Citation | Evans, Katherine Louise. 2017. Contributions to Semiparametric Methods for Incomplete Data. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Citable link | http://nrs.harvard.edu/urn-3:HUL.InstRepos:41141508

Terms of Use | This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
Contributions to Semiparametric Methods for Incomplete Data

A thesis presented
by
Katherine Louise Evans
to
The Department of Biostatistics
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics
Harvard University
Cambridge, Massachusetts
May 2017
Contributions to Semiparametric Methods for Incomplete Data

Abstract

Chapter 1: The effect of treatment on the treated (ETT) is a common parameter of interest in causal inference. Identification of ETT typically relies on an assumption of no unobserved confounding. When information on a subset of potential confounders is not observed in a main study, external data from a validation study with more detailed confounding information may be used, under certain assumptions, to help mitigate confounding. In the absence of missing data, a common approach to account for confounding is based on the propensity score. Recently such methods have been extended to address missing confounder data in a main-validation study context; however existing methods rely on overly restrictive assumptions that are unlikely to hold in practice. To address this problem, we develop a novel approach which entails constructing an extended propensity score (EPS) which preserves essential properties of a standard propensity score, but with the additional advantage that it can be evaluated even for subjects with missing confounders. The approach is universal in the sense that it applies to virtually any outcome scale, whether binary, polytomous, or continuous. The finite sample performance of the proposed approach is carefully evaluated and compared to several existing methods in extensive simulation studies. The proposed EPS approach is also illustrated in an application examining the effect of surgical resection on survival time among 14,312 Medicare beneficiaries with malignant neoplasm of the brain using 2,391 patients in SEER-Medicare for the validation study.

Chapter 2: Missing data and confounding are two problems researchers face in observational studies for comparative effectiveness. Williamson et al (2012) recently proposed a unified approach to handle both issues concurrently using a multiply-robust (MR) methodology under the assumption that confounders are missing at random. Their approach considers a union of models in which any submodel has a parametric component while the remaining models are unrestricted. We show that while their estimating function is MR in theory, the possibility for multiply robust inference is complicated by the fact that parametric models for different components of the union model are not variation independent and therefore the MR property is unlikely to hold in practice. To address this, we propose an alternative transparent parametrization of the likelihood function, which makes explicit the model dependencies between various nuisance functions needed to evaluate the MR efficient score. The proposed method is genuinely
doubly-robust (DR) in that it is consistent and asymptotic normal if one of two sets of modeling assumptions holds, and we establish that in a sense, this is the best one can achieve in this framework. We evaluate the performance of the DR method via a simulation study.

Chapter 3: This chapter investigates the problem of making inference about a parametric model for the regression of an outcome variable $Y$ on covariates $(V, L)$ when data are fused from two separate sources, one which contains information only on $(V, Y)$ while the other contains information only on covariates $(V, L)$. This data fusion setting may be viewed as an extreme form of missing data in which the probability of observing complete data $(V, L, Y)$ on any given subject is zero. We have developed a large class of semiparametric twin inverse probability weighting (TIPW) estimators, which includes doubly robust (DR) estimators, of the regression coefficients in fused data. The proposed method is DR in that it is consistent and asymptotically normal if, in addition to the model of interest, we correctly specify a model for either the data source process under an ignorability assumption, or the distribution of unobserved covariates. We evaluate the performance of the proposed methodologies via an extensive simulation study.
An Extended Propensity Score Approach to Account for Missing Confounders when Estimating Causal Effects

1.1 Introduction .................................................. 2
1.2 Existing Methods ............................................. 3
1.3 Preliminaries and the Full Data Problem ...................... 6
1.4 Main/Validation Study Design for Identification .............. 7
   1.4.1 ETT Identification Conditions .......................... 7
   1.4.2 Estimation and Inference .............................. 8
   1.4.3 Simulation Study ..................................... 10
1.5 Main Study-only ETT .......................................... 13
   1.5.1 Identification and Inference .......................... 13
   1.5.2 Simulation Study ..................................... 15
1.6 Data Application .............................................. 16
   1.6.1 Study Populations .................................... 16
1.7 Discussion .................................................. 20

A Coherent Likelihood Parameterization for Doubly Robust Estimation of a Causal Effect with Missing Confounders

2.1 Introduction .................................................. 23
2.2 Preliminaries ............................................... 25
   2.2.1 Full Data Setting .................................... 25
   2.2.2 Missing Data Setting ................................. 27
2.3 Reparametrization .............................................. 29
2.4 Implementation .............................................. 30
2.5 Simulation Study ............................................. 33
2.6 Discussion ...................................................... 38

Doubly Robust Regression Analysis for Data Fusion 39
3.1 Introduction .................................................... 40
3.2 Preliminaries ................................................... 43
3.3 Twin IPW Estimating Functions and DR Estimating Functions .... 44
3.4 Feasible Twin IPW and DR Estimating Equations .................. 45
3.5 Simulation Study .............................................. 48
3.6 Local Efficiency ............................................... 53
3.7 Discussion ...................................................... 54

References .......................................................... 56

A Proofs and results for Chapter 1 61
A.1 Proof of (1.2): ................................................. 61
A.2 Estimation ....................................................... 61
A.3 Data Simulation ............................................... 64
  A.3.1 Description .................................................. 64
  A.3.2 Requirements .............................................. 64
  A.3.3 Data Generation ........................................... 65
  A.3.4 Additional relationship ................................... 71
A.4 Asymptotic Distribution of Estimator .............................. 73
  A.4.1 For $\eta$: .................................................... 73
  A.4.2 For $\lambda$: ................................................. 73
  A.4.3 For $\Psi$: .................................................. 74

B Proofs and results for Chapter 2 74
B.1 Influence Function Derivation .................................... 74
B.2 Proof of Double Robustness for Full Data ......................... 77
C Proofs and results for Chapter 3

C.1 Proof of Double Robustness: ........................................ 85
C.2 Proof of 3.13 .......................................................... 86
C.3 Closed form for DR estimator ....................................... 87
List of Figures

1.1 Simulations Results using $\Psi$ .................................................. 12
1.2 Results of simulation study using $\Psi^*$ ..................................... 15
1.3 Results of data analysis ................................................................. 19
3.1 Simulation results ................................................................. 49
3.2 Simulation results when $AC$ interaction is small ....................... 50
3.3 Simulation results when $g(V)$ is chosen poorly ......................... 51
3.4 Simulation results for a small sample size ................................. 52
List of Tables

1.1 Characteristics of the primary data and validation data .................. 18
Acknowledgments

I would like to express my sincere gratitude to my advisor, Eric Tchetgen Tchetgen for his guidance, genius, enthusiasm, and eternal patience. Words cannot describe the impact he has had on my life.

Further, I thank my dissertation committee, Jamie Robins and Miguel Hernán, for their contributions of scientific expertise, thoughtful comments, and hard questions. I would also like to thank my other coauthor, Francesca Dominici, for her encouragement and invaluable insights.

I am deeply appreciative of the wonderful friends I have had the fortune to know throughout my graduate experience. In particular, I thank Anna Decker, James Russell, and Tianyi “Tai” Cai for their friendship, unwavering support, and countless ways they have made my life better.

Finally, I must thank my family. I thank my brother, Christopher, for first encouraging me to take a statistics class well over a decade ago. And of course, I thank my loving parents, Craig and Claudia, without whom I would not be the woman I am today.
An Extended Propensity Score Approach to Account for Missing Confounders when Estimating Causal Effects

with

Francesca Dominici

and

Eric J. Tchetgen Tchetgen
1.1 Introduction

The effect of treatment on the treated (ETT) is a common parameter of interest in comparative effectiveness research. In randomized trials ETT coincides with the average causal effect (ACE). However for cases with potentially high barriers for treatment, it may be unrealistic to estimate the effect of treatment if it were applied to all patients. Instead, greater clinical interest may lie in the effect of treatment on those who elect to be treated. Under the assumption of no unmeasured confounding, it is well known that ETT can be estimated by re-weighting the untreated by the estimated odds of being treated [2]. Unfortunately, this method breaks down when important potential confounders are not available for all subjects in the sample constituting the main study sample. A possible remedy considered in the literature is to augment the main study sample with a validation sample that contains complete information on potential confounders missing from the main study [57, 30, 25]. This information sometimes can be used to correct for confounding bias in estimating ETT. This strategy is commonly used when the main study population comes from a large observational database with a large number of subjects but limited information on potential confounders. Validation data can come from internal secondary samples, from the source population, or from external sources.

To adjust for bias, we propose a method based on the propensity score, first proposed by Rosenbaum and Rubin [47], as it is well understood, easy to implement, and favored in modern studies of comparative effectiveness for practical considerations [56, 60]. Propensity scores allow the analyst to account for multiple measured potential confounders without having to specify an outcome regression model relating potential confounders to the outcome, which is advantageous in finite samples particularly if the outcome is rare. Instead, propensity score methods rely on correct specification of a regression model for the treatment given potentially confounding covariates. This method also prevents selectively choosing covariates in the outcome regression in favor of an anticipated treatment effect.

In this chapter we distinguish between two ETT parameters, "verify-in-sample" ETT and "verify-out-of-sample" ETT. The "verify-in-sample" ETT is the ETT where the validation population is a sample from the main population and thus ETT is calculated for the population from which the pooled main-validation sample is an i.i.d sample. The "verify-out-of-sample" ETT is the
ETT specific to the population from which the main study is an i.i.d. sample [8]. We refer to the "verify-in-sample" ETT as ETT and the "verify-out-of-sample" ETT as Main Study ETT.

Our approach constructs an extended propensity score (EPS) that is simple to implement and allows one to recover and incorporate confounder information from a validation sample under certain fairly general conditions formalized below. EPS can be used to calculate ETT or Main Study ETT. The EPS approach preserves some of the advantages of standard propensity scores, but has the additional advantage that it can be evaluated even for subjects in the main study with missing potential confounders. When the main study is much larger than the validation study, traditional missing data methods break down. For example, inverse probability of censoring weighting method (IPCW), a standard propensity score approach, is likely to break down, as it only uses complete cases, and multiple imputation by chained equations (MICE) may tend to over-extrapolate. EPS avoids this problem by preserving the large sample size from the main study while borrowing information from the validation study.

In section 2 we give a brief overview of several existing methods to address the missing confounder problem. In section 3, we review existing results on identification and inference about ETT when confounders are fully observed. In section 4 we develop the extended propensity score, discuss implementation, and evaluate the finite sample performance of the EPS approach and compare it to several existing methods using simulation. In section 5 we develop an EPS-based approach to evaluating the Main Study ETT. Section 6 contains an illustration of EPS on the effect of surgical resection on survival time among an elderly population in Medicare data using SEER-Medicare as the validation study. We conclude in section 7 with a discussion. Throughout, proofs and derivations can be found in the appendix.

1.2 Existing Methods

The analysis of main/validation studies has drawn tremendous attention in biostatistics and epidemiology in recent years, primarily due to the ubiquity of missing potential confounder information in large studies of comparative effectiveness. Here we focus primarily on three recent analytic techniques in the literature: Propensity Score Calibration [57], Bayesian Propensity Score [30], and Two-Stage Calibration [25].

Propensity Score Calibration (PSC) considers the propensity score based solely on potential
confounders measured in the main study as a mismeasured version of the propensity score based on more complete potential confounder information available in the validation sample. [57]. This “error-prone” propensity score, constructed using only the reduced set of potential confounders is related to the “gold standard” propensity score which uses data on all potential confounders available in the validation study. Specifically, Stürmer et al. propose the use of regression calibration, a well-established and widely used method for measurement error correction [48], to adjust the regression coefficients in the outcome model in the main study by using the gold standard propensity score. The primary advantage to using PSC is that it is concerned only with the propensity score and therefore does not require outcome information in the validation study.

However, the validity of PSC is questionable in most observational settings because PSC requires the assumption that the error-prone score is a surrogate for the gold-standard. Specifically, the surrogacy assumption, on which it relies, states that the error-prone score is independent of the outcome given the gold-standard score and the treatment [55]. Surrogacy is a strong assumption because it implies non-differential measurement error, an assumption that is unlikely to hold. When outcome information is available in the validation study, Stürmer et al. suggest evaluating surrogacy by checking whether the direction of confounding due to any observed confounders is in the same direction as the confounding due to the unobserved confounders; surrogacy is violated when the direction of the confounders differs. However, surrogacy may not hold even if the direction test is satisfied. One final limitation is that PSC assumes that the main and validation populations are exchangeable. This assumption is stringent and limits options for the choice of validation study.

McCandless et al [30] proposed a method called Bayesian adjustment for missing potential confounders using propensity scores (BayesPS). This method uses a Bayesian framework to incorporate external potential confounding information through propensity scores. In this framework the propensity score is unknown; a joint model is specified for the main and validation data with the propensity score as a missing covariate. The uncertainty of the propensity score is integrated out of the likelihood using Markov Chain Monte Carlo simulation. BayesPS is flexible in that it can be used when the potential confounders are both continuous and categorical. Additionally, BayesPS avoids the surrogacy assumption required by PSC. However, BayesPS requires a strong modeling assumption about the marginal relationship between measured and unmeasured potential confounders, as it requires that they be independent of each other or marginally follow some
pre-specified relationship. Strong correlation between the measured and unmeasured potential confounders may cause over-adjustment. Moreover, BayesPS performs poorly when the validation study sample size is small, and incorporating any survey weights into the BayesPS framework is difficult, according to McCandless et al. These constraints limit the generality of BayesPS [30].

Two-stage calibration (TSC) is similar to PSC, in that it combines results from the main and validation studies to adjust the propensity score from the main sample [25]. However, unlike PSC and BayesPS, TSC does not rely on the surrogacy assumption or any measurement error model assumptions about the relationship between measured and unmeasured potential confounders. Like PSC and BayesPS, TSC requires the representative assumption, namely that the validation study is a simple random sample of the main study, and thus the unmeasured potential confounders are missing completely at random (MCAR) [9]. Additionally, two-stage calibration makes a strong assumption that the treatment effect is constant conditional on both the gold-standard and error-prone propensity scores. Finally, like BayesPS (and like IPCW and MICE), TSC is sensitive to the size of the validation study.

Our proposed EPS estimator relaxes several of the assumptions made by PSC, BayesPS and TSC. EPS does not rely on a measurement error model. Most importantly, EPS does not assume that the main study and validation populations are unconditionally exchangeable; that is, it does not assume the validation study is a simple random sample of the main study. EPS makes the weaker assumption that, conditional on completely observed variables in the main study (including the outcome), the probability of being in the validation study is independent of the unmeasured potential confounders. This is in fact a missing at random (MAR) assumption, which is less stringent than MCAR. When the validation sample is from an internal secondary source, it may be reasonable to assume the validation and main study samples are representative of the same underlying population. In such a situation, methods that only assume an MCAR structure would be valid. However, when the validation sample is external, it may be more prudent to use methods which only require MAR.

In this chapter we also will consider the following well-established missing data methods: multiple imputation by chained equations (MICE) [15] and inverse probability of censoring weighting (IPCW) [46] both in simulation and in our application.
1.3 Preliminaries and the Full Data Problem

Let \( A \) denote a binary treatment, \( A \in \{0, 1\} \), and \( Y_1 \) and \( Y_0 \) denote the potential outcomes under treatment and control conditions respectively. Let \( W \) denote pre-treatment covariates assumed to predict the outcome and the treatment. Therefore the \( W \) covariates are confounders. The effect of treatment on the treated (ETT) on the additive scale is defined as \( E[Y_{a=1} - Y_{a=0} | A = 1] = \theta - \Psi \) where \( \theta = E[Y_{a=1} | A = 1] \) and \( \Psi = E[Y_{a=0} | A = 1] \).

Throughout, we make the following standard causal assumptions in order to identify ETT:

1. Consistency: \( Y = Y_A \) almost surely;
2. No unmeasured confounding: \( A \perp Y_0 | W \);
3. Positivity: \( \frac{P(A=0|W)}{P(A=1|W)} > 0 \) almost surely.

In the absence of missing data, it is well known that under assumption 1, \( \theta = E[Y | A = 1] \) [2, 23]. Under assumptions 1-3, \( \Psi \) can be written as

\[
\Psi = E[Y_{a=0} | A = 1] = \frac{1}{P(A = 1)} E \left[ (1 - A) \frac{P(A = 1 | W)}{P(A = 0 | W)} Y \right]. \tag{1.1}
\]

Next, suppose that in the main study only a subset of confounders \( C \) is observed, so that \( W = \{L, C\} \) where \( L \) is not observed in the main study. Therefore, the main study sample are i.i.d. observations on \( (Y, A, C) \). We assume we have a validation sample with i.i.d. observations on \( (Y, A, W) \). As motivation for the missing data methods described in the next section, we wish to establish a representation of \( \Psi \) in this setting which does not depend on \( L \). This is given in the following result:

**Result:**

For all \( c \) and \( y \), define \( q(c, y) = P(A = 1 | C = c, Y_0 = y) \). Under assumptions 1-3

\[
\Psi = \frac{1}{P(A = 1)} E \left[ (1 - A) \frac{q(C, Y)}{1 - q(C, Y)} Y \right]. \tag{1.2}
\]

This expression for \( \Psi \) does not rely on \( L \), but does require identification of \( q \) which is not directly observable. Therefore, the above expression cannot be computed without additional assumptions because it conditions on \( Y_0 \) which is only partially observed. The quantity \( q(C, Y) \) shall be referred to as the extended propensity score (EPS), and generalizes the standard propensity score to a setting...
with unobserved confounders. Unlike the standard propensity score, which typically only conditions on fully observed pretreatment confounders, the extended propensity score conditions on a subset of confounders and depends on the potential outcome \( Y_0 \). This reflects the fact that while \( Y_0 \) is independent of \( A \) given \( L \) and \( C \), \( Y_0 \) is not independent of \( A \) given \( C \) only. While \( Y_0 \) is conceptually a pretreatment variable, it is not observed for treated persons. Note that equation (1.2) is analogous to equation (1.1), with \( Y_0 \) substituted in for \( L \) to account for confounding.

The extended propensity score has been used previously for sensitivity analysis for unmeasured confounding or non-ignorable nonresponse methods in Robins et al [45] and Vansteelandt et al [63] as well as the methods in Liu et al [28] in that \( Y_0 \) can be captured by a known sensitivity parameter. Although the identifying assumptions are different for the first two papers in that the authors assume the dependence on \( Y \) is known. In Liu et al [28], the authors use instrumental variable assumptions to identify \( q \).

1.4 Main/Validation Study Design for Identification

1.4.1 ETT Identification Conditions

Both treated and untreated subjects are needed in order to estimate \( q(C,Y) \). However, because the extended propensity score uses \( Y_0 \), we cannot observe it for every subject. To make progress, suppose that in addition to the main study we have a validation sample in which \((A,W,Y)\) are observed. Let \( R \) denote an indicator of being in the validation sample (\( R = 0 \) if in the main study sample and \( R = 1 \) if in the validation sample). We make the following additional assumptions:

4. \( P(R = 1|A,C,Y,L) > 0 \) almost surely;

5. Conditional exchangeability: \( L \perp R|A,C,Y \).

Assumption 4 is a positivity assumption for \( R \) and states that there is a positive probability of observing any possible value of \((A,W,Y)\) in the validation sample. Assumption 5 is an MAR assumption in that it states that the conditional distribution of \( L \) given \( A,C,Y \) is identical in the main and the validation samples.

These additional assumptions allow us to identify the extended propensity score from the observed data using the following lemma.
Lemma:
Under assumptions 1-5, \( q(C, Y) \) is non-parametrically identified by
\[
\frac{q(C, Y)}{1 - q(C, Y)} = E \left[ \frac{P(A = 1|W)}{P(A = 0|W)} \right] \middle| R = 1, Y, A = 0, C. \tag{1.3}
\]

Therefore it follows that, under our assumptions, the extended propensity score, \( q(C, Y) \), can be recovered as a function of the observed data.

1.4.2 Estimation and Inference

According to (1.3), EPS requires a model for the density of the missing potential confounders \( L \) conditional on \( C, Y, A = 0, R = 1 \) and a model for the propensity score, \( P(A = 1|W) \). Let \( t(L|C, Y, A = 0, R = 1; \eta) \) denote a model for the conditional density of \( L|C, Y, A = 0, R = 1 \) with unknown parameter \( \eta \). For instance, in the simple case where \( t \) is univariate normal with mean \( \mu_L \) and variance \( \sigma^2_L \). One may take
\[
\mu_L = E[L|A = 0, C, Y, R = 1; \eta] = \eta_0 + \eta_1 C + \eta_2 Y. \tag{1.4}
\]
Efficient estimators, \( \hat{\eta} \) and \( \hat{\sigma}^2_L \), can be found using maximum likelihood in the validation sample by regressing \( L \) on \( C, Y \) and \( A \).

When \( L \) is multivariate it often may be difficult to specify a multivariate model for the joint density of all missing potential confounders. Instead it is convenient to factorize the joint density into a series of conditional densities. For example, suppose \( L = (L_1, L_2, L_3) \), with \( L_1 \) and \( L_2 \) binary and \( L_3 \) normal, we might use the following models
\[
\text{logit} \left[ Pr(L_1 = 1|A = 0, C, Y, R = 1; \eta_1) = \eta_{10} + \eta_{11} C + \eta_{12} Y, \right. \tag{1.5}
\]
\[
\text{logit} \left[ Pr(L_2 = 1|A = 0, C, Y, L_1, R = 1; \eta_2) = \eta_{20} + \eta_{21} C + \eta_{22} Y + \eta_{23} L_1, \right. \tag{1.6}
\]
\[
\mu_3(C, Y, L_1, L_2) = \mu_{30} + \mu_{31} C + \mu_{32} Y + \mu_{33} L_1 + \mu_{34} L_2, \tag{1.7}
\]
where \( L_3|A = 0, C, Y, L_1, L_2, R = 1 \sim N(\mu_3(C, Y, L_1, L_2), \sigma^2_3) \).

Estimates for \( \eta = (\eta_1, \eta_2, \eta_3) \), \( \hat{\eta} \), can be found using maximum likelihood in the validation sample and the above estimated models can be combined to recover an estimate of their joint distribution.

We denote the standard propensity score as \( p(W; \lambda) = Pr(A = 1|W; \lambda) \), with unknown param-
eter $\Lambda$ and corresponding individual score function
\[
S(\Lambda) = \frac{d}{d\Lambda} \log \left[ p(W; \Lambda)^A (1 - p(W; \Lambda))^{1-A} \right].
\]

Then an unbiased estimating equation for $\hat{\Lambda}$ can be obtained from the expected score equation
\[
P_n \left[ \tilde{S}(\hat{\Lambda}, \hat{\eta}) \right] = 0,
\]
where $P_n(.) = \frac{1}{n} \sum_i(.)_i$ and $\tilde{S}(\hat{\Lambda}, \hat{\eta}) = (1 - R_i) E \left[ S(\hat{\Lambda})|C, Y, A, R = 1; \hat{\eta} \right] + R_i S(\hat{\Lambda})$.

For example in the case of a standard logistic model where
\[
\logit[P(A = 1|L, C; \Lambda)] = \lambda_0 + \lambda_1 L + \lambda_2 C,
\]
the corresponding score function is then
\[
S(\Lambda) = \begin{pmatrix} 1 \\ L \\ C \end{pmatrix} (A - P(A = 1|W; \Lambda)).
\]

The extended propensity score is found by evaluating
\[
\frac{q(C,Y; \hat{\eta}, \hat{\Lambda})}{1 - q(C,Y; \hat{\eta}, \hat{\Lambda})} = E \left[ \frac{p(W; \hat{\Lambda})}{1 - p(W; \hat{\Lambda})} | C, Y, A = 0, R = 1; \hat{\eta} \right].
\]

Under the model given by equations 4 and 9 where $L$ is univariate normal, we obtain
\[
\logit[\hat{q}(C,Y_0)] = \hat{\lambda}_0 + \log \left( E \left[ \exp \left( \hat{\lambda}_1 (L - \hat{\mu}_L)|C, Y = y, A = 0, R = 1; \hat{\eta} \right) \right] \right) + \hat{\lambda}_2 C + \hat{\lambda}_1 \hat{\mu}_L(a,y)
\]
\[
= \hat{\lambda}_0^* + \hat{\lambda}_1^* C + \hat{\lambda}_2^* Y,
\]
where $\hat{\lambda}_0^* = \hat{\lambda}_0 + \frac{1}{2} \hat{\sigma}_L^2 \hat{\lambda}_1^2 + \hat{\lambda}_1 \hat{\eta}_0$, $\hat{\lambda}_1^* = \hat{\lambda}_2 + \hat{\lambda}_1 \hat{\eta}_1$, and $\hat{\lambda}_2^* = \hat{\lambda}_1 \hat{\eta}_2$.

