제13장 단순회귀분석

- 회귀분석과 상관분석
- 단순회귀분석
- 최소자승법
- 표본회귀선의 적합도 검정
- 표본회귀선의 유의성 검정

Ⅰ. 회귀분석과 상관분석

➢ 회귀분석(regression analysis)

두 변수(종속변수, 독립변수) 사이의 함수적 관계를 기술하는 수학적 방정식을 구하는데 사용된다. 이 식은 독립변수의 값이 주어질 때 종속변수의 값을 추정하거나 예측하는데 사용됨.

➢ 상관분석(correlation analysis)

두 변수 사이의 선형관계의 강도와 방향을 요약하는 수치를 구하는데 사용된다. 즉 상관분석은 두 변수가 얼마나 밀접하게 연관되어 있는가 하는 정도를 나타냄.

➢ 회귀분석과 상관분석은 두 변수 사이의 인과관계를 밝히는 것이 아니라, 두 변수가 서로 어떻게 , 어느 정도 관련되어 있는가를 나타낸다.

II. 단순회귀분석

➢ 종속변수 Y가 독립변수 X와 오차항(error term)에 어떻게 관련되어 있는가를 나타내는 방정식을 회귀모형(regression model)이라고 한다.

단순회귀모형

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

y_i : i번째 관측치에 대한 종속변수의 값

x_i : 이미 알려진 독립변수의 i번째 값

α, β : 회귀계수

ε_i : i번째 관측치에 대한 오차항

표본회귀선

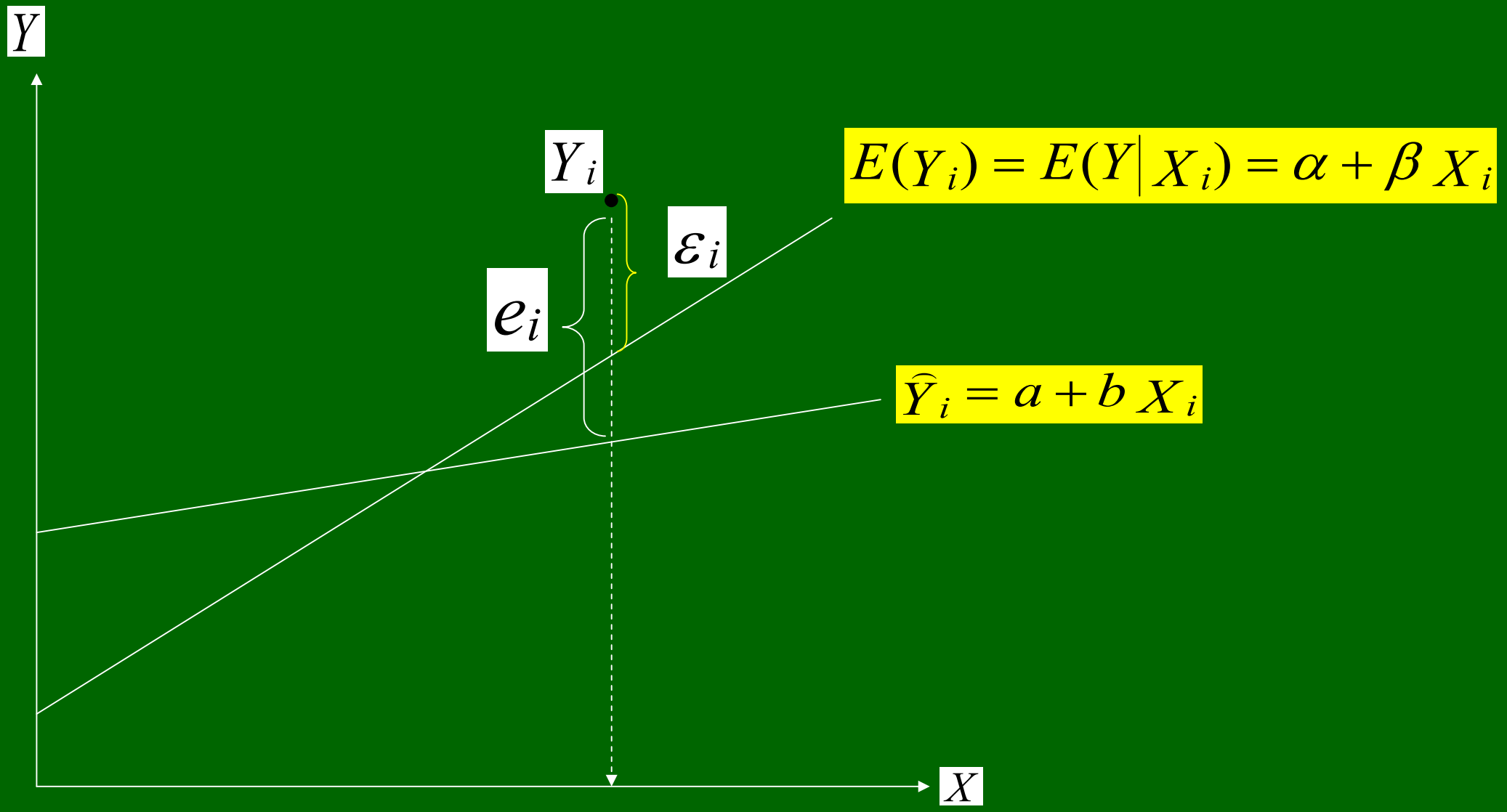
$$\hat{Y}_i = a + b X_i$$

표본회귀모형

$$Y_i = a + b X_i + e_i$$

$$e_i = Y_i - \hat{Y}_i$$

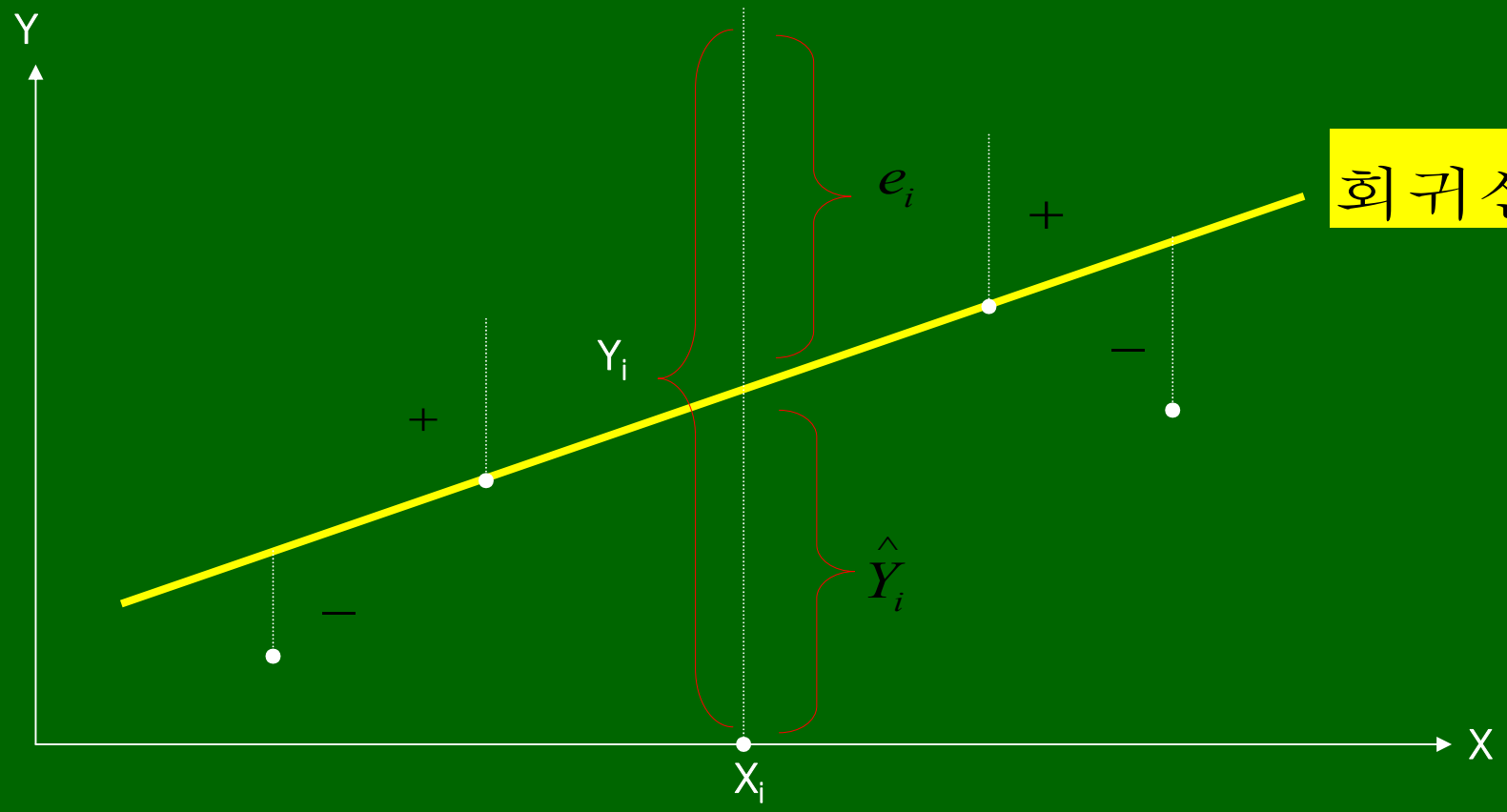
II. 오차항 ε_i 와 잔차 e_i 의 관계



➢ 오차항에 대한 가정

- 오차의 확률분포의 평균은 0 이다.
- 오차는 근사하게 정규분포를 따른다.
- 오차의 확률분포의 분산 σ_e^2 은 독립변수 X의 모든 값에 대해 동일하다
- 오차는 서로 독립적이다.

III. 최소자승법



회귀선 $\hat{Y} = a + bX$

