WORKING EXAMPLE 2

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

- data reading

```
dat <- read.table("WExample2.csv", header=TRUE, sep=";", dec=",")
names(dat) #
```

```
##  [1] "Diagnostic" "Sex"        "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)
```

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: $S = R^2_{WI} + R^2_{Sex} + R^2_{Diagnostic} + R^2_{Sex/Diagnostic}$.

According to S, list the PRSs in decreasing order.

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

```
##            Model1     Model2    Model3    Model4       Sum
## PRS.5  0.06590991 0.14386357 0.1167506 0.1730818 0.4996059
## PRS.10 0.04263225 0.05908320 0.1400714 0.1838914 0.4256782
```

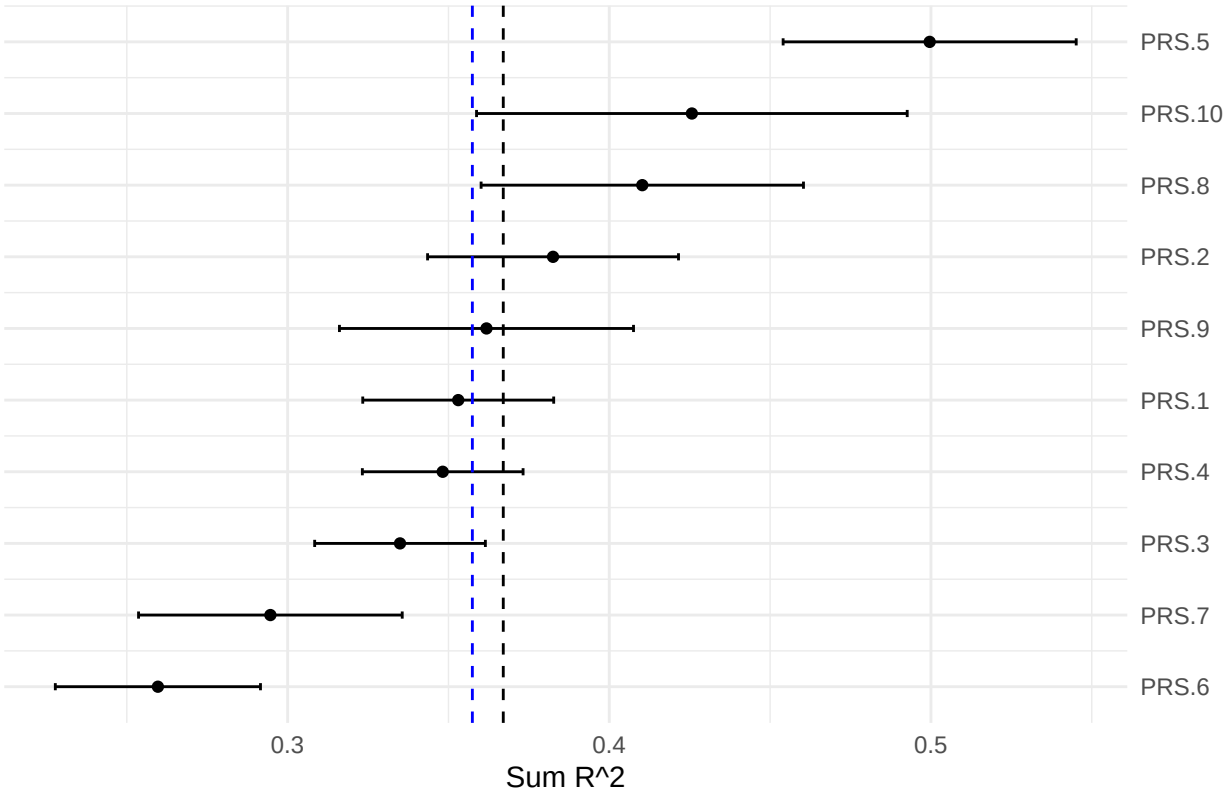
```
## PRS.8  0.05330296 0.07034766 0.1239488 0.1626609 0.4102603
## PRS.2  0.06157642 0.06212837 0.1293514 0.1294614 0.3825176
## PRS.9  0.03691860 0.07445585 0.1061419 0.1443078 0.3618242
## PRS.1  0.06250443 0.06260526 0.1138417 0.1140869 0.3530384
```

```
mainfilename <- "WExample2"
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)
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 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()
```



