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

In this paper, we propose to use a model average method to improve the estimation performance of Hsiao et al. (2012) panel data approach for program evaluation. Instead of using the two-step model selection strategy which chooses one best model according to a criterion such as AIC or AICC, we average over a set of candidate models. Simulation results show that the model average estimator exhibits smaller estimation errors in post-treatment prediction than AIC or AICC method.

1. Introduction

Econometric models for treatment effect estimation have been extensively discussed in literature these days. To estimate the treatment effect of a program or a policy, one needs to measure the difference between the outcomes under treatment and the outcomes when treatment is absent, since we do not observe both outcomes simultaneously. Hsiao et al. (2012) propose a panel data approach to estimate the counterfactual outcomes. This method relies on the correlation between the treatment and the control units. Hsiao et al. (2012) argue that it is the presence of common factors that cause the cross-sectional dependence and drive co-movement of all the relevant cross-sectional outcome variables. Bai et al. (2014) further extend Hsiao et al. (2012) method to allow for non-stationary data. In finite sample applications, Hsiao et al. (2012) propose to select the best model through a two-step strategy, which is based on comparing goodness-of-fit statistic and widely used model selection criterion such as Akaike information criterion (AIC, Akaike, 1970) and corrected Akaike information criterion (AICC, Hurvich and Tsai, 1989).

In this paper, we contribute to improve the estimating performance of Hsiao et al. (2012) method by using the Jackknife model average (JMA) method, which is proposed by Hansen and Racine (2012). Instead of selecting one specific “best” model based on a criterion, model average method addresses the model uncertainty problem by averaging over the set of candidate models. We direct interested readers to Buckland et al. (1997), Hansen (2007) and Wan et al. (2010) for frequentist method for model averaging. Hoeting et al. (1999) for Bayesian model averaging. Under model average framework, we select weights for each candidate model by minimizing a cross-validation criterion function. Hansen and Racine (2012) show that the computing procedure of JMA is an application of the quadratic programming technique. They further prove that JMA estimator is asymptotically optimal in the sense of achieving the lowest possible expected squared estimation loss. By
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 $y_{0t} = (x_{0t1}, \ldots, x_{0tn})'$ be the outcomes of $N$ control units at time $t$. 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 $t = 1, \ldots, T_1$:

\[
y_{0t} = \alpha + B f_{0t} + u_t,
\]

where $y_{0t} = (y_{0t1}, \ldots, y_{0tn})'$, $\alpha = (\alpha_1, \ldots, \alpha_{n+1})'$, $f_t$ 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, $u_t = (u_{0t1}, \ldots, u_{0tn})'$ is a vector of idiosyncratic error. Let $y_t$ and $y_{0t}$ 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 $\Delta_1 = E(y_t - y_{0t})$. The difficulty is that we cannot observe $y_{0t}$ for $t \geq T_1+1$. Hsiao et al. (2012) suggest using control units' outcomes $y_{0t}$ to estimate $y_t$ when $t \geq T_1+1$. This can be done by replacing $f_t$ by $x_t$ in the treatment unit’s equation $y_t = \alpha + B x_t + u_t$ to obtain

\[
y_t = y_0 + x_t' \gamma + v_t,
\]

for $t = 1, \ldots, T_1$, where $y_0$ is intercept, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)'$, $v_t$ satisfies that $E(v_t) = 0$, $E(v_t x_t) = 0$ and var($v_t$) is finite. Let $\hat{y}_0$ and $\hat{y}$ denote the least square estimators of $y_0$ and $\gamma$ based on (2), then we estimate the counterfactual outcome of $y_0$ by

\[
\hat{y}_0^t = \hat{y}_0 + x_t' \hat{\gamma},
\]

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

\[
\hat{\Delta}_1 = \frac{1}{T_2} \sum_{t=T_1+1}^{T} (y_t - \hat{y}_0^t).
\]

