회 귀 분 석

ToBig's 9기 임소정

Linear Regression Analysis

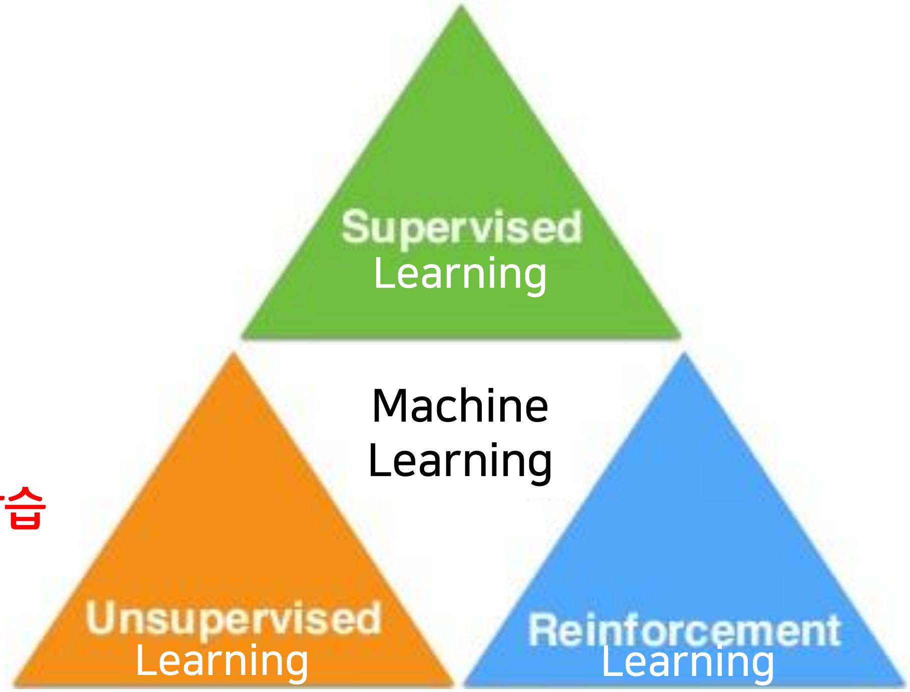
선형회귀분석

Contents

Copyright ToBig's

Machine Learning 알고리즘 분류

- labeled data 이용한 학습
- Direct feedback
- Predict outcome/future



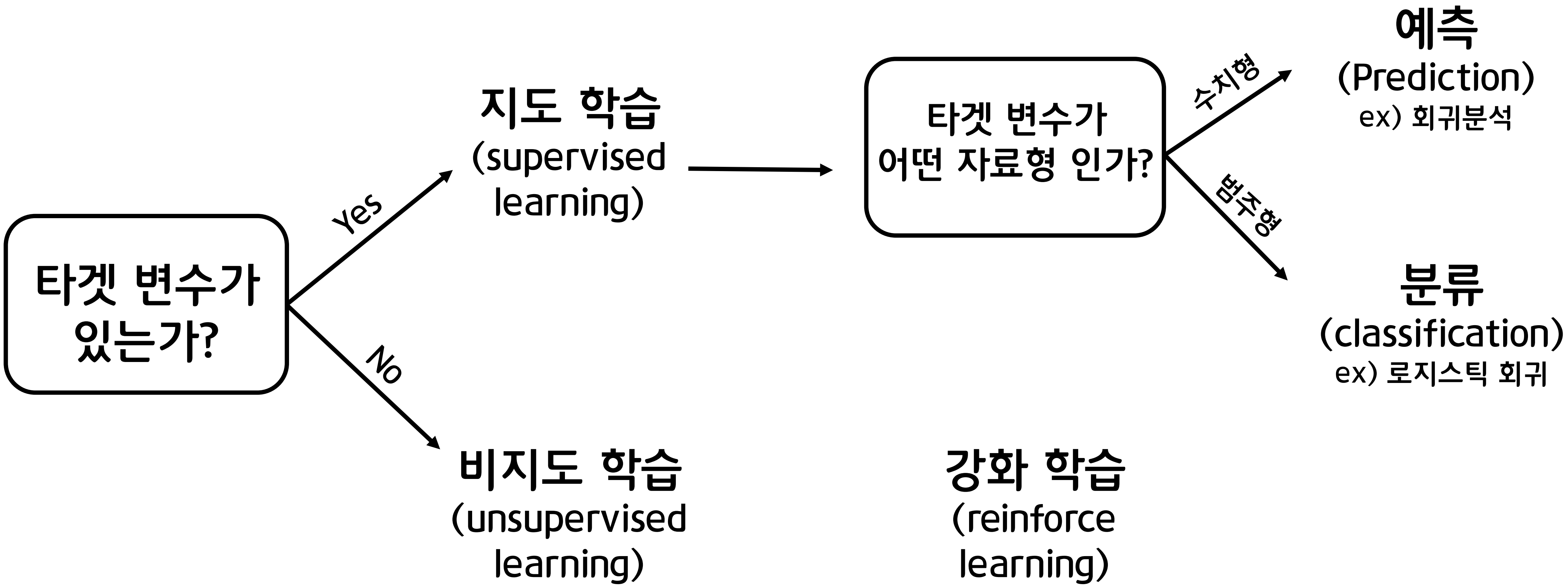
- Unlabeled data 이용한 학습
- No feedback
- “Find hidden structure”