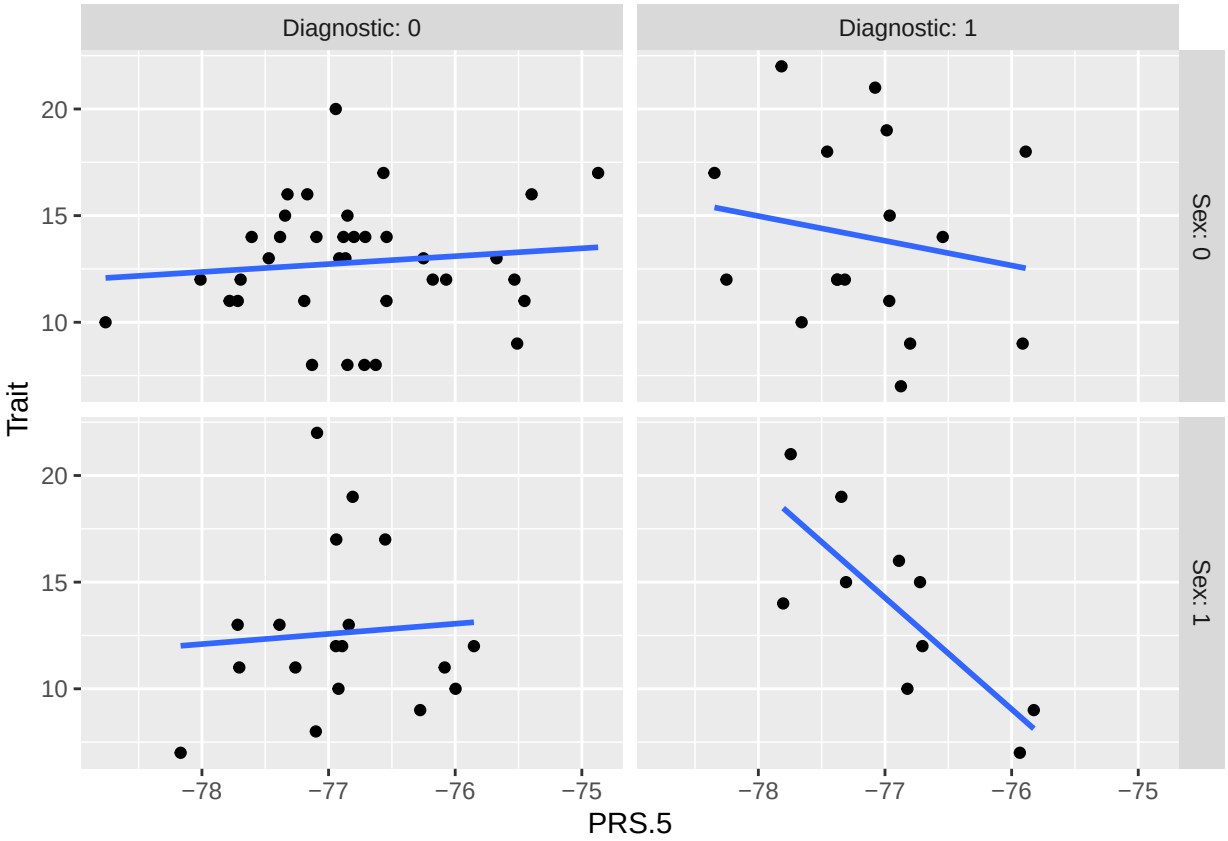
According to the obtained results, first PRS.5 is selected to analyse its association with the Trait.

