Regularization & GLM

분석 D조 곽현석, 김태희, 민선우, 전윤희, 조보금, 최혜린

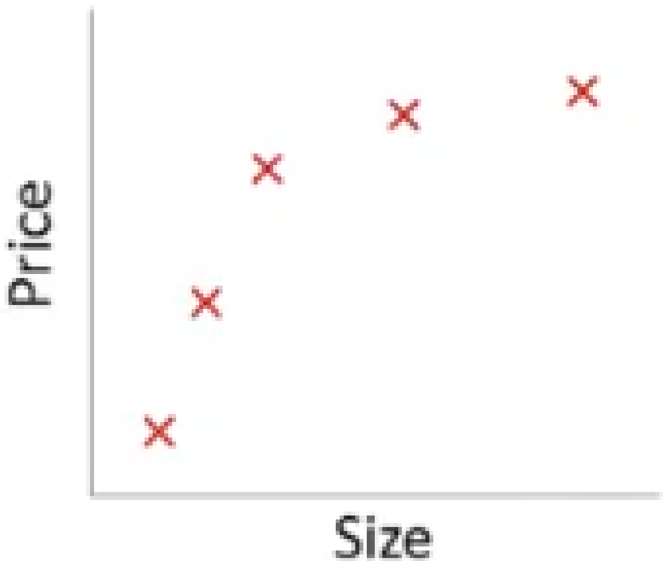
Regularization

문제 제기

Regularization

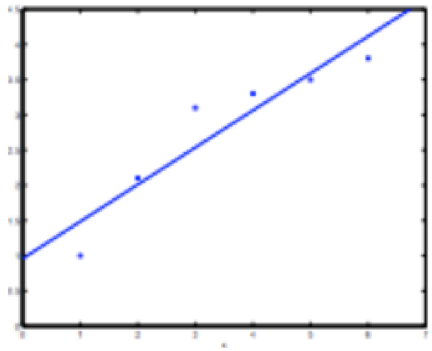
회귀식을 통해 집의 크기에 따른

집값을 어떻게 예측할 수 있을까?



