



Published in final edited form as:

J R Stat Soc Ser A Stat Soc. 2021 October ; 184(4): 1390–1413. doi:10.1111/rssa.12716.

Propensity score analysis for a semi-continuous exposure variable: a study of gestational alcohol exposure and childhood cognition

Tugba Akkaya Hocagil¹, Richard J. Cook¹, Sandra W. Jacobson², Joseph L. Jacobson², Louise M. Ryan^{3,4}

¹Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

²Department of Psychiatry and Behavioral Neurosciences, Wayne State University, Detroit, USA

³School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, Australia

⁴Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, USA

Abstract

Propensity score methodology has become increasingly popular in recent years as a tool for estimating causal effects in observational studies. Much of the related research has been directed at settings with binary or discrete exposure variables with more recent work involving continuous exposure variables. In environmental epidemiology, a substantial proportion of individuals is often completely unexposed while others may experience heavy exposure leading to an exposure distribution with a point mass at zero and a heavy right tail. We suggest a new approach to handle this type of exposure data by constructing a propensity score based on a two-part model and show how this model can be used to more reliably adjust for covariates of a semi-continuous exposure variable. We also consider the case when a misspecified propensity score is used in a regression adjustment and derive an explicit form of the bias. We show that the potential bias gets smaller as the estimated propensity score gets closer to the true expectation of the exposure variable given a set of observed covariates. While this result pertains to a more general setting, we use it to evaluate the potential bias in settings in which the true exposure has a semi-continuous structure. We also evaluate and compare the performance of our proposed method through simulation studies relative to a simpler linear regression-based propensity score for a continuous exposure variable as well as through direct covariate adjustment. Overall, we find that using a propensity score constructed via a two-part model significantly improves the regression estimate when the exposure variable is semi-continuous in nature. Specifically when the proportion of non-exposed subjects is high and the effects of covariates on exposure and outcome are strong, the proposed two-part propensity score method outperforms the more standard competing methods. We illustrate our method using data from the Detroit Longitudinal Cohort Study in which the exposure variable reflects gestational alcohol exposure featuring zero values and a long tail.

Correspondence: Tugba Akkaya Hocagil, University of Waterloo, Department of Statistics and Actuarial Science, Waterloo, Canada. takkayahocagil@uwaterloo.ca.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

Keywords

causal effect; confounding; generalized propensity score; propensity score; two-part model

1 | INTRODUCTION

1.1 | Background and literature review

Environmental epidemiologists often seek to assess the impact of a potentially toxic exposure on a health outcome, while adjusting statistically for possible confounders. Failure to account for the effect of the confounders can result in biased effect estimates that either overstate or obscure the true causal effect of the exposure. In practice, regression models that include potential confounders are most commonly used to adjust for the effects of confounding variables. However, particularly when the number of potential confounders is large, it can be difficult to determine whether the regression model relating the exposure variable and confounding variables to the outcome has been correctly specified (Austin, 2011). Use of a propensity score has become a popular alternative strategy to reduce or eliminate the effects of confounding in observational settings (Austin, 2011; Rosenbaum & Rubin, 1983).

A propensity score is typically obtained via a regression analysis in which the exposure is modelled as a function of potential confounding variables. Propensity score methodology was originally developed to estimate the causal effect of a binary exposure variable (Rosenbaum & Rubin, 1984, 1985) and consequently propensity scores have typically been based on binary regression models. Imbens (1999), Imai and van Dyk (2004) and Hirano and Imbens (2004) broadened this framework to accommodate continuous exposure variables and coined the term the generalized propensity score for such settings. (Hirano & Imbens, 2004; Imai & van Dyk, 2004; Imbens, 1999). With a continuous exposure variable, Hirano and Imbens (2004) define the generalized propensity score for a particular individual as the conditional density function of the exposure model given covariates, including potential confounders, evaluated at the observed exposure and covariate values. In contrast, Imai and van Dyk (2004) define the generalized propensity score to be the entire conditional density function of the treatment. Such propensity score-based methods allow practitioners to estimate a full dose–response function rather than simply the average treatment effect (Fichera et al., 2016; Moodie & Stephens, 2012). Zhao et al. (2020) evaluated these two methods for estimation of the full dose–response function and explored the performance of two extensions they proposed to improve the robustness of analyses based on the generalized propensity score (Zhao et al., 2020).

Generalized propensity scores have been used in settings involving continuous exposure variables related to labour earnings (Bia & Mattei, 2008; Imai & van Dyk, 2004), medical expenditure (Imai & van Dyk, 2004; Zhao et al., 2020), and birth weight (Zhang et al., 2016), among others.

In environmental and behavioural epidemiology, exposure may refer to chemical exposures, pollution or personal use of substances, such as alcohol, tobacco or illicit drugs. A number

of challenging problems can arise in such settings. Often, the exposure variable is semi-continuous in that the data may feature a large proportion of individuals who have not been exposed at all. The exposure distribution may also have a long right tail due to a small number of individuals with very high exposure. Examples are ubiquitous and include counts of cigarettes smoked per day in a population made up of smokers and non-smokers (Wang & Heitjan, 2008), concentrations of toxins in environmental epidemiology where levels below the detection limit are recorded as zero (Choi & Hwang, 2018), and alcohol intake per day in a population made up of drinkers and non-drinkers (Choi & Hwang, 2018). Other settings include health services research (Neelon et al., 2015), studies of the microbiome (Chai et al., 2018), when modelling functional ability in chronic diseases (Su et al., 2015) and when modelling occlusion therapy in the treatment of childhood amblyopia (Moodie & Stephens, 2012). One strategy for modelling the effects of such exposure variables is to discretize them. In the present setting, for example, we could classify subjects as having zero, low, moderate or high exposure. While this can address confounding, this strategy is not appropriate when scientific interest lies in a parsimonious characterization of the dose–response effect. In environmental risk assessment, for instance, the goal is often to identify the specific exposure level associated with a specified risk above a background level (Crump, 1984). In the project funded by the National Institutes of Health in the United States that motivates our work, investigators seek to understand whether there are levels of gestational alcohol exposure that are not associated with clinically significant risks of adverse developmental effects for the children. Study findings could have important implications for the diagnosis and treatment of affected children.

Semi-continuous variables can be viewed as arising from two distinct processes: (i) a binary process, reflecting whether or not the subject was exposed, and (ii) a continuous process, determining the level of exposure given that some exposure occurred. Two-part models have been developed to handle such data. Often the binary component is modelled via logistic regression, and a log transformation is applied to model the continuous part using standard linear regression (Smith et al., 2017). As an alternative to two-part models, analysts often fit a standard generalized linear model with a log link to ensure that the predicted values are not negative (Smith et al., 2017). Several researchers have compared the performance of two-part models with various classical approaches in settings that involve semi-continuous response variables. While findings are context dependent, two-part models often exhibit better performance (Duan et al., 1983; Madden et al., 2000; Smith et al., 2017). Smith et al. (2017) found that the standard generalized linear models are not appropriate when data are zero-heavy. Specifically, they reported that inferences based on standard generalized linear models incurred increased bias, lower than nominal coverage, and increased type I error rates in settings with non-negligible proportions of individuals with zero values (Smith et al., 2017).

Despite the recent interest in generalized propensity scores, there has been relatively little work examining modelling issues with zero-inflated exposure variables. Imai and van Dyk (2004) argued that generalized propensity scores can be used to establish causal effects in observational studies when the exposure has a mixture of zeros and continuously distributed positive values, but did not explore this issue further. Naimi et al. (2014) explored a variety of approaches for modelling the propensity score for normal and contaminated

Poisson exposure variables and investigated the sensitivity of misspecification of propensity score models within the framework of inverse probability weighted estimating functions. Specifically, they constructed inverse probability weights using: a normal distribution, a normal distribution with heteroscedastic variance, a gamma distribution, a t-distribution and using a quantile binning approach. These authors found that when the exposure variable is zero-inflated, all of these methods yielded slightly biased estimators with the gamma and quantile binning approach giving the best estimators (Naimi et al., 2014).

In this paper, we propose use of a two-part model for the construction of a generalized propensity score for a semi-continuous exposure variable. Unlike the existing methods, our method allows covariates to act differently on the exposure indicator (exposed vs. unexposed) and the extent of exposure among those who are exposed. Given the known problems of inferences based on standard generalized linear models when the response variable is semi-continuous, application of a two-part modelling approach to propensity score construction would seem to be a more appealing approach for propensity score construction in this setting. We evaluate and compare the performance of the proposed method with the established generalized propensity score approach (Imai & van Dyk, 2004).

The seminal work by Rosenbaum and Rubin (1983) showed that once the propensity score is estimated, effects of confounding can be eliminated through propensity score matching, stratification or covariance adjustment in regression. We focus on covariate adjustment using the propensity score estimated by the two-part model. We derive an explicit functional form of the bias that arises when a misspecified propensity score is used in covariate adjustment. In our simulation studies, we calculate this bias for settings where the exposure variable is governed by a two-part model but the propensity score is estimated by ordinary least square regression as it is suggested in generalized propensity score approach.

The remainder of this article is organized as follows. In the following sub-section we introduce the motivating study which explores the relationship between prenatal alcohol exposure and cognitive function in children from the Detroit Longitudinal Cohort study. In Section 2, we introduce notation and describe how to estimate the two-part generalized propensity score. We then provide the explicit form of bias associated with use of a misspecified propensity score as a covariate in regression adjustment. Section 3 describes and reports on a simulation study designed to evaluate the performance of the proposed method relative to methods based on a naive unadjusted analysis, conventional linear regression and the generalized propensity score approach. In Section 4, we apply these methods to the Detroit Longitudinal Cohort Study. In Section 5 we discuss the strengths and limitations of the proposed methods, and outline areas warranting further research.

