

Real data: CAPE Negative

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 real data set there are 106 PRS, and a **continuous Trait** ($CAPE_{Negative}$), with sex, age, and two Principal Components as covariates. For more details see Section 7 in the paper.

- Data reading

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

##	[1]	"ID"	"Sex"	"Age"	"CAPE_Negative"
##	[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]
```

Important! 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)
```

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 + Age + PC1 + PC2
- FM_{Sex} : Trait 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 coefficient of determination R^2 and calculate the sum: $S = R_{WI}^2 + R_{Sex}^2$.

According to S, list the PRSs in decreasing order:

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

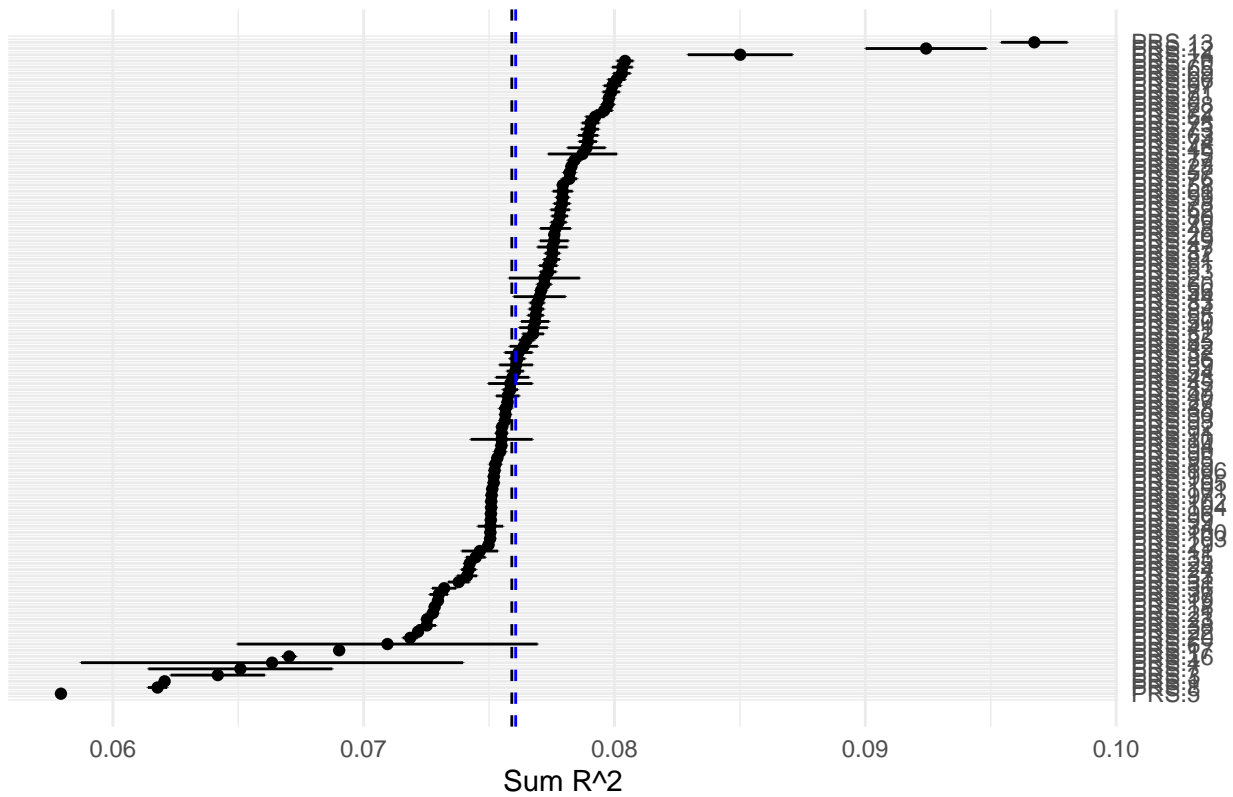
```
##           Model1      Model2      Sum
## PRS.13 0.04743339 0.04930427 0.09673766
## PRS.12 0.04450825 0.04791485 0.09242311
## PRS.14 0.04102906 0.04398352 0.08501259
## PRS.70 0.03997513 0.04044623 0.08042136
## PRS.65 0.03986674 0.04044801 0.08031474
## PRS.69 0.03990297 0.04039181 0.08029478
```

