**빅데이터 - 14주차 과제 - 2018301034 서 용 주**

**8장 연습문제**

**1번.**

**활용할 데이터셋, 시계열 데이터로 생성**

**# 데이터셋 만들기**

data1 <- c(23.47359, 18.5600, 22.75696, 15.23990, 18.34075, 12.99859, 18.41205, 19.41997, 17.08437, 17.44508,

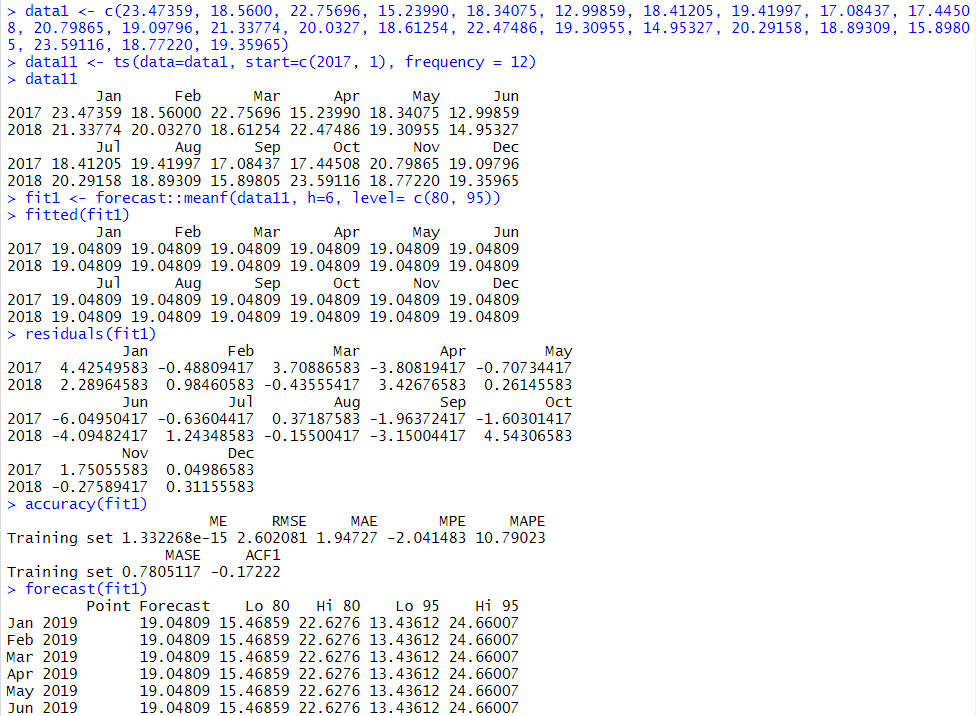
20.79865, 19.09786, 21.33774, 10.0327, 18.61254, 22.47486, 19.30955, 14.95327, 20.29158, 18.89309, 15.89805,

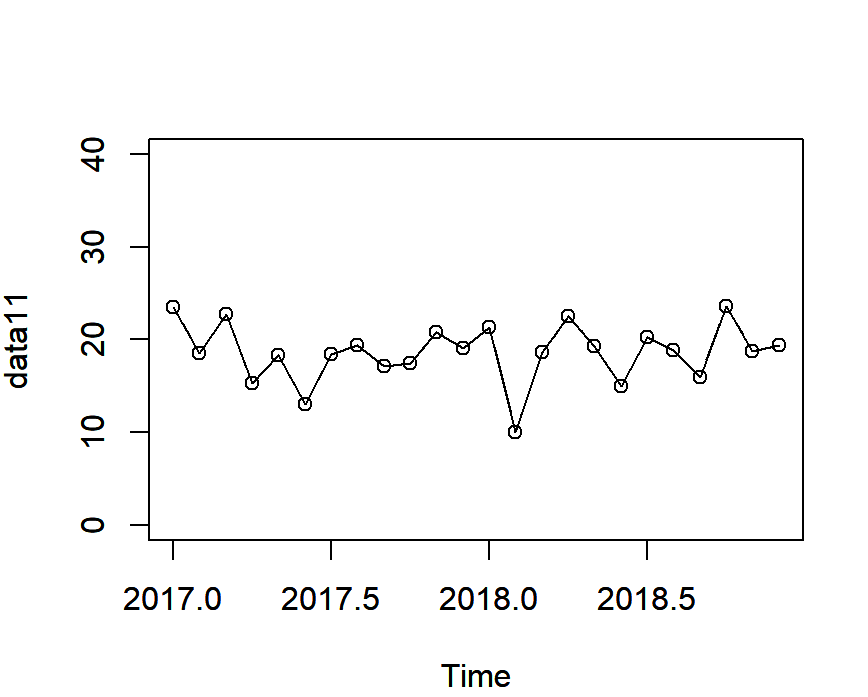
23.59116, 18.777220, 19.35965)

**# 시계열 데이터 형식으로 생성**

data11 <- ts(data= data1, start=c(2017, 1), frequency = 12)

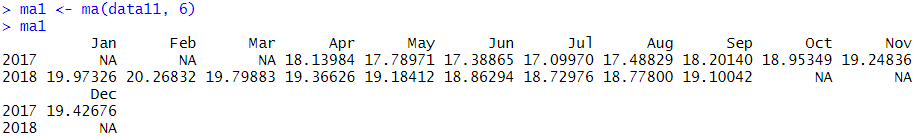
**(1) 전체 데이터 평균으로 예측 : meanf() 함수**



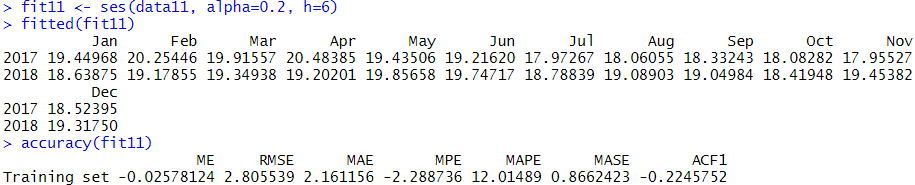


**(2) 이동평균 구하기 : ma() 함수 사용**

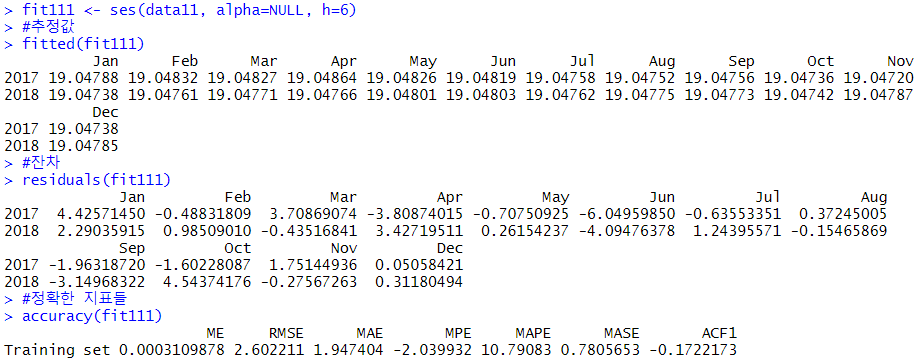
**- ma(data11, 6) 함수의 계산 결과는 6개월 간 이동평균 값이다.**



