03-02

02 기초 통계분석

3-64. 회귀 분석(Regression Analysis)

출 : 20, 21

용어 정리

독립변수

- 다른 변수에 영향을 받지 않고 독립적으로 변화하는 수, 설명 변수라고도 함
- 입력 값이나 원인을 나타내는 변수, y = f(x) 에서 x에 해당하는 것

종속변수

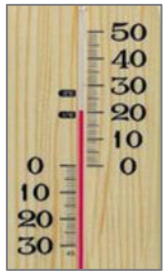
- 독립변수의 영향을 받아 값이 변화하는 수, 분석의 대상이 되는 변수
- 결과물이나 효과를 나타내는 변수 , y = f(x) 에서 y에 해당하는 것

잔차(오차항)

- 계산에 의해 얻어진 이론 값과 실제 관측이나 측정에 의해 얻어진 값의 차이
- 오차(Error) - 모집단, 잔차(Residual) - 표본집단

회귀 분석

- 변수와 변수 사이의 관계를 알아보기 위한 통계적 분석 방법
- 독립변수의 값에 의해 종속변수의 값을 예측하기 위함
- 일반 선형회귀는 종속변수가 연속형 변수일 때 가능함
- 이산형(범주형) - 명목, 서열척도, 연속형 - 구간, 비율척도

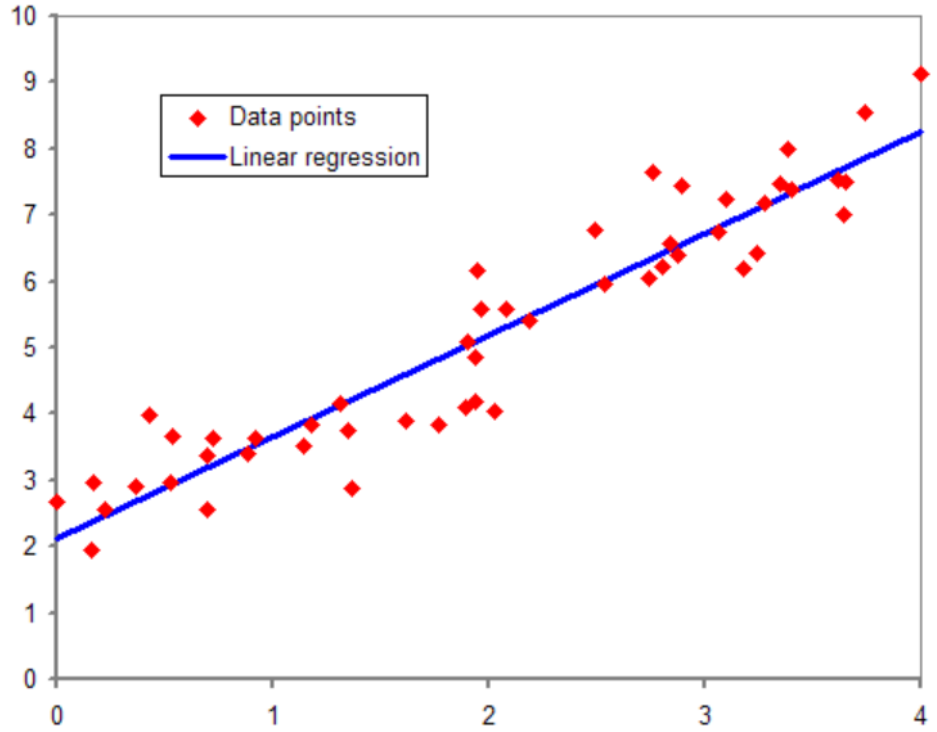


3-64. 회귀 모형

- 선형회귀모형
  - 종속변수 y와 한 개 이상의 독립변수 X와의 선형 상관 관계를 모델링하는 회귀분석 기법
  - 한 개의 독립변수 : 단순 선형회귀, 둘 이상의 독립변수 : 다중 선형 회귀
- 단순회귀모형 (독립변수 1개일 때)

모집단

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i=1, 2, \cdots, n$
- Y_i : 종속변수
- X_i : 독립변수
- ε_i : 에러(error)
- β_0 : 선형회귀식의 절편
- β_1 : 기울기, 회귀계수(coefficient)

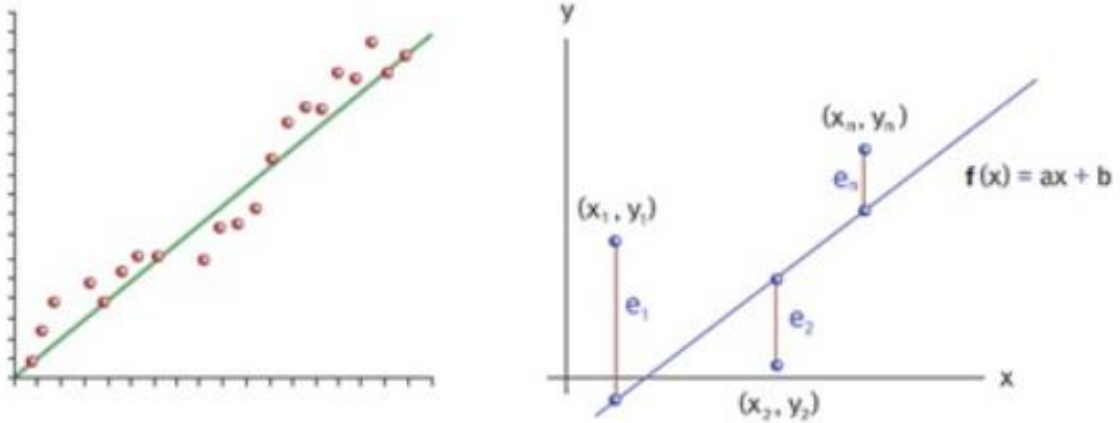


영어 위키백과의 Amatulic (same as Anachronist on Wikimedia) - en.wikipedia 에서 공용으로 옮겨왔습니다.

출 : 20, 21

최소자승법(Least Square Method)

- Y= f(X)의 측정값 y_i와 함수값 $f(x_i)$의 차이를 제곱한 것의 합이 최소가 되도록 Y=f(X)를 구하는 것
- Y = aX + b 일 때 잔차를 제곱한 것의 합이 최소가 되도록 하는 상수 a, b를 찾는 것
- 즉, (측정값 - 함수값)^2 의 합이 최소가 되는 직선의 그래프를 찾는 것
- 큰 폭의 잔차에 대해 보다 더 큰 가중치를 부여하여, 독립변수 값이 동일한 평균치를 갖는 경우 가능한 한 변동 폭이 적은 표본회귀선을 도출하기 위한 것



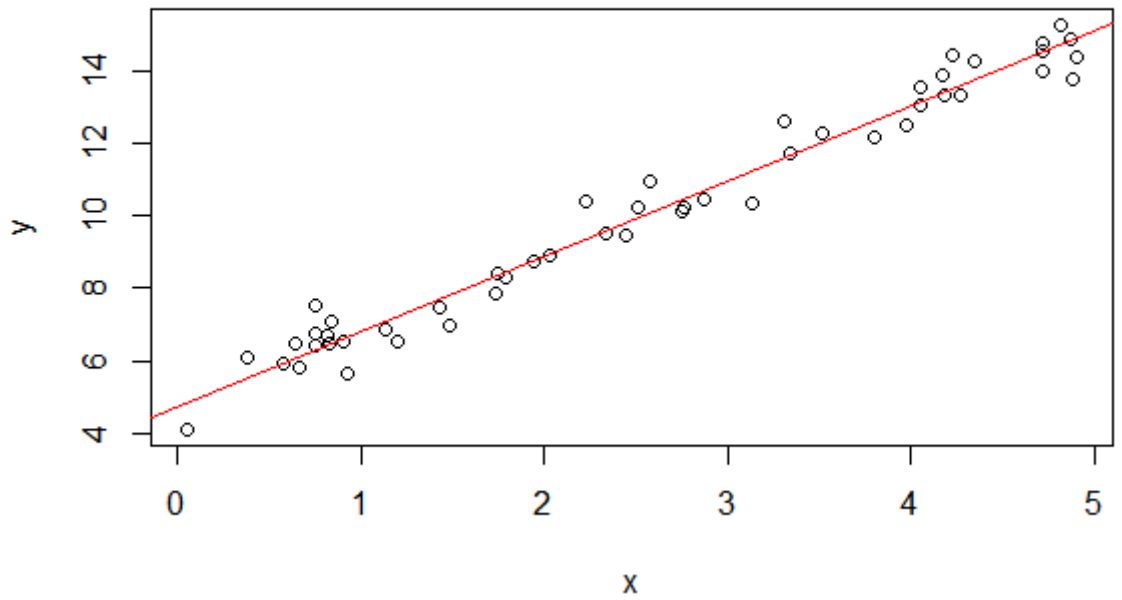
3-64. 단일 회귀 모형의 예


```
2 set.seed(2)
3 x = runif(50, 0, 5)
4 y = 5 + 2 * x + rnorm(50, 0, 0.5)
5 df <- data.frame(x, y)
6 fit <- lm(y~x, data=df)
7 fit

 Call:
 lm(formula = y ~ x, data = dfrm)

 Coefficients:
 (Intercept)           x
       4.748       2.072
```

