分位数控制法及Stata应用

Quantile Control Method (QCM) with Stata

陈强
山东大学经济学院
qiang2chen2@126.com
www.econometrics-stata.com

Outline

1. Introduction
2. Model
3. Point Estimations
4. Confidence Interval
5. Quantile Random Forest (QRF)
6. Simulations
7. Stata Application with qcm Command: Economic Integration of HK with China

References

- Chen, Qiang, Zhijie Xiao and Qingsong Yao, 2021, “Robust Nonparametric Confidence Intervals for Treatment Effects in Panel Data Using Quantile Random Forest,” *Shandong University working paper*.

- Yan, Guanpeng, and Qiang Chen, 2021, “Quantile Control Method with Stata,” *Shandong University working paper*.

- Yan, Guanpeng, and Qiang Chen, 2021, “qcm: A R Package for Quantile Control Method,” *Shandong University working paper*.

1. Introduction

- Approaches to estimate treatment effects in panel data with only one treated unit have become popular in applied works, which include synthetic control method (Abadie and Gardeazabal, 2003, Abadie et al., 2010), and regression control method (Hsiao et al., 2012).

- However, no pointwise standard errors or confidence intervals for the treatment effects have been provided or rigorously proven in the literature yet.

Our Contributions

- We propose a direct nonparametric construction of pointwise robust confidence intervals using quantile random forest (QRF), i.e. quantile regression via random forest.
- This is called "Quantile Control Method"(QCM).
- Monte Carlos simulations show good coverage probability for the confidence intervals, which are robust to heteroskedasticity, autocorrelation, and model misspecification.

2. Model

- We observe panel data with an outcome variable $\{y_{it}\}$ for individuals $i=1,\cdots,N$ (often known as "regions" in regional policy analysis), and over periods

$$t = \underbrace{1,\cdots,T_0}_{pre-treatment}, \underbrace{T_0+1,\cdots,T_0+T_1}_{post-treatment}$$

- T_0 is the number of pretreatment periods (from period 1 through period T_0), and T_1 is the number of posttreatment periods (from period T_0+1 through period T_0+T_1).

- Assume that the first individual is the only treated unit, while all other individuals are control units.

Potential Outcomes

- Following Rubin's causal model, denote y_{it}^{1} as the potential outcome with treatment, and y_{it}^{0} as the potential outcome without treatment.
- Following Hsiao et al. (2012), assume a linear factor model for the potential outcome without treatment

$$y_{it}^{0} = \alpha_i + \mathbf{b}_i'\mathbf{f}_t + u_{it} \quad (i = 1, \cdots, N; t = 1, \cdots, T_0 + T_1)$$

- For a particular pretreatment period $t = \{1, \cdots, T_0\}$, stack the above equations for all individuals:

$$\mathbf{y}_t = \mathbf{y}_t^{0} = \boldsymbol{\alpha} + \mathbf{B}\mathbf{f}_t + \mathbf{u}_t \qquad (t = 1, \cdots, T_0)$$

Interactive Fixed Effects

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

- $\mathbf{f}_t$ 为不可观测的共同因子（common factors），可理解为不同地区所面临的共同冲击（common shocks），比如两个分量分别表示技术冲击（technological shocks）与金融危机（financial shocks）

- 各地区对于共同冲击 $\mathbf{f}_t$ 的反应不相同，以向量 $\mathbf{b}_i$ 表示。称 $\mathbf{b}_i'\mathbf{f}_t$ 为“交互固定效应”（interactive fixed effects）

Partitions

- To derive the relationship between the first and the rest units, do the following partitions:

$$\mathbf{y}_t = \begin{pmatrix} y_{1t} \\ \tilde{\mathbf{y}}_t \end{pmatrix}, \quad \text{where} \quad \tilde{\mathbf{y}}_t = \begin{pmatrix} y_{2t} \\ \vdots \\ y_{Nt} \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1' \\ \tilde{\mathbf{B}} \end{pmatrix}, \text{ where } \quad \mathbf{u}_t = \begin{pmatrix} u_{1t} \\ \tilde{\mathbf{u}}_t \end{pmatrix}$$

3. Estimation

- Hsiao et al. (2012) proposes multiplying the equation by some row vector $\mathbf{a}'$ to get rid of the interactive term $\mathbf{Bf}_t$

- Li and Bell (2017) offers a particular solution of $\mathbf{a}'$:

$$\mathbf{a}' \equiv \begin{pmatrix} 1 & -\boldsymbol{\gamma}' \end{pmatrix}, \text{ where } \boldsymbol{\gamma}' \equiv \mathbf{b}_1'(\tilde{\mathbf{B}}'\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{B}}'$$

Counterfactuals

- We end up with

$$y_{1t} = \delta_1 + \boldsymbol{\delta}'\tilde{\mathbf{y}}_t + \varepsilon_{1t} \quad (t = 1, \cdots, T_0)$$

where ε_{1t} is uncorrelated with $\tilde{\mathbf{y}}_t$.

- Assume that the linear factor model are stable, we could estimate the post-treatment counterfactual outcome by

$$\hat{y}_{1t}^0 = \hat{\delta}_1 + \hat{\boldsymbol{\delta}}'\tilde{\mathbf{y}}_t \quad (t = T_0 + 1, \cdots, T_0 + T_1)$$

Treatment Effects

- The treatment effect on unit 1 in period t is estimated by

$$\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^{0} \quad (t = T_0 + 1, \cdots, T_0 + T_1)$$

- Average treatment effect on unit 1 from period 1 to period T_1 can be estimated as

$$\hat{\Delta}_{1} = \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \hat{\Delta}_{1t}$$

Point Estimations

- Hsiao et al. (2012) uses AIC and AICC criteria for best subset selection, followed by OLS regression. Li and Bell (2017) proposes using lasso for variable selection in the high-dimensional case, and estimate by post-lasso OLS. Carvalho et al. (2018) proposes a direct estimation by lasso, ("ArCo" for "Artificial Counterfactual ").
- Possible approaches: best subset OLS, post-lasso OLS, lasso, ridge regression, elastic net regression, random forest, boosting, or neural networks. Which approach provides the best point estimate with the lowest MSE is still an unsettled issue, which might be problem-specific and data-dependent.

4. Inference

- No pointwise standard errors or confidence intervals have been provided yet. A notable exception is Fujiki and Hsiao (2015), which relies on a strong *iid* assumption.
- Li and Bell (2017) and Carvalho et al. (2018) provide large sample inference on the estimated average treatment effect, by letting both pretreatment periods T_0 and posttreatment periods T_1 tend to infinity.
- While a moderate or large T_0 is needed for any inference, the requirement of large T_1 may be problematic in practice

Case Study in Hsiao et al. (2012): Political Integration of HK with China in 1997Q3

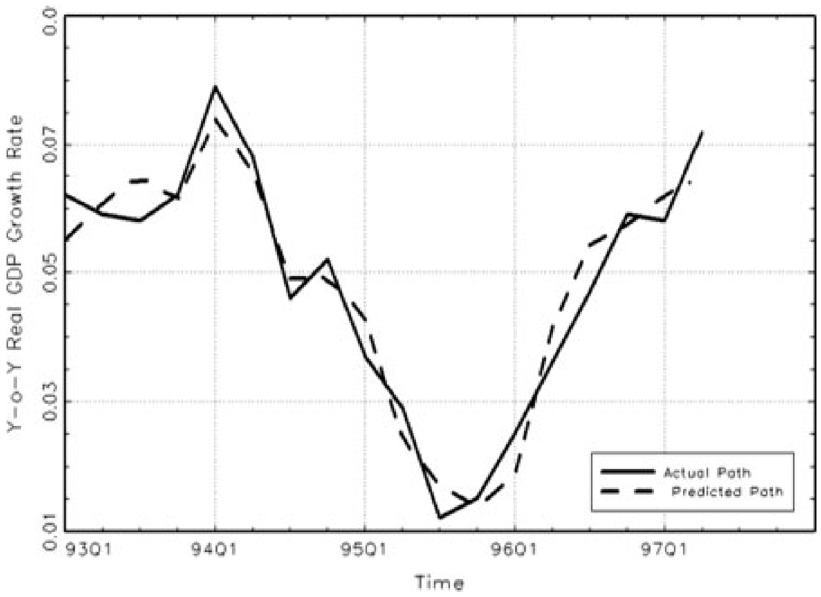


Figure 1. AICC: actual and predicted real GDP from 1993:Q1 to 1997:Q2

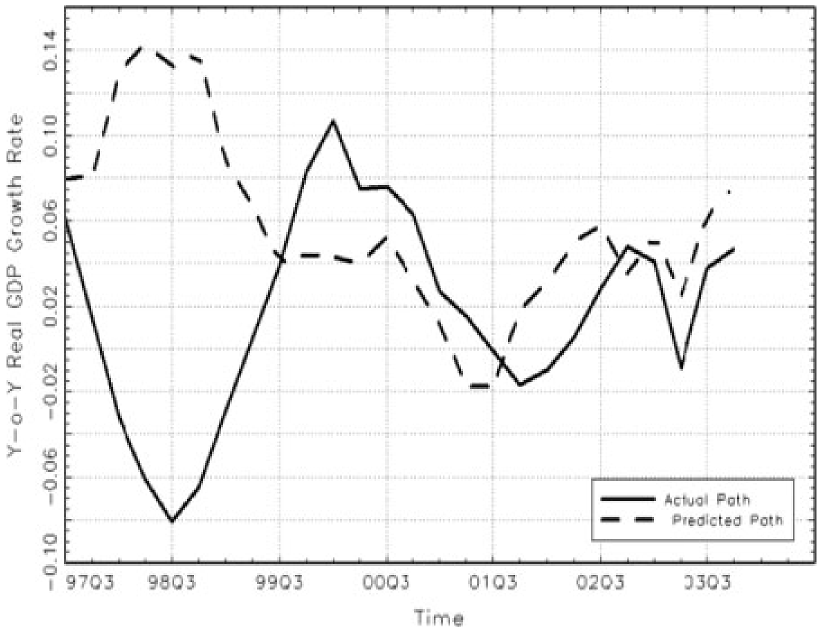


Figure 2. AICC : actual and counterfactual real GDP from 1997:Q3 to 2003:Q4

Case Study in Hsiao et al. (2012): Economic Integration of HK with China in 2004Q1

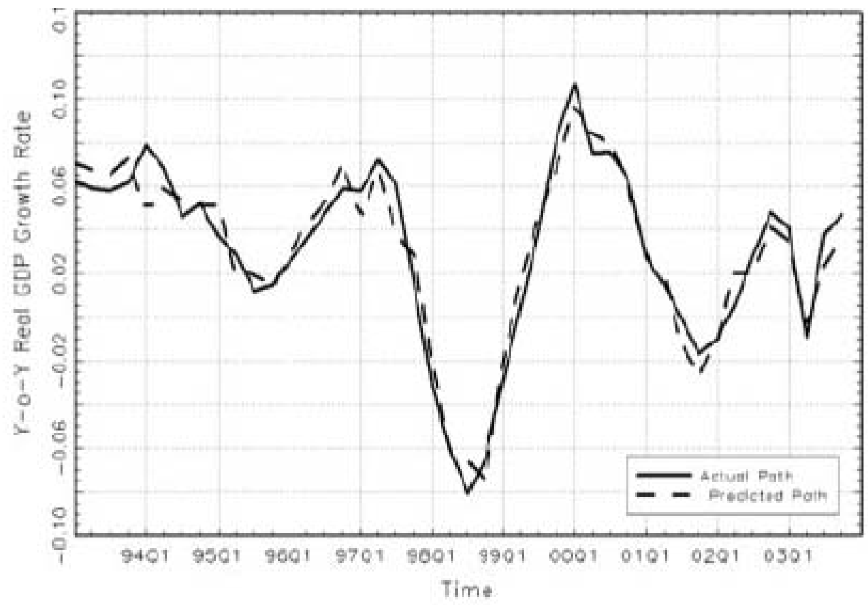


Figure 7. AICC: actual and predicted real GDP from 1993:Q1 to 2003:Q4

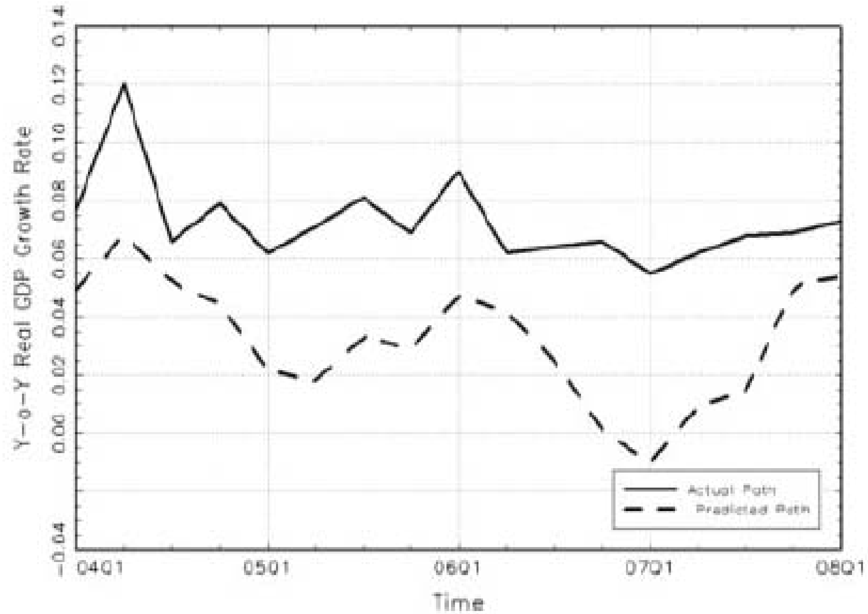


Figure 8. AICC: actual and counterfactual real GDP from 2004:Q1 to 2008:Q1

A Quantile Regression Approach

- Inspired by Zhou and Portnoy (1996), we use regression quantiles of the post-treatment counterfactual outcome to directly construct valid confidence intervals of the treatment effects

$$P\left(Q_{y_{1t}^0}(\alpha/2) \le y_{1t}^0 \le Q_{y_{1t}^0}(1-\alpha/2)\right) = 1-\alpha$$

- Since $\Delta_{1t} = y_{1t}^1 - y_{1t}^0$, we could plug in the expression $y_{1t}^0 = y_{1t}^1 - \Delta_{1t}$ to get

$$P\left(Q_{y_{1t}^0}(\alpha/2) \le y_{1t}^1 - \Delta_{1t} \le Q_{y_{1t}^0}(1-\alpha/2)\right) = 1-\alpha$$

Confidence Interval for Treatment Effects

- Since $y_{1t}^{1} = y_{1t}$ for $t = T_0 + 1, \cdots, T_0 + T_1$, rearranging ,

$$P\left(y_{1t} - Q_{y_{1t}^{0}}(1-\alpha/2) \le \Delta_{1t} \le y_{1t} - Q_{y_{1t}^{0}}(\alpha/2)\right) = 1-\alpha$$

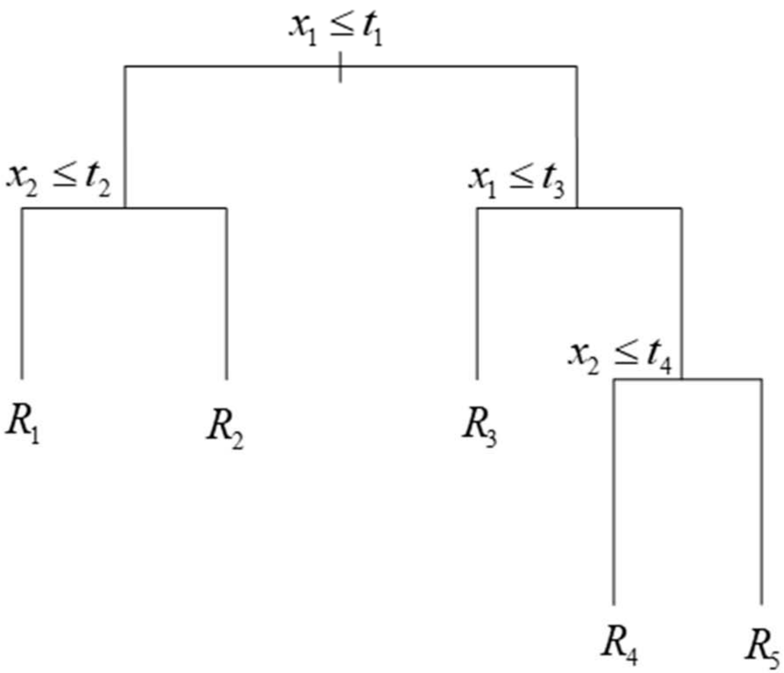
- With consistent estimators:

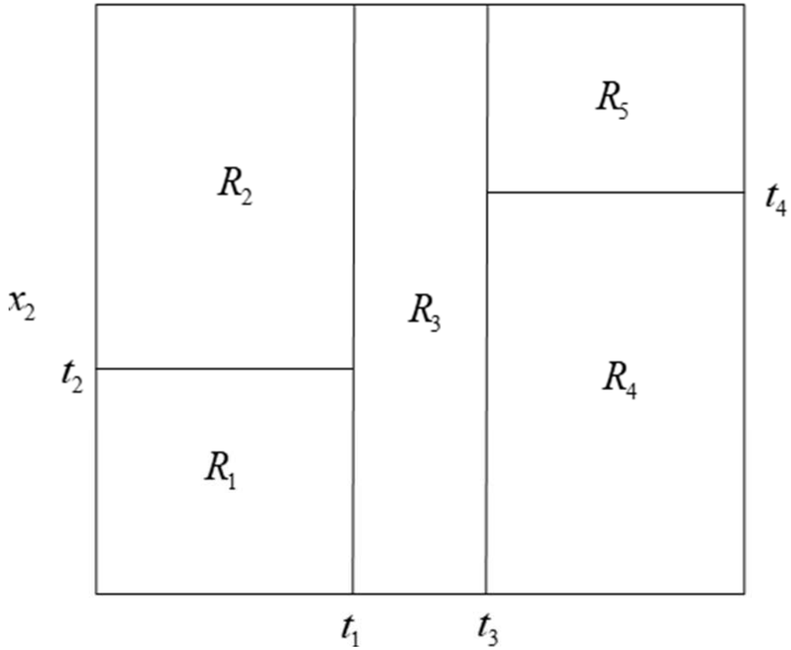
$$P\left(y_{1t} - \hat{Q}_{y_{1t}^{0}}(1-\alpha/2) \le \Delta_{1t} \le y_{1t} - \hat{Q}_{y_{1t}^{0}}(\alpha/2)\right) \xrightarrow{p} 1-\alpha$$

