Accident Analysis and Prevention 129 (2019) 148–155

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Comparison of empirical Bayes and propensity score methods for road safety evaluation: A simulation study

Haojie Li[a,c,d,*], Daniel J. Graham[b], Hongliang Ding[a,c,d], Gang Ren[a,c,d]

[a] School of Transportation, Southeast University, China
[b] Transport Strategy Centre, Imperial College London, UK
[c] Jiangsu Key Laboratory of Urban ITS, China
[d] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, China

ARTICLE INFO

Keywords:
Road safety evaluation
Causal inference
Empirical Bayes
Propensity score
Doubly robust

ABSTRACT

Statistical evaluation of road safety interventions can be undertaken using a variety of different approaches, typically requiring different assumptions to obtain causal identification. In this paper, we conduct a simulation study to compare the performance of empirical Bayes (EB) and propensity score (PS) based methods, which have featured prominently in the recent literature, in settings with and without violation of key assumptions. The estimators considered include EB, inverse probability weighting (IPW), and Doubly Robust (DR) estimation. We find that while the EB approach has good finite sample properties when model assumptions are met, the consistency of this estimator is substantially diminished when the reference and treated sites follow different functions. The IPW estimator performs well in large samples, but requires a correctly specified PS model with sufficient overlap in covariate distributions between treated and control units. The DR estimator allows for violation of assumptions in either the regression or PS model, but not both. We find that this added level of robustness affords overall better performance than attained via EB or IPW estimation.

1. Introduction

Road accidents place a great burden on individuals, property and society. Over several decades, a considerable body of research has developed with the aim of identifying key factors underpinning the incidence of road accidents, including traffic characteristics, road characteristics, socio-economic and environmental factors (e.g. Gu et al., 2019; Wang et al., 2018; Xu et al., 2019; Lee et al., 2015). In recent years, attention has increasingly turned to evaluation of the effects of road safety interventions. The motivation behind this focus on quantification of the effectiveness of specific safety measures is to help policy makers choose the most appropriate course of action for accident mitigation.

Several approaches have been used in previous road safety evaluation studies. The earliest work was mainly based on simple before–after control methods and cross-sectional regression approaches. It became evident, however, that these approaches can fail to fully address issues such as regression to the mean and confounding, particularity in the presence of site selection bias and unobserved confounding (Hauer, 1997; Tarko et al., 1998; Sasidharan and Donnell, 2013). Widespread adoption of the empirical Bayes (EB) approach arose as a response to the limitation of before–after and cross-sectional regression models. EB is viewed as a statistically defensible means of increasing the precision of estimation and correcting for the regression to the mean bias. However, EB relies on inference relative to a reference group that must be similar to the treatment group in baseline characteristics, and recent studies have shown that the performance of the EB approach can be adversely affected when this assumption is violated (Wood and Donnell, 2017; Lord and Kuo, 2012). In response to this potential limitation of EB, a small number of recent studies have proposed use of propensity score (PS) methods to evaluate the effects of road safety measures (e.g. Karwa et al., 2011; Wood et al., 2015a,b; Sasidharan and Donnell, 2013; Li et al., 2013; Li and Graham, 2016). The PS approach provides a credible mean of ensuring that treated and control units are matched in their baseline characteristics, but it relies on the key identifying assumption that the probability of treatment given confounders is correctly represented in the PS model.

In this paper we conduct a simulation study to compare the performance of EB and PS methods in settings with and without violation of key assumptions. The paper is organized as follows. In section two we describe methods for road safety evaluation based on EB, PS, and outcome regression (OR). We do so using the potential outcomes

* Corresponding author at: School of Transportation, Southeast University, China.
E-mail address: h.li@seu.edu.cn (H. Li).

https://doi.org/10.1016/j.aap.2019.05.015
Received 3 April 2019; Received in revised form 16 May 2019; Accepted 20 May 2019
Available online 28 May 2019
0001-4575/

framework for causal inference, sometimes referred to as the "Rubin causal model". Our simulation study is described in Section 3, followed by presentation of results in Section 4. Discussion and conclusions are given in the final section.

2. Literature review

In this section, we first discuss the estimands in road safety evaluation studies. Then we review the fundamentals of EB and PS methods, and explain the conditions under which they can be applied, as well as their key identifying assumptions and limitations.

2.1. Estimands in road safety evaluation studies

In road safety evaluation studies, traffic interventions or 'treatments' can be policies, legislation and enforcement, physical changes to the network; and other general-purpose measures which directly or indirectly affect traffic conditions, driver behavior and the travel environment. In most cases, treatment is assigned to 'units' (i.e. locations or links etc.) to address specific concerns and is therefore non-randomly assigned. Crucially, under a non-random assignment the treatment is allocated in relation to some baseline characteristics of the unit under study, thus producing a situation in which units differ by treatment status regardless of the effect of the treatment.

Consider a binary treatment denoted by random variable T_i for unit i, where $i = 1, \ldots, n$, and n denote the total population. T_i takes a value of 1 if unit i receives the treatment and a value of 0 otherwise. Let $Y_i(T)$ be the potential outcome of unit i under treatment T, then we define the average treatment effect (ATE) as

$$\varphi_{\text{ATE}} = \frac{\mathbb{E}[Y_i(1)]}{\mathbb{E}[Y_i(0)]}$$

The fundamental problem of causal inference is that it is impossible to observe the outcomes of any i under both treatment status at the same time (Holland, 1986). So the dual expectations in the expression above are not simultaneously observed and have to be estimated. We now review three such estimation methods that have been used recently in safety evaluation studies.

