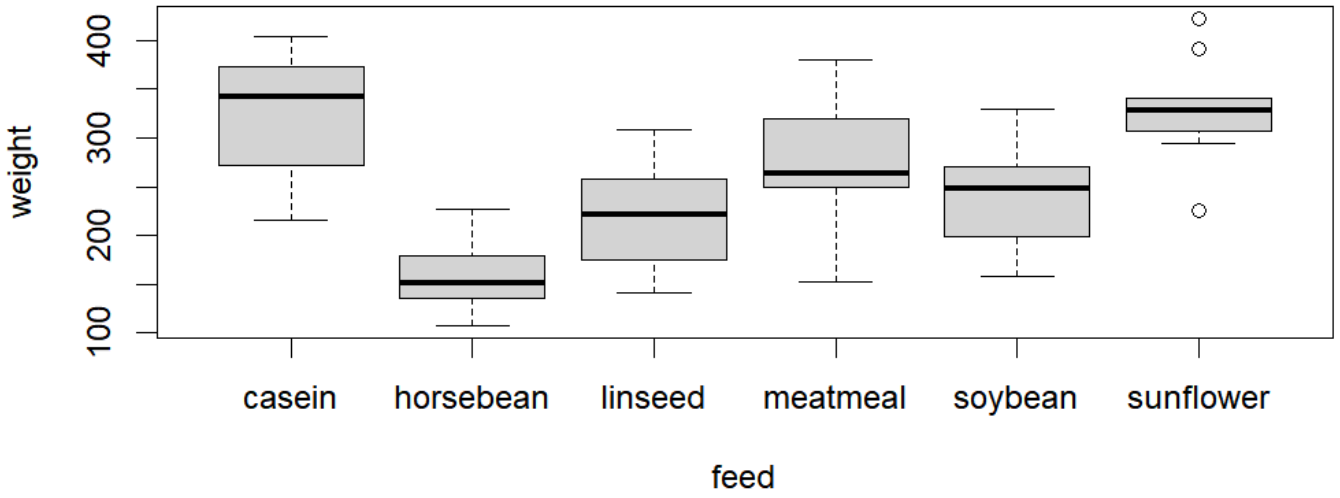
결과 해석

2020년, 2021년 기출 관련

Boxplot



- sunflower : 이상치를 가지고 있음
- casein : median 이 다른 것들보다 큼
- horsebean : 다른 것들보다 weight가 작음

summary

```
> summary(Hitters)
     AtBat            Hits         HmRun          NewLeague     Salary       
 Min.   : 16.0   Min.   :  1   Min.   : 0.00   A:176     Min.   :  67.5  
 1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   N:146     1st Qu.: 190.0  
 Median :379.5   Median : 96   Median : 8.00             Median : 425.0  
 Mean   :380.9   Mean   :101   Mean   :10.77             Mean   : 535.9  
 3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00             3rd Qu.: 750.0  
 Max.   :687.0   Max.   :238   Max.   :40.00             Max.   :2460.0  
                                                         NA's   :59      
```

범주형

결측치 59개

- AtBat, Hits, HuRun : **연속형 변수**, Min, 1st Qu., Median, Mean, 3rd Qu., Max. 값이 표시되어 있음
- Salary : **연속형 변수 + 결측치**, 결측치는 **NA's : 59**로 표시되어 있음
- NewLeague : **범주형**, A 범주가 176개, N 범주가 146개 있음

