필기

본 내용은 cic (https://platform.samsungcic.com/) 에서 Senior DS(Professional Pro DS) 자격 시험 필기 교육 영상을 기반하여 작성되었습니다.

통계를 전공하였기때문에 강의 내용 중 많이 익숙한 내용에 대해선 생략한 부분이 많습니다. 이 부분 참고하시기 바랍니다.

중간 중간에 '한 스푼'이라는 타이틀을 가지고 나오는 내용이 있는데 중간중간에 참고하기 좋은 내용을 적어두었습니다.

모수, 비모수 및 다변량 통계

▼ 확률 분포

데이터에 따라서 특정 구간의 분포가 달라져 각 분포의 확률이 달라지는 것을 확률 분포

1. 이산형 확률 분포
    - 베르누이, 이항, 포아송, 기하 , .. 분포
2. 연속형 확률 분포
    - 지수, 정규 ,… 분포
        - 정규분포

          실제 데이터와 가장 유사한 형태를 가진 분포

          (실제 현상을 비교적 가장 설명을 잘하는 분포)

통계적 추론(모수적 접근)

모수를 알 수 없으므로 가지고 있는 데이터(표본)을 통해서 모수를 추정한다.

모수를 추정하기 위해선 통계적 추론을 하는데 이는 크게 추정과 가설검정 두개로 나눠질 수 있다.

▼ (한 스푼) 통계량이란?

샘플들의 함수 (즉, 샘플들을 조합하여 도출한 수식)

통계량은 임의로 선택된 샘플에 의해서 계산되는 값이므로 확률변수이다.

추정하고싶은 분포의 파라미터를 나타내는 통계량을 추정량인거고, 가설검정에 활용되는 통계량을 검정통계량이라고 하는것이다.

Ex. 표본 평균, 분산

추정

> 샘플을 통해 모분포의 파라미터에 대한 정보를 도출하는 과정

- **점 추정**

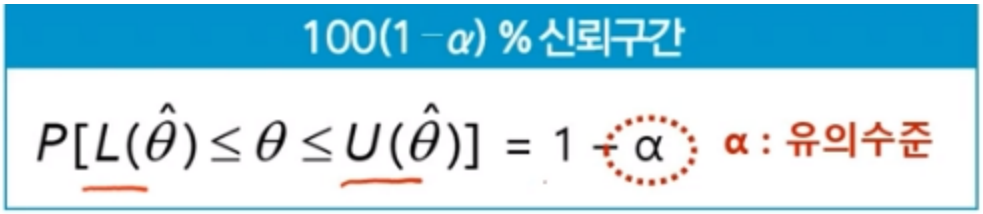
  파라미터 추정결과를 하나의 값으로 도출하는 방식

  보통 세타로 쓰며 추정량의 기대값이 파라미터와 같으면 불편 추정량이라고 한다.

- **구간 추정**

파라미터 추정을 구간으로 도출하는 방식

이 구간을 신뢰구간이라고 한다.



신뢰 구간의 폭은 추정량의 분포와 분산,유의수준에 의해 결정된다.

여기서 **신뢰도**는 샘플링을 무한히 반복하여 각 샘플별로 신뢰구간을 구했을 때, 신뢰구간들이 실제 파라미터를 포함할 확률을 의미한다.

표본 평균 및 분산을 샘플로부터 직접 추정량을 계산하는 위 방식들 외에도 **최대 우도 추정법**, **적률 생성 함수(mgf)** 이 있다.

▼ 최대 우도 추정법

> **우도(likelihood)** 란?
> 주어진 표본 x 정보들을 바탕으로 모집단의 모수에 대한 추정이 그럴듯한 정도
> 이 우도를 최대화 시키는 모평균/모분산에 대한 추정량을 구한 것이 Maximum Likelihood Estimator

데이터를 최대한 잘 설명할 수 있는 모분포의 파라미터를 추정하기 위해 개별 샘플들의 확률분포 함수 값들의 곱(우도함수)를 최대화

▼ 적률 생성 함수법

확률분포의 적률 생성 함수를 정의하고 이를 통해 확률 분포의 파라미터를 추정

> **적률** 이란?
> 확률변수 X^n의 기댓값을 의미한다. X^n의 기댓값을 n차 적률이라 하며, n이 1이면 1차 적률, 2이면 2차 적률이라 한다. 1차 적률은 평균(mean), 2차 적률은 분산(variance), 3차 적률은 왜도(skewness), 4차 적률은 첨도(kurtosis)를 구하는 데 사용될 수 있다.

