DSL Seminar: 통입+통방 (6)

Kyung-han Kim

Data Science Lab

February, 2023

목차

- 각종 가설 검정의 예시
- 회귀 분석
  - 단순 선형 회귀분석
  - 회귀분석의 5가지 표준 가정
  - 회귀계수의 OLS 추정량

가설 검정의 구조

1. 귀무가설과 대립가설을 설정한다.
2. 적절한 검정통계량을 설정하고, 그 값을 계산한다.
3. 유의 수준과 p-value를 비교한다.
   혹은, 임계치와 검정통계량 값을 비교한다.
   (p-value: 현재보다 더 극단적인 검정통계량 값이 계산될 확률)
4. (3)에서의 결과에 따라 귀무가설 기각 여부를 결정한다.

- 대립가설의 형태에 따라 단측 검정과 양측 검정으로 나뉜다.

가설 검정 (1): 모평균 검정

- 대한민국 성인 남성 평균 키가 190cm인지 검정. (유의수준: 0.05, N)
- H_0: $\mu = 190$, H_1: $\mu < 190$
- 임의로 뽑은 100명의 평균 키가 185cm였고,
  성인 남성 키의 표준편차는 10cm로 알려져 있다고 해 봅시다.
- 검정통계량 $T = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{185-190}{10/\sqrt{100}} = \frac{-5}{1} = -5$
- $T \sim Z$이므로 (p-value) $= P(T < -5) = 2.8 \times 10^{-7} < 0.05$.
- p-value가 유의수준보다 작음을 확인한다.
  $\therefore$ Reject H_0.

- 만약 모표준편차를 모른다면 표본표준편차 (S)를 사용하는 대신,
  $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} = t_9$가 된다.
- 표본표준편차도 10cm였다면,
  (p-value) $= P(T < -5) = 0.0003 < 0.05$. $\therefore$ Reject H_0.

가설 검정 (2): 두 집단의 평균 차이 검정 (σ^2 known)

- 대한민국 성인 남성의 키 (X_1)와 여성의 키 (X_2)의 평균이 다른가? (정규)
- H_0: $\mu_1 = \mu_2$, H_1: not H_0
- 임의로 뽑은 남성, 여성 각각 100명의 평균 키가 185cm, 175cm였고 표준편차는 모두 10cm로 알려져 있다고 하자.
- 검정통계량:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{10}{\sqrt{2}}$$

- $T \sim Z$이므로 (p-value) $= P(T > 5\sqrt{2}) = 7.7 \times 10^{-13} < 0.025$.
  $\therefore$ Reject H_0.

가설 검정 (2): 두 집단의 평균 차이 검정 (σ^2 unknown, but equal)

- 이번에는 모분산을 모르기 때문에 σ_1^2, σ_2^2를 쓸 수 없습니다.
- 모분산을 모르는 경우 표본분산으로 모분산을 대체해서 사용합니다.
- 단, $\sigma_1^2 = \sigma_2^2$임을 아는 경우, $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ 를 사용합니다. (pooled variance)
- 검정통계량:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad T \sim t_{n_1+n_2-2}$$

가설 검정 (2): 두 집단의 평균 차이 검정 (σ^2 unknown, and inequal)

- 이번에도 역시 표본분산으로 모분산을 대체해서 사용합니다.
- 단, 이번에는 단순히 σ_1^2을 s_1^2, σ_2^2을 s_2^2으로 대체합니다.
- 검정통계량:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \quad T \sim t_\nu, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_1)^2}{n_2 - 1}}$$

- 분산을 모르는 경우, 검정통계량이 t분포를 따르게 되지만 대표본인 경우 t분포가 표준정규분포와 비슷해지기 때문에 편의상 표준정규분포 (Z)를 사용하기도 합니다.

단순선형회귀분석

- 회귀분석 (Regression Analysis): 하나 또는 여러 개의 독립변수가 종속변수에 어떻게 영향을 미치는지 분석하는 기법

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim^{iid} N(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i, \qquad i = 1, 2, \cdots, n$$

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \text{ (matrix notation)}$$

- X: 독립 (Independent) 변수, 설명 (Explanatory) 변수
- Y: 종속 (Dependent) 변수, 반응 (Response) 변수
- ϵ: 오차 (error) 항
- 회귀분석의 목표는 회귀계수 (β_i) 를 알아내는 것입니다!
- 다양한 회귀분석 모형 가운데,
  - x와 β의 선형결합으로 y가 구해지고, (선형)
  - 독립변수가 1개인 (단순)

  모형을 단순선형회귀라고 합니다.

