BST227 - Homework 3

Due: Wednesday, December 5

Introduction

In the last homework, we plotted summary statistics for LDL using a large cohort. In this homework, we'll perform several regression analyses to understand how we can generate summary statistics. Rather than use the large dataset from the last homework we'll restrict our analysis to chromosome 10. You can find the simulated case-control data in the 'genotype.txt.gz', 'legend.txt', and 'phenotype.txt' files on the course website.

Problem 1

A) Measure the association between each SNP and LDL status without adjusting for any other covariates. Note that Y=1 denotes having a high level of LDL (the bad kind of cholesterol) .

B) Include a qq-plot and Manhattan plot as we did in Homework 2 using the `qqman` package.

C) Compute λ_{GC} as we did in Homework 2. What does the computed value of λ_{GC} say about the degree of inflation for this chromosome?

Problem 2

A) Rerun the analysis now adjusting for self-reported ancestry and sex.

B) Rerun the analysis now adjusting for the first four principal components rather than using self-reported ancestry. Continue to include sex as a covariate in the model.

C) Summarize the results at a per-SNP level using a scatterplot where each point on the scatterplot represents a single SNP. Plot the $-log_{10}(p)$ for each SNP from 2A on the x-axis and the $-log_{10}(p)$ for each SNP from 2B on the y-axis. What would the scatterplot look like if adjusting for self-reported ancestry were equivalent to adjusting for the first four principal components?

Problem 3

Using the summary statistics for the unadjusted associations that you calculated in Problem 1 and either model (your choice) from Problem 2, plot the region +/- 100,000 bp around the top SNP discovered from your adjusted model using Locuszoom. Include the plots in your submission as part of a single document.