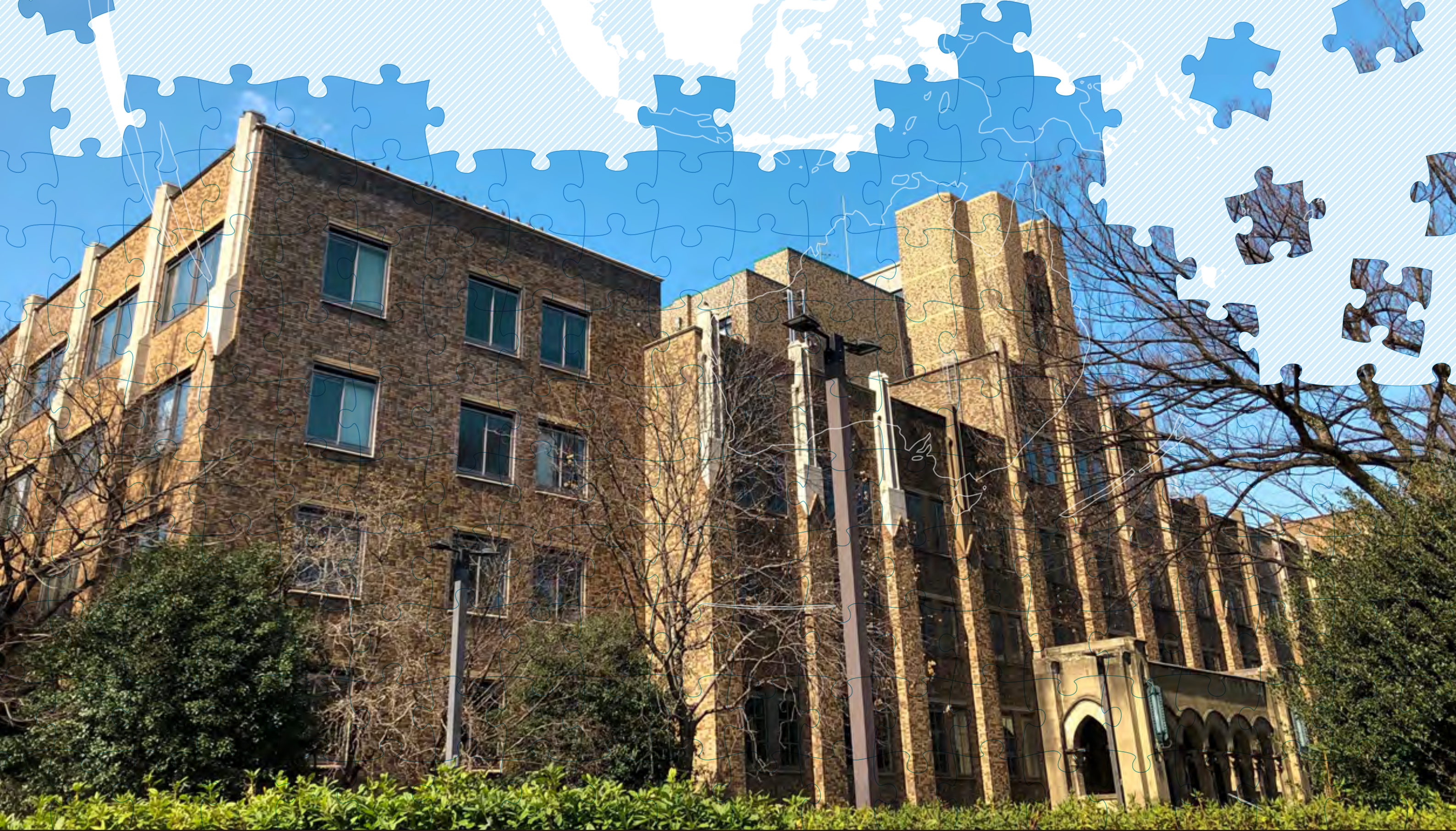
CSRDA
Center for Social Research and Data Archives,
Institute of Social Science, The University of Tokyo

SUSTAINABLE DEVELOPMENT GOALS

CSRDA supports the Sustainable Development Goals

CSRDA Discussion Paper

Double/debiased Machine Learning for Causal Inference on Survival Function



No.	Date	SDGs
84	May. 2024	8 DECENT WORK AND ECONOMIC GROWTH

Name
Daijiro Kabata, Mototsugu Shintani

Double/debiased Machine Learning for Causal Inference on Survival Function

Daijiro Kabata[1] and Mototsugu Shintani*[2]

[1]*Department of Medical Statistics, Osaka Metropolitan University*

[2]*Faculty of Economics, The University of Tokyo*

This version: May 2024

Abstract

This paper discusses the use of double/debiased machine learning (DML) for estimating the average treatment effect (ATE) on a survival function using pseudo-observations. Through simulations, we demonstrate the double robustness property of our method and its improved performance, compared to existing estimators in the presence of many covariates. In our empirical example, the method is applied in evaluating the effect of the e-learning program participation on the job-finding rate among individuals who are seeking employment.

