Regression Control Method with Stata

回归控制法及Stata应用

School of Economics
Shandong University
颜冠鹏
Guanpengyan@mail.sdu.edu.cn

Outline

1. Introduction
2. Model
3. Extension
4. The rcm command
5. Examples

1. Introduction

- Regression control method (RCM)
  - Aka a panel data approach for program evaluation (Hsiao et al. 2012)
  - Exploits cross-sectional correlation to construct counterfactual outcomes for a single treated unit
  - Hsiao and Zhou (2019) propose to add covariates in model for helping prediction

1. Introduction

- Stata command `rcm` :
  - Implement regression control method
  - Facilitate model selection and estimation
  - Support to add covariates
  - Support placebo tests for statistical inferences

2. Model

2.1 Basic Model

- N cross-sectional units indexed by $i = 1, \ldots, N$ over $t = 1, \ldots, T_0 + 1, \ldots, T$ periods
- $i = 1$ indexes the treated unit, whereas $\{2, \ldots, N\}$ is the index set of $N - 1$ control units (donor pool)
- Policy intervention occurs at time $T_0 + 1$, which partitions the time series into two sections:
    - Pre-treatment periods: $1, \ldots, T_0$
    - Post-treatment periods: $T_0 + 1, \ldots, T$

2.1 Basic Model

- y_{it}^1 and y_{it}^0 be the outcomes of the unit i in period t with and without intervention respectively
- y_{it} is observed that in the form:

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

- $d_{it} = 1$ if $i = 1$ and unit i is under intervention in period t, or $d_{it} = 0$ if not
- The treatment effect for can be expressed as

$$\Delta_{it} = y_{it}^1 - y_{it}^0 \quad t = T_0 + 1, \ldots, T$$

2.1 Basic Model

- Assume that y_{it}^0 is generated by a **pure** factor model of the form

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

- f_t: $K \times 1$ (unobserved) common factors
- $\mathbf{b}_i'$: $1 \times K$ (unobserved) factor loadings
- ε_{it}: Random idiosyncratic component with $E\left(\varepsilon_{it}\right) = 0$.

2.1 Basic Model

Stack y_{it}^0 for $i \in \{2, \ldots, N\}$, we get

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

- $\widetilde{\mathbf{y}}_t = \left(y_{2t}^0, \ldots, y_{Nt}^0\right)'$
- $\widetilde{\boldsymbol{\varepsilon}}_t = \left(\varepsilon_{2t}, \ldots, \varepsilon_{Nt}\right)'$
- $\widetilde{\mathbf{B}}$: $(N-1) \times K$ factor loading matrix $\left(\mathbf{b}_2, \ldots, \mathbf{b}_N\right)'$

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

2.1 Basic Model

Then y_{1t}^0 can be represented as

$$\begin{aligned} y_{1t}^0 &= \mathbf{b}_1' \mathbf{f}_t + \varepsilon_{1t} \\ &= \mathbf{b}_1' \left(\widetilde{\mathbf{B}}' \widetilde{\mathbf{B}} \right)^{-1} \widetilde{\mathbf{B}}' \left(\widetilde{\mathbf{y}}_t - \widetilde{\boldsymbol{\varepsilon}}_t \right) + \varepsilon_{1t} \\ &= \boldsymbol{\gamma}' \widetilde{\mathbf{y}}_t + \varepsilon_{1t} - \boldsymbol{\gamma}' \widetilde{\boldsymbol{\varepsilon}}_t \end{aligned}$$

- $\boldsymbol{\gamma}' = \mathbf{b}_1' \left(\widetilde{\mathbf{B}}' \widetilde{\mathbf{B}} \right)^{-1} \widetilde{\mathbf{B}}'$
- $\widetilde{\mathbf{y}}_t$ is correlated to $\varepsilon_{1t} - \boldsymbol{\gamma}' \widetilde{\boldsymbol{\varepsilon}}_t$

2.1 Basic Model

$$
\begin{aligned}
y_{1t}^0 &= \gamma'\widetilde{\mathbf{y}}_t + \varepsilon_{1t} - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\varepsilon}}_t \\
&= \gamma'\widetilde{\mathbf{y}}_t + L\left(\varepsilon_{1t} - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\varepsilon}}_t \middle| \widetilde{\mathbf{y}}_t\right) + \upsilon_{1t} \\
&= \gamma'\widetilde{\mathbf{y}}_t + c_1 + \mathbf{c}'\widetilde{\mathbf{y}}_t + \upsilon_{1t}
\end{aligned}
$$

- $L\left(\varepsilon_{1t} - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\varepsilon}}_t \middle| \widetilde{\mathbf{y}}_t\right)$ is the linear projection of $\varepsilon_{1t} - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\varepsilon}}_t$ onto $(1, \widetilde{\mathbf{y}}_t')$, and c_1 and $\mathbf{c}$ are the minimizers of

$$
\min_{c_1, \mathbf{c}} E\left[\left(\varepsilon_{1t} - \boldsymbol{\gamma}'\widetilde{\boldsymbol{\varepsilon}}_t - c_1 - \mathbf{c}'\widetilde{\mathbf{y}}_t\right)^2\right]
$$

2.1 Basic Model

$$\begin{aligned} y_{1t}^0 &= \boldsymbol{\gamma}'\widetilde{\mathbf{y}}_t + c_1 + \mathbf{c}'\widetilde{\mathbf{y}}_t + v_{1t} \\ &= \delta_1 + \boldsymbol{\delta}'\widetilde{\mathbf{y}}_t + v_{1t} \end{aligned}$$

- $\delta_1 = c_1, \boldsymbol{\delta}' = \boldsymbol{\gamma}' + \mathbf{c}'$
- Hsiao et al. (2012) advocate estimating $\hat{\delta}_1$ and $\hat{\boldsymbol{\delta}}'$ by OLS with the pre-treatment subsample, and predicting the counterfactual outcomes as

$$\hat{y}_{1t}^0 = \hat{\delta}_1 + \hat{\boldsymbol{\delta}}'\widetilde{\mathbf{y}}_t$$

2.1 Basic Model

- So the treatment effect Δ_{1t} can be predicted using

$$\hat{\Delta}_{1t} = y_{1t}^1 - \hat{y}_{1t}^0 \quad t = T_0 + 1, \ldots, T$$

- Average treatment effect (ATE) is estimated by averaging $\hat{\Delta}_{1t}$ over the post-treatment periods:

$$\hat{\Delta}_1 = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \hat{\Delta}_{1t}$$

2.1 Basic Model

- Use all of control units for estimation may not be the best choice
- Large estimation variance in turn leads to poor out-of-sample predictions
- Hsiao et at. (2012) suggest using information criterion approach to select control units in exhaustive search (best subset selection)
- Use $\widetilde{\mathbf{y}}_t^*$ instead of $\widetilde{\mathbf{y}}_t$, where $\widetilde{\mathbf{y}}_t^*$ is the best subset of $(y_{2t}, \ldots, y_{Nt})'$

2.2 Model with Covariates (Hsiao and Zhou, 2019)

- Assume y_{it}^0 is a function of p observable variables $\mathbf{x}_{it}$:

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

- y_{1t}^0 can be predicted as

$$\hat{y}_{1t}^0 = \hat{\delta}_1 + \hat{\boldsymbol{\delta}}'\mathbf{z}_t^*$$

- $\mathbf{z}_t^*$ includes any subset of $\mathbf{z}_t = (y_{2t}, \ldots, y_{Nt}, \mathbf{x}_{1t} \ldots, \mathbf{x}_{Nt})'$ that helps to predict y_{1t}^0
(forward stepwise method or Lasso method suggested)

3 Extension

- Step 1 - Select the Suboptimal Model
  Select the suboptimal model $M^*_{OLS}(j)$ using best subset selection, forward stepwise selection or backward stepwise selection, or $M^*_{lasso}(\lambda)$ using Lasso seletion $(j \in \{1, \ldots, T_0 - 1\})$
- Step 2 - Select the Optimal Model
  Choose the optimal model M^* from all of suboptimal models in terms of information criterion or cross-validation
- After model selection, OLS or Lasso regression is used to fit the optimal model for counterfactual prediction

3.1 Select the Suboptimal Model

3.1.1 Best Subset Selection

- Hsiao et. al. (2012) suggest using best subset selection
- $exhaustion$:

  For $k = 1, 2, \ldots p$:

  1. Fit all $\binom{p}{k}$ models that contain k predictors using OLS regression
  2. Choose the smallest RSS model as $M^*_{OLS}(k)$
- $2^p - 1$ possible combination of the p predictors

3.1.1 Best Subset Selection

- $leaps\ and\ bounds$: quickly calculate best subsets without examining all possible subsets
- Fundamental inequality:

$$RSS(A) \leq RSS(B) \quad B \subset A$$

- If $RSS(\{3,4\}) \geq RSS(1)$, $RSS(\{3\})$ and $RSS(\{4\})$ do not need to be calculated

3.1.1 Best Subset Selection

- $\textbf{Initial : Reorder the variables by their impact on RSS}$
- Regression and bound tree (pair tree)
- Traverse all the subsets of the root node in level 1
- Traverse the tree from right to left

