Real data: CAPE Positive

In: A guide to test association between Polygenic Risk Scores and psychological and psychiatric traits: practical examples

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

Working flow and code

In this real data set there are 106 PRS, and a **binary Trait** ($CAPE_{Positive}$), with gender, age, and two Principal Components as covariates.

- Data reading

```
dat <- read.table("Real_data_Positive.csv", header=TRUE, sep="\t", dec=".")
names(dat) #
```

```
##   [1] "ID"           "Sex"          "Age"          "CAPE_Positive"
##   [5] "PRS.1"        "PRS.2"        "PRS.3"        "PRS.4"
##   [9] "PRS.5"        "PRS.6"        "PRS.7"        "PRS.8"
##  [13] "PRS.9"        "PRS.10"       "PRS.11"       "PRS.12"
##  [17] "PRS.13"       "PRS.14"       "PRS.15"       "PRS.16"
##  [21] "PRS.17"       "PRS.18"       "PRS.19"       "PRS.20"
##  [25] "PRS.21"       "PRS.22"       "PRS.23"       "PRS.24"
##  [29] "PRS.25"       "PRS.26"       "PRS.27"       "PRS.28"
##  [33] "PRS.29"       "PRS.30"       "PRS.31"       "PRS.32"
##  [37] "PRS.33"       "PRS.34"       "PRS.35"       "PRS.36"
##  [41] "PRS.37"       "PRS.38"       "PRS.39"       "PRS.40"
##  [45] "PRS.41"       "PRS.42"       "PRS.43"       "PRS.44"
##  [49] "PRS.45"       "PRS.46"       "PRS.47"       "PRS.48"
##  [53] "PRS.49"       "PRS.50"       "PRS.51"       "PRS.52"
##  [57] "PRS.53"       "PRS.54"       "PRS.55"       "PRS.56"
##  [61] "PRS.57"       "PRS.58"       "PRS.59"       "PRS.60"
##  [65] "PRS.61"       "PRS.62"       "PRS.63"       "PRS.64"
##  [69] "PRS.65"       "PRS.66"       "PRS.67"       "PRS.68"
##  [73] "PRS.69"       "PRS.70"       "PRS.71"       "PRS.72"
##  [77] "PRS.73"       "PRS.74"       "PRS.75"       "PRS.76"
##  [81] "PRS.77"       "PRS.78"       "PRS.79"       "PRS.80"
##  [85] "PRS.81"       "PRS.82"       "PRS.83"       "PRS.84"
##  [89] "PRS.85"       "PRS.86"       "PRS.87"       "PRS.88"
##  [93] "PRS.89"       "PRS.90"       "PRS.91"       "PRS.92"
##  [97] "PRS.93"       "PRS.94"       "PRS.95"       "PRS.96"
## [101] "PRS.97"       "PRS.98"       "PRS.99"       "PRS.100"
## [105] "PRS.101"      "PRS.102"      "PRS.103"      "PRS.104"
## [109] "PRS.105"      "PRS.106"      "PC1"          "PC2"
```

```
dat <- dat[, -1]
```

Check that all variables you are interested in are properly read and that there are not other variables you do not need.

- Do not forget to declare the categorical variables as factors

```
dat$Sex <- as.factor(dat$Sex)
dat$CAPE_Positive <- as.factor(dat$CAPE_Positive)
```