**(3) 지수평활 : ses() 함수 사용**



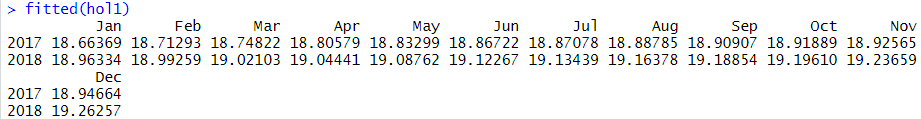
**-> ses() 함수 사용 시, alpha 값 지정안하고 NULL을 주면 자동으로 alpha값을 추정해준다.**



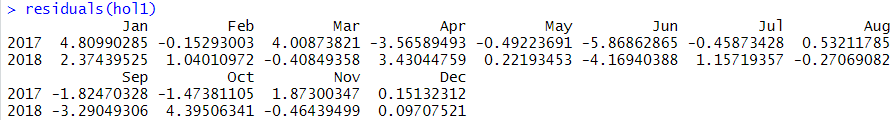
**(4) 추세패턴 예측 : holt() 함수 사용**



**#추정값**



**#잔차**



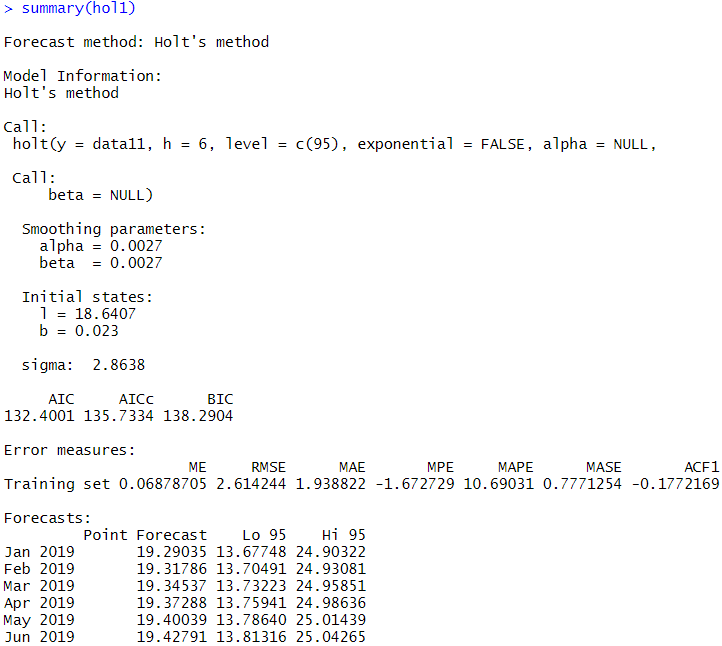
**#정확성 척도**

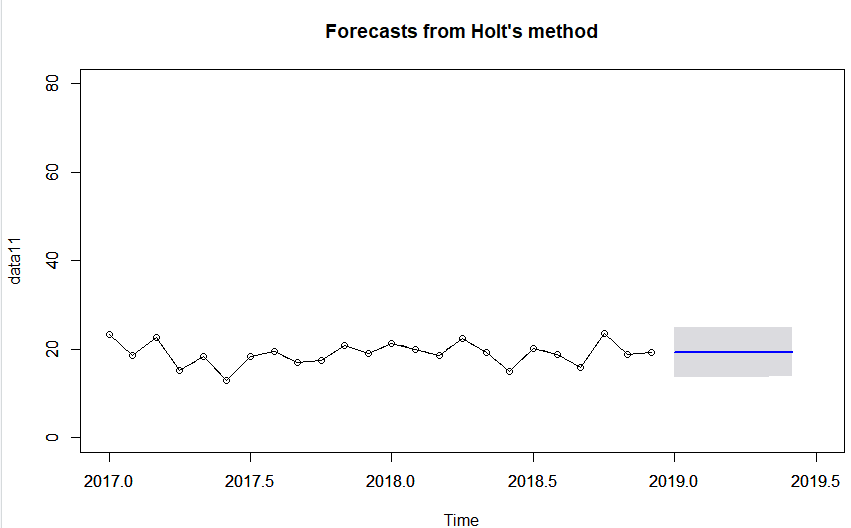


**#그래프**



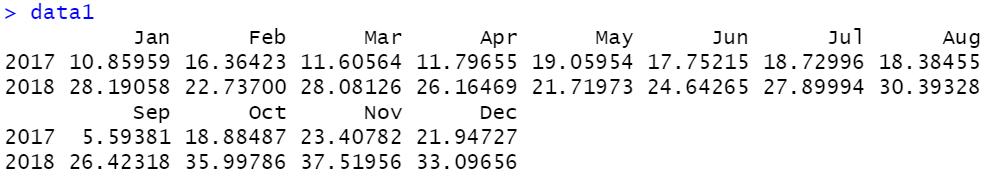
**(5) 계절패턴 예측**

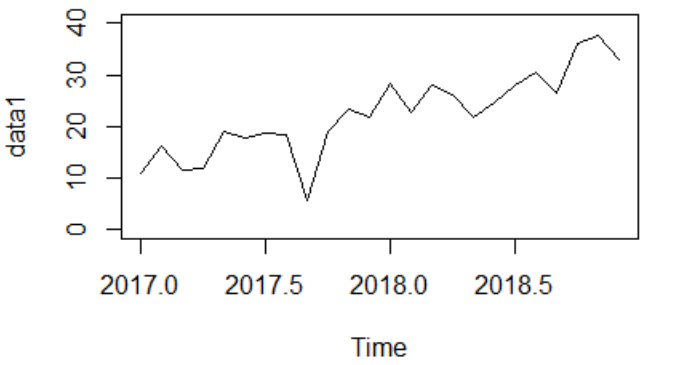




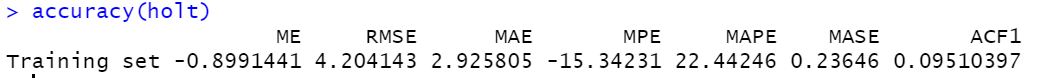
