MULTILEVEL 1 - Data set: CASCHOOL

INTRODUZIONE

Il data set contiene informazioni sulle performance dei test, sulle caratteristiche delle scuole e sulla situazione demografica di 420 studenti nei diversi distretti scolastici della California. Ci sono 14 variabili:

1. DISTRICT: codice del distretto
2. SCHOOL: nome della scuola
3. COUNTRY: nome della contea
4. GRADES: metodo di voto utilizzato nella contea
5. STUDENTS: totale degli studenti nella scuola
6. TEACHERS: totale degli insegnanti a tempo pieno
7. CALWORKS: percentuale di studenti che rientrano nel programma pubblico assistenziale CalWorks
8. LUNCH: percentuale di studenti che hanno diritto ad una riduzione sul prezzo del pranzo
9. COMPUTERS: numero di computer per classe
10. EXPENDITURE: spesa per studente
11. INCOME: reddito medio del distretto (migliaia di USD)
12. ENGLISH: percentuale di studenti per cui l'inglese è la seconda lingua
13. READ: punteggio megio nel test di lettura
14. MATH: punteggio megio nel test di matematica

Variabile dipendente: MATH

Analisi proposte:

1. Statistiche descrittive
2. Analisi multilevel

```r
#-- R CODE
library(car)
library(sjstats)
library(plotrix)
library(sjPlot)
library(sjmisc)
library(lme4)
library(pander)
library(car)
library(olsrr)
library(systemfit)
library(het.test)
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)
}
```

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

#-- import dei dati
ABSOLUTE_PATH <- "C:\\Users\\sbarberis\\Dropbox\\MODELLI STATISTICI"
d <- read.csv(paste0(ABSOLUTE_PATH,"\\esercizi (3) copia\\1.multilevel\\CASchools.txt"),sep=" ")

#-- vettore di variabili numeriche presenti nei dati
VAR_NUMERIC <- names(d)[6:ncol(d)]

#-- print delle prime 6 righe del dataset
pander(head(d),big.mark=",")
```

Table 1: Table continues below

id	district	school	county	grades	students
1	75,119	Sunol Glen Unified	Alameda	KK-08	195
2	61,499	Manzanita Elementary	Butte	KK-08	240
3	61,549	Thermalito Union Elementary	Butte	KK-08	1,550
4	61,457	Golden Feather Union Elementary	Butte	KK-08	243
5	61,523	Palermo Union Elementary	Butte	KK-08	1,335
6	62,042	Burrel Union Elementary	Fresno	KK-08	137

Table 2: Table continues below

teachers	calworks	lunch	computer	expenditure	income	english
10.9	0.5102	2.041	67	6,385	22.69	0
11.15	15.42	47.92	101	5,099	9.824	4.583
82.9	55.03	76.32	169	5,502	8.978	30
14	36.48	77.05	85	7,102	8.978	0
71.5	33.11	78.43	171	5,236	9.08	13.86
6.4	12.32	86.96	25	5,580	10.41	12.41

read	math
691.6	690
660.5	661.9
636.3	650.9
651.9	643.5
641.8	639.9
605.7	605.4

STATISTICHE DESCRITTIVE

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

```r
#-- R CODE
pander(summary(d[,VAR_NUMERIC]),big.mark=",") #-- statistiche descrittive
```

Table 4: Table continues below

students	teachers	calworks	lunch
Min. : 81.0	Min. : 4.85	Min. : 0.000	Min. : 0.00
1st Qu.: 379.0	1st Qu.: 19.66	1st Qu.: 4.395	1st Qu.: 23.28
Median : 950.5	Median : 48.56	Median :10.520	Median : 41.75
Mean : 2628.8	Mean : 129.07	Mean :13.246	Mean : 44.71
3rd Qu.: 3008.0	3rd Qu.: 146.35	3rd Qu.:18.981	3rd Qu.: 66.86
Max. :27176.0	Max. :1429.00	Max. :78.994	Max. :100.00

Table 5: Table continues below

computer	expenditure	income	english
Min. : 0.0	Min. :3926	Min. : 5.335	Min. : 0.000
1st Qu.: 46.0	1st Qu.:4906	1st Qu.:10.639	1st Qu.: 1.941
Median : 117.5	Median :5215	Median :13.728	Median : 8.778
Mean : 303.4	Mean :5312	Mean :15.317	Mean :15.768
3rd Qu.: 375.2	3rd Qu.:5601	3rd Qu.:17.629	3rd Qu.:22.970
Max. :3324.0	Max. :7712	Max. :55.328	Max. :85.540

read	math
Min. :604.5	Min. :605.4
1st Qu.:640.4	1st Qu.:639.4
Median :655.8	Median :652.5
Mean :655.0	Mean :653.3
3rd Qu.:668.7	3rd Qu.:665.9
Max. :704.0	Max. :709.5

