제목 : 로지스틱 회귀분석을 이용한 서울 강수 확률예측

과 목 명 : 회귀분석2

담당교수님 : 김규성 교수님

학 과 : 통계학과

학 번 : 2017580002

성 명 : 권휘성

제 출 일 : 20211128

서울시립대학교
UNIVERSITY OF SEOUL

차례

1. 서론

1.1 연구목적

기상과 관련된 정보는 예전부터 많은 사람들의 관심이 되는 대상이다. 작게는 여행 일정이나 야외 행사를 계획하는 순간부터 크게는 기상이변으로 인한 피해를 막기 위해 예측을 하고 그에 대한 대응을 하고 있다. 예측을 잘하여 사람들에게 정확한 정보를 제공하기 위해 기상청에서는 슈퍼 컴퓨터를 활용하는 등 오차를 가능하면 작게 하여 예측 정보를 제공하기 위해 각기에서 노력 중이다. 하지만 예측은 언제나 정확하기는 불가능하며 종종 틀린 예측을 하기도 한다. 기상정보는 생활과 밀접하게 연관되어 있기에 예측이 틀렸을 때 전체적으로 큰 피해를 준다. 특히, 농부와 같이 생업이 걸린 사람들에게 더욱 중요한 사안이다. 더 정확한 기상정보를 줄 수 있는 방안이 없을까 생각을 해보았다. 기상 관련 정보 중에서 비나 눈이 올지 안 올지에 대해 사람들이 평소 관심을 많이 가질 거라고 생각을 했고 비나 눈이 올 확률을 예측하여 미리 대비할 수 있도록 정보를 제공하고 싶었다. 더 정확한 예측을 위해 기온이나 풍속과 같은 다른 기상정보를 이용하여 비나 눈이 올 확률을 예측해 사람들에게 비가 오거나 눈이 올 상황에 대해 예측하고 대비할 수 있도록 도움을 주고 싶어 이 연구를 시작해본다.

1.2 문헌연구

동아사이언스에서 2021년 10월 1일에 발표한 '구글 딥마인드, 90분 뒤 강수 예측 AI 개발'이라는 기사에서 바둑 AI 알파고 개발업체로 유명한 딥마인드가 당장 한두 시간 뒤 비가 올지 예측하는 인공지능을 개발하였다고 말하고 있다. 여기에서는 학습 방법을 생성모델링(generative modeling)방식으로 진행하였고 그 결과는 상당히 높은 정확도를 보였다고 한다. 딥마인드 수석연구원인 샤키르 모하메드는 최근 기후변화로 기상이변이 잦아졌는데 AI가 향후 인명과 재산 피해를 줄이는 데 필수적인 도구가 될 거라고 말하고 있다. 이렇듯 인공지능 분야에서도 기상정보를 예측하여 피해를 줄이고자 하는 것을 알 수 있고 비록 생성 모델링을 구현할 수는 없지만 회귀분석 방법을 이용해도 예측이 가능하고 피해를 줄이는 데 도움이 될 것이라 생각하였다.

1.3 데이터 설명

먼저 올 한 해(2021년 1월 1일~) 서울의 강수량 정보를 기상자료 개방 포털에 있는 데이터를 통해 받아왔다. 다음은 받아온 서울의 강수량 정보의 일부이다. 관련된 변수들에 대해 설명을 해보겠다.

지점	지점명	일시	일강수량(mm)
108	서울	2021-01-04	0
108	서울	2021-01-05	0
108	서울	2021-01-06	2.3
108	서울	2021-01-12	2.6
108	서울	2021-01-13	0
108	서울	2021-01-15	0.2
108	서울	2021-01-17	0
108	서울	2021-01-18	0.3
108	서울	2021-01-21	9.2
108	서울	2021-01-22	0.1

지점, 지점명 : 측정을 한 장소(서울을 대상으로 연구할 것이기에 서울만 뽑았다.)
일시 : 강수량 측정한 날짜
일강수량(mm) : 측정한 강수량
다음으로 올 한 해(2021년 1월 1일~) 서울의 기온 정보를 기상자료 개방 포털에 있는 데이터를 통해 받아왔다. 다음은 받아온 서울의 기온 정보의 일부이다. 관련된 변수들에 대해 설명을 해보겠다.

지점	지점명	일시	평균기온(°	최저기온(°	최고기온(°(
108	서울	2021-01-01	-4.2	-9.8	1.6
108	서울	2021-01-02	-5	-8.4	-1.4
108	서울	2021-01-03	-5.6	-9.1	-2
108	서울	2021-01-04	-3.5	-8.4	0.3
108	서울	2021-01-05	-5.5	-9.9	-2.1
108	서울	2021-01-06	-7.4	-12	-1.9
108	서울	2021-01-07	-14.5	-16.5	-8.4
108	서울	2021-01-08	-14.9	-18.6	-10.7
108	서울	2021-01-09	-12.2	-16.6	-7.5
108	서울	2021-01-10	-7.7	-12.8	-2.7

날짜 : 각 날짜
지점 : 기온을 조사한 위치 (2호선들은 모두 서울에 위치하므로 서울을 기준으로 조사한 자료를 조사했다.)
평균 기온 : 각 날짜의 평균 기온
최저 기온 : 각 날짜의 최저 기온
최고 기온 : 각 날짜의 최고 기온

다음으로 올 한 해(2021년 1월 1일~) 서울의 평균 상대습도 정보를 기상자료 개방 포털에 있는 데이터를 통해 받아왔다. 다음은 받아온 서울의 평균 상대습도 정보의 일부이다. 관련된 변수들에 대해 설명을 해보겠다.

