선형회귀 (Linear Regression) 모델

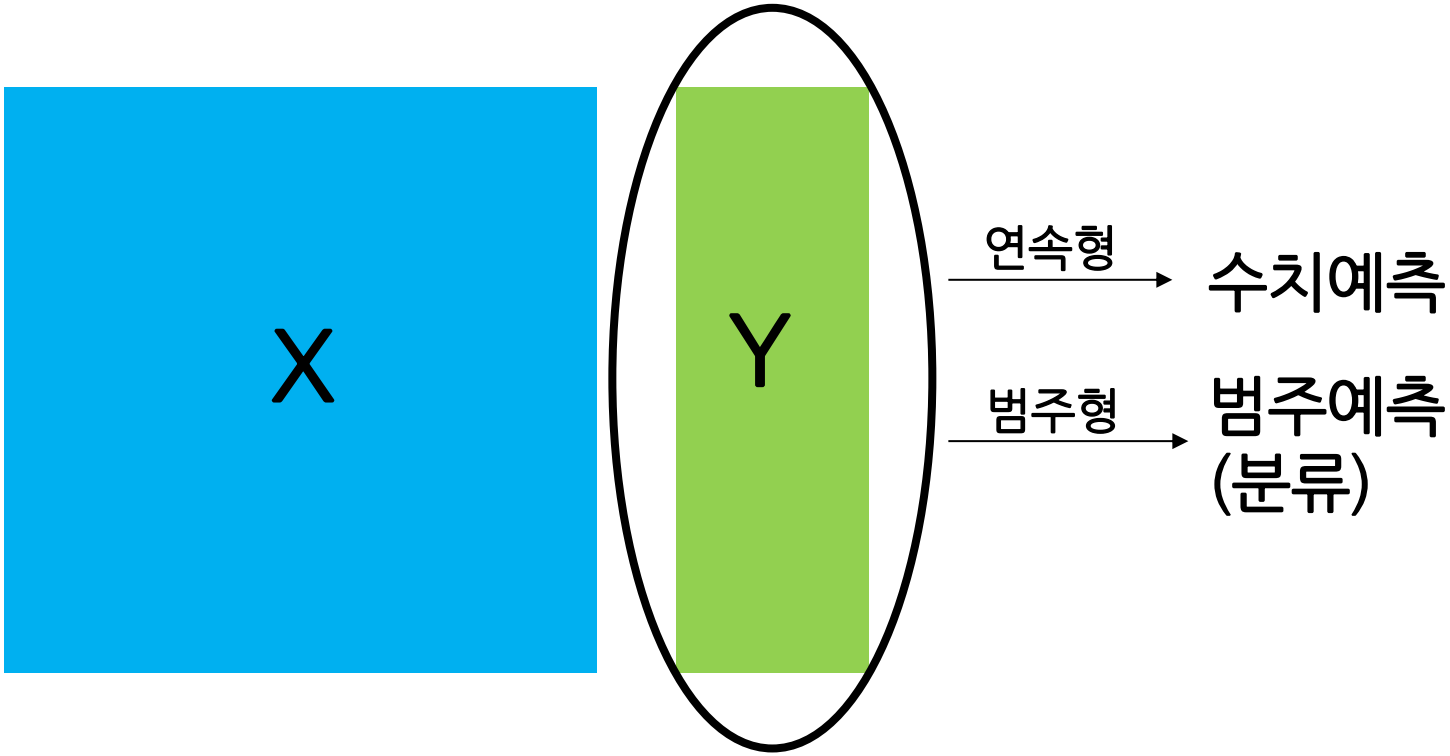
데이터

Y (결과): 종속변수, 반응변수, 출력변수

X (원인): 독립변수, 예측변수, 입력변수

변수 / 관측치	X_1	…	X_i	…	X_p	Y
N_1	x_{11}	…	x_{1i}	…	x_{1p}	20.5
N_2	x_{21}	…	x_{2i}	…	x_{2p}	22.2
…	…	…	…	…	…	…
N_{n-1}	x_{n-11}	…	x_{n-1i}	…	x_{n-1p}	72.3
N_n	x_{n1}	…	x_{ni}	…	x_{np}	82.8

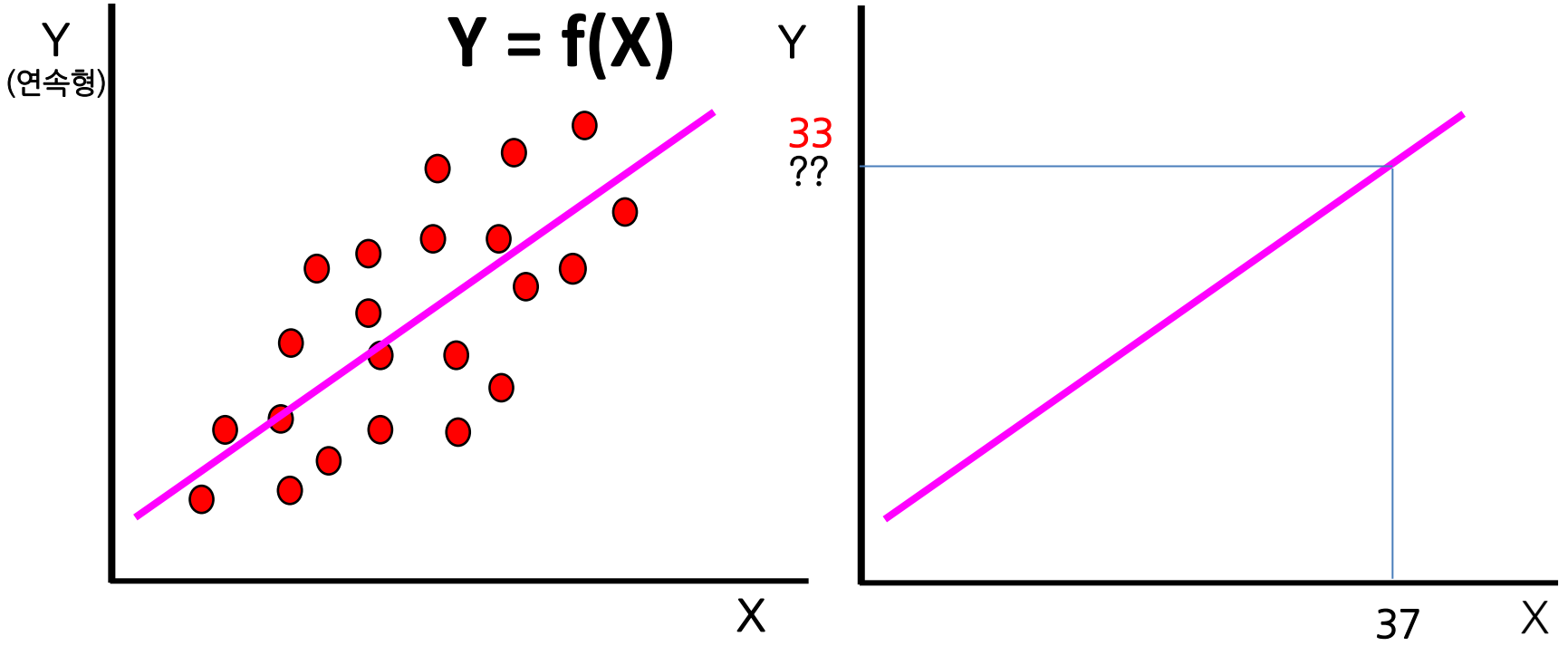
예측 (Prediction)



용어 정리

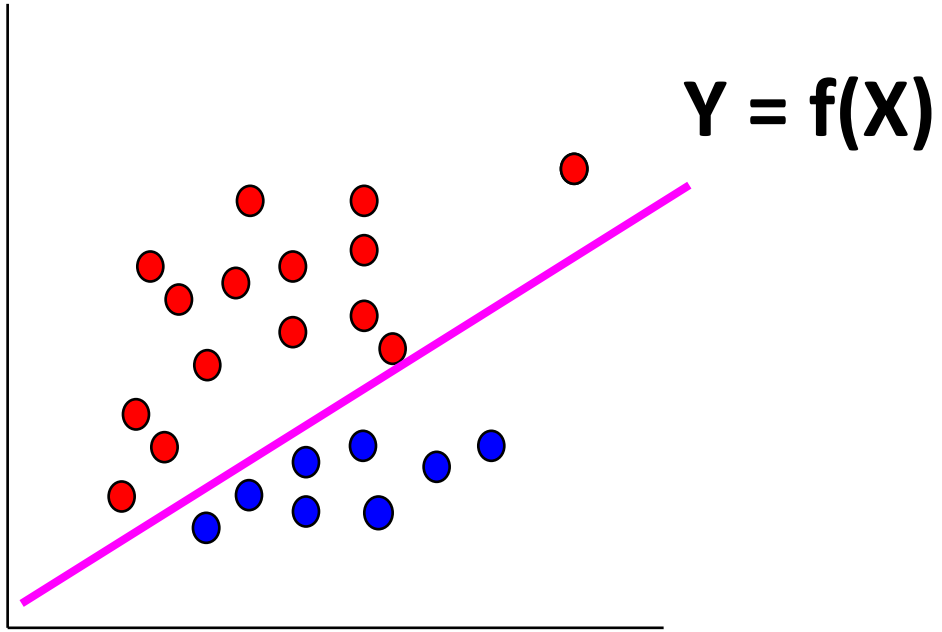
- Prediction: 예측
- Regression: 수치예측 (주의: 회귀모델과 다른 의미)
- Classification: 범주예측(분류)
- Forecasting (예측): 시계열데이터 예측 때 주로 사용
- Class: 범주

