WORKING EXAMPLE 3

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

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

Working flow and code

In this example we simulate 5 PRSs, and a continuous trait, with sex, clinical diagnosis (with 2 categories), age, and two Principal Components as covariates.

- data reading

```
dat <- read.table("WExample3.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
- FM$_{Diagnostic}$: Trait versus PRS.i + Sex + Diagnostic + PRS.i $\cdot$ Diagnostic + Age + PC1 + PC2
- 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: $S = R^2_{WI} + R^2_{Sex} + R^2_{Diagnostic} + R^2_{Sex \cdot Diagnostic}$.

According to S, list the PRSs in decreasing order:

```
out <- orderR2(dat, yname="Trait", prsname = "PRS.")
head(out)
```

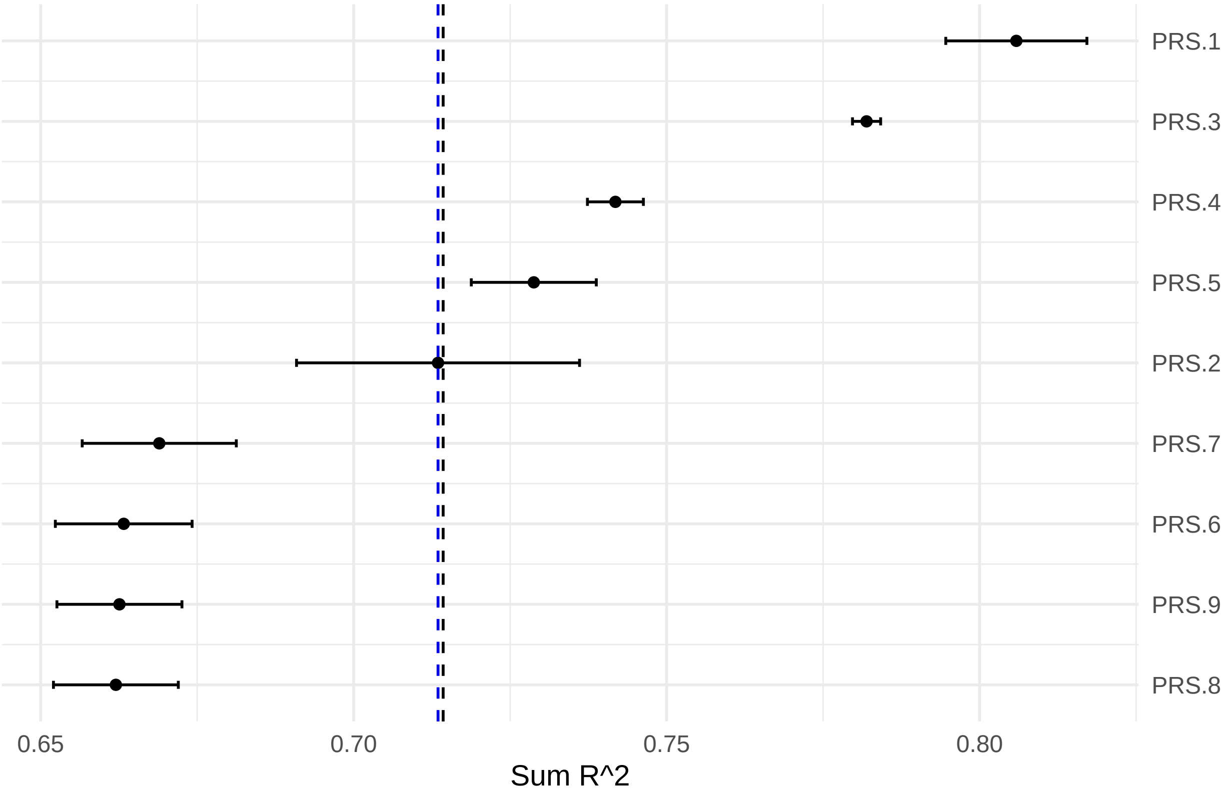
```
##          Model1    Model2    Model3    Model4       Sum
## PRS.1 0.1906479 0.2106171 0.1928382 0.2117647 0.8058678
## PRS.3 0.1933149 0.1972267 0.1937877 0.1976130 0.7819423
## PRS.4 0.1815838 0.1893220 0.1815840 0.1893272 0.7418170
## PRS.5 0.1734391 0.1896543 0.1737373 0.1919494 0.7287801
## PRS.2 0.1536939 0.1924057 0.1653767 0.2020034 0.7134798
## PRS.7 0.1526951 0.1747579 0.1617509 0.1797416 0.6689455
```

