Analisi Inferenziale sulla Magnitudine Assoluta dei Clusters Globulari della Via Lattea

Chiara Peppicelli

2024-01-23

Introduzione

Globular Clusters Dataset

Il dataset oggetto del mio studio è "Globclus_prop" che fornisce le misurazioni di 20 proprietà astronomiche e astrofisiche relative a 147 ammassi globulari presenti nella Via Lattea. Queste misurazioni sono state estratte dal catalogo di Webbink (1985), il quale rappresenta un punto di riferimento fondamentale nell'analisi di tali strutture celesti.

Per via della complessità intrinseca degli studi astrofisici ho deciso di procedere inizialmente con un'approfondimento delle variabili presenti nel dataset al fine di comprenderne il significato base. Di seguito è riportata una lista esauriente delle variabili coinvolte, con le loro unità di misura.

1. Name: Nome comune
2. Gal.long: Longitudine galattica (gradi)
3. Gal.lat: Latitudine galattica (gradi)
4. R.sol: Distanza dal Sole (kiloparsecs, kpc)
5. R.GC: Distanza dal Centro Galattico (kpc)
6. Metal: Logaritmo della metallicità rispetto a quella solare
7. Mv: Magnitudine assoluta, è una misura della luminosità del cluster, osservato da una distanza standard
8. r.core: Raggio del "nucleo" (parsecs, pc), è la distanza dal centro del cluster entro la quale la densità superficiale delle stelle si dimezza rispetto al valore medio
9. r.tidal: Raggio di marea (pc), è il raggio dal centro del cluster entro il quale le forze di marea causate da un corpo esterno diventano dominanti rispetto alle forze gravitazionali interne
10. Conc: Parametro di concentrazione del nucleo
11. log.t: Logaritmo del tempo di rilassamento centrale (anni), è il tempo necessario affinchè il cluster raggiunga uno stato di equilibrio dinamico a seguito delle interazioni gravitazionali
12. log.rho: Logaritmo della densità centrale (Masse solari per pc cubo)
13. S0: Velocità di dispersione centrale (km/s), la velocità media con cui le stelle del cluster si muovono attorno al centro

14. V.esc: Velocità di fuga centrale (km/s), la velocità minima che una stella deve avere per sfuggire dalla gravità del cluster
15. VHB: Livello del ramo orizzontale (mag), rappresenta una specifica fase evolutiva delle stelle
16. E.B-V: Eccesso di colore (mag), è una misura della differenza tra il colore apparente di un oggetto celeste e il suo colore atteso sulla base di modelli teorici o stime, dovuta alla presenza di polvere interstellare
17. B-V: Indice di colore (mag)
18. Ellipt: Ellitticità
19. V.t: Magnitudine V integrata (mag), è la luminosità totale di un cluster, come appare dalla Terra, nella fascia di luce visibile
20. CSB: Luminosità superficiale centrale (mag per arcsec al quadrato), descrive la quantità di luce emmessa per unità di area, calcolata al centro del cluster

Dopo essermi informata sul significato scientifico di tali proprietà ho deciso di prendere come obiettivo di questo progetto uno studio della proprietà di magnitudine assoluta degli ammassi globulari, descritta dalla variabile "Mv".

La magnitudine assoluta, in astronomia, è la magnitudine apparente che un oggetto avrebbe se si trovasse ad una distanza dall'osservatore di 10 parsec o 1 unità astronomica a seconda del tipo di oggetto (stellare/galattico o corpo del Sistema solare), in altri termini è una misura della luminosità intrinseca di un oggetto. Più un oggetto è intrinsecamente luminoso, più la sua magnitudine assoluta è numericamente bassa, anche negativa (nel nostro dataset scorre tra -10.400 a -3.300, con media : -7.431).

Lo scopo che mi sono dunque prefissata per questo elaborato è capire quali ed in che modo le variabili nel dataset "Globclus_prop" abbiano effetti sulla magnitudine assoluta. Il tutto approcciato nell'ottica di selezionare un modello efficiente per la previsione, in relazione alla sua parsimonia e semplicità.

Operazioni preliminari iniziali

Prima di passare alla vera e propria fase di analisi del dataset, ho iniziato con alcune operazioni preliminari.

Data l'elevata presenza di osservazioni nelle quali non erano presenti le misurazioni di una o più variabili ho deciso di rimuovere tali osservazioni tramite il comando "na.omit" di R, riducendo così il dataset da 147 a 113 osservazioni, risultando in un dataset più significativo per il mio fine ultimo di fare previsione. Inoltre ho cambiato il nome del dataset in modo da poterlo utilizzare con più facilità (ho evitato invece di cambiare i nomi delle variabili in quanto risultavano già esaustivamente chiare).

```
# Caricamento del dataset, rinominazione e rimozione na
library(ggplot2)
require(graphics)
require(astrodatR)
```

```
## Caricamento del pacchetto richiesto: astrodatR
```

```
data(GlobClus_prop)
clusters <- GlobClus_prop
clusters <- na.omit(clusters)

# Sommario

summary(clusters)
```

```
##        Name        Gal.long         Gal.lat            R.sol
##  AM_1    :  1   Min.   :  0.07   Min.   :-89.3800   Min.   :  2.10
##  Eri     :  1   1st Qu.: 15.14   1st Qu.:-14.0900   1st Qu.:  6.60
##  IC_4499 :  1   Median :151.15   Median : -3.8700   Median :  9.20
##  Lil_1   :  1   Mean   :169.23   Mean   : -0.2057   Mean   : 13.85
##  NGC_104 :  1   3rd Qu.:332.97   3rd Qu.: 10.7100   3rd Qu.: 13.80
##  NGC_1261:  1   Max.   :359.59   Max.   : 79.7600   Max.   :116.40
##  (Other) :107
##       R.GC            Metal              Mv             r.core
##  Min.   :  0.90   Min.   :-2.400   Min.   :-10.400   Min.   : 0.100
##  1st Qu.:  3.10   1st Qu.:-1.800   1st Qu.: -8.300   1st Qu.: 0.500
##  Median :  6.00   Median :-1.600   Median : -7.400   Median : 0.900
##  Mean   : 11.66   Mean   :-1.418   Mean   : -7.431   Mean   : 1.795
##  3rd Qu.: 12.10   3rd Qu.:-1.000   3rd Qu.: -6.600   3rd Qu.: 1.900
##  Max.   :117.90   Max.   :-0.100   Max.   : -3.300   Max.   :12.000
##
##     r.tidal           Conc           log.t           log.rho
##  Min.   :  6.5   Min.   :0.700   Min.   : 6.200   Min.   :0.000
##  1st Qu.: 21.9   1st Qu.:1.300   1st Qu.: 7.500   1st Qu.:3.100
##  Median : 32.5   Median :1.500   Median : 8.100   Median :4.000
##  Mean   : 41.5   Mean   :1.545   Mean   : 8.093   Mean   :3.692
##  3rd Qu.: 51.7   3rd Qu.:1.800   3rd Qu.: 8.600   3rd Qu.:4.600
##  Max.   :284.8   Max.   :2.500   Max.   :10.100   Max.   :6.100
##
##        S0             V.esc            VHB            E.B.V
##  Min.   : 0.700   Min.   : 2.40   Min.   :12.90   Min.   :0.0000
##  1st Qu.: 3.900   1st Qu.:14.90   1st Qu.:15.60   1st Qu.:0.1000
##  Median : 5.600   Median :22.40   Median :16.50   Median :0.2000
##  Mean   : 6.228   Mean   :25.07   Mean   :16.64   Mean   :0.3301
##  3rd Qu.: 8.200   3rd Qu.:33.00   3rd Qu.:17.40   3rd Qu.:0.5000
##  Max.   :19.100   Max.   :78.20   Max.   :24.40   Max.   :2.9000
##
##       B.V            Ellipt            V.t             CSBt
##  Min.   :0.700   Min.   : 3.500   Min.   :0.000   Min.   : 5.20
##  1st Qu.:0.900   1st Qu.: 7.200   1st Qu.:2.000   1st Qu.: 7.70
##  Median :1.000   Median : 8.300   Median :5.000   Median : 9.00
##  Mean   :1.162   Mean   : 8.554   Mean   :4.717   Mean   : 9.15
##  3rd Qu.:1.300   3rd Qu.: 9.400   3rd Qu.:7.000   3rd Qu.:10.20
##  Max.   :4.000   Max.   :15.800   Max.   :9.000   Max.   :15.00
##
```

Questo dataset è caratterizzato da un'elevata complessità dovuta al gran numero di variabili presenti. Per semplificare l'analisi dunque, ho avviato un'esplorazione preliminare mirata alla rimozione di alcune di esse. La prima variabile eliminata è stata "Name", considerata insignificante per qualsiasi analisi inferenziale. Inoltre, dato che mi sto focalizzando sulla variabile "Mv", ho scelto di trascurare "V.t", la magnitudine integrata, poiché rappresenta semplicemente un altro modo di descrivere la magnitudine di un ammasso globulare (in particolare è la magnitudine che l'oggetto avrebbe se fosse compresso in un singolo punto luminoso) e non aggiunge rilevanza al mio studio.

Dopo un'analisi della matrice di covarianza, di cui l'output è omesso per via delle sue dimensioni considerevoli, ho deciso di eliminare una delle due variabili che rappresentavano la distanza del corpo celeste, ovvero "R.sol" o "R.GC", in quanto risultavano altamente correlate (0.97358443).

Per il mio studio, ho scelto di mantenere come misura della distanza la variabile "R.sol", distanza dal sole (in quanto approssimabile a quella dalla Terra), anziché quella dal centro galattico.

Qui di seguito è riportato il dataset dopo l'eliminazione di tali variabili.

```
clusters <- clusters[c(-1,-5,-19)]
summary(clusters)
```

```
##     Gal.long         Gal.lat             R.sol            Metal
##  Min.   :  0.07   Min.   :-89.3800   Min.   :  2.10   Min.   :-2.400
##  1st Qu.: 15.14   1st Qu.:-14.0900   1st Qu.:  6.60   1st Qu.:-1.800
##  Median :151.15   Median : -3.8700   Median :  9.20   Median :-1.600
##  Mean   :169.23   Mean   : -0.2057   Mean   : 13.85   Mean   :-1.418
##  3rd Qu.:332.97   3rd Qu.: 10.7100   3rd Qu.: 13.80   3rd Qu.:-1.000
##  Max.   :359.59   Max.   : 79.7600   Max.   :116.40   Max.   :-0.100
##        Mv              r.core          r.tidal           Conc
##  Min.   :-10.400   Min.   : 0.100   Min.   :  6.5   Min.   :0.700
##  1st Qu.: -8.300   1st Qu.: 0.500   1st Qu.: 21.9   1st Qu.:1.300
##  Median : -7.400   Median : 0.900   Median : 32.5   Median :1.500
##  Mean   : -7.431   Mean   : 1.795   Mean   : 41.5   Mean   :1.545
##  3rd Qu.: -6.600   3rd Qu.: 1.900   3rd Qu.: 51.7   3rd Qu.:1.800
##  Max.   : -3.300   Max.   :12.000   Max.   :284.8   Max.   :2.500
##      log.t           log.rho            S0             V.esc
##  Min.   : 6.200   Min.   :0.000   Min.   : 0.700   Min.   : 2.40
##  1st Qu.: 7.500   1st Qu.:3.100   1st Qu.: 3.900   1st Qu.:14.90
##  Median : 8.100   Median :4.000   Median : 5.600   Median :22.40
##  Mean   : 8.093   Mean   :3.692   Mean   : 6.228   Mean   :25.07
##  3rd Qu.: 8.600   3rd Qu.:4.600   3rd Qu.: 8.200   3rd Qu.:33.00
##  Max.   :10.100   Max.   :6.100   Max.   :19.100   Max.   :78.20
##       VHB            E.B.V             B.V            Ellipt
##  Min.   :12.90   Min.   :0.0000   Min.   :0.700   Min.   : 3.500
##  1st Qu.:15.60   1st Qu.:0.1000   1st Qu.:0.900   1st Qu.: 7.200
##  Median :16.50   Median :0.2000   Median :1.000   Median : 8.300
##  Mean   :16.64   Mean   :0.3301   Mean   :1.162   Mean   : 8.554
##  3rd Qu.:17.40   3rd Qu.:0.5000   3rd Qu.:1.300   3rd Qu.: 9.400
##  Max.   :24.40   Max.   :2.9000   Max.   :4.000   Max.   :15.800
##       CSBt
##  Min.   : 5.20
##  1st Qu.: 7.70
##  Median : 9.00
##  Mean   : 9.15
##  3rd Qu.:10.20
##  Max.   :15.00
```

La fase successiva per esplorare le relazioni tra le variabili ha coinvolto la creazione di scatterplot al fine di ottenere una comprensione visuale delle associazioni. Tuttavia, data la complessità dovuta al gran numero di variabili, ho deciso di suddividere ulteriormente il dataset in diverse categorie, quali variabili di posizione, dinamiche e fisiche. Questo approccio è stato adottato per facilitare la visualizzazione e l'analisi di ciascuna categoria, mantenendo in ognuna di esse la variabile "Mv" come riferimento.

Ho deciso di realizzare tale analisi preliminare per comprendere la natura delle relazioni esistenti e per guidare le decisioni successive riguardo all'inclusione o all'esclusione di specifiche variabili nel modello statistico.

Variabili di posizione

Dopo aver isolato il sottodataset "clusters_posizione", composto esclusivamente dalle variabili "Mv" e le coordinate di posizione, quali "Gal.long", "Gal.lat" e "R.sol", ho utilizzato il comando "pairs" di R per

creare uno scatterplot rappresentante le relazioni bivariate tra le variabili selezionate.

```
clusters_posizione <- clusters[,c(-4,-6,-7,-8,-9,-10,-11,-12,-13,-14,-15,-16,-17)]
summary(clusters_posizione)
```

```
##     Gal.long          Gal.lat              R.sol              Mv
##  Min.   :  0.07   Min.   :-89.3800   Min.   :  2.10   Min.   :-10.400
##  1st Qu.: 15.14   1st Qu.:-14.0900   1st Qu.:  6.60   1st Qu.: -8.300
##  Median :151.15   Median : -3.8700   Median :  9.20   Median : -7.400
##  Mean   :169.23   Mean   : -0.2057   Mean   : 13.85   Mean   : -7.431
##  3rd Qu.:332.97   3rd Qu.: 10.7100   3rd Qu.: 13.80   3rd Qu.: -6.600
##  Max.   :359.59   Max.   : 79.7600   Max.   :116.40   Max.   : -3.300
```

```
# pairs(clusters_posizione, panel = panel.smooth)
```