수치예측 (Regression)



범주예측, 분류 (Classification)

- 불량 클래스
- 양품 클래스



변수 사이의 관계

- X변수(원인)과 Y변수(결과) 사이의 관계
  - 확정적 관계
  - 확률적 관계

- 확정적 관계: X변수만으로 Y를 100% 표현 (오차항 없음)

$$Y = f(X)$$

예) 힘 = f(질량, 가속도), 주행거리 = f(속도, 시간)

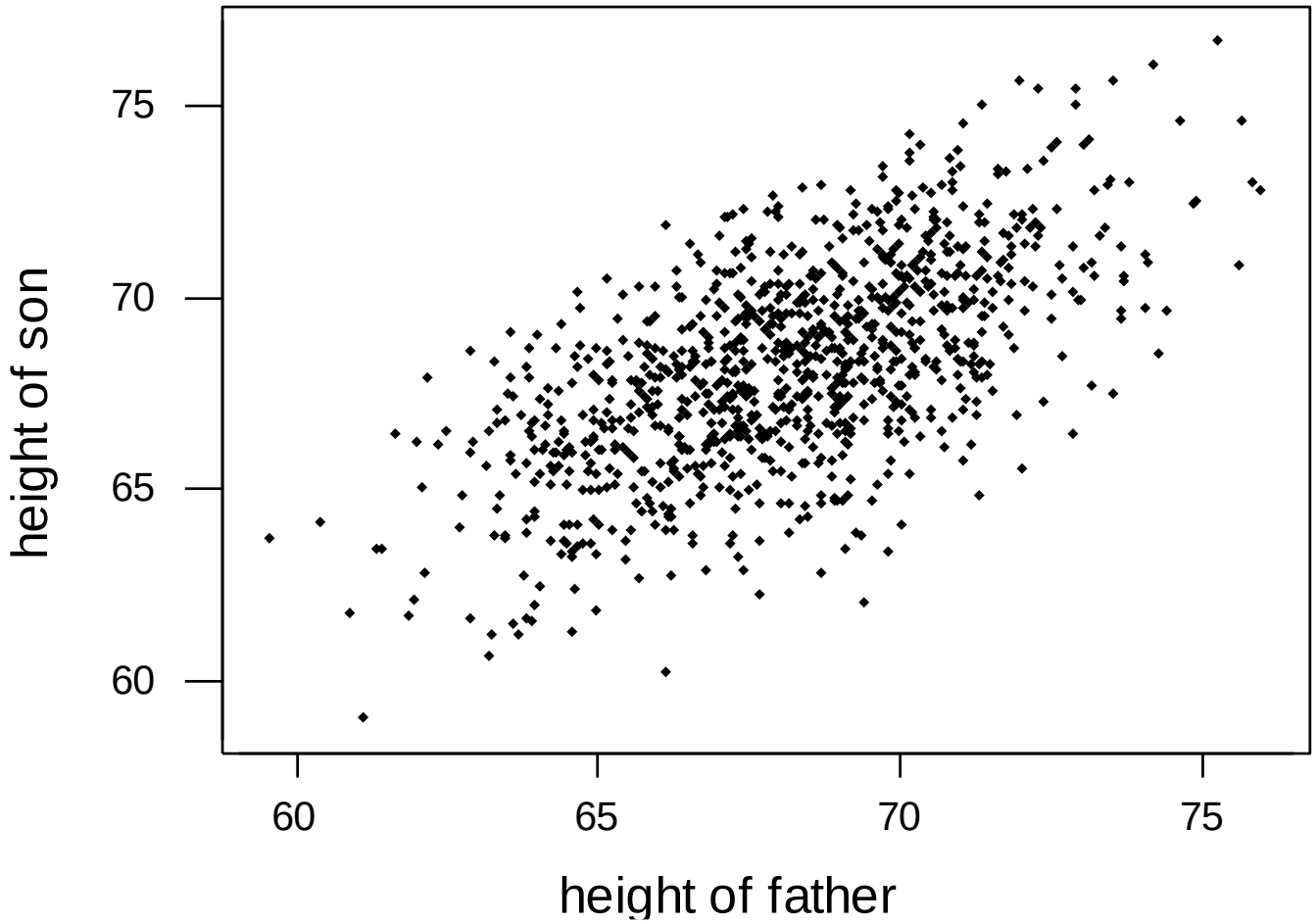
- 확률적 관계: X변수와 오차항이 Y를 표현 (오차항 있음)

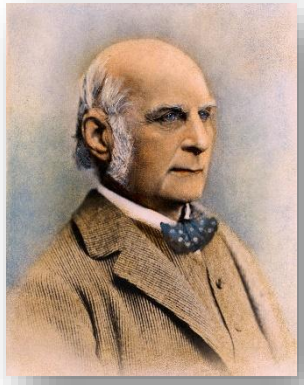
$$Y = f(X) + \varepsilon$$

예) 제품품질 = f(설비 파라미터들의 상태, 온도, 습도) + ε

포도주 가격 = f(강우량, 온도, 포도품종) + ε

선형회귀모델





Francis Galton
(1822~1911)

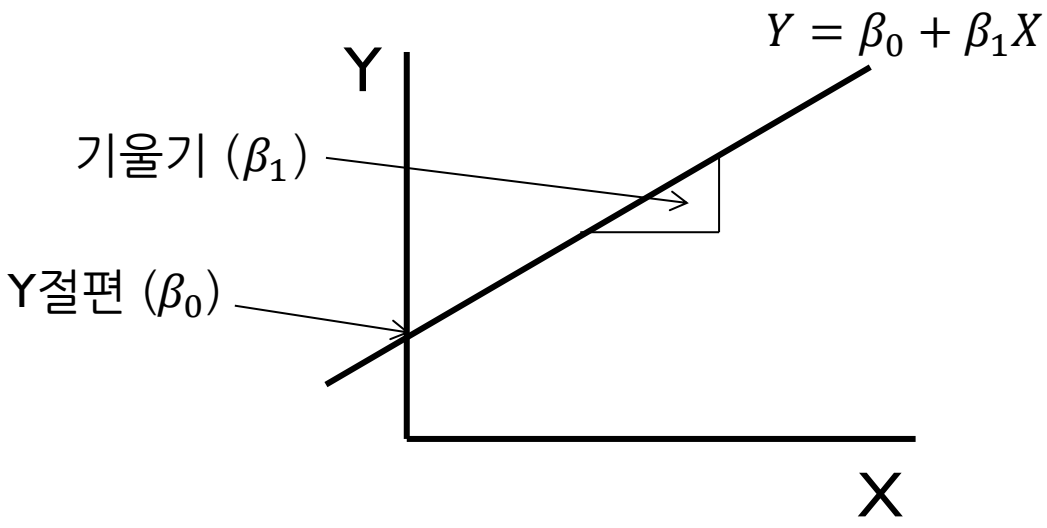
선형회귀모델

- 선형회귀모델: 출력변수 Y를 입력변수 X들의 선형결합으로 표현한 모델

  ☝ 선형결합: 변수들을 (상수 배와) 더하기 빼기를 통해 결합

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- X변수 한 개가 Y를 표현하는 경우: $Y = \beta_0 + \beta_1 X$ (직선 식)



$\beta_1 = 2$

$\beta_1 = -2$

$\beta_1 = 0$

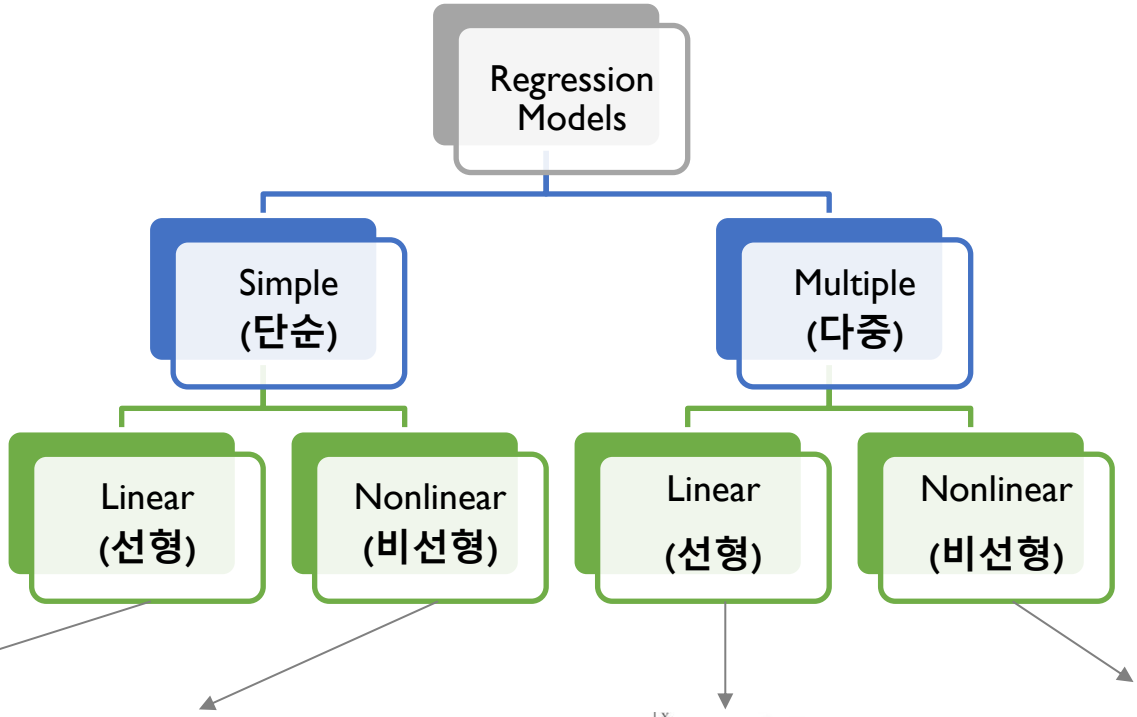
선형회귀모델링 목적

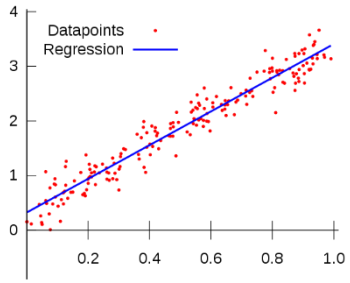
$$Y = \beta_0 + \beta_1 X$$

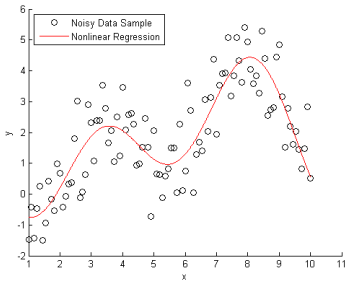
- X변수와 Y변수 사이의 관계를 수치로 설명
- 미래의 반응변수 (Y) 값을 예측

