REPORT

TITLE : IDENTIFYING SNP ASSOCIATED WITH THE RISK

INTRODUCTION:

Genome-wide association studies also known as GWAS is analysis of genome-wide set of genetic variants is associated with a trait. It focuses on association between single nucleotide polymorphisms and traits like major human disease.

Here, our objective is to identify a single nucleotide polymorphism associated with a disease. The given dataset had information about 100 individuals (assignment.5.csv), genetic ancestry information , a disease susceptibility score and allele for SNP1, SNP2, SNP3, SNP4, SNP5. Each SNP was recorded as either 0 or 1

Based on the previous Research it was said that the phenotype under investigation is monogenic which means that only one SNP is associated with the risk score and it was mentioned that the phenotype followed a recessive mode of inheritance which means only the presence or absence of risk allele is considered, So genotype is not basically considered for this analysis. The most important thing here is that genetic ancestry had to be treated as confounding factor for the analysis.

- Loaded the dataset and inspected SNPs, risk-score and ensured that ancestry is treated as a factor and made sure that data is consistent.
- Found that ANOVA is well suited for this task because it is specifically designed to test whether the means of risk score are significantly different by including the ancestry. The output like F-statistic and p-values, clearly indicates whether the SNPs have statistically significant effect on the risk-score.
- Applying Bonferroni correction and adjusting the p-values made sure that it reduces the likelihood of identifying false positives and finally visually using boxplot to illustrate the relationship.

This Approach ensured that robust identification of SNP associated with the disease susceptibility by minimising confounding effects.

METHADOLOGY:

1) Loading the libraries and inspecting the data.
    - dplyr , data.table and ggplot2 were the libraries.
    - The dataset assignment.5.csv was loaded.
    - Variables like SNPs, individual, ancestry, risk-score were confirmed with the help of summary.
    - Variables like ancestry and SNP1, SNP2, SNP3, SNP4, and SNP5 were converted into factor for statistical analysis.

2) Data Analysis
    - Calculated mean and standard deviation using data-table.
    - SNP frequencies were calculated which counted occurrence of allele for each SNPs which tells about the frequency of allele (0 or 1) for each of the five SNPs.

3) Statistical Analysis
    - Performed the one-way ANOVA for each of the SNP with the model:

      Risk score ~ SNP + Ancestry
    - Adjusted the p-value and upon that for verification I performed Bonferroni correction to check for any false values.

4) Visualization
    - Boxplot was created to visualize the risk-score associated between SNP2 and allele group.

5) Linear regression
    - Finally, linear regression model was used to find out the effect size of SNP2 and ancestry on the risk.

RESULTS: INTERPRETATION AND ANALYSIS

1) Dataset interpretation:
    - The dataset (assignment.5.csv) had 100 observation each representing individual sample.
    - There are 8 variables, individual, ancestry, risk-score, SNP1, SNP2, SNP3, SNP4, SNP5.
    - There are 2 ancestry groups.
    - Lowest risk-score is -1.814 and highest score is 6.742.
    - Mean is 2.852, which indicates the average risk-score in moderately low.
    - Each SNP is recorded as 0 or 1, indicating whether the allele is present or not.

2) Summary Statistics:

ANCESTRY	MEAN RISK SCORE	STANDARD DEVIATION
1	2.634	1.819
2	3.151	1.642

    - Ancestry group 1 has an average risk score of 2.634 and ancestry group 2 has an average risk score of 3.151 which is higher.
    - Standard deviation measures the spread of risk scores with each ancestry group and we can see that group1 has larger Standard

deviation and it will suggest that risk scores are more diverged within group 1.

3) SNP frequency

SNP	FREQUENCY
SNP1	61
SNP2	51
SNP3	60
SNP4	60
SNP5	61

- The above SNP-frequencies table talks about the frequency of the allele (0 or 1) for each of the five SNPs. Over 100 samples from the data we can say that for SNP1, SNP3, SNP4, SNP5 the most common allele is present in 61% of the individuals while for SNP2 it is present in 51% of the individuals. There is a slight difference in the frequency for SNP2.

4) ANOVA

Variable	SNP1	SNP2	SNP3	SNP4	SNP5
SNP	0.192	4.44e-11	0.174	0.3686	0.57670
ancestry	0.522	0.00384	0.617	0.0318	0.00538

Above is the p-value result after the ANOVA test and below are the observation.

- SNP1 : Here p value is 0.192, which is greater than the significant threshold (0.05). This indicates SNP1 has no effect on risk-score.

- SNP2 : p-value is 4.44e-11, which is highly significant ($p < 0.001$). This conforms that SNP2 has high effect on the risk score.