절편

회귀 계수

회귀 방정식 ← y=2.072*x + 4.748



- runif(개수, 시작, 끝) : 시작 ~ 끝 범위에서 개수 만큼의 균일분포를 따르는 난수 발생
- rnorm(개수, 평균, 표준편차) : 특정 평균 및 표준편차를 갖으며 정규분포를 따르는 난수 발생
  평균, 표준편차 생략 시 평균 0, 표준편차 1
- lm(y~x, data = df) : df에서 y를 종속변수, x를 독립변수로 회귀 모형 생성

3-64. 다중 회귀 모형의 예

```
14 rm(list=ls())
15 set.seed(10)
16 u <- runif(50, 0, 6)
17 v <- runif(50, 6, 12)
18 w <- runif(50, 3, 25)
19 y = 3 + 0.5 * u + 1 * v - 2*w + rnorm(50, 0, 0.5)
20 df <- data.frame(y, u, v, w)
```

```
> a <- lm(y~u+v+w, df)
> a

Call:
lm(formula = y ~ u + v + w, data = df)

Coefficients:
(Intercept)
     3.4374

      u         v         w
 0.4676    0.9556   -1.9923
```

종속변수

독립변수

회귀 방정식

y = 3.4374 + 0.4676*u + 0.9556*v - 1.9923*w

회귀 계수

절편

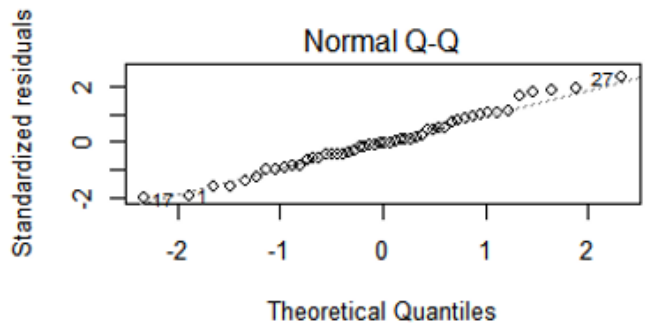
3-64. 회귀 모형의 가정

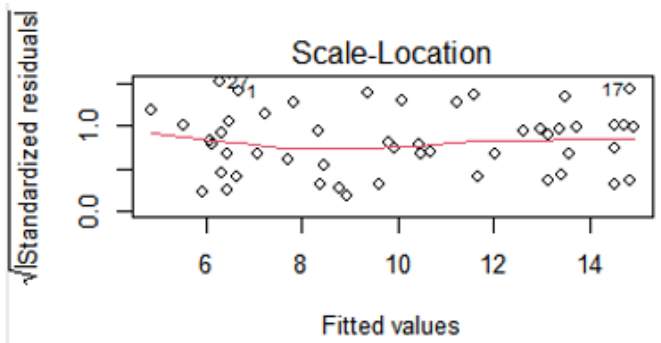
출 : 16, 28, 34

회귀 모형의 가정, 모델진단 그래프

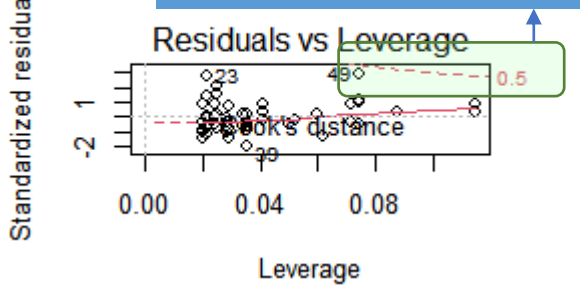
- 선형성 : 독립변수의 변화에 따라 종속변수도 변화하는 선형(linear) 모형이다
- 독립성 : 잔차와 독립변수의 값이 관련되어 있지 않다 (Durbin-Watson 통계량 확인)
- 정규성 : 잔차항이 정규분포를 이뤄야 한다
- 등분산성 : 잔차항들의 분포는 동일한 분산을 갖는다
- 비상관성 : 잔차들끼리 상관이 없어야 한다

Normal Q-Q plot	▪ 정규성(정상성), 잔차가 정규분포를 잘 따르고 있는지를 확인하는 그래프 ▪ 잔차들이 그래프 선상에 있어야 이상적임
Scale-Location	▪ 등분산성, y축이 표준화 잔차를 나타내며, 기울기 0인 직선이 이상적임
Cook's Distance	▪ 일반적으로 1값이 넘어가면 관측치를 영향점(influence points)로 판별



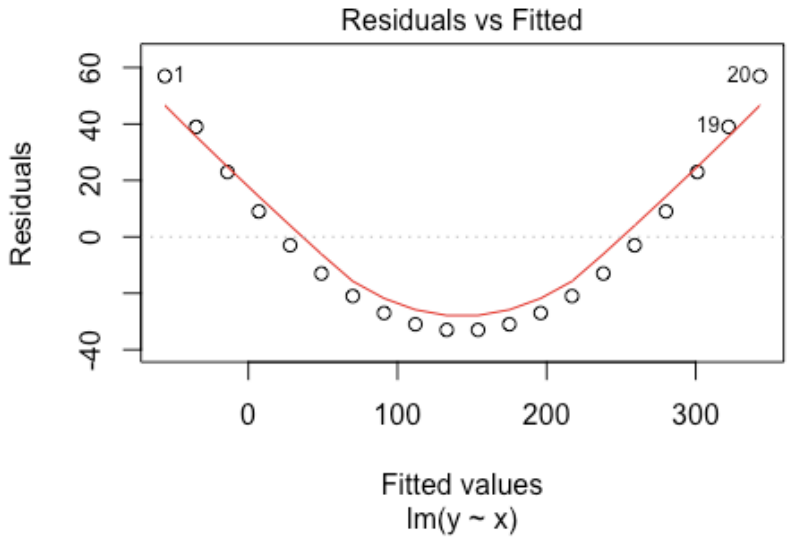


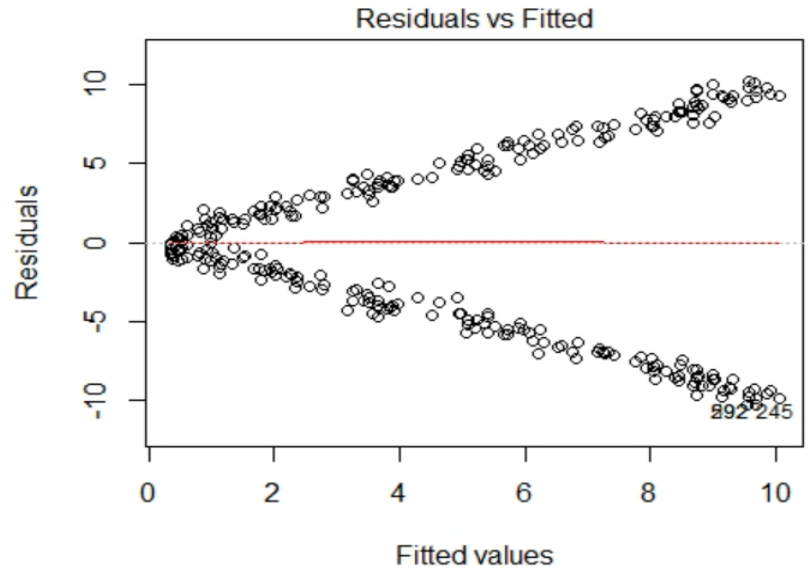
cook's distance



3-64. Residuals vs Fitted

- Residuals vs Fitted는 선형성, 등분산성에 대해 알아 볼 수 있는 그래프
  - 선형성 : y값의 기울기가 0인 직선이 이상적, 등분산성 : 점의 위치가 전체 그래프에 고르게 분포하는 것이 이상적





- lm(y~x)가 선형성, 잔차가 등분산성을 만족하지 않음
- U자 모형으로 제곱항을 넣어 보거나, 비선형으로 변환해 볼 수 있음

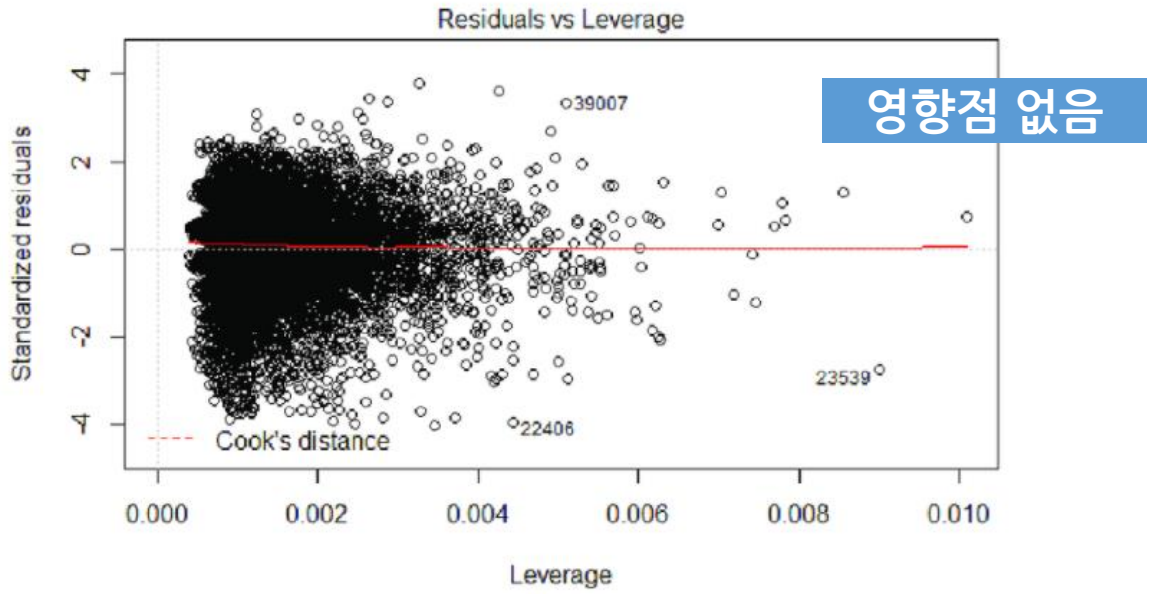
- 잔차가 등분산성을 만족하지 않음 (이분산성)
- 분산이 증가하고 있음
- 종속변수를 log로 변환하여 사용

