Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2021	MATH3051-WE01

Title: Statistical Methods III

Time (for guidance only):	2 hours 30 minutes	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	Yes	Models Permitted: There is no restriction on the model of calculator which may be used.

Instructions to Candidates:	Credit will be given for your answers to all questions. All questions carry the same marks. Please start each question on a new page. Please write your CIS username at the top of each page. To receive credit, your answers must show your working and explain your reasoning.

Revision:	

ED01/2021

Q1 **1.1** A study conducted in central Virginia focused on the prevalence of obesity, diabetes and other cardiovascular risk factors. We consider a subset of the study's data set, consisting of 103 individual records. The response variable under consideration is `glycosolated hemoglobin`, which is a common diagnostic measure of diabetes (values greater than 7 indicate a positive diagnosis of the disease). We consider two explanatory variables on logarithmic scale; namely, the ratio `Cholesterol/HDL` (`HDL`: High Density Lipoprotein) and `stabilised glucose` levels.

Let Y denote the response `glycosolated hemoglobin`, X_1 the logarithm of `Cholesterol/HDL` and X_2 the logarithm of `stabilised glucose`. Fitting the model

$$y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \epsilon_i, \quad \text{for } i = 1, \ldots, 103,$$

yields the following quantities:

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \begin{pmatrix} 0.41808 & -0.00500 & -0.08881 \\ -0.00500 & 0.02301 & -0.00616 \\ -0.08881 & -0.00616 & 0.02126 \end{pmatrix},$$

$$\boldsymbol{X}^T\boldsymbol{Y} = (2175.730 \quad 3236.672 \quad 10232.817)^T.$$

We have assumed that the errors are independent and identically distributed, following a normal distribution; namely, $\epsilon_i \sim N(0, \sigma^2)$. The error variance σ^2 estimate of the model is $s^2 = 1.5185^2$.

(i) Find the 95% confidence intervals of β_1, β_2 and β_3.

(ii) Suppose that we obtain new measurements from an individual within the cohort of the study. The new measurements are $x_{01} = 1.447$ and $x_{02} = 4.682$. Find:

A. a 95% confidence interval for the expected `glycosolated hemoglobin` level $E(y_0 \mid x_{01} = 1.447, x_{02} = 4.682)$.

B. a 95% prediction interval for the individual's `glycosolated hemoglobin` level y_0.

Show the required calculations in parts (i) and (ii).

1.2 Figure 1 is the output of using the R function boxcox{MASS} for a particular Normal linear regression model. By eye, deduce a 95% confidence interval for λ. Does this suggest a need for a transformation to be applied to the response? If yes, what transformation would you choose? What features would suggest a logarithmic transformation as an appropriate one?

