



Universidade de Brasília
Departamento de Estatística
EST0077 - 2023/1

SOLUÇÃO DOS EXERCÍCIOS PARA ENTREGA 5

Igor de Oliveira Barros Faluhelyi
Prof. George

Brasília
2023

Sumário

1 Exercício 32 da Lista 5	4
1.1 Ajuste no banco de dados	4
1.2 Resumo dos resultados.	5
1.2.1 <i>Leave-One-Out</i>	5

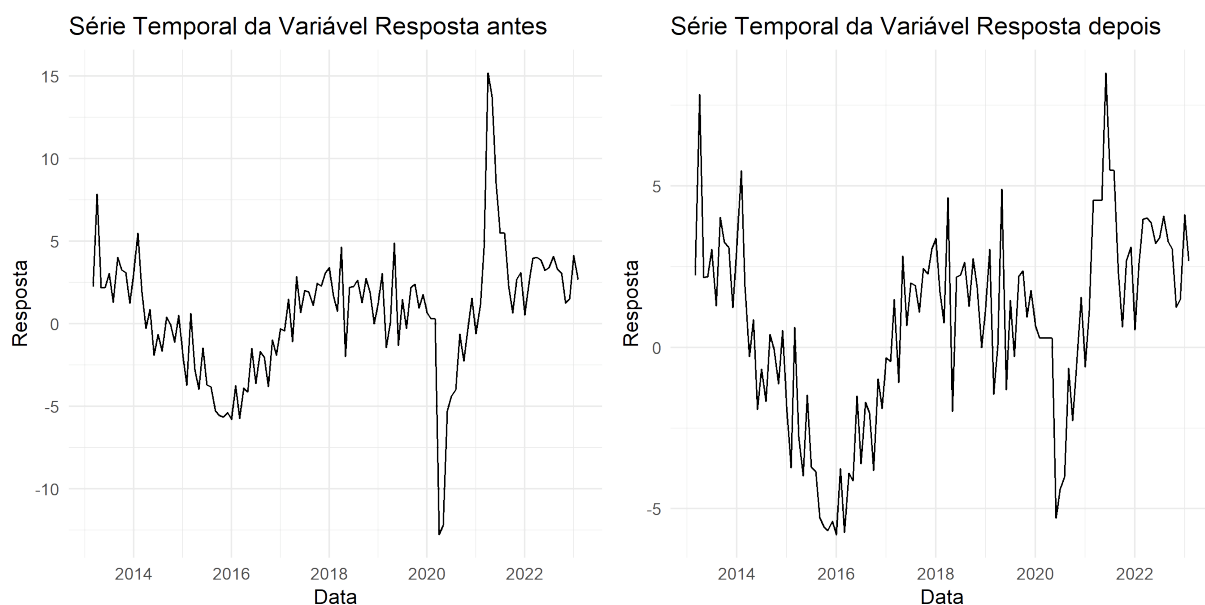
1 Exercício 32 da Lista 5

Exercício 14 do Capítulo 3 de James et al. (2021).

1.1 Ajuste no banco de dados

Antes de propor a solução para o exercício, foram feitas algumas alterações em períodos específicos nas séries temporais. Fixados os períodos, as mudanças seguem a mesma regra para todas as variáveis. No primeiro passo, os valores observados para todas variáveis nos meses nos anos "2020-04", "2020-05", "2021-04" e "2021-05" foram substituídos por *missing values*. No segundo momento, é feita a imputação desses "dados faltantes" pelo respectivo valor observado no mês anterior.

Para ilustrar o resultado do procedimento, segue o plot para a série temporal da variável resposta, respectivamente, antes e depois:



Veja que o comportamento geral da série se mantém, porém o efeito dos períodos, que antes geravam dados discrepantes, mencionados vai poder ser melhor captado pelo modelo linear ao decorrer do exercício e o ajuste aos dados vai ser melhor. A argumentação segue válida para todas as covariáveis do exercício.

1.2 Resumo dos resultados

Os códigos R foram escritos e executados em um computador com Windows 10 Pro; Intel(R) Core(TM) i5-10500 CPU @ 3.10GHz 3.10 GHz como Processador e 8.00 GB (utilizável: 7.72 GB) de RAM instalada.

1.2.1 *Leave-One-Out*

Segue, na tabela, o respectivo RMSEP para as diferentes retenções de informação em **X** considerando PCR ou PLS.

Retenção	# Componentes	RMSEP		
		PCR próprio	PCR no pls do R	PCR no pls do R
70%	4	1.1386270	1.1351774	0.9153081
80%	8	0.9768882	0.9768882	0.7565003
90%	17	0.9811538	0.9811014	0.6518926
95%	29	0.9760095	0.9444652	0.7177428

O RMSEP pela implementação do pacote pls do R não bate exatamente com a implementação manual do PCR, pois, esta, considera um número de CPs variável para cada iteração do LOOCV, visto a retenção especificada. Então a coluna de Números de componentes principais da tabela anterior é exatamente válida para os RMSEPs calculados pela implementação no pacote pls e, como *Trade off*, a coluna de Retenção não vale exatamente para eles. O exato contrário vale para a implementação manual, isto é, a coluna de Retenção vale exatamente, porém a coluna de Número de componentes seria uma média dos números de componentes utilizadas nas iteração do LOOCV para a respectiva retenção. Sendo essas médias: 3.991667, 8, 16.95 e 29.05. Por isso na retenção de 80% os RMSEP "batem na vírgula".

O número de componentes utilizados no modelo linear em cada iteração do LOOCV com a retenção especificada segue na relação:

1. Retenção de 70% 99% das iterações usaram 4 CPs e 1% usaram 3;
2. Retenção de 80% 100% das iterações usaram 8 CPs;
3. Retenção de 90% 99% das iterações usaram 17 CPs e 1% usaram 16;
4. Retenção de 95% 95% das iterações usaram 29 CPs e 5% usaram 30.

Retenção de 70% 99% das iterações usaram 4 CPs e 1% usaram 3

70% 99% 4 e 1% 3 80% 100% 8