REGULARIZATION & GENERALIZED LINEAR MODEL

Towards Work And Life Balance

Organizational Behavior

01 전통적 회귀분석의 과제

02 제약식에 의한 회귀계수의 조정

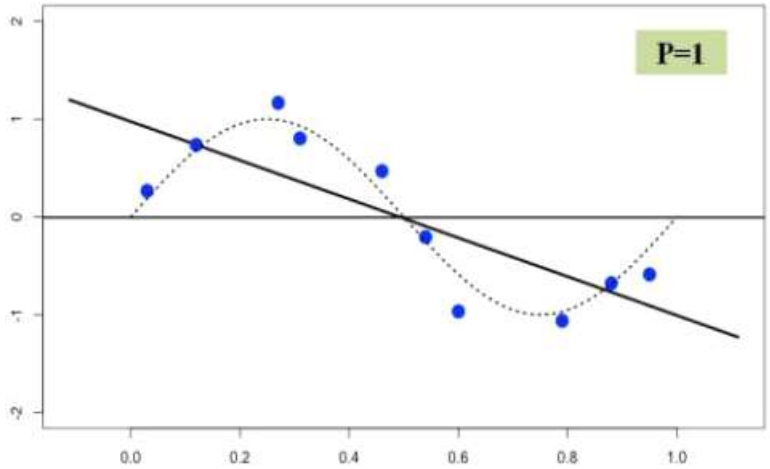
- Ridge
- LASSO
- Elastic Net

REGULARIZATIO

01 Regularization
전통적 회귀 분석의 과제

- 다항 회귀 모형 : $y = \beta_0 + \beta_1 X_1 + \beta_2 {X_1}^2 + \cdots + \beta_p {X_1}^p$

주어진 데이터에 적합한 함수를 추정해보자!



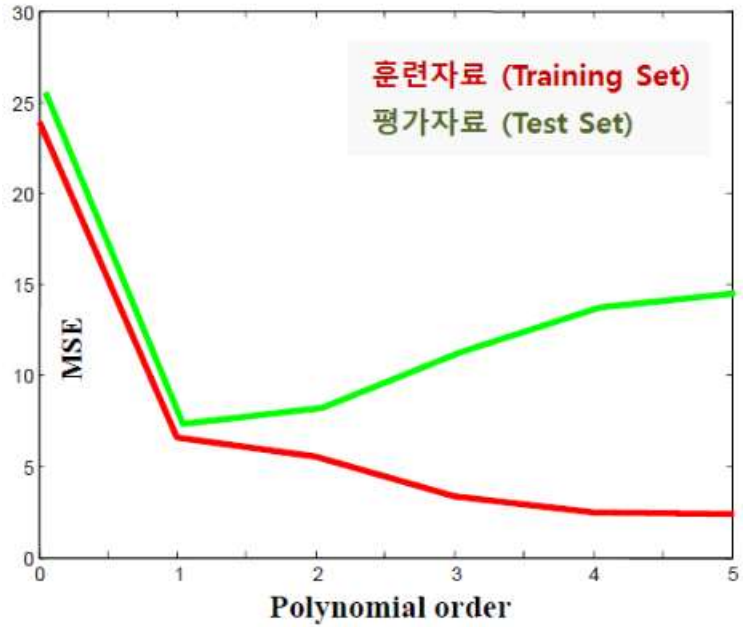
계수	p=1	p=3	p=5	p=7	p=9
β_0	0.975251	-0.12568	0.296001	-0.36812	5.906723
β_1	-1.98201	12.27827	-1.60235	28.4562	-306.064
β_2		-36.2189	63.53375	-289.686	4857.43
β_3		24.15899	-246.735	1494.172	-34459.8
β_4			309.7399	-3961.34	132901.4
β_5			-125.887	5370.273	-302706
β_6				-3549.17	417736
β_7				906.729	-342516
β_8					153250.7
β_9					-28762.9
MSE	0.2141	0.0166	0.0200	0.0075	0.0

01 Regularization
전통적 회귀 분석의 과제

입력 변수의 차수가 높아지면,
- 훈련 자료에의 적합도는 높아지지만,
  추정된 함수의 변동성이 커지게 되어
  평가자료에 측정한 모형 성능이
  매우 저하되는 현상이 나타난다!

반대로 낮은 차수의 함수를 사용하면,
- 변동성을 줄일 수 있지만
  모형의 왜곡도가 높아지는 상충효과가 나타난다.

