MA40198: Coursework sheet

Karim Anaya-Izquierdo

21/11/2019

General instructions

You should work in groups for this assignment, submitting only one report per group. Guidelines on the structure and submission of the report appear at the end of this document and in the unit's Moodle page. The coursework is worth 40% of the mark for the unit.

Breaking strength of fibres

A tensile strength experiment on single carbon fibres produced the following data which gives the stress values (in giga-pascals) at which each single carbon fibres failed. Fibres were of four different lengths: 1 mm, 10 mm, 20 mm and 50 mm. The data can be downloaded as follows:

```
strength.data<-
  read.table(url("http://people.bath.ac.uk/kai21/ASI/CW2019/strength.txt"),header = TRUE)
library(mvtnorm)
```

Mechanical Engineers are interested in establising a relationship between length of fibres and the corresponding breaking strength. The so-called *weakest link hypothesis* establishes the notion that a fibre consists of independent links, whose weakest member determines the breaking of the fibre. Let Y_L be the random variable describing the breaking stress of a fibre of length L with corresponding survival function $S_L(y) = P[Y_L \geq y]$. The weakest link hypothesis can then be expressed as follows:

$$S_L(y) = [S_1(y)]^L$$

One family of distributions that satisfies the weakest link hypothesis is the Weibull family, namely:

$$S_L(y|b,\sigma) = \exp\left(-L\left(\frac{y}{\sigma}\right)^{1/b}\right), \quad y > 0$$

where $b > 0$ and $\sigma > 0$ are shape and scale parameters, respectively. We will denote this model by M_0.

1. Assuming that the strength data provided corresponds to an independent sample from an unknown member of M_0, estimate the unknown true parameter values b^t and σ^t by maximum likelihood. To perform the maximisation, you should transform the parameters to $\boldsymbol{\theta}^T = (\theta_1, \theta_2)$ where

$$\theta_1 = \log(b), \quad \theta_2 = \log(\sigma)$$

   You should use the `R` function `optim` with the version of the BFGS algorithm that requires the gradient function. You should report back:

- The maximum likelihood estimates (mle's) of the unknown parameters $(\theta_1^t, \theta_2^t)^T$. To have reassurance that a maximum has been found, you should simulate random initial points according to `rmvnorm(1,c(-1,1),diag(c(1,1))` until you find 10 identical estimates with the highest possible log-likelihood value.
- The maximum value of the log-likelihood found.
- An estimate of the asymptotic variance-covariance matrix of the mle's.

2. Consider now the extended model

$$S_L(y|b,\sigma,\xi) = \exp\left(-L^{\xi}\,\left(\frac{y}{\sigma}\right)^{1/b}\right)\,, \quad y>0$$

where $\sigma>0$ is a scale parameter and $b>0$ and $\xi\in(0,1)$ are shape parameters. We will denote this model by M_1. Assuming that the strength data provided corresponds to an independent sample from an unknown member of M_1, estimate the unknown parameters b^t, σ^t and ξ^t by maximum likelihood. To perform the maximisation, you should transform the parameters to $\boldsymbol{\theta}^T=(\theta_1,\theta_2,\theta_3)$ where

$$\theta_1=\log(b)\,, \quad \theta_2=\log(\sigma)\,, \quad \theta_3=\log\left(\frac{\xi}{1-\xi}\right)$$

You should use the `R` function `optim` with the version of the BFGS method that requires the gradient function. You should report back:

- The maximum likelihood estimates of the unknown parameters $(\theta_1^t,\theta_2^t,\theta_3^t)^T$. To have reassurance that a maximum has been found, you should simulate random initial points according to `rmvnorm(1,c(-1,1,1),diag(c(1,1,2)))` until you find 10 identical estimates with the highest possible log-likelihood value.
- The maximum value of the log-likelihood.
- An estimate of the asymptotic variance-covariance matrix of the mle's.

Use the Metropolis-Hastings algorithm to sample from the posterior distribution of the parameter $\boldsymbol{\theta}^T=(\theta_1,\theta_2,\theta_3)$ using the following prior specification:

- The marginal prior densities of θ_1 and θ_2 are constant, that is

$$\pi_0(\theta_1)\propto 1\,, \qquad \pi_0(\theta_2)\propto 1$$

- The prior distribution for $\log(\xi)$ is exponential with known mean parameter $\mu_0>0$.
- The parameters θ_1, θ_2 and θ_3 are independent under the prior.

