Feng Chia University

Predicting and Exploring Factors Influencing the Magnitude of Forest Fires and Burned Area

International Business

D083209

Shu-Xi Chen

Regression Analysis STAT406

Jan 09, 2023

Chapter 1. Information description

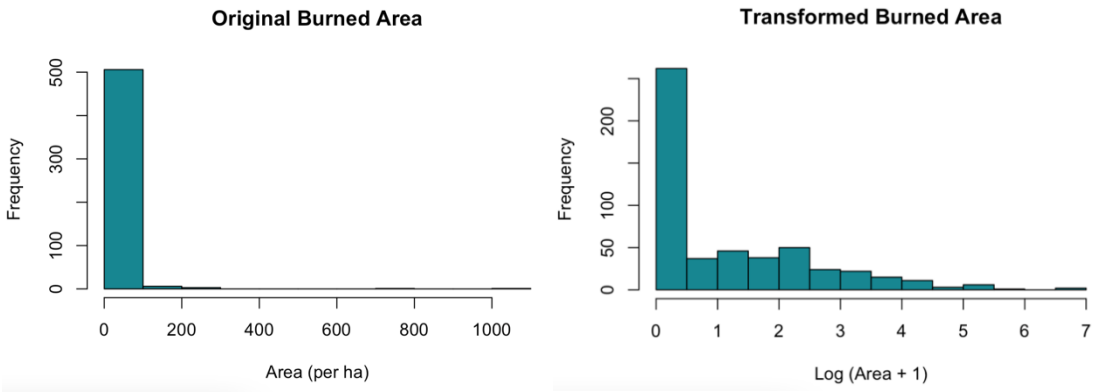
1-1 Source

The data for this study comes from the machine learning database of the University of California, Irvine (UCI) , recording northeastern Portugal Tra´s - os -Montes region Montesinho There are a total of 517 data related to each fire event in the natural park between January 2000 and December 2003 . The park is located in a Mediterranean climate zone with rich ecological diversity , and the average annual temperature ranges from 8 to 12°C .

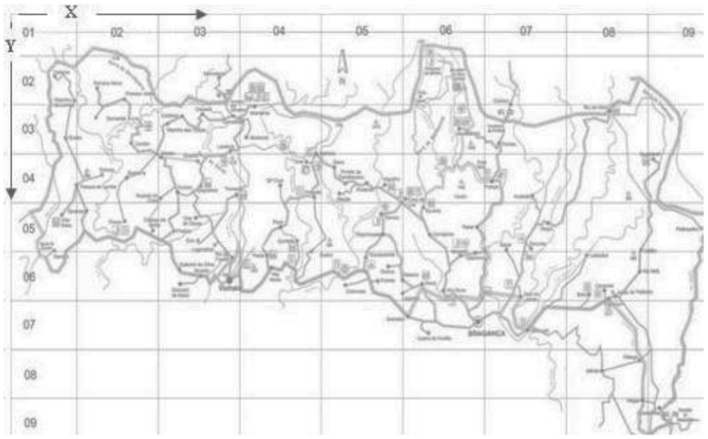
1-2 Variable description

1. Dependent variable

   Area : The burned area when the fire occurred, in hectares (ha) . a rea = 0 means that the fire burning area is less than 0.01 hectares (100 square meters). However, this data set has a total of 247 samples with a value of 0 , which is a highly right-skewed data type. In order to reduce the skewness and improve symmetry, we perform variable transformation on this field and take the natural logarithm of the original data. $ln(area + 1)$, use this result to conduct subsequent regression analysis.



2. Independent variables



A. time and space

X: Montesinho The x- axis spatial coordinate within the park map , ranging from 1 to 9 .

Y: Montesinho within the park map y Axis space coordinate, ranging from 1 to 9 .

Month : The month in which the fire event occurred, ranging from January to December .

Day : The week of the fire event, ranging from Monday to Sunday.

Month and Day are both categorical variables, and subsequent analysis will use dummy variables to establish a regression model.

B. weather system detection data

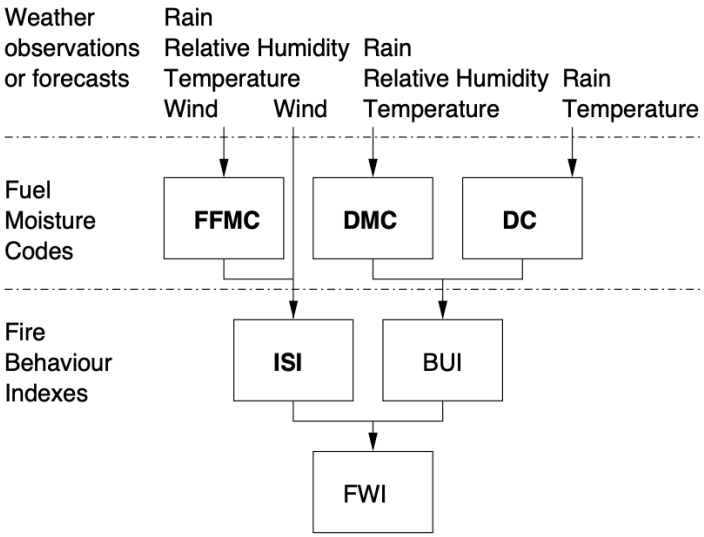
Temp : The temperature in degrees Celsius when the fire event occurred, in units of ℃.

RH : Relative humidity at the time of fire event, unit: %.

Wind : The wind speed when the fire event occurs, in units of km/h.

Rain : The amount of outdoor rain when the fire event occurred, in units of mm/m^2.

C. Forest Fire Weather Index (FWI)



It is Canada's fire danger rating system, including six parts. The first three are flammable materials index (FFMC , BMC , DC), which represent the humidity of different strata in the forest respectively. The last three are fire behavior indicators (ISI , BUI , FWI) .), the greater the value of the six indicators, the higher the fire risk level. This data set field only records FFMC , DMC , DC , and ISI . The following only briefly describes these four indicators.

FFMC (surface organic layer): represents a dry mass of 0.25 $kg/m^2$and a thickness of 1.2 cmin the forest ground cover Moisture content of litter and other solidified fine fuels such as needles, moss, leaves, lichens and other small loose debris. The minimum value is 0 (moisture content 100%), and the maximum value is 101 (moisture content 0). The larger the value, the higher the fire danger level.

DMC (shallow organic layer): represents the moisture content of organic matter in the forest ground cover with a dry mass of 5 $kg/m^2$and a thickness of 7 . cmThe value changes with the change of moisture content. The minimum value is 0 (moisture content 100%), and the maximum value has no upper limit (usually within 150).

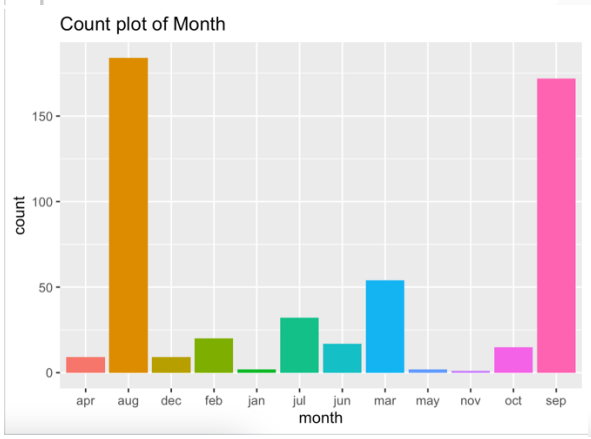
DC (deep organic layer): represents the moisture content of organic matter in the forest floor with a dry mass of 25% $kg/m^2$and a thickness of 18% cm. Calculate the impact index of long-term drought on forest combustibles. The minimum value is 0 (moisture content 100%), and the maximum value has no upper limit (usually within 1,000).

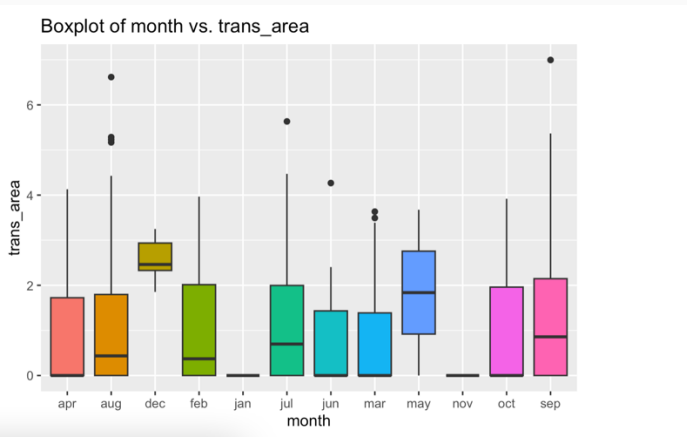
ISI : calculated from FFMC and wind speed, representing the level related to fire spread.

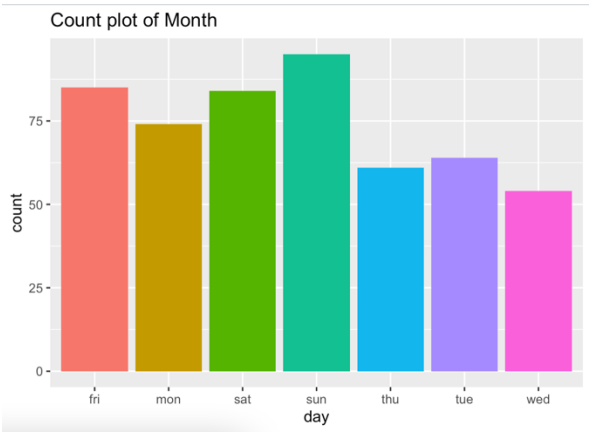
Chapter 2. Basic statistical data analysis

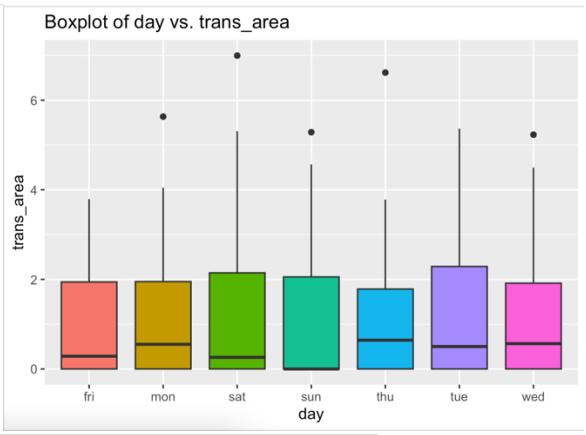
2-1 Descriptive Statistics

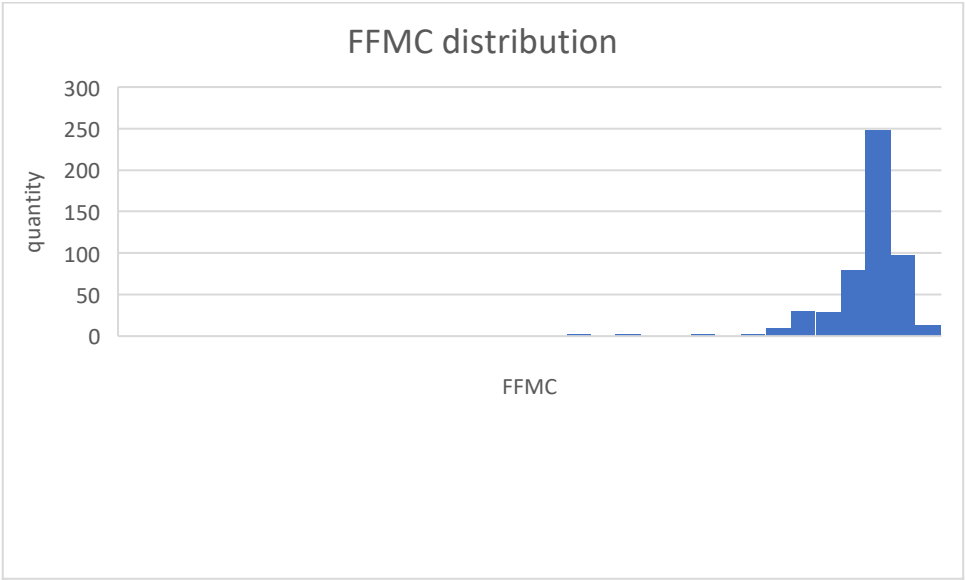
```
> summary(df_new[,c(1, 2, 5:13)])
       X               Y              FFMC             DMC              DC             ISI        
 Min.   :1.000   Min.   :2.0   Min.   :18.70   Min.   :  1.1   Min.   :  7.9   Min.   : 0.000  
 1st Qu.:3.000   1st Qu.:4.0   1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500  
 Median :4.000   Median :4.0   Median :91.60   Median :108.3   Median :664.2   Median : 8.400  
 Mean   :4.669   Mean   :4.3   Mean   :90.64   Mean   :110.9   Mean   :547.9   Mean   : 9.022  
 3rd Qu.:7.000   3rd Qu.:5.0   3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800  
 Max.   :9.000   Max.   :9.0   Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :56.100  
      temp             RH              wind             rain          trans_area      
 Min.   : 2.20   Min.   : 15.00   Min.   :0.400   Min.   :0.00000   Min.   :0.0000  
 1st Qu.:15.50   1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000   1st Qu.:0.0000  
 Median :19.30   Median : 42.00   Median :4.000   Median :0.00000   Median :0.4187  
 Mean   :18.89   Mean   : 44.29   Mean   :4.018   Mean   :0.02166   Mean   :1.1110  
 3rd Qu.:22.80   3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000   3rd Qu.:2.0242  
 Max.   :33.30   Max.   :100.00   Max.   :9.400   Max.   :6.40000   Max.   :6.9956  
```