$$최소\sum e_i^2 = 최소\sum(Y_i - \hat{Y}_i)^2 = 최소\sum(Y_i - a - bX_i)^2$$

$$\therefore \quad \sum Y_i = na + b\sum X_i$$

$$\sum X_i Y_i = a\sum X_i + b\sum X_i^2$$

표본회귀선의 회귀계수

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sum X_i^2 - n\overline{X}^2}$$

$$a = \overline{Y} - b\overline{X}$$

Ⅳ. 표본회귀선의 적합도 검정

➢적합도검정(goodness-of-fit test)
회귀모형 자체에 대하여 회귀선이 모든 관측치들을 적합하도록 도출되었는지 밝히는 것이다.

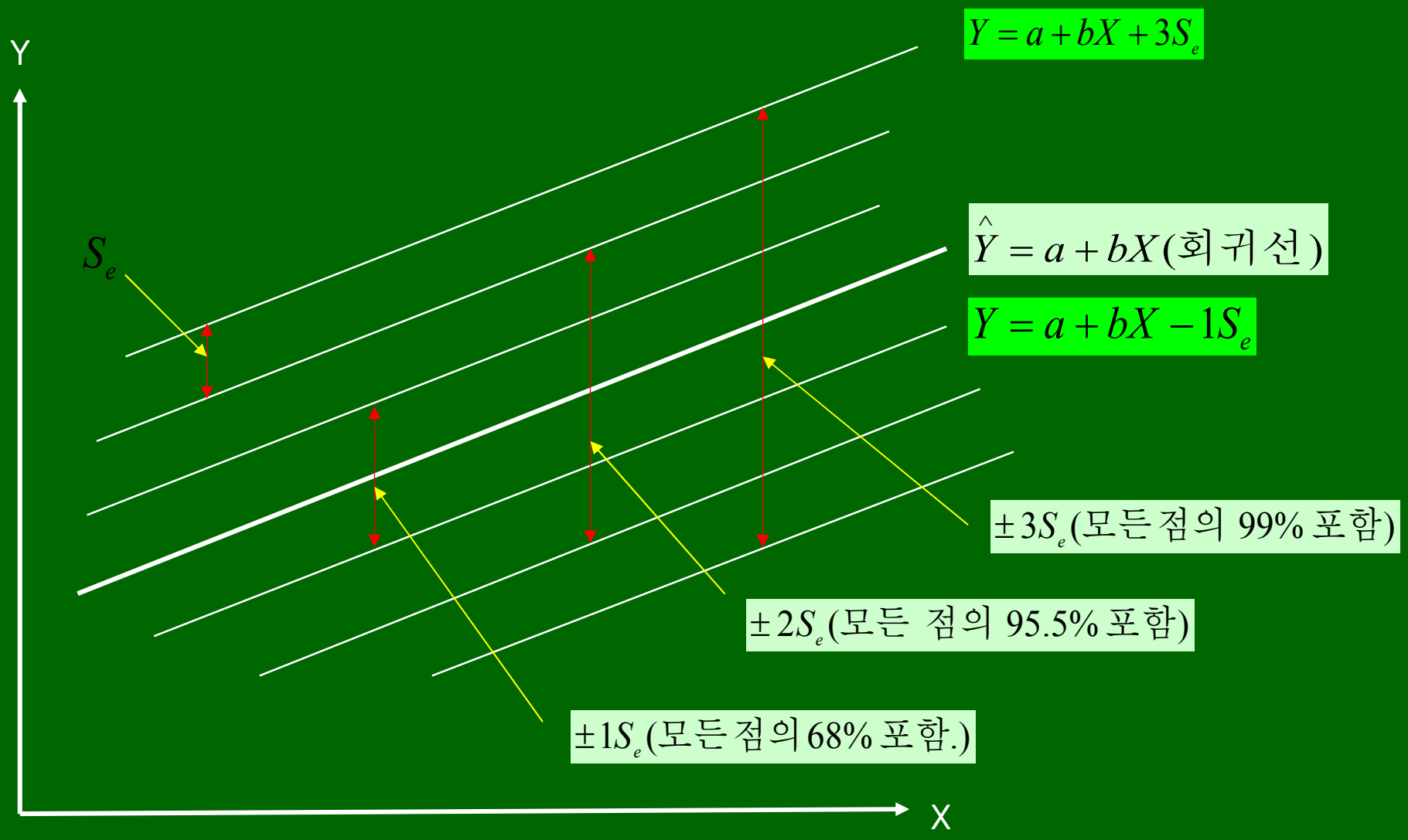
➢유의성검정(significance test)
각 독립변수와 종속변수의 관련도가 유의한 지 또는 종속변수에 대한 설명력을 가지고 있는가를 밝히는 것이다.

