**빅데이터소프트웨어**

기말고사 대체 프로젝트

담당교수: 류귀열교수님

학과: 컴퓨터과학과

학번: 2017301086

이름: 황인애

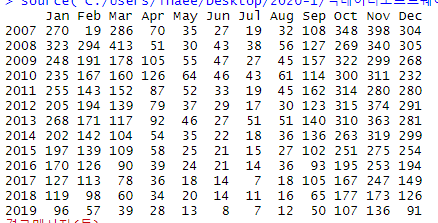
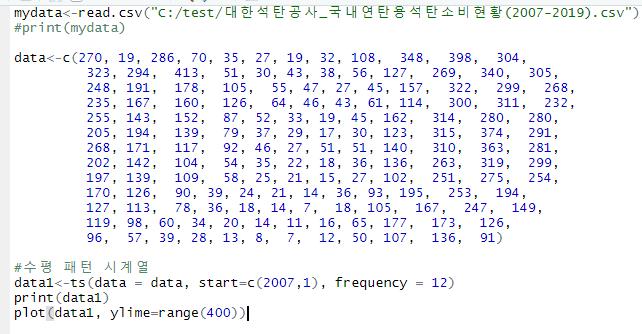
1. 시계열 분석 및 예측

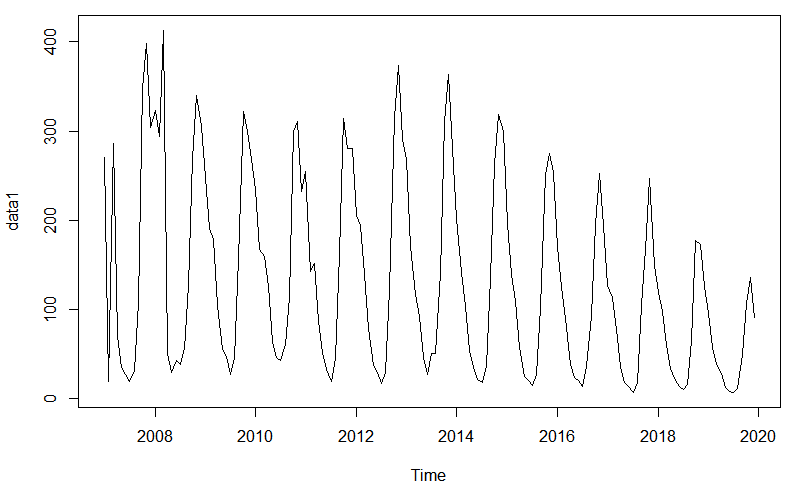
공공데이터 포털에서 국내 연탄용 석탄 소비 현황의 데이터셋을 가지고 시계열 분석을 하였습니다. 대한석탄공사에서 제공하는 국내 생산 무연탄의 월별 가정용 소비량의 데이터는 연도별에 따른 월별 소비량을 나타냅니다. 기본 단위는 천톤이며, 또한 연간 소비량과 전년대비증감율(%)의 데이터도 있습니다.

시계열 데이터는 시간에 따라 측정된 데이터입니다. 시계열 분석 방법은 분석보다는 예측에 많이 사용되기 때문에 활용도가 높습니다.

1. 수평 패턴(Horizontal Pattern) 시계열

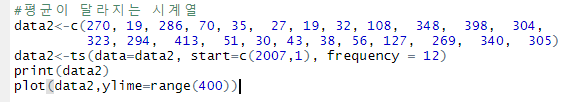
평균을 중심으로 랜덤하게 변화하는 패턴입니다. 다음의 데이터는 2007년 1월 ~ 2019년 12월까지 조사된 국내 연탄용 석탄 소비 현황입니다.