- Decision Process
- Reward system
- Learn series of actions

http://solarisailab.com/archives/1785

Copyright ToBig's

Machine Learning 알고리즘 분류

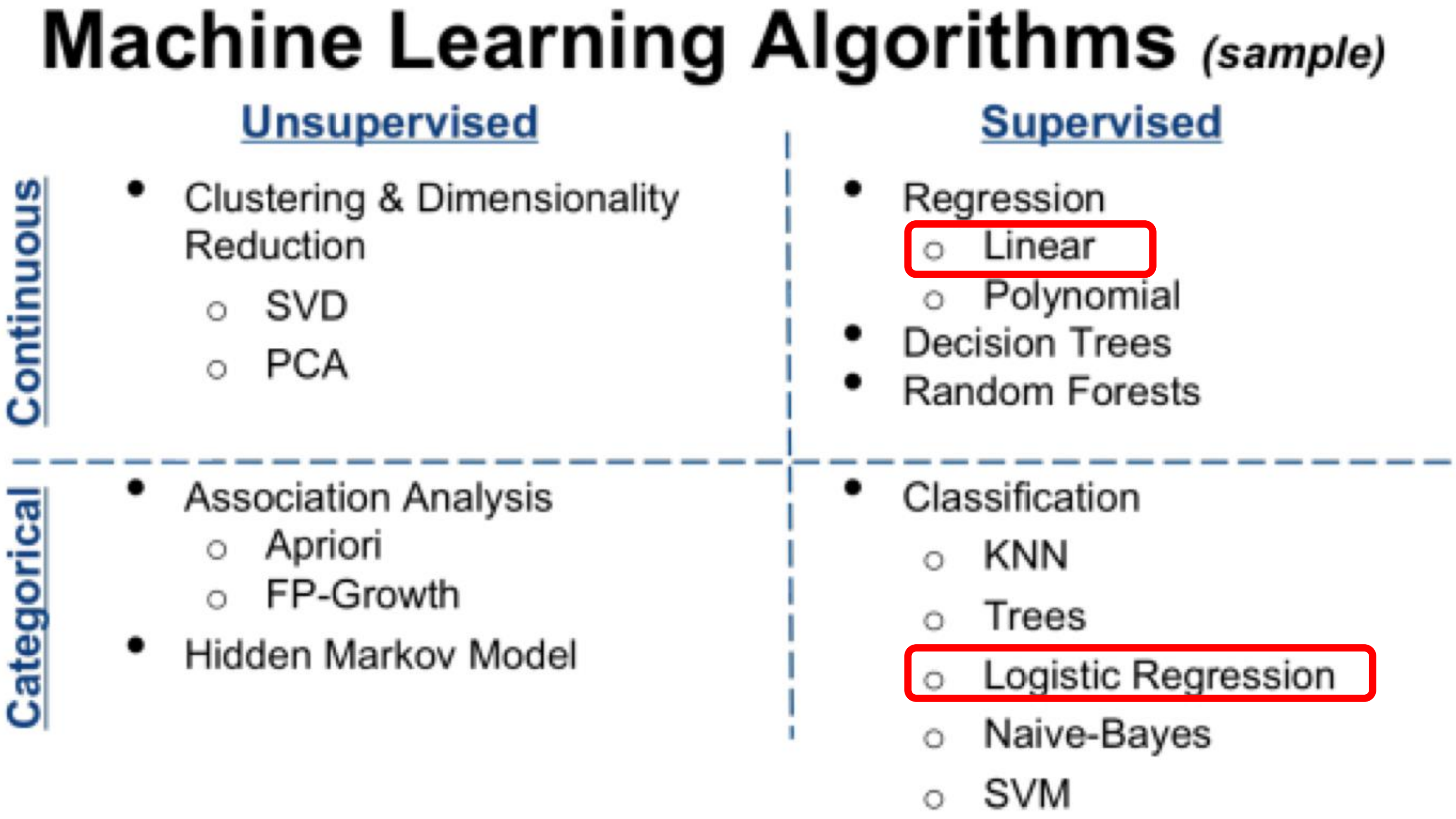


Ⓒopyright ToBig's

Machine Learning 알고리즘 분류

예측 ←

분류 ←



https://swalloow.github.io/pyml-intro1

Copyright ToBig's

Unit 02 | 선형 회귀

회귀분석

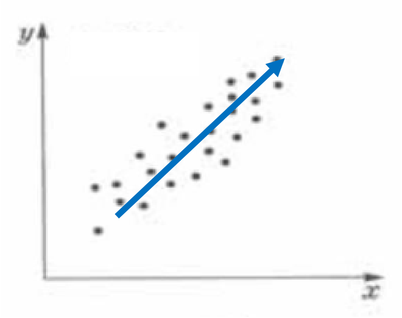
하나 이상의 독립변수 $X_1, X_2, \ldots, X_p$의 종속 변수 Y 에 대한 영향의 추정을 할 수 있는 통계기법.
선형 회귀는 변수들 사이의 **선형 상관 관계**를 모델링하는 **회귀분석** 기법이다.