추정치의 표준오차

$$S_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum Y_i^2 - a\sum Y_i - b\sum X_i Y_i}{n-2}}$$

SSE : 오차제곱합

회귀선과 표준오차


Y
X
S_e
$Y = a + bX + 3S_e$
$\hat{Y} = a + bX$ (회귀선)
$Y = a + bX - 1S_e$
$\pm 3S_e$ (모든 점의 99% 포함)
$\pm 2S_e$ (모든 점의 95.5% 포함)
$\pm 1S_e$ (모든 점의 68% 포함.)

결정계수

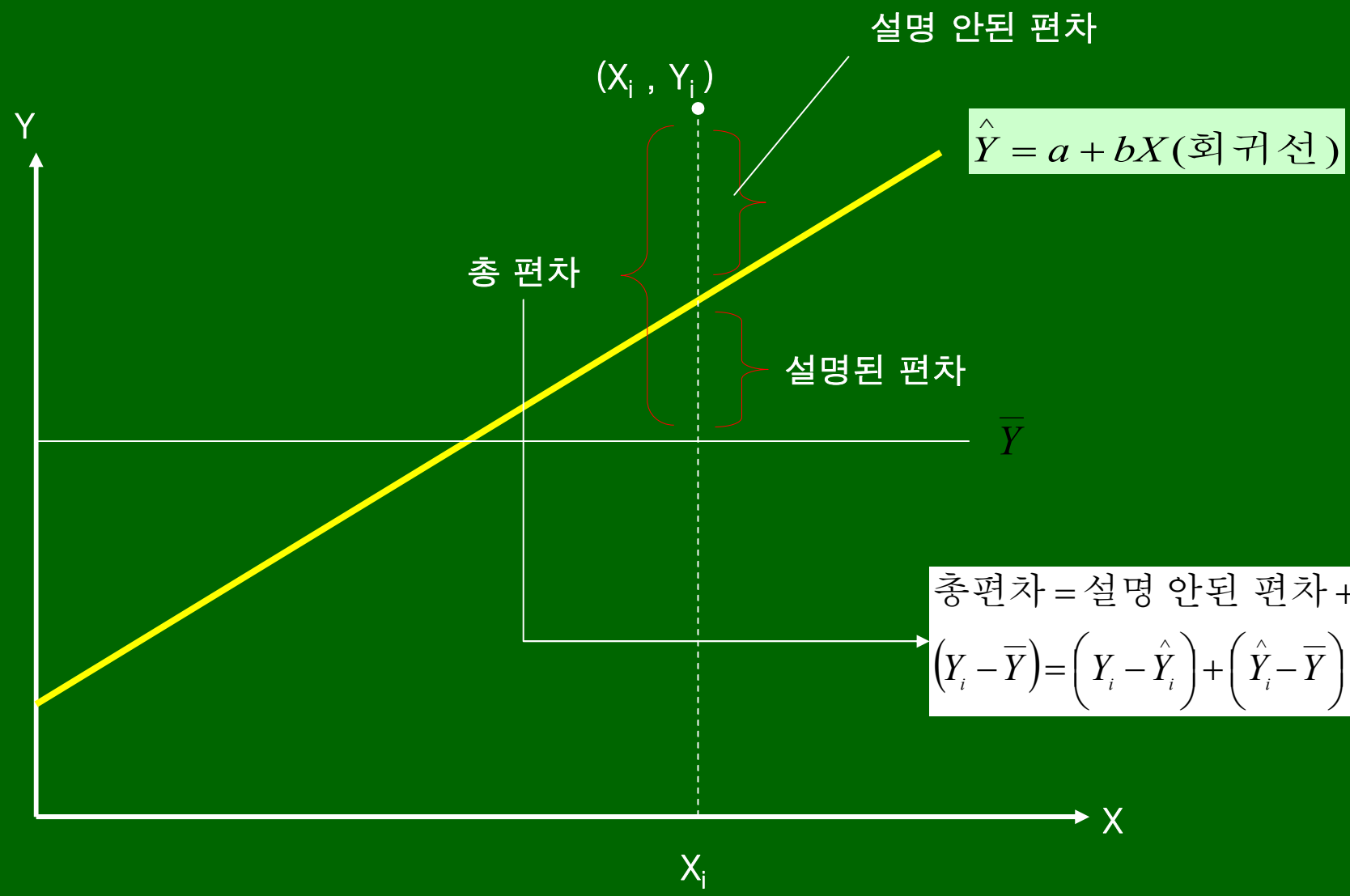
표본회귀선이 표본자료를 얼마나 잘 설명하는가를 평가하는 기준의 하나가 결정계수(coefficient of determination)이다.

총편차의 구성

$$\text{총편차} = \text{설명 안된 편차} + \text{설명된 편차}$$

$$\left(Y_i - \overline{Y}\right) = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \overline{Y}\right)$$

총 편차의 구성



총편차 = 설명 안된 편차 + 설명된 편차

$$\left(Y_i - \overline{Y}\right) = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \overline{Y}\right)$$

결정계수

결정계수는 0부터 1까지의 값을 갖는데 표본회귀선이 모든 자료에 완전히 적합하면 결정계수는 1이 된다.

총제곱합 = 오차제곱합 + 회귀제곱합

