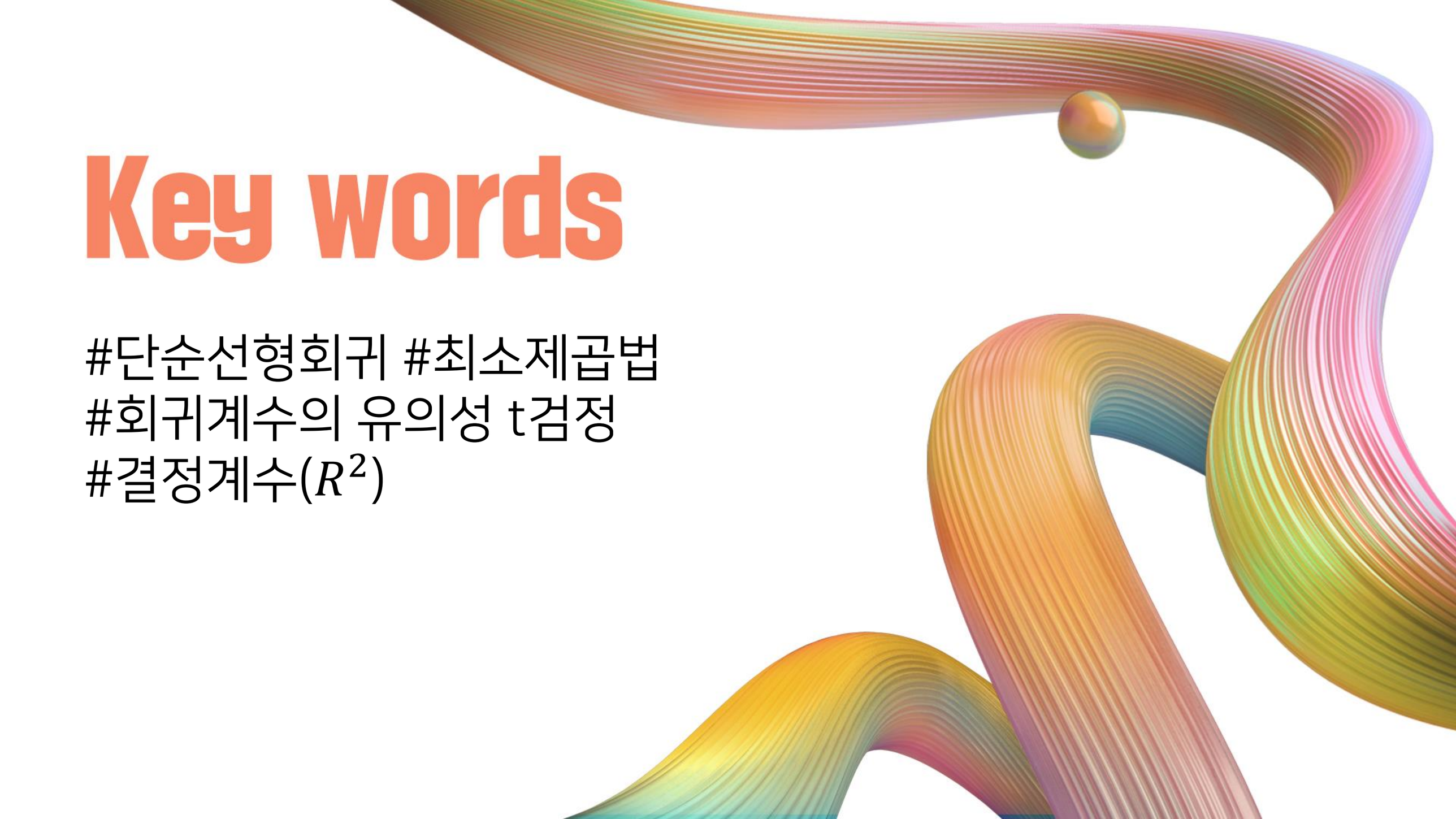
단순회귀분석
(Simple Linear Regression)

Key words

#단순선형회귀 #최소제곱법
#회귀계수의 유의성 t검정
#결정계수(R^2)



회귀분석 개요

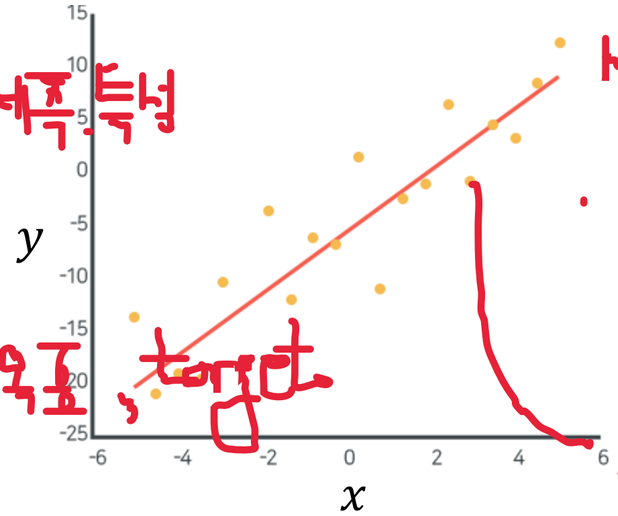
회귀분석

독립변수와 종속변수 간의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 자료로부터 통계적으로 추정하는 분석방법.

회귀분석 개요

회귀분석

- $y=f(x)$의 함수 관계가 있을 때,
  - x를 설명변수(explanatory variable) 또는 독립변수(independent variable)
    - 단순 회귀 : 독립변수가 1개
    - 다중 회귀 : 독립변수가 2개 이상
  - y를 반응변수(response variable) 또는 종속변수(dependent variable)



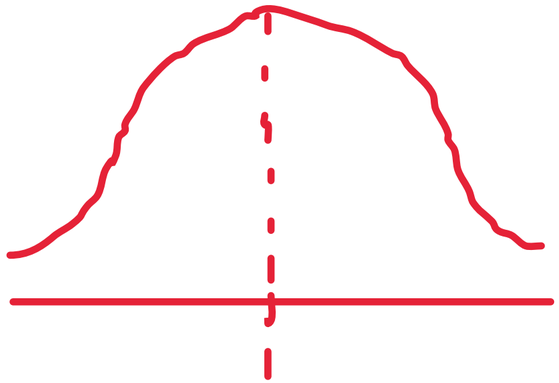
단순선형회귀모형

모형 정의 및 가정

- 자료(x_i, Y_i), i=1,⋯, n에 다음의 관계식이 성립한다고 가정함.

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad , i = 1,2,\dots,n$$

- 오차항인 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$는 서로 독립인 확률변수로, $\varepsilon_i \sim N[0, \sigma^2]$
  : 정규, 등분산, 독립 가정
- α,β 는 회귀계수라 부르며 α는 절편, β은 기울기를 나타냄.
- α,β, σ^2은 미지의 모수로, 상수임.

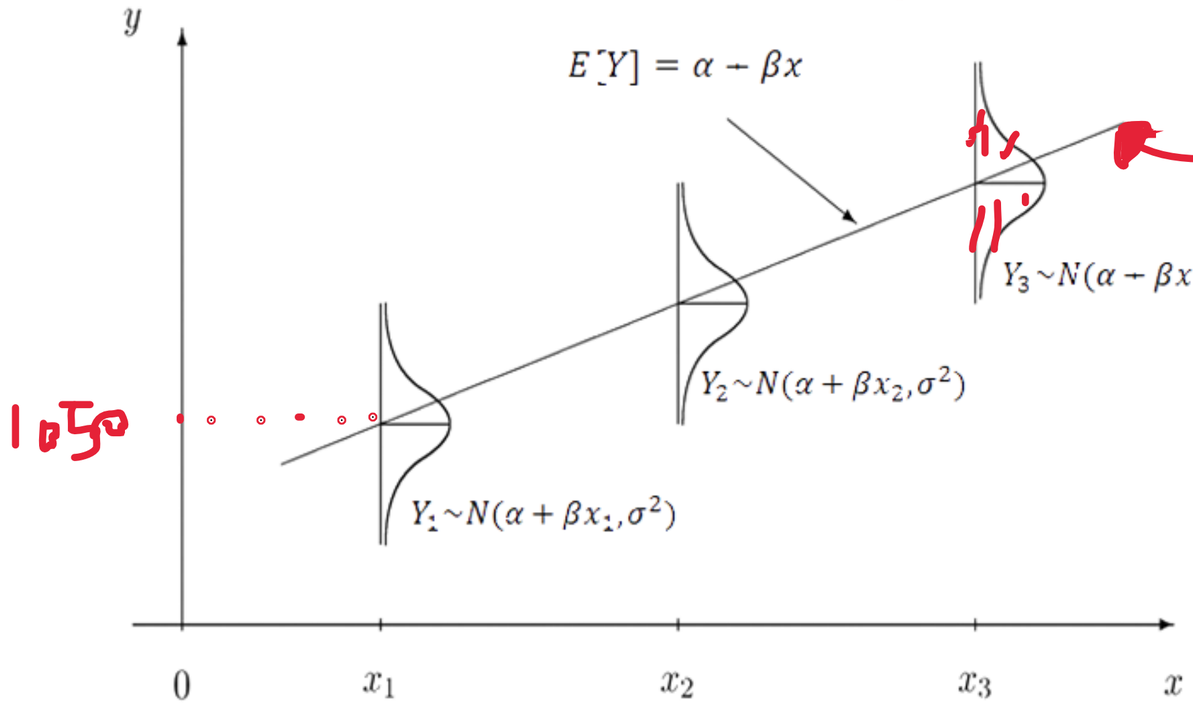


단순선형회귀모형

모형 정의 및 가정

- 자료(x_i, Y_i), $i=1,\cdots,n$ 에 다음의 관계식이 성립한다고 가정함.

