

Esame_Scritto_27/04/2020

Luca Testa 816000

27/4/2020

```
library(readr)
library(car)

## Loading required package: carData

library(sjstats)

## Registered S3 methods overwritten by 'lme4':
##   method                                from
##   cooks.distance.influence.merMod      car
##   influence.merMod                     car
##   dfbeta.influence.merMod              car
##   dfbetas.influence.merMod             car

library(plotrix)
library(sjPlot)
library(sjmisc)
library(lme4)

## Loading required package: Matrix

library(pander)
library(car)
library(olsrr)

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##   rivers

library(systemfit)

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##   as.Date, as.Date.numeric
```

```

##
## Please cite the 'systemfit' package as:
## Arne Henningsen and Jeff D. Hamann (2007). systemfit: A Package for Estimating
## Systems of Simultaneous Equations in R. Journal of Statistical Software 23(4), 1-
## 40. http://www.jstatsoft.org/v23/i04/.
##
## If you have questions, suggestions, or comments regarding the 'systemfit'
## package, please use a forum or 'tracker' at systemfit's R-Forge site:
## https://r-forge.r-project.org/projects/systemfit/

library(het.test)

## Loading required package: vars

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##      cement

## Loading required package: strucchange

## Loading required package: sandwich

## Loading required package: urca

library(ppcor)
library(snakecase)
library(performance)

##
## Attaching package: 'performance'

## The following objects are masked from 'package:sjstats':
##
##      icc, r2

library(glmmTMB)

panderOptions('knitr.auto.asis', FALSE)

##-- White test function
white.test <- function(lmod, data=d){
  u2 <- lmod$residuals^2
  y <- fitted(lmod)
  Ru2 <- summary(lm(u2 ~ y + I(y^2)))$r.squared
  LM <- nrow(data)*Ru2
  p.value <- 1-pchisq(LM, 2)
  data.frame("Test statistic"=LM, "P value"=p.value)
}

```

```

# funzione per ottenere osservazioni outlier univariate
FIND_EXTREME_OBSERVATION <- function(x,sd_factor=2){
  which(x>mean(x)+sd_factor*sd(x) | x<mean(x)-sd_factor*sd(x))
}

# import del dataset
d <- read_delim("C:/Users/lucat/Desktop/PISA_USA.csv",
  ";", escape_double = FALSE, trim_ws = TRUE)

## Parsed with column specification:
## cols(
##   SCHOOLID = col_double(),
##   StIDStd = col_double(),
##   AGE = col_double(),
##   SCHOOL_SIZE = col_double(),
##   SOCIO_ECONOMIC_STATUS = col_double(),
##   SCHOOL_TYPE = col_double(),
##   JOYREAD = col_double(),
##   SCHOOL_SELECTIVITY = col_double(),
##   STUDENT_TEACHER_RATIO = col_double()
## )

# vista delle prima 6 righe del dataset
head(d)

## # A tibble: 6 x 9
##   SCHOOLID StIDStd   AGE SCHOOL_SIZE SOCIO_ECONOMIC_~ SCHOOL_TYPE JOYREAD
##   <dbl>   <dbl> <dbl>         <dbl>         <dbl>         <dbl>   <dbl>
## 1       1       1  15.8           808           0.183           1   -1.16
## 2       1       2  15.3           808           0.976           1    0.674
## 3       1       3   16           808          -0.474           1   -1.30
## 4       1       4  15.7           808          -0.574           1  -0.802
## 5       1       5  15.4           808           0.162           1   -1.45
## 6       1       7  15.9           808          -0.916           1   -2.45
## # ... with 2 more variables: SCHOOL_SELECTIVITY <dbl>,
## #   STUDENT_TEACHER_RATIO <dbl>

# controllo tipo di ogni colonna del dataset
str(d)

## tibble [5,101 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ SCHOOLID      : num [1:5101] 1 1 1 1 1 1 1 1 1 1 ...
##  $ StIDStd       : num [1:5101] 1 2 3 4 5 7 8 9 10 11 ...
##  $ AGE           : num [1:5101] 15.8 15.3 16 15.7 15.4 ...
##  $ SCHOOL_SIZE   : num [1:5101] 808 808 808 808 808 808 808 808 808 808
##  ...
##  $ SOCIO_ECONOMIC_STATUS: num [1:5101] 0.183 0.976 -0.474 -0.574 0.162 ...
##  $ SCHOOL_TYPE    : num [1:5101] 1 1 1 1 1 1 1 1 1 1 ...
##  $ JOYREAD        : num [1:5101] -1.162 0.674 -1.297 -0.802 -1.446 ...
##  $ SCHOOL_SELECTIVITY : num [1:5101] 2 2 2 2 2 2 2 2 2 2 ...
##  $ STUDENT_TEACHER_RATIO: num [1:5101] 18.4 18.4 18.4 18.4 18.4 ...

```

```
## - attr(*, "spec")=
## .. cols(
## .. SCHOOLID = col_double(),
## .. StIDStd = col_double(),
## .. AGE = col_double(),
## .. SCHOOL_SIZE = col_double(),
## .. SOCIO_ECONOMIC_STATUS = col_double(),
## .. SCHOOL_TYPE = col_double(),
## .. JOYREAD = col_double(),
## .. SCHOOL_SELECTIVITY = col_double(),
## .. STUDENT_TEACHER_RATIO = col_double()
## .. )
```

#-- vettore di variabili numeriche presenti nei dati per attuare statistiche descrittive

```
VAR_NUMERIC <- c("JOYREAD", "SOCIO_ECONOMIC_STATUS", "AGE", "SCHOOL_SELECTIVITY")
```

#-- STATISTICHE DESCRITTIVE

Si propongono la matrice di correlazione tra le variabili e alcune descrittive di base.

Qui posso notare come le variabili sembrano essere tutte quasi simmetriche

```
summary(d[,VAR_NUMERIC])
```

##	JOYREAD	SOCIO_ECONOMIC_STATUS	AGE	SCHOOL_SELECTIVITY
## Min.	:-3.22650	Min. :-3.4232	Min. :15.25	Min. :1.000
## 1st Qu.:	-0.69090	1st Qu.: -0.4524	1st Qu.:15.58	1st Qu.:1.000
## Median	:-0.15330	Median : 0.1811	Median :15.83	Median :2.000
## Mean	:-0.04298	Mean : 0.1559	Mean :15.79	Mean :1.788
## 3rd Qu.:	0.54210	3rd Qu.: 0.8235	3rd Qu.:16.08	3rd Qu.:3.000
## Max.	: 3.49500	Max. : 2.8914	Max. :16.33	Max. :3.000

Non sembrano esserci correlazioni di nessun tipo, tutti i valori sono molto bassi