Report on the sensitivity of the posterior distribution when changing the prior parameter μ_0.

Plot a histogram of the posterior samples of the parameter ξ and compute a 95% posterior probability interval for ξ. What conclusions can you draw in terms of the weakest link hypothesis? In this case, we cannot use the generalised likelihood ratio to test the weakest link hypothesis, what conclusion we would have obtained from using it?

3. Consider now the model

$$S_L(y|\eta,\sigma,\xi,\tau) = \exp\left(-L^{\xi}\,\left(\frac{y}{\sigma}\right)^{\eta\,L^{\tau}}\right)\,, \quad y>0$$

where $\sigma>0$ is the scale parameter and $\eta>0$, $\tau\in\mathbb{R}$ and $\xi\in(0,1)$ are shape parameters. We will denote this model as M_2. Assuming that the strength data provided corresponds to an independent sample from an unknown member of the extended Weibull family described above, estimate the unknown parameters η^t, σ^t, ξ^t and τ^t by maximum likelihood. To perform the maximisation, you should transform the parameters to $\boldsymbol{\theta}^T=(\theta_1,\theta_2,\theta_3)$ where

$$\theta_1=\log(\eta)\,, \quad \theta_2=\log(\sigma)\,, \quad \theta_3=\log\left(\frac{\xi}{1-\xi}\right)\,, \quad \theta_4=\tau$$

You should use the `R` function `optim` with the version of the BFGS method that requires the gradient function. You should report back:

- the maximum likelihood estimates of the unknown parameters $(\theta_1^t, \theta_2^t, \theta_3^t, \theta_4^t)^T$. To have reassurance that a maximum has been found, you should simulate random initial points according to `rmvnorm(1,c(1,1,1,0),diag(c(1,1,2,1)))` until you find 10 identical estimates with the highest possible log-likelihood value.
- The maximum value of the log-likelihood.
- An estimate of the asymptotic variance-covariance matrix of the mle's.

Use the Metropolis-Hastings algorithm to sample from the posterior distribution of the parameter $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \theta_3, \theta_4)$ using the following prior specification.

- The marginal prior densities of θ_1 and θ_2 are constant, that is

$$\pi_0(\theta_1) \propto 1\,, \qquad \pi_0(\theta_2) \propto 1$$

- The prior distribution for $\log(\xi)$ is exponential with known mean parameter $\mu_0 > 0$.
- The prior distribution of τ is double-exponential with known scale parameter $\mu_1 > 0$. The corresponding density is given by:

$$\pi_0(\tau) \propto \exp\left(-\frac{|\tau|}{\mu_1}\right)$$

- The parameters θ_1, θ_2, θ_3 and θ_4 are independent under the prior.

Report on the sensitivity of the posterior when changing the prior parameters μ_0 and μ_1.

Using the generated posterior samples, what conclusions can you draw in terms of the weakest link hypothesis?

4. Consider another extended model, called the Burr type XII model, with survival function:

$$S_L(y|b, \sigma, k, \nu) = \left(1 + \frac{1}{k}\left(\frac{y}{\sigma L^\nu}\right)^{1/b}\right)^{-k}$$

where $\sigma > 0$ is a scale parameter and $b > 0$, $k > 0$ and $\nu \in \mathbb{R}$ are shape parameters. We will denote this model by M_3. A slightly more general version of the weakest-link hypothesis is obtained from the above model when $\nu = -b$. In particular, note that when $k \to \infty$ then we obtain the original weakest link Weibull model M_0.

Assuming that the strength data provided corresponds to an independent sample from an unknown member of the Burr model described above, estimate the unknown parameters b^t, σ^t, k^t and ν^t by maximum likelihood. To perform the maximisation, you should transform the parameters to $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \theta_3, \theta_4)$ where

$$\theta_1 = \log(b)\,, \quad \theta_2 = \log(\sigma)\,, \quad \theta_3 = \log(k)\,, \quad \theta_4 = \nu$$

You should use the `R` function `optim` with the version of the BFGS method that requires the gradient function. You should report back:

- The maximum likelihood estimates of the unknown parameters $(\theta_1^t, \theta_2^t, \theta_3^t, \theta_4^t)^T$. To have reassurance that a maximum has been found, you should simulate random initial points according to `rmvnorm(1,c(-2,1,1,0),diag(c(1,1,1,1)))` until you find 10 identical estimates with the highest possible log-likelihood value.
- The maximum value of the log-likelihood.
- An estimate of the asymptotic variance-covariance matrix of the mle's.

