[통계 분석 방법론 최종 과제]

Forest Fire 데이터를 활용한 Montesinho 국립공원의 화재로 인한 연소면적 predictor 만들기

2020020324 임형준

1.서론

UCI machine learning repository에서 제공하는 'Forest Fire' 데이터셋[Cortez and Morais, 2007]은 포르투갈의 Montesinho 국립공원에서 발생한 숲 화재에 대한 정보를 담고 있다. 이 데이터셋은 총 12개의 독립변수와 1개의 종속변수(burned area: 연소면적)로 이루어져 있다. 이 프로젝트의 목적은 종속변수와 독립변수의 상관관계를 알아내어 화재 발생시 공원의 연소면적을 예측하는 predictor를 만드는 것이다. 이를 통해 화재로 인한 자연 피해를 줄이기 위해 어떤 요인을 통제하여야 하는지 또 그 통제가 어떤 효과를 갖는지 알 수 있게 될 것이라 전망된다.

종속변수와 독립변수의 유의한 연관성을 찾기 위해 linear regression, robust regression(M regression, LMS regression), quantile regression, non-linear regression(LOESS and GAM) 그리고 Support Vector regression analysis 등이 활용되었다. 먼저 전체 데이터에 모델들을 적합하여 데이터에 대한 적합성을 판단하였다. 그리고 데이터를 observation을 기준으로 5개의 파트로 나눠 한 파트가 test data가 되고 나머지 네 파트가 train data가 되는 5-fold cross validation 방법론을 활용하여 train data를 통해 학습된 모델의 prediction accuracy(MSE)들의 평균을 구해 모델별로 예측기로서의 성능을 평가하였다.

또한 Principal Component Analysis를 활용하여 12개의 설명변수 중 10개의 연속형 변수를 5개의 연속형 변수로 줄이는 차원 축소를 진행하였다. 이렇게 축소된 데이터에도 종전과 같은 과정의 통계적 모델을 구축하면 예측기로서의 성능 보전이 되는지 확인하였다. 또한 PCA를 통해 새로 생성된 다섯개의 Principal Component 변수에 대해, 기존 변수와의 연관성을 확인하여 잠재적 변수로서의 가능성과 특성을 파악, 정리하였다.

2. Explorative Data Analysis

데이터의 특성을 파악하기 위해 EDA를 진행하였다. 변수들의 형태와 성질을 먼저 파악하고 필요에 따라 변수 변환까지도 실시하여 추후 통계 모델을 적용함에 있어 용이하게 하였다. 또한 결측 발생 여부를 확인하였고, 연속형 변수의 경우 시각화, 통계적 검정 방법(콜모고로프-스미르노프 검정) 등을 통해 특정 분포를 따르는 지 확인하였다. 마지막으로 Pearson 상관계수 검정, Kendall의 tau검정, Kruskal-Wallis 검정 등을 활용하여 각 변수의 종속변수(burned area) 연관성 대해서도 측정하였다.

[변수 설명]

13개의 변수 & 517개의 observation

1. X (연속형) – 공원 내의 위도 정보 [1-9] |2. Y (연속형) – 공원 내의 경도 정도 [2-9] |3. month (이산형) – 월 정보 |4. day (이산형) – 요일 정보 |5. FFMC (연속형) - fine fuel moisture code; FWI 시스템의 지표 [18.7-96.2] |6. DMC (연속형) - duff moisture code; FWI 시스템의 지표 [1.1-291.3] |7. DC (연속형)- drought code; FWI 시스템의 지표 [7.9-860.6] |8. ISI (연속형)- initial spread index; FWI 시스템의 지표 [0-56.1] |9. temp (연속형)- temperature in Celsius degrees: [2.2-33.3] |10. RH (연속형)- relative humidity in %: [15-100] |11. wind (연속형)- wind speed in km/h: [0.4-9.4] |12. rain (연속형)- outside rain in mm/m2 : [0-6.4] |13. area (연속형, 종속변수) - the burned area of the forest (in ha) [0-1090.84]

[변수 별 결측 여부 파악]

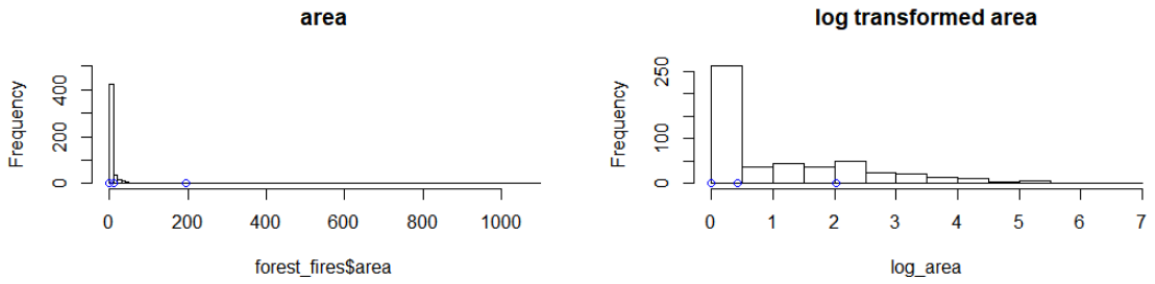
R code	Output

`# missing check`
`missing <-c()`
`for(i in 1:13){`
`  missing <- c(missing, sum(is.na(forest_fires[,i])))`
`}`
`print(missing)`
	`> missing <-c()`
	`> for(i in 1:13){`
	`+  missing <- c(missing, sum(is.na(forest_fires[,i])))`
	`+ }`
	`> print(missing)`
	`[1] 0 0 0 0 0 0 0 0 0 0 0 0 0`

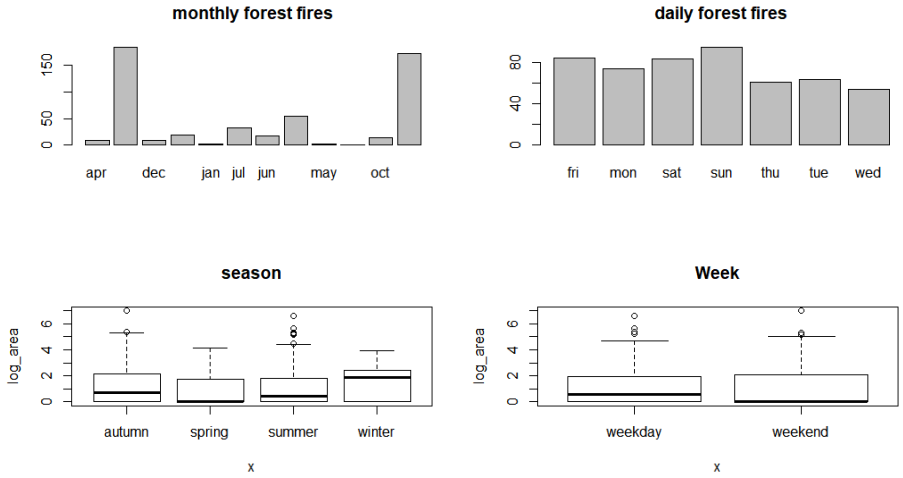
-13개의 변수에서 결측 발생한 observation 전무.

