Una formula per la selezione di algoritmi per lo speedcubing

Tommaso Raposio

26 Luglio 2022

Glossario e Background

Speedcubing
Lo "sport" che consiste nella risoluzione del Cubo di Rubik nel minor tempo possibile.

Algoritmo
Sequenza di mosse che porta da una configurazione ad un altra.

CFOP
Il metodo più popolare. Consiste nella risoluzione del cubo procedendo per strati:

1. I primi due strati si risolvono intuitivamente.
2. Il terzo strato richiede la conoscenza di **78 algoritmi**.

Obiettivo

Ciascuna delle 78 configurazioni si può risolvere con un numero **infinito** di algoritmi.

La scelta ricade però nel dominio delle preferenze personali. Non è sufficiente per esempio basarsi esclusivamente sul numero di mosse perché un algortimo ottimale per una configurazione non è sempre "speed-optimal".

L'obiettivo di questa analisi è quindi quello di sviluppare un modello che sia in grado di valutare gli algoritmi in maniera più oggettiva, per soddisfare esigenze di:

1. **Confronto**: posto di fronte alla scelta di imparare un algoritmo nuovo, voglio poter valutare a priori se sia migliore di quello attualmente in uso.

2. **Previsione del tempo richiesto** per eseguire l'algoritmo a monte del processo di apprendimento.

Il dataset

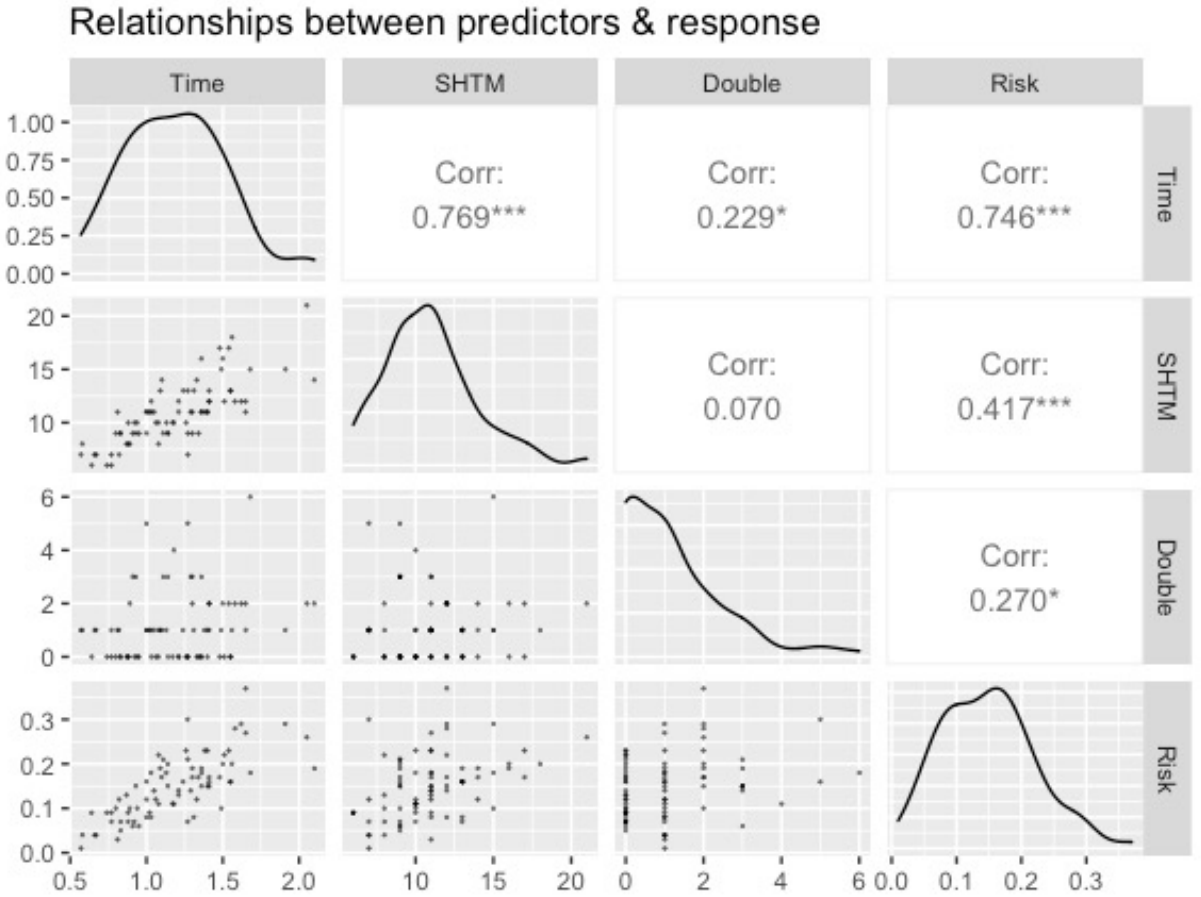
Data Collection
Per ottenere i dati richiesti, ho eseguito una media di 12 tentativi su ciascun algoritmo, eliminando il tempo migliore e il tempo peggiore. Per i predittori che non sono legati al tempo, ho eseguito un'analisi del mio *turning style*.