〈Housing Prices〉

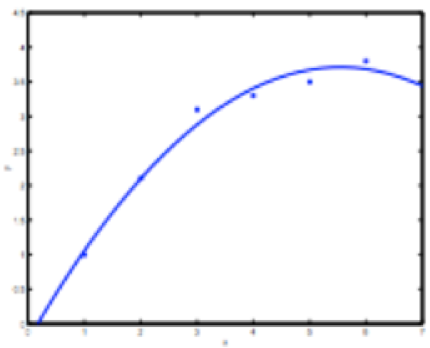
1. 문제 제기 Regularization



(degree = 1) underfitting high bias

$$y = \theta_0 + \theta_1 x$$

→ 데이터 내 모든 정보를 고려할 수 없게됨

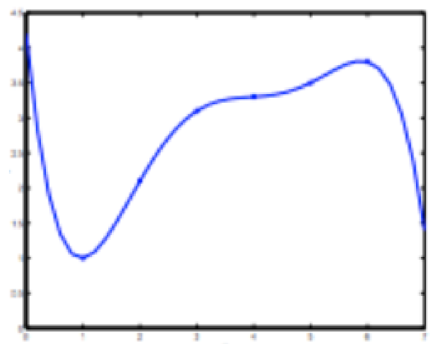


(degree = 2) just right

$$y = \theta_0 + \theta_1 x + \theta_1 x^2$$

→ bias와 variance가 잘 고려된 모델

그러나 실제 상황에서 두 가지를 동시에 만족하는 것은 거의 불가능하다

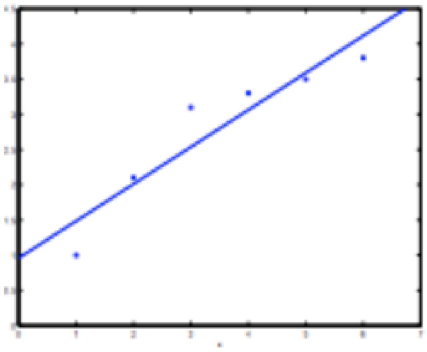


(degree = 4) overfitting high variance

$$y = \theta_0 + \theta_1 x + \theta_1 x^2 + \theta_1 x^3 + \theta_1 x^4$$

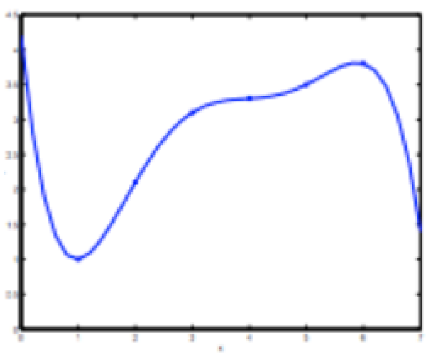
→ 새 데이터가 들어오게 된다면 모델이 완전히 변해, 일반성을 잃음

1. 문제 제기 Regularization



underfitting 해결방안

- 모형에 새로운 변수를 추가
- model의 복잡도는 증가하고 high bias를 줄이는 방법



overfitting 해결방안

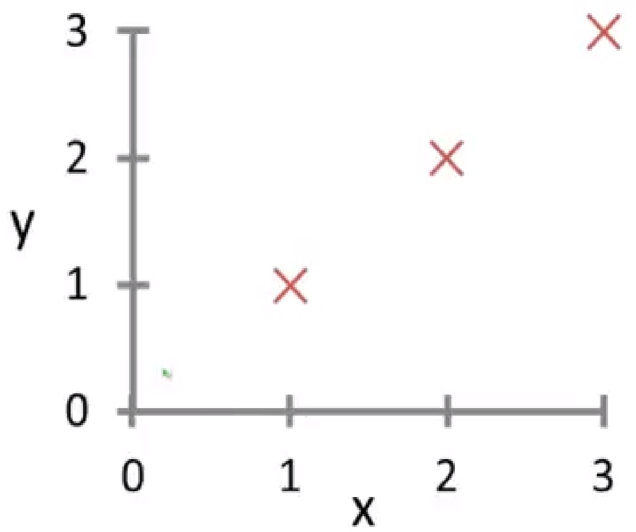
- 정규화
- 모형의 복잡도 줄이기

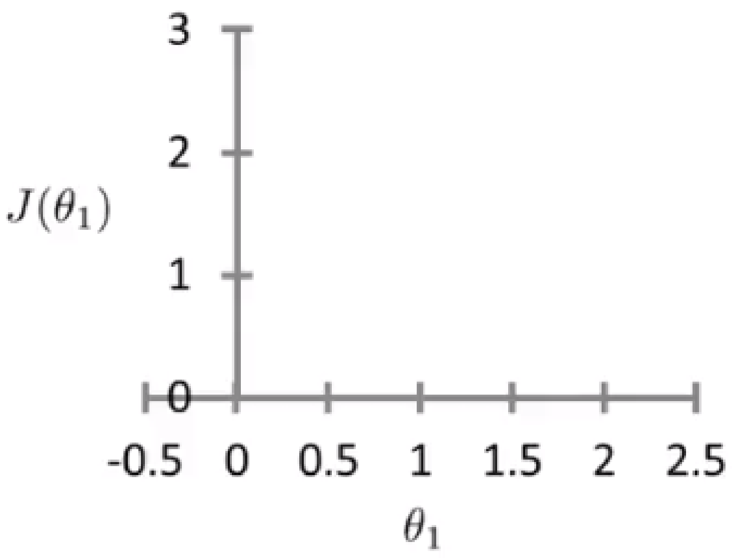
2. Cost function(비용 함수)
Gradient Descent(경사 하강법)

Regularization

$$y = \theta_1 x$$

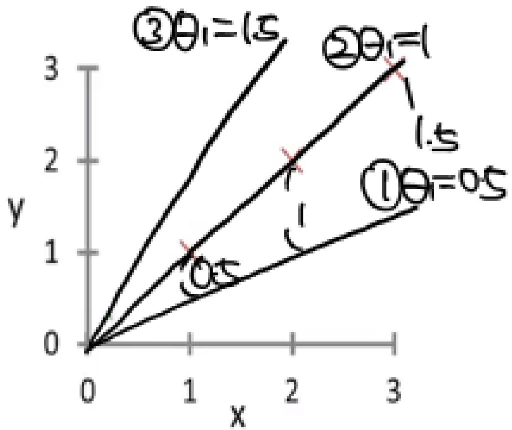
$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

