WORKING EXAMPLE 4

In: Association Analysis Between Polygenic Risk Scores and Traits: Practical Guidelines and Tutorial with an Illustrative Data Set of Schizophrenia

Itziar Irigoien, Patricia Mas-Bermejo, Sergi Papiol, Neus Barrantes-Vidal,
Araceli Rosa, and Concepción Arenas

Working flow and code

In this example we simulate 10 PRSs, and a binary trait, with sex, clinical diagnosis (with 2 categories), age, and two Principal Components as covariates.

- data reading

```
dat <- read.table("WExample4.csv", header=TRUE, sep=";", dec=",")
names(dat) #
```

```
##  [1] "Sex"        "Diagnostic" "Age"        "Trait"      "PRS.1"
##  [6] "PRS.2"      "PRS.3"      "PRS.4"      "PRS.5"      "PRS.6"
## [11] "PRS.7"      "PRS.8"      "PRS.9"      "PRS.10"     "PC1"
## [16] "PC2"
```

- do not forget to declare the categorical variables as factors

```
dat$Sex <- as.factor(dat$Sex)
dat$Diagnostic <- as.factor(dat$Diagnostic)
dat$Trait <- as.factor(dat$Trait)
```

1. What full model should be considered?

First, given a particular PRS (named PRS.i), consider all the possible full models:

- FM_{WI}: log(p/1-p) versus PRS.i + Sex + Diagnostic + Age + PC1 + PC2
- FM_{Sex}: log(p/1-p) versus PRS.i + Sex + PRS.i $\cdot$ Sex + Diagnostic + Age + PC1 + PC2
- $FM_{Diagnostic}$: log(p/1-p) versus PRS.i + Sex + Diagnostic + PRS.i $\cdot$ Diagnostic + Age + PC1 + PC2
- $FM_{Sex/Diagnostic}$: log(p/1-p) versus PRS.i + Sex + PRS.i $\cdot$ Sex + Diagnostic + PRS.i $\cdot$ Diagnostic + Age + PC1 + PC2

2. How to make a PRS ranking to find the important ones?

As is described in the paper, for each model, calculate the Tjur's coefficients of discrimination. If Nagelkerke's R^2 is prefered, set statistic="PseudoR2" in function orderBin(), and calculate their sum, S.

According to S, list the PRSs in decreasing order:

```
# Order the PRSs
out <- orderBin(dat, yname="Trait", prsname = "PRS.", statistic = "D") # Note that
# this function is included in the customized file via source("Functions.R")
head(out)
```

