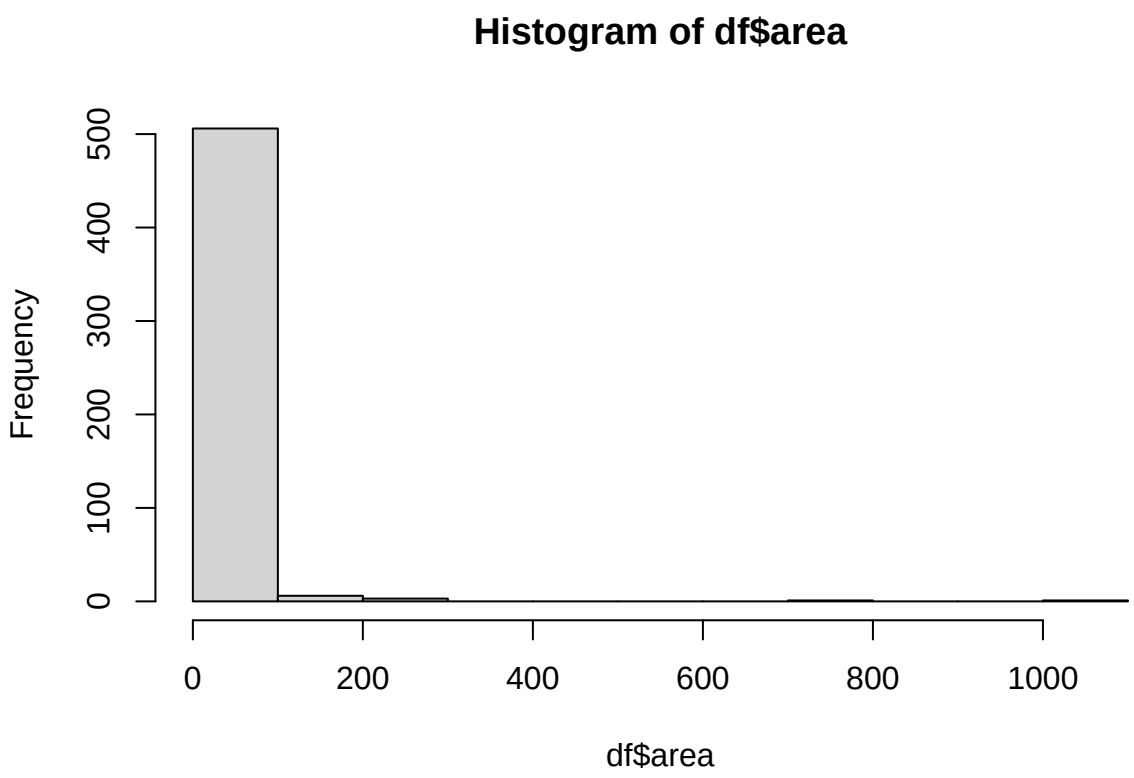
Linear Regression Model

Bowei Zhang

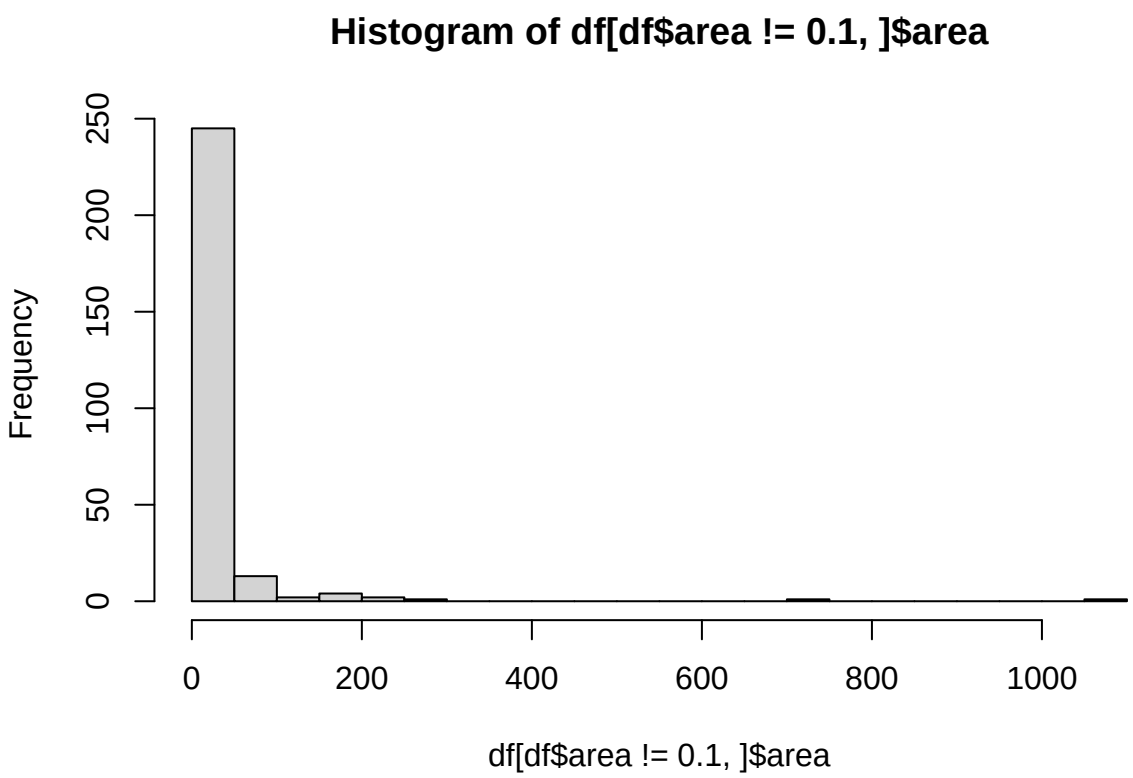
2023-03-27

Data Wrangling

```
# See distribution
hist(df$area)
```




```
# Distribution without area = 0.1
hist(df[df$area != 0.1,]$area, breaks = 20)
```