Niente fa presagire che ci siano delle variabili collineari tra di loro, ovvero che

due o più variabili siano correlate tra di loro andando a formare una matrice X'X singolare

La cosa, appunto, è esclusa

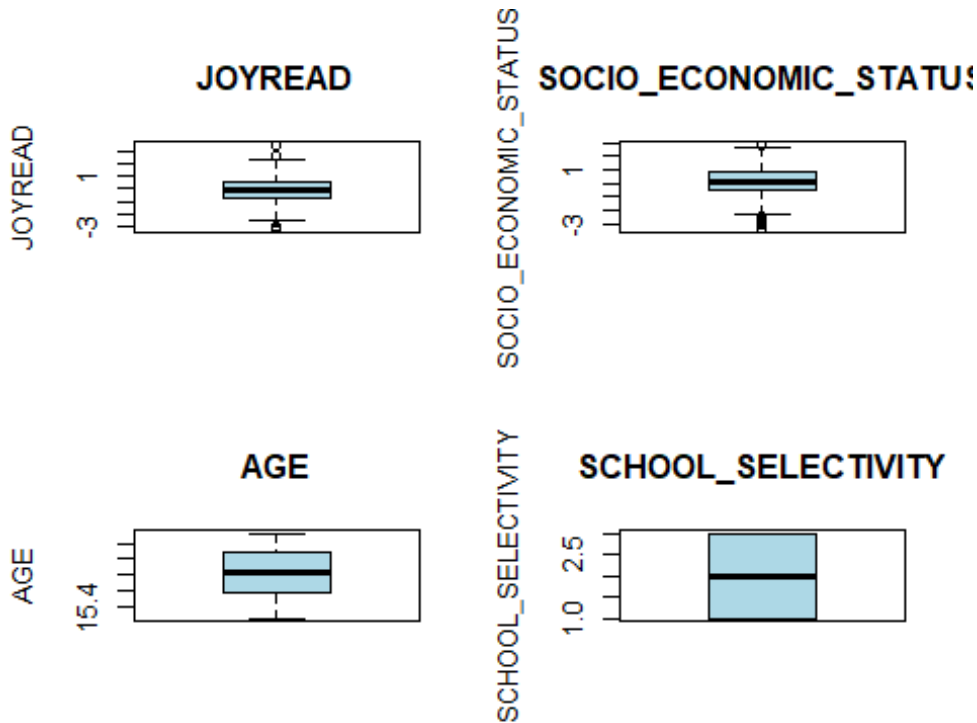
```
cor(d[,VAR_NUMERIC])
```

##	JOYREAD	SOCIO_ECONOMIC_STATUS	AGE
## JOYREAD	1.00000000	0.182777872	0.010281832
## SOCIO_ECONOMIC_STATUS	0.18277787	1.000000000	0.007714375
## AGE	0.01028183	0.007714375	1.000000000
## SCHOOL_SELECTIVITY	0.01315069	0.085757743	-0.003058384
## SCHOOL_SELECTIVITY	0.013150691	0.085757743	-0.003058384
## JOYREAD	0.013150691	0.085757743	-0.003058384
## SOCIO_ECONOMIC_STATUS	0.085757743	1.000000000	-0.003058384
## AGE	-0.003058384	-0.003058384	1.000000000
## SCHOOL_SELECTIVITY	1.000000000	-0.003058384	-0.003058384

Le variabili sembrano essere tutte molto vicine alla simmetria, soltanto le prime due hanno una coda

più pesante a destra e a sinistra e segnalano la presenza di outlier

```
par(mfrow=c(2,2))
for(i in VAR_NUMERIC){
  boxplot(d[,i],main=i,col="lightblue",ylab=i)
}
```



REGRESSIONE

*# Si utilizzano come variabili esplicative rispetto a JOYREAD:
SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELECTIVITY.*

#-- R CODE

Stima modello Lineare semplice

```
mod1 <- lm(JOYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELECTIVITY, d)
```

*# L'unica variabile che sembra essere significativa per questo modello è
"SOCIO_ECONOMIC_STATUS".*

*# Il modello ha un fitting veramente basso, quasi insignificante anche se, il p-
value mostra che il modello*

viene spiegato bene, si prosegue nei vari test.

```
summary(mod1)
```

```
##
```

```
## Call:
```

```
## lm(formula = JOYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELECTIVITY,
```