Keywords: doubly robust estimator, survival analysis, pseudo-observations.

JEL Classification: C410, J640.

*Correspondence: Mototsugu Shintani, Faculty of Economics, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8654, JAPAN (e-mail: shintani@e.u-tokyo.ac.jp).

1 Introduction

In the field of causal inference, doubly robust (DR) estimators have been widely used as a workhorse method because of their robustness against potential misspecifications in either propensity scores or outcome equations. Within the class of DR estimators, Wang (2018) has proposed an estimator for the survival function using pseudo-observations. In this paper, we utilize double/debiased machine learning (DML), initially developed by Chernozhukov et al. (2018), to estimate a survival function using pseudo-observations. Through simulations, we investigate its performance compared to existing estimators, such as the inverse probability weighted (IPW) estimator and the DR estimator, particularly in the presence of many covariates. We also apply our method in estimating the effect of the e-learning program participation on reducing unemployment duration.

2 Estimators of the Average Treatment Effect on Survival Probability using Pseudo-observations

Let T be the survival time to the first event, C be the censoring time, $X = (X_1,\ X_2, \ldots,\ X_p)$ be a p-dimensional vector of covariates with distribution $F(X)$, and $D \in \{0, 1\}$ be a binary treatment variable, where $D = 1$ signifies the treatment group. The conditional survival probability with covariate X under $D = d$ at time

t is given by $S_d(t|X) = E[I\{T > t\}|D = d, X]$, and the unconditional survival function is given by $S_d(t) = \int S_d(t|X)dF(X)$. Our target is the average treatment effect (ATE) on survival probability defined as

$$\theta(t) = S_1(t) - S_0(t).$$

The survival function $S_d(t)$ can be calculated by the Kaplan-Meier estimator $\hat{S}_d(t)$ using observation $(T_i,\ C_i,\ X_i)$ for $i = 1, \ldots,\ N$. However, in general, a propensity score $m(X) = \Pr[D = 1|X]$ is a function of X, and the dependence of X and D implies that $\hat{\theta}(t) = \hat{S}_1(t) - \hat{S}_0(t)$ is a biased estimator of the ATE.

To reduce bias from confounding, one can utilize a standard causal inference procedure applied to outcome $S_d(t|X_i)$, namely, the individual survival function of each individual i. While individual outcome $S_d(t|X_i)$ is not directly observed, Andersen, Klein, and Rosthøj (2003) and Klein et al. (2007) proposed using a pseudo-observation of individual i defined by

$$\hat{S}_d^i(t) = N\hat{S}_d(t) - (N-1)\hat{S}_d^{-i}(t)$$

where $\hat{S}_d(t)$ is the Kaplan-Meier estimator using all observations $\{(T_i,\ C_i,\ X_i)\}_{i=1}^N$, and $\hat{S}_d^{-i}(t)$ is the leave-one-out estimator using $\{(T_j,\ C_j,\ X_j)\}_{j=1, j\neq i}^N$. As shown by Graw, Gerds, and Schumacher (2009), $E[\hat{S}_d^i(t)|X_i] \to S_d(t|X_i)$ as $N \to \infty$. Relying on this asymptotic property of the pseudo-observations, Andersen, Syriopoulou, and Parner (2017) propose the IPW estimator of the ATE on survival probability given by

$$\hat{\theta}_{IPW}(t) = \frac{\sum_{i=1}^{N} D_i \hat{S}_1^i(t)/\hat{m}(X_i)}{\sum_{i=1}^{N} D_i/\hat{m}(X_i)} - \frac{\sum_{i=1}^{N} (1-D_i)\hat{S}_0^i(t)/(1-\hat{m}(X_i))}{\sum_{i=1}^{N} (1-D_i)/(1-\hat{m}(X_i))}$$

where $\hat{m}(X)$ is an estimator of $m(X)$. Furthermore, Wang (2018) also uses pseudo-observations and considers the DR estimator of the ATE on survival probability given by

$$\begin{aligned}\hat{\theta}_{DR}(t) = &\frac{1}{N}\sum_{i=1}^{N}\{\hat{S}_1(t|X_i) - \hat{S}_0(t|X_i)\} \\ &+ \frac{\sum_{i=1}^{N} D_i(\hat{S}_1^i(t) - \hat{S}_1(t|X_i))/\hat{m}(X_i)}{\sum_{i=1}^{N} D_i/\hat{m}(X_i)} \\ &- \frac{\sum_{i=1}^{N}(1-D_i)(\hat{S}_0^i(t) - \hat{S}_0(t|X_i))/(1-\hat{m}(X_i))}{\sum_{i=1}^{N} (1-D_i)/(1-\hat{m}(X_i))}.\end{aligned}$$

where $\hat{S}_d(t|X)$ is an estimator of outcome equation $S_d(t|X)$ as a function of X. For example, we can employ the Cox regression model or a generalized estimating equation for $\hat{S}_d(t|X)$. The IPW estimator is asymptotically unbiased when $\hat{m}(X)$ is correctly specified. Furthermore, the DR estimator is asymptotically unbiased when either $\hat{m}(X)$ or $\hat{S}_d(t|X)$ is correctly specified (double robustness).