1.2 | The Detroit study of foetal alcohol exposure

While evidence from animal models and epidemiological studies has linked prenatal alcohol exposure (PAE) to a broad range of cognitive and behavioural deficits (Jacobson et al., 2011; Lewis et al., 2016; Mattson et al., 1997) the exact nature of the dose–response effect is not well understood. In particular, there is virtually no information in the scientific literature regarding the levels of PAE associated with an increased risk of clinically significant adverse effects. It is widely known that women should stop drinking during pregnancy, but there

remain important scientific questions regarding whether or not low-level exposures will result in adverse developmental effects, and more generally about the dose–response effects. Insights into these questions could have important clinical implications for the diagnosis and treatment of affected children. Moreover this critical public health question cannot be addressed by studies that have used samples of children identified in clinics and referred to specialists for assessment since in such cases the level of maternal alcohol use, and hence gestational alcohol exposure has not been quantified.

The Detroit Longitudinal Cohort study is a large prospective study of children with a broad range of prenatal alcohol exposure levels, who were followed from birth through age 19 years (Jacobson et al., 2004). Four hundred and eighty pregnant, African-American, inner-city women were interviewed about their alcohol use during pregnancy using a timeline follow-back interview (Jacobson et al., 2002). Data from each visit regarding maternal alcohol consumption were summarized in terms of daily alcohol intake (standard drinks or ounces of absolute alcohol (AA)/day) averaged across pregnancy.

For this paper, we analysed data for a single outcome measure, the Wechsler Intelligence Scales for Children (3rd edition) (WISC-III) Freedom from Distractibility Index, for the 337 children who participated in the 7-year follow-up assessment of this cohort in the presence of a broad set of potential confounding variables covariates (see Tables 1 and 2). Figure 1 displays the relation between the outcome variable and the log of prenatal alcohol exposure. The figure shows a long-tailed exposure distribution, even after transformation to the log scale. Given that 23.8% of the women reported no alcohol use during pregnancy, characterizing this variable as two distinct processes (i.e. one determining whether the mother consumed alcohol, the other determining the actual level if alcohol was consumed), is theoretically appealing and provides a richer description than a model with a single function. It is also reasonable to suspect that covariates might operate differently in terms of predicting whether or not a mother consumes alcohol, as compared to the amount of alcohol used by those who consumed it. Indeed, initial exploratory analysis revealed that prenatal marijuana exposure and gravidity were strong predictors of whether or not a woman drank alcohol during pregnancy but were not predictors of the amount consumed by those who reported drinking at least some alcohol.

2 | PROPENSITY SCORE ADJUSTMENT WITH SEMI-CONTINUOUS EXPOSURE

We consider the case in which the goal is to model the causal effect of a semi-continuous exposure variables X on an outcome Y in the presence of a set of confounding variables $Z = (Z_1, \dots, Z_p)'$. The effect of confounding can be addressed by regressing the quantitative exposure on the set of observed covariates as it is suggested by Imai and van Dyk (2004). Specifically, the propensity score can be estimated using ordinary least squares (OLS) regression. Then the predicted mean of X obtained from a model predicting X given all confounding variables Z can be used as the estimated propensity score in the outcome model:

$$Y_i = \eta_0 + \eta_1 X_i + \eta_2 S_i^* + \epsilon_i \quad (1)$$

where S_i^* is the estimated propensity score for subject i . In the next subsection, we introduce a two-part generalized propensity score for the settings where the exposure variable has a substantial proportion of zero values. In this approach, the semi-continuous exposure variable is considered as arising from two distinct processes: a binary part (exposed vs. not exposed) and a continuous part (characterizing the extent of exposure among those exposed).

2.1. A two-part generalized propensity score model

Consider a non-negative semi-continuous variable X with a point mass at zero and a density characterizing the probability distribution of non-zero values on the positive real line. Let $Z = (Z_1, \dots, Z_p)'$ denote a $p \times 1$ set of covariates and consider a two-part model for characterizing the association between Z and X . Specifically, if $X^+ = I(X > 0)$ indicates a positive value for X , we let $\pi(Z) = E(X^+ | Z)$ and consider a binary regression model defined by the link function $g(\cdot)$ mapping the interval $[0, 1]$ onto the real line and setting $g(\pi(Z; \alpha_1)) = \bar{Z}'\alpha_1$ where $\bar{Z} = (1, Z_1, \dots, Z_p)'$ and $\alpha_1 = (\alpha_{10}, \alpha_{11}, \dots, \alpha_{1p})'$. In what follows we use the canonical logistic link function in which case α_{1j} , $j = 0, 1, \dots, p$, are log odds ratios.

We also let $P(X \leq x | X^+ = 1, Z; \alpha_2) = F^+(x | Z; \alpha_2)$ denote the cumulative distribution function for the positive part of X , $X | X > 0$, given Z , which we take to be indexed by a $(p + 1) \times 1$ parameter α_2 . The full distribution for X is therefore indexed by $\alpha = (\alpha_1, \alpha_2)'$. A key requirement of the model for $X | X^+ = 1$ given Z is that it involve a simple way to compute $E(X | X^+ = 1, Z; \alpha_2) = \mu(Z; \alpha_2)$. One approach would be to adopt a generalized linear model, but any non-negative distribution can be used. Ultimately we may compute

$$E(X | Z; \alpha) = \pi(Z; \alpha_1)\mu(Z; \alpha_2) \quad (2)$$

as the marginal mean for $X | Z$ based on the two-part model formulation. For example, we consider a logistic model for $X^+ | Z$ and let

$$\log(\pi(Z; \alpha_1)/(1 - \pi(Z; \alpha_1))) = \bar{Z}'\alpha_1, \quad (3)$$

while any location-scale regression model can be adopted for $X | X > 0, Z$. These models are specified by setting

$$\log X = \eta + \sigma R \quad (4)$$

where $\eta = \bar{Z}'\gamma$ is the linear predictor, σ is a dispersion parameter, and R is an error term which has any standard distribution on the real line; we write $\alpha_2 = (\gamma', \sigma)'$. We assume R are independently distributed with $R \perp Z$. If $R \sim N(0, 1)$ for example this implies a log-normal regression model for which $\mu(Z; \alpha_2) = \exp(\bar{Z}'\gamma + \sigma^2/2)$. Together the two models based on (3) and (4) give an expression for (2) as

$$E(X | Z; \alpha) = \text{expit}(\bar{Z}'\alpha_1) \cdot \exp(\bar{Z}'\gamma + \sigma^2/2) \quad (5)$$

where $\text{expit}(u) = \exp(u)/(1 + \exp(u))$.

With a sample of observations on n independent individuals, we introduce a subscript i to label individuals and write the data for fitting the semi-continuous propensity score as $\{(X_i, Z_i), i = 1, \dots, n\}$. The full likelihood for α based on this sample is,

$$L(\alpha) \propto \prod_{i=1}^n [P(X_i^+ = 0 | Z_i; \alpha_1)]^{1 - X_i^+} [P(X_i^+ = 1 | Z_i; \alpha_1) dF^+(X_i | X_i^+ = 1, Z_i; \alpha_2)]^{X_i^+} \quad (6)$$

where $dF^+(x | X_i^+ = 1, Z_i; \alpha_2) = f(x | X_i^+ = 1, Z_i; \alpha_2)dx$. Note that (6) can be factored as $L(\alpha) = L_1(\alpha_1)L_2(\alpha_2)$ where

$$L_1(\alpha_1) \propto \prod_{i=1}^n [1 - \pi(Z_i; \alpha_1)]^{1 - X_i^+} [\pi(Z_i; \alpha_1)]^{X_i^+} \quad (7)$$

and

$$L_2(\alpha_2) \propto \prod_{i=1}^n dF^+(X_i | Z_i; \alpha_2)^{X_i^+} \quad (8)$$

which can be maximized separately as they are functionally independent. Then substitution of $\hat{\alpha}$ into (2.2) yields the estimated propensity score $S_i = E(X_i | Z_i; \hat{\alpha})$ for individuals $i, i = 1, \dots, n$. We make the stable unit treatment value assumption, which posits that Y_i , the response by the i th subject under exposure $X = x$, is not affected by what exposures are received by other subjects (Hernán & Robins, 2006; Winship & Morgan, 1999). We also assumed weak unconfoundedness (Hirano & Imbens, 2004).

Instead of including an extensive list of potential confounders in the primary regression analysis, we address the confounding effect of the covariates via covariate adjustment using the propensity score S_i . We fit a linear regression model of the response over the exposure variable and the propensity score. We then fit the response model by controlling for S_i in

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 S_i + \tau R_i \quad (9)$$

R_i are assumed to be independent and identically distributed with mean zero and variance one, $i = 1, \dots, n$ and we also assume $R_i \perp (X_i, S_i), i = 1, \dots, n$. Note also that the purpose of adjusting for the propensity score is to render $X_i \perp Z_i | S_i$.

2.2 | Bias associated with using a misspecified propensity score as a regression covariate

In this sub-section, we derive a formula of the limiting bias of an estimator of an exposure effect when a propensity score is used via regression adjustment but is estimated based on

a misspecified propensity score model. While our general result facilitates the calculation of the asymptotic bias for any type of misspecified propensity score model, we use our result to explore the consequences of using a standard generalized propensity score based on a linear regression model when the confounder–exposure relationship is governed by a two-part model.