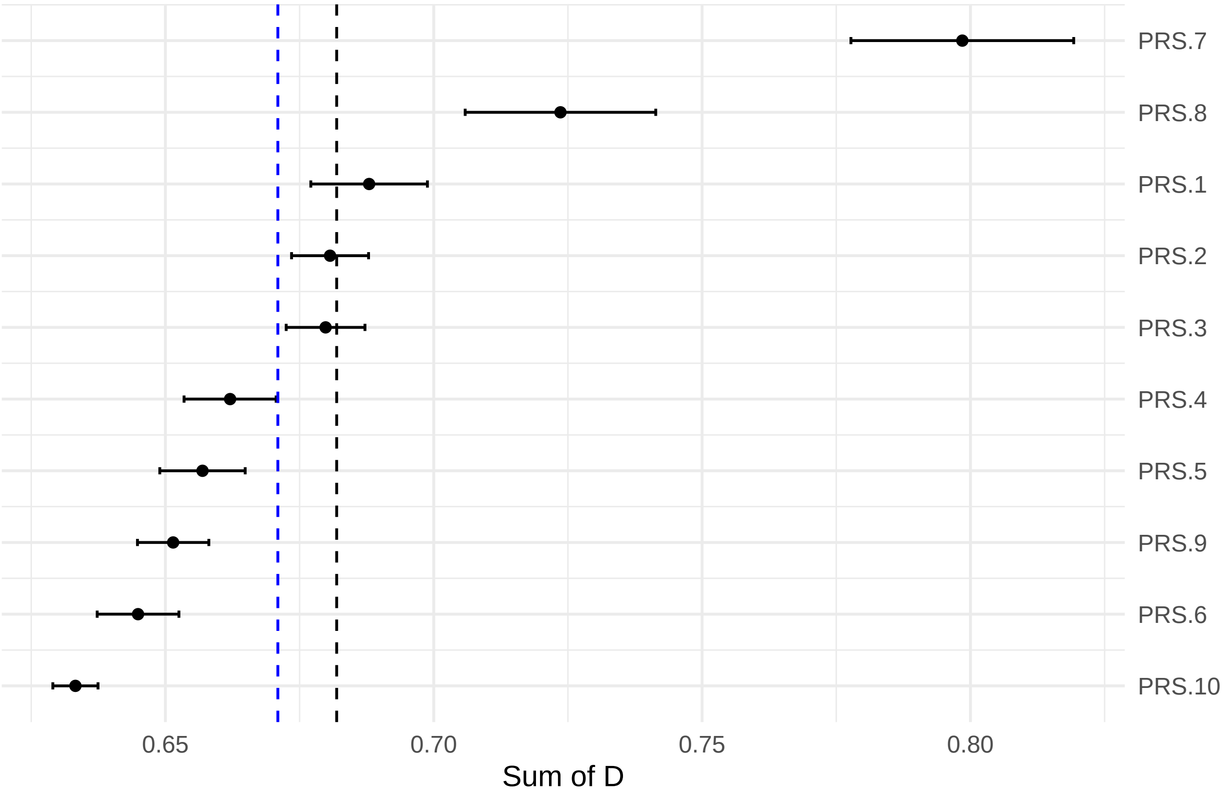
```
##          Model1    Model2    Model3    Model4       Sum
## PRS.7 0.1816453 0.1816453 0.2175212 0.2176745 0.7984862
## PRS.8 0.1657769 0.1657769 0.1921405 0.1999160 0.7236103
## PRS.1 0.1644174 0.1644174 0.1716650 0.1874602 0.6879599
## PRS.2 0.1648077 0.1648077 0.1710979 0.1799683 0.6806815
## PRS.3 0.1638651 0.1638651 0.1735185 0.1785984 0.6798472
## PRS.4 0.1581438 0.1581438 0.1715735 0.1741957 0.6620568
```

```
mainfilename <- "WExample4"
filename <- paste0(mainfilename, "_Ordered_PRS.csv")
write.csv2(out,file=filename)
```

Plot the sum of coefficients of discrimination coefficients D. Lines: in blue the median; in black the mean.

```
out <- data.frame(out)
n <- dim(out)[1]
select <- grep("Model", names(out), value=FALSE)
out$effect <- out$Sum
sds <- apply(out[, select], 1, sd)
out$lower <- out$effect - sds
out$upper <- out$effect + sds
out$rank <- n:1


ggplot(data=out, aes(y=rank, x=effect, xmin=lower, xmax=upper)) +
  geom_point() +
  geom_errorbarh(height=.1) +
  scale_y_continuous(name=NULL, breaks= n:1, labels=row.names(out), position="right") +
  labs(title='', x='Sum of D', y = 'PRS') +
  geom_vline(xintercept=mean(out$effect), color='black', linetype='dashed') +
  geom_vline(xintercept=median(out$effect), color='blue', linetype='dashed') +
  theme_minimal()
```



