposure-specific OLS regression functional value when the true regression model follows a partially linear additive model. The line $\theta_{\text{OLS}} = 1$ (black) denotes the exposure-specific regression coefficient for the true data-generating mechanism. The other lines represent the value of the OLS regression functional for different strengths of the exposure-confounder relationship.

In the particular example considered, inclusion of polynomial confounder terms of sufficient degree in the linear model would have resolved the issue. However, this would likely not have been known a priori. The issue may have been discovered in a post-fit diagnostic analysis, but model revisions based on diagnostics are known to render calibrated inference difficult to perform. As an alternative, it would have been possible to consider a model with more flexible confounding control. In other words, a model without unnecessary (and possibly harmful) parsimony could have been used instead. For instance, as a flexible alternative to the linear model-based regression functionals, the partially linear additive model (PLAM) specifying that $E_P(Y \mid X = x, W = w) = \theta x + g(w)$ for some scalar θ and univariate

real-valued function g could be considered. The estimand would then be the index θ_* of the least-squares projection of the true regression function onto the PLAM.

This simple example underscores that more flexible semi-parametric models can lead to estimands with a more useful interpretation than provided by restrictive parametric models, without sacrificing parsimony relative to the association of interest. Nevertheless, this improved model-robust interpretation can come at a cost, as more involved procedures may be required to achieve valid inference. Indeed, performing calibrated model-robust inference requires additional considerations when data-adaptive techniques are used in the construction of the regression coefficient estimator. We discuss these challenges in Section 3.

2.1.3 The impact of coarsening in the data collection mechanism

In many applications, the observed data consist of a coarsening of the full data, for instance, due to missingness or censoring. Regression models are typically imposed on the full data distribution, since it is a feature of this distribution that is generally of scientific interest. Although in this context estimands resulting from misspecified models can still be interpreted as projections, the latter generally involve the coarsening mechanism.

As an illustration, it is instructive to consider the use of maximum likelihood (ML) with coarsened data. When the full data are available, the ML approach is known to yield consistent estimators of the index of the model element closest to the true data-generating distribution in a Kullback-Leibler sense. When instead the data are subject to coarsening, we must distinguish between the space of distributions for the observed versus full data. The ML approach in this case will identify the member of the model for the observed data (as induced by the model for the full data and the coarsening mechanism) closest to the true distribution of the observed data (as induced by the true distribution for the full data and the coarsening mechanism). As such, in these settings, model-based estimators often correspond to regression functionals that depend not only on the full data distribution but also the coarsening mechanism.

This phenomenon generalizes the notion of miss/well-specification introduced by the aut-

hors to include the distribution of coarsening variables in addition to that of regressors. However, the coarsening mechanism is usually a study-specific nuisance rather than an inherent feature of the population of interest. As such, dependence of the regression functional on the coarsening mechanism is particularly troublesome. Indeed, two investigators studying the same population and fitting the same regression models may be estimating very different quantities simply because of differences in the coarsening affecting their study samples.

2.1.4 Example: relationship between the proportional hazards regression estimator and conditional censoring distribution

As a concrete illustration of this phenomenon, it is informative to consider the case of proportional hazards (PH) regression under right-censoring. In the simplest of scenarios, where the full data consist of observations on the time-to-event variable T and a single binary covariate X, the PH model stipulates that the conditional hazard function h_x of the distribution of T given $X = x$ satisfies

$$h_x(t) = h_0(t) \exp(\theta x) \text{ for all } t > 0 \ ,$$

where h_0 is an unspecified baseline hazard and θ is the scalar regression coefficient of interest. Instead of complete observations, it is common to observe possibly right-censored event times. When the censoring variable is conditionally independent of T given X, and the PH model indeed holds, the maximizer of the partial likelihood is known to be a consistent estimator of the true regression coefficient.

When instead the PH model is misspecified, one may hope that the resulting regression functional perhaps represents an average of the time-varying hazard ratio (on a logarithmic scale). It has been shown that this is indeed approximately true, though the limit in probability θ_* of the maximum partial likelihood estimator (MPLE) depends not only on the conditional time-to-event and marginal covariate distributions but also on the conditional censoring distribution (Struthers and Kalbfleisch 1986) in a complicated manner. The fact that the censoring distribution defines the estimand is particularly alarming. In commenting

on this finding, O'Quigley (2008) states that the partial likelihood-based regression functional is not itself particularly useful nor interpretable — we agree with this viewpoint.

To emphasize this point numerically, we may consider the hazard functions $h_1(t) := \alpha t^{\alpha-1}$ for arbitrary $\alpha \geq 1$ and $h_0(t) := 1$. In such case, the PH model holds if and only if $\alpha = 1$. The further α is from this value, the more time-varying the hazard ratio $h_1(t)/h_0(t)$ becomes, thereby increasingly violating the PH model assumption. In Figure 2.2, we display the value θ_* of the partial likelihood regression functional as a function of α for various censoring distributions. For simplicity, we have considered exponential distributions for the conditional censoring distribution, with exposure-specific rate parameters $\gamma_0, \gamma_1 \in \{0.2, 1.1, 2.0\}$. As is readily apparent, the dependence of θ_* on the censoring distribution increases with α.

Noting this dependence, under differing independence assumptions, Xu and O'Quigley (2000), Schemper et al. (2009) and Hattori and Henmi (2012) have studied weighted partial likelihood-based estimators whose corresponding estimands do not depend on the censoring distribution. Nevertheless, their estimands represent interpretable weighted averages of log-hazard ratios only in an *approximate* sense. In Section 4, we propose a novel estimator *exactly* targeting a weighted average log-hazard ratio.

2.2 Valid inference in the presence of irregular nuisances

In Buja et al. (2019b), the authors define the regression functional broadly as the solution of (a set of) population-level model-derived estimating equation(s), possibly arising from the minimization of a risk function. For the examples explicitly considered, the estimating function is entirely parametric, being indexed by a vector including the parameter of interest and possibly nuisance parameters. In such cases, under regularity conditions, the resulting estimator can be shown to be asymptotically linear using a standard Taylor expansion; thus, asymptotic normality at the parametric rate holds, even when the model is misspecified. In contrast, in the context of certain semi-parametric models, some of the indexing nuisances may be infinite-dimensional and irregular, in the sense that they are not estimable at the parametric rate without strong (e.g., parametric) assumptions. The asymptotic linearity of

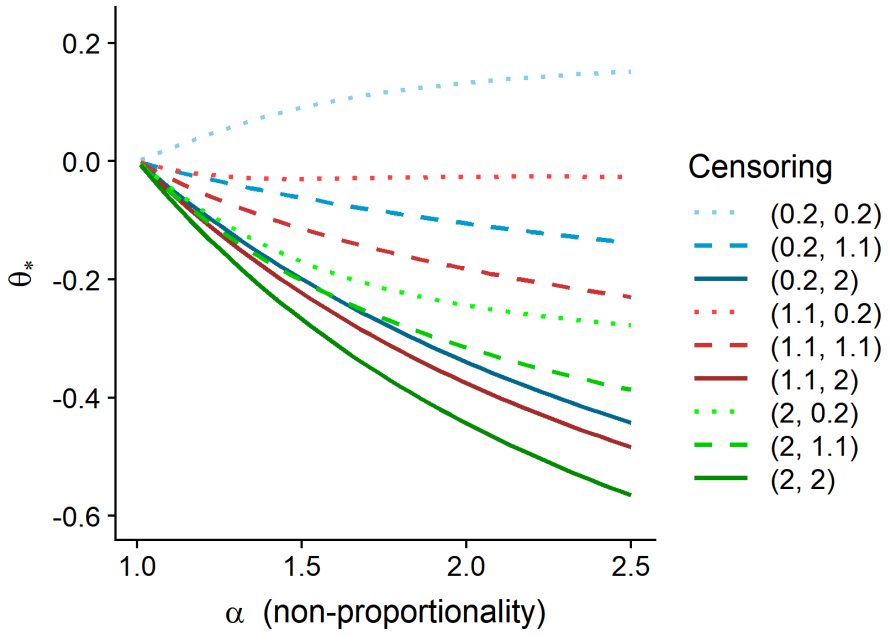

Figure 2.2: The model-robust interpretation of the partial likelihood regression functional depends on the censoring distribution. For a binary covariate, the population hazard ratio at time t is assumed to be $\alpha t^{\alpha-1}$, with larger α values corresponding to greater departures from the PH model. The exposure group-specific censoring distributions are exponential with rates γ_0 and γ_1.

model-based estimators may in such cases rely on correct specification of the model. This happens because the nuisance estimator may not contribute in first order to the behavior of the regression functional estimator *when the model is correctly specified*, but indeed does so *when the model is misspecified*. In the latter case, the regression functional estimator may inherit the slow convergence rate of the nuisance estimator, and fail to be asymptotically normal at the parametric rate, thereby rendering inference difficult.

2.2.1 Large sample behavior in the estimating equations framework

For concreteness, suppose that an estimating function ψ is available for the parametric index θ of a semi-parametric regression model. Suppose further that this estimating function is indexed by an infinite-dimensional nuisance η. To simplify notation, we consider θ to be scalar. Suppose that η_n is a consistent estimator of the true nuisance value η_0, defined unambiguously when the semi-parametric model holds, and that η_n tends to some $\eta_*(P)$ in general, where P denotes the data-generating distribution. If P is in the model, then $\eta_*(P) = \eta_0$. In this case, the model-robust regression functional is the solution $\theta_*(P)$ of the population equation $E_P\{\psi(\theta, \eta_*(P); Z)\} = 0$. In practice, any solution θ_n of the empirical equation

$$\frac{1}{n}\sum_{i=1}^{n}\psi(\theta, \eta_n; Z_i) = 0$$

may be taken as estimator of $\theta_*(P)$. In what follows, we will simply write θ_* and η_*, dropping the explicit dependence on P for convenience. As before, under regularity conditions, a Taylor expansion results in the first-order approximation

$$\theta_n - \theta_* \approx -\left[\frac{\partial}{\partial\theta}E_P\left\{\psi(\theta, \eta_*; Z)\right\}\Big|_{\theta=\theta_*}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\psi(\theta_*, \eta_*; Z_i) + \Phi(\eta_n)\right],$$

where Φ is the functional $\eta \mapsto \int \psi(\theta_*, \eta; z)dP(z)$. If $\Phi(\eta_n)$ is asymptotically negligible in the sense that $\Phi(\eta_n) = o_P(n^{-1/2})$, then θ_n is an asymptotically linear estimator with influence function proportional to the estimating function. Otherwise, $n^{1/2}(\theta_n - \theta_*)$ may fail to converge in law to a non-degenerate limit.

If Φ is sufficiently smooth, we may use the first-order approximation

$$\Phi(\eta_n) = \Phi(\eta_n) - \Phi(\eta_*) \approx \dot{\Phi}(\eta_*; \eta_n - \eta_*)\ ,$$

where the expression on the right-hand side denotes the Gâteaux derivative of Φ at η_* in the direction of $\eta_n - \eta_*$. If $h \mapsto \dot{\Phi}(\eta_*; h)$ is identically zero for h ranging in a set in which $\eta_n - \eta_*$ concentrates, then $\Phi(\eta_n)$ can be expected to be $o_P(n^{-1/2})$ provided $\eta_n - \eta_*$ vanishes quickly enough, as may be needed to guarantee the asymptotic linearity of θ_n. In

such case, the estimating function is said to be orthogonalized with respect to the nuisance η. In contrast, if the estimating function is not orthogonalized, then $\dot{\Phi}(\eta_*; \eta_n - \eta_*)$ will generally contribute in first-order to the behavior of $\theta_n - \theta_*$. Since nonparametric estimators of nuisance parameters often have a rate of convergence that is slower than the parametric rate, convergence of $n^{1/2}(\theta_n - \theta_*)$ to a nondegenerate limit distribution cannot be expected, at least with standard tuning of the involved data-adaptive nuisance estimators.

The above discussion relates to key ideas in efficiency theory. In that literature, influence functions often serve as estimating functions, in part because they are typically pre-orthogonalized relative to the nuisances involved — see, e.g., Lemma 1.3 of van der Laan and Robins (2003) for results in the finite-dimensional case. However, this orthogonalization generally only holds in the model under which the influence function is derived. As such, it may well be that $\Phi(\eta_n)$ is a higher-order term when the semi-parametric regression model holds but is a first-order term otherwise.

2.2.2 Example: agnostic and model-specific behavior of estimators based on the PLAM

We highlight the above phenomenon in the context of an example. Suppose that we wish to evaluate the association between outcome Y and binary exposure X adjusting for W, and we denote the data unit by $Z := (W, X, Y)$. We focus on the coefficient θ_0 in the PLAM $E_P(Y \mid X = x, W = w) = \theta_0 x + g_0(w)$ mentioned in Section 2.1.1. Defining $\pi_0(w) := E_P(X \mid W = w)$, we consider model-based estimation approaches built upon two candidate estimating functions

$$\psi_0(\theta, \pi; z) := \{x - \pi(w)\}(y - \theta x) \text{ and } \psi(\theta, g, \pi; z) := \{x - \pi(w)\}\{y - \theta x - g(w)\}.$$

It can be shown that ψ is proportional to the efficient influence function for θ_0 in the PLAM under homoscedasticity, but that ψ_0 is not even an influence function (Yu and van der Laan 2003). These estimating functions require estimation of π_0 and g_0. The nuisance π_0 is defined irrespective of whether the PLAM holds and can be estimated using a non-parametric estimator π_n. This allows us to define the model-based estimator $\theta_{0,n}$ based on ψ_0 and π_n,

namely

$$\theta_{0,n} := \frac{\sum_{i=1}^{n} Y_i\{X_i - \pi_n(W_i)\}}{\sum_{i=1}^{n} X_i\{X_i - \pi_n(W_i)\}} ,$$

with corresponding model-robust estimand $\theta_* := E_P[Y\{X - \pi_0(W)\}]/E_P[X\{X - \pi_0(W)\}]$. The nuisance g_0 is only well-defined under the PLAM. Two different estimators of g_0 consistent under the PLAM, say $g_{1,n}$ and $g_{2,n}$, may each converge to distinct limits $g_{1,*}$ and $g_{2,*}$ outside the PLAM. Noting that $g_{1,*}(w) := E_P(Y \mid X = 0, W = w) = g_0(w)$ under the PLAM, any non-parametric estimator $g_{1,n}(w)$ of $g_{1,*}(w)$ could be used as estimator of $g_0(w)$. Alternatively, we may consider the more elaborate back-fitting approach of Buja et al. (1989), setting $g_{2,n}(w)$ to be a non-parametric estimator of the regression of $Y - \theta_{0,n}X$ onto W evaluated at w. We note that $g_{2,n}(w)$ is then a consistent estimator of $g_{2,*}(w) := E_P(Y \mid W = w) - \theta_*\pi_0(w)$, which also coincides with $g_0(w)$ under the PLAM. The resulting model-based estimators of θ_0 based on ψ and $(\pi_n, g_{j,n})$ are then

$$\theta_{j,n} := \frac{\sum_{i=1}^{n}\{Y_i - g_{j,n}(W_i)\}\{X_i - \pi_n(W_i)\}}{\sum_{i=1}^{n} X_i\{X_i - \pi_n(W_i)\}}$$

for $j = 1, 2$. Both $\theta_{1,n}$ and $\theta_{2,n}$ also tend to θ_* outside the PLAM.

The respective (first-order) nuisance contributions of $(\pi_n, g_{j,n})$ when using ψ_0 and ψ are

$$\Phi_0(\pi_n) \approx -\int \{\pi_n(w) - \pi_0(w)\}E_P(Y - \theta_* X \mid W = w)dP(w) \quad \text{and}$$
$$\Phi(\pi_n, g_{j,n}) \approx \int \{\pi_n(w) - \pi_0(w)\}\{g_{j,*}(w) - E_P(Y - \theta_* X \mid W = w)\}dP(w) .$$

We provide the details of these approximation in Appendix 2.C. It is clear that, irrespective of whether the PLAM holds, $\Phi_0(\pi_n)$ makes a first-order contribution to the behavior of $\theta_{0,n} - \theta_*$. This fact is not surprising since ψ_0 is not orthogonalized relative to π. If the PLAM holds, then $E_P(Y - \theta_* X \mid W = w) = E_P(Y - \theta_0 X \mid W = w) = g_0(w)$ and the first-order approximation of $\Phi(\pi_n, g_n)$ is zero with g_n taken to be either $g_{1,n}$ or $g_{2,n}$. If instead the PLAM does not hold, the situation is more complex. In general, $g_{1,*}(w) - E_P(Y - \theta_* X \mid W = w) \neq 0$, whereas $g_{2,*}(w) - E_P(Y - \theta_* X \mid W = w) = 0$ for each w. Thus, when the PLAM does not hold, the nuisance estimator will make a first-order contribution to the behavior of $\theta_{1,n} - \theta_*$

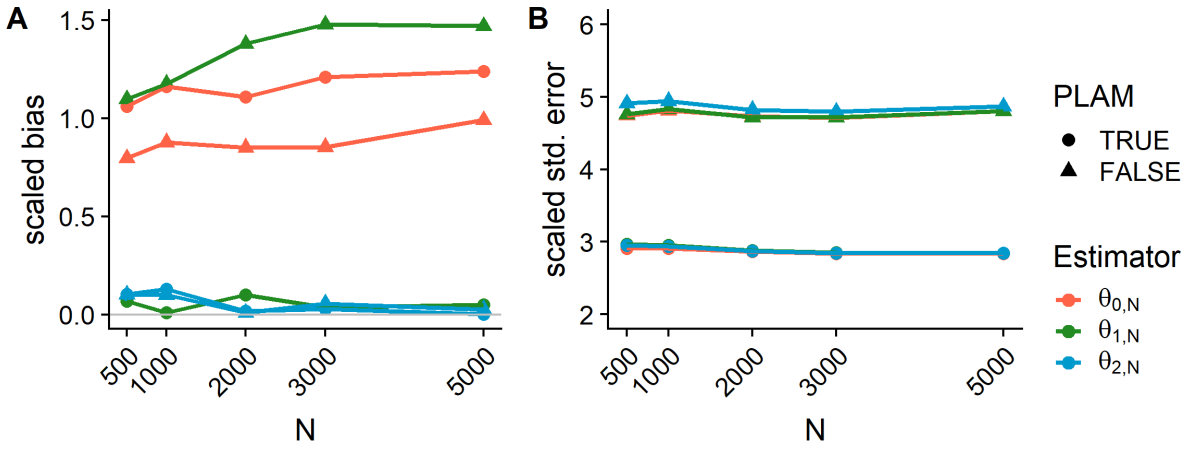

Figure 2.3: Empirical bias and standard error of model-based estimators $\theta_{0,n}$, $\theta_{1,n}$ and $\theta_{2,n}$ scaled by $n^{1/2}$ for sample sizes $n \in \{500, 1000, 2000, 3000, 5000\}$ computed using 5000 simulated datasets for each sample size under correct and incorrect model specifications.

but not to that of $\theta_{2,n} - \theta_*$. In other words, valid model-robust inference can be easily carried out using $\theta_{2,n}$ but not $\theta_{1,n}$.

2.2.3 Simulation study

We illustrate this phenomenon in a simulation study. We set $W \sim U(-2, 2)$, $X|W \sim$ Bernoulli$(\pi_0(W))$ with $\pi_0(w) = \text{expit}\,(0.5 + 2w - w^2)$, and $Y|(W, X) \sim \mathcal{N}(\mu(X, W), 1)$, where $\mu(x, w) = 0.2x + w^2$ or $\mu(x, w) = (2x - 1)w$ to simulate a valid versus misspecified PLAM, respectively. We generated 5000 datasets for each sample size $n \in \{500, 1000, 2000, 3000, 5000\}$ and computed estimators $\theta_{0,n}$, $\theta_{1,n}$ and $\theta_{2,n}$ with Nadaraya-Watson kernel estimator (with cross-validated bandwidth selection) used for nuisance estimation whenever non-parametric regression was required. Results are depicted in Figure 2.3.

This simulation study confirms the expected behavior of the estimators considered. The bias of $\theta_{0,n}$ does not tend to zero sufficiently fast to allow convergence of $n^{1/2}(\theta_{0,n} - \theta_*)$ to a nondegenerate distribution regardless of whether the PLAM holds — this occurs because ψ_0 is

not orthogonalized, and so, the behavior of the kernel regression estimator dominates. When the PLAM holds, the bias of both $\theta_{1,n}$ and $\theta_{2,n}$ tends to zero faster than $n^{-1/2}$. However, only the bias of $\theta_{2,n}$ remains small when the PLAM is misspecified. These results demonstrate the importance of estimating function orthogonality when evaluating the model-agnostic sampling behavior of regression model-based estimators. Additionally, they highlight that if a model-based procedure is not suitably orthogonalized, it does not suffice to devise an improved variance estimator — bias that does not vanish quickly enough results in the lack of a nondegenerate distribution at the parametric rate.

2.3 Deliberate model-agnostic parameter extensions

In Buja et al. (2019b), the authors focus on regression functionals that arise as the limit of model-based procedures, and propose strategies for model-robust inference. As an alternative, it may be fruitful to first define a model-agnostic parameter extension (or projection) based on the considered regression model, and then develop robust inferential procedures for this estimand. By deliberately defining the projection of interest rather than letting it be dictated by some model-based estimator, issues pertaining to parameter interpretation, as highlighted in Section 2.1, can largely be circumvented. Furthermore, by using non- or semi-parametric tools, valid inference can be performed for this projection parameter while avoiding the potentially poor behavior of model-based procedures in the presence of irregular nuisances and model misspecification, as discussed in Section 2.2.

To illustrate what we mean by model-agnostic parameter extension, we return to the proportional hazards model — we refer interested readers to Chambaz et al. (2012), Graham and Pinto (2018) and references therein for a treatment of projections onto the PLAM.

2.3.1 Example: an agnostic parameter extension for PH regression

Under proportional hazards, the regression coefficient θ_0 is the (constant) hazard ratio value. A natural summary of a time-varying hazard ratio is

$$\theta_{**} := \int \log \left\{ \frac{h_1(t)}{h_0(t)} \right\} \nu(dt) \ ,$$

where h_x is the true hazard function corresponding to $X = x$, and ν represents a weight function, possibly dependent on components of the data-generating distribution. If the PH model holds, θ_{**} coincides with the usual PH regression coefficient θ_0. If the PH model does not hold, θ_{**} remains a transparent and interpretable estimand. While the usual Cox estimand is often claimed to represent a quantity such as θ_{**} when the PH model fails to hold, we see from Figure 2.2 that the weight function depends to a large extent on the interplay between the censoring distribution and degree of model misspecification.

Consider n independent triples $Z_i := (Y_i, \Delta_i, X_i)$, where Y_i is the follow-up time, Δ_i the event indicator, and X_i the exposure group indicator for study participant i. Suppose that the right-censoring mechanism is uninformative within exposure groups. In that case, under identification conditions, θ_{**} represents a pathwise differentiable parameter of the data-generating distribution. A regular and asymptotically linear estimator of θ_{**} can thus be constructed. For example, for the important case in which ν is the marginal time-to-event distribution, we may consider the one-step bias-corrected estimator

$$\theta_{\mathrm{OS},n} := \int \theta_n(t) F_n(dt) + \frac{1}{n} \sum_{i=1}^{n} \phi_n(Z_i) \ ,$$

where $\theta_n(t) := \log h_{1,n}(t) - \log h_{0,n}(t)$ with $h_{x,n}$ a non-parametric estimator of h_x, $F_n := (1-\pi_n)F_{0,n} + \pi_n F_{1,n}$ is a non-parametric estimator of the marginal time-to-event distribution function, $F_{x,n}$ is the Kaplan-Meier estimator of the distribution function corresponding to $X = x$, and π_n is the proportion of study participants with $X = 1$. Here, ϕ_n is a plug-in estimator of the (non-parametric) efficient influence function of θ_{**} when the weight function

is considered fixed:

$$\begin{aligned}
\phi_n(z) &:= \left(\frac{x}{\pi_n}\right)\gamma_{1,n}(y,\delta) - \left(\frac{1-x}{1-\pi_n}\right)\gamma_{0,n}(y,\delta) \\
\gamma_{x,n}(y,\delta) &:= \frac{\delta \exp\{-x\theta_n(y)\}Q_n(y)}{R_{x,n}(y)} - \int_{u\leq y} \frac{1}{R_{x,n}(u)} F_n(du) \\
Q_n(y) &:= (1-\pi_n)\{1-F_{0,n}(y)\} + \pi_n\{1-F_{1,n}(y)\}\exp\{\theta_n(y)\} \\
R_{x,n}(y) &:= \frac{1}{n}\sum_{i=1}^{n} I(Y_i \geq y)
\end{aligned}$$

It can be shown (see Appendix 2.D) that $n^{1/2}(\theta_{\mathrm{OS},n} - \theta_{**})$ tends to a mean-zero Gaussian variable under regularity conditions and rate conditions on $h_{x,n}$.

2.3.2 Simulation study

A simulation study was conducted to evaluate the finite-sample behavior of this estimator. We generated exposure $X \sim \text{Bernoulli}(2/3)$ and time-to-event $T|X \sim \text{Weibull}(1, 1+\alpha X/2)$, with $\alpha \in \{0,1\}$ yielding correct and incorrect PH specifications. Censoring time $C \sim$ exponential(0.2) was generated independently of (T,X). For each scenario, we generated 5000 datasets of size $n \in \{500, 1000, 2000, 3000, 5000\}$, and evaluated the empirical bias (relative to the projection estimand) and standard error of $\theta_{\mathrm{OS},n}$ and of the maximum partial likelihood estimator $\theta_{\mathrm{MPLE},n}$. Hazard functions were estimated with kernel regression. Results are displayed in Figure 2.4. Under correct PH specification, both estimators have negligible bias; interestingly, they also have similar standard errors. When the PH assumption fails, only $\theta_{\mathrm{OS},n}$ tends to the projection parameter. It also has bias tending to zero faster than $n^{-1/2}$ and variance stabilizing at rate n^{-1}. In contrast to the Cox estimand, the projection parameter is an interpretable summary of the hazard ratio invariant to the censoring distribution.

