

다중회귀분석 & 로지스틱 회귀분석을 이용한 기상환경에 따른 모기지수 분석

데이터분석응용통계 기말 프로젝트

유경빈(팀장) 202104246 ICT융합공학부

정지용 201804108 소프트웨어응용학부

문지혜 202004042 소프트웨어응용학부

INDEX

1. 팀 & 프로젝트 주제 소개

2. 분석

2-1. 데이터 전처리

2-2. 다중회귀분석

2-3. 로지스틱 회귀분석

3. 분석 결과

4. 마무리

1. 팀 & 프로젝트 주제 소개

- ① 팀원 소개
- ② 프로젝트 주제 소개

01. 팀 소개



유경빈 (팀장)

- 학과: ICT 융합공학부
- 학번: 202104246
- 역할
 - (공동) 데이터 분석 및 해석
 - 발표자료 제작
 - 발표



정지용

- 학과: 소프트웨어응용학부
- 학번: 201804108
- 역할
 - (공동) 데이터 분석 및 해석
 - 데이터 분석 리드 개발
 - R 데이터 전처리



문지혜

- 학과: 소프트웨어응용학부
- 학번: 202004042
- 역할
 - (공동) 데이터 분석 및 해석
 - 데이터 수집
 - 발표자료 제작

02. 프로젝트 주제 소개

" 다중회귀분석 & 로지스틱 회귀분석을 이용한 **기상 환경에 따른 모기지수 분석** "

가설 설정

모기지수에 영향을 주는 것은 무엇일까?

⋮

- **귀무가설:** 모기 지수는 기상 환경에 영향을 받지 않는다.

- **대립가설:** 모기 지수는 기상 환경에 큰 영향을 받는다.

⋮

분석 진행

모기지수 데이터와 기상 환경 데이터를 사용하여

다중회귀분석, 로지스틱 회귀분석을 진행하여 결과 비교

02. 프로젝트 주제 소개

" 다중회귀분석 & 로지스틱 회귀분석을 이용한 **기상 환경에 따른 모기지수 분석** "

모기지수란?

모기발생 상황을 지수화하여 단계 별 시민행동 요령을 전달하기 위한 지수.



보건

활용사례(갤러리) 등록

URL 복사

목록 이동

서울시 모기예보제 정보

서울지역 모기발생 상황을 지수화하여 모기발생 단계별 시민행동요령을 알려주는 일일 모기발생 예보서비스입니다.

※ 모기 활동 지수는 5.1~10.31 기간에만 측정되오니, 이용에 참고해주시기 바랍니다.

※ 모기지수값이 아래와같이 3가지로 변경됩니다. (2020.4.14)

- mosquito_value(모기지수) -> mosquito_value_water(모기지수(수변부))/mosquito_value_house/모기지수(주거지)),


[전체 설명보기](#)

02. 프로젝트 주제 소개

" 다중회귀분석 & 로지스틱 회귀분석을 이용한 **기상 환경에 따른 모기지수 분석** "

서울 열린데이터 광장
서울시 모기예보제 정보 Data

공공데이터



보건

활용사례(갤러리) 등록

URL 복사

목록 이동

서울시 모기예보제 정보

서울지역 모기발생 상황을 지수화하여 모기발생 단계별 시민행동요령을 알려주는 일일 모기발생 예보서비스입니다.

※ 모기 활동 지수는 5.1~10.31 기간에만 측정되오니, 이용에 참고해주시기 바랍니다.

※ 모기지수값이 아래와같이 3가지로 변경됩니다. (2020.4.14)

[전체 설명보기](#)

기상자료개방포털
종관기상관측(ASOS) 서울시 자료

종관기상관측(ASOS) - 자료

자료설명

자료설명

• 자료설명

종관기상관측이란 종관규모의 날씨를 파악하기 위하여 정해진 시각에 모든 관측소에서 같은 시각에 실시하는 지상관측을 말합니다. 종관규모는 일기도에 표현되어 있는 보통의 고기압이나 저기압의 공간적 크기 및 수명을 말하며, 주로 매일의 날씨 현상을 뜻합니다.

자료형태	분, 시간(매정시), 일, 월, 연	제공기간	1904년~(지점별, 요소별 다름)
제공지점	103개 * 원하는 지점이 없는 경우, 방재기상관측(AWS) 메뉴 이용	제공요소	기온, 강수, 바람, 기압, 습도, 일사, 일조, 눈, 구름, 시정, 지면상태, 지면·초상온도, 일기현상, 증발량, 현상번호
유의사항	- 1회 조회 가능 최대 기간: 분 1일, 시간 1년, 일 10년, 월·연 제한 없음(장기간 자료는 '파일셋 조회' 메뉴 이용) - 시간/분 자료에 대해 관측값의 정상 여부를 판단하는 품질검사 플래그(QC FLAG) 정보 제공 * 제공 요소: 기온, 습도, 기압, 지면온도, 풍향, 풍속, 일조 / 플래그 종류(의미): 0(정상), 1(오류), 9(결측) - 전일 자료는 당일 10시 이후 확인 가능		
비고	- 10분 또는 1시간 최다강수시각은 최다강수가 나타난 시작 시간으로, (-) 표기가 있는 경우 전날을 뜻함 - 강수량은 겨울철(11월~익년 3월) 3시간 간격으로 제공		
지침	요소별 관측방법이나 자료 산출방식에 대한 상세 설명은 ☞ [지상기상관측지침] 참조		

2. 분석

- ① 데이터 전처리
- ② 다중회귀분석
- ③ 로지스틱 회귀분석

01. 데이터 전처리

1-1. 데이터 전처리 과정 소개

01 단계 수집 영역 설정 및 데이터 수집
주제에 대한 데이터를 선정하고 수집할 범위를 선정
공공데이터, 대회 등의 자료를 통해 분석할 데이터를 수집

02 단계 R 프로젝트로 데이터 불러오기

03 단계 R을 활용한 중복값, 결측치 제거
R을 사용하여 중복값이나 결측치를 제거하여 분석을 진행

01. 데이터 전처리

1-2. 데이터 수집 영역 설정 및 데이터 수집

서울 열린데이터 광장에서 '서울시 모기예보제 정보' 데이터를 수집을 진행

공공데이터

활용사례(갤러리) 등록

URL 복사

목록 이동

보건

서울시 모기예보제 정보

서울지역 모기발생 상황을 지수화하여 모기발생 단계별 시민행동요령을 알려주는 일일 모기 발생 예보서비스입니다.
※ 모기 활동 지수는 5.1~10.31 기간에만 측정되오니, 이용에 참고해주시기 바랍니다.
※ 모기지수값이 아래와같이 3가지로 변경됩니다. (2020.4.14)