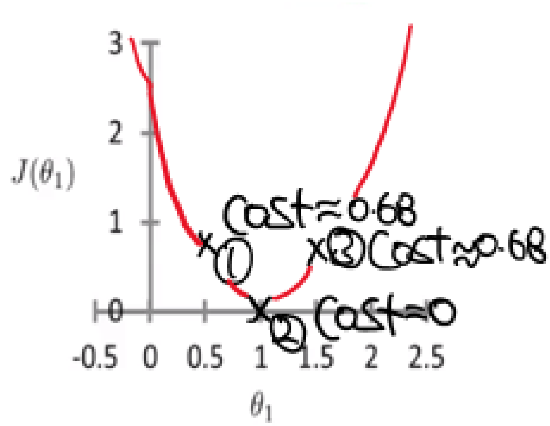




2. Cost function(비용 함수)
Gradient Descent(경사 하강법)

Regularization





모델의 목표 : 비용함수를 최소로 하는 것

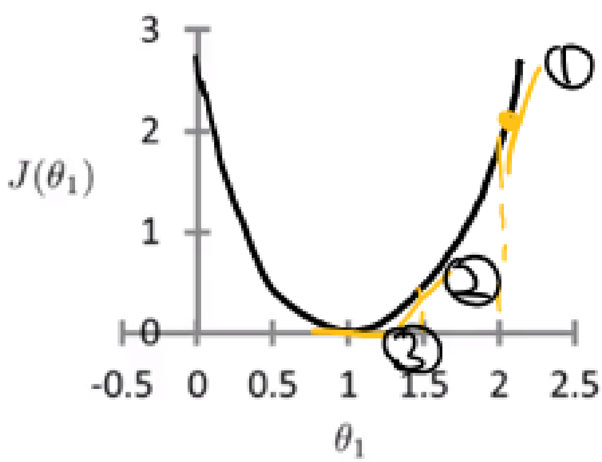
이 데이터의 경우,
계수가 1과 멀어진다면, 비용이 증가

2. Cost function (비용 함수)
Gradient Descent (경사 하강법)

Regularization

[경사 하강법]

$$J(\theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$



1. 기울기 : 1 〉2 〉3

2. Gradient Descent
   → 미분을 활용하여 계수를 갱신

$$\theta_1 := \theta_1 - \alpha\frac{\partial}{\partial\theta_1}J(\theta_1)$$

(α 는 가중치를 의미)