In general, the overfitting in the nuisance function estimation can lead to bias in the estimation of the ATE. Chernozhukov et al. (2018) have introduced the

DML method based on cross-fitting to address this overfitting issue. Here, we also apply the DML to pseudo-observations in the estimation of ATE on survival probability.

For simplification, we assume that N is a multiple number of integer K. Consider a K-fold random partition $(I_k)_{k=1}^{K}$ of $\{1, \ldots, N\}$ such that the size of each fold I_k is fixed at $n = N/K$. For each subsample I_k, define its complement as $I_k^c = \{1, \ldots, N\} \setminus I_k$. In the first step, estimate the ATE using each subsample I_k $(k = 1, \ldots, K)$ by

$$
\begin{aligned}
\hat{\psi}_{DML}(t; I_k, I_k^c) = &\frac{1}{n}\sum_{i \in I_k}\{\hat{S}_1(t|X_i; I_k^c) - \hat{S}_0(t|X_i; I_k^c)\} \\
&+ \frac{\sum_{i \in I_k} D_i(\hat{S}_1^i(t) - \hat{S}_1(t|X_i; I_k^c))/\hat{m}_(X_i; I_k^c)}{\sum_{i \in I_k} D_i/\hat{m}(X_i; I_k^c)} \\
&- \frac{\sum_{i \in I_k}(1 - D_i)(\hat{S}_0^i(t) - \hat{S}_0(t|X_i; I_k^c))/(1 - \hat{m}(X_i; I_k^c))}{\sum_{i \in I_k}(1 - D_i)/(1 - \hat{m}(X_i; I_k^c))}
\end{aligned}
$$

where $\hat{S}_d(t|X_i; I_k^c)$ and $\hat{m}(X_i; I_k^c)$ are the estimators of $S_d(t|X_i; I_k^c)$ and $m(X_i; I_k^c)$, respectively. In the second step, aggregate $\hat{\psi}_{DML}(t; I_k, I_k^c)$ for all $k \in \{1, \ldots, K\}$, and the final ATE estimator based on DML is given by

$$
\hat{\theta}_{DML}(t) = \frac{1}{K}\sum_{k=1}^{K}\hat{\psi}_{DML}(t; I_k, I_k^c).
$$

Chernozhukov et al. (2018) suggest iteratively performing cross-fitting and utilizing the mean or median value to enhance the ATE estimator's stability against data-splitting randomness.

3 Simulation Experiments

We conduct two simulation experiments to assess the performance of the proposed estimator. The first experiment (DGP1) evaluates the effect of misspecification on the IPW, DR, and DML estimators. The second experiment (DGP2) compares the sensitivity to overfitting between the DR estimator and the DML estimator.

3.1 DGP1

In the first simulation experiment, we fix the sample size at $N = 200$ and the number of covariates at $p = 8$. First, we generate the covariates $X_i = (X_{1i}, X_{2i}, \dots, X_{8i})$ for $i = 1, \dots, 200$ from multivariate standard normal distribution with unit variance and covariance where only pairs (X_1, X_2), (X_3, X_4), (X_5, X_6), and (X_7, X_8) are correlated with a correlation coefficient of 0.2. Then, the binary treatment variable D_i is generated by a Bernoulli distribution with the true propensity score given by

$$p_i = \left\{ 1 + \exp\left(-\alpha_0 - \sum_{j=1}^{p} \alpha_j X_{ji} \right) \right\}^{-1}$$

where $(\alpha_1, .., \alpha_8) = (1.0,\ 1.0,\ 0.5,\ 0.5,\ 0.0,\ 0.0,\ 0.0,\ 0.0)$. To fix the treatment prevalence at around 50 percent, the intercept α_0 is set at -0.7. The continuous time variable T_i is generated from the exponential distribution with an event rate

$$h_i = \exp\left(\beta_0 + \gamma D_i + \sum_{j=1}^{p} \beta_j X_{ji}\right)$$

where $(\beta_1, .., \beta_8) = (1.0,\ 1.0,\ 0.0,\ 0.0,\ 0.5,\ 0.5,\ 0.0,\ 0.0)$ and $\gamma = 0$. The intercept β_0 is set at -0.7 so that the event rate is fixed at around 50 percent. In the above setup, X_1 and X_2 can be considered confounders that affect both the treatment selection and the outcome. On the other hand, X_3 and X_4 are covariates only affecting the treatment, while X_5 and X_6 are covariates only affecting the outcome. Furthermore, X_7 and X_8 do not relate to the treatment and outcome.

