GWAS course exam 10.5.2023

Return to Moodle by 16.00 (strict!) on 10-May-2023 preferably as a PDF and your document can be either knitted using Rstudio or compiled using Word or any other word processor. It will be OK as long as it is easy to open and read.

The rules:

1. You are free to use all material from the course notes and exercise solutions (and elsewhere from the web for that matter).
2. You are not allowed to communicate with other course participants during the exam.
3. You are not allowed to ask anyone for help; whether the other person has taken this course or not makes no difference.
4. You are not allowed to help anyone else during the exam.
5. If you have any problems during the exam, send email to matti.pirinen@helsinki.fi.

Problem 1. (6 points)

Read in the genotype-phenotype data from file "exam100523_qt.txt", for example by a command

```
data = read.table("https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/material/exam100523_qt.txt",
           as.is = T, header = T)
```

or by first downloading the file from Moodle to your own computer.

The file has 51 columns and 1000 rows. The first column is a quantitative phenotype `y` and the columns `x1`,...`x50` are genotypes of biallelic SNPs for $n = 1000$ individuals. In this problem, you will only need data on variant `x10` and the phenotype `y`.

(a) Use R to compute a marginal effect size estimate of allele 1 of variant `x10` on the phenotype `y` under the additive model. (1p)

(b) Is this variant `x10` genome-wide significant at a typical GWAS significance threshold? (Justify your answer instead of just giving a yes/no answer.) (1p)

(c) Compute an exact P-value for `x10` in R using the effect size estimate and its SE. (1p)

(d) Give an approximate 95% confidence interval for the effect size of `x10`. (1p)

(e) Compute the allele 1 frequency of `x10`. (1p)

(f) Compute the variance of phenotype `y` explained by variant `x10` in these data. (1p)

Solution.

Problem 2. (6p)

Consider a case-control GWAS. In all questions, justify your answers.

(a) Which of the following two study designs would be more powerful to detect genotype-disease associations:

(1) $S_1 = 13,000$ cases and $R_1 = 17,000$ controls OR
(2) $S_2 = 17,500$ cases and $R_2 = 12,500$ controls?

(b) For which of the following two biallelic SNPs would we have more power to detect disease associations:

(1) allele 1 frequency is 0.93 in the case-control sample OR
(2) allele 1 frequency is 0.96 in the case-control sample?

(c) What is the power to detect an effect of $\beta = 0.15$ on the log-odds scale at significance level of 5e-8 when there are $S = 13,000$ cases and $R = 17,000$ controls and allele 1 frequency is 0.93 in the case-control sample? (2p)

(d) What is the smallest (positive) effect size value β that gives 90% power when the other properties of the study remained the same as in part (c) except that the effect size β can vary? (It is enough to give the answer up to the accuracy of two digits.) (2p)

Problem 3. (6p)

(a) Describe an example of a GWAS study where genetic population structure could confound the GWAS results if the population structure was not accounted for? (Your study need not exist in real life.) (1p)

(b) Explain how you could account for the population structure in your example (a) in such a way that the confounding is largely removed. (1p)

You are studying genetics of biomarker B that you have measured on a continuous scale. Biological sex has a strong association with the average value of B and the leading principal component (PC) of genetic structure is strongly predictive of the levels of B. You want to carry out a GWAS on B that does not produce false positives and has the maximal power.

(c) Explain whether you would use biological sex as a covariate in the GWAS and what was your reasoning? (1p)

(d) Explain whether you would use the leading PC as a covariate in the GWAS and what was your reasoning? (1p)

Prevalence of MS-disease is about 15/10000 and being female is a strong risk factor for MS-disease. Suppose that you have available a case-control sample for MS-disease. The leading principal component (PC) of genetic structure is strongly predictive of the MS status in your sample. You want to carry out a GWAS that does not produce false positives and has the maximal power.

(e) Explain whether you would use biological sex as a covariate in the GWAS and what was your reasoning? (1p)

(f) Explain whether you would use the leading PC as a covariate in the GWAS and what was your reasoning? (1p)

Problem 4. (6p)

File "exam100523_figures.pdf" in Moodle has 3 Figures. Answer the following question based on the figures.

(a) Figure 1 shows QQ-plots of chi-square association test statistics from two GWAS. Explain what you can infer about these two GWAS based on the information visible in the two plots of Figure 1. (2p)

(b) Figure 2 shows plots from LD-score regression applied to two sets of GWAS results. Explain what you can infer about these two GWAS based on the two LD-score regression plots of Figure 2. (2p)

(c) Figure 3 shows 10 samples from Finland and 10 from Sweden analyzed by principal components analysis (PCA) using a genome-wide set of variants. What is a possible source of the problem in this PCA when our goal is to capture the population structure via PCA? (1p) What could you do to overcome the problem when you do not want to completely exclude any samples from further analyses? (1p)

Problem 5. (6p)

Read in the genotype-phenotype data from file "exam100523_qt.txt", for example by a command

```
data = read.table("https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/material/exam100523_qt.txt",
           as.is = T, header = T)
```

or by first downloading the file from Moodle to your own computer.

The file has 51 columns and 1000 rows. The first column is a quantitative phenotype `y` and the columns `x1`,...`x50` are genotypes of biallelic SNPs for $n = 1000$ individuals.

(a) Use R to run marginal association tests between the phenotype `y` and each of the SNPs separately using a linear model. Which variants have P-values below 5e-8? From now on, we will call the variant with the smallest P-value as SNP A. (1p)

(b) Run conditional analyses for the remaining 49 SNPs after you have included the SNP A with the smallest P-value from part (a) in the model as a covariate. What is the top SNP in this conditional analysis? From now on, we will call this variant as SNP B. (1p)

(c) Compute a 95% credible set for the signal represented by SNP B by assuming that there are exactly two causal variants in this region and that the other causal variant is represented by SNP A. Use $\mathcal{N}(0, 1^2)$ as the prior distribution for the causal effect size and assume that every SNP is a priori equally probable to be a causal variant. (2p)

(d) Assume that the SNPs A and B are the only causal variants in this region and their causal effect sizes (for allele 1) are 0.5 for SNP A and -0.25 for SNP B. Compute the scaled marginal effect sizes for all 50 SNPs in the region and plot their absolute value with SNP index on x-axis and the absolute value of the scaled effect size on y-axis. (2p)