2.4 Discussion

The simple example above highlights that it is possible to define deliberate model-agnostic extensions of regression coefficients, and to construct (non-parametric efficient) estimators

that minimize the need for unrealistic assumptions about the data-generating mechanism. We emphasize that if knowledge on the data-generating distribution is available (e.g., known moment conditions or conditional independences), it should be incorporated into the inferential process. In such case, the efficient influence function used to construct the estimator of the regression functional should be relative to this greater state of knowledge. While we used the simple one-step construction in our illustration, more recent strategies with possibly improved performance also exist — see, e.g., van der Laan and Rose (2011). These strategies naturally allow the use of flexible, data-adaptive tools (e.g., machine learning) for nuisance estimation yet allow valid inference for the deliberate target of scientific interest. A potential challenge is the reliance of these strategies on analytic objects whose derivation requires specialized skills, though there have been recent efforts to overcome this difficulty using computational tools (Carone et al. in press). In addition to better understanding the ramifications of regression model misspecification, and devising model-robust inferential procedures for available estimators, as Buja and co-authors have done, we hope to see further efforts to develop and vet estimators of natural model-agnostic parameter extensions based upon common regression models.

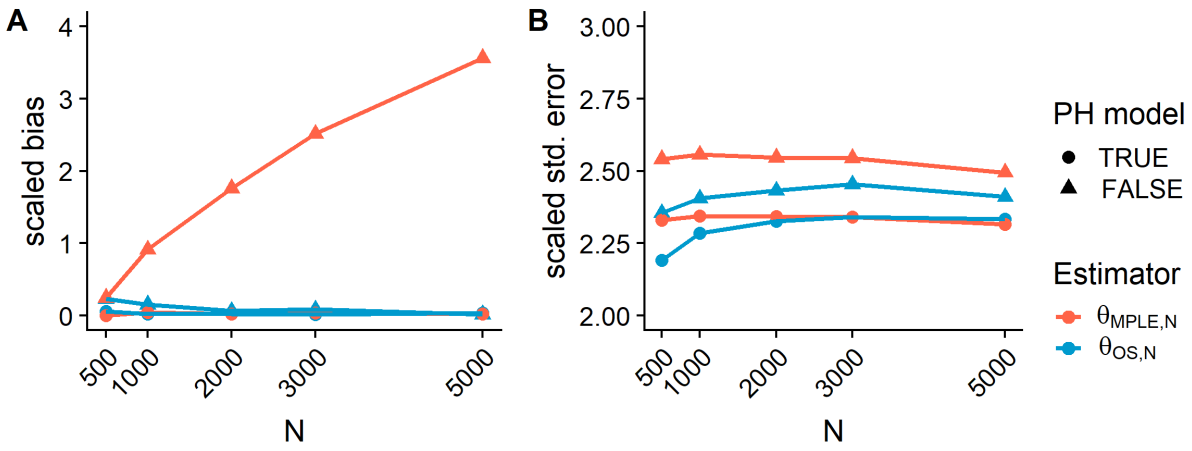

Figure 2.4: Empirical bias and standard error of estimators $\theta_{\text{MPLE},n}$ and $\theta_{\text{OS},n}$ scaled by $n^{1/2}$ for sample sizes $n \in \{500, 1000, 2000, 3000, 5000\}$ computed using 5000 simulated datasets for each sample size under correct and incorrect model specifications.

APPENDIX

2.A Computing the OLS slopes for Figure 2.1

In Figure 2.1, the data unit is (W, X, Y), with data-generating distribution P. Specifically, we consider P to be specified by $W \sim U(-2, 2)$,

$$X|W \sim \mathcal{N}(b_X W^4, 1) \text{ and } Y|(X, W) \sim \mathcal{N}(X + b_Y W^2, 1) \ .$$

Coefficients b_X and b_Y control the strength of the exposure-confounder and (nonlinear) outcome-confounder relationships, respectively. Let $(W_1, X_1, Y_1), \ldots, (W_n, X_n, Y_n)$ denote an iid sample of size n from P. We define the design matrix

$$Z = \begin{pmatrix} 1 & X_1 & W_1 \\ \vdots & \vdots & \vdots \\ 1 & X_n & W_n \end{pmatrix} .$$

We assume that $Z^\top Z$ is non-singular, so that the OLS estimator for the regression of Y on X and W is $(Z^\top Z)^{-1} Z^\top Y$. We then define $\theta_{\text{OLS},n}$ to be the coefficient for X. We are interested in the value of the limit (in probability) of $\theta_{\text{OLS},n}$, corresponding to the OLS regression functional. The following theorem establishes the OLS regression functional as a function of b_X and b_Y, as displayed in the figure.

Theorem 2.1. *The OLS estimator $\theta_{OLS,n}$ converges in probability to*

$$\theta_{OLS}(b_X, b_Y) = 1 + \frac{1}{7}\left(\frac{7680 b_X b_Y}{225 + 4096 b_X^2}\right) .$$

Proof. Without loss of generality, assume that the data have been centered. We have

$$(Z^\top Z)^{-1} Z^\top Y = \begin{pmatrix} 0 \\ \frac{\text{cov}_n(X,Y)\text{cov}_n(W,W) - \text{cov}_n(W,X)\text{cov}_n(W,Y)}{\text{cov}_n(X,X)\text{cov}_n(W,W) - \text{cov}_n(X,W)^2} \\ \frac{\text{cov}_n(W,Y)\text{cov}_n(X,X) - \text{cov}_n(W,X)\text{cov}_n(X,Y)}{\text{cov}_n(X,X)\text{cov}_n(W,W) - \text{cov}_n(X,W)^2} \end{pmatrix} ,$$

where $\text{cov}_n(a,b) := \frac{1}{n}\sum_{i=1}^n a_i b_i$ is the sample covariance for (centered) vectors $a, b \in \mathbf{R}^n$. From this, we see that the estimated coefficient for X is

$$\theta_{\text{OLS},n} = \frac{\text{cov}_n(X,Y)\text{cov}_n(W,W) - \text{cov}_n(W,X)\text{cov}_n(W,Y)}{\text{cov}_n(X,X)\text{cov}_n(W,W) - \text{cov}_n(X,W)^2}.$$

By the law of large numbers and the continuous mapping theorem, $\theta_{\text{OLS},n} \to_p \theta_{\text{OLS},*}$ with

$$\theta_{\text{OLS},*} := \frac{\text{cov}_P(X,Y)\text{cov}_P(W,W) - \text{cov}_P(W,X)\text{cov}_P(W,Y)}{\text{cov}_P(X,X)\text{cov}_P(W,W) - \text{cov}_P(X,W)^2}.$$

Since the density of W is symmetric, we have $\text{cov}_P(W,X) = 0$ for all b_X and b_Y. This yields the simplification $\theta_{\text{OLS},*} = \text{cov}_P(X,Y)/\text{cov}_P(X,X)$. Standard moment calculations lead to

$$\begin{aligned}
\text{cov}_P(X,Y) &= 1 + b_X^2\{E_P(W^8) - E_P(W^4)^2\} + b_X b_Y\{E_P(W^6) - E_P(W^2)E_P(W^4)\}; \\
\text{cov}_P(X,X) &= 1 + b_X^2\{E_P(W^8) - E_P(W^4)^2\}.
\end{aligned}$$

The even moments of W are given by $E_P(W^{2k}) = \int_{-2}^2 (w^{2k}/4)dw = 4^k/(2k+1)$ for each positive integer k. Setting $k \in \{1,2,3,4\}$, we compute the required moments

$$E_P(W^2) = \frac{4}{3}; \quad E_P(W^4) = \frac{16}{5}; \quad E_P(W^6) = \frac{64}{7}; \quad E_P(W^8) = \frac{256}{9}.$$

Combining the above calculations with our expression for $\theta_{\text{OLS},*}$, we find that

$$\begin{aligned}
\theta_{\text{OLS},*} &= \frac{\text{cov}_P(X,Y)}{\text{cov}_P(X,X)} \\
&= 1 + \frac{b_X b_Y\{E_P(W^6) - E_P(W^2)E_P(W^4)\}}{1 + b_X^2\{E_P(W^8) - E_P(W^4)^2\}} \\
&= 1 + \frac{(512/105)b_X b_Y}{1 + (4096/225)b_X^2} \\
&= 1 + \frac{1}{7}\left(\frac{7680 b_X b_Y}{225 + 4096 b_X^2}\right).
\end{aligned}$$

Setting $\theta_{\text{OLS}}(b_X, b_Y) := \theta_{\text{OLS},*}$ completes the proof.

□

2.B Solving the partial likelihood equations for Figure 2.2

In this setting, Struthers and Kalbfleisch (1986) showed that θ_* is the solution to the population estimating equation

$$\int_0^\infty \left\{ \frac{h_1(t)}{h_0(t)} - e^\theta \right\} \left\{ \frac{h_0(t) R_0(t) R_1(t)}{(1-\pi) R_0(t) + \pi R_1(t) e^\theta} \right\} dt = 0.$$

For the example in Figure 2.2, we take $h_0(t) := 1$ and $h_1(t) := \alpha t^{\alpha-1}$. This corresponds to $T|X = x \sim \text{Weibull}(1, \alpha^x)$. The censoring distributions were taken to be $C|X = x \sim$ exponential(γ_x), where $\gamma_x > 0$ is the hazard parameter. Under this specification, $R_x(t) = \exp\{-(\gamma_x t + t^{1+\alpha x})\}$. Substituting these quantities and solving the population estimating equation numerically yields the values of θ_* displayed in the figure.

2.C Details of the PLAM nuisance contributions

We define $\Phi_0 : \pi \mapsto \int \psi_0(\theta_*, \pi; z) dP(z)$ and $\Phi : (g, \pi) \mapsto \int \psi(\theta_*, g, \pi; z) dP(z)$.

Theorem 2.2. *The following approximations hold up to second-order remainder terms:*

1. $\Phi_0(\pi_n) \approx -\int \{\pi_n(w) - \pi_0(w)\} E_P(Y - \theta_* X \mid W = w) dP(z)$;

2. $\Phi(\pi_n, g_{j,n}) \approx \int \{\pi_n(w) - \pi_0(w)\} \{g_{j,*}(w) - E_P(Y - \theta_* X \mid W = w)\} dP(z)$.

Proof. 1. Set $\pi_{\epsilon,n} := \pi_0 + \epsilon(\pi_n - \pi_0)$. A first-order Taylor expansion yields

$$\Phi_0(\pi_{\epsilon,n})|_{\epsilon=1} = \Phi_0(\pi_{\epsilon,n})|_{\epsilon=0} + \left. \frac{d}{d\epsilon} \Phi_0(\pi_{\epsilon,n}) \right|_{\epsilon=0} + r_n$$

for a second-order remainder term r_n. The Gâteaux derivative is

$$\begin{aligned} \dot{\Phi}_0(\pi_0; \pi_n - \pi_0) &= \left. \frac{d}{d\epsilon} \Phi_0(\pi_{\epsilon,n}) \right|_{\epsilon=0} \\ &= \left. \frac{d}{d\epsilon} \int \{x - \pi_{\epsilon,n}(w)\}(y - \theta_* x) dP(z) \right|_{\epsilon=0} \\ &= -\int \{\pi_n(w) - \pi_0(w)\} E_P(Y - \theta_* X \mid W = w) dP(z), \end{aligned}$$

where the last equality follows by taking conditional expectations. Noting that $\Phi_0(\pi_n) \equiv \Phi_0(\pi_\epsilon)|_{\epsilon=1}$ and $\Phi_0(\pi_0) \equiv \Phi_0(\pi_\epsilon)|_{\epsilon=0} \equiv 0$, we obtain the desired approximation.

2. Define the vector of nuisance parameters $\eta = (g, \pi)$. Set $g_{\epsilon,n} := g_{j,*} + \epsilon(g_{j,n} - g_{j,*})$ so that we may take $\eta_{\epsilon,n} = (g_{\epsilon,n}, \pi_{\epsilon,n})$ to be a single (partitioned) nuisance parameter. The Gâteaux derivative is

$$\begin{aligned}
\dot{\Phi}(\eta_*; \eta_n - \eta_*) &= \frac{d}{d\epsilon}\Phi(\pi_{\epsilon,n}, g_{\epsilon,n})\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\int \{x - \pi_{\epsilon,n}(w)\}(y - \theta_* x - g_{\epsilon,n}(w))dP(z)\Big|_{\epsilon=0} \\
&= -\int \{x - \pi_0(w)\}\{g_{j,n}(w) - g_{j,*}(w)\}dP(z) \\
&\quad - \int \{\pi_n(w) - \pi_0(w)\}\{y - \theta_* x - g_{j,*}(w)\}dP(z) \\
&= 0 - \int \{\pi_n(w) - \pi_0(w)\}\{y - \theta_* x - g_{j,*}(w)\}dP(z) \\
&= \int \{\pi_n(w) - \pi_0(w)\}\{g_{j,*}(w) - E_P(Y - \theta_* X \mid W = w)\}dP(z).
\end{aligned}$$

The last two equalities follow by taking conditional expectations inside the respective integrals. As above, substituting the relevant quantities into a first-order Taylor expansion of $\Phi(\eta_\epsilon)$ at $\epsilon = 0$ yields the desired approximation.

□

2.D Sampling distribution of the one-step estimator

As θ_{**} is the difference of two integrals, it suffices for us to consider explicit results for $\beta_\nu(h_1) = \int \log h_1(t)\nu(dt)$ only (nearly identical results hold for the $X = 0$ integral). We first derive the (non-parametric) efficient influence function of $\beta_\nu(h_1)$. Then, we derive an asymptotically linear one-step estimator. We conclude by detailing construction of valid Wald confidence intervals for the summary of interest, θ_{**}.

In addition to the assumption that right-censoring is uninformative within stratum defined by X, identification of $\beta_\nu(h_1)$ requires the (standard) technical assumption that no mass is placed beyond the support of the observation times. We define $\tau := \min_x \inf_y \{y : R_x(y) = 0\}$, which is equivalently the least upper bound for the support of the conditional observed time distribution. For the remainder of this section, we consider all integrals to be taken

over the time interval $(0, \tau)$. This identification condition can be viewed as a requirement on the weight function ν that zero mass be placed beyond time τ.

Theorem 2.3. *The non-parametric efficient influence function for $\beta_\nu(h_1)$ for fixed $\nu \geq 0$ is*

$$\phi_\nu^{(1)}(h_1) : z \mapsto \frac{x}{\pi}\left[\frac{\delta\nu(dy)}{h_1(y)R_1(y)} - \int_{u\leq y}\frac{\nu(du)}{R_1(u)}\right],$$

where $\nu(dy)$ denotes either the jump $\nu(y) - \nu(y-)$ at y or the density with respect to an appropriate dominating measure, if ν is continuous at y.

Proof. Suppose that the observation times are discrete. The extension to continuous time follows by a limiting argument similar to that employed by Chamberlain (1987). In discrete time (provided the aforementioned identification assumptions are met), we may estimate $h_1(t)$ consistently from the observed data with the plug-in estimator

$$h_{1,n}(t) := \frac{\sum_{i=1}^n X_i\Delta_i I(Y_i = t)}{\sum_{i=1}^n X_i I(Y_i \geq t)} = \frac{\frac{1}{n}\sum_{i=1}^n X_i\Delta_i I(Y_i = t)}{\frac{1}{n}\sum_{i=1}^n X_i I(Y_i \geq t)}.$$

Since the numerator of $h_{1,n}(t)$ is the mean of a fixed function, it is an asymptotically linear estimator of $E\{X\Delta I(Y = t)\} = \pi P(Y = t, \Delta = 1 \mid X = 1)$ with influence function

$$D^{(1)}(t) : z \mapsto x\delta I(y = t) - \pi P(Y = t, \Delta = 1 \mid X = 1).$$

Similarly, the denominator of $h_{1,n}(t)$ is an asymptotically linear estimator of $E\{XI(Y \geq t)\} = \pi P(Y \geq t \mid X = 1)$ with influence function

$$D^{(2)}(t) : z \mapsto xI(y \geq t) - \pi P(Y \geq t \mid X = 1).$$

By the delta method for influence functions, $h_{1,n}(t)$ is an asymptotically linear estimator of $h_1(t)$ with influence function

$$D^{(3)}(t) : z \mapsto \frac{x}{\pi}\frac{I(y \geq t)}{R_1(t)}\{\delta I(y = t) - h_1(t)\},$$

for each t such that $P(Y \geq t, \Delta = 1 \mid X = 1) > 0$. Similarly, $\log h_{1,n}(t)$ is asymptotically linear with influence function

$$D^{(4)}(t) : z \mapsto \frac{x}{\pi}\frac{I(y \geq t)}{h_1(t)R_1(t)}\{\delta I(y = t) - h_1(t)\}.$$

To complete the proof, we note that $\int \log h_{1,n}(u)\nu(du)$ satisfies the linearization

$$\int \{\log h_{1,n}(u) - \log h_1(u)\}\nu(du) = \frac{1}{n}\sum_{i=1}^{n}\int D^{(4)}(u; Z_i)\nu(du) + o_p(n^{-1/2}).$$

From the above linearization, we have that the influence function for $\beta_\nu(h_1)$ is

$$\begin{aligned}\phi_\nu^{(1)}(h_1) : z \mapsto &\int D^{(4)}(u; z)\nu(du) \\ &= \frac{x}{\pi}\int_{y\geq u}\frac{\nu(du)}{h_1(u)R_1(u)}\{\delta I(y = u) - h_1(u)\} \\ &= \frac{x}{\pi}\left[\frac{\delta\nu(dy)}{h_1(y)R_1(y)} - \int_{u\leq y}\frac{\nu(du)}{R_1(u)}\right].\end{aligned}$$

For the non-parametric model considered, $\phi_\nu^{(1)}(h_1)$ is the only influence function for $\beta_\nu(h_1)$. In particular, it is the nonparametric efficient influence function. □

In the case where $\nu = F$, the marginal time-to-event distribution, we must estimate the weight function from the observed data. This estimation of F alters the efficient influence function. We define $o \mapsto m_{x,h}^{(k)}(z)$ for $k \in \{1, 2, 3\}$, $x \in \{0, 1\}$ and h a fixed, positive function:

$$\begin{aligned} m_{x,h}^{(1)}(z) &:= \frac{\delta}{1 - G_x(y)}\log h(y) - \int \log h(u)F_x(du); \\ m_{x,h}^{(2)}(z) &:= \frac{1-\delta}{R_x(y)}\int_{u>y}\log h(u)F_x(du); \\ m_{x,h}^{(3)}(z) &:= \int C_x(u \wedge y)\log h(u)F_x(du), \end{aligned}$$

where $G_x(t) := P(C \leq t \mid X = x)$ and $C_x(t) := \int_{u<t} G_x(du)/R_x(u-)$. The functions $m_{x,h}^{(k)}$ arise as terms in the influence function of Kaplan-Meier integrals (Stute 1995).

Theorem 2.4. *Let $\nu := F$ be estimated from the data, where F is the marginal time-to-event distribution function. The non-parametric efficient influence function for $\beta_F(h_1)$ is*

$$\tilde{\phi}^{(1)}(h_1; z) := \phi_F^{(1)}(h_1; z) + \sum_{k=1}^{3} x\ m_{1,h_1}^{(k)}(z).$$

Proof. We will again conduct the proof in the setting of discrete observation times. First, we define $F_n := (1 - \pi_n)F_{0,n} + \pi_n F_{1,n}$ to be an estimator of F. Here, we take $F_{x,n}$ to be the

Kaplan-Meier estimator of the distribution function corresponding to $X = x$, and π_n to be the proportion of study participants with $X = 1$. For an estimator $h_{1,n}$ that is consistent for h_1, the plug-in estimator

$$\beta_{F_n}(h_{1,n}) := \int \log h_{1,n}(u) F_n(du)$$

can be expanded as

$$\begin{aligned}\beta_{F_n}(h_{1,n}) - \beta_F(h_1) &= \int \{\log h_{1,n}(u) - \log h_1(u)\} F(du) \\ &\quad + \int \log h_1(u)\{F_n(du) - F(du)\} + \mathbb{G}_{2,n},\end{aligned}$$

where $\mathbb{G}_{2,n} := \int \{\log h_{1,n}(u) - \log h_1(u)\}\{F_n(du) - F(du)\}$. For now, we assume $\mathbb{G}_{2,n} = o_p(n^{-1/2})$. This assumption certainly holds when plug-in estimators are utilized in a discrete-time setting (as both $h_{1,n}$ and F_n are root-n consistent). By Theorem 2.3, the first integral contributes $\phi_\nu^{(1)}(h_1)$ to the influence function; we will set formal conditions for asymptotic linearity below. For the second integral, we find

$$\begin{aligned}&\int \log h_1(u)\{F_n(du) - F(du)\} \\ &= \pi_n \int \log h_1(u)\{F_{1,n}(du) - F_1(du)\} + (1-\pi_n) \int \log h_1(u)\{F_{0,n}(du) - F_0(du)\} \\ &= \pi \int \log h_1(u)\{F_{1,n}(du) - F_1(du)\} + (1-\pi) \int \log h_1(u)\{F_{0,n}(du) - F_0(du)\} + o_p(n^{-1/2}) \\ &= \pi \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{3} \frac{X_i}{\pi} m_{1,h_1}^{(k)}(Z_i) + (1-\pi) \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{3} \frac{1-X_i}{1-\pi} m_{0,h_1}^{(k)}(Z_i) + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{3} m_{X_i,h_1}^{(k)}(Z_i) + o_p(n^{-1/2}),\end{aligned}$$

where we applied the linearization of Stute (1995) to the stratum-specific Kaplan-Meier integrals of h_1 (which is fixed in this expansion). Adding the components of the influence function, we find that the non-parametric efficient influence function of $\beta_F(h_1)$ is $\tilde{\phi}^{(1)}(h_1)$. □

Having derived the (non-parametric) efficient influence function for $\beta_F(h_1)$, we turn to construction of efficient estimators. For initial hazard estimator $h_{1,n}$, pathwise differentiabi-

lity of $\beta_F(h_1)$ permits us to write

$$\int \log h_{1,n}(t)F_n(dt) - \int \log h_1(t)F(dt) = -\int \tilde{\phi}^{(1)}(h_{1,n}; z)dP(z) + r_n^{(1)}$$
$$= \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}^{(1)}(h_1; Z_i) \; - \; \mathbb{B}_n^{(1)} + \mathbb{G}_n^{(1)} + r_n^{(1)}$$

with $r_n^{(1)}$ a second-order remainder, $\mathbb{B}_n^{(1)} := \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}^{(1)}(h_{1,n}; Z_i)$ a first-order bias, and

$$\mathbb{G}_n^{(1)} := \frac{1}{n}\sum_{i=1}^{n}\{\tilde{\phi}^{(1)}(h_{1,n}; Z_i) - \tilde{\phi}^{(1)}(h_1; Z_i)\} - \int\{\tilde{\phi}^{(1)}(h_{1,n}; z) - \tilde{\phi}^{(1)}(h_{1,n}; z)\}dP(z).$$

We define the one-step bias corrected estimator $\beta_{n,\mathrm{OS}}^{(1)} := \int \log h_{1,n}(t)F_n(dt) + \mathbb{B}_n^{(1)}$. The following result provides one set of conditions under which $\beta_{n,\mathrm{OS}}^{(1)}$ is an asymptotically linear estimator of $\beta_F(h_1)$.

Theorem 2.5. *Suppose that we observe n independent triples $Z_i := (Y_i, \Delta_i, X_i)$, where $Y_i = \min\{T_i, C_i\}$ is the observation time, $\Delta_i = I(T_i \leq C_i)$ is the event indicator, and $X_i \in \{0, 1\}$ denotes the exposure group for study participant i. Suppose that the right-censoring mechanism is uninformative within exposure groups: $T_i \perp\!\!\!\perp C_i \mid X_i$. Suppose we have constructed estimators, $h_{1,n}(t)$ and $R_{1,n}(t)$ of the conditional hazard function, $h_1(t)$, and conditional observation time distribution function $R_1(t)$ for the $X = 1$ stratum and that*

1. *the (estimated) influence functions have $\int\{\tilde{\phi}^{(1)}(h_{1,n}; z) - \tilde{\phi}^{(1)}(h_1; z)\}^2 dP(z) = o_p(1)$ and the maps $\tilde{\phi}^{(1)}(h_{1,n})$ belong to a P-Donsker class with probability increasing to one,*

2. *the nuisance parameter estimators converge sufficiently fast:*

$$\int\{R_{1,n}(t) - R_1(t)\}^2 F(dt) = o_p(n^{-1/2}) \quad \textit{and} \quad \int\{h_{1,n}(t) - h_1(t)\}^2 F(dt) = o_p(n^{-1/2})$$

and the maps $h_{1,n}$ belong to a P-Donsker class with probability increasing to one.

Then, the one-step estimator is consistent asymptotically normal,

$$n^{1/2}\{\beta_{n,OS}^{(1)} - \beta_F(h_1)\} \rightsquigarrow \mathcal{N}(0, \sigma_P^2)$$

with $\sigma_P^2 := \int \tilde{\phi}^{(1)}(h_{1,n}; z)^2 dP(z)$.