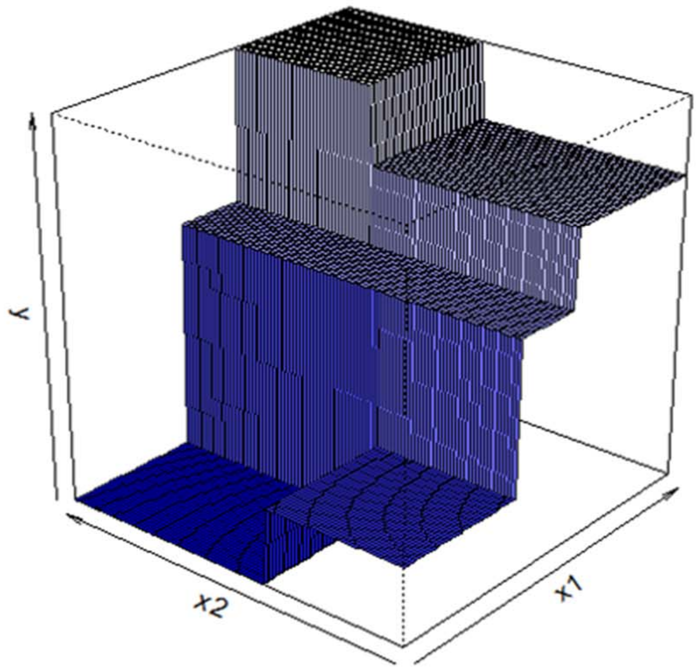
5. Quantile Random Forest (QRF)

- Linear quantile regression converges too slowly, and is not robust to model misspecification.
- QRF (also called "Quantile Regression Forest") uses random forest for quantile regression (Meinshausen, 2006)
- Random forests (Breiman, 2001) is basically an ensemble of decision trees, or the CART algorithm (Breiman et al., 1984)

Decision Tree







Regression Tree

- 将决策树应用于回归问题，则为回归树。
- 由于回归树的响应变量为连续变量，故使用最小化“残差平方和”（SSR）作为节点的分裂准则。
- 在进行节点分裂时，希望分裂后，残差平方和下降最多，即两个子节点的残差平方和之总和最小。

决策树的优缺点

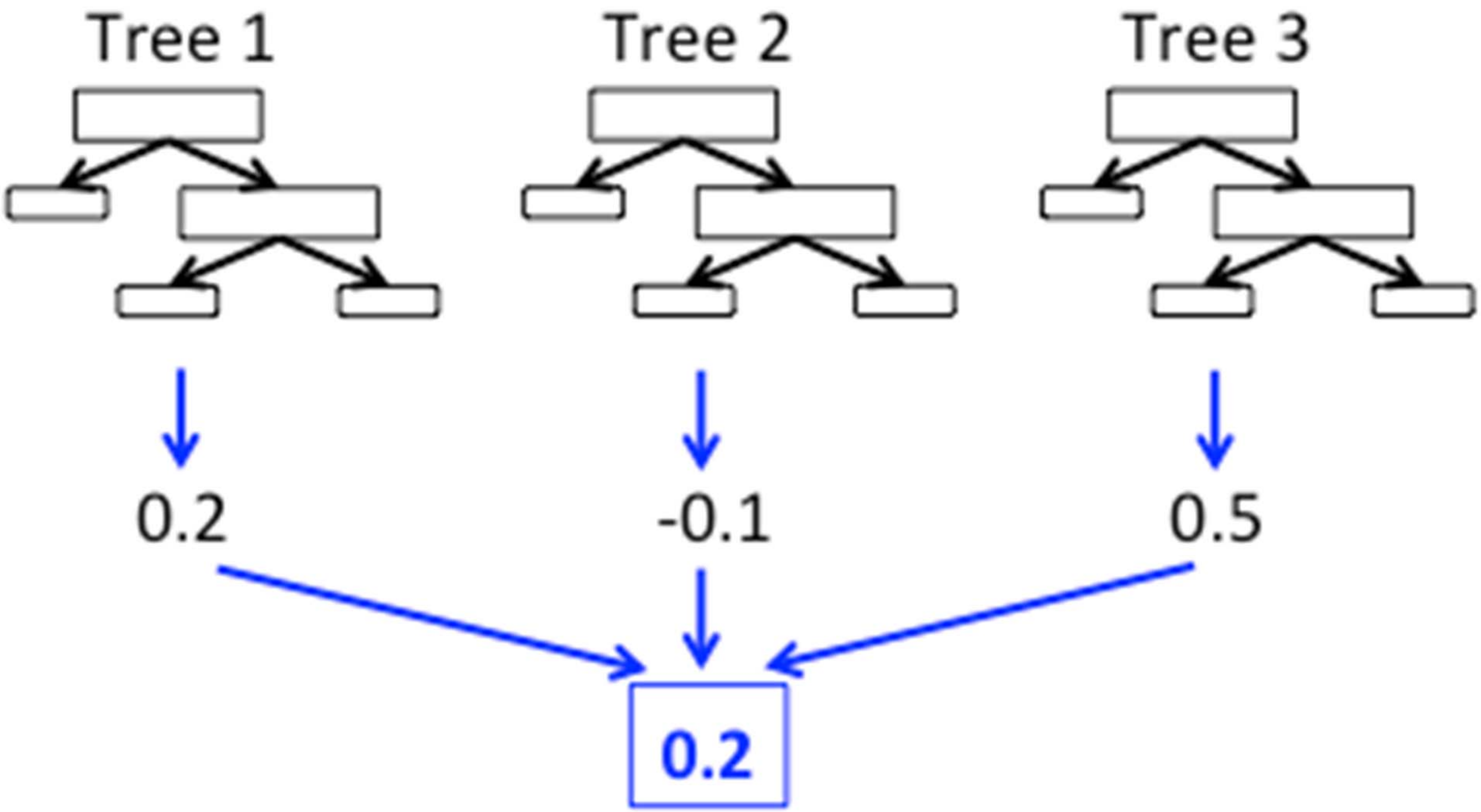
- 优点：由于每次仅使用一个分裂变量，故决策树天然适用于高维数据（high-dimensional data），且不受噪音变量的影响（噪音变量一般不会被选为分裂变量）。

- 缺点：作为“分段常值函数”（piecewise constant function），决策树既不光滑，也不连续，故单颗决策树（single tree）的预测效果可能不佳。

Random Forest

- Grow many trees (e.g., 1000 trees) by random feature selection, i.e., at each node of every tree, only 1/3 features are randomly chosen for splitting
- This helps to de-correlate trees for variance reduction
- Average over these 1000 trees for prediction

Ensemble Model:
example for regression



随机森林的变量重要性

- 由于在每次节点分裂，仅使用一个变量，故容易区分每个变量的贡献。
- 对于每个变量，在随机森林的每棵决策树，可度量由该变量所导致的残差平方之下降幅度。将此下降幅度，对每棵决策树进行平均，即为对该变量重要性的度量。
- 将每个特征变量的重要性依次排列画图，即为变量重要性图（Variable Importance Plot）

Consistency of QRF

- Meinshausen (2006) uses random forest to estimate conditional CDF, then invert it to get QR.
- Meinshausen (2006) provides a sketch of proof for the consistency of QRF under *iid* assumptions
- We extend the proof to the panel/time series case, see Chen, Xiao and Yao (2021).

6. Simulations

- We conduct simulations for the four DGPs in Hsiao et al. (2012)

- Also simulations for DGPs under heteroskedasticity, autocorrelation, and model misspecification.

- Good coverage probability in finite samples

Table 1. Coverage Probability for DGP 1

$$y_{it}^0 = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + u_{it} \qquad f_{1t} = 0.3 f_{1,t-1} + \varepsilon_{1t} \qquad f_{2t} = 0.6 f_{2,t-1} + \varepsilon_{2t}$$

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.803	0.897	0.926	0.922	0.936	0.934	0.948	0.943	0.951
$N = 20$	0.828	0.883	0.902	0.924	0.931	0.946	0.949	0.954	0.955
$N = 30$	0.794	0.9	0.919	0.931	0.951	0.943	0.966	0.96	0.962
$N = 40$	0.786	0.875	0.912	0.941	0.945	0.951	0.947	0.962	0.952
$N = 50$	0.811	0.904	0.911	0.933	0.947	0.957	0.935	0.957	0.955
$N = 60$	0.794	0.89	0.925	0.944	0.929	0.945	0.944	0.962	0.95

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period.
The nominal coverage rate is 95%.

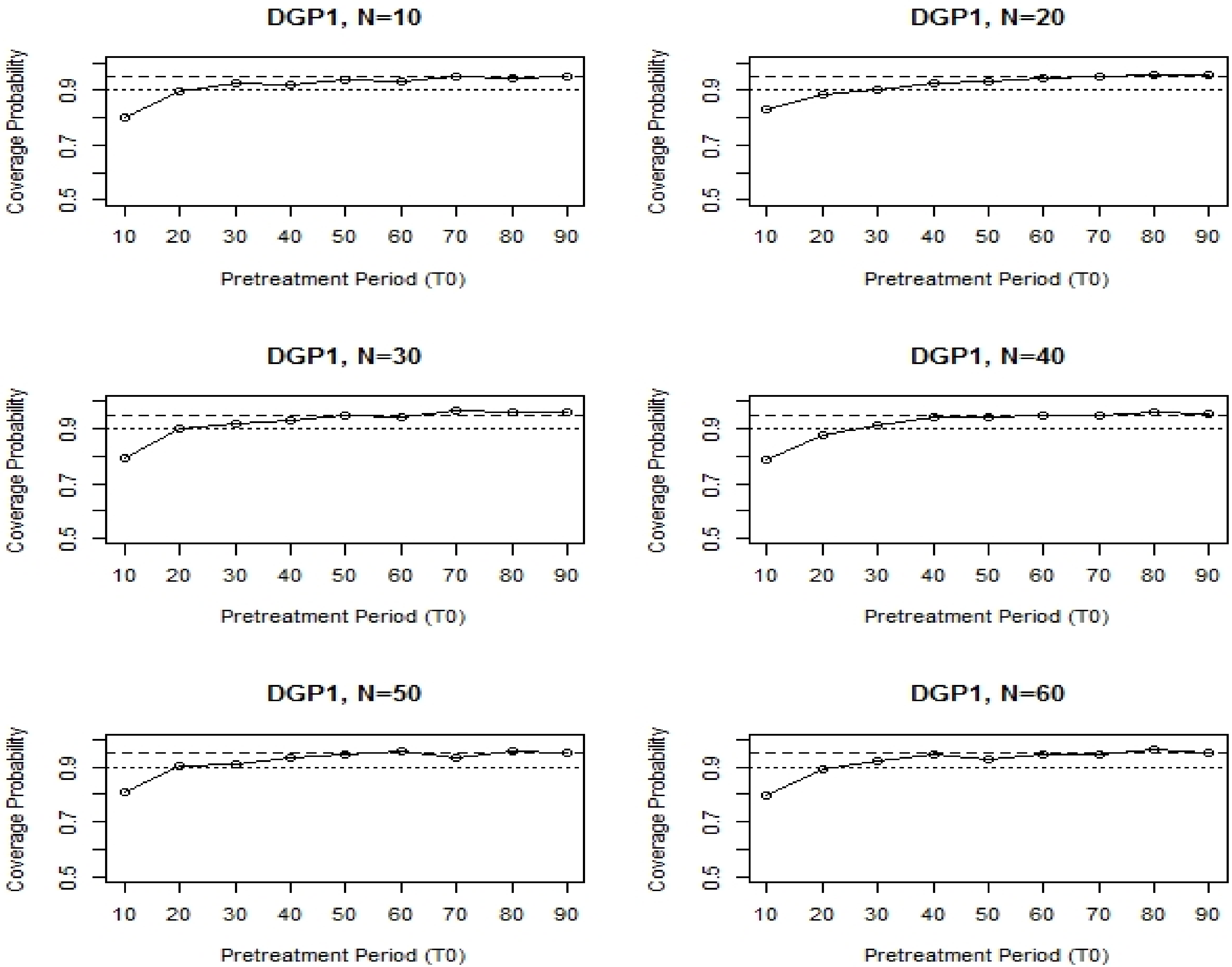

DGP1, N=10
DGP1, N=20
DGP1, N=30
DGP1, N=40
DGP1, N=50
DGP1, N=60
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)

Table 2. Coverage Probability for DGP 2

$$y_{it}^0 = \alpha_i + b_{i1} f_{1t} + b_{i2} f_{2t} + b_{i3} f_{3t} + u_{it} \qquad f_{1t} = 0.8 f_{1,t-1} + \varepsilon_{1t}$$

$$f_{2t} = -0.6 f_{2,t-1} + \varepsilon_{2t} + 0.8\varepsilon_{2,t-1} \qquad f_{3t} = \varepsilon_{3t} + 0.9\varepsilon_{3,t-1} + 0.4\varepsilon_{3,t-2}$$

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.781	0.883	0.921	0.917	0.948	0.938	0.955	0.957	0.949
$N = 20$	0.789	0.878	0.924	0.927	0.947	0.951	0.95	0.965	0.963
$N = 30$	0.765	0.875	0.907	0.938	0.936	0.951	0.956	0.956	0.95
$N = 40$	0.799	0.875	0.917	0.934	0.945	0.946	0.954	0.953	0.969
$N = 50$	0.786	0.875	0.918	0.941	0.949	0.947	0.963	0.967	0.955
$N = 60$	0.756	0.874	0.931	0.936	0.938	0.953	0.955	0.96	0.955

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.

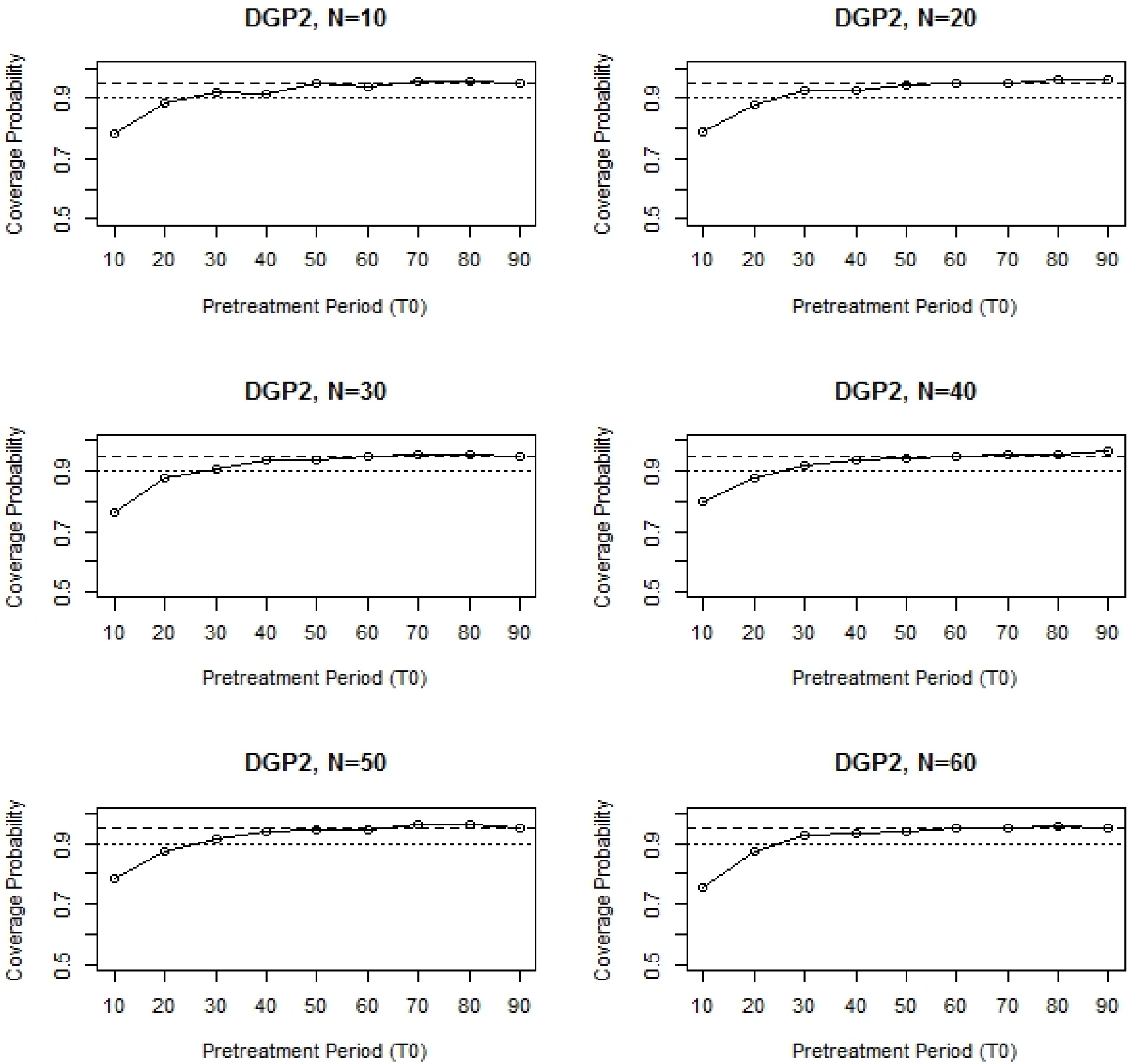
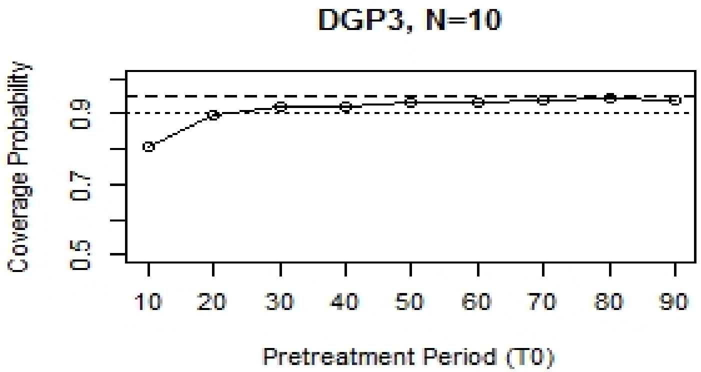

DGP2, N=10
DGP2, N=20
DGP2, N=30
DGP2, N=40
DGP2, N=50
DGP2, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

