# 제 4 장 단순회귀분석 (simple regression analysis)

## 1. 단순회귀분석의 기본가정

- 일반적으로 최소자승법에 의거 추론할 경우 다음과 같은 기본적인 가정이 필요하다.

가정 1) 회귀모형은 다음과같이 모수에 대해 선형(linear)인 모형이다: $Yi = \alpha + \beta Xi + \varepsilon_i$

→ 선형모형(linear model): 모수 $\alpha$ 와 $\beta$ 에 대하여 1 차 미분이 모수 $\alpha$ 와 $\beta$ 의 함수가 아니라 일정한 상수가 된다

가정 2) 전체표본에 있어서 독립변수 X 는 적어도 서로 다른 두 값을 가져야 한다.

가정 3) 오차항의 평균은 0 이다: $E(\varepsilon_i) = 0$

→ 오차항들이 회귀선을 중심으로 대칭적으로 분포되어있어야 한다

→ 이 가정은 최소자승법에 의한 추정과정상의 계산을 간편하게 하기 위해 만들어진 편의상의 조건식

가정 4) 독립변수 X 는 비확률(nonstochastic) 변수이다: $E[\varepsilon_i Xi] = Xi\, E[\varepsilon_i] = 0$

→ 독립변수는 비확률적 변수로 일반적인 상수(constant)로 취급한다.

가정 5)오차항은 모든 관찰치에 대해 $\sigma^2$ 의 일정한 분산을 갖는다: $var(\varepsilon_i) =$

$E(\varepsilon_i^{\ 2})=\sigma^2$

→ 관측시점에 관계없이(i 값에 관계없이) 오차항들의 분포는 동일한 분산을 갖는다

→ 각독립변수 X 의 값에 있어서 종속변수 Y 가 그 평균을 중심으로 분포되어 있는 정도가 같다

→ "동분산(homoskedasticity)" 가정이라 한다

↔ 이분산(heteroskedasticity): 시계열 데이터(time series data)나 횡단면 데이터 (cross-section data)경우 분석대상의 기간이나 크기따라 분산의 크기가 달라지는 경우.

가정 6) 서로 다른 관찰치간의 오차항은 상관이 없다: $Cov(\varepsilon_i, \varepsilon_j) = E[\varepsilon_i, \varepsilon_j] = 0$ $(i \neq j)$

→ 오차항끼리는 서로 1 차독립적인임을 의미한다.

- 가정 7): 오차항이 정규분포를 따르며, $\varepsilon_i \sim N(0,\sigma^2)$

2. 최소자승법(Ordinary Least Squares Method: OLS)에 의한 회귀모형의 추정

- 모집단 회귀모형을 추정하거나, 또는 모집단회귀선의 특성치인 모수 α 와 β 에 대한 가장 좋은 추정치를 구하는 것은 결국 어떻게 표본회귀선을 도출하는 것이 최선인가 하는 문제로 귀착된다.

➔ 적합도가 가장 큰 표본회귀선이란 오차항($e_i$ = Yi - Y^i)의 합이 가장 작은 회귀선이다

- 이러한 최적의 표본회귀선을 구하는 방법중 가장 많이 사용되는 추정법이 최소자승법(OLS: ordinary least squares method)이다

   ➔ 오차항을 자승한 값들의 합이 최소기 되도록 하는 회귀선을 구하는 방법:

   $\mathrm{Min}\sum_{i=1,n} e_i^2 = \mathrm{Min}\sum_{i=1,n} (Y_i - \hat{Y}_i)^2 = \mathrm{Min}\sum_{i=1,n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$

- 최소자승 추정량의 도출

   ➔ 식 $\mathrm{Min}\ L = \mathrm{Min}\sum_{i=1,n} e_i^2 = \mathrm{Min}\sum_{i=1,n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$ 에 나타난 오차의 최소자승값을 구하려면 이식을 α^ 와 β^에 대해 1 차미분한 값이 0 이 되어야한다:

   $(\partial L/\partial \hat{\alpha}) = -2\sum (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$

   $(\partial L/\partial \hat{\beta}) = -2\sum X_i (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$