In this paper, we avoid choosing one “best” model out of the $2^N$ models as it could be time consuming when $N$ is large. On the contrary, we apply the model average method. First, we regress $y_t$ on the outcomes of each individual control unit $x_{0t1}, \ldots, x_{0tn}$, $t = 1, 2, \ldots, T_1$, and then obtain their respective goodness-of-fit statistics $R_{11}^2, R_{12}^2, \ldots, R_{1n}^2$. We order these goodness-of-fit statistics so that $R_{11}^2 \geq R_{12}^2 \geq \cdots \geq R_{1n}^2 \geq R_{N}^2$, where $R_{j}^2$ is the order statistic, and $x_{0t1}, x_{0t2}, \ldots, x_{0tn}$ correspond to the regressor in each model, i.e., $x_{0jt}$ is the regressor that has a goodness-of-fit $R_{j}^2$, $j = 1, \ldots, N$. Second, we form our first candidate model by regressing $y_{0t}$ on $(x_{0tj})$ and then denote the first model’s linear fit by a $T_1 \times 1$ vector $\hat{\mu}_1 = (\hat{\mu}_{11}, \hat{\mu}_{12}, \ldots, \hat{\mu}_{1n})'$, where $\hat{\mu}_{1t} = (x_{0t1}, x_{0t2}, \ldots, x_{0tn}) \hat{\beta}_{1t}^1$, with $t = 1, 2, \ldots, T_1$ and $x_{0tj}$ represents the observation for $x_{0jt}$ at time $t$. $\hat{\mu}_1^1 = (\hat{\mu}_1^1, \hat{\mu}_2^1, \ldots, \hat{\mu}_n^1)$ denotes the OLS estimator based on regressing $y_{0t}$ on $(x_{0tj})$ with $t = 1, \ldots, T_1$, the second candidate model regresses $y_{0t}$ on $(x_{0t1}, x_{0t2}, \ldots, x_{0tn})$ with $t = 1, \ldots, T_1$, and the corresponding linear estimate is $\hat{\mu}_2 = (\hat{\mu}_2^1, \hat{\mu}_2^2, \ldots, \hat{\mu}_n^2)$, where $\hat{\mu}_2 = (x_{0t1}, x_{0t2}, \ldots, x_{0tn}) \hat{\beta}_{2}^1$, for $t = 1, 2, \ldots, T_1$ and $\hat{\beta}_{2}$ denotes the OLS estimator obtained from regressing $y_{0t}$ on $(x_{0t1}, x_{0t2}, \ldots, x_{0tn})$ with $t = 1, \ldots, T_1$. We continue this procedure until all $N$ control units are included, i.e., the last one is to regress $y_{0t}$ on $(x_{0t1}, x_{0t2}, \ldots, x_{0tn})$ and we denote the fit as $\hat{\mu}_N = (\hat{\mu}_N^1, \hat{\mu}_N^2, \ldots, \hat{\mu}_N^n)$, where $\hat{\mu}_N^t = (x_{0t1}, x_{0t2}, \ldots, x_{0tn}) \hat{\beta}_{N}^1$, denotes the OLS estimator obtained from regressing $y_{0t}$ on $(x_{0t1}, x_{0t2}, \ldots, x_{0tn})$ with $t = 1, \ldots, T_1$. Thus, we obtain $N$ candidate models’ fits with the corresponding set of estimates denoted as $[\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N]$. Let $w = (w^1, w^2, \ldots, w^N)$ be a set of weights which are non-negative and $\sum_{i=1}^{N} w^i = 1$, given $w$, a model averaging estimator could be formulated as

\[
\hat{\mu}(w) = \sum_{i=1}^{N} w^i \hat{\mu}_i = \hat{\mu} \cdot w,
\]

where $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N)$ is a $T_1 \times N$ matrix.

Hansen and Racine (2012) propose to select the empirical weights $w$ in the Jackknife manner (leave-one-out cross-validation). For the ith model’s estimate $\hat{\mu}_i$, we denote the Jackknife estimator as a $T_1 \times 1$ vector $\hat{\mu}_i = (\hat{\mu}_{i1}, \hat{\mu}_{i2}, \ldots, \hat{\mu}_{i(T_1-1)}$, where $\hat{\mu}_{ik}$ represents the estimator of $\hat{\mu}_i$ with the kth observation deleted. Thus, the Jackknife residual for the ith model, written in a $T_1 \times 1$ vector form, is $\hat{e}_i = y - \hat{\mu}_i$, where $y = (y_1, y_2, \ldots, y_{T_1})$. We further define the Jackknife version of the averaging estimator as