지점	지점명	일시	평균 상대습도(%)
108	서울	2021-01-01	64
108	서울	2021-01-02	38.5
108	서울	2021-01-03	45
108	서울	2021-01-04	51.4
108	서울	2021-01-05	52.8
108	서울	2021-01-06	54.6
108	서울	2021-01-07	49.9
108	서울	2021-01-08	44
108	서울	2021-01-09	46.3

지점, 지점명 : 측정을 한 장소(서울을 대상으로 연구할 것이기에 서울만 뽑았다.)
일시 : 평균 상대습도 측정한 날짜
평균 상대습도(%) : 측정한 평균 상대습도
다음으로 올 한 해(2021년 1월 1일~) 서울의 평균 풍속 정보를 기상자료 개방 포털에 있는 데이터를 통해 받아왔다. 다음은 받아온 서울의 평균 풍속 정보의 일부이다. 관련된 변수들에 대해 설명을 해보겠다.

지점	지점명	일시	평균 풍속(m/s)
108	서울	2021-01-01	2
108	서울	2021-01-02	2.6
108	서울	2021-01-03	2
108	서울	2021-01-04	1.7
108	서울	2021-01-05	2.9
108	서울	2021-01-06	2.4
108	서울	2021-01-07	4.1
108	서울	2021-01-08	3.3
108	서울	2021-01-09	2.6

지점, 지점명 : 측정을 한 장소(서울을 대상으로 연구할 것이기에 서울만 뽑았다.)
일시 : 평균 풍속 측정한 날짜
평균 풍속(m/s) : 측정한 평균 풍속
사용할 데이터들은 위와 같다.

1.4 분석 방법

평균 기온, 평균 상대습도와 평균 풍속을 독립변수로 설정할 것이다. 반응변수로는

비가 왔을 때를 1로, 비가 오지 않았을 때(강수량이 0일 때)는 0으로 이항변수를 만들 것이다. 독립변수와 반응변수를 만든 후 로지스틱 회귀분석을 적합하여 독립변수들이 주어졌을 때 비가 올 확률을 예측하는 모델을 만들어볼 예정이다. 적합을 한 후 각 회귀계수들의 적합성을 파악해보고 잔차 그래프를 그려봄으로써 모형이 적합한지 파악할 것이다. 적합하다면 그대로 모델을 활용할 것이고 적합하지 않다면 이차항을 추가해보는 등 최대한 가정에 부합하고 적합한 모형을 향해 수정해나갈 것이다. 그리고 모형이 정해진다면 적합에 사용하지 않은 데이터를 이용해 예측을 해보고 오차가 어느 정도일지 실제로 파악해볼 예정이다.

1.5 결과 활용 및 기대효과

여름철에 장마 기간이 있기에 이 기간 비가 오는 날이 많을 것이다. 따라서 평균 기온과 습도가 높을 때 비나 눈이 올 확률이 높을 거라고 예상된다. 물론 비가 오는지 안 오는지의 여부에만 관심있기에 분석결과가 예상과 다를 수 있다는 생각도 든다. 비나 눈이 오면 바람이 많이 강하게 풀기에 평균 풍속이 강할수록 비나 눈이 올 확률이 높을 것이라고 예상한다. 물론 예측값이 언제나 정확하다고는 보장할 수는 없다. 하지만 이러한 결과들을 토대로 모델을 만들어 미래에 비나 눈이 올 확률을 예측할 수 있다면 일상의 크고 작은 부분부터 삶의 질이 많이 상승할 효과를 기대하고 있다.

2. 본론

2.1 분석방법 소개

모든 분석 과정은 SAS 프로그램을 활용하여 진행할 것이다. 비가 오는 것을 1로 오지 않는 사건을 0으로 하는 이항변수를 반응변수로 사용할 것이다. 독립변수로는 비가 오는 것과 관련이 있을 법한 변수들을 택하였다. 그리고 그 변수들은 평균 기온, 평균 상대습도와 평균 풍속이다. 먼저 이 변수들 간 관계를 상관계수를 통해 간단히 볼 것이다. 그 후, 반응변수로 정한 비가 오는 사건을 바탕으로 로지스틱 회귀분석을 실시해볼 것이다. 적합결과를 토대로 모형이 적합한지, 각 회귀계수들이 유의한지를 확인한 후 각 변수들이 어떤 영향을 주는지 파악해볼 것이다. 그후 잔차그래프를 그려봄으로써 이분산성 같은 모형의 가정을 위배하는 사항이 나온다면 변환을 하거나 적절한 조치를 취하여 가정이 지켜지도록 할 것이다. 또한, 이차항이나 교호작용을 추가하였을 때 더 나은 설명력을 보인다면 그 모형을 선택할 것이다. 이러한 과정을 통해 최종 모형이 선택되면 학습에 쓰이지 않은 자료들을 불러와 어떠한 날의 평균 기온, 평균 상대습도와 평균 풍속을 토대로 특정한 날의 비가 올 확률을 예측해볼 것이다.

