Forest Fires - Gruppo Capri

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

15-01-2024

Contents

1 Dataset

Questo dataset (A Data Mining Approach to Predict Forest Fires using Meteorological Data)[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

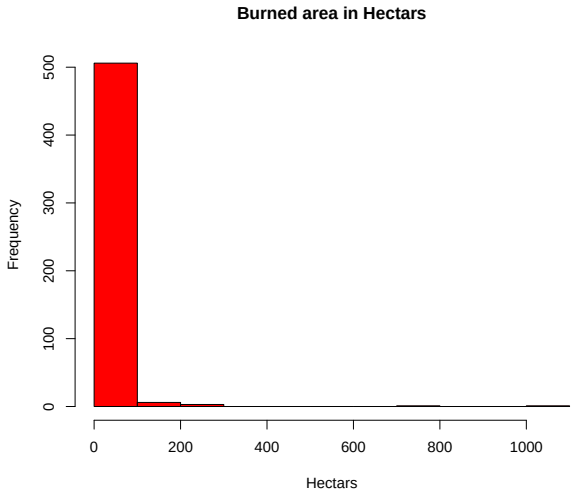
Il dataset è composto dalle seguenti rilevazioni:

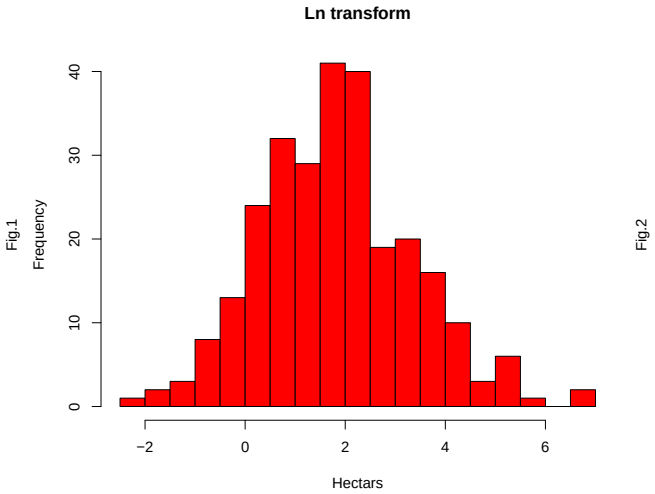
```
##       Days
## Months mon tue wed thu fri sat sun
##    jan   0   0   0   0   0   1   1
##    feb   3   2   1   1   5   4   4
##    mar  12   5   4   5  11  10   7
##    apr   1   0   1   2   1   1   3
##    may   0   0   0   0   1   1   0
##    jun   3   0   3   2   3   2   4
##    jul   4   6   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
```

La tabella rappresenta il numero di osservazioni per mese(righe) e per giorno(colonne) della settimana.

[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 log-trasformiamo i dati, cosifacendo otteniamo un grafico più simile ad una normale.

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

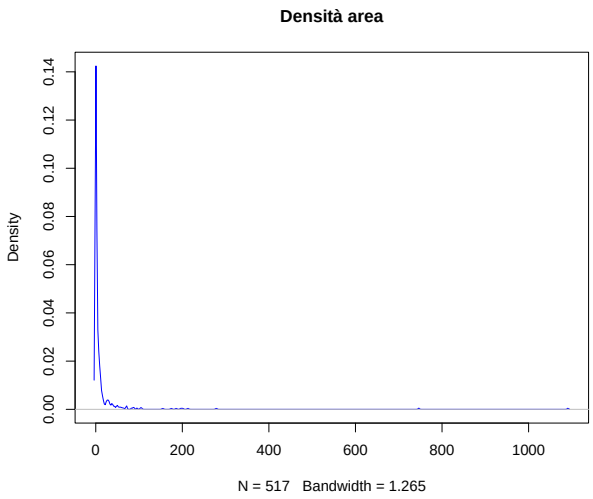
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.52   12.85    6.57 1090.84
```

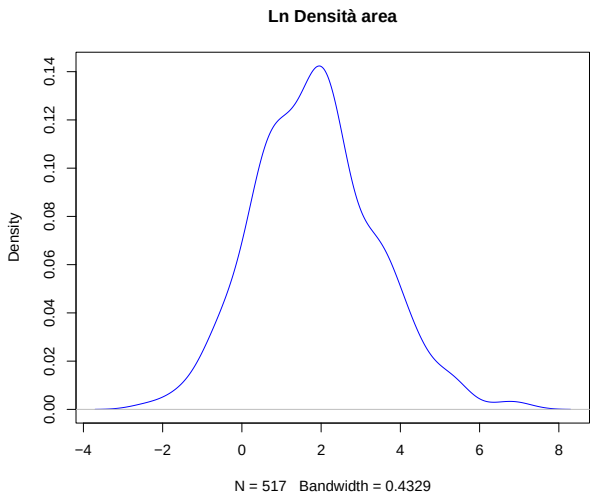
Facendo il summary dei dati originali(non quelli log-trasformati) otteniamo i valori di: min, max, media, mediana, possiamo considerare i nostri dati molto vicini allo zero.

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

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

1.2 Grafici densità area bruciata