As before, X denotes the exposure variable and \mathbf{Z} represents the $p \times 1$ vector of confounding variables. We denote $\mathbf{W}(\mathbf{Z})$ to be a $q \times 1$ function of the confounders. If we let $\mathbf{V} = (1, X, \mathbf{W}'(\mathbf{Z}))'$ be a $(q + 2) \times 1$ vector of covariates and $\boldsymbol{\theta} = (\theta_0, \theta_1, \boldsymbol{\theta}_2)'$ be a $(q + 2) \times 1$ vector of parameters, we can consider a linear model of the form:

$$Y = \theta_0 + \theta_1 X + \boldsymbol{\theta}_2' \mathbf{W}(\mathbf{Z}) + E \quad (10)$$

where we assume $(X, \mathbf{Z}) \perp\!\!\!\perp E$ and $E \sim N(0, \sigma^2)$. We assume that a possibly misspecified propensity score has been computed from a model for X given all of the covariates \mathbf{Z} and we denote this propensity score by

$$\hat{\mu} = \mu(\mathbf{Z}; \hat{\boldsymbol{\alpha}}).$$

Then we regress the continuous outcome Y on the semi-continuous exposure variable X and the estimated assumed propensity score obtained from the assumed model:

$$Y = \beta_0 + \beta_1 X + \beta_2 \mu(\mathbf{Z}; \hat{\boldsymbol{\alpha}}) + \epsilon, \quad (11)$$

under the assumption $(X, \mu(\mathbf{Z}); \hat{\boldsymbol{\alpha}}) \perp\!\!\!\perp \epsilon$ with $\epsilon \sim N(0, \tau^2)$; We let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and assume that estimates are obtained as follows. Let $\bar{\mathbf{V}} = (1, X, \mu(\mathbf{Z}))'$. The contribution to the score equation for $\boldsymbol{\beta}$ from a single individual based on (2.11) is

$$\mathbf{U}(Y, X, \mu(\mathbf{Z}); \boldsymbol{\beta}) = \bar{\mathbf{V}}(Y - \bar{\mathbf{V}}' \boldsymbol{\beta})$$

and $\hat{\boldsymbol{\beta}}$ is consistent for the solution to $E\{\mathbf{U}(Y, X, \mathbf{Z}; \mu(\mathbf{Z}), \boldsymbol{\beta})\} = 0$ where the expectation is taken with respect to true distribution involving Y, X, \mathbf{Z} (Rotnitzky & Wypij, 1994; White, 1982). We show in the Appendix that the explicit form of the asymptotic bias of the estimator $\hat{\beta}_1$ for θ_1 can be expressed as

$$\beta_1 = \theta_1 + \frac{\boldsymbol{\theta}_2' [\boldsymbol{\eta}(\mathbf{Z}) - \boldsymbol{\phi}(\mathbf{Z}) \text{corr}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) \sqrt{V(E(X | \mathbf{Z})) / V(\mu(\mathbf{Z}))}]}{\text{var}(X) - \text{var}(E(X | \mathbf{Z})) \text{corr}^2(E(X | \mathbf{Z}), \mu(\mathbf{Z}))}, \quad (12)$$

where, $\boldsymbol{\eta}(\mathbf{Z}) = (\eta_1(\mathbf{Z}), \eta_1(\mathbf{Z}), \dots, \eta_q(\mathbf{Z}))'$ where $\eta_j(\mathbf{Z}) = \text{cov}(E(X | \mathbf{Z}), W_j(\mathbf{Z}))$, $j = 1, \dots, q$, and $\boldsymbol{\phi}(\mathbf{Z}) = (\phi_1(\mathbf{Z}), \phi_2(\mathbf{Z}), \dots, \phi_q(\mathbf{Z}))'$ where $\phi_j(\mathbf{Z}) = \text{cov}(\mu(\mathbf{Z}), W_j(\mathbf{Z}))$, $j = 1, \dots, q$.

From (2.12), we see that the correct solution is θ_1 , so the propensity score-based estimation will be consistent when the second term in (2.12) is zero. The bias will be smaller as the assumed propensity score model gets closer to characterizing the true conditional expectation of $X | \mathbf{Z}$. Through (2.12) the bias can be seen to depend on the effect of

covariates on the continuous outcome Y as well as the covariance between $\mathbf{W}(Z)$, $E(X | Z)$ and $E(\mu(Z))$ and related variances.

3 | EMPIRICAL STUDIES OF FINITE SAMPLE PROPERTIES

In Section 2, we derived an expression for the limiting bias of an estimator of an exposure effect when a propensity score is used via regression adjustment but it is estimated based on a misspecified propensity score model. The purpose of this derivation is to examine the sensitivity of causal inference to the validity of the propensity score model, and in particular to lay the foundation for the empirical studies that follow relating to the two-part model and its importance with semi-continuous exposure data.

Here we conduct simulation studies to investigate the finite sample performance of the propensity score adjustment based on the two-part model for the exposure variable. We consider $Z = (Z_1, Z_2)'$ where we let Z_1 and Z_2 be standard normal with $E(Z_k) = 0$ and $\text{var}(Z_k) = 1$, $k = 1, 2$.

We specify the distribution for the semi-continuous exposure variable using the framework of Section 2.1. We let $(\text{logit})(\pi(Z)) = \alpha_{10} + \alpha_{11}Z_1 + \alpha_{12}Z_2$ where $\alpha_{11} = \log 2$ and $\alpha_{12} = 0$ and solve for α_{10} in order to set the marginal probability $P(X^+ = 1) = p^\circ$ is at a desired level. We then consider a log-normal regression model as in Equation (4) where we set $\bar{Z} = (1, Z_1, Z_2)'$ in Equation (4) to $\gamma_1 = 0.5$, $\gamma_2 = 0.5$, $\gamma_0 = 1.8$; we set $\sigma = 0.5$. Here, we specified the regression coefficients to ensure that covariates operate differently in terms of predicting whether or not a subject is exposed, versus the amount of exposure for subjects who are exposed.

We also explore the performance of procedure based on fitting a log-normal model when the true distribution for exposed individuals follows a Weibull regression model. Specifically we generate positive exposure levels according to $\log X = \eta^* + \phi R^*$ where $\eta^* = \bar{Z}'\gamma^*$ is the linear predictor, ϕ is a dispersion parameter, and R^* is an error term. We assume R^* has extreme value distribution which implies a Weibull regression model.

We generated the outcome variable using various known functions of the covariates and exposure variable. In particular, we follow the simulation studies described by Imai and van Dyk (2004). First we constructed an additive model of the form:

$$Y_1 = \theta_0 + \theta_1 X + \theta_2 \sum_{k=1}^K \lambda_k^{\kappa_k Z_k} \quad (13)$$

and a multiplicative model of the form:

$$Y_2 = \theta_0^* + \theta_1^* X + \theta_2^* \sum_{k=1}^K \lambda_k Z_k. \quad (14)$$

In both models the coefficient vector for covariates is represented by $\lambda = (\lambda_1, \dots, \lambda_p)$, the constants θ_2 and θ_2^* determine the relative influence of the exposure and the covariates on the continuous outcome, and each component of $\kappa = (\kappa_1, \dots, \kappa_p)$ is 1.

We considered scenarios in which 25% or 50% of subjects are unexposed. For each scenario, we carried out a naive unadjusted analysis, a conventional covariate adjustment, an analysis in which we adjusted for the generalized propensity score function via Gaussian linear regression (Imai & van Dyk, 2004) and finally our proposed propensity function based on a two-part model. We evaluate the performance of our method in two settings, one where the exposure levels for exposed individuals follow a log-normal regression model and one where it follows a Weibull regression model. For each parameter setting we simulated 1000 datasets of size $n = 500$, a number comparable to the sample size of 480 in our motivating cohort study. We recorded the average estimate across the 1000 datasets, the empirical bias, the empirical standard error, the average model-based standard error and the empirical coverage probability. Tables 3 and 4 display the results to facilitate a comparison of the estimators based on the two-part generalized propensity score with the methods above for the additive and multiplicative outcome models. We find that use of the two-part propensity score model significantly improves the regression estimate when the exposure variable is semi-continuous in nature. Even in cases when the assumptions of the direct regression model are appropriate (i.e. additive models with a constant treatment effect), using a conventional generalized propensity score with regression adjustment produced estimators with substantial bias compared to estimators based on the two-part model propensity score. Moreover, our simulation results demonstrate that use of the two-part propensity score model is increasingly important as the per cent of unexposed individuals increases. When the continuous exposure data are generated from an alternative error distribution within the location-scale family of models the model describe in Equation (4) is misspecified. Figure 2 shows the bias resulting from each method of regression adjustment for the scenario in which 50% of the subjects are unexposed and the positive exposure values are generated from log-normal distribution. Similarly, Figure 3 shows the bias for the scenario where 50% of the subjects are unexposed and the positive exposure values arise from a Weibull regression model. It is clear that the two-part propensity score model leads to better performance than the alternative approaches when the true propensity score model involves an extreme value error distribution.

We also considered scenarios where we varied the percentage of unexposed subjects, the effect of Z_1 on the exposure status, the effect of Z_2 on the extent of exposure among exposed individuals and the effect of covariates on the continuous outcome Y_1 and Y_2 . We thus constructed 8 different scenarios for the two outcome models. Table S1–S2 provided in the supplementary material compares the performance of the two-part generalized propensity score with the methods above for the additive and multiplicative outcome models for the scenarios considered above. These results also showed that improvement offered by our method as the per cent of subjects who are not exposed increases in the sample and the effects of covariates Z_1 and Z_2 on the exposure variable and the outcome variables become stronger.

4 | APPLICATION

We evaluated the performance of our method using data from the Detroit Longitudinal Cohort study. We focus on the subset of 232 subjects with complete data since our purpose is to illustrate the use of propensity score estimation for long-tailed, semi-continuous prenatal alcohol exposure variables. In practice, it would be important to incorporate appropriate strategies for handling missing data. However, further discussion on this issue is beyond the scope of this paper. The effect of observed covariates on the association between prenatal alcohol exposure and the continuous outcome variable, the WISC Freedom from Distractibility measure at age 7 years was examined using four methods: (1) conventional regression adjustment, (2) covariate adjustment using the generalized propensity score, (3) covariate adjustment using a two-part generalized propensity score and (4) adjustment using a two-part generalized propensity score via stratification (5 and 10 strata).

