통계 - 회귀모형

목차

1-1 회귀 분석이란?

한 나라에서 국민소득이 증가하면 자동차 보유대수는 어느 정도 증가할까?

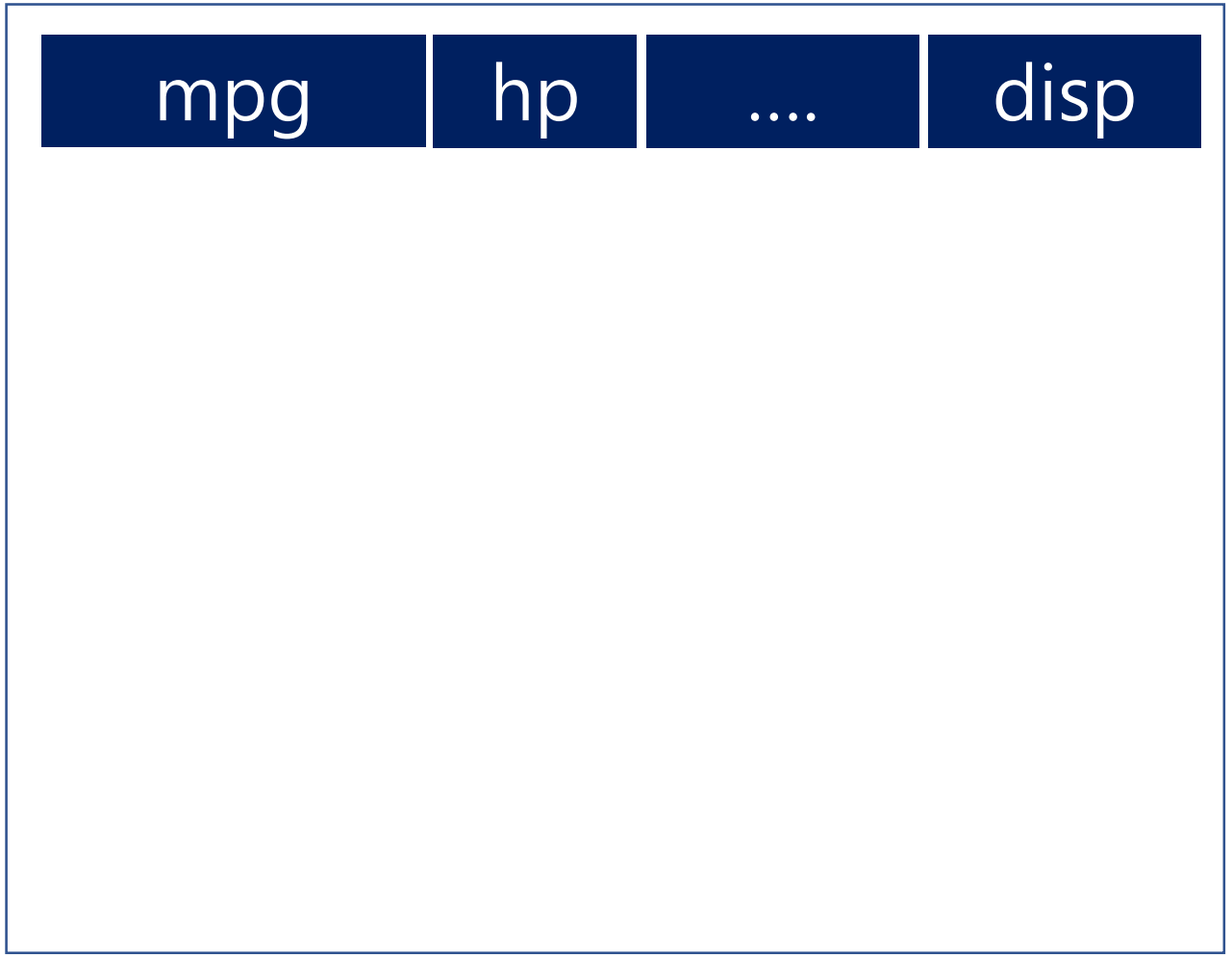
자동차 사고 **발생건수**가 증가하면 **병원의 입원 환자수**는 어느 정도 증가하는가?

날씨에 따라 자전거 대여 수는 어떻게 변화할까?

=> 연구 조사 결과에 대해 변수 간의 **상호 관련성**을 찾고 싶다.

=> 영향을 끼치는 변수(독립변수), 영향을 받는 변수(종속변수)에 대해 이를 확인할 수 있는 함수 관계로 규명할 수 없을까?

1-1 회귀 분석이란?



