회귀분석

목차

목차 **1**

CLRM(Classical Linear Regression Model)의 가정 **1**

최소자승법(OLS : Ordinary Linear Squared) **2**

[회귀계수 0 , 1에 대한 검정] 2

[결정계수 (coefficient of determinant) = R2] 3

Gause - Markov Theorethm **3**

BLUE 4

최대우도추정(maximum likelihood estimation) **4**

정의 4

최대우도추정 vs 최소제곱오차 6

참조
Jay님 블로그 https://m.blog.naver.com/jhkang8420/221291682151
이기창님 블로그 https://ratsgo.github.io/statistics/2017/09/23/MLE/

CLRM(Classical Linear Regression Model)의 가정

1. **두 변수간 선형관계가 있어야 한다.**
2. **표본추출이 무작위로 일어나야 한다.**
3. **설명변수 X의 값이 두 개이상이어야 한다.**
   [직관적]
   적어도 두 점이 있어야 한 직선이 만들어진다.
   [수식적]
   만약 x의 값이 하나라 가정하자.
   최소자승법(OLS : Ordinary Linear Squared)에서 기울기 $\alpha = S_{XY} / S_X^2$ 이다.
   가정에 의해 x의 값이 하나이므로 $S_X = 0$ 이 되므로 기울기의 분모 또한 0 이 된다.
   분모가 0 이 될 수 없으므로 x의 값이 하나라는 가정은 틀리다.
   따라서 x의 값은 두 개 이상이다.

4. **Zero - Conditional Mean**
   $E(\varepsilon_i \mid x_i) = 0 \, for \, \forall \, i = 1, 2, \ldots, n$

5. **등분산성**
   모든 X_i 에 대하여 오차들이 같은 정도로 퍼져 있다.
   $Var(\varepsilon_i) = \sigma^2 \, for \, \forall \, i = 1, 2, \ldots, n$

6. **독립성**
   $cov(\varepsilon_i, \varepsilon_j) = 0 \, for \, i \neq j$
   모 회귀모형 : $Y_i = \alpha X_i + \beta + \varepsilon_i$ 일 때, Y_i 와 ε_i 이 분산이 같은 이유
   $X_i = x_i$ 가 주어졌을 때, Y_i 의 평균인 $\alpha x_i + \beta \, [Y_i \mid X = X_i \sim N(\alpha X_i + \beta, \sigma^2)]$ 는
   상수값으로 분산 σ^2 을 구하는데 영향을 미치지 않음
   즉, Y_i, Y_j 는 서로 독립 ⇒ $\varepsilon_i, \varepsilon_j$ 는 서로 독립

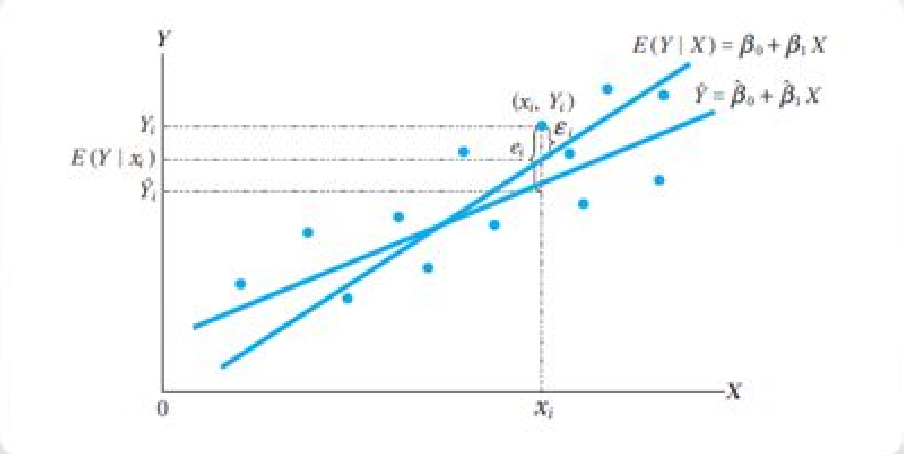
7. **정규성**
   $\varepsilon_i \sim N(0, \sigma^2)$

*가정 4에서 7까지는 오차항 ε 대한 내용

(cf) 편향되지 않는다. ⇔ $E(\widehat{\theta}) = \theta \, where \, \theta$: 모수, $\widehat{\theta}$: 모수의 추정치

최소자승법(OLS : Ordinary Linear Squared)

그림 10-9 모회귀선과 추정된 회귀선



$$min \sum_{i=1}^{n}(y_i - \hat{y_i})^2 \ for\ \hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i \ where\ \beta_0, \beta_1\ are\ chosen\ to\ minimize$$

[증명]
https://drive.google.com/file/d/1kyLC6OKV-zLjvstSUN2g2clrrowf6Ny_/view?usp=sharing

[회귀계수 $\hat{\beta_0}$, $\hat{\beta_1}$ 에 대한 검정]

