STAT 621 HOMEWORK 5
Due: Monday October 21

1. The data set `heart.txt` is posted on blackboard and contains the ages and survival times after surgery (days) of 184 heart transplant patients.

   (a) Compute and plot a kernel density function for the variable `survival`. Choose any kernel and any method of bandwidth estimation. Interpret the result – what information does the estimate contain about the population of survival times for transplant patients?

   (b) Do the same for the variable `age`. Interpret the result.

   (c) Estimate the *joint* density function for the two variables `survival` and `age`. To do this, load the R library MASS (this library usually comes with R so you don't need to install it). Use the function `kde2d` as described below to estimate and plot the bivariate density of two variables x and y. Just use the default Gaussian product kernel and normal reference bandwidth.

```
library(MASS)
fhat = kde2d(x,y)
contour(fhat, xlab=’x’, ylab=’y’)         # make contour plot
image(fhat, xlab=’x’, ylab=’y’)             # image plot
contour(fhat, add=T)                # overlay contours
```

   Interpret your results. What additional information do you get from the joint density that you don't get from the marginals?

2. <u>MISE and bandwidth:</u> Recall the the expression for MISE

$$\text{MISE}\{\widehat{f}(x)\} = \frac{1}{n\lambda}\int K^2(x)dx + \frac{\lambda^4}{4}\mu_2\int\{f''(x)\}^2dx \tag{1}$$

   where $\mu_2 = \int x^2K(x)dx$.

   (a) For a fixed sample size n, describe in words how MISE behaves as a function of bandwidth λ.

   (b) Investigate this relationship using simulation. Write and R function that estimates MISE for different values of λ. You may want to use the `mise.chi` function in the Density Estimation Part 2 lecture notes as a template. Plot the relationship between MISE and λ, and identify the optimal value of λ in your simulation. Some Suggestions:

      i. Pick a relatively large n, at least 100 or 200

      ii. Use any distribution you like to generate the data

      iii. Generate samples and plot a few estimates first to get an idea of the range of bandwidths to consider. You want to make sure that the MISE is minimized over your range of λs.

      iv. **Extra Credit:** Repeat for several n, or several different distributions of the data. Summarize your results and draw some conclusions.

3. Kernel Regression: The data set `motor.txt` contains measurements on motorcycle test crashes. Here the objective is to model Y= acceleration (in g) as a function of X= time since impact (in millisec). Estimate the regression function in two ways:

   (a) Use a nonparametric kernel regression estimator to fit the function. Plot the data and estimated function. It's fine to just use the `ksmooth` function with defalut kernel and bandwidth (actually the span here which gives the nearest neighbor fraction).

   (b) Try estimating the function using linear regression with a suitable degree polynomial. Plot the data and estimated function.

   Compare your two estimators. Which seems to do a better job? Discuss advantages and disadvantages of each approach.

4. Local Polynomial Regression: The data set `lidar.txt` has 221 observations from a light detection and ranging (LIDAR) experiment. In this problem you will use local polynomial regression to model Y=`logratio` as a function of X=`range`.

   (a) Write down the model and note any assumptions you will make for fitting the model.

   (b) Use the `locfit` function in the package `locfit` to fit the local constant (Nadaraya-Watson) model where `deg=0`. Use GCV to select the optimal value of h; see the example on pages 10-12 of the Nonparametric Regression lecutre notes. Report your h and plot the estimated function.

   (c) Repeat this to estimate the local linear model. You will need to adjust the range of h in the above code, and also the `deg=` argument. Report h. Use the `lines` command to overlay the estimated function on the same plot.

   (d) Repeat again to estimate the local quadratic model. Report h and overlay the estimate on the same plot.

   (e) Briefly discuss the results.

5. Select one of the models you fit in Problem 4 and compute both 95% pointwise and simultaneous confidence bands on the estimated function. Plot each. Discuss the differences in the type of inferences you can make with the two types of intervals. Can you draw any conclutions about the true function or function values?