기상청

[STEP4 분석모형 구축]

일반화 선형 모형(GLM) 활용 분석

기상청

기상기후
빅데이터 분석 플랫폼

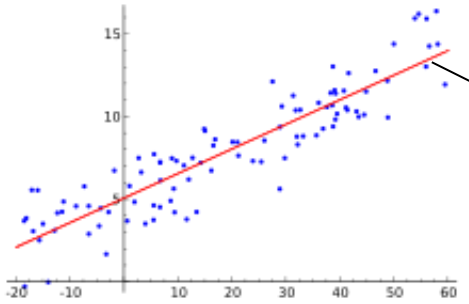
[분석교육] 일반화 선형 모형(GLM)

1. 일반화 선형 모형(GLM) 분석 수행 함수

[분석 교육] 일반화 선형 모형(GLM) (1/2)

일반화 선형 모형(General Linear Model : GLM)은 선형 모형 상에서 하나 이상의 변수를 대상으로 일반화된 모형을 구축하는 알고리즘이다.

● 일반화 선형 모형(GLM) 개념



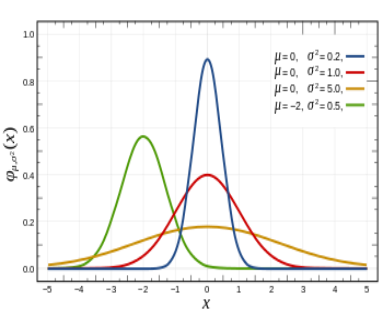
$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}$$

〈Linear Regression Model〉

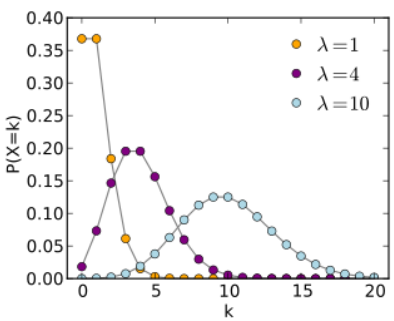
- 전통적인 선형모형이 갖는 종속변수의 정규분포와 분산의 동등성 가정을 배제하고 자료의 독립성 가정과 모형의 가법성(additivity)원리에 기초한 통계모형이다.

● 선형 회귀 분석

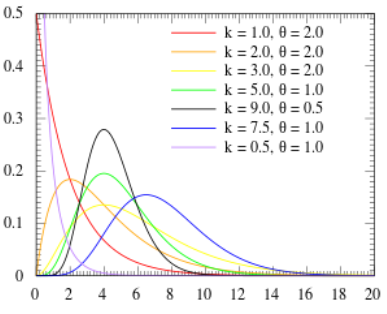
- 일반화 선형 모형은 종속변수의 분포에 따라 다음과 같은 선형회귀 기법을 사용할 수 있다.
  - Gaussian Regression : 종속변수가 정규분포인 경우
  - Poisson Regression : 종속변수가 포아송분포인 경우
  - Binomial Regression : 종속변수가 이항분포인 경우(0 or 1)
  - Gamma Regression : 종속변수가 감마분포인 경우



〈Gaussian Distribution〉



〈Poisson Distribution〉



〈Gamma Distribution〉

[분석 교육] 일반화 선형 모형(GLM) (2/2)

● 로지스틱 회귀 분석

- 로지스틱 회귀는 종속변수가 이항변수(0 or 1) 인 경우에 사용하는 선형 모형이다.
- 로지스틱 회귀는 선형 회귀와 유사하지만 차이점을 지니고 있는데, 첫번째 차이점은 이항형인 데이터에 적용하였을 때 종속변수의 결과가 범위[0,1]로 제한된다는 것이고, 두번째 차이점은 종속변수가 이진적이기 때문에 조건부 확률의 분포가 정규분포가 아닌 이항분포를 따른다는 점이다.
- 로지스틱 회귀 모형 식은 범위가 [-∞,∞]인 독립변수의 값과 상관없이 종속변수의 값이 항상 범위 [0,1] 사이에 있도록 하기 위해 오즈비(odds ratio)를 로짓(logit)변환을 수행한다.

$$ODDs = \frac{p}{(1-p)} \qquad logit(p) = log\frac{p}{(1-p)}$$

〈ODDs : 임의의 사건이 발생할 확률〉 〈오즈비에 로그를 취해 범위를 [0,1]로 조정〉

- 데이터 X의 분류가 Y일 확률을 p, N일 확률을 1-p라 할 때의 로지스틱 함수는 다음과 같다.

$$\log\left(\frac{p}{(1-p)}\right) = \beta_0 + \beta_1 X$$

- 로지스틱 회귀는 다항형 로지스틱 회귀(Multinomial Logistic Regression), 순서형 로지스틱 회귀(Ordered Logistic Regression), 혼합형 로지스틱 회귀(Mized Logit) 등이 있다.

● R에서의 GLM

- R에서는 glm() 함수를 사용하여 쉽게 일반화 선형 모형 분석을 수행할 수 있다. glm() 함수의 family 옵션 설정을 통해 분포에 따른 선형 회귀와 로지스틱 회귀 분석이 가능하다.

▪ **Usage**

glm(formula, **family = family**, data)

▪ **family 설정값**

- 로지스틱 회귀 : binomial, quasibinomial
- 분포에 따른 선형 회귀 : Gaussian, Gamma, inverse.Gaussian, poisson, quasi, quasipoisson

참고 자료 : 위키백과

1. 일반화 선형 모형(GLM) 분석 수행 함수

- **h2o::h2o.glm**

  - 일반선형모델을 구축

▪ **Usage**

- training_frame : 학습 데이터 셋
- validation_frame : 검증 데이터 셋, 기본값은 NULL
- x : 독립변수
- y : 종속변수
- family : 수행할 일반화 선형 모델의 분포 종류
- link : 링크 함수 설정, 기본값은 family에 종속된 링크 함수 적용
- lambda_search : 모형의 람다 값을 산출
- lambda : 람다 값 설정

▪ **Examples**

```
Created_Model <- h2o.glm(
                    y=1, x=2:ncol(Train),
                    training_frame = Train,
                    family = "gaussian",
                    link = "family_default",
                    lambda_search = TRUE,
                    fold_assignment = "AUTO")
Created_Model
```

```
> Created_Model <- h2o.glm(
+   y=1, x=2:ncol(Train),
+   training_frame = Train,
+   family = "gaussian",
+   link = "family_default",
+   lambda_search = TRUE,
+   fold_assignment = "AUTO")
|=========================================================================| 100%
```

본 문서의 내용은 기상청의 날씨마루(http://big.kma.go.kr) 내
R 프로그래밍 교육 자료입니다.