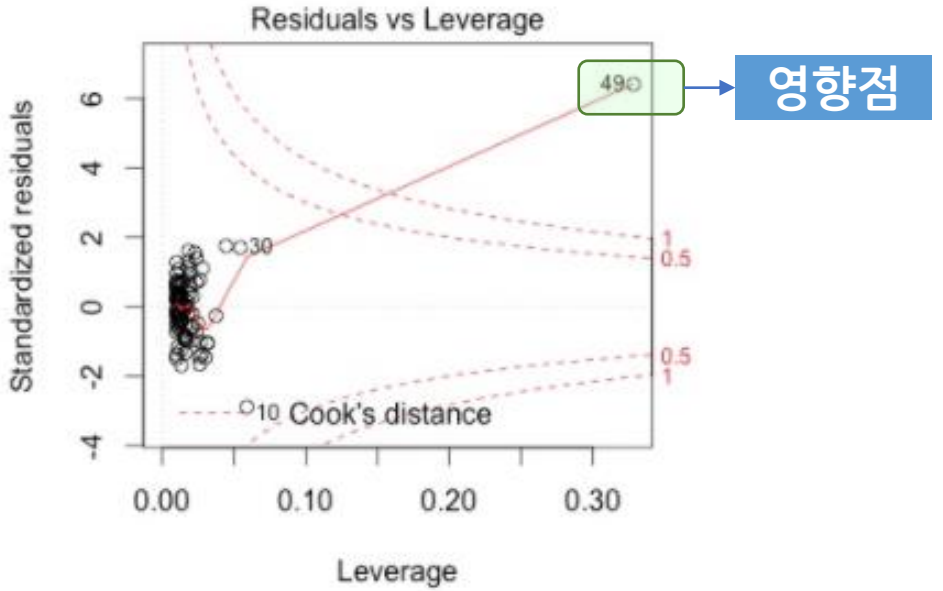
이상값(Outliler)

- 숫자와 함께 표시된 것

3-64. Residuals vs Leverage

- 회귀 분석에는 잔차(Residual)의 크기가 큰 데이터가 Outlier가 되는데 이 중에서도 주로 관심을 가지는 것은 Leverage와 Residual의 크기가 모두 큰 데이터임
- Leverage : 종속변수 값이 예측 값에 미치는 영향을 나타낸 값
- Cook's distance는 Leverage와 Residual를 동시에 보기 위한 기준으로 그림에서 빨간색 점선으로 표시임
- Leverage가 커지거나 Residual의 크기가 커지면 Cook's distance 값이 커짐
- 일반적으로 1값이 넘어가면 관측치를 영향점(influence points)로 판별





출처 : https://www.researchgate.net/

3-65. 회귀모형 해석(평가방법)

출 : 18*2, 19, 22, 23, 25

표본 회귀선의 유의성 검정

- 두 변수 사이에 선형관계가 성립하는지 검정하는 것으로 회귀식의 기울기 계수 $\beta_1 = 0$ 일 때 귀무가설, $\beta_1 \neq 0$일 때 대립가설로 설정한다

회귀모형 해석

모형이 통계적으로 유의미한가?	F 통계량, 유의확률(p-value)로 확인
회귀계수들이 유의미한가?	회귀계수의 t값, 유의확률(p-value)로 확인
모형이 얼마나 설명력을 갖는가?	결정계수(R^2) 확인
모형이 데이터를 잘 적합하고 있는가?	잔차 통계량 확인, 회귀진단 진행(선형성~ 정상성)

F 통계량, p-value

F 통계량 = 회귀제곱평균(MSR) / 잔차제곱평균(MSE)

F 통계량에 대한 p-value < 0.05

t 값, p-value

t 값 = Estimate(회귀계수) / Std.Error(표준오차)

t 값에 대한 p-value < 0.05

결정계수(R^2)

70~90%

3-65. 회귀모형 해석(평가방법)

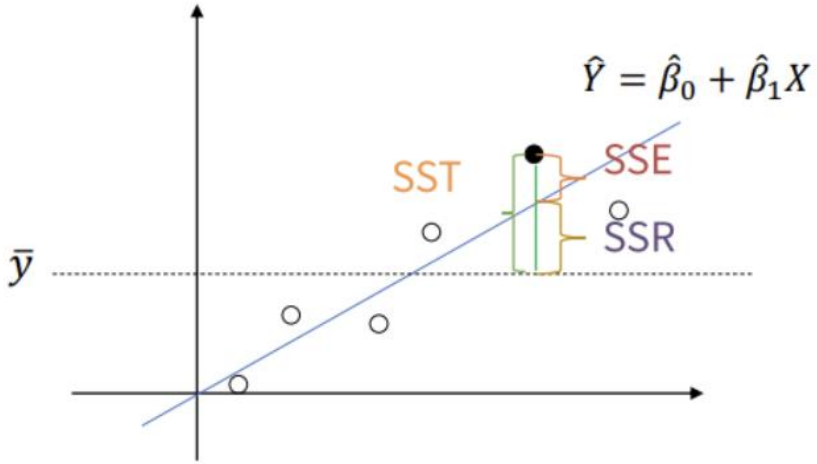
출 : 18, 19, 28, 35

F 통계량

- 모델의 통계적 유의성을 검정하기 위한 검정 통계량 (분산 분석)
- F통계량 = 회귀제곱평균(MSR) / 잔차제곱평균(MSE)
- F통계량이 클수록 회귀 모형은 통계적으로 유의하다, p-value < 0.05 일 때 유의함

결정계수
$R^2 = SSR/SST$

- 회귀식의 적합도를 재는 척도
- 결정계수(R^2) = 회귀제곱합(SSR) / 총제곱합(SST), 1-(SSE/SST)
- 결정계수는 0~1 사이의 범위를 갖음
- 전체 분산 중 모델에 의해 설명되는 분산의 양
- 결정계수가 커질수록 회귀방정식의 설명력이 높아짐



- SST : Total Sum of Squares, Y의 변동성
- SSE : Error Sum of Squares, X,Y를 통해 설명하지 못하는 변동성
- SSR : Regression Sum of Squares, Y를 설명하는 X의 변동성

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

3-65. 회귀모형 해석(평가방법)

출 : 17, 18*2, 19, 20, 23, 25, 26, 28, 30, 31, 33, 34

```
> a <- lm(y~u+v+w, df)
> summary(a)

Call:
lm(formula = y ~ u + v + w, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.06096 -0.31857  0.06092  0.32280  1.03220

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  3.43742    0.46949    7.322 3.01e-09 ***
u            0.46762    0.04419   10.581 6.55e-14 ***
v            0.95558    0.04546   21.019  < 2e-16 ***
w           -1.99230    0.01052 -189.459  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4695 on 46 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9987
F-statistic: 1.254e+04 on 3 and 46 DF,  p-value: < 2.2e-16
```

- t 통계량 = Estimate(회귀계수) / Std.Error(표준오차)
- t 통계량이 크다는 것은 표준오차가 작다는 의미
- t 통계량이 클수록 회귀계수가 유의하다

t 통계량

회귀계수들이 유의미한가?

R^2 모형이 얼마나 설명력을 갖는가?

F 통계량, 모형이 통계적으로 유의미한가?

