2017.12.10

금융통계학 분석

: 수출입 통계 중 우리나라 수출금액을 통한 모형설정

정보통계학과 2015016015 조희연

Ⅰ 서 론
Ⅱ 모형설명 및 실증분석방법
Ⅲ 결 론

Ⅰ 서론

한국의 월별 수출금액 자료를 바탕으로 모형을 예측하고 분석방법 별로 비교하여 최종모형을 설정해 2017년도 10월 이후의 수출금액을 예측한다.

자료는 관세청에서 제공하는 수출입 무역통계 중 한국의 2000년 01월부터 2017년 10월까지의 수출금액이다. 수출금액은 수출면장에 나타난 본선인도가격(FOB) 기준의 제품 수출총액을 말한다. 제품에는 원료, 중간재, 최종재가 있으며 FOB(Free on Board)는 물품 자체가격 + 운송비용(생산지에서 물품을 수출항구까지 운송하는 데 드는 비용)를 뜻한다. 한국은 세계 10대 수출국 중의 하나로 17년 상반기 수출액 규모는 세계 6위, 무역 규모는 9위이며, 2008년부터 국내총생산(GDP)에서 수출 비중이 46% 이상일 정도로 국민 경제에서 많은 부분을 차지하고 있다. 수출입 거래량이 상당한 만큼 국제 무역환경에 많은 영향을 받기 때문에 국내 기업들은 무역 동향을 파악해야 한다. 필요성에 따라 관세청은 수출입무역에 대한 통계를 70년대 후반부터 국제 무역통계 기준(IMTS)에 따라 월, 분기, 반기 또는 년 단위로 작성하여 공표하고 있다. 수출입무역통계는 사전적인 의미로 국민경제가 다른 나라와의 사이에서 행하는 화물 교류에 관한 통계이며, 광의적으로는 컨테이너 물동량, 선박(항공기) 입출항 등 수출입 화물통계와 여행자 입출국 통계 등 일국(一國)의 경제영역을 통과한 모든 교류내역이 포함된 통계를 말한다. 개방되는 수출입무역통계정보는 국가관세종합정보망의 수출입통관 분야 정보를 활용하여 대국민 서비스용으로 제공되는 다양한 항목별 수출입 건수·금액·중량 등에 대한 수출입실적 정보이다. 국가관세종합정보망은 우리나라와 외국 간 거래되는 모든 물품의 반출입과 관련된 각종 신고·신청 자료 등 외부(국민+기업+정부기관)으로부터 입수되는 물품의 수출입통관 전 과정을 정보화한 시스템이다. 우리나라는 수입은 CIF 금액기준으로, 수출은 FOB 금액기준으로 집계한다. 이때 CIF(Cost Insurance Feight)는 운임보험료 포함가격의 의미로 수출입 지급조건의 하나로, 판매자가 화물의 선적에서 보내는 목적지까지의 모든 운임과 보험료를 부담하는 방식이다. 또한 관세청은 무역통계부호, 수출입 화물 총괄, FTA 발효국과의 수출실적, 남북교역 반출입 실적, 10대 수출입 품목, 품목별 수출입 실적, 국가별 수출입 실적 등 총 43종 파일데이터(xml, csv 등), Open API로 공공데이터포털 및 수출입무역통계사이트에 제공하고 있다. 수출입 무역통계자료는 매월 15일에 전월 자료를 반영한다. 원 데이터는 기간, 수출건수, 수출금액, 수입건수, 수입금액, 무역수지로 단위는 천 불(USD 1,000)이며 이 중 수출금액만 따로 정리하여 사용하였다.

Ⅱ모형설명 및 실증분석방법

Ⅰ. 모형설명

시계열모형에 의한 예측은 예측될 변수 자체의 과거의 자료에서 어떠한 패턴을 발견하여 미래에도 그러한 패턴이 특성을 잃지 않고 반복될 것이라는 가정 하에서 모형을 확립하여 예측하는 방법이다. 그러므로 확립된 시계열모형은 특정한 자료의 집합에 모형이 얼마나 잘 적합한가에 전적으로 평가되어 진다(Newbold & Reed,1979). 이러한 시계열모형에 의한 예측방법에는 평활법(smoothing methods), 분해법(decomposition methods) 등과 같은 전통적 시계열분석 방법과 확률적 시계열분석 방법이 있다. 확률적 시계열분석에는 ARIMA(Auto- Regressive integrated moving average) 모형이 있다. 설명모형에 의한 예측방법은 예측되어야 할 변수를 하나 이상의 설명변수들과 인과관계가 존재한다고 가정하고, 과거의 자료에서부터 이들 간의 인과적 연관성을 추정하여 미래에도 그러한 관련성이 지속될 것이라는 가정 하에 미래의 값을 예측하는 방법이다. 이러한 인과모형에는 회귀모형, 시계열모형, 투입-산출모형 등이 있다. 모형 적합 후 예측의 사후평가는 예측값과 실제값과의 차이인 예측오차를 사용하는 평균제곱오차(Mean squared error : MSE)와 평균절대편차(Mean absolute deviation : MAE)를 주로 사용한다. 작을수록 정확한 예측값임을 나타낸다. 여러 예측 방법 중 확률적 시계열 분석방법인 ARIMA 모형을 통한 예측법, 평활법을 이용한 예측법, 인공신경망을 이용한 예측법을 소개하려 한다.

1. ARIMA 모형

Box- Jenkins(1970)는 시계열 분석에 대한 접근 방법을 정립했고, 또 이 방법을 더욱 더 일반화하여 통합자기회귀이동평균 (Autoregressive integrated moving average : ARIMA) 모형화를 토대로 하는 통계적인 이론 체계를 구축하게 되었다. 이 분석 방법은 관찰된 시계열 자료를 하나의 시계열 모집단(모형)으로부터 구축된 표본으로 간주하여 이들이 어떤 확률적 성질을 만족하는가를 조사하고 통계적 추정 및 검정을 통하여 적절한 시계열 모형을 수립하는 것이다. 구체적으로 Box-Jenkins방법은 관찰된 시계열자료가 어떤 확률적 성질을 가지고 있으며 어떤 시계열 모형(ARIMA model)이 적합한가를 찾기 위한 모형 식별(model identification), 모수 추정(parameter estimation) 그리고 모형의 적합성 진단(model diagnostic checking)의 세 단계를 거친다. 모형식별은 관찰된 시계열 자료에 대하여 적절할 것이라고 생각되는 몇 개의 모형을 선정한 후 자료의 그래프, 관련 통계량 등을 비교하여 가장 적절한 하나의 모형을 선정하는 것이다. 이렇게 선정된 모형은 잠정적인 것으로 계속적인 분석을 통하여 개선시켜야 한다. 모형을 선정하는데 주의해야 할 점은 모수 축약의 원칙(principle of parsimony)에 충실해야 한다는 것이다. 이는 관찰된 자료를 적절히 표현하면서 모수의 수가 가장 작은 모형을 선정하는 것으로 가능한 한 단순하게 표현하려는 것이다. 선정된 모형은 하나 이상의 모수를 포함하고 있는데 이 모수는 관찰된 자료로부터 추정할 수 있다. 모수의 추정방법으로는 적률법, 최우추정법 그리고 최소제곱법등이 주로 사용된다. 모형이 선정되고 나면 이것이 원래의 시계열 자료를 얼마나 잘 표

현하는가를 검정해야 하는데 이를 모형의 적합성진단이라고 한다. 만일 모든 가정과 조건이 만족하면 그 모형은 적절한 것이고 따라서 그 모형을 기초로 미래값을 예측할 수 있다. 그러나 가정 및 조건이 만족되지 않는다면 다시 모형식별 단계로 되돌아가서 적절한 모형이 찾아질 때까지 위의 세 단계를 반복하여야 할 것이다. 우선 시계열 모형 종류에 대해서 먼저 설명하도록 한다.

가. 시계열 모형 종류

ㄱ. 일반 선형 시계열 모형

일반선형모형(general linear model)은 정상시계열 모형을 모두 포함하여 자기상관이 존재하는 시계열 X_t 는 식 (1.1)과 같이 서로 독립적인 확률변수들의 선형결합으로 나타낼 수 있다.(Wald 1938).

$$X_t = a_t + \phi_1 a_{t-1} + \phi_2 a_{t-2} + \cdots = \sum_{i=0} \phi_i a_{t-i} \quad (1.1)$$

여기서 $\phi_0 = 1$, a_t 는 평균이 0이고 분산이 σ_{a2} 그리고 $cov\ (a_t\ ,\ a_{t+k}\ ) = E\ (a_t\ ,\ a_{t+k}\ )$을 갖는 백색잡음 과정이고 $\sum_{i=0} \phi_i\ a_{t-i} < \infty$이다. 또한 X_t 의 평균은 편의상 0으로간주한다.