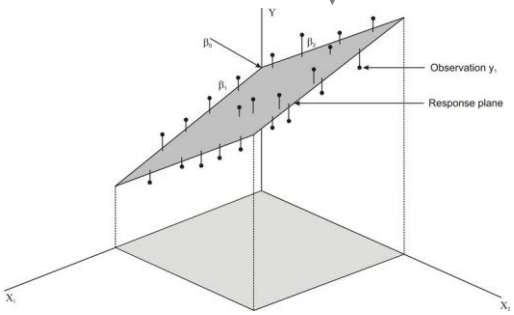
선형회귀모델 분류

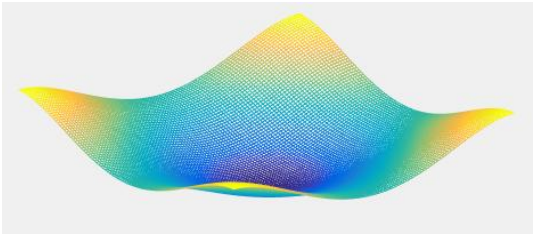
X변수의 수, X변수와 Y변수의 선형성 여부에 따라 구분



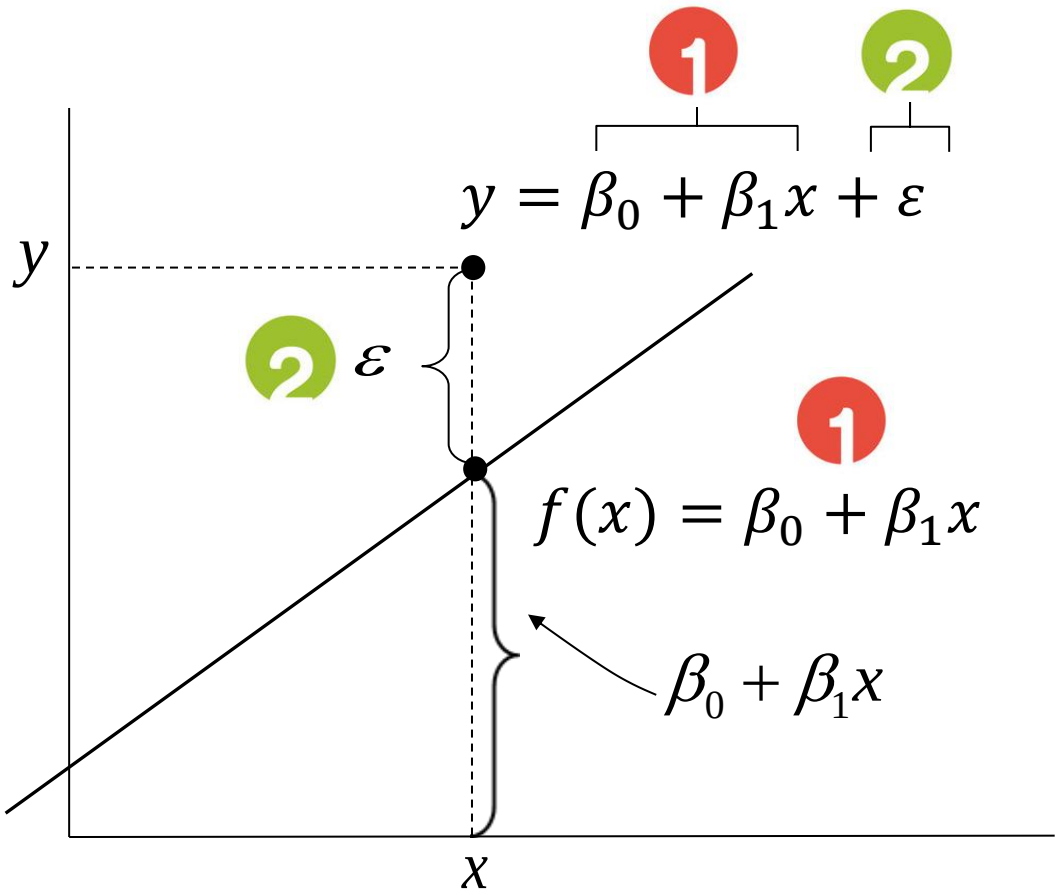








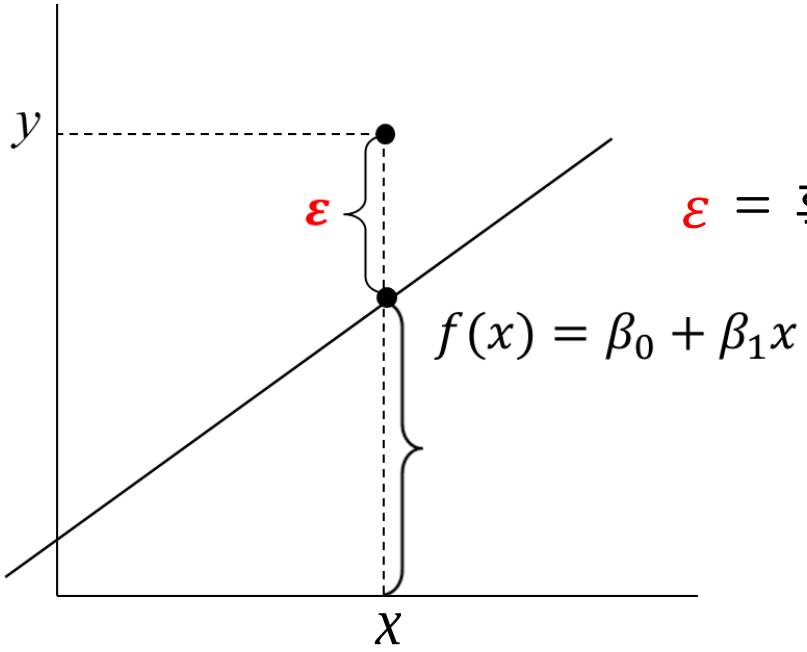
선형회귀 모델



Y = X로 부터 설명되는 부분 + 그렇지 않은 부분

1 2

선형회귀 모델 가정



ε = 확률오차 (random error)

확률오차 가정 : ε_i~정규분포 $E(\varepsilon_i) = 0$ $V(\varepsilon_i) = \sigma^2$ for all *i*.

$$\varepsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n$$

선형회귀 모델 가정

$$\varepsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon \quad E(Y_i) = ? \quad V(Y_i) = ?$$

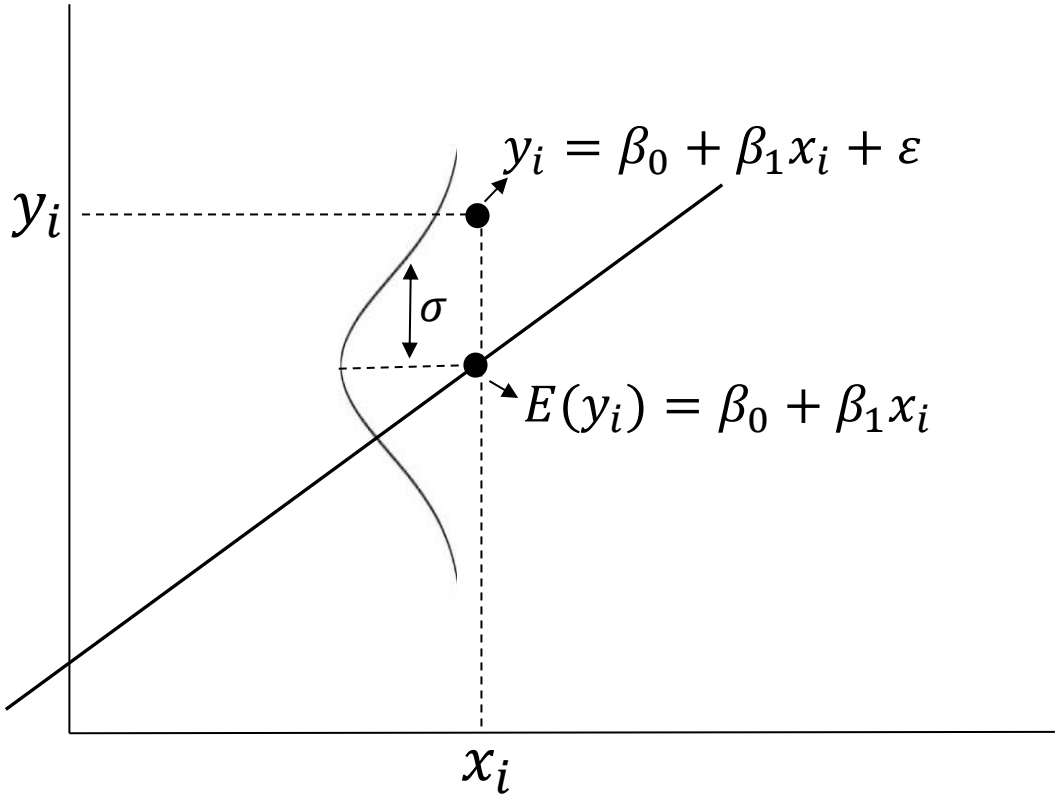
$$E(Y_i) = E(\beta_0 + \beta_1 X_i) + E(\varepsilon) = \beta_0 + \beta_1 X_i$$

$$V(Y_i) = V(\beta_0 + \beta_1 X_i) + V(\varepsilon) = \sigma^2$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2),\ i = 1, 2, \ldots, n$$