- 다중회귀모형의 자유도 df = n - k - 1 # n은 sample 수, k는 독립변수의 수

3-66. 다중 공선성

출 : 18, 19, 22, 25, 28, 31

다중공선성(Multicollinearity)

- 모형의 일부 설명변수(=예측변수)가 다른 설명변수와 상관되어 있을 때 발생하는 조건
- 중대한 다중공선성은 회귀계수의 분산을 증가시켜 불안정하고 해석하기 어렵게 만들기 때문에 문제가 됨
- R의 vif 함수를 사용해 구할 수 있으며, VIF 값이 10이 넘으면 다중공선성이 존재한다고 봄

variance inflation factor

해결방법

- 높은 상관 관계가 있는 설명변수를 모형에서 제거하는 것으로 해결함
- 설명변수를 제거하면 대부분 R-square가 감소함
- 단계적 회귀분석을 이용하여 제거함

설명변수의 선택 원칙

- y 에 영향을 미칠 수 있는 모든 설명변수 x 들은 y의 값을 예측하는 데 참여시킴
- 설명변수 x 들의 수가 많아지면 관리에 많은 노력이 요구되므로 가능한 범위 내에서 적은 수의 설명변수를 포함시켜야 함
- 두 원칙이 이율배반적이므로 적절한 설명변수 선택이 필요함

3-67. 설명 변수 선택 방법

출 : 18, 19, 29, 20, 33, 35

모든 가능한 조합	▪ 모든 가능한 독립변수들의 조합에 대한 회귀모형을 고려해 AIC, BIC의 기준으로 가장 적합한 회귀 모형 선택 ▪ AIC, BIC : 최소자승법의 R^2와 비슷한 역할을 하며, 적합성을 측정해주는 지표로, R^2는 큰 값이 좋지만, AIC, BIC는 작은 값이 좋음
후진제거법	Backward Elimination, 독립변수 후보 모두를 포함한 모형에서 출발해 제곱합의 기준으로 가장적은 영향을 주는 변수로부터 하나씩 제거하면서 더 이상 유의하지 않은 변수가 없을 때까지 설명변수를 제거하고, 이때 모형을 선택
전진선택법	Forward Selection, 절편만 있는 모델에서 출발해 기준 통계치를 가장 많이 개선시키는 변수를 차례로 추가하는 방법
단계별 선택법	Stepwise method, 모든 변수가 포함된 모델에서 출발해 기준 통계치에 가장 도움이 되지 않는 변수를 삭제하거나, 모델에서 빠져 있는 변수 중에서 기준 통계치를 가장 개선 시키는 변수를 추가함
회귀모델에서 변수 선택을 위한 판단 기준	Cp, AIC, BIC 등이 있으며, 값이 작을 수록 좋음

3-67. 설명 변수 선택 방법

- 유의 확률 확인해 가며 하나씩 제거하는 후진제거법

유의확률이 가장 높은 x3 제거

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441
```

유의확률이 가장 높은 x4 제거

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .
x4           -0.2365     0.1733  -1.365 0.205395
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
x1           1.46831    0.12130   12.11 2.69e-07 ***
x2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

- F 통계량 및 p-value가 유의수준 5% 아래로 통계적으로 유의함
- 설명변수 $x1, x2$ 유의 확률 값이 유의함
- 최종 회귀식
  $y = 52.57735 + 1.46831x1 + 0.66225x2$

3-67. 설명 변수 선택 방법

step 함수를 사용한 후진제거법

```
> step(lm(y~x1+x2+x3+x4, df), direction='backward')
Start:  AIC=26.94
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq    RSS    AIC
- x3    1    0.1091 47.973 24.974
- x4    1    0.2470 48.111 25.011
- x2    1    2.9725 50.836 25.728
<none>              47.864 26.944
- x1    1   25.9509 73.815 30.576

Step:  AIC=24.97
y ~ x1 + x2 + x4

       Df Sum of Sq    RSS    AIC
<none>              47.97 24.974
- x4    1      9.93  57.90 25.420
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629
```

- 후진제거법 : direction = 'backward'
- 전진선택법 : direction = 'forward'
- 단계선택법 : direction = 'both'

최종 선택 설명 변수 : x1, x2, x4

```
Call:
lm(formula = y ~ x1 + x2 + x4, data = df)

Coefficients:
(Intercept)           x1           x2           x4
    71.6483       1.4519       0.4161      -0.2365
```

3-67. 설명 변수 선택 방법

step 함수를 사용한 전진선택법

```
> step(lm(y~1, df),
+      scope = list(lower=~1, upper=~x1+x2+x3+x4, direction='forward'))
```

```
Start:  AIC=71.44
y ~ 1

       Df Sum of Sq     RSS    AIC
+ x4    1   1831.90  883.87 58.852
+ x2    1   1809.43  906.34 59.178
+ x1    1   1450.08 1265.69 63.519
+ x3    1    776.36 1939.40 69.067
<none>              2715.76 71.444
```

```
Step:  AIC=58.85
y ~ x4

       Df Sum of Sq     RSS    AIC
+ x1    1    809.10   74.76 28.742
+ x3    1    708.13  175.74 39.853
<none>               883.87 58.852
+ x2    1     14.99  868.88 60.629
- x4    1   1831.90 2715.76 71.444
```

```
Step:  AIC=28.74
y ~ x4 + x1

       Df Sum of Sq     RSS    AIC
+ x2    1     26.79   47.97 24.974
+ x3    1     23.93   50.84 25.728
<none>               74.76 28.742
- x1    1    809.10  883.87 58.852
- x4    1   1190.92 1265.69 63.519
```

```
Step:  AIC=24.97
y ~ x4 + x1 + x2

       Df Sum of Sq    RSS    AIC
<none>              47.97 24.974
- x4    1      9.93  57.90 25.420
+ x3    1      0.11  47.86 26.944
- x2    1     26.79  74.76 28.742
- x1    1    820.91 868.88 60.629
```

최종 선택 설명 변수 : x1, x2, x4

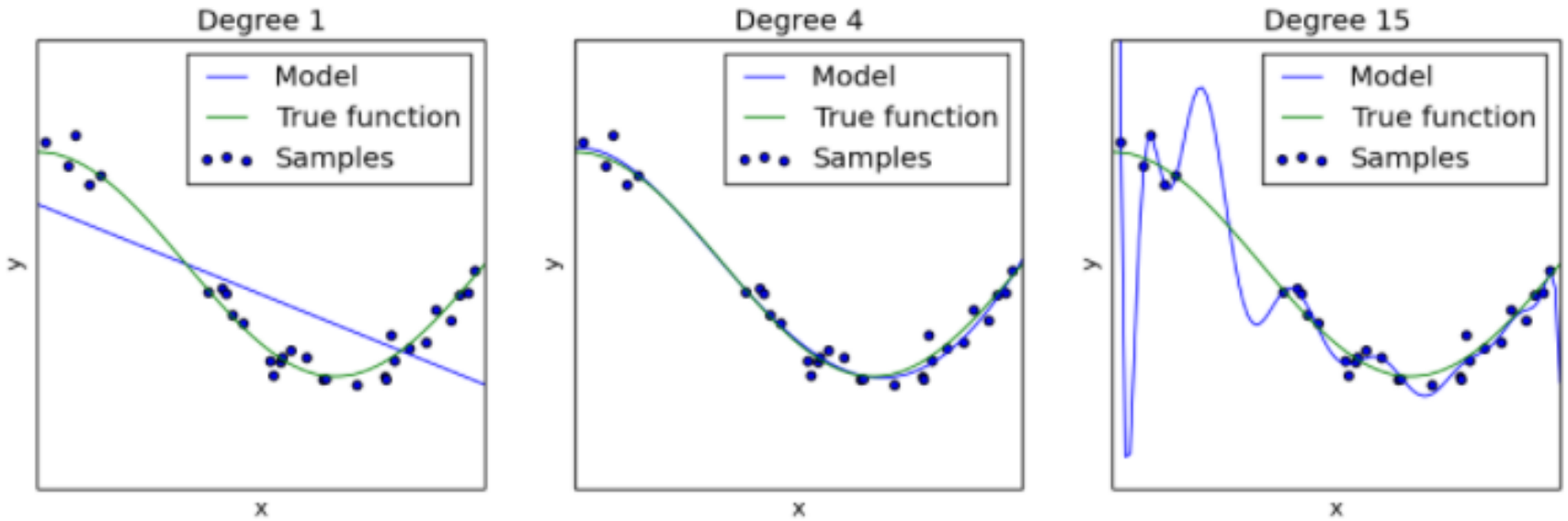
```
Call:
lm(formula = y ~ x4 + x1 + x2, data = df)