회귀분석의 표준 가정

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ 에서
회귀분석은 아래의 5가지 가정 사항을 전제로 합니다.

- 1) $E[\epsilon_i] = 0,\ \forall i$
- 2) $V[\epsilon_i] = \sigma^2,\ \forall i$
- 3) $Cov[\epsilon_i, \epsilon_j] = 0,\ \forall i \neq j$
- 4) Explanatory variables are non-stochastic.
- 5) Explanatory variables are not collinear.

Assumption (1): $E[\epsilon_i] = 0,\ \forall i$

- 기본적으로 이는 오차항의 평균은 0임을 의미합니다.
- 단, 이는 모든 오차항들을 전부 더했을 때 0이 된다는 의미가 아니라 하나하나의 데이터에서, 각각의 오차의 평균이 0이라는 의미입니다.
- $\sum_{i=1}^{n} \epsilon_i = 0$과 $\quad E[\epsilon_1] = 0, ..., E[\epsilon_n] = 0$의 차이입니다.

- 추가적 해석: 우리의 모형은 틀리지 않았다! (No Misspecification)
- $E[\epsilon_i] = 0 \leftrightarrow E[y_i] = \beta_0 + \beta_1 x_i$ 이므로,
  이는 우리의 모형이 누락한 설명변수는 없다, 모형이 완벽하다는 의미가 됩니다.

Assumption (2): $V[\epsilon_i] = \sigma^2, \forall i$

- 흔히 **등분산 가정** (Homoskedasticity) 이라고 부르는 것으로, 오차항의 분산은 모든 경우에서 일정하다는 의미를 가집니다.
- 하지만 이는 실제 상황에서는 잘 지켜지지 않는 가정입니다.
- 자료 자체의 절대적 수치에 따라서 분산은 달라지는 경우가 많기 때문입니다. (값이 커지면 분산이 커지는 것이 일반적)
- 이러한 현상은 시계열 자료에서 더욱 두드러진다고 합니다.
- 즉, 다른 가정의 경우에도 그러하나, 등분산 가정은 분석을 용이하게 만들기 위해 가정하는 것일 뿐입니다.
  이러한 가정사항이 현실적이라는 것은 아닙니다.
- 이 가정이 위배되는 상황을 이분산성 (Heteroskedasticity) 이라고 합니다.
- (1) 과 (2) 번 가정을 합치면 y_i 들이 iid라는 말이 됩니다.
  iid: **i**ndependently and **i**dentically **d**istributed.

Assumption (3): $Cov[\epsilon_i, \epsilon_j] = 0,\ \forall i \neq j$

- 서로 다른 관측치가 서로에게 영향을 주지 않는다는 의미가 된다.
- 이 역시 분석을 편하게 하기 위해 유지하는 가정사항이다.
- 시계열 자료에서 가정이 깨지는 경우가 특히 많고,
  그러한 경우를 자기상관성 (Autocorrelation) 이라고 한다.

Assumption (4): Explanatory variables are non-stochastic.

- 설명변수는 stochastic하지 않다는 가정 사항이다.
- 이는 설명변수 (X) 들에는 불확실성이 전혀 없다는 의미로,
  이 또한 상당히 비현실적인 가정 사항이다.
- 나이, 날짜 등의 변수는 Non-stochastic하다. 하지만 그 외의 다른
  설명변수들은 Randomness를 가지고 있는 경우가 더 많다고 한다.
- 경우에 따라 "설명변수가 stochastic하긴 하지만 오차항과는 독립"이라고
  가정하는 경우도 있지만,
  결과는 크게 달라지지 않는다고 한다.
- 이 가정이 깨지면 내생성 (Endogeneity) 이 있다고 한다.

Assumption (5): Explanatory variables are not collinear.

- **계산 불능**을 막기 위한 가정 사항.
- 만약 하나의 설명변수가 다른 설명변수와 perfect correlation을 가지면, 회귀계수의 추정 자체가 불가능해진다. (분산이 무한대로 발산하기 때문)
- 보다 관념적으로 보자면, 한 설명변수와 상당한 상관관계를 갖는 설명변수는 사실 모형에 추가하는 것이 그닥 효과적이지 않다.
- 추가되는 가치가 적거나 없는 변수는 포함할 필요가 없다고 볼 수 있다.
- 이 가정이 깨지는 경우 다중공선성 (Multicollienarity) 이 발생했다고 한다.