선형회귀 모델 가정

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2),\ i = 1, 2, \ldots, n$$



선형회귀 모델

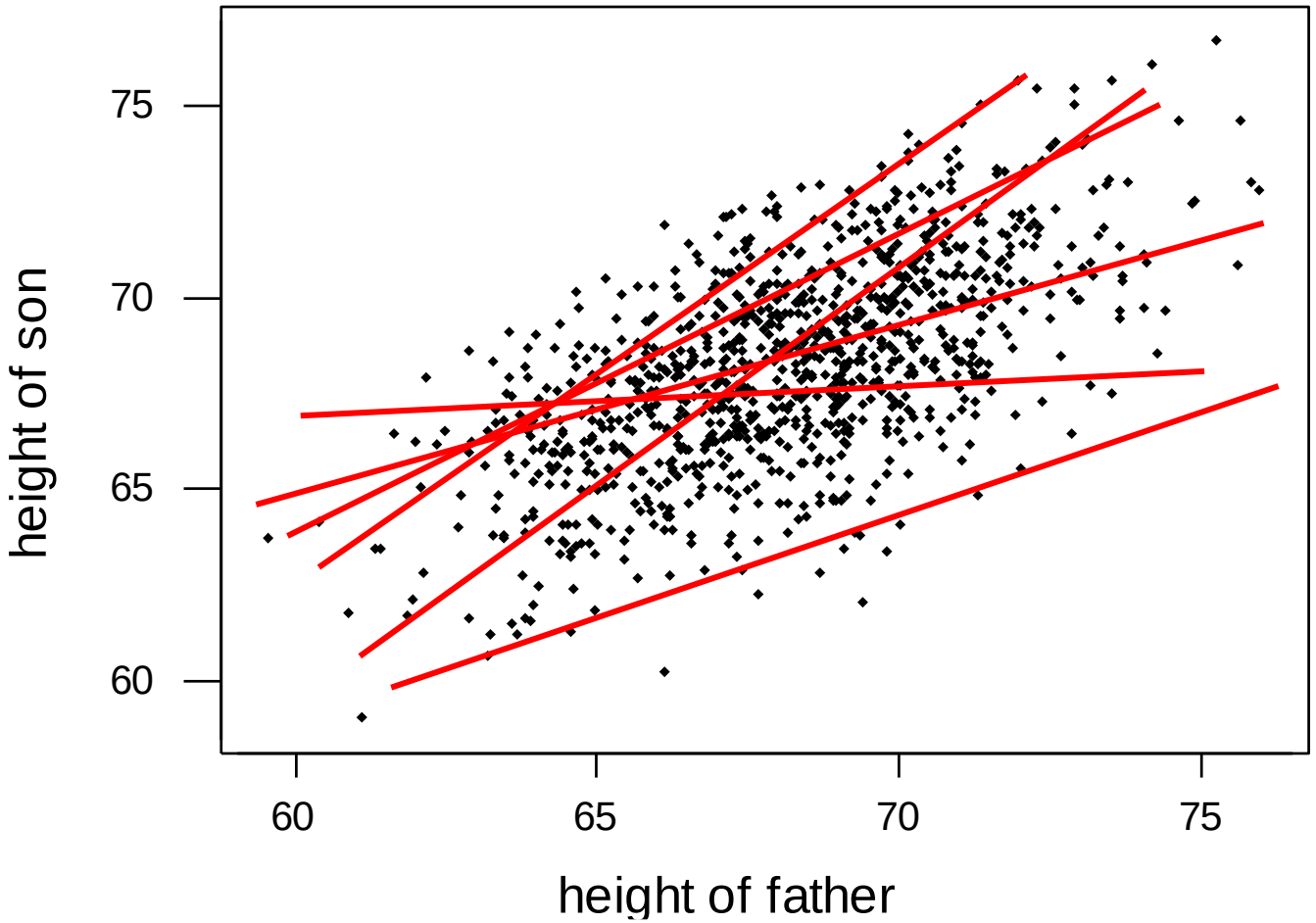
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

입력변수(X)와 출력변수(Y) 평균과의 관계를 설명하는 선형식 찾기!!

선형회귀 모델

$$E(Y) = f(X) = \beta_0 + \beta_1 X_1$$



선형회귀 모델

$$E(Y) = f(X) = \beta_0 + \beta_1 X_1$$

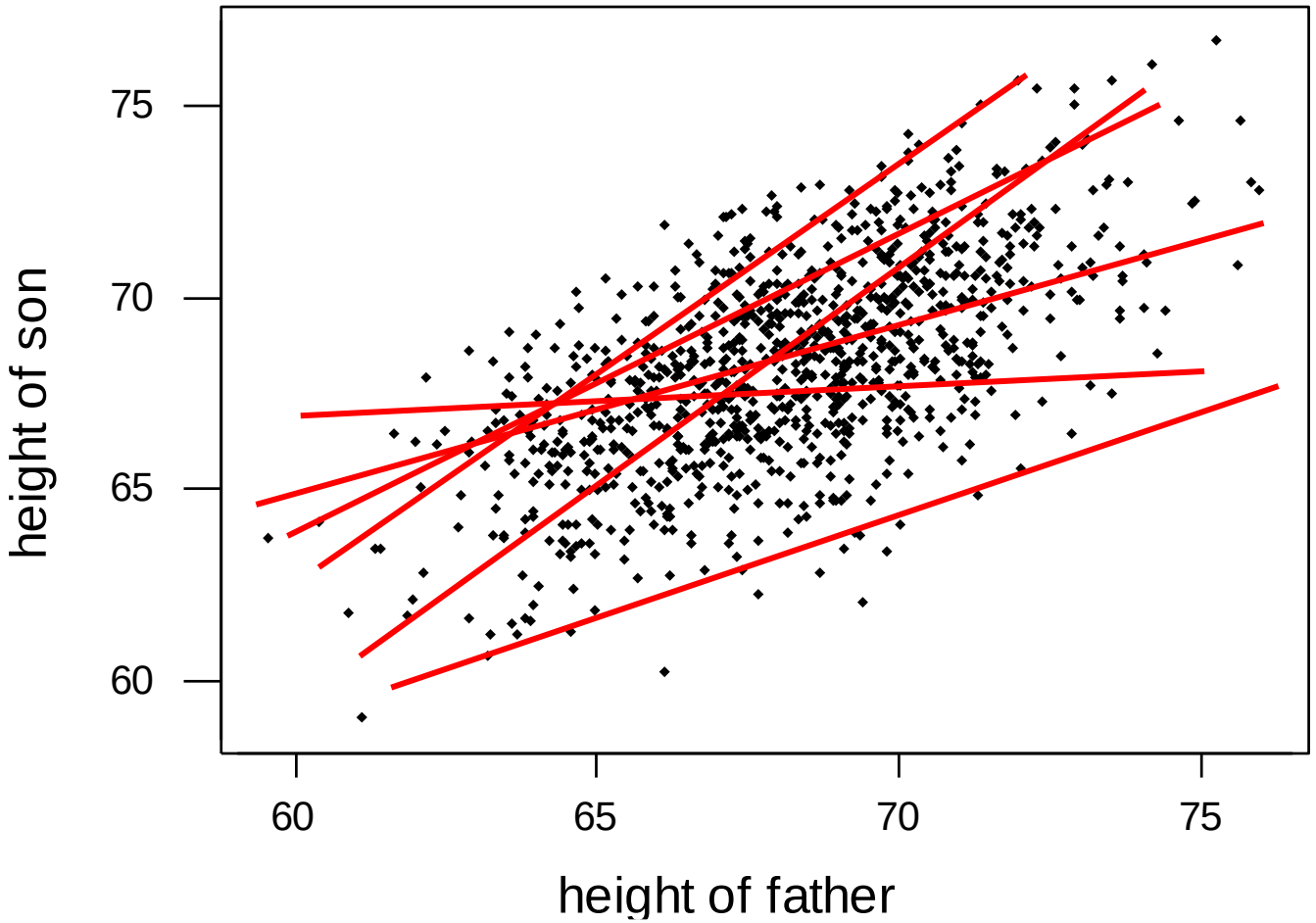
파라미터 (Parameter)

파라미터를 찾자 (추정하자)

가지고 있는 데이터들의 함수식으로!

