Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May	2017	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:	

ED01/2017

SECTION A

1. An air pollution monitoring station in the city of Munich has recorded daily average SO_2 concentrations over a period of 14 consecutive days. We denote the logarithms of these daily averages by $y_1, \ldots, y_{14}$, which will form our response variable to which we refer as "pollution" in what follows. We are interested in modelling the responses y_i in dependence of the daily average temperatures $x_1, \ldots, x_{14}$ (recorded in degrees Celsius on the same 14 days), as well as an indicator z_i which takes the value 0 if day i is a weekday, and 1 if day i is a Saturday or Sunday. The full data set is provided below.

i	1	2	3	4	5	6	7
y_i	-3.147	-2.830	-3.016	-3.079	-3.541	-2.976	-2.781
x_i	16.47	16.02	16.81	22.87	21.68	21.23	20.55
z_i	0	0	0	1	1	0	0
i	8	9	10	11	12	13	14
y_i	-3.352	-2.765	-1.897	-2.120	-2.453	-1.973	-2.235
x_i	18.32	15.96	15.36	12.47	12.46	11.77	11.72
z_i	0	0	0	1	1	0	0

   (a) We are fitting the linear model $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i$. Write down the first four rows of the design matrix, $\boldsymbol{X}$.

   (b) Denote $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and s^2 the usual unbiased estimator of the error variance. You can use in what follows that

$$\boldsymbol{C} = \begin{pmatrix} 1.54887 & -0.08823 & -0.01627 \\ -0.08823 & 0.00537 & -0.00510 \\ -0.01627 & -0.00510 & 0.35484 \end{pmatrix},$$

   $s^2 = 0.1338985$, and $\boldsymbol{X}^T(y_1, \ldots, y_{14})^T = (-38.1650, -656.4754, -11.1930)^T$. Find $\hat{\beta}_j$, $j = 1, 2, 3$, and their standard errors $SE(\hat{\beta}_j)$, $j = 1, 2, 3$.

   (c) Assume that on a particular Tuesday one observes an average temperature $x_0 = 16.5°$. We would like to predict the true, unknown pollution y_0 on this day, using the fitted model. Hence, find

      (i) the predicted pollution, $\hat{y}_0$, on that day;

      (ii) a 95% confidence interval for the expected pollution $E(y_0|x_0, z_0 = 0)$ on that day;

      (iii) a 95% prediction interval for the actual pollution y_0 on that day.

2. For a linear model of type $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\beta} \in \mathbb{R}^p$, the *hat matrix* is given by

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

(a) Show $\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{H}$ and $\mathrm{Tr}(\boldsymbol{H}) = p$.

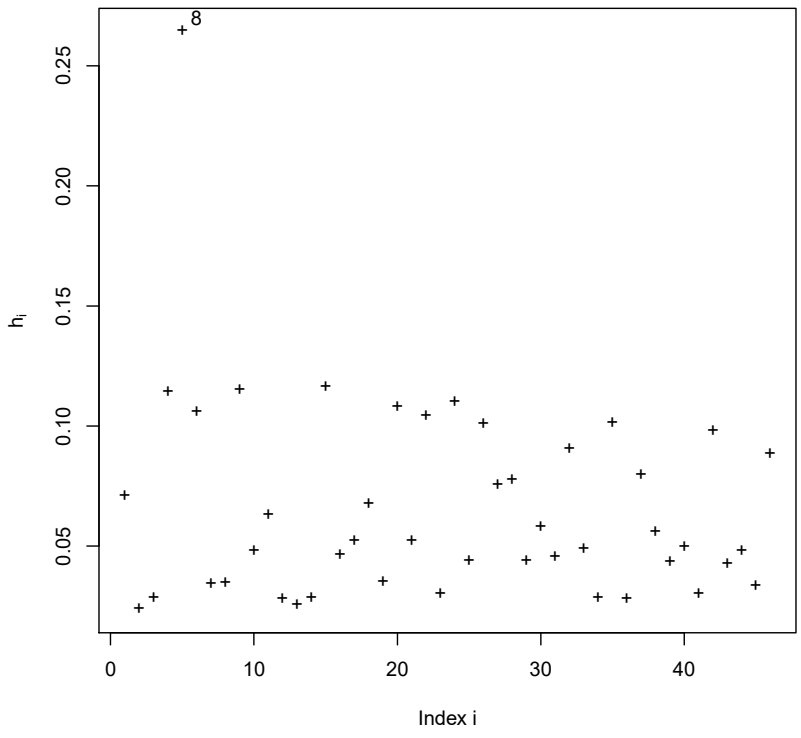
(b) Of particular interest are the diagonal values of $\boldsymbol{H}$, the so-called *leverage values* $h_i, i = 1, \ldots, n$. Show $0 \leq h_i \leq 1$.