Proof. We start with the expansion

$$\beta_{n,\mathrm{OS}}^{(1)} - \beta_F(h_1) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\phi}^{(1)}(h_1; Z_i) + \mathbb{G}_n^{(1)} + r_n^{(1)}.$$

Condition 1 ensures that $\mathbb{G}_n^{(1)} = o_p(n^{-1/2})$, by a standard result in the theory of empirical processes (van der Vaart 1998, Lemma 19.24). To study the remainder term, we recall that

$$\begin{aligned} &\beta_{F_n}(h_{1,n}) - \beta_F(h_1) \\ &= \int \{\log h_{1,n}(u) - \log h_1(u)\} F(du) + \int \log h_1(u)\{F_n(du) - F(du)\} + \mathbb{G}_{2,n}^{(1)}, \end{aligned} \tag{2.1}$$

where $\mathbb{G}_{2,n}^{(1)} := \int \{\log h_{1,n}(u) - \log h_1(u)\}\{F_n(du) - F(du)\}$. We again have $\mathbb{G}_{2,n}^{(1)} = o_p(n^{-1/2})$ by standard empirical process results. Using this representation, we will write $r_n^{(1)} = r_{n,1}^{(1)} + r_{n,2}^{(1)}$ for remainders $r_{n,1}^{(1)}$ and $r_{n,2}^{(1)}$ defined by expansion of the integrals on the right-hand side of (2.1). First, we have $\int \{\log h_{1,n}(u) - \log h_1(u)\} F(du) = -\int \phi_F^{(1)}(h_{1,n}; z) dP(z) + r_{n,1}^{(1)}$ by pathwise differentiability of $\beta_\nu(h_1)$ for fixed ν. Rearranging terms, we find

$$r_{n,1}^{(1)} = \int \{\log h_{1,n}(u) - \log h_1(u)\} F(du) + \int \phi_F^{(1)}(h_{1,n}; z) dP(z). \tag{2.2}$$

We use a first-order expansion around $\alpha = 1$ for $\alpha \mapsto \log(\alpha)$ for the first term of (2.2):

$$\int \{\log h_{1,n}(u) - \log h_1(u)\} F(du) = -\int \left\{ \left(\frac{h_1(u)}{h_{1,n}(u)} - 1 \right) - \frac{1}{2\alpha_n(u)} \left(\frac{h_{1,n}(u)}{h_1(u)} - 1 \right)^2 \right\} F(du)$$

with $\alpha_n(u)$ a value between 1 and $h_{1,n}(u)/h_1(u)$. Expanding the second term of (2.2), we find that

$$\begin{aligned} \int \phi_F^{(1)}(h_{1,n}; z) dP(z) &= \int \frac{x}{\pi} \left[\frac{\delta\nu(dy)}{R_{1,n}(y) h_{1,n}(y)} - \int_{u \le y} \frac{\nu(du)}{R_{1,n}(u)} \right] dP(z) \\ &= \int \left[\frac{\delta\nu(dy)}{R_{1,n}(y) h_{1,n}(y)} - \int_{u \le y} \frac{\nu(du)}{R_{1,n}(u)} \right] dP_1(z) \\ &= \int \frac{R_1(y) h_1(y) \nu(dy)}{R_{1,n}(y) h_{1,n}(y)} - \int \int_{u \le y} \frac{\nu(du)}{R_{1,n}(u)} dP_1(z) \\ &= \int \frac{R_1(y) h_1(y)}{R_{1,n}(y) h_{1,n}(y)} \nu(dy) - \int \frac{R_1(u)}{R_{1,n}(u)} \nu(du) \\ &= \int \frac{R_1(u)}{R_{1,n}(u)} \left\{ \frac{h_1(u)}{h_{1,n}(u)} - 1 \right\} \nu(du). \end{aligned}$$

From our simplification of the terms in $r_{n,1}^{(1)}$, we conclude that this remainder is

$$r_{n,1}^{(1)} = \int \left\{ \frac{R_1(u)}{R_{1,n}(u)} - 1 \right\} \left\{ \frac{h_1(u)}{h_{1,n}(u)} - 1 \right\} \nu(du) - \int \frac{1}{2\alpha_n(u)} \left(\frac{h_{1,n}(u)}{h_1(u)} - 1 \right)^2 F(du).$$

Under the rates of Condition 2, $r_{n,1}^{(1)} = o_p(n^{-1/2})$. We now consider the second term of (2.1). By pathwise differentiability, this term will also contribute a second-order term $r_{n,2}^{(1)}$ to the remainder. This contribution arises in a similar expansion. In practice, (asymptotically linear) Kaplan-Meier estimators will be used to construct F_n, so this remainder will be expected to be $o_p(n^{-1/2})$ as well. Thus, we conclude that

$$\beta_{n,\mathrm{OS}}^{(1)} - \beta_F(h_1) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}^{(1)}(h_1; Z_i) + o_p(n^{-1/2}),$$

completing the proof by application of Slutsky's theorem and the central limit theorem. □

Entirely similar results can be established for the one-step estimator of $\beta_F(h_0)$. Hence, we can write

$$\theta_{\mathrm{OS},n} := \int \theta_n(t) F_n(dt) + \frac{1}{n} \sum_{i=1}^{n} \phi_n(Z_i) \ = \beta_{n,\mathrm{OS}}^{(1)} - \beta_{n,\mathrm{OS}}^{(0)},$$

with $\phi_n(z) := \phi_{F_n}^{(1)}(h_{1,n}; z) - \phi_{F_n}^{(0)}(h_{0,n}; z)$. Note that, due to the use of the Kaplan-Meier estimator in construction of F_n, the sum $n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{3} m_{X_i,h_{1,n}}^{(k)}(Z_i) = 0$. This is a well-known property of non-parametric maximum likelihood estimators. Wald confidence intervals can be constructed for θ_{**} based on the standard error estimator

$$\frac{1}{n} \left[\sum_{i=1}^{n} \{ \tilde{\phi}^{(1)}(h_1; Z_i) - \tilde{\phi}^{(0)}(h_0; Z_i) \}^2 \right]^{1/2}.$$

2.E Sample R code for computing PLAM estimators

```
library(SuperLearner)
library(np)
```

```
# Small sample library of density estimators to select from by cross-
   validation
bwList <- exp(seq(log(.05), log(20), length=5))

SL.npreg <- function (Y, X, newX,
                      family = gaussian(),
                      obsWeights = rep(1, length(Y)), bwx = 1,
                      rangeThresh = 1e-07, ...) {
  options(np.messages = FALSE)
  if (abs(diff(range(Y))) <= rangeThresh) {
    thisMod <- glm(Y ~ 1, data = X)
  } else {
    thisMod <- npreg(as.formula(paste('Y ~', paste(names(X), collapse = '+')
   )),
                     data = X, newdata = X, bws = bwx*1.06*sd(Y)*length(Y)
   ^-.2,
                     remin = FALSE)
  }
  pred <- predict(thisMod, newdata = newX)
  fit <- list(object = thisMod)
  class(fit) <- 'SL.npreg'
  out <- list(pred = pred, fit = fit)
  return(out)
}

create.Learner("SL.npreg", tune = list(bwx = bwList))
SL.lib <- paste0("SL.npreg_", seq_along(bwList))
```

```
# One run of the simulations described in Section 2.2 of the chapter
do.one <- function(n, PLAM=TRUE){
  # Generate a set of observations
  w <- runif(n, min=-2, max=2)
  x <- rbinom(n, size = 1, prob=plogis(.5 + 2*w - .5*w^2))
  if(!PLAM) y <- x*w + (1-x)*(-w) + rnorm(n)
  if(PLAM) y <- .2*x + w^2 + rnorm(n)

  # Estimators of X-W regression function
  g.n <- SuperLearner(Y=x, X=data.frame(X=w), SL.library=SL.lib, cvControl=
   list(V=5L))
  g.n$est <- g.n$SL.predict

  x.g <- x-g.n$est

  # Modelbased estimator of nonparametric function
  Q0.n <- SuperLearner(Y=y[x==0], X=data.frame(X=w[x==0]), newX=data.frame(X
   =w), SL.library=SL.lib, cvControl=list(V=5L))
  Q0.n$est <- Q0.n$SL.predict

  # Initial estimator
  xx.g <- x.g*x
  B.g <- mean(x.g*y)/mean(xx.g)
  y.g <- y-B.g*x

  # Backfit estimator of nonparametric function
```

```
  Qe.g <- SuperLearner(Y=y.g, X=data.frame(X=w), SL.library=SL.lib,
   cvControl=list(V=5L))
  Qe.g$est <- Qe.g$SL.predict

  phi.Q0.g <- x.g*(y.g - Q0.n$est)/mean(xx.g) # EIF based on Q0
  phi.Qe.g <- x.g*(y.g - Qe.g$est)/mean(xx.g) # EIF based on eta

  # One-step estimators
  B.Q0.g <- B.g+mean(phi.Q0.g)
  B.Qe.g <- B.g+mean(phi.Qe.g)

  return(c(B.g = B.g , B.Q0.g = B.Q0.g , B.Qe.g = B.Qe.g))
}

set.seed(21082019)
do.one(n=100)
#       B.g    B.Q0.g    B.Qe.g
# 0.3631308 0.2951544 0.5406715
```

2.F Sample R code for computing the one-step estimator

```
library(survival)
library(muhaz)

# One run of the simulations described in Section 2.3
do.one <- function(n, pi=2/3, MISSPEC=TRUE, censoRate=.2){
  x <- rbinom(size = 1, prob = pi, n = n)
  if(MISSPEC){
    t.event <- rweibull(n, scale=1, shape=1+.5*x)
```

```
} else {
  t.event <- rweibull(n, scale=1, shape=1)
}

t.cens <- rexp(n, rate = censoRate)
y <- pmin(t.event, t.cens)
d <- t.event <= t.cens

tauX <- c(max(y[x==0 & d==1]), max(y[x==0 & d==0]),
          max(y[x==1 & d==1]), max(y[x==1 & d==0]))
tau <- min(tauX[c(1,3)])
tau <- min(tauX)

## Estimating the difference in log-hazards
haz0 <- muhaz(y, d, subset={x==0},
              min.time=min(y),
              max.time=tau,
              n.est.grid=length(y[d==1]))
haz1 <- muhaz(y, d, subset={x==1},
              min.time=min(y),
              max.time=tau, n.est.grid=length(y[d==1]))
beta.n <- approxfun(x=haz0$est.grid, y=log(haz1$haz.est)-log(haz0$haz.est)
 )
beta.fitted <- beta.n(y)
beta.fitted[is.na(beta.fitted)] <- 0
beta.fitted[is.infinite(beta.fitted)] <- 0
```

```
## Estimating the censoring distributions
cens <- survfit(Surv(y,1-d)~x)
cens0.n <- with(cens[1], approxfun(x=time, y=surv, method = "constant",
 yleft = 1, yright = 0))
cens1.n <- with(cens[2], approxfun(x=time, y=surv, method = "constant",
 yleft = 1, yright = 0))
G0.fitted <- cens0.n(y)
G1.fitted <- cens1.n(y)

G0.stable <- ifelse(G0.fitted>0, 1/G0.fitted, 0)*(y<=tau) #Average over
 support of distributions only
G1.stable <- ifelse(G1.fitted>0, 1/G1.fitted, 0)*(y<=tau)

## Initial estimators
psi.n <- mean((d * ((1-x)*G0.stable+x*G1.stable) * beta.fitted)*(y<=tau))
 #est beta

cox.n <- coef(coxph(Surv(y,d)~x))

## Estimating the event distributions
evnt <- survfit(Surv(y,d)~x)
evnt0.n <- with(evnt[1], approxfun(x=time, y=surv, method = "constant",
 yleft = 1, yright = 0))
evnt1.n <- with(evnt[2], approxfun(x=time, y=surv, method = "constant",
 yleft = 1, yright = 0))
S0.fitted <- evnt0.n(y)*(y<=tau)
S1.fitted <- evnt1.n(y)*(y<=tau)
```

```
## Estimating the observed time distributions
H <- survfit(Surv(y, rep(1, length(y)))~x)
H0.n <- with(H[1], approxfun(x=time, y=surv, method = "constant", yleft =
 1, yright = 0))
H1.n <- with(H[2], approxfun(x=time, y=surv, method = "constant", yleft =
 1, yright = 0))
H0.fitted <- H0.n(y)
H1.fitted <- H1.n(y)
H0.stable <- ifelse(H0.fitted>0, 1/H0.fitted, 0)*(y<=tau) #Average over
 support of distributions only
H1.stable <- ifelse(H1.fitted>0, 1/H1.fitted, 0)*(y<=tau)

## Computing EIC terms

gamma0.0 <- d*beta.fitted*G0.stable*(1-x)*{y<=tau}
gamma0.1 <- d*beta.fitted*G1.stable*x*{y<=tau}

y0 <- y1 <- y2 <- y
gamma1.0 <- (1-x)*(1-d)*H0.stable * sapply(y0, function(t) {
  mean({t<=pmin(y,tau)}*d*{x==0}*beta.fitted*G0.stable) / (1-pi)
}
)
gamma1.1 <- x*(1-d)*H1.stable * sapply(y0, function(t) {
  mean({t<=pmin(y,tau)}*d*{x==1}*beta.fitted*G1.stable) / pi
}
)
```

```
gamma2.0 <- (1-x)*sapply(y, function(t) {
  mean(d*{x==0}*G0.stable*beta.fitted*
         sapply(
           y1, function(u) {
             mean((1-d)*{x==0}*{y2<pmin(u,pmin(t, tau))}*H0.stable^2)
           }
         ))/((1-pi)^2)
}
)
gamma2.1 <- x*sapply(y, function(t) {
  mean(d*{x==1}*G1.stable*beta.fitted*
         sapply(
           y1, function(u) {
             mean((1-d)*{x==1}*{y2<pmin(u,pmin(t, tau))}*H1.stable^2)
           }
         ))/(pi^2)
}
)

gamma3.0 <- -{x==0}*(
  d*H0.stable*{pi*S1.fitted*exp(beta.fitted)+(1-pi)*S0.fitted}-
    sapply(y, function(t){
      mean(d*{{x==1}*G1.stable+{x==0}*G0.stable}*H0.stable*{pmin(tau,t)>=
 y0})
    })
)/(1-pi)
```

```
  gamma3.1 <- {x==1}*(
    d*H1.stable*{pi*S1.fitted+(1-pi)*S0.fitted*exp(-beta.fitted)}-
      sapply(y, function(t){
        mean(d*{{x==1}*G1.stable+{x==0}*G0.stable}*H1.stable*{pmin(tau,t)>=
   y0})
      })
  )/pi

  influence <- (gamma0.0+gamma1.0-gamma2.0+gamma3.0)+
    (gamma0.1+gamma1.1-gamma2.1+gamma3.1) - psi.n
  # influence <- (gamma0.0+gamma3.0)+
  #   (gamma0.1+gamma3.1) - psi.n

  b.n <- mean(influence)
  names(cox.n) <- NULL

  c(init=psi.n, one_step=psi.n+b.n, Cox=cox.n)
}

set.seed(21082019)
do.one(n = 500, MISSPEC = TRUE, censoRate = .2)
#       init   one_step        Cox
# 0.11891482 0.06355955 0.09371771
```

Chapter 3

A GENERAL APPROACH TO DOUBLY ROBUST INFERENCE

3.1 Introduction

The study of doubly robust estimators is an area of great interest in the causal inference literature. An estimator is said to be doubly robust if it remains consistent for the target parameter provided at least one of two nuisance parameters involved in its construction is consistently estimated. Doubly robust estimators often arise in the pursuit of efficient estimators (van der Laan and Robins 2003). In such cases, when both nuisances are estimated consistently, the doubly robust estimator is efficient. That is, the estimator is regular and asymptotically linear with variance equal to the variance of the efficient influence function evaluated at an observation. Asymptotic linearity of the estimator ensures that standard Wald methods of inference based on a normal approximation to the sampling distribution will be asymptotically valid.

Unfortunately, the notion of double robustness discussed so far does not typically confer doubly robust asymptotic linearity. When data-adaptive nuisance estimators (e.g. nonparametric regression) are used in construction of the estimator, inconsistent estimation of one nuisance results in non-negligible first order contributions to the sampling distribution. These terms contribute bias with respect to the target parameter that invalidates the use of standard Wald inference. We will call an estimator consistent asymptotically normal doubly robust (CANDoR) if it is both consistent and asymptotically linear for the target parameter provided at least one of two nuisance parameters is estimated consistently. Hence, a CANDoR estimator provides us with both doubly robust estimation (i.e., consistency) and inference (based on asymptotic normality). As quantifying uncertainty about point estima-

tes is of paramount importance, the construction of CANDoR estimators represents a key development in causal inference and other research areas in which doubly robust estimators.

Pioneering work by van der Laan (2014) and Benkeser et al. (2017) detailed and evaluated the construction of CANDoR estimators for the average treatment effect. Their estimators are based on linearizations of a remainder term arising from an expansion for pathwise differentiable parameters. This approach has also been successfully applied to estimate the average treatment effect in randomized trials with outcomes missing at random (Díaz and van der Laan 2017) and observational studies of survival with informative right-censoring (Díaz 2019). Notably, each of these target parameters and the associated remainder terms have a closed form that simplifies the analysis.

Many parameters (and doubly robust estimators) are defined implicitly as solutions to optimization problems or sets of equations, potentially prohibiting direct study of the estimation remainder. Examples include quantile treatment effects and the coefficients from the nonparametric projection onto a marginal structural model (Neugebauer and van der Laan 2007). For such parameters, one may ask whether a similar approach may be used to construct a CANDoR estimator. The primary contribution of this chapter is to demonstrate that the answer is affirmative for a large class of parameters. Our approach first requires a tool for deriving a tractable closed-form first-order approximation to the estimation remainder.

In Díaz and van der Laan (2017) and Díaz (2019), the authors note that the linearization of the remainder term in van der Laan (2014) and Benkeser et al. (2017) is equivalent to a first-order approximation of the so-called "drift" term arising in expansions for estimating equations-based estimators (van der Vaart 1998). While the drift term need not be more tractable than the aforementioned remainder, we may (as in Chapter 2) study the drift through first-order expansion based on a particular Gâteaux derivative. In this article, we introduce a representation for the estimation remainder that instead leverages pathwise differentiability of the target parameter. We show that this term is equal in first-order to the expansion of the drift term using Gâteaux derivatives. Having replaced the implicitly-defined remainder with a closed form approximation, we return to the problem of constructing

CANDoR estimators.

We utilize our novel representation to linearize the remainder for doubly robust estimators in terms of score functions that depend on additional nuisance parameters. We derive a general result that applies to target parameters when observations are generated subject to missingness at random. In this setting, we detail the implementation of targeted minimum loss-based estimators that update a set of initial nuisance estimators (van der Laan and Rubin 2006). The resulting nuisance estimators solve additional score equations defined by terms in our linearization of the remainder. We show that, under weak conditions, the plug-in estimator of the target parameter based on the updated nuisance estimators is asymptotically linear with known influence function provided at least one of the two nuisance parameters is estimated consistently. Hence, our estimators lead naturally to doubly robust inference in the form of valid Wald confidence intervals and hypothesis tests. We illustrate our proposed methodology in simulation studies for counterfactual quantiles, which can be utilized to define quantile treatment effects.

3.2 Doubly robustness

3.2.1 Doubly robust estimation

In this section we formally define what we mean by a doubly robust estimator and discuss the commonly used one-step and targeted minimum loss-based approaches for constructing such an estimator.

Let P_0 denote the probability measure for observed data unit Z. We suppose that P_0 is contained in a nonparametric model $\mathcal{M}$. We wish to estimate the evaluation $\psi_0 := \Psi(P_0)$ of a pathwise differentiable parameter $\Psi : \mathcal{M} \to \mathbb{K}$, where $\mathbb{K}$ may be a Euclidean or function space. In the remainder, we focus on Euclidean parameters explicitly, taking $\mathbb{K} = \mathbb{R}^p$. We further suppose that ψ_0 depends on P_0 only through a nuisance parameter $\bar{Q}_0 := \bar{Q}(P_0)$ and write $\Psi(\bar{Q}_0)$ instead of $\Psi(P_0)$.

In the nonparametric model under consideration, all regular asymptotically linear estima-

tors of ψ_0 have the same (efficient) influence function (van der Vaart 1998). We suppose that the efficient influence function, $\phi^*_{\bar{Q},g}$, depends on P_0 only through $\bar{Q}_0$ and an additional (variation independent) nuisance parameter $g_0 := g(P_0)$. For such parameters, it is often the case that the efficient influence function is doubly robust so that $P_0\phi^*_{\bar{Q},g} = \int \phi^*_{\bar{Q},g}(z)dP_0(z) = 0$ provided either $\bar{Q} = \bar{Q}_0$ or $g = g_0$ (almost surely-P_0). We will continue to use the notation $Pf = \int f(z)dP(z)$ and $\|f\|^2_{L_2(P)} = P(f^2)$ throughout the remainder of the chapter. Many examples of parameters with doubly robust influence functions arise in coarsening at random models and closely related causal inference settings (van der Laan and Robins 2003).

We now review standard methods to construct doubly robust estimators of parameters with doubly robust influence functions. To start, we suppose that $\bar{Q}_n$ and g_n are estimators of the nuisance parameters $\bar{Q}_0$ and g_0, and define the plug-in estimator $\psi_{n,\text{PLUGIN}} := \Psi(\bar{Q}_n)$. Following Pfanzagl (1982), pathwise differentiability leads to the asymptotic expansion

$$\begin{aligned}\psi_{n,\text{PLUGIN}} - \psi_0 &= -P_0\phi^*_{\bar{Q}_n,g_n} + R_{\bar{Q}_n,g_n} \\ &= \mathbb{P}_n\phi^*_{\bar{Q}_*,g_*} - \mathbb{P}_n\phi^*_{\bar{Q}_n,g_n} + (\mathbb{P}_n - P_0)(\phi^*_{\bar{Q}_n,g_n} - \phi^*_{\bar{Q}_*,g_*}) + R_{\bar{Q}_n,g_n},\end{aligned} \tag{3.1}$$

where $\bar{Q}_*$ and g_* are limits of the nuisance estimators such that $\|\bar{Q}_n - \bar{Q}_*\|^2_{L_2(P_0)} = o_p(1)$ and $\|g_n - g_*\|^2_{L_2(P_0)} = o_p(1)$, and $R_{\bar{Q}_n,g_n}$ is a remainder term that is $o(\rho((\bar{Q}_n, g_n), (\bar{Q}_0, g_0)))$ for an appropriate distance ρ defined on the nuisance parameter space. We wish to use (3.1) to derive an asymptotically linear estimator of ψ_0 with influence function $\phi^*_{\bar{Q}_*,g_*}$, so that the estimator is equal to $\psi_0 + \mathbb{P}_n\phi^*_{\bar{Q}_*,g_*} + o_p(n^{-1/2})$.

The term $\mathbb{P}_n\phi^*_{\bar{Q}_n,g_n}$ is a first-order bias that may preclude $\psi_{n,\text{PLUGIN}}$ from being regular and asymptotically linear. Adjustment for this bias commonly proceeds by one of two methods of updating the initial plug-in estimator. In particular, we consider replacing the initial plug-in estimator $\psi_{n,\text{PLUGIN}}$ in the expansion by a one-step or targeted maximum likelihood estimator. The one-step bias-correction $\psi_{n,\text{1STEP}} := \psi_{n,\text{PLUGIN}} + \mathbb{P}_n\phi^*_{\bar{Q}_n,g_n}$ is obtained by adding the bias term to the plug-in estimator. Alternatively, one may utilize the targeted maximum likelihood estimation framework (van der Laan and Rubin 2006) to replace $\bar{Q}_n$ by a targeted version $\bar{Q}^*_n$ that solves the score equation $\mathbb{P}_n\phi^*_{\bar{Q}^*_n,g_n} = 0$, thereby annihilating the bias term.

The targeted maximum likelihood estimator $\psi_{n,\text{TMLE}} := \Psi(\bar{Q}_n^*)$ is then a plug-in estimator of ψ_0.

Returning to the expansion, the term $(\mathbb{P}_n - P_0)(\phi^*_{\bar{Q}_n,g_n} - \phi^*_{\bar{Q}_*,g_*})$ can be studied by empirical process methods. If, for instance, $\phi^*_{\bar{Q}_n,g_n} - \phi^*_{\bar{Q}_*,g_*}$ belongs (asymptotically) almost surely-P_0 to a P_0-Donsker class and $\|\phi^*_{\bar{Q}_n,g_n} - \phi^*_{\bar{Q}_*,g_*}\|^2_{L_2(P_0)} = o_p(1)$, then, by Lemma 19.24 in van der Vaart (1998), we have that $(\mathbb{P}_n - P_0)(\phi^*_{\bar{Q}_n,g_n} - \phi^*_{\bar{Q}_*,g_*}) = o_p(n^{-1/2})$. Rather than imposing conditions on the class of functions containing $\phi^*_{\bar{Q}_n,g_n} - \phi^*_{\bar{Q}_*,g_*}$, cross-fitting techniques that ensure this term is $o_p(n^{-1/2})$ may be used in the construction of the one-step or targeted maximum likelihood estimator (van der Vaart 1998, Zheng and van der Laan 2011).

