WORKING EXAMPLE 1

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 9 PRSs, and a continuous trait, with sex, clinical diagnosis (with 2 categories), age, and two Principal Components as covariates.

- data reading

```
dat <- read.table("WExample1.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"      "PC1"        "PC2"
```

- do not forget to declare the categorical variables as factors

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

1. What full model should be considered?

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

- FM_{WI}: Trait versus PRS.i + Sex + Diagnostic + Age + PC1 + PC2
- FM_{Sex}: Trait versus PRS.i + Sex + PRS.i $\cdot$ Sex + Diagnostic + Age + PC1 + PC2
- $\text{FM}_{Diagnostic}$: Trait versus PRS.i + Sex + Diagnostic + PRS.i $\cdot$ Diagnostic + Age + PC1 + PC2
- $\text{FM}_{Sex/Diagnostic}$: Trait 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 coefficient of determination R^2 and calculate the sum: $S = R^2_{WI} + R^2_{Sex} + R^2_{Diagnostic} + R^2_{Sex/Diagnostic}$.

According to S, list the PRSs in decreasing order:

```
out <- orderR2(dat, yname="Trait", prsname = "PRS.") # Note this function is included
# in the customized file via source("Functions.R")
head(out)
```

```
##           Model1     Model2     Model3     Model4       Sum
## PRS.4 0.06561329 0.07420427 0.12600084 0.15146070 0.4172791
## PRS.6 0.05596966 0.06220536 0.13638253 0.15759204 0.4121496
## PRS.2 0.04973955 0.05309214 0.11850006 0.12356753 0.3448993
```

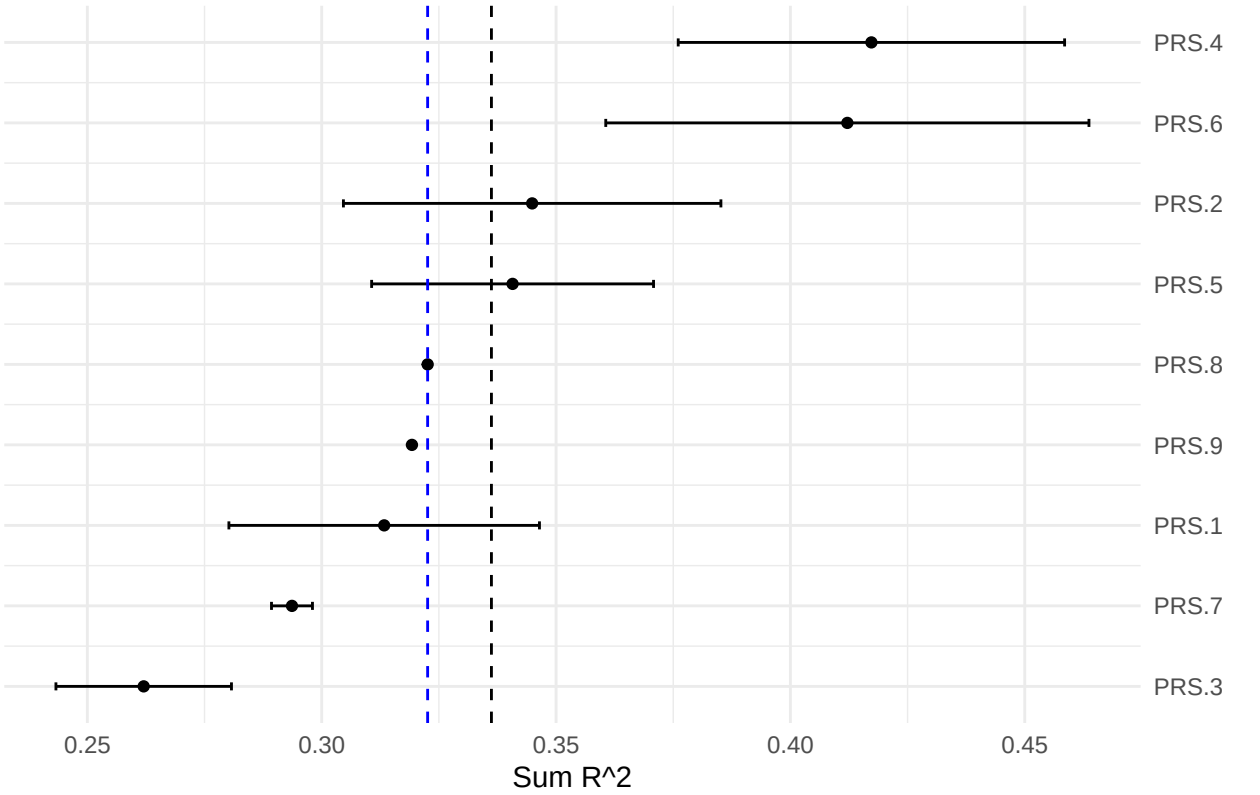
```
## PRS.5 0.05308815 0.06728312 0.10264612 0.11770778 0.3407252
## PRS.8 0.07995274 0.08034903 0.08088220 0.08141782 0.3226018
## PRS.9 0.07932875 0.08028067 0.07933969 0.08030066 0.3192498
```

```
write.csv2(out,file="WExample1_Ordered_PRS.csv")
```