종속변수 ➡ 독립변수들에 의해 설명되는 변수 (반응변수)

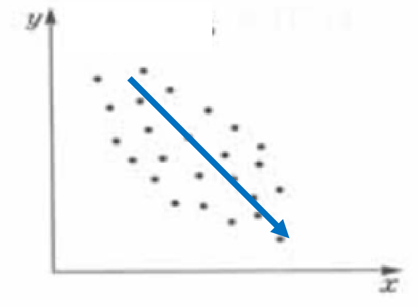
독립변수 ➡ 종속변수를 설명하기 위해 쓰이는 변수 (설명변수)

Copyright ToBig's

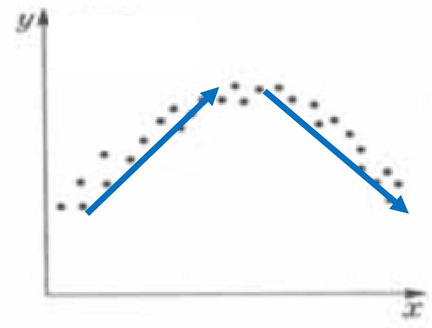
상관 관계



x가 증가할수록 y가 증가한다.



x가 증가할수록 y가 감소한다.



x가 증가할수록, y가 증가 했다가 감소한다.

x와 y는 상관 관계가 없어 보인다.

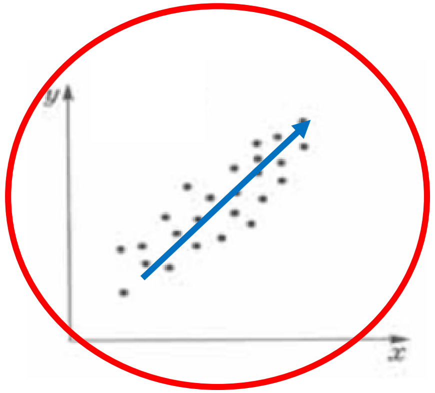
한 변수가 변할 때 다른 변수도 변화하면 상관관계가 있다!

그러나 두 변수의 인과적 선후관계는 없다. 즉, 어느 쪽이 원인인지 알 수 없다.

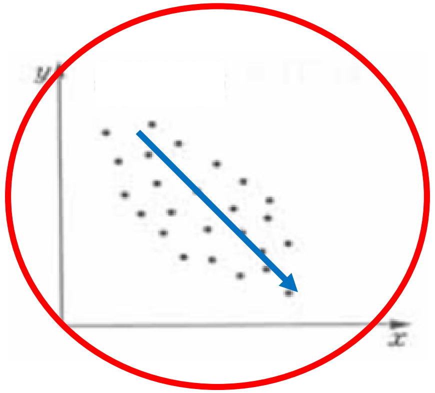
http://biocourse.org/index.php/%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D

Copyright ToBig's

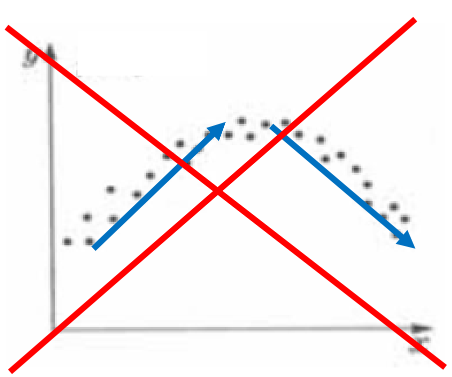
선형 상관 관계



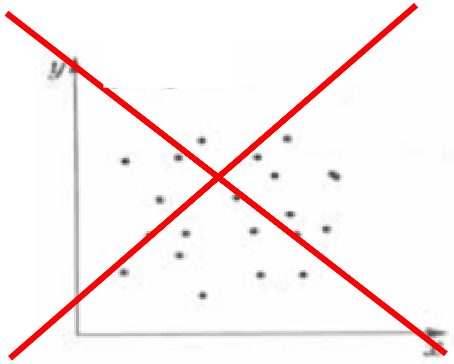
x가 증가할수록 y가 증가한다.



x가 증가할수록 y가 감소한다.



x가 증가할수록, y가 증가 했다가 감소한다.



x와 y는 상관 관계가 없어 보인다.

