P8139 HW4

Brian Jo Hsuan Lee

4/18/2022

Set chunk parameters

```
knitr::opts_chunk$set(
  fig.width = 6,
  fig.asp = .6,
  out.width = "90%"
)
```

Load packages

```
library(tidyverse)
library(genetics)
library(knitr)
```

Load data

```
data = read.csv("FMS_data.csv") %>%
  janitor::clean_names()
attach(data)
```

a, b) **Association Analysis and Significance**

We are interested in determining whether there is an association between any SNP within the resistin gene and a person's height at baseline according to the FAMuSS study.

```
# Find all 6 SNP sites on the resistin gene and create a genotype matrix with them
names_resistin_snps = names(data)[substr(names(data), 1, 4) == "resi"]; names_resistin_snps
```

```
## [1] "resistin_c30t"  "resistin_c398t" "resistin_g540a" "resistin_c980g"
## [5] "resistin_c180g" "resistin_a537c"
```

```
data_resistin = data[,is.element(names(data), names_resistin_snps)]

# Set a categorical and a continuous height trait to be responses:
# the first trait is true if the sample is over 6' (= 72"),
# the second trait is the sample's height in inches
trait_1 = as.numeric(pre_height>=72)
trait_2 = pre_height
```

Use a χ^2 test to examine if there exists any SNP in the resistin gene that are associated with the sample being over 6' at baseline.

```
# Run a chi-squared test on the c30t SNP site
chisq.test(table(trait_1, data_resistin$resistin_c30t))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  table(trait_1, data_resistin$resistin_c30t)
## X-squared = 0.039213, df = 1, p-value = 0.843
```

```
# Define a function that runs a chi-squared test on the 2 × 6 contingency table corresponding
# to each SNP and the categorical height trait, and then extracts their p-values
pval_cat_func = function(geno){
  res = chisq.test(table(trait_1, geno))
  return(res$p.value)
}

# Print the corresponding p-values of associations for each SNP
over72_pval = apply(data_resistin, 2, pval_cat_func)
col_names = gsub("resistin_", "", names(over72_pval))
over72_pval_mat =
  over72_pval%>%
  data.matrix() %>%
  t()
kable(over72_pval_mat, col.names = col_names, "simple")
```

c30t	c398t	g540a	c980g	c180g	a537c
0.8430263	0.0327967	0.1388105	0.0700069	0.2276695	0.8836756

```
# Print the corresponding BH-corrected p-values of associations for each SNP
over72_adj_pval = p.adjust(over72_pval, method="BH")
over72_adj_pval_mat =
  over72_adj_pval%>%
  data.matrix() %>%
  t()
kable(over72_adj_pval_mat, col.names = col_names, "simple")
```

c30t	c398t	g540a	c980g	c180g	a537c
0.8836756	0.1967802	0.2776211	0.2100207	0.3415042	0.8836756

Use an ANOVA test to examine whether any of the mean heights associated with a SNP is different from the overall mean height at the SNP site. It does so by checking the variance of a SNP against the overall variance.

```
# Run an ANOVA test on the c30t SNP site
aov(trait_2 ~ data_resistin$resistin_c30t)
```

```
## Call:
##    aov(formula = trait_2 ~ data_resistin$resistin_c30t)
##
## Terms:
##                 data_resistin$resistin_c30t Residuals
## Sum of Squares                         1.76   7727.37
## Deg. of Freedom                           1       612
##
## Residual standard error: 3.553368
## Estimated effects may be unbalanced
## 783 observations deleted due to missingness
```

```r
# Define a function that performs an anova test on each SNP site, then extracts the p-values
pval_cont_func = function(geno){
  res = aov(trait_2 ~ geno)
  pval = summary(res)[[1]][["Pr(>F)"]][1]
  return(pval)
}

# Print the corresponding p-values of associations for each SNP
height_pval = apply(data_resistin, 2, pval_cont_func)
height_pval_mat =
  height_pval%>%
  data.matrix() %>%
  t()
kable(height_pval_mat, col.names = col_names, "simple")
```

c30t	c398t	g540a	c980g	c180g	a537c
0.7089844	0.0128936	0.1413522	0.4220756	0.2416083	0.3494802

