Statistical Genetics and Bioinformatics Coursework 1

Kyriacos Xanthos
CID: 01389741

March 11, 2022

Introduction

In this report we consider a genome-wide association study (GWAS) which includes the genotypes of Type II Diabetes patients and a corresponding matching set of controls. This is encoded in the y column of our dataset with the number 0 corresponding to control and the number 1 corresponding to patient. The study also includes and environmental factor variable defined as 1 for individuals with Body Mass Index (BMI) greater than 25 and zero otherwise. In total, we consider 1,980 observations (individuals), 1,029 features (SNPs), variable E which is our environmental factor variable and our feature y which is the binary variable signaling if the individual has the disease or not.

GWAS are considered as hypothesis free studies because they do not make any hypothesis for the genotype SNPs and we can test them for association with the Type II Diabetes disease [1]. SNPs are variations of single nucleotides of the genome. Each SNP represents a difference in a single DNA building block. If alternative forms from one copy of the chromosome to another copy of the same chromosome is found, then these different forms are called alleles. The allele that has the lowest observed frequency in a population is called the minor allele, and the Minor Allele Frequency (MAF) is the frequency at which the minor allele is observed in the selected population [1]. In our dataset, for each SNP we have access to the allele of that SNP for each individual, with 2 representing the minor allele, 1 representing the heterozygous alleles and 0 representing the major allele.

The objective of this study is to identify if there is an association between our available SNPs and Type II Diabetes patients. We conduct a Frequentist univariate analysis aswell as a Bayesian one to compare the results.

1 Frequentist Analysis

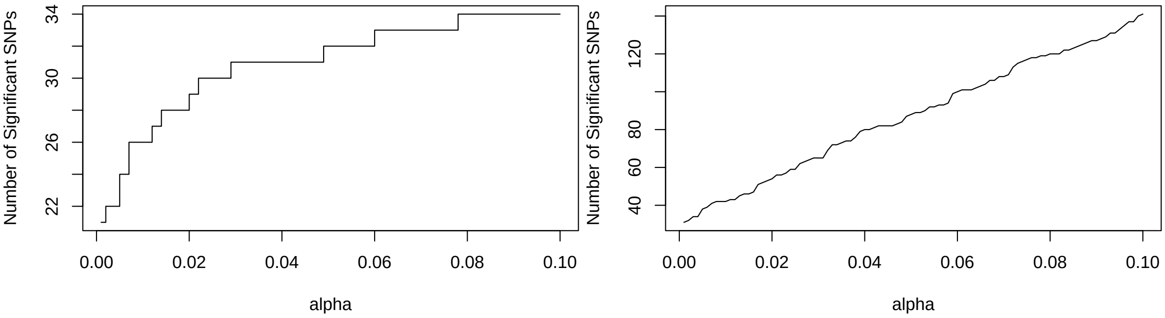
Before beginning our analysis we need to conduct Quality control of our GWA data where we will clean the data in order to keep the number of false-positive and false-negative rates as low as possible. We remove any individuals with missing response y, which accounts to 3 individuals. We only keep individuals that have less than 5% missing genotypes, which removes another 8 individuals. We keep the SNPs with less than 5% missing genotypes which removes 3 SNPs. We also remove any SNPs with MAF $<$ 5% and MAF $>$ 50% (because by definition of MAF that would correspond to the major allele and therefore they were labeled wrong) which equates to 79 SNPs. The last property we check is if the SNPs exibit the Hardy-Weinberg equilibrium. To check this we use a χ^2 test [2] with 1 degree of freedom and use the Benjamini-Hockberg Procedure at level $\alpha = 0.05$ to control the False Discovery Rate (FDR). Running the above steps multiple times to make sure quality control has removed any points that will increase our false-positive and false-negative rate, we are left with 1,971 individuals and 948 SNPs.