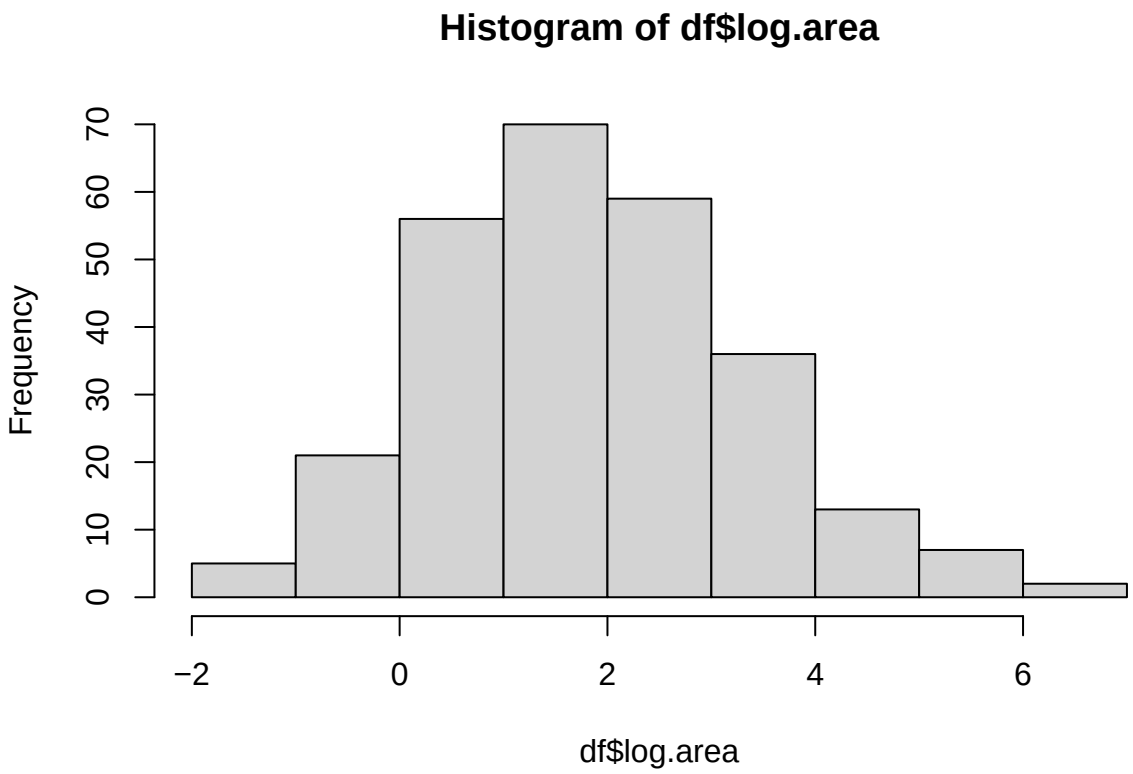


update 'df'

```
df <- df %>%
  dplyr::select(-area_size) %>%
  filter(area != 0.1) %>%
  mutate(log.area = log(area))
head(df)
```

```
## # A tibble: 6 x 14
##    area X     Y     month day    FFMC   DMC    DC   ISI  temp    RH  wind  rain
##   <dbl> <fct> <fct> <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  0.36 9     9     jul   tue    85.8  48.3  313.   3.9  18      42   2.7     0
## 2  0.43 1     4     sep   tue    91   130.   693.   7    21.7    38   2.2     0
## 3  0.47 2     5     sep   mon    90.9 126.   686.   7    21.9    39   1.8     0
## 4  0.55 1     2     aug   wed    95.5  99.9  513.  13.2  23.3    31   4.5     0
## 5  0.61 8     6     aug   fri    90.1 108    530.  12.5  21.2    51   8.9     0
## 6  0.71 1     2     jul   sat    90    51.3  296.   8.7  16.6    53   5.4     0
## # ... with 1 more variable: log.area <dbl>
```

```
hist(df$log.area)
```



Data Analysis

variable selection

```
step(lm(log.area ~ 1, data = df %>% dplyr::select(-area)),
     ~ X + Y + FFMC + DMC + DC + ISI + temp + RH + wind, direction="both", trace = 0)
```

```