```
##    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
```

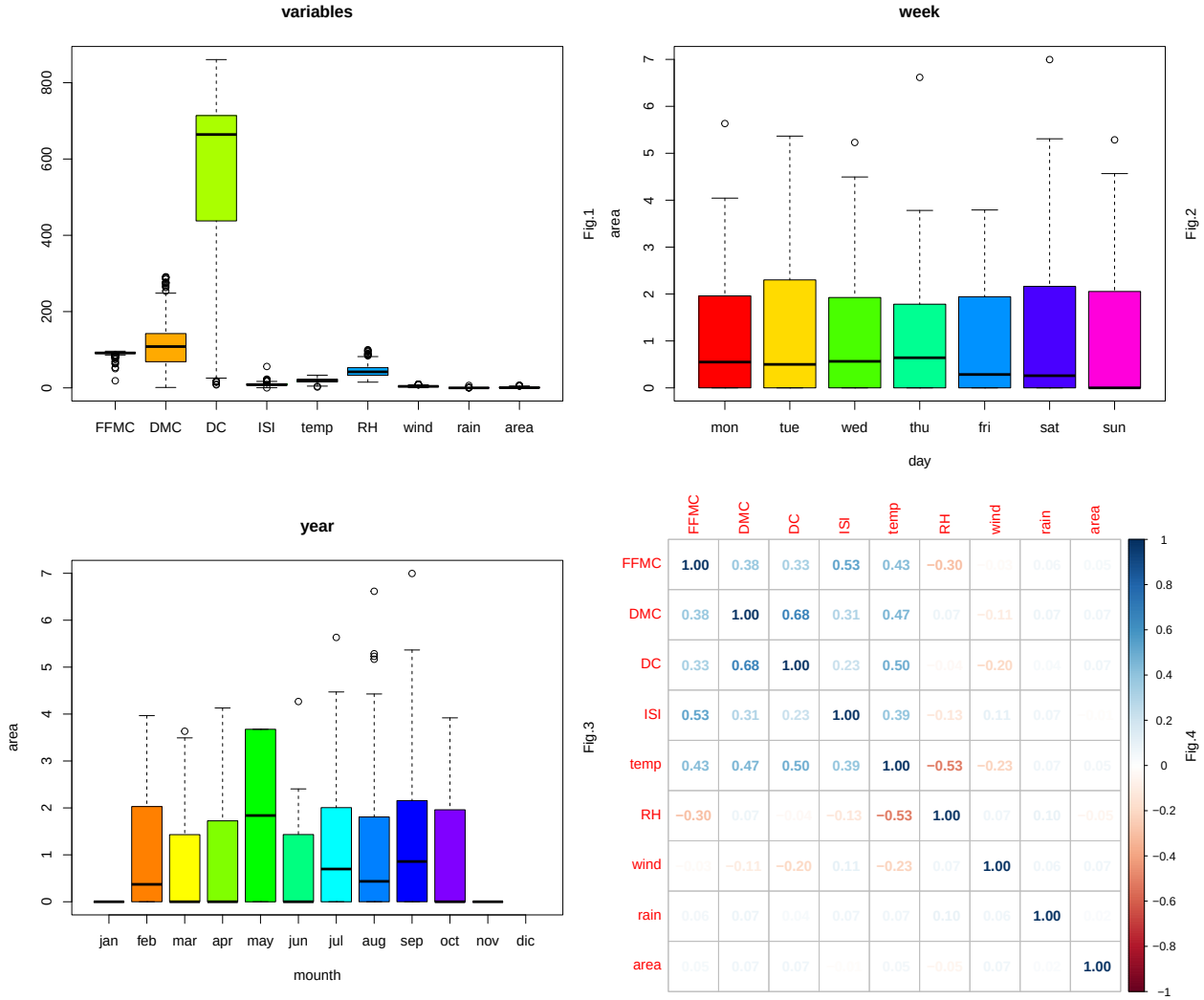
Possiamo vedere l'indice di asimmetria(skewness), l'indice di dispersione(curtosi) e la deviazione standard.

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

Si tratta di un test di normalità che verifica, come ipotesi nulla, simultaneamente se l'asimmetria(skewness) e la curtosi sono coerenti con i valori che dovrebbero assumere sotto l'ipotesi nulla di normalità, ossia rispettivamente 0 e 3. Sotto l'ipotesi nulla H0 il test si distribuisce asintoticamente come una chi-quadro con 2 gradi di libertà. Tale ipotesi viene rifiutata per valori di JB troppo grandi.

Il Jarque Bera Test indica che la statistica del test è 145.49, con un valore p di 2.2e-16. Rifiuteremo, per quando detto in precedenza, 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

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