2.2. Empirical Bayes

The EB method combines prior and observed data to derive an estimate for the ATE. In road safety applications, prior information is obtained from a group of similar sites and the observed information is the accident frequency for the treated site. A safety performance function (SPF) is applied to model the relationship between the accident frequency of the control sites and covariates deemed to be relevant. An improved estimate of the long-term accident frequency can be obtained by combining the model-predicted number of accidents with the recorded accident number of treated sites. EB methods have been widely used and discussed in traffic safety studies over the last two decades, especially in before–after evaluations (Wood and Donnell, 2017; Elvik et al., 2017; Hauer, 1992, 1997; Hauer et al., 2002; Li et al., 2008; Miaou and Lord, 2003; Park and Lord, 2007; Persaud et al., 1997, 2004, 2010; Persaud and Lyon, 2007; Quddus, 2008; Aguero-Valverde and Jovanis, 2006; El-Basyouny and Sayed, 2011).

The insight underpinning the EB approach is that "accident counts are not the only clue to the safety of an entity. Another clue is in what is known about the safety of similar entities" (Hauer, 1997). Accordingly, the predicted number of crashes without treatment is derived by combining the observed crash counts in the before treatment period and expected number of crashes from the safety performance functions (SPFs), which relates the accident frequency of the control group to their characteristics (Hauer, 1995). There has been an extensive discussion of the regression methods used for developing SPFs in the EB (e.g. Wood et al., 2015a,b; Shin and Washington, 2012), among which standard negative binomial regression (NB) has been widely used (Shin and Washington, 2012). In this study a Poisson-Gamma (negative binomial) model is considered:

$$Y \sim \text{Poisson}(\mu\varepsilon)$$

$$\ln(\mu) = \alpha + \delta T + \beta X$$

where Y is the observed number of accidents, μ is the expected number of accidents, ϵ is a Gamma distributed random effect, (α, β) are the regression coefficients and X is the vector of covariates.

Using this model, the expected number of accidents in the before period, $\hat{M}_B$ can be obtained by

$$\hat{M}_B = \rho\hat{\mu}_B + (1 - \rho)X_B$$

where X_B is the observed number of accidents in the before period, $\hat{\mu}_B$ is the predicted number of accidents based on SPF before treatment, and the weight

$$\rho = \left(1 + \frac{\hat{\mu}_B}{\phi}\right)^{-1}$$

uses the shape parameter ϕ from the NB distribution.

To account for the trend in accidents between the before and after periods, the expected accidents in the after period is calibrated using a reference group. The estimate of accidents in the after period had treatment not occurred, $\hat{M}_A$, can be calculated after adjusting the time trend effect using

$$\hat{M}_A = \left(\frac{N_{A,\text{POP}}}{N_{B,\text{POP}}}\right)\hat{M}_B,$$

where $N_{A,POP}$ and $N_{B,POP}$ are the numbers of accidents for total population in the before and after periods.

The treatment effect under the EB approach can then be estimated as

$$\hat{\tau}_{\text{EB}} = \frac{X_A/\hat{M}_A}{1 + [\text{Var}(\hat{M}_A)/\hat{M}_A^2]},$$

where X_A is the observed number of accidents in the after period.

Control or reference groups are usually employed to calibrate the SPF to account for the trend in accidents, as well as the effects of changes in flow, between the before and after periods. Ideally control groups should have the same or similar traffic flow and road characteristics, i.e. the control group should be representative of the treated sites. However, The EB method is a quasi-empirical Bayesian method, which uses a weighted combination of observed and predicted accidents frequencies to estimate potential outcomes based on an outcome regression. It does not use an explicit mechanism for assessing common support between treated and control groups.

Another important characteristic of the EB approach relates to the specification of the SPF, which may strongly affect model development and consequently the results of before and after evaluation. In the EB model it is assumed that all the reference and treated units follow the same SPF, which is typically hard to defend in practice. Furthermore, it is not necessarily obvious which covariates should be included in the SPFs and some deemed relevant may not be observed. For these reasons, it is important to understand how the misspecification of the SPF, or the omission of covariates, can affect the estimates of the EB approaches.

2.3. Propensity score methods

Rosenbaum and Rubin (1983) introduced the use of propensity scores to systematically address the issue of similarity in treated and control group characteristics. The PS is a scalar value measuring the probability that a unit is selected into treatment conditional on observed covariates. For treatment evaluation, "similar" groups can be

defined as those with similar propensity score values, and by using this simple principle we can adjust for selection bias and ensure that the difference between the treated and control groups can be attributed solely to the treatment. PS methods have been widely studied and used in many evaluations of social, economic and medical programs (Heckman et al., 1998; Rudner and Peyton, 2006; Hirano and Imbens, 2001; Dehejia, 2005; Dehejia and Wahba, 2002; Kurth et al., 2006; Lechner, 2001; Abadie and Imbens, 2006, 2016).

In recent years, the PS method has also been applied in road safety analysis. For example, Wood et al. (2015a,b) compared EB, PS and regression methods with cross-sectional data. Their results indicated that all three methods yield consistent results, and the PS approach is a viable alternative to the EB method for road safety evaluation studies. Another study by Li et al. (2013) examined the impacts of speed limit enforcement cameras on road accidents. The PS method is compared with the naive before and after approach and the EB method. Although the PS and the EB methods show similar results, it is suggested that the PS is superior for selection of an appropriate control group.

2.3.1. Assumptions

Three crucial assumptions underlying the PS method are introduced by Rosenbaum and Rubin (1983)

1. Stable Unit Treatment Value Assumption (SUTVA) – the observed response under a given treatment allocation must be equivalent to the potential response under that treatment allocation.

   $$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

   for all $i = 1, \ldots, n$.
2. Conditional Independence Assumption (CIA) – the potential outcomes for unit i must be conditionally independent of the treatment assignment given the PS.

   $$Y_i(0), Y_i(1) \perp T_i | P(T_i | X_i)$$

   In other words, the treatment assignment can be considered as a random assignment conditional on $P(T_i|X_i)$. The CIA ensures that differences between treated and untreated units can be accounted for and the untreated units can be used to estimate a counterfactual outcome for the treatment group.
3. Common support condition (CSC) – units with the same X values have a positive probability of being both treated and untreated.

   $$0 < P(T_i = 1 | X_i = x) < 1, \quad \forall \ x$$

   This assumption ensures that The CSC is also known as the overlap condition, because there is sufficient overlap in the X of the treated and untreated units to find adequate matches.