##
## Call:
## lm(formula = log.area ~ Y + ISI, data = df %>% dplyr::select(-area))
##
## Coefficients:
## (Intercept)           Y3           Y4           Y5           Y6           Y8
##     2.50839     -0.15182     -0.38403     -0.42221      0.05598      3.02269
##          Y9          ISI
##    -2.13931     -0.03931
```

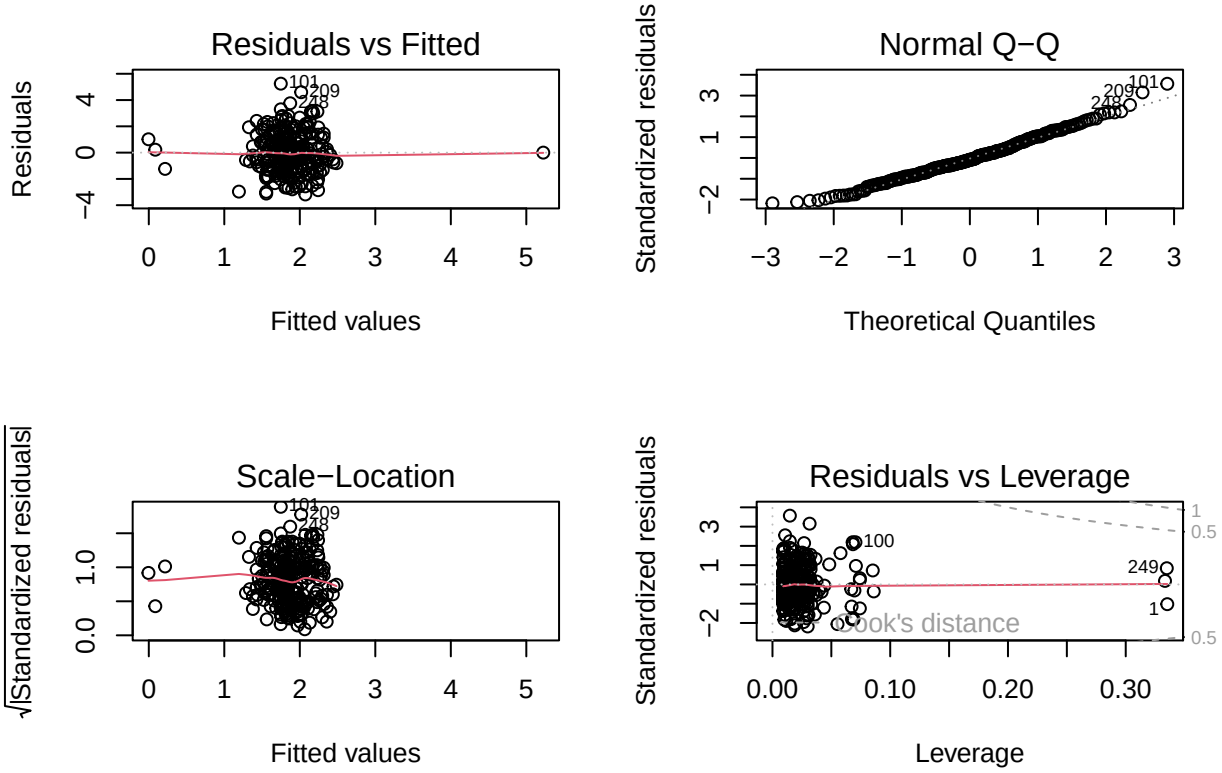
```
m1 = lm(formula = log.area ~ Y + ISI, data = df)
summary(m1)
```

```
##
## Call:
## lm(formula = log.area ~ Y + ISI, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1822 -1.0342 -0.1236  0.9544  5.2466
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.50839    0.45470   5.517 8.32e-08 ***
## Y3          -0.15182    0.46248  -0.328   0.7430
## Y4          -0.38403    0.41016  -0.936   0.3500
## Y5          -0.42221    0.42489  -0.994   0.3213
## Y6           0.05598    0.45417   0.123   0.9020
## Y8           3.02269    1.53193   1.973   0.0495 *
## Y9          -2.13931    0.94177  -2.272   0.0239 *
## ISI         -0.03931    0.02204  -1.784   0.0756 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.482 on 261 degrees of freedom
## Multiple R-squared:  0.05904,    Adjusted R-squared:  0.03381
## F-statistic:  2.34 on 7 and 261 DF,  p-value: 0.0248
```

```
par(mfrow = c(2,2))
plot(m1)
```

```
## Warning: not plotting observations with leverage one:
##   213
```