[변수변환]

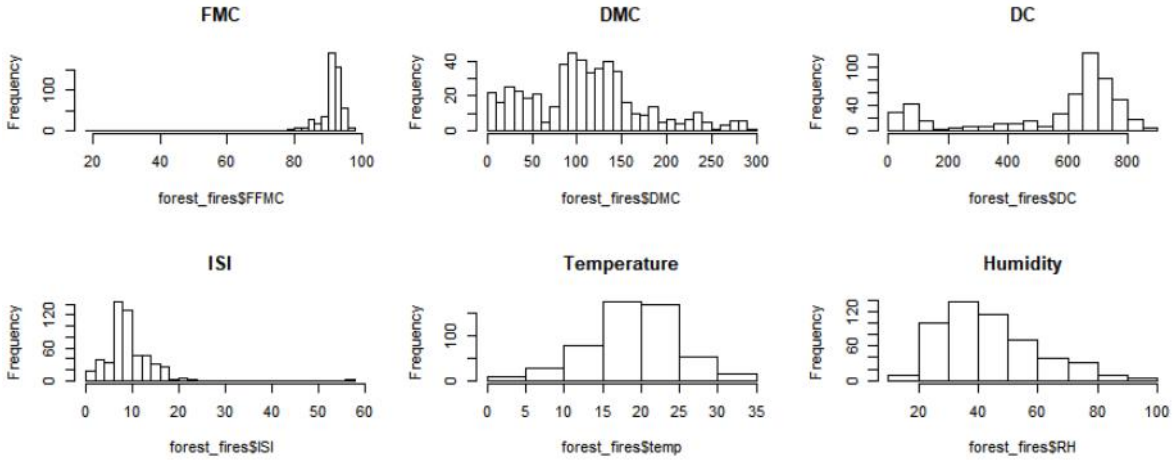
- 종속변수 burned area는 0부터 1090.84의 범위를 갖는 것을 감안하면 0으로 너무 치우친 분포를 가진다. 따라서 데이터 제공처(UCI machine learning repository)의 권고사항에 따라 log(area+1) 변수로 변환을 실시하여 추후의 모델 적합에 활용하도록 한다.



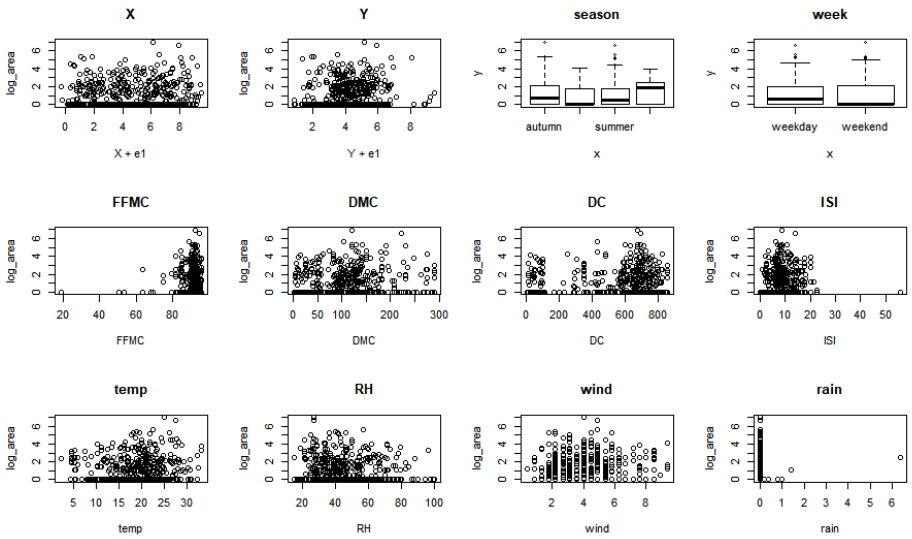
- 두 이산형 변수 month와 day는 각각 범주의 수가 너무 많아서 시각화 자료상 종속변수와의 연관성이 뚜렷하게 보이지 않았다. 따라서 month의 데이터는 3월,4월,5월은 spring, 6월,7월,8월은 summer, 9월,10월,11월은 autumn, 12월,1월,2월은 winter로 category를 통합하여 season(계절)변수로 변환하였다. 또 day의 데이터는 월~목요일은 weekdays, 금~일요일은 weekend변수로 category를 통합하여 week변수로 변환하였다.



[기타 변수 시각화]



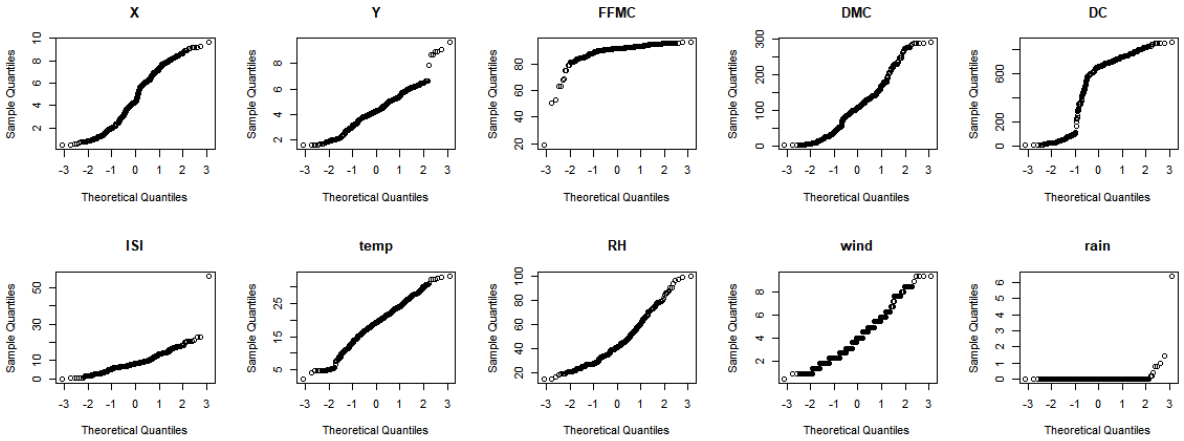
[종속변수와의 연관성]



- 각 12개의 설명변수를 log(area+1)에 대해 plotting한 결과이다. 이렇게만 봐서는 어떠한 연관성이 있는지 확인하기가 어려워 우선적으로 연속형 변수와 종속변수 간의 선형적인 연관성 체크

를 위해 상관계수 검정을 실시하였다. 정규분포를 따르는 연속형 변수에 대해서는 Pearson Correlation 검정이 가능하기 때문에 우선 정규성 검정을 위해 10개의 변수에 대해 qqplot을 그리고, 콜모고로프-스미르노프 검정을 실시하였다. (K-S 검정 시 동점 제거를 위해 $N(0,0.3^2)$에서 뽑은 난수를 추가하는 jittering을 실시하였다)

[qqplot과 K-S검정 p-value]



변수	X	Y	FFMC	DMC	DC
p-value	0.003	0.0774	<0.001	0.007	<0.001
변수	ISI	temp	RH	wind	rain
p-value	<0.001	0.1349	<0.001	0.0233	<0.001

- p-value가 0.05보다 작은 변수들은 normal과 다른 누적분포를 갖는다는 결과를 낸다. K-S 검정상 근사적으로 정규분포를 따른다고 할 수 있는 변수는 Y와 temp 정도이다. 이 두 변수에 대해서는 Pearson의 상관계수 검정을, 나머지 변수에 대해서는 Kendall의 tau 상관계수 검정을 실시한다. 그 결과는 다음과 같다.