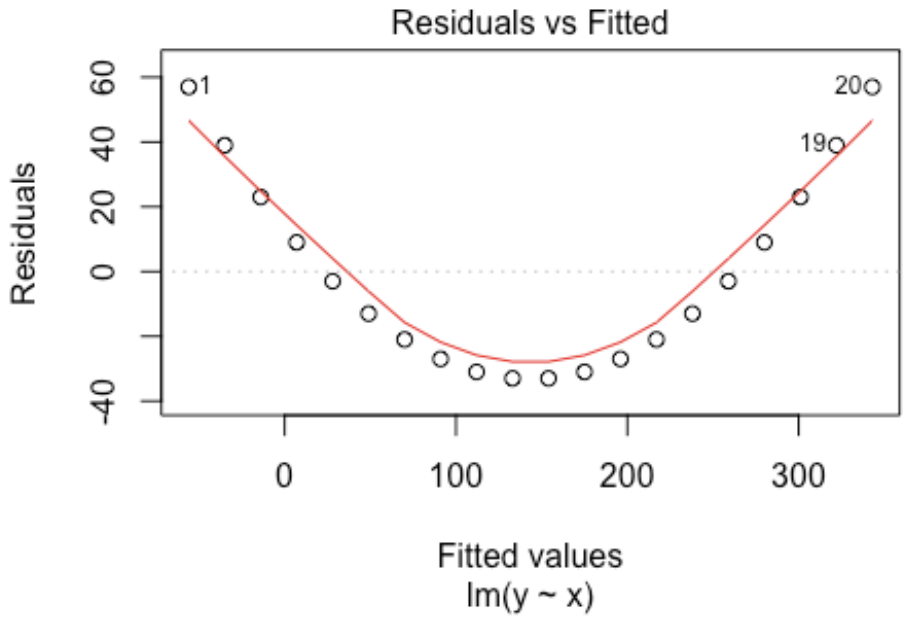
summary

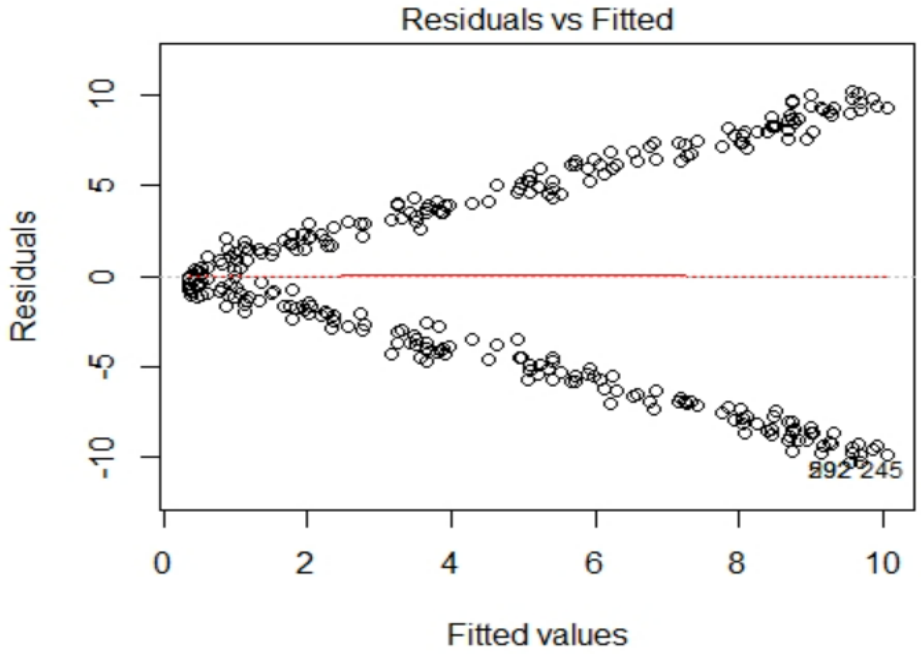
```
> summary(Orange)
 Tree       age         circumference
 3:7   Min.   : 118.0   Min.   : 30.0
 1:7   1st Qu.: 484.0   1st Qu.: 65.5
 5:7   Median :1004.0   Median :115.0
 2:7   Mean   : 922.1   Mean   :115.9
 4:7   3rd Qu.:1372.0   3rd Qu.:161.5
       Max.   :1582.0   Max.   :214.0
```

- age 평균 : 922.1
- 35개의 관측치 포함 (Tree가 5가지 종류이며 7개씩임)
  Factor 타입이며 순서가 3 > 1 > 5 > 2 > 4로 되어 있음
- 약 50%의 나무가 115보다 큰 둘레를 가지고 있음

- 틀린 지문 : age는 Tree별로 평균이 유의한 차이가 없다고 할 수 있다 (X)

잔차도(Residuals vs Fitted)





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

잔차의 등분산성을 만족하지 않음

주성분 분석(PCA) 해석

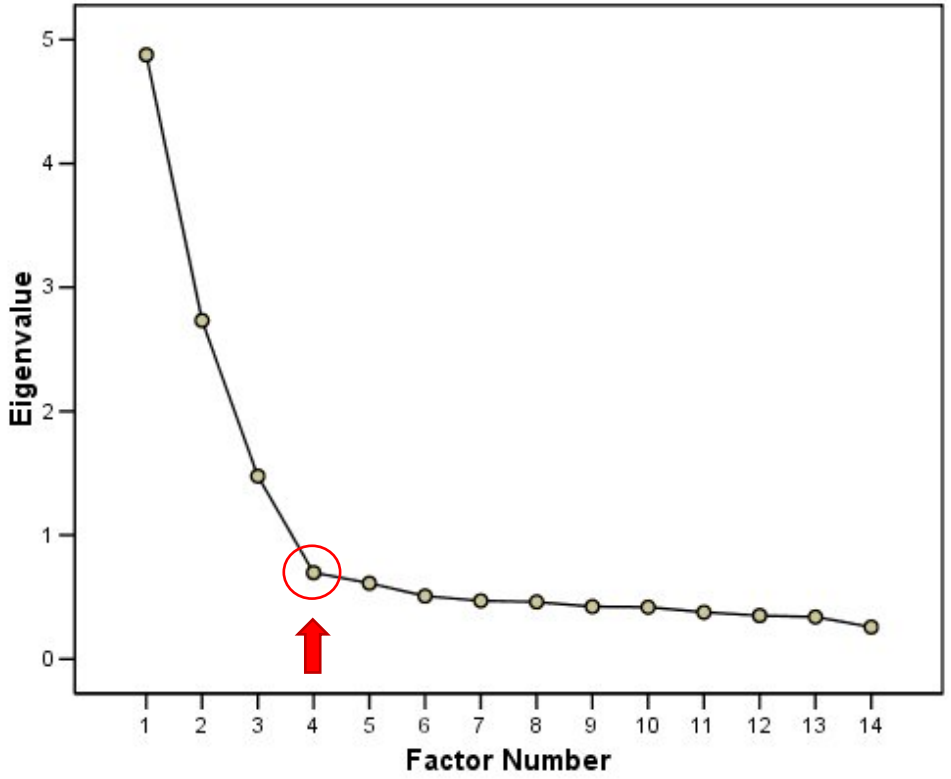
```
> data3 <- princomp(data1, cor=TRUE)   # ISLR 패키지 data (Hitters)
> data3
Call :
princomp(x = data1, cor = TRUE)

Standard deviations:
                    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7
                2.77339679 2.03026013 1.31485574 0.95454099 0.84109683  0.7237422 0.69841796
생략
17 variables and 263 observations.
> summary(data3)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.7733967  2.0302601  1.3148557  0.9575410 0.84109683
Proportion of Variance 0.4524547  0.2424680  0.1016968  0.0539344 0.04161435
Cumulative Proportion  0.4524547  0.6949227  0.7966195  0.8505539 0.89216822
```

- 옆의 결과에서 princomp(data1, cor=TRUE)라는 것이 주성분분석의 함수입니다. cor=TRUE라는 것은 상관계수 행렬을 사용하겠다는 것입니다.
- 이것은 prcomp(data1, scale=TRUE)라는 것과 동일한 동작을 합니다. prcomp의 경우 scale=TRUE라는 것이 상관행렬을 사용하겠다는 것을 의미합니다.
- cor=FALSE, scale=FALSE를 사용하거나 생략하면 공분산 행렬을 사용하겠다는 의미가 됩니다.
- 공분산 행렬은 변수의 측정 단위를 그대로 반영한 것이고,
- 상관행렬을 사용하는 경우 모든 측정 단위를 표준화 한 것입니다.
- 따라서, 공분산 행렬을 이용한 분석의 경우 변수들의 측정 단위에 민감한 특성이 있습니다.