전체 설명보기

데이터 정보

공개일자	2016.07.01.
최신수정일자	2023.05.26.
갱신주기	매일
분류	보건
원본시스템	모기예보제
저작권자	서울특별시
제공기관	서울특별시
제공부서	시민건강국 감염병관리과
담당자	송미영 (02-2133-7612)
원본형태	DB
제3저작권자	없음
라이선스	 저작권자표시(BY): 이용이나 변경 및 2차적 저작물의 작성을 포함한 자유이용을 허락합니다.
관련태그	모기, 모기발생, 모기지수, 모기유충, 모기예보



DATA 출처

- 공공데이터 명: 서울시 모기예보제 정보
- 출처
 - 사이트 명: 서울 열린데이터 광장
 - 링크: <https://data.seoul.go.kr/>
- 제공처: 서울특별시 시민건강국 감염병관리과
- 제공일: 2016.07.01 ~ 현재
- 제공 확장자: .csv, .json

01. 데이터 전처리

1-2. 데이터 수집 영역 설정 및 데이터 수집

서울 열린데이터 광장에서 '서울시 모기예보제 정보' 데이터를 수집을 진행

모기지수 발생일 ▼	모기지수(수변부)	모기지수(주거지)	
2023-05-26	94.0	37.3	33.6
2023-05-25	90.7	37.8	33.5
2023-05-24	100.0	39.4	33.1
2023-05-23	100.0	30.0	27.6
2023-05-22	100.0	39.6	33.3
2023-05-21	100.0	40.4	35.1
2023-05-20	100.0	40.8	33.3
2023-05-19	100.0	40.7	35.7
2023-05-18	100.0	41.5	36.5
2023-05-17	81.1	37.1	30.9
2023-05-16	77.8	32.4	26.7
2023-05-15	72.2	31.8	24.9
2023-05-14	74.0	30.9	25.9
2023-05-13	79.2	33.3	27.7
2023-05-12	61.0	31.2	24.5
2023-05-11	61.0	29.5	22.2
2023-05-10	53.8	26.4	19.4



Column 정보

- Columns
 - 모기지수 발생일
 - 모기지수(수변부)
 - 모기지수(주거지)
 - 모기지수(공원)
- DataSet: 1918개 데이터

01. 데이터 전처리

1-2. 데이터 수집 영역 설정 및 데이터 수집

기상자료개방포털에서 종관기상관측(ASOS) 데이터 중 서울시 자료를 수집

종관기상관측(ASOS) - 자료

자료설명

종관기상관측이란 종관규모의 날씨를 파악하기 위하여 정해진 시각에 모든 관측소에서 같은 시각에 실시하는 지상관측을 말합니다. 종관규모는 일기도에 표현되어 있는 보통의 고기압이나 저기압의 공간적 크기 및 수명을 말하며, 주로 매일의 날씨 현상을 뜻합니다.

자료형태	분, 시간(매정시), 일, 월, 연	제공기간	1904년~(지점별, 요소별 다름)
제공지점	103개 * 원하는 지점이 없는 경우, 방재기상관측(AWS) 메뉴 이용	제공요소	기온, 강수, 바람, 기압, 습도, 일사, 일조, 눈, 구름, 시정, 지면상태, 지면·초상온도, 일기현상, 증발량, 현상번호
유의사항	- 1회 조회 가능 최대 기간: 분 1일, 시간 1년, 일 10년, 월·연 제한 없음(장기간 자료는 '파일셋 조회' 메뉴 이용) - 시간/분 자료에 대해 관측값의 정상 여부를 판단하는 품질검사 플래그(QC FLAG) 정보 제공 * 제공 요소: 기온, 습도, 기압, 지면온도, 풍향, 풍속, 일조 / 플래그 종류(의미): 0(정상), 1(오류), 9(결측) - 전일 자료는 당일 10시 이후 확인 가능		
비고	- 10분 또는 1시간 최다강수시각은 최다강수가 나타난 시작 시간으로, (-) 표기가 있는 경우 전날을 뜻함 - 강수량은 겨울철(11월~익년 3월) 3시간 간격으로 제공		
지침	요소별 관측방법이나 자료 산출방식에 대한 상세 설명은 [지상기상관측지침] 참조		

DATA 필터링

전체

강원도

경기도

경상남도

경상북도

광주광역시

대구광역시

대전광역시

부산광역시

서울특별시

세종특별자치시

울산광역시

인천광역시

기온

강수

바람

습도

기압

일조, 일사

눈

구름

지면.지중온도

증발량

일기현상

01. 데이터 전처리

1-2. 데이터 수집 영역 설정 및 데이터 수집

기상자료개방포털에서 종관기상관측(ASOS) 데이터 중 서울시 자료를 수집

일시	평균기온	일강수량	평균상대습도	합계일사량
2018-01-01	-1.3		39.1	6.14
2018-01-02	-1.8		42	5.36
2018-01-03	-4.7		42.3	6.56
2018-01-04	-4.7		43	4.73
2018-01-05	-3		48.4	5.96
2018-01-06	-2.8		45.8	6.79
2018-01-07	-0.8		42.1	3.75
2018-01-08	1.3	0.9	51.9	2.39
2018-01-09	-4.2	0.5	59.4	5.6
2018-01-10	-7.5	0.3	52	6.85
2018-01-11	-11.1		49.8	6.8
2018-01-12	-10.2	0	35.4	7.11
2018-01-13	-4.4	0.4	67.3	2.8
2018-01-14	0.6		64.1	4.55
2018-01-15	4.7	0.2	63.9	4.67
2018-01-16	3.5	0	59.1	2.4
2018-01-17	4.5		64.1	4.17



Column 정보

- Columns
 - 평균기온
 - 일강수량
 - 평균상대습도
 - 합계일사량
- DataSet: 1827개 데이터

01. 데이터 전처리

1-3. R 프로젝트에 데이터 세팅

분석을 위해 R 프로젝트로 '모기지수' 데이터와 '기상 관측' 데이터를 불러와 분석을 위해 데이터를 세팅

```
10 # 데이터 파일 불러오기
11 weather_data <- read.csv("C:\\mosquito\\OBS_ASOS_DD.csv",
12                           header = TRUE, stringsAsFactors = TRUE, sep = ",", fileEncoding = 'euc-kr')
13 mosq_data <- read.csv("C:\\mosquito\\mosquito.csv",
14                       header = TRUE, stringsAsFactors = TRUE, sep = ",", fileEncoding = 'euc-kr')
15 View(weather_data)
16 View(mosq_data)
```

