The Stata Journal (2022)
22, Number 4, pp. 842–883 DOI: 10.1177/1536867X221140960

rcm: A command for the regression control method

Guanpeng Yan
Shandong University
Jinan, China
guanpengyan@yeah.net

Qiang Chen
Shandong University
Jinan, China
qiang2chen2@126.com

Abstract. The regression control method, also known as the panel-data approach for program evaluation (Hsiao, Ching, and Wan, 2012, *Journal of Applied Econometrics* 27: 705–740; Hsiao and Zhou, 2019, *Journal of Applied Econometrics* 34: 463–481), is a convenient method for causal inference in panel data that exploits cross-sectional correlation to construct counterfactual outcomes for a single treated unit by linear regression. In this article, we present the `rcm` command, which efficiently implements the regression control method with or without covariates. Available methods for model selection include best subset, lasso, and forward stepwise and backward stepwise regression, while available selection criteria include the corrected Akaike information criterion, the Akaike information criterion, the Bayesian information criterion, the modified Bayesian information criterion, and cross-validation. Estimation and counterfactual predictions can be made by ordinary least squares, lasso, or postlasso ordinary least squares. For statistical inference, both the in-space placebo test using fake treatment units and the in-time placebo test using a fake treatment time can be implemented. The `rcm` command produces a series of graphs for visualization along the way. We demonstrate the use of the `rcm` command by revisiting classic examples of political and economic integration between Hong Kong and mainland China (Hsiao, Ching, and Wan 2012) and German reunification (Abadie, Diamond, and Hainmueller, 2015, *American Journal of Political Science* 59: 495–510).

Keywords: st0693, rcm, regression control method, panel-data approach, program evaluation, causal inference, counterfactual outcomes

1 Introduction

The regression control method (RCM), also known as the panel-data approach (PDA) for program evaluation (Hsiao, Ching, and Wan 2012), is a convenient method for causal inference that exploits cross-sectional correlation to construct counterfactual outcomes for a single treated unit by linear regression. Essentially, the RCM uses control units to predict the counterfactual outcomes of the treated unit via linear regression. Its basic panel-data setting is similar to the synthetic control method (SCM) (Abadie and Gardeazabal 2003; Abadie, Diamond, and Hainmueller 2010), which constructs counterfactual outcomes for the treated unit by a linear combination of control units with optimal weights obtained by solving two nested optimization problems. Because Hsiao, Ching, and Wan (2012) use regression to construct the counterfactual control unit, we coin the term "regression control method" in the same spirit as the SCM.

Since its introduction, the RCM has seen widespread and growing applications in applied work (for example, Ouyang and Peng [2015], Du and Zhang [2015], Ke et al. [2017]). An advantage of the RCM over the SCM is its ease of computation via ordinary least squares (OLS) or lasso, whereas the SCM relies on numerical methods to solve for optimal weights constrained to be nonnegative and summed to one. Moreover, no covariates are necessary for the RCM, which reduces the cost of data collection and widens its scope of potential application. On the other hand, covariates are critical for the SCM to construct optimal weights, and applied researchers with no clear guidance as to what covariates to include must often resort to robustness checks with different sets of covariates. Nevertheless, Hsiao and Zhou (2019) introduce covariates to the RCM to further improve its counterfactual prediction.

To our knowledge, while the RCM can be implemented in R by the `pampe` package (Vega-Bayo 2015), there is no Stata command for the RCM yet. Therefore, in this article, we present the command `rcm`, which efficiently implements the RCM with or without covariates and offers significantly more functionalities than the R package `pampe`. First, `rcm` deals with cases with or without covariates, whereas no covariates are allowed in `pampe`. Second, while `pampe` relies solely on best subset regression with the Akaike information criterion (AIC) or corrected Akaike information criterion (AICc) for model selection, `rcm` offers best subset, lasso, and forward and backward stepwise regressions with the AIC, the AICc, the Bayesian information criterion (BIC), the modified Bayesian information criterion (MBIC) and cross-validation when appropriate. Third, while `pampe` uses only OLS for estimation and prediction, `rcm` provides OLS, lasso, and postlasso OLS. For statistical inference, both the in-space placebo test using fake treatment units and the in-time placebo test using a fake treatment time can be implemented. The `rcm` command produces a series of graphs for visualization along the way. We demonstrate the use of `rcm` by revisiting classic examples of political and economic integration between Hong Kong and mainland China (Hsiao, Ching, and Wan 2012) and German reunification (Abadie, Diamond, and Hainmueller 2015).

The rest of the article is organized as follows. Section 2 presents the model for the RCM with or without covariates. Section 3 introduces methods for model selection, estimation, and prediction. Section 4 discusses statistical inference via placebo tests. Section 5 presents the command `rcm`. Section 6 illustrates `rcm` by revisiting classic examples in Hsiao, Ching, and Wan (2012) and Abadie, Diamond, and Hainmueller (2015). Section 7 concludes.

2 The model

Suppose there are N cross-sectional units for $i = 1, \ldots, N$, observed over periods $t = 1, \ldots, T_0, T_0+1, \ldots, T$. Without loss of generality, assume the first unit with $i = 1$ is the treated unit, whereas all other units with $i = 2, \ldots, N$ are control units that form the so-called "donor pool". The policy intervention occurs at time $t = T_0+1$ and thereafter, which partitions the time series into two sections, that is, the pretreatment periods for $t = 1, \ldots, T_0$ and the posttreatment periods for $t = T_0 + 1, \ldots, T$.

Following Rubin's counterfactual framework (Rubin 1974), let y_{it}^1 and y_{it}^0 be the potential outcomes of unit i in period t with and without intervention, respectively. The fundamental problem of causal inference is that y_{it}^1 and y_{it}^0 cannot be observed at the same time. Instead, the outcome variable y_{it} is observed and takes the form

$$y_{it} = d_{it}y_{it}^1 + (1 - d_{it})\, y_{it}^0$$

where the treatment variable $d_{it} = 1$ if unit i is treated in period t and $d_{it} = 0$ otherwise. The treatment effect can be expressed as

$$\Delta_{it} = y_{it}^1 - y_{it}^0$$

The goal of the RCM is to obtain estimates of counterfactual outcomes $\widehat{y}_{it}^0$ during the posttreatment periods and predict the treatment effects for the first unit by $\widehat{\Delta}_{1t} = y_{1t}^1 - \widehat{y}_{1t}^0$.

2.1 The basic model

Suppose the cross-sectional correlation across units is driven by some unobserved common factors (that is, common shocks such as technology shocks, trade shocks, financial shocks, pandemic shocks), while their impacts on each unit are allowed to be heterogeneous. Specifically, assume the potential outcome of unit i at time t without treatment y_{it}^0 is determined by a pure linear factor model of the form

$$y_{it}^0 = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + \varepsilon_{it} \tag{1}$$

where α_i is an individual fixed effect, $\mathbf{f}_t$ is a $(r \times 1)$ vector of unobserved common factors, $\mathbf{b}_i'$ is a $(1 \times r)$ vector of unobserved factor loadings, and ε_{it} is an idiosyncratic shock. The strategy of Hsiao, Ching, and Wan (2012) is to eliminate the common factors $\mathbf{f}_t$ and express y_{1t}^0 for the treated unit as a function of $\left(y_{2t}^0, \ldots, y_{Nt}^0\right)$ for the control units. To this end, for units with $i = 2, \ldots, N$, stacking (1) into a vector yields

$$\widetilde{\mathbf{y}}_t = \widetilde{\boldsymbol{\alpha}} + \widetilde{\mathbf{B}}\mathbf{f}_t + \widetilde{\boldsymbol{\varepsilon}}_t \tag{2}$$

Here $\widetilde{\mathbf{y}}_t = \left(y_{2t}^0, \ldots, y_{Nt}^0\right)'$, $\widetilde{\boldsymbol{\alpha}} = (\alpha_2, \ldots, \alpha_N)'$, $\widetilde{\boldsymbol{\varepsilon}}_t = (\varepsilon_{2t}, \ldots, \varepsilon_{Nt})'$, and $\widetilde{\mathbf{B}} = (b_2, \ldots, b_N)'$ is an $\{(N-1) \times r\}$ factor loading matrix. We can back out the information contained in $\mathbf{f}_t$ by multiplying (2) by $\widetilde{\mathbf{B}}'$ and solving for $\mathbf{f}_t$:

$$\mathbf{f}_t = \left(\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}}\right)^{-1} \widetilde{\mathbf{B}}' \left(\widetilde{\mathbf{y}}_t - \widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\varepsilon}}_t\right) \tag{3}$$

Substituting (3) into (1) for $i = 1$, we get

$$y_{1t}^0 = \gamma_1 + \boldsymbol{\gamma}'\widetilde{\mathbf{y}}_t + \varepsilon_{1t}^* \tag{4}$$

where $\boldsymbol{\gamma}' = \mathbf{b}_1' \left(\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}}\right)^{-1} \widetilde{\mathbf{B}}'$, $\gamma_1 = \alpha_1 - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\alpha}}$, and $\varepsilon_{1t}^* = \varepsilon_{1t} - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\varepsilon}}_t$. This solution was first obtained by Li and Bell (2017). However, because $\widetilde{\mathbf{y}}_t$ is correlated with ε_{1t}^*, (4)

cannot be estimated consistently. Nevertheless, a linear projection solves the problem. Specifically, we can decompose ε_{1t}^* into $\varepsilon_{1t}^* = c_1 + \mathbf{c}'\widetilde{\mathbf{y}}_t + v_{1t}$, where $\widetilde{\mathbf{y}}_t$ is orthogonal to v_{1t} by design. Plugging this decomposition into (4), we end up with

$$y_{1t}^0 = \delta_1 + \boldsymbol{\delta}'\widetilde{\mathbf{y}}_t + v_{1t}$$

where $\delta_1 = \gamma_1 + c_1$ and $\boldsymbol{\delta} = \boldsymbol{\gamma} + \mathbf{c}$. Hsiao, Ching, and Wan (2012) advocates estimating $\widehat{\delta}_1$ and $\widehat{\boldsymbol{\delta}}'$ by OLS for $t = 1, \ldots, T_0$ and predicting the counterfactual outcomes by $\widehat{y}_{1t}^0 = \widehat{\delta}_1 + \widehat{\boldsymbol{\delta}}'\widetilde{\mathbf{y}}_t$ for $t = T_0+1, \ldots, T$. Clearly, a reasonably large number of pretreatment periods (for example, $T_0 \geq 20$ or more) are usually needed for trustworthy estimation and subsequent counterfactual prediction. However, a complication is that when the number of potential controls (that is, the number of regressors) is large relative to the number of pretreatment periods (that is, the sample size), OLS estimation of the above equation may be overfitting and thus compromise its ability for out-of-sample prediction during the posttreatment periods. As a solution, Hsiao, Ching, and Wan (2012) propose the best subset approach for model selection and use information criteria for regularization, such as the AIC or the AICc. See more details in section 3.

However, when the number of control units is large, the best subset approach is often too time consuming and may not be feasible in the high-dimensional setting, where the number of control units is larger than the number of pretreatment periods. To overcome this issue, Li and Bell (2017) and Carvalho, Masini, and Medeiros (2018) suggest using lasso for model selection, while Hsiao and Zhou (2019) and Shi and Huang (Forthcoming) propose using forward stepwise regression for model selection. All of these approaches for model selection are implemented in the `rcm` command, in addition to backward stepwise regression, which is provided as another convenient method for model selection.

2.2 The model with covariates

Hsiao and Zhou (2019) introduce covariates into the RCM to further improve the performance of counterfactual prediction. The above pure linear factor model (1) can be augmented by including K observable variables $\mathbf{x}_{it} = (x_{it,1}, \ldots, x_{it,K})'$ into the data-generating process of y_{it}^0:

$$y_{it}^0 = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{b}_i'\mathbf{f}_t + \varepsilon_{it}$$

$\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown parameters. Using a similar approach to eliminate the common factors $\mathbf{f}_t$ as above, we end up with

$$y_{1t}^0 = \delta_1 + \boldsymbol{\delta}'\mathbf{z}_t + v_{1t}$$

where $\mathbf{z}_t = (y_{2t}, \ldots, y_{Nt}, \mathbf{x}_{1t}', \ldots, \mathbf{x}_{Nt}')'$ is a $\{(KN + N - 1) \times 1\}$ vector that is used to predict y_{1t}^0. The counterfactual outcomes during the posttreatment periods are predicted by $\widehat{y}_{1t}^0 = \widehat{\delta}_1 + \widehat{\boldsymbol{\delta}}'\mathbf{z}_t$, where $\widehat{\delta}_1$ and $\widehat{\boldsymbol{\delta}}'$ are obtained by OLS or lasso. It is clear that the RCM model without covariates is a special case of the model with covariates, where $\mathbf{z}_t = \widetilde{\mathbf{y}}_t$ for the case without covariates. Therefore, in the following discussion, we always use $\mathbf{z}_t$ to denote predictors (regressors) in the RCM model.