➡ **선형 회귀로는 직선 관계를 파악할 수 있는 왼쪽 2가지의 경우만 분석을 할 수 있다.**

http://biocourse.org/index.php/%ED%9A%8C%EA%B7%80%EB%B6%84%EC%84%9D

Ⓒopyright ToBig's

회귀 분석 종류

1. 단순선형회귀 ➡ 반응변수와 설명변수가 1:1 관계일 때

2. 다중선형회귀 ➡ 반응변수와 설명변수가 1:N 관계일 때

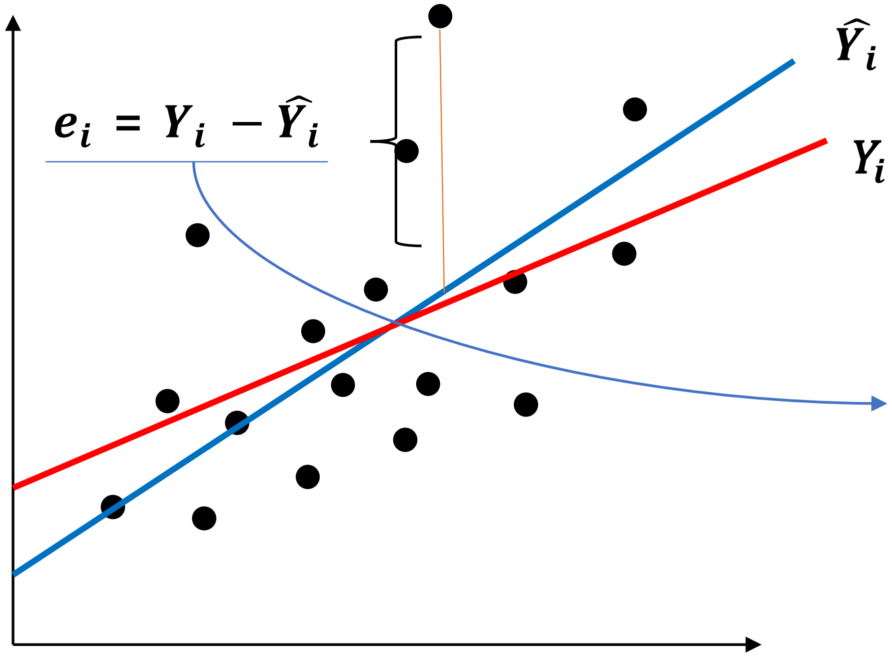
3. 비선형회귀 ➡ $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

4. 로지스틱회귀 ➡ 다음 시간에 설명할 부분 ~!

Copyright ToBig's

Unit 02 | 선형 회귀

1. 단순 선형 회귀 분석 - 설명 변수가 1개



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

$\widehat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_i$: 추정회귀선

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

잔차 : 관측치와 추정회귀선의 수직거리

어떤 직선이 가장 관측치들의 패턴을 잘 표현하는 직선이지 ??!

잔차의 제곱의 합이 최소가 되도록! ➡ 최소제곱법

Copyright ToBig's

* 최소제곱법

잔차의 제곱의 합이 최소가 되도록 모수를 추정하는 방법

$$Q(\beta_0, \beta_1) = \sum {e_i}^2 = \sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\beta_0 \rightarrow \widehat{\beta_0}$$
$$\beta_1 \rightarrow \widehat{\beta_1}$$

편미분

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum(Y_i - \beta_0 - \beta_1 X_i)X_i = 0 \end{cases}$$

$$\widehat{\beta_1} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

< 모수 β_0, β_1 의 최소제곱추정량 >

Copyright ToBig's

회귀 분석 종류

1. 단순선형회귀 ➡ 반응변수와 설명변수가 1:1 관계일 때

2. 다중선형회귀 ➡ 반응변수와 설명변수가 1:N 관계일 때

설명변수의 수 증가

$Y = \beta_0 + \beta_1 X + \varepsilon$ ➡ $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon$

Copyright ToBig's

2. 다중 선형 회귀 분석 – 설명 변수가 2개 이상

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon$ 행렬로 이해해 봅시다 !

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} \quad \Rightarrow \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & & \cdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

N개의 종속변수	P개의 반응변수	P+1개의 회귀계수	N개의 오차
N x 1	N x (p+1)	(p+1) x 1	N x 1

Copyright ToBig's

2. 다중 선형 회귀 분석 – 설명 변수가 2개 이상

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P + \varepsilon$ **행렬로 이해해 봅시다 !**

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdot \cdot + \beta_n x_{1n} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdot \cdot + \beta_n x_{2n} \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdot \cdot + \beta_n x_{nn} \end{pmatrix}$$

간단히 행렬화하면 $\hat{Y} = X'\beta$ 꼴로 나타낼 수 있다.

따라서 잔차 ($e_i = Y_i - \hat{Y}_i$)의 제곱합은 $Q(\beta) = \sum(Y_i - x_i'\beta)^2 = (Y - X\beta)'(Y - X\beta)$

β에 대해 편미분한 결과를 영벡터로 놓고 앞과 같은 원리로 β의 최소 제곱 추정량 $\hat{\beta} = (X'X)^{-1}X'Y$

Copyright ToBig's

Unit 02 | 선형 회귀

* 회귀분석 결과표 보는 법

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64117 -0.77977 -0.01839  0.68191  2.88420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1919     0.1084  -1.769    0.080 .
x1            0.9782     0.1059   9.240 5.86e-15 ***
x2           -0.1371     0.1104  -1.242    0.217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.083 on 97 degrees of freedom
Multiple R-squared:  0.4683,    Adjusted R-squared:  0.4573
F-statistic: 42.72 on 2 and 97 DF,  p-value: 4.951e-14
```

제일 먼저 봐야 할 부분!

회귀식 전체에 대한 유의성 검정

H_0 : 모든 회귀계수가 0이다.
H_1 : 회귀식이 유의하다.

유의수준 0.05일 때
p-value가 0.05보다 작으므로
H_0 기각 (회귀식이 유의하다)

Copyright ToBig's

* 회귀분석 결과표 보는 법

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64117 -0.77977 -0.01839  0.68191  2.88420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1919     0.1084  -1.769    0.080 .
x1            0.9782     0.1059   9.240 5.86e-15 ***
x2           -0.1371     0.1104  -1.242    0.217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.083 on 97 degrees of freedom
Multiple R-squared:  0.4683,	Adjusted R-squared:  0.4573
F-statistic: 42.72 on 2 and 97 DF,  p-value: 4.951e-14
```

