Relazione progetto Statistical Modeling

Mirko Bicchierai

2024-06-28

Contents

1 Introduzione

1.1 Introduzione al Dataset

Il dataset analizzato nel mio studio è "Globclus_prop", che contiene le misurazioni di 20 proprietà astronomiche e astrofisiche relative a 147 ammassi globulari nella Via Lattea. Queste misurazioni provengono dal catalogo di Webbink (1985), considerato un punto di riferimento essenziale per l'analisi di queste strutture celesti.

1.2 Variabili coinvolte e metodologia di Studio

Data la complessità degli studi astrofisici, ho deciso di iniziare con un approfondimento delle variabili presenti nel dataset, per comprenderne il significato fondamentale. Di seguito, è riportata una lista dettagliata delle variabili coinvolte, insieme alle rispettive 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

1.3 Obbiettivo dello Studio

Dopo aver approfondito il significato scientifico delle proprietà presenti nel dataset, ho deciso di focalizzare questo progetto sull'analisi della magnitudine assoluta degli ammassi globulari, rappresentata dalla variabile "Mv", seguendo quindi l'indicazione dell'esercizio 1 dell'appendice C.7 del libro "Modern statistical methods

with R applications", per la quale si specificava la richiesta di effettuare regressioni sulle variabili strutturali del cluster nel dataset, ponendo la variabile "Mv" come risposta dei modelli utilizzati.

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

L'obiettivo di questo studio è capire quali variabili nel dataset "Globclus_prop" influenzino la magnitudine assoluta e in che modo.

La magnitudine apparente in una determinata banda x è definita come: $m_x = -2.5log(F_x) + C$, dove F è il flusso osservato nella banda x, e C è una costante dipendente dalla banda in cui l'oggetto è osservato e nel visibile ha un valore di circa 0.941. Per calcolare la magnitudine assoluta (Mv) data quella apparente (m), è necessario ricordare che la luminosità di un oggetto è inversamente proporzionale al quadrato della sua distanza. Ne segue che la differenza fra la magnitudine apparente e la magnitudine assoluta di un oggetto sarà espressa dalla seguente formula: $Mv = m + 5 - 5log(d)$ dove d è la distanza della stella espressa in parsec.

1.4 Operazioni Preliminari

Prima di procedere alla fase di analisi vera e propria del dataset, sono state effettuate alcune operazioni preliminari. Data l'elevata presenza di osservazioni con misurazioni mancanti per una o più variabili, si è deciso di rimuovere tali osservazioni utilizzando il comando "na.omit" di R. Questo ha ridotto il dataset da 147 a 113 osservazioni, rendendolo più significativo per l'obiettivo finale di previsione.

Inoltre, è stato rinominato il dataset per facilitarne l'utilizzo durante le analisi successive. I nomi delle variabili non sono stati modificati, in quanto risultavano già sufficientemente chiari e descrittivi.

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

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
##
```

1.5 Selezione delle Variabili

Nel contesto di questo dataset caratterizzato da una complessità elevata dovuta al gran numero di variabili, è stata avviata un'esplorazione preliminare mirata alla loro riduzione per semplificare l'analisi.

La prima variabile eliminata è stata "Name", ritenuta insignificante per qualsiasi analisi inferenziale. Inoltre, dato che l'attenzione è focalizzata sulla variabile "Mv" (magnitudine assoluta), è stata scelta di trascurare "V.t", la magnitudine integrata, poiché rappresenta semplicemente un altro modo di descrivere la magnitudine di un ammasso globulare (specificamente la magnitudine che l'oggetto avrebbe se fosse compresso in un singolo punto luminoso), senza aggiungere rilevanza allo studio.

Successivamente, dopo un'analisi della matrice di covarianza (di cui l'output è omesso per le dimensioni considerevoli), è stata eliminata una delle due variabili correlate che rappresentavano la distanza del corpo celeste: "R.sol" o "R.GC", poiché presentavano una correlazione molto alta (0.97). Per questo studio, è stata mantenuta come misura della distanza la variabile "R.GC", che rappresenta la distanza dal centro galattico, anziché quella rispetto al sole.

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

```
clusters <- clusters[c(-1,-4,-19)]
```

Nella fase successiva, verranno analizzate ed esplorate le relazioni tra le varie variabili del dataset. Per agevolare questo studio, il dataset è stato suddiviso in diverse sottocategorie, data la sua alta complessità. Questo approccio è stato adottato per facilitare la visualizzazione e l'analisi, mantenendo in ogni gruppo la variabile "Mv" come riferimento. Questa analisi permetterà di comprendere le relazioni esistenti tra le variabili, con l'obiettivo di eliminare alcune di esse dal modello statistico che verrà utilizzato.

1.6 Primo cluster di variabili, posizione dei corpi celesti

Questo cluster è stato progettato per raggruppare tutte le variabili posizionali dei corpi celesti registrati nel dataset. È composto esclusivamente dalle variabili "Mv" e dalle coordinate di posizione, quali "Gal.long",

"Gal.lat" e "R.GC". Attraverso la stampa degli scatter plot delle variabili a coppie, è possibile osservare se esiste qualche tipo di relazione tra le variabili, effettuando una semplice analisi grafica.

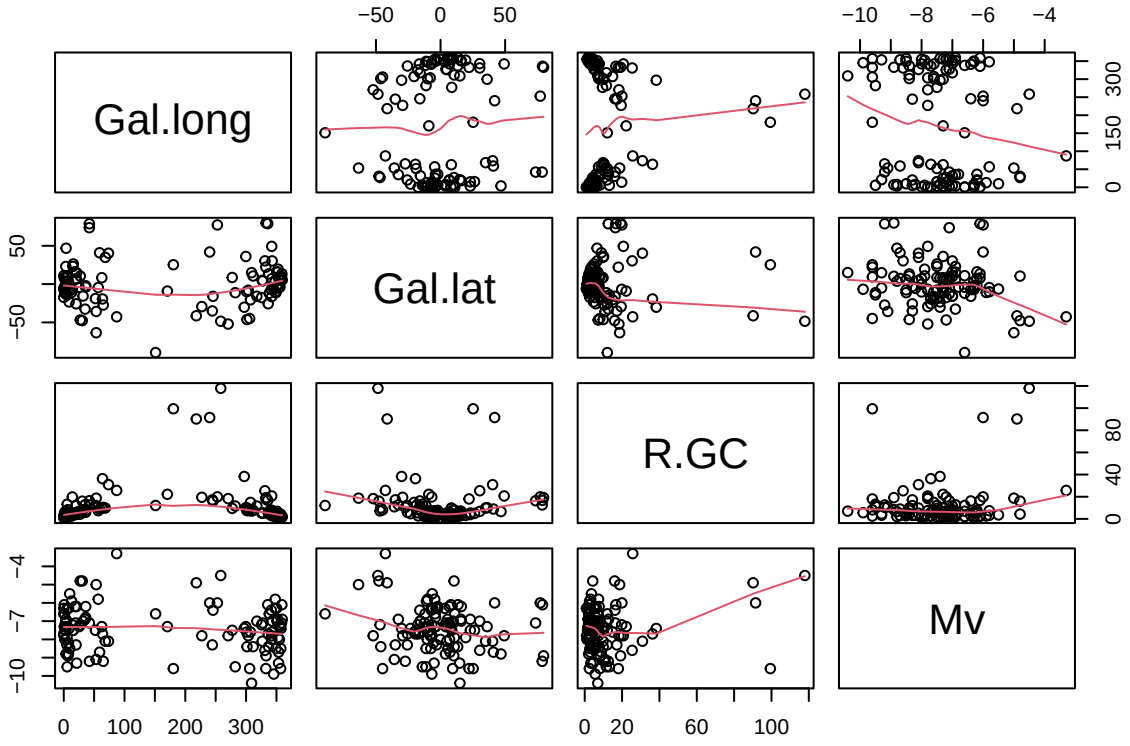
Di seguito è riportato il codice necessario per creare il primo cluster a partire dal dataset originario. Questo cluster raggruppa tutte le variabili posizionali dei corpi celesti:

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

```
##     Gal.long          Gal.lat              R.GC              Mv
##  Min.   :  0.07   Min.   :-89.3800   Min.   :  0.90   Min.   :-10.400
##  1st Qu.: 15.14   1st Qu.:-14.0900   1st Qu.:  3.10   1st Qu.: -8.300
##  Median :151.15   Median : -3.8700   Median :  6.00   Median : -7.400
##  Mean   :169.23   Mean   : -0.2057   Mean   : 11.66   Mean   : -7.431
##  3rd Qu.:332.97   3rd Qu.: 10.7100   3rd Qu.: 12.10   3rd Qu.: -6.600
##  Max.   :359.59   Max.   : 79.7600   Max.   :117.90   Max.   : -3.300
```

A questo punto, si procederà con una visualizzazione grafica di questo gruppo di variabili, partendo dagli scatter plot, al fine di analizzare e individuare possibili relazioni tra le variabili.

```
pairs(c_pos, panel = panel.smooth)
```



Dall'analisi grafica non emerge alcun tipo di correlazione tra queste variabili posizionali, e soprattutto risulta evidente l'assenza di correlazioni lineari tra di esse. Tuttavia, anche in questi casi, conviene tentare comunque di costruire un modello statistico, in questo caso lineare, per effettuare un'analisi più approfondita e cercare di ottenere maggiori informazioni prima di escludere queste variabili dall'analisi finale. La variabile di risposta, che rispecchia l'obiettivo preposto, sarà sempre "Mv".