The remainder term $R_{\bar{Q}_n,g_n} := \psi_{n,\text{PLUGIN}} - \psi_0 + P_0\phi^*_{\bar{Q}_n,g_n}$ will play a key role in later sections of the chapter. For a large class of problems with doubly robust influence functions, the remainder takes the product form

$$R_{\bar{Q}_n,g_n} = P_0\{s_{ab}(a_n - a_0)(b_n - b_0)\}, \tag{3.2}$$

where $a_n := \alpha(\bar{Q}_n)$, $a_0 := \alpha(\bar{Q}_0)$, $b_n := \beta(g_n)$, and $b_0 := \beta(g_0)$ for mappings α and β of the nuisance parameters, and s_{ab} is a known function from $\mathcal{Z}$ to $\mathbb{R}$, where $\mathcal{Z}$ is the support of Z under P_0 (Rotnitzky et al. 2019). This structure occurs naturally in many important causal inference applications. When the remainder has this product form, some notable properties follow:

Proposition 3.1. *If the remainder takes the product form* (3.2)*, the following implications hold:*

1. *if either* $\bar{Q}_* = \bar{Q}_0$ *or* $g_* = g_0$*, then* $R_{\bar{Q}_n,g_n} = o_p(1)$*; and*

2. *if* $\|\bar{Q}_n - \bar{Q}_0\|^2_{L_2(P_0)} = \|g_n - g_0\|^2_{L_2(P_0)} = o_p(n^{-1/2})$*, then* $R_{\bar{Q}_n,g_n} = o_p(n^{-1/2})$.

The first part of this proposition allows us to conclude that both $\psi_{n,\text{1STEP}} = \psi_0 + o_p(1)$ and $\psi_{n,\text{TMLE}} = \psi_0 + o_p(1)$ provided *either* $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. That is, the one-step and targeted maximum likelihood estimators are doubly robust estimators of ψ_0. Moreover, if

both nuisance parameters are estimated consistently at an appropriate rate, and the empirical process term is negligible, then the second part of Proposition 3.1 tells us that one-step and targeted maximum likelihood estimators are asymptotically efficient estimators of ψ_0. In particular, $\psi_{n,\text{1STEP}} - \psi_0 = \mathbb{P}_n \phi^*_{\bar{Q}_0,g_0} + o_p(n^{-1/2})$ and $\psi_{n,\text{TMLE}} - \psi_0 = \mathbb{P}_n \phi^*_{\bar{Q}_0,g_0} + o_p(n^{-1/2})$, so that both estimator sequences are regular asymptotically linear with influence function equal to the efficient influence function. Hence, normal-based inference wherein the sampling distribution of either estimator is estimated by $\mathcal{N}(\psi_0, n^{-1}\mathbf{\Sigma}_0)$ is valid for sufficiently large sample sizes, where the asymptotic covariance matrix is given by $\mathbf{\Sigma}_0 := P_0\left\{\phi^*_{\bar{Q}_0,g_0}\phi^{*\top}_{\bar{Q}_0,g_0}\right\}$.

3.2.2 Doubly robust inference

The second guarantee from Proposition 3.1 no longer holds when merely one of $\bar{Q}_0$ and g_0 is estimated consistently. Heuristically, this is because the remainder $R_{\bar{Q}_n,g_n}$ then inherits the convergence rate of the consistent nuisance estimator. This precludes the $n^{1/2}$-consistency of the one-step and targeted maximum likelihood estimators, leading to complex sampling distributions, and significantly complicating inference beyond point estimation. In such cases, one may wonder whether and when it is possible to construct a CANDoR estimator, say $\psi_{n,\text{CANDR}}$, provided the nuisance estimators converge fast enough to their limits. When such a construction is possible, and provided at least one of $\bar{Q}_0$ and g_0 is estimated consistently, then the hypothetical estimator satisfies

$$n^{1/2}(\psi_{n,\text{CANDR}} - \psi_0) \rightsquigarrow \mathcal{N}(0, \mathbf{\Sigma}_*)$$

for some covariance matrix $\mathbf{\Sigma}_*$ that depends upon the limits $\bar{Q}_*$ and g_* of the nuisance estimators.

The doubly robust asymptotic normality of such an estimator has immediate practical implications for inference. Suppose interest lies in $\tau(\psi_0)$ for some differentiable function $\tau : \mathbb{R}^p \mapsto \mathbb{R}$. If one of $\bar{Q}_0$ and g_0 is estimated consistently, and the analyst constructs a consistent estimator $\mathbf{\Sigma}_n$ of $\mathbf{\Sigma}_*$, then standard Wald-type $(1-\alpha)\times 100\%$ confidence intervals

defined by the limits

$$\tau(\psi_{n,\text{CANDR}}) \mp z_{1-\alpha/2} \left\{ \frac{1}{n} \dot{\tau}(\psi_{n,\text{CANDR}})^\top \mathbf{\Sigma}_n \dot{\tau}(\psi_{n,\text{CANDR}}) \right\}^{1/2}$$

will be asymptotically valid. This follows readily from the delta method and Slutsky's lemma. In the above display, $\dot{\tau}$ denotes the gradient of τ and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Similarly, α-level Wald hypothesis tests based on $\psi_{n,\text{CANDR}}$ are valid – the type I error rate of these tests will not exceed their nominal α-level asymptotically.

Hence, the extension from doubly robust estimators to CANDoR estimators has important ramifications beyond point estimation. In particular, such estimators maintain the validity of standard inferential techniques even when one of the nuisance parameters is estimated inconsistently.

Having reviewed the desirable properties of CANDoR estimators, we now turn to expanding upon the current examples of such estimators. The existing constructions typically require identifying and approximating the first-order behavior of the estimation remainder $R_{\bar{Q}_n,g_n}$. We first review current approaches for obtaining such approximations and then propose a novel expression for the remainder.

3.3 Approximating estimation remainders

3.3.1 Existing, direct approach to approximation

If only one of the nuisance estimators is consistent in the above sense, then the remainder $R_{\bar{Q}_n,g_n}$ will comprise both first-order contributions $\dot{R}_{\bar{Q}_n,g_n}$ and higher-order contributions $\ddot{R}_{\bar{Q}_n,g_n}$. In the analysis of the average treatment effect, one such decomposition

$$R_{\bar{Q}_n,g_n} = \dot{R}_{\bar{Q}_n,g_n} + \ddot{R}_{\bar{Q}_n,g_n} \tag{3.3}$$

was derived by van der Laan (2014). Similar, explicit decompositions feature in the consistent asymptotically normal doubly robust estimator constructions of Benkeser et al. (2017), Díaz and van der Laan (2017), and Díaz (2019). Common to each of the settings considered is a closed-form expression for the estimation remainder and also for $\dot{R}_{\bar{Q}_n,g_n}$ and $\ddot{R}_{\bar{Q}_n,g_n}$.

To illustrate this direct approach in some generality, we suppose for the moment that $R_{\bar{Q}_n,g_n}$ is of the product form (3.2), and that either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. In this setting, $R_{\bar{Q}_n,g_n}$ can be expressed as a sum as in (3.3) with

$$\dot{R}_{\bar{Q}_n,g_n} := P_0\{s_{ab}(a_* - a_0)(b_n - b_*)\} + P_0\{s_{ab}(a_n - a_*)(b_* - b_0)\} \tag{3.4}$$

$$\ddot{R}_{\bar{Q}_n,g_n} := P_0\{s_{ab}(a_n - a_*)(b_n - b_*)\}. \tag{3.5}$$

We note that at most one of the terms in $\dot{R}_{\bar{Q}_n,g_n}$ will be nonzero. Moreover, when both nuisance parameters are estimated consistently, $\dot{R}_{\bar{Q}_n,g_n} = 0$ and $\ddot{R}_{\bar{Q}_n,g_n} = R_{\bar{Q}_n,g_n}$. In close analogy to the second part of Proposition 3.1, we have that $\|\bar{Q}_n - \bar{Q}_*\|^2_{L_2(P_0)} = \|g_n - g_*\|^2_{L_2(P_0)} = o_p(n^{-1/2})$ implies that the second-order remainder $\ddot{R}_{\bar{Q}_n,g_n}$ also converges to zero in probability at $n^{1/2}$-rate. Thus, we have a firm grasp on both the first- and second-order remainder terms.

3.3.2 Approximations based on Gâteaux derivatives

The direct approach above is applicable whenever a closed-form expression for the estimation remainder exists. However, many parameters are defined only as the population solution to optimization problems or systems of estimating equations. This leads naturally to the use of M-estimators and Z-estimators, respectively. In either paradigm, estimators of the implicitly-defined parameters, as well as the corresponding remainder term, are not typically available in closed form. This necessitates more general approaches to achieving the decomposition in (3.3). One possible strategy has been hinted at by Díaz and van der Laan (2017). There, the authors note that the approximations derived by van der Laan (2014) are equivalent to linearizing the so-called "drift" term that arises in standard asymptotic analysis of Z-estimators (as in Theorem 5.31 of van der Vaart (1998)).

Suppose that an unbiased estimating function $\phi(\psi; z)$ is available for the parameter ψ_0. We note that (un-normalized) influence functions often serve as estimating functions. Here, we consider the case that $\phi(\psi; z) = \phi_{Q,g}(\psi; z)$ is proportional to the doubly robust efficient influence function $\phi^*_{Q,g}(z)$. This implies that $\phi_{Q,g}(\psi; z)$ is an unbiased estimating

function provided either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. Thus, ψ_0 is the unique solution in ψ to $E_{P_0}\{\phi_{\bar{Q}_*,g_*}(\psi; Z)\} = 0$ if $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. Of course, $\bar{Q}_0$ and g_0 are estimated by $\bar{Q}_n$ and g_n in practice. Any solution ψ_n of the equation

$$\frac{1}{n}\sum_{i=1}^{n} \phi_{\bar{Q}_n,g_n}(\psi; Z_i) = o_p(n^{-1/2})$$

may be taken as an estimator of ψ_0. Under the empirical process and rate conditions of the previous section, a Taylor expansion results in the first-order approximation

$$\psi_n - \psi_0 = -\mathbf{V}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} \phi_{\bar{Q}_*,g_*}(\psi_0; Z_i) + \Phi(\bar{Q}_n, g_n)\right] + o_p(n^{-1/2} + |\Phi(\bar{Q}_n, g_n)|) \ ,$$

where Φ is the functional $(\bar{Q}, g) \mapsto \int \phi_{\bar{Q},g}(\psi_0; z) dP_0(z)$ and

$$\mathbf{V} := \frac{\partial}{\partial \psi} E_{P_0}\left\{\phi_{\bar{Q}_*,g_*}(\psi; Z)\right\}\Big|_{\psi=\psi_0}$$

denotes the (non-singular) matrix of partial derivatives. The drift term referenced by Díaz and van der Laan (2017) and van der Vaart (1998) is given by the value $\Phi(\bar{Q}_n, g_n)$ above.

If Φ is sufficiently smooth, and the nuisance parameters are contained in a convex set, we may approximate the drift term in first-order by

$$\Phi(\bar{Q}_n, g_n) - \Phi(\bar{Q}_*, g_*) = \dot{\Phi}_{\bar{Q}}(\bar{Q}_*, g_*; \bar{Q}_n - \bar{Q}_*) + \dot{\Phi}_g(\bar{Q}_*, g_*; g_n - g_*) + \ddot{R}_{\bar{Q}_n,g_n} \ ,$$

where the first two terms on the right-hand side denote the partial Gâteaux derivatives of Φ at $(\bar{Q}_*, g_*)$ in the directions $\bar{Q}_n - \bar{Q}_*$ and $g_n - g_*$, respectively, and $\ddot{R}_{\bar{Q}_n,g_n}$ is a second-order remainder term. In this case, $\dot{R}_{\bar{Q}_n,g_n} := \dot{\Phi}_{\bar{Q}}(\bar{Q}_*, g_*; \bar{Q}_n - \bar{Q}_*) + \dot{\Phi}_g(\bar{Q}_*, g_*; g_n - g_*)$ denotes the first-order contribution to the remainder. When $\bar{Q}_* = \bar{Q}_0$ and $g_* = g_0$, we typically expect a Neyman orthogonality property (van der Laan and Robins 2003, Chernozhukov et al. 2018) for $\phi_{\bar{Q}_0,g_0}$ such that

$$\dot{\Phi}_{\bar{Q}}(\bar{Q}_0, g_0; \bar{Q}_n - \bar{Q}_0) = \dot{\Phi}_g(\bar{Q}_0, g_0; g_n - g_0) = 0.$$

When just one of the nuisance estimators is consistent, this orthogonality will tend to fail for directions involving the consistent nuisance estimator. In such case, we expect that the

consistent nuisance estimator will make a first-order contribution to the estimation remainder.

It is often possible to evaluate the required Gâteaux derivatives using standard derivatives. If $\phi_{\bar{Q},g}$ depends on $\bar{Q}$ and g only through their values $\bar{Q}(z)$ and $g(z)$, and the regularity conditions of Newey (1994) hold, then

$$\dot{\Phi}_{\bar{Q}}(\bar{Q}_*, g_*; \bar{Q}_n - \bar{Q}_*) = \int \frac{\partial}{\partial \bar{Q}(z)} \phi_{\bar{Q},g_*}(\psi_0; z)\Big|_{Q(z)=\bar{Q}_*(z)} \{\bar{Q}_n(z) - \bar{Q}_*(z)\} dP_0(z) \; ; \tag{3.6}$$

$$\dot{\Phi}_{g}(\bar{Q}_*, g_*; g_n - g_*) = \int \frac{\partial}{\partial g(z)} \phi_{\bar{Q}_*,g}(\psi_0; z)\Big|_{g(z)=g_*(z)} \{g_n(z) - g_*(z)\} dP_0(z) \; . \tag{3.7}$$

Here, we see that Neyman orthogonality of $\phi_{\bar{Q}_*,g_*}$ is equivalent to orthogonality of the paths defined by $(\bar{Q}_n - \bar{Q}_*)$ and $(g_n - g_*)$ to the respective partial derivatives, where orthogonality is defined with respect to the covariance inner-product, $\langle f_1, f_2 \rangle_{P_0} := P_0(f_1 f_2)$.

To illustrate how this orthogonality fails under inconsistent estimation of only one nuisance parameter, we consider the class of mixed-bias parameters studied by Rotnitzky et al. (2019). For this class of parameters, the influence function is given by

$$\phi_{a_0,b_0}(Z) := s_{ab}(Z) a_0(Z) b_0(Z) + m_1(Z, a_0) + m_2(Z, b_0) + s_0(Z) - \Psi(P_0)$$

where s_{ab} and s_0 are known maps, and m_1 and m_2 are known up the nuisance parameters a_0 and b_0. We let a_n and b_n be estimators of a_0 and b_0 that are consistent for limits a_* and b_*, respectively. Then, under the conditions of their Theorem 1,

$$\dot{\Phi}_a(a_0, b_*; a_n - a_0) = P_0\{s_{ab}(a_n - a_0)(b_* - b_0)\},$$

and, similarly

$$\dot{\Phi}_b(a_*, b_0; b_n - b_0) = P_0\{s_{ab}(a_* - a_0)(b_n - b_0)\},$$

with $\dot{\Phi}_a(a_0, b_*; a_n - a_0) = 0$ if $b_* = b_0$ and $\dot{\Phi}_b(a_*, b_0; b_n - b_0) = 0$ if $a_* = a_0$. This finding is in agreement with the direct calculations at the end of the previous section.

3.3.3 Approximations requiring only pathwise differentiability

We now consider an alternative approach to decomposing the remainder into its first-order and higher-order components. This method will utilize only the definition of pathwise differentiability of the target parameter, requiring no further functional derivatives. We first define the submodels $\mathcal{M}_{\bar{Q}_0} = \{P \in \mathcal{M} : \bar{Q}(P) = \bar{Q}_0\}$ and $\mathcal{M}_{g_0} = \{P \in \mathcal{M} : g(P) = g_0\}$. In particular, P_0 is an element of the intersection model $\mathcal{M}_{\bar{Q}_0,g_0} = \mathcal{M}_{\bar{Q}_0} \cap \mathcal{M}_{g_0}$ and an element $P_* \in (\mathcal{M}_{\bar{Q}_0} \cup \mathcal{M}_{g_0}) \setminus \mathcal{M}_{\bar{Q}_0,g_0}$ has either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$, but not both. We continue to assume that Ψ is pathwise differentiable as a function of the nuisance parameter $\bar{Q}(P)$ with doubly robust efficient influence function $\phi^*_{\bar{Q},g}$.

Let $P_* \in (\mathcal{M}_{\bar{Q}_0} \cup \mathcal{M}_{g_0}) \setminus \mathcal{M}_{\bar{Q}_0,g_0}$ be arbitrary. Denote the nuisance parameter values at P_* by $\bar{Q}(P_*) = \bar{Q}_*$ and $g(P_*) = g_*$. We consider a smooth submodel

$$\{P_t : t\} \subset (\mathcal{M}_{\bar{Q}_0} \cup \mathcal{M}_{g_0}) \setminus \mathcal{M}_{\bar{Q}_0,g_0}$$

with nuisance values $\bar{Q}(P_t) = \bar{Q}_t$ and $g(P_t) = g_t$, such that $\bar{Q}_t \to \bar{Q}_*$ and $g_t \to g_*$ as $t \to 0$ (in P_0-norm). By pathwise differentiability of Ψ, we have $\Psi(P_t) - \Psi(P_0) = -P_0\phi^*_{\bar{Q}_t,g_t} + R_{\bar{Q}_t,g_t}$ and $\Psi(P_t) - \Psi(P_*) = -P_*\phi^*_{\bar{Q}_t,g_t} + \ddot{R}_{\bar{Q}_t,g_t}$. Under these conditions, $R_{\bar{Q}_t,g_t} := \Psi(P_t) - \Psi(P_0) + P_0\phi^*_{\bar{Q}_t,g_t}$ is not strictly a second-order remainder term since only one of the nuisance parameters converges to its corresponding value at P_0. However, $\ddot{R}_{\bar{Q}_t,g_t} := \Psi(P_t) - \Psi(P_*) + P_*\phi^*_{\bar{Q}_t,g_t}$ is second-order by construction since both nuisance parameters converge by definition to their value at P_*. This observation is a simple consequence of pathwise differentiability, but provides additional insight into the first-order remainder. We encapsulate this insight in the following result.

Theorem 3.1. *Let P_*, P_0, $\bar{Q}_t$, $\bar{Q}_*$, g_t, g_*, $R_{\bar{Q}_t,g_t}$ and $\ddot{R}_{\bar{Q}_t,g_t}$ be defined as above. Then,*

$$R_{\bar{Q}_t,g_t} = \dot{R}_{\bar{Q}_t,g_t} + \ddot{R}_{\bar{Q}_t,g_t} \quad \textit{where} \quad \dot{R}_{\bar{Q}_t,g_t} := (P_0 - P_*)(\phi^*_{\bar{Q}_t,g_t} - \phi^*_{\bar{Q}_*,g_*}) \ .$$

The proof is remarkably simple, relying only on the definition of pathwise differentiability and the fact that the efficient influence function of the map Ψ is doubly robust. In spite of

this simplicity, this result demonstrates the remarkable fact that the first-order contributions to the remainder can *always* be represented as a difference-in-differences. This representation for the first-order remainder can be related immediately to Neyman orthogonality of $\phi^*_{\bar{Q},g}$ by noting that if $P_0 \ll P_*$, then

$$(P_0 - P_*)\phi^*_{\bar{Q}_t,g_t} = -\int \phi^*_{\bar{Q}_t,g_t}(z)\left(\frac{dP_*}{dP_0}(z) - 1\right) dP_0(z) \ .$$

As noted in the proof of Lemma 1.8 of van der Laan and Robins (2003), we can interpret the function $dP_*/dP_0 - 1$ as a nuisance score with respect to estimation of ψ_0. Hence, we again conclude that $\dot{R}_{\bar{Q}_t,g_t} = 0$ provided that $\phi^*_{\bar{Q},g}$ is suitably orthogonalized to particular paths for the nuisance parameters.

In fact, knowledge of which nuisance parameter is consistently estimated (while not usually known in practice, outside of controlled experiments) tells us more about the nature of the first-order contribution. Invoking Lemma 1.8 of van der Laan and Robins (2003), we can anticipate that $\phi^*_{\bar{Q}_0,g_*}$ remains orthogonalized for paths involving g_* only. This implies that, provided $\bar{Q}_n$ is consistent for $\bar{Q}_0$, the inconsistent nuisance estimator g_n will not contribute to the asymptotic distribution of the estimator of ψ_0. However, no such theoretical guarantees exist for orthogonality of $\phi^*_{\bar{Q}_0,g_*}$ to paths involving $\bar{Q}_0$. Similar considerations apply for $\phi^*_{\bar{Q}_*,g_0}$. In summary, when just one nuisance parameter is consistently estimated, we expect that the first-order contribution $\dot{R}_{\bar{Q}_n,g_n}$ of the estimation remainder will correspond to the nuisance parameter that is consistently estimated. This is, of course, in agreement with what was demonstrated in the preceding subsections.

3.4 Achieving doubly robust inference under missingness at random

3.4.1 Remainder from doubly robust estimation when outcomes are missing at random

We are now prepared to describe a general framework for CANDoR estimation when outcomes are coarsened according to a missingness at random mechanism. Consider a full data unit of the form $Z^* = (W, X, Y)$ distributed according to F_0 belonging to some model $\mathcal{F}$, where W is a vector of covariates, $X \in \{0, 1\}$ is a binary missingness indicator, and Y

is an outcome of interest. Instead of directly observing Z^*, we suppose that we observe $Z = (W, X, XY)$ distributed according to some probability measure P_0. Hence, we only observe the value of Y when $X = 1$. In this section, the only restriction we place on the model for the distribution of Z is that the outcomes are missing at random, in the sense that Y and X are independent given W. In other words, given W, the conditional probability of observing Y does not depend on the value of Y. We denote this probability by $g_0(w) := P_0(X = 1 \mid Y = y, W = w) = P_0(X = 1 \mid W = w)$.

We may equivalently consider the binary point-treatment counterfactual framework of causal inference. For this setting, we suppose that a typical observation is $Z = (W, X, Y)$ where $W \in \mathbb{R}^d$ remains a vector of covariates, $X \in \{0, 1\}$ now represents a binary treatment variable, and Y denotes a univariate outcome of interest. Here, we rely on the consistency, randomization and positivity assumptions, which state that

- the observation unit is $(W, X, Y) = (W, X, Y(X))$, where $Y(0)$ and $Y(1)$ are potential outcomes;

- the potential outcomes $Y(0)$ and $Y(1)$ are independent of X conditional on W, and the propensity score $g_0(W) = P_0(X = 1 \mid W)$ depends only on W;

- the propensity score $g_0(W) \in \{0, 1\}$ almost surely, but is otherwise unrestricted.

An observation Z can then be thought of as a coarsened version of the full, unobservable, data unit $(W, X, Y(0), Y(1))$ having distribution F_0 in an otherwise unrestricted model $\mathcal{F}$. The results about missingness at random models below can then be interpreted as applying to functionals of the marginal distribution of $Y(1)$.

Let $\Gamma : \mathcal{F} \to \mathbb{R}^p$ be a pathwise differentiable parameter at each F in $\mathcal{F}$ with efficient influence function τ_F^*. If Γ is identifiable in the observed data model $\mathcal{M}$, then there exists a map $\Psi : \mathcal{M} \to \mathbb{R}^p$ such that the value $\Psi(P_0) = \Gamma(F_0)$. Then, it follows from results in van der Laan and Robins (2003) that in the missingness at random model, Ψ is a pathwise

differentiable mapping on the observed data model $\mathcal{M}$ with efficient influence function

$$\phi^*_{\bar{Q}_0,g_0}(z) = \frac{I(x=1)}{g_0(w)}\tau^*_{F_0}(z) + \left(1 - \frac{I(x=1)}{g_0(w)}\right)\bar{Q}_0(w) \tag{3.8}$$

where $\bar{Q}_0(w) := E_{P_0}\{\tau^*_F(z) \mid X = 1, W = w\}$ is a second nuisance parameter. Note that we can frequently re-parametrize $\bar{Q}_0$ depending on the particular form of $\tau^*_{F_0}$, as in the following example.

Example: Mean of an outcome missing at random If the target parameter is the mean outcome $\Gamma(F_0) = E_{F_0}(Y)$, then $\tau^*_{F_0}(z^*) = y - \Gamma(F_0)$. Since $E_{P_0}\{\tau^*_F(Z) \mid X = 1, W = w\} = E_{P_0}(Y \mid X = 1, W = w) - \Psi(P_0)$, it suffices to define the nuisance parameter $\bar{Q}_0$ as the observed outcome regression $w \mapsto E_{P_0}(Y \mid X = 1, W = w)$. We can then write the efficient influence function in the observed data model as

$$\phi^*_{\bar{Q}_0,g_0}(z) = \phi^*_{\bar{Q}_0,g_0}(z) = \frac{I(x=1)}{g_0(w)}(y - \Psi(P_0)) + \left(1 - \frac{I(x=1)}{g_0(w)}\right)\{\bar{Q}_0(w) - \Psi(P_0)\} \ .$$

While this re-parametrization for $\bar{Q}_0(w)$ is convenient in specific cases, we will state our primary results in terms of the former – more general – definition.

We also note that $\tau^*_{F_0}$ may itself involve additional nuisance parameters of F_0 beyond g_0 and $\bar{Q}_0$. To simplify the writing, we will continue to suppress this in our notation (unless otherwise stated) and assume that any such nuisance parameters are either known or can be estimated at an appropriate rate. For instance, it is often the case $\phi^*_{\bar{Q}_0,g_0}$ also depends on P_0 through the marginal distribution of W. However, estimating $P_0(W \leq w)$ consistently at $n^{1/2}$-rate is trivial, as the nonparametric maximum likelihood estimator for this nuisance parameter is just the empirical cumulative distribution function.