2.2 데이터 분석 및 결과 설명

먼저 전처리과정에 대해 설명해보겠다.

```
proc import datafile="C:₩Users₩0518h₩Desktop₩서울시립대학교₩3학년₩2학기₩회귀분석 2₩학기말과제₩data₩강수량.csv"
dbms=csv replace out=rainfall;
getnames=yes;
run;

data rainfall;
set rainfall;
keep VAR3 VAR4;
run;

proc import datafile="C:₩Users₩0518h₩Desktop₩서울시립대학교₩3학년₩2학기₩회귀분석 2₩학기말과제₩data₩기온.csv"
dbms=csv replace out=temp;
getnames=yes;
run;

data temp;
set temp;
keep VAR3 VAR4;
run;

proc import datafile="C:₩Users₩0518h₩Desktop₩서울시립대학교₩3학년₩2학기₩회귀분석 2₩학기말과제₩data₩평균상대습도.csv"
dbms=csv replace out=hum;
getnames=yes;
run;

data hum;
set hum;
keep VAR3 VAR4;
run;

proc import datafile="C:₩Users₩0518h₩Desktop₩서울시립대학교₩3학년₩2학기₩회귀분석 2₩학기말과제₩data₩평균풍속.csv"
dbms=csv replace out=wind;
getnames=yes;
run;

data wind;
set wind;
keep VAR3 VAR4;
run;
```

데이터들을 각각 불러와 rainfall, temp, hum, wind에 저장을 하였고 temp, 지점명과 같은 필요가 없는 칼럼들을 탈락시키고 각각 날짜와 관심있는 변수만 keep을 통해 저장하였다.

```
data rainfall;
set rainfall(rename=(VAR4=rainfall));
run;

data temp;
set temp(rename=(VAR4=temp));
run;

data hum;
set hum(rename=(VAR4=hum));
run;

data wind;
set wind(rename=(VAR4=wind));
run;
```

또한 var4의 경우에는 각각 관심있는 변수의 칼럼들로 칼럼명을 rename 해주었다. Var3는 date에 해당하여 굳이 칼럼명을 바꿔주지 않았다.

```
data m;
merge temp rainfall hum wind;
by VAR3;
run;
```

그리고 date인 var3를 가지고 가진 데이터들을 병합하였고 간단히 보면 다음과 같다.

	VAR3	temp	rainfall	hum	wind
1	2021-01-01	-4.2	.	64	2
2	2021-01-02	-5	.	38.5	2.6
3	2021-01-03	-5.6	.	45	2
4	2021-01-04	-3.5	0	51.4	1.7
5	2021-01-05	-5.5	0	52.8	2.9
6	2021-01-06	-7.4	2.3	54.6	2.4
7	2021-01-07	-14.5	.	49.9	4.1

위는 병합된 데이터의 일부이다. 결측치가 관측되는데 이는 이 날에는 비나 눈이 안 왔다고 판단하고 진행하였다.

결측치를 0으로 대체하고 상관계수를 파악해보았다.

```
data m;
set m;
if rainfall=. then rainfall=0;
run;
proc corr data=m;
run;
```

단순 통계량						
변수	N	평균	표준편차	합	최솟값	최댓값
VAR3	317	22439	91.65424	7113163	22281	22597
temp	317	15.40000	10.05997	4882	-14.90000	31.70000
rainfall	317	3.60946	11.30671	1144	0	77.40000
hum	317	65.90789	13.54738	20893	33.00000	98.10000
wind	317	2.31514	0.61538	733.90000	1.30000	5.00000

피어슨 상관 계수, N = 317 H0: Rho=0 가정하에서 Prob > |r|					
	VAR3	temp	rainfall	hum	wind
VAR3	1.00000	0.64462 <.0001	0.06910 0.2199	0.40216 <.0001	-0.25166 <.0001
temp	0.64462 <.0001	1.00000	0.09525 0.0905	0.38105 <.0001	-0.29139 <.0001
rainfall	0.06910 0.2199	0.09525 0.0905	1.00000	0.40175 <.0001	0.11226 0.0458
hum	0.40216 <.0001	0.38105 <.0001	0.40175 <.0001	1.00000	-0.20041 0.0003
wind	-0.25166 <.0001	-0.29139 <.0001	0.11226 0.0458	-0.20041 0.0003	1.00000

변수들의 단순 통계량들을 확인할 수 있고 변수들간의 상관계수도 파악할 수 있다. 우리가 관심있는 변수인 rainfall에 대해서 보면 평균 온도와의 상관계수는 약 0.1 hum과의 상관계수는 약 0.4 wind와는 약 0.11임을 알 수 있다. 선형관계는 평균 상대습도와 가장 강하다고 볼 수 있겠다.