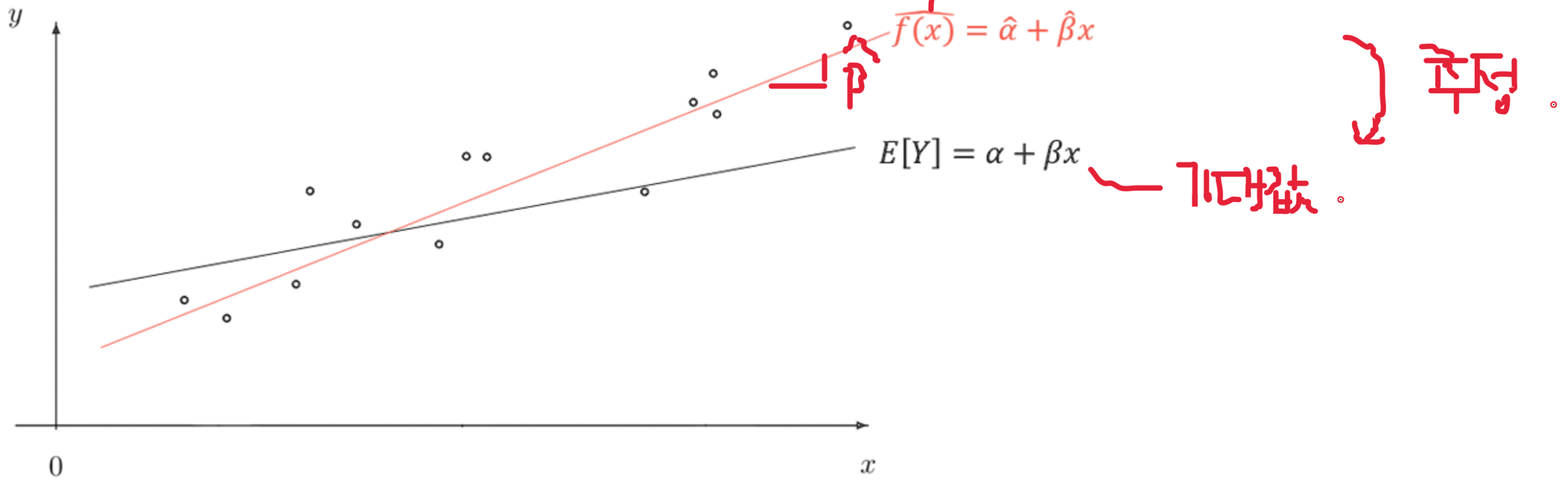
$$Y_i \sim N[\alpha+\beta x_i, \sigma^2] \rightarrow E[Y_i] = \alpha+\beta x_i$$



단순선형회귀모형의 모수 추정

모수 추정

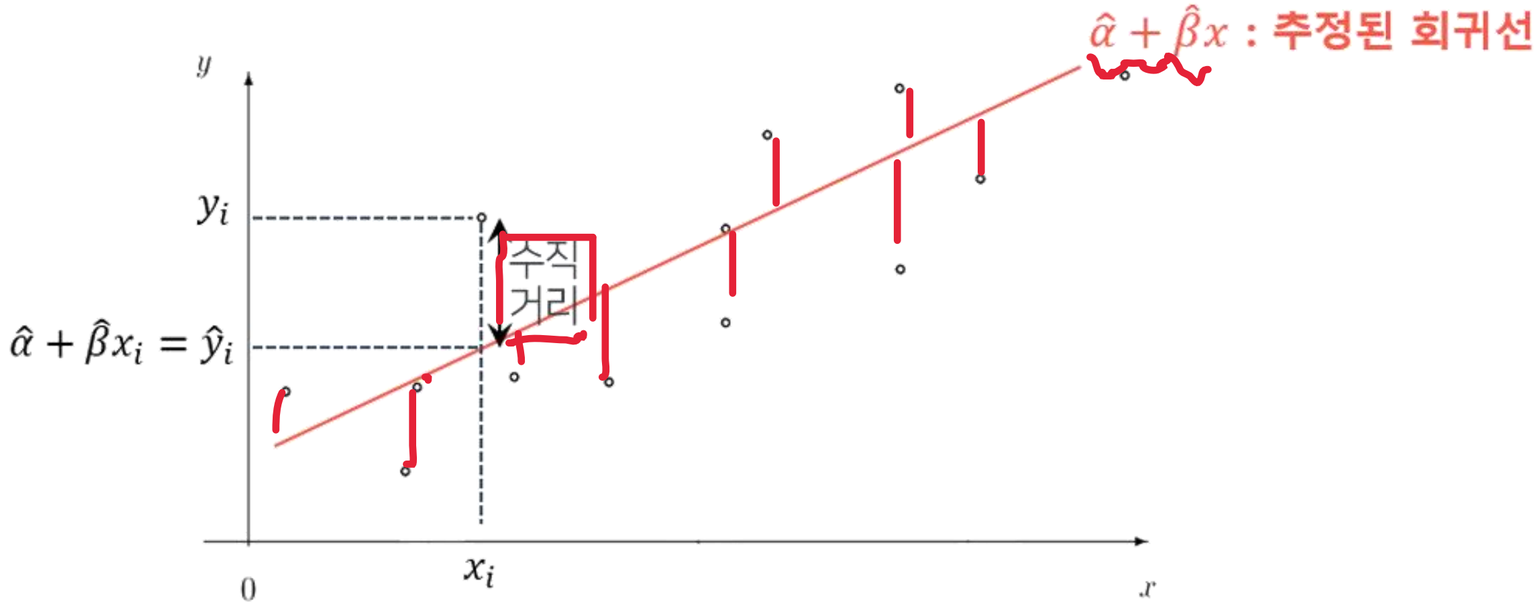
- 모형이 포함한 미지의 모수 α, β 를 추정하기 위하여 각 독립변수 x_i에 대응하는 종속변수 y_i로 짝지어진 n 개의 표본인 관측치 (x_i, y_i) 가 주어짐.



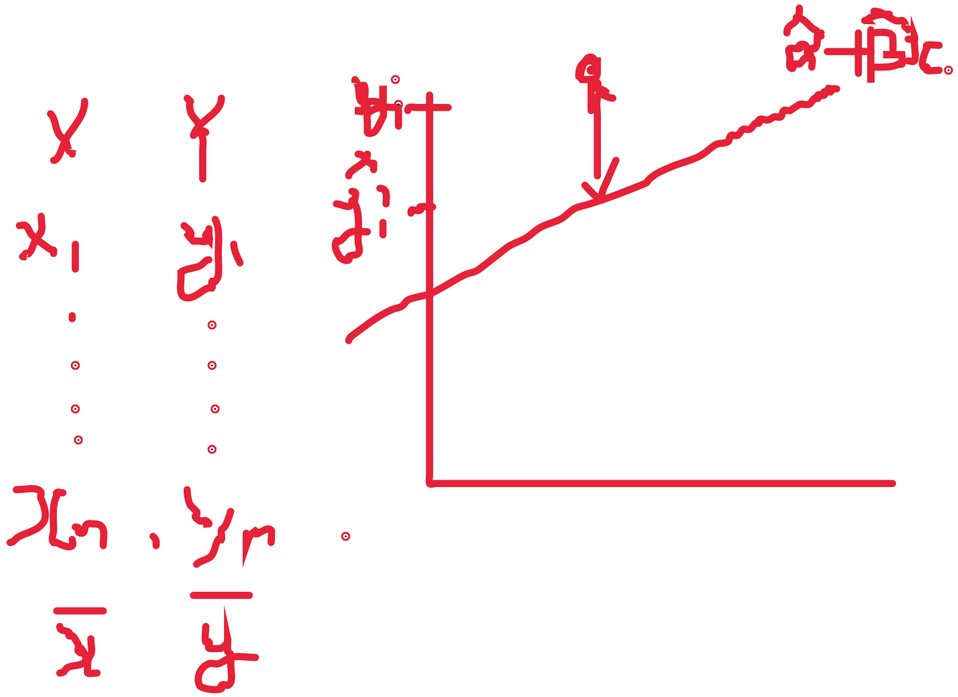
단순선형회귀모형의 모수 추정

최소제곱법

- 단순회귀모형 $Y_i = \alpha + \beta x_i + \varepsilon_i$에서 자료점과 회귀선 간의 수직거리 제곱합 $SS(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$이 최소가 되도록 α와 β를 추정하는 방법



단순선형회귀모형의 모수 추정



최소제곱법

- α에 대한 최소제곱 추정량 : $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
- β에 대한 최소제곱 추정량 : $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}$

(단, $\bar{x}$는 x_i의 평균, $\bar{y}$는 y_i의 평균)

- y_i의 추정치 : $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \quad , i = 1,2,\ldots,n$
- 잔차 : $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i \quad , i = 1,2,\ldots,n$

단순선형회귀모형 사례

예제

- 자동차의 주행속도와 정지거리에 관한 50개의 표본 자료를 이용하여, 주행거리를 독립변수로, 정지거리를 종속변수로 두고 단순선형회귀모형을 적합하고자 함.

사례	1	2	3	4	…	47	48	49	50
주행속도	4	4	7	7	…	24	24	24	25
정지거리	2	10	4	22	…	92	93	120	85

단순선형회귀모형 사례

예제

- 회귀계수의 추정 $\bar{x} = 15.4,\ \bar{y} = 42.98$이므로,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = 3.932$$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -17.579$ 로 계산됨.

