Assignment 2 (Due Date is 23:59pm 23rd Sep 2022)

1 Data Introduction

This assignment considers an empirical study on the relationship between health-care expenditure (HE) and income (i.e. GDP). In health economics, the income elasticity of health-care expenditure is defined as the percentage change in health-care expenditure in response to the percentage change in income. This elasticity can provide important policy implications since if its value is lower than one, health-care belongs to necessity good. Otherwise, it is a luxury good and therefore more sensitive to the market forces. There has been substantial literature on estimating the income elasticity of health-care expenditure in the last four decades.

In the literature, many variables have been identified as potential drivers of health-care expenditure (HE) in addition to the income variable (GDP). We select three additional explanatory variables which are widely accepted by the literature. They are the proportions of population aged below 14 and above 65 over all population (DR_{young} and DR_{old}, respectively), which reflect the dependency rates on age structure; and the proportion of government fundings invested on health care industry (GHE).

For these five variables (response variable: HE; predictors: GDP, DR_{young}, DR_{old}, GHE), we collect an annual data set from 1971 to 2013 ($T = 43$) on 18 OECD countries ($N = 18$). Use the observations from the year 1971 to the year 2000 as training data and annual observations from the year 2001 to the year 2013 as test data.

Please attach all the R codes (for solving the following questions) to your submission.

2 Questions

Question 1 [20 marks]: For each county i (here $i = 1, 2, \ldots, 18$), consider a general nonparametric regression model

$$y_t^{(i)} = g_i\left(x_{1t}^{(i)}, x_{2t}^{(i)}, x_{3t}^{(i)}, x_{4t}^{(i)}\right) + \varepsilon_t^{(i)}, \quad t = 1, 2, \ldots, T. \tag{2.1}$$

where $y_t^{(i)}$ is the observation for the response variable HE in the year t for country i; $x_{1t}^{(i)}, x_{2t}^{(i)}, x_{3t}^{(i)}, x_{4t}^{(i)}$ are observations for predictors GDP, DR_{young}, DR_{old}, GHE, respectively, in the year t for country i; and $\varepsilon_t^{(i)}$ is the error component in year t for country i. The unknown function $g_i(\cdot) : \mathbb{R}^4 \to \mathbb{R}^1$ is regression function for country i.

For each country, first apply kernel smoothing method and smoothing spline method, respectively, to estimating the regression function $g_i(\cdot)$. Next, obtain the prediction values for the response variable $y_t^{(i)}$ for the training data and test data. At last, please calculate the training MSE and test MSE for predicted response variable.

Question 2 [20 marks]: For model (2.1), we assume that all the 18 countries share the same regression function, i.e.

$$g(\cdot) := g_1(\cdot) = g_2(\cdot) = \cdots = g_{18}(\cdot). \tag{2.2}$$

Under this homogeneous regression assumption, please conduct smoothing spline method to estimate the common regression function $g(\cdot)$. Please calculate the test MSE with the test data.

Question 3 [10 marks]: Compare the test MSE derived in Question 2 and Question 1 (for smoothing spline method). Which model ((2.1) in Question 1 or (2.2) in Question 2) results in smaller test MSE? Please interpret your conclusion.

Question 4 [20 marks]: For model (2.2), suppose the regression function satisfies an additive structure below

$$g\left(x_{1t}^{(i)}, x_{2t}^{(i)}, x_{3t}^{(i)}, x_{4t}^{(i)}\right) := f_1\left(x_{1t}^{(i)}\right) + f_2\left(x_{2t}^{(i)}\right) + f_3\left(x_{3t}^{(i)}\right) + f_4\left(x_{4t}^{(i)}\right). \tag{2.3}$$

Please estimate the unknown functions f_1, f_2, f_3, f_4 with the training data and then predict the response variable $y_t^{(i)}$ for the test data. Please provide the test MSE.

Question 5 [20 marks]: For model (2.2), suppose the regression function satisfies an additive structure below

$$g\left(x_{1t}^{(i)}, x_{2t}^{(i)}, x_{3t}^{(i)}, x_{4t}^{(i)}\right) := \gamma_1\left(x_{1t}^{(i)}, x_{2t}^{(i)}\right) + \gamma_2\left(x_{3t}^{(i)}\right) + \gamma_3\left(x_{4t}^{(i)}\right). \tag{2.4}$$

Please estimate the unknown functions $\gamma_1, \gamma_2, \gamma_3$ with the training data and then predict the response variable $y_t^{(i)}$ for the test data. Please provide the test MSE.

Question 6 [10 marks]: Compare the test MSE derived in Question 4 and Question 5. Which model results in smaller test MSE? Please interpret your conclusion.