> $$\mu_n = E[X^n]$$
>
> 적률, μ로도 표기

적률 생성 함수는 특정 확률 분포에 대한 "적률을 생성하는 함수"이다. 다음과 같이 정의되고 값을 계산하므로써 적률을 구할 수 있다.

$$M_X(t) = E[e^{tX}]$$

가설 검정

> 샘플을 통해 집단간의 차이에 대해 추론하는 과정

검정은 모수적 검점와 비모수적 검정으로 나눠지는데, 여기선 모수적 접근을 다루므로 모수적 검정에 대해 알아보자.

모수적 검정의 경우 정규분포를 따른다는 가정하에 이뤄지는데, 만약 정규분포를 따르지 않는 데이터의 경우에도 log,square root,지수 함수 적용 등을 통해 데이터 변환을 하거나 데이터를 더 수집하거나 이상치 제거 등을 수행하여 정규분포에 근사시켜 모수적 검정을 수행할 수 도 있다.

하지만 그런 작업을 수행하여도 정규분포를 따르지 않거나 극단적인 이상치가 존재하는 경우 등의 문제로 모수적 검정을 수행할 수 없는 경우가 있다. 이럴 경우 비모수적 검정을 수행한다.

- 모수적 검정

  모집단의 분포가 정규분포라는 가정 하에서 이뤄짐

  1. One sample t-test (집단이 1개이고 실험요인이 1개인 경우)

     보통 모수(평균)가 추정치라는 가설에서 쓰임

  2. Two sample t-test (집단이 2개이고 실험요인이 1개인 경우)

     두집단의 평균이 같다 다르다라는 가설에 쓰임

     Ex. 렌즈의 뒤틀림(**실험요인**)이 글라스 두께에 영향을 주는가?

     - 영향이 없다 (그룹 1,2가 동일) - 귀무가설

     ▼ 검정 절차

두 그룹의 평균을 뺀 값을 정규화 한 것이 T 라는 통계량이 되고, 귀무가설이 참이라는 가정 하에 만들어진 확률 분포(두 그룹 차의 분포 - 정규분포라는 가정과 평균이 같다는 가정이 있으니 두 그룹 차의 분포는 평균이 0 인 정규분포가 될 것)에서 P(Y>T) 가 pvalue가 되어 이 값과 유의수준을 비교하여 가설 검정한다.

3. One-way ANOVA (집단이 3개 이상이고 실험요인이 1개인 경우)

   u1=u2=u3 가설(요인에 효과가 없다 - 귀무가설)

   ▼ 검정 절차

$$x_{i1}, \cdots, x_{ir} \sim iid \;\; N(\mu_i, \sigma^2)$$

가정 : Ai 데이터는 정규 분포를 따름

SST = SSA + SSE

- SST : 총 제곱합

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{\bar{x}}\right)^2$$

- SSA : 그룹 영향의 제곱합

$$SS_A = r\sum_{i=1}^{a}\left(\bar{x}_{i\cdot} - \bar{\bar{x}}\right)^2$$

- SSE : 오차 제곱합

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{r}\left(x_{ij} - \bar{x}_{i\cdot}\right)^2$$

SSA >> SSE 면 그룹의 영향에 의한 변동이 오차에 의한 변동보다 총 변동을 상대적으로 많이 설명한다는 것이므로 반응 변수의 평균에 차이가 있다는 뜻이다.

즉, 귀무가설이 참이 아닐 확률이 높다는것이다.

그러나 제곱합의 경우 그룹의 갯수와 샘플의 크기에 영향을 받으므로 two-way 에서 정규화를 한것처럼 anova는 제곱합을 자유도로 나눈 평균 제곱합을 사용하여 검정을 수행한다.