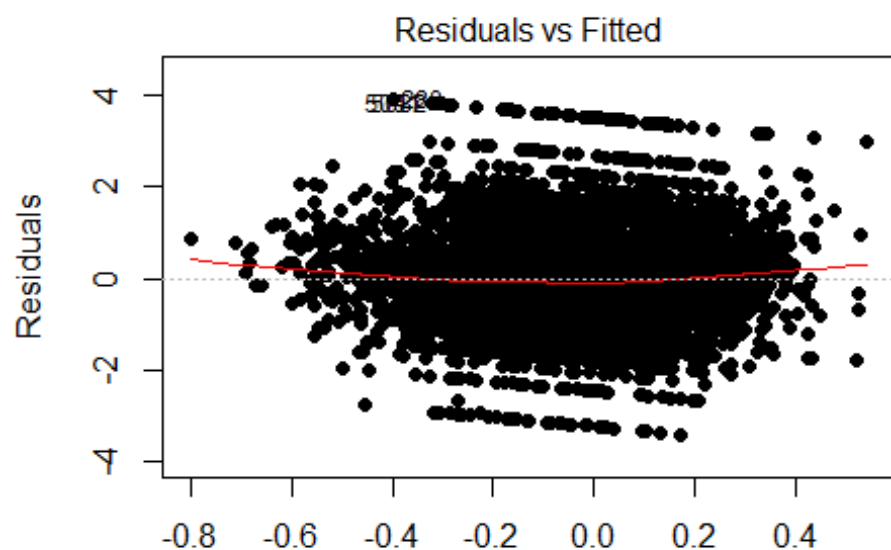
```
## data = d)
```

```
##
```

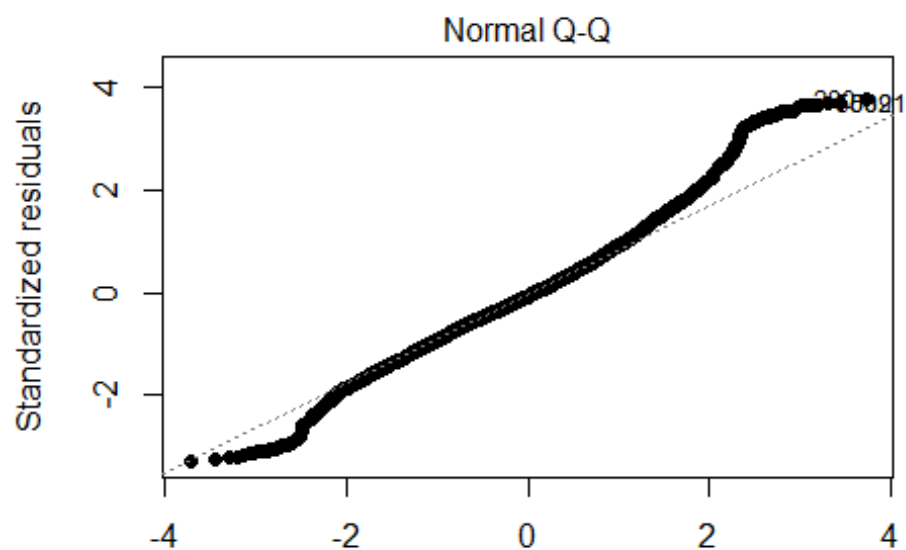
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.3959 -0.6349 -0.0679  0.5809  3.8948
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.567473    0.773875  -0.733    0.463
## SOCIO_ECONOMIC_STATUS  0.209622    0.015839  13.235 <2e-16 ***
## AGE            0.031500    0.048942   0.644    0.520
## SCHOOL_SELECTIVITY -0.003187    0.017553  -0.182    0.856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.035 on 5097 degrees of freedom
## Multiple R-squared:  0.03349,    Adjusted R-squared:  0.03292
## F-statistic: 58.88 on 3 and 5097 DF,  p-value: < 2.2e-16
```

*# Il primo plot sembra segnalare una leggera eteroschedasticità dovuta alla non completa ottima distribuzione
dei residui. Il secondo grafico segnala che siamo in presenza di non normalità.
Il terzo grafico segnala un paio di outlier non del tutti però significativi,
verrà approfondita la questione tramite test ad hoc. Anche il quarto grafico
sembra segnalare la presenza di leverage poco significativi.*

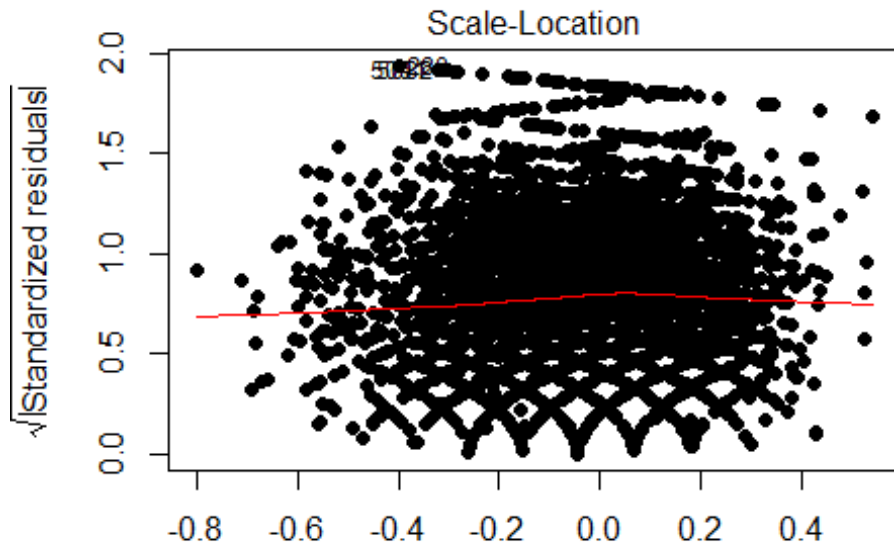
```
plot(mod1,pch=19)
```



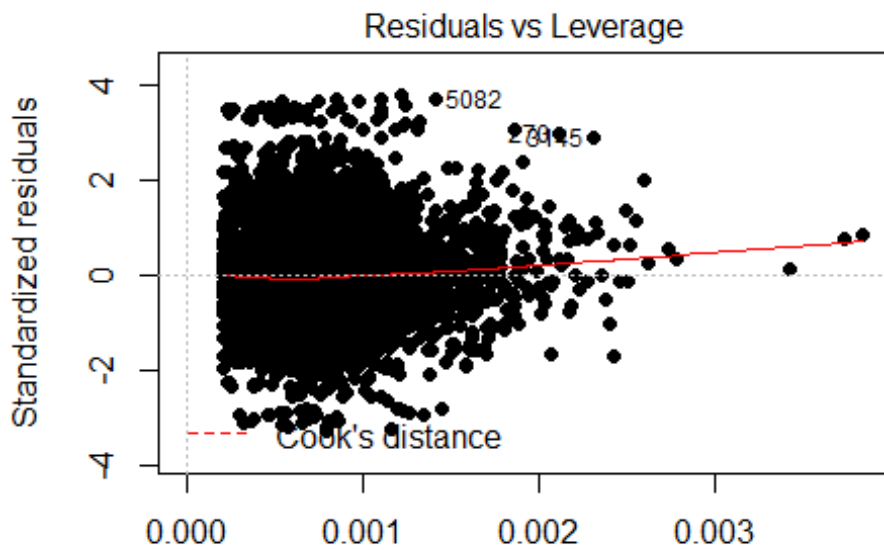
Fitted values
 OYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELE



Theoretical Quantiles
 OYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELE



Fitted values
OYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELE



Leverage
OYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + SCHOOL_SELE

Varianza multifattoriale o $Vif(x_j) = 1/Tol(x_j) = 1/(1 - R^2(x_j/x_1, \dots, x_{(j-1)}, x_{(j+1)}, \dots, x_p))$, ovvero il reciproco della tolleranza. Valori di tale indice variano tra 0 e 1 perciò se superiori a 20 indicano uno stretto rapporto tra la variabile considerata e le altre ovvero un eccessivo grado di multicollinearità. Anche oltre 10 andrebbero controllate

In questo modello non ne viene però segnalata la presenza.

```
ols_vif_tol(mod1)
```

```
##          Variables Tolerance      VIF
## 1 SOCIO_ECONOMIC_STATUS 0.9925820 1.007473
## 2                AGE 0.9999265 1.000073
## 3   SCHOOL_SELECTIVITY 0.9926318 1.007423
```

Questo test si basa sull'assunzione di omoschedasticità dei residui, viene perciò definita l'ipotesi nulla come $H_0 : \text{Var}(e_i) = \sigma^2$ e l'ipotesi alternativa come $H_1 : \text{Var}(e_i) = \sigma_i^2$. Il test sfrutta la regressione OLS del quadrato dei residui e_i^2 sui regressori x_j , i regressori al quadrato x_j^2 e le loro interazioni. # Con un valore pari a 0.08 possiamo dire di avere una leggera eteroschedasticità da migliorare tramite una stima WLS, per essere sicuri dell'omoschedasticità, come segnalato durante le lezioni, il valore dev'essere superiore o uguale a 0.1.