이제 로지스틱 회귀분석을 위해 rainfall이 0보다 클 때는 1, 0일 때는 0으로 만들어보겠다.

```
data m;
set m;
if rainfall>0 then rainfall=1;
run;
proc logistic data=m descending;
model rainfall=temp hum wind;
run;
```

그리고 그 결과는 아래와 같다.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	157.1806	3	<.0001
Score	120.4233	3	<.0001
Wald	70.6384	3	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.8009	1.9390	66.4078	<.0001
temp	1	-0.0536	0.0191	7.8903	0.0050
hum	1	0.1889	0.0226	69.7285	<.0001
wind	1	1.2682	0.2959	18.3637	<.0001

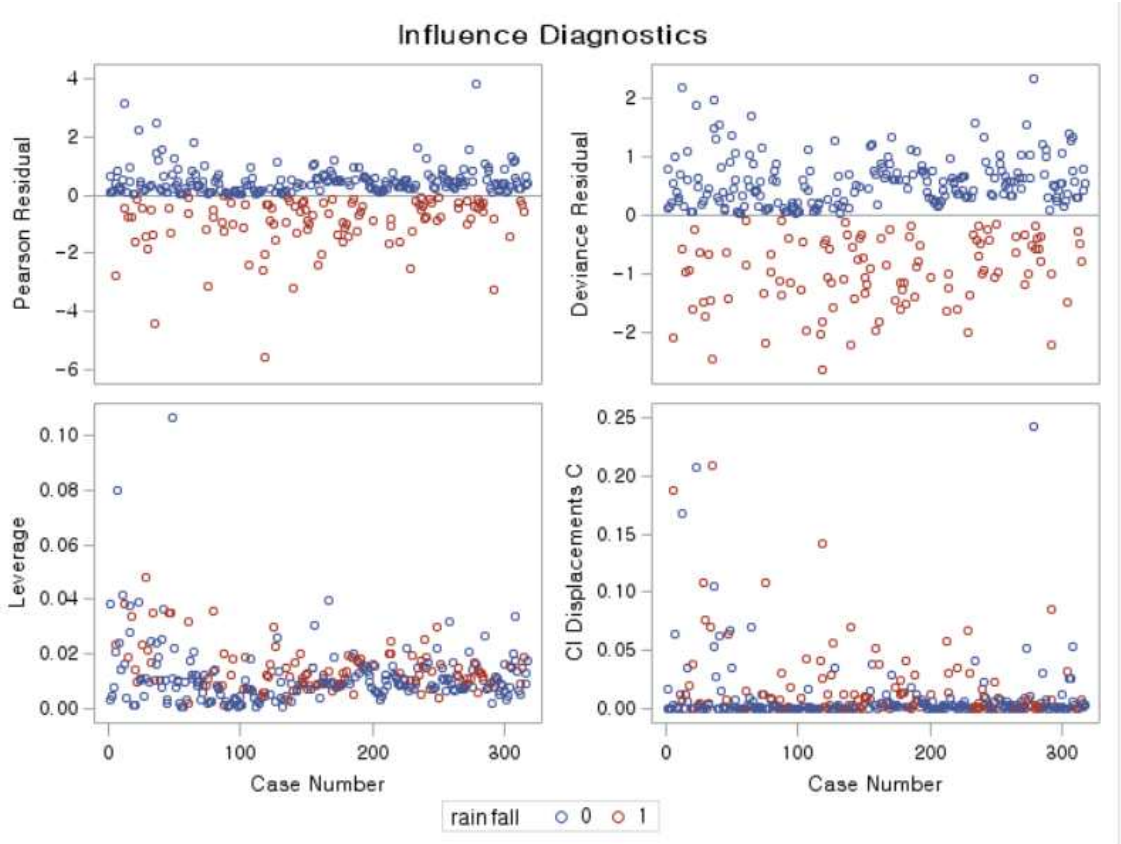
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
temp	0.948	0.913	0.984
hum	1.208	1.156	1.263
wind	3.554	1.990	6.349

회귀계수들을 보면 기울기는 -15.8009, temp의 경우에는 -0.0536, hum의 경우에는 0.1889, wind의 경우에는 1.2682인 것을 알 수 있다. 또한 모든 회귀계수들의 유의확률은 0.05보다 작으며 5% 유의수준하에서 모두 유의하다고 말할 수 있다. 그중

