Statistical Genetics and Bioinformatics Coursework 1 (Weight: 20% of the module)

Marina Evangelou

General Instructions

Deadline: 11 March 2022 1pm (UK Time)

Write a report (up to 10 pages, including appendix; smallest font allowed 11pt).

The report should be in portable document format (PDF) and should be uploaded to blackboard (URL: bb.imperial.ac.uk) at the **Coursework 1** folder of the Statistical Genetics and Bioinformatics module page. Once the report is uploaded at blackboard there is no option for re-uploading so you should upload your final version only. Avoid last minute uploads, because the system can crash if it receives too many requests simultaneously.

In addition, do upload your code source files in the **Coursework 1 Source Code** folder of the Statistical Genetics and Bioinformatics module page.

As this is assessed work you need to work on it individually. It must be your own and unaided work. You are not allowed to discuss the assessed coursework with your fellow students or anybody else. All rules regarding academic integrity and plagiarism apply. Violations of this will be treated as an examination offence. In particular, letting somebody else copy your work constitutes an examination offence.

All questions that you may have concerning the coursework must be addressed to the lecturer via **e-mail**. Any resulting clarifications will be communicated to the entire cohort via Blackboard announcements/email. The use of the discussion forum for any questions related to the coursework is not allowed.

Genome-wide association study

The data for this question are available at:

`https://www.ma.ic.ac.uk/~me208/StatisticalGeneticsCoursework1_2022/data_CID.txt`

where you replace CID with your CID number (no zeros needed).

The genome-wide association study (GWAS) presented in the data of the file include the genotypes of Type II Diabetes patients and a corresponding matching set of controls (column y of the file). The individuals have been genotyped on a number of single nucleotide polymorphisms (SNPs) that correspond from columns 3 of the dataset onwards.

1. The researchers of the study have asked you to conduct a Frequentist univariate analysis for identifying the SNPs that are associated with Type II Diabetes. Present in your report the following:
   (a) The data of the study including any characteristics
   (b) Any steps conducted for the analysis of the data
   (c) The statistical model/method implemented; including the null hypothesis tested
   (d) The findings of the conducted analysis
2. Following this, the researchers have asked you to translate the Frequentist analysis into a Bayesian one. Present in your report the following:
   (a) The statistical model/method implemented including the prior distribution of the SNPs and any assumptions made
   (b) The effect of the chosen prior on the results
   (c) The findings of the conducted analysis and how they compare with the findings of the previous part
3. The researchers of the study believe that SNPs might interact with the environmental factor E which is defined as 1 for individuals with BMI greater than 25, and zero otherwise. They have therefore asked you to investigate if there are any SNPs that interact with the environmental factor. Present in your report the following:
   (a) The statistical model/method implemented; including the null hypothesis tested
   (b) The findings of the conducted analysis and how they compare with the findings of first part