Coefficients:
(Intercept)           x4           x1           x2
    71.6483      -0.2365       1.4519       0.4161
```

3-68. 과적합(Overfitting)

출 : 23, 26, 32, 33, 35

과적합의 문제와 해결방법

- 주어진 샘플들의 설명변수와 종속변수의 관계를 필요이상 너무 자세하고 복잡하게 분석
- 샘플에 심취한 모델로 새로운 데이터가 주어졌을 때 제대로 예측해내기 어려울 수 있음
- 해결 방법으로 Feature의 개수를 줄이거나, Regularization을 수행하는 방법이 있음



출처 : **https://scikit-learn.org/0.15/auto_examples/plot_underfitting_overfitting.html**

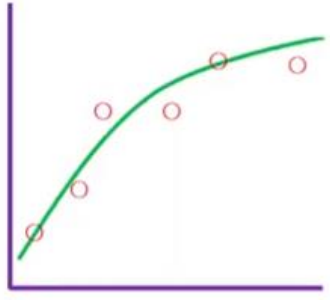
Underfitting

Overfitting

3-69. Regularization

정규화(Regularization) 개념

$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$



$\beta_0 + \beta_1 x + \beta_2 x^2$

$$\beta_1, \beta_2, \dots, \beta_p$$

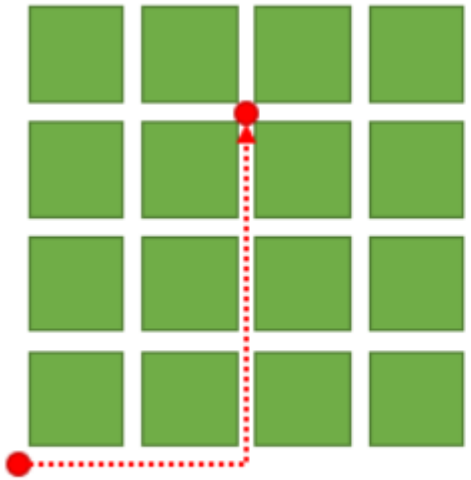
$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1} (y_i - \hat{y}_i)^2}_{\text{(1) Training accuracy}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{(2) Generalization accuracy}}$$

- 베타(β)값에 제약(penalty)을 주어 모델에 변화를 주는 것
- λ 값은 정규화 모형을 조정하는 hyper parameter
- λ 값이 클수록 제약이 많아져 적은 변수가 사용되고, 해석이 쉬워지지만 underfitting 됨
- λ 값이 작아질수록 제약이 적어 많은 변수가 사용되고, 해석이 어려워지며 overfitting 됨

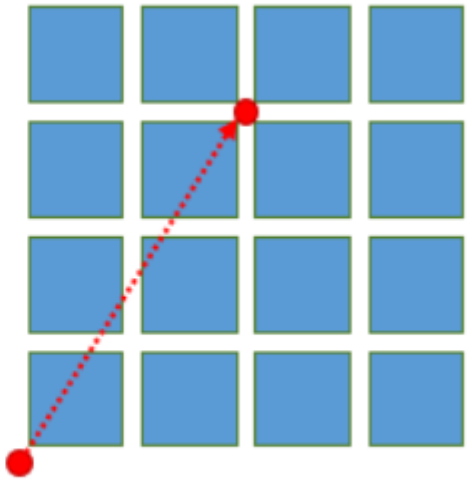
출처 : https://rk1993.tistory.com/entry/Ridge-regression%EC%99%80-Lasso-regression-%EC%89%BD%EA%B2%8C-%EC%9D%B4%ED%95%B4%ED%95%98%EA%B8%B0

3-69. L_1, L_2 Norm

- norm : 선형대수학에서 벡터의 크기(magnitude) 또는 길이(length)를 측정하는 방법
  - L_1 norm(=Manhattan norm) : 벡터의 모든 성분의 절대값을 더함
  - L_2 norm(=Euclidean norm) : 출발점에서 도착점까지의 거리를 직선거리로 측정함



L_1 norm



L_2 norm

L_1 norm $\quad \|*\|_1$

$$x = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad \|x\|_1 = |2| + |3| = 5$$

L_2 norm $\quad \|*\|_2$

$$x = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \qquad \|x\|_2 = \sqrt{(2)^2 + (3)^2} = \sqrt{13}$$

3-70. Regularized Linear Regression

출 : 16, 23, 29, 35

라쏘(Lasso) 회귀 특징

- 변수 선택이 가능하며, 변수간 상관관계가 높으면 성능이 떨어짐
- L_1 norm을 패널티를 가진 선형 회귀 방법, 회귀계수의 절댓값이 클수록 패널티 부여
- MSE가 최소가 되게 하는 w, b를 찾는 동시에 w의 절대값들의 합이 최소가 되게 해야함
- w의 모든 원소가 0이 되거나 0에 가깝게 되게 해야 함 => 불필요 특성 제거
- 어떤 특성은 모델을 만들 때 사용되지 않게 됨

라쏘(Lasso) 회귀 장점

- 제약 조건을 통해 일반화된 모형을 찾는다
- 가중치들이 0이 되게 함으로써 그에 해당하는 특성들을 제외해준다
- 모델에서 가장 중요한 특성이 무엇인지 알게 되는 등 모델 해석력이 좋아진다

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

3-70. Regularized Linear Regression

Ridge 회귀 특성

- L_2 norm을 사용해 패널티를 주는 방식
- 변수 선택이 불가능, 변수 간 상관관계가 높아도 좋은 성능
- Lasso 는 가중치들이 0이 되지만, Ridge의 가중치들은 0에 가까워질 뿐 0이 되지는 않음
- 특성이 많은데 특성의 중요도가 전체적으로 비슷하다면 Ridge 가 좀 더 괜찮은 모델을 찾아줄 것이다.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

엘라스틱넷 특성

- L_1, L_2 norm regularization
- 변수 선택 가능
- 변수 간 상관관계를 반영한 정규화

3-71. 데이터 스케일링(Scaling)

출 : 21, 28

- 데이터 단위의 불일치 문제를 해결하는 방법
- 분석에 사용되는 변수들에 사용 단위가 다를 때 데이터를 같은 기준으로 만듦
- 원 데이터의 분포를 유지하는 정규화 방법

정규화 normalization

- 값의 범위를 [0, 1]로 변환하는 것, min-max nomalization
- $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$ = 0, 100점일 경우 50점은? 50 - 0 / (100) = 0.5

표준화 standardization

- 특성의 값이 정규분포를 갖도록 변환하는 것, 평균 0, 표준편차 1
- $Z' = \frac{X - \mu}{\sigma}$ = 평균 80, 표준편차 10 일 경우 90점은? 90 - 80 / 10 = 1

```
> iris_s = cbind(as.data.frame(scale(iris[1:4])), iris$Species)
> head(iris_s)
  Sepal.Length Sepal.Width Petal.Length Petal.Width iris$Species
1   -0.8976739  1.01560199    -1.335752   -1.311052       setosa
2   -1.1392005 -0.13153881    -1.335752   -1.311052       setosa
3   -1.3807271  0.32731751    -1.392399   -1.311052       setosa
```

3-71. 회귀 모델 평가 지표

- MAPE(Mean Absolute Percentage Errors)
  - MSE, RMSE와 같이 큰 에러에 패널티를 부여하는 평가지표의 단점을 극복하기 위한 방법
  - 각 항의 '실제값과 예측값의 차이/실제값'에 대한 절대값을 모두 합하여 데이터의 개수 n으로 나누고 100을 곱한 값

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

y_i : 실제 값, actual value
$\hat{y}_i$: 예측 값, forecast value

actual	1	2	4	8
forecast	0.7	2.5	3.6	10

- n = 4
- |(1-0.7)/1| + |(2-2.5)/2| + |(4-3.6)/4| + |(8-10)/8|
- 0.3/1 + 0.5/2 + 0.4/4 + 2/8 = 0.3 + 0.25 + 0.1 + 0.25
- MAPE = 0.9 / 4 = 0.225 * 100% = 22.5%

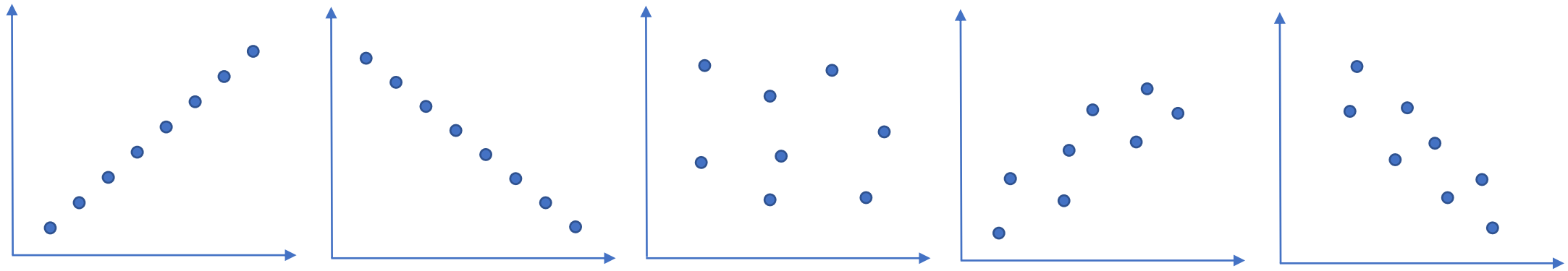
03-02

03 상관관계를 이용하는 다변량 분석

출 : 18, 19, 22*2, 23, 26, 29, 34*2

상관 계수의 이해

- 상관계수는 두 변수의 관련성의 정도를 의미함 (-1 ~ 1의 값으로 나타냄)
- 두 변수의 상관관계가 존재하지 않을 경우 상관계수는 '0 ' 임
- 상관관계가 높다고 인과관계가 있다고 할 수는 없음
- 피어슨 상관계수와 스피어만 상관계수가 있음
- 피어슨 상관계수는 두 변수 간의 선형적인 크기만 측정 가능하며 스피어만 상관계수는 두 변수 간의 비선형적인 관계도 나타낼 수 있음
- R의 cor.test() 함수를 사용해 상관계수 검정을 수행하고, 유의성검정을 판단할 수 있음
- 이때 귀무가설은 '상관계수가 0이다'. 대립가설은 '상관계수가 0이 아니다'



3-72. 상관 분석

출 : 17, 18, 21, 23, 27, 29, 31, 33

스피어만 상관계수

- 대상자료는 서열척도 사용, 두 변수 간의 비선형적인 관계를 나타낼 수 있음
- 연속형 외에 이산형도 가능함
- 스피어만 상관 계수는 원시 데이터가 아니라 각 변수에 대해 순위를 매긴 값을 기반으로 함
- 두 변수 안의 순위가 완전 일치하면 1, 완전 반대이면 -1
- 예) 수학 잘하는 학생이 영어도 잘하는 것과 상관있는지 알아보는데 사용될 수 있음

피어슨 상관계수

- 대상자료는 등간척도, 비율척도 사용, 두 변수 간의 선형적인 크기만 측정 가능
- 피어슨 상관계수 : x, y 의 공분산을 x, y의 표준편차의 곱으로 나눈 값 $corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$
- 응답자1의 표준편차 2, 응답자2의 표준편차 2, 두 응답자의 공분산 값 4 이면
  피어슨 상관계수(p) = 4 / (2 * 2) = 1

공분산

- Covariance, 2개의 확률변수의 선형 관계를 나타내는 값
- $cov(x,y) = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{n}$
- 하나의 변수가 상승하는 경향을 보일 때 다른 값도 상승하는 선형 상관성이 있다면 양의 공분산을 갖음
- 공분산이 0이면 서로 독립이며, 관측값들이 4면에 균일하게 분포되어 있다고 추정할 수 있음

3.72. 상관분석의 예

- 귀무가설 : 상관계수가 0이다.