For the two-part propensity score method, we first modelled the probability of drinking any alcohol during pregnancy on all covariates using logistic regression. We then modelled the conditional distribution of the log of alcohol consumption given all covariates using Gaussian linear regression. The covariates, which are listed in Table 1 and 2, were selected because they are expected to affect child outcomes and also to be likely to correlate with prenatal alcohol exposure. If these variables are left uncontrolled, they could confound the estimated exposure effect on the outcome in question. We included covariates that may be unrelated to the exposure but related to the outcome in the propensity score model. It has been reported that inclusion of these variables increases the precision of the estimated exposure effect without increasing bias (Brookhart et al., 2006). To evaluate the relation of these covariates to the quantity of alcohol consumed by the mother during pregnancy, we computed the absolute values of the Pearson correlations between each covariate and the exposure variable for those cases in which the mother used alcohol during pregnancy (Fong et al., 2018). We found that the correlation between the exposure and each covariate was reduced substantially after conditioning on the propensity score, indicating that the covariate balance is significantly improved. We further evaluated the balance of covariates by employing the graphical method suggested by Imai and van Dyk (2004). Figure 4(a) shows the standard normal quantile plot of the t-statistics obtained by regressing each covariate on the exposure variable using linear or logistic regression for continuous and indicator covariates respectively. The magnitude of the t-statistics shows an obvious lack of balance. Figure 4(b) shows the magnitude of t-statistics after we control for the estimated propensity score. The figure shows the substantial reduction in the t-statistics obtained by conditioning on the estimated propensity score. These figures are constructed including the square of the prenatal visit and gestational age variables. Inclusion of the square of these variables in the estimation of the propensity score improves the balance. The WISC Freedom from Distractibility Index was then regressed on the prenatal alcohol exposure variable and the estimated propensity score to estimate the average effect of prenatal alcohol exposure on the response. We conducted regression diagnostics for the outcome model to assess whether there is evidence that we need to expand the main effects model. There was insufficient evidence to reject the adequacy of the linearity assumption. (See diagnostic plots provided in the supplementary material). We tested our model for a possible

interaction between the prenatal alcohol exposure and the propensity score. To fit this model, we centred the exposure variable and the propensity score so that the coefficient of the exposure variable represents the effect of a one unit increase in the logarithm of the volume of alcohol consumed per day during pregnancy. It turns out that there was no significant interaction between the exposure variable and the propensity score. To allow for the possibility of a more complex relationship between the propensity score and the response, we fitted a generalized additive model. In this model we let the exposure effect and the effect of propensity score vary smoothly. Figure 5 illustrates the smooth curves for the prenatal alcohol exposure and the estimated propensity score. We did not find evidence of a non-linear relationship between the estimated propensity score and the response.

As an alternative to regression adjustment, we considered stratification on the estimated propensity score. Specifically, we considered the unexposed as one category and the quartiles of the continuous value as defining four more categories to create five strata. We then create four indicator variables with each one indicating a distinct quartile of the positive exposure variable, and regress on these along with the exposure variable itself. To assess the sensitivity of stratification schemes, we repeated the same process with 10 strata.

Table 5 presents the results from the methods based on unadjusted regression analysis, the covariate-adjusted regression model, the regression model adjusted for the generalized propensity score, the regression adjusted for the two-part propensity score and the adjusting for the two-part propensity score via stratification (5 and 10 strata). With the exception of the unadjusted regression method and the stratification with five strata, all of the methods provided very similar estimates, indicating that the WISC Freedom from Distractibility score decreases as the amount of alcohol consumption during pregnancy increases. Adjusting for the estimated propensity score via stratification using five strata tends to result in a smaller effect size than did estimates obtained using other adjustment methods. Indeed the estimated effect size obtained via stratification using five strata is closer to the effect size obtained from the unadjusted analysis. This finding highlights the importance of creating adequate number of strata when adjusting via stratification. Table 6 presents the regression estimates and standard errors from the two-part model. Even though some covariates operated differently in terms of predicting whether or not a mother consumes alcohol, as compared to the amount of alcohol used by those who consumed it, the parameter estimates were fairly similar for the logistic model and for the log-normal model. This reflects the fact that the percentage of zeros was not large enough, and the correlation between the estimated generalized propensity score and the estimated generalized two-part propensity score was quite strong ($\rho = 0.83$); hence, the estimated propensity score was primarily driven by the non-zero values.

As suggested by a referee, we extend our analysis and examine the estimation of full dose–response function using a more flexible approach suggested by Zhao et al. (2020). Specifically, we fit a smooth coefficient model using R package ‘mgcv’ (Wood, 2017) to predict the response variable as a smooth function of the prenatal alcohol exposure and a smooth function of the estimated propensity score along with their interaction through a tensor product smooth. We then estimated the dose–response by averaging predicted values over the empirical distribution of the propensity score, for a set of equally spaced values

of exposure. A comparison of this model fit compared with the one that omitted the tensor product term suggests that adding the interaction did not significantly improve model fit (p -value = 0.10). However it is important to note that because only partial ranges of the propensity score values were observed for most exposure levels, the power to detect an interaction effect was low. Similarly, it means that when we estimated the dose–response curve by averaging over the empirical distribution of propensity score values, the results were subject to bias because they involved extrapolation outside the range of observed data. Indeed, confidence intervals were very wide for the resulting predicted dose–response. Zhao et al. (2020) found a similar issue with their application and recommended caution in such settings.

5 | DISCUSSION

In this paper, we developed a propensity score methodology for use with a semi-continuous exposure/treatment variable. In particular, our strategy allows researchers to estimate causal effects after adjustment for potential confounding variables based on an estimated propensity score rather than on a large number of potential confounders when the exposure/treatment variable includes both a large number of non-drinking mothers and continuously distributed levels of alcohol consumption. Our approach also allows covariates to operate differently in terms of predicting whether or not the subject is exposed, as compared to the level of exposure. We evaluate and compare our method with conventional covariate adjustment and the generalized propensity score method.

In our simulations we observed that the use of the propensity score based on a two-part model as an additional covariate can successfully reduce bias relative to standard regression adjustment and the generalized propensity score method in all scenarios considered here. Our simulation results suggested that when the exposure variable is semi-continuous, a propensity score based on a two-part model provides a better alternative to both conventional regression adjustment and a generalized propensity score. Our work and that of a number of other authors suggests that propensity score methodologies can be sensitive to model misspecification. In this paper, we derived an explicit expression for the asymptotic bias in the estimator of the causal parameter arising from use of a misspecified propensity score. Specifically, we calculated the asymptotic bias associated with using the generalized propensity score in settings where the exposure variable is semi-continuous in nature and a two-part regression model is more appropriate. We found that the asymptotic bias was aligned with the empirical bias from the simulation studies as expected (Table S1–S2). The expression for the asymptotic bias from the misspecified propensity score model generalizes the findings of Hade and Lu (2014), who investigated the impact of using estimated propensity score in lieu of the true propensity score, which is, of course, necessary in such applications. We emphasize that while our derivations were carried out primarily to investigate the importance of two-part models for semi-continuous exposure data, they offer general insight into the impact of propensity score model specification. We evaluate the performance of our method when there is some form of model misspecification. Specifically, we explore the performance of the log-normal model when the true propensity score model involves a Weibull regression model. We found that even in this misspecified setting, our method performs better than the other approaches considered here.

We estimated the average effect of prenatal alcohol exposure on the WISC Freedom from Distractibility at age 7 using four methods: (1) conventional regression adjustment, (2) covariate adjustment using the generalized propensity score, (3) covariate adjustment using a two-part generalized propensity score and (4) adjustment using a two-part generalized propensity score via stratification (5 and 10 strata). All methods agreed that the WISC Freedom from Distractibility score decreases as the amount of alcohol consumption during pregnancy increases. Even though stratification is more flexible way to adjust for the propensity score, our results underlined that creating smaller number of strata may cause bias on the estimated effect size.

In our response model we assumed a linear relationship between the estimated propensity score and the response. As noted by Schafer and Kang (2008), it is important to check the assumption that the propensity score is linearly related to the response. To assess the need for a more complex relationship between the propensity score and the response, we tested our model for a possible interaction between the propensity score and the exposure variable. Furthermore, we fit a smooth coefficient model by including a smooth estimated propensity score. We found no evidence of non-linear relationship between the propensity score and the outcome.

In this paper, we extended the use of generalized propensity score to settings where the exposure variable is semi-continuous in nature. Specifically, we use the mean obtained from the two-part model as the estimated propensity score to generalize the propensity score method suggested by Imai and van Dyk (2004). However, one can consider to use two elements of the two-part propensity score separately in the outcome model and facilitate causal inference regarding the effects of any alcohol consumption and the incremental change in the volume of alcohol consumed. As a future research, we are exploring approaches for propensity scores for multivariate exposure data.

As in many studies involving environmental exposures, a complication for our application was the very long tail of the prenatal alcohol exposure distribution. In practice we have found that it works well to use a log transformation of alcohol to reduce the effect of outliers when predicting the developmental outcomes, but this approach complicates construction of the propensity score based on two-part models. To overcome this complication we used a smooth function of the propensity score using the R package ‘mgcv’ (Wood, 2017) rather than a linear function of the propensity score function. Missing data are another complication that needs to be addressed in our data application. For the purpose of this paper we have conducted a complete case analysis yielding a sample of 232 subjects. However, in practice, it will be important to properly take account of missing data since complete case analysis involving the propensity score produces biased causal inference unless the data are missing completely at random (D’Agostino & Rubin, 2000).

It is also important to point out that our application involves a cohort of socio-economically disadvantaged women, who were asked to provide sensitive information regarding their exposure to alcohol and other substances of abuse in pregnancy. For this reason, there can be a concern regarding under-reporting. This concern is ameliorated in part by the fact that study investigators used a state-of-the-art timeline follow-back interview method, which

has been validated in relation to levels of fatty acid ethyl esters in meconium samples (Bearer et al., 2003) and successfully predicted a broad range of developmental outcomes (Jacobson et al., 2002; Lindinger et al., 2016). There were also other sources of information regarding prenatal exposure, including toxicological screens performed during pregnancy that can identify women who used cocaine or other substances but denied using them when interviewed (Jacobson et al., 2002). Developing analytical strategies to take account of these kinds of issues is part of our ongoing work on the project.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENT

We thank Neil Dodge, Ph.D., for his assistance with data management.

Support:

This research was funded by grants to Sandra W. Jacobson and Joseph L. Jacobson from the National Institutes of Health/National Institute on Alcohol Abuse and Alcoholism (NIH/NIAAA; R01-AA025905) and the Lycaki-Young Fund from the State of Michigan. Richard J. Cook was supported by the Natural Sciences and Engineering Research Council of Canada through grants RGPIN 155849 and RGPIN 04207. Louise Ryan was also supported by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) CE140100049. Data collection for the Detroit Longitudinal Study was supported by grants from NIH/NIAAA (R01-AA06966, R01-AA09524, and P50-AA07606) and NIH/National Institute on Drug Abuse (R21-DA021034).