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). Seguiremo i seguenti step:

-Step1. Dato che non sarebbe sufficientemente esaustivo usare solo una porzione dei dati, verranno utilizzati tutti i dati a nostra disposizione. Alleneremo quindi 3 modelli di diversi di complessità.

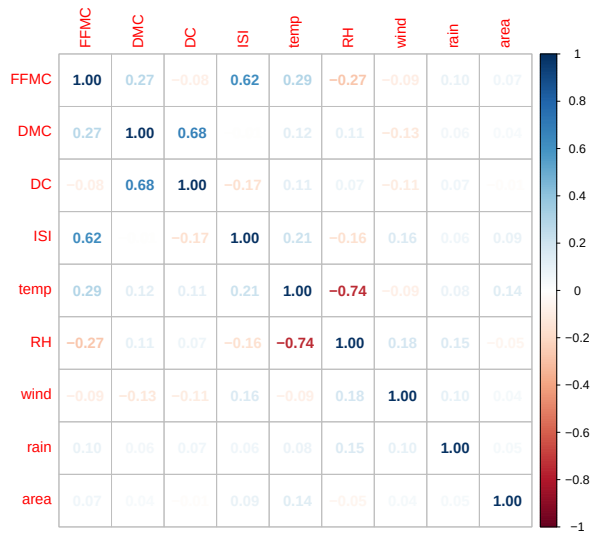
-Step2. Rimuoveremo gli outliers per vedere come cambiano gli R2, cosi facendo scegliamo il miglior modello in base a chi ha l'RMSE piu basso.

-Step3. Utilizzeremo la regressione Lasso che ci permette di penalizzare alcuni regressori, questo per rendere il modello piu semplice(facciamo model selection), successivamente useremo la k-fold cross-validation per trovare un valore di lambda ottimale, la cross-validation è un ottimo modo per non overfittare e trovare il miglior modello.

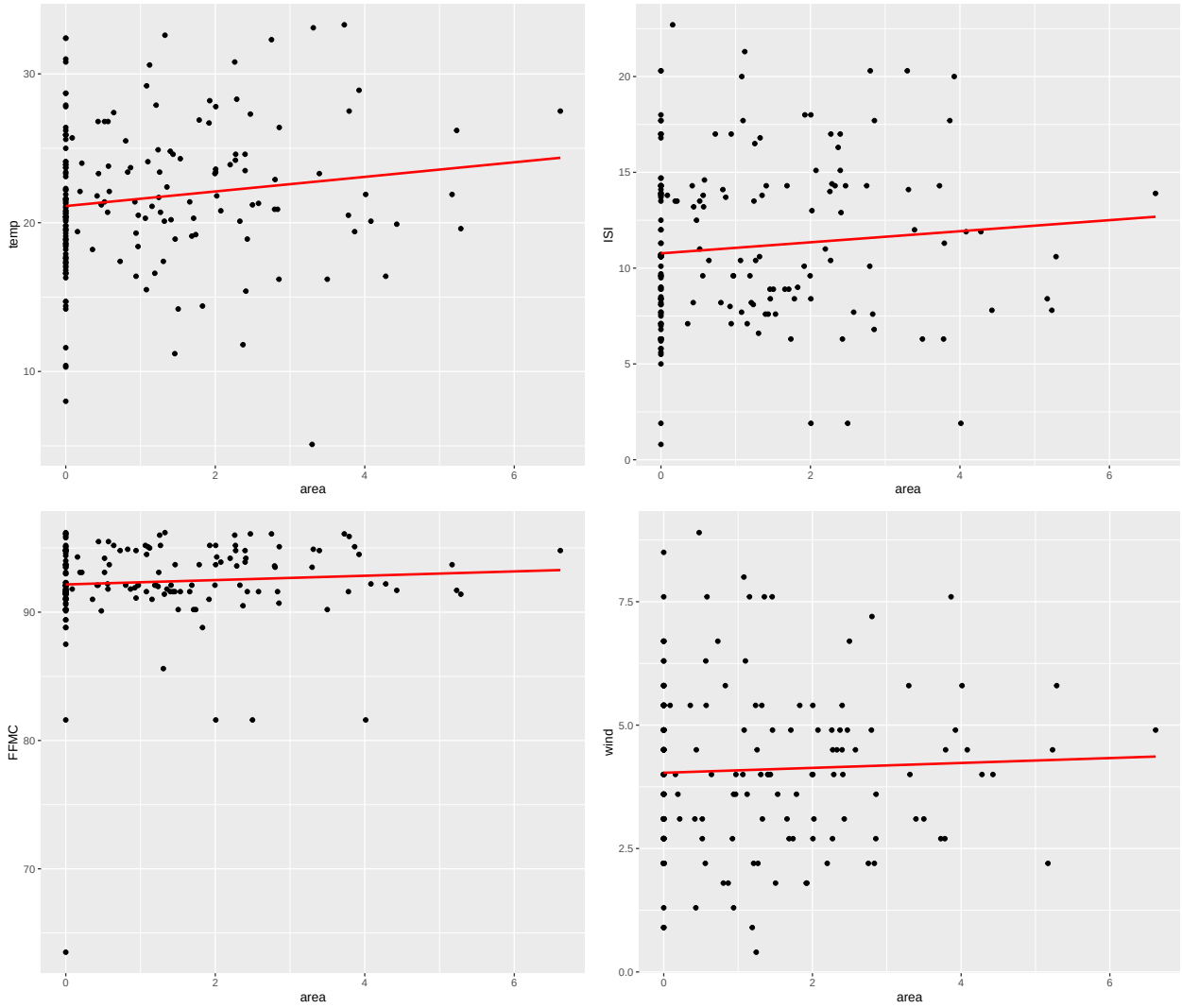
4 Analisi dati e discussione risultati

Prendiamo in considerazione solamente il mese di agosto per poter spiegare come varia l'area bruciata in relazione alle altre variabili. Questa fase sarà svolta su tutti i campioni a nostra disposizione(non ha senso prendere il 70% dei campioni per via del basso numero di dati). La generalizzazione dei modelli viene fatta nello Step2 tramite cross-validazione, alleneremo perciò diversi modelli con tutti i dati per poi semplificarli tramite regressione Lasso.

Viene riportato il correlation plot riferito al mese di agosto:



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 alcuni scatterplot.



Come possiamo vedere le correlazioni non sono molto significative, l'R2 tende a 0 quindi c'è una forte mancanza di correlazione. Visto il gran numero di zeri deicidiamo di togliere la maggioraparte di essi per vedere come performano i modelli. Proviamo ora a creare 3 modelli(semplice, polinomiale e polinomiale con interazioni).