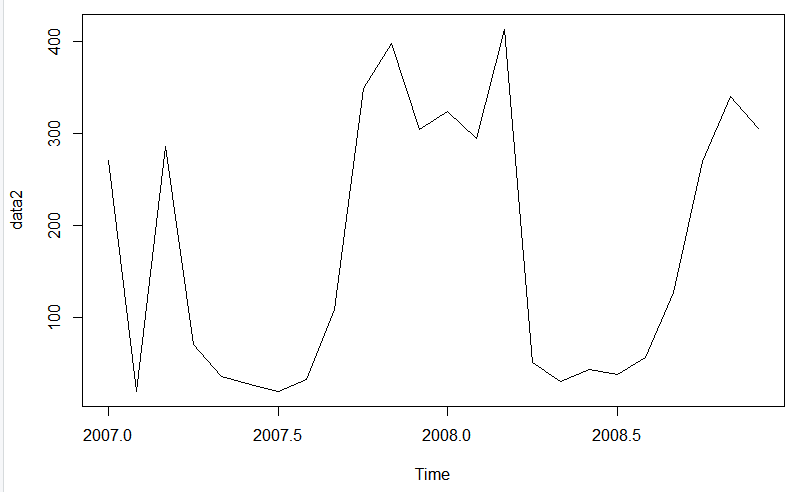


1. 평균이 달라지는 시계열 (Different Horizontal Pattern)

2007년과 2008년의 데이터를 가지고 작성한 시계열입니다. 이러한 그래프는 환경변화나 조건의 변동이 생기면 평균수준의 변화를 유발하여 다음과 같은 그래프 특징을 보입니다.



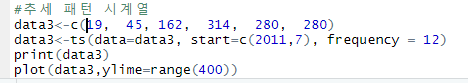


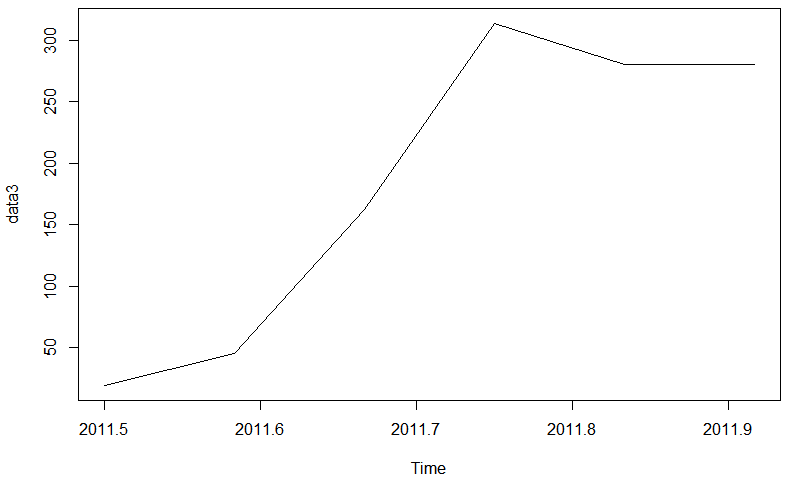


1. 추세 패턴 (Trend Pattern) 시계열

경제관련 시계열은 시간이 지남에 따라 커지는 경향을 보입니다. 이런 시계열 패턴을 추세 패턴이라고 합니다.

2011년 7월부터 12월까지의 데이터를 가지고 작성한 시계열입니다.