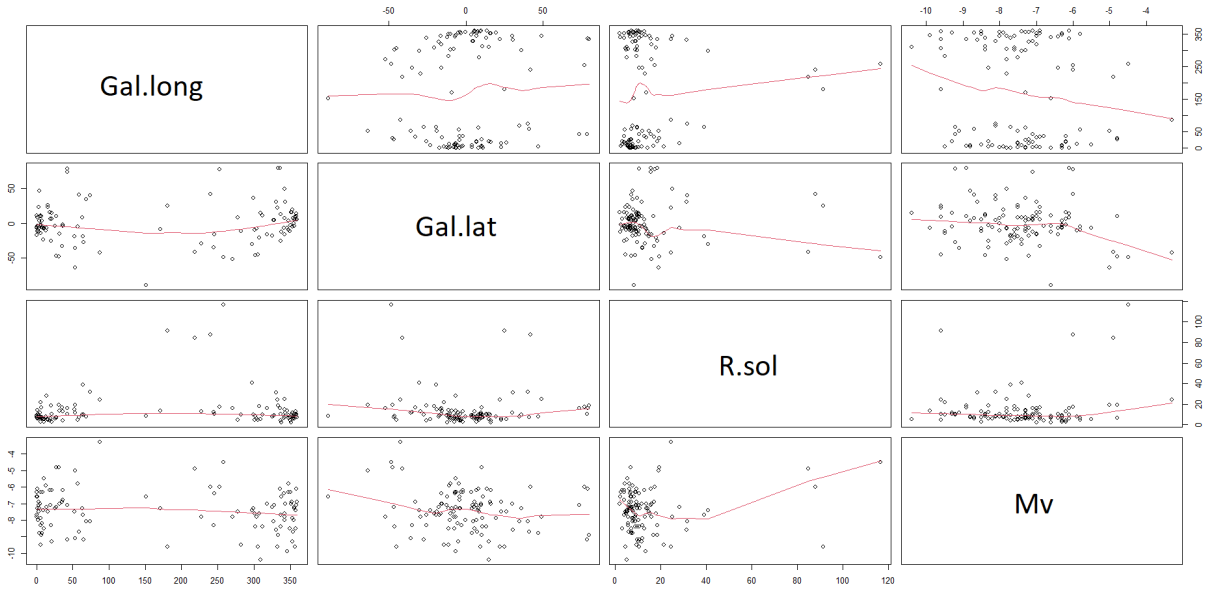


Figure 1: **Scatterplot tra "Mv" e le variabili di posizione**

Graficamente emerge l'assenza di una correlazione lineare tra le variabili di posizione e la magnitudine assoluta. Tuttavia per avere più informazioni ho tentato di costruire un modello lineare utilizzando solamente queste variabili, il quale è riportato di seguito.

```
qm <- lm(Mv ~ Gal.long + Gal.lat + R.sol, data = clusters)
summary(qm)
```

```
##
## Call:
## lm(formula = Mv ~ Gal.long + Gal.lat + R.sol, data = clusters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.8682 -0.7746  0.0178  0.9014  3.5882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.3794679  0.1907787 -38.681   <2e-16 ***
## Gal.long    -0.0012655  0.0007685  -1.647   0.1025
## Gal.lat     -0.0074362  0.0040209  -1.849   0.0671 .
## R.sol        0.0116369  0.0066802   1.742   0.0843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.217 on 109 degrees of freedom
## Multiple R-squared:  0.08101,    Adjusted R-squared:  0.05571
## F-statistic: 3.203 on 3 and 109 DF,  p-value: 0.02615
```

Le variabili di posizione non risultano statisticamente significative in un modello di regressione lineare per la mia variabile risposta "Mv". Inizialmente, avevo già considerato l'opzione di escluderle, poiché la magnitudine assoluta è una proprietà intrinseca del cluster globulare e, di conseguenza, non dipende dalla sua posizione spaziale. Il fatto che la regressione lineare non abbia significatività, in questo caso, conferma le ipotesi iniziali di escludere queste variabili.

È importante notare che queste variabili potrebbero avere un potenziale impatto su altri parametri ma per via delle dimensioni considerevoli del dataset ho scelto di ometterle.

Variabili dinamiche

Definisco il sottodataset "clusters_dinamico", composto esclusivamente dalle variabili "Mv" e le proprietà dinamiche dei clusters ovvero "r.core", "r.tidal", "Conc", "log.t", "log.rho", "V.esc", "S0", e studio lo scatterplot come fatto precedentemente nel caso delle variabili di posizione.

Di seguito il sottodataset e la matrice di correlazione tra le variabili dinamiche.

```
clusters_dinamico <- clusters[,c(-1,-2,-3,-4,-13,-14,-15,-16,-17)]
summary(clusters_dinamico)
```

```
##        Mv              r.core          r.tidal           Conc
##  Min.   :-10.400   Min.   : 0.100   Min.   :  6.5   Min.   :0.700
##  1st Qu.: -8.300   1st Qu.: 0.500   1st Qu.: 21.9   1st Qu.:1.300
##  Median : -7.400   Median : 0.900   Median : 32.5   Median :1.500
##  Mean   : -7.431   Mean   : 1.795   Mean   : 41.5   Mean   :1.545
##  3rd Qu.: -6.600   3rd Qu.: 1.900   3rd Qu.: 51.7   3rd Qu.:1.800
##  Max.   : -3.300   Max.   :12.000   Max.   :284.8   Max.   :2.500
##      log.t           log.rho            S0             V.esc
##  Min.   : 6.200   Min.   :0.000   Min.   : 0.700   Min.   : 2.40
##  1st Qu.: 7.500   1st Qu.:3.100   1st Qu.: 3.900   1st Qu.:14.90
##  Median : 8.100   Median :4.000   Median : 5.600   Median :22.40
##  Mean   : 8.093   Mean   :3.692   Mean   : 6.228   Mean   :25.07
##  3rd Qu.: 8.600   3rd Qu.:4.600   3rd Qu.: 8.200   3rd Qu.:33.00
##  Max.   :10.100   Max.   :6.100   Max.   :19.100   Max.   :78.20
```

```
cor(clusters_dinamico[-1])
```

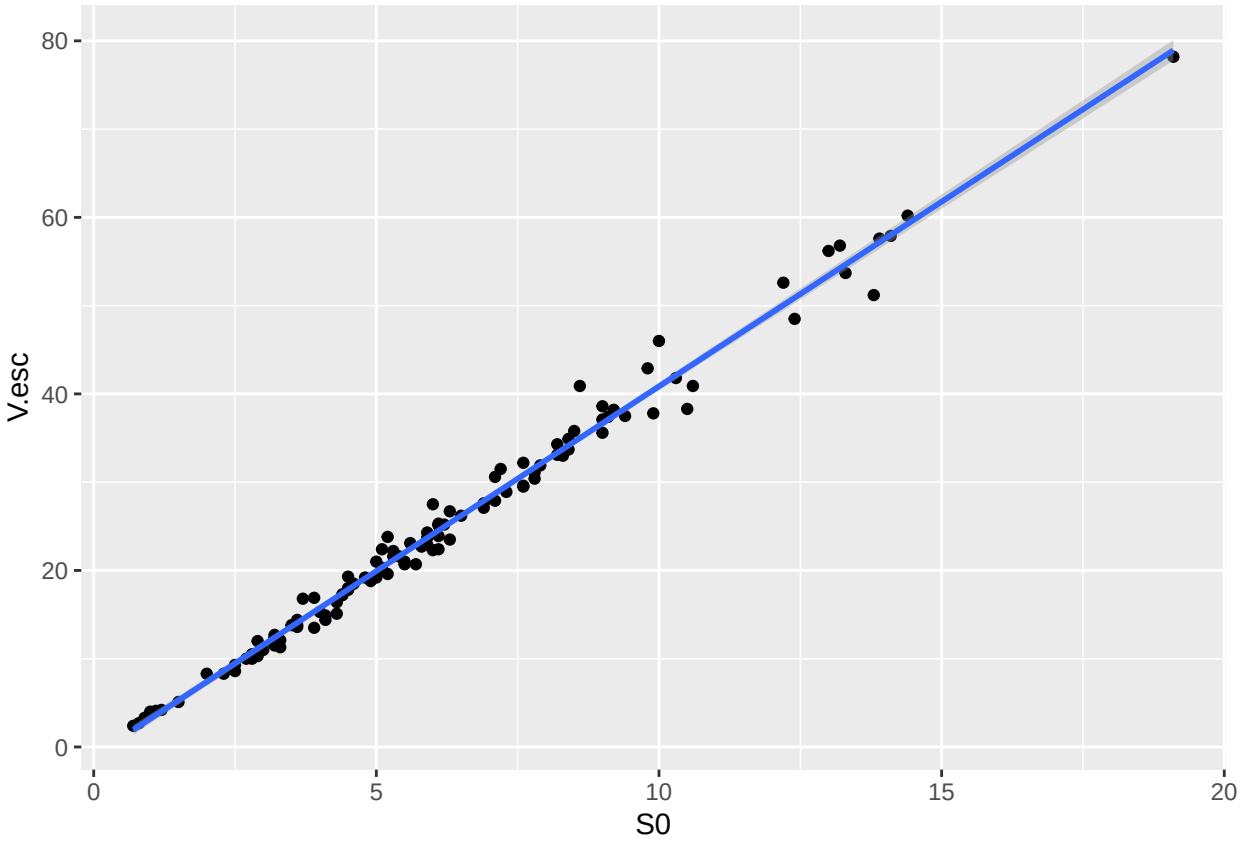
```
##             r.core     r.tidal        Conc      log.t    log.rho         S0
```

```
## r.core   1.0000000  0.57491876 -0.59924936  0.7880193 -0.8455641 -0.4455143
## r.tidal  0.5749188  1.00000000  0.02275233  0.4876358 -0.3882028 -0.0750999
## Conc    -0.5992494  0.02275233  1.00000000 -0.7632399  0.7715935  0.4555030
## log.t    0.7880193  0.48763579 -0.76323992  1.0000000 -0.8293925 -0.2480824
## log.rho -0.8455641 -0.38820283  0.77159351 -0.8293925  1.0000000  0.6959956
## S0      -0.4455143 -0.07509990  0.45550300 -0.2480824  0.6959956  1.0000000
## V.esc   -0.4688192 -0.05891512  0.53544334 -0.3096412  0.7299670  0.9939332
##                V.esc
## r.core  -0.46881916
## r.tidal -0.05891512
## Conc     0.53544334
## log.t   -0.30964120
## log.rho  0.72996697
## S0       0.99393317
## V.esc    1.00000000
```

Dall'osservazione della matrice è possibile notare un'alta correlazione tra "S0" e "V.esc" (circa 0.99), e di conseguenza, ho deciso di mantenere soltanto una delle due variabili. Di seguito riportato il plot tra le due variabili che evidenzia una forte dipendenza lineare.

```
v <- ggplot(clusters, aes(x=S0,y= V.esc))+ geom_point()+ geom_smooth(method="lm")
plot(v)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



In particolare, ho scelto di eliminare la velocità di uscita "V.esc" come variabile, poiché è una misura derivabile dalle altre (essendo dipendente dalla massa e dal raggio dell'ammasso globulare) e poichè sembrava più interessante la variabile dispersione centrale delle velocità, che fornisce informazioni sulla distribuzione delle velocità delle stelle all'interno dell'ammasso. Un'alta dispersione centrale "S0" indica una maggiore diversità nelle velocità delle stelle, che può essere correlata a una dinamica più complessa dell'ammasso globulare.

Come fatto precedentemente per avere più informazioni ho tentato di costruire un modello lineare utilizzando solamente queste variabili, il quale è riportato di seguito.

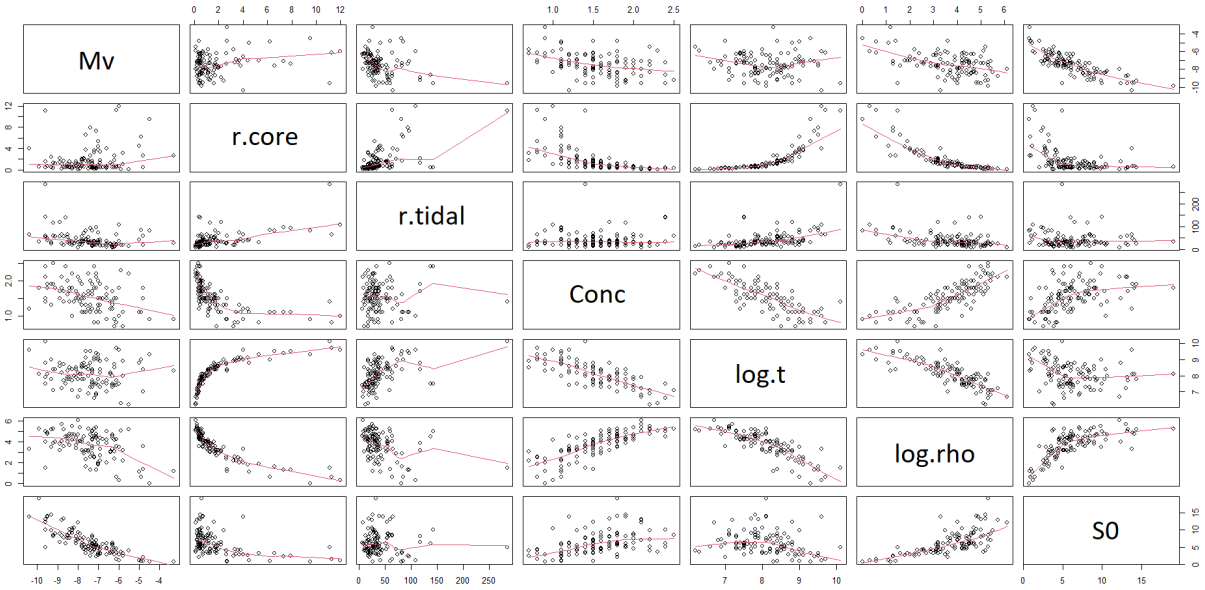


Figure 2: **Scatterplot tra "Mv" e le variabili dinamiche**

```
# pairs(clusters_dinamico[-8], panel = panel.smooth)
qm <- lm(Mv ~ r.core + r.tidal + Conc + log.t + log.rho + S0, data = clusters)
summary(qm)
```

```
##
## Call:
## lm(formula = Mv ~ r.core + r.tidal + Conc + log.t + log.rho +
##     S0, data = clusters)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56744 -0.08186  0.00293  0.08453  0.27803
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.4508312  0.6838828  32.828   <2e-16 ***
## r.core      -0.0014457  0.0124712  -0.116    0.908
## r.tidal     -0.0009552  0.0006870  -1.390    0.167
## Conc        -1.7287242  0.0818267 -21.127   <2e-16 ***
## log.t       -2.7223156  0.0658978 -41.311   <2e-16 ***
```

```
## log.rho     -1.4012985  0.0486062 -28.830   <2e-16 ***
## S0           0.0058871  0.0098114   0.600    0.550
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1353 on 106 degrees of freedom
## Multiple R-squared:  0.9889, Adjusted R-squared:  0.9883
## F-statistic:  1581 on 6 and 106 DF,  p-value: < 2.2e-16
```

Le variabili "Conc" e "log.rho" sono altamente significative in un modello di regressione lineare con la magnitudine come variabile di risposta.

Sorprendentemente, il logaritmo del tempo di rilassamento ("log.t") mostra una significatività elevata, anche se inizialmente consideravo di scartarlo. Questa variabile rappresenta una misura della scala temporale entro cui le interazioni gravitazionali tra le singole stelle all'interno dell'ammasso portano a cambiamenti nei loro parametri orbitali e nelle loro velocità, e inizialmente sembrava avere poco a che fare con la magnitudine e quindi la luminosità del cluster. Tuttavia, alla luce dei risultati, ho deciso di mantenerla come possibile variabile esplicativa nella ricerca del mio modello predittivo finale.

Al contrario, i raggi che descrivono l'ammasso globulare non risultano significativi, tuttavia ho deciso di eliminare solo uno di essi, "r.core" (la variabile con il p-value più grande nel modello). Questa decisione è coerente con ipotesi fisiche in quanto la variabile "r.core" rappresenta il raggio entro cui si dimezza la magnitudine delle stelle presenti nell'ammasso globulare e quindi non mi da informazioni sulla magnitudine assoluta del cluster. Ho deciso di tenere in considerazione per la ricerca del mio modello la variabile "r.tidal" che mi rappresenta il raggio entro il quale le stelle dell'ammasso globulare appartengono al bacino gravitazionale del cluster, perchè, pur non essendo risultata significativa in questo modello, è risultata altamente significativa in un modello lineare con solo le variabili non significative (omesso per rendere la lettura più fluida), e poichè per motivi di interpretazione, ero interessata a mantenere una misura della dimensione del cluster.

Per quanto riguarda la variabile "S0", essa, in questo modello con solo proprietà dinamiche, non è risultata molto significativa , ma comunque ho scelto di non escluderla in questa fase preliminare.

Variabili fisiche

In modo analogo a ciò che ho fatto precedentemente, definisco il sottodataset "clusters_fisico", composto esclusivamente dalle variabili "Mv" e le proprietà fisiche dei clusters ovvero "Metal", "E.B.V", "B.V", "Ellipt", "VHB" e "CSBt".

Di seguito il sottodataset e la matrice di correlazione tra le variabili fisiche.

