# 14.386 Problem Set 1
Spring 2007

1. Consider an i.i.d. sample of $(y_i, x_i)$ of log of earnings $y_i$ and explanatory variables $x_i$. Suppose that the data earnings has been topcoded, so that any true log-earnings above a limit $L$ has been recorded as $L$.

   (a) Now suppose that all you are interested in is $E[y|x]$. How could you estimate this without imposing any functional form?

   (b) Is $E[y|x]$ an interesting object in this application? State why or why not.

   (c) Suppose $median(y_i^*|x) = g_0(x)$. What is $median(y_i|x)$? Is $median(y_i|x)$ an interesting object in this application?

   (d) How oculd you nonparametrically estimate $median(y_i|x)$? (Hint: $median(y|x) = \arg\min_\mu E[|y_i - \mu||x]$)

2. A kernel density estimator is $\hat{f}_h(z) = \sum_{i=1}^n K_h(z - z_i)/n, K_h(u) = h^{-r}k(u/h)$ where $r$ is the dimension of $z$. Consider $z$ as fixed at some value. Since $\hat{f}_h(z)$ is just a sample average of a function of the data, when the data are i.i.d. we have

$$Var(\hat{f}_h(z)) = n^{-1}Var\left(K_h(z - z_i)\right).$$

   (a) Give a simple estimator of $Var(\hat{f}_h(z))$. How could you use this to form a confidence interval for $f_0(z)$?

   (b) Let $\hat{m}_h(z) = \sum_{j=1}^n y_i K_h(z - z_i)/n$. How could you estimate the joint covariance matrix of $\hat{m}_h(z)$ and $\hat{f}_h(z)$?

   (c) Use part (b) and the delta method to form an estimator of the asymptotic variance of the kernel regression estimator $\hat{g}_h(z) = \hat{m}_h(z)/\hat{f}_h(z)$.

3. This is an empirical exercise using the gasoline data supplied by the TA.

   (a) Graph an estimator of the density of gasoline consumption (note NOT the natural log of gasoline consumption) for each individual year. Are there any discernible patterns over time?

   (b) For each year separately, estimate a series estimator of a partially linear model

$$\ln q_i = g(\ln p_i, lnI_i) + D_i^{'}\beta + \epsilon_i,$$

   where $\ln q_i$, $\ln p_i$, and $\ln I_i$ are the natural logs of gasoline consumption, price, and income, respectively, and $D_i$ are the location dummies for the year. Choose the number of terms using cross-validation and Mallows criterion. Are there any differences?

   (c) Pick any two years and plot $\hat{g}(\ln p, \ln I)$ as a function of $\ln p$ for your estimator from (b) with $\ln I$ fixed at some value. Do they have similar shape for the two years?

   (d) For your estimates from (c) also plot plus or minus two standard error bands. Do you think that there are statistical differences among the estimates?

   (e) Construct and carry out a formal test of stability across time of your estimates from (c) and (d).