회귀 분석은 변수 사이에 함수적 관계를 조사하는 통계적 방법이다.

**관계는 목표(종속) 변수와
독립(설명)변수를
연결하는 방정식 또는 모형의
형태로 표현된다.**

1-2 회귀 모형 종류

▶ **목표(종속)변수의 연속형인지 이항형인지에 따라 구분된다.**

mpg	hp	….	Target 목표변수

연속형변수

목표 변수가 연속 형인 경우,
[다중, 단순]선형회귀모형
(Linear Regression)

이항형 변수

목표 변수가 이항형인 경우,
로지스틱 회귀 모형

1-3 회귀 모델의 유래

- **1885년 영국의 과학자 갈톤(F. Galton)이 발표했다.**

  Regression toward Meiocrity in Hereditary Stature
  유전에 의해 보통사람의 신장으로 회귀

- **내용**

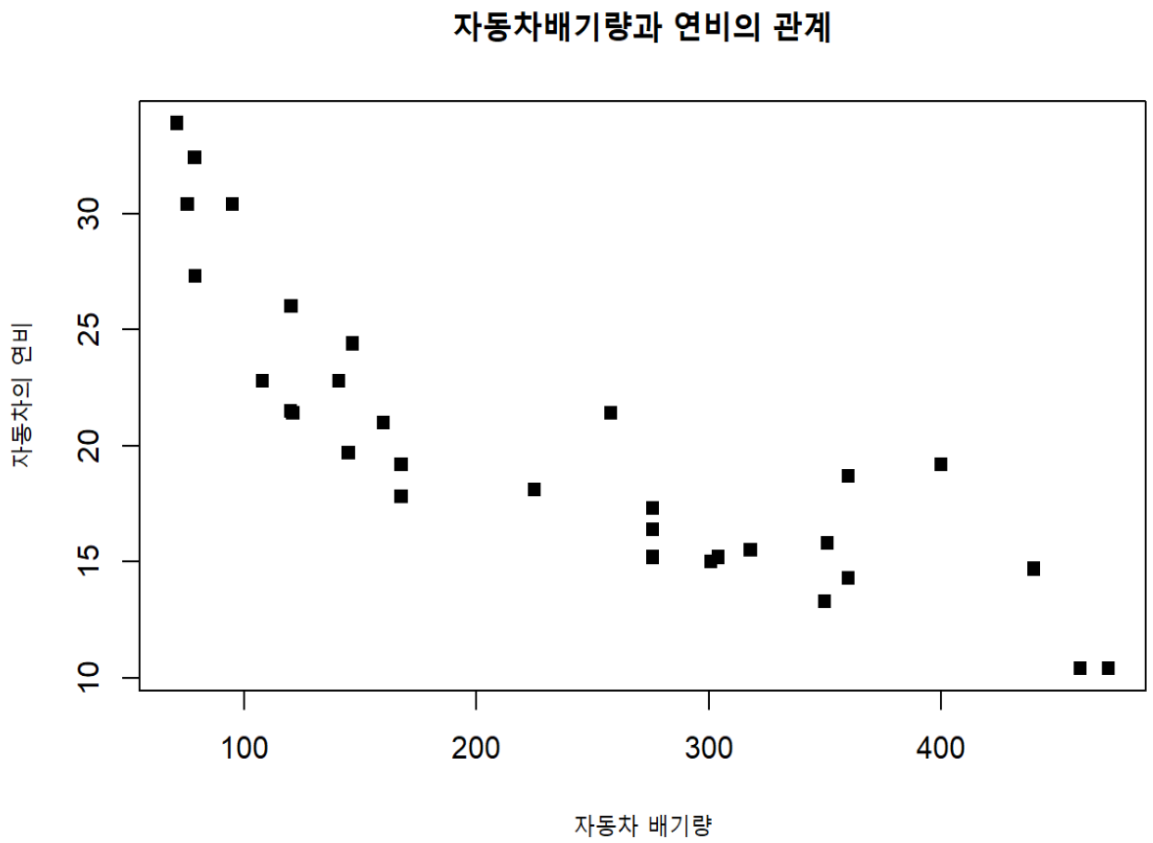
  부모의 키가 **매우 클 때** 아들의 키는 일반적으로 평균키보다 **크지만** 그들의 **부모만큼 크지는 않다**는 결론.

  부모의 키가 **매우 작을 때** 아들의 키는 일반적으로 평균키보다 **작지만** 그들의 **부모만큼 작지는 않다**는 결론.

  부모의 키가 매우 크든 작든 그 자식들은 결국 보통키로 돌아간다.(회귀한다.)

