<다변량분석 리포트#5>

2014170852 산업경영공학부 조영관

<Training set과 Validation set으로 나누기>

250개의 training set과 71개의 validation set으로 나누었다.

<모든 변수를 사용해 MLR 모형 학습하기>

모든 변수를 이용하여 MLR 모형을 학습시킨 결과는 다음과 같았다.

```
Call:
lm(formula = Mean_temperature ~ ., data = trn)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9198 -0.9771 -0.1229  0.9206  5.7278

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       73.57716   23.65907   3.110  0.00210 **
Max_termperature   0.55839    0.01563  35.716  < 2e-16 ***
Min_temperature    0.31742    0.02596  12.227  < 2e-16 ***
Dewpoint           0.07527    0.02735   2.752  0.00638 **
Precipitation     -0.05895    0.80743  -0.073  0.94186
Sea_level_pressure -8.37101   3.63043  -2.306  0.02198 *
Standard_pressure  6.60741    4.14384   1.595  0.11214
Visibility         0.06689    0.08099   0.826  0.40966
Wind_speed         0.10581    0.05708   1.854  0.06503 .
Max_wind_speed     0.01984    0.02473   0.802  0.42326
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.618 on 240 degrees of freedom
Multiple R-squared:  0.9893,    Adjusted R-squared:  0.9889
F-statistic:  2463 on 9 and 240 DF,  p-value: < 2.2e-16
```

Adjusted R2 값은 0.9889로 매우 높게 나왔다. (데이터셋의 품질이 매우 우수한 듯 보인다)

위의 개별 변수들에 대한 P-VALUE 값을 살펴보면, 유의수준 0.01에서 유의미한 변수는 max_temperature, min_temperature, dewpoint 이렇게 3가지가 있다.

학습한 모형을 이용해 validation set에 대한 RMSE, MAE, MAPE를 산출해보았다.

```
                        RMSE      MAE      MAPE
All                 1.577273 1.255487 2.910357
Exhaustive search   1.569378 1.245601 2.910040
Forward Selection   1.569378 1.245601 2.910040
Backward Elimination 1.569378 1.245601 2.910040
Stepwise Selection  1.569378 1.245601 2.910040
Genetic Algorithm   1.569378 1.245601 2.910040
```

맨 위 ALL 항목 (모든 변수를 사용)을 살펴보면, RMSE는 1.577, MAE는 1.255, MAPE는 2.91 이 나온 것을 확인할 수 있다.

<Exhaustive search를 통해 얻어진 R2 값이 가장 높은 변수 집합 추출, Performance 측정>

직접 함수를 구현하여 Exhaustive search를 수행하였다.

그 결과 제일 R2가 높은 변수 조합은 아래 결과와 같았다.

```
Call:
lm(formula = best_formula, data = trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0046 -0.9892 -0.1080  0.9725  5.7280

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.63148   23.24824   3.210  0.00151 **
Max_termperature   0.56279    0.01474  38.193  < 2e-16 ***
Min_temperature    0.31395    0.02549  12.318  < 2e-16 ***
Dewpoint           0.07331    0.02637   2.780  0.00586 **
Sea_level_pressure -8.83805    3.54336  -2.494  0.01329 *
Standard_pressure  7.10933    4.05478   1.753  0.08081 .
Wind_speed         0.14508    0.04403   3.295  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.612 on 243 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.989
F-statistic:  3720 on 6 and 243 DF,  p-value: < 2.2e-16
```

3개의 입력변수를 제외한 위의 6개의 입력 변수 조합이 R2 값이 제일 높았다.

Max_temperature, min_temperature, dewpoint, sea_level_pressure, standard_pressure, wind_speed가 유효한 변수로 선택되었다.

```
사용자  시스템 elapsed
  1.10    0.00    4.35
```

위 결과는 소요시간이다. 원래 Exhaustive Search는 제일 오래 걸리는 search 방식이지만 위의 데이터의 경우 데이터셋의 크기가 작고, 데이터셋의 품질도 좋아서 Exhaustive search를 하는데 그리 많은 시간이 소요되지 않았다.