Covariata	Significato
Label	Uno dei 78 algoritmi
Time	Tempo richiesto
SHTM	Numero di mosse
Double	Numero di mosse doppie
Overwork	Volte in cui un dito deve eseguire due mosse consecutivamente

Covariata	Significato
Slice	Mosse interne
Risk	Deviazione standard nei tempi
SoftRegr	Volte in cui bisogna cambiare impugnatura *contemporaneamente ad una mossa*
HardRegr	Volte in cui bisogna cambiare impugnatura e fare una pausa

Analisi Esplorativa

Summary	Time	SHTM	Risk
Min	0.5700	6.0000	0.0100
Mediana	1.1750	11.0000	0.1500
Media	1.1869	10.9100	0.1485
Max	2.1000	21.0000	0.3700



Il Modello

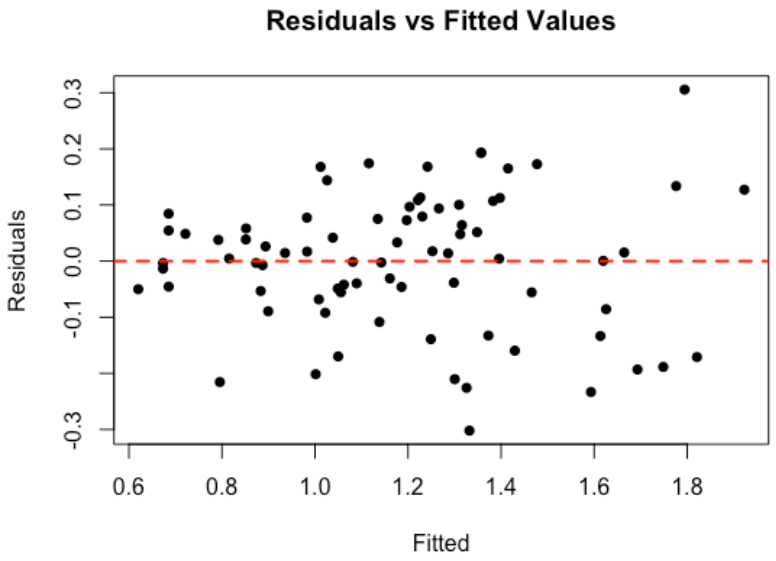
```
Call:
lm(formula = Time ~ SHTM + Double + Risk + Overwork + Slice +
    SoftRegr + HardRegr, data = df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.301958 -0.055688  0.009229  0.078879  0.305562

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.159912   0.057876   2.763  0.00731 **
SHTM         0.060956   0.005877  10.372 8.56e-16 ***
Double       0.015694   0.012654   1.240  0.21904
Risk         1.776173   0.258036   6.883 2.04e-09 ***
Overwork    -0.005178   0.016429  -0.315  0.75357
Slice        0.050391   0.025414   1.983  0.05131 .
SoftRegr     0.061165   0.028904   2.116  0.03789 *
HardRegr     0.120489   0.024639   4.890 6.19e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1255 on 70 degrees of freedom
Multiple R-squared:  0.8674,    Adjusted R-squared:  0.8541
F-statistic: 65.41 on 7 and 70 DF,  p-value: < 2.2e-16
```

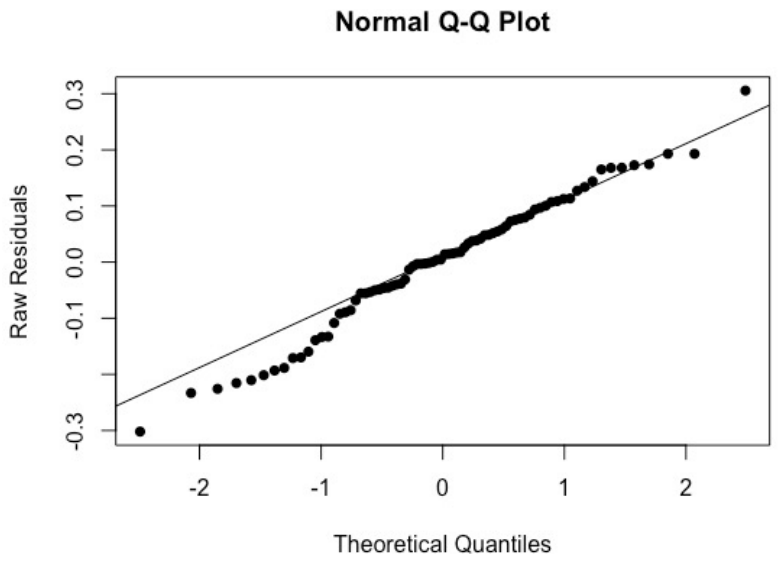
R^2	R^2_{adj}	p-value F-statistic
0.8674	0.8541	< 2.2e-16

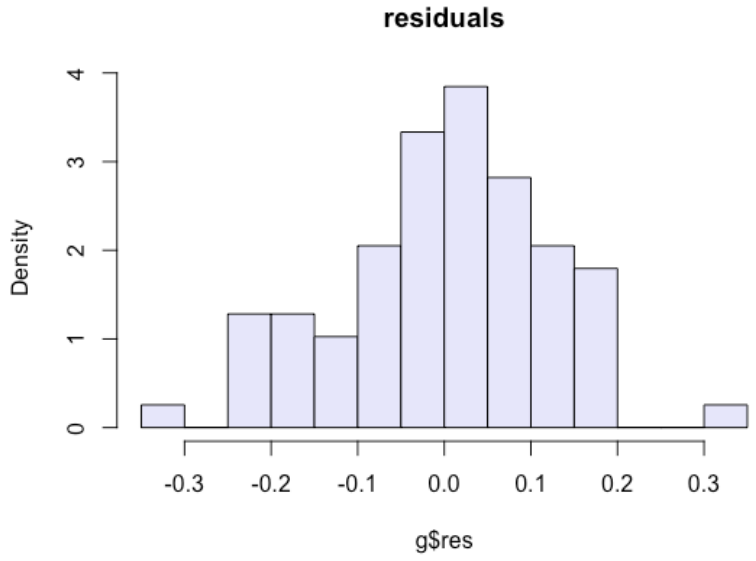
- La distribuzione dei residui è centrata in 0
- Ho evidenza per rifiutare l'ipotesi nulla per l'intercetta β_0
- Non tutte le covariate sono significative, ma ho evidenza statistica per rifiutare l'ipotesi che *tutti* i β (p-value del test F molto basso)
- I valori di R^2 e R^2_{adj} sono buoni
- Il residual standard error è 0.1255