```r
qm <- lm(Mv ~ Gal.long + Gal.lat + R.GC, data = c_pos)
summary(qm)
```

```
##
## Call:
## lm(formula = Mv ~ Gal.long + Gal.lat + R.GC, data = c_pos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0237 -0.7450  0.0007  0.8703  3.5443
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.3560054  0.1825965 -40.286   <2e-16 ***
## Gal.long    -0.0012815  0.0007661  -1.673   0.0973 .
## Gal.lat     -0.0073533  0.0040091  -1.834   0.0694 .
## R.GC         0.0120365  0.0062060   1.939   0.0550 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.213 on 109 degrees of freedom
## Multiple R-squared:  0.08693,    Adjusted R-squared:  0.0618
## F-statistic: 3.459 on 3 and 109 DF,  p-value: 0.01893
```

Dai risultati ottenuti con questo modello lineare di base, il più semplice, non emerge una significatività statistica per la variabile di risposta "Mv" rispetto alle variabili posizionali. Questo risultato era comunque atteso ed è in linea con le proprietà del dataset e delle variabili stesse, poiché la magnitudine assoluta è una proprietà intrinseca del cluster globulare e non dipende dalla sua posizione spaziale. L'analisi pratica effettuata, quindi, conferma le proprietà teoriche e fisiche delle variabili presenti nel dataset.

1.7 Secondo cluster di variabili

A questo punto, si definisce un secondo gruppo di variabili del dataset, distinto dal primo, che include tutte le variabili rimanenti escludendo quelle relative alle posizioni analizzate nel primo cluster ad eccezione della nostra variabile di risposta "Mv".

```r
c_res <- clusters[,c(-1,-2,-3)]
summary(c_res)
```

```
##      Metal               Mv              r.core          r.tidal
##  Min.   :-2.400   Min.   :-10.400   Min.   : 0.100   Min.   :  6.5
##  1st Qu.:-1.800   1st Qu.: -8.300   1st Qu.: 0.500   1st Qu.: 21.9
##  Median :-1.600   Median : -7.400   Median : 0.900   Median : 32.5
##  Mean   :-1.418   Mean   : -7.431   Mean   : 1.795   Mean   : 41.5
##  3rd Qu.:-1.000   3rd Qu.: -6.600   3rd Qu.: 1.900   3rd Qu.: 51.7
##  Max.   :-0.100   Max.   : -3.300   Max.   :12.000   Max.   :284.8
##       Conc           log.t           log.rho            S0
##  Min.   :0.700   Min.   : 6.200   Min.   :0.000   Min.   : 0.700
##  1st Qu.:1.300   1st Qu.: 7.500   1st Qu.:3.100   1st Qu.: 3.900
##  Median :1.500   Median : 8.100   Median :4.000   Median : 5.600
##  Mean   :1.545   Mean   : 8.093   Mean   :3.692   Mean   : 6.228
```

```
##  3rd Qu.:1.800   3rd Qu.: 8.600   3rd Qu.:4.600   3rd Qu.: 8.200
##  Max.   :2.500   Max.   :10.100   Max.   :6.100   Max.   :19.100
##      V.esc            VHB            E.B.V            B.V
##  Min.   : 2.40   Min.   :12.90   Min.   :0.0000   Min.   :0.700
##  1st Qu.:14.90   1st Qu.:15.60   1st Qu.:0.1000   1st Qu.:0.900
##  Median :22.40   Median :16.50   Median :0.2000   Median :1.000
##  Mean   :25.07   Mean   :16.64   Mean   :0.3301   Mean   :1.162
##  3rd Qu.:33.00   3rd Qu.:17.40   3rd Qu.:0.5000   3rd Qu.:1.300
##  Max.   :78.20   Max.   :24.40   Max.   :2.9000   Max.   :4.000
##      Ellipt           CSBt
##  Min.   : 3.500   Min.   : 5.20
##  1st Qu.: 7.200   1st Qu.: 7.70
##  Median : 8.300   Median : 9.00
##  Mean   : 8.554   Mean   : 9.15
##  3rd Qu.: 9.400   3rd Qu.:10.20
##  Max.   :15.800   Max.   :15.00
```

In questo ampio gruppo di variabili rimanenti, è possibile identificare un ulteriore sottogruppo che presenta proprietà fisiche comuni, similmente al precedente cluster relativo alle posizioni. In particolare, le variabili "Metal", "E.B.V", "B.V", "Ellipt", "VHB" e "CSBt" condividono caratteristiche fisiche del corpo celeste in esame.

- **Metal** rappresenta il logaritmo della metallicità rispetto a quella solare.
- **E.B.V** misura la differenza tra il colore apparente di un oggetto celeste e il suo colore atteso.
- **B.V** è un indice di colore che riflette la differenza tra la magnitudine in banda blu e quella in banda visibile.
- **Ellipt** indica l'eccentricità della forma dell'oggetto.
- **VHB** rappresenta la magnitudine assoluta in banda B per oggetti della popolazione HB (Horizontal Branch).
- **CSBt** è una misura della classificazione spettrale.

Per ulteriori dettagli sulle variabili, si rimanda alla prima sezione del report. Come per il primo cluster, anche questo secondo gruppo di variabili sarà suddiviso in due sottogruppi: il primo riguarderà le proprietà fisiche dei corpi celesti, mentre il secondo includerà le variabili rimanenti.

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

```
##      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
```

Si procederà ora ad analizzare la correlazione tra le varie variabili, poiché alcune di esse potrebbero indicare aspetti simili. Questa analisi ha l'obiettivo di identificare variabili ad alta correlazione. Se tali variabili presentano significati sovrapponibili, si potrebbe considerare la rimozione per semplificare il modello.

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

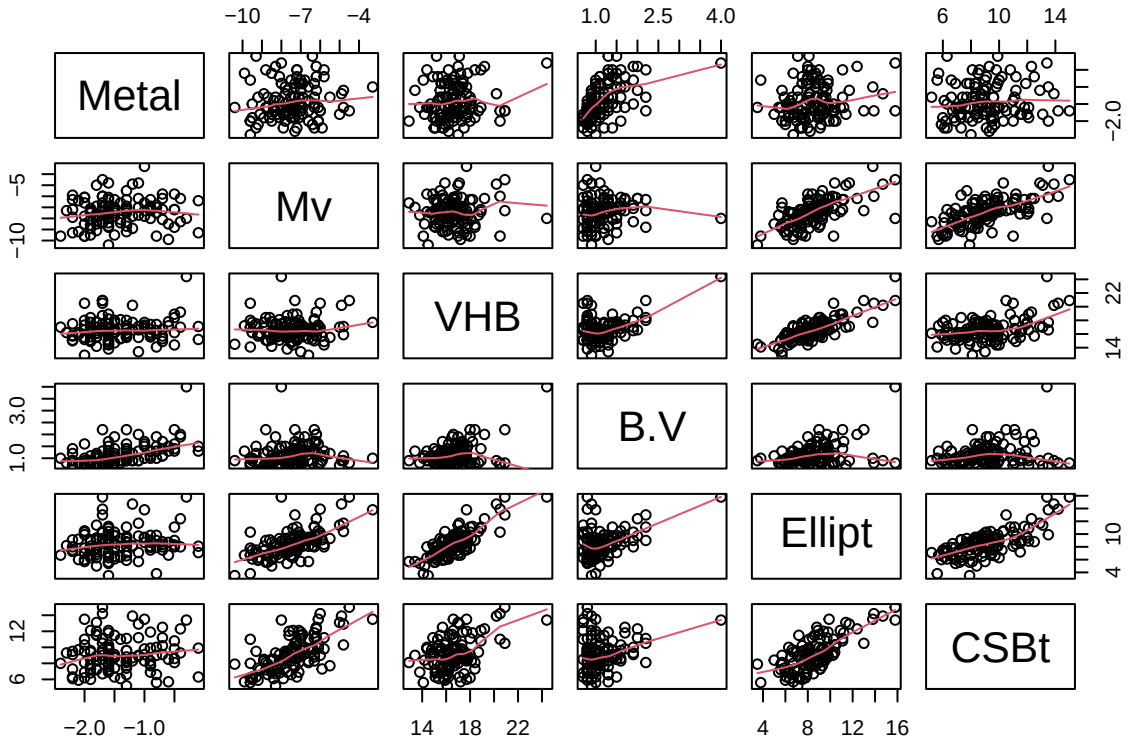
```
##             Metal       VHB     E.B.V       B.V    Ellipt       CSBt
## Metal  1.00000000 0.1205876 0.2903893 0.5431760 0.1516224 0.09212871
## VHB    0.12058757 1.0000000 0.4555454 0.4279573 0.8139601 0.44822417
## E.B.V  0.29038929 0.4555454 1.0000000 0.9565375 0.3833592 0.20830574
## B.V    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
```

Esaminando la tabella delle correlazioni, emerge chiaramente un valore particolarmente elevato: la correlazione tra le variabili "E.B.V" e "B.V", che è pari a circa 0.95, di base misurano concetti simili:

- **E.B.V** rappresenta l'eccesso di colore, ovvero la differenza tra l'indice di colore osservato dell'ammasso e il suo indice di colore intrinseco o atteso.
- **B.V** è un semplice indice di colore.

Poiché entrambe le variabili esprimono lo stesso significato, seppur da angolazioni diverse, e data la loro alta correlazione, è stato deciso di mantenere la variabile "B.V" per motivi di semplicità di interpretazione.

```
c_fis <- c_fis[-4]
pairs(c_fis, panel = panel.smooth)
```



Analizzando gli scatter plot delle variabili a coppie, emergono alcune relazioni significative tra le variabili. Pertanto, si procederà ora con l'applicazione di un semplice modello di regressione lineare su questi dati, utilizzando sempre la variabile di risposta "Mv". Questo modello aiuterà a esplorare ulteriormente le relazioni identificate e a valutare l'effetto delle variabili esplicative sulla variabile di risposta.

```
qm <- lm(Mv ~ Metal + VHB + B.V + Ellipt+ CSBt, data = c_fis)
summary(qm)
```

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

Dal semplice modello di regressione lineare emerge un'importanza statistica elevata per le variabili "VHB" e "Ellipt". Tuttavia, per l'analisi finale, si è deciso di mantenere tutte le variabili ad eccezione di "CSBt", che è risultata meno significativa rispetto alle altre. Questo approccio si basa sul fatto che la variabile di risposta "Mv" è anch'essa una proprietà fisica, e quindi le variabili rimaste potrebbero rivelarsi utili per l'analisi, contribuendo a una comprensione più completa delle dinamiche in gioco, evitando di perdere meno informazioni possibili.