→ **'편의-분산 딜레마'**
→ 따라서 회귀 계수들을 제한하여
분산을 제어하려는 것이
Regularized Regression의 기본 아이디어!



02 Regularization
제약식에 의한 회귀 계수의 조정

회귀 계수가 가질 수 있는 값의 범위를 제한하는
Regularized Regression 기법들 중에서 가장 많이 사용
→ Ridge, LASSO, Elastic Net 회귀분석

1) **Ridge Regression** : 제약식을 부가하여 회귀 계수 조정

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_i z_i)^2, \quad s.t. \sum_{j=1}^{p} \beta_j^2 \le t$$

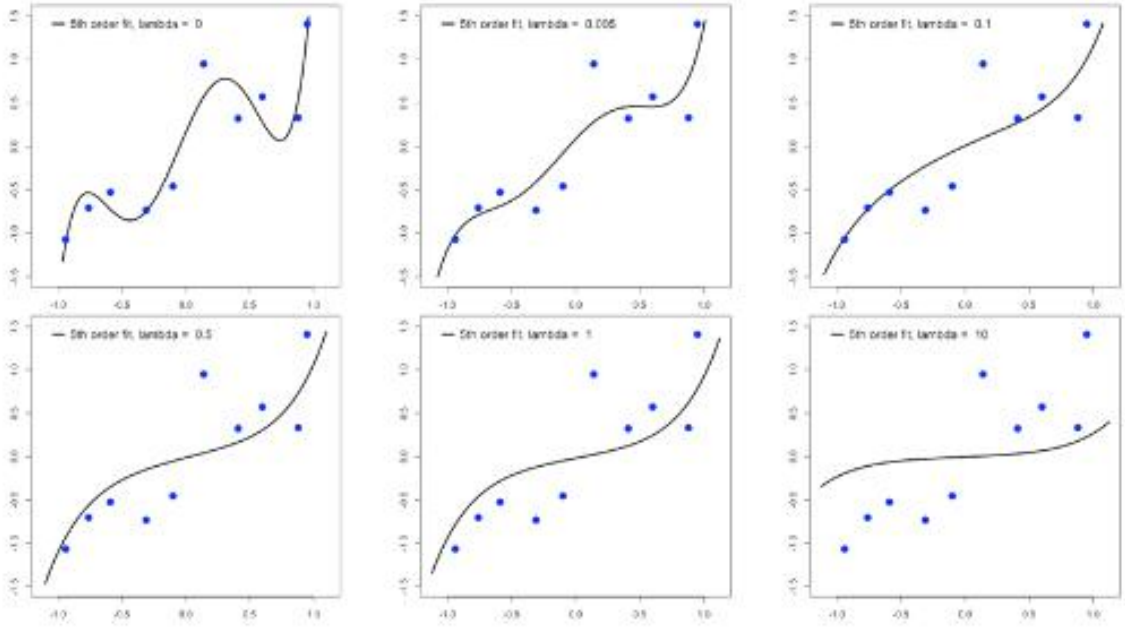
$$Min\ PRSS(\beta)_{L_2} = \sum_{i=1}^{n}(y_i - \beta_i z_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

02 Regularization
제약식에 의한 회귀 계수의 조정

* Ridge 회귀분석을 이용한 다항 회귀

10개의 데이터에 대하여 5차 다항 회귀식을 이용한 함수적합을 해보았다.
(λ =0,0.005,0.1,0.5,1,10 사용)

λ =0인 경우 기존과 동일 하므로
5차 다항식의 변동성이 상당히 큼.
하지만 λ 가 커짐에 따라
변동성이 축소되고 있으며
λ =10이 되면 평균값으로
접근하였음을 알 수 있음.
→ 이러한 현상은
λ 가 페널티의 양을
조절하였기 때문 (회귀계수 수축)
λ =10에서는 $\lambda \to \infty$가 되어도
회귀계수들이 0이 되지 않는다!



λ 에 따른 Ridge 회귀분석 결과의 변화

02 Regularization
제약식에 의한 회귀 계수의 조정

* LASSO

- Ridge : 회귀계수의 제곱합을 제약식으로 사용
- LASSO : 회귀계수의 절대값의 합을 제약식으로 사용

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_i z_i)^2, \quad s.t. \sum_{j=1}^{p}|\beta_j| \le t$$

$$Min\ PRSS(\beta)_{L_1} = \sum_{i=1}^{n}(y_i - \beta_i z_i)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

λ 는 Ridge 회귀에서와 마찬가지로 페널티 양을 조정하는 역할

But,

$\lambda \to \infty$ 이면 Ridge와 달리 완벽한 평균 모형이 된다.