summary(data3) 에 대한 해석

- Comp.1, Comp.2, ... Comp.5 가 주성분이며 뒤쪽이 생략되어 있습니다.
  (Cumulative Proportion)이 1이 보여야 모든 성분에 대한 내용이 표시된 것입니다.
- Proportion of Variance 가 각 성분의 분산으로 설명력을 의미합니다.
- 주성분은 가장 분산이 높은 것부터 작은 순서로 주성분1, 2, 3 ... 이 됩니다.
- Comp.1 의 경우 0.4524547, Comp.2 의 경우 0.2424680 ... 등으로 표시되고 있습니다
- Cumulative Proportion은 Proportion of Variance를 누적한 것입니다.
  Comp.1은 그대로 표시되고 Comp.2의 경우 Comp.1 + Comp.2의 Proportion of Variance를 더한 것이고,
  Comp.3의 Comp.1 + Comp.2 + Comp.3 의 Proportion of Variance를 더한 것이 됩니다.
- 설명력을 이야기 할때는 Cumulative Proportion을 보면 됩니다.
- 주성분은 1번부터 순서대로 사용됩니다. 따라서 주성분을 4개 사용한다면 Comp.1에서 Comp.4까지 사용한 것이 됩니다.
- 주성분을 4개 사용했을 때의 설명력은 Comp.4의 Cumulative Proportion을 보면됩니다. (위의 그림에서는 0.8505539가 됩니다. 약 85.05% 입니다.)
- 차원을 2차원으로 줄였다는 것은 2개의 주성분만 사용하겠다는 것입니다.
- 설명력은 전체 주성분을 사용해야 100%가 됩니다.
- 따라서 차원을 2차원으로 줄인 경우 (1 - 0.6949227) 이 되어서 0.3050773이 되며, 이것을 %로 표현하면 약 30.51% 손실이 됩니다.
- X차원으로 줄였을 때 손실율은 (1 - X차원의 Cumulative Proportion) 입니다.

주성분 개수 선택 - Elbow 기법



Scree Plot에서 최적의 요소 수를 찾으라는 문제가 나오면 팔꿈치 부분을 찾아야 합니다.
경사가 완만해지기 시작하는 부분입니다.

t.test

```
> t.test(x=Default$income, mu=33000)

        One Sample t-test

data:  Default$income
t = 3.8764, df = 9999, p-value = 0.0001067
alternative hypothesis: true mean is not equal to 33000
95 percent confidence interval:
 33255.56 33778.41
sample estimates:
mean of x
 33516.98
```

p-value

- 귀무가설 : income의 평균이 33000과 같다
- 대립가설 : income의 평균이 33000과 같지 않다
- df = 9999, n = df + 1 = 10000
- 95% 신뢰구간 : 33255.56 ~ 33778.41
- p-value : 0.05보다 작으므로 귀무가설 기각, 대립가설 채택
- x의 평균 (점추정 값) : 33516.98

t.test

```
> t.test(x=chickwts$weight, mu=260)

        One Sample t-test

data:  chickwts$weight
t = 0.14137, df = 70, p-value = 0.888
alternative hypothesis: true mean is not equal to 260
95 percent confidence interval:
 242.8301 279.7896
sample estimates:
mean of x
 261.3099
```

p-value

- 귀무가설 : weight의 평균이 260과 같다
- 대립가설 : weight의 평균이 260과 같지 않다
- df = 70 (degree of freedom), n = df + 1 = 71 (관측치 개수)
- 95% 신뢰구간 : 242.8301 ~ 279.7896
- p-value : 0.05보다 큰 값으로 귀무가설을 채택, 대립가설을 기각함
- x의 평균(점추정 값) : 261.3099

다중 선형 회귀

```
> temp <- lm(Fertility~., data=swiss)
> summary(temp)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

데이터셋 : swiss, 종속변수 : Fertility

y절편

독립변수

- 각 변수의 회귀계수의 p-value
- 유의수준 95%에서 0.05 보다 작을 때 유의미
- Examination만 무의미

- df = 41
- n = df + 6(변수개수) = 47

결정계수

수정결정계수

F 통계량의 p-value, 결과의 유의미

다중 선형 회귀

```
> summary(Wage)
      year           age                     maritl           race           education     
 Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480   1. < HS Grad      :268  
 1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293   2. HS Grad        :971  
 Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190   3. Some College   :650  
 Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37   4. College Grad   :685  
 3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55                   5. Advanced Degree:426  
 Max.   :2009   Max.   :80.00                                                                   
```

education은 범주형 변수

다중 선형 회귀

```
> result <- lm(wage ~ education, Wage)
> summary(result)

Call:
lm(formula = wage ~ education, data = Wage)

Residuals:
    Min      1Q  Median      3Q     Max
-112.31  -19.94   -3.09   15.33  222.56

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 84.104      2.231  37.695  < 2e-16 ***
education2. HS Grad         11.679      2.520   4.634 3.74e-06 ***
education3. Some College    23.651      2.652   8.920  < 2e-16 ***
education4. College Grad    40.323      2.632  15.322  < 2e-16 ***
education5. Advanced Degree 66.813      2.848  23.462  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.53 on 2995 degrees of freedom
Multiple R-squared:  0.2348,    Adjusted R-squared:  0.2338
F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

데이터셋 : Wage, 종속변수 : wage, 독립변수 : education

회귀계수 : 종속변수 wage와의 관계를 나타내는 값

y절편

독립변수

회귀식의 모든 변수가 통계적으로 유의미

education에 대한 더미 변수
더미 변수 개수 = 범주개수 - 1

n = 2995 + 5

다중 선형 회귀

```
> summary(lm(Salary ~., data=Hitters))

Call:
lm(formula = Salary ~ ., data = Hitters

Residuals:
    Min      1Q  Median      3Q     Max
-907.62 -178.35  -31.11  139.09 1877.04
```

Division이 범주형 변수이기 때문에
더미변수(dummy)로 만들어져 사용됨
DivisionW 일 때 1, DivisionE 일 때 0

- DivisionW의 Estimate가 음수이기 때문에 E인 선수에 비해 W인 선수가 평균적으로 Salary가 낮게 됨
- 만일, DivisionW가 양수였다면 E인 선수에 비해 평균적으로 Salary가 높게 됨

- lm을 이용해서 선형 회귀 분석을 했는데,
  Multiple R-squared 가 0.5461로 점수가 매우 낮음
- 따라서 선형인지 아닌지 알 수 없음

범주형

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  163.10359   90.77854   1.797 0.073622 .
AtBat         -1.97987    0.63398  -3.123 0.002008 **
Hits           7.50077    2.37753   3.155 0.001808 **
HmRun          4.33088    6.20145   0.698 0.485616
Runs          -2.37621    2.98076  -0.797 0.426122
RBI           -1.04496    2.60088  -0.402 0.688204
Walks          6.23129    1.82850   3.408 0.000766 ***
Years         -3.48905   12.41219  -0.281 0.778874
CAtBat        -0.17134    0.13524  -1.267 0.206380
CHits          0.13399    0.67455   0.199 0.842713
CHmRun        -0.17286    1.61724  -0.107 0.914967
CRuns          1.45430    0.75046   1.938 0.053795 .
CRBI           0.80771    0.69262   1.166 0.244691
CWalks        -0.81157    0.32808  -2.474 0.014057 *
LeagueN       62.59942   79.26140   0.790 0.430424
DivisionW   -116.84925   40.36695  -2.895 0.004141 **
PutOuts        0.28189    0.07744   3.640 0.000333 ***
Assists        0.37107    0.22120   1.678 0.094723 .
Errors        -3.36076    4.39163  -0.765 0.444857
NewLeagueN   -24.76233   79.00263  -0.313 0.754218
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 315.6 on 243 degrees of freedom
  (결측으로 인하여 59개의 관측치가 삭제되었습니다.)
Multiple R-squared:  0.5461,    Adjusted R-squared:  0.5106
F-statistic: 15.39 on 19 and 243 DF,  p-value: < 2.2e-16
```

다중 선형 회귀 - 변수 선택