```
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()
```



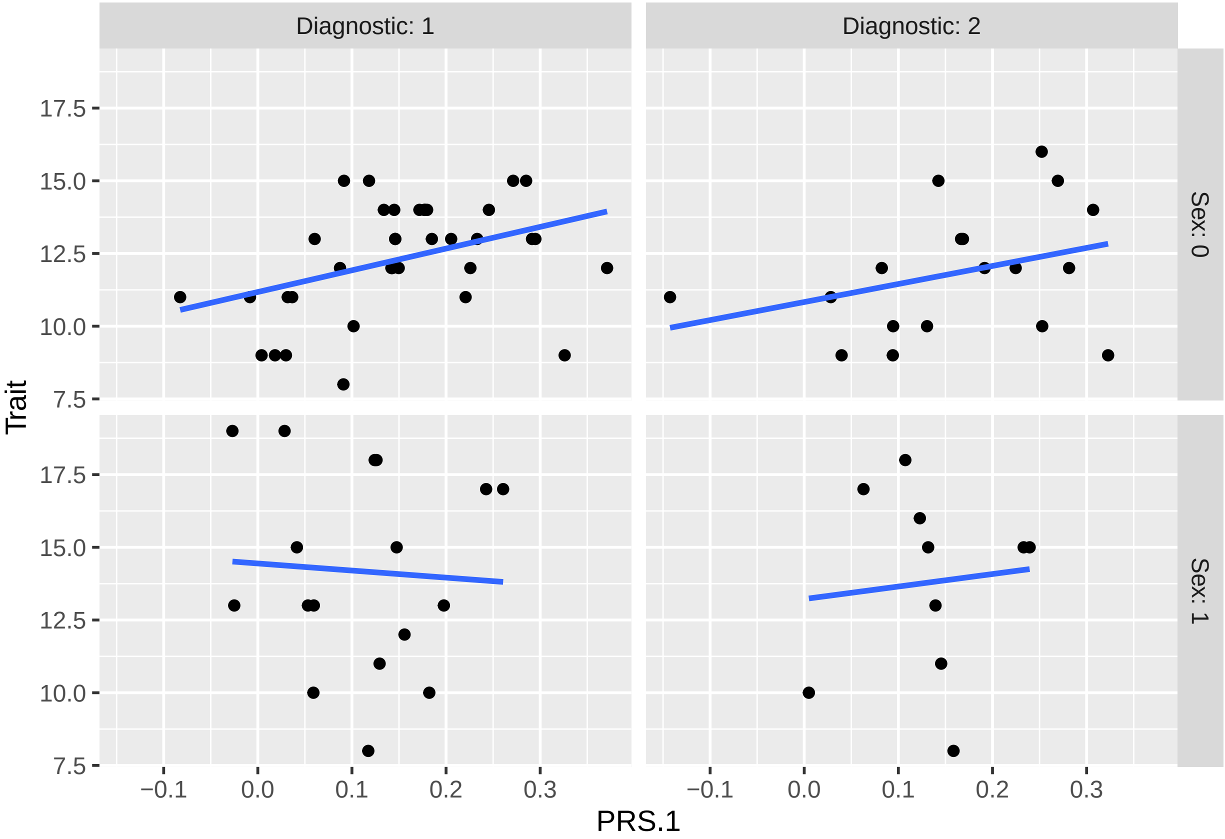
According to the obtained results, first PRS.1 is selected to analyze its association with the Trait.

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

The following figure represents the scatter plot of Trait versus PRS.1 separated by Sex and Diagnostic groups.