2.3.2. Inverse probability weighting

Inverse probability weighting (IPW) uses the PS to effectively adjust for non-random assignment. Under a non-random assignment certain sub-populations may be over- or under-sampled. The idea underpinning the IPW estimator is that a pseudo population can be created in which the distributions of confounders among the treated and untreated are the same as the overall distribution of those in the original total population (Sturmer et al., 2006). This is achieved by using a function of estimated PS to effectively weight the sample observations according to their conditional probability of being treated.

For a binary treatment, the PS is estimated using an appropriate model. In this paper, a logit model as follows is used

$$P(T = 1|X) = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)}$$

where α is the intercept and β' is the vector of regression coefficients. The IPW estimator is then formed as the inverse of the conditional probability of the observed treatment status of units. The IPW is $1/P'$ for the treated and $1/(1 - P')$ for the untreated and the IPW estimator of the ATE is computed as

$$\tau_{\text{IPW}} = \frac{N^{-1}\Sigma_i^N \frac{T_i Y_i}{P_i'}}{N^{-1}\Sigma_i^N \frac{1 - T_i Y_i}{1 - P_i'}}$$

As with other PS based estimators, the IPW estimator can be biased if the model for calculating the PS is misspecified.

2.3.3. Doubly robust estimation

The DR estimator combines PS and outcome regression (OR) models to provide an additional level of robustness to model misspecification that can arise via violation of identifying assumptions. For the Poisson-Gamma model,

$$Y \sim \text{Poisson}(\mu\varepsilon)$$

we can estimate an OR

$$\ln(\mu) = \alpha + \delta T + \beta X$$

and form an estimator of the ATE as

$$\hat{\tau}_{\text{OR}} = \exp(\hat{\delta} + (\hat{\varepsilon}(1) - \hat{\varepsilon}(0))$$

This estimator will be unbiased if all potential confounders X are observed and correctly specified in the regression model. In other words the treatment assignment T is independent of the error term ε, $T \perp \varepsilon \,|\, X$. Proper specification of this model, however, can be difficult when multiple potential confounders exist.

The Doubly Robust (DR) estimator proposed by Robins et al. (1995) combine OR and IPW in a single model. The DR estimator can be written

$$\tau_{\text{DR}} = \frac{N^{-1}\Sigma_i^N \left[\frac{T_i Y_i}{P_i'} - \frac{(T_i - P_i')Y_{i,T=1}'}{P_i'} \right]}{N^{-1}\Sigma_i^N \left[\frac{(1 - T_i Y_i)}{1 - P_i'} - \frac{(T_i - P_i')Y_{i,T=0}'}{1 - P_i'} \right]}$$

where $Y_i' = E(Y|T, X)$ is the predicted value from the OR model given $T = 0, 1$ and the baseline covariates X. The two average terms are estimates of the mean potential outcomes, $Y_{X=1}$ and $Y_{X=0}$, if all units were treated or untreated. As a consequence, the difference in means is the effect due to the treatment. In the above equation, the first terms in each average are the IPW estimators for $E(Y_{X=1})$ and $E(Y_{X=0})$ respectively. The second terms are called "augmentations" (Funk et al., 2011) as this component is formed by taking the product of two bias terms: one from the PS model and one from the outcome regression model. If either bias term equals zero, then it excludes the other non-zero bias term from the incorrect model. That is the DR estimator will be consistent for the true ATE, if either the PS model or OR model is correctly specified. (for more details see Lunceford and Davidian, 2004; Graham et al., 2015).

3. Simulations

Previous studies have focused on the ability of the EB and the PS methods to address the issue of selection bias. However, they have not investigated how the assumptions required for consistent identification are met, and the performance of these methods when these assumptions are violated. The main challenge when making comparisons of different safety evaluation methods based on observational data is that the true ATE of the safety measures is not known. Recently a "no treatment" evaluation has been suggested as an effective way to assess the performance of Bayesian methods for before–after observational studies (Sacchi and Sayed, 2015; Wood and Donnell, 2017; Kuo and Lord, 2017). However, this method requires that no safety treatment or other operational changes were implemented in the time framework (Sacchi and Sayed, 2015), which usually cannot be fully guaranteed. Instead this paper compares the effectiveness of each method based on a simulation study.

Table 1
Summary of the models and simulation scenarios.

	Scenario 1 (N = 5000): stable weights + single SPF	Scenario 2 (N = 500): stable weights + single SPF	Scenario 3 (N = 5000): stable weights + two SPFs with minor differences	Scenario 4 (N = 5000): stable weights + two SPFs with major differences	Scenario 5 (N = 5000): unstable weights + single SPF
EB1	Correct OR and correct PS				
EB2	Minor misspecification with OR and correct PS				
EB3	Major misspecification with OR and correct PS				
IPW1	Correct OR and correct PS				
IPW2	Correct OR and minor misspecification with PS				
IPW3	Correct OR and major misspecification with PS				
DR1	Correct OR and correct PS				
DR2	Major misspecification with OR and correct PS				
DR3	Correct OR and major misspecification with PS				
DR4	Major misspecification with OR and major misspecification with PS				