```
> model <- lm(Salary~., data=Hitters)
> step(model, direction='backward')
Start:  AIC=3046.02
Salary ~ AtBat + Hits + HmRun + Runs + RBI + Walks + Years +
    CAtBat + CHits + CHmRun + CRuns + CRBI + CWalks + League +
    Division + PutOuts + Assists + Errors + NewLeague

             Df Sum of Sq      RSS    AIC
 - CHmRun     1      1138 24201837 3044.0
 - CHits      1      3930 24204629 3044.1
 - Years      1      7869 24208569 3044.1
 - NewLeague  1      9784 24210484 3044.1
 - RBI        1     16076 24216776 3044.2
 - HmRun      1     48572 24249272 3044.6
 - Errors     1     58324 24259023 3044.7
 - League     1     62121 24262821 3044.7
 - Runs       1     63291 24263990 3044.7
 - CRBI       1    135439 24336138 3045.5
 - CAtBat     1    159864 24360564 3045.8
 <none>                   24200700 3046.0
 - Assists    1    280263 24480963 3047.1
 - CRuns      1    374007 24574707 3048.1
 - CWalks     1    609408 24810108 3050.6
 - Division   1    834491 25035190 3052.9
 - AtBat      1    971288 25171987 3054.4
 - Hits       1    991242 25191941 3054.6
 - Walks      1   1156606 25357305 3056.3
 - PutOuts    1   1319628 25520328 3058.0
```

- direction='backward' 이므로 후진 제거법
- 후진 제거법은 모든 설명변수가 포함된 모형에서 시작
- 한 번 제거된 변수는 다시 모형에 포함될 수 없음
- 변수 선택에 있어 AIC가 작을 수록 좋은 평가이므로, AIC가 작아지게 되는 변수를 제거하게 된다

로지스틱 회귀

```
> model = glm(default ~ ., data=Default, family=binomial)
> summary(model)

Call:
glm(formula = default ~ ., family = binomial, data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

로지스틱 회귀 (분류)

종속변수 : default, 2항분류

범주형

n = 10000

student가 범주형 변수이기 때문에 더미변수(dummy)로 만들어져 사용됨 studentYes 일 때 1, studentNo 일 때 0

student 값이 Yes 일 때, default를 감소 시킴

```
default = -1.087e+01 -6.468e-01*studentYes
                     + 5.737e-03*balance
                     + 3.033e-06*income
```

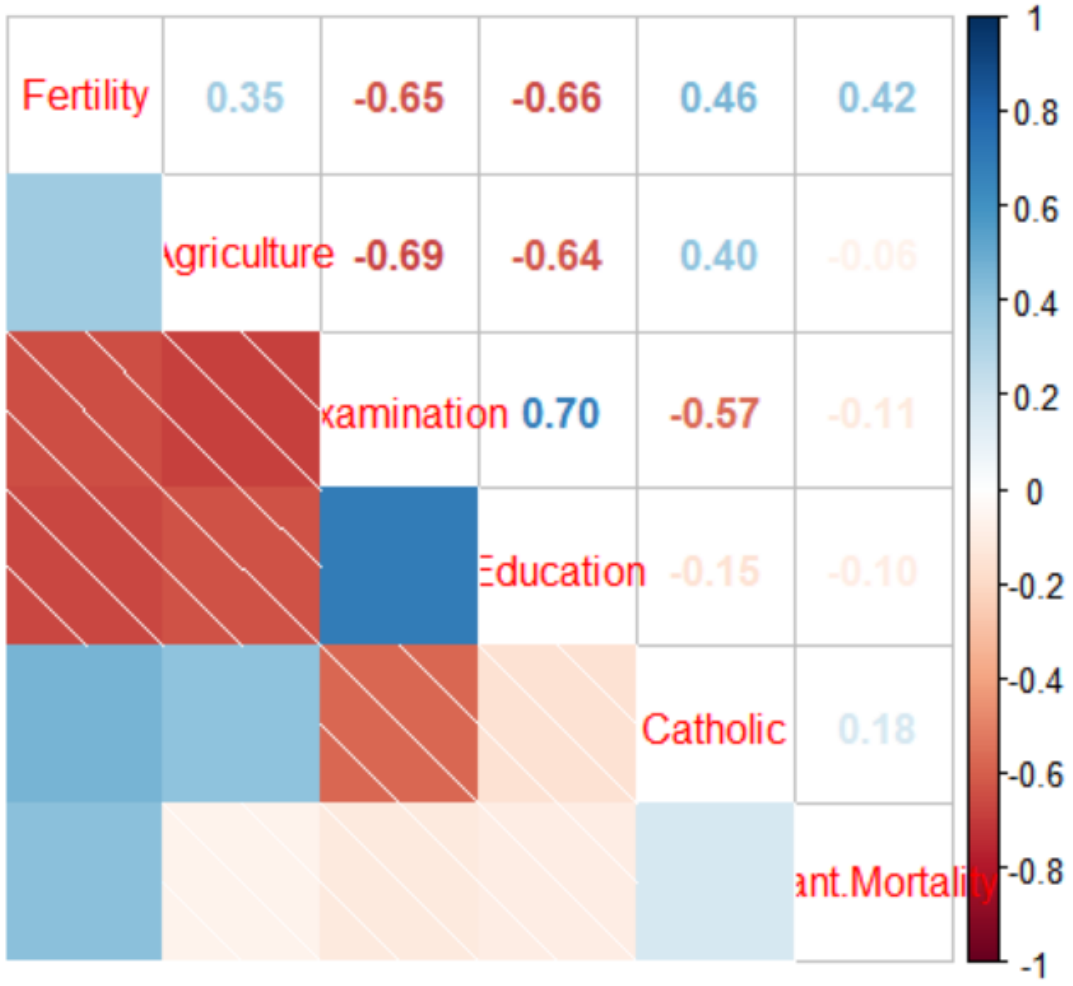
student값이 No 일 때, default를 변화시키지 않음

```
default = -1.087e+01
                     + 5.737e-03 * balance
                     + 3.033e-06*income
```