[연속형 변수와 종속변수의 상관계수 검정]

변수	X	Y	FFMC	DMC	DC
p-value	0.1729	0.3782	0.5687	0.0865	0.1575
변수	ISI	temp	RH	wind	rain
p-value	0.7738	0.2247	0.5773	0.2282	0.1459

- 그나마 DMC가 작은 p-value(0.0865)를 가지지만 그마저 0.05보다 크다. 즉, 설명변수들 중 종속변수와 선형적인 연관성을 가지는 변수는 DMC를 제외하면 전무하다고 판단된다.
- 이산형 설명변수 season과 week는 종속변수와 어떠한 연관성을 가지는 검정하기위해 ANOVA의 비모수적 방법인 Kruskal-Wallis 검정을 실시하였다.

[이산형 변수와 종속변수의 연관성 검정]

변수	Season	Week
p-value	0.06363	0.6385

- Season의 경우 p-value 기준으로 0.05보다는 크지만 종속변수와 어느정도 유의한 연관성을 지니는 것을 확인할 수 있다.

[EDA 결과]

- 결론적으로 연속형 변수 중에는 단순히 선형적인 연관성만 놓고 보자면 DMC만이, 이산형 변수 중에는 season만이 종속변수와 상대적으로 유의한 연관성을 가진다고 할 수 있다. 하지만 비선형적인 연관성이나 눈으로 감지할 수 없는 연관성이 있을 수 있으므로 추후 통계적 모델을 구축할 시에는 모든 변수를 다 활용하도록 한다. Input에 제약이 있을 경우에는 유의하게 연관성이 있는 변수들을 우선적으로 선택, 적용하도록 한다.

3. Statistical Model Fitting

[5-fold data split]

- 통계적 모델을 기반한 예측기를 만들기 위해서는 data를 training set과 test set으로 나누어야 한다. Training set을 사용하여 fitting된 통계적 모델이 predictor가 되어 test set 데이터의 설명변수를 통해 종속변수를 예측하고 이를 실제 종속변수 값과 비교하여 prediction accuracy를 구하는 것이 일반적인 predictor의 성능평가 과정이다. 하지만 train-test를 1회만 실시해 모델을 비교할 경우 일반적인 우열을 평가하기 힘들기 때문에 데이터를 5개의 set으로 나누어 각 set이 돌아가며 test set의 역할을 하는 5-fold validation의 발상을 차용하여 각 모델마다 이렇게 구해진 5개의 prediction accuracy(MSE)의 평균값을 비교하도록 하였다.

[Linear Regression Predictor]

- 가장 일반적으로 생각할 수 있는 통계 모델인 선형회귀 모델이다.
- 먼저 전체 데이터 Forest Fire를 모델에 적합하여 모델 적합성을 평가해본다.

Fitted Model

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.312e+00  1.558e+00  -0.842   0.4002
X             4.305e-02  3.161e-02   1.362   0.1738
Y             7.177e-03  5.970e-02   0.120   0.9043
seasonspring -1.277e-01  4.907e-01  -0.260   0.7948
seasonsummer -3.665e-01  1.983e-01  -1.849   0.0651 .
seasonwinter  7.437e-01  5.034e-01   1.478   0.1402
weekweekend  -1.891e-02  1.254e-01  -0.151   0.8803
FFMC          1.499e-02  1.479e-02   1.014   0.3111
DMC           2.479e-03  1.677e-03   1.478   0.1400
DC           -8.422e-06  7.437e-04  -0.011   0.9910
ISI          -1.695e-02  1.719e-02  -0.986   0.3246
temp          2.980e-02  2.064e-02   1.444   0.1494
RH           -9.011e-04  5.729e-03  -0.157   0.8751
wind          8.404e-02  3.687e-02   2.279   0.0231 *
rain          3.694e-02  2.128e-01   0.174   0.8623

Residual standard error: 1.385 on 502 degrees of freedom
Multiple R-squared:  0.04636,   Adjusted R-squared:  0.01976
F-statistic: 1.743 on 14 and 502 DF,  p-value: 0.04443
```

- Season의 summer범주에 해당하는 더미변수와 wind 변수가 종속변수 log(area)에 유의한 영향을 지닌다는 것을 확인할 수 있다. 하지만 전체적으로 단순 선형 모델로는 유의한 설명력을 가지는 변수를 특정 짓기 쉽지 않다. R^2값 또한 0.04 수준으로 아쉬운 결과를 보인다.
- 다음으로 5개의 train, test set에 대한 모델의 prediction accuracy를 구해본다. 앞서 언급한 바와 같이 prediction accuracy는 Mean Squared Error($\sum(True\ value - Predicted\ value)^2 / N$)를 활용하여 구했다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.808308 1.873899 1.893652 1.80175 1.862369 ]
5-fold test prediction MSEs are [ 2.090801 1.861928 1.766396 2.424066 1.903235 ]
```

Average MSEs for Training set & Test set

```
> cv_result_lin_reg
$Train_prediction_MSE
[1] 1.847995

$Test_prediction_MSE
[1] 2.009286
```

- MSE값에 대해서는 추후 모든 모델의 MSE를 다 구한 후 일괄적으로 비교하기로 한다.

[robust m Regression Predictor]

- 일반 선형 회귀에 비해 outlier에 robust한 M-회귀 모델이다.