   이를 다시 α^ 와 β^에 대하여 정리를 하면,

   $\hat{\beta} = (n\sum X_iY_i - \sum X_i \sum Y_i)/ [n\sum X_i^2 - (\sum X_i)^2] = \sum(X_i - X)(Y_i - Y)/ \sum(X_i - X)^2$

   $= (\sum X_iY_i - n XY)/ (\sum X_i^2 - nX^2) = (\sum x_i y_i)/ \sum x_i^2$

   ⇨ X (표본평균)= $(1/n) \sum X_i$, Y (표본평균)= $(1/n)\sum Y_i$,

⇨ $x_i$ (편차)= Xi – X, $y_i$ (편차) = Yi - Y

α^ = Y - β^ X

➔ α^ 과 β^ 은 알려지지 않은 모집단의 상수 α 과 β 의 최소자승추정량 (least squares estimator: LSE)이며, 이들은 확률함수형태의 확률변수이다.

## 3. 최소자승추정량의 통계적 특성

- 최소자승법에 의해 산출된 최소자승추정량 (α^ 과 β^)이 확률변수이라면, 이들의 통계적 특성 (분포, 평균, 분산)을 파악할 필요가 있다.

- 최소자승추정량은 기본가정들이 충족되면 통계적으로 바람직한 특성, 불편성(unbiasedness), 효율성(efficiency), 선형성(linearity) 및 일관성(consitency)을 갖게 된다

1) 불편성(unbiasedness)

- 최소승차법에 의해 도출된 추정량 (least squares estimator: LSE)은 불편성(unbiasedness)를 갖게된다: E[α^ ] = α, E[ β^] = β

  → 추정량 α^ 와 β^의 평균값은 모수인 특성치 α, β 의 실제값과 일치하게되

는 특성을 갖는다

→ 하나의 표본으로부터 구한 추정량은 평균적으로 모집단의 회귀계수와 같다.

2) 효율성

- LSE 는 효율성을 갖는다: 모든 가능한 불편추정량(unbiased estimators)중에서 최소의 분산을 갖는다
- LSE 는 다음의 분산과 공분산을 갖는다

→ $Var(\alpha\hat{}) = \sigma^2[\sum_{i=1,n} X_i^2 / n\sum_{i=1,n} (X_i - X)^2]$,

→ $Var(\beta\hat{}) = [\sigma^2 / n\sum_{i=1,n} (X_i - X)^2]$,

→ $Cov(\alpha\hat{},\beta\hat{}) = \sigma^2[- X / n\sum_{i=1,n} (X_i - X)^2]$

➔ a) 오차항의 분산값($\sigma^2$)이 커질수록 LSE 의 분산은 커지고 LSE 는 덜 정확한 추정치를 낳게된다.

b) 독립변수 X 의 값이 넓게 퍼져있을수록($\sum_{i=1,n} (X_i - X)^2$ 이 클수록) LSE 의 분산은 작아지고 독립변수의 변화에 의한 종속변수의 변화를 상대적으로 잘 설명할수있게된다

c) 표본의 수(n)가 증가할수록 $\sum_{i=1,n} (X_i - X)^2$ 의 값이 증가하게 되어 LSE 의

분산과 공분산이 작아진다. 이는 표본의수가 증가할수록 전체 모집단의 수에 근접하게 되어 모집단의 모수에 대한 정보를 더 정확하게 구할 수 있기 때문이다.

d) 공분산은 독립변수(X)의 평균(X)과 반대의 부호를 가진다.

3) 선형성(linearity)

- LSE 를 정의하면,

$\beta^\wedge = (\sum x_i y_i)/ \sum x_i^2 = \sum C_i y_i)$, [$Ci = \sum x_i / \sum x_i^2$]

$\alpha^\wedge = Y - \beta^\wedge X = Y - X\sum C_i y_i$,

→ 두개의 LSE 모두 $y_i$ 의 1 차함수관계인 선형결합형태를 가진 선형추정량(linear estimator)이라 한다

4) 가우스-마르코프 정리 (Gauss-Markov Theorem)

- 최소자승추정량(LSE)은 최량선형불편추정량(Best Linear Unbiased Estimator :BLUE)이다

  → 최소자승추정량(LSE)은 모든 선형이고 불편의인 추정량들 가운데 최량(best)인 추정량이다: 최량(best)는 최소의분산을 갖는 것을 의미한다

- 모집단 회귀계수 α, β 의 추정량(estimator)으로서 선형(linear)이고 불편

(unbiased)인 추정량중에서는 최소자승추정량(LSE)이 분산이 가장 작은 최상의 추정량임을 말한다.