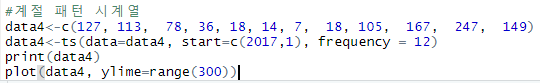


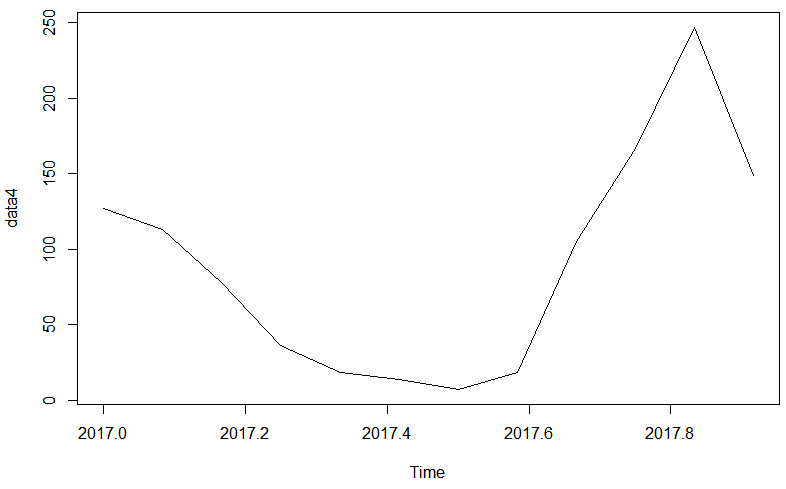


1. 계절 패턴 (Seasonal Pattern) 시계열

시계열 데이터는 시간이 지남에 따라 상승하고 하락하는 패턴을 반복할 수 있습니다. 계절에 따라 변하는 패턴을 계절 패턴이라고 합니다.

계절 패턴 시계열에 사용할 데이터는 2017년 1월부터 12월까지의 데이터입니다.



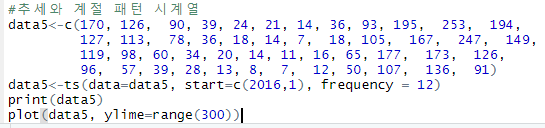


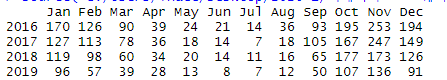
이 그래프의 특징은 1분기는 점점 감소하는 추세이고 2분기는 매우 낮고 3분기는 점점 증가하는 추세이고 4분기는 매우 높습니다.

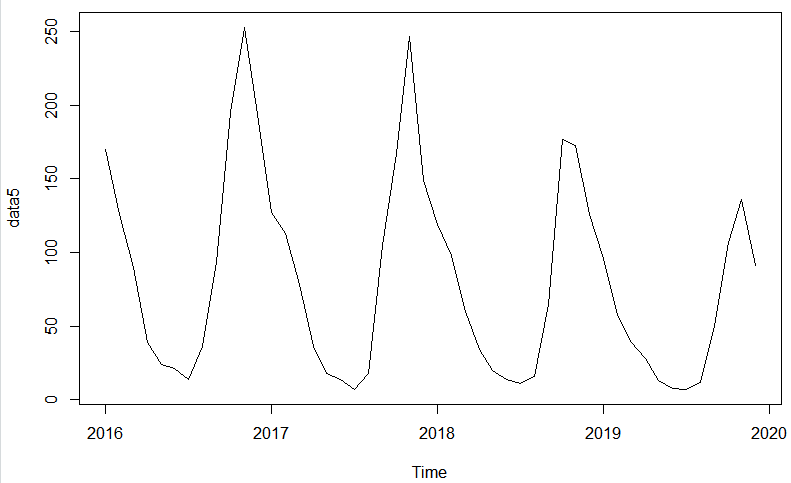
1. 추세와 계절 패턴 (Trend and Seasonal Pattern) 시계열

추세 패턴과 계절 패턴이 동시에 있는 시계열자료로 경제 데이터나 SNS 데이터에 나타나는 특징입니다.

추세와 계절 패턴에 사용할 데이터는 2016년부터 2019년도까지의 데이터입니다.





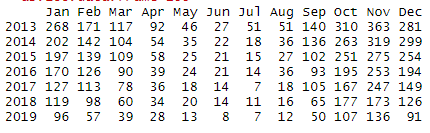
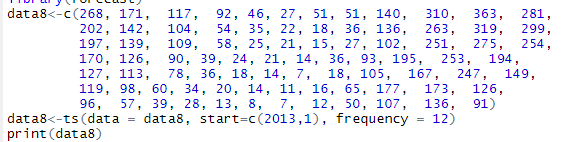


시간이 지남에 따라 감소하는 특징을 나타냅니다.