Plot the sum of coefficients of determination S_{R^2}. Lines: in blue the median; in black the mean

```
out <- data.frame(out)
nPRS <- 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 <- nPRS:1

n <- dim(out)[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 R^2', y = 'PRS') +
  geom_vline(xintercept=mean(out$effect), color='black', linetype='dashed') +
  geom_vline(xintercept=median(out$effect), color='blue', linetype='dashed') +
  theme_minimal()
```



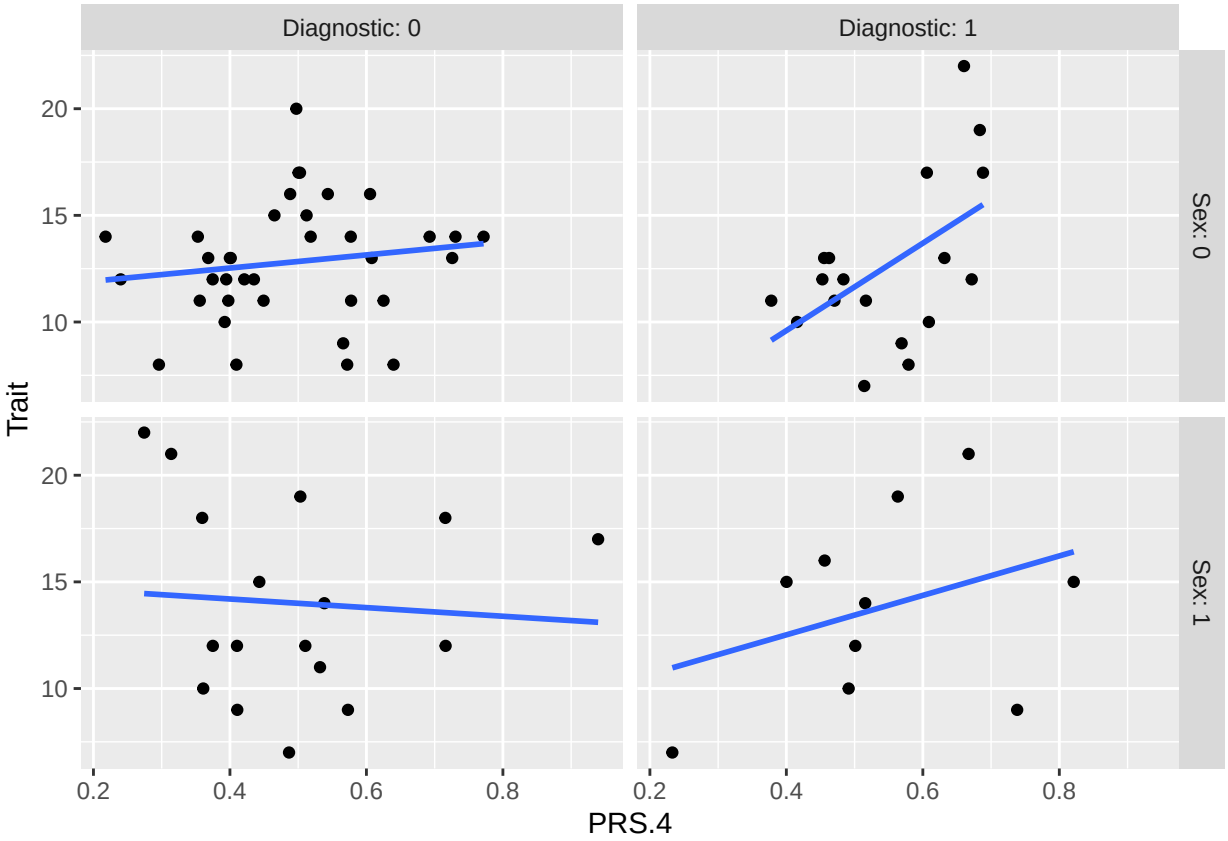
According to the obtained results, first, PRS.4 is selected to analyse its possible association with the Trait.