This produces a simple closed-form estimator of the extended propensity score, $\hat{q}(C,Y)$, as a function of the observed data, from which we can obtain the EPS estimator for $\Psi$, $\hat{\Psi} = \left(1/\hat{P}(A = 1)\right) P_n[Y_i(1 - A_i)\hat{q}(C_i, Y_i)/(1 - \hat{q}(C_i, Y_i))]$. In general, one would use Monte Carlo expectations in (1.8) and (1.10) to evaluate EPS.

The asymptotic distribution of the estimator can be found as follows. Let $U(\hat{\eta})$ be an individual contribution to the score for $\eta$, $\tilde{S}(\hat{\Lambda}, \hat{\eta})$ an individual contribution to the estimated observed score equation for $\Lambda$ (from above), and $Z(\hat{\Psi}, \hat{\eta}, \hat{\Lambda}) = I(A_i = 0)(Y_i - \hat{\Psi}) [\hat{q}(C_i, Y_i)/(1 - \hat{q}(C_i, Y_i))]$ an
individual contribution to the estimating equation for $\Psi$. Let $\hat{\Phi} = \left(\hat{\eta}, \hat{\lambda}, \hat{\Psi} \right)$ and define

$$V(\hat{\Phi}) = \begin{pmatrix} U(\hat{\eta}) \\ S(\hat{\lambda}, \hat{\eta}) \\ Z(\hat{\Psi}, \hat{\eta}, \hat{\lambda}) \end{pmatrix}.$$ 

Then, under standard regularity conditions,

$$\sqrt{n}(\hat{\Phi} - \Phi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{dV}{d\Phi} \right]^{-1} V(\Phi) + op(1).$$

Therefore, a consistent estimator of the asymptotic variance is

$$\left[ \mathbb{P}_n \frac{dV}{d\Phi} \right]^{-1} \mathbb{P}_n \left[ V(\hat{\Phi})V'(\hat{\Phi}) \right] \left[ \mathbb{P}_n \left( \frac{dV}{d\Phi'} \right) \right]^{-1}.$$ 

Alternatively, we recommend using the nonparametric bootstrap to obtain estimates of the variance.

### 1.4.3 Simulation Study

In this section, we report two simulation studies to compare the finite sample performance of our EPS estimator to a number of existing methods described above.

In the both simulations, $C$ was Bernoulli with $Pr(C = 1) = p_C = 0.45$. The treatment, $A$, was Bernoulli with $Pr(A = 1|C) = p_A = \omega_0 + \omega_1 C$. For this simulation we chose $(\omega_0, \omega_1) = (0.40, 0.54)$. The outcome, $Y$, was chosen to be normal conditional on $A$ and $C$, with $Y = \beta_0 + \beta_1 A + \beta_2 C + \epsilon_Y$ where $\epsilon_Y \sim N(0, \sigma_Y^2)$, $(\beta_0, \beta_1, \beta_2, \sigma_Y^2) = (0.2, 0.38, 0.3, 0.51)$. Similarly, $L$ was chosen to be normal conditional on $A$ and $C$, with $L = \alpha_0 + \alpha_1 A + \alpha_2 C + \epsilon_L$ where $\epsilon_L \sim N(0, \sigma_L^2)$ and such that $Cov(\epsilon_Y, \epsilon_L) = \sigma_{YL}$ and $(\alpha_0, \alpha_1, \alpha_2, \sigma_Y^2, \sigma_{YL}) = (-0.15, 0.215, 0.14, 0.43, 0.21)$. As a consequence, the distribution of $L$ given $C$, $Y$ and $A = 0$ follows equation (4) with $(\eta_0, \eta_1, \eta_2) = (-0.23, 0.01, 0.41)$ and equation (9) holds for the propensity score with $(\lambda_0, \lambda_1, \lambda_2) = (-0.42, 0.5, 0.36)$. For the first simulation, $R$ was Bernoulli with $Pr(R = 1) = p_R = 0.30$ in order to simulate data under MCAR. For the second simulation, in order to examine the MAR scenario, $R$ was Bernoulli with logit $[Pr(R = 0|A, C, Y)] = \xi_0 + \xi_1 A + \xi_2 C + \xi_3 Y$ where $(\xi_0, \xi_1, \xi_2, \xi_3) = (0.7, 0.46, -0.41, 1.2)$. On average, $P(R = 1) \approx 0.30$. The observed data is given by i.i.d. realizations of $(R, RL, Y, A, C)$ with sample size equal to 2,500.
In simulations, in order to gauge the loss of efficiency due to missing data, we computed the full data estimator (Full) of $\Psi$ using the complete data $(A, Y, L, C)$ based on equation (1.1). We also implemented PSC, TSC and BayesPS as well as standard IPCW estimation, MICE (using predictive mean matching), complete-case analysis (CC), and a naive estimator (Naive) that drops the missing confounder $L$ completely and evaluates (1.1) upon substituting an estimate of $P(A = 1|C)$ for $P(A = 1|C, L)$.

PSC requires models for the “gold standard” propensity score,

$$p(C, L; \Lambda) = Pr(A = 1|C, L; \Lambda)$$

and “error-prone” propensity score,

$$\tilde{p}(C; \tilde{\Lambda}) = Pr(A = 1|C; \tilde{\Lambda})$$

as well as for the relationship between the two propensity scores and an outcome regression. For the propensity scores, we used a logistic regression on $A$ with main effects for $(L, C)$ and $C$ respectively. Following Stürmer et al. (2005) [57], to relate the two propensity scores we used a linear regression for the gold standard propensity score with main effects for the error-prone propensity score and $A$. For the outcome we used a main effects only linear regression on the gold standard propensity score and $A$. For TSC we used the same models for the gold standard and error-prone propensity scores as in PSC, $p(C, L; \Lambda)$ and $\tilde{p}(C; \tilde{\Lambda})$. Additionally, the method required main effect outcome regression models on $A$ and the gold standard propensity score as well as on $A$ and the error-prone propensity score. BayesPS required models for the gold standard propensity score, $p(C, L; \Lambda)$, as well as an outcome regression. To implement the IPCW estimator, we had to specify a logistic regression for $R$ with main effects for $(A, Y, C)$ as well as a logistic model for $A$ on $(C, L), p(C, L; \Lambda)$. MICE used a main effects model for $L$ as a function of $(A, C, Y)$ as well as a logistic regression for $A$ with main effects for $L$ and $C$, $p(C, L; \Lambda)$. The complete-case estimator only required a logistic regression for $A$ with main effects for $L$ and $C$, $p(C, L; \Lambda)$. Finally the Naive estimator required a logistic regression for $A$ with main effects for $C$ alone, $\tilde{p}(C; \tilde{\Lambda})$. All these methods were compared to the proposed EPS estimator. For the CC, Naive, MICE, IPCW, and Full estimators, we calculated ETT for each method using equation (1.1) for $\Psi$. For the naive estimator the odds, $p(C, L; \Lambda)/1 - p(C, L; \Lambda)$,
were replaced with $\tilde{p}(C; \tilde{A}) / 1 - \tilde{p}(C; \tilde{A})$ and for the IPCW estimator the odds were estimated with inverse probability weighting. For the estimators using PSC, BayesPS, and TSC, we calculated the average causal effect (ACE) rather than ETT. In this simulation there are no interactions with treatment on the additive scale, and therefore ACE and ETT coincide.

Figure 1.1 summarizes the results in the form of Monte Carlo average point estimates and empirical 95% confidence intervals for the estimated population Effect of Treatment on the Treated for 250 MC samples of 2500 subjects.

We can see from Figure 1.1a that under MCAR, many of the estimators perform similarly. The EPS estimator showed better efficiency than IPCW, TSC and CC, but not markedly better than BayesPS or MICE. The surrogacy assumption required by PSC was not met, which is the likely cause of the bias seen for that estimator. The naive estimator also was biased, which was to be expected.

Figure 1.1b illustrates the performance of the estimators when the probability of being in the main sample is a function of $(Y, A, C)$. EPS clearly outperforms all other estimators and performs remarkably well compared with the full data estimator (Full). Inverse probability weighting, PSC and multiple imputation cover the true parameter value. The complete-case, naive, BayesPS and TSC estimators are all biased. This bias is likely a result of the simulated data violating the MCAR
assumption that these methods require. Additionally, the surrogacy assumption required by PSC was not met.

When it is reasonable to assume that the missingness mechanism is MCAR, such as when the validation study is conducted on a random sample from the main study, most of the methods discussed will be valid. This includes the complete-case estimator, which is very simple to calculate. However, even under the MCAR assumption, IPCW may be more efficient. We see this in Figure 1.1b, which agrees with theory from Robins et al. (1994) [46]. In general, when the MCAR assumption is in doubt, and only MAR can be safely assumed, EPS will be consistent whereas PSC, TSC, and BayesPS break down.

1.5 Main Study-only ETT

1.5.1 Identification and Inference

We note that all of the previous methods assume that there is a larger well-defined population with joint density \( f(Y, A, W) = f(Y, A, W|R = 0) P(R = 0) + f(Y, A, W|R = 1) P(R = 1) \), and the target ETT parameter of interest is evaluated under this population law. An example would be when the validation study is internal and thus from the same population as the main study, by design. However, when the main and validation study populations are different, it may be of greater clinical interest to estimate ETT only in the main study population, since the main study is considered to be a random sample from the overall population of interest. This situation is more likely to occur when the validation sample is taken from an external source. EPS is advantageous in this situation, since it uses the validation study to calculate the extended propensity score for each subject, but calculating ETT can be limited only to the main study.

ETT in the main study on the additive scale is defined as

\[
E[Y_{a=1} - Y_{a=0}|A = 1, R = 0] = \theta^* - \Psi^*
\]

where \( \theta^* = E[Y_{a=1}|A = 1, R = 0] \) and \( \Psi^* = E[Y_{a=0}|A = 1, R = 0] \).

Under assumption 1, it is straightforward to calculate

\[
\theta^* = E[Y_{a=1}|A = 1, R = 0] = E[Y|A = 1, R = 0].
\]
Identification of $\Psi^*$ requires a modified version of assumption 2. Namely we require:

6. $A \perp Y_0|W, R = 0$;

7. Positivity: $\frac{P(A=0|W)}{P(A=1|W)} > 0$ almost surely.

Assumptions 6 and 7 are weaker than assumptions 2 and 3 which assume no unmeasured confounding and positivity in the larger pooled population while 6 and 7 only require the assumptions to hold in the main study.

Under assumptions 1,3 and 6, $\Psi^*$ can be written as

$$\Psi^* = E[Y_0|A = 1, R = 0] = \frac{1}{P(A = 1|R = 0)} E \left[ I(A = 0, R = 0) \frac{P(A = 1|L, C, R = 0)}{P(A = 0|L, C, R = 0)} Y \right].$$

Now consider the same missing data setting as in section 3. As before, we wish to establish a representation of $\Psi^*$ in this setting that does not depend on $L$.

This representation is given in the following result.

**Result:**

Let $q^*(c, y) = P(A = 1|C = c, Y_0 = y, R = 0)$ for all $c$ and $y$, then under assumptions 1,3 and 6,

$$\Psi^* = \frac{1}{P(A = 1|R = 0)} E \left[ I(A = 0, R = 0) \frac{q^*(C, Y)}{1 - q^*(C, Y)} Y \right].$$

This expression for $\Psi^*$ does not rely on $L$, but does require identification of $q^*$ which is not directly observable.

The addition of assumption 6 also allow us to identify $q^*$ from the observed data using the following lemma:

**Lemma:**

Under assumptions 1 and 3-6, $q^*(C, Y)$ is non-parametrically identified by

$$\frac{q^*(C, Y)}{1 - q^*(C, Y)} = E \left[ \frac{P(A = 1|L, C, R = 0)}{P(A = 0|L, C, R = 0)} | R = 1, Y, A = 0, C \right].$$

Therefore it follows that, under our assumptions, the modified extended propensity score, $q^*(C, Y)$, can be recovered as a function of the observed data.

To estimate the main study standard propensity score, $P(A = 1|L, C, R = 0)$, we solve the score
equation for the propensity score in the main study using assumption 5.

Denote the propensity score model as 
\[
p(W; \tilde{A}) = \Pr(A = 1 | W, R = 0; \tilde{A}),
\] 
with corresponding individual score function
\[
S(\tilde{A}) = \frac{d}{d\tilde{A}} \log \left[ p(W; \tilde{A})^A (1 - p(W; \tilde{A}))^{1-A} \right].
\]
Then an unbiased estimating equation for \( \hat{\tilde{A}} \) can be obtained from the expected score equation
\[
\mathbb{E}_n \left[ \hat{S}(\hat{\tilde{A}}, \hat{\eta}) \right] = 0,
\]
where \( \hat{S}(\hat{\tilde{A}}, \hat{\eta}) = (1 - R_i)E \left[ S_A(A, W; \tilde{A}) | C, Y, A, R = 1; \hat{\eta} \right] \) and \( \hat{\eta} \) are found as in section 4.2. An empirical evaluation of (1.13) parallels previous sections, and large sample inference can likewise be derived as in previous sections.

### 1.5.2 Simulation Study

We performed a simulation study in which the main study data were generated as described in section 4.3. However, the validation data was generated using the following modified mechanism.

In the validation data, we chose \( C \) to be Bernoulli with \( \Pr(C = 1) = p_C = 0.75 \). This choice is in contrast to the main study data, for which we set \( \Pr(C = 1) = p_C = 0.45 \). In the validation data, we chose \( A \) to be Bernoulli with \( \Pr(A = 1) = p_A = 0.65 \). In the main study, \( A \) was generated as a

![Comparison of Estimators](image)

Figure 1.2: Results of simulation study using \( \Psi^* \)
function of $C$. In the validation data, the outcome, $Y$, was generated $Normal(\mu_L = 2, \sigma_L = 1.7)$. By contrast, in the main study $Y$ was generated as a function of $A$ and $C$. Then, to preserve assumption 5, $L$ was generated in the validation data so as to follow the same distribution as $L|A, Y, C, R = 0$ in the main study. Figure 1.2 displays results for the various estimators. The PSC estimator was $-0.867$ (-26.180, 24.445).

We found that the amount of bias increases for all the previous estimators, including EPS. However, the EPS_Main estimator, which implements the above method, performs the best if ETT in the main study is the target parameter of interest as would likely be the case when missing data are not MCAR. This simulation was performed under an MAR missingness mechanism, however under MCAR there is no distinction between ETT and Main Study ETT. Therefore under MCAR we would expect results similar to the MCAR simulation in Section 4.

1.6 Data Application

1.6.1 Study Populations

To illustrate the proposed method, we studied the effect of surgical resection of brain tumors on log survival time among patients with malignant neoplasm of the brain. Specifically, we considered a large Medicare population with data on survival time, and on whether a patient had surgical resection or biopsy alone, and with numerous, though incomplete, potential confounding variables. We supplemented this sample with a smaller validation sample from the SEER-Medicare program of cancer registries that contains richer, more complete, cancer-specific information, listed in Table 1.1.

Main study sample. Our main study population comes from Medicare and consisted of 40,776 patients who were over the age of 64 with a diagnosis of malignant neoplasm of brain (MNB) between the years 2000 and 2007, who represented a ICD-9 principal discharge diagnosis of 191.xx, who were continuous in fee for service (FFS) from diagnosis to death, who were enrolled in a managed care organization (MCO) continuously for at least 12-months prior to the index MNB diagnosis and whose tumors were first and malignant. We further limited the population to those who had either surgical resection of their tumor or biopsy.

All subjects had information on 28 potential confounders listed in Table 1.1. The outcome of
interest was the log of survival time, and the treatment was surgical resection of the tumor [35, 64].

Validation sample. Our validation study sample comes from the linked Surveillance, Epidemiology and End-Results (SEER)-Medicare data and consists of 5,474 patients meeting the same criteria as the main study. Subjects in SEER-Medicare had to meet two inclusion criteria: MNB was reported/confirmed during a hospitalization, and all tumors were first and malignant. Further, patients missing the month of MNB diagnosis were excluded. In addition to variables shared with the main study subjects, SEER-Medicare subjects had information available on the following potential confounders that may provide additional insight into demographics as well as the nature of the cancer: number of tumors (dichotomized to one/more than one), tumor location (dichotomized to Supratentorial/other), tumor size (dichotomized to greater than 3 inches/other), marital status (married/other), income group (low/high), cancer stage (dichotomized to localized/other), and an indicator for glioblastoma multiforme (GBM) [64, 35]. We prescreened all binary variables to ensure they were present in at least 0.1% of the subjects in both the main and validation samples.

Data Linkage. Linkage of SEER-Medicare and Medicare datasets was performed by the National Cancer Institute (NCI) and the Centers for Medicare and Medicaid Services (CMS) with collaboration from the SEER-Medicare registries. The linkage is based on a matching algorithm that uses individual identifiers for all persons. The linkage is described in more detail elsewhere [64, 36].

It is unclear if patients in the validation data are also present in the main study dataset. Therefore, to prevent potential overlap, we examined only even numbered years for the main study and odd numbered years for the validation study. We believe this eliminates any overlap as survival times were sufficiently short. In order to make the two samples as comparable as possible, we further restricted both datasets to states that were present in both the Medicare and SEER-Medicare samples. These restrictions yielded 14,312 and 2,391 patients in the main study and validation study respectively.

Table 1.1 compares the two populations. The age and sex distributions are similar for the two samples, while the SEER-Medicare subjects are more diverse racially. However, the distributions of the other comorbidities reflect the fact that the subjects from the SEER-Medicare validation study are healthier overall than those in Medicare. Moreover, more subjects in the validation study elected to have surgical resection (60.5%) versus those in the main study (42.9%). It is well known that the SEER-Medicare population is not marginally exchangeable with the Medicare population.
residing outside of SEER-Medicare areas [64, 14]. SEER-Medicare areas tend to be more affluent and more urban than non-SEER-Medicare areas [64, 14]. SEER-Medicare areas also tend to have a lower proportion of white residents than non-SEER-Medicare areas. There were no censored observations.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Medicare</th>
<th>SEER Medicare</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>14312</td>
<td>2391</td>
</tr>
<tr>
<td>Survival Time in Days</td>
<td>141.22 (120.94)</td>
<td>283.79 (346.96)</td>
</tr>
<tr>
<td>Resection</td>
<td>8988 (62.8)</td>
<td>1676 (70.1)</td>
</tr>
<tr>
<td>Age</td>
<td>75.09 (6.34)</td>
<td>74.03 (6.07)</td>
</tr>
<tr>
<td>Female</td>
<td>6583 (46.0)</td>
<td>1145 (47.9)</td>
</tr>
<tr>
<td>White</td>
<td>13289 (92.9)</td>
<td>2108 (88.2)</td>
</tr>
<tr>
<td>Brain MRI</td>
<td>1379 (9.6)</td>
<td>133 (5.6)</td>
</tr>
<tr>
<td>Head CT scan</td>
<td>1355 (9.5)</td>
<td>359 (15.0)</td>
</tr>
<tr>
<td>Chronic Atherosclerosis</td>
<td>998 (7.0)</td>
<td>197 (8.2)</td>
</tr>
<tr>
<td>Substance Abuse</td>
<td>3142 (22.0)</td>
<td>362 (15.1)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1177 (8.2)</td>
<td>154 (6.4)</td>
</tr>
<tr>
<td>COPD</td>
<td>1813 (12.7)</td>
<td>218 (9.1)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>622 (4.3)</td>
<td>43 (1.8)</td>
</tr>
<tr>
<td>Functional Disability</td>
<td>1806 (5.6)</td>
<td>8 (0.3)</td>
</tr>
<tr>
<td>Seizure Disorder</td>
<td>2764 (19.3)</td>
<td>171 (7.2)</td>
</tr>
<tr>
<td>Trauma</td>
<td>655 (4.6)</td>
<td>66 (2.8)</td>
</tr>
<tr>
<td>Depression</td>
<td>1097 (7.7)</td>
<td>147 (6.1)</td>
</tr>
<tr>
<td>Cancer</td>
<td>2969 (20.9)</td>
<td>503 (21.6)</td>
</tr>
<tr>
<td>Major Psychiatric Disorder</td>
<td>601 (4.2)</td>
<td>57 (2.4)</td>
</tr>
<tr>
<td>Asthma</td>
<td>443 (3.1)</td>
<td>56 (2.3)</td>
</tr>
<tr>
<td>Hypertensive Heart Disease</td>
<td>252 (1.6)</td>
<td>28 (1.2)</td>
</tr>
<tr>
<td>Parkinson’s or Huntington’s</td>
<td>176 (1.2)</td>
<td>13 (0.5)</td>
</tr>
<tr>
<td>Major Cancer</td>
<td>2432 (17.0)</td>
<td>146 (6.1)</td>
</tr>
<tr>
<td>Chronic Lung Disorders</td>
<td>185 (1.3)</td>
<td>20 (0.8)</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>1097 (7.7)</td>
<td>147 (6.1)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2945 (20.6)</td>
<td>403 (16.9)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>1506 (63.0)</td>
<td>1577 (66.0)</td>
</tr>
<tr>
<td>Metastatic Cancer</td>
<td>303 (12.7)</td>
<td>1808 (75.6)</td>
</tr>
<tr>
<td>Low Income</td>
<td>655 (4.6)</td>
<td>66 (2.8)</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>1097 (7.7)</td>
<td>147 (6.1)</td>
</tr>
<tr>
<td>GBM</td>
<td>1971 (82.4)</td>
<td>218 (9.1)</td>
</tr>
<tr>
<td>Asthma</td>
<td>443 (3.1)</td>
<td>56 (2.3)</td>
</tr>
<tr>
<td>Cancer</td>
<td>252 (1.6)</td>
<td>28 (1.2)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2945 (20.6)</td>
<td>403 (16.9)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>1506 (63.0)</td>
<td>1577 (66.0)</td>
</tr>
<tr>
<td>Asthma</td>
<td>443 (3.1)</td>
<td>56 (2.3)</td>
</tr>
<tr>
<td>Cancer</td>
<td>252 (1.6)</td>
<td>28 (1.2)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2945 (20.6)</td>
<td>403 (16.9)</td>
</tr>
<tr>
<td>Marital Status</td>
<td>1506 (63.0)</td>
<td>1577 (66.0)</td>
</tr>
</tbody>
</table>

Table 1.1: Characteristics of the primary data and validation data. Rows contain frequencies (percentages) for dichotomous variables and means (standard deviation) for ordinal variables.

We implemented the proposed EPS method described in Section 3 to assess the effect of surgical resection vs biopsy (A) on the log of survival time (Y) among those who received surgical resection. For comparison, we also implemented the following methods: IPCW, MICE, PSC, BayesPS, and TSC, as well as complete-case and naive analyses.

Figure 1.3 gives results for the various estimators. The complete-case estimator was 0.163 (-1.11, 1.435). Standard deviations for all estimators aside from BayesPS were captured using 500 weighted bootstrap samples, with weights sampled from an exponential distribution with mean 1 [62]. The BayesPS estimate and credible interval were obtained from 2,500 samples from the posterior distribution.
Results across all estimators seem to agree to some extent and all indicate a positive effect estimate, with EPS and Bayes PS having the shortest confidence intervals and MICE the widest. The complete-case estimator is the most unstable, with a very large variance, which is likely due to the relatively small size of the SEER-Medicare validation sample \( m = 2,391 \) compared to the Medicare main sample \( m = 14,312 \). We note that EPS, IPCW, MICE, complete-case and naive methods estimate ETT; however PSC, BayesPS and TSC estimate the average treatment effect (ATE). These two quantities are only the same when there are no interactions between confounders and treatment on the additive scale. While none of the outcome regressions used in this application included such interactions, we cannot be certain of their absence, which might explain potential differences between estimators.

Furthermore, we acknowledge that the dataset is not as rich as might be desired, despite the additional information provided by the SEER-Medicare validation study. Neither dataset includes information on venous thrombosis, venous thromboembolism, coagulopathy, or platelet dysfunction, which can be important confounding variables. Moreover, the decision process for primary (such as glioblastoma) and secondary (metastatic) tumors differ significantly. The variables considered in (or relevant to) each decision process may differ. For example, the number of tumors is not very relevant to glioma and metastasis tumors are not localized by definition (Peleg Horowitz, PhD, MD, personal communication, September 1, 2016). As such we do not make any clinical recommendations.
Because the SEER-Medicare population is generally healthier, we may also be interested in reporting the ETT in the Medicare main study alone. This estimate (first in Figure 1.3) was found to be 0.391 with 95% confidence interval (0.343, 0.439), which is less than the EPS estimate of 0.430, though not significantly different. Since the SEER-Medicare population is generally healthier, it is expected that the main study ETT would be lower, given the lower survival time in general.