Validation set에 대한 RMSE, MAE, MAPE 값을 산출해보았다.

```
                          RMSE      MAE      MAPE
All                   1.577273 1.255487 2.910357
Exhaustive search     1.569378 1.245601 2.910040
Forward Selection     1.569378 1.245601 2.910040
Backward Elimination 1.569378 1.245601 2.910040
Stepwise Selection    1.569378 1.245601 2.910040
Genetic Algorithm     1.569378 1.245601 2.910040
```

두 번째 항목인 Exhaustive search를 보면 된다. 결과는 모든 변수를 선택했을 때보다 RMSE, MAE, MAPE 모두 더 작게 나왔다. (참고로 MAPE는 100을 곱한 값)

<Forward selection, Backward Elimination, Stepwise Selection 방식 사용하여 변수 선택하기>

Forward selection 먼저 살펴보자.

```
Call:
lm(formula = Mean_temperature ~ Max_termperature + Min_temperature +
    Sea_level_pressure + Standard_pressure + Wind_speed + Dewpoint,
    data = trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0046 -0.9892 -0.1080  0.9725  5.7280

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.63148   23.24824   3.210  0.00151 **
Max_termperature   0.56279    0.01474  38.193  < 2e-16 ***
Min_temperature    0.31395    0.02549  12.318  < 2e-16 ***
Sea_level_pressure -8.83805    3.54336  -2.494  0.01329 *
Standard_pressure  7.10933    4.05478   1.753  0.08081 .
Wind_speed         0.14508    0.04403   3.295  0.00113 **
Dewpoint           0.07331    0.02637   2.780  0.00586 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.612 on 243 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.989
F-statistic:  3720 on 6 and 243 DF,  p-value: < 2.2e-16
```

Forward selection 역시, Exhaustive search를 통해 나온 결과와 동일한 결과가 나왔다.

동일하게 6가지 변수를 선택하였다. 선택된 변수도 동일하며 개수도 동일하다.

따라서 Adjusted R2 값도 동일하게 0.989 값이 나왔다.

```
 사용자  시스템 elapsed
   0.14    0.02    0.75
```

위의 결과를 보면 수행시간은 exhaustive search 보다 작게 나왔다.

Validation set을 이용해 RMSE, MAE, MAPE를 산출해보았다.

```
                         RMSE      MAE      MAPE
All                  1.577273 1.255487 2.910357
Exhaustive search    1.569378 1.245601 2.910040
Forward Selection    1.569378 1.245601 2.910040
Backward Elimination 1.569378 1.245601 2.910040
Stepwise Selection   1.569378 1.245601 2.910040
Genetic Algorithm    1.569378 1.245601 2.910040
```

3번째 항목의 Forward selection을 보면 된다. 동일한 변수가 선택되었으므로 결과값도 exhaustive search와 동일하게 나왔다.

다음은 backward elimination을 보자.

```
Call:
lm(formula = Mean_temperature ~ Max_termperature + Min_temperature +
    Dewpoint + Sea_level_pressure + Standard_pressure + Wind_speed,
    data = trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0046 -0.9892 -0.1080  0.9725  5.7280

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.63148   23.24824   3.210  0.00151 **
Max_termperature   0.56279    0.01474  38.193  < 2e-16 ***
Min_temperature    0.31395    0.02549  12.318  < 2e-16 ***
Dewpoint           0.07331    0.02637   2.780  0.00586 **
Sea_level_pressure -8.83805    3.54336  -2.494  0.01329 *
Standard_pressure   7.10933    4.05478   1.753  0.08081 .
Wind_speed          0.14508    0.04403   3.295  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.612 on 243 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.989
F-statistic:  3720 on 6 and 243 DF,  p-value: < 2.2e-16
```

Backward elimination도 동일하게 6개의 변수를 선택하였다. 선택된 변수들도 동일하다.

따라서 adjusted R2값은 0.989이다.