(LASSO가 필요에 따라 회귀계수의 일부를 정확히 0으로 만들어 주기 때문)

→LASSO는 변수 선택의 기능이 있다!

02 Regularization
제약식에 의한 회귀 계수의 조정

Ridge vs. LASSO

- 회귀 계수 추정의 목적함수를 제약하는 패널티 항을 부가 > 분산 축소 효과
- 예측 성능에서는 큰 차이가 없다.

- Ridge Regression :

  회귀계수들이 0이 되지는 않는다.
  다중공선성이 존재할 때, 상관관계에 있는 변수들은 유사한 크기의 계수를 갖는다.
  → 그룹 선정 : 동일 변수들의 영향력은 1/k

- LASSO Regression :

  회귀계수들이 0이 될 수 있다.
  상관관계에 있는 변수들 중에서 임의로 하나만 선택되고 나머지 계수들은 0이 된다.
  → 변수 선택 : 해석이 유리하다.

02 Regularization
제약식에 의한 회귀 계수의 조정

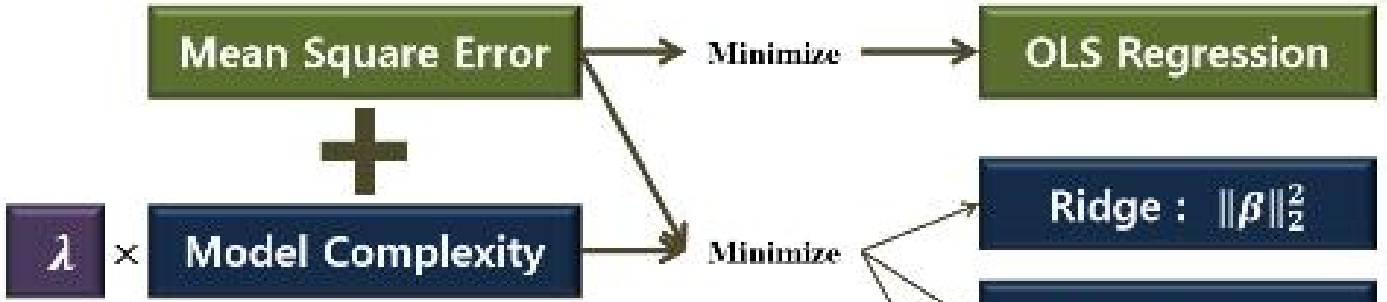
* Elastic Net

: Ridge와 LASSO 모형을 합친 것

$$\hat{\beta} = \arg\min_{\beta}(y - Z\beta)^T(y - Z\beta) + \lambda(\alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1)$$

- λ & α 두가지 파라미터를 사용해서 해 찾기가 복잡해진다!

- Summary :



❖ 조정된 회귀분석은 모형의 성능과
모형의 복잡성 사이의 균형을 추구한다 :

◆ Regularization Parameter λ를 추정하여야 한다 : CV 이용

01 GLM의 필요성
- GLM은 왜 필요한가..

02 GLM의 구성요소
- Random Component
- Systematic Component
- Link Fucntion

03 포아송 회귀 모형

04 로지스틱 회귀모형

G L M

01 Generalized Linear Model
일반화 선형모형(GLM)의 필요성

- 회귀분석, 분산분석 등 선형모형
  → 종속변수가 정규분포 되어 있는 연속형 변수

- 그렇지 않은 경우,
  - 종속변수가 범주형 변수인 경우
    : 이항변수(합격, 불합격)인 경우나,
    다형변수 (공화당/민주당/무소속 등)인 경우
  - 종속변수가 COUNT인 경우
    (ex. 한 주간 교통사고 발생건수, 하루에 마시는 물이 몇 잔인지)

- 확률 값을 예측 하는게 의미 있는데, 확률 값은 0과 1 사이에 존재하므로 종속변수 값의 범위가 제한이 됨.