위 모형이 정상적이기 위해서는 평균, 분산 및 공분산이 시점 t에 의존하지 않아야 한다. 일반선형모형의 공분산 구조(covariance structure)인 평균, 분산, 자기공분산함수, 그리고 자기상관함수는 아래와 같다.

$$\begin{aligned} E\ (X_t\ ) &= E\ (a_t + \phi_1 a_{t-1} + \phi_2 a_{t-2} + \cdots) \\ &= E\ (a_t\ ) + \phi_1 E\ (a_{t-1}\ ) + \phi_2 E\ (a_{t-2}\ ) + \cdots \\ &= 0 \end{aligned}$$

$$\begin{aligned} Var(X_t\ ) &= E\ (a_t + \phi_1 a_{t-1} + \phi_2 a_{t-2} + \cdots) \\ &= E\ (a_{t2}\ ) + \phi_{12}\ E\ (a_{t-12}\ ) + \phi_{22}\ E\ (a_{t-22}\ ) + \cdots \\ &= \sigma_{a2} \sum_{i=1} \phi_{i2} \\ &= \gamma_0 \end{aligned}$$

$$\begin{aligned} \gamma_k &= E\ (X_t - \mu)(X_{t+k} - \mu) \\ &= E\ (X_t X_{t+k}\ ) \\ &= E\ (\phi_k a_{t2} + \phi_1 \phi_{k+1} a_{t-12} + \phi_2 \phi_{k+2} a_{t-2} + \cdots) \\ &= \sigma_{a2} \sum_{i=0} \phi_i \phi_{t+k} \end{aligned}$$

여기서 $\sum_{i=0} \phi_{i2} < \infty$은 일반선형모형으로 정의된 모든 시계열의 정상성(stationarity)을 만족하는 필요조건이다.

ㄴ. 자기회귀 이동평균 모형 : ARMA(p,q) model

Box와 Jenkins(1976)는 시계열 모형에서 자기회귀항과 이동평균항을 혼합한 자기회귀 이동평균 모형(Autoregressive Moving Average Model : ARIMA)을 제안하였으며, ARIMA(p,q)이라 하고 모형은 식 (1.2)와 같다.

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \quad (1.2)$$

여기서 오차항 $a_t \sim \mathrm{WN}(0, \sigma_{a2})$이다.
또는

$$\phi(B) X_t = \theta(B) a_t .$$

여기에서 $\phi(B)$와 $\theta(B)$는 각각 자기회귀(auto regressive ;AR) 연산자와 이동평균(moving average ; MA) 연산자이며 식 (1.3)과 같이 나타낼 수 있다.

$$\phi(B) = 1 - \phi_1 B - \phi_2 B_2 - \cdots - \phi_p B_p,$$
$$\theta(B) = 1 - \theta_1 B - \theta_2 B_2 - \cdots - \theta_q B \quad (1.3)$$

여기서 B는 후진연산자(backward shift operator)이다. 정상성과 가역성을 만족하기 위하여 $|\phi(B)| = 0$ 과 $|\theta(B)| = 0$의 모든 근들이 단위원 밖에 존재하해야 한다. 여기서 가역성(invertible)이란 이동평균모형이 자기회귀모형으로 가역될 수 있다는 것이다.

ㄷ. 누적 자기회귀 이동평균 모형 : ARIMA(p,d,q) model

정상성을 만족하지 않는 시계열은 적당한 차분을 취하게 된다. 이렇게 정상성을 만족하지 않은 자료를 비정상 시계열이라 하며 차수가 p, d, q인 비정상 시계열 모형을 누적자기회귀 이동평균 모형(Autoregressive Integrated Moving Averagemodel)이라 하며 식 (1.4)와 같이 나타낼 수 있다.

$$\phi_p(B)(1 - B)_d X_t = \theta_0 + \theta_q(B)(1 - B)_d a_t , \; a_t \sim \mathrm{WN}(0, \sigma_{a2}) \quad (1.4)$$

여기서 $\phi_p(B) = 1 - \phi_1 B - \phi_2 B_2 - \cdots - \phi_p B_p$, $\theta_q(B) = 1 - \theta_1 B - \theta_2 B_2 - \cdots - \theta_q B_q$ 이며 각각 자기회귀(autoregressive ;AR)연산자와 이동평균(moving average ; MA)연산자이다. $\theta_0 = \mu(1 - \phi_1 - \cdots - \phi_p)$로 계열의 평균과 관련되며 모수 θ_0는 d의 값에 따라 매우 다른 역할을 하게 된다. $d = 0$인 경우 원 시계열은 정상적이고, $d \geq 1$인 경우에는 θ_0는 결정적 추세항이 된다.

나. 모형식별 과정

관찰된 한 시계열에 대하여 여러 가지 유사한 모형을 세울 수 있지만 중요한 것은 모수 절약의 원칙(principles of parsimony)에 따라서 가급적 모수의 수를 적게 포함하는 모형을 설정하여야 한다 .이를 토대로 Box-Jenkins의 시계열 분석방법에서 첫 번째 단계인 관찰된 자료가 어떤 시계열 모형으로부터 추출된 표본으로 볼 수 있는지를 밝히는 모형식별(model identification) 단계에 대하여 알아보고자 한다.

모형식별이란 관찰된 시계열 자료에 의하여 적합한 잠정적인 누적 자기회귀이동평균모형(ARIMA model)을 찾아내는 단계를 말하며 먼저 다음의 ARIMA$(p,\ d,\ q)$일반모형을 고려하자.

$$\left(1 - \phi_1 \mathrm{B} - \cdots - \phi_p \mathrm{B}^{\ p} \right) \left(1 - \mathrm{B} \right)^d \mathrm{Z}t = \theta_0 + \left(1 - \theta_1 \mathrm{B} - \cdots - \theta_q \mathrm{B}^{\ q} \right) at$$

구체적으로 모형식별은 주어진 자료에 대하여 가정된 위 모형 (p,d,q)의 차수를 결정하는 방법을 말하며 다음과 같은 단계를 거쳐 모형을 식별한다.

[단계 1] 시계열 자료의 산점도를 나타내고 적당한 변환을 선택한다.

모든 시계열 분석에서의 첫 번째 단계는 자료의 산점도를 통하여 시계열의 경향, 계절성, 이상치, 일정하지 않은 분산, 비정규성 등 비정상성의 현상을 갖고 있는지에 대한 견해들을 얻게 된다. 이러한 작업은 종종 자료 변환에 대한 기초 정보를 제공한다. 시계열 분석에서 가장 일반적으로 사용되는 변환은 분산안정화 변환과 분이다. 차분은 음의 값을 만들 수 있기 때문에 차분을 하기 전에 항상 분산 안정화 변환을 먼저 해야 한다. 일정한 분산을 갖지 않는 시계열은 종종 대수 변환을 요로 한다. 좀 더 일반적으로 분산을 안정시키기 위하여 Box-Cox's의 멱변환을 용할 수 있다.

[단계 2] 필요한 차분의 정도를 확인하기 위하여 시계열의 SACF와 SPACF를 계산하고 관찰한다.

일반적인 규칙은 다음과 같다.

1. SACF가 완만하게 줄어들고 SPACF가 시차 1이후에 절단현상이 나타나면 이것은 차분이 필요함을 의미한다. 즉, 일차차분 $(1\text{-B})\,\mathrm{Z}_t$를 취한다.
2. 좀 더 일반적인 비정상성을 제거하기 위해서는 $\mathrm{d} > 1$에 대한 차분 $(1\text{-B})^d \mathrm{Z}_t$ 이 필요할 수도 있다. 그러나 대부분의 경우, d는 0, 1 또는 2이다. 하지만 필요한 차분은 과모수화를 일으키므로 피해야 한다.

[단계 3] SACF와 SPACF로 p와 q를 식별한다.

차수 p와 q의 식별을 위하여 적당하게 변환하고 차분된 시계열의 SACF와 SPACF를 계산하고

살펴본다. 하지만 SACF와 SPACF로 p와 q를 식별하는 것이 애매한 경우가 많고, AR과 MA가 혼합된 ARMA 모형에 대해서도 식별이 어려울 수 있다. 이때, 모형 식별을 위한 통계량으로 AIC와 BIC 통계량이나 SBC 통계량을 사용하게 되는데 각 모형의 통계량을 비교하여 가장 작은 값을 가지는 모형으로 식별하는 것이 보다 정확할 수 있다. 선택 p는 AR 특성함수$(1 - \phi_1 \mathrm{B} - \cdots - \phi_p \mathrm{B}^{\ p})$에서 가장 상위의 차수이고 q는 MA 특성함수 $(1 - \theta_1 \mathrm{B} - \cdots - \theta_q \mathrm{B}^{\ q})$에서 가장 상위의 차수이다. 일반적으로 이러한 p와 q의 필요한 차수들은 3보다 작거나 같다. AR과 MA 모형 사이에는 강한 상대적인 면이 존재한다는 것으로 알려져 있다. 적당한 ARIMA 모형을 세우기 위해서 이상적으로는 최소 n = 50인 관측치가 필요하고 SACF와 SPACF의 개수는 약 n/4가 필요하다. 물론 때때로 양질의 데이터에 대하여 아주 적은 표본수로 적당한 모형을 식별할 수도 있지만 일반적으로 50개 이상이 요구된다.

ㄱ. ARMA 모형의 식별

비정상 시계열 모형은 차분을 사용하여 정상시계열로 변환시킬 수 있음을 보았다. 즉, d −차분을 통하여 $\mathrm{ARIMA}(p,\ d,\ q)$모형은 $\mathrm{ARMA}(p,\ q)$모형으로 변환할 수 있는 것이다. 실제 주어진 시계열이 정상적이냐 비정상적이냐의 판단 기준은 일반적으로 시계열 산점도와 ACF에 의존한다. 즉, 시계열이 정상적이면 시차가 증가할 때 ACF의 절대값이 감소하여 0에 수렴하게 되지만 비정상 시계열의 경우는 시차가 증가하더라도 매우 서서히 감소한다. 따라서 시계열 자료가 주어졌을 때 모형식별을 하기 위해서는 먼저 차분의 차수 d를 결정해야 한다. 이 때, d는 ACF가 급격하게 감소할 때의 차수로 결정하면 된다. d가 결정된 후에는 $\mathrm{ARMA}(p,\ q)$의 차수인 p와 q를 결정해야 하는데 이때는 각 모형의 형태를 표본의 ACF와 표본 PACF가 어떤 차수의 모형에 맞는가를 살펴보면 된다. 일반적으로 $\mathrm{AR}(p)$모형, $\mathrm{MA}(q)$모형 및 $\mathrm{ARMA}(p,\ q)$모형의 각 모형들에서 나타나는 다음과 같은 ACF와 PACF의 형태를 바탕으로 차수 p와 q를 결정한다.

