**LIFE752 Assessment 2 Part 2 of 3**

**GAI declaration**

I did use generative ai in the production of this report, specifically, I used ChatGPT 3.5 for the interpretation of error messages and to aid understanding for deeper annotation of code.

**Disclaimer: All code was run in Ubuntu.**

**Question 1 (12 points)**

a.

(base) ricefarmer02@LAPTOP-LCHBMJ9L:~$ cd /mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset # Changing the directory

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ plink --bfile blood\_pressure --missing --out output/blood\_pressure\_summary # Using PLINK to calculate statistics for the blood pressure data (PLINK, 1.9)

PLINK v1.90b7 64-bit (16 Jan 2023) www.cog-genomics.org/plink/1.9/

(C) 2005-2023 Shaun Purcell, Christopher Chang GNU General Public License v3

Logging to output/blood\_pressure\_summary.log.

Options in effect:

--bfile blood\_pressure

--missing

--out output/blood\_pressure\_summary

5752 MB RAM detected; reserving 2876 MB for main workspace.

50255 variants loaded from .bim file.

1989 people (975 males, 1014 females) loaded from .fam.

1989 phenotype values loaded from .fam.

Using 1 thread (no multithreaded calculations invoked).

Before main variant filters, 1989 founders and 0 nonfounders present.

Calculating allele frequencies... done.

Total genotyping rate is 0.989996.

--missing: Sample missing data report written to

output/blood\_pressure\_summary.imiss, and variant-based missing data report

written to output/blood\_pressure\_summary.lmiss.

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ awk 'END {print NR}' blood\_pressure.bim > output/snp\_count.txt #Counting the number of genetic variants in the dataset (SNPs)

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ awk 'END {print NR}' blood\_pressure.fam > output/sample\_count.txt # Counting the total number of samples

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ awk '{print $5}' blood\_pressure.fam | sort | uniq -c > output/sex\_count.txt # Counting the number of males and females

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$

plink --bfile blood\_pressure --missing --out output/blood\_pressure\_missing # Calculating genotype data per sample using PLINK for detailed analysis

PLINK v1.90b7 64-bit (16 Jan 2023) www.cog-genomics.org/plink/1.9/

(C) 2005-2023 Shaun Purcell, Christopher Chang GNU General Public License v3

Logging to output/blood\_pressure\_missing.log.

Options in effect:

--bfile blood\_pressure

--missing

--out output/blood\_pressure\_missing

5752 MB RAM detected; reserving 2876 MB for main workspace.

50255 variants loaded from .bim file.

1989 people (975 males, 1014 females) loaded from .fam.

1989 phenotype values loaded from .fam.

Using 1 thread (no multithreaded calculations invoked).

Before main variant filters, 1989 founders and 0 nonfounders present.

Calculating allele frequencies... done.

Total genotyping rate is 0.989996.

--missing: Sample missing data report written to

output/blood\_pressure\_missing.imiss, and variant-based missing data report

written to output/blood\_pressure\_missing.lmiss.

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ awk '$6 > 0.05' output/blood\_pressure\_missing.imiss | wc -l > output/failed\_samples.txt # Counting the number of samples with more than 5% missing genotypes

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$

awk 'NR > 1' output/blood\_pressure\_missing.imiss | wc -l >> output/failed\_samples.txt # Counting samples assessed for genotype quality

b. 50255

c. 1989

d. 975/1989 = 0.49 (49%)

e. 1/1989 = 0.0005 (0.05%)

**Question 2 *[17 points total]***

a. (base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ plink --bfile blood\_pressure \ # Using PLINK for LD clumping on specific criteria

--clump blood\_pressure\_base.txt \ # Initiates LD clumping and defines file

--clump-p1 0.9 \ # Keeping SNPs with p-value<0.09

--clump-r2 0.3 \ # Removing SNPs with high linkage disequilibrium (R2>0.3)

--clump-kb 350 \ # Considering SNPs within 350 kb of the index SNP for clumping capacity

--clump-snp-field SNP \ # Specifying the SNP column to be used as identifiers for SNPs

--clump-field P \ # Specifying the P-value is used for filtering

--out output/blood\_pressure\_clumped # Defines output

awk 'NR > 1 {print $3}' output/blood\_pressure\_clumped.clumped > output/blood\_pressure\_index\_snps # Extracting index SNPs after LD clumping

# OUTPUT:

PLINK v1.90b7 64-bit (16 Jan 2023) www.cog-genomics.org/plink/1.9/

(C) 2005-2023 Shaun Purcell, Christopher Chang GNU General Public License v3

Logging to output/blood\_pressure\_clumped.log.