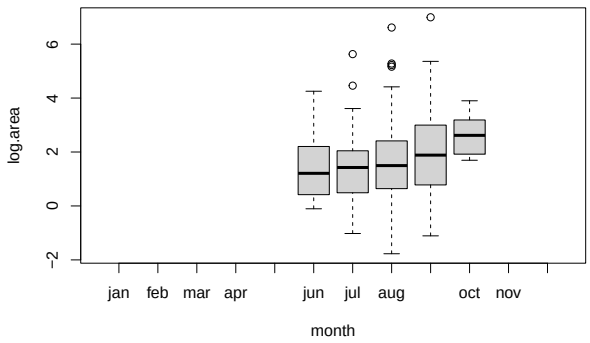
Adjusted R-square is too low.

PCA

```
## Importance of components:
##                             PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.7735 1.2541 1.1188 1.0847 1.01169 0.95259 0.81380
## Proportion of Variance 0.2859 0.1430 0.1138 0.1070 0.09305 0.08249 0.06021
## Cumulative Proportion  0.2859 0.4289 0.5427 0.6497 0.74273 0.82522 0.88543
##                            PC8     PC9    PC10    PC11
## Standard deviation     0.67150 0.57171 0.54327 0.43287
## Proportion of Variance 0.04099 0.02971 0.02683 0.01703
## Cumulative Proportion  0.92642 0.95614 0.98297 1.00000
```

```
##              PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9   PC10
## area       0.071 -0.684 -0.078 -0.029  0.007 -0.029  0.179 -0.689  0.065  0.086
## Y          0.000 -0.072  0.113  0.255 -0.811 -0.491  0.061  0.071  0.060 -0.078
## FFMC       0.463  0.065 -0.057 -0.263 -0.040 -0.064  0.321  0.157  0.315  0.497
## DMC        0.410 -0.013 -0.317  0.331  0.088 -0.189 -0.193  0.070 -0.419  0.469
## DC         0.401  0.025 -0.155  0.394  0.189 -0.068 -0.350 -0.059  0.602 -0.359
## ISI        0.377  0.140 -0.185 -0.481 -0.026 -0.164  0.357  0.007 -0.011 -0.425
## temp       0.475 -0.038  0.216 -0.020 -0.081  0.140 -0.162 -0.104 -0.560 -0.366
## RH        -0.220  0.158 -0.608  0.325  0.132 -0.148  0.451 -0.059 -0.185 -0.217
## wind      -0.179 -0.019 -0.442 -0.513 -0.067 -0.340 -0.581 -0.076 -0.029  0.012
## rain       0.054  0.044 -0.432 -0.006 -0.512  0.726 -0.077 -0.031  0.074  0.031
## log.area -0.011 -0.688 -0.144 -0.023  0.077  0.056  0.033  0.683 -0.019 -0.166
##              PC11
## area        0.029
## Y          -0.007
## FFMC       -0.479
## DMC         0.375
## DC         -0.043
## ISI         0.486
## temp       -0.463
## RH         -0.353
## wind       -0.212
## rain        0.078
## log.area -0.027
```