소요시간은 다음과 같다.

```
사용자  시스템 elapsed
  0.10    0.00    0.66
```

Forward와 거의 동일하다.

마지막으로 RMSE, MAE, MAPE 값 역시 Forward와 동일하다.

```
                       RMSE      MAE     MAPE
All                1.577273 1.255487 2.910357
Exhaustive search  1.569378 1.245601 2.910040
Forward Selection  1.569378 1.245601 2.910040
Backward Elimination 1.569378 1.245601 2.910040
Stepwise Selection 1.569378 1.245601 2.910040
Genetic Algorithm  1.569378 1.245601 2.910040
```

마지막으로 stepwise selection을 해보자.

```
Call:
lm(formula = Mean_temperature ~ Max_termperature + Min_temperature +
    Sea_level_pressure + Standard_pressure + Wind_speed + Dewpoint,
    data = trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0046 -0.9892 -0.1080  0.9725  5.7280

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.63148   23.24824   3.210  0.00151 **
Max_termperature   0.56279    0.01474  38.193  < 2e-16 ***
Min_temperature    0.31395    0.02549  12.318  < 2e-16 ***
Sea_level_pressure -8.83805    3.54336  -2.494  0.01329 *
Standard_pressure  7.10933    4.05478   1.753  0.08081 .
Wind_speed         0.14508    0.04403   3.295  0.00113 **
Dewpoint           0.07331    0.02637   2.780  0.00586 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.612 on 243 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.989
F-statistic:  3720 on 6 and 243 DF,  p-value: < 2.2e-16
```

선택된 변수는 exhaustive search, forward selection, backward elimination과 동일하다.

Adjusted R2값은 0.989 이다.

```
 사용자  시스템 elapsed
   0.13    0.00    0.62
```

수행시간은 위와 같다. Forward selection, backward elimination과 큰 차이는 없고, exhaustive search 보다는 짧게 걸리는 것을 알 수 있다.

마지막으로 RMSE, MAE, MAPE 값을 산출한 결과는 FS, BE, ES와 동일하였다.

```
                          RMSE      MAE      MAPE
All                   1.577273 1.255487 2.910357
Exhaustive search     1.569378 1.245601 2.910040
Forward Selection     1.569378 1.245601 2.910040
Backward Elimination 1.569378 1.245601 2.910040
Stepwise Selection    1.569378 1.245601 2.910040
Genetic Algorithm     1.569378 1.245601 2.910040
```

모든 변수를 선택했을 때 보다는 RMSE, MAE, MAPE 값이 작아졌다.

ES, FS, BE를 선택한 결과와는 동일하다.

<Adjusted R2를 Fitness function으로 지정 GA 기반 변수 선택 함수>

함수를 작성한 후 변수 선택을 해보았다.

```
GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
            names = colnames(x), popSize = 50, pcrossover = 0.5,
            pmutation = 0.01, maxiter = 100, elitism = 2, seed = 123)
```

옵션은 위와 같이 주었다.

소요시간은 다음과 같았다.

```
 사용자  시스템 elapsed
   4.92    0.09    5.91
```

FS, BE, SS 보다 길게 소요되었다.

원래는 Exhaustive search 보다는 적게 소요되어야 하는데, 데이터셋의 특성상 (데이터셋 크기가 작고, 품질이 좋고, R 자체 구동 시간?) exhaustive search 보다 오래 걸렸다.

그리고 변수 선택을 통해 MLR 모형을 학습한 결과를 살펴보자.

```
Call:
lm(formula = Mean_temperature ~ ., data = GA_trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0046 -0.9892 -0.1080  0.9725  5.7280

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.63148   23.24824   3.210  0.00151 **
Max_termperature   0.56279    0.01474  38.193  < 2e-16 ***
Min_temperature    0.31395    0.02549  12.318  < 2e-16 ***
Dewpoint           0.07331    0.02637   2.780  0.00586 **
Sea_level_pressure -8.83805    3.54336  -2.494  0.01329 *
Standard_pressure   7.10933    4.05478   1.753  0.08081 .
Wind_speed          0.14508    0.04403   3.295  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.612 on 243 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.989
F-statistic:  3720 on 6 and 243 DF,  p-value: < 2.2e-16
```

ES, FS, BE, SS과 같이 동일한 변수 6개를 선택하였다.