Options in effect:

--bfile blood\_pressure

--clump blood\_pressure\_base.txt

--clump-field P

--clump-kb 350

--clump-p1 0.9

--clump-r2 0.3

--clump-snp-field SNP

--out output/blood\_pressure\_clumped

5752 MB RAM detected; reserving 2876 MB for main workspace.

50255 variants loaded from .bim file.

1989 people (975 males, 1014 females) loaded from .fam.

1989 phenotype values loaded from .fam.

Using 1 thread (no multithreaded calculations invoked).

Before main variant filters, 1989 founders and 0 nonfounders present.

Calculating allele frequencies... done.

Total genotyping rate is 0.989996.

50255 variants and 1989 people pass filters and QC.

Among remaining phenotypes, 996 are cases and 993 are controls.

b. 16,263

c. awk 'NR > 1 {print $3}' output/blood\_pressure\_clumped.clumped > output/blood\_pressure\_index\_snps

d. Clumping helps PRS calculation by only including SNPs that contribute unique information. If two SNPs have high linkage disequilibrium (LD), they provide redundant information and would overestimate that genetic region (Chasman et al., 2020).

**Question 3  *[24 points total]***

1. Write code to build polygenic risk scores for blood pressure using the target dataset and information from the base dataset. Build scores including SNPs with p-values equal to or less than the following thresholds: i) 1; ii)0.8; iii)0.6; iv)0.4, v) 0.2, vi) 0.05; vii) 0.01. Copy and paste your code and annotate it. [8 points]
2. Looking at the polygenic risk score which included p-values of 0.05 or less, what is the highest score for an individual? Copy and paste the code you used to work this out and annotate it, and provide the highest score [5 points]
3. Looking at the polygenic risk score which included p-values of 1 or less, what is the lowest score for an individual ? Copy and paste the code you used to work this out and annotate it, and provide the lowest score [5 points]
4. Explain in no more than 30 words why researchers often filter based on a p-value threshold when developing a polygenic risk score [6 points]

Question 3

a.

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ awk 'NR > 1 {print $2, $4, $7, $6}' blood\_pressure\_base.txt > output/blood\_pressure\_base\_corrected.txt # Extracting required columns (SNP ID, effect allele, effect size, p-value)

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ awk '{print $1, $4}' output/blood\_pressure\_base\_corrected.txt > output/blood\_pressure\_snp\_pvals.txt # Creating a file just containing SNP IDs and p values

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$ plink --bfile blood\_pressure \ # Calculating PRS with PLINK

--score output/blood\_pressure\_base\_corrected.txt 1 2 3 \ # Defines the columns (1: SNP ID, 2: effect allele, 3: effect size

--extract output/blood\_pressure\_index\_snps \ # Only considers SNPs retained after LD clumping

--q-score-range output/p\_thresh.txt output/blood\_pressure\_snp\_pvals.txt \ # Uses SNP p-value file to create thresholds

--out output/blood\_pressure\_prs # Output file

# Ubuntu output

PLINK v1.90b7 64-bit (16 Jan 2023) www.cog-genomics.org/plink/1.9/

(C) 2005-2023 Shaun Purcell, Christopher Chang GNU General Public License v3

Logging to output/blood\_pressure\_prs.log.