Correct OR: $Y \sim \text{Poisson}(\exp(\alpha_0 + \delta T + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_2^2)*\epsilon)$
Correct PS: $P(T = 1|X) \sim \text{Logit}^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2)$
Minor misspecification with OR: $Y \sim Poisson(exp(\alpha_0 + \delta T + \alpha_1 X_1 + \alpha_2 X_2) * \epsilon)$
Major misspecification with OR: $Y \sim Poisson(exp(\alpha_0 + \delta T + \alpha_1 X_1) * \epsilon)$
Minor misspecification with PS: $P(T = 1 \mid X) \sim Logit^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$
Major misspecification with PS: $P(T = 1 \mid X) \sim Logit^{-1}(\beta_0 + \beta_1 X_1$

Table 1 summarizes the models and simulation scenarios investigated in this paper. While Scenario 1 is the basic scenario, alternative models and scenarios are set up to test the key assumptions/issues in the EB and PS methods including:

(1) Unmeasured confounding. Besides the correct models, models with minor and major misspecification are simulated for the EB, IPW and DR models separately.
(2) Sample size. The model performances with small sample size are tested in Scenario 2.
(3) Similarity of the control group. Two SPFs with minor/major differences are tested separately in Scenario 3 and 4, while a PS with unstable weights is tested in Scenario 5.

The following models are tested in the simulations:

1 τ_{EB1} – an empirical Bayes model based on a correctly specified SPF:

$$Y \sim \text{Poisson}(\exp(\alpha_0 + \delta T + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_2^2)*\epsilon)$$

The EB estimator is calculated as:

$$\hat{\tau}_{\text{EB}} = \frac{X_A/\hat{M}_A}{1 + [\text{Var}(\hat{M}_A)/\hat{M}_A^2]}$$

2 τ_{EB2} – same as [1.] except based on a "minor" misspecified SPF with X_2^2 excluded.
3 τ_{EB3} – same as [1.] except based on a "major" misspecified SPF with X_2, X_2^2 excluded.
4 τ_{IPW1} – an IPW model based on a correctly specified PS model:

$$P(T = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2)}$$

The IPW estimator is calculated as:

$$\tau_{\text{IPW}} = \frac{N^{-1}\Sigma_i^N \frac{T_i Y_i}{P_i'}}{N^{-1}\Sigma_i^N \frac{1 - T_i Y_i}{1 - P_i'}}$$

5 τ_{IPW2} – same as [4.] except based on a "minor" misspecified PS model with X_2^2 excluded.
6 τ_{IPW3} – same as [4.] except based on a "major" misspecified PS model with X_2, X_2^2 excluded.
7 τ_{DR1} – a DR model based on correctly specified PS and OR models. The DR estimator is

$$\tau_{\text{DR}} = \frac{N^{-1}\Sigma_i^N [\frac{T_i Y_i}{P_i'} - \frac{(T_i - P_i')Y_{i,T=1}'}{P_i'}]}{N^{-1}\Sigma_i^N [\frac{(1 - T_i Y_i)}{1 - P_i'} - \frac{(T_i - P_i')Y_{i,T=0}'}{1 - P_i'}]}$$

8 τ_{DR2} – same as [7.] except based on a "major" misspecified OR model (X_2, X_2^2 excluded) but with weights based on the correct PS model.
9 τ_{DR3} – same as [7.] except based on a correctly specified OR model but with weights based on a "major" misspecified PS model (X_2, X_2^2 excluded).
10 τ_{DR4} – same as [7.] except based on a "major" misspecified OR model (X_2, X_2^2 excluded) with weights based on a "major" misspecified PS model (X_2, X_2^2 excluded).

The simulations are conducted in five scenarios, two basic scenarios and three alternative scenarios (sensitivity tests).

Scenario 1. The data generating process (DGP_1) for a sample of 5000 is

$$X_{1,\text{pre}} \sim \text{Normal}(0, 1)$$

$$X_{1,\text{post}} \sim \text{Uniform}(0, 1) + X_{1,\text{pre}}$$

$$X_2 \sim \text{Normal}(1, 1)$$

A binary treatment T is assigned as a function of covariates X_1 and X_2.

$$T \sim \text{Bernoulli}(\text{expit}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2))$$

The SPF_1 can be described as:

$$Y_{\text{pre}} \sim \text{Poisson}(\exp(\alpha_0 + \alpha_1 X_{1,\text{pre}} + \alpha_2 X_2 + \alpha_3 X_2^2)*\epsilon)$$

$$Y_{\text{post}} \sim \text{Poisson}(\exp(\alpha_0 + \delta T + \alpha_1 X_{1,\text{post}} + \alpha_2 X_2 + \alpha_3 X_2^2)*\epsilon)$$

$$\epsilon \sim \text{Gamma}(2, 0.5)$$

where ϵ is a Gamma distributed random error term, $\beta_0 = -1$, $\beta_1 = 0.1$, $\beta_2 = 0.1$, $\beta_3 = 0.1$, $\alpha_0 = 1$, $\alpha_1 = 0.1$, $\alpha_2 = 0.1$, $\alpha_3 = 0.01$ and $\delta = 1$. The true value of treatment effect is $\tau = exp(\delta) = 2.718$.

*Scenario*2. Same as scenario 1 except the sample size is 500.

*Scenario*3. Same as scenario 1 except that the outcome Y is generated based on a different SPF (SPF_2) for a sample of 2500:

$$Y_{\text{pre}} \sim \text{Poisson}(\exp(\alpha_0' + \alpha_1' X_{1,\text{pre}} + \alpha_2' X_2 + \alpha_3' X_2^2)*\epsilon)$$

$$Y_{\text{post}} \sim \text{Poisson}(\exp(\alpha_0' + \delta T + \alpha_1' X_{1,\text{post}} + \alpha_2' X_2 + \alpha_3' X_2^2)*\epsilon)$$

$$\epsilon \sim \text{Gamma}(2, 0.5)$$

where $\alpha_0' = 1$, $\alpha_1' = 0.5$, $\alpha_2' = 0.1$, $\alpha_3' = 0.01$ and $\theta = 1$.