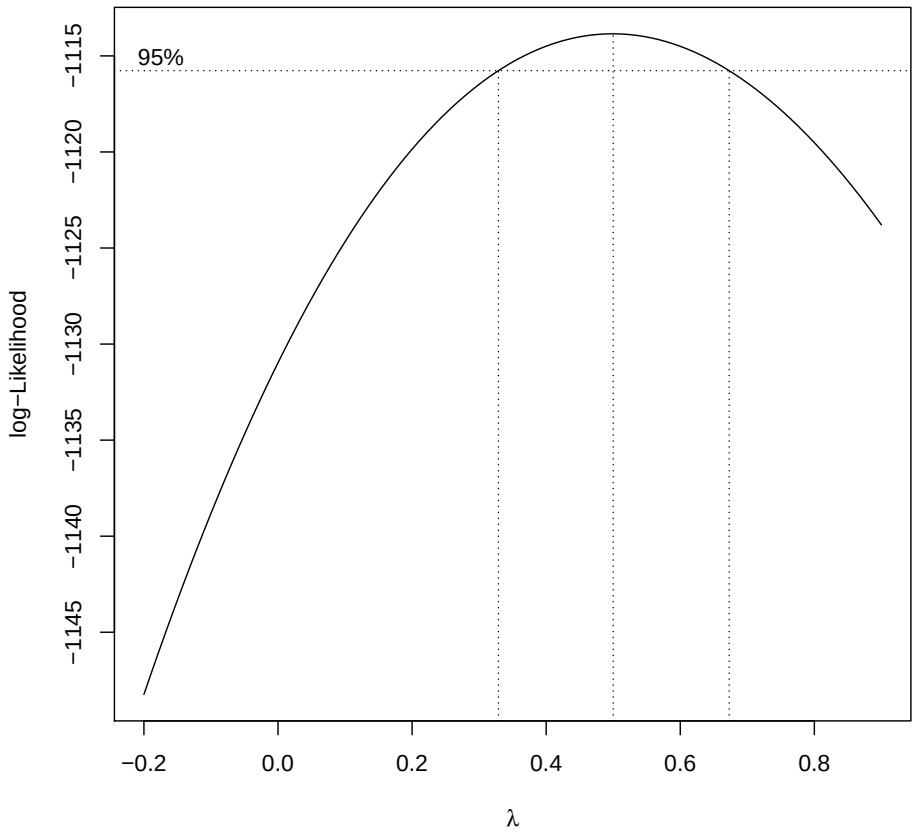
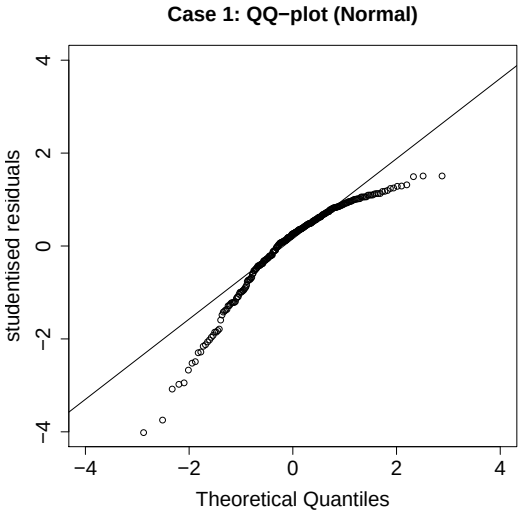
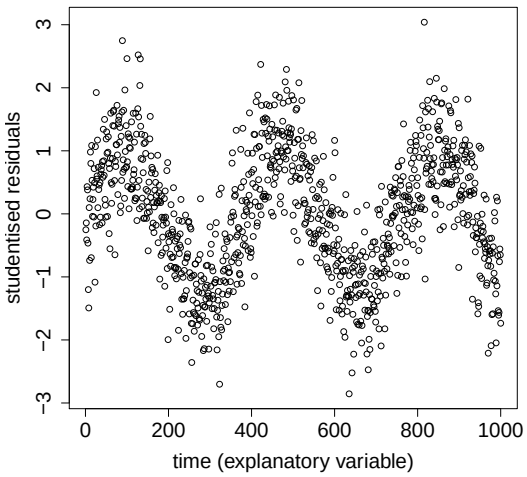


Figure 1:

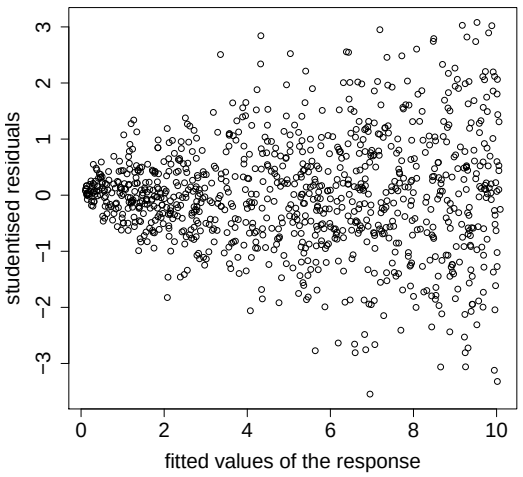
1.3 We consider four cases (Cases I, II, III, IV) where specific Normal linear regression models are fitted against different data-sets. Figures 2a, 2b, 2c, and 2d present diagnostic plots for Cases I, II, III, IV correspondingly. In Figure 2b, the values in the horizontal axis refer to a potential explanatory variable. For each case, explain if the corresponding plots indicate anything meriting investigation, and if they do, explain how you would deal with it.



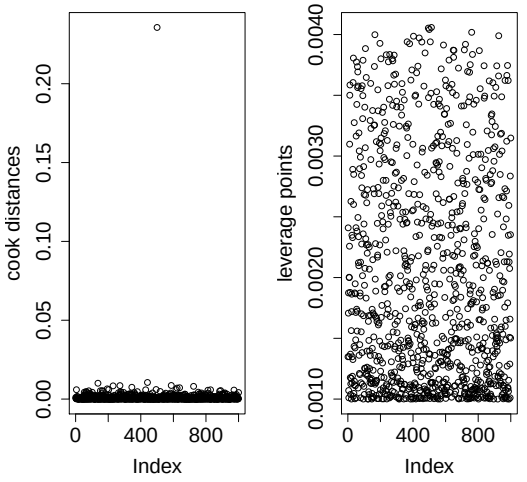
(a) Case I



(b) Case II



(c) Case III



(d) Case IV

Figure 2:

Q2 Let $X \sim N(0,1)$ and $Y = WX + 7$, where the random variable W is independent of X and has the following probability mass function:

$$W = \begin{cases} -1, & \text{with probability } \pi, \\ 1, & \text{with probability } 1-\pi. \end{cases}$$

2.1 Consider the following cases

(i) $\pi = 0.5$;

(ii) $\pi = 0.4$,

and examine whether X and Y are correlated or not by calculating the corresponding covariance of X and Y under each case.