## 4. 최소자승 추정량(LSE)의 확률분포

- 가정 7): $\varepsilon_i \sim N(0,\sigma^2)$은 LSE 의 확률본포를 도출하기위해 필요한 가정

→ (가정 2)와 함께 종속변수(Y)가 정규분포를 가짐 의미: $Y_i \sim N(\alpha + \beta Xi, \sigma^2)$

1) 오차항의 분산($\sigma^2$) 값을 알 때의 최소자승 추정량(LSE)의 확률분포

- 가정(1)~(7)하에서 최소자승 추정량(LSE)은 다음과 같은 정규분포를 가진다.

→ $\beta^\wedge \sim N(\beta,\sigma_\beta{}^2)$ , $\alpha^\wedge \sim N(\alpha,\sigma_\alpha{}^2)$,

단, $\sigma_\alpha{}^2 = \sigma^2[\sum_{i=1,n} X_i^2 / n\sum_{i=1,n} (X_i - X)^2]$, $\sigma_\beta{}^2 = [\sigma^2 / n\sum_{i=1,n} (X_i - X)^2]$,

➔ 일치추정량 (consistent estimator): 표본크기가 무한히 커질 경우에 추정량의 분산값이 "0'에 근접함에 따라 추정량($\alpha^\wedge$ , $\beta^\wedge$ )이 실제 모수($\alpha, \beta$)에 수렴하게 된다 (일치성)

- 이를 표준화하면, $[(\beta^\wedge - \beta)/\sigma_\beta] \sim N(0,1)$, $[(\alpha^\wedge - \alpha)/\sigma_\alpha] \sim N(0,1)$

→ 이를 토대로 모수 $\alpha, \beta$ 의 검증을 위해 표준정규분포를 이용할 수 있으나 일반적으로 오차항의 분산인$\sigma^2$ 값이 알려지지 않아 검증을 하지 못한다

→ 이를 위해 $\sigma^2$ 값을 표본자료로부터 표본분산을 이용하여 추정해야 한다.

2) $\sigma^2$ 과 $\sigma_\alpha{}^2, \sigma_\beta{}^2$ 의 추정량

- 표본자료와 LSE 인 β^ 와 α^ 을 이용하여, 잔차항 $e_i$ 을 도출할 수 있으므로 이를 토대로 오차항의 분산 $\sigma^2$ 의 추정량 $s^2$ 을 구할 수 있다

→ $s^2 = (1/n-2) \sum_{i=1,n} e_i^2$, s 는 회귀식의 표준오차(standard error of regression)

→ (n-2)로 나누는 이유는 n 개의 표본자료로부터 2 개의 회귀계수(β 와 α)을 추정하는데 따르는 자유도의 감소를 고려한 것이다.

- 추정량 $s^2$ 을 이용한 LSE 의 분산 추정치는,

→ $s^2_\alpha = s^2[\ \sum_{i=1,n} X_i^2 / n\sum_{i=1,n} (X_i - X)^2\ ]$.

→ $s^2_\beta = [\ s^2 / \sum_{i=1,n} (X_i - X)^2\ ]$.

→ 최소자승 추정량(LSE)의 확률분포: (n-2)의 자유도를 갖는 t-분포.

: $(\alpha\hat{} - \alpha)/ s_\alpha \sim t(n-2)$, $(\beta\hat{} - \beta)/ s_\beta \sim t(n-2)$

<u>5. 회귀계수에 대한 통계적 추론</u>

- 통계적 추론의 목적

a) 표본회귀에 대한 통계적 유의성 검정

b) 실증분석에서 회귀계수에 대한 통계적 추론 또는 가설검증은 모집단 회귀계수에 대해 가지는 막연한 추측이나 가설이 맞는지를 표본으로부터 구한 결과를 토대로 판단 할 수 있게 해준다.

