vrije Universiteit amsterdam

Studentnumber:

Name:

School of Business and Economics

Exam:	Big Data Statistics
Code:	E_EORM_BDS
Examinator:	E.A. Beutner
Co-reader:	Y. Lin
Date:	May 17, 2021
Time:	12:55hrs
Duration:	2 hours
Calculator allowed:	**Yes**
Graphical calculator allowed:	**No**
Scrap paper	**Yes**
Number of questions:	7
Type of questions:	Open
Answer in:	English

Remarks:

Credit score:	You can get in total 100 points. To pass you need 55 points or more.
Grades:	The grades will be made public within 10 working days.
Inspection:	Will be announced on the course Canvas page.
Number of pages:	3 (including front page).

Good luck!

Re-sit-A-L

Problem 1 (10 points)
Assume we have 15 hypotheses $H_1, \ldots, H_{15}$ and for each hypothesis we use a test statistic T_i, $i = 1, \ldots, 15$, to test hypothesis H_i at level $\alpha = 0.025$. Assume that the test statistics are independent. Find the probability that we reject at least one true hypothesis, i.e. calculate the following $\mathbb{P}(\text{reject at least one true hypothesis})$.

Problem 2 (15 points)
A friend reports you the following five **Bonferroni** adjusted p-values for testing $H_1, \ldots, H_5$

$$1.00, 0.00788, 0.00062, 1.00, 0.07025,$$

where 1.00 is the Bonferroni adjusted p-value for H_1, 0.00788 is the Bonferroni adjusted p-value for H_2 and so on. Your friend used the convention to report a Bonferroni adjusted p-value of 1 if the Bonferroni adjusted p-value is equal to 1 or larger than 1. Testing at a significance level of 0.05 your friend rejects hypotheses 2 and 3.
Use your friend's Bonferroni adjusted p-values to calculate **Holm** adjusted p-values and decide which hypotheses are rejected if you use Holm adjusted p-values and the same significance level as your friend.

Problem 3 (10 points)
You are given the following eight unadjusted p-values for testing $H_1, \ldots, H_8$

$$0.07823, 0.00503, 0.000063, 0.17914, 0.12575, 0.004971, 0.000399, 0.000076,$$

where 0.07823 is the p-value for H_1, 0.00503 is the p-value for H_2 and so on. Decide which hypotheses we reject if we use the k-FWER modified Bonferroni procedure with $k = 3$ and level $\alpha = 0.05$.

Problem 4 (20 points)
Assume that the distribution of Y_i, $1 \leq i \leq n$, is given by

$$\mathbb{P}(Y_i = k) = (1 - p_i)^{k-1} p_i, \text{ for } k = 1, 2, \ldots,$$

where given the explanatory variables x_{i1} and x_{i2}, $1 \leq i \leq n$, we have $p_i = \Phi(\beta_1 x_{i1} + \beta_2 x_{i2})$, $1 \leq i \leq n$. Here Φ is the cumulative distribution function of a standard normal. For a sample $y_1, \ldots, y_n$ and explanatory variables $((x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{n1}, x_{n2}))$ give the log-likelihood function and find the first order conditions for β_1 and β_2.
Remark: There is of course no need to solve the first order conditions; giving them is enough.

Problem 5 (7.5+7.5 points)
In class we related the expectation of a random variable Y to a linear function $\sum_{j=1}^{d} \beta_j x_j$ using a link function h. Assume that the distribution of Y is given by

$$\mathbb{P}(Y = -1) = (1 - p) \text{ and } \mathbb{P}(Y = 1) = p$$

where we have for the parameter p that $0 < p < 1$. This implies that the expectation of $\mathbb{E}[Y]$ equals

$$\mathbb{E}[Y] = 2p - 1.$$

For each of the following alternative choices of the link function h, argue if it is meaningful to use them to relate $\mathbb{E}[Y]$ and $\sum_{j=1}^{d} \beta_j\, x_j$ by $\mathbb{E}[Y] = h(\sum_{j=1}^{d} \beta_j\, x_j)$. Explain your answer.

(i) $h_1(x) = \frac{x}{|x|+1}$, $x \in \mathbb{R}$;

(ii) $h_2(x) = \Phi(x)$, $x \in \mathbb{R}$, where Φ is the cumulative distribution function of a standard normal.

Problem 6 (10 + 5 points)
(i) For any vector $\mathbf{b} = (b_1, \ldots, b_d) \in \mathbb{R}^d$ define the set $S(\mathbf{b}) = \{j \,|\, b_j \leq 2,\ j = 1, 2, \ldots, d\}$ which is a subset of $\{1, \ldots, d\}$. Given a sequence $(\mathbf{a}_n)$ of d-dimensional vectors that converges to the vector $\mathbf{a}$ does this imply that $S(\mathbf{a}_n)$ converge to $S(\mathbf{a})$? If the statement is true argue briefly why this convergence holds. If it is false give a counterexample.
(ii) Explain why the question considered in part (i) is of interest when studying the LASSO estimator.

Problem 7 (10 + 5 points)
(i) Assume our data come from the linear model

$$Y_i = \sum_{j=1}^{60} \beta_j\, x_{ij} + \epsilon_i,\ i = 1, \ldots, 10, \tag{1}$$

with ϵ_i, $1 \leq i \leq 10$, independent and normally distributed with expectation zero and variance σ^2. Unfortunately the observations $\mathbf{y} = (y_1, \ldots, y_{10})$, and $\mathbf{x}_1 = (x_{11}, \ldots, x_{10\,1}), \ldots,\ \mathbf{x}_{60} = (x_{1\,60}, \ldots, x_{10\,60})$ were lost. The only thing known is that

$$\frac{2}{10} \max_{1 \leq j \leq 60} | < \mathbf{x}_j, \mathbf{y} > | \ = 4.5.$$

Here as always $< ., . >$ denotes the Euclidean scalar product. Given this information only can you find the solution of the following minimization problem

$$\text{minimize w.r.t. } \boldsymbol{\beta} : \ \frac{1}{10} \sum_{i=1}^{10} \left(y_i - \sum_{j=1}^{60} \beta_j\, x_{ij} \right)^2 + 5 \sum_{j=1}^{60} |\beta_j|, \tag{2}$$

where $y_1, \ldots, y_{10}$ and $x_{11}, \ldots, x_{10\,60}$ are the unknown observations?
Explain your answer briefly and in case you can find the solution give the solution.
(ii) Consider again the model in Equation (1). Assume you have all observations. How would you choose λ in the following optimization problem

$$\text{minimize w.r.t. } \boldsymbol{\beta} : \ \frac{1}{10} \sum_{i=1}^{10} \left(y_i - \sum_{j=1}^{60} \beta_j\, x_{ij} \right)^2 + \lambda \sum_{j=1}^{60} |\beta_j|\,? \tag{3}$$

Explain your choice of λ briefly.