Coursework

AUTHOR
MA40198: Applied Statistical Inference

General instructions

Deadline: 5:00 pm Friday 8 December 2023

Submission: Submission is via the coursework submission point on Moodle. A single student per group can submit the group's work. Submissions up to five working days late (without a DoS-approved extension) will receive a maximum mark of 40% and after that will be 0%. Guidelines on the structure and submission of the report appear at the end of this document.

Marking: The coursework is marked out of 40, and the number of marks for each part is stated in the heading of each question. The coursework is worth 40% of the marks for the unit.

Estimated time: This assignment should take roughly the amount of time you would spend on 2 lab sheets.

Support: This is a group assignment, and you should not discuss it with anyone other than your lecturer or the other members of your group. You can ask questions in the dedicated Moodle Forum.

Preliminaries

Consider the following parametric family of probability density functions:

$$\mathcal{F}_1 = \left\{ f(y|\boldsymbol{\lambda}) = \exp(\lambda_1\, y + \lambda_2 \log(\cos(\lambda_1)))\, g(y|\lambda_2) \;:\; -\frac{\pi}{2} < \lambda_1 < \frac{\pi}{2}, \lambda_2 > 0 \right\}$$

The function $g(y|\lambda_2)$ is itself a probability density function given by:

$$g(y|\lambda_2) = \frac{2^{\lambda_2 - 2}\Gamma^2\left(\frac{\lambda_2}{2}\right)}{\pi\Gamma(\lambda_2)} \prod_{j=0}^{\infty} \left\{ 1 + \left(\frac{y}{\lambda_2 + 2j} \right)^2 \right\}^{-1}, \quad y \in \mathbb{R}$$

Here

$$\Gamma(x) = \int_0^{\infty} t^{x-1} \exp(-t) dt\,, \quad x > 0$$

is the standard Gamma function.

For any computations involving $g(y|\lambda_2)$ you should use the following approximation:

$$g_N(y|\lambda_2) = \frac{2^{\lambda_2 - 2}\Gamma^2\left(\frac{\lambda_2}{2}\right)}{\pi\Gamma(\lambda_2)} \prod_{j=0}^{N} \left\{ 1 + \left(\frac{y}{\lambda_2 + 2j} \right)^2 \right\}^{-1}, \quad y \in \mathbb{R}$$

Unless otherwise stated, you should use $N = 10,000$.

Questions

Question 1 [4 marks]

Consider the following observed sample:

▼ Code

```
y_sample_q1 <- scan("http://people.bath.ac.uk/kai21/ASI/CW_2023/y_sample_q1.txt")
```

Draw a contour plot of the negative loglikelihood with 40 levels over the region defined by $-\pi/2 < \lambda_1 < \pi/2$ and $0 < \lambda_2 < 50$. The contours should be sufficiently smooth and cover the entire region. You should indicate a smaller region delimited by a contour that contains the global minimum.

Question 2 [6 marks]

Find the maximum likelihood estimate $\widehat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)^T$ by picking the best out of 100 optimisations (using the BFGS algorithm) where each optimisation uses a different initial value. The following data frame gives the list of initial values to be used.

▼ Code

```
L0 <-read.table("http://people.bath.ac.uk/kai21/ASI/CW_2023/starting_vals_q2.txt")
```

Question 3 [4 marks]

Check the sensitivity of the MLE to the choice of N by plotting (separately) the values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ as function of $\log_{10}(N)$. You should use the values $10^1, 10^2, 10^3, 10^4, 10^5, 10^6$ for N. What conclusions can you make from these two plots?

Question 4 [4 marks]

Compute the maximum likelihood estimate of the mean parameter

$$\mu(\boldsymbol{\lambda}_*) = E[Y|\boldsymbol{\lambda}_*] = \int_{\mathbb{R}} y\, f(y|\boldsymbol{\lambda}_*) dy\,.$$

Also compute an asymptotic 95% confidence interval for $\mu(\boldsymbol{\lambda}_*)$. State clearly any assumptions you have made.

Question 5 [4 marks]

Compute an asymptotic 95% confidence interval for the unknown parameter λ_2^* using:

- the asymptotic normal approximation to the distribution $\hat{\lambda}_2$

- the asymptotic normal approximation to the distribution $\log(\hat{\lambda}_2)$

Question 6 [4 marks]

Use the generalised likelihood ratio to test the hypotheses:

$$H_0 : \mu(\boldsymbol{\lambda}_*) = 5 \qquad \text{vs} \qquad H_a : \mu(\boldsymbol{\lambda}_*) \neq 5$$

using a significance level $\alpha = 0.05$.

Separately, also test

$$H_0 : \lambda_2^* = 5 \qquad \text{vs} \qquad H_a : \lambda_2^* \neq 5$$

using a significance level $\alpha = 0.05$.

