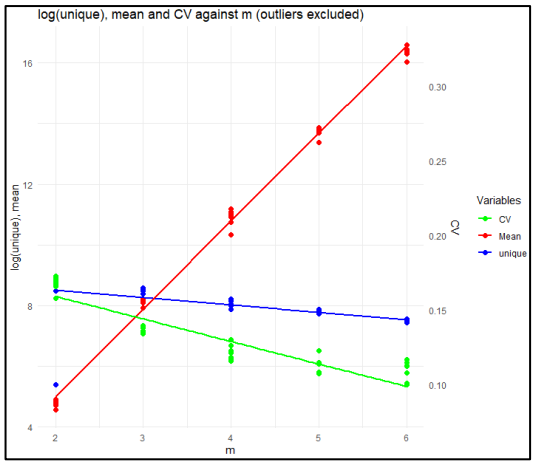
12/03 - Altri tentativi

Il grafico che si potrebbe inserire (previo modifiche per migliorarne l'estetica) per riassumere l'azione di m è questo qua sotto con tre linee colorate. Che ne dici? Ho dovuto utilizzare log(unique) e scalare CV altrimenti non avremmo visto niente dato che unique è nell'ordine delle migliaia.



Ho cercato anche di cambiare un po' le regressioni per provare a trovare un collegamento tra SNPs/unique con M ed n. Un risvolto interessante è che, utilizzando log(SNPs) e log(unique), M ed n migliorano leggermente i modelli.

SNPs

- Vecchio: **lm(SNPs ~ log(m)**, data = df). Su 30 repeats di 10-fold cross validation, R-squared: 0.74 ± 0.29

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    22869       3029   7.549 1.81e-09 ***
log(m)        -12708       2249  -5.650 1.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6515 on 44 degrees of freedom
Multiple R-squared:  0.4204,    Adjusted R-squared:  0.4073
F-statistic: 31.92 on 1 and 44 DF,  p-value: 1.102e-06
```

- Prova 1: **lm(log(SNPs) ~ m**, data = df). Su 30 repeats di 10-fold cross validation, R-squared:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.54698    0.29653  32.196  < 2e-16 ***
m           -0.27356    0.07151  -3.826 0.000408 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.695 on 44 degrees of freedom
Multiple R-squared:  0.2496,    Adjusted R-squared:  0.2325
F-statistic: 14.63 on 1 and 44 DF,  p-value: 0.0004081
```

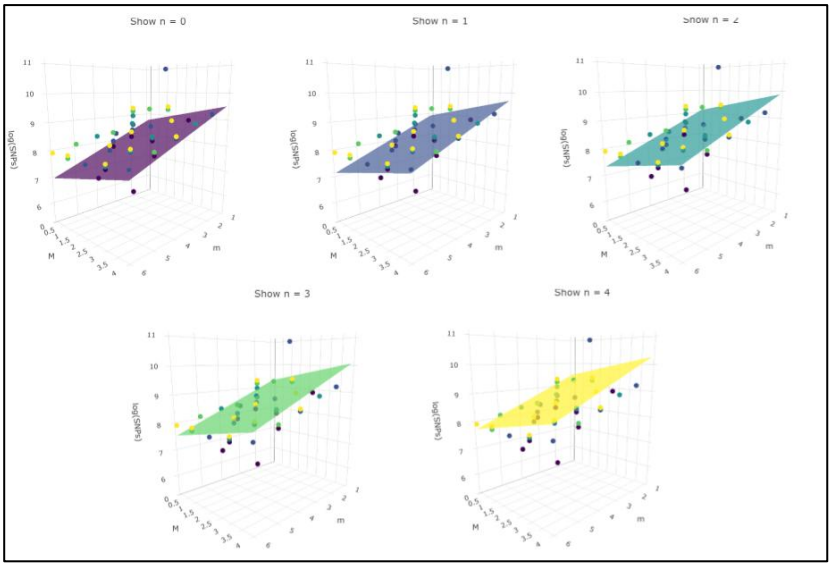
0.73 ± 0.30

- Prova 2: n ed M sono abbastanza importanti con **lm(log(SNPs) ~ m+M+n**, data = df). Su 30 repeats di 10-fold cross validation, R-squared: 0.76 ± 0.28