For all estimators, we estimated the propensity score using the lasso. For DML, the conditional average survival function is estimated using the regularized Cox model. In the cross-fitting part of DML, we fixed $K = 5$. To incorporate the uncertainty induced by sample splitting, we iterate the estimating procedure 5 times and aggregate these estimates as the mean value. We fix the time point at $t = 3$ and compute $\hat{\theta}_{IPW}(3)$, $\hat{\theta}_{DR}(3)$, and $\hat{\theta}_{DML}(3)$ to estimate the true ATE $\theta(3) = 0$.

To assess the robustness against the misspecification of the nuisance functions, we consider four cases: (1) both the propensity score and the survival function are correctly specified; (2) the propensity score is correctly specified but the survival function is misspecified; (3) the propensity score is misspecified but the survival function is correctly specified; and (4) both models are misspecified. We provide the misspecified models by excluding confounders (X_1 and X_2) from each function.

The results of experiments are provided in Figure 1, which shows the empirical distribution from $1,000$ replications, and in Table 1, which presents the absolute bias, standard deviation (SD) and root mean square error (RMSE). In case 1, where nuisance functions for treatment and survival are correctly specified, all estimators perform relatively well. However, bias, SD, and RMSE of the IPW estimator are slightly larger than those of the DR and DML estimators. In case 2, with a correctly specified propensity score model, all estimators perform similarly. In case 3, with a misspecified propensity score model, the bias of the IPW estimator becomes much larger than that of two other estimators. In case 4, with both nuisance functions misspecified, all estimators lead to large biases. These results confirm the doubly robust properties of the DR and DML estimators. Both doubly robust estimators perform equally well when the number of covariates is relatively small.

3.2 DGP2

We now consider the effect of a relatively large number of covariates on the performance of two doubly robust estimators, the DR and DML estimators. The simulation setting is similar to case 1 of DGP1, except for the sample sizes and the number of confounders. In particular, the number of confounders have increased from 2 to 94 with the correlation coefficients of all confounders at 0.2. The corresponding parameters are $(\alpha_1, .., \alpha_{94}) = (1.0, ..., 1.0)$ for the propensity score and $(\beta_1, .., \beta_{94}) = (1.0, ..., 1.0)$ for the survival function. The other 6 covariates are generated in the same way as in DGP1 with the same set of parameters. With the number of covariates fixed at $p = 100$, we consider sample sizes N of 1000, 500, 300, 250, and 200 so that the corresponding ratio of covariate parameters to the number of subjects, namely p/N, is 0.1, 0.2, 0.3, 0.4, and 0.5. The intercepts α_0 and β_0 are set at -20 to fix the prevalence proportion of the treatment and the event at around 50 percent.

The results of experiments are provided in Figure 2, which shows how the RMSEs of $\hat{\theta}_{DR}(3)$ and $\hat{\theta}_{DML}(3)$ respond to p/N. As the p/N ratio increases, the RMSEs of both estimators increase. However, the RMSE of the DML estimator increases much more slowly than the DR estimator. This difference is mainly due to the smaller bias of the DML estimator as the SD of the two estimators are almost the same (see Supplementary Table 1 for the details). This result suggests the DML suffers less from the bias caused by overfitting, compared to the DR estimator. This finding is consistent with the advantage of the DML as emphasized in Chernozhukov et al. (2018). This observation also provides the rationale for the use of cross-fitting in estimating the nuisance function in the DR estimator of Wang (2018).

4 An Empirical Application

We apply the proposed method to estimate the ATE on unemployment duration to evaluate the effect of participating in an e-learning program. We utilize the Japanese Panel Study of Employment Dynamics (JPSED) dataset, which is provided by the Recruit Works Institute. The JPSED collects data on employment status among Japanese individuals, including information on individual characteristics such as gender, age, occupation, residential area, and education level. From the 2020 survey, we extract 2,833 individuals who resigned from their previous work in 2019. We investigate whether the experience of participating in an e-learning program helped to reduce the unemployment duration in 2019. In our sample of 2,833 individuals, 141 participated in the e-learning program and will be considered as the treatment group. The data indicate that the treatment group comprises more males and individuals with higher education levels than the control group (see Supplementary Table 2 for the details). Figure 3 shows the survival curves of two groups, estimated using the IPW, DR, and DML estimators, with 45 individual characteristics as covariates. In Table 2, the estimated ATEs at 3, 6, and 9 months are all negative, indicating that the unemployment duration tends to be shorter for the treatment group. The 95 percent confidence intervals, calculated using the procedure described in Chernozhukov et al. (2018), exclude the zero and positive regions. Hence, we conclude that participating in an e-learning program significantly increases the job-finding rate among individuals seeking employment.

5 Concluding Remarks

This paper discusses the use of DML for estimating the ATE on a survival function using pseudo-observations. Through simulations, we have demonstrated the double robustness property of our method, as well as that of the DR estimator. We also have confirmed the improved performance, compared to DR estimators in the presence of many covariates. Our results show the advantage of using DML in the context of the survival analysis.