(c) The figure below shows a plot of h_i versus i for a linear model fitted to a particular data set with $n = 46$.

i. Give an interpretation of this plot.

ii. Is this plot useful to judge whether the linear model fit would change considerably if the observation labelled "8" were removed from the data set? If so, provide this judgement. Otherwise, suggest an alternative measure to deal with this question (no formulae necessary).

iii. The mean of the plotted leverage values is 0.06521739. Find p.



3. The data below are from an experiment on the tensile strength of cement (Hald, 1952). We wish to model the relationship between $y =$ "tensile strength" (in kg/cm^2) and $z =$ "curing time" (days).

z	y	$\sum(y_i - \bar{y})^2$
1	13.0 13.3 11.8	1.26
2	21.9 24.5 24.7	4.88
3	29.8 28.0 24.1 24.2 26.2	24.272
7	32.4 30.4 34.5 33.1 35.7	16.428
28	41.8 42.6 40.3 35.7 37.3	34.812

We fit a simple linear regression model of y vs. z, yielding the residual sum of squares RSS $= 625.84$. We denote the error variance of the linear model by σ^2.

(a) For the data above, calculate the "pure estimate" of σ^2. Compare it to the model-based estimate (using the RSS), and give a brief interpretation.

(b) Next, carry out formally the F-test for lack-of-fit at the 1% level of significance.

(c) In the light of your results from (a) and/or (b), decide whether or not you would transform the covariate z, and if so, suggest a suitable transformation.

4. We are given a random vector $X = (X_1, \ldots, X_q)^T$ with $\mathrm{Var}(X) = \mathbf{\Sigma} = [\Sigma_{ij}]_{1 \le i \le q, 1 \le j \le q}$.

(a) Define the correlation matrix $\boldsymbol{R}$ of X in terms of the values $[\Sigma_{ij}]_{1 \le i \le q, 1 \le j \le q}$.

(b) Let $\tilde{X}_j = \frac{X_j}{\sqrt{\mathrm{Var}(X_j)}}$ and $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_q)^T$. Show that $\mathrm{Var}(\tilde{X}) = \boldsymbol{R}$.

(c) From a 46×3 dimensional data set, an estimated correlation matrix has been produced by firstly computing the sample variance matrix $\hat{\mathbf{\Sigma}}$, and then applying the definition requested in part (a), yielding

$$\hat{\boldsymbol{R}} = \begin{pmatrix} 1.000 & 0.901 & -0.685 \\ 0.901 & 1.000 & -0.559 \\ -0.685 & -0.559 & 1.000 \end{pmatrix}.$$

(i) How would $\hat{\boldsymbol{R}}$ change if the maximum likelihood–based variance matrix had been used instead of the sample variance matrix? Explain your answer.

(ii) Assume we know additionally that $\hat{\Sigma}_{11} = 27.6740$, $\hat{\Sigma}_{22} = 0.0283$, $\hat{\Sigma}_{33} = 0.2346$. Reconstruct the full matrix $\hat{\mathbf{\Sigma}}$, and verify that it is a valid variance matrix.

SECTION B

5. We consider a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, where $\mathbf{0}$ denotes a vector of appropriate length consisting only of zeros, and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Denote by $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ the least squares estimator of $\boldsymbol{\beta}$, and $s^2 = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ the unbiased estimator of σ^2.

   (a) Derive the expectation and variance of $\hat{\boldsymbol{\beta}}$. Hence, give the sampling distribution of $\hat{\boldsymbol{\beta}}$.

   (b) Write down the expression for the (squared) Mahalanobis distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$, and give its distribution.

   (c) Write down the expression for the (squared) Mahalanobis distance between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, and give its distribution.

   (d) Prove the decomposition

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

   (e) Using (b), (c), and (d), justify that, after appropriate standardization, the sampling distribution of s^2 is given by a χ^2 distribution, that is

$$cs^2 \sim \chi^2_k, \tag{1}$$

   and give the constant c, as well as the degrees of freedom k. [*Note:* Please explain your line of reasoning, but no formal proof is required. In particular, you do not need to show that $\hat{\boldsymbol{\beta}}$ and s^2 are independent.]