```
> cor.test(c(1,3,5,7,9), c(1,2,4,6,8), method='pearson')

        Pearson's product-moment correlation

data:  c(1, 3, 5, 7, 9) and c(1, 2, 4, 6, 8)
t = 15.588, df = 3, p-value = 0.0005737
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9065015 0.9996163
sample estimates:
      cor
0.9938837
```

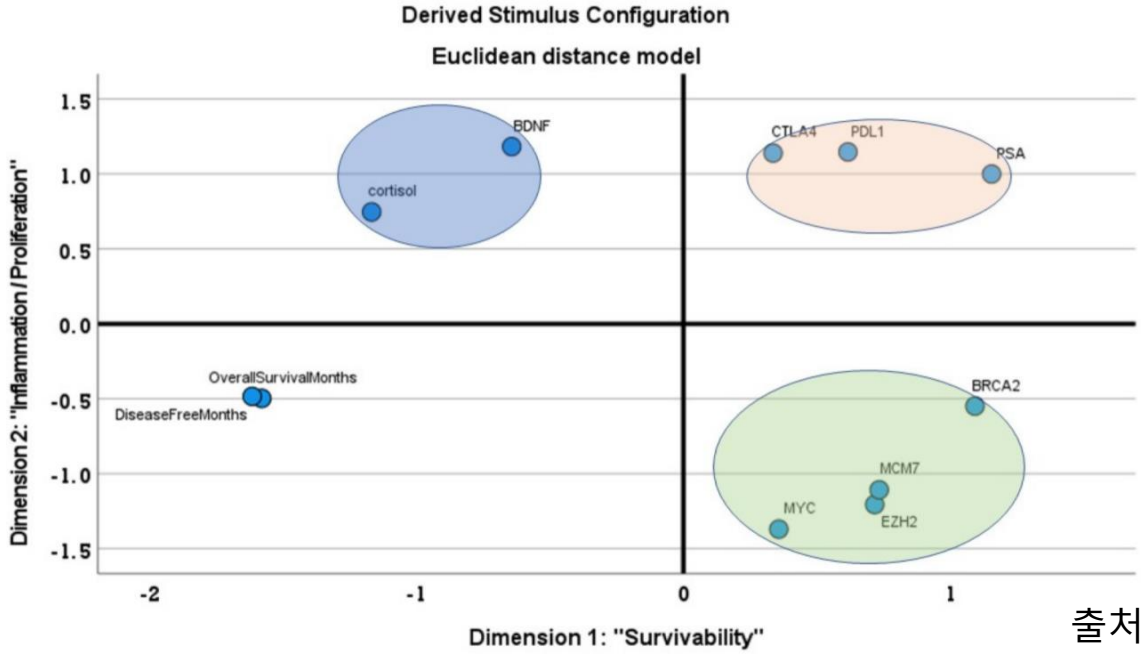
3-73. 차원축소 목표를 위해 개발된 분석 방법

1. 주성분분석(Principal Component Analysis)
2. 요인분석(Factor Analysis)
3. 판별분석(Discriminant Analysis)
4. 군집분석(Cluster Analysis)
5. 정준상관분석(Canonical Correlation Analysis)
6. 다차원척도법(Multi-dimensional scaling)

출 : 21, 35

다차원 척도법, MDS(Multi-Dimensional Scaling)

- 객체 간 근접성을 시각화 하는 기법
- 개체들 사이의 유사성, 비유사성을 2차원 혹은 3차원 공간상에 점으로 표현하여 개체 사이의 군집을 시각적으로 표현하는 분석방법
- 개체들의 거리는 유클리드(Euclidean) 거리와 유사도를 이용하여 구함
- 관측 대상의 상대적 거리의 정확도를 높이기 위해 적합 정도를 스트레스 값(Stress Value)로 나타내며, 0 에 가까울수록 적합도가 좋음



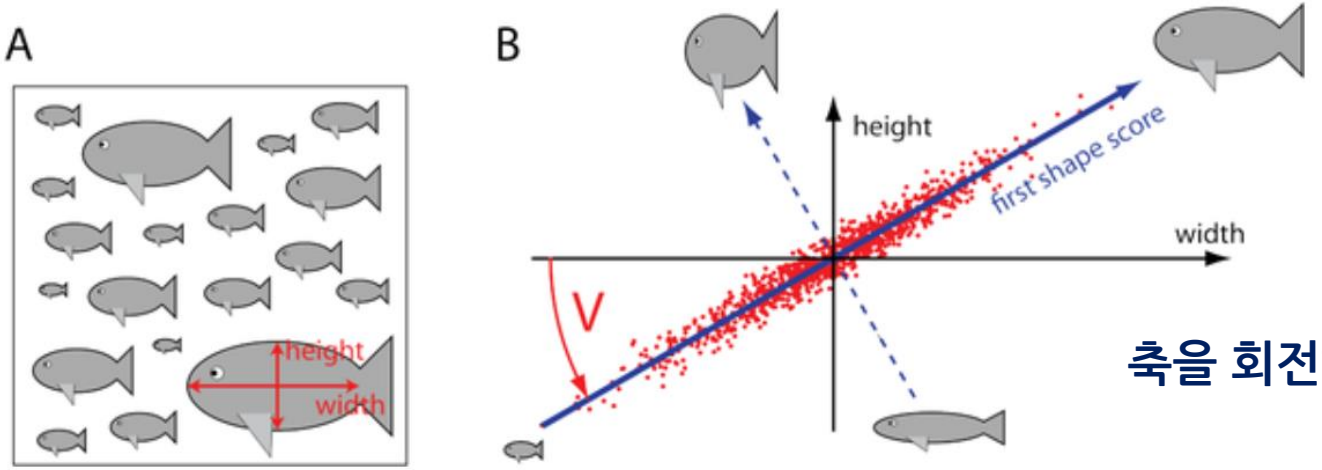
출처 : https://www.mdpi.com/

3-74. 주성분 분석(PCA)

출 : 19, 20, 26, 32*2, 33*2, 34
(434, 435 페이지)

주성분 분석(PCA, Principal Component Analysis)

- 데이터를 분석할 때 변수의 개수가 많다고 모두 활용하는 것이 꼭 좋은 것은 아님
- 오히려 변수가 '다중공선성'이 있을 경우 분석 결과에 영향을 줄 수도 있음
- 공분산행렬 또는 상관계수 행렬을 사용해 모든 변수들을 가장 잘 설명하는 주성분을 찾는 방법
- 상관관계가 있는 변수들을 선형 결합에 의해 상관관계가 없는 새로운 변수(주성분)를 만들고 분산을 극대화하는 변수로 축약함
- 주성분은 변수들의 선형결합으로 이루어져 있음
- 독립변수들과 주성분과의 거리인 '정보손실량'을 최소화하거나 분산을 최대화 함



축을 회전하여 직교하는 성분으로 변환

3-74. 주성분 분석(PCA)

주성분분석 할 때 고민해야 하는 것

- 공분산행렬과 상관계수행렬 중 어떤 것을 선택할 것인가?
- 주성분의 개수를 몇 개로 할 것인가?
- 주성분에 영향을 미치는 변수로 어떤 변수를 선택할 것인가?

공분산 행렬(default) VS 상관계수 행렬

- 공분산 행렬은 변수의 측정단위를 그대로 반영한 것이고, 상관계수 행렬은 모든 변수의 측정단위를 표준화한 것이다
- 공분산행렬을 이용한 경우 측정 단위를 그대로 반영하였기 때문에 변수들의 측정 단위에 민감하다
- 주성분 분석은 거리를 사용하기 때문에 척도에 영향을 받는다 (정규화 전후의 결과가 다르다)
- 설문조사처럼 모든 변수들이 같은 수준으로 점수화 된 경우 공분산행렬을 사용한다
- 변수들의 scale이 서로 많이 다른 경우에는 상관계수행렬(correlation matrix)을 사용한다

주성분 분석에서 상관계수 행렬 사용

- prcomp(data, scale=TRUE)
- princomp(data, cor=TRUE)

3-74. 주성분 분석(PCA)

출 : 23, 27, 29, 30*2, 31, 33, 34

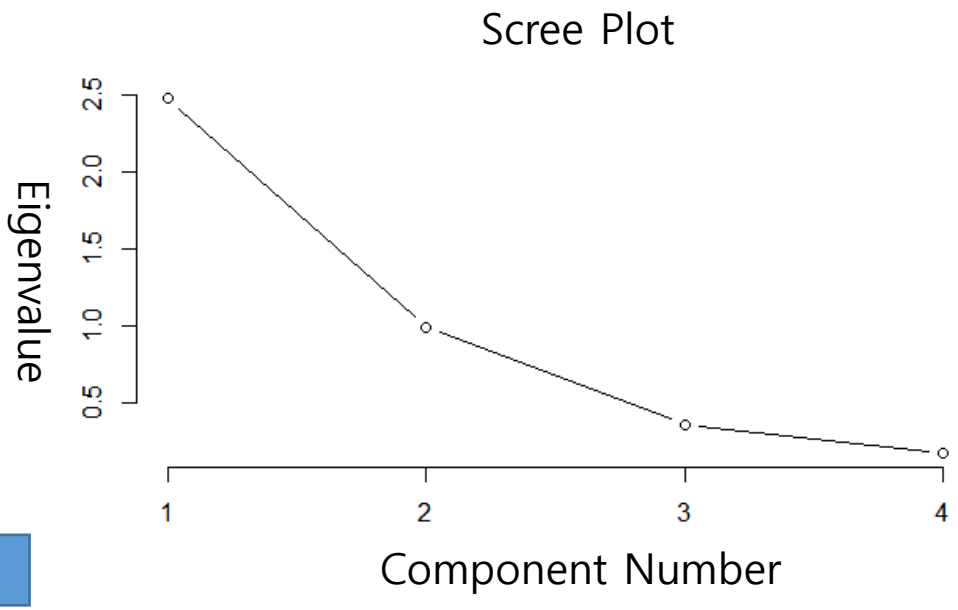
주성분 결정 기준

성분들이 설명하는 분산의 비율	▪ 누적 분산 비율을 확인하면 주성분들이 설명하는 전체 분산 양을 알 수 있음 ▪ 누적 분산 비율이 70~90% 사이가 되는 주성분 개수 선택
고윳값(Eigenvalue)	▪ 분산의 크기(=중요도 기준)를 나타내며, 고윳값이 1보다 큰 주성분만 사용함
Scree Plot	▪ 고윳값을 가장 큰 값에서 가장 작은 값을 순서로 정렬해 보여줌 (1보다 큰 값 사용)