3 Model selection and estimation

A crucial step of the RCM is the selection of predictors (regressors) to be included in the model. Suppose there are a total of P predictors available (not including the constant term). For the case without covariates, $P = N - 1$ (the number of control units in the donor pool). On the other hand, for the case with covariates, $P = KN + N - 1$, where K is the number of covariates.

Model selection is a process to choose the optimal number of predictors p^*, where $0 \leq p^* \leq P$. Conceptually, this proceeds in two stages. In the first stage, given a specific p with $0 \leq p \leq P$, a best model is chosen by maximizing the R^2 or minimizing the sum of squared residuals among all models with p predictors. The resulting best model conditioning on having p predictors is called the "suboptimal model". In the second stage, the optimal model with p^* predictors is chosen among all suboptimal models using an information criterion. However, if lasso is used for model selection, the role of p is replaced by the tuning parameter λ, also known as the penalty parameter, while the general process of model selection is still similar. See details below.

3.1 Select the suboptimal models

3.1.1 The best subset regression

Given a specific p with $0 \leq p \leq P$, the best subset regression fits OLS regressions for all possible models containing p predictors and finds the suboptimal model with the largest R^2 or the smallest sum of squared residuals. Because there are 2^P possible combinations of predictors to be considered, it is often very time consuming when P is large. A better approach is the leaps and bounds algorithm by Furnival and Wilson (1974) and later improved by Narendra and Fukunaga (1977) and Ni and Huo (2006), which is implemented in the `rcm` command. The leaps and bounds algorithm not only reduces the amount of computation in examining a subset but also finds the best subset without examining all possible subsets, which greatly speeds up the best subset approach.

However, if the number of available predictors P is large, the best subset approach with the leaps and bounds algorithm may still be slow. In that case, one may try forward stepwise regression, backward stepwise regression, or lasso for model selection.

3.1.2 Forward stepwise regression

Forward stepwise regression, advocated by Hsiao and Zhou (2019) and Shi and Huang (Forthcoming), is a computationally efficient alternative to best subset regression. It starts with the smallest model containing no predictors and includes an additional predictor at a time, where the additional predictor to be included is chosen such that it yields the highest R^2 or the smallest sum of squared residuals. In this way, a series of suboptimal models is obtained after P iterations.

Forward stepwise regression computes a total of $1+\sum_{p=0}^{P-1}(P-p) = 1+\left(P^2+P\right)/2$ models, which is much smaller than the total of 2^P models for the best subset regression. Moreover, in the high-dimensional setting, where the number of predictors exceeds the number of pretreatment periods, forward stepwise regression has a clear advantage over the best subset regression because the latter does not have a unique solution for the suboptimal model when $p > T_0 - 1$, where T_0 is the number of pretreatment periods.

3.1.3 Backward stepwise regression

Backward stepwise regression provides another efficient alternative to the best subset regression but applies only in the case of $P \leq T_0 - 1$. It starts with the largest possible model containing all P predictors and considers dropping one predictor at a time, where the predictor to be dropped is chosen such that it yields the highest R^2 or the smallest sum of squared residuals. In this way, a series of suboptimal models is obtained after P iterations.

Backward stepwise regression computes a total of $1 + \left(P^2+P\right)/2$ models, which is the same as forward stepwise regression. However, backward stepwise regression is only applicable in the case of $P \leq T_0 - 1$ so that the full model can be fit. In contrast, forward stepwise regression may still be used even in the high-dimensional case with $P > T_0 - 1$.

3.1.4 Lasso regression

In the high-dimensional case with $P > T_0 - 1$, one could use lasso (Tibshirani 1996) for model selection, which is a popular method of high-dimensional regression. Lasso includes all P predictors in a single regression while imposing an L_1 penalty on the absolute values of regression coefficients, which shrinks these coefficients toward zero, resulting in a sparse model. Specifically, lasso regression minimizes the mean squared error with penalty as follows:

$$\min_{\delta_1,\boldsymbol{\delta}} \left\{ \sum_{t=1}^{T_0} \left(y_{1t} - \delta_1 - \boldsymbol{\delta}'\mathbf{z}_t\right)^2 + \lambda\|\boldsymbol{\delta}\|_1 \right\}$$

$\lambda \geq 0$ is a tuning (penalty) parameter in the scope of $[\lambda_{\text{gmin}}, \lambda_{\text{gmax}}]$ (the subscripts "gmin" and "gmax" stand for "grid min" and "grid max", respectively), and $\|\boldsymbol{\delta}\|_1$ is the L_1 norm of the coefficient vector $\boldsymbol{\delta}$ (that is, the sum of the absolute values of all its components). There are two extreme cases where $\lambda = 0$ yields the OLS regression and $\lambda \to \infty$ yields a null model with $\boldsymbol{\delta} = \mathbf{0}$. Moreover, a coefficient path can be computed as λ changes. Specifically, suppose the scope $[\lambda_{\text{gmin}}, \lambda_{\text{gmax}}]$ is divided into a grid $\{\lambda_{\text{gmax}}, \lambda_{\text{gmax}-1}, \ldots, \lambda_{\text{gmin}}\}$. For each λ_l in the grid, we compute the lasso coefficients and consider it as the suboptimal model given $\lambda = \lambda_l$. In this way, a series of suboptimal models is obtained for each λ_l in the grid.

3.2 Select the optimal model

In the second stage of model selection, we choose an optimal model from all suboptimal models by an information criterion or cross-validation, the latter of which is available only for lasso.

3.2.1 Information criterion

The `rcm` command provides four information criteria for model selection, that is, AIC, AICc, BIC, and MBIC (Wang, Li, and Leng 2009; Shi and Huang Forthcoming), that are computed as follows:

$$\begin{aligned}
\text{AIC}(p) &= T_0 \ln\left(\frac{\mathbf{e}_0'\mathbf{e}_0}{T_0}\right) + 2(p+2) \\
\text{AICc}(p) &= \text{AIC}(p) + \frac{2(p+2)(p+3)}{T_0-(p+1)-2} \\
\text{BIC}(p) &= T_0 \ln\left(\frac{\mathbf{e}_0'\mathbf{e}_0}{T_0}\right) + (p+2)\ln(T_0) \\
\text{MBIC}(p) &= T_0 \ln\left(\frac{\mathbf{e}_0'\mathbf{e}_0}{T_0}\right) + (p+2)\ln(T_0)\left[\ln\{\ln(p+1)\}\right]
\end{aligned}$$

p is the number of predictors in the suboptimal model, T_0 is the number of pretreatment periods, and $\mathbf{e}_0$ is the OLS or lasso residuals in pretreatment periods (hence, $\mathbf{e}_0'\mathbf{e}_0$ is the sum of squared residuals). Basically, the model with the minimized AIC, AICc, BIC, or MBIC is chosen as the optimal model among all suboptimal models. Hsiao, Ching, and Wan (2012) recommend AICc, which performs better in small samples and is set as the default in the `rcm` command. Also, note that in the high-dimensional case with $P > T_0 - 1$, these information criteria may run into difficulty without a lasso-type penalty. In particular, the sum of squared residuals may be reduced to zero in the high-dimensional case, which makes the above information criteria undefined.

3.2.2 Cross-validation

For lasso regression, it is customary to choose the optimal penalty parameter λ by cross-validation, although information criteria are also available in the `rcm` command. For K-fold cross-validation (for example, $K = 5$ or 10),[1] the data in the pretreatment periods are randomly split into K folds (parts) of approximately equal sizes. The basic idea is to leave out data in fold k, use the rest of the data to predict the outcomes in fold k, and repeat this for all folds $k = 1, \ldots, K$. Specifically, for a given penalty parameter λ, the cross-validation coefficients $\delta_{1,k}$ and $\boldsymbol{\delta}_k$ for fold $k = 1, \ldots, K$ are estimated using all pretreatment data except those data in fold k:

1. Here K is a positive integer such as 5 or 10 chosen by the researcher, not to be confused with the number of covariates K above.

$$\min_{\delta_{1,k},\boldsymbol{\delta}_k} \left\{ \sum_{t=1,t\notin\mathcal{F}_k}^{T_0} \left(y_{1t} - \delta_{1,k} - \boldsymbol{\delta}_k' \mathbf{z}_t\right)^2 + \lambda \left\|\boldsymbol{\delta}_k\right\|_1 \right\}$$

$\mathcal{F}_k$ denotes the time periods in fold k. After obtaining $\widehat{\delta}_{1,k}$ and $\widehat{\boldsymbol{\delta}}_k$ for all $k = 1, \ldots, K$, we choose the optimal penalty parameter λ by minimizing the cross-validation mean squared error (CVMSE) (λ), defined as

$$\min_{\lambda} \text{CVMSE}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \left\{ \frac{1}{n_k} \sum_{t\in\mathcal{F}_k} \left(y_{1t} - \widehat{\delta}_{1,k} - \widehat{\boldsymbol{\delta}}_k' \mathbf{z}_t\right)^2 \right\}$$

where n_k is the number of time periods in fold k. Essentially, the model with minimized CVMSE(λ) is selected as the optimal model among all suboptimal models with λ in the grid $\{\lambda_{\text{gmax}}, \lambda_{\text{gmax}-1}, \ldots, \lambda_{\text{gmin}}\}$. Because the RCM is typically applied to small samples in practice, the default value of K in the `rcm` command is set to be the number of the pretreatment period T_0, which is also known as leave-one-out cross-validation (LOOCV).

3.3 Estimation and prediction

After the step of model selection described above, there are two options for the steps of estimation and prediction. The first option is OLS, which is available after all methods of model selection, including best subset, forward stepwise, backward stepwise, and lasso regression. In particular, if OLS is used for estimation following lasso for model selection, this is known as "postlasso OLS". The second option for estimation is lasso, which is available only after using lasso for model selection.

After one fits the model using the pretreatment data with OLS or lasso, the fitted model is then used to predict the counterfactual outcomes for the posttreatment periods. The treatment effects are simply the differences between the observed outcomes and counterfactual outcomes for the treated unit during the posttreatment periods.

4 Statistical inference via placebo tests

4.1 In-space placebo test

Statistical inference for the RCM is still an unsettled business. Li and Bell (2017) and Shi and Huang (Forthcoming) consider statistical inference for the average treatment effect over the entire posttreatment period, which requires many posttreatment periods. For pointwise inference, Chen, Xiao, and Yao (2022) propose using the quantile random forest (that is, quantile regression via random forest) to construct robust nonparametric confidence intervals for treatment effects and demonstrate their excellent properties even in small samples. In the `rcm` command, we focus on the popular method of inference via placebo tests, which have been proposed by Abadie, Diamond, and Hainmueller (2010, 2015) for the SCM but are equally applicable to the RCM.

The placebo tests for the RCM come in two forms: in-space and in-time placebo tests. The in-space placebo test uses "fake treatment units", while the in-time placebo test uses a "fake treatment time" (see details below for the latter). Specifically, the in-space placebo test compares the estimated treatment effects with a distribution of placebo effects obtained by iteratively assigning the treatment to control units in the donor pool (that is, fake treatment units) and estimating placebo effects in each iteration. As a technical detail, we may require the fake treatment units to have a pretreatment mean squared prediction error (MSPE, which is the same as mean squared error) not too much larger (say, 5 or 20 times more) than that of the treated unit because there is not much information contained in fake treatment units with poor pretreatment fits (Abadie, Diamond, and Hainmueller 2010).

If the treatment effects are "unusually extreme" (unusually large, small, or large in absolute value) relative to the distribution of placebo effects, then the treatment effects are considered significant. Otherwise, if the treatment effects are not extreme relative to the distribution of placebo effects, then we accept the null hypothesis of no treatment effects. Depending on how one measures unusual extremeness, the `rcm` command computes right-sided p-values (for "unusually large"), left-sided p-values (for "unusually small", for example, negative numbers with large absolute values), and two-sided p-values (for "unusually large in absolute values") for each posttreatment period. Conducting hypothesis tests using one-sided p-values (including right-sided and left-sided p-values) generally has more power than using two-sided p-values. If the treatment effects are mostly positive, then one should use right-sided p-values, whereas left-sided p-values are recommended for mostly negative treatment effects.

