MA308: Statistical Calculation and Software

Assignment 3 (Nov 11, 2021 - Dec 07, 2021)

3.1 [28 points] For the *"cars"* dataset in *R*, first clean up the data so that it doesn't contain any NA values. Next in order to understand how cars' speed effect the stopping distances,

- **(a)** [7 points] First obtain a scatter plot for *dist* against *speed*, then obtain the Nadaraya-Watson kernel estimator with the choice of two different kernels and the bandwidth of 1 by implementing Nadaraya-Watson Kernel Regression analysis.
- **(b)** [7 points] Fit the local polynomial regression model with the degree of the polynomial up to 2 and Gausssian kernel via the LOESS technique.
- **(c)** [7 points] Compare the fitted error of the Nadaraya-Watson Kernel Regression model and the local polynomial model, which one fits better?
- **(d)** [7 points] Show a scatter plot of *dist* versus *speed*, add on the fitted curve in (a) and (b) with different colors, use legends for illustration.

3.2 [10 points] The *"galaxies"* data set from *MASS* package the velocities of 82 galaxies from six well-separated conic sections of space (Postman et al., 1986, Roeder, 1990). The data are intended to shed light on whether or not the observable universe contains superclusters of galaxies surrounded by large voids. The evidence for the existence of superclusters would be the multimodality of the distribution of velocities. Construct a histogram of the data and add a variety of kernel estimates of the density function. Estimate the density function of the *"galaxies"* data using histogram smoothing, and uniform, triangle, Epanechnikov and Gaussian kernels. For histogram smoothing, please derive and use the best bandwidth. For uniform, triangle, Epanechnikov and Gaussian kernels, the smoothing bandwidth values are set to $1700, 1800, 1900, 2000$ respectively. What do you conclude about the possible existence of superclusters of galaxies?

3.3 [34 points] For the *"foster"* dataset from *HSAUR3* package, the data arise from a foster feeding experiment with rat mothers and litters of four different genotypes: A, B, I and J. The measurement is the litter weight (in grams) after a trial feeding period. Here the investigators interest lies in uncovering the effect of genotype of mother and litter on litter weight.

(a) [6 points] Summarize the main features of the data by calculating group means and standard deviations.

(b) [7 points] Use *interaction2wt()* function in the *HH* package to produce a plot of both main effects and two-way interactions for any factorial design of any order. Explain whether there exists interaction between *litgen* and *motgen*.

(c) [7 points] Carry out two-way factorial ANOVA analysis with and without interaction terms respectively, explain the corresponding results.

(d) [7 points] What are the assumptions that our data need to satisfy when we implement one-way ANOVA? Now if we use one-way ANOVA to examine the difference of *weight* between different genotypes of the litter, are these assumptions satisfied?

(e) [7 points] Carry out the permutation test version of the two-way factorial ANOVA analysis of *weight∼litgen*motgen* with the *lmPerm* package, compare the result with that in (c).

3.4 [28 points] For the *"Default"* dataset from *ISLR* pacakge, we consider how to predict *default* for any given value of *student*, *balance* and *income*. In particular, we will now compute estimates for the standard errors of the *student*, *balance* and *income* logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the *glm()* function. Do not forget to set a random seed before beginning your analysis.

(a) [7 points] Using the *summary()* and *glm()* functions, determine the estimated standard errors for the coefficients associated with *student*, *income* and *balance* in a multiple logistic regression model that uses both predictors.

(b) [7 points] Write a function, *boot.fn()*, that takes as input the "*Default*" data set as well as an index of the observations, and that outputs the coefficient estimates for *student*, *income* and *balance* in the multiple logistic regression model.

(c) [7 points] Use the *boot()* function together with your *boot.fn()* function to estimate the standard errors of the logistic regression coefficients for *student*, *income* and *balance*.

(d) [7 points] Comment on the estimated standard errors obtained using the *glm()* function and using your bootstrap function.