*Scenario*4. Same as scenario 1 except that the outcome Y is generated based on a different SPF (SPF_3) for a sample of 2500:

$$Y_{\text{pre}} \sim \text{Poisson}(\exp(\alpha_0'' + \alpha_1'' X_{1,\text{pre}} + \alpha_2'' X_2 + \alpha_3'' X_2^2)*\epsilon)$$

$$Y_{\text{post}} \sim \text{Poisson}(\exp(\alpha_0'' + \delta T + \alpha_1'' X_{1,\text{post}} + \alpha_2'' X_2 + \alpha_3'' X_2^2)*\epsilon)$$

$$\epsilon \sim \text{Gamma}(2, 0.5)$$

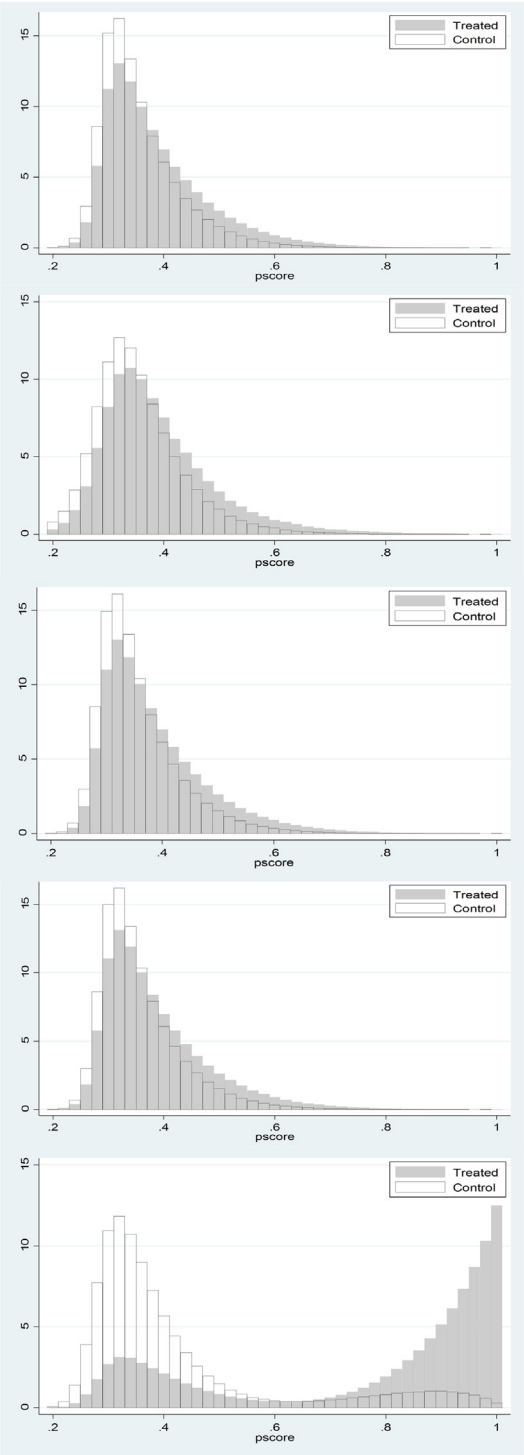
where $\alpha_0'' = 0.1$, $\alpha_1'' = 1$, $\alpha_2'' = 0.01$, $\alpha_3'' = 0.01$ and $\theta = 1$.

*Scenario*5. Same as scenario 1 except that covariates X_1, X_1 are generated based on a different DGP (DGP_2) for 2500 of the data sample:

$$X_{1,\text{pre}} \sim \text{Normal}(5, 1)$$

$$X_{1,\text{post}} \sim \text{Uniform}(0, 1) + X_{1,\text{pre}}$$

$$X_2 \sim \text{Normal}(5, 1)$$



	Standardized difference		Variance ratio	
Scenario 1	Raw	Weighted	Raw	Weighted
x1_pre	0.098	0.000	1.001	1.001
x2	0.301	0.000	1.202	1.000
x3	0.322	0.000	1.875	1.002
	Standardized difference		Variance ratio	
Scenario 2	Raw	Weighted	Raw	Weighted
x1_pre	0.100	0.000	1.001	0.999
x2	0.301	0.000	1.220	1.001
x3	0.324	0.001	1.935	1.042
	Standardized difference		Variance ratio	
Scenario 3	Raw	Weighted	Raw	Weighted
x1_pre	0.098	0.000	1.003	1.003
x2	0.301	0.000	1.201	1.000
x3	0.322	0.000	1.871	1.002
	Standardized difference		Variance ratio	
Scenario 4	Raw	Weighted	Raw	Weighted
x1_pre	0.097	0.000	0.998	0.998
x2	0.303	0.000	1.201	1.000
x3	0.323	0.000	1.884	1.008
	Standardized difference		Variance ratio	
Scenario 5	Raw	Weighted	Raw	Weighted
x1_pre	1.276	0.002	1.474	1.006
x2	1.493	0.003	1.866	1.013
x3	1.459	0.008	4.367	1.093

Fig. 1. Tests of overlap and covariates balance.

The models are simulated for 1000 iterations in each scenario. Mean values, relative bias in percentage, variances of the estimates of the treatment effect $\hat{\tau}$ and the mean squared error (MSE) are reported.

