Example Take-home Assignment

StatBI/E

Formalities

- *This is an example of what the questions in a final take-home assignment in the StatBI/E course may be like.*
- The assignment is to be completed individually and without any help from others. Cheating (including working in groups) is forbidden.
- You need to submit one single PDF file with all your solutions to the exam problems. The PDF file must contain all written answers, mathematical derivations, R code, and figures. You may include legible scans/photos, for example, of handwritten derivations (which can be accomplished in Rmd via `![figurecaption](/path/to/image.png)`). Yet, the preferable and easiest way is to prepare your solutions in Rmarkdown to produce your PDF (in the RStudio editor toolbar click to expand the "Knit" dropdown menu and select "Knit to PDF"). Refer to the Exam Knit Test (cf. Absalon) for possible solutions to common knitting problems.
  (As fallback, you may knit to HTML / DOC, open the resulting document, and convert this document to PDF via "print to PDF".)
- Clearly indicate which question you are answering and formulate complete answers with justifications, that is, only stating a code block in the PDF and its output is not sufficient and does not earn points.
  For example, if in Problem 12 you were given a sample of n iid copies of a random variable X and asked to provide a point estimate of its expected value, then the following

  ```
  > mean(data)
  [1] 39.183
  ```

  is insufficient and a better alternative could be

  ```
  Answer to Problem 12
  The empirical average is a point estimator of the expected
  value.  We use R to calculate the empirical average
  of the given sample and print the result below
   > mean(data)
   [1] 39.183
  ```

  (The code in the above example is arguably self-explanatory as it is a single code line calling a single command with a self-explanatory name,

`mean`. Longer code blocks are usually not self-explanatory. Do not rely on us guessing what you are trying to achieve/implement/show in a plot, but instead provide a full self-contained answer with text+code+output. This also enables us to possibly award partial points if the code has a bug resulting in the wrong numerical result/plot, but the remaining answer explains correctly what you intended to implement with your code.)

- If you find a problem with any of the problem statements, please clearly state and describe so in your solution. For example, if you suspect that some aspect of a problem statement may be ambiguous, explicitly state what appears possibly ambiguous to you and what you took the statement to mean so you can proceed with answering the question.
- If you use R packages that are neither part of the base R distribution nor have been introduced during the course for (the solutions of) the weekly exercises, you are responsible for these packages to work correctly.
- Always make sure to double check your PDF before submitting it.

Data

The data sets to be used in the exam can be downloaded from Absalon. The file is called `practice_exam.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 three data sets in this RData file: `radiation` , `bodyfat` and `wind`. The data `radiation` and `bodyfat` are data frame objects while `wind` is a simple vector of values.

Problems

Problem 1 (30 points)

The data set `radiation` contains the result of experiments on the survival of mice exposed to different level of radiation and treated with streptomycin. In particular the data set contains the results of experiments on 522 mice, where three variables are recorded:

- `dose` the neutron dose $(201, 220, 243, 260)$
- `treatment` the treatment (1 = Streptomycin, 0 = saline control)
- `dead` if mouse died in 3-10 days (1 = died, 0 = lived)

Question 1.1 Compute the estimated probability of death for the 8 different combinations of neutron dose and treatment. Plot the estimated probabilities as a function of the neutron dose, in blue for mice treated with Streptomycin and in red for mice treated with the saline control (one single plot with all the probabilities).

Question 1.2 We investigate now if the Streptomycin is effective in mitigate the effect of radiation. First of all we test if the Streptomycin has some relevant effect in the probability of death for the different dose levels of radiation. Perform the chi squared test for independence for the contingency tables of `treatment` and `dead` for different neutron doses. In R you can use the function `table` to build contingency tables, as an example the `treatment-dead` contingency table for `dose`= 201 is obtained with the following code:

```
table(radiation[radiation$dose == 201, c(2, 3)])
```

In particular perform chi squared test for the following null hypothesis and explain the results (we reject at $\alpha = 0.05$?):

- For neutron dose equal to 201, `dead` is independent of `treatment`.
- For neutron dose equal to 220, `dead` is independent of `treatment`.
- For neutron dose equal to 243, `dead` is independent of `treatment`.
- For neutron dose equal to 260, `dead` is independent of `treatment`.

