Forest Fires - Gruppo Capri

Andrei Petrisor 1085993, Antonio Radu 1085992, Lorenzo Medici 1085852, Andrea Rusconi 1086646

15-01-2024

Contents

1 Dataset

Questo dataset[1] è pubblico e a disposizione per la ricerca. I dettagli sul dataset possono essere trovati in Cortez e Morais (2007). Il dataset è composto dalle seguenti variabili:

1. Coordinata spaziale dell'asse X all'interno della mappa del parco Montesinho: da 1 a 9
2. Y coordinata spaziale dell'asse y all'interno della mappa del parco Montesinho: da 2 a 9
3. mese dell'anno: da "jan" a "dec"
4. giorno della settimana: da "mon" a "sun"
5. Indice FFMC dal sistema FWI: da 18,7 a 96,20
6. Indice DMC dal sistema FWI: da 1,1 a 291,3
7. Indice DC dal sistema FWI: da 7,9 a 860,6
8. Indice ISI del sistema FWI: da 0,0 a 56,10
9. temperatura temporanea in gradi Celsius: da 2,2 a 33,30
10. Umidità relativa RH in %: da 15,0 a 100
11. velocità del vento in km/h: da 0,40 a 9,40
12. pioggia in mm/m2: da 0,0 a 6,4
13. area della superficie bruciata della foresta (in ettari): da 0,00 a 1090,84.

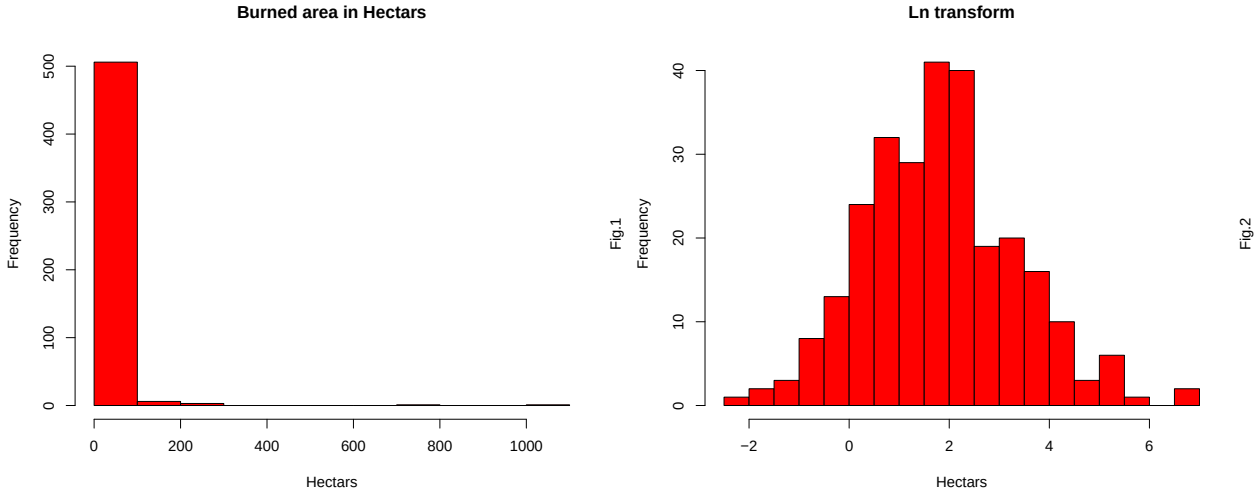
Il Forest Fire Weather Index (FWI) è il sistema canadese per la classificazione del pericolo di incendio e comprende sei componenti: Indice di umidità del combustibile (FFMC), indice di umidità (DMC), indice di siccità (DC), indice di dispersione iniziale (ISI) nel nostro caso indica la velocità della diffusione del fuoco, indice di accumulo (BUI) e FWI

Il dataset è composto dalle seguenti rilevazioni:

```
##       Days
## Months mon tue wed thu fri sat sun
##    jan   0   0   0   0   0   1   1
##    feb   3   2   1   1   5   4   4
##    mar  12   5   4   5  11  10   7
##    apr   1   0   1   2   1   1   3
##    may   0   0   0   0   1   1   0
##    jun   3   0   3   2   3   2   4
##    jul   4   6   3   3   3   8   5
##    aug  15  28  25  26  21  29  40
##    sep  28  19  14  21  38  25  27
##    oct   4   2   2   0   1   3   3
##    nov   0   1   0   0   0   0   0
##    dic   0   0   0   0   0   0   0
```