Il Modello – Verifica delle ipotesi

Shapiro-Wilk normality test
W = 0.98428, p-value = 0.4519



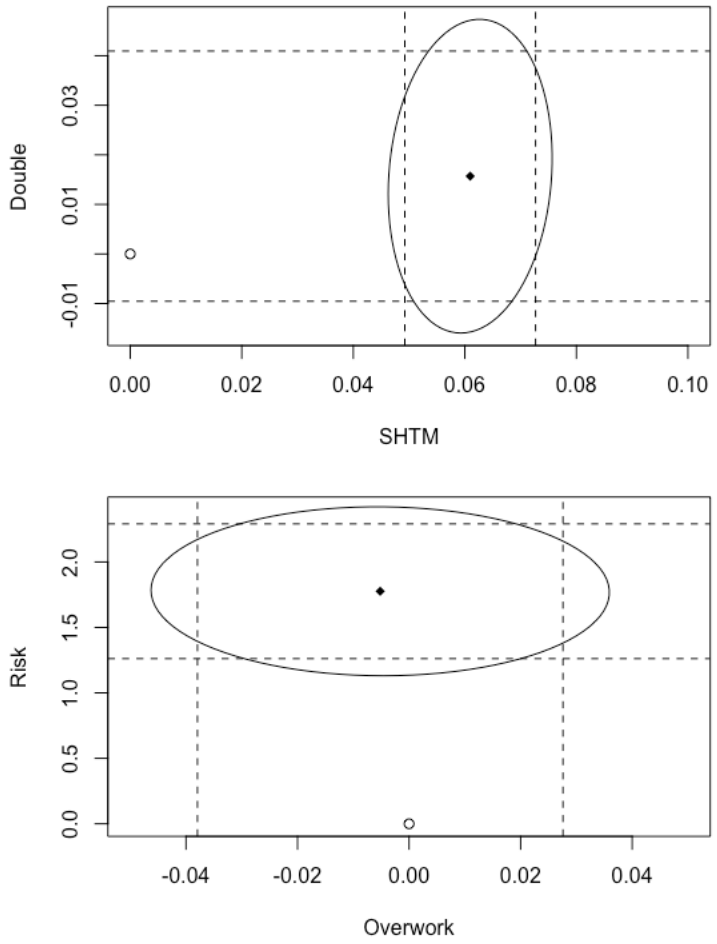


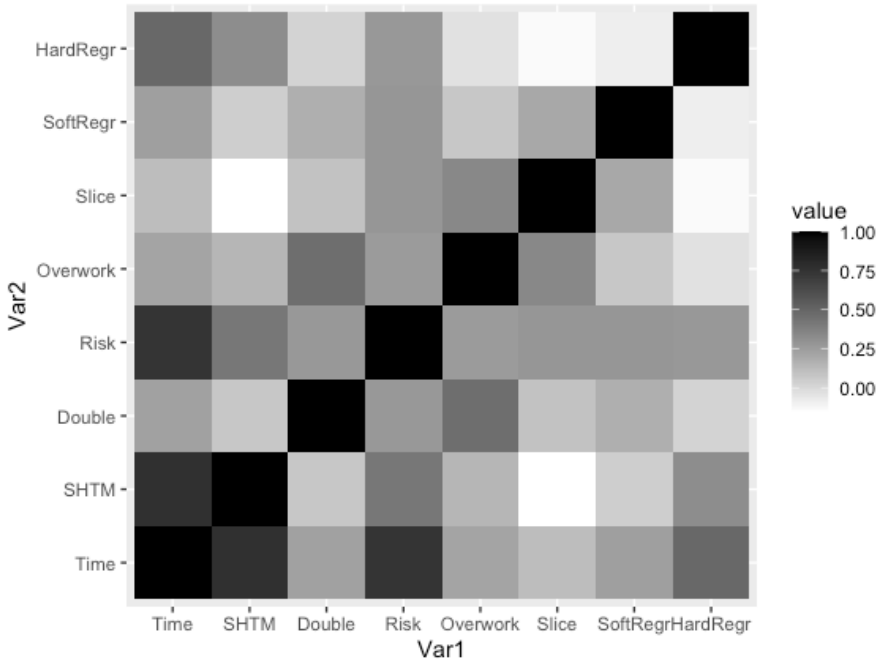


Possiamo assumere
omoschedasticità

La combinazione di QQ plot, test di Shapiro e istogramma dei residui suggerisce **distribuzione normale**

Il Modello – Analisi delle covariate





Guardando il VIF non saltano all'occhio variabili da eliminare, ma ci potrebbero essere variabili che ne mascherano altre.