```
clusters_fisico <- clusters[,c(-1,-2,-3,-6,-7,-8,-9,-10,-11,-12)]
summary(clusters_fisico)
```

```
##      Metal              Mv              VHB            E.B.V
##  Min.   :-2.400   Min.   :-10.400   Min.   :12.90   Min.   :0.0000
##  1st Qu.:-1.800   1st Qu.: -8.300   1st Qu.:15.60   1st Qu.:0.1000
##  Median :-1.600   Median : -7.400   Median :16.50   Median :0.2000
##  Mean   :-1.418   Mean   : -7.431   Mean   :16.64   Mean   :0.3301
##  3rd Qu.:-1.000   3rd Qu.: -6.600   3rd Qu.:17.40   3rd Qu.:0.5000
##  Max.   :-0.100   Max.   : -3.300   Max.   :24.40   Max.   :2.9000
##       B.V            Ellipt            CSBt
##  Min.   :0.700   Min.   : 3.500   Min.   : 5.20
##  1st Qu.:0.900   1st Qu.: 7.200   1st Qu.: 7.70
##  Median :1.000   Median : 8.300   Median : 9.00
```

```
##  Mean   :1.162   Mean   : 8.554   Mean   : 9.15
##  3rd Qu.:1.300   3rd Qu.: 9.400   3rd Qu.:10.20
##  Max.   :4.000   Max.   :15.800   Max.   :15.00
```

```
cor(clusters_fisico[-2])
```

```
##              Metal       VHB     E.B.V       B.V    Ellipt       CSBt
## Metal   1.00000000 0.1205876 0.2903893 0.5431760 0.1516224 0.09212871
## VHB     0.12058757 1.0000000 0.4555454 0.4279573 0.8139601 0.44822417
## E.B.V   0.29038929 0.4555454 1.0000000 0.9565375 0.3833592 0.20830574
## B.V     0.54317601 0.4279573 0.9565375 1.0000000 0.3752533 0.20884296
## Ellipt  0.15162244 0.8139601 0.3833592 0.3752533 1.0000000 0.72792978
## CSBt    0.09212871 0.4482242 0.2083057 0.2088430 0.7279298 1.00000000
```

Per prima cosa ho deciso di escludere la variabile "E.B.V", l'eccesso di colore, che si riferisce alla differenza tra l'indice di colore osservato dell'ammasso e il suo indice di colore intrinseco o atteso. La motivazione alla base di questa esclusione è che "E.B.V" è risultata altamente correlata (0.95) con la variabile "B.V", l'indice di colore, e tra le due "B.V" è risultata più significativa, più intuitiva e facilmente comprensibile dal punto di vista interpretativo.

Di seguito un modello di regressione lineare semplice tra le variabili fisiche.

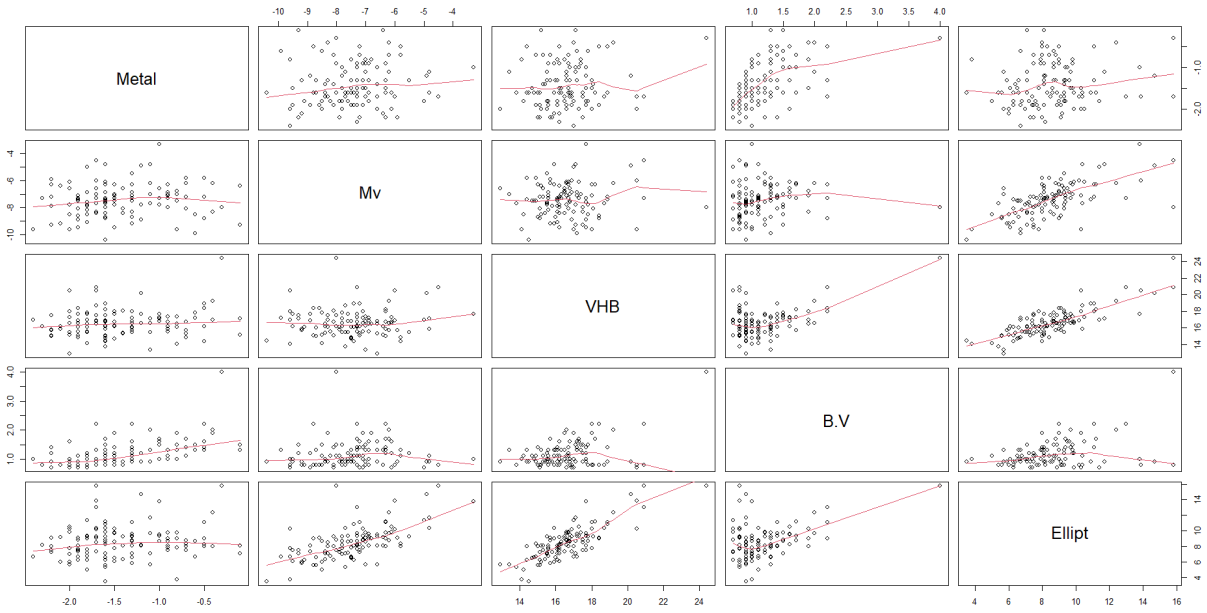


Figure 3: **Scatterplot tra "Mv" e le variabili fisiche**

```
# pairs(clusters_fisico[c(-3,-7)], panel = panel.smooth)
qm <- lm(Mv ~ Metal + VHB + B.V + Ellipt+ CSBt, data = clusters)
summary(qm)
```

```
##
## Call:
## lm(formula = Mv ~ Metal + VHB + B.V + Ellipt + CSBt, data = clusters)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08986 -0.05028  0.02030  0.04314  0.06387
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.765100   0.066500   11.505   <2e-16 ***
## Metal       -0.011711   0.010495   -1.116    0.267
## VHB         -1.006510   0.005263 -191.226   <2e-16 ***
## B.V          0.017874   0.013297    1.344    0.182
## Ellipt       0.997020   0.005087  196.000   <2e-16 ***
## CSBt        -0.001702   0.003401   -0.500    0.618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04808 on 107 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9985
## F-statistic: 1.518e+04 on 5 and 107 DF,  p-value: < 2.2e-16
```

Nonostante alcune variabili non abbiano mostrato significatività ho scelto di conservarle tutte per costruire un modello predittivo, dato che il mio parametro d'interesse riguarda una proprietà fisica, ad esclusione della variabile "CSBt" (la luminosità superficiale centrale); questa è stata esclusa poiché è risultata la meno significativa del modello. Inoltre, rappresentando una misura della luminosità concentrata nella regione centrale dell'ammasso, ed essendo espressa in magnitudini per arcsec al quadrato, inserirla nello studio avrebbe portato a una duplicazione dei dati riferiti alla luminosità, risultando meno rilevante ai fini dell'inferenza sulla stessa.

Modello di regressione lineare

Al termine dell'analisi preliminare, ho selezionato un insieme specifico di variabili rilevanti. Per evitare confusione e semplificare l'analisi successiva, procedo con la creazione di un nuovo dataset rinominato "gb", contenente solo le variabili di interesse che sono risultate significative nel contesto del mio studio.

```
gb <- clusters[c(-1,-2,-3,-6,-12,-14,-17)]
summary(gb)
```

```
##      Metal              Mv             r.tidal           Conc
##  Min.   :-2.400   Min.   :-10.400   Min.   :  6.5   Min.   :0.700
##  1st Qu.:-1.800   1st Qu.: -8.300   1st Qu.: 21.9   1st Qu.:1.300
##  Median :-1.600   Median : -7.400   Median : 32.5   Median :1.500
##  Mean   :-1.418   Mean   : -7.431   Mean   : 41.5   Mean   :1.545
##  3rd Qu.:-1.000   3rd Qu.: -6.600   3rd Qu.: 51.7   3rd Qu.:1.800
##  Max.   :-0.100   Max.   : -3.300   Max.   :284.8   Max.   :2.500
##      log.t           log.rho            S0              VHB
##  Min.   : 6.200   Min.   :0.000   Min.   : 0.700   Min.   :12.90
##  1st Qu.: 7.500   1st Qu.:3.100   1st Qu.: 3.900   1st Qu.:15.60
##  Median : 8.100   Median :4.000   Median : 5.600   Median :16.50
##  Mean   : 8.093   Mean   :3.692   Mean   : 6.228   Mean   :16.64
##  3rd Qu.: 8.600   3rd Qu.:4.600   3rd Qu.: 8.200   3rd Qu.:17.40
##  Max.   :10.100   Max.   :6.100   Max.   :19.100   Max.   :24.40
##       B.V            Ellipt
```

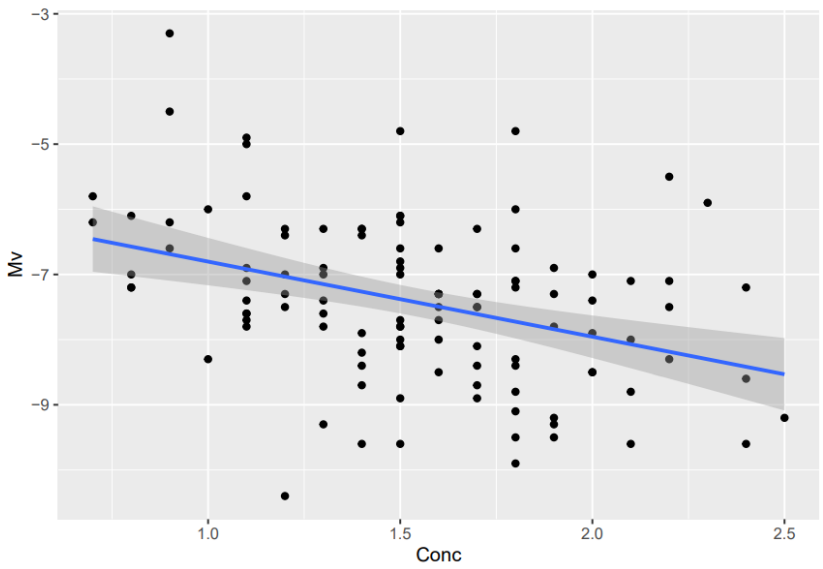
```
##  Min.   :0.700   Min.   : 3.500
##  1st Qu.:0.900   1st Qu.: 7.200
##  Median :1.000   Median : 8.300
##  Mean   :1.162   Mean   : 8.554
##  3rd Qu.:1.300   3rd Qu.: 9.400
##  Max.   :4.000   Max.   :15.800
```

Regressione lineare semplice

Inizialmente, al fine di identificare possibili connessioni e interazioni all'interno del mio dataset ridotto, ho optato per l'implementazione di un modello di regressione lineare semplice. Nello specifico, osservando i grafici precedenti, avevo individuato potenziali effetti e relazioni lineari tra la mia variabile obiettivo "Mv" e le altre variabili.

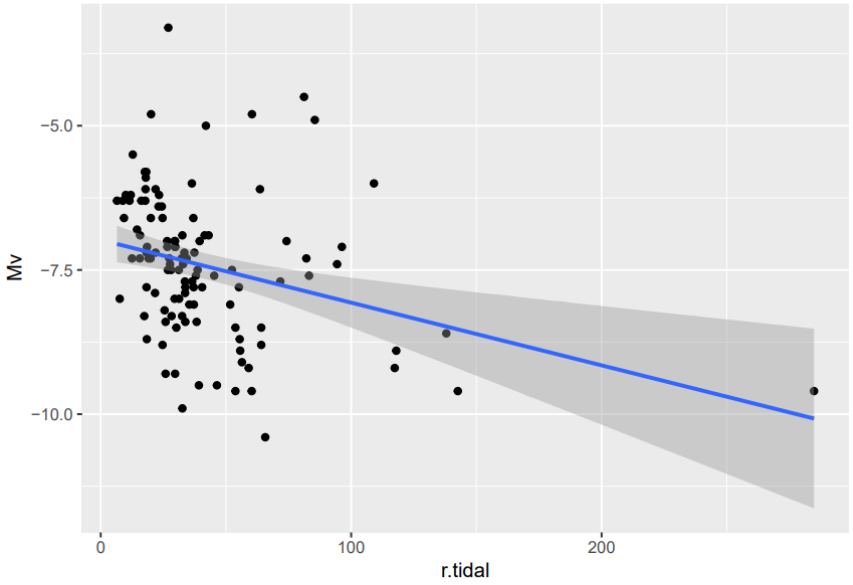
Di seguito, sono presentati i grafici di confronto tra "Mv" e le singole variabili, oltre a una tabella che rappresenta il riepilogo ottenuto da una regressione lineare semplice con ciascuna variabile. I valori del coefficiente di determinazione e altri dettagli sono inclusi nel resoconto, mentre il "summary" completo è stato omesso per facilitare la lettura.

```
library(ggplot2)

Qplot1 <- ggplot(gb, aes(x=Conc,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm1 <- lm( Mv ~ Conc, data=gb)
plot(Qplot1)
```



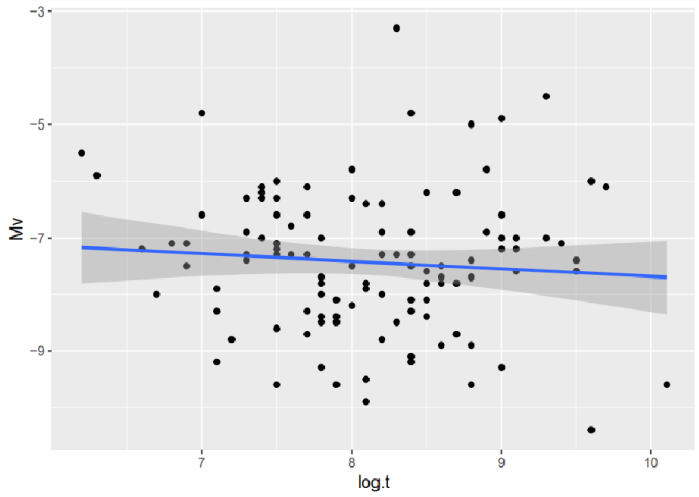
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-5.6497	0.4312	-13.104	< 2e-16 ***
Conc	-1.1529	0.2699	-4.272	4.11e-05 ***

```
Qplot2 <-  ggplot(clusters, aes(x=r.tidal,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm2 <- lm(Mv ~r.tidal, data=gb)
plot(Qplot2)
```



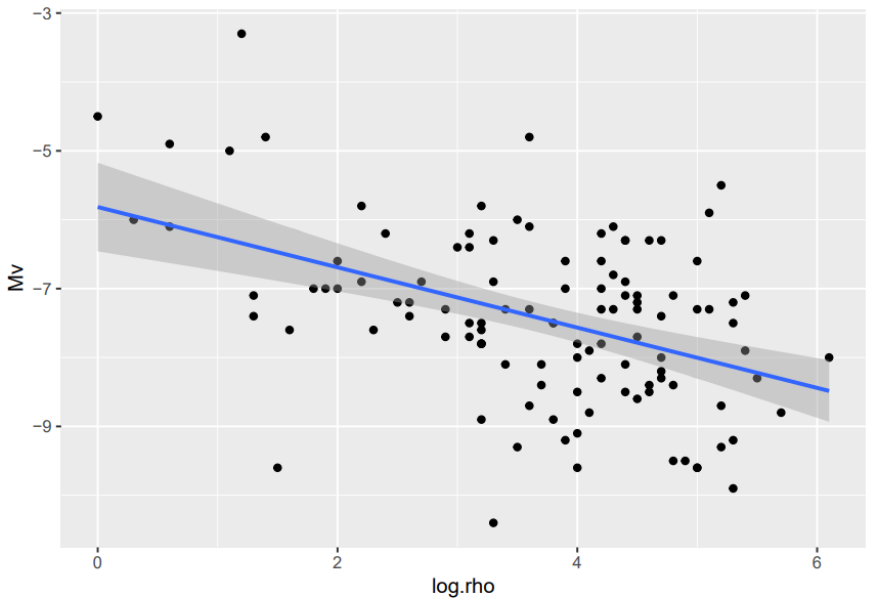
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-6.979754	0.173999		-40.114
r.tidal	-0.010872	0.003196	-3.402	0.000932 ***

