Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2018	MATH3051-WE01

Title: Statistical Methods III

Time Allowed:	2 hours 30 minutes	
Additional Material provided:	Tables: Normal, t-distribution, F-distribution, χ^2-distribution; Graph paper.	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio fx-83 GTPLUS or Casio fx-85 GTPLUS.
Visiting Students may use dictionaries: No		

Instructions to Candidates:	Credit will be given for: the best **THREE** answers from Section A and the best **TWO** answers from Section B. Questions in Section B carry **TWICE** as many marks as those in Section A.

Revision:	

ED01/2018

SECTION A

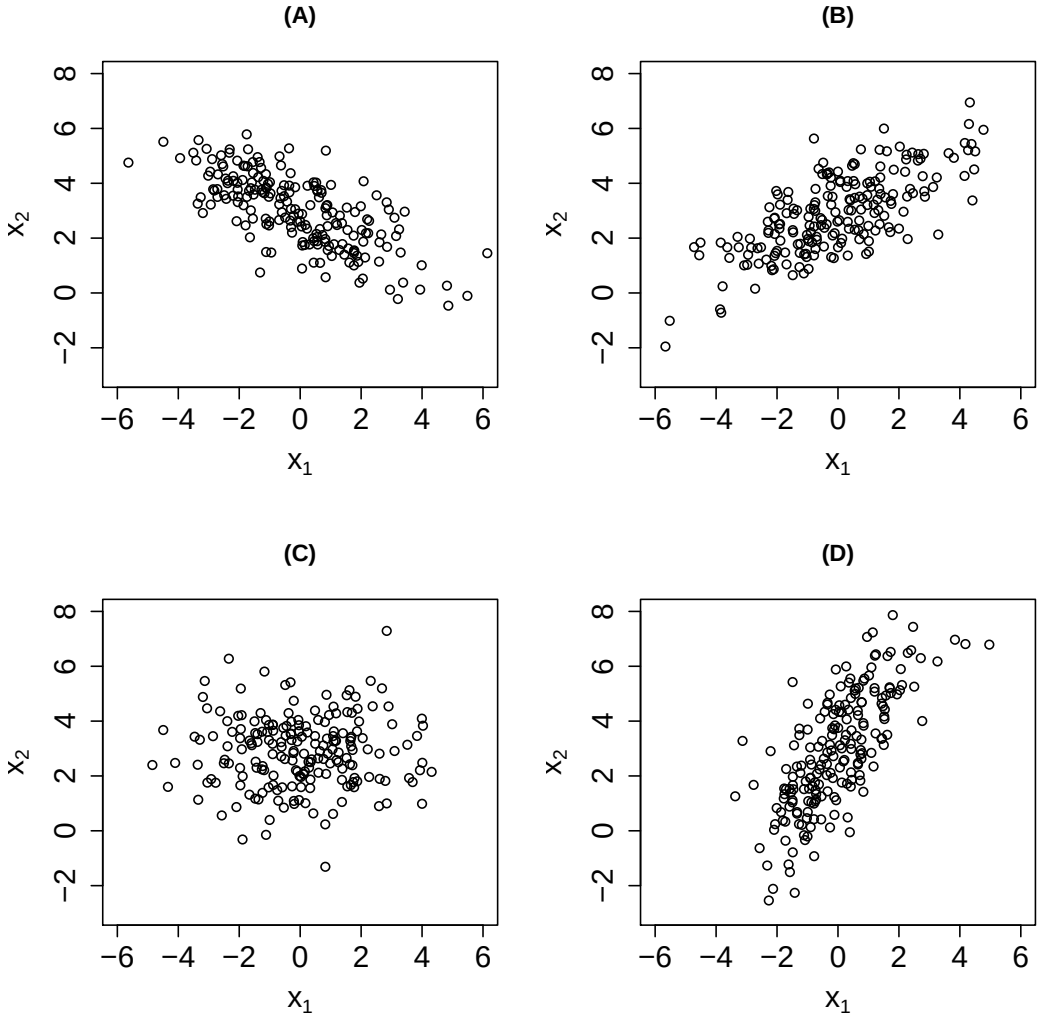
1. Consider the random vector

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix} \right).$$

   (a) Give the probability density function of X. Simplify the expression where possible.