```r
ggplot(dat, aes(x=PRS.1, y=Trait)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)+
  facet_grid(Sex ~ Diagnostic, labeller=label_both)
```

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



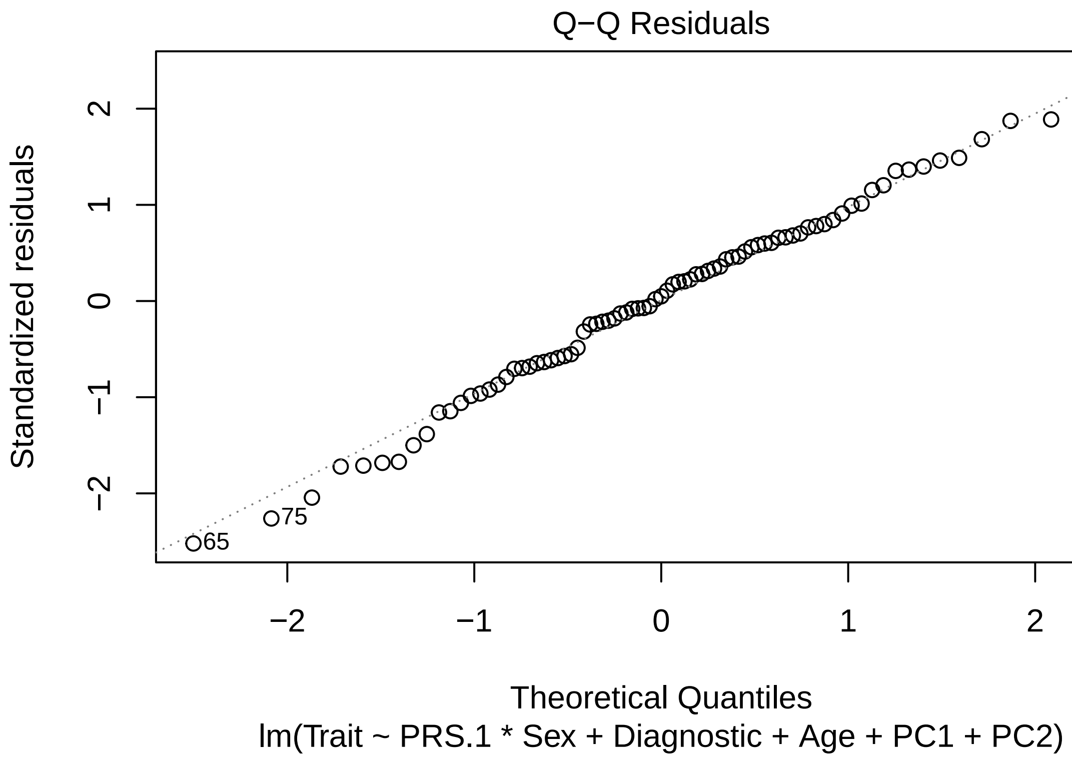
The plots suggest that the interaction between the PRS.1 and sex is relevant. Thus, we set the full model candidate (FM): $Trait \sim PRS + Sex + Diagnostic + PRS \cdot Sex + PC1 + PC2$.

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).

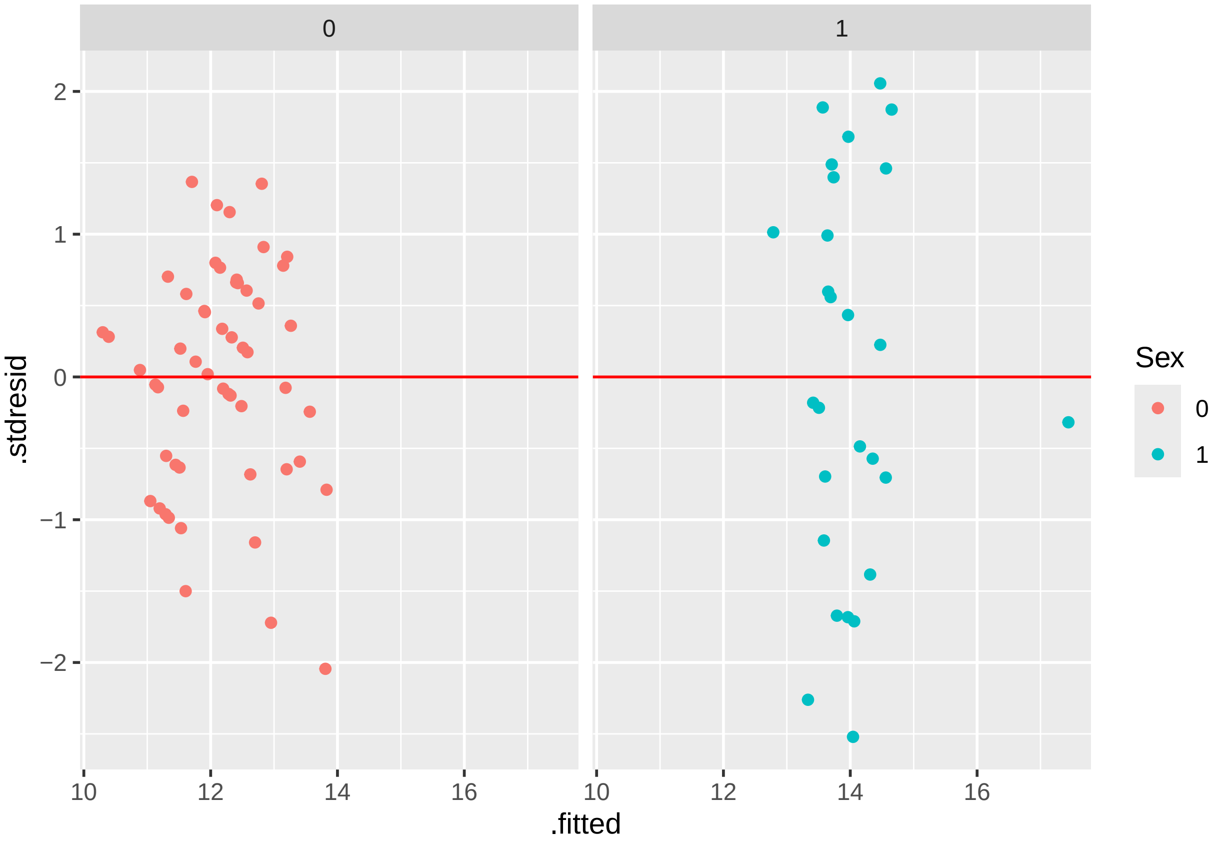
```r
#model
FM <- lm(Trait ~ PRS.1*Sex + Diagnostic + Age + PC1 + PC2, data=dat)
#qq-plot for normality
plot(FM,2)
```



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