3. Regularization

Regularization

회귀 계수가 가질 수 있는 값의 범위를 제한하는
Regularized Regression 기법들 중에서 가장 많이 사용

→ Ridge, LASSO, Elastic Net 회귀분석

Ridge

제약조건에 가중치들의 제곱합(squared sum of weights)을 최소화하는 것을 추가적으로 설정

$$\text{cost} = \sum e_i^2 + \lambda \sum w_i^2$$

Lasso

제약조건에 가중치의 절대값의 합을 최소화하는 것을 추가적으로 설정

$$\text{cost} = \sum e_i^2 + \lambda \sum |w_i|$$

Elastic Net

제약조건에 가중치의 절대값의 합과 제곱합을 동시에 제약 조건으로 둠

$$\text{cost} = \sum e_i^2 + \lambda_1 \sum |w_i| + \lambda_2 \sum w_i^2$$

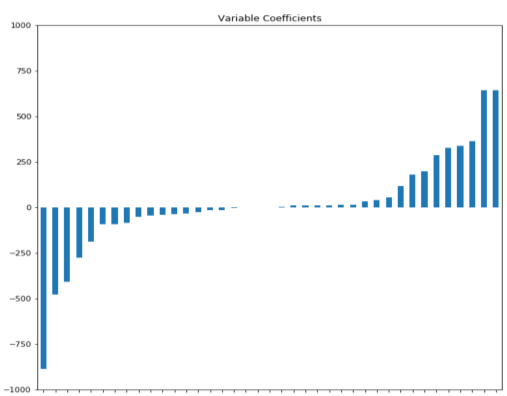
3. Regularization

Regularization

Ridge $\text{cost} = \sum e_i^2 + \lambda \sum w_i^2$

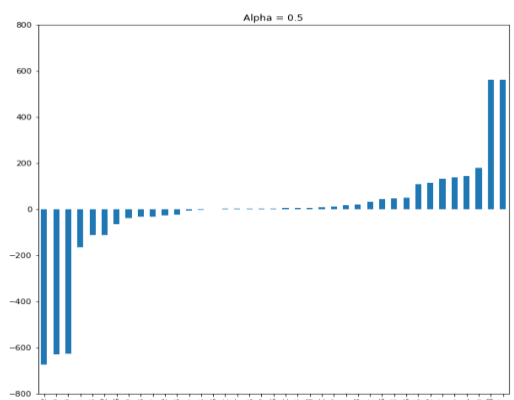
가중치 값이 커질수록 계수의 크기는 감소

가중치 값을 고려해 가장 낮은 오류를 산출하도록 접근해야 함

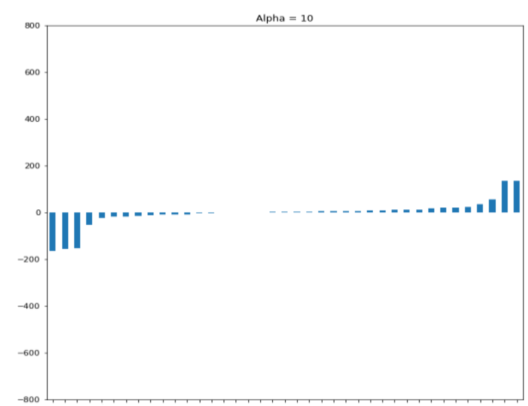


가중치(lambda)=0

*선형회귀와 동일



가중치(lambda)=0.5



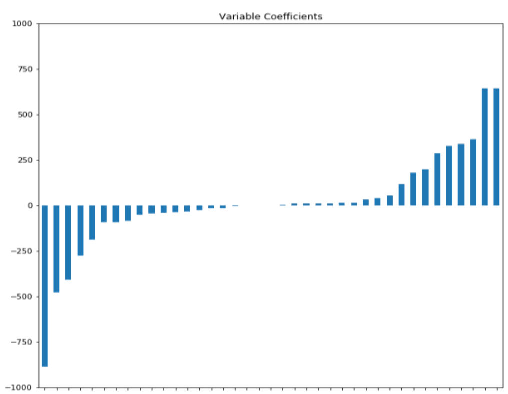
가중치(lambda)=10

3. Regularization Regularization

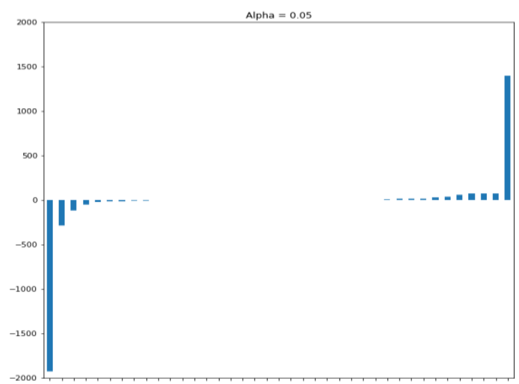
Lasso $\text{cost} = \sum e_i^2 + \lambda \sum |w_i|$

가중치 값이 커질수록 계수의 크기는 감소

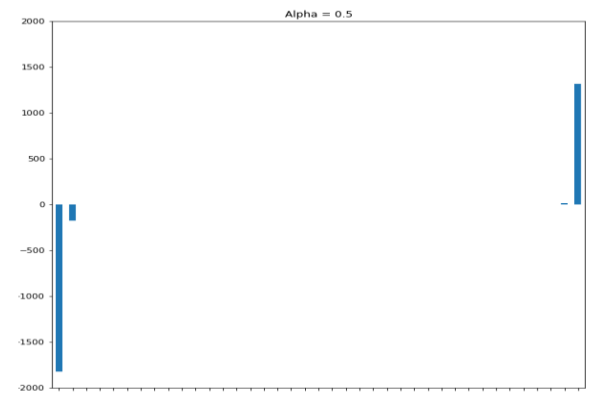
가중치 값을 고려해 가장 낮을 오류를 산출하도록 접근해야함



가중치(lambda)=0



가중치(lambda)=0.05



가중치(lambda)=0.5

3. Regularization

[RIDGE vs LASSO]

공통점 : 제약식을 통해 회귀계수가 가질 수 있는 값의 범위를 제한
→ 다중공선성 방지
→ 모델의 복잡도 감소

RIDGE
가중치가 커져도 회귀계수들이 0이 되지 않는다.
→ 계수를 줄여도 모델은 여전히 복잡한 상태

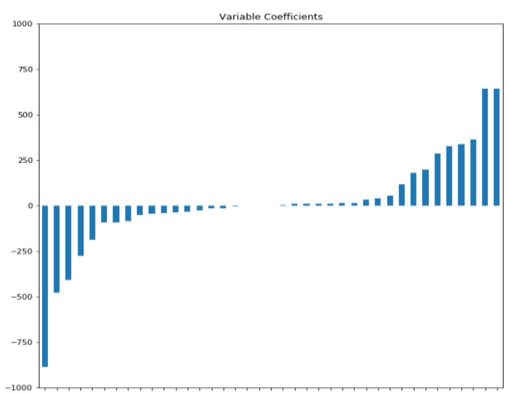
LASSO
필요에 따라 회귀계수의 일부를 0으로 만들어준다
(=가중치가 ∞이면 완벽한 평균모형이 된다, 변수선택의 기능이 있다)
→ 변수들끼리 상관도가 높다면, 한개의 변수만 채택하고,
다른 변수들의 계수는 0으로 바뀔 것이다.