1.7 Discussion

In this chapter, we propose the extended propensity score method to control for unobserved confounding. EPS preserves many properties of standard propensity scores, with the additional advantage that it can be evaluated even for subjects with missing confounders. This approach avoids the problems seen in other methods whenever the validation study is decidedly smaller than the main study. Moreover, EPS is universal as it applies virtually to any outcome scale. EPS does not require some of the restrictive assumptions of other existing methods such as MCAR or surrogacy. However, like TSC and BayesPS, our approach does require that the outcome be observed in the validation study. PSC does not have this requirement, but still relies on the stringent surrogacy assumption. In future work, we plan to extend the EPS method to allow for cases where the outcome is not available in the validation study and when surrogacy cannot be assumed. Care must also be taken to ensure that the main and validation datasets contain \((Y, A, C)\) and that they are similarly defined and measured. Additionally, EPS, like other methods, cannot account for bias due to unmeasured confounders which are not included in the main or validation studies.

An estimator based solely on the propensity score potentially may allow for more sophisticated statistical analysis of data from a distributed research network (DRN), without requiring the transfer of individual-level data. Distributed research networks are networks of organizations that hold clinical or research datasets, such as hospitals or clinics, that share information through a common, centralized infrastructure [60, 56]. Queries are made through the central hub and retrieved by the individual sites. As a result, large amounts of data can be gathered quickly and inexpensively from many sources. DRNs allow data partners to maintain control and analysis of their data, while providing structure to standardize data across sites. For studies that draw from multiple populations, such as those utilizing distributed research networks, it is prudent to have each center report as little information as possible on each subject [60, 56]. The primary methods discussed in this chapter are
based on the propensity score. Each site in the network estimates and reports propensity scores for subjects in their data set allowing investigators to adjust for potential confounders without having direct access to potentially identifiable information [60, 56]. DRNs are an important potential area of application for the proposed methodology.

In simulations we saw that most of the methods performed well under MCAR, except the Naive estimator, which drops $L$ entirely, and PSC, for which surrogacy was not met. Under MAR, TSC, and BayesPS were no longer unbiased. Moreover, if it is unreasonable to make the MCAR assumption, EPS is the only estimator among those considered that recovers the Main Study (or "verify-in-sample") ETT. This is a major advantage of EPS. Future work investigating ETT might include sensitivity analysis, bounds, or other existing methods. For example it would be straightforward to adopt the sensitivity methods of Robins et al (2000) [45] to the extended propensity score.

All of the estimators considered here require very specific assumptions, not just regarding the data structure but also regarding correct specification of a number of models (outcome, propensity score etc). We might expect these assumptions to breakdown in some cases. Therefore it would be important to develop a doubly robust estimator in this setting.
A Coherent Likelihood Parameterization for Doubly Robust Estimation of a Causal Effect with Missing Confounders

with

Eric J. Tchetgen Tchetgen
2.1 Introduction

Confounding bias and missing data are two major analytic challenges in comparative effectiveness research using observational data such as electronic medical records. While each problem has been thoroughly studied separately, consolidated approaches for addressing both issues are lacking. In the absence of missing data, confounding bias must still be adjusted for in order to evaluate causal effects [19]. Researchers often use the g-formula for identifying the distribution of counterfactuals from the observed data distribution [41, 54]. Inverse probability weighting estimators are commonly used and involve modeling the propensity score [47, 41, 18]. Doubly robust (DR) estimators for causal effects have been well established and widely studied [4, 29, 42]. These estimators are doubly robust in the sense that they are consistent and asymptotic normal if either the treatment mechanism (propensity score) or outcome model is correctly specified, but not necessarily both. These methods are also locally semiparametric efficient because they achieve the semiparametric efficiency bound for the nonparametric model when model misspecification is absent, that is, at the intersection submodel of the union of specified models.

Multiple imputation and inverse probability of censoring weighting are increasingly popular methods for addressing missing data [51, 26, 53]. In the context of regression analysis, various weighting schemes to account for missing covariates have been examined previously in the literature [31, 27, 34, 58]. Semiparametric locally efficient methods are also available to address data missing at random (MAR) i.e., the probability of observing the full data depends on the fully observed data only [22, 4]. Robins, Rotnitzky and others examined improved augmented inverse weighted estimators within the semiparametric framework [46, 44, 49, 52]. Additionally, *Tsiatis (2006)* is a textbook that provides an extensive overview of the state of the art for applying semiparametric theory to missing data.

However, to date, few methods have considered joint inferences about causal effects that are doubly or multiply robust, in the sense of *Vansteelandt et al (2007)* [63], in the presence of missing data and confounding. This setting presents a special challenge in that it involves the nesting of causal inference in the missing data setting, each of which requires, to obtain the parameter of interest, estimating a nuisance parameter while appropriately accounting for the fact that nuisance parameters needed to adjust for selection bias are entangled with nuisance parameters needed to
address confounding bias. *Davidian et al. (2005)* [10] presented a DR augmented inverse weighted estimator of the causal effect of exposure when the outcome was missing. In their 2003 textbook, Robins and van der Laan give a unified theory for addressing causal inference in the presence of missing data but do not address specific challenges with identifying an appropriate parametrization for the observed data when addressing both confounding adjustment and incomplete confounder data [61]. This chapter addresses a special case of that general theory.

*Williamson et al. [66]* attempt to combine existing methods in order to create a multiply robust (MR) estimator. The authors consider a union of four semiparametric models each of which specifies parametric working models for either the missingness mechanism or the missing covariates to account for missing data, and for either the treatment mechanism or the outcome to account for confounding. MR estimation requires that each submodel of the union model is a semiparametric model in that the correctly specified part is parametric, while the remaining submodels are unrestricted. However, we will show in this chapter that the rest of the likelihood is, in fact, restricted in at least one submodel of the union model and therefore the MR property claimed by *Williamson et al.* may not be achievable in reality. An immediate implication of this phenomenon is that in addition to possible lack of compatibility across submodels of the union model, the intersection submodel of the union model may in fact be empty. Therefore, unless one explicitly acknowledges the overlap between components of the union model in the process of model specification, one may in fact rule out the possibility of achieving local efficiency.

In this chapter we discuss the difficulty of achieving double robustness in semiparametric missing data when full data nuisance parameters are entangled with nuisance parameters needed to account for data missing at random. We carefully examine the previously suggested MR method and explain why it may fail to achieve the claimed MR property. We then propose a solution that carefully identifies the modeling assumptions through an alternative transparent parametrization of the likelihood function, which makes explicit model dependencies between various nuisance functions needed to evaluate the MR estimating equation for the causal effect of interest. The proposed method is genuinely doubly-robust (DR) in that it is consistent and asymptotically normal if one of two sets of modeling assumptions holds, and we establish that this may be the most robustness to model misspecification one can achieve. This chapter suggests an approach that could easily be adopted in other settings where one may wish to obtain a DR estimator in the presence of entangled
nuisance parameters. While the chapter focuses on the effect of treatment on the treated (ETT), the proposed approach equally applies to the average causal effect (ACE).

In section 2 we review existing results on identification and inference about ETT in the absence and presence of missing data. In section 3 we develop a new doubly robust method by using an alternative coherent parametrization of the likelihood function in terms of nuisance parameters needed for inference. In section 4 we discuss implementation. In section 5 we compare the new approach to several existing methods in an extensive simulation study and evaluate the finite sample performance of the proposed double robust (DR) approach. We conclude in section 6 with a discussion. Proofs and additional derivations referenced throughout the text can be found in the appendix.

2.2 Preliminaries

2.2.1 Full Data Setting

Let $A$ denote a binary treatment, $A \in \{0, 1\}$, and let $Y$ be the outcome in view with $Y_1$ and $Y_0$ denoting the potential outcomes under treatment and control conditions respectively. Let $W$ denote a set of pre-treatment covariates. The parameter of interest is the effect of treatment on the treated (ETT) on the additive scale, defined as $E[Y_{a=1} - Y_{a=0}|A = 1] = \theta - \Psi$ where $\theta = E[Y_{a=1}|A = 1]$ and $\Psi = E[Y_{a=0}|A = 1]$.

Throughout, we make the following standard causal assumptions in order to identify ETT:

1. Consistency: $Y = Y_A$ almost surely;
2. No unmeasured confounding: $A \perp Y_0|W$;
3. Positivity: $\frac{P(A=0|W)}{P(A=1|W)} > 0$ almost surely.

Assumption 1 states that a person’s observed outcome corresponds to her potential outcome for the observed treatment. Assumption 2 states that the treatment assignment is ignorable conditional on covariates $W$, i.e. $W$ includes all common causes of $A$ and $Y$. And assumption 3 states that there is no treated subject without an untreated counterpart.
Under assumption 1, \( \theta = E[Y|A = 1] \) \([2, 23]\). Under assumptions 1-3, \( \Psi \) is well known to be non-parametrically identified and is given by

\[
\Psi = E[Y_{a=0}|A = 1]
\]

\[
= \frac{1}{P(A = 1)} E \left[ (1 - A) \frac{P(A = 1|W)}{P(A = 0|W)} Y \right].
\]  

(2.1)

The following dual representation of (1.1) is also of interest

\[
\Psi = E \left[ E[Y|A = 0, W]|A = 1 \right]
\]

\[
= \int f(w|A = 1) \int y f(Y|A = 0, w) dy dw.
\]

Note that any regular and asymptotically linear estimator \( \hat{\Psi} \) of \( \Psi \) satisfies

\[
\sqrt{n} \left( \hat{\Psi} - \Psi \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \iota_i (A_i, W_i, Y_i; \Psi) + o_p(1)
\]

where \( \iota_i (A_i, W_i, Y_i; \Psi) \) is a zero-mean function, called the \( i^{th} \) influence function for \( \Psi \). The influence function characterizes the behavior of the estimator (such as the asymptotic distribution) and under certain condition, may also be used to define an estimating equation to obtain an estimator with the corresponding influence function. For functionals defined on nonparametric models, as will be considered in this chapter, there exists a unique influence function which is semiparametric efficient under the nonparametric model. Influence functions were first introduced by Huber [20] in the context of robust statistics and later developed in semiparametric theory, in the sense of Bickel et al. [5]. In either context, influence functions represent the influence of a single observation on the estimator.

The efficient influence function of \( \Psi \) in the nonparametric model in which assumptions 1-3 hold, but the observed data likelihood is otherwise unrestricted, is given by [16],

\[
\iota_{full}(\Psi) = \frac{I(A = 0)}{P(A = 1)} \frac{f(A = 1|W)}{f(A = 0|W)} (Y - E[Y|A = 0, W]) + \frac{I(A = 1)}{P(A = 1)} (E[Y|A = 0, W] - \Psi).
\]  

(2.2)

In order to use the efficient influence function as an estimating function for \( \Psi \) when, as is typically the case in observational studies, \( W \) is high dimensional, one must estimate the nuisance functions \( f(A|W) \) and \( E[Y|A = 0, W] \) using low dimensional parametric models. The solution to the resulting estimating equation is doubly robust for \( \Psi \) in that it is consistent provided we consistently estimate
the propensity score, \( f(A|W) \), or the outcome model, \( f(Y|A = 0, W) \), but not necessarily both. Additionally, the estimator achieves the semiparametric efficiency bound in the absence of model misspecification.

### 2.2.2 Missing Data Setting

Next, suppose that only a subset of covariates, \( C \), of \( W \) is fully observed while \( L \) is missing for a subset of participants where \( W = \{L, C\} \). Therefore the observed data can be written as \((Y, A, C, RL, R)\), where \( R \) is an indicator function which is equal to 1 when \( L \) is observed and is otherwise equal to 0. Define \( O = (Y, A, C) \), the fully observed data. Furthermore, suppose that \( L \) is missing at random (MAR). The data now presents a non-monotone missingness pattern with respect to missing confounders and missing counterfactual outcomes.

In order to address missing data, we make the following additional assumptions:

4. \( \pi = P(R = 1|A, L, C, Y) > 0 \) almost surely;
5. Conditional exchangeability: \( L \perp R|A, C, Y \).

Assumption 4 is a positivity assumption and states that there is a positive probability of observing any possible value of \((A, C, L, Y)\) in the complete cases. Assumption 5 is an MAR assumption and states that the conditional distribution of \( L \) given \( A, C, Y \) is the same in incomplete and complete cases.

Under assumptions 1-5, the efficient influence function of \( \Psi \) in the nonparametric model in which the observed data distribution is unrestricted is given by

\[
\tau_{\text{Miss}}(\Psi) = \frac{R}{\pi} \tau_{\text{Full}}(\Psi) - \left( \frac{R}{\pi} - 1 \right) E[\tau_{\text{Full}}(\Psi)|O].
\]  

The efficient influence function in equation (2.3) depends on the following functions: the propensity score, \( p = P(A = 1|L, C) \), the outcome model, \( m = m(Y|A, L, C) \), the missing data mechanism, \( \pi = P(R = 1|A, Y, C) \), and the density of \( L \) given \( A, C \) and \( Y \), \( t = t(L|A, C, Y) \).

The efficient influence function is appealing as a basis for obtaining inferences about \( \Psi \), mainly because of the following multiple robust property:

\[
E[\tau_{\text{Miss}}(p, m, \pi, t; \Psi)] = 0
\]  

\[2.4\]
if $\Psi$ is evaluated at the truth, and one of the following statements hold:

(i) $p$ and $\pi$ are evaluated at the truth;

(ii) $m$ and $\pi$ are evaluated at the truth;

(iii) $p$ and $t$ are evaluated at the truth;

(iv) $m$ and $t$ are evaluated at the truth.

Additionally, at the intersection submodel where all of the models are evaluated at the truth, the variance of $\ell_{Miss}(\Psi)$ achieves the semiparametric efficiency bound for the union of models (i)-(iv) at the intersection submodel.

A closely related multiply robust property of the efficient influence function to account for missing confounders was established for ACE by Williamson et al (Statistics in Medicine 2012) [66].

However, because in practice one must estimate $p$, $m$, $\pi$, and $t$ under corresponding low dimensional working model, model incompatibility may render the MR property given above infeasible, as we show next. The approach considers four submodels, $p$, $m$, $\pi$, and $t$ and four unions of those submodels. These submodels are semiparametric in the sense that in practice, within each submodel of the union model, two models are parametrically specified, but the remaining two are not modeled and left unrestricted. However, this cannot hold as the specified models in at least one of the submodels of the union model places restrictions on components of the likelihood not explicitly modeled in the submodel. Because each component of the likelihood is eventually modeled separately, conflict may arise in two separate models for the same component of the likelihood, therefore ruling out the possibility for MR and local efficiency. To explain how this potential conflict in model specification may arise, note that the joint likelihood of all four models is $\left\{ f (Y, A, L, C) \right\} R \left\{ \int f (Y, A, l, C) dl \right\}^{1 - R} f (R|Y, A, L, C)$. The submodels $t(L|A, C, Y)$ and $m(Y|A, C, L)$ are not variation independent, as they both encode an association between $Y$ and $L$ given $A$ and $C$. Similarly, the models for $t(L|A, C, Y)$ and $p = P(A = 1|L, C)$ are not variation independent because both densities encode the association between $A$ and $L$ given $C$. Because these various functions are not variation independent, a choice of model for one may restrict modeling options for the other. As a result of the lack of variation independence, MR cannot be achieved under a coherent parametrization of the likelihood which acknowledges the model dependence revealed above. Furthermore, unless one such parametrization can be established, local efficiency may also not be attainable because the intersection submodel may be empty in presence of conflicting mod-
els. The following section provides a carefully crafted coherent parametrization of the observed data likelihood under which a certain degree of double robustness can be achieved and local efficiency remains a genuine possibility.

2.3 Reparametrization

We propose a transparent parametrization of the likelihood function, \( f(L, Y|A, C) \), which makes explicit the model dependencies between nuisance functions that are needed to evaluate the efficient score given by (2.4) in order to obtain an estimator, as described later in section 4. The proposed approach is based on a generalized conditional odds ratio symmetric parametrization of a joint conditional distribution.

Following Chen (2007) [7] and Tchetgen Tchetgen et al. (2009) [59] we define the generalized conditional odds ratio function of \( A \) and \( Y \) given \( L \) as

\[
\chi(A, Y|L) = \frac{f(A|Y, L) f(a_0|y_0, L)}{f(a_0|Y, L) f(A|y_0, L)}
\]

where \((a_0, y_0)\) is a reference value.

Chen (2007) [7] established that the joint distribution of \( A \) and \( Y \) given \( L \) can be written as

\[
\begin{align*}
\int \int \chi(a, y|L) f(a|y_0, L) f(y|a_0, L) \, dy \, da < \infty.
\end{align*}
\]

This parametrization is attractive because \( \chi(A, Y|L) \), \( f(A|y_0, L) \), and \( f(Y|a_0, L) \) are variation independent in that the choice of a parametric model for one component does not restrict available model choices for another and their joint parameter space is the product space of their respective parameter spaces. We repeatedly make use of this result in the appendix to prove the following result.

Result 1:

Let \( f = f(L, Y|A, C) \) be the distribution of \( L \) and \( Y \) given \( A \) and \( C \). \( f \) can be written as

\[
f = \frac{\chi(L, Y|A, C) f(Y|L = 0, A, C)}{K(A, C)} \frac{f(L|A = 0, C) \chi(A, L|C)}{\int \chi(L, y|A, C) f(y|A, C, L = 0) \, dy},
\]

where \( K(A, C) = [f(L = 0|Y = 0, A, C)]^{-1} \int \int \chi(l, y|A, C) f(l|y_0, A, C) f(y|l_0, A, C) \, dy \, dl \) and \( f(L|A = 0, C) \), \( f(Y|L = 0, A, C) \), \( \chi(L, Y|A, C) \), and \( \chi(A, L|C) \) are variation independent. This reparametriza-
tion makes explicit the fact that to model \( f \) we must model \( \chi(L, Y|A, C) \), \( \chi(A, L|C) \), \( f(L|A = 0, C) \), and \( f(Y|L = 0, A, C) \).

Similarly we can parameterize the propensity score, \( p \) as

\[
f(A|L, C) = \frac{\chi(A, L|C)f(A|L = 0, C)}{\tilde{K}(C)},
\]

where \( \tilde{K}(C) = \sum_a P(A = a|L = 0, C)\chi(a, L|C) \). This reparametrization makes explicit the fact that the propensity score can be expressed in terms of \( \chi(A, L|C) \) and \( f(A|L = 0, C) \), which are variation independent.

Therefore, both \( f \) and \( p \) require the same specification of \( \chi(A, L|C) \). Furthermore both \( m \) and \( t \) share the odds ratio of \( L \) and \( Y \) given \( C \) and \( A \). This implies that assuming a submodel for \( m \) also places a restriction on the submodel for \( t \) which cannot remain unrestricted as assumed by Williamson et al in their claim to achieve MR [66].

### 2.4 Implementation

Let \( j = j (L|A = 0, C) \) denote the distribution of \( L \) given \( A = 0 \) and \( C \) and let \( j (\alpha) = j (L|A = 0, C; \alpha) \) be a parametric model for \( j \). Define \( w = w (Y, L|A, C) \) such that \( w (0, L|A, C) = w (Y, 0|A, C) = 1 \) and \( W \geq 0 \) so that \( W \) is the true generalized conditional odds ratio function for \( Y \) and \( L \) given \( A \) and \( C \) [7] and let \( w (\omega) = w (Y, L|A, C; \omega) \) be a parametric model for \( w \). Let \( \chi (\beta) = \chi (A, L|C, \beta) \) be a parametric model for \( \chi (A, L|C) \) and let \( \pi (\eta) = P(R = 1|A, C, Y; \eta) \) be a parametric model for \( \pi \). Let \( r = r (Y|A, L = 0, C) \) be a model for the distribution of \( Y \) given \( A, L = 0 \), and \( C \) and let \( r (\theta) = r (Y|A, L = 0, C; \theta) \) be a parametric model for \( r \). Additionally, let \( h = P(A = 1|L = 0, C) \) and let \( h (\kappa) = P(A = 1|L = 0, C; \kappa) \) be a parametric model for \( h \).

Estimates for the parameters \((\alpha, \omega, \beta, \theta) \), \( (\hat{\alpha}, \hat{\omega}, \hat{\beta}, \hat{\theta}) \), can be found by using direct likelihood maximization of the observed data. This entails maximizing the observed data likelihood,

\[
\prod [f (Y, L, |A, C; \alpha, \omega, \beta, \theta)]^R \left[ \int f (Y, l, |A, C; \alpha, \omega, \beta, \theta) \, dl \right]^{1-R}.
\]

An estimate of \( \eta \), \( \hat{\eta} \), can be found by fitting \( \pi (\eta) \) to the observed data \( \hat{\eta} = \arg \max_\eta \left[ \sum R_l \log \pi (\eta) + (n - \sum R_l) \log (1 - \pi (\eta)) \right] \) where \( n \) is the total number of subjects. Finally, \( \kappa \) can be estimated using inverse probability weighting using \( \frac{1}{\pi (\hat{\eta})} \) as weights in the complete cases.
Result 2:

Define $\hat{\Psi}$ the solution to

$$\mathbb{P}_{n}\left(t_{\text{Miss}}(\hat{\Psi}; \hat{\alpha}, \hat{\omega}, \hat{\beta}, \hat{\theta}, \hat{\eta}, \hat{\kappa})\right) = 0,$$

where $\mathbb{P}_{n}(\cdot) = \frac{1}{n} \sum_{i}(\cdot)$. Then under standard regularity conditions, $\hat{\Psi}$ is consistent and asymptotically normal if $\chi(\beta)$ is correctly specified and in addition either (i) $\pi(\hat{\eta})$ and $h(\hat{k})$ are consistent for $\pi$ and $h$ or (ii) $j(\hat{\alpha})$, $w(\hat{\omega})$, and $r(\hat{\theta})$ are consistent for $j$, $w$, and $r$. Additionally, at the intersection submodel where all of the models are evaluated at the truth, the variance of $\hat{\Psi}$ achieves the semiparametric efficiency bound for the union of models (i) and (ii).

An alternative approach using a more standard parametrization can sometimes be used, provided that the parametrization can be shown to satisfy the variation dependence described in Result 1, which ensures the existence of a joint distribution for $(L, A, Y|C)$. A standard parametrization in this case implies specifying parametric models for $p$, $t$, $\pi$, and $m$ needed to evaluate the efficient influence function (2.3).

Let $p(\lambda)$ be a parametric model for $p$, $t(\phi)$ be a parametric model for $t$, $\pi(\eta)$ be a parametric model for $\pi$, and $m(\nu)$ be a parametric model for $m$. An estimator of $\pi, \hat{\eta}$, can be found by fitting $\pi(\eta)$ on the observed data by using, using, for example, a logistic regression of $R$ on $A$, $C$, and $Y$. We can estimate $\lambda$ by using inverse probability of censoring weighting using $\frac{1}{\pi(\eta)}$ as weights in the complete cases. For example we might fit a weighted logistic regression of $A$ on $C$, and $L$. By specifying $m$ and $t$ as normal with constant variance, one may ensure the existence of a corresponding joint distribution of $(Y, L)$ given $A$ and $C$, provided the mean of $Y$ given $L$, $A$, and $C$ is linear in $L$ and likewise the mean model for $L$ given $Y$, $A$, and $C$ is linear in $Y$. Assumption 5 (MAR) allows us to estimate $t(\phi)$ using the complete cases by using, for example, a standard linear regression of $L$ on $A$, $C$, and $Y$. Finally, we describe a simple Monte Carlo algorithm to estimate $m(\hat{\nu})$:

1. Create $M$ duplicates of the data.

2. Where $L$ is missing, “fill in” the missing variable with a random draw from $t(\hat{\phi})$.

3. Stack all $M$ datasets in long format and estimate $m(\hat{\nu})$. In practice this will involve fitting a standard model for $Y$ on $A$, $L$, and $C$. For example, a standard main effects linear model.
We then have the following result.

