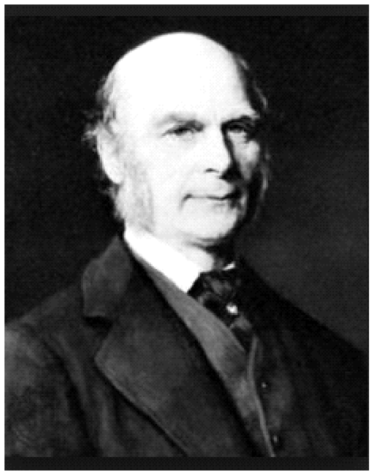
3. 회귀분석

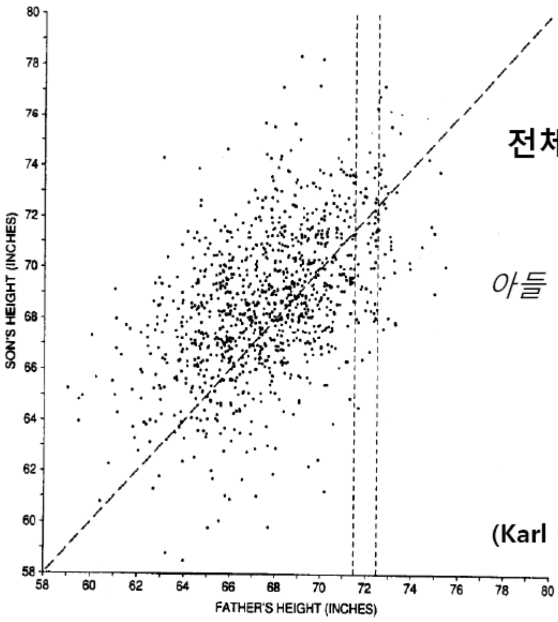
회귀란 ?

- 연어 : 다시 고향으로 돌아와 알을 낳고 죽음.
- "다시 본디 상태로 되돌아 온다"

<Francis Galton, 1822~1911>
- 영국, 인류학자
- "회귀(regression)"라는 용어의 사용
- 일반적으로 키가 큰 부모에게서 키 큰 자녀가, 작은 부모에게서 작은 자녀가 태어나지만 전체적으로 전체 인구의 평균 키로 접근하는 현상을 보인다. (보편적 회귀의 법칙, Law of universal regression)



<Karl Pearson, 1857~1936>
- 영국, 통계학자
- 체계적인 회귀분석이론 정립
- 아버지의 키와 아들의 키 사이에 존재하는 회귀법칙 규명
- 키 큰 아버지 집단에서 태어난 아들들의 평균 키는 아버지들의 평균 키보다 작으며 키 작은 아버지 집단에서 태어난 아들들의 평균 키는 아버지들의 평균 키보다 크다.



전체 1978쌍의 아버지와 성숙한 아들의 신장 간의 연구

아들 신장 = 33.73 + 0.516 X 아버지 신장
단위: (inch)

(Karl Pearson, 1903, *Biometrika*, Vol2, p357)

6

A. 종속변수와 독립변수

1) 종속 변수 (dependent variable) : 반응변수 (response variable)
결과변수 (outcome variable)

2) 독립 변수 (independent variable) : 설명 변수 (explanatory variable)
예측변수 (predictor)

독립변수 (Independent variable)	**종속변수 (dependent variable)**
수학능력시험 점수	1학년 학기말 성적
홍보비용 (만원)	예금 유치액 (만원)
수면제의 용량 (M/kg)	수면시간 (시간)
수축기 혈압 (mmHg)	나이 (age)
출생시 신생아의 체중 (kg)	초음파검사시 태아의 배 (abdomen)둘레 (cm)

B. 변수들 간의 관계

1) 결정적 관계 (deterministic relationship)

$Y = f(x)$ 와 같은 함수식으로 정의되는 관계

원의 면적(S)과 반지름(r)과의 관계 : $S = \pi r^2$

2) 통계적 관계 (statistical relationship)

$Y \approx f(x) + \epsilon$ 과 같이 오차를 포함하는 확률적 모형으로 예측되는 관계

IQ에 따른 성적, 소득수준에 따른 소비지출액 등.