1) 제 1 종 과오 (type I error), 제 2 종 과오(type II error)및 유의수준(significance level)

- 제 1 종 과오 (type I error): 귀무가설이 맞는데도 이를 기각하는 오류
- 제 2 종 과오(type II error): 귀무가설이 틀리는데도 이를 기각하지 못하는 오류
- 유의수준(significance level): 제 1 종 과오를 저지를 확률

  → 일반적으로 1%(0.01),5%(0.05) ,10%(0.1)를 많이 채택한다

2) 귀무가설(null hypothesis)과 대립가설(alternative hypothesis)

- 귀무가설 ($H_0$): 모집단 회귀계수에 대해 특정값을 부여한 가설

- 대립가설 ($H_1$): 귀무가설에 대응하는 가설

예) $H_0 : \beta\hat{} = \beta_0$ vs. $H_1 : \beta\hat{} \neq \beta_0$

3) 가설검정의 절차와 해석

a) 가설을 설정: $H_0 : \beta\hat{} = \beta_0$ vs. $H_1 : \beta\hat{} \neq \beta_0$

b) 통계량의 표본분포를 파악: $(\beta^{\wedge} - \beta_0)/ s_\beta \sim t(n-2)$

c) 귀무가설하에서 검정통계량을 계산: $(\beta^{\wedge} - \beta)/ s_\beta \sim t(n-2)$ 분포

d) 유의수준하에서 임계치 (critical value) 계산: $t(n-2; \alpha/2)$

e) 기각역(rejection region)을 산정

f) 판정

4) 단측검정(one-side test)과 양측검정(two-side test)

- 귀무가설, $H_0 : \beta\hat{} = \beta_0$ 에 대하여 대립가설은 설정방향에 따라 다를 수 가 있다

  → $H_1 : \beta\hat{} \neq \beta_0$ , $H_1 : \beta\hat{} > \beta_0$ , $H_1 : \beta\hat{} < \beta_0$

- 대립가설에서 모수값의 방향에 따라 양측검정 또는 단측검정이 결정된다.

  → 단측검정: $H_0$ : 대립가설하에서 무수값의 부등호가 한쪽방향일 경우 사용

한다

⇨ $H_0 : \hat{\beta} \geq \beta_0$ vs. $H_1 : \hat{\beta} < \beta_0$ : 검정통계량, $t = (\hat{\beta} - \beta)/ s_{\beta} >$ 임계치 $(-t\ (\alpha))$일때 귀무가설을 기각하지 못한다

⇨ $H_0 : \hat{\beta} \leq \beta_0$ vs. $H_1 : \hat{\beta} > \beta_0$ : 검정통계량, $t = (\hat{\beta} - \beta)/ s_{\beta} <$ 임계치 $(t\ (\alpha))$일때 귀무가설을 기각하지 못한다

→ 양측검정: $H_0 : \hat{\beta} = \beta_0$ vs. $H_1 : \hat{\beta} \neq \beta_0$

⇨ $|t=(\hat{\beta} - \beta)/ s_{\beta}| \geq t(n-2; \alpha/2)$이면, 귀무가설 기각

⇨ $|t=(\hat{\beta} - \beta)/ s_{\beta}| < t(n-2; \alpha/2)$이면, 귀무가설 채택(기각하지 못한다)

5) 유의성 검증

- $H_0 : \hat{\beta} = 0$ vs. $H_1 : \hat{\beta} \neq 0$ 의 가설검정은 추정량의 유의성 검증 (significance test)에 중요한 의미를 부여한다.

→ 귀무가설이 기각되면, 독립변수 X 가 종속변수 Y 에 유의적인 (significant) 영향을 주지 못함: 추정량이 통계적으로 유의하지 못하다(statistically insignificant),

→ 귀무가설이 기각되지 못하면 독립변수 X 가 종속변수 Y 에 유의적인 (significant) 영향을 줌: 추정량이 통계적으로 유의하다(statistically significant)

- 귀무가설하에서의 통계치는, $t = (\hat{\beta} - \beta_0)/ s_{\beta} = \hat{\beta} / s_{\beta}$ $(\beta_0 = 0)$: t-통계치(t-ratio)

→ 2-t 으뜸법칙: 통계치의 표본분포는 t-분포를 가지나 표본수가 충분히 클 때 (n→∞) 정규분포에 근접한다

- ⇨ 5%의 유의수준에서 임계치는 1,96, 대략 2 가 되면, 만일 |t| < 2 (-2< t <2)이면 귀무가설이 기각되지 못하여 변수 X 가 변수 Y 에 유의적인 영향을 주지 못함을 의미한다.