```r
pander(cor(d[,VAR_NUMERIC]),big.mark=",") #-- matrice di correlazione
```

Table 7: Table continues below

	students	teachers	calworks	lunch	computer
students	1	0.9971	0.09016	0.1292	0.9289
teachers	0.9971	1	0.09265	0.1243	0.9372
calworks	0.09016	0.09265	1	0.7394	0.05916
lunch	0.1292	0.1243	0.7394	1	0.06139
computer	0.9289	0.9372	0.05916	0.06139	1
expenditure	-0.1123	-0.09519	0.06789	-0.06104	-0.07131
income	0.02839	0.04301	-0.5127	-0.6844	0.09434
english	0.3549	0.3514	0.3196	0.6531	0.2913
read	-0.1884	-0.1791	-0.6118	-0.8788	-0.109
math	-0.1109	-0.1023	-0.6177	-0.823	-0.03295

	expenditure	income	english	read	math
students	-0.1123	0.02839	0.3549	-0.1884	-0.1109
teachers	-0.09519	0.04301	0.3514	-0.1791	-0.1023
calworks	0.06789	-0.5127	0.3196	-0.6118	-0.6177
lunch	-0.06104	-0.6844	0.6531	-0.8788	-0.823
computer	-0.07131	0.09434	0.2913	-0.109	-0.03295
expenditure	1	0.3145	-0.0714	0.2179	0.155
income	0.3145	1	-0.3074	0.6978	0.6994
english	-0.0714	-0.3074	1	-0.6903	-0.5687
read	0.2179	0.6978	-0.6903	1	0.9229
math	0.155	0.6994	-0.5687	0.9229	1

REGRESSIONE MULTILEVEL: Empty Model

Il primo modello proposto è l'empty model.

```
#-- R CODE
mod1 <- lmer(math ~ 1 + (1 | county),d,REML=T) #-- empty model
summary(mod1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: math ~ 1 + (1 | county)
##    Data: d
##
## REML criterion at convergence: 3576.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.38073 -0.70007 -0.09327  0.63614  2.85586
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  county   (Intercept)  95.96    9.796
##  Residual             255.95   15.998
## Number of obs: 420, groups:  county, 45
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  653.696      1.759   371.6
```

```
pander(Anova(mod1, type="III"))
```

```
##
## ----------------------------------------------
##                Chisq    Df   Pr(>Chisq)
## ----------------- -------- ---- ------------
##  **(Intercept)**   138114   1         0
## ----------------------------------------------
##
## Table: Analysis of Deviance Table (Type III Wald chisquare tests)
```

```
mod1_null <- lm(math ~ 1,d) #-- modello nullo
pander(anova(mod1,mod1_null),big.mark=",") #-- test del rapporto di verosimiglianza
```

```
##
## ------------------------------------------------------------------------------------------
##              Df     AIC      BIC     logLik   deviance   Chisq   Chi Df   Pr(>Chisq)
## ---------------- ---- -------- -------- --------- ---------- ------- -------- ------------
##  **mod1_null**    2    3,657    3,665    -1,827     3,653       NA       NA          NA
##
##    **mod1**       3    3,586    3,598    -1,790     3,580      73.42     1       1.048e-17
## ------------------------------------------------------------------------------------------
##
## Table: Data: d
```

