다변량분석 Assignment 5

2014170824 황태민

[Q1] All Variables

모든 변수를 사용하여 Multiple Linear Regression (MLR) 모형을 학습한 결과는 다음과 같다.

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0070 -1.1036 -0.1057  0.9213  5.5629

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.42371   23.71798   3.475 0.000606 ***
Max_termperature   0.54929    0.01669  32.916  < 2e-16 ***
Min_temperature    0.33083    0.02703  12.241  < 2e-16 ***
Dewpoint           0.06663    0.02836   2.349 0.019625 *
Precipitation      0.22227    0.80914   0.275 0.783783
Sea_level_pressure -9.36521   3.83076  -2.445 0.015216 *
Standard_pressure  7.39153    4.34684   1.700 0.090344 .
Visibility         0.17062    0.08806   1.938 0.053854 .
Wind_speed         0.08272    0.05928   1.395 0.164189
Max_wind_speed    -0.01336    0.02833  -0.471 0.637823
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.667 on 240 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9877
F-statistic:  2216 on 9 and 240 DF,  p-value: < 2.2e-16
```

모든 변수를 사용하여 구축한 모형의 Adjusted R-sq 값은 0.9877로 상당히 높은 값을 가지는 것을 볼 수 있다. Multiple R-sq 값이 0.9881이므로 수정 결정계수와 비교했을 때 크게 차이가 나지 않는다는 것을 알 수 있다. 이는 데이터가 가지는 전체 변동중 약 98.7%가 회귀식에 의해 설명된다는 의미이다. 유의수준 1%에서 유의미한 변수는 Max_temperature, Min_temperature 두 개의 변수만이 이에 해당한다.

모든 변수를 사용하여 학습한 모형을 통해 validation dataset에 대한 평가 지표들은 다음과 같다.

	RMSE	MAE	MAPE
All	1.363005	1.092225	2.342326

RMSE 값은 1.363005이고, MAE 값은 1.092225로 평균적으로 예측값과 실제값이 1.09정도 차이를 보인다는 것이다. MAPE는 2.342326으로 예측값과 실제값이 약 2.34% 차이를 보인다.

[Q2] Exhaustive Search

Exhaustive Search를 통해 모든 경우의 회귀 모형을 만들어 Adjusted R-sq 값을 구해본 결과는 다음과 같다.

	var1	var2	var3	var4	var5	var6	var7	var8	var9	ADJ R
X.477	1	1	1	0	1	1	1	1	0	0.98775
X.478	1	1	1	0	1	1	1	1	1	0.98771
X.475	1	1	1	0	1	1	1	0	0	0.98770
X.509	1	1	1	1	1	1	1	1	0	0.98770
X.476	1	1	1	0	1	1	1	0	1	0.98766
X.510	1	1	1	1	1	1	1	1	1	0.98766
X.469	1	1	1	0	1	0	1	1	0	0.98765
X.507	1	1	1	1	1	1	1	0	0	0.98765
X.508	1	1	1	1	1	1	1	0	1	0.98762
X.470	1	1	1	0	1	0	1	1	1	0.98761
X.473	1	1	1	0	1	1	0	1	0	0.98761
X.501	1	1	1	1	1	0	1	1	0	0.98761
X.467	1	1	1	0	1	0	1	0	0	0.98758
X.474	1	1	1	0	1	1	0	1	1	0.98757

약 상위 10개 경우에 대한 변수 조합과 Adjusted R-sq값은 위와 같다. 상위 4개의 Adjusted R-sq 값이 소수점 4번째 자리까지는 0.9877로 같은 것을 볼 수 있다. 또한 첨부한 표의 조합들의 Adj.R 값이 거의 차이가 나지 않는다는 것을 알 수 있다. 가장 높은 Adjusted R-sq 값은 0.98775이고 이 값을 가지는 변수들의 조합은 1, 2, 3, 5, 6, 7, 8번 변수들이다.

소요시간은 1.408233초로 다소 짧은 시간내에 연산이 가능했다. 변수가 하나씩 많아질 때마다 2배의 양을 연산해야 하므로 변수의 개수가 5개만 늘어나도 약 1분이 걸릴 것으로 예상된다.

```
Time difference of 1.408233 secs
```

위의 결과를 통해 얻은 조합을 통해 회귀식을 구성한 결과는 다음과 같다.

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0023 -1.0844 -0.1001  0.9463  5.5837

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.32339   23.30264   3.533 0.000492 ***
Max_termperature   0.54727    0.01600  34.213  < 2e-16 ***
Min_temperature    0.33255    0.02659  12.506  < 2e-16 ***
Dewpoint           0.06634    0.02777   2.389 0.017666 *
Sea_level_pressure -9.44216   3.80204  -2.483 0.013690 *
Standard_pressure  7.48168    4.32352   1.730 0.084823 .
Visibility         0.16940    0.08764   1.933 0.054405 .
Wind_speed         0.06560    0.04653   1.410 0.159863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.661 on 242 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9877
F-statistic:  2869 on 7 and 242 DF,  p-value: < 2.2e-16
```

Adjusted R-sq 값은 0.9877이고 모든 변수를 사용했을 때와 동일하게 유의수준 0.01에서 max_temperature와 min_temperature만이 유의미한 변수이다.

