DSL Seminar: 통입+통방 (7)

Kyung-han Kim

Data Science Lab

March, 2023

목차

- 회귀 분석 (2)
  - Gauss-Markov Theorem
  - BLUE (Best Linear Unbiased Estimator)
  - 회귀분석에서의 가설 검정
  - Bayesian Regression

지난 시간에 언급한 내용

- 회귀분석: 독립변수가 종속변수에 어떻게 영향을 미치는지 분석하는 기법

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

- 회귀분석의 목표는 회귀계수를 알아내는 것!
- 5가지 표준 가정
  - 오차항의 평균이 0이다
  - 오차항의 분산이 항상 일정하다
  - 오차항끼리의 공분산은 0이다
  - 독립변수는 Non-stochastic하다
  - 독립변수는 colinear하지 않다
- OLS

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

Gauss-Markov Theorem

- 왜 OLS 추정량을 써야 할까?
  그게 가장 좋은 추정량이기 때문이다.
- 좋은 추정량의 3가지 기준: 일치성, 비편향성, 효율성
- Gauss-Markov Theorem: 5가지 표준 가정 하에서,
  OLS 추정량이 BLUE이다!
  - BLUE: Best(Minimum-variance) Linear Unbiased Estimator
- Linear는 쉽게 확인 가능하므로, 비편향성(Unbiasedness)과 효율성(Efficiency, "Best"?)을 확인하도록 한다.
- 특히 $\hat{\beta}_0$ 보다도 $\hat{\beta}_1$ 을 확인한다.

Gauss-Markov Theorem (1): Unbiasedness

Prove that $E[\hat{\beta}_1] = \beta_1$.

$$\text{Let } c_i := \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}.$$

$$\text{Then } \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \sum c_i y_i = \sum c_i(\beta_0 + \beta_1 x_i + \epsilon_i).$$

$$\text{By the way, } \sum c_i = 0, \sum c_i x_i = 1.$$

$$\text{Thereefore, } \sum c_i(\beta_0 + \beta_1 x_i + \epsilon_i) = 0 + \beta_1 + 0 = \beta_1$$

$$\therefore E[\hat{\beta}_1] = \beta_1$$

Gauss-Markov Theorem (2): Efficiency

- 비편향성을 보이기 위해서는 $E[\hat{\beta}_1] = \beta_1$ 임을 보이기만 하면 됐지만, 효율성을 보이기 위해서는 $V[\hat{\beta}_1]$ 을 구하는 것뿐만 아니라, 그 분산이 다른 β_1 의 선형 추정량의 분산보다 반드시 작거나 같다는 것을 보여야 합니다.
- 이 과정에서 라그랑주 승수법이 필요합니다.
- 라그랑주 승수법: 최적화(Optimization)의 일종으로, 제약(Constraint)이 있는 상황에서 어떤 목적식의 최솟값을 찾아내는 방법.
(단, 함수가 미분 가능할 때 사용 가능함)
- 참고로,

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}} \text{ 입니다.}$$

라그랑주 승수법 (Lagrange Multiplier)

- f : 목적식 - 우리가 최솟값을 찾고 싶은 식
- $g = 0$: 등호 제약식

- $L = f - \lambda g$ 일 때 L의 모든 변수로 각각 편미분한 식이 모두 0이 되게 하는 λ에서 식이 최소가 된다.
- Ex] $x + 2y = 1$ 일 때 $f(x, y) = x^2 + y^2$ 의 최솟값은?

라그랑주 승수법 (Lagrange Multipler)의 예시

Ex] $x + 2y = 1$ 일 때 $f(x, y) = x^2 + y^2$ 의 최솟값은?

$$\text{min } x^2 + y^2 \text{ s.t. } x + 2y - 1 = 0$$

$$L(x, y) = f(x, y) - \lambda g(x, y) = x^2 + y^2 - \lambda(x + 2y - 1)$$

$$\frac{\partial L}{\partial x} = 2x - \lambda = 0$$

$$\frac{\partial L}{\partial y} = 2y - 2\lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = x + 2y - 1 = 0$$

$$\lambda = 2x = y, \; x + 2(2x) - 1 = 0, \; 5x = 1$$

$$\text{That is, } x = \frac{1}{5}, y = \frac{2}{5}.$$

$$\therefore \min(x^2 + y^2) = \frac{1}{25} + \frac{4}{25} = \frac{1}{5}.$$

라그랑주 승수법 (Lagrange Multiplier)의 예시 (결과 확인)