- $\hat{\beta_0}$ 에 대한 검정

  1. 가설 설정
     $H_0 : \beta_0 = 0$
     $H_1 : \beta_0 \neq 0$

  2. 검정통계량

     $$T(X) = \hat{\beta_0} / \hat{\sigma} \sqrt{1/n + \overline{x^2} / \sum_{i=1}^{n}(X_i - \overline{X})^2} \sim t_{n-2}$$

- $\hat{\beta_1}$ 에 대한 검정

  1. 가설 설정
     $H_0 : \beta_1 = 0$ (⇔ x는 y를 설명하는 중요변수가 아니다.)
     $H_1 : \beta_1 \neq 0$ (⇔ x는 유의하다.(significant))

  2. 검정통계량

     $$T(X) = (\widehat{\beta_1} - \beta_1) / \hat{\sigma} \sqrt{1 / \sum_{i=1}^{n}(X_i - \overline{X})^2} \sim t_{n-2}$$

[결정계수 (coefficient of determinant) = R^2]

$$\sum(Y_i - \overline{Y})^2 = \sum(\widehat{Y_i} - \overline{Y})^2 + \sum(Y_i - \widehat{Y_i})^2$$

SST = SSR + SSE

- 총제곱합 SST $\sum(Y_i - \overline{Y})^2$
- 회귀제곱합 SSR $\sum(\widehat{Y_i} - \overline{Y})^2$
- 잔차제곱합 SSE $\sum(Y_i - \widehat{Y_i})^2$

결정계수 : 총변동 중 회귀선에 의해 설명되는 변동SSR의 비율을 측정하는 방법

1. $R^2 = SSR / SST = 1 - (SSE / SST)$
2. $0 \le R^2 \le 1$ (1에 가까울수록 회귀선 설명력이 큼)
3. $R^2 = \gamma^2 \ where\, \gamma = S_{XY} / \sqrt{S_X \cdot S_Y}$

[조정된 R^2 (Adjusted R^2)] 설명변수가 다른 회귀식 중에서 가장 적합한 것을 고를때 이용.

다중회귀분석에서 쓰인다. (설명계수 X가 여러 개인 경우)

결정계수는 설명계수가 늘어나면 값이 상승하는 결점을 가지고 있으므로 자유도 조정 결정계수 $\overline{R}^2$이 이용하며 이를 보정.

$\overline{R^2} = Adjusted\, R^2 = 1 - ((SSE / (n-k-1)) / (SST / (n-1)))\, for\ k = x$의 개수

$$= 1 - (1-R^2)\frac{n-1}{n-k-1}$$

Gause - Markov Theorethm

1. 오차변수의 기대값은 0 이다.
   $E(\varepsilon) = 0$
2. 오차변수와 독립변수의 공분산은 0 이다.
   $cov(X, \varepsilon) = 0$
3. 오차변수의 분산은 일정한 상수이다.
   $var(\varepsilon) = E(\varepsilon - E(\varepsilon))^2 = E(\varepsilon^2) - E(\varepsilon)^2 = \sigma^2$
4. 오차변수 사이의 공분산은 0 이다.
   $cov(\varepsilon_i, \varepsilon_j) = 0\, for\ i \neq j$

   $\because cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i\ \varepsilon_j) - E(\varepsilon_i)E(\varepsilon_j) = E(\varepsilon_i\ \varepsilon_j) = 0$

5. 오차변수는 정규분포를 따른다.

여기서 1~4까지 성립되면 BLUE(Best Linear Unbiased Estimator)이고
1~5까지 모두 성립되면 MVUE(Maximum Value Unbiased Estimator)이다.

BLUE

선형회귀에서 BLUE는 다음과 같이 볼 수 있다.

1) Linear

   $\widehat{\beta_1} = \sum x_i y_i / \sum x_i^2 = (1/ \sum x_i^2)\,[x_1 y_1 + \ldots + x_n y_n]$ 인

   $(1/A)\,[a_1 y_1 + \ldots + a_n y_n$ 꼴의 선형함수

2) Unbiased

   $E\,(\widehat{\beta_1}) = \beta_1$

3) Best

   $var\,(\widehat{\beta_1}) = \sigma^2 / \sum x_i^2$

   ⇒ 선형이고 불편이 추정량 중에서 분산이 가장 작다.

최대우도추정(maximum likelihood estimation)

정의

: 모수(parameter)가 미지의 θ인 확률분포에서 뽑은 표본(관측치) x들을 바탕으로 θ를 추정하는 기법

- **θ에 대해 편미분을 해 0이 되는 지점**을 구하면 우도를 최대화하는 θ를 단박에 구할 수 있다.

- θ에 대해 **미분이 불가능**할 경우에는 **그래디언트 디센트** 등 반복적이고 점진적인 방식으로 θ를 추정하게 된다.