   (b) Find the distributions of X_1, X_2, and $X_1 + X_2$.

   (c) Below you find four data sets (A), (B), (C), and (D), each of size $n = 200$. All of them are generated from a bivariate normal random vector, but only one of them is generated from X. Which one is it? Explain your answer.



2. (a) Let $X = (X_1, \ldots, X_q)^T$ be a q–dimensional random vector with density f. We know that the expectation $\boldsymbol{m}$=$E(X)$ is defined through the q–dimensional integral

$$\boldsymbol{m} = \int \boldsymbol{x} f(\boldsymbol{x})\, d\boldsymbol{x}$$

where $\boldsymbol{x} = (x_1, \ldots, x_q)^T$ and $\boldsymbol{m} = (m_1, \ldots, m_q)^T$.
Show that $m_j = \int x_j f(x_j) dx_j$, where $f(x_j)$ is the marginal density of variable X_j, $j = 1, \ldots, q$.

(b) Consider the bivariate random vector $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with density

$$f(x_1, x_2) = \begin{cases} 2, & 0 < x_2 < x_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$=Var(X) through explicit calculation. Verify that your result for $\boldsymbol{\Sigma}$ is a valid variance matrix.

3. The normal equations for the linear model can be written in our usual notation as

$$\boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{Y}, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{Y} \in \mathbb{R}^{n \times 1}$, $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{p \times 1}$. Consider now the special case of the simple linear regression model, that is

$$y_i = a + b x_i + \epsilon_i, \quad i = 1, \ldots, n.$$

(a) Write down the matrices $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ for this scenario.

(b) Using (1), or otherwise, derive explicit expressions for the least squares regression estimators $\hat{a}$ and $\hat{b}$.

(c) Assume $n \geq p$. We are interested in identifying scenarios under which the matrix $\boldsymbol{X}^T \boldsymbol{X}$ might not be invertible. Therefore, work out the determinant of $\boldsymbol{X}^T \boldsymbol{X}$ for the considered simple linear regression model. Discuss for which choices of x_i, $i = 1, \ldots n$, this determinant would (or could) take the value 0. Give an interpretation of your results.

(d) Assume now $n < p$. Is there any choice of predictors x_i, $i = 1, \ldots n$, for which $\boldsymbol{X}^T \boldsymbol{X}$ is invertible?

4. An experiment was carried out to understand the strength of wool as a function of three factors relating to its production. Specifically, we investigate the relationship between the number of cycles to failure y of a worsted yarn and the factors: the length of test specimen (x_1 : 250, 300, 350 mm), amplitude of loading cycle (x_2 : 8, 9, 10 mm), and load (x_3 : 40, 45, 50 gm). Each of the $3\times3\times3$ factor combinations was used only once.

   In what follows, we will use the logarithm of failure, that is $\log y$, as response variable. Concerning the specification of the predictor terms, there are two possible views on the data:

   (A) This is a designed factorial experiment involving three factors with three levels each; so we can build a model involving an intercept and two dummy variables for each factor level.

   (B) Noting that the possible values for each factor are actually *numeric* and *equidistant*, we can deal with them as if they were continuous variables, and build a model of type

$$E(\log y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

   (a) Write down the dimension of the design matrix for each of scenarios (A) and (B).

   (b) Fitting the models leads to values of R^2 of 0.7692 for model (A) and 0.7291 for model (B). Interpret these values briefly.

   In order to allow better comparability for models involving a different number of parameters, it has been suggested to consider *adjusted* R^2, which is defined by

$$R^2_{adj} = 1 - \frac{n-1}{n-p}(1-R^2)$$

   with n being the sample size and p denoting the number of parameters in the respective model. Compute R^2_{adj} for both models and interpret the result.

   (c) State how R^2_{adj} behaves when $R^2 \longrightarrow 1$, and discuss the relevance of this result with view towards using R^2_{adj} as a model selection tool.

   (d) State the three main principles of experimental design and explain their purpose in the context of this data set. Do you detect any violations of these principles?