```
white.test(mod1)
```

```
##      Test.statistic      P.value
## 1           5.036008 0.08062036
```

L'ipotesi nulla è: $H_0: \rho = \text{Corr}[\varepsilon_i, \varepsilon_{(i-1)}] = 0$

Con valori della statistica compresi tra 1 e 3 si può dire di avere incorrelazione.

Nel caso di autocorrelazione, il teorema di Aitken stabilisce che nella classe degli stimatori lineari per il modello # di regressione generalizzato lo stimatore GLS è efficiente in quanto caratterizzato dalla minima varianza.

Il valore di 1.9486 ci segnala la presenza di incorrelazione tra i residui.

```
dwtest(mod1)
```

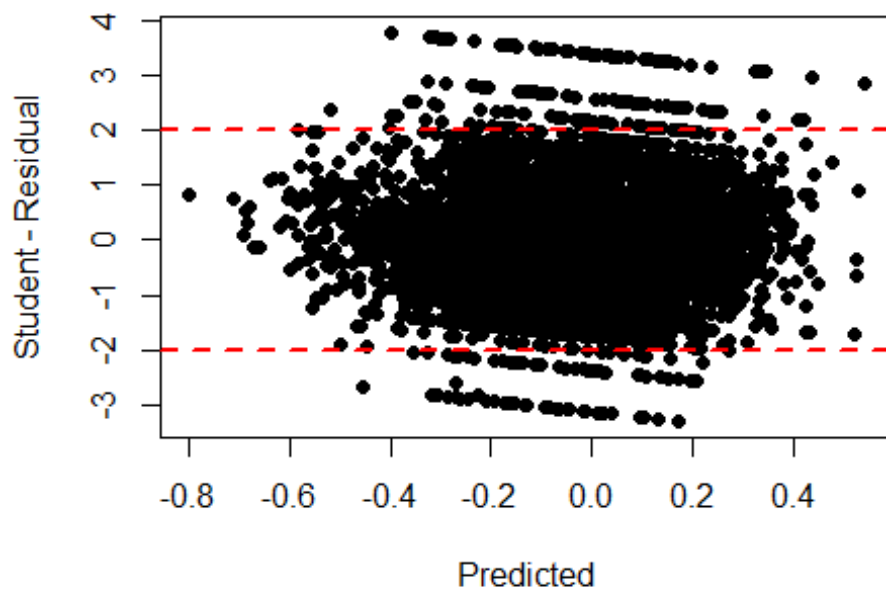
```
##
## Durbin-Watson test
##
## data:  mod1
## DW = 1.9486, p-value = 0.03198
## alternative hypothesis: true autocorrelation is greater than 0
```

I test seguenti sono stati fatti per vedere la quantità di outlier presenti all'interno del modello.

I grafici sottolineano un'elevata presenza di outlier.

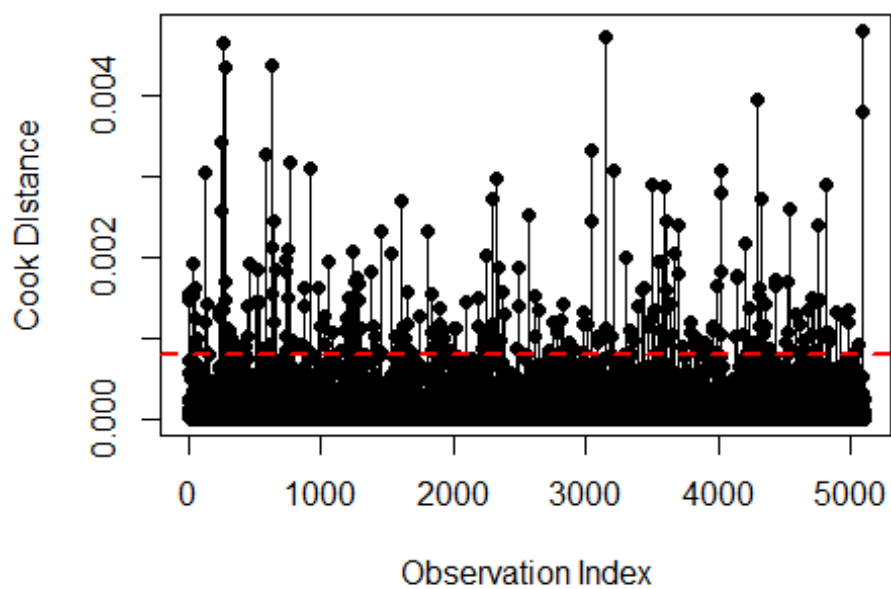
Residui studentizzati: è la versione dei residui standardizzati ma relativamente al campione. Di conseguenza le forme analitiche saranno le medesime facendo però riferimento non alla varianza σ^2 ma alla varianza campionaria s^2 .

```
{plot(fitted(mod1), rstudent(mod1), pch=19, xlab="Predicted", ylab="Student -
Residual")
abline(h=-2,col=2,lty=2,lwd=2)
abline(h=2,col=2,lty=2,lwd=2)}
```

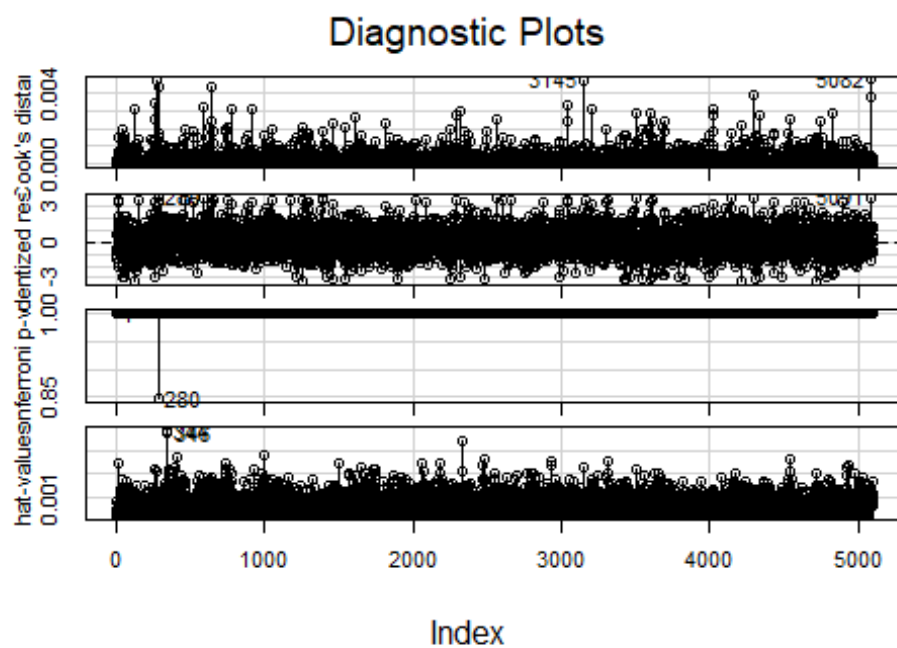


Misura l'influenza dell'i-esima osservazione sulla stima dei coefficienti di regressione nel loro complesso, in termini di capacità del modello di predire tutti i casi quando la singola osservazione viene rimossa dal dataset. Valori superiori a 1 (o eventualmente a $4/n$, essendo n il numero di osservazioni) indicano che il punto è influente.