[1] Per ulteriori informazioni sul dataset visitare il link: http://www3.dsi.uminho.pt/pcortez/fires.pdf

1.1 Istogrammi area bruciata



Possiamo vedere come l'istogramma(Fig.1) presenti una asimmetria positiva(obliqua a destra) , di conseguenza valutiamo i dati facendo la trasformata grazie al logaritmo, cosifacendo otteniamo un grafico più simile ad una normale.

```
summary(forest$area)
```

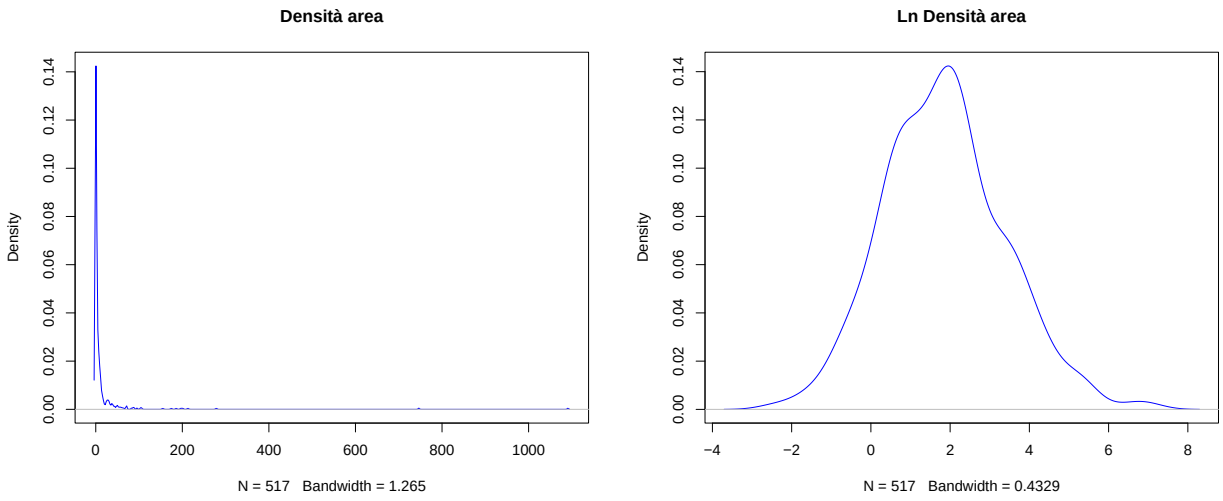
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.52   12.85    6.57 1090.84
```

Facendo il summary otteniamo i valori di: min, max, media, mediana, possiamo considerare i nostri dati molto vicini allo zero.

```
## [1] 0.4777563
```

Si evidenzia che il dataset ha il 48% dei valori che valgono 0.

1.2 Grafici densità area bruciata



Sembra esserci anche una forte distorsione rispetto ai residui che non sono normalmente distribuiti. Ciò sembra essere dovuto al gran numero di 0 nel database. Quando questi 0 vengono rimossi, possiamo vedere che i residui diventano più normalmente distribuiti.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.4187  1.1110  2.0242  6.9956
```

come possiamo vedere la mediana e la media sono diminuite.