3. Regularization

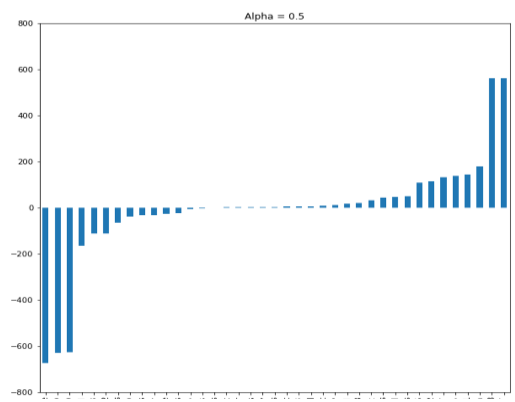
Elastic net

alpha : 제곱합에 대한 가중치(ridge)

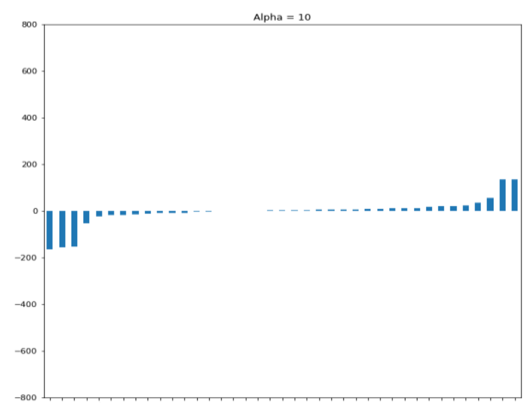
l1_ratio : 절대값의 합에 대한 가중치(lasso)



alpha=0, l1_ratio=0



alpha=1, l1_ratio=0.5



alpha=1, l1_ratio=1

[Generalized Linear Model]

선형회귀의 기본 가정 Generalized Linear Model

가정 1) 회귀모형은 모수에 대해 선형(linear)인 모형 : 직선 형태의 모형 Y=ax+b

가정 2) 수집된 데이터는 정규분포를 따른다

가정 3) 독립변수간에는 상관관계가 없다

가정 4) 오차항은 다음과 같은 가정을 한다.

- 정규성:오차의 평균이 0
- 등분산성:오차항은 분산이 σ^2인 정규분포
- 독립성:오차항은 서로 독립적으로 존재

선형회귀의 기본 가정 Generalized Linear Model

종속변수가 정규분포 되어 있는 연속형 변수일 경우
=> 선형모형 사용

BUT
1) y가 범주형
2) y가 count data
=> 일반화 선형 모형(GLM) 사용

1. GLM의 구성요소 Generalized Linear Model

[Generalized Linear Model]

종속변수가 어떤 연결함수를 통해 요인들이 선형적으로 관련되도록 일반선형모형을 확장한 것

대표적으로 Logistic Regression과 Poisson Regression이 있음

1. GLM의 구성요소 Generalized Linear Model

1) 임의 요소(Random component) : Y를 정의+Y의 확률분포를 가정
2) 시스템 요소(Systematic component): 설명변수 수식의 형태를 정의
3) 연결함수(Link function): 임의 요소의 기댓값과 시스템 요소를 연결하는 함수

Identity Link	$\varphi(\mu) = \mu = \mathbb{X}\beta$	전통적 회귀모형
Log Link	$\varphi(\mu) = \log(\mu) = \mathbb{X}\beta$	Poisson 회귀
Logit Link	$\varphi(\mu) = log\frac{\mu}{1-\mu} = \mathbb{X}\beta \quad \Rightarrow \mu = \frac{e^{\mathbb{X}\beta}}{1+e^{\mathbb{X}\beta}}$	Logistic 회귀
Probit Link	$\varphi(\mu) = \Phi^{-1}(\mu) = \mathbb{X}\beta,\ \Phi^{-1} \sim N(0,1)$	
Complementary	$\varphi(\mu) = log(-log(1-\mu)) = \mathbb{X}\beta \Rightarrow \mu = 1 - exp(-e^{\mathbb{X}\beta})$	

1. GLM의 구성요소 Generalized Linear Model

ex) y가 count data일 경우

$y = ax+b \Rightarrow E(y) = \mu$

시스템 요소 = ax+b

연결 함수 = $g(\cdot)$

$g(\mu) = ax+b$

$\log(\mu) = ax+b$

1. GLM의 구성요소 Generalized Linear Model

정준연결함수 : 가장 일반적인 연결 함수
평균을 자연 대수 모수로 변환하는 연결 함수

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)]$$

ex)

$$f(y_i;\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = (1-\pi_i)[\pi_i/(1-\pi_i)]^{y_i} = (1-\pi_i)\exp[y_i \ln(\frac{\pi_i}{1-\pi_i})]$$: NE Family

1. GLM의 구성요소 Generalized Linear Model

[Logistic Regression]