**2번.**

**아래 그래프를 보면 이 데이터들은 '추세패턴 시계열'임을 확인할 수 있다.**

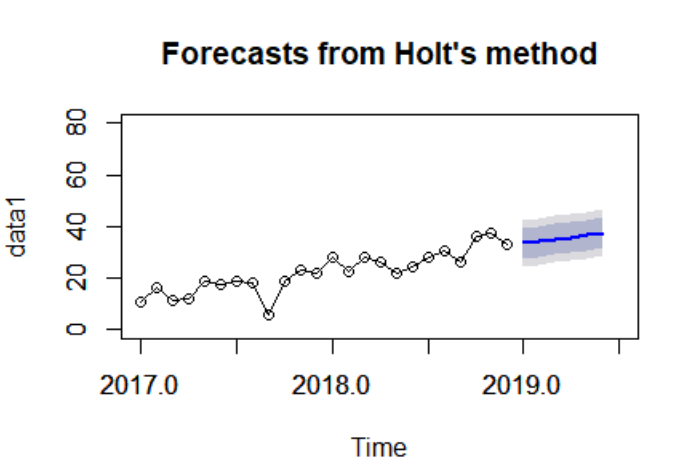




**#정확성 척도**



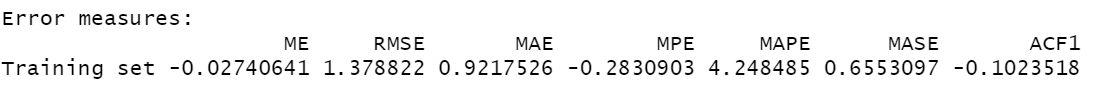
**#예측값과 신뢰구간**



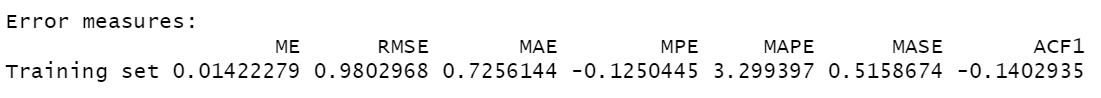
**3번.**

**- 문제에서 주어진 데이터로 그래프를 먼저 그려보면 '계절패턴'이 적합한 데이터임을 확인할 수 있다.**

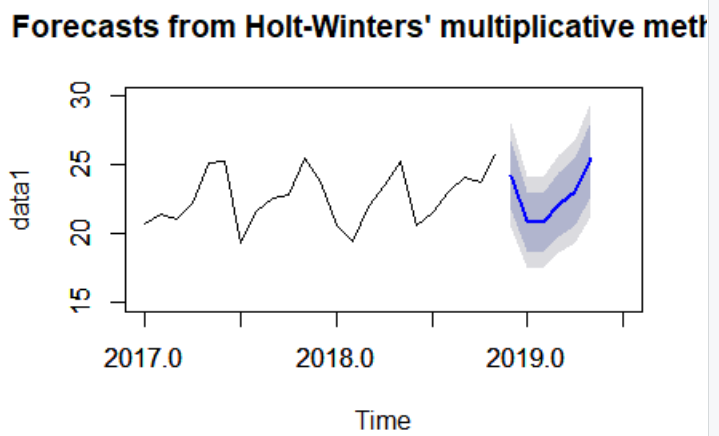
**#가법모형 정확도**

****

**#승법모형 정확도**

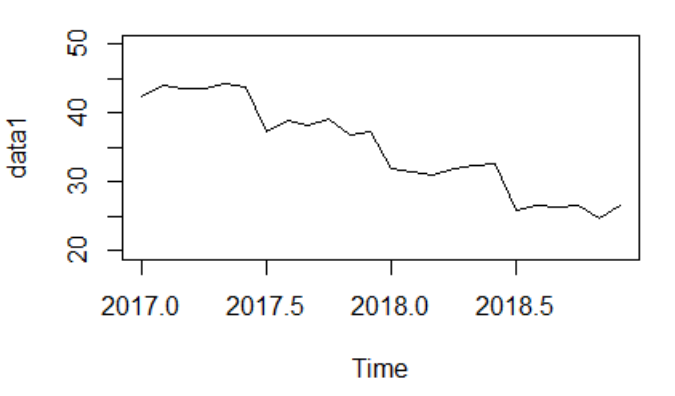
****

**#6개 예측값과 신뢰구간**

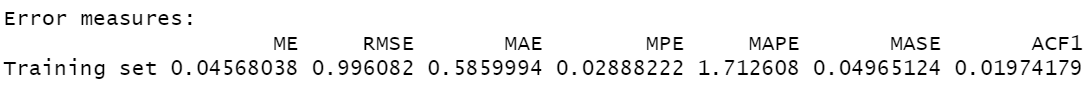


**4번.**

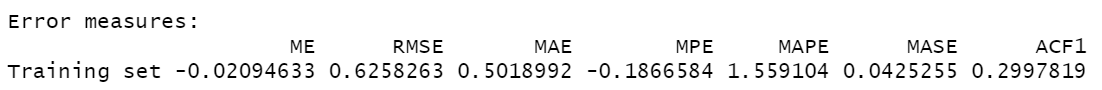
**-아래의 그래프를 확인해보면 문제에서 주어진 데이터들이 '추세와 계절 패턴'임을 알 수 있다.**