Options in effect:

--bfile blood\_pressure

--extract output/blood\_pressure\_index\_snps

--out output/blood\_pressure\_prs

--q-score-range output/p\_thresh.txt output/blood\_pressure\_snp\_pvals.txt

--score output/blood\_pressure\_base\_corrected.txt 1 2 3

5752 MB RAM detected; reserving 2876 MB for main workspace.

50255 variants loaded from .bim file.

1989 people (975 males, 1014 females) loaded from .fam.

1989 phenotype values loaded from .fam.

--extract: 16263 variants remaining.

Using 1 thread (no multithreaded calculations invoked).

Before main variant filters, 1989 founders and 0 nonfounders present.

Calculating allele frequencies... done.

Total genotyping rate is 0.989964.

16263 variants and 1989 people pass filters and QC.

Among remaining phenotypes, 996 are cases and 993 are controls.

Warning: 33992 lines skipped in --score file (33992 due to variant ID mismatch,

0 due to allele code mismatch); see output/blood\_pressure\_prs.nopred for

details.

--score: 16263 valid predictors loaded.

Warning: 33992 lines skipped in --q-score-range data file.

--score: 8 ranges processed.

Results written to output/blood\_pressure\_prs.\*.profile.

(base) ricefarmer02@LAPTOP-LCHBMJ9L:/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset$

b. prs\_data <- read.table("output/blood\_pressure\_prs.range3.profile", header = TRUE) # Loading the PRS data for range 3 (p ≤ 0.05)

max(prs\_data$SCORE) # Identifying the highest PRS score

**ANSWER =** 0.0280242

c.

prs\_data <- read.table("output/blood\_pressure\_prs.range8.profile", header = TRUE)# Loading the PRS data for range 8 (p ≤ 1)

min(prs\_data$SCORE) # Identifying the lowest PRS score

**ANSWER**=-0.010944

d. P-value thresholds help filter SNPs by their association strength with the train, enhancing the PRS predictive performance by removing irrelevant SNPs.

**Question 4 *[20 points total]***

1. Write some code to explore how well each of the polygenic risk scores that you developed predicts phenotype of blood pressure in the target dataset.  Copy and paste the code – you do not need to annotate this one.[4 points]
2. Prepare a summary table to show how well each of the polygenic risk scores predict the phenotype of blood pressure. Ensure the table includes a legend explaining what the table is showing [5 points]
3. Present the data in your table graphically. Make sure you provide a legend for your graph to explain what it is summarising [8 points]
4. Based on the results of using the polygenic risk scores to predict blood pressure in the target dataset, which score do you think is best ? Why ?[3 points]

Question 4

a.

library(ggplot2) # Loading libraries

p.thresholds <- c("range1", "range2", "range3", "range4",

"range5", "range6", "range7", "range8") # Defining p-value thresholds so we change them to numeric e.g. ≤ 0.05

prs.results <- data.frame( # Initialising empty result data frame for:

Threshold = character(), # p-value threshold range

R2 = numeric(), # R2 (predictive power)

P = numeric(), # p-value for PRS model

BETA = numeric(), # Effect size (regression coefficient)

SE = numeric(), # Standard error of effect size

stringsAsFactors = FALSE) # Make sure strings aren’t factors

phenotype <- read.table("../blood\_pressure.fam", header = FALSE) # Loading the phenotype data (blood pressure measurements)

colnames(phenotype) <- c("FID", "IID", "FatherID", "MotherID", "Sex", "Phenotype")# Setting the column names for clarity, e.g. FID = Family ID

for (p in p.thresholds) {# Looping through each threshold PRS file

prs <- read.table(paste0("blood\_pressure\_prs.", p, ".profile"), header = TRUE) # Load the PRS file for current threshold

prs.data <- merge(phenotype, prs[, c("FID", "IID", "SCORE")], by = c("FID", "IID")) # Merging PRS with phenotype data

model <- lm(Phenotype ~ SCORE, data = prs.data) # Performing linear regression (PRS ~ phenotype)

summary.model <- summary(model) # Summarising model

# Extracting model statistics

r2 <- summary.model$r.squared # R2 (predictive power)

p\_value <- coef(summary.model)["SCORE", "Pr(>|t|)"] # p-value of PRS

beta <- coef(summary.model)["SCORE", "Estimate"] # Regression coefficiet

se <- coef(summary.model)["SCORE", "Std. Error"] # Standard error

prs.results <- rbind(prs.results, # Storing the results as a data frame for future use

data.frame(Threshold = p, R2 = r2, P = p\_value,

BETA = beta, SE = se))