[*Note: Take for granted that when X and W are independent X^2 and W are also independent; formal proof of that is not required.*]

2.2 What is the marginal distribution of Y and what are its parameters under cases (i) and (ii) in **2.1**?

2.3 Assume that you are given a univariate sample of size 20 which contains samples from the marginal distribution of X as well as from the marginal distribution of Y. You are asked to detect possible outliers based on Mahalanobis distances using all observations for estimating the sample mean and the sample variance. The resulting squared Mahalanobis distances are given below.

1	2	3	4	5	6	7	8	9	10
0.408	7.592	0.521	0.059	0.067	0.512	0.039	0.009	0.026	7.295
11	12	13	14	15	16	17	18	19	20
0.044	0.055	0.405	1.612	0.003	0.166	0.157	0.001	0.004	0.024

Test H_0 : "sample$_i$ is an outlier" vs. H_1 : "sample$_i$ is not an outlier" for $i = 1, \ldots, 20$ using the chi-square distribution with one degree of freedom at a significance level of 5%. Report the critical value and explain how the test works. Which sample points are detected as outliers? Apply the Bonferroni correction, explain how the test works in this case, and report the resulting critical value. Which sample points are detected as outliers in this case?

2.4 You are subsequently informed that 90% of the samples in **2.3** originate from the marginal distribution of X and that the remaining samples originate from the marginal distribution of Y. Comment on the efficacy of the test you implemented in **2.3** (without the Bonferroni correction) in view of the new information. Would the efficacy of the test be affected if you were given only the first 10 samples above and, if yes, how?

Q3 **3.1** Consider the simple linear model

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

for $i = 1, \ldots, n$. Identify a case where the matrix $\boldsymbol{X}^T\boldsymbol{X}$, where $\boldsymbol{X}$ is the $n \times 2$ design matrix, does not have an inverse and, thus, least-squares estimation does not lead to a unique solution. Explain what this means for the values of the predictor variable.

3.2 A $100 \times (1-\alpha)\%$ prediction interval for y_0 given a new observation $\boldsymbol{x}_0$ has limits

$$\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \times s\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0},$$

where $\hat{\boldsymbol{\beta}}$ is the p-dimensional vector of least-squares estimates, s is the error standard deviation estimate and $t_{n-p,\alpha/2}$ is the quantile of a t-distribution with $n-p$ degrees of freedom which has right tail probability $\alpha/2$. An accurate approximation to the above prediction interval, for $n \to \infty$, is given by

$$\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}} \pm z_{\alpha/2} \times s,$$

where $z_{\alpha/2}$ is the Gaussian quantile with right tail probability $\alpha/2$. Show analytically how this approximation is derived, considering again the simple linear model in **3.1** where $p = 2$. Assume that for $n \to \infty$: (i) $\sum_{i=1}^n x_i^2/n$ is bounded and (ii) the sample variance of the predictor variable does not converge to zero.

3.3 The least-squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ minimises the following objective function

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

Suppose that we want to minimise $R(\boldsymbol{\beta})$ subject to a constraint of the form $\boldsymbol{q}^T\boldsymbol{\beta} = c$, where $\boldsymbol{q}$ is a $p \times 1$ vector and c is some constant. This is equivalent to minimising the corresponding Lagrangian form

$$R^c(\boldsymbol{\beta}, \lambda) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) - 2\lambda(\boldsymbol{q}^T\boldsymbol{\beta} - c),$$

where scalar λ is the Lagrange multiplier. Show that the least-squares estimator in this case is given by

$$\hat{\boldsymbol{\beta}}^c = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{q}\left[\boldsymbol{q}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{q}\right]^{-1}(c - \boldsymbol{q}^T\hat{\boldsymbol{\beta}}).$$

[*Hints: Work with the gradient of* $R^c(\boldsymbol{\beta}, \lambda)$ *and solve with respect to* $\boldsymbol{\beta}$ *and* λ. *You do not need to use second partial derivatives in order to show that* $\hat{\boldsymbol{\beta}}^c$ *is a minimum as* $R^c(\boldsymbol{\beta}, \lambda)$ *is convex*.]

Q4 **4.1** Consider a Normal linear regression statistical model $y_i = \left[1,\ \boldsymbol{x}_i^\top\right]\boldsymbol{\beta} + \epsilon_i$ with $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ for $i = 1, ..., .n$, with p explanatory variables $\{x_1, ..., x_p\}$, unknown regression coefficients $\boldsymbol{\beta}$, unknown variance σ^2 and n observations.