	자유도	편차의 개수	선형 제약조건
SS_T (총 제곱합)에 관한 자유도	$\Phi_T = ar - 1$	ar	$\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}}) = 0$
SS_A (요인 A의 제곱합)에 관한 자유도	$\Phi_A = a - 1$	a	$\sum_{i=1}^{a}(\overline{x_{i\cdot}} - \bar{\bar{x}}) = 0$
SS_E (오차 제곱합)에 관한 자유도	$\Phi_E = a(r - 1)$	ar	$\sum_{i=1}^{a}\sum_{j=1}^{r}(x_{ij} - \overline{x_{i\cdot}}) = 0, i = 1, 2, \dots, a$

각각 자유도로 나눈 것을 MST,MSA,MSE 라고 한다.

여기서 MSE는 수식을 보면 결국 데이터의 분산과 같다. 그러면 MSE와 MSA는 아래와 같은 식으로 나타낼 수 있다.

$$E(MS_A) = \sigma^2 + \frac{r}{a-1}\sum_{i=1}^{a}\alpha_i^2 \qquad \alpha_i = \mu_i - \mu \quad \text{(그룹 } A_i \text{ 의 영향력)}$$

$$E(MS_E) = \sigma^2$$

이 정리를 바탕으로 검정할려는 귀무가설을 아래와 같이 변경 할 수 있다.

u= u1=u2=u3.. ⇒ **a1(x1평균 - x평균) = a2 = a3 = 0**

귀무가설이 참이라면 ai의 제곱의 합 또한 0이여야 하므로 MSA의 평균은 MSE의 평균과 동일하게 분산이 된다.

$$\sum_{i=1}^{a}\alpha_i^2 = 0 \Rightarrow E(MS_A) = \sigma^2 + \frac{r}{a-1}\sum_{i=1}^{a}\alpha_i^2 = \boldsymbol{\sigma^2} \Rightarrow \frac{E(MS_A)}{E(MS_E)} = 1$$

따라서 **MSA/MSE 이 1에 가까우면 귀무가설이 성립하고 1보다 커질수록 (MSA>>MSE) 대립가설이 성립하는 것**이다.

그럼 이 통계량을 어떤 분포를 기반하여 pvalue를 설정해야할까?

앞의 가정을 짚어보면 정규분포 베이스인 샘플이고 통계량은 평균제곱의 비율이므로 이때는 **F분포를 기반**하여 결정한다. (F검정 - 두 요소간 분산 비율)

$$F_0 = \frac{MS_A}{MS_E} = \frac{\frac{SS_A}{(a-1)}}{\frac{SS_E}{(ar-a)}} = \frac{\frac{SS_A}{\sigma^2}\frac{1}{(a-1)}}{\frac{SS_E}{\sigma^2}\frac{1}{(ar-a)}} \sim F((a-1),(ar-a))$$

- 만약 $F_0 > F_\alpha((a-1),(ar-a))$ 라면, H_0 기각

4. Two-way ANOVA (집단이 3개 이상이고 실험요인이 2개인 경우)

   A 요인으로 인한 집단간 평균의 차이가 없다.

   B 요인으로 인한 집단간 평균의 차이가 없다.

   ▼ 교호작용

   .두 요인 이상의 특정한 요인 수준 조합에서 일어나는 효과로 요인 A의 효과가 B의 수준의 변화에 따라 변하는 경우 교호작용 A*B가 존재한다고 한다.

- 가설 검정의 오류

  가설 검정은 결국 임의로 뽑힌 샘플에 의해서 이루어지는데, 해당 샘플이 특정 현상에 치우친 것들로만 이뤄지거나 하는 즉, 실제 현상을 반영하지 못하는 경우에 잘못된 판단을 할 수 있다. 이 때 발생되는 오류를 가설 검정의 오류라고 한다.

  - 1종 오류 / 2종 오류

    1종 오류의 경우는 분석자가 분석 전에 최대허용치를 유의수준으로 설정을 하고 진행한다.

- **P-value**

  귀무 가설이 맞다는 전제하에, 관측된 검정 통계량보다 귀무가설의 반대 방향으로 나타날 확률이다. 즉, **귀무가설이 참이라는 가정 아래 얻은 통계량이 귀무가설을 얼마나 지지하는 확률**인것이다. 이것이 유의수준(1종 오류를 겪을 최대 허용치)보다 낮으면 귀무가설이 틀릴 확률이 굉장히 크다고 판단하여 귀무가설을 기각한다.

다변량 통계

변수가 두개 이상인 데이터를 다변량 데이터라고 한다. 다변량 데이터의 경우 변수들간의 상관성 및 교호작용 또한 고려하여 데이터 분석을 수행해야 한다.