- y가 범주형일 경우 사용하는 대표적인 회귀 분석 방법
- y가 이항분포 B(n, p)를 따른다고 가정
- p = p(y=1 | x)

ex) 오늘 대중교통을 이용했는가?, 지난해 회귀분석 강의를 수강했는가?
부도예측, 신용평가, 고객이탈예측 등

$$\log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

$$P(Y=1|x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}$$

2. Logistic Regression 모형 Generalized Linear Model

[Logistic Regression]

- y가 범주형일 경우 사용하는 대표적인 회귀 분석 방법
- y가 이항분포 B(n, p)를 따른다고 가정
- p = p(y=1 | x)

ex) 오늘 대중교통을 이용했는가?, 지난해 회귀분석 강의를 수강했는가?
부도예측, 신용평가, 고객이탈예측 등

$$\log\boxed{\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)} = \beta_0 + \sum_{j=1}^{p}\beta_j x_j.$$

$$P(Y=1|x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p}\beta_j x_j)}{1+\exp(\beta_0 + \sum_{j=1}^{p}\beta_j x_j)}$$

2. Logistic Regression 모형 Generalized Linear Model

[Odds]

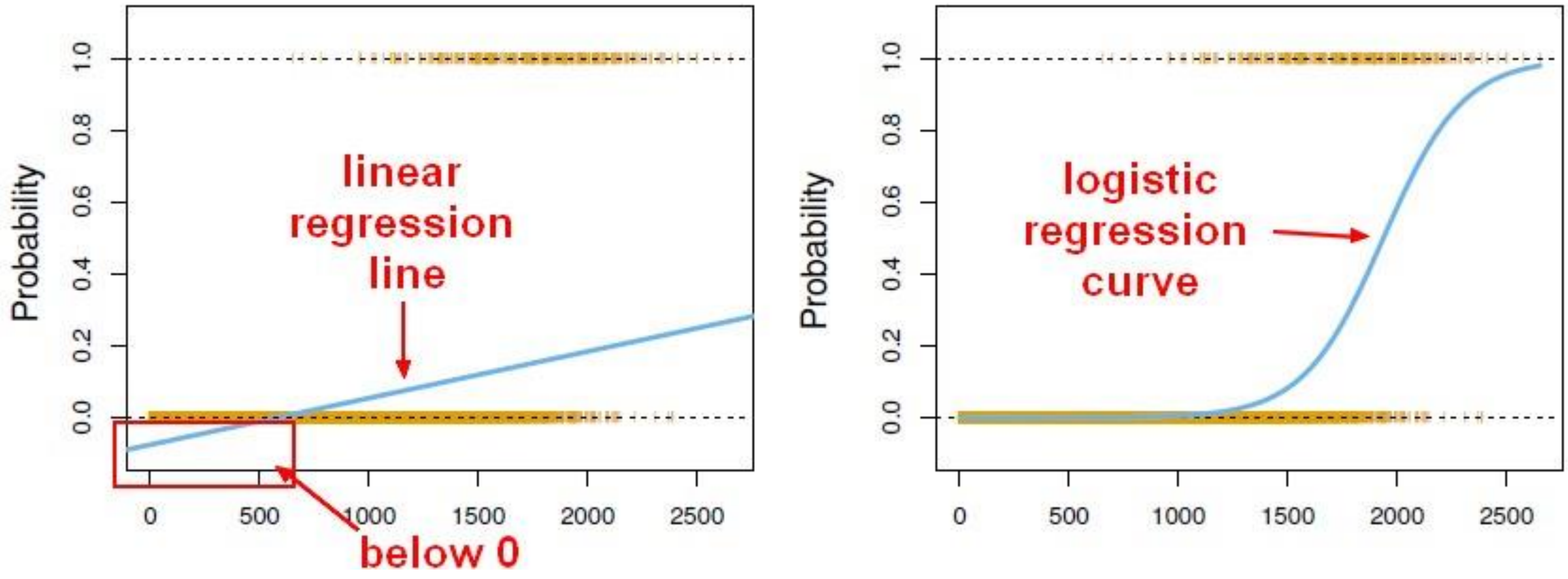
오즈(odds) = p/1−p
특정 리스크에 노출될 경우의 위험도로 해석

$$0 \le Odds = \frac{\text{성공확률}}{\text{실패확률}} = \frac{P(Y=1|x)}{1-P(Y=1|x)} \le \infty$$

$$-\infty \le \log(Odds) = Log\frac{P(Y=1|x)}{1-P(Y=1|x)} \le \infty$$

ex) X = 흡연 유무, Y = 폐질환 여부 (1, 0)
a = 3.72
→ odds = exp(3.72) = 42
→ 흡연자의 폐질환에 대한 위험이 비흡연자의 위험에 비해 42배 증가하는 것으로 해석