SECTION B

5. There is considerable interest in modelling (and predicting) water supply from precipitation. For this study, precipitation measurements (in inches per annum) were taken at six sites in the Owens Valley in Southern California between 1948 and 1990. Water supply is represented by 'stream runoff volume' (hereafter coded `RunoffVol`), measured in acre–feet at a site near Bishop, California. Four of the six sites were situated at lakes (`MammothLake`, `SabrinaLake`, `SouthLake`, `UnnamedLake`), and two at creeks (`BigpineCreek`, `RockCreek`). A linear regression model was fitted with `RunoffVol` as response variable, and the measurements at the six sites (plus an intercept) as predictors.

   The first two columns of a sequential Analysis of Variance table, where the predictors are included into the model in the order as given above, are provided below. For better readability, the SS values have been divided by 10^6. That is, for example, the actual contribution of `SabrinaLake` to the Sum of Squares (SS) is 17,427,400.

	Df	SS/10^6
`MammothLake`	1	1556.6948
`SabrinaLake`	1	17.4274
`SouthLake`	1	661.2574
`UnnamedLake`	1	22549.8347
`BigpineCreek`	1	0.0789
`RockCreek`	1	509.8944
Error	36	2055.8307

   (a) At the 5% level of significance, carry out adequate F–tests which address the following questions:

      i. Do the six precipitation measurements, as a whole, contribute significantly to the variation of stream runoff volume?

      ii. Given the inclusion of `MammothLake`, do the measurements from `SabrinaLake` contribute significantly to the variation of stream runoff volume?

      iii. Given the inclusion of the four measurements at the lakes, do the two measurements at the creeks contribute significantly to the variation of stream runoff volume?

   (b) Model selection criteria can be used to decide between different models involving different configurations (and numbers) of variables. A well–known model selection criterion is Mallows' $C_{\mathcal{I}}$, which is given by

$$C_{\mathcal{I}} = \frac{\mathrm{RSS}_{\mathcal{I}}}{s^2} + 2p_{\mathcal{I}} - n$$

   where s is the residual standard error of the 'full' model, $\mathcal{I}$ is the index set of included variables, $p_{\mathcal{I}}$ denotes the cardinality of this set, and $\mathrm{RSS}_{\mathcal{I}}$ the sum of squares of a model fitted using only the variables corresponding to the index set $\mathcal{I}$.

[Question 5 continues on the next page]

i. Show that Mallows' $C_{\mathcal{I}}$ can be expressed in the form $a_{\mathcal{I}} + b_{\mathcal{I}} F_{\mathcal{D}}$, where $a_{\mathcal{I}}$ and $b_{\mathcal{I}}$ are constants which depend on $\mathcal{I}$ through $p_{\mathcal{I}}$, $\mathcal{D}$ is the index set of variables *not* included in the model, and $F_{\mathcal{D}}$ is the test statistic for testing H_0: 'Given the inclusion of $\mathcal{I}$, the variables in $\mathcal{D}$ do not contribute to the variation in the response'.
Use this connection to the F–test to deduce which values of $C_{\mathcal{I}}$ one can expect for 'good' submodels (that is, submodels those for which the hypothesis H_0 is true).

ii. For the data set considered, compute $C_{\mathcal{I}}$ for the 'empty' model (including only the intercept), the 'full' model (including measurements at all six sites), and the model involving only the measurements taken at the lakes (of course, both latter models also include the intercept). Select a suitable model from these three options.

iii. Discuss briefly strategies to efficiently select a best submodel from *all* possible submodels of the full model.

6. We consider a data set recorded in 1971 in Canada. The data set has $n = 102$ rows, each of which gives an observation that relates to an occupation. Specifically, we have measurements on the the following four variables:

`income`	Average income of incumbents, dollars, in 1971.
`education`	Average education of incumbents, years, in 1971.
`women`	Percentage of incumbents who are women.
`prestige`	Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

For instance, the rows $i = 1, 2$ and 53 of this data set are displayed below.

i		`income`	`education`	`women`	`prestige`
1	`gov.administrators`	12351	13.11	11.16	68.8
2	`general.managers`	25879	12.26	4.02	69.1
53	`newsboys`	918	9.62	7.00	14.8

We are fitting initially a linear model to this data set,

$$\texttt{income} = \beta_1 + \beta_2 \texttt{education} + \beta_3 \texttt{women} + \beta_4 \texttt{prestige} + \epsilon \tag{2}$$