```
Qplot3 <- ggplot(clusters, aes(x=log.t,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm3 <- lm(Mv ~log.t, data=gb)

plot(Qplot3)
```



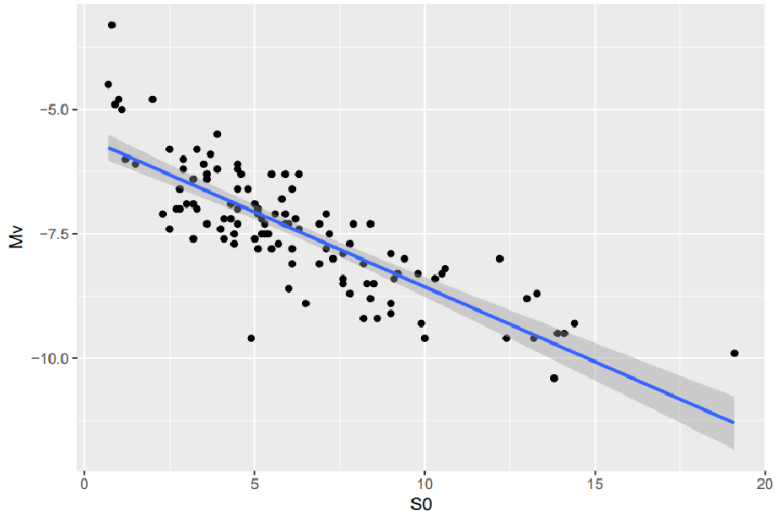
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-6.3274	1.2436	-5.088	1.49e-06 ***
log.t	-0.1364	0.1530	-0.891	0.375

```
Qplot4 <- ggplot(clusters, aes(x=log.rho,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm4 <- lm(Mv ~log.rho , data=gb)
plot(Qplot4)
```



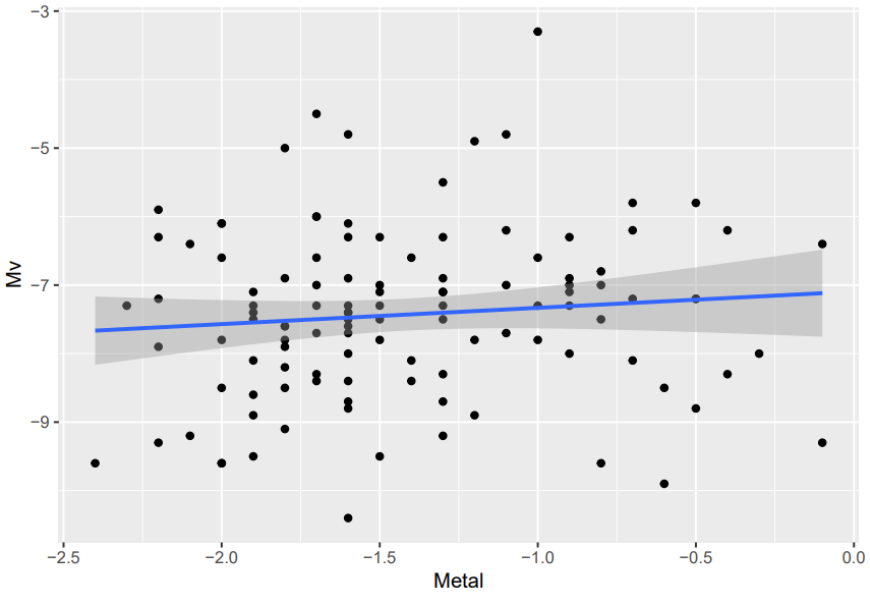
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-5.81540	0.32439	-17.927	< 2e-16 ***
log.rho	-0.43758	0.08305	-5.269	6.82e-07 ***

```
Qplot5 <- ggplot(clusters, aes(x=S0,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm5 <- lm(Mv ~S0, data=gb)
plot(Qplot5)
```



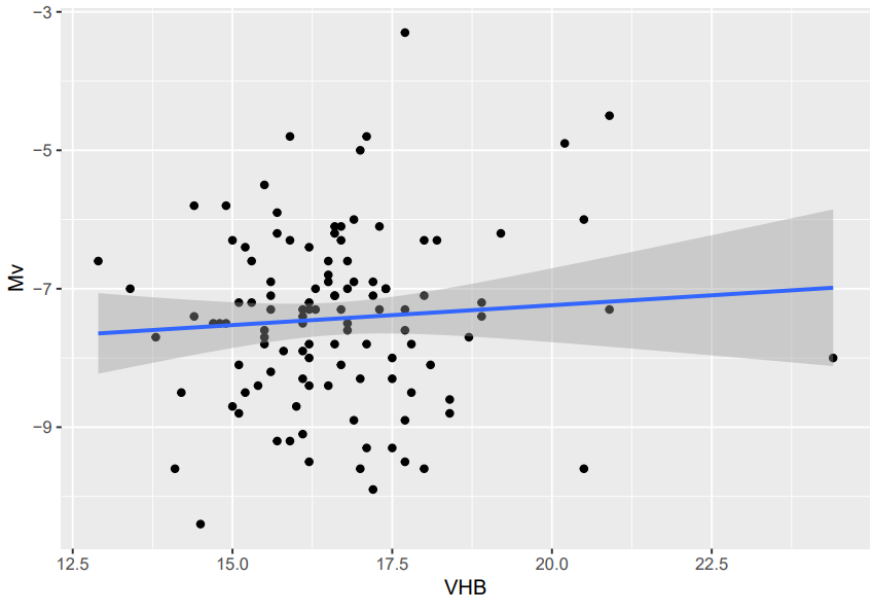
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-5.55847	0.14291	-38.90	<2e-16 ***
S0	-0.30064	0.02016	-14.91	<2e-16 ***

```
Qplot6 <- ggplot(clusters, aes(x=Metal,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm6 <- lm(Mv ~Metal, data=gb)

plot(Qplot6)
```



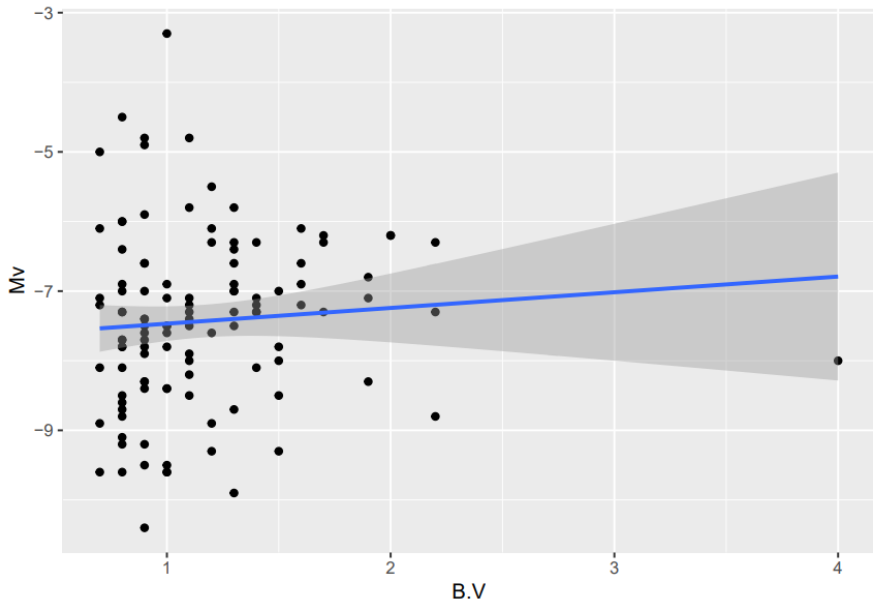
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-7.0941	0.3416	-20.767	<2e-16 ***
Metal	0.2377	0.2262	1.051	0.296

```
Qplot7 <- ggplot(clusters, aes(x=VHB,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm7 <- lm(Mv ~VHB, data=gb)

plot(Qplot7)
```



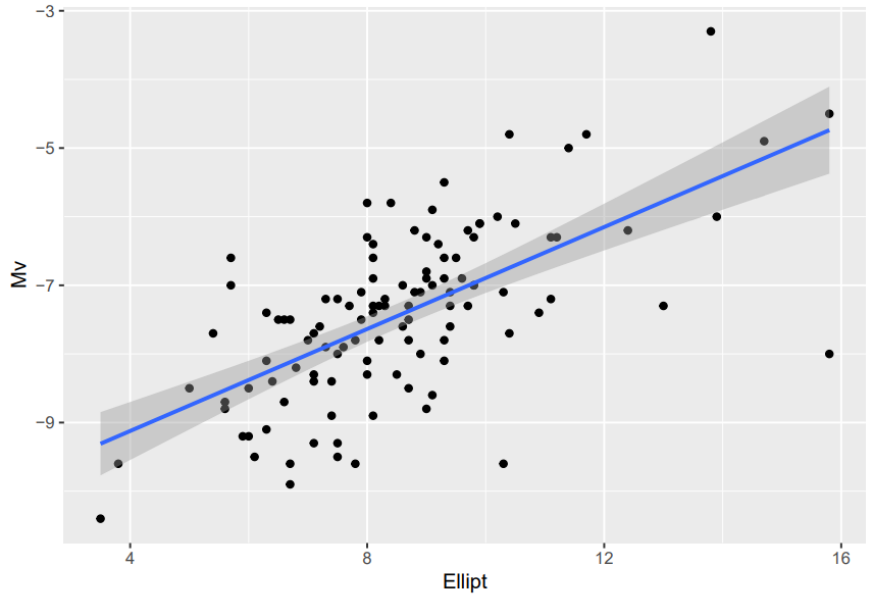
Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-8.38550	1.20266	-6.972	2.36e-10 ***
VHB	0.05737	0.07194	0.798	0.427

```
Qplot8 <- ggplot(clusters, aes(x=B.V,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm8 <- lm(Mv ~B.V, data=gb)
plot(Qplot8)
```



Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-7.6930	0.3267	-23.55	<2e-16 ***
B.V	0.2255	0.2622	0.86	0.392

```
Qplot9 <- ggplot(clusters, aes(x=Ellipt,y= Mv))+ geom_point()+ geom_smooth(method="lm")
lm9 <- lm(Mv ~Ellipt, data=gb)
plot(Qplot9)
```



Coefficients:	Estimate	Std. Error	t value	Pr(>|z|)
(Intercept)	-10.6084	0.3731	-28.437	< 2e-16 ***
Ellipt	0.3715	0.0423	8.782	2.24e-14 ***

Confronto tra i modelli lineari precedenti :

1. lm1: Si nota un effetto lineare negativo di "Conc" su "Mv" (coefficiente di -1.1529) con una evidente significatività (p-value 4.11e-05), Adjusted R-squared: 0.1335
2. lm2: Si osserva un effetto lineare negativo piccolo di "r.tidal" su "Mv" (coefficiente di -0.010872), la variabile risulta significativa (p-value 0.000932) e Adjusted R-squared in peggioramento rispetto a lm1 (0.08625)
3. lm3: "log.t" presenta un effetto lineare negativo più marcato rispetto a "r.tidal" (coefficiente di -0.1364). Tuttavia, il p-value non è significativo (p-value 0.375) e l'Adjusted-R-squared è negativo (-0.001837).
4. lm4: "log.rho" mostra un effetto lineare negativo più accentuato di "log.t" (coefficiente di -0.43758) ed è altamente significativo (p-value 6.82e-07) con un Adjusted-R-squared di 0.1928.
5. lm5: "S0" ha un effetto lineare negativo minore rispetto al precedente (coefficiente di -0.30064), con p-value molto significativo (p-value <2e-16 *) e Adjusted-R-squared molto buono (0.6641) .
6. lm6: "Metal" è la prima variabile ad avere un effetto lineare positivo (in linea con le conoscenze teoriche per cui più è alta la metallicita di un cluster meno è luminoso) (coefficiente 0.2377) però ha p-value non significativo (p-value 0.296) e Adjusted-R-squared molto piccolo (0.0009274).
7. lm7: "VHB" ha un un effetto lineare positivo più basso di "Metal" (coefficiente 0.05737), e ha un p-value meno significativo (p-value 0.427) e Adjusted-R-squared negativo (-0.00326) .
8. lm8: "Bv" ha un effetto lineare positivo (coefficiente di 0.2255), p-value non significativo (0.392) e Adjusted-R-squared negativo (-0.002328).
9. lm9: "Ellipt" mostra un effetto lineare positivo (coefficiente 0.3715) con p value molto significativo (p-value 2.24e-14) e Adjusted-R-squared 0.4046.

Nei modelli di regressione lineare diretta, le variabili "VHB", "BV", "Metal" e "log.t" hanno mostrato una bassa significatività.

Tuttavia, prima di procedere con l'eliminazione di queste variabili, che potrebbero rivelarsi altamente significative in un modello più completo, ho scelto di esaminare il comportamento in un contesto più ampio.

Per fare ciò, ho iniziato con un modello completo utilizzando tutte le variabili nel dataset "gb" e successivamente ho rimosso gradualmente le variabili che mostravano una minore significatività. Ho interrotto questo processo quando tutte le variabili rimanenti risultavano significative, seguendo il metodo "backwards" di eliminazione delle variabili basato sul p-value. Questo approccio mira a garantire che le variabili mantenute siano statisticamente rilevanti nel contesto di un modello più completo.