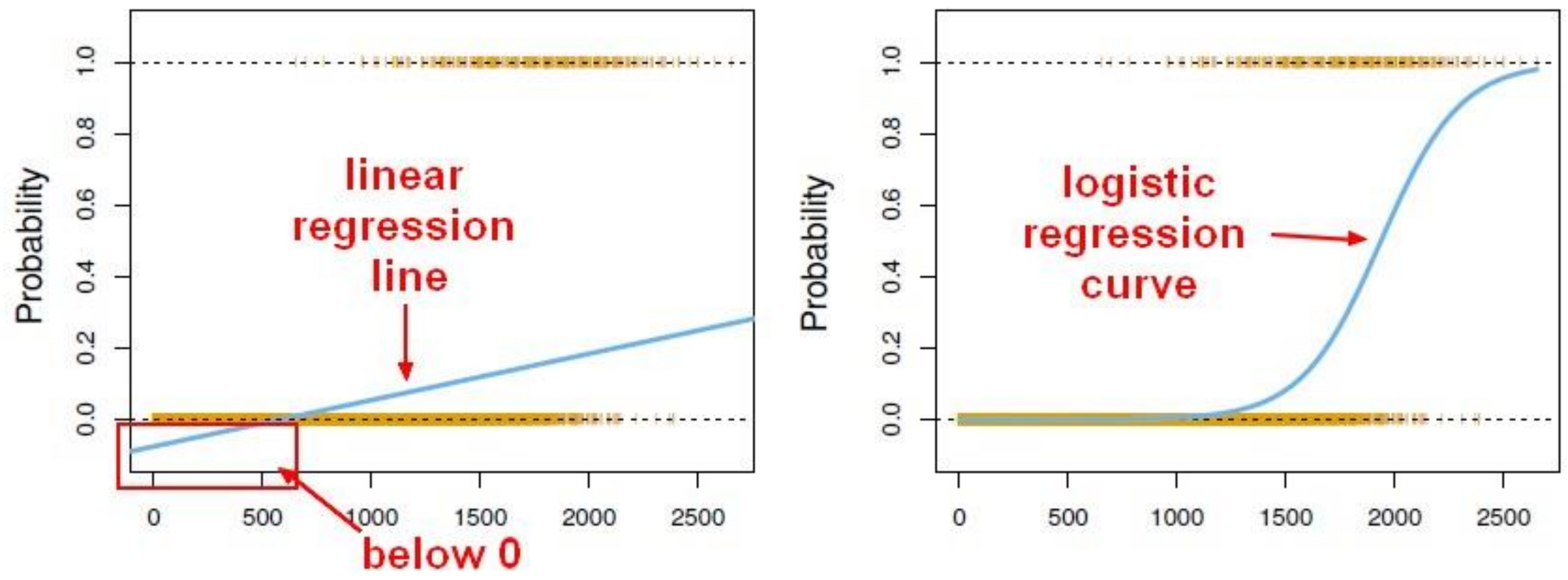
2. Logistic Regression 모형 Generalized Linear Model



추정하고자 하는 종속변수 p는 (0, 1)의 값을 가짐

→ 선형회귀 불가능

2. Logistic Regression 모형 Generalized Linear Model



비선형인 확률값을 선형으로 만들어야 함

→ p 자체가 아닌 p/1−p라는 Odds를 이용

3. Poisson Regression 모형 Generalized Linear Model

[Poisson Regression]