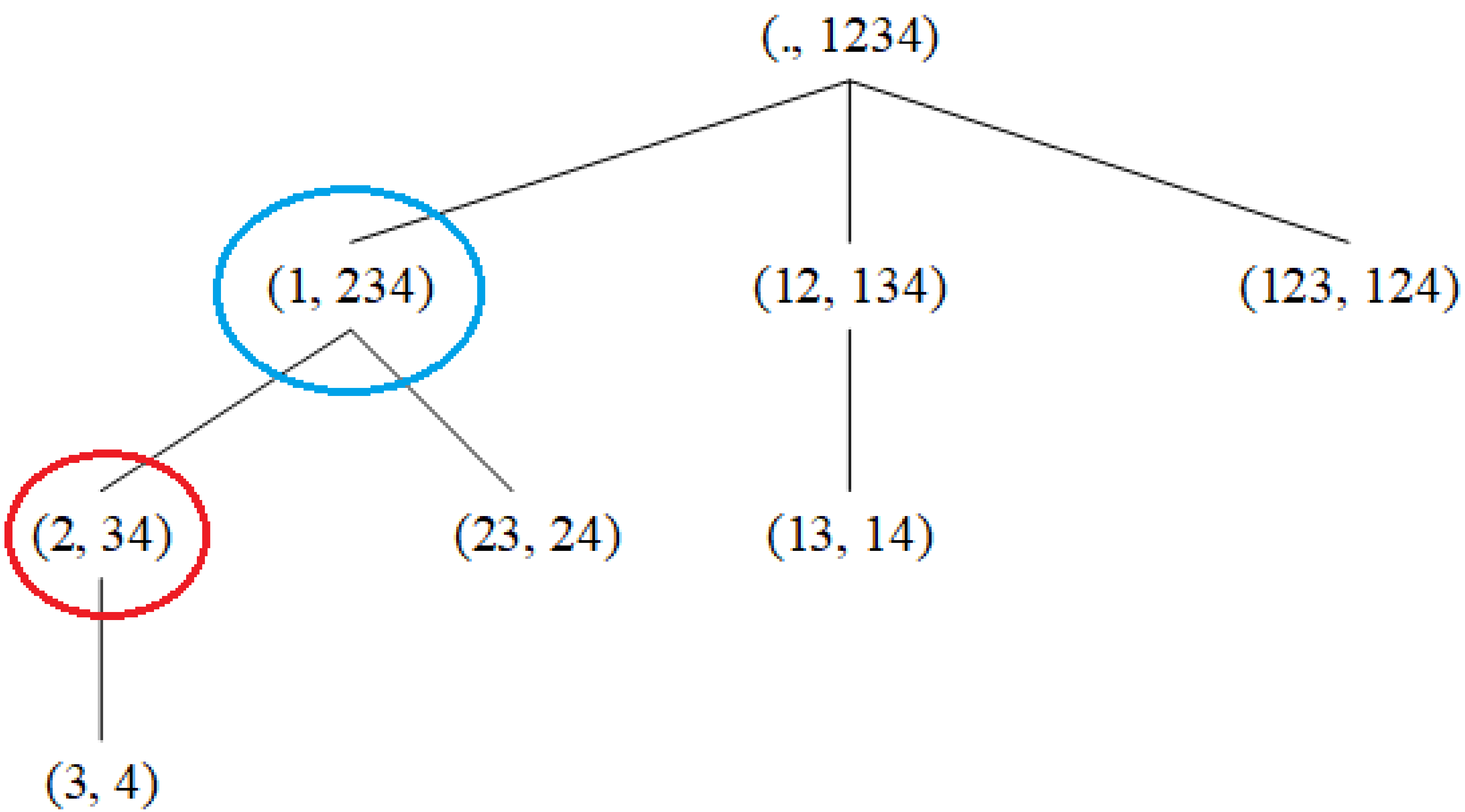


Figure : The pair tree

3.1.1 Best Subset Selection

- Suppose subset $\Omega \subset \{1, 2, \ldots, p+1\}$, we have

$$\mathrm{RSS}(\Omega) = y'y - (\Phi'_\Omega \cdot y)' \left(\Phi'_\Omega \Phi_\Omega\right)^{-1} (\Phi'_\Omega \cdot y)$$

- Φ is $T_0 \times (p+1)$ matrix with a constant column and predictors
- Precomputed matrix :

$$(y, \Phi)'(y, \Phi) = \begin{pmatrix} y'y & y'\Phi \\ \Phi' y & \Phi'\Phi \end{pmatrix}$$

3.1.2 Forward Stepwise Selection

- Hsiao and Zhou (2019) suggest using forward stepwise selection
- $M^*_{OLS}(0)$ denote the smallest model, which contains no predictors
- For $k = 0, \ldots, p-1:$
  1. Consider models that augment the predictors in $M^*_{OLS}(k)$ with one additional predictor, fit OLS regression
  2. Fit these models with OLS regression
  3. Choose the smallest $\boldsymbol{RSS}$ model as $M^*_{OLS}(k+1)$
- Shi and Huang (2021) : Forward-Selected PDA (fsPDa)

3.1.3 Backward Stepwise Selection

- $M^*_{OLS}(p)$ denote the largest possible model, which contains all p predictors
- For $k = p, \ldots, 1:$
  1. Consider models that contain all but one of the predictors in $M^*_{OLS}(k)$
  2. Fit these models with OLS regression
  3. Choose the smallest RSS model as $M^*_{OLS}(k-1)$
- Require that $T_0 > p$ (the largest possible model can be fit)

3.1.4 Lasso Selection

- Lasso selects $\boldsymbol{\beta}$ to minimize:

$$\hat{\boldsymbol{\beta}}_{lasso}(\lambda) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- $\lambda \geq 0$ is a tuning parameter
- Two extreme cases : $\hat{\boldsymbol{\beta}}_{lasso}(\lambda = 0) = \hat{\boldsymbol{\beta}}_{ols}, \hat{\boldsymbol{\beta}}_{lasso}(\lambda = \infty) = \mathbf{0}$

3.1.4 Lasso Selection

- For $\lambda_l = \lambda_1, \ldots, \lambda_L$:

1. Calculate lasso estimator $\hat{\boldsymbol{\beta}}_{lasso}(\lambda_l)$
2. Consider the model with a set of predictors $\Omega(\lambda_l)$, where

$$\Omega(\lambda_l) = \{predictor_j | j \in \{1, \ldots, p\}, \beta_{lasso,j}(\lambda_l) \neq 0\}$$

3. Choose the model as $M^*_{lasso}(\lambda_l)$ if $\Omega(\lambda_l)$ is different from $\Omega(\lambda_1), \ldots, \Omega(\lambda_{l-1})$

3.2 Select Optimal Model

3.2.1 Information criterion

$$\mathrm{AIC}(p) = T_0 \ln\left(\frac{\mathbf{e}_0'\mathbf{e}_0}{T_0}\right) + 2(p+2)$$

$$\mathrm{BIC}(p) = T_0 \ln\left(\frac{\mathbf{e}_0'\mathbf{e}_0}{T_0}\right) + (p+2)\ln(T_0)$$

$$\mathrm{AICc}(p) = \mathrm{AIC}(p) + \frac{2(p+2)(p+3)}{T_1 - (p+1) - 2}$$

- $\mathbf{e}_0'$: OLS or Lasso residuals fitted in pre-treatment periods.

3.2.2 Cross-Validation

- Common practice after Lasso selection
- Optimize the out-of-sample prediction performance
- The mean squared error for each fold is computed as

$$\mathrm{MSE}(\lambda, k) = \frac{1}{n_k} \sum_{i \in \mathcal{F}_k} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_{lasso,k}(\lambda) \right)^2$$

- K : Number of groups in which pre-treatment data is splitted
- n_k : Size of pre-treatment data partition k for $k = 1, \ldots, K$
- $\mathcal{F}_k$: Set of observations in k-fold

3.2.2 Cross-Validation

- K-fold Cross-Validation estimate of $\mathrm{CV\ MSE}$, which serves as a measure of prediction performance, is

$$
\begin{aligned}
\mathrm{CV\ MSE}(\lambda, K) &= \frac{1}{K}\sum_{k=1}^{K}\mathrm{MSE}(\lambda, k) \\
&= \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i\in\mathcal{F}_k}\left(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{lasso,k}(\lambda)\right)\right)
\end{aligned}
$$

3.3 Post-Estimation of the Optimal Model

- Post-Estimation OLS :
  - Common practice
  - Fit OLS regression to the optimal model and obtain β_{ols}
  - Use β_{ols} for counterfactual prediction
  - **Post-Lasso OLS** : Fit OLS regression after Lasso selected
- Post-Estimation Lasso:
  - Obtain $\beta_{lasso}(\lambda)$ in **Lasso selection process**
  - Use $\beta_{lasso}(\lambda)$ for counterfactual prediction
  - **Can only be used after Lasso selection**

3.4 Placebo Test

3.4.1 Placebo Test Using Fake Treatment Unit

- Reassign the treatment to control units (donor pool) where no intervention actually occurred
- Determine statistical significance of treatment effect
- $p\text{-}val(t)$: p-value of estimated effect for a particular period is

$$p\text{-}val(t) = \frac{1}{N}\sum_{i=1}^{N} 1\left(\left|\hat{\Delta}_{it}\right| \geq \left|\hat{\Delta}_{1t}\right|\right) \qquad t = T_0 + 1, \ldots, T$$

3.4.1 Placebo Test Using Fake Treatment Unit

- $p\text{-}val$: The probability of obtaining a post/pre-MSPE ratio as large as that of treated unit, is

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

- Cutoff : Discard the fake units $i \in \{2, \ldots, N\}$ with extreme values of $\mathrm{MSPE}_{i,pre}$ (Abadie et al. , 2010)

3.4.2 Placebo Test Using Fake Treatment Time

- Reassign the treatment to periods previous to the intervention when no treatment actually ocurred
- Whether there is a perceivable effect durning $1, \ldots, T_0 + 1$ periods?

4 The rcm command

- Implement "Regression Control Method (RCM)" in Stata
- Install : `ssc install rcm, all replace` with Stata version >= 16
- Installed files :
  - rcm.ado and rcm.sthlp: Stata ado and help file
- Ancillary files :
  - growth.dta : The dataset obtained from Hsiao et al. (2012) which has been reshaped to long form
  - repgermany.dta : The dataset obtained from Abadie et al. (2015) of which panel variable has been reencoded

4 The rcm command

- Syntax:

```
rcm depvar [indepvars] , trunit(#) trperiod(#) [options]
```

  - `xtset panelvar timevar` must be used to declare a panel dataset in the usual long form
  - `depvar` and `indepvars` must be numeric variables, and abbreviations are not allowed.
- Without `indepvars` : Basic model (Section 2.1)
- With `indepvars` : Model with covariates (Section 2.2)

5 Examples

5.1 Replicate Hsiao et al.(2012)

- Consider the impact on Hong Kong real GDP growth rate with the reversion of sovereignty on 1 July 1997 from the UK to China
    - Treatment period : 1997Q3
    - Pre-treatment preiods : 1993Q1-1997Q2
    - Post-treatment preiods : 1997Q3-2004Q1
    - Treated unit : Hong Kong
    - Control units : 10 countries/regions that are either in the region or economically closely associated with Hong Kong

5.1 Replicate Hsiao et al.(2012)

```
use growth, clear
xtset region time
des
```

```
 Observations:         1,525
    Variables:             3                 16 Jun 2019 00:03
```

Variable name	Storage type	Display format	Value label	Variable label
time	float	%tq		
gdp	float	%8.0g		
region	long	%13.0g	region	

Sorted by: **region** **time**

5.1 Replicate Hsiao et al.(2012)

```
label list
```

```
region:
          1 Australia      14 Malaysia
          2 Austria        15 Mexico
          3 Canada         16 Netherlands
          4 China          17 NewZealand
          5 Denmark        18 Norway
          6 Finland        19 Philippines
          7 France         20 Singapore
          8 Germany        21 Switzerland
          9 HongKong       22 Taiwan
         10 Indonesia      23 Thailand
         11 Italy          24 UnitedKingdom
         12 Japan          25 UnitedStates
         13 Korea
```

5.1 Replicate Hsiao et al.(2012)

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

- `gdp` : Specifies "gdp" as dependent variable (outcome variable)
- `trunit(9)` : Specifies "HongKong" as the treated unit
- `trperiod(150)` : Specifies "1997q3" as the treatment period (`150` is obtained from `di tq(1997q3)`
- `ctrlunit(4 10 12 13 14 19 20 22 23 25)` : Specifies 10 countries/regions as the control units
- `postperiod(150/175)` : Specifies "1997q3-2003q4" as the post-treatment periods (`175` is obtained from `di tq(2003q4)`)