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

The following figure represents the scatter plot of Trait versus PRS.5 separated by Sex and Diagnostic groups.

```
ggplot(dat, aes(x=PRS.5, y=Trait)) +
  geom_point() +
  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 + PRS*Diagnostic + PC1 +PC2
```

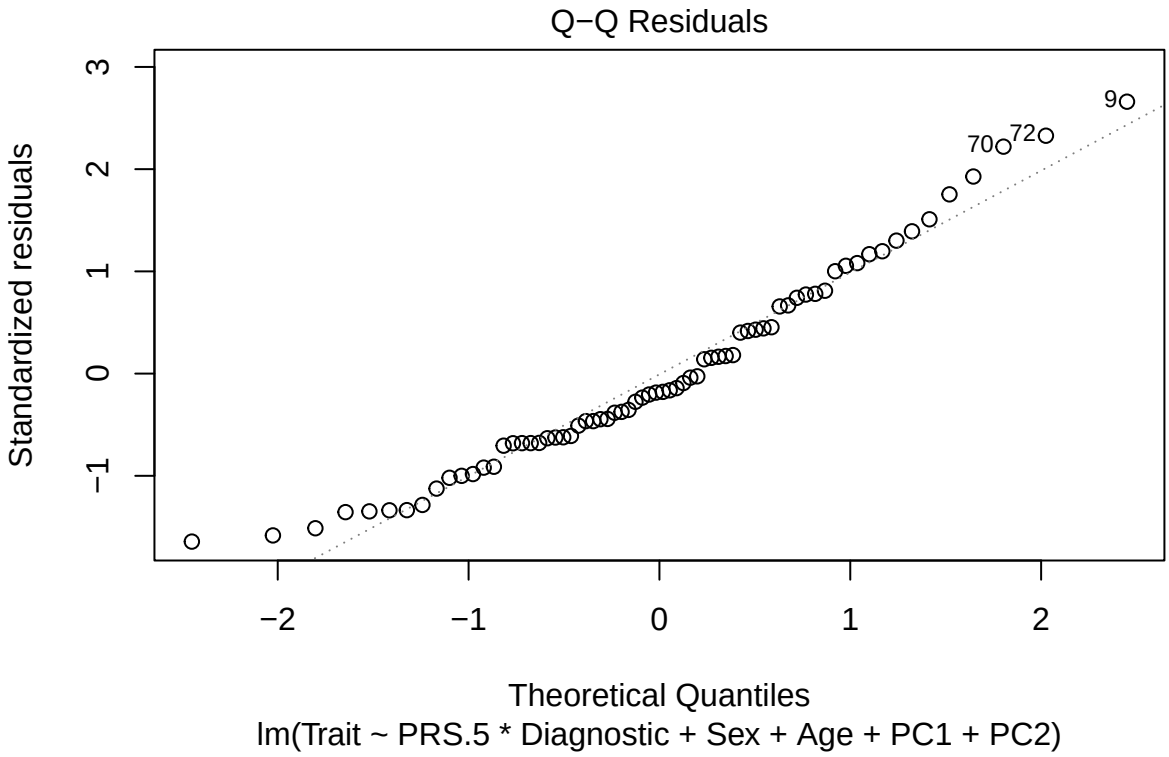
The plots suggest that the interaction between the PRS.5 and the diagnostic is relevant. Thus, we set the full model candidate (FM): $Trait \sim PRS + Sex + Diagnostic + Sex + PRS \cdot Diagnostic + PC1 + PC2$.