	RMSE	MAE	MAPE
All	1.363005	1.092225	2.342326
Exhaustive	1.352704	1.089025	2.328155

위 회귀 모형을 통해 validation을 진행하였을 때의 평가지표들은 위와 같다. 세 가지 지표 모두에서 모든 변수를 사용한 것보다 좋은 수치를 가진다는 것을 알 수 있다. 세 가지 지표 모두 큰 차이를 보이는 것은 아니지만 그래도 변수를 7개만 사용한 모델의 예측력이 모든 변수를 사용한 모델보다 좋은 수치를 가지는 것은 모든 변수를 사용하는 것이 예측에 적합하지 않다는 것을 시사한다.

[Q3] Forward Selection, Backward Elimination, Stepwise Selection

Forward Selection의 결과를 먼저 살펴보자.

```
> summary(forward_model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0023 -1.0844 -0.1001  0.9463  5.5837

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.32339   23.30264   3.533 0.000492 ***
Max_termperature   0.54727    0.01600  34.213  < 2e-16 ***
Min_temperature    0.33255    0.02659  12.506  < 2e-16 ***
Sea_level_pressure -9.44216    3.80204  -2.483 0.013690 *
Standard_pressure  7.48168    4.32352   1.730 0.084823 .
Visibility         0.16940    0.08764   1.933 0.054405 .
Dewpoint           0.06634    0.02777   2.389 0.017666 *
Wind_speed         0.06560    0.04653   1.410 0.159863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.661 on 242 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9877
F-statistic:  2869 on 7 and 242 DF,  p-value: < 2.2e-16
```

먼저 Forward selection의 결과로 선택된 변수들은 정확히 Exhaustive Search를 사용하여 선택한 변수들과 일치한다. 이에 따라 당연히 회귀 모델의 결과도 같다. Adjusted R-sq 값은 0.9877이다.

소요시간은 0.3470678초로 Exhaustive Search에 비해 거의 1/3 이상 적게 소요되었다.

```
Time difference of 0.3470678 secs
```

예측에 대한 평가는 F.S, B.E, S.S의 결과를 본 후 설명하도록 하자.

다음은 Backward Elimination의 결과를 살펴보자.

```
> summary(backward_model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0023 -1.0844 -0.1001  0.9463  5.5837

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.32339   23.30264   3.533 0.000492 ***
Max_termperature   0.54727    0.01600  34.213  < 2e-16 ***
Min_temperature    0.33255    0.02659  12.506  < 2e-16 ***
Dewpoint           0.06634    0.02777   2.389 0.017666 *
Sea_level_pressure -9.44216   3.80204  -2.483 0.013690 *
Standard_pressure  7.48168    4.32352   1.730 0.084823 .
Visibility         0.16940    0.08764   1.933 0.054405 .
Wind_speed         0.06560    0.04653   1.410 0.159863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.661 on 242 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9877
F-statistic:  2869 on 7 and 242 DF,  p-value: < 2.2e-16
```

Backward Elimination의 결과로 선택된 변수들도 정확히 Exhaustive Search를 사용하여 선택한 변수들과 일치한다. 이에 따라 당연히 회귀 모델의 결과도 같다. Adjusted R-sq 값은 0.9877이다.

소요시간은 0.257746초로 Forward Selection 보다도 적게 소요되었다.

```
Time difference of 0.257746 secs
```

마지막으로 Stepwise Selection의 결과를 살펴보자.

```
> summary(stepwise_model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0023 -1.0844 -0.1001  0.9463  5.5837

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.32339   23.30264   3.533 0.000492 ***
Max_termperature   0.54727    0.01600  34.213  < 2e-16 ***
Min_temperature    0.33255    0.02659  12.506  < 2e-16 ***
Sea_level_pressure -9.44216    3.80204  -2.483 0.013690 *
Standard_pressure  7.48168    4.32352   1.730 0.084823 .
Visibility         0.16940    0.08764   1.933 0.054405 .
Dewpoint           0.06634    0.02777   2.389 0.017666 *
Wind_speed         0.06560    0.04653   1.410 0.159863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.661 on 242 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9877
F-statistic:  2869 on 7 and 242 DF,  p-value: < 2.2e-16
```

Stepwise Selection의 결과로 선택된 변수들도 정확히 Exhaustive Search를 사용하여 선택한 변수들과 일치한다. 이에 따라 당연히 회귀 모델의 결과도 같다. Adjusted R-sq 값은 0.9877이다.