Question 1.3 To compare probability of death for different treatment under the four possible neutron doses we perform now one-sided Wald tests, using the (asymptotically normal) statistic:

$$\delta = \hat{p}_1 - \hat{p}_0$$

where $\hat{p}_1$ is the estimated probability of death for mice treated with Streptomycin and $\hat{p}_0$ for mice treated with saline solution, under the same neutron dose.

In particular test the following null hypothesis, report the p-values and comment if we reject the null hypothesis at $\alpha = 0.05$.

- For neutron dose equal to 201, mice treated with Streptomycin are less or equal probable to survive than mice treated with the saline control.
- For neutron dose equal to 220, mice treated with Streptomycin are less or equal probable to survive than mice treated with the saline control.
- For neutron dose equal to 243, mice treated with Streptomycin are less or equal probable to survive than mice treated with the saline control.
- For neutron dose equal to 260, mice treated with Streptomycin are less or equal probable to survive than mice treated with the saline control.

When we can say that Streptomycin treatment is effective in contrasting radiations effects?

Hint: To perform the Wald test using the statistic $\delta = \hat{p}_1 - \hat{p}_0$ defined above we need to estimate the standard error of δ. Since $\hat{p}_0$ and $\hat{p}_1$ are estimations of the parameters for Bernoulli random variables we can obtain a closed formula for the variance of δ and thus of the standard error. Remember that if $X \sim Bernoulli(p)$ we have that $\mathbb{E}(X) = p$ and $\mathbb{V}(X) = p(1-p)$.

If you are not able to obtain a closed form expression for the standard error of δ you can use bootstrap.

Question 1.4 We now perform logistic regression to predict the probability of death as a function of the neutron dose. Fit, using only the observations from mice treated with Streptomycin, the model

$$\text{logit}(\mathbb{E}(\texttt{dead}|\texttt{dose})) = \beta_0 + \beta_1 \texttt{dose} \qquad \text{(log-regr 1)}$$

where `dead|dose` is a Bernoulli random variable. As in Question 1.1 plot the probability of death predicted by the model (log-regr 1) as a function of the neutron dose. Add, with a different color and in the same plot, the probability estimated from the data (for mice treated with Streptomycin).

Fit, using the data from mice treated with Streptomycin, the other two logistic regressions models, now with polynomial terms,

$$\text{logit}(\mathbb{E}(\texttt{dead}|\texttt{dose})) = \beta_0 + \beta_1 \texttt{dose} + \beta_2 \texttt{dose}^2 \qquad \text{(log-regr 2)}$$

$$\text{logit}(\mathbb{E}(\texttt{dead}|\texttt{dose})) = \beta_0 + \beta_1 \texttt{dose}i + \beta_2 \texttt{dose}^2 + \beta_3 \texttt{dose}^3 \qquad \text{(log-regr 3)}$$

For each of the three models above show the value of the estimated coefficients.

Perform model selection using AIC to choose between the three logistic regression models. Moreover perform the likelihood-ratio test to select between log-regr 1 and log-regr 2.

Problem 2 (30 points)

Percentage of body fat for an individual can be estimated once body density has been determined. Volume, and hence body density, can be accurately measured in a variety of ways. The technique of underwater weighing computes body volume as the difference between body weight measured in air and weight measured during water submersion. Accurate measurement of body fat is thus inconvenient and costly and it is desirable to have easy methods of estimating body fat that are not inconvenient/costly. The idea is to be able to estimate body fat from simple body measurements.

The data set `bodyfat` contains estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men.

The goal is to find a regression model to estimate the percentage of body fat from the body measurements only.