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.5*Diagnostic + Sex + Age + PC1 + PC2, data=dat)
#qq-plot for normality
plot(FM,2)
```

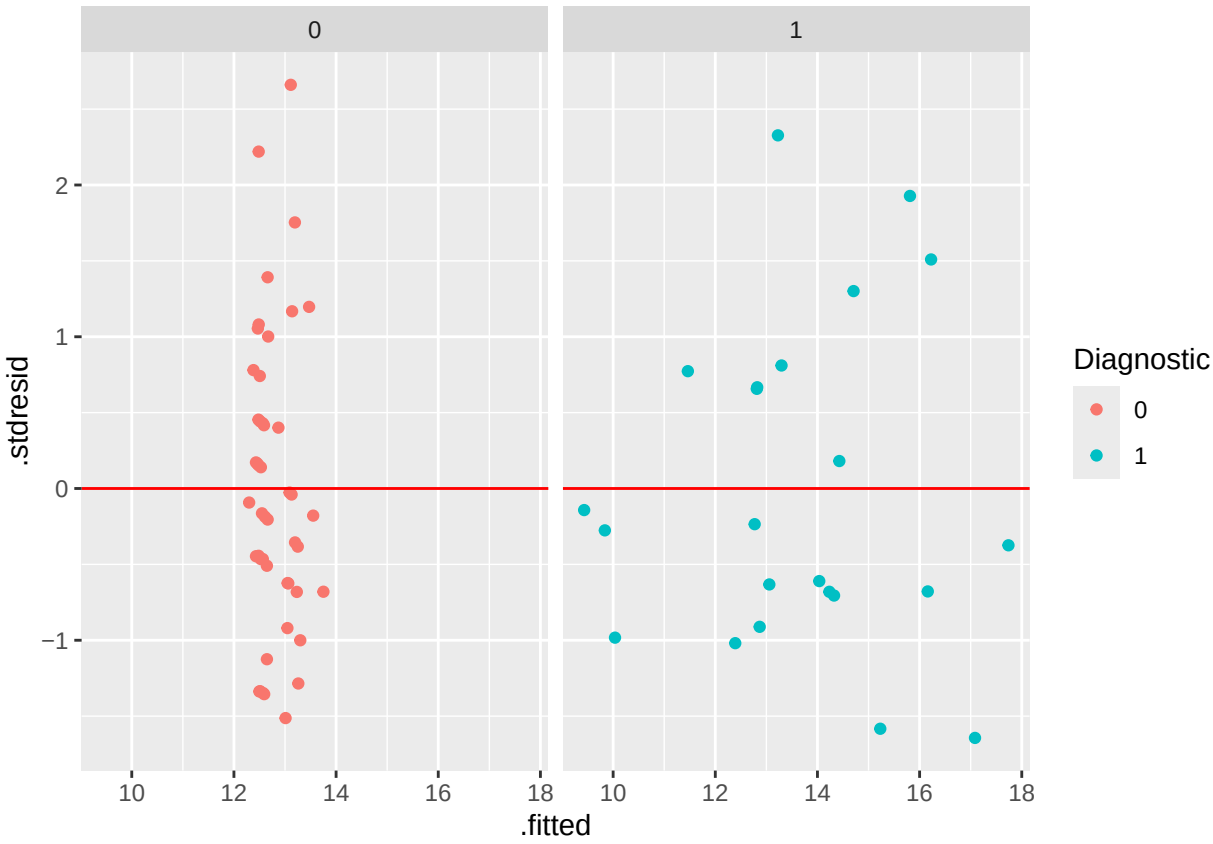


Lack of normality of residuals is suggested and this is supported by Shapiro's test.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  FM$residuals
## W = 0.96257, p-value = 0.03475
```

```
#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  0.4042 0.5271
##       68
```

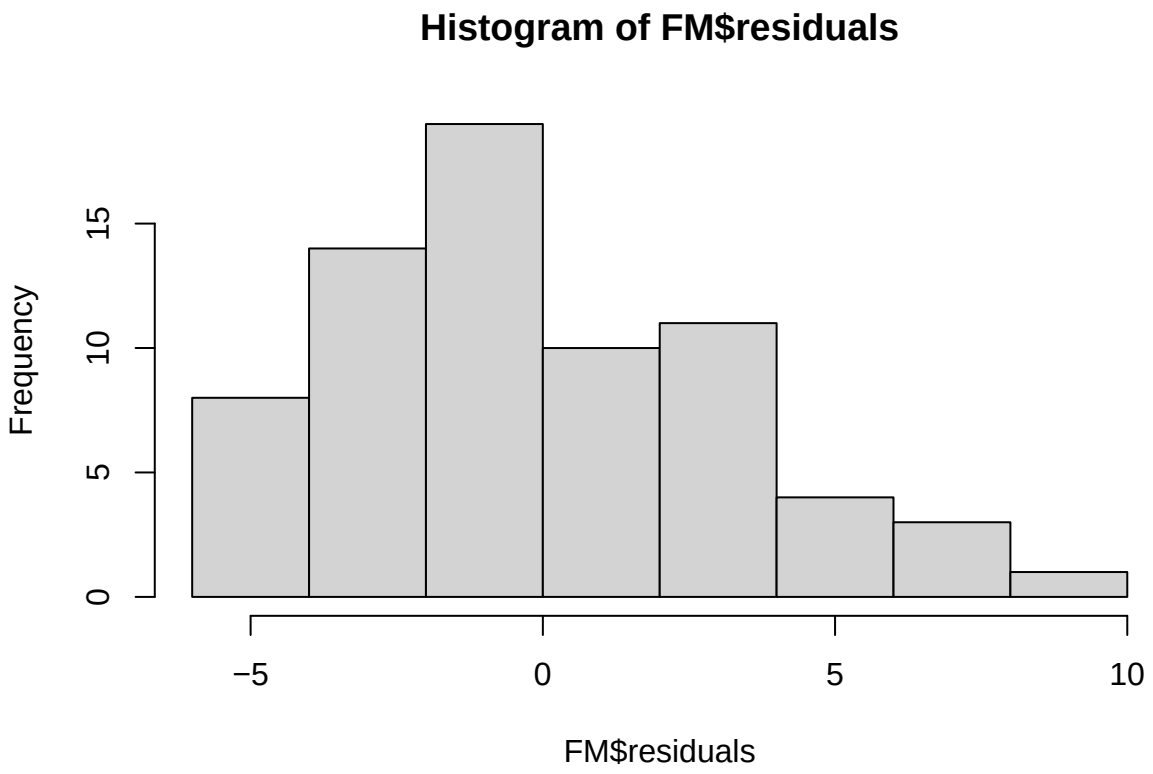
The figure does not show problems with the residuals, although it can be seen that for Diagnostic group 1 their variability increases as the fitted values do so, suggesting possible problems with the homogeneity of variances. However, Levene's test does not indicate evidence against homoscedasticity, neither a trend is observed that could indicate a lack of linearity.