****

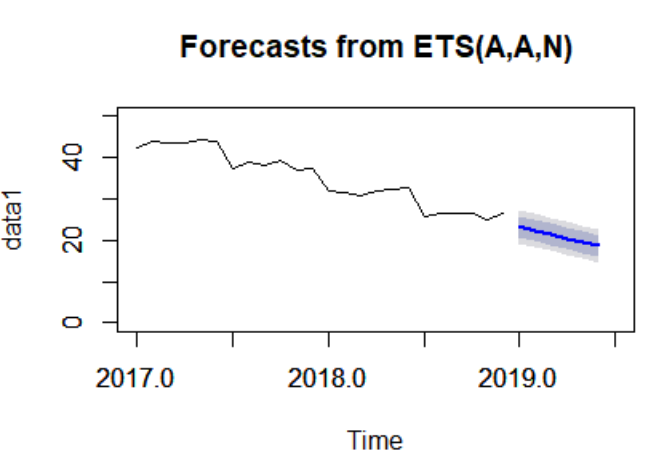
**#가법모형 정확도**

****

**#승법모형 정확도**

****

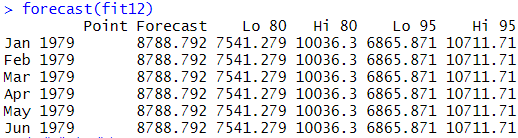
**#6개의 예측값과 신뢰구간**

****

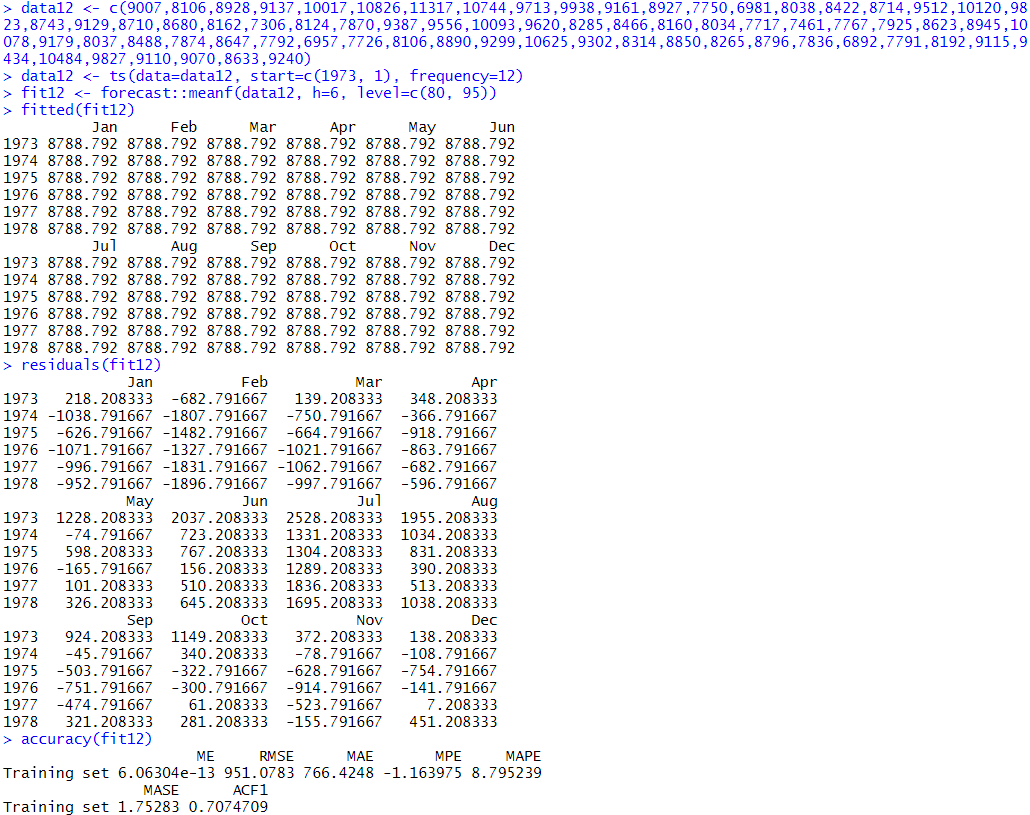
**12번.**

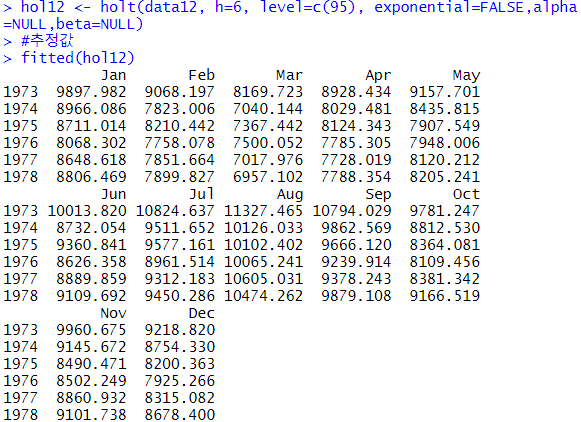
**- USAccDeaths 데이터셋은 타임 시리즈 미국에서 우연한 죽음의 월별 합계를 데이터화한 것이다.**

**# 데이터를 forecast() 함수로 확인해보면 다음과 같다.**

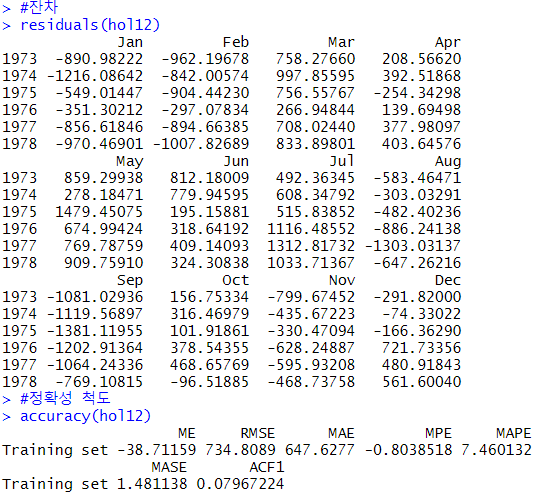
****

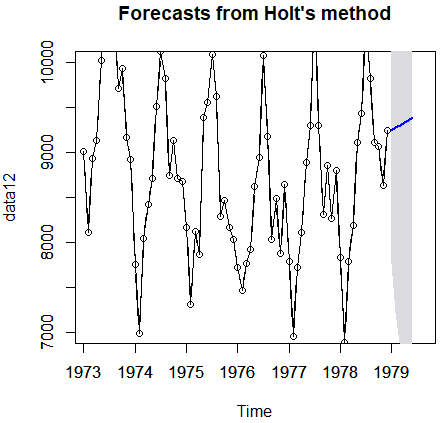
**#holt함수를 이용하여 예측하면 다음과 같다.**





#잔차





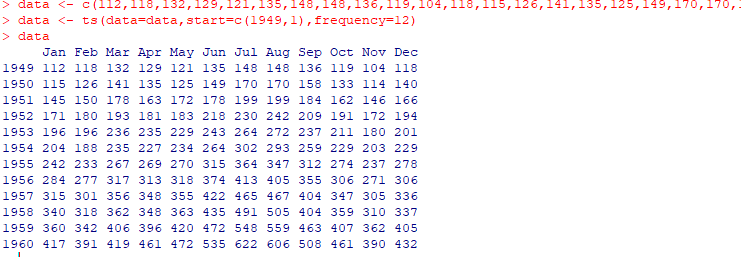
**- 위와 같이 데이터와 추정값, 6개의 예측값, 예측값 신뢰구간들을 그려보았다.**

**추세패턴 예측으로 시계열 패턴을 예측하였으며, 수평패턴의 지수평활을 이용한 방법의 RMSE, MAE, MPE, MAPE값을 비교하여 보았을 때, 수평패탄 예측으로 시계열 패턴을 예측한 값이 더 우수하였음을 보였다.**

**12장 연습문제 풀이**

AirPassengers 데이터셋은 1937년에서 1960년까지 매년 미국의 상업용 항공사가 운항하는 수익 승객 마일리지를 데이터화한 것이다.

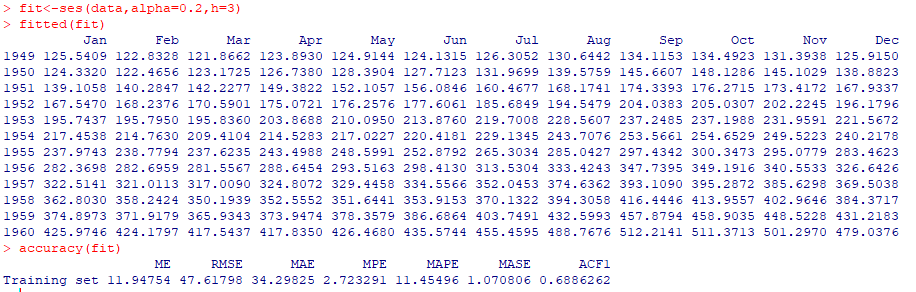
**#데이터 대입**



**1) 지수평활법 사용**

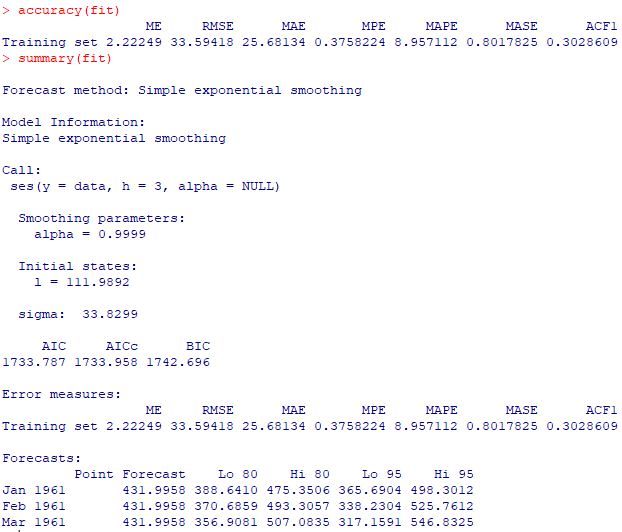
좀 더 수학적인 방법으로 과거예측 값과 전 달의 값의 가중평균으로 예측값으로 사용할 수 있다. 왜냐하면 현재의 데이터는 가장 최근의 데이터가 영향을 많이 주고 과거로 갈수록 영향이 적으므로 가중치를 최근의 데이터에 가장 많이 과거로 갈수록 가중치를 적게 주는 방법을 선택랗 수 있다. 이러한 방법이 지수평활법이다.