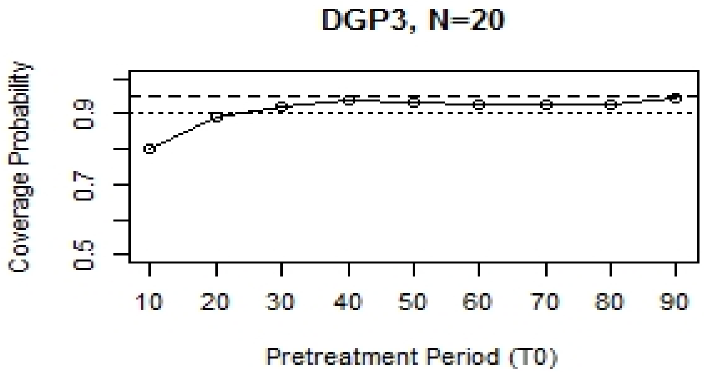
Table 3. Coverage Probability for DGP 3

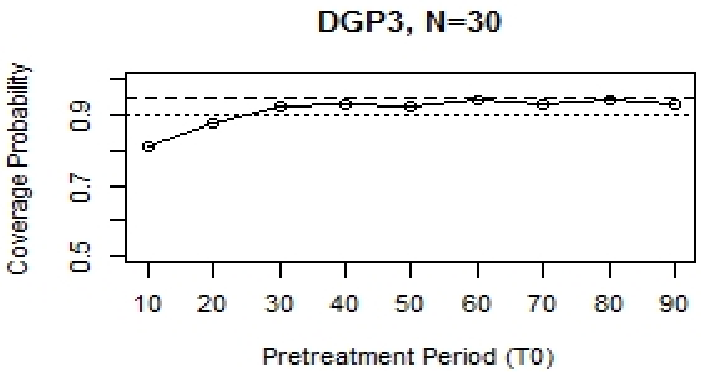
$$y_{it}^{0} = \alpha_i + b_i f_t + u_{it} \qquad f_t = i.i.d.\ N(0,1)$$

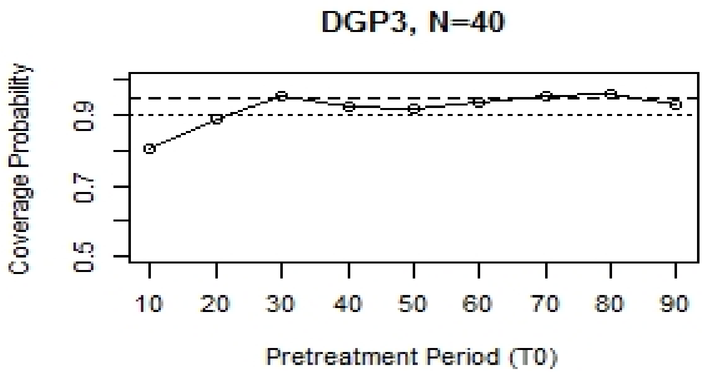
CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.804	0.897	0.921	0.921	0.935	0.934	0.94	0.942	0.938
$N = 20$	0.802	0.892	0.919	0.939	0.93	0.928	0.929	0.925	0.945
$N = 30$	0.813	0.876	0.927	0.929	0.923	0.941	0.933	0.945	0.929
$N = 40$	0.804	0.887	0.952	0.924	0.917	0.935	0.956	0.962	0.93
$N = 50$	0.827	0.896	0.92	0.926	0.935	0.948	0.941	0.938	0.951
$N = 60$	0.814	0.895	0.916	0.938	0.935	0.947	0.937	0.932	0.947

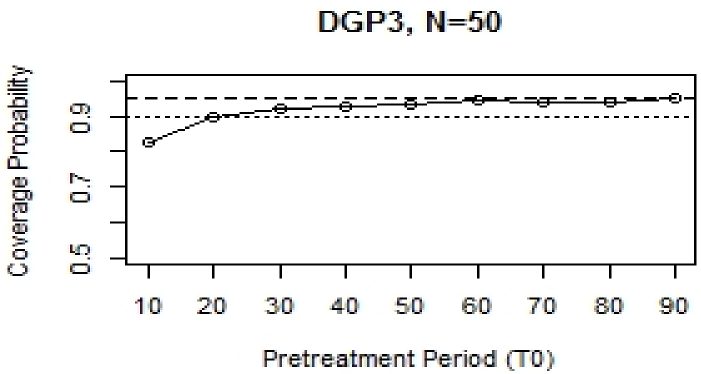
Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP3, N=10
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=20
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=30
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=40
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)


DGP3, N=50
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

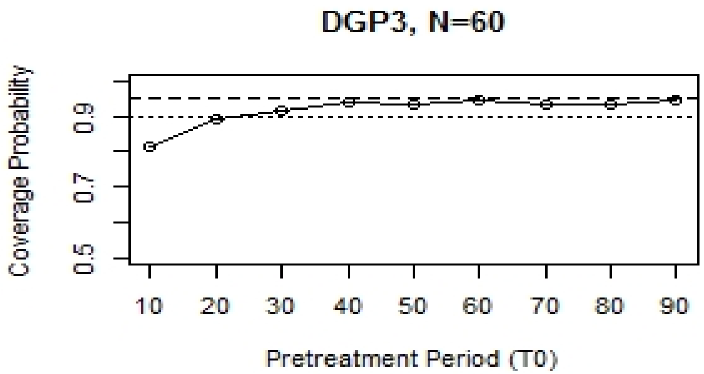

DGP3, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

Table 4. Coverage Probability for DGP 4

$$y_{it}^{0} = \alpha_i + b_i f_t + u_{it} \qquad f_t = 0.95 f_{t-1} + \varepsilon_{1t}$$

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.727	0.848	0.843	0.893	0.895	0.915	0.901	0.921	0.914
$N = 20$	0.723	0.835	0.868	0.887	0.889	0.926	0.919	0.921	0.925
$N = 30$	0.72	0.842	0.863	0.874	0.902	0.913	0.92	0.925	0.924
$N = 40$	0.758	0.817	0.863	0.877	0.899	0.892	0.933	0.931	0.92
$N = 50$	0.761	0.831	0.874	0.882	0.893	0.902	0.923	0.924	0.923
$N = 60$	0.748	0.816	0.875	0.889	0.899	0.925	0.907	0.913	0.939

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.

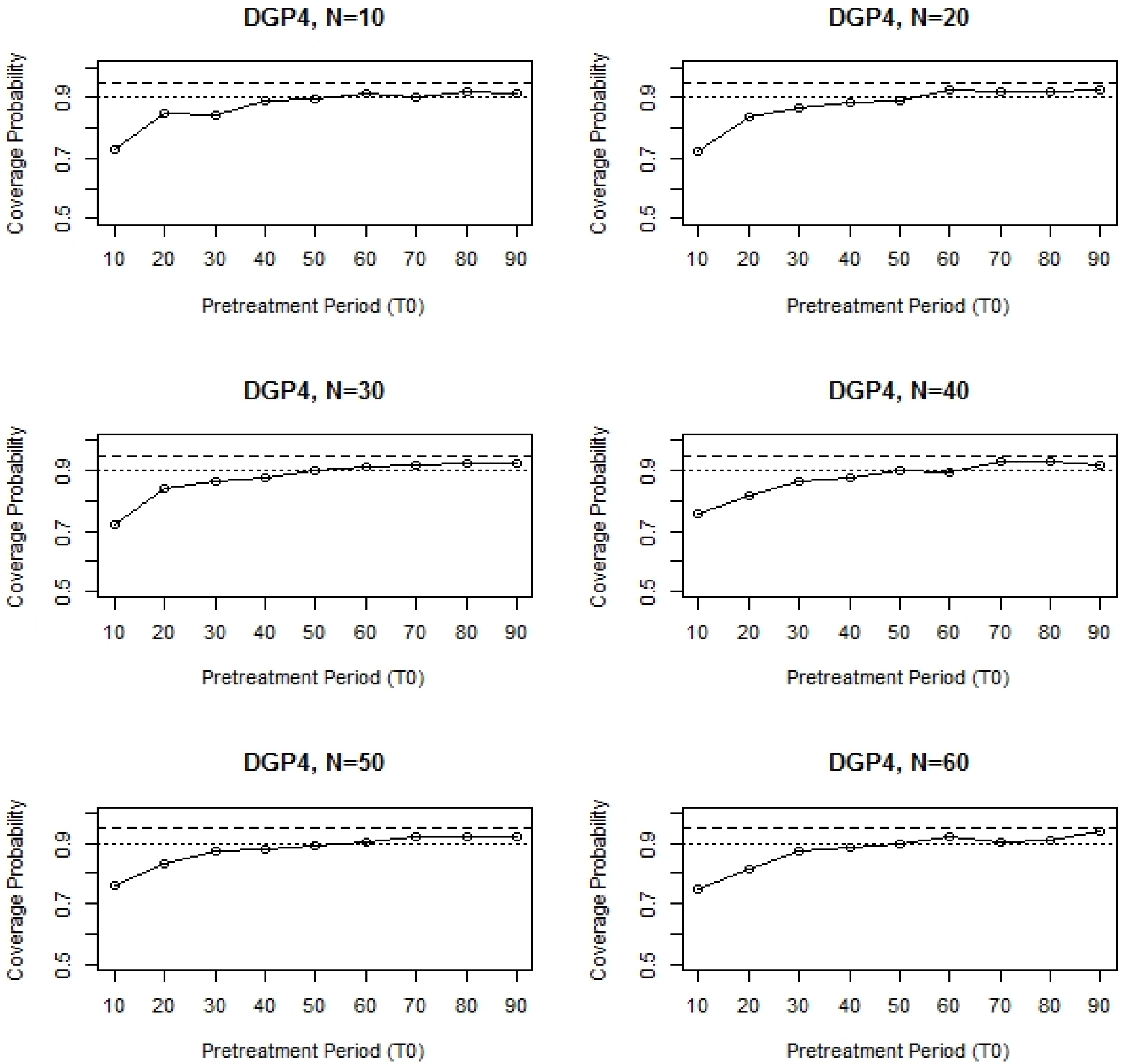

DGP4, N=10
DGP4, N=20
DGP4, N=30
DGP4, N=40
DGP4, N=50
DGP4, N=60
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)

Table 5. Coverage Prob. for DGP 1 with Cross-sectional Heteroskedasticity

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.788	0.888	0.906	0.912	0.933	0.933	0.932	0.943	0.944
$N = 20$	0.798	0.897	0.91	0.939	0.946	0.939	0.955	0.954	0.955
$N = 30$	0.784	0.869	0.917	0.939	0.933	0.938	0.948	0.945	0.952
$N = 40$	0.792	0.903	0.914	0.937	0.942	0.944	0.944	0.956	0.96
$N = 50$	0.805	0.879	0.93	0.935	0.929	0.936	0.95	0.945	0.96
$N = 60$	0.8	0.9	0.919	0.938	0.941	0.938	0.939	0.944	0.967

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period.
The nominal coverage rate is 95%.

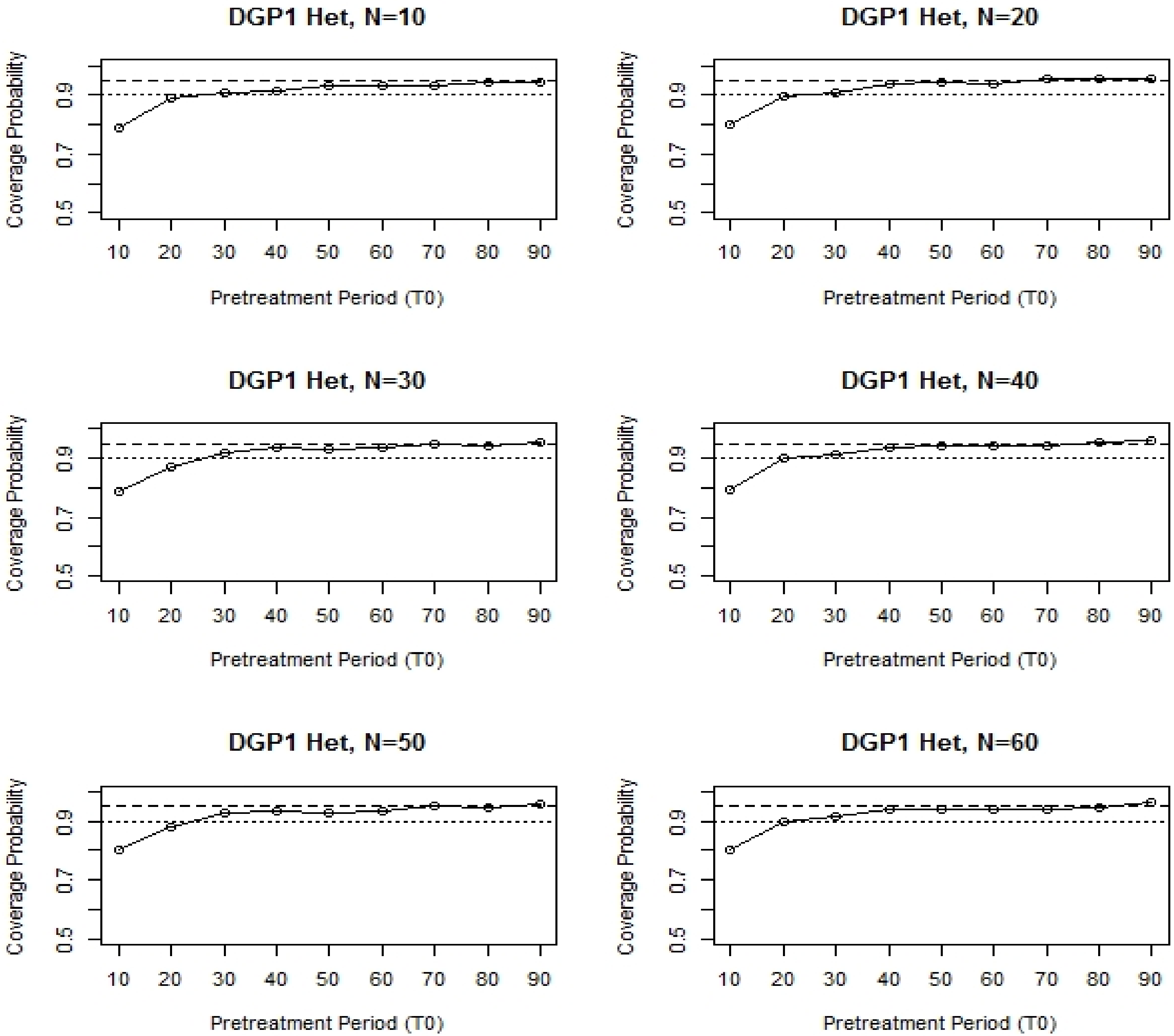

DGP1 Het, N=10
DGP1 Het, N=20
DGP1 Het, N=30
DGP1 Het, N=40
DGP1 Het, N=50
DGP1 Het, N=60
Coverage Probability
0.5 0.7 0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

Table 6. Coverage Prob. for DGP 1 with Within-panel Autocorrelation

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.791	0.897	0.915	0.926	0.934	0.94	0.941	0.941	0.948
$N = 20$	0.79	0.87	0.909	0.923	0.939	0.942	0.953	0.957	0.949
$N = 30$	0.789	0.885	0.924	0.936	0.926	0.938	0.95	0.945	0.961
$N = 40$	0.808	0.879	0.915	0.921	0.938	0.945	0.95	0.947	0.957
$N = 50$	0.772	0.887	0.917	0.918	0.952	0.949	0.937	0.953	0.959
$N = 60$	0.801	0.898	0.919	0.935	0.944	0.937	0.939	0.954	0.95

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.

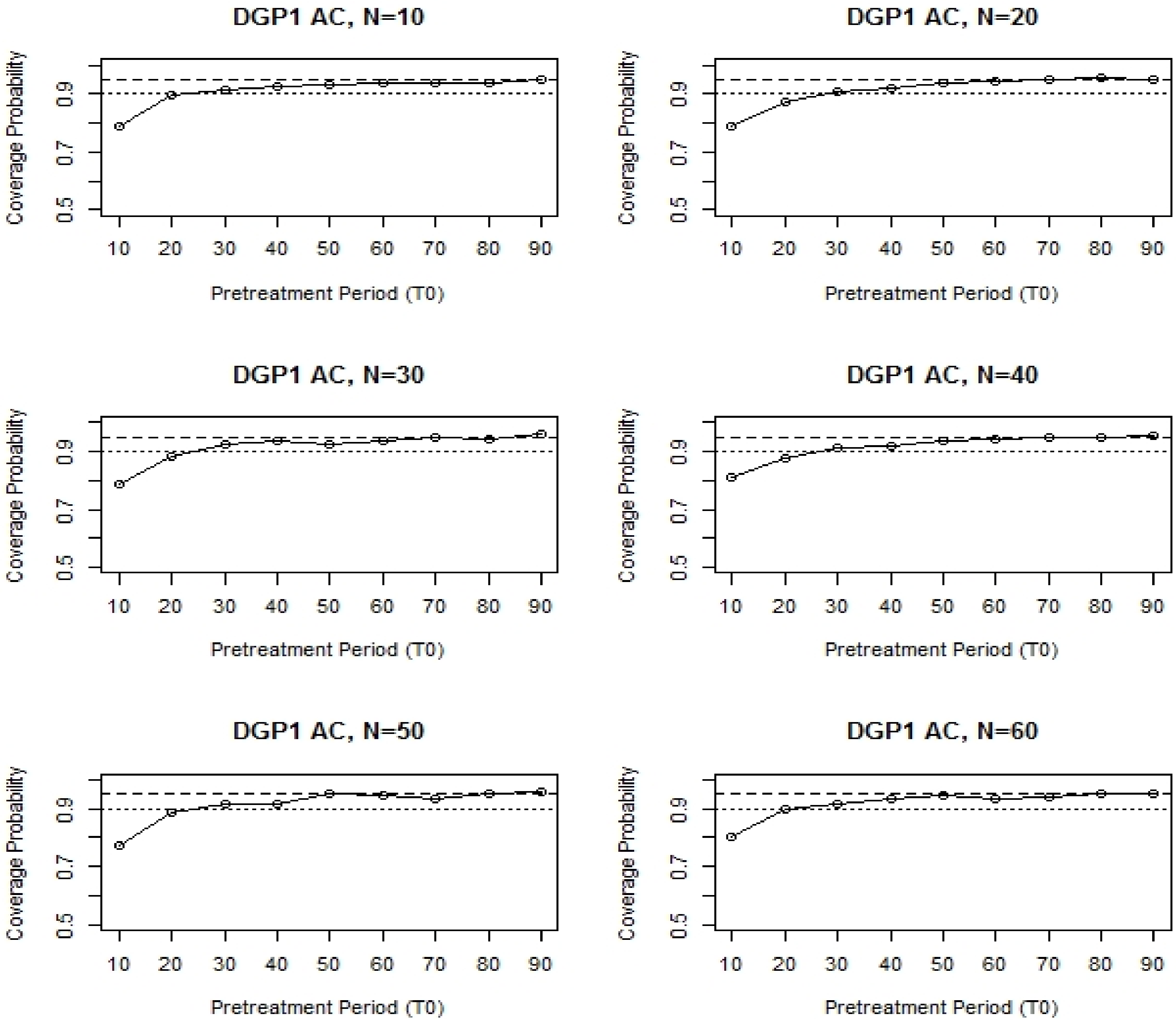

DGP1 AC, N=10
DGP1 AC, N=20
DGP1 AC, N=30
DGP1 AC, N=40
DGP1 AC, N=50
DGP1 AC, N=60
Coverage Probability
0.5
0.7
0.9
10
20
30
40
50
60
70
80
90
Pretreatment Period (T0)

Table 7. Coverage Prob. for DGP 1 with Heteroskedasticity and Autocorrelation

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.791	0.87	0.915	0.905	0.93	0.937	0.953	0.944	0.948
$N = 20$	0.767	0.89	0.906	0.927	0.95	0.928	0.947	0.94	0.949
$N = 30$	0.791	0.863	0.927	0.921	0.922	0.948	0.95	0.937	0.947
$N = 40$	0.765	0.883	0.92	0.926	0.945	0.943	0.947	0.952	0.958
$N = 50$	0.757	0.88	0.927	0.929	0.923	0.939	0.951	0.961	0.949
$N = 60$	0.783	0.895	0.896	0.919	0.94	0.936	0.945	0.952	0.959

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.