```
##
##  Shapiro-Wilk normality test
##
## data:  FM$residuals
## W = 0.98615, p-value = 0.5353
```

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



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

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)    
## group  1  15.673 0.0001636 ***
##       79                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

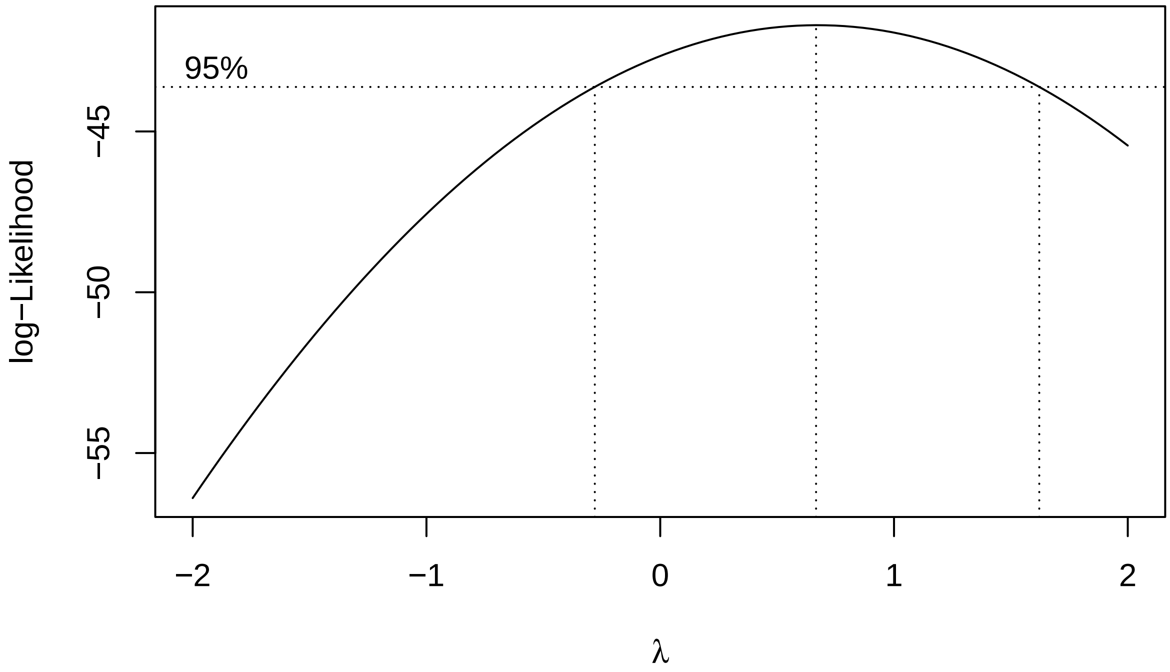
It seems that homocedasticity does not hold.

- **4.2. What can be done if any validation condition fails?**

We have two approaches to assess the possible association with PRS.1 and the Trait: try a Box-Cox transformation or perform a weighted permutation test.

First, we try a Box-Cox transformation of the dependent variable: - Determine the lambda value.

