Part 1: Theory

To estimate the conditional mean $m(x)$ using local linear regression smoothing at a given point x one begins by fitting the model

$$y_i = \theta_1 + \theta_2 x_i + e_i, \quad i = 1, 2, \ldots, n,$$

(with the usual assumptions) using the method of weighted least squares, where the weights, w_i, are determined by some weighting $w(x_i - x, h)$. The estimator is then given by $\hat{m}(x) = \hat{\theta}_1 + \hat{\theta}_2 x$

(a) Show that

$$\hat{\theta}_1 = \frac{(\Sigma w_i y_i)(\Sigma w_i x_i^2) - (\Sigma w_i x_i y_i)(\Sigma w_i x_i)}{(\Sigma w_i)(\Sigma w_i x_i^2) - (\Sigma w_i x_i)^2}$$

$$\hat{\theta}_2 = \frac{(\Sigma w_i x_i y_i)(\Sigma w_i) - (\Sigma w_i x_i)(\Sigma w_i y_i)}{(\Sigma w_i)(\Sigma w_i x_i^2) - (\Sigma w_i x_i)^2},$$

(b) Bowmann and Azzalini (1997) consider the parameterization:

$$y_i = \theta_1 + \theta_2 (x_i - x) + e_i, \quad i = 1, 2, \ldots, n,$$

and give the following estimator of $m(x)$:

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{[s_2(x;h) - s_1(x;h)(x_i - x)]\, w(x_i - x; h) y_i}{s_2(x;h) s_0(x;h) - s_1(x;h)^2},$$

where $s_r(x,h) = \frac{1}{n} \sum (x_i - x)^r w(x_i - x; h)$. Show that this is equivalent to the estimator given in (a).

Part 2: Practical

(a) The point of this exercise is investigate the behaviour of the above estimator for a particular case in which the model is known, namely:

$$y_i = m(x_i) + e_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \theta_4 x_i^3 + e_i,$$

where $e_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, 2, \ldots, n.$

(i) Write an **R**–function `genreg(x,theta,sigma)` that generates realizations from the model, and then use it to generate a sample of size 20 with $x_i = i, i = 1, 2, ..., 20$; $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)' = (100, 15, -2.5, 0.10)'$; $\sigma^2 = 10^2$. Plot the data and m(x). [Hint: Use your function `genreg` with $\sigma^2 = 0$ to generate "observations" $y_i' = m(x_i)$.]

(ii) To assess the effect of changing the bandwith use the function `sm.regression` in the library `sm` to estimate $m(x)$ using each of the following bandwidths: $h = 0.5, 1, 3, 10$.

(iii) Now assess the effect of changing the bandwith on the expectation of the estimator , $\mathrm{E}\hat{m}(x)$ using each of the following bandwidths: $h = 0.5, 1, 3, 10$. [Hint: A simple way to do this is to repeat (ii) using y_i' (that you computed in (i)) instead of y_i .]

(b) The file "corruption.dat" contains 6 indicators for each of 102 countries. The indicators (columns) are CPI (Corruption Perception Index), Investment/GDP, Government deficit/GDP, GDP/Population in 1970, Reserve/GDP, Fuels & minerals/GDP. The purpose of this exercise is for you to explore the relationship between CPI (the target variable) and the other indicators, both individually and also in pairs.

(i) Read the data in the file "corruption.dat" into a 102×6 matrix X, and the abbreviated names of the countries that are contained in the file "cabbrev.txt". [Hint: Use `cn<-scan("<path>cabbrev.txt",what="character")` to read the names.]

(ii) Use the command "`pairs`" to display the scatterplots of all the pairs of variables.

(iii) Use "`sm.regression`" assess how CPI depends on some of the covariates, in particular on "GDP/Population in 1970". Identify "interesting" points on the plot using the function "`identify`" making use of the country names in "`cn`".

(iv) Use "`sm.regression`" assess how CPI depends on "GDP/Population in 1970" and "Investment / GDP", i.e. an example with two covariates. [Note that you need to give two smoothing constants in this case.]