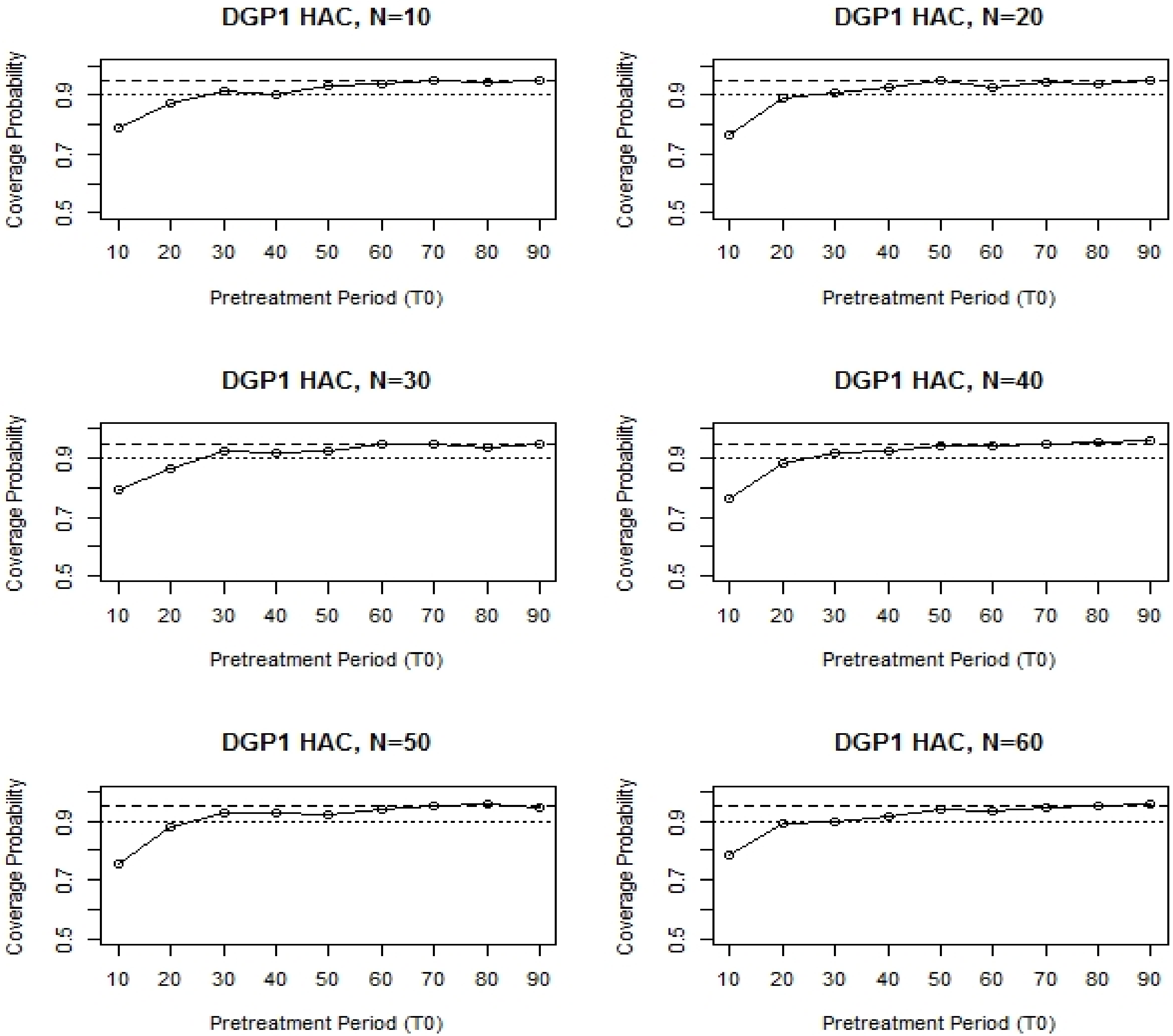
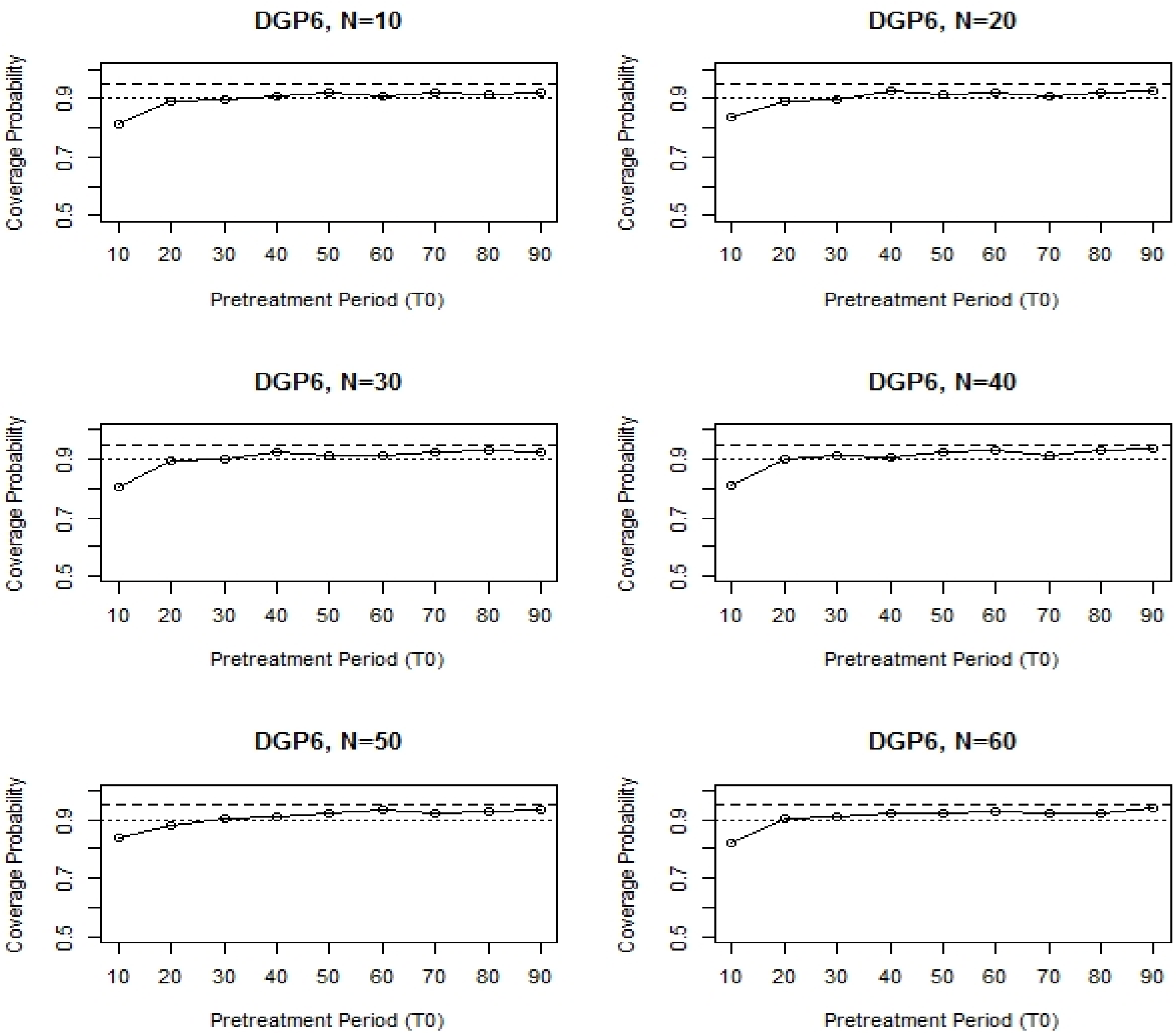

DGP1 HAC, N=10
DGP1 HAC, N=20
DGP1 HAC, N=30
DGP1 HAC, N=40
DGP1 HAC, N=50
DGP1 HAC, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

Table 8. Coverage Prob. for DGP 1 with Nonlinear Logit Transformation

CP	$T_0 = 10$	$T_0 = 20$	$T_0 = 30$	$T_0 = 40$	$T_0 = 50$	$T_0 = 60$	$T_0 = 70$	$T_0 = 80$	$T_0 = 90$
$N = 10$	0.814	0.888	0.896	0.907	0.919	0.909	0.918	0.913	0.92
$N = 20$	0.834	0.888	0.894	0.924	0.914	0.921	0.908	0.92	0.927
$N = 30$	0.807	0.892	0.901	0.927	0.915	0.915	0.924	0.933	0.924
$N = 40$	0.812	0.903	0.913	0.909	0.922	0.932	0.915	0.928	0.938
$N = 50$	0.84	0.879	0.902	0.913	0.921	0.934	0.921	0.93	0.933
$N = 60$	0.821	0.903	0.91	0.924	0.921	0.931	0.92	0.924	0.938

Notes: N is the number of cross-sectional units. T_0 is the pretreatment period. The nominal coverage rate is 95%.


DGP6, N=10
DGP6, N=20
DGP6, N=30
DGP6, N=40
DGP6, N=50
DGP6, N=60
Coverage Probability
0.5
0.7
0.9
10 20 30 40 50 60 70 80 90
Pretreatment Period (T0)

7. Stata Application with the qcm Command

- We provide a forthcoming qcm command to implement “Quantile Control Method” (QCM) in Stata

- Due the convenient integration between Stata and Python, qcm calls Python’s scikit-garden module to implement quantile random forest.

Installation of qcm

- Install Python 3.7 or lower version
- Install the Python module scikit-garden: executing the command "conda install -c conda-forge scikit-garden" at Anaconda prompt (recommended).
- ssc install qcm, all replace
(coming soon)

Example: Economic Integration of HK with Mainland China in 2004Q1

- 《内地与香港关于建立更紧密经贸关系的安排》（英文简称CEPA）的主要内容包括：允许众多香港产品零关税进入内地，放宽内地对香港服务业的准入领域以及贸易便利化三方面。
- Pre-treatment period: 1993Q1 - 2003Q4 (44 quarters)
- Treatment period: 2004Q1 - 2008Q1 (17 quarters)

自带示例数据集 growth.dta

```
. use growth, clear

. describe

Contains data from growth.dta
 Observations:         1,525
    Variables:             3                  16 Jun 2019 00:03
--------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------
time            float   %tq
gdp             float   %8.0g
region          long    %13.0g     region
--------------------------------------------------------------------
Sorted by:

. xtset region time

Panel variable: region (strongly balanced)
 Time variable: time, 1993q1 to 2008q1
         Delta: 1 quarter
```

命令qcm的基本句型

- display tq(2004q1)
- 176

- qcm gdp, trunit(9) trperiod(176) importance

- 其中，必选项trunit()指定处理单位(treated unit)，trperiod()指定处理期(treatment period，必须为整数)；选择项importance计算变量重要性

Result 1: Basic Info

```
treatment unit:
    HongKong
Control Units:
    Australia Austria Canada China Denmark Finland France Germany Indonesia Italy Japan Korea
    Malaysia Mexico Netherlands NewZealand Norway Philippines Singapore Switzerland Taiwan Thailand
    UnitedKingdom UnitedStates
Pre-treatment Periods:
    1993q1 1993q2 1993q3 1993q4 1994q1 1994q2 1994q3 1994q4 1995q1 1995q2 1995q3 1995q4 1996q1 1996q2
    1996q3 1996q4 1997q1 1997q2 1997q3 1997q4 1998q1 1998q2 1998q3 1998q4 1999q1 1999q2 1999q3 1999q4
    2000q1 2000q2 2000q3 2000q4 2001q1 2001q2 2001q3 2001q4 2002q1 2002q2 2002q3 2002q4 2003q1 2003q2
    2003q3 2003q4
Post-treatment Periods:
    2004q1 2004q2 2004q3 2004q4 2005q1 2005q2 2005q3 2005q4 2006q1 2006q2 2006q3 2006q4 2007q1 2007q2
    2007q3 2007q4 2008q1
Dependent Variable:
    gdp
Response:
    gdp·HongKong
Predictors:
    gdp·Australia gdp·Austria gdp·Canada gdp·China gdp·Denmark gdp·Finland gdp·France gdp·Germany
    gdp·Indonesia gdp·Italy gdp·Japan gdp·Korea gdp·Malaysia gdp·Mexico gdp·Netherlands
    gdp·NewZealand gdp·Norway gdp·Philippines gdp·Singapore gdp·Switzerland gdp·Taiwan gdp·Thailand
    gdp·UnitedKingdom gdp·UnitedStates
```

Result 2: Counterfactual Predictions

Fitting results in the pre-treatment periods:

Mean Absolute Error	=	0.00758	Number of Observations	=	44
Mean Squared Error	=	0.00010	Number of Predictors	=	24
Root Mean Squared Error	=	0.00998	R-squared	=	0.94032

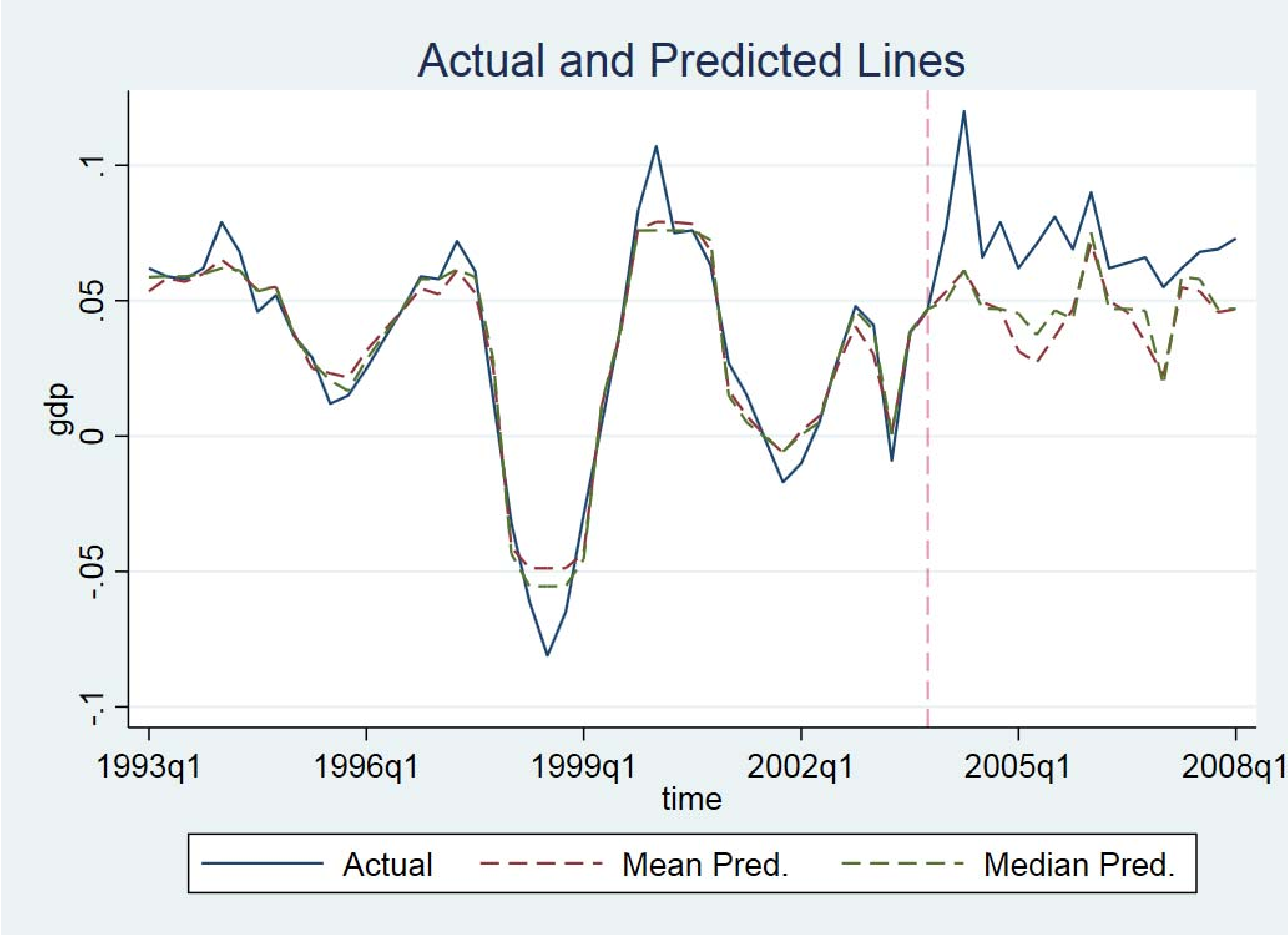
Prediction results in the post-treatment periods:

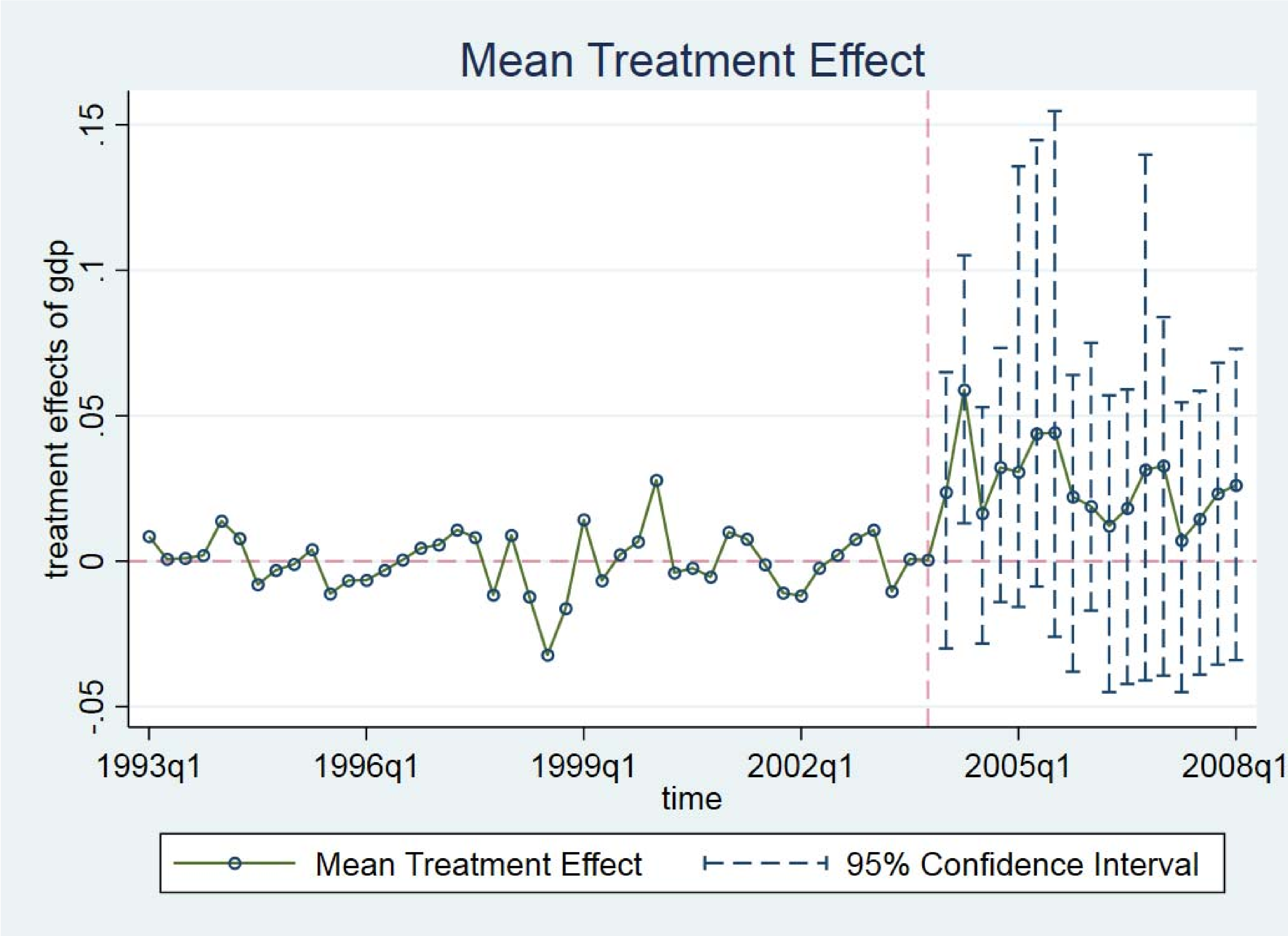
Time	Treated	Mean Pred.	Median Pred.	[95% Conf.	Interval of Pred.]
2004q1	0.0770	0.0533	0.0500	0.0121	0.1070
2004q2	0.1200	0.0611	0.0613	0.0149	0.1070
2004q3	0.0660	0.0495	0.0471	0.0130	0.0943
2004q4	0.0790	0.0467	0.0470	0.0057	0.0931
2005q1	0.0620	0.0314	0.0453	-0.0737	0.0777
2005q2	0.0710	0.0272	0.0374	-0.0737	0.0797
2005q3	0.0810	0.0367	0.0465	-0.0737	0.1070
2005q4	0.0690	0.0469	0.0434	0.0050	0.1070
2006q1	0.0900	0.0711	0.0751	0.0150	0.10[illegible]0
2006q2	0.0620	0.0499	0.0470	0.0050	0.1070
2006q3	0.0640	0.0458	0.0470	0.0050	0.1062
2006q4	0.0660	0.0346	0.0463	-0.0737	0.1070
2007q1	0.0550	0.0222	0.0189	-0.0289	0.0944
2007q2	0.0620	0.0549	0.0588	0.0074	0.1070
2007q3	0.0680	0.0535	0.0580	0.0095	0.1070
2007q4	0.0690	0.0458	0.0470	0.0008	0.1045
2008q1	0.0730	0.0469	0.0470	0.0000	0.1070

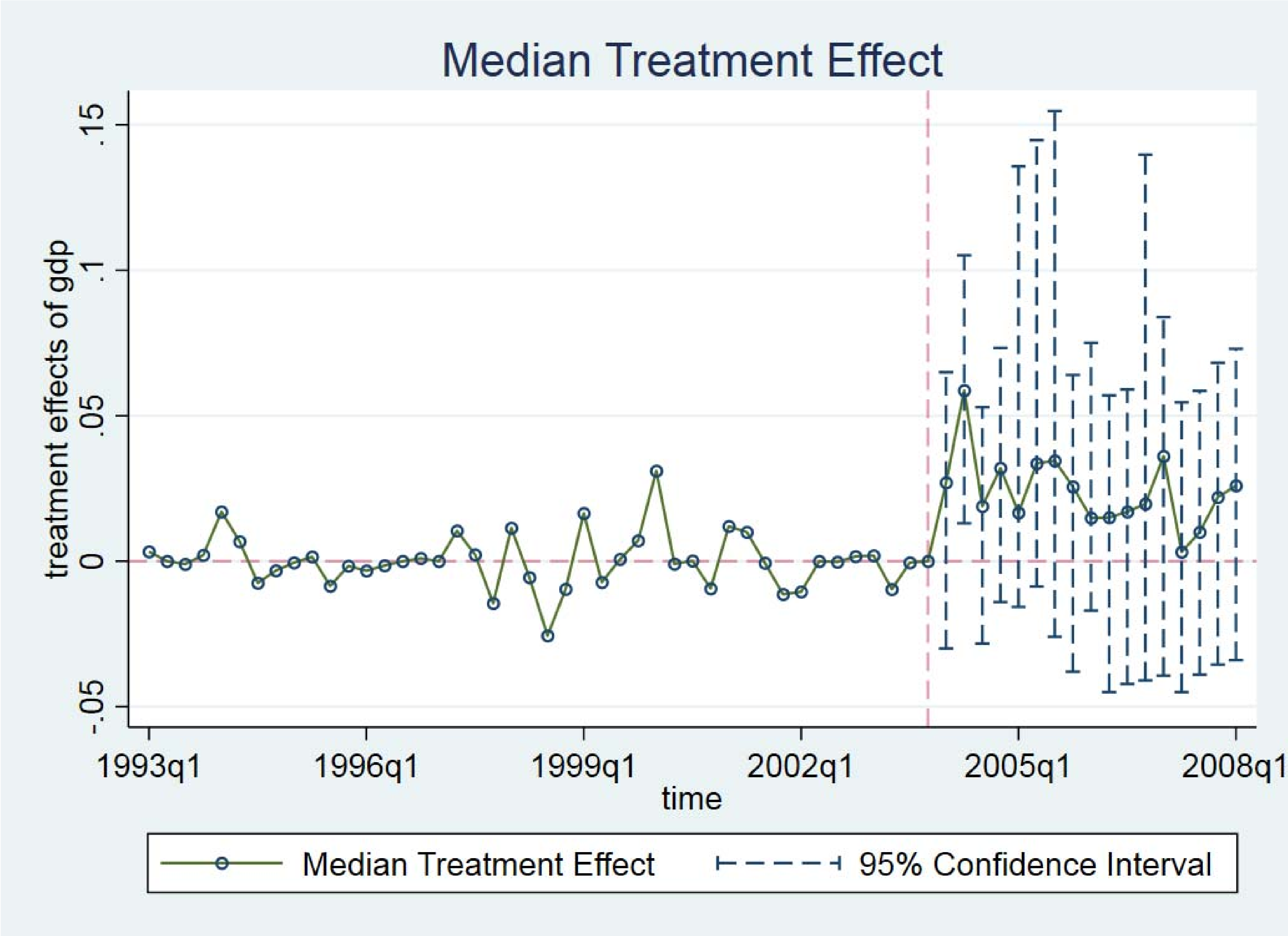
Result 3: Treatment Effects

Time	Mean Tr. Eff.	Median Tr. Eff.	[95% Conf. Interval	of Tr. Eff.]
2004q1	0.0237	0.0270	-0.0300	0.0649
2004q2	0.0589	0.0587	0.0130	0.1051
2004q3	0.0165	0.0189	-0.0283	0.0530
2004q4	0.0323	0.0320	-0.0141	0.0733
2005q1	0.0306	0.0167	-0.0157	0.1357
2005q2	0.0438	0.0336	-0.0087	0.1447
2005q3	0.0443	0.0345	-0.0260	0.1547
2005q4	0.0221	0.0256	-0.0380	0.0640
2006q1	0.0189	0.0149	-0.0170	0.0750
2006q2	0.0121	0.0150	-0.0450	0.0570
2006q3	0.0182	0.0170	-0.0422	0.0590
2006q4	0.0314	0.0197	-0.0410	0.1397
2007q1	0.0328	0.0361	-0.0394	0.0839
2007q2	0.0071	0.0032	-0.0450	0.0546
2007q3	0.0145	0.0100	-0.0390	0.0585
2007q4	0.0232	0.0220	-0.0355	0.0682
2008q1	0.0261	0.0260	-0.0340	0.0730
Mean	0.0269	0.0242	-0.0286	0.0861

Note: The average treatment effect over the post-treatment periods is .026852739.


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

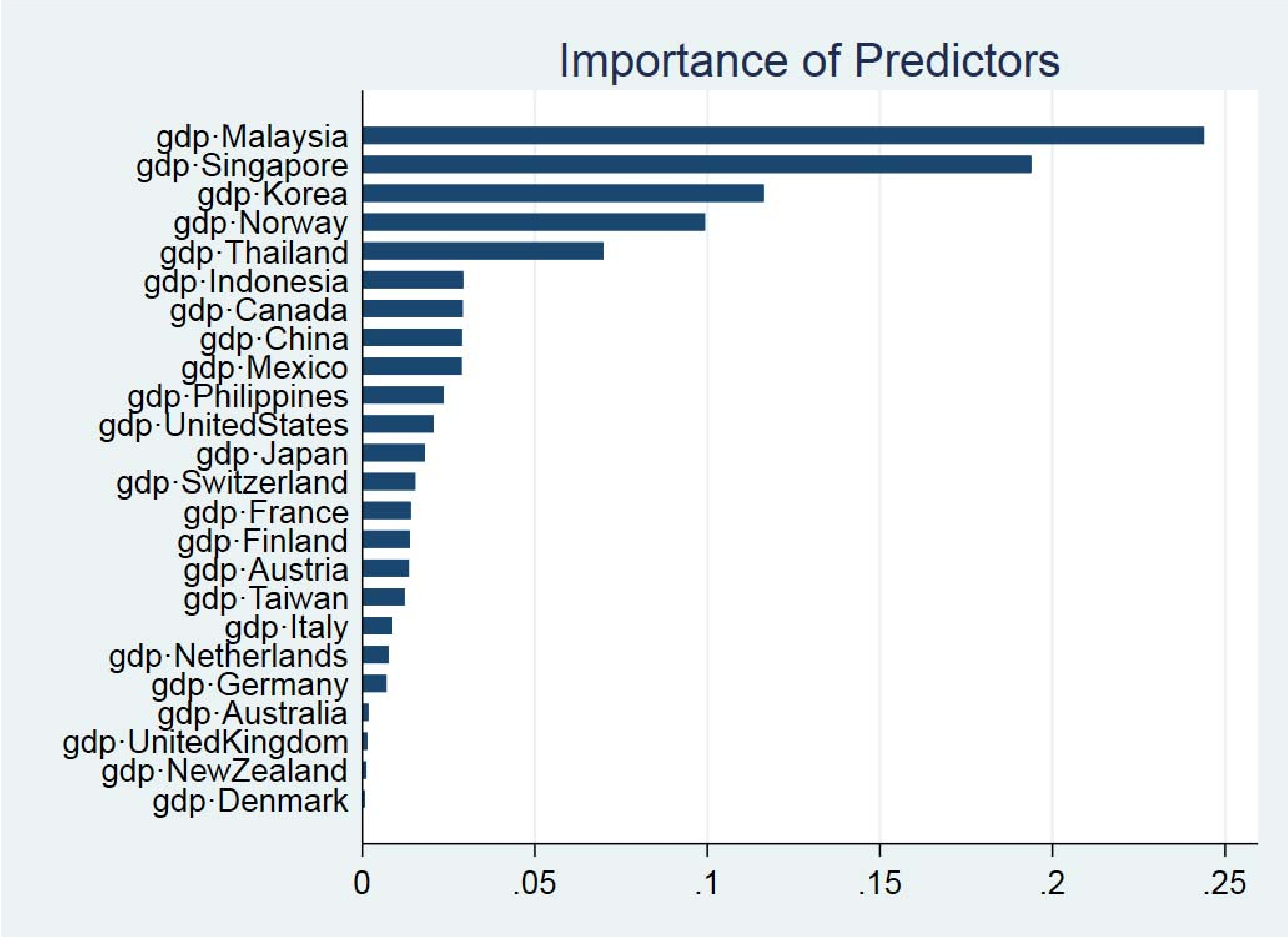

Mean Treatment Effect
treatment effects of gdp
.15
.1
.05
0
-.05
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Mean Treatment Effect
95% Confidence Interval


Median Treatment Effect
treatment effects of gdp
-.05
0
.05
.1
.15
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Median Treatment Effect
95% Confidence Interval

Result 4: Variable Importance

Importance of predictors:

Predictor	Importance
gdp·Malaysia	0.2439
gdp·Singapore	0.1939
gdp·Korea	0.1164
gdp·Norway	0.0993
gdp·Thailand	0.0699
gdp·Indonesia	0.0293
gdp·Canada	0.0291
gdp·China	0.0289
gdp·Mexico	0.0288
gdp·Philippines	0.0236
gdp·UnitedStates	0.0207
gdp·Japan	0.0182
gdp·Switzerland	0.0154
gdp·France	0.0141
gdp·Finland	0.0138
gdp·Austria	0.0135
gdp·Taiwan	0.0123
gdp·Italy	0.0088
gdp·Netherlands	0.0077
gdp·Germany	0.0070
gdp·Australia	0.0018
gdp·UnitedKingdom	0.0015
gdp·NewZealand	0.0012
gdp·Denmark	0.0007


Importance of Predictors
gdp·Malaysia
gdp·Singapore
gdp·Korea
gdp·Norway
gdp·Thailand
gdp·Indonesia
gdp·Canada
gdp·China
gdp·Mexico
gdp·Philippines
gdp·UnitedStates
gdp·Japan
gdp·Switzerland
gdp·France
gdp·Finland
gdp·Austria
gdp·Taiwan
gdp·Italy
gdp·Netherlands
gdp·Germany
gdp·Australia
gdp·UnitedKingdom
gdp·NewZealand
gdp·Denmark
0
.05
.1
.15
.2
.25

安慰剂检验（Placebo Tests）

- `qcm gdp, trunit(9) trperiod(176) placebo(unit period(168 172))`

- 其中，“unit”表示使用所有控制组个体作为“fake treatment units”，而“period(168 172)”分别以2002Q1与2003Q1作为“fake treatment times”。