Fitted Model

```
Call: rlm(formula = log_area ~ X + Y + season + week + FFMC + DMC +
    DC + ISI + temp + RH + wind + rain, data = forest_fires)
Residuals:
    Min      1Q  Median      3Q     Max
-1.6664 -0.9424 -0.4613  0.9510  5.6989

Coefficients:
             Value   Std. Error t value
(Intercept)  -1.4169  1.4151    -1.0013
X             0.0298  0.0287     1.0370
Y             0.0225  0.0542     0.4160
seasonspring  0.1587  0.4455     0.3562
seasonsummer -0.2313  0.1800    -1.2848
seasonwinter  0.9960  0.4570     2.1792
weekweekend  -0.0380  0.1139    -0.3333
FFMC          0.0132  0.0134     0.9822
DMC           0.0013  0.0015     0.8698
DC            0.0005  0.0007     0.8044
ISI          -0.0097  0.0156    -0.6195
temp          0.0208  0.0187     1.1090
RH           -0.0008  0.0052    -0.1453
wind          0.0739  0.0335     2.2085
rain          0.0767  0.1932     0.3969

Residual standard error: 1.398 on 502 degrees of freedom
```

- p-value가 나와있지 않지만 회귀계수의 t-value가 자유도 502인 t분포값들이 나왔기 때문에 근사 정규분포라 가정하고 1.96과 |t|값들을 비교해보면 변수의 유의성을 알 수 있다.
- 이번에는 season winter범주의 더미와 wind변수가 유의한 영향력을 가지게 되었다. 여전히 season변수와 wind변수가 유의하다는 점에서 앞선 선형회귀의 결과와 같다고 할 수 있다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.824827 1.897031 1.915335 1.829826 1.889676 ]
5-fold test prediction MSEs are [ 2.109917 1.860285 1.731038 2.37398 1.947608 ]>
```

Average MSEs for Training set & Test set

```
> cv_result_m_reg
$Train_prediction_MSE
[1] 1.871339

$Test_prediction_MSE
[1] 2.004565
```

[robust LMS Regression Predictor]

- M-회귀 모델보다 더 이상치에 robust한 Least Median of Squares회귀이다.
- r_i를 i 번째 개체의 residual이라 정의했을 때, LMS 회귀는 $median(\sum {r_i}^2)$을 최소화하는 coefficient를 찾는다. 그렇게 구한 회귀계수는 다음과 같다.

Fitted Model Coefficients

```
> lms_reg$coefficients
  (Intercept)             X             Y    seasonspring    seasonsummer    seasonwinter     weekweekend            FFMC             DMC
-20.075717788   0.062923663   0.551194368   1.623588309   1.170201708   3.808987963  -0.487319315   0.128820284  -0.015165065
           DC           ISI          temp            RH          wind          rain
  0.005257734  -0.051861936   0.120504197   0.054491002   0.091288542  -4.056009792
```

- 알고리즘의 특성에 따라 각 변수의 유의성 검정을 할 수 없다.
- 다음은 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 3.028648 3.221465 3.523548 4.089457 4.845039 ]
5-fold test prediction MSEs are [ 3.200413 3.218615 4.084075 5.076959 3.798557 ]>
```

Average MSEs for Training set & Test set

```
> cv_result_lms_reg
$Train_prediction_MSE
[1] 3.741631

$Test_prediction_MSE
[1] 3.875724
```

- 아직 모든 모델의 MSE가 구해진 것은 아니나 이미 앞의 두 모델에 비해 열등한 결과가 나왔다.

[robust LTS Regression Predictor]

- 마지막 robust 회귀모델인 Least Trimmed Squares 회귀이다.
- r_i를 i 번째 개체의 residual이라 정의했을 때, LTS 회귀는 가장 큰 잔차 중 일정 개수를 제외한 잔차들의 합을 최소화하는 coefficient를 찾는다. 그렇게 구한 회귀계수는 다음과 같다.

Fitted Model Coefficients

```
  (Intercept)             X             Y    seasonspring    seasonsummer    seasonwinter     weekweekend            FFMC             DMC
11.3409467348 -0.1802826548 -0.0895358510   0.1818658798   0.7054485064   1.2597278582   1.1631043718 -0.1147613887   0.0002175376
           DC           ISI          temp            RH          wind          rain
 0.0006096049   0.0895100239 -0.0551889175 -0.0281806559   0.1733816624   2.2850219490
```

- 알고리즘의 특성에 따라 각 변수의 유의성 검정을 할 수 없다.
- 다음은 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 3.383389 4.344592 3.187911 5.408859 4.150139 ]
5-fold test prediction MSEs are [ 4.018065 4.565047 2.94193 7.24513 4.285379 ]>
```

Average MSEs for Training set & Test set

```
> cv_result_lts_reg
$Train_prediction_MSE
[1] 4.094978

$Test_prediction_MSE
[1] 4.61111
```

- 앞선 LMS의 결과보다도 열등한 결과가 나왔다.

[Quantile Regression Predictor]

- median을 예측하는 분위수 회귀 모델이다. 그 결과는 다음과 같다.

Fitted Model

```
Call: rq(formula = log_area ~ X + Y + season + week + FFMC + DMC +
    DC + ISI + temp + RH + wind + rain, tau = 0.5, data = forest_fires)

tau: [1] 0.5

Coefficients:
             coefficients lower bd upper bd
(Intercept)  -1.99502     -3.64229 -0.15847
X             0.02621     -0.05377  0.09143
Y             0.04935     -0.06173  0.14549
seasonspring  0.00764     -1.31588  0.76042
seasonsummer -0.30725     -0.67030  0.46302
seasonwinter  1.60965      0.14049  2.48750
weekweekend  -0.12079     -0.38014  0.27228
FFMC          0.01338     -0.00784  0.02261
DMC           0.00205     -0.00470  0.00828
DC            0.00058     -0.00146  0.00194
ISI          -0.00450     -0.04624  0.03865
temp          0.01792     -0.02509  0.05121
RH           -0.00068     -0.00569  0.01178
wind          0.08267      0.01548  0.17526
rain          0.23528     -3.40964  0.24316
```

- 알고리즘의 특성에 따라 각 변수의 유의성 검정을 할 수 없다.
- 다음은 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 2.152117 2.116039 2.190261 2.126661 2.18833 ]
5-fold test prediction MSEs are [ 2.390955 2.112401 2.01335 2.629205 2.304598 ]>
```

Average MSEs for Training set & Test set

```
> cv_result_quant_reg
$Train_prediction_MSE
[1] 2.154682

$Test_prediction_MSE
[1] 2.290102
```

- robust 회귀들 보다는 양질의 prediction이 이루었다.

[LOESS Predictor]

- 국소적 회귀 Local Polynomial Regression 기법이다. 추정을 함에 있어 데이터의 구간을 나누어 미분 가능한 식을 적합하는 형태로 모델이 적합된다.
- LOESS에는 이산형 변수가 들어갈 수 없고 변수의 수 또한 3개를 넘을 수 없기에 앞선 EDA결과 가장 종속변수와 유의한 관계를 가졌던 X, DMC, wind 변수를 설명변수로 입력하였다.

Fitted Model

```
Call:
loess(formula = log_area ~ X + DMC + wind, data = forest_fires)

Number of Observations: 517
Equivalent Number of Parameters: 18.53
Residual Standard Error: 1.389
Trace of smoother matrix: 22.75  (exact)

Control settings:
  span     :  0.75
  degree   :  2
  family   :  gaussian
```

- 실제 사용된 모수의 개수는 18.53개로 선형회귀에서 통상적으로 더미변수를 포함해 15개의 모수가 추정된 것에 비해 많은 모수가 사용되었다. 구간 별로 다항식이 적합되었기 때문이다.