(1) $\mathrm{AR}(p)$ 모형 : ACF는 지수적으로 감소하나 PACF는 시차 $(p + 1)$이후 절단현상이 일어난다.

(2) $\mathrm{MA}(q)$ 모형 : $\mathrm{AR}(p)$모형의 경우와 반대로 ACF가 시차 $(q + 1)$이후 절단현상이 나타나고 PACF는 지수적으로 감소한다.

(3) $\mathrm{ARMA}(p,\ q)$ 모형 : ACF가 시차 $(q - p)$이후에 지수적으로 감소하며 PACF는 시차 $(p - q)$이후에 지수적으로 감소한다. 그러나 ARMA$(p,\ q)$모형의 식별은 일반적으로 쉽지 않고 이것이 Box-Jenkins 분석 방법의 큰 약점이기도 하다. 따라서 식별을 위한 통계량 등을 고려해서 종합적으로 판단을 해야 한다.

SAS나 Minitab등과 같은 통계패키지나 R에서 ACF나 PACF의 존재 유무에 대해서는 점선으로 제공하고 있다. 즉, 점선 내에 ACF나 PACF의 막대그래프가 들어가있으면 ACF의 값이 존재한다는 가설이 기각되어 ACF나 PACF가 0인지 아닌지는 점선 내에 있는가 아닌가로 판단하면 된다.

지금까지 설명한 각각의 ARMA 모형의 ACF와 PACF의 형태를 요약하면 다음과 같다.

모형		ACF ρ_k	PACF ϕ_{kk}
AR(1)	(1,0)	지수감소함수	$\phi_{kk}=0$, k>1
AR(2)	(2,0)	지수감소함수 혹은 점차적으로 진폭이 줄어드는 사인곡선을 보인다.	$\phi_{kk}=0$, k>2
AR(p)	(p,0)	지수감소함수 혹은 점차적으로 진폭이 줄어드는 사인곡선을 보인다.	$\phi_{kk}=0$, k>p
MA(1)	(0,1)	$\rho_k=0$, k>1	지수감소함수
MA(2)	(0,2)	$\rho_k=0$, k>2	지수감소함수 또는 점차적으로 진폭이 줄어드는 사인곡선의 양상으로 나타난다.
MA(q)	(0,q)	$\rho_k=0$, k>q	지수감소함수 또는 점차적으로 진폭이 줄어드는 사인곡선의 양상으로 나타난다.
ARMA(1,1)	(1,1)	시차 k=1에서부터 지수감소함수	시차 k=1에서부터 지수감소함수
ARMA(p,q)	(p,q)	시차 (q-p)이후부터 지수감소함수 또는 감소사인곡선을 보인다.	시차 (p-q)이후부터 지수감소함수 또는 감소사인곡선을 보인다.

① 모형식별을 위한 통계량

ACF와 PACF를 이용한 ARMA 모형식별 방법은 AR모형이나 MA모형에는 비교적 식별이 잘되고 효과적이지만 그것들이 혼합된 모형인 ARMA모형에 대한 식별은 일반적으로 어렵다고 알려져 있고 주관적인 결정에 의존하게 되어 Box-Jenkins방법의 약점으로 알려져 있다. 이러한 약점을 극복하기 위한 노력이 1970년대 이래 계속되고 있다. Akaike(1970)의 FPE(Final Prediction Error Function), kaike(1972)의 AIC(Akaike Information Criterion), BIC(Baysian Information Criterion) 그리고 chwartz(1978)의 SBC(Schwartz's Bayesian Criterion)등이 모형식별을 위한 통계량들로써 AIC에대해서만 간략히 소개한다.

AIC는 모수들의 개수가 M개인 통계적 모형이 자료에 적합하다고 가정 했을때, 모형 및 차수의 적합을 위한 통계적인 근거를 위하여 Akaike(1973, 1974a)는 소개한 정보기준이다. 이 기준은 AIC(Akaike's information criteria)로 불리우고

$$\mathrm{AIC}\ (\mathrm{M}) = -\ 2\ \ln\ [\mathrm{max}imum\ likelihood] + 2\mathrm{M}$$

와 같이 정의된다. 여기서 M은 모형에서의 모수들의 개수이다. ARMA(p, q)모형의 최적 차수는 AIC(M)이 최소화되는 p와 q의 함수인 M(즉, $\mathrm{M} = p + q$)의 값에 의하여 선택된다.

다. 모수추정

모형식별 후 관찰된 시계열 $\mathrm{Z}_1, \mathrm{Z}_2, \cdots, \mathrm{Z}_n$으로 ARIMA 모형의 모수를 추정하는 문제를 다룬다. 자료에 알맞은 모형으로 잘 식별되었다고 가정하고 그래서 차수 p, d, q가 결정되었다고 하자. 비정상 시계열도 d만큼 차분하면 정상시계열로 고려될 수 있으므로 정상시계열로서 여러 가지 모수 추정방법 중 가장 많이 사용되는 적률추정법, 최우추정법 그리고 최소제곱 추정법에 의한 모수 추정이 가장 많이 사용된다.

라. 모형진단

일단 ARIMA모형의 계수에 대한 정밀한 추정값을 얻은 후, 우리가 가지고 있는 자료가 그 모형에 잘 적합 되었는지를 알아보는 모형진단의 단계를 거쳐야 한다. 이 단계에서는 추정된 모형이 통계적으로 적합한가의 여부를 결정한다. 모형진단 단계는 식별단계와 연관되어 있다. 모형진단에서 모형이 부적합하다고 판명되면 우리는 다른 모형을 선택하기 위하여 식별단계로 되돌아간다. ARIMA모형의 통계적 적합성을 진단하는 가장 중요한 검정은 잔차들이 독립이라는 가정에 대한 검정이다. 이 장에서는 먼저 이 가정을 만족하느냐의 여부를 검정하기 위하여 잔차의 자기상관함수(ACF)에 관하여 논의한 다음 여러 가지 모형진단기법들에 관하여 고찰해 보기로 한다.

ㄱ. 잔차의 독립성에 대한 검토

관찰치들의 실제값과 적합된 모형으로부터 얻어진 적합된 값(fitted value)과의 차이를 잔차(residuals)라 한다. 잔차를 수식으로 표현하면 $\hat{a}_t = \mathrm{Z}_t - \hat{\mathrm{Z}}_t$, $t = 1, 2, \cdots, n$ 가 된다. 통계적으로 적합한 모형은 잔차들이 서로 독립이다. 이 말은 잔차가 서로 자기상관관계에 있지 않다는 사실을 의미한다. 실제로 백색잡음 a_t를 관찰할 수 없으므로 추정된 백색잡음인 $\hat{a}_t$를 관찰한다. 모형진단 단계에서는 잔차들의 독립성에관한 가설을 검정하기 위하여 추정된 모형으로부터 얻어진 관찰된 잔차에 대한 분석을 실시한다. 백색잡음은 우리가 모형구축을 하는 변수인 Z_t의 요소이다. 그러므로 만약 백색잡음이 서로 연속적이고 상관관계가 있으면 모형의 AR이나 MA부분에 의하여 설명되지 않는 Z_t 상에 자기상관관계가 존재하게 된다. ARIMA모형구축의 전반적인 아이디어는 Z_t에 포함된 모든 형태의 자기상관관계를 AR이나 MA부분을 이용하여 가장 간결한

모형을 구축하는 것이다. 만약 잔차가 자기상관관계가 있으면 이는 백색잡음이 아니므로 독립성 가설에 부합하는 잔차를 가진 다른 모형을 찾아야만 한다. 잔차가 자기상관관계가 있을 때는 추정된 ARIMA모형을 어떻게 재형성할 것인가를 고려해야 한다. 이것은 초기의 추정된 자기상관함수와 편자기상관함수를 재검토해야 한다는 사실을 의미한다.

① 잔차 자기상관함수(ACF)

모형진단단계의 가장 기본적인 분석 도구는 잔차의 자기상관함수(ACF)이다. 잔차의 ACF는 근본적으로 다른 추정된 ACF와 동일하다. 유일한 차이점이라면 자기상관계수를 계산하기 위하여 실현값의 관찰값인 Z_t를 사용하는 대신에 추정된 모형에서의 잔차인 $\hat{a}_t$를 사용한다는 점이다. 잔차들의 ACF는 앞 장들에서 정의한 바와 같이 다음과 같이 주어진다.

$$r_k(\hat{a}_t) = \frac{\sum_{t=1}^{n-k}(\hat{a}_t - \bar{a})(\hat{a}_{t+k} - \bar{a})}{\sum_{t=1}^{n}(\hat{a}_t - \bar{a})^2}$$

만약 추정된 모형이 적당하다면 백색잡음 a_t는 서로 상관관계가 없어야 한다. 백색잡음이 서로 상관되어 있지 않으면 추정된 잔차 $\hat{a}_t$도 역시 상관관계가 없다. 그러므로 적합하게 식별된 ARIMA모형의 잔차의 ACF는 이상적으로 모두 통계적으로 0인 자기상관계수를 가질 것이다. 그러나 잔차는 추정된 ARIMA계수를 사용하여 추정값을 사용하므로 설사 우리가 좋은 모형을 찾았다 할지라도 표준오차에 기인한 0이 아닌 잔차의 ACF가 있을 수 있을 것이다.