```
##
## Call:
## lm(formula = area ~ DMC + ISI, data = forest_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9149 -0.9234 -0.2672  0.6182  4.4195
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

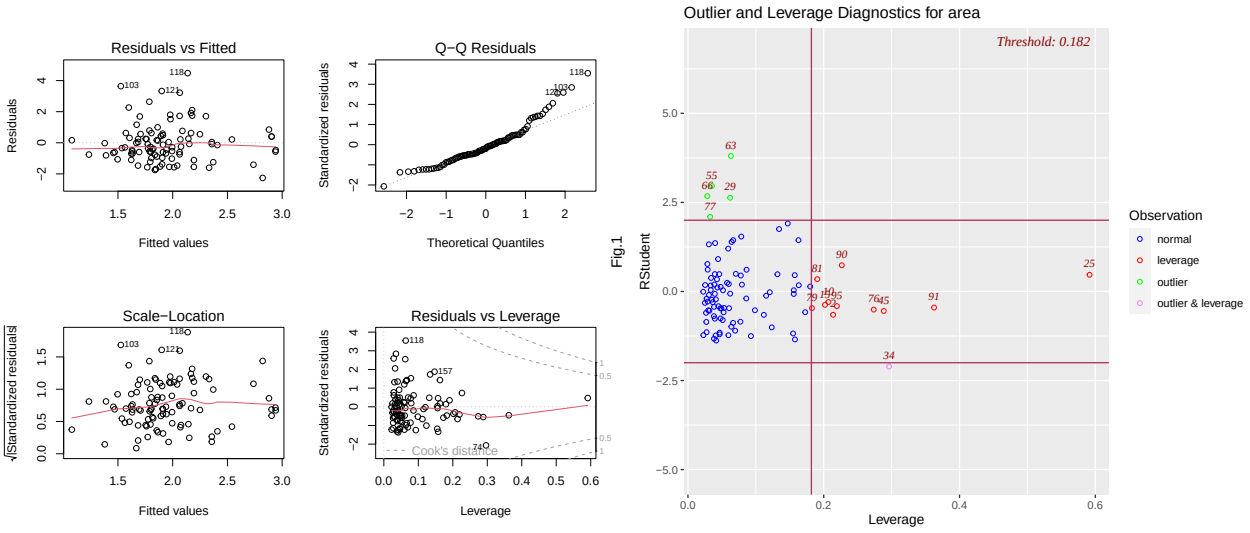
```
## (Intercept)  1.542864   0.576518   2.676  0.00876 **
## DMC          0.004183   0.002802   1.493  0.13874
## ISI         -0.019877   0.030220  -0.658  0.51227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.298 on 96 degrees of freedom
## Multiple R-squared:  0.02752,    Adjusted R-squared:  0.007262
## F-statistic: 1.358 on 2 and 96 DF,  p-value: 0.262
```

```
##
## Call:
## lm(formula = area ~ I(temp^2) + I(temp^3) + wind + I(wind^2),
##     data = forest_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9499 -0.8379 -0.2535  0.5018  4.4406
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7420641  1.0104621   1.724   0.0880 .
## I(temp^2)   -0.0072428  0.0043222  -1.676   0.0971 .
## I(temp^3)    0.0002099  0.0001201   1.747   0.0839 .
## wind         0.6584183  0.3102548   2.122   0.0365 *
## I(wind^2)   -0.0699695  0.0335749  -2.084   0.0399 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.278 on 94 degrees of freedom
## Multiple R-squared:  0.07668,    Adjusted R-squared:  0.03739
## F-statistic: 1.952 on 4 and 94 DF,  p-value: 0.1083
```

```
##
## Call:
## lm(formula = area ~ temp + I(temp^3) + wind + RH + I(RH^2) +
##     I(RH * wind) + I(temp^2 * wind^2) + I(RH * temp), data = forest_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2624 -0.7515 -0.2469  0.5201  4.4795
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.144e+01  9.992e+00   1.145    0.255
## temp               -5.201e-01  4.264e-01  -1.220    0.226
## I(temp^3)           2.066e-04  1.400e-04   1.476    0.143
## wind                7.238e-01  4.549e-01   1.591    0.115
## RH                 -2.160e-01  1.986e-01  -1.088    0.280
## I(RH^2)             1.488e-03  1.068e-03   1.393    0.167
## I(RH * wind)       -1.048e-02  6.562e-03  -1.597    0.114
## I(temp^2 * wind^2) -5.223e-05  5.760e-05  -0.907    0.367
## I(RH * temp)        5.713e-03  4.791e-03   1.192    0.236
##
```

```
## Residual standard error: 1.305 on 90 degrees of freedom
## Multiple R-squared:  0.07838,    Adjusted R-squared:  -0.00354
## F-statistic: 0.9568 on 8 and 90 DF,  p-value: 0.4748
```