(a) The two top panels of the figure **provided on the next page** contain some diagnostics for this model. Specifically, the top left plot contains the leverage values, h_i, and the top right plot the ('internally') studentised residuals, r_i, which are defined as

$$h_i = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i, \qquad r_i = \frac{\hat{\epsilon}_i}{s\sqrt{1 - h_i}},$$

respectively, where $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$ denotes the design matrix and s the residual standard error.

i. Explain qualitatively for which aspects (of the model and/or the data) h_i and r_i serve as diagnostic tools.

ii. For both diagnostics, provide some statistical arguments which indicate which magnitude of values of h_i and r_i one would typically expect. Based on these, suggest rules of thumb which guide the data analyst when applying h_i and r_i as diagnostic devices.

iii. Now, using the information provided in the two images, as well as your rules of thumb, carry out the diagnostics for the provided data set, and report your conclusions.

(b) If violations of model assumptions are diagnosed, a frequently applied solution is to transform the data adequately. One such transformation is the Box–Cox–transformation,

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0; \\ \log y & \lambda = 0, \end{cases}$$

which attempts to find the parameter, λ, such that $y^{(\lambda)}$ fulfils the linear model assumptions,

$$y^{(\lambda)} | \boldsymbol{x} \sim N(\boldsymbol{x}^T \boldsymbol{\beta}, \sigma^2). \tag{3}$$

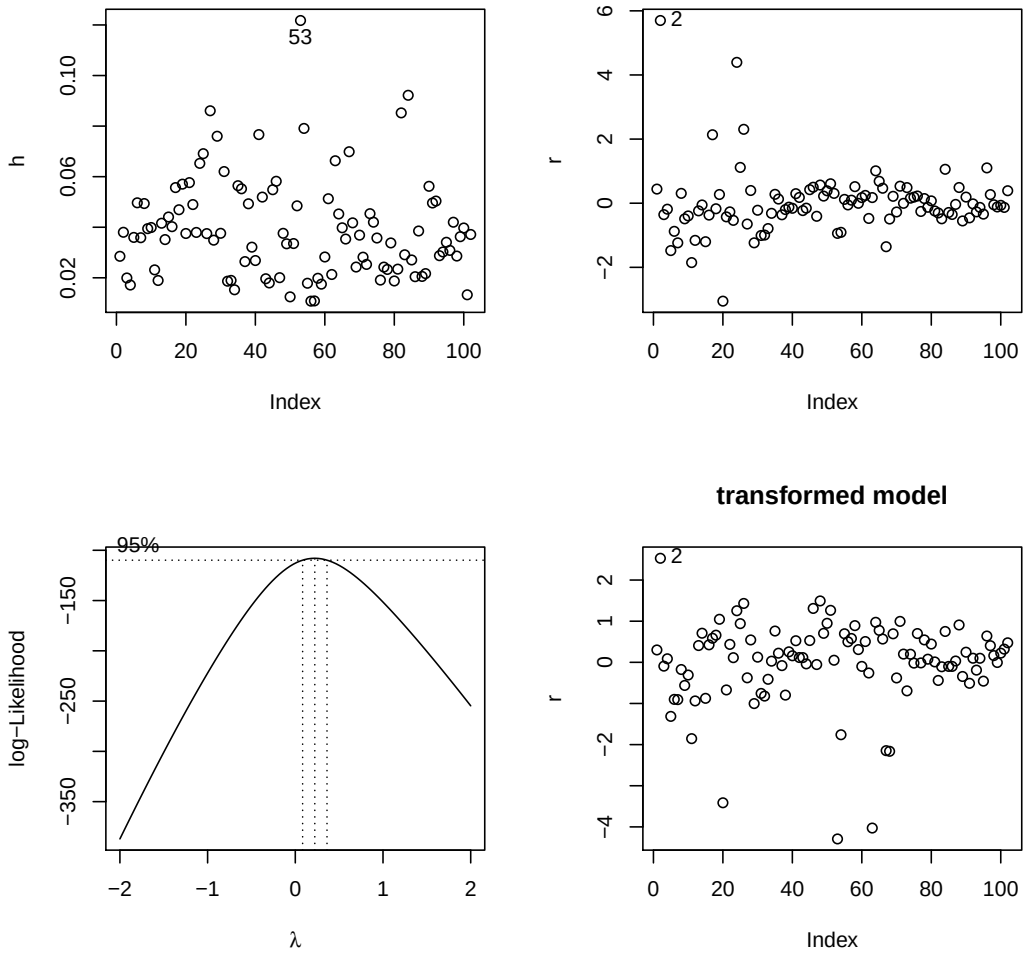
[Question 6 continues on the next page]