그래서 Adjusted R2값도 동일하게 0.989 이다.

Validation set에 대한 RMSE, MAE, MAPE 값을 산출해보았다.

```
                          RMSE      MAE      MAPE
All                   1.577273 1.255487 2.910357
Exhaustive search     1.569378 1.245601 2.910040
Forward Selection     1.569378 1.245601 2.910040
Backward Elimination 1.569378 1.245601 2.910040
Stepwise Selection    1.569378 1.245601 2.910040
Genetic Algorithm     1.569378 1.245601 2.910040
```

마지막의 Genetic Algorithm 항목을 보면 된다. 결과값이 모든 변수를 사용했을 때 보다는 RMSE, MAE, MAPE 값이 작게 나왔다. ES, FS, BE, SS와는 선택된 변수가 동일해서 결과값이 동일하게 나왔다.

<Genetic Algorithm에서 하이퍼 파라미터 변경해보기>

먼저 다른 옵션은 고정시키고 population size를 변화시켜보았다.

```
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 10, pcrossover = 0.1,
+             pmutation = 0.01, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.9117016 | Best = 0.9886382
GA | iter = 2 | Mean = 0.9351893 | Best = 0.9886382
GA | iter = 3 | Mean = 0.9884567 | Best = 0.9886382
GA | iter = 4 | Mean = 0.9885032 | Best = 0.9886382
GA | iter = 5 | Mean = 0.9885257 | Best = 0.9886382
GA | iter = 6 | Mean = 0.9885707 | Best = 0.9886382
GA | iter = 7 | Mean = 0.9885932 | Best = 0.9886382
GA | iter = 8 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 9 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 10 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 11 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 12 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 13 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 14 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 15 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 16 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 17 | Mean = 0.9886157 | Best = 0.9886382
GA | iter = 18 | Mean = 0.9885932 | Best = 0.9886382
GA | iter = 19 | Mean = 0.9885959 | Best = 0.9886382
GA | iter = 20 | Mean = 0.9886184 | Best = 0.9886382
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 50, pcrossover = 0.1,
+             pmutation = 0.01, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.9051567 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9395405 | Best = 0.9888888
GA | iter = 4 | Mean = 0.9549521 | Best = 0.9888888
GA | iter = 5 | Mean = 0.9648784 | Best = 0.9888888
GA | iter = 6 | Mean = 0.9647177 | Best = 0.9888888
GA | iter = 7 | Mean = 0.9739832 | Best = 0.9888888
GA | iter = 8 | Mean = 0.9810065 | Best = 0.9889053
GA | iter = 9 | Mean = 0.9778540 | Best = 0.9889053
GA | iter = 10 | Mean = 0.9818591 | Best = 0.9889053
GA | iter = 11 | Mean = 0.9845454 | Best = 0.9889347
GA | iter = 12 | Mean = 0.9847312 | Best = 0.9889347
GA | iter = 13 | Mean = 0.9880008 | Best = 0.9889503
GA | iter = 14 | Mean = 0.9883375 | Best = 0.9889503
GA | iter = 15 | Mean = 0.9887652 | Best = 0.9889503
GA | iter = 16 | Mean = 0.9888026 | Best = 0.9889639
GA | iter = 17 | Mean = 0.9888627 | Best = 0.9889639
GA | iter = 18 | Mean = 0.9888648 | Best = 0.9889639
GA | iter = 19 | Mean = 0.9888799 | Best = 0.9889639
GA | iter = 20 | Mean = 0.9888812 | Best = 0.9889639
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 100, pcrossover = 0.1,
+             pmutation = 0.01, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.8793135 | Best = 0.9888831
GA | iter = 2 | Mean = 0.9110572 | Best = 0.9888831
GA | iter = 3 | Mean = 0.9372310 | Best = 0.9888831
GA | iter = 4 | Mean = 0.9425003 | Best = 0.9888831
GA | iter = 5 | Mean = 0.9434097 | Best = 0.9888831
GA | iter = 6 | Mean = 0.9479034 | Best = 0.9888831
GA | iter = 7 | Mean = 0.9487419 | Best = 0.9889639
GA | iter = 8 | Mean = 0.9488115 | Best = 0.9889639
GA | iter = 9 | Mean = 0.9533347 | Best = 0.9889639
GA | iter = 10 | Mean = 0.9541570 | Best = 0.9889639
GA | iter = 11 | Mean = 0.9622597 | Best = 0.9889639
GA | iter = 12 | Mean = 0.9725374 | Best = 0.9889639
GA | iter = 13 | Mean = 0.9768423 | Best = 0.9889639
GA | iter = 14 | Mean = 0.9806991 | Best = 0.9889639
GA | iter = 15 | Mean = 0.9826561 | Best = 0.9889639
GA | iter = 16 | Mean = 0.9814939 | Best = 0.9889639
GA | iter = 17 | Mean = 0.9847555 | Best = 0.9889639
GA | iter = 18 | Mean = 0.9860718 | Best = 0.9889639
GA | iter = 19 | Mean = 0.9862913 | Best = 0.9889639
GA | iter = 20 | Mean = 0.9874262 | Best = 0.9889639
```