단순선형회귀모형 사례

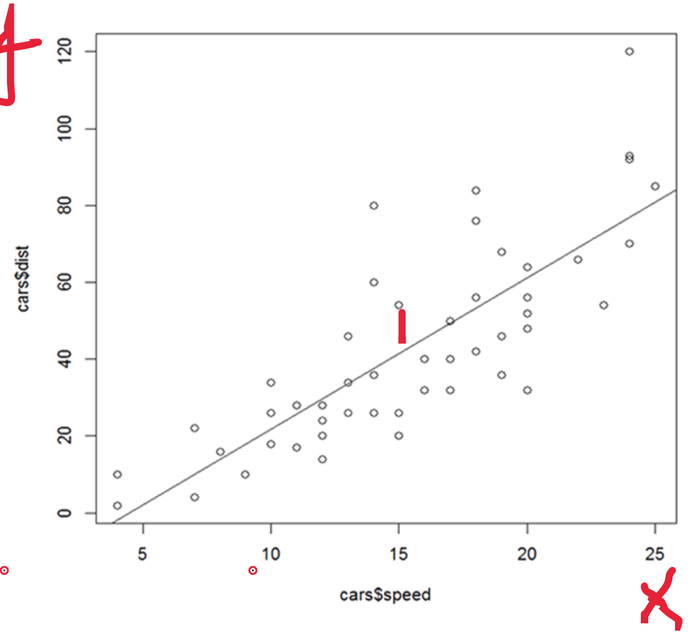
예제

- 추정된 회귀선

$$\hat{y} = -17.579 + 3.932x$$

- 결정계수

$$R^2 = \frac{SSR}{SST} = 0.6511$$



단순선형회귀모형의 유의성 검정

모형의 유의성 t 검정

- 독립변수 x가 종속변수 Y를 설명하기에 유용한 변수인가에 대한 통계적 추론은 회귀계수 β에 대한 검정을 통해 파악할 수 있음.

- 가설

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

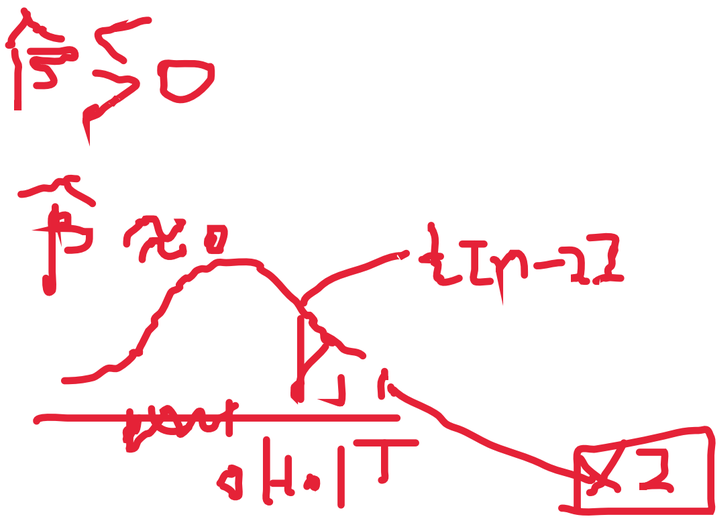
단순선형회귀모형의 유의성 검정

모형의 유의성 t 검정

- 검정통계량과 표본분포

  - 귀무가설 H_0이 사실일 때,

$$T = \frac{\hat{\beta}}{\widehat{S.E.}[\hat{\beta}]} \sim t[n-2]$$



$$|T| = \left|\frac{\hat{\beta}}{\widehat{S.E.}[\hat{\beta}]}\right| > t_{\alpha/2,n-2}$$ 또는 p-value (=P[T>|t_0|] × 2) < α 면

귀무가설을 기각.

→ 독립변수 x가 종속변수 Y를 설명하기에 유용한 변수라고 해석할 수 있음.

단순선형회귀모형 사례

예제

- 베타에 관한 유의성을 유의수준 5%로 검정할 것.

	추정치	표준오차	T 통계량	p-value
절편	-17.5791	6.7584	-2.601	0.0123
주행속도	3.9324	0.4155	9.464	1.49E-12

- 가설 $H_0 : \beta = 0$
  $H_1 : \beta \neq 0$
- 귀무가설(H_0)이 사실일 때,
- 검정통계량의 관찰값(x_0)는 9.464로 $t[48]$의 분포에서 유의확률은 1.49E-12
- p-value(=1.49E-12) ≤ α(=0.05) 이므로, H_0를 기각.

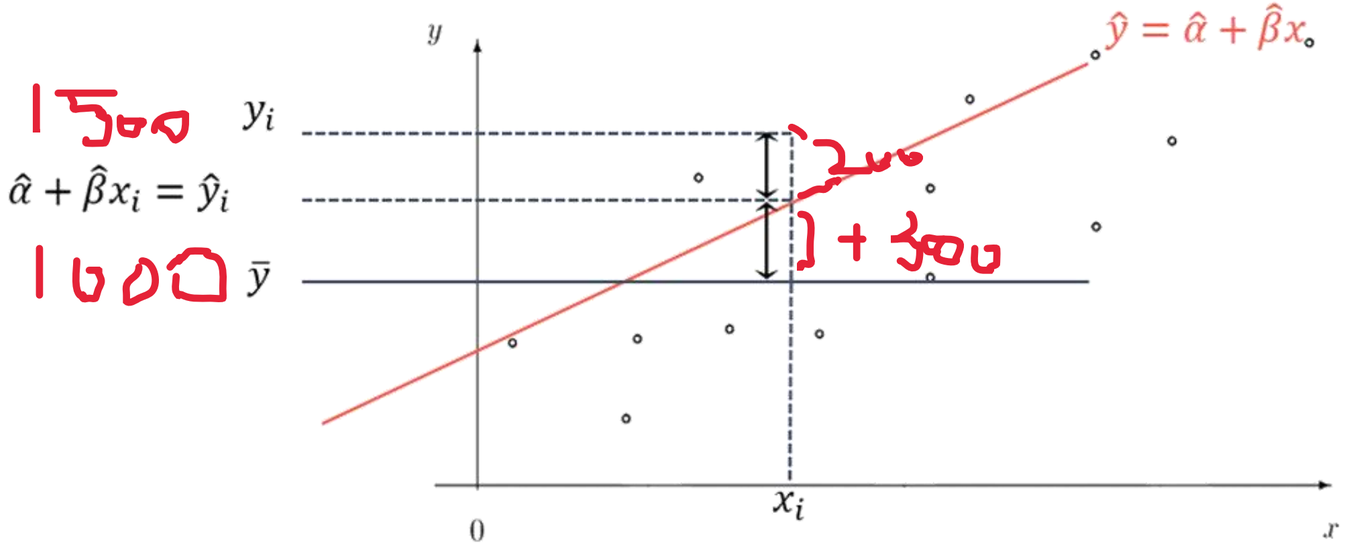
단순선형회귀모형의 적합도

Y의 변동성 분해

- 제곱합 :

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

SST (y_i의 변동) = SSR (모형으로 설명되는 변동) + SSE (모형으로 설명되지 않는 변동)



단순선형회귀모형의 적합도

모형의 적합성

- 결정계수 R^2

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

- $SST = SSR + SSE$ 이므로 항상 0과 1 사이의 값을 가짐 ($0 \le R^2 \le 1$).
- y_i의 변동 가운데 추정된 회귀모형으로 통해 설명되는 변동의 비중을 의미함.
- 0에 가까울 수록 추정된 모형의 설명력이 떨어지는 것으로,
  1에 가까울수록 추정된 모형이 y_i의 변동을 완벽하게 설명하는 것으로 해석할 수 있음.
- R^2는 두 변수 간의 상관계수 r의 제곱과 같음.

다중회귀분석
(Multiple Linear Regression)

Key words

#다중선형회귀모형
#범주형 독립변수 #더미변수

다중선형회귀모형

다중선형회귀모형으로의 확장

- 다중 선형회귀모형
  - 독립변수가 두 개 이상인 선형회귀모형.
  - 여러 개의 독립변수를 이용하면 종속변수의 변화를 더 잘 설명할 수 있을 것임.
  - 자료 $((x_{1i}, x_{2i}, \cdots, x_{ki}), Y_i),\ i = 1, \cdots, n$ 에 다음의 관계식이 성립한다고 가정함.

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \ (i = 1,2, \ldots, n)$$

다중선형회귀모형

다중선형회귀모형으로의 확장

- 다중 선형회귀모형
  - 오차항인 $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ 서로 독립인 확률변수로, $\varepsilon_i \sim N[0, \sigma^2]$
    : 정규, 등분산, 독립.
  - 회귀계수 $\alpha, \beta_1, \ldots, \beta_k$, 와 σ^2 은 미지인 모수로 상수임.
    - β_j 의 해석 : x_j 를 제외한 나머지 모든 예측변수들을 상수로 고정시킨 상태에서 x_i 의 한 단위 증가에 따른 $E[Y]$ 의 증분을 의미 $(j = 1, \ldots, k)$.