```
{plot(cooks.distance(mod1), pch=19, xlab="Observation Index", ylab="Cook
DDistance", type="h")
points(cooks.distance(mod1), pch=19)
abline(h=4/nrow(d), col=2, lty=2, lwd=2)}
```



```
influenceIndexPlot(mod1)
```



```
# REGRESSION MULTILEVEL: Empty Model
```

```
# Modello ANOVA ad effetti casuali (empty model)
```

Modello

```
mod2 <- lmer(JOYREAD ~ (1 | SCHOOLID), d, REML=T)
summary(mod2)
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: JOYREAD ~ (1 | SCHOOLID)
```

```
## Data: d
```

```
##
```

```
## REML criterion at convergence: 14961.7
```

```
##
```

```
## Scaled residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -3.2398 -0.6292 -0.0860  0.5800  3.6265
```

```
##
```

```
## Random effects:
```

```
## Groups   Name      Variance Std.Dev.
```

```
## SCHOOLID (Intercept) 0.03314  0.182
```

```
## Residual              1.07561  1.037
```

```
## Number of obs: 5101, groups: SCHOOLID, 165
```

```
##
```

```
## Fixed effects:
```

```
##              Estimate Std. Error t value
```

```
## (Intercept) -0.04073    0.02050  -1.987
```

```
mod1_null <- lm(JOYREAD ~ 1, d)
```

```
anova(mod2, mod1_null)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: d
```

```
## Models:
```

```
## mod1_null: JOYREAD ~ 1
```

```
## mod2: JOYREAD ~ (1 | SCHOOLID)
```

```
##      npar    AIC    BIC logLik deviance Chisq Df Pr(>Chisq)
```

```
## mod1_null    2 15005 15018 -7500.7    15001
```

```
## mod2         3 14962 14981 -7477.9    14956 45.555  1 1.484e-11 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
performance::icc(mod2)
```

```
## # Intraclass Correlation Coefficient
```

```
##
```

```
##      Adjusted ICC: 0.030
```

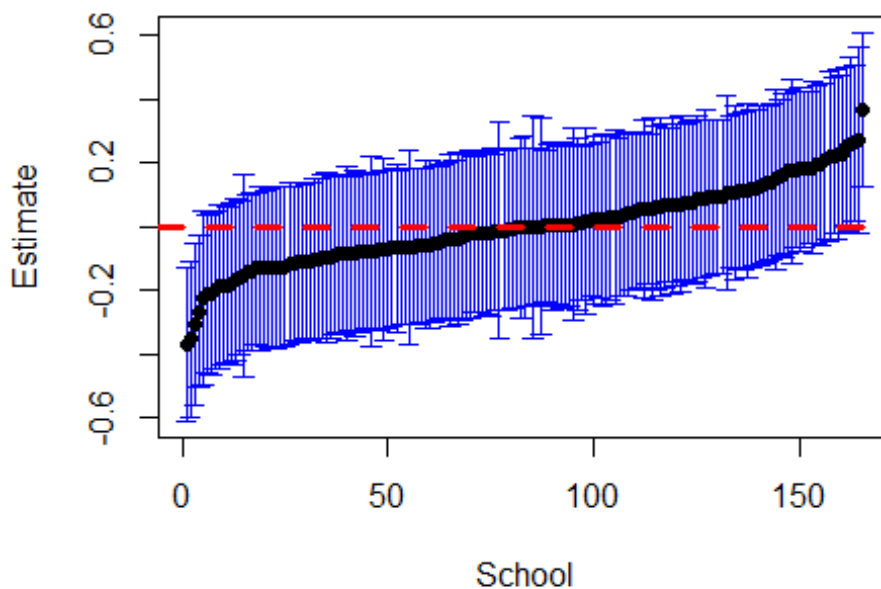
```
##      Conditional ICC: 0.030
```

```
res <- plot_model(mod2, type = "re", grid = FALSE,
show.values=T,title="T",prnt.plot=F, sort.est = "sort.all" )
```

```
data_1 =res$data[order(res$data$estimate),]
```

```
{plotCI(1:nrow(data_1),data_1$estimate,ui=data_1$conf.high, li=data_1$conf.low
,pch=19,scol="blue",xlab="School",ylab="Estimate")
```

```
abline(h=mean(data_1$estimate),col=2,lwd=3,lty=2)}
```



E' respinta l'ipotesi che il modello non interpreti i dati e dal rapporto tra varianza spiegata e totale si ricava
un coefficiente intraclasse pari a 0.03 che è molto basso, e segnala una bassa variabilità fra le scuole nei punteggi
di JOYREAD. Si propongono poi i valori attesi e gli intervalli di confidenza dei parametri casuali inerenti le
singole scuole. Per i parametri casuali il modello postula graduatorie basate su valori attesi e intervalli di
confidenza.
Come è noto una scuola A può ritenersi superiore a una scuola B in termini di efficacia solo se l'estremo
inferiore dell'intervallo di confidenza di A sia superiore all'estremo superiore dell'intervallo di confidenza di B.
Dal grafico posso anche affermare che ci sono poche scuole significativamente diverse da 0, cioè che, non hanno nel loro intervallo di confidenza appunto lo 0.

Modello ANOVA ad effetti casuali detto anche empty model: $y_{ij} = \lambda_0 + v_j + \epsilon_{ij}$; $\epsilon_{ij} \sim N(0, \sigma^2)$; $v_j \sim N(0, \tau^2)$
In questo caso la variabile dipendente y dipende dagli effetti casuali:
1) a livello di gruppo, V_j , distribuiti in modo normale $N(\lambda_0; \tau^2)$
2) a livello individuale, dai residui R_{ij} , distribuiti in modo normale $N(0; \sigma^2)$

REGRESSIONE MULTILEVEL: Random Intercept
Modello empty al quale si inserisce una variabile esplicativa (random intercept model - mixed model)
Multilevel perchè tiene conto sia della parte di regressione che di analisi della varianza ed è la sintesi dei modelli empty e OLS

##-- R CODE

