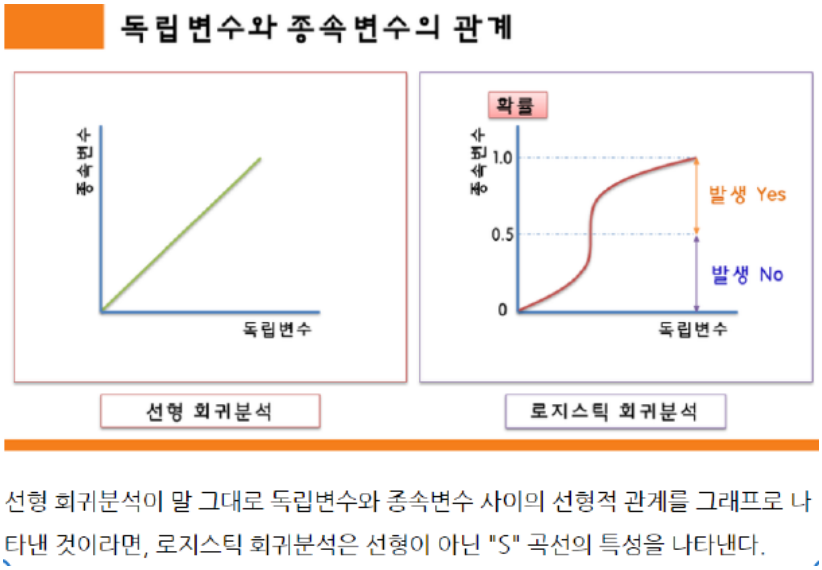
6.3 로지스틱 회귀



선형 회귀분석이 말 그대로 독립변수와 종속변수 사이의 선형적 관계를 그래프로 나타낸 것이라면, 로지스틱 회귀분석은 선형이 아닌 "S" 곡선의 특성을 나타낸다.

로지스틱 회귀는

- 확률분포에 이항분포를 사용
- 링크함수에 로짓함수를 사용한 일반선형모델이다.
- 독립변수는 여러 개 있어도 상관x
- 연속형과 카테고리형이 섞여 있어도 상관x

6.3.1 이 절의 예제

`시험에 합/불 을 예측한다.`

- 선형예측자는 `B0+B1x공부시간`

6.3.2 두 값의 판별 문제

- **종속변수**는 합(1), 불(0)을 취하는 **이항**확률변수.
- 공부시간은 **연속형** 변수이므로 예측값인 시험 합격 여부가 소수점 이하의 값이 되기도 한다.

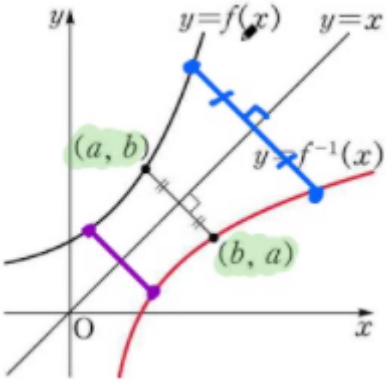
`시험의 합불(합1,불0) = B0+B1x공부시간` ←이렇게 하면 안되겠죠? 왜냐면 합불이 연속형(심지어 마이너스도 가능)으로 나오니까!!

6.3.3 로짓함수

로짓함수는 아래와 같은 함수를 가리킵니다. 로그의 밑은 e입니다.

f(x)=log(x/(1-x))

6.3.4 역함수



6.3.5 로지스틱 함수

로지스틱함수는 로짓함수의 역함수이다.

- 로짓함수는 f(x)
- 로지스틱 함수를 g(x)라 하면, g(f(x))=x 가 된다.
- g(y)-1/(1+exp(-y))

6.3.6 로지스틱 함수의 특징

지수함수인 exp(-y) 는 음수가 안된다. 로지스틱함수의

- 분모가 1 이하로 내려가지 않는다.
- y가 작아질 수록 큰 값이 된다.
- 분모가 커지면 출력이 점점0이 된다.

`그래서 로지스틱함수의 출력은 0 미만이거나 1을 초과하지 않는다.`

6.3.7 로지스틱 회귀의 구조

로지스틱 회귀는 확률분포에 **이항분포를** 사용하고 링크함수에 **로짓함수를** 사용한 **일반선형모델입니다**.

성공확률 p, 링크함수에 로짓함수를 사용하면 시험의 ***합격률과 공부시간의 관계***를 아래와 같이 나타낼 수 있다.

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

양변에 로지스틱 함수를 적용하면 아래와 같이 변형할 수 있다.

$$p = \frac{1}{1+\exp[-(\beta_0 + \beta_1 \times \text{공부시간})]}$$

합불 데이터를 이용해서 ***공부시간이 시험의 합격률에 영향을 주는지*** 조사한다. 공부시간이 5시간이던 학생이 10명 있다고 하자, 이때 **합격자수 M**은 **성공확률이 p**에서 **시행횟수가 10**인 이항분포를 따른다고 상정한다.

$$\text{합격자수} : M \sim \text{Bin}\left(m \mid 10, \frac{1}{1+\exp[-(\beta_0+\beta_1\times 5)]}\right)$$

과 같이 확률분포를 따르는 표본을 얻었다고 생각하는 것이 **로지스틱 회귀**이다.

그러면 이항분포의 확률질량함수는 다음과 같다.

$$\text{Bin}(m \mid N, p) = {}_N C_m \cdot p^m \cdot (1-p)^{N-m}$$

6.3.8 로지스틱 회귀의 우도함수

앞 절에서는 계수 B0, B1과 공부시간을 알고 있을 때 ***시험의 합격률과 합격자수의 분포를 추측하는 방법***을 배웠다. 이제는 이 **계수 추정**을 배우겠다. GLM은 **최대우도법으로 파라미터를 추정**한다.

예)

- 공부시간이3시간인 학생9명 중 4명이 합격했습니다.
- 공부시간이5시간인 학생8명 중 6명이 합격했습니다.
- 공부시간이8시간인 학생1명 중 1명이 합격했습니다.

이때의 우도함수는

$$\begin{aligned}\mathcal{L}(\beta_0,\beta_1;N,m) = &\text{Bin}\left(4 \mid 9, \frac{1}{1+\exp[-(\beta_0+\beta_1\times 3)]}\right)\\ &\times \text{Bin}\left(6 \mid 8, \frac{1}{1+\exp[-(\beta_0+\beta_1\times 5)]}\right)\\ &\times \text{Bin}\left(1 \mid 1, \frac{1}{1+\exp[-(\beta_0+\beta_1\times 8)]}\right)\end{aligned}$$

이다.

시험자수가 늘어나면 수식이 복잡해지지만 구조는 안 변함.

(주피터로!)

6.3.16 용어 설명

오즈(odds)는 실패하는 것보다 ***성공하는 것이 몇 배 더 쉬운가***를 나타낸다.

`오즈=p/(1-p)` 여기에 로그를 취한게 **로그오즈**. **로짓함수는** 성공확률을 로그오즈로 변환하는 함수로도 볼 수 있다.

오즈비(odd rate)는 오즈 간에 비율을 취한 것, 오즈비에 로그를 취한 것을 **로그오즈비** 라고 부른다.

로지스틱 회귀의 계수와 오즈는 밀접한 관계, **회귀계수는 독립변수를 1단위 변화시켰을 때의 로그오즈비**라고 해석할 수 있음.

(주피터로!)