```
c_fis <- c_fis[-6]
summary(c_fis)
```

```
##      Metal              Mv               VHB             B.V
##  Min.   :-2.400   Min.   :-10.400   Min.   :12.90   Min.   :0.700
##  1st Qu.:-1.800   1st Qu.: -8.300   1st Qu.:15.60   1st Qu.:0.900
##  Median :-1.600   Median : -7.400   Median :16.50   Median :1.000
##  Mean   :-1.418   Mean   : -7.431   Mean   :16.64   Mean   :1.162
##  3rd Qu.:-1.000   3rd Qu.: -6.600   3rd Qu.:17.40   3rd Qu.:1.300
##  Max.   :-0.100   Max.   : -3.300   Max.   :24.40   Max.   :4.000
##      Ellipt
##  Min.   : 3.500
##  1st Qu.: 7.200
##  Median : 8.300
##  Mean   : 8.554
```

```
##  3rd Qu.: 9.400
##  Max.   :15.800
```

1.8 Terzo cluster di variabili

A questo punto, rimane da analizzare l'ultimo set di variabili rimaste nel dataset, che include: "r.core", "r.tidal", "Conc", "log.t", "log.rho", "V.esc" e "S0". Queste variabili saranno esaminate insieme alla variabile di risposta scelta, "Mv". Questa analisi finale permetterà di valutare il contributo di ciascuna di queste variabili al modello e di identificare eventuali relazioni significative con la variabile di risposta.

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

```
##        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
```

Anlizziamo adesso l'indice di correlazione tra le variabili di questo gruppo:

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

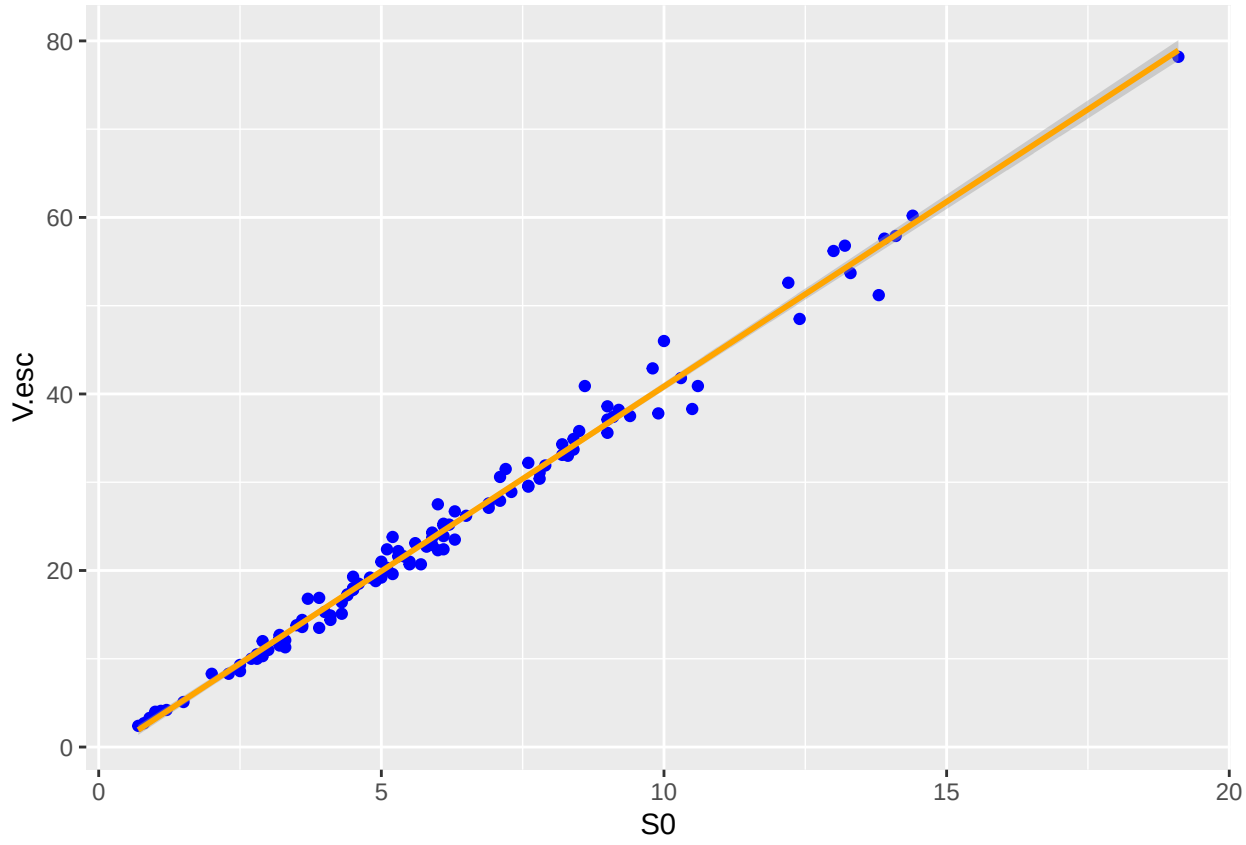
```
##             r.core     r.tidal        Conc      log.t    log.rho         S0
## r.core   1.0000000  0.57491876 -0.59924936  0.7880193 -0.8455641 -0.4455143
## r.tidal  0.5749188  1.00000000  0.02275233  0.4876358 -0.3882028 -0.0750999
## Conc    -0.5992494  0.02275233  1.00000000 -0.7632399  0.7715935  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
```

Analizzando la matrice di correlazioni, si osservano alcuni indici di correlazione molto elevati tra le variabili. In particolare, due variabili, "V.esc" e "S0", presentano un valore di correlazione estremamente alto, pari a

0.99. Pertanto, è opportuno considerare la possibilità di mantenere solo una delle due variabili per evitare ridondanze nel modello. Questa decisione permetterà di semplificare l'analisi senza compromettere la qualità delle informazioni fornite.

Dopo un'analisi approfondita delle proprietà fisiche espresse dalle variabili, è stato deciso di mantenere "S0" rispetto a "V.esc". La variabile "S0" fornisce informazioni sulla distribuzione delle velocità delle stelle all'interno dell'ammasso, risultando quindi particolarmente utile per comprendere le caratteristiche dinamiche del sistema. Di conseguenza, "V.esc" sarà esclusa dall'analisi finale.

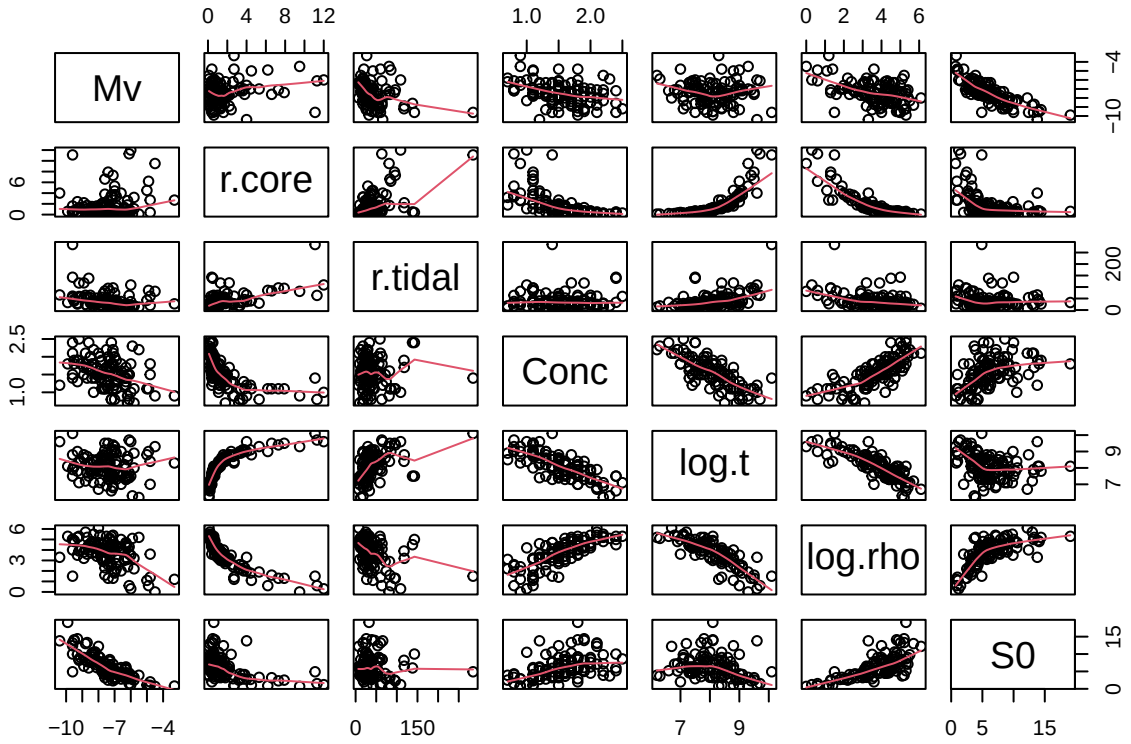
```
library(ggplot2)
ggplot(clusters, aes(x=S0,y= V.esc)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")
```




```
c_last <- c_last[-8]
```

Si procederà ora, come fatto in precedenza, ad analizzare gli scatter plot delle variabili a coppie per esaminare graficamente eventuali altre relazioni tra le variabili. Questa analisi visiva aiuterà a identificare ulteriori tendenze o correlazioni che potrebbero non essere evidenti tramite l'analisi delle sole correlazioni numeriche.