	일시	평균 기온	일강 수량	평균 상대 습도	합계 일사 량
1	2018-01-01	-1.3	NA	39.1	6.14
2	2018-01-02	-1.8	NA	42.0	5.36
3	2018-01-03	-4.7	NA	42.3	6.56

weather_data

	일시	수변 부모 기지 수	주거 지모 기지 수
1	2018-01-01	5.5	5.5
2	2018-01-02	5.5	5.5
3	2018-01-03	5.5	5.5

mosq_data

01. 데이터 전처리

1-3. R 프로젝트에 데이터 세팅

컬럼명 변경 및 '일시'를 기준으로 데이터 병합

```
18 # 컬럼명 변경
19 names(weather_data) <- c('일시', '평균기온', '일강수량', '평균상대습도', '합계일사량')
20 names(weather_data)
21 names(mosq_data) <- c('일시', '수변부모기지수', '주거지모기지수')
22 names(mosq_data)
23
24 # 날짜를 키값으로 데이터 병합
25 mg_data <- merge(weather_data, mosq_data, by='일시')
26 View(mg_data)
```



	일시	평균 기온	일강 수량	평균 상대 습도	합계 일사 량	수변 부모 기지 수	주거 지모 기지 수
1	2018-01-01	-1.3	NA	39.1	6.14	5.5	5.5
2	2018-01-02	-1.8	NA	42.0	5.36	5.5	5.5
3	2018-01-03	-4.7	NA	42.3	6.56	5.5	5.5

mg_data

01. 데이터 전처리

1-4. R 언어를 활용한 중복값, 결측치 제거

데이터의 중복값 확인, 결측치 확인 및 제거

1. 중복값 확인

```
28 # -----EDA전처리-----  
29 #중복값 확인  
30 duplicates <- mg_data %>% duplicated() %>% table()  
31 duplicates
```

> duplicates

	FALSE	TRUE
	1804	113

중복된 데이터가 113개 존재

2. 결측치 확인 및 제거

```
33 #결측치 확인  
34 table(is.na(mg_data))  
  
40 # 결측치 확인 후 제거  
41 is.na(c_data)  
42 c_data <- na.omit(c_data)  
43 table(is.na(c_data))
```

> table(is.na(mg_data))

	FALSE	TRUE
	12226	1193

결측된 데이터가 1193개 존재

* 분석 목표 *

" 다중회귀분석 & 로지스틱 회귀분석을 이용한 **기상 환경에 따른 모기지수 분석** "

다중회귀분석

" 여러 기상 환경 변수들로부터 수변부와 주거지의 모기지수를 예측한다. "

로지스틱 회귀분석

" 여러 기상 환경 변수들을 통해
수변부와 주거지의 모기 번식 비율이 증가할 확률을 예측한다"

② 다중회귀분석

- ① 데이터 전처리
- ② 다중회귀분석
- ③ 로지스틱 회귀분석

02. 다중회귀분석

2-1. 다중회귀분석의 상세 분석 단계

다중 회귀 분석

하나의 모형을 자료에
적합시킨 뒤 이를 이용하여

여러 개의 예측변수로부터
한 종속 변수를 예측하는 것

01 단계

상관관계 확인 (시각화)

02 단계

다중회귀분석 진행

> 03 단계

다중공선성 AIC 확인

04 단계

각 데이터간의 회귀선, 상관계수, 상관계수검정 값 (시각화)

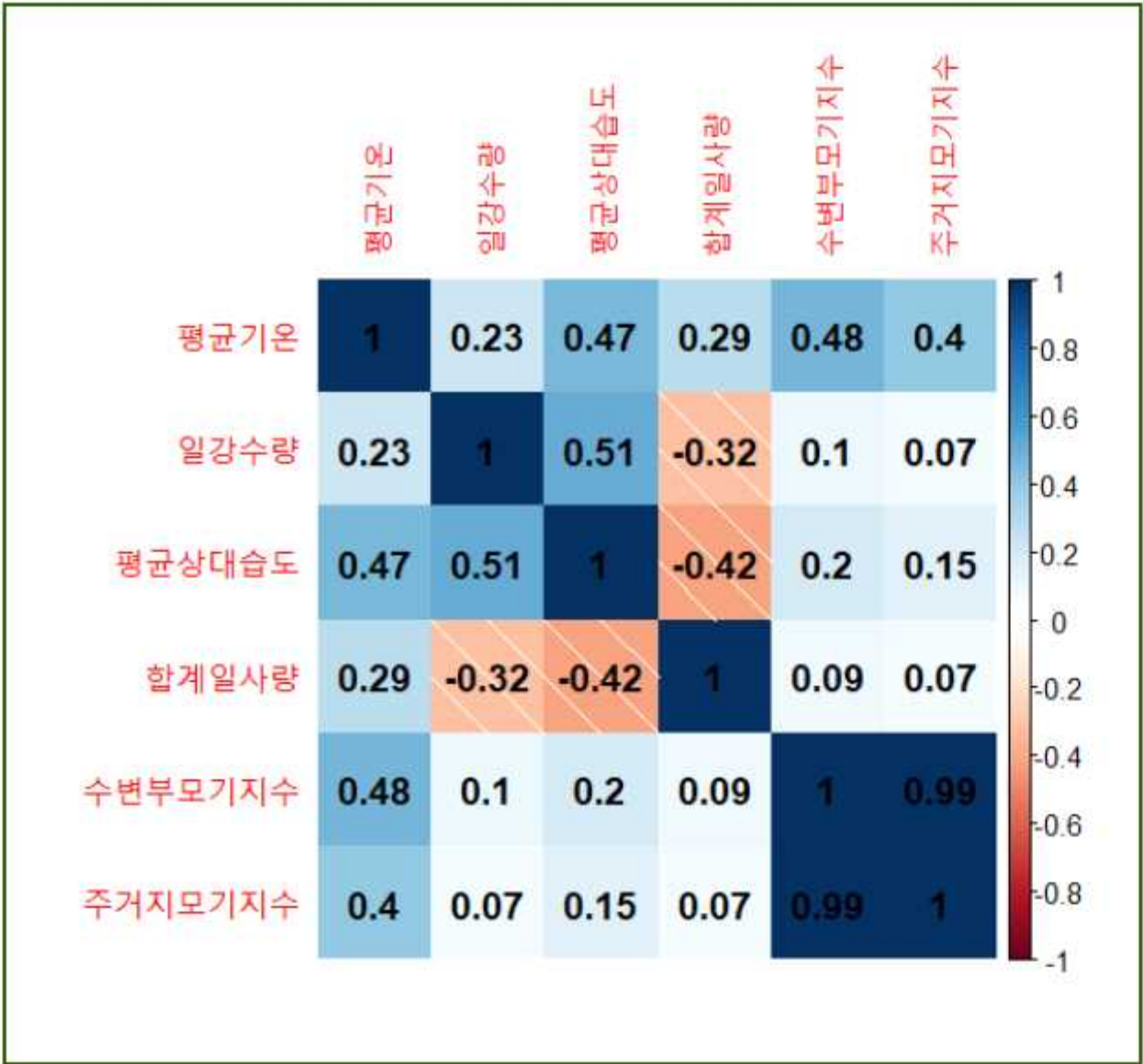
05 단계

결과 해석

02. 다중회귀분석

2-2. 상관관계 확인 (시각화)

상관계수 행렬을 시각화 하기 위하여 corrplot 패키지를 사용하여 그래프를 제작



```
45 # 상관관계 확인 (시각화)
46 M = cor(c_data)
47 corrplot(M, method = 'shade', addCoef.col = "black")
```

46: cor(c_data)를 통해 상관 계수 행렬을 계산

47: corrplot 함수를 통해 그래프 제작.