**# alpha 값을 0.2로 지정할 경우**



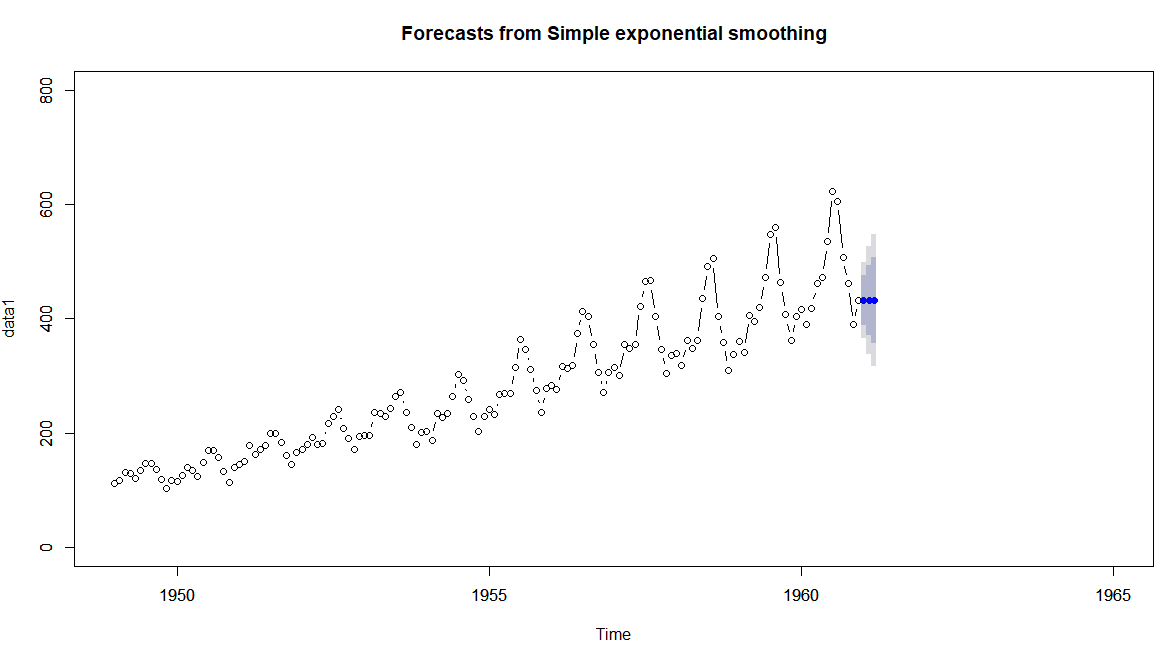
**- 최적의 alpha값을 찾아서 예측할 경우**

**함수: ses()에서 alpha=NULL로 설정하면 을 최적의 값을 찾는다.**



**#예측 그래프**





**# 예측의 정확도 비교**

|  |  |  |  |
| --- | --- | --- | --- |
| 정확도 측정기준 | 지수평활 | HoltWinters() | ets() |
| MAE | 25.68134 | 예측결과에 대한 정확성 측정 어려움 | 12.81203 |
| MAPE | 80.17% | 예측결과에 대한 정확성 측정 어려움 | 39.99% |

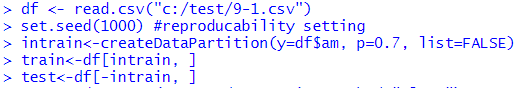
- ets()의 MAPE가 가장 낮기 때문에 예측의 정확도가 제일 높아서 우수하다.

**9장 연습문제**

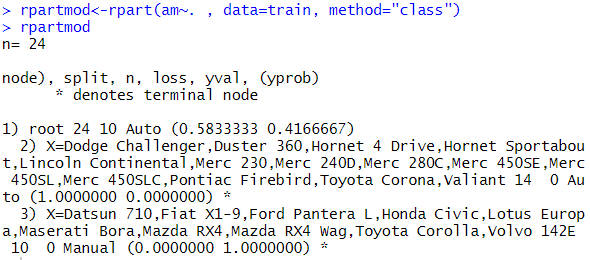
**1번.**

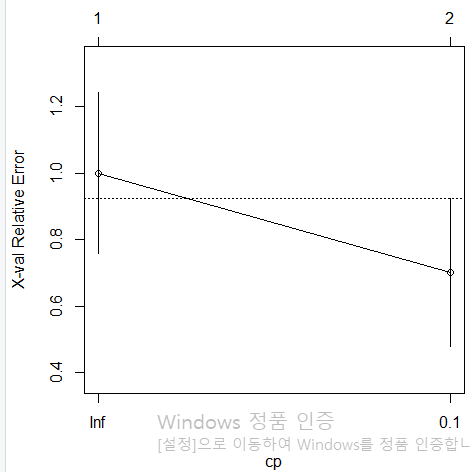
**#rpart 패키지를 이용한 의사결정나무**

**#데이터 삽입**

****

**#학습/실험 데이터 설정**

****

**#그래프로 확인 **

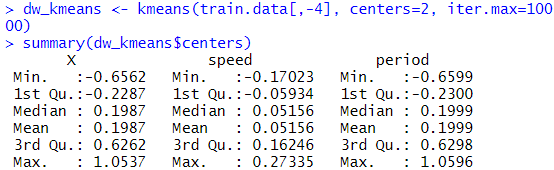
**결과)**

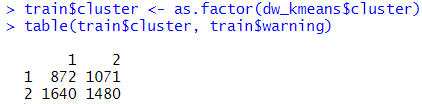
rpart를 이용했을 때의 정확성은 75%

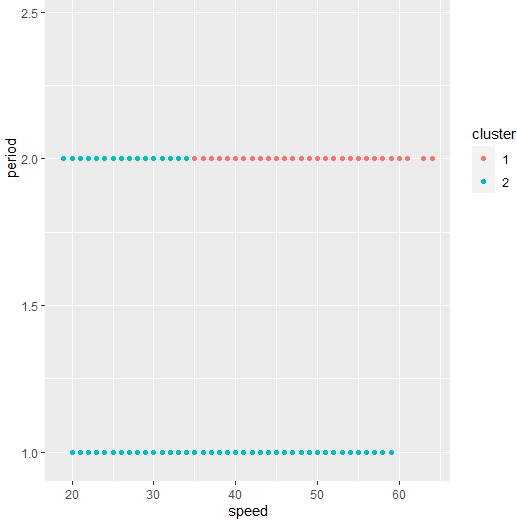
party를 이용했을 때의 정확성은 87.5%이기 때문에 두 모델 다 우수하긴 하지만 party를 이용했을때가 정확성이 훨씬 높기 때문에 pary를 이용한 것이 군집을 가장 잘 구분 할 수 있는 의사결정나무이다.