```
pander(data.frame("ICC"=icc(mod1)),big.mark=",") #-- ICC
```

```
##
## ---------------------
##            ICC
## ------------ --------
##  **county**   0.2727
## ---------------------
```

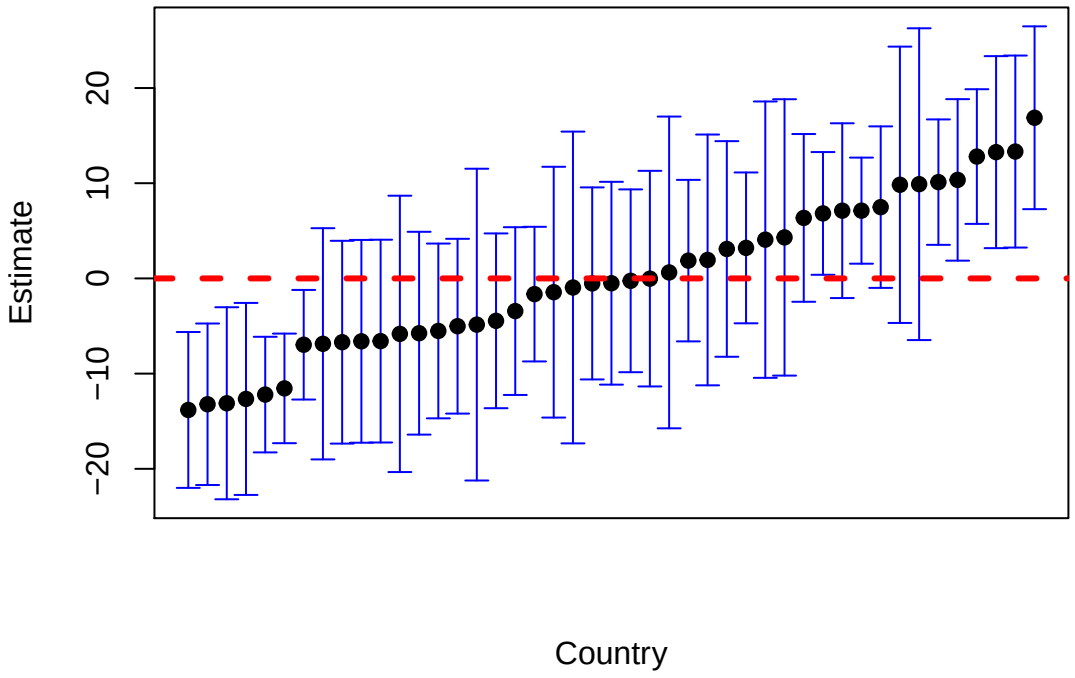
In questo caso come è noto non esistono variabili esplicative e si vede che il coefficiente interclasse è rilevante e pari a 0.273. Il test di verosimiglianza respinge l'ipotesi che il modello non interpreti la variabile dipendente. Si propongono poi gli intervalli di confidenza dei parametri casuali inerenti i distretti e quindi la relativa rappresentazione grafica.

```
#-- R CODE
res <- sjp.lmer(mod1, type = "re.qq", sort.est = "sort.all",show.values=T,title="T",prnt.plot=F)
res$data$lower <- res$data$y-res$data$ci
res$data$upper <- res$data$y+res$data$ci

pander(res$data[1:10,c("ID","y","upper","lower")],big.mark=",")
```

	ID	y	upper	lower
(Intercept)6	Fresno	-13.82	-5.629	-22
(Intercept)19	Merced	-13.22	-4.737	-21.7
(Intercept)20	Monterey	-13.11	-3.029	-23.2
(Intercept)25	Sacramento	-12.66	-2.574	-22.74
(Intercept)42	Tulare	-12.2	-6.13	-18.27
(Intercept)11	Kern	-11.55	-5.794	-17.31
(Intercept)15	Los Angeles	-6.967	-1.21	-12.72
(Intercept)24	Riverside	-6.867	5.277	-19.01
(Intercept)2	Butte	-6.701	3.95	-17.35
(Intercept)29	San Joaquin	-6.609	4.042	-17.26

```
plotCI(1:nrow(res$data),res$data$y,ui=res$data$upper, li=res$data$lower,pch=19,scol="blue",xlab="Country
abline(h=mean(res$data$y),col=2,lwd=3,lty=2)
```



Si osserva che per pochi distretti si può affermare una chiara superiorità in termini di efficacia rispetto ad altri perché l'estremo inferiore di molti si interseca con l'estremo superiore di altri.

REGRESSIONE MULTILEVEL: Random Intercept

Si propone ora un random intercept model con variabili di primo livello "calworks"" e "read" e la loro interazione.

Si è visto che il coefficiente intraclasse si dimezza perché una buona parte della varianza complessiva viene spiegata dalla variabili esplicative di primo livello.
Tutte le variabili risultano essere significative.