결정계수 (R^2) : 총 변동 중에서 회귀 모형에 의해 설명되어지는 변동의 크기로
추정된 회귀식이 얼마나 해당자료를 잘 설명하고 있는지 알려준다. $(0 \leq R^2 \leq 1)$

단, 변수가 늘어나면 결정계수는 무조건 증가!

수정결정계수 : 변수의 개수를 고려한 결정계수

Copyright ToBig's

* 회귀분석 결과표 보는 법

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64117 -0.77977 -0.01839  0.68191  2.88420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1919     0.1084  -1.769    0.080 .
x1            0.9782     0.1059   9.240 5.86e-15 ***
x2           -0.1371     0.1104  -1.242    0.217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.083 on 97 degrees of freedom
Multiple R-squared:  0.4683,	Adjusted R-squared:  0.4573
F-statistic: 42.72 on 2 and 97 DF,  p-value: 4.951e-14
```

모수($\beta_0, \beta_1, \beta_2$)의 추정값

각 회귀계수에 대한 유의성 검정

H_0 : 회귀계수가 0이다.
H_1 : 회귀계수가 유의하다.(유의한 변수이다)

유의수준 0.05일 때 p-value가 0.05보다 작으면 H_0 기각 (회귀계수가 유의하다)

여기서, x1의 p-value는 0.05보다 작으므로 x1은 유의한 변수이고, x2의 p-value는 0.05보다 크기 때문에 x1이 존재할 때, x2의 설명력은 유의하지 않은 것으로 판단 할 수 있다.

Copyright ToBig's

* 지시 변수와 범주형 변수

지시 변수 : 그룹을 분류 해주는 설명변수

1) 명목형 변수를 수치형으로 표현을 해주되 0과 1로만 표현한다.
2) 원래 범주의 개수가 N이라면 N-1개의 지시변수 만든다.

EX) 학교 : 초등학교, 중학교, 고등학교, 대학교 (총 4개의 범주) => 3개의 변수 필요

X1	X2	X3	
0	0	0	초등학교
1	0	0	중학교
0	1	0	고등학교
0	0	1	대학교

Copyright ToBig's

* 지시 변수와 범주형 변수