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

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()
```

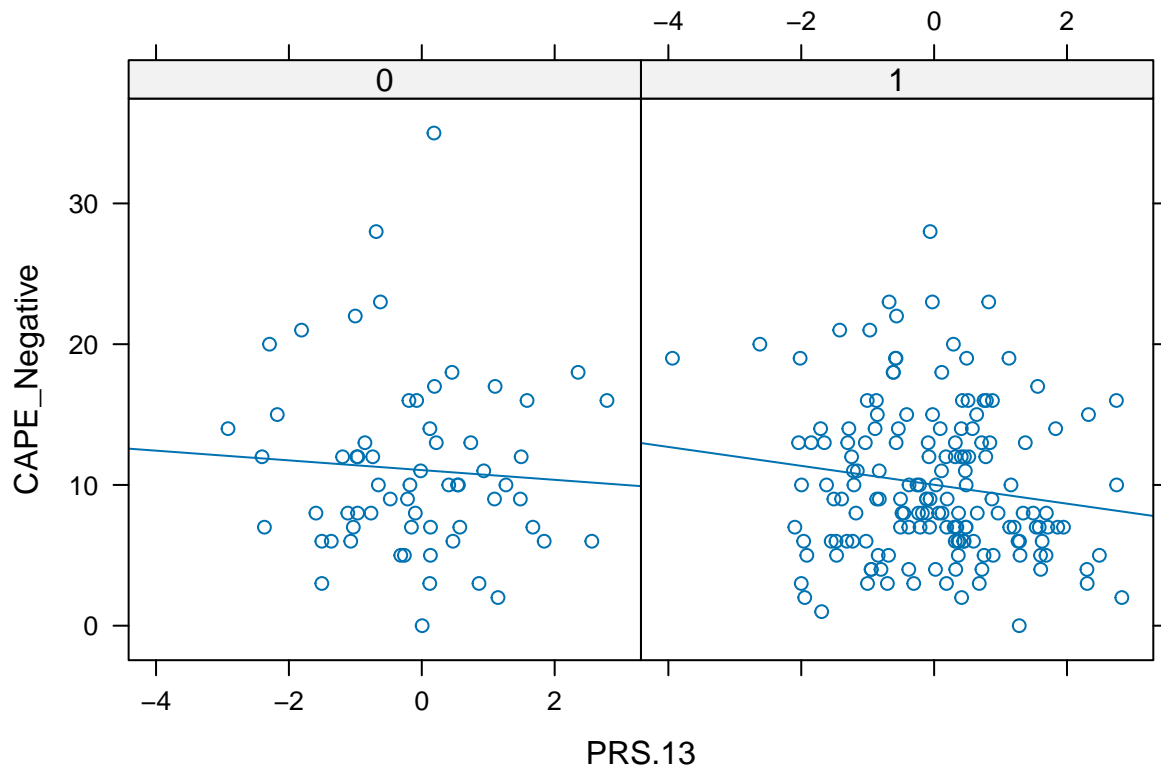


According to the obtained results, first, PRS.13 is selected to analyse its possible association with the $CAPE_{Negative}$.

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

The following Figure represents the scatter plot of $CAPE_{Negative}$ versus PRS.13 separated by Sex group.

