Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2026	MATH2801-WE01

Title: Data Science and Statistical Modelling II

Time:	2 hours	
Additional Material provided:	Statistical tables	
Materials Permitted:	None	
Calculators Permitted:	Yes	Models Permitted: Casio FX83 series or FX85 series.

Instructions to Candidates:	Answer all questions. The indicative marks shown in brackets for the main parts of each question are given as a guide to the weighting the markers expect to apply. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2026

1. The Shear-o-Matic automatic sheep shearing machine invented by Shaun the Sheep has now been in use for a year. Shaun has been analysing the reliability of the machine using a model that assumes failure times follow an Exponential distribution, which has 'Coefficient of Variation' (CV) equal to 1, where CV := σ/μ. However, he suspects the Shear-o-Matic might be suffering from "wear-out", which would cause failures to cluster more around the mean lifetime than under an Exponential model, leading to a CV less than 1. Five local farmers reported failure times for the Shear-o-Matic (in days):

$$\mathbf{x} = (131, \quad 51, \quad 76, \quad 95, \quad 142)$$

(a) Shaun wants to test the hypothesis $H_0 : \text{CV} = 1$ against $H_1 : \text{CV} < 1$. The natural test statistic is the sample Coefficient of Variation, $T = s/\bar{x}$, where s is the sample standard deviation and $\bar{x}$ is the sample mean.

Calculate the observed test statistic t_{obs} for the failure data provided and describe in detail how you would perform a Monte Carlo hypothesis test for this specific problem. Be precise about what distributions you simulate from. *Hint: The distribution of the sample Coefficient of Variation for an Exponential distribution does not depend on the parameter, λ.* **[6]**

(b) 100 simulations of the test statistic under the null (ordered, 10 per line) follow:

0.34, 0.36, 0.37, 0.39, 0.43, 0.44, 0.47, 0.49, 0.50, 0.50
0.51, 0.51, 0.51, 0.51, 0.53, 0.54, 0.55, 0.55, 0.55, 0.56
$\cdots$ 60 other simulations $\cdots$
1.12, 1.14, 1.15, 1.17, 1.17, 1.18, 1.21, 1.21, 1.23, 1.23
1.25, 1.26, 1.31, 1.34, 1.36, 1.38, 1.38, 1.39, 1.39, 1.50

Estimate the p-value based on this (small) Monte Carlo simulation. Would you advise that the Shear-o-Matic does, or does not, have "wear-out"? **[4]**

(c) Shaun wants an estimate of the CV for all Shear-o-Matics, without relying on the Exponential assumption. He defines $S(\mathbf{x}) := s/\bar{x}$.

Perform one manual iteration of the non-parametric bootstrap by:

(i) Writing down a possible bootstrap resample, $\mathbf{x}^\star$, that contains the value 142 exactly three times.

(ii) Calculating the bootstrap statistic for this resample. **[4]**

(d) A total of 1000 bootstrap resamples are taken. Shaun calculates that $\bar{S}^\star = 0.3388, \widehat{\text{Var}}(S(\mathbf{x})) = 0.0096$, and that the 95% percentile confidence interval is $(0.1299, 0.5023)$. Compute the 95% Normal confidence interval and estimate the bias in the CV estimator. Comment on what you find. **[6]**

(e) Finally, Shaun needs to quantify the uncertainty in the mean lifetime of the Shear-o-Matic, because farmers demand evidence it is at least 65 days. Alas, Shaun computes a confidence interval for the mean as,

$$\bar{x} \pm t_4 \frac{s}{\sqrt{n}} = 99 \pm 2.776 \frac{37.82}{\sqrt{5}} = (52.05, 145.95)$$

where t_4 is the critical value of a $t-$distribution with 4 degrees of freedom at 95% confidence. As Shaun's statistical consultant, you discuss with him and discover he only ever made 10 Shear-o-Matics and will *never* make any more. What should the interval be and can the farmers complain? **[5]**

2. The Epanechnikov kernel is often used in non-parametric statistics and has the probability density function (pdf):

$$f(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if } x \in [-1,1] \\ 0 & \text{otherwise} \end{cases}$$

You want to estimate $\mu = \mathbb{E}_f[X^2]$ and so construct a rejection sampler to simulate from $f(\cdot)$. You run the algorithm and end up with accepted samples from the proposal $\{x_1, \ldots, x_n\}$, and rejected samples from the proposal $\{x'_1, \ldots, x'_m\}$. Then, you use the usual Monte Carlo estimator, $\hat{\mu} = n^{-1} \sum_{i=1}^{n} x_i^2$, based only on the accepted samples.