It is well-known that $\phi^*_{\bar{Q}_0,g_0}$ is a doubly robust influence function. This follows by a

standard conditioning argument. Assuming that either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$, we have

$$\begin{aligned} E_{P_0}\{\phi^*_{\bar{Q}_*,g_*}(Z)\} &= E_{P_0}\{\phi^*_{\bar{Q}_*,g_*}(Z) - \phi^*_{\bar{Q}_0,g_0}(Z)\} \\ &= E_{P_0}\left\{\frac{g_0(W)}{g_*(W)}\bar{Q}_0(W) - \bar{Q}_0(W) + \left[1 - \frac{g_0(W)}{g_*(W)}\right]\bar{Q}_*(W)\right\} \\ &= P_0\left\{\frac{(g_0 - g_*)\bar{Q}_0}{g_*} + \frac{(g_* - g_0)\bar{Q}_*}{g_*}\right\} \\ &= P_0\left\{\frac{(g_0 - g_*)(\bar{Q}_0 - \bar{Q}_*)}{g_*}\right\} \\ &= 0 \ . \end{aligned}$$

Hence, we may apply Theorem 3.1. For estimators $\bar{Q}_n$ and g_n, similar manipulations as in the above verification of doubly robustness lead to

$$(P_0 - P_*)\phi^*_{\bar{Q}_n,g_n} = P_0\left\{\frac{(g_0 - g_n)(\bar{Q}_0 - \bar{Q}_n)}{g_n}\right\} - P_*\left\{\frac{(g_* - g_n)(\bar{Q}_* - \bar{Q}_n)}{g_n}\right\}$$

for any $P_* \in (\mathcal{M}_{\bar{Q}_0} \cup \mathcal{M}_{g_0}) \setminus \mathcal{M}_{\bar{Q}_0,g_0}$. Since $\bar{Q}_0$ and g_0 are variationally independent of the marginal distribution of W, we may take $P_*(W \leq w) = P_0(W \leq w)$. Then, we have

$$\begin{aligned} (P_0 - P_*)\phi^*_{\bar{Q}_n,g_n} &= P_0\left\{\frac{(g_0 - g_n)(\bar{Q}_0 - \bar{Q}_n)}{g_n} - \frac{(g_* - g_n)(\bar{Q}_* - \bar{Q}_n)}{g_n}\right\} \\ &= P_0\left\{\frac{(g_0 - g_*)(\bar{Q}_0 - \bar{Q}_n)}{g_n}\right\} + P_0\left\{\frac{(g_0 - g_n)(\bar{Q}_0 - \bar{Q}_*)}{g_n}\right\} \ . \end{aligned}$$

This provides us with a representation of the first-order remainder contributions for estimation of ψ_0 by standard one-step or targeted maximum likelihood approaches. Had we instead taken the Gâteaux derivative of Φ in the direction defined by $\bar{Q}_n - \bar{Q}_*$ and $g_n - g_*$, we would have identified the first-order term $\dot{R}_{\bar{Q}_n,g_n} = P_0\left\{(g_0 - g_*)(\bar{Q}_0 - \bar{Q}_n)/g_*\right\} + P_0\left\{(g_0 - g_n)(\bar{Q}_0 - \bar{Q}_*)/g_*\right\}$. These forms for the first-order estimation remainder are readily seen to be equal up to a second-order term, so we may use either expression freely.

3.4.2 Consistent asymptotically normal doubly robust estimators

Now taking $\dot{R}_{\bar{Q}_n,g_n} := P_0\left\{(g_0 - g_*)(\bar{Q}_0 - \bar{Q}_n)/g_*\right\} + P_0\left\{(g_0 - g_n)(\bar{Q}_0 - \bar{Q}_*)/g_*\right\}$, our interest lies in linearizing this first-order term. We will see that this amounts to re-orthogonalization

of $\phi^*_{\bar{Q}_*,g_*}$ with respect to fluctuations involving the consistently estimated nuisance parameter. We define $\dot{R}_{\bar{Q}_n,g_*} := P_0\left\{(g_0 - g_*)(\bar{Q}_0 - \bar{Q}_n)/g_*\right\}$ and $\dot{R}_{\bar{Q}_*,g_n} := P_0\left\{(g_0 - g_n)(\bar{Q}_0 - \bar{Q}_*)/g_*\right\}$, so that $\dot{R}_{\bar{Q}_n,g_n} = \dot{R}_{\bar{Q}_n,g_*} + \dot{R}_{\bar{Q}_*,g_n}$. Linearizing these two terms separately is justified since $\dot{R}_{\bar{Q}_n,g_*} = 0$ when $g_* = g_0$ and $\dot{R}_{\bar{Q}_*,g_n} = 0$ when $\bar{Q}_* = \bar{Q}_0$.

We introduce the following nuisance parameters, which are immediately seen to generalize the three univariate regressions of Benkeser et al. (2017):

$$\mu_{\bar{Q}_*}(g_0(w)) = E_{P_0}\{\tau^*_{F_0} - \bar{Q}_*(W) \mid X = 1, g_0(W) = g_0(w)\} \tag{3.9}$$

$$\pi_1(\bar{Q}_0(w)) = P_0\{X = 1 \mid \bar{Q}_0(W) = \bar{Q}_0(w)\} \tag{3.10}$$

$$\pi_{2,g_*}(\bar{Q}_0(w)) = E_{P_0}\left\{\frac{X - g_*(W)}{g_*(W)} \,\middle|\, \bar{Q}_0(W) = \bar{Q}_0(w)\right\} . \tag{3.11}$$

We also define functions

$$\delta^{\bar{Q}_0}_{\bar{Q},g,\pi_1,\pi_2}(Z) := -\frac{\pi_{2,g}(\bar{Q}(W))}{\pi_1(\bar{Q}(W))} X\{\tau^*_{F_0}(Z) - \bar{Q}(W)\} \tag{3.12}$$

$$\delta^{g_0}_{\bar{Q},g,\mu}(Z) := -\frac{\mu_{\bar{Q}}(g(W))}{g(W)}\{X - g(W)\} . \tag{3.13}$$

Let μ_n, π_{1n}, and π_{2n} be estimators of the regressions $\mu_{\bar{Q}_*}$, π_1, and π_{2,g_*} above, based on the estimators $\bar{Q}_n$ and g_n. From the following lemma, we find that $\delta^{\bar{Q}_0}_{\bar{Q},g,\pi_1,\pi_2}$ and $\delta^{g_0}_{\bar{Q},g,\mu}$ play the role of influence functions in the linear expansion of $\dot{R}_{\bar{Q}_n,g_*}$ and $\dot{R}_{\bar{Q}_*,g_n}$ respectively.

Lemma 3.1. *If conditions (A1)-(A4) in the Appendix hold, then*

$$\dot{R}_{\bar{Q}_n,g_*} = \dot{R}_{\bar{Q}_n,g_*} - \dot{R}_{\bar{Q}_0,g_*} = -\mathbb{P}_n\delta^{\bar{Q}_0}_{\bar{Q}_n,g_n,\pi_{1n},\pi_{2n}} + \mathbb{P}_n\delta^{\bar{Q}_0}_{\bar{Q}_0,g_*,\pi_1,\pi_{2,g_*}} + o_p(n^{-1/2}) \tag{3.14}$$

$$\dot{R}_{\bar{Q}_*,g_n} = \dot{R}_{\bar{Q}_*,g_n} - \dot{R}_{\bar{Q}_*,g_0} = -\mathbb{P}_n\delta^{g_0}_{\bar{Q}_n,g_n,\mu_n} + \mathbb{P}_n\delta^{g_0}_{\bar{Q}_*,g_0,\mu_{\bar{Q}_*}} + o_p(n^{-1/2}) . \tag{3.15}$$

In tandem with the asymptotic expansion for $\psi_n - \psi_0$ studied in (3.1), Lemma 3.1 establishes the following result.

Theorem 3.2. *Suppose that either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. If the nuisance estimators $\bar{Q}^*_n, g^*_n, \pi_{1n}, \pi_{2n}$ and μ_n satisfy*

$$|\mathbb{P}_n\phi^*_{\bar{Q}^*_n,g^*_n}| = |\mathbb{P}_n\delta^{\bar{Q}_0}_{\bar{Q}^*_n,g^*_n,\pi_{1n},\pi_{2n}}| = |\mathbb{P}_n\delta^{g_0}_{\bar{Q}^*_n,g^*_n,\mu_n}| = o_p(n^{-1/2}) \tag{3.16}$$

where $|v|$ denotes the Euclidean norm, and if the empirical process and second-order remainder conditions (A1)-(A6) in the Appendix are met, then the plug-in estimator $\psi_{n,\text{CANDR}} := \Psi(\bar{Q}_n^*)$ *is asymptotically linear at* ψ_0 *with influence function* $\tilde{\phi}_{\bar{Q}_*,g_*} := \phi^*_{\bar{Q}_*,g_*} + \delta^{\bar{Q}_0}_{\bar{Q}_0,g_*,\pi_1,\pi_{2,g_*}} + \delta^{g_0}_{\bar{Q}_*,g_0,\mu_{\bar{Q}_*}}$. *Hence,* $\psi_{n,\text{CANDR}}$ *is consistent and asymptotically normal with*

$$n^{1/2}(\psi_{n,\text{CANDR}} - \psi_0) \rightsquigarrow \mathcal{N}(0, \boldsymbol{\Sigma}_*) \;\; \textit{for} \;\; \boldsymbol{\Sigma}_* := P_0(\tilde{\phi}_{\bar{Q}_*,g_*}\tilde{\phi}^\top_{\bar{Q}_*,g_*}) \; .$$

This theorem generalizes the constructions of van der Laan (2014) and Benkeser et al. (2017) to allow for consistent asymptotically normal doubly robust estimators of any pathwise differentiable parameter in the missingness at random setting. The proof of the result is left to the Appendix. We note that the statement of Theorem 3.2 is well-suited to construction of plug-in estimators based on the targeted minimum loss-based framework. We detail an algorithm for updating initial nuisance parameter estimators such that they satisfy the "small-bias" condition (3.16) in the Theorem. First, we suppose that $0 \leq \tau^*_{F_0} \leq 1$. Then, we implement the following iterative scheme to construct targeted nuisance estimators.

Step 1. Set $k = 0$. Let $\bar{Q}_n^0$ and g_n^0 denote initial estimates of $\bar{Q}_0$ and g_0.

Step 2. Fit a logistic regression on the subset of data having $X = 1$ with outcome $\tau^*_{F_0}(Z)$, no intercept, a single covariate $X/g_n^k(W)$ and offset logit $\bar{Q}_n^k(W)$. Define ϵ_{1n}^k to be the estimated coefficient and set $\bar{Q}_n^{k+0.5}(W) := \text{expit}\{\text{logit } \bar{Q}_n^k(W) + \epsilon_{1n}^k X/g_n^k(W)\}$.

Step 3. Let π_{1n}^k and π_{2n}^k be estimates of $\pi_1(\bar{Q}_0(w_i))$ and $\pi_{2,g_*}(\bar{Q}_0(w_i))$ based on g_n^k and $\bar{Q}_n^{k+0.5}$.

Step 4. Fit a logistic regression on the subset of data having $X = 1$ with outcome $\tau^*_{F_0}(Z)$, no intercept, a single covariate $X\pi_{2n}^k(W)/\pi_{1n}^k(W)$ and offset logit $\bar{Q}_n^{k+0.5}(W)$. Define ϵ_{2n}^k to be the estimated coefficient and set $\bar{Q}_n^{k+1}(W) := \text{expit}\{\text{logit } \bar{Q}_n^{k+0.5}(W) + \epsilon_{2n}^k X\pi_{2n}^k(W)/\pi_{1n}^k(W)\}$.

Step 5. Let μ_n^k be an estimate of $\mu_{\bar{Q}_*}(g_0(w))$ based on g_n^k and $\bar{Q}_n^{k+1}$.

Step 6. Fit a logistic regression with outcome X, no intercept, a single covariate $\mu_n^k(W)/g_n^k(W)$ and offset logit $g_n^k(W)$. Define ϵ_{3n}^k to be the estimated coefficient and set $g_n^{k+1}(W) := \text{expit}\{\text{logit } g_n^k(W) + \epsilon_{3n}^k \mu_n^k(W)/g_n^k(W)\}$.

Step 7. Set $k = k+1$ and repeat Steps 2-6 until $k = k^*$ such that (3.16) holds. Then, define the targeted nuisance estimators $\bar{Q}_n^* := \bar{Q}_n^{k^*}, g_n^* := g_n^{k^*}, \pi_{1n} := \pi_{1n}^{k^*}, \pi_{2n} := \pi_{2n}^{k^*}$ and $\mu_n := \mu_n^{k^*}$.

The targeted minimum loss-based estimator $\psi_{n,\text{CANDR}} := \Psi(\bar{Q}_n^*)$ is an asymptotically linear estimator of ψ_0 by Theorem 3.2. The asymptotic covariance matrix $\mathbf{\Sigma}_*$ can be consistently estimated by the plug-in estimator $\mathbf{\Sigma}_n := \mathbb{P}_n(\tilde{\phi}_n \tilde{\phi}_n^\top)$ for $\tilde{\phi}_n := \phi^*_{\bar{Q}_n^*, g_n^*} + \delta^{\bar{Q}_0}_{\bar{Q}_n^*, g_n^*, \pi_{1n}, \pi_{2n}} + \delta^{g_0}_{\bar{Q}_n^*, g_n^*, \mu_n}$.

The above requirement that $0 \leq \tau_{F_0}^* \leq 1$ is necessary in order to use logistic regression in the algorithm. However, this condition can be circumvented for bounded $\tau_{F_0}^*$ by linearly mapping the space of possible values for $\tau_{F_0}^*$ into the interval $[0,1]$, as described in Chapter 7 of van der Laan and Rose (2011). After the above algorithm converges, the transformed values are mapped back to their original scale.

We note that, here, one-step estimators are no longer a suitable alternative to targeted minimum loss-based estimators. In Benkeser et al. (2017) it is argued that one-step consistent asymptotically normal doubly robust estimators are only justified theoretically in the (typically) unrealistic setting where it is known which nuisance parameter is being estimated consistently. If (3.16) does not hold for the nuisance estimators, the one-step estimator $\Psi(\bar{Q}_n^*) + \mathbb{P}_n \tilde{\phi}_n$ contains an extra term that is not generally $o_p(n^{-1/2})$. When (3.16) holds, the above one-step construction is asymptotically equivalent to the targeted minimum loss-based estimator (and will typically rely on a similar algorithm to that above). Hence, to obtain valid inference for ψ_0, we generally suggest using targeted minimum loss-based estimation.

3.4.3 Examples

Example: Quantiles of an outcome missing at random. It is natural to define the quantiles in terms of the underlying distribution function. First, we define $H_{P_0}(t, w) :=$

$P_0(Y \leq t \mid X = 1, W = w)$ to be the conditional distribution function of Y. Hence, the marginal distribution function is $F_0(t) := E_{P_0}\{H_{P_0}(t, W)\}$. The q-th quantile is then

$$\Psi_{F_0}(q) := \inf\{y : q \leq F_0(y)\}.$$

The efficient influence function for estimating $\Psi_{F_0}(q)$ in the missing at random model is

$$\phi^*_{P_0}(z; t) := \frac{1}{f_0(\Psi_{F_0}(q))}\left[\frac{x}{g_0(w)}\{I(y \leq t) - H_{P_0}(t, w)\} + H_{P_0}(t, w) - F_0(t)\right],$$

where f_0 is the density function of F_0.

In the following examples, we consider the binary point-treatment counterfactual framework reviewed above using our results to develop consistent asymptotically normal doubly robust estimators for treatment effect parameters. We encourage readers interested specifically in consistent asymptotically normal doubly robust estimators of the average treatment effect (ATE), $E_{F_0}(Y^1 - Y^0)$, to read the motivating works by van der Laan (2014) and Benkeser et al. (2017). Below, unless otherwise stated, we take $\bar{Q}_0(x, x) := E_{P_0}(Y \mid X = x, W = w)$, while $g_0(w)$ continues to denote $P_0(X = 1 \mid W = w)$.

Example: Average treatment effect on the treated (ATT). The ATT defined by

$$\Gamma(F_0) := E_{F_0}(Y(1) - Y((0) \mid X = 1)$$

is often of interest in observational studies. Under the causal assumptions discussed above, the ATT is identified from the observed data distribution by the functional

$$\mathrm{ATT}(P_0) := \arg\min_{\psi} E_{P_0}[\{Y - \bar{Q}_0(0, W) - X\psi\}^2] = \frac{E_{P_0}[X\{Y - \bar{Q}_0(0, W)\}]}{E_{P_0}\{X\}}\ .$$

The ATT is pathwise differentiable with (nonparametric) efficient influence function

$$\phi^*_{\bar{Q}_0, g_0}(z) := \frac{g_0(w)}{E_{P_0}\{X\}}\left[\left\{\frac{x}{g_0(w)} - \frac{1 - x}{1 - g_0(w)}\right\}\{y - \bar{Q}_0(0, w)\} - \frac{x}{g_0(w)}\mathrm{ATT}(P_0)\right]\ .$$

It is easily verified that $\phi^*_{\bar{Q}_0, g_0}$ is a doubly robust influence function and that plug-in estimators can be based on estimates of $\bar{Q}_0(0, \cdot)$ and g_0 only. See, e.g., Hahn (1998) or Hirano et al. (2003) for a detailed treatment of these properties.

The first-order estimation remainder for the ATT is readily seen to be

$$\dot{R}_{\bar{Q}_n,g_n} = P_0\left[\{\bar{Q}_n(0,\cdot) - \bar{Q}_0(0,\cdot)\}\left(\frac{g_* - g_0}{1-g_*}\right)\right] + P_0\left[\{\bar{Q}_0(0,\cdot) - \bar{Q}_*(0,\cdot)\}\left(\frac{g_0 - g_n}{1-g_0}\right)\right] ,$$

which is equivalent to the expression found by Benkeser et al. (2017) for the estimation remainder corresponding to the mean of $Y(0)$. This immediately implies that any nuisance estimators $\bar{Q}_n^*, g_n^*, \pi_{1n}, \pi_{2n}$ and μ_n that satisfy the conditions of Theorem 3.2 for $E_{F_0}\{Y(0)\}$ will also ensure that the plug-in estimator

$$\mathrm{ATT}(\bar{Q}_n^*, g_n^*) := \frac{\sum_{i=1}^n X_i\{Y_i - \bar{Q}_n^*(0, W_i)\}}{\sum_{i=1}^n X_i}$$

is a consistent asymptotically normal doubly robust estimator of the ATT. The influence function of $\mathrm{ATT}(\bar{Q}_n^*, g_n^*)$ is

$$\phi^*_{\bar{Q}_0,g_0}(z) + \delta^{\bar{Q}_0}_{\bar{Q}_0,g_*,\pi_1,\pi_{2,g_*}}(z) + \delta^{g_0}_{\bar{Q}_*,g_0,\mu_{\bar{Q}_*}}(z) ,$$

where

$$\delta^{\bar{Q}_0}_{\bar{Q}_0,g_*,\pi_1,\pi_{2,g_*}}(z) = -\frac{\pi_{2,g_*}(\bar{Q}_0(0,w))}{\pi_1(\bar{Q}_0(0,w))}(1-x)\{y - \bar{Q}_0(0,w)\}$$

and

$$\delta^{g_0}_{\bar{Q}_*,g_0,\mu_{\bar{Q}_*}}(z) = -\frac{\mu_{\bar{Q}_*}(1-g_0(w))}{1-g_0(w)}\{g_0(w) - x\} .$$

Example: V-smoothed average treatment effects (V-ATE). It is often desired to evaluate the ATE among relevant sub-populations defined by a subset of the covariate vector, say V. Conditioning on V, the V-adjusted ATE is

$$\mathrm{ATE}(v; P_0) := E_{P_0}\{\bar{Q}_0(1,W) - \bar{Q}_0(0,W) \mid V = v\} .$$

For a general V (possibly not strictly discrete), $\mathrm{ATE}(v; P_0)$ is not pathwise differentiable in a nonparametric model. As an alternative to making strong assumptions on the form of $\mathrm{ATE}(v; P_0)$ (e.g., that it belongs to a sufficiently low-dimensional parametric model), we consider the class of projection parameters defined by van der Laan (2006). In particular, let $m_\beta(v)$ denote a working parametric model indexed by $\beta \in \mathbb{R}^{\dim(V)}$. We define

$$\beta(P_0) := \arg\min_{\beta \in \mathbb{R}^{\dim(V)}} E_{P_0}\{\mathrm{ATE}(V; P_0) - m_\beta(V)\}^2 .$$

Hence, we can interpret $\beta_0 := \beta(P_0)$ as the least-squares projection of $\text{ATE}(v; P_0)$ onto the working model $m_\beta(V)$ and $m_{\beta_0}(V)$ can be interpreted as a V-smoothed average treatment effect. In van der Laan (2006), it was shown that $P \mapsto \beta(P)$ is pathwise differentiable at P_0 with efficient influence function

$$\phi^*_{\bar{Q}_0,g_0}(z) := -c^{-1}(P_0)\ \dot{m}_{\beta_0}(v)\{\xi_{\bar{Q}_0,g_0}(z) - m_{\beta_0}(v)\}\ ,$$

where

$$\xi_{\bar{Q}_0,g_0}(z) := \bar{Q}_0(1,w) - \bar{Q}_0(0,w) + \frac{x}{g_0(w)}\{y - \bar{Q}_0(1,w)\} - \frac{1-x}{1-g_0(w)}\{y - \bar{Q}_0(0,w)\},$$

and

$$c(P_0) := P_0\{\ddot{m}_{\beta_0}(\xi_{\bar{Q}_0,g_0} - m_{\beta_0})\} - P_0(\dot{m}_{\beta_0}\dot{m}_{\beta_0}^\top)\ ,$$

with $\dot{m}_{\beta_0}$ and $\ddot{m}_{\beta_0}$ denoting, respectively, the gradient and Hessian of m_β evaluated at β_0. The random variable $\xi_{\bar{Q}_0,g_0}(Z)$ can be interpreted as a doubly robust pseudo-outcome, as $P_0\xi_{\bar{Q}_*,g_*} = \text{ATE}(P_0)$, provided either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. Note that $\xi_{\bar{Q}_0,g_0}(z) - \text{ATE}(P_0)$ is in fact the efficient influence function of the ATE. Indeed, if a constant is used as the working model, $m_\beta \equiv \beta \in \mathbb{R}^1$, then we find $\beta_0 = \text{ATE}(P_0)$ as a special (and, perhaps, trivial) case of V-smoothed average treatment effect.

It follows from Theorem 3.1 that the first-order remainder for estimating $\beta(P_0)$ is

$$\begin{aligned}\dot{R}_{\bar{Q}_n,g_n} = &- c^{-1}(P_0)P_0\left(\dot{m}_{\beta_0}\left[\{\bar{Q}_n(1) - \bar{Q}_0(1)\}\left(\frac{g_* - g_0}{g_*}\right) - \{\bar{Q}_n(0) - \bar{Q}_0(0)\}\left(\frac{g_0 - g_*}{1 - g_*}\right)\right]\right)\\ &- c^{-1}(P_0)P_0\left(\dot{m}_{\beta_0}\left[\{\bar{Q}_*(1) - \bar{Q}_0(1)\}\left(\frac{g_n - g_0}{g_0}\right) - \{\bar{Q}_*(0) - \bar{Q}_0(0)\}\left(\frac{g_0 - g_n}{1 - g_0}\right)\right]\right).\end{aligned}$$

Using this expression, we can prove the following corollary of Theorem 3.2.

Corollary 3.1. *Suppose that either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$. If the estimators $\bar{Q}^*_n, g^*_n, \pi_{1n}, \pi_{2n}$ and μ_n satisfy* (3.16) *for estimation of* $\text{ATE}(P_0)$*, and m_β is such that $|c(P_0)^{-1}P_0\dot{m}_{\beta_0}| < \infty$, then $\beta(\bar{Q}^*_n)$ is a CANDoR estimator of β_0.*

This result means that nuisance estimators that have been targeted for the ATE are also sufficiently targeted for the vector-valued parameter β_0. This drastically simplifies estimation and inference for V-smoothed treatment effects.

3.5 Discussion

We note that the Donsker conditions, such as A2, can be considerably weakened. For example, we may utilize various sample-splitting and cross-fitting techniques as discussed by Zheng and van der Laan (2011) and Chernozhukov et al. (2018).

The influence function arising in Theorem 3.2 merits additional comment. In particular, the influence function $\tilde{\phi}_{\bar{Q}_*,g_*}$ is now orthogonalized in the sense that (i) the Gâteaux derivative of $g \mapsto P_0\tilde{\phi}_{\bar{Q}_*,g}$ at g_0 in the direction $g_n - g_0$ is zero, and (ii) the Gâteaux derivative of $\bar{Q} \mapsto P_0\tilde{\phi}_{\bar{Q},g_*}$ at $\bar{Q}_0$ in the direction $\bar{Q}_n - \bar{Q}_0$ is zero. It is important to note that although $\tilde{\phi}_{\bar{Q}_*,g_*}$ exhibits a property similar to the Neyman orthogonality of $\phi^*_{\bar{Q}_0,g_0}$, it is not generally the case that $\psi_{n,\text{CANDR}}$ is a regular estimator. This follows immediately from the fact that any regular asymptotically linear estimator of ψ_0 in a nonparametric model must have influence function equal to the efficient influence function, $\phi^*_{\bar{Q}_0,g_0}$. By inspection, the influence function of $\psi_{n,\text{CANDR}}$ only equals the efficient influence function when both nuisance parameters are consistently estimated, so that $\bar{Q}_* = \bar{Q}_0$ and $g_* = g_0$. Further exploring the irregular nature of the estimators when one nuisance estimator is inconsistent remains an important direction for future work.