상관계수 그래프 출력 결과

02. 다중회귀분석

2-3. 다중회귀분석 진행

1) 다중회귀분석 진행 결과 '일강수량' 회귀계수가 유의미하지 않음이 판별됨.

```
52 lm_1 <- lm(수변 부모기 지수 ~ 평균기온 + 일강수량 + 평균상대습도 + 합계일사량, data = c_data)
53 summary(lm_1)
54 lm_2 <- lm(주거지모기지수 ~ 평균기온 + 일강수량 + 평균상대습도 + 합계일사량, data = c_data)
55 summary(lm_2)
```

Call:

```
lm(formula = 수변부모기 지수 ~ 평균기온 + 일강수량 +
    평균상대습도 + 합계일사량, data = c_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-277.48	-100.75	-51.00	16.96	784.69

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	115.5491	55.0124	2.100	0.0360 *
평균기온	11.4526	0.9303	12.310	<2e-16 ***
일강수량	-0.1182	0.4134	-0.286	0.7750
평균상대습도	-1.5350	0.7586	-2.024	0.0434 *
합계일사량	-3.7462	1.4596	-2.567	0.0105 *

Call:

```
lm(formula = 주거지모기지수 ~ 평균기온 + 일강수량 +
    평균상대습도 + 합계일사량, data = c_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-240.88	-118.87	-57.98	22.56	823.28

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	149.6724	59.0057	2.537	0.01140 *
평균기온	10.7108	0.9979	10.734	< 2e-16 ***
일강수량	-0.1559	0.4434	-0.352	0.72529
평균상대습도	-2.1016	0.8136	-2.583	0.00998 **
합계일사량	-4.5634	1.5655	-2.915	0.00366 **

- 평균 기온이 가장 유의하다는 결과를 보였으며, 일 강수량이 가장 무의하다는 결과를 보임.

02. 다중회귀분석

2-3. 다중회귀분석 진행

2) '일강수량' 회귀계수를 제거 후 다중회귀분석을 진행.

```
57 # 일강수량의 값이 유의미하지 않기때문에 제거
58 lm_3 <- lm(수변 부모기 지수 ~ 평균기온 + 평균상대습도 + 합계일사량, data = c_data)
59 summary(lm_3)
60 lm_4 <- lm(주거지모기 지수 ~ 평균기온 + 평균상대습도 + 합계일사량, data = c_data)
61 summary(lm_4)
```

Call:

lm(formula = 수변부모기 지수 ~ 평균기온 + 평균상대습도 + 합계일사량, data = c_data)

Residuals:

Min	1Q	Median	3Q	Max
-276.46	-100.77	-51.98	16.37	785.72

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	118.7604	53.8206	2.207	0.0276 *
평균기온	11.4326	0.9271	12.331	<2e-16 ***
평균상대습도	-1.6001	0.7231	-2.213	0.0272 *
합계일사량	-3.6830	1.4419	-2.554	0.0108 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 183.2 on 741 degrees of freedom
Multiple R-squared: 0.2358, Adjusted R-squared: 0.2328
F-statistic: 76.23 on 3 and 741 DF, p-value: < 2.2e-16

Call:

lm(formula = 주거지모기 지수 ~ 평균기온 + 평균상대습도 + 합계일사량, data = c_data)

Residuals:

Min	1Q	Median	3Q	Max
-239.52	-118.16	-57.49	22.54	824.63

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	153.9067	57.7290	2.666	0.00784 **
평균기온	10.6844	0.9945	10.744	< 2e-16 ***
평균상대습도	-2.1875	0.7756	-2.820	0.00493 **
합계일사량	-4.4802	1.5466	-2.897	0.00388 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 196.5 on 741 degrees of freedom
Multiple R-squared: 0.1739, Adjusted R-squared: 0.1705
F-statistic: 51.98 on 3 and 741 DF, p-value: < 2.2e-16

- 모든 회귀계수가 유의미한 결과를 된다는 것이 확인됨.

02. 다중회귀분석

2-3. 다중회귀분석 진행

3) 다중회귀분석 결과

Call:

```
lm(formula = 수변부모기지수 ~ 평균기온 + 평균상대습도 +  
    합계일사량, data = c_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-276.46	-100.77	-51.98	16.37	785.72

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	118.7604	53.8206	2.207	0.0276	*
평균기온	11.4326	0.9271	12.331	<2e-16	***
평균상대습도	-1.6001	0.7231	-2.213	0.0272	*
합계일사량	-3.6830	1.4419	-2.554	0.0108	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 183.2 on 741 degrees of freedom

Multiple R-squared: 0.2358, Adjusted R-squared: 0.2328

F-statistic: 76.23 on 3 and 741 DF, p-value: < 2.2e-16

- 모든 회귀 계수가 유의한 것을 알 수 있음.

- $R^2 = 0.23 \rightarrow$ 설명력이 다소 부족할 수 있음.
- $p\text{-value} < 0.05 \rightarrow$ 모델이 유의함.

02. 다중회귀분석

2-4. 다중공선성 확인

다중공선성 확인을 통해 독립변수 사이에 상관성이 있는 지 판별함.