- 다음은 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.764041 1.813484 1.856781 1.796107 1.808089 ]
5-fold test prediction MSEs are [ 2.156686 2.064901 1.812828 2.150201 1.955262 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.8077

$Test_prediction_MSE
[1] 2.027976
```

- train set prediction accuracy에서 지금까지 중 가장 좋은 결과가 나왔다. 이는 선형적인 관계로 파악하지 못한 설명변수와 종속변수 간의 관계성이 구간별 회귀로 잘 설명된 결과일 것이다.
- test set prediction accuracy 또한 아주 좋은 결과가 나왔다.

[LOESS Predictor 2 (adjusted model)]

- LOESS 회귀에 option을 수정하여 데이터에 좀 더 적합한 형태로 바꾼 모델이다.
- 구간을 앞선 모델보다 더 크게 설정하였고 구간 별로 2차식이 아닌 선형 1차식을 적합하였다.

Fitted Model

```
> summary(loess2_reg)
Call:
loess(formula = log_area ~ X + DMC + wind, data = forest_fires,
    span = 1, degree = 1)

Number of Observations: 517
Equivalent Number of Parameters: 4.83
Residual Standard Error: 1.389
Trace of smoother matrix: 5.89  (exact)

Control settings:
  span     :  1
  degree   :  1
  family   :  gaussian
```

- 실제 사용된 모수의 개수는 4.83개로 굉장히 적은 수의 모수만이 추정 되었다.
- 다음은 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.845432 1.905318 1.94336 1.85795 1.917409 ]
5-fold test prediction MSEs are [ 2.136375 1.930709 1.747844 2.134093 1.858742 ]>
```

Average MSEs for Training set & Test set

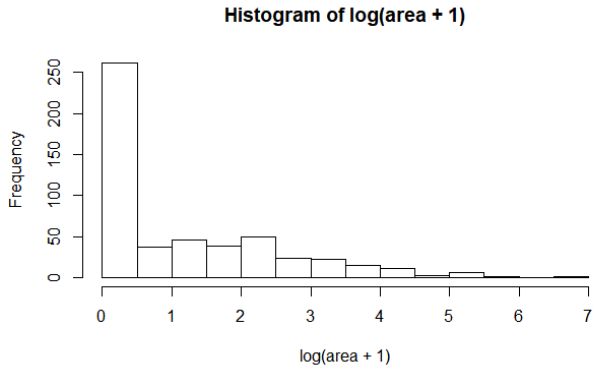
```
$Train_prediction_MSE
[1] 1.893894

$Test_prediction_MSE
[1] 1.961553
```

- 굉장히 적은 수의 모수만이 추정되었음에도 불구하고 train set prediction accuracy는 앞선 LOESS 모델과 비슷한 수준에, test set prediction accuracy는 개량된 결과가 나왔다.
- 이는 아마 구간별 회귀 적합 시 앞선 모델처럼 구간을 촘촘하게 잡을 필요도 없을 뿐더러 복잡한 다항식이 아닌 선형 식으로만 적합하여도 충분히 데이터가 잘 설명된다는 뜻이다.

[Generalized Additive Model Predictor]

- 일반화 가법 모형 일명 GAM을 활용한 semi-parametric regression 모델이다. 종속변수 log(area+1)의 분포가 감마 분포의 형태를 띄고 있기 때문에 gamma분포를 가정한 log regression GAM을 적합하였다. 또한 연속형 변수에 대해서는 smooth 비선형 변환을 해주고 이산형 변수에 대해서는 변환 없이 그대로 입력하였기에 semi-parametric GAM 모델이 된다.

Non-linear gamma shape target variable



- 다음은 전체 데이터를 모델에 적합한 결과이다.

Fitted Model

```
Family: Gamma
Link function: log

Formula:
(area + 1) ~ s(X, k = 9) + s(Y, k = 6) + season + week + s(FFMC) +
    s(DMC) + s(DC) + s(ISI) + s(temp) + s(RH) + s(wind) + s(rain,
    k = 4)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.5680     0.3046   8.430 4.13e-16 ***
seasonspring   0.3326     0.9464   0.351   0.7254
seasonsummer  -1.6152     0.4058  -3.981 7.95e-05 ***
seasonwinter   1.2309     0.8511   1.446   0.1488
weekweekend    0.2899     0.1744   1.662   0.0971 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df     F  p-value
s(X)    7.708  7.970 3.714 0.000314 ***
s(Y)    4.669  4.911 2.005 0.085932 .
s(FFMC) 1.000  1.000 2.030 0.154919
s(DMC)  8.233  8.802 3.917 0.000123 ***
s(DC)   5.591  6.736 2.345 0.029688 *
s(ISI)  1.000  1.000 3.418 0.065091 .
s(temp) 1.000  1.000 7.642 0.005924 **
s(RH)   1.000  1.000 1.587 0.208381
s(wind) 2.613  3.293 4.097 0.005256 **
s(rain) 2.891  2.988 2.537 0.057661 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0257   Deviance explained = 34.6%
GCV = 2.3394  Scale est. = 3.4012    n = 517
```

- 이산형 변수에서는 season summer에 해당하는 변수가 연속형 변수에서는 비선형 변환 덕분인지 X, DMC, DC, temp, wind 등 많은 변수가 유의한 영향력을 가지는 것이 확인되었다. 확실히 forest fires의 설명변수는 종속변수와 비선형적 관계를 띄는 것을 확인할 수 있다.
- 다음은 다섯 train-test set에 대한 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 2.537335 2.95578 3.28843 3.039655 2.982559 ]
5-fold test prediction MSEs are [ 3.798601 3.425191 3.20318 4.349078 3.708556 ]
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 2.960752

$Test_prediction_MSE
[1] 3.696921
```

- 많은 변수가 유의성을 띈 것에 비해 train set과 test set의 평균 prediction accuracy의 결과는 좋은 편이 아니다.

[Generalized Additive Model Predictor 2 (adjusted)]

- 비록 종속변수의 형태가 gaussian 분포의 형태를 띈다고 보기는 힘들지만 log link function 대신 identity link를 사용하는 semi-parametric GAM을 적합하였다. 설명변수의 입력형태는 첫번째 GAM의 그것과 같다.

Fitted Model