APPENDIX A

BIAS FROM A MISSPECIFIED PROPENSITY SCORE IN REGRESSION

Let \mathbf{Z} be an $r \times 1$ vector of confounders and $\mathbf{W}(\mathbf{Z})$ be a $q \times 1$ function of these. If we let $\mathbf{V} = (1, X, \mathbf{W}'(\mathbf{Z}))'$ be a $(q+2) \times 1$ vector of covariates and $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \dots)'$ be a $(q+2) \times 1$ vector of parameters, we can consider a linear model of the form:

$$Y = \theta_0 + \theta_1 X + \boldsymbol{\theta}_2 \mathbf{W}(\mathbf{Z}) + E \quad (\text{A1})$$

where we assume $(X, \mathbf{Z}) \perp \perp E$ and $E \sim N(0, \sigma^2)$. Now consider a propensity score estimated based on a linear model for $X|\mathbf{Z}$ with estimation of the regression coefficients via using ordinary least squares (OLS) to give an estimate:

$$\hat{\mu} = \mu(\mathbf{Z}; \hat{\boldsymbol{\alpha}})$$

We then regress on Y on X and the true propensity score, we fit a linear model

$$Y = \beta_0 + \beta_1 X + \beta_2 \mu(\mathbf{Z}; \hat{\boldsymbol{\alpha}}) + \epsilon \quad (\text{A2})$$

under the assumption $(X, \mu(\mathbf{Z}; \boldsymbol{\alpha})) \perp \perp \epsilon$ with $\epsilon \sim N(0, \tau^2)$. We aim to derive the limiting value of the estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ from fitting (A2), when the linear propensity score is misspecified. To do so, we let $\bar{\mathbf{V}} = (1, X, \mu(\mathbf{Z}))'$ and $\mu(\mathbf{Z}) = E(X | \mathbf{Z}; \boldsymbol{\alpha})$ and write the least squares equations for $\boldsymbol{\beta}$ in vector form with the contribution from single individual given by

$$\mathbf{U}(Y, X, \mu(\mathbf{Z}); \boldsymbol{\beta}) = \bar{\mathbf{V}}(Y - \bar{\mathbf{V}}' \boldsymbol{\beta}) \quad (\text{A3})$$

Note that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$, the solution to $E\{\mathbf{U}(Y, X, \mathbf{Z}; \mu(\mathbf{Z}), \boldsymbol{\beta})\} = 0$, where the expectation is taken with respect to true distribution involving Y, X, \mathbf{Z} . We evaluate this expectation in stages in the following subsections.

A.1 The expectation with respect to $Y \mid X, \mathbf{Z}$

We first take the expectation with respect to Y given X, \mathbf{Z} so require

$$E_Y\{\mathbf{U}(Y, X, \mu(\mathbf{Z}); \boldsymbol{\beta}) \mid X, \mathbf{Z}\} = E_Y\{\bar{\mathbf{V}}(Y - \bar{\mathbf{V}}' \boldsymbol{\beta}) \mid X, \mathbf{Z}\}$$

This gives elements,

$$\mathbf{D}(\bar{\mathbf{V}})(E(Y \mid X, \mathbf{Z}; \theta) - \bar{\mathbf{V}}' \boldsymbol{\beta}), \quad (\text{A4})$$

where $\mathbf{D}(\bar{\mathbf{V}}) = \frac{\partial \bar{\mathbf{V}}' \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}$

A.2 The expectation with respect to $X \mid \mathbf{Z}$

The next step requires computation of $E\{\mathbf{D}(\bar{\mathbf{V}})[E(Y \mid X, \mathbf{Z}; \theta) - \bar{\mathbf{V}}' \boldsymbol{\beta}] \mid \mathbf{Z}\}$. We give the form of each entry in this 3×1 vector explicitly as follows:

$$\begin{aligned} \theta_0 + \theta_1 E(X \mid \mathbf{Z}) + \theta_2 \mathbf{W}(\mathbf{Z}) - (\beta_0 + \beta_1 E(X \mid \mathbf{Z}) + \beta_2 \mu(\mathbf{Z})) &= 0 \\ \theta_0 E(X \mid \mathbf{Z}) + \theta_1 E(X^2 \mid \mathbf{Z}) + \theta_2 E((X \mid \mathbf{Z}) \mathbf{W}(\mathbf{Z})) - (\beta_0 E(X \mid \mathbf{Z}) + \beta_1 E(X^2 \mid \mathbf{Z}) + \beta_2 E(X \mid \mathbf{Z}) \mu(\mathbf{Z})) &= 0 \\ \theta_0 \mu(\mathbf{Z}) + \theta_1 E(X \mid \mathbf{Z}) \mu(\mathbf{Z}) + \theta_2 (\mu(\mathbf{Z}) \mathbf{W}(\mathbf{Z})) - (\beta_0 \mu(\mathbf{Z}) + \beta_1 E(X \mid \mathbf{Z}) \mu(\mathbf{Z}) + \beta_2 \mu(\mathbf{Z})^2) &= 0 \end{aligned}$$

The Expectation with Respect to \mathbf{Z} From the first line of (A5) we get

$$\theta_0 + \theta_1 E(X) + \theta_2 E(\mathbf{W}(\mathbf{Z})) - (\beta_0 + \beta_1 E(X) + \beta_2 E(\mu(\mathbf{Z}))) = 0 \quad (\text{A6})$$

From the second line we get

$$\begin{aligned} \theta_0 E(X) + \theta_1 E(X^2) + \theta_2 [\boldsymbol{\eta}(\mathbf{Z}) + E(X) E(\mathbf{W}(\mathbf{Z}))] \\ - (\beta_0 E(X) + \beta_1 E(X^2) + \beta_2 [\text{cov}(E(X \mid \mathbf{Z}), \mu(\mathbf{Z})) + E(X) E(\mu(\mathbf{Z}))]) = 0 \end{aligned} \quad (\text{A7})$$

where $\boldsymbol{\eta}(\mathbf{Z}) = (\eta_1(\mathbf{Z}), \eta_2(\mathbf{Z}), \dots, \eta_q(\mathbf{Z}))'$ where $\eta_j(\mathbf{Z}) = \text{cov}(E(X \mid \mathbf{Z}), W_j(\mathbf{Z}))$, $j = 1, \dots, q$.

From the last line of (A5) we get

$$\begin{aligned} \theta_0 E(\mu(\mathbf{Z})) + \theta_1 [\text{cov}(E(X \mid \mathbf{Z}) \mu(\mathbf{Z})) + E(X) E(\mu(\mathbf{Z}))] + \theta_2 [\boldsymbol{\phi}(\mathbf{Z}) + E(\mu(\mathbf{Z})) E(\mathbf{W}(\mathbf{Z}))] \\ - (\beta_0 E(\mu(\mathbf{Z})) + \beta_1 [\text{cov}(E(X \mid \mathbf{Z}) \mu(\mathbf{Z})) + E(X) E(\mu(\mathbf{Z}))] + \beta_2 E(\mu(\mathbf{Z})^2)) = 0 \end{aligned} \quad (\text{A8})$$

where $\boldsymbol{\phi}(\mathbf{Z}) = \phi_1(\mathbf{Z}), \phi_2(\mathbf{Z}), \dots, \phi_q(\mathbf{Z})$ where $\phi_j(\mathbf{Z}) = \text{cov}(\mu(\mathbf{Z}), \mathbf{W}_j(\mathbf{Z}))$, $j = 1, \dots, q$. Solve equation (A6) to get β_0 :

$$\beta_0 = \theta_0 + \theta_1 E(X) + \theta_2' E(\mathbf{W}(\mathbf{Z})) - \beta_1 E(X) - \beta_2 E(\mu(\mathbf{Z}))$$

We next replace β_0 in (A7) to get

$$\begin{aligned} & \theta_0 E(X) + \theta_1 E(X^2) + \theta_2' [\boldsymbol{\eta}(\mathbf{Z}) + E(X)E(\mathbf{W}(\mathbf{Z}))] \\ & - (E(X)(\theta_0 + \theta_1 E(X) + \theta_2' E(\mathbf{W}(\mathbf{Z})) - \beta_1 E(X) \\ & - \beta_2 E(\mu(\mathbf{Z}))) + \beta_1 E(X^2) + \beta_2 [\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) + E(X)E(\mu(\mathbf{Z}))]) = 0 \end{aligned}$$

$$\begin{aligned} & \theta_0 E(X) + \theta_1 E(X^2) + \theta_2' [\boldsymbol{\eta}(\mathbf{Z}) + E(X)E(\mathbf{W}(\mathbf{Z}))] \\ & - (\theta_0 E(X) + \theta_1 E(X)^2 + \theta_2' E(\mathbf{W}(\mathbf{Z}))E(X) - \beta_1 E(X)^2 \\ & - \beta_2 E(\mu(\mathbf{Z}))E(X) + \beta_1 E(X^2) + \beta_2 [\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) + E(X)E(\mu(\mathbf{Z}))]) = 0 \end{aligned}$$

$$\begin{aligned} & \theta_0 E(X) + \theta_1 E(X^2) + \theta_2' \text{cov}(E(X | \mathbf{Z}), \mathbf{W}(\mathbf{Z})) + \theta_2' E(X)E(\mathbf{W}(\mathbf{Z})) \\ & - \theta_0 E(X) - \theta_1 E(X)^2 - \theta_2' E(\mathbf{W}(\mathbf{Z}))E(X) + \beta_1 E(X)^2 \\ & + \beta_2 E(\mu(\mathbf{Z}))E(X) - \beta_1 E(X^2) - \beta_2 \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) - \beta_2 E(X)E(\mu(\mathbf{Z})) = 0 \end{aligned}$$

$$\begin{aligned} & \theta_1 [E(X^2) - E(X)^2] + \theta_2' \boldsymbol{\eta}(\mathbf{Z}) - \beta_1 [E(X)^2 - E(X)^2] \\ & - \beta_2 \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) = 0 \end{aligned} \tag{A9}$$