Popsize가 10일 때는 다른 경우의 최고의 값 0.9889639에 도달하지 못한다.

Popsize 50일 때와 100일 때를 비교해보면 둘 다 0.9889639에 도달하지만, 도달하는 iteration에서 차이가 있다. 50일 경우에 더 많은 16 번의 iteration을 수행해야 하는 반면에, 100일 경우에는 7 iteration만 수행하면 0.9889639에 도달하고 더 이상 변화하지 않는다.

```
Call:
lm(formula = Mean_temperature ~ ., data = GA_trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8506 -1.0813 -0.0830  0.8877  6.3015

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         90.13464   22.89125   3.938 0.000107 ***
Max_termperature     0.55186    0.01450  38.053  < 2e-16 ***
Min_temperature      0.36529    0.02040  17.904  < 2e-16 ***
Sea_level_pressure -11.66207    3.45600  -3.374 0.000860 ***
Standard_pressure    9.75193    4.00300   2.436 0.015561 *
Max_wind_speed       0.04550    0.01981   2.297 0.022455 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.636 on 244 degrees of freedom
Multiple R-squared:  0.9889,    Adjusted R-squared:  0.9886
F-statistic:  4334 on 5 and 244 DF,  p-value: < 2.2e-16
```

변수 선택 결과를 보면 dewpoint 변수가 제거된 5개의 변수가 선택된 것을 확인할 수 있다.