```
pairs(c_last, panel = panel.smooth)
```



Di seguito il modello di regressione lineare con le variabili rimaste in gioco:

```
qm <- lm(Mv ~ r.core + r.tidal + Conc + log.t + log.rho + S0, data = c_last)
summary(qm)
```

```
##
## Call:
## lm(formula = Mv ~ r.core + r.tidal + Conc + log.t + log.rho +
##     S0, data = c_last)
##
## 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
```

Nel contesto del modello di regressione lineare con la magnitudine come variabile di risposta, le variabili "Conc", "log.rho" e "log.t" si sono dimostrate altamente significative. Al contrario, i raggi che descrivono l'ammasso globulare non risultano significativi. Tuttavia, è stata presa la decisione di eliminare solo una di queste variabili, ovvero "r.core", poiché ha il p-value più alto nel modello. La variabile "S0", sebbene abbia mostrato una minore significatività statistica nel modello, è stata mantenuta per il valore informativo che apporta riguardo alla distribuzione delle velocità delle stelle all'interno dell'ammasso.

2 Modelli di regressione lineare

2.1 Setup

A questo punto, per evitare confusioni, si procederà a ricostruire il dataset escludendo le variabili risultate poco significative nell'analisi precedente. Le variabili escluse sono:

- **E.B.V** (a causa della sua alta correlazione con "B.V")
- **CSBt** (a causa della sua bassa significatività)
- **V.esc** (a causa dell'alta correlazione con "S0")
- **r.core** (a causa del p-value più alto nel modello di regressione utilizzato precedentemente)
- **Variabili posizionali del primo cluster** (a causa della loro bassa significatività)

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

```
##      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
```

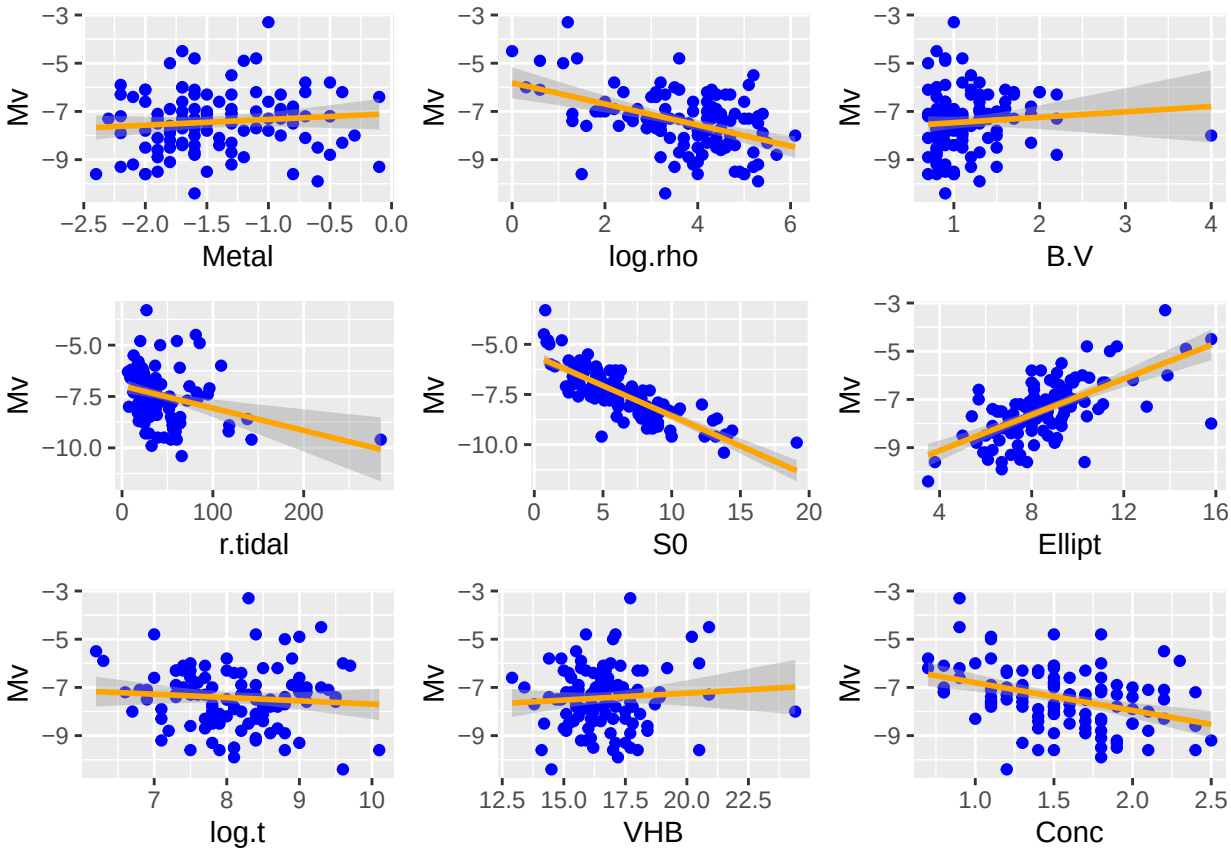
A questo punto, dopo una prima selezione delle variabili utili, il dataset è stato ridotto a 10 variabili, con "Mv" sempre come variabile di risposta.

2.2 Regressione lineare

Per identificare possibili connessioni e interazioni all'interno del dataset ridotto, è stata implementata una serie di modelli di regressione lineare semplice, uno per ciascuna delle variabili indipendenti con la variabile di risposta "Mv". Di seguito sono presentati i grafici di confronto tra "Mv" e le singole variabili, insieme al summary dei singoli modelli.

```
library(gridExtra)
p1 <- ggplot(data, aes(x=Metal,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p2 <- ggplot(data, aes(x=r.tidal,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p3 <- ggplot(data, aes(x=log.t,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p4 <- ggplot(data, aes(x=log.rho,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p5 <- ggplot(data, aes(x=S0,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p6 <- ggplot(data, aes(x=VHB,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p7 <- ggplot(data, aes(x=B.V,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p8 <- ggplot(data, aes(x=Ellipt,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

p9 <- ggplot(data, aes(x=Conc,y= Mv)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", color = "orange")

grid.arrange(
  arrangeGrob(
    p1,
    p2,
    p3,
    ncol = 1
  ),
  arrangeGrob(
    p4,
    p5,
```

```
    p6,
    ncol = 1
  ),
  arrangeGrob(
    p7,
    p8,
    p9,
    ncol = 1
  ),
  widths = c(1,1,1)
)
```




```
lm <- lm(Mv ~ Metal, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ Metal, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9257 -0.8396  0.0317  0.7178  4.0317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -7.0941     0.3416 -20.767   <2e-16 ***
## Metal         0.2377     0.2262   1.051    0.296
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.252 on 111 degrees of freedom
## Multiple R-squared:  0.009848,   Adjusted R-squared:  0.0009274
## F-statistic: 1.104 on 1 and 111 DF,  p-value: 0.2957
```

```
lm <- lm(Mv ~ log.rho, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ log.rho, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1406 -0.6468 -0.1219  0.7719  3.0405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.81540    0.32439 -17.927  < 2e-16 ***
## log.rho     -0.43758    0.08305  -5.269 6.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.125 on 111 degrees of freedom
## Multiple R-squared:    0.2,  Adjusted R-squared:  0.1928
## F-statistic: 27.76 on 1 and 111 DF,  p-value: 6.816e-07
```

```
lm <- lm(Mv ~ B.V, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ B.V, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9100 -0.8100  0.0449  0.7322  4.1675
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.6930     0.3267  -23.55   <2e-16 ***
## B.V           0.2255     0.2622    0.86    0.392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.254 on 111 degrees of freedom
## Multiple R-squared:  0.006621,   Adjusted R-squared:  -0.002328
## F-statistic: 0.7399 on 1 and 111 DF,  p-value: 0.3916
```

```r
lm <- lm(Mv ~ r.tidal, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ r.tidal, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7059 -0.7354 -0.0223  0.6483  3.9733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.979754   0.173999 -40.114  < 2e-16 ***
## r.tidal     -0.010872   0.003196  -3.402 0.000932 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 111 degrees of freedom
## Multiple R-squared:  0.09441,    Adjusted R-squared:  0.08625
## F-statistic: 11.57 on 1 and 111 DF,  p-value: 0.0009318
```

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

```
##
## Call:
## lm(formula = Mv ~ S0, data = data)
##
## 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
```

```r
lm <- lm(Mv ~ Ellipt, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt, data = data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.2606 -0.5432  0.1424  0.6196  2.1823
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.6084     0.3731 -28.437  < 2e-16 ***
## Ellipt        0.3715     0.0423   8.782 2.24e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9663 on 111 degrees of freedom
## Multiple R-squared:   0.41,  Adjusted R-squared:  0.4046
## F-statistic: 77.12 on 1 and 111 DF,  p-value: 2.241e-14
```

```
lm <- lm(Mv ~ log.t, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ log.t, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7635 -0.8272  0.0228  0.6819  4.1592
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.3274     1.2436  -5.088 1.49e-06 ***
## log.t        -0.1364     0.1530  -0.891    0.375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.253 on 111 degrees of freedom
## Multiple R-squared:  0.007108,   Adjusted R-squared:  -0.001837
## F-statistic: 0.7946 on 1 and 111 DF,  p-value: 0.3746
```

```
lm <- lm(Mv ~ VHB, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ VHB, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8464 -0.8898  0.0618  0.8217  4.0700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.38550    1.20266  -6.972 2.36e-10 ***
## VHB          0.05737    0.07194   0.798    0.427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.254 on 111 degrees of freedom
## Multiple R-squared:  0.005698,   Adjusted R-squared:  -0.00326
## F-statistic: 0.636 on 1 and 111 DF,  p-value: 0.4268
```

```
lm <- lm(Mv ~ Conc, data = data)
summary(lm)
```

```
##
## Call:
## lm(formula = Mv ~ Conc, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3669 -0.6822  0.0331  0.6859  3.3872
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.6497     0.4312 -13.104  < 2e-16 ***
## Conc         -1.1529     0.2699  -4.272 4.11e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.166 on 111 degrees of freedom
## Multiple R-squared:  0.1412, Adjusted R-squared:  0.1335
## F-statistic: 18.25 on 1 and 111 DF,  p-value: 4.107e-05
```

Dai modelli di regressione lineare semplice emerge che alcune variabili influenzano la variabile di risposta "Mv" in modo positivo e altre in modo negativo, come evidenziato dai coefficienti. In particolare, le variabili "VHB", "BV", "Metal" e "log.t" hanno mostrato una bassa significatività singolarmente. Tuttavia, prima di procedere all'eliminazione di queste variabili, è opportuno costruire un modello di regressione lineare multipla che includa tutte le variabili selezionate. Questo modello completo potrebbe rivelare interazioni tra variabili che non sono visibili nei modelli semplici con una sola variabile.

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

```
##
## Call:
## lm(formula = Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + Metal +
##     log.t + VHB + B.V, data = data)
##
## 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 **
## 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 ***
## 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
## ---
## 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
```

Per affinare il modello di regressione lineare multipla,si procede eliminando le variabili meno significative in modo graduale, procedendo per step anziché eliminarle tutte in blocco. Questo metodo permette di valutare l'impatto di ogni variabile sull'accuratezza del modello e garantisce che nessuna variabile significativa venga rimossa prematuramente.

Procediamo quindi con eliminare le variabili "Metal" e B.V le quali risultano le meno significative.