선형회귀 모델

$$E(Y) = f(X) = \beta_0 + \beta_1 X$$

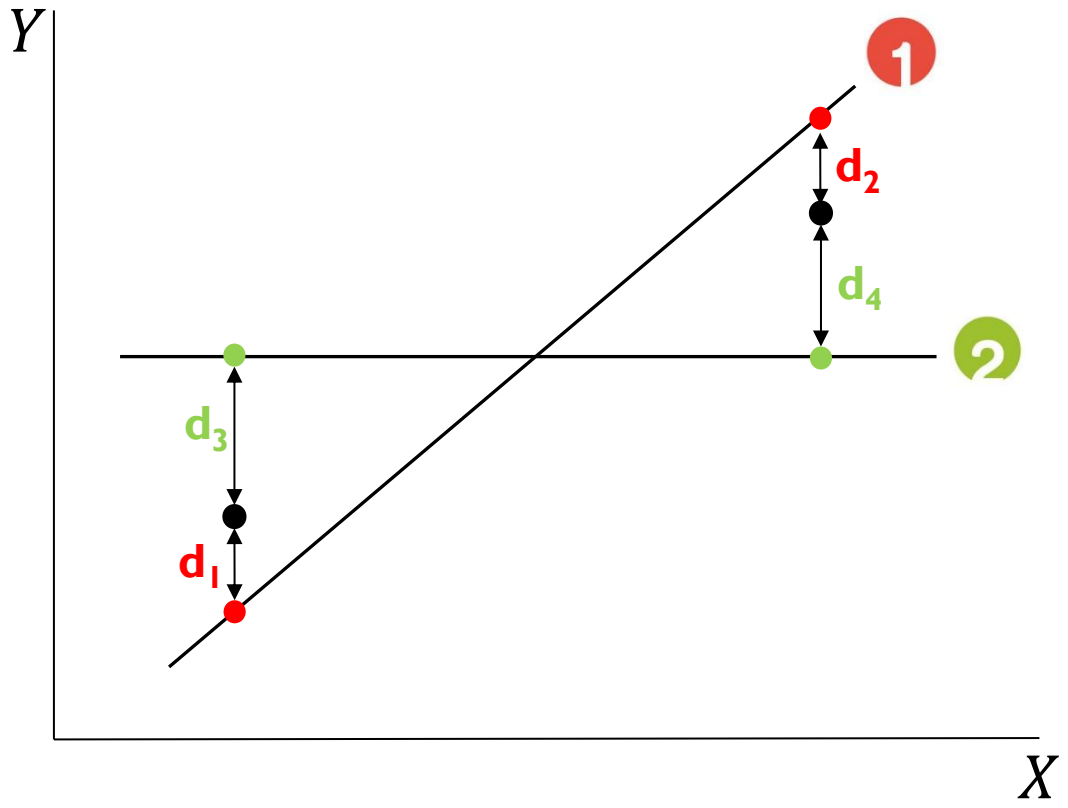


선형회귀 모델

$$E(Y) = f(X) = \beta_0 + \beta_1 X$$

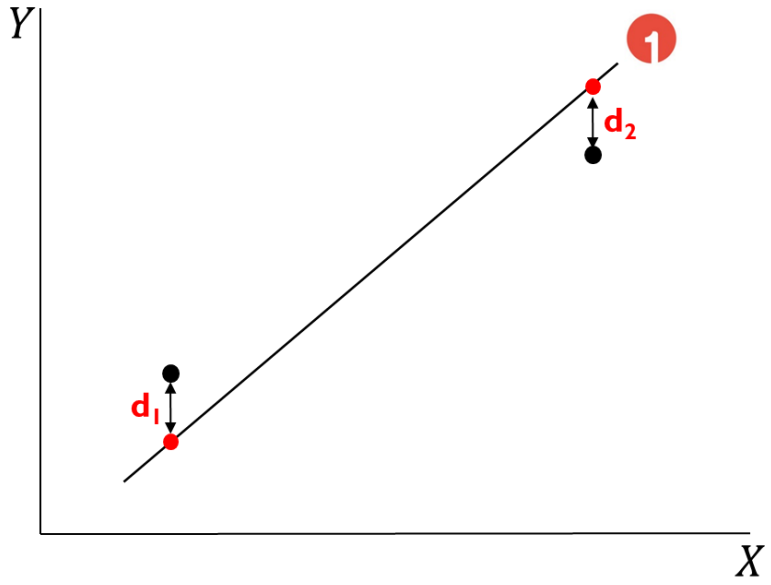
파라미터 (**Parameter**)

파라미터 추정



$$d_1+d_2 < d_3+d_4$$

파라미터 추정



$$d_1+d_2+\cdots+d_n=0$$

$$d_1^2+d_2^2+\cdots+d_n^2\geq 0$$

$$d_1=Y_1-E(Y_1)$$
$$=Y_1-(\beta_0+\beta_1X_1)$$

$$\sum_{i=1}^{n}d_i^2=\sum_{i=1}^{n}\{Y_i-(\beta_0+\beta_1X_i)\}^2$$

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{n}\{Y_i-(\beta_0+\beta_1X_i)\}^2$$

Cost function
(비용함수)

파라미터 추정

Cost function (비용함수)

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

Algorithm

$$\hat{\beta}_0,\ \hat{\beta}_1$$

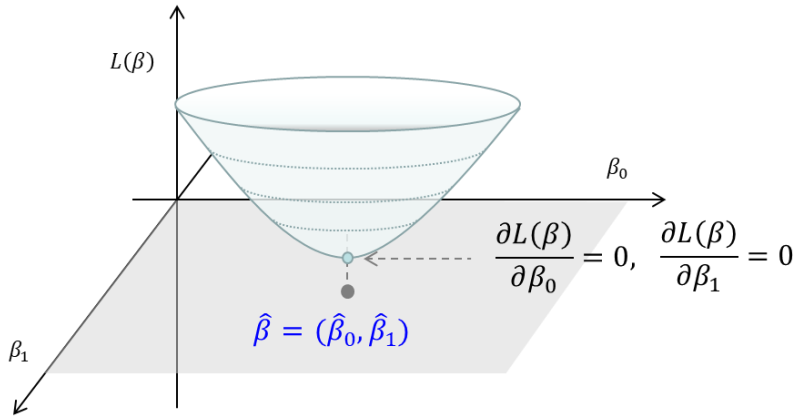
$$f(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

파라미터 추정 알고리즘

Cost function (비용함수)

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

- Cost function is convex → globally optimal solution exists (전역 최적해 존재)



$$\frac{\partial C(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) = 0$$

$$\frac{\partial C(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i)X_i = 0$$

파라미터 추정 알고리즘

$$\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) = 0$$

$$\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) X_i = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$f(X) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

파라미터 추정 알고리즘

Question. Find estimator of β_0 and β_1 $(i.e., \hat{\beta}_0$ and $\hat{\beta}_1)$

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

Step 2. $$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

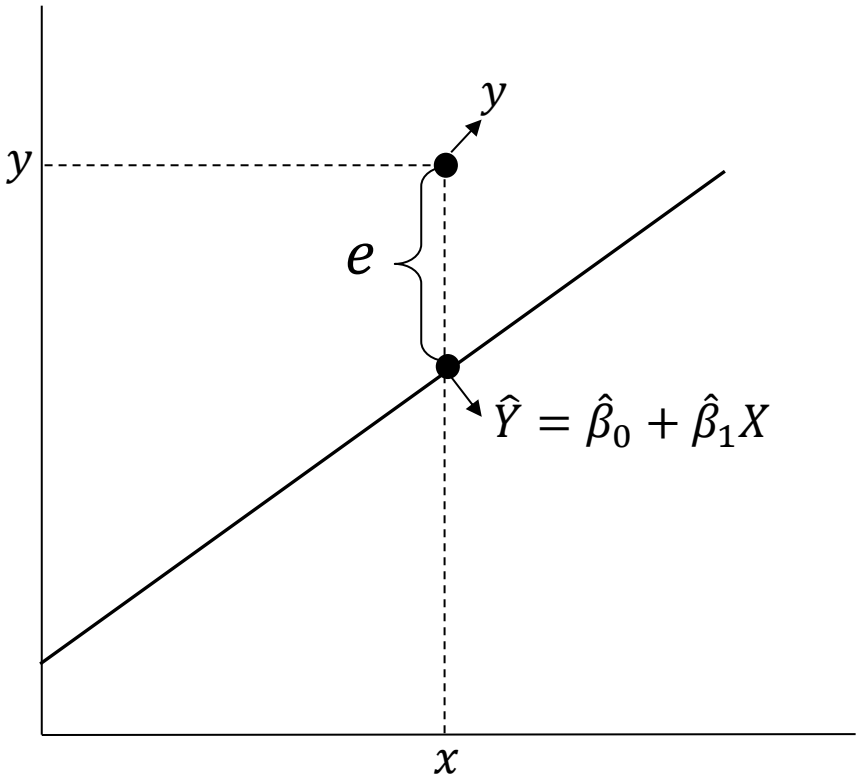
Step 3. $$\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) = 0$$

$$\frac{\partial C(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) X_i = 0$$