다중선형회귀모형

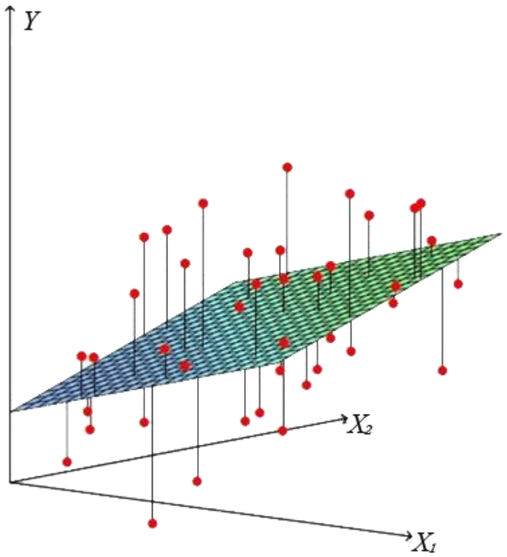
회귀계수 $\alpha, \beta_1, \dots, \beta_k$ 의 추정

- 수직거리 제곱합

$$SS(\alpha, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$$

이 최소가 되도록 $\alpha, \beta_1, \dots, \beta_k$ 를 추정.

- 최소제곱 추정량 : $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_k$



다중선형회귀모형

다중회귀모형 분석 예시

- 현재 영업중인 La Quinta Inn 호텔 중에서 랜덤하게 100곳의 영업자료를 수집.
- 다중회귀 모형의 설정

$$Margin = \alpha + \beta_1 Number + \beta_2 Nearest + \beta_3 Office + \beta_4 College + \beta_5 Income + \beta_6 Disttwn + \varepsilon$$

Margin	Number	Nearest	Office Space	Enrollment	Income	Distance
55.5	3203	4.2	549	8	37	2.7
33.8	2810	2.8	496	17.5	35	14.4
49	2890	2.4	254	20	35	2.6
31.9	3422	3.3	434	15.5	38	12.1
57.4	2687	0.9	678	15.5	42	6.9
49	3759	2.9	635	19	33	10.8
…	…	…	…	…	…	…

다중선형회귀모형

다중회귀모형 분석 예시

- 다중선형회귀모형 추정결과

$$Margin = 38.14 - 0.0076\,Number + 1.65\,Nearest + 0.020\,Office\,Space + 0.21\,Enrollment + 0.41\,Income - 0.23\,Distance$$

- $\hat{\alpha}$= 38.14. 해석하지 않음.
- $\hat{\beta}_1$ = −0.0076. 다른 변수는 고정되어 있고, 반경 3마일이내 호텔 객실이 1개 증가하면, 영업이익율은 평균 0.0076% 감소.
- $\hat{\beta}_2$ = 1.65. 다른 변수는 고정되어 있고, 경쟁호텔과의 거리가 1마일 증가하면, 영업이익율은 평균 1.65% 증가.
- $\hat{\beta}_3$ = 0.02. 다른 변수는 고정되어 있고, 사무실 면적이 100평방피트 증가하면, 영업이익율은 평균 0.02% 증가.

다중선형회귀모형

다중회귀모형 분석 예시

- 다중선형회귀모형 추정결과

$$Margin = 38.14 - 0.0076\,Number + 1.65\,Nearest + 0.020\,Office\,Space + 0.21\,Enrollment + 0.41\,Income - 0.23\,Distance$$

- $\hat{\beta}_4$ = 0.21. 다른 변수는 고정되어 있고, 대학생 거주자수가 1000명 증가하면, 영업이익율은 평균 0.21% 증가.
- $\hat{\beta}_5$ = 0.41. 다른 변수는 고정되어 있고, 중간 가구소득이 $1000 높은 지역은, 영업이익율이 평균 0.41% 증가.
- $\hat{\beta}_6$ = -0.23. 다른 변수는 고정되어 있고, 다운타운 중심으로 부터 1마일 멀어질수록, 영업이익율은 평균 0.23% 감소.

다중선형회귀모형

다중회귀모형 분석 예시

- 다중선형회귀모형 추정결과

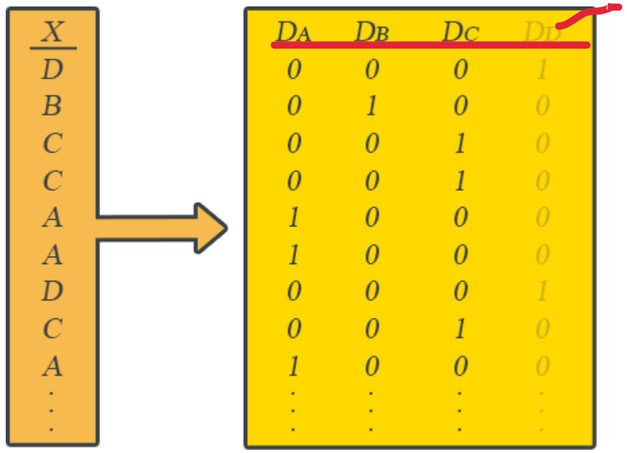
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 0.5251$$

- 영업이익율은 6개 설명변수에 의해 52.51% 설명됨.

범주형 독립변수가 포함된 회귀모형

범주형 독립변수가 포함된 회귀모형

- 범주형 독립변수를 회귀모형에 포함하기 위해서는 더미변수 (dummy variable) 기법을 사용.
- 더미변수는 0 또는 1의 값을 갖는 변수로 아래와 같이 정의됨.



X
D
B
C
C
A
A
D
C
A
⋮

D_A	D_B	D_C	D_D
0	0	0	1
0	1	0	0
0	0	1	0
0	0	1	0
1	0	0	0
1	0	0	0
0	0	0	1
0	0	1	0
1	0	0	0
⋮	⋮	⋮	⋮

더미변수의 개수
= 범주의 개수 - 1

범주형 독립변수가 포함된 회귀모형

범주형 독립변수가 포함된 회귀모형

- 범주형 독립변수 예시 : 중고차 가격에 관한 예측 모형
  - 중고차 시장에서 차량의 주행거리와 색상이 차량의 가격에 어떤 영향을 미치는지를 파악하고자 2013년형 A 브랜드 중고차 100대에 관한 자료를 수집.
  - 종속변수 : Price(가격)
  - 독립변수 : Odometer(주행거리)

범주형 독립변수가 포함된 회귀모형

범주형 독립변수가 포함된 회귀모형

- 범주형 독립변수 예시 : 중고차 가격에 관한 예측 모형
  - Color(차량색상, 범주형: white/silver/other)
  - Color의 더미변수 : 범주의 수가 3개
    → 2개의 더미변수(D_1, D_2)를 생성.