```
lm_clear1 <-lm(Mv ~ log.t + VHB + r.tidal + Conc + log.rho + S0 + Ellipt, data=data)
summary(lm_clear1)
```

```
##
## Call:
## lm(formula = Mv ~ log.t + VHB + r.tidal + Conc + log.rho + S0 +
##     Ellipt, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08653 -0.04675  0.01977  0.04039  0.06719
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1697951  0.7631117   2.843  0.00537 **
## log.t       -0.1915311  0.0924683  -2.071  0.04078 *
## VHB         -0.9191409  0.0324855 -28.294  < 2e-16 ***
## r.tidal     -0.0004823  0.0002257  -2.137  0.03495 *
## Conc        -0.1006469  0.0640553  -1.571  0.11913
## log.rho     -0.1010265  0.0482843  -2.092  0.03882 *
## S0          -0.0040245  0.0034050  -1.182  0.23990
## Ellipt       0.9136772  0.0322220  28.356  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0462 on 105 degrees of freedom
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9986
## F-statistic: 1.174e+04 on 7 and 105 DF,  p-value: < 2.2e-16
```

A questo punto, si decide di eliminare anche le variabili "Conc" e "S0". Anche in questo modello, tali variabili risultano essere poco significative, nonostante la rimozione di altre due variabili. Tale insignificanza era già evidente nel modello completo, confermando quindi la correttezza della scelta di escluderle.

```r
lm_clear2 <-lm(Mv ~ log.t + VHB + r.tidal + log.rho + Ellipt, data=data)
summary(lm_clear2)
```

```
##
## Call:
## lm(formula = Mv ~ log.t + VHB + r.tidal + log.rho + Ellipt, data = data)
##
## 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
```

Questo modello di regressione può essere considerato adeguato, poiché tutte le variabili incluse mostrano una certa rilevanza e risultano significative. Il modello in questione è il seguente:

$$\text{Mv} \sim \text{log.t} + \text{VHB} + \text{r.tidal} + \text{log.rho} + \text{Ellipt}$$

Tuttavia, tre variabili, "log.t", "log.rho" e "r.tidal", presentano una significatività inferiore rispetto alle altre due rimaste. Pertanto, potrebbe essere conveniente continuare con il metodo di eliminazione graduale delle variabili meno significative. Per questioni di spazio, i vari summary dei modelli non verranno riportati.

```r
# Esclusione di log.rho
lm_clear3 <-lm(Mv ~ log.t + VHB + r.tidal + Ellipt, data=data)
# Esclusione di log.t (p-value 0.553 su lm_clear3)
lm_clear4 <-lm(Mv ~ VHB + r.tidal + Ellipt, data=data)
# Esclusione di r.tidal (p-value 0.0844 su lm_clear4)
lm_clear5 <-lm(Mv ~ VHB + Ellipt, data=data)
summary(lm_clear5)
```

```
##
## Call:
## lm(formula = Mv ~ VHB + Ellipt, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08596 -0.05064  0.02072  0.04438  0.06578
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

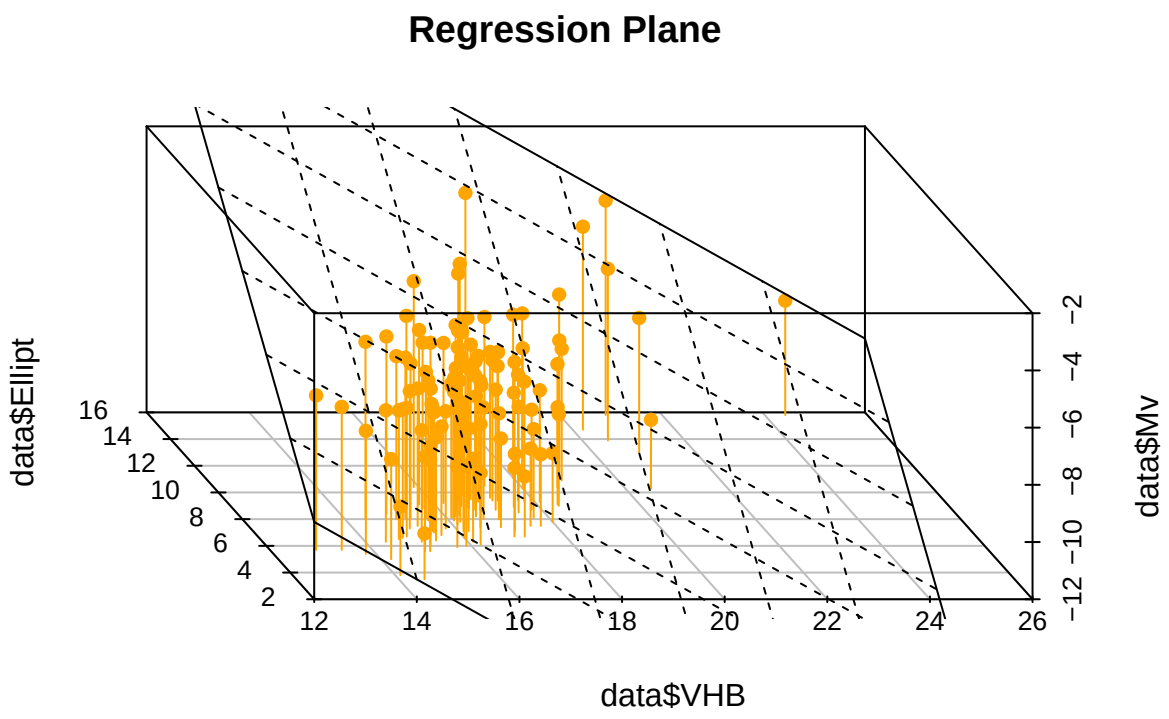
```
## (Intercept)  0.75787    0.05667   13.38   <2e-16 ***
## VHB         -1.00378    0.00473 -212.21   <2e-16 ***
## Ellipt       0.99509    0.00361  275.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04791 on 110 degrees of 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
```

A questo punto, abbiamo ottenuto un modello di regressione molto semplice, che utilizza soltanto due variabili. Questo modello è stato sviluppato tramite un approccio "backwards", e entrambe le variabili risultano altamente significative. Inoltre, il modello presenta un valore di Adjusted R-squared molto elevato, indicativo di una buona qualità di adattamento. Il modello finale ottenuto è il seguente:

Mv ~ VHB + Ellipt

Ho quindi deciso di rimuovere le variabili che non sono risultate significative nel confronto diretto. La variabile "BV" è stata esclusa in quanto è risultata la meno significativa tra tutte. La variabile "log.t", era stata considerata per l'eliminazione a priori, poiché intuitivamente non sembrava influenzare direttamente la magnitudine assoluta. Anche da questa ultima analisi, la variabile risulta poco significativa ed è quindi stata rimossa. Successivamente viene mostrato un plot del piano di regressione del modello:

```
library(scatterplot3d)
plot3d <- scatterplot3d(data$VHB,data$Ellipt,data$Mv, angle=120, scale.y=0.7, pch=16,
                        color ="orange", main ="Regression Plane" ,type = "h")
plot3d$plane3d(lm_clear5, lty.box = "solid")
```



2.3 Selezione del modello con metodi iterativi (AIC e BIC)

In questo paragrafo si cercherà di stimare il modello statistico che meglio descrive la variabile target che si vuole studiare, "Mv". La selezione del modello verrà effettuata utilizzando le procedure AIC e BIC nelle direzioni "forward", "backward" e "both".

2.3.1 Procedura AIC

Di seguito vengono riportati i risultati ottenuti con la procedura AIC prima nella direzione "forward", nella direzione "backward" e successivamente "both".

```
qm_complete <-lm(Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + Metal + VHB, data=data)
qm_1 <-lm(Mv ~ 1, data=data)

# AIC FORWARD
aic_for = step(qm_1, scope=formula(qm_complete), 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
## + Metal    1     1.730 173.912  52.721
## + 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
## + Metal    1    5.3647 53.101  -79.337
## + 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
## + Metal    1    0.7649 33.702 -128.71
## <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
## + Metal    1    0.0038 22.4593 -172.57
##
## 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
## + Metal    1 0.0018565 0.23626 -685.23
##
## Step:  AIC=-686.55
## Mv ~ S0 + r.tidal + Ellipt + VHB + Conc
##
##           Df  Sum of Sq     RSS     AIC
## <none>                  0.23353 -686.55
## + Metal    1 0.00108774 0.23244 -685.07
## + log.rho  1 0.00021603 0.23331 -684.65
```

```
summary(aic_for)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB + Conc, data = data)
##
## 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
# AIC BACKWARD
aic_back <- step(qm_complete, scope=formula(qm_1), direction ="backward", k=2)
```

```
## Start:  AIC=-683.16
## Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + Metal + VHB
##
##           Df Sum of Sq     RSS     AIC
## - log.rho  1    0.0002  0.2324 -685.07
## - Metal    1    0.0010  0.2333 -684.65
## - Conc     1    0.0020  0.2343 -684.19
## <none>                  0.2323 -683.16
## - S0       1    0.0069  0.2392 -681.84
## - r.tidal  1    0.0120  0.2443 -679.46
## - VHB      1   21.9719 22.2042 -169.86
## - Ellipt   1   29.7047 29.9370 -136.10
##
## Step:  AIC=-685.07
## Mv ~ r.tidal + Conc + S0 + Ellipt + Metal + VHB
##
##           Df Sum of Sq     RSS     AIC
## - Metal    1    0.0011  0.2335 -686.55
## - Conc     1    0.0038  0.2363 -685.23
## <none>                  0.2324 -685.07
## - S0       1    0.0092  0.2416 -682.69
## - r.tidal  1    0.0166  0.2491 -679.27
## - VHB      1   22.0512 22.2836 -171.46
## - Ellipt   1   29.9004 30.1328 -137.36
##
## 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
```

```r
summary(aic_back)
```

```
##
## Call:
## lm(formula = Mv ~ r.tidal + Conc + S0 + Ellipt + VHB, data = data)
##
## 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
```

```r
# AIC BOTH
aic_both <- step(qm_1, scope=formula(qm_complete), 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
## + Metal    1     1.730 173.912  52.721
## + 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
## + Metal    1     5.365  53.101  -79.337
## + 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
## + Metal    1     0.765  33.702 -128.712
## <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
## + Metal    1     0.004 22.459 -172.57
## - 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
## + Metal    1    0.0019  0.2363 -685.23
## - 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
## + Metal    1    0.0011  0.2324 -685.07
## + 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(aic_both)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB + Conc, data = data)
##
## 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
```

Modelli ottenuti:

- AIC direzione "forward" : Mv ~ S0 + r.tidal + Ellipt + VHB + Conc
- AIC direzione "backward" : Mv ~ S0 + r.tidal + Ellipt + VHB + Conc
- AIC direzione "both": Mv ~ S0 + r.tidal + Ellipt + VHB + Conc

Con tutte le varianti del metodo AIC si ottiene lo stesso modello per la variabile di risposta "Mv".