상관 계수

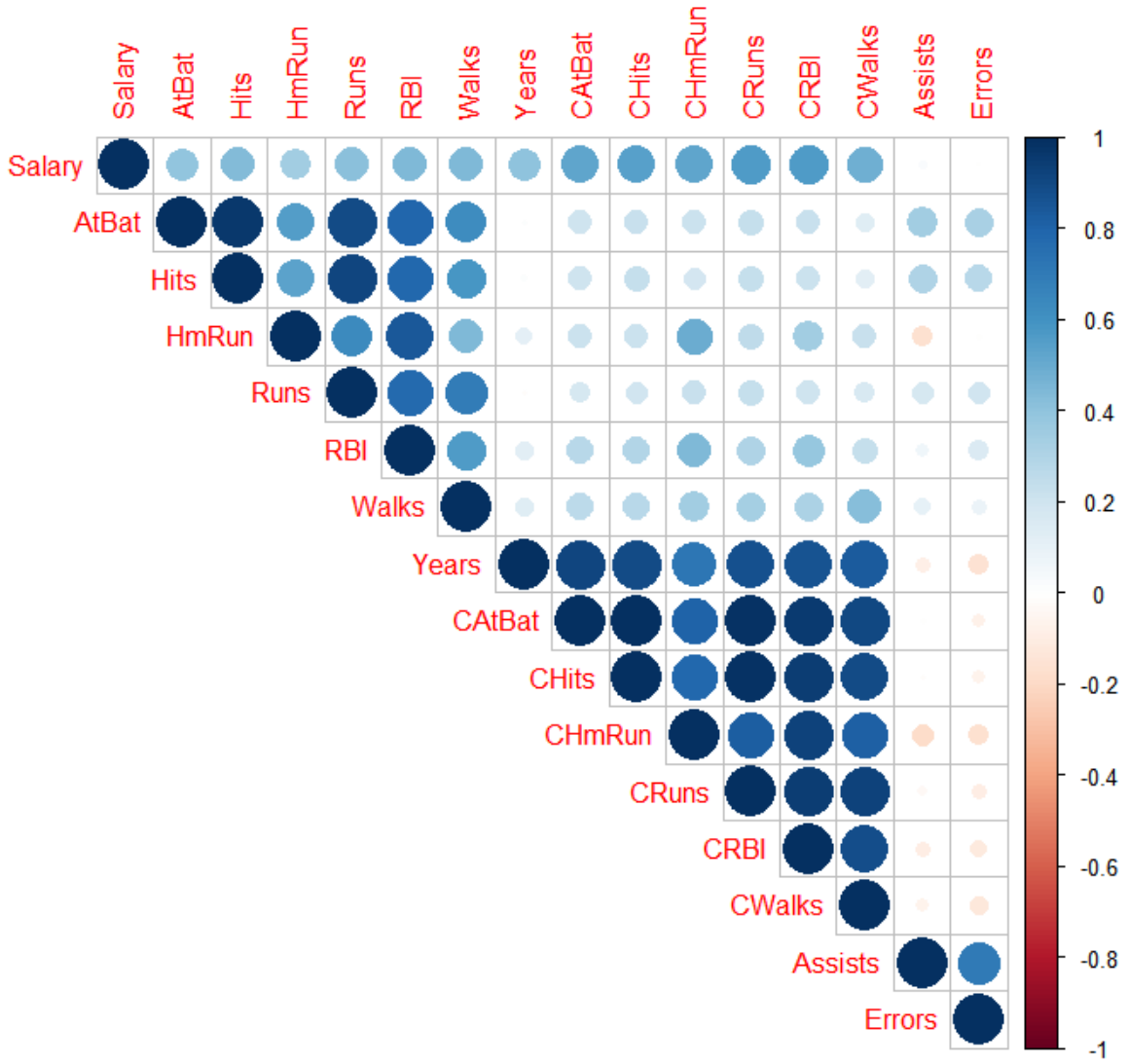
- **상관 계수** 값의 범위는 –1부터 +1까지입니다.
- **계수**의 절대값이 클수록 변수 사이에 강한 **관계**가 있습니다.
- **피어슨 상관**의 경우 절대값 1은 완전한 **선형 관계**를 나타냅니다.
- 0에 가까운 **상관** 값은 변수 사이에 **선형의 상관 관계**가 없음을 나타냅니다.
- 상관계수가 0이라는 것은 아무런 관계가 없다는 것이 아니라 , **'선형의 상관 관계가 아니다**'선로 해석

상관계수 그래프 (corrplot)



- 시험(Examination)과 가장 상관관계가 높은 변수는 교육수준(Education)이다
- 교육수준(Education)이 높을수록 출산율(Fertility)은 낮아진다
- 출산율(Fertility)과 농업인구 비율(Agriculture)은 선형관계를 보인다
- 틀림 : 출산율(Fertility)은 시험(Examination)과 가장 높은 음의 상관관계를 가진다

상관계수 그래프 (corrplot)



동그라미의 크기가 크고, 짙은 색상일 수록 높은 상관관계

- Salary와의 상관계수가 작은 변수 중 하나는 Errors이다
- Salary와 Errors의 산점도에서는 선형성이 나타나지 않을 것이다.
- Salary를 종속변수로 나머지 변수들을 독립변수로 하는 회귀모형을 적합할 때 다중공선성이 존재할 가능성이 크다
- 틀림:Salary와 CRuns의 상관계수는 통계적으로 유의하다

Salary, Errors 의 상관계수를 보면 거의 흰색으로 보이지 않습니다. 즉, 0에 가깝다는 것이며, 이런 경우 선형성이 없다고 판단할 수 있습니다.

여러 변수들이 진하고 검은 동그라미로 색칠된 관계 (-1 또는 1) 인 것을 볼 수 있습니다.
이런 경우 다중 공선성이 존재한다고 할 수 있습니다.