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

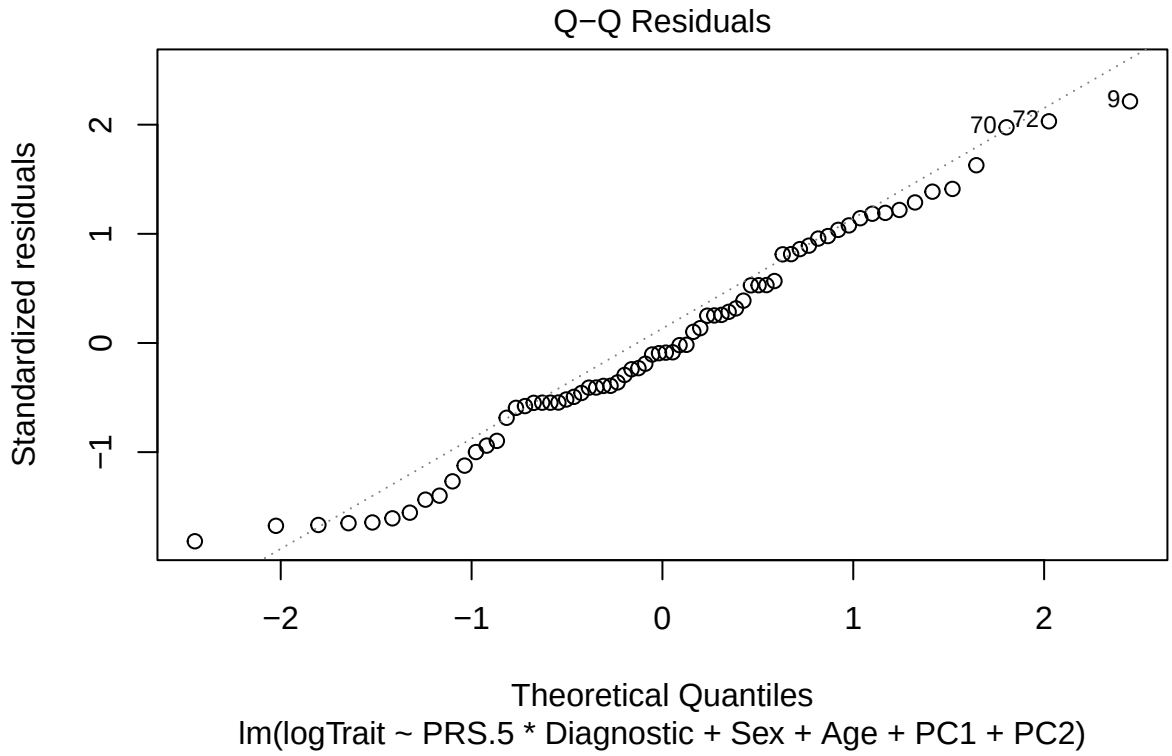
We have two approaches to assess the possible association with PRS.5 and the Trait: try a transformation of the trait or perform a permutation test.