2.3.2 Procedura BIC

Di seguito vengono riportati i risultati ottenuti con la procedura BIC prima nella direzione "forward", nella direzione "backward" e successivamente "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.

```r
# BIC FORWARD
bic_for = step(qm_1, scope=formula(qm_complete), 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
## + Metal    1     1.730 173.912  58.176
## + 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
## + Metal    1    5.3647 53.101  -71.155
## + 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
## + Metal    1    0.7649 33.702 -117.80
## + 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
## + Metal    1    0.0038 22.4593 -158.94
##
## 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
## + Metal    1 0.0018565 0.23626 -668.87
```

```
summary(bic_for)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB, data = data)
##
## 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
```

```r
# BIC BACKWARD
bic_back <- step(qm_complete, scope=formula(qm_1), direction ="backward", k=log(113))
```

```
## Start:  AIC=-661.34
## Mv ~ r.tidal + Conc + log.rho + S0 + Ellipt + Metal + VHB
##
##           Df Sum of Sq     RSS     AIC
## - log.rho  1    0.0002  0.2324 -665.98
## - Metal    1    0.0010  0.2333 -665.56
## - Conc     1    0.0020  0.2343 -665.10
## - S0       1    0.0069  0.2392 -662.75
## <none>                  0.2323 -661.34
## - r.tidal  1    0.0120  0.2443 -660.37
## - VHB      1   21.9719 22.2042 -150.77
## - Ellipt   1   29.7047 29.9370 -117.01
##
## Step:  AIC=-665.98
## Mv ~ r.tidal + Conc + S0 + Ellipt + Metal + VHB
##
##           Df Sum of Sq     RSS     AIC
## - Metal    1    0.0011  0.2335 -670.18
## - Conc     1    0.0038  0.2363 -668.87
## - S0       1    0.0092  0.2416 -666.33
## <none>                  0.2324 -665.98
## - r.tidal  1    0.0166  0.2491 -662.90
## - VHB      1   22.0512 22.2836 -155.10
## - Ellipt   1   29.9004 30.1328 -121.00
##
## 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(bic_back)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = data)
##
## 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
```

```
#BIC BOTH
bic_both <- step(qm_1, scope=formula(qm_complete), 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
## + Metal    1     1.730 173.912  58.176
## + 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
## + Metal    1     5.365  53.101  -71.155
## + 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
## + Metal    1     0.765  33.702 -117.803
## + 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
## + Metal    1     0.004 22.459 -158.94
## - 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
## + Metal    1    0.0019  0.2363 -668.87
## - 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
## + Metal    1     0.003   0.243 -670.47
## + 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
## + Metal    1     0.000   0.252 -671.01
## + S0       1     0.000   0.252 -670.91
## - VHB      1   103.382 103.635   -0.32
## - Ellipt   1   174.388 174.641   58.65
```

```
summary(bic_both)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = data)
##
## 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
```

Modelli ottenuti:

- BIC direzione "forward" : Mv ~ S0 + r.tidal + Ellipt + VHB
- BIC direzione "backward" :Mv ~ Ellipt + VHB
- BIC direzione "both" : Mv ~ Ellipt + VHB

2.4 Contronto tra i modelli ottenuti mediante le diverse tecniche utilizzate

Il modello AIC in tutte e tre le modalità viste, forward, backward e both, seleziona lo stesso modello:

- Mv ~ S0 + r.tidal + Ellipt + VHB + Conc

Il modello è descritto da:

```
summary(aic_both)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB + Conc, data = data)
##
## 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
```

In questo modello, la variabile "Conc" è risultata non significativa (p-value = 0.15016). Inoltre, le variabili "S0" e "r.tidal" hanno mostrato p-value relativamente elevati, rispettivamente 0.15016 e 0.00868, pur essendo statisticamente significativi. Pertanto, ho deciso di escludere questo modello nella scelta finale, sia a causa della scarsa significatività di alcune variabili, sia perché la presenza di molteplici variabili che rendono il modello di difficile rappresentazione.

Il modello BIC invece seleziona due modelli diversi:

- BIC direzione "forward" : Mv ~ S0 + r.tidal + Ellipt + VHB

```
summary(bic_for)
```

```
##
## Call:
## lm(formula = Mv ~ S0 + r.tidal + Ellipt + VHB, data = data)
##
## 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
```

La variabile "S0" non risulta significativa (p-value 0.0667) e "r.tidal" ha un p-value di 0.0127, che seppur significativo non è troppo basso come valore. Anche questo modello quindi è stato scartato.

- BIC direzione "backward" e "both" :Mv ~ Ellipt + VHB

```
summary(bic_both)
```

```
##
## Call:
## lm(formula = Mv ~ Ellipt + VHB, data = data)
##
## 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
```

Questo modello corrisponde a quello ottenuto in precedenza tramite la rimozione manuale delle variabili meno significative, fino a lasciare solo quelle più rilevanti. Entrambe le variabili rimanenti risultano essere molto significative a livello statistico per la variabile di risposta "Mv". Inoltre, questo modello, oltre ad avere un altissimo valore di Adjusted R-squared pari a 0.9985, si distingue per la sua semplicità e facilità di interpretazione, poiché è composto da sole due variabili.

Il modello quindi che risulta essere più semplice e di facile intrepretazione, oltre ad avere tutte le variabili altamente significative è quello selezionato minimizzando "BIC" (direzione "backward" e "both").

In generale il fatto che minimizzando "BIC" si ottengano dei modelli di regressione lineare più semplici, parsimoniosi è un risultato atteso ed era il mio obbiettivo principale, ho voluto comunque testare la metodologia AIC per vedere se selezionava comunque qualche modello interesante ma così non è stato.

3 Metodi grafici

In questo paragrafo verranno costruiti i grafi (non direzionato e DAG) per individuare le relazioni causali tra le variabili del dataset. Per questa analisi, verranno utilizzate tutte le variabili considerate nella sezione sulla regressione, comprese quelle eliminate nelle fasi iniziali delle regressioni semplici. Le variabili incluse sono: "Mv", "r.tidal", "Conc", "log.rho", "S0", "Ellipt", "Metal", "log.t", "VHB" e "B.V".

3.1 Matrice di concentrazione e matrice delle correlazioni parziali

Per la costruzione di questi grafi, è importante notare che tutte le variabili sono continue. Di conseguenza, sarà necessario calcolare la matrice di concentrazione e, successivamente, la matrice delle correlazioni parziali.

```
library(gRbase)

# Matrice di covarianza
m_cov <- cov.wt(data, method="ML")$cov
# Matrice di concentrazione
m_conc <- solve(m_cov)
# Correlazioni parziali
m_pc <-cov2pcor(m_conc)
m_pc
```