```
Family: gaussian
Link function: identity

Formula:
log_area ~ s(X, k = 9) + s(Y, k = 6) + season + week + s(FFMC) +
    s(DMC) + s(DC) + s(ISI) + s(temp) + s(RH) + s(wind) + s(rain,
    k = 4)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.292484   0.183238   7.054 5.86e-12 ***
seasonspring  0.101729   0.524696   0.194   0.8463
seasonsummer -0.532387   0.219617  -2.424   0.0157 *
seasonwinter  0.788453   0.559495   1.409   0.1594
weekweekend  -0.003113   0.124943  -0.025   0.9801
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(X)    1.741  2.171 1.744  0.1766
s(Y)    1.506  1.854 0.353  0.7231
s(FFMC) 1.000  1.000 1.084  0.2983
s(DMC)  1.512  1.869 2.081  0.1208
s(DC)   1.762  2.245 0.837  0.3983
s(ISI)  1.000  1.000 1.491  0.2226
s(temp) 1.788  2.277 2.207  0.1135
s(RH)   1.000  1.000 0.020  0.8867
s(wind) 1.000  1.000 5.217  0.0228 *
s(rain) 1.766  1.995 1.200  0.2988
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0417   Deviance explained = 7.52%
GCV = 1.9459  Scale est. = 1.8741    n = 517
```

- 첫번째 GAM 모델에 비해 유의성을 띄는 변수의 수도 적고 Deviance explained 또한 낮아졌지만 adjusted R^2값만은 증가하였다. 통상적으로 adjusted R^2 값은 모델이 얼마나 적합한지와 설명변수의 redundancy를 함께 측정하므로 좀 더 효율적인 모델이라 할 수 있다.
- 다음은 다섯 train-test set에 대한 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.742259 1.79546 1.859695 1.631028 1.792323 ]
5-fold test prediction MSEs are [ 2.104275 1.890528 1.695161 2.219513 1.904601 ]
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.764153

$Test_prediction_MSE
[1] 1.962816
```

- 최적의 두번째 LOESS 보다 좋은 prediction 결과가 나왔다.
- 지금까지 적합된 모델 중 설명변수와 종속변수의 관계성을 가장 최적으로 설명하는 모델이라 할 수 있다.

[Support Vector Regression predictor]

- 대표적인 지도학습기계인 Support Vector Machine 방법론을 활용한 회귀 적합 모델이다. 정확히는 가우스 커널(일명 radial kernel)을 활용한 Kernel SVM Regression 방법이다. Regularization term에 부과되는 cost는 1로 지정하였고 kernel의 gamma는 0.15로 조정하였다.

Fitted Model

```
Call:
svm(formula = log_area ~ X + Y + season + week + FFMC + DMC + DC + ISI + temp + RH + wind + rain, data = train_data,
    kernel = "radial", gamma = 0.15, cost = 0.75)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  0.75
      gamma:  0.15
    epsilon:  0.1


Number of Support Vectors:  382
```

- 다음은 다섯 train-test set에 대한 prediction accuracy이다.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.43052 1.525687 1.617377 1.637805 1.531342 ]
5-fold test prediction MSEs are [ 2.476185 2.015512 1.895581 2.223201 2.193219 ]
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.548546

$Test_prediction_MSE
[1] 2.16074
```

- training set prediction accuracy가 지금까지의 모델 중 가장 높았고 test set prediction accuracy 또한 준수한 결과를 냈다.

[방법론 총 비교]

모든 방법론의 prediction accuracy를 한번에 비교하여 보자.

```
> Result_data
                    Linear   M-reg.   LMS-reg. LTS-reg. Quantile LOESS    LOESS(adjusted) GAM(Gamma) GAM      SVM
Train_prediction_MSE 1.847995 1.871339 3.741631 4.094978 2.154682 1.8077   1.893894        2.960752   1.764153 1.548546
Test_prediction_MSE  2.009286 2.004565 3.875724 4.61111  2.290102 2.027976 1.961553        3.696921   1.962816 2.16074
```

-아까도 언급하였듯이 adjusted LOESS 모델이 predictor로서 가장 좋은 성능을 보였다.

4. PCA를 통한 데이터 차원 축소 및 통계적 모델 적합

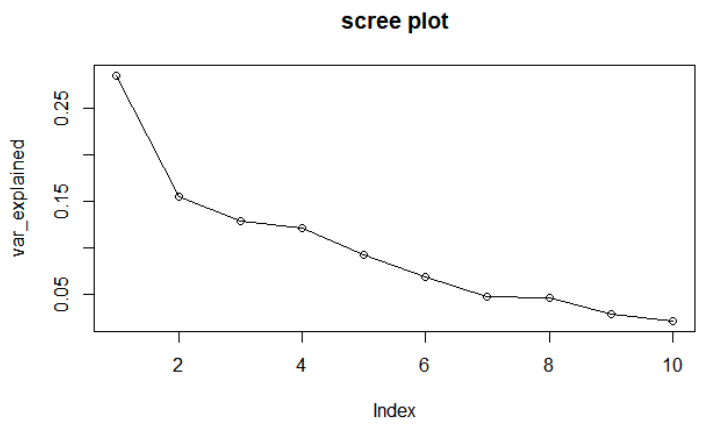
[Principal Component Analysis]

- 통계적 차원 축소 알고리즘으로 가장 자주 거론되는 것이 바로 classical multidimensional scaling이라고도 불리는 주성분 분석(Principal Component Analysis)일 것이다. 그 이유는 PCA가 가지는 많은 장점들 때문인데 먼저 PCA는 데이터 행렬에 대한 고유값-고유행렬 분해를 통해 이루어지기 때문에 선형대수 이론적 근거가 확실하다. 또한 PCA를 통해 생성된 주성분들은 새로운 변수로 쓰일 때, PC loadings 정보를 통해 기존 변수와의 연관성을 확인할 수 있다. 이를 통해 새로운 변수로서 주성분들이 어떠한 의미를 가지는지에 대한 interpretability가 생긴다. 즉 잠재변수 측정 방법으로서 PCA가 활용될 수 있는 것이다.
- EDA와 앞선 통계모형 적합을 통해 알아낸 바와 같이 설명변수 중에서 연속형 변수가 아주 많은 것에 비해 그 설명력이 너무나도 낮게 나온다. 따라서 연속형 설명변수에 대해 PCA를 적용하여 축소된 차원의 잠재변수를 활용한 통계적 모델링 결과는 어떻게 나오는지에 대해 확인하였다.

Principal Components generated by PCA

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9     PC
Standard deviation     1.691 1.2488 1.1385 1.1011 0.96515 0.83224 0.68963 0.68051 0.54286 0.463
Proportion of Variance 0.286 0.1559 0.1296 0.1212 0.09315 0.06926 0.04756 0.04631 0.02947 0.021
Cumulative Proportion  0.286 0.4419 0.5715 0.6928 0.78592 0.85518 0.90274 0.94905 0.97852 1.000
```

Scree Plot



PC loadings

	PC1	PC2	PC3	PC4	PC5
X	-0.07916725	0.67783265	-0.1251601	0.081582786	-0.05471802
Y	-0.07187754	0.67031117	-0.1181043	0.168962834	-0.07391155
FFMC	0.42442495	0.06128114	-0.2032123	-0.232991392	-0.09698043
DMC	0.42938051	0.12272472	0.4362328	0.006512088	-0.15887594
DC	0.43358454	0.01268424	0.3982471	0.155859707	-0.08502588
ISI	0.35990798	0.10979234	-0.1694431	-0.433785355	-0.19202801
temp	0.48475331	0.01579246	-0.2023952	0.181924300	0.18122423
RH	-0.23192607	0.15547790	0.6771104	-0.193305099	-0.19197759
wind	-0.12384067	0.02193112	-0.1157121	-0.708102520	-0.24010742
rain	0.04920077	0.18785657	0.1965926	-0.355153405	0.88625217

- 연속형 변수에 대한 PCA를 통해 생성된 주성분 중 다섯개의 주성분이 전체 분산의 약 79% 이

상을 설명하므로 다섯개의 PC에 대한 PC-scores를 새로운 변수로서 분석에 활용하도록 한다.