(i) Show that the Akaike Information Criterion can be written as

$$\mathrm{AIC} = n\log\left(\frac{\mathrm{SSE}}{n}\right) + n\log(2\pi) + n + 2(p+2)$$

where SSE is the residual sum of squares.

(ii) Consider the full regression model denoted as A and a reduced regression model denoted as B. B is nested within A. Show that

$$\mathrm{AIC_B} - \mathrm{AIC_A} = n\log\left(\frac{p_\mathrm{A} - p_B}{n - p_A}F + 1\right) - 2(p_\mathrm{A} - p_B)$$

where indexes $\cdot_\mathrm{A}$ and $\cdot_\mathrm{B}$ indicate that the associated quantities are based on the "Full model" and "Reduced model" correspondingly. F denotes the test statistic of the corresponding sequential ANOVA table for testing the adequacy of model B against model A.

(iii) For large number of observations, i.e. $n \to \infty$, show that removing one continuous regressor from the full regression model based on the AIC criterion is equivalent to removing it based on a marginal t-test on the associated regression coefficient, and specify the rejection area.

4.2 Consider a Normal linear regression statistical model $\boldsymbol{y} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \boldsymbol{I}_n\sigma^2)$, where $\beta \in \mathbb{R}^q$, with n observations $\{(y_i, x_i)\}_{i=1}^n$. Consider the following statistic

$$M_i = \frac{\det\left(\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{(i)}\right)\right)}{\det\left(\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\right)},\quad i = 1, ..., n;$$

where the hat $\hat{\cdot}$ indicates the MLE of the associated parameter, and the index $\cdot_{(i)}$ indicates that the associated quantity has been computed by omitting the i-th observation.

(i) Derive that

$$M_i = \left(\frac{\mathrm{MSE}_{(i)}}{\mathrm{MSE}}\right)^q \frac{1}{1 - h_i},\quad i = 1, ..., n;$$

where $h_i = \boldsymbol{x}_i^\top(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{x}_i$ and MSE denotes the Mean Square Error in the corresponding regression.

(ii) Assume that the number of observations n is large. First, describe how the magnitude of M_i behaves when the i-th observation is an outlier but not influential point. Then, describe how the magnitude of M_i behaves when the i-th observation is an influential point but not an outlier. Finally, describe around what value M_i may be if the i-th observation is neither an outlier nor an influential point.

Q5 **5.1** Let ξ_1 and ξ_2 be two independent uniform random variables on $[0, 1]$. Suppose $X_1 = \xi_1$, $X_2 = \xi_2$, $X_3 = \xi_1 + \xi_2$, and $X_4 = \xi_1 - \xi_2$.

(i) Compute the correlation matrix $\boldsymbol{R}$ of $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^\top$, and show that $\gamma_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 1, 0\right)^\top$ and $\gamma_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 1\right)^\top$ are eigenvectors of $\boldsymbol{R}$ corresponding to the nontrivial eigenvalue.

(ii) How many normalised Principal Components (PCs) are important? Compute the normalised PCs of $\boldsymbol{X}$, compute the proportion of variance that each explain, and interpret them.

(iii) Explain which potential issue you could address if you perform a PC regression where Y is regressed on the important normalised PC's of $\boldsymbol{X}$, instead of regressing Y on $\{X_1, X_2, X_3, X_4\}$.

5.2

(i) Consider a Normal linear regression model $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with normaly distributed errors $\boldsymbol{\epsilon}$, where the arithmetic average of $\boldsymbol{Y}$ is zero ($\bar{Y} = 0$) and the $n \times p$ design matrix $\boldsymbol{Z}$ has columns with arithmetic averages zero $\bar{Z}_j = 0$, $j = 1, ..., p$. Let $\boldsymbol{W} = (\boldsymbol{W}_1, ..., \boldsymbol{W}_p)$ be the Principal Components (PCs) of $\boldsymbol{Z}$. Assume that the true vector of the regression coefficients $\boldsymbol{\beta}$ is in the direction of the j-th eigenvector of $\boldsymbol{Z}^\top\boldsymbol{Z}$. Show that when Y is regressed on $\boldsymbol{W}$, apart from the random error term ϵ, the j-th PC $\boldsymbol{W}_j$ alone will contribute everything to the fit, while the remaining PCs will contribute nothing.

(ii) Based on part (i), comment in detail about the use of the naive implementation of the principal component regression. Naive implementation of PC regression describes the procedure where: one computes the PCs of $\boldsymbol{Z}$, chooses a subset of PCs corresponding to eigenvalues greater than a threshold value, and regresses $\boldsymbol{Y}$ against this subset of PCs.

END