UNIVERSITY OF EXETER

COLLEGE OF ENGINEERING, MATHEMATICS, AND PHYSICAL SCIENCES

MTH3028 Statistical Inference

Summative coursework 2

Date set:	8 March 2021
Hand-in date:	26 March 2021 by noon
Return date:	12 April 2021

This coursework comprises 20% of the overall module assessment. Submit your answers by the date shown above. This is an **individual** exercise and **you must not collaborate when working on this assessment.** Please read the College guidelines on collaboration and plagiarism[1].

Background

Explosions of stars as Type 1A Supernovae are among the most violent events in the universe. We can observe these events in distant galaxies millions of lightyears away. The physics of Supernova events are fairly well understood, and we know how bright they are, and how much light they emit at what wavelengths.

From the apparent luminosity of observed supernovae, their distance from Earth can be estimated. Further, from the Doppler shift of the wavelength spectrum we can estimate the velocity of the supernova relative to our own galaxy.

The file `supernovae_mth3028.csv`, available on the module's ELE page, contains distances and velocities of 36 Type 1A supernovae. Distances are measured in Megaparsecs (Mpc), and velocities are measured in Megaparsecs per billion years (Mpc/Ga), with positive values indicating movement away from Earth.

Denote by $(x_1, \ldots, x_n)$ the distances in Mpc and by $(y_1, \ldots, y_n)$ the corresponding velocities in Mpc/Ga. Consider modelling the velocities as random variables Y_i that are related to the distances by

$$Y_i = \beta x_i + \epsilon_i,$$

where ϵ_i are independent Normal random variables with mean zero and variance e^τ (in other words, τ is the log-variance). That is, Y_i is modelled as a random variable with probability density function

$$f(y_i; \beta, \tau) = (2\pi e^\tau)^{-1/2} e^{-\frac{1}{2}e^{-\tau}(y_i - \beta x_i)^2}$$

The x_i are considered to be fixed covariates, and β and τ are parameters. Further define the quantity $\gamma = 1/\beta$.

Questions

1. Preliminaries

   (a) Load the data, and create a plot of velocity against distance. The plot should include the origin, and should have properly annotated axes.

   (b) For a body that travels at constant velocity v, the relationship between time t and distance d travelled is $v = d/t$. Assume for a moment that the supernova data satisfy $y_i = \beta x_i$ exactly. Explain why the quantity γ can be interpreted as "the time since the big bang", or "the age of the universe".

(6 marks)

[1] `http://intranet.exeter.ac.uk/emps/studentinfo/subjects/mathematics/assessment/academicmisconduct/`

2. Derive the following functions for the assumed statistical model:

   (a) the likelihood function;
   (b) the log-likelihood function;
   (c) the score function;
   (d) the observed information; and
   (e) the expected information.

(8 marks)

3. Estimation of β and τ.

   (a) Derive the maximum likelihood estimators of β and τ, and state the estimated values of β and τ for the supernova data.
   (b) State the asymptotic distribution of $\hat{\beta}$ and $\hat{\tau}$ in the limit of large n, and calculate their approximate standard errors under the asymptotic distribution.

(15 marks)

4. How old is our universe?

   (a) Derive a point estimator $\hat{\gamma}$ for γ, and state your estimate.
   (b) Apply the delta method, using the asymptotic distribution of $\hat{\beta}$, to approximate the variance of $\hat{\gamma}$.
   (c) Answer the question "How old is our universe?", including an indication of the standard error of your estimate.

(8 marks)

5. A commonly quoted value for the age of the universe is 13.5 billion years. Is this value compatible with our data?

   (a) Carry out a Score test at size 0.05 of the simple null hypothesis $H_0 : (\beta, \tau) = (1/13.5, 0)$ against the composite alternative hypothesis $H_1 : (\beta, \tau) \neq (1/13.5, 0)$. State the value of the test statistic, the critical region, and the outcome of the test.
   (b) Carry out a profile likelihood ratio test of the composite null hypothesis $H_0 : \beta = 1/13.5$ against the composite alternative hypothesis $H_1 : \beta \neq 1/13.5$. State the value of the test statistic, the critical region, and the outcome of the test.
   (c) Derive a two-sided 95% confidence interval for β by inverting the profile likelihood ratio test from question 5b.
   (d) Use a Monte-Carlo simulation with parameters $\beta = 1/13.5$ and $\tau = 0$ to check whether your confidence interval for β derived in question 5c has the correct coverage. Report your Monte-Carlo estimate of the coverage frequency.

(28 marks)

6. How robust is our estimate of the age of the universe?

   (a) Simulate a (sufficiently large) number of bootstrap versions of $\hat{\gamma}$, using

      i. parametric bootstrapping
      ii. semi-parametric bootstrapping (error resampling), and
      iii. non-parametric bootstrapping (case resampling).

      Derive a percentile bootstrap 95% confidence interval for γ from each of the 3 bootstrap samples.
   (b) Plot density plots of the 3 bootstrap distributions into one plotting panel, and comment on any similarities or differences in their location and spread.
   (c) Compare your bootstrap estimates and confidence intervals to the currently best estimate of the age of the universe of $13.799 \pm 0.021 Ga$. Suggest an explanation for any differences.

(25 marks)

Submission instructions

Submit your answers to the questions via eBART

`https://bart.exeter.ac.uk`

under the coursework titled "1056670 Set Questions 2".

Use one of the following methods. Independent of which method you use, **you will have to submit a zip archive containing exactly two files**.

Method 1: Prepare your answers using R markdown. Include all written up answers, any equations and proofs, as well as all your R code and figures. Submit both, the uncompiled R markdown file (plain text, file extension `.Rmd`), as well as a compiled version in either pdf or html format.

Method 2: Submit all your answers, equations, and figures as a single pdf document. Submit the R code as one additional source code file (plain text, file extension `.R`) that contains all R code used to answer the questions. Use code comments to indicate clearly what code was used to answer which question, and remove any code that is irrelevant to answer the questions.

For both methods, use the function `set.seed` to initialise the random number generator at the beginning of your R markdown or R source code file, to ensure that your Monte-Carlo simulations and bootstrap runs are reproducible.

Marking criteria

The 90 marks listed above are awarded for the accuracy of your mathematical calculations and computer code, and for your understanding and interpretation. A further 10 marks are awarded for the clarity of your presentation: correct and precise notation, clear and logical arguments, readable, well-commented and concise computer code. There will be deductions for any deviations from the submission methods (e.g. wrong number of files or wrong file format).