\[
\hat{\mu}_i(w) = \sum_{i=1}^{N} w^i \hat{\mu}_i = \hat{\mu}_i \cdot w.
\]

where $\hat{\mu}_i = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N)$ is a $T_1 \times N$ matrix, and the Jackknife residuals as

\[
\hat{e}(w) = y - \hat{\mu}(w) = \sum_{i=1}^{N} w^i y - \sum_{i=1}^{N} w^i \hat{\mu}_i = \sum_{i=1}^{N} w^i (y - \hat{\mu}_i) = \sum_{i=1}^{N} w^i \hat{e}_i = \hat{e} \cdot w,
\]

where $\hat{e} = (\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_N)$ is a $T_1 \times N$ matrix. Then, the least square cross-validation criterion function could be written as

\[
CV_{T_1}(w) = \frac{1}{T_1} \hat{e}(w)' \hat{e}(w) = \textbf{w}' S_{T_1} \textbf{w},
\]

1 Under Hsiao et al. (2012), there are $N+1$ units $y_1, y_2, \ldots, y_{N+1}$. The first identity is assumed to be the treatment unit and the remained $N$ are control units. So under our notations, $y = (y_1, x_1, x_2 = (y_2, \ldots, x_N)$. Under Hsiao et al. (2012), there are $N+1$ units $y_1, y_2, \ldots, y_{N+1}$. The first identity is assumed to be the treatment unit and the remained $N$ are control units. So under our notations, $y = (y_1, x_1, x_2 = (y_2, \ldots, x_N)$. Under Hsiao et al. (2012), there are $N+1$ units $y_1, y_2, \ldots, y_{N+1}$. The first identity is assumed to be the treatment unit and the remained $N$ are control units. So under our notations, $y = (y_1, x_1, x_2 = (y_2, \ldots, x_N)$.
where $S_{T_1} = \frac{1}{T_1} \mathbf{e}^\top \mathbf{e}$ is a $N \times N$ matrix. The Jackknife weight $\mathbf{w}$ 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 $\mathbf{w}$ by applying the standard quadratic programming technique which requires short computing time. With the selected weight $\mathbf{w}$ above, the Jackknife model average (JMA) estimator of $\mu$ could be written as $\hat{\mu} (\mathbf{w}) = \mathbf{w} \hat{\mu}$, and our model averaging estimation of the treatment effect defined in (4) could be calculated through $\bar{y}_t^{\mathbf{w}} = \hat{\mu} (\mathbf{w}) = \sum_{j=1}^{N} \hat{\mu}_t^j$ for $t = T_1 + 1, \ldots, T$.

Let $\mu_t = E(y_t|x_t)$ denotes the conditional mean of $y_t$ given $x_t$, $t = 1, 2, \ldots, T$, so that $\mu = (\mu_1, \mu_2, \ldots, \mu_T)^\top$ is a $T_1 \times 1$ vector. Define $L_{T_1} (\mathbf{w}) = \frac{1}{T_1} (\mu_1 - \hat{\mu} (\mathbf{w})) (\mu_2 - \hat{\mu} (\mathbf{w}))$ as a measure for the fit of the pre-treatment period, and $R_{T_1} (\mathbf{w}) = E (\frac{1}{T_1} (\mathbf{y}^0 - \hat{\mu} (\mathbf{w}))^\top (\mathbf{y}^0 - \hat{\mu} (\mathbf{w})))$ as a measure for the prediction errors of the counterfactual outcomes for the post-treatment period, where $\mathbf{y}^0 = (y_{t+1}^{0,1}, y_{t+2}^{0,1}, \ldots, y_T^{0,1})$, $\hat{\mu}^0 = (\hat{\mu}_1^{0,1}, \hat{\mu}_2^{0,1}, \ldots, \hat{\mu}_N^{0,1})$ is a $T_2 \times N$ matrix with $\hat{\mu}_t^{0,i} = (\hat{\mu}_1^{0,i}, \hat{\mu}_2^{0,i}, \ldots, \hat{\mu}_N^{0,i})$ as a $T_2 \times 1$ vector, where for $j = 1, \ldots, N$, $\hat{\mu}_t^{0,j} = (1, x_{(1,t)}, \ldots, x_{(N,t)})^\top$ (for $t = T_1 + 1, T_1 + 2, \ldots, T$) with $\mathbf{y}_t^{0,i}$ being the OLS estimate of the coefficient vector from regressing $y_t$ on $(1, x_{(1,t)}, \ldots, x_{(N,t)})$ using the pre-treatment period data $t = 1, \ldots, T$. Thus, $\hat{\mu}^0$ represents the out-sample predictions of the counterfactual outcome from the $N$ candidate models. Therefore, $\hat{\mu}^0 (\mathbf{w})$ is a $T_2 \times 1$ vector the model averaging estimate of the counterfactual outcome $\mathbf{y}^0 = (y_{t+1}^{0,1}, y_{t+2}^{0,1}, \ldots, y_T^{0,1})$. Under mild assumptions, Hansen and Racine (2012) prove the asymptotic optimality of the Jackknife selected weights by showing the following:

