Applied Nonparametric Econometrics

Empirical Problem Set #1 (Density Estimation)

1. Consider data on the waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We have measurements for the time interval (in minutes) between starts of successive eruptions. The data were collected continuously from August 1st until August 15th, 1985 and resulted in $n = 299$ observations (see Azzalini and Bowman, 1990). Using this data

   (a) Plot a histogram with 15 bins and a histogram with 50 bins. What can be said about the difference in binwidth? What shape does the density seem to possess?

   (b) Estimate a kernel density with a Epanechnikov kernel and Silverman rule-of-thumb (2.345) bandwidth. What shape does the density seem to possess?

   (c) Calculate the LSCV bandwidth for the density and plot the density with this bandwidth. How does the figure change from part (b)?

   (d) Calculate the LCV bandwidth for the density and plot the density with this bandwidth. How does this bandwidth compare to part (b)?

2. Attention in the growth empirics literature has focused on the fact that the distribution of output, as measured by some form of GDP, has become increasingly bimodal as time has passed. Henderson, Parmeter and Russell (2008) tested this phenomena using data from the Penn World Table version 6.2 (available on the JAE data archive website). You will be using it to analyze the robustness of their results to a variety of features. Use only the 1970 and 2000 RGDPCH (rgdpch70 and rgdpch00 in the file) series from their paper. For your plots please print both the 1970 and 2000 plots on the same graph but make sure that you have no more than two plots per graph. Also, prior to estimating any densities convert the data my dividing by the mean output in each year. That is instead of plotting x plot $x/\overline{x}$.

   (a) Using the Silverman rule-of-thumb (1.06) please plot the distributions of output using a Gaussian kernel. Is the bimodal feature apparent in either year?

   (b) Using Silverman's appropriate rule-of-thumb bandwidth plot out these two distributions using the Epanechnikov kernel. Is the bimodal feature still visually apparent?

   (c) Calculate the LSCV bandwidths for the 2000 density for a Gaussian kernel. Is the bimodal feature still apparent. List the value of the bandwidth chosen via LSCV.

   (d) Instead of using the leave-one-out estimator, show that when you fail to use the leave-one-out estimator that the bandwidth tends towards zero (essentially this is asking you to list the bandwidth chosen via LSCV when you do not use the leave-one-out estimator).

Applied Nonparametric Econometrics

Empirical Problem Set #2 (Inference about the Density)

1. Consider data on the waiting time between eruptions for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. We have measurements for the time interval (in minutes) between starts of successive eruptions. The data were collected continuously from August 1st until August 15th, 1985 and resulted in $n = 299$ observations (see Azzalini and Bowman, 1990). Using this data

   (a) Using the Silverman rule-of-thumb (1.06) please plot the distributions of output using a Gaussian kernel.

   (b) Calculate the LSCV bandwidth for the density and plot the density with this bandwidth (using a Gaussian kernel).

   (c) Using the Fan (1994) test, test the null that the distribution is Gaussian (using a Gaussian kernel, the Silverman rule-of-thumb bandwidth and 399 bootstraps).

   (d) Using the test of Ahmad and Li (1997a), test the null that the distribution is symmetric (using a Gaussian kernel, the Silverman rule-of-thumb bandwidth and 399 bootstraps).

   (e) Repeat parts (c) and (d) using the LSCV bandwidth from part (b). What happens to the p-values? What happens to the p-values?

2. Attention in the growth empirics literature has focused on the fact that the distribution of output, as measured by some form of GDP, has become increasingly bimodal as time has passed. Henderson, Parmeter and Russell (2008) tested this phenomena using data from the Penn World Table version 6.2 (available on the JAE data archive website). You will be using it to analyze the robustness of their results to a variety of features. Use only the 1970 and 2000 RGDPCH (rgdpch70 and rgdpch00 in the file) series from their paper. For your plots please print both the 1970 and 2000 plots on the same graph but make sure that you have no more than two plots per graph. Also, prior to estimating any densities convert the data my dividing by the mean output in each year. That is instead of plotting x plot $x/\overline{x}$.

   (a) Using the Li (1996) test, test that the 1970 and 2000 distributions are equal (using a Gaussian kernel, the Silverman rule-of-thumb bandwidth and 399 bootstraps).

   (b) Using the Fan (1994) test, test the null that the 2000 distribution is Gaussian (using a Gaussian kernel, the Silverman rule-of-thumb bandwidth and 399 bootstraps).

   (c) Using the test of Ahmad and Li (1997a), test the null that the 2000 distribution of real GDP per capita (rgdpch00) is independent of real GDP per worker (rgdpwok00) (using a Gaussian kernel, the Silverman rule-of-thumb bandwidth and 399 bootstraps).