```
64 vif(lm_3)
65 vif(lm_4)
66 extractAIC(lm_3)
67 extractAIC(lm_4)
```

```
> vif(lm_3)
    평균기온    평균상대습도    합계일사량
    1.999391    2.225338    1.890916
> vif(lm_4)
    평균기온    평균상대습도    합계일사량
    1.999391    2.225338    1.890916
> extractAIC(lm_3)
[1]    4.000 7767.537
> extractAIC(lm_4)
[1]    4.000 7871.991
```

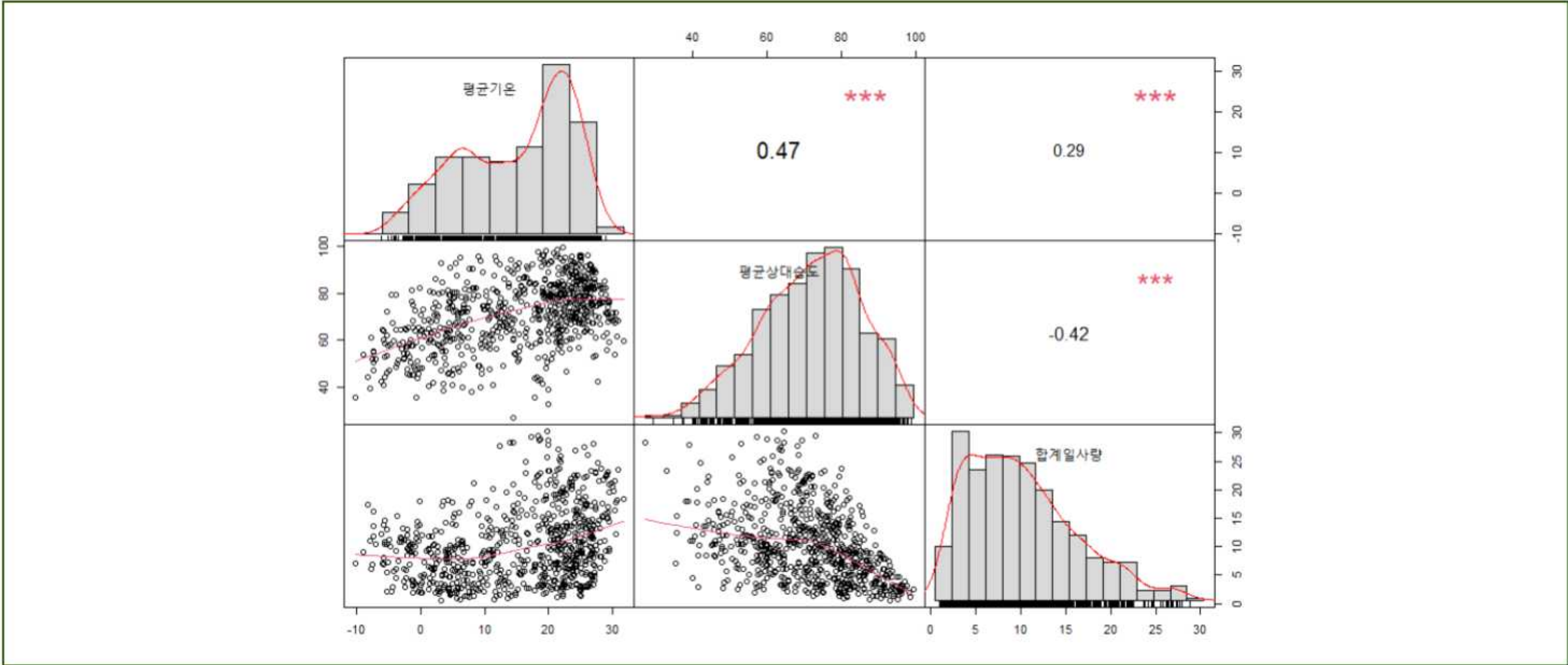
- 평균상대습도를 제외한 설명변수는 $VIF < 2$ 이므로 다중공선성의 영향 거의 없음.
- 평균 상대습도 설명변수는 $2 < VIF < 5$ 이므로 영향은 거의 없으나 주의가 필요함.

- 대부분의 설명 변수가 다중공선성의 영향이 거의 없으나 '평균상대습도' 만 낮은 수치로 주의가 필요함.

02. 다중회귀분석

2-5. 각 데이터간의 회귀선, 상관계수, 상관계수검정 값(시각화)

1) 회귀선, 상관계수, 상관계수검정 값(시각화) 전체 그래프



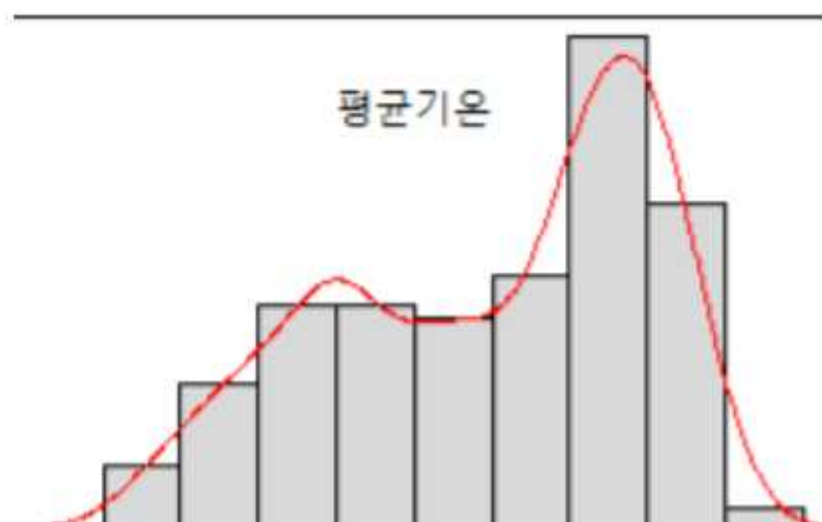
회귀선, 상관계수, 상관계수검정 값(시각화)

02. 다중회귀분석

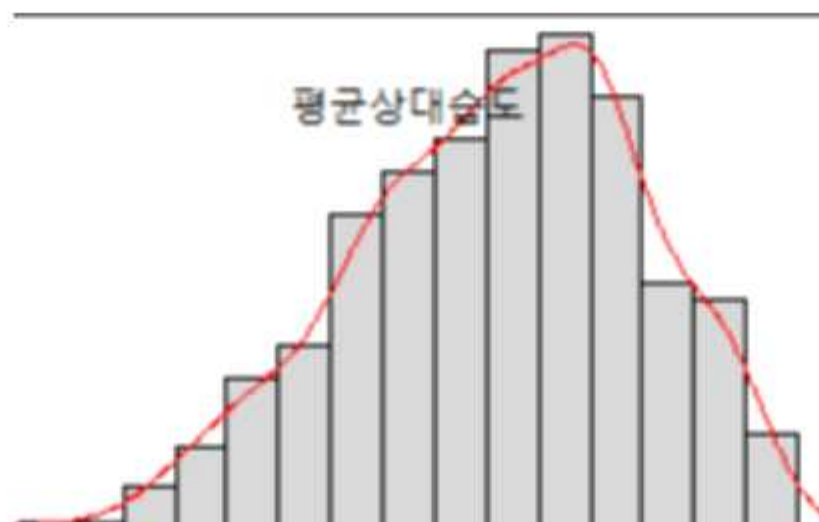