Covariata	VIF
SHTM	1.4435
Double	1.3693
Risk	1.6763
Overwork	1.5097
Slice	1.4233
SoftRegr	1.1558
HardRegr	1.2119

Rimozione covariate non significative

1. Eliminando la covariata *Overwork*:
   La covariata *Slice* aumenta di significatività.
   Non variano $\boldsymbol{R^2}$ e $\boldsymbol{R^2_{adj}}$.
   Non varia il p-value del test F del modello.

2. Eliminando la covariata *Double*:
   Aumentano di significatività *SHTM* e *Risk*.
   Non variano $\boldsymbol{R^2}$ e $\boldsymbol{R^2_{adj}}$.
   Non varia il p-value del test F del modello.

3. Eliminando la covariata *Slice*:
   Variazioni nella significatività non apprezzabili

4. Eliminando le covariate *Overwork, Double* e *Slice*:
   Il modello **migliora globalmente**.

Elimino in ordine di p-value decrescente

Covariata
SHTM
Double (2)
Risk
Overwork (1)
Slice (3)
SoftRegr
HardRegr

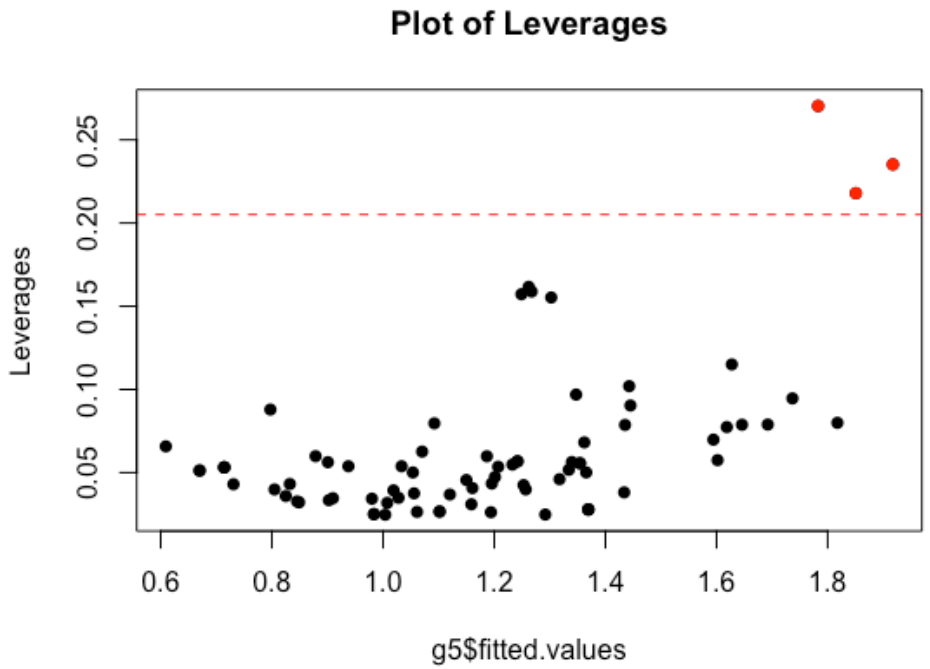
Il Modello migliorato

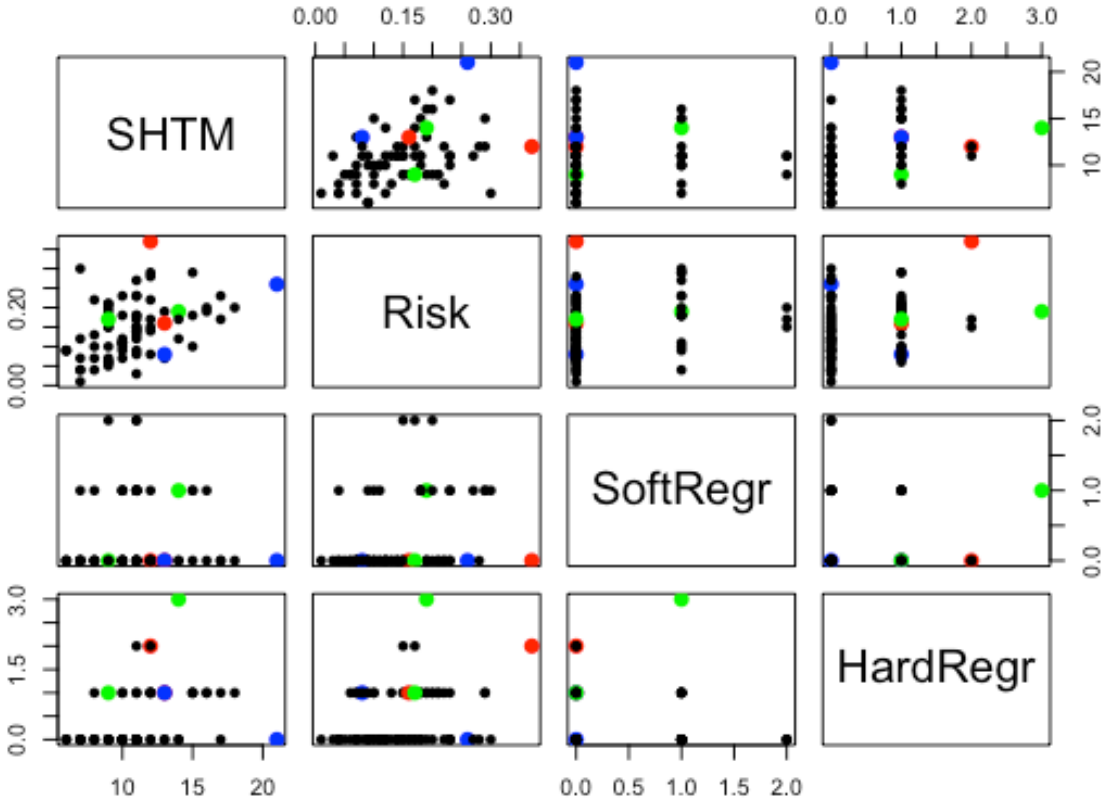
```
Call:
lm(formula = Time ~ SHTM + Risk + SoftRegr + HardRegr, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33553 -0.07279  0.01412  0.09076  0.31719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.188320   0.057422   3.280  0.00159 **
SHTM        0.057221   0.005645  10.137 1.43e-15 ***
Risk        2.028217   0.238026   8.521 1.48e-12 ***
SoftRegr    0.070030   0.029038   2.412  0.01839 *
HardRegr    0.112672   0.024689   4.564 1.99e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1277 on 73 degrees of freedom
Multiple R-squared:  0.8569,    Adjusted R-squared:  0.8491
F-statistic: 109.3 on 4 and 73 DF,  p-value: < 2.2e-16
```