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 + Age + PC1 + PC2
- FM_{Sex}: log(p/1-p) versus PRS.i + Sex + PRS.i,Sex + 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="CAPE_Positive", prsname = "PRS.", statistic = "D")
head(out)
```

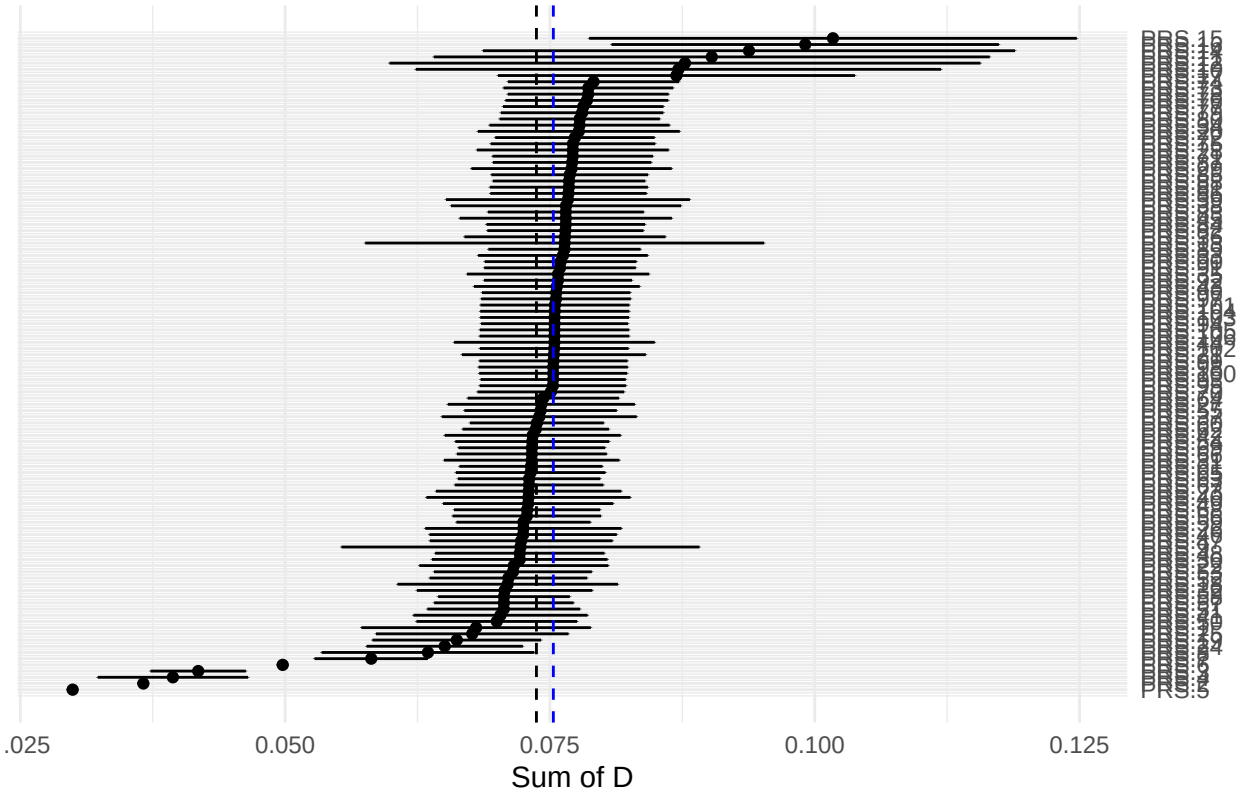
```
##            Model1     Model2        Sum
## PRS.15 0.03458250 0.06713832 0.10172083
## PRS.16 0.03662211 0.06247882 0.09910094
## PRS.14 0.02913388 0.06465976 0.09379364
## PRS.11 0.02658133 0.06369724 0.09027858
## PRS.12 0.02415016 0.06359970 0.08774985
## PRS.10 0.02603126 0.06110069 0.08713195
```

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

Plot the sum 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()
```

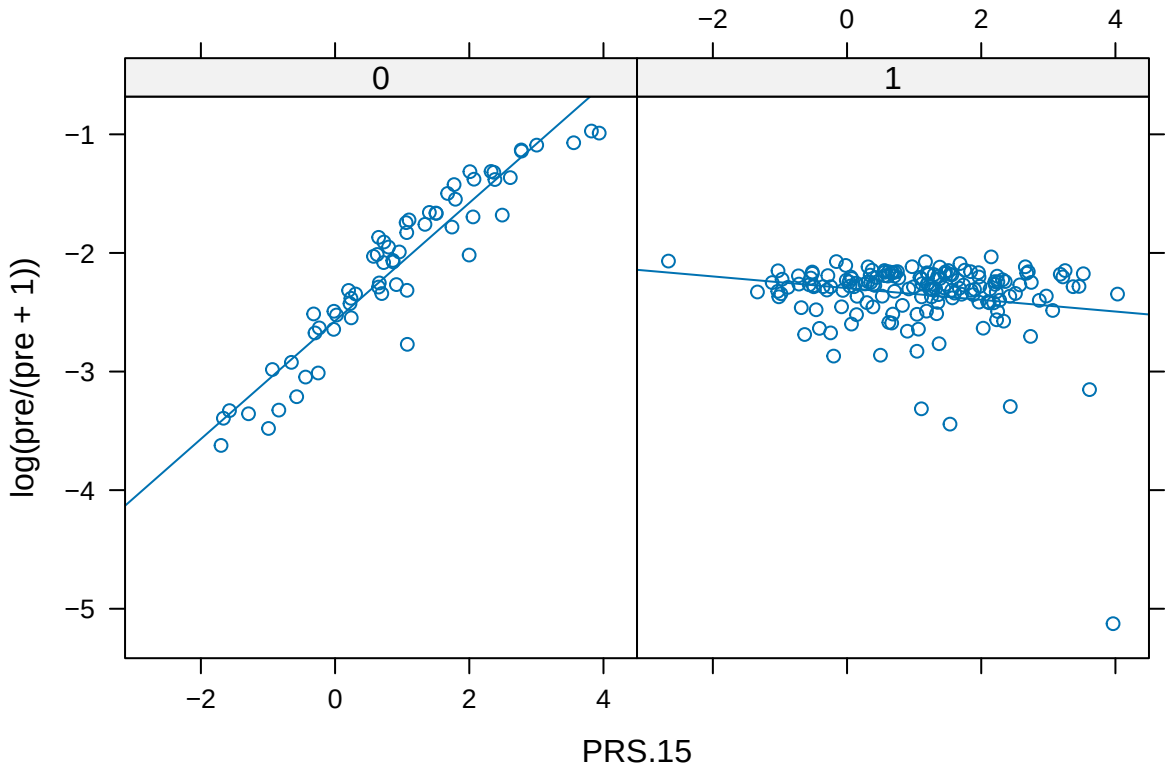


According to the obtained results, first PRS.15 is selected to analyse its association with the $CAPE_{Positive}$.