상관 분석 중 하나의 방법으로, 산점도는 변수들 간에 어떤 관계를 가지는지, 이상점을 가지는지, 몇개의 그룹으로 나뉘어지는지 등을 그래프로 그려 눈으로 직관적으로 확인 할 수 있다.

그리고 상관 관계를 정량적이고 객관적으로 파악하기 위해서 상관 계수를 계산할 수 있다.

- 상관 계수 종류
  1. 모수적 상관분석 - 피어슨 상관계수(=샘플 상관계수)
     a. 두 변수의 형태가 등간형 또는 비율형일 경우 사용
     b. 두 변수의 선형정도를 파악

     상관계수 값 = x,y의 공분산 / x 표준편차 * y 표준편차
  2. 비모수적 상관분석 - 스피어만 **순위** 상관계수 / 켄달 **순위** 상관계수
     a. 두 변수의 형태가 순서형인 경우(서열 변수) 사용
     b. 변수의 순위가 증가할 때, 다른 변수의 순위도 같이 증가 또는 감소하는지의 관계

     스피어만 순위 상관계수도 피어슨 상관계수와 동일한 계산식으로 나온다. 사실상 순위형 데이터 버전의 피어슨 상관계수이다.

     하지만 켄달의 경우 계산식이 다르다.

공분산

> 두 변수들의 연관성 뿐만 아니라 개별 변수의 분산 정도까지 동시에 고려하는 척도

두 변수 X,Y 간의 공분산은 다음과 같이 정의한다.

Cov(X,Y) = E[(X-E(X))(Y-E(Y))] = 상관계수(X,Y) * X 표준편차 * Y 표준편차

모집단의 공분산은 알려져있지 않은 경우엔 아래의 표본 공분산을 이용한다.

$$S(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \quad \Rightarrow \quad E[S(X,Y)] = Cov(X,Y)$$

(표본 공분산의 평균은 모집단의 공분산 즉, 불편추정량임이 통계적으로 증명되었다.)

시계열 분석

시계열 자료란 **일정한 시간 간격으로 측정되었거나 특정 시간 간격으로 수집된 주기적인 시간 간격을 따르는 자료를** 뜻합니다. 즉, 시계열은 단순히 시간에 따라 정렬된 일련의 데이터 지점이며 시계열 분석은 이 데이터를 이해하는 프로세스입니다.

- 시계열의 변동 형태
  - 체계적 변동
    - 추세 변동

      장기간에 걸쳐 나타나는 추세

      지속적으로 증가 또는 감소를 한다거나 일정한 상태를 유지하려는 성향
    - 순환 변동

      추세선에 따라 주기적으로 오르고 내림을 반복하는 변동
    - 계절 변동

      계절적 요인을 작용하여 특정 주기로 나타나는 변동
  - 불규칙적 변동

    규칙성 없이 예측 불가능하게 나타나는 변동

시계열의 특성 - 정상성 vs 비정상성

정상성은 시계열 분석에서 가장 중요한 요인중에 하나인데, 일부 시계열 분석 모형들은 정상성 가정하에 수립되는 경우가 있다. 예를 들면 시계열 회귀 모형과 ARMA 모형이 있다. 따라서 이러한 모형에 적용하기 위해선 비정상 자료를 정상성 조건을 만족시키도록 하여 정상 시계열 자료로 만든 후 (변환 또는 차분 수행) 분석을 실시해야합니다.

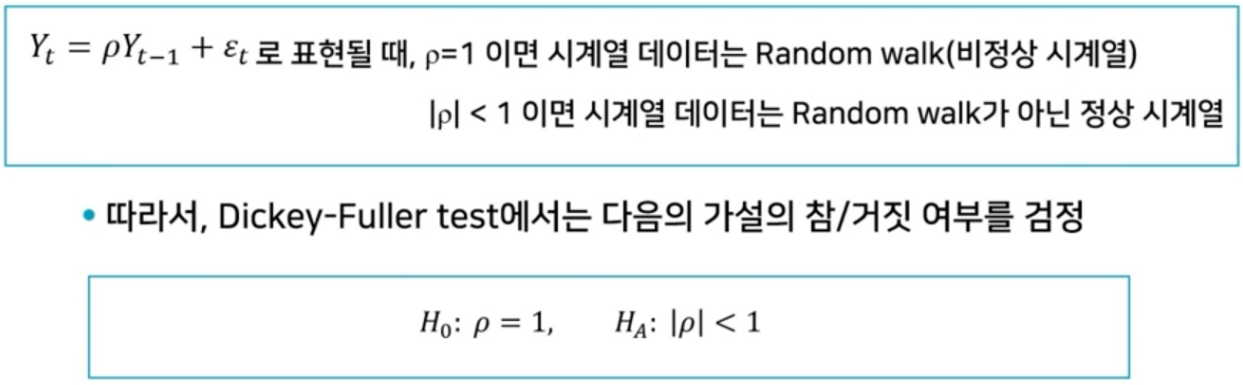
정상성이란?

