Economics Letters 135 (2015) 42 45

Contents lists available at ScienceDirect

Economics Letters

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

Estimating average treatment effect by model averaging

Yichen Gao [a], Wei Long [b,*], Zhengwei Wang [c]

[a] ISEM, the Capital University of Economics & Business, Beijing, 100070, China

[b] Department of Economics, Tulane University, New Orleans, LA 70118, United States

[c] PBC School of Finance, Tsinghua University, Beijing, 100083, China

HIGHLIGHTS

- We propose to use a model average method to improve the estimation of average treatment effects.
- The proposed model average estimator selects weight optimally to minimize estimation mean squared errors.
- Simulation results show that the model average estimator exhibits smaller estimation mean squared errors in post-treatment prediction than AIC or AICC methods.

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
17 July 2015
Accepted 2 August 2015
Available online 8 August 2015

JEL classification:
C5

Keywords:
Average treatment effect
Counterfactual
Model average

ABSTRACT

replacing the two-step selection strategy with JMA method, we improve the post-treatment prediction of Hsiao et al. (2012) in terms of mean squared prediction errors (PMSE).

The rest of the paper is organized as follows. Section 2 briefly reviews both Hsiao et al. (2012) method and the JMA method. Section 3 reports simulation results to examine the finite sample performance of our proposed method. Section 4 concludes the paper.

2. Theoretical model

In this section we briefly discuss the estimation method in Hsiao et al. (2012). Suppose there is no treatment to all units up to T_1. At time $T_1 + 1$, there is only one unit that receives a treatment. Let y_t be the treatment unit's outcome at time t. Correspondingly, let $x_t = (x_{1t}, \ldots, x_{Nt})'$ be the outcomes of N control units at time t.[1] Hsiao et al. (2012) consider the case that both treatment and control units' outcomes are generated by a factor model (e.g., Bai and Ng, 2002) in the absence of treatment for $t = 1, \ldots, T_1$:

$$\tilde{y}_t = a + Bf_t + u_t, \tag{1}$$

where $\tilde{y}_t = (y_t, x_{1t}, \ldots, x_{Nt})'$, $a = (a_1, \ldots, a_{N+1})'$, f_t is a $K \times 1$ vector of common factors (they may be unobservable) that affect outcomes, B is a $(N + 1) \times K$ matrix of factor loading, $u_t = \left(u_{1t}, \ldots, u_{(N+1)t}\right)'$ is a vector of idiosyncratic error. Let y_t^1 and y_t^0 denote the outcomes of the treated unit with and without the policy intervention, respectively. Given that there is a treatment at time $T_1 + 1$, we are interested in estimating the average treatment effects $\Delta_1 = E(y_t^1 - y_t^0)$. The difficulty is that we cannot observe y_t^0 for $t \geq T_1 + 1$. Hsiao et al. (2012) suggest using control units' outcomes x_t to estimate y_t^0 when $t \geq T_1 + 1$. This can be done by replacing f_t by x_t in the treatment unit's equation $y_t = a_1 + b_1' f_t + u_t$ to obtain

$$y_t = \gamma_0 + x_t'\gamma + \upsilon_t, \tag{2}$$

for $t = 1, \ldots, T_1$, where γ_0 is intercept, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_N)'$, υ_t satisfies that $E(\upsilon_t) = 0$, $E(\upsilon_t x_t) = 0$ and $var(\upsilon_t)$ is finite. Let $\hat{\gamma}_0$ and $\hat{\gamma}$ denote the least square estimators of γ_0 and γ based on (2), then we estimate the counterfactual outcome of y_t^0 by

$$\hat{y}_t^0 = \hat{\gamma}_0 + x_t'\hat{\gamma}, \tag{3}$$

for $t = T_1 + 1, \ldots, T$. Let $T_2 = T - T_1$, then the average treatment effect is estimated by

$$\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} \left(y_t - \hat{y}_t^0\right). \tag{4}$$

In application, N may not be small relative to T_1. Thus, it is advantageous to use only a subset of the N control units rather than all of them to predict the counterfactuals. For N control units, there are 2^N different models, and the most appropriate model should balance the within-sample fit with the out-sample prediction error. Hsiao et al. (2012) propose a two-step model selection procedure to find out which model is the most appropriate. Specifically, in the first step, they use R^2 to select the best predictor for y_t^0 using k control units out of N control units, denoted by $M(k)^*$, for $k = 1, \ldots, N$. Then, in the second step, from $M(1)^*, M(2)^*, \ldots, M(N)^*$, they pin down one best model from the N candidate models in terms of model selection criterion such as Akaike information criterion (Tf   4.799   0   Td   [(/)]TJ/F174   6.6948   Tf   3.568   3.472   Td   [()]TJ/Fen

Table 1
Comparison of PMSE between model average method (MA) and HCW method (A[illegible]