First, we try a transformation. Given the shape of the density of residuals given by the next figure, the logarithmic transformation seems adequate to achieve normality.

```
#Histogram of the residuals
hist(FM$residuals)
```




```
dat$logTrait <- log(dat$Trait)
FM <- lm(logTrait ~ PRS.5*Diagnostic + Sex + Age + PC1 + PC2, data=dat)
#qq-plot for normality
plot(FM,2)
```

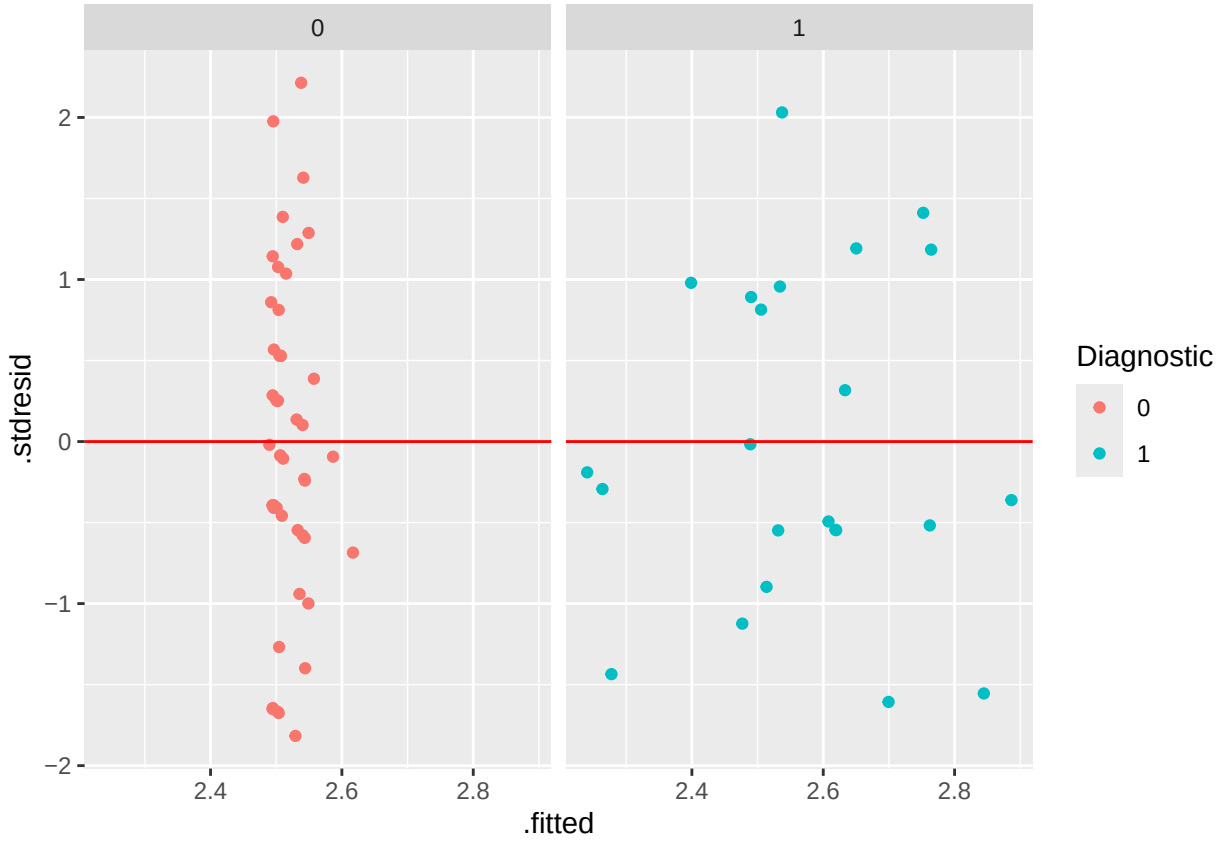


```
#Shapiro-Wilk test
shapiro.test(FM$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  FM$residuals
## W = 0.97729, p-value = 0.2331
```