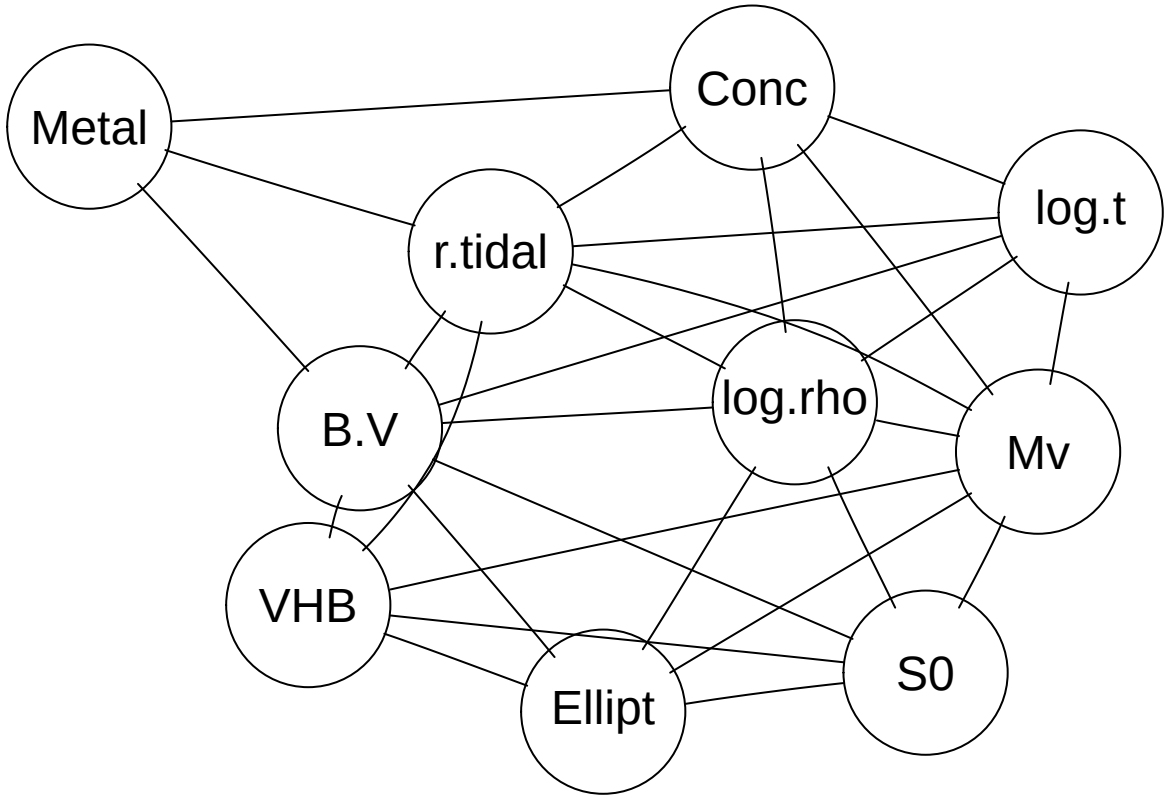
```
##               Metal          Mv     r.tidal        Conc       log.t     log.rho
## Metal     1.00000000 -0.09923533  0.30739218  0.10917101  0.09426341 -0.09329823
## Mv       -0.09923533  1.00000000  0.30725942  0.37576304  0.08430665  0.44726920
## r.tidal   0.30739218  0.30725942  1.00000000 -0.02275233 -0.48763579  0.38820283
## Conc      0.10917101  0.37576304 -0.02275233  1.00000000  0.76323992 -0.77159351
## log.t     0.09426341  0.08430665 -0.48763579  0.76323992  1.00000000  0.82939253
## log.rho  -0.09329823  0.44726920  0.38820283 -0.77159351  0.82939253  1.00000000
## S0       -0.09157562  0.81678113  0.07509990 -0.45550300  0.24808238 -0.69599562
## VHB      -0.12058757 -0.07548179 -0.25868316  0.05169877 -0.10107372  0.15810657
## B.V      -0.54317601 -0.08137091  0.44997833 -0.05006081  0.38055686 -0.38062582
## Ellipt   -0.15162244 -0.64028370 -0.02335289  0.26089578 -0.03148477  0.38464750
##                  S0         VHB         B.V      Ellipt
## Metal    -0.09157562 -0.12058757 -0.54317601 -0.15162244
## Mv        0.81678113 -0.07548179 -0.08137091 -0.64028370
## r.tidal   0.07509990 -0.25868316  0.44997833 -0.02335289
## Conc     -0.45550300  0.05169877 -0.05006081  0.26089578
## log.t     0.24808238 -0.10107372  0.38055686 -0.03148477
## log.rho  -0.69599562  0.15810657 -0.38062582  0.38464750
## S0        1.00000000  0.03193854 -0.21096025  0.50038188
## VHB       0.03193854  1.00000000 -0.42795733 -0.81396014
## B.V      -0.21096025 -0.42795733  1.00000000 -0.37525328
## Ellipt    0.50038188 -0.81396014 -0.37525328  1.00000000
```

È importante notare che alcune variabili presentano tra di loro un coefficiente di correlazione piuttosto elevato. Ad esempio, la variabile "S0" ha una correlazione di 0.816 con "Mv", indicando una forte relazione lineare positiva tra queste due variabili. Analogamente, "log.rho" e "log.t" hanno una correlazione di 0.829, suggerendo una forte relazione positiva tra di loro. Al contrario, "VHB" e "Ellipt" mostrano una correlazione di -0.813, indicando una forte relazione lineare negativa. Questi alti valori di correlazione devono essere tenuti in considerazione durante l'analisi causale, poiché potrebbero influenzare la costruzione e l'interpretazione dei grafi.

3.2 Grafo non direzionato

Per la rappresentazione del grafo non direzionato si utilizza la procedura GLASSO (Graphical Lasso). Questa tecnica mira a ridurre l'overfitting e gli effetti indesiderati dovuti a una eventuale collinearità tra le variabili altamente correlate. Inoltre, GLASSO permette di ottenere un grafo sparso, semplificando così la trattazione riducendo le connessioni tra le variabili.

```
df_data = data.frame(data)
library(gRain)
library(gRim)
library(glasso)
library(igraph)
c <- cov2cor(m_cov)
res_lasso <- glasso(c, rho=0.1)
AM <- res_lasso$wi != 0
diag(AM) <- F
colnames(AM) <- names(df_data)
rownames(AM) <- names(df_data)
gl <- as_graphnel(as(AM, "igraph"))
plot(gl, "neato")
```




```
graph::degree(gl)
```

```
##   Metal      Mv r.tidal    Conc   log.t log.rho      S0     VHB     B.V  Ellipt
##       3       7       7       5       5       7       5       5       7       5
```

La lettura del grafo è utile per individuare, in maniera preliminare, le connessioni causali tra le variabili. Il grafo è stato ottenuto impostando il parametro di regolarizzazione a 0.1; di conseguenza, il grafo presenta molte connessioni tra le variabili. Concentrandosi sulla variabile "Mv", l'obiettivo di questo studio, essa risulta connessa con altre sette variabili: "S0", "Ellipt", "log.rho", "Conc", "log.t", "r.tidal" e "VHB". Tutti questi collegamenti sono sicuramente corretti da un punto di vista fisico almeno a primo impatto in quanto:

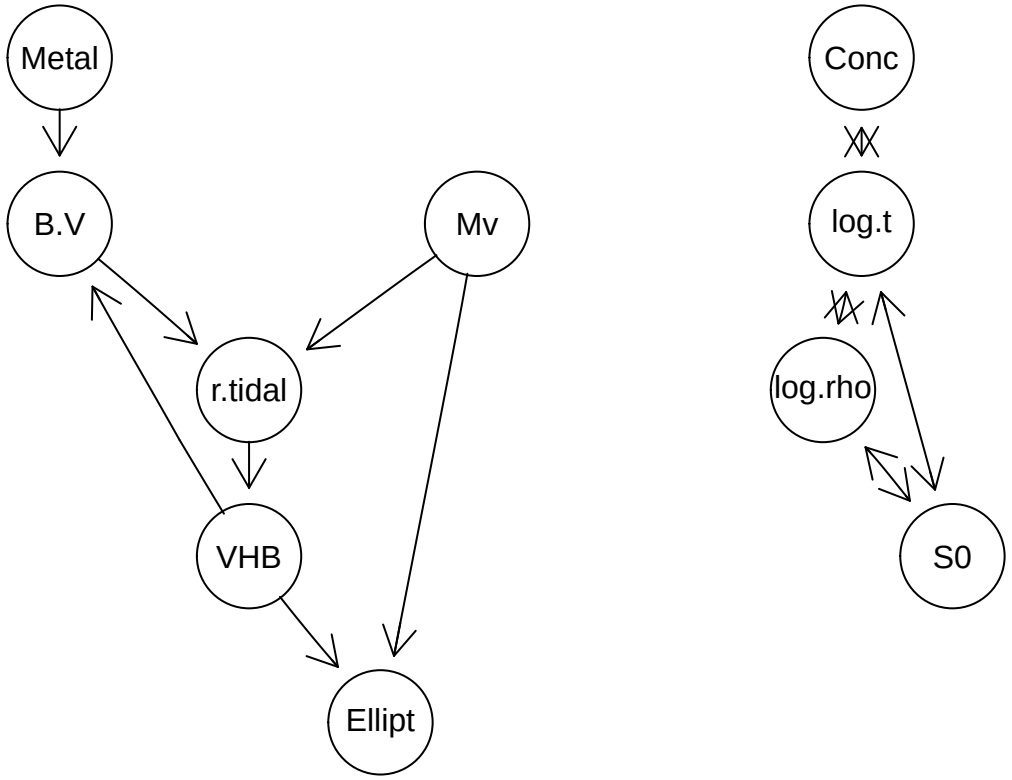
- **r.tidal** (Raggio di marea): Indica il raggio entro il quale le forze di marea esterne diventano dominanti. Un raggio di marea maggiore suggerisce un cluster più grande o più massiccio, il che può aumentare la luminosità totale del cluster, influenzando quindi la magnitudine assoluta (Mv).
- **Conc** (Parametro di concentrazione del nucleo): Misura quanto è concentrato il nucleo del cluster. Un nucleo più concentrato implica una maggiore densità stellare centrale, che può aumentare la luminosità osservata del cluster.
- **log.t** (Logaritmo del tempo di rilassamento centrale): Rappresenta il tempo necessario affinché il cluster raggiunga un equilibrio dinamico. Un tempo di rilassamento più lungo indica un cluster più massiccio e meno evoluto dinamicamente, il che può in alcuni casi influenzare positivamente la luminosità totale del cluster.
- **log.rho** (Logaritmo della densità centrale): Indica la densità stellare centrale. Una maggiore densità centrale implica più stelle per unità di volume, può aumentare così la luminosità del cluster ma non è detto.
- **S0** (Velocità di dispersione centrale): Rappresenta la velocità media con cui le stelle si muovono nel cluster. Una velocità di dispersione maggiore può essere correlata a un cluster più dinamico e massiccio, contribuendo a una maggiore luminosità totale.
- **VHB** (Livello del ramo orizzontale): Indica la luminosità delle stelle in una fase evolutiva specifica. Un livello del ramo orizzontale più alto significa stelle più luminose in quella fase, aumentando la luminosità complessiva del cluster.
- **Ellipt** (Ellitticità): Misura la forma del cluster. Un cluster più ellittico può influenzare la distribuzione della luce, alterando la percezione della luminosità e quindi la magnitudine assoluta.

In sintesi, queste variabili influenzano la magnitudine assoluta (Mv) del cluster poiché determinano la distribuzione, la densità e la luminosità delle stelle all'interno del cluster.