Come possiamo vedere i nostri modelli non performano benissimo, con degli R2 (2%,7.7%,7.8%) così bassi i modelli hanno un elevato margine di errore. Il p-value del modello 1 ci suggerisce che il modello migliore è quello senza nessuna covariata e quindi avrebbe senso tenere soltanto l'intercetta(p-value > 5%). Il modello 2 ci evidenzia come i termini polinomiali rendano la correlazione tra area e le covariate un pò più significativa (rispetto al modello 1). Il modello 3 ha un R2 simile al modello 2, ma le sue covariate non sono significative. Togliamo dal primo modello le covariate con i p-value piu grandi(ISI). Proviamo a vedere se ci sono degli Outliers(valori estremi) nei modelli.



Possiamo vedere come nei 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 ~ DMC, data = forest_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7878 -0.8403 -0.2883  0.5930  3.4458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.691290   0.420815   4.019  0.00012 ***
## DMC         0.001043   0.002684   0.389  0.69839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.189 on 91 degrees of freedom
## Multiple R-squared:  0.001658,   Adjusted R-squared:  -0.009313
## F-statistic: 0.1511 on 1 and 91 DF,  p-value: 0.6984
```

```
##
## Call:
## lm(formula = area ~ I(temp^2) + I(temp^3) + wind + I(wind^2),
##     data = forest_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8587 -0.7725 -0.1380  0.5307  3.3284
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.0437390  1.0521795   1.942   0.0553 .
## I(temp^2)   -0.0086600  0.0049543  -1.748   0.0840 .
## I(temp^3)    0.0002422  0.0001345   1.800   0.0752 .
## wind         0.6357422  0.2901282   2.191   0.0311 *
## I(wind^2)   -0.0674893  0.0311581  -2.166   0.0330 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.161 on 88 degrees of freedom
## Multiple R-squared:  0.07954,    Adjusted R-squared:  0.0377
## F-statistic: 1.901 on 4 and 88 DF,  p-value: 0.1173
```

```
##
## Call:
## lm(formula = area ~ temp + I(temp^3) + wind + RH + I(RH^2) +
##     I(RH * wind) + I(temp^2 * wind^2) + I(RH * temp) + I(temp *
##     ISI), data = forest_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8098 -0.6908 -0.1467  0.5197  3.2626
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.366e+01  1.041e+01   2.272  0.02569 *
## temp               -1.103e+00  4.599e-01  -2.398  0.01871 *
## I(temp^3)           3.903e-04  1.436e-04   2.718  0.00799 **
## wind                1.230e+00  4.207e-01   2.924  0.00445 **
## RH                 -5.099e-01  2.139e-01  -2.384  0.01943 *
## I(RH^2)             3.307e-03  1.174e-03   2.816  0.00608 **
## I(RH * wind)       -2.003e-02  6.930e-03  -2.891  0.00491 **
## I(temp^2 * wind^2) -7.563e-05  5.186e-05  -1.458  0.14854
## I(RH * temp)        1.419e-02  5.789e-03   2.451  0.01633 *
## I(temp * ISI)      -1.800e-03  1.361e-03  -1.323  0.18956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.122 on 83 degrees of freedom
## Multiple R-squared:   0.19,  Adjusted R-squared:  0.1022
## F-statistic: 2.164 on 9 and 83 DF,  p-value: 0.0328
```

I modelli presentano un leggero miglioramento. Il modello 1 peggiora e quindi ci dice in definitiva che un modello lineare semplice non conviene per studiare questo dataset. Il modello 2 migliora leggermente, un

R2 cosi piccolo e un R2 corretto cosi diverso ci indica che ciò è dovuto al fatto che il secondo corregge il primo tenendo conto anche di n(numero campioni) e di p(parametri). Se n non è molto più grande di p, allora l'R2 corretto penalizza R2 poiché c'è il forte rischio di andare in over-fitting. Il modello 3 si è rilevato essere il miglior modello tra i 3, con un R2 di 0.19(comunque basso) e le covariate che diventano un minimo significative, è il modello che si adatta meglio ai nostri dati.

4.1 Scelta del modello

Per valutare quale dei 3 modelli sia meglio abbiamo analizzato in precedenza i diversi R2 per capire chi avesse un RMSE piu piccolo(modello 3). Vogliamo mettere i 3 modelli a confronto anche 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
## 300.1716 307.7694 298.6177 313.8133 296.7243 324.5829
```