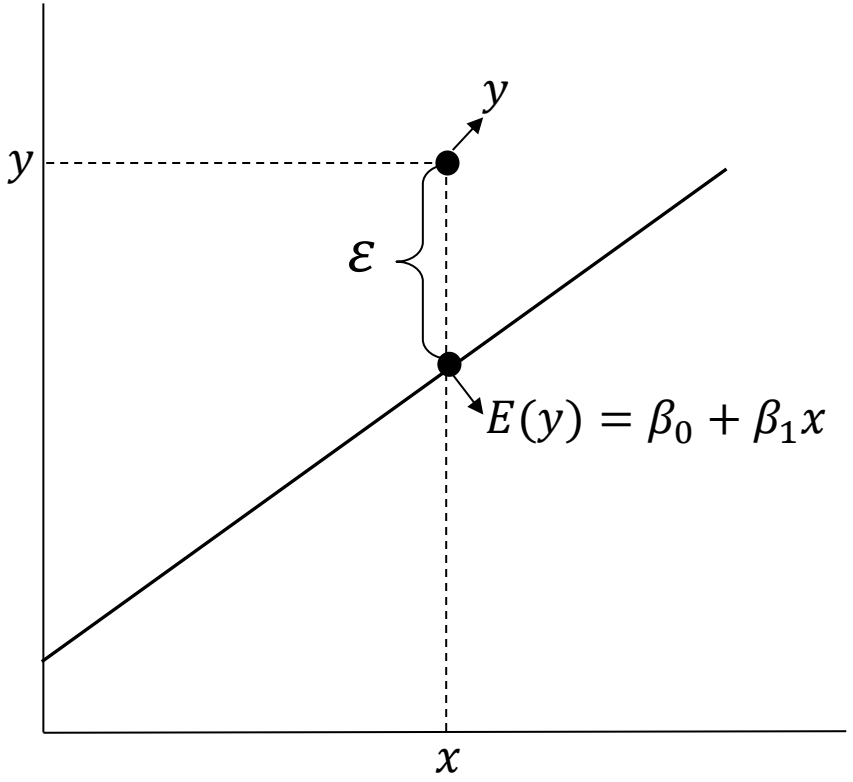
Solutions. $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Algorithm: Least Squares Estimation Algorithm (최소제곱법, 최소자승법)

잔차 (Residual)



$$e = y - \hat{y}$$

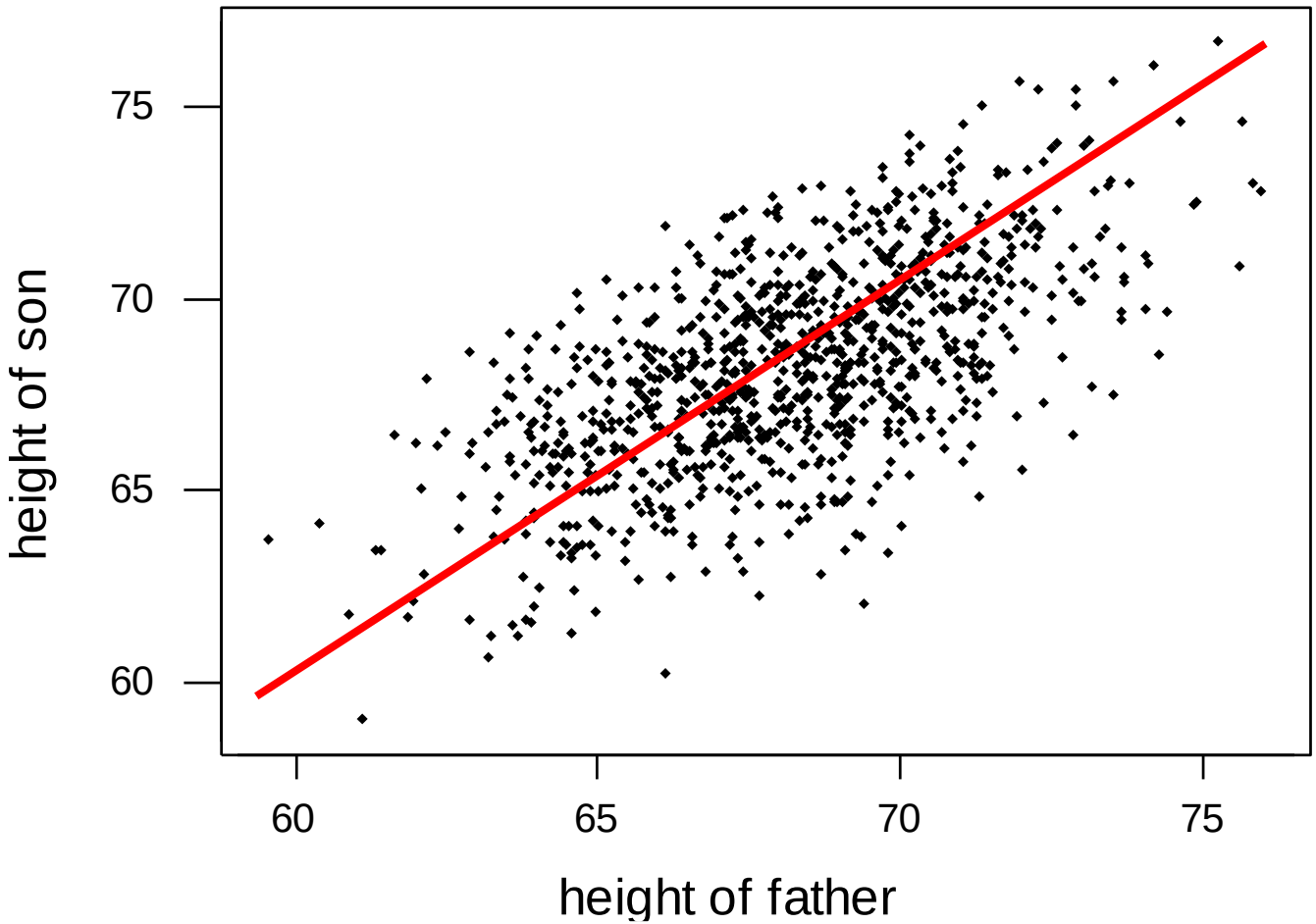


$$\varepsilon = y - E(y)$$

- 잔차 e 는 확률오차 ε이 실제로 구현된 값

선형회귀 모델

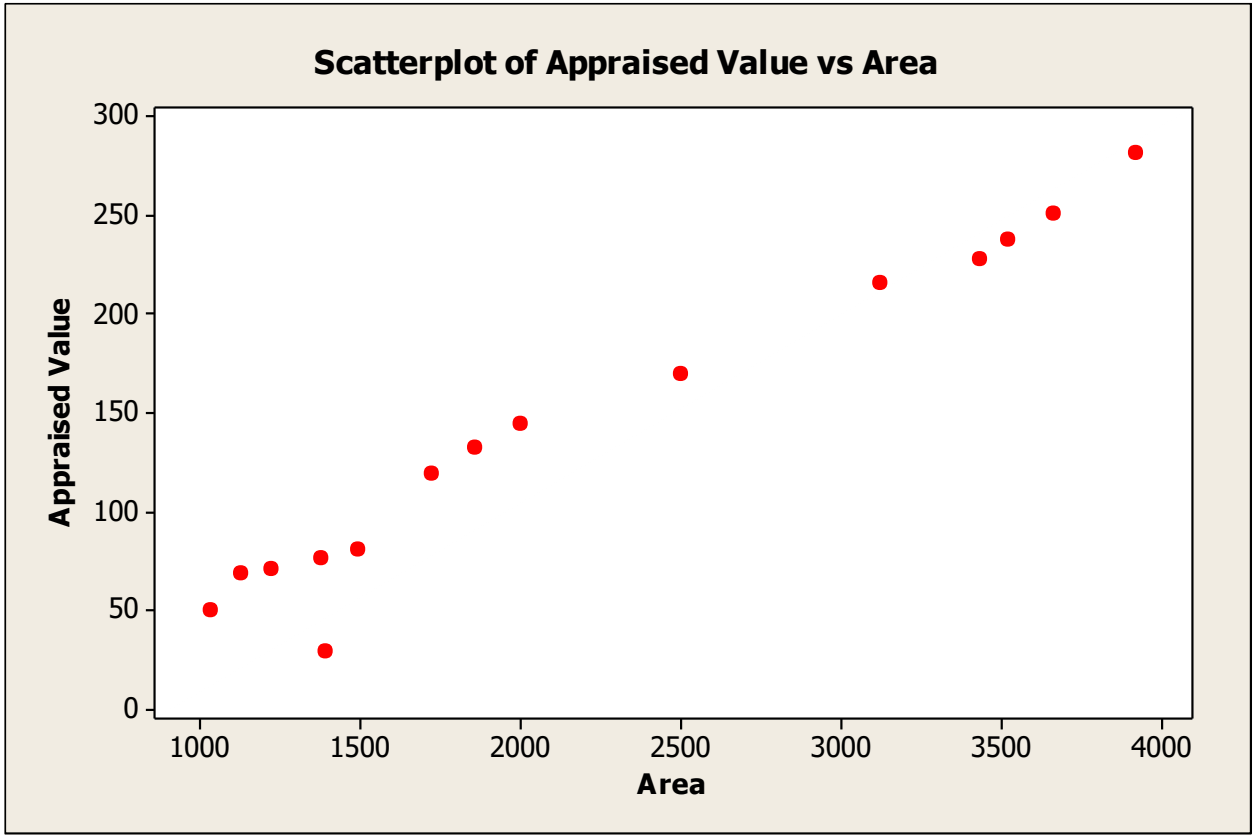
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$



선형회귀 모델 예제

관측치	Area, 집크기 (X)	Appraised Value, 집가격 (Y)
1	1380	76
2	3120	216
3	3520	238
4	1130	69
5	1030	50
6	1720	119
7	3920	282
8	1490	81
9	1860	132
10	3430	228
11	2000	145
12	3660	251
13	2500	170
14	1220	71
15	1390	29

선형회귀분석 모델 예제


Scatterplot of Appraised Value vs Area
Appraised Value
300
250
200
150
100
50
0
1000
1500
2000
2500
3000
3500
4000
Area