Specifically, the two-sided p-value for a particular posttreatment period t is defined as the frequency that the absolute values of the placebo effects are greater than or equal to the absolute value of the estimated treatment effect:

$$\text{two-sided } p\text{-value}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\left|\widehat{\Delta}_{it}\right| \geq \left|\widehat{\Delta}_{1t}\right|\right), \quad t = T_0+1,\ldots,T$$

$\widehat{\Delta}_{it}$ is the estimated treatment (placebo) effect for unit i in period t (that is, $\widehat{\Delta}_{1t}$ is the treatment effect, whereas $\widehat{\Delta}_{it}$ is the placebo effect for unit $i \neq 1$); and $\mathbf{1}(\cdot)$ is the indicator function, which equals 1 if the expression inside is true and 0 otherwise. Similarly, the right-sided and left-sided p-values are defined as follows:

$$\text{right-sided } p\text{-value}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\widehat{\Delta}_{it} \geq \widehat{\Delta}_{1t}\right) \quad t = T_0+1,\ldots,T$$

$$\text{left-sided } p\text{-value}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\widehat{\Delta}_{it} \leq \widehat{\Delta}_{1t}\right) \quad t = T_0+1,\ldots,T$$

The above p-values measure pointwise significance of the treatment effects. As an overall measure of the significance of treatment effects over the entire posttreatment

periods, we can compare the ratio of posttreatment MSPEs to pretreatment MSPEs (denoted as "post/pre MSPE ratio" for short) for the treated unit with a placebo distribution of this ratio obtained by the above in-space placebo test. Intuitively, if the post/pre MSPE ratio for the treated unit is unusually large relative to the placebo distribution of this ratio, then we are more confident that the overall treatment effects are significant. Specifically, the `rcm` command computes the probability (that is, p-value) of obtaining a post/pre MSPE ratio as large as that of the treated unit as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left(\frac{\text{MSPE}_{i,\text{post}}}{\text{MSPE}_{i,\text{pre}}} \geq \frac{\text{MSPE}_{1,\text{post}}}{\text{MSPE}_{1,\text{pre}}}\right)$$

$\text{MSPE}_{i,\text{post}}$ and $\text{MSPE}_{i,\text{pre}}$ are the posttreatment MSPE and the pretreatment MSPE for unit i, respectively. For example, if the post/pre MSPE ratio for the treated unit is larger than all other units, then the corresponding p-value is $1/N$.

4.2 In-time placebo test

In contrast to the in-space placebo test using fake treatment units, the in-time placebo test uses a fake treatment time before the treatment actually starts. Specifically, a fake treatment time in the pretreatment periods is chosen, say, $\widetilde{T}_0 < T_0 + 1$ (the actual treatment starts in $T_0 + 1$). We then assign the treatment to periods from $\widetilde{T}_0$ on, where no treatment actually occurred during the periods $\left[\widetilde{T}_0, T_0\right]$.

The intuition is that, if the estimated placebo effects during the periods $\left[\widetilde{T}_0, T_0\right]$ turn out to be "significant" or "large" in some sense, then our confidence in the significance of the actual treatment effects, if any, will be eroded. Note that no p-value is computed for the in-time placebo test and one typically uses a graph to present the results from an in-time placebo test. In addition, a researcher can choose multiple fake treatment times and conduct in-time placebo tests for each fake treatment time separately.

5 The rcm command

5.1 Syntax

The syntax for `rcm` is

```
rcm depvar [indepvars], trunit(#) trperiod(#) [ctrlunit(numlist)
   preperiod(numlist) postperiod(numlist) scope(p_min p_max)
   method(sel_method) criterion(sel_criterion) estimate(est_method)
   grid(#g[, ratio(#) min(#)]) fold(#k) seed(int) fill(fill_method)
   placebo([[unit|unit(numlist)] period(numlist) cutoff(#c)])
   frame(framename) nofigure]
```

`xtset` *panelvar timevar* must be used to declare a panel dataset in the usual long form; see [XT] **xtset**. `rcm` automatically reshapes the panel dataset from long to wide form, suitable for implementing RCM.

depvar and *indepvars* must be numeric variables, and abbreviations are not allowed.

5.2 Options

`rcm` automatically reshapes the panel dataset from long to wide form before implementation, where the *depvar* of the treated unit is transformed to be the response and the *depvar* of the control units are transformed to be predictors. If *indepvars* are specified, the *indepvars* of all units are transformed to be predictors during this process.

`trunit(#)` specifies the unit number of the treated unit (that is, the unit affected by the intervention) as given in the panel variable specified in `xtset` *panelvar*. Note that only a single unit number can be specified. `trunit()` is required.

`trperiod(#)` specifies the time period when the intervention occurred. The time period refers to the time variable specified in `xtset` *timevar* and must be an integer (see examples below). Note that only a single time period can be specified. `trperiod()` is required.

The model selection consists of two steps that `rcm` performs automatically. Understanding the steps is helpful for specifying options.

- Step 1: Select the suboptimal models

  `rcm` selects a series of suboptimal models; each contains a unique subset of predictors. The exact procedure for selecting the suboptimal model depends on the selection method specified by `method()`. Available selection methods include best subset, lasso, and forward stepwise and backward stepwise regression; see below for details.

- Step 2: Select the optimal model from all suboptimal models

  `rcm` selects the optimal model from all suboptimal models by an information criterion or cross-validation as specified by `criterion()`. The allowable criteria include `aicc`, `aic`, `bic`, `mbic`, and `cv` (only available for `method(lasso)`). By default, there is no restriction on the number of predictors in selecting the optimal model, but the allowable number of predictors can be specified by `scope()` to limit its range.

After model selection, `rcm` uses the optimal model for counterfactual prediction and estimation of treatment effects. `estimate()` specifies the method used to fit the optimal model, and the allowable criteria include `ols` (such as OLS or postlasso OLS) and `lasso` (directly uses lasso for prediction); see details below.

`ctrlunit(`*numlist*`)` specifies a list of unit numbers for the control units as *numlist* given in the panel variable specified in `xtset` *panelvar*. The list of specified control units constitutes what is known as the "donor pool". The donor pool defaults to all available units other than the treated unit.

`preperiod(`*numlist*`)` specifies a list of pretreatment periods as *numlist* given in the time variable specified in `xtset` *timevar*. `preperiod()` defaults to the entire preintervention period, which ranges from the earliest time period available in the time variable to the period immediately prior to the intervention.

`postperiod(`*numlist*`)` specifies a list of posttreatment periods (when and after the intervention occurred) as *numlist* given in the time variable specified in `xtset` *timevar*. `postperiod()` defaults to the entire postintervention period, which ranges from the time period when the intervention occurred to the last time period available in the time variable.

`scope(`*p_min p_max*`)` specifies the allowable range for the number of predictors in the optimal model. `rcm` selects the optimal model from the suboptimal models containing *p_min* to *p_max* predictors. *p_min* and *p_max* are two numbers that specify the lower and upper bounds of the number of predictors, and the defaults are 1 and the number of all predictors, respectively. If there is no model with the number of predictors in the specified range, *p_min* and *p_max* are automatically changed to the default to expand the selection.

`method(`*sel_method*`)` specifies the method used for selecting the suboptimal model. *sel_method* may be `best` (the default), `lasso`, `forward`, or `backward`.

`best` (best subset regression) is the default, which considers different numbers of predictors in each iteration of OLS estimation, and selects the suboptimal model with the highest R^2 for each specified number of predictors. We use the "leaps and bounds" algorithm (Furnival and Wilson 1974) to speed up the process of best subset regression. Nevertheless, it may still be too time consuming when there are many predictors or more predictors than the number of pretreatment periods. In that case, you may wish to try `method(lasso)` (recommended), `method(forward)`, or `method(backward)`. Alternatively, you may restrict *indepvars* or the donor pool by the option `ctrlunit()`.

`lasso` (lasso regression) sets a grid for λ (known as the tuning or penalty parameter) and fits the corresponding lasso regressions on that grid as the suboptimal models. Specifically, λ iterates from λ_{gmax} to λ_{gmin}; see [LASSO] **lasso**.

`forward` (forward stepwise regression) starts with the smallest model, adds a predictor in each iteration of OLS estimation, and selects the model with the highest R^2 as the suboptimal model for each iteration. If `method(best)` is feasible, then `method(forward)` is not recommended.

`backward` (backward stepwise regression) starts with the largest possible model, drops a predictor in each iteration of OLS estimation, and selects the model with the highest R^2 as the suboptimal model for each iteration. If `method(best)` is feasible, `method(backward)` is not recommended. Note that `method(backward)` is not applicable in the high-dimensional case where the number of predictors exceeds the number of pretreatment periods.

`criterion(`*sel_criterion*`)` specifies the criterion for selecting the optimal model from all suboptimal models, which may be `aicc` (the default), `aic`, `bic`, `mbic`, or `cv`.

`aicc`, the default, specifies the AICc as the criterion for selecting the optimal model; see Hsiao, Ching, and Wan (2012) for details.

`aic` specifies AIC as the selection criterion.

`bic` specifies BIC as the selection criterion.

`mbic` specifies MBIC as the selection criterion; see Wang, Li, and Leng (2009) and Shi and Huang (Forthcoming) for details.

`cv` specifies CVMSE as the selection criterion. Note that `criterion(cv)` applies only to `method(lasso)`, and the option `fold()` determines the number of folds for cross-validation (see details below).

`estimate(`*est_method*`)` specifies the method used to fit the optimal model for counterfactual prediction, which may be `ols` (the default) or `lasso`.

`ols`, the default, fits the optimal model by either OLS or postlasso OLS, whichever is applicable. The latter corresponds to the combination of `method(lasso)` and `estimate(ols)`.

`lasso` directly uses lasso to fit the optimal model for counterfactual prediction. Note that `estimate(lasso)` applies only to `method(lasso)`.

`grid(`$\#_g$`[ , ratio(#) min(#) ])` is a rarely used option specifying the set of possible lambdas with $\#_g$ grid points, where `ratio()` specifies $\lambda_{\text{gmin}}/\lambda_{\text{gmax}}$ and `min()` specifies λ_{gmin}. These parameters are transmitted to the Stata command `lasso`; see [LASSO] **lasso** for details. Note that this option applies only to `method(lasso)`.

`fold(`$\#_k$`)` specifies cross-validation with $\#_k$ folds, where $\#_k$ must be an integer $\geq$ 3 and $\leq T_0$ (the number of pretreatment periods). This option applies only to the combination of `method(lasso)` and `criterion(cv)`. The default is `fold(`T_0`)`, which corresponds to LOOCV.

`seed(`*int*`)` specifies the seed used by the random-number generator for reproducible results. The default is `seed(1)`. This option is useful only for `criterion(cv)`.

`fill(`*fill_method*`)` is a rarely used option that specifies the method to fill in missing values. If `fill(mean)` is specified, missing values are replaced by sample means for each unit. If `fill(linear)` is specified, then missing values are replaced by linear interpolation for each unit. Beware that these two methods for filling in missing values are rough and provided only for convenience. By default, missing values are left unchanged.

Note that `rcm` generally allows for missing values in the pretreatment periods, although it may be difficult to perform cross-validation for lasso. However, if the selected predictors include missing values in the posttreatment periods, then there will be missing values in the counterfactual predictions and treatment effects as well.

`placebo(`[[`unit` | `unit(`*numlist*`)`] `period(`*numlist*`) cutoff(`$\#_c$`)`]`)` specifies the placebo tests to be performed; otherwise, no placebo test will be implemented.

`unit` and `unit(`*numlist*`)` specify placebo tests using fake treatment units in the donor pool, where `unit` uses all fake treatment units and `unit(`*numlist*`)` uses a list of fake treatment units specified by *numlist*. These two options iteratively assign the treatment to control units where no intervention actually occurred and calculate the p-value of the treatment effect. Note that only one of `unit` and `unit()` can be specified.

`period(`*numlist*`)` specifies placebo tests using fake treatment times. This option assigns the treatment to time periods previous to the intervention, when no treatment actually occurred.

`cutoff(`$\#_c$`)` specifies a cutoff threshold that discards fake treatment units with pretreatment MSPE $\#_c$ times larger than that of the treated unit, where $\#_c$ must be a real number greater than or equal to 1. This option applies only when `unit` or `unit()` is specified. By default, no fake treatment units are discarded.