② C-1-2. t-검정

잔차의 ACF가 유의적으로 0과 다른가를 검정하기 위하여 t −검정을 실시한다. 이 t −검정을 실시하기 위한 잔차의 자기상관계수의 표준오차는 다음과 같다.

$$S[r_k(\hat{a}_t)] = \sqrt{1 + 2\sum_{j=1}^{k-j} r_j(\hat{a}_t)^2}\ \sqrt{\frac{1}{n}}$$

잔차 ACF의 추정된 표준오차를 계산한 다음 우리는 각 잔차 ACF에 대하여 다음과 같은 귀무가설 $H_0 : \rho_k(a) = 0$을 검정할 수 있다. 현실에서 우리는 ρ와 a값을 관찰할 수 없기 때문에 $\rho_k(a)$대신에 추정치인 $r_k(\hat{a})$를 사용한다. 얼마나 많은 ACF가 통계적으로 과 다르냐를 검정하기 위해 우리는 일반적으로 t값을 계산하여야 한다.

$$t = \frac{r_k(\hat{a}) - 0}{S[r_k(\hat{a})]}$$

③ C-1-3. 포트맨토 검정(Portmanteau test)

잔차의 독립성을 검정하기 위해서는 먼저 모형의 식별과 추정이 올바르게 되어야 한다. 그러나 일반적으로 ARMA모형의 경우 식별과 추정이 간단하지 않을 때가 많으므로 잔차의 독립성을 검정하는 보다 통계적이고 과학적인 방법이 필요하다. Box와 Pierce(1970)는 잔차들의 독립성을 검정하는 모수적인 방법으로 다음과 같은 포트맨토 검정통계량(Portmanteau statistic)을 제시하였다. 다음과 같은 일련의 귀무가설

$$\mathrm{H0} : \rho 1 = \rho 2 = \cdots = \rho \mathrm{K} = 0$$

즉, H_0 : '모형이 적합하다.' 라는 가설을 검정한다.

제안된 포트맨토 검정통계량은 다음과 같다.

$$Q = n \sum_{k=1}^{K} r_k^2$$

여기서 r_k는 시차 k인 백색잡음의 SACF로서

$$r_k = \frac{\sum_{t=1}^{n} \hat{a}_t \widehat{a_{t-k}}}{\sum_{t=1}^{n} \hat{a}_{\text{ㅓ}}^2}$$

이며 이 Q통계량은 $(\mathrm{K} - p - q)$를 자유도로 갖는 χ^2분포를 따른다고 주장하였다. 그러나 이 통계량은 그 후 많은 학자들에 의해 이 작을 때 분포를 잘 따르지 않는 등 여러 가지 문제가 있음이 밝혀지자 Ljung과 Box(1978)는 수정된 포트맨토우 검정통계량(Modified portmanteau statistic)을 다음과 같이 제안하였고 SAS 등 대부분의 통계패키지는 수정된 포트맨토 검정을 채택하고 있다.

$$Q^* = n(n+2) \sum_{k=1}^{K} \frac{r_k^2}{n-k}$$

이 수정된 포트맨토우 검정통계량 Q^{*}는 일반 ARMA(p,q)모형에 대하여 좋은 모형진단을 내려주며 이 Q^{*} 통계량은 $(\mathrm{K}-p-q)$를 자유도로 가지는 χ^2분포에 잘 따르는 것이 증명되었다. 따라서 Q^{*}값이 χ^2분포표에 나타난 기준값과 비교하여 그보다 크다면 잔차의 ACF는 유의적으로 0과 다르며 추정된 모형의 백색잡음은 자기상관관계가 있다고 할 수 있다. 따라서 모형을 재검토하고 새로운 모형을 고려해야한다.

마. 시계열 예측

시계열모형을 수립하는 중요한 목적 중의 하나는 미래시점의 시계열 값을 예측할 수 있도록 하는 것이다. 또한 이 예측값들이 얼마나 정확한가를 확인하는 것도 중요하다. 시계열 예측에서는 주어진 모형은 모든 모수에 대해 정확하게 알려져 있다고 가정한다. 이것이 실제적으로는 사실이 아니라 할지라도 표본의 크기가 큰 경우에는 모수 대신 추정값을 사용해도 예측에 심각하게 영향을 미치지는 않는다. 따라서 우리는 모수값이 알려져 있다고 가정하고 예측에 대하여 설명한다. 실제로 시계열을 예측할 때는 모수의 추정값을 사용하여 예측값을 구하면 된다.

2. 평활법

관측된 과거 및 현재의 시계열 자료에 대해서 이동평균(Moving Average), 가중이동평균(Weighted Moving Average) 등의 방법을 사용하여 불규칙변동을 평활해서 시계열의 미래의 값을 예측하는 방법이 평활법에 의한 시계열의 분석방법이다. 이러한 평활법에는 이동평균 평활법과 가중평균 평활법인 지수평활법 등이 있으며, 시계열의 유형에 따라서 다른 평활방법이 적용된다. 즉, 관측된 시계열들의 평균수준이 시간대에 관계없이 거의 변하지 않는 불규칙변동만을 포함하는 시계열인 수평적 계열(horizontal series)에 대해서는 단순이동평균법과 단순지수평활법이, 추세성을 갖는 시계열에 대해서는 선형이동평균법과 선형 및 이차지수평활법이, 그리고 추세 및 계절성을 갖는 시계열에 대해서는 계절지수평활법이 적용된다.

가. 이동평균 평활법

① 단순이동평균법(Simple Moving Average Method)

단순이동평균에 의한 시계열 평활방법은 수평적 계열(horizontal series), 즉 시간의 경과에 따라 평균수준이 변하지 않는 계열에 적용되며, 가장 최근의 m-기간동안의 자료들의 단순평균을 다음

기간의 예측값으로 추정하는 방법이다. 이 때 m은 분석자에 의하여 사전적으로 결정되어야 하고 또 시계열의 새로운 관측값이 추가되면 단순이동평균은 달라진다. Z_t를 시점 t에서의 실제값이라고 하고 F_{n+1}을 시점 n에서 추정한 시점 $n+1$의 예측값이라고 하면 단순이동평균법으로 F_{n+1}을 구하는 계산식은 다음과 같이 나타난다.

단순이동평균법에 의한 예측

$$F_{n+1} = \frac{1}{m}(Z_n + Z_{n-1} + \ldots + Z_{n-m+1})$$

$$= \frac{1}{m}\sum_{t=n-m+1}^{n} Z_i \ (= MA_n)$$

혹은, $F_{n+1} = F_n + \frac{1}{m}(Z_n - Z_{n-m})$

즉, 현재 시점이 n인 경우에 다음 시점의 예측값은 최근의 M-기간 동안의 이동평균(MA_n)으로 예측하는 방법이 된다. 여기서 만약 $m = n$ 이면 F_{n+1}은 $n-$기간 전체의 평균이 되고, $m = 1$이면 $F_{n+1} = Z_n$이 된다. 이 때, 이동평균 기간 m은 관측기간 내에서의 한 기간후의 예측오차(one-step-ahead forecast errors)의 평균제곱합을 최소로 하는 값으로 결정할 수 있다.

$$MSE(m) = \frac{1}{n}\sum_{1}^{n}(Z_t - F_t)^2$$

② 선형이동평균법(Linear Moving Average)

관측된 시계열이 선형 추세성을 갖는 경우에 적합시킬 수 있는 평활법인 선형이동평균 방법은 이중이동평균(double moving average : MA')을 이용하여 현재 시점이 n인 경우에 $n+l$ 시점의 값을 다음과 같이 예측하는 방법이다.

선형이동평균법에 의한 예측

$$F_{n+l} = a_n + b_n \cdot l$$

$$= (2MA_n - MA_n') + \frac{2}{m-1}(MA_n - MA_n') \cdot l$$

여기서 MA_n는 이동평균을, MA_n' 은 이중이동평균을 나타낸다. 이중이동평균이란 원계열의 이동평균을 다시 이동평균한 것으로 다음과 같이 정의된다.

이중이동평균

$$MA_n' = \frac{1}{m}(MA_n + MA_{n-1} + \cdots + MA_{n-m+1})$$

나. 지수평활법

ㄱ. 단순지수평활법(Simple Exponential Smoothing Method)

단순이동평균법과 마찬가지로 단순지수평활법도 수평적 시계열 자료에 적용되는 방법이지만 단순이동평균법과는 달리 최근의 자료들에 대해 더 많은 가중치를 부여하는 방법이다. 단순이동평균법에서는 이동평균을 계산하기 위해 최근의 m-기간의 관측값만 사용하였으며 또한 이들에 대하여 동일한 가중치를 부여하였다. 단순지수평활법은 단순이동평균법의 이러한 단점을 보안한 방법으로 예측식은 다음과 같이 된다.