We consider the null hypothesis H_0 that SNP$_j$, $j = 1, \ldots 948$, is not associated with Type II Diabetes. We will use the χ^2 test of independence to check the H_0 at a level $\alpha = 0.05$. Since we have Multiple testing occurring, (similar hypothesis are being tested simultaneously), we need to control the FDR using the Benjamini-Hockberg procedure [3]. After adjusting the p-values using this procedure, we find that there are 32 SNPs that appear to be dependent on the target y at the 5% significance level. These SNPs are shown in table 1.

rs1389741442	rs1389741262	rs1389741570	rs1389741261	rs1389741385
rs13897411028	rs1389741738	rs138974113	rs1389741324	rs138974190
rs1389741952	rs1389741374	rs1389741744	rs1389741488	rs1389741532
rs1389741904	rs1389741681	rs138974181	rs1389741890	rs1389741595
rs1389741750	rs1389741350	rs1389741455	rs1389741660	rs1389741830
rs1389741707	rs1389741829	rs1389741339	rs1389741136	rs1389741925
rs1389741903	rs1389741450			

Table 1: SNPs that were identified as significant in testing the Null Hypothesis that the SNPs are not associated with Type II Diabetes at the 5% significance level. They are sorted with the lowest p-value of the test on the top row. All of the p-values were adjusted using the Benjamini-Hockberg procedure.

We can check how sensitive the number of SNPs associated with y is with the significance level α by varying α on a grid from 0.001 to 0.1. We also compare the number of significant SNPs we get without using the Benjamini-Hockberg procedure. This can be seen in figure 1.



(a) Adjusted p-values using Benjamini-Hockberg procedure.

(b) No adjustment on p-values

Figure 1: Results from Frequentist Analysis for number of significant snips against significance levels α.

We can see in 1a that there is a decrease of the number of features as the significance level decreases as expected. The lowest value we tried was 0.001 which picked out 21 significant SNPs. This gives us confidence that the 32 SNPs we found from our analysis at the 5% level is not an unreasonable amount of SNPs. Moreover, we can confirm from figure 1b that the adjustment of p-values was necessary, and there is a clear multiple testing problem if we did not adjust the p-values.

2 Bayesian Analysis

In this part we consider a Bayesian Statistical model to test the same Null Hypothesis as in the last section, namely the hypothesis that there is no association between a SNP and Type II Diabetes. To do this we will use the Posterior Probability of Association (PPA) which is the bayesian analogue of a p-value. PPA is the probability that a single SNP in the GWAS study is associated with the disease. To calculate PPA we conduct the following steps:

1. Choose the value of π, which is the prior probability of the alternative hypothesis H_1, that there is association between the disease and the SNP. We can choose different priors depending on our belief of association for each SNP, or we can use the same prior for all SNPs. If there is strong belief that there is association of the SNP with the disease we could use a π close to 1.

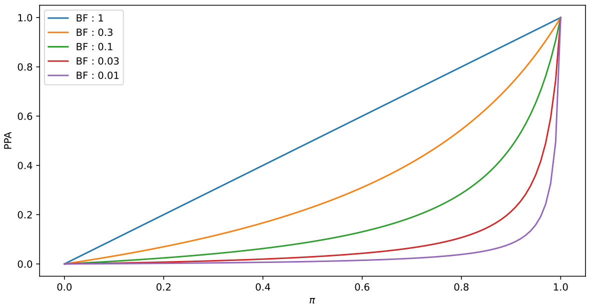
2. Compute the Bayes Factor (BF) for each SNP. BF is defined as a likelihood ratio of the marginal likelihood of the Null and Alternative Hypothesis [4]. The closer this value

is to 1, the higher the probability that the data are equally likely under H_0 and H_1.

3. Calculate the posterior odds (PO) on H_1. This is defined as PO $= \frac{\pi}{1-\pi}$ BF.

4. Then PPA is given by: PPA $= \frac{\text{PO}}{\text{PO}+1}$.

The definition of PPA shows that it is heavily dependent on the prior π we choose. To illustrate this we can explore what the PPA is for different prior probabilities and Bayes factors in figure 2a. Figure 2b shows the interpretation of Bayes factor values in terms of evidence against H_0, which motivated the BF values we checked in 2a.



(a) PPA against prior probabilities for different Bayes Factors

Bayes factor B	Interpretation
$B > 1$	Evidence supports H_0
$1 > B > 10^{-1/2}$	Slight evidence against H_0
$10^{-1/2} > B > 10^{-1}$	Substantial evidence against H_0
$10^{-1} > B > 10^{-3/2}$	Strong evidence against H_0
$10^{-3/2} > B > 10^{-2}$	Very strong evidence against H_0
$10^{-2} > B$	Decisive evidence against H_0

(b) Table for the interpretation of p-values adapted from [5]

Figure 2: Exploration of PPA dependence on prior beliefs.

When sample sizes are large it is very computationally expensive to compute BFs. Instead we approximate BFs using Wakefield's Approximate Bayes Factors (ABF) [6].

