Problem 1

Consider a nonparametric regression problem where we observe the data $\{(x_i, y_i)\}_{i=1}^n$. Suppose $x_i = i/n$, and y_i follows the model

$$y_i = f(x_i) + \epsilon_i, \tag{1}$$

with unknown f, and $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Consider the local polynomial regression model

$$\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i \le n} K_h(x_i - x_0) \left(y_i - \left[\beta_0 + \beta_1 \left(\frac{x_i - x_0}{h} \right) + \cdots + \beta_k \left(\frac{x_i - x_0}{h} \right)^k \right] \right)^2,$$

where we estimate $f(x_0)$ with

$$\hat{f}(x_0) := \hat{\beta}_0$$

(a) Verify that the local polynomial regression of order k is a *linear estimator*, i.e.,

$$\hat{f}(x_0) = s_{x_0}^\top y.$$

Define $z_i(x_0) := \left(1, \frac{x_i - x_0}{h}, \ldots, \left(\frac{x_i - x_0}{h}\right)^k\right)^\top$, $Z_{x_0} := [z_{x_0}(x_1), \ldots, z_{x_0}(x_n)]^\top$, $W_{x_0} = \operatorname{diag}[K_h(x_1 - x_0), \ldots, K_h(x_n - x_0)]$ and $e_1 = (\underbrace{1, 0, \ldots, 0}_{k+1})^T$ and compute the explicit form of s_{x_0}.

(b) Consider Model (1), with $n = 100$ and $k \in \{1, 2\}$. Code the explicit formula of s_{x_0} in `R`, with $K_\sigma(z)$ a Gaussian kernel with bandwidth (standard deviation) $\sigma = 0.2$. For each $x_0 \in \{0.1, 0.5, 0.9\}$ compute the n dimensional vector s_{x_0}, and make the scatter plot $(x_i, s_{x_0,i})_i$. This displays how the data are linearly 'combined' to return an estimate at the point x_0. Compare with the coefficients you would get from a Nadaraya-Watson estimator with a Gaussian kernel. Give a short writeup stating comparisons.

Problem 2

This is a follow up of Problem 2 in Assignment #2.
Suppose $x_i \overset{iid}{\sim} U[-1, 1]$, and

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n = 100$$

where $\epsilon_i \sim N(0, 1)$. Consider the following three unknown functions

(a) $f(x) = sin(x * \pi)$.

(b) $f(x) = 2x + x^3 - 6x^4$.

(c) $f(x) = \frac{1}{1 + (5x)^2}$.

Generate one dataset for each of the given functions and apply the following estimators.

- Nadaraya-Watson with a "gaussian" kernel. Fix the bandwidth to be the 'oracle' bandwidth (The 'oracle' bandwidth is the h that minimizes the misfit from f, computed by assuming f known)
- Local Polynomial of degree 2 with a "gaussian" kernel. Fix the bandwidth to be the 'oracle' bandwidth. Use the function `locpoly` in the package `KernSmooth`.
- Natural cubic B-spline. Choose the oracle number of breakpoints. Let `R` choose the locations of these breakpoints by setting `knots=NULL` as a parameter of `ns`.

Plot the estimated functions and give a brief discussion of the results.