에서도 왈드 통계량이 제일 큰 평균 상대습도가 가장 유의하다. 또한 독립변수들과 반응변수와의 관계를 보도록 하겠다. 평균기온이 1도 증가하면 비나 눈이 올 확률은 0.9배가 된다. 평균 상대습도가 1%p 증가하면 비나 눈이 올 확률은 1.2배가 된다. 평균 풍속 1m/s 증가하면 비나 눈이 올 확률은 3.554배가 된다. 평균 풍속이 비나 눈이 오는 사건에 미치는 영향이 가장 크다고 볼 수 있다. 또한 오즈비의 95% 왈드 신뢰구간을 보았을 때 1을 포함하는 변수가 하나도 없으므로 모든 변수가 유의하다고 최종적으로 말할 수 있다. 하지만 LR 통계량이나 score 통계량을 보게 되면 모두 자유도 3하에서 유의확률이 0.0001보다 작으며 현재 고려하고 있는 모형이 적절하지 않다는 결론이 나온다.

```
proc logistic data = m;
model rainfall=temp hum wind/influence;
run;
quit;
```

위의 모형을 기반으로 잔차 그래프를 그려보았다.



잔차들 값을 보았을 때 크게 패턴을 보이지 않아 모형의 가정을 만족시킨다고 말할 수 있다. 하지만 잔차가 4 정도이거나 4를 넘어가는 관측치들도 있고 이상치라고 판단할 수 있었다. 또한, leverage를 보였을 때 유독 큰 값이 두 개가 보였고 모형

에 영향을 주는 영향치라고 판단할 수 있었다.

```
proc logistic data = m;
model rainfall=temp hum wind/rsquare;
run;
quit;
```

R-Square	0.3909	Max-rescaled R-Square	0.5436

이번엔 R-Square 값을 보았는데 0.39로 설명력이 크다곤 확실히 말할 순 없지만 나름 설명력이 있다고 판단했다.

```
proc logistic data = m;
model rainfall=temp hum wind/selection=backward;
run;
quit;
```

```
proc logistic data = m;
model rainfall=temp hum wind/selection=forward;
run;
quit;
```

이어서 적절한 모형에 대해 backward,forward stepwise selection을 진행하였고 두 방법 모두 모든 변수를 넣은 모형이 더 적합하다고 말하고 있다.

Note: All effects have been entered into the model.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	hum	1	1	95.7621	<.0001
2	wind	1	2	30.3188	<.0001
3	temp	1	3	8.1983	0.0042

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.8009	1.9390	66.4078	<.0001
temp	1	0.0536	0.0191	7.8903	0.0050
hum	1	-0.1889	0.0226	69.7285	<.0001
wind	1	-1.2682	0.2959	18.3637	<.0001

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.8009	1.9390	66.4078	<.0001
temp	1	0.0536	0.0191	7.8903	0.0050
hum	1	-0.1889	0.0226	69.7285	<.0001
wind	1	-1.2682	0.2959	18.3637	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
temp	1.055	1.016	1.095
hum	0.828	0.792	0.865
wind	0.281	0.158	0.502

다음으로는 적합성 검정을 해보았다.

```
proc logistic data = m;
model rainfall=temp hum wind/lackfit;
run;
quit;
```

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
7.3781	8	0.4964

호스머-램쇼 통계량을 보았을 때 자유도 8하에서 7.3781값을 보이므로 유의확률은 0.4964이고 위 모형은 적절하다고 말하고 있다. LR 통계량에 따르면 적절하지 않다고 하지만 회귀계수들이 모두 유의하고 호스머 램쇼 또한 귀무가설을 채택하기에 위 모형을 가지고 설명을 해보려 한다.

다음으로는 temp의 2차 항을 모형에 넣었을 때 어떤 모형이 더 적합할지 판단해보겠다.

```
data m;
set m;
temp2=temp**2;
run;

proc logistic data=m descending;
model rainfall=temp temp2 hum wind/rsquare;
run;
```

R-Square	0.3912	Max-rescaled R-Square	0.5439

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	157.3103	4	<.0001
Score	121.7597	4	<.0001
Wald	70.9542	4	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-15.7190	1.9454	65.2853	<.0001
temp	1	-0.0398	0.0431	0.8535	0.3556
temp2	1	-0.00054	0.00150	0.1280	0.7206
hum	1	0.1873	0.0230	66.5337	<.0001
wind	1	1.2681	0.2962	18.3241	<.0001

그 결과 모형의 설명력은 비슷한 수준이었지만 이차항을 넣으니 temp와 temp의 이차항 모두 유의하지 않다는 결과를 얻을 수 있었다. 따라서 변수항을 추가하였지만 설명력은 좋아지지 않고 유의하지 않은 변수들만 나타났다는 걸 알 수 있었다. 이런 과정과 비슷하게 최대 이차항으로 모든 가능한 변수들의 조합에 대해 가장 적합한 모형을 찾아보려 한다.

```
data m;
set m;
wind2=wind**2;
hum2=hum**2;
wh=wind*hum;
th=temp*hum;
tw=temp*wind;
run;

proc logistic data = m descending;
model rainfall=temp hum wind temp2 hum2 wind2 wh th tw/selection=backward;
run;
quit;
```

모든 가능한 변수들을 만들었고 그것을 기반으로 backward 스텝으로 가장 적합한 모형을 찾아보도록 한다.

