MATH70071 Applied Statistics
2025-2026 end-of-module assignment

Submission deadline:
12:00 (noon) on Friday, 12/12/2025

Preparing your assignment

1. Submit a **single PDF report** containing your answers to all questions. The code used to produce your answers should be included only in the appendix; no code should appear in the main body of the report. The **main report should not exceed 10 pages**; the code appendix does not count towards the page count.

2. There are several acceptable options for preparing the report: the provided **R Markdown template**; other tools such as **Quarto**; or standard **LaTeX**.

3. **Two or three sentences** are usually sufficient for the verbal/explanatory parts of questions; longer answers are likely to be less clear. Only the **main steps of calculations** need to be shown. Avoid wasted space around figures and unnecessarily large plots (but make sure labels in figures are large enough to be clearly legible).

4. The report should be properly structured, and should be written using complete sentences. Marks are given both for the **content of the report** (correctness of code, numerical answers, *etc.*) and the **quality of the presentation** (clarity of plots, explanations, *etc.*).

5. At the beginning of your report you must include this **statement of originality**:

   "I, [YOUR CID], certify that this assessed coursework is my own work, unless otherwise acknowledged, and includes no plagiarism. I have not discussed my coursework with anyone else except when seeking clarification with the module lecturer via email or on MS Teams. I have not shared any code underlying my coursework with anyone else prior to submission."

Submitting your assignment

1. Before the above deadline **submit a single PDF report via Blackboard** (with, as above, your `R` code included as an appendix).

2. The filename should be **MScStatistics_AppliedStatistics_[YOUR CID].pdf** so, *e.g.*, `MScStatistics_AppliedStatistics_00123456.pdf` .

3. Your report will be **checked for plagiarism**.

The results of a medical survey aimed at assessing the influence of smoking on lung function and lung cancer are in the table `smoking.csv`, which, for each of the $N = 36$ people in the survey, lists these quantities:

column name	parameter	notation	units/values
	i	row number/index	{1, 2, . . . , N}
`age`:	A	age at the start of the survey	years
`weight`:	W	weight at the start of the survey	kg
`cigarettes`:	η	cigarettes smoked per day	
`capacity`:	L	lung capacity at the start of the survey	litres
`cancer`:	C	diagnosed with lung cancer during the survey	{true, false}

None of the participants in the survey had lung cancer initially.

The above units are used implicitly for all numerical quantities, so, *e.g.*, $\log(A)$ is the log of the age in years; but units should be included explicitly in any plots. $\Theta(\cdot)$ is the Heaviside step function; $\delta_{\mathrm{D}}(\cdot)$ is the Dirac delta function.

1. For the below analyses assume that the survey participants are completely independent of each other. Identify three ways in which this assumption could, in reality, be violated for a survey like this, and briefly explain the problem with the assumption of independence in each case.

   (6 marks)

2. The survey participants can be divided into smokers (*i.e.*, those with $\eta > 0$) and non-smokers, with S_i introduced as the proposition that participant i is a smoker (with $i \in \{1, 2, \ldots, N\}$). It would hence be possible to analyse the relationship between smoking (at any level) and lung cancer in terms of two probabilities: the probability of a smoker getting cancer, $p = \mathbb{P}(\mathsf{C}|\mathsf{S})$; and the probability of a non-smoker getting cancer, $q = \mathbb{P}(\mathsf{C}|\neg\, \mathsf{S})$. In the below sub-questions it is implicit that $0 \leq p \leq 1$ and $0 \leq q \leq 1$.

   (a) Assess qualitatively whether using $\mathsf{S}_{1:N}$ in place of $\eta_{1:N}$ would represent a loss of information in the analysis of this survey.

   (b) The full likelihood for inferring the values of p and q has the form $\mathbb{P}(\mathsf{C}_{1:N}|p, q, \mathsf{S}_{1:N})$. But it would also be possible to work with these summary statistics: the number of smokers who got cancer, $N_{\mathrm{s,c}}$; the number of non-smokers who got cancer, $N_{\neg\mathrm{s,c}}$; the total number of smokers, N_{s}; and the total number of non-smokers, $N_{\neg\mathrm{s}}$. In this case the likelihood would have the form $\mathbb{P}(N_{\mathrm{s,c}}, N_{\neg\mathrm{s,c}}, |p, q, N_{\mathrm{s}}, N_{\neg\mathrm{s}})$. Find expressions for both likelihoods and explain the relationship between them. Hence assess whether $N_{\mathrm{s,c}}$, $N_{\neg\mathrm{s,c}}$, N_{s} and $N_{\neg\mathrm{s}}$ are jointly sufficient statistics.