R^2	R^2_{adj}	p-value statistica F
0.8569	0.8491	< 2.2e-16

- La distribuzione dei residui è sempre centrata in 0
- Ho maggiore evidenza per rifiutare l'ipotesi nulla per l'intercetta β_0
- Tutte le covariate sono significative e ho evidenza statistica per rifiutare l'ipotesi che tutti i β (p-value del test F molto basso) siano nulli.
- I valori di R^2 e R^2_{adj} sono buoni
- Il residual standard error è invariato

Il Modello – Dati Influenti





Il Modello – Dati Influenti

```
Call:
lm(formula = Time ~ SHTM + Risk + SoftRegr + HardRegr, data = df,
    subset = (lev < 0.2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.30938 -0.07192  0.01644  0.08955  0.20044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.21439    0.05861   3.658 0.000489 ***
SHTM         0.05317    0.00589   9.027 2.36e-13 ***
Risk         2.23609    0.24291   9.205 1.11e-13 ***
SoftRegr     0.04656    0.02876   1.619 0.109972
HardRegr     0.09106    0.02829   3.220 0.001948 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1197 on 70 degrees of freedom
Multiple R-squared:  0.8446,    Adjusted R-squared:  0.8357
F-statistic: 95.12 on 4 and 70 DF,  p-value: < 2.2e-16
```

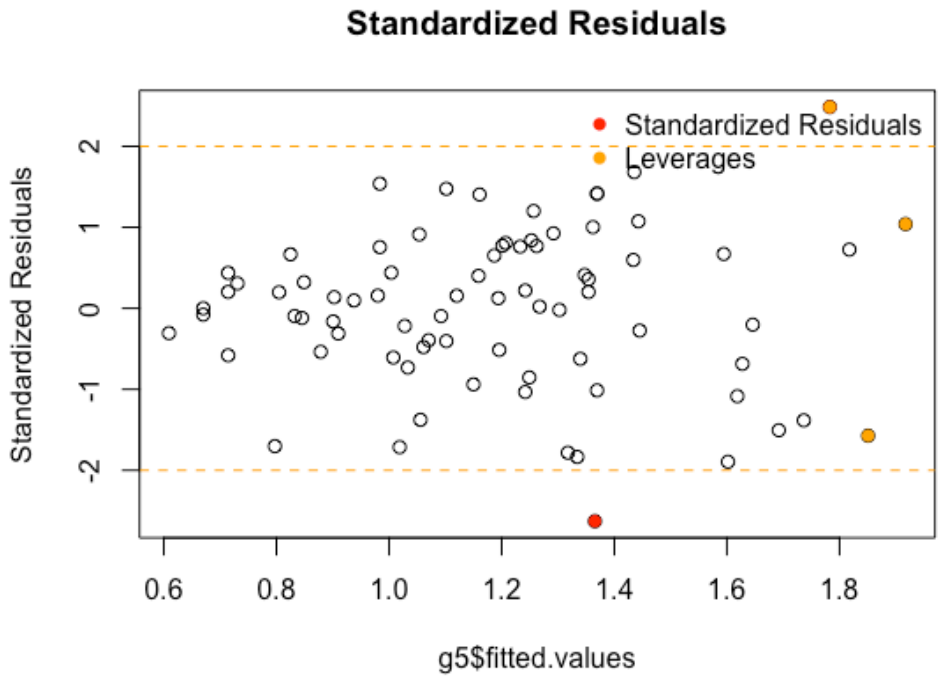
Fittiamo un modello senza leverages e valutiamo l'impatto sui coefficienti:

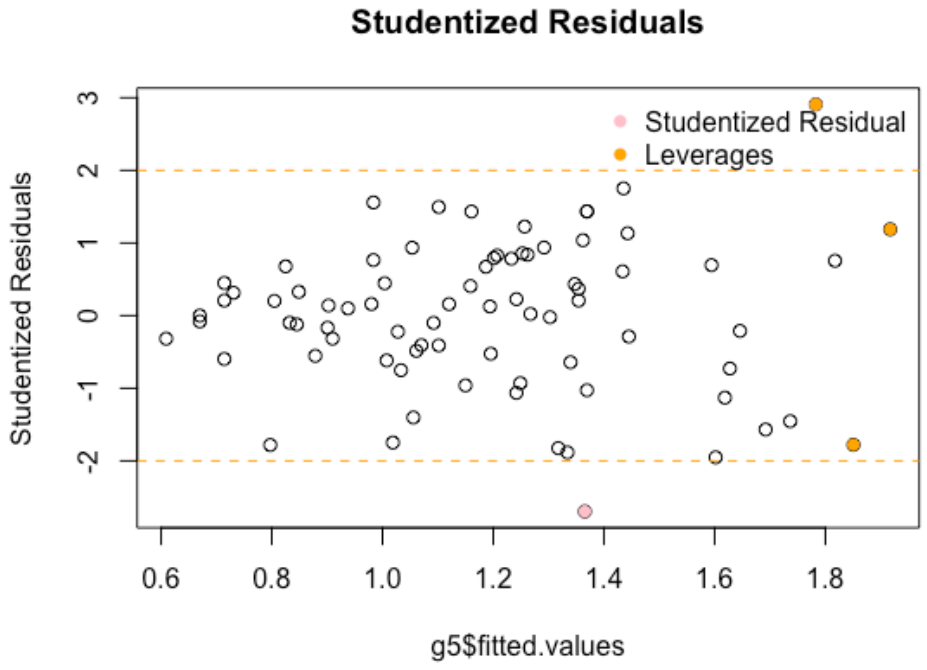
Intercept	SHTM	Risk	SoftRegr	HardRegr
0.1384	0.0708	0.1025	0.3351	0.1918

Rispetto al modello precedente, si ha un aumento netto della significatività di β_0 a fronte di una perdita di significatività della variabile *SoftRegr*

Il Modello – Dati Influenti

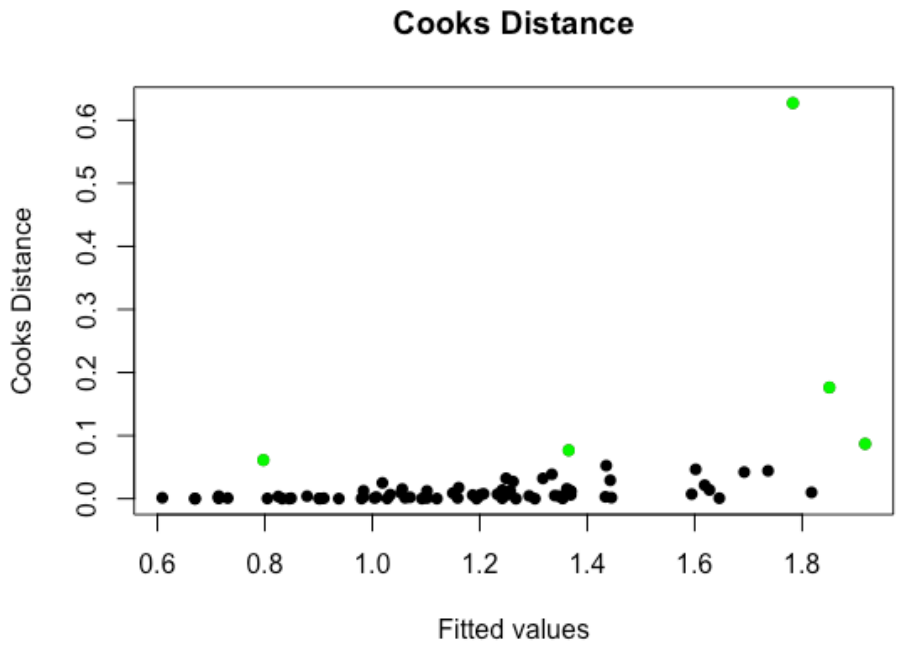
Valuto i **residui standardizzati** e i **residui studentizzati**, per cui solo pochi punti superano il treshold imposto