이런 경우에서는 선형모형이 적절하지 않다!
대안 모형이 필요! GLM (일반화 선형 모형)

01 Generalized Linear Model
GLM의 구성요소

$$\varphi(\mu) = \mathbb{X}\beta$$

1. Random Component(임의 요소)

   : 종속변수 Y의 분포를 대변한다. 이항분포, 포아송, 감마, 정규분포 등

2. Systematic Component(시스템 요소)

   : 독립변수의 선형결합 부분

3. Link Function(연결 함수)

   : 반응변수 Y와 시스템 요소를 연결하는 함수

01 Generalized Linear Model
GLM의 구성요소

GLM은 연결함수에 따라 회귀모형이 달라진다.

종속변수
- 이진형, 범주형 또는 서수형 변수 → 로지스틱 회귀 모형
- 개수(COUNT) → 포아송 회귀 모형

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

02 Generalized Linear Model
포아송 회귀Poisson Regression

포아송 회귀

: 종속 변수 Y가 도수(COUNT)인 경우에 적용되는 일반화 선형 모형

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

02 Generalized Linear Model
포아송 회귀Poisson Regression

예시) 호주의 2002년의 당뇨병 사망 자료

표 1. 당뇨병 사망 자료 (파일명: diabetes.csv)

gender	age	deaths	popn	l_popn	agemidpt
Male	<25	3	1141100	13.9475	20
Male	25-34	0	485571	13.09308	30
Male	35-44	12	504312	13.13095	40
Male	45-54	25	447315	13.01102	50
Male	55-64	61	330902	12.70958	60
Male	65-74	130	226403	12.33007	70
Male	75-84	192	130527	11.77934	80
Male	85+	102	29785	10.30176	90
Female	<25	2	1086408	13.89839	20
Female	25-34	1	489948	13.10205	30
Female	35-44	3	504030	13.13039	40
Female	45-54	11	445763	13.00754	50
Female	55-64	30	323669	12.68748	60
Female	65-74	63	241488	12.39458	70
Female	75-84	174	179686	12.09897	80
Female	85+	159	67203	11.11547	90

deaths(사망자수)은 종속변수 Y

gender(성)은 설명변수 X1

age(연령) 은 설명변수 X2

성*나이의 조합마다 popn(인구) s가 붙어있다.

L_popn은 popn의 로그

02 Generalized Linear Model
포아송 회귀Poisson Regression

Y가 도수이고, s에 비례하여 커지는 경향이 있을 것이므로 Y의 평균을 다음과 같이 모형화

$Y \sim \text{Poisson}(\mu)$, 즉 평균 μ인 포아송 분포를 따르며[3)]

$\log_e \frac{\mu}{s}$가 성 효과와 연령 효과의 선형결합으로 표현된다.

따라서 모형식을 아래와 같이 쓸 수 있다.

$$\log_e \mu = \log_e s + \text{절편} + \text{성효과} + \text{연령효과}$$

03 Generalized Linear Model
로지스틱 회귀Logistic Regression 모형

로지스틱 회귀는

→ 결과가 범주형일 때 사용 !
ex. 학생이 문제를 맞을 것인지 틀릴 것인지, 내일 비가 올지 안 올지, 과목의 학점이 A인지 B인지 C인지)

→ 로지스틱 회귀는 종속변수가 이항분포 ($B\ \ n, p$) 를 따른다고 가정하므로,
p = $P(Y = 1|x)$ 인데, 베르누이 시행에 입력벡터 X가 주어졌을 때 \

성공할 확률을 의미

적용사례:
- 부도예측, 신용평가, 고객이탈예측, 목표고객 선정
- 정당 지지여부, 질병 진단

03 Generalized Linear Model
로지스틱 회귀Logistic Regression 모형

로지스틱 회귀
= GLM에서 Logit 연결함수를 사용한 것으로 표현.

$$\eta = log\frac{P(Y=1|X_1,\cdots,X_p)}{1-p(Y=1|X_1,\cdots,X_p)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

추정하고자하는 종속변수는 성공확률을 나타내므로 0과 1 사이에 있다.
→ 선형회귀 불가능

03 Generalized Linear Model
로지스틱 회귀Logistic Regression 모형

선형회귀가 불가능하기 때문에,
직접 성공확률을 예측하지 않고 **오즈Odds**를 이용한다.

1. Odds, 성공확률과 실패확률의 비를 예측한다.
2. Odds는 0과 무한대 사이의 값을 갖게 된다.
3. 비선형인 확률값을 선형으로 만들기 위해 **자연 로그**를 취한다!

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

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

로그를 취하면 모든 범위의 값을 취하게 하여
선형 회귀 모형의 적용이 가능하도록 한다!