**Result 3:**

Define $\hat{\Psi}$ the solution to

$$P_n \left( t_{Miss} \left( \hat{\Psi}; \hat{\lambda}, \hat{\nu}, \hat{\eta}, \hat{\phi} \right) \right) = 0.$$ 

Then under standard regularity conditions, $\hat{\Psi}$ is consistent and asymptotically normal if the implied form of $\chi (\beta)$ is correctly specified and in addition either (i) $\pi (\hat{\eta})$ and $p \left( \hat{\lambda} \right)$ are consistent for $\pi$ and $p$ or (ii) $m (\hat{\nu})$ and $t \left( \hat{\phi} \right)$ are consistent for $m$ and $t$, but not necessarily both. As before, at the intersection submodel where all of the models are evaluated at the truth, the variance of $\hat{\Psi}$ achieves the semiparametric efficiency bound for the union of models (i) and (ii).

Then, it is straightforward to show that the solution to equation (2.3) is given by

$$\hat{\Psi} = P_n \left\{ \frac{R}{\pi} \left\{ \frac{I(A = 0)}{P(A = 1)} \frac{\hat{p}}{1 - \hat{p}} (Y - \hat{\mu}_Y^0) + \frac{I(A = 1)}{P(A = 1)} \hat{\mu}_Y^0 \right\} \right.$$ 

$$- \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 0)}{P(A = 1)} YE \left[ \frac{\hat{p}}{1 - \hat{p}} \hat{\mu}_Y^0 | Y, A = 0, C \right] \right\} \right.$$ 

$$+ \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 0)}{P(A = 1)} E \left[ \hat{p} | Y, A = 0, C \right] \right\} \right.$$ 

$$- \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 1)}{P(A = 1)} E \left[ \hat{\mu}_Y^0 | Y, A = 1, C \right] \right\} \right\},$$

where $\hat{\mu}_Y^0 (\nu) = E[Y | A = 0, L, C; \hat{\nu}]$, $\hat{\pi} = \pi (\hat{\eta})$, and $\hat{p} = p \left( \hat{\lambda} \right)$.

The asymptotic distribution of the estimator can be found as follows. Let $Q_R (\hat{\eta})$ be an individual contribution to the score for $\eta$, $Q_A (\hat{\lambda})$ be an individual contribution to the score for $\lambda$, $Q_L (\hat{\phi})$ be an individual contribution to the score for $\phi$, and $Q_Y (\hat{\nu})$ be an individual contribution to the score for $\nu$. For example,

$$Q_R (\eta) = \frac{d}{d\eta} \log \left[ \pi(\eta)^R (1 - \pi(\eta))^{1-R} \right].$$

Also let $Z \left( \hat{\Psi}; \hat{\lambda}, \hat{\nu}, \hat{\eta}, \hat{\phi} \right)$ be an individual contribution to the estimating equation for $\Psi$. Let
$\Xi = (\eta, \lambda, \phi, \nu)$ and define

$$ Q(\Xi) = \begin{pmatrix} Q_R(\hat{\eta}) \\ Q_A(\hat{\lambda}) \\ Q_L(\hat{\phi}) \\ Q_Y(\hat{\nu}) \end{pmatrix}. $$

Then, under standard regularity conditions,

$$ \sqrt{n}(\hat{\Psi} - \Psi) = \frac{1}{\sqrt{n}} E \left[ \frac{dZ}{d\Psi} \right]^{-1} \sum_{i=1}^{n} \left\{ Z(\Psi, \Xi) - \frac{d}{d\Xi} E[Z(\Psi, \Xi)] \right\} E \left[ \frac{dQ}{d\Xi} \right]^{-1} Q(\Xi) + o_p(1). $$

Therefore, a consistent estimator of the asymptotic variance of $\sqrt{n}(\hat{\Psi} - \Psi)$ is

$$ \left[ \mathbb{P}_n \left( \frac{dZ}{d\Psi} \right)^{-1} \mathbb{P}_n \left[ V \left( \hat{\Psi}, \hat{\Xi} \right) V' \left( \hat{\Psi}, \hat{\Xi} \right) \right] \left[ \mathbb{P}_n \left( \frac{dZ}{d\Psi} \right) \right]^{-1} \right]^{-1}. $$

where $V \left( \hat{\Psi}, \hat{\Xi} \right) = Z(\Psi, \Xi) - \frac{d}{d\Xi} E[Z(\Psi, \Xi)] \right\} E \left[ \frac{dQ}{d\Xi} \right]^{-1} Q(\Xi)$.

Alternatively, we recommend using the nonparametric bootstrap to obtain estimates of the variance.

In their application to the B-Aware trial, Williamson et al. [66] use models that are compatible with this new parametrization. As a result, the estimating equation under those choices of models is DR, though not MR as they claim. However, their simulation models fail to be compatible, thus there is no chance of it being even DR. Moreover, the favorable simulation results obtained by the authors can be explained by two reasons. The first reason is that the missing confounder has little effect on the exposure compared the other, fully observed confounders. As a result, the complete case estimator performs well. The second reason for their favorable simulation results is that the model for the missing confounder is not badly misspecified. In order to misspecify the model for the missing confounder, the authors omit variables that have little effect in the model while retaining variables that have a large effect [66].

### 2.5 Simulation Study

We report a simulation study comparing finite sample performance of our DR estimator to a number of existing methods. We compared our DR estimator to an estimator that used Monte Carlo direct likelihood maximization, one using inverse probability of censoring weights, as well as a complete
case estimator, a naive estimator that drops the missing confounder, and an estimator calculated from the complete dataset where \( L \) was observed for all subjects. The last estimator is obviously not feasible in the presence of missing data however provides a benchmark to assess efficiency loss due to missing data.

In the first set of simulations, we simulated \( C \) by summing draws from a \( \text{Normal}(0,1) \) distribution and a \( \text{Uniform}(-1,1) \) distribution. The treatment, \( A \), was Bernoulli with \( P_r(A = 1|C; \zeta) = p_A = \zeta_0 + \zeta_1 C \). For this simulation we chose \((\zeta_0, \zeta_1) = (-0.44, 0.40)\). The outcome, \( Y \), was chosen to be Normal conditional on \( A \) and \( C \), with \( Y = v_0 + v_1 A + v_2 C + \epsilon_y \) where \( \epsilon_Y \sim N(0, \sigma_Y^2) \), \((v_0, v_1, v_2, \sigma_Y^2) = (0.2, 0.38, 0.3, 0.51)\). Similarly, \( L \) was chosen to be Normal conditional on \( A \) and \( C \), with \( L = \alpha_0 + \alpha_1 A + \alpha_2 C + \epsilon_L \) where \( \epsilon_L \sim N(0, \sigma_L^2) \) and such that \( \text{Cov}(\epsilon_Y, \epsilon_L) = \sigma_{YL} \) and \((\alpha_0, \alpha_1, \alpha_2, \sigma_L^2, \sigma_{YL}) = (-0.15, 0.215, 0.14, 0.43, 0.21)\). As a consequence, the distribution of \( L \) given \( A, Y \) and \( C \), \( t(\phi) \), was Normal such that \( E[L|A,Y,C] = \phi_0 + \phi_1 A + \phi_2 Y + \phi_3 C = \mu_L \) where \((\phi_0, \phi_1, \phi_2, \phi_3) = (-0.23, 0.058, 0.41, 0.016)\), the distribution of \( Y \) given \( C, L \) and \( A \), \( m(\nu) \), was Normal such that \( E[Y|A,L,C] = \nu_0 + \nu_1 A + \nu_2 L + \nu_3 C = \mu_Y \) where \((\nu_0, \nu_1, \nu_2, \nu_3) = (0.27, 0.275, 0.49, 0.23)\), and \( p(\lambda) = \text{logit}[P(A = 1|L,C; \lambda)] = \lambda_0 + \lambda_1 L + \lambda_2 C \) was the propensity score with \((\lambda_0, \lambda_1, \lambda_2) = (-0.42, 0.5, 0.36)\). These models satisfy the variation dependence described in Result 1 and ensure the existence of a joint distribution of \((L, A, Y|C)\).

These simulations were used for the first 6 figures below (a-f) and have only a moderate relationship between \( L \) and \( C \). This setting is especially useful when we want to make changes to the propensity score which will be explained further below.

In the second set of simulations, \( C \) was chosen as in the previous simulation. The treatment, \( A \), was Bernoulli with \( P_r(A = 1|C) = p_A = \zeta_0 + \zeta_1 C \) as above. However, for this simulation we chose \((\zeta_0, \zeta_1) = (-0.44, 0.38)\). The outcome, \( Y \), was chosen to be Normal conditional on \( A \) and \( C \) as above. \( L \) was chosen to be Normal conditional on \( A \) and \( C \) similarly to the previous simulation but instead with and \( \alpha_2 = 0.914 \) in order to have a strong relationship between \( L \) and \( C \). As a consequence, \( t(\phi) \) was Normal such that \( E[L|A,Y,C] = \phi_0 + \phi_1 A + \phi_2 Y + \phi_3 C = \mu_L \) where \((\phi_0, \phi_1, \phi_2, \phi_3) = (-0.23, 0.058, 0.41, 0.79)\), \( m(\nu) \) was Normal such that \( E[Y|A,L,C] = \nu_0 + \nu_1 A + \nu_2 L + \nu_3 C = \mu_Y \) where \((\nu_0, \nu_1, \nu_2, \nu_3) = (0.27, 0.275, 0.49, -0.146)\) and \( p(\lambda) = \text{logit}[P(A = 1|L,C; \lambda)] = \lambda_0 + \lambda_1 L + \lambda_2 C \) was the propensity score with \((\lambda_0, \lambda_1, \lambda_2) = (-0.42, 0.5, 0.10)\). These simulations were used for the final 2 figures below (g and h) and have a strong relationship between \( L \) and \( C \). This setting is
especially useful when we want to make changes to the joint distribution of \( Y \) and \( L \), which will be explained further below.

In both simulations, \( R \) was Bernoulli (\( \pi \)) with \( \pi (\eta) = \logit [Pr(R = 1|A, C, Y)] = \eta_0 + \eta_1 A + \eta_2 C + \eta_3 Y \) where \((\eta_0, \eta_1, \eta_2, \eta_3) = (1, -1.75, -1.75, 1.25)\). In both simulations, on average, \( P(R = 1) \approx 0.61 \). The observed data was given by i.i.d. realizations of \((R, RL, Y, A, C)\) with sample size equal to 2,500. Many more details concerning the simulation can be found in the appendix.

In simulations, in order to gauge the loss of efficiency due to missing data, we computed the full data estimator (Full) of \( \Psi \) using the complete data \((A, Y, L, C)\) based on equation (1.1). We also implemented standard IPCW estimation, Monte Carlo direct likelihood maximization using 100 imputed datasets (MCDLM), complete-case analysis (CC), and a naive estimator (Naive) that drops the missing confounder \( L \) completely and evaluates (1.1) upon substituting an estimate of \( P(A = 1|C) \) for \( p \).

For the various methods we fitted the following models. For the missingness mechanism \( \pi (\eta) \), we fitted a logistic regression. Similarly, we fitted a logistic regression for the propensity score, \( p (\lambda) \), using inverse probability weighting with \( \frac{1}{\pi(\eta)} \) as weights in the complete cases. For the distribution of the missing variable, \( t (\phi) \), we fitted a main effects linear model. Finally, for the outcome model, \( m (\nu) \), we fitted a main effects linear model.

The IPCW estimator required \( \pi (\eta) \) as well as \( p (\lambda) \). MCDLM used \( t (\phi) \) as well as \( p (\lambda) \). The complete case estimator only required \( p (\lambda) \). The Naive estimator required a logistic regression for \( A \) with main effects for \( C \) alone, \( \tilde{p} (\tilde{\lambda}) = Pr (A = 1|C; \tilde{\lambda}) \). All these methods were compared to the proposed DR estimator which required \( p (\lambda), \pi (\eta), t (\phi), \) and \( m (\nu) \).

For the CC, Naive, MCDLM, IPCW, and Full estimators, we calculated ETT for each method using equation (1.1) for \( \Psi \). For the naive estimator the odds, \( p (\lambda) / [1 - p (\lambda)] \), were replaced with \( \tilde{p} (\tilde{\lambda}) / [1 - \tilde{p} (\tilde{\lambda})] \) and for the IPCW estimator the odds were estimated with inverse probability weighting. Our proposed estimator was calculated using equation (2.7).

The misspecified versions of each model were as follows. The missingness mechanism, \( \pi \), was misspecified by only using \( C \) in the regression, \( \pi^* = P(R = 1|C; \lambda^*) \). In order to misspecify a model for \( p \) or \( f \) we simply (incorrectly) set the coefficient on \( C \) to 0 in the working model. This form of misspecification was chosen in order to preserve the structure of the odds ratio between \( A \) and \( L \) given \( C, \chi(A, L|C) \), wherever it is required as seen in Section 3.
It is important to note that if $L$ and $C$ are strongly correlated, particularly when the coefficient on $C$ in the propensity score model, $\lambda_2$, is small, then not including $C$ in the propensity score will not be far off from the truth as $L$ will likely suffice to account for confounding. However if $L$ and $C$ are weakly correlated, then any imputation of $L$ that sets the coefficient on $C$ to zero will not be far off from the true model that includes $C$. Therefore we impose a weak correlation for the simulations exploring misspecification of $p$ and a strong correlation for those misspecifying $f$. For settings where both are misspecified, we impose a weak correlation $L$ and $C$. We denote the incorrect propensity score as $p^*$ and the incorrect joint distribution of $Y$ and $L$ given $A$ and $C$ as $f^*$.

Figure 2.1 summarizes the results in the form of Monte Carlo boxplots for the estimated population effect of treatment on the treated for 250 MC samples of 2,500 subjects for each of the following scenarios: (a) all models were correctly specified, (b) $\pi^*$ used in place of $\pi$, (c) $p^*$ used in place of $p$, (d) $\pi^*$ and $p^*$ used in place of $\pi$ and $p$, (e) $f^*$ and $p^*$ used in place of $f$ and $p$, (f) $f^*$, $\pi^*$, and $p^*$ used in place of $f, \pi$, and $p$, (g) $f^*$ used in place of $f$, and (h) $f^*$ and $\pi^*$ used in place of $f$ and $\pi$.

Regardless of model misspecification, the Naive and CC estimators are biased ($a-f$) as expected. The Full estimator is only biased when the propensity score is misspecified ($c-f$). This result is expected as it uses the full data and therefore does not require a model for the missingness mechanism or the missing covariate. Similarly, the IPCW estimator is biased when $p^*$ or $\pi^*$ are used in place of $p$ or $\pi$ ($b-f,h$) as it requires both and not a model for $f$. Additionally, the IPCW estimator tended to have large variance compared to the other estimators, even under correct specification for $p$ and $\pi$. The MCDLM estimator is biased when $p^*$ is used in place of $p(\cdot,f)$. When $f^*$ is used in place of $f$, but $p$ is correctly used ($g, h$) the MCDLM estimator is biased, but not overly so. This is likely an artifact of the simulation design regarding the correlation between $L$ and $C$ as explained above. Finally, our DR estimator is only biased under the settings we expected, namely when $f$ is misspecified along with $p$ or $\pi$ or both misspecified ($e,f,h$). Even in settings where the DR estimator is biased, the bias is less than that of the other biased estimators. We note that in the setting where $\pi^*$ and $f^*$ are used in place of $\pi$ and $f$, the not all that different from the bias for the MCDLM estimator. However, this may be an artifact of the simulation design. In the setting where $f^*$ is used in place of $f$, there is large variability for all estimators except the Naive one. This is unexpected because the CC, Full and IPCW don’t use $f$ in their calculations; these findings may be a result of the relatively small sample size of 2,500. Overall, despite a few
anomalies, the simulations are in line with expectations.

(a) All models correct  
(b) Model for $\pi$ incorrect  
(c) Model for $p$ incorrect  
(d) Models for $p$ and $\pi$ incorrect  
(e) Models for $p$ and $f$ incorrect  
(f) Models for $p$, $\pi$, and $f$ incorrect  
(g) Model for $f$ incorrect  
(h) Models for $\pi$ and $f$ incorrect

Figure 2.1: Simulation Results
The simulation used in Williamson et al. [66] did not allow for the range of settings we have explored. Furthermore, only 30% of the subjects had a missing covariate and the model for $R$ did not include $Y$. As a result the complete case estimator was in fact consistent in their simulation settings artificially obviating the need for modeling either $\pi$ or the density of $L$ given $Y$, $A$, and $C$ correctly.

2.6 Discussion

Analysts are commonly faced with missing data when using observational data to estimate a causal effect. This is particularly true in the setting of two stage non-monotone missingness, such as when potential confounding information is missing along with counterfactual outcomes. The difficulty arises when full data nuisance parameters are entangled with nuisance parameters which are needed to account for data missing at random. In such a setting it is unlikely that researchers will know the underlying mechanisms for the missingness and confounding. Therefore, model misspecification is a likely source of bias when using standard statistical analysis methods. In this chapter we have explained why the proposed method of Williamson et al. [66] fails to achieve the claimed MR property by carefully examining model dependencies. We identified the modeling assumptions through an alternative parametrization of the joint distribution of the outcome and missing confounder in order to understand the nuisance parameter entanglements. Furthermore, we presented an estimator of the effect of treatment on the treated that accounts for both missingness and potential confounding and that is robust to partial model misspecification. The chapter focuses on the effect of treatment on the treated, however the proposed method can be applied to the average causal effect and may be extended to other settings where one may wish to obtain a DR estimator in the presence of entangled nuisance parameters.

The simulation study confirmed our proposed estimator is doubly robust and outperformed existing methods but still failed to be multiply robust as we argued on theoretical basis. Moreover, we only considered a setting in which a single confounder had missing data. It is more common that several variables may be missing possibly in arbitrary patterns across individuals and therefore it would be important to extend our approach to allow for arbitrary missing data patterns.
Doubly Robust Regression Analysis for Data Fusion

with

James M. Robins

and

Eric J. Tchetgen Tchetgen
3.1 Introduction

Parametric likelihood based inference for regression analysis is a well-developed area of modern statistical theory. In recent years, fairly complete theory has also developed to account for incomplete outcome or covariate information in regression analysis. Inverse probability weighting (IPW) of complete cases and multiple imputation are two prominent methods that stand out in modern missing data theory [15, 46]. A fundamental assumption on which most missing data methods rely is that the probability of observing a subject with complete data is bounded away from zero, also known as the positivity assumption [65]. In this chapter, we consider a more extreme form of incomplete data, in which the positivity assumption does not hold, i.e. the probability of observing complete data is zero for all units in the population.

This situation may arise, for instance, when two data sets from separate sources are fused together such that no unit belongs to both sources and some variables obtained from one source are not available in the other source. For instance, as we consider throughout in the chapter, it may be that the outcome of interest $Y$ is collected only in the first data set but not in the second, and likewise, a subset of regressors $L$ are only observed in the second data set but not in the first. Both data sets contain information on common variables $V$. A prominent example of such missing data structure concerns the main/validation study design in comparative effectiveness studies. In these designs, a main study sample in which outcome, treatment variable and a relatively limited subset of confounders are available, is enriched with an external validation sample which contains extensive potential confounders together with treatment information, but lacks outcome information [57]. The two datasets are then fused together in the hope that information available in the validation sample can somehow be leveraged to reduce confounding bias.

Another example, somewhat related to meta-analysis for prediction model evaluation [11, 12, 40], might involve enriching a data set of a clinical study with covariate information from a separate source, say the census, containing socio-demographic information, but no outcome data, for the purpose of improving clinical risk prediction [6, 9]. Clearly, in both of these examples, a regression model for the outcome on the combined set of covariates can be identified only under fairly stringent parametric assumptions and, as we discuss below, provided that there is a non-trivial overlap in the amount of information available from both sources of data. We shall refer to this general framework
as regression analysis for data fusion.

The missing data literature has previously described the data fusion problem as that of "data matching." The textbooks by Rässler (2013) [38] and D’Orazio et al. (2007) [13] provide an extensive overview of the state of the art for data fusion. D’Orazio et al. (2010) [37] give an overview and comparison of many of the existing data matching methods in the literature and their assumptions needed to recover valid inferences using these methods. A fundamental assumption on which much of this literature relies on is that conditional independence between \( Y \) and \( L \) given \( V \), an assumption which is likely untenable in practice. This assumption is particularly problematic in the two settings described above where a potential non-null association between \( Y \) and \( L \) given \( V \) is an important part of the scientific hypothesis under consideration. When the samples are drawn from a finite population according to a complex survey design, concatenation [50] and calibration [39, 67] are two commonly used methods for statistical matching. Concatenation [50] involves modifying the sample weights of the samples in order to get a unique sample given by the union of the original sample with new weights that represent the population of interest. The new weights require computing the probability of the subjects in one sample under the survey design of the other sample, which requires detailed knowledge of the survey designs. Calibration [39] preserves both samples and calibrates the two sets of survey weights. The method obtains a unique estimate of the common variable, \( V \), by combining the estimates of the distribution of \( V \) from both samples and then calibrating the original sample weights to the obtained estimate. The weights are then used to estimate the distribution of \( Y|V \) in the sample with \( L \) and the distribution of \( L|V \) in the sample with \( Y \). Wu (2004) [67] suggests similar approaches with different constraints for the sample weights, such as forbidding negative final weights.

Data fusion designs have also been prominent in literature on instrumental variable (IV) methods for causal inference. An instrumental variable is an exogenous variable known to be associated with a treatment or exposure variable of interest but only to be associated with an outcome of interest through its association with treatment. The IV approach can, under certain conditions, be used to recover an unbiased estimator of a causal effect in the presence of unmeasured confounding. The most common IV approach assumes a linear model relating the outcome to exposure and observed covariates, together with a linear model relating exposure to IV and covariates. Angrist and Krueger (1992) [1] examine estimation and inference about the causal effect of exposure under
such linear models, when IV and exposure are available from one data source, however outcome and IV are available in a separate data source, and therefore no subject has available data on all three variables, IV, exposure and outcome. These two-sample instrumental variable estimators deliver point identification and inference by explicitly leveraging parametric assumptions. Regression using Two-Sample Two-Stage Least Squares was introduced by Klevmarken (1982) [24] and shown by Inoue and Solon (2010) [21] to be more efficient than the two-sample instrumental variable estimator. These methods assume that both samples are i.i.d. random samples from the same population with finite fourth moments and are independent. Pacini (2016) [33] assumes independence of the samples and makes use of the marginal distributions to provide a characterization of the identified set of the coefficients of interest when no assumption on the joint distribution of \((Y, V, L)\) is imposed.

Robins et al. (1995) [43] consider a missing data setting closely related to ours. The main contribution of their paper, is to characterize a large class of semiparametric estimators of a parametric conditional density of \(Y\) given \(L\) and \(V\) when \(L\) is missing at random. They characterize the efficient influence function for the parameters of the parametric model which is the solution to an integral equation that is not generally available in closed form. They also point out in a remark that Bickel et al. (1993) [5] and Hasminskii and Ibragimov (1983) [17] obtained results similar to theirs when \(Y\) and \(L\) are never observed together, which is the data fusion setting with which the current chapter is concerned. An important contribution of our chapter is to show that, in fact, there exists a large class of twin IPW estimating functions for the parameters of the conditional density \(f(Y|L, V)\) available in closed form and therefore convenient candidates as estimating functions. The proposed semiparametric estimating functions include a large class of estimating functions that are in fact doubly robust (DR) in that they produce estimators which are consistent and asymptotic normal if, in addition to the outcome model of interest, one correctly specifies a model for either the data source process or the distribution of unobserved covariates. To the best of our knowledge, our proposed estimators are new and make an important contribution to growing literature on data fusion inference. In addition, we show that the efficient influence function for the parameters of the conditional density is available in closed form in the special case where the outcome is polytomous.

In section 2 we lay out notation and preliminaries. In section 3 we develop the general class of estimators as well as a new semi-parametric doubly robust method. In section 4 we discuss implementation. We examine and evaluate the finite sample performance of the double robust
approach in an extensive simulation study summarized in section 5. In section 6 we discuss local efficiency in the special case of binary outcome and provide a general expression for the efficient influence function in the case of polytomous outcome in the appendix. We conclude in section 7 with a discussion. Throughout, proofs and derivations can be found in the appendix.

3.2 Preliminaries

Let $R$ be an indicator that a subject is observed in data source $A$ ($R = 1$) or in data source $B$ ($R = 0$). Let $V$ denote covariates which are observed in both sources, $Y$ denote the outcome only observed in source $A$, and $L$ denote covariates only observed in source $B$. The full data $(Y, L, V)$ are i.i.d realization from a common law $f(Y, L, V)$. Let $f = f(Y|V, L)$ be the true conditional distribution of $Y$ given $V$ and $L$. Let $\pi = Pr(R = 1|V)$ be the probability that a subject is in data source $A$. Therefore, the observed data are $O = (R, RY, (1 - R)L, V)$, so that $Y$ and $L$ are never observed together and one either observes $(V, Y)$ with probability $\pi$ or $(V, L)$ with probability $1 - \pi$. Our goal is to make inferences about a parametric model $f(Y|V, L; \theta)$ for the conditional density $f$.