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

The following Figure represents the scatter plot of Trait versus PRS.4 separated by Sex and Diagnostic groups.

```
# First candidate PRS.4
# Plot it
ggplot(dat, aes(x=PRS.4, y=Trait)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)+
  facet_grid(Sex ~ Diagnostic, labeller=label_both)
```

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



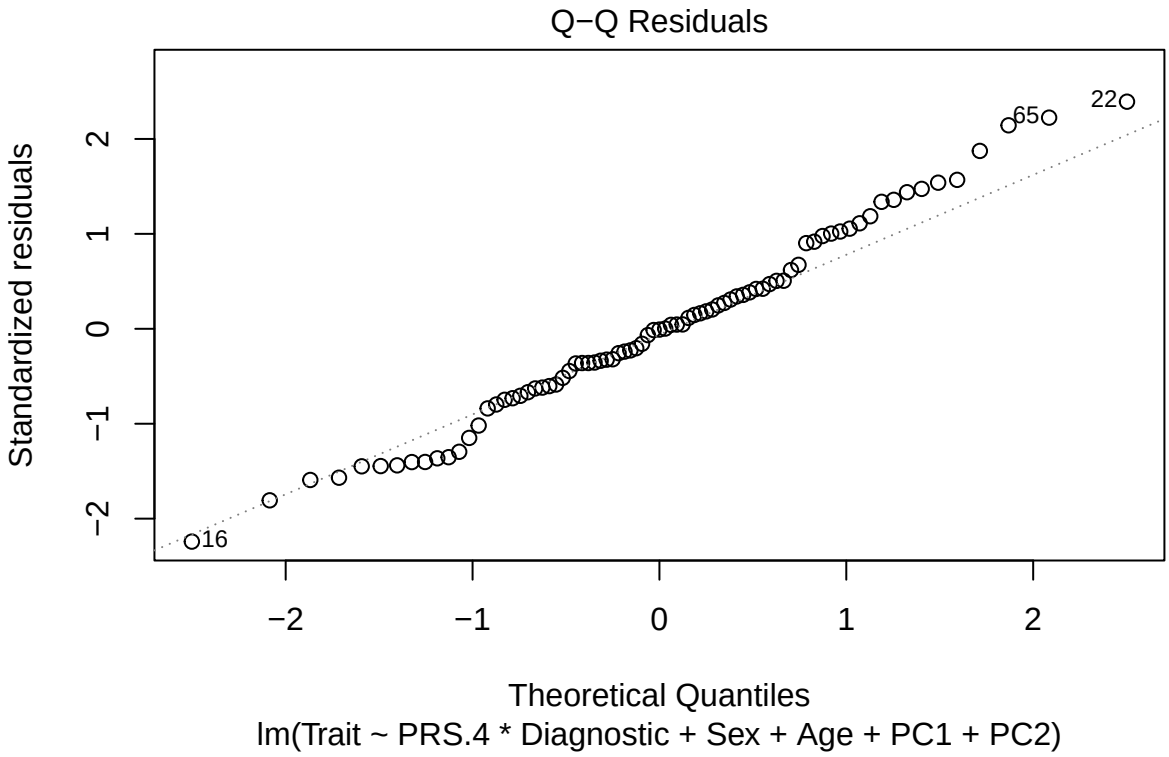
Observing the slopes of the lines, it seems that a different relationship between Trait and PRS.4 could be expected depending on the diagnostic group, but not depending on the sex group.Thus, we set the full model candidate (FM): $Trait \sim PRS + Sex + Diagnostic + PRS \cdot Diagnostic + PC1 + PC2$.