선형회귀 모델 예제

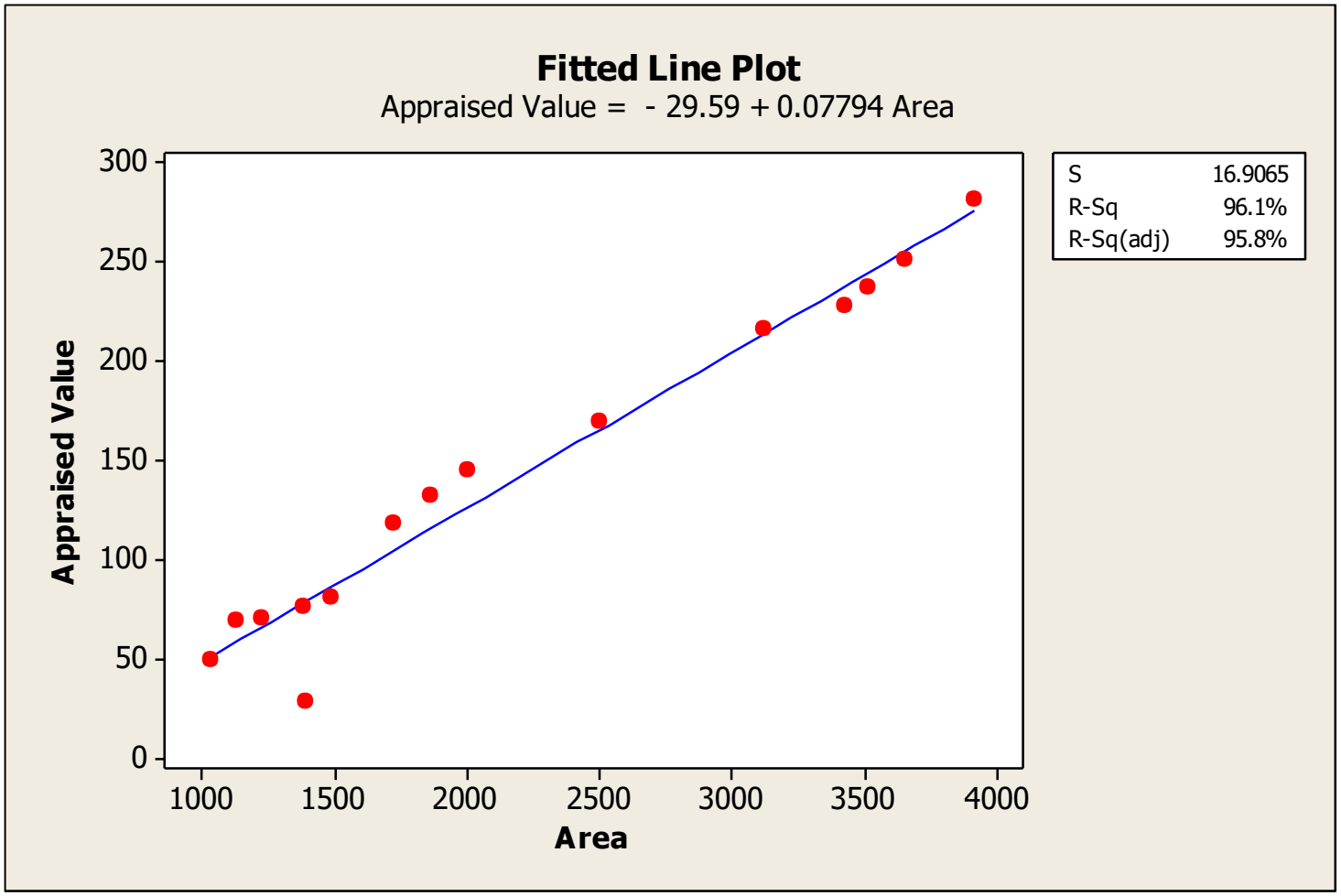
The regression equation is

Appraised Value (집 가격)= - 29.6 + 0.0779 Area (집 크기)

$\hat{\beta}_1$ = 0.077939

집 크기가 1 square feet 증가할 때마다 집 가격은 0.077 ($ thousand) 증가

선형회귀 모델 예제


Fitted Line Plot
Appraised Value = - 29.59 + 0.07794 Area
S 16.9065
R-Sq 96.1%
R-Sq(adj) 95.8%
Appraised Value
Area
300
250
200
150
100
50
0
1000
1500
2000
2500
3000
3500
4000

최소제곱법을 구한 파라미터

Cost function (비용함수)

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

Least Squares Estimation Algorithm
(최소제곱법)

$$\frac{\partial C(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i) = 0$$

$$\frac{\partial C(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} Y_i - (\beta_0 + \beta_1 X_i)X_i = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Least square estimator

파라미터 추정 알고리즘

Least square estimator

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- Estimator(추정량): 샘플의 함수 (a function of the samples)
- 추정량의 용도: 알려지지 않은 파라미터를 추정
- 추정량의 종류

  (1) 점추정 (point estimator), (2) 구간추정 (interval estimator)

파라미터에 대한 점추정 (Point Estimator)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n$$

β_0 에 대한 점추정 식: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

β_1 에 대한 점추정 식: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

σ^2 에 대한 점추정 식: $\hat{\sigma}^2 = \left(\dfrac{1}{n-2}\right)\sum_{i=1}^{n} e_i^2$

최소제곱법 추정량 성질

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Gauss-Markov Theorem: Least square estimator is the best linear unbiased estimator (BLUE)

BLUE: The BLUE is (1) unbiased estimator and (2) has the smallest average squared error (variance) compared to any unbiased estimators

(1) unbiased estimator $E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$

(2) smallest variance estimator $V(\hat{\beta}_0) \leq V(\tilde{\beta}_0), \quad V(\hat{\beta}_1) \leq V(\tilde{\beta}_1)$

$\tilde{\beta}$: any other linear unbiased estimators

파라미터에 대한 구간추정

- 점추정 $\hat{\beta}_0 \rightarrow \beta_0$, $\hat{\beta}_1 \rightarrow \beta_1$
- 구간추정 (Interval Estimation)
- 구간으로 추정하여 보다 유연한 정보 제공

θ (파라미터)에 대한 구간추정 기본 형태

$\hat{\theta}$ −상수값·표준편차($\hat{\theta}$) $\leq \theta \leq$ $\hat{\theta}$ +상수값·표준편차($\hat{\theta}$)

$\hat{\theta}$: point estimator of θ

기울기에 대한 신뢰구간

β_1 에 대한100(1-α)% 신뢰구간, n=관측치 수

$$\hat{\beta}_1 - t_{\alpha/2,n-2} \cdot sd\{\hat{\beta}_1\} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} \cdot sd\{\hat{\beta}_1\}$$

1 2 3

1. $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$: β_1의 point estimator

2. $t_{\alpha/2,n-2}$: 유의수준 1-α하에서 자유도가 n-2인 t 분포의 값

3. $sd\{\hat{\beta}_1\} = \sqrt{\dfrac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$: $\hat{\beta}_1$의 표준편자

Y절편에 대한 신뢰구간

β_0 에 대한100(1-α)% 신뢰구간, n=관측치 수

$$\hat{\beta}_0 - t\alpha_{/2,n-2} \cdot sd\{\hat{\beta}_0\} \leq \beta_0 \leq \hat{\beta}_0 + t\alpha_{/2,n-2} \cdot sd\{\hat{\beta}_0\}$$

1 2 3

1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$: β_0의 point estimator

2. $t\alpha_{/2,n-2}$: 유의수준 1-α하에서 자유도가 n-2인 t 분포의 값

3. $sd\{\hat{\beta}_0\} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]}$: $\hat{\beta}_0$의 표준편자