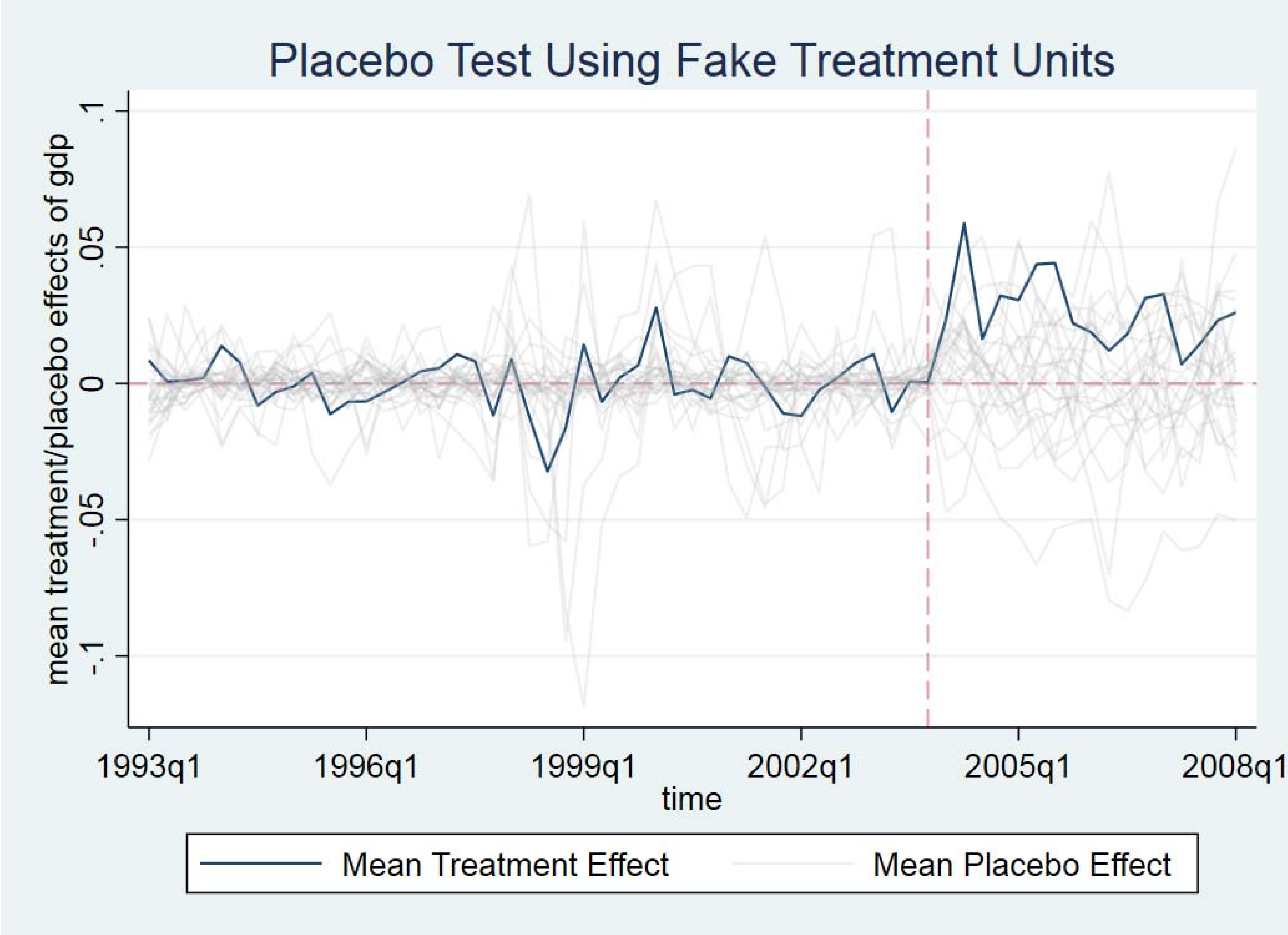
Result 1: Fake Treatment Units

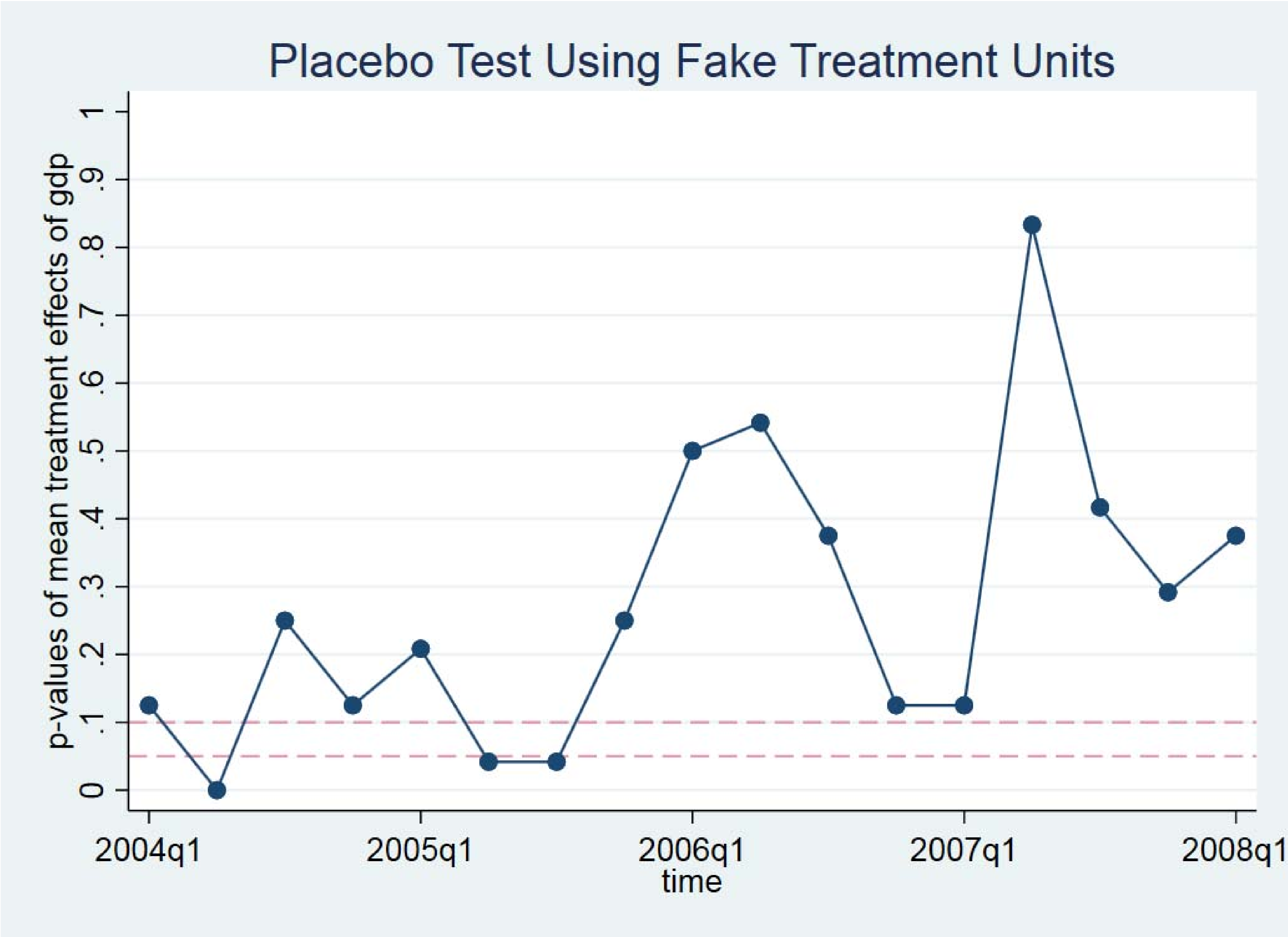
```
Implementing placebo test using fake treatment unit Australia...Austria...Canada...China...Denmark.
> ..Finland...France...Germany...Indonesia...Italy...Japan...Korea...Malaysia...Mexico...Netherland
> s...NewZealand...Norway...Philippines...Singapore...Switzerland...Taiwan...Thailand...UnitedKingd
> om...UnitedStates...
```

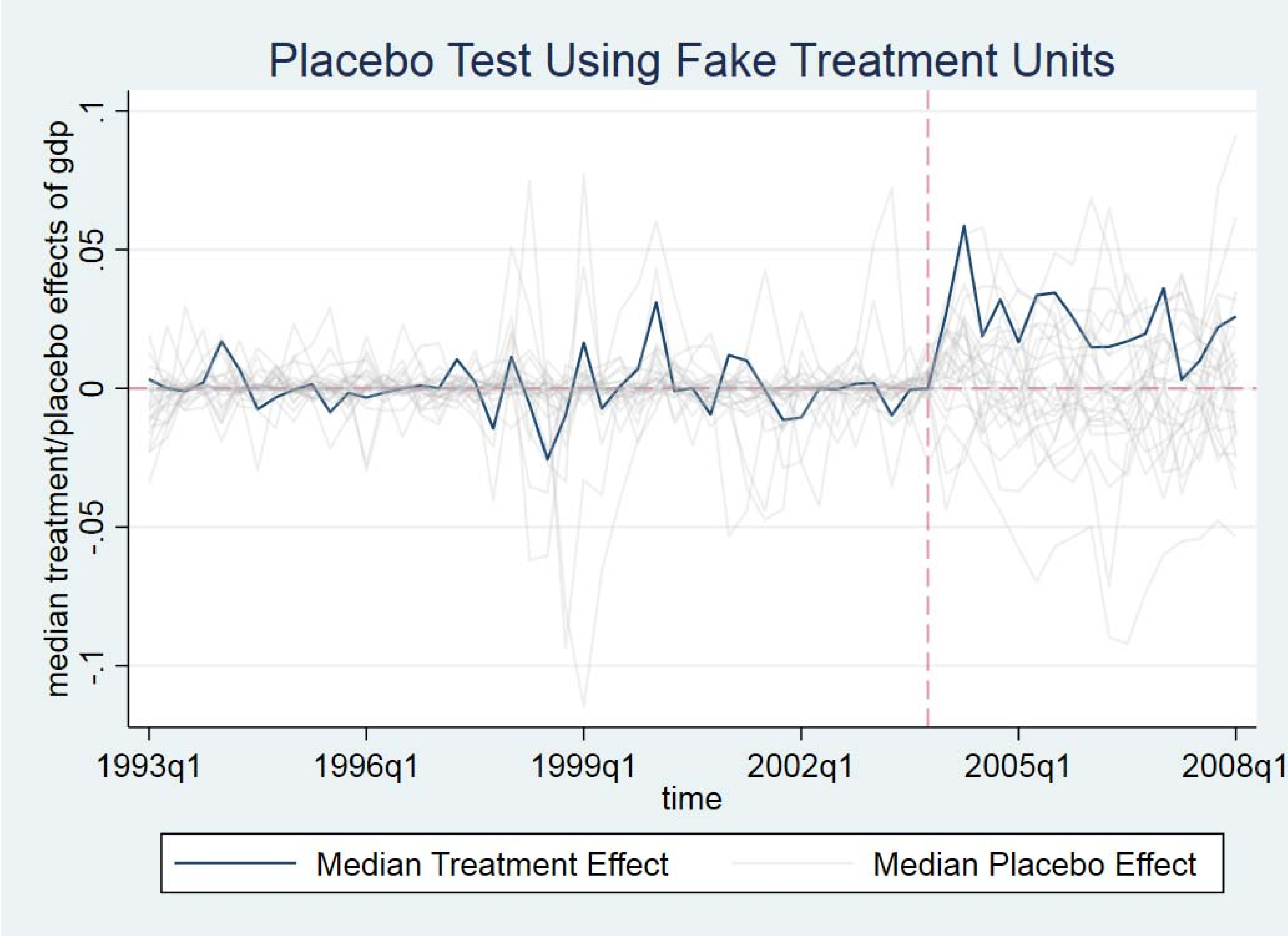
Placebo test results using fake treatment units:

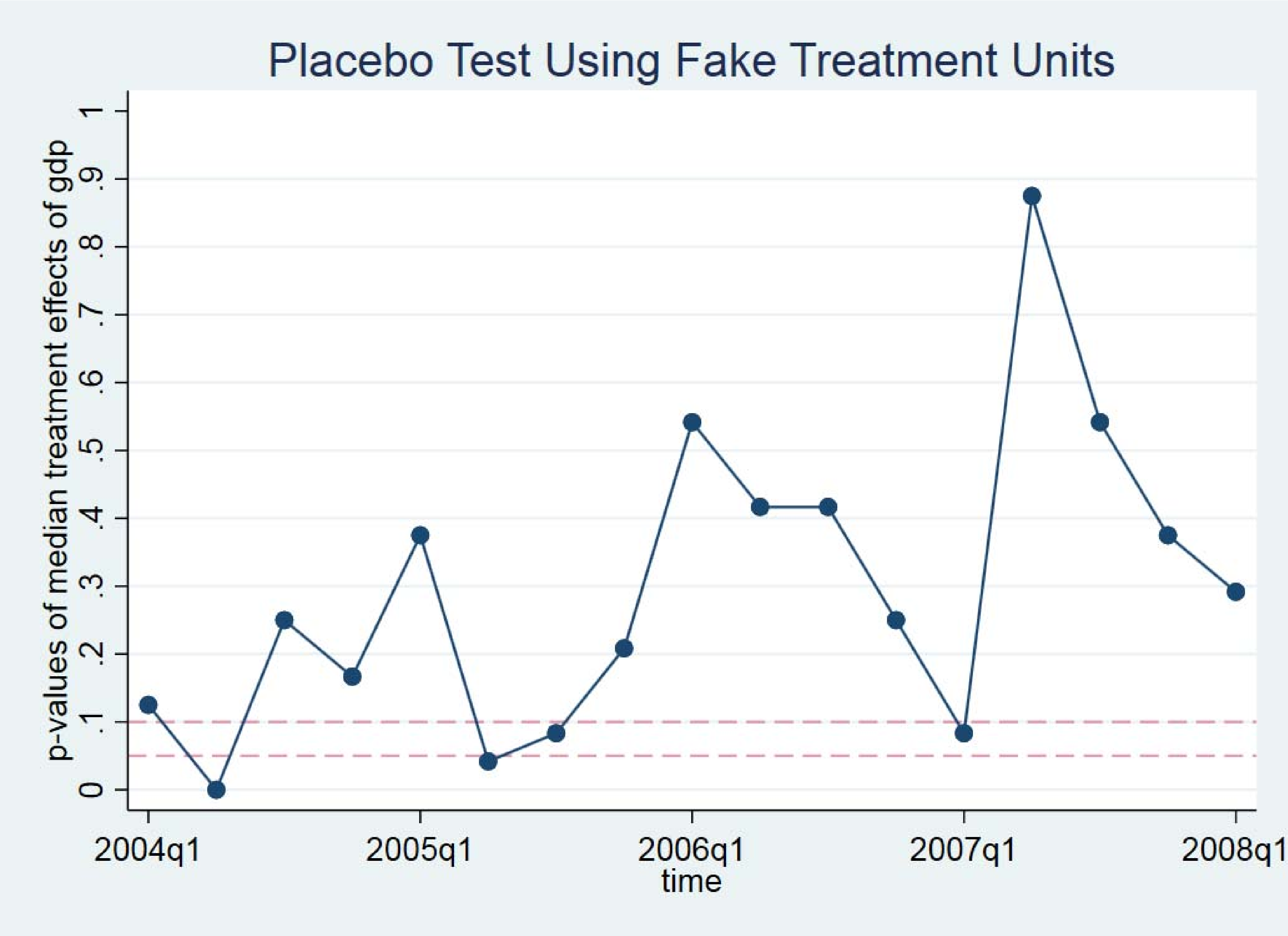
Time	Mean Tr. Eff.	P-value of Mean	Median Tr. Eff.	P-value of Median
2004q1	0.0237	0.1250	0.0270	0.1250
2004q2	0.0589	0.0000	0.0587	0.0000
2004q3	0.0165	0.2500	0.0189	0.2500
2004q4	0.0323	0.1250	0.0320	0.1667
2005q1	0.0306	0.2083	0.0167	0.3750
2005q2	0.0438	0.0417	0.0336	0.0417
2005q3	0.0443	0.0417	0.0345	0.0833
2005q4	0.0221	0.2500	0.0256	0.2083
2006q1	0.0189	0.5000	0.0149	0.5417
2006q2	0.0121	0.5417	0.0150	0.4167
2006q3	0.0182	0.3750	0.0170	0.4167
2006q4	0.0314	0.1250	0.0197	0.2500
2007q1	0.0328	0.1250	0.0361	0.0833
2007q2	0.0071	0.8333	0.0032	0.8750
2007q3	0.0145	0.4167	0.0100	0.5417
2007q4	0.0232	0.2917	0.0220	0.3750
2008q1	0.0261	0.3750	0.0260	0.2917

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
mean treatment/placebo effects of gdp
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Mean Treatment Effect
Mean Placebo Effect


Placebo Test Using Fake Treatment Units
p-values of mean treatment effects of gdp
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
2004q1
2005q1
2006q1
2007q1
2008q1
time

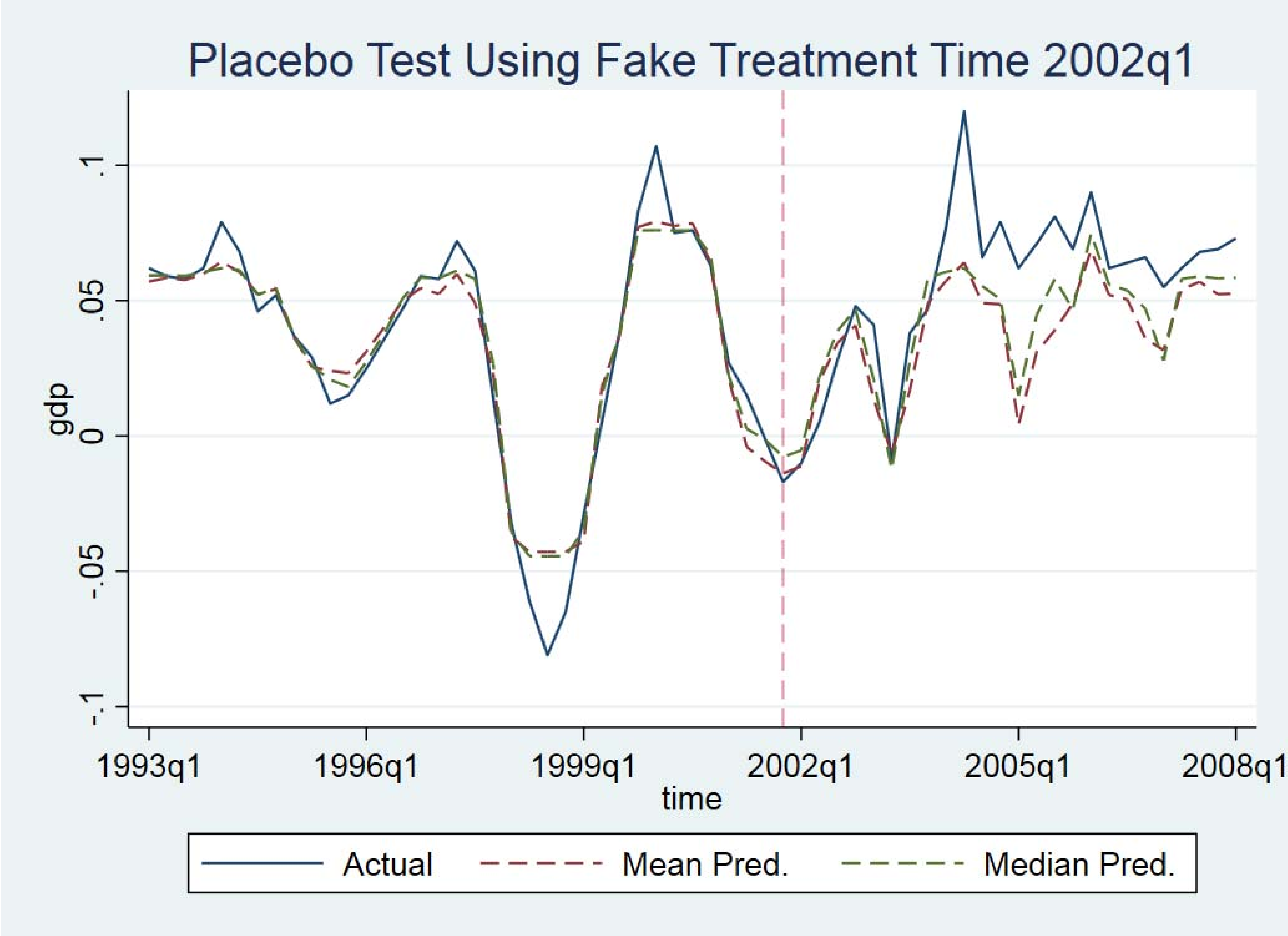

Placebo Test Using Fake Treatment Units
median treatment/placebo effects of gdp
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Median Treatment Effect
Median Placebo Effect

