Practical Lecture 6: investigate differences in prediction accuracy across 3 samples

Loic Yengo

2022-06-17

Preparation

Create a directory where to run the practical

```
mkdir prac6/
```

and sub-directory where to create PLINK files

```
mkdir prac6/PLINKdata/
```

Run the practical script to load all required R objects.

```
source("/data/module4/prac6/prac6.R")
```

Question 1. Calculate the prediction accuracy of each PGS in each sample. What can you conclude?

```
## [TIP] PLINK command to calculate PGS:
## plink --bfile myData0 --score GWAS.txt 1 2 5 sum --out myPGS0
Rsq0 <- cor(y0,PGS0)^2
Rsq1 <- cor(y1,PGS1)^2
Rsq2 <- cor(y2,PGS2)^2
```

Display results

```
cat(paste0("\tPopulation 0: Rsq(PGS0) = ",round(mean(Rsq0),3),".\n"))
cat(paste0("\tPopulation 1: Rsq(PGS1) = ",round(mean(Rsq1),3),".\n"))
cat(paste0("\tPopulation 2: Rsq(PGS2) = ",round(mean(Rsq2),3),".\n\n"))
```

Question 2. Do you see significant differences in PGS distributions (mean and variance) across these samples?

Visualize the distribution of PGS across the 3 samples

```
plot(density(PGS0),main="",ylim=c(0,1))
lines(density(PGS1),col=2)
lines(density(PGS2),col=3)
```

Display mean and standard deviation of PGS

```
cat(paste0("\tPopulation 0: Mean(PGS) = ",round(mean(PGS0),3)," - SD(PGS) = ",round(sd(PGS0),3),".\n"))
cat(paste0("\tPopulation 1: Mean(PGS) = ",round(mean(PGS1),3)," - SD(PGS) = ",round(sd(PGS1),3),".\n"))
cat(paste0("\tPopulation 2: Mean(PGS) = ",round(mean(PGS2),3)," - SD(PGS) = ",round(sd(PGS2),3),".\n"))
```

What could explain those differences? Are they statistically significant?

```
t.test(PGS0,PGS1)
t.test(PGS0,PGS2)
t.test(PGS1,PGS2)
```

Question 3. Compare allele frequencies across these samples with that in the GWAS. Which dataset is genetically closer to GWAS participants? Does this confirm your intuition from **Question 2**?

```
## [TIP] PLINK command to calculate allele frequencies:
## plink --bfile myData0 --freq --out AF0
freq0 <- colMeans(x0)/2
freq1 <- colMeans(x1)/2
freq2 <- colMeans(x2)/2
```

Visualize frequencies differences

```
op <- par(mfrow=c(3,2))
plot(GWAS[,"FREQA1"],freq0,pch=19);hist(GWAS[,"FREQA1"]-freq0,main="Frequency Difference")
plot(GWAS[,"FREQA1"],freq1,pch=19);hist(GWAS[,"FREQA1"]-freq1,main="Frequency Difference")
plot(GWAS[,"FREQA1"],freq2,pch=19);hist(GWAS[,"FREQA1"]-freq2,main="Frequency Difference")
par(op)
```

Question 4. Estimate SNP effects in each dataset.

```
## [TIP] PLINK command to estimate SNP effect (simple GWAS)
## plink --bfile myData0 --pheno myData0.phen --linear --out GWAS0
SNPeffect0 <- cov(y0,x0)[1,]/apply(x0,2,var)
SNPeffect1 <- cov(y1,x1)[1,]/apply(x1,2,var)
SNPeffect2 <- cov(y2,x2)[1,]/apply(x2,2,var)
```

Regress SNP effect in each sample onto that of the reference GWAS. What can you conclude on the source of differences in prediction accuracy across these samples?

```
summary( lm(SNPeffect0~GWAS[,"BETA"]) )
summary( lm(SNPeffect1~GWAS[,"BETA"]) )
summary( lm(SNPeffect2~GWAS[,"BETA"]) )
```