1. 예측모형 선택

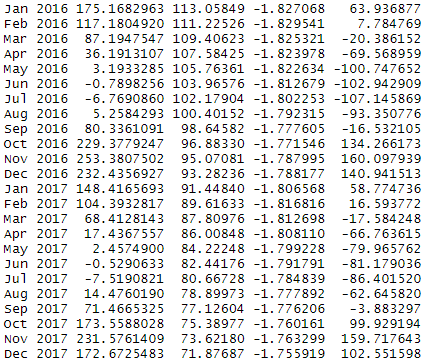
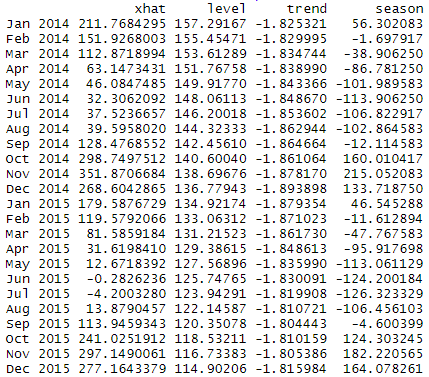
시계열 분석 방법은 여러 가지가 있으므로 시계열 자료를 이해하고 그에 적합한 시계열 분석 방법을 선택하는 것이 가장 중요합니다. 왜냐하면 시계열 패턴에 따라 분석방법이 다르기 때문입니다. 국내 연탄용 석탄 소비 현황의 데이터는 추세와 계절 패턴을 이용하여 분석 및 예측해보도록 하겠습니다.

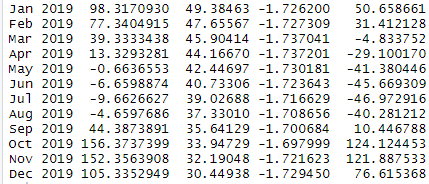
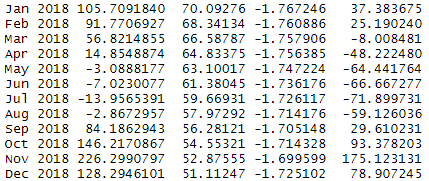
추세와 계절 패턴 두 가지 요소가 모두 있는 시계열 데이터를 추정하기 위해서는 기본적으로 추세 요인과 1월,2월, …, 12월의 계절 요인을 모두 추정하므로 최소한 2년 정도의 데이터가 필요합니다. 따라서 2013년부터 2019년까지의 데이터를 사용하도록 하겠습니다.



HolterWinters() 함수를 이용하겠습니다.

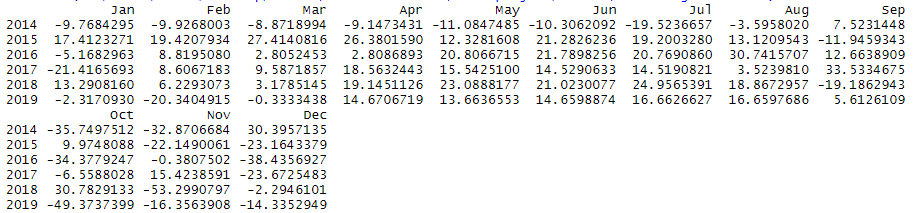




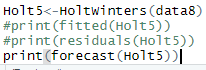


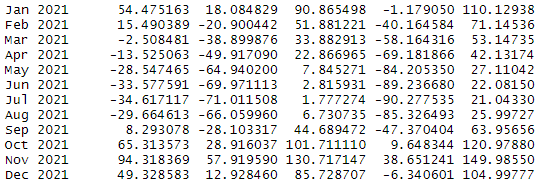
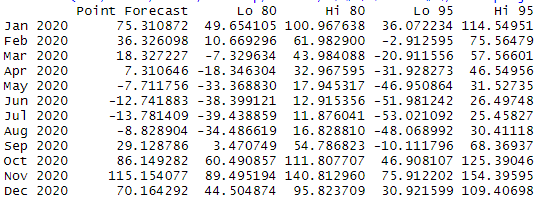
다음은 잔차입니다.