- PC loadings를 통해 새롭게 생성된 다섯개의 주성분들이 기존 변수와 어떠한 관계를 가지는 지 알 수 있다. 이로써 새로운 다섯개의 잠재변수의 특성을 파악하거나 명명을 할 수도 있지만 여기서는 우선 주성분 활용에만 초점을 두도록 한다.

[PCA data에 모델 적합]

- 위의 원 데이터에 적용했던 모든 방법론을 전부 새로운 데이터에 적용한다.
- 각 통계적 모델에 대한 설명과 해석은 위에서 했으므로 여기서는 결과만 간단하게 정리하고 최종 결과를 비교하는 단계로 넘어가도록 한다.

[Linear Regression Predictor]

Fitted Model

```
Call:
lm(formula = log_area ~ PC1 + PC2 + PC3 + PC4 + PC5 + season +
    week, data = pc_var)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8504 -1.0749 -0.5993  0.8717  5.6037

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.219610   0.128370   9.501   <2e-16 ***
PC1           0.129212   0.064292   2.010   0.0450 *
PC2           0.091910   0.049924   1.841   0.0662 .
PC3          -0.009319   0.064448  -0.145   0.8851
PC4          -0.046160   0.058885  -0.784   0.4335
PC5          -0.014638   0.064211  -0.228   0.8198
seasonspring -0.099559   0.311134  -0.320   0.7491
seasonsummer -0.305770   0.142160  -2.151   0.0320 *
seasonwinter  0.720435   0.387816   1.858   0.0638 .
weekweekend  -0.002863   0.123439  -0.023   0.9815
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.39 on 507 degrees of freedom
Multiple R-squared:  0.02928,   Adjusted R-squared:  0.01205
F-statistic: 1.699 on 9 and 507 DF,  p-value: 0.08637
```

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.763521 1.995803 2.032021 1.748124 1.883379 ]
5-fold test prediction MSEs are [ 2.213753 1.375412 1.272515 2.612927 1.964253 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.88457

$Test_prediction_MSE
[1] 1.887772
```

[robust m Regression Predictor]

Fitted Model

```
Call: rlm(formula = log_area ~ PC1 + PC2 + PC3 + PC4 + PC5 + season +
    week, data = pc_var)
Residuals:
    Min      1Q  Median      3Q     Max
-1.7275 -0.9336 -0.4554  0.9770  5.8032

Coefficients:
             Value   Std. Error t value
(Intercept)   1.0860  0.1161     9.3573
PC1           0.1157  0.0581     1.9907
PC2           0.0819  0.0451     1.8147
PC3           0.0050  0.0583     0.0856
PC4          -0.0601  0.0532    -1.1292
PC5          -0.0128  0.0581    -0.2209
seasonspring -0.0754  0.2813    -0.2680
seasonsummer -0.2836  0.1285    -2.2066
seasonwinter  0.7956  0.3506     2.2690
weekweekend  -0.0271  0.1116    -0.2427

Residual standard error: 1.398 on 507 degrees of freedom
```

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.781107 2.011705 2.056834 1.763473 1.902314 ]
5-fold test prediction MSEs are [ 2.305824 1.375735 1.277162 2.675272 1.948387 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.903086

$Test_prediction_MSE
[1] 1.916476
```

[robust LMS Regression Predictor] .

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 2.695351 2.93466 2.972185 2.373916 2.642011 ]
5-fold test prediction MSEs are [ 3.498788 2.447339 3.456042 11.85895 3.438789 ]
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 2.723624

$Test_prediction_MSE
[1] 4.939982
```

[robust LTS Regression Predictor]

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 2.756696 3.311972 3.368969 3.014798 3.005463 ]
5-fold test prediction MSEs are [ 4.767565 3.507679 3.381874 5.690134 3.704096 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 3.091579

$Test_prediction_MSE
[1] 4.21027
```

[Quantile Regression Predictor]

Fitted Model

```
Call: rq(formula = log_area ~ PC1 + PC2 + PC3 + PC4 + PC5 + season +
    week, tau = 0.5, data = pc_var)

tau: [1] 0.5

Coefficients:
             coefficients lower bd upper bd
(Intercept)   0.79340      0.32554  1.17134
PC1           0.11676     -0.01631  0.30209
PC2           0.13698      0.02446  0.22853
PC3           0.03495     -0.13710  0.16939
PC4          -0.07484     -0.19710  0.01046
PC5           0.00809     -0.52698  0.14415
seasonspring -0.35042     -0.93539  0.28056
seasonsummer -0.37491     -0.71733  0.13462
seasonwinter  1.38853     -0.06304  2.27433
weekweekend  -0.08062     -0.37824  0.19875
```

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 2.069769 2.362475 2.347073 1.983172 2.211748 ]
5-fold test prediction MSEs are [ 3.011155 1.534005 1.722165 3.503223 2.178436 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 2.194847

$Test_prediction_MSE
[1] 2.389797
```

[LOESS Predictor]

Fitted Model

```
Call:
loess(formula = log_area ~ PC1 + PC2 + PC3, data = pc_var)

Number of Observations: 517
Equivalent Number of Parameters: 17.32
Residual Standard Error: 1.414
Trace of smoother matrix: 21.16  (exact)

Control settings:
  span     :  0.75
  degree   :  2
  family   :  gaussian
```

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.738108 1.990836 1.99725 1.751297 1.954409 ]
5-fold test prediction MSEs are [ 2.504814 1.498481 1.765063 2.796652 1.91876 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.88638

$Test_prediction_MSE
[1] 2.096754
```

[LOESS Predictor 2 (adjusted model)]

Fitted Model

```
loess(formula = log_area ~ PC1 + PC2 + PC3, data = pc_var, span = 1,
    degree = 1)

Number of Observations: 517
Equivalent Number of Parameters: 4.48
Residual Standard Error: 1.399
Trace of smoother matrix: 5.26  (exact)

Control settings:
  span     :  1
  degree   :  1
  family   :  gaussian
```