We note, as did Benkeser et al. (2017), that when $\bar{Q}_* = \bar{Q}_0$ and $g_* \neq g_0$, $\delta^{\bar{Q}_0}_{\bar{Q},g,\pi_1,\pi_2}$ is not the only possible influence function for which Lemma 3.1 holds. See, for example, the influence function based on the bivariate regression parameter of van der Laan (2014). That we can exhibit multiple influence functions for $\dot{R}_{\bar{Q}_n,g_*}$ in a non-parametric model implies that the asymptotic relative efficiency of the resulting estimators may vary.

While we detailed the construction of consistent asymptotically normal doubly robust estimators for general parameters when outcomes are missing at random, similar results can easily be derived for other forms of coarsening. However, our method of construction still relies on having an expression for the efficient influence function in order to study the remainder. It is often the case for more complex observed data structures that the efficient influence function is not available in closed form (van der Laan and Robins 2003).

While various methods, such as Carone et al. (in press), may be implemented to evaluate the influence function without a closed form, how to linearly approximate the first-order remainder in such settings remains an open question.

Throughout this chapter, we have assumed that the limits $\bar{Q}_*$ and g_* exist in the sense that the norms $\|\bar{Q}_n - \bar{Q}_*\|_{L_2(P_0)}$ and $\|g_n - g_*\|_{L_2(P_0)}$ converge to zero in probability sufficiently fast. In practice, this is an assumption that requires careful thought. If the second-order partial derivatives of $f : \mathbb{R}^p \mapsto \mathbb{R}$ exist, an optimistic rate for nonparametric estimation of f by f_n (e.g., a kernel regression estimator) is $\|f_n - f\|_{L_2(P_0)} = o_p(n^{-4/(4+p)})$ (Wasserman 2006). This suggests that for $p \geq 4$ covariates W, one or both of $\|\bar{Q}_n - \bar{Q}_*\|_{L_2(P_0)}$ and $\|g_n - g_*\|_{L_2(P_0)}$ may fail to be $o_p(n^{-1/2})$. Hence, care needs to be taken when using our methods due to the curse of dimensionality.

Even so, there has been a recent sustained interest in utilizing doubly robust estimators in high-dimensional data settings. Work by authors such as Dukes et al. (2018), Tan (2018), Ning et al. (2018), Smucler et al. (2019), and Bradic et al. (2019) study a variant of doubly robust inference that arises when there is a large number of covariates. In their work, the authors study several approaches to inference for parameters (such as the ATE) that are doubly robust to misspecification of the sparsity of the nuisance estimators. However, the specific estimators considered by these authors typically make strong assumptions (e.g., linearity) about the functional form of the nuisance parameters. It will therefore be of great interest to reconcile their approach to doubly robust inference with the targeted learning approach we have extended in the present article.

APPENDIX

3.A Proof of Theorem 3.1

As an immediate consequence of pathwise differentiability, $\Psi(P_t)-\Psi(P_0) = -P_0\phi^*_{\bar{Q}_t,g_t}+R_{\bar{Q}_t,g_t}$ and $\Psi(P_t) - \Psi(P_*) = -P_*\phi^*_{\bar{Q}_t,g_t} + \ddot{R}_{\bar{Q}_t,g_t}$. Combining these equalities, we have

$$R_{\bar{Q}_t,g_t} = \Psi(P_*) - \Psi(P_0) + (P_0 - P_*)\phi^*_{\bar{Q}_t,g_t} + \ddot{R}_{\bar{Q}_t,g_t} \ .$$

To conclude the proof we argue that since either $\bar{Q}_* = \bar{Q}_0$ or $g_* = g_0$, and $\phi^*_{\bar{Q},g}$ is a doubly robust influence function, it follows that $\Psi(P_*) - \Psi(P_0) = -P_0\phi^*_{\bar{Q}_*,g_*} = (P_* - P_0)\phi^*_{\bar{Q}_*,g_*}$.

3.B Proof of Lemma 3.1

We linearize $\dot{R}_{\bar{Q}_n,g_*}$ and $\dot{R}_{\bar{Q}_*,g_n}$ separately, as discussed in the main body of the chapter. Starting with $\dot{R}_{\bar{Q}_n,g_*}$, we suppose that $\bar{Q}_* = \bar{Q}_0$ while $g_* \neq g_0$ and define the bivariate regressions

$$\begin{aligned}
\pi'_1(\bar{Q}(W), \bar{Q}'(W)) &= E_{P_0}\{X \mid \bar{Q}(W), \bar{Q}'(W)\} \\
\pi'_{2,g_*}(\bar{Q}(W), \bar{Q}'(W)) &= E_{P_0}\left\{ \frac{X - g_*(W)}{g_*(W)} \,\middle|\, \bar{Q}(W), \bar{Q}'(w) \right\} \ ,
\end{aligned}$$

noting that the univariate regressions (3.10) and (3.11) are recovered by taking $\bar{Q} = \bar{Q}' = \bar{Q}_0$ above, respectively. Then, letting $x : Z \mapsto X$ denote the coordinate map for X, we can write this first-order remainder term as:

$$\begin{aligned}
\dot{R}_{\bar{Q}_n,g_*} &= P_0\left\{ \left(\frac{g_* - g_0}{g_*} \right) (\bar{Q}_0 - \bar{Q}_n) \right\} \\
&= -P_0\left\{ \left(\frac{x - g_*}{g_*} \right) (\bar{Q}_0 - \bar{Q}_n) \right\} \\
&= -P_0\left\{ \pi'_{2,g_*}(\bar{Q}_0, \bar{Q}_n)(\bar{Q}_0 - \bar{Q}_n) \right\} \\
&= -P_0\left\{ \frac{\pi'_{2,g_*}(\bar{Q}_0, \bar{Q}_n)}{\pi'_1(\bar{Q}_0, \bar{Q}_n)} x(\tau^*_{F_0} - \bar{Q}_n) \right\} \ .
\end{aligned}$$

Letting π_{1n} and π_{2n} be consistent estimators of (3.10) and (3.11), respectively, we define

$$M_n^{\bar{Q}_0} := -P_0 \left\{ \left(\frac{\pi'_{2,g_*}(\bar{Q}_0, \bar{Q}_n)}{\pi'_1(\bar{Q}_0, \bar{Q}_n)} - \frac{\pi_{2n}}{\pi_{1n}} \right) (\bar{Q}_0 - \bar{Q}_n) g_0 \right\} .$$

We now have $\dot{R}_{\bar{Q}_n, g_*} = -P_0 \delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}} + M_n^{\bar{Q}_0}$. Then, defining

$$G_n^{\bar{Q}_0} := (\mathbb{P}_n - P_0) \left(\delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}} - \delta^{\bar{Q}_0}_{\bar{Q}_0, g_*, \pi_1, \pi_{2,g_*}} \right),$$

we have that $\dot{R}_{\bar{Q}_*, g_n} = -\mathbb{P}_n \delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}} + \mathbb{P}_n \delta^{\bar{Q}_0}_{\bar{Q}_0, g_*, \pi_1, \pi_{2,g_*}} + G_n^{\bar{Q}_0} + M_n^{\bar{Q}_0}$. Now, we introduce regularity conditions to ensure the terms $G_n^{\bar{Q}_0}$ and $M_n^{\bar{Q}_0}$ are $o_p(n^{-1/2})$. We suppose that:

Condition A1. *The convergence rates of the estimators π_{1n}, π_{2n} and $\bar{Q}_n$ are such that*

$$\left\| \frac{\pi'_{2,g_*}(\bar{Q}_0, \bar{Q}_n)}{\pi'_1(\bar{Q}_0, \bar{Q}_n)} - \frac{\pi_{2n}}{\pi_{1n}} \right\|_{L_2(P_0)} \|\bar{Q}_n - \bar{Q}_0\|_{L_2(P_0)} = o_p(n^{-1/2}) .$$

Condition A2. *The functions $\delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}}$ and $\delta^{\bar{Q}_0}_{\bar{Q}_0, g_*, \pi_1, \pi_{2,g_*}}$ belong to a P_0-Donsker class and $\|\delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}} - \delta^{\bar{Q}_0}_{\bar{Q}_0, g_*, \pi_1, \pi_{2,g_*}}\|_{L_2(P_0)} = o_p(1)$.*

Then, applying the Cauchy-Schwartz inequality with A1, we find that $M_n^{\bar{Q}} = o_p(n^{-1/2})$. In conjunction with Lemma 19.24 of van der Vaart (1998), the assumption A2 ensures that $G_n^{\bar{Q}_0} = o_p(n^{-1/2})$. Then, we have the desired result:

$$\dot{R}_{\bar{Q}_n, g_*} = \dot{R}_{\bar{Q}_n, g_*} - \dot{R}_{\bar{Q}_0, g_*} = -\mathbb{P}_n \delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}} + \mathbb{P}_n \delta^{\bar{Q}_0}_{\bar{Q}_0, g_*, \pi_1, \pi_{2,g_*}} + o_p(n^{-1/2}) .$$

We note that the condition A2 can be considerably weakened, for example, by utilizing various sample-splitting and cross-fitting techniques as discussed by Zheng and van der Laan (2011) and Chernozhukov et al. (2018).

To linearize $\dot{R}_{\bar{Q}_*, g_n}$, we suppose that $\bar{Q}_* \neq \bar{Q}_0$ while $g_* = g_0$ and define the bivariate regression

$$\mu'_{\bar{Q}_*}(g(W), g'(W)) := E_{P_0}\{\tau^*_{F_0}(Z) - \bar{Q}_*(Z) \mid X = 1, g(W), g'(W)\} ,$$

noting that $\mu_{\bar{Q}_*}(g_0(W)) \equiv \mu'_{\bar{Q}_*}(g_0(W), g_0(W))$. Again letting $x : Z \mapsto X$ denote the coordinate map for X, we can write this first-order remainder term as:

$$\begin{aligned}
\dot{R}_{\bar{Q}_*, g_n} &= P_0 \left\{ \left(\frac{g_0 - g_n}{g_0} \right) (\bar{Q}_* - \bar{Q}_0) \right\} \\
&= -P_0 \left\{ \left(\frac{g_0 - g_n}{g_0} \right) \frac{x}{g_0} (\tau^*_{F_0} - \bar{Q}_*) \right\} \\
&= -P_0 \left\{ \left(\frac{g_0 - g_n}{g_0} \right) \frac{x}{g_0} \mu'_{\bar{Q}_*}(g_0, g_n) \right\} \\
&= -P_0 \left\{ \left(\frac{x - g_n}{g_0} \right) \mu'_{\bar{Q}_*}(g_0, g_n) \right\} \ .
\end{aligned}$$

Letting $\mu_n(W)$ be a consistent estimator of $\mu_{\bar{Q}_*}(g_0(W))$ we define the second-order term

$$M_n^{g_0} := -P_0[\{\mu'_{\bar{Q}_*}(g_0, g_n) - \mu_n\}(g_0 - g_n)/g_0] + P_0\{\mu_n(g_n - g_0)^2/(g_0 g_n)\} \ .$$

We now have $\dot{R}_{\bar{Q}_*, g_n} = -P_0 \delta^{g_0}_{\bar{Q}_n, g_n, \mu_n} + M_n^{g_0}$. Then, defining

$$G_n^{g_0} := (\mathbb{P}_n - P_0) \left(\delta^{g_0}_{\bar{Q}_n, g_n, \mu_n} - \delta^{g_0}_{\bar{Q}_*, g_0, \mu_{\bar{Q}_*}} \right),$$

we have that $\dot{R}_{\bar{Q}_*, g_n} = -\mathbb{P}_n \delta^{g_0}_{\bar{Q}_n, g_n, \mu_n} + \mathbb{P}_n \delta^{g_0}_{\bar{Q}_*, g_0, \mu_{\bar{Q}_*}} + G_n^{g_0} + M_n^{g_0}$. Now, we introduce regularity conditions to ensure the terms $G_n^{g_0}$ and $M_n^{g_0}$ are $o_p(n^{-1/2})$. We suppose that:

Condition A3. *The convergence rates of the estimators μ_n and g_n are such that* $\|\mu'_{\bar{Q}_*}(g_0, g_n) - \mu_n\|_{L_2(P_0)} \ \|g_n - g_0\|_{L_2(P_0)} = o_p(n^{-1/2})$ *and* $\|g_n - g_0\|_{L_2(P_0)} = o_p(n^{-1/4})$.

Condition A4. *The functions $\delta^{g_0}_{\bar{Q}_n, g_n, \mu_n}$ and $\delta^{g_0}_{\bar{Q}_*, g_0, \mu_{\bar{Q}_*}}$ belong to a P_0-Donsker class and* $\|\delta^{g_0}_{\bar{Q}_n, g_n, \mu_n} - \delta^{g_0}_{\bar{Q}_*, g_0, \mu_{\bar{Q}_*}}\|_{L_2(P_0)} = o_p(1)$.

Then, applying the Cauchy-Schwartz inequality with A3, we find that $M_n^{g_0} = o_p(n^{-1/2})$. In conjunction with Lemma 19.24 of van der Vaart (1998), the assumption A4 ensures that $G_n^{g_0} = o_p(n^{-1/2})$. Then, we have the desired result:

$$\dot{R}_{\bar{Q}_*, g_n} = \dot{R}_{\bar{Q}_*, g_n} - \dot{R}_{\bar{Q}_*, g_0} = -\mathbb{P}_n \delta^{g_0}_{\bar{Q}_n, g_n, \mu_n} + \mathbb{P}_n \delta^{g_0}_{\bar{Q}_*, g_0, \mu_{\bar{Q}_*}} + o_p(n^{-1/2}) \ .$$

As before, we note that the condition A4 can be considerably weakened. □

3.C Proof of Theorem 3.2

Defining $G_n^* := (\mathbb{P}_n - P_0)(\phi^*_{\bar{Q}_n^*, g_n^*} - \phi^*_{\bar{Q}_*, g_*})$, we have from (3.1) that

$$\psi_{n,\text{CANDR}} - \psi_0 = \mathbb{P}_n \phi^*_{\bar{Q}_*, g_*} - \mathbb{P}_n \phi^*_{\bar{Q}_n^*, g_n^*} + G_n^* + R_{\bar{Q}_n^*, g_n^*} \ .$$

By application of Theorem 3.1 and Lemma 3.1, we find

$$\begin{aligned}
\psi_{n,\text{CANDR}} - \psi_0 &= \mathbb{P}_n \phi^*_{\bar{Q}_*, g_*} - \mathbb{P}_n \phi^*_{\bar{Q}_n^*, g_n^*} + G_n^* + \dot{R}_{\bar{Q}_n^*, g_n^*} + \ddot{R}_{\bar{Q}_n^*, g_n^*} \\
&= \mathbb{P}_n \phi^*_{\bar{Q}_*, g_*} - \mathbb{P}_n \phi^*_{\bar{Q}_n^*, g_n^*} + G_n^* + \dot{R}_{\bar{Q}_n^*, g_*} + \dot{R}_{\bar{Q}_*, g_n^*} + \ddot{R}_{\bar{Q}_n^*, g_n^*} \\
&= \mathbb{P}_n \tilde{\phi}_{\bar{Q}_*, g_*} - \mathbb{P}_n \phi^*_{\bar{Q}_n^*, g_n^*} - \mathbb{P}_n \delta^{\bar{Q}_0}_{\bar{Q}_n, g_n, \pi_{1n}, \pi_{2n}} - \mathbb{P}_n \delta^{g_0}_{\bar{Q}_n, g_n, \mu_n} + G_n^* + \ddot{R}_{\bar{Q}_n^*, g_n^*} + o_p(n^{-1/2}) \\
&= \mathbb{P}_n \tilde{\phi}_{\bar{Q}_*, g_*} + G_n^* + \ddot{R}_{\bar{Q}_n^*, g_n^*} + o_p(n^{-1/2}) \ ,
\end{aligned}$$

where the last equality holds by assumption (3.16). We suppose that:

Condition A5. *The convergence rates of estimators $\bar{Q}_n^*$ and g_n^* satisfy*

$$\|\bar{Q}_n^* - \bar{Q}_*\|_{L_2(P_0)} \ \|g_n^* - g_*\|_{L_2(P_0)} = o_p(n^{-1/2}) \ .$$

Condition A6. *The functions $\phi^*_{\bar{Q}_n^*, g_n^*}$ and $\phi^*_{\bar{Q}_*, g_*}$ belong to a P_0-Donsker class and*

$$\|\phi^*_{\bar{Q}_n^*, g_n^*} - \phi^*_{\bar{Q}_*, g_*}\|_{L_2(P_0)} = o_p(1) \ .$$

Then, applying the Cauchy-Schwartz inequality with A5, we find that $\ddot{R}_{\bar{Q}_n^*, g_n^*} = o_p(n^{-1/2})$. In conjunction with Lemma 19.24 of van der Vaart (1998), condition A6 ensures that $G_n^* = o_p(n^{-1/2})$. Then, we have the desired result:

$$n^{1/2}(\psi_{n,\text{CANDR}} - \psi_0) = n^{1/2} \mathbb{P}_n \tilde{\phi}_{\bar{Q}_*, g_*} + o_p(1) \rightsquigarrow \mathcal{N}(0, \mathbf{\Sigma}_*) \ \text{ for } \ \mathbf{\Sigma}_* := P_0(\tilde{\phi}_{\bar{Q}_*, g_*} \tilde{\phi}^\top_{\bar{Q}_*, g_*}) \ .$$

As before, we note that condition A6 can be relaxed by utilizing cross-fitting techniques in the construction of $\psi_{n,\text{CANDR}}$. □

Chapter 4

ROBUST INFERENCE FOR QUANTILES

Quantile functions are often of scientific interest. For instance, in time to event analyses, the median survival time is a common measure of survivorship used to compare exposure groups. In epidemiology and econometric applications, counterfactual quantiles offer alternatives to mean-based causal inference. While there has been extensive work on estimation of quantiles, in both areas comparatively less attention has been paid to deriving valid and robust inference. In this chapter, we compare several model-agnostic approaches to statistical inference for quantiles and related functionals. We apply and evaluate these approaches in the context of the important application of estimation of counterfactual quantile contrasts in the potential outcomes framework.

Specifically, we consider the following three proposals for the construction of valid $(1 - \alpha) \times 100\%$ confidence regions for generalized quantiles:

1. pointwise Wald-based confidence intervals;

2. pointwise confidence intervals for quantiles based on inverting pointwise confidence intervals for the underlying distribution function;

3. pointwise Wald and percentile confidence intervals based on resampling quantiles using influence function-based multiplier bootstrap samples.

The construction of pointwise *Wald*-based $(1 - \alpha) \times 100\%$ confidence intervals for a fixed quantile or contrasts of quantiles is intuitively simplest and likely most familiar, but requires accurately estimating the standard error of the quantile(s) of interest. In practice, this entails estimating a density function at the desired quantile, itself an involved estimation problem.

As a first alternative, we consider a formulation for pointwise $(1-\alpha)\times 100\%$ confidence intervals for fixed quantiles that leverages the fact that the distribution function and quantile functions are inverse mappings. Unlike the Wald-based approach, these *inversion* intervals circumvent estimation of additional nuisances, such as density values. We show, through an argument similar to that in Brookmeyer and Crowley (1982), that the inversion procedure results in valid confidence intervals for the quantile of interest. However, we note that this procedure for pointwise intervals can not readily be used to construct intervals for contrasts of quantiles, as may be of interest in practice.

This motivates us to consider a further alternative, again based on first constructing an asymptotically linear estimator for the distribution function. We then repeatedly resample this distribution function estimator using the multiplier bootstrap (van der Vaart and Wellner 1996), and for each resample, we compute the quantiles of interest. We then apply a Wald or percentile approach to construct an interval for the quantile or contrast of quantiles of interest. The validity of this approach is based on results described in Hsu (2016). Unlike the inversion construction, the multiplier bootstrap approach naturally allows for contrasts of quantiles.

We will establish the theoretical validity of each of these inferential methods and evaluate their finite sample performance in simulation studies.

4.1 Efficiency theory for quantile estimators

4.1.1 Pathwise differentiable distribution function parameters

To study efficient estimation of general quantile parameters, we find it useful to first consider the corresponding distribution function parameters. We suppose a typical observation $Z \in \mathcal{Z}$ is distributed according to P_0 in the observed data model $\mathcal{M}$. We consider parameter mappings $F : \mathcal{M} \to \mathcal{D}$ that map elements of $\mathcal{M}$ to the space $\mathcal{D}$ of cadlag functions $\Gamma : \mathbf{R} \to [0,1]$ that satisfy $\Gamma(t) \to 0$ as $t \to -\infty$ and $\Gamma(t) \to 1$ as $t \to \infty$. Typically, the value $F(P) = F_P$ can be considered a distribution function. In our motivating example, we

are interested in quantiles derived from the full data event time distribution. However, the observed data units are subject to coarsening. Hence, we restrict our attention to parameters F that are identifiable and pathwise differentiable at every t in a set $\mathcal{T} \subset \mathbf{R}$ at each $P \in \mathcal{M}$. We let $\tau_P^* : \mathcal{Z} \times \mathcal{T} \to \mathbf{R}$ denote the efficient influence curve of F at P relative to the observed data model $\mathcal{M}$.

Before continuing our discussion of quantile parameters, we first provide the efficient influence curve for the distribution function of an event time in several common settings. In particular, we consider nonparametric models for data structures of varying complexity.

Example 1: Complete data in a nonparametric model. We first consider the familiar case where we observe i.i.d. real-valued random variables $Z_1, \ldots, Z_n$ with distribution function $F_0(z) = P_0(Z \leq z)$. In an unrestricted model for F_0, we have that any regular asymptotically linear estimator F_n of F_0 has influence curve $\tau_{P_0}^*(z;t) = I(z \leq t) - F_0(t)$.

The next example is the classic time-to-event setting with non-informative right-censoring.

Example 2: Observations subject to independent right-censoring. We now consider full data units (T, C), where T is the true event time and C is the right-censoring time. We suppose that C is independent of T and let $F_0 := F(P_0)$ denote the distribution function of T, and define $S_{P_0}(t) := 1 - F_0(t)$ and $G_{P_0}(t) := P_0(C > t)$. The observed data units of the form $Z = (Y, \Delta) = (\min(T, C), I(T \leq C))$.

We consider estimation of $F_0(t)$, restricting attention to times $t \in \mathcal{T} := \{t : 0 \leq t \leq t_{\max}\}$, where $t_{\max}$ is defined such that $G_{P_0}(t_{\max})$ is bounded away from 0.

In this setting, the survival curve $1 - F_0(t)$ is pathwise differentiable in an otherwise unrestricted (i.e. nonparametric) model with efficient influence curve

$$\tilde{\tau}_{P_0}^*(z;t) := -S_{P_0}(t) \left\{ \delta \frac{I(y \leq t)}{G_{P_0}(y) S_{P_0}(y)} - \int_0^{\min(y, t_{\max})} \frac{d\Lambda_{P_0}(u)}{G_{P_0}(u) S_{P_0}(u)} \right\} ,$$

where $\Lambda_{P_0}(y)$ is the cumulative hazard function of T at time y. Thus, the efficient influence curve for $F_0(t)$ is $\tau_{P_0}^* = -\tilde{\tau}_{P_0}^*$.

Example 3: Observation times missing at random. We suppose that the observed data unit is $Z = (W, X, XY)$, where W is a vector of covariates and X is a binary indicator of observing the time-to-event of interest Y. Hence, we only observe the value of Y when $X = 1$. In this example, the only restriction we place on the model for the distribution of Z is that the outcomes are missing at random, in the sense that Y and X are independent given W. In other words, given W, the conditional probability of observing Y does not depend on the value of Y. We denote this probability by $g_0(w) := P_0(X = 1 \mid Y = y, W = w) = P_0(X = 1 \mid W = w)$. We define the conditional distribution function $H_{P_0}(t, w) := P_0(Y \leq t \mid X = 1, W = w)$. The efficient influence function for estimating the marginal distribution function $F_0(t) := E_{P_0}\{H_{P_0}(t, W)\}$ in this setting is

$$\tau^*_{P_0}(z; t) := \frac{x}{g_0(w)}\{I(y \leq t) - H_{P_0}(t, w)\} + H_{P_0}(t, w) - F_0(t).$$

The following example covers more general time-to-event settings for causal inference in which the event times are right-censored and the censoring is possibly informative unless we condition on baseline covariates.

Example 4: Observed times are right-censored in a causal model. We consider observations of the form $Z = (W, X, Y, \Delta)$. Here W is a vector of baseline covariates and X is a binary indicator of exposure. The observation times Y are right-censored, so that $(Y, \Delta) = (\min(T, C), I(T \leq C))$, where the uncensored event time is $T = XT(1) + (1 - X)T(0)$ for potential outcome event times $T(1)$ and $T(0)$, and C is the censoring time. We suppose T and C are discrete random variables and that, given W, X and T are independent and, given X and W, C and T are independent. Then, we may write $g_{P_0}(W) := P_0(X = 1 \mid Z) = P_0(X = 1 \mid W)$. We assume that $g_{P_0}(W)$ is bounded away from both zero and one P_0-almost surely. We let $F_0 := F(P_0)$ denote the marginal distribution function of $T(1)$, and define $S_{P_0}(t, W) := P_0(T > t \mid X = 1, W)$ and $G_{P_0}(t, W) := P_0(C > t \mid X = 1, W)$.