```
b <- boxcox(FM)
```



```
# Exact lambda
lambda <- b$x[which.max(b$y)]
lambda
```

```
## [1] 0.6666667
```

- Transform the dependent variable and establish the new model.

```
dat$newTrait <- (dat$Trait ^ lambda - 1) / lambda
FM <- lm(newTrait ~ PRS.1*Sex + Diagnostic + Age + PC1 + PC2, data=dat)
```

- Check the normality and homocedasticity conditions.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  FM$residuals
## W = 0.98029, p-value = 0.2465
```

```
#Levene's test
d <- fortify(FM)
leveneTest(.stdresid ~ Sex, data=d)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)
## group  1  13.368 0.0004595 ***
##       79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the suggested Box-Cox transformation with $\lambda = 0.667$ does not solve the heteroscedasticity problem. For this reason we perform a weighted-permutation test.

```
NM <- lm(Trait ~ Sex + Diagnostic + Age + PC1+PC2, data=dat)
FM <- lm(Trait ~ PRS.1*Sex + Diagnostic + Age + PC1+PC2, data=dat)
outperm <- dR2(NullModel=NM, FullModel=FM, B=1000, seed=165,  weights=TRUE)
outperm
```

```
## $dR2
## [1] 0.09217286
##
## $pvalue
## [1] 0
```

We observe an increase of 0.092 in the coefficient of determination when the PRS.1 is included in the model and the permutation test indicates it is significant.

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