- 로지스틱 회귀나 딥러닝 등 모델의 θ를 최대우도추정 기법으로 추정할 때 자주 쓰는 기법

$$\theta_{ML} = \arg\max_{\theta} P_{model}(X|\theta)$$
$$= \arg\max_{\theta} \left\{ E_{X\sim\hat{P}_{data}} [\log P_{model}(x|\theta)] \right\}$$

1. 확률이 1보다 작기 때문에 계속 곱해주면 값이 지나치게 작아지게 되어underflow 발생
⇒ 로그 취해주기
2. 전체 값에 로그를 취하거나 스케일을 하여도 대소관계 변화 無

쿨백-라이블러 발산(Kullback-Leibler divergence, KLD)
: 두 확률분포의 차이를 계산하는 데 사용하는 함수

$$D_{KL}\left(P||Q\right) = E_{X\sim\hat{P}_{data}}\left[\log\hat{P}_{data}(x) - \log P_{model}(x)\right]$$

- 딥러닝 모델을 만들 때 예로 들면 우리가 가지고 있는 데이터의 분포 P_{data}와 모델이 추정한 데이터의 분포 P_{model} 간에 차이를 KLD를 활용해 구할 수 있다.

- KLD를 최소화하는 것이 모델의 학습 과정
  ⇒ P_{data} 는 변하지 않으므로 P_{model} 을 최소화 하는 값을 찾아야 한다.

따라서 크로스 엔트로피(Cross Entropy)는 아래와 같고
크로스 엔트로피(혹은 KLD) 최소화가 **우도의 최대화**와 본질적으로 같다.

$$-E_{X\sim\hat{P}_{data}}\left[\log P_{model}(x)\right]$$

때문에 최대우도추정은 우리가 가지고 있는 데이터의 분포와 모델이 추정한 데이터의 분포를 가장 유사하게 만들어주는 모수(파라메터)를 찾아내는 방법이라고 볼 수 있다.

$$-E_{X\sim\hat{P}_{data}}\left[\log P_{model}(x)\right]$$

$$= E_{X\sim P}\left[-\log Q(x)\right] = -\sum_x P(x)\log Q(x)$$

$$= -\sum_x P(x)\log\left(\frac{Q\left(x\right)}{P(x)}\right)$$

$$= -\sum_x P(x)\left\{\log Q(x) - \log P(x)\right\}$$

$$= -\sum_x \left\{P(x)\log Q(x) - P(x)\log P(x)\right\}$$

$$= -\sum_x P(x)\log Q(x) + \sum_x P(x)\log P(x)$$

$$= H(P,Q) - H(P)$$

최대우도추정 vs 최소제곱오차

다시말해서 최대우도추정은 입력값 X와 모델의 파라메터 θ가 주어졌을 때 정답 Y가 나타날 확률을 최대화하는 θ를 찾는 것을 말한다.

m개의 모든 관측치가 i.i.d(independent and identically distributed)라고 가정하고,
언더플로우 방지를 위해 우도에 로그를 취하면

$$\begin{aligned}\theta_{ML} &= arg\max_{\theta} P_{model}\left(Y|X;\theta\right) \\ &= arg\max_{\theta} \sum_{i=1}^{m} \log P_{model}\left(y_i|x_i;\theta\right)\end{aligned}$$

P_{model}이 가우시안 확률함수라고(X와 Y가 정규분포를 따를 것이라고) 가정해보면

$$\sum_{i=1}^{m} \log P_{model}\left(y_i|x_i;\theta\right) = \sum_{i=1}^{m} logf(x_i)$$

$where$ 정규분포곡선 함수 $f(x_i) = (1/\sigma\sqrt{2\pi})exp(-(x-\mu)^2/2\sigma^2)$

$$= -m\log\sigma - \frac{m}{2}\log 2\pi - \sum_{i=1}^{m} \frac{\|\hat{y}_i - y_i\|^2}{2\sigma^2}$$

평균제곱오차(Mean Squared Error)

$$MSE = \frac{1}{m}\sum_{i=1}^{m} \|\hat{y}_i - y_i\|^2$$

∴ 가정하는 확률모델이 정규분포일 경우, 우도를 최대화하는 모수(파라메터)와
평균제곱오차를 최소화하는 모수가 본질적으로 동일

[최대우도추정 기법으로 추정한 모수의 특성]

1. **일치성(consistency)**
   추정에 사용하는 표본의 크기가 커질 수록 진짜 모수값에 수렴하는 특성
2. **효율성(efficiency)**
   일치성 등에서 같은 추정량 가운데서도 분산이 작은 특성

- 추정량의 **효율성**을 따질 때는 보통 **평균제곱오차(MSE)를 기준**으로 한다.

- 크래머-라오 하한 정리에 따르면, **일치성**을 가진 추정량 가운데 **최대우도추정**량보다 낮은 MSE를 지닌 추정량이 존재하지 않는다.