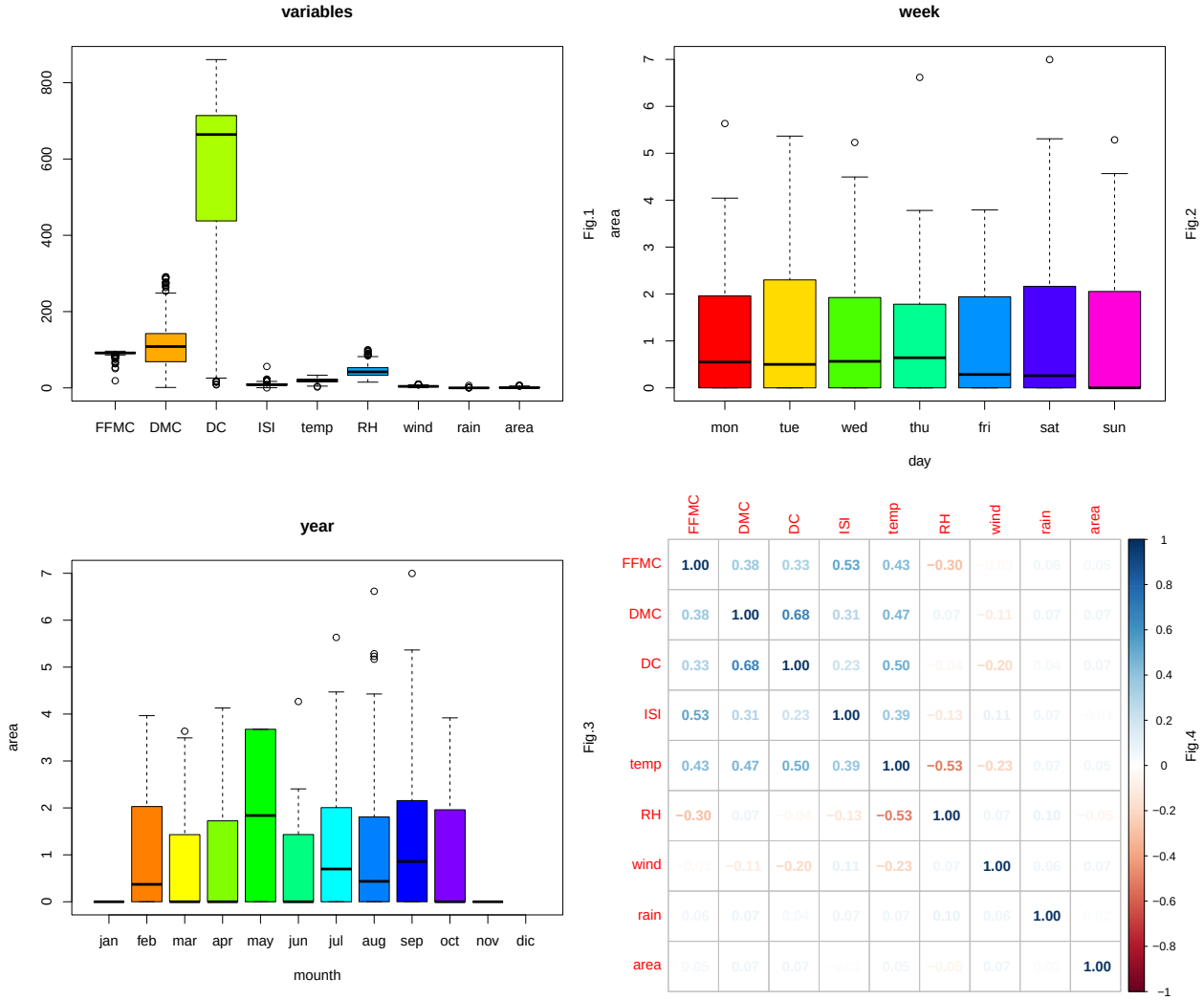
```
##   skewness deviazione  curtosi
## 1 1.214301   1.398436 3.924964
```

Possiamo vedere l'indice di simmetria(skew), l'indice di dispersione(curtosi) e la deviazione standard.

```
##
##  Jarque Bera Test
##
## data:  forest$area
## X-squared = 145.49, df = 2, p-value < 2.2e-16
```

Il Jarque Bera Test indica che la statistica del test è 145.49, con un valore p di 2.2e-16. Rifiuteremmo l'ipotesi nulla che i dati siano distribuiti normalmente in questa circostanza.

1.3 Boxplot



-Come possiamo vedere da questo correlation plot(fig.4) le variabili che sono correlate positivamente di più sono: DC con DMC, ISI e FFMC, temp e DC. Le variabili correlate negativamente di più sono, invece: RH e temp. Concentrandoci sulla variabile risposta (area) non ci sono correlazioni forti con nessuna delle covariate.