단순지수평활법에 의한 예측

$$\begin{aligned} F_{n+1} &= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + \cdots \\ &= SM_n \end{aligned}$$

여기서 α는 평활상수(smoothing constant)라고 하고 0과 1 사이의 값을 갖으며, SM_n은 시점 n에서의 평활값(smoothed value)이라고 한다. 그러므로 단순지수평활법이란 시계열 관측값들에 대한 가중평균인 평활값으로 미래의 관측값을 예측하는 방법이다. 이는 아래와 같은 방법으로도 나타낼 수 있다.

단순지수평활법에 의한 예측

$$F_{n+1} = \alpha Z_n + (1-\alpha)F_{(n-1)+1}$$

혹은, $\quad SM_n = \alpha Z_n + (1-\alpha)SM_{n-1}$

여기서 $\mathrm{F}_{(n-1)+1}$는 시점 $n-1$에서 예측한 시점 n의 예측값을 의미하고, 시점 n에서의 예측오차는 $\mathrm{E}_n = \mathrm{Z}_n - \mathrm{F}_n$이 된다. 따라서 α, Z_n, 그리고 F_n이 주어지면 F_{n+1}이 계산될 수 있다. 지수평활법에 의한 예측에서는 관측값들에 대한 가중치의 역할을 하는 평활상수 α의 결정이 매우 중요한 문제가 된다. 일반적으로 시계열자료가 안정적이고 변동이완만한 자료에서는 작은 값의 α가, 그리고 변동이 심한 자료에서는 큰 값의 α가 적절하게 된다. 실제 문제에서 α는 0.01과 0.30 사이의 값을 주로 사용하고 있다. 적절한 α의 결정은 단순이동평균법에서의 의 결정방법과 마찬가지로 다음과 같은 예측오차의 제곱합을 최소로 하는 α로 결정하면 된다.

$$MSE(\alpha) = \frac{1}{n}\sum_{1}^{n}(Z_t - F_t)^2$$

ㄴ. 선형지수평활법(Linear Exponential Smoothing Method)

선형지수평활법은 추세성을 갖는 시계열의 예측을 위한 지수평활법이다. 앞 절에서 설명된 것처럼 선형이동평균법은 단순이동평균법이 갖고 있는 단점들을 거의 그대로 갖고 있다. 즉, 선형이동평균법에서는 한 시점 이후의 미래의 값을 예측하기 위하여 최근의 2m-1개의 관측값만을 사용하고 또 2m-1개의 각 관측값들은 미래의 값에 대하여 동일한 크기의 영향을 미친다는 전제를 하게 된다. 선형지수평활법은 이와 같은 선형이동평균법의 문제점들을 개선시킨 방법으로 브라운(Brown,1963)과 홀트(Hollt, 1957)의 방법이 있다.

① 브라운의 선형지수평활법(Brown's Linear Exponential Smoothing Method)

브라운의 선형지수평활법은 지수평활값과 이중지수평활값(double exponentialsmoothed value)을 이용하여 미래의 값을 예측하는 방법이다.

브라운의 선형지수평활법

$$F_{n+l} = a_n + b_n \cdot l = 2SM_n - SM_n^{'} + \frac{\alpha}{1-\alpha}(SM_n - SM_n^{'}) \cdot l$$

여기서 F_{n+l}은 현재시점 n에서 시점 $n+l$에 대한 예측값을 나타내고, α는 평활상수이고 SM_n과 $SM_{n'}$ 는 각각 다음과 같이 정의되는 지수평활값과 이중지수평활값이다.

지수평활값과 이중지수평활값

$$SM_n = \alpha Z_n + (1-\alpha)SM_{n-1}$$

$$SM_n^{'} = \alpha SM_n + (1-\alpha)SM_{n-1}^{'}$$

② 홀트의 선형지수평활법(Holt's two-parameter linear xponential smoothing method)

홀트의 두 모수 선형지수평활법은 이중지수평활값을 사용하지 않는다는 것을 제외하고는 브라운의 방법과 유사하다. 홀트의 모형에서는 추세를 먼저 추정하여 그것을 이용하여 예측하는 방법으로 다음과 같이 3개의 방정식과 2개의 평활상수(0 < α,β < 1)가 포함된다.

홀트의 선형지수평활법

$$F_{n+l} = SM_n + T_n \cdot l$$
$$SM_n = \alpha Z_n + (1-\alpha)(SM_{n-1} + T_{n-1})$$
$$T_n = \beta(SM_n - SM_{n-1}) + (1-\alpha)T_{n-1}$$

여기서 SM_n은 자료의 평활(smoothing of data), T_n은 추세의 평활(smoothingof trend)을 나타낸다. 그런데 홀트의 방법에서는 2개의 상수를 사전에 결정해야 하므로 다소 복잡한 계산과정을 거쳐야 한다. 물론 이 경우에도 상수 α, β는 평균제곱예측오차를 최소로 하는 값으로 결정할 수 있다. 실제 계산에서의 초기값 SM_1,T_1은 다음과 같은 값으로 사용할 수 있다.

$$SM_1 = Z_1, \quad T_1 = [(Z_2 - Z_1) + (Z_4 - Z_3)]/2$$

③ 이차지수평활법(Quadratic Exponential Smoothing Method)

브라운(Brown,1963)에 의하여 제안된 이차지수평활법은 시계열이 이차곡선 형태의 추세를 갖는 경우에 적용하는 예측방법이다. 이 경우에 이차추세는 삼중지수평활값 (triple exponential smoothed value: SM''_n)을 사용하여 평활한다. 브라운의 이차지수평활 예측식은 다음과 같이 한 개의 평활상수 α와 4개의 방정식으로 구성된다.

이차지수평활 예측식

$$F_{n+l} = a_n + b_n \cdot l + \frac{1}{2} c_n \cdot l^2$$
$$a_n = 3SM_n - 3SM'_n + SM''_{n-1}$$
$$b_n = \frac{\alpha}{2(1-\alpha)^2}[(6-5\alpha)SM_n - (10-8\alpha)SM'_n + (4-3\alpha)SM''_n]$$
$$c_n = \frac{\alpha^2}{(1-\alpha)^2}(SM_n - 2SM'_n + SM''_n)$$

여기서, $SM_{n'} = \alpha Z_n + (1-\alpha)SM_{n-1'}$, $SM_{n''} = \alpha Z_n + (1-\alpha)SM_{n-1''}$ 이다. 이차평활을 위한 방정식들은 단순평활이나 선형평활 보다 훨씬 더 복잡하지만 그 원리는 모두 동일하다.

ㄷ. 계절지수평활법

대부분의 월별 혹은 분기별 자료들은 계절변동을 포함하게 되므로 이러한 계열에대한 예측을 할 때는 계절변동을 고려할 수 있는 계절지수평활법을 사용할 수 있다. 그런데 계절변동은 가법적 계절변동(additive seasonal variation)과 승법적 계절변동(multiplicative seasonal variation)의 두 형태가 있기 때문에 계절변동의 형태에 따라서 서로 다른 예측방법이 적용되어야 한다. 가법적 계절변동의 특성은 시계열의 계절적 진폭이 시간의 흐름에 따라 일정하지만 승법적 계절변동은 시계열의 진폭이 점차적으로 증가 혹은 감소하게 된다.

① 승법적 계절지수평활법(Multiplicative Seasonal Exponential Smoothing Method)

윈터스(Winters, 1960)의 승법적 계절지수평활에 의한 예측방법은 홀트의 선형지수평활법을 확장시킨 방법으로 관측된 시계열이 선형추세성과 승법적 계절변동을 나타낼 때 적용하는 방법이다. 윈터스의 승법적 계절지수평활법은 다음과 같이 시계열 패턴의 세 가지 성분인 수평성, 추세성, 그리고 계절성을 평활하는 세 개의 방정식과 예측식으로 구성된다. 즉, 현재 시점이 n인 경우에 l 시점 후의 예측값 F_{n+l} 은 다음과 같다.

승법적 계절지수평활에 의한 예측

$$F_{n+l} = (a_n + b_n l) S_{n+l-L} \qquad l = 1,2,\ldots,L$$

$$a_n = \alpha \frac{Z_n}{S_{n-L}} + (1-\alpha)(a_{n-1} + b_{n-1}) \quad : \text{수평성분}$$

$$b_n = \beta(a_n - a_{n-1}) + (1-\beta) b_{n-1} \quad : \text{추세성분}$$

$$S_n = \gamma \frac{Z_n}{a_n} + (1-\gamma) S_{n-L} \quad : \text{계절성분}$$

여기서, L 은 계절성의 길이, S_n은 계절인자, α, β, γ는 평활상수를 나타낸다.

② 가법적 계절지수평활법(Additive Seasonal Exponential Smoothing Method)

윈터스(Winters, 1960)의 가법적 계절지수평활에 의한 예측방법은 선형추세성과 가법계절변동을 갖는 시계열에 적용하는 방법이다. 이 때 예측식은 시계열 패턴의세 가지 성분인 수평성, 추세성, 그리고 계절성을 평활한 세 성분의 합으로 나타난다. 즉, 현재 시점이 n인 경우에 l시점 후의 예측값은 다음과 같이 나타난다.