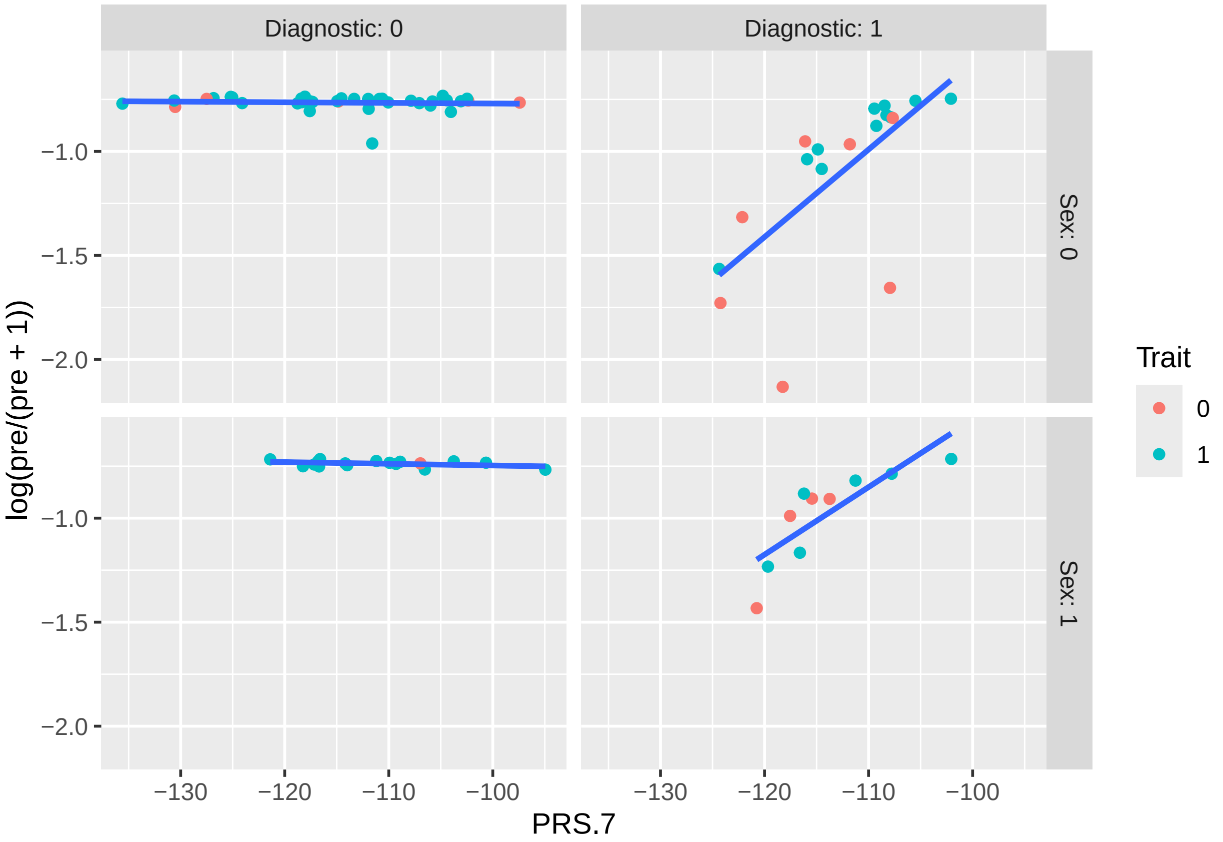
According to the obtained results, first PRS.7 is selected to analyse its association with the Trait.

3. Which model, of all the possible ones, should be used?

The following figure represents the scatter plot separated by Sex and Diagnostic groups.

```r
# First candidate PRS.7
# Plot it
M <- glm(Trait ~ PRS.7*Sex + PRS.7*Diagnostic + Age + PC1 + PC2, data=dat, family=binomial())
pre <- M$fitted.values #predict(M,type='response')
ggplot(dat, aes(x=PRS.7, y=log(pre/(pre+1)))) +
  geom_point(aes(color=Trait)) +
  geom_smooth(method=lm, se=FALSE)+
  facet_grid(Sex ~ Diagnostic, labeller=label_both)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# Candidate FM Trait ~ PRS + Sex + Diagnostic + PRS*Diagnostic + C1 +C2
```

The plots suggest that the interaction between the PRS.7 and the diagnostic is relevant. Thus, we set the full model candidate (FM): $\log(p/1-p) \sim PRS + Sex + Diagnostic + PRS \cdot Diagnostic + Age + PC1 + PC2$.