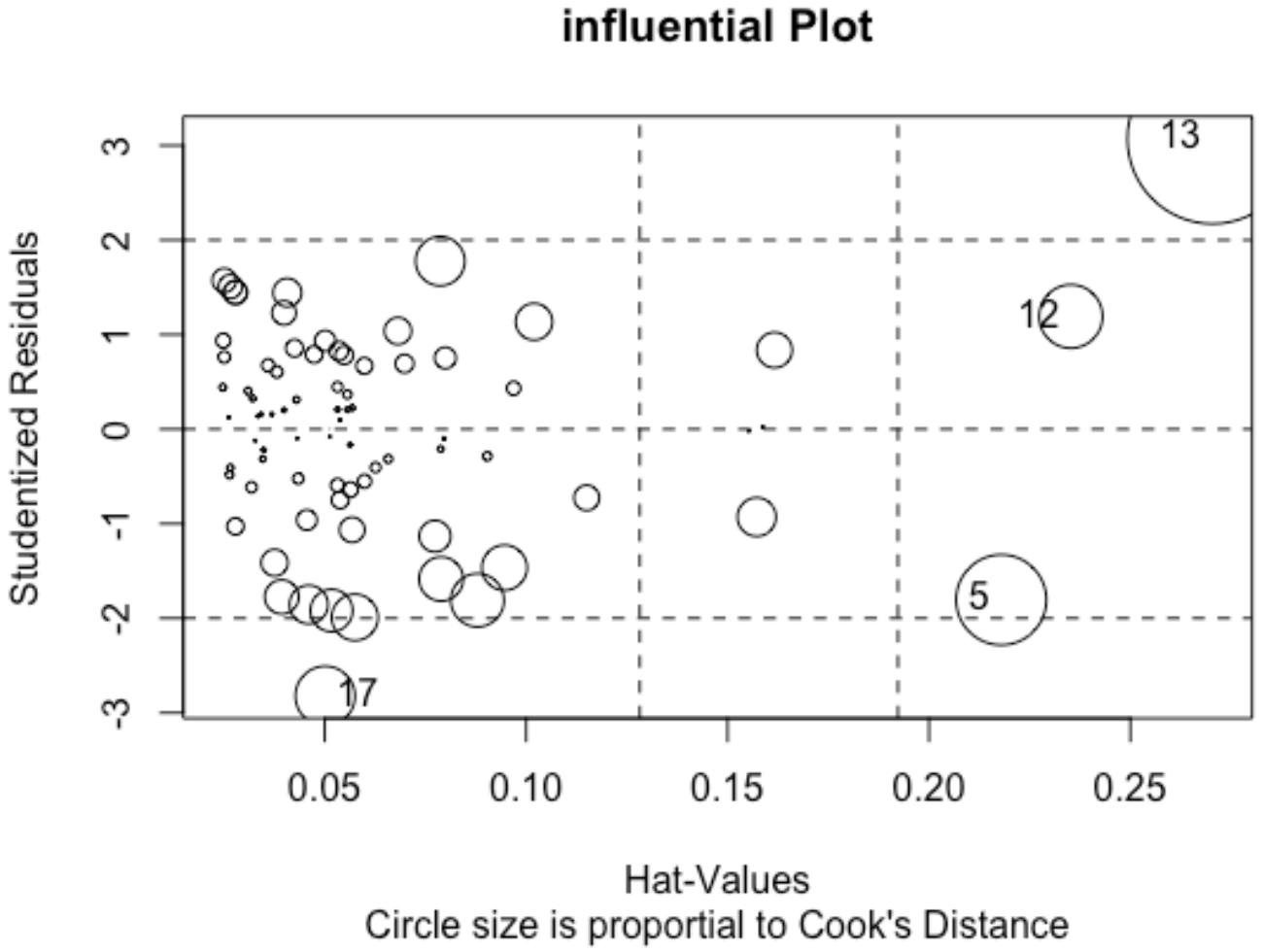
Il Modello – Dati Influenti

Valuto la **distanza di Cook** come metro per i dati influenti. Anche in questo caso solo pochi punti superano il treshold imposto. Rimuovendoli, il miglioramento del modello è netto sotto ogni aspetto.




```
Call:
lm(formula = Time ~ SHTM + Risk + SoftRegr + HardRegr, data = df[id_to_keep,
    ])

Residuals:
      Min        1Q    Median        3Q       Max
-0.226259 -0.063789  0.008691  0.086012  0.193717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.238163   0.055630   4.281 5.97e-05 ***
SHTM        0.051882   0.005521   9.397 6.53e-14 ***
Risk        2.124367   0.233061   9.115 2.09e-13 ***
SoftRegr    0.073839   0.028145   2.623 0.010735 *
HardRegr    0.105257   0.026842   3.921 0.000207 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.112 on 68 degrees of freedom
Multiple R-squared:  0.8603,    Adjusted R-squared:  0.852
F-statistic: 104.7 on 4 and 68 DF,  p-value: < 2.2e-16
```

Il Modello – Dati Influenti



Il Modello – Confronto

```
Analysis of Variance Table

Model 1: Time ~ SHTM + Risk + SoftRegr + HardRegr
Model 2: Time ~ SHTM + Double + Risk + Overwork + Slice + SoftRegr + HardRegr
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     73 1.1897
2     70 1.1027  3  0.087005 1.841 0.1477
```

Modello	R^2	R^2_{adj}	AIC	BIC
Completo	0.8674	0.8541	-92.8409	-71.6305
Predittori significativi	0.8569	0.8491	-92.9175	-78.7772
Predittori significativi e no dati influenti	0.8603	0.8520	-105.6955	-91.9527

Potenziali Problemi

1. Il numero di Soft e Hard Regrip e la varianza nei tempi di esecuzione dipendono fortemente dal "turning style" della persona. Senza l'aggiunta di nuovi dati (più tempi per i singoli algoritmi oppure tempi di diverse persone ma dello stesso livello) mi aspetto che il modello **non abbia una buona capacità predittiva generale (overfitting)**.

2. Ho raccolto i tempi in una sola sessione: verso la fine delle quasi 1000 esecuzioni ero piuttosto stanco e l'accuratezza e la velocità ne hanno risentito.

3. Avere hardware migliore potrebbe portare a dati più vicini a quello che è il mio massimo potenziale.

4. 78 dati sono pochi per avere una previsione robusta

Conclusioni

Il modello sembra in grado di catturare molto bene il comportamento della variabile di target, basandosi principalmente su tre covariate:

- **Numero di mosse** dell'algoritmo – È ragionevole pensare che al crescere della lunghezza aumenti anche il tempo necessario per eseguire le mosse

- **Rischiosità dell'algoritmo** / variabilità nel tempo di esecuzione – Un algoritmo con mosse "più comode" porta a meno errori e ad una esecuzione più fluida

