일반화 선형모형(Generalized Linear Model)

회귀분석이나 분산분석은 종속변수가 정규분포되어 있는 연속형 변수이다. 하지만 많은 경우에 있어서 종속변수가 정규분포되어 있다는 가정을 할 수 없는 경우도 있으며 범주형 변수가 종속변수인 경우도 있다. 다음과 같은 경우에 일반화 선형모형을 사용한다.

1. 종속변수가 범주형변수인 경우: 이항변수(0 또는 1, 합격/불합격, 사망.생존 등)인 경우도 있으며 다항변수(예를 들어 poor/good/excellent 또는 공화당/민주당/무소속 등)인 경우 정규분포 하지 않는다.
2. 종속변수가 count(예를 들면 한 주간 교통사고 발생 건수, 하루에 마시는 물이 몇잔인지 등)인 경우. 이들 값은 매우 제한적이며 음수가 되지 않고 평균과 분산이 밀접하게 관련되어 있고 정규분포하지 않는다.

일반화 선형 모형은 종속변수가 정규분포하지 않는 경우를 포함하는 선형모형의 확장이며 glm()함수를 사용한다. 이 장에서는 대표적으로 로지스틱회귀(Logistic regression)와 포아송회귀(Poisson regression)를 다룬다.

glm() 함수

일반화선형모형은 glm()함수를 사용한다. glm() 함수의 사용방법은 lm()함수와 유사하나 추가로 **family**라는 인수를 지정해준다. family에 따라 연결된 함수가 달라지는데 사용법은 다음과 같다.

> glm(formula, **family=family(link=function)**, data)

family는 종속변수의 분포에 따라 다음과 같은 것들을 사용할 수 있다. 종속변수의 분포가 정규분포인 경우 gaussian, 이항분포인 경우 binomial, 포아송분포인 경우 poisson, 역정규분포인 경우 inverse.gaussian, 감마분포인 경우 gamma, 그리고 응답분포가 확실하지 않은 때를 위한 유사가능도 모형인 경우 quasi를 사용할 수 있다. glm()함수의 결과를 anova()와 조합하면 분산분석표를 생성할 수 있고 summary()에 넣어서 잔차와 추정값 등을 얻을 수 있다. coef() 함수를 사용하여 모형 인수들의 절편과 기울기 등을 얻을수 있으며 residual()함수는 잔차를 얻을 수 있다. plot()함수를 사용하여 회귀진단 plot을 얻을 수 있고 회귀모형을 사용하여 predict() 함수로 새로운 데이타에 대한 예측치를 추정할 수 있다.

오차항의 확률분포가 정규분포가 아닌 경우

Generalized Linear Model (GLM)

일반화 선형 회귀 모델은 종속 변수에 적절한 함수를 적용하는 회귀 모델링 기법입니다.

$$g(\hat{y}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i \ \ (g: link\ function)$$

이렇게 종속변수에 적용하는 함수를 link function 이라고 부르는데 오차항의 확률 분포가 무엇이냐에 따라 일반적으로 사용하는 link function 이 정해져 있습니다. 대표적인 것 몇 가지만 소개하면 아래와 같습니다.

오차항의 확률 분포	사용하는 Link function
binomial	Logit function
exponential	Inverse function
Poisson	Log function

보통 GLM은 종속 변수의 특성에 따라 세부적인 명칭을 구분하기도 합니다.

- 종속 변수가 0 아니면 1인 경우: Logistic regression
- 종속 변수가 순위나 선호도와 같이 순서만 있는 데이터인 경우: Ordinal regression
- 종속 변수가 개수(count)를 나타내는 경우: Poisson regression