5.1 Replicate Hsiao et al.(2012)

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	R-squared
1	**-144.7514**	**-146.4657**	**-143.7946**	**0.4034**
2	**-160.5063**	**-163.5832**	**-160.0217**	**0.7937**
3	**-170.6492**	**-175.6492**	**-171.1973**	**0.9056**
4	**-171.7725**	**-179.4088**	**-174.0666**	**0.9314**
5	**-169.7878**	**-180.9878**	**-174.7552**	**0.9438**
6	**-164.2937**	**-180.2937**	**-173.1707**	**0.9477**
7	**-156.6834**	**-179.1834**	**-171.1701**	**0.9503**
8	**-146.2921**	**-177.7207**	**-168.8169**	**0.9517**
9	**-131.7464**	**-175.7464**	**-165.9523**	**0.9518**
10	**-111.3603**	**-173.7603**	**-163.0758**	**0.9518**

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

5.1 Replicate Hsiao et al.(2012)

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

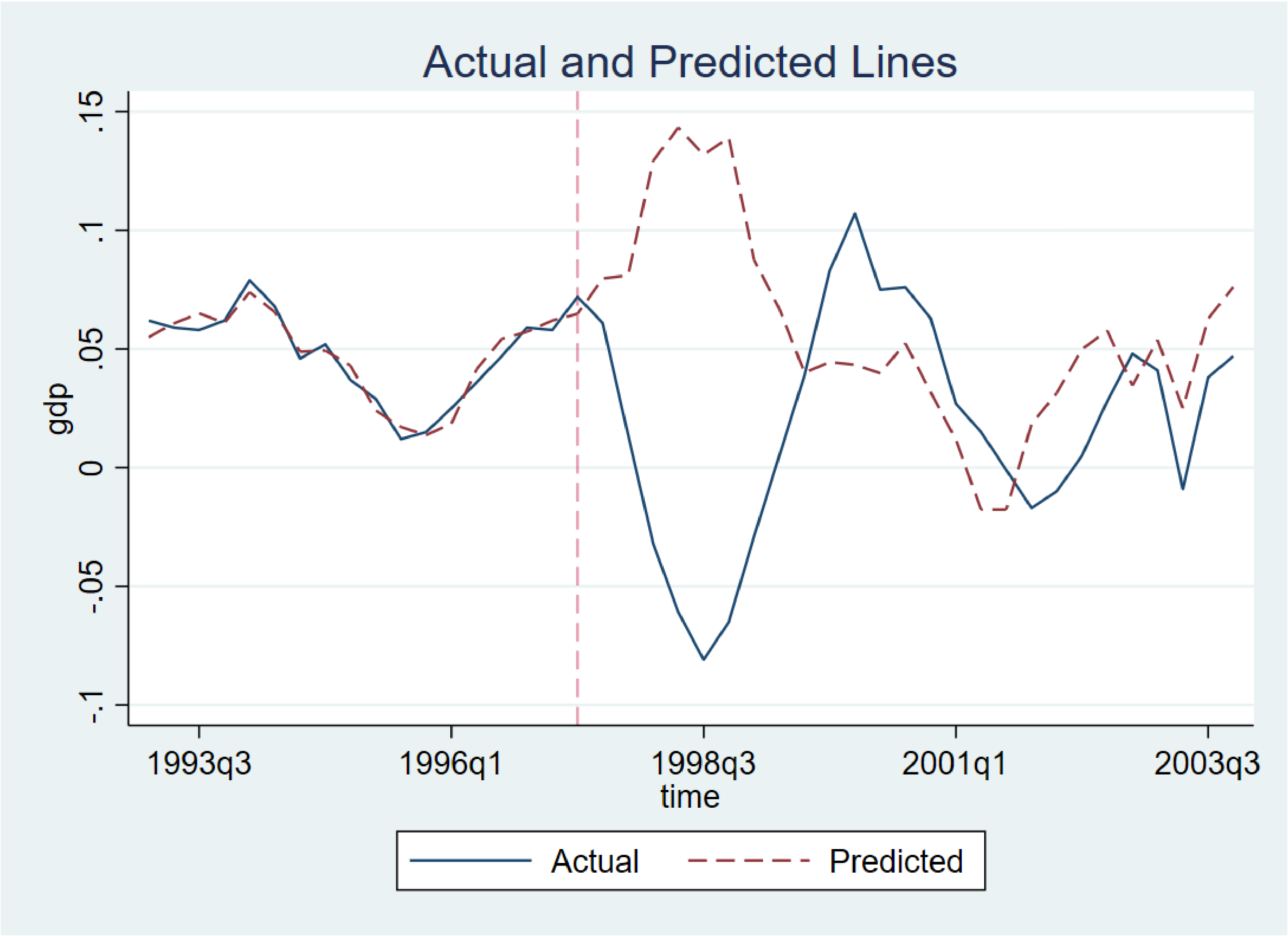
gdp·HongKong	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Korea	-0.4323	0.0634	-6.82	0.000	-0.5692	-0.2954
gdp·Japan	-0.6760	0.1117	-6.05	0.000	-0.9172	-0.4347
gdp·Taiwan	0.7926	0.3099	2.56	0.024	0.1231	1.4621
gdp·UnitedStates	0.4860	0.2195	2.21	0.045	0.0118	0.9603
_cons	0.0263	0.0170	1.54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

5.1 Replicate Hsiao et al.(2012)

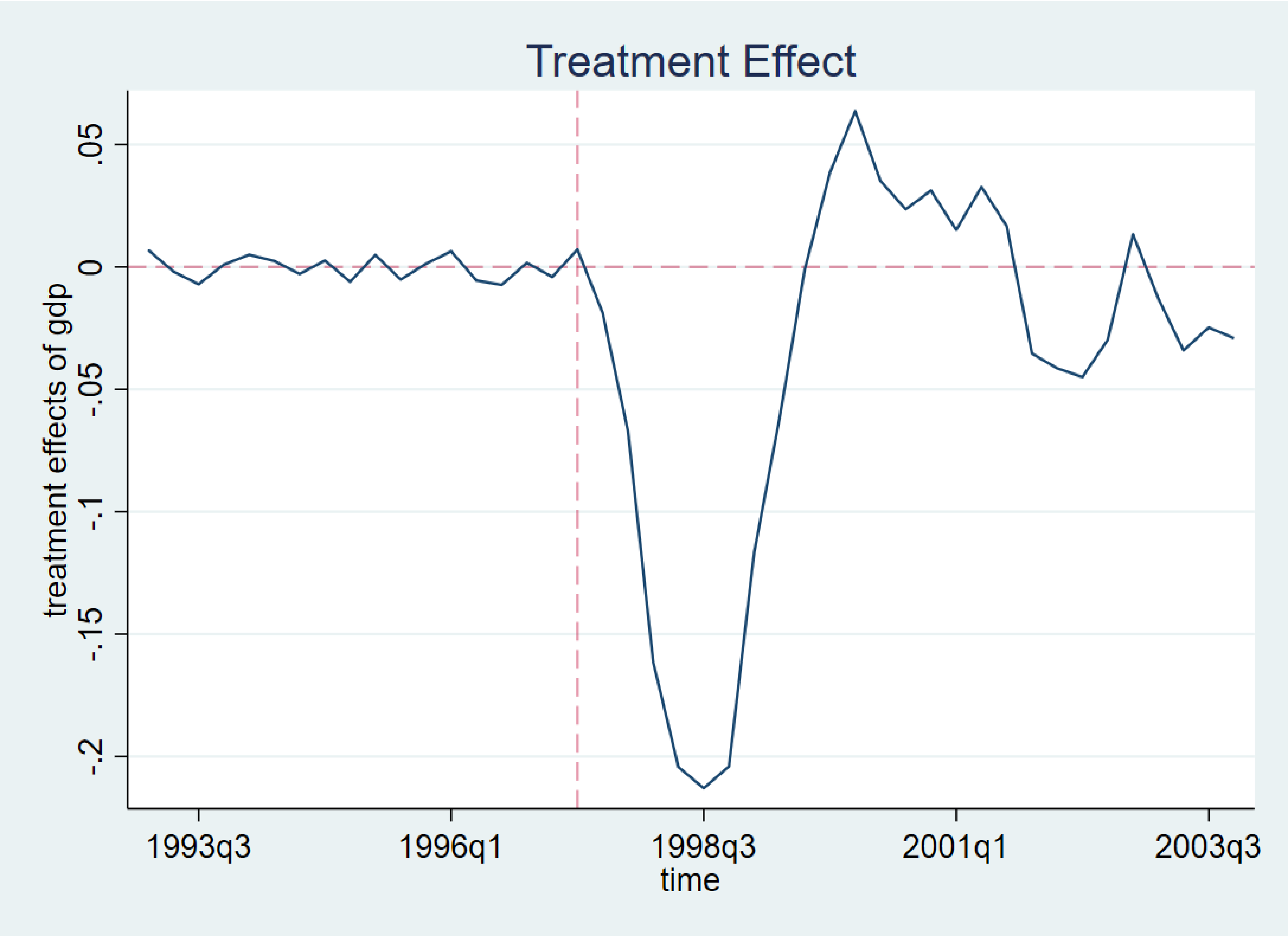
Prediction results in the post-treatment periods using OLS:

Time	Treated	Predicted	Tr. 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**.



Actual and Predicted Lines
gdp
.15
.1
.05
0
-.05
-.1
1993q3
1996q1
1998q3
2001q1
2003q3
time
Actual
Predicted



Treatment Effect
treatment effects of gdp
.05
0
-.05
-.1
-.15
-.2
1993q3
1996q1
1998q3
2001q1
2003q3
time

5.1 Replicate Hsiao et al.(2012)

- Consider the impact on Hong Kong real GDP growth rate with the implementation of CEPA starting in 2004Q1 between mainland China and Hong Kong.
  - Treatment period : 2004Q1
  - Pre-treatment preiods : 1993Q1-2003Q4
  - Post-treatment preiods : 2004Q1-2008Q1
  - Treated unit : Hong Kong
  - Control units : All countries/regions in dataset except Hong Kong

5.1 Replicate Hsiao et al.(2012)

```
rcm gdp, trunit(9) trperiod(176)
```

- `trperiod(176)` : Specifies "2004q1" as the treatment period (`176` is obtained from `di tq(2004q1)`)