모든 시점에 대해 평균이 일정한 상태이면서 (추세 x) 분산도 시점에 의존하지 않고 공분산은 단지 시차에만 의존하며 시점에 의존하지 않는 상태 즉, 시간에 따라 시계열의 확률구조가 변하지 않는 성질을 뜻합니다. 이는 매우 강한 가정인데, 만족하지 않는 자료가 대부분이므로 위에서 말한바와 같이 비정상 자료는 변환 또는 차분을 통해 정상 시계열로 만들어주는 것이 필요합니다.

정상성을 만족하는 시계열 데이터 종류중에 가장 대표적인 것은 백색잡음과정(white noise) , 1차 이동평균과정, 1차 자기회귀과정 이 있습니다.

그렇다면 비정상성을 갖는 시계열 데이터의 종류는 뭐가 있을까요? 대표적으로 random walk 모형과 random walk with drift 모형(random walk 모형에 상수값 추가된 모형)이 있다.

이 정상성을 판별하기 위해선 그래프를 보고 판단하는 방법과 검정 방법이 있다. 검정 방법중엔 단위근 검정으로 Dickey-Fuller test 와 KPSS(Kwiatkowski-Phillips-Schmidt-Shin) test가 있다



Dieckey-Fuller Test

비정상 시계열을 정상화하는 방법으론 두가지가 있는데, 위에서 말한바와 같이 **로그 변환** 과 **차분**이 있다.

- 로그변환

  값을 감소시킴으로써 시계열의 분산 변화를 일정하게 만듦

- 차분

  계열의 평균 변화를 일정하게 만듦

  - 1차 차분

    Y[t] - Y[t-1]

  - 2차 차분

    (Y[t] - Y[t-1]) - (Y[t-1] - Y[t-2])

  - 계절성 차분

    Y[t] - Y[t-m]

* 후방이동 연산자 B 는 한단계 전 단계를 표현하고 싶을때 BY[t] (= Y[t-1]) 를 쓴다

⇒ 차분 값 = (1-B)Y[t]

자기상관(Autocorrelation)

> **자기상관성**이란?
> 순차적으로 관측되는 자료들이 서로 종속적인 관계를 가지는것을 의미함. 예를 들면 그전 시점의 값이 1000을 가지는데 갑자기 -1000이 되지 않고 전의 큰 값을 가지니 그다음 값도 어느정도 큰값을 가지는 것을 자기상관이라고 함.
> 즉, **시계열의 공분산 행렬을 함수화 한 것**

자기상관은 보통 아래의 2가지 함수로 그 정도를 확인할 수 있다. 두 함수의 값에 대한 그래프를 이용하여 해당 시계열 데이터가 현재 정상성을 가지는지 계절성을 가지는지 등을 확인 할 수 있다.

자기 상관 함수(ACF)

- 시차가 다른(k) 시계열의 상관계수

$$\rho_k = Corr(Z_t, Z_{t+k}) = \frac{COV(Z_t, Z_{t+k})}{\sqrt{Var(Z_t)Var(Z_{t+k})}}$$

※ 백색 잡음 과정은 자기상관이 없다

부분자기상관함수(PACF)

시차가 다른 시계열의 종속성을 계산할때, **시차 사이의 시계열로 인한 종속성은 제외**한 상관계수

* 추가 모델로는 시계열 회귀 분석 / 비선형 시계열 회귀 분석이 있다.