```
#-- R CODE
mod1 <- lmer(math ~ calworks + read + calworks*read + (1| county),d,REML=F) #-- empty model
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(mod1)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: math ~ calworks + read + calworks * read + (1 | county)
##    Data: d
##
##      AIC      BIC   logLik deviance df.resid
##   2811.3   2835.5  -1399.6   2799.3      414
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0544 -0.5807 -0.0197  0.5804  3.3553
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  county   (Intercept)  5.686   2.385
##  Residual             42.471   6.517
## Number of obs: 420, groups:  county, 45
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)   66.144119  17.178843    3.85
## calworks       3.790746   1.029885    3.68
## read           0.899224   0.025886   34.74
## calworks:read -0.006111   0.001613   -3.79
##
## Correlation of Fixed Effects:
##             (Intr) clwrks read
## calworks    -0.480
## read        -0.999  0.494
## calworks:rd  0.459 -0.999 -0.474
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```
pander(Anova(mod1, type="III"),big.mark=",")
```

```
##
## ---------------------------------------------
##                  Chisq   Df   Pr(>Chisq)
## ------------------- ------- ---- ------------
##   **(Intercept)**     14.82   1     0.000118
##
##    **calworks**       13.55   1    0.0002326
##
##      **read**         1,207   1    2.144e-264
##
##  **calworks:read**    14.35   1    0.0001514
## ---------------------------------------------
##
## Table: Analysis of Deviance Table (Type III Wald chisquare tests)
```

```
pander(data.frame("ICC"=icc(mod1)),big.mark=",") #-- ICC
```

```
##
## ----------------------
##            ICC
## ------------ --------
##  **county**   0.1181
## ----------------------
```

REGRESSIONE MULTILEVEL: Random Slope

```r
#-- R CODE
mod1 <- lmer(math ~ calworks + read + calworks*read + (calworks| county),d,REML=T) #-- empty model
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```r
summary(mod1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: math ~ calworks + read + calworks * read + (calworks | county)
##    Data: d
##
## REML criterion at convergence: 2818.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0568 -0.5719 -0.0226  0.5805  3.3751
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  county   (Intercept) 10.973478 3.31262
##           calworks     0.006747 0.08214  -0.86
##  Residual             42.130374 6.49079
## Number of obs: 420, groups:  county, 45
##
## Fixed effects:
##                Estimate Std. Error t value
## (Intercept)   59.336004  18.057454    3.29
## calworks       4.175200   1.075153    3.88
## read           0.909315   0.027257   33.36
## calworks:read -0.006679   0.001680   -3.98
##
## Correlation of Fixed Effects:
##             (Intr) clwrks read
## calworks    -0.540
## read        -0.999  0.553
## calworks:rd  0.521 -0.999 -0.535
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```r
pander(Anova(mod1, type="III"),big.mark=",")
```

```
##
## ------------------------------------------
##                  Chisq   Df   Pr(>Chisq)
## ------------------- ------- ---- ------------
##   **(Intercept)**     10.8    1     0.001016
##
##    **calworks**       15.08   1     0.000103
##
##      **read**         1,113   1    5.078e-244
##
##  **calworks:read**    15.81   1    6.987e-05
## ------------------------------------------
##
```

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

```
pander(data.frame("ICC"=icc(mod1)),big.mark=",") #-- ICC
```

```
##
## ---------------------
##            ICC
## ------------ --------
##  **county**   0.2066
## ---------------------
```

Il coefficiente di correlazione dovrebbe essere calcolato in modo diverso tenendo conto della correlazione tra i coefficienti casuali di 1 e 2 livello che risulta negativa. Il coefficiente intraclasse calcolato in modo da tenere conto della correlazione fra coefficienti casuali di 1° e 2° livello vale 0.207. Il modello rimane significativo come ogni variabile.

Si presentano ora i grafici per l'intercetta e il parametro di regressione casuale inerente "calworks".