-Dal box plot delle variabili(fig.1) possiamo determinare che le mediane di quasi tutte le variabili sono più o meno simili quindi ci sono poche differenze, inoltre possiamo notare come in tutte le variabili ci siano degli outliers sia al minimo che al massimo. La covariata DMC, invece, presenta dei baffi più lunghi il che implica che tale variabile ha valori più incoerenti rispetto alle altre, la mediana è lontana da tutte le altre ed è spostata molto verso il terzo quartile. La covariata DC ha valori molto piu elevati rispetto alle altre covariate, percio non terremo conto di questa covariata.

-Dal box plot dei giorni della settimana(fig.2) possiamo vedere come le mediane siano simili tra di loro, come anche i baffi, questo dimostra come i giorni della settimana abbiano tutti valori coerenti tra di loro.

-Dal box plot dei mesi dell'anno(fig.3) notiamo come il mese di dicembre e quello di maggio siano quelli più incoerenti rispetto agli altri mesi. Il mese di agosto presenta molti più outliers rispetto agli altri mesi.

2 Domande specifiche

In questo dataset siamo interessati a modellare l'area bruciata della foresta come funzione delle altre variabili. Siamo in particolare interessati a capire nel mese di agosto come possiamo spiegare l'area bruciata in funzione delle altre variabili, vogliamo capire anche come si comportano i vari modelli(semplice, polinomiale e interazioni tra variabili) e interpretare i loro risultati. Nello specifico vogliamo scoprire da quali variabili dipende l'area bruciata maggiormente.

3 Metodologia

I metodi usati per raggiungere gli obiettivi sono dei metodi di regressione lineare, ci permettono di spiegare una variabile risposta (Y=area bruciata) in funzione delle altre variabili esplicative a disposizione nel dataset(covariate, X). Useremo due metodi di regressione:

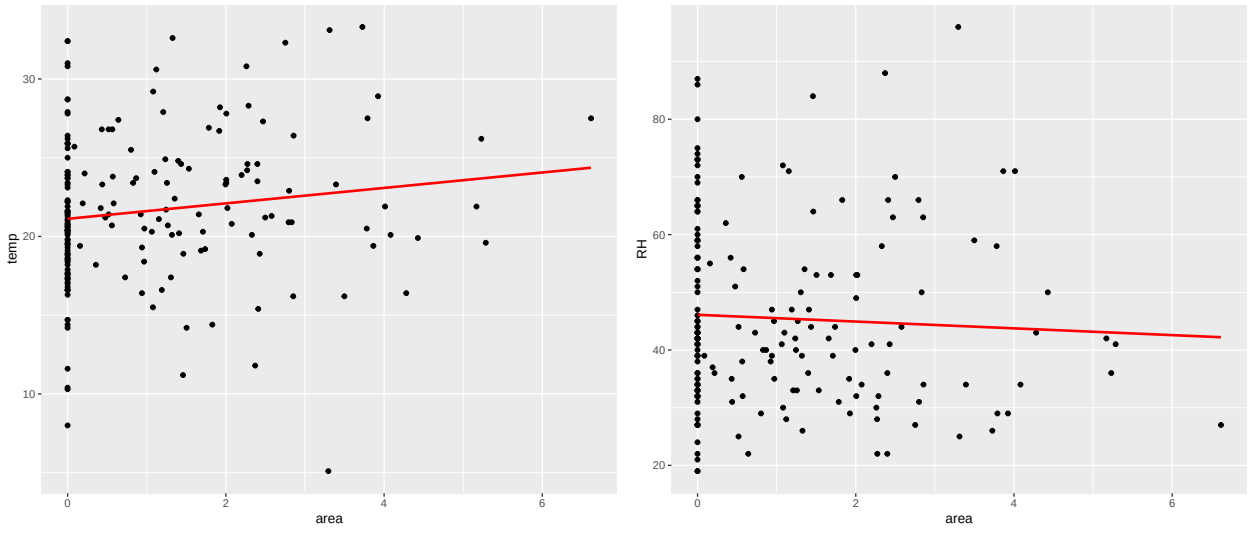
-LASSO regression: metodo usato per trovare una best-lambda che ci permette di penalizzare alcuni regressori, questo per rendere il modello piu semplice(facciamo model selection).