기계 학습

Model Centric AI VS. Data Centric AI

Model Centric 은 모델 선택으로써 성능을 향상시킬려는 것이고, Data Centric 은 데이터의 질을 향상시키는것이 주이다. 이젠 Model Centric 은 많이 개선되었고, Data Centric 의 연구를 해야한다는 의견이 있다.

기계 학습이란 결국 함수를 찾는 것이다. (x,y) 데이터에서 f(x) = y(유사 y) 가 되는 함수 f(x)를 구하는것이 기계학습이다. 이것이 **알고리즘과 다른 점은 알고리즘은 인풋과 함수를 함께 컴퓨터**

에 넣으면 답이 나오는 방식이지만 기계학습은 인풋 데이터로 학습을 시켜 나온 함수를 컴퓨터에 넣어 문제를 푸는 방식이다. 즉, Training 과 Testing 이 함께 존재한다.

중략

릿지와 라쏘

모델을 생성할때 train에 학습이 덜 된 Bias가 큰 undefitted 된 모델을 만들수도 있고 train 맞춤형으로 Variance가 큰 overffited 된 모델을 만들 수도 있습니다. Bias와 Variance를 조절하여 최적화된 모델을 만드는것이 중요한데, bias를 줄이고 variance도 최소화 하기 위한 즉, overfitting을 해결할 수 있는 방법이 크게 두가지가 있습니다.

1. Feature 줄이기

   model selection 알고리즘을 사용하여 주요 특성만 선택하고 나머지는 버리는 방법이 있습니다. PCA를 사용하는 것도 결국 Feature를 줄이기 위한 방법입니다.

2. Regularization 수행

   모든 특성을 사용하되 파라미터의 값을 줄입니다. 베타값(파라미터값)에 제약을 줌으로써 모델을 정돈하는 것입니다. 과적합이 아닌 일반성을 띄게 해주는 해주는 것이죠.

$$\beta_1, \beta_2, \dots, \beta_p$$

$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1} (y_i - \hat{y}_i)^2}_{\text{(1) Training accuracy}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{(2) Generalization accuracy}}$$

현재 데이터에 대한 예측력도 중요하지만 미래에 예측할 데이터도 중요하기 때문에 일반화가 중요합니다. (1)Training accuray만 있으면 최소제곱법과 다른게 없는데 (2)Generalization

accuaracy가 추가되면서 베타에 제약을 줄 수 있어 정규화가 가능해지게 됩니다. 이렇게 계수 추정치를 줄여주는 정규화 방법을 shrinkage method라고 말하기도 합니다.

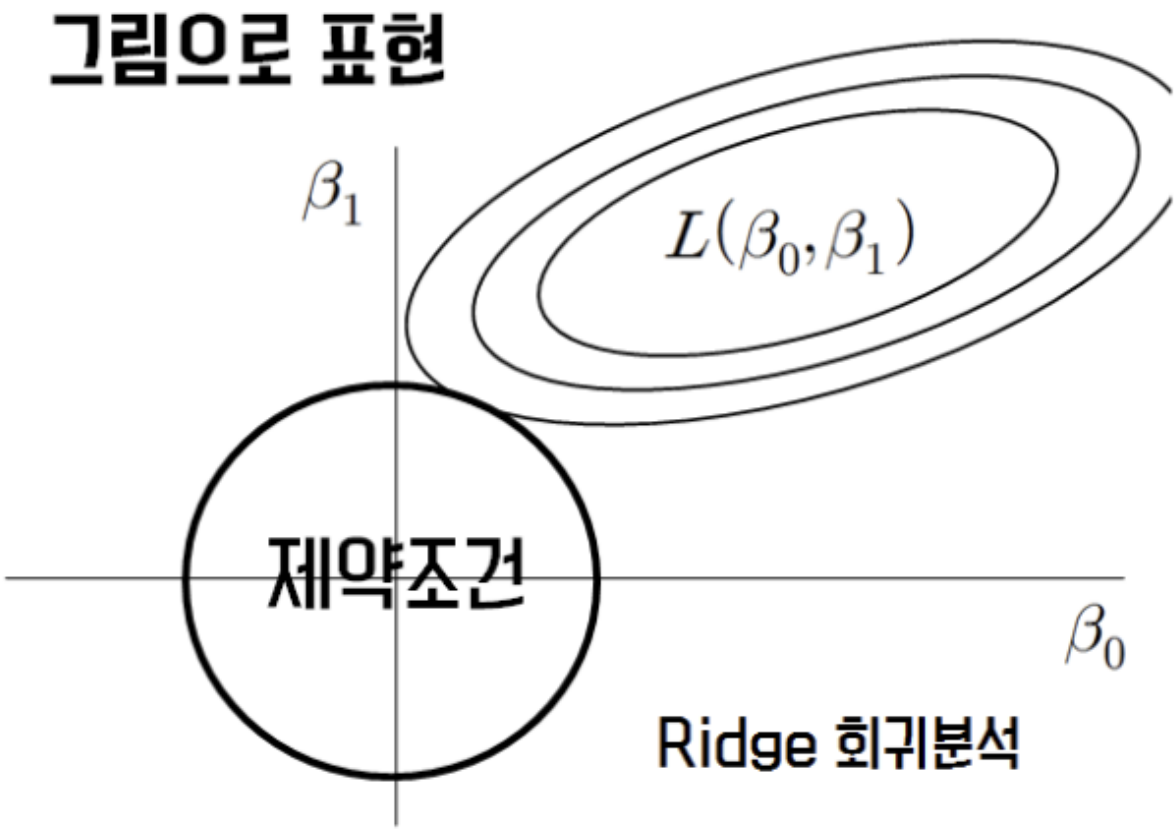
이런 정규화 컨셉을 처음 도입한 모델이 Ridge Regression 입니다.