D_1 = { 1 white인 경우 / 0 white가 아닌 경우

D_2 = { 1 silver인 경우 / 0 silver가 아닌 경우

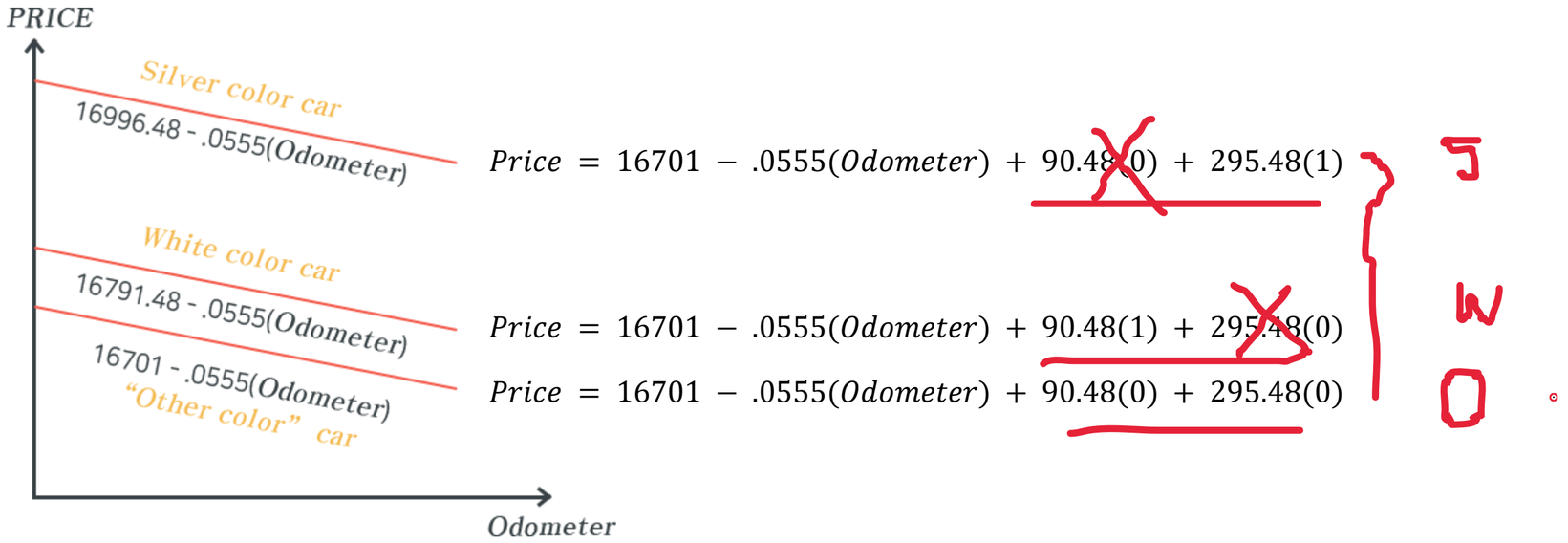
  - Other인 경우: $D_1 = 0$ & $D_2 = 0$

범주형 독립변수가 포함된 회귀모형

범주형 독립변수가 포함된 회귀모형

- 다중선형회귀모형 추정결과

$$PRICE = 16701 - .0555(Odometer) + 90.48(D_1) + 295.48(D_2)$$



$$Price = 16701 - .0555(Odometer) + 90.48(0) + 295.48(1)$$

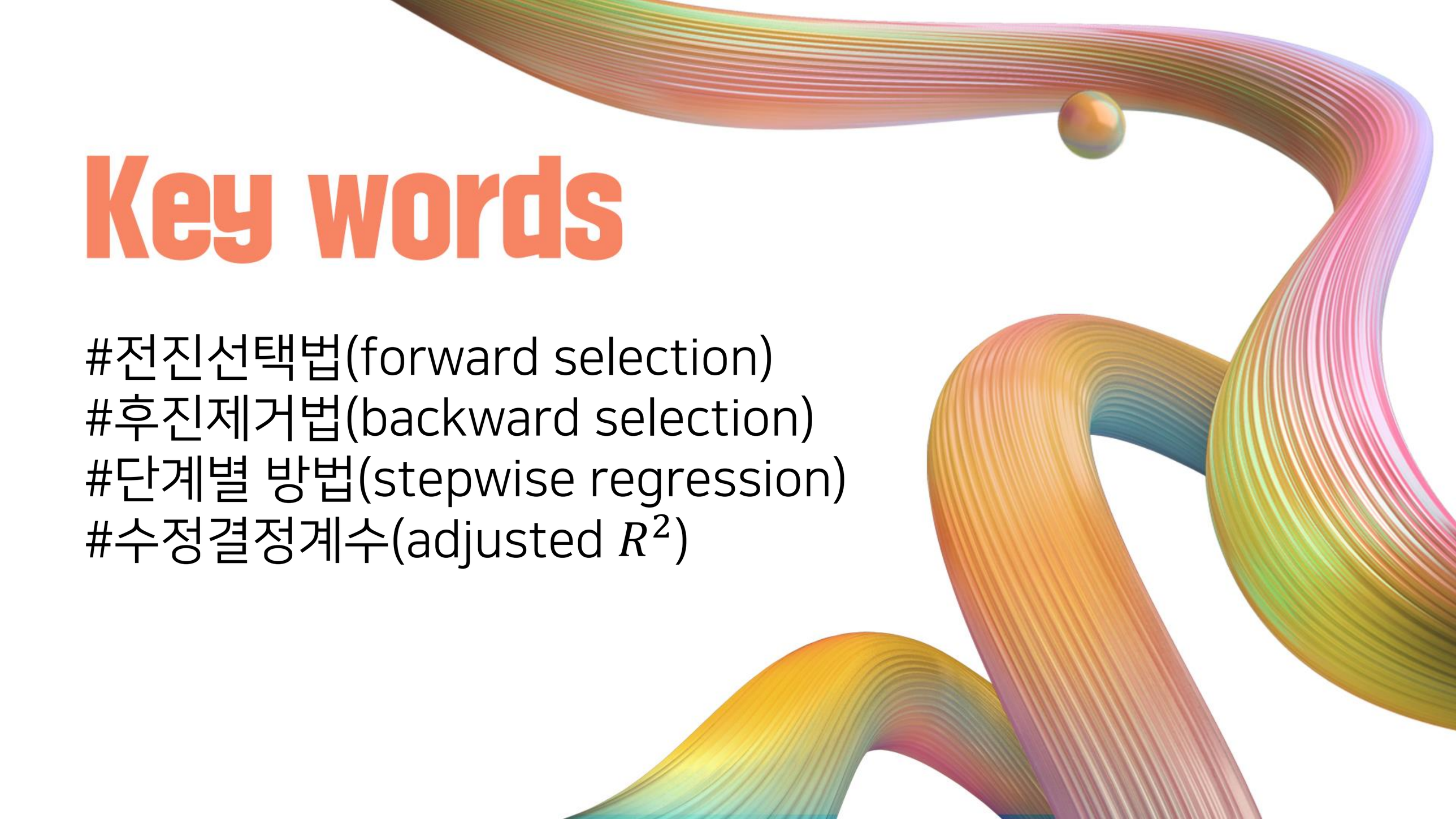
$$Price = 16701 - .0555(Odometer) + 90.48(1) + 295.48(0)$$

$$Price = 16701 - .0555(Odometer) + 90.48(0) + 295.48(0)$$

다중회귀분석(변수선택)

Key words

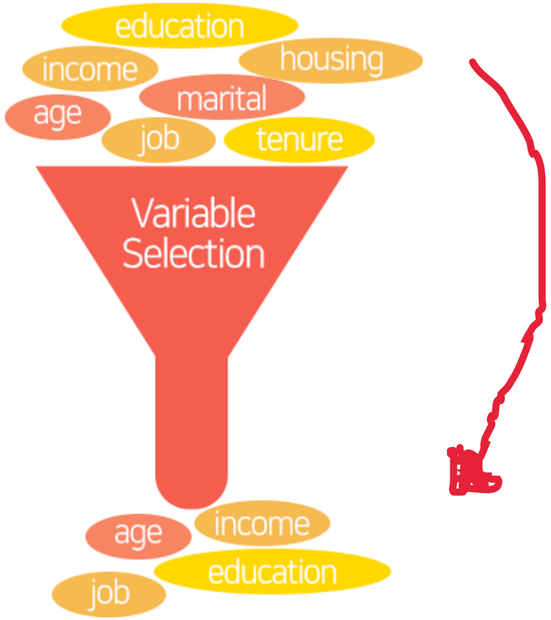
#전진선택법(forward selection)
#후진제거법(backward selection)
#단계별 방법(stepwise regression)
#수정결정계수(adjusted R^2)



다중회귀모형의 변수선택

다중회귀모형의 변수선택 개요

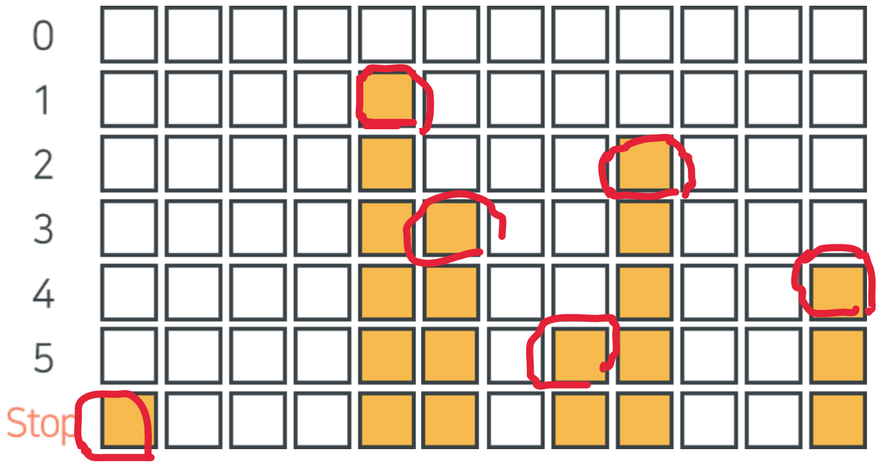
- 가능한 적은 수의 설명변수로 좋은 예측력을 가지는 모형을 찾고자 함.
- 변수선택법
  - 전진선택법(foward selection)
  - 후진제거법(backward elimination)
  - 단계선택법(stepwise method)
  - 모든 가능한 조합의 회귀분석 : 모든 가능한 독립변수들의 조합에 대한 회귀모형을 생성한 뒤 가장 적합한 회귀모형을 선택.



전진선택법

변수 선택 방법

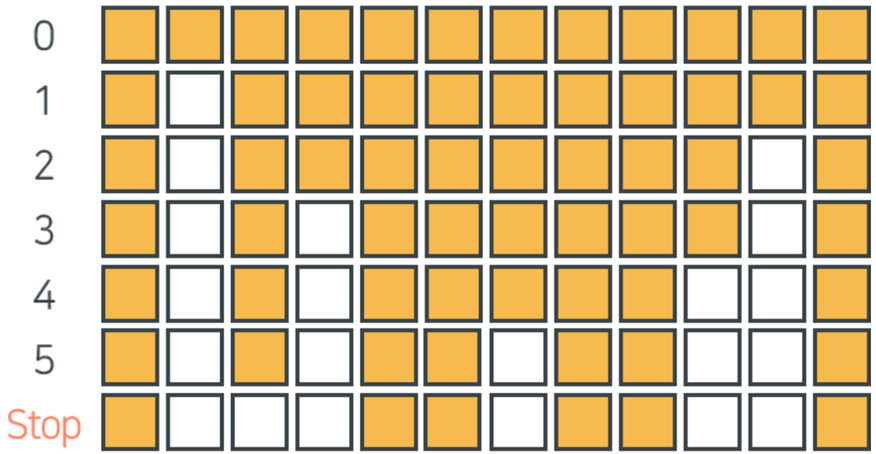
- 전진선택법(Forward selection)
  - 절편만 있는 모델에서 출발하여 중요한 변수를 하나씩 추가하는 방식.
  - 한번 선택된 변수는 제거되지 않는 단점이 있음.