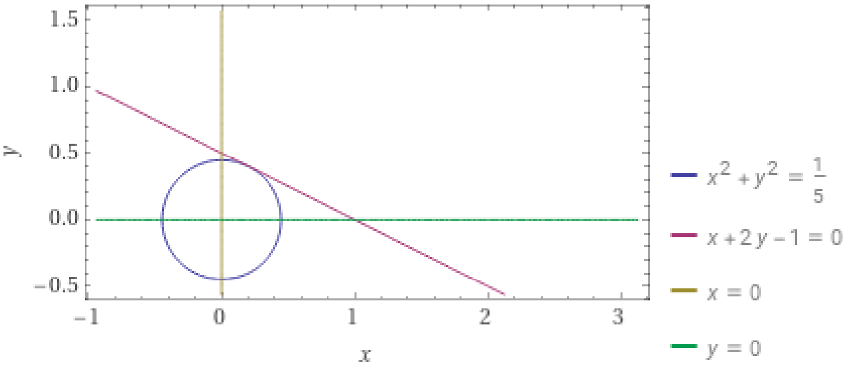

Figure 1: $x^2 + y^2 = 1$ and $x + 2y = 1$

Lagrange Multiplier with Gauss-Markov Theorem

- $\hat{\beta}_1$ 이 가장 효율적인 추정량임을 라그랑주 승수법으로 증명합니다.
- 임의의 선형 추정량 $\beta^* = \sum d_i y_i$ 로 놓을 수 있습니다.
- 이 추정량 또한 비편향 추정량이어야 하므로,
  $\sum d_i = 0, \sum d_i y_i = 1$ 을 만족해야 합니다.
- 우리는 선형 추정량 중 분산이 가장 작은 상황이 궁금하기 때문에 아래와 같이 목적식과 제약식을 설정합니다.

$$\text{minimize } V[\beta^*] \text{ s.t. } \sum d_i = 0 \text{ and } \sum d_i x_i = 1$$

- $V[\beta^*] = V[\sum d_i y_i] = V[\sum d_i \epsilon_i] = \sum d_i^2 V[\epsilon_i] = \sigma^2 \sum d_i^2$ 이고,
  σ^2 는 상수이므로 아래와 같이 정리됩니다.

$$\text{minimize } \sum d_i^2 \text{ s.t. } \sum d_i = 0 \text{ and } \sum d_i x_i = 1$$

Lagrange Multiplier with Gauss-Markov Theorem (cont'd)

- 따라서, $L = \sum d_i^2 - \lambda_1 \sum d_i - \lambda_2(\sum d_i x_i - 1)$ 입니다.
- 각각의 편미분 값이 모두 0이 되어야 하므로,

$$\frac{\partial L}{\partial d_k} = 2d_k - \lambda_1 - \lambda_2 x_k = 0, \;\; \forall k = 1, 2, ..., n$$

$$\frac{\partial L}{\partial \lambda_1} = \sum d_i = 0, \;\; \frac{\partial L}{\partial \lambda_2} = \sum d_i x_i - 1 = 0$$

- 첫 n 개의 식을 모두 더하면, $2\sum d_i - n\lambda_1 - \lambda_2 \sum x_i = 0$.
- 그런데 $\sum d_i = 0$ 이므로 $n\lambda_1 = -\lambda_2 \sum x_i, \quad \lambda_1 = -\lambda_2 \bar{x}$.
- 또한 $2\sum d_i x_i - \lambda_1 \sum x_i - \lambda_2 \sum x_i^2 = 0$ 이므로,

$$2 + \lambda_2 n\bar{x}^2 - \lambda_2 \sum x_i^2 = 0, \;\; \lambda_2(\sum x_i^2 - n\bar{x}^2) = 2$$

$$\therefore \lambda_2 = \frac{2}{\sum (x_i - \bar{x})^2}$$

- 앞서 구한 λ_1과 λ_2를 첫 식에 대입하면,

$$2d_i = -\lambda_2\bar{x} + \lambda_2 x_i = \lambda_2(x_i - \bar{x}) = \frac{2(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}.$$

$$\therefore d_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}.$$

- 그런데 이 값은 앞서 구했던

$$c_i := \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$$ 와 정확히 일치합니다.

- 따라서, OLS 추정량인 $\hat{\beta}_1 = \sum c_i x_i$가 가장 작은 분산을 가지는 선형 비편향 추정량임을 알 수 있습니다. $Q.E.D.$

회귀분석에서의 가설 검정

- 회귀계수 (β_1) 가 0이라면 해당 독립변수는 회귀 모형에서 사라지는 꼴이 됩니다.
- 따라서 회귀계수가 0인지 아닌지를 통계적으로 검정하는 절차가 반드시 필요합니다.
- 가설 검정을 위해서는
  - 귀무가설 (H_0) 과 대립가설 (H_1)
  - 검정 통계량과 그 확률분포

  를 알아야 합니다.

회귀계수의 유의성 검정

- $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$
- $\hat{\beta}_1 = \sum c_i y_i$이므로 $\hat{\beta}_1 \sim N$이다.
- 그렇다면
$$\frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{V[\hat{\beta}_1]} \sim Z$$
여야 하는데, 여기에서
$$\text{under } H_0, E[\hat{\beta}_1] = \beta_1 = 0, \quad \text{and } V[\beta_1] \text{ unknown.}$$
- 그러므로 아래와 같은 검정통계량을 사용한다.
$$T = \frac{\hat{\beta}_1}{SE[\hat{\beta}_1]} \sim t_{n-2}$$
- 여기서 2를 빼주는 이유는 추정하는 회귀 계수가 2개이기 때문입니다.