Acknowledgments

The authors express their gratitude to Jixian Wang for kindly sharing the R source code. The authors thank the Social Science Japan Data Archive, the Center for Social Research and Data Archives, the Institute of Social Science, the University of Tokyo, for permission to use the data.

Funding Sources

This work is supported by JSPS KAKENHI (JP20H01482 and JP23K17245) and the Uehara Memorial Foundation.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could appear to have influenced the work reported in this paper.

References

Andersen, P. K., E. Syriopoulou, and E. T. Parner. 2017. "Causal inference in survival analysis using pseudo-observations." *Statistics in Medicine* 36: 2669–2681.

Andersen, P. K., J. P. Klein, and S. Rosthøj. 2003. "Generalised linear models for correlated pseudo-observations, with applications to multi-state models." *Biometrika* 90: 15–27.

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

Graw, F., T. A. Gerds, and M. Schumacher. 2009. "On pseudo-values for regression analysis in competing risks models." *Lifetime Data Analysis* 15: 241–255.

Klein, J. P., B. Logan, M. Harhoff, and P. K. Andersen. 2007. "Analyzing survival curves at a fixed point in time." *Statistics in Medicine* 26: 4505–4519.

Wang, J. 2018. "A simple, doubly robust, efficient estimator for survival functions using pseudo observations." *Pharmaceutical statistics* 17: 38–48.

Tables

Table 1. Performance of ATE estimators

Case	Propensity Score	Survival Function	Metrics	IPW	DR	DML
1	Correct	Correct	Absolute Bias	0.057	0.002	0.006
			SD	0.176	0.168	0.163
			RMSE	0.185	0.168	0.163
2	Correct	Incorrect	Absolute Bias	0.057	0.065	0.062
			SD	0.176	0.177	0.167
			RMSE	0.185	0.189	0.179
3	Incorrect	Correct	Absolute Bias	0.267	0.037	0.025
			SD	0.135	0.132	0.131
			RMSE	0.299	0.138	0.133
4	Incorrect	Incorrect	Absolute Bias	0.267	0.271	0.269
			SD	0.135	0.134	0.133
			RMSE	0.299	0.302	0.300

Table 2. The ATE of e-learning program participation on unempolyment probability

	Time after becoming unemployed		
Estimators	**3 months**	**6 months**	**9 months**
IPW	-0.256 [-0.260, -0.252]	-0.219 [-0.222, -0.215]	-0.159 [-0.162, -0.156]
DR	-0.253 [-0.257, -0.249]	-0.215 [-0.218, -0.212]	-0.156 [-0.159, -0.154]
DML	-0.248 [-0.252, -0.243]	-0.209 [-0.213, -0.206]	-0.154 [-0.158, -0.151]

Notes: The 95 percent confidence intervals are shown in parentheses.

