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:	March 23, 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:	4 (including front page).

Good luck!

Exam-0-4

Problem 1 (20 points)
You are given the following five unadjusted p-values for testing $H_1, \ldots, H_5$

$$0.99734, 0.60008, 0.13896, 0.00773, 0.00097,$$

where 0.99734 is the p-value for H_1, 0.60008 is the p-value for H_2 and so on. Calculate Bonferroni, Holm, and Benjamini and Hochberg adjusted p-values. For each of the three methods decide which hypotheses we reject if we use $\alpha = 0.05$.

Problem 2 (10 + 10 points)
Assume that the cumulative distribution function of the random variable X is given by

$$F_\lambda(x) = 1 - \exp\left(-\left(\frac{x}{\lambda}\right)^2\right), \; x > 0, \text{ and zero otherwise,}$$

where $\lambda > 0$. For testing

$$H: \lambda^2 \leq 2, \quad A: \lambda^2 > 2,$$

we use, based on a sample X of size 1, the test statistic $T(X) = X^2$ and we reject H at level α if X exceeds the critical value $c(\alpha)$.

(i) Find the critical value $c(\alpha)$ if we test at $\alpha = 0.05$;

(ii) Calculate the power of the test at $\lambda = 4$.

Problem 3 (10 points)
For $0 < p < 1$ consider the following probability distribution

$$\mathbb{P}(Y = y) = \frac{(1-p)^{y-1}p}{1-(1-p)^{10}}, \text{ for } y = 1, \ldots, 10,$$

which takes only the values $1, 2, \ldots, 10$. In other words the probability mass function g_Y^p of this random variable is

$$g_Y^p(y) = \frac{(1-p)^{y-1}p}{1-(1-p)^{10}}, \text{ for } y = 1, \ldots, 10.$$

In class (Lecture 6) we discussed a particular form for probability mass functions given by

$$f_\theta^Y(y) = \exp\left(\frac{y\theta - b(\theta)}{\psi} - c(\psi, y)\right), \; y \in D,$$

where $\theta \in \Theta$ and ψ are real-valued parameters, D is the support of the distribution of Y, and b and c are real-valued functions. Is it possible to write g_Y^p in this form?

Problem 4 (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

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

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

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

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

(ii) $h_2(x) = \int_0^{|x|} y^2 \, dy$, $x \in \mathbb{R}$.

Problem 5 (7.5+7.5 points)
Assume our data come from the linear model

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

with ϵ_i, $1 \leq i \leq 10$, independent and normally distributed with expectation zero and variance σ^2. Unfortunately the observations $y_1, \ldots, y_{10}$, and $x_{11}, \ldots, x_{10\,60}$ were lost. What is known is that the X_{ij} were independent and each normally distributed with expectation 2 and variance 10. A friend of you tells you that he was additionally given the following two vectors

(i) $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_{60})$ with $\bar{\beta}_j = 1 + j$ for $j = 1, \ldots, 12$ and $\bar{\beta}_j = 0$ otherwise;

(ii) $\breve{\boldsymbol{\beta}} = (\breve{\beta}_1, \ldots, \breve{\beta}_{60})$ with $\breve{\beta}_j = 1 + j$ for $j = 1, \ldots, 5$, $\breve{\beta}_{59} = 2.5$ and $\breve{\beta}_j = 0$ otherwise.

Given this information only decide for both $\bar{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ whether they could potentially be the solution to the following minimization problem

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

where $y_1, \ldots, y_{10}$ and $x_{11}, \ldots, x_{10\,60}$ are the unknown observations.
Explain your answers briefly.

Problem 6 (10 points)
We discussed in class that there can be multiple solutions to the (LASSO) minimization problem

$$\text{minimize w.r.t. } \boldsymbol{\beta}: \; \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j|.$$

Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be two different solutions for this minimization problem. Prove or disprove that $X\hat{\boldsymbol{\beta}}_1 = X\hat{\boldsymbol{\beta}}_2$ where, as usual, X is the design matrix.
Hints: $X\hat{\boldsymbol{\beta}}_1$ and $X\hat{\boldsymbol{\beta}}_2$ are in $\mathbb{R}^n$, the mapping $z \to ||y - z||_2^2$ is strictly ...

Problem 7 (10 points)
For $k = 1, \ldots, 5$ let I_k be a confidence interval for θ_k, $k = 1, \ldots, 5$, with coverage probability $1 - \alpha$. Assuming that the I_k are independent how do we need to choose α such that $I_1 \times \ldots \times I_5$ is a simultaneous confidence interval for $(\theta_1, \ldots, \theta_5)$ at level 95% (, i.e. $\mathbb{P}((\theta_1, \ldots, \theta_5) \in I_1 \times \ldots \times I_5 = 0.95)$)?