   (c) Adopting a uniform prior probability distribution for p and q, plot the two marginal posterior densities for p and q given $N_{\mathrm{s,c}}$, $N_{\neg\mathrm{s,c}}$, N_{s} and $N_{\neg\mathrm{s}}$ and assess whether there is evidence from this data that smoking is linked to lung cancer.

   (d) Consider these three models:

   - M_1: Smoking has a positive link with lung cancer and $P(p, q|\mathsf{M}_1) = \Theta(p - q)/2$.
   - M_2: Smoking is unrelated to lung cancer and $P(p, q|\mathsf{M}_2) = \delta_{\mathrm{D}}(p - q)$.
   - M_3: Smoking has a negative link with lung cancer and $P(p, q|\mathsf{M}_3) = \Theta(q - p)/2$.

   Calculate the three marginal liklihoods $\mathbb{P}(N_{\mathrm{s,c}}, N_{\neg\mathrm{s,c}}, |N_{\mathrm{s}}, N_{\neg\mathrm{s}}, \mathsf{M}_m)$, with $m \in \{1, 2, 3\}$, and again assess the evidence for a link between smoking and lung cancer.

   (18 marks)

3. The machine used for the lung capacity measurements is known to produce very noisy data, so $M = 20$ lung capacity measurements, $L_{1:M}$, were made of a volunteer, resulting in the calibration data-set in `calibration.csv`; these measurements can be considered independent of each other. The volunteer also had their lung capacity measured using a high-precision device which provided an effective true value of $L^* = 3.44$ litres.

   A suggested noise model for the lung capacity measurements is given by the (unnormalised) probability density

$$\tilde{P}(L|L^*, \Delta, \sigma, \beta) = \left\{1 + \frac{[L - (L^* + \Delta)]^2}{\sigma^2}\right\}^{-\beta},$$

   where Δ characterises any (possible) systematic bias or offset, $\sigma > 0$ characterises the width of the distribution and $\beta > 0$ is a shape parameter, with the distribution having heavy tails for low β and approaching a normal in the limit that $\beta \to \infty$.

   (a) Identify the values of β for which this distribution i) is normaliseable, ii) has a defined mean and iii) has a defined variance. Which, if any, of these criteria are necessary for this distribution to be a valid noise model?

   (b) The above sampling distribution could yield negative measured values. Briefly assess whether this is i) a problem in principle and ii) likely to be important in practice given $L_{1:M}$.

   (c) Assess whether using an improper prior distribution of the form

$$\tilde{\pi}_1(\Delta, \sigma, \beta) = \Theta(\sigma)\,\Theta(\beta - 2)$$

   would yield a valid posterior distribution for Δ, σ and β given $L_{1:M}$. Briefly explain your answer.

   (d) Use the `metrop` function in the `mcmc` package to draw samples from the posterior distribution $P(\Delta, \sigma, \beta|L_{1:M}, L^*)$ given the improper prior probability density

$$\tilde{\pi}_2(\Delta, \sigma, \beta) = \Theta(\sigma)\,\Theta(\beta - 2)\,\beta^{-2}.$$

   Plot i) the joint distribution in Δ and β and ii) the three marginal distributions for the individual parameters. Briefly summarise the results, in particular whether there is evidence for i) systematic bias in the measurements and ii) heavy tails producing outliers.

(16 marks)

4. One possible model is that lung capacity is related to age, weight and the number of cigarettes smoked per day according to

   $$L = \beta_0 + \beta_1 \log(A) + \beta_2 \log(W) + \beta_3 \eta,$$

   where β_0, β_1, β_2 and β_3 are unspecified regression coefficients. These can be constrained by the data in `smoking.csv`.

   (a) Fit this regression model using the `glm()` function assuming normal uncertainties of standard deviation 0.3 litres for all the lung capacity measurements. Use the standard `glm()` diagnostics to assess whether there is evidence that any of these quantities (*i.e.*, A, W and η) influence lung capacity, taking care to be precise in the interpretation of the tail probabilities `glm()` provides.

   (b) Consider a more robust version of the above analysis but now assuming errors drawn from the heavy-tailed distribution introduced in the previous question with $\Delta = 0$, $\sigma = 0.3$ and $\beta = 3$. Explain why this cannot be done using `glm()`. Instead sample from the posterior distribution (assuming uniform priors on $\beta_{0:3}$) using `RStan`. Plot the marginal distributions in β_0, β_1, β_2 and β_3. Compare the results to those obtained using the normal model in part (a) and assess which of the two analyses is likely to be more reliable.

   (c) Describe how you would modify this analysis to include both lung capacity and lung cancer as output quantities. (Point-form answers are fine, as is using psuedo-code or mathematics; this question is deliberately open-ended.)

(20 marks)

(total: 60 marks)