Exam-2019-20

Problem 1
Let $\mathbf{X} = (X_1, \ldots, X_d)$ be multivariate normally distributed with unknown expectation vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and known covariance matrix Σ. The d hypotheses are

$$H_1 : \mu_1 \leq 0, \ldots, H_d : \mu_d \leq 0.$$

Based on a sample $(X_{11}, \ldots, X_{1d}), \ldots, (X_{n1}, \ldots, X_{nd})$ of size n from $\mathbf{X}$ we reject hypothesis i, $1 \leq i \leq d$, at level α if

$$\sum_{\ell=1}^{n} X_{\ell i} > \sqrt{n}\, \sigma_i q_{1-\alpha},$$

where σ_i is the standard deviation of X_i and $q_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal.

(i) For $d = 15$ which multiple testing procedure could you use to test the d hypotheses?

(ii) For $d = 500$ which multiple testing procedure would you use to test the d hypotheses?

Please motivate your answers in (i) and (ii) briefly.

Problem 2
Assume that X is exponentially distributed with parameter λ, i.e. the cumulative distribution function of X is given by

$$F_\lambda(x) = 1 - \exp(-\lambda\, x),\ x > 0,\ \text{and zero otherwise.}$$

For testing

$$H : \lambda \leq 1, \quad A : \lambda > 1,$$

we use a sample X of size 1. Our test statistic is then simply $T(X) = X$ and our p-value is simply $\hat{p}(X) = 1 - \exp(-X)$.

(i) Show that $\hat{p}(X)$ is uniformly distributed on $(0, 1)$ if the cumulative distribution function of X is

$$F_1(x) = 1 - \exp(-x),\ x > 0,\ \text{and zero otherwise.}$$

(ii) Find the cumulative distribution function of $\hat{p}(X)$ if the cumulative distribution function of X is F_2.

Problem 3
Assume that Y_i, $1 \leq i \leq n$, is Poisson distributed given the explanatory variables x_{i1}, x_{i2} and x_{i3}, $1 \leq i \leq n$, with parameter $\lambda_i = \exp(\beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3 x_{i3})$, $1 \leq i \leq n$. For a sample $y_1, \ldots, y_n$ and explanatory variables $((x_{11}, x_{12}, x_{13}), (x_{21}, x_{22}, x_{23}), \ldots, (x_{n1}, x_{n2}, x_{n3}))$ give the log-likelihood function and find the first order conditions for β_1, β_2, and β_3.

Problem 4
In Problem 3 we related λ_i, i.e. the expectation of Y_i, to $\beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3 x_{i3}$ using the function

$h : \mathbb{R} \to \mathbb{R}_+$ defined by

$$h(x) := \exp(x).$$

For each of the following alternative choices, argue if it is meaningful to use them instead of the above h. Explain your answer.

(i) $h_1(x) = x^3$, $x \in \mathbb{R}$;

(ii) $h_2(x) = \frac{x^2}{1+|x|}$, $x \in \mathbb{R}$.

Problem 5

Assume our data come from the linear model

$$Y_i = \sum_{j=1}^{d} \beta_j \, X_{ij} + \epsilon_i, \; i = 1, \ldots, n,$$

with ϵ_i, $1 \leq i \leq n$, independent and normally distributed with expectation zero and variance σ^2. Consider the following choices for d as a function of the sample size

(i) $d = n^n$;

(ii) $d = n^{15} \log(n)$;

(iii) $d = \sqrt{n} \, \exp\left(n^{0.8}\right)$.

For which of these choices do we have consistency of the LASSO estimator as $n \to \infty$ (you can assume that the design matrix fulfills the restricted eigenvalue condition and that β has only k non-zero entries for all d and n)? **Please motivate your answers in (i), (ii) and (iii) briefly.**

Problem 6

Consider the following set-up:

- X_i, $1 \leq i \leq 50$, is binomially distributed with success probability $p_i = p$, $1 \leq i \leq 50$;
- The hypotheses H_i, $1 \leq i \leq 50$, are $H_i : p_i = 0.5$;
- From each X_i we have a sample $(X_{i1}, \ldots, X_{i10})$ of size 10, and reject H_i if

$$\left| \frac{\sqrt{10}}{0.5} \, (\hat{p}_i - 0.5) \right| > 1.645,$$

  where $\hat{p}_i = \frac{1}{10} \sum_{j=1}^{10} X_{ij}$.
- If H_i is rejected we construct the following confidence interval for p_i

$$CI_i = \left[\hat{p}_i - 1.645 \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{10}} ; \hat{p}_i + 1.645 \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{10}} \right].$$

Assume $p = 0.6$, put $S = \{ i \in \{1, \ldots, 50\} \, | \, H_i \text{ was rejected } \}$ and let $p_S = (p_{i_1}, \ldots, p_{i_{|S|}})$, $i_1 < \ldots < i_{|S|}$, $i_j \in S$, be the vector of selected parameters (selected here means associated with the hypotheses rejected). Write pseudo code (or code) that could be used to calculate the probability that p_S is contained in $CI_{i_1} \times \ldots \times CI_{i_{|S|}}$.