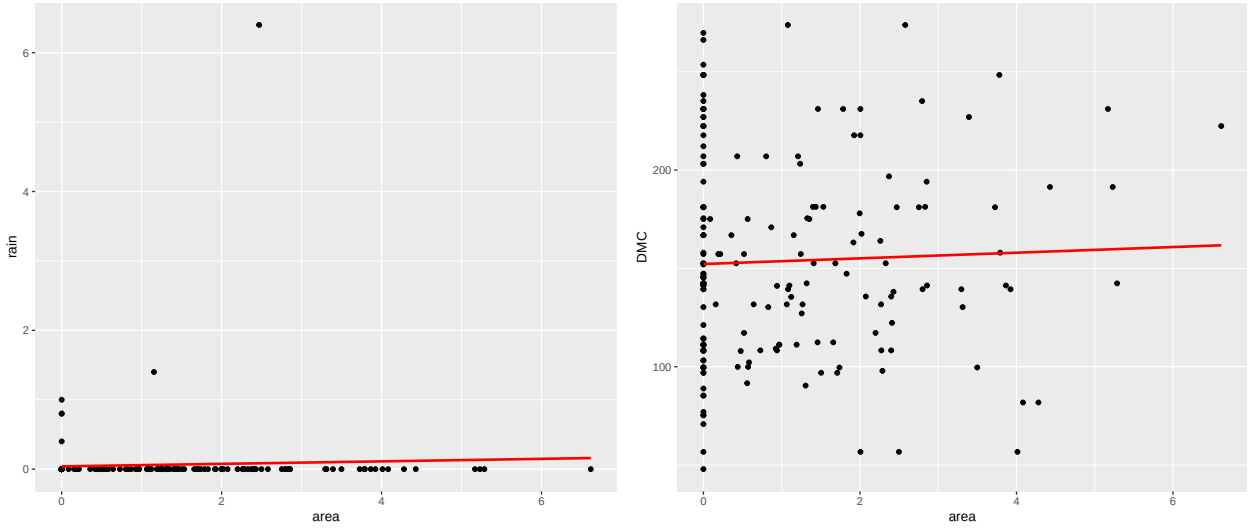
-K-fold cross validation: metodo usato per trattare un modello piu complesso(visto che abbiamo pochi dati), questo per non andare in overfitting cioe il modello performa bene in train, ma male in test.

4 Analisi dati e discussione risultati

Per la fase di testing prendiamo in considerazione solamente il mese di agosto per poter spiegare come varia l'area bruciata in relazione alle altre variabili. Dividiamo la fase di Testing in: train e test, con il train al 70% e il test al 30%(cioe con il 70% dei dati alleno il modello, mentre con il restante 30% faccio le previsioni con i metodi).

Proviamo a fare un modello lineare con i dati del mese di agosto. Le variabili che secondo noi ha piu senso considerare nel modello sono: rain ,RH ,temp, FFMC, e DMC. RH-FFMC-DMC sono degli indici che tengono conto dell'umidità. Nessuno di questi parametri agisce in modo diretto sullo sviluppo dell'incendio, ma sono tutti fattori predisponenti, perciò consideriamo rilevante studiare come influenzano l'area incendiata. Per capire la correlazione lineare tra la variabile risposta e le covariate mostriamo dei scatterplot.