가법적 계절지수평활에 의한 예측

$$F_{n+l} = a_n + b_n l + S_{n+l-L} \qquad l = 1,2,\ldots,L$$

$$a_n = \alpha(Z_n - S_{n-L}) + (1-\alpha)(a_{n-1} + b_{n-1}) \quad : \text{수평성분}$$

$$b_n = \beta(a_n - a_{n-1}) + (1-\beta)b_{n-1} \quad : \text{추세성분}$$

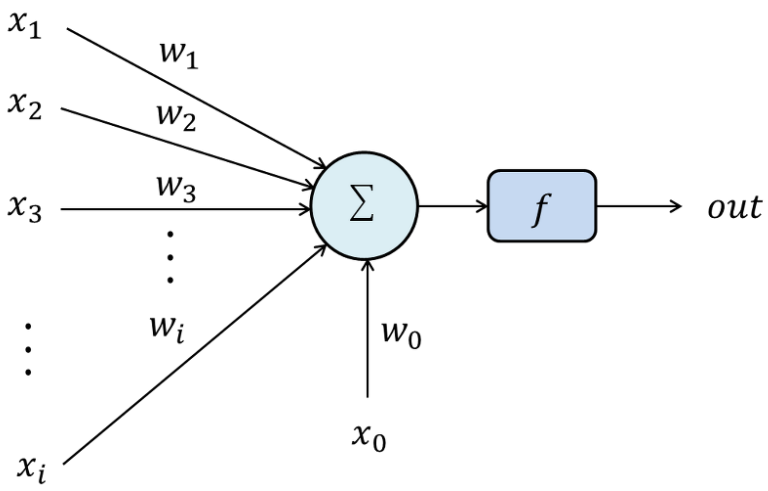
$$S_n = \gamma(Z_n - a_n) + (1-\gamma)S_{n-L} \quad : \text{계절성분}$$

여기서, L 은 계절성의 길이, S $_n$은 계절인자, α, β, γ는 평활상수를 나타낸다.

3. 인공신경망 ANN(artificial neural network)

가. 단층 퍼셉트론(Single Layer Perceptron)

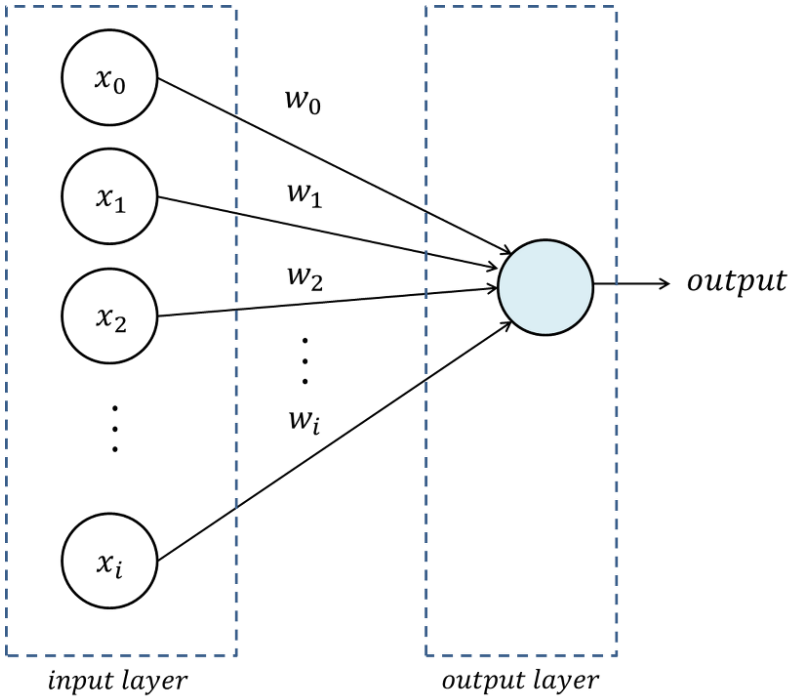
퍼셉트론은 1957년 코넬 항공 연구소(Cornell Aeronautical Lab)의 프랑크 로젠블라트(Frank Rosenblatt)에 의해 고안된 인공신경망이다. 로젠블라트에 의해 제안된 것은 가장 간단한 형태의 단층 퍼셉트론(single-layer perceptron)으로 입력 벡터를 두 부류로 구분하는 선형분류기이다. 간단하게 자주 사용하는 용어들에 대한 정의를 아래에서 내린다. 임계치(threshold)는 어떠한 값이 활성화되기 위한 최소값을 임계치라고 한다. 바이어스(bias)는 선형 경계의 절편을 나타내는 값으로써, 직선의 경우는 y절편을 나타낸다. 가중치(weight)는 퍼셉트론의 학습 목표는 학습 벡터를 두 부류로 선형 분류하기 위한 선형 경계를 찾는 것이다. 가중치는 이러한 선형 경계의 방향성 또는 형태를 나타내는 값이다. Net 값은 입력값과 가중치의 곱을 모두 합한 값으로써, 기하학적으로 해석하면 선형 경계의 방정식과 같다. 활성함수(Activation Function)는 뉴런에서 계산된 net값이 임계치보다 크면 1을 출력하고, 임계치보다 작은 경우에는 0을 출력하는 함수이다. 이 정의는 단층 퍼셉트론에서만 유효하며, 다층 퍼셉트론에서는 다른 형태의 활성함수를 이용한다. 뉴런(neuron)은 인공신경망을 구성하는 가장 작은 요소로써, net값이 임계치보다 크면 활성화되면서 1을 출력하고, 반대의 경우에는 비활성화되면서 0을 출력한다.



위의 그림은 뉴런의 구조를 나타낸다. 뉴런에서 x는 입력 벡터의 값을 나타내고 w는 가중치를 나타낸다. 바이어스 입력값은 x_0, 바이어스 기울기는 w_0로 표기했으며, f는 활성함수를 나타낸다.

ㄱ. 알고리즘 구조

단층 퍼셉트론은 입력층(input layer)과 출력층(output layer)으로 구성된다. 입력층은 학습 벡터 또는 입력 벡터가 입력되는 계층으로써, 입력된 데이터는 출력층 뉴런으로 전달되어 활성함수에 따라 값이 출력된다. 출력층은 퍼셉트론 설계 시 임의의 n개의 뉴런으로 구성할 수 있으며, 아래의 그림은 1개의 뉴런으로 구성된 단층 퍼셉트론을 나타낸다.



① 단층 퍼셉트론의 학습 알고리즘.

1) 가중치와 바이어스 가중치를 -0.5와 0.5 사이의 임의의 값으로, 바이어스 입력값을 임의의 값으로 초기화한다.

2) 하나의 학습 벡터에 대한 출력층 뉴런의 net값을 계산한다.

3) 활성함수를 통해 계산된 net값으로부터 뉴런의 실제 출력값을 계산한다.

4-1) 뉴런의 출력값과 목표값의 차이가 허용 오차보다 작으면 5)로 이동한다.

4-2) 뉴런의 출력값과 목표값의 차이가 허용 오차보다 크면 학습을 진행한다.

5-1) 현재 학습 벡터가 마지막 학습 벡터가 아니면, 현재 학습 벡터를 다음 학습 벡터로 설정하고 2)로 이동하여 반복한다.

5-2-1) 현재 학습 벡터가 마지막 학습 벡터이고, 모든 학습 벡터에 대해 출력값과 목표값이 허용 오차보다 작으면 알고리즘을 종료한다.

5-2-2) 현재 학습 벡터가 마지막 학습 벡터이지만 출력값과 목표값이 허용 오차보다 큰 학습 벡터가 존재하면, 현재 학습 벡터를 처음 학습 벡터로 설정하고 2)로 이동하여 반복한다.

ㄴ. 연산 정의

① 뉴런의 net값 계산

뉴런의 net값은 아래와 같이 계산된다. 아래의 식에서 N은 입력 벡터의 크기를 나타낸다.

$$net = \sum_{i}^{N} w_i x_i + w_0 x_0$$

② 활성함수의 정의

활성함수는 net값이 임계치보다 크면 뉴런의 출력값을 활성화하고, 그렇지 않으면 뉴런의 출력값을 비활성화하는 함수이다. 퍼셉트론에서 사용하는 가장 기본적인 활성함수는 계단 함수(step function)를 이용한다. 퍼셉트론에서 이용하는 계단 함수의 정의는 아래의 식과 같다. 활성함수로 계단함수를 이용할 때는 뉴런의 출력값은 0과 1만을 갖기 때문에 목표값과 출력값의 차이라는 개념보다는 목표값과 출력값의 일치, 불일치라는 개념을 이용한다.

$$f(net) = \begin{cases} 1, & net \geq threshold \\ 0, & net < threshold \end{cases}$$

③ 학습 연산(Learning rule) 정의

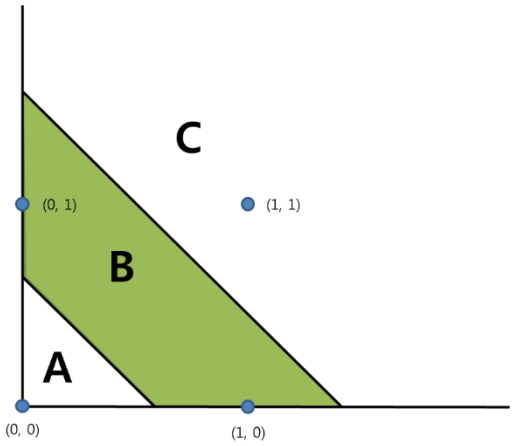
출력층 뉴런의 출력값과 목표값의 차이가 허용오차보다 크면 출력층 뉴런의 가중치를 조정해야한다. 가중치를 조정하는 식은 아래와 같다..