```
# First candidate PRS.13
# Plot it
library(lattice)
xyplot(CAPE_Negative~PRS.13|Sex, data=dat, type=c("p", "r"))
```



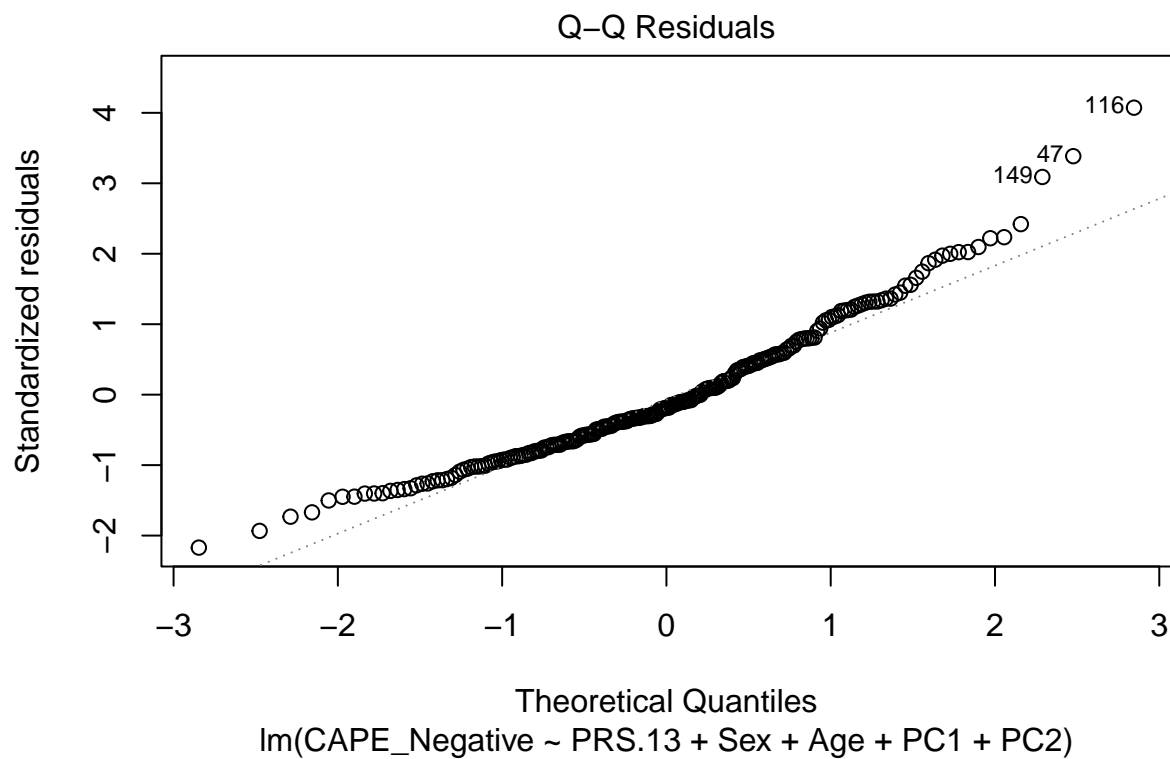
The plots suggest that the interaction between the PRS.13 and the sex is not relevant. Thus, we set the full model candidate (FM): $CAPE_{Negative} \sim PRS + Sex + Age + PC1 + PC2$.

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(CAPE_Negative ~ PRS.13 + Sex + Age + PC1 + PC2, data=dat)
#qq-plot for normality
plot(FM,2)
```