We consider estimation of $F_0(t)$, restricting attention to times $t \in \mathcal{T} := \{0, 1, ..., t_{\max}\}$, where $t_{\max}$ is defined such that $G_{P_0}(t_{\max}, W)$ is bounded away from 0 P_0-almost surely.

In this setting, van der Laan and Rubin (2007) show that the marginal survival curve $1 - F_0(t)$ is pathwise differentiable in an otherwise unrestricted (i.e. nonparametric) model with efficient influence curve

$$\tilde{\tau}^*_{P_0}(Z;t) := -\frac{I(X=1)}{g_{P_0}(W)} \sum_{y \in \mathcal{T}} \frac{I(t \geq y) I(Y \geq y) S_{P_0}(t,W)}{G_{P_0}(y,W) S_{P_0}(y,W)} \{\Delta I(Y=y) - \lambda_{P_0}(y,X,W)\}$$
$$+ S_{P_0}(t,W) - \{1 - F_0(t)\} \ ,$$

where $\lambda_{P_0}(y,X,W) := P_0(Y = y, \Delta = 1 \mid Y \geq y, X, W)$ reduces to the conditional hazard of T given X and W at time y. Thus, the efficient influence curve for $F_0(t)$ is $\tau^*_{P_0} = -\tilde{\tau}^*_{P_0}$. We note that this setting can be extended to cover continuous time variables.

Note that to obtain the efficient influence function for the marginal distribution function of $T(0)$, we need only replace all instances of $X = 1$ in the previous definitions of g_{P_0}, S_{P_0}, G_{P_0}, and $\tau^*_{P_0}$ by $X = 0$.

4.1.2 Differentiability of the quantile function

Our present interest is in estimating a the quantile function or contrasts of quantiles corresponding to a distribution function F_P, rather than F_P itself. The quantile function Ψ_Γ of $\Gamma \in \mathcal{D}$ is defined as the inverse

$$\Psi_\Gamma(q) = \inf_{t \in \mathcal{T}} \{q \leq \Gamma(t)\} \ .$$

Under weak conditions, the inverse map is Hadamard differentiable. Through application of the functional delta method, this differentiability implies that the quantile function $\Psi_{F_P}(q)$ is itself pathwise differentiable at P in the model $\mathcal{M}$ with efficient influence curve

$$\phi^*_P(z;q) = -\frac{\tau^*_P(z; \Psi_{F_P}(q))}{f_P\left(\Psi_{F_P}(q)\right)}$$

where f_P is the density function corresponding to F_P. Formally, we have the following proposition.

Proposition 4.1 (Lemma 3.9.23 (van der Vaart and Wellner 1996)). *Let $\mathcal{D}_1[a,b]$ be the set of all restrictions of distribution functions on $\mathbf{R}$ to an interval $[a,b]$, and let $\mathcal{D}_2 \subset \mathcal{D}_1[a,b]$ be the collection of distribution functions for probability measures on interval $(a,b]$.*

(i) Let $0 < p_1 < p_2 < 1$, and let F be continuously differentiable on the interval $[a,b] = [F^{-1}(p_1) - \epsilon, F^{-1}(p_2) + \epsilon]$ for some $\epsilon > 0$, with derivative $f > 0$. Then the inverse map $G \mapsto G^{-1}$ as a mapping from $\mathcal{D}_1[a,b] \to \ell^\infty[p_1,p_2]$ is Hadamard differentiable at F tangentially to $C[a,b]$, the space of continuous functions from $[a,b]$ to $\mathbf{R}$.

(ii) Let F have compact support $[a,b]$ and be continuously differentiable on its support with derivative $f > 0$. Then the inverse map $G \mapsto G^{-1}$ as a mapping from $\mathcal{D}_2 \to \ell^\infty(0,1)$ is Hadamard differentiable at F tangentially to $C[a,b]$.

For both cases, the derivative is the mapping $\phi : h \mapsto -(h/f) \circ F^{-1}$.

4.1.3 Standard approaches to estimating quantiles

In light of the above result, an important consequence of the functional delta method is that if F_n is any asymptotically linear estimator of F_0 with influence function τ_{P_0} so that

$$n^{1/2}\{F_n(t) - F_0(t)\} \rightsquigarrow \mathbb{G}_0 \tau_{P_0}(t)$$

as a process indexed by $t \in \mathcal{T}$, then the plug-in estimator of the quantile function $\Psi_{F_n}(q)$ is itself asymptotically linear with influence function $z \mapsto \phi_{P_0}(z;q)$, and

$$n^{1/2}\{\Psi_{F_n}(q) - \Psi_{F_0}(q)\} \rightsquigarrow \mathbb{G}_0 \phi_{P_0}(q) \ .$$

Hence, one natural approach to estimating the quantile function is to first obtain an estimator of the distribution function on the region $\mathcal{T}$.

Alternatively, we may formulate $\psi_0(q)$ as the solution to an optimization problem. Let $L_q(t) = t\{q - I(t < 0)\}$ denote the tilted absolute value loss function. We define the

corresponding complete data risk function to be $R_q(\psi) = \int_{\mathcal{T}} L_q(t-\psi)dF_0(t)$. Then, the q-th quantile is

$$\psi_0(q) = \arg\min_{\psi} R_q(\psi) \ ,$$

suggesting that an estimator $\psi_n(q)$ can be constructed by minimizing an estimator $R_{q,n}(\psi)$ of $R_q(\psi)$. The sampling behavior of $\psi_n(q)$ can be studied subsequently using standard techniques in the theory of M-estimation (van der Vaart 1998). In the presence of data that are subject to censoring, missingness or other forms of coarsening, the methods of van der Laan and Robins (2003) can be used to obtain a suitable loss function for the observed data.

Regardless of how it is constructed, ψ_n is an asymptotically efficient estimator of $\psi_0 = \Psi_{F_{P_0}}$ if

$$n^{1/2}\{\psi_n(q) - \psi_0(q)\} = n^{1/2}\mathbb{P}_n\phi^*_{P_0}(q) + o_p(1) \rightsquigarrow \mathbb{G}_{P_0}\phi^*_{P_0}$$

where $o_p(1)$ holds uniformly in q and the limiting process $\mathbb{G}_{P_0}\phi^*_{P_0}$ is a P_0-Brownian bridge that has covariance function $(q_u, q_v) \mapsto P_0\{\phi^*_{P_0}(q_u)\phi^*_{P_0}(q_v)\}$ for $0 \leq q_u, q_v \leq 1$. In particular, for the q_0-th quantile, an efficient estimator $\psi_n(q_0)$ of $\psi_0(q_0)$ satisfies

$$n^{1/2}\{\psi_n(q_0) - \psi_0(q_0)\} \rightarrow_d \mathcal{N}(0, V_0^*(q_0)) \ , \tag{4.1}$$

where $V_0^*(q_0) = P_0\{\phi^*_{P_0}(q_0)^2\}$. This convergence in distribution provides the basis for the theoretical validity of the methods for constructing confidence intervals described in the following section.

4.2 Methods for obtaining pointwise confidence intervals for a fixed quantile

4.2.1 Pointwise Wald confidence intervals for the quantile

For an efficient estimator $\psi_n(q_0)$ of $\psi_0(q_0)$, an approximate $(1-\alpha)\times 100\%$ Wald confidence interval is given by the limits

$$\psi_n(q_0) \pm z_{1-\alpha/2}\left\{\frac{V_0^*(q_0)}{n}\right\}^{1/2} .$$

In practice, $V_0^*(q_0)$ is replaced by any consistent estimator, such as $V_n^*(q_0) = \mathbb{P}_n\{\phi_n^*(q_0)^2\}$ with

$$\phi_n^*(q_0) = -\frac{\tau_{P_n}^*(\psi_n(q_0))}{f_n\left(\psi_n(q_0)\right)} \ ,$$

where $\tau_{P_n}^*$ is the efficient influence function for the distribution function F_{P_0} evaluated at some estimator P_n of P_0 and f_n is an estimator of the density corresponding to F_{P_0}. The approximate Wald confidence limits based on this standard error estimator are

$$\psi_n(q_0) \pm z_{1-\alpha/2} \left\{ \frac{V_n^*(q_0)}{n} \right\}^{1/2} .$$

While asymptotically valid, in practice, constructing a consistent estimator $f_n(\psi_n(q_0))$ of $f_{P_0}(\psi_0(q_0))$ often depends on the particular data-generating mechanism. A natural choice is to construct a kernel density estimator of the form

$$f_{n,h}(\psi_n(q_0)) := \int \frac{1}{h} K\left(\frac{t - \psi_n(q_0)}{h}\right) dF_n(t) \ ,$$

where K is a second-order kernel function and $h > 0$ is the bandwidth tuning parameter. Selection of h by cross-validation is straightforward in full data models, but is made complicated when the observed data units are subject to coarsening. In such case, selection of the bandwidth h depends on constructing an appropriate loss function specific to the setting. For example, Padgett and McNichols (1984) and Marron et al. (1987) study density function estimation in the presence of independent right-censoring.

4.2.2 Inverting pointwise Wald confidence intervals for the distribution function

Whereas direct construction of pointwise Wald confidence intervals for $\psi_0(q)$ requires estimating the density $f_{P_0}(\psi_0(q))$ as an additional nuisance parameter, the Wald intervals for the distribution function do not. An approximate $(1-\alpha) \times 100\%$ Wald confidence interval for the distribution function F_0 at a fixed point $t \in \mathbf{R}$ is given by $[F_{n,\alpha}^-(t), F_{n,\alpha}^+(t)]$, with limits

$$F_{n,\alpha}^{\pm}(t) = F_n(t) \pm z_{1-\alpha/2} \left\{ \frac{\mathbb{P}_n\{\tau_{P_n}^*(t)^2\}}{n} \right\}^{1/2} .$$

Here, we suppose that F_n is an efficient estimator of F_0. Now, for any $0 < q < 1$ we define

$$\begin{aligned}
\mathcal{T}_n(\alpha, q) &= \{t \in \mathcal{T} : F^-_{n,\alpha}(t) \leq q \leq F^+_{n,\alpha}(t)\}, \\
\psi^-_{n,\alpha}(q) &= \inf \mathcal{T}_n(\alpha, q), \text{ and} \\
\psi^+_{n,\alpha}(q) &= \sup \mathcal{T}_n(\alpha, q) \ .
\end{aligned}$$

We note that $\mathcal{T}_n(\alpha, q) \subset \mathcal{T}$ is the collection of values t so that the $(1-\alpha) \times 100\%$ Wald confidence interval for $F_0(t)$ contains q. Then, we define the approximate $(1-\alpha) \times 100\%$ inversion confidence interval for $\psi_0(q_0)$ to be $[\psi^-_{n,\alpha}(q), \psi^+_{n,\alpha}(q)]$.

Theorem 4.1. *The $(1-\alpha) \times 100\%$ inversion confidence interval for $\psi_0(q_0)$ is asymptotically valid; the interval has coverage probability that is asymptotically $1-\alpha$.*

Proof. By definition, we may write the coverage probability of the inversion interval as

$$P_0\{\psi^-_{n,\alpha}(q) \leq \psi_0(q) \leq \psi^+_{n,\alpha}(q)\} = P_0\{\psi_0(q) \in \mathcal{T}_n(\alpha, q)\} \ ,$$

since the events $\{\psi^-_{n,\alpha}(q) \leq \psi_0(q) \leq \psi^+_{n,\alpha}(q)\}$ and $\{\psi_0(q) \in \mathcal{T}_n(\alpha, q)\}$ differ by a set of measure zero. The event $\{\psi_0(q) \in \mathcal{T}_n(\alpha, q)\}$ is equivalent to $\{F^-_{n,\alpha}(\psi_0(q)) \leq q \leq F^+_{n,\alpha}(\psi_0(q))\}$. However, $P_0\{F^-_{n,\alpha}(\psi_0(q)) \leq q \leq F^+_{n,\alpha}(\psi_0(q))\}$ is exactly the coverage probability for the approximate $(1-\alpha) \times 100\%$ Wald confidence interval for the distribution function at $\psi_0(q)$. Since F_n is an efficient estimator of F_0, we certainly have

$$P_0\{F^-_{n,\alpha}(\psi_0(q)) \leq q \leq F^+_{n,\alpha}(\psi_0(q))\} = P_0\left[n\frac{\{F_n(t) - q\}^2}{\mathbb{P}_n\{\tau^*_{P_n}(t)^2\}} \leq z^2_{1-\alpha}\right] \to Pr(\chi^2_1 \leq z^2_{1-\alpha_2}) = 1-\alpha \ .$$

□

This approach mirrors that of Brookmeyer and Crowley (1982), who inverted a sign test based on the Kaplan-Meier estimator of the distribution function. We expect that any similar confidence region for a quantile that is determined by inverting tests/intervals for the distribution function will avoid the need to estimate the density function $f_{P_0}(\psi_0(q))$. However, this complicates matters if we wish to subsequently estimate a contrast of quantiles

across multiple groups (e.g. quantile treatment effects). Inverting intervals for the distribution function does not (or, at least, does not obviously) facilitate inference on contrasts of quantile estimators.

4.2.3 Multiplier bootstrap confidence intervals

An alternative to inverting confidence intervals for an asymptotically linear estimator of the distribution function is to instead bootstrap F_n and use the samples to estimate the desired quantile(s). The use of bootstrap methods is often suggested as an alternative to Wald-based inference when estimating the standard error of an estimator is challenging. The most familiar version of the bootstrap takes the empirical distribution of the observations as estimator of the data-generating mechanism. We refer to this as the nonparametric bootstrap. Practical implementations of the nonparametric bootstrap typically proceed by drawing i.i.d. samples $Z_1^{(b)}, \ldots, Z_n^{(b)}$ from the original data $Z_1, \ldots, Z_n$, with $b = 1, \ldots, B$ for some large integer B. For each of the bootstrap samples, the estimator $\psi_n^{(b)}$ is computed. Inference is then reported based on the empirical distribution of the $\psi_n^{(b)}$.

To obtain nonparametric bootstrap inference for quantile or contrast of quantiles in a general setting, we do not advocate for the bootstrap that samples from the observations directly. There are both computational and theoretical reasons for this, particularly when the estimator F_n of the underlying distribution function is based on flexible nonparametric estimators. Computationally, construction of F_n may be quite demanding. This demand often depends on both the number and complexity of nuisance parameters involved in the construction of F_n. For instance, in the missing data setting of Example 3, an asymptotically efficient estimator F_n of the distribution function F_0 requires estimating both the propensity score and conditional distribution function. Even if sufficient computing resources are available to implement the nonparametric bootstrap, it has only been shown to be theoretically justified in settings where the estimator F_n is a smooth transformation of the empirical distribution of $Z_1, \ldots, Z_n$. Nuisance estimators often involve techniques such as cross-validation to select among candidate tuning parameters, so that F_n is not a smooth function of the

empirical distribution.

In light of the above considerations for the nonparametric bootstrap, we instead proceed by generating B *multiplier bootstrap* sample quantiles $\psi_n^{(1)}, \ldots, \psi_n^{(B)}$ by perturbing F_n and computing the quantile function of the perturbation $F_n^{(b)} := F_n + \frac{1}{n}\sum_{i=1}^n \xi_n^{(b)} \tau_{P_n}^*(Z_i)$. Here, $\xi_1^{(b)}, \ldots, \xi_n^{(b)}$ can be arbitrary independent random variables that are independent of $Z_1, \ldots, Z_n$ with $E(\xi) = 0$, $\text{var}(\xi) = 1$ and $E(\xi^{2+\alpha}) < \infty$ for some $\alpha > 0$. The perturbation random variables $\xi_1^{(b)}, \ldots, \xi_n^{(b)}$ are commonly taken to be generated from a standard normal or Rademacher distribution. Following this, percentile or Wald-based bootstrap confidence intervals may be constructed from the bootstrap quantile realizations $\psi_n^{(1)}, \ldots, \psi_n^{(B)}$.

The wild or multiplier bootstrap can refer to a number of procedures relating to conditional multiplier central limit theorems such as those found in Chapter 2.9 of van der Vaart and Wellner (1996). In the remainder of this section, we provide an overview of the developments that apply to drawing inference about quantiles in an arbitrary model.

First, we introduce some standard notation. For a set $\mathcal{F}$, we let $\ell^\infty(\mathcal{F})$ be the space of all bounded functions $l : \mathcal{F} \to \mathbb{R}$ and adopt the norm $\|l\|_\mathcal{F} = \sup_{f\in\mathcal{F}} |l(f)|$. We denote by BL_1 the collection of bounded Lipschitz-1 functions $h : \ell^\infty(\mathcal{F}) \to [0, 1]$. Hence, for any $h \in \text{BL}_1$, we have $|h(l_1) - h(l_2)| \leq \|l_1 - l_2\|_\mathcal{F}$. We suppose that the empirical process $\mathbb{G}_n = n^{-1/2}\sum_{i=1}^n (\delta_{Z_i} - P_0) \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where δ_{Z_i} is the dirac measure at Z_i and the limiting process $\mathbb{G}$ is a Brownian bridge. We re-state a version of the "in probability" multiplier central limit theorem in the next proposition.

Proposition 4.2 (Theorem 2.9.6 of van der Vaart and Wellner, 1996)**.** *Let $\mathcal{F}$ be a class of measurable functions. Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables with $E_\xi(\xi_1) = 0$, $var_\xi(\xi_1) = 1$, and $E(|\xi_1|^{2+\alpha}) < \infty$ for some $\alpha > 0$, independent of $Z_1, \ldots, Z_n$. Let $\mathbb{G}_{\xi,n} = n^{-1/2}\sum_{i=1}^n \xi_i(\delta_{Z_i} - P_0)$. Then, $\mathcal{F}$ is Donsker if and only if*

$$\sup_{h\in BL_1} |E_\xi h(\mathbb{G}_{\xi,n}) - Eh(\mathbb{G})| \to 0$$

in outer probability, and the sequence $\mathbb{G}_{\xi,n}$ is asymptotically measurable so that

$$E^* h(\mathbb{G}_{\xi,n}) - E_* h(\mathbb{G}_{\xi,n}) \to 0, \quad \textit{for all } h \in BL_1 \ .$$

The convergence $\sup_{h \in \mathrm{BL}_1} |E_\xi h(\mathbb{G}_{\xi,n}) - Eh(\mathbb{G})| \to 0$ in the above proposition is equivalent to the weak convergence $\mathbb{G}_{\xi,n} \rightsquigarrow \mathbb{G}$ (van der Vaart and Wellner 1996). We are interested in using this result to approximate the asymptotic process $\mathbb{G}\phi^*_{P_0}(q)$ corresponding to an efficient estimator of the quantile function. However, both $\mathbb{G}_{\xi,n}$ and $\phi^*_{P_0}$ involve unknown parameters. A natural estimator of the process $\mathbb{G}_{\xi,n}\phi^*_{P_0}(q)$ is given by

$$\hat{\mathbb{G}}_{\xi,n}\phi^*_{P_n}(q) := n^{-1/2} \sum_{i=1}^{n} \xi_i \{\phi^*_{P_n}(Z_i; q) - \mathbb{P}_n \phi^*_{P_n}(q)\} \ .$$

We note, however, that the conditional multiplier central limit theorem does not apply directly to the process $\hat{\mathbb{G}}_{\xi,n}\phi^*_{P_n}(q)$.

The multiplier bootstrap first appears in the literature as the wild bootstrap of Liu et al. (1988), based on ideas due to Wu et al. (1986) and the discussion of Beran (1986). In these original articles, the wild bootstrap was applied to obtain valid inference for the coefficients of heteroskedastic linear regression models. Kline and Santos (2012) studied a score bootstrap for M-estimators more generally. The class of estimators treated by Kline and Santos (2012), however, is restricted to finite-dimensional parameters that do not require data-adaptive estimation of nuisance parameters to evaluate the score (and hence, influence) function. These limitations are addressed in the technical report of Hsu (2016). We state the following weaker version of Theorem 3.1 of Hsu (2016) as a proposition below.

In the following proposition, we use the definition of manageability given in Pollard (1990). Let $a \cdot b$ denote the Hadamard (element-wise) product of two vectors. Then, we say that $\{\tau^*_{P_n}(Z_i; t) : i = 1, \ldots, n, \ t \in \mathcal{T}\}$ is manageable with respect to the sequence of envelopes M_n if there is a deterministic function λ such that (i) $\int_0^1 \sqrt{\log \lambda(u)} du < \infty$ and (ii) the random packing number

$$D(x|\nu \cdot M_n(z)|, \nu \cdot \{\tau^*_{P_n}(Z_i; t) : i = 1, \ldots, n, \ t \in \mathcal{T}\}) \leq \lambda(x)$$

for any $0 < x \leq 1$ and all $z \in \mathcal{Z}$, for any $\nu \in \mathbf{R}^n$ with non-negative entries.

The proposition depends on the following conditions:

Condition A1. *Suppose that $Z_1, \ldots, Z_n$ are i.i.d. random variables taking values in $\mathcal{Z}$ and distributed according to P_0. Let $F_n(t)$ be an asymptotically linear estimator of the distribution function parameter $F_0(t)$ with influence function $\tau^*_{P_0}(t)$ such that*

$$n^{1/2}\{F_n(t) - F_0(t)\} = \mathbb{G}_n \tau^*_{P_0} + o_p(1) \ ,$$

where the $o_p(1)$ term holds uniformly in t.

Condition A2. *Suppose further that there exists an envelope function M_0 such that $|\tau^*_{P_0}(z;t)| \leq M_0(z)$ for all $z \in \mathcal{Z}$ and all $t \in \mathcal{T}$, the set $\{\tau^*_{P_0}(Z_i;t) : i = 1, \ldots, n, \ t \in \mathcal{T}\}$ is manageable with respect to M_0, and that there exists a value $\alpha > 0$ such that $P_0 M_0^{2+\alpha} < \infty$.*

Condition A3. *Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables, independent of $Z_1, \ldots, Z_n$. Suppose that $E_\xi(\xi_1) = 0$, $var_\xi(\xi_1) = 1$, and $E(|\xi_1|^{2+\alpha}) < \infty$ for the same α as above.*

Condition A4. *Let $\tau^*_{P_n}(t)$ be an estimator of $\tau^*_{P_0}(t)$ with envelope M_n, and suppose that the $\{\tau^*_{P_n}(Z_i;t) : i = 1, \ldots, n, \ t \in \mathcal{T}\}$ are manageable with respect to the M_n. Suppose that*

$$\sup_{t_1, t_2 \in \mathcal{T}} |\mathbb{P}_n \tau^*_{P_n}(t_1)\tau^*_{P_n}(t_2) - P_0 \tau^*_{P_0}(t_1)\tau^*_{P_0}(t_2)| = o_p(1),$$

and that there exists $0 < \alpha' < \alpha$ such that

$$\mathbb{P}_n(M_n^2 - M_0^2) = \mathbb{P}_n(M_n^{2+\alpha'} - M_0^{2+\alpha'}) = o_p(1) \ .$$

The first condition requires that F_n is asymptotically linear at F_0. Conditions (2A) and (4A) require that the true influence function and the estimators $\tau^*_{P_n}$ are not overly complex. Condition (3A) ensures that the sequence of perturbation random variables $\xi_1, \ldots, \xi_n$ satisfy the requirements of the conditional multiplier central limit theorem.

Proposition 4.3 (Theorem 3.1 of Hsu, 2016)**.** *Suppose that conditions (A1)-(A4) hold. Let $\hat{\mathbb{G}}_{\xi,n}\tau^*_{P_n}(t) = n^{-1/2}\sum_{i=1}^n \xi_i\{\tau^*_{P_n}(Z_i;t) - \mathbb{P}_n\tau^*_{P_n}(t)\}$. Then, $\sup_{h \in BL_1} |E_\xi h(\hat{\mathbb{G}}_{\xi,n}\tau^*_{P_n}) - Eh(\mathbb{G}_0\tau^*_{P_0})| \to 0$ in outer probability, and the sequence $\hat{\mathbb{G}}_{\xi,n}\tau^*_{P_n}$ is asymptotically measurable. Moreover, $\hat{\mathbb{G}}_{\xi,n}\tau^*_{P_n}(t) \rightsquigarrow \mathbb{G}_0\tau^*_{P_0}(t)$ in probability, conditionally on the sample path $Z_1, \ldots, Z_n$.*

Under the conditions of Proposition 4.3, we find that a perturbed distribution function estimator $F_n^{(b)} = F_n + \frac{1}{n}\sum_{i=1}^n \xi_n^{(b)} \tau_{P_n}^*(Z_i)$ satisfies

$$n^{1/2}\{F_n^{(b)}(t) - F_n(t)\} = \hat{\mathbb{G}}_{\xi,n}\tau_{P_n}^*(t) \rightsquigarrow \mathbb{G}_0\tau_{P_0}^*(t),$$

in probability, conditionally on the sample path $Z_1, \ldots, Z_n$. For the quantile process, we find by application of the functional delta method that

$$n^{1/2}\{\psi_n^{(b)}(q) - \psi_n(q)\} \rightsquigarrow \mathbb{G}_0\phi_{P_0}^*(t)$$

in probability, conditionally on the sample path $Z_1, \ldots, Z_n$. Hence, both Wald and percentile bootstrap confidence intervals constructed from $\psi_n^{(1)}(q), \ldots, \psi_n^{(B)}(q)$ are asymptotically valid, and avoid having to construct an estimator of the density function. Moreover, implementation of the multiplier bootstrap avoids the need to repeat the estimation procedure used in construction of F_n itself.