```
# Parto un modello lineare con tutte le variabili del dataset, includendo anche quelle che erano
# risultate non significative nei modelli di regressione lineare semplici

qm_dataset <-lm(Mv ~ Metal+  log.t + VHB + B.V + r.tidal + Conc + log.rho +  S0 +Ellipt, data=clusters)
summary(qm_dataset)
```

```
##
## Call:
## lm(formula = Mv ~ Metal + log.t + VHB + B.V + r.tidal + Conc +
##     log.rho + S0 + Ellipt, data = clusters)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08471 -0.04416  0.01830  0.04170  0.07068
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1363483  0.7700067   2.774  0.00657 **
## Metal       -0.0076684  0.0106288  -0.721  0.47225
## log.t       -0.1877464  0.0933549  -2.011  0.04693 *
## VHB         -0.9215240  0.0330587 -27.875  < 2e-16 ***
## B.V          0.0066127  0.0192174   0.344  0.73147
## r.tidal     -0.0004902  0.0002313  -2.120  0.03644 *
## Conc        -0.0963970  0.0660610  -1.459  0.14755
## log.rho     -0.1009284  0.0487674  -2.070  0.04099 *
## S0          -0.0036363  0.0034772  -1.046  0.29812
## Ellipt       0.9154185  0.0325485  28.125  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04653 on 103 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic:  9001 on 9 and 103 DF,  p-value: < 2.2e-16
```

```
# La variabile risultata meno significativa è "B.V" (p-value 0.73147)

qm_dataset1 <-lm(Mv ~ Metal+  log.t + VHB + r.tidal + Conc + log.rho +  S0 +Ellipt, data=clusters)

# Ometto i prossimi summary per facilitare la lettura

# La variabile meno significativa è "Metal" (p-value 0.52553)

qm_dataset2 <-lm(Mv ~  log.t + VHB  + r.tidal + Conc + log.rho +  S0 +Ellipt, data=clusters)

# La variabile meno significativa è "S0" (p-value 0.23990)

qm_dataset3 <-lm(Mv ~  log.t + VHB +  r.tidal + Conc + log.rho  +Ellipt, data=clusters)

# La variabile meno significativa è "Conc" (p-value 0.064067)

qm_dataset4 <-lm(Mv ~  log.t + VHB +  r.tidal  + log.rho  +Ellipt, data=clusters)
summary(qm_dataset4)
```

```
##
## Call:
## lm(formula = Mv ~ log.t + VHB + r.tidal + log.rho + Ellipt, data = clusters)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08032 -0.04814  0.02164  0.03983  0.06770
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2457409  0.2393519   5.205 9.45e-07 ***
## log.t       -0.0700707  0.0312904  -2.239   0.0272 *
## VHB         -0.9703622  0.0147869 -65.623  < 2e-16 ***
## r.tidal     -0.0005228  0.0002173  -2.405   0.0179 *
## log.rho     -0.0506694  0.0234522  -2.161   0.0330 *
```

```
## Ellipt       0.9637697  0.0140809  68.445  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04683 on 107 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.6e+04 on 5 and 107 DF,  p-value: < 2.2e-16
```

```
# Ho lasciato il summary di questo modello in quanto le variabili risultano tutte significative
# ma non altamente significative (ordine 0.01) quindi ho deciso di continuare eliminando
# la variabile "log.rho" (p-value 0.0330)

qm_dataset5 <-lm(Mv ~  log.t + VHB +  r.tidal   +Ellipt, data=clusters)

# La variabile meno significativa è "log.t" (p-value 0.553)

qm_dataset6 <-lm(Mv ~   VHB +  r.tidal   +Ellipt, data=clusters)

# La variabile meno significativa è "r.tidal" (p-value 0.0844)

qm_dataset7 <-lm(Mv ~  VHB + Ellipt, data=clusters)
summary(qm_dataset7)
```

```
##
## Call:
## lm(formula = Mv ~ VHB + Ellipt, data = clusters)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08596 -0.05064  0.02072  0.04438  0.06578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75787    0.05667   13.38   <2e-16 ***
## VHB         -1.00378    0.00473 -212.21   <2e-16 ***
## Ellipt       0.99509    0.00361  275.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04791 on 110 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9985
## F-statistic: 3.82e+04 on 2 and 110 DF,  p-value: < 2.2e-16
```

```
# Rimangono due variabili altamente significative
```

Grazie a questo approccio di analisi "backwards", a partire da tutte le variabili, arrivo alla definizione del modello :

Mv ~ VHB + Ellipt

Entrambe le variabili sono risultate altamente significative (p-value: $< 2.2e\text{-}16$), e il valore dell'Adjusted R-squared è notevolmente elevato (0.9985). Nonostante nella regressione lineare semplice la variabile "VHB" sembrasse poco significativa, qua è apparsa come l'unica insieme ad "Ellipt" ad avere un'alta significatività e quindi ho scelto di mantenerla nella ricerca del modello finale.

D'altra parte, ho deciso di rimuovere le altre variabili che non sono risultate significative nel confronto diretto. La variabile "BV", che rappresenta l'indice di colore, è stata esclusa in quanto è risultata la meno significativa tra tutte. La variabile "log.t", che rappresenta il tempo di rilassamento, era stata considerata a priori per l'eliminazione, poiché, intuitivamente, non sembrava influenzare direttamente la magnitudine assoluta.

Le variabili rimanenti nel mio studio, che sono risultate significative individualmente, compongono un elenco significativo di sei parametri, "r.tidal", "Conc", "log.rho", "S0", "Ellipt" e "VHB" i quali contribuiscono collettivamente alla spiegazione della variabilità della magnitudine assoluta.

Scelta del modello lineare

Selezione modello con metodo p-value direzione "backward"

Partendo dal modello completo andiamo ogni volta ad eliminare la variabile a cui è associato il p-value più alto.

Modello completo:

```
# Modello completo
mq <-lm(Mv ~  r.tidal + Conc + log.rho + S0 + Ellipt + VHB, data=gb)
summary(mq)
```

```
##
## Call:
## lm(formula = Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + VHB,
##     data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08054 -0.05050  0.02552  0.04021  0.06041
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.5979267  0.0815458   7.332 4.72e-11 ***
## r.tidal     -0.0005181  0.0002285  -2.267   0.0254 *
## Conc         0.0236887  0.0227006   1.044   0.2991
## log.rho     -0.0031768  0.0101401  -0.313   0.7547
## S0          -0.0061681  0.0032938  -1.873   0.0639 .
## Ellipt       0.9781925  0.0083809 116.717  < 2e-16 ***
## VHB         -0.9833659  0.0098384 -99.952  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04692 on 106 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.328e+04 on 6 and 106 DF,  p-value: < 2.2e-16
```

```
#  La variabile meno significativa è log.rho (p-value  0.7547)
# Ometto i promissi summary per facilitare la lettura

mq1 <-lm(Mv ~  r.tidal + Conc + S0 + Ellipt + VHB, data=gb)
```

```r
# La variabile meno significativa è Conc (0.15016)
mq2 <-lm(Mv ~  r.tidal + S0 + Ellipt + VHB, data=gb)

# La variabile meno significativa è S0 (0.0667)

mq3 <- lm(Mv ~  r.tidal + Ellipt + VHB, data=gb)
summary(mq3)
```

```
##
## Call:
## lm(formula = Mv ~ r.tidal + Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08333 -0.05045  0.02305  0.04002  0.07432
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.7252633  0.0591881   12.254   <2e-16 ***
## r.tidal     -0.0002424  0.0001392   -1.742   0.0844 .
## Ellipt       0.9928901  0.0037949  261.639   <2e-16 ***
## VHB         -1.0000773  0.0051455 -194.360   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04748 on 109 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 2.594e+04 on 3 and 109 DF,  p-value: < 2.2e-16
```

```r
# La variabile meno significativa è r.tidal (0.0844)

mq4 <- lm(Mv ~  Ellipt + VHB, data=gb)
summary(mq4)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08596 -0.05064  0.02072  0.04438  0.06578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75787    0.05667   13.38   <2e-16 ***
## Ellipt       0.99509    0.00361  275.62   <2e-16 ***
## VHB         -1.00378    0.00473 -212.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04791 on 110 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9985
## F-statistic: 3.82e+04 on 2 and 110 DF,  p-value: < 2.2e-16
```

Terminiamo la procedura quando tutte le variabili risultano significative ovvero quando otteniamo il modello:

$Mv \sim Ellipt + VHB$

Selezione modello con metodo p-value, direzione "forward"

Iniziamo la procedura considerando il modello nullo con sola intercetta.

```
qm0 <-lm(Mv ~  1, data=gb)
summary(qm0)
```

```
##
## Call:
## lm(formula = Mv ~ 1, data = gb)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.969 -0.869  0.031  0.831  4.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.4310     0.1178  -63.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.252 on 112 degrees of freedom
```

Già precedentemente abbiamo stimato la significatività dei modelli semplici e avevamo osservato che quello con la variabile più significativa è quello dove è stata aggiunto il parametro "S0".

```
qm1 <-lm(Mv ~  S0 , data=gb)
summary(qm1)
```

```
##
## Call:
## lm(formula = Mv ~ S0, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56838 -0.46709 -0.04876  0.50220  2.49899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.55847    0.14291  -38.90   <2e-16 ***
## S0          -0.30064    0.02016  -14.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7258 on 111 degrees of freedom
## Multiple R-squared:  0.6671, Adjusted R-squared:  0.6641
## F-statistic: 222.5 on 1 and 111 DF,  p-value: < 2.2e-16
```

Iteriamo aggiungendo le variabili in ordine, dal p-value più basso, fintanto che risultano tutte significative.

```
qm2 <-lm(Mv ~  S0 + Ellipt, data=gb)

qm3 <-lm(Mv ~  S0 + Ellipt + log.rho, data=gb)

qm4 <-lm(Mv ~  S0 + Ellipt + log.rho + Conc, data=gb)
summary(qm4)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + Ellipt + log.rho + Conc, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96486 -0.28944 -0.04868  0.22667  2.07170
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.49997    0.37270 -20.123  < 2e-16 ***
## S0          -0.32392    0.02367 -13.687  < 2e-16 ***
## Ellipt       0.18629    0.02860   6.513 2.40e-09 ***
## log.rho      0.49500    0.08265   5.989 2.82e-08 ***
## Conc        -0.86375    0.20889  -4.135 7.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5648 on 108 degrees of freedom
## Multiple R-squared:  0.8039, Adjusted R-squared:  0.7966
## F-statistic: 110.7 on 4 and 108 DF,  p-value: < 2.2e-16
```

```
qm5 <-lm(Mv ~ Conc + log.rho + S0 + Ellipt + r.tidal, data=gb)
summary(qm5)
```

```
##
## Call:
## lm(formula = Mv ~ Conc + log.rho + S0 + Ellipt + r.tidal, data = gb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9249 -0.3392 -0.0111  0.2114  1.6352
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.899332   0.310762 -22.201  < 2e-16 ***
## Conc         0.217456   0.219705   0.990    0.325
## log.rho     -0.059555   0.098347  -0.606    0.546
## S0          -0.252177   0.021263 -11.860  < 2e-16 ***
## Ellipt       0.175015   0.023128   7.567 1.41e-11 ***
## r.tidal     -0.013835   0.001803  -7.672 8.30e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4557 on 107 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8676
## F-statistic: 147.7 on 5 and 107 DF,  p-value: < 2.2e-16
```

Se aggiungiamo la variabile "r.tidal", ossia la successiva con p-value più basso, si ottiene la perdità di significatività di alcune variabili ("Conc" e "log.rho"), quindi il modello dove sono tutte significative rimane quello precedente alla sua aggiunta, ovvero :

Mv ~ S0 + Ellipt + log.rho + Conc

Selezione del modello con metodo p-value misto

Il metodo misto p-value è una procedura che combina sia l'approccio forward che backward nella costruzione del modello. Il processo inizia come il metodo p-value forward con l'aggiunta sequenziale delle variabili risultate più significative; poi continua fino a quando l'aggiunta di ulteriori variabili non comporta la perdita di significatività di quelle già presenti. In questo caso, esse, vengono rimosse. La procedura di aggiunta e rimozione si ripete fino a quando non ci saranno ulteriori variabili da aggiungere.

```
# Dato che l'aggiunta iterativa delle variabili è stata già vista per il metodo forward
# ho deciso di omettere i summary

qm1 <-lm(Mv ~  S0 , data=gb)

qm2 <-lm(Mv ~  S0 + Ellipt, data=gb)


qm3 <-lm(Mv ~  S0 + Ellipt + log.rho, data=gb)

qm4 <-lm(Mv ~  S0 + Ellipt + log.rho + Conc, data=gb)

# Con l'aggiunta di r.tidal si ha una perdità di significatività per le variabili Conc e log.rho

qm5 <-lm(Mv ~ Conc + log.rho + S0 + Ellipt + r.tidal, data=gb)
summary(qm5)
```

```
##
## Call:
## lm(formula = Mv ~ Conc + log.rho + S0 + Ellipt + r.tidal, data = gb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9249 -0.3392 -0.0111  0.2114  1.6352
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.899332   0.310762 -22.201  < 2e-16 ***
## Conc         0.217456   0.219705   0.990    0.325
## log.rho     -0.059555   0.098347  -0.606    0.546
## S0          -0.252177   0.021263 -11.860  < 2e-16 ***
## Ellipt       0.175015   0.023128   7.567 1.41e-11 ***
## r.tidal     -0.013835   0.001803  -7.672 8.30e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4557 on 107 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8676
## F-statistic: 147.7 on 5 and 107 DF,  p-value: < 2.2e-16
```

```
qm6 <-lm(Mv ~  S0 + Ellipt + r.tidal, data=gb)
summary(qm6)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + Ellipt + r.tidal, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00787 -0.32163 -0.01945  0.22724  1.63887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.802370   0.264688 -25.700  < 2e-16 ***
## S0          -0.255152   0.014602 -17.474  < 2e-16 ***
## Ellipt       0.175192   0.022955   7.632 9.36e-12 ***
## r.tidal     -0.012964   0.001216 -10.664  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.454 on 109 degrees of freedom
## Multiple R-squared:  0.8721, Adjusted R-squared:  0.8686
## F-statistic: 247.8 on 3 and 109 DF,  p-value: < 2.2e-16
```

```
# Risultano tutte significative così
# Potrei pensare di fermarmi al modello

# Mv ~  S0 + Ellipt + r.tidal


# Continuo aggiungendo ultima variabile rimasta, VHB

qm7 <-lm(Mv ~  S0 + Ellipt + r.tidal + VHB, data=gb)
summary(qm7)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + Ellipt + r.tidal + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07417 -0.05089  0.02477  0.04192  0.07317
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.6272764  0.0789016    7.950 1.95e-12 ***
## S0          -0.0053904  0.0029102   -1.852   0.0667 .
## Ellipt       0.9790718  0.0083513  117.236  < 2e-16 ***
```

```
## r.tidal     -0.0004487  0.0001771   -2.534   0.0127 *
## VHB         -0.9845515  0.0098062 -100.401  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04696 on 108 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.989e+04 on 4 and 108 DF,  p-value: < 2.2e-16
```

```
# SO perde di significatività


qm8 <-lm(Mv ~  Ellipt + r.tidal + VHB, data=gb)
summary(qm8)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + r.tidal + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08333 -0.05045  0.02305  0.04002  0.07432
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.7252633  0.0591881   12.254   <2e-16 ***
## Ellipt       0.9928901  0.0037949  261.639   <2e-16 ***
## r.tidal     -0.0002424  0.0001392   -1.742   0.0844 .
## VHB         -1.0000773  0.0051455 -194.360   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04748 on 109 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 2.594e+04 on 3 and 109 DF,  p-value: < 2.2e-16
```

```
# r.tidal perde di significatività

qm9 <-lm(Mv ~  Ellipt  + VHB, data=gb)
summary(qm9)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08596 -0.05064  0.02072  0.04438  0.06578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75787    0.05667   13.38   <2e-16 ***
## Ellipt       0.99509    0.00361  275.62   <2e-16 ***
```

```
## VHB         -1.00378    0.00473 -212.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04791 on 110 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9985
## F-statistic: 3.82e+04 on 2 and 110 DF,  p-value: < 2.2e-16
```

Alla fine poichè non ho altre variabili da aggiungere mi ritrovo con il modello :

$Mv \sim Ellipt + VHB$

Selezione modello con metodo di penalizzazione AIC (direzione forward, backward e both)

Successivamente ho deciso di utilizzare algoritmi iterativi per esplorare lo spazio dei modelli per trovare quelli più promettenti. Il criterio che ho utilizzato in questo caso è denominato "AIC" e l'esplorazione è stata fatta partendo dai modelli completo e con solo l'intercetta, in direzioni "forward", "backward" e "both".