}

print(prs.results) # Viewing the results

1. # b

print(prs.results # View the summary table

write.csv(prs.results, "prs\_results\_summary.csv", row.names = FALSE) ) # Saving table as CSV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Threshold** | **R2** | **P** | **BETA** | **SE** |
| P ≤ 0.001 | 0.002975 | 0.014974 | 0.666076 | 0.273528 |
| P ≤ 0.01 | 0.001989 | 0.046736 | 1.325576 | 0.666144 |
| P ≤ 0.05 | 0.002351 | 0.030584 | 2.515925 | 1.162633 |
| P ≤ 0.1 | 0.001622 | 0.07249 | 2.711666 | 1.509011 |
| P ≤ 0.2 | 0.00156 | 0.078209 | 3.528617 | 2.002528 |
| P ≤ 0.3 | 0.001236 | 0.117076 | 3.872733 | 2.470096 |
| P ≤ 0.4 | 0.001327 | 0.104366 | 4.734445 | 2.913879 |
| P ≤ 0.5 | 0.001306 | 0.107072 | 5.395585 | 3.346667 |

1. ggplot(prs.results, aes(x = Threshold, y = R2)) + # Visualising PRS performance (R² vs threshold)

geom\_bar(stat = "identity", fill = "skyblue") + # Barplot with custom colour

theme\_minimal() + # Minimal theme

labs(title = "PRS Performance by P-Value Threshold", # Creating a title for the figure

x = "P-Value Threshold", # x-axis label

y = "R² (Predictive Power)") + # y-axis label

geom\_text(aes(label = round(R2, 4)), # Rounding to 4 decimal places

vjust = -0.5, size = 3.5) # Adjusting size

**A graph of a performance

AI-generated content may be incorrect.**

**Figure x. Predictive performance of polygenic risk scores (PRS) across different p-value thresholds.** This bargraph visualises the predictive performance (R2) using SNPs filtered at different thresholds. Higher R2 values indicate more effective predictive performance.

1. The best PRS is calculated using p ≤ 0.001, as it has the highest R2 value, suggesting the strongest statistical association with blood pressure.

**Question 5** *[10 points total]*

1. Have a look at the predict() function in R. Use this function as well as a plotting function of your choice to create a plot to help visualise how similar the blood pressure predicted using the best polygenic risk score is to the actual blood pressure measurements. Copy and paste the code you have used, annotate it, and include the plot you have created below. [8 points]
2. Using the best polygenic risk score that you chose in question (4) above, what is your predicted blood pressure for the 10th sample in your dataset ? Give your answer and also copy and paste the code you used. [2 points]

a.

library(ggplot2) # Loading libraries

setwd("/mnt/c/Users/samri/Documents/101\_bioinformatics/COMPUTATIONAL/EXAM2/PART2/assignment\_dataset/output") # Set directory

phenotype <- read.table("../blood\_pressure.fam", header = FALSE) # Re-loading the phenotype data

colnames(phenotype) <- c("FID", "IID", "FatherID", "MotherID", "Sex", "BloodPressure") # Set column names

prs\_data <- read.table("blood\_pressure\_prs.range1.profile", header = TRUE) # Loading the best identified PRS (P ≤ 0.001)

prs\_data <- merge(phenotype, prs\_data[, c("FID", "IID", "SCORE")], by = c("FID", "IID")) # Merging the PRS with phenotype data again (reloaded script)

prs\_data$SCORE <- scale(prs\_data$SCORE) # Standardising PRS scores (mean = 0, standard deviation = 1)

model <- glm(BloodPressure ~ SCORE, data = prs\_data, family = binomial()) # Logistic regression with a binary outcome

