Princeton University **Spring 2021**
Department of Economics **ECO 518**

Exercise 7

Version 1 (March 26, 2021)

Mikkel Plagborg-Møller

Due on Canvas by 2.59 pm on Monday, April 5

The problem set should be submitted as a single PDF file on the course website. You may turn in one problem set per group of at most 3 students. You may discuss the exercises with any of your classmates. Please attach your code to the problem set. Use any software you like, although you learn the most by avoiding canned statistics/econometrics packages. If you find any errors or require clarification, please let me know right away.

Question 1

The data set `engel.csv`, which is available on the course website, was used by Ernst Engel in 1857 to support the proposition that the food expenditure share declines with personal income (the "Engel curve"). This file contains 235 observations on income and food expenditure for 19th century Belgian working class households. Each line contains the variables for one household. The first variable is annual household food expenditure in Belgian francs (`foodexp`). The second variable is annual household income in Belgian francs (`income`).

i) Using a normal kernel and the normal reference rule, estimate the density of log income at each value of X, where X is the log income data. Plot the estimated density of log income.

ii) Estimate the regression of *share* of food expenditures on log income (i.e. the Engel curve) using Kernel Regression with a normal kernel in three steps:

   (a) For Kernel Regression, calculate and plot the cross-validation criterion as a function of h (the bandwidth) on the interval $(0, 1]$ in 0.01 increments (e.g. $h = 0.01, 0.02, \ldots, 0.99, 1$).

   (b) Approximate the optimal bandwidth h_{CV} using a grid search over the interval $(0, 1]$ in 0.01 increments. (A grid search is a way to approximate the critical value at which an objective function obtains an extremum. To find the critical value of an

objective function using a grid search, evaluate the objective function at incremental values over the relevant interval and then choose as the critical value the point at which the function obtains its extremum on that interval.)

(c) Use the optimal bandwidth to estimate the Kernel Regression of share of food expenditures on log income at each value of X, where X is the log income data. Plot the estimated regression curve and a scatterplot of the data in the same graph.

iii) Estimate the regression of share of food expenditures on log income (i.e. the Engel curve) using Local Linear Regression with a normal kernel in three steps:

(a) For Local Linear Regression, calculate and plot the cross-validation criterion as a function of h (the bandwidth) on the interval $(1, 2]$ at 0.01 increments (e.g. $h = 1.01, 1.02, \ldots, 1.99, 2$).

(b) Approximate the optimal bandwidth h_{CV} using a grid search over the interval $(1, 2]$ in 0.01 increments.

(c) Use the optimal bandwidth to estimate the Local Linear Regression of share of food expenditures on log income at each value of X, where X is the log income data. Plot the estimated regression curve and a scatterplot of the data in the same graph.

iv) Estimate the regression of share of food expenditures on log income (i.e. the Engel curve) using Polynomial Series Regression in three steps:

(a) For Polynomial Series Regression, calculate and plot the cross-validation criterion as a function of p (the order of the polynomial) on the grid $\{1, 2, \ldots, 10\}$.

(b) Select the polynomial order that minimizes the cross-validation criterion.

(c) Use the optimal polynomial order to estimate the Polynomial Series Regression of share of food expenditures on log income at each value of X, where X is the log income data. Plot the estimated regression curve and a scatterplot of the data in the same graph.

v) Interpret the source and nature of the differences between the plots in parts (ii)–(iv) above. Be brief and concrete.

Question 2

We are interested in measuring the predictive effect of years of education X on later-in-life (log) earnings Y, but we also wish to control for the effect of years of work experience R.

Since work experience might theoretically have a nonlinear effect on earnings, we want to allow R to affect Y in a flexible way.

The Partially Linear Regression Model assumes

$$Y = \beta_0 X + g_0(R) + \varepsilon, \quad E[\varepsilon \mid X, R] = 0. \tag{1}$$

Here $\beta_0 \in \mathbb{R}$ and $g_0 \colon \mathbb{R}_+ \to \mathbb{R}$ are an unknown scalar and unknown function, respectively. We observe data on (Y, X, R).

The data set `nls.csv`, available on the course website, contains data for 929 individuals from the National Longitudinal Survey in a particular year. `luwe` is log weekly earnings (Y), `educ` is years of education (X), and `exper` is years of experience (R).

i) Assume that $g(\cdot)$ is a second-order polynomial. Compute an estimate of β_0 using OLS. This is the traditional Mincer (1974) approach to estimating the returns to schooling, while allowing for nonlinear effects of experience on earnings.

ii) Assuming only that $g(\cdot)$ is a smooth function, provide an estimate of β_0 using the following implication of the model (1):

$$Y - E[Y \mid R] = \beta_0 (X - E[X \mid R]) + \varepsilon.$$

To do this, first estimate the two conditional expectation functions using Nadaraya-Watson Kernel Regression. Use any kernel and bandwidth selection procedure you like, but make sure to specify your choices. [Hint: If you need further details on this approach, google Robinson's (*Econometrica* 1988) "double residual" method.]

This is an example of a "semi-parametric" method. Although the model contains an infinite-dimensional parameter $g(\cdot)$, ultimately we are interested in doing inference on a finite-dimensional parameter β_0.

iii) (OPTIONAL!) Provide standard errors for your point estimates in (i) and (ii) using the nonparametric boostrap. For simplicity, when bootstrapping the estimator in (ii), fix the bandwidth at the value that you used in the actual data.