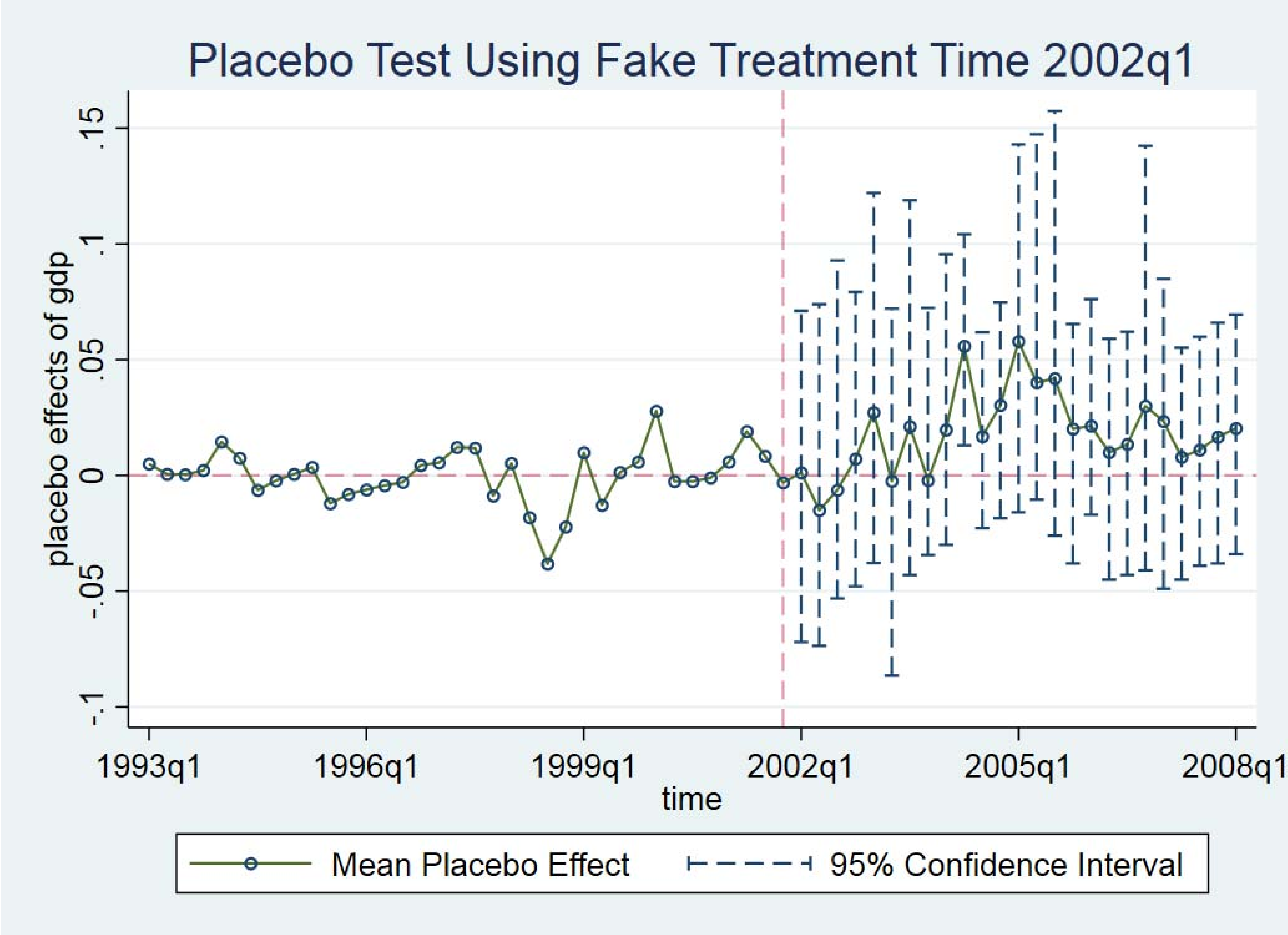

Placebo Test Using Fake Treatment Units
p-values of median treatment effects of gdp
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
2004q1
2005q1
2006q1
2007q1
2008q1
time

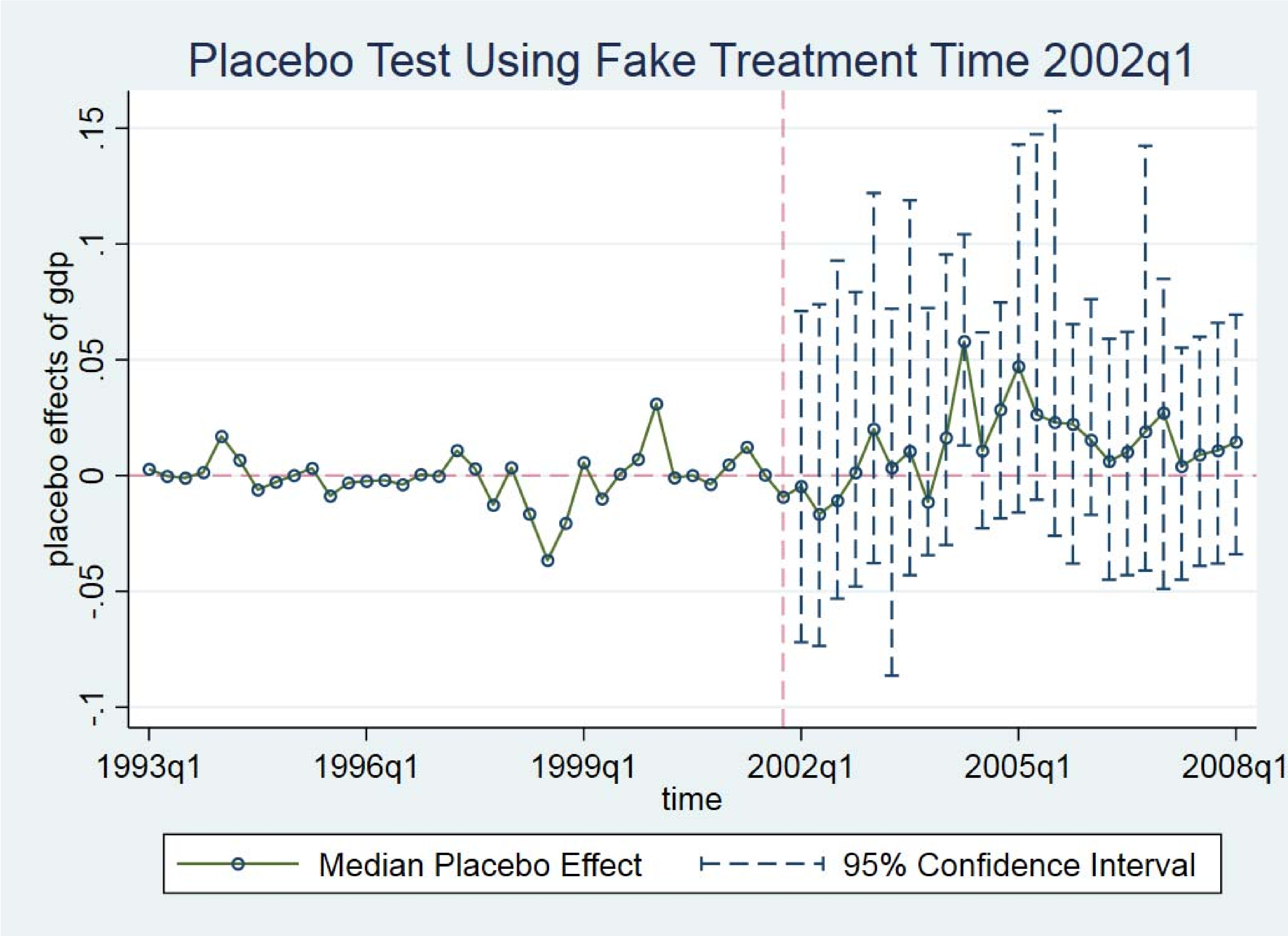
Result 2: Fake Treatment Time 2002Q1

Placebo test results using fake treatment time **2002q1**:

Time	Mean Tr. Eff.	Median Tr. Eff.	[95% Conf. Interval of Tr. Eff.]	
2002q1	0.0012	-0.0046	-0.0720	0.0710
2002q2	-0.0150	-0.0166	-0.0736	0.0739
2002q3	-0.0063	-0.0108	-0.0532	0.0928
2002q4	0.0072	0.0013	-0.0479	0.0792
2003q1	0.0272	0.0201	-0.0378	0.1220
2003q2	-0.0024	0.0033	-0.0864	0.0720
2003q3	0.0212	0.0105	-0.0430	0.1189
2003q4	-0.0021	-0.0114	-0.0344	0.0723
2004q1	0.0197	0.0164	-0.0300	0.0955
2004q2	0.0559	0.0579	0.0130	0.1042
2004q3	0.0169	0.0107	-0.0228	0.0618
2004q4	0.0304	0.0285	-0.0185	0.0747
2005q1	0.0579	0.0471	-0.0159	0.1430
2005q2	0.0401	0.0264	-0.0105	0.1473
2005q3	0.0419	0.0230	-0.0260	0.1573
2005q4	0.0201	0.0222	-0.0380	0.0654
2006q1	0.0214	0.0154	-0.0170	0.0761
2006q2	0.0100	0.0062	-0.0450	0.0590
2006q3	0.0135	0.0102	-0.0430	0.0621
2006q4	0.0300	0.0190	-0.0410	0.1423
2007q1	0.0234	0.0271	-0.0490	0.0849
2007q2	0.0079	0.0040	-0.0450	0.0552
2007q3	0.0111	0.0090	-0.0390	0.0599
2007q4	0.0166	0.0108	-0.0380	0.0660
2008q1	0.0204	0.0145	-0.0340	0.0695


Placebo Test Using Fake Treatment Time 2002q1
gdp
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Actual
Mean Pred.
Median Pred.

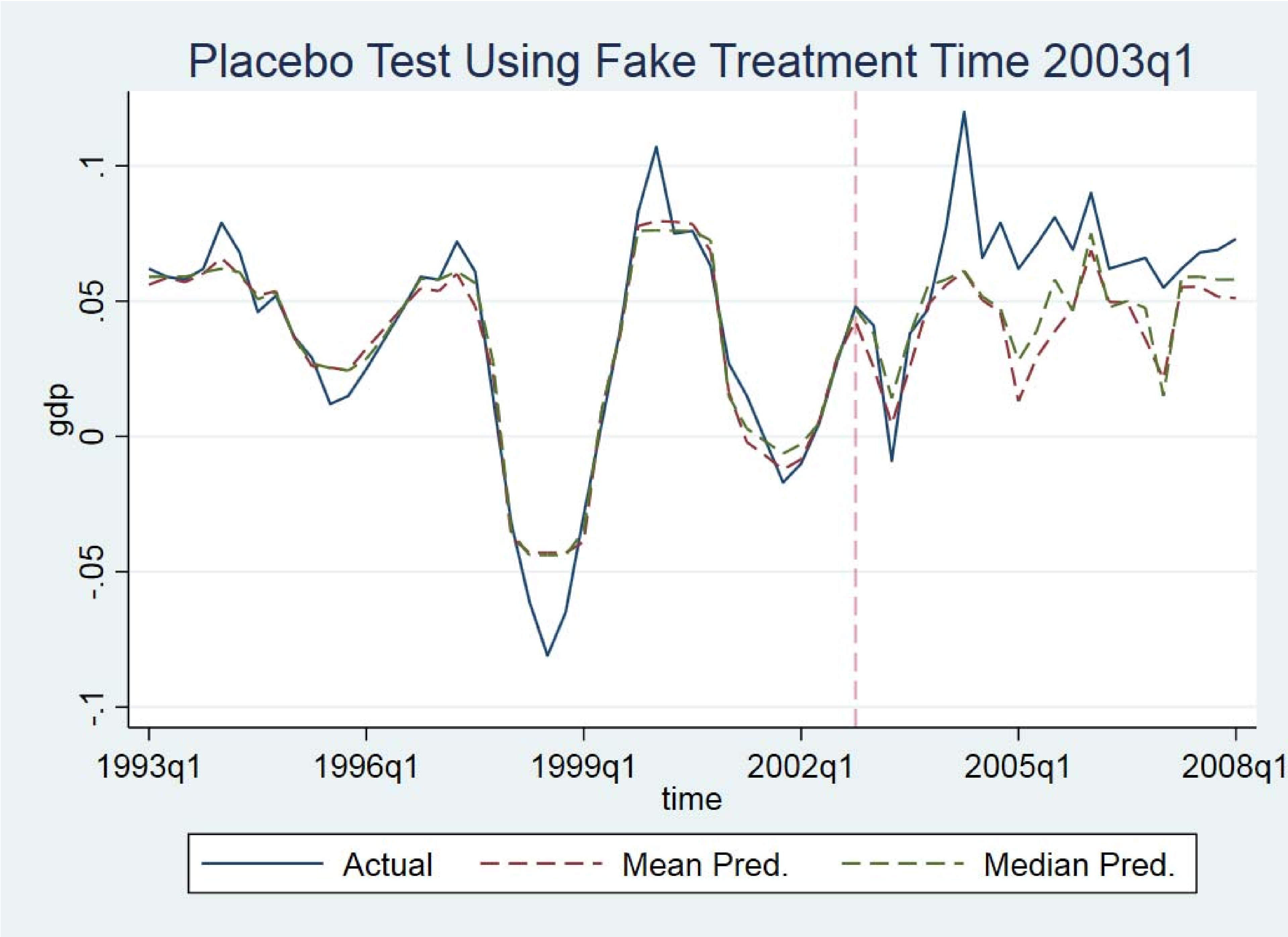

Placebo Test Using Fake Treatment Time 2002q1
placebo effects of gdp
.15
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Mean Placebo Effect
95% Confidence Interval

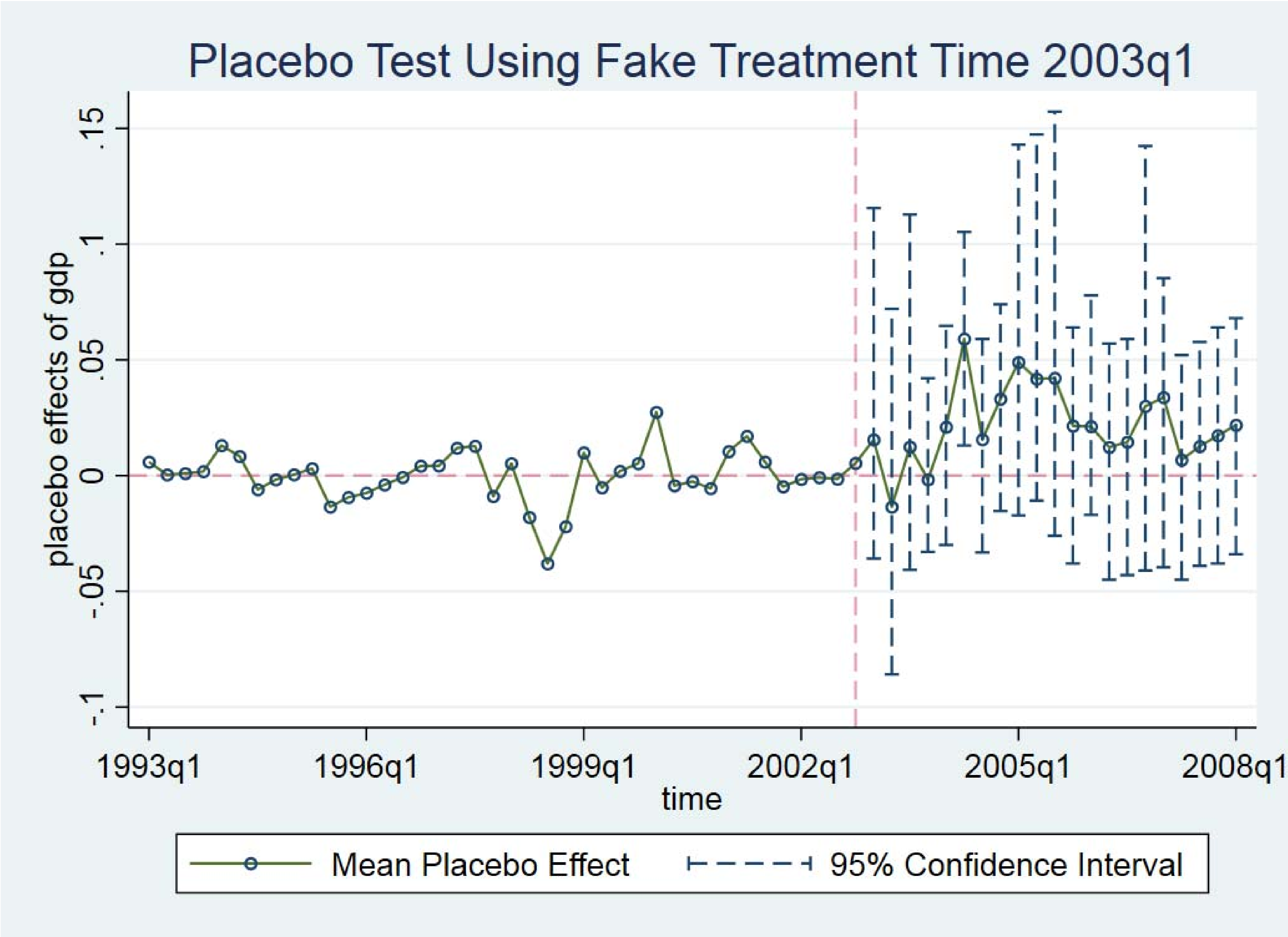

Placebo Test Using Fake Treatment Time 2002q1
placebo effects of gdp
.15
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Median Placebo Effect
95% Confidence Interval