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.08343    0.28250  32.154  < 2e-16 ***
m           -0.34923    0.06518  -5.358 3.31e-06 ***
M            0.19751    0.06577   3.003  0.00449 **
n            0.18783    0.06299   2.982  0.00475 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6021 on 42 degrees of freedom
Multiple R-squared:  0.4623,    Adjusted R-squared:  0.4239
F-statistic: 12.04 on 3 and 42 DF,  p-value: 8.047e-06
```

Ho creato un grafico di quest'ultima regressione contenente tutti i parametri, ma ho dovuto creare un piano di regressione per ogni valore di n perché non sapevo come plottare 4 dimensioni, ho il file html se vuoi. In generale, anche da come si vede dal summary, m fa diminuire il numero di SNPs, mentre M ed n sembra lo facciano aumentare. È il modello migliore fino ad ora anche se solo leggermente rispetto a quello contenente solo m. L'ho inserito nell'algoritmo su Python che ho creato.



UNIQUE

Unique invece continua a dipendere solo da m (e forse da n):

- con m, M ed n: 0.76 ± 0.28

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.98514    0.28241  31.816  < 2e-16 ***
m           -0.36534    0.06516  -5.606 1.46e-06 ***
M            0.13229    0.06575   2.012   0.0507 .
n            0.13531    0.06297   2.149   0.0375 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6019 on 42 degrees of freedom
Multiple R-squared:  0.4451,    Adjusted R-squared:  0.4055
F-statistic: 11.23 on 3 and 42 DF,  p-value: 1.534e-05
```

- con m ed n: 0.77 ± 0.27

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.11348    0.28470  32.011  < 2e-16 ***
m           -0.32680    0.06445  -5.070 8.04e-06 ***
n            0.12546    0.06496   1.931   0.0601 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6229 on 43 degrees of freedom
Multiple R-squared:  0.3916,    Adjusted R-squared:  0.3633
F-statistic: 13.84 on 2 and 43 DF,  p-value: 2.29e-05
```

- con solo m: 0.76 ± 0.28

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31055    0.27390  33.992  < 2e-16 ***
m           -0.31366    0.06605  -4.749  2.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6419 on 44 degrees of freedom
Multiple R-squared:  0.3388,    Adjusted R-squared:  0.3238
F-statistic: 22.55 on 1 and 44 DF,  p-value: 2.201e-05
```

```
Model for SNPs:
  Features: ['m' 'M' 'n']
  Coefficients: [-0.3492325   0.19750855  0.18782753]
  Intercept: 9.083434280047847
---------------------------------------
Model for unique:
  Features: ['m' 'n']
  Coefficients: [-0.32679541  0.12545854]
  Intercept: 9.113478809885112
---------------------------------------
Model for sd:
  Features: ['m']
  Coefficients: [0.25126343]
  Intercept: 0.31958304533947046
---------------------------------------
Model for mean:
  Features: ['m' 'M']
  Coefficients: [2.88444349 0.08466461]
  Intercept: -0.9411739202873637
---------------------------------------
```

Ho deciso di inserire il modello migliore (quello con m ed n) sull’algoritmo in Python. (Ho controllato che scikit scelga gli stessi coefficienti di R) quindi il programmino Python contiene i seguenti modelli:

$\log(SNPs) = 9.0834 - 0.3492m + 0.1975M + 0.1878n$

$\log(unique) = 9.1135 - 0.3268m + 0.1254n$

$sd = 0.3196 + 0.2513m$

$mean = -0.9412 + 2.8844m + 0.0847M$

Quando faccio andare l’algoritmo su Python ottengo che la soluzione migliore è “m= 4, M=0, n=4”. Quando la provo ottengo:

- **Mean:** 10.8323 (predetto = 10.5964)
- **Sd:** 1.42895 (predetto = 1.3248)
- **SNPs:** 5264 (predetto = 4618)
- **Unique**: 3274 (predetto = 4055)

Ha sottostimato la media (che darebbe un risultato peggiore), ma ha sottostimato di gran lunga gli SNPs per locus (SNPs/unique) ed ha anche sottostimato la standard deviation (che ha anche sottostimato CV, predicendo 0.1250 contro il vero 0.1319), portandolo erroneamente a definire questa soluzione “ottimale”. Non è una brutta soluzione, ma semplicemente non è quella ottimale, quindi direi che questo modello da delle buone stime ma non perfette.

In generale la mia conclusione è che i modelli che includono ‘M’ ed ‘n’ (negli SNPs e unique) non migliorano moltissimo rispetto a quelli contenete solo ‘m’ (potrei farci un test d’ipotesi di t-student sulle 30 ripetizioni del 10-fold cross validation, ora non mi va ma sono sicura darebbe come risultato che le due soluzioni non siano differenti), quindi continua a scegliere a caso M ed n.