.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.797615 2.02849 2.066625 1.799016 1.93965 ]
5-fold test prediction MSEs are [ 2.396507 1.462987 1.518757 2.605126 1.910331 ]
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.926279

$Test_prediction_MSE
[1] 1.978741
```

[Generalized Additive Model Predictor]

Fitted Model

```
Family: Gamma
Link function: log

Formula:
(area + 1) ~ s(PC1) + s(PC2) + s(PC3) + s(PC4) + s(PC5) + season +
    week

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.63228    0.23386  11.256  < 2e-16 ***
seasonspring -0.52500    0.63267  -0.830  0.40705
seasonsummer -0.79164    0.25811  -3.067  0.00228 **
seasonwinter  0.80001    0.81191   0.985  0.32495
weekweekend   0.01219    0.21383   0.057  0.95456
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(PC1) 2.464  3.148 0.657   0.563
s(PC2) 7.798  8.512 1.760   0.140
s(PC3) 4.884  5.958 1.047   0.413
s(PC4) 7.014  7.730 1.532   0.144
s(PC5) 4.725  5.809 0.716   0.634

R-sq.(adj) =  0.00871   Deviance explained =   24%
GCV = 2.6183  Scale est. = 5.5374    n = 517
```

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 2.725166 3.568209 3.618276 3.033779 3.351518 ]
5-fold test prediction MSEs are [ 3.345004 3.997672 3.961354 16.73209 4.253914 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 3.259389

$Test_prediction_MSE
[1] 6.458006
```

[Generalized Additive Model Predictor 2 (adjusted)]

Fitted Model

```
Family: gaussian
Link function: identity

Formula:
log_area ~ s(PC1) + s(PC2) + s(PC3) + s(PC4) + s(PC5) + season +
    week

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.194306   0.128964   9.261   <2e-16 ***
seasonspring  0.001449   0.320595   0.005   0.9964
seasonsummer -0.300782   0.142751  -2.107   0.0356 *
seasonwinter  0.717402   0.386664   1.855   0.0641 .
weekweekend   0.017775   0.123736   0.144   0.8858
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(PC1) 1.000  1.000 3.762   0.053 .
s(PC2) 1.000  1.000 1.781   0.183
s(PC3) 3.992  5.028 0.817   0.537
s(PC4) 1.000  1.000 0.026   0.871
s(PC5) 1.581  1.901 0.418   0.587
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0198   Deviance explained = 4.37%
GCV = 1.9686  Scale est. = 1.9169    n = 517
```

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.73099 1.982399 1.995267 1.748124 1.784658 ]
5-fold test prediction MSEs are [ 2.351014 1.442195 1.526589 2.879236 2.132343 ]
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.848288

$Test_prediction_MSE
[1] 2.066275
```

[Support Vector Regression predictor]

Fitted Model

```
Call:
svm(formula = log_area ~ PC1 + PC2 + PC3 + PC4 + PC5 + season + week, data = pc_var, kernel = "radial", gamma = 0.15,
    cost = 0.75)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  0.75
      gamma:  0.15
    epsilon:  0.1


Number of Support Vectors:  479
```

.

5 Prediction Accuracies (MSEs)

```
5-fold train prediction MSEs are [ 1.83954 2.082743 2.039701 1.723442 1.889962 ]
5-fold test prediction MSEs are [ 2.169061 1.251591 1.335119 2.40718 2.132343 ]>
```

Average MSEs for Training set & Test set

```
$Train_prediction_MSE
[1] 1.915078

$Test_prediction_MSE
[1] 1.859059
```

[방법론 총 비교]

- 원본 Forest Fires 데이터를 사용하여 모델을 적합해 구한 prediction accuracy와 PCA로 구한 축소된 차원의 데이터를 사용하여 모델에 적합해 구한 예측기의 prediction accuracy를 한번에 정리하면 다음과 같다.

```
> Result_data
                    Linear   M-reg.   LMS-reg. LTS-reg. Quantile LOESS    LOESS(adjusted) GAM(Gamma) GAM      SVM
Train_prediction_MSE 1.847995 1.871339 3.741631 4.094978 2.154682 1.8077   1.893894        2.960752   1.764153 1.548546
Test_prediction_MSE  2.009286 2.004565 3.875724 4.61111  2.290102 2.027976 1.961553        3.696921   1.962816 2.16074
> Result_PC
                    Linear   M-reg.   LMS-reg. LTS-reg. Quantile LOESS    LOESS(adjusted) GAM(Gamma) GAM      SVM
Train_prediction_MSE 1.88457  1.903086 2.723624 3.091579 2.194847 1.88638  1.926279        3.259389   1.848288 1.915078
Test_prediction_MSE  1.887772 1.916476 4.939982 4.21027  2.389797 2.096754 1.978741        6.458006   2.066275 1.859059
```

4 토의점과 결론

전체 비교 결과를 보면 linear regression과 robust regression 등에서는 PCA를 통해 축소된 데이터의 prediction accuracy가 train set에서도 test set에서도 낮게 나온 것을 확인할 수 있다. 하지만 비선형적 특성을 설명하는 통계적 모델을 활용한 predictor 들에서는 딱히 축소된 데이터에서 더 나은 결과를 냈다고 보기 어렵다.

이는 아까도 말했듯이 PCA가 선형대수적 배경을 가지고 있어서 선형적 정보에 대한 정보 축약은 성공적으로 해냈지만 비선형적 특성에 대한 정보는 오히려 보존하지 못하고 유실된 결과라고 추측된다. 하지만 또 결과적으로 가장 좋은 prediction accuracy를 낸 predictor는 PCA로 축소된 데이터를 사용하여 kernel Support Vector Machine을 활용한 예측기이다.

따라서 통계적 모델을 활용한 predictor를 만들 때에는 최적의 모델과 데이터의 입력 형태가 정해진 것이 아니다. 그러므로 다양한 형태의 시도를 통해 경험적으로 최적의 모델을 찾아가는 노력 또한 중요한 요소중의 하나라고 할 수 있다.

추가적으로 SVM이나 LOESS와 같은 모델들은 분류기로서의 성능은 좋으나 설명변수들의 coefficient가 구해지지 않는다는 점에서 interpretability가 전무한 수준이다. 따라서 서론에 기술한 요인에 대한 분석과 통제를 위해서는 이를 고려하려 prediction accuracy만 맹목적으로 최적화하는 것이 아닌 어느 정도의 설명력을 가지는 통계적 모델도 함께 고려함이 마땅하다.

[Reference]

- 이재원/박미라/유한나, 2005, *생명과학연구를 위한 통계적 방법*, 자유아카데미 출판
- 허명회, 2014, *응용통계데이터분석 (Applied Data Analysis Using R),* 자유아카데미 출판
- 송문섭/박창순/김홍기, 2015, *R과 함께 비모수 통계학*, 자유아카데미 출판