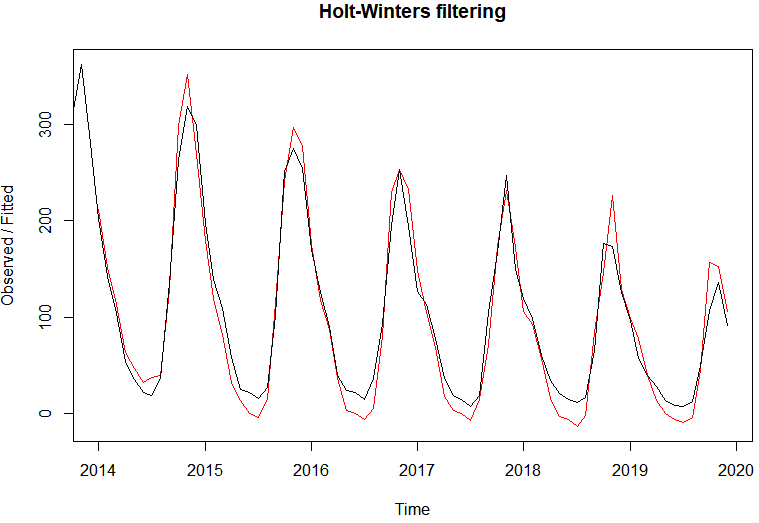


다음은 예측 값입니다.



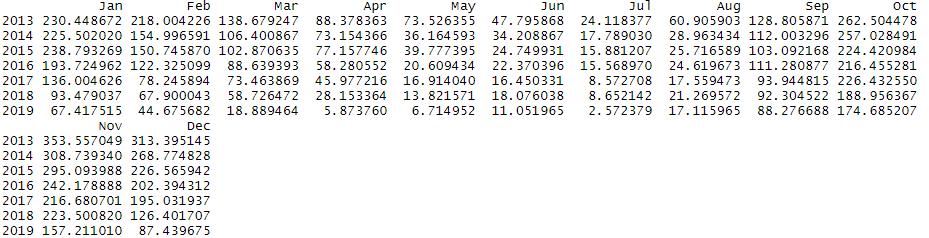


다음은 예측 값을 포함한 그래프입니다.



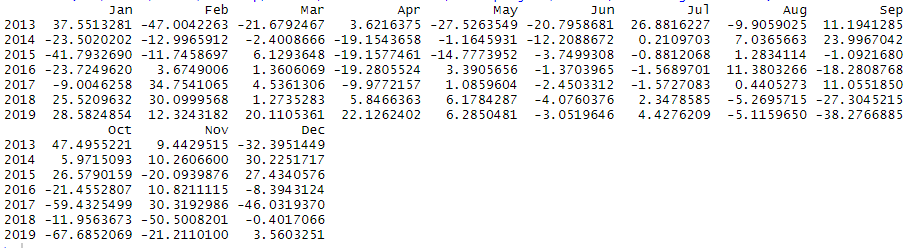
추세와 계절성 데이터를 적용하도록 하겠습니다.





다음은 잔차입니다.



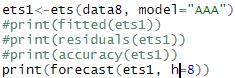


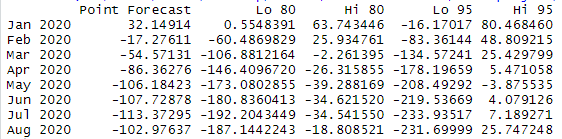
다음은 정확성 척도입니다.