4. Results

4.1. Tests for PS methods

We first check the validity of the PS based methods. We perform balancing tests to verify that treatment is independent of the covariates after weighting. The PS methods aim to balance characteristics between the treated and control groups, i.e., there should be no significant differences between covariate means. Fig. 1 shows diagnostic statistics for the raw and the weighted data based on 1000 iterations. For *scenario* 1, 3 and 4 the standardized differences are all close to zero and the variance ratios are all close to one, indicating that weighted on the estimated propensity score the covariates are all well balanced. For a small sample of 500 in *scenario* 2, although the results also indicate balanced covariates, the variance ratios for X_3 is not sufficiently close to 1. For *scenario* 5, covariates are generated based on two different DGPs to

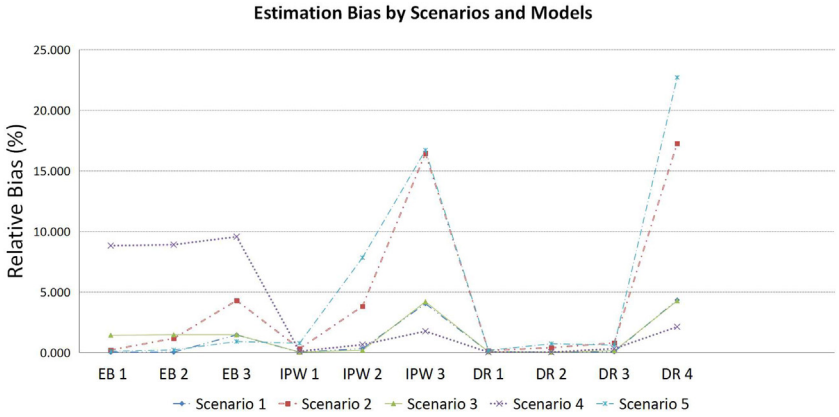


Fig. 2. Estimation bias of treatment effects in different scenarios by EB, IPW and DR models.

increase the imbalance in the characteristics of the original data. As expected, the results show that the scale of imbalance in the raw data is larger than those in previous scenarios. For X_3, the weighted standardized difference is close to zero, but the weighted variance ratio still appears to be considerably larger than one, indicating a lack of balance in X_3.

We further conduct a visual inspection of the propensity score distributions for both the treatment and control groups. From the histograms of propensity scores for both groups, the extent to which there is overlap in the scores between the treatment and control groups is apparent. Fig. 1 shows the densities of propensity scores for both groups in different scenarios. For *scenario* 1, 2, 3 and 4, the propensity scores for both treated and untreated groups appear to have sufficient overlap, indicating that the overlap condition is not violated. For *scenario* 5, although the two estimated densities overlap each other to a certain extent, a possible problem is that the support of the density gets very close to 0 for most treated units and 1 for the untreated.

4.2. Simulation results

The simulation results are summarized in Fig. 2. We first investigate the robustness to unmeasured confounding for different models. For misspecified models, the estimates are almost unaffected when only the quadratic term X_2^2 is excluded. However, both the IPW and EB models fails to provide precise estimation of the true treatment effect when confounder X_2 is excluded from the models, while the DR models consistently provide less unbiased estimates, indicating that the DR models can offer additional robustness to model misspecification.

We then compare the model performances across different sample sizes (*scenario* 1 and 2) as shown in Table 2. For *scenario* 1, the correctly specified EB, IPW and DR models all show similar results regarding the precision of effect estimates (relative bias (%) = 0.037, 0.037 and 0.074 respectively). For models with misspecification, the EB and IPW models with the quadratic term excluded results in similar effect estimates with small degree of bias. However, both models fail to provide precise approximation to the true value of τ when the confounder X_2 is further excluded. Although the DR models consistently produce valid estimates of the treatment effect when the OR and/or IPW models are wrongly specified, the estimate is biased when both models are incorrect. For *scenario* 2 with a small sample of 500, the performance of the EB and DR models is similar to the ones in *scenario* 1 except that the bias and MSE are slightly increased. However, the bias of the effect estimates increases dramatically for misspecified IPW models, suggesting that the PS methods are “data hungry”.

Table 2
Results for basic scenarios with sample size of 5000 and 500 (τ = 2.718).

Model	Scenario 1				Scenario 2			
	Av. Est.	Relative bias (%)	Var.	MSE	Av. Est.	Relative bias (%)	Var.	MSE
EB1	2.719	0.037	0.002	0.002	2.724	0.221	0.023	0.023
EB2	2.719	0.037	0.002	0.002	2.750	1.177	0.015	0.016
EB3	2.758	1.472	0.045	0.047	2.835	4.305	0.022	0.035
IPW1	2.719	0.037	0.004	0.004	2.727	0.331	0.052	0.053
IPW2	2.727	0.331	0.004	0.004	2.822	3.826	0.064	0.075
IPW3	2.828	4.047	0.005	0.017	3.165	16.446	0.093	0.293
DR1	2.720	0.074	0.004	0.004	2.723	0.184	0.056	0.056
DR2	2.717	0.037	0.005	0.005	2.729	0.405	0.049	0.049
DR3	2.716	0.074	0.004	0.004	2.740	0.809	0.053	0.053
DR4	2.836	4.341	0.005	0.019	3.187	17.255	0.099	0.319

In the EB approach, the expected accident number of a site is usually estimated based on observed and prior information. A reference group of similar characteristics to the treated group is used to calibrate a SPF to obtain the prior information. However, in practice it is not always feasible to assume that the accident frequency of the reference sites follow the same SPF. To further investigate this issue, in *scenario* 3 and 4, the outcome Y is generated based on different SPFs (SPF_2 and SPF_3) for half of the sample. As shown in Table 3, for *scenario* 3 minor differences exist between SPF_1 and SPF_2, resulting in slight increases in the estimation bias, while higher bias is reported when half of the sample is generated based on SPF_3. These findings demonstrate that the performance of the EB methods relies heavily on the assumption that all the reference and treated units follow the same SPF, and that estimation can be seriously biased when this assumption is violated. In terms of the IPW and DR models, similar patterns are observed as in *scenario* 1, although the estimation variance is increased slightly.