```
# Do nome al modello nullo e il modello completo
qmc <-lm(Mv ~  r.tidal + Conc + log.rho + S0 + Ellipt + VHB, data=gb)
qm0 <-lm(Mv ~  1, data=gb)

back_aic <- step(qmc, scope=formula(qm0), direction ="backward", k=2)
```

```
## Start:  AIC=-684.65
## Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## - log.rho  1    0.0002  0.2335 -686.55
## - Conc     1    0.0024  0.2357 -685.50
## <none>                  0.2333 -684.65
## - S0       1    0.0077  0.2410 -682.97
## - r.tidal  1    0.0113  0.2446 -681.30
## - VHB      1   21.9898 22.2232 -171.77
## - Ellipt   1   29.9849 30.2182 -137.04
##
## Step:  AIC=-686.55
## Mv ~ r.tidal + Conc + S0 + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## <none>                  0.2335 -686.55
## - Conc     1    0.0046  0.2381 -686.35
## - S0       1    0.0104  0.2440 -683.61
## - r.tidal  1    0.0156  0.2491 -681.24
## - VHB      1   22.0658 22.2993 -173.38
## - Ellipt   1   30.1632 30.3967 -138.38
```

```
# Summary del modello ottenuto con direzione backward
summary(back_aic)
```

```
##
```

```
## Call:
## lm(formula = Mv ~ r.tidal + Conc + S0 + Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08061 -0.05092  0.02709  0.04034  0.05872
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.5973176  0.0811783    7.358    4e-11 ***
## r.tidal     -0.0004731  0.0001770   -2.673  0.00868 **
## Conc         0.0177047  0.0122155    1.449  0.15016
## S0          -0.0065786  0.0030093   -2.186  0.03099 *
## Ellipt       0.9783874  0.0083225  117.560  < 2e-16 ***
## VHB         -0.9835374  0.0097816 -100.549  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04672 on 107 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.607e+04 on 5 and 107 DF,  p-value: < 2.2e-16
```

```
forw_aic <- step(qm0, scope=formula(qmc), direction="forward", k=2)
```

```
## Start:  AIC=51.84
## Mv ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + S0       1   117.176  58.466 -70.461
## + Ellipt   1    72.007 103.635  -5.776
## + log.rho  1    35.137 140.505  28.617
## + Conc     1    24.800 150.841  36.639
## + r.tidal  1    16.582 159.060  42.634
## <none>                 175.642  51.840
## + VHB      1     1.001 174.641  53.194
##
## Step:  AIC=-70.46
## Mv ~ S0
##
##           Df Sum of Sq    RSS      AIC
## + r.tidal  1   23.9990 34.467 -128.176
## + Ellipt   1   12.5659 45.900  -95.805
## + log.rho  1    5.0047 53.461  -78.573
## <none>                 58.466  -70.461
## + VHB      1    0.4290 58.037  -69.293
## + Conc     1    0.0031 58.463  -68.467
##
## Step:  AIC=-128.18
## Mv ~ S0 + r.tidal
##
##           Df Sum of Sq    RSS     AIC
## + Ellipt   1   12.0034 22.463 -174.55
## + VHB      1    3.9255 30.541 -139.84
## <none>                 34.467 -128.18
```

```
## + Conc     1    0.0674 34.399 -126.40
## + log.rho  1    0.0048 34.462 -126.19
##
## Step:  AIC=-174.55
## Mv ~ S0 + r.tidal + Ellipt
##
##           Df Sum of Sq     RSS     AIC
## + VHB      1   22.2250  0.2381 -686.35
## <none>                 22.4631 -174.55
## + Conc     1    0.1638 22.2993 -173.38
## + log.rho  1    0.0365 22.4266 -172.74
##
## Step:  AIC=-686.35
## Mv ~ S0 + r.tidal + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## + Conc     1 0.0045847 0.23353 -686.55
## <none>                 0.23812 -686.35
## + log.rho  1 0.0024039 0.23571 -685.50
##
## Step:  AIC=-686.55
## Mv ~ S0 + r.tidal + Ellipt + VHB + Conc
##
##           Df  Sum of Sq     RSS     AIC
## <none>                  0.23353 -686.55
## + log.rho  1 0.00021603 0.23331 -684.65
```

```
# Summary del modello ottenuto con direzione forward
summary(forw_aic)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB + Conc, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08061 -0.05092  0.02709  0.04034  0.05872
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.5973176  0.0811783    7.358    4e-11 ***
## S0          -0.0065786  0.0030093   -2.186  0.03099 *
## r.tidal     -0.0004731  0.0001770   -2.673  0.00868 **
## Ellipt       0.9783874  0.0083225  117.560  < 2e-16 ***
## VHB         -0.9835374  0.0097816 -100.549  < 2e-16 ***
## Conc         0.0177047  0.0122155    1.449  0.15016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04672 on 107 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.607e+04 on 5 and 107 DF,  p-value: < 2.2e-16
```

```r
both_aic <- step(qm0, scope=formula(qmc), direction="both", k=2)
```

```
## Start:  AIC=51.84
## Mv ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + S0       1   117.176  58.466 -70.461
## + Ellipt   1    72.007 103.635  -5.776
## + log.rho  1    35.137 140.505  28.617
## + Conc     1    24.800 150.841  36.639
## + r.tidal  1    16.582 159.060  42.634
## <none>                 175.642  51.840
## + VHB      1     1.001 174.641  53.194
##
## Step:  AIC=-70.46
## Mv ~ S0
##
##           Df Sum of Sq     RSS      AIC
## + r.tidal  1    23.999  34.467 -128.176
## + Ellipt   1    12.566  45.900  -95.805
## + log.rho  1     5.005  53.461  -78.573
## <none>                 58.466  -70.461
## + VHB      1     0.429  58.037  -69.293
## + Conc     1     0.003  58.463  -68.467
## - S0       1   117.176 175.642   51.840
##
## Step:  AIC=-128.18
## Mv ~ S0 + r.tidal
##
##           Df Sum of Sq     RSS      AIC
## + Ellipt   1    12.003  22.463 -174.553
## + VHB      1     3.926  30.541 -139.840
## <none>                 34.467 -128.176
## + Conc     1     0.067  34.399 -126.397
## + log.rho  1     0.005  34.462 -126.192
## - r.tidal  1    23.999  58.466  -70.461
## - S0       1   124.593 159.060   42.634
##
## Step:  AIC=-174.55
## Mv ~ S0 + r.tidal + Ellipt
##
##           Df Sum of Sq    RSS     AIC
## + VHB      1    22.225  0.238 -686.35
## <none>                 22.463 -174.55
## + Conc     1     0.164 22.299 -173.38
## + log.rho  1     0.037 22.427 -172.74
## - Ellipt   1    12.003 34.467 -128.18
## - r.tidal  1    23.437 45.900  -95.81
## - S0       1    62.927 85.390  -25.66
##
## Step:  AIC=-686.35
## Mv ~ S0 + r.tidal + Ellipt + VHB
##
```

```
##           Df Sum of Sq     RSS     AIC
## + Conc     1    0.0046  0.2335 -686.55
## <none>                  0.2381 -686.35
## + log.rho  1    0.0024  0.2357 -685.50
## - S0       1    0.0076  0.2457 -684.82
## - r.tidal  1    0.0142  0.2523 -681.82
## - VHB      1   22.2250 22.4631 -174.55
## - Ellipt   1   30.3029 30.5410 -139.84
##
## Step:  AIC=-686.55
## Mv ~ S0 + r.tidal + Ellipt + VHB + Conc
##
##           Df Sum of Sq     RSS     AIC
## <none>                  0.2335 -686.55
## - Conc     1    0.0046  0.2381 -686.35
## + log.rho  1    0.0002  0.2333 -684.65
## - S0       1    0.0104  0.2440 -683.61
## - r.tidal  1    0.0156  0.2491 -681.24
## - VHB      1   22.0658 22.2993 -173.38
## - Ellipt   1   30.1632 30.3967 -138.38
```

```
# Summary del modello ottenuto con direzione both
summary(both_aic)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB + Conc, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08061 -0.05092  0.02709  0.04034  0.05872
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.5973176  0.0811783    7.358    4e-11 ***
## S0          -0.0065786  0.0030093   -2.186  0.03099 *
## r.tidal     -0.0004731  0.0001770   -2.673  0.00868 **
## Ellipt       0.9783874  0.0083225  117.560  < 2e-16 ***
## VHB         -0.9835374  0.0097816 -100.549  < 2e-16 ***
## Conc         0.0177047  0.0122155    1.449  0.15016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04672 on 107 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.607e+04 on 5 and 107 DF,  p-value: < 2.2e-16
```

Alla fine ottengo questi modelli:

- Modello direzione "backward" : $Mv \sim S0 + r.tidal + Ellipt + VHB + Conc$
- Modello direzione "forward" : $Mv \sim S0 + r.tidal + Ellipt + VHB + Conc$
- Modello direzione "both" : $Mv \sim S0 + r.tidal + Ellipt + VHB + Conc$

Selezione modello con metodo di penalizzazione BIC (direzione forward, backward e both)

Analogamente a quanto fatto in precedenza, ho analizzato i modelli ottenuti con il criterio "BIC", tramite un'esplorazione in direzioni "forward", "backward" e "both". Nel criterio di penalizzazione BIC si seleziona k = log(numero di osservazioni) che, quindi, dopo aver omesso i valori na, è pari a 113 per il dataset.

```
# Do nome al modello nullo e il modello completo
qmc <-lm(Mv ~  r.tidal + Conc + log.rho + S0 + Ellipt + VHB, data=gb)
qm0 <-lm(Mv ~  1, data=gb)

back_bic <- step(qmc, scope=formula(qm0), direction ="backward", k=log(113))
```

```
## Start:  AIC=-665.56
## Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## - log.rho  1    0.0002  0.2335 -670.18
## - Conc     1    0.0024  0.2357 -669.13
## - S0       1    0.0077  0.2410 -666.61
## <none>                  0.2333 -665.56
## - r.tidal  1    0.0113  0.2446 -664.93
## - VHB      1   21.9898 22.2232 -155.40
## - Ellipt   1   29.9849 30.2182 -120.68
##
## Step:  AIC=-670.18
## Mv ~ r.tidal + Conc + S0 + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## - Conc     1    0.0046  0.2381 -672.71
## <none>                  0.2335 -670.18
## - S0       1    0.0104  0.2440 -669.97
## - r.tidal  1    0.0156  0.2491 -667.60
## - VHB      1   22.0658 22.2993 -159.74
## - Ellipt   1   30.1632 30.3967 -124.74
##
## Step:  AIC=-672.71
## Mv ~ r.tidal + S0 + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## - S0       1    0.0076  0.2457 -673.91
## <none>                  0.2381 -672.71
## - r.tidal  1    0.0142  0.2523 -670.91
## - VHB      1   22.2250 22.4631 -163.64
## - Ellipt   1   30.3029 30.5410 -128.93
##
## Step:  AIC=-673.91
## Mv ~ r.tidal + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## - r.tidal  1     0.007   0.253 -675.53
## <none>                  0.246 -673.91
## - VHB      1    85.144  85.390  -17.48
## - Ellipt   1   154.294 154.539   49.56
```

```
##
## Step:  AIC=-675.53
## Mv ~ Ellipt + VHB
##
##          Df Sum of Sq     RSS     AIC
## <none>                  0.253 -675.53
## - VHB     1    103.38 103.635   -0.32
## - Ellipt  1    174.39 174.641   58.65
```

```
# Summary del modello ottenuto con direzione backward
summary(back_bic)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08596 -0.05064  0.02072  0.04438  0.06578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75787    0.05667   13.38   <2e-16 ***
## Ellipt       0.99509    0.00361  275.62   <2e-16 ***
## VHB         -1.00378    0.00473 -212.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04791 on 110 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9985
## F-statistic: 3.82e+04 on 2 and 110 DF,  p-value: < 2.2e-16
```

```
forw_bic <- step(qm0, scope=formula(qmc), direction="forward", k = log(113))
```

```
## Start:  AIC=54.57
## Mv ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + S0       1   117.176  58.466 -65.007
## + Ellipt   1    72.007 103.635  -0.321
## + log.rho  1    35.137 140.505  34.072
## + Conc     1    24.800 150.841  42.094
## + r.tidal  1    16.582 159.060  48.088
## <none>                 175.642  54.567
## + VHB      1     1.001 174.641  58.649
##
## Step:  AIC=-65.01
## Mv ~ S0
##
##           Df Sum of Sq    RSS      AIC
## + r.tidal  1   23.9990 34.467 -119.994
## + Ellipt   1   12.5659 45.900  -87.623
## + log.rho  1    5.0047 53.461  -70.391
```

```
## <none>                 58.466  -65.007
## + VHB      1    0.4290 58.037  -61.111
## + Conc     1    0.0031 58.463  -60.285
##
## Step:  AIC=-119.99
## Mv ~ S0 + r.tidal
##
##           Df Sum of Sq    RSS     AIC
## + Ellipt   1   12.0034 22.463 -163.64
## + VHB      1    3.9255 30.541 -128.93
## <none>                 34.467 -119.99
## + Conc     1    0.0674 34.399 -115.49
## + log.rho  1    0.0048 34.462 -115.28
##
## Step:  AIC=-163.64
## Mv ~ S0 + r.tidal + Ellipt
##
##           Df Sum of Sq     RSS     AIC
## + VHB      1   22.2250  0.2381 -672.71
## <none>                 22.4631 -163.64
## + Conc     1    0.1638 22.2993 -159.74
## + log.rho  1    0.0365 22.4266 -159.10
##
## Step:  AIC=-672.71
## Mv ~ S0 + r.tidal + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## <none>                 0.23812 -672.71
## + Conc     1 0.0045847 0.23353 -670.18
## + log.rho  1 0.0024039 0.23571 -669.13
```

```
# Summary del modello ottenuto con direzione forward
summary(forw_bic)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07417 -0.05089  0.02477  0.04192  0.07317
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  0.6272764  0.0789016    7.950 1.95e-12 ***
## S0          -0.0053904  0.0029102   -1.852   0.0667 .
## r.tidal     -0.0004487  0.0001771   -2.534   0.0127 *
## Ellipt       0.9790718  0.0083513  117.236  < 2e-16 ***
## VHB         -0.9845515  0.0098062 -100.401  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04696 on 108 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
```

```
## F-statistic: 1.989e+04 on 4 and 108 DF,  p-value: < 2.2e-16
```

```
both_bic <- step(qm0, scope=formula(qmc), direction="both", k=log(113))
```

```
## Start:  AIC=54.57
## Mv ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + S0       1   117.176  58.466 -65.007
## + Ellipt   1    72.007 103.635  -0.321
## + log.rho  1    35.137 140.505  34.072
## + Conc     1    24.800 150.841  42.094
## + r.tidal  1    16.582 159.060  48.088
## <none>                 175.642  54.567
## + VHB      1     1.001 174.641  58.649
##
## Step:  AIC=-65.01
## Mv ~ S0
##
##           Df Sum of Sq     RSS      AIC
## + r.tidal  1    23.999  34.467 -119.994
## + Ellipt   1    12.566  45.900  -87.623
## + log.rho  1     5.005  53.461  -70.391
## <none>                 58.466  -65.007
## + VHB      1     0.429  58.037  -61.111
## + Conc     1     0.003  58.463  -60.285
## - S0       1   117.176 175.642   54.567
##
## Step:  AIC=-119.99
## Mv ~ S0 + r.tidal
##
##           Df Sum of Sq     RSS      AIC
## + Ellipt   1    12.003  22.463 -163.643
## + VHB      1     3.926  30.541 -128.930
## <none>                 34.467 -119.994
## + Conc     1     0.067  34.399 -115.488
## + log.rho  1     0.005  34.462 -115.282
## - r.tidal  1    23.999  58.466  -65.007
## - S0       1   124.593 159.060   48.088
##
## Step:  AIC=-163.64
## Mv ~ S0 + r.tidal + Ellipt
##
##           Df Sum of Sq    RSS     AIC
## + VHB      1    22.225  0.238 -672.71
## <none>                 22.463 -163.64
## + Conc     1     0.164 22.299 -159.74
## + log.rho  1     0.037 22.427 -159.10
## - Ellipt   1    12.003 34.467 -119.99
## - r.tidal  1    23.437 45.900  -87.62
## - S0       1    62.927 85.390  -17.48
##
## Step:  AIC=-672.71
## Mv ~ S0 + r.tidal + Ellipt + VHB
```