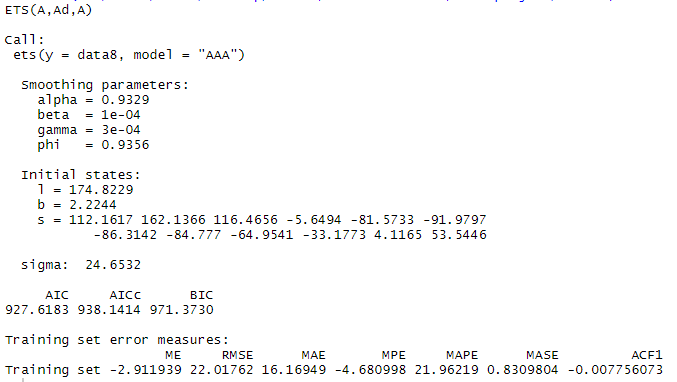


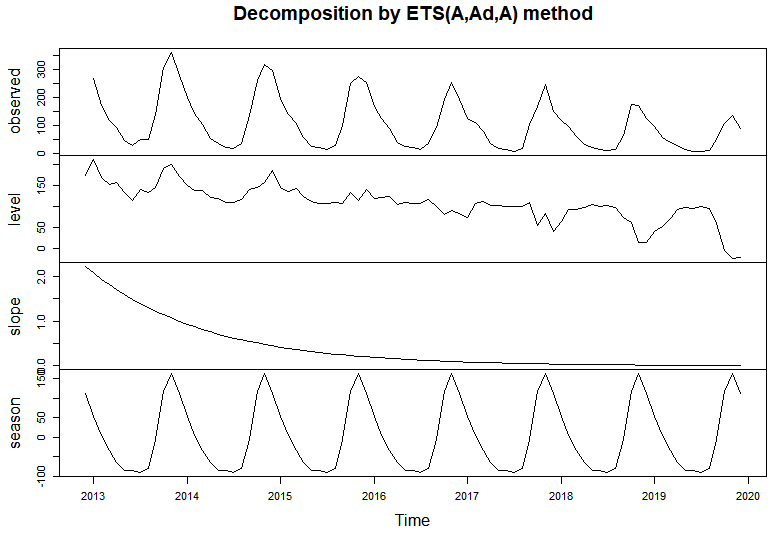
다음은 예측 값입니다.



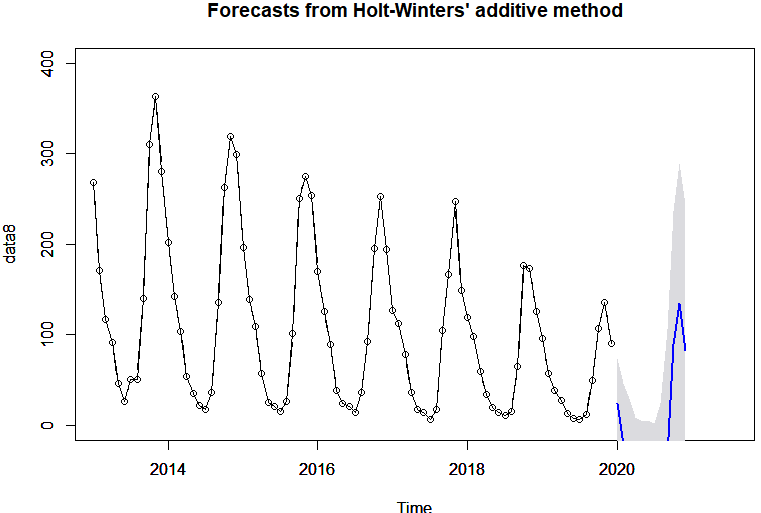


다음은 예측 요약 값과 그래프입니다.





다음은 예측값 그래프입니다.



이와 같은 분석과 예측을 통해 국내 연탄용 석탄 소비량은 시간이 지날수록 점차적으로 감소하는 경향을 보일 것입니다. 여름에 해당되는 달에는 석탄 소비량은 겨울에 비해 상대적으로 적고, 겨울에 해당되는 달에는 석탄 소비량은 여름에 비해 상대적으로 많습니다.

