Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
7.6.2023
J Kohonen

Instructions: Answer in English. Write clearly and give reasons for your answers. A number only as an answer does not yield points. The exam has 4 problems, each worth 6 points.

Allowed equipment: writing tools, calculator (symbolic and graphic OK), at most A4-size cheat sheet written on one side.

P1 All of the following claims are false. Explain very briefly (in 1–2 sentences) why. 1 point for each.

(a) If a distribution is symmetric around zero, it is called the normal distribution.

(b) Positive skewness indicates that both tails of the distribution are long.

(c) In linear regression, correlation coefficient is the slope of the regression line.

(d) Confidence level indicates how much confidence you have in the model assumptions.

(e) In hypothesis testing, the null hypothesis is rejected when the p-value is greater than the significance level.

(f) Bonferroni correction is used for correcting outlier points in linear regression.

P2 (a) Draw a scatter plot of two variables that have:

(i) perfect linear dependence **(1p)**

(ii) perfect monotonic dependence but not perfect linear dependence **(1p)**

(b) Is it possible for two variables to have perfect linear dependence but not perfect monotonic dependence? Explain why or why not. **(2p)**

(c) Explain Spearman's rank correlation coefficient. When is it used and how? **(2p)**

P3 In each of the following scenarios you have an i.i.d. (independent and identically distributed) sample $x_1, \ldots, x_n$ from some distribution F. Describe (in 3–5 sentences, and including also the assumptions that your chosen methods make) how you would investigate the following research questions.

(a) Does the sample come from a distribution with median equal 0? **(2p)**

(b) Does the sample come from the standard normal distribution? **(2p)**

(c) Does the sample come from a distribution whose standard deviation is 5? **(2p)**

P4

(a) How does backward selection conduct variable selection? List also two of its drawbacks. **(3p)**

(b) Why is simply picking the model which gives the largest value of R^2 not a good idea with respect to variable selection? **(1p)**

(c) The following plot shows the LASSO coefficient profiles in a regression problem with a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$. (If the colors are not well visible, note that the order of the profiles from top to bottom in the left end of the plot is `x5`, `x4`, `x2`, `x3`, `x1`). The optimal value of $\log(\lambda)$ given by cross-validation is shown as a vertical line.
(i) Which variable does LASSO hold as the most important one? Which is the second most important? **(1p)**
(ii) Write down (approximately) the estimated coefficients of the five predictors in the model corresponding to the optimal value of $\log(\lambda)$, and explain which variables (if any) have been left out of the model by this stage. **(1p)**