1-4 Linear Regression

▶ 두 변수 간의 함수 관계를 밝히는 첫번째 단계는 산점도 그리기



기본 plot이용 산점도 그리기 R

```
plot(mtcars$disp, mtcars$mpg,
     xlab="자동차 배기량",
     ylab="자동차의 연비",
     pch = 15, main="자동차배기량과 연비의 관계")
```

1-5 Simple Linear Regression(단순선형회귀)

▶ **독립변수가 한 개인 경우,**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$ ⟵ 단순선형회귀 모형

Y_i : i번째 측정된 종속(반응)변수 Y의 값

β_0 : 절편 coefficients(회귀계수)

β_1 : 기울기 coefficients(회귀계수)

X_i : i번째 주어진 상수 X값

ε_i : i번째 측정된 오차항

우리는 어떻게 회귀 모델을 구할 것인가?

1-6 최소 제곱법

▶ **가장 많이 사용되는 방법은 최소 제곱법**
-> (method of least squares)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\varepsilon_i = \beta_0 + \beta_1 X_i - Y_i$$

$$\varepsilon_i^2 = (\beta_0 + \beta_1 X_i - Y_i)^2$$

$$\Rightarrow \quad S = \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i - Y_i)^2$$

S(오차 제곱의 합)를 최소로 하는 β_0, β_1 의 값들을 추측하는 방법이다.
이를 위해서는 오차 제곱 합 S를 β_0, β_1 으로 각각 편 미분하여 이를 0으로 하여 구하면 된다.

http://bitly.kr/JArwFQC : 최소제곱법 네이버 캐스터 참조

1-6 최소 제곱법

▶ **표본 자료(sample data)로 부터 모델을 추정하여 얻은 식**

$$\hat{Y} = b_0 + b_1 X$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

ε_i**(오차) 제곱들의 합이다.**

$$S = \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i - Y_i)^2$$

1-6 최소 제곱법

▶ **표본 자료(sample data)로 부터 모델을 추정하여 얻은 식**

$$\hat{Y} = b_0 + b_1 X$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$S_{XX} = \sum (X_i - \bar{X})^2$$

$$S_{YY} = \sum (Y_i - \bar{Y})^2$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

1-7 Residual(잔차)

▶ **적합(학습)된 회귀 직선 –fitting Regression Line**

$$\hat{Y} = b_0 + b_1 X \quad \Rightarrow \quad \hat{Y}_i = b_0 + b_1 X_i$$

X대신에 i번째 X의 값 X_i 를 사용하여 표현

▶ **잔차(Residual) : i번째의 실제값과 예측(추정)된 값의 차이**

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

*** 잔차를 전부 더해주게 되면 회귀 모형은 최소제곱법에 의해 구해진 값이므로 전체 잔차를 더하면 거의 0에 가까운 값이 된다.**

1-8 Regression model precision(회귀 모형 정도)

