B. Applied Statistics II

4. Consider a one-dimensional nonparametric regression problem of the form

$$Y_i = f(x_i) + \sigma\epsilon_i, \quad i = 1, \dots, n$$

where ϵ_i denote i.i.d. random vaiables with zero mean and standard deviation of 1. Assume that

$$0 \leqslant x_1 < x_2 < \cdots < x_n \leqslant 1.$$

(a) [5 marks] Define the *Nadaraya-Watson kernel estimator* and its parameters. Give the definition of a *linear smoother*. Show that the Nadaraya-Watson kernel estimator is a linear smoother.

(b) [2 marks] Show that the estimator $\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$ is a linear smoother. What are its degrees of freedom?

(c) [5 marks] Let $\widehat{f}$ be a linear smoother. State one estimator $\widehat{\sigma}^2$ for σ^2 and compute its expectation. Comment on how the bias of $\widehat{\sigma}^2$ depends on f and $\widehat{f}$.

(d) [4 marks] The `lm()` command in R stands for "linear model". Nevertheless, it can be of use for some nonparametric regression methods. Define a *regression spline* and say briefly why `lm()` can be used to calculate a regression spline.

(e) [6 marks] Consider the following estimator

$$\widehat{f}(x) = \text{median}\left\{Y_i \text{ such that } |x_i - x| \leqslant \lambda\right\}.$$

(i) State one desirable and one undesirable property of this estimator.

(ii) Show that $\widehat{f}$ is not a linear smoother for all λ, by considering $\lambda = 2$ and a dataset $\{i, y_i\}_{1 \leqslant i \leqslant n}$, or otherwise.

5. (a) [11 marks]

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ where F is some unknown distribution and let $\widehat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of θ.

(i) Describe how $\mathrm{Var}\big(\widehat{\theta}_n\big)$ is estimated using the bootstrap method.

(ii) Suppose that an estimator $\widehat{\theta}\,(X_1, \ldots, X_n)$ of θ has an expectation equal to $\theta\,(1+\gamma)$ so that the bias is $\theta\gamma$. The bias factor γ can be estimated by $\widehat{\gamma} = \frac{\mathbb{E}_\tau \widehat{\theta}(X^\star)}{\widehat{\theta}} - 1$ where

$$X^\star = (X_{\tau_1}, \ldots, X_{\tau_n})$$

with $\tau_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}\left(\{1, \ldots, n\}\right)$ (i.e. each τ_i has a uniform distribution on the discrete values $1, \ldots, n$). Show that $\widehat{\gamma}$ is exactly equal to γ in the case of the variance estimate $\widehat{\sigma}^2(X) = \frac{1}{n}\sum(X_i - \overline{X})^2$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

(b) [11 marks]

(i) Define the *breakdown point* and the *asymptotic breakdown point* of an estimator $\widehat{\theta}$. State the asymptotic breakdown point of the median.

(ii) Consider the data points $x_1, \ldots, x_n$ with $x_i \in \mathbb{R}$ for all i. The following estimator

$$\widehat{p} = \text{median}\left\{\frac{x_i + x_j}{2} : i \leqslant j\right\}$$

is called the pseudomedian. Calculate its asymptotic breakdown point.