(a) (i) Show that the Uniform distribution on $[-1, 1]$ can be used as a proposal for this rejection sampler to produce samples from $f(x)$. Find the optimal value of the constant, c, required for rejection sampling.

(ii) What is the probability that a proposal is accepted in any iteration? We usually fix the desired number of accepted samples, meaning we observe a random number of rejections, m. That is, we fix n and m is the realisation of a random variable M. Determine the expected number of rejected samples, $\mathbb{E}[M]$, as a function of n. **[6]**

Your colleague worries rejected samples are wasted, but does not know importance sampling, so they think up a new way to use the rejected samples.

They write down a decomposition of the Uniform$[-1, 1]$ proposal pdf, $g(\cdot)$, as a mixture of the Epanechnikov pdf, $f(\cdot)$, and (unknown) 'rejection' pdf, $f_{\text{rej}}(\cdot)$,

$$g(x) = \frac{1}{c} f(x) + \left(1 - \frac{1}{c}\right) f_{\text{rej}}(x)$$

(b) In a few sentences, explain why writing $g(\cdot)$ in this way would make sense. Then solve to find $f_{\text{rej}}(x)$ and prove that what you find is a valid pdf. **[3]**

(c) (i) Using the decomposition above, derive an expression for $\mu = \mathbb{E}_f[X^2]$ in terms of the second moment of the proposal, $\mathbb{E}_g[X^2]$, and the second moment of the rejection distribution, $\mathbb{E}_{f_{\text{rej}}}[X^2]$, <u>without</u> calculating the expectations.

(ii) Calculate $\mathbb{E}_g[X^2]$ and substitute the value into this expression, but leave $\mathbb{E}_{f_{\text{rej}}}[X^2]$ unevaluated (since in real-world problems it would usually be intractable). **[6]**

(d) Hence or otherwise, propose a new estimator, $\hat{\mu}_{\text{rej}}$, for $\mathbb{E}_f[X^2]$ that uses only the *rejected* samples instead of the accepted samples. **[3]**

(e) The variance of the usual estimator is $\text{Var}(\hat{\mu}) = \frac{8}{175n}$. Find $\text{Var}(\hat{\mu}_{\text{rej}})$ as a function of m (compute the integrals required, which would usually not be possible to do in real-world problems). Based on the expected size of m, prove whether it is usually better to use the estimator $\hat{\mu}$ or $\hat{\mu}_{\text{rej}}$. **[7]**

3. Suppose there are n observations $(x_i, y_i) \in \mathbb{R}^2$ related through the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the errors ε_i are independent and satisfy $\varepsilon_i \sim N(0, \sigma^2)$. Let $\bar{x}$ and s_x denote the sample mean and sample standard deviation, respectively, of the observed predictor values $x_1, \ldots, x_n$.

(a) Write down the design matrix $\mathbf{X}$, derive $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$, and hence show that the entries h_{ij} of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ are given by

$$h_{ij} = \frac{1}{n\left(\sum_{\ell=1}^{n} x_\ell^2 - n\bar{x}^2\right)} \left[n x_i x_j - n(x_i + x_j)\bar{x} + \sum_{\ell=1}^{n} x_\ell^2 \right]. \qquad (3.1)$$

[6]

(b) Using Equation (3.1), show that the diagonal elements $h_i \equiv h_{ii}$ of $\mathbf{H}$ can be written as

$$h_i = \frac{1}{n} + \frac{1}{n-1}\, d_M^2(x_i; \bar{x}, s_x^2),$$

where $d_M^2(x_i; \bar{x}, s_x^2)$ is the squared Mahalanobis distance between x_i and their mean $\bar{x}$. **[5]**

(c) Cooks distance for the ith observation is defined as

$$D_i = \frac{1}{2}\, r_i^2 \frac{h_i}{1 - h_i},$$

where h_i is the leverage value, and $r_i = \hat{\varepsilon}_i / (s\sqrt{1 - h_i})$ is the standardised residual. Here $\hat{\varepsilon}_i$ denotes the fitted residual and s is the residual standard deviation.

