B. Applied Statistics II

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

\[ Y_i = f(x_i) + \sigma \epsilon_i, \quad i = 1, \ldots, n \]

where \( \epsilon_i \) denote i.i.d. random variables with zero mean and standard deviation of 1. Assume that

\[ 0 \leq x_1 < x_2 < \cdots < x_n \leq 1. \]

(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 \( \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y} \) is a linear smoother. What are its degrees of freedom?

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

(d) [4 marks] The \texttt{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 \texttt{lm()} can be used to calculate a regression spline.

(e) [6 marks] Consider the following estimator

\[ \hat{f}(x) = \text{median} \{ Y_i \text{ such that } |x_i - x| \leq \lambda \}. \]

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

(ii) Show that \( \hat{f} \) is not a linear smoother for all \( \lambda \), by considering \( \lambda = 2 \) and a dataset \( \{i, y_i\}_{1 \leq i \leq n} \), 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 \( \hat{\theta}_n(X_1, \ldots, X_n) \) be an estimator of \( \theta \).

(i) Describe how \( \text{Var}(\hat{\theta}_n) \) is estimated using the bootstrap method.

(ii) Suppose that an estimator \( \hat{\theta}(X_1, \ldots, X_n) \) of \( \theta \) has an expectation equal to \( \theta (1 + \gamma) \) so that the bias is \( \theta \gamma \). The bias factor \( \gamma \) can be estimated by \( \hat{\gamma} = \frac{E[\hat{\theta}(X^*)]}{\theta} - 1 \) where \( X^* = (X_{\tau_1}, \ldots, X_{\tau_n}) \) with \( \tau_i \overset{\text{i.i.d.}}{\sim} U\{1, \ldots, n\} \) (i.e. each \( \tau_i \) has a uniform distribution on the discrete values \( 1, \ldots, n \)). Show that \( \hat{\gamma} \) is exactly equal to \( \gamma \) in the case of the variance estimate \( \hat{\sigma}^2(X) = \frac{1}{n} \sum (X_i - \overline{X})^2 \) 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 \( \hat{\theta} \). 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

\[
\hat{\rho} = \text{median} \left\{ \frac{x_i + x_j}{2} : i \leq j \right\}
\]

is called the pseudomedian. Calculate its asymptotic breakdown point.