i. Find the density of the response variable, $f(y|\boldsymbol{x})$, given the distributional assumption (3). Hence, find the log–likelihood $L(\boldsymbol{\beta}, \sigma^2, \lambda)$ of (independent) observations $y_1, \ldots, y_n$.

ii. Without providing further calculations, explain the strategy which is used to arrive from here at the profile–log–likelihood $L_p(\lambda)$.

iii. Consider the bottom left panel of the provided figure, which shows, for model (2), a graph of $L_p(\lambda)$ versus λ, peaking at some value $\hat{\lambda}$ with a log–likelihood of $L_p(\hat{\lambda}) = -108.0$. The plot also contains a 95% confidence interval for the true value of λ, which is of the type $\{\lambda | L_p(\lambda) > c\}$. Give the exact numerical value of c. Discuss whether this plot gives evidence to transform the response.

iv. The bottom right panel shows the studentised residuals after applying a log–transformation onto the response. Discuss whether this specific transformation is consistent with the result from the Box–Cox analysis, and whether this residual plot suggests that the transformation has worked well. Explain in words how the corresponding plot for the leverage values would look after this response transformation.



7. We consider a q–variate random vector $X = (X_1, \ldots, X_q)^T \sim (\mathbf{0}, \mathbf{\Sigma})$, where $q \geq 2$, and $\mathbf{0}$ denotes a vector of dimension q containing only zeros. Denote as usual by $\lambda_1 > ... > \lambda_q$ the q ordered eigenvalues of $\mathbf{\Sigma}$, and by $\boldsymbol{\gamma}_j = (\gamma_{j1} \ldots, \gamma_{jq})^T$ the corresponding j–th eigenvector, $j = 1, \ldots, q$. Denote further

$$\mathbf{\Lambda}_r = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{pmatrix}$$

a diagonal matrix containing the first r ordered eigenvalues of $\mathbf{\Sigma}$, and by $\mathbf{\Gamma}_r$ a $q \times r$ matrix which has the corresponding eigenvector $\boldsymbol{\gamma}_j$ in its $j-$th column, $j = 1, \ldots, r$.

We are interested in approximating X by the best fitting bivariate linear subspace. We know that the solution to this problem is the plane through $\mathbf{0}$ spanned by the vectors $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, with coordinates on this plane given by the principal component scores $\boldsymbol{\gamma}_1^T X$ and $\boldsymbol{\gamma}_2^T X$. We can combine these to form a bivariate random vector $T = \mathbf{\Gamma}_2^T X$.

(a) The random vector T can be written as

$$T = \boldsymbol{v}_1 X_1 + \ldots + \boldsymbol{v}_q X_q,$$

with appropriate vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q \in \mathbb{R}^2$. Give expressions for the $\boldsymbol{v}_j, j = 1, \ldots, q$.

(b) Show $\mathbf{\Sigma}\mathbf{\Gamma}_2 = \mathbf{\Gamma}_2\mathbf{\Lambda}_2$.

(c) Derive $E(T)$ and $\mathrm{Var}(T)$.

(d) The bivariate coordinates T can be decompressed into the original (q–variate) data space via the operation $X' = \mathbf{\Gamma}_2 T$.

i. Derive $E(X')$ and $\mathrm{Var}(X')$.

ii. Show that, if $q = 2$, then $X' = X$.

(e) The following 'scree plot' was obtained through a principal component analysis of a data set of dimension $q = 60$, where each of the 60 variables measure the energy of sonar signals within a certain frequency band, after these signals bounced off from metallic objects or rocks. (The sonar signals were originally given on a scale from 0 to 1, but were mean–centered for this analysis in order to comply with the framework of this question).

i. Explain what we see in this plot, referring to the notation outlined above.

ii. Would you deem it adequate to approximate this data set by a plane?