Practice Test

Gherardo Varando, gherardo.varando@math.ku.dk

19/12/2018

Formalities

The answers should be submitted as a PDF file. You can either type the mathematical equations in your computer, or write them down by hand and scan/take photos. Make sure that the scanned versions are clear and readable. Combine everything including the text, plotted figures, scanned files/photos, and R code into a single PDF file.

The easier way is to use Rmarkdown and produce the pdf, it will automatically include all the r code and the results, but you should check that the final PDf you submit is correct.

Please identify clearly the problems in the solutions.

In general you can not use other packages that are not in the base R distribution. Exceptions to the above rule are: ggplot2 and tidyr

Data

The data sets to be used in the exam can be downloaded from Absalon. The file is called `exam.RData`. Load the data by opening the file with R, or with the `load` function. There are three data sets in this RData file: `neuron` , `ToothGrowth`, `quakes` and `beerfoam`

Problems

Problem 1 (15 points)

Consider the `ToothGrowth` data set, which describes the effect of vitamin C on tooth growth in Guinea pigs. The length of odontoblasts (cells responsible for tooth growth) were recorded in 60 Guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods: orange juice (OJ) and ascorbic acid (VC). The data frame contains three columns:

- `len`: tooth length

- `supp`: supplement type (VC or OJ)
- `dose`: dose level

Question 1.1 Compute the mean tooth length for all six combinations of supplement types and levels. Also provide the standard error of the mean for each situation.

Question 1.2 We will investigate whether different dose levels have the same effect. Perform 0.05-level two sample t-tests with unequal variances to check whether to reject the following null hypotheses, and explain the result for each hypothesis:

- With the OJ method, the dose levels 0.5 and 1.0 mg/day have the same effect in tooth length.
- With the OJ method, the dose levels 1.0 and 2.0 mg/day have the same effect in tooth length.
- With the VC method, the dose levels 0.5 and 1.0 mg/day have the same effect in tooth length.
- With the VC method, the dose levels 1.0 and 2.0 mg/day have the same effect in tooth length.

Question 1.3 We are interested in whether OJ is more effective than VC. Perform 0.05-level two sample t-tests with unequal variances to check whether to reject the following null hypotheses:

- With 0.5 mg/day dose level, OJ is less effective than or as effective as VC in tooth growth.
- With 1.0 mg/day dose level, OJ is less effective than or as effective as VC in tooth growth.
- With 2.0 mg/day dose level, OJ is less effective than or as effective as VC in tooth growth.

Under which dose level(s) can we say OJ is more effective than VC?

Problem 2 (25 points)

Consider the Weibull distribution with the probability density function (PDF) given by

$$f_{WB}(x|k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \quad x \geq 0$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter.

The PDF of the exponential distribution with rate parameter $r > 0$ is given by

$$f_{Exp}(x|r) = re^{-rx}, \quad x \geq 0$$

Consider a Weibull statistical model, where we observe n i.i.d. observations from a Weibull distribution with shape k and scale λ.

$$X_1, \ldots, X_n \sim Weibull(k, \lambda)$$

In the exponential model instead we consider the n observations following an exponential distribution

$$X_1, \ldots, X_n \sim Exp(R)$$

Question 2.1 Show that when $k = 1$, the Weibull distribution with a parameters $k = 1, \lambda$ reduces to the exponential distribution. What is the rate parameter of the obtained exponential distribution? You can also plot the exponential density and the Weibull density and graphically check that they coincide.

Question 2.2 We want to fit the Weibull model to the `neuron` ISI data. Implement the minus log-likelihood function in R, and find the parameter estimates for the shape and the scale using `optim`. You can use the built-in function dweibull

Question 2.3 Investigate how the Weibull model fits the nueron data by a Q-Q plot and comparing with the kernel density estimation.

Question 2.4 Compute confidence intervals for k and λ using parametric and non-parametric bootstrap, use both normal confidence interval and percentile confidence intervals.

Question 2.5 Fit the exponential distribution to the neuron ISI data. Estimate the rate parameter r by maximum likelihood. Perform model selection using AIC and BIC. In this case to select between Weibull and exponential model can we also use the likelihood ratio test? (justify your answer) If yes, perform the likelihood ratio test.

Problem 3 (40 points)

In this Problem, we will analyze the `quakes` data set, which contains the information of 1000 earthquakes that occurred near Fiji since 1964. It is a 1000×5 data frame with 1000 observations and 5 columns. `lat` , `long` and `depth` are the latitude, the longitude and the depth (km) of the earthquake location, corresponding to the three coordinates in a three-dimensional space. `mag` is the Richter magnitude of the earthquake, and `stations` is the number of stations that detected and reported each earthquake

In the first part of the problem we study the locations of the earthquakes using a Gaussian mixture model.

