STAT8178-STAT7178 Assignment 3

Due Date is 2nd of June 2023

Question 1: 10 marks

1. Generate a sample of size $N = 100$ in R from a standard Normal distribution, $X \sim N(0, 1)$. Obtain a sample for $Y = X^2$. (1 mark)
2. Obtain two sample estimates of $P_0 = P(Y \leq 1.2)$ by using the original sample (call this $\hat{P}_0$) and by using the bootstrapping technique with 500 replications (call this $\hat{P}_0^\star$). Provide the description of steps undertaken for computing $\hat{P}_0$ and $\hat{P}_0^\star$. (3 marks)
3. Find the exact probability of $P_0 = P(Y \leq 1.2)$. (2 marks)
4. Comment on the discrepancy between $\hat{P}_0$, $\hat{P}_0^\star$, and P_0. Provide your reasoning. (1 mark)
5. Design a simulation study to show that $P(\hat{P}_0^\star = \hat{P}_0) \to 1$ as sample size increases. (4 marks)

   Hint: For several sample sizes like $n = 100, 250, 500, 1000, 2000, 5000$, compute the approximation of $P(\hat{P}_0^\star = \hat{P}_0)$.

Question 2: 15 marks

1. Find constant c such that $K(x) = c(1 - x^2), \quad -1 < x < 1$, becomes a valid kernel function. (2 marks)
2. Estimate probability density function with a data set of 10; 12; 15; 25 based on the kernel function in (1) with $c = 0.75$ and the standard normal reference (rule of thumb) window width. Find the estimated probability density function evaluated at $x = 15$.
3. Show that your density function estimator in (2) is a valid density function. Find the corresponding expected value and variance. (3 marks)

4. Estimate p = P (A random observation is less than 20), using the obtained kernel density function estimator in (2). First, use hand calculation and then write a code to estimate p. (6 marks)

5. How we can find a 95% Confidence interval for the true probability density function evaluated at $x = 15$? (2 marks)

Question 3: 18 marks

Assume that the random variable X has a below probability density function with $\alpha = 0.4$.

$g_\alpha(x) = \alpha g_1(x) + (1 - \alpha) g_2(x)$

Let $g_1(x)$ be a Chi-Squared density function with three degrees of freedom and $g_2(x)$ be the density function of Beta distribution Beta$(5, 2)$.

1. Verify that g_α is a valid probability density function. (2 marks)

2. Generate 500 random observations from $g_{0.4}(x)$. (1 mark)

3. Use the sample generated in part (2) to plot the true density function, Histogram, and kernel density function estimators. You should use the standard normal reference (rule of thumb) to find an appropriate bandwidth with Epanechnikov kernel function. (3 marks)

4. Plot the estimates of MSE based on two density function estimators in (3), using Monte Carlo simulation with 500 replications. (4 marks)

5. Calculate the estimates of $MISE$ based on two estimators in (3) using Monte Carlo simulation with 1000 replications. (4 marks)

6. Compare the performance of the two competitors. Your answers must be based on the estimates of MSE and $MISE$. (4 marks)

Question 4: 17 marks

The objects *floorspace* and *saleprice* in *data.csv* describe the floor space, in square meters, and the inflation-adjusted sale price of a random sample of 2930 homes in Iowa sold between 2006 and 2010.

In this question, you will use these data to predict the expectation, median, and 5th and 95th percentiles of the sale price of a house with a given floor space. (This is somewhat simplified, for your convenience. In practice we would want to use several other predictors as well.)

1. Use Nadaraya-Watson to estimate the expected sale price as a function of *floorspace*. Produce a scatterplot of the data with the estimated curve overlaid. You can use trial-and-error and visual inspection, or any other method, to choose the bandwidth h, but you should also plot curves estimated using the bandwidths $5h$ and $h/5$, to demonstrate that h is a good choice. (4 marks)

2. Repeat this for the local linear kernel estimator, again plotting curves for bandwidths h, $5h$ and $h/5$. (4 marks)

3. Use nonparametric (locally linear) quantile regression to estimate the 5th, 50th and 95th percentiles of the conditional distribution of sale price as functions of *floorspace*. Again, choose a bandwidth h, and plot the curves for h, $5h$ and $h/5$. (4 marks)

   (This will give nine curves in all, so you'll probably want to full-screen the plot window before copying it into your assignment submission, to capture the detail.)

4. Consider a house for sale with $200\,\mathrm{m}^2$ of *floorspace*. Use the results of previous parts to find two estimates of the corresponding expected sale price, and estimates of the 5th, 50th and 95th percentiles of its sale price distribution. How do you interpret your findings? (5 marks)

   Hint: You should log-transform the data, that is, use the logarithm of *floorspace* to predict the logarithm of the sale price. You just need to remember to reverse the transformation when interpreting the results.