C. 회귀분석의 종류

1) 단순 선형회귀분석

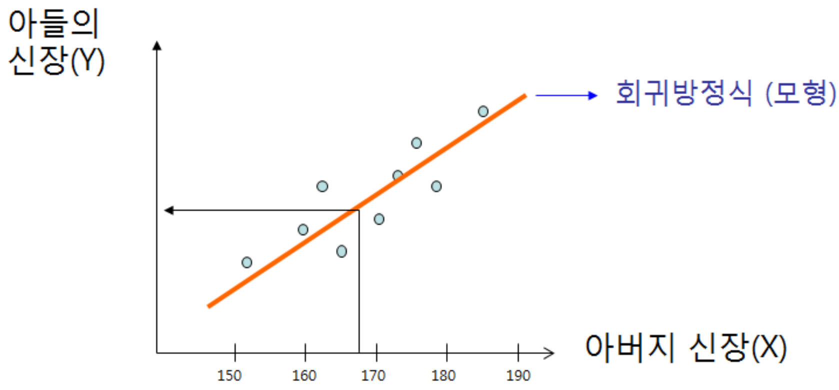
2) 다중 회귀분석

3) 다변량 회귀분석

4) 로지스틱 회귀분석

5) 비선형 회귀분석

D. 회귀분석의 개념

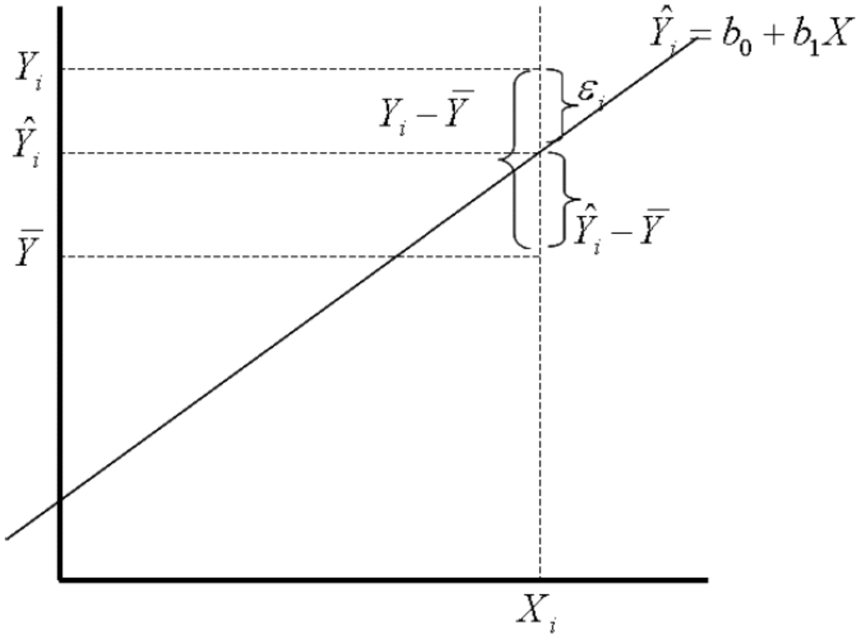


ex) 키가 168인 남자가 결혼하여 아들을 낳으면, 그 아들의 키는 아마도...

➔ 예측(prediction)

어떤 회귀직선이 가장 좋은가?

⇒ 최소제곱법 (Least Squares Estimation : LSE)



$$SST = SSR + SSE$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

E. 검정통계량

<분산분석표>

요인	자유도	제곱합	평균제곱합	F값
처리	1	SSR	$MSR=\dfrac{SSR}{1}$	$F=\dfrac{MSR}{MSE}$
오차	$n-2$	SSE	$MSE=\dfrac{SSE}{n-2}$	
전체	$n-1$	SST		

※ 주의 : ANOVA table과 반드시 비교해 볼 것 !!