Note: No (additional) effects met the 0.05 significance level for removal from the model.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	hum2	1	7	0.0395	0.8425
2	wh	1	6	0.8537	0.3555
3	temp2	1	5	2.4723	0.1159

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-18.0388	2.9573	37.2066	<.0001
temp	1	-0.3006	0.1059	8.0527	0.0045
hum	1	0.1499	0.0274	29.8178	<.0001
wind	1	5.0522	1.7508	8.3272	0.0039
wind2	1	-0.7164	0.3250	4.8580	0.0275
th	1	0.00369	0.00159	5.4158	0.0200

그 결과 temp, hum, wind, wind^2, temp*wind가 독립변수로 들어간 모형이 가장 적합하다는 것을 알 수 있었다. 그 변수들만을 이용해 로지스틱 회귀적합을 해보겠다.

```
proc logistic data=m descending;
model rainfall=temp hum wind wind2 th/rsquare influence lackfit;
run;
```

R-Square	0.4059	Max-rescaled R-Square	0.5644

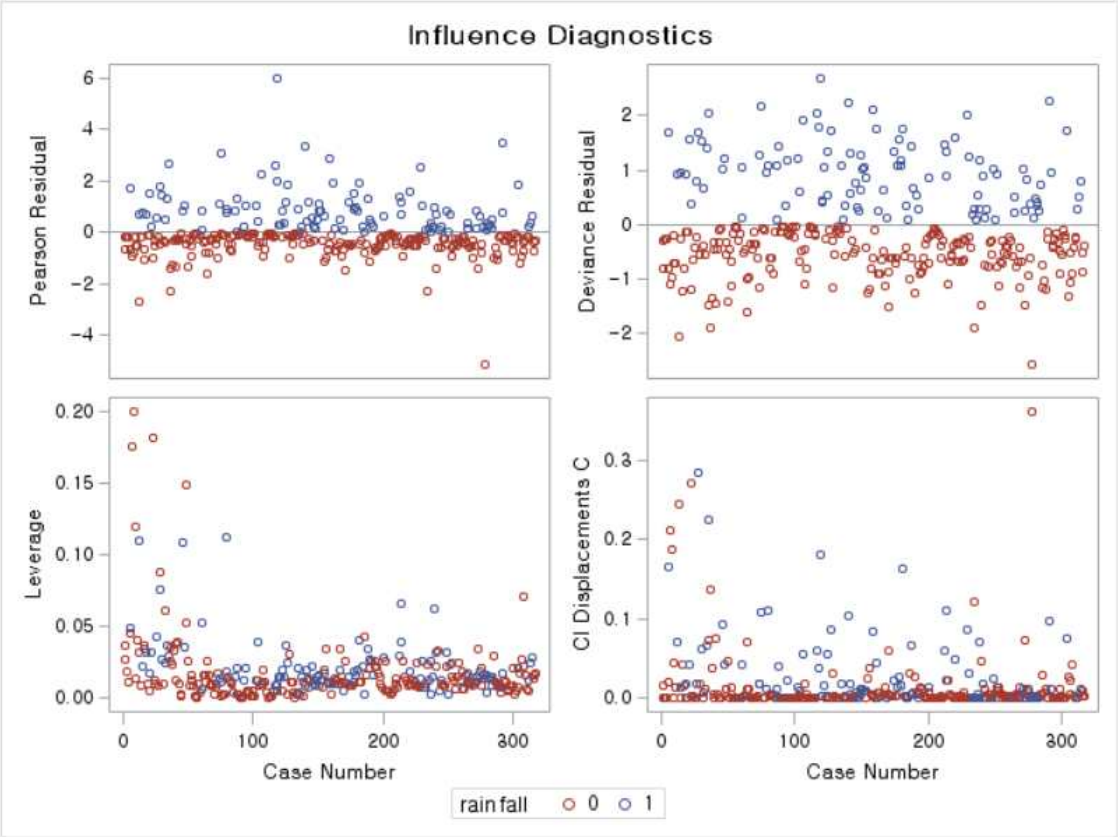
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	165.0551	5	<.0001
Score	124.0949	5	<.0001
Wald	67.6353	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-18.0388	2.9573	37.2066	<.0001
temp	1	-0.3006	0.1059	8.0527	0.0045
hum	1	0.1499	0.0274	29.8178	<.0001
wind	1	5.0522	1.7508	8.3272	0.0039
wind2	1	-0.7164	0.3250	4.8580	0.0275
th	1	0.00369	0.00159	5.4158	0.0200

