Math 575 HW 3

Washington University in St. Louis, University College

11/21/2021

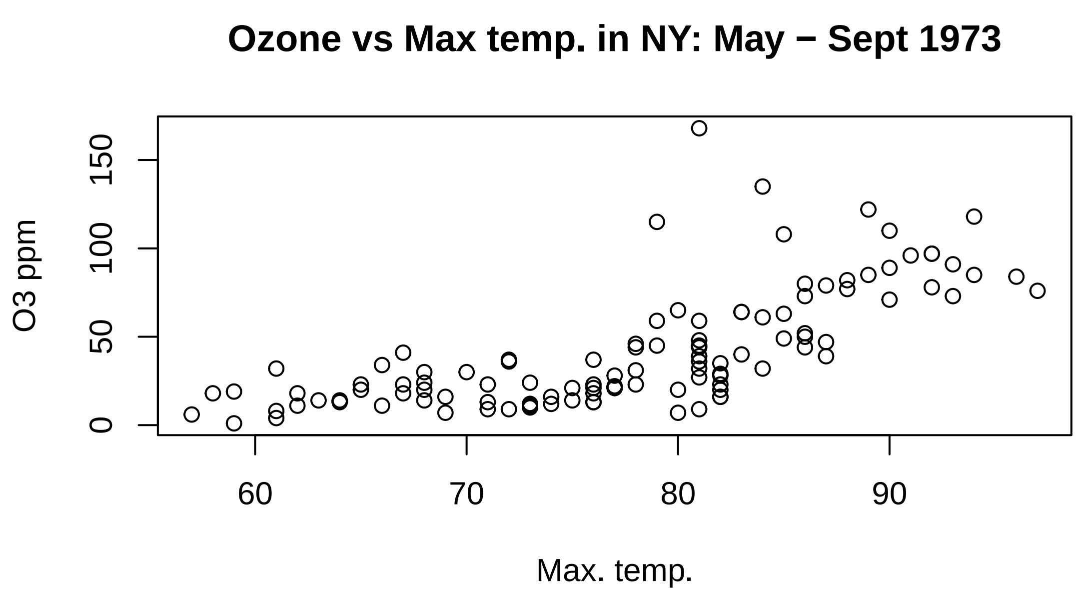
Instruction:

Please type your answers clearly and show your work neatly. You are encouraged to use the Rmarkdown version of this assignment as a template to submit your work. Unless stated otherwise, all programming references in the assignment will be in R. For this assignment, problems roughly covers content from lecture 7-9. Please note, you are free to use the code created in class. For example, the `boostrap_example.Rmd`, `lecture7_code.R` and `lecture9_code.R` will be useful in this assignment.

Problem 1

The R dataset `Airquality` is a set of daily air quality measurements in New York, collected from May to September in 1973.Below is a code snippet showing the relationship between the variables `Ozone` and `Temp`. You can read more about them by using the R command `? datasets::airquality`

```
# Removing some NA's in the data
aq <- na.omit(datasets::airquality)
plot(aq$Ozone~aq$Temp,
     main = 'Ozone vs Max temp. in NY: May - Sept 1973',
     xlab = 'Max. temp.', ylab = 'O3 ppm')
```



In this problem, we are interested in estimating the effect of temperature on ozone measurement. Consider the model

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $y =$ mean ozone level,and $x =$ daily max. temp. Let β_1 be our quantity of interest.

a. Implement the model above using the R function `lm()`. Report the estimated effect from temperature, as well as its standard error. (Note: you can get that by using the `summary()` command)

b. Implement a bootstrap procedure by sampling from the model residuals from a) repeatedly with replacement using 10,000 iterations. Let's call them $\hat{\beta}_{boostrap}$. Plot your results as histogram.

c. The95% confidence interval of $\hat{\beta}$ can be calculated as $[\hat{\beta} - 1.96 * se(\hat{\beta}), \hat{\beta} + 1.96 * se(\hat{\beta})]$, where $\hat{\beta}$ and $se(\hat{\beta})$ are the estimates we got from a). Compare this value with the 95% confidence interval from $\hat{\beta}_{boostrap}$. Comment.

Problem 2

Consider the same dataset above, let X_i represent on the variable `Ozone`, where $i = 1, 2, ...n$. We can use non-parametric method to fit a density function $\hat{f}(x)$ for Ozone level in general, with

$$\hat{f}(x) = \frac{1}{nh} \sum_{i}^{n} K(\frac{X_i - x}{h}),$$

where $K(\cdot)$ is a kernel function that is non-negative, symmetric and integrates to 1.

a. Plot the histogram of `Ozone`.

b. Find the optimal bandwidth h for a Gaussian kernel by minimizing the unbiased cross-validation (UCV) criterion. Using this optimal h, generate a density function using the R command `density()`, and plot it over the histogram created from a).

c. Compare the density fitted using a Gaussian kernel with one generated by a rectangular kernel (provided as an option in `density()`, use R command `?density` to find out more). Comment on your findings.