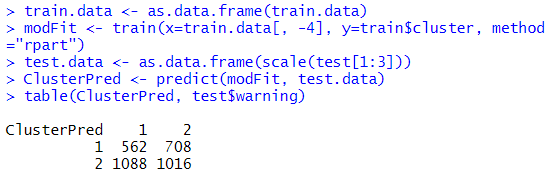
**8번.**

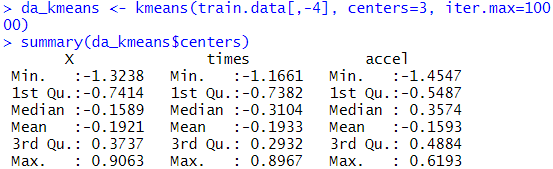
**#학습/실험 데이터 생성**

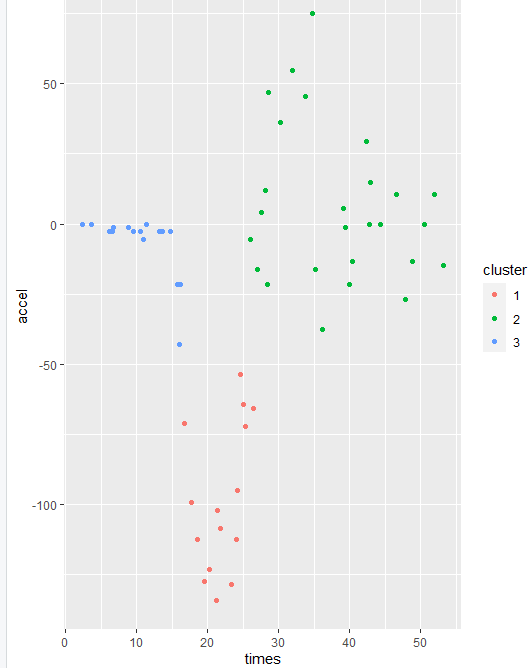
****

****

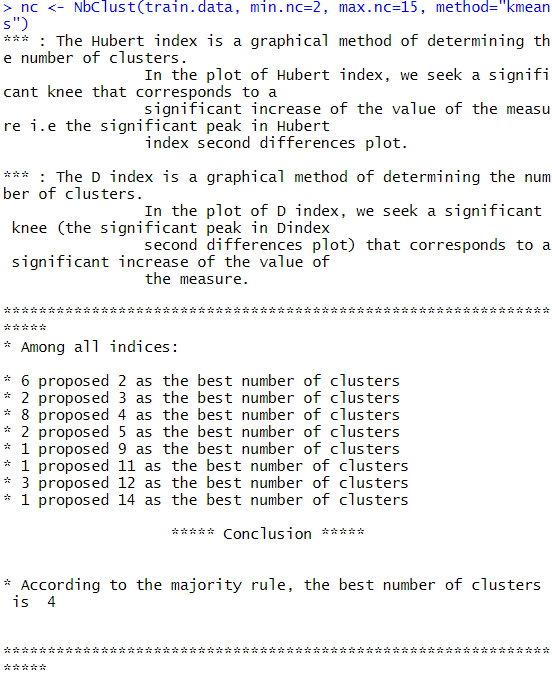
****

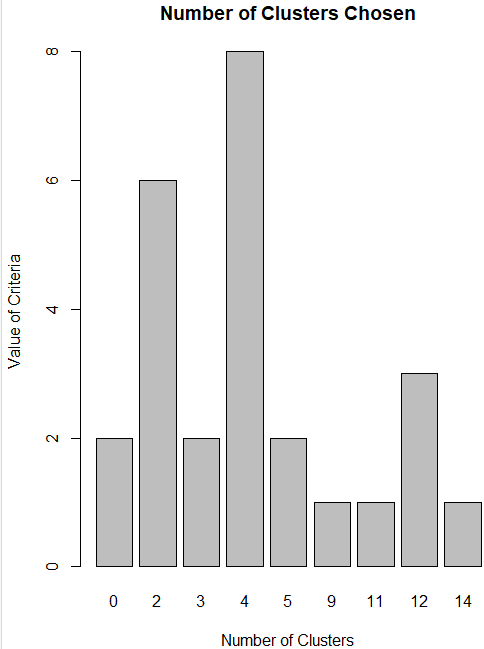
****

****

****

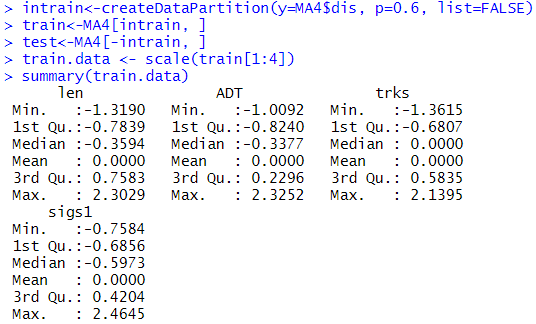
**#학습 모형 생성하여 실험데이터에 적용**

****

****

**#그룹의 수 k 정하기 : NbClust()**

****

****

**9장 실습과제 풀이**

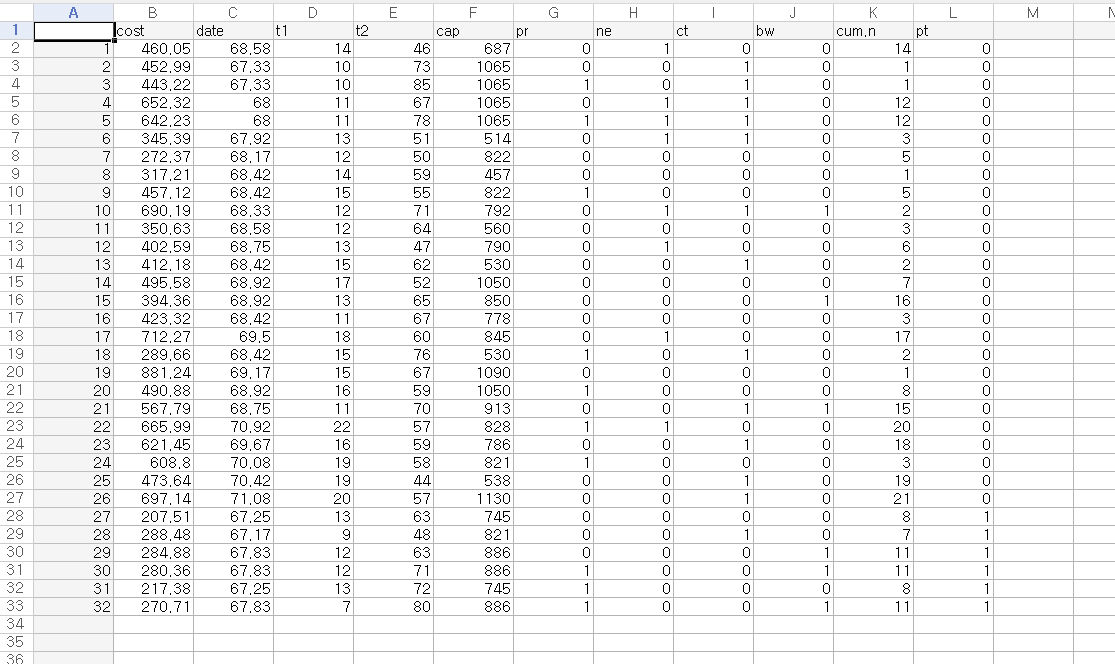
1. **의사결정나무를 적용할 수 있는 데이터 셋을 찾아서 의사결정나무 모델을 구하여라. 그리고 실험 데이터로 결과를 확인해 보아라.**

**① 데이터 삽입: boot패키지의 nuclear 데이터 셋 이용**

**원자력 발전소 건설 데이터이다.**

**ne: 1은 식물이 미국 북동부 지역에 건설되었음을 나타냅니다.**

**ne에 관한 의사결정나무 모델을 구하겠습니다.**



비용: 수백만 달러의 건설 비용은 1976 년 기준으로 조정되었다.