3.3 DAG

Per la costruzione del grafo direzionato verrà utilizzato l'algoritmo PC (Peter-Clark). Questo metodo parte dalla ricerca delle correlazioni tra le variabili e, attraverso test di indipendenza, consente di disegnare un grafo direzionato che rappresenta le relazioni causali tra le variabili.

```
library(ggm)
library(pcalg)
suffStat <- list(C=c, n=nrow(df_data))
indepTest <- gaussCItest
cpdag_dyn <- pc(suffStat, indepTest, p=ncol(df_data), alpha=0.1)
nodes(cpdag_dyn@graph) <- names(df_data)
plot(cpdag_dyn@graph)
```



Si noti che il grafo ottenuto non è propriamente un DAG (Directed Acyclic Graph), poiché sono presenti alcuni cicli tra le variabili. Per eliminare questi cicli, si modificano le relazioni ottenute. Questo è possibile poiché l'algoritmo PC, per sua natura, può non essere in grado di distinguere completamente tra tutte le possibili strutture causali, e potrebbero esserci casi in cui l'orientamento degli archi non è definito con certezza assoluta. Si utilizzano quindi le conoscenze a priori sull'argomento per selezionare il grafo che meglio rappresenta le relazioni tra le variabili del dataset. Alcune connessioni potrebbero non riflettere correttamente le relazioni causali fisiche, pertanto si apportano le seguenti modifiche:

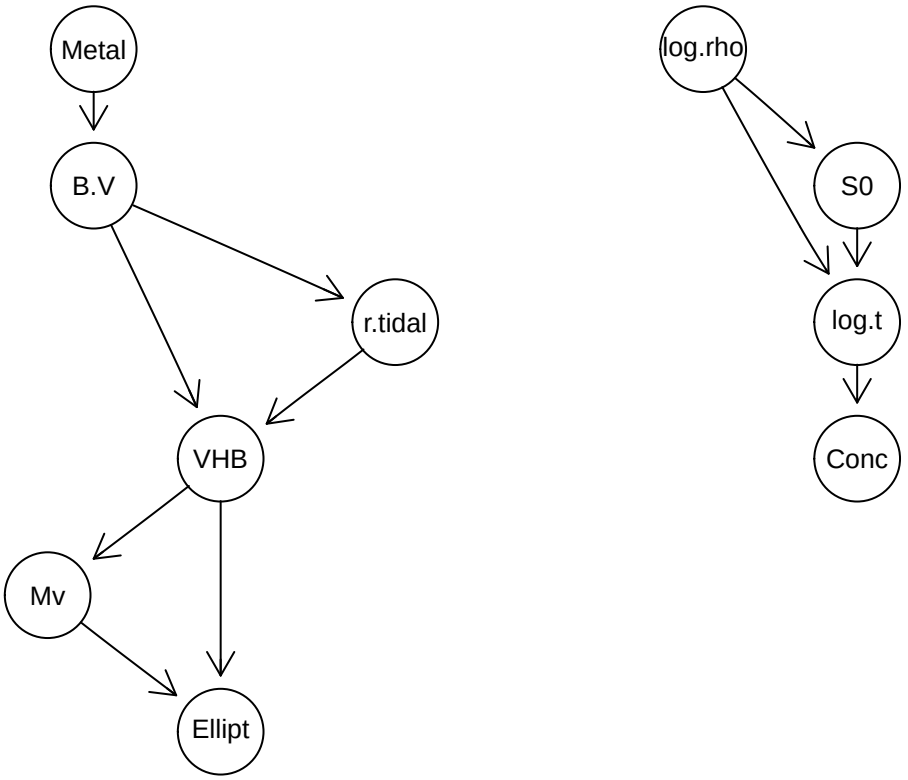
- Modifica: La connessione VHB -> B.V potrebbe non essere intuitiva. Sebbene la luminosità del ramo orizzontale (VHB) possa influenzare l'indice di colore (B.V), è più comune che l'indice di colore (B.V) influenzi la determinazione delle stelle nel ramo orizzontale. Pertanto, si inverte la direzione a B.V -> VHB.
- Aggiunta: Il livello del ramo orizzontale (VHB) è una misura diretta della luminosità delle stelle in una fase evolutiva specifica. Un valore più basso di VHB (indicando maggiore luminosità) contribuisce a una maggiore luminosità complessiva dell'ammasso. Per questo motivo, si aggiunge la connessione VHB -> Mv.
- Eliminazione: La connessione Mv -> r.tidal è stata rimossa, questo perchè da un punto di vista astrofico non c'è infuenza diretta tra le due variabili in quanto sono proprietà diverse dell'ammasso.Tuttavia, come conferma anche il grafo, esistono invece delle dipendenze indirette tramite VHB come che sono giustificabili anche da un punto di vista fisico.

Anche una connessione del tipo Ellipt -> Mv sarebbe fisicamente plausibile, tuttavia difficile da interpretare e spiegare, per evitare di alterare eccessivamente il grafo originale, ho preferito non modificarla.

La parte destra del grafo, che include le variabili log.rho, log.t, S0 e Conc, non è stata trattata con la stessa attenzione della parte sinistra, poiché queste variabili risultano scollegate dalla variabile Mv e, sotto un

profilo fisico, anche parzialmente non rilevanti per l'obiettivo di questo progetto oltre ad essere di difficile interpetazione. Sono stati comunque sistemati i vari cicli in modo da rederlo un vero e proprio DAG cercando un iterpretazione fisica adeguata.

```
correct_dag <- DAG(B.V~Metal, VHB~B.V:r.tidal, r.tidal~B.V, Mv~VHB, Ellipt~VHB:Mv, Conc~log.t, log.t~log
plot(as(correct_dag, "graphNEL"))
```



Da notare da questo risultato finale che con l'algoritmo PC si escludono alcune relazioni individuate precedentemente dal grafo non direzionato. Inoltre questa divisione tra questi due blocchi di variabili rispetto ad "Mv" è qualcosa che ci si aspettava avendo già visto i risultati ottenuti con i modelli di regressione lineare.

3.4 Risultati ottenuti

Utilizzando il grafo non direzionato ottenuto tramite la procedura GLASSO, abbiamo inizialmente identificato delle relazioni tra la variabile Mv e le altre variabili rimanenti. Alcune di queste relazioni sono state confermate nella costruzione del grafo direzionato (DAG) utilizzando l'algoritmo PC, mentre altre sono state escluse. Il grafo non direzionato aveva suggerito più connessioni, ma molte di esse non erano chiare né particolarmente dirette dal punto di vista fisico.

Analizzando il DAG ottenuto, si osserva che Mv sembra dipendere direttamente solo dalla variabile VHB. Quando si considera VHB, Mv mostra una dipendenza condizionata anche da B.V e r.tidal. Inoltre, se ci si condiziona anche su B.V, Mv appare influenzata anche da Metal, sebbene questa relazione sia più distante e meno diretta.

In sintesi, il grafo direzionato chiarisce che Mv è principalmente influenzata da VHB e, condizionatamente, anche da B.V e r.tidal. L'inclusione di B.V nella condizione aggiunge una dipendenza più remota da Metal, suggerendo una struttura causale complessa e multi-livello per la variabile Mv.

4 Conclusioni

In queste pagine sono stati discussi i modelli ottenuti attraverso tecniche di regressione lineare per descrivere la magnitudine assoluta di un ammasso globulare. Sono stati esplorati diversi modelli, sia tramite analisi manuale che mediante metodi iterativi. Tra i vari modelli analizzati, il modello che è risultato migliore da un punto di vista di efficacia, semplicità e rilevanza delle variabili coinvolte è stato:

Mv ~ Ellipt + VHB

Lo studio ha affrontato diverse difficoltà, in particolare nelle fasi iniziali del progetto, a causa della complessità dell'argomento e dell'elevato numero di variabili nel dataset, la maggior parte delle quali di natura fisica, con alcune eccezioni. Nonostante queste sfide, i risultati ottenuti sono stati soddisfacenti e coerenti con l'obiettivo prefissato.

Successivamente, è stata effettuata un'analisi grafica per individuare le relazioni causali tra le variabili del dataset. La creazione del grafo non direzionato ha permesso di identificare preliminarmente le connessioni causali tra le variabili, con un focus particolare sulla magnitudine assoluta. Successivamente, è stato costruito il grafo direzionato (DAG) utilizzando l'algoritmo PC. Tuttavia, l'algoritmo non è riuscito a catturare correttamente tutte le relazioni causali e ha richiesto alcune modifiche.

Questa analisi ha confermato i risultati ottenuti inizialmente tramite i modelli di regressione. In particolare, è stata confermata una forte influenza della variabile VHB sulla magnitudine assoluta (Mv). Inoltre, le altre variabili che influenzano Mv sono state riscontrate nei modelli lineari, anche se con un impatto minore rispetto a VHB.

Per quanto riguarda la variabile Ellipt, l'analisi grafica suggerisce che sia Mv a influenzarla, piuttosto che il contrario. Tuttavia, dal punto di vista fisico, questa relazione è complessa e le due variabili dipendono parzialmente l'una dall'altra. In pratica, Ellipt e Mv sono correlate in modo bidirezionale, con entrambe le variabili che possono influenzarsi reciprocamente.

Nel contesto della fisica degli ammassi globulari, l'ellitticità (Ellipt) del cluster e la magnitudine assoluta (Mv) possono essere legate attraverso meccanismi indiretti. Ad esempio, la forma ellittica di un cluster potrebbe riflettere aspetti della sua evoluzione e distribuzione stellare, che a loro volta influenzano la sua luminosità complessiva. Tuttavia, la magnitudine assoluta può anche fornire informazioni sullo stato evolutivo del cluster che potrebbe influenzare la sua ellitticità osservata. Quindi, mentre il grafo suggerisce una direzione causale specifica, la realtà fisica implica una relazione più complessa e bidirezionale tra Ellipt e Mv.

Tutti i risultati ottenuti, sebbene utili, rappresentano solo il punto di partenza per ulteriori ricerche e studi sperimentali. Questi risultati offrono ottime basi per approfondire la comprensione della struttura e delle dinamiche degli ammassi globulari.

Le connessioni e le influenze identificate tra le variabili possono servire come guida per indagini più dettagliate, che potrebbero includere studi empirici più approfonditi e modelli teorici avanzati. Questi ulteriori studi potrebbero contribuire a chiarire le relazioni causali tra le variabili, esplorare meccanismi fisici sottostanti e migliorare la nostra comprensione della formazione e dell'evoluzione degli ammassi globulari. In definitiva, le scoperte fatte forniscono un'importante base di partenza per ampliare la nostra conoscenza in questo campo affascinante e complesso.