	$\sigma^2 = 1$			$\sigma^2 = 0.5$			[illegible].1		
	MA	AICC	AIC	MA	[illegible]	[illegible]	[illegible]	AICC	AIC
$T_1 = 25, T = 35$									
Avg. #		3.56	10.92		[illegible]	[illegible]		3.66	10.95
PMSE	1.8404	2.0967	6.0158	0.9199	[illegible]	[illegible]	.1838	0.2101	0.7461
$T_1 = 40, T = 50$									
Avg. #		3.71	5.91		[illegible]	[illegible]		3.76	5.97
PMSE	1.3841	1.7310	1.8935	0.6919	[illegible]	[illegible]	0.1383	0.1728	0.1915
$T_1 = 60, T = 70$									
Avg. #		4.14	5.60		[illegible]	[illegible]		4.19	5.39
PMSE	1.2553	1.4180	1.4722	0.6282	[illegible]	[illegible]	0.1258	0.1453	0.1496

Table 2
Comparison of PMSE between model average method (MA) and HCW method (A[illegible]

	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
	MA	AICC	AIC	MA	[illegible]	[illegible]	MA	AICC	AIC
$T_1 = 25, T = 35$									
Avg. #		4.16	11.09		[illegible]	[illegible]		4.24	11.15
PMSE	2.0749	2.4358	6.2314	1.0626	[illegible]	[illegible]	0.2195	0.2372	0.6337
$T_1 = 40, T = 50$									
Avg. #		3.94	6.25		[illegible]	[illegible]		3.98	6.35
PMSE	1.6196	1.8450	1.9634	0.8160	[illegible]	[illegible]	0.1639	0.1865	0.1970
$T_1 = 60, T = 70$									
Avg. #		4.35	5.58		[illegible]	[illegible]		4.54	5.83
PMSE	1.3829	1.6194	1.6502	0.6962	[illegible]	[illegible]	.1400	0.1578	0.1605

where $\mathbf{S}_{T_1} = \frac{1}{T_1} \tilde{\mathbf{e}}' \tilde{\mathbf{e}}$ is a $N \times N$ matrix. The Jackknife weight $\hat{\mathbf{w}}$ is the value that minimizes (6) under the restrictions that each weight is between 0 and 1 and their summation equals to 1. Since Eq. (6) is quadratic in $\mathbf{w}$, we could get $\hat{\mathbf{w}}$ by applying the standard quadratic programming technique which requires short computing time. With the selected weight $\hat{\mathbf{w}}$ above, the Jackknife model average (JMA) estimator of μ could be written as $\hat{\mu}(\hat{\mathbf{w}}) = \hat{\mu}\hat{\mathbf{w}}$, and our model averaging estimation of the treatment effect defined in (4) could be calculated through $\hat{y}_t^0$ [illegible]

Table 3
Comparison of PMSE between model average method (MA) and HCW method (AICC and AIC): 3 factors.

	$\sigma^2 = 1$			$\sigma^2 = 0.5$			$\sigma^2 = 0.1$		
	MA	AICC	AIC	MA	AICC	AIC	MA	AICC	AIC
$T_1 = 25; T = 35$									
Avg. #		4.40	11.18		4.54	11.70		4.74	11.72
PMSE	2.4758	2.8493	6.5231	1.2956	1.4640	3.2716	0.2732	0.3051	0.7024
$T_1 = 40; T = 50$									
Avg. #		4.90	7.10		5.08	7.22		5.04	7.24
PMSE	1.7900	1.9705	2.1136	0.9051	0.9716	1.0804	0.1829	0.1987	0.2095
$T_1 = 60; T = 70$									
Avg. #		5.12	6.34		5.18	6.36		5.24	6.42
PMSE	1.5205	1.6437	1.6842	0.7659	0.8302	0.8516	0.1542	0.1677	0.1685

method with Hsiao et al. (2012) by comparing the post-treatment mean squared prediction errors (PMSE), which is defined as

$$\mathrm{PMSE} = \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} (y_t^0 - \hat{y}_t^0)^2;$$

where $\hat{y}_t^0$ is the estimated counterfactual outcome by using the AIC method, or the AICC method, or our proposed Jackknife model average method.

We repeat each of the structures 1000 times. The results are displayed in Tables 1 3. The Avg. # is the average number of control units selected by the AIC or the AICC methods. Simulation results show that our method has smaller PMSE in all cases, indicating improved predicting performance by replacing the two-step model selection strategy with the JMA method.

4. Conclusion

In this paper we suggest to replace the two-step model selection strategy in Hsiao et al. (2012) with the Jackknife model average (JMA) method to estimate average treatment effect of a program or a policy. By applying the JMA, we show that the post-treatment predicting performance improves in terms of prediction mean squared error.

Acknowledgment

Zhengwei Wang's work is partially supported by the National Science Foundation of China (Project No. 71472100).

References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22 (1), 203 217.

Bai, C., Li, Q., Ouyang, M., 2014. Property taxes and home prices: A tale of two cities. J. Econometrics 180 (1), 1 15.

Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191 221.

Buckland, S., Burnham, K., Augustin, N., 1997. Model selection: an integral part of inference. Biometrics 53 (2), 603 618.

Hansen, B., 2007. Least squares model averaging. Econometrica 75 (4), 1175 1189.

Hansen, B., 2008. Least-squares forecast averaging. J. Econometrics 146 (2), 342 350.

Hansen, B., Racine, J., 2012. Jackknife model averaging. J. Econometrics 167 (1), 38 46.

Hoeting, J., Madigan, D., Raftery, A., Volinsky, C., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 14 (4), 382 417.

Hsiao, C., Ching, H. ... 2012. ... approach ... J. Econometrics