Figures

Figure 1. Distribution of ATE estimates

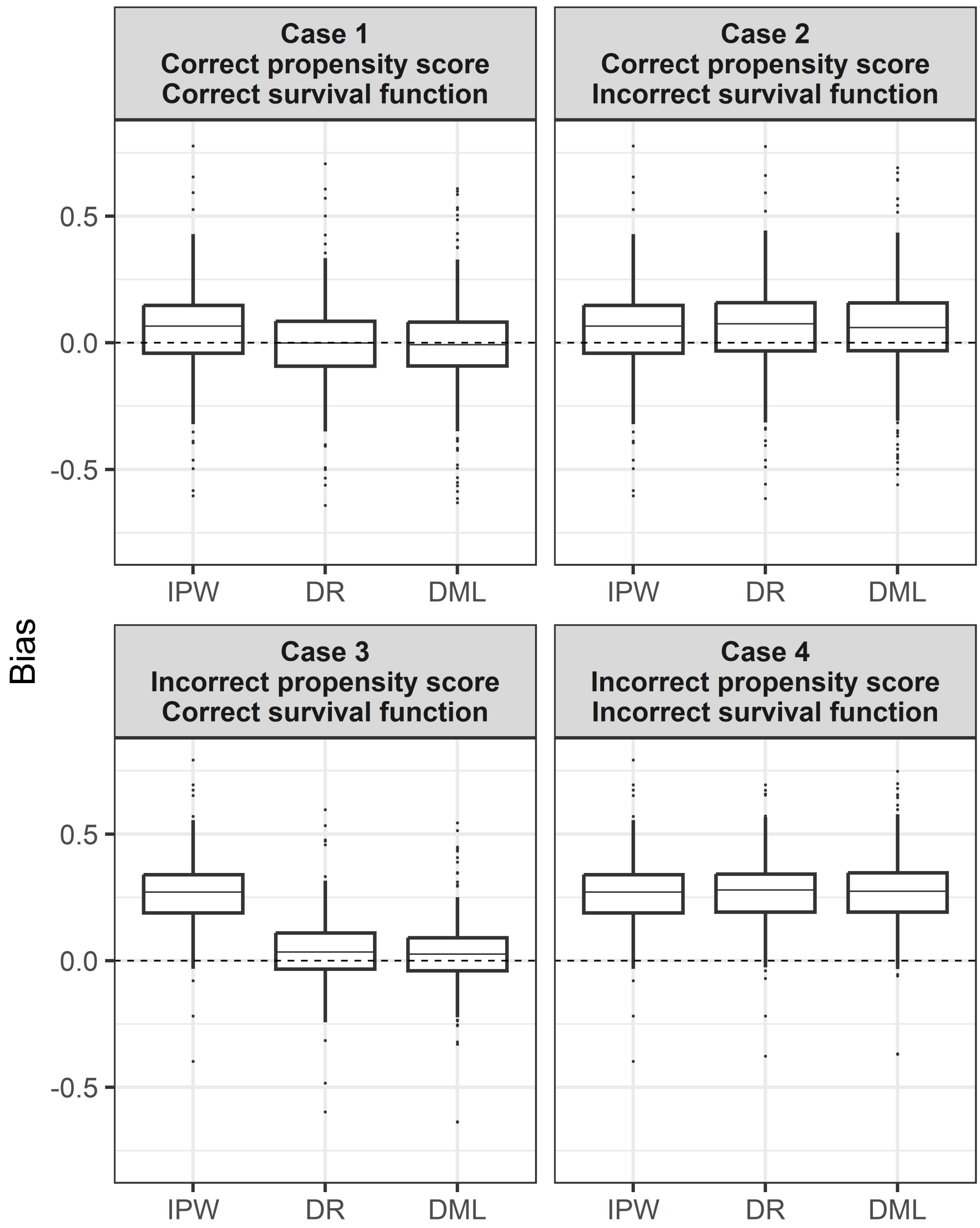


Figure 2. Effect of increasing relative number of covariates

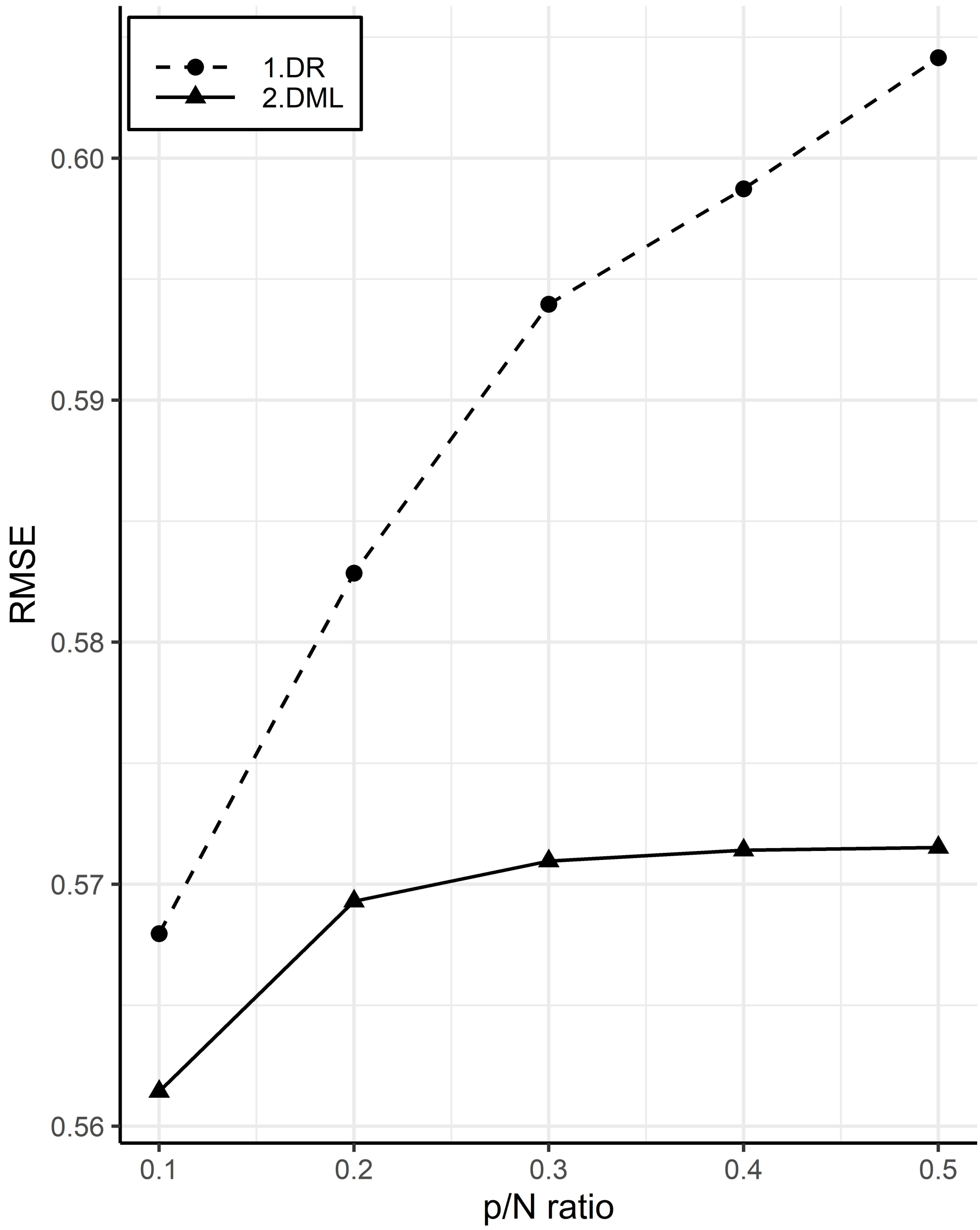
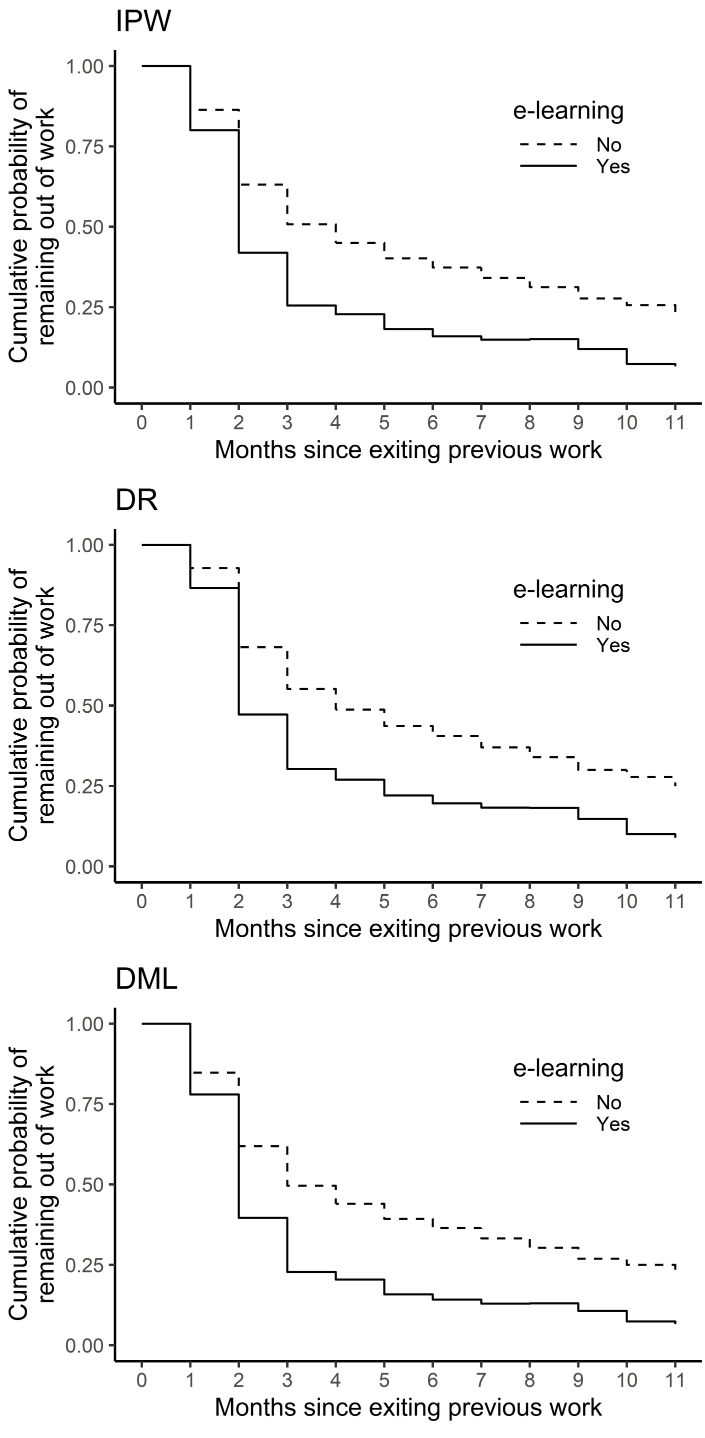


Figure 3. Effect of e-learning program participation on survival curves for unemployment duration



Supplementary material

Supplementary Table 1. Performance of ATE estimators and relative number of covariates

p/N ratios	Metrics	DR	DML
	Absolute Bias	0.568	0.561
0.1	SD	0.019	0.019
	RMSE	0.568	0.561
	Absolute Bias	0.582	0.569
0.2	SD	0.025	0.025
	RMSE	0.583	0.569
	Absolute Bias	0.593	0.570
0.3	SD	0.033	0.033
	RMSE	0.594	0.571
	Absolute Bias	0.598	0.570
0.4	SD	0.032	0.034
	RMSE	0.599	0.571
	Absolute Bias	0.603	0.570
0.5	SD	0.041	0.044
	RMSE	0.604	0.572