Likewise replace β_0 in equation (A8) to obtain a simplified expression in β_1 and β_2

$$\begin{aligned} & \theta_0 E(\mu(\mathbf{Z})) + \theta_1 [\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) + E(X)E(\mu(\mathbf{Z}))] + \theta_2' [\boldsymbol{\phi}(\mathbf{Z}) + E(\mu(\mathbf{Z}))E(\mathbf{W}(\mathbf{Z}))] \\ & - [(E(\mu(\mathbf{Z}))(\theta_0 + \theta_1 E(X) + \theta_2' E(\mathbf{W}(\mathbf{Z})) - \beta_1 E(X) - \beta_2 E(\mu(\mathbf{Z}))) \\ & + \beta_1 [\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) + E(X)E(\mu(\mathbf{Z}))] + \beta_2 E(\mu(\mathbf{Z})^2)] = 0 \end{aligned}$$

$$\begin{aligned} & \theta_0 E(\mu(\mathbf{Z})) + \theta_1 \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) + \theta_1 E(X)E(\mu(\mathbf{Z})) + \theta_2' \boldsymbol{\phi}(\mathbf{Z}) + \theta_2' E(\mu(\mathbf{Z}))E(\mathbf{W}(\mathbf{Z})) \\ & - \theta_0 E(\mu(\mathbf{Z})) - \theta_1 E(X)E(\mu(\mathbf{Z})) - \theta_2' E(\mathbf{W}(\mathbf{Z}))E(\mu(\mathbf{Z})) + \beta_1 E(X)E(\mu(\mathbf{Z})) + \beta_2 E(\mu(\mathbf{Z}))^2 \\ & - \beta_1 \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) - \beta_1 E(X)E(\mu(\mathbf{Z})) - \beta_2 E(\mu(\mathbf{Z})^2) = 0 \end{aligned}$$

$$\begin{aligned} & \theta_1 \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) + \theta_2' \boldsymbol{\phi}(\mathbf{Z}) \\ & - \beta_1 \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) - \beta_2 [E(\mu(\mathbf{Z})^2) - E(\mu(\mathbf{Z}))^2] = 0 \end{aligned}$$

Solve for β_2 we obtain

$$\beta_2 = \frac{[(\theta_1 - \beta_1)\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z}))) + \theta_2\phi(\mathbf{Z})]}{V(\mu(\mathbf{Z}))}$$

Replacing β_2 in (A9) we obtain

$$\begin{aligned} & \theta_1 V(X) + \theta_2 \eta(\mathbf{Z}) - \beta_1 V(X) \\ & - \left[\frac{[(\theta_1 - \beta_1)\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z}))) + \theta_2\phi(\mathbf{Z})]}{V(\mu(\mathbf{Z}))} \right] \text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z})) = 0 \end{aligned}$$

Re-organize terms:

$$\begin{aligned} & \theta_1 V(X) + \theta_2 \eta(\mathbf{Z}) - \left\{ \theta_1 \left[\frac{\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z}))}{V(\mu(\mathbf{Z}))} \right] + \theta_2 \left[\frac{\phi(\mathbf{Z})}{V(\mu(\mathbf{Z}))} \right] \right\} \text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z}))) \\ & = \beta_1 \left[V(X) - \frac{(\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z})))\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z}))))}{V(\mu(\mathbf{Z}))} \right] \end{aligned}$$

The whole expression can be re-written as:

$$\theta_1 \left[V(X) - \left[\frac{\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z}))\text{cov}(E(X | \mathbf{Z}), \mu(\mathbf{Z}))}{V(\mu(\mathbf{Z}))} \right] \right] \quad (\text{A})$$

$$+ \theta_2 \left[\eta(\mathbf{Z}) - \frac{\phi(\mathbf{Z})\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z})))}{V(\mu(\mathbf{Z}))} \right] \quad (\text{B})$$

$$= \beta_1 \left[V(X) - \frac{\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z}))\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z}))))}{V(\mu(\mathbf{Z}))} \right] \quad (\text{C})$$

Term (A) can be re-written as

$$\begin{aligned} & \theta_1 \left[V(X) - V(E(X | \mathbf{Z}))\text{corr}^2(E(X | \mathbf{Z}), (\mu(\mathbf{Z}))) \right] \\ & = \theta_1 \left[V(X) \left[1 - \text{corr}^2(E(X | \mathbf{Z}), (\mu(\mathbf{Z}))) \right] + E(V(X | \mathbf{Z}))\text{corr}^2(E(X | \mathbf{Z}), (\mu(\mathbf{Z}))) \right] \end{aligned}$$

And term (C) can be re-written as:

$$\beta_1 \left[V(X) \left[1 - \text{corr}^2(E(X | \mathbf{Z}), \mu(\mathbf{Z})) \right] + E(V(X | \mathbf{Z}))\text{corr}^2(E(X | \mathbf{Z}), \mu(\mathbf{Z})) \right]$$

So the whole expression can be written as:

$$\begin{aligned} & \theta_1 \left[V(X) - V(E(X | \mathbf{Z}))\text{corr}^2(E(X | \mathbf{Z}), (\mu(\mathbf{Z}))) \right] \\ & + \theta_2 \left[\eta(\mathbf{Z}) - \frac{\phi(\mathbf{Z})\text{cov}(E((X | \mathbf{Z})\mu(\mathbf{Z})))}{V(\mu(\mathbf{Z}))} \right] \\ & = \beta_1 \left[V(X) - V(E(X | \mathbf{Z}))\text{corr}^2(E(X | \mathbf{Z}), (\mu(\mathbf{Z}))) \right] \end{aligned}$$

We can re-express this as

$$\beta_1 = \theta_1 + \frac{\theta_2[\eta(\mathbf{Z}) - \phi(\mathbf{Z})\text{corr}(E(X | \mathbf{Z}), \mu(\mathbf{Z}))\sqrt{\frac{V(E(X | \mathbf{Z}))}{V(\mu(\mathbf{Z}))}}]}{V(X) - V(E(X | \mathbf{Z}))\text{corr}^2(E(X | \mathbf{Z}), (\mu(\mathbf{Z})))}$$

To help interpret this expression, it is helpful to define ρ to represent the correlation between the propensity score, $\mu(\mathbf{Z})$ and the true expected value of X given Z , namely:

$$\rho = \text{corr}(E(X | \mathbf{Z}), \mu(\mathbf{Z})).$$

In this case, we can rewrite the last expression as follows:

$$\beta_1 = \theta_1 + \frac{\theta_2[\eta(\mathbf{Z}) - \phi(\mathbf{Z})\rho\sqrt{\frac{V(E(X | \mathbf{Z}))}{V(\mu(\mathbf{Z}))}}]}{V(X) - V(E(X | \mathbf{Z}))\rho^2}.$$

REFERENCES

- Austin PC (2011) An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3), 399–424. [PubMed: 21818162]
- Bearer CF, Jacobson JL, Jacobson SW, Barr D, Croxford J, Molteno CD et al. (2003) Validation of a new biomarker of fetal exposure to alcohol. *The Journal of Pediatrics*, 143(4), 463–469. [PubMed: 14571221]
- Bia M & Mattei A (2008) A stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. *Stata Journal*, 8(3), 354–373(20).
- Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J & Stürmer T (2006) Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12), 1149–1156. [PubMed: 16624967]
- Chai H, Jiang H, Lin L & Liu L (2018) A marginalized two-part beta regression model for microbiome compositional data. *PLOS Computational Biology*, 14(7), 1–16.
- Choi Y & Hwang B (2018) A bayesian latent class model for effect of environmental pollutants on female infertility. *Journal of the Korean Data and Information Science Society*, 29, 1257–1268.
- Crump K (1984) An improved procedure for low-dose carcinogenic risk assessment from animal data. *Journal of Environmental Pathology, Toxicology and Oncology: Official Organ of the International Society for Environmental Toxicology and Cancer*, 5(4–5), 339–348. [PubMed: 6520736]
- D’Agostino RB & Rubin DB (2000) Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451), 749–759.
- Duan N, Manning WG, Morris CN & Newhouse JP (1983) A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, 1(2), 115–126.
- Fichera E, Emsley R & Sutton M (2016) Is treatment intensity associated with healthier lifestyle choices? an application of the dose response function. *Economics and Human Biology*, 23, 149–163. [PubMed: 27648973]
- Fong C, Hazlett C & Imai K (2018) Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12, 156–177.
- Hade EM & Lu B (2014) Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine*, 33(1), 74–87. 10.1002/sim.5884 [PubMed: 23787715]
- Hernán MA & Robins JM (2006) Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7), 578–86. [PubMed: 16790829]