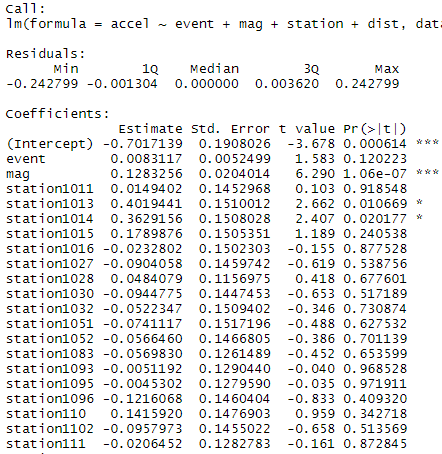
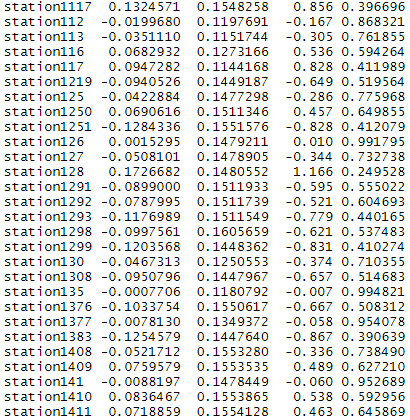
1. 판별분석

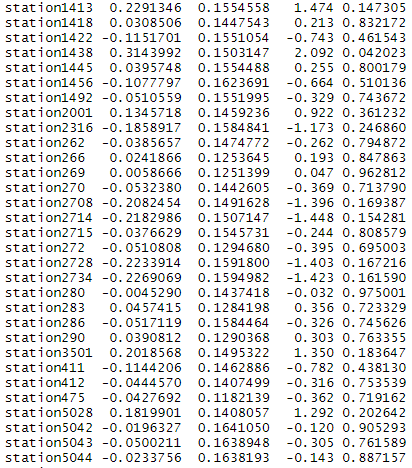
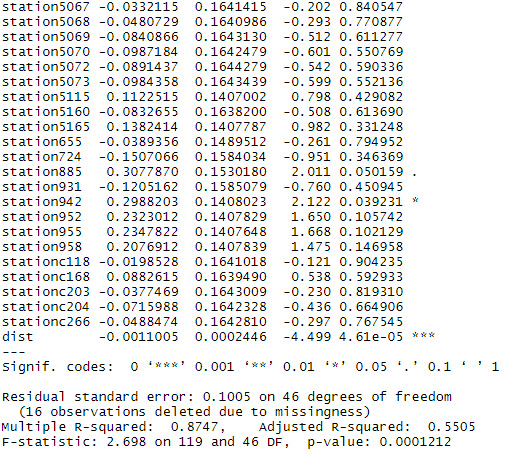
판별분석에 이용할 데이터셋은 attenu 데이터입니다. 이 데이터는 캘리포니아의 23개 지진에 대한 다양한 관측소에서 측정된 최대 가속도를 제공합니다. 이 데이터는 다양한 작업자가 지상 가속 거리의 감쇠 영향을 추정하는 데 사용되었습니다. 5개의 변수가 존재합니다. 변수 event는 번호입니다. 변수 mag는 순간 크기입니다. 변수 station는 스테이션 번호입니다. 변수 dist는 역 중심 거리입니다. 변수 accel는 피크 가속도입니다.

Attenu 데이터셋을 이용하여 다중회귀분석을 통해 변수 accel에 영향을 주는 독립변수를 단계별 제거법을 이용하여 선택하고, 선형관계가 유의하다면 오차를 검토하여 선형관계의 타당성을 설명하도록 하겠습니다.