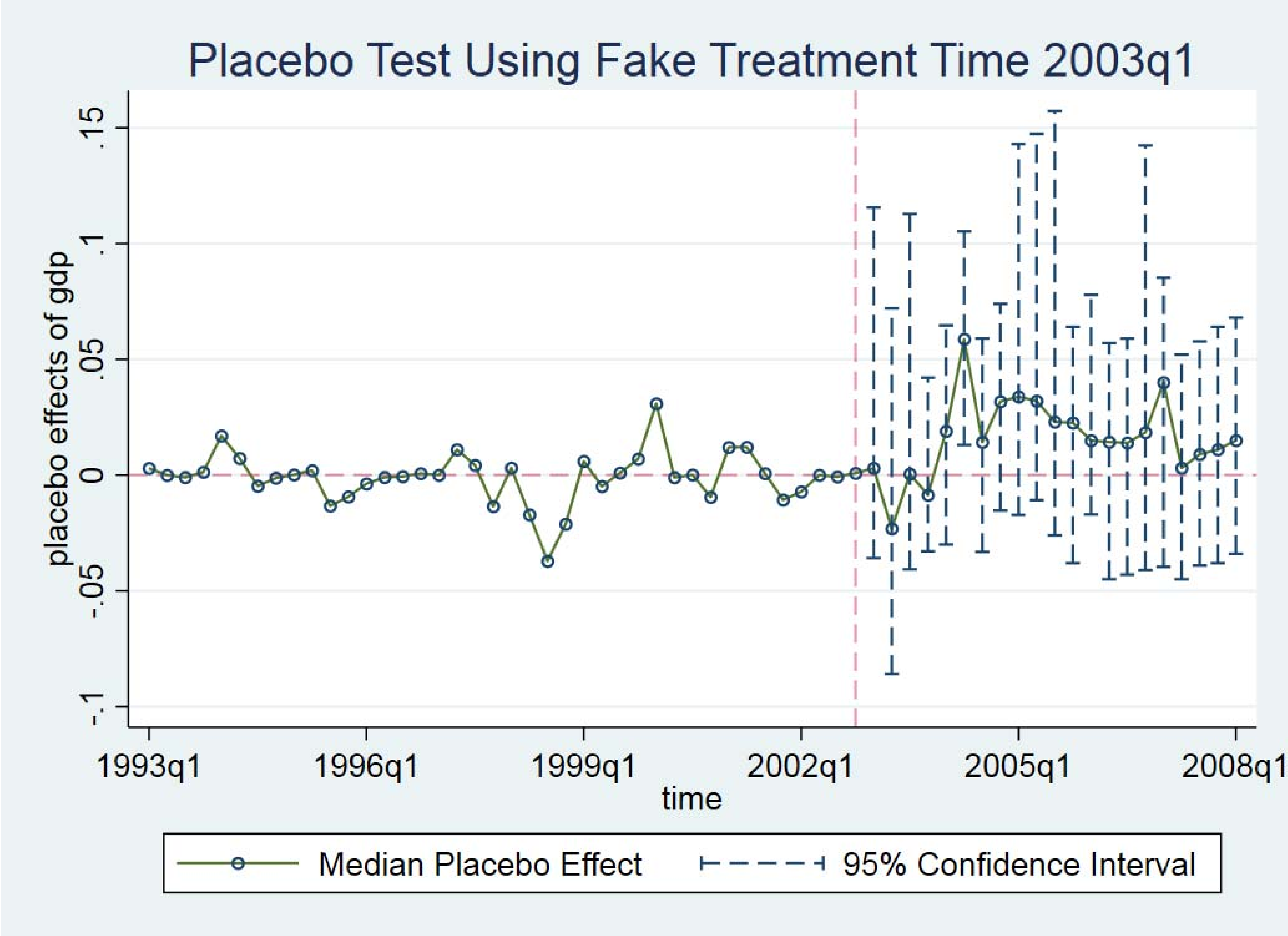
Result 3: Fake Treatment Time 2003Q1

Placebo test results using fake treatment time **2003q1**:

Time	Mean Tr. Eff.	Median Tr. Eff.	[95% Conf. Interval of Tr. Eff.]	
2003q1	**0.0155**	**0.0030**	**-0.0359**	**0.1156**
2003q2	**-0.0135**	**-0.0231**	**-0.0858**	**0.0720**
2003q3	**0.0125**	**0.0005**	**-0.0407**	**0.1128**
2003q4	**-0.0017**	**-0.0086**	**-0.0329**	**0.0420**
2004q1	**0.0210**	**0.0190**	**-0.0300**	**0.0646**
2004q2	**0.0591**	**0.0588**	**0.0130**	**0.1053**
2004q3	**0.0156**	**0.0143**	**-0.0332**	**0.0590**
2004q4	**0.0331**	**0.0318**	**-0.0153**	**0.0740**
2005q1	**0.0490**	**0.0339**	**-0.0172**	**0.1430**
2005q2	**0.0418**	**0.0321**	**-0.0108**	**0.1474**
2005q3	**0.0421**	**0.0230**	**-0.0260**	**0.1572**
2005q4	**0.0215**	**0.0226**	**-0.0380**	**0.0640**
2006q1	**0.0212**	**0.0150**	**-0.0170**	**0.0778**
2006q2	**0.0123**	**0.0143**	**-0.0450**	**0.0570**
2006q3	**0.0145**	**0.0139**	**-0.0430**	**0.0590**
2006q4	**0.0300**	**0.0184**	**-0.0410**	**0.1424**
2007q1	**0.0338**	**0.0400**	**-0.0396**	**0.0853**
2007q2	**0.0068**	**0.0031**	**-0.0450**	**0.0520**
2007q3	**0.0127**	**0.0090**	**-0.0390**	**0.0578**
2007q4	**0.0173**	**0.0111**	**-0.0380**	**0.0640**
2008q1	**0.0219**	**0.0150**	**-0.0340**	**0.0680**


Placebo Test Using Fake Treatment Time 2003q1
gdp
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Actual
Mean Pred.
Median Pred.


Placebo Test Using Fake Treatment Time 2003q1
placebo effects of gdp
.15
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Mean Placebo Effect
95% Confidence Interval

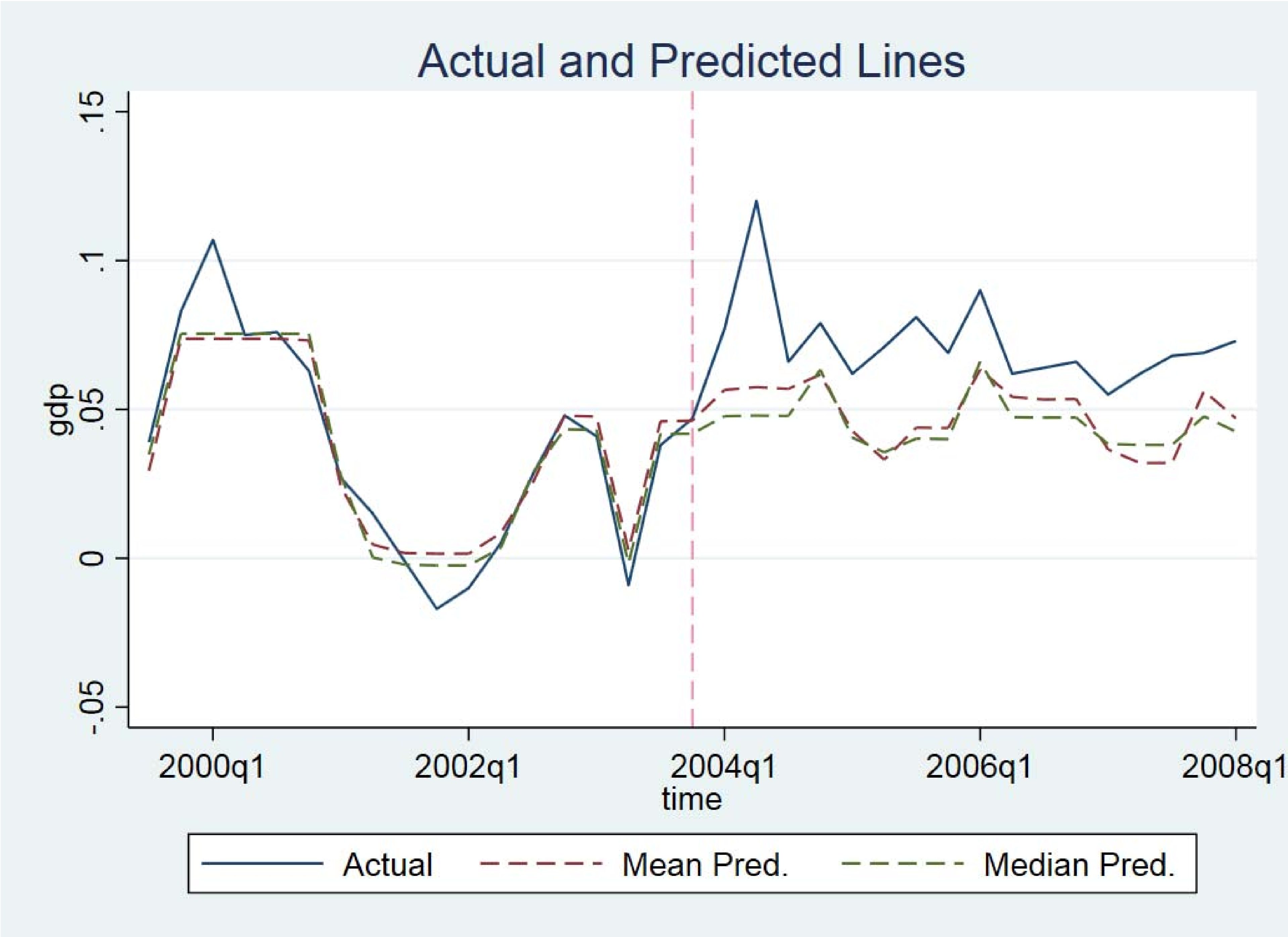

Placebo Test Using Fake Treatment Time 2003q1
placebo effects of gdp
.15
.1
.05
0
-.05
-.1
1993q1
1996q1
1999q1
2002q1
2005q1
2008q1
time
Median Placebo Effect
95% Confidence Interval

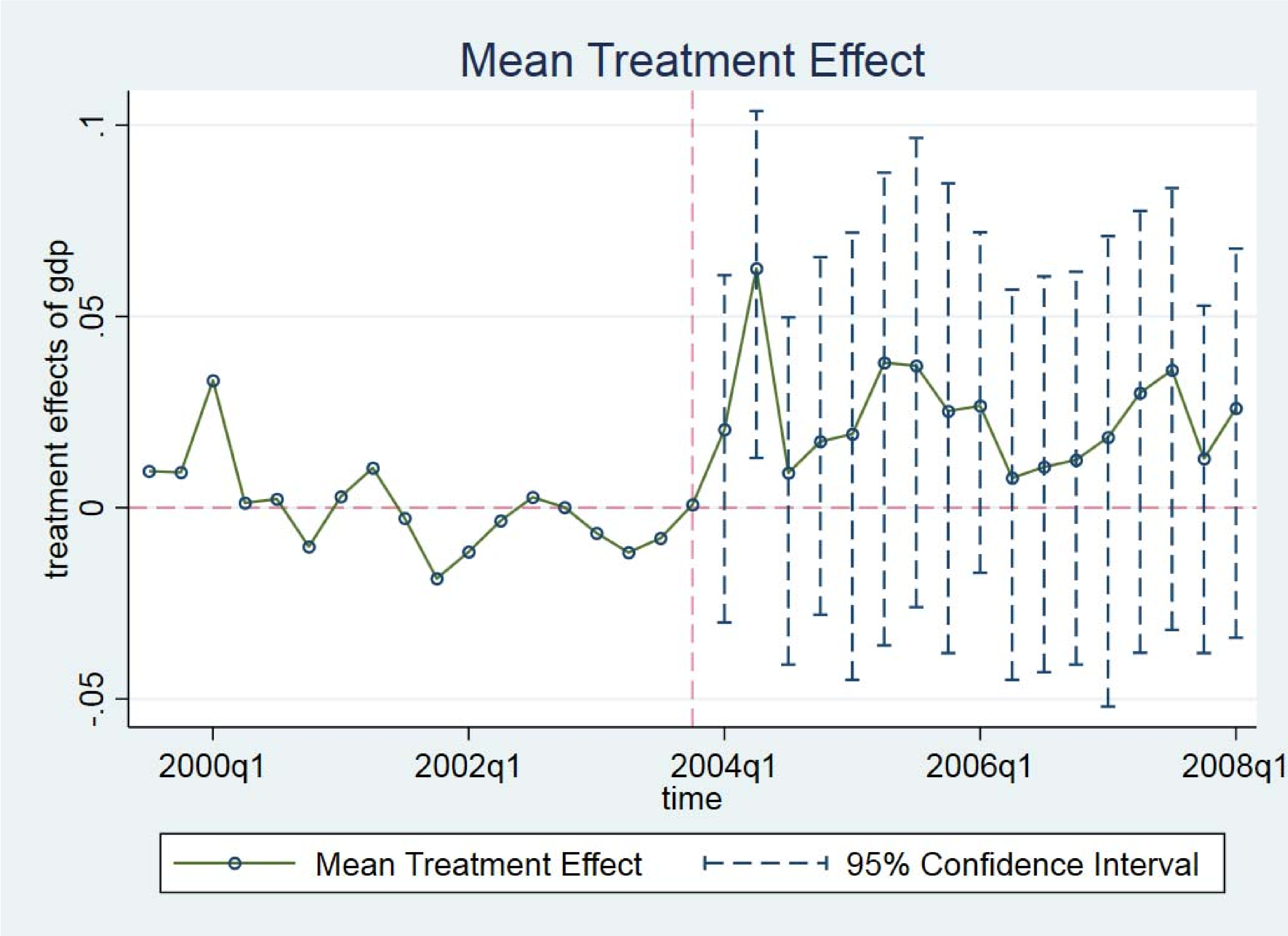
Specify Pre-treatment Periods

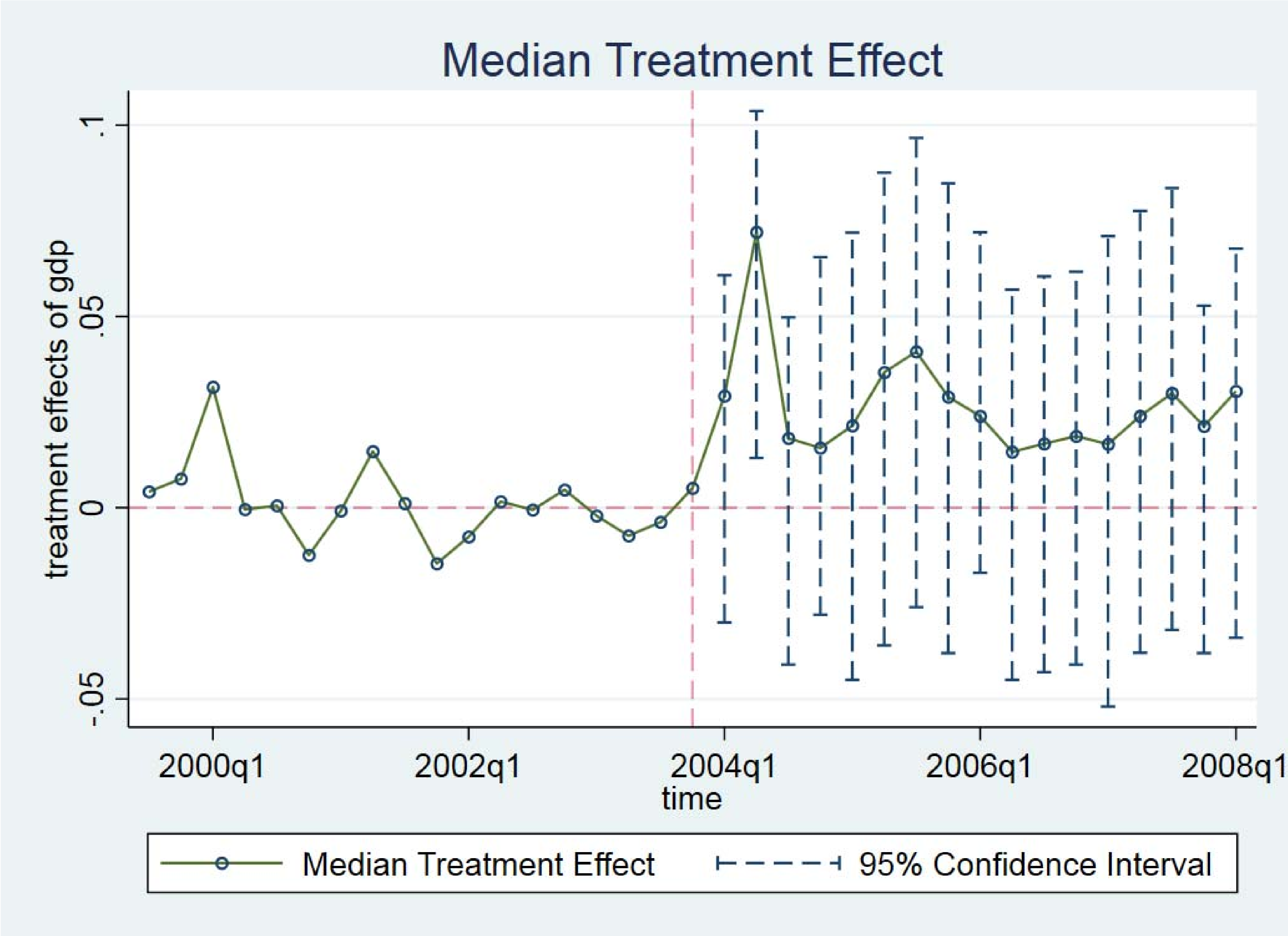
- Let pre-treatment periods start from 1999Q3 (2 years after 1997Q3) instead of 1993Q1

- `display tq(1999q3)`
- `158`

- `qcm gdp, trunit(9) trperiod(176) preperiod(158/175)`


Actual and Predicted Lines
gdp
.15
.1
.05
0
-.05
2000q1
2002q1
2004q1
2006q1
2008q1
time
Actual
Mean Pred.
Median Pred.


Mean Treatment Effect
treatment effects of gdp
.1
.05
0
-.05
2000q1
2002q1
2004q1
2006q1
2008q1
time
Mean Treatment Effect
95% Confidence Interval


Median Treatment Effect
.1
.05
0
-.05
treatment effects of gdp
2000q1
2002q1
2004q1
2006q1
2008q1
time
Median Treatment Effect
95% Confidence Interval

More Options

Title

qcm — implementation of the quantile control method for program evaluation

Syntax

qcm *depvar* [*indepvars*] [*if*] **,** **trunit(#)** **trperiod(#)** [*options*]

options	Description
Model	
ctrlunit(*numlist*)	control units to be used as the donor pool
preperiod(*numlist*)	pre-treatment periods before the intervention occurred
postperiod(*numlist*)	post-treatment periods when and after the intervention occurred
Optimization	
ntree(*int*)	number of trees to grow
mtry(*int*)	number of predictors randomly selected as candidate splitting va
maxdepth(*int*)	maximum depth of the tree
minlsize(*int*)	minimum number of observations at each leaf node
seed(*int*)	seed used by the random number generator
thread(*int*)	number of threads to run in parallel
Placebo Test	
placebo([**unit** **period(*numlist*)**])	placebo test using fake treatment unit and/or time
Reporting	
frame(*framename*)	create a Stata frame storing dataset with generated variables ir
importance	display feature importances
nofigure	Do not display figures. The default is to display all figures.

xtset *panelvar timevar* must be used to declare a (strongly balanced) panel dataset before **qcm** is impl
depvar and *indepvars* must be numeric variables, and abbreviations are not allowed.

Welcome feedbacks!
Thank You ☺

陈强
山东大学经济学院
qiang2chen2@126.com
www.econometrics-stata.com