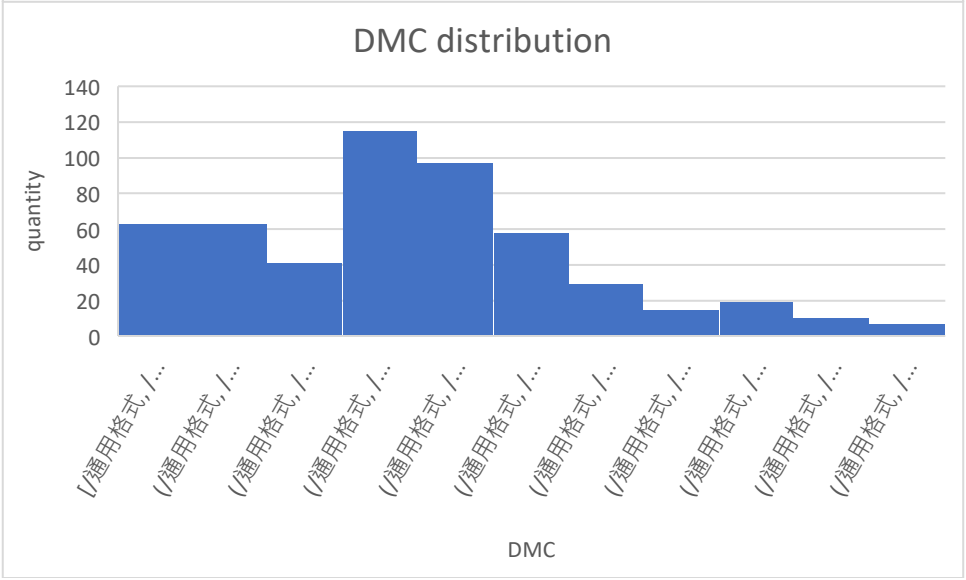


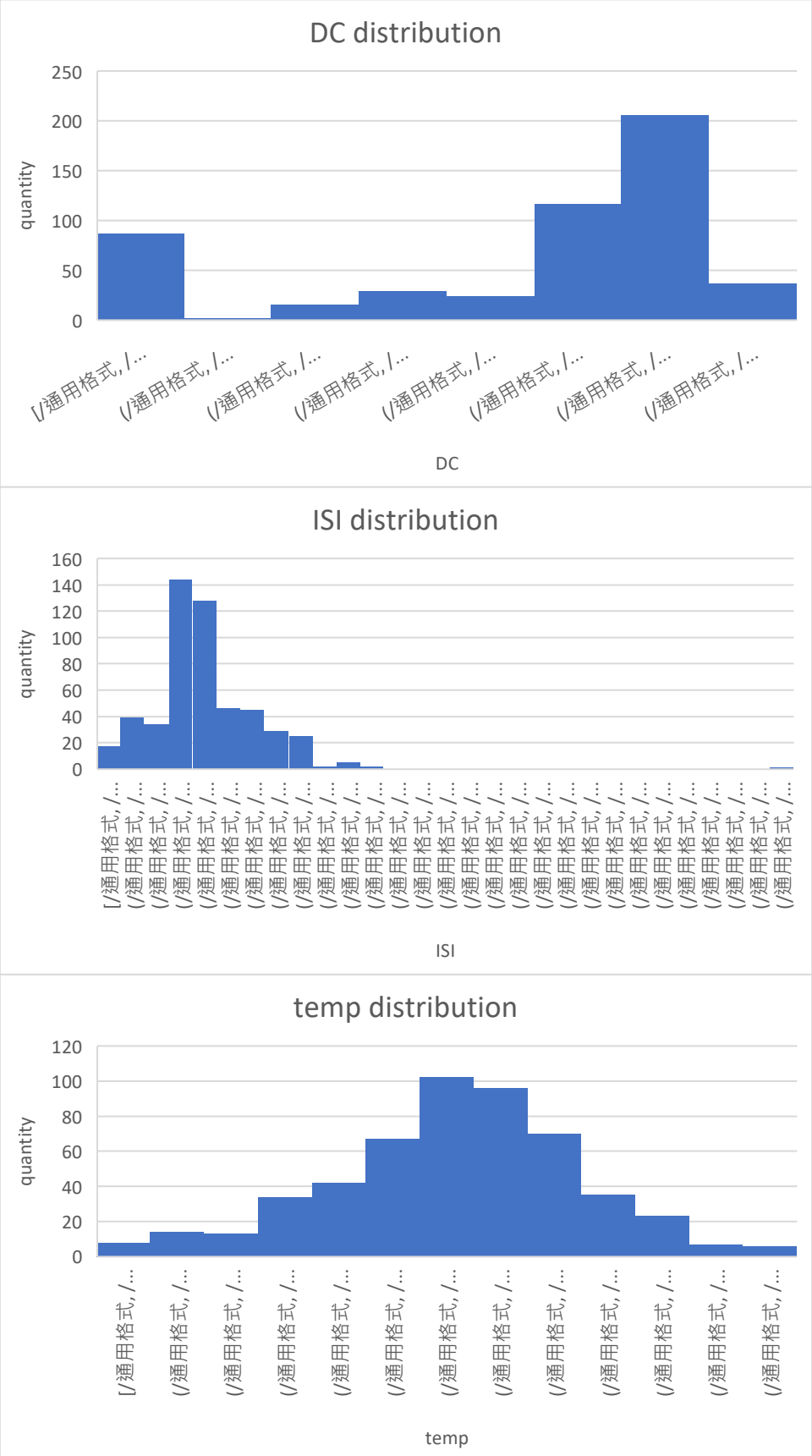


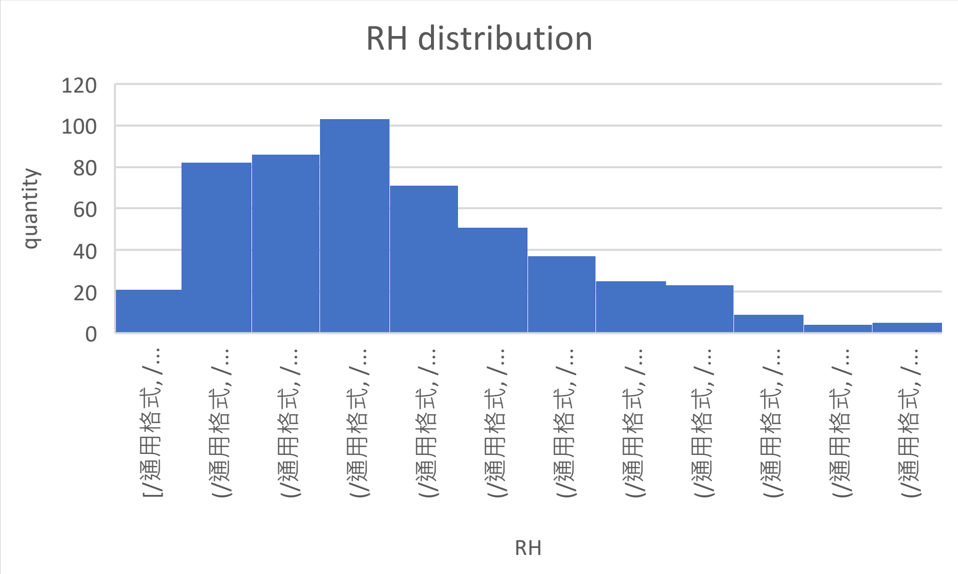

Count plot of Month
count
75
50
25
0
fri
mon
sat
sun
thu
tue
wed
day

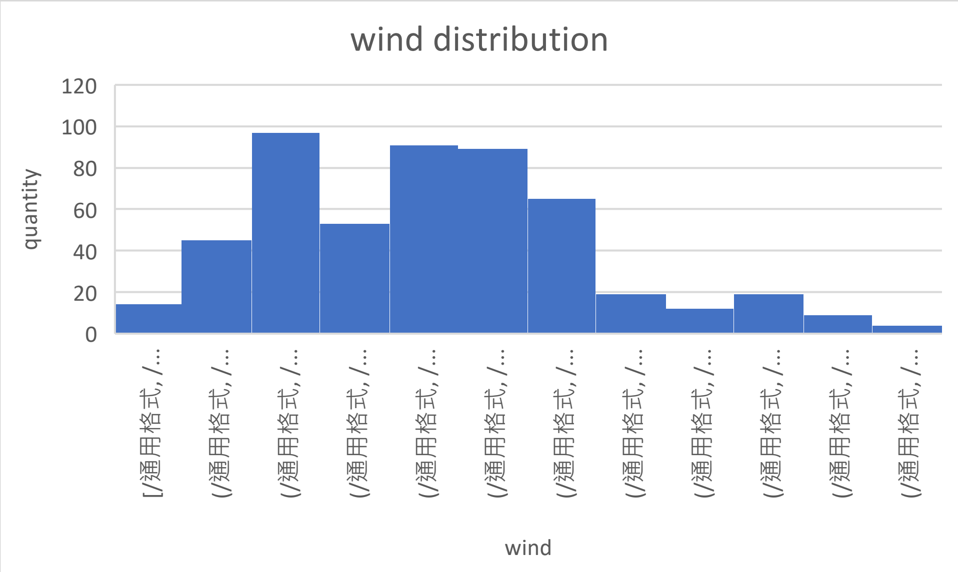

Boxplot of day vs. trans_area
trans_area
6
4
2
0
fri
mon
sat
sun
thu
tue
wed
day

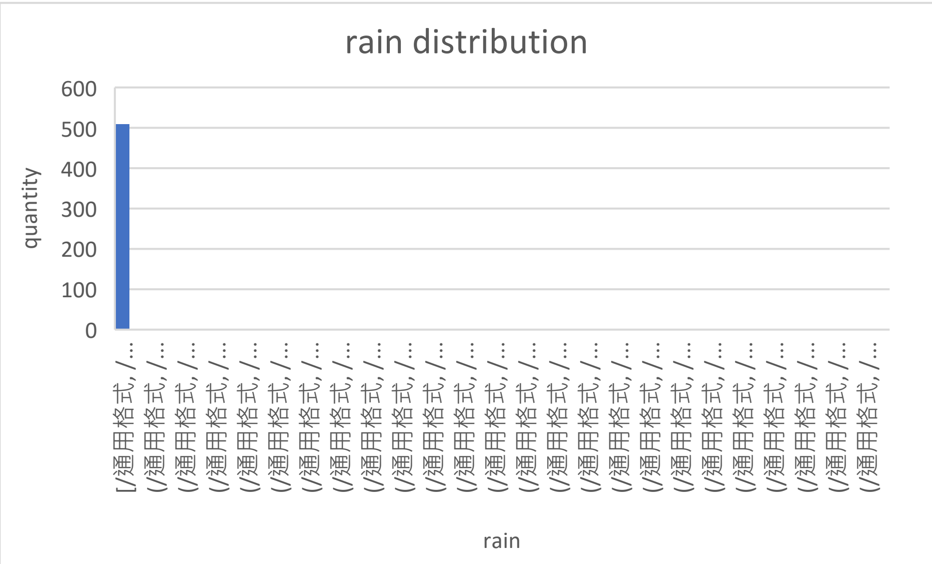

FFMC distribution
quantity
300
250
200
150
100
50
0
FFMC

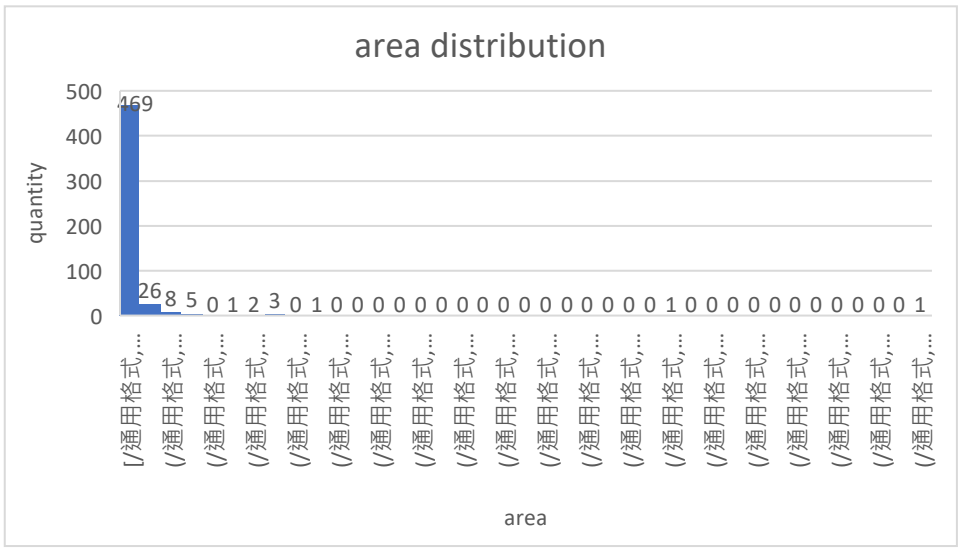

DMC distribution
quantity
140
120
100
80
60
40
20
0
DMC

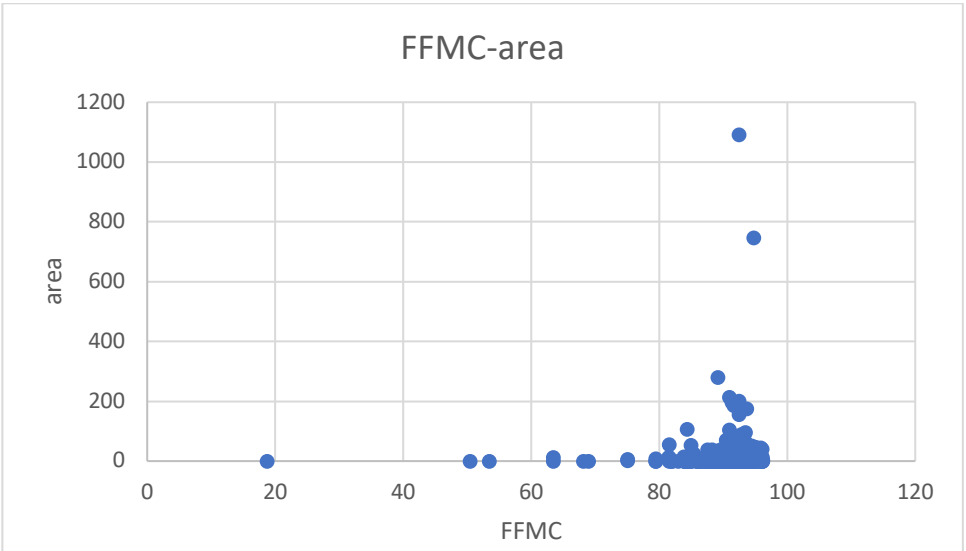

DC distribution
quantity
250
200
150
100
50
0
[/通用格式,/...
(/通用格式,/...
(/通用格式,/...
(/通用格式,/...
(/通用格式,/...
(/通用格式,/...
(/通用格式,/...
(/通用格式,/...
DC
ISI distribution
quantity
160
140
120
100
80
60
40
20
0
ISI
temp distribution
quantity
120
100
80
60
40
20
0
temp

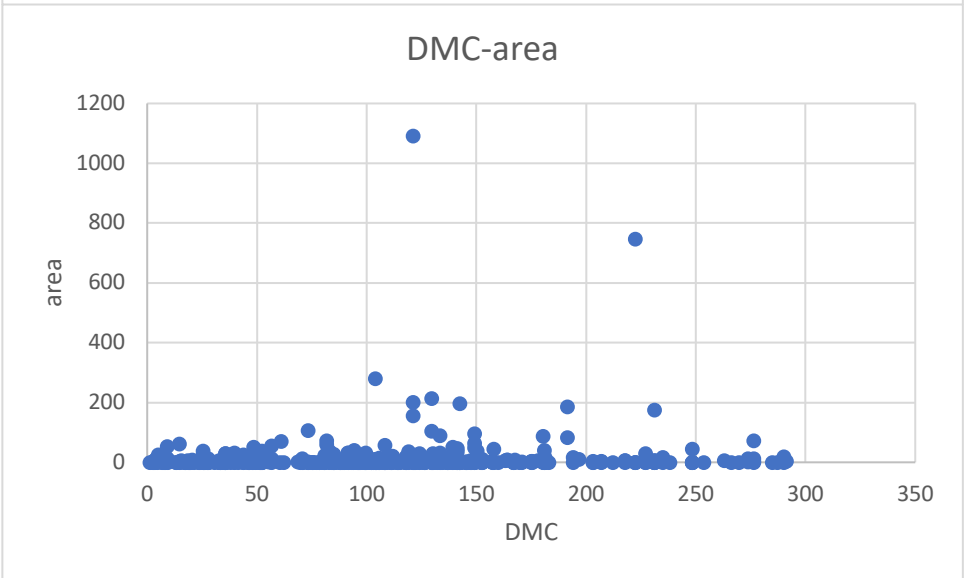

RH distribution
quantity
120
100
80
60
40
20
0
[/通用格式, /...
(/通用格式, /...
RH

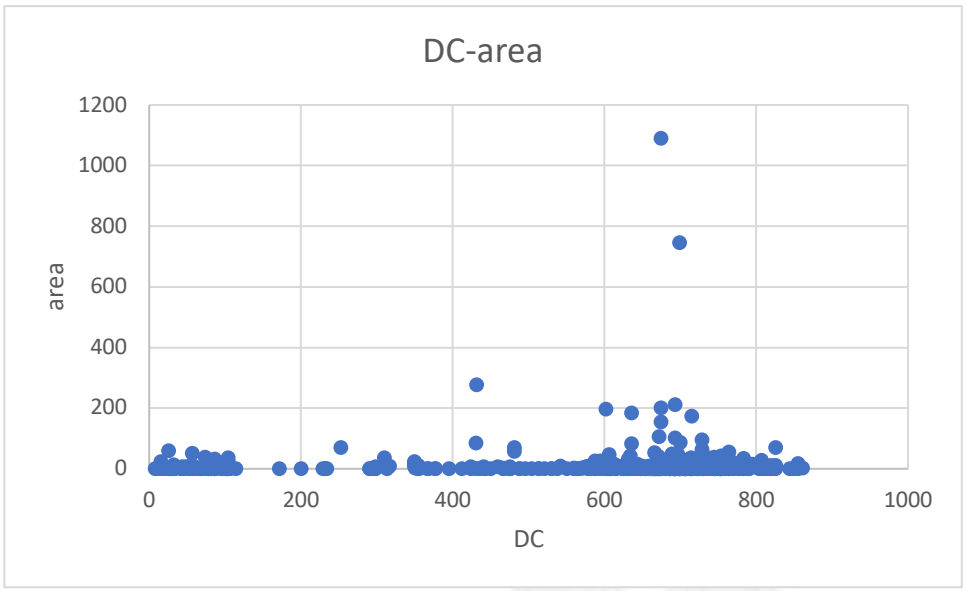

wind distribution
quantity
120
100
80
60
40
20
0
[/通用格式, /...
(/通用格式, /...
wind

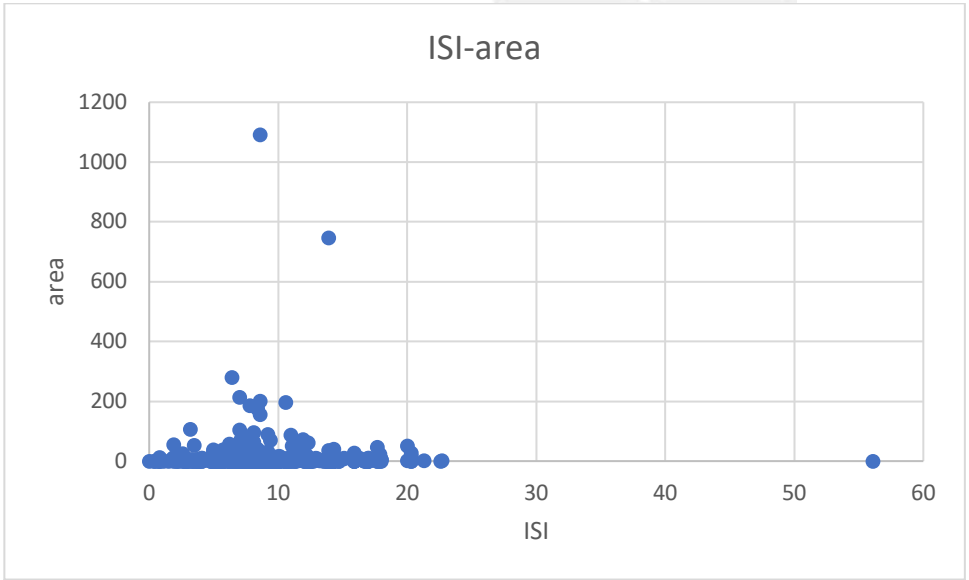

rain distribution
quantity
600
500
400
300
200
100
0
[/通用格式, /...
(/通用格式, /...
rain