$$L_{T_1} (\mathbf{w}) \rightarrow \inf_{w \in \mathcal{D}_w} L_{T_1} (\mathbf{w}) \quad \text{as} \quad T_1 \rightarrow \infty,$$

(7)

where $\mathcal{D}_w = \{w \in \mathbb{R}_+^N \mid \sum_{j=1}^{N} w_j = 1\}$. Eq. (7) indicates that the mean of squared estimation loss of the Jackknife model average (JMA) estimator is asymptotically identical to that of the infeasible best possible model averaging estimator. Let $\mu_0 = (\mu_{T_1+1}, \mu_{T_1+2}, \ldots, \mu_{T_1+T_2})^\top$ so $R_{T_1} (\mathbf{w}) = E (\frac{1}{T_1} (\mu_0 - \hat{\mu} (\mathbf{w})) (\mu_0 - \hat{\mu} (\mathbf{w})) + \frac{1}{T_1} \sum_{t=1}^{T_1} E (\epsilon_t^2)$, where the second term is unrelated to $\mathbf{w}$. Hansen (2008) proves $E (\frac{1}{T_1} (\mu_0 - \hat{\mu} (\mathbf{w})) (\mu_0 - \hat{\mu} (\mathbf{w})) = E (l_{T_1} (\mathbf{w})) (1+ o_p (1)) \approx E (l_{T_1} (\mathbf{w}))$, which together with the asymptotic optimality shown by (7) prompts us to use the JMA estimator for out-sample prediction in the current study.

### 3. Numerical studies

We compare the performance of the Jackknife model average method with the method in Hsiao et al. (2012) through Monte Carlo simulations. For comparing purposes, we generally follow the settings of their work: we generate the model with 1 treatment unit and 20 control units ($N = 20$), and the pre-treatment period $T_1 = 25, 40$ and 60. The post-treatment period includes $T - T_1 = 10$ observations. As in Hsiao et al. (2012), we consider 1-factor, 2-factor and 3-factor structures as the following:

1-factor: $f_{1t} = 0.95 f_{1t-1} + \epsilon_{1t},$

2-factor: $f_{1t} = 0.3 f_{1t-1} + \epsilon_{1t},$

$f_{2t} = 0.6 f_{2t-1} + \epsilon_{2t},$

3-factor: $f_{1t} = 0.8 f_{1t-1} + \epsilon_{1t},$

$f_{2t} = -0.6 f_{2t-1} + \epsilon_{2t} + 0.8 \epsilon_{2t-1},$

$f_{3t} = \epsilon_{3t} + 0.9 \epsilon_{3t-1} + 0.4 \epsilon_{3t-2},$

where $\epsilon_t$ is distributed to $N(0,1), i = 1, 2, 3, \mu_0$ in (1) is generated from $N(0, \sigma^2)$ where $\sigma^2 = 1, 0.5$ and 0.1, and $B$ is generated from $N(1, 1)$. We examine the performance of our model average...
method with Hsiao et al. (2012) by comparing the post-treatment mean squared prediction errors (PMSE), which is defined as

$$PMSE = \frac{1}{T - T_1} \sum_{t = T_1 + 1}^T \left( \hat{y}_t^0 - \bar{y}_t^0 \right)^2,$$

where \( \bar{y}_t^0 \) 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

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