Supplementary Table 2. Variables in the data

Characteristic	Overall (N = 2,833)	e-learning program participation No (N = 2,692)	Yes (N = 141)
Gender			
Female	1,558 (55%)	1,494 (55%)	64 (45%)
Male	1,275 (45%)	1,198 (45%)	77 (55%)
Age at retirement	42.29 (14.87)	42.28 (14.93)	42.51 (13.77)
Current residential area			
Hokkaido region	151 (5.3%)	145 (5.4%)	6 (4.3%)
Tohoku region	214 (7.6%)	204 (7.6%)	10 (7.1%)
North Kanto region	137 (4.8%)	128 (4.8%)	9 (6.4%)
South Kanto region	880 (31%)	827 (31%)	53 (38%)
Hokuriku region	112 (4.0%)	106 (3.9%)	6 (4.3%)
Tokai region	328 (12%)	318 (12%)	10 (7.1%)
Kansai region	495 (17%)	471 (17%)	24 (17%)
Chugoku region	141 (5.0%)	136 (5.1%)	5 (3.5%)
Shikoku region	77 (2.7%)	72 (2.7%)	5 (3.5%)
Kyushu region	298 (11%)	285 (11%)	13 (9.2%)
Final education			
Completed primary/junior high school	76 (2.7%)	76 (2.8%)	0 (0%)
Completed high school	997 (35%)	957 (36%)	40 (28%)
Completed vocational school (technical college)	424 (15%)	409 (15%)	15 (11%)
Completed junior college	294 (10%)	282 (10%)	12 (8.5%)
Completed technical college	39 (1.4%)	34 (1.3%)	5 (3.5%)
Completed university	870 (31%)	812 (30%)	58 (41%)
Completed graduate school (master's/doctoral program)	94 (3.3%)	84 (3.1%)	10 (7.1%)
Currently enrolled	39 (1.4%)	38 (1.4%)	1 (0.7%)
Presence of spouse			
No spouse	1,432 (51%)	1,359 (50%)	73 (52%)
Spouse	1,401 (49%)	1,333 (50%)	68 (48%)
Presence of children			
No children	1,605 (57%)	1,522 (57%)	83 (59%)
Children	1,228 (43%)	1,170 (43%)	58 (41%)
Residential status			
Own home	1,599 (56%)	1,520 (56%)	79 (56%)
Rental/Other	1,234 (44%)	1,172 (44%)	62 (44%)
Main earner			
Self	1,391 (49%)	1,301 (48%)	90 (64%)
Spouse	821 (29%)	794 (29%)	27 (19%)
Other	621 (22%)	597 (22%)	24 (17%)
Reason for leaving previous job			
End of contract period	421 (15%)	401 (15%)	20 (14%)
Retirement	158 (5.6%)	147 (5.5%)	11 (7.8%)
Company bankruptcy/business closure	111 (3.9%)	106 (3.9%)	5 (3.5%)
Retirement recommendation	61 (2.2%)	58 (2.2%)	3 (2.1%)
Dismissal	50 (1.8%)	50 (1.9%)	0 (0%)
Transfer	17 (0.6%)	16 (0.6%)	1 (0.7%)
Early retirement	43 (1.5%)	42 (1.6%)	1 (0.7%)
Dissatisfaction with wage	194 (6.8%)	181 (6.7%)	13 (9.2%)
Dissatisfaction with working conditions or workplace	218 (7.7%)	203 (7.5%)	15 (11%)
Dissatisfaction with human relationships	352 (12%)	343 (13%)	9 (6.4%)
Dissatisfaction with job content	264 (9.3%)	250 (9.3%)	14 (9.9%)
Anxiety about company future or employment stability	158 (5.6%)	144 (5.3%)	14 (9.9%)
Personal physical injury or illness	120 (4.2%)	119 (4.4%)	1 (0.7%)
Personal mental illness	137 (4.8%)	133 (4.9%)	4 (2.8%)
Marriage	63 (2.2%)	62 (2.3%)	1 (0.7%)
Pregnancy/Childbirth	70 (2.5%)	69 (2.6%)	1 (0.7%)
Child-rearing	35 (1.2%)	34 (1.3%)	1 (0.7%)
Caretaking	41 (1.4%)	40 (1.5%)	1 (0.7%)
Spouse's transfer	39 (1.4%)	36 (1.3%)	3 (2.1%)
Independence	22 (0.8%)	22 (0.8%)	0 (0%)
Taking over family business or assisting family's work	14 (0.5%)	12 (0.4%)	2 (1.4%)
Pursuing education or obtaining qualification	35 (1.2%)	32 (1.2%)	3 (2.1%)
Other	210 (7.4%)	192 (7.1%)	18 (13%)

Notes: Mean for age at retirement with standard deviation in parenthesis. Count for other variables with proportion in parenthesis.