```
mod3 <- lmer(JOYREAD ~ SOCIO_ECONOMIC_STATUS + (1 | SCHOOLID), d, REML=T)
summary(mod3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: JOYREAD ~ SOCIO_ECONOMIC_STATUS + (1 | SCHOOLID)
## Data: d
##
## REML criterion at convergence: 14803.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.2794 -0.6220 -0.0655  0.5640  3.9140
##
## Random effects:
## Groups Name Variance Std.Dev.
## SCHOOLID (Intercept) 0.02847 0.1687
## Residual 1.04357 1.0216
## Number of obs: 5101, groups: SCHOOLID, 165
##
## Fixed effects:
## Estimate Std. Error t value
## (Intercept) -0.07320 0.01978 -3.701
## SOCIO_ECONOMIC_STATUS 0.21735 0.01680 12.938
##
## Correlation of Fixed Effects:
## (Intr)
## SOCIO_ECONO -0.126
```

```
Anova(mod3, type="III")
```

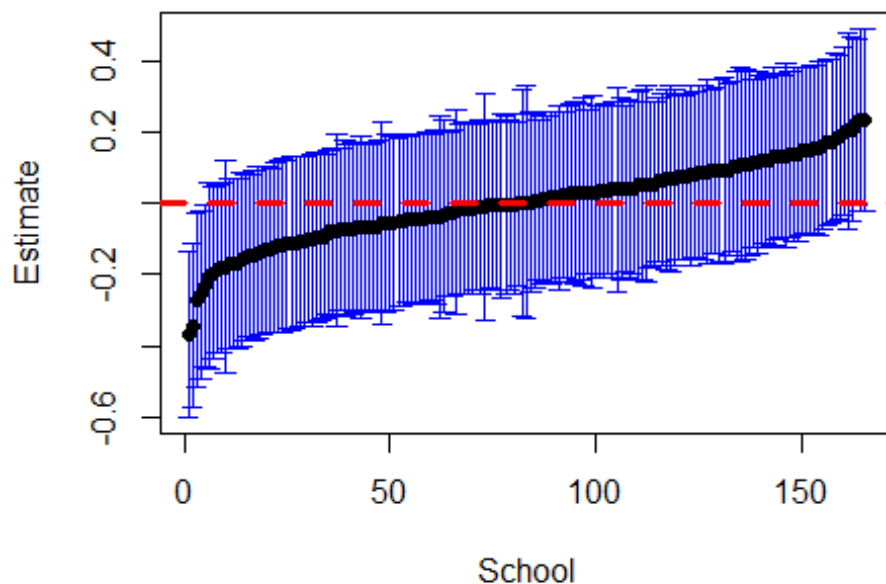
```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: JOYREAD
## Chisq Df Pr(>Chisq)
## (Intercept) 13.699 1 0.0002145 ***
## SOCIO_ECONOMIC_STATUS 167.386 1 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
performance::icc(mod3)
```

```
## # Intraclass Correlation Coefficient
##
## Adjusted ICC: 0.027
## Conditional ICC: 0.026
```

```
res <- plot_model(mod3, type = "re", grid = FALSE,
show.values=T,title="T",prnt.plot=F, sort.est = "sort.all" )
data_1 =res$data[order(res$data$estimate),]
{plotCI(1:nrow(data_1),data_1$estimate,ui=data_1$conf.high, li=data_1$conf.low
```

```
,pch=19,scol="blue",xlab="School",ylab="Estimate")
abline(h=mean(data_1$estimate),col=2,lwd=3,lty=2)}
```



Il coefficiente di correlazione intraclasse si abbassa di pochissimo (0.027), rimanendo veramente basso, in quanto si abbassano in uguale proporzione varianza spiegata e residua. Il modello interpreta bene i dati e la variabile SOCIO_ECONOMIC_STATUS risulta significativa. Anche il test di 3° tipo degli effetti fissi conferma questa significatività. Si propongono i valori attesi e gli intervalli di confidenza dei parametri casuali inerenti i gruppi. Si vede come il ranking non muti in modo rilevante al caso empty. Ciò mostra che la diversa distribuzione fra le scuole della variabile SOCIO_ECONOMIC_STATUS non è all'origine di parte della variabilità di "JOYREAD" attribuito in prima istanza nel modello empty alla efficacia delle scuole.

$y_{ij} = \lambda \theta_0 + \beta_1 x_{ij} + v_j + \epsilon_{ij}; \epsilon_{ij} \sim N(0, \sigma^2);$
 # Dove v_j è la determinazione della variabile casuale V_j distribuita normalmente $N(\lambda \theta_0; \tau^2)$ a rappresentazione dei residui di secondo livello. Essi sono indipendenti e quindi incorrelati con i residui di primo livello ϵ_{ij}
 determinazioni della variabile casuale normalmente distribuita $E_{ij} \sim N(0; \sigma^2)$

2° REGRESSIONE MULTILEVEL

```
mod4 <- lmer(JOYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + (1 | SCHOOLID), d, REML=T)
summary(mod4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: JOYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + (1 | SCHOOLID)
## Data: d
```

```

##
## REML criterion at convergence: 14807.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.2843 -0.6208 -0.0660  0.5624  3.9170
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
## SCHOOLID (Intercept) 0.02838  0.1685
## Residual              1.04380  1.0217
## Number of obs: 5101, groups: SCHOOLID, 165
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    -0.35084    0.77674  -0.452
## SOCIO_ECONOMIC_STATUS 0.21731    0.01680  12.936
## AGE              0.01758    0.04916   0.358
##
## Correlation of Fixed Effects:
##              (Intr) SOCIO_
## SOCIO_ECONO  -0.001
## AGE           -1.000 -0.003

```

Anova(mod4, type="III")

```

## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: JOYREAD
##              Chisq Df Pr(>Chisq)
## (Intercept)    0.2040  1    0.6515
## SOCIO_ECONOMIC_STATUS 167.3477  1    <2e-16 ***
## AGE              0.1279  1    0.7207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

performance::icc(mod4)

```

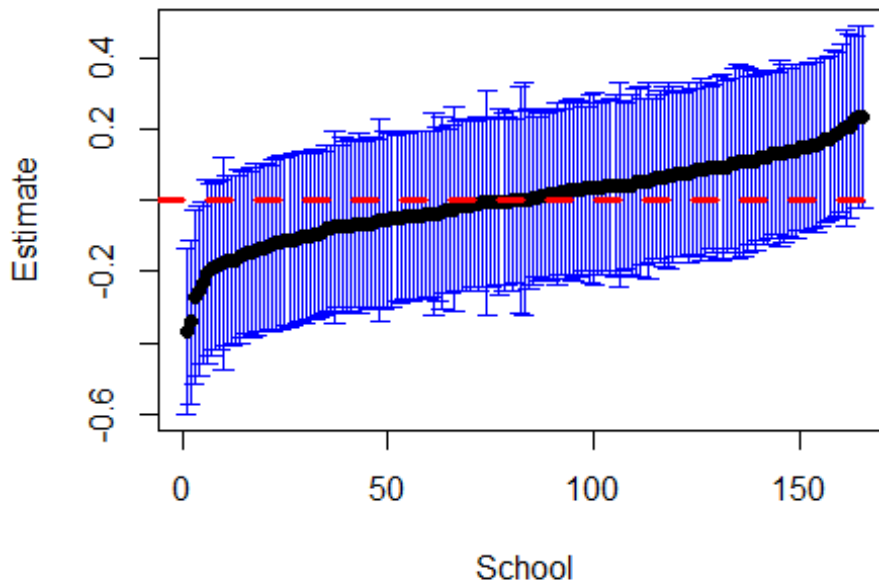
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.026
##      Conditional ICC: 0.026

