

R 包 fsPDA

史震涛、王一澍 (王珏瑛翻译)

2022 年 8 月 1 日



Illustration of fsPDA, by Iris Shi

前言

项目评估是一个重要的计量经济学主题。Hsiao, Ching, and Wan (2012) 的面板方法 (panel data approach, PDA) 是项目评估的主要方法之一。为了把 PDA 扩展到大数据环境, Shi and Huang (2021) (arXiv: 1908.05894) 提出了前向选择面板数据方法 (fsPDA), 它使用前向选择算法来选择少量的控制单元。前向选择是一种众所周知的贪婪变量选择方法 (Hastie, Tibshirani, and Friedman 2009)。在前向选择完成之后, fsPDA 使用选定的控制单元来运行 OLS, 以预测反事实, 并计算标准 t 统计量以进行平均处理效果 (average treatment effect ATE) 的假设检验。

Shi and Huang (2021) 提出整个过程并建立其渐进性质。本文档介绍了 R 包 fsPDA (<https://github.com/zhentaoshi/fsPDA>), 它可以自动进行估计和推断。

用法

要安装这个 R 包, 请运行

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

该软件包记录了所有函数和数据集的完整帮助文件。其中的核心函数是 `est.fsPDA`:

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

该函数的参数是:

- `treated`: 被处理单元时间序列的 T 维向量。
- `control`: $T \times N$ 的面板矩阵, 其中每列是一个控制单元的时间序列。
- `treatment_start`: 一个指定处理/干预开始时间段的整数。
- `date`: 日期类的 T 维向量, 或任何有意义的数字序列。默认设置 `NULL` 使用所有时间观察 `1:length(treated)`。
- `lrvar_lag`: 用于 Newey-West 长期方差估计当中 Bartlett 核的最大滞后期数。它是一个非负整数。默认设置 `NULL` 指定 `floor((length(treated)-treatment_start+1)^(1/4))`。

运行该函数后将产生一个 `fsPDA` 类 (class) 的对象, 其包含以下组件:

- `select`: 当前向选择被贝叶斯信息准则 (BIC) 终止时, 所选控制单元的身份。
- `in_sample`: 处理日期前的样本内拟合。
- `out_of_sample`: 样本外反事实预测以及处理日期后随时间变化的处理效果。
- `ATE`: 估计的 ATE、相对应的 (长期) 标准误差、零 ATE 检验的 t 统计量, 以及相关的 p 值。

为了帮助可视化原始数据、拟合值以及反事实预测, 我们为类 `fsPDA` 提供了 `plot` 方法。有关其用法, 请参阅 `plot.fsPDA` 的帮助文件。

数据集

该软件包包含两个数据集。`china_import` 是 Shi and Huang (2021) 第 5 节实证应用中的中国奢侈手表进口数据。HCW 是来自 Hsiao, Ching, and Wan (2012) 的原始数据集。这两个数据集被用来演示此软件包的使用。

示例

我们首先尝试复制了对于中国奢侈品手表进口的研究。统计推断结果如下。

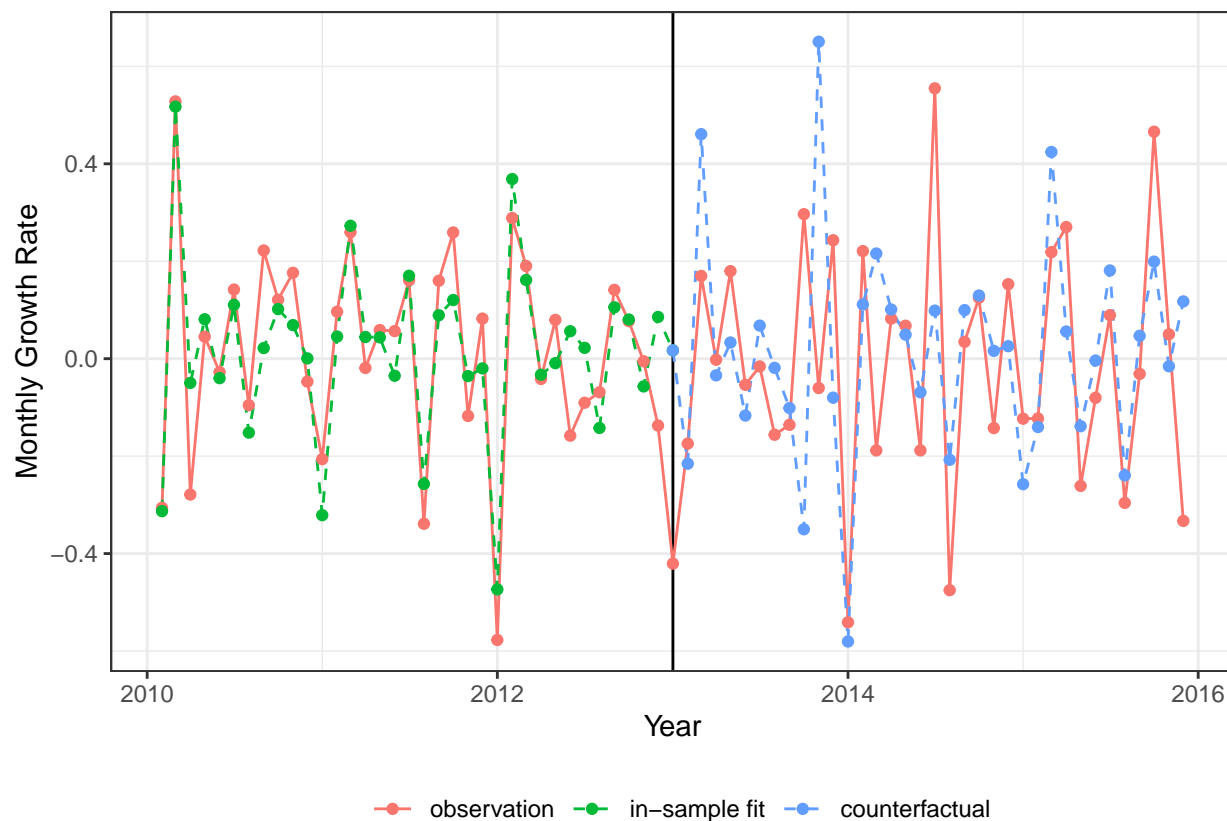
```
library(fsPDA)
data("china_import")
date_import <- names(china_import$treated)

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      se    t_stat    p_value
#> -0.03089581 0.02743310 -1.12622398 0.26007073
```