Pcrossover 옵션을 변경해보자.

```
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 50, pcrossover = 0.1,
+             pmutation = 0.01, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.9051567 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9395405 | Best = 0.9888888
GA | iter = 4 | Mean = 0.9549521 | Best = 0.9888888
GA | iter = 5 | Mean = 0.9648784 | Best = 0.9888888
GA | iter = 6 | Mean = 0.9647177 | Best = 0.9888888
GA | iter = 7 | Mean = 0.9739832 | Best = 0.9888888
GA | iter = 8 | Mean = 0.9810065 | Best = 0.9889053
GA | iter = 9 | Mean = 0.9778540 | Best = 0.9889053
GA | iter = 10 | Mean = 0.9818591 | Best = 0.9889053
GA | iter = 11 | Mean = 0.9845454 | Best = 0.9889347
GA | iter = 12 | Mean = 0.9847312 | Best = 0.9889347
GA | iter = 13 | Mean = 0.9880008 | Best = 0.9889503
GA | iter = 14 | Mean = 0.9883375 | Best = 0.9889503
GA | iter = 15 | Mean = 0.9887652 | Best = 0.9889503
GA | iter = 16 | Mean = 0.9888026 | Best = 0.9889639
GA | iter = 17 | Mean = 0.9888627 | Best = 0.9889639
GA | iter = 18 | Mean = 0.9888648 | Best = 0.9889639
GA | iter = 19 | Mean = 0.9888799 | Best = 0.9889639
GA | iter = 20 | Mean = 0.9888812 | Best = 0.9889639
```

```
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 50, pcrossover = 0.5,
+             pmutation = 0.01, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.9038099 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9409802 | Best = 0.9888888
GA | iter = 4 | Mean = 0.9496711 | Best = 0.9888888
GA | iter = 5 | Mean = 0.9675970 | Best = 0.9888888
GA | iter = 6 | Mean = 0.9746978 | Best = 0.9889639
GA | iter = 7 | Mean = 0.9772644 | Best = 0.9889639
GA | iter = 8 | Mean = 0.9780216 | Best = 0.9889639
GA | iter = 9 | Mean = 0.9795839 | Best = 0.9889639
GA | iter = 10 | Mean = 0.9844732 | Best = 0.9889639
GA | iter = 11 | Mean = 0.9853749 | Best = 0.9889639
GA | iter = 12 | Mean = 0.9857525 | Best = 0.9889639
GA | iter = 13 | Mean = 0.9869629 | Best = 0.9889639
GA | iter = 14 | Mean = 0.9879114 | Best = 0.9889639
GA | iter = 15 | Mean = 0.9885822 | Best = 0.9889639
GA | iter = 16 | Mean = 0.9886529 | Best = 0.9889639
GA | iter = 17 | Mean = 0.9887368 | Best = 0.9889639
GA | iter = 18 | Mean = 0.9887919 | Best = 0.9889639
GA | iter = 19 | Mean = 0.9887898 | Best = 0.9889639
GA | iter = 20 | Mean = 0.9888388 | Best = 0.9889639
```

```
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.9100275 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9513927 | Best = 0.9889053
GA | iter = 4 | Mean = 0.9279270 | Best = 0.9889053
GA | iter = 5 | Mean = 0.9342988 | Best = 0.9889347
GA | iter = 6 | Mean = 0.9463711 | Best = 0.9889347
GA | iter = 7 | Mean = 0.9329151 | Best = 0.9889503
GA | iter = 8 | Mean = 0.9676232 | Best = 0.9889503
GA | iter = 9 | Mean = 0.9737465 | Best = 0.9889503
GA | iter = 10 | Mean = 0.9745460 | Best = 0.9889503
GA | iter = 11 | Mean = 0.9728380 | Best = 0.9889503
GA | iter = 12 | Mean = 0.9788108 | Best = 0.9889503
GA | iter = 13 | Mean = 0.9813061 | Best = 0.9889503
GA | iter = 14 | Mean = 0.9849716 | Best = 0.9889503
GA | iter = 15 | Mean = 0.9865345 | Best = 0.9889503
GA | iter = 16 | Mean = 0.9863958 | Best = 0.9889503
GA | iter = 17 | Mean = 0.9865246 | Best = 0.9889503
GA | iter = 18 | Mean = 0.9875570 | Best = 0.9889503
GA | iter = 19 | Mean = 0.9860467 | Best = 0.9889503
GA | iter = 20 | Mean = 0.9822407 | Best = 0.9889503
```

눈에 띄는 변화를 확인하기 위해, 차이를 크게 두었다.

0.1 ,0.5, 0.8 일 때 각각을 살펴보자.