기울기에 대한 가설검정

- 알려지지 않은 파라미터에 대한 가설을 세우고 이를 검정
- 일종오류 α 하에서 기울기가 0인지 아닌지 검정

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0$$

$$t^* = \frac{\hat{\beta}_1 - 0}{sd\{\hat{\beta}_1\}}$$

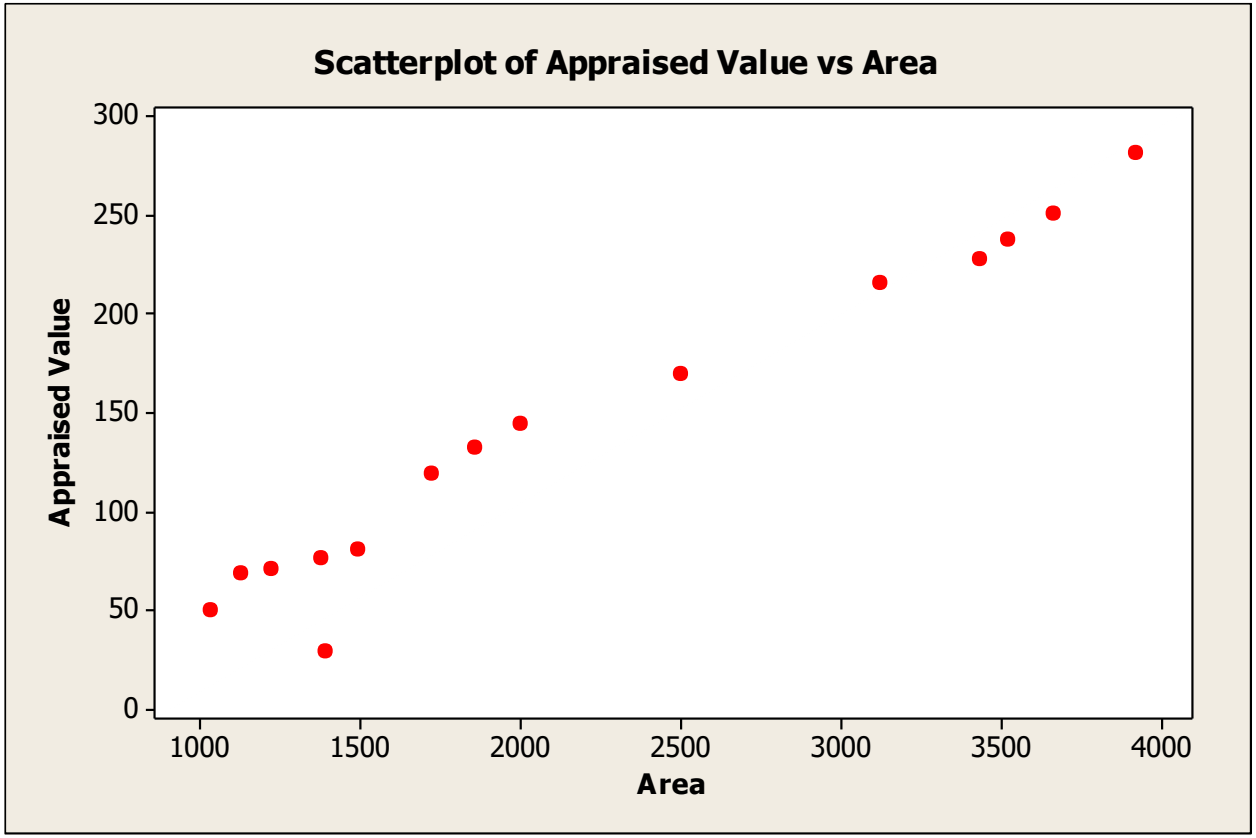
If $|t^*| > t_{\alpha/2,n-2}$, we reject H_0

P-value = $2 \cdot P(T > |t^*|)$ where $T \sim t(n-2)$

선형회귀 모델 예제

관측치	Area, 집크기 (X)	Appraised Value, 집가격 (Y)
1	1380	76
2	3120	216
3	3520	238
4	1130	69
5	1030	50
6	1720	119
7	3920	282
8	1490	81
9	1860	132
10	3430	228
11	2000	145
12	3660	251
13	2500	170
14	1220	71
15	1390	29

선형회귀분석 모델 예제


Scatterplot of Appraised Value vs Area
Appraised Value
300
250
200
150
100
50
0
1000
1500
2000
2500
3000
3500
4000
Area

선형회귀 모델 예제

The regression equation is
Appraised Value (집가격)= - 29.6 + 0.0779 Area (집크기)

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.000

S = 16.9065

- What are the parameters?
- What are the point estimates of the parameters?
- What is the standard deviation (standard error) of the parameter?
- What is the T in the above table?
- What is the P in the above table?
- What is the S in the above table?

선형회귀 모델 예제

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.000

S = 16.9065

- What are the parameters?

- What are the point estimates of the parameters?

선형회귀 모델 예제

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.000

S = 16.9065

- What are the standard deviations(standard errors) of the point estimator?

선형회귀 모델 예제

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.000

S = 16.9065

- What is the T in the above table?

선형회귀 모델 예제

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.000

S = 16.9065

- What is the P in the above table?

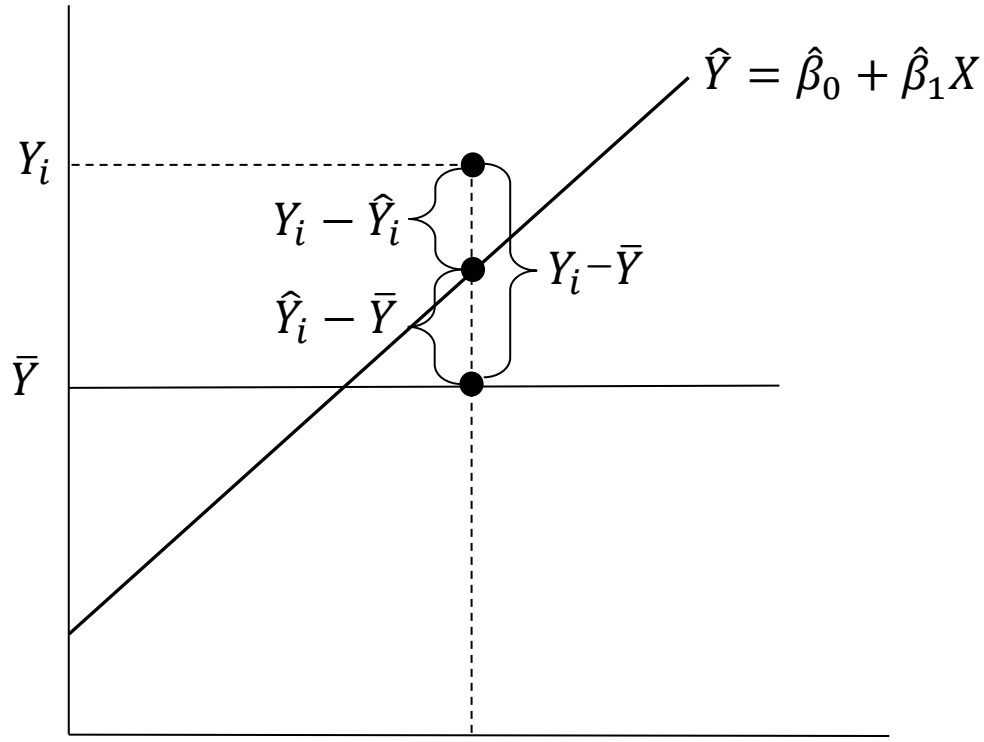
선형회귀 모델 예제

Predictor	Coef	SE Coef	T	P
Constant	-29.59	10.66	-2.78	0.016
Area	0.077939	0.004370	17.83	0.000

S = 16.9065

- What is the S in the above table?

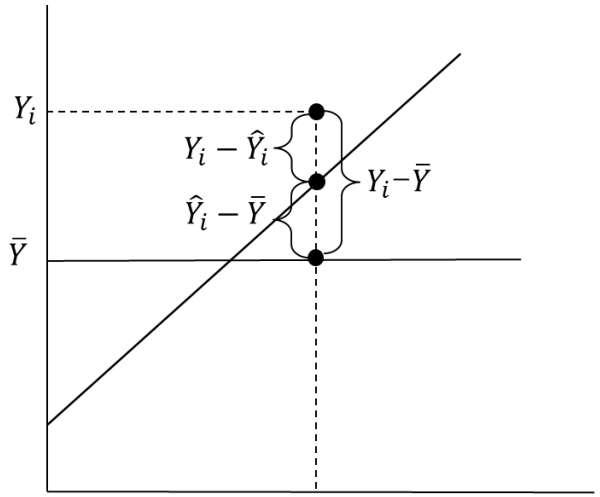
결정계수 (Coefficient of Determination: R^2)



$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \qquad SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 \qquad SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SST = SSR + SSE$$

결정계수 (Coefficient of Determination: R^2)



$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \qquad SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$SST = SSR + SSE$$

$$\frac{SSR}{SST} = 1 \qquad \frac{SSR}{SST} = 0$$

$$\frac{SSR}{SST} = R^2$$

결정계수 (R^2)

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

- R^2는 0과 1사이에 존재
- R^2=1: 현재 가지고 있는 X변수로 Y를 100% 설명. 즉, 모든 관측치가 회귀직선 위에 있음
- R^2=0: 현재 가지고 있는 X변수는Y설명(예측)에 전혀 도움이 되지 않음
- 사용하고 있는 X변수가 Y변수의 분산을 얼마나 줄였는지 정도
- 단순히Y의 평균값을 사용했을 때 대비 X 정보를 사용함으로써 얻는 성능향상 정도
- 사용하고 있는 X변수의 품질

수정 결정계수 (Adjusted R^2)

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

$$R^2_{adj} = 1 - \left[\frac{n-1}{n-(p+1)}\right]\frac{SSE}{SST}$$

$$R^2_{adj} \leq R^2$$

- R^2는 유의하지 않은 변수가 추가되어도 항상 증가
- 수정 R^2 는 앞에 특정 계수를 곱해 줌으로써 (보정) 유의하지 않은 변수가 추가 될 경우 증가하지 않게 함
- 설명변수가 서로 다른 회귀모형의 설명력을 비교할 때 사용

선형회귀모델 예제

전국적으로 500개의 대리점을 가지고 있는 요플레 제조회사에서
각 대리점의 판매원 수와 광고비 지출이 매출액에 어떤 영향을 미치는가를
알아 보기 위해 10개의 대리점에 대한 자료를 수집하였다

판매원 수 (X_1)	광고비 (X_2)	월간 매출액 (Y)
14	37	850
16	43	970
13	38	730
10	42	940
18	36	920
17	33	830
16	40	940
15	35	900
11	34	760
10	29	710

선형회귀모델 예제

Variable	추정치	T	P-value
(Constant)	141.516	.706	.472
판매원 수, X_1	13.035	1.854	.106
광고비, X_2	14.469	3.025	.019

SSR=54809.18, SSE=25440.82, SST=80250.00

$$R^2 = \frac{SSR}{SST} = \frac{54809.18}{80250.00} = 0.683$$

- 판매원 수와 광고비 변수에 의해 매출액 변수의 변동성을 68.3% 감소
- 매출액의 (단순)평균 대비 판매원 수와 광고비를 이용하면 설명력이 68.3%증가
- 현재 분석에 사용하고 있는 판매원 수와 광고비의 “변수 품질”정도가 68.3(100점 기준)

EOD