```
> fit<-prcomp(USArrests, scale=TRUE)
> summary(fit)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.5749 0.9949 0.59713 0.41645
Proportion of Variance 0.6201 0.2474 0.08914 0.04336
Cumulative Proportion  0.6201 0.8675 0.95664 1.00000

> fit$rotation
                PC1        PC2        PC3         PC4
Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

Y1 = -0.536Murder – 0.583Assault – 0.278UrbanPop – 0.543Rape

```
> plot(fit,type='lines')
```



3-74. 주성분 분석(PCA) 해석

출 : 19*2, 20, 26, 32*2, 33*2, 34

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.5748783	0.9948694	0.5971291	0.41644938
Proportion of Variance	0.6200604	0.2474413	0.0891408	0.04335752
Cumulative Proportion	0.6200604	0.8675017	0.9566425	1.00000000

- Standard deviation (표준편차) : 자료의 산포도를 나타내는 수치로, 분산의 양의 제곱근, 표준편차가 작을수록 평균값에서 변량들의 거리가 가깝다.
- Proportion of Variance(분산비율) : 각 분산이 전체 분산에서 차지하는 비중
- Cumulative Proportion(누적비율) : 분산의 누적 비율

- 첫 번째 주성분 분석 하나가 전체 분산의 62%를 설명하고 있다
- 두 번째는 24.7%를 설명하고 있다
- 반대로 이야기 하면 첫 번째 주성분 부분만 수용했을 때 정보 손실은 (100-62) = 38%가 된다

03-02

04 시계열 예측

3-75. 시계열 자료(time series)

출 : 16, 18, 22, 23, 25*2, 29, 32, 34, 35

시계열 자료

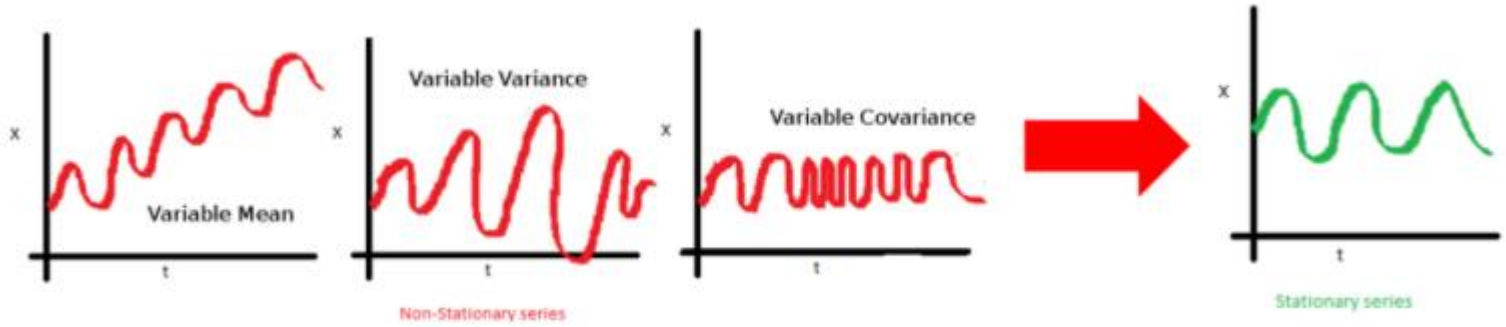
- 시간의 흐름에 따라 관측된 데이터
- 시계열 분석을 위해서는 정상성을 만족해야 함

정상성(stationary)

- 시계열의 수준과 분산에 체계적인 변화가 없고, 주기적 변동이 없다는 것
- 미래는 확률적으로 과거와 동일하다는 것

정상 시계열의 조건

- 평균은 모든 시점(시간 t)에 대해 일정하다
- 분산은 모든 시점(시간 t)에 대해 일정하다
- 공분산은 시점(시간 t) 에 의존하지 않고, 단지 시차에만 의존한다



이미지 출처 : https://sodayeong.tistory.com/34

3-75. 정상 시계열 전환

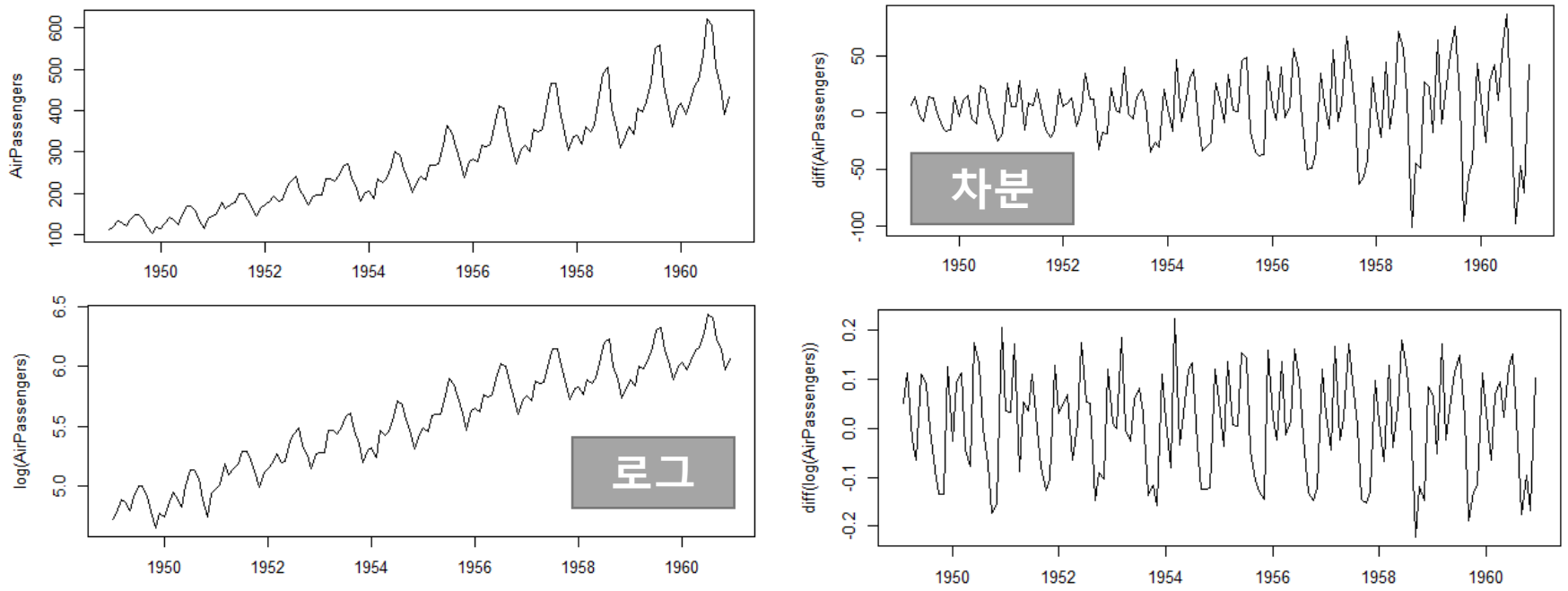
출 : 18, 23, 31

정상시계열로 전환하는 방법

- 비정상시계열 자료는 정상성을 만족하도록 데이터를 정상시계열로 만든 후 시계열 분석을 수행한다
- 평균이 일정하지 않은 경우 : 원계열에 차분 사용
- 계절성을 갖는 비정상시계열 : 계절 차분 사용
- 분산이 일정하지 않은 경우 : 원계열에 자연로그(변환) 사용