A Gaussian mixture is a mixture distribution consisting of two Gaussian distributions with a weight parameter. In particular, if a random variable X follows a Gaussian mixture containing two distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ with weight $w \in [0, 1]$ then there is a w probability that X follows $N(\mu_1, \sigma_1^2)$ and $1 - w$ probability that X follows $N(\mu_2, \sigma_2^2)$, we denote the Gaussian mixture by

$$X \sim GM(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w)$$

with five parameters $\mu_1 \in \mathbb{R}, \sigma_1 > 0, \mu_2 \in \mathbb{R}, \sigma_2 > 0$ and $0 \leq w \leq 1$. The probability density function of a Gaussian mixture is given by,

$$f_{GM}(x|\mu_1, \sigma_1, \mu_2, \sigma_2, w) = w f_N(x|\mu_1, \sigma_1) + (1 - w) f_N(\mu_2, \sigma_2)$$

where $f_N(x|\mu, \sigma)$ is the Gaussian PDF of $N(\mu, \sigma^2)$.

Question 3.1 Implement an R function for the PDF of the Gaussian mixture distribution, Plot the PDF of $GM(2, 1, 5, 1, 0.3)$. You can use the built-in `dnorm` function for the PDF of the Gaussian distribution.

Question 3.2 Initially we only look at the longitude data, and we assume that the longitude locations are i.i.d. following a Gaussian mixture. Estimate the five parameters of the Gaussian mixture using the 1000 observed longitude values. You can done this numerically in R with the `optim` function. Plot the fitted Gaussian mixture on top of the histogram of the longitude data.

To find a good initial guess for the parameters you can simply have a look at the histogram of the data and try to guess the location of the means μ_1 and μ_2. An initial guess for w is the proportion of the size of the two data clusters (or use $w = 0.5$ as initial guess). You can also try different initial values and report the results with the smallest minus log-likelihood.

Since there are many parameters the optimization can be long and you need probably to increase the maximum number of iterations of the algorithm otherwise it will exit before it reaches a good optimum. You can do it with `control = list(maxit = 10000)` in the `optim` function. Probably there will be also a lot of warning, mainly because the parameters should be constrained w especially, you can ignore the warnings.

Question 3.3 Consider now another model where the longitude locations are i.i.d. Gaussian distributed $N(\mu, \sigma^2)$. Fit this model to the observed longitude data.

Question 3.4 Compute the AIC and BIC values for the simple Gaussian model and the Gaussian mixture model for the longitude data. Which model should be selected?

Question 3.5 Repeat the above fitting procedure for the latitude and the depth data, and perform as usual model selection using AIC and BIC, which model should be used?

In the following questions we want to make a regression model to describe the number of stations from the three coordinates and the magnitude. Such a model is not immediately useful (mainly because we are using the location and magnitude of earthquakes as predictors), but it may be useful in Bayesian methods for earthquake predictions.

We consider the generalized linear model given by

$$l(\mathbb{E}(\texttt{stations}|\texttt{lat},\texttt{long},\texttt{depth},\texttt{mag})) = g_\beta(\texttt{lat},\texttt{long},\texttt{depth},\texttt{mag})$$

Where l is the link function.

Question 3.6 In this question we consider a generalized linear model with the log link and stations follows a Gaussian distribution.

$$\log(\mathbb{E}(\texttt{stations}|\texttt{lat},\texttt{long},\texttt{depth},\texttt{mag}) = \beta_0+\beta_1\texttt{lat}+\beta_2\texttt{long}+\beta_3\texttt{depth}+\beta_4\texttt{mag} \quad \text{(model 1)}$$

Fit the model using the quake data.

Since it is intuitive that stronger earthquakes are more likely to be detected, we assume that *stations* is more related to `mag`. Fit the following model:

$$\log(\mathbb{E}(\texttt{stations}|\texttt{lat},\texttt{long},\texttt{depth},\texttt{mag})) = \beta_0+\beta_1\texttt{lat}+\beta_2\texttt{long}+\beta_3\texttt{depth}+\beta_4\texttt{mag}+\beta_5\texttt{mag}^2 \quad \text{(model 2)}$$

The recorded magnitude is actually in the Richter scale which is a log scale of the earthquake wave amplitude. We thus transform now the Richter scale back to the original scale. Fit the model:

$$\log(\mathbb{E}(\texttt{stations}|\texttt{lat},\texttt{long},\texttt{depth},\texttt{mag})) = \beta_0+\beta_1\texttt{lat}+\beta_2\texttt{long}+\beta_3\texttt{depth}+\beta_4\exp(\texttt{mag})+\beta_5(\exp(\texttt{mag}))^2 \quad \text{(model 3)}$$

Question 3.7 Perform the log likelihood ratio test selection between model 1 and model 2. Use instead AIC and BIC to perform model selection between model 1, model 2 and model 3.