5. For a binary trait, what steps should be followed for a correct analysis?

Check for overdispersion

```
#model
FM <- glm(Trait ~ Sex + PRS.7*Diagnostic + Age + PC1 + PC2, data=dat, family=binomial())
#Residual Deviance
FM$deviance
```

```
## [1] 67.46733
```

```
# Ratio
FM$deviance/FM$df.residual
```

```
## [1] 0.92421
```

Since this ratio is close to 1, there is not evidence of overdispersion.

```
#With chi-squared test
FM.od <- glm(Trait ~ Sex + PRS.7*Diagnostic + Age + PC1 + PC2, data=dat, family=quasibinomial())
pchisq(summary(FM.od)$dispersion * FM$df.residual,
       FM$df.residual, lower = FALSE)
```

```
## [1] 0.3389642
```

With this p-value = 0.3389, we conclude that there is not evidence of overdispersion.

Based on the following table…

```r
summary(FM)
```

```
##
## Call:
## glm(formula = Trait ~ Sex + PRS.7 * Diagnostic + Age + PC1 +
##     PC2, family = binomial(), data = dat)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.953493   5.754567   0.339   0.7343
## Sex1                 0.445672   0.698490   0.638   0.5234
## PRS.7                0.008372   0.051088   0.164   0.8698
## Diagnostic1         16.714171  11.146266   1.500   0.1337
## Age                  0.043589   0.065843   0.662   0.5080
## PC1                 -5.565987   4.517987  -1.232   0.2180
## PC2                  3.784619   5.604881   0.675   0.4995
## PRS.7:Diagnostic1    0.162405   0.097745   1.662   0.0966 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.234  on 80  degrees of freedom
## Residual deviance: 67.467  on 73  degrees of freedom
## AIC: 83.467
##
## Number of Fisher Scoring iterations: 5
```

…the results show that PRS.7 is related with the $\log(p/1-p)$ in the following way:

- $\log(\widehat{p/1-p}) = 1.953 + 0.008 \times PRS.7 + 0.446 \times Sex + 0.044 \times Age - 5.566 \times PC1 + 3.785 \times PC2$, if Diagnostic $= 0$.
- $\log(\widehat{p/1-p}) = (1.953 + 16.714) + (0.008 + 0.162) \times PRS.7 + 0.446 \times Sex + 0.044 \times Age - 5.566 \times PC1 + 3.785 \times PC2$, if Diagnostic $= 1$.

Check whether the respective PRS coefficients under each group are significant or not.

```r
summary(glht(FM, "PRS.7 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = Trait ~ Sex + PRS.7 * Diagnostic + Age + PC1 +
##     PC2, family = binomial(), data = dat)
##
## Linear Hypotheses:
##              Estimate Std. Error z value Pr(>|z|)
## PRS.7 == 0 0.008372   0.051088   0.164     0.87
## (Adjusted p values reported -- single-step method)
```

```r
summary(glht(FM, "PRS.7  + PRS.7:Diagnostic1 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
```

```
## Fit: glm(formula = Trait ~ Sex + PRS.7 * Diagnostic + Age + PC1 +
##     PC2, family = binomial(), data = dat)
##
## Linear Hypotheses:
##                                   Estimate Std. Error z value Pr(>|z|)
## PRS.7 + PRS.7:Diagnostic1 == 0  0.17078    0.08476   2.015   0.0439 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

That means that for those with Diagnostic=0, it seems that the PRS.7 is not related to the Trait with odds = exp(0.0082) = 1.008, but for those with Diagnosis =1 the model indicates that the coefficient of PRS.7 is 0.0082 + 0.162 = 0.1702, so the odds increase exp(0.1702) = 1.186 for an incremental of one unit in PRS.7 with a p-value=0.0439.

It is also possible to compute a permutation test to assess whether the increase in the coefficient of determination D is significative.

```
# Null model
NM <- glm(Trait ~  Sex + Diagnostic + Age + PC1 + PC2, data=dat, family=binomial() )
permtest <- dD(NM, FM, seed=1236)
permtest
```

```
## $dD
##          1
## 0.07083991
##
## $pvalue
## [1] 0.09
```

In this particular case, it can be seen that the coefficient of discrimination of the FM model is 0.07 units bigger than the corresponding to the *Null Model* NM, but it is not statistically significant.

- **Last step: We move to the next PRS.**