`frame(`*framename*`)` creates a frame storing generated variables in wide form, including counterfactual predictions, treatment effects, and results from placebo tests if implemented. The frame named *framename* is replaced if it already exists or is created if not.

`nofigure` specifies to not display figures. The default is to display all figures for estimation results and placebo tests if available.

5.3 Stored results

`rcm` stores the following in `e()`:

Scalars

`e(T)`	number of observations in the dataset in wide form
`e(T0)`	number of observations in the pretreatment periods with the dataset in wide form
`e(T1)`	number of observations in the posttreatment periods with the dataset in wide form
`e(K_preds_all)`	number of all predictors
`e(K_preds_sel)`	number of predictors selected for the optimal model
`e(aicc)`	AICc of the optimal model fit in the pretreatment periods
`e(aic)`	AIC of the optimal model fit in the pretreatment periods
`e(bic)`	BIC of the optimal model fit in the pretreatment periods
`e(mbic)`	MBIC of the optimal model fit in the pretreatment periods
`e(cvmse)`	CVMSE of the optimal model fit in the pretreatment periods
`e(mae)`	mean absolute error of the model fit in the pretreatment periods
`e(mse)`	mean squared error of the model fit in the pretreatment periods
`e(rmse)`	root mean squared error of the model fit in the pretreatment periods
`e(r2)`	R^2 of the model fit in the pretreatment periods

Macros

`e(panelvar)`	name of the panel variable
`e(timevar)`	name of the time variable
`e(varlist)`	names of the dependent variable and independent variables
`e(respo)`	name of the response
`e(preds_all)`	names of all predictors
`e(preds_sel)`	names of the predictors selected for the optimal model
`e(unit_all)`	all units
`e(unit_tr)`	treatment unit
`e(unit_ctrl)`	control units
`e(time_all)`	entire periods
`e(time_tr)`	treatment period
`e(time_pre)`	pretreatment periods
`e(time_post)`	posttreatment periods
`e(regcmd)`	`regress`
`e(regcmdline)`	regression command of the optimal model
`e(scope)`	allowable range for the number of predictors to be selected
`e(method)`	method for selecting the suboptimal models
`e(criterion)`	criterion for selecting the optimal model from all suboptimal models
`e(estimate)`	method for fitting the optimal model for counterfactual predictions
`e(seed)`	seed used by the random-number generator for reproducible results
`e(frame)`	name of frame storing generated variables in wide form
`e(properties)`	`b V`

Matrices

`e(b)`	coefficient vector of the optimal model fit in the pretreatment periods
`e(V)`	variance–covariance matrix of the coefficient estimators of the optimal model fit in the pretreatment periods
`e(info)`	matrix containing information of the suboptimal models
`e(mspe)`	matrix containing pretreatment MSPE, posttreatment MSPE, ratios of posttreatment MSPE to pretreatment MSPE, and ratios of pretreatment MSPE of control units to that of the treatment unit
`e(pval)`	matrix containing estimated "treatment effects" and p-values from placebo tests using fake treatment units

6 Examples

6.1 Example 1: Political and economic integration between Hong Kong and mainland China (Hsiao, Ching, and Wan 2012)

In this example, we replicate the results in Hsiao, Ching, and Wan (2012), who evaluate the effects of political and economic integration between Hong Kong and mainland China on the economy of Hong Kong. `growth.dta` is attached to the `rcm` command and contains information on the quarterly real gross domestic product (GDP) growth rates of Hong Kong and 24 other countries or regions from 1993q1 to 2008q1.

After loading `growth.dta` and declaring it a panel dataset by `xtset region time`, we use the command `label list` to find the unit number for the treated unit Hong Kong:

```
. use growth
. xtset region time
Panel variable: region (strongly balanced)
 Time variable: time, 1993q1 to 2008q1
         Delta: 1 quarter
. label list
region:
           1 Australia
           2 Austria
           3 Canada
           4 China
           5 Denmark
           6 Finland
           7 France
           8 Germany
           9 HongKong
          10 Indonesia
          11 Italy
          12 Japan
          13 Korea
          14 Malaysia
          15 Mexico
          16 Netherlands
          17 NewZealand
          18 Norway
          19 Philippines
          20 Singapore
          21 Switzerland
          22 Taiwan
          23 Thailand
          24 UnitedKingdom
          25 UnitedStates
```

The results show that the unit number for Hong Kong is 9. Hence, we shall use the option `trunit(9)` to specify Hong Kong as the treated unit.

We first consider the case of political integration between Hong Kong and mainland China, which happened on July 1, 1997 (that is, 1997q3), when the sovereignty of Hong Kong was reverted from the United Kingdom to China. To find the treatment period, we need to convert `1997q3` into a numeric value, which is accomplished by the following command:

```
. display tq(1997q3)
150
```

`tq()` means "time in the quarterly format". Basically, the result shows that 1997q3 is the 150th quarter since 1960q1, according to the convention of Stata. Thus, we shall use the option `trperiod(150)` to specify 1997q3 as the treatment period.

Following Hsiao, Ching, and Wan (2012), we restrict the donor pool to be the following 10 countries or regions, that is, China, Indonesia, Japan, Korea, Malaysia, Philippines, Singapore, Taiwan, Thailand, and the United States, which are either geographically or economically closely associated with Hong Kong. The unit numbers for these 10 regions can be obtained from the above results following the command `label list`. Therefore, we shall use the option `ctrlunit(4 10 12 13 14 19 20 22 23 25)` to specify the donor pool.

Again following Hsiao, Ching, and Wan (2012), we restrict the posttreatment periods to end in 2003q4 because the economic integration between Hong Kong and mainland China happened in 2004q1, when the Closer Economic Partnership Arrangement between the two parties went into effect. To find the numeric value for `2003q4`, we use the following command:

```
. display tq(2003q4)
175
```

Hence, we shall use the option `postperiod(150/175)` to specify the posttreatment periods from 1997q3 to 2003q4. After collecting all the above information, we can use the `rcm` command to replicate the results of Hsiao, Ching, and Wan (2012) for the case of political integration with the default setting of model selection by best subset with the AICc:[2]

```
. rcm gdp, trunit(9) trperiod(150) ctrlunit(4 10 12 13 14 19 20 22 23 25)
> postperiod(150/175)

Step 1: Select the suboptimal models
(method best specified)
Note: If this takes too long, you may wish to try method(lasso)(recommended),
      method(forward) or method(backward). Alternatively, you may restrict
      indepvars, and/or the donor pool by the option ctrlunit().

Selecting the suboptimal model with number of predictors 1-10...

Step 2: Select the optimal model from the suboptimal models
(criterion aicc specified)

Comparing the suboptimal models containing different set of predictors:
```

K	AICc	AIC	BIC	MBIC	R-squared
1	-144.7514	-146.4657	-143.7946	-155.6437	0.4034
2	-160.5063	-163.5832	-160.0217	-170.4959	0.7937
3	-170.6492	-175.6492	-171.1973	-180.9287	0.9056
4	-171.7725	-179.4088	-174.0666	-183.1559	0.9314
5	-169.7878	-180.9878	-174.7552	-183.1882	0.9438
6	-164.2937	-180.2937	-173.1707	-180.9000	0.9477
7	-156.6834	-179.1834	-171.1701	-178.1391	0.9503
8	-146.2921	-177.7207	-168.8169	-174.9678	0.9517
9	-131.7464	-175.7464	-165.9523	-171.2291	0.9518
10	-111.3603	-173.7603	-163.0758	-167.4256	0.9518

```
Among models with 1-10 predictors, the optimal model contains 4 predictors
with AICc = -171.7725.

Fitting results in the pre-treatment periods using OLS:
```

Mean Absolute Error	= 0.00611	Number of Observations	= 18
Mean Squared Error	= 0.00003	Number of Predictors	= 4
Root Mean Squared Error	= 0.00578	R-squared	= 0.93144

gdp·HongKong	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Korea	-0.4323	0.0634	-6.82	0.000	-0.5692	-0.2954
gdp·Japan	-0.6760	0.1117	-6.05	0.000	-0.9172	-0.4347
gdp·Taiwan	0.7926	0.3099	2.56	0.024	0.1231	1.4621
gdp·UnitedS_s	0.4860	0.2195	2.21	0.045	0.0118	0.9603
_cons	0.0263	0.0170	1.54	0.147	-0.0105	0.0631

2. If there are many covariates, using the best subset approach (or even the forward and backward stepwise regressions) for model selection can be computationally demanding. One possibility is to use sure independence screening to speed up the process of selecting important variables (Fan and Lv 2008). We leave this as a possible direction for future research.

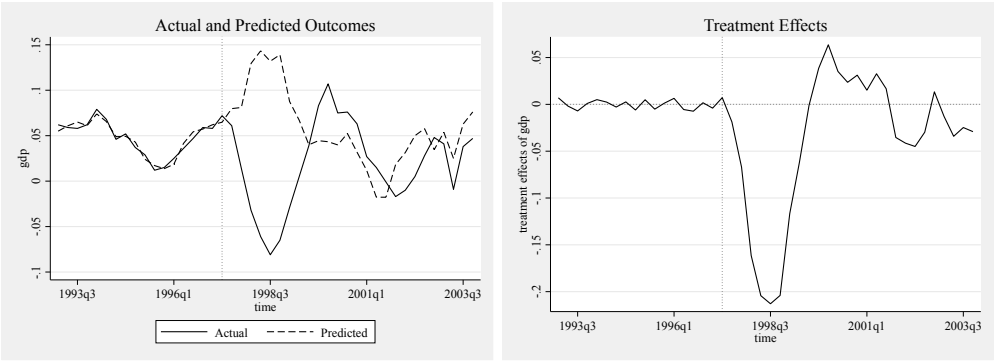
```
Prediction results in the post-treatment periods using OLS:
```

Time	Actual Outcome	Predicted Outcome	Treatment Effect
1997q3	0.0610	0.0798	-0.0188
1997q4	0.0140	0.0810	-0.0670
1998q1	-0.0320	0.1294	-0.1614
1998q2	-0.0610	0.1433	-0.2043
1998q3	-0.0810	0.1319	-0.2129
1998q4	-0.0650	0.1390	-0.2040
1999q1	-0.0290	0.0876	-0.1166
1999q2	0.0050	0.0670	-0.0620
1999q3	0.0390	0.0400	-0.0010
1999q4	0.0830	0.0445	0.0385
2000q1	0.1070	0.0434	0.0636
2000q2	0.0750	0.0398	0.0352
2000q3	0.0760	0.0524	0.0236
2000q4	0.0630	0.0318	0.0312
2001q1	0.0270	0.0118	0.0152
2001q2	0.0150	-0.0177	0.0327
2001q3	-0.0010	-0.0177	0.0167
2001q4	-0.0170	0.0184	-0.0354
2002q1	-0.0100	0.0314	-0.0414
2002q2	0.0050	0.0500	-0.0450
2002q3	0.0280	0.0577	-0.0297
2002q4	0.0480	0.0346	0.0134
2003q1	0.0410	0.0538	-0.0128
2003q2	-0.0090	0.0251	-0.0341
2003q3	0.0380	0.0628	-0.0248
2003q4	0.0470	0.0761	-0.0291
Mean	0.0180	0.0576	-0.0396

```
Note: The average treatment effect over the post-treatment periods is -0.0396.
Finished.
```

The results show that the optimal model contains four predictors, which are the GDPs for Korea, Japan, Taiwan, and the United States. The pretreatment R^2 is 0.93144, indicating a good pretreatment fit. The predicted outcomes and treatment effects are also reported for each posttreatment period.

In the meantime, the above `rcm` command produces the following two graphs. The first graph [figure 1(a)] depicts the actual and counterfactual outcomes, also known as the "gap graph"; the second graph [figure 1(b)] presents a visualization of the treatment effects.



(a) Actual and predicted paths (b) Treatment effects

Figure 1. Graphs for political integration in example 1

Notice that the dotted vertical lines in figure 1 are drawn at the last pretreatment period (that is, at T_0) instead of the first posttreatment period (that is, at T_0+1) for better visual appearance. Also note that if we were to replicate the results in Hsiao, Ching, and Wan (2012) using the AIC, we could simply include the option `criterion(aic)` instead of using the default option, `criterion(aicc)`. The results are omitted to save space.