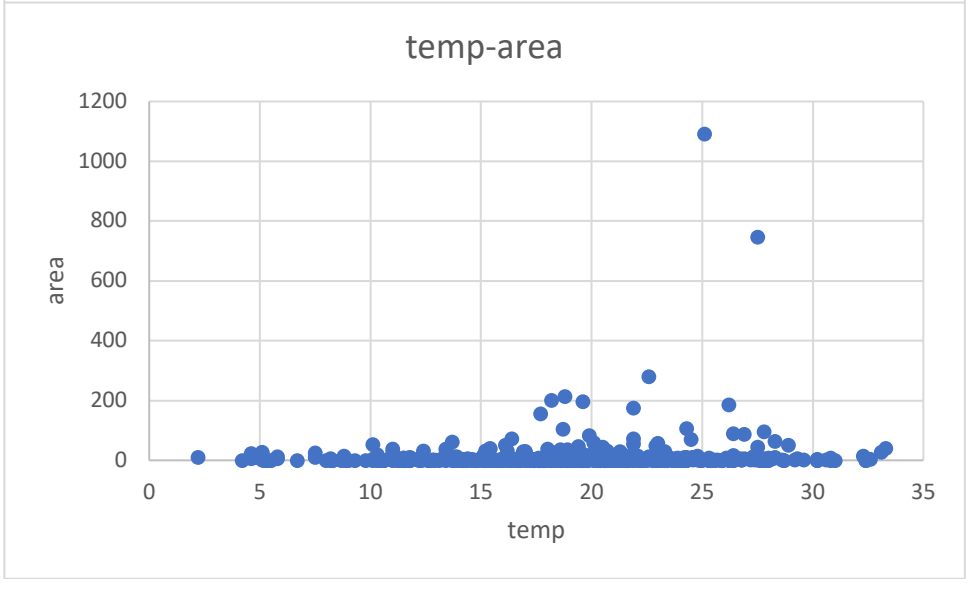

area distribution
quantity
area
500
400
300
200
100
0
469
26
8
5
0 1 2 3 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1
/通用格式,...

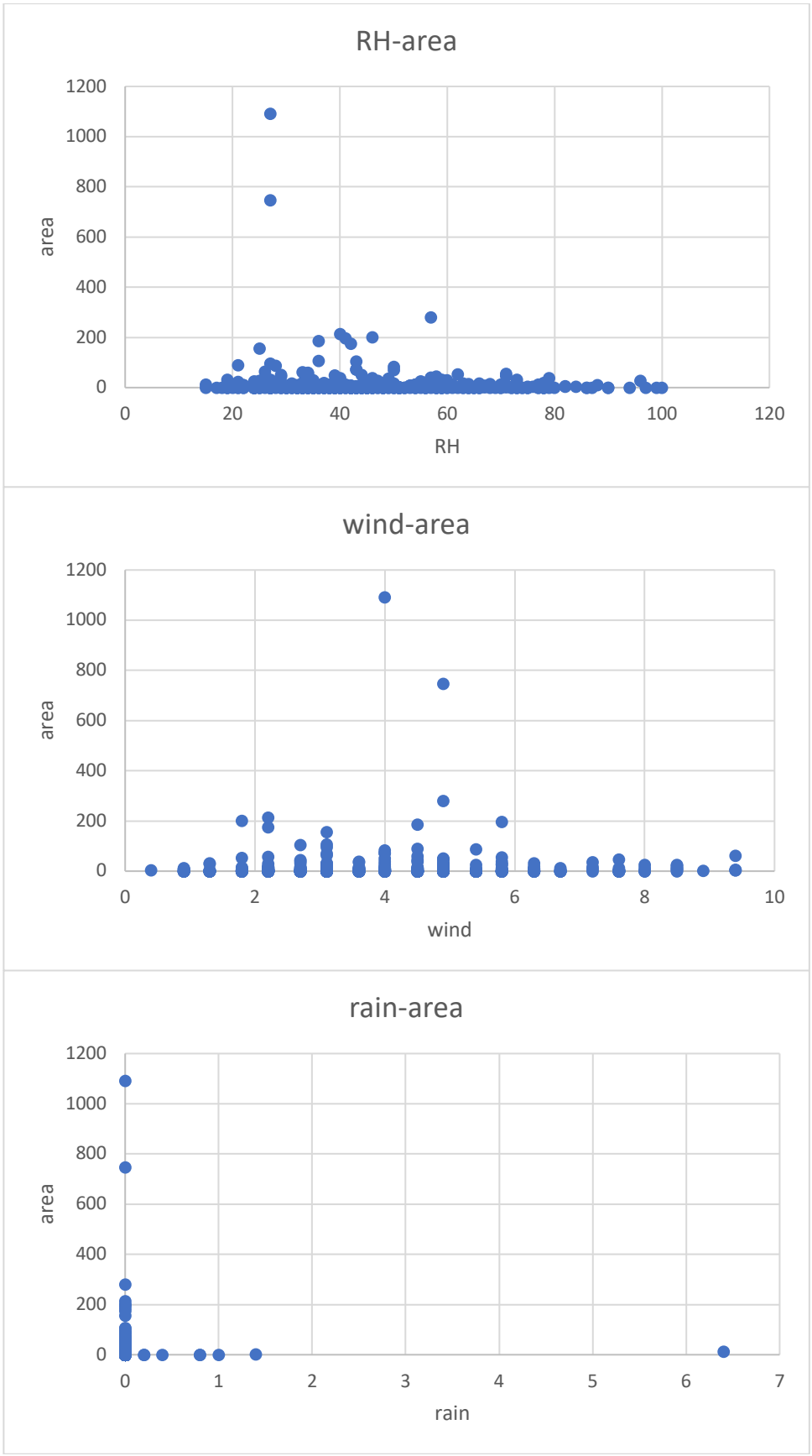

FFMC-area
area
FFMC
1200
1000
800
600
400
200
0
0
20
40
60
80
100
120

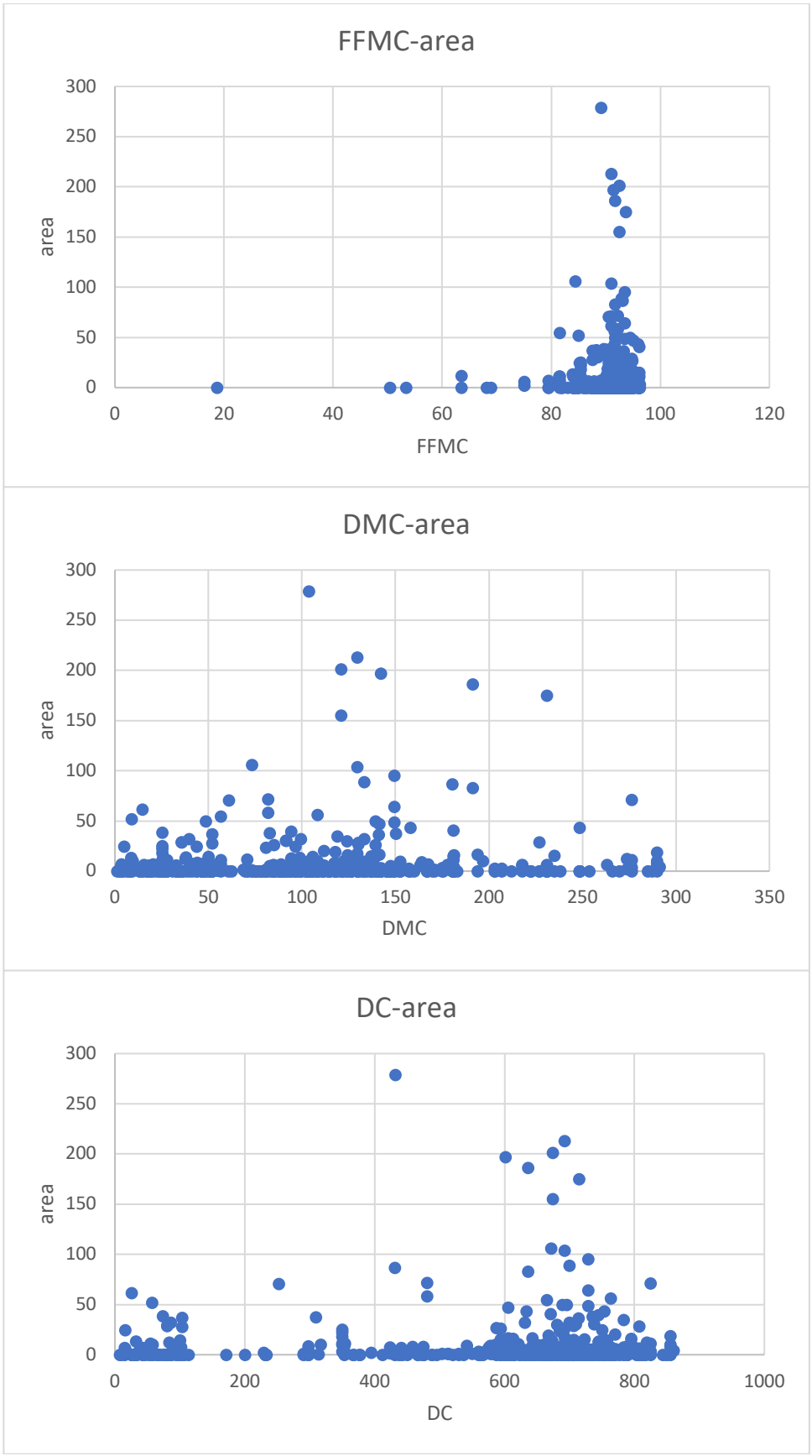

DMC-area
area
DMC
1200
1000
800
600
400
200
0
0
50
100
150
200
250
300
350

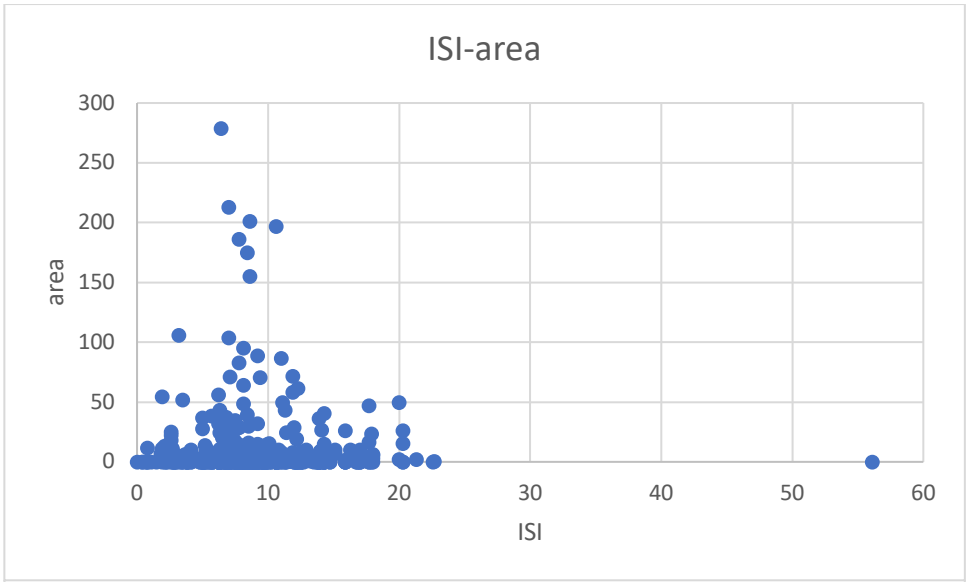

DC-area
area
DC
0
200
400
600
800
1000
1200

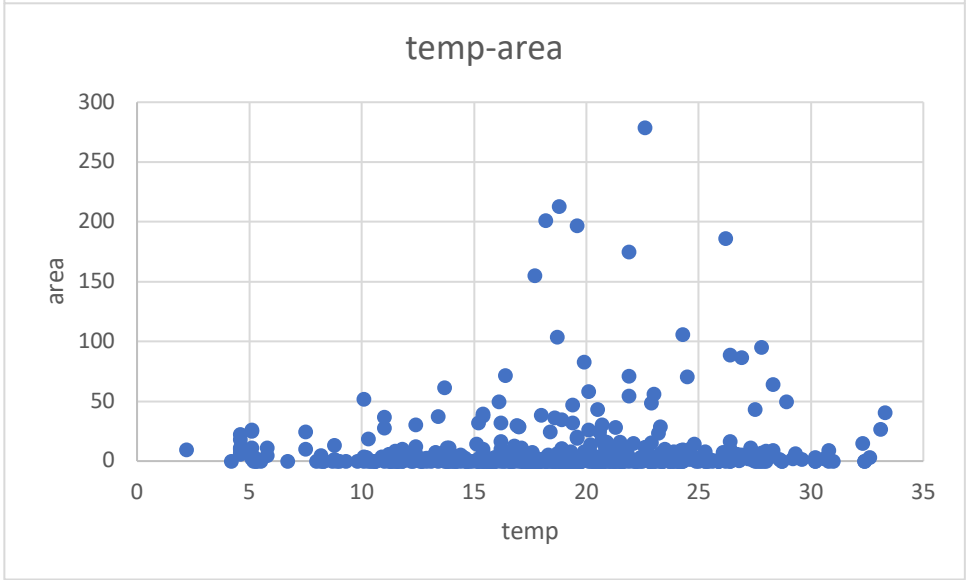

ISI-area
area
ISI
0
10
20
30
40
50
60
200
400
600
800
1000
1200

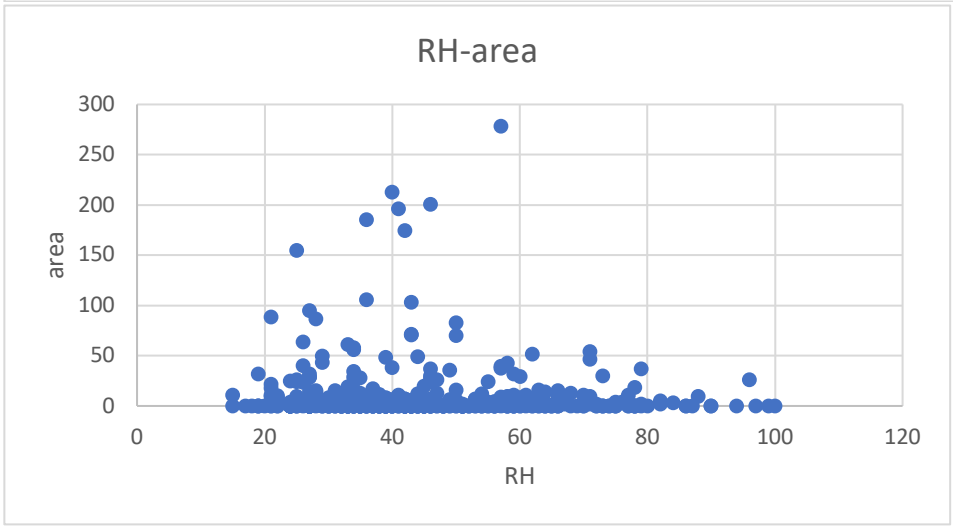

temp-area
area
temp
0
5
10
15
20
25
30
35
200
400
600
800
1000
1200

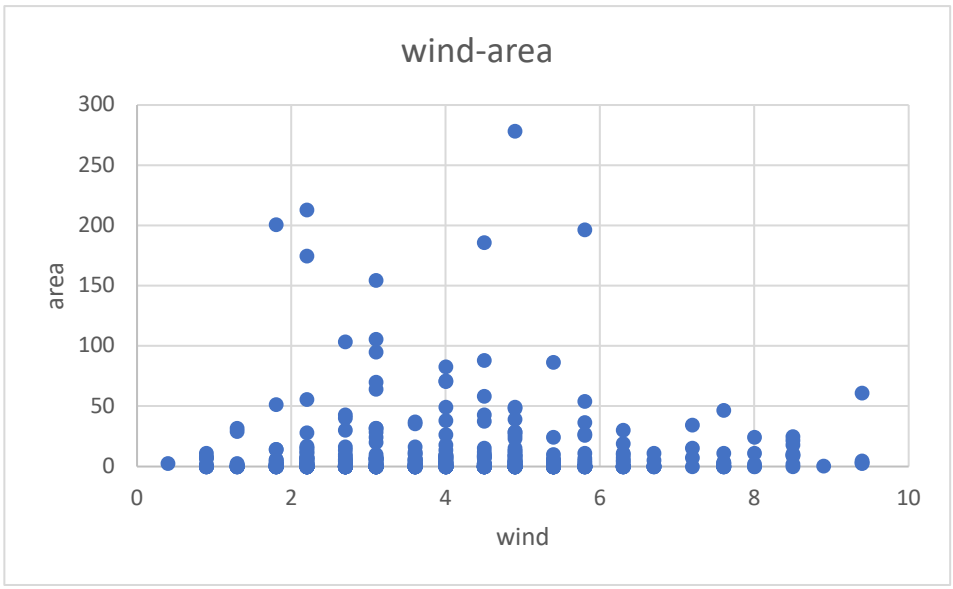

RH-area
area
RH
0
200
400
600
800
1000
1200
0
20
40
60
80
100
120
wind-area
area
wind
0
200
400
600
800
1000
1200
0
2
4
6
8
10
rain-area
area
rain
0
200
400
600
800
1000
1200
0
1
2
3
4
5
6
7