5.1 Replicate Hsiao et al.(2012)

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

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

5.1 Replicate Hsiao et al.(2012)

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

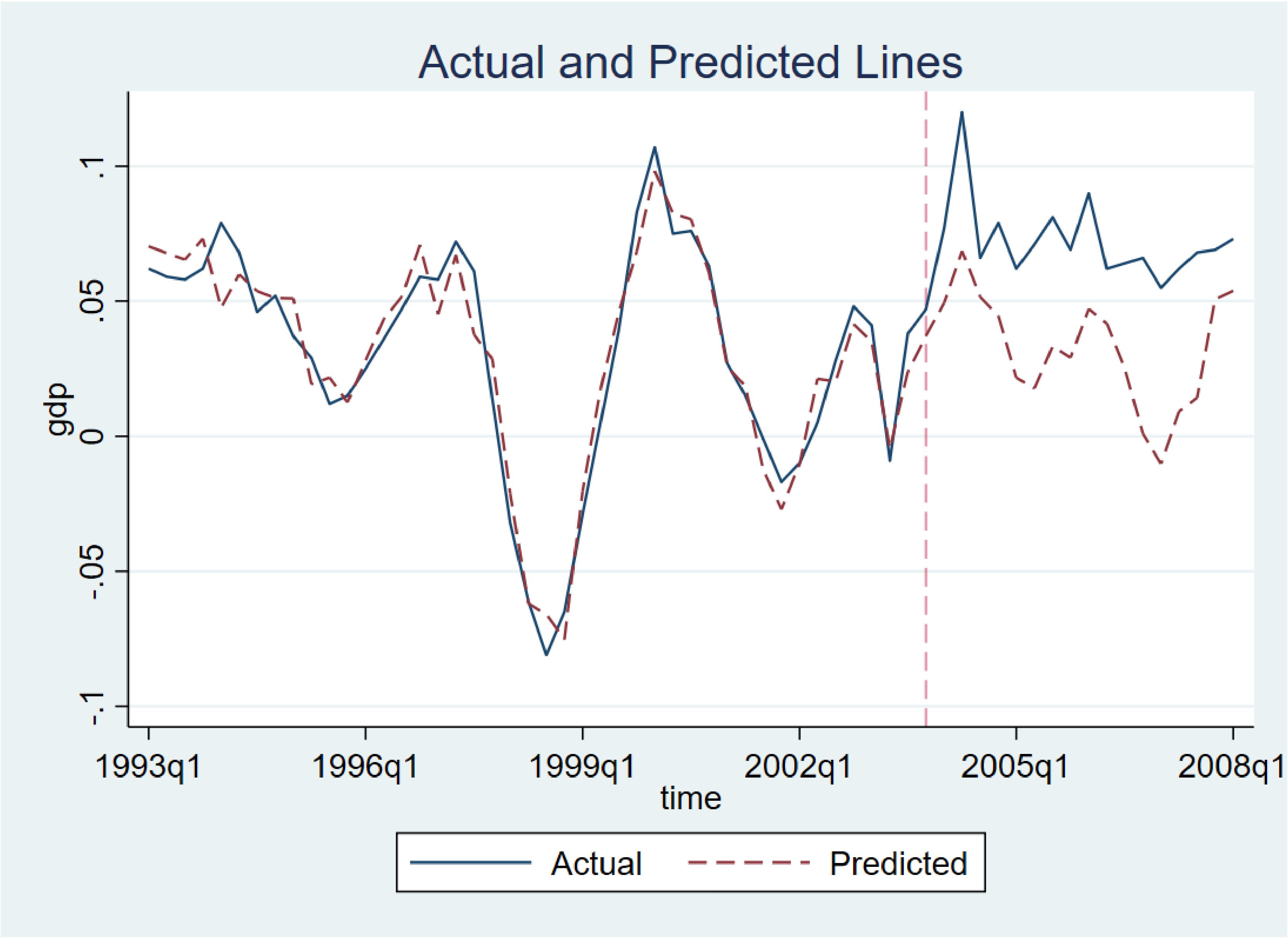
gdp·HongKong	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Norway	0.3222	0.0538	5.99	0.000	0.2132	0.4311
gdp·Austria	-1.0115	0.1682	-6.01	0.000	-1.3524	-0.6707
gdp·Korea	0.3447	0.0469	7.35	0.000	0.2497	0.4398
gdp·Mexico	0.3129	0.0510	6.13	0.000	0.2095	0.4162
gdp·Italy	-0.3177	0.1591	-2.00	0.053	-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

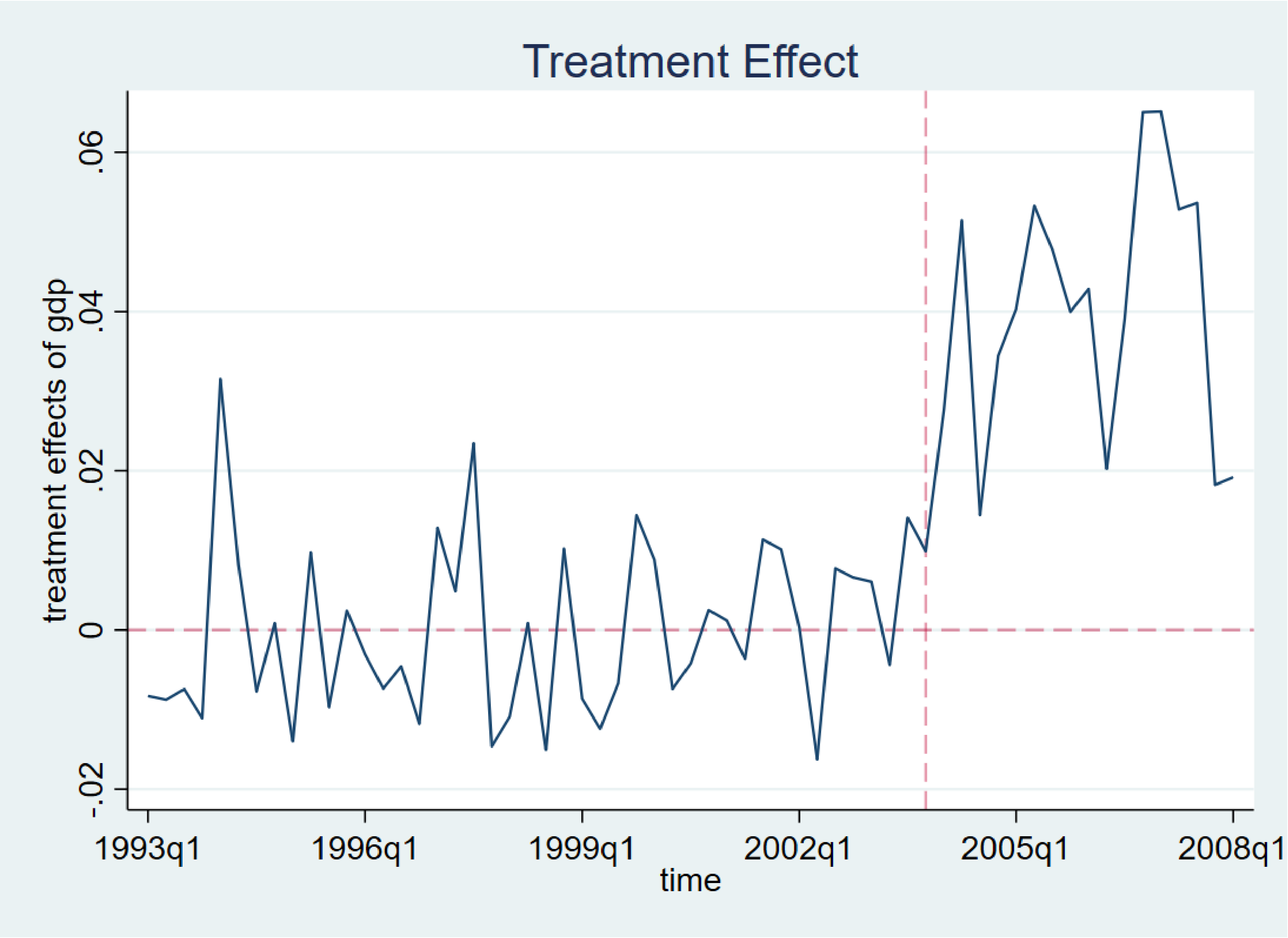
5.1 Replicate Hsiao et al.(2012)

Prediction results in the post-treatment periods using OLS:

Time	Treated	Predicted	Tr. 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**.


Actual and Predicted Lines
gdp
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Actual
Predicted



Treatment Effect
treatment effects of gdp
.06
.04
.02
0
-.02
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time

5.2 Illustrate RCM with Covariates and Placebo Test

- Use the same dataset as Abadie et al. (2015)
- Estimate the economic impact of the 1990 German reunification
  - Treatment period : 1990
  - Pre-treatment preiods : 1960-1989
  - Post-treatment preiods : 1990-2003
  - Treated unit : West Germany
  - Control units : 16 OECD member countries

5.2 Illustrate RCM with Covariates and Placebo Test

```
use repgermany.dta, clear
xtset country year
des
```

```
Contains data from repgermany.dta
 Observations:           748
    Variables:            10                  12 Aug 2021 08:25
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
year            float   %8.0g                 Year
gdp             long    %8.0g                 GDP per-capita (annual)
infrate         float   %9.0g                 Inflation Rate (annual)
trade           float   %9.0g                 Trade openness (annual)
schooling       float   %9.0g                 Schooling (every 5 years)
invest60        float   %9.0g                 Investment rate (average 60-65)
invest70        float   %9.0g                 Investment rate (average 70-75)
invest80        float   %9.0g                 Investment rate (average 80-85)
industry        float   %9.0g                 Industry Share (annual)
country         long    %12.0g     country    Country Name
--------------------------------------------------------------------------------
Sorted by: country  year
```

5.2 Illustrate RCM with Covariates and Placebo Test

```
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
```

5.2 Illustrate RCM with Covariates and Placebo Test

```
rcm gdp infrate trade industry, tru(17) trp(1990) me(lasso) cr(cv) fold(10)
```

- `infrate trade industry` : Specifies three covariates
- `tru(17)` : Abbr. for `trunit(17)`
- `trp(1990)` : Abbr. for `trperiod(1990)`
- `me(lasso)` : Abbr. for `method(lasso)`, which specifies Lasso as selection method
- `cr(cv)` : Abbr. for `criterion(cv)`, which specifies cross-validation as the selection criterion
- `fold(10)` : Specifies that cross-validation with 10 folds