```
#-- R CODE
res <- sjp.lmer(mod1, type = "re.qq", sort.est = "sort.all",show.values=T,title="T",prnt.plot=F)
res$data$upper <- res$data$y+res$data$ci
res$data$lower <- res$data$y-res$data$ci

res_int <- subset(res$data,ind=="(Intercept)")
res_hw <- subset(res$data,ind=="calworks")

pander(res_int[1:10,c("ID","y","upper","lower")],big.mark=",")
```

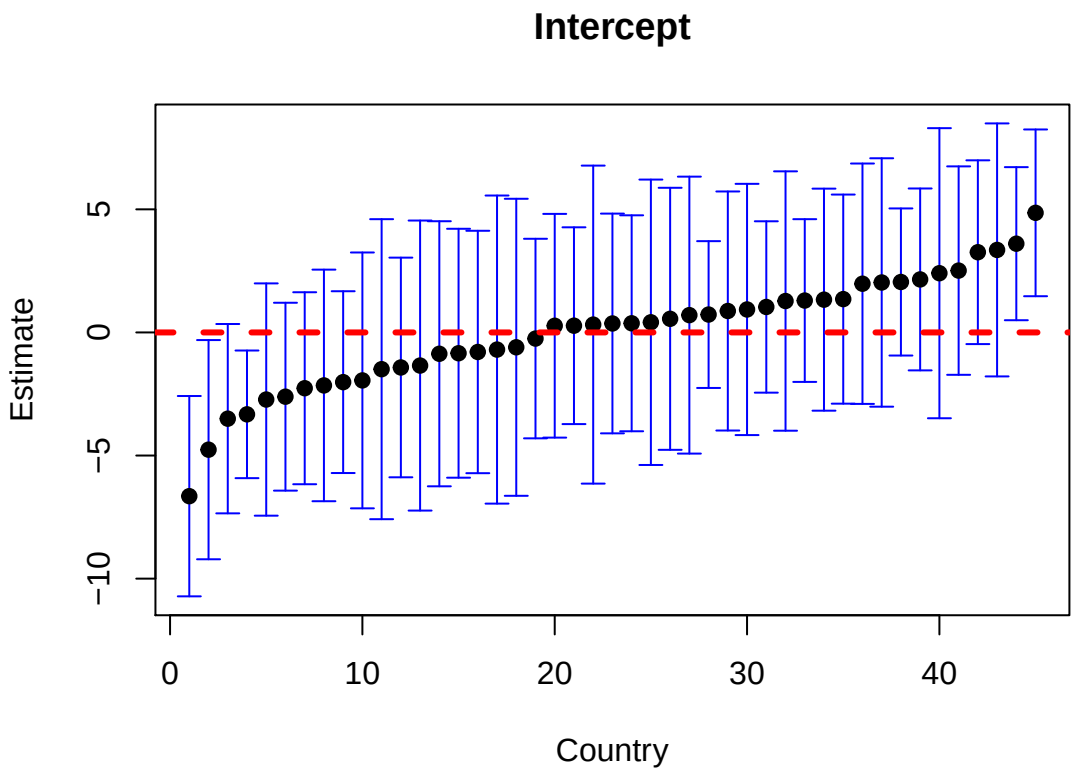
	ID	y	upper	lower
(Intercept)8	Humboldt	-6.651	-2.582	-10.72
(Intercept)39	Sutter	-4.762	-0.3106	-9.214
(Intercept)17	Marin	-3.503	0.3423	-7.348
(Intercept)37	Sonoma	-3.327	-0.735	-5.919
(Intercept)36	Siskiyou	-2.725	1.992	-7.442
(Intercept)5	El Dorado	-2.61	1.206	-6.425
(Intercept)21	Nevada	-2.267	1.632	-6.165
(Intercept)20	Monterey	-2.152	2.551	-6.856
(Intercept)23	Placer	-2.019	1.673	-5.71
(Intercept)7	Glenn	-1.949	3.247	-7.145