Throughout, we make the following assumptions:

1. Correct outcome model: $f(V, L; \theta) = f(Y|V, L; \theta)$ is correctly specified such that $f(V, L; \theta^1) = f(Y|V, L)$ for some value $\theta^1$;

2. Positivity: For $\delta > 0$, $P(\delta < \pi(V) < 1 - \delta) = 1$;


Assumption 1 is an assumption that the outcome model proposed for $f$ is correctly specified. Assumption 2 is a positivity assumption and states that the probability of observing a subject in either data source is bounded away from both 0 and 1. We note that assumption 2 is strictly weaker that the usual positivity assumption typically assumed in missing data problems which requires a positive probability of observing complete data for each subject. Assumption 3 is an ignorable data source process assumption and states that the probability that a unit is observed in either data source only depends on $V$ and does not further depend on $Y$ or $L$. This assumption is akin to missing at random and is weaker than conditional independence between $Y$ and $L$ given $V$ required by many existing methods, such as matching [37].

We also note that assumption 1 is necessary for identification. To see this, consider that our
problem is solving

$$E [Y|V] = \int f^\dagger (Y|V,l) t(l|V) \, dl$$  \hspace{1cm} (3.1)$$

where we can estimate $E [Y|V]$ and $t(L|V)$ from data sources A and B respectively. If $f^\dagger (Y|V,L)$ were left unrestricted, we could always add to it any function $h(L,V)$ such that $E [h(L,V)|V] = 0$ and equation (3.1) would still be satisfied. Therefore there could not be a unique solution for $f^\dagger (Y|V,L)$. Thus we must know the specification of $f$ a priori.

### 3.3 Twin IPW Estimating Functions and DR Estimating Functions

In this section we describe a large class of twin inverse probability weighted (TIPW) estimating functions. Let $\pi (V; \eta) = P(R = 1|V; \eta)$ be a parametric model for the data source process indexed by a finite dimensional parameter $\eta$. We shall make use of the following assumption:

4. $\pi(V; \eta)$ is correctly specified such that $\pi (V; \eta^\dagger) = \pi$ for some value $\eta^\dagger$.

Then, for any choice of $g(Y,V)$, let

$$U_g(\theta; \eta) = \frac{R}{\pi(V; \eta)} g(Y,V) - \frac{1-R}{1-\pi(V; \eta)} E_\theta [g(Y,V)|V,L].$$  \hspace{1cm} (3.2)$$

**Result 1:**

Under assumptions 1-4,

$$E \left[ U_g \left( \theta^\dagger; \eta^\dagger \right) \right] = 0.$$  \hspace{1cm} (3.3)$$

These estimating functions are twin IPW estimating functions in the sense that they assign to every subject the inverse probability of observing the subject from the data source in which he or she was indeed observed.

To illustrate, suppose that $Y$, $L$, and $V$ are continuous and suppose that $\theta$ is of dimension three. Then we might choose $g(Y,V) = (Y, YV, YV^2)^T$ in order to account for incomplete information on $L$, provided that the expected derivative of the corresponding estimating function, $E \left[ \frac{d}{\theta} U_g (\theta, \eta) \right]$, is nonsingular which essentially requires that $V^2$ is associated with $L$. We also require that $g(Y,V)$ is of at least the same dimension as $\theta$ and that $E \left[ U_g(\theta)^T U_g(\theta) \right] < \infty$.

Interestingly, this general class of estimating functions includes a large set of doubly robust
estimating functions. First, consider the density of $L$ given $V$, $t = t(L|V)$, and suppose that one has specified a parametric model for $t$, $t(V; \alpha) = t(L|V; \alpha)$. Then we make the final additional assumption:

5. $t(V; \alpha)$ is correctly specified such that $t(V; \alpha^\dagger) = t$ for some value $\alpha^\dagger$.

We then obtain a subclass of doubly robust estimating equations, similar to (3.2), by setting $g(Y,V)$ to be a residual, $\tilde{g}(Y,V) - E_{\theta,\alpha} [\tilde{g}(Y,V)|V]$ for a choice of $\tilde{g}(Y,V)$. Hereafter we rename $\tilde{g}(Y,V)$ as $g(Y,V)$ to simplify notation.

Thus, for any choice of $g(Y,V)$, let

$$U_g^{DR}(\theta; \eta, \alpha) = \frac{R}{\pi(V; \eta)} \{ g(Y,V) - E_{\theta,\alpha} [g(Y,V)|V] \} + \frac{1 - R}{1 - \pi(V; \eta)} \{ E_{\theta,\alpha} [g(Y,V)|V] - E_\theta [g(Y,V)|V,L] \}.$$  \hfill (3.4)

Thus, for any choice of $g(Y,V)$, let

$$E \left[ U_g^{DR}(\theta^\dagger; \eta, \alpha) \right] = 0$$  \hfill (3.5)

if either (i) $\eta = \eta^\dagger$ and therefore assumption 4 holds; or (ii) $\alpha = \alpha^\dagger$ and therefore assumption 5 holds, but not necessarily both. The resulting estimating functions are then said to be doubly robust for $\theta$ in that they will be consistent provided we correctly specify a model for $t(V; \alpha)$ or $\pi(V; \eta)$, but not necessarily both. Additionally, at the intersection submodel where both models are evaluated at the truth, the variance of $U_g^{DR}(\theta; \eta, \alpha)$ achieves the semiparametric efficiency bound for estimating $\theta$ for the union of models $t(V; \alpha)$ and $\pi(V; \eta)$ at the intersection submodel.

In the next section, we construct feasible TIPW and DR estimating equations based on results (3.3) and (3.5) and describe the large sample behavior of the resulting estimators of $\theta$.

### 3.4 Feasible Twin IPW and DR Estimating Equations

A feasible TIPW estimator is recovered by setting $\mathbb{P}_n(U_g(\theta; \eta))$ equal to zero and a a feasible DR estimator is recovered by setting $\mathbb{P}_n(U_g^{DR}(\theta; \eta, \alpha))$ equal to zero where $\mathbb{P}_n(.) = \frac{1}{n} \sum_i(.)_i$. Solving these empirical version of equations (3.3) and (3.5) requires first estimating the nuisance parameters.

45
\[ \hat{\eta} = \arg \max \left[ \sum R_i \log \pi(V; \eta) + \left( n - \sum R_i \right) \log (1 - \pi(V; \eta)) \right] \] (3.6)

where \( n \) is the total number of subjects. An estimate of \( \alpha \), can be found by likelihood maximization restricted to sample \( B \) (\( R = 0 \)). That is,

\[ \hat{\alpha} = \arg \max \sum_n (1 - R_i) N_i(\alpha) \] (3.7)

where \( N_i(\alpha) = \log t(L_i|V_i; \alpha) \). Then \( E_\theta [g(Y,V) | V, L] = \int g(y, V) f(y|V, L; \theta) dy \).

To further ground ideas, consider the simple example where all components of \( L \) are continuous, and \( L \) is of dimension \( k \). Suppose the conditional density of \( Y \) given \( L \) and \( V \) is normal with

\[ E[Y|V, L; \theta] = \theta_0 + \theta_1 V + \theta_2 L \] (3.8)

and constant variance \( \sigma_Y^2 \). Suppose \( R \) follows a logistic model such that \( \logit \{ \pi(V; \eta) \} = \eta_0 + \eta_1 V \).

Then we estimate \( \hat{\eta} \) by solving equation (3.6) which could be done by fitting a logistic regression to the observed data. The DR estimator is calculated by further estimating \( \hat{\alpha} \) by solving equation (3.7). Suppose that the conditional density of \( (L|V) \) is multivariate normal where \( L = (L_1, L_2, ..., L_k) \) with \( \mu_{L_i} = E[L_j|V; \alpha] = \alpha_{j0} + \alpha_{j1} V \) and constant variance \( \sigma_{L_j}^2 \) for \( j \in \{1, 2, ..., k\} \). We can estimate \( \hat{\alpha} \) by fitting linear regressions by standard OLS restricted to subjects in sample \( B \).

Then the TIPW and DR estimates of \( \theta \) can be found by solving equations

\[ P_n(U_g(\theta; \eta)) = 0 \] (3.9)

and

\[ P_n(U_g^{DR}(\theta; \eta, \alpha)) = 0 \] (3.10)

respectively by a quasi-Newton procedure such as that which is implemented in the R function `optim()`. The resulting empirical TIPW estimators of \( \theta \) are unbiased if \( \pi(V; \eta) \) is correctly specified and the DR estimators are unbiased if \( \pi(V; \eta) \) or \( t(V; \alpha) \) is correctly specified. We note that if \( g(Y,V) \) is of larger dimension than \( \theta \), we cannot solve (3.9) or (3.10) directly and must use the generalized method of moments (GMM).
The asymptotic distribution of both estimators can be found as follows. Let \( \phi \) denote the set of nuisance parameters on which a given estimator of \( \theta \) depends. For the TIPW estimator, \( \phi = \eta \) and for the DR estimator \( \phi = \{\eta, \alpha\} \). Let \( W(\hat{\phi}) \) be an individual contribution to the score for \( \phi \). For example for the TIPW estimator,

\[
W(\eta) = \frac{d}{d\eta} \log \left[ \pi(V; \eta)^R (1 - \pi(V; \eta))^{1-R} \right].
\]

Also let \( U(\hat{\theta}, \hat{\phi}) \) be an individual contribution to the estimating equation for \( \theta \).

Then, under standard regularity conditions,

\[
\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} E \left[ \frac{dU}{d\theta} \right]^{-1} \sum_{i=1}^{n} \left\{ U(\theta, \phi) - \frac{d}{d\phi} E[U(\theta, \phi)] E \left[ \frac{dW}{d\phi} \right] \right\} W(\phi) + o_p(1).
\]

Therefore, a consistent estimator of the asymptotic variance of \( \sqrt{n}(\hat{\theta} - \theta) \) is

\[
\Sigma_\theta = \left[ \mathbb{P}_n \frac{dU}{d\theta} \right]^{-1} \mathbb{P}_n \left[ V(\hat{\theta}, \hat{\phi}) V^T(\hat{\theta}, \hat{\phi}) \right] \left[ \mathbb{P}_n \left( \frac{dU}{d\theta} \right)^T \right]^{-1}
\]

where \( V(\hat{\theta}, \hat{\phi}) = U(\theta, \phi) - \frac{d}{d\phi} E[U(\theta, \phi)] E \left[ \frac{dW}{d\phi} \right] \). Then a 95% Wald confidence interval for \( \theta_j \) is given by \( \hat{\theta}_j \pm 1.96 \text{se}_j \) where \( \text{se}_j \) is the \( j \)th component of the diagonal of \( \frac{1}{\sqrt{n}}\Sigma_\theta \). Alternatively, we recommend using the nonparametric bootstrap for inference.

We note that when \( Y \) is normal, as in our example, and one is primarily interested in the mean parameter and not in the variance component of the normal model (3.8), a convenient choice for \( g(Y, V) \) is given by \( g(V) = \left(1, V, \tilde{V}\right)^T \) where \( \tilde{V} \) represents a \( k \)-dimensional vector comprising polynomials in components of \( V \) and interactions in said components to address multivariate \( L \).

This choice of \( g(V) \) gives rise to the following set of estimating functions:

\[
U_g(\theta) = g(V) \left( \frac{R}{\pi(V; \eta)} Y - \frac{1-R}{1-\pi(V; \eta)} E_\theta [Y|V,L] \right)
\]

and

\[
U_g^{DR}(\theta) = g(V) \left[ \frac{R}{\pi(V)} \{Y - E_{\theta,\alpha}[Y|V]\} \right. \]

\[
+ \frac{1-R}{1-\pi(V)} \{E_{\theta,\alpha}[Y|V] - E_\theta [Y|V,L]\} \]

(3.11)

(3.12)
Then it is simple to evaluate $E_{\theta, \alpha}[g(Y, V)|V]$. For example, when using the examples given, and after reducing (3.4) to (3.12), we have that $E_{\theta, \alpha}[Y|V] = \theta_0 + \theta_1 V + \theta_2 E[L|V; \alpha]$.

### 3.5 Simulation Study

In this section, we report a simulation study evaluating the finite sample performance of our proposed estimators.

We defined $V = (A, C)$ such that $C$ was $\text{Normal}(0, 0.5)$ and $A$ was normal conditional on $C$ with $A = \phi_0 + \phi_1 C + \epsilon_A$ where $\epsilon_A \sim N(0, \sigma_A)$ with $(\phi_0, \phi_1, \sigma_A) = (0.27, 0.3, 0.3)$. Similarly, $L$ was chosen to be normal conditional on $A$ and $C$ with $L = \alpha_0 + \alpha_1 A + \alpha_2 C + \alpha_3 AC + \epsilon_L$ where $\epsilon_L \sim N(0, \sigma_L)$ with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_L) = (-0.27, 1.5, 1.3, 1.74, 0.3)$. We chose $Y$ to be normal conditional on $A$, $C$, and $L$ with $Y = \theta_0 + \theta_1 A + \theta_2 C + \theta_3 L + \epsilon_Y$ where $\epsilon_Y \sim N(0, \sigma_Y)$ with $(\theta_0, \theta_1, \theta_2, \theta_3, \sigma_Y) = (0.27, -0.43, 0.38, 0.52, 0.4)$. Finally $R$ was Bernoulli with logit $[Pr(R = 1|A, C)] = \eta_0 + \eta_1 A + \eta_2 C$ where $(\eta_0, \eta_1, \eta_2) = (0.175, -0.75, -0.75)$. On average, $P(R = 1) \approx 0.50$ so that each data source roughly contributes equally to sample size. The observed data is given by i.i.d. realizations of $(R, \text{RY}, (1 - R)L, A, C)$ with sample size equal to 1,500. For both the TIPW and DR estimators, we specified $g(V) = (1, A, C, AC)^T$.

In order to illustrate the DR property, we considered the following forms of model misspecification. The data source process, $\pi$, was misspecified by dropping $C$ from the logistic model, that is, $\text{logit}(\pi^*) = \text{logit}[Pr(R = 1|A; \eta^*)] = \eta_0^* + \eta_1^* A$. The density of the missing covariates, $t$, was misspecified by fitting a standard regression using only $C^2$ as a regressor, $t^* = t(L|C^2)$ where $E[L|C^2; \alpha^*] = \alpha_0^* + \alpha_1^* C^2$.

Figure 3.1 summarizes the results in the form of Monte Carlo boxplots for the estimates of $(\theta_0, \theta_1, \theta_2, \theta_3), (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$, for 1000 MC samples of 1,500 subjects for each of the following scenarios: (3.1a) all models were correctly specified, (3.1b) $\pi^*$ used in place of $\pi$, (3.1c) $t^*$ used in place of $t$, and (3.1d) $\pi^*$ and $t^*$ used in place of $\pi$ and $t$. For all results shown, the TIPW estimator is shown first followed by the DR estimator. The different parameter estimates are coded by color, with dashed lines on the true value of each parameter, by color. We can see that across all scenarios and parameters the DR estimator has less variability than the TIPW estimator (3.1a-3.1d). As expected, when $\pi$ is modeled incorrectly and $t$ is modeled correctly, the TIPW estimates are biased.
Figure 3.1: Simulation results

while the DR estimates are not (3.1b). When $t$ is modeled incorrectly and $\pi$ is modeled correctly, (3.1c), the variance of the DR estimates increases, relative to the scenario when all models are correct (3.1a). Additionally, there is no significant bias in any of the estimators and no significant bias reduction for the DR estimator compared to the TIPW estimator. This is likely because the TIPW estimator does not require a model for $t$, so incorrectly specifying $t$ will not affect it. However there is a moderate gain in efficiency by using the DR estimators. Finally, when $\pi$ and $t$ are modeled incorrectly (3.1d), all estimators are biased with relatively wide confidence intervals. Graphs were truncated for scaling issues, resulting in a loss of 30 data points out of 32,000 in total.
For the second set of simulations, shown in Figure 3.2, we relaxed the coefficient for the interaction between $A$ and $C$, $\alpha_3$, in the model for generating $L$. All models were chosen as above with the same coefficient values except we set $\alpha_3 = 0.3$. We lowered the level of interaction in order to show how the strength of the relationship between $L$ and $V$ can affect identifiability. When there is a strong $AC$ interaction in the model that generates $L$, using $AC$ in place of $L$ in $g(V)$ performed well. But when that interaction is weak, we begin to run into collinearity issues with $g(V)$ which effectively violates our requirement that $E \left[ \frac{\partial}{\partial \theta} U_g(\theta, \eta) \right]$ be nonsingular. We see that across all scenarios, we no longer have unbiasedness. As in the first simulation, across all scenarios (3.2a-3.2d),

Figure 3.2: Simulation results when $AC$ interaction is small
Bias and variance decrease when using the DR estimator instead of the TIPW estimator. When $t$ was modeled correctly and $\pi$ was modeled incorrectly (3.2b), the bias of all estimators is worse than in (3.2a) but bias and variance decreased significantly when using the DR estimator instead of the TIPW estimator. When $t$ was modeled incorrectly and $\pi$ was modeled correctly (3.2c), the bias for both estimators was similar to that in (3.2a). In fact, the bias for the DR estimator is similar across the first three scenarios (3.2a-3.2c) indicating that the doubly robust property holds for the initial biased estimate. When both $t$ and $\pi$ were modeled incorrectly (3.2d), there was large bias, but a decrease in both variance and bias when using the DR estimates. Graphs were truncated for scaling issues, resulting in a loss of 2,115 data points out of 32,000 in total.

Figure 3.3: Simulation results when $g(V)$ is chosen poorly
For the third set of simulations, shown in Figure 3.3, we chose $g(V) = (1, A, C, A^2)^T$ but kept all other facets of the simulation the same as the original. We made this choice in order to illustrate why $g(V)$ must be chosen carefully to reflect the relationship between $L$ and $V$. In general, across all scenarios the TIPW and DR estimators are quite biased with the only exception being the DR estimator for the intercept term, $\theta_0$. As with the previous simulations, regardless of the scenario, the DR estimator has lower variance than the TIPW estimator. Similarly, the DR estimators tend
to have less bias than the TIPW estimators, while still displaying significant bias. These results demonstrate the importance of choosing $g(V)$ carefully in order to do a good job reflecting the

![Simulation results for a small sample size](image)

Figure 3.4: Simulation results for a small sample size
relationship between $L$ and $V$. Graphs were truncated for scaling issues, resulting in a loss of 2,435 data points out of 32,000 in total.

For the final set of simulations, shown in Figure 3.4, we used the same parameters as in the first simulation, but lowered the sample size for each simulation from 1,500 to 250. Estimators performed similarly to the first simulation, but with more bias and variance, as is to be expected with a smaller sample size. Graphs were truncated for scaling issues, resulting in a loss of 1,106 data points out of 32,000 in total.

### 3.6 Local Efficiency

Consider the case where $Y$ is binary. Then the class of DR estimating functions derived in section 3 is equivalently written as follows.

Let

$$\Upsilon (\theta) = \frac{R}{\pi (V)} (Y - M (V; \theta, \alpha)) + \frac{1 - R}{1 - \pi (V)} (M (V; \theta, \alpha) - J (L, V; \theta))$$

where $J (L, V; \theta) = E [Y | L, V; \theta]$ and $M (V; \theta, \alpha) = E_{\Theta, \alpha} [Y | V; \theta, \alpha] = E [J (L, V; \theta) | V]$. Then let

$$\Lambda_1 = \{ h (V) \Upsilon (\theta) : h (V) \} \cap \mathcal{L}_2^0$$

**Result 3:**

Suppose $\hat{\theta}$ is a regular and asymptotically linear estimator of $\theta$ in the semiparametric model which only assumes assumptions 1-3. Then,

$$\sqrt{n} \left( \hat{\theta} - \theta^i \right) \sim \frac{1}{\sqrt{n}} E \left( \frac{\delta \Omega}{\delta \theta^i} \right)^{-1} \sum_{i=1}^{n} \Omega \left( \theta^i \right)$$

for some $\Omega (\theta, h)$ in $\Lambda_1$. $\hat{\theta}_{opt}$ achieves the efficiency bound for the model if and only if $\Omega (\theta, h) = h_{opt} (V) \Upsilon (\theta)$ where

$$h_{opt} (V) = -E [\nabla_{\theta^i} U_{DR} (\theta) | V] E \left[ U_{DR}^2 (\theta) | V \right]^{-1} .$$

(3.13)

$-E [\nabla_{\theta^i} \Upsilon (\theta) | V]$ and $E [\Upsilon^2 (\theta) | V]^{-1}$ have the following closed forms:

$$-E [\nabla_{\theta^i} \Upsilon (\theta) | V] = E \left[ (1, V, L)^T J (L, V; \theta) \{ 1 - J (L, V; \theta) \} | V \right]$$
and
\[
E \left[ \mathcal{Y}^2 (\theta) \bigg| V \right] = \frac{1}{\pi (V)} E \left[ M (V; \theta, \alpha) (1 - M (V; \theta, \alpha)) \bigg| V \right] \\
+ \frac{1}{1 - \pi (V)} E \left[ - \left( M (V; \theta, \alpha) \right)^2 + J (L, V; \theta) \right] ^2 \bigg| V \right].
\]

One could then construct a locally efficient estimator by substituting the unknown quantities with their empirical estimates including a preliminary TIPW estimator of \( \theta \) in order to evaluate the optimal index.

### 3.7 Discussion

Traditional regression models break down when two data sources are fused together such that no subject has complete data. Investigators often consider parametric models for a given outcome regressed on a number of independent variables, but existing missing data methods for parametric models do not adequately deal with the missing data structure that arises from data fusion. In this chapter we have developed a general class of semiparametric twin inverse probability weighting estimating functions, whose resulting estimators are consistent if the outcome regression and data source process are correctly specified. This general class of estimating functions includes a large set of doubly robust estimating functions which additionally require a model for the covariates that are only observed in one data source. The proposed method is doubly robust in that it is consistent and asymptotic normal if we correctly specify a model for either the data source process or the distribution of unobserved covariates, but not necessarily both.

It can be difficult to achieve identifiability when using our proposed method. Our method requires that the index \( g(Y, V) \) be of at least the same dimension as \( \theta \), which can be impossible if \( V \) is of low dimension while \( L \) is of high dimension or if \( V \) is categorical with few categories. Therefore it is advantageous to have a rich set of covariates measured in both data sources. We also require that \( E \left[ \frac{d}{d\theta} U_g (\theta, \eta) \right] \) be nonsingular which will generally require \( L \) and \( V \) to be highly correlated and we require that \( g(Y, V) \) be chosen carefully. In extensive simulations, we saw that weakening the relationship between \( L \) and \( V \) (by decreasing the coefficient on an interaction term) lead to overall increases in variability and bias for both the TIPW and DR estimates of \( \theta \). The bias further increased when the models for \( \pi \) or \( t \) were misspecified and the double robustness no
longer appeared to hold. When we made a poor choice for $g(V)$, we also saw increases in bias and a loss of double robustness. Across several scenarios we explored, the TIPW estimator was less biased than the DR estimator, though the DR estimates always had lower variance than their TIPW counterparts. Future work investigating this method might closely examine the identification criteria and the consequences when those criteria break down.

There are also several clear areas for additional research, notably the open question of how to generalize this method to other settings.