Ridge Regression (릿지 회귀, L2 Regression)

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

릿지 회귀식을 보면 잔자제곱합(RSS)과 패널티 항(베타 값)의 합으로 이루어져있습니다. 릿지 회귀의 패널티항은 파라미터의 제곱을 더해준 것입니다. 이것은 미분이 가능해 Gradient Descent(경사 하강) 최적화가 가능하고, 파라미터의 크기가 작은 것보다 큰 것을 더 빠른 속도로 줄여줍니다. 다시 말하면 λ(람다)가 크면 클수록 리지회귀의 계수 추정치(베타)는 0에 가까워지는 것입니다. (λ = 0 일 때는 패널티 항은 효과가 없고, 따라서 리지 회귀는 일반적인 회귀와 같이 최소제곱추정치를 생성) 즉, λ(람다)가 패널티를 얼마나 부과하는가를 조절하는 조절버튼이라고 생각하면 되겠네요.

어떻게 이 모델이 일반 모델보다 일반화 된건지 이해하기 어려우실 수 있는데 아래의 그림을 보면 직관적으로 이해가 되실겁니다.



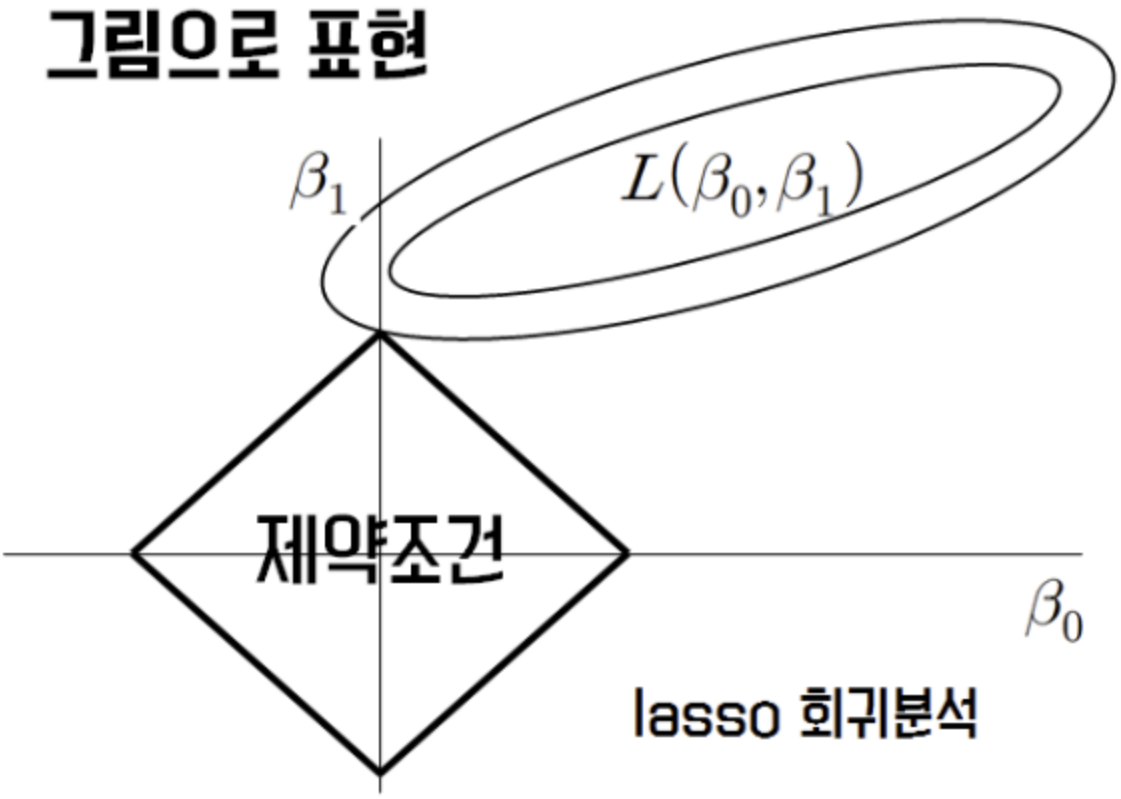
릿지회귀의 β0^2+β1^2의 제약조건인 원이 있습니다. 기존의 OLS(Ordinary Least Squares)가 아래에 보이는 제약조건까지 와야지 최적값이라고 할 수 있는 것입니다. 즉, loss function이 바뀌므로 기본 모델에서 최적화된 베타값이 아닌 제약조건에 맞는 다른 베타 값을 가지는(기존 오차가 좀 더 커질 수 있는?) 모델이 되는 것입니다. OLS가 제약조건까지 오기 위해서 RSS(RSS : residual sum of squares) 크기를 키워주는거죠. bias가 약간의 희생은 하지만 variance를 줄이기 위해서 아래의 그림처럼 제약조건까지 오는 가장 작은 RSS를 고르면 되는 것이죠.

Lasso Regression (라쏘 회귀, L1 Regression)

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