Question 7 [10 marks]

Consider the following data frame

▼ Code

```
data_q7 <-read.table("http://people.bath.ac.uk/kai21/ASI/CW_2023/data_q7.txt")
```

that contains a bivariate sample

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

of size $n = 300$.

Use the parametric family $\mathcal{F}_1$ defined in Question 1 to find an appropriate model for the unknown conditional distribution of $\mathcal{Y}$ given $\mathcal{X} = x$, that is $f_*(y|x)$. The model should be defined by specifying the mean function $\mu(\boldsymbol{\theta}^{(1)}, x)$ as follows:

$$\mu(\boldsymbol{\theta}^{(1)}, x) = g^{-1}(\theta_1 + \theta_2\, x + \theta_3\, x^2 + \theta_4\, x^3 + \cdots + \theta_{p+1}\, x^p)$$

for some choice of link function g and some choice of integer $p \geq 1$.

From a set of candidate models (that is for different choices of g and p), choose the model with the smallest AIC (Akaike Information Criterion). Only present the results from the maximum likelihood estimation from the best chosen model and simply comment on the other models considered.

Now, repeat the same process above to find an appropriate model for the unknown conditional distribution of $\mathcal{Y}$ given $\mathcal{X} = x$ but now based on the Gamma parametric family:

$$\mathcal{F}_{gamma} = \left\{ f(y|\lambda_1, \lambda_2) = \frac{\lambda_2^{\lambda_1}}{\Gamma(\lambda_1)} y^{\lambda_1 - 1} \exp(-\lambda_2\, y) \ :\ \lambda_1 > 0\,, \lambda_2 > 0, y > 0 \right\}$$

Finally, find an appropriate model for the unknown conditional distribution of $\mathcal{Y}$ given $\mathcal{X} = x$ but now based on the Normal parametric family:

$$\mathcal{F}_{normal} = \left\{ f(y|\lambda_1, \lambda_2) = \frac{1}{\lambda_2\sqrt{2\pi}} \exp\left(-\frac{(y-\lambda_1)^2}{2\lambda_2^2} \right) \; : \; \lambda_1 \in \mathbb{R},\, \lambda_2 > 0, y \in \mathbb{R} \right\}$$

For each of the three chosen models, you should plot the data together with the maximum likelihood estimate of the mean function as well as corresponding asymptotic 95% confidence bands in the range $x \in (-3, 3)$. Comment on the differences between the confidence bands and the mean function estimates. You must select the best model out of the three, based on the Akaike Information Criterion.

Question 8 [4 marks]

Use the data in Question 7 to compute 95% confidence intervals for the least worse value of the mean function at each x, that is $\mu(\boldsymbol{\theta}_\dagger^{(1)}, x)$ for each of the three parametric families: $\mathcal{F}_1$, the Gamma and the Normal. Plot the computed confidence bands in the range $x \in (-3, 3)$ for each parametric family and comment on the differences obtained.

Submission guidelines

- You should submit only two files: your completed R Markdown file (which should be named `CW23.Rmd`) and the created HTML file (which should be named `CW23.html`) to the coursework submission point in Moodle
- You should complete your answers in the template file `CW23.Rmd` which can be downloaded from the Moodle page. You should open this file in RStudio and you should not change the general structure of the file.
- In the supporting information for your answers, please make clear reference (equation numbers, Theorem numbers, sections, etc) to any unit material such as the lecture notes, question sheets, etc. You can make reference to other external material in which case you should cite the source appropriately.
- The `R` code submitted in the `CW23.Rmd` file, should be reproducible in the sense that your lecturer should be able to reproduce your numerical results, plots etc if necessary. Adding comments to your code is specially useful for this purpose.
- Clearly label the axes of your plots as well as providing them with a title or caption where possible.
- You can only use external `R` packages to verify that the answers of your questions are correct. You will be downmarked (see below) for making calculations using an external R package where you can use the tools developed in the course.
- Your submitted work may be subject to electronic plagiarism screening, unfair use of AI tools and other investigations.

Optimisation guidelines

- Unless otherwise indicated, you should use at least 100 different random starting points for any optimisation performed. This is to make sure you obtain the smallest value possible (for

minimisation) of the objective function. To generate random initial points you should use normal random numbers.

- For each optimisation, you should use the BFGS algorithm and you should provide the gradient function to the algorithm. You can use the R function `optim`. You can use pre-existing R functions (such as `lm`, `glm` or external packages) to check your answers but if you only report the outputs from such functions (where you can use the tools developed in the course) then the maximum mark for the question will be 50% of the original mark.

- You should use appropriate reparametrisations to perform unconstrained optimisation.