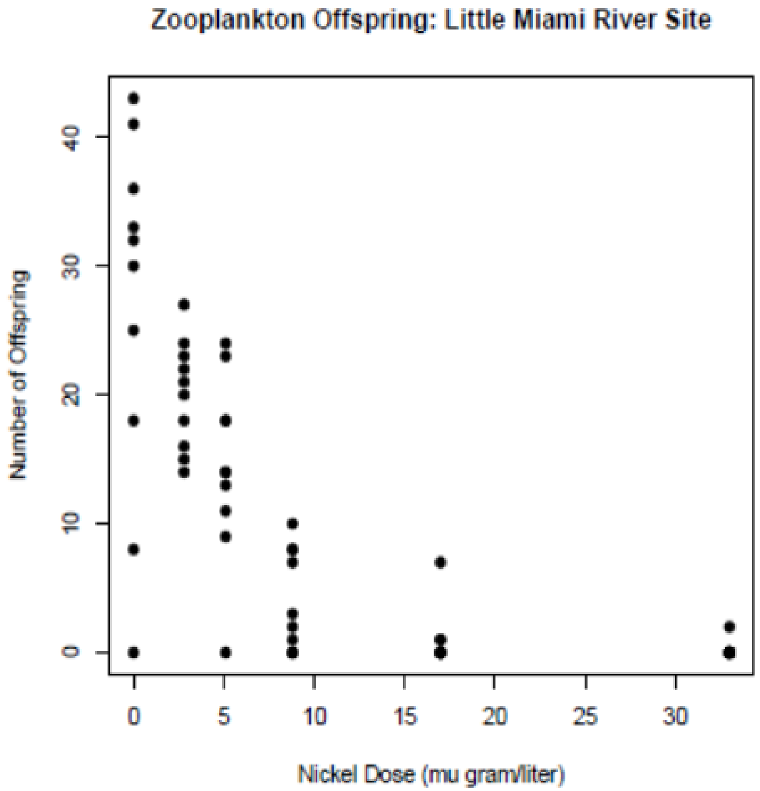
y가 도수(count data)일 경우 사용하는 일반화 선형 모형

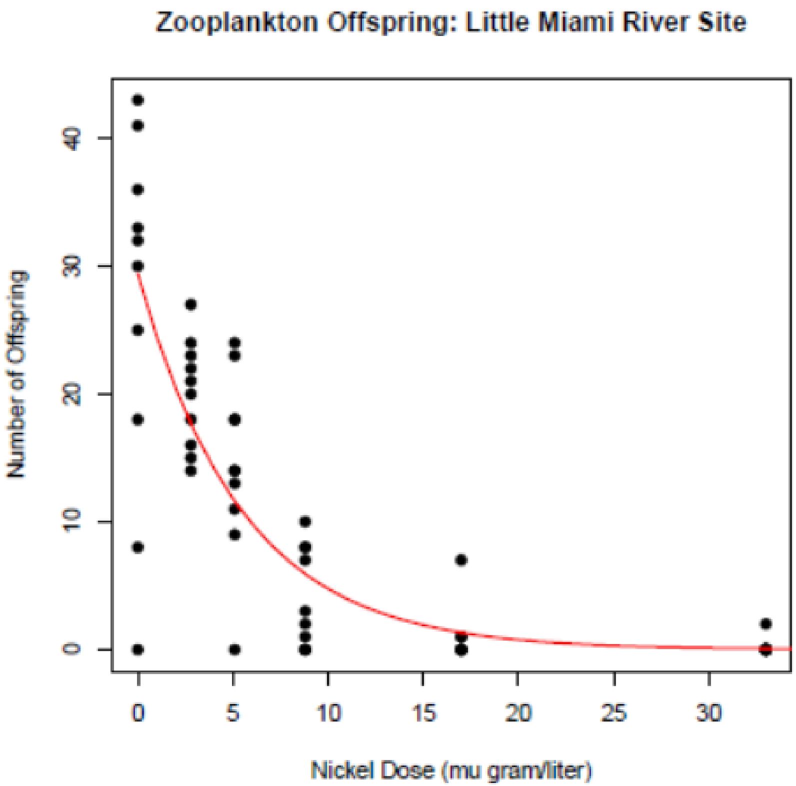
ex) A 고속도로에서 일어난 교통사고 횟수, 콜센터에 걸려온 전화 횟수 등

$$\log(\mu) = \beta_0+\beta_1 x_1+\ldots+\beta_p x_p$$

3. Poisson Regression 모형 Generalized Linear Model

ex)





3. Poisson Regression 모형

G e n e r a l i z e d   L i n e a r   M o d e l

```
Call:
glm(formula = y ~ dose, family = poisson)

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   3.38150     0.04959    68.19    <2e-16 ***
dose         -0.18225     0.01092   -16.69    <2e-16 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 872.94 on 59 degrees of freedom
Residual deviance: 251.32 on 58 degrees of freedom
AIC: 424.80

Number of Fisher Scoring iterations: 5
```

$$\log(\mu) = 3.38150 - 0.18225X \Rightarrow \mu = e^{3.38 - 0.182X}$$

3. Poisson Regression 모형 Generalized Linear Model

[Overdispersion]

포아송 분포의 경우 mean=variance=λ

하나의 모수를 공유하고 있으므로 overdispersion이 일어날 가능성 有

ex) crab data

3. Poisson Regression 모형

ex)

Table 3.3. Sample Mean and Variance of Number of Satellites

Width	No. Cases	No. Satellites	Sample Mean	Sample Variance
<23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
>29.25	14	72	5.14	8.29

3. Poisson Regression 모형

ex)

Table 3.3. Sample Mean and Variance of Number of Satellites

Width	No. Cases	No. Satellites	Sample Mean	Sample Variance
<23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
>29.25	14	72	5.14	8.29

3. Poisson Regression 모형 Generalized Linear Model

ex)

Table 3.3. Sample Mean and Variance of Number of Satellites

Width	No. Cases	No. Satellites	Sample Mean	Sample Variance
<23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
>29.25	14	72	5.14	8.29

THANK YOU !