prs\_data$Predicted\_Prob <- predict(model, prs\_data, type = "response") # Predicting probabilities

ggplot(prs\_data, aes(x = as.factor(BloodPressure), y = Predicted\_Prob)) + # Plotting actual vs predicted probability (x = control vs case, y = predicted probability of high blood pressure from the logistic model)

geom\_boxplot(fill = "skyblue") + # Setting colour

geom\_jitter(width = 0.2, color = "blue", alpha = 0.4) + # Adding individual points, adjusting alpha for semi-transparent points with better visibility

theme\_minimal() + # setting minimal theme

labs(title = "Predicted Probability of Blood Pressure (Case) by PRS (P ≤ 0.001)", # Setting title

x = "Actual Blood Pressure (0 = Control, 1 = Case)", # x axis labels

y = "Predicted Probability (Logistic Model)") + # y axis labels

theme(plot.title = element\_text(hjust = 0.5)) # Adjusting title

A graph showing a number of blue squares

AI-generated content may be incorrect.

b.

predicted\_prob\_10 <- prs\_data$Predicted\_Prob[10] # Extracting the predicted probability for the 10th sample

Predicted\_prob\_10

ANSWER: Predicted probability of blood pressure (case) for the 10th Sample: 0.5093462

**Question 6** *[7 points total]*

1. No, the logistic model does not definitively separate the cases from controls, and the PRS has limited predictive power (R2 < 0.003).
2. The predictive ability could be improved by incorporating a larger sample size (GWAS), non-genetic metadata factors or using more sophisticated models, that include metadata (Choi and O’Reilly, 2019)
3. Cross-validation or bootstrapping can perform internal validation to assess the model’s robustness. Additionally, testing the model in an independent population can ensure generalisability through external validation (Osterman et al., 2024).

**Question 7 [*10 points total*]**

1. The frequency of SNPs can differ between populations, potentially causing the effect sizes to be inaccurate in the target population, as they were derived from the base population. Additionally, populations can have different LD structure, meaning the SNPs used in the PRS may not identify the causal variants as well as in the target population. Finally, the environmental and lifestyle factors influencing blood pressure may differ between populations, impacting the prediction capabilities of the genetic variants (Wray et al., 2013).
2. Heritability describes the proportion of variance within a population in a trait explained by genetic factors. Contextualised to PRS, moderate to high heritability (≥ 20-30%) is generally recommended to validate that the genetic factors have sufficient predictive power.
3. The area under the receiver characteristic curve (AUC or AUROC) is commonly used to quantify how well the model distinguishes between case (trait-positive) and control (trait-negative).

**Reference list**

Chasman, D.I., Giulianini, F., Demler, O.V. and Udler, M.S. (2020). Pleiotropy-Based Decomposition of Genetic Risk Scores: Association and Interaction Analysis for Type 2 Diabetes and CAD. *The American Journal of Human Genetics*, 106(5), pp.646–658. doi:https://doi.org/10.1016/j.ajhg.2020.03.011.

Choi, S.W. and O’Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. *GigaScience*, 8(7). doi:https://doi.org/10.1093/gigascience/giz082.

LINK, P. (n.d.). *PLINK 1.9*. [online] www.cog-genomics.org. Available at: https://www.cog-genomics.org/plink/.

Osterman, M.D., Song, Y.E., Lynn, A., Miskimen, K., Wheeler, N.R., Bartlett, J., Farrer, L.A., Schellenberg, G.D., Martin, E.R., Pericak-Vance, M.A., Bush, W.S. and Haines, J.L. (2024). Examining the Performance of Polygenic Risk Scores for Alzheimer Disease Within and Across Populations Using *k* -Fold Cross-Validation. *Neurology Genetics*, 10(6). doi:https://doi.org/10.1212/nxg.0000000000200198.

Wray, N.R., Yang, J., Hayes, B.J., Price, A.L., Goddard, M.E. and Visscher, P.M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Reviews. Genetics*, [online] 14(7), pp.507–515. doi:https://doi.org/10.1038/nrg3457.