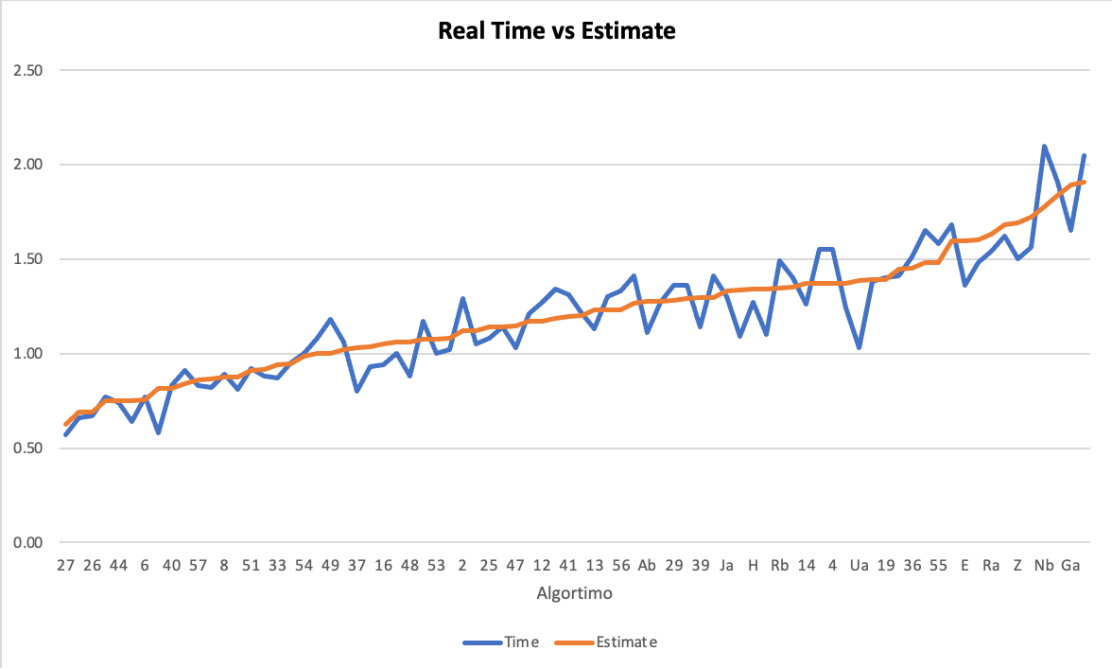
- **Numero di Hard Regrip** – Meno pause nell'esecuzione (ovviamente) corrispondono ad una esecuzione più veloce.

Variabili scartate
Gli "overworked fingers" sono un concetto di alto livello: è importante tenerli in considerazione solo se si è ai vertici dello sport, quindi è ragionevole pensare che non si applichino al mio caso. Le mosse doppie non sono un problema quando sono poco frequenti e non consecutive. Utilizzare le slice è solo marginalmente più efficiente che usare le mosse esterne.

Conclusioni

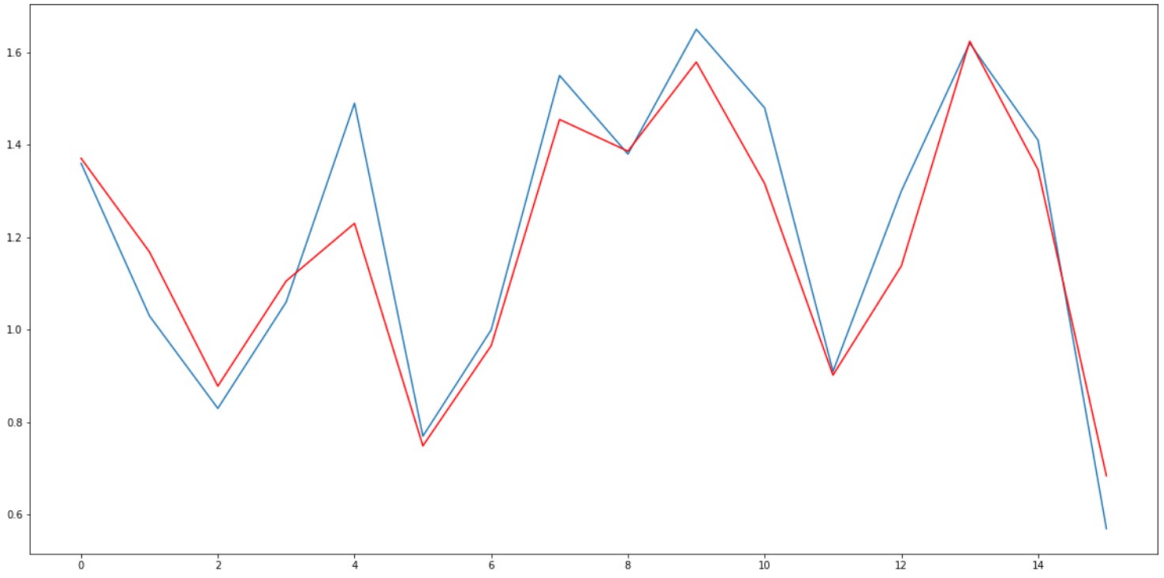
Si ottiene quindi la seguente formula per lo score di un algoritmo:

$$\widehat{Tempo} = 0.238 + 0.052 \cdot Mosse + 2.124 \cdot \text{Rischio} + 0.074 \cdot MiniPause + 0.105 \cdot Pause$$



Machine Learning

Affrontando il problema analogo con una **Foresta Casuale di Regressione**, i risultati sono leggermente migliori e il modello sembra possedere una buona capacità predittiva (linea rossa), catturando bene la variabilità del target.



$\boldsymbol{R^2}$ = 0.8935

Mean Absolute Error 0.07788
Mean Squared Error 0.01107
Root Mean Squared Error 0.10522

Feature Importances
SHTM = 0.47
Risk = 0.41