Question 3.8 We observe now that `stations` are actually positive counts. It is thus natural to use the Poisson regression model. Fit then the Poisson regression models with the log link function:

$$\log(\mathbb{E}(\texttt{stations}|\ldots)) = \beta_0 + \beta_1\texttt{lat} + \beta_2\texttt{long} + \beta_3\texttt{depth} + \beta_4\texttt{mag} \quad \text{(model 4)}$$

$$\log(\mathbb{E}(\texttt{stations}|\ldots)) = \beta_0 + \beta_1\texttt{lat} + \beta_2\texttt{long} + \beta_3\texttt{depth} + \beta_4\texttt{mag} + \beta_6\texttt{mag}^2 \quad \text{(model 5)}$$

$$\log(\mathbb{E}(\texttt{stations}|\ldots)) = \beta_0+\beta_1\texttt{lat}+\beta_2\texttt{long}+\beta_3\texttt{depth}+\beta_4\exp(\texttt{mag})+\beta_5(\exp(\texttt{mag}))^2 \quad \text{(model 6)}$$

Perform model selection between model 4 and model 5 using the `anova` function. Perform model selection between the three Poisson regression models using AIC and BIC.

Question 3.9 Consider the generalized linear regression model with the inverse link function $l(y) = 1/y$:

$$\frac{1}{\mathbb{E}(\texttt{stations}|\dots)} = \beta_0 + \beta_1 \texttt{lat} + \beta_2 \texttt{long} + \beta_3 \texttt{depth} + \beta_4 \texttt{mag} + \beta_6 \texttt{mag}^2$$

Where $\texttt{stations}|\texttt{lat}, \texttt{long}, \texttt{depth}, \texttt{mag}$ follows a gamma distribution.

Fit this model to the `quakes` data. Take a look to the relevant information about the distribution and the link function `?family`.

Problem 4 (20 points)

In this problem we analyze the `beerfoam` data. The data set contains 13 observations of measurements of wet foam height and beer height at various time points for Shiner Bock at 20C. In particular we have three variables:

- `t`: the time from pour (in seconds)
- `foam`: the wet foam height (cm)
- `beer`: the beer height (cm)

We want to estimate how the height of the foam decrease with time.

Question 4.1 Fit the simple linear regression

$$\mathbb{E}(\texttt{foam}|\texttt{t}) = \beta_0 + \beta_1 \texttt{t} \tag{SL}$$

Plot the observations $(\texttt{t}, \texttt{foam})$ and the linear regression obtained. Do you think the model is correct? (hint: plot `t` versus the residuals)

Question 4.2 Fit now the quadratic regression

$$\mathbb{E}(\texttt{foam}|\texttt{t}) = \beta_0 + \beta_1 \texttt{t} + \beta_2 \texttt{t}^2 \tag{Q}$$

Plot the observed point and the fitted quadratic regression, what you can observe? The model seems better than the simple linear regression? Perform the F-test and the model selection using AIC and BIC between the simple linear regression (SL) and the quadratic regression (Q).

What does the quadratic model predict when the time increase (e.g. $\texttt{t} = 600$ seconds) ? Does the prediction of the model make sense? Explain and comment (remember which is the physical experiment we are analyzing).

Question 4.3 Fit now the simple linear regression for the log(`foam`), this can be done since `foam` is always a positive variable.

$$\mathbb{E}(\log(\texttt{foam})|\texttt{t}) = \beta_0 + \beta_1 \texttt{t} \tag{E}$$

(this is a simple linear regression over a transformed variable, it is not a generalized linear regression)

Plot the regression function on top of the observed points and perform model selection using AIC and BIC for all the models (SL, Q and E).

Question 4.4 We want now to investigate the relationship between `foam` and `beer` that is the relationship between the height of the foam and the height of the beer. Try to fit different type of linear regression (and generalized linear regression models) and comment on the results, you can use AIC and BIC to preform model selection. (hint: plot the observations first and try to guess possible models)