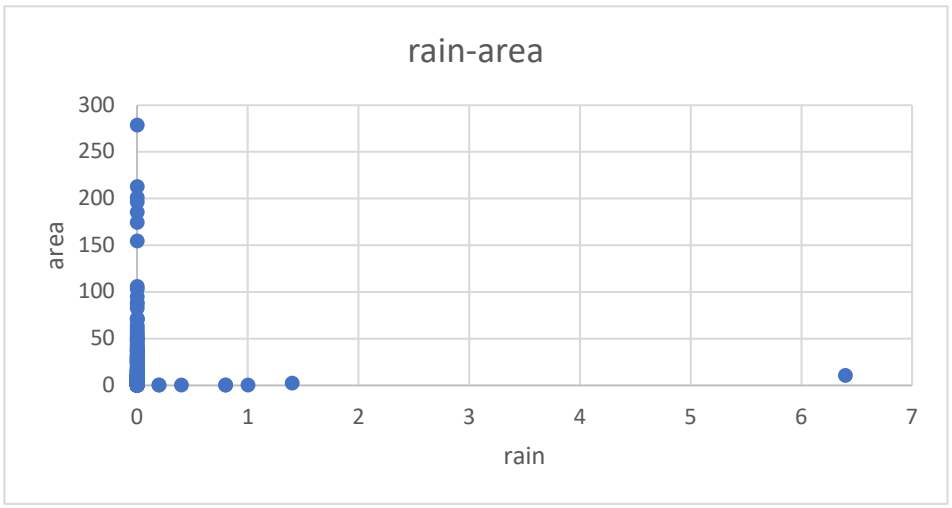

FFMC-area
area
FFMC
0
50
100
150
200
250
300
0
20
40
60
80
100
120
DMC-area
area
DMC
0
50
100
150
200
250
300
350
DC-area
area
DC
0
50
100
150
200
250
300
0
200
400
600
800
1000


ISI-area
area
ISI
0
50
100
150
200
250
300
0
10
20
30
40
50
60

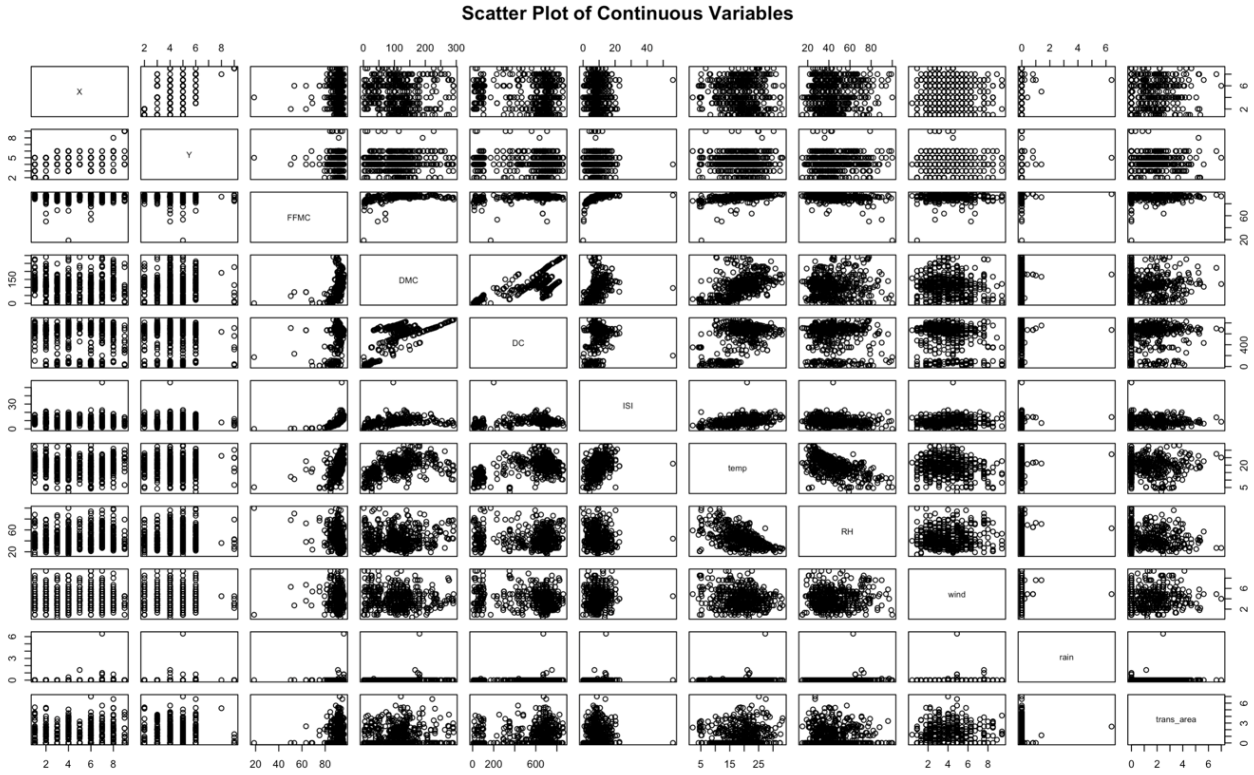

temp-area
area
temp
0
50
100
150
200
250
300
0
5
10
15
20
25
30
35


RH-area
area
RH
0
50
100
150
200
250
300
0
20
40
60
80
100
120


wind-area
300
250
200
150
100
50
0
area
0
2
4
6
8
10
wind


rain-area
300
250
200
150
100
50
0
area
0
1
2
3
4
5
6
7
rain

2-2 Correlation




```
> rcorr(as.matrix(df_new[, c(1:2, 5:13)]))
              X     Y  FFMC   DMC    DC   ISI  temp    RH  wind rain trans_area
X          1.00  0.54 -0.02 -0.05 -0.09  0.01 -0.05  0.09  0.02 0.07       0.06
Y          0.54  1.00 -0.05  0.01 -0.10 -0.02 -0.02  0.06 -0.02 0.03       0.04
FFMC      -0.02 -0.05  1.00  0.38  0.33  0.53  0.43 -0.30 -0.03 0.06       0.05
DMC       -0.05  0.01  0.38  1.00  0.68  0.31  0.47  0.07 -0.11 0.07       0.07
DC        -0.09 -0.10  0.33  0.68  1.00  0.23  0.50 -0.04 -0.20 0.04       0.07
ISI        0.01 -0.02  0.53  0.31  0.23  1.00  0.39 -0.13  0.11 0.07      -0.01
temp      -0.05 -0.02  0.43  0.47  0.50  0.39  1.00 -0.53 -0.23 0.07       0.05
RH         0.09  0.06 -0.30  0.07 -0.04 -0.13 -0.53  1.00  0.07 0.10      -0.05
wind       0.02 -0.02 -0.03 -0.11 -0.20  0.11 -0.23  0.07  1.00 0.06       0.07
rain       0.07  0.03  0.06  0.07  0.04  0.07  0.07  0.10  0.06 1.00       0.02
trans_area 0.06  0.04  0.05  0.07  0.07 -0.01  0.05 -0.05  0.07 0.02       1.00

n= 517


P
           X      Y      FFMC   DMC    DC     ISI    temp   RH     wind   rain   trans_area
X                 0.0000 0.6332 0.2722 0.0509 0.8880 0.2447 0.0528 0.6698 0.1376 0.1593
Y          0.0000        0.2933 0.8599 0.0214 0.5785 0.5845 0.1577 0.6445 0.4508 0.3782
FFMC       0.6332 0.2933        0.0000 0.0000 0.0000 0.0000 0.0000 0.5181 0.1980 0.2882
DMC        0.2722 0.8599 0.0000        0.0000 0.0000 0.0000 0.0937 0.0166 0.0894 0.1273
DC         0.0509 0.0214 0.0000 0.0000        0.0000 0.0000 0.3738 0.0000 0.4158 0.1318
ISI        0.8880 0.5785 0.0000 0.0000 0.0000        0.0000 0.0025 0.0151 0.1244 0.8144
temp       0.2447 0.5845 0.0000 0.0000 0.0000 0.0000        0.0000 0.0000 0.1145 0.2247
RH         0.0528 0.1577 0.0000 0.0937 0.3738 0.0025 0.0000        0.1150 0.0233 0.2232
wind       0.6698 0.6445 0.5181 0.0166 0.0000 0.0151 0.0000 0.1150        0.1652 0.1283
rain       0.1376 0.4508 0.1980 0.0894 0.4158 0.1244 0.1145 0.0233 0.1652        0.5969
trans_area 0.1593 0.3782 0.2882 0.1273 0.1318 0.8144 0.2247 0.2232 0.1283 0.5969
```

Chapter 3. Original Model Verification

3-1 _ Build a regression model

First, a first-order initial model is established and all explanatory variables are included in the model. Assume that it consists of month (month), week (day), surface organic layer (FFMC), shallow organic layer (DMC) , deep organic layer (DC), fire speed level (ISI), temperature (temp), relative humidity (RH)), wind speed (wind), and rainfall (rain) to predict the fire burned area. Month (month) and week (day) are categorical variables, which are processed using dummy variables. Taking April (Month_apr) and Friday (Day_fri) as the base, the regression relationship established is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = \widehat{\beta_0} + \widehat{\beta_1}X_i + \widehat{\beta_2}Y_i + \widehat{\beta_3}month_{jan_i} + \cdots + \widehat{\beta_{13}}month_{dec_i} + \widehat{\beta_{14}}day_{mon_i} + \cdots + \widehat{\beta_{19}}day_{sun_i} + \widehat{\beta_{20}}FFMC_i + \widehat{\beta_{21}}DMC_i + \widehat{\beta_{22}}DC_i + \widehat{\beta_{23}}ISI_i + \widehat{\beta_{24}}Temp_i + \widehat{\beta_{25}}RH_i + \widehat{\beta_{26}}Wind_i + \widehat{\beta_{27}}Rain_i$, $\boldsymbol{i = 1, \ldots, 517}$

3-2 Parameter estimation

Parameter Estimation						
variable	Label	DF	parameter estimates	standard error	t value	Pr>|t|
Intercept	Intercept		-0.5705460	1.6566550	-0.344	0.73070
X	space coordinates	1	0.0524204	0.0324114	1.617	0.10645
Y	space coordinates	1	-0.1886078	0.0609946	-0.303	0.76216
Month_jan	January	1	-0.3163816	1.2205397	-0.259	0.79558

Month_feb	February	1	0.1886078	0.5614767	0.336	0.73708
Month_mar	March	1	-0.3416243	0.5066518	-0.674	0.50045
Month_may	May	1	0.7175267	1.1017352	0.651	0.51518
Month_jun	June	1	-0.2862231	0.6564584	-0.436	0.66302
Month_jul	July	1	0.0991694	0.7151995	0.139	0.88978
Month_aug	August	1	0.3274391	0.8241619	0.397	0.69132
Month_sep	September	1	0.9934196	0.9234562	1.076	0.28256
Month_oct	October	1	0.8232625	0.9828184	0.838	0.40263
Month_nov	November	1	-1.1031443	1.4795838	-0.746	0.45628
Month_dec	December	1	2.2050797	0.7995023	2.758	0.00603
Day_mon	Monday	1	0.1457734	0.2268038	0.643	0.52070
Day_tue	Tuesday	1	0.3222933	0.2354888	1.369	0.17175
Day_wed	Wednesday	1	0.1978808	0.2467916	0.802	0.42305
Day_thu	Thursday	1	0.0722394	0.2403369	0.301	0.76387
Day_sat	Saturday	1	0.3099153	0.2177296	1.423	0.15526
Day_sun	Sunday	1	0.2109897	0.2118118	0.996	0.31969
FFMC	There is a base layer on the surface	1	0.0074547	0.0166582	0.448	0.65471
DMC	shallow organic layer	1	0.0041790	0.0018785	2.225	0.02656
DC	deep organic layer	1	-0.0020052	0.0012706	-1.578	0.11516
ISI	Fire speed level	1	-0.0147970	0.0179825	-0.823	0.41099
temp	temperature	1	0.0360374	0.0223054	1.616	0.10682

RH	Relative humidity	1	0.0006673	0.0062416	0.107	0.91490
Wind	wind speed	1	0.0603127	0.0384782	1.567	0.11766
Rain	rainfall	1	0.0309440	0.2147931	0.144	0.88551

From the above parameter estimation table, we can know the estimated value of each parameter, so we can know that the Model_1 model is as follows:

Model_all : $\boldsymbol{ln}(\widehat{area+1}) = -0.5705460 + 0.0524204X_i - 0.1886078Y_i -$

$0.3163816month_{jan_i} + \cdots + 2.2050797month_{dec_i} +$

$0.1457734day_{mon_i} + \cdots + 0.2109897day_{sun_i} + 0.0074547FFMC_i +$

$0.0041790DMC_i - 0.0020052DC_i - 0.0147970ISI_i +$

$0.0360374Temp_i + 0.0006673RH_i + 0.0603127Wind_i +$

$0.0309440Rain_i$ ，$\boldsymbol{i = 1, 2, \ldots, 517}$

3-3 Model fitness test

The ANOVA table is as follows:

ANOVA table				
source	degrees of freedom	SSE	MSE	Pr (>F)
Reg	27	11.5636	0.4283	0.06765
Error	489	934.17	1.9104	

The statistical assumptions are:

$$H_0 : \beta_1 = \cdots = \beta_{27} = 0$$

$$H_1 : \beta_k \; is \; not \; all \; equal \; to \; 0, k = 1, \ldots, 27$$

It can be seen from the above table that because P -value = 0.06765 $> \alpha =$0.05, it is not rejected H_0. It is inferred that the original data of the explanatory variables and the combustion area after taking the logarithm (ln(area+1)) are not linearly related, and variable conversion may be required.

3-4 Model explanation ability

Model explanation analysis table			
Multiple R-squared	0.072426	Adjusted R-squared	0.02315

Under this model, Multiple R-squared = 0.072426 , indicating that the total variation of ln(area+1) can be explained by all explanatory variables 7.2426% ; and the corrected R-squared = 0.02315 , indicating that this model only has 2.315% Explanation ability.

3-5 Parameter verification

we get the values, we start to test βwhether there is a linear correlation between each explanatory variable and the logarithm of the burning area (ln(area+1)).

（1） X and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$$

Because p-value = 0.10645 > $\alpha = 0.05$, it is not rejected . It is inferred that there is no linear relationship between $H_0$the X- axis of the spatial coordinates and the logarithmic combustion area.

（2） Y and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$$

Because p-value = 0.76216 > $\alpha = 0.05$, it is not rejected . It is inferred that there is no linear relationship between $H_0$the Y- axis of the spatial coordinates and the logarithmic combustion area.

（3） Month (month) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_k = 0 ; H_1 : \beta_k \neq 0 , \boldsymbol{k = 3, \dots, 13}$$

Except β_{13} for p-value = 0.0.00603 < $\alpha = 0.05$, the parameter estimated p-values of other months are all greater than the significance level 0.05 , reject H_0, it is speculated that there is a linear relationship between the spatial coordinates in December and the burned area after taking the logarithm. In other months, there is no linear relationship with the logarithm of the burned area.

（4） Week (day) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_k = 0 ; H_1 : \beta_k \neq 0 , \boldsymbol{k = 14, \dots, 19}$$

Because the p-values of all weeks $> \alpha = 0.05$are not rejected H_0, it is inferred that there is no linear relationship between the day of the week and the logarithm of the burned area.