2-5. 각 데이터간의 회귀선, 상관계수, 상관계수검정 값(시각화)

2) 전체그래프 세부 내용

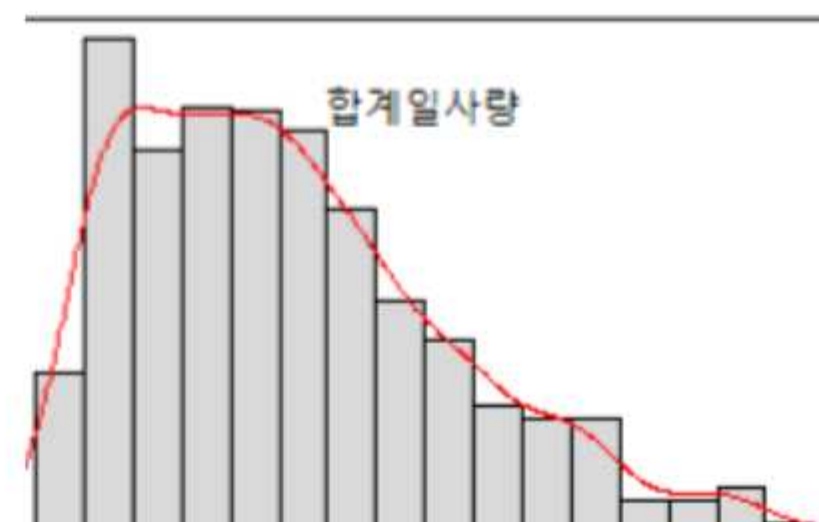
- 평균기온의 히스토그램



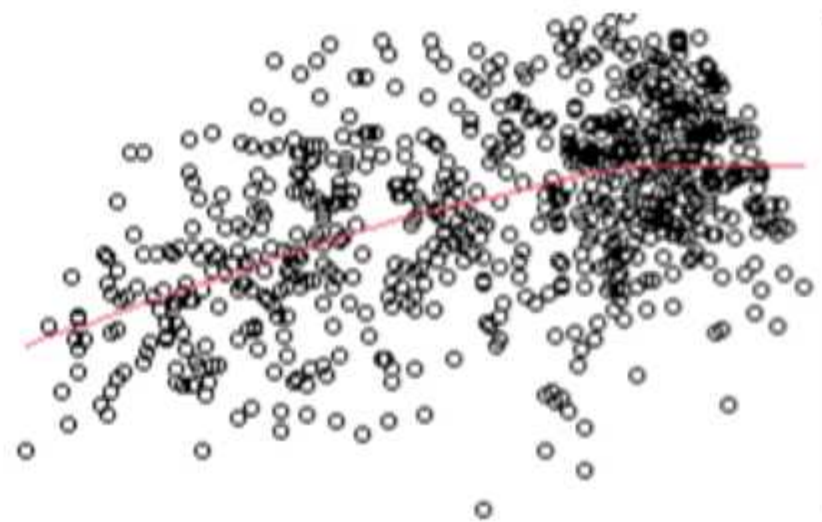
- 평균상대습도의 히스토그램



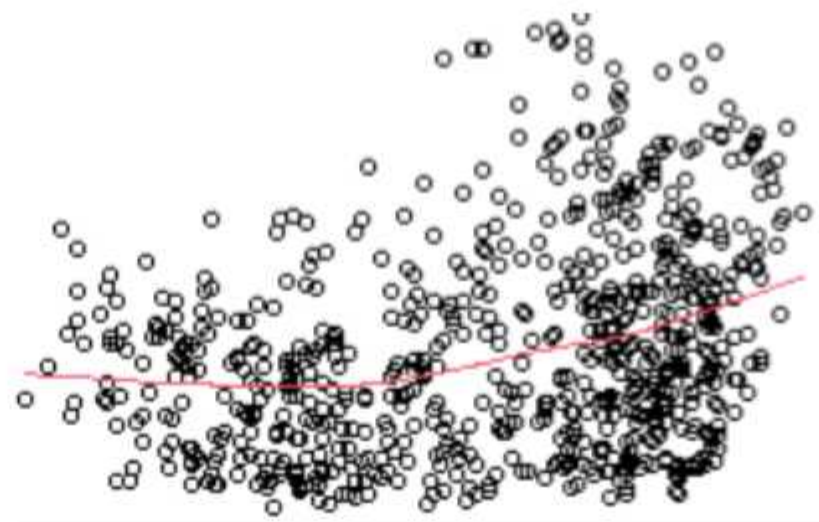
- 합계일사량의 히스토그램



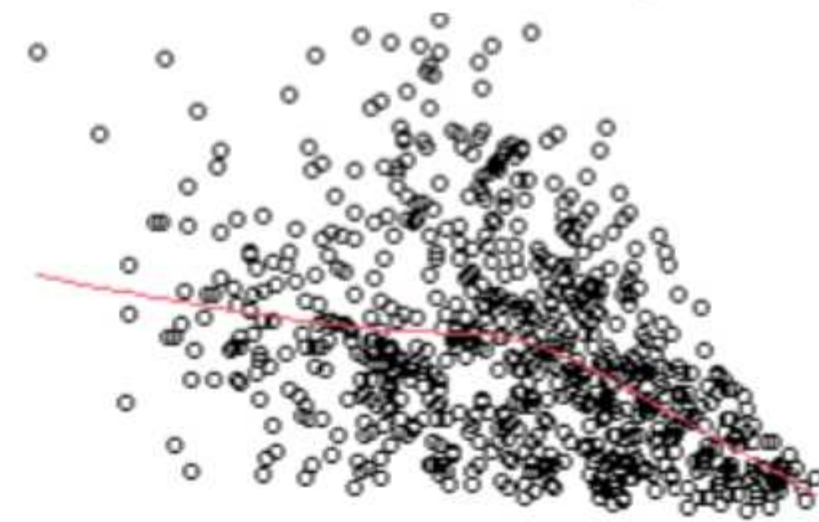
- Scale-Location 그래프



- Scale-Location 그래프



- Residual vs Fitted 그래프



③ 로지스틱 회귀분석

- ① 데이터 전처리
- ② 다중회귀분석
- ③ 로지스틱 회귀분석

03. 로지스틱 회귀분석

3-1. 로지스틱 회귀분석 상세 단계

로지스틱 회귀분석

하나의 모형을 자료에
적합시킨 뒤 이를 이용하여

여러 개의 예측변수로부터
한 종속 변수를 예측하는 것

01 단계 로지스틱 회귀 분석 진행

02 단계 지수변환 값 산출

> 03 단계 AIC & 다중공선성 확인

04 단계 로지스틱 회귀 분석 모델 최적화 및 분석

05 단계 지수변환 값 산출 및 H-L 적합도 검정

03. 로지스틱 회귀분석

3-2. 로지스틱 회귀분석 진행

1) glm을 위해 종속변수 y값을 0, 1로 변환 작업 진행