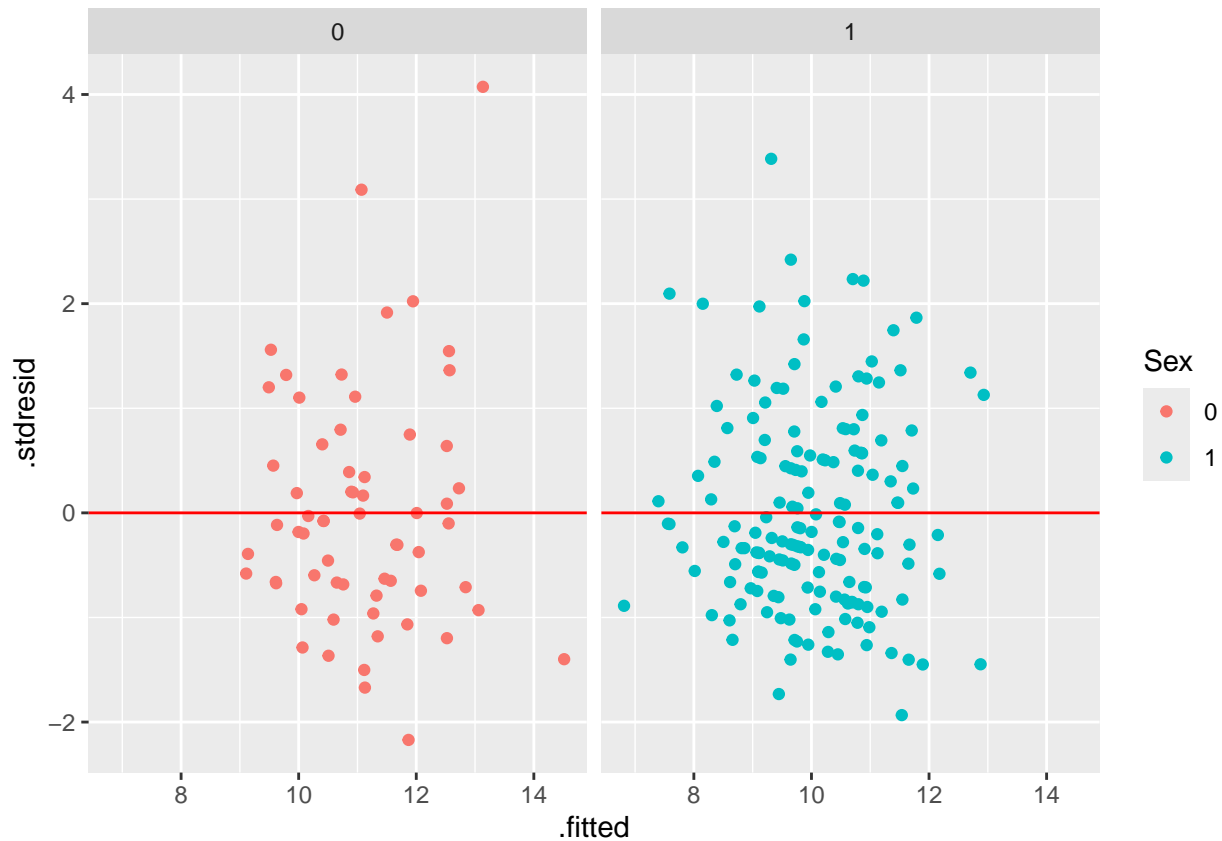
The lack of normality of residuals is suggested by this last plot.

This supported by Shapiro's test:

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

```
##
##  Shapiro-Wilk normality test
##
## data:  FM$residuals
## W = 0.95885, p-value = 4.335e-06
```

```
#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  0.7613 0.3839
##      224
```

