14.385B Problem Set 2

Fall 2007

Due Wednesday, November 28

1. 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

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

a) Give a simple estimator of $Var(\hat{f}_h(z))$ based just on the fact that $\hat{f}_h(x)$ is a sample average of a random variable. How could you use this to form a confidence interval for $f_0(z)$?

b) Let $\hat{m}_h(z) = \sum_{i=1}^n y_i K_h(z - z_i)/n$. How could you estimate the joint covariance matrix of $\hat{m}_h(z)$ and $\hat{f}_h(z)$ based on them both being sample averages?

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

2. Consider panel data $(y_{it}, x_{it}), (t = 1, 2; i = 1, ..., n)$, where the data are independent across i. Suppose that the data follow a nonparametric model of the form

$$y_{it} = g_0(x_{it}) + \alpha_i + \eta_{it}, E[\eta_{it}|x_{i1}, x_{i2}] = 0.$$

a) If $E[\alpha_i|x_{i1}, x_{i2}] = 0$, how could you estimate $g_0(x_{it})$ from a nonparametric regression of y_{it} on x_{it} for all i and t?

b) If $E[\alpha_i|x_{i1}, x_{i2}] = h(x_{i1}, x_{i2}) \neq 0$, what would $E[y_{i1}|x_{i1}, x_{i2}] - E[y_{i2}|x_{i1}, x_{i2}]$ be? How could you use this formula and nonparametric estimators of $E[y_{i1}|x_{i1}, x_{i2}]$ and $E[y_{i2}|x_{i1}, x_{i2}]$ to estimate $g_0(x)$ up to an additive constant ?

c) Consider a series estimator obtained by regressing $y_{i2} - y_{i1}$ on approximating functions of the form $p_1(x_{i2}) - p_1(x_{i1})$, $p_2(x_{i2}) - p_2(x_{i1})$, ..., $p_K(x_{i2}) - p_K(x_{i1})$. How could you use this series estimator to construct an estimator of $g_0(x)$ up to an additive constant?

d) Which of the estimators from b) or c) do you prefer, and why?

3. Consider the partially linear model $E[y|x,w] = x\beta_0 + h_0(w)$, where x is scalar.

a) Show that if $Var(x|w) = 0$ for all w then there is a $\beta \neq \beta_0$ and a function $h(w)$ that is not equal to $h_0(w)$ such that $E[y|x,w] = x\beta + h(w)$. What does this imply about the ability to consistently estimate β_0?

b) Consider a series estimator of β obtained by regressing y on x and a $K \times 1$ vector of approximating functions $p^K(w)$. Show that this can be written as an OLS regression of $y - \hat{E}[y|w]$ on $x - \hat{E}[x|w]$, where $\hat{E}[y|w]$ and $\hat{E}[x|w]$ are series estimators obtained from regressing y and x respectively on $p^K(w)$.

c) Assume that $E[y_i^2]$ and $E[x_i^2]$ exist (implying existence of $E[E[y|w]^2]$ and $E[E[x|w]^2]$). Also assume that $p^K(w)$ has the property that for any function $g(w)$ with $E[g(w)^2]$ existing there is γ_K with $\lim_{K\longrightarrow\infty} E[\{g(w) - p^K(w)'\gamma_K\}^2] = 0$. Give conditions on the growth rate of K as a function of n such that

$$\sum_{i=1}^{n}\{E[y_i|w_i] - \hat{E}[y_i|w_i]\}^2/n \xrightarrow{p} 0, \sum_{i=1}^{n}\{E[y_i|w_i] - \hat{E}[y_i|w_i]\}^2/n \xrightarrow{p} 0.$$

d) Give conditions for consistency of the $\hat{\beta}$ from part b).

4. Let D be a binary variable and consider the nonparametric regression model

$$E[Y|X,D] = \alpha(X) + \beta(X)D.$$

a) For a given value of δ, what function $h(x,\delta)$ minimizes $E[\{E[Y|X,D] - h(X) - D\delta\}^2]$?

b) Find the value of δ that minimizes $E[\{E[Y|X,D] - h(X) - D\delta\}^2]$ over all δ and over all functions $h(x)$. Interpret this as a weighted average of $\beta(X)$.

c) Show that δ also minimizes $E[\{Y - h(X) - D\delta\}^2]$ over all δ and functions $h(x)$.

d) Suppose that $E[Y|X,D] = \alpha(X) + \beta(X)D$ continues to hold. What would be the probability limit of a series estimator of the coefficient of D from regressing Y_i on D_i and a vector of approximating functions $p^K(X_i)$ if K is allowed to grow with n in the way described in question 1) (and the other conditions from question 1) are satisfied)?

5. Using the gasoline demand data from Hausman and Newey (1995), consider a locally linear nonparametric regression estimator of the conditional expectation of the $\ln(q)$ given $\ln(p)$ and $\ln(y)$.

a. Using 1983 data and Silverman's rule of thumb for the bandwidth, plot the estimate of $E[\ln(q)|\ln(p), \ln(y)]$ as a function of $\ln(p)$ for two different values of $\ln(y)$. Does it appear that $E[\ln(q)|\ln(p), \ln(y)]$ is additively separable in $\ln(p)$ and $\ln(y)$?

b. Plot the estimate as a function of $\ln(p)$ for fixed $\ln(y)$ for three different bandwidths, Silverman's rule, one that is 50 percent larger, and one that is 50 percent smaller. Which do you prefer and why?

c. Using the 1985 data repeat a. and b. Do the results appear to depend on which data set is used?