```

```

res <- plot_model(mod4, type = "re", grid = FALSE,
show.values=T,title="T",prnt.plot=F, sort.est = "sort.all" )
data_1 =res$data[order(res$data$estimate),]
{plotCI(1:nrow(data_1),data_1$estimate,ui=data_1$conf.high, li=data_1$conf.low
,pch=19,scol="blue",xlab="School",ylab="Estimate")
abline(h=mean(data_1$estimate),col=2,lwd=3,lty=2)}

```

Il coefficiente di correlazione intraclasse si abbassa di pochissimo ancora (0.026). Il modello interpreta bene i dati e la variabile SOCIO_ECONOMIC_STATUS risulta significativa mentre la variabile AGE no con un valore di Chisq veramente basso, questo vuol dire che non "ruba" parte della variabilità. Si propongono i valori attesi e gli intervalli di confidenza dei parametri casuali inerenti i gruppi. Si vede come il ranking non muti in modo rilevante, anzi, peggiori.

3° REGRESSIONE MULTILEVEL

2 variabili di primo livello e 1 variabile di secondo 2 livello

#-- RIMUOVO L'OUTLIER -999

```
summary(d$STUDENT_TEACHER_RATIO)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -999.00   11.97   14.56  -84.67   18.02   40.36
```

```
d1 <- d[!(d$STUDENT_TEACHER_RATIO == -999),]
```

```
mod5 <- lmer(JOYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + STUDENT_TEACHER_RATIO + (1 | SCHOOLID), d1, REML=T)
```

```
summary(mod5)
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: JOYREAD ~ SOCIO_ECONOMIC_STATUS + AGE + STUDENT_TEACHER_RATIO +
```

```
##      (1 | SCHOOLID)
```

```
##      Data: d1
```

```
##
```

```
## REML criterion at convergence: 13374.9
```

```
##
```

```

## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.7399 -0.5606  0.0680  0.6197  3.2696
##
## Random effects:
##   Groups   Name                Variance Std.Dev.
## SCHOOLID (Intercept) 0.03167  0.178
## Residual              1.04837  1.024
## Number of obs: 4595, groups: SCHOOLID, 146
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)      0.069679   0.823224   0.085
## SOCIO_ECONOMIC_STATUS 0.222488   0.017946  12.398
## AGE               0.001177   0.051901   0.023
## STUDENT_TEACHER_RATIO -0.001724   0.004149  -0.415
##
## Correlation of Fixed Effects:
##              (Intr) SOCIO_ AGE
## SOCIO_ECONO  0.000
## AGE          0.996 -0.008
## STUDENT_TEA  0.089  0.049  0.008

Anova(mod5, type="III")

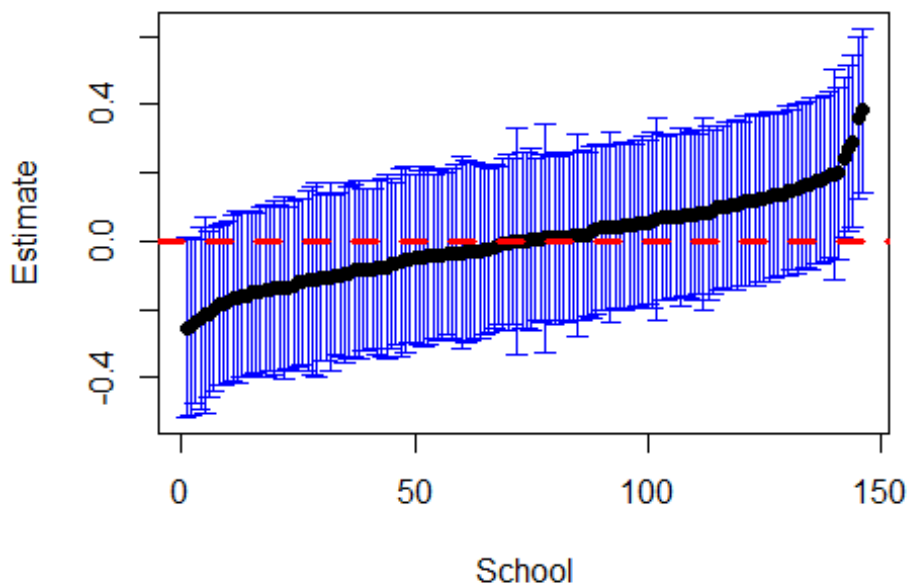
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: JOYREAD
##              Chisq Df Pr(>Chisq)
## (Intercept)      0.0072  1    0.9325
## SOCIO_ECONOMIC_STATUS 153.6993  1    <2e-16 ***
## AGE               0.0005  1    0.9819
## STUDENT_TEACHER_RATIO 0.1726  1    0.6778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

performance::icc(mod5)

## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.029
##      Conditional ICC: 0.028

res <- plot_model(mod5, type = "re", grid = FALSE,
show.values=T,title="T",prnt.plot=F, sort.est = "sort.all" )
data_1 =res$data[order(res$data$estimate),]
{plotCI(1:nrow(data_1),data_1$estimate,ui=data_1$conf.high, li=data_1$conf.low
,pch=19,scol="blue",xlab="School",ylab="Estimate")
abline(h=mean(data_1$estimate),col=2,lwd=3,lty=2)}

```



Il coefficiente di correlazione intraclasse rimane invariato (0.027), rimanendo veramente basso, in quanto si abbassano in uguale proporzione varianza spiegata e residua. Il modello interpreta bene i dati e la variabile SOCIO_ECONOMIC_STATUS risulta significativa, AGE no e nemmeno la variabile di secondo livello. Entrambe non catturano parte della variabilità. Si propongono i valori attesi e gli intervalli di confidenza dei parametri casuali inerenti i gruppi. Si vede come il ranking non muti in modo rilevante e come non sia possibile comunque stilare un ranking.