카이제곱 독립성 검정

```
> table(Default$default, Default$student)

default    No  Yes  student
    No   6850 2817
    Yes   206  127
> chisq.test(Default$default, Default$student)

        Pearson's Chi-squared test with Yates' continuity correction

data:  Default$default and Default$student
X-squared = 12.117, df = 1, p-value = 0.0004997
```

default(연체)에 대한 학생, 비학생의 독립성 검정

p-value가 0.05보다 작으므로 귀무가설 기각 대립가설 채택
➔ 연체와 학생은 서로 독립이 아니다.
➔ 연체와 학생의 차이가 5% 유의 수준에서 존재한다!

카이제곱 검정 (Chi square test)

- 범주형 자료로 구성된 데이터 분석에 사용
- 관찰된 빈도가 기대되는 빈도와 유의한 차이가 있는지 검증
- 카이제곱값(χ^2) = $\sum$(관측값 − 기댓값)^2/기댓값

카이제곱 검정 중 독립성 검정

- Contingency table에 있는 두 개 이상의 변수가 서로 독립인지 검정

- 귀무가설 : 두 변수는 차이가 없음 (독립O, 관계X)
- 대립가설 : 두 변수는 차이가 있음 (독립X, 관계O)

귀무가설 : 연체와 학생은 서로 독립이다.
대립가설 : 연체와 학생은 서로 독립이 아니다.

회귀모형의 anova

Cars 데이터에서 속도(speed)와 제동거리(dist)의 관계를 회귀모형으로 추정한 것이다.
(회귀모형의 유의성 분석)

```
> out <- lm(dist~speed, data=cars)
> anova(out)
Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value

MSE = 오차 분산의 불편추정량

- 회귀계수는 5% 수준에서 유의하다
- 관측치는 48 + 2 이다 (Residuals df + 변수 2개)
- 결정계수 = SSR/SST= ((21186) / (21186 +11354)) = 0.651
- 오차 분산의 불편추정량은 'MSE' 오차제곱평균으로, Mean Sq 와 Residuals가 교차되는 지점에 235.6 이라고 써 있음

의사결정나무 - 불순도 측정

- 지니 지수, 엔트로피 지수 :
  - '자체가 불순도를 의미'함
  - 따라서 작은 값을 선택해야 함, 불순도가 작다는 것이 순수도가 높아지며
  - '순수도가 높은 것을 분류를 잘한 것'으로 봄
  - 엔트로피 지수는 p=0.5 (확률이 0.5)일 때 가장 불순도가 높은 것임

- 카이제곱통계량의 p-value :
  - 카이제곱통계량을 구할 때 '동질성 검사'를 사용함, 즉 동질성 검사의 결과임
  - 동질성 검사의 가설
    귀무 : A의 종류에 따라 B가 같다
    대립 : A의 종류에 따라 B가 다르다 - '대립가설이 채택' 되어야 함
    (대립가설이 채택되면 : 어떤 독립변수를 사용해서, 종속변수의 구분이 잘 된다는 뜻)
  - Chi-Square 값이 클 수록 p-value가 작음의 의미
  - p값이 작을수록 자식노드 내의 이질성이 큼을 나타냄 ('카이제곱 이용시, 이질성이 커야 좋음')

의사결정나무 - 불순도 측정

의사결정나무에서 [A], [B], [C] 노드가 있고 [A노드 Good:50, Bad:50], [B노드 Good:10, Bad:40], [C노드 Good:20, Bad:30] 일 때, B노드의 지니 지수를 구하시오.

- B노드의 지니지수를 구하라고 했으니 [B노드 Good:10, Bad:40] 으로 구하시면 됩니다.
- 지니지수 = $1 - \sum P(i)2 = 1 - ((1/5)^2 + (4/5)^2) = 1 - (1/25 + 16/25) = 1-17/25 = 8/25 = 0.32$

- Good 과 Bad의 합이 50 이고, Good이 10, Bad 가 40 이어서
  Good은 10/50 = 1/5 로 보고 Bad 는 40/50 = 4/5 로 한 것입니다.

정밀도(Precision)

오분류표에서 실제/예측 True와 실제/예측 False가 100으로 동일하다고 한다.
민감도가 0.8이라고 할 때, 정밀도(Precision)은 얼마인가?

민감도(Sensitivity) = 실제 True인 것 중에 예측도 True로 한 것

$= \frac{TP}{TP+FN} = 0.8$ $\frac{10*(TP+FN)*TP}{TP+FN} = \frac{8}{10} * (TP+FN) * 10$, $10TP = 8TP + 8FN$ ➔ TP = 4FN

민감도가 0.8이면 TP=4FN이다.
실제/예측 True와 실제/예측 False가 100으로 통일이라면 아래와 같이 된다. TP+FP=100, FN+TN=100, TP+FN=100, FP+TN=100 이 중 TP+FN=100에서 FN=20, TP=80이다. 그러면 FP=20, TN=80이므로 Precision(예측이 True인 것 중 실제도 True인 것) 을 구하면 TP/(TP+FP)=80/100=0.8이다.

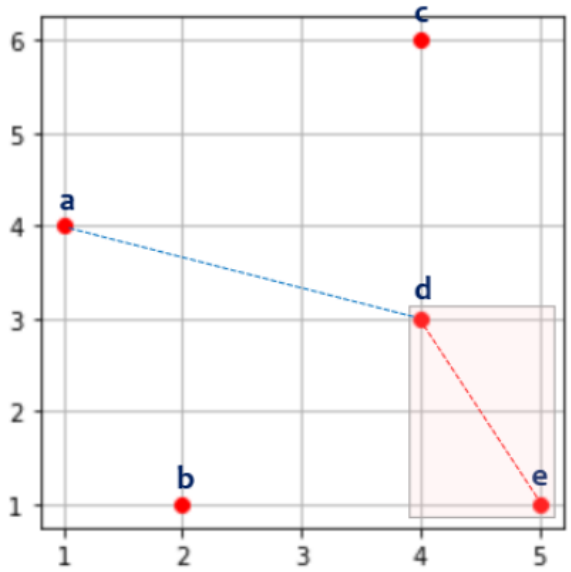
실\예	True	False
True	TP	FN
False	FP	TN

계층적 군집 - 최단연결법

데이터셋 x는 두 개의 변수와 5개의 관측치를 가지며 아래는 데이터와 관측치 간의 유클리드 거리를 나타낸다. 최단연결법을 사용하여 계층적 군집화를 할 때 첫 단계에서 형성되는 군집과 관측치 a와의 거리를 구하시오