소요시간은 0.4527791초로 Forward Selection이나 Backward Elimination보다는 오래 소요되었다는 것을 알 수 있다.

```
Time difference of 0.4527791 secs
```

세 가지 변수 선택 방법의 결과가 모두 같고, 또한 Exhaustive Search의 결과와도 같기 때문에 예측력에 관한 수치 또한 모두 같다.

```
                RMSE      MAE     MAPE
All         1.363005 1.092225 2.342326
Exhaustive  1.352704 1.089025 2.328155
Forward     1.352704 1.089025 2.328155
Backward    1.352704 1.089025 2.328155
Stepwise    1.352704 1.089025 2.328155
```

위와 마찬가지로 모든 변수를 사용했을 때 보다 좋은 결과를 보이고 있는 것을 알 수 있다.

[Q4] Genetic Algorithm

Genetic Algorithm을 이용하여 변수 선택을 실행해보았다.

Fitness function으로는 Adjusted R-sq를 사용하였고, population size는 50, cross over rate는 0.5, mutation rate는 0.01로 설정하였으며 최대 iteration은 100, elitism은 2로 설정하였다.

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

```
> summary(GA_model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.0023 -1.0844 -0.1001  0.9463  5.5837

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       82.32339   23.30264   3.533 0.000492 ***
Max_termperature   0.54727    0.01600  34.213  < 2e-16 ***
Min_temperature    0.33255    0.02659  12.506  < 2e-16 ***
Dewpoint           0.06634    0.02777   2.389 0.017666 *
Sea_level_pressure -9.44216   3.80204  -2.483 0.013690 *
Standard_pressure  7.48168    4.32352   1.730 0.084823 .
Visibility         0.16940    0.08764   1.933 0.054405 .
Wind_speed         0.06560    0.04653   1.410 0.159863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.661 on 242 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9877
F-statistic:  2869 on 7 and 242 DF,  p-value: < 2.2e-16
```

Genetic Algorithm을 사용하여 선택된 변수들을 사용한 회귀 모델의 결과도 앞서 실행해보았던 E.S.의 결과와 같으며 Adjusted R-sq 값도 0.9877로 동일하다.

소요시간은 9.740937초로 위의 사용했던 모든 방법들 보다도 훨씬 오랜 시간이 소요되었다. 변수의 개수가 많지 않을 때에 Genetic Algorithm은 시간 소요 측면에서는 비효율적이라고 볼 수 있다.

```
Time difference of 9.740937 secs
```

선택된 변수들이 동일하기 때문에 예측력 또한 위의 방법들과 같은 수치를 가진다. 정리해보면 예측 지표 관점에서 보면 모든 변수를 사용하는 방법을 제외한 5가지 방법이 동일한 수치를 가지고 이 5가지 방법을 시간 소요 순으로 나열해보면 B.E. < F.S. < S.S. < E.S. < G.A. 순으로 시간이 소요되는 것을 알 수 있다. 항상 Backward elimination 방법이 우월한 것은 아니지만 이 dataset에 대해서는 backward elimination 방법이 가장 효율적이라고 할 수 있다.

	RMSE	MAE	MAPE
All	1.363005	1.092225	2.342326
Exhaustive	1.352704	1.089025	2.328155
Forward	1.352704	1.089025	2.328155
Backward	1.352704	1.089025	2.328155
Stepwise	1.352704	1.089025	2.328155
GA	1.352704	1.089025	2.328155

[Q5] GA hyper parameter variation

-cross over rate 변화

먼저 cross over rate를 바꾸어 보았다. cross over rate가 0.5인 경우가 원래 값이었으므로 0.25의 경우와 0.75의 경우를 실행해보았다.

```
best_var_idx2    int [1:7] 1 2 3 5 6 7 8
best_var_idx3    int [1:7] 1 2 3 5 6 7 8
```

그 결과 두 경우 모두 1 2 3 5 6 7 8 변수를 선택하였으며 이 결과는 원래의 결과와 다르지 않고 Adj R-sq 값도 역시 동일하다.

-mutation rate 변화

다음은 mutation rate를 변화시켜보았다. 보통 mutation rate는 0.01로 설정하는데 훨씬 높은 값인 0.1과 0.5로 설정해보았다.

```
best_var_idx4    int [1:7] 1 2 3 5 6 7 8
best_var_idx5    int [1:7] 1 2 3 5 6 7 8
```

이 경우에도 원래의 결과와 다르지 않은 변수 선택 결과를 보였다. mutation rate를 0.5라는 아주 높은 수치로 설정했음에도 결과는 달라지지 않았다.