EX) 학교 : 초등학교, 중학교, 고등학교, 대학교 (총 4개의 범주) => 3개의 변수 필요

X1	X2	X3	
0	0	0	초등학교
1	0	0	중학교
0	1	0	고등학교
0	0	1	대학교

Y를 임금, X_4 를 경력 이라고 할 때,

추정된 회귀식 : $\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\beta_3}X_3 + \widehat{\beta_4}X_4$

➡ 여기서, 경력이 1단위 증가할 때 초등학교 졸업이라면 Y는 $\widehat{\beta_0}$ + $\widehat{\beta_4}$ 만큼 증가하고
중학교 졸업이라면 Y는 $\widehat{\beta_0}+\widehat{\beta_1}+\widehat{\beta_4}$만큼 증가하고,
고등학교 졸업이라면 Y는 $\widehat{\beta_0}+\widehat{\beta_2}+\widehat{\beta_4}$만큼 증가하고,
대학교 졸업이라면 Y는 $\widehat{\beta_0}+\widehat{\beta_3}+\widehat{\beta_4}$만큼 증가한다.

Copyright ToBig's

회귀진단

만들어진 회귀 모형이 적절한가에 대한 평가.
각각의 관측값이 모형에 어떠한 영향을 미치는지(자료진단),
회귀모형과 가정이 타당한지(모형진단) 검토하는 것.

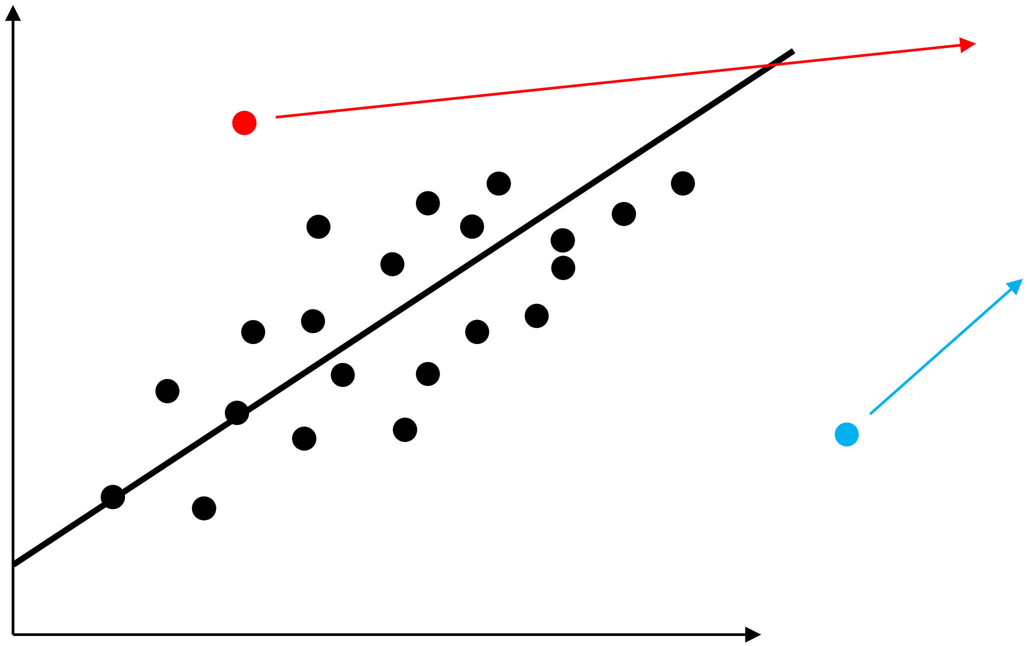
1. 자료 진단 : 이상점, 영향력 관측값

2. 모형 진단 : 선형성, 정규성, 등분산성, 독립성, 비상관성

Copyright ToBig's

Unit 03 | 회귀 진단

1. 자료 진단

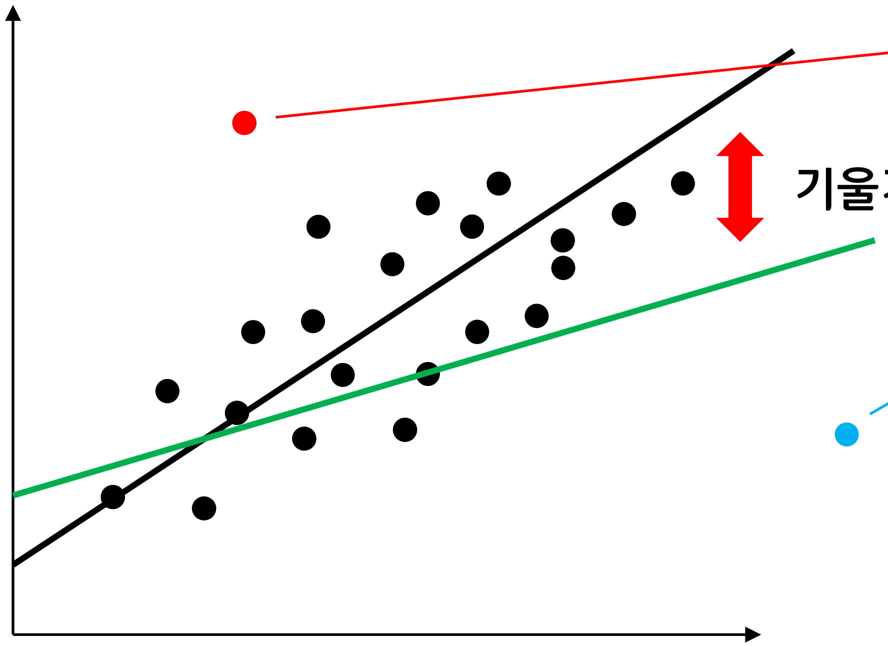


이상점 : 다른 관측값과는 다른 경향을 가지는 관측값

영향력 관측값 : 관측값을 제거 했을 때 모수 추정에 큰 영향을 미치는 관측값

Copyright ToBig's

1. 자료 진단



이상점 : 다른 관측값과는 다른 경향을 가지는 관측값

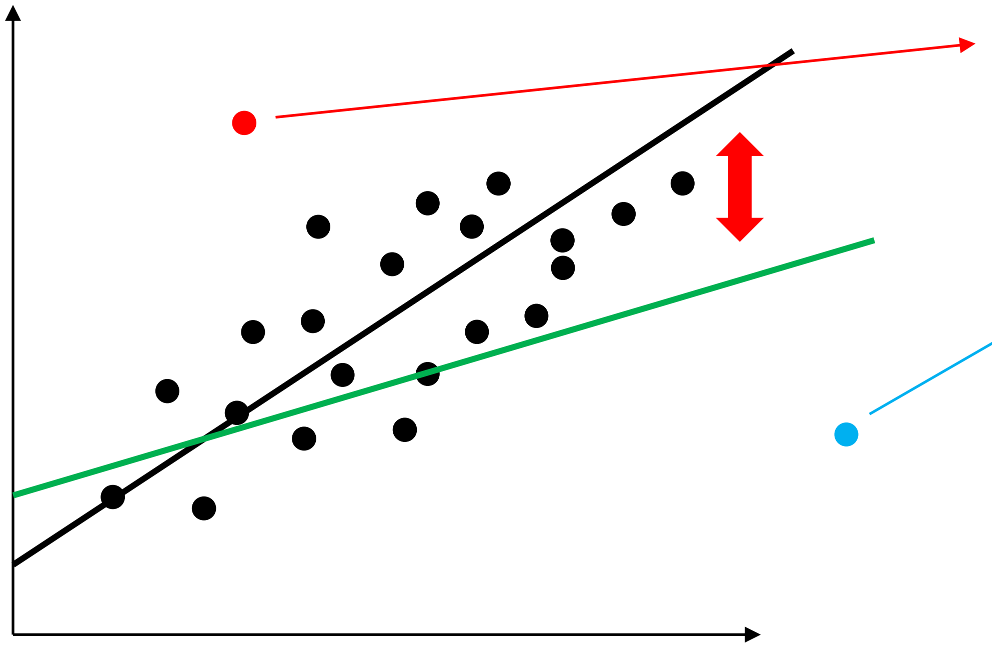
기울기 ($\widehat{B_1}$)가 크게 변화 !!

영향력 관측값 : 관측값을 제거 했을 때 모수 추정에 큰 영향을 미치는 관측값

Copyright ToBig's

Unit 03 | 회귀 진단

1. 자료 진단



이상점 : 다른 관측값과는 다른 경향을 가지는 관측값

➡ 잔차산점도, Bonferroni t-분포표 이용해 검정