데이터:건설 허가가 발행 된 날짜 데이터는 1990 년 1 월 1 일 이후 가장 가까운 달부터 몇 년 단위로 측정되었다.

t1:건축 허가 신청과 발급 사이의 시간.

t2: 운전 면허 발급과 건축 허가 발행 사이의 시간.

캡: 발전소의 순 용량 (MWe).

홍보: 1이 같은 사이트에 LWR 설비가 존재 함을 나타내는 2 진 변수.

ne: 1이 2 인 변수는 식물이 미국 북동부 지역에 건설되었음을 나타냄.

ct: 공장에서 냉각탑 사용을 나타내는 1의 이진 변수임.

bw: 1이 2 진 변수 인 경우 원자력 증기 공급 시스템이 Babcock-Wilcox에 의해 제조되었음을 나타냄.

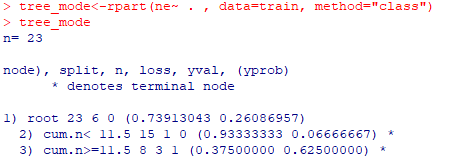
cum.n: 각 설계 엔지니어가 건설 한 누적 발전소 수.

pt: 1은 부분 턴키가 보장되는 시스템을 나타내는 이진 변수.

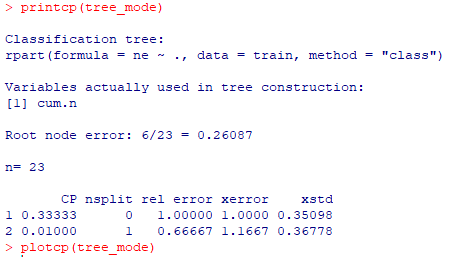
**② 학습(Train)/실험(Test) 데이터 설정: 학습 데이터 70%, 실험데이터 30%**

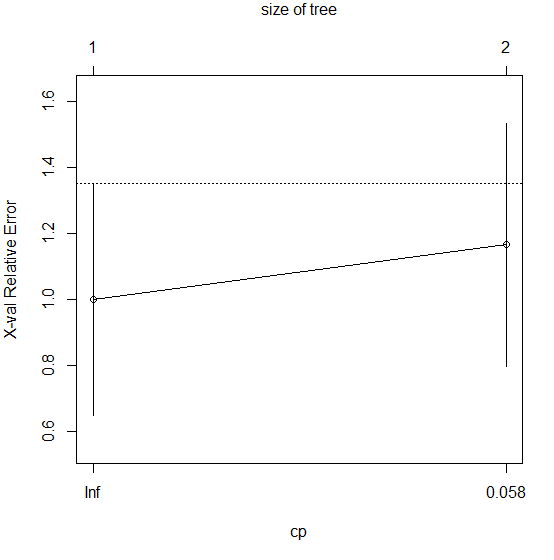


**③ 의사결정나무 모형 설정: rpart( 종속변수 ~. data=데이터셋)**



**- 트리 크기 결정: printcp() cp= complexity parameter 값 계산, plotcp()**



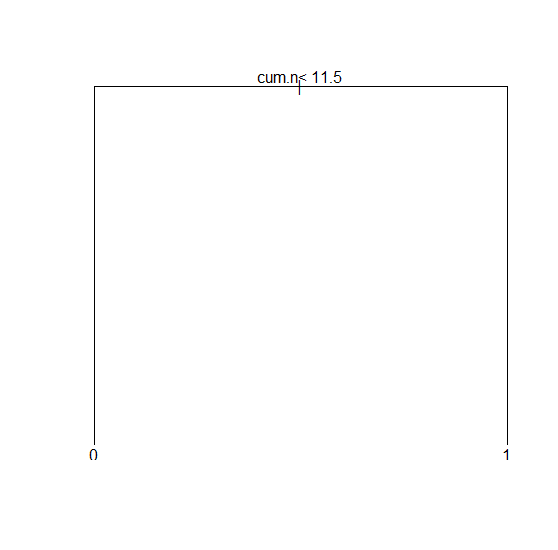


**④ 분류실시: 가지치기를 이용한 분류실시 prune()**

**cp는 xerror가 가장 낮은 split 개수를 선택하면 된다.**

**cp=0.058이다.**

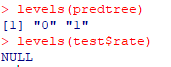




**⑤ 성과분석**



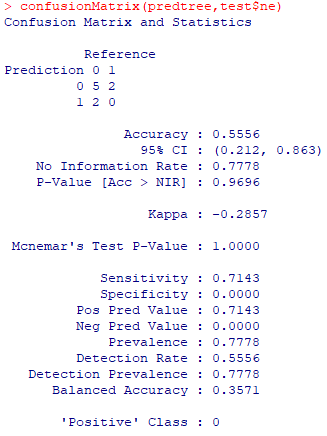
**- level을 같은 형식으로 맞춰주지 않으면 오류가 난다.**







**- 레벨을 맞추면 오류가 없다.**



**0 rpart패키지는 55.56%의 전체 예측 성공률을 보인다.**

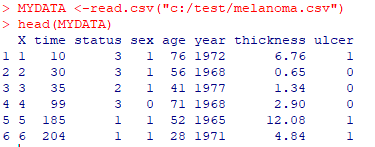
**2. kmeans 모델을 적용할 수 있는 데이터 셋을 찾아서 의사결정나무 모델을 구하여라. 그리고 실험 데이터로 결과를 확인해 보아라. 단 사전에 군집이 정해지지 않은 데이터 셋을 이용하여라.**

**- 2-1kmeans 모델 적용가능한지 확인하기.**

**①데이터 설정 : melanoma**

**melanoma: 악성 흑색 종에서 생존데이터**

**sex 군집을 구분해보겠다.**



**시각: 수술 후 며칠 동안 생존 한 시간으로, 검열되었을 수 있음.**

**상태: 연구 종료 시점의 환자 상태 1은 흑색 종으로 사망했음을 나타내고, 2는 여전히 살아 있음을 나타내고 3은 흑색 종과 관련이없는 원인으로 사망했음을 나타냄.**

**sex: 환자 성별; 1 = 남성, 0 = 여성.**

**나이: 수술 당시 연령 (년)**

**년: 작년.**

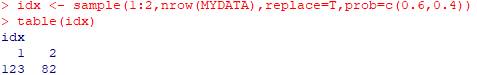
**두께: mm의 종양 두께.**

**궤양: 궤양의 지표; 1 = 현재, 0 = 없음**

**② 중심위치 설정을 위한 난수 생성**



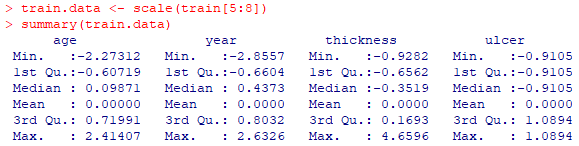
**③ 학습 데이터(train)와 실험 데이터(test) 분리: 60%, 40%**

