Finance Analytics
Chapter3. Linear Regression Model
Part 3. Inference in Model

권태연

한국외대 국제금융학과

선형회귀모형에 대한 추론

선형회귀모형의 가정

$$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + ... + B_k X_{ki} + \epsilon_i$$

$\star$ 위의 모형을 OLS로 추정(estimation)하는 과정에서는 선형회귀모형에 대한 가정이 필요하지 않다.

$\star$ 하지만

- OLS가 모회귀계수를 추정하는 좋은, 최적의 추정방법일까?
- OLS추정하는 값이 모수를 얼마나 잘 추정할까? 그 값이 틀릴 가능성은 얼마나 될까?

에 대한 문제에 있어서는 선형회귀모형에 대한 가정이 필요하다.

선형회귀모형의 가정

- 가정 1. 변수 Y와 X의 관계는 선형(Linear)이다.
- 가정 2. X는 확률변수가 아닌 주어진 상수값이다.
- 가정 3. X값이 주어져 있을 때, 오차항의 평균은 0이다.

$$E(u_i|X) = 0$$

- 가정 3'. X값이 주어져 있을 때, Y의 평균은 $B_0 + B_1X_{1i} + B_2X_{2i} + ... + B_kX_{ki}$이다.

$$E(Y_i|X) = B_0 + B_1X_{1i} + B_2X_{2i} + ... + B_kX_{ki}$$

선형회귀모형의 가정

- 가정 4. X값이 주어져 있을 때, 오차항의 분산은 σ^2로 모든 개체 i에 대해 동일하다.= 등분산(homoscedastoc)가정.

$$var(u_i|X) = \sigma^2$$

- 가정 4'. X값이 주어져 있을 때, Y의 분산은 σ^2로 모든 개체 i에 대해 동일하다.= 등분산(homoscedastoc)가정.

$$var(Y_i|X) = \sigma^2$$

- 가정 5. 서로 다른 개체간 오차항들은 상관되어 있지 않다. = 자기상관(autocorrelation)이 없다.

$$cor(u_i, u_j|X) = 0$$

- 가정 5'. 서로 다른 개체간 Y변수들은 상관되어 있지 않다. = 자기상관(autocorrelation)이 없다.

$$cor(Y_i, Y_j|X) = 0$$

선형회귀모형의 가정

- 가정 6. X변수들이 여러개 있을때, X변수들 사이에는 선형관계가 없다. = 다중공선성(Multicollinearity) 문제가 없음을 가정
- 가정 7. 모형 설정 오류가 없음
- 가정 8. 오차항은 정규분포를 따름을 가정

$$u_i|X \sim N(0, \sigma^2)$$

- 가정 8'. Y는 정규분포를 따름을 가정

$$Y_i|X \sim N(BX, \sigma^2)$$

BLUE (Best Linear Unbiased Estimator)

⋆ 하지만

- OLS가 모회귀계수를 추정하는 좋은, 최적의 추정방법일까? ✓

  ⋆ 가우스-마르코프 정리 **(GAUSS-MARKOV THEOREM)**
  : 모든 가정을 만족하면, OLS는 BLUE
  (Best Linear Unbiased Estimator)!

- OLS추정하는 값이 모수를 얼마나 잘 추정할까? 그 값이 틀릴 가능성은 얼마나 될까?
  : 확률의 개념, 확률변수이용

여러번 추정? : "Random" "Variable"에 대한 논의

- 관찰한 (하나의) 표본을 가지고 추정한 표본회귀모형의 $b_1, ..., b_k$,
- 즉 모회귀계수의 추정치들을 하나씩이지만..
- 다른 표본을 가지고 추정한다면?

  추정치가 달라질 수 있다.
- 모집단의 모든 개체를 이용하여 추정한 것이 아니기 때문에
  추정치는 실제 모수인 $B_1, ..B_k$값과 매우 유사한 값일 수도,

  혹은 그렇지 않을 수 있다.

  $\rightarrow$ 즉 추정량 $b_1, ..., b_k$ 역시 확률변수 **(random variable)**이다.

- 내가 현재의 표본을 가지고 추정한 추정치가 틀릴 가능성은 얼마나 되는가?

  이에 대한 계산은 추정량 $b_1, ..., b_k$의 확률분포에 근거한다.

  ==INFERENCE

OLS추정량의 확률분포

- 워드파일 coefficient is random variable 자료 참조
  그러나 현실에서는 위와 같은 반복실험 할수 없음.
- 만약 앞서 제시된 가정 1- 8을 모두 만족하면