$$w_i = w_i + \eta x_i (t - f(net))$$

$\eta = learning\ rate$
$t = target\ value$

나. 다층 퍼셉트론 (Multi-layer Perceptron)

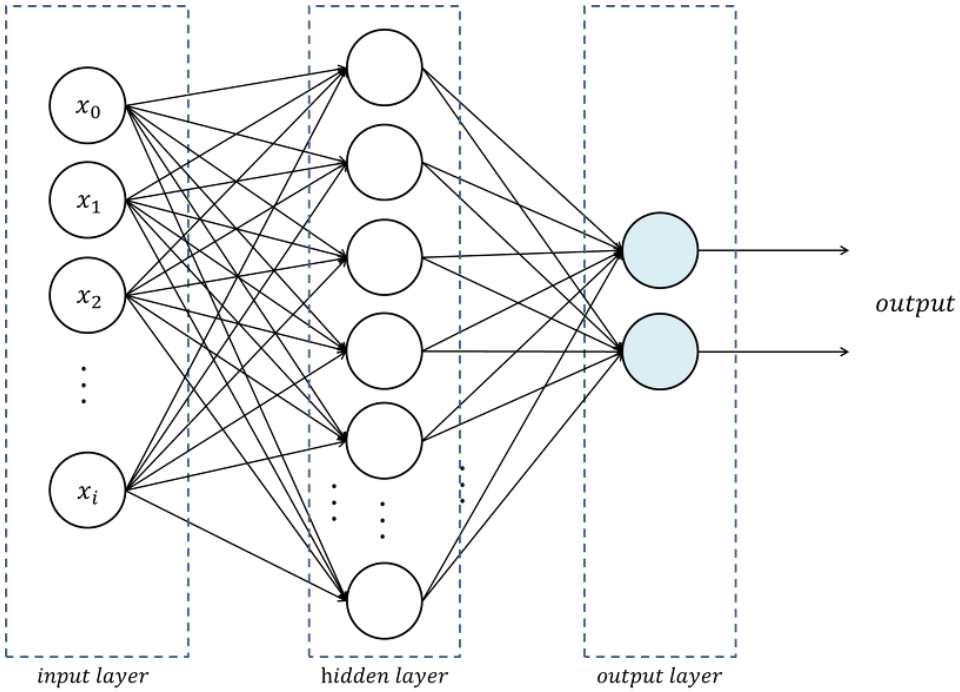
다층 퍼셉트론은 단층 퍼셉트론이 해결할 수 없는 문제를 해결할 수 있다. 단층 퍼셉트론으로 해결할 수 없는 대표적인 문제로는 "Minsky M. L. and Papert S. A. 1969. Perceptrons"에서 소개된 XOR (exclusive OR) 문제가 있다. XOR 이라는 것은 ture 값이 하나만 있을 때 ture가 되는 논리연산자이다. XOR 연산의 정의를 기하학적으로 표현하면 아래의 그림과 같다.



XOR 문제를 기하학적으로 해석하면 A와 C 영역은 0, B 영역은 1의 값을 가져야 한다. 그러나 2차원 평면상에서 직선 하나로 이러한 분리를 하는 것은 불가능하다. 따라서, 선형 분리만 가능한 단층 퍼셉트론으로는 XOR 문제를 해결할 수 없다는 것이 자명하다. 이러한 문제점을 해결하기 위해 입력층과 출력층 사이에 은닉층(hidden layer)을 추가하고, 역전파 (backpropagation) 알고리즘을 이용하여 학습을 하는 다층 퍼셉트론이 고안되었다.

ㄱ. 은닉층(Hidden Layer)

다층 퍼셉트론에서 은닉층은 서포트 벡터 머신 (support vector machine)에서의 커널 함수 (kernel function)와 비슷하다. 은닉층은 입력층으로 표현되는 입력 벡터가 은닉층으로 표현되는 새로운 벡터로 변환되는 중간 계층이다. 아래의 그림은 다층 퍼셉트론의 구조이다.



은닉층은 선형 분리가 불가능했던 데이터를 새로운 공간으로 사상 (morphism 또는 mapping)함으로써 선형 분리가 불가능했던 데이터를 선형 분리가 가능하도록 변환한다.

ㄴ. 다층 퍼셉트론의 동작

다층 퍼셉트론은 단층 퍼셉트론에 은닉층을 추가한 형태로써 입력층에서 전달되는 출력값이 은닉층으로 전달되고, 은닉층의 출력값이 출력층으로 전달되는 구조이다. 하나의 학습 벡터는 아래의 그림과 같이 <입력값, 목표값>의 쌍으로 구성된다. 단층 퍼셉트론처럼 다층 퍼셉트론에서도 마지막에 계산되는 출력층의 출력값과 학습 벡터의 목표값을 비교하여 출력값과 목표값의 차이가 허용 오차보다 크면 가중치를 학습 규칙에 따라 조정한다.

$$<x_1, x_2, x_3, \ldots, x_n, t_1, t_2, t_3, \ldots, t_m>$$

input values　　target values

I. 다층 퍼셉트론의 학습 알고리즘.

1) 가중치와 바이어스 가중치를 -0.5와 0.5 사이의 임의의 값으로, 바이어스 입력값을 -1 또는 1로 초기화한다.

2) 하나의 학습 벡터에 대한 은닉층 뉴런의 net값을 계산하고, 활성함수를 통해 실제 출력값을 연산한다.

3) 은닉층의 출력 벡터에 대한 출력층 뉴런의 net값을 계산하고, 활성함수를 통해 실제 출력값을 연산한다.

4) 학습 규칙에 따라 입력층-은닉층, 은닉층-출력층 가중치를 수정한다.

5-1) 현재 학습 벡터가 마지막 학습 벡터가 아니면, 현재 학습 벡터를 다음 학습 벡터로 설정하고 2)로 이동하여 반복한다.

5-2-1) 현재 학습 벡터가 마지막 학습 벡터이고, 오차의 총합이 허용 오차보다 작으면 학습을 종료한다.

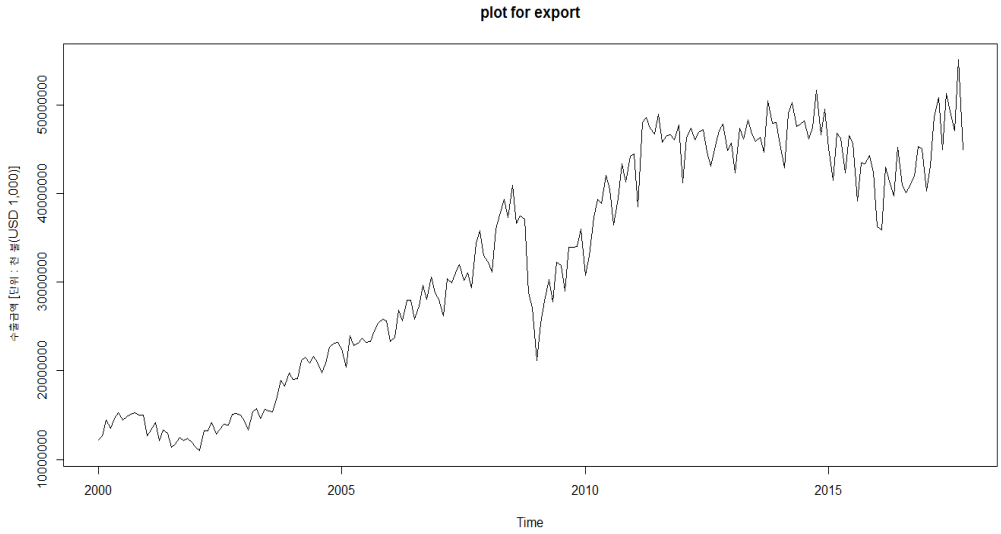
5-2-2) 현재 학습 벡터가 마지막 학습 벡터이지만 오차의 총합이 허용 오차보다 크면, 학습 벡터를 처음 학습 벡터로 설정하고 2)로 이동하여 학습을 반복한다.

II. 실증분석

1. ARIMA 모형

```
#####################
#####ARIMA model#####
#####################
##ready
#install.packages("forecast")
library(forecast)
library(lmtest)
library(tseries)

##importing the data
export <- read.csv("D:/금융통계학/export.csv", header=F)
export <- ts(export, start=c(2000,1), end=c(2017,10),
             frequency = 12)

#options("scipen" = 100)

plot.ts(export, ylab="수출금액 [단위 : 천 불(USD 1,000)]",
        main="plot for export")
```

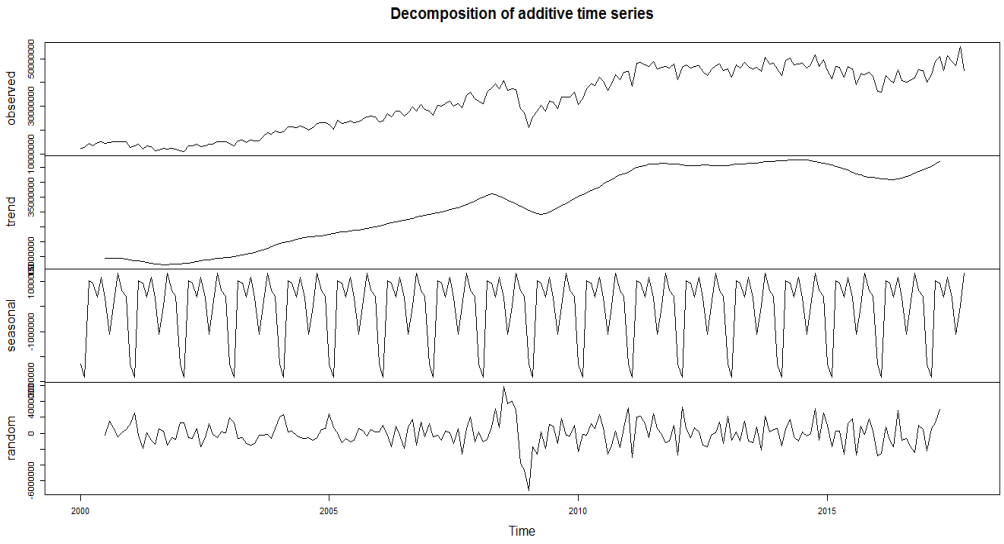
먼저 2000년 1월부터 2017년 10월까지의 한국의 수출금액 시계열 데이터 자료를 불러온다. 시계열 그림을 그려서 확인해보면 계절성과 추세가 존재하는 것을 확인할 수 있다.