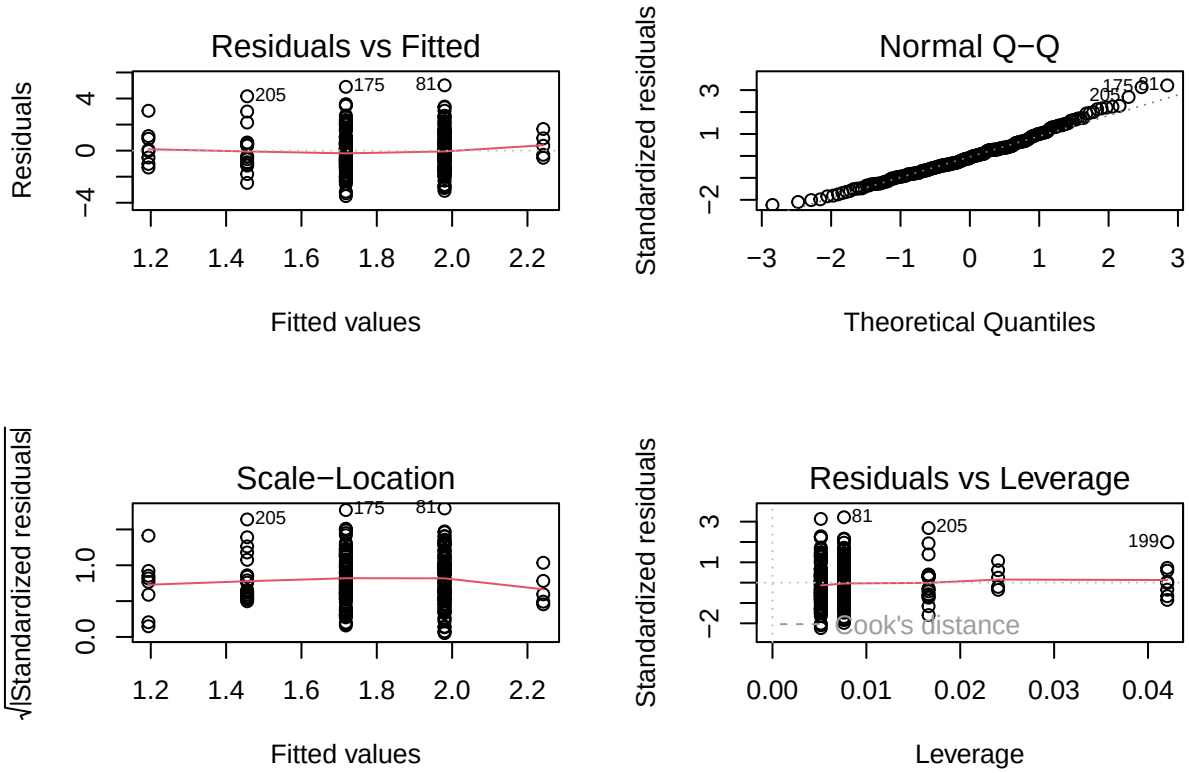
During Summer




```
##
## Call:
## lm(formula = log.area ~ month, data = df_summer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4899 -1.0539 -0.1441  0.9216  5.0148
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3782     1.0937  -0.346   0.7298
## month         0.2620     0.1308   2.003   0.0464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.567 on 224 degrees of freedom
## Multiple R-squared:  0.0176, Adjusted R-squared:  0.01321
## F-statistic: 4.012 on 1 and 224 DF,  p-value: 0.04638
```

```
##
## Call:
## lm(formula = log.area ~ poly(month, 2), data = df_summer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4064 -1.0483 -0.1389  0.8214  4.9857
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.8025     0.1042  17.299   <2e-16 ***
## poly(month, 2)1   3.1379     1.5664   2.003   0.0464 *
## poly(month, 2)2   1.6013     1.5664   1.022   0.3078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.566 on 223 degrees of freedom
## Multiple R-squared:  0.02218,    Adjusted R-squared:  0.01341
```

```
## F-statistic: 2.529 on 2 and 223 DF,  p-value: 0.08203
```



Conclusion

- According to our results, the area burned from forest fires can not be predicted using linear regression model from the variables that we havein the dataset.
- Other variables - such as elevation, levels of human activity - may be necessary to predict area burned
- Other statistical reports have had similar diffuculties predicting area burned.(source)
- Some suggest that neural network and other Tree based models could give a better result.