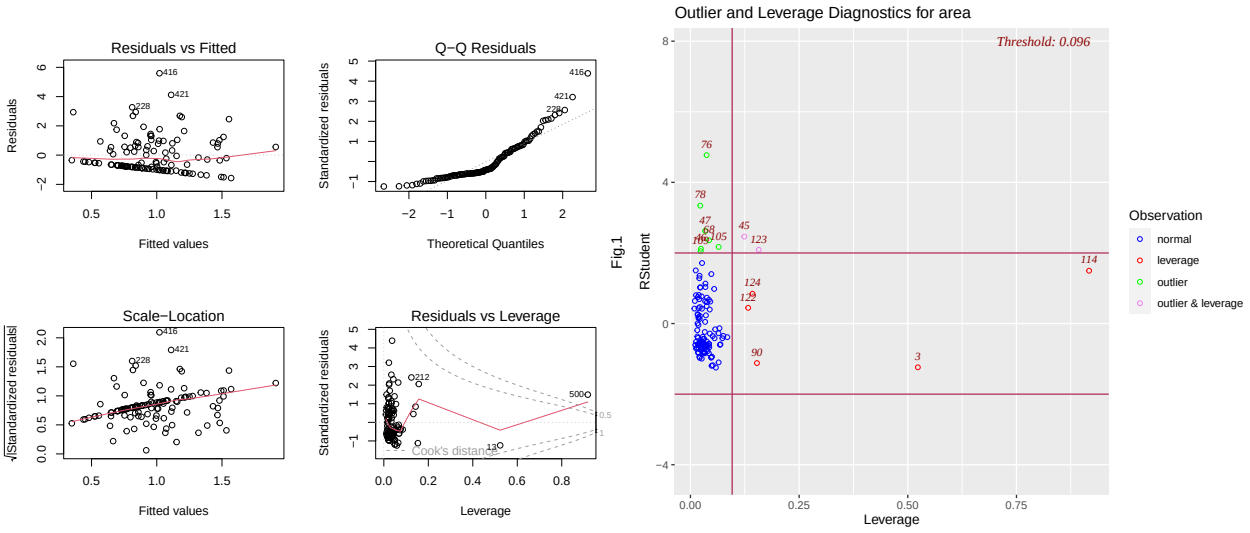
Come possiamo vedere le correlazioni non sono molto significative, l'r^2 tende a 0 quindi c'è una forte mancanza di correlazione che ci suggerisce di usare il modello vuoto per i nostri test. Le correlazioni tra area-DMC, area-rain e area-RH sono le piu evidenti a tal proposito. Proviamo ora a creare 3 modelli(semplice, polinomiale e interazionale)(Per realizzare il modello interazionale abbiamo usato la tecnica step-wise).

```
##
## Call:
## lm(formula = area ~ rain + RH + temp + FFMC + DMC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5702 -0.8345 -0.5287  0.6288  5.5955
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.292937   3.540255  -0.365   0.7156
## rain         0.044054   0.212911   0.207   0.8364
## RH           0.014859   0.012451   1.193   0.2351
## temp         0.083263   0.039326   2.117   0.0363 *
## FFMC         0.001123   0.035157   0.032   0.9746
## DMC         -0.002174   0.002545  -0.854   0.3947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.3 on 119 degrees of freedom
## Multiple R-squared:  0.04681,    Adjusted R-squared:  0.006758
## F-statistic: 1.169 on 5 and 119 DF,  p-value: 0.3286
```

```
##
## Call:
## lm(formula = area ~ temp + (rain * temp) + (temp * FFMC) + (RH *
##     FFMC) + (train$day == "tue") + (train$day == "wed") + (train$day ==
##     "thu"), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.3428 -0.8996 -0.4608  0.7033  5.2668
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -79.417342  41.424874  -1.917   0.0577 .
## temp                       2.473365   1.204105   2.054   0.0423 *
## rain                      -4.701083   3.246733  -1.448   0.1504
## FFMC                       0.841054   0.450289   1.868   0.0644 .
## RH                         0.524843   0.318901   1.646   0.1026
## train$day == "tue"TRUE     0.155754   0.357349   0.436   0.6638
## train$day == "wed"TRUE     0.066586   0.366138   0.182   0.8560
## train$day == "thu"TRUE     0.244584   0.342235   0.715   0.4763
## temp:rain                  0.179877   0.119921   1.500   0.1364
## temp:FFMC                 -0.025822   0.013036  -1.981   0.0500 .
## FFMC:RH                   -0.005501   0.003473  -1.584   0.1160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.294 on 114 degrees of freedom
## Multiple R-squared:  0.09437,    Adjusted R-squared:  0.01493
## F-statistic: 1.188 on 10 and 114 DF,  p-value: 0.3063
```

Come possiamo vedere i nostri modelli non performano benissimo, con un 4% e 9% come r^2 i modelli hanno un grande margine di errore. Tra le varie covariate scelte notiamo come la temperatura abbia il valore piu grande, quindi, all'aumentare unitario dell'area, la temperatura media aumenta di 0.08 nel primo e 2.47 nel secondo. Il modello polinomiale non ci ha dato risultati positivi quindi abbiamo deciso di scartatlo a priori. Togliamo dal modello le covariate con i p-values piu grandi(FFMC per il primo). Proviamo a vedere se ci sono degli Outliers(valori estremi) nei modelli.