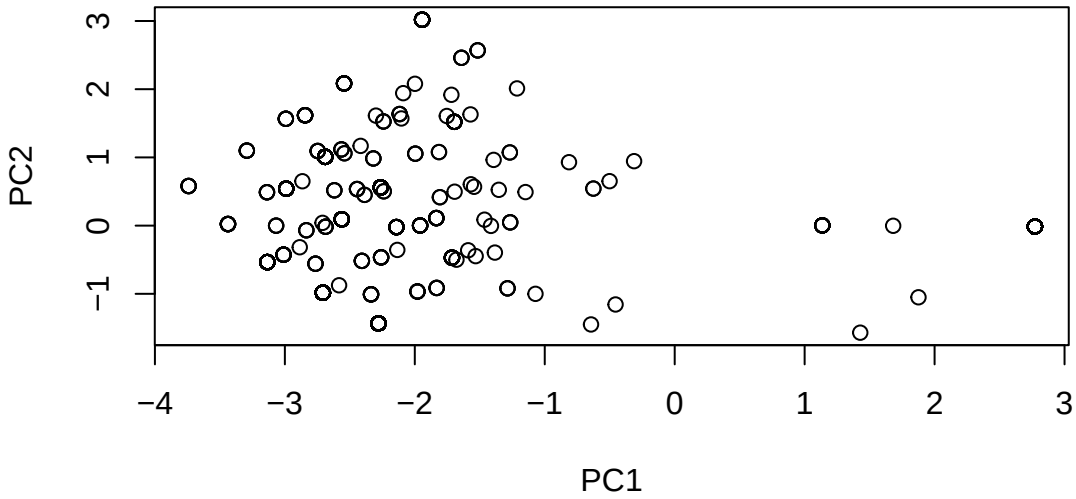
```r
# Print the corresponding BH-corrected p-values of associations for each SNP
height_adj_pval = p.adjust(height_pval, method="BH")
height_adj_pval_mat =
  height_adj_pval%>%
  data.matrix() %>%
  t()
kable(height_adj_pval_mat, col.names = col_names, "simple")
```

c30t	c398t	g540a	c980g	c180g	a537c
0.7089844	0.0773614	0.4240567	0.5064907	0.4832166	0.5064907

c) **Investigate Population Stratification via PCA**

```r
# Convert the genotype data from factor variables to numeric variables
# and additionally assign the missing data a number
data_resistin_mat =
  data_resistin %>%
  mutate(across(everything(), ~replace(., is.na(.) , "NA"))) %>%
  data.matrix()

# Apply and plot PCA
data_pc = prcomp(data_resistin_mat)
plot(data_pc$"x"[,1], data_pc$"x"[,2], xlab = "PC1", ylab = "PC2")
```




```
# calculate variance for each PC, get total variance, and divide each by total variance
# to acquire percentage variance
pc_vars =
  data.frame(data_pc$x) %>%
  summarize(PC1 = var(PC1),
            PC2 = var(PC2),
            PC3 = var(PC3),
            PC4 = var(PC4),
            PC5 = var(PC5),
            PC6 = var(PC6)) %>%
  mutate(
    tot = sum(PC1, PC2, PC3, PC4, PC5, PC6)
  ) %>%
  summarize(PC1 = scales::percent(PC1/tot, 0.01),
            PC2 = scales::percent(PC2/tot, 0.01),
            PC3 = scales::percent(PC3/tot, 0.01))

kable(pc_vars, "simple")
```

PC1	PC2	PC3
86.56%	7.78%	2.64%

d) **Summary**

When population stratification is not accounted for, no significant associations were found between a SNP at an annotated site on the resistin gene and the sample's height. While some p-values generated by the Pearson's χ^2 or the F test statistics were lower then the $\alpha = 0.05$ level, our FDR adjusted p-values demonstrated that the difference in height could still be subjected to randomness. For instance, the SNPs at c398t had a 0.033 p-value when associated with being over or under 6' (72") at baseline, and a 0.013 p-value when associated with height as a continuous variable, but those values increased to 0.20 and 0.077, respectively, after the correction. Note that our correction approach is more lenient than some other multiple comparison adjustment methods such as Bonferroni's FWER.

And indeed, the sub-populations are detectable under PCA. Our analysis suggests that our first principle component alone accounts for over 86% of the variance in our sample data, and the first 2 PCs explain close to 95% of the total variance. Categorical or continuous, performing an association test with these 2 PCs as the predictors would be the more appropriate approach instead of treating each SNP site independently.