- 어떠한 독립변수가 종속변수에 중요한 변수인지 아닌지를 판단하기위해 t-통계치와 2 의 값을 비교하여 유의성을 검증하는 것이 필요하다

6) 유의확률(P-value) 에 의한 가설검증

- 유의성 검증을 위한 가설검정에 있어서, 유의확률(P-value)가 사용되기도 한다

→ 이경우 일정한 t-통계치에서 유의수준(α)를 최소한 얼마로 하여야 귀무가설을 기각할 수 있는가하는 확률이 필요하다

→ 주어진 귀무가설을 기각할 최소수준의 유의수준,또는 t-분포에서 t-통계치의 절대값보다 큰값을 가질 확률을 유의확률(P-value)라 한다:

P[(t(n-2) > |t-통계치|)]

- ⇨ 유의확률 > 유의수준: 귀무가설 채택
- ⇨ 유의확률 < 유의수준: 귀무가설 기각

## 6. 회귀분석 모형의 적합도(goodness of fit) 평가: 결정계수 (coefficient of determination)

- 적합도(goodness of fit) 란 도출된 표본회귀선이 각 관측점들을 얼마나 잘나타낼수 있는지, 또는 주어진 자료로부터 독립변수가 종속변수를 얼마나 잘 설명하는지의 여부를 말한다
- 결정계수(coefficient of determination): 이러한 회귀모형의 설명력, 회귀선의 적합도를 나타내는 지표 및 기준
- Y 의 변동에 대한 설명: $Yi = α^ + β^Xi + e_i = Y_i^{\wedge} + e_i$

→ $(Y_i - Y) = (Y_i\wedge - Y) + e_i$, (평균을 기준으로)

→ $\sum_{i=1,n} (Y_i - Y)^2 = \sum_{i=1,n} (Y_i\wedge - Y)^2 + \sum_{i=1,n} e_i^2$,

- ⇨ TSS(Total Sum of Squares): $\sum_{i=1,n} (Yi - Y)^2$: 종속변수 Y 가 그 평균인 Y 로부터 얼마나 변동하는가를 나타태는 Y 의 전체변동
- ⇨ ESS(Explained Sum of Squares): $\sum_{i=1,n} (Y\wedge - Y)^2$ :Y 중에서 X 에의해 설명되는 부분의 변동
- ⇨ RSS(Residual Sum of Squares): $\sum_{i=1,n} e_i^2$ :Y 중에서 X 에 의해 설명되지 않는 잔차의 변동
- ⇨ TSS = ESS + RSS

→ 결정계수, $R^2$ = ESS/TSS = 1- (RSS/TSS): $Y_i$ 의 변동 중 얼마만큼의 부분이 설

명되고 있는가를 보여준다

⇨ $0 \le R^2 \le 1$: $R^2$ 이 1 에 가까워질수록 주어진 모형이 자료에 더 적합하다고 판정된다

⇨ $R^2$ 는 단순히 모형의 적합도에 대한 수치적 척도를 나타낼 뿐 질적인 척도를 나타내지 못한다: $R^2$ 를 극대화 하는 것에 초점을 두는 것은 옳지 못함.

## 7. 예측(forecast, prediction)

- 계량경제학의 중요한 목적중 하나가 미래의 특정시점에서 설명변수(독립변수)의 값을 알 때 종속변수의 값을 예측하는 것이다.

  → 예측: 두 변수가 인과관계에 의한 만족할만한 함수관계(선형, 1 차함수관계)를 갖고 있는 것으로 판정될 때 표본회귀선에 의거하여 미래의 종속변수 값을 판정하는 것.