```
> x
  x1 x2
a  1  4
b  2  1
c  4  6
d  4  3
e  5  1
```

```
> round(dist(x), 1)
    a   b   c   d
b 3.2
c 3.6 5.4
d 3.2 2.8 3.0
e 5.0 3.0 5.1 2.2
```

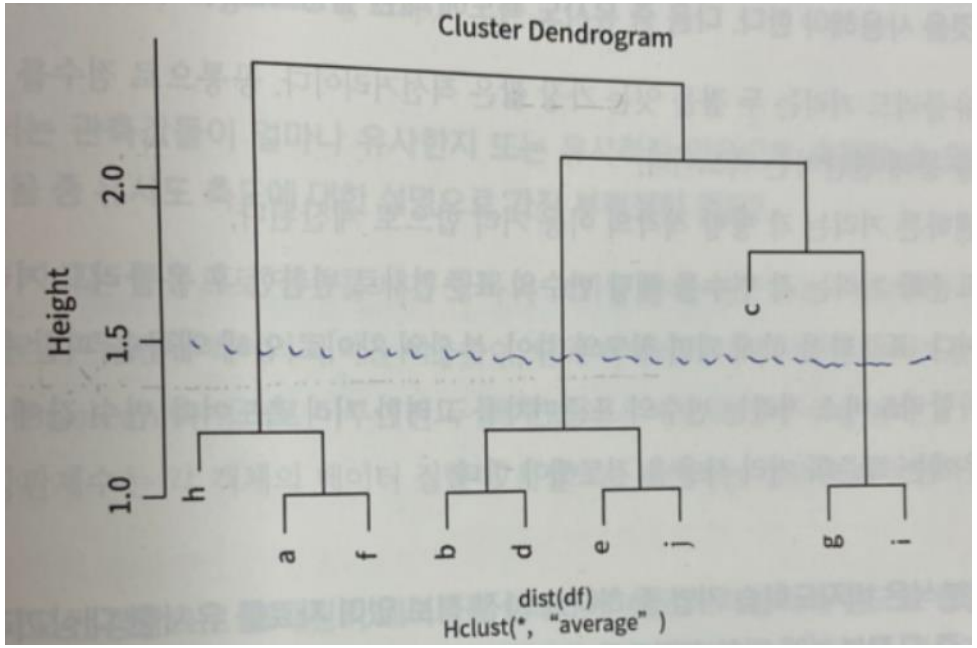


1. 첫 단계에서 형성되는 군집을 dist(x)의 결과에서 찾습니다.
   ➔ 값이 가장 작은 것 {d, e} : 2.0
2. 문제에서 '최단연결법'이라고 했기 때문에 {d, e} 군집과 a의 거리는 d 와 e 중에서 a와 더 가까운 것과의 거리가 됩니다.
   ➔ dist(x)의 결과를 보면 a와 d의 유클리드 거리가 3.2 라고 표기 되어 있습니다.

답 : 3.2

계층적 군집 - 덴드로그램

아래 그림은 평균연결법을 통한 계층적 군집화 예제이다. 데이터 분석 목적 상 Height값을 1.5를 기준으로 하위 군집을 구성할 때 다음 중 생성된 하위 군집을 가장 잘 나타낸 것은?

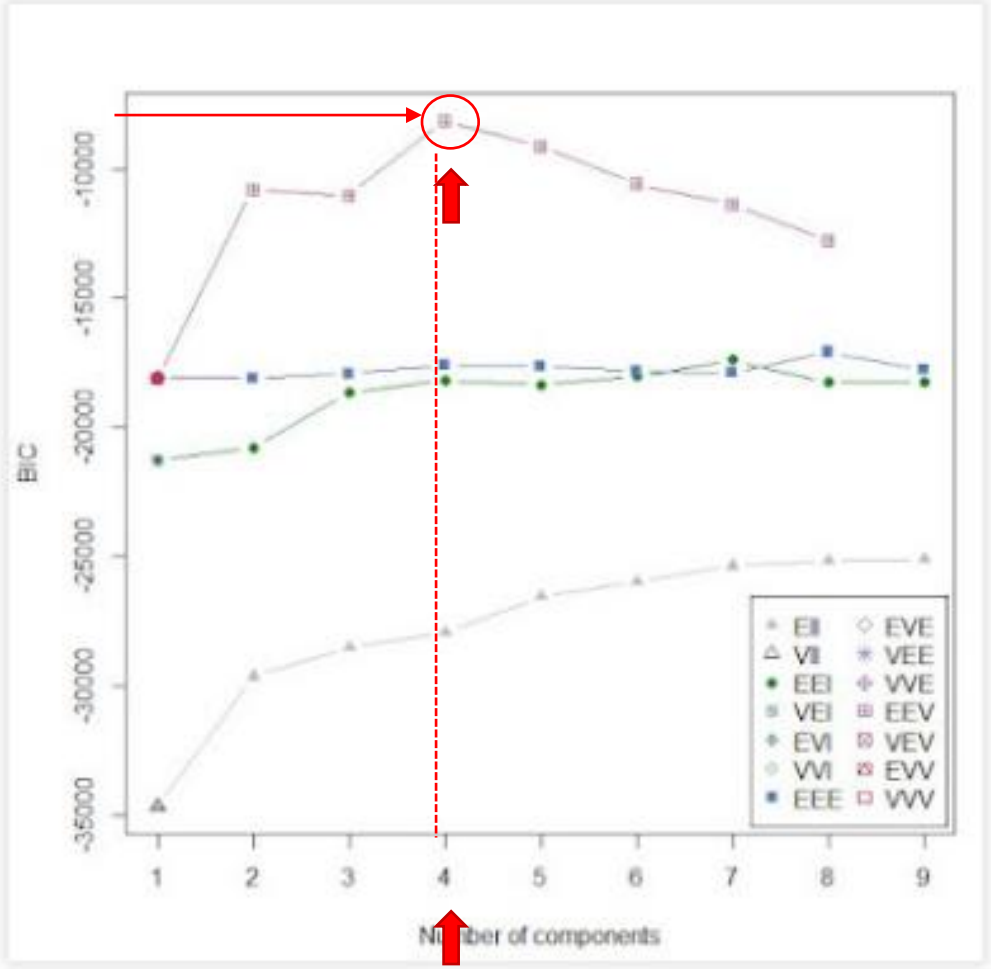


- 덴드로그램에서 c 쪽이 위에 걸려 있지만 아래로 선을 그려서 생각해 주셔야 합니다.
- 즉, Height 1.5의 경우 4개의 군집이 되는 것입니다.
- 그래프가 그려질 때 더 이상 분리 되지 않는 경우 선이 아래까지 그려지지 않아서 그런 것입니다.
- 문제 오류 아님

{c} 가 1.5 선 위에 있지만 그렇다고 군집이 아닌 것은 아니다 (= 군집이다!)

답: {h,a,f}, {b,d,e,j}, {c}, {g,i}

군집분석 결과



최적의 군집수는?

BIC 값이 가장 큰 것의 x축 값을 읽어 주시면 됩니다.
BIC 값이 Y 축에 있으며, 그림에서 위쪽으로 갈 수록 값이 커지는 것을 볼 수 있습니다.

군집, 분류 분석의 구분

분류분석과 군집분석을 어떻게 구분 할 수 있나요??

- 이질적인 모집단을 세분화 ==> 군집, 서로 다른 특성을 갖는 것으로 나누는 것입니다.
- 분류 분석은 '지도 학습'이기 때문에 정답에 맞는 특성을 학습하고 그 구분되는 특성에 따라 class를 나누는 것입니다.
- 군집은 기준이 없고, 분류는 기준이 있습니다.

다음 데이터 마이닝의 대표적인 기능 중 이질적인 모집단을 세분화하는 기능으로 적절한 것은?
➔ 군집분석

에어컨 회사에서 지역별 온도, 습도에 따라 고객군을 나눠서 마케팅 전략을 수립할 때
분석 방법은?
➔ 분류분석

시계열 모형

시계열 모형의 여러 종류 중 아래에서 설명하는 것은 무엇인가?

가) 시계열 모델 중 자기 자신의 과거 값을 사용하여 설명하는 모형임
나) 백색 잡음의 현재 값과 자기 자신의 과거 값의 선형 가중합으로 이루어진 정상 확률 모형
다) 모형에 사용하는 시계열 자료의 시점에 따라 1차, 2차, …, p차 등을 사용하나
정상시계열 모형에서는 주로 1, 2차를 사용함

AR 모델

7.2.1 AR 모델

AR(AutoRegression)(자기 회귀) 모델은 이전 관측 값이 이후 관측 값에 영향을 준다는 아이디어에 대한 모형으로 자기 회귀 모델이라고도 합니다. AR에 대한 수식은 다음과 같습니다.

$$\underset{①}{Z_t} = \underbrace{\Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p}}_{②} + \underset{③}{a_t}$$