후진제거법

변수 선택 방법

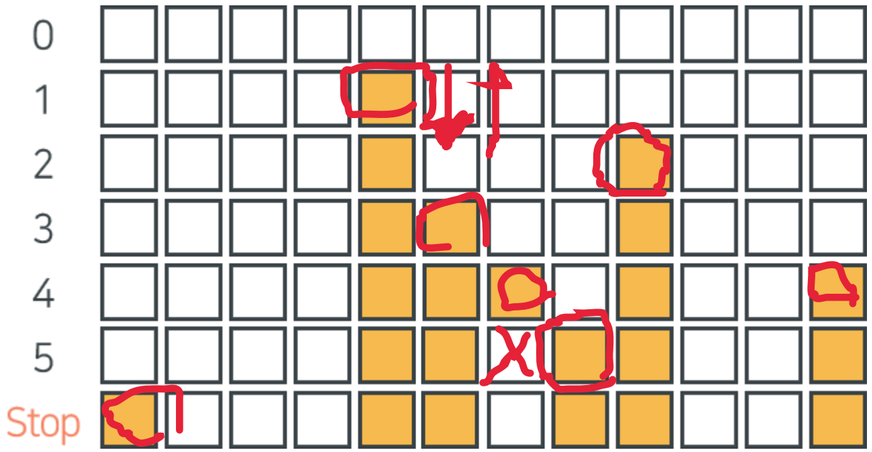
- 후진제거법(Backward elimination)
  - 모든 변수가 포함된 모델에서 가장 중요하지 않은 변수부터 하나씩 제거.
  - 한번 제거된 변수는 선택되지 않는 단점이 있음. - 같은맥락



단계별 방법

변수 선택 방법

- 단계별 방법(Stepwise method)
  - 절편만 포함된 모델에서 출발해 가장 중요한 변수부터 추가하고, 모델에 포함되어 있는 변수 중에서 중요하지 않은 변수를 제거함.
  - 더 이상 새롭게 추가되는 변수가 없을 때까지 변수의 추가 또는 삭제를 반복함.



모형 선택의 기준

모형선택의 기준

- 수정된 결정계수(Adjusted R^2)
  - 결정계수 R^2 는 새로운 독립변수가 추가되면 항상 증가함.
  - 이를 보완한 수정결정계수 $Adjusted\ R^2$ 는 추가된 독립변수가 종속변수를 설명하는데 기여하는 바가 큰 경우에만 증가함.

$$Adjusted \quad R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

- 그 밖에 AIC, BIC, Mallow's Cp 등의 다양한 적합도 지표를 이용할 수 있음.

다중회귀분석
(잔차분석, 다중공선성)

Key words

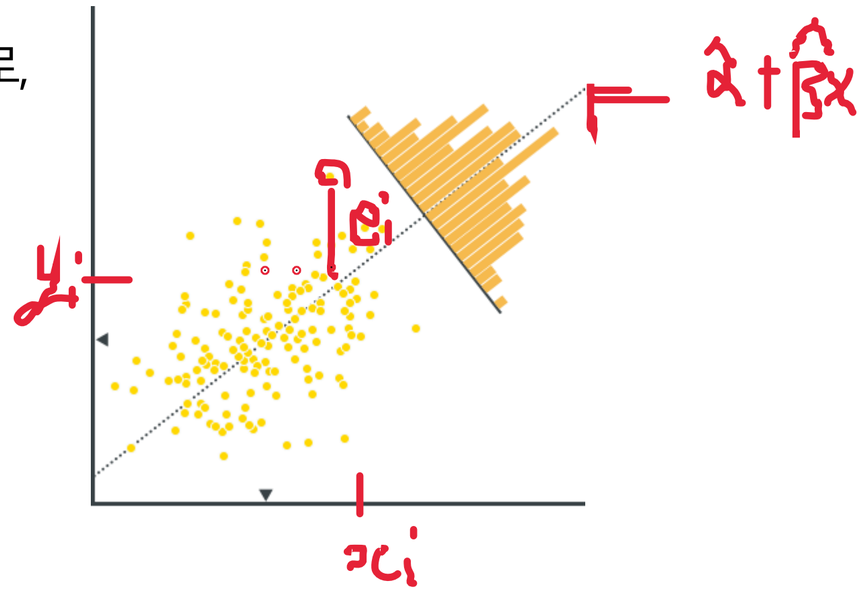
#잔차분석 #잔차산점도
#QQ플롯 #다중공선성 #VIF계수

적은수 설명변수 여러개 중요
↳ 차원축소

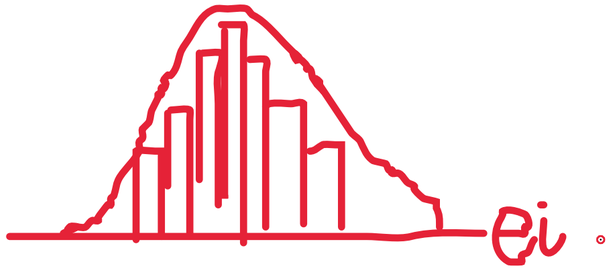
가정 위반 검토 및 해결

다중회귀모형의 가정 위반 검토 및 해결

- 잔차분석
  - 회귀 모형에서의 가정이 적절한 것인가에 대한 평가
    1) 오차의 정규성
    2) 오차의 등분산성
    3) 오차의 독립성
  - 오차는 확률변수로 관찰되지 않는 값이므로, 각 오차에 대응되는 잔차를 관찰한 뒤 잔차들의 분포를 통해 오차에 대한 가정의 적정성을 확인 할 수 있음.



가정 위반 검토 및 해결



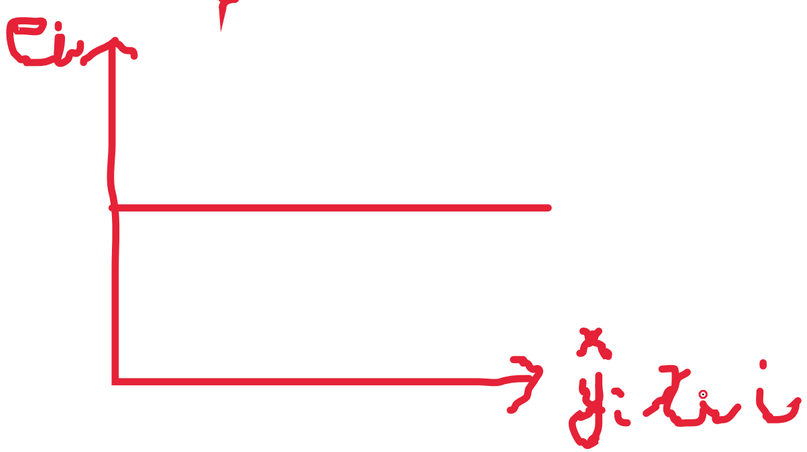
다중회귀모형의 가정 위반 검토 및 해결

- 잔차분석 방법

  - 각 가정 별로, 검정을 통한 방법과 그래프를 통한 시각적인 확인 방법이 가능.

  시각적 방법을 이용할 경우,

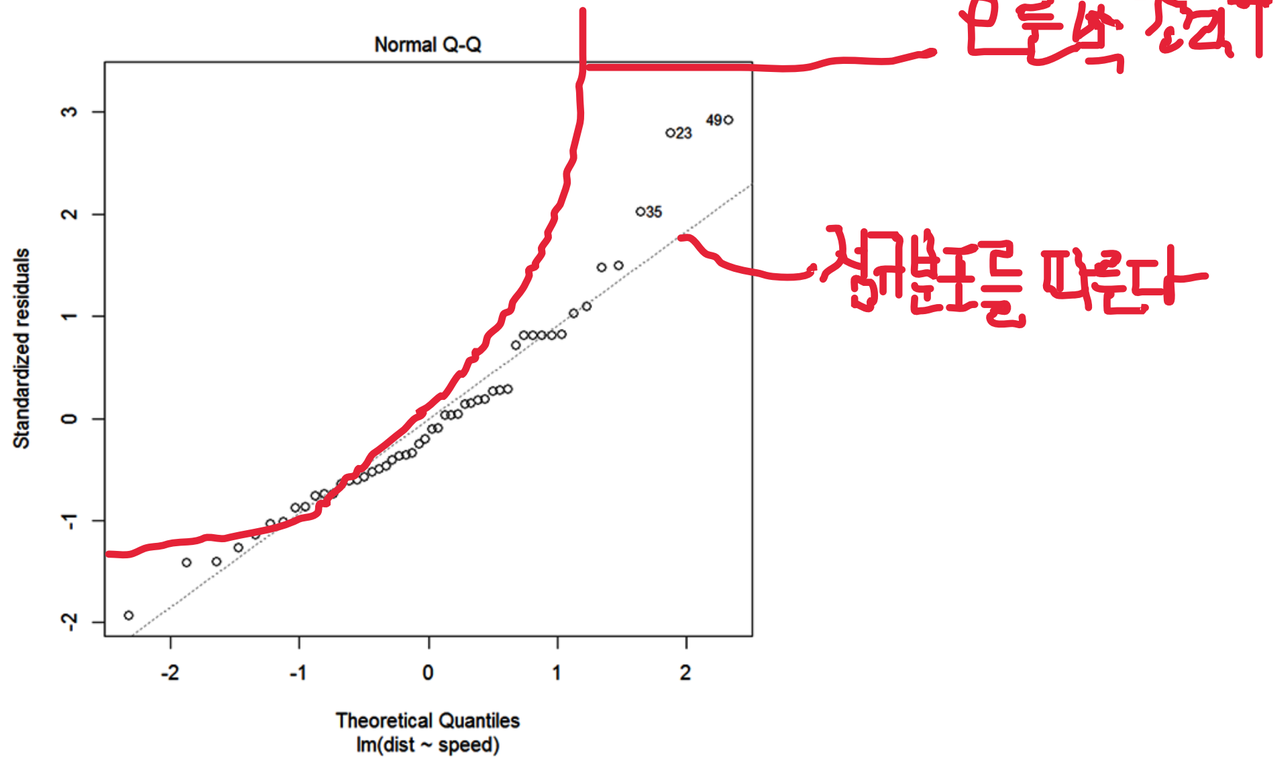
  - 오차의 정규성 위반 : 히스토그램, QQ플롯
  - 오차의 등분산성 : 잔차산점도
  - 오차의 독립성 : 잔차산점도