In particular the data set `bodyfat` contains the following variables:

- `fat` the percentage of body fat estimated from underwater body measurement
- `age` the age in years
- `weight` Weight in kg
- `height` Height in m
- `neck` Neck circumference in cm
- `chest` Chest circumference in cm
- `abdomen` Abdomen (2) circumference in cm
- `hip` Hip circumference in cm
- `thigh` Thigh circumference in cm
- `knee` Knee circumference in cm
- `ankle` Ankle circumference in cm
- `biceps` Biceps extended circumference in cm
- `forearm` Forearm circumference in cm
- `wrist` Wrist circumference in cm

Question 2.1 Fit a linear regression model for the variable `fat` (percentage of body fat), using all the other body measurements in the data set as predictor variables. Use the function `summary` to extract informations on the coefficients of the linear regression. Which features seem to be the most relevant to predict the percentage of body fat? Motivate the answer.

Can we reject at $\alpha = 0.05$ that the coefficient for `knee` (knee circumference) is equal to 0 ? Explain and motivate using the information obtained with the `summary` function.

Question 2.2 We want now to obtain simpler regression models for the percentage of fat (`fat`). Perform model selection using both forward and backward stepwise regression using BIC as score. You can use the built-in function `step`.

Question 2.3 A common way to estimate percentage of body fat by body measurements is through the body mass index (BMI).

The BMI is defined as the weight (in kg) of an individual divided by the square of the body height (in m). Compute the body mass index for all the individuals in the data set.

Fit the following linear model to estimate the percentage of body fat (`fat`):

$$\mathbb{E}(\texttt{fat}|\texttt{bmi}, \texttt{age}) = \beta_0 + \beta_1 \texttt{bmi} + \beta_2 \texttt{age} \qquad \text{(BMI model)}$$

where `bmi` is the body mass index. Report the fitted coefficients.

Question 2.4 Compute 95% percentile confidence intervals for the coefficients β_0, β_1 and β_2 in the model of Question 2.3 using non-parametric bootstrap. Compare the obtained confidence intervals with the intervals obtained with the R built-in function `confint`.

Question 2.5 Compare the models obtained in Question 2.2 and the model obtained in Question 2.3 using BIC and errors estimated with leave-one-out cross validation ($\hat{R}_{CV} = \sum_{i=1}^{n}(Y_i - \hat{Y}_{(i)})^2$ where $\hat{Y}_{(i)}$ is the prediction for Y_i obtained by the model fitted omitting the observation (X_i, Y_i)).

Problem 3 (40 points)

In this problem we will study the Gumbel distribution.

The Gumbel distribution is a continuous distribution with density function (PDF) given by the following expression,

$$f_{Gumbel(\mu,\beta)}(x) = f(x|\mu,\beta) = \frac{1}{\beta}\exp\left(-\frac{(x-\mu)}{\beta} - e^{-\frac{(x-\mu)}{\beta}}\right)$$

or alternatively

$$f(x|\mu,\beta) = \frac{1}{\beta}e^{-(z+e^{-z})}$$

where $z = \frac{x-\mu}{\beta}$ and $\mu \in \mathbb{R}$ is the location parameter and $\beta > 0$ the scale parameter.

Moreover if $X \sim Gumbel(\mu,\beta)$ we have expressions for the expected value and the variance:

$$\mathbb{E}(X) = \mu + \beta\gamma \qquad \mathbb{V}(X) = \frac{\pi^2}{6}\beta^2$$

where γ is the Euler-Mascheroni constant and in R can be computed as

$$\gamma = -\texttt{digamma}(1)$$

The approximate value is $\gamma \approx 0.5772$.

The cumulative distribution function (CDF) of the Gumbel distribution is,

$$F_{Gumbel(\mu,\beta)}(x) = \exp\left(-e^{-(x-\mu)/\beta}\right)$$

And the quantile function is given by,

$$Q_{Gumbel(\mu,\beta)}(p) = \mu - \beta\log(-\log(p))$$

All the above formulas and expressions can be used in the solution.