（5） FFMC and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{20} = 0 ; H_1 : \beta_{20} \neq 0$$

Because p-value = 0.65471 > $\alpha = 0.05$, it is not rejected H_0, and it is inferred that there is no linear relationship between F FMC and the logarithmic combustion area.

（6） DMC and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{21} = 0 ; H_1 : \beta_{21} \neq 0$$

Because p-value = 0.02656 , $< \alpha = 0.05$it is rejected H_0, and it is inferred that there is a linear relationship between D MC and the logarithmic burning area.

（7） D C and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{22} = 0 ; H_1 : \beta_{22} \neq 0$$

Because p-value = 0.11516 > , $\alpha = 0.05$it is not rejected H_0. It is speculated that there is no linear relationship between DC and the logarithm of the burning area.

（8） I SI and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{23} = 0 ; H_1 : \beta_{23} \neq 0$$

Because p-value = 0.41099 > $\alpha = 0.05$, it is not rejected H_0, and it is inferred that there is no linear relationship between ISI and logarithmic burning area.

（9） Temperature (temp) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{24} = 0 ; H_1 : \beta_{24} \neq 0$$

Because p-value = 0.10682 > $\alpha = 0.05$, it is not rejected H_0, and it is inferred that there is no linear relationship between temperature (temp) and logarithmic combustion area.

（10） Relative humidity (RH) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{25} = 0 ; H_1 : \beta_{25} \neq 0$$

Because p-value = 0.91490 > , $\alpha = 0.05$it is not rejected H_0, and it is inferred that there is no linear relationship between relative humidity (RH) and the logarithm of the burned area.

（11） Wind speed (wind) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{26} = 0 ; H_1 : \beta_{26} \neq 0$$

Because p-value = 0.11766 > $\alpha = 0.05$, it is not rejected H_0. It is inferred that there is no linear relationship between wind speed (wind) and the logarithmic burning area.

（12） Rainfall (rain) and burned area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{27} = 0 ; H_1 : \beta_{27} \neq 0$$

Because p-value = 0.88551 > , $\alpha = 0.05$it is not rejected H_0. It is inferred that there is no linear relationship between rainfall (rain) and the logarithm of the burned area.

Chapter 4. Model Selection

4-1 Forward selection method

We used the ols package of R software to select models using the p-value selection method. It is set that each explanatory variable entered into the model must meet the set standard (P -value = 0.15) . The variables with high significance are given priority, and variables are selected one by one to enter the model until no variable meets the standard. The following are the execution results:

```
> forward_model <- ols_step_forward_p(all_model, penter = 0.15)
> forward_model
```

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	month	0.0370	0.0161	-4.3343	1819.4300	1.3872
2	temp	0.0459	0.0232	-7.0057	1816.6601	1.3821
3	X	0.0518	0.0273	-8.1596	1815.4147	1.3792
4	DMC	0.0574	0.0311	-9.0752	1814.3962	1.3765
5	DC	0.0628	0.0347	-9.9418	1813.4112	1.3739

```
> forward_model$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun
   0.463792     0.362821     2.369202     0.145480    -0.595731     0.137374    -0.349367
   monthmar     monthmay     monthnov     monthoct     monthsep         temp            X
  -0.321630     0.750746    -0.915995     0.874552     0.995536     0.032001     0.047405
        DMC           DC
   0.004234    -0.002086
```

Final Model Output

Model Summary

R	0.251	RMSE	1.374
R-Squared	0.063	Coef. Var	123.664
Adj. R-Squared	0.035	MSE	1.888
Pred R-Squared	-Inf	MAE	1.106

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	63.362	15	4.224	2.238	0.0048
Residual	945.739	501	1.888		
Total	1009.102	516			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.464	0.516		0.899	0.369	-0.550	1.478
monthaug	0.363	0.779	0.124	0.466	0.642	-1.168	1.893
monthdec	2.369	0.751	0.222	3.153	0.002	0.893	3.845
monthfeb	0.145	0.553	0.020	0.263	0.793	-0.941	1.232
monthjan	-0.596	1.084	-0.026	-0.550	0.583	-2.725	1.533
monthjul	0.137	0.676	0.024	0.203	0.839	-1.190	1.465
monthjun	-0.349	0.628	-0.045	-0.556	0.578	-1.583	0.885
monthmar	-0.322	0.497	-0.070	-0.648	0.517	-1.297	0.654
monthmay	0.751	1.076	0.033	0.698	0.486	-1.363	2.865
monthnov	-0.916	1.451	-0.029	-0.631	0.528	-3.767	1.935
monthoct	0.875	0.948	0.105	0.923	0.357	-0.988	2.737
monthsep	0.996	0.884	0.336	1.126	0.261	-0.741	2.732
temp	0.032	0.014	0.133	2.235	0.026	0.004	0.060
X	0.047	0.027	0.078	1.771	0.077	-0.005	0.100
DMC	0.004	0.002	0.194	2.400	0.017	0.001	0.008
DC	-0.002	0.001	-0.370	-1.703	0.089	-0.004	0.000

The final model selected using the forward selection method is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.464 + 0.047X_i - 0.596month_{jan_i} + \cdots + 2.369month_{dec_i} +$

$0.004DMC_i - 0.002DC_i + 0.032Temp_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

4-2 backward selection method

We used the ols package of R software to select models using the p-value selection method. Set the standard for each explanatory variable to be kicked out of the model (P -value = 0.15) . The least significant (P-value is the largest) explanatory variable will be eliminated first until all explanatory variables in the model are significant. So far . The following are the execution results:

```
> backward_model <- ols_step_backward_p(all_model, prem = 0.15)
> backward_model


                         Elimination Summary
-------------------------------------------------------------------------
        Variable                 Adj.
Step    Removed     R-Square    R-Square     C(p)         AIC        RMSE
-------------------------------------------------------------------------
   1    RH            0.0742      0.0251    -3.9886    1829.0575    1.3808
   2    rain          0.0742       0.027    -5.9599    1827.0878    1.3794
   3    day           0.0681      0.0324    -4.7330    1818.4881    1.3756
   4    Y             0.0679      0.0342    -6.6260    1816.6004    1.3743
   5    FFMC          0.0676      0.0359    -8.4952    1814.7377    1.3731
   6    ISI           0.0666      0.0367    -9.9625    1813.2967    1.3725
   7    wind          0.0628      0.0347    -9.9418    1813.4112    1.3739
-------------------------------------------------------------------------
> backward_model$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)            X     monthaug     monthdec     monthfeb     monthjan     monthjul
   0.463792     0.047405     0.362821     2.369202     0.145480    -0.595731     0.137374
   monthjun     monthmar     monthmay     monthnov     monthoct     monthsep          DMC
  -0.349367    -0.321630     0.750746    -0.915995     0.874552     0.995536     0.004234
         DC         temp
  -0.002086     0.032001
```

Final Model Output

Model Summary

R	0.251	RMSE	1.374
R-Squared	0.063	Coef. Var	123.664
Adj. R-Squared	0.035	MSE	1.888
Pred R-Squared	-Inf	MAE	1.106

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	63.362	15	4.224	2.238	0.0048
Residual	945.739	501	1.888		
Total	1009.102	516			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.464	0.516		0.899	0.369	-0.550	1.478
X	0.047	0.027	0.078	1.771	0.077	-0.005	0.100
monthaug	0.363	0.779	0.124	0.466	0.642	-1.168	1.893
monthdec	2.369	0.751	0.222	3.153	0.002	0.893	3.845
monthfeb	0.145	0.553	0.020	0.263	0.793	-0.941	1.232
monthjan	-0.596	1.084	-0.026	-0.550	0.583	-2.725	1.533
monthjul	0.137	0.676	0.024	0.203	0.839	-1.190	1.465
monthjun	-0.349	0.628	-0.045	-0.556	0.578	-1.583	0.885
monthmar	-0.322	0.497	-0.070	-0.648	0.517	-1.297	0.654
monthmay	0.751	1.076	0.033	0.698	0.486	-1.363	2.865
monthnov	-0.916	1.451	-0.029	-0.631	0.528	-3.767	1.935
monthoct	0.875	0.948	0.105	0.923	0.357	-0.988	2.737
monthsep	0.996	0.884	0.336	1.126	0.261	-0.741	2.732
DMC	0.004	0.002	0.194	2.400	0.017	0.001	0.008
DC	-0.002	0.001	-0.370	-1.703	0.089	-0.004	0.000
temp	0.032	0.014	0.133	2.235	0.026	0.004	0.060

The final model selected using the backward selection method is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area} + 1) = 0.464 + 0.047X_i - 0.596month_{jan_i} + \cdots + 2.369month_{dec_i} +$

$0.004DMC_i - 0.002DC_i + 0.032Temp_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

4-3 Stepwise regression method

We used the ols package of R software to select models using the p-value selection method. By comprehensively using the backward selection method and the forward selection method, each explanatory variable entering and exiting the model must meet the set standard (P-value = 0. 1 5). First use the forward selection method to put the significant explanatory variables outside the model into the model , then use the backward selection method to test all variables in the model, and kick out the non-significant variables from the model. Repeat this action until the model No external variables are significant or this variable has entered the model . The following are the execution results:

```
> stepwise_model <- ols_step_both_p(all_model, prent = 0.15, prem = 0.15)
> stepwise_model

                            Stepwise Selection Summary
--------------------------------------------------------------------------------------
                     Added/                   Adj.
Step    Variable     Removed     R-Square    R-Square     C(p)         AIC         RMSE
--------------------------------------------------------------------------------------
   1     month      addition       0.037       0.016    -4.3340    1819.4300    1.3872
   2      temp      addition       0.046       0.023    -7.0060    1816.6601    1.3821
   3       X        addition       0.052       0.027    -8.1600    1815.4147    1.3792
   4      DMC       addition       0.057       0.031    -9.0750    1814.3962    1.3765
   5       DC       addition       0.063       0.035    -9.9420    1813.4112    1.3739
--------------------------------------------------------------------------------------
> stepwise_model$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun
   0.463792     0.362821     2.369202     0.145480    -0.595731     0.137374    -0.349367
   monthmar     monthmay     monthnov     monthoct     monthsep         temp            X
  -0.321630     0.750746    -0.915995     0.874552     0.995536     0.032001     0.047405
        DMC           DC
   0.004234    -0.002086
```

```
Stepwise Selection: Step 1

+ month
```

Model Summary

R	0.192	RMSE	1.387
R-Squared	0.037	Coef. Var	124.854
Adj. R-Squared	0.016	MSE	1.924
Pred R-Squared	-Inf	MAE	1.118

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	37.367	11	3.397	1.765	0.0572
Residual	971.735	505	1.924		
Total	1009.102	516			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	1.089	0.462		2.356	0.019	0.181	1.998
monthaug	-0.044	0.474	-0.015	-0.093	0.926	-0.974	0.886
monthdec	1.482	0.654	0.139	2.267	0.024	0.198	2.767
monthfeb	-0.001	0.557	0.000	-0.002	0.998	-1.095	1.093
monthjan	-1.089	1.084	-0.048	-1.004	0.316	-3.220	1.041
monthjul	-0.006	0.523	-0.001	-0.011	0.991	-1.034	1.023
monthjun	-0.246	0.572	-0.031	-0.430	0.667	-1.370	0.877
monthmar	-0.317	0.499	-0.069	-0.634	0.526	-1.298	0.665
monthmay	0.749	1.084	0.033	0.690	0.490	-1.382	2.879
monthnov	-1.089	1.462	-0.034	-0.745	0.457	-3.962	1.784
monthoct	-0.172	0.585	-0.021	-0.294	0.769	-1.321	0.977
monthsep	0.185	0.474	0.062	0.391	0.696	-0.747	1.117

The final model selected using the stepwise regression method is as follows:

Model_2 :ln $(\widehat{area} + 1) = 1.089 - 1.089month_{jan_i} + \cdots + 1.482month_{dec_i}$

$$\boldsymbol{i = 1, 2, \ldots, 517}$$

4-4 Summary

Selected by the above three methods , two temporary models are produced. The results of the regression equation are as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.464 + 0.047X_i - 0.596month_{jan_i} + \cdots + 2.369month_{dec_i} + 0.004DMC_i - 0.002DC_i + 0.032Temp_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

Model_2 :ln $(\widehat{area+1}) = 1.089 - 1.089month_{jan_i} + \cdots + 1.482month_{dec_i}$

$$\boldsymbol{i = 1, 2, \ldots, 517}$$

4-5 Collinearity test

Since Model_2 contains multiple explanatory variables, it can also be seen from the correlation coefficient table that there is a high correlation between some variables. Therefore, a collinearity test was performed on this temporary model. The results are as follows:

Collinearity test table (Model_2)	
variable name	**VIF value**
X	1.049082
Month	41.720393
DMC	3.489188
DC	25.224133
Temp	1.889160

It can be known from the test results that the VIF values of the two variables month (month) and deep organic layer (DC) exceed 10 , indicating serious collinearity problems. Since the model performance after removing the month is very poor , we decided to remove the deep organic layer

(DC) with the second largest VIF value from the model to solve the collinearity problem of the model.

4-6 Conclusion

In order to solve the collinearity problem, we removed the deep organic layer (DC) variable from the model and re-carried out the model selection step.

Chapter 5. Rebuilding the model

(Procedure 1)

5-1 Rebuild regression model verification

Removing the deep organic layer (DC), a new model is built. Establish a first-order initial model and include all explanatory variables into the model. Assume that it consists of month (month), week (day), surface organic layer (FFMC), shallow organic layer (DMC) , fire speed level (ISI), temperature (temp), relative humidity (RH), wind speed (wind), Rainfall (rain) predicts the fire burned area after taking the logarithm. Month (month) and week (day) are categorical variables, which are processed using dummy variables. Taking April (Month_apr) and Friday (Day_fri) as the base, the regression relationship established is as follows:

Model_all1 : $\boldsymbol{ln}(\widehat{area+1}) = \widehat{\beta_0} + \widehat{\beta_1}X_i + \widehat{\beta_2}Y_i + \widehat{\beta_3}month_{jan_i} + \cdots + \widehat{\beta_{13}}month_{dec_i} + \widehat{\beta_{14}}day_{mon_i} + \cdots + \widehat{\beta_{19}}day_{sun_i} + \widehat{\beta_{20}}FFMC_i + \widehat{\beta_{21}}DMC_i + \widehat{\beta_{22}}ISI_i + \widehat{\beta_{23}}Temp_i + \widehat{\beta_{24}}RH_i + \widehat{\beta_{25}}Wind_i + \widehat{\beta_{26}}Rain_i$ ，$\boldsymbol{i = 1, \ldots, 517}$

5-1-1 Parameter estimation

Parameter Estimation						
variable	Label	DF	parameter estimates	standard error	t value	Pr>|t|
Intercept	Intercept		-0.935099	1.642966	-0.569	0.5695
X	space coordinates	1	0.047008	0.032278	1.456	0.1459
Y	space coordinates	1	-0.004772	0.060466	-0.079	0.9371
Month_jan	January	1	-0.361391	1.222061	-0.296	0.7676
Month_feb	February	1	0.168688	0.562188	0.300	0.7643
Month_mar	March	1	-0.383904	0.506712	-0.758	0.4490
Month_may	May	1	0.638182	1.10226	0.579	0.5629
Month_jun	June	1	-0.669571	0.610793	-1.096	0.2735
Month_jul	July	1	-0.56234	0.580382	-0.969	0.3331
Month_aug	August	1	-0.642747	0.549778	-1.169	0.2429
Month_sep	September	1	-0.211024	0.520733	-0.405	0.6855
Month_oct	October	1	-0.41102	0.596124	-0.689	0.4908
Month_nov	November	1	-1.199871	1.480561	-0.810	0.4181
Month_dec	December	1	1.599703	0.702541	2.277	0.0232
Day_mon	Monday	1	0.154318	0.227084	0.680	0.4971
Day_tue	Tuesday	1	0.340057	0.235577	1.444	0.1495
Day_wed	Wednesday	1	0.19716	0.247166	0.798	0.4254
Day_thu	Thursday	1	0.074064	0.240699	0.308	0.7584

Day_sat	Saturday	1	0.326504	0.217806	1.499	0.1345
Day_sun	Sunday	1	0.227486	0.211875	1.074	0.2835
FFMC	There is a base layer on the surface	1	0.010039	0.016603	0.605	0.5457
DMC	shallow organic layer	1	0.002348	0.00148	1.587	0.1132
ISI	Fire speed level	1	-0.011967	0.01792	-0.668	0.5046
temp	temperature	1	0.035552	0.022337	1.592	0.1121
RH	Relative humidity	1	0.001047	0.006246	0.168	0.8670
Wind	wind speed	1	0.062939	0.038501	1.635	0.1027
Rain	rainfall	1	0.015617	0.2149	0.073	0.9421

From the above parameter estimation table, we can know the estimated value of each parameter, so we can know that the Model_1 model is as follows:

Model_all1 : $\boldsymbol{ln}(\widehat{area+1}) = -0.935099 + 0.047008X_i - 0.004772Y_i -$

$0.361391month_{jan_i} + \cdots + 1.599703month_{dec_i} +$

$0.154318day_{mon_i} + \cdots + 0.227486day_{sun_i} + 0.010039FFMC_i +$

$0.002348DMC_i - 0.011967ISI_i + 0.035552Temp_i + 0.001047RH_i +$

$0.062939Wind_i + 0.015617Rain_i$ ，$\boldsymbol{i = 1, 2, \ldots, 517}$

5-1-2 Model fitness test

The ANOVA table is as follows:

ANOVA table				
source	degrees of freedom	SSE	MSE	Pr (>F)
Reg	26	70.19	2.6996	0.08849
Error	490	938.92	1.9162	

The statistical assumptions are:

$$H_0 : \beta_1 = \cdots = \beta_{26} = 0$$

$$H_1 : \beta_k \; is \; not \; all \; equal \; to \; 0, k = 1, \ldots, 26$$

It can be seen from the above table that because P -value = 0.08849 $> \alpha =$0.05, it is not rejected H_0. It is inferred that the original data of the explanatory variables and the combustion area after taking the logarithm (ln(area+1)) are not linearly related, and variable conversion may be required.