REGRESSIONE MULTILEVEL: Random Slope

Questo modello prevede che i coefficienti della regressione varino da gruppo a gruppo

```
mod6 <- lmer(JOYREAD ~ 1 + (SOCIO_ECONOMIC_STATUS | SCHOOLID), d, REML=T)
summary(mod6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: JOYREAD ~ 1 + (SOCIO_ECONOMIC_STATUS | SCHOOLID)
## Data: d
##
## REML criterion at convergence: 14904.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.2083 -0.6183 -0.0705  0.5529  3.6427
##
## Random effects:
##      Groups      Name                Variance Std.Dev. Corr
```

```
## SCHOOLID (Intercept)          0.02515  0.1586
##          SOCIO_ECONOMIC_STATUS 0.05057  0.2249  0.04
## Residual                    1.04252  1.0210
## Number of obs: 5101, groups: SCHOOLID, 165
##
## Fixed effects:
##          Estimate Std. Error t value
## (Intercept) -0.07561    0.02012  -3.757
```

```
Anova(mod6, type="III")
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
```

```
##
```

```
## Response: JOYREAD
```

```
##          Chisq Df Pr(>Chisq)
```

```
## (Intercept) 14.115  1  0.000172 ***
```

```
## ---
```

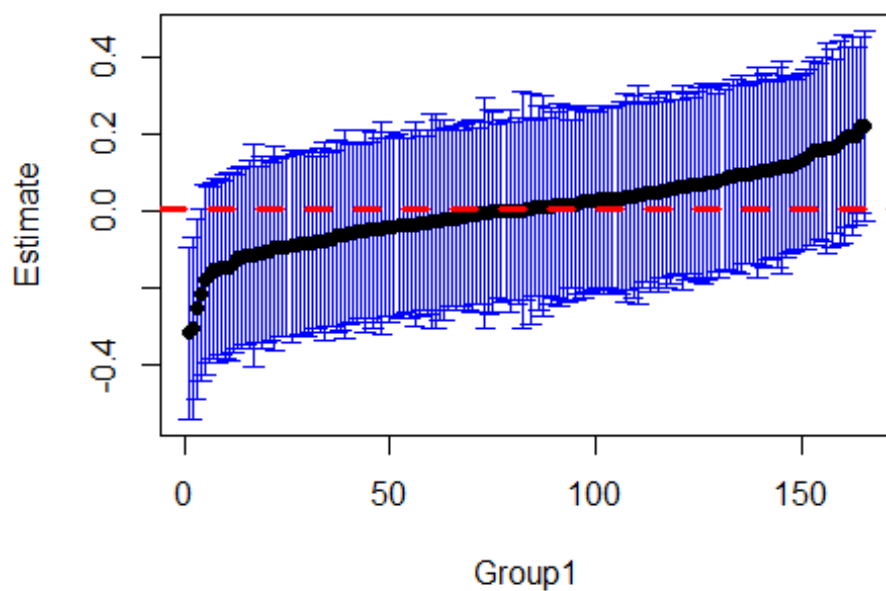
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res <- plot_model(mod6, type = "re", grid = FALSE,
show.values=T,title="T",prnt.plot=F, sort.est = "sort.all" )
```

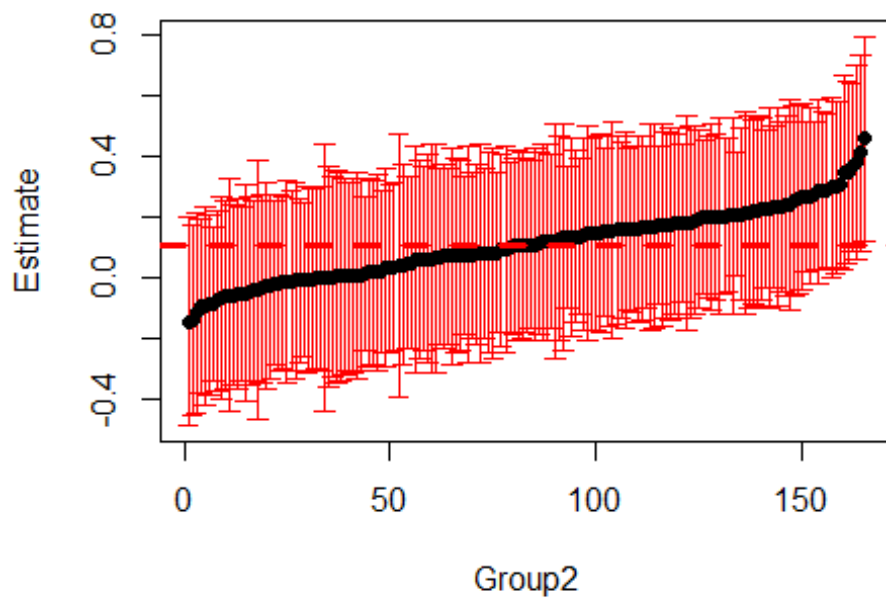
```
data_coef = res[[1]]$data[order(res[[1]]$data$estimate),]
```

```
data_int  = res[[2]]$data[order(res[[2]]$data$estimate),]
```

```
{plotCI(1:nrow(data_coef),data_coef$estimate,ui=data_coef$conf.high,
li=data_coef$conf.low ,pch=19,scol="blue",xlab="Group1",ylab="Estimate")
abline(h=mean(data_coef$estimate),col=2,lwd=3,lty=2)}
```



```
{plotCI(1:nrow(data_int),data_int$estimate,ui=data_int$conf.high,
li=data_int$conf.low ,pch=19,scol="red",xlab="Group2",ylab="Estimate")
abline(h=mean(data_int$estimate),col=2,lwd=3,lty=2)}
```



Il coefficiente intraclasse non può più essere calcolato nel modo semplice precedente perché si deve tener conto della correlazione tra effetti casuali relativi a "JOYREAD" e alle scuole nel loro complesso. Gli effetti casuali complessivi relativi all'efficacia delle scuole nel loro complesso e "JOYREAD" sono significativi. La correlazione tra effetti relativi alle scuole e a "JOYREAD" è leggermente positiva.

Sia per il coefficiente che per l'intercetta non possiamo stilare un ranking vero e proprio in quanto non siamo in presenza di differenze significativamente diverse fra le scuole.

Sono poche quelle (come la 157) ad essere significativamente diverse da 0.

Dato il modello random intercept si può proporre il modello random slope (total effects): $y_{ij} = \lambda_{00} + \beta_{1j} * x_{ij} + v_j + \epsilon_{ij}$ $y_{ij} = \lambda_{00} + (\beta_{1j} \theta + \beta_{1j})x_{ij} + v_j + \epsilon_{ij}$ ove i coefficiente β_{1j} ($j=1, \dots, p$) variano da gruppo a gruppo poiché l'effetto di X su Y è diverso nei diversi gruppi.

In questo caso non si è in presenza di un total effects.