An interesting direction would be to extend the methods to the setting of fusing multiple dataset together, not just two. Consider $n$ data sources with $V$ observed for all and each of $(L_1, L_2, ..., L_{n-1}, Y)$ observed in only a subset of sources with respective indicators of observation $(R_1, R_2, ..., R_{n-1}, R_n)$ and respective inclusion probabilities $(\pi_1, \pi_2, ..., \pi_{n-1}, \pi_n)$. Therefore the observed data are $O = (V, R_1 L_1, R_2 L_2, ..., R_{n-1} L_{n-1}, R_n Y)$. Then, provided $V$ is rich enough for identification, it is easy to extend (3.11) for linear models to be

$$U^*_g(\theta) = g(V) \left( \frac{R_n}{\pi_n} Y - \left[ \theta_0 + \frac{R_1}{\pi_1} \theta_1 L_1 + \frac{R_2}{\pi_2} \theta_2 L_2 + \ldots + \frac{R_{n-1}}{\pi_{n-1}} \theta_{n-1} L_{n-1} + \theta_n V \right] \right).$$

The above estimating function can easily be modified if $Y$ were available in more than one source of data. However it is unclear how to further extend our method to non-linear models, for example right censored survival data such as Cox proportional hazards models for estimation of causal effects or prediction under data fusion. Even with only two sources of data, a problem arises with our method because the estimating functions for Cox model regressions are much more complicated in the sense that $Y$ and $L$ are intrinsically tied to each other because the estimating function is a counting process multiplied by a form of a risk specific residual of the covariates. It is unclear how to proceed when $Y$ and $L$ are never observed together in this context.
References


A  Proofs and results for Chapter 1

A.1  Proof of (1.2):

Under assumptions of consistency, positivity, and no unmeasured confounding, and recalling that
$A \perp Y_0|C, Y_0$ holds trivially :

$$\Psi = E[Y_0|A = 1]$$

$$= \frac{P(A = 1)}{P(A = 1)} E[Y_0|A = 1]$$

$$= \frac{1}{P(A = 1)} E[Y_0]$$

$$= \frac{1}{P(A = 1)} E \left[ E[Y_0|A, Y_0, C] \right]$$

$$= \frac{1}{P(A = 1)} E \left[ \left(1 - A\right) \frac{P(A = 1|C, Y_0)}{P(A = 0|C, Y_0)} Y_0 \right].$$

A.2  Estimation

The estimator is found using a weighted approach.

EPS requires a model for $L|C, Y, A = 0, R = 1$. We can get an estimate of $t(L|C, Y, A = 0, R = 1)$
using the complete cases from the validation study.

It also requires estimating the propensity score, $P(A = 1|L, C)$ such that $\logit P(A = 1|L, C) = \lambda_0 + \lambda_1 L + \lambda_2 C$, which then creates another modeling assumption (namely that the probability of $A$
is logistic with main effects $L$ and $C$). In many scenarios, we would estimate the propensity score by
using the inverse probability of censoring as weights. However, ESP does not require a true model
for the missingness mechanism, only models for the propensity score and the density of the missing
variable. Thus we must use $t(L|C, Y, A = 0, R = 1)$ to find $\Lambda = (\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2)$. Therefore, we estimate
the propensity score as follows:

For those subjects with $R = 1$, we use the score function and for those with $R = 0$, we use the
expected score function.
The score is

\[ S_A(A, W; \Lambda) = \begin{pmatrix} 1 \\ L \\ C \end{pmatrix} (A - P(A = 1|W; \Lambda)). \]

Now we aim to find \( \Lambda = (\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2) \) by solving

\[
\sum_i (1 - R_i) E \left[ \begin{pmatrix} 1 \\ L_i \\ C_i \end{pmatrix} (A_i - Pr(A_i = 1|L, C)) \right]_{C, Y, A, R = 1} + \\
\sum_i R_i \begin{pmatrix} 1 \\ L_i \\ C_i \end{pmatrix} (A_i - Pr(A_i = 1|L_i, C_i)) = 0.
\]

This implies

\[
\sum_i (1 - R_i) E \left[ \begin{pmatrix} 1 \\ L_i \\ C_i \end{pmatrix} (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) \right]_{C, Y, A, R = 1} + \\
\sum_i R_i \begin{pmatrix} 1 \\ L_i \\ C_i \end{pmatrix} (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0.
\]

Estimates are obtained by using these score functions, leading us to a system of equations,

\[
\sum_i (1 - R_i) E \left[ (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) \right]_{C, Y, A, R = 1} + \\
\sum_i R_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0,
\]

\[
\sum_i (1 - R_i) E [L_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i))]_{C, Y, A, R = 1} + \\
\sum_i R_i L_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0,
\]

\[
\sum_i (1 - R_i) E [C_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i))]_{C, Y, A, R = 1} + \\
\sum_i R_i C_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0.
\]
These can be written as

$$
\sum_i (1 - R_i) \int (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) t(l_i|C, Y, A, R = 1)dl + \\
\sum_i R_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0,
$$

$$
\sum_i (1 - R_i) \int L_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) t(l_i|C, Y, A, R = 1)dl + \\
\sum_i R_i L_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0,
$$

$$
\sum_i (1 - R_i) \int C_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) t(l_i|C, Y, A, R = 1)dl + \\
\sum_i R_i C_i (A_i - \expit(\lambda_0 + \lambda_1 L_i + \lambda_2 C_i)) = 0.
$$

We must solve this system of equations to find $\Lambda = (\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2)$. Very often there is no closed form expression for $\Lambda$ and we must use methods such as the Newton-Raphson algorithm. The integral in the expected score function can prove difficult, but is easily approximated for low dimension $L$ through numerical integration methods such as Gaussian quadrature.

Then the extended propensity score can be found by evaluating

$$
P(A = 1|C, Y_0) = \frac{P(A = 1|L, C) \cdot P(A = 0|L, C)}{P(A = 1|C, Y, A = 0, R = 1)} = \\
E \left[ e^{\hat{\lambda}_0 + \hat{\lambda}_1 L + \hat{\lambda}_2 C} | C, Y, A = 0, R = 1 \right] = \\
e^{\hat{\lambda}_0 + \hat{\lambda}_1 C} E \left[ e^{\hat{\lambda}_1 L} | C, Y, A = 0, R = 1 \right] = \\
e^{\hat{\lambda}_0 + \hat{\lambda}_2 C} e^{\hat{\lambda}_1 \mu(C, Y)} M_{\text{resid}(L)|C, Y, A = 0, R = 1}(\lambda_1).
$$

Therefore

$$
\logit(P(A = 1|C = c, Y_0 = y)) = \hat{\lambda}_0 + \log M_L(\hat{\lambda}_1) + \hat{\lambda}_2 C + \hat{\lambda}_1 \mu(C, y) = \\
\hat{\lambda}_0 + \log (E[e^{\hat{\lambda}_1 \epsilon}]) + \hat{\lambda}_2 C + \hat{\lambda}_1 \mu(C, y) = \\
\hat{\lambda}_0^* + \hat{\lambda}_2 C + \hat{\lambda}_1 \mu(c, y).
$$

With $\lambda_0^* = \lambda_0 + \log M_L(\lambda_1)$, $M_L(\lambda_1) = E \left[ e^{\lambda_1 \epsilon} \right]$, $\mu(C, y) = E[L|C = C, Y = y, A = 0]$, and $\epsilon = (L - \mu)|R = 1$. Once we have the propensity score, $P(A = 1|C, Y_0)$, as a function of the observed data, we can calculate $\mu = \frac{1}{P(A = 1)} E \left[ (1 - A) \frac{P(A = 1|C, Y_0)}{P(A = 0|C, Y_0)} \right]$. 

63
In short, we calculate $P(A = 1|C, Y_0)$ as follows.

- Regress $L$ on $C$, $Y$, and $A$. Use that model to predict $L$ when $A = 0$, call the prediction $\mu_L$
  then calculate $\epsilon = L - \mu_L$ for all subjects with $L$ observed.
- Fit logit[$P(A = 1)|L, C] = \lambda_0 + \lambda_1 L + \lambda_2 C$ to get estimates for the lambdas, $\hat{\lambda}_0, \hat{\lambda}_1, \hat{\lambda}_2$.
- Calculate $\hat{\pi} = \exp(\hat{\lambda}_0 + \log (\mathbb{E}[\exp(\hat{\lambda}_1 (L - \mu_L) | R = 1)])) + \hat{\lambda}_2 C + \hat{\lambda}_1 \mu_L$
- Then $\Psi = \frac{1}{P(A=1)} E \left[ (1 - A) \frac{P(A=1|W,Y_0)}{P(A=0|W,Y_0)} Y_0 \right] = \frac{1}{P(A=1)} E[(1 - A) \ast \exp(\hat{\pi}) \ast Y]$

A.3 Data Simulation

A.3.1 Description

We wish to simulate data such that we have a binary treatment $A$, normal outcome $Y$, and with two confounders $L$ and $C$ such that $C$ is binary and $L$ is normal. We want to generate the data in such a way that we have closed forms for many of the relationships between the variables. These requirements are stated in the follow section.

A.3.2 Requirements

Simulate data such that the following relationships hold:

- $C \sim Bernouli(p_C)$
- $(L,Y|A) \sim BivariateNormal$
- $Y = \beta_0 + \beta_1 A + \beta_2 C + \epsilon_y$ where $\epsilon_Y \sim N(0, \sigma_Y^2)$
- $L = \alpha_0 + \alpha_1 A + \alpha_2 C + \epsilon_L$ where $\epsilon_L \sim N(0, \sigma_L^2)$
- $E[L|A=0, C, Y] = \eta_0 + \eta_1 C + \eta_2 Y$
- logit[$P(A = 1|L, Y, C) = \gamma_0 + \gamma_1 L + \gamma_2 C + \gamma_3 Y$
- logit[$P(A = 1|L, C) = \lambda_0 + \lambda_1 L + \lambda_2 C$

These basic requirements will necessitate other relationships.

First, $C \sim Bernouli(p_C)$ with probability chosen by the user.

Next we choose $(L,Y|A) \sim BivariateNormal$ with:
• $Y = \beta_0 + \beta_1 A + \beta_2 C + \epsilon_Y$ where $\epsilon_Y \sim N(0, \sigma_Y^2)$

and

• $L = \alpha_0 + \alpha_1 A + \alpha_2 C + \epsilon_L$ where $\epsilon_L \sim N(0, \sigma_L^2)$

Thus the following relationships hold as well, where $\rho = \frac{\sigma_{YL}}{\sigma_Y \sigma_L}$ and $\text{Cov}(\epsilon_Y, \epsilon_L) = \sigma_{YL}$.

• $E[Y|A, L, C] = \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_Y} (L - \alpha_0 - \alpha_1 A - \alpha_2 C)$

• $\text{Var}[Y|A, L, C] = \sigma_Y^2 - \rho^2 \sigma_Y^2$

• $E[L|A, Y, C] = \alpha_0 + \alpha_1 A + \alpha_2 C + \frac{\sigma_{YL}}{\sigma_Y} (Y - \beta_0 - \beta_1 A - \beta_2 C)$

• $\text{Var}[L|A, Y, C] = \sigma_L^2 - \rho^2 \sigma_L^2$

Finally, as stated above, we require our binary treatment, $A$, to be generated such that the following two relationships hold,

• $\text{logit}[P(A = 1|L, C, Y)] = \gamma_0 + \gamma_1 L + \gamma_2 C + \gamma_3 Y$

• $\text{logit}[P(A = 1|L, C)] = \lambda_0 + \lambda_1 L + \lambda_2 C$

We do this so that we have a closed form for the standard propensity score $P(A = 1|L, C)$ and well as $\text{logit}[P(A = 1|L = 0, C, Y = 0)] = \gamma_0 + \gamma_2 C$ which will allow for ease of implementation.

This is not a simple simulation as we need the conditional distribution of $A$ to be logistic conditional on $L, C$ and $Y$ as well as logistic conditional on $L = 0, C$ and $Y = 0$. To achieve this, we will take the joint distribution $f(A, Y, L|C)$ and integrate over $Y$. We also need to find $Pr(A = 1|C)$ in order to actually generate the data. Again we can integrate $f(A, Y, L|C)$ over $Y$ and $L$ to get to this probability.

A.3.3 Data Generation

We will first generate $C$. Then $A$ as a function of $C$, then $Y$ and $L$ from a joint normal. We need to find $f(A|L, C, Y)$ as a function of only $C$. To do this we will get the joint distribution of $f(Y, A, L|C)$ and marginalize over $Y$ and $L$,

$$f(Y, A, L|C) = \frac{1}{K} f(Y|A = 0, L = 0, C) \text{OR}(Y, A|L = 0, C) f(A|Y = 0, L = 0, C)$$
Next we need to calculate the odds ratios

\[
\begin{align*}
\times \ OR(A, L|Y = 0, C)f(L|A = 0, Y = 0, C)OR(L, Y|A = 0, C)\theta(Y, A, L),
\end{align*}
\]

where

\[
K = \sum_A \int_Y \int_{L|C} f(y, A, l|C)dl\ dy
\]

and

\[
\theta(Y, A, L) = \frac{OR(Y, A|L, C)}{OR(Y, A|L = 0, C)} \frac{OR(L, A|Y = 0, C)}{OR(L, A|Y, C)} \frac{OR(Y, L|A, C)}{OR(Y, L|A = 0, C)} = 1.
\]

Next we need to calculate the odds ratios

\[
\begin{align*}
OR(Y, L|C, A = 0) &= \frac{f(Y|L, A, C) \times f(Y = 0|L = 0, A, C)}{f(Y = 0|L, A, C) \times f(Y|L = 0, A, C)} \\
&= \exp \left\{ \frac{-1}{2(\sigma_Y^2 - \rho^2 \sigma_L^2)} \left[ (Y - \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right])^2 \\
&\quad + \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (-\alpha_0 - \alpha_1 A - \alpha C) \right)\right]^2 \\
&\quad - \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right) \right] \\
&\quad - \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (-\alpha_0 - \alpha_1 A - \alpha C) \right)^2 \\
&\quad - \left( Y - \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right] \right)^2 \right) \\
&= \exp \left\{ \frac{-1}{2(\sigma_Y^2 - \rho^2 \sigma_L^2)} \left[ (Y - 2Y \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right])^2 \\
&+ \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (-\alpha_0 - \alpha_1 A - \alpha C) \right)^2 \\
&- \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right)^2 \\
&+ \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (-\alpha_0 - \alpha_1 A - \alpha C) \right)^2 \right] \right\} \\
&= \exp \left\{ \frac{-1}{2(\sigma_Y^2 - \rho^2 \sigma_L^2)} \left[ (Y - 2Y \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right])^2 \\
&- \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right)^2 \right] \right\} \\
&+ 2Y \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (-\alpha_0 - \alpha_1 A - \alpha C) \right] \\
&= \exp \left\{ \frac{-1}{2(\sigma_Y^2 - \rho^2 \sigma_L^2)} \left[ (Y - 2Y \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right])^2 \\
&- \left( \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_1 A - \alpha C) \right)^2 \right] \right\} \\
&+ 2Y \left[ \beta_0 + \beta_1 A + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (-\alpha_0 - \alpha_1 A - \alpha C) \right] \\
&= 1.
\end{align*}
\]
\[
\begin{align*}
&= \exp \left\{ \frac{1}{\sigma_Y^2 - \rho^2 \sigma_Y^2} \left( \frac{\sigma_{YL}}{\sigma_L^2} \right) \right\} \\
&= \exp \left\{ \frac{\rho}{(1 - \rho^2) \sigma_Y \sigma_L} YL \right\}.
\end{align*}
\]

Therefore,

\[ OR(Y, L|C, A = 0) = \exp\{\omega_{12} YL\} \]

where

\[ \omega_{12} = \frac{\rho}{(1 - \rho^2) \sigma_Y \sigma_L}. \]

Similarly,

\[ OR(Y, A|C, L = 0) = \frac{f(Y|L, A, C) \times f(Y = 0|L, A = 0, C)}{f(Y = 0|L, A, C) \times f(Y|L, A = 0, C)} \]

\[ = \exp \left\{ \frac{\beta_1 - \frac{\sigma_{YL}}{\sigma_Y^2} \alpha_1}{(1 - \rho^2) \sigma_Y^2} AY \right\} \]

\[ = \exp \{\omega_{13} AY\} \]

where

\[ \omega_{13} = \frac{\beta_1 - \frac{\sigma_{YL}}{\sigma_Y^2} \alpha_1}{(1 - \rho^2) \sigma_Y^2}. \]

And

\[ OR(L, A|C, Y = 0) = \frac{f(L|Y, A, C) \times f(L = 0|Y, A = 0, C)}{f(L = 0|Y, A, C) \times f(L|Y, A = 0, C)} \]

\[ = \exp \left\{ \alpha_1 - \frac{\sigma_{YL}}{\sigma_Y^2} \beta_1 \right\} \]

\[ = \exp \{\omega_{23} AL\} \]

where

\[ \omega_{23} = \frac{\alpha_1 - \frac{\sigma_{YL}}{\sigma_Y^2} \beta_1}{(1 - \rho^2) \sigma_Y^2}. \]

Alternatively, we could find \( OR(L, A|C, Y = 0) \) as follows,

\[ OR(L, A|C, Y = 0) = \frac{f(A|Y, L, C) \times f(A = 0|Y, L = 0, C)}{f(A = 0|Y, L, C) \times f(A|Y, L = 0, C)} \]
Thus we know that \( \gamma_1 = \frac{\alpha_L - \sigma_L}{\sigma_L} \beta_1 \). Similarly, by using looking at \( OR(Y, A|C, L = 0) \) we can find that \( \gamma_3 = \frac{\beta_2 - \sigma_L}{\sigma_L} \). We don’t ever actually use these coefficients in our estimator, but they are useful for checking that the data has been generated correctly.

Next we derive a closed form for the probability of \( A \) that only depends on \( C \).

The numerator of our intended probability of \( A \) is

\[
\text{numerator} = \int \int f(y, A, l| C) \, dy \, dl = \int \int f(y|A = 0, L = 0, C) f(\theta|A = 0, Y = 0, L = 0, C) f(A|Y = 0, L = 0, C) \, dy \, dl
\]

\[
= \int \int f(y|A = 0, L = 0, C) OR(y, A|L = 0, C) f(A|Y = 0, L = 0, C) \, dy \, dl
\]

\[
\times OR(A, l|Y = 0, C) f(l|A = 0, Y = 0, L = 0, C) OR(l, y|A = 0, C) \theta(y, A, l) \, dy \, dl
\]

\[
= \left( \frac{1}{1 + e^{-\gamma_0 - \gamma_1 C}} \right)^{A} \left( 1 - \frac{1}{1 + e^{-\gamma_0 - \gamma_1 C}} \right)^{1-A}
\]

\[
\times \int \exp \left[ \omega_{21C} N_L \left( \alpha_0 + \alpha_2 C - \frac{\sigma_{2L}}{\sigma_L} (\beta_0 + \beta_2 C) \right) \right] (\beta_0 + \beta_2 C - \frac{\sigma_{2L}}{\sigma_L} (\alpha_0 + \alpha_2 C)) \sigma_L^2 - \rho^2 \sigma_L^2 \, dy \, dl
\]

\[
= \left( \frac{1}{1 + e^{-\gamma_0 - \gamma_1 C}} \right)^{A} \left( 1 - \frac{1}{1 + e^{-\gamma_0 - \gamma_1 C}} \right)^{1-A}
\]

\[
\times \int \frac{1}{\sqrt{2\pi} \sigma_{1L} \sigma_{2L}} \exp \left[ \frac{-1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right)^2 \right] \, dy \, dl
\]

\[
\times \exp \left[ \omega_{21A} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right] \exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \beta_0 + \beta_2 C \right)^2 \right]
\]

\[
\times \exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \omega_{12A} + \omega_{12L} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right) \sigma_{1L}^2 \right]
\]

\[
\times \exp \left[ \omega_{13A} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right] \exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \beta_0 + \beta_2 C \right)^2 \right]
\]

\[
\times \exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \omega_{13A} + \omega_{13L} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right) \sigma_{1L}^2 \right]
\]

\[
\times \int \exp \left[ \frac{-1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right] \sigma_{1L}^2 \right]
\]

\[
\left( \frac{1}{1 + e^{-\gamma_0 - \gamma_1 C}} \right)^{A} \left( 1 - \frac{1}{1 + e^{-\gamma_0 - \gamma_1 C}} \right)^{1-A}
\]

\[
\times \frac{1}{\sqrt{2\pi} \sigma_{1L} \sigma_{2L}} \exp \left[ \frac{-1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right)^2 \right] \, dy \, dl
\]

\[
\exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \omega_{13A} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right) \sigma_{1L}^2 \right]
\]

\[
\exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \omega_{13A} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right) \sigma_{1L}^2 \right]
\]

\[
\exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \omega_{13A} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right) \sigma_{1L}^2 \right]
\]

\[
\exp \left[ \frac{1}{2} \left( \frac{1}{\sigma_{1L}^2} - \rho^2 \sigma_{1L}^2 \right) \left( \omega_{13A} \left( \beta_0 + \beta_2 C - \frac{\sigma_{1L}}{\sigma_L} (\alpha_0 + \alpha_2 C) \right) \right) \sigma_{1L}^2 \right]
\]
\[ l \left( \frac{1}{2} \omega_{12}^2 \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \right) \, dl \]

\[ = \left( \frac{1}{1 + e^{-\gamma - \gamma_2^C}} \right)^A \left( 1 - \frac{1}{1 + e^{-\gamma - \gamma_2^C}} \right)^{1-A} \times \exp \left[ \frac{1}{2 \gamma_2^C} \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \omega_{12}^2 A^2 \right] \]

\[ \times \exp \left[ \omega_{13} A \left( \beta_0 + \beta_2 C - \frac{\gamma L}{\gamma} (\alpha_0 + \alpha_2 C) \right) \right] \]

\[ \times \int \exp \left\{ -\frac{1}{2} \omega_{12}^2 \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \right\} \left[ -\left( 1 + \frac{1}{2} \omega_{12}^2 \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \right) \cdot \left[ L + \frac{\alpha_0 + \alpha_2 C - \frac{\gamma L}{\gamma} (\beta_0 + \beta_2 C)^2}{[1 + \frac{1}{2} \omega_{12}^2 (\sigma^2_Y - \rho^2 \sigma^2_Y)]} \right] \right] \]

\[-2 \int \omega_{13} A \omega_{12} \left[ \gamma_2^C + \omega_{12} \left( \beta_0 + \beta_2 C - \frac{\gamma L}{\gamma} (\alpha_0 + \alpha_2 C) \right) + (\sigma^2_Y - \rho^2 \sigma^2_Y) \right] \right] \omega_{12}^2 A \]

\[ + \alpha_0 + \alpha_2 C - \frac{\gamma L}{\gamma} (\beta_0 + \beta_2 C) \right) \right] \right\} \, dl \]

\[ = \left( \frac{1}{1 + e^{-\gamma - \gamma_2^C}} \right)^A \left( 1 - \frac{1}{1 + e^{-\gamma - \gamma_2^C}} \right)^{1-A} \times \exp \left[ \frac{1}{2 \gamma_2^C} \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \omega_{12}^2 A^2 \right] \]

\[ \times \exp \left[ \omega_{13} A \left( \beta_0 + \beta_2 C - \frac{\gamma L}{\gamma} (\alpha_0 + \alpha_2 C) \right) \right] \times \exp \left[ \frac{1}{2 (\sigma^2_Y - \rho^2 \sigma^2_Y)} \left[ -\left( \alpha_0 + \alpha_2 C - \frac{\gamma L}{\gamma} (\beta_0 + \beta_2 C)^2 \right) \right] \right] \]

\[ \times \exp \left\{ -\frac{1}{2} \omega_{12}^2 \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \right\} \left[ L + \frac{\alpha_0 + \alpha_2 C - \frac{\gamma L}{\gamma} (\beta_0 + \beta_2 C)^2}{[1 + \frac{1}{2} \omega_{12}^2 (\sigma^2_Y - \rho^2 \sigma^2_Y)]} \right] \]

\[ \times \int \omega_{13} A \omega_{12} \left[ \gamma_2^C + \omega_{12} \left( \beta_0 + \beta_2 C - \frac{\gamma L}{\gamma} (\alpha_0 + \alpha_2 C) \right) + (\sigma^2_Y - \rho^2 \sigma^2_Y) \right] \omega_{12}^2 A \]

\[ + \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \left[ \omega_{13} A \omega_{12} \left( \beta_0 + \beta_2 C - \frac{\gamma L}{\gamma} (\alpha_0 + \alpha_2 C) \right) + (\sigma^2_Y - \rho^2 \sigma^2_Y) \omega_{12} A \right] \]

\[ + \left( \sigma^2_Y - \rho^2 \sigma^2_Y \right) \left[ \omega_{13} A \omega_{12} \left( \beta_0 + \beta_2 C - \frac{\gamma L}{\gamma} (\alpha_0 + \alpha_2 C) \right) + (\sigma^2_Y - \rho^2 \sigma^2_Y) \right] \omega_{12} A \]
We can now simulate our data as follows:

\[
\begin{align*}
\text{Next, generate } Y, L \\
\text{The probability of interest is then}
\end{align*}
\]

And

\[
\begin{align*}
g(A, C)|_{A=0} &= \left(1 - \frac{1}{1 + e^{-\gamma_0 - \gamma_2 C}}\right) \frac{1}{\sqrt{1 + \frac{1}{2} \omega_{12}^2 (\sigma_Y^2 - \rho^2 \sigma_Y^2)}} \\
&\times \exp \left[-\frac{(\omega_{13})}{2(\sigma_L^2 - \rho^2 \sigma_L^2)} \frac{1}{\sqrt{1 + \frac{1}{2} \omega_{12}^2 (\sigma_Y^2 - \rho^2 \sigma_Y^2)}} \frac{1}{\sqrt{1 + \frac{1}{2} \omega_{12}^2 (\sigma_Y^2 - \rho^2 \sigma_Y^2)}} \right] \\
&\times \exp \left[\frac{\omega_{13}}{2(\sigma_L^2 - \rho^2 \sigma_L^2)} \left((\alpha_0 + \alpha_2 C - \frac{\sigma_Y}{\sigma_L} (\omega_{13} + \omega_{12}) (\beta_0 + \beta_2 C - 1) = 1) \right) \right] \\
&\times \left[1 + \frac{1}{2} \omega_{12}^2 (\sigma_Y^2 - \rho^2 \sigma_Y^2) \right]^{-2}
\end{align*}
\]

The probability of interest is then

\[
P(A = 1| C) = \frac{g(A, C)|_{A=1}}{g(A, C)|_{A=0} + g(A, C)|_{A=1}} = p_A.
\]

This is a way of writing \(P(A = 1|L, C, Y)\) that only depends on \(C\). This can be used to generate \(A\).