```
##
##           Df Sum of Sq     RSS     AIC
## - S0       1    0.0076  0.2457 -673.91
## <none>                  0.2381 -672.71
## - r.tidal  1    0.0142  0.2523 -670.91
## + Conc     1    0.0046  0.2335 -670.18
## + log.rho  1    0.0024  0.2357 -669.13
## - VHB      1   22.2250 22.4631 -163.64
## - Ellipt   1   30.3029 30.5410 -128.93
##
## Step:  AIC=-673.91
## Mv ~ r.tidal + Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## - r.tidal  1     0.007   0.253 -675.53
## <none>                   0.246 -673.91
## + S0       1     0.008   0.238 -672.71
## + Conc     1     0.002   0.244 -669.97
## + log.rho  1     0.000   0.246 -669.20
## - VHB      1    85.144  85.390  -17.48
## - Ellipt   1   154.294 154.539   49.56
##
## Step:  AIC=-675.53
## Mv ~ Ellipt + VHB
##
##           Df Sum of Sq     RSS     AIC
## <none>                   0.253 -675.53
## + r.tidal  1     0.007   0.246 -673.91
## + log.rho  1     0.003   0.249 -672.19
## + Conc     1     0.002   0.250 -671.91
## + S0       1     0.000   0.252 -670.91
## - VHB      1   103.382 103.635   -0.32
## - Ellipt   1   174.388 174.641   58.65
```

```r
# Summary del modello ottenuto con direzione both
summary(both_bic)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = gb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08596 -0.05064  0.02072  0.04438  0.06578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75787    0.05667   13.38   <2e-16 ***
## Ellipt       0.99509    0.00361  275.62   <2e-16 ***
## VHB         -1.00378    0.00473 -212.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04791 on 110 degrees of freedom
```

```
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9985
## F-statistic: 3.82e+04 on 2 and 110 DF,  p-value: < 2.2e-16
```

Alla fine ottengo questi modelli:

- Modello direzione "backward" :$Mv \sim Ellipt + VHB$
- Modello direzione "forward" : $Mv \sim S0 + r.tidal + Ellipt + VHB$
- Modello direzione "both" : $Mv \sim Ellipt + VHB$

Confronto modelli

I modelli ottenuti con criterio del p-value sono:

- Modello direzione "backward" e "both" $Mv \sim Ellipt + VHB$: le variabili risultano estremamente significative (p-value <2e-16) con un elevato Adjusted R-squared (0.9985). La variabile "VHB" ha un effetto negativo (-1.00378) mentre la variabile "Ellipt" (0.99509) un effetto positivo, sulla magnitudine, relativamente elevati.

  Questo modello è sicuramente uno dei migliori dal punto di vista della significatività e interpretabilità, uno degli scopi di questo elaborato. Ciò lo rende uno dei possibili candidati come scelta finale del caso di studio.

- Modello direzione "forward" $Mv \quad S0 + Ellipt + log.rho + Conc$: le variabili risultano tutte estremamente significative. I parametri "S0" e "Conc" hanno un effetto negativo sulla variabile obiettivo ("S0" -0.32392, "Conc" -0.86375) mentre Ellipt (0.18629) e "log.rho" (0.49500) hanno un effetto positivo. Il più significativo è "S0". Questo modello appare come molto significativo ed esplicativo e rappresenta una possibile scelta ai fini dello studio.

I modelli ottenuto con criterio "AIC" sono caratterizzati dalla bassa significatività e l'elevato numero di variabili. Utilizzando il criterio di penalizzazione "AIC", con direzioni "both","forward" e "backward" ho ottenuto lo stesso metodo $Mv \sim S0 + r.tidal + Ellipt + VHB + Conc$. In questo modello, la variabile "Conc" è risultata non significativa (0.15016) e altre due variabili risultavano con un p-value significativo ma alto, "S0"(0.15016) e "r.tidal" (0.00868). Quindi, è stato scartato per la mia scelta finale, sia per la scarsa significatività di alcune variabili, sia perchè non rappresenta un modello facilmente interpretabile per la grande quantità di variabili.

I modelli ottenuti con criterio "BIC" :

- Modello direzione "backward" e "both" $Mv \sim Ellipt + VHB$: stesso modello ottenuto con il criterio del p-value, direzione "both", quindi esposto precedentemente.
- Modello direzione "forward" : $Mv \sim S0 + r.tidal + Ellipt + VHB$: la variabile "S0" non risulta significativa (p-value 0.0667) e "r.tidal" ha un p-value di 0.0127, che seppur significativo non lo è altamente. Anche questo modello quindi è stato scartato.

Alla fine i modelli più promettenti, in quanto gli unici dove tutte le variabili sono significative, sono $Mv \sim Ellipt + VHB$ e $Mv \sim S0 + Ellipt + log.rho + Conc$. Il primo è caratterizzato dalla sua semplicità e alta significatività. Il secondo è risultato molto significativo ma allo stesso tempo ha un Adjusted R-squared definitivamente peggiore (0.7966 rispetto a 0.9985 del modello con solo due variabili).

Il modello che quindi rispecchia maggiormente l'obiettivo di interpretabilità, semplicità e parsimonia è quello ottenuto minimizzando "BIC" (direzione "backward" e "both") e p-value (anche esso direzione "backward" e "both"). Per confermare tale scelta ho deciso di fare uno studio grafico.

Modelli grafici

Grafi orientati: Bayesian Networks

Come secondo approccio di studio ho deciso di utilizzare il modello della rete bayesiana, modello grafico probabilistico che rappresenta un insieme di variabili stocastiche con le loro dipendenze condizionali attraverso l'uso di un grafo aciclico diretto (DAG).

Ho optato per l'utilizzo dell'algoritmo Hill Climbing, algoritmo greedy che massimizza una funzione di score. In particolare, dato che il mio obiettivo è quello di ottenere un grafo coerente con l'analisi effettuata precedentemente ho deciso di concentrarmi sul criterio di penalizzazione "BIC" ma ho analizzato anche i risultati ottenuti con il criterio "AIC".

In una prima fase esplorativa ho costruito il DAG a partire da tutte le variabili del dataset "gb", includendo anche quelle che, dopo la regressione semplice, erano state escluse dal modello completo.

Per prima cosa ho provato a costruire il DAG dicotomizzando le variabili continue presenti nel mio dataset. Tuttavia dicotomizzare le variabili ha restituito un grafo non interpretabile.

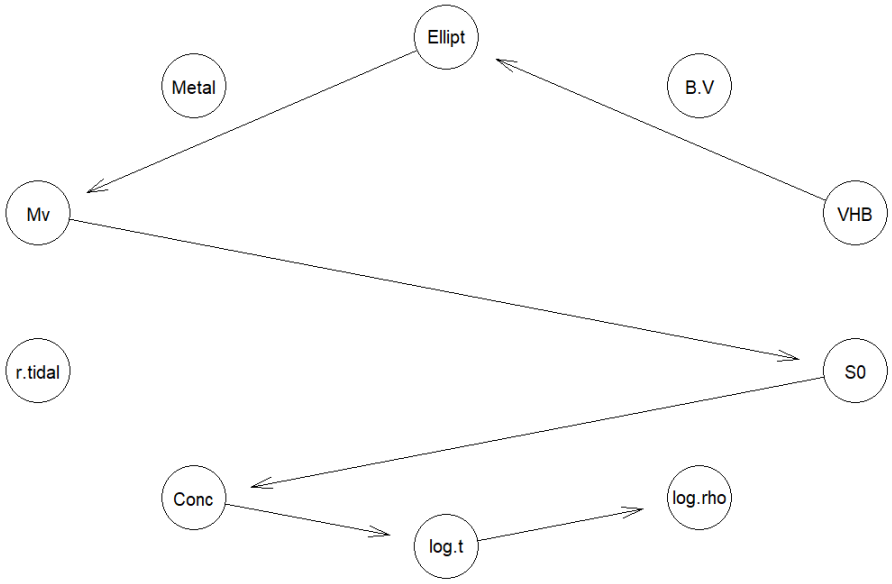
Ho deciso dunque di provare altre suddivisioni di seguito il metodo utilizzato.

```
library(bnlearn)
gb_disc3 <- gb
gb_disc3 <- data.frame(gb_disc3)

# Discretizzazione delle variabili nel dataframe gb
num_bins <- 3

# Ciclo attraverso le colonne e discretizzazione
for (col in names(gb_disc3)) {
  gb_disc3[[col]] <- cut(gb_disc3[[col]], breaks = num_bins, labels = FALSE)
}

gb_disc3 <- lapply(gb_disc3, as.factor)
gb_disc3 <- as.data.frame(gb_disc3)

# Summary del dataframe discretizzato
summary(gb_disc3)
```

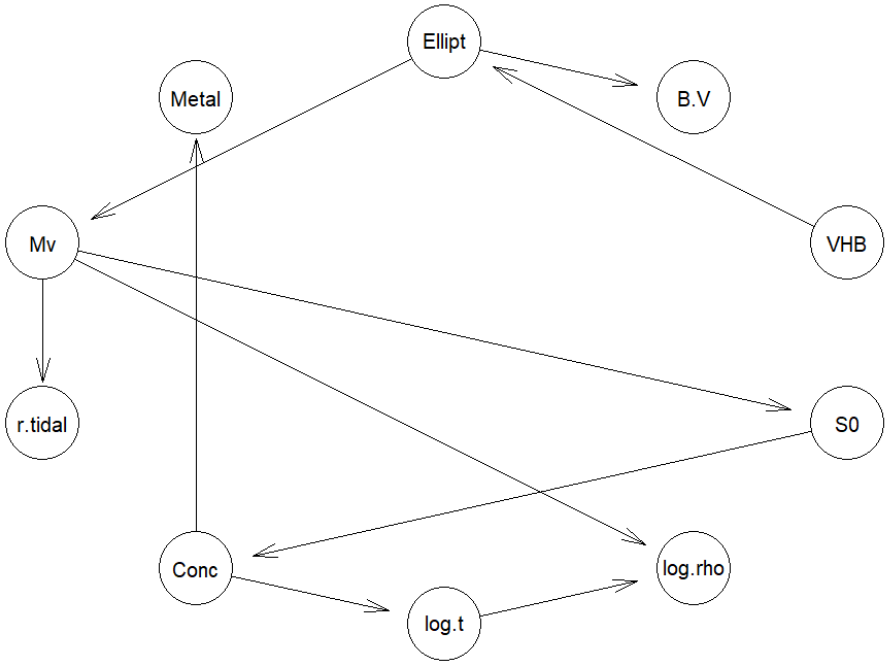
```
##  Metal   Mv     r.tidal Conc   log.t  log.rho S0     VHB    B.V     Ellipt
##  1:42   1:33   1:107   1:28   1:32   1:15    1:72   1:66   1:104   1:38
##  2:52   2:73   2:  5   2:67   2:62   2:46    2:33   2:44   2:  8   2:68
##  3:19   3: 7   3:  1   3:18   3:19   3:52    3: 8   3: 3   3:  1   3: 7
```

```
gb_disc3.bn <- hc(gb_disc3, score = 'bic')  # hill climbing
```

Di seguito la Bayesian Network ottenuta tramite criterio “BIC”, con le variabili continue divise in terzili.



Di seguito la Bayesian Network generata con criterio “AIC”, suddividendo le variabili continue in terzili.



Anche se sono stati utilizzati criteri diversi, l’unica apparente dipendenza di “MV” è quella dalla variabile “Ellipt” a sua volta dipendente dalla variabile “VHB”. Curiosamente, in questo grafico, “Mv” è solo indirettamente collegato a “VHB”, mentre il modello di regressione lineare sembrava premiare la sua presenza diretta.

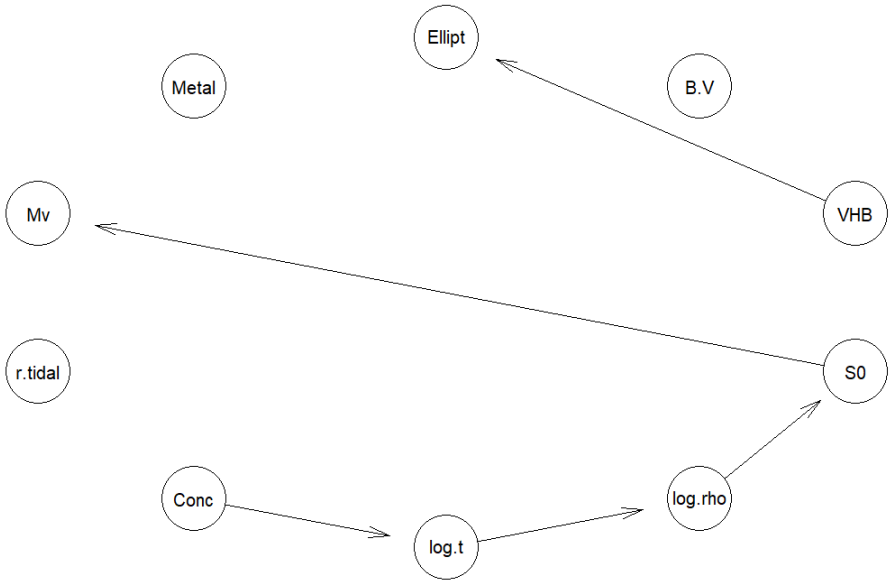
Ho provato ad aumentare il numero di suddivisioni in cui sono state discretizzate le variabili.