①은 시계열 데이터에서 현재 시점을 의미하며, ②는 과거가 현재에 미치는 영향을 나타내는 모수(Φ)에 시계열 데이터의 과거 시점을 곱한 것입니다. 마지막으로 ③은 시계열 분석에서 오차 항을 의미하며 백색 잡음이라고도 합니다. 따라서 수식은 p 시점을 기준으로 그 이전의 데이터에 의해 현재 시점의 데이터가 영향을 받는 모형이라고 할 수 있습니다.

퍼스널 빅데이터

- 퍼스널 빅데이터는 나와 관련된 주변 데이터까지 모두 수집합니다.
- 이동, 구매, 식사 같은 실생활 패턴 외에도 웹이나 소셜 로그 같은 온라인 활동을 포함합니다.
- 혈압, 간 수치, 혈액형 같은 건강검진 기록부터
  끼니별 식사 메뉴, 수면 시간, 일하는 시간, 카드 기록, 소비 습관 같은 생활 데이터
  책이나 TV를 볼 때 눈동자의 시선이 어떻게 움직이는지 같은 생체 데이터 등
  사람에게서 나올 수 있는 모든 데이터를 퍼스널 빅데이터라고 합니다.

기업들이 퍼스널 빅데이터로 가치를 창출하려고 한다. 퍼스널 빅데이터에는 건강정보, 행태정보, 감정정보 등이 있다. 다음 중 행태정보에 속하지 않는 것은?
: 주간별 운동량, 하루 중 통화 빈도, 여름에 판매량이 느는 상품, 연중 정당별 선호도 변화

- 행태정보는 개인 이용자의 활동에 따른 정보라고 할 수 있습니다.
  ➔ 그런데 연중 정당별 선호도 변화는 개인의 활동정보로 보기에 적합하지 않는 내용으로 보입니다.

KDD, CRISP-DM

KDD 분석 방법론에서 잡음, 이상치, 결측치를 식별하여 분석용 데이터셋을 선택하고, 분석에 필요한 변수 등을 선정하는 단계와 유사한 CRISP-DM 방법론의 단계는?

위에서 이상치, 결측치를 식별하고 분석용 데이터셋을 선택하고 분석에 필요한 변수를 선정하는 단계라는 말이 두 가지 단계를 설명하고 있는 것 같습니다.

KDD

- Preprocessing : 데이터셋에 포함되어 있는 잡음(noise)과 이상값(outlier), 결측치(missing value)를 식별하고 필요시 제거하거나 의미 있는 데이터로 처리한다. 좋은 결과를 위해 지저분한 데이터를 깨끗하게 정제하는 단계!
- Transformation : 분석 목적에 맞는 데이터를 선택하거나 데이터 차원을 축소하여 데이터 마이닝을 효율적으로 적용될 수 있도록 데이터셋을 변경한다. 내가 가지고 있는 데이터를 그대로 가져와서 바로 분석할 수는 없다! 분석에 용이하도록 데이터를 예쁘게 다듬는 과정!

- Preprocessing, Transformation 두가지를 모두 설명하고 있어서, Transformation 기능과 관련 있는 데이터 준비로 답을 한 것 같습니다

분산 구하기

100명의 키를 cm으로 측정한 데이터의 분산이 225였다. 동일한 100명의 키를 m로 측정한다면 데이터의 분산은 얼마인가?

예를 들어 cm 데이터가 [100, 200] 이라고 생각하고 분산을 구합니다.
cm와 m의 단위 차이가 100이기 때문에 100 단위로 만들었습니다. 그러면 평균이 150 이기 때문에
분산은 = ((150-100)^2 + (150-200)^2) / 2 = (50^2 + 50^2) / 2 = 2500

이제, 단위를 m로 변경하면 [1, 2] 가 됩니다.
평균이 1.5 이기 때문에 분산은 = ((1.5-1)^2 + (1.5-2)^2) / 2 = (0.5^2 + 0.5^2) / 2 = 0.25

즉, 10000 배의 차이를 보인다는 것을 알 수 있습니다.
분산 식에 제곱이 들어가기 때문이죠. => 100^2, 1^2 => 10000, 1

답 : 225를 10000으로 나눈 값 0.0225 가 됩니다.