```
84 c_data$수변 부모기 지 수01 <- ifelse(c_data$수변 부모기 지 수 >= 1, 1,0)
85 c_data$주거 지 모기 지 수01 <- ifelse(c_data$주거 지 모기 지 수 >= 1, 1,0)
```

2) 로지스틱 회귀 분석 진행

```
86 glm_1 <- glm(수변 부모기 지 수01 ~ 평균기온 + 일강수량 + 평균상대습도 + 합계일사량,
87              family = binomial, data = c_data)
88 glm_2 <- glm(주거 지 모기 지 수01 ~ 평균기온 + 일강수량 + 평균상대습도 + 합계일사량,
89              family = binomial, data = c_data)
```

3) 출력

```
90 summary(glm_1)
91 summary(glm_2)
```


03. 로지스틱 회귀분석

3-2. 로지스틱 회귀분석 진행

4) 로지스틱 회귀분석 결과

```
> summary(glm_1)

Call:
glm(formula = 수변부모기지수01 ~ 평균기온 + 일강수량 +
    평균상대습도 + 합계일사량, family = binomial,
    data = c_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2572   0.1163   0.1918   0.3693   1.0965

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.46553    1.29625   2.673  0.00751 **
평균기온       0.14081    0.02409   5.845 5.06e-09 ***
일강수량       0.09559    0.05070   1.886  0.05936 .
평균상대습도  -0.03288    0.01796  -1.831  0.06709 .
합계일사량    -0.02724    0.03799  -0.717  0.47340
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 377.15  on 744  degrees of freedom
Residual deviance: 293.03  on 740  degrees of freedom
AIC: 303.03

Number of Fisher Scoring iterations: 8
```

수변부모기지수 로지스틱 회귀분석 결과

```
> summary(glm_2)

Call:
glm(formula = 주거지모기지수01 ~ 평균기온 + 일강수량 +
    평균상대습도 + 합계일사량, family = binomial,
    data = c_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3019   0.1084   0.1847   0.3759   1.1338

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.01721    1.31048   3.065  0.00217 **
평균기온       0.15105    0.02439   6.193 5.89e-10 ***
일강수량       0.10555    0.05157   2.047  0.04067 *
평균상대습도  -0.04294    0.01809  -2.374  0.01759 *
합계일사량    -0.02814    0.03821  -0.737  0.46139
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 387.42  on 744  degrees of freedom
Residual deviance: 294.78  on 740  degrees of freedom
AIC: 304.78

Number of Fisher Scoring iterations: 8
```

주거지모기지수 로지스틱 회귀분석 결과

03. 로지스틱 회귀분석

3-2. 로지스틱 회귀분석 진행

4) 로지스틱 회귀분석 결과

```
> summary(glm_1)

Call:
glm(formula = 수변부모기지수01 ~ 평균기온 + 일강수량 +
    평균상대습도 + 합계일사량, family = binomial,
    data = c_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2572   0.1163   0.1918   0.3693   1.0965

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.46553    1.29625   2.673  0.00751 **
평균기온       0.14081    0.02409   5.845 5.06e-09 ***
일강수량       0.09559    0.05070   1.886  0.05936 .
평균상대습도  -0.03288    0.01796  -1.831  0.06709 .
합계일사량    -0.02724    0.03799  -0.717  0.47340

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 377.15  on 744  degrees of freedom
Residual deviance: 293.03  on 740  degrees of freedom
AIC: 303.03

Number of Fisher Scoring iterations: 8
```

- 목적변수가 1이 될 확률을 높이는 요인: '평균기온', '일강수량'
- 목적변수가 0이 될 확률을 높이는 요인: '평균상대습도', '합계일사량'

수변부모기지수 로지스틱 회귀분석 결과

03. 로지스틱 회귀분석

3-3. 지수변환 값을 산출

지수변환 값을 산출

```
92 exp(glm_1$coefficients)
93 exp(glm_2$coefficients)
```

```
> # 지수변환 값 산출
> exp(glm_1$coefficients)
(Intercept)    평균기온    일강수량    평균상대습도    합계일수량
31.9933963    1.1512059    1.1003058    0.9676561    0.9731320
> exp(glm_2$coefficients)
(Intercept)    평균기온    일강수량    평균상대습도    합계일수량
55.5458834    1.1630501    1.1113212    0.9579681    0.9722515
```

- 전체 설명 변수값이 0일 때, 모기 번식 비율
 - 수변부: 31.99% / 주거지: 55.54%
- 평균기온, 일강수량이 증가했을 때 모기 번식 비율
 - 수변부: 1.15% (평균기온), 1.16% (일강수량) / 주거지: 1.16% (평균기온), 1.11% (일강수량)

03. 로지스틱 회귀분석

3-4. AIC & 다중공선성 확인

1) AIC가 가장작은 모델을 탐색

```
96 step(glm_1, direction = "both",
97     scope = (~ 평균기온 + 일강수량 + 평균상대습도 + 합계일사량))
98
99 step(glm_2, direction = "both",
100     scope = (~ 평균기온 + 일강수량 + 평균상대습도 + 합계일사량))
```

	Df	Deviance	AIC
- 합계일사량	1	293.53	301.53
<none>		293.03	303.03
- 평균상대습도	1	296.49	304.49
- 일강수량	1	299.98	307.98
- 평균기온	1	336.02	344.02

Step: AIC=301.53
수변부모기지수01 ~ 평균기온 + 일강수량 + 평균상대습도

	Df	Deviance	AIC
<none>		293.53	301.53
- 평균상대습도	1	296.58	302.58
+ 합계일사량	1	293.03	303.03
- 일강수량	1	301.11	307.11
- 평균기온	1	350.54	356.54

<수변부 분석>

	Df	Deviance	AIC
- 합계일사량	1	295.31	303.31
<none>		294.78	304.78
- 평균상대습도	1	300.70	308.70
- 일강수량	1	303.16	311.16
- 평균기온	1	344.33	352.33

Step: AIC=303.31
주거지모기지수01 ~ 평균기온 + 일강수량 + 평균상대습도

	Df	Deviance	AIC
<none>		295.31	303.31
+ 합계일사량	1	294.78	304.78
- 평균상대습도	1	301.07	307.07
- 일강수량	1	304.43	310.43
- 평균기온	1	361.28	367.28

<주거지 분석>

- 합계 일사량을 제외한 모델이 AIC 값이 더 낮음 → 제외한 모델의 품질이 더 좋음.

03. 로지스틱 회귀분석

3-4. AIC & 다중공선성 확인

2) 다중공선성 확인을 통해 독립변수 사이에 상관성이 있는 지 판별함.

