fsPDA: R Package

Zhentao Shi and Yishu Wang

May 12, 2021

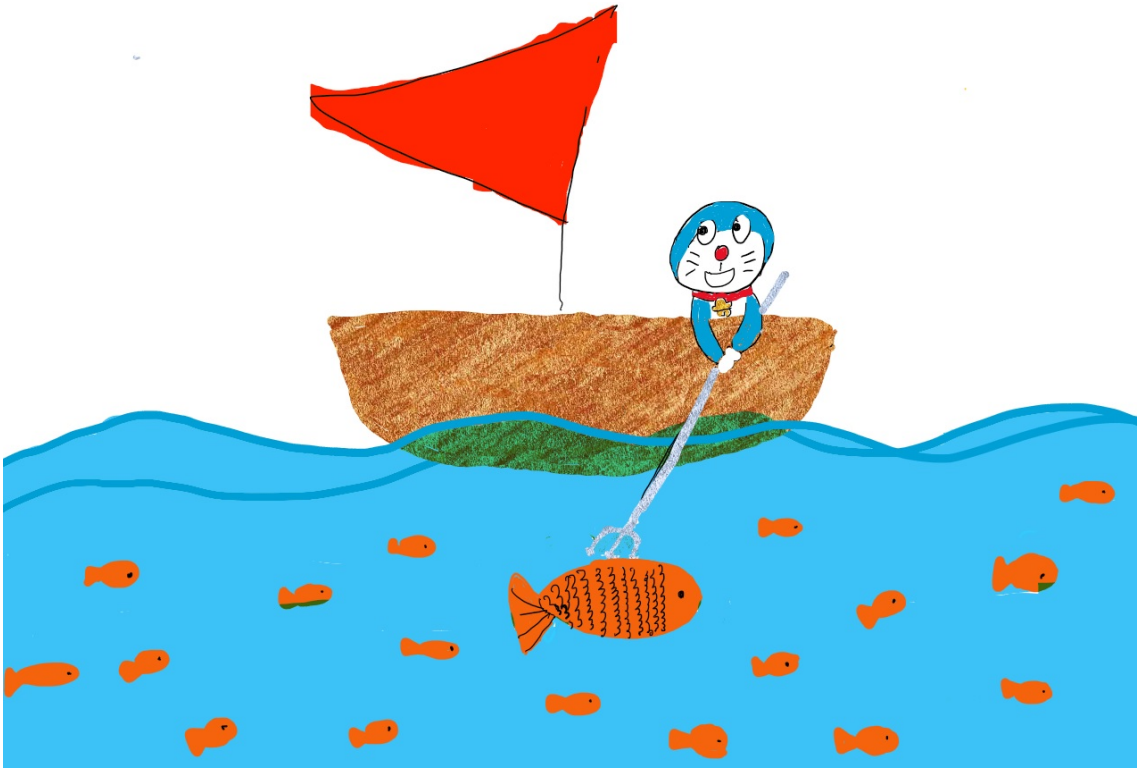


Illustration of fsPDA, by Iris Shi

Introduction

Program evaluation is an important econometric topic. Hsiao, Ching, and Wan (2012)'s *panel data approach* (PDA) is one of the leading methods for program evaluation. To extend PDA to big data environments, Shi and Huang (2021) (arXiv: 1908.05894) propose the *forward-selected panel data approach* (fsPDA) which uses the forward selection algorithm to select a small number of control units. Forward selection is a well-known greedy variable selection method (Hastie, Tibshirani, and Friedman 2009). After forward selection, fsPDA uses the selected control units to run OLS, predicting the counterfactual, and calculating the standard t-statistic for hypothesis testing for the average treatment effect (ATE).

Shi and Huang (2021) bring forth the procedure and establish its asymptotic guarantee. This document introduces the R package `fsPDA` (https://github.com/zhentaoshi/fsPDA) which automates the estimation and inference procedure.

Usage

To install the R package, run

```
devtools::install_github("zhentaoshi/fsPDA/R_pkg_fsPDA")
```

The package is documented with complete help files for all functions and datasets. The work horse function is `est.fsPDA`:

```
est.fsPDA <- function(treated, control, treatment_start,
                      date = NULL, lrvar_lag = NULL)
```

The arguments of the function are

- `treated`: A T-dimensional vector of time series of the treated unit.
- `control`: A $T \times N$ panel matrix with each column being a control unit.
- `treatment_start`: An integer specifying the period when the treatment / intervention starts.
- `date`: A T-dimensional vector of date class or any meaningful numerical sequence. The default setting `NULL` uses all time observations `1:length(treated)`.
- `lrvar_lag`: A non-negative integer for the maximum lag with the Bartlett kernel for the Newey-West long-run variance estimator. The default choice `NULL` specifies `floor((length(treated)-treatment_start+1)^(1/4))`.

