Problem Set 2

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: Suppose that you would like to approximate logarithmic functions with an orthogonal basis.

i Begin with the Taylor expansion of $\log x$ and orthonormalize (by the Gram-Schmidt method) the first 5 elements of the basis.

ii What are the coefficients of a series regression of Y_i onto X_i using the first five elements of the orthonormal basis?

Question 2:

i State the assumptions that guarantee the convergence in distribution of the tensor spline series estimator in the model

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

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

ii Suggest an estimator of the quantity

$$E(g(X, Z)|X > 0).$$

iii Suppose that $g(X, Z) = g_1(X) \cdot g_2(Z)$. Are g_1 and g_2 identifiable? If not, can you propose restrictions in order to guarantee their identifiability?

iv Propose an estimator of g_1 and g_2 using a spline series estimator. Can you use a regular spline basis, or do you need to make any modification?

v How are the assumptions required for the convergence in distribution of this estimator different from the assumptions required in item (i)?

vi If you double the frequency of the knot structure in the spline basis, how do you expect your estimates to change and why?

vii Compare the likely finite sample bias and variance of the estimators in items (i) and (iv) under the assumptions of item (i).

Question 3: How would you estimate the quantity:

$$\theta = E\left[\frac{\partial}{\partial x} g(x, Z)\right]$$

using series estimators? What kind of condition would you need to impose on g in order to guarantee the convergence of your estimator?

2 Empirical Part

Question 4: (This question focuses on multivariate power basis regression)

Use the data in SLEEP75.RAW from Biddle and Hamermesh (1990). 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* using a power basis of degree 1, 2 and 3.

ii Plot the variance of $\hat{g}(x, educ^m, age^m)$ (for the power basis of degree 2), where $educ^m$ is the median education, and age^m is the median age.

iii Test the hypothesis:

$$H_0 : g(sleep^m, educ^m, age^m) = 3360.$$

Question 5: (This exercise is about spline regression)

Use the same data set from the last question 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* using a B-spline basis with degree 3, smoothness 2, and knot structure such that there are about 100 observations per subinterval. Report the knot structure and plot the expected *totwrk* per *sleep*, as well as the fitted curve.

(ii) Supposing that you will use uniform splines of degree 3 and smoothness 2, perform the three methods for the choice of the optimal K. Report the optimal K under each method.

(iii) Choose the average K among the three found on the last question, and regress *totwrk* on *sleep* at using a uniform B-spline basis of degree 3 and smoothness 2 using the average K. Add this regression to the plot on (i).

(iv) Use the same knot vector as in (iii) and regress *totwrk* on *sleep* at using a uniform B-spline basis of degree 3 and smoothness 0. Add this regression to the plot on (iii).

Question 6: (This question is about wavelet regression.)

Use the same data set as in the last question.

(i) Program a Haar basis and plot it with different colors for resolution levels 1 to 4.

(ii) Scale the data so that it is all inside of a unit interval. To do that, you will eliminate the outliers (*sleep*), determine the range after this elimination and divide each *sleep* value by this range. Report your procedure.

(iii) Generate a discontinuity. Determine the median value in the remaining range. To all the observations for which *sleep* is above the median, add the equivalent to 10 hours of work per day. Plot the expected *totwrk* per *sleep*.

(iv) Regress *totwrk* on *sleep* using the Haar basis with resolution level 5. Plot the expected *totwrk* per *sleep*, as well as the fitted curve.

(v) Regress *totwrk* on *sleep* using the Haar basis with increasing resolution levels from 1 to 8. Plot the expected *totwrk* per *sleep*, as well as the fitted curve for each resolution level. Plot each resolution level in one separate plot and show them all.