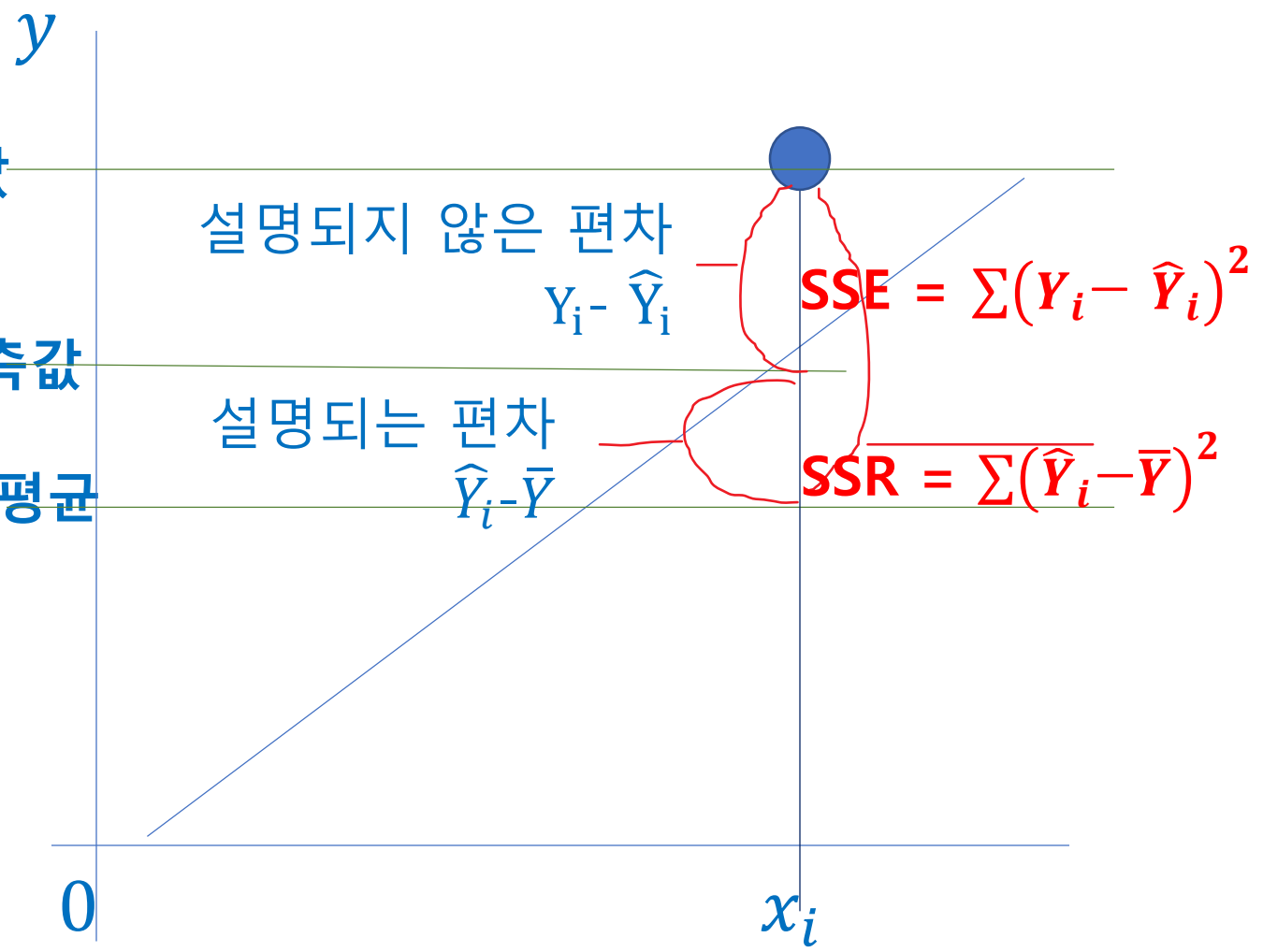
- 추정된 회귀선의 정도(모델이 잘 되었나?)를 측정하는 여러가지 측도(measure)들 중에서 널리 이용되는 세가지 살펴보기

▶ **(1) 분산 분석표에 의한 F-검정**

▶ **(2) 결정계수**

▶ **(3) 예측(추정)값의 표준오차**

1-8 Regression model precision(회귀 모형 정도)



SSE = $\sum(Y_i - \hat{Y}_i)^2$

잔차에 의한 제곱 합

SSR = $\sum(\hat{Y}_i - \bar{Y})^2$

회귀에 의한 제곱합

SST = SSE(sum of squares due to residual errors) **+**
SSR(sum of squares due to regression)

1-9 분산 분석표

다중 회귀의 분산 분석표

요인	제곱합	자유도	평균제곱	F-값	P-값
회귀	SSR(회귀에 의한 제곱합)	k	MSR=SSR/k	MSR/MSE	
잔차	SSE(잔차에 의한 제곱합)	n-k-1	MSE=SSE/(n-k-1)		
전체	SST(총 편차의 제곱합)	n-1			

회귀 직선이 유의한가는 분산 분석표(analysis of variance table)를 만들어 살펴본다.
(1) K는 독립변수의 개수
(2) 전체의 자유도는 회귀와 잔차의 자유도를 더한값과 같다.

1-10 검정 통계량 F0

- **회귀 계수에 대한 가설**

  H0 : $\beta 1 = 0$ 귀무가설

  H1 : $\beta 1 \neq 0$ 대립가설

- **검정 통계량**

$$F_0 = \frac{MSR}{MSE}$$

  F_0의 값은 높으면 높을 수록 좋다. 높다는 것은 귀무 가설을 기각하고, 내가 가설을 세운 대립 가설이 맞을 확률이 높아진다.

- **기각의 기준 – 유의 수준 α와 자유도(1, n-2)을 이용**

  F(1, n-2; α)를 구한다.

  F_0 > F(1, n-2; α) 이면 귀무 가설 기각하고 회귀 직선이 유의함

1-11 변수 선택법

- 모든 가능한 회귀(all possible regression)
- 앞으로부터 선택법(forward selection)
- 뒤로부터 제거법(backward elimination)
- 단계별 회귀방법(stepwise regression)