Next, generate \((Y, L) \sim \text{MVN}(\begin{pmatrix} \beta_0 + \beta_1 A + \beta_2 C \\ \alpha_0 + \alpha_1 A + \alpha_2 C \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YL} \\ \sigma_{YL} & \sigma_L^2 \end{pmatrix})\)

We can now simulate our data as follows:

- Generate \(C\) according to a binomial distribution with \(P(C = 1) = p_C\)
• Generate $A$ according to binomial with probability $p_A$

• Generate $(Y,L) \sim \text{MVN} \left( \begin{pmatrix} \beta_0 + \beta_1 A + \beta_2 C \\ \alpha_0 + \alpha_1 A + \alpha_2 C \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YL} \\ \sigma_{YL} & \sigma_L \end{pmatrix} \right)$

### A.3.4 Additional relationship

We are also interested in finding

$$\logit[P(A = 1|L, C)] = \lambda_0 + \lambda_1 L + \lambda_2 C$$

so that we can know the true model for the standard propensity score.

This can be achieved by finding a closed form for $f(A,Y|L,C)$ and integrating over $Y$ to get $f(A|L,C)$,

$$\frac{1}{C} \int_y f(Y,A|L,C)dy = \frac{1}{C} \int_y f(Y|A=0, L, C) OR(Y, A|L, C) f(A|Y = 0, L, C)dy$$

$$= \frac{1}{C} \left( \frac{1}{1 + e^{-\gamma_0 - \gamma_1 L - \gamma_2 C}} \right)^A \left( 1 - \frac{1}{1 + e^{-\gamma_0 - \gamma_1 L - \gamma_2 C}} \right)^{1-A}$$

$$\times \int_y \exp(\omega_{13}AY) N_Y \left( \beta_0 + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L^2} (L - \alpha_0 - \alpha_2 C), \sigma_Y^2 - \rho^2 \sigma_Y^2 \right) dy$$

$$= \frac{1}{C} \left( \frac{1}{1 + e^{-\gamma_0 - \gamma_1 L - \gamma_2 C}} \right)^A \left( 1 - \frac{1}{1 + e^{-\gamma_0 - \gamma_1 L - \gamma_2 C}} \right)^{1-A}$$

$$\times \exp \left[ A\omega_{13} \left( \beta_0 + \beta_2 C + \frac{\sigma_{YL}}{\sigma_L} (L - \alpha_0 - \alpha_2 C) \right) \right]$$

$$\times \exp \left[ \frac{1}{2} A^2 \omega_{13}^2 \left( \sigma_Y^2 - \rho^2 \sigma_Y^2 \right) \right]$$

$$= f(A|L,C)$$

where $C = \sum_{a=0}^1 \int_y f(Y|A = 0, L, C) OR(Y, A|L, C) f(A|Y = 0, L, C)dy$. We see that

$$\logit[P(A = 1|L, C)] = \lambda_0 + \lambda_1 L + \lambda_2 C,$$

where

$$\lambda_0 = \gamma_0 + \omega_{13} \beta_0 - \omega_{13} \frac{\sigma_{YL}}{\sigma_L^2} \alpha_0 + \frac{1}{2} \omega_{13}^2 (\sigma_Y^2 - \rho^2 \sigma_Y^2),$$

$$\lambda_1 = \gamma_1 + \omega_{13} \frac{\sigma_{YL}}{\sigma_L^2},$$

$$\lambda_2 = \gamma_2 + \omega_{13} \beta_2 - \omega_{13} \frac{\sigma_{YL}}{\sigma_L^2} \alpha_2.$$

This will give us the true coefficients of interest.
In our simulation we chose the following values:

$$\alpha_0 = -0.15$$
$$\alpha_1 = 0.215$$
$$\alpha_2 = 0.14$$
$$\beta_0 = 0.2$$
$$\beta_1 = 0.38$$
$$\beta_2 = 0.3$$
$$\sigma_Y^2 = 0.51$$
$$\sigma_{YL}^2 = 0.43$$
$$\rho = \frac{\sigma_{YL}}{\sigma_L \sigma_Y} = 0.448$$
$$\eta_0 = \alpha_0 - \beta_0 \frac{\sigma_{YL}}{\sigma_Y^2} = -0.2323529$$
$$\eta_1 = \alpha_2 - \beta_2 \frac{\sigma_{YL}}{\sigma_Y^2} = 0.01647059$$
$$\eta_2 = \frac{\sigma_{YL}}{\sigma_Y^2} = 0.4117647$$
$$\eta_3 = \alpha_1 - \beta_1 \frac{\sigma_{YL}}{\sigma_Y^2}$$
$$\omega_{12} = \frac{\rho}{(1-\rho^2)\sigma_L \sigma_Y} = 1.19863$$
$$\omega_{23} = \frac{\beta_1 - \frac{\sigma_{YL}}{\sigma_Y^2} \alpha_1}{(1-\rho^2)\sigma_Y^2} = 0.1703767$$
$$\omega_{13} = \frac{\frac{\sigma_{YL}}{\sigma_Y^2}}{(1-\rho^2)\sigma_Y^2} = 0.6749429$$
$$\gamma_0 = -0.7$$
$$\gamma_1 = \omega_{23} = 0.1703767$$
$$\gamma_2 = -0.2$$
$$\gamma_3 = \omega_{13} = 0.6749429$$
$$\lambda_0 = \gamma_0 + \omega_{13} \beta_0 - \omega_{13} \frac{\sigma_{YL}}{\sigma_Y^2} \alpha_0 + \frac{1}{2} \omega_{13}^2 (\sigma_Y^2 - \rho^2 \sigma_Y^2) = -0.4227633$$
$$\lambda_1 = \gamma_1 + \omega_{13} \frac{\sigma_{YL}}{\sigma_Y^2} = 0.5$$
$$\lambda_2 = \gamma_2 + \omega_{13} \beta_2 - \omega_{13} \frac{\sigma_{YL}}{\sigma_Y^2} \alpha_2 = 0.3563356$$

Moreover,

$$\theta = E[Y_{a=1} | A = 1] = \beta_1 - \alpha_1 \frac{\sigma_{YL}}{\sigma_Y^2}$$
$$E[L | A = 0, C, Y, R = 1] = \eta_0 + \eta_1 C + \eta_2 Y$$
A.4 Asymptotic Distribution of Estimator

A.4.1 For $\eta$:

From $\mu_L = E[L|A = 0, C, Y, R = 1; \eta] = \eta_0 + \eta_1 C + \eta_2 Y$,

$$P_n(U(\hat{\eta})) = 0$$

For $L$ normal,

$$\begin{pmatrix} P_n \left( \frac{1}{\sigma^2}(L_i - \mu_L) \right) \\ P_n \left( \frac{1}{2\sigma^2} + \frac{1}{2\sigma^2}(L_i - \mu_L)^2 \right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

A.4.2 For $\lambda$:

We denote the propensity score as $p(W; \Lambda) = Pr(A = 1|W; \Lambda)$. We consider $\tilde{S}(\hat{\Lambda}, \hat{\eta})$ an individual contribution to the estimated observed score equation for $\Lambda$. Then the unbiased estimating equations for $\hat{\Lambda}$ can be calculated from the expected score. Namely,

$$P_n \left[ (1 - R_i)E \left[ \tilde{S}(\hat{\Lambda}, \hat{\eta})|C, Y, A = 0, R = 1; \eta \right] + R_i \tilde{S}(\hat{\Lambda}, \hat{\eta}) \right] = 0,$$

that is,

$$P_n \left( \tilde{S}(\hat{\Lambda}, \hat{\eta}) \right) = 0.$$
A.4.3 For $\Psi$:

$$\mathbb{P}_n \left( I(A_i = 0) \frac{P(A = 1|C, Y_0; \hat{n}, \hat{\lambda})}{P(A = 0|C, Y_0; \hat{n}, \hat{\lambda})} (Y_i - \hat{\Psi}) \right) = 0.$$ 

If we let $Z \left( \hat{\Psi}, \hat{n}, \hat{\lambda} \right) = I(A_i = 0) \frac{P(A=1|C_i,Y_0;\hat{n}_i,\hat{\lambda})}{P(A=0|C_i,Y_0;\hat{n}_i,\hat{\lambda})} (Y_i - \hat{\Psi})$ and $\hat{\Phi} = \left( \hat{n}, \hat{\lambda}, \hat{\Psi} \right)$. Then,

$$0 = \mathbb{P}_n \left( V \left( \hat{n}, \hat{\lambda}, \hat{\Psi} \right) \right) = \mathbb{P}_n \left( V \left( \hat{\Phi} \right) \right) = \left( \begin{array}{c} \mathbb{P}_n \left( U_\eta \left( \hat{\mu}_L \right) \right) \\ \mathbb{P}_n \left( S \left( \hat{\Lambda} \right) \right) \\ \mathbb{P}_n Z \left( \hat{\Psi} \right) \end{array} \right).$$

This is simply solving a system of score equations.

Then, under standard regularity conditions,

$$\sqrt{n}(\hat{\Phi} - \Phi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{dV}{d\Phi} \right]^{-1} V(\Phi) + op(1).$$

Therefore, a consistent estimator of the asymptotic variance is

$$\left[ \mathbb{P}_n \frac{dV}{d\Phi} \right]^{-1} \mathbb{P}_n \left[ V(\hat{\Phi})V'(\hat{\Phi}) \right] \left[ \mathbb{P}_n \left( \frac{dV}{d\Phi} \right) \right]^{-1}.$$

B Proofs and results for Chapter 2

B.1 Influence Function Derivation

Assumptions:

1. Consistency: $Y = Y_A$ almost surely;

2. No unmeasured confounding: $A \perp Y_0|W$, where $W = (C, L)$;

Then, $\Psi = E[Y_0|A = 1] = E[E[Y|A = 0, W]|A = 1] = \int f(W|A = 1) \int yf(y|A = 0, W)dydw.$

Consider a function of the observed data, $O$, $F_t(O)$ such that $F_0(O) = F(O)$. Then $\Psi_t = \Psi(F_t) = \int f_t(W|A = 1) \int yf_t(y|A = 0, W)dydw.$

Our goal is to write $\frac{d\Psi_t}{dt}$ as $E[\iota \times S(O)]$ where $\iota$ is the influence function and $S(O)$ is the score
\[ S(O) = \frac{d}{dt} \log f_t(O) \]
\[ = \frac{d}{dt} \left[ \log f_t(Y|A,W) + \log f_t(A|W) + \log f_t(W) \right] \]
\[ = \frac{d}{dt} \left[ \log f_t(Y|A,W) + \log f_t(W|A) + \log f_t(A) \right]. \]

Then,
\[
\frac{d \Psi_t}{dt} = \int \int \frac{d}{dt} f_t(W|A = 1) y f(Y|A = 0,W) \right) dwdy \]
\[ + \frac{d}{dt} f_t(Y|A = 0,W) y f(W|A = 1) dwdy. \]

Let us examine each term separately. Denote:

- **Term 1** = \[ \int \int \frac{d}{dt} f_t(Y|A = 0,W) f(W|A = 1) y dwdy \]
- **Term 2** = \[ \int \int \frac{d}{dt} f_t(W|A = 1) y f(Y|A = 0,W) dwdy. \]

We will drop the subscript for ease of reading.

**Term 1**

\[
\text{Term 1} = \int \int \frac{d}{dt} f_t(Y|A = 0,W) f(W|A = 1) y dwdy \\
= \int \int \int \int \frac{I(A = 0)}{f(A = 0|W)} \frac{d}{dt} f_t(Y|A,W) \frac{f(W|A = 1)}{f(W)} \frac{P(A = 1)}{f(O) dwdy d\mu_a} \\
\times f(Y|A,W) f(A|W) f(W) dwdy d\mu_a \\
= \int \int \int \int \frac{I(A = 0)}{f(A = 0|W)} \frac{f(W,A = 1)}{f(W) P(A = 1)} \frac{S(Y|A,W) f(O) dwdy d\mu_a}{f(W)} \\
= \frac{\int \int \int \int \frac{I(A = 0)}{f(A = 0|W)} \frac{f(A = 1|W) f(W)}{f(W) P(A = 1)} \frac{S(Y|A,W) f(O) dwdy d\mu_a}{f(W)}}{f(O) dwdy d\mu_a} \\
= \frac{\int \int \int \int \frac{I(A = 0)}{f(A = 0|W)} \frac{f(A = 1|W) f(W)}{f(W) P(A = 1)} \frac{S(Y|A,W) f(O) dwdy d\mu_a}{f(W)}}{f(O) dwdy d\mu_a} \\
\times f(O) [S(Y|A,W) + S(A|W) + S(W)] dwdy d\mu_a \]

where we note the final equality holds from adding and subtracting functions of \( A \) and \( W \) which
are orthogonal to $E[Y|A = 0, W]$ and therefore disappear in the inner product.

Therefore, the influence function for Term 1 is

$$
\iota_1 = (Y - E[Y|A = 0, W]) \frac{I(A = 0)}{f(A = 0|W)} \frac{f(A = 1|W)}{P(A = 1)}.
$$

**Term 2**

$$
\begin{align*}
\text{Term 2} = & \int \int \frac{d}{dt} f_t(W|A = 1)yf(Y|A = 0, W)dwdy \\
= & \int \int \frac{d}{dt} \frac{f_t(W|A)}{f(W|A)} E[Y|A = 0, W] \frac{I(A = 1)}{f(A)} f(A) f(W|A) dwd\mu_a \\
= & \int \int \int [S(Y|W, A) + S(W|A)] f(Y|W, A) f(W|A) E[Y|A = 0, W] \frac{I(A = 1)}{f(A)} dwdy d\mu_a \\
= & \int \int \int [S(Y|W, A) + S(W|A)] f(0) \frac{I(A = 1)}{f(A)} \left\{ E[Y|A = 0, W] - \Psi \right\} dwdy d\mu_a.
\end{align*}
$$

Therefore the influence function for Term 2 is

$$
\iota_2 = (E[Y|A = 0, W] - \Psi) \frac{I(A = 1)}{f(A)}.
$$

Now we can see that

$$
\frac{d\Psi_t}{dt} = E[t \times S(O)]
$$

where

$$
\iota = \frac{I(A = 0)}{P(A = 1)} \frac{P(A = 1|W)}{P(A = 0|W)} (Y - E[Y|A = 0, W]) + \frac{I(A = 1)}{P(A = 1)} (E[Y|A = 0, W] - \Psi).
$$

If one were to use the efficient influence function as an estimating equation for $\Psi$, set it equal to zero and, solve for $\Psi$, then we would have

$$
0 = \frac{I(A = 0)}{P(A = 1)} \frac{P(A = 1|W)}{P(A = 0|W)} (Y - E[Y|A = 0, W]) + \frac{I(A = 1)}{P(A = 1)} (E[Y|A = 0, W] - \Psi).
$$

This equation implies that

$$
\psi I(A = 1) = I(A = 0) \frac{P(A = 1|W)}{P(A = 0|W)} Y
$$

76
Taking the expectation of both sides gives our final result as the second and third terms will cancel and thus

\[ \Psi = \frac{1}{P(A = 1)} E \left[ (1 - A) \frac{P(A = 1|W)}{P(A = 0|W)} Y \right]. \]

**B.2 Proof of Double Robustness for Full Data**

If one were to use the efficient influence function, (2.2), as an estimating equation for \( \Psi \), one would need to estimate the nuisance functions \( f(A|W) \) and \( m(Y|A = 0, L, C) \). The resulting estimator is double robust for \( \Psi \) in that it will be consistent provided we correctly specify a model for the propensity score, \( f(A|W) \), or the outcome model, \( m(Y|A = 0, L, C) \), but not necessarily both.

**Case 1- \( f^*(A|W) \) is wrong but \( m^*(Y|A = 0, L, C) \) is correct** Then

\[
E[IF] = E \left[ \frac{I(A = 0)}{P(A = 1)} \frac{f^*(A = 1|W)}{f^*(A = 0|W)} (Y - E[Y|A = 0, W]) + \frac{I(A = 1)}{f(A = 1)} (E[Y|A = 0, W] - \Psi)|A, W \right]
\]

\[
= E \left[ \frac{I(A = 1)}{f(A = 1)} (E[Y|A = 0, W] - \Psi) \right]
\]

\[
= E \left[ \frac{I(A = 1)}{f(A = 1)} (E[Y|A = 0, W]|A = 1) - \Psi \right]
\]

\[
= E \left[ \frac{I(A = 1)}{f(A = 1)} (\Psi - \Psi) \right]
\]

\[
= 0.
\]

**Case 2- \( f(A|W) \) is correct but \( m^*(Y|A = 0, L, C) \) is wrong** Here we let \( E^*[Y|A = 0, W] = b^*(W) \) and \( f(A|W) = p(W) \) for ease of notation. Then

\[
E[IF|W] = E \left[ \frac{I(A = 0)}{P(A = 1)} \frac{p(W)}{1 - p(W)} (Y - b^*(W)) + \frac{I(A = 1)}{f(A = 1)} (b^*(W) - \Psi)|Y, A, W \right]
\]

\[
= \frac{1}{P(A = 1)} E \left[ (1 - A) \frac{p(W)}{1 - p(W)} (E[Y|A = 0] - b^*(W)) + A(b^*(W) - \Psi)|Y, A, W \right]
\]

\[
= \frac{1}{P(A = 1)} E \left[ p(W) (E[Y|A = 0] - b^*(W)) + p(W) (b^*(W) - \Psi)|Y, A, W \right]
\]

\[
= \frac{1}{P(A = 1)} E \left[ p(W)(E[Y|A = 0] - \Psi)|Y, A, W \right]
\]
\[
= \frac{1}{P(A = 1)} E \left[ p(W) (E[Y|A = 0, W] - \Psi) \right]
\]
\[
= \frac{1}{P(A = 1)} E \left[ p(W) (E \left[ E[Y|A = 0, W] | A = 1 \right] - \Psi) \right]
\]
\[
= \frac{1}{P(A = 1)} E \left[ \Psi - \Psi \right]
\]
\[
= 0.
\]

B.3 Missing Data Influence Function

Let
\[
\tau_{\text{Full}}(O) = \frac{I(A = 0)}{P(A = 1)} \frac{f(A = 1|W)}{f(A = 0|W)} (Y - E[Y|A = 0, W]) + \frac{I(A = 1)}{f(A = 1)} (E[Y|A = 0, W] - \Psi).
\]

Then the influence function for the missing data problem is
\[
\tau_{\text{Miss}} = \frac{R}{\pi^*} \tau_{\text{Full}}(O) - \left( \frac{R}{\pi^*} - 1 \right) E^*[\tau_{\text{Full}}(O)|O]
\]
\[
= \frac{R}{\pi^*} (\tau_{\text{Full}}(O) - E^*[\tau_{\text{Full}}(O)|O]) + E^*[\tau_{\text{Full}}(O)|O],
\]

where \( \pi = P(R = 1|A, Y, C) \).

We, in theory, need to correctly specify:

\( p = P(A = 1|W) \) or \( m(Y|A, C, L) \)

and

\( \pi = P(R = 1|A, Y, C) \) or \( t(L|A, C, Y) \).

Notes: If \( t \) is correctly specified, then \( E[\tau_{\text{Full}}(O)] = E[E^*[\tau_{\text{Full}}(O)|O]] \). The \( t \) shows up in the right hand side (RHS) of the equation.

If \( p \) or \( m \) are correct then we saw above that \( E[\tau_{\text{Full}}(O)] = 0 \) and in turn, \( E[\tau_{\text{Full}}(O)] = E[E[\tau_{\text{Full}}(O)|O]] \)

From the above expressions for \( \tau_{\text{Miss}} \), let:

\[
\text{Term 1} = \frac{R}{\pi^*} \tau_{\text{Full}}(O)
\]
\[
\text{Term 2} = \left( \frac{R}{\pi^*} - 1 \right) E^*[\tau_{\text{Full}}(O)|O]
\]
\[
\text{Term 3} = \frac{R}{\pi^*} (\tau_{\text{Full}}(O) - E^*[\tau_{\text{Full}}(O)|O])
\]
\[
\text{Term 4} = E^*[\tau_{\text{Full}}(O)|O]
\]
Case 1 - \( p \) and \( \pi \) correct. i.e. \( p^* = p \) and \( \pi^* = \pi = P(R = 1|A, Y, C) \)

\[
E[\text{Term 1}] = E\left[ \frac{R}{\pi^*} t_{\text{Full}}(O) \right] \\
= E\left[ \frac{R}{\pi} t_{\text{Full}}(O) \right] \\
= E\left[ E\left[ \frac{R}{\pi} t_{\text{Full}}(O)|Y, A, C \right] \right] \\
= E\left[ \frac{P(R = 1|Y, A, C)}{P(R = 1|Y, A, C)} E[t_{\text{Full}}(O)|Y, A, C] \right] \\
= E\left[ E[t_{\text{Full}}(O)|O] \right] \\
= 0
\]

And

\[
E[\text{Term 2}] = E\left[ \left( \frac{R}{\pi^*} - 1 \right) E^*[t_{\text{Full}}(O)|O] \right] \\
= E\left[ \left( \frac{R}{\pi} - 1 \right) E^*[t_{\text{Full}}(O)|O]|A, Y, C \right] \\
= E\left[ \left( \frac{P(R = 1|Y, A, C)}{P(R = 1|Y, A, C)} - 1 \right) E^*[t_{\text{Full}}(O)|O]|A, Y, C \right] \\
= E\left[ 0 \times E^*[t_{\text{Full}}(O)|O]|A, Y, C \right] \\
= 0
\]

Case 2 - \( m \) and \( \pi \) correct. Same as above. Term 2 only relies on \( \pi \) being correct. Term 1 relies on \( E[IF_{\text{Full}}(O)] = 0 \) when \( m \) is correct.

Case 3 - \( p \) and \( t \) correct. i.e. \( p^* = p \) and \( E^*[IF_{\text{Full}}(O)|O] = E[IF_{\text{Full}}(O)|O] \)

\[
E[\text{Term 3}] = E\left[ \frac{R}{\pi^*} (t_{\text{Full}}(O) - E^*[t_{\text{Full}}(O)|O]) \right] \\
= E\left[ \frac{R}{\pi} (t_{\text{Full}}(O) - E[t_{\text{Full}}(O)|O]) \right] \\
= E\left[ \frac{P(R = 1|O)}{P^*(R = 1|O)} E[(t_{\text{Full}}(O) - E[t_{\text{Full}}(O)|O])|O] \right] \\
= E\left[ \frac{P(R = 1|O)}{P^*(R = 1|O)} \times 0 \right]
\]
\begin{align*}
E[\text{Term 4}] &= E[E^*[t_{\text{Full}}(O)|O]] \\
&= E[E^*[t_{\text{Full}}(O)|O]] \\
&= E[t_{\text{Full}}(O)] \\
&= 0
\end{align*}

\textbf{Case 4 - m and t correct}  \enspace \text{Follows similarly as the above case.}

\subsection*{B.4 Details on Failure of Multiple Robustness}

The problem with first method is a problem with model congeniality

The first MR method we established was one wherein we must correctly specify:

\[ p = P(A = 1|L, C) \text{ or } m(Y|A, L, C) \]

and

\[ \pi = P(R = 1|A, Y, C) \text{ or } t(L|A, C, Y) \]

The problem with this is that it assumes we estimate each model independent of these others and that they are not related. However, they are closely related quantities.