Use the generalised likelihood ratio test to test the null hypothesis that $\nu_1 = -b$, that is that the model follows the generalised weakest link hypothesis.

Use the Metropolis-Hastings algorithm to sample from the posterior distribution of the parameter $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \theta_3, \theta_4)$ using the following prior specification.

- The marginal prior densities of θ_1 and θ_2 are constant, that is

$$\pi_0(\theta_1) \propto 1\,, \qquad \pi_0(\theta_2) \propto 1$$

- The prior distribution for $1/k$ is exponential with known mean parameter $\mu_0 > 0$.
- the prior distribution of $\nu + b$ is exponential with known mean parameter $\mu_1 > 0$.
- The parameters θ_1, θ_2, θ_3 and $\nu + b$ are independent under the prior.

Report on the sensitivity of the posterior when changing the prior parameters μ_0 and μ_1.

Using the generated posterior samples, what conclusions can you draw in terms of the original and generalised weakest link hypothesis?

Report on the sensitivity of the posterior when changing the prior parameters μ_0 and μ_1.

5. Use the Akaike information criterion to select the best model from all of the models considered above.

Report

- You should produce a report that discusses the validity of the Weakest Link hypothesis based on the experimental data given.
- Write the report in a way that is suitable for another statistician to read.
- The report should contain all the material requested in the questions in a coherent way.
- Combine written explanations, code (with comments), plots and mathematical formulas, following a style similar to the one used in the solutions to the lab sheets.
- Remember to clearly label the axes of your plots as well as the relevant parts of any table shown.
- The R code submitted should be reproducible in the sense that the lecturer should be able to reproduce your numerical results if necessary. Therefore, you should include all the necessary code in your report so use the space wisely. To aid reproducibility you may want to use the command `set.seed`
- There is a 25 page limit for the report.
- Preferably, use R-markdown in RStudio to produce the pdf file of the report. The R-markdown source file used to produce this document will be available in the Moodle page to give you an idea. If R-markdown is used to produce the report, please also upload it also in the submission as a supplementary file.

Please use the following guidelines for the posterior samples:

- Specify clearly the proposal distributions used.
- Choose suitable starting values for the parameters and determine a suitable burn-in period in which all samples are discarded.
- Tune the proposal distribution iteratively to reach an acceptance rate of about 25%.
- In general, you should give substantial empirical evidence that the Markov chain constructed has converged to its stationary limiting distribution.
- Produce plots to learn about the shape of the marginal posterior densities of the parameters. You should also investigate the correlation between parameters.
- In every case you should aim to obtain effective sample sizes (per parameter) of at least 2000.

Coursework submission

- Your report must be submitted electronically on the MA40198 Applied Statistical Inference Moodle page as a pdf file. There will be a section dedicated to the electronic submission.
- You should submit two files only. These correspond to the source R-markdown file together with the corresponding knitted PDF file.
- Include the names of the members of the group at the beginning of the report.
- Only one group member should submit the work, and check that it has been uploaded properly by the deadline.
- Work may be subject to electronic plagiarism screening, and other investigations to check that it is your own.
- The coursework files must be uploaded by midnight on Friday 13th December 2019.

Marking scheme

A mark between 0 amd 100% will be given. The following criteria only defines marks in the usual class categories.

- **First class** A report without substantive errors, which could be used as the basis for a consultancy meeting with the scientists who gathered the original data. No major omissions. Sensible well backed up conclusions. Work could be repeated easily on basis of what is written. Reasons for choices are explained and reasonable. Clear tables and well labelled plots. Code well structured and commented.
- **2.1** Would require some relatively minor work before being usable as the basis for a consultancy meeting with the scientists who gathered the original data. Some errors or omissions, but basically sound and well put together. Most issues covered well.
- **2.2** Would require substantial revisions before being usable as the basis for a consultancy meeting with the scientists who gathered the original data. Some serious errors or omissions, but some other components of a good standard.
- **3** Not suitable as the basis for a consultancy meeting with the scientists who gathered the original data, without re-doing. Major errors or omissions, but some understanding of the material demonstrated.
- **Fail** Enough major errors that little understanding of the material has been demonstrated in the project work.