4. For a continuous trait, what steps should be followed for a correct analysis?

- **4.1. How is the candidate model validated?**

First, we validate the normality of the errors and the constant variance conditions (see the figures and the results of Shapiro test and Levene test.

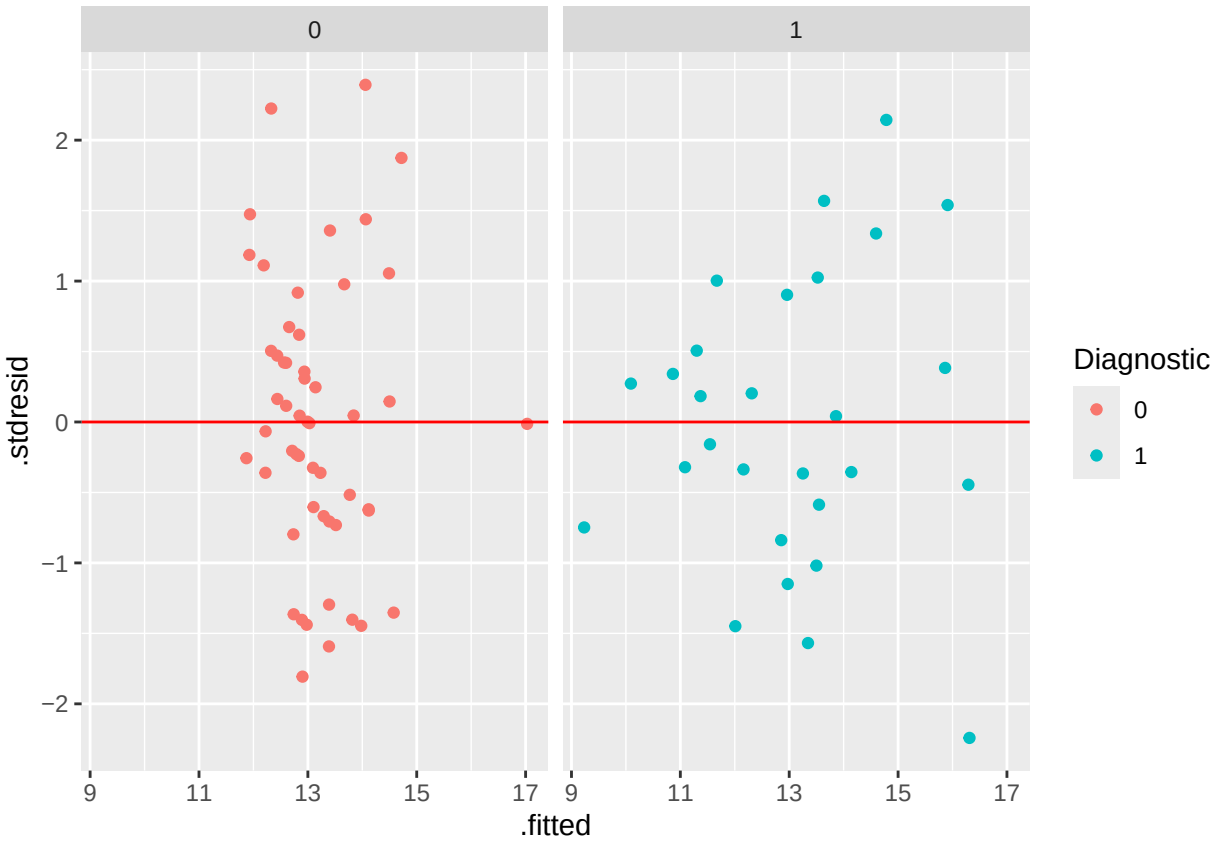
```
#model
FM <- lm(Trait ~ PRS.4*Diagnostic + Sex + Age + PC1 + PC2, data=dat)
#qq-plot for normality
plot(FM,2)
```




```
#Shapiro-Wilk test
shapiro.test(FM$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  FM$residuals
## W = 0.98413, p-value = 0.4166
```

```
#plot for variances
d <- fortify(FM)
ggplot(d,aes(x=.fitted, y=.stdresid, colour=Diagnostic)) +
  geom_point() +
  geom_hline(yintercept=0, col="red") +
  facet_wrap(.~Diagnostic)
```




```
#Levene's test
leveneTest(.stdresid ~ Diagnostic, data=d)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.1193 0.7307
##       79
```

Having the full model validated, we can use it.