-population size 변화

다음은 population size를 변화시켜보았다. 원래의 경우 100으로 설정되어있었지만 25와 5라는 아주 작은 숫자로 설정하고 실행시켜보았다.

```
best_var_idx6    int [1:7] 1 2 3 5 6 7 8
best_var_idx7    int [1:5] 1 2 3 7 8
```

popsize가 25인 경우는 원래의 결과와 다르지 않게 나왔지만 popsize를 5로 설정한 결과가 처음으로 원래의 결과와 다른 결과를 보였다. 1 2 3 7 8 번째 변수만을 선택하였고 fitness function으로 설정한 Adj R-sq 값은 Best = 0.9869169 로 원래의 결과와 큰 차이를 보이진 않는다. popsize를 10보다 밑인 값을 설정하면 실행은 되지만 R에서 다음과 같은 경고 메시지를 보여준다. 이에 따르면 10보다 작은 값은 설정하지 않는 것을 권장한다는 것이다.

```
Warning message:
In ga(type = "binary", fitness = fit_R, nBits = ncol(x), names = colnames(x),  :
  The population size is less than 10.
```

-max iteration 변화

다음은 max iteration을 변화시켜보았다. 원래의 경우 100으로 설정되어있었지만 50, 10, 5로 설정하여 유전 알고리즘을 실행시켜보았다.

```
best_var_idx8    int [1:7] 1 2 3 5 6 7 8
best_var_idx9    int [1:6] 1 2 3 5 6 7
best_var_idx10   int [1:6] 1 2 3 5 6 7
```

max iteration = 50의 경우 원래의 결과와 달라지지 않았지만 극단적으로 낮춘 10과 5의 경우는 8번 변수를 선택하지 않은 결과를 나타내었다. 이에 따른 Adj R-sq 값은 Best = 0.9877002로 원래 변수 선택의 결과와 크게 다르지는 않다.

위에서 네 가지 hyper parameter를 변화시켜가며 genetic algorithm을 실행시켜본 결과 population size와 max iteration이 충분히 클 경우, cross over rate와 mutation rate는 변수 선택에 큰 영향을 미치지 못하였다. 그렇다면 population size와 max iteration을 임의로 낮춘 후 cross over rate와 mutation rate를 조정해보도록 하자.

먼저 maxiter = 5, popsize = 10인 아주 극단적인 parameter 값을 설정한 결과 다음과 같은 변수 선택 결과를 보였다.

```
best_var_idx11      int [1:7] 1 2 5 6 7 8 9
```

7개의 변수가 선택되었고 이는 원래 GA 방법으로 선택한 결과와 다르다. 하지만 Adj R-sq는 Best = 0.9874636 로 충분히 높은 값을 보인다.

이제 maxiter = 5, popsize = 10, mutation rate = 0.5로 설정해보자.

```
best_var_idx12      int [1:6] 1 2 5 6 7 9
```

이 경우도 새로운 변수 선택 결과를 보인다. 하지만 여전히 Adj R-sq 값은 Best = 0.9874894 로 충분히 높은 값을 보인다.

마지막으로 maxiter = 5, popsize = 10, cross over rate = 0.75로 설정해보자.

```
best_var_idx13      int [1:6] 1 2 4 5 6 7
```

7개의 변수가 선택되었고 이는 위의 결과들과 또 다른 변수 선택을 보여준다. Adj R-sq 값은 Best = 0.9875047로 높은 수치를 보인다.

총 13가지의 hyper parameter 변화 조합을 통해 알 수 있는 것은 변수 선택에 있어서 가장 큰 영향을 미치는 파라미터는 population size와 max iteration이다. 당연한 결과일지도 모르지만 population size와 max iteration이 커질수록 더 많은 chromosome에 대한 시도를 해볼 것이므로 더 나은 결과를 낳게 된다. 반면에 cross over rate나 mutation rate의 경우 탐색하는 chromosome의 개수의 차이는 없다. 이에 따라 결과에 큰 차이를 보이지 않지만, population size와 max iteration이 작을 경우 변수 선택에 있어서 차이를 보이곤 한다.

이 문제의 경우 입력변수의 개수가 9개 밖에 되지 않는 점을 고려했을 때 극단적인 파라미터 선정에도 꽤 좋은 변수 선택의 결과를 보이지만 현실에서 GA를 활용하는 분야는 입력변수의 개수가 훨씬 더 많을 것이다. 이에 따라 컴퓨터의 연산이 허용하는 내로 충분한 population size와 max iteration을 선정해 최선의 변수 선택 결과를 도출해야 할 것이라고 생각한다.