Final Assignment

Statistics BI/E, 2019-2020

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

15/01/2020

Formalities

This assignment is the exam assignment for the course Statistics for Bioinformatics and eScience, 2019-2020. The assignment must be completed individually and without any help from others. The exam is available on-line at the digital exam platform `https://eksamen.ku.dk/` from 09.00 Wednesday, 15 of January 2020. The answers must be handed in via the same digital exam platform **before the deadline at 15.00 on Thursday, 16th of January 2020.**

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.

All the R code must be submitted in the pdf file.

Please identify clearly the problems and questions in the solutions.

Data

The data sets to be used in the exam can be downloaded from Absalon. The file is called `exam1920.RData` and a link to this file is placed in the home page of the course. Load the data by opening the file with Rstudio, or with the `load` function. Hint: clean the R session before loading the data and after use the command `ls()` to check the loaded objects. There are four data sets in this RData file: `applejuice` , `applejuice_test`, `obs` and `brainhead`. The data `applejuice`, `applejuice_test` and `brainhead` are data frame, while `obs` is a vector of numerical values.

Problems

Problem 1 (30 points)

In this problem we will study the distribution of the data in `obs` that contains 100 observations from an unknown distribution.

Question 1.1 Plot the histogram of the values and the kernel density estimation (you can try different bandwith and bandwith rules and/or different type of kernel, report the one that you think obtain a good result).

Moreover plot the empirical distribution function and compute and report the first, second and third empirical quantile (Hint: the `quantile` function in R).

Question 1.2 Fit the following models to the data in obs:

- Gaussian distribution.
- Log-normal distribution.
- Gamma distribution.
- Exponential distribution.

Summarize the results in a single plot, showing the histogram of the observations and the fitted densities with different colors, use a legend. Hint: We saw all these models and how to fit them to data during the course, for some of them we know closed form expression of the maximum-likelihood estimators, otherwise numerical method must be used.

Question 1.3 Judge how the Gaussian distribution and the exponential distribution fit the data using Q-Q plots; What can you observe? Do you think the Gaussian distribution is a good candidate model? Why? Do you think the exponential model fit well the data judging from the Q-Q plot?

Question 1.4 Can we use the likelihood-ratio test to compare the Exponential and Gamma model ? Justify your answer and eventually perform the likelihood ratio test, report the obtained p-value and comment the results.

Question 1.5 Use AIC to select among all the models fitted in Question 1.2, report the computed AIC scores and report the selected model.

Question 1.6 Consider now the exponential model for the data in `obs`, build a 95% confidence interval for the rate parameter (λ) and answer the following question:

Is there enough evidence to say that $\lambda \neq 0.5$?

Moreover compute and report the p-value for the corresponding Wald test.

Question 1.7 Obtain 95% confidence intervals for the parameters of the Gamma distribution using non-parametric bootstrap.

Problem 2 (35 points)

We study here the effects of temperature, ph and Nisin (a polycyclic antibacterial peptide used as food preservative) concentration in the growth of Alicyclobacillus Acidoterrestris in apple juice.

The data for this problem are the two data frame objects `applejuice` and `applejuice_test`.

The data frame `applejuice` contains 74 observations (rows) on the presence or absence of growth (after 16 days) of Alicyclobacillus Acidoterrestris under different conditions, in particular the following measurements (columns) are present in the data frame:

- `ph` the ph of the apple juice.
- `nisin` the Nisin concentration.
- `temp` the temperature of the juice.
- `brix` the Brix concentration (the sugar content, an approximation of the dissolved solid contents).
- `growth` a binary variable indicating presence (1) or absence (0) of growth of Alicyclobacillus Acidoterrestris.

The data frame `applejuice_test` contains the results of verification trials; A. Acidoterrestris were incubated in apple juice for 16 days at 30C under fixed Brix concentration and varying ph and Nisin concentration. For each combination of ph and Nisin concentration five samples were observed. The `applejuice_test` data frame contains thus 6 rows corresponding to the different ph-Nisin conditions. The columns in `applejuice_test` correspond to the same variable as in the `applejuice` data frame for all but the `growth` column which is replaced by the `growth_p` variable indicating the proportion of samples where growth was observed.

Question 2.1 Using the data in `applejuice` fit a simple logistic regression model for the variable `growth` using all the other variables as covariates. That is the model:

$$\text{logit}\left(\mathbb{E}(\texttt{growth}|\texttt{ph},\texttt{nisin},\texttt{temp},\texttt{brix})\right) = \beta_0+\beta_1\texttt{ph}+\beta_2\texttt{nisin}+\beta_3\texttt{temp}+\beta_3\texttt{brix}.$$

Report the estimated coefficients, the standard errors and comment on which coefficient is significant at a level 0.01.