- **4.3 How a possible association is established?**

```
summary(FM)
```

```
##
## Call:
## lm(formula = Trait ~ PRS.4 * Diagnostic + Sex + Age + PC1 + PC2,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3117 -2.1180 -0.0277  1.6987  7.9418
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.81613    2.80605   5.280 1.28e-06 ***
## PRS.4             -1.04058    3.71031  -0.280   0.7799
## Diagnostic1       -8.32066    3.62154  -2.298   0.0245 *
```

```
## Sex1                1.29952    0.83922   1.548   0.1258
## Age                -0.07070    0.08635  -0.819   0.4156
## PC1                 0.92650    6.45497   0.144   0.8863
## PC2                 8.03272    6.22984   1.289   0.2013
## PRS.4:Diagnostic1 15.02560    6.69040   2.246   0.0277 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.541 on 73 degrees of freedom
## Multiple R-squared:  0.126,  Adjusted R-squared:  0.04219
## F-statistic: 1.503 on 7 and 73 DF,  p-value: 0.1796
```

Next table shows the parameters and hypothesis (columns 1-3), together with the standard output for this type of analysis (columns 4-7). This output allows the model equation to be built and it would be used if the objective of the study is to predict the Trait values.

	Parameter	Null Hypothesis	Estimate	Std. Error	t value	p-value
Intercept	b_0	$b_0 = 0$	14.816	2.806	5.280	1.28e-06
PRS.4	b_1	$b_1 = 0$	-1.041	3.710	-0.280	0.780
Diagnostic1	b_2	$b_2 = 0$	-8.321	3.622	-2.298	0.025
Sex1	b_3	$b_3 = 0$	1.300	0.839	1.548	0.126
Age	b_4	$b_4 = 0$	-0.071	0.086	-0.819	0.416
PC1	b_5	$b_5 = 0$	0.927	6.455	0.144	0.886
PC2	b_6	$b_6 = 0$	8.033	6.230	1.289	0.201
PRS.4:Diagnostic1	b_7	$b_7 = 0$	15.026	6.690	2.246	0.028

Table 1: : Working example 1. Parameters, null hypothesis, estimates, standard errors, t statistics, and p-values for the regression coefficients in the $\text{FM}_{Diagnostic}$.

The PRS.4 coefficient will vary depending on the diagnostic group each individual belongs to, being:

- if Diagnostic $= 0$, $\widehat{Trait} = 14.816 + -1.041\, PRS.4 + 1.3\, Sex - 0.071\, Age + 0.927\, PC1 + 8.033\, PC2$
- if Diagnostic $= 1$, $\widehat{Trait} = (14.816 - 8.321) + (-1.041 + 15.022)\, PRS.4 + 1.3\, Sex - 0.071\, Age + 0.927\, PC1 + 8.033\, PC2$

where Sex takes values 0 or 1, depending on whether the individual under study is male or female, affecting only the value of the intercept.

However, if the objective is to evaluate the possible association between Trait and PRS.4, it can be interesting to check whether the respective PRS.4 coefficients under each Diagnostic group are considerable or not. Note that these coefficients vary according to the Diagnosis group being b_1 for the Diagnostic group 0, which has been taken as the basal group; $b_1 + b_7$ for the Diagnostic 1:

```
summary(glht(FM, "PRS.4 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Trait ~ PRS.4 * Diagnostic + Sex + Age + PC1 + PC2,
##     data = dat)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## PRS.4 == 0   -1.041      3.710   -0.28     0.78
## (Adjusted p values reported -- single-step method)
```

```r
summary(glht(FM, "PRS.4  + PRS.4:Diagnostic1 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Trait ~ PRS.4 * Diagnostic + Sex + Age + PC1 + PC2,
##     data = dat)
##
## Linear Hypotheses:
##                                Estimate Std. Error t value Pr(>|t|)
## PRS.4 + PRS.4:Diagnostic1 == 0    13.98       5.55    2.52   0.0139 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

That means that for Diagnostic group 0 there is no relevant association (coeff = -1.041), but for Diagnostic group 1, the association is strong, positive (coeff = 13.980), and significant (p-value = 0.014).

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