Thus, 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=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  0.1157 0.7348
##       68
```

it does not indicate evidence against homoscedasticity, neither a trend is observed that could indicate a lack of linearity.

Then, the model we build is given by:

```r
summary(FM)
```

```
##
## Call:
## lm(formula = logTrait ~ PRS.5 * Diagnostic + Sex + Age + PC1 +
##     PC2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44983 -0.13453 -0.02147  0.20389  0.55307
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.585e+00  4.118e+00   0.628   0.5325
## PRS.5               8.892e-04  5.365e-02   0.017   0.9868
## Diagnostic1        -2.005e+01  8.027e+00  -2.498   0.0152 *
## Sex1                4.141e-02  6.937e-02   0.597   0.5527
## Age                -6.155e-04  6.528e-03  -0.094   0.9252
## PC1                 1.971e-01  4.669e-01   0.422   0.6744
## PC2                -5.958e-02  4.440e-01  -0.134   0.8937
## PRS.5:Diagnostic1 -2.607e-01  1.042e-01  -2.502   0.0150 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2581 on 62 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.1518, Adjusted R-squared:  0.05609
## F-statistic: 1.586 on 7 and 62 DF,  p-value: 0.1564
```

The results show that PRS.5 is related with the log(Trait) in the following way:

- $\widehat{log(Trait)} = 2.585 + 0.00089 \times PRS.5 + 0.041 \times Sex - 0.0006 \times Age + 0.197 \times PC1 - 0.0596 \times PC2$, if Diagnostic $= 0$;
- $\widehat{log(Trait)} = (2.585 - 20.052) + (0.00089 - 0.26068) \times PRS.5 + 0.041 \times Sex + -0.0006 \times Age + 0.197 \times PC1 - 0.0596 \times PC2$, if Diagnostic $= 1$;

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.

If the objective is to evaluate the possible association between Trait and PRS.5, it can be interesting to check whether the respective PRS.4 coefficients under each Diagnostic group are considerable or not.

```r
summary(glht(FM, "PRS.5 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = logTrait ~ PRS.5 * Diagnostic + Sex + Age + PC1 +
##     PC2, data = dat)
##
## Linear Hypotheses:
##             Estimate Std. Error t value Pr(>|t|)
## PRS.5 == 0 0.0008892  0.0536465   0.017    0.987
## (Adjusted p values reported -- single-step method)
```

```r
summary(glht(FM, "PRS.5  + PRS.5:Diagnostic1 = 0"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = logTrait ~ PRS.5 * Diagnostic + Sex + Age + PC1 +
##     PC2, data = dat)
##
## Linear Hypotheses:
##                                 Estimate Std. Error t value Pr(>|t|)
## PRS.5 + PRS.5:Diagnostic1 == 0 -0.25979    0.08808  -2.949  0.00449 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

This means that for Diagnostic 1 the association is significant (p-value= 0.004487) and negative and if PRS increases in one unit while keeping the other predictors constant, the change in the Trait is obtained multiplying by coeff = $\exp(-0.25979) = 0.771$; therefore it will decrease. For Diagnostic 0 there is no significant association, and if PRS increases in one unit while keeping the other predictors constant, the expected change in the Trait is obtained multiplying by coeff = $\exp(0.00089) = 1.001$; thus no change is expected. The ratio $\widehat{log(Trait)}|_{PRS.5=1}/\widehat{log(Trait)}|_{PRS.5=0} = \exp(-0.25979) = 0.771 < 1$, thus, gives a mean decrease of about 22.9%, for any given value (PRS.5= prs value).

On the other hand, with the permutation approach:

```r
NM <- lm(Trait ~ Diagnostic + Sex + Age + PC1 + PC2, data=dat)
FM <- lm(Trait ~ PRS.5*Diagnostic + Sex + Age + PC1 + PC2, data=dat)
outperm <- dR2(NullModel=NM, FullModel=FM, B=1000, seed=165) # Note that this function
# is included in the customized file via source("Functions.R")
outperm
```

```
## $dR2
## [1] 0.1078214
##
## $pvalue
## [1] 0.014
```

We observe an increase of 0.108 in the coefficient of determination when the PRS.5 is included in the model and the permutation test indicates it is significant.

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