그 결과 rsqaure값은 약 0.4로 단순한 모형에 비해 조금 상승한 모습을 보였고 각 회귀계수들에 대해 살펴보면 temp와 wind의 이차항을 제하고는 모두 비나 눈이 올 확률이 커지는 데 영향을 준다는 것을 알 수 있고 모든 회귀계수들이 5% 유의수준 하에서 모두 유의하다고 말할 수 있다. 기울기를 제외하고는 평균 상대습도가 가장 유의하게 나타난다. 하지만 여전히 LR 통계량이나 score 통계량을 보았을 때는 유의확률이 매우 작은 값으로 나타나 모형의 설명력에 의문이 생기긴 한다. 다음으로는 호스머-램쇼 검정을 진행해보았다.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
3.7738	8	0.8769

적합결여 검정 결과를 보면 유의확률이 0.8을 넘으며 굉장히 높은 값을 보인다. LR 통계량을 보았을 때는 적합하지 않지만 호스머-램쇼 통계량을 보았을 때는 유의확률이 매우 높으므로 위 모형을 가지고 예측을 진행해보려 한다.



또한 잔차들 그림을 보게 되면 몇몇 이상치나 영향을 주는 값들은 보이지만 조절해줘야 할 정도로 많다고 판단하진 않았다. 따라서 위 모형을 예측에 활용할 모형으로 선택하고자 한다. 이 모형을 토대로 새로운 데이터들에 대해 예측해보고 얼마나 정확하게 예측할 수 있는지 보도록 하겠다. 모형 적합에 사용하지 않은 11월 14일부터 11월 18까지의 데이터를 이용하도록 하겠다. 아래에 나온 데이터를 이용할 것이다.

지점	지점명	일시	평균기온(°	일강수량(r	평균 풍속(	평균 상대습도(%)
108	서울	2021-11-14	11.3		2.2	69.8
108	서울	2021-11-15	10.1		1.6	71.3
108	서울	2021-11-16	8.5		1.9	65.5
108	서울	2021-11-17	7.5		1.6	64.1
108	서울	2021-11-18	11.9	0	2.4	70.8

14일부터 18일은 모두 비나 눈이 오지 않았다. 위에서 채택한 모형에 변수값을 넣어 적합값을 구해보았다.

```
data test;
input temp wind hum @@;
cards;
11.3 2.2 69.8 10.1 1.6 71.3
8.5 1.9 65.5 7.5 1.6 64.1
11.9 2.4 70.8
;
run;

data test;
set test;
wind2=wind**2;
th=temp*hum;
run;

data test;
set test;
prob=1/(1+exp(18.0388+0.3006*temp-0.1499*hum-5.0522*wind+0.7164*wind2-0.00369*th));
run;
```

	temp	wind	hum	wind2	th	prob
1	11.3	2.2	69.8	4.84	788.74	0.3977987662
2	10.1	1.6	71.3	2.56	720.13	0.1854010938
3	8.5	1.9	65.5	3.61	556.75	0.1534146678
4	7.5	1.6	64.1	2.56	480.75	0.0653016532
5	11.9	2.4	70.8	5.76	842.52	0.5261498187

참값은 모두 0이지만 추정값은 각각 약 0.4, 0.18, 0.15, 0.06, 0.53으로 나타났다. 비록 다섯 번째 추정값이 0.5보다 크지만 나름대로 비나 눈이 올 확률을 잘 예상한다고 판단하였다. 따라서 해당 모형이 비나 눈이 올 확률을 잘 예상한다고 생각해 비나 눈이 올 확률을 잘 설명해주는 모형이라고 생각한다.

2.3 분석의 타당성 설명

처음에 설계하였던 대로 독립변수에 평균 기온, 평균 상대습도와 평균 풍속을 넣고 비가 오는지 여부를 분석을 해보았다. 등분산성과 같은 모형의 가정들은 지켜지고 있었고 회귀계수들 또한 유의한 것으로 나왔다. rsquar 값이 0.39정도로 매우 설명력 있다고 말할 수 있는 수준은 아니었지만 나름대로 설명력이 있다고 판단하였다. 그렇지만 조금이라도 설명력을 높힐 수 있는 방안에 대해 생각해보았고 각 변수들

의 이차항과 변수들끼리의 교호작용을 추가한 가장 복잡한 모형부터 변수가 없는 가장 단순한 모형까지 모두 고려해 가장 타당한 모형을 선택하기로 하였다. 그리고 그 결과는 기존 세 개의 변수들과 평균 풍속의 이차항, 평균 기온과 평균 상대습도의 교호작용을 독립변수로 추가한 모형이 가장 적합하다고 말하고 있다. 하지만 모든 가능한 모형들의 LR 통계량과 score 값들의 유의확률은 모두 0.0001보다 작은 값을 보였고 위 모형의 적합성에 대해 의구심이 들기도 하였다. 그렇지만 호스머-램쇼 통계량은 항상 모형이 적합하다고 말해주고 있고 이 모형을 보았을 때 회귀계수들도 유의하고 오차들의 분포에도 큰 문제가 없다고 판단하여 이 모형을 선택하기로 하였다. 또한, 적합에 사용하지 않은 데이터들을 이용해 예측을 해보았을 때도 큰 문제가 없어 보여 예측에 잘 사용할 수 있다고 생각한다.