- SNP3 : p-value is 0.1739, it is greater than significant threshold (0.05), SNP3 has no effect on the risk-score.

- SNP4 : The p-value is 0.3686, which is greater than 0.05, SNP4 has no effect on the risk-scores.

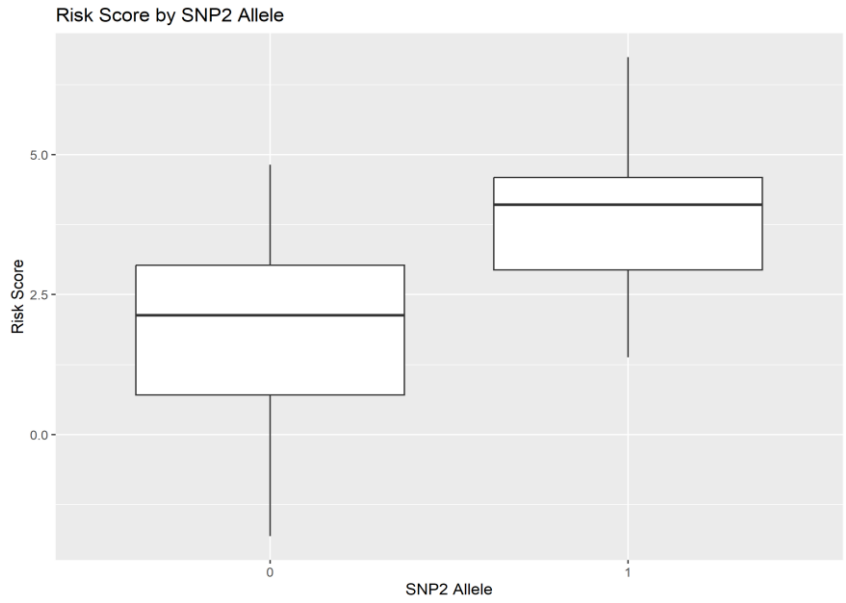
- SNP5 : The p-value is 0.5767, which is higher than 0.05, it does not significantly affect the risk score.

5) Adjusted P-values:

Variable	Raw p-value	Adjusted p-value	Significance (Adj)
SNP1	0.192	0.9586	Not Significant
SNP2	4.44e-11	2.22e-10	Significant
SNP3	0.174	0.8699	Not Significant
SNP4	0.3686	1.0000	Not Significant
SNP5	0.5767	1.0000	Not Significant

- After adjusting the p-value and additionally validating by Bonferroni correction, the SNP2 remains significant with the value of 2.22e-10 which conforms that SNP2 is associated with the risk score.
- Other SNPs do not show any effect on risk-score even after correction as their p-value is greater than 0.05.

6) VISUALIZATION:



- In, the above plot, X-axis represents the different allele of SNP2 loaded as 0 and 1, Y-axis represents the risk score.
- There is a median line inside the box and line extending from box shows the range of risk scores.
- From the above plot, we can interpret that SNP2 (0) has large box than SNP2 (1) which suggests that SNP2 (0) has greater variability and SNP2 (1) has less variability.
- We can also observe that the median line for SNP2 (0) is slightly above the midpoint which suggest that there are more individuals with risk scores that are below the median.

- SNP2(1) has median line significantly above the midpoint of the box meaning that majority of the risk score in this group is higher.
- The larger box for SNP2 (0) suggests that larger IQR and SNP2 (1) it is more compressed which indicates that less variation in risk among individuals with this allele.
- SNP2 (1) shows higher and more consistent risk scores and SNP2 (0) shows that individual with this allele have a great diversity in terms of risk.

7) Linear Regression model:

- Here, Residual Standard Error is 1.378 which is lower and this indicates the models accuracy in predicting the risk-scores.
- R-squared value is 0.3846 which accounts for inclusion of multiple predictors. This conforms the effectiveness of the model.
- F-statistic is 31.93 which suggest that model is statistically significant.
- P-value id 2.216e -11 which is lower than 0.05 which indicates that model significantly explains the variance in risk-score.

CONCLUSION:

In this study, We found that SNP2 is the genetic variant associated with the disease. SNP2 has p-value of 4.44e-11 which remains highly significant even after Bonferroni correction. With the help of data analysis we also found that genetic ancestry play major role in influencing the risk-score.

Later, the visual analysis showed variation in risk-score where SNP2 (0) showing more variability compared to SNP2 (1). Finally, with the help of linear regression model we could conform the statistical significance of SNP2 and ancestry in predicting the risk score. Overall, the analysis helped us to understand that SNP2 is the reason for the disease.