5-1-3 Model explanation ability

Model explanation analysis table			
Multiple R-squared	0.06954	Adjusted R-squared	0.02017

Under this model, Multiple R-squared = 0.06954 , which means that 6.954% of the total variation of ln(area+1) can be explained by all explanatory variables ; and the corrected R-squared = 0.02017 , which means that this model only has 2.017% Explanation ability.

5-1-4 Parameter verification

we get the values, we start to test βwhether there is a linear correlation between each explanatory variable and the logarithm of the burning area (ln(area+1)).

（1） X and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$$

Because p-value = 0.1459 > $\alpha = 0.05$, it is not rejected . It is inferred that there is no linear relationship between $H_0$the X- axis of the spatial coordinates and the logarithmic combustion area.

(2) Y and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$$

Because p-value = 0.9371 > , $\alpha = 0.05$it is not rejected . It is inferred that there is no linear relationship between $H_0$the Y- axis of the spatial coordinates and the logarithmic combustion area.

(3) Month (month) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_k = 0 ; H_1 : \beta_k \neq 0 , \boldsymbol{k = 3, \dots, 13}$$

Except $\beta_{13}$for p-value = 0.0232 , $< \alpha = 0.05$rejected H_0, the parameter estimated p-values of other months are all greater than the significance level 0.05 , rejected H_0, it is speculated that there is a linear relationship between the spatial coordinates in December and the logarithm of the burned area, and There is no linear relationship between other months and logarithmic burned area.

(4) Week (day) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_k = 0 ; H_1 : \beta_k \neq 0 , \boldsymbol{k = 14, \dots, 19}$$

Because the p-values of all weeks $> \alpha = 0.05$are not rejected H_0, it is inferred that there is no linear relationship between the day of the week and the logarithm of the burned area.

(5) FFMC and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{20} = 0 ; H_1 : \beta_{20} \neq 0$$

Because p-value = 0.5457 > , $\alpha = 0.05$it is not rejected H_0, and it is inferred that there is no linear relationship between F FMC and the logarithmic combustion area.

(6) DMC and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{21} = 0 ; H_1 : \beta_{21} \neq 0$$

Because p-value = 0.1132 , $> \alpha = 0.05$it is not rejected H_0, and it is speculated that there is no linear relationship between D MC and the logarithmic burning area.

(7) I SI and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{23} = 0 ; H_1 : \beta_{23} \neq 0$$

Because p-value = 0.5046 > , $\alpha = 0.05$it is not rejected H_0, and it is inferred that there is no linear relationship between ISI and logarithmic burning area.

(8) Temperature (temp) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{24} = 0 ; H_1 : \beta_{24} \neq 0$$

Because p-value = 0.1121 > , $\alpha = 0.05$it is not rejected H_0, and it is inferred that there is no linear relationship between temperature (temp) and logarithmic combustion area.

(9) Relative humidity (RH) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{25} = 0 ; H_1 : \beta_{25} \neq 0$$

Because p-value = 0.8670 > , $\alpha = 0.05$it is not rejected H_0, and it is inferred that there is no linear relationship between relative humidity (RH) and the logarithm of the burned area.

(10) Wind speed (wind) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{26} = 0 ; H_1 : \beta_{26} \neq 0$$

Because p-value = 0.1027 > , $\alpha = 0.05$it is not rejected H_0. It is inferred that there is no linear relationship between wind speed (wind) and logarithmic burning area.

(11) Rainfall (rain) and burned area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{27} = 0 ; H_1 : \beta_{27} \neq 0$$

Because p-value = 0.9421 > , $\alpha = 0.05$it is not rejected H_0. It is inferred that there is no linear relationship between rainfall (rain) and the logarithm of the burned area.

5-1-5 Summary

Parameter estimation tests were performed on each variable separately. The results showed that most of the variables were not rejected H_0. It is speculated that there is no linear relationship between the explanatory variables and the response variables of the original data, and it is very likely that subsequent variable conversion will be required.

5-2 Choice of remodeling

5-2-1 forward selection method

According to the instructions in Section 4-1 , we again use the same rules for variable selection. The analysis results are as follows:

```
> forward_model1
```

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	month	0.0370	0.0161	-5.8771	1819.4300	1.3872
2	temp	0.0459	0.0232	-8.5344	1816.6601	1.3821
3	X	0.0518	0.0273	-9.6787	1815.4147	1.3792
4	DMC	0.0574	0.0311	-10.5855	1814.3962	1.3765
5	wind	0.0617	0.0336	-10.8703	1814.0111	1.3747

```
> forward_model1$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun
   0.113041    -0.612499     1.605671     0.176324    -0.547777    -0.500259    -0.716123
   monthmar     monthmay     monthnov     monthoct     monthsep         temp            X
  -0.365609     0.681061    -1.050677    -0.343384    -0.200953     0.034299     0.045764
        DMC         wind
   0.002403     0.055803
```

Final Model Output

Model Summary

R	0.248	RMSE	1.375
R-Squared	0.062	Coef. Var	123.736
Adj. R-Squared	0.034	MSE	1.890
Pred R-Squared	-Inf	MAE	1.099

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	62.264	15	4.151	2.196	0.0058
Residual	946.837	501	1.890		
Total	1009.102	516			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.113	0.550		0.205	0.837	-0.968	1.194
monthaug	-0.612	0.526	-0.210	-1.165	0.245	-1.646	0.421
monthdec	1.606	0.665	0.150	2.414	0.016	0.299	2.912
monthfeb	0.176	0.555	0.024	0.318	0.751	-0.913	1.266
monthjan	-0.548	1.088	-0.024	-0.504	0.615	-2.684	1.589
monthjul	-0.500	0.554	-0.086	-0.903	0.367	-1.589	0.588
monthjun	-0.716	0.589	-0.091	-1.215	0.225	-1.874	0.442
monthmar	-0.366	0.497	-0.080	-0.736	0.462	-1.342	0.611
monthmay	0.681	1.076	0.030	0.633	0.527	-1.432	2.794
monthnov	-1.051	1.449	-0.033	-0.725	0.469	-3.898	1.797
monthoct	-0.343	0.586	-0.041	-0.586	0.558	-1.495	0.808
monthsep	-0.201	0.506	-0.068	-0.397	0.691	-1.195	0.793
temp	0.034	0.014	0.142	2.374	0.018	0.006	0.063
X	0.046	0.027	0.076	1.709	0.088	-0.007	0.098
DMC	0.002	0.001	0.110	1.719	0.086	0.000	0.005
wind	0.056	0.037	0.071	1.522	0.129	-0.016	0.128

The final model selected using the forward selection method is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.113 + 0.046X_i - 0.548month_{jan_i} + \cdots + 1.606month_{dec_i} +$

$0.002DMC_i + 0.034Temp_i + 0.056Wind_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

5-2-2 Backward selection method

According to the instructions in Section 4-2 , we again use the same rules for variable selection. The analysis results are as follows:

```
> backward_model1
```

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	rain	0.0695	0.0222	-4.9947	1829.6777	1.3829
2	Y	0.0695	0.0241	-6.9884	1827.6844	1.3815
3	RH	0.0695	0.026	-8.9553	1825.7193	1.3801
4	day	0.0626	0.0307	-7.3445	1817.5148	1.3768
5	FFMC	0.0621	0.0321	-9.1012	1815.7695	1.3758
6	ISI	0.0617	0.0336	-10.8703	1814.0111	1.3747

```
> backward_model1$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)            X     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun     monthmar
   0.113041     0.045764    -0.612499     1.605671     0.176324    -0.547777    -0.500259    -0.716123    -0.365609
   monthmay     monthnov     monthoct     monthsep          DMC         temp         wind
   0.681061    -1.050677    -0.343384    -0.200953     0.002403     0.034299     0.055803
```

Final Model Output

Model Summary

R	0.248	RMSE	1.375
R-Squared	0.062	Coef. Var	123.736
Adj. R-Squared	0.034	MSE	1.890
Pred R-Squared	-Inf	MAE	1.099

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	62.264	15	4.151	2.196	0.0058
Residual	946.837	501	1.890		
Total	1009.102	516			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.113	0.550		0.205	0.837	-0.968	1.194
X	0.046	0.027	0.076	1.709	0.088	-0.007	0.098
monthaug	-0.612	0.526	-0.210	-1.165	0.245	-1.646	0.421
monthdec	1.606	0.665	0.150	2.414	0.016	0.299	2.912
monthfeb	0.176	0.555	0.024	0.318	0.751	-0.913	1.266
monthjan	-0.548	1.088	-0.024	-0.504	0.615	-2.684	1.589
monthjul	-0.500	0.554	-0.086	-0.903	0.367	-1.589	0.588
monthjun	-0.716	0.589	-0.091	-1.215	0.225	-1.874	0.442
monthmar	-0.366	0.497	-0.080	-0.736	0.462	-1.342	0.611
monthmay	0.681	1.076	0.030	0.633	0.527	-1.432	2.794
monthnov	-1.051	1.449	-0.033	-0.725	0.469	-3.898	1.797
monthoct	-0.343	0.586	-0.041	-0.586	0.558	-1.495	0.808
monthsep	-0.201	0.506	-0.068	-0.397	0.691	-1.195	0.793
DMC	0.002	0.001	0.110	1.719	0.086	0.000	0.005
temp	0.034	0.014	0.142	2.374	0.018	0.006	0.063
wind	0.056	0.037	0.071	1.522	0.129	-0.016	0.128

The final model selected using the backward selection method is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.113 + 0.046X_i - 0.548month_{jan_i} + \cdots + 1.606month_{dec_i} +$

$0.002DMC_i + 0.034Temp_i + 0.056Wind_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

5-2-3 Stepwise regression method

According to the instructions in Section 4-3 , we again use the same rules for variable selection. The analysis results are as follows:

```
> stepwise_model1

                          Stepwise Selection Summary
-----------------------------------------------------------------------------------------
                    Added/                   Adj.
Step    Variable    Removed     R-Square    R-Square      C(p)         AIC         RMSE
-----------------------------------------------------------------------------------------
   1     month      addition       0.037       0.016     -5.8770    1819.4300    1.3872
   2      temp      addition       0.046       0.023     -8.5340    1816.6601    1.3821
   3       X        addition       0.052       0.027     -9.6790    1815.4147    1.3792
   4      DMC       addition       0.057       0.031    -10.5850    1814.3962    1.3765
-----------------------------------------------------------------------------------------
> stepwise_model1$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun     monthmar     monthmay
   0.406520    -0.616656     1.750619     0.118687    -0.715377    -0.523007    -0.721461    -0.345455     0.676678
   monthnov     monthoct     monthsep         temp            X          DMC
  -1.060774    -0.396192    -0.240479     0.031431     0.046028     0.002401
```

Final Model Output

Model Summary

R	0.240	RMSE	1.377
R-Squared	0.057	Coef. Var	123.898
Adj. R-Squared	0.031	MSE	1.895
Pred R-Squared	-Inf	MAE	1.103

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	57.886	14	4.135	2.182	0.0076
Residual	951.216	502	1.895		
Total	1009.102	516			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.407	0.516		0.788	0.431	-0.607	1.420
monthaug	-0.617	0.526	-0.211	-1.171	0.242	-1.651	0.418
monthdec	1.751	0.659	0.164	2.656	0.008	0.456	3.045
monthfeb	0.119	0.554	0.016	0.214	0.830	-0.970	1.207
monthjan	-0.715	1.083	-0.032	-0.660	0.509	-2.844	1.413
monthjul	-0.523	0.555	-0.090	-0.943	0.346	-1.612	0.566
monthjun	-0.721	0.590	-0.092	-1.223	0.222	-1.881	0.438
monthmar	-0.345	0.497	-0.076	-0.695	0.488	-1.323	0.632
monthmay	0.677	1.077	0.030	0.628	0.530	-1.439	2.793
monthnov	-1.061	1.451	-0.033	-0.731	0.465	-3.912	1.790
monthoct	-0.396	0.586	-0.048	-0.676	0.499	-1.547	0.755
monthsep	-0.240	0.506	-0.081	-0.475	0.635	-1.234	0.753
temp	0.031	0.014	0.131	2.192	0.029	0.003	0.060
X	0.046	0.027	0.076	1.717	0.087	-0.007	0.099
DMC	0.002	0.001	0.110	1.715	0.087	0.000	0.005

The final model selected using the stepwise regression method is as follows:

Model_2 : $\boldsymbol{ln}(\widehat{area+1}) = 0.407 + 0.046X_i - 0.715month_{jan_i} + \cdots + 1.751month_{dec_i} +$

$0.002DMC_i + 0.031Temp_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

5-2-4 Summary

Combining the models selected by the above three methods , two temporary models are produced. The results of the regression equation are as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.113 + 0.046X_i - 0.548month_{jan_i} + \cdots + 1.606month_{dec_i} + 0.002DMC_i + 0.034Temp_i + 0.056Wind_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

Model_2 : $\boldsymbol{ln}(\widehat{area+1}) = 0.407 + 0.046X_i - 0.715month_{jan_i} + \cdots + 1.751month_{dec_i} + 0.002DMC_i + 0.031Temp_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

5-3 Collinearity test again

Model_2		Model_1	
variable name	VIF value	variable name	VIF value
Month	3.356483	Month	3.750094
Temp	1.888125	Temp	1.920810
X	1.048126	X	1.048169
DMC	2.190053	DMC	2.190057
		wind	1.178101

As can be seen from the table above, after excluding the deep organic layer (DC), the VIF values of the explanatory variables in the two temporary models did not exceed 10 , which solved the problem of model collinearity.

5-4 Rebuild model residual analysis

General linear regression analysis has three major assumptions : the residuals are normally distributed and the expected value is 0 , the number of variations is σ^2, and the residuals are

independent of each other. Therefore, after selecting a model, it is necessary to conduct normality test, variation homogeneity test and independence test. Only if the three tests pass, the model will be more convincing . The models to be tested are as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.113 + 0.046X_i - 0.548month_{jan_i} + \cdots + 1.606month_{dec_i} + 0.002DMC_i + 0.034Temp_i + 0.056Wind_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

Model_2 : $\boldsymbol{ln}(\widehat{area+1}) = 0.407 + 0.046X_i - 0.715month_{jan_i} + \cdots + 1.751month_{dec_i} + 0.002DMC_i + 0.031Temp_i$ ， $\boldsymbol{i = 1, 2, \ldots, 517}$

5-4-1 Normality test

In the R software, we use shapiro.test to test the normality of the residuals. The assumptions and results are as follows:

$$H_0 : Residuals\ are\ normally\ distributed.$$

$$H_1 : Residuals\ are\ not\ normally\ distributed.$$

Shapiro-Walk Normality Test (Model_2)			
W	0.87861	**P-value**	< 2.2e-16
Shapiro-Walk Normality Test (Model_1)			
W	0.88162	**P-value**	< 2.2e-16

The p-value $< \alpha = 0.05$ ，of the two temporary models is rejected < 2.2e-16 H_0, which means that the residuals do not meet the assumption of normal distribution.

5-4-2 Independence Check

In the R software, we use dubinWatsonTest to perform the independence test of the residuals. The assumptions and results are as follows:

$$H_0 : Residuals\ are\ mutually\ independent.$$

$$H_1 : Residuals\ are\ not\ mutually\ independent.$$

Durbin Watson Test (Model_2)			
Lag	**Autocorrelation**	**DW Statistics**	**p-value**
1	0.5162079	0.9670634	0
Durbin Watson Test (Model_1)			
Lag	**Autocorrelation**	**DW Statistics**	**p-value**
1	0.5188712	0.9616054	0

As can be seen from the table above, the p-value of the two temporary models approaches $0 < \alpha = 0.05$, and is rejected H_0, which means that the residuals do not meet the assumption of independence.