Next we consider the case of economic integration between Hong Kong and mainland China, that is, the implementation of Closer Economic Partnership Arrangement starting from 2004q1. Following Hsiao, Ching, and Wan (2012), this time we place no restriction on the donor pool or posttreatment periods, and the treatment period can be obtained by

```
. display tq(2004q1)
176
```

We could replicate the results in Hsiao, Ching, and Wan (2012) using the best subset regression and the AICc with the command

```
. rcm gdp, trunit(9) trperiod(176) nofigure frame(growth_wide)
```

where the option `nofigure` suppresses the default production of figures and the option `frame(growth_wide)` creates a frame called `growth_wide` that stores generated variables (including counterfactual outcomes and treatment effects) in wide form such that users may find them useful later on (for example, to draw their own figures).

```
Step 1: Select the suboptimal models
(method best specified)
Note: If this takes too long, you may wish to try method(lasso)(recommended),
      method(forward) or method(backward). Alternatively, you may restrict
      indepvars, and/or the donor pool by the option ctrlunit().

Selecting the suboptimal model with number of predictors 1-24...

Step 2: Select the optimal model from the suboptimal models
(criterion aicc specified)

Comparing the suboptimal models containing different set of predictors:
```

K	AICc	AIC	BIC	MBIC	R-squared
1	-313.8269	-314.4269	-309.0743	-324.5878	0.5877
2	-335.2386	-336.2642	-329.1275	-342.8407	0.7602
3	-348.2800	-349.8590	-340.9380	-353.6787	0.8318
4	-365.6420	-367.9122	-357.2071	-369.1072	0.8933
5	-377.4412	-380.5523	-368.0630	-379.1038	0.9235
6	-378.9426	-383.0569	-368.7833	-378.9029	0.9310
7	-378.9074	-384.2016	-368.1439	-377.2679	0.9357
8	-378.5854	-385.2521	-367.4102	-375.4631	0.9400
9	-377.5003	-385.7503	-366.1242	-373.0328	0.9433
10	-375.0098	-385.0744	-363.6641	-369.3589	0.9450
11	-372.4606	-384.5939	-361.3994	-365.8154	0.9469
12	-369.2578	-383.7405	-358.7619	-361.8379	0.9483
13	-365.9158	-383.0586	-356.2958	-357.9747	0.9498
14	-362.5660	-382.7142	-354.1671	-354.3955	0.9516
15	-358.3157	-381.8542	-351.5230	-350.2504	0.9529
16	-353.3736	-380.7336	-348.6182	-345.7974	0.9538
17	-348.1579	-379.8246	-345.9250	-341.5114	0.9549
18	-342.4931	-379.0149	-343.3311	-337.2826	0.9561
19	-335.8492	-377.8492	-340.3812	-332.6578	0.9570
20	-328.0881	-376.2785	-337.0264	-327.5902	0.9574
21	-319.2286	-374.4286	-333.3922	-322.2073	0.9575
22	-309.3373	-372.4952	-329.6747	-316.7067	0.9576
23	-298.3113	-370.5335	-325.9288	-311.1450	0.9576
24	-285.9617	-368.5499	-322.1610	-305.5301	0.9576

```
Among models with 1-24 predictors, the optimal model contains 6 predictors
with AICc = -378.9426.

Fitting results in the pre-treatment periods using OLS:

Mean Absolute Error      =   0.01070      Number of Observations  =         44
Mean Squared Error       =   0.00014      Number of Predictors    =          6
Root Mean Squared Error  =   0.01170      R-squared               =    0.93097
```

gdp·HongKong	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Norway	0.3222	0.0538	5.99	0.000	0.2132	0.4311
gdp·Austria	-1.0115	0.1682	-6.01	0.000	-1.3524	-0.6707
gdp·Korea	0.3447	0.0469	7.35	0.000	0.2497	0.4398
gdp·Mexico	0.3129	0.0510	6.13	0.000	0.2095	0.4162
gdp·Italy	-0.3177	0.1591	-2.00	0.053	-0.6400	0.0046
gdp·Singapore	0.1845	0.0546	3.38	0.002	0.0739	0.2951
_cons	-0.0019	0.0037	-0.52	0.603	-0.0094	0.0056

Prediction results in the post-treatment periods using OLS:

Time	Actual Outcome	Predicted Outcome	Treatment Effect
2004q1	0.0770	0.0493	0.0277
2004q2	0.1200	0.0686	0.0514
2004q3	0.0660	0.0515	0.0145
2004q4	0.0790	0.0446	0.0344
2005q1	0.0620	0.0217	0.0403
2005q2	0.0710	0.0177	0.0533
2005q3	0.0810	0.0333	0.0477
2005q4	0.0690	0.0290	0.0400
2006q1	0.0900	0.0471	0.0429
2006q2	0.0620	0.0417	0.0203
2006q3	0.0640	0.0250	0.0390
2006q4	0.0660	0.0009	0.0651
2007q1	0.0550	-0.0101	0.0651
2007q2	0.0620	0.0092	0.0528
2007q3	0.0680	0.0143	0.0537
2007q4	0.0690	0.0508	0.0182
2008q1	0.0730	0.0538	0.0192
Mean	0.0726	0.0323	0.0403

Note: The average treatment effect over the post-treatment periods is 0.0403.

Finished.

To access the generated frame `growth_wide`, we may use the following command:

```
. frame change growth_wide
. describe
Contains data
 Observations:            61
    Variables:            28
```

Variable name	Storage type	Display format	Value label	Variable label
time	float	%tq		time
gdp·Australia	float	%8.0g		gdp in Australia
gdp·Austria	float	%8.0g		gdp in Austria
gdp·Canada	float	%8.0g		gdp in Canada
gdp·China	float	%8.0g		gdp in China
gdp·Denmark	float	%8.0g		gdp in Denmark
gdp·Finland	float	%8.0g		gdp in Finland
gdp·France	float	%8.0g		gdp in France
gdp·Germany	float	%8.0g		gdp in Germany
gdp·HongKong	float	%8.0g		gdp in HongKong
gdp·Indonesia	float	%8.0g		gdp in Indonesia
gdp·Italy	float	%8.0g		gdp in Italy
gdp·Japan	float	%8.0g		gdp in Japan
gdp·Korea	float	%8.0g		gdp in Korea
gdp·Malaysia	float	%8.0g		gdp in Malaysia
gdp·Mexico	float	%8.0g		gdp in Mexico
gdp·Netherlands	float	%8.0g		gdp in Netherlands
gdp·NewZealand	float	%8.0g		gdp in NewZealand
gdp·Norway	float	%8.0g		gdp in Norway
gdp·Philippines	float	%8.0g		gdp in Philippines
gdp·Singapore	float	%8.0g		gdp in Singapore
gdp·Switzerland	float	%8.0g		gdp in Switzerland
gdp·Taiwan	float	%8.0g		gdp in Taiwan
gdp·Thailand	float	%8.0g		gdp in Thailand
gdp·UnitedKin~m	float	%8.0g		gdp in UnitedKingdom
gdp·UnitedSta~s	float	%8.0g		gdp in UnitedStates
pred·gdp·Hong~g	float	%9.0g		prediction of gdp in HongKong
tr·gdp·HongKong	float	%9.0g		treatment effect of gdp in HongKong

```
Sorted by: time
     Note: Dataset has changed since last saved.
```

It is easy to switch back to the default frame containing `growth.dta` by the following command:

```
. frame change default
```

As a further illustration, below we use lasso with LOOCV for model selection, followed by postlasso OLS for estimation and prediction. In addition, we request both in-space and in-time placebo tests. For the in-time place test, the fake treatment time is chosen to be 2002q1, that is, two years before the actual treatment time in 2004q1:

```
. display tq(2002q1)
168
. rcm gdp, trunit(9) trperiod(176) method(lasso) criterion(cv)
> placebo(unit cutoff(5) period(168))
```

`method(lasso)` specifies lasso for model selection, and `criterion(cv)` uses the cross-validation criterion, which defaults to LOOCV. The option `placebo(unit cutoff(5))` requests the in-space placebo test using all fake treatment units except those with a pretreatment MSPE 5 times larger than that of the treated unit. `placebo(period(168))` conducts the in-time placebo test using 168 (that is, 2002q1) as the fake treatment time. This command returns a wealth of information, starting with model selection, estimation, and prediction.

```
Step 1: Select the suboptimal models
(method lasso specified)
Selecting the suboptimal model...
Step 2: Select the optimal model from the suboptimal models
(criterion cv specified for leave-one-out cross-validation)
Comparing the suboptimal models containing different set of predictors:
```

K	lambda	CVMSE	R-squared	Operation
2	0.0285	0.0017	0.1145	add gdp·Malaysia gdp·Singapore
3	0.0216	0.0013	0.3576	add gdp·Norway
5	0.0197	0.0012	0.4252	add gdp·Korea gdp·Thailand
6	0.0179	0.0011	0.4940	add gdp·Indonesia
7	0.0123	0.0008	0.6725	add gdp·Philippines
8	0.0093	0.0006	0.7449	add gdp·Finland
9	0.0064	0.0005	0.8264	add gdp·Mexico
10	0.0044	0.0004	0.8731	add gdp·Austria
11	0.0037	0.0003	0.8919	add gdp·NewZealand
10	0.0034	0.0003	0.8993	drop gdp·Malaysia
11	0.0028	0.0003	0.9104	add gdp·France
12	0.0018	0.0003	0.9273	add gdp·Italy
11	0.0012	0.0002	0.9346	drop gdp·Finland
11	0.0009	0.0002	0.9367	.
12	0.0008	0.0002	0.9377	add gdp·Canada
14	0.0006	0.0002	0.9388	add gdp·Germany gdp·Switzerland
15	0.0005	0.0003	0.9420	add gdp·China
16	0.0004	0.0003	0.9431	add gdp·Australia
18	0.0003	0.0003	0.9456	add gdp·Japan gdp·UnitedKingdom
19	0.0002	0.0003	0.9485	add gdp·Finland
18	0.0002	0.0003	0.9519	drop gdp·Indonesia
20	0.0001	0.0003	0.9526	add gdp·Taiwan gdp·UnitedStates
21	0.0001	0.0003	0.9531	add gdp·Netherlands
23	0.0001	0.0003	0.9544	add gdp·Denmark gdp·Indonesia
24	0.0000	0.0004	0.9571	add gdp·Malaysia
23	0.0000	0.0004	0.9573	drop gdp·Philippines
24	0.0000	0.0004	0.9575	add gdp·Philippines
24	0.0000	0.0004	0.9576	.

```
Among models with 1-24 predictors, the optimal model contains 11 predictors
with CVMSE = 0.0002.
```

Fitting results in the pre-treatment periods using post-lasso OLS:

Mean Absolute Error	= 0.01163	Number of Observations	= 44
Mean Squared Error	= 0.00014	Number of Predictors	= 11
Root Mean Squared Error	= 0.01177	R-squared	= 0.93955

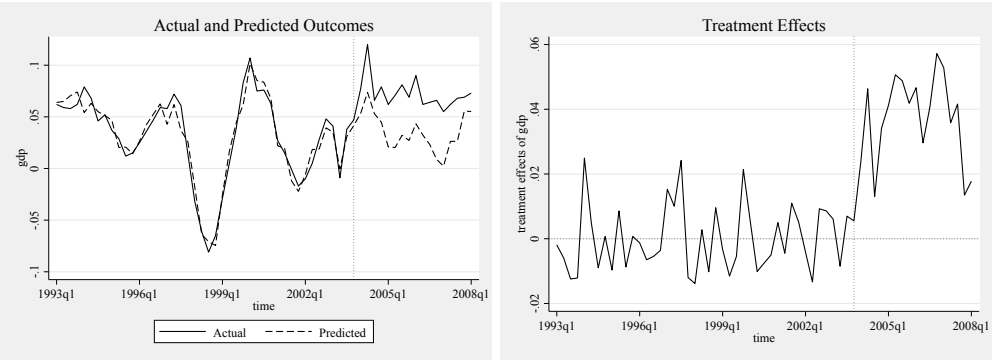
gdp·HongKong	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Austria	-0.8596	0.2734	-3.14	0.004	-1.4166	-0.3027
gdp·France	0.0054	0.3774	0.01	0.989	-0.7633	0.7741
gdp·Indonesia	0.0331	0.0414	0.80	0.430	-0.0512	0.1174
gdp·Italy	-0.3447	0.3225	-1.07	0.293	-1.0016	0.3122
gdp·Korea	0.2501	0.0757	3.30	0.002	0.0958	0.4044
gdp·Mexico	0.2501	0.0746	3.35	0.002	0.0981	0.4021
gdp·NewZeal~d	0.1310	0.1157	1.13	0.266	-0.1047	0.3667
gdp·Norway	0.2535	0.0758	3.34	0.002	0.0991	0.4079
gdp·Philipp~s	0.1540	0.1240	1.24	0.223	-0.0987	0.4066
gdp·Singapore	0.2105	0.0649	3.24	0.003	0.0783	0.3427
gdp·Thailand	0.0232	0.0668	0.35	0.730	-0.1129	0.1594
_cons	-0.0100	0.0064	-1.57	0.126	-0.0230	0.0030

Prediction results in the post-treatment periods using post-lasso OLS:

Time	Actual Outcome	Predicted Outcome	Treatment Effect
2004q1	0.0770	0.0533	0.0237
2004q2	0.1200	0.0737	0.0463
2004q3	0.0660	0.0530	0.0130
2004q4	0.0790	0.0448	0.0342
2005q1	0.0620	0.0208	0.0412
2005q2	0.0710	0.0204	0.0506
2005q3	0.0810	0.0322	0.0488
2005q4	0.0690	0.0271	0.0419
2006q1	0.0900	0.0433	0.0467
2006q2	0.0620	0.0324	0.0296
2006q3	0.0640	0.0233	0.0407
2006q4	0.0660	0.0088	0.0572
2007q1	0.0550	0.0022	0.0528
2007q2	0.0620	0.0262	0.0358
2007q3	0.0680	0.0264	0.0416
2007q4	0.0690	0.0555	0.0135
2008q1	0.0730	0.0552	0.0178
Mean	0.0726	0.0352	0.0374

Note: The average treatment effect over the post-treatment periods is 0.0374.