通用的 plot 方法提供了一个带有图例的时间序列图表。

```
plot(result, tlab = "Year", ylab = "Monthly Growth Rate")
#> Warning: package 'ggplot2' was built under R version 4.1.3
```



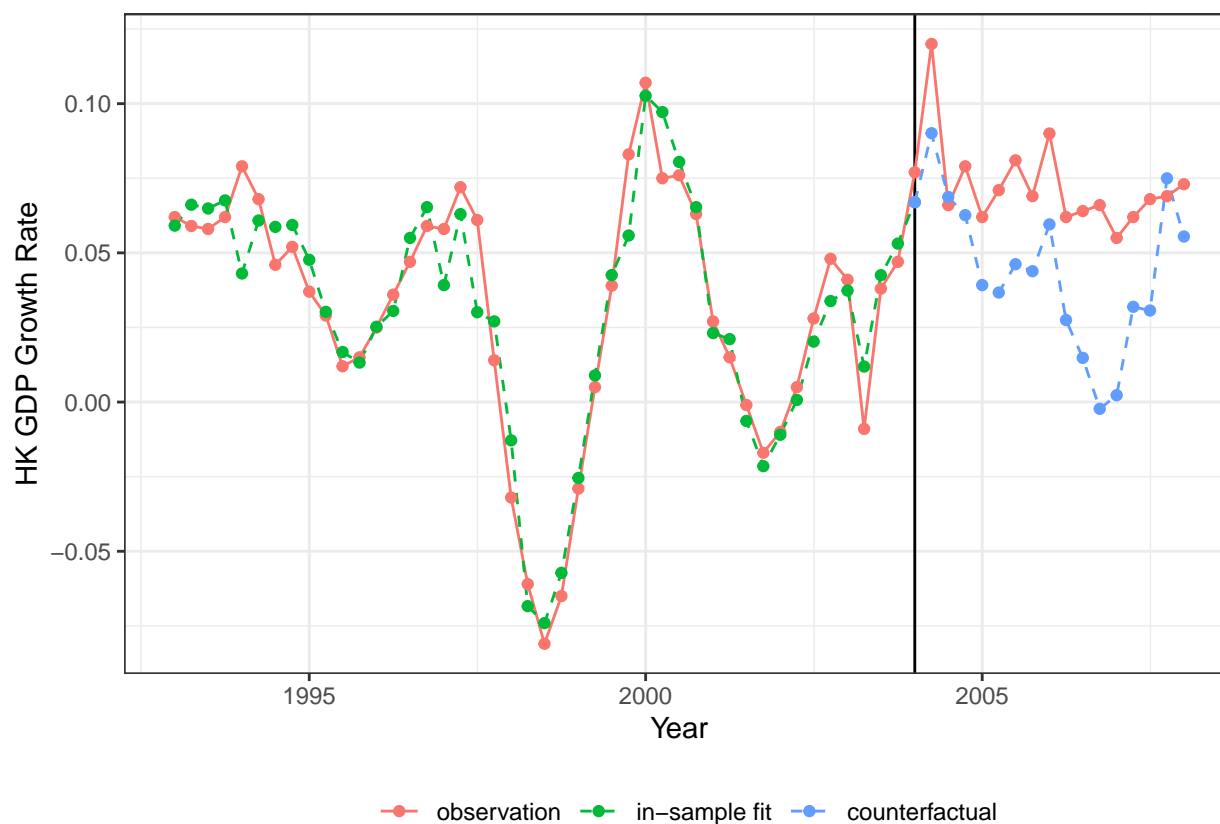
接着，我们将 fsPDA 应用于包含了 24 个国家和地区的 HCW 数据集，以评估贸易协定对香港 GDP 增长率的影响。

```
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"
```

选择七个经济体后，前向选择将由 BIC 自动终止。已实现的香港实际 GDP 增长率与估计的反事实之间的差异是显而易见的。

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



致谢

我们感谢黄静宜和高展对开发这个 R 包的帮助。

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*.