가정 위반 검토 및 해결

다중회귀모형의 가정 위반 검토 및 해결

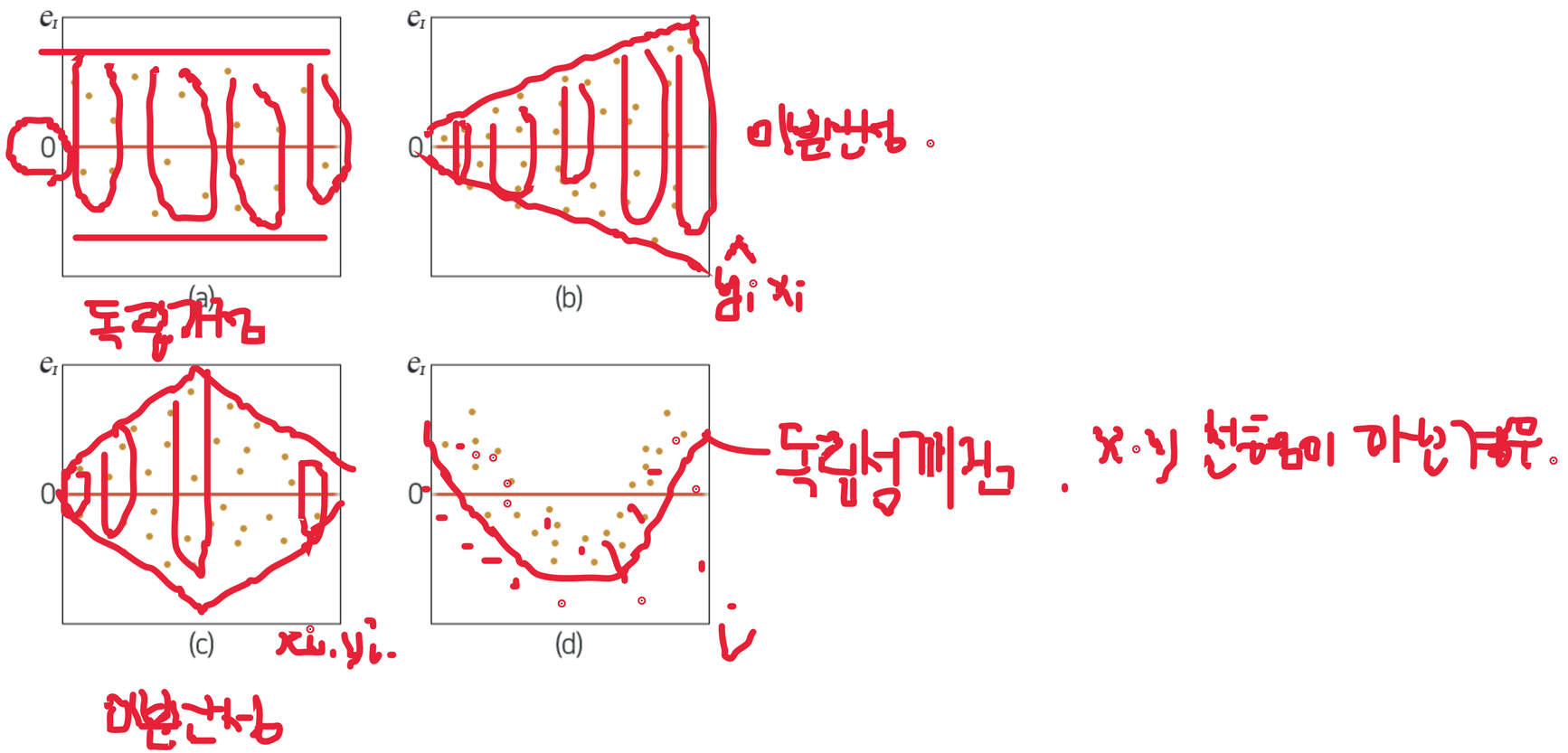
- QQ 플롯을 이용한 오차의 정규성 검토.



가정 위반 검토 및 해결

다중회귀모형의 가정 위반 검토 및 해결

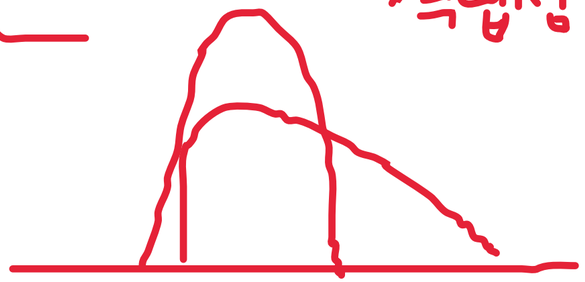
- 잔차산점도를 이용한 오차의 독립성, 등분산성 검토.



가정 위반 검토 및 해결

다중회귀모형의 가정 위반 검토 및 해결

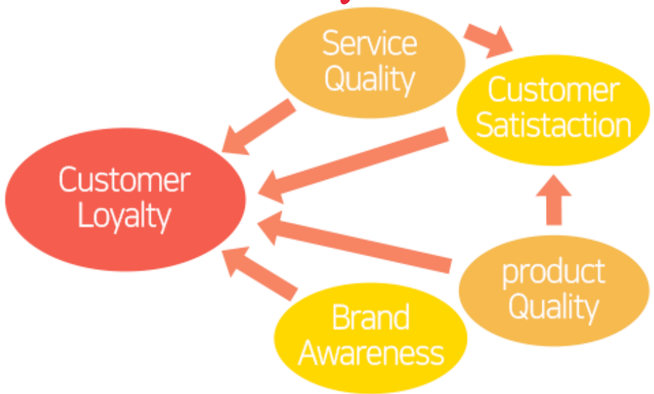
- 가정 위반 시 해결방안
  - 오차의 정규성 위반 : 변수변환
  - 오차의 등분산성 : 가중최소제곱회귀
  - 오차의 독립성 : 시계열 분석



다중공선성 진단 및 해결

다중공선성

- 다중공선성이란
  - 독립변수들 간에 강한 선형관계가 존재하는 경우
  - 다중회귀모형 분석 시 자주 발생하는 문제 중 하나임.
  - 다중회귀모형에서 회귀계수 추정에 부정적인 영향을 미침.
    1) 개별적인 회귀계수 추정의 신뢰성이 떨어져 추정치를 믿을 수 없게 만듦.
    2) 전반적인 모형의 적합성이나 정확도는 크게 변하지 않음.



다중공선성 진단 및 해결

다중공선성

- 다중공선성 진단방법

  - VIF 계수 도출

$$VIF = \frac{1}{1 - R_j^2}$$

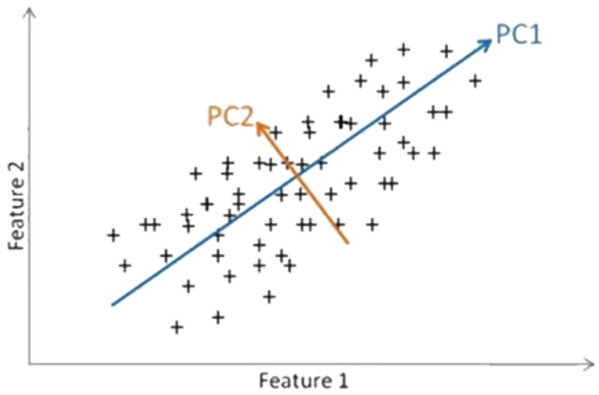
  1) R_j^2 : x_j 종속변수로 두고 나머지 독립변수로 설명하는 다중선형회귀모델에서의 결정계수.

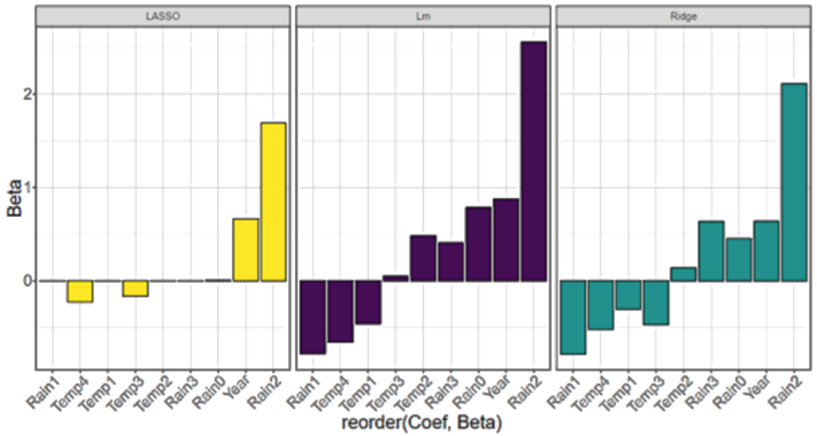
  - VIF 계수가 5 또는 10이상인 경우 다중 공선성이 심각한 것으로 봄.