일반적인 경우에서의 유의성 검정

- 지금은 검정하고자 하는 회귀 계수가 1개이기 때문에 t검정을 활용할 수 있지만, 다중선형회귀에서는 검정 대상이 여러 개이기 때문에 그들이 모두 0이 아니라는 것을 한번에 보여야 합니다.
- 이런 경우에는 F검정을 사용하게 됩니다.
  자세한 검정 절차는 생략합니다.

Miscellaneous Topics

- Structural Break
  - Dummy Variable
- R^2
- Multicollinearity
  - Why Multicollinearity matters?
  - VIF
  - How to solve multicollinearity

MT (1): Structural Break

- 간혹 매우 중대한 사건 (코로나-19, 대공황, 우크라이나 전쟁 등) 이 발생해 기존의 트렌드를 어그러뜨리는 일이 발생할 수 있습니다.
- 이런 경우 해당 사건이 발생하기 전과 후에 차이가 있는지를 검정할 수 있습니다.
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$를 가정합니다.
- 가변수 (Dummy variable) 를 활용해 이를 검정할 수 있습니다.

MT (1): Structural Break - Dummy Variable

- 가변수(Dummy Variable): 본래 수치형이 아닌 변수를 0과 1의 형태로 변환해 임의로 만들어 낸 변수
- 우리의 예시에서는
  $D_i = 0$: Before the structural break
  $D_i = 1$: After the structural break 로 설정합니다.
- 그러면 모형을 아래와 같이 설정할 수 있습니다.

$$y_i = \beta_0 + \gamma_1 D_i + \beta_1 x_i + \gamma_2 D_i x_i + \epsilon_i$$

- 이 모형은

$$\text{Before: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\text{After: } y_i = (\beta_0 + \gamma_1) + (\beta_1 + \gamma_2) x_i + \epsilon_i$$

  를 한번에 표현했다고 볼 수 있습니다.
- 최종적으로 F 검정을 통해 $\gamma_1 = \gamma_2 = 0$인지를 검정합니다.

MT (2): R^2

- 설명력의 지표인 R^2는,

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

  로 정의됩니다.
- SST: y 전체의 분산 (필연적으로 존재하는 값)
- SSR: 회귀 모형이 분산을 설명하는 정도 (조절 가능한 값)
- SSR이 SST에 최대한 가까운 것이 바람직한 모형입니다.

MT (2): R^2 의 문제점

- 그런데 R^2 에는 몇 가지 결함이 있습니다.
  1. **독립변수가 많아지면 R^2 는 항상 커진다.**
  2. 강건 (Robust) 하지 못해 이상치 (outlier) 에 민감하다.
  3. 특정 상황에서는 0 보다 작아지거나, 1 보다 커지기도 한다.
  4. R^2 은 분포를 구할 수 없다. (non-linear transformation 필요?)
- 이 중 첫 번째 문제점을 해결하기 위해 adjusted-R^2 를 사용합니다.
- Adjusted-R^2:

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

- 독립변수의 개수인 k 가 커지면 뒤 항의 분모가 커지므로 전체 adjusted-R^2 값은 작아집니다. 즉, 회귀 모형의 독립 변수 개수에 일정 정도의 penalty를 부여한 형태가 됩니다.

- 하나의 독립변수가, 다른 독립변수들의 선형결합으로 표현될 때
- 독립변수가 collinear하지 않다는 가정이 깨지는 상황
- 완전다중공선성이 발생할 경우, $\hat{\beta}_1$ 의 분산을 계산할 수 없습니다.
- 상당한 정도의 다중공선성이 발생할 경우, $\hat{\beta}_1$ 의 분산이 커집니다.
- 즉, 회귀계수의 추정량 자체를 그다지 신뢰할 수 없게 됩니다.
- 단 다른 가정이 깨지는 상황과 다르게 다중공선성은 Gauss-Markov Theorem을 해치지는 않습니다. 다시 말해, OLS 추정량 자체는 여전히 최선의 선택입니다.

MT (3): VIF (Variance Inflation Factor)

- 분산이 커진다 = 다중공선성 의심

$$\hat{V}[\hat{\beta}_j] = \frac{s^2}{(n-1)\hat{V}(x_j)} \frac{1}{1-R_j^2}$$

$$VIF = \frac{1}{1-R_j^2}$$

- R_j^2: $x_j = \beta_0 + \beta_1 x_1 + ... + \beta_{j-1}x_{j-1} + \beta_{j+1}x_{j+1} + ... + \beta_n x_n$ 의 R^2
- R_j^2 값이 1에 가깝다
  = x_j 가 다른 독립변수로 설명되는 정도가 강하다 & $\hat{\beta}_j$ 의 분산이 크다
  = 다중공선성이 발생했다!
- VIF가 10 이상이면 다중공선성이 발생했다고 봅니다.
  단, VIF 자체에는 가설검정이 존재하지 않으므로 정확한 기각역을 설정하는 경우는 없습니다.
  - 다중공선성은 유무의 문제가 아니라 정도의 문제?
  - Gauss-Markov Theorem을 깨트리지 않는다

다음주 예고

- ANOVA