The above results show that the optimal model chosen by lasso with LOOCV contains 11 predictors with a pretreatment R^2 of 0.93955 for postlasso OLS. The corresponding graphs for counterfactual prediction and treatment effects are shown in figure 2. Apparently, the pretreatment fit is great, while the estimated treatment effects are all positive in the posttreatment periods.



(a) Actual and predicted paths (b) Treatment effects

Figure 2. Graphs for economic integration in example 1

Results are then reported from the in-space placebo test, including the overall measure based on the post/pre MSPE ratio and pointwise p-values for each posttreatment period.

```
Implementing placebo test using fake treatment unit Australia...Austria...Canada
> ...China...Denmark...Finland...France...Germany...Indonesia...Italy...Japan...
> Korea...Malaysia...Mexico...Netherlands...NewZealand...Norway...Philippines...
> Singapore...Switzerland...Taiwan...Thailand...UnitedKingdom...UnitedStates...
Placebo test results using fake treatment units:
```

Unit	Pre MSPE	Post MSPE	Post/Pre MSPE	Pre MSPE of Fake Unit/ Pre MSPE of Treated Unit
HongKong	0.0001	0.0016	11.3444	1.0000
Australia	0.0000	0.0005	20.3885	0.1787
Austria	0.0000	0.0001	7.4453	0.1200
Canada	0.0000	0.0004	24.8804	0.1211
China	0.0001	0.0006	8.3822	0.4917
Denmark	0.0001	0.0009	12.2336	0.5599
Finland	0.0001	0.0004	5.1043	0.5812
France	0.0000	0.0002	13.9916	0.0882
Germany	0.0000	0.0004	17.2181	0.1558
Indonesia	0.0009	0.0074	8.0772	6.5869
Italy	0.0000	0.0001	1.8399	0.2103
Japan	0.0001	0.0009	10.2775	0.6274
Korea	0.0004	0.0035	9.2554	2.7445
Malaysia	0.0004	0.0024	6.4060	2.6866
Mexico	0.0001	0.0010	7.1496	0.9758
Netherlands	0.0000	0.0001	2.4530	0.2757
NewZealand	0.0001	0.0005	3.7067	1.0239
Norway	0.0003	0.0014	4.2723	2.3216
Philippines	0.0002	0.0004	1.6699	1.6748
Singapore	0.0002	0.0024	14.8665	1.1473
Switzerland	0.0000	0.0004	8.4799	0.3053
Taiwan	0.0001	0.0004	4.0779	0.6745
Thailand	0.0003	0.0011	3.6558	2.1797
UnitedKingdom	0.0000	0.0001	7.1206	0.1284
UnitedStates	0.0000	0.0002	16.5412	0.0860

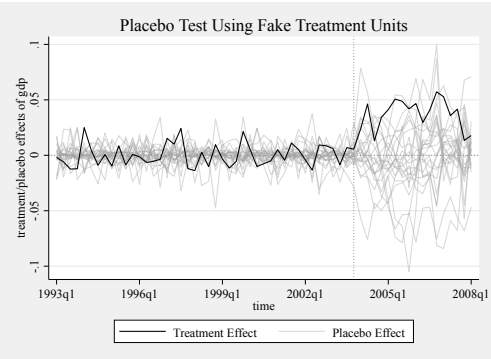
```
Note: (1) The probability of obtaining a post/pre-treatment MSPE ratio as
      large as HongKong's is 0.3200.
      (2) Total 1 unit with pre-treatment MSPE 5 times larger than the treated
      unit is excluded in computing pointwise p-values, including Indonesia.
```

Placebo test results using fake treatment units (continued, cutoff = 5):

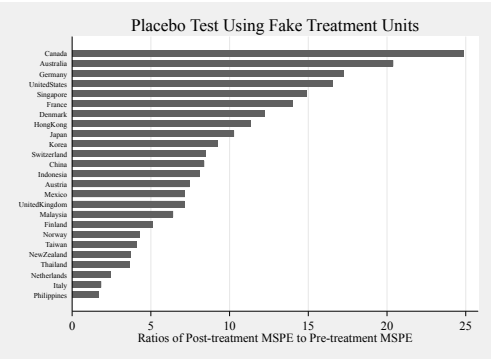
Time	Treatment Effect	p-value of Treatment Effect		
		Two-sided	Right-sided	Left-sided
2004q1	0.0237	0.2917	0.2500	0.7917
2004q2	0.0463	0.1250	0.0833	0.9583
2004q3	0.0130	0.7500	0.3750	0.6667
2004q4	0.0342	0.2083	0.0833	0.9583
2005q1	0.0412	0.1250	0.0417	1.0000
2005q2	0.0506	0.2083	0.0833	0.9583
2005q3	0.0488	0.2083	0.0833	0.9583
2005q4	0.0419	0.2500	0.1250	0.9167
2006q1	0.0467	0.1667	0.0833	0.9583
2006q2	0.0296	0.2917	0.2083	0.8333
2006q3	0.0407	0.1250	0.0833	0.9583
2006q4	0.0572	0.2083	0.1250	0.9167
2007q1	0.0528	0.1250	0.0833	0.9583
2007q2	0.0358	0.2083	0.1250	0.9167
2007q3	0.0416	0.1250	0.0417	1.0000
2007q4	0.0135	0.7500	0.4167	0.6250
2008q1	0.0178	0.1667	0.1250	0.9167

Note: (1) The two-sided p-value of the treatment effect for a particular period is defined as the frequency that the absolute values of the placebo effects are greater than or equal to the absolute value of treatment effect.
(2) The right-sided (left-sided) p-value of the treatment effect for a particular period is defined as the frequency that the placebo effects are greater (smaller) than or equal to the treatment effect.
(3) If the treatment effects are mostly positive, then the right-sided p-values are recommended; whereas the left-sided p-values are recommended if the treatment effects are mostly negative.

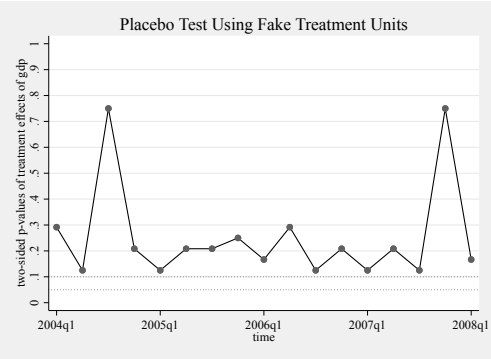
The above results show that the overall p-value based on the post/pre MSPE ratio is 0.3200, which is greater than any conventional significance level. However, if we look at the pointwise right-sided p-values (because the estimated treatment effects are all positive), the right-sided p-values for both 2005q1 and 2007q3 are 0.0417, which are significant at the 5% level. Moreover, the right-sided p-values are 0.0833 for 7 posttreatment periods, which are significant at the 10% level. The above results from the in-space placebo test are presented graphically in figure 3.



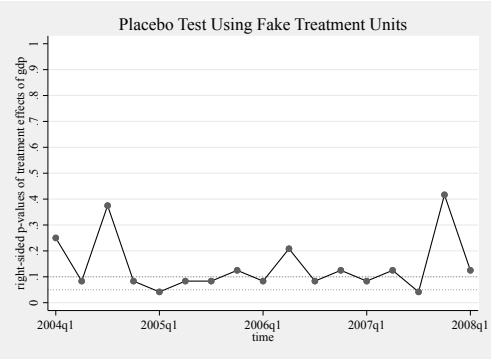
(a) Treatment and placebo effects



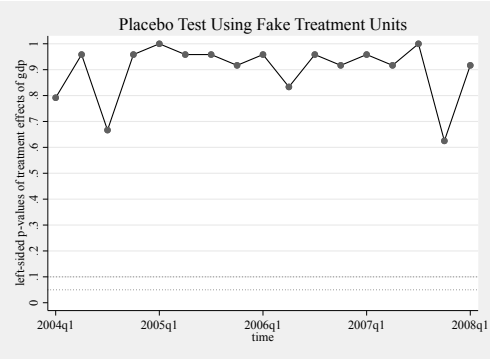
(b) Post/pre MSPE ratios



(c) Two-sided p-values of treatment effects



(d) Right-sided p-values of treatment effects



(e) Left-sided p-values of treatment effects

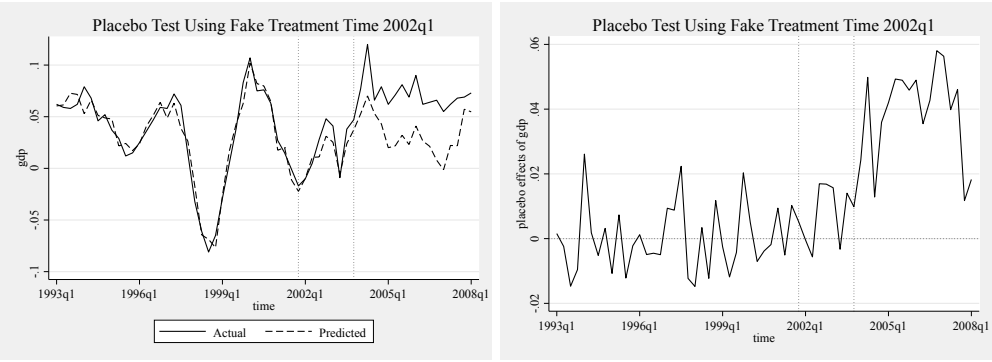
Figure 3. Graphs for in-space placebo test in example 1

Last, the results from the in-time placebo test using 2002q1 as the fake treatment time are reported, as well as the corresponding graphs shown in figure 4.

```
Implementing placebo test using fake treatment time 2002q1...
Placebo test results using fake treatment time 2002q1:
```

Time	Actual Outcome	Predicted Outcome	Treatment Effect
2002q1	-0.0100	-0.0096	-0.0004
2002q2	0.0050	0.0106	-0.0056
2002q3	0.0280	0.0110	0.0170
2002q4	0.0480	0.0312	0.0168
2003q1	0.0410	0.0253	0.0157
2003q2	-0.0090	-0.0057	-0.0033
2003q3	0.0380	0.0240	0.0140
2003q4	0.0470	0.0372	0.0098
2004q1	0.0770	0.0527	0.0243
2004q2	0.1200	0.0702	0.0498
2004q3	0.0660	0.0531	0.0129
2004q4	0.0790	0.0431	0.0359
2005q1	0.0620	0.0200	0.0420
2005q2	0.0710	0.0217	0.0493
2005q3	0.0810	0.0321	0.0489
2005q4	0.0690	0.0231	0.0459
2006q1	0.0900	0.0410	0.0490
2006q2	0.0620	0.0265	0.0355
2006q3	0.0640	0.0213	0.0427
2006q4	0.0660	0.0079	0.0581
2007q1	0.0550	-0.0013	0.0563
2007q2	0.0620	0.0222	0.0398
2007q3	0.0680	0.0219	0.0461
2007q4	0.0690	0.0573	0.0117
2008q1	0.0730	0.0547	0.0183
Mean	0.0569	0.0277	0.0292

```
Note: The average treatment effect over the post-treatment periods is 0.0292.
Finished.
```



(a) Actual and predicted paths

(b) Treatment effects

Figure 4. Graphs for in-time placebo test in example 1

From figure 4, it is apparent that there are no "significant" placebo effects during the "fake posttreatment periods" from 2002q1 to 2004q1, the durations of which are indicated by the two vertical dashed lines in figure 4. This gives us more confidence in the significance of the actual treatment effects, if any.