3.결론

3.1 분석 결과 요약

비나 눈이 올 확률을 예측하고자 평균기온, 평균 상대습도와 평균 풍속을 예측변수로 설정하여 사용하였다. 여러 가능한 모형들을 고려해본 결과 평균기온, 평균 상대습도, 평균 풍속, 평균 풍속의 이차항과 평균기온과 평균 상대습도와 교호작용까지 포함하는 모형이 가장 설명력이 높고 적절하다고 판단하였다. 모형을 수식화해서 나타내 보겠다.

$$P(y=1|tmp,hum,wind)=\frac{1}{1+e^{18.0388+0.3006*tmp-0.1499*hum-5.8522*wind+0.7164*wind^2-0.00369*temp*hum}}$$

와 같이 표현이 가능하며 각 변수들과 비나 눈이 올 확률의 관계에 대해 말해보면 평균 기온이 낮을수록, 평균 상대습도가 높을수록, 평균 풍속이 높을수록, 평균 풍속의 제곱이 낮을수록, 평균 기온과 평균 상대습도의 교호작용이 클수록 비나 눈이 올 확률이 커진다고 말할 수 있다. 여기서 평균 풍속이 높을수록 비나 눈이 올 확률이 커지지만 평균 풍속의 제곱은 낮을수록 비나 눈이 올 확률이 커지므로 둘의 관계가 모순적이라고는 생각한다. 하지만 회귀계수가 유의하게 나왔기 때문에 그대로 쓰기로 결정하였고 평균 풍속이 너무 높거나 낮으면 비나 눈이 올 확률이 낮아진다고 추측해본다.

3.2 분석의 장점 및 한계점 설명

비나 눈이 올 확률을 평균기온, 평균 상대습도와 평균 풍속 총 세 가지 변수로 예측할 수 있다는 건 어마어마한 장점이다. 물론 변수를 많이 사용하는 것이 예측의 정확성을 높힐 순 있지만, 변수를 더 많이 사용하는 비용 대비 더 많은 정확성을 가져갈 수 있을지는 의문인 부분이다. 그렇지만 유의미한 변수들을 더 많이 사용한다면 지금보다 성능이 좋아질 것은 분명하다. 한계점도 존재한다. 우선 모형의 적합성에 대해 한계가 있을 수도 있다. 물론 회귀계수와 호스머-램쇼 통계량이 모형이 적합한 쪽으로 값들이 도출됐지만 LR 통계량과 score값 모두 유의확률이 매우 낮으며 모형이 적절하다는 귀무가설을 충분히 기각시킬 수 있다. 우선 예측을 할 수 있다고 판단하여 예측을 진행하였지만 이러한 한계점을 보완해나가야 한다고 생각한다. 이 외에도 한계점은 또 존재한다. 모델링을 사용할 때 사용한 변수 모두 실제로는 미래를 예측한 변수이다. 예측한 변수들을 토대로 비나 눈이 올 확률을 예측하는 것이다. 예측한 변수들을 토대로 예측하는 것이니 성능이 좋다고 보기 애매하다. 게다가 다른 기상정보를 예측하여 얻을 수 있으면 비나 눈이 올 확률을 손쉽게 예측해서 얻을 수 있을 거 같다.

3.3 추가 연구사항 제안

3.2절에서도 말했듯이 평균기온, 평균 상대습도와 평균 풍속 이외에도 다양한 변수를 더 추가한다면 설명력이 더 좋아질 것 같다. 이번 연구에서는 최대 2차 항까지만 고려하여 모형을 선택하는 과정을 거쳤지만, 더 높은 차수를 고려한다면 더 설명력이 좋은 모형을 만들 수 있다. 또한, 모형의 적합성을 보장하기 위해 다른 통계량의 결과들은 괜찮았지만, LR 통계량도 모형이 적합하다는 결론이 나올 수 있도록 보완을 할 필요가 있어 보인다. 로지스틱 회귀분석을 사용하고 테스트 데이터를 통해서도 예측을 시도하였지만, 그 모델의 성능을 평가하는 ROC 커브와 같은 평가지표들을 활용하지 못하였다. 로지스틱 회귀분석을 평가하는 지표를 이용해 모형의 설명력을 더욱 입증하였다면 여러 비교를 통해 더 좋은 모형을 선택할 수 있을 거 같다. 로지스틱 회귀분석의 기법이 성능이 나쁜 기법이 아니지만, 앙상블과 같은 다른 머신러닝 기법들을 적용해보면 어떤 학습을 할지에 호기심이 생기며 기회가 된다면 다른 머신러닝 기법들을 이용해 학습시켜보고 싶다.

－참고문헌

https://www.dongascience.com/news.php?idx=49611

https://data.kma.go.kr/cmmn/static/staticPage.do?page=intro