6.2 일반선형모델의 기본

있다/ 없다 (***뭘까요?***)   /   1개,2개,3개 등 0 이상의 정수만 취하는 (***뭘까요?***)

를 따르는 데이터가 있다면 모집단분포가 정규분포라고 가정하기에는 무리가 있다. 이럴때 쓰는게 일반선형모델(GLM, General Linear Models)이다.

<u>장점</u>

- 분류 문제와 회귀 문제를 통일성 있게 취급가능

6.2.1 일반선형모델의 구성요소

1. 모집단이 따르는 확률분포
2. 선형예측자
3. 링크함수

구성요소가 많다? 데이터에 따라 **유연하게** 변화시킬 수 있다! 적용할 수 있는 **데이터가 많아**진다.

6.2.2 확률분포

일반선형모델은 정규분포나 이항분포, 푸아송 분포 등에 적용가능.

6.2.3 선형예측자

선형예측자란 독립변수를 선형의 관계식으로 표현한 것입니다.

예) **맥주 매상**이라는 **종속변수**를 **기온**이라는 **독립변수**에서 예측하는 경우는 아래와 같습니다.

`B0+B1x기온(℃)`

예) **시험합격**이라는 **종속변수**를 **공부시간**이라는 **독립변수**에서 예측하는 경우는 아래와 같습니다.

`B0+B1x공부시간`

- 종속 변수가 연속형일 때
  - 차이를 보고자 할 때 : T 검정(T-test), 분산분석(Anova)
  - 관계를 보고자 할 때 : 회귀분석(Regression)

- 종속 변수가 이산형일 때
  - 연관을 보고자 할 때 : 카이제곱 독립성 검정(Chi square Independent Test)
  - 관계를 보고자 할 때 : 로지스틱 회귀분석(Logisitc Regression)

6.2.4 링크함수

링크함수는 *종속변수*와 *선형예측자*를 서로 대응시키기 위해 사용한다.

예) `맥주 판매 개수= B0+B1x기온(℃)` 라고 예측하면, **마이너스** 가 될 가능성이 있음. 그래서 링크함수로 **로그함수를 쓴다.**

그러면, `log[맥주 판매 개수] = B0+B1x기온(℃)`

exp를 취해서, `맥주 판매 개수 = exp[B0+B1x기온(℃)]`

자 이러면 마이너스는 안되겠죠?

이렇게 종속변수에 링크함수를 적용함으로써 0 이상의 카운트 데이터나 [0,1] 범위를 취하는 성공확률 등을 대상으로 예측 할 수 있다

6.2.5 링크함수와 확률분포의 대응

확률분포	링크함수	모델명
정규분포	항등함수	정규선형모델
이항분포	로짓함수	로지스틱 회귀
푸아송 분포	로그함수	푸아송 회귀

항등함수란 f(x)=x가 되는 함수입니다.

즉, 아무런 변환도 하지 않는 함수, 정규선형모델에서는 변환이 x, 그래서 일반선형모델의 틀 안에서 항등함수라고 부른다.

로짓함수란

(뒤에 나온데요)

로그함수란정규분포에서 종속변수가 마이너스 값이 되지 않게 한 모델. 음이항분포에서도 링크함수로 로그함수가 자주 사용된다. 감마분포에 대해서는 역수**도** 사용하기도 함.

6.2.6 일반선형모델의 파라미터 추정

GLM에서는 정규분포 이외의 확률분포가 사용되는 경우도 있기 때문에 **최대우도법에 의한 파라미터 추정**을 한다. **최소제곱법**이 이용되는 경우가 많다.

6.2.7 일반선형모델을 이용한 검정 방법

GLM에 보통 세 가지 검정 방법이 있다.

t검정이 안되서, Wald검정을 쓴다.(statmodels 출력에서도 볼수 o)

Wald는

- 샘플 사이즈가 클때
- 추정값이 정규분포를 따르는 것을 이용

이책은 AIC를 쓴대요

Akaike's Information Criterion

Likelihood Max

- **AIC는 "penalized likelihood":** 모델의 파라미터 (β) 추정 방식이 MLE일 경우, 최대화된 Log-Likelihood에 변수 추가에 따른 패널티 항을 추가한 것이 AIC.

$$AIC = -2\ LogLikelihood + 2p$$

$$BIC = -2\ LogLikelihood + \log(n)p$$

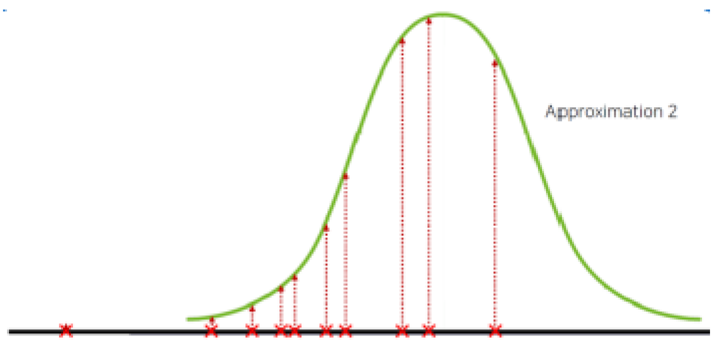
- 선형회귀에서의 MLE:

OLS

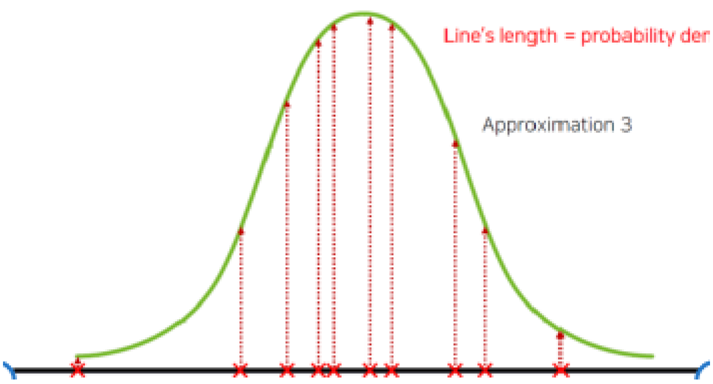
$$L(\beta, \sigma^2|X, Y) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n exp(-\frac{1}{2}\sum(\frac{Y_i - X_i^T\beta}{\sigma})^2)$$

$$\log L(\beta, \sigma^2|X, Y) = -\frac{n}{2}log(2\pi) - nlog\sigma - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

- MLE Estimate는 베타는 OLS와 같으며 $\hat{\sigma}^2_{MLE} = \frac{RSS}{n}$ ($\hat{\sigma}^2_{OLS} = \frac{RSS}{n-p}$)



그리고 세번째 그래프에서는 중심점에 가깝게 분포를 많이했죠? 그러니깐, 즉 이전 2개의 그래프보다 데이터가 몰 아졌죠?



우도비 검정은 모델의 적합도를 비교하는 방법. Type II ANOVA와 같은 해석이 가능한 계산법도 제안되고 있습니다.

스코어 검정이라는 방법도 있는데 잘 사용x