Possiamo vedere come nel grafici (Fig.1), un numero considerevole di osservazioni non segue l'andamento desiderato, poniamo maggiore attenzione sul grafico Q-Q Res. dove possiamo vedere che i vari valori alle code tendono a spostarsi di molto, l'andamento non tende per niente alla normale. Per quanto riguarda il grafico a destra possiamo notare come alcune rilevazioni siano parecchio "lontane" rispetto alle altre, poniamo attenzione soprattuto a quelle che sono Outlier&Leverage e le rimuoviamo.

```
##
## Call:
```

```
## lm(formula = area ~ rain + RH + temp + DMC, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7785 -0.7302 -0.4259  0.6797  3.0669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.238822   1.080414  -2.072  0.04056 *
## rain        -1.614104   0.756201  -2.134  0.03500 *
## RH           0.028180   0.010600   2.658  0.00901 **
## temp         0.113035   0.033566   3.368  0.00104 **
## DMC         -0.004020   0.002037  -1.974  0.05092 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.052 on 111 degrees of freedom
## Multiple R-squared:  0.1136, Adjusted R-squared:  0.0817
## F-statistic: 3.558 on 4 and 111 DF,  p-value: 0.009039
```

```
##
## Call:
## lm(formula = area ~ temp + (rain * temp) + (temp * FFMC) + (RH *
##     FFMC) + (train$day == "tue") + (train$day == "wed") + (train$day ==
##     "thu"), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5245 -0.7476 -0.3644  0.7575  2.8817
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -49.269550  35.410430  -1.391   0.1670
## temp                     1.223043   1.075007   1.138   0.2578
## rain                    -0.580570 106.932174  -0.005   0.9957
## FFMC                     0.513098   0.385028   1.333   0.1855
## RH                       0.571426   0.276782   2.065   0.0414 *
## train$day == "tue"TRUE   0.524335   0.315860   1.660   0.0999 .
## train$day == "wed"TRUE  -0.045037   0.308423  -0.146   0.8842
## train$day == "thu"TRUE   0.093087   0.286881   0.324   0.7462
## temp:rain               -0.036039   4.917039  -0.007   0.9942
## temp:FFMC               -0.012312   0.011632  -1.058   0.2923
## FFMC:RH                 -0.006000   0.003019  -1.987   0.0495 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 105 degrees of freedom
## Multiple R-squared:  0.1609, Adjusted R-squared:  0.08103
## F-statistic: 2.014 on 10 and 105 DF,  p-value: 0.03889
```

Il modello presenta un leggero miglioramento, l'r^2 aumenta fino a 12% e 16% rendendo i modelli sono un minimo piu precisi, quindi la correlazione tra la risposta e le covariate è migliore. La covariata piu significativa è la temperatura, quindi all'aumentare unitario dell'area, la temperatura media aumenta di 0.11 per il primo e 1.22 per il secondo. I coefficienti associati alle variabili rain, DMC non sono significativi, quindi

supponiamo non influiscano sull'area. Passando da non avere la rain ad averla, a parità di temperatura, l'area diminuisce di 1.61 per il primo e 0.58 per il secondo.

4.1 Scelta modello

Per valutare quale dei 3 modelli (1o modello con variabili scelte da noi, il 2o modello con tutte le variabili e il 3o modello presenta le interazioni tra le variabili) sia meglio li mettiamo a confronto tramite la tecnica AIC(Akaike information criterion). Consiste in un metodo per la valutazione e il confronto tra modelli statistici. Fornisce una misura della qualità della stima di un modello statistico tenendo conto sia della bontà di adattamento che della complessità del modello. La regola è quella di preferire i modelli con l'AIC più basso.

```
## AIC lm 1 BIC lm 1 AIC lm 2 BIC lm 2 AIC lm 3 BIC lm 3
## 347.9498 364.4713 360.0412 404.0986 353.5877 386.6308
```

I risultati ci mostrano come il modello 1 abbia l'AIC piu basso, useremo questo modello per fare le previsioni.