Cooks distance measures the *influence* of an observation on the fitted regression model. **Explain** what is meant by influence and how it differs from leverage. Then **describe** situations in which:

(i) Cooks distance D_i is equal to zero;

(ii) an observation has high leverage but low influence;

(iii) an observation has low leverage but high influence. **[6]**

(d) Suppose that $n = 25$, $p = 2$, $s = 7.674$. The following statistics have been obtained for four observations:

$\hat{\epsilon}_i$	h_i	r_i	D_i
-13.65	0.042	-1.817	?
-9.59	0.269	-1.462	0.393
0.51	0.040	?	0.000
-7.95	0.405	-1.343	0.614

Compute any missing values in the table above. Then, for each observation, state whether it appears to be: (i) potentially influential, (ii) an outlier, or (iii) influential. *Hint:* Use appropriate rules of thumb discussed in the lectures for leverage, outliers, and Cooks distance. **[8]**

4. Bluegill fish were sampled from two lakes, Lake Mary and Camp Lake. For each fish, length (in millimetres) and age (in years, treated as numeric) were recorded. The sample consists of 78 fish from Lake Mary and 140 fish from Camp Lake.

The relationship between fish `length` and `age` is modelled using the regression function:

$$\mathrm{E}[\texttt{length} \mid \texttt{age}, \texttt{lake}] = \beta_1 + \beta_2\,\texttt{age} + \beta_3\,\texttt{age}^2 + \beta_4\,\texttt{lake} + \beta_5\,\texttt{lake}{\cdot}\texttt{age} + \beta_6\,\texttt{lake}{\cdot}\texttt{age}^2,$$

where `lake` is an indicator variable taking the value 0 for Lake Mary and 1 for Camp Lake.

The `R` output provided below may be used to answer the following questions.

```
> full.model <- lm(length ~ (age + I(age^2)) * lake, data = bluegill)
> summary(full.model)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      13.622     11.787   1.156   0.2491
age              54.049      6.943   7.785 3.04e-13 ***
I(age^2)         -4.719      1.010  -4.672 5.30e-06 ***
lake1            22.527     14.126   1.595   0.1123
age:lake1       -13.614      8.095  -1.682   0.0941 .
I(age^2):lake1    1.631      1.146   1.423   0.1561
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> anova(full.model)
Analysis of Variance Table

Response: length
               Df Sum Sq Mean Sq  F value    Pr(>F)
age             1  84211   84211 618.4420 < 2.2e-16 ***
I(age^2)        1   9089    9089  66.7515 2.761e-14 ***
lake            1    751     751   5.5171   0.01975 *
age:lake        1    256     256   1.8795   0.17184
I(age^2):lake   1    276     276   2.0262   0.15607
Residuals     212  28867     136
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) Write down the regression models implied by the above regression function for the expected length, separately for Lake Mary and Camp Lake. Then **predict** the length of a 4-year-old fish from each lake, and comment on any differences between the predictions. **[7]**

(b) State, from the ANOVA table, the usual estimate of the common error variance σ^2. Then construct a 95% prediction interval for the length of a 4-year-old fish from either lake, using a large-sample approximation. If you were unable to answer part (a), explain clearly how such a prediction interval would be constructed. **[6]**

(c) Based on the ANOVA table, terms with a p-value (`Pr(>F)`) greater than 0.05 are considered negligible and are removed, resulting in the following reduced model:

$$\mathrm{E}[\texttt{length} \mid \texttt{age}, \texttt{lake}] = \beta_1 + \beta_2\,\texttt{age} + \beta_3\,\texttt{age}^2 + \beta_4\,\texttt{lake}.$$

Perform an F-test to assess whether the reduced model provides an adequate fit to the data compared with the full model. Clearly state the null and alternative hypotheses, and summarise your conclusions. **[8]**

(d) Residual plots are often used to assess model adequacy. Consider the residual plots for the model with a linear `age` term only (left) and for the model with a quadratic `age` term (right). Based on these residual plots, discuss whether including a quadratic term in `age` is necessary. **[4]**