Pcrossover 옵션의 경우 최적의 특정 값이 존재하는 듯 보인다. 세 개의 후보에 관해 수행하였는데 0.5일 때 제일 iteration을 적게 수행하고 최적값에 도달하였다. 즉 이 crossover option은 높을수록, 낮을수록 좋은 것이 아닌 그 사이 어딘가에 적합한 값이 존재한다.

```
Call:
lm(formula = Mean_temperature ~ ., data = GA_trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9454 -0.9941 -0.1025  0.9078  5.7227

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.38562   23.26440   3.197  0.00157 **
Max_termperature   0.56043    0.01501  37.337  < 2e-16 ***
Min_temperature    0.31420    0.02550  12.319  < 2e-16 ***
Dewpoint           0.07704    0.02676   2.879  0.00435 **
Sea_level_pressure -8.29239   3.60493  -2.300  0.02228 *
Standard_pressure   6.49102   4.12392   1.574  0.11680
Visibility          0.06759   0.08072   0.837  0.40323
Wind_speed          0.13193   0.04677   2.821  0.00519 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.613 on 242 degrees of freedom
Multiple R-squared:  0.9893,    Adjusted R-squared:  0.989
F-statistic:  3185 on 7 and 242 DF,  p-value: < 2.2e-16
```

변수 선택 결과를 보면 visibility 변수가 유효한 변수로 추가되었음을 확인할 수 있다.

마지막으로 pmutation 옵션을 변경해보자.

```
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 50, pcrossover = 0.5,
+             pmutation = 0.01, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.9038099 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9409802 | Best = 0.9888888
GA | iter = 4 | Mean = 0.9496711 | Best = 0.9888888
GA | iter = 5 | Mean = 0.9675970 | Best = 0.9888888
GA | iter = 6 | Mean = 0.9746978 | Best = 0.9889639
GA | iter = 7 | Mean = 0.9772644 | Best = 0.9889639
GA | iter = 8 | Mean = 0.9780216 | Best = 0.9889639
GA | iter = 9 | Mean = 0.9795839 | Best = 0.9889639
GA | iter = 10 | Mean = 0.9844732 | Best = 0.9889639
GA | iter = 11 | Mean = 0.9853749 | Best = 0.9889639
GA | iter = 12 | Mean = 0.9857525 | Best = 0.9889639
GA | iter = 13 | Mean = 0.9869629 | Best = 0.9889639
GA | iter = 14 | Mean = 0.9879114 | Best = 0.9889639
GA | iter = 15 | Mean = 0.9885822 | Best = 0.9889639
GA | iter = 16 | Mean = 0.9886529 | Best = 0.9889639
GA | iter = 17 | Mean = 0.9887368 | Best = 0.9889639
GA | iter = 18 | Mean = 0.9887919 | Best = 0.9889639
GA | iter = 19 | Mean = 0.9887898 | Best = 0.9889639
GA | iter = 20 | Mean = 0.9888388 | Best = 0.9889639
```

```
> GA_r2 <- ga(type = "binary", fitness = fit_r2, nBits = ncol(x),
+             names = colnames(x), popSize = 50, pcrossover = 0.5,
+             pmutation = 0.1, maxiter = 20, elitism = 2, seed = 123)
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.8903937 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9315793 | Best = 0.9888888
GA | iter = 4 | Mean = 0.9552792 | Best = 0.9888888
GA | iter = 5 | Mean = 0.9457533 | Best = 0.9888888
GA | iter = 6 | Mean = 0.9671356 | Best = 0.9889503
GA | iter = 7 | Mean = 0.9685248 | Best = 0.9889503
GA | iter = 8 | Mean = 0.9760404 | Best = 0.9889503
GA | iter = 9 | Mean = 0.9742513 | Best = 0.9889503
GA | iter = 10 | Mean = 0.9602760 | Best = 0.9889503
GA | iter = 11 | Mean = 0.9424168 | Best = 0.9889503
GA | iter = 12 | Mean = 0.9678809 | Best = 0.9889503
GA | iter = 13 | Mean = 0.9782509 | Best = 0.9889503
GA | iter = 14 | Mean = 0.9772289 | Best = 0.9889503
GA | iter = 15 | Mean = 0.9765090 | Best = 0.9889503
GA | iter = 16 | Mean = 0.9806204 | Best = 0.9889503
GA | iter = 17 | Mean = 0.9830991 | Best = 0.9889503
GA | iter = 18 | Mean = 0.9823773 | Best = 0.9889503
GA | iter = 19 | Mean = 0.9795209 | Best = 0.9889503
GA | iter = 20 | Mean = 0.9790417 | Best = 0.9889503
```

```
GA | iter = 1 | Mean = 0.8457345 | Best = 0.9888888
GA | iter = 2 | Mean = 0.8872166 | Best = 0.9888888
GA | iter = 3 | Mean = 0.9188532 | Best = 0.9889035
GA | iter = 4 | Mean = 0.9377460 | Best = 0.9889503
GA | iter = 5 | Mean = 0.9550680 | Best = 0.9889503
GA | iter = 6 | Mean = 0.9618408 | Best = 0.9889503
GA | iter = 7 | Mean = 0.9557553 | Best = 0.9889503
GA | iter = 8 | Mean = 0.9538369 | Best = 0.9889503
GA | iter = 9 | Mean = 0.9574194 | Best = 0.9889503
GA | iter = 10 | Mean = 0.9588799 | Best = 0.9889503
GA | iter = 11 | Mean = 0.9677900 | Best = 0.9889639
GA | iter = 12 | Mean = 0.9529862 | Best = 0.9889639
GA | iter = 13 | Mean = 0.9435903 | Best = 0.9889639
GA | iter = 14 | Mean = 0.9669255 | Best = 0.9889639
GA | iter = 15 | Mean = 0.9699238 | Best = 0.9889639
GA | iter = 16 | Mean = 0.9620488 | Best = 0.9889639
GA | iter = 17 | Mean = 0.9643991 | Best = 0.9889639
GA | iter = 18 | Mean = 0.9629497 | Best = 0.9889639
GA | iter = 19 | Mean = 0.9757429 | Best = 0.9889639
GA | iter = 20 | Mean = 0.9772101 | Best = 0.9889639
```