```
gb_disc4 <- gb
gb_disc4 <- data.frame(gb_disc4)
# Discretizzazione delle variabili nel dataframe gb
num_bins <- 4

# Ciclo attraverso le colonne e discretizzazione
for (col in names(gb_disc4)) {
  gb_disc4[[col]] <- cut(gb_disc4[[col]], breaks = num_bins, labels = FALSE)
}

gb_disc4 <- lapply(gb_disc4, as.factor)
gb_disc4 <- as.data.frame(gb_disc4)

# Summary del dataframe discretizzato
summary(gb_disc4)
```

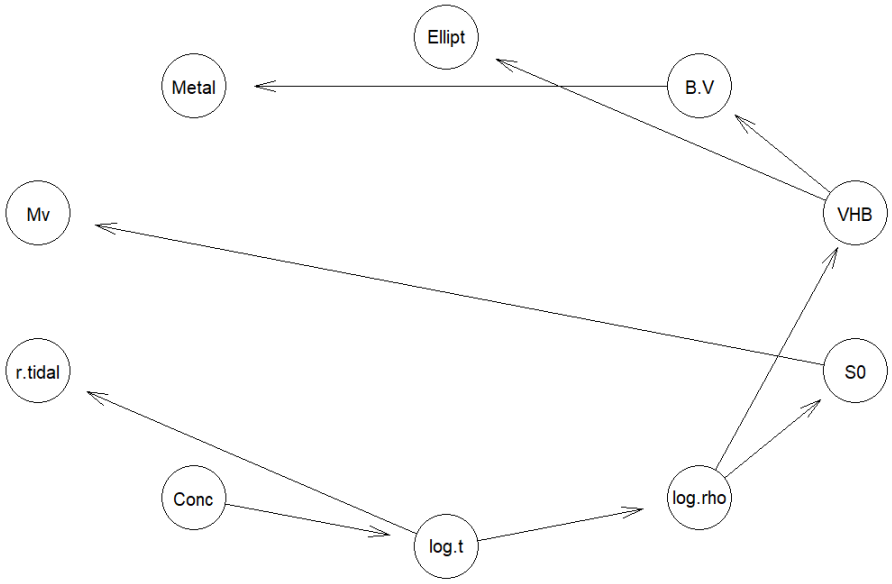
```
##  Metal  Mv     r.tidal Conc   log.t  log.rho S0     VHB    B.V    Ellipt
##  1:24   1:19   1:101   1:22   1:12   1:10    1:53   1:33   1:97   1:17
##  2:53   2:64   2: 11   2:46   2:49   2:16    2:46   2:70   2:15   2:70
##  3:25   3:24   4:  1   3:32   3:43   3:57    3:13   3: 9   4: 1   3:20
##  4:11   4: 6           4:13   4: 9   4:30    4: 1   4: 1          4: 6
```

```
gb_disc4.bn <- hc(gb_disc4, score = 'bic')  # hill climbing
```

Di seguito la Bayesian Network generata con criterio "BIC" e le variabili continue divise in quartili.



Di seguito la Bayesian Network generata con criterio "AIC" e le variabili continue divise in quartili.



Si nota come, aumentando il numero di suddivisioni, le dipendenze cambiano. Nello specifico non si ha più la dipendenza da "Ellipt" ma emerge la sola dipendenza dalla variabile "S0". Quest'ultima risultava altamente significativa nel modello di regressione lineare semplice per la variabile obiettivo "Mv".

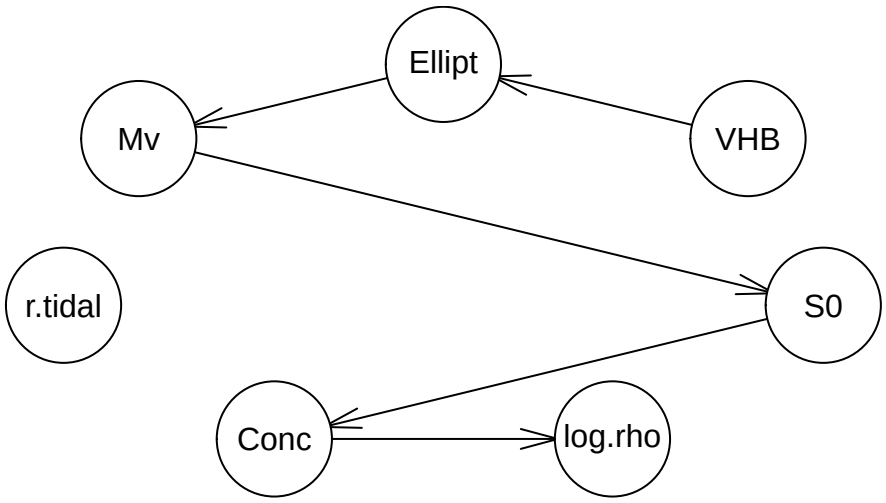
Successivamente, ho costruito la Bayesian Network utilizzando soltanto i parametri presenti nel modello completo, ossia quelli che erano risultati più significativi.

```
gb_di3 <- gb[c(-1,-5,-9)]
gb_di3 <- data.frame(gb_di3)
# Discretizzazione delle variabili nel dataframe ridotto gb
num_bins <- 3

# Ciclo attraverso le colonne e discretizzazione
for (col in names(gb_di3)) {
  gb_di3[[col]] <- cut(gb_di3[[col]], breaks = num_bins, labels = FALSE)
}

gb_di3 <- lapply(gb_di3, as.factor)
gb_di3 <- as.data.frame(gb_di3)

# Summary del dataframe discretizzato
summary(gb_di3)
```

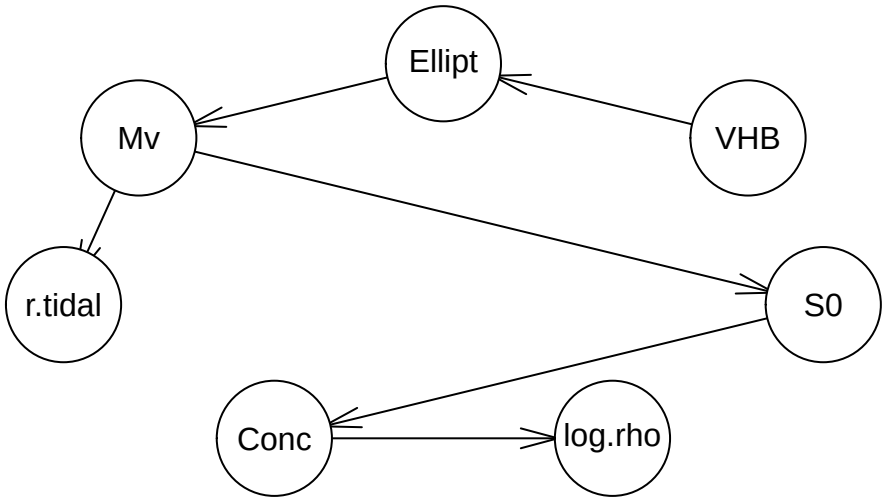
```
##  Mv     r.tidal Conc   log.rho S0     VHB    Ellipt
##  1:33   1:107   1:28   1:15    1:72   1:66   1:38
##  2:73   2:  5   2:67   2:46    2:33   2:44   2:68
```

```
##  3: 7   3:  1   3:18   3:52    3: 8   3: 3   3: 7
```

```r
gb_bic.bn <- hc(gb_di3, score = 'bic')  # hill climbing

# Di seguito la Bayesian Network generata con criterio "BIC"
# costruita con le variabili risultate più significative, suddivise in terzili

plot(gb_bic.bn)
```




```r
gb_aic.bn <- hc(gb_di3, score = 'aic')

# Di seguito la Bayesian Network generata con criterio "AIC"
# costruita con le variabili risultate più significative, suddivise in terzili

plot(gb_aic.bn)
```

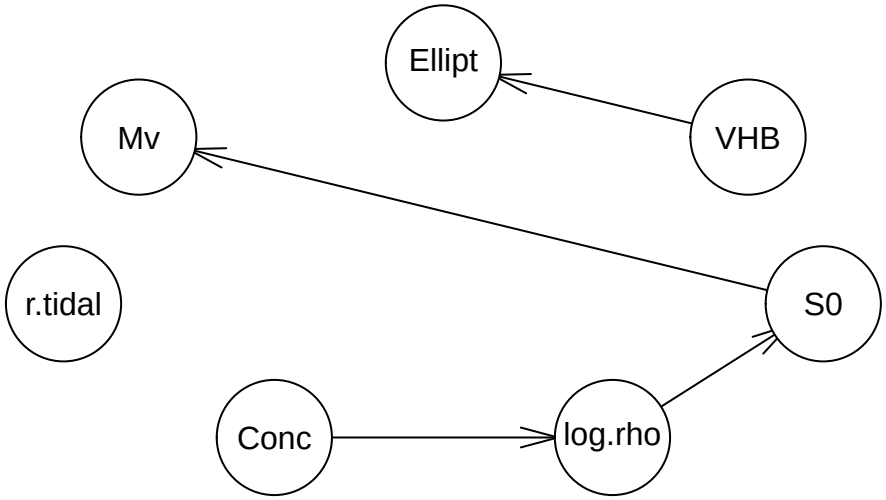


Come avevo fatto precedentemente per tutto il dataset, ho aumentato il numero di suddivisioni delle mie variabili continue.

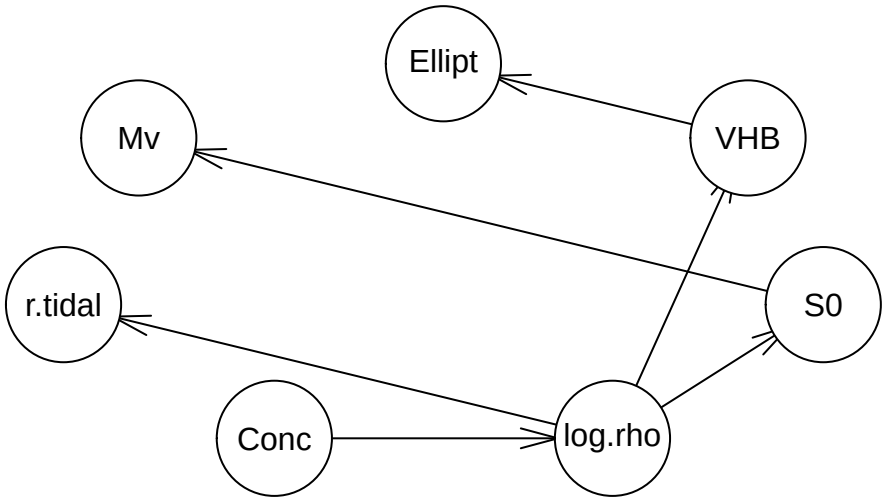
```
gb_di4 <- gb[c(-1,-5,-9)]
gb_di4 <- data.frame(gb_di4)
# Discretizzazione delle variabili nel dataframe ridotto gb
num_bins <- 4

# Ciclo attraverso le colonne e discretizzazione
for (col in names(gb_di4)) {
  gb_di4[[col]] <- cut(gb_di4[[col]], breaks = num_bins, labels = FALSE)
}

gb_di4 <- lapply(gb_di4, as.factor)
gb_di4 <- as.data.frame(gb_di4)

# Summary del dataframe discretizzato
summary(gb_di4)
```

```
##  Mv     r.tidal Conc   log.rho S0     VHB    Ellipt
##  1:19   1:101   1:22   1:10    1:53   1:33   1:17
##  2:64   2: 11   2:46   2:16    2:46   2:70   2:70
##  3:24   4:  1   3:32   3:57    3:13   3: 9   3:20
##  4: 6           4:13   4:30    4: 1   4: 1   4: 6
```

```r
gb_bic4.bn <- hc(gb_di4, score = 'bic')  # hill climbing

# Di seguito la Bayesian Network generata con criterio "BIC"
# costruita con le variabili risultate più significative, suddivise in quartili

plot(gb_bic4.bn)
```




```r
gb_aic4.bn <- hc(gb_di4, score = 'aic')

# Di seguito la Bayesian Network generata con criterio "AIC"
# costruita con le variabili risultate più significative, suddivise in quartili

plot(gb_aic4.bn)
```



Osservo che aumentando le suddivisioni, anche nel DAG costruito a partire da solo le variabili più significative, sparisce la dipendenza da "Ellipt" per essere sostituita da una dipendenza da "S0".

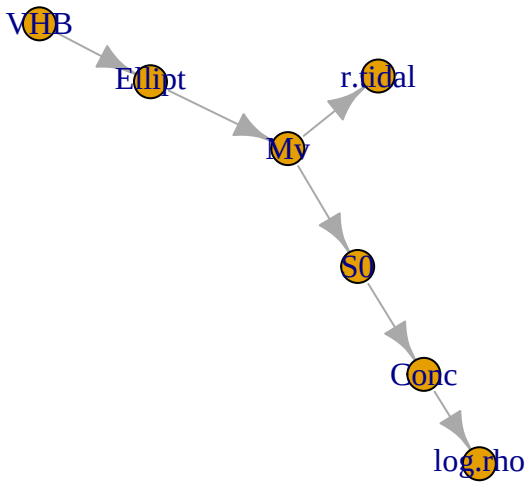
E' interessante osservare come piccole variazioni nelle suddivisioni hanno portato a cambiamenti notevoli, per via delle differenti sensibilità di ciascuna suddivisione rispetto alla distribuzione dei dati del dataset.

Dopo aver osservato il grafo ottenuto con il metodo Hill Climbing, con suddivisione delle variabili continue in terzili (quello che meglio rispecchia il modello lineare), ho utilizzato il comando "dag" per ricostruire il DAG ottenuto. La motivazione è che tale comando offre una rappresentazione esplicita del modello probabilistico appreso dall'algoritmo di Hill Climbing e in particolare consente di sfruttare altre funzioni statistiche, come il comando "querygrain", per effettuare osservazioni specifiche sulla distribuzione delle variabili.

```
library(gRain)
```

```
## Caricamento del pacchetto richiesto: gRbase

##
## Caricamento pacchetto: 'gRbase'

## I seguenti oggetti sono mascherati da 'package:bnlearn':
##
##     ancestors, children, nodes, parents
```

```
library(gRim)
library(ggm)
gdag1 <- dag(~Mv*Ellipt + Ellipt*VHB + S0*Mv + r.tidal*Mv + Conc*S0 + log.rho*Conc)
plot(gdag1)
```




```
gb_bn <- grain(gdag1, data = gb_di3)
```

La funzione "querygrain" viene utilizzata per eseguire una query sulla rete bayesiana "gb_bn". Il parametro "nodes" specifica le variabili le cui probabilità sono di interesse. Il parametro "type = conditional" indica che l'obiettivo è osservare le probabilità condizionali del DAG.

Le tabelle ottenute riportano come la probabilità di un valore della prima variabile può variare in base ai diversi valori possibili della seconda variabile (o viceversa) secondo le relazioni specifiche definite nella struttura della rete bayesiana.

```
querygrain(gb_bn, nodes=c("Mv", "Ellipt"), type="conditional")
```

```
##    Ellipt
## Mv         1         2         3
##   1 0.6052632 0.14705882 0.0000000
##   2 0.3947368 0.79411765 0.5714286
##   3 0.0000000 0.05882353 0.4285714
```

La tabella fornisce un'indicazione delle relazioni probabilistiche tra "Mv" ed "Ellipt". Se i valori di "Ellipt" sono bassi (nel primo terzile) la probabilità che "Mv" abbia valori bassi è 0.6052632. Si nota come, effettivamente, al crescere dei valori di "Ellipt" la probabilità che il valore di "Mv" sia basso diminuisce, rispettando l'ipotesi di dipendenza fra i parametri. Osservo, in particolare, che se il valore di "Ellipt" appartiene al terzo terzile la probabilità che il valore di "Mv" appartenga al primo terzile risulta pari a 0.

```r
querygrain(gb_bn, nodes=c("Mv", "VHB"), type="conditional")
```

```
##    VHB
## Mv           1          2         3
##   1 0.36921850 0.19617225 0.0000000
##   2 0.60047847 0.71941217 0.5714286
##   3 0.03030303 0.08441558 0.4285714
```

```r
querygrain(gb_bn, nodes=c("Ellipt", "VHB"), type="conditional")
```

```
##       VHB
## Ellipt         1          2 3
##      1 0.4848485 0.13636364 0
##      2 0.5151515 0.77272727 0
##      3 0.0000000 0.09090909 1
```

Il modello ottenuto da questo studio grafico sembra confermare la plausibilità del modello $Mv \sim Ellipt + VHB$ ottenuto con i metodi precedenti, anche se nel caso grafico la dipendenza di "VHB" non è diretta. Tuttavia anche se sono stato utilizzati due approcci differenti è emersa l'importanza di queste due variabili in uno studio della stima di "Mv".

Conclusione

Lo studio effettuato ha presentato delle difficoltà soprattutto durante le fasi iniziali del progetto, per via della complessità dell'argomento e dell'elevato numero di variabili nel dataset. Nonostante ciò, ha prodotto risultati soddisfacenti, allineati con l'obiettivo prefissato. Il modello identificato, si è dimostrato un equilibrato compromesso tra espressività e parsimonia.

L'osservazione di una dipendenza lineare tra la magnitudine assoluta, l'ellitticità ("Ellipt") e il livello del ramo orizzontale ("VHB") fornisce importanti intuizioni sulla natura intrinseca dei cluster globulari. L'ellitticità emerge come un fattore chiave nella determinazione della magnitudine assoluta, suggerendo che la geometria del cluster globulare può influire direttamente sulla sua luminosità intrinseca. Inoltre, la relazione con il livello del ramo orizzontale evidenzia il ruolo significativo dell'età e dell'evoluzione stellare nella determinazione della magnitudine assoluta dei cluster. In particolare, all'aumentare dell'ellitticità del cluster, la luminosità diminuisce, mentre all'aumentare del livello del ramo orizzontale, la luminosità aumenta.