Question 2.2 Using the data in `applejuice`, fit the logistic regression model where we add three more terms corresponding to the products between `brix` and the remaining variables:

$$\begin{aligned}\text{logit}\left(\mathbb{E}(\texttt{growth}|\texttt{ph},\texttt{nisin},\texttt{temp},\texttt{brix})\right) = \beta_0 &+ \beta_1\texttt{ph} + \beta_2\texttt{nisin} + \beta_3\texttt{temp} + \beta_4\texttt{brix} \\ &+ \beta_5\texttt{brix}\cdot\texttt{ph} + \beta_6\texttt{brix}\cdot\texttt{temp} \\ &+ \beta_7\texttt{brix}\cdot\texttt{nisin}\end{aligned}$$

Remember to use `I()` in the formula to specify the model in R when writing the product terms such as $\texttt{brix}\cdot\texttt{ph}$ (`I(brix * ph)` in R).

Report the estimated coefficients and the corresponding standard errors.

Question 2.3 Perform model selection using AIC and BIC between the two models obtained in Questions 2.1 and 2.2.

Can we use the likelihood ratio test to select between the two models ? If yes perform the likelihood ratio test and comment the results, otherwise justify your answer.

Question 2.4 Compute 99% confidence intervals for the coefficients in the simple logistic regression model fitted in Question 2.1. (You can use the built-in function `confint`). Then, based on the computed confidence intervals, answer the following questions:

- Can we state, with 99% confidence, that the `ph` coefficient is greater than zero ?
- What can we say about the `brix` coefficient?
- Can we say, that an higher value of `ph` will result in a predicted higher probability of growth of Alicyclobacillus ?

Question 2.5 For both models fitted in Question 2.1 compute the predicted probabilities of growth for the data in `applejuice_test`. Compare the predicted probabilities obtained from the two models with the probabilities estimated (`growth_p` in the data frame `applejuice_test`). Which model performs better in estimating the probability of growth under the conditions in the `applejuice_test` data? Comment the results also with respect to the model selection performed in Question 2.3.

Question 2.6 Use the model obtained in Question 2.2 to estimate the probability of growth of Alicyclobacillus Acidoterrestris as a function of the temperature and the Nisin concentration when the Brix concentration is fixed at 14 and the ph is equal to 5. In particular you can obtain the estimated probabilities of growth for temperatures varying between 20 and 60, and Nisisn concentrations between 0 and 80 (You will thus obtain estimated probabilities for a grid of

temperature and nisin concentration values). Resume the different estimated probabilities in a plot. You can use a contour plot or different colours.

Problem 3 (35 points)

In this problem we study the distribution of the weight of human brain in adults and the relationship with the head size. The data is contained in the data frame `brainhead` which contains the following measurement for 237 subjects:

- `agerange`, 1 if age $\in [20, 46]$, 2 if age > 46
- `headsize`, volume of the head in cubic cm.
- `brainweight`, weight of the brain in grams.

Question 3.1 Check (and comment) if the distribution of the brain weight can be assumed Gaussian, use the tools we have seen in the course (Q-Q plot, histogram).

Question 3.2 Fit now the Gaussian distribution to the brain weight values using all the observations in the data frame `brainhead`. Report the estimated parameters and plot the obtained density on top of the histogram of the data.

Obtain and report a 95% confidence interval for the mean parameter, use the method you think is the most appropriate and comment the choice (you can also use different methods and comment the differences).

Question 3.3 Consider the following question:

> *Is there a significant difference between the mean value of the brain weight for old (age > 46) and young (age ≤ 46) subjects?*

Answer the above question with an appropriate statistical procedure and comment the results.

Question 3.4 Fit the simple linear regression model

$$\mathbb{E}(\texttt{brainweight}|\texttt{headsize}) = \beta_0 + \beta_1 \texttt{headsize},$$

using three different data sets: (1) the full data in `brainhead`, (2) the observations from the young subjects (`agerange` $= 1$) and (3) the observations from the old subjects (`agerange` $= 2$).

Resume the three linear fitted models in a single plot: draw a scatter plot of the observations differentiating with two colors young and old subjects; add then the plot of the linear regressions obtained (use the same colors to differentiate the regressions for old and young subjects).

Question 3.5 Investigate if the linear model with Gaussian noise is appropriate by plotting

- the QQ-plot of the residuals against the normal distribution and
- the residuals versus the `headsize` variable.

Comment the obtained plots.

Question 3.6 Consider the linear regression in Question 4.3 using all the data (you can forget the `agerange` variable here). Obtain an estimation of the mean square error using 10-fold cross validation. That is:

1. split the data in 10 random group of equal size (approximately).
2. For $i = 1, .., 10$ fit a linear regression using all but the data in the i group, and estimate the mean square error over the observations in group i.
3. Average the 10 mean squared error obtained.

Question 3.7 Consider the polynomial regression model

$$\mathbb{E}(\texttt{brainweight}|\texttt{headsize}) = \beta_0 + \beta_1 \texttt{headsize} + \beta_2 \texttt{headsize}^2.$$

Perform the appropriate model selection methods between the above polynomial regression and the simple linear regression, is the simpler model sufficient?