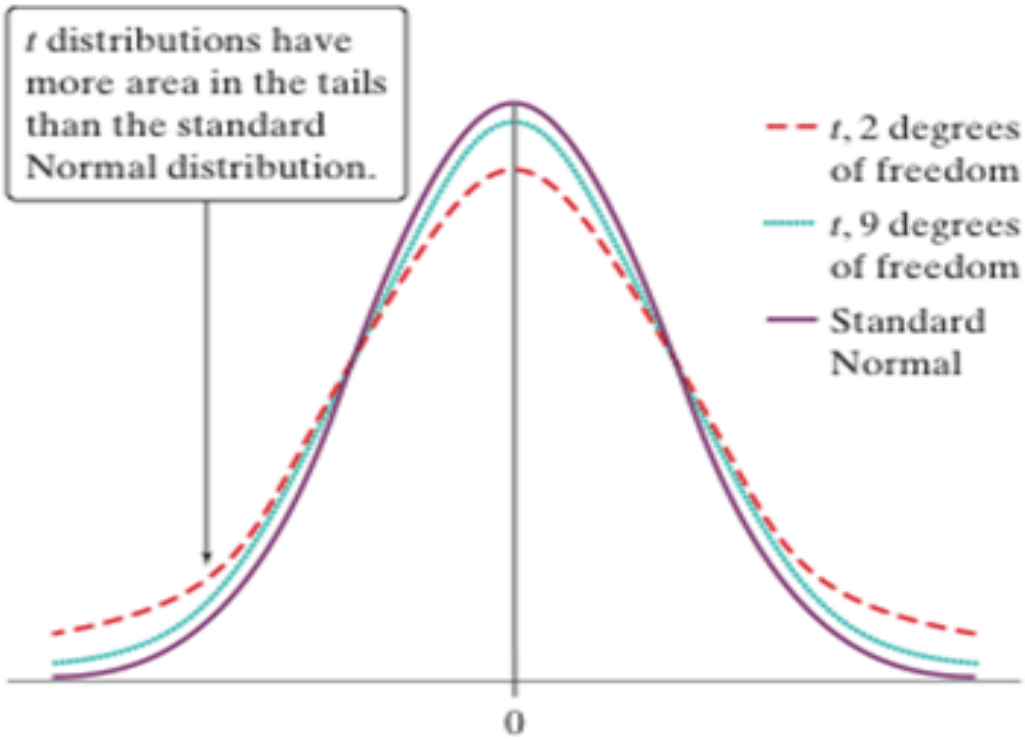
$$(b_k - B_k)/s.e.(b_k) \sim t - distribution(df = n - k)$$

  임이 알려져 있다.

  - n은 표본의 크기 (표수 내 개체의 수), k는 b의 갯수
  - $s.e.(b_k)$는 OLS 추정량 b_k의 표준오차(standard error)이다.
  - 추정량의 표준편차=표준오차

t-distribution

- t-분포는 정규분포와 매우 유사한 형태의 분포로 평균은 0이다.
- 정규분포보다 꼬리가 두꺼운 분포함수
- 정규분포의 모양이 μ와 σ에 의해 결정된다면
  t-분포는 "자유도(degree of freedom, df)"에 의해서 결정된다.



모회귀계수 진위에 대한 가설검정

모 회귀계수 B_k값이 '0'이라는 가설(Hypothesis)을 검정(Test)하고자 한다. (Why 0?)

- 가설 (Hypothesis) : 모수에 대한 예상, 주장 또는 단순한 추측
  - 귀무가설 (Null) : "=" 가설
  - 대립가설 (Alternative): "$\neq$", ">", "<", 연구자가 밝히고 싶은 가설
  - 귀무가설 (H0) : B_k=0
  - 대립가설 (Ha) : $B_k \neq 0$
- 가설검정(Test): 모집단에 대하여 주장하는 가설의 타당성을 표본을 이용하여 검토해보는 과정, $H0$ 기각, $H0$기각하지 못함으로 결정!
- 잘못된 의사결정 가능, 그 가능성(확률) 제시.

  $\rightarrow$ 유의성 검정 (Significant Test)

유의성 검정 (Significant Test)의 기본 원리

- 귀무가설이 사실이라고 받아들였을때 내가 현재 관찰한 b_k가 얼마나 관찰하기 힘든 값인지를 계산한다
- 관찰하기 매우 힘든 값이라면? 즉 관찰할 확률이 매우 작다면 사실이라고 받아들인 귀무가설 기각
- 관찰하기 매우 쉬운 값이라면? 즉 관찰할 확률이 매우 크다면 사실이라고 받아들인 귀무가설 채택

유의성 검정 (Significant Test)

$$\text{H0} : B_k = 0 \text{ vs } H_a : B_k \neq 0$$

- 표본을 이용하여 모집단의 모수에 대한 테스트
- b_k의 확률분포 이용

- (시간당 임금함수 예제)

1. 관찰된 한개의 표본을 이용하여 $b_k = 1.37$, $se(b_k) = 0.065$를 계산하였다.
2. 여러개의 표본을 이용하여 여러개의 b_k를 구할 수 있다면 모든 가정 (1-8)이 만족하고, $H0$가 사실이면

$$b_k / s.e.(b_k) \sim t - distribution(df = n - k)$$

관찰된 한개의 표본으로부터 구한 $b_k = 1.37$는 그 중 한 값

유의성 검정 (Significant Test)

3. 실제 $B_k = 0$이라면 관찰된 $b_k = 1.37$는 얼마나 관찰하기 힘든 값인가?

   $\rightarrow$ 이에 대한 확률을 구하여 이 확률이 매우 작다면 $H0$을 기각한다.

   - P-value: $H0$가 사실일 때, 현재 표본으로 부터 구한 $b_k = 1.37$ 값 혹은 그 보다 더 관찰하기 힘든 값으로 B_k가 추정될 확률
   - 유의수준 (significant level) : P-value가 얼마나 작으면 귀무가설을 기각할지에 대한 기준 확률, 일반적으로 0.05(5%) 혹은 0.01 (1%)이용

시간당 임금함수 예제에서 유의성 검정 (Significant Test)

- H0:
- Ha:
- 추정
- b_k의 확률분포는?
- n-k=
- p-value=
- 유의수준이 0.05일때 의사결정
- 결론

** R을 이용하여 위의 과정 실시하고 결과를 살펴보자.