In terms of the PS based models, in *scenario* 5 covariates are generated based on a different DGP (DGP_2) for half of the data sample to increase the imbalance in the characteristics of the original data. The simulation results show that under unstable weights the IPW models provide relatively high bias even with a correctly specified PS model. The bias and MSE are even higher for the IPW models with unmeasured covariates. The EB approaches based on a correctly specified OR model provide a good approximation to the true value τ. Compared to the IPW models, the DR model consistently provide good estimation with small bias when either the PS model or the OR is correctly specified. However, in most scenarios except *scenario* 4, the DR estimates are the least accurate when both the PS and OR models are wrongly specified.

5. Discussion and conclusions

The EB approach has been widely used in before-and-after road safety studies due to its ability to address regression to the mean bias. In recent years, causal models based on PS methods have been proposed as an alternative to the conventional EB method. Both methods are based

Table 3
Results of sensitivity tests among various models (τ = 2.718).

Model	Scenario 3				Scenario 4				Scenario 5			
	Av. est.	Relative bias (%)	Var.	MSE	Av. est.	Relative bias (%)	Var.	MSE	Av. est.	Relative bias (%)	Var.	MSE
EB1	2.757	1.435	0.002	0.004	2.958	8.830	0.004	0.062	2.721	0.110	0.001	0.001
EB2	2.758	1.472	0.002	0.004	2.960	8.904	0.004	0.063	2.724	0.221	0.002	0.002
EB3	2.758	1.472	0.002	0.004	2.978	9.566	0.004	0.072	2.743	0.920	0.001	0.002
IPW1	2.717	0.037	0.007	0.007	2.715	0.110	0.019	0.019	2.739	0.773	0.041	0.042
IPW2	2.723	0.184	0.007	0.007	2.736	0.662	0.021	0.022	2.931	7.837	0.013	0.059
IPW3	2.832	4.194	0.008	0.021	2.766	1.766	0.021	0.024	3.172	16.703	0.015	0.221
DR1	2.716	0.074	0.007	0.007	2.717	0.037	0.018	0.018	2.713	0.184	0.044	0.044
DR2	2.719	0.037	0.007	0.007	2.719	0.037	0.016	0.016	2.738	0.736	0.044	0.044
DR3	2.722	0.147	0.007	0.007	2.727	0.331	0.017	0.017	2.735	0.625	0.027	0.027
DR4	2.834	4.268	0.007	0.021	2.776	2.134	0.018	0.022	3.335	22.701	0.025	0.406

on a series of key identifying assumptions. Since in practice these assumptions are often violated, this study investigates and compares the performance of the evaluation methods via simulation in settings with different model specifications and data conditions.

We investigated three critical assumptions concerning: (i) confounding, (ii) sample size, and (iii) similarity between treated and control groups. The EB and IPW approaches fail to provide consistent estimation in the presence of unmeasured confounding, but their performances are less affected by minor model misspecifications (e.g. quadratic term excluded). The DR estimator allows for violation of the unconfoundness assumption in either the OR or PS model, but not both. We further compared the model performances with data sample sizes of 5000 and 500 in *scenarios* 1 and 2. Both the estimation bias and variance of the PS based methods (except for the full DR) are increased with a small sample size (*scenario* 2), indicating that the propensity score methods perform better in large samples.

In terms of the similarity of the control group, it is usually assumed in the EB approach that the accident records of the reference sites follow a single SPF to those of the treated sites. In practice, this assumption can be easily violated. In *scenarios* 3 and 4, two different SPFs are specified for the outcomes. While the IPW and DR methods are less affected, due to their ability to successfully match treated and control units, the bias is dramatically increased for EB estimates especially when there are major differences between the two SPFs. In terms of the IPW and DR approaches, two DGPs are also employed in *scenario* 5 to increase the imbalance of the data, resulting in unstable weights. While the performance of the full model is less affected, the precision of the estimates is greatly reduced for the IPW models with incorrect specification. In contrast, the DR models consistently provide precise estimation.

In summary, this study contributes to the literature by investigating and comparing the performance of the EB and the PS approaches in settings with various model specifications and data conditions via simulation studies. The results suggest that the DR methods are superior to the EB and IPW in most cases. Specifically, the DR methods can provide consistent and unbiased estimates when either the OR or IPW model is correctly specified. The IPW models can provide unbiased estimates only with large data sample and sufficient overlap in covariate distributions between treated and control units. The EB approach performs better than the IPW and DR models with data sets of small sample size (e.g. 500). However, the performance of the EB models heavily relies on correctly specified SPF functions. In addition, the EB approaches require that all the reference sites follow the same SPF. Our results show that the precision of the estimates can be seriously affected if this condition is not satisfied.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 71701042), and the Key Project of National Natural Science Foundation of China (Grant No. 51638004), and the National Natural Science Foundation of China (Grant No. 51578149).