회귀 계수의 추정

- 회귀 계수를 추정하기 위해 여러 방식을 생각할 수 있습니다:
  - Maximum Likelihood Estimation (MLE) - R. Fisher (1912)
  - Method of Moments (MM) - K. Pearson (1895)
  - Least Square Estimation (LSE) - C. F. Gauss (1795), Legendre (1805)
- 각각의 장단점이 있지만, 회귀분석에서는 주로 LSE를 사용해 회귀계수를 추정합니다!
- 5가지 표준 가정이 전제된 상황에서는 LSE 방식으로 구한 회귀계수의 추정치가 **가장 좋은 선형 비편향 추정량** (BLUE, Best Linear Unbiased Estimator) 임이 알려져 있습니다.

LSE의 아이디어

- 우리는 $y = \beta_0 + \beta_1 x + \epsilon$의 계수를 알고 싶습니다.
- 하지만 실제 계수인 β_0, β_1은 Unknown constant이므로 접근이 불가하고, 대신 표본을 통해 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + e$의 추정식을 만들어, 회귀계수를 추정하고자 합니다.

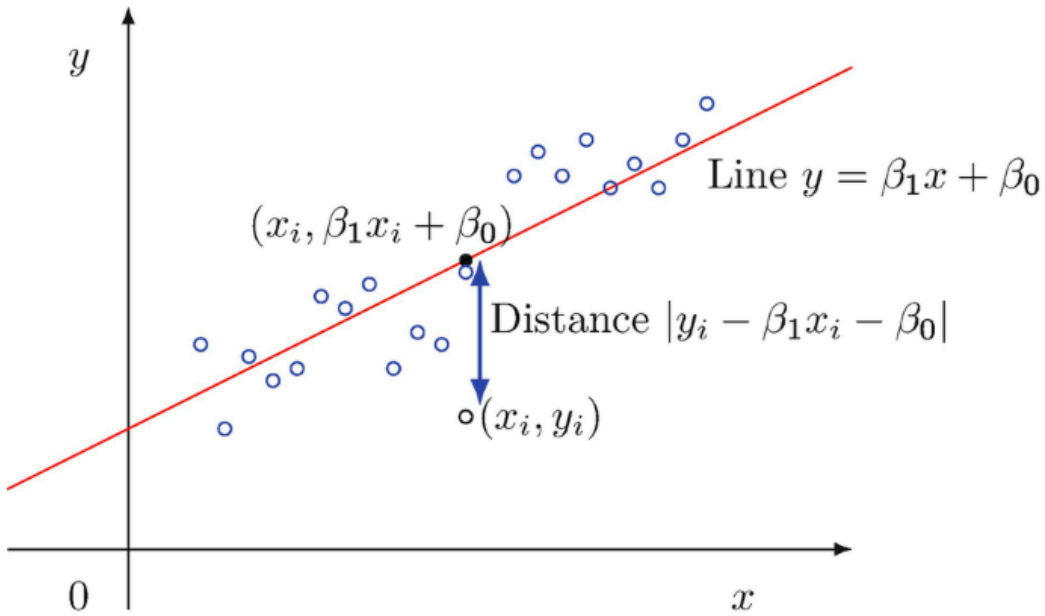

Figure 1: Illustration of LSE on Regression Model

LSE는 어떻게 계산되는가?

- 우리의 목표는 오차를 최소화하는 것입니다.
- 단순선형회귀모형에서, $\epsilon = y - \beta_0 - \beta_1 x$ 이므로
  우리는 $y - \beta_0 - \beta_1 x$ 를 최대한 줄여야 할 것입니다.
- LSE는 각각의 오차의 제곱을 모두 더해, 그 값을 최소화하는 방식으로
  회귀계수 추정치를 얻습니다.

$$\text{minimize } Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q}{\partial \beta_0} = 0 \text{ and } \frac{\partial Q}{\partial \beta_1} = 0 \text{을 만족시키는 } \beta_0, \beta_1 \text{을 찾는다!}$$

$$\therefore \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \equiv \frac{S_{xy}}{S_{xx}}$$

- 이 추정량을 OLS(Ordinary Least Square) 추정량이라고 합니다.

다음주 예고

- 회귀분석 2주차
  - BLUE (Gauss-Markov Theorem)
  - 회귀분석에서의 가설 검정
  - Bayesian Regression
- 분산 분석 (ANOVA, if possible)