다중공선성 진단 및 해결

다중공선성

- 다중공선성의 해결책
  - 변수선택으로 중복된 변수를 제거
  - 주성분 분석 등을 이용하여 중복된 변수를 변환하여 새로운 변수 생성
  - 릿지, 라쏘 등으로 중복된 변수의 영향력을 일부만 사용

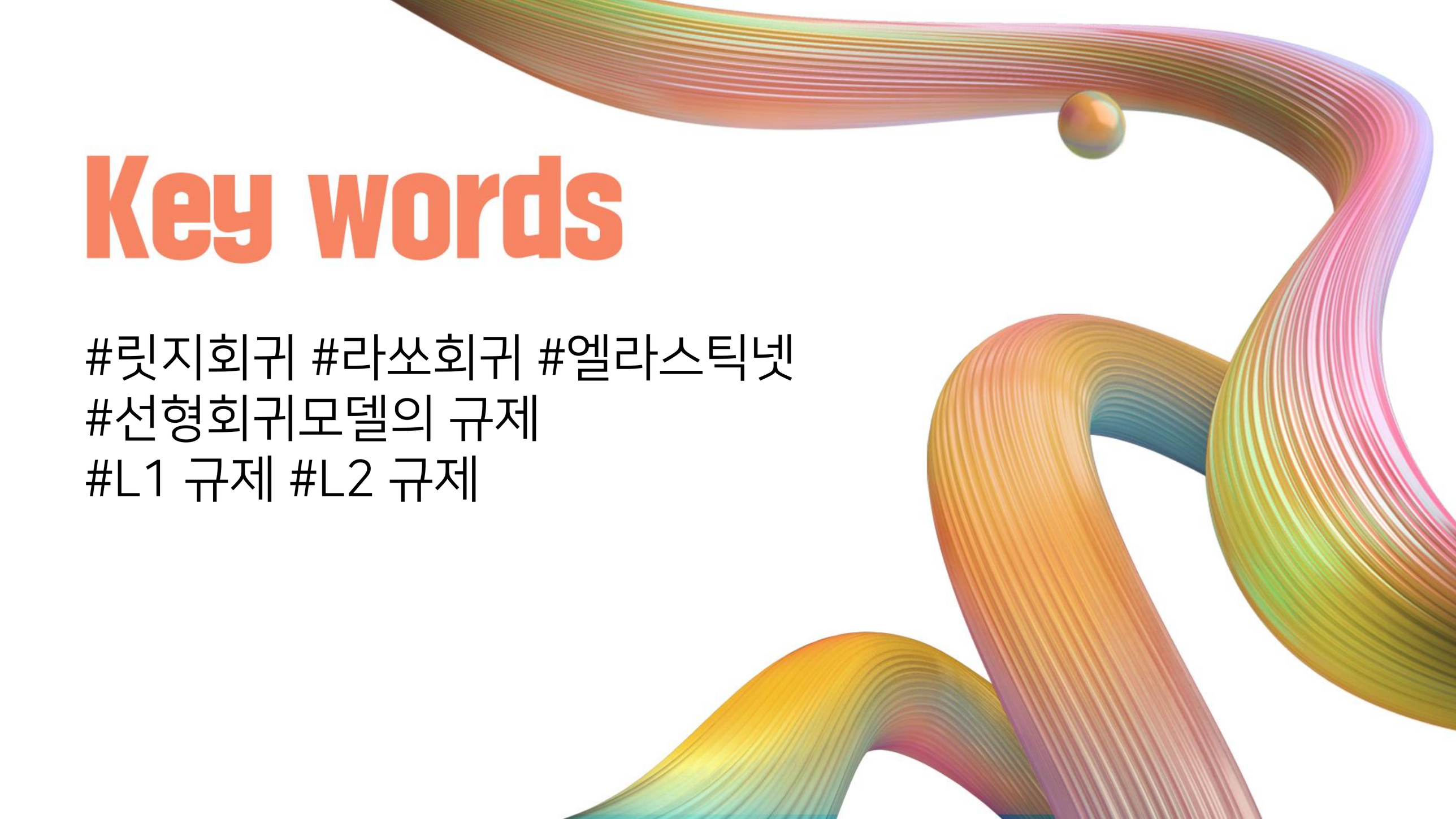




규제가 있는 선형회귀모델
(Ridge, Lasso, Elastic Net)

Key words

#릿지회귀 #라쏘회귀 #엘라스틱넷
#선형회귀모델의 규제
#L1 규제 #L2 규제



규제가 있는 선형회귀모델

선형회귀모델의 규제

- 모형의 과대적합을 막기 위한 규제 방법(regularization) 으로
  선형회귀모형에서는 보통 모델의 가중치를 제한하는 방법을 사용함.

  - 선형 회귀모델의 비용함수

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2$$

  - 규제가 있는 경우 비용함수

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2 + \lambda \cdot penalty(\beta_1, \ldots, \beta_k)$$

규제가 있는 선형회귀모델

선형회귀모델의 규제

- 가중치를 제한하는 방법에 따른 규제 선형회귀모델의 종류.
  - 릿지회귀(Ridge Regression)
  - 라쏘회귀(Lasso Regression)
  - 엘라스틱넷(Elastic Net)

릿지회귀

릿지회귀(Ridge Regression)와 L2 규제

- 릿지회귀의 비용함수
  - 비용함수 $J(\boldsymbol{\beta})$ 에 규제항 $\lambda \cdot \sum_{j=1}^{k} \beta_j^2$ 이 추가된 선형회귀모형.

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k} \beta_j^2$$

  - λ(규제정도를 결정하는 하이퍼파라미터)
    - λ가 크면 규제 많음. 회귀계수 추정치가 작아짐.
    - λ가 0이면 일반 선형회귀모델과 동일한 결과.
    - 적절한 λ는 교차검증(cross-validation)등으로 최적화.

- 릿지회귀의 훈련
  - 비용함수 $J(\beta)$ 를 최소로 하는 회귀계수 $\hat{\beta}^R = (\hat{\beta}_0^R, \hat{\beta}_1^R, \ldots, \hat{\beta}_k^R)$ 를 찾는 문제.

$$\hat{\beta}^R = arg\min_{\beta} J(\beta)$$

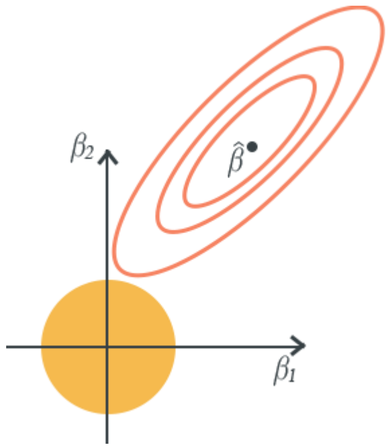
릿지회귀

릿지회귀(Ridge Regression)와 L2 규제

- Alternative Fomulation

  - 어떤 임의의 λ에 대해 이에 대응하는 하나의 t가 존재하여,
    아래 식으로 동일한 해 $\hat{\beta}^R$를 얻게 됨.

$$\hat{\beta}^R = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2 \right\} \; subject\ to \sum_{j=1}^{k} \beta_j^2 \leq t$$



라쏘회귀

라쏘회귀(LASSO Regression)와 L1규제

- 라쏘회귀의 비용함수

  - 비용함수 $J(\beta)$ 에 규제항 $\lambda \cdot \sum_{j=1}^{k}|\beta_j|$이 추가된 선형회귀모형.

$$J(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k}|\beta_j|$$

  - Λ : 규제정도를 결정하는 하이퍼파라미터, 교차검증(cross-validation)등으로 최적화.

  - 비용함수 $J(\beta)$를 최소로 하는 회귀계수 $\hat{\beta}^L = (\hat{\beta}_0^L, \hat{\beta}_1^L, \ldots, \hat{\beta}_k^L)$ 를 찾는 문제.

$$\hat{\beta}^L = arg\min_{\beta} J(\beta)$$

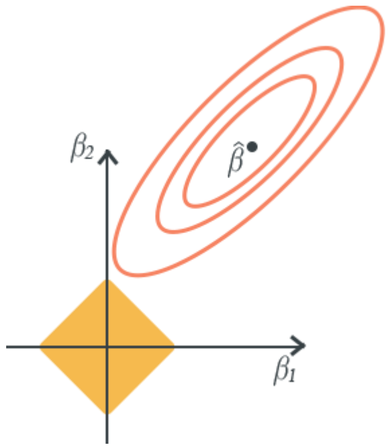
라쏘회귀

라쏘회귀(LASSO Regression)와 L1규제

- Alternative Fomulation

  - 어떤 임의의 λ에 대해 이에 대응하는 하나의 t가 존재하여,
    아래 식으로 동일한 해 $\hat{\beta}^{L}$를 얻게 됨.

$$\hat{\beta}^{L} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2 \right\} \; subject\ to \sum_{j=1}^{k} |\beta_j| \le t$$



릿지회귀와 라쏘회귀의 특징

릿지회귀와 라쏘회귀의 특징

- 두 방식 모두 추정치는 일반선형회귀모형과는 달리
  편의가 발생하지만, 분산은 더 작아지게 됨.
  → λ에 따라 일반화오차가 더 작아질 수 있음.
- 라쏘 회귀의 경우 제약 범위가 각진 형태
  → 파라미터의 일부가 0이 되는 경향이 있음. (sparse model)
- 릿지 회귀의 경우 제약 범위가 원의 형태
  → 파라미터가 0이 되지 않고 전반적으로 줄어드는 경향이 있음.

엘라스틱 넷

엘라스틱 넷(Elastic Net)

- L1과 L2 규제를 혼합한 방식.
- 엘라스틱 넷의 비용함수.

$$J(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2 + \lambda_1 \cdot \sum_{j=1}^{k}\beta_j^2 + \lambda_2 \cdot \sum_{j=1}^{k}|\beta_j|$$

- 릿지회귀와 라쏘회귀의 장점을 모두 가짐.