I risultati ci mostrano come il modello 3 abbia l'AIC piu basso nonostante sia molto piu complesso rispetto al primo e al secondo modello, questo significa che il primo e il secondo modello performano molto male. Scegliamo(tramite R2) il 3 modello per svolgere le previsioni.

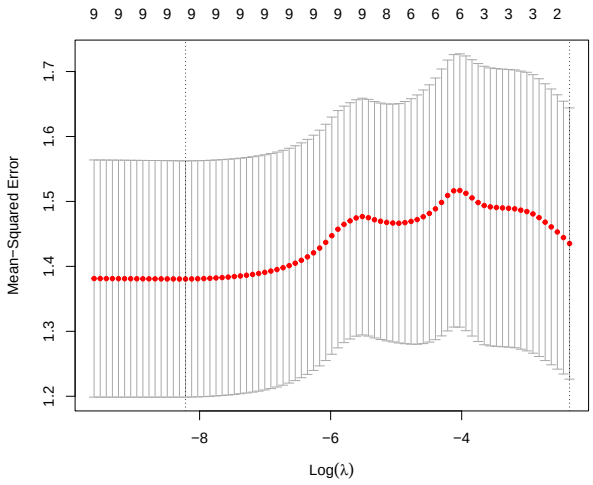
4.2 Errore medio commesso dal modello scelto

Calcoliamo il RMSE del modello 3 in training(cioè calcolata con gli stessi dati utilizzati per allenare il modello).

```
## [1] 1.059793
```

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. Diversamente dalla tecnica Ridge, la regressione Lasso manda a esattamente a zero i coefficienti non significativi, eseguendo i questo modo una vera e propria semplificazione del modello.



Il grafico dovrebbe assomigliare ad una curva esponenziale, nel nostro caso il mean-squared error tende ad un andamento esponenziale, ma alla fine torna al punto di partenza. Questo sta ad indicare che il modello diventa troppo semplice (under-fitting), quindi le sue performance peggiorano.

4.4 Previsioni con Lasso

```
## [1] 1.224745
```

Il RMSE di test calcolato in cross-validazione durante la procedura di scelta di lambda ((1.5)^(1/2) = 1.22) è peggiore di quello di training (1.05). (E' normale che un modello si comporti meglio sui dati sui quali è stato allenato piuttosto che su quelli nuovi.)

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, i regressori non hanno una potenza espressiva tale da permettere un'adeguata descrizione della variabile in uscita, oggetto dell'analisi, tramite un modello lineare. Gli sviluppi futuri potrebbero essere superare la linearità e optare su modelli statistici più complessi, per esempio:

-Avere piu dati e/o aggiungere qualche altra variabile (e.g. variabili spaziali) per una maggior accuratezza nei risultati.

-Si potrebbero implementare delle reti neurali per analizzare i dati.

-Si potrebbero utilizzare altre tecniche come la SVM o la random forest(non overfitta, è molto accurata ed parecchio complessa). [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/