References

Abadie, A., Imbens, G., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74, 235–267.
Abadie, A., Imbens, G., 2016. Matching on the estimated propensity score. Econometrica 84 (2), 781–807.
Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38, 618–625.
Dehejia, R.H., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. Rev. Econ. Stat. 84 (1), 151–161.
Dehejia, R., 2005. Practical propensity score matching: a reply to Smith and Todd. J. Econom. 125, 355–364.
El-Basyouny, K., Sayed, T., 2011. A full Bayes multivariate intervention model with random parameters among matched pairs for before–after safety evaluation. Accid. Anal. Prev. 43, 87–94.
Elvik, R., et al., 2017. An empirical Bayes before–after evaluation of road safety effects of a new motorway in Norway. Accid. Anal. Prev. 108, 285–296.
Funk, M., Westreich, D., Wiesen, C., Sturmer, T., Brookhart, M., Davidian, M., 2011. Doubly robust estimation of causal effects. Am. J. Epidemiol. 173 (7), 761–767.
Graham, D.J., McCoy, E.J., Stephens, D.A., 2015. Approximate Bayesian inference for doubly robust estimation. Bayesian Anal. 11, 47–69.
Gu, X., Abdel-Aty, M., et al., 2019. Utilizing UAV video data for in-depth analysis of drivers' crash risk at interchange merging areas. Accid. Anal. Prev. 123, 159–169.
Hauer, E., 1992. Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. Accid. Anal. Prev. 24 (5), 457–477.
Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., 2002. The empirical Bayes method for estimating safety: a tutorial. Transp. Res. Rec. 1784, 126–131.
Hauer, E., 1997. Observational Before–After Studies in Road Safety. Pergamon Publication, England.
Hauer, E., 1995. On exposure and accident rate. Traffic Eng. Control 36 (3), 134–138.
Heckman, J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Rev. Econ. Stud. 65, 261–294.
Hirano, K., Imbens, G.W., 2001. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Serv. Outcomes Res. Methodol. 2, 259–278.
Holland, P., 1986. Statistics and causal inference (with discussion). J. Am. Stat. Assoc. 81, 945–970.
Karwa, V., Slavkovic, A.B., Donnell, E.T., 2011. Causal inference in transportation safety studies: comparison of potential outcomes and causal diagrams. Ann. Appl. Stat. 5, 1428–1455.
Kuo, P., Lord, D., 2017. Estimating the safety impacts in before–after studies using the Naïve Adjustment Method. Transportmetrica A: Transp. Sci. 13, 915–931.
Kurth, T., Walker, A.M., et al., 2006. Results of multivariable logistic regression, propensity matching propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am. J. Epidemiol. 163, 262–270.
Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Econometric Evaluations of Active Labor Market Policies in Europe. Heidelberg, Physica.
Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accid. Anal. Prev. 78, 146–154.
Li, H., Graham, D.J., Majumdar, A., 2013. The impacts of speed cameras on road accidents: an application of propensity score matching methods. Accid. Anal. Prev. 60, 148–157.
Li, H., Graham, D.J., 2016. Quantifying the causal effects of 20 mph zones on road casualties in London via doubly robust estimation. Accid. Anal. Prev. 93, 65–74.
Li, W., Carriquiry, A., Pawlovich, M., Welch, T., 2008. The choice of statistical models in road safety countermeasure effectiveness studies in Iowa. Accid. Anal. Prev. 40,

1531–1542.
Lord, D., Kuo, P., 2012. Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. Accid. Anal. Prev. 47, 52–63.
Lunceford, J.K., Davidian, M., 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat. Med. 23, 2937–2960.
Miaou, S.P., Lord, D., 2003. Modelling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. Transp. Res. Rec. 1840, 31–40.
Park, E.S., Lord, D., 2007. Multivariate Poisson–lognormal models for jointly modelling crash frequency by severity. Transp. Res. Rec. 2019, 1–6.
Persaud, B., Lyon, C., 2007. Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. Accid. Anal. Prev. 39, 546–555.
Persaud, B., Lan, B., Lyon, C., Bhim, R., 2010. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. Accid. Anal. Prev. 42, 38–43.
Persaud, B., Retting, R., Lyon, C., 2004. Crash reductions following installation of centre-line rumble strips. Accid. Anal. Prev. 36, 1073–1079.
Persaud, B.N., Hauer, E., Retting, R., Vallurapalli, R., Mucsi, K., 1997. Crash reductions following traffic signal removal in Philadelphia. Accid. Anal. Prev. 29, 803–810.
Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accid. Anal. Prev. 40, 1486–1497.
Robins, J.M., Rotnitzky, A., Zhao, L.P., 1995. Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data. J. Am. Stat. Assoc. 90, 106–121.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rudner, L.M., Peyton, J., 2006. Consider propensity scores to compare treatments. Pract. Assess. Res. Eval. 11 (9), 2.
Sacchi, E., Sayed, T., 2015. Investigating the accuracy of Bayesian techniques for before–after safety studies: the case of a “no treatment” evaluation. Accid. Anal. Prev. 78, 138–145.
Sasidharan, L., Donnell, E., 2013. Application of propensity scores and potential outcomes to estimate effectiveness of traffic safety countermeasures: exploratory analysis using intersection lighting data. Accid. Anal. Prev. 50, 539–553.
Shin, K., Washington, S., 2012. Empirical Bayes method in the study of traffic safety via heterogenous negative multinomial model. Transpormetrica 8, 131–147.
Sturmer, T., Rothman, K.J., Glynn, R.J., 2006. Insights into different results from different causal contrasts in the presence of effect measure modification. Pharmacoepidemiol. Drug Saf. 15 (10), 698–709.
Tarko, A., Eranky, S., Sinha, K., 1998. Methodological considerations in the development and use of crash reduction factors. In: Preprint CD. 77th Annual Meeting of the Transportation Research Board. Washington, DC.
Wang, X., et al., 2018. Investigating the safety impact of roadway network features of suburban arterials in Shanghai. Accid. Anal. Prev. 113, 137–148.
Wood, J.S., Donnell, E.T., Porter, R.J., 2015a. Comparison of safety effect estimates obtained from empirical Bayes before–after study, propensity scores-potential outcomes framework, and regression model with cross-sectional data. Accid. Anal. Prev. 75, 144–154.
Wood, J., Donnell, E., 2017. Causal inference framework for generalizable safety effect estimates. Accid. Anal. Prev. 104, 74–87.
Wood, J., Gooch, J., Donnell, E., 2015b. Estimating the safety effects of lane widths on urban streets in Nebraska using the propensity scores-potential outcomes framework. Accid. Anal. Prev. 82, 180–191.
Xu, C., et al., 2019. Investigating the factors affecting secondary crash frequency caused by one primary crash using zero-inflated ordered probit regression. Physica A: Stat. Mech. Appl. 524, 121–129.