03 Generalized Linear Model
로지스틱 회귀Logistic Regression 모형

※주의점※ - 회귀계수의 의미를 해석할 때

- 독립변수의 한 단위 증가가
  조건부 확률의 선형적 증가를 가지고 오지 않는다!
- Log(오즈), 오즈와 조건부 확률의 관계를 다시한번 확인하자!

	로짓(Logits)	오즈(Odds)	조건부 확률
모형	$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1$	$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_1}$	$p_i = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_0 x}}$
예	$\log(Odds) = -2 + 1.3x$	$Odds = e^{-2}e^{1.3x}$	$p = \frac{e^{-2+1.3x}}{1+e^{-2+1.3x}}$
해석	X 값 한 단위 증가시 $\log(Odds)$이 1.3 단위 증가	X 값 한 단위 증가 시 오즈(Odds)가 $e^{1.3} = 3.67$ 단위 증가	특정 x 값이 주어졌을 때 성공확률 $p(y=1\|x)$

※주의점※ - 회귀계수의 의미를 해석할 때

- 독립변수 **X1의 한 단위**의 증가
  = Log(오즈)의 **1.3단위** 증가 = 오즈가 **3.67배**(e^(1.3)=3.67) 증가

	로짓(Logits)	오즈(Odds)	조건부 확률
모형	$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1$	$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_1}$	$p_i = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_0 x}}$
예	$\log(Odds) = -2 + 1.3x$	$Odds = e^{-2}e^{1.3x}$	$p = \frac{e^{-2+1.3x}}{1+e^{-2+1.3x}}$
해석	X 값 한 단위 증가시 $\log(Odds)$이 1.3 단위 증가	X 값 한 단위 증가 시 오즈(Odds)가 $e^{1.3} = 3.67$ 단위 증가	특정 x 값이 주어졌을 때 성공확률 $p(y=1\|x)$

로지스틱 회귀Logistic Regression 모형

오즈(Odds)와 오즈비(Odds Ratio)의 확실한 구분이 필요해!

⇒ 오즈비는 독립변수 X가 한 단위 증가했을 때, 오즈의 변화율

$$OR = \frac{(p/(1-p))|_{X=x^{(2)}}}{(p/(1-p))|_{X=x^{(1)}}} :$$

IF $x \rightarrow x+1$,

$$OR = \frac{exp(\beta_0+\beta_1(x+1))}{exp(\beta_0+\beta_1 x)} = e^{\beta_1}$$

QUIZ-로지스틱 회귀식의 활용

심장병으로 10년 후 사망할 확률을 추정하기 위하여 100명의 연령, 성별 및 콜레스테롤 수치를 조사하여 다음과 같은 로지스틱 회귀식을 얻었다.

설명변수: $X_1 =$ 50세 초과 연령; $X_2 = 0$ (남자), 1(여자) ;

$X_3 =$ 5.0를 초과하는 콜레스테롤 수치

$$Risk\ of\ Death = \frac{1}{1+e^{-z}},$$

$$where\ \ z = -5.0 + 2.0X_1 - 1.0X_2 + 1.2X_3$$

1) 50세 남성의 콜레스테롤 수치가 7.0일 경우 사망 확률을 구하라.
2) 콜레스테롤 수치가 한 단위 증가했을 때의 사망 확률과 오즈비를 구하라.

QUIZ-로지스틱 회귀식의 활용

1) 50세 남성의 콜레스테롤 수치가 7.0일 경우 사망 확률을 구하라.

z = -5.0+2.0(50-50)-1.0(0)+1.2(7-5)=-2.6
➔1/(1+exp(2.6))=0.069(7%)

2) 콜레스테롤 수치가 한 단위 증가했을 때의 사망 확률과 오즈비를 구하라.
z = -5.0+2.0(50-50)-1.0(0)+1.2(8-5)=-1.4
➔1/(1+exp(1.4))=0.1978(20%)
오즈비: OR=exp(1.2)=3.32
오즈는 X3=2일 때, 0.0743, X3=3일 때, 0.2466

→ 따라서 사망위험이 3.32배 높아졌다는 표현은 가능해도 사망확률이 3.32로 높아진 것은 아니라는 점에 주의하자!

끝

Generalized Linear Model

감사합니다.

질문해주세요