5-4-3 Homogeneity test

In the R software, we use ncv Test to perform the homogeneity test of the residuals. The assumptions and results are as follows:

$$H_0 : Homoscedasticity\ of\ Variance$$

$$H_1 : Heteroscedasticity\ of\ Variance$$

NON-CONSTANT VARIANCE SCORE TEST (Model_2)					
Chisquare	8.501109	**DF**	1	**p-value**	0.0035493
NON-CONSTANT VARIANCE SCORE TEST (Model_1)					
Chisquare	9.309687	**DF**	1	**p-value**	0.0022795

As can be seen from the table above, the p-values of the two temporary models are 0.0035493 and 0.0022795 respectively , both of which $< \alpha = 0.05$ ，are rejected H_0, indicating that the variables do not meet the assumption of homogeneity.

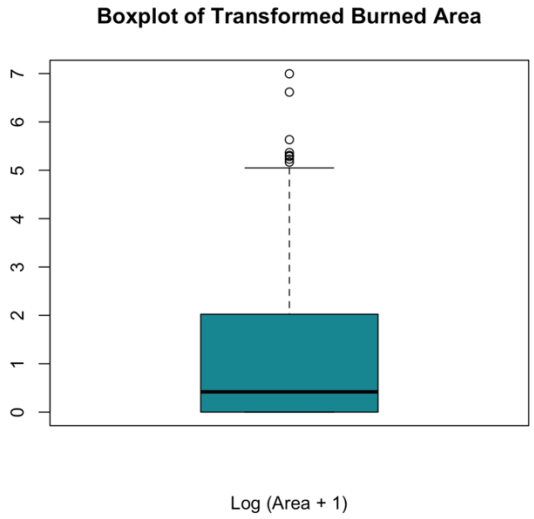
5-4-4 Conclusion

Neither of the two tentative models passed the three major model assumptions, and it can be seen from the above analysis results that the degree of linear correlation is low, so we decided to transform the variables and refit the new model.

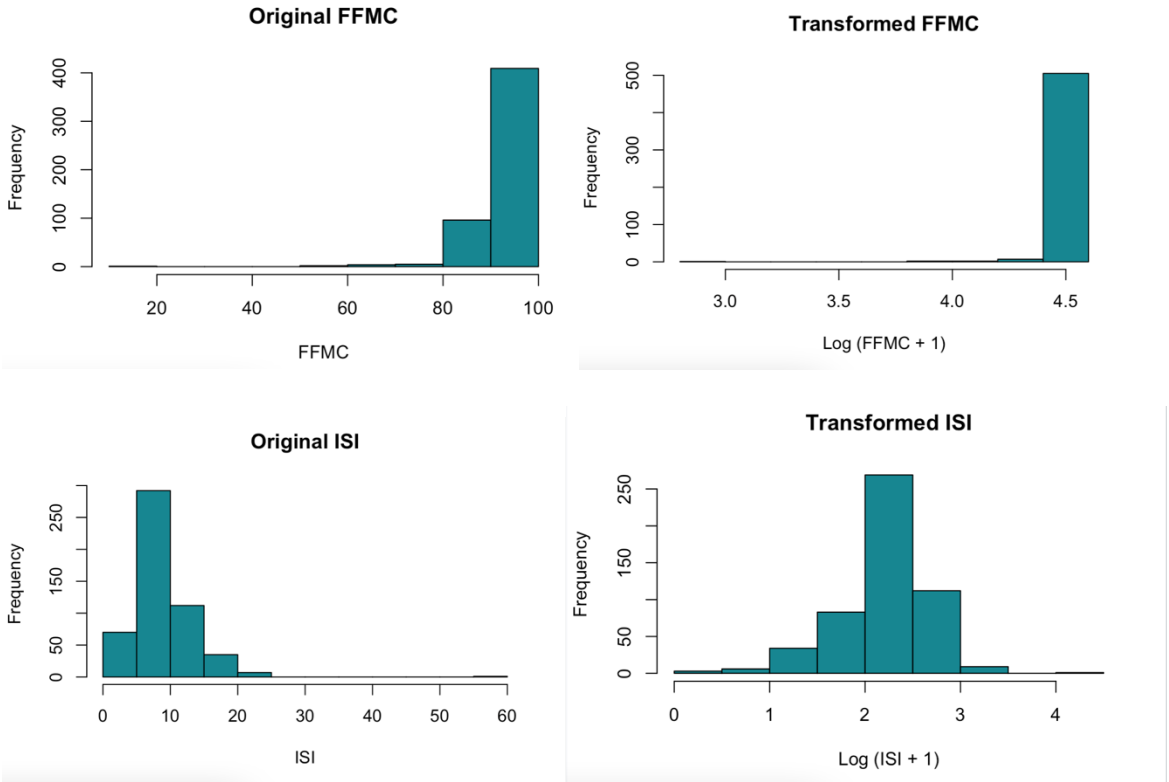
(Procedure 2)

Since the residual analysis results of the rebuilt model (Procedure 1) were unsatisfactory, we looked back at the data and found that 509 values in the rainfall (rain) variable were all 0. Considering that there was insufficient information in this field, we decided to delete this variable.

From the box -and-whisker plot of the logarithmic burning area (ln(area+1)) , we can also find that there are many outliers in the data and memory . We decided to delete extreme values that exceed 3 times the IQR .



F FMC and ISI have very high skewness. We perform variable transformation on this field, take the natural logarithm $\ln(FFMC+1)$and , respectively, of the original data $\ln(ISI+1)$, and delete outliers greater than three times the standard deviation.



DC and DMC are much larger than other field data . We try to open the root sign, take the general logarithm, and take the natural logarithm of the model explanation variables to balance the overall

field values and try to reduce the The level of the two variables month and day (for example: change month into four seasons, and day into weekday , weekend) until the model with the highest R squared value is found.

All processed field information is as follows:

```
> summary(df_new_transformed[, c(trans_x, trans_xx, "trans_area")])
      FFMC             ISI               DC              DMC           trans_area
 Min.   :2.981   Min.   :0.000   Min.   : 2.811   Min.   : 1.049   Min.   :0.0000
 1st Qu.:4.513   1st Qu.:2.015   1st Qu.:20.921   1st Qu.: 8.283   1st Qu.:0.0000
 Median :4.528   Median :2.241   Median :25.772   Median :10.407   Median :0.4187
 Mean   :4.515   Mean   :2.204   Mean   :22.316   Mean   : 9.991   Mean   :1.1110
 3rd Qu.:4.542   3rd Qu.:2.468   3rd Qu.:26.719   3rd Qu.:11.933   3rd Qu.:2.0242
 Max.   :4.577   Max.   :4.045   Max.   :29.336   Max.   :17.068   Max.   :6.9956
```

5-1 Rebuild regression model verification

The total deletion ratio is 15 , and the total deletion field is 1. Build the model again with this new data. Month (month) and holiday (weekday , weeke n d) are categorical variables, which are processed using dummy variables. Using April (Month_apr) and weekday (Day_weekday) as the basis, a new regression model relationship is established as follows:

Model_all2 : $\boldsymbol{ln}(\widehat{area+1}) = \widehat{\beta_0} + \widehat{\beta_1}X_i + \widehat{\beta_2}Y_i + \widehat{\beta_3}month_{jan_i} + \cdots + \widehat{\beta_{13}}month_{dec_i} + \widehat{\beta_{14}}day_{weekand_i} + \widehat{\beta_{15}}\boldsymbol{ln}(\widehat{FFMC+1})_i + \widehat{\beta_{16}}\sqrt{DMC}_i + \widehat{\beta_{17}}\sqrt{DC}_i + \widehat{\beta_{18}}\boldsymbol{ln}(\widehat{ISI+1})_i + \widehat{\beta_{19}}Temp_i + \widehat{\beta_{20}}RH_i + \widehat{\beta_{21}}Wind_i$ ，$\boldsymbol{i = 1, \ldots, 502}$

5-1-1 Parameter estimation

Parameter Estimation						
variable	Label	DF	parameter estimates	standard error	t value	Pr>|t|
Intercept	Intercept		1.743E+01	1.806E+01	0.965	0.33489

X	space coordinates	1	4.286E-02	3.185E-02	1.345	0.17913
Y	space coordinates	1	-2.687E-02	5.950E-02	-0.452	0.65179
Month_jan	January	1	-9.449E-01	1.440E+00	-0.656	0.51213
Month_feb	February	1	2.294E-01	5.507E-01	0.417	0.67722
Month_mar	March	1	-1.587E-01	5.083E-01	-0.312	0.75505
Month_may	May	1	8.539E-01	1.075E+00	0.794	0.42738
Month_jun	June	1	7.945E-02	7.483E-01	0.106	0.91549
Month_jul	July	1	4.500E-01	8.224E-01	0.547	0.58446
Month_aug	August	1	5.299E-01	-9.248E-01	0.573	0.56692
Month_sep	September	1	1.095E+00	1.002E+00	1.093	0.27486
Month_oct	October	1	1.098E+00	1.072E+00	1.025	0.30600
Month_dec	December	1	2.444E+00	8.866E-01	2.757	0.00606
Day_weekend	weekend	1	1.435E-01	1.294E-01	1.109	0.26794
FFMC	There is a base layer on the surface	1	-3.839E+00	4.112E+00	-0.934	0.35099
DMC	shallow organic layer	1	9.624E-02	4.352E-02	2.212	0.02747

DC	deep organic layer	1	-7.341E-02	4.859E-02	-1.511	0.13149
ISI	Fire speed level	1	4.546E-02	2.864E-01	0.159	0.87395
temp	temperature	1	3.259E-02	2.167E-02	1.504	0.13325
RH	Relative humidity	1	9.982E-05	6.019E-03	0.017	0.98678
Wind	wind speed	1	5.023E-02	3.822E-02	1.314	0.18942

From the above parameter estimation table, we can know the estimated value of each parameter, so we can know that the Model_1 model is as follows:

Model_all2 : $\boldsymbol{ln}(\widehat{area+1}) = 0.1743 + 0.004286X_i - 0.002687Y_i - 0.9449month_{jan_i} + \cdots + 2.444month_{dec_i} + 0.1435day_{weekand_i} - 3.839\boldsymbol{ln}(\widehat{FFMC+1})_{\boldsymbol{i}} + 0.09624\sqrt{DMC_i} - 0.07341\widehat{\beta_{17}}\sqrt{DC_i} + 0.04546\boldsymbol{ln}(\widehat{ISI+1})_{\boldsymbol{i}} + 0.03259Temp_i + 0.0001RH_i + 0.05023Wind_i$ ，$\boldsymbol{i = 1, \ldots, 502}$

5-1-2 Model fitness test

The ANOVA table is as follows:

ANOVA table				
source	degrees of freedom	SSE	MSE	Pr (>F)
Reg	20	57.38	2.869	0.0519
Error	481	870.64	1.8101	

The statistical assumptions are:

$$H_0 : \beta_1 = \cdots = \beta_{26} = 0$$

$$H_1 : \beta_k \; is \; not \; all \; equal \; to \; 0, k = 1, \ldots, 26$$

It can be seen from the above table that because P -value = 0.0519 $> \alpha =$0.05, it is not rejected H_0, and the original data of the inferred explanatory variables has no linear correlation with the logarithm of the burning area (ln(area+1)) . Although it did not pass the hypothesis test, it was very close to the rejection critical value and was more representative than the previous model.

5-1-3 Model explanation ability

Model explanation analysis table			
Multiple R-squared	0.06181	Adjusted R-squared	0.0228

Under this model, Multiple R-squared = 0.06181 , indicating that the total variation of ln(area+1) can be explained by all explanatory variables 6.181% ; and the corrected R-squared = 0.0228 , indicating that this model only has 2.28% Explanation ability.

5-1-4 Parameter verification

we get the values, we start to test βwhether there is a linear correlation between each explanatory variable after variable transformation and the logarithm of the burning area (ln(area+1)).

（2） X and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$$

Because p-value = 0.17913 > $\alpha = 0.05$, it is not rejected , and it is inferred that there is no linear relationship between $H_0$the X- axis of the spatial coordinates and the logarithmic combustion area.

(2) Y and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$$

Because p-value = 0.65179 > , $\alpha = 0.05$it is not rejected . It is inferred that there is no linear relationship between $H_0$the Y- axis of the spatial coordinates and the logarithmic combustion area.

(3) Month (month) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_k = 0 ; H_1 : \beta_k \neq 0 , \boldsymbol{k = 3, \dots, 13}$$

Except $\beta_{13}$for p-value = 0.00606 , $< \alpha = 0.05$rejected H_0, the parameter estimated p-values of other months are all greater than the significance level 0.05 , rejected H_0, it is speculated that there is a linear relationship between the spatial coordinates in December and the logarithm of the burned area, and There is no linear relationship between other months and logarithmic burned area.

(4) Weekend or not (Day_weekend) and burning area (ln(area+1)) , assuming that other variables are fixed:

$$H_0 : \beta_{14} = 0 ; H_1 : \beta_{14} \neq 0$$

Because the p-value of weekend (Day_weekend) = 0.26794 $> \alpha = 0.05$, it is not rejected H_0. It is speculated that there is no linear relationship between whether there is a weekend and the logarithm of the burned area .

(5) FFMC (ln(FFMC+1)) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{15} = 0 ; H_1 : \beta_{15} \neq 0$$

Because p-value = 0.35099 > , $\alpha = 0.05$it is not rejected H_0. It is speculated that there is no linear relationship between the logarithmic F FMC and the logarithmic combustion area.

(6) DMC ($\sqrt{DMC}$) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{16} = 0 ; H_1 : \beta_{16} \neq 0$$

Because p-value = 0.02747 , $< \alpha = 0.05$it is rejected . It is speculated that there is a linear relationship between D H_0MC after taking the root sign and the burning area after taking the logarithm.

(6) DC ($\sqrt{DC}$) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{16} = 0 ; H_1 : \beta_{16} \neq 0$$

Because p-value = 0.13149 , $> \alpha = 0.05$it is not rejected . It is speculated that there is a linear relationship between D H_0C after taking the root sign and the burning area after taking the logarithm.

(7) I SI (ln(ISI+1)) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{17} = 0 ; H_1 : \beta_{17} \neq 0$$

Because p-value = 0.87395 > , $\alpha = 0.05$it is not rejected H_0. It is speculated that there is no linear relationship between the logarithmic ISI and the logarithmic burning area.

(8) Temperature (temp) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{18} = 0 ; H_1 : \beta_{18} \neq 0$$

Because p-value = 0.13325 > $\alpha = 0.05$, it is not rejected H_0, and it is inferred that there is no linear relationship between temperature (temp) and the logarithm of the combustion area.

(9) Relative humidity (RH) and combustion area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{19} = 0 ; H_1 : \beta_{19} \neq 0$$

Because p-value = 0.98678 > $\alpha = 0.05$, it is not rejected H_0, and it is inferred that there is no linear relationship between relative humidity (RH) and the logarithm of the burned area.

(10) Wind speed (wind) and burning area (ln(area+1)), assuming that other variables are fixed:

$$H_0 : \beta_{20} = 0 ; H_1 : \beta_{20} \neq 0$$

Because p-value = 0.18942 > , $\alpha = 0.05$it is not rejected H_0, and it is inferred that there is no linear relationship between wind speed (wind) and logarithmic burning area.

5-1-5 Summary

Perform parameter estimation tests on each variable separately. Although most of the explanatory variables are not rejected $H_0$and there is no significant linear relationship between the explanatory variables and the response variables, the model performance after variable transformation is better than before the transformation, and the month is twelve There is a significant linear relationship between the moon and the shallow organic layer (DC) and the reaction variables. The relationship between the variables can be further explored through model selection.