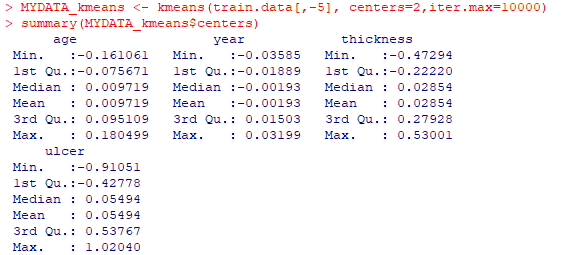


**학습 데이터 train, 실험 데이터 test**



**④학습 데이터 표준화: scale(데이터 셋)**

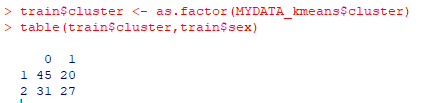


**⑤ k-means() 함수구동**

**⑥ k-means() 결과 보기**

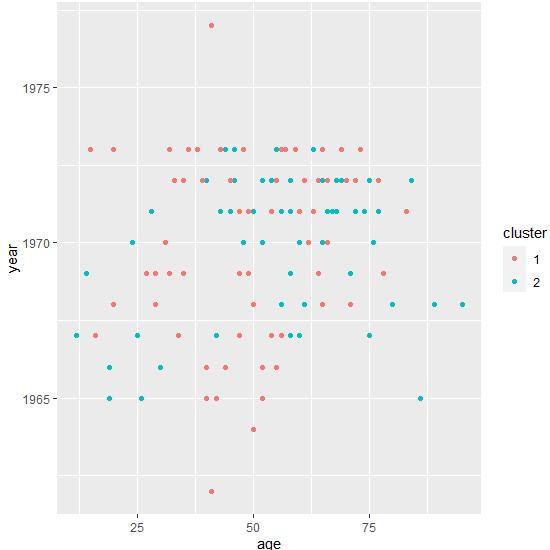
**0=여성**

**1=남성**

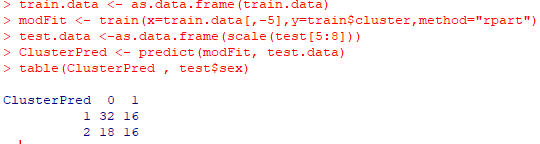


**- 정확도: 58.53% (72/123)**

**⑦ 데이터 셋 그래프로 확인하기**



**⑧실험 데이터에 적용**



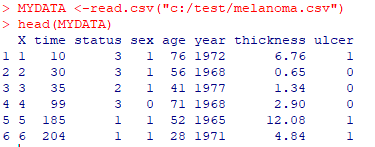
**정확도: 58.53% (48/82)**

**- kmeans모델을 사용했을 경우 정확도가 58.53%이 나왔으며 실험데이터에 적용 했을경우에도 놀랍게도 58.53%의 정확도가 나왔다. kmeans모델이 적용가능하였고 의사결정나무 모델을 구하겠다.**

**2-2 의사결정나무모델 구하기**

**rpart를 사용시**

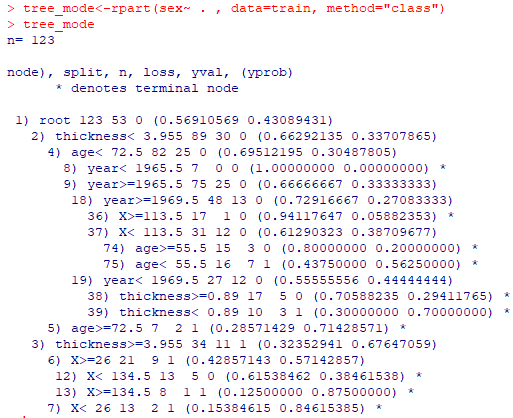
**① 데이터 삽입: melanoma 데이터 이용**



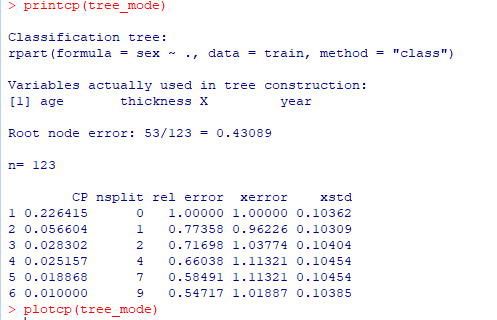
**② 학습(Train)/실험(Test) 데이터 설정: 학습 데이터 60%, 실험데이터 40%**

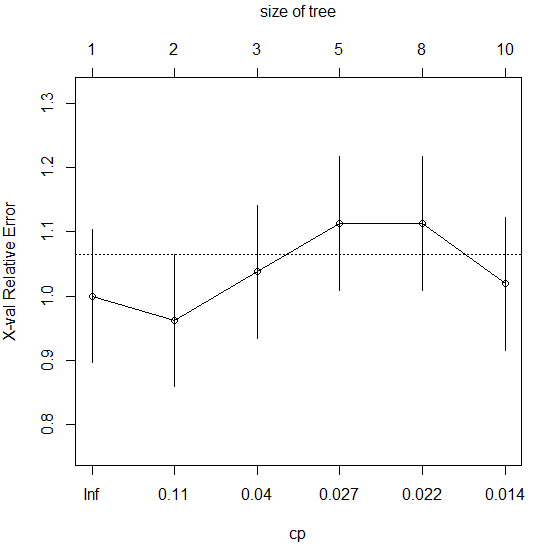


**③ 의사결정나무 모형 설정: rpart( 종속변수 ~. data=데이터셋)**



**- 트리 크기 결정: printcp() cp= complexity parameter 값 계산, plotcp()**

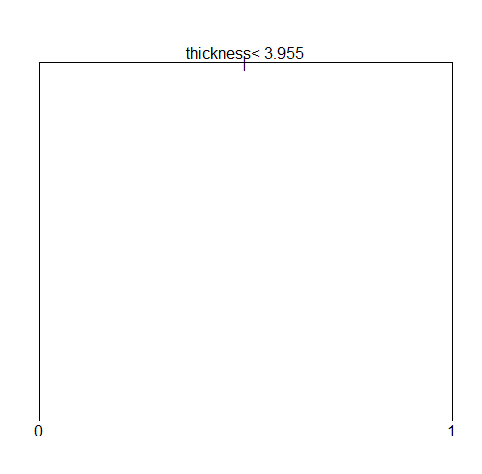




**④ 분류실시: 가지치기를 이용한 분류실시 prune()**

**cp는 xerror가 가장 낮은 split 개수를 선택하면 된다.**



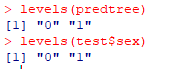


**⑤ 성과분석**

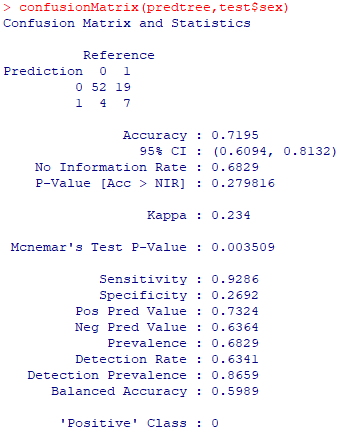
**level을 같은 형식으로 맞춰주지 않으면 오류가 난다.**







**같은 level 형식이면 오류가 안난다.**



**- rpart패키지는 71.95%의 전체 예측 성공률을 보인다.**