```
pander(res_hw[1:10,c("ID","y","upper","lower")],big.mark=",")
```

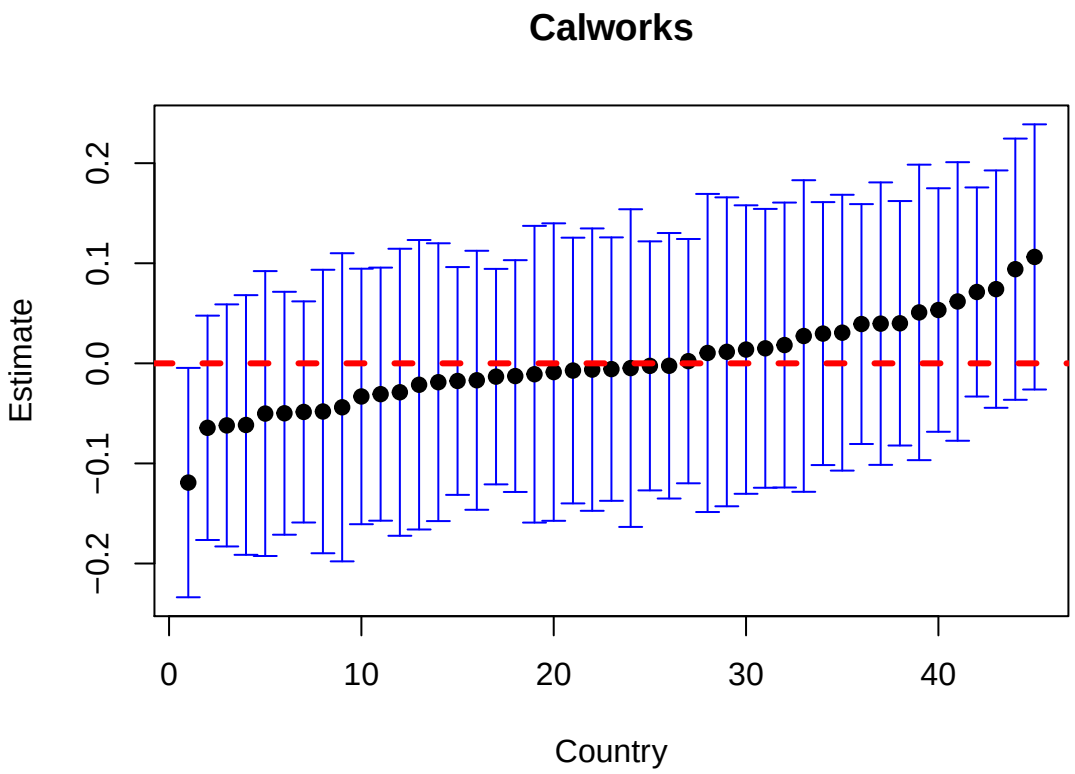
	ID	y	upper	lower
calworks11	Kern	-0.1191	-0.004548	-0.2337
calworks28	San Diego	-0.06442	0.04766	-0.1765
calworks22	Orange	-0.06201	0.05892	-0.1829
calworks27	San Bernardino	-0.06159	0.06814	-0.1913
calworks9	Imperial	-0.05016	0.09218	-0.1925
calworks6	Fresno	-0.04986	0.07151	-0.1712
calworks33	Santa Clara	-0.04855	0.06193	-0.159
calworks16	Madera	-0.04811	0.09351	-0.1897
calworks41	Trinity	-0.04392	0.11	-0.1979
calworks25	Sacramento	-0.03309	0.09456	-0.1607

	ID	y	upper	lower

```
plotCI(1:nrow(res_int),res_int$y,ui=res_int$upper, li=res_int$lower,pch=19,scol="blue",xlab="Country",y
abline(h=mean(res_int$y),col=2,lwd=3,lty=2)
```




```
plotCI(1:nrow(res_hw),res_hw$y,ui=res_hw$upper, li=res_hw$lower,pch=19,scol="blue",xlab="Country",ylab=
abline(h=mean(res_hw$y),col=2,lwd=3,lty=2)
```



Si vede come gli intervalli di confidenza si intersecano in gran parte in entrambi i casi e rendono difficile la costruzione di una graduatoria.