```
104 vif(glm_1)
105 vif(glm_2)
```

1. 수변부 분석 모델의 다중공선성 분석

```
> vif(glm_1)
      평균기온      일강수량 평균상대습도      합계일사량
      1.811780      1.266486      2.108442      1.616683
```

2. 주거지 분석 모델의 다중공선성 분석

```
> vif(glm_2)
      평균기온      일강수량 평균상대습도      합계일사량
      1.819445      1.276775      2.167883      1.609258
```

- 수변부와 주거지 모두 $VIF < 10$ 이므로 다중공선성 가능성이 낮음

03. 로지스틱 회귀분석

3-5. 최적의 모델을 이용하여 재분석

로지스틱 회귀분석 재실시

```
108 glm_3 <- glm(수변부모기지수01 ~ 평균기온 + 일강수량 + 평균상대습도,
109               family = binomial, data = c_data)
110 glm_4 <- glm(주거지모기지수01 ~ 평균기온 + 일강수량 + 평균상대습도,
111               family = binomial, data = c_data)
112 summary(glm_3)
113 summary(glm_4)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.83339	0.93074	3.044	0.00233 **
평균기온	0.13208	0.02058	6.416	1.4e-10 ***
일강수량	0.09903	0.05080	1.949	0.05126 .
평균상대습도	-0.02633	0.01530	-1.721	0.08523 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 377.15 on 744 degrees of freedom
Residual deviance: 293.53 on 741 degrees of freedom
AIC: 301.53

Number of Fisher Scoring iterations: 8

수변부모기지수 로지스틱 회귀분석 결과

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.36305	0.94282	3.567	0.000361 ***
평균기온	0.14214	0.02097	6.777	1.23e-11 ***
일강수량	0.10916	0.05164	2.114	0.034521 *
평균상대습도	-0.03616	0.01540	-2.349	0.018829 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 387.42 on 744 degrees of freedom
Residual deviance: 295.31 on 741 degrees of freedom
AIC: 303.31

Number of Fisher Scoring iterations: 8

주거지모기지수 로지스틱 회귀분석 결과

03. 로지스틱 회귀분석

3-6. 지수변환 값 산출 및 H-L 적합도 검정

1) 지수변환 값을 산출

```
116 exp(glm_3$coefficients)
117 exp(glm_4$coefficients)
```

```
> exp(glm_3$coefficients)
(Intercept)    평균기온    일강수량    평균상대습도
17.0029230    1.1411983    1.1041017    0.9740133
> exp(glm_4$coefficients)
(Intercept)    평균기온    일강수량    평균상대습도
28.8772000    1.1527353    1.1153459    0.9644817
```

- 전체 설명 변수값이 0일 때, 모기 번식 비율
 - 수변부: 17.002% / 주거지: 28.87%
- 평균기온, 일강수량이 증가했을 때 모기 번식 비율
 - 수변부: 1.14% (평균기온), 1.10% (일강수량) / 주거지: 1.15% (평균기온), 1.11% (일강수량)

03. 로지스틱 회귀분석

3-6. 지수변환 값 산출 및 H-L 적합도 검정

2) H-L 적합도 검증

```
118 hoslem.test(x = glm_1$y , y = fitted(glm_3))
119 hoslem.test(x = glm_2$y , y = fitted(glm_4))
```

```
> # H-L 적합도 검정
> hoslem.test(x = glm_1$y , y = fitted(glm_3))

      Hosmer and Lemeshow goodness of fit (GOF) test

data:  glm_1$y, fitted(glm_3)
X-squared = 18.108, df = 8, p-value = 0.02043

> hoslem.test(x = glm_2$y , y = fitted(glm_4))

      Hosmer and Lemeshow goodness of fit (GOF) test

data:  glm_2$y, fitted(glm_4)
X-squared = 20.433, df = 8, p-value = 0.008817
```

- 수변부와 주거지 분석 결과 둘 다 P-value < 0.05로 모델이 적합하지 않음

3. 분석 결과

- ① 다중회귀분석 결과 요약
- ② 로지스틱 회귀분석 결과 요약

01. 분석목표

" 다중회귀분석 & 로지스틱 회귀분석을 이용한 **기상 환경에 따른 모기지수 분석** "

다중회귀분석

" 여러 기상 환경 변수들로부터 수변부와 주거지의 모기지수를 예측한다. "

로지스틱 회귀분석

" 여러 기상 환경 변수들을 통해
수변부와 주거지의 모기 번식 비율이 증가할 확률을 예측한다"

01. 다중회귀분석 결과 보고서

1-1. 다중회귀분석 결과

	추정값	표준오차
절편	118.7604	53.8206
평균기온	11.4326	0.9271
평균상대습도	-1.6001	0.7231
합계일사량	-3.6830	1.4419

수변부모기지수 분석 결과

	추정값	표준오차
절편	153.9067	57.7290
평균기온	10.6844	0.9945
평균상대습도	-2.1875	0.7756
합계일사량	-4.4802	1.5466

주거지모기지수 분석 결과

1-2. 해석

- 수변부와 주거지의 모기지수 데이터를 목적 변수로 설정하고, 기후데이터를 설명 변수로 설정하여 분석 진행.
- VIF를 이용하여 다중공선성의 가능성을 검토한 결과, 평균상대습도를 제외한 설명 변수가 2 미만으로 다중공선성 가능성 낮음.
 - 평균상대습도 또한 $2 < VIF < 5$ 로 주의가 필요한 정도로 가능성이 낮음.

02. 로지스틱 회귀분석 결과 보고서

2-1. 로지스틱 회귀분석 결과

	추정값	표준오차
절편	2.83339	0.93074
평균기온	0.13208	0.02058
일강수량	0.09903	0.05080
합계일사량	-0.02633	0.01530

수변부모기지수 분석 결과

	추정값	표준오차
절편	3.36305	0.94282
평균기온	0.14214	0.02097
일강수량	0.10916	0.05164
합계일사량	-0.03616	0.01540

주거지모기지수 분석 결과

2-2. 해석

- 수변부와 주거지의 모기지수 데이터를 목적 변수로 설정하고, 기후데이터를 설명 변수로 설정하여 로지스틱 회귀분석 진행.
- VIF를 이용하여 다중공선성의 가능성을 검토한 결과, 모든 설명 변수가 2 미만으로 다중공선성 가능성 낮음.
- H-L 적합도 검정 결과, 적합성이 P-value < 0.05로 모델이 적합하지 않음.

감사합니다

유경빈(팀장)	20210424 6	ICT융합공학부
정지용	201804108	소프트웨어응용학부
문지혜	202004042	소프트웨어응용학부