0.01 , 0.1, 0.3 일 때로 나누어 비교해보았다.

결과를 보면 0.01일 때 빠른 iteration 내에 최적 값에 도달하는 것과 달리 0.1로 변화시켰는데, 20 iteration 내에 최적 값에 도달하지 못했다. 즉, 매우 민감하게 반응하는 모습을 보인다.

변수 선택 결과를 살펴보면 다음과 같다.

```
Call:
lm(formula = Mean_temperature ~ ., data = GA_trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9454 -0.9941 -0.1025  0.9078  5.7227

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       74.38562   23.26440   3.197  0.00157 **
Max_termperature   0.56043    0.01501  37.337  < 2e-16 ***
Min_temperature    0.31420    0.02550  12.319  < 2e-16 ***
Dewpoint           0.07704    0.02676   2.879  0.00435 **
Sea_level_pressure -8.29239   3.60493  -2.300  0.02228 *
Standard_pressure  6.49102    4.12392   1.574  0.11680
Visibility         0.06759    0.08072   0.837  0.40323
Wind_speed         0.13193    0.04677   2.821  0.00519 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.613 on 242 degrees of freedom
Multiple R-squared:  0.9893,    Adjusted R-squared:  0.989
F-statistic:  3185 on 7 and 242 DF,  p-value: < 2.2e-16
```

Pcrossover 옵션과 동일하게 visibility 변수가 추가된 7개의 변수가 선택되었음을 확인할 수 있다.

종합적으로 정리하자면, 일단 위의 데이터셋으로는 사실 어떤 옵션에 민감하게 반응하는지에 대한 정확한 판단을 내리기가 어렵다. 옵션을 변경해도 R2값의 변동이 크지 않기 때문이다.

그래도 위의 후보 셋들을 통해 추정을 해보자면, MUTATION RATE 옵션이 제일 영향을 많이 미치며 그 다음이 Population size, Cross-over rate 순이다.

Mutation rate이 크다는 것은 기존의 조합 외의 새로운 돌연변이가 생성되는 확률이므로 mutation rate이 커지면 R2값(Fitness function)이 급격하게 작아질 것이다. 따라서 mutation rate 값은 큰 변화를 주면 안된다.