라쏘회귀는 릿지회귀와 비슷하게 생겼지만 패널티 항에 절대값의 합을 주었습니다.

라쏘는 제약조건이 절대값이라 아래의 그림처럼 마름모꼴의 형태로 나타납니다. 릿지회귀와 비슷하게 OLS의 RSS 값을 크게 늘려줍니다. 라쏘회귀의 경우 최적값은 모서리 부분에서 나타날 확률이 릿지에 비해 높아 몇몇 유의미하지 않은 변수들에 대해 계수를 0에 가깝게 추정해 주어 변수 선택 효과를 가져오게 됩니다. 라쏘회귀는 파라미터의 크기에 관계없이 같은 수준의 Regularization을 적용하기 때문에 작은 값의 파라미터를 0으로 만들어(아래 그림에서와 같이 β0 를 0으로 만듭니다) 해당 변수를 모델에서 삭제하고 따라서 모델을 단순하게 만들어주고 해석에 용이하게 만들어줍니다.



반면 릿지의 경우 어느정도 상관성을 가지는 변수들에 대해서 pulling이 되는 효과를 보여줘 변수 선택보다는 상관성이 있는 변수들에 대해서 적절한 가중치 배분을 하게 됩니다. 따라서 릿지의 경우 PCA와 상당한 관련성이 있게 됩니다.

Ridge	Lasso
L_2-norm regularization	L_1-norm regularization
변수 선택 불가능	변수 선택 가능
Closed form solution 존재 (미분으로 구함)	Closed form solution이 존재하지 않음 (numerical optimization 이용)
변수 간 상관관계가 높은 상황 (collinearity)에서 좋은 예측 성능	변수 간 상관관계가 높은 상황에서 ridge에 비해 상대적으로 예측 성능이 떨어짐
크기가 큰 변수를 우선적으로 줄이는 경향이 있음	

- https://rk1993.tistory.com/entry/Ridge-regression와-Lasso-regression-쉽게-이해하기 참고

모의 시험 정리

부록

1. 정규성 확인 방법

- 그래프 활용
  - 히스토그램
  - QQ Plot
- 통계적 검정 수행
  - Shapiro-Wilk
  - D’Agostino’s K2