   (f) Give $E(s^2)$, and develop a formula for $\mathrm{Var}(s^2)$. [*Hint:* You can use that $\mathrm{Var}(\chi^2_k) = 2k$, for $k \in \mathbb{Z}^+$. If you could not solve part (e), please work with equation (1) as displayed.]

6. (a) We are given monthly records $\boldsymbol{x}_i = (x_{i1}, x_{i2})^T, i = 1, \ldots, 12$, of excess returns of the durables industry (X_1) and construction industry (X_2) in the year 2002:

i	1	2	3	4	5	6
x_{i1}	-3.82	3.09	3.58	-0.91	-1.26	-7.85
x_{i2}	-0.20	1.52	-0.47	-0.85	-3.46	-5.77
i	7	8	9	10	11	12
x_{i1}	-12.97	1.87	-11.07	4.58	15.63	-4.89
x_{i2}	-12.69	1.85	-12.58	5.00	0.99	-5.13

Some summary statistics are: $\bar{x}_1 = -1.168$, $\bar{x}_2 = -2.649$, $\sum x_{i1}^2 = 684.4$, $\sum x_{i2}^2 = 423.6$, $\sum x_{i1}x_{i2} = 425.0$.

(i) Provide a scatterplot of the data.

(ii) Compute the sample variance matrix $\hat{\boldsymbol{\Sigma}}$.

(iii) Compute the Mahalanobis distances to the mean for cases 1 and 11.

(iv) Stating any assumptions made, test whether observations 1 and 11 correspond to outliers at the 5% level of significance.

(b) Principal component analysis (PCA) is a statistical technique which provides a sequence of "best" linear approximations to a data cloud.

[*Note:* This part of the question is **not** to be answered in the context of the data in part (a).]

(i) Explain briefly in which sense the approximation provided by the principal components is "best".

(ii) Given a $q-$dimensional random vector X, we know that the principal components correspond to orthonormal eigenvectors $\boldsymbol{\gamma}_j$ of the variance matrix $\boldsymbol{\Sigma} = \mathrm{Var}(X)$, i.e. $\boldsymbol{\Sigma}\boldsymbol{\gamma}_j = \lambda_j\boldsymbol{\gamma}_j$, with $\boldsymbol{\gamma}_i^T\boldsymbol{\gamma}_j = 1$ if $i = j$ and 0 otherwise, and λ_j ordered from largest to smallest.
Deduce from this the eigen decomposition of $\boldsymbol{\Sigma}$.

(iii) Show that the sum of the eigenvalues λ_j equals the total variance of X.

(c) For the data from part (a), a principal component analysis gives $\lambda_2 = 7.50$. From this, find λ_1, and show that $\boldsymbol{\gamma}_1 = (0.834, 0.552)^T$. Draw the resulting first principal component line into the scatterplot produced in part (a).

7. We are given a multiple linear regression model in the form $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$, where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$.

   (a) Show that, for models involving an intercept, one has $\boldsymbol{X}^T\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$ and $\hat{\boldsymbol{Y}}^T\hat{\boldsymbol{\epsilon}} = \boldsymbol{0}$, where $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\epsilon}}$ are the vectors of fitted values and residuals, respectively, after the usual least squares fit;

   (b) Hence, for models involving an intercept, show that

$$SST = SSR + SSE \tag{2}$$

   where $SST = \sum_{i=1}^n (y_i - \bar{y})^2$, $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, and $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Also explain why equation (2) is generally *not* correct if there is *no* intercept in the model.

   (c) The statistic F for the overall $F-$test is defined by

$$F = \frac{SSR/(p-1)}{SSE/(n-p)}.$$

   Define the coefficient of determination (R^2) in terms of the quantities introduced in part (b), and find an expression for R^2 which only depends on F, n and p.

   (d) We are given a real data set with $n = 14$, which after fitting the linear model with $p = 3$ (including the intercept) yields the value $F = 7.539$.

      i. Carry out the overall F–test at the 0.01 level of significance.

      ii. Compute R^2, and interpret the result. [*Note:* If you could not solve part (c), you can make use of the information $SSR = 1.999$.]

      iii. Assume that, for subject–matter considerations, the data analyst decides to remove the intercept. They refit the model using some statistical software, which reports a value $R^2 = 0.9803$ for the fitted model. Does this give evidence that the model without intercept is preferable to the model with intercept? Explain your answer carefully.