Computer Intensive Statistics
Group HWA3: Kernel Estimation and Regression

1. (1p) Consider estimation of the density of bill length in the **Penguin** data set.

   (a) Write your own code to produce the histogram. For simplicity, take the number of bins to be 20 .

   (b) Write your own code to produce the Gaussian kernel density estimator. For simplicity, use the Silverman's rule-of-thumb to determine the bandwidth.

   Plot both estimates in the same figure.

2. (1p) Consider the kernel density estimate obtain above. Suppose that we want to simulate a data set from the density estimate. Write your own code to generate a sample of size $n = 10000$. Explain also the idea of your random number generator using text and equations. In case you haven't finished Task 1, you can use the built-in function to estimate the kernel density estimate.

3. (1p) Is the univariate kernel density estimator always a probability density?

4. (2p) Suppose that we want to regress the body mass on the flipper length using the local linear regression.

   (a) Write your own code to perform such regression. Plot the fitted function in the scatter plot of the observed data. For simplicity, choose an arbitrary bandwidth yourself.

   (b) Write your own code to perform case bootstrap for such local linear regression.

5. (1p) Consider a simple linear regression with no intercept

   $$y_i = \beta x_i + \epsilon_i.$$

   The lasso estimator for this simple model minimizes

   $$\sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda |\beta| .$$

   In this simple model, we can obtain a closed form expression that is used in the coordinate gradient descent algorithm as

   $$\hat{\beta}^{\text{lasso}} = \begin{cases} \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} - \frac{\lambda}{2\sum_{i=1}^n x_i^2}, & \text{if } \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} - \frac{\lambda}{2\sum_{i=1}^n x_i^2} > 0, \\ \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} + \frac{\lambda}{2\sum_{i=1}^n x_i^2}, & \text{if } \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2} + \frac{\lambda}{2\sum_{i=1}^n x_i^2} < 0, \\ 0, & \text{otherwise.} \end{cases}$$

The residual bootstrap samples the lasso residual $e_i = y_i - \hat{\beta}^{\text{lasso}} x_i$. Investigate whether the residual bootstrap works.

6. (2p) Consider again the local linear regression.
   (a) Develop a leave-one-out cross validation procedure to select the bandwidth.
   (b) Write your own code to implement your leave-one-out cross validation to the regression body mass on flipper length.
7. (1p) Show that the splines is a linear smoother.
8. (1p) Write your own code to estimate the joint density of bill length and bill depth. Present the estimated joint density using contour plot or a 3D plot.