5-2 Choice of remodeling

5-2-1 Forward selection method

According to the instructions in Section 4-1 , we again use the same rules for variable selection. The analysis results are as follows:

```
> forward_model

                        Selection Summary
-------------------------------------------------------------------------
        Variable                 Adj.
Step    Entered     R-Square    R-Square     C(p)         AIC         RMSE
-------------------------------------------------------------------------
   1    month         0.0349      0.0153    -3.2213    1739.2057    1.3505
   2    DMC           0.0398      0.0182    -3.6999    1738.6846    1.3485
   3    temp          0.0444      0.0209    -4.0604    1738.2718    1.3467
   4    wind          0.0487      0.0233    -4.2699    1738.0028    1.3450
-------------------------------------------------------------------------
> forward_model$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun     monthmar
    0.33809     -0.70820      1.44141      0.22062     -0.73743     -0.50918     -0.63114     -0.44487
   monthmay     monthoct     monthsep          DMC         temp         wind
    0.62734     -0.36552     -0.33898      0.05193      0.02476      0.05426
```

Final Model Output

Model Summary

R	0.221	RMSE	1.345
R-Squared	0.049	Coef. Var	121.733
Adj. R-Squared	0.023	MSE	1.809
Pred R-Squared	-Inf	MAE	1.091

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	45.174	13	3.475	1.921	0.0258
Residual	882.818	488	1.809		
Total	927.992	501			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.338	0.533		0.635	0.526	-0.709	1.385
monthaug	-0.708	0.553	-0.249	-1.281	0.201	-1.795	0.378
monthdec	1.441	0.652	0.141	2.212	0.027	0.161	2.722
monthfeb	0.221	0.548	0.031	0.403	0.687	-0.856	1.297
monthjan	-0.737	1.424	-0.024	-0.518	0.605	-3.536	2.061
monthjul	-0.509	0.571	-0.090	-0.892	0.373	-1.631	0.613
monthjun	-0.631	0.611	-0.079	-1.032	0.302	-1.832	0.570
monthmar	-0.445	0.490	-0.101	-0.908	0.364	-1.407	0.517
monthmay	0.627	1.053	0.029	0.596	0.552	-1.442	2.696
monthoct	-0.366	0.579	-0.046	-0.631	0.528	-1.503	0.772
monthsep	-0.339	0.527	-0.118	-0.643	0.520	-1.374	0.696
DMC	0.052	0.033	0.125	1.558	0.120	-0.014	0.117
temp	0.025	0.014	0.105	1.720	0.086	-0.004	0.053
wind	0.054	0.036	0.071	1.487	0.138	-0.017	0.126

The final model selected using the forward selection method is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.338 - 0.737month_{jan_i} + \cdots + 1.441month_{dec_i} + 0.052\sqrt{DMC_i} +$

$0.025Temp_i + 0.054Wind_i$, $\boldsymbol{i = 1, \ldots, 502}$

5-2-2 backward selection method

According to the instructions in Section 4-2 , we again use the same rules for variable selection. The analysis results are as follows:

```
> backward_model


                         Elimination Summary
-----------------------------------------------------------------------------
        Variable                 Adj.
Step    Removed     R-Square    R-Square     C(p)         AIC         RMSE
-----------------------------------------------------------------------------
   1    RH            0.0618      0.0248     1.0003    1743.0278    1.3440
   2    ISI           0.0618      0.0268    -0.9739    1741.0547    1.3426
   3    Y             0.0614      0.0284    -2.7706    1739.2669    1.3415
   4    day           0.0589      0.0279    -3.5314    1738.5579    1.3419
   5    FFMC          0.0556      0.0264    -3.7954    1738.3610    1.3429
   6    DC            0.0522       0.025    -4.0994    1738.1162    1.3439
   7    X             0.0487      0.0233    -4.2699    1738.0028    1.3450
-----------------------------------------------------------------------------
> backward_model$model

Call:
lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
    data = l)

Coefficients:
(Intercept)     monthaug     monthdec     monthfeb     monthjan     monthjul     monthjun     monthmar     monthmay
    0.33809     -0.70820      1.44141      0.22062     -0.73743     -0.50918     -0.63114     -0.44487      0.62734
   monthoct     monthsep          DMC         temp         wind
   -0.36552     -0.33898      0.05193      0.02476      0.05426
```

Final Model Output

Model Summary

R	0.221	RMSE	1.345
R-Squared	0.049	Coef. Var	121.733
Adj. R-Squared	0.023	MSE	1.809
Pred R-Squared	-Inf	MAE	1.091

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	45.174	13	3.475	1.921	0.0258
Residual	882.818	488	1.809		
Total	927.992	501			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	0.338	0.533		0.635	0.526	-0.709	1.385
monthaug	-0.708	0.553	-0.249	-1.281	0.201	-1.795	0.378
monthdec	1.441	0.652	0.141	2.212	0.027	0.161	2.722
monthfeb	0.221	0.548	0.031	0.403	0.687	-0.856	1.297
monthjan	-0.737	1.424	-0.024	-0.518	0.605	-3.536	2.061
monthjul	-0.509	0.571	-0.090	-0.892	0.373	-1.631	0.613
monthjun	-0.631	0.611	-0.079	-1.032	0.302	-1.832	0.570
monthmar	-0.445	0.490	-0.101	-0.908	0.364	-1.407	0.517
monthmay	0.627	1.053	0.029	0.596	0.552	-1.442	2.696
monthoct	-0.366	0.579	-0.046	-0.631	0.528	-1.503	0.772
monthsep	-0.339	0.527	-0.118	-0.643	0.520	-1.374	0.696
DMC	0.052	0.033	0.125	1.558	0.120	-0.014	0.117
temp	0.025	0.014	0.105	1.720	0.086	-0.004	0.053
wind	0.054	0.036	0.071	1.487	0.138	-0.017	0.126

The final model selected using the backward selection method is as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.338 - 0.737month_{jan_i} + \cdots + 1.441month_{dec_i} + 0.052\sqrt{DMC}_i +$

$0.025Temp_i + 0.054Wind_i$, $\boldsymbol{i = 1, \ldots, 502}$

5-2-3 Stepwise regression method

According to the instructions in Section 4-3 , we again use the same rules for variable selection. The analysis results are as follows:

```
> stepwise_model

                          Stepwise Selection Summary
--------------------------------------------------------------------------------
                    Added/                   Adj.
Step    Variable    Removed     R-Square    R-Square     C(p)         AIC         RMSE
--------------------------------------------------------------------------------
   1     month      addition       0.035       0.015    -3.2210    1739.2057    1.3505
--------------------------------------------------------------------------------
```

```
Final Model Output
------------------

                        Model Summary
-----------------------------------------------------------
R                       0.187       RMSE                1.351
R-Squared               0.035       Coef. Var         122.234
Adj. R-Squared          0.015       MSE                 1.824
Pred R-Squared           -Inf       MAE                 1.105
-----------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                             ANOVA
---------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F        Sig.
---------------------------------------------------------------------
Regression     32.416         10          3.242    1.777    0.0621
Residual      895.576        491          1.824
Total         927.992        501
---------------------------------------------------------------------

                                Parameter Estimates
--------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower    upper
--------------------------------------------------------------------------------------
(Intercept)     1.089         0.450                2.420    0.016     0.205    1.974
   monthaug    -0.072         0.461       -0.025   -0.156    0.876    -0.979    0.834
   monthdec     1.482         0.637        0.145    2.328    0.020     0.232    2.733
   monthfeb     0.056         0.547        0.008    0.102    0.919    -1.018    1.130
   monthjan    -1.089         1.424       -0.036   -0.765    0.445    -3.886    1.708
   monthjul     0.029         0.511        0.005    0.057    0.954    -0.975    1.034
   monthjun    -0.134         0.569       -0.017   -0.235    0.814    -1.253    0.985
   monthmar    -0.302         0.487       -0.068   -0.620    0.535    -1.259    0.655
   monthmay     0.749         1.056        0.035    0.709    0.479    -1.326    2.823
   monthoct    -0.172         0.569       -0.022   -0.302    0.763    -1.291    0.947
   monthsep     0.152         0.462        0.053    0.328    0.743    -0.756    1.059
--------------------------------------------------------------------------------------
```

Model_2 : $\boldsymbol{ln}(\widehat{area}+1) = 1.089 - 1.089month_{jan_i} + \cdots + 1.482month_{dec_i}$, $\boldsymbol{i = 1, \ldots, 502}$

5-2-4 Summary

Combining the models selected by the above three methods , two temporary models are produced. The results of the regression equation are as follows:

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.338 - 0.737month_{jan_i} + \cdots + 1.441month_{dec_i} + 0.052\sqrt{DMC}_i + 0.025Temp_i + 0.054Wind_i$ **，** $\boldsymbol{i = 1, \ldots, 502}$

Model_2 :ln $(\widehat{area+1}) = 1.089 - 1.089month_{jan_i} + \cdots + 1.482month_{dec_i}$ **，** $\boldsymbol{i = 1, \ldots, 502}$

5-3 Collinearity test again

Collinearity diagnosis is performed on the Model_1 temporary model with multiple variables . The results are as follows:

Collinearity test				
variable name	Month	DMC	Temp	Wind
VIF value	4.896305	3.311605	1.924123	1.177972

The VIF values of the explanatory variables in the temporary model do not exceed 10 , and there is no problem of model collinearity.

5.4 Residual analysis of reconstructed model

According to the instructions in Section 5.4 , the model to be tested is as follows :

Model_1 : $\boldsymbol{ln}(\widehat{area+1}) = 0.338 - 0.737month_{jan_i} + \cdots + 1.441month_{dec_i} + 0.052\sqrt{DMC}_i + 0.025Temp_i + 0.054Wind_i$ **，** $\boldsymbol{i = 1, \ldots, 502}$

Model_2 :ln $(\widehat{area+1}) = 1.089 - 1.089month_{jan_i} + \cdots + 1.482month_{dec_i}$

， $\boldsymbol{i = 1, \ldots, 502}$

5-4-1 Normality test

In the R software, we use shapiro.test to test the normality of the residuals. The assumptions and results are as follows:

$$H_0 : Residuals\ are\ normally\ distributed.$$

$$H_1 : Residuals\ are\ not\ normally\ distributed.$$

Shapiro-Walk Normality Test (Model_1)			
W	0.87207	**P-value**	< 2.2e-16
Shapiro-Walk Normality Test (Model_2)			
W	0.84745	**P-value**	< 2.2e-16

The p-value $< \alpha = 0.05$, of the two temporary models is rejected < 2.2e-16 H_0, which means that the residuals do not meet the assumption of normal distribution.

5-4-2 Independence Check

In the R software, we use dubinWatsonTest to perform the independence test of the residuals. The assumptions and results are as follows:

$$H_0 : Residuals\ are\ mutually\ independent.$$

$$H_1 : Residuals\ are\ not\ mutually\ independent.$$

Durbin Watson Test (Model_1)		Durbin Watson Test (Model_2)	
DW Statistics	**p-value**	**DW Statistics**	**p-value**
1.7904469	0.01267	1.6919894	0.009023

The p-values of the two temporary models are $< \alpha = 0.05$, rejected H_0, which means that the residuals do not meet the assumption of independence.

5-4-3 Homogeneity test

In the R software, we use ncv Test to perform the homogeneity test of the residuals. The assumptions and results are as follows:

$$H_0 : Homoscedasticity\ of\ Variance$$

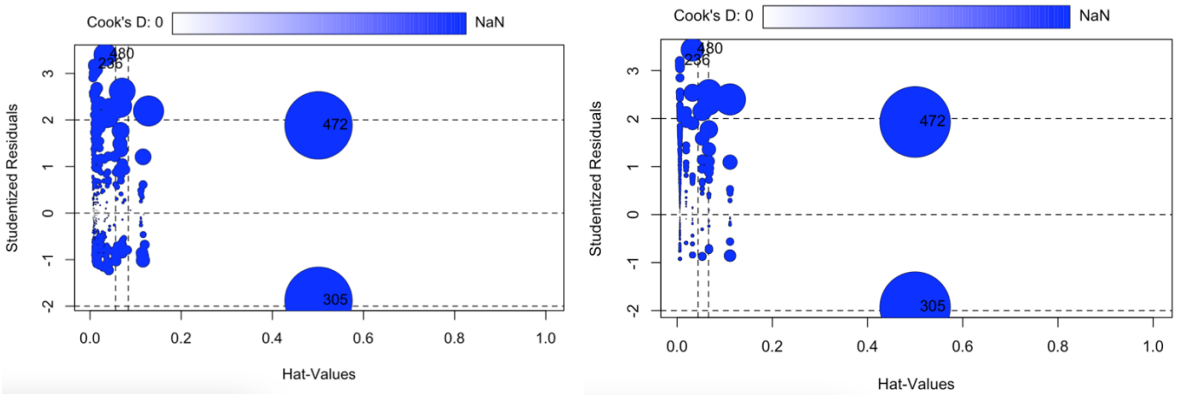
$$H_1 : Heteroscedasticity\ of\ Variance$$

NON-CONSTANT VARIANCE SCORE TEST (Model_1)					
Chisquare	1.063679	**DF**	1	**p-value**	0.30238
NON-CONSTANT VARIANCE SCORE TEST (Model_2)					
Chisquare	0.02354126	**DF**	1	**p-value**	0.87806

The p-values of the two temporary models are 0.30238 and 0.87806 respectively , both of which are not $> \alpha = 0.05$, rejected H_0, which means that the variables meet the assumption of homogeneity.

5-4-4 Delete impact points

Both temporary models only passed the homogeneity test, and the influence points (2 strokes each) were deleted and the regression test was performed again.

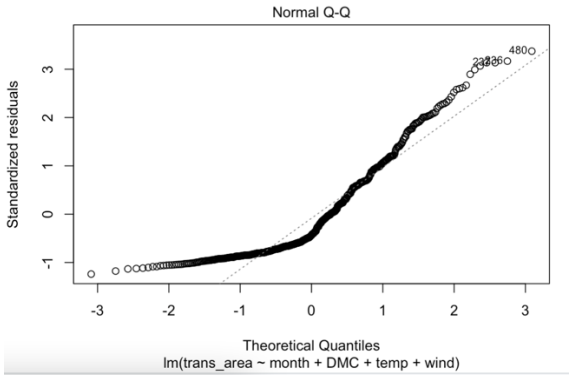


5-4-5 Final Normality Test

In the R software, we use shapiro.test to test the normality of the residuals. The assumptions and results are as follows:

$$H_0 : Residuals\ are\ normally\ distributed.$$

$$H_1 : Residuals\ are\ not\ normally\ distributed.$$



Shapiro-Walk Normality Test (Model_1)			
W	0.86916	**P-value**	< 2.2e-16
Shapiro-Walk Normality Test (Model_2)			
W	0.84347	**P-value**	< 2.2e-16

The p-value $< \alpha = 0.05$, of the two temporary models is rejected < 2.2e-16 H_0, which means that the residuals do not meet the assumption of normal distribution. It can also be seen from QQ-plot that the residual value deviates from the 45 -degree line.

5-4-6 Final Independence Check

In the R software, we use dubinWatsonTest to perform the independence test of the residuals. The assumptions and results are as follows:

$$H_0 : Residuals\ are\ mutually\ independent.$$

H_1 ：$Residuals\ are\ not\ mutually\ independent.$

Durbin Watson Test (Model_1)		Durbin Watson Test (Model_2)	
DW Statistics	**p-value**	**DW Statistics**	**p-value**
1.8336011	0.05132	1.8096540	0.03845

It can be seen from the above table that the p-value of Model_1 = 0.05132 is rejected H_0, which means that the residuals meet the assumption of mutual independence; the p- value $> \alpha = 0.05$ ，of Model_2 = 0.03845 is not rejected H_0, which means that the residuals do not meet the assumption of mutual independence . $< \alpha = 0.05$ ，.

5-4-7 Final Homogeneity test

In the R software, we use ncv Test to perform the homogeneity test of the residuals. The assumptions and results are as follows:

H_0 ：$Homoscedasticity\ of\ Variance$

H_1 ：$Heteroscedasticity\ of\ Variance$

NON-CONSTANT VARIANCE SCORE TEST (Model_1)					
Chisquare	0.8118436	**DF**	1	**p-value**	0.36758
NON-CONSTANT VARIANCE SCORE TEST (Model_2)					
Chisquare	0.104529	**DF**	1	**p-value**	0.74646

The p-values of the two temporary models are 0.36758 and 0.74646 respectively , both of which are not $> \alpha = 0.05$ ，rejected H_0, which means that the variation numbers meet the assumption of homogeneity.

5-5 Conclusion

After deleting the influence points (2 strokes each), Model_2 still only passed the homogeneity test, while Model_1 could pass the independence test and the homogeneity test. However, the normality test still failed. We speculate that the sample number may not be large enough. the result of.

Therefore, we finally used Model_1 as our final model. The corrected R-squared = 0.03167 . Although the value is low, the p-value of this overall model = 0.005988 $< \alpha = 0.05$, which represents a linear relationship. And although the normality test failed, it should be able to pass if the sample size is large enough. The final model regression relationship is as follows:

Model : $\boldsymbol{ln}(\widehat{area} + 1) = 0.18611 - 0.56544 month_{jan} + \cdots + 1.46619 month_{dec} + 0.065\sqrt{DMC} + 0.03174 Temp + 0.05803 Wind$, $\boldsymbol{i = 1, \ldots, 500}$

Chapter 6. Reference Materials

1. data set

https://archive.ics.uci.edu/ml/datasets/Forest+Fires

2. A Data Mining Approach to Predict Forest Fires using Meteorological Data

https://www.researchgate.net/publication/238767143_A_Data_Mining_Approach_to_Predict_Forest_Fires_using_Meteorological_Data

3. Kaggle

https://www.kaggle.com/code/psvishnu/forestfire-impact-prediction-stats-and-ml