April 6, 2020

Course “Nonparametric statistics”. Home exam

Duration: 3 hours. The solutions should be sent to 2 email addresses
Polina Bogomolova [pmbogomolova@gmail.com]
Vladimir Panov [vpanov@hse.ru].

If I receive your solutions by email, I send immediately a short reply (”Accepted”), which serves as a confirmation of your submission. If you have any questions during the exam, you can phone me (8-965-2886458) or contact me on skype (panov1984).

The exam consists of practical tasks (1,2) and questions (Q1-Q10). Each practical task (1,2) - 9 points, each question (Q1 - Q10) - 2 points. The maximal score is equal to

$$2 * 9 + 10 * 2 = 38.$$

The total scores will be converted into marks (1-10) in accordance to the rule, which will be announced after the exam (e.g., 36-38 can be converted into 10, etc.).

The solutions to the numerical part should contain the programming code and several pictures (at least 1 picture for each item), which serve as an evidence that the solution is correct. It is allowed to write the answers to the questions by hand and to take pictures of them. Important hint: the answers to the questions are typically short (less than a half of a page).

Please do not forget to copy all solutions into one pdf file.

Part I. Numerical tasks

Consider the dataset ”Prestige” from the package ”carData”. To access this dataset, type
install.packages(”carData”)
library(carData)
data(Prestige)

The observations in this database are occupations (jobs). The first 4 variables are:

- education: average education of occupational incumbents, years, in 1971;
- income: average income of incumbents, dollars, in 1971;
- women: percentage of incumbents who are women.
- prestige: a prestige score for occupation, from a social survey conducted in the mid-1960s.

1. (Tests.)
   (i) Compute the Kendall correlation coefficients between the variable ”prestige” and all other variables. Test the hypothesis of independence using **exact** distribution of the corresponding statistics. If the calculation of the exact distribution is not possible, make the appropriate changes in the database (e.g., delete repeated and/or missing values).

   (ii) Divide all occupations into 2 groups: with percentage of women less than 50 (”male jobs”) and more than 50 (”female jobs”). Using nonparametric test, check the hypothesis that the incomes in these two groups follow the same distribution.

   (iii) For each occupation from the first group, find an occupation in the second group with the closest value of the variable ”education”. Using nonparametric test, check the hypothesis that the incomes in these two groups (”male jobs” and ”female jobs” for persons with similar education) follow the same distribution.

2. (Regression.) The aim of this exercise is to analyse whether the prestige of a job can be better explained by the level of income or by the level of education.

The exact tasks are listed below.

(i) Prepare the data for construction of regression dependences between "prestige" and "income":

- find the outliers of the variable "income", delete the corresponding observations from the data;
- find the average "prestige" between the observations with the same "income".

Further analysis should be provided with these new data.

(ii) Fit the supsmu (Super Smoother) model describing the dependence between "prestige" (as y-variable) and "income" (as x-variable). Construct the estimators under various choices of span parameter (0.05, 0.2, 0.5, cross-validation). Find the best model in the sense that the mean-squared error is minimal.

(iii) For the same variables, fit the kernel regression under various choices of kernels (Gaussian, Epanechnikov), and various methods for bandwidth selection (Akaike criterion, least-squares cross-validation). Find the best model in the sense that the mean-squared error is minimal.

(iv) For the same variables, fit the projection estimates to the basis of Legendre polynomials with the number of basis functions varying from 1 to 3. Find the best model in the sense that the mean-squared error is minimal.

(v) Prepare the data for construction of regression dependences between "prestige" and "education":

- select the "middle-class" of occupations in terms of education: find the occupations with "education" lying between 0.25 and 0.75 quantiles of this variable;
- find the average "prestige" between the observations with the same "education".

Further analysis should be provided with these new data.

(vi) Provide the steps (ii)-(iv) for the variables "prestige" and "education".

(vii) Compare the results obtained on previous steps.

Part II. Questions

Density estimation

Q1 Which kernel yields the minimal AMISE (asymptotic mean integrated squared error) of the kernel density estimator? Formulate the corresponding theorem. What is the kernel efficiency?

Q2 In which sense the histogram is an optimal estimate of the density?

Q3 Why the histogram is an asymptotically unbiased and consistent estimate of the density?

Regression

Q4 What is the difference between cross-validation and generalized cross-validation (in the context of regression problems)? Which theoretical fact lies in the core of the generalized cross-validation approach?

Q5 Describe the scheme for the construction of the projection estimator for regression function.

Q6 What is the effective degrees of freedom? What does the value of this characteristic mean for linear regression?

Statistical tests

Q7 How one can test the hypothesis of independence using large-sample approximation of the distribution of Spearman's rho? Explain the mathematical idea.

Q8 Explain the relation between the Kruskal-Wallis statistics and the ANOVA test.

Q9 Explain why the Kendall tau is equal to -1 if and only if the Pearson correlation coefficient is equal to -1, provided that the data follow the multivariate normal distribution.

Bonus question

Q10 The chi-squared test for contingency tables is based on a theorem, which guarantees the convergence of some statistics to the chi-squared distribution. Which value of the degree of freedom of this limiting distribution should be used for contingency tables ? Explain your answer.