5.2 Illustrate RCM with Covariates and Placebo Test

```
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 10-fold cross-validation)

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

K	AICc	AIC	BIC	CV MSE	R-squared	lambda	Operation
1	**526.5586**	**525.6355**	**529.8391**	**2.187e+07**	**0.0888**	**4287.1880**	add **gdp·Italy**
2	**506.8945**	**505.2945**	**510.8993**	**1.038e+07**	**0.5668**	**2954.9924**	add **gdp·Netherlands**
3	**504.0514**	**501.5514**	**508.5574**	**8.616e+06**	**0.6403**	**2692.4790**	add **gdp·Austria**
4	**484.0258**	**480.3737**	**488.7808**	**4.061e+06**	**0.8289**	**1855.8213**	add **gdp·Denmark**
5	**478.7213**	**473.6304**	**483.4388**	**3.058e+06**	**0.8706**	**1614.0987**	add **gdp·USA**
7	**407.2748**	**398.2748**	**410.8855**	**2.420e+05**	**0.9894**	**459.7011**	add **gdp·Greece gdp·Norway**
6	**357.2235**	**350.3663**	**361.5759**	**58902.4925**	**0.9976**	**218.3953**	drop **gdp·Netherlands**
7	**334.6504**	**325.6504**	**338.2611**	**29433.5466**	**0.9988**	**150.5314**	add **gdp·Switzerland**
8	**328.6835**	**317.1046**	**331.1166**	**25005.4310**	**0.9990**	**137.1586**	add **industry·Spain**
9	**285.8035**	**271.1368**	**286.5500**	**6852.9127**	**0.9998**	**59.3727**	add **gdp·Netherlands**
9	**270.7571**	**256.0904**	**271.5036**	**4898.0417**	**0.9999**	**44.9133**	**.**

Among models with **1-67** predictors, the optimal model contains **9** predictors with **CV MSE = 4898.0417.**

5.2 Illustrate RCM with Covariates and Placebo Test

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

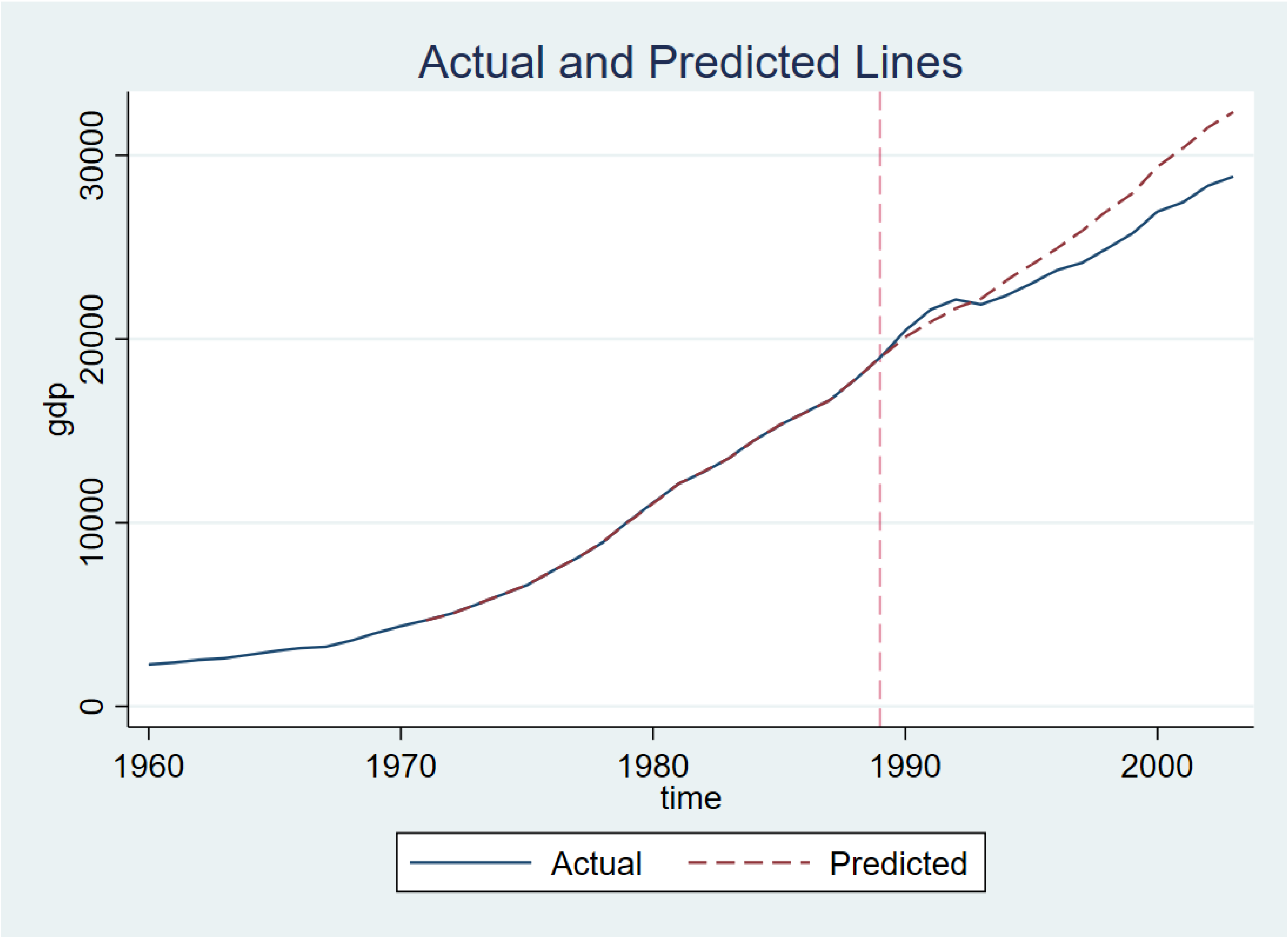
gdp·WestGermany	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gdp·Austria	0.1331	0.0708	1.88	0.093	-0.0271	0.2934
gdp·Denmark	0.1046	0.0788	1.33	0.217	-0.0735	0.2828
gdp·Greece	0.1513	0.0422	3.58	0.006	0.0557	0.2468
gdp·Italy	0.2914	0.0838	3.48	0.007	0.1018	0.4809
gdp·Netherlands	0.1334	0.1097	1.22	0.255	-0.1148	0.3816
gdp·Norway	-0.0313	0.0598	-0.52	0.613	-0.1665	0.1039
industry·Spain	-44.6926	11.6744	-3.83	0.004	-71.1019	-18.2833
gdp·Switzerland	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

5.2 Illustrate RCM with Covariates and Placebo Test

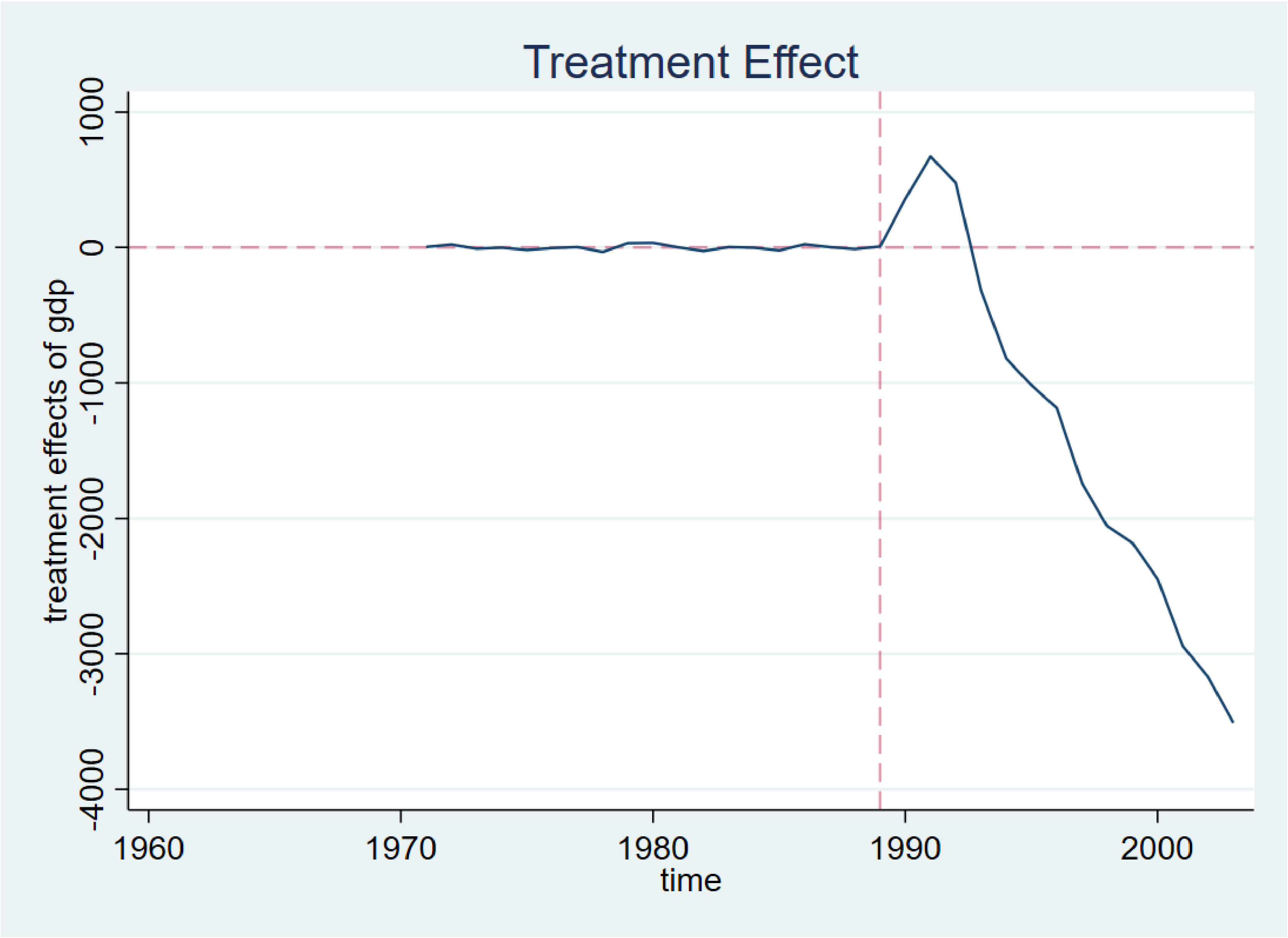
Prediction results in the post-treatment periods using **post-lasso OLS**:

Time	Treated	Predicted	Tr. 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**	**24052.6563**	**-1017.6563**
1996	**23742.0000**	**24926.1309**	**-1184.1309**
1997	**24156.0000**	**25896.0313**	**-1740.0313**
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**.



Actual and Predicted Lines
gdp
30000
20000
10000
0
1960
1970
1980
1990
2000
time
Actual
Predicted



Treatment Effect
1000
0
-1000
-2000
-3000
-4000
treatment effects of gdp
1960
1970
1980
1990
2000
time

5.2 Illustrate RCM with Covariates and Placebo Test

```
rcm gdp infrate trade industry, tru(17) trp(1990) me(lasso) cr(cv) fold(10) fill(mean) placebo(unit cut(10))
```

- `fill(mean)` : Fill in missing values by sample means for each units
- `placebo(unit cut(10))` :
  - Implement placebo tests using the fake treatment units
  - Discard the fake treatment units of which pre-treatment MSPE 10 times smaller than or equal to that of the treated unit

5.2 Illustrate RCM with Covariates and Placebo Test

```
Implementing placebo effects using fake treatment unit Australia...Austria...Belgium...Denmark...France...Greece...Italy...Japan...Netherlands...Ne
> wZealand...Norway...Portugal...Spain...Switzerland...UK...USA...