For example, suppose \( L \) and \( A \) were binary. Then

\[
\logit[P(L = 1|A, C)] = \phi_0^* + \phi_1^* A + \phi_2^* C
\]

and

\[
\logit[P(A = 1|L, C)] = \lambda_0 + \lambda_1 L + \lambda_2 C.
\]

Then \( \phi_1^* = \lambda_1 = \text{OR}(A, L|C) \). However we aren’t interested in \( \logit[P(L = 1|A, C)] \) but rather \( \logit[P(L = 1|A, C, Y)] = \phi_0 + \phi_1 A + \phi_2 Y + \phi_3 C \) which will not marginalize over \( Y \) to a logistic regression, but rather a mixture of two logistic regressions for \( Y = 0 \) and \( Y = 1 \). Therefore we could not specify models for \( \logit[P(L = 1|A, C, Y)] = \phi_0 + \phi_1 A + \phi_2 Y + \phi_3 C \) and \( \logit[P(A = 1|L, C)] = \lambda_0 + \lambda_1 L + \lambda_2 C \) that are compatible with each other.
Then we can use the logit link function to model both $t$ and $p$:

\[
\begin{align*}
\text{logit} \left[ P(L = 1|A,Y,C) \right] &= \text{logit} \left[ P(L = 1|Y,C) \right] + \log \text{OR}(L = 1, A|Y,C) - \log E \left[ \text{OR}(L = 1, A|Y,C) \mid L = 0, Y, C \right] \\
\text{logit} \left[ P(A = 1|L,C) \right] &= \text{logit} \left[ P(A = 1|C) \right] + \log \text{OR}(A = 1, L|C) - \log E \left[ \text{OR}(A = 1, L|C) \mid A = 0, C \right]
\end{align*}
\]

Thus we see that both model the association between $L$ and $A$, but the former is conditional on $Y$ and $C$, while the later is only conditional on $C$. There may not be an intersection submodel for the particular choice of the nuisance models. In simulation we can ensure these models are compatible, but in practice we won’t realistically be able to make this assumption.

We can similarly relate $m(Y|A,C,L)$ and $t(L|A,C,Y)$, which may be easier for illustrative purposes. The proposed MR solution assumes we can specify $t(L|A,C,Y)$ and $m(Y|A,C,L)$ independently of each other. For example suppose we propose that:

\[
\begin{align*}
Y|A,C,L & \sim N \left( \nu_0 + \nu_1 A + \nu_2 L^3 + \nu_3 C, \sigma_Y^2 \right) \\
L|A,C,Y & \sim N \left( \phi_0 + \phi_1 A + \phi_2 Y^2 + \phi_3 C, \sigma_L^2 \right)
\end{align*}
\]

Unless $\nu_2 = 0$, these models are not compatible in that there does not exist a joint distribution for $(L,Y)$ with the given families as its conditional distributions [3]. Therefore, for this example, we could never have $t$ and $b$ both be correct. So case 4 above could never be true. It is therefore necessary to understand the implicit dependencies between various nuisance functions.

### B.5 Detailed Reparameterization of the Likelihood

We must reparameterize the likelihood because the nuisance parameters overlap.

**Lemma:**

\[
\frac{f(X_1|X_2)}{f(X_1 = 0|X_2)} = \int \frac{f(X_1|X_2, X_3)}{f(X_1 = 0|X_2, X_3)} df(X_3|X_2, X_1 = 0)
\]

Following Chen (2007) [7] and Tchetgen et al. (2009) [59] we define the generalized conditional odds ratio function of $A$ and $Y$ given $L$ as

\[
\chi(A, Y|L) = \frac{f(A|Y, L) f(a_0|y_0, L)}{f(a_0|Y, L) f(A|y_0, L)}
\]
where \((a_0, y_0)\) is a reference value.

Reparameterization:

\[
\frac{f(L|Y, A, C)}{f(L = 0|Y, A, C)} = \frac{f(L|Y, A, C)}{f(L = 0|Y, A, C)} \left\{ \frac{f(L|Y = 0, A, C)}{f(L = 0|A, C)} \right\}^{-1} \frac{f(L|A, C)}{f(L = 0|A, C)} \\
= \frac{f(L|Y, A, C)}{f(L = 0|Y, A, C)} \left\{ \int \frac{f(L|y, A, C)}{f(L = 0|y, A, C)} f(y|A, C, L = 0) dy \right\}^{-1} \frac{f(L|A, C)}{f(L = 0|A, C)} \\
	imes \left\{ \int \frac{f(L|y, A, C)}{f(L = 0|y, A, C)} \frac{f(L = 0|y = 0, A, C)}{f(L|y = 0, A, C)} f(y|A, C, L = 0) dy \right\}^{-1} \\
= \chi(L, Y|A, C) \frac{f(L|A, C)}{f(L = 0|A, C)} \\
	imes \left\{ \int \chi(L, y|A, C) f(y|A, C, L = 0) dy \right\}^{-1} \frac{f(L|A, C)}{f(L = 0|A, C)}.
\]

Following Chen (2007) [7] the joint distribution of \(L\) and \(Y\) given \(A\) and \(C\) can be written as

\[
f(L, Y|A, C) = \frac{f(L|Y = 0, A, C) \chi(L, Y|A, C) f(Y|L = 0, A, C)}{\int \chi(l, y|A, C) f(l|y_0, A, C) f(y|l_0, A, C) \ dy dl}.
\]

Then,

\[
f(L, Y|A, C) = \frac{f(L|Y = 0, A, C)}{f(L = 0|Y = 0, A, C)} \chi(L, Y|A, C) f(Y|L = 0, A, C) \\
= \frac{f(L|Y = 0, A, C)}{f(L = 0|Y = 0, A, C)} \left\{ \int \chi(l, y|A, C) f(y|A, C, L = 0) \ dy \right\}^{-1} \frac{f(L|A, C)}{K(A, C)} \\
= \chi(L, Y = 0|A, C) \left\{ \int \chi(l, y|A, C) f(y|A, C, L = 0) \ dy \right\}^{-1} \frac{f(L|A, C)}{K(A, C)} \\
= \frac{f(L|A, C)}{f(L = 0|A, C)} \left\{ \int \chi(l, y|A, C) f(y|A, C, L = 0) \ dy \right\}^{-1} \frac{f(L|A, C)}{K(A, C)} \chi(L, Y|A, C) f(Y|L = 0, A, C)
\]

where \(K(A, C) = \left\{ f(L|L = 0|Y = 0, A, C) \right\}^{-1} \int \chi(l, y|A, C) f(l|y_0, A, C) f(y|l_0, A, C) \ dy dl \) and because \(\chi(L, Y = 0|A, C) = 1\).

We can see that the joint distribution of \(L\) and \(Y\) can be expressed in terms of \(f(L|A = 0, C)\)
, \( \chi(A, L|C) \), \( \chi(L, Y|A, C) \), and \( f(Y|L = 0, A, C) \). This allows us to estimate \( f(L, Y|A, C) \) with direct maximum likelihood.

Similarly we can write:

\[
f(A|L, C) = \frac{f(A|L = 0, C)}{f(A = 0|L = 0, C)} \chi(A, L|C) \left\{ \int \frac{f(A|L = 0, C)}{f(A = 0|L = 0, C)} \chi(A, L|C) da \right\}^{-1}
\]

The propensity score can be expressed in terms of \( \chi(A, L|C) \) and \( f(A|L = 0, C) \).

Therefore both terms require correct specification of \( \chi(A, L|C) \).

**Details on Estimation** Recall:

\[
\tau_{\text{Full}}(O) = \frac{I(A = 0)}{Pr(A = 1)} \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} (Y - E[Y|A = 0, L, C]) + \frac{I(A = 1)}{Pr(A = 1)} (E[Y|A = 0, L, C] - \Psi)
\]

where \( O = (Y, A, C) \) are the fully observed variables.

Therefore,

\[
\tau_{\text{Miss}}(\Psi) = R \left\{ \frac{I(A = 0)}{Pr(A = 1)} \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} (Y - E[Y|A = 0, L, C]) + \frac{I(A = 1)}{Pr(A = 1)} (E[Y|A = 0, L, C] - \Psi) \right\}
\]

Thus,

\[
= R \left\{ \frac{I(A = 0)}{Pr(A = 1)} \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} (Y - E[Y|A = 0, L, C]) + \frac{I(A = 1)}{Pr(A = 1)} E[Y|A = 0, L, C] \right\}
\]

\[
= R \left\{ \frac{I(A = 0)}{Pr(A = 1)} \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} (Y - E[Y|A = 0, L, C]) + \frac{I(A = 1)}{Pr(A = 1)} E[Y|A = 0, L, C] \right\}
\]

\[
- \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 0)}{Pr(A = 1)} \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} (Y - E[Y|A = 0, L, C]) + \frac{I(A = 1)}{Pr(A = 1)} E[Y|A = 0, L, C] \right\}
\]

\[
- \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 0)}{Pr(A = 1)} Y E \left[ \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} | Y, A = 0, C \right] \right\}
\]

\[
+ \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 0)}{Pr(A = 1)} E \left[ \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} E[Y|A = 0, L, C] | Y, A = 0, C \right] \right\}
\]

\[
- \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A = 1)}{Pr(A = 1)} E [E[Y|A = 0, L, C]|Y, A = 1, C] \right\}
\]

83
\[
+ \left( \frac{R}{\pi} - 1 \right) \frac{I(A = 1)}{Pr(A = 1)} \Psi
\]

We set the previous expression equal to zero and solve for \( \Psi \). Note that \( \frac{R}{\pi} \frac{I(A=1)}{Pr(A=1)} \Psi - \left( \frac{R}{\pi} - 1 \right) \frac{I(A=1)}{Pr(A=1)} \Psi = \Psi \frac{I(A=1)}{Pr(A=1)} \).

\[
\Psi \frac{I(A = 1)}{Pr(A = 1)} = \frac{R}{\pi} \left\{ \frac{I(A = 0)}{Pr(A = 1)} \frac{Pr(A = 1|L, C)}{Pr(A = 0|L, C)} (Y - E[Y|A = 0, L, C]) + \frac{I(A = 1)}{Pr(A = 1)} E[Y|A = 0, L, C] \right\}
\]

We can then look at each term on the right hand side of the equation separately and consider the models proposed in the main body of the chapter.

**Term 1:** \( \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A=0)}{Pr(A=1)} p \frac{Y}{1-p} (Y - m(0, L, C)) \right\} \) Term 1 requires models for \( \pi, p \), and for \( m(0, L, C) = \mu_Y^0 \), which are easily estimated as described in the main body of the chapter.

**Term 2:** \( \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A=0)}{Pr(A=1)} YE \left[ \frac{p}{1-p}, Y, A = 0, C \right] \right\} \) Term 2 also requires models for \( \pi \) and \( p \) in addition to \( g(A, Y, C) \). Using that fact that, for \( X \sim N(\mu, \sigma^2) \), the MGF is \( E[e^{tX}] = e^{\mu t + \frac{1}{2} \sigma^2 t^2} \), we have that:

\[
E \left[ \frac{p}{1-p} | Y, A = 0, C \right] = E \left[ e^{\lambda_0 + \lambda_1 L + \lambda_2 C} | Y, A = 0, C \right] = e^{\lambda_0 + \lambda_2 C} E \left[ e^{\lambda_1 L} | Y, A = 0, C \right] = e^{\lambda_0 + \lambda_2 C} \exp(\lambda_1 \mu_L^0 + \frac{1}{2} \sigma_L^2 \lambda_1^2).
\]

Thus Term 2 can be expressed as \( \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A=0)}{Pr(A=1)} Y \exp(\lambda_0 + \lambda_2 C + \lambda_1 \mu_L^0 + \frac{1}{2} \sigma_L^2 \lambda_1^2) \right\} \).

**Term 3** \( \left( \frac{R}{\pi} - 1 \right) \left\{ \frac{I(A=0)}{Pr(A=1)} E \left[ \frac{p}{1-p} E[Y|A = 0, L, C] | Y, A = 0, C \right] \right\} \) Term 3 requires models for \( \pi, p \), \( g(0, Y, C) \) and \( m(0, L, C) \).
Let \( E \left[ \frac{p}{1-p} m(0, L, C) \mid Y, A = 0, C \right] = \zeta \). Then,

\[
\zeta = E \left[ e^{\lambda_0 + \lambda_1 L + \lambda_2 C (\nu_0 + \nu_2 L + \nu_3 C)} \mid Y, A = 0, C \right] \\
= e^{\lambda_0 + \lambda_2 C (\nu_0 + \nu_3 C)} \exp \left( \lambda_1 \mu_1^0 + \frac{1}{2} \sigma_1^2 \lambda_1^2 \right) + e^{\lambda_0 + \lambda_2 C} \nu_2 \exp \left( \lambda_1 \mu_L^0 + \frac{1}{2} \sigma_L^2 \lambda_1^2 \right) (\mu_L^0 + \sigma_L^2 \lambda_1)
\]

Thus Term 3 can be expressed as

\[
\left( \frac{R}{\pi} - 1 \right) \{ I(A=0) \left[ (\nu_0 + \nu_3 C) \exp \left( \lambda_0 + \lambda_2 C + \lambda_1 \mu_L^0 + \frac{1}{2} \sigma_L^2 \lambda_1^2 \right) + (\mu_L^0 + \sigma_L^2 \lambda_1) \right] \}
\]

**Term 4** requires models for \( \pi, g(A, Y, C) \) and \( m(0, L, C) \).

\[
E \left[ \frac{R}{\pi} m(0, L, C) \mid Y, A = 1, C \right] = E \left[ \nu_0 + \nu_2 L + \nu_3 C \mid Y, A = 1, C \right]
\]

Thus, in our example, Term 4 can be expressed as \( (\frac{R}{\pi} - 1) \left\{ I(A=1) \frac{E \left[ m(0, L, C) \mid Y, A = 1, C \right]}{E \left[ m(0, L, C) \mid Y, A = 1, C \right]} \right\} \)

To estimate \( \Psi \), we calculate the sum of the four terms for each subject, then take the sample mean across all subjects.

### C Proofs and Results for Chapter 3

**C.1 Proof of Double Robustness:**

Case 1 - \( \pi^*(V; \eta) \) **is incorrect but** \( t(L|A, C; \alpha) \) **is correct**  Recall that we always require \( f(Y|L, V) \) to be correctly specified, and therefore \( E_{\Theta} \left[ g(Y, V) \mid L, V \right] \) is always correct. If \( t(L|A, C; \alpha) \) is also correct, then \( E \left\{ E_{\Theta} \left[ g(Y, V) \mid L, V \right] \right\} = E_{\Theta, \alpha} \left[ g(Y, V) \mid V \right] \)

\[
E[U_{DR}(\Theta)|L, V] = E_{\Theta} \left[ \frac{R}{\pi^*(V; \eta)} \{ g(Y, V) - E_{\Theta, \alpha} [g(Y, V)|V] \} \right]
\]
\[
\begin{align*}
&+ \frac{1 - R}{1 - \pi^*(V; \eta)} \{E_{\Theta, \alpha} [g(Y, V)|V] - E_{\Theta} [g(Y, V)|L, V]\} |L, V

&= \frac{\pi(V; \eta)}{\pi^*(V; \eta)} E\Theta [g(Y, V) - E_{\Theta, \alpha} [g(Y, V)|V] |L, V] + \\
&+ \frac{1 - \pi(V; \eta)}{1 - \pi^*(V; \eta)} E\Theta \{[E_{\Theta, \alpha} [g(Y, V)|V] - E_{\Theta} [g(Y, V)|L, V]\} |L, V

&= \frac{\pi(V; \eta)}{\pi^*(V; \eta)} \{E_{\Theta} [g(Y, V)|L, V] - E_{\Theta, \alpha} [g(Y, V)|V]\} + \\
&+ \frac{1 - \pi(V; \eta)}{1 - \pi^*(V; \eta)} \{E_{\Theta, \alpha} [g(Y, V)|V] - E_{\Theta} [g(Y, V)|L, V]\}

&= \frac{\pi(V; \eta)}{\pi^*(V; \eta)} \{0\} + \frac{1 - \pi(V; \eta)}{1 - \pi^*(V; \eta)} \{0\}

&= 0
\end{align*}
\]

**Case 2 - \(\pi(V; \eta)\) is correct but \(t^*(L|A, C; \alpha)\) is incorrect**  
Recall that we always require \(f(Y|L, V)\) to be correctly specified, and therefore \(E_{\Theta} [g(Y, V)|L, V]\) is always correct.

\[
E[U_{DR}(\Theta)|L, V] = E_{\Theta} \left[ \frac{R}{\pi(V; \eta)} \{g(Y, V) - E_{\Theta, \alpha} [g(Y, V)|V]\} \right]
\]

\[
+ \frac{1 - R}{1 - \pi(V; \eta)} \{E_{\Theta, \alpha}^* [g(Y, V)|V] - E_{\Theta} [g(Y, V)|L, V]\} |L, V
\]

\[
= \frac{\pi(V; \eta)}{\pi^*(V; \eta)} E_{\Theta} [g(Y, V) - E_{\Theta, \alpha}^* [g(Y, V)|V] |L, V] + \\
+ \frac{1 - \pi(V; \eta)}{1 - \pi^*(V; \eta)} E \{E_{\Theta, \alpha}^* [g(Y, V)|V] - E_{\Theta} [g(Y, V)|L, V]\} |L, V

= E_{\Theta} [g(Y, V)|L, V] - E_{\Theta, \alpha}^* [g(Y, V)|V] + E_{\Theta, \alpha}^* [g(Y, V)|V] - E_{\Theta} [g(Y, V)|L, V]

= 0
\]

**C.2 Proof of 3.13**

From *Newey & McFadden (1994)* [32] we can find the closed form for \(h^{opt}(V)\). \(h^{opt}(V)\) solves the following equation:

\[
\forall h(V),
\]

\[
-E \left[ \frac{\delta \Omega(\theta, h)}{\delta \theta} \right] = E \left[ \Omega(\theta, h) \Omega(\theta, h^{opt})^T \right]
\]
\[ \forall h(V), \]
\[ -E \left[ h(V) \nabla_{\theta} \mathcal{Y} (\theta) \right] = E \left[ \mathcal{Y}^2 (\theta) h(V) h^{opt} (V) \right] \]
\[ \Rightarrow \]
\[ E \left[ h(V) (\mathcal{Y}^2 (\theta) h^{opt} (V) + \nabla_{\theta} \mathcal{Y} (\theta)) \right] = 0 \]
\[ \Leftarrow \Rightarrow \]
\[ E \left[ h(V) E \left[ \mathcal{Y}^2 (\theta) h^{opt} (V) + \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right] \right] = 0 \]

Choose \( h(V) = E \left[ \mathcal{Y}^2 (\theta) h^{opt} (V) + \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right] \).

Thus,
\[ E \left[ (E \left[ \mathcal{Y}^2 (\theta) h^{opt} (V) + \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right])^2 \right] = 0. \]

Then we note that \( (E \left[ \mathcal{Y}^2 (\theta) h^{opt} (V) + \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right])^2 = 0 \implies E \left[ \mathcal{Y}^2 (\theta) h^{opt} (V) + \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right] = 0. \)

Thus,
\[ E \left[ \mathcal{Y}^2 (\theta) \mid V \right] h^{opt} (V) + E \left[ \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right] = 0 \]
\[ \Rightarrow \]
\[ h^{opt} (V) = -E \left[ \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right] E \left[ \mathcal{Y}^2 (\theta) \mid V \right]^{-1} \]

### C.3 Closed form for DR estimator

Recall that \( \frac{d}{dx} \expit (ax) = a (1 - \expit (ax)) \expit (ax) \) and that \( J(L, V; \theta) = E \left[ Y \mid L, V; \theta \right] \) for \( Y \) binary and \( M(V; \theta, \alpha) = E_{\Theta, \alpha} [g(Y, V) \mid V]. \)

Then,
\[ -E \left[ \nabla_{\theta} \mathcal{Y} (\theta) \mid V \right] = -E \left[ \nabla_{\theta} \left( \frac{R}{\pi(V)} (Y - M(V; \theta, \alpha)) + \frac{1 - R}{1 - \pi(V)} (M(V; \theta, \alpha) - J(L, V; \theta)) \right) \mid V \right] \]
\[ = \ E \left[ \frac{R}{\pi(V)} \nabla_{\theta} M(V; \theta, \alpha) - \frac{1 - R}{1 - \pi(V)} \nabla_{\theta} (M(V; \theta, \alpha) - J(L, V; \theta)) \mid V \right] \]
\[ = \ E \left[ \frac{R}{\pi(V)} \nabla_{\theta} E \left[ J(L, V; \theta) \mid V \right] \mid V \right] - E \left[ \frac{1 - R}{1 - \pi(V)} \nabla_{\theta} \left\{ E \left[ J(L, V; \theta) \mid V \right] - J(L, V; \theta) \right\} \mid V \right] \]
\[
\begin{align*}
E \left[ \frac{R}{\pi(V)} E[\nabla_{\theta} J (L, V; \theta) \mid V] \right] & = E \left[ \frac{1 - R}{1 - \pi(V)} \{ E[\nabla_{\theta} J (L, V; \theta) \mid V] - \nabla_{\theta} J (L, V; \theta) \} \mid V \right] \\
& = E \left[ \frac{R}{\pi(V)} \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) J (L, V; \theta) \left\{ 1 - J (L, V; \theta) \right\} \mid V \right] \\
& - E \left[ \frac{1 - R}{1 - \pi(V)} \left\{ E \left[ \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) J (L, V; \theta) \left\{ 1 - J (L, V; \theta) \right\} \mid V \right] - \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) J (L, V; \theta) \left\{ 1 - J (L, V; \theta) \right\} \right\} \mid V \right] \\
& = E \left[ \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) J (L, V; \theta) \left\{ 1 - J (L, V; \theta) \right\} \mid V \right] \\
& - E \left[ \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) J (L, V; \theta) \left\{ 1 - J (L, V; \theta) \right\} \mid V \right] + E \left[ \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) J (L, V; \theta) \left\{ 1 - J (L, V; \theta) \right\} \mid V \right] \\
& = E \left[ \left( \begin{array}{c} 1 \\ V \\ L \end{array} \right) \exp \left( \theta_0 + \theta_1 V + \theta_2 L \right) \frac{1}{1 + \exp \left( \theta_0 + \theta_1 V + \theta_2 L \right)} \mid V \right].
\end{align*}
\]

Next,

\[
E \left[ Y^2 \mid V \right] = E \left[ \left( \frac{R}{\pi(V)} \left( Y - M (V; \theta, \alpha) \right) + \frac{1 - R}{1 - \pi(V)} \left( M (V; \theta, \alpha) - J (L, V; \theta) \right) \right)^2 \mid V \right]
\]

\[
= E \left[ \frac{R}{\pi(V)} \left( Y - M (V; \theta, \alpha) \right) \right]^2 + 2 \left\{ \frac{1 - R}{1 - \pi(V)} \left( M (V; \theta, \alpha) - J (L, V; \theta) \right) \right\} \left[ \frac{R}{\pi(V)} \left( Y - M (V; \theta, \alpha) \right) \right] \\
+ \left\{ \frac{1 - R}{1 - \pi(V)} \left( M (V; \theta, \alpha) - J (L, V; \theta) \right) \right\}^2 \mid V \right] \\
+ \frac{(1 - R)^2}{\left(1 - \pi(V) \right)^2} \left\{ \left[ M (V; \theta, \alpha) \right]^2 - 2 M (V; \theta, \alpha) J (L, V; \theta) + J (L, V; \theta)^2 \right\} \mid V \right] \\
= \frac{1}{\pi(V)} E \left[ Y - 2YM (V; \theta, \alpha) + (M (V; \theta, \alpha))^2 \mid V \right]
\]

88
\[
+ \frac{1}{1 - \pi(V)} \mathbb{E} \left[ [M(V; \theta, \alpha)]^2 - 2M(V; \theta, \alpha) J(L, V; \theta) + [J(L, V; \theta)]^2 \right] \\
= \frac{1}{\pi(V)} \mathbb{E} \left[ [M(V; \theta, \alpha)]^2 - 2M(V; \theta, \alpha) M(V; \theta, \alpha) + [J(L, V; \theta)]^2 \right] + \frac{1}{1 - \pi(V)} \mathbb{E} \left[ - [M(V; \theta, \alpha)]^2 + [J(L, V; \theta)]^2 \right]
\]