Computer Intensive Statistics
Individual HWA3: Kernel Estimation and Regression

1. (1p) Consider the histogram estimator for univariate density. Based on the expression of the bias and variance, derive the expression of MSE and AMISE.

2. (1p) Leave-one-out cross validation can be used to choose the bandwidth in kernel density estimation. We have seen in the slides that, for each h in a pre-specified grid of candidate bandwidths,

$$\mathrm{CV}(h) = \int \hat{f}_h^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i),$$

where

$$\hat{f}_{h,-i}(x) = \frac{1}{(n-1)h}\sum_{j\neq i} K\left(\frac{x-X_j}{h}\right).$$

The second term in $\mathrm{CV}(h)$ can be easily computed from the kernel density estimate. Explain how the first term can be computed in practice. You can take the density estimation problem for the variable bill length in the **Penguin** data set as an example to illustrate your proposal.

3. (1p) We have seen that the kNN estimator of density is

$$\hat{f}_k(x) = \frac{k}{n\mathrm{Vol}\left(B\left(x, R_k(x)\right)\right)}.$$

Explain how such kNN estimator is extended to the kernel kNN estimator

$$\hat{f}_k(x) = \frac{1}{nR_k^d(x)}\sum_{i=1}^{n} K\left(\frac{x-X_i}{R_k(x)}\right).$$

4. (1p) We present the multivariate kernel regression as

$$\hat{m}_H(x) = \int y\frac{\hat{f}_{(X,Y)}(x,y)}{\hat{f}_X(x)}dy.$$

Explain how $f_X(x)$ and $f_{(X,Y)}(x,y)$ are estimated, including the corresponding bandwidth. Derive also the expression of $\hat{m}_H(x)$ as

$$\hat{m}_H(x) = \frac{\sum_{i=1}^{n} K\left[H^{-1}(x-X_i)\right]Y_i}{\sum_{i=1}^{n} K\left[H^{-1}(x-X_i)\right]}.$$

5. (1p) Suppose that we want to regress body mass on bill length and flipper length by local linear regression. Write your own code to perform such regression. Present the estimated regression model using contour plot or a surface plot.