A potential issue with this approach in finite samples is that the perturbed distribution function estimators $F_n^{(b)}$ may fail to be monotone. Such an estimator is not contained in the parameter space. When this occurs, the corresponding quantile value is not well-defined. We consider two possible remedies for defining the $\psi_n^{(1)}(q)$. First, we consider correction of $F_n^{(b)}$ to enforce monotonicity. In particular, Westling et al. (2018) demonstrate that under weak conditions, projection of $F_n^{(b)}$ onto the space of non-decreasing functions via isotonic regression does not change the limiting process. Denoting by $\tilde{F}_n^{(b)}$ the isotonic projection of $F_n^{(b)}$, we then define the b-th quantile estimator to be $\psi_{n,I}^{(b)}(q) := \Psi_{\tilde{F}_n^{(b)}}(q)$. The second method we consider is to define the b-th quantile estimator as the average

$$\psi_{n,A}^{(b)}(q) := \frac{1}{2}\left(\inf_{t\in\mathcal{T}}\{q \leq F_n^{(b)}(t)\} + \sup_{t\in\mathcal{T}}\{q \geq F_n^{(b)}(t-)\}\right) .$$

As $F_n^{(b)}$ is typically a step-function, we note that in this case $\psi_{n,A}^{(b)}(q)$ is simply the average of the first time index and last time index at which the graph of $t \mapsto F_n^{(b)}(t)$ crosses q. The length of this interval, $\sup_{t\in\mathcal{T}}\{q \geq F_n^{(b)}(t-)\} - \inf_{t\in\mathcal{T}}\{q \leq F_n^{(b)}(t)\}$, will tend in probability to zero at faster than $n^{-1/2}$ rate, so taking its midpoint as the quantile is asymptotically valid.

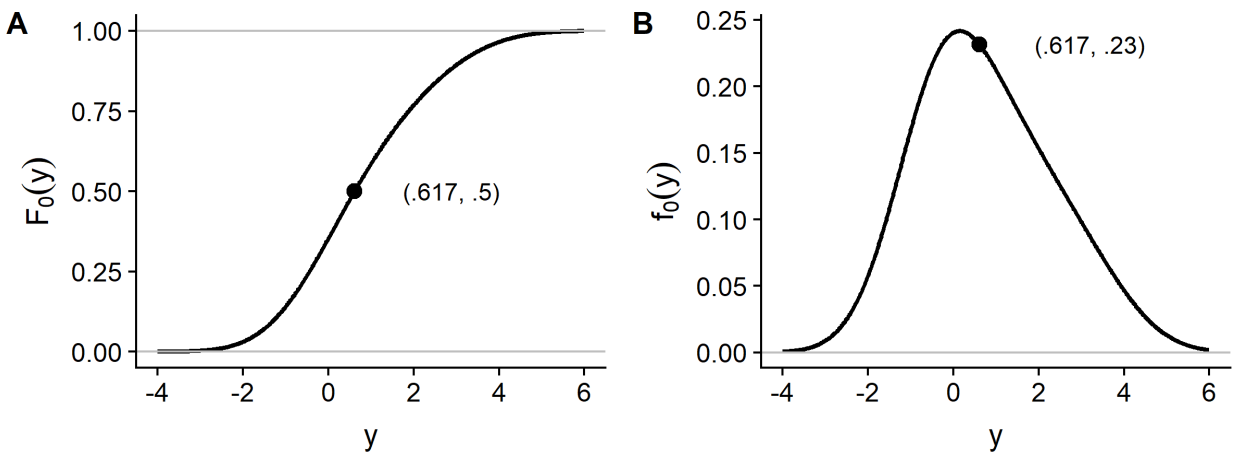

Figure 4.1: (A) The marginal distribution function F_0 and (B) marginal density f_0 for Y implied by the data-generating mechanism. The median of Y is plotted in each panel.

4.3 Simulations

In this section, we evaluate the construction of the Wald, inversion, and multiplier bootstrap confidence intervals for the median in the setting of our example 3, in which the outcomes are missing at random.

4.3.1 Data generation

For $n \in \{100, 200, 500, 1000, 2000\}$, we generated i.i.d. observations $Z_i = (W_i, X_i, Y_i)$ for $i = 1, \ldots, n$. The covariates $W_i = (W_{1i}, W_{2i})$ are such that W_{1i} is independent of W_{2i}, $W_{1i} \sim \text{Uniform}(-2, 2)$ and $W_{2i} \sim \text{Uniform}(0, 1)$. The outcomes $Y_i = \bar{Q}_0(W_i) + \epsilon_i$, where $\bar{Q}_0(w) := w_1^2 + (w_2 - .5)(w_2 - .2)w_2 - \exp(-w_2) + \sin(w_1)\exp(-w_2)$, and $\epsilon_i \sim \mathcal{N}(0, 1)$. The marginal distribution function of Y is $F_0(t) := P_0(Y \leq t) = E_{P_0}\{H_0(t, W)\}$, where the conditional distribution function $H_0(t, w) := \Phi(t - \bar{Q}(W))$ with Φ denoting the distribution function of a standard normal random variable. For this data-generating mechanism, the median is $\psi_0(1/2) \approx 0.617$. We plot F_0, its density f_0, and $\psi_0(1/2)$ in Figure 4.1.

4.3.2 Notes on construction of estimators and confidence intervals

Initial nuisance estimators $\bar{Q}_n$ and g_n were constructed using bivariate kernel regression, with all bandwidths selected by cross-validation. Then, to estimate the conditional distribution function $H_0(t, w)$, we defined $\epsilon_{n,i} := Y_i - \bar{Q}_n(W_i)$ and set

$$H_n(t, w) := \frac{1}{n} \sum_{i=1}^{n} I(\epsilon_{n,i} \leq t - \bar{Q}_n(w))$$

to be the empirical distribution function of the $\epsilon_{n,i}$ evaluated at $t - \bar{Q}_n(w)$. Our estimator of F_0 is the augmented inverse probability weighted (AIPTW) estimator

$$F_n(t) := \frac{1}{n} \sum_{i=1}^{n} H_n(t, W_i) + \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{g_n(W_i)} \{I(Y_i \leq t) - H_n(t, W_i)\} \ .$$

For a given sample, we define a uniformly-spaced mesh $\min(Y_i) = t_1 < t_2 < \ldots < t_n = \max(Y_i)$. We evaluate $F_n(t_1), \ldots, F_n(t_n)$ and compute $\psi_n(1/2) = \inf\{t_i : 1/2 \leq F_n(t_i)\}$.

We then construct 95% Wald confidence intervals for the median in two ways, varying the estimator of the density function. First, similar to above, we compute the conditional density $h_n(t, w)$ by estimating the density function of the $\epsilon_{n,i}$ by kernel density estimation (with bandwidth selected by cross-validation) and evaluate the density estimator at $t - \bar{Q}_n(w)$. We then set $f_n(t) := n^{-1} \sum_{i=1}^{n} h_n(t, W_i)$ as an estimator of f_0. We refer to the Wald intervals based on $f_n(t)$ as *Wald-KDE* intervals. An alternative estimator of the density is the to approximate the derivative of $F_n(t)$ by a difference quotient. We define

$$\tilde{f}_n(\psi_n(1/2)) := \frac{F^+_{n,\alpha}(\psi^+_{n,\alpha}) - F^-_{n,\alpha}(\psi^-_{n,\alpha})}{\psi^+_{n,\alpha} - \psi^-_{n,\alpha}} \ ,$$

where $\psi^+_{n,\alpha}$, $\psi^-_{n,\alpha}$, $F^+_{n,\alpha}(\psi^+_{n,\alpha})$, and $F^-_{n,\alpha}(\psi^-_{n,\alpha})$ are as defined above in the description of the inversion confidence intervals. We refer to the Wald intervals based on this difference quotient as *Wald-DQ* intervals.

The inversion confidence intervals are based on $\psi^+_{n,\alpha}$, $\psi^-_{n,\alpha}$, $F^+_{n,\alpha}(\psi^+_{n,\alpha})$, and $F^-_{n,\alpha}(\psi^-_{n,\alpha})$ computed based on the mesh.

We implemented the multiplier bootstrap by drawing the perturbation random variables $\xi_i^{(b)}$ from a $\mathcal{N}(0, 1)$ distribution and computing both $\psi^{(b)}_{n,I}(1/2)$ and $\psi^{(b)}_{n,A}(1/2)$ for $b =$

$1, \ldots, 2000$. Then, we used the $B = 2000$ bootstrap samples for each method to compute 95% Wald and percentile confidence intervals for the median. We refer to the multiplier bootstrap intervals based on $\psi^{(b)}_{n,I}(1/2)$ as *MBS-I, Wald* intervals and *MBS-I, %* intervals, respectively. Similarly, we refer to the intervals based on $\psi^{(b)}_{n,A}(1/2)$ as *MBS-A, Wald* and *MBS-A, %* intervals.

4.3.3 Results

We repeated the data generation and confidence interval construction in 1000 independent simulation replicates, and plot the empirical coverage and length of the 95% confidence intervals in Figure 4.2. We find that the coverage of both the Wald-KDE and inversion confidence intervals outperformed the Wald-DQ and multiplier bootstrap estimators at lower sample sizes. However, all of the methods for interval construction approach the nominal 95% coverage as the sample size increases. Of interest is that the confidence interval lengths tended to increase with the coverage probability. The performance of the multiplier bootstrap confidence intervals did not appear sensitive to the use of isotonic regression or the averaging procedure for defining the median. However, we do note that the multiplier bootstrap-based Wald confidence intervals appear to outperform the percentile multiplier bootstrap intervals. We might expect this performance gap to shrink as the number of bootstrap replicates increases. That the performance of the Wald-DQ confidence intervals is competitive with the multiplier bootstrap confidence intervals is also of interest, as the difference quotient is typically considered to be a poor estimator of the density function.

4.4 Discussion

We have shown evidence that there are several methods for constructing valid confidence intervals for quantiles and related quantities. In our simulation study, we found that the Wald-KDE and inversion intervals had stronger performance at lower sample sizes. However, we were able to leverage the known regression structure in this problem to ease estimation of the density function. In other complex data settings, this exact method may not be jus-

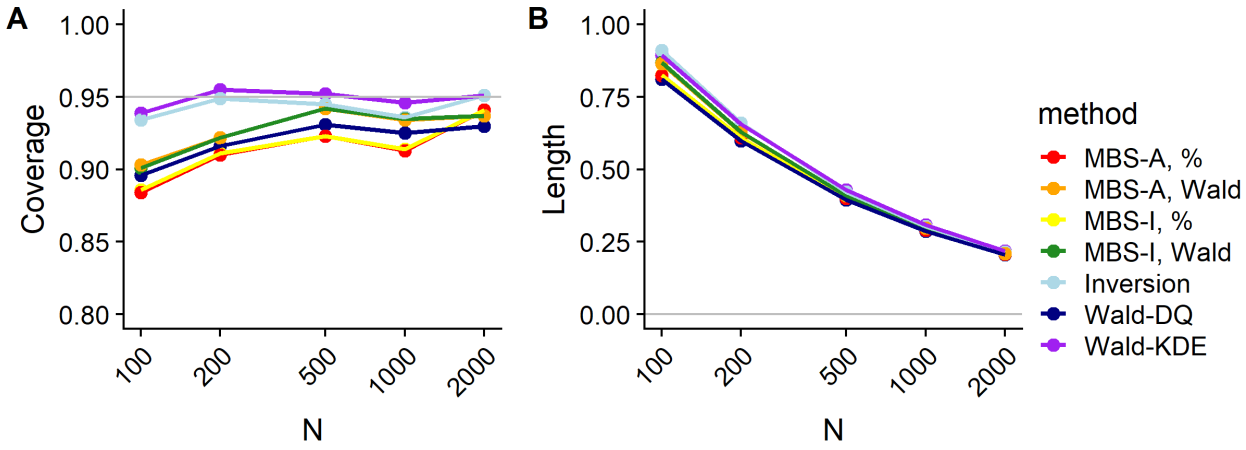

Figure 4.2: Empirical coverage and length of approximate 95% confidence intervals for sample sizes $n \in \{100, 200, 500, 1000, 2000\}$ computed using 1000 simulated datasets.

tifiable. In such cases, potentially complex cross-validation procedures are often required in order to select from among candidate density estimators, which may push the analyst toward using a different approach. If contrasts of quantiles are of interest, the multiplier bootstrap approach to confidence interval construction appears promising, though it may be more computationally intensive. Unlike the nonparametric bootstrap, multiplier bootstrap confidence intervals are based on resampling the (estimated) influence function for the parameter of interest. Hence, it is not clear how the higher order behavior of the multiplier bootstrap compares to other procedures. A possible direction towards this understanding may stem from arguments similar to those in Kline and Santos (2012) where it is shown that, in terms of Edgworth expansions of studentized statistics, the score and wild bootstrap are equivalent to higher-order than the wild bootstrap and standard Wald confidence interval. Showing the equivalence of the multiplier bootstrap to another method for constructing confidence intervals may facilitate theoretical comparisons to the inversion and Wald methods of interval construction.

APPENDIX

4.A Sample R code for constructing confidence intervals

```
## Helper function to perturb cdf using multiplier bootstrap
a_wild_cdf_appears <- function(cdf, influence.curve, dist = rnorm) {
  n <- nrow(influence.curve)
  M <- ncol(influence.curve)
  U <- dist(n)
  U.mat <- matrix(U, nrow = n, ncol = M, byrow = T)
  wild_process <- apply(U * influence.curve, 2, mean)
  wild_cdf <- cdf + wild_process
  return(wild_cdf)
}

## If an estimated cdf is non-monotone ("wiggly"), either
## (1) compute its monotone projection via isotonic regression
##     and then compute the quantile; or
## (2) check the first and last time the wiggly curve crosses
##     the corresponding probability, define the quantile as
##     the average of the abscissae
wiggly_quant <- function(mesh, wiggly, prob=0.5){
  # Quantile from isotonic regression fit
  iso_wiggly <- isoreg(x = mesh, y = wiggly)
  quant_iso <- inv(mesh, iso_wiggly$yf, p = prob)
```

```
  # Quantile as average of first/last crossing of prob
  first_ab <- mesh[which(mesh == min(mesh[wiggly >= prob]))] #First crossing
  last_ab <- mesh[which(mesh == max(mesh[wiggly < prob])) + 1] #Last
   crossing
  quant_avg <- (first_ab + last_ab) / 2

  out <- c("quant_iso"=quant_iso, "quant_avg"=quant_avg)
  return(out)
}

## Generate empirical multiplier bootstrap quantile draws
bs_quantile <- function(mesh, cdf, influence.curve, dist = rnorm, B=200,
   prob = 0.5){
  replicate(B, expr = {
    wcdf <- a_wild_cdf_appears(cdf, influence.curve, dist)
    wiggly_quant(mesh, wcdf, prob)
  })
}

## Quantile confidence intervals

# cdf -- a vector of the (non-decreasing!) cdf values on a mesh of length M
# influence.function -- an n x M matrix with each row denoting an
   observation and each column denoting that observation's influence
   function evaluation at a distinct value of the mesh for cdf
```

```
qcdf <- function(mesh, cdf, influence.curve, q = 0.50, alpha = 0.05, reps =
   500,
                pdf=NULL) {
  n <- nrow(influence.curve)
  M <- length(cdf)
  if( M != ncol(influence.curve) ) warning("length of mesh implied by cdf
   and influence.function differ!")
  if( M != length(mesh) ) warning("length of mesh implied by mesh and cdf
   differ!")

  ## Target quantile, point estimate
  q_index <- max(which(cdf <= q))
  quant <- mesh[q_index]

  inv_time <- system.time({
    ## Compute Wald confidence intervals for cdf at each point of mesh
    n <- nrow(influence.curve)
    z_a <- abs(qnorm(p = alpha/2))
    cdf.SE <- sqrt(colSums(influence.curve^2)) / n
    cdf.lo <- cdf - z_a * cdf.SE
    cdf.up <- cdf + z_a * cdf.SE

    # Compute mesh point for quantile and
    # which mesh point has first/last cdf interval to include quantile
    lo <- max(which(cdf.up <= q))
    up <- min(which(cdf.lo >= q))
```

```
  ## Inversion-based confidence interval for q-th quantile
  inv.lo <- mesh[lo]
  inv.up <- mesh[up]
})

## Wald-based confidence interval for q-th quantile
slope <- diff(cdf[c(lo,up)]) / diff(mesh[c(lo,up)]) # naive density
 estimate at q-th quantile
wald.lo <- quant - z_a * cdf.SE[q_index] / slope
wald.up <- quant + z_a * cdf.SE[q_index] / slope

if(!is.null(pdf)){
  # pdf should be a list with vectors pdf and mesh, mesh should be
  # the same as the mesh used elsewhere in this function and pdf
  # consists of estimates of the pdf at each point in mesh
  qpdf <- pdf[q_index]
  wald.f.lo <- quant - z_a * cdf.SE[q_index] / qpdf
  wald.f.up <- quant + z_a * cdf.SE[q_index] / qpdf
}

## Multiplier-bootstrap confidence intervals for q-th quantile
bs_time <- system.time({
  # generating the bootstrap quantiles
  bs_quants <- bs_quantile(mesh, cdf, influence.curve,
                           dist = rnorm, B = reps, prob = q)

  # 1: Percentile bootstrap confidence interval
```

```
  bs_iso.pct.lo <- quantile(bs_quants[1,], probs = alpha / 2, names = F)
  bs_iso.pct.up <- quantile(bs_quants[1,], probs = 1 - alpha / 2, names =
 F)

  bs_avg.pct.lo <- quantile(bs_quants[2,], probs = alpha / 2, names = F)
  bs_avg.pct.up <- quantile(bs_quants[2,], probs = 1 - alpha / 2, names =
 F)

  # 2: Wald bootstrap confidence interval
  bs_iso_SE <- sd(bs_quants[1,])
  bs_iso.wald.lo <- quant - z_a * bs_iso_SE
  bs_iso.wald.up <- quant + z_a * bs_iso_SE

  bs_avg_SE <- sd(bs_quants[2,])
  bs_avg.wald.lo <- quant - z_a * bs_avg_SE
  bs_avg.wald.up <- quant + z_a * bs_avg_SE
})
if(is.null(pdf)){
  out <- c(wald.dq.lo = wald.lo, wald.dq.up = wald.up,
           inverse.lo = inv.lo, inverse.up = inv.up,
           bs.iso.pct.lo = bs_iso.pct.lo, bs.iso.pct.up = bs_iso.pct.up,
           bs.avg.pct.lo = bs_avg.pct.lo, bs.avg.pct.up = bs_avg.pct.up,
           bs.iso.wald.lo = bs_iso.wald.lo, bs.iso.wald.up = bs_iso.wald.
 up,
           bs.avg.wald.lo = bs_avg.wald.lo, bs.avg.wald.up = bs_avg.wald.
 up)
} else {
```

```
    out <- c(wald.dq.lo = wald.lo, wald.dq.up = wald.up,
             wald.f.lo = wald.f.lo, wald.f.up = wald.f.up,
             inverse.lo = inv.lo, inverse.up = inv.up,
             bs.iso.pct.lo = bs_iso.pct.lo, bs.iso.pct.up = bs_iso.pct.up,
             bs.avg.pct.lo = bs_avg.pct.lo, bs.avg.pct.up = bs_avg.pct.up,
             bs.iso.wald.lo = bs_iso.wald.lo, bs.iso.wald.up = bs_iso.wald.
   up,
             bs.avg.wald.lo = bs_avg.wald.lo, bs.avg.wald.up = bs_avg.wald.
   up)
  }
  attr(out, "times") <- c(wald = attr(pdf, "pdf_time")["elapsed"], inv=
   inv_time["elapsed"], bs=bs_time["elapsed"])
  return(out)
}

# Function to check length and coverage of one realization of CIs
ci.eval <- function(ci, param=0, pdf=TRUE){
  if(pdf==TRUE){
    rnames <- c("Wald-DiffQuo", "Wald-pdf", "Inverse",
                "BS-Iso-Pct", "BS-Avg-Pct",
                "BS-Iso-Wald", "BS-Avg-Wald")
  } else {
    rnames <- c("Wald-DiffQuo", "Inverse",
                "BS-Iso-Pct", "BS-Avg-Pct",
                "BS-Iso-Wald", "BS-Avg-Wald")
  }
  ci.mat <- matrix(ci, ncol=2, byrow = TRUE,
```

```
                    dimnames = list(rnames,
                                    c("Lower", "Upper")))
  len <- ci.mat[,"Upper"]-ci.mat[,"Lower"]
  cov <- ci.mat[,"Lower"] <= param & param <= ci.mat[,"Upper"]
  out <- c(Length = len, Coverage = cov)
  return(out)
}

set.seed(2019)
n <- 100
X <- rnorm(n)
mesh <- seq(min(X), max(X), by = 1/n)
cdf <- ecdf(X)(mesh)
influence.curve <- t(sapply(X, FUN = function(x) {(x <= cdf) - cdf}))
plot(mesh, cdf, type="l")
CI <- qcdf(mesh, cdf, influence.curve, reps = 5e2)
ci.eval(CI)
```

BIBLIOGRAPHY

D. Benkeser, M. Carone, M. J. van der Laan, and P. B. Gilbert. Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880, 2017.

R. Beran. Discussion: Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4):1295–1298, 1986.

J. Bradic, S. Wager, and Y. Zhu. Sparsity double robust inference of average treatment effects. *arXiv preprint arXiv:1905.00744*, 2019.

R. Brookmeyer and J. Crowley. A confidence interval for the median survival time. *Biometrics*, pages 29–41, 1982.

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang. Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 2019a.

A. Buja, L. Brown, A. K. Kuchibhotla, R. Berk, E. George, and L. Zhao. Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, 2019b.

M. Carone, A. R. Luedtke, and M. J. van der Laan. Toward computerized efficient estimation in infinite-dimensional models. *Journal of the American Statistical Association*, in press.

A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6:1059, 2012.

G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

I. Díaz. Statistical inference for data-adaptive doubly robust estimators with survival outcomes. *Statistics in Medicine*, 2019.

I. Díaz and M. J. van der Laan. Doubly robust inference for targeted minimum loss–based estimation in randomized trials with missing outcome data. *Statistics in medicine*, 36(24): 3807–3819, 2017.

O. Dukes, V. Avagyan, and S. Vansteelandt. High-dimensional doubly robust tests for regression parameters. *arXiv preprint arXiv:1805.06714*, 2018.

B. S. Graham and C. C. d. X. Pinto. Semiparametrically efficient estimation of the average linear regression function. Technical report, National Bureau of Economic Research, 2018.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

S. Hattori and M. Henmi. Estimation of treatment effects based on possibly misspecified cox regression. *Lifetime data analysis*, pages 1–26, 2012.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Y.-C. Hsu. Multiplier bootstrap for empirical processes. Technical report, Institute of Economics, Academia Sinica, Taipei, Taiwan, 2016.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. The Regents of the University of California, 1967.

C. A. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pages 1548–1562, 1987.

P. Kline and A. Santos. A score based approach to wild bootstrap inference. *Journal of Econometric Methods*, 1(1):23–41, 2012.

R. Y. Liu et al. Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16(4):1696–1708, 1988.

J. Marron, W. Padgett, et al. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics*, 15(4):1520–1535, 1987.

R. Neugebauer and M. van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382, 1994.

Y. Ning, S. Peng, and K. Imai. Robust estimation of causal effects via high-dimensional covariate balancing propensity score. *arXiv preprint arXiv:1812.08683*, 2018.

J. O'Quigley. *Proportional hazards regression*. Springer, 2008.

W. Padgett and D. T. McNichols. Nonparametric density estimation from censored data. *Communications in Statistics-Theory and Methods*, 13(13):1581–1611, 1984.

J. Pfanzagl. *Contributions to a General Asymptotic Statistical Theory*. 1982.

D. Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

A. Rotnitzky, E. Smucler, and J. M. Robins. Characterization of parameters with a mixed bias property. *arXiv preprint arXiv:1904.03725*, 2019.

M. Schemper, S. Wakounig, and G. Heinze. The estimation of average hazard ratios by weighted cox regression. *Statistics in medicine*, 28(19):2473–2489, 2009.

E. Smucler, A. Rotnitzky, and J. M. Robins. A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

C. A. Struthers and J. D. Kalbfleisch. Misspecified proportional hazard models. *Biometrika*, 73(2):363–369, 1986.

W. Stute. The statistical analysis of kaplan-meier integrals. *Lecture Notes-Monograph Series*, pages 231–254, 1995.

Z. Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *arXiv preprint arXiv:1801.09817*, 2018.

M. J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. van der Laan. Targeted estimation of nuisance parameters to obtain valid statistical inference. *The International Journal of Biostatistics*, 10(1):29–57, 2014.

M. J. van der Laan and J. Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer, 2011.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. van der Laan and D. Rubin. A note on targeted maximum likelihood and right censored data. 2007.

A. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998.

A. van der Vaart and J. A. Wellner. Weak convergence and empirical processes: with applications to statistics. 1996.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

T. Westling, M. van der Laan, and M. Carone. Correcting an estimator of a multivariate monotone function with isotonic regression. Technical report, 10 2018.

D. Whitney, A. Shojaie, and M. Carone. Models as (deliberate) approximations. *Statistical Science*, in press.

C.-F. J. Wu et al. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics*, 14(4):1261–1295, 1986.

R. Xu and J. O'Quigley. Estimating average regression effect under non-proportional hazards. *Biostatistics*, 1(4):423–439, 2000.

Z. Yu and M. J. van der Laan. Measuring treatment effects using semiparametric models. Technical report, UC Berkeley Department of Biostatistics Working Paper Series, 2003.

W. Zheng and M. J. van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.