This function returns an objected of the class `fsPDA` with the following components:

- `select`: The number and the identities of the selected units when the forward selection is terminated by the modified Bayesian information criterion (BIC).
- `in_sample`: In-sample fitting before the treatment date.
- `out_of_sample`: Out-of-sample counterfactual prediction and the time-varying treatment effect after the treatment date.
- `ATE`: The estimated ATE, the corresponding (long-run) standard error, t-statistic for the test of zero ATE, and the associated p-value.

To help visualize the raw data, the fitted values and the counterfactual predictions, a `plot` method is provided for the class `fsPDA`. See the help file of `plot.fsPDA` about its usage.

Datasets

The package contains two datasets. `china_import` is the China's luxury watch import in the empirical application of Section 5 of Shi and Huang (2021). `HCW` is original dataset from in Hsiao, Ching, and Wan (2012). These two datasets are used to demonstrate the usage of the package.

Demonstrations

First, we replicate the study of China's luxury watch import. The statistical inference outcomes are printed.

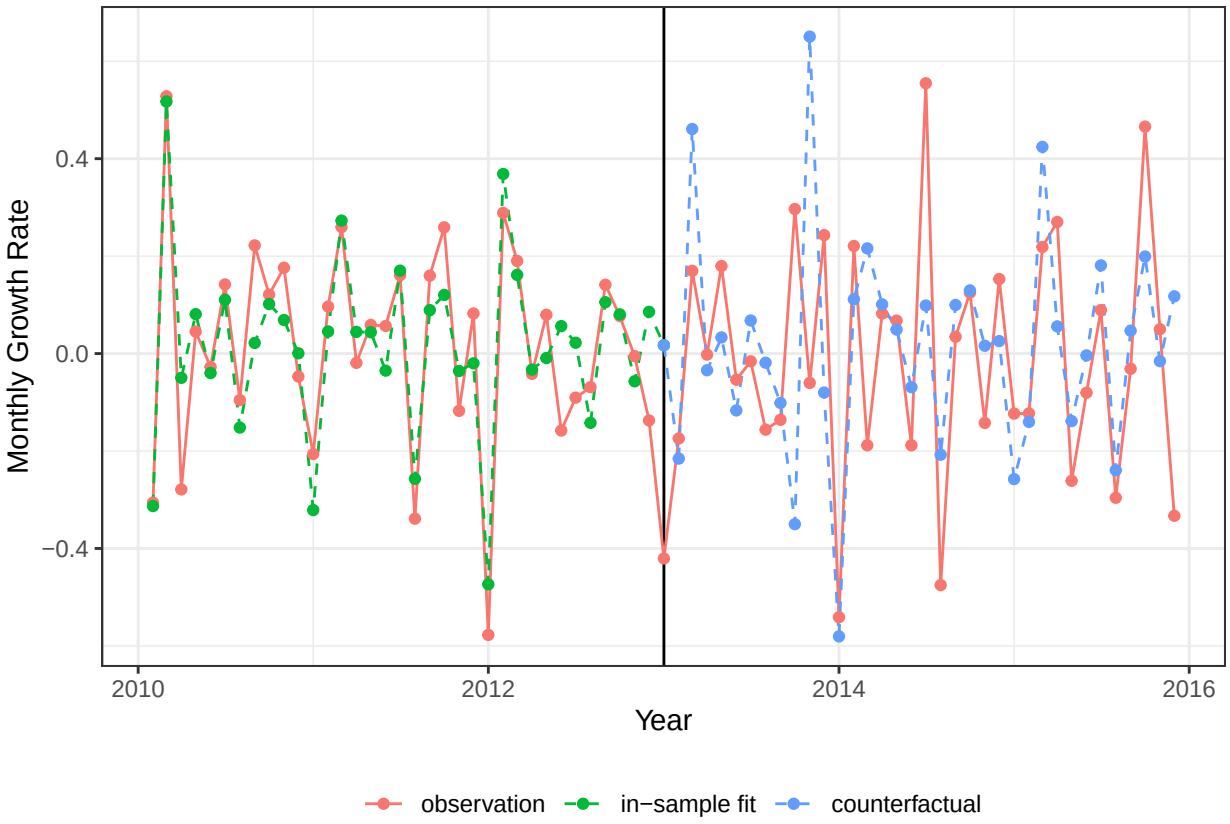
```
library(fsPDA)
data("china_import")
date_import <- names(china_import$treated)

result <- est.fsPDA(
  treated = china_import$treated,
  control = china_import$control,
  treatment_start = which(date_import == china_import$intervention_time),
  date = as.Date(paste(substr(date_import, 1, 4), "-",
                       substr(date_import, 5, 6), "-01", sep = ""))
)

print(result$ATE)
#>          ATE       lrVar      t_stat     p_value
#> -0.03089581  0.02709269 -1.12622398  0.26007073
```