To calculate ABF we need to consider first a logistic regression model of the form:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}\left(\pi_i\right), i = 1, \ldots, N \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \beta X_i \end{aligned} \tag{1}$$

where we consider a Normal prior distribution for β with mean zero and variance W. In our model Y_i is the binary variable when the inividial has the disease, X_i is each SNP and π_i the prior probability that the SNP has an association with the disease.

Then, the ABF is defined as:

$$\text{ABF} = \frac{P\left(\hat{\beta} \mid H_0\right)}{P\left(\hat{\beta} \mid H_1\right)} = \sqrt{\frac{V+W}{V}} \exp\left\{-\frac{\hat{\beta}^2}{2}\frac{W}{V(V+W)}\right\} \tag{2}$$

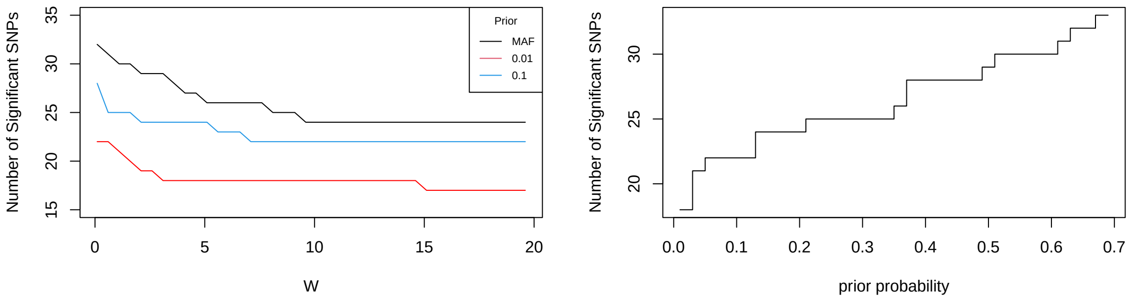
where $\hat{\beta}$ is the MLE of the logistic regression above and V is the variance of $\hat{\beta}$. Note that BFs can be approximated using the inverse of ABF in the definition above.

Now we are ready to build our Bayesian Model. We need to choose the variance W used for the prior distribution of β. There are several ways to specify W which have been proposed by Wakefield, for example the Effect-MAF independence which is the simplest choice and we only have to specify an upper value RR_u above which we believe that the relative risks $\exp\beta$ will occur with low probability. Then we compute W using $W = \left\{\log \text{RR}_\text{u}/\Phi^{-1}(1-q)\right\}^2$ where $\Phi^{-1}(1-q)$ is the 1 - q quantile of a standard normal distribution. We choose a value for RR_u at 2, which was chosen to be a reasonable upper value in the paper by Wakefield and this would give us $W = 0.02$. However, in order to explore how much effect W has in our analysis, we will run the model for different Ws later on.

In deciding the prior π we usually consider the position of a SNP on the DNA, if it is near a gene or not, and use it to compute different π for each tested SNP [1]. However, since we do not have access to that information now, we will use the MAF as the prior for each SNP. This means that the prior will be changing for each SNP and it will never be more than 0.5 (by definition of MAF). Since PPA is analogous to p-values we will use again the

$\alpha = 0.05$ significance level for our analysis. Using MAF as the prior, we find that 28 SNPs are significant. In fact, 24 of these SNPs were also found in our frequentist analysis.

We will first explore how the choice of variance W affects the amount of significant features we get from this Bayesian analysis. We consider a grid of W [0.01, 20] and check the number of significant SNPs we find. This can be shown in figure 3a. Clearly there is a change between the number of significant SNPs for different Ws. This number ranges between 17 - 32. This is not a very wide range given that we are testing in the same plot three different priors π, two of each are the same for every SNP. The value of W used with Wakeffield's Effect-MAF independence seems small given that there are quite big jumps for all priors between for Ws between 0 and 5, but since it is the only theory-based choice, we will stick with 0.2. In figure 3b we can see the sensitivity of the number of significant SNPs with the prior probability, and it is clear that there is a steady increase. This is because large prior probability values assume that there is a larger probability of SNP association with the disease. The MAF estimate that ranges between 0 and 0.5 seems like a reasonable prior to use, since we want to be conservative for the association of SNPs with disease.



(a) Number of Significant SNPs at the 5 % significance level against prior variance W for the normal distribution of β as defined in 1

(b) Number of Significant SNPs at the 5 % significance level against prior probability π for the prior belief of association of SNP with the disease.