  → 도출된 추정결과식 (회귀분석식)이 적합도나 통계적 유의검정에서 바람직한 것으로 평가되면 이를 바탕으로 미래예측을 할 수 있게 된다

- 예측대상 시기에 따라 사후적 예측(ex-post forecast)과 사전적 예측(ex-ante forecast)으로 구분

  a) 사후적 예측:

→ 예측하고자 하는 기간이 회귀분석 대상기간 이후에 속하더라도 이미 과거에 속한 기간이라면 비록 설명변수와 종속변수의 값을 알고 있다 하더라도 설명변수 값을 표본회귀선에 대입하여 종속변수를 예측하는 것

→ 이러한 사후적 예측과정은 예측자체보다는 추정된 표본회귀선에 의해 예측된 종속변수 값이 실제 실현된 종속변수 값과 얼마나 근접된 값으로 나타나는지 확인함을 써 표본회귀선의 설명력을 평가하는 또 하나의 방법으로 주로 이용된다

b) 사전적 예측: 현시점이후에 속하는 미래를 대상으로 하되 가정된 설명변수 값을 가지고 종속변수 값을 추정하는 것

- 미래에 대한 예측을 시행할 때 설명변수의 값이 알려져 있는지의 여부가 중요한 역할을 한다: 무조건부 예측(unconditional forecast)와 조건부 예측(conditional forecast)

→ 무조건부 예측: 설명변수가 알려진 상태에서의 예측으로 사후적 예측이 주로 이에 속한다

→ 조건부 예측: 설명변수가 미지인 상태에서의 예측으로 사전적 예측이 이에 속한다

- 현실적으로 주로 사용하는 예측방법은 사전적 예측에 속하나 설명변수 값이 알려져 있다는 가정하에 무조건부 예측을 주로 사용한다.

→ 무조건부 예측은 크게 점예측(point forecast)과 구간예측(interval forecast)의 두가지가 있다

a) 점예측:

- ⇨ 단순히 표본관측점들을 근거로 최소자승법에 의해 도출된 표본회귀선을 이용하여 가상적인 미래의 설명변수값 $X_f$ (단, 이 X 값은 표본회귀선 도출을 위한 회귀분석에사용되지 않은 별도의 X 값) 에 대응하는 종속변수값을 하나의 수치로 표현하는 방법
- ⇨ 추정된 회귀선이 $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ 이라하면, 새로운 관찰치 f 의 설명변수값, $X_f$ 을 알 때 $Y_f$ 의 예측치 (predicted value)는 $\hat{Y}_f = \hat{\alpha} + \hat{\beta}X_f$
- ⇨ 예측오차(prediction error), $e_f = \hat{Y}_f - Y_f$ 는 다음과 같은 특성을 갖는다

    i) $E(e_f) = 0$

    ii) $Var(e_f) = \sigma^2[1+(1/n)+\{(X_f - X)^2 / \sum (X_i - X)^2\}]$

    iii) 만일 오차항 분산값($\sigma^2$)이 알려져 있지 않을 경우에는

    추정치 $s^2$ 를 이용한다:

    $$Var(\hat{e}_f) = s_f^2 = s^2[1+(1/n)+\{(X_f - X)^2 / \sum (X_i - X)^2\}]$$

⇨ $e_f / s_f^2 = (\hat{Y}_f - Y_f) / s_f^2 \sim t(n-2)$

b) 구간 예측:

⇨ 점예측의 경우 아무리 가까운 미래에 대한 예측이라 하더라도 점예측치와 실제 종속변수값이 일치할 가능성은 희박하므로 대개 이러한 점예측치를 중심으로 일정한 구간을 설정하여 제시하는 구간예측을 동시에 실시한다

⇨ 실제 $Y_f$에 대한 100(1-α)% 신뢰구간예측은, $[Y_f \pm t(\alpha/2)\ s_f]$

→ 예측에 있어서 관찰 표본의 수(n)이 클수록 예측오차의 분산이 작아지고 일정한 신뢰수준 하에서 신뢰구간이 좁아지게 되어 정확한 예측치를 얻을 수 있다.

→ 예측하고자 하는 $X_f$ 값이 평균값(X)과 멀어질수록 예측오차는 커지고 예측구간 폭은 점차 넓어지므로 예측의 신뢰성이 낮아진다.