6.2 Example 2: German reunification (Abadie, Diamond, and Hainmueller 2015)

Because example 1 from Hsiao, Ching, and Wan (2012) contains no covariates, we turn to the case of German reunification in Abadie, Diamond, and Hainmueller (2015) to demonstrate the use of the `rcm` command in the presence of covariates. Abadie, Diamond, and Hainmueller (2015) study the effect of the 1990 German reunification on the economy of West Germany using the SCM. As we shall see, applying the RCM to the same data yields similar results.

`repgermany.dta` is attached to the `rcm` command and includes the following variables for West Germany and 16 other Organisation for Economic Co-operation and Development member countries from 1960 to 2003: the outcome variable `gdp` (GDP per capita) and covariates `infrate` (inflation rate defined as annual percentage change in consumer prices), `trade` (trade openness defined as export plus imports as percentage of GDP), and `industry` (industry share of value added). We ignore other covariates, which contain too many missing values.

After loading `repgermany.dta` and declaring it a panel dataset with `xtset country year`, we use the command `xtsum` to look at summary statistics for the relevant variables:

```
. use repgermany, clear
. xtset country year
Panel variable: country (strongly balanced)
 Time variable: year, 1960 to 2003
         Delta: 1 unit
. xtsum gdp infrate trade industry
```

Variable		Mean	Std. dev.	Min	Max	Observations
gdp	overall	12144.14	8951.553	707	37548	N = 748
	between		2346.311	7267.5	16063.09	n = 17
	within		8656.906	-890.8128	35869.26	T = 44
infrate	overall	5.867715	5.127335	-.9151205	28.78333	N = 727
	between		2.340064	3.117853	10.71496	n = 17
	within		4.598013	-5.226749	24.93645	T = 42.7647
trade	overall	53.12414	26.4594	9.429324	149.6824	N = 646
	between		25.39686	16.7063	113.8836	n = 17
	within		9.020864	16.58236	88.92298	T = 38
industry	overall	33.23844	5.161249	21.59255	48.00126	N = 541
	between		3.427973	27.35454	39.68952	n = 17
	within		4.024088	23.4932	42.88638	T-bar = 31.8235

The above results show that all three covariates have more or less missing values, which may affect the prediction of posttreatment counterfactual outcomes. We use the command `label list` to find the unit number for the treated unit West Germany:

```
. label list
country:
           1 Australia
           2 Austria
           3 Belgium
           4 Denmark
           5 France
           6 Greece
           7 Italy
           8 Japan
           9 Netherlands
          10 New Zealand
          11 Norway
          12 Portugal
          13 Spain
          14 Switzerland
          15 UK
          16 USA
          17 West Germany
```

The results show that the unit number for West Germany is 17. Hence, we shall use the option `trunit(17)` to specify West Germany as the treated unit. To specify the treatment period, we use the option `trperiod(1990)` because the German reunification occurred in 1990. In the presence of three covariates, we have many predictors. There-

fore, we use lasso with 5-fold cross-validation for model selection, followed by postlasso OLS for estimation and prediction:

```
. rcm gdp infrate trade industry, trunit(17) trperiod(1990) method(lasso)
> criterion(cv) fold(5) placebo(unit cutoff(20) period(1980))
```

The options `method(lasso)`, `criterion(cv)`, and `fold(5)` specify model selection by lasso with 5-fold cross-validation. The option `placebo(unit cutoff(20))` requests an in-space placebo test while requiring the pretreatment MSPE of fake treatment units to be no more than 20 times that of the treated unit.[3] The option `placebo(period(1980))` specifies the in-time placebo test with 1980 as the fake treatment period, which is 10 years earlier than the actual treatment period of 1990.

This command returns a rich set of information, starting with model selection, estimation, and prediction. The optimal model contains nine predictors, including `industry·Spain` (industry for Spain), which showcases the value of adding covariates to RCM. The pretreatment fit is excellent, with an R^2 of 0.99999 by postlasso OLS. The corresponding graphs for counterfactual outcomes and treatment effects are shown in figure 5.

```
Step 1: Select the suboptimal models
(method lasso specified)
Selecting the suboptimal model...
Step 2: Select the optimal model from the suboptimal models
(criterion cv specified for 5-fold cross-validation)
Comparing the suboptimal models containing different set of predictors:
```

K	lambda	CVMSE	R-squared	Operation
1	4287.1880	2.178e+07	0.0888	add gdp·Italy
2	2954.9924	1.037e+07	0.5668	add gdp·Netherlands
3	2692.4790	8.618e+06	0.6403	add gdp·Austria
4	1855.8213	4.096e+06	0.8289	add gdp·Denmark
5	1614.0987	3.091e+06	0.8706	add gdp·USA
7	459.7011	2.517e+05	0.9894	add gdp·Greece gdp·Norway
6	218.3953	61772.7118	0.9976	drop gdp·Netherlands
7	150.5314	31060.5191	0.9988	add gdp·Switzerland
8	137.1586	26263.3771	0.9990	add industry·Spain
9	59.3727	7092.8250	0.9998	add gdp·Netherlands
9	44.9133	4923.4104	0.9999	.

```
Among models with 1-67 predictors, the optimal model contains 9 predictors
with CVMSE = 4923.4104.
```

3. Because the donor pool is small, with only 16 control units, we set a higher threshold of pretreatment MSPE to preserve the size of the donor pool.

Fitting results in the pre-treatment periods using post-lasso OLS:

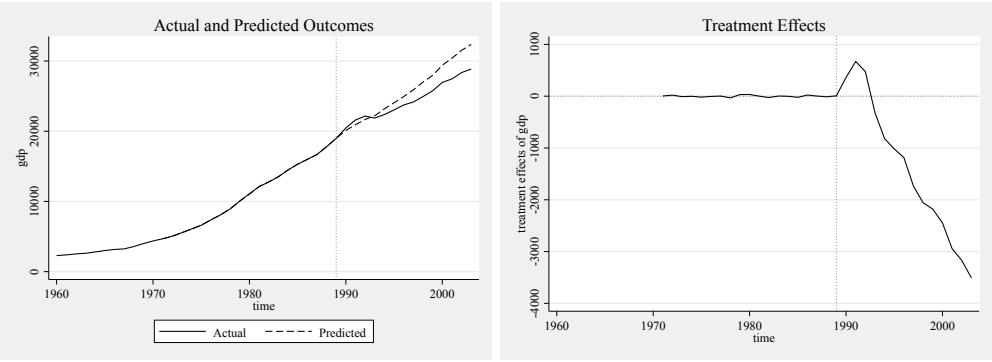
Mean Absolute Error	= 12.87119	Number of Observations	= 30
Mean Squared Error	= 3.0e+02	Number of Predictors	= 9
Root Mean Squared Error	= 17.30368	R-squared	= 0.99999

gdp·WestGer_y	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Austria	0.1331	0.0708	1.88	0.093	-0.0271	0.2934
gdp·Denmark	0.1046	0.0788	1.33	0.217	-0.0735	0.2828
gdp·Greece	0.1513	0.0422	3.58	0.006	0.0557	0.2468
gdp·Italy	0.2914	0.0838	3.48	0.007	0.1018	0.4809
gdp·Netherl_s	0.1334	0.1097	1.22	0.255	-0.1148	0.3816
gdp·Norway	-0.0313	0.0598	-0.52	0.613	-0.1665	0.1039
industry·Sp_n	-44.6926	11.6744	-3.83	0.004	-71.1019	-18.2833
gdp·Switzer_d	0.0548	0.0343	1.60	0.144	-0.0228	0.1324
gdp·USA	0.2282	0.0439	5.20	0.001	0.1289	0.3276
_cons	1755.2090	419.7465	4.18	0.002	805.6764	2704.7417

Prediction results in the post-treatment periods using post-lasso OLS:

Time	Actual Outcome	Predicted Outcome	Treatment Effect
1990	20465.0000	20104.6133	360.3867
1991	21602.0000	20930.3809	671.6191
1992	22154.0000	21677.2500	476.7500
1993	21878.0000	22194.6797	-316.6797
1994	22371.0000	23190.9297	-819.9297
1995	23035.0000	24052.6562	-1017.6562
1996	23742.0000	24926.1309	-1184.1309
1997	24156.0000	25896.0312	-1740.0312
1998	24931.0000	26988.5430	-2057.5430
1999	25755.0000	27935.7734	-2180.7734
2000	26943.0000	29389.8184	-2446.8184
2001	27449.0000	30392.2207	-2943.2207
2002	28348.0000	31518.2871	-3170.2871
2003	28855.0000	32363.5508	-3508.5508
Mean	24406.0000	25825.7761	-1419.7761

Note: The average treatment effect over the post-treatment periods is -1419.7761.



(a) Actual and predicted paths

(b) Treatment effects

Figure 5. Graphs for German reunification in example 2

It is apparent from figure 5 that the treatment effects were positive for only three years (possibly because of a boom in demand following the German reunification) but turned increasingly negative thereafter. In fact, these results are very similar to those originally reported by Abadie, Diamond, and Hainmueller (2015) using the SCM. Results are then reported from the in-space placebo test:

```
Implementing placebo test using fake treatment unit Australia...Austria...
> Belgium...Denmark...France...Greece...Italy...Japan...Netherlands...NewZealand
> ...Norway...Portugal...Spain...Switzerland...UK...USA...
Placebo test results using fake treatment units:
```

Unit	Pre MSPE	Post MSPE	Post/Pre MSPE	Pre MSPE of Fake Unit/ Pre MSPE of Treated Unit
WestGermany	299.4172	3.79e+06	12654.5085	1.0000
Australia	7823.7705	.	.	26.1300
Austria	2715.1534	4.34e+05	160.0022	9.0681
Belgium	543.8757	.	.	1.8164
Denmark	4200.7224	1.09e+06	258.5140	14.0297
France	365.3031	2.93e+05	802.7585	1.2200
Greece	8905.4055	.	.	29.7425
Italy	4666.7858	1.17e+05	25.1081	15.5862
Japan	1656.3602	1.10e+06	664.9375	5.5319
Netherlands	1675.8367	5.79e+05	345.3352	5.5970
NewZealand	5861.0887	7.17e+05	122.3009	19.5750
Norway	2575.4635	.	.	8.6016
Portugal	5330.0930	.	.	17.8016
Spain	476.2446	2.10e+05	441.4192	1.5906
Switzerland	6176.4133	.	.	20.6281
UK	2613.4331	4.13e+05	157.8457	8.7284
USA	1.11e+04	2.59e+06	234.1799	36.9503

```
Note: (1) The probability of obtaining a post/pre-treatment MSPE ratio as
      large as WestGermany's is 0.4118.
      (2) Total 4 units with pre-treatment MSPE 20 times larger than the
      treated unit are excluded in computing pointwise p-values, including
      Australia Greece Switzerland USA.
```

Placebo test results using fake treatment units (continued, cutoff = 20):

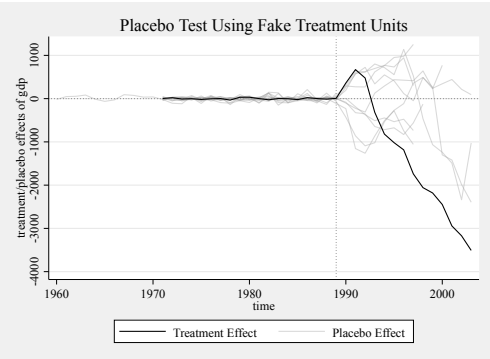
Time	Treatment Effect	p-value of Treatment Effect		
		Two-sided	Right-sided	Left-sided
1990	360.3867	0.3846	0.3077	0.7692
1991	671.6191	0.4615	0.3077	0.7692
1992	476.7500	0.6923	0.5385	0.5385
1993	-316.6797	0.9231	0.6923	0.3846
1994	-819.9297	0.3077	1.0000	0.0769
1995	-1017.6562	0.3077	1.0000	0.0769
1996	-1184.1309	0.3077	1.0000	0.0769
1997	-1740.0312	0.3077	1.0000	0.0769
1998	-2057.5430	0.6154	1.0000	0.0769
1999	-2180.7734	0.6923	1.0000	0.0769
2000	-2446.8184	0.6923	1.0000	0.0769
2001	-2943.2207	0.7692	1.0000	0.0769
2002	-3170.2871	0.7692	1.0000	0.0769
2003	-3508.5508	0.7692	1.0000	0.0769

Note: (1) The two-sided p-value of the treatment effect for a particular period is defined as the frequency that the absolute values of the placebo effects are greater than or equal to the absolute value of treatment effect.
(2) The right-sided (left-sided) p-value of the treatment effect for a particular period is defined as the frequency that the placebo effects are greater (smaller) than or equal to the treatment effect.
(3) If the treatment effects are mostly positive, then the right-sided p-values are recommended; whereas the left-sided p-values are recommended if the treatment effects are mostly negative.