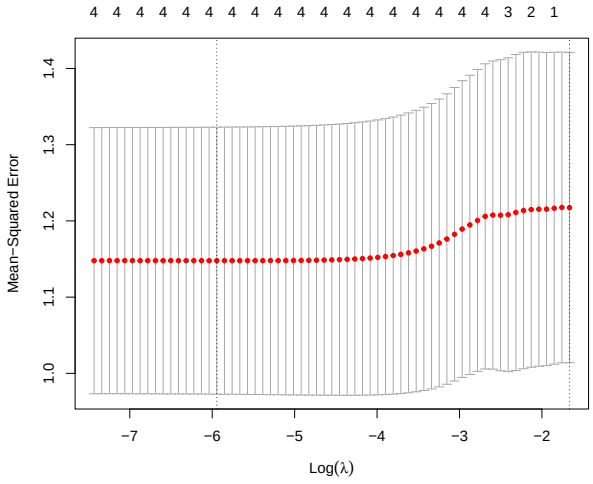
4.2 Previsioni puntuale

Per stima puntuale s'intende l'insieme dei metodi inferenziali che permettono di attribuire un valore ad un parametro della popolazione, utilizzando i dati di un campione casuale osservato (x1, x2,. . . ,xn) ed elaborandoli.Per valutare la bontà di uno stimatore è necessario considerare le stime ottenute ripetendo un grande numero di volte il processo impiegato per eseguire una stima.

```
## [1] 1.541404
```

4.3 LASSO regression

La tecnica LASSO(least absolute shrinkage and selection operator) è un metodo di analisi della regressione che esegue sia la selezione delle variabili sia la regolarizzazione per migliorare l'accuratezza della previsione e l'interpretabilità del modello statistico risultante. Il metodo lasso presuppone che i coefficienti del modello lineare siano sparsi, ossia che pochi di essi siano non nulli.



Il grafico dovrebbe assomigliare ad una curva esponenziale, nel nostro caso il mean-squared error assomiglia ad una retta. Questo sta ad indicare che indipendentemente dal lambda non cambia la penalizzazione, dovuto al fatto che i regressori non hanno molto senso.

4.4 Previsioni con Lasso

```
#use lasso regression model to predict response value
new = data.matrix(test[,c("temp","rain","RH","DMC")])
previsioni = predict(best_model, s = best_lambda, newx = new)

# Root Mean Squared Error (RMSE): è una misura dell'errore che compiamo
sqrt(mean((test$area - previsioni)^2))#errore minimo
```

```
## [1] 1.539185
```

Come possiamo vedere la previsione con la Lasso è poco migliore rispetto alla previsone puntuale, decidiamo quindi di usare la Lasso regression per prevedere i dati.

```
media_area=mean(test$area)
print(media_area)
```

```
## [1] 1.220827
```

Il nostro RMSE è molto vicino alla media dell'area bruciata, questo vuol dire che il modello non performa bene (è dovuto al dataset).

5 Conclusioni

Concentrandosi sul mese di agosto, spiegare l'area bruciata in funzionde delle altre variabili risulta non facile visto la poca quantità di dati a disposizione. Poiché il valore r2 del modello è molto basso, probabilmente i predittori avranno un errore significativo e potrebbero non essere affidabili. Per una maggior accuratezza nei risultati ci servirebbero piu dati o aggiungere qualche altra variabile (e.g. variabili spaziali). Abbiamo valutato anche il modello 6(senza mostrare i risultati) tramite la cross-validation per capire come si comportasse, il risultato è stato che anche quest'ultimo non performa bene.[2]

[2]Abbiamo consultato diverse fonti per realizzare questo report: Per il layout- https://bookdown.org/yihui/rmarkdown/. Per confrontare i risultati- https://rstudio-pubs-static.s3.amazonaws.com/419751_b251adb1ab8e40f7aeab8b5c4a739c4f.html. Per risolvere problemi di natura tecnica- https://stackoverflow.com/