- Hirano K & Imbens GW (2004) The propensity score with continuous treatments, chapter The propensity score with continuous treatments, by Hirano K and Imbens GW. (Eds), Hoboken: John Wiley and Sons, Ltd, pp. 73–84.
- Imai K & van Dyk DA (2004) Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467), 854–866.
- Imbens GW (1999) The role of the propensity score in estimating dose-response functions. NBER Technical Working Papers 0237, National Bureau of Economic Research, Inc.
- Jacobson SW, Chiodo LM, Sokol RJ & Jacobson JL (2002) Validity of maternal report of prenatal alcohol, cocaine, and smoking in relation to neurobehavioral outcome. *Pediatrics*, 109(5), 815–825. [PubMed: 11986441]
- Jacobson SW, Jacobson JL, Sokol RJ, Chiodo LM & Corobana R (2004) Maternal age, alcohol abuse history, and quality of parenting as moderators of the effects of prenatal alcohol exposure on 7.5-years intellectual function. *Alcoholism: Clinical and Experimental Research*, 28(11), 1732–1745. [PubMed: 15547461]
- Jacobson JL, Dodge NC, Burden MJ, Klorman R & Jacobson SW (2011) Number processing in adolescents with prenatal alcohol exposure and adhd: Differences in the neurobehavioral phenotype. *Alcoholism: Clinical and Experimental Research*, 35(3), 431–442. [PubMed: 21158874]
- Lewis CE, Thomas KGF, Molteno CD, Kliegel M, Meintjes EM, Jacobson JL et al. (2016) Prospective memory impairment in children with prenatal alcohol exposure. *Alcoholism: Clinical and Experimental Research*, 40(5), 969–978. [PubMed: 27028983]
- Lindinger NM, Malcolm-Smith S, Dodge NC, Molteno CD, Thomas KGF, Meintjes EM et al. (2016) Theory of mind in children with fetal alcohol spectrum disorders. *Alcoholism: Clinical and Experimental Research*, 40(2), 367–376. [PubMed: 26842255]
- Madden CW, Mackay BP, Skillman SM, Ciol M & Diehr PK (2000) Risk adjusting capitation: Applications in employed and disabled populations. *Health Care Management Science*, 3(2), 101–109. [PubMed: 10780278]
- Mattson SN, Riley EP, Gramling L, Delis DC & Jones KL (1997) Heavy prenatal alcohol exposure with or without physical features of fetal alcohol syndrome leads to iq deficits. *The Journal of Pediatrics*, 131(5), 718–721. [PubMed: 9403652]
- Moodie EE & Stephens DA (2012) Estimation of dose-response functions for longitudinal data using the generalised propensity score. *Statistical Methods in Medical Research*, 21(2), 149–166. [PubMed: 20442194]
- Naimi A, Moodie E, Auger N & Kaufman J (2014) Constructing inverse probability weights for continuous exposures a comparison of methods. *Epidemiology*, 25, 292–9. [PubMed: 24487212]
- Neelon B, Zhu L & Neelon SEB (2015) Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures. *Biostatistics*, 16(3), 465–479. [PubMed: 25649743]
- Rosenbaum PR & Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Rosenbaum PR & Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387), 516–524.
- Rosenbaum PR & Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1), 33–38.
- Rotnitzky A & Wypij D (1994) A note on the bias of estimators with missing data. *Biometrics*, 50(4), 1163–1170. [PubMed: 7786998]
- Schafer JL & Kang J (2008) Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*, 13(4), 279–313. 10.1037/a0014268 [PubMed: 19071996]
- Smith VA, Neelon B, Maciejewski ML & Preisser JS (2017) Two parts are better than one: modeling marginal means of semicontinuous data. *Health Services and Outcomes Research Methodology*, 17(3), 198–218.

- Su L, Tom BDM & Farewell VT (2015) A likelihood-based two-part marginal model for longitudinal semicontinuous data. *Statistical Methods in Medical Research*, 24(2), 194–205. [PubMed: 21873302]
- Wang H & Heitjan DF (2008) Modeling heaping in self-reported cigarette counts. *Statistics in Medicine*, 27(19), 3789–3804. [PubMed: 18407584]
- White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1–25.
- Winship C & Morgan SL (1999) The estimation of causal effects from observational data. *Annual Review of Sociology*, 25(1), 659–706.
- Wood S (2017) *Generalized additive models: an introduction with R*, 2nd edn. Boca Raton: Chapman and Hall/CRC.
- Zhang Z, Zhou J, Cao W & Zhang J (2016) Causal inference with a quantitative exposure. *Statistical Methods in Medical Research*, 25(1), 315–335. [PubMed: 22729475]
- Zhao S, van Dyk DA & Imai K (2020) Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Statistical Methods in Medical Research*, 29(3), 709–727. [PubMed: 32186266]

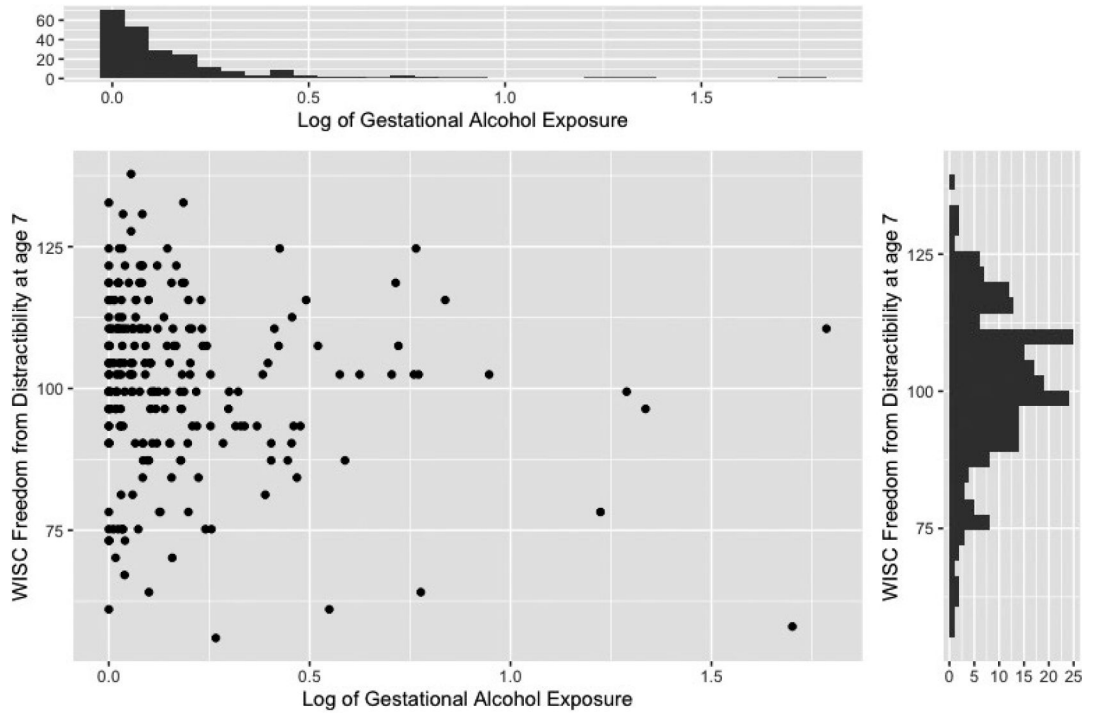


FIGURE 1.
The relation between the Wechsler Intelligence Scale for Children (WISC) Freedom from Distractibility Score at age 7 and log of gestational alcohol exposure

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

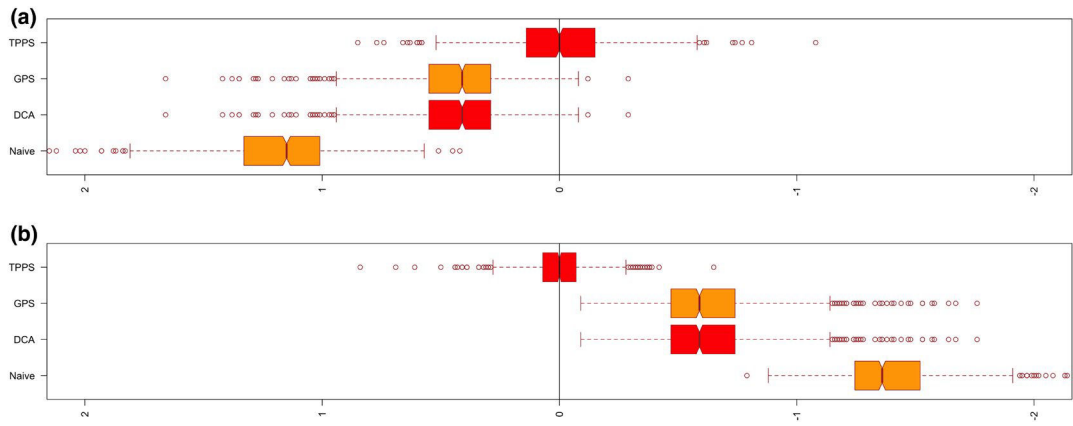


FIGURE 2. Bias of regression adjustment methods from the simulation study where the true propensity score is generated from log-normal distribution. A, Additive outcome model; B, Multiplicative outcome model. Naive, Unadjusted model; DCA, Direct covariate adjustment; GPS, Generalized propensity score; TPPS, Two-part propensity score

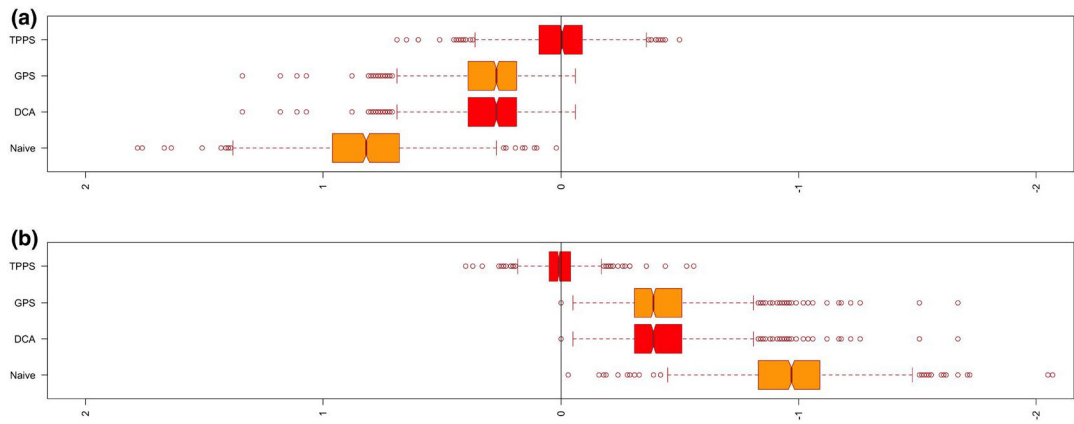


FIGURE 3. Bias of regression adjustment methods from the simulation study where the true propensity score involves extreme value error distribution. A, Additive outcome model; B, Multiplicative outcome model. Naive, Unadjusted model; DCA, Direct covariate adjustment; GPS, Generalized propensity score; TPPS, Two-part propensity score