Figure 3: Exploration of the variability of number of snips with W and π.

The 24 SNPs that were found in both Frequentist and Bayesian analysis are found in table 2.

rs138974113	rs138974181	rs138974190	rs1389741261	rs1389741262
rs1389741324	rs1389741350	rs1389741374	rs1389741385	rs1389741442
rs1389741488	rs1389741532	rs1389741570	rs1389741595	rs1389741660
rs1389741681	rs1389741707	rs1389741738	rs1389741744	rs1389741750
rs1389741890	rs1389741904	rs1389741952	rs13897411028	

Table 2: The intersection of SNPs that were identified as significant in testing the Null Hypothesis that the SNPs are not associated with Type II Diabetes at the 5% significance level from both Frequentist and Bayesian procedures. For the Bayesian procedure PPA was used as analogue for p-values with prior probability $\pi =$ MAF and $W = 0.02$.

3 Environmental Factors

SNPs might interact with environmental factors and an extension to our above models would be to include this. Specifically we have access to the binary variable for individuals having a BMI more than 25 taking the value one, and zero otherwise. To investigate if the SNPs interact with the environmental factor and assess whether that affects the SNPs that are flagged as significant for their association of Type II Diabetes, we will use a logistic regression model.

Logistic Regression models are more flexible than χ^2 tests in defining the genetic model that we would like to test, and we could easily use it for analysis in section 1, but we wanted to explore as many models possible so we explored both.

This model will be defined exactly like equation 1 with the addition of the variable E as a multiplication with X_i. This creates an interaction term and we can assess the effect of this factor from the coefficient of this new variable in our regression, EX_i. The null hypothesis we will be testing is that there is no association of SNP with the disease, given the environmental factor E.

We can therefore extract the p-value from each of the SNPs that is associated with the significance of this interaction term at the 5% significance level. Again, we have a multiple testing problem here so we adjust our p-values using the Bonferroni procedure [7] which controls the familywise error rate. This is an alternative to the Benjamini-Hockberg Procedure which controls the false discovery rate, and we use it here because we want to try different procedures and make sure we get similar results.

Running this logistic regression model we find that only 3 SNPs seem to be associated with the disease. Comparing with our frequentist analysis we see that 2 of those were included as associated and 1 of those was associated with the Bayesian method aswell. Namely, we see that SNPs rs1389741903 and rs1389741952 appeared in both the general frequentist method and the one which included the interaction term, and rs1389741952 was also found in the Bayesian analysis.

It is interesting that rs1389741952 was not among the 10 smallest p-values seen in table 1 for the frequentist approach but was the only one that appeared in all three models. This eludes the power of using different models on the same study and how the confidence in our results can increase when we try different approaches from different schools of statistical philosophy.

In conclusion, we are confident that rs1389741952 is a SNP that is associated with Type II Diabetes when including the BMI factor. In addition, we found 27 more SNPs shown in table 2 that were significant at the 5% level from both Frequentist and Bayesian approaches and they should therefore be studied further for their association with Type II Diabetes.

References

[1] Marina Evangelou. Statistical Genetics and Bioinformatics Lecture Notes. *Imperial College London*, pages 30–100, March 2022.

[2] Wikipedia. Chi-squared test — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Chi-squared%20test&oldid=1072395297, 2022. [Online; accessed 10-March-2022].

[3] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

[4] James Walker Hardin. *Strengths and Limitations of Some Miscellaneous Statistical Procedures*, chapter 7, pages 119–137. John Wiley Sons, Ltd, 2012. ISBN 9781118360125. doi: https://doi.org/10.1002/9781118360125.ch7. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118360125.ch7.

[5] Young G. A. Fundamentals of Statistical Inference Lecture Notes. *Imperial College London*, page 56, March 2022.

[6] Jon Wakefield. Commentary: Genome-wide significance thresholds via Bayes factors. *International Journal of Epidemiology*, 41(1):286–291, 02 2012. ISSN 0300-5771. doi: 10.1093/ije/dyr241. URL https://doi.org/10.1093/ije/dyr241.

[7] M. E. Dewey and E. Seneta. *Carlo Emilio Bonferroni*, pages 411–414. Springer New York, New York, NY, 2001. ISBN 978-1-4613-0179-0. doi: 10.1007/978-1-4613-0179-0_88. URL https://doi.org/10.1007/978-1-4613-0179-0_88.