Problem Set 1

Eco 523 - Nonparametric Econometrics

Carolina Caetano

The data sets used in this problem set can be found in the website: http://www.wadsworth.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9780324581621&discipline_number=413&token=. Click on the "Data Sets" link on the left.

1 Theoretical Part

Question 1:

(i) Write down the least squares optimization problem that yields the Nadaraya-Watson estimator as the optimizing estimator.

(ii) In a parametric context (that is, assuming that we know that $g(x) = \alpha_0 + x^T\alpha_1$ is linear in x), what is the expected difference between $\hat{g}(x)$ calculated with the Nadaraya-Watson estimator and $\hat{g}(x)$ calculated with the OLS estimator?

Question 2: Use the fact that $Var(y|x) = E(y^2|x) - E(y|x)^2$ to suggest a non-parametric estimator of $Var(y|x)$.

Question 3:

(i) Show that if one uses $h_s = \infty$ for all $s = 1, \ldots, q$, then the local linear estimator $\hat{g}(x)$ is identical to the least squares estimator $\hat{\alpha}_0 + x^T\hat{\alpha}_1$, where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are the least squares estimators of α_0 and α_1 based on $Y_i = \alpha_0 + X_i^T\alpha_1 + error$. (Note that we do not assume that the true regression function $g(x)$ is linear in x for this problem.)

(ii) Show when $g(x) = \alpha_0 + x^T\alpha_1$ is linear in x, then the local linear estimator is unbiased (note that you should not assume $h_s = \infty$ in this problem.)

Question 4:

(i) State the assumptions that guarantee the convergence in distribution of the binary local linear estimator $\hat{g}(x, z)$ to the true value $g(x, z)$ in the model

$$Y = g(X, Z) + u,$$

where $E(u|X, Z) = 0$, and X and Z are scalar random variables.

(ii) Suppose that in the previous item you used a kernel

$$k_h(X - x, Z - z) = k_X\left(\frac{X - x}{h_X}\right) \cdot k_Z\left(\frac{Z - z}{h_Z}\right)$$

where k_Z is an epanechnikov kernel, and k_X is an asymmetric kernel, which puts considerably more weight on the negative values than on the positive. Suppose that g is increasing in X for all the values of Z. How would you expect the bias of this estimator to be, compared to the estimator based on a symmetric kernel?

(iii) If you believe that at x, $Y = g_1(X) + g_2(Z) + u$. How would you incorporate this information into a cubic local polynomial estimator? Would you expect efficiency gains from incorporating the additivity information into the bivariate cubic local polynomial estimator? Why?

2 Empirical Part

Question 5: (This question allows you to compare the comparative performance of various kernel-based methods amongst themselves and with the OLS.)

The data set BWGHT.RAW contains data on births to women in the united states. Two variables of interest are the dependent variable, infant birth weight in ounces ($bwght$), and an explanatory variable, average number of cigarettes the mother smoked per day during pregnancy ($cigs$).

(i) Regress $bwght$ on $cigs$ using the OLS method and find the predicted value of $g(cigs)$ for $cigs = 0, 1, 2, \ldots, 20$.

(ii) Regress $bwght$ on $cigs$ for $cigs = 0, 1, 2, \ldots, 20$ using the Nadaraya-Watson estimator with the epanechnikov kernel and $h = 5$.

(iii) Regress $bwght$ on $cigs$ for $cigs = 0, 1, 2, \ldots, 20$ using the local linear estimator, using the epanechnikov kernel and $h = 5$.

(iv) Regress *bwght* on *cigs* for $cigs = 0, 1, 2, \ldots, 20$ using the local polynomial estimator, using the epanechnikov kernel, $h = 5$ and polynomial degree equal to 3.

(v) Plot the average birth weight per point for $cigs = 0, 1, 2, \ldots, 20$. In the same plot, incorporate the predicted *bwght* calculated in (i), (ii), (iii) and (iv) with connecting lines. How do the predictions compare to each other? How does the prediction at and near $cigs = 0$ compare across the different methods?

(v) Regress *bwght* on *cigs* for $cigs = 0.01$ with the local polynomial estimator using only data such that $cigs > 0$, the epanechnikov kernel, $h = 5$ and polynomial degree equal to 3. Compare your finding to the average birth weight at $cigs = 0$. What may you conclude about the continuity of g?

Question 6: (This question asks you to calculate a bandwidth by cross-validation and Silverman's rule-of-thumb method, as well as evidencing the smoothing nature of the kernel method)

Use the data in SLEEP75.RAW from Biddle and Hamermesh (1990) to study whether there is a tradeoff between the time spent sleeping per week and the time spent in paid work. Estimate the model

$$totwrk = g(sleep) + u$$

where *sleep* is minutes spent sleeping at night per week and *totwrk* is total minutes worked during the week.

(i) Regress *totwrk* on *sleep* at $sleep = 3360$ using the Nadaraya-Watson estimator with the epanechnikov kernel and bandwidth chosen with Silverman's rule-of-thumb method.

(ii) Regress *totwrk* on *sleep* at $sleep = 3360$ using the Nadaraya-Watson estimator with the epanechnikov kernel and bandwidth chosen with the cross-validation method.

(iii) Regress *totwrk* on *sleep* at $sleep = 3360.1, 3360.2, 3360.3, \ldots, 3360.9, 3361$ using the Nadaraya-Watson estimator with the epanechnikov kernel and the bandwidth derived in item (ii).

(iv) Plot the average *totwrk* per point for $sleep = 3356, 3357, \ldots, 3364, 3365$, and then, in the same plot, incorporate the predicted value of g for $sleep = 3360$, 3360.1, 3360.2, $3360.3, \ldots, 3360.9$, 3361. Observe that even though there is no data at the points $3360.1, \ldots$ $\ldots, 3360.9$, it is still possible to estimate the value of g at these points, and it is not constant. Why do you suppose this happens?

Question 7: (This question focuses on multivariate regression, estimation of partial derivarives and the bootstrap)

Using the same data set as in the last question, consider the model

$$totwrk = g(sleep, educ, age) + u$$

where *educ* and *age* are education and age measured in years respectively.

(i) Regress *totwrk* on *sleep*, *educ* and *age* at $(sleep, educ, age) = (3360, 13, 25)$, $(3360, 13, 26)$, $(3360, 13, 27)$, $(3360, 13, 28)$, $(3360, 13, 29)$ and $(3360, 13, 30)$ using the local linear estimator with the epanechnikov kernel with bandwidths $h_{sleep} = 11$, $h_{educ} = 4$, and $h_{age} = 5$.

(ii) Estimate the variance of $\hat{g}(3360, 13, 30)$ using the bootstrap method with 100 repetitions.

(iii) Estimate the mean derivative

$$E\left(\frac{\partial}{\partial sleep} g(3360, 13, age) \;\middle|\; 25 \leqslant age \leqslant 30\right)$$

Interpret the value you found in economic terms.