Applied Nonparametric Econometrics

Empirical Problem Set #3 (Regression Estimation)

1. Heckman and Polachek (1974) suggest a quadratic parametric relationship between earnings and age

$$y_i = \alpha + \beta z_i + \gamma x_i + \delta x_i^2 + \varepsilon_i$$

where y_i is the logarithm of earnings, z_i is education and x_i is age. Mincer (1974) finds that earnings increase with age through much of the working life but the rate of increase diminishes with age. Pagan and Ullah (1999) present a local-constant kernel estimate of an age earnings profile based on Canadian data (cps71 – available in the np package in R) for $n = 205$ males having common education (high school)

$$y_i = m(\overline{z}, x_i) + \varepsilon_i.$$

Note: given that all of the males have the same value for z (education), you only need to run a regression of log earnings on age (and the square of age in the parametric specification).

   (a) Compute and plot the parametric quadratic as well as local-constant and local-linear estimates using a standard normal kernel with

      1. Rule of thumb bandwidth $h = 1.06\sigma_x n^{-1/(4+q)}$
      2. Bandwidth calculated using least-squares cross-validation

   (b) Is the dip present in the resulting nonparametric estimates?

   (c) Plot the nonparametric estimates along with their 95% confidence bounds (use a wild bootstrap – 399 bootstraps). Without conducting a formal test, does the dip appear to be significant?

   (d) Which nonparametric estimator appears to provide the most "appropriate" fit to this data?

2. Anglin and Gencay (1996) look at the relationship between housing prices and lot size in Windsor, Canada. We consider the overly simplistic case of regressing the log price of the home ($\ln(price)$) on the log of the square footage of the property ($\ln(lot)$) as

$$\ln(price_i) = m[\ln(lot_i)] + \varepsilon_i$$

where we have $n = 546$ homes (note: you must take the natural logarithm yourself).

   (a) Compute and plot the local-constant estimate using an Epanechnikov kernel with least-squares cross-validation bandwidth.

   (b) Plot the nonparametric estimates along with their 95% confidence bounds (use a wild bootstrap – 399 bootstraps).

(c) Consider the binary variable for a full-finished basement ($ffin$), run separate LCLS regressions for the cases where $ffin = 1$ and $ffin = 0$ using an Epanechnikov kernel and the Silverman rule-of-thumb bandwidth (hint: the samples sizes should be 191 and 355, respectively). Plot these estimates on top of one another on a single graph.

(d) What can be said about the relationship between the curves for low levels of lot size? What can be said about the relationship at high levels of lot size? Are these results intuitive?

Applied Nonparametric Econometrics

Empirical Problem Set #4 (Testing in Regression)

1. Heckman and Polachek (1974) suggest a quadratic parametric relationship between earnings and age

$$y_i = \alpha + \beta z_i + \gamma x_i + \delta x_i^2 + \varepsilon_i$$

where y_i is the logarithm of earnings, z_i is education and x_i is age. Mincer (1974) finds that earnings increase with age through much of the working life but the rate of increase diminishes with age. Pagan and Ullah (1999) present a local-constant kernel estimate of an age earnings profile based on Canadian data (cps71 – available in the np package in R) for $n = 205$ males having common education (high school)

$$y_i = m(\overline{z}, x_i) + \varepsilon_i.$$

Note: given that all of the males have the same value for z (education), you only need to run a regression of log earnings on age (and the square of age in the parametric specification).

   (a) Using the Li and Wang (1999) test, test that the quadratic parametric specificiation is appropriate (use a wild bootstrap – 399 bootstraps). Use the
      1. Rule of thumb bandwidth $h = 1.06\sigma_x n^{-1/5}$
      2. Cross-validated bandwidth from the local-constant least-squares regresion

   (b) Using the LCLS version of the Ullah (1985) test, test that the quadratic parametric specificiation is appropriate (use a wild bootstrap – 399 bootstraps). Use the
      1. Rule of thumb bandwidth $h = 1.06\sigma_x n^{-1/5}$
      2. Cross-validated bandwidth from the local-constant least-squares regresion

   (c) Using the Zheng (2009) test, test that the errors from the nonparametric model are homoskedastic (note: estimate the residuals via LCLS with a rule-of-thumb bandwidth and use a wild bootstrap – 399 bootstraps). Use the
      1. Rule of thumb bandwidth $h = 1.06\sigma_x n^{-1/5}$
      2. Cross-validated bandwidth from the local-constant least-squares regresion

1. Anglin and Gencay (1996) look at the relationship between housing prices and lot size in Windsor, Canada. We consider the overly simplistic case of regressing the log price of the home ($\ln(sell)$) on the log of the square footage of the property ($\ln(lot)$) as

$$\ln(sell_i) = m[\ln(lot_i)] + \varepsilon_i$$

where we have $n = 546$ homes (note: you must take the natural logarithm yourself).

   (a) Compute and plot a simple linear parametric least-squares estimate for the above relationship.

(b) Compute and plot (on top of the figure in part (a)), the local-constant estimate using an Epanechnikov kernel with Silverman rule-of-thumb bandwidth.

(c) Using the Li and Wang (1999) test, test that the linear parametric specificiation in part (a) is appropriate (use a wild bootstrap – 399 bootstraps). Use the Silverman rule-of-thumb bandwidth.

(d) Using the LCLS version of the Ullah (1985) test, test that the linear parametric specificiation in part (a) is appropriate (use a wild bootstrap – 399 bootstraps). Use the Silverman rule-of-thumb bandwidth.