영향력 관측값 : 관측값을 제거 했을 때 모수 추정에 큰 영향을 미치는 관측값

➡ cook의 D통계량, DFFITS, DFBETAS 이용

https://rstudio-pubs-static.s3.amazonaws.com/190997_40fa09db8e344b19b14a687ea5de914b.html

Copyright ToBig's

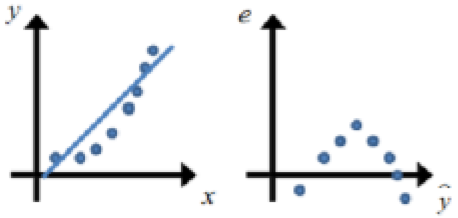
2. 모형 진단

1. 선형성
2. 정규성
3. 등분산성
4. 독립성
5. 비상관성

☞ 선형 회귀 분석이므로 종속 변수와 설명 변수 간의 관계가 선형성을 띄어야 한다.

☞ 종속 변수와 독립 변수가 선형 관계이면 잔차와 예측치 사이에는 어떠한 체계적인 관계가 없기 때문에 Residual vs Fitted plot의 분포가 random한지 확인한다.

☞ 아래의 예시들처럼 비선형적 관계가 보이면 2차, 3차 등의 다항식을 포함하는 비선형 회귀를 해야 한다. 혹은 변수에 log나 $\sqrt{\ }$ 변환을 해줄 수 있다.



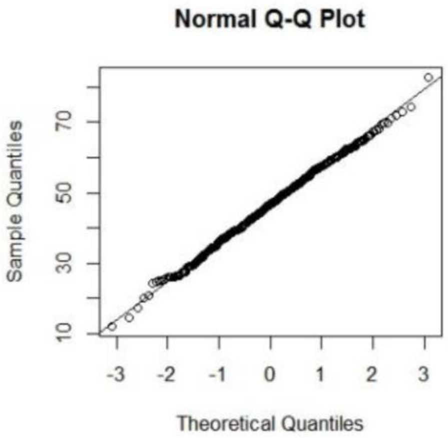
Copyright ToBig's

2. 모형 진단

1. 선형성
2. 정규성
3. 등분산성
4. 독립성
5. 비상관성

☞ 오차가 정규분포를 따른다는 가정

☞ normal Q-Q plot을 그렸을 때, 점들이 45도 각도의 직선 위에 있으면 정규성 가정을 만족한다.



Copyright ToBig's

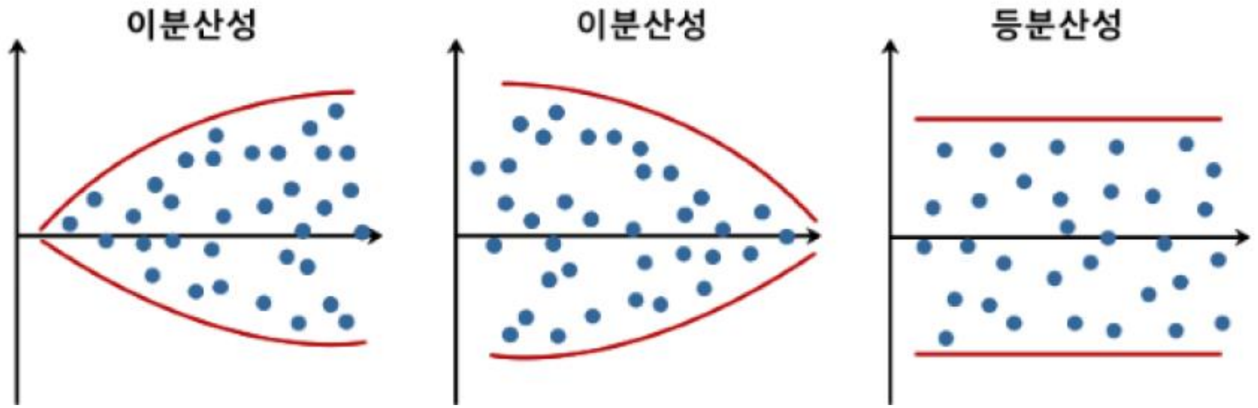
2. 모형 진단

1. 선형성
2. 정규성
3. 등분산성
4. 독립성
5. 비상관성

☞ 오차의 분산이 일정하다는 가정

☞ Residual vs Fitted plot의 분포가 random하게 나타나는지 확인한다. 스코어 검정으로 확인한다.

☞ 변수에 log나 변환 등을 해줄 수 있다. 또는 $wi \propto 1/\ Var(yi)$가 되도록 가중치를 주는 가중최소제곱법을 이용 할 수 있다.



Copyright ToBig's

2. 모형 진단

1. 선형성
2. 정규성
3. 등분산성
4. 독립성
5. 비상관성

☞ 오차항들이 독립적 이어야 한다.

☞ 주식 시장에서 어제의 주가가 오늘의 주가에 영향을 미치는 것처럼 잔차의 변화에 어떠한 패턴이 있는 경우 독립성을 위배했다고 한다. 99% 시계열 자료일 경우가 많다.

Copyright ToBig's