$$\sum (Y_i - \overline{Y})^2 = \sum \left(Y_i - \hat{Y}_i \right)^2 + \sum \left(\hat{Y}_i - \overline{Y} \right)^2$$

$$SST = SSE + SSR$$

결정계수:

$$R^2 = \frac{\text{설명되는 변동}}{\text{총변동}} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{a\sum Y_i + b\sum X_i Y_i - n\overline{Y}^2}{\sum Y_i^2 - n\overline{Y}^2}$$

상관계수

- 상관계수(correlation coefficient)

두 변수의 짝 지은 관측치 사이에 존재하는 관계의 강도와 방향을 결정.

- 표본상관계수

회귀분석과 관련된 문제에 있어서 결정계수의 제곱근.

표본상관계수

$$r = \sqrt{R^2} = \sqrt{\frac{SSR}{SST}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\sqrt{\sum (Y_i - \overline{Y})^2}}$$

표본상관계수의 특징

- r 은 –1.0부터 +.10까지의 값을 갖는다.
- | r |이 클수록 두 변수 사이의 선형관계는 더욱 강하다.
- r 이 0에 가깝다는 것은 두 변수 사이에 선형관계가 없음을 의미한다.
- r =1 또는 r =–1은 두 변수 사이에 완전한 선형관계가 있음을 의미한다. 즉 표본회귀선은 모든 표본점들을 통과한다.
- r 이 0, –1, +1의 값을 갖는 경우는 실제로 흔치 않다.
- r 의 부호가 +이면 두 변수의관계가 정의관계이고 –이면 부의 관계이다.