Placebo test results using fake treatment units:
```

Unit	Pre MSPE	Post MSPE	Post/Pre MSPE	Pre MSPE of Unit/Treated Unit
WestGermany	2009.4254	4912795.2300	2444.8757	1.0000
Austria	6616.8915	191860.4677	28.9956	3.2929
Belgium	2640.2164	81819.2402	30.9896	1.3139
France	3509.5265	241028.1689	68.6783	1.7465
Greece	16088.6537	1218658.0663	75.7464	8.0066
Italy	2180.6256	181301.3038	83.1419	1.0852
Japan	10414.5510	2085867.0714	200.2839	5.1829
Netherlands	5532.6214	1278504.6038	231.0848	2.7533
Portugal	7468.8217	858321.5550	114.9206	3.7169
Spain	8020.4273	1327015.9387	165.4545	3.9914
UK	6950.5984	443053.1737	63.7432	3.4590
USA	14405.5364	5967054.2037	414.2195	7.1690

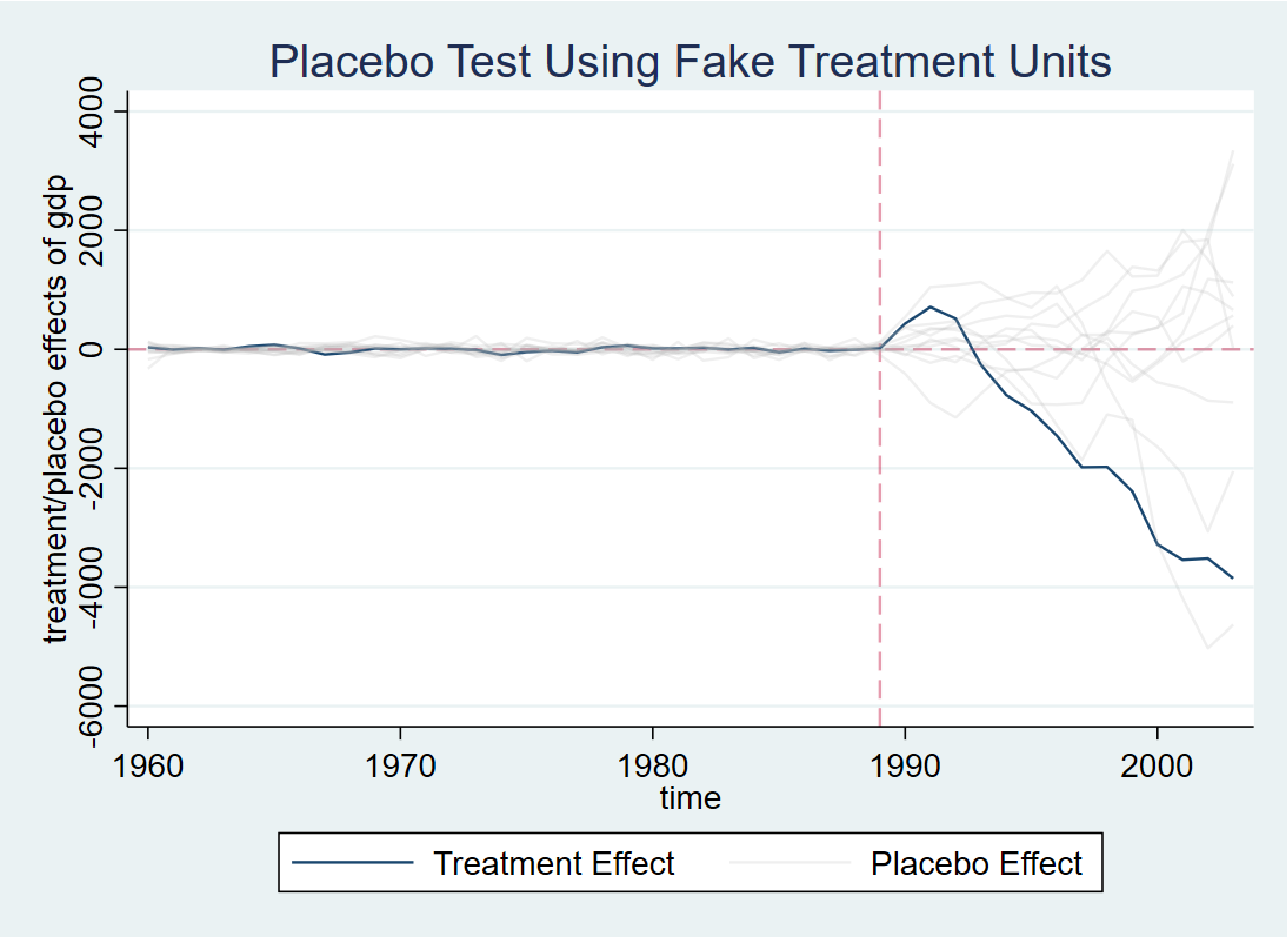
```
Note: The units Australia Denmark NewZealand Norway Switzerland with Pre-Treatment MSPE 10 times higher than WestGermany's are excluded.
      The probability of obtaining a post/pre-treatment MSPE ratio as large as WestGermany's is 0.0833.
```

5.2 Illustrate RCM with Covariates and Placebo Test

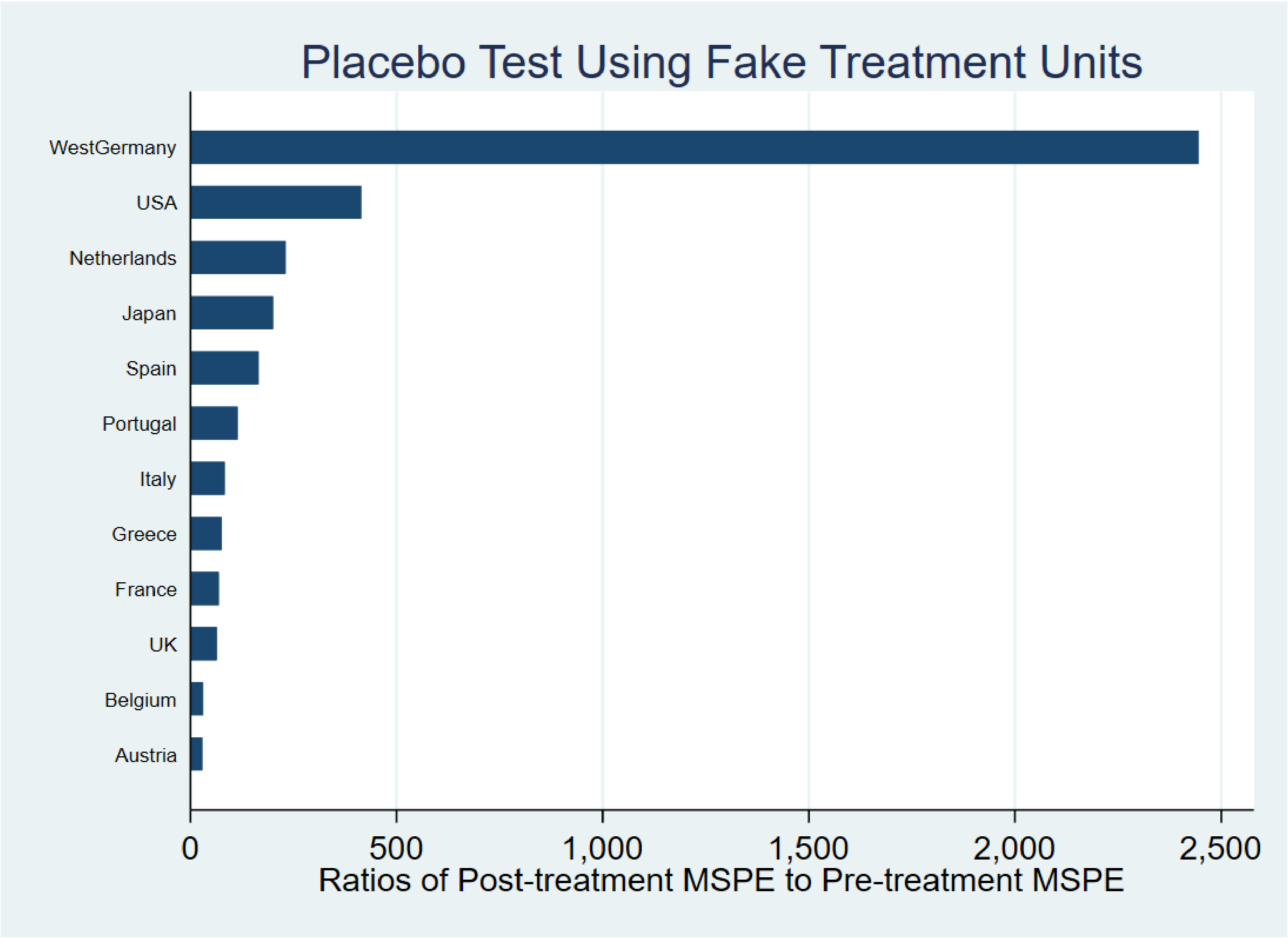
Placebo test results using fake treatment units (continued):

Time	Tr. Eff.	P-value
1990	**432.1543**	**0.1667**
1991	**709.9277**	**0.2500**
1992	**515.8379**	**0.2500**
1993	**-256.7402**	**0.5000**
1994	**-766.9336**	**0.2500**
1995	**-1035.2422**	**0.0833**
1996	**-1448.2324**	**0.0833**
1997	**-1981.4160**	**0.0833**
1998	**-1975.5234**	**0.0833**
1999	**-2385.0527**	**0.0833**
2000	**-3280.4258**	**0.0833**
2001	**-3540.6426**	**0.1667**
2002	**-3515.3242**	**0.1667**
2003	**-3850.5918**	**0.1667**

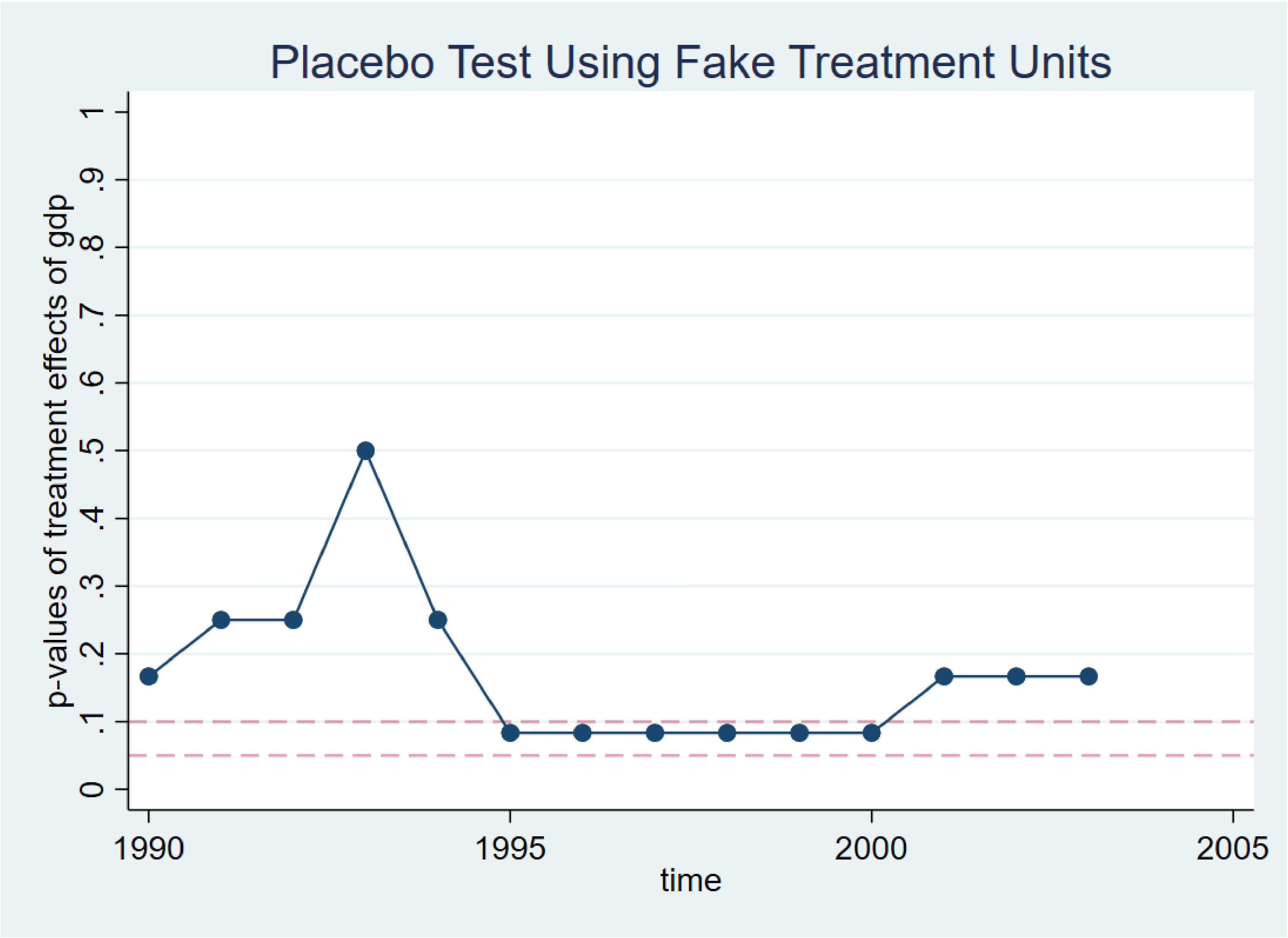
Note: The p-value of the treatment effect for a particular period is definded as the frequency that the absolute values of the placebo effects are greater or equal to the absolute value of treatment effect.



Placebo Test Using Fake Treatment Units
treatment/placebo effects of gdp
4000
2000
0
-2000
-4000
-6000
1960
1970
1980
1990
2000
time
Treatment Effect
Placebo Effect



Placebo Test Using Fake Treatment Units
WestGermany
USA
Netherlands
Japan
Spain
Portugal
Italy
Greece
France
UK
Belgium
Austria
0
500
1,000
1,500
2,000
2,500
Ratios of Post-treatment MSPE to Pre-treatment MSPE


Placebo Test Using Fake Treatment Units
p-values of treatment effects of gdp
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
1990
1995
2000
2005
time

5.2 Illustrate RCM with Covariates and Placebo Test