<기각역>

$$F > F_{\alpha}(1,\ n-2)$$

MSR이 클수록 모형이 의미 있음.
(모형으로 설명할 수 있는 변동이 큼)

$F \uparrow \Leftrightarrow P-$값 $\downarrow \ \Rightarrow$ Reject H_0

$F \downarrow \Leftrightarrow P-$값 $\uparrow \ \Rightarrow$ Do /Reject H_0

F. 결정계수 (Determination Coefficient)

$$\begin{array}{lcccccc} SST & = & SSE & + & SSR & & \\ 1 & = & \dfrac{SSE}{SST} & + & \dfrac{SSR}{SST} & & \\ 1 & = & \uparrow & + & \downarrow & \Rightarrow & \text{나쁜 모형} \\ 1 & = & \downarrow & + & \uparrow & \Rightarrow & \text{좋은 모형} \end{array}$$

$$R^2 = \frac{SSR}{SST} = \frac{\text{회귀모형에 의한 변동}}{\text{자료의 총변동}}$$

모형으로 변동을 설명하는 정도를 나타내는 측도

일반적으로,

$$0 < R^2 < 1$$

$R^2 \approx 1 \Rightarrow$ 적합한 회귀모형(설명력이 높음)

$R^2 \approx 0 \Rightarrow$ 부적합한 회귀모형(설명력이 낮음)

(i) **정의**

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

(ii) $0 \le r^2 \le 1$

$$\begin{array}{lccl} r^2 & \downarrow & \Rightarrow & \text{나쁜 모형} \\ r^2 & \uparrow & \Rightarrow & \text{좋은 모형} \end{array}$$

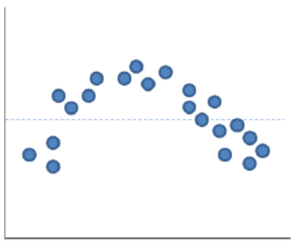
(iii)
$$r^2 = \frac{SSR}{SST} = \frac{S_{(xy)}^2}{S_{(xx)}\,S_{(yy)}} = \left(\frac{S_{(xy)}}{\sqrt{S_{(xx)}}\,\sqrt{S_{(yy)}}}\right)^2$$

$$= (\text{표본상관계수})^2$$

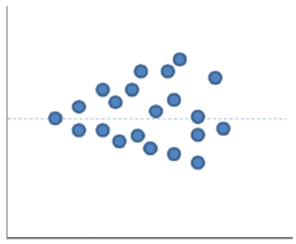
G. 잔차에 대한 가정

1) 선형성 (linearity)
2) 독립성 (independency)
3) 정규성 (normality)
4) 등분산성

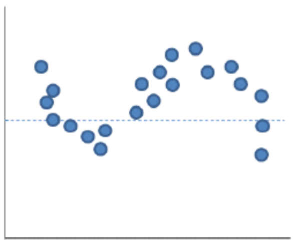
<잔차그림 이용>



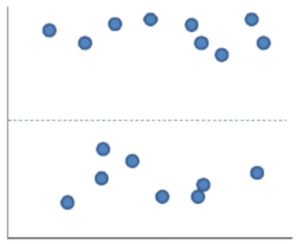
(a) 선형성이 벗어나는 경우



(b) 등분산성이 벗어난 경우



(c) 등분산성이 벗어난 경우



(d) 정규성이 벗어난 경우

H. 회귀계수에 대한 검정

1) β_1 에 대한 검정

- 귀무가설 : 기울기가 0이다 ($\beta_1 = 0$)

  ⇒ 독립변수에 의한 효과가 없다.

- 대립가설 : 기울기가 0이 아니다.

  ⇒ 독립변수에 의한 효과가 없다.

2) β_0 에 대한 검정

- 귀무가설 : 절편(intercept)이 0이다 ($\beta_0 = 0$)

  ⇒ 기저효과가 없다

- 대립가설 : 절편(intercept)이 0이 아니다

  ⇒ 기저효과가 있다.