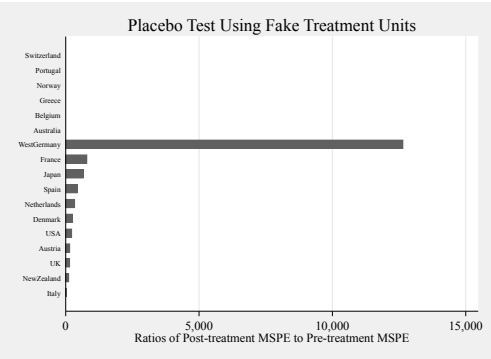
Notice that the posttreatment MSPEs for six countries (that is, Australia, Belgium, Greece, Norway, Portugal, and Switzerland) are missing, which is due to missing observations for making posttreatment predictions. If one wishes to fill in missing values before applying the RCM, the `rcm` command provides convenient utilities: one can use either the sample mean with the option `fill(mean)` or linear interpolation with the option `fill(linear)`.

The above results show that, as an overall measure of significance of the treatment effects over the entire posttreatment periods, the p-value based on the post/pre MSPE ratio is 0.4118, which is larger than any conventional level of significance. Then the four countries (that is, Australia, Greece, Switzerland, and the United States) with pretreatment MSPEs at least 20 times more than that of West Germany are excluded, and the rest will continue with the pointwise in-space placebo test.

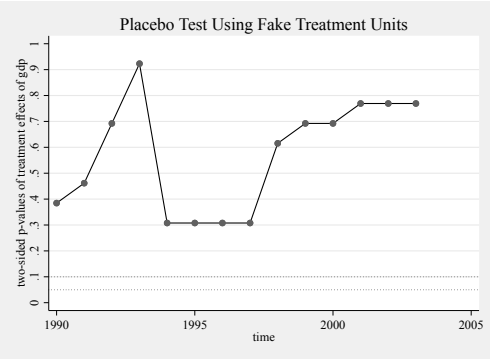
However, if we look at pointwise left-sided p-values (because the treatment effects are mostly negative), the left-sided p-values are 0.0769 for 10 posttreatment periods starting from 1994, which is significant at the 10% level. Also, bear in mind that the smallest p-value attainable for the dataset is $(1/13) \approx 0.0769$ because there are only 12 countries kept in the donor pool. These results from the in-space placebo test are presented graphically in figure 6.



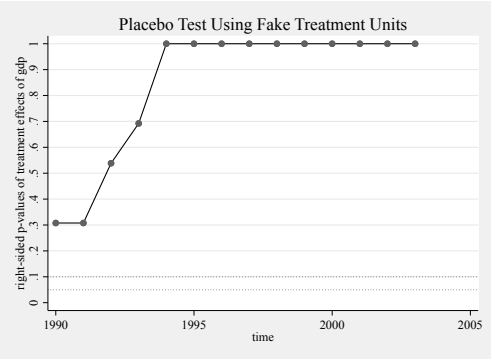
(a) Treatment and placebo effects



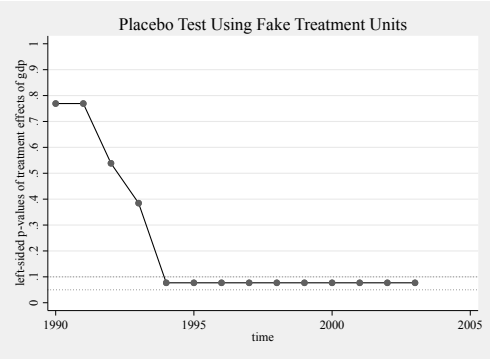
(b) Post/pre MSPE ratios



(c) Two-sided p-values of treatment effects



(d) Right-sided p-values of treatment effects



(e) Left-sided p-values of treatment effects

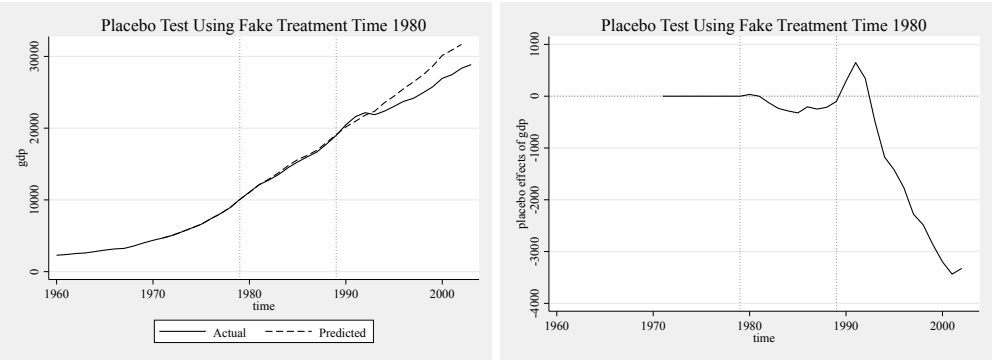
Figure 6. Graphs for in-space placebo test in example 2

Last, the results from the in-time placebo test using 1980 as the fake treatment time are reported, as well as the corresponding graphs shown in figure 7.

```
Implementing placebo test using fake treatment time 1980...
Placebo test results using fake treatment time 1980:
```

Time	Actual Outcome	Predicted Outcome	Treatment Effect
1980	11083.0000	11048.8652	34.1348
1981	12115.0000	12114.9834	0.0166
1982	12761.0000	12888.2305	-127.2305
1983	13519.0000	13755.9736	-236.9736
1984	14481.0000	14766.0557	-285.0557
1985	15291.0000	15613.6670	-322.6670
1986	15998.0000	16203.2393	-205.2393
1987	16679.0000	16926.9902	-247.9902
1988	17786.0000	17999.3027	-213.3027
1989	18994.0000	19095.3828	-101.3828
1990	20465.0000	20174.4883	290.5117
1991	21602.0000	20953.9238	648.0762
1992	22154.0000	21812.3711	341.6289
1993	21878.0000	22367.1992	-489.1992
1994	22371.0000	23546.5918	-1175.5918
1995	23035.0000	24460.1660	-1425.1660
1996	23742.0000	25509.9199	-1767.9199
1997	24156.0000	26435.2461	-2279.2461
1998	24931.0000	27414.1738	-2483.1738
1999	25755.0000	28618.4395	-2863.4395
2000	26943.0000	30140.9590	-3197.9590
2001	27449.0000	30884.2793	-3435.2793
2002	28348.0000	31672.4336	-3324.4336
2003	28855.0000	.	.
Mean	20066.7826	21060.9949	-994.2123

```
Note: The average treatment effect over the post-treatment periods is
      -994.2123.
Finished.
```



(a) Actual and predicted paths

(b) Treatment effects

Figure 7. Graphs for in-time placebo test in example 2

Apparently, the placebo effects during the "fake posttreatment periods" from 1980 to 1989 are negligible, which supports our confidence in the significance of actual treatment effects, if any.

7 Conclusion

The RCM is a convenient approach for causal inference in panel data with a single treated unit that exploits cross-sectional correlation to construct counterfactual outcomes by linear regression. In this article, we reviewed the RCM methodology and presented the command `rcm` for efficient implementation of the RCM with or without covariates.

Available methods for model selection include best subset, lasso, and forward stepwise and backward stepwise regression, while available selection criteria include the AICc, the AIC, the BIC, the MBIC, and cross-validation. Estimation and prediction can be made by OLS, lasso, or postlasso OLS. For statistical inference, both in-space and in-time placebo tests can be implemented. The `rcm` command produces a series of graphs for visualization along the way. We demonstrated the use of `rcm` by revisiting classic examples of political and economic integration between Hong Kong and mainland China (Hsiao, Ching, and Wan 2012) and German reunification (Abadie, Diamond, and Hainmueller 2015).

8 Acknowledgments

The scientific calculation in this article has been done on the HPC Cloud Platform of Shandong University.[4] We thank editor Stephen P. Jenkins, Zhentao Shi, Congshan Zhou, and an anonymous referee for very helpful comments. All remaining errors are ours.

9 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-4
. net install st0693     (to install program files, if available)
. net get st0693         (to install ancillary files, if available)
```

The `rcm` command also is available on the Statistical Software Components Archive and can be installed directly in Stata with the command

```
. ssc install rcm
```

10 References

Abadie, A., A. Diamond, and J. Hainmueller. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association* 105: 493–505. https://doi.org/10.1198/jasa.2009.ap08746.

———. 2015. Comparative politics and the synthetic control method. *American Journal of Political Science* 59: 495–510. https://doi.org/10.1111/ajps.12116.

Abadie, A., and J. Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque country. *American Economic Review* 93: 113–132. https://doi.org/10.1257/000282803321455188.

Carvalho, C., R. Masini, and M. C. Medeiros. 2018. ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. *Journal of Econometrics* 207: 352–380. https://doi.org/10.1016/j.jeconom.2018.07.005.

Chen, Q., Z. Xiao, and Q. Yao. 2022. Quantile control via random forest. Working Paper, Shandong University.

Du, Z., and L. Zhang. 2015. Home-purchase restriction, property tax and housing price in China: A counterfactual analysis. *Journal of Econometrics* 188: 558–568. https://doi.org/10.1016/j.jeconom.2015.03.018.

4. In fact, the computation for the RCM is not demanding, because it is based on OLS and lasso regressions. In addition, we have used Mata to optimize the codes, which enables the `rcm` command to run smoothly on computers with either high or low configurations.

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B* 70: 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x.

Furnival, G. M., and R. W. Wilson. 1974. Regressions by leaps and bounds. *Technometrics* 16: 499–511. https://doi.org/10.2307/1267601.

Hsiao, C., H. S. Ching, and S. K. Wan. 2012. A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics* 27: 705–740. https://doi.org/10.1002/jae.1230.

Hsiao, C., and Q. Zhou. 2019. Panel parametric, semiparametric, and nonparametric construction of counterfactuals. *Journal of Applied Econometrics* 34: 463–481. https://doi.org/10.1002/jae.2702.

Ke, X., H. Chen, Y. Hong, and C. Hsiao. 2017. Do China's high-speed-rail projects promote local economy?—New evidence from a panel data approach. *China Economic Review* 44: 203–226. https://doi.org/10.1016/j.chieco.2017.02.008.

Li, K. T., and D. R. Bell. 2017. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics* 197: 65–75. https://doi.org/10.1016/j.jeconom.2016.01.011.

Narendra, P. M., and K. Fukunaga. 1977. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers* 26: 917–922. https://doi.org/10.1109/TC.1977.1674939.

Ni, X. S., and X. Huo. 2006. Regressions by enhanced leaps-and-bounds via additional optimality tests (LBOT).

Ouyang, M., and Y. Peng. 2015. The treatment-effect estimation: A case study of the 2008 economic stimulus package of China. *Journal of Econometrics* 188: 545–557. https://doi.org/10.1016/j.jeconom.2015.03.017.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66: 688–701. https://doi.org/10.1037/h0037350.

Shi, Z., and J. Huang. Forthcoming. Forward-selected panel data approach for program evaluation. *Journal of Econometrics*. https://doi.org/10.1016/j.jeconom.2021.04.009.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Vega-Bayo, A. 2015. An R package for the panel approach method for program evaluation: pampe. *R Journal* 7: 105–121. https://doi.org/10.32614/RJ-2015-024.

Wang, H., B. Li, and C. Leng. 2009. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B* 71: 671–683. https://doi.org/10.1111/j.1467-9868.2008.00693.x.

About the authors

Guanpeng Yan is a PhD student at the School of Economics, Shandong University, Jinan, Shandong Province, China.

Qiang Chen is a professor at the School of Economics, Shandong University, Jinan, Shandong Province, China.