```
rcm gdp infrate trade industry, tru(17) trp(1990) me(lasso) cr(cv) fold(10) fill(linear) placebo(period(1980 1985))
```

- `fill(linear)` : Fill in missing values by linear interpolation for each units
- `placebo(period(1980 1985))` : Implement placebo tests with fake treatment time 1980 and 1985

5.2 Illustrate RCM with Covariates and Placebo Test

```
Implementing placebo effects using fake treatment time 1980...1985...

Placebo test results using fake treatment time 1980:
```

Time	Treated	Predicted	Tr. Eff.
1980	11083.0000	11062.1953	20.8047
1981	12115.0000	11937.9014	177.0986
1982	12761.0000	12584.5449	176.4551
1983	13519.0000	13487.0322	31.9678
1984	14481.0000	14485.5996	-4.5996
1985	15291.0000	15414.0498	-123.0498
1986	15998.0000	15827.1543	170.8457
1987	16679.0000	16503.3066	175.6934
1988	17786.0000	17469.7988	316.2012
1989	18994.0000	18590.4004	403.5996
1990	20465.0000	19414.7520	1050.2480
1991	21602.0000	20036.1699	1565.8301
1992	22154.0000	20710.7656	1443.2344
1993	21878.0000	21346.8164	531.1836
1994	22371.0000	22564.2402	-193.2402
1995	23035.0000	23573.0605	-538.0605
1996	23742.0000	25172.0996	-1430.0996
1997	24156.0000	26662.7520	-2506.7520
1998	24931.0000	27212.3789	-2281.3789
1999	25755.0000	28998.8145	-3243.8145
2000	26943.0000	32415.6914	-5472.6914
2001	27449.0000	33603.8672	-6154.8672
2002	28348.0000	34962.1016	-6614.1016
2003	28855.0000	36147.2969	-7292.2969
Mean	20432.9583	21674.2829	-1241.3246

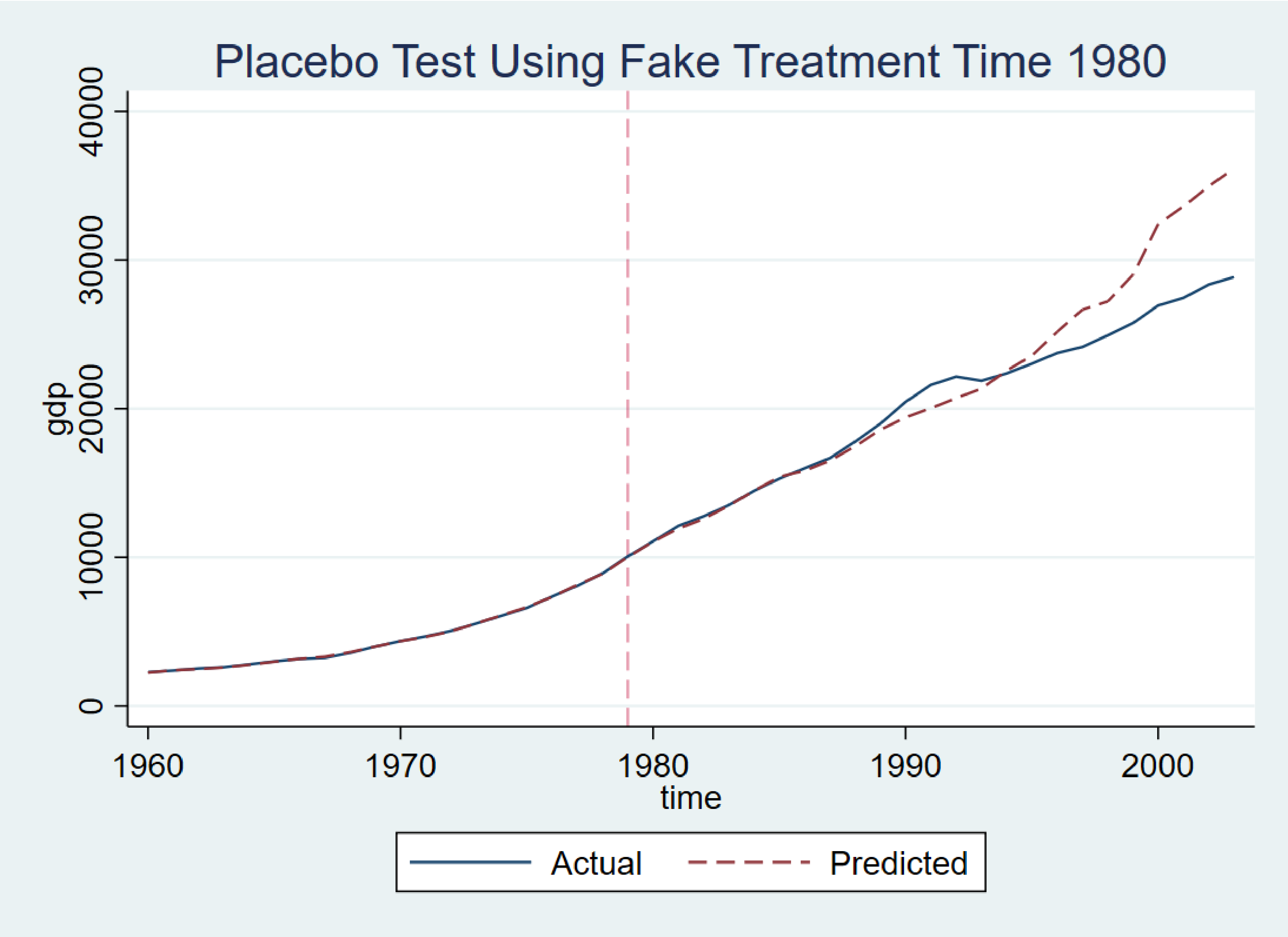
```
Note: The average treatment effect over the post-treatment periods is -1241.3246.
```

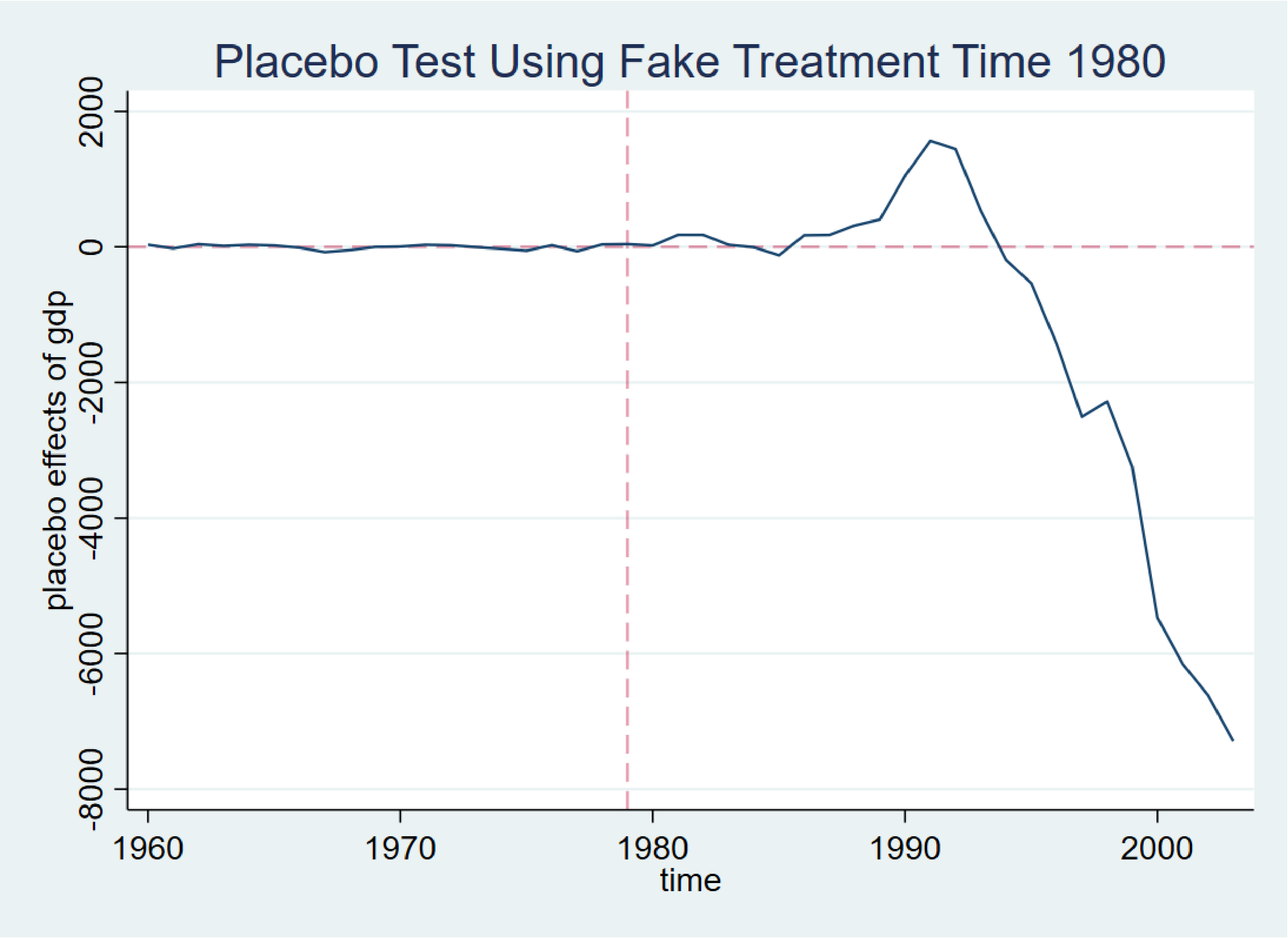
5.2 Illustrate RCM with Covariates and Placebo Test

Placebo test results using fake treatment time **1985**:

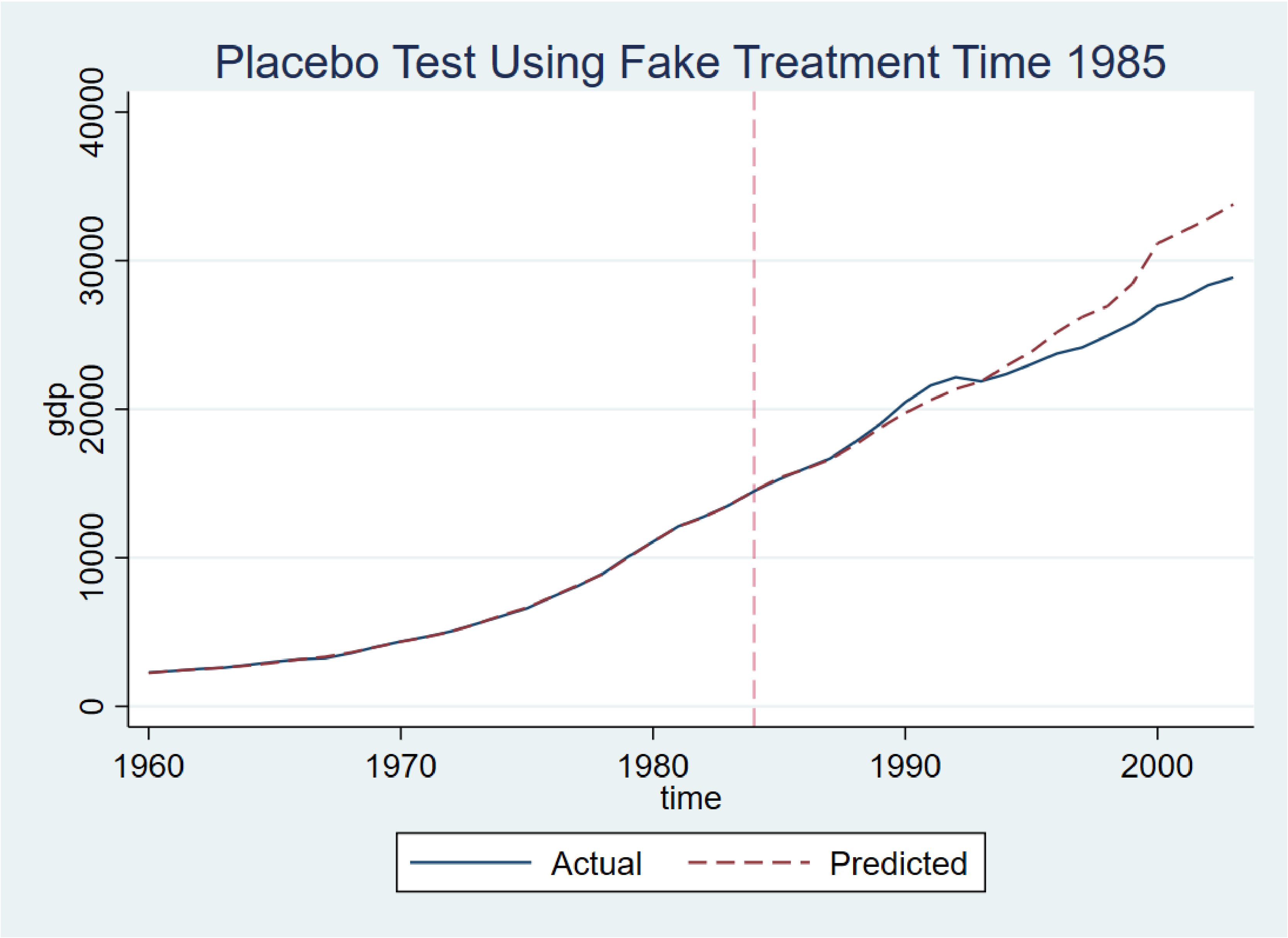
Time	Treated	Predicted	Tr. Eff.
1985	**15291.0000**	**15399.4453**	**-108.4453**
1986	**15998.0000**	**15950.9844**	**47.0156**
1987	**16679.0000**	**16604.2344**	**74.7656**
1988	**17786.0000**	**17587.4297**	**198.5703**
1989	**18994.0000**	**18738.6387**	**255.3613**
1990	**20465.0000**	**19745.6875**	**719.3125**