2. 모형 진단

1. 선형성
2. 정규성
3. 등분산성
4. 독립성
5. 비상관성

☞ 설명 변수들 사이에 상관성이 없어야 한다.

☞ 설명 변수들 사이에 강한 상관 관계가 존재 한다면 매우 부적절한 모형이 되고 다중공선성 문제 발생!

https://rstudio-pubs-static.s3.amazonaws.com/190997_40fa09db8e344b19b14a687ea5de914b.html

Copyright ToBig's

다중공선성

임의의 상수Ci에 대해 $C_1X_1 + C_2X_2 + \cdots + C_PX_P = C_0$ 이 성립하거나 근사적으로 성립 할 때,
설명 변수 사이에 다중공선성이 존재한다.
다중공선성이 존재 하면 어떤 설명변수가 다른 설명변수들에 의해 결정, 설명 가능하다.

➡ 추정된 회귀계수의 분산인 $\mathrm{Var}(\hat{\beta}_j) = \sigma^2\left(\frac{1}{1-R_j^2}\right)S_{x_ix_j} = \sigma^2\ Vif_j\ S_{x_ix_j}$ 이 매우 커지게 되어
적합된 모형의 안정성과 신뢰성↓

Vif_j(분산팽창인자)

☞ 5 ~ 10 이상이면 다중공선성이 존재한다. 최소는1 $(\because Vif_j = \left(\frac{1}{1-R_j^2}\right))$

☞ 일반적으로 다중공선성을 가진 변수 중 1개를 제거해 문제 해결

©opyright ToBig's

변수선택법

1. 전진선택법(Forward Selection) : 아무 것도 없는 상태에서 변수를 고를 때, 변수를 하나씩 추가하면서 AIC 값이 낮아지는 변수를 고른다.

2. 후진제거법(Backward Elimination) : Full model 에서 변수를 고를 때, 변수를 하나씩 제거하면서 AIC 값이 낮아지는 변수를 고른다.

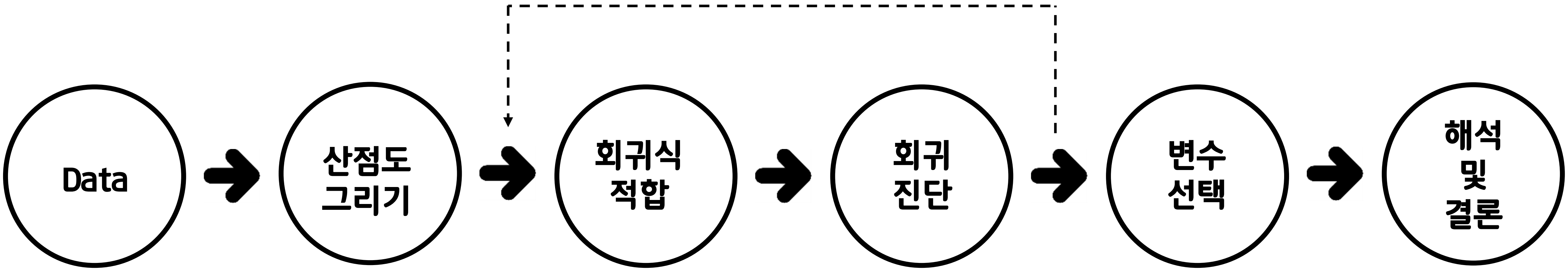
3. 단계적회귀방법(Stepwise) : 1번과 2번 방법을 혼합한 것으로 처음엔 전진선택법으로 한 변수를 추가한 뒤, 제거했다가 추가했다가를 반복하면서 AIC 값이 낮아지는 변수를 선택하는 방법

☞ 이미 선택된 변수가 새로운 변수 추가에 의해 중요도를 상실해 제거될 필요가 있는지 매 단계 검토함

모델 선정 기준 통계량 ➡ AIC↓,BIC↓, 결정계수↑, 수정결정계수↑

Copyright ToBig's

Unit 05 | 정리



Copyright ToBig's

Q & A

들어주셔서 감사합니다.