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

The following Figure represents the scatter plot separated by Sex group.

```
# First candidate PRS.15
# Plot it
M <- glm(CAPE_Positive ~ PRS.15*Sex + Age + PC1 + PC2, data=dat, family=binomial())
pre <- M$fitted.values #predict(M,type='response')
xyplot(log(pre/(pre+1))~PRS.15|Sex, data=dat,  type=c("p", "r"))
```



The plots suggest that the interaction between the PRS.15 and the diagnostic is relevant. Thus, we set the full model candidate (FM): $CAPE_{Positive} \sim PRS + Sex + PRS \cdot Sex + Age + PC1 + PC2$

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

Check for overdispersion

```
#model
FM <- glm(CAPE_Positive ~ Sex + PRS.15*Sex + Age + PC1 + PC2, data=dat, family=binomial())
#Residual Deviance
FM$deviance
```

```
## [1] 165.7134
```

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

```
## [1] 0.7532427
```

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

```
#With chi-squared test
FM.od <- glm(CAPE_Positive ~ Sex + PRS.15*Sex + Age + PC1 + PC2, data=dat, family=quasibinomial())
pchisq(summary(FM.od)$dispersion * FM$df.residual,
       FM$df.residual, lower = F)
```

```
## [1] 0.2651312
```

With this p-value = 0.2651 we conclude that there is not evedence of overdispersion.

Based on the estimations given in this table:

```
summary(FM)
```

```
##
```

```
## Call:
## glm(formula = CAPE_Positive ~ Sex + PRS.15 * Sex + Age + PC1 +
##     PC2, family = binomial(), data = dat)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05492    2.22416   0.025   0.9803
## Sex1         0.13895    0.66711   0.208   0.8350
## PRS.15       0.74628    0.28997   2.574   0.0101 *
## Age         -0.11833    0.10697  -1.106   0.2686
## PC1          1.95129   14.54023   0.134   0.8932
## PC2          2.86709   14.94056   0.192   0.8478
## Sex1:PRS.15 -0.74871    0.35877  -2.087   0.0369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 177.27  on 226  degrees of freedom
## Residual deviance: 165.71  on 220  degrees of freedom
## AIC: 179.71
##
## Number of Fisher Scoring iterations: 5
```

The PRS.15 coefficient will vary depending on the gender, being:

- $\widehat{log(p/1-p)} = 0.055 + 0.746 \times PRS.15 - 0.118 \times Age + 1.951 \times PC1 + 2.867 \times PC2$, if Sex = 0.
- $\widehat{log(p/1-p)} = (0.055 + 0.139) + (0.746 - 0.749) \times PRS.15 - 0.118 \times Age + 1.951 \times PC1 + 2.867 \times PC2$, if Sex = 1.

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

```
summary(glht(FM, "PRS.15 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = CAPE_Positive ~ Sex + PRS.15 * Sex + Age + PC1 +
##     PC2, family = binomial(), data = dat)
##
## Linear Hypotheses:
##             Estimate Std. Error z value Pr(>|z|)
## PRS.15 == 0   0.7463     0.2900   2.574   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
summary(glht(FM, "PRS.15  + Sex1:PRS.15 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = CAPE_Positive ~ Sex + PRS.15 * Sex + Age + PC1 +
##     PC2, family = binomial(), data = dat)
##
## Linear Hypotheses:
```

```
##                             Estimate Std. Error z value Pr(>|z|)
## PRS.15 + Sex1:PRS.15 == 0 -0.002425   0.211998  -0.011    0.991
## (Adjusted p values reported -- single-step method)
```

That means that for those with Sex=1, it seems that the PRS.15 is not related to the Trait with odds = exp(-0.002) = 0.991, but for those with Sex = 0 the model indicates that the coefficient of PRS.15 is 0.7463, so the odds increase exp(0.7463) = 2.109 for an incremental of one unit in PRS.15 with a p-value=0.0101.

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

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

```
## $dD
##          1
## 0.05246593
##
## $pvalue
## [1] 0.008
```

In this particular case, it can be seen that the coefficient of discrimination of the FM model is 0.05 units bigger than the corresponding to the *Null Model* NM, and it is statistically significant.

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