1991	**21602.0000**	**20602.5215**	**999.4785**
1992	**22154.0000**	**21363.2012**	**790.7988**
1993	**21878.0000**	**21871.5781**	**6.4219**
1994	**22371.0000**	**22915.0332**	**-544.0332**
1995	**23035.0000**	**23870.9434**	**-835.9434**
1996	**23742.0000**	**25165.6582**	**-1423.6582**
1997	**24156.0000**	**26213.7539**	**-2057.7539**
1998	**24931.0000**	**26917.2168**	**-1986.2168**
1999	**25755.0000**	**28444.0781**	**-2689.0781**
2000	**26943.0000**	**31154.6328**	**-4211.6328**
2001	**27449.0000**	**31974.5781**	**-4525.5781**
2002	**28348.0000**	**32806.0039**	**-4458.0039**
2003	**28855.0000**	**33786.4219**	**-4931.4219**
Mean	**22443.7895**	**23742.7390**	**-1298.9495**

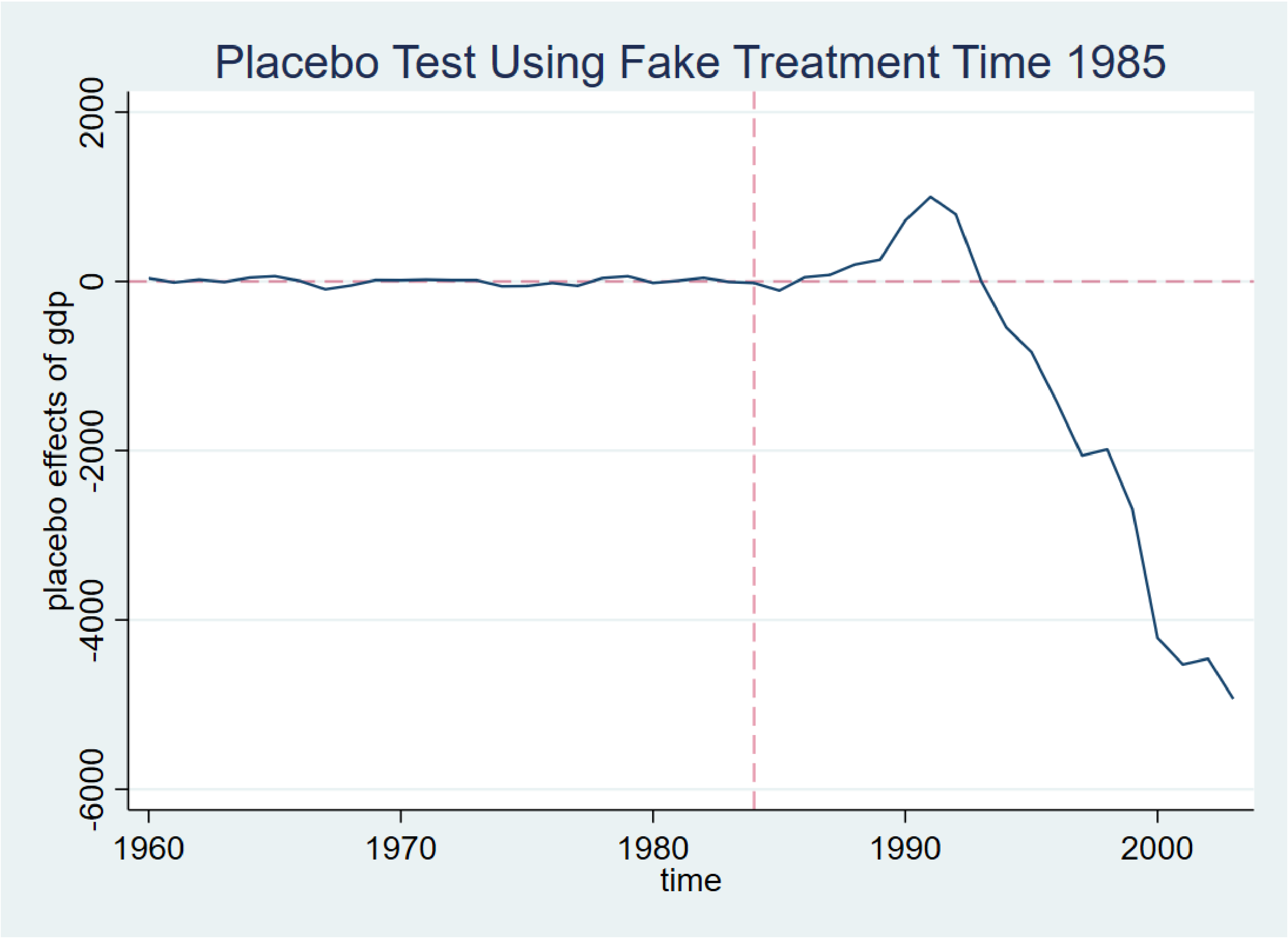
Note: The average treatment effect over the post-treatment periods is **-1298.9495**.


Placebo Test Using Fake Treatment Time 1980
gdp
0
10000
20000
30000
40000
1960
1970
1980
1990
2000
time
Actual
Predicted



Placebo Test Using Fake Treatment Time 1980
placebo effects of gdp
2000
0
-2000
-4000
-6000
-8000
1960
1970
1980
1990
2000
time


Placebo Test Using Fake Treatment Time 1985
gdp
0
10000
20000
30000
40000
1960
1970
1980
1990
2000
time
Actual
Predicted



Placebo Test Using Fake Treatment Time 1985
placebo effects of gdp
2000
0
-2000
-4000
-6000
1960
1970
1980
1990
2000
time

Reference

- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. Comparative Politics and the Synthetic Control Method. American Journal of Political Science 59(2): 495-510.
- Hsiao, Cheng, H. Steve Ching, and Shui Ki 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(5): 705-740.
- Hsiao, Cheng, and Qiankun Zhou. 2019. Panel Parametric, Semiparametric, and Nonparametric Construction of Counterfactuals. Journal of Applied Econometrics 34(4): 463-481.

Reference

- Furnival, George M., and Robert W. Wilson, Jr. 1974. Regressions by Leaps and Bounds. Technometrics 16(4): 499-511.
- Guanpeng Yan, and Qiang Chen. 2021. Regression Control Method with Stata. Shandong University working paper.

Thank for Listening

School of Economics
Shandong University
颜冠鹏
Guanpengyan@mail.sdu.edu.cn