A time series graph with a clear legend is provided by the generic `plot` method.

```
plot(result, tlab = "Year", ylab = "Monthly Growth Rate")
```



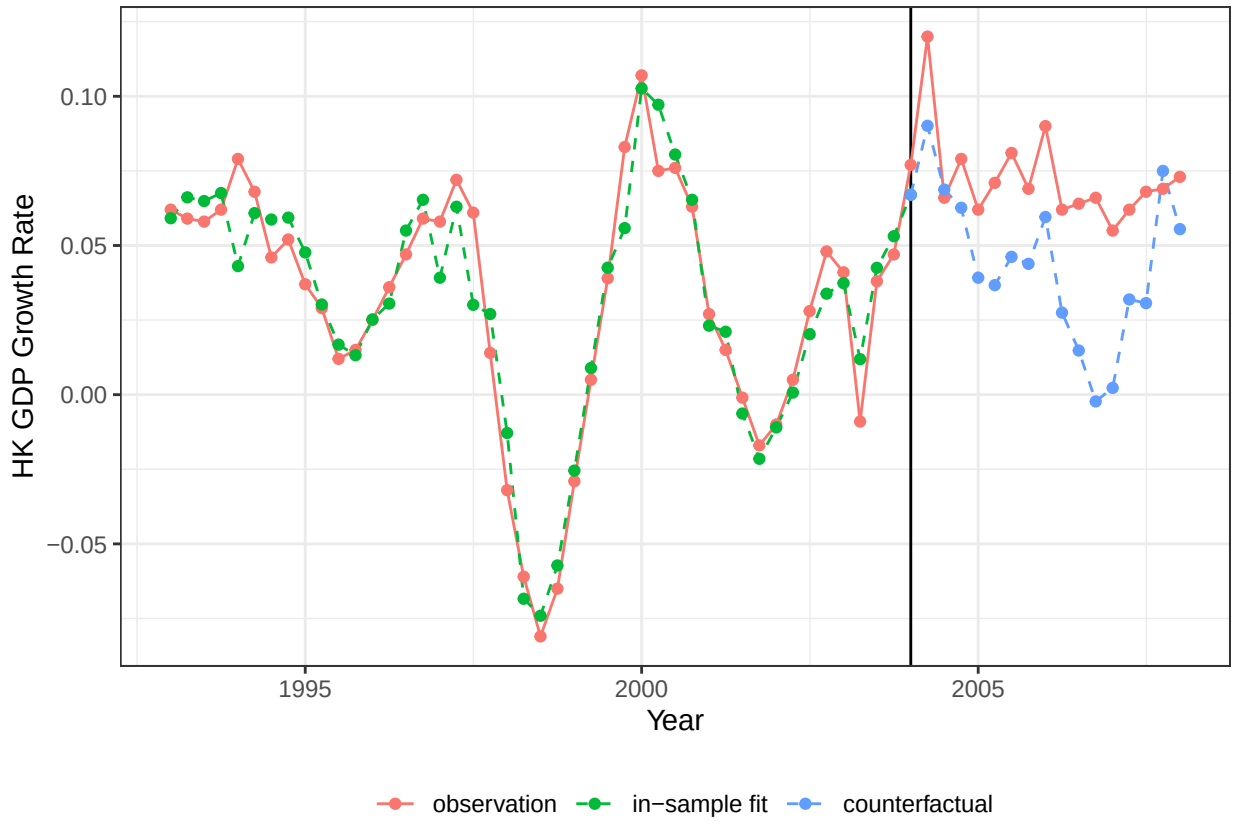
Next, we apply fsPDA to the `HCW` dataset of 24 countries and territories to evaluate the effect of the trade treaty on Hong Kong's GDP growth rate.

```
data("HCW")
result <- est.fsPDA(
  treated = HCW$panel[, 1],
  control = HCW$panel[, -1],
  treatment_start = HCW$T1 + 1,
  date = as.Date(paste(substr(HCW$quarter, 1, 4), "-",
            (as.numeric(substr(HCW$quarter, 6, 6)) - 1) * 3 + 1, "-1", sep = ""))
)

print(result$select$control)
#> [1] "Malaysia"    "New Zealand" "Norway"      "Austria"     "Canada"
#> [6] "Thailand"    "Australia"
```

The forward selection is automatically terminated by the modified BIC after selecting 7 economies. The discrepancy between the realized Hong Kong real GDP growth and the estimated counterfactual can is discernible.

```
plot(result, tlab = "Year", ylab = "HK GDP Growth Rate")
```



Acknowledgement

Shi acknowledges the financial support from the Hong Kong Research Grants Council No.24614817. We thank Jingyi Huang and Zhen Gao for their assistance in developing this R package.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

Hsiao, Cheng, Steve H 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): 705–40.

Shi, Zhentao, and Jingyi Huang. 2021. "Forward-Selected Panel Data Approach for Program Evaluation." *arXiv Preprint arXiv:1908.05894*.