1. 첫 번째 변수 제거



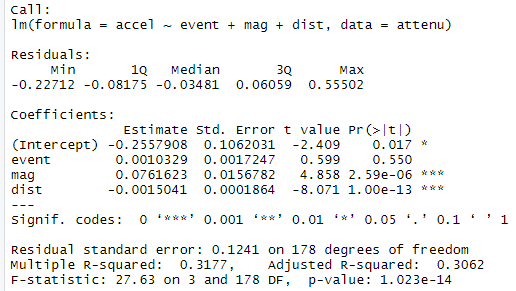
 

가장 높은 p-값을 가지는 station을 제거

1. 두 번째 변수 제거

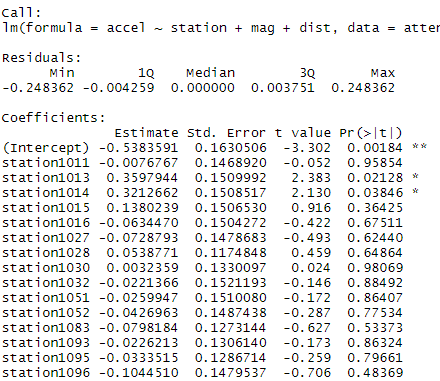
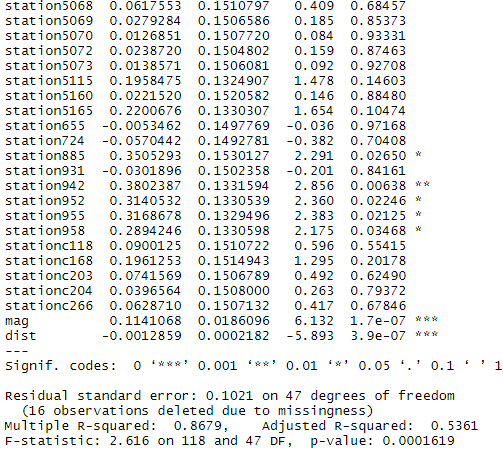




가장 높은 p-값을 가지는 event 제거

앞에서 제거한 station 변수를 포함여부 판단

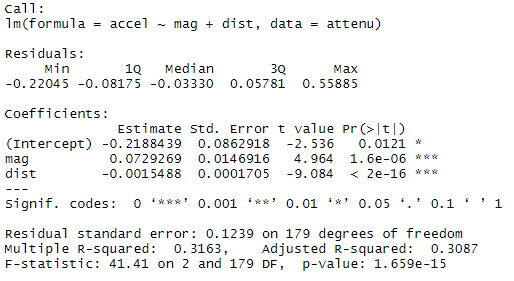


Station의 p-값이 유의하지 않으므로 제거합니다.

1. 세 번째 변수 제거





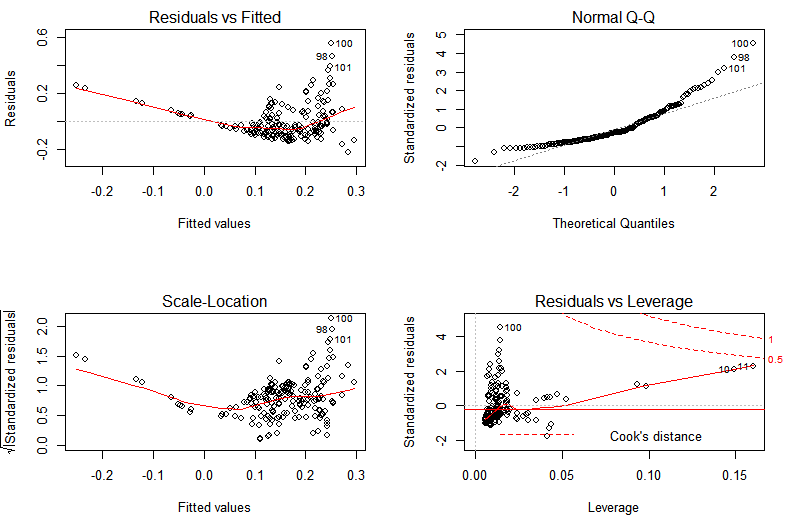
mag의 p-값이 0.0121로 유의수준 =10%으로 유의하므로 변수를 추가합니다.

유의하지 않은 변수가 없습니다.

따라서 선택된 모형은 accel=-0.2188439 + 0.0729269mag – 0.0015488dist이며, F-값은 41.41이고 p-값은 1.659\*10-15로 매우 유의합니다. 이 모형의 수정된 R^2=30.87%를 설명합니다.

이 모형의 타당성을 검정하기 위해서 잔차 그래프를 그려보겠습니다.





모형선택은 다중회귀분석에서 매우 중요한 문제입니다. 정해진 규칙이 없으므로 가장 좋은 모형을 찾기 위한 많은 노력이 필요합니다. 전방추가법, 후방제거법, 단계별제거법 등이 있습니다.