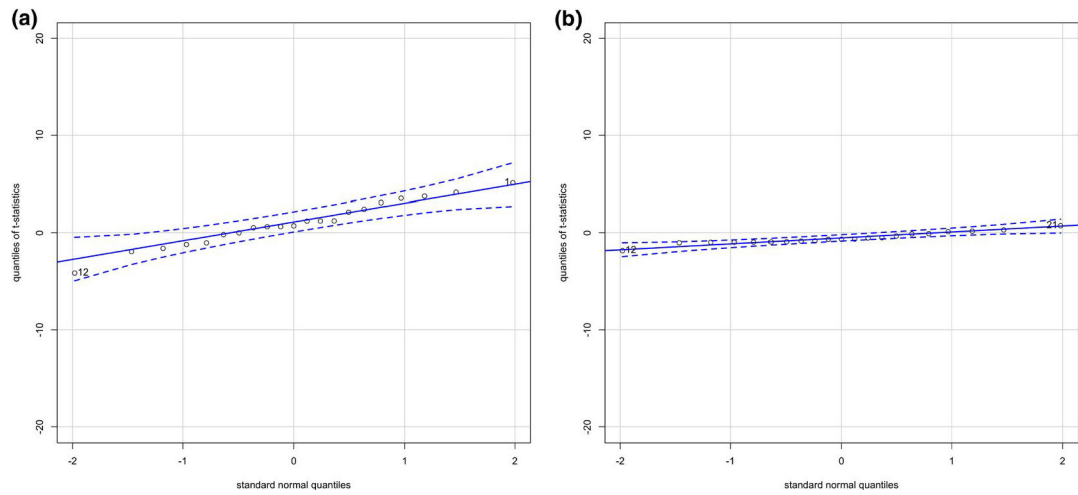


FIGURE 4. Standard normal quantile plots of t-statistics for the coefficient of log of gestational alcohol exposure in the models predicting each covariate including the square of the number of prenatal visits and the gestational age at screening, (a) Without controlling for the estimated propensity score and (b) Controlling for the estimated propensity score

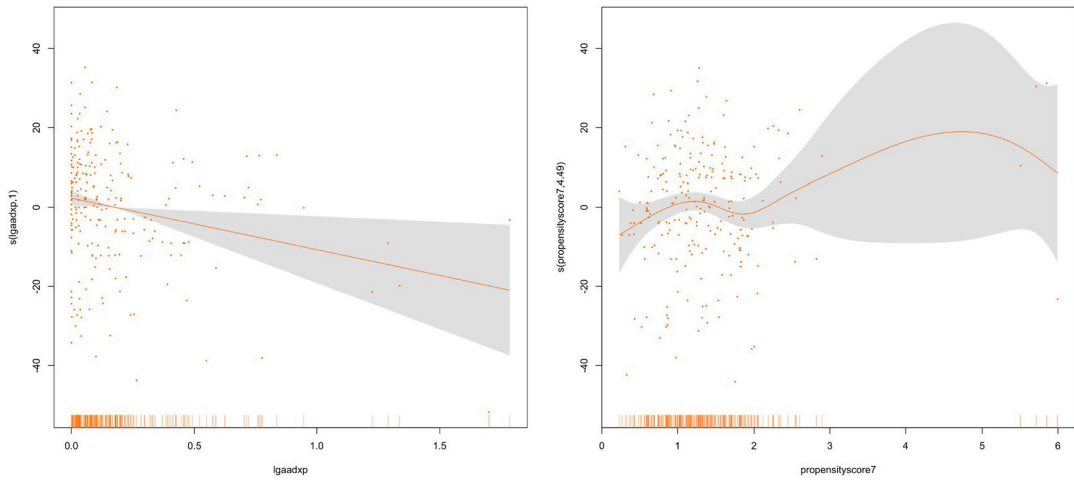


FIGURE 5. Generalized Additive Model (GAM) plots showing the partial effects of log of gestational alcohol exposure and the estimated propensity score on the Wechsler Intelligence Scale for Children (WISC) Freedom from Distractibility Score at age 7

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 1

Summary statistics of continuous covariates in the Detroit Longitudinal Cohort study (N=232)

	Mean	SD
Gestational age at recruitment (weeks)	23.8	7.5
Number of prenatal visits	5.8	3.0
Smoking during pregnancy (cigarettes/day)	9.5	11.2
Prenatal marijuana exposure (days/month)	0.9	2.9
Prenatal cocaine exposure (days/month)	1.0	2.9
Prenatal opiate exposure (days/month)	0.2	1.0
Biological mother's education (years)	11.8	1.6
Biological mother's verbal IQ (PPVT score)	71.7	11.7
Beck Depression Inventory at 6-month postpartum follow-up	11.2	7.5
Age of a child at 7-year follow-up visit (years)	7.9	0.3
Socioeconomic status at age 7-year follow-up ^a	26.2	10.4
Primary caregiver's verbal IQ (PPVT score) at 7-year follow-up	72.9	12.5
HOME score at 7-year follow-up	33.7	6.9
Beck Depression Inventory at 7-year follow-up	8.3	7.4
Number of stressful events at 7-year follow-up ^b	10.2	5.5
Perceived life stress at 7-year follow-up ^b	36.7	27.8

^aHollingshead Four Factor Index of Social Status (Hollingshead, 2011)^bLife Events Scale (Holmes and Rahe, 1967)

TABLE 2

Summary statistics of categorical covariates in the Detroit Longitudinal Cohort

	N	%
Marital status		
Biological mother's		
Married	19	8.2
Not married	213	91.8
Primary care giver's at 7-year follow-up		
Married	205	88.4
Not married	27	11.6
Parity		
0	88	37.9
1	63	27.2
2	51	22.0
3	17	7.3
4	8	3.4
>=5	5	2.2
Gravidity		
1	39	16.8
2	53	22.8
3	47	20.3
4	33	14.2
>=5	60	25.8

TABLE 3

Empirical results for estimators of causal effect of each additional unit exposure on outcome for the scenario where $P(X = 0) = 0.25$

Method	Log-normal				Weibull			
	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
Additive outcome model								
Unadjusted	0.83	0.24	0.16	0.05	0.77	0.22	0.17	0.05
Regression adjustment	0.3	0.18	0.13	0.34	0.27	0.18	0.14	0.47
Generalized propensity score	0.3	0.18	0.16	0.53	0.27	0.18	0.17	0.66
Propensity score based on two-part model	0.0	0.15	0.15	0.96	0.01	0.16	0.15	0.97
Multiplicative outcome model								
Unadjusted	-0.98	0.23	0.17	0.03	-0.99	0.23	0.17	0.03
Regression adjustment	-0.43	0.19	0.14	0.1	-0.45	0.2	0.14	0.1
Generalized propensity score	-0.43	0.19	0.14	0.12	-0.45	0.2	0.15	0.11
Propensity score based on two-part model	0.0	0.09	0.07	0.95	0	0.1	0.08	0.96

^aEBIAS: Empirical bias

^bESE: Empirical standard error

^cASE: Asymptotic standard error

^dECP: Empirical coverage probability

TABLE 4

Empirical results for estimators of causal effect of each additional unit exposure on outcome for the scenario where $P(X = 0) = 0.50$

Method	Log-normal				Weibull			
	EBIAS	ESE	ASE	ECP	EBIAS	ESE	ASE	ECP
Additive outcome model								
Unadjusted	1.18	0.26	0.23	0.01	0.82	0.23	0.16	0.05
Regression adjustment	0.44	0.23	0.18	0.27	0.30	0.17	0.13	0.32
Generalized propensity score	0.44	0.23	0.23	0.48	0.30	0.17	0.16	0.53
Propensity score based on two-part model	0.00	0.22	0.21	0.97	0.00	0.15	0.14	0.96
Multiplicative outcome model								
Unadjusted	-1.39	0.23	0.19	0.00	-0.97	0.23	0.17	0.03
Regression adjustment	-0.63	0.24	0.18	0.03	-0.42	0.18	0.14	0.10
Generalized propensity score	-0.63	0.24	0.19	0.04	-0.42	0.18	0.14	0.11
Propensity score based on two-part model	0.00	0.13	0.10	0.95	0.00	0.08	0.07	0.95

^aEBIAS: Empirical bias

^bESE: Empirical standard error

^cASE: Asymptotic standard error

^dECP: Empirical coverage probability

TABLE 5

Estimated effects of prenatal alcohol exposure on the WISC Freedom based on six models: (1) an unadjusted model, (2) a model using traditional covariate adjustment via linear regression, (3) a model using the generalized propensity score, (4) the proposed two-part generalized propensity score, (5) two-part generalized propensity score via stratification (5 strata) and (6) two-part generalized propensity score via stratification (10 strata)

Method	Average causal effect	SE
Unadjusted model	-7.90	3.65
Covariate-adjusted model	-11.21	4.43
Generalized propensity score	-11.21	4.29
Two-part generalized propensity score	-11.89	5.03
Two-part generalized propensity score via stratification (5 strata)	-8.50	3.96
Two-part generalized propensity score via stratification (10 strata)	-11.04	4.11

TABLE 6

Estimates for the effects of each of the components of the two-part model

	$P(X_i > 0)$		$E(\ln(X_i) X_i > 0)$	
	Estimate	Standard error	Estimate	Standard error
Intercept	0.3823	6.100	3.142	2.804
Gestational age at recruitment (week)	0.008	0.035	0.016	0.016
Number of prenatal visits	0.063	0.088	-0.100	0.041
Smoking during pregnancy (cigarettes/day)	0.113	0.039	0.031	0.009
Prenatal marijuana exposure (days/month)	0.336	0.220	-0.014	0.033
Prenatal cocaine exposure (days/month)	-0.074	0.094	-0.044	0.038
Prenatal opiate exposure (days/month)	-0.050	0.248	0.316	0.101
Biological mother's education (years)	0.196	0.166	0.106	0.072
Biological mother's verbal IQ (PPVT score)	-0.029	0.056	-0.023	0.027
Beck Depression Inventory at 6-month postpartum follow-up	-0.010	0.042	0.000	0.016
Age of a child at 7-year follow-up (years)	-0.000	0.002	-0.002	0.000
Socioeconomic status at age 7-year follow-up	0.023	0.027	0.018	0.011
Primary caregiver's verbal IQ (PPVT score) at 7-year follow-up	0.016	0.057	0.003	0.026
HOME score at 7-year follow-up	-0.063	0.041	0.018	0.017
Beck Depression Inventory at 7-year follow-up	0.029	0.050	0.047	0.019
Number of stressful events at 7-year follow-up	0.191	0.115	0.037	0.043
Perceived life stress at 7-year follow-up	-0.020	0.024	-0.016	0.009
Biological mother's marital status	-0.401	0.750	0.213	0.408
Primary care giver's marital status at 7-year follow-up	-0.109	0.636	-0.273	0.303
Parity	-0.6705	0.338	0.131	0.122
Gravidity	0.3653	0.215	0.016	0.016