Question 3.1 Implement in R the functions `dgumbel` (PDF), `pgumbel` (CDF), `qgumbel` (quantile function) and `rgumbel` (sampling). This functions should behave and have the signature similar to `dnorm`, `pnorm`, `qnorm`, `rnorm` and the other built-in functions for PDF, CDF, quantile functions end random number generation implemented in R. In particular,

- For `dgumbel`, the R function should work as follow: `dgumbel(x, mu, b)`. Where `x` can be a single numerical value or a vector of numerical values. The function should return a vector with the values of the density of the Gumbel distribution in the points `x`. Optionally you can implement also the additional parameter `gumbel(x, mu, b, log)` where `log` can be `TRUE` or `FALSE` and behaves as in `dnorm`.

- For `pgumbel` The R function should work as `pgumbel(q, mu, b)`, where `q` is the vector of quantiles and `mu,b` are the parameters of the Gumbel distribution. Optionally you can implement the additional parameter and associate behaviour `lower.tail` and `log.p` (check `pnorm` behaviour).

- `qgumbel(p, mu, b)` should return the corresponding vector of quantiles for each vector of probabilities `p`. Optionally add the argument `lower.tail`.

- `rgumbel(n, mu, b)` should return a sample of size `n` distributed as independent Gumbel random variables with parameters μ=`mu` and β =`b`. Remember the inverse transform sampling.

In all functions, the parameters `mu` (the location parameter μ) and `b` (the scale parameter β) should have default values of `mu = 0`, `b = 1`.

To test the above functions generate a sample of size 10000 from a Gumbel distribution using the function `rgumbel` implemented above. Plot the histogram and on top the true density `dgumbel` (you can use the `curve` function for this). The histogram should approximate well the true density. You can use the default values $\mu = 0$ and $\beta = 1$.

How can you check in R that the functions `pgumbel` and `qgumbel` are correct ? Which are the sanity checks that they should pass? Think also on some numerical checks to test if the implemented function `pgumbel` is the cumulative distribution function of `dgumbel` (hint: in R we can perform numerical integration). Comment and explain all the tests and checks you perform.

Question 3.2 Using the formulas for the expected value and the variance of the Gumbel distribution given above, write the method of moments estimators for the parameters μ and β (you have to write down mathematical expressions not R code).

Then apply the method of moments estimators to fit a Gumbel distribution to the observations in the data set `wind` containing the measurements of the maximum of wind speed in different days in the city of St Martin-En-Haut (France). Plot the histogram of the data in `wind` and the estimated Gumbel density corresponding to the method of moments estimators. Judge the estimation with a Q-Q plot.

Question 3.3 Implement in R the minus log-likelihood for the Gumbel model and obtain numerically the maximum-likelihood estimation of the parameters μ and β for the `wind` data set. Use the method of moments parameters as initial parameters for the optimization algorithm, (if you did not solve Question 3.2 try different initial conditions e.g. $\mu = 0, \beta = 1$). Plot the estimated density on top of the histogram and compare it with the method of moments estimates.

Question 3.4 Fit also a Gaussian model to the data in `wind` using maximum likelihood. Check how well the Gaussian model fits the data using the histogram and a Q-Q plot. Do you think the Gaussian model is appropriate? Compare the Gaussian and the Gumbel models for the `wind` data using both AIC and BIC. Can we use likelihood ratio test to compare Gaussian and Gumbel models? If yes compute the p-value otherwise comment the reason we can not perform the test.

Question 3.5 Use non-parametric bootstrap to estimate the standard error of $\hat{\mu}$ and $\hat{\beta}$ (the MLE estimators for the Gumbel distribution) over the `wind` dataset. Compute 95% confidence intervals for the parameters using normal quantiles. Compute also the percentile confidence intervals from the bootstrap samples.

Question 3.6 In this last question we want to perform Bayesian inference on the parameters μ and β of the Gumbel distribution. We choose a Gaussian prior for the parameter μ, in particular

$$\mu \sim N(0, 10)$$

and for the parameter $\beta > 0$ we will use an exponential prior,

$$\beta \sim Exponential(1)$$

Compute using `optim` the MAP estimator (be careful that `optim` perform minimization by default, to perform maximization you should set `control = list(fnscale = -1)` as one of the parameter of `optim`, check `optim` documentation, otherwise you can change the sign of the objective function). Obtain also the posterior expected values of the two parameters μ and β using the Monte Carlo method.