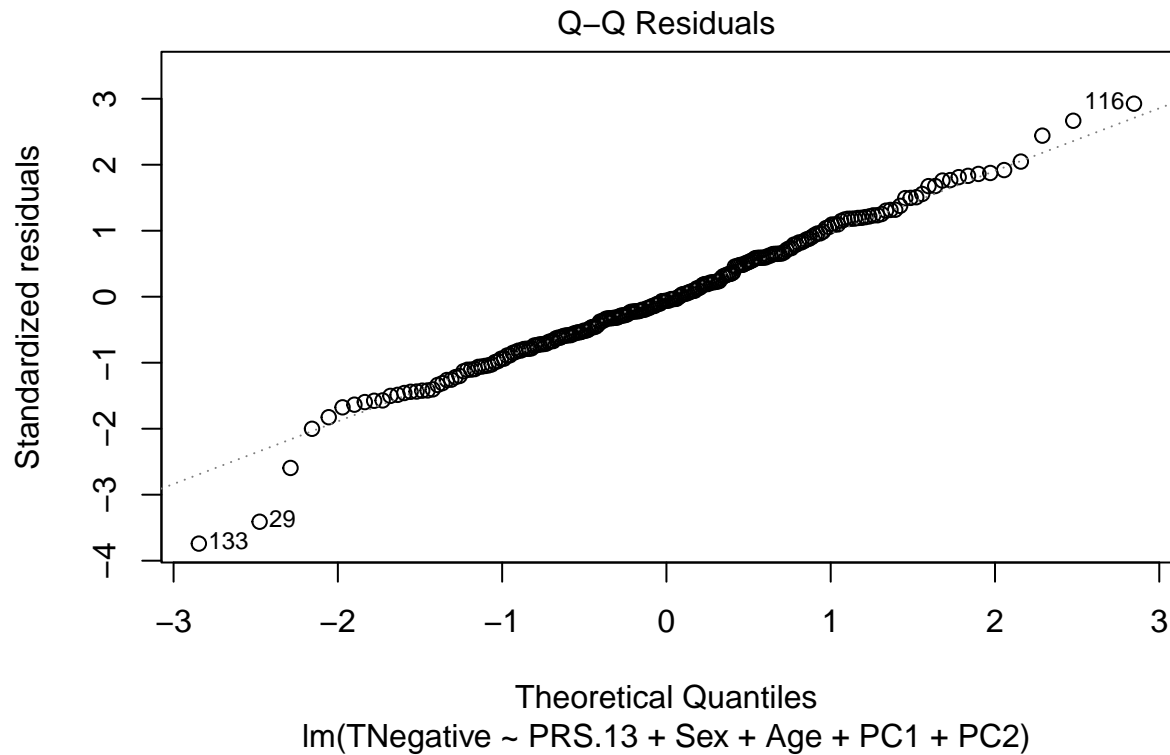
All in all, it seems that there is lack of normality of residuals but linearity and homocedasticity assumptions hold.

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

We have two approaches to assess the possible association with $PRS.13$ and $CAPE_{Negative}$: try a transformation or perform a permutation test.

First we try the squared root transformation for the dependent variable:

```
dat$TNegative <- sqrt(dat$CAPE_Negative)
FM <- lm(TNegative ~ PRS.13 + 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 = 0.98898, p-value = 0.08108
```

It is suggested the transformation offered a solution for the lack of normality. Furthermore,...

```
#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  0.2998 0.5846
##      224
```

...results do not indicate evidence against homoscedasticity, neither a pattern is observed that could indicate a lack of linearity.

Therefore, the model we build is given by:

```
summary(FM)
```

```
##
## Call:
## lm(formula = TNegative ~ PRS.13 + Sex + Age + PC1 + PC2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2590 -0.5534 -0.0529  0.5640  2.5020
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.013198   0.467855   6.440 7.39e-10 ***
## PRS.13        -0.101131   0.050141  -2.017  0.0449 *
## Sex1          -0.142137   0.133729  -1.063  0.2890
## Age           0.008634   0.021935   0.394  0.6943
```



```
## PC1          3.796970    4.232105    0.897    0.3706
## PC2          8.408721    4.206204    1.999    0.0468 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8837 on 220 degrees of freedom
## (1 observation deleted due to missingness)
## Multiple R-squared:  0.04105,    Adjusted R-squared:  0.01926
## F-statistic: 1.884 on 5 and 220 DF,  p-value: 0.09824
```

The results show that PRS.13 is related with the \sqrt{Trait} in the following way:

$$\widehat{\sqrt{Trait}} = 3.013 - 0.101 \times PRS.13 - 0.142 \times Sex + 0.009 \times Age + 3.797 \times PC1 + 8.409 \times PC2,$$

where Sex takes values 0 or 1, depending on whether the individual under study is male or female. See section 8.1 in the paper for more details.

On the other hand, based on the permutation approach:

```
NM <- lm(CAPE_Negative ~ Sex + Age + PC1 + PC2, data=dat)
FM <- lm(CAPE_Negative ~ PRS.13 + Sex + Age + PC1 + PC2, data=dat)
outperm <- dR2(NullModel=NM, FullModel=FM, B=5000, seed=165)
outperm
```

```
## $dR2
## [1] 0.01964777
##
## $pvalue
## [1] 0.033
```

We observe an increase of 0.0196 in the coefficient of determination when the PRS.13 is included in the model and the permutation test indicates it is significant.

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