차분

현 시점의 자료 값에서 전 시점의 자료 값을 빼 주는 것 의미함



이미지 출처 : https://sodayeong.tistory.com/34

3-76. 시계열 모형

출 : 19, 22, 26*2, 32

AR 모형
자기회귀모형

- AR(p) : 현 시점의 자료를 p 시점 전의 유한 개의 자기 자신의 과거 값을 사용하여 설명
- 백색 잡음의 현재 값과 자기 자신의 과거 값의 선형 가중 값으로 이루어진 정상 확률 모형
- 현 시점의 시계열 자료에 과거 1시점 이전의 자료만 영향을 주면 이를 1차 AR 모형이라 하고 AR(1) 라고 표기함

MA 모형
이동평균 모형

- MA(q) : 과거 q 시점 이전 오차들에서 현재항의 상태를 추론한다
- 최근 데이터의 평균을 예측치로 사용하는 방법, 각 과거치는 동일 가중치가 주어짐
- 현시점의 자료가 유한 개의 과거 백색잡음의 선형결합으로 표현되었기 때문에 항상 정상성을 만족함

AR $y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$

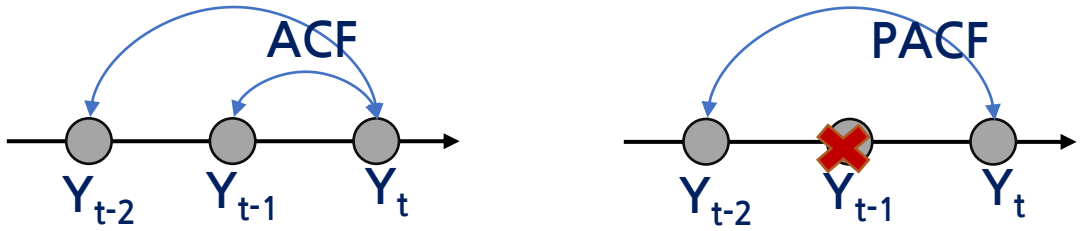
MA $y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$

ARIMA 모형
자기회귀 누적
이동평균 모형
Auto-
Regressive
Integrated
Moving
Average

- 현재와 추세간의 관계를 정의한 것, 많은 시계열 자료가 ARIMA모형을 따름
- ARIMA 모형은 비정상시계열 모형이며 차분/변환을 통해 AR, MA, ARMA 모형으로 정상화 할 수 있음
- ARIMA(p, d, q) ➔ p : AR모형 차수, d : 차분, q: MA모형 차수
- ARIMA(1, 2, 3) 이라면 2번 차분해서 ARMA 모형이 될 수 있음
- ARIMA(0, 1, 3) : IMA(1, 3) 모형이고 이것을 1번 차분하면 MA(3) 모형이 됨
- ARIMA(2, 3, 0) : ARI(2, 3) 모형이고, 이것을 3번 차분하면 AR(2) 모형이 됨

3-76. ACF, PACF, White Noise

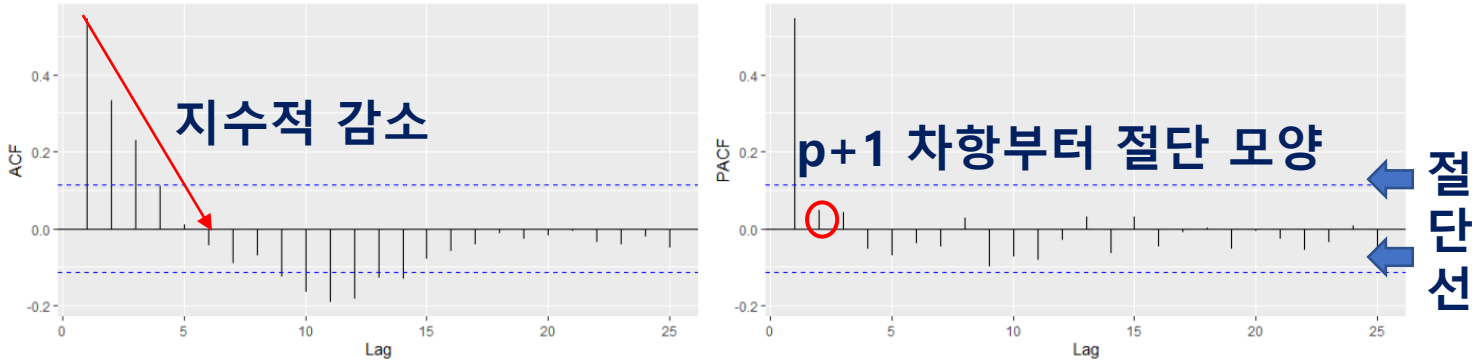
자기 상관 함수 ACF	▪ Auto-Correlation Function, 시계열 데이터의 자기상관성을 파악하기 위한 함수 ▪ 시계열의 관측치 Y_t와 Y_{t-k} 간 상관계수를 k의 함수 형태로 표시한 것, k : 시간단위 ▪ $-1 \leq autocorr(Y_t, Y_{t-k}) \leq 1$, k가 커질 수록 ACF는 0으로 수렴함
부분 자기 상관 함수 PACF (Partial ACF)	▪ Y_t와 Y_{t-k} 중간에 있는 값들의 영향을 제외시킨 Y_t와 Y_{t-k} 사이의 직접적 상관관계를 파악하기 위한 함수
백색잡음 (White Noise)	▪ 시계열 자료 중 자기상관이 전혀 없는 특별한 경우 ▪ 시계열의 평균이 0, 분산이 일정한 값, 자기공분산이 0인 경우 ▪ 현재 값이 미래 예측에 전혀 도움이 되지 못함, 회귀분석의 오차항과 비슷한 개념



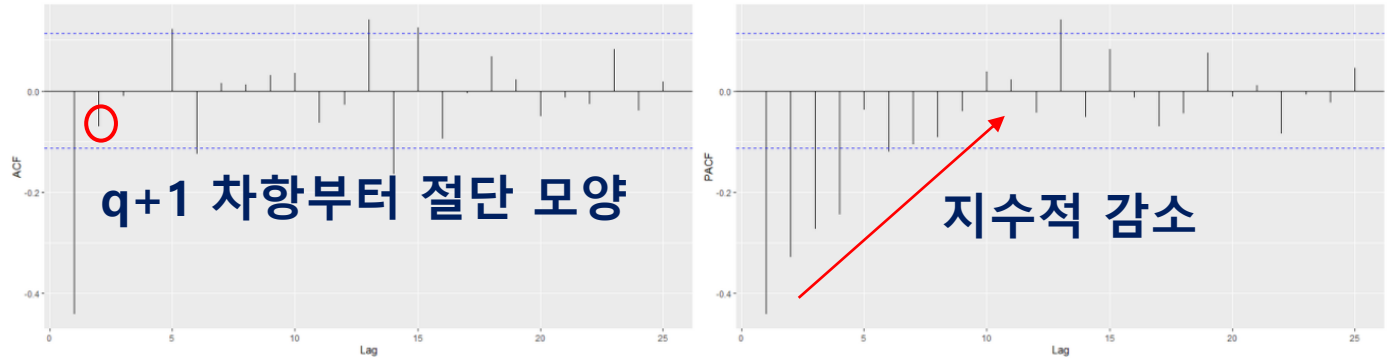
3-77. 시계열 모형

시계열 모형	자기 상관 함수(ACF)	부분 자기 상관 함수(PACF)
자기회귀(AR)	지수적 감소, 0에 접근	p+1차항부터 절단 모양
이동평균(MA)	q+1차항부터 절단 모양	지수적 감소, 0에 접근
자기회귀이동평균(ARMA)	p+1차항부터 절단 모양	q+1차항부터 절단 모양

AR(1) 모형의 예



MA(1) 모형의 예



이미지 출처 : https://sodayeong.tistory.com/36

3-78. 분해 시계열

출 : 17, 19, 22, 23, 30, 31, 32, 33

분해 시계열	시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법

분해시계열 분해 요인 (종류는 반드시 암기!)

추세요인 Trend Factor	자료의 그림을 그렸을 때 그 형태가 오르거나 내리는 등 자료가 어떤 특정한 형태를 취할 때
계절요인 Seasonal Factor	계절에 따라, 고정된 주기에 따라 자료가 변화하는 경우
순환요인 Cyclical Factor	물가상승률, 급격한 인구 증가 등의 이유로 알려지지 않은 주기를 가지고 자료가 변화하는 경우
불규칙요인 Irregular Factor	위 세 가지 요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인에 의해 발생하는 경우