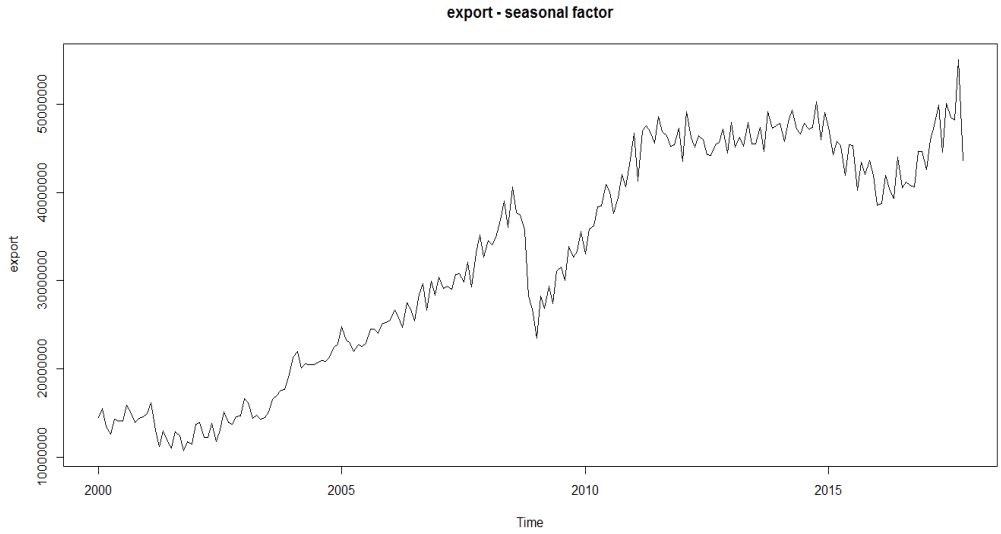
[그림 1.1] 한국의 수출금액 시계열 그림

```r
##데이터 분해 - trend, seasonal, random 데이터 추세 확인
export_comp <- decompose(export)
plot(export_comp)


##시계열 데이터에서 계절성 요인 제거
export_adjusted <- export - export_comp$seasonal
plot.ts(export_adjusted, main = "export - seasonal factor")
```

위의 코드를 사용하여 데이터를 분해한 그림은 아래 [그림1.2]에서와 같이 계절성과 추세가 존재함을 확인할 수 있다. 분해한 계절성 요인을 제거해서 export_adjusted에 적합 시켜 플롯을 그려보면 [그림 1.3]이 된다.



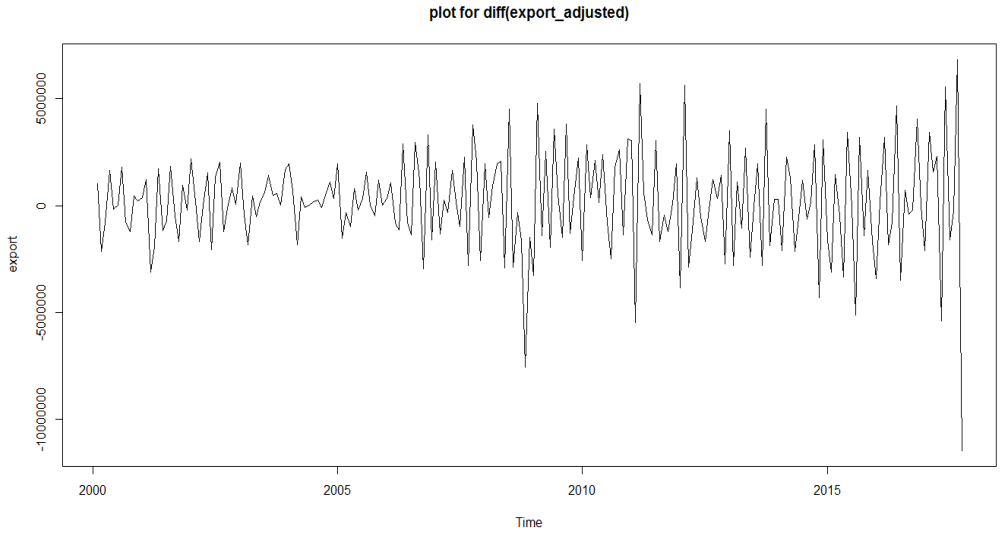
[그림 1.2] 분해한 시계열 그림



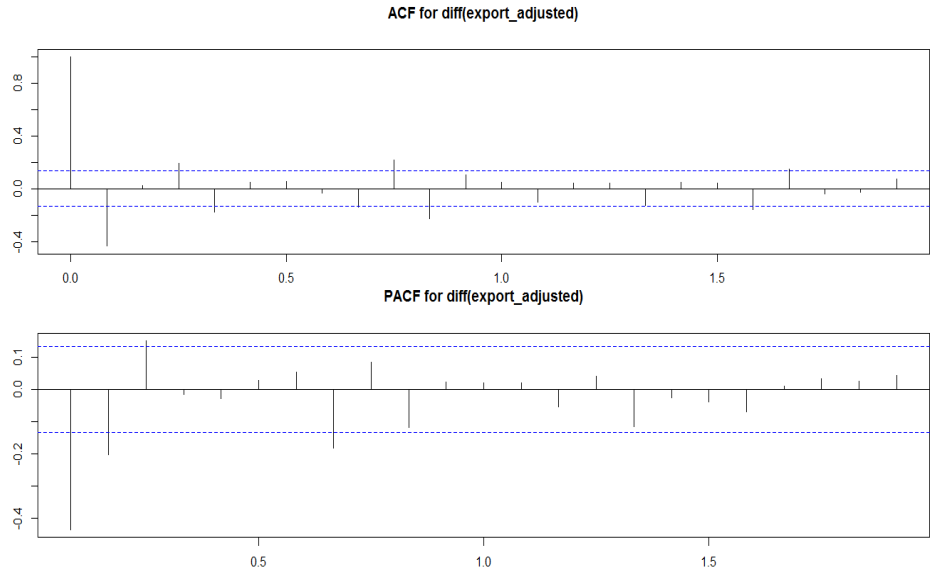
[그림 1.3] 계절성을 제거한 시계열 그림

```
##차분을 통해 정상성 확인
dexport <- diff(export_adjusted)
par(mfrow=c(1,1))
plot(dexport, main="plot for diff(export_adjusted)")

#par(mar=c(2,3,3,2))

par(mfrow=c(2,1))
acf(dexport, main="ACF for diff(export_adjusted)")
pacf(dexport, main="PACF for diff(export_adjusted)")
```

계절성을 제거한 후 차분을 통해 추세를 제거한 후 그린 시계열 그림은 [그림 1.4]와 같다. 그 후 계절성과 추세를 제거한 데이터 dexport의 ACF와 PACF를 확인해본다.



[그림 1.4] 차분한 시계열 그림



[그림 1.5] 변환한 데이터의 ACF, PACF

[그림 1.5]를 보면 ACF는 2차 이후 절단 혹은 지수적 감소를 보이고 PACF는 3차 이후 절단 혹은 지수적 감소를 보임을 확인할 수 있다. 이에 따라 다음과 같은 ARIMA(p,d,q) 모델에 대해서 고려해볼 수 있다.

```r
##Box-Jenkins
#identification
arima(dexport, order=c(2,0,0))
arima(dexport, order=c(0,0,3))
arima(dexport, order=c(2,0,3))
arima(dexport, order=c(0,0,2))
arima(dexport, order=c(1,0,2))
arima(dexport, order=c(2,0,2))
```

각각의 arima를 적합 시켜보면 AIC값들을 얻을 수 있다. 아래의 [표 1.1]에서 확인해 보면 ARIMA(2,0,3)의 AIC값이 가장 작고, ARIMA(2,0,2)의 AIC값이 2밖에 차이 나지 않음을 확인할 수 있다. 모수축약의 원칙에 따라 ARIMA(2,0,2)도 충분히 모형을 설명할 수 있을 것으로 기대되므로 이에 따라 ARIMA(2,0,3)을 fit_a1, ARIMA(2,0,2)를 fit_a2로 적합 시켜본다.

Model	AIC값
arima(dexport, order=c(2,0,0))	6796.09
arima(dexport, order=c(0,0,3))	6797.88
arima(dexport, order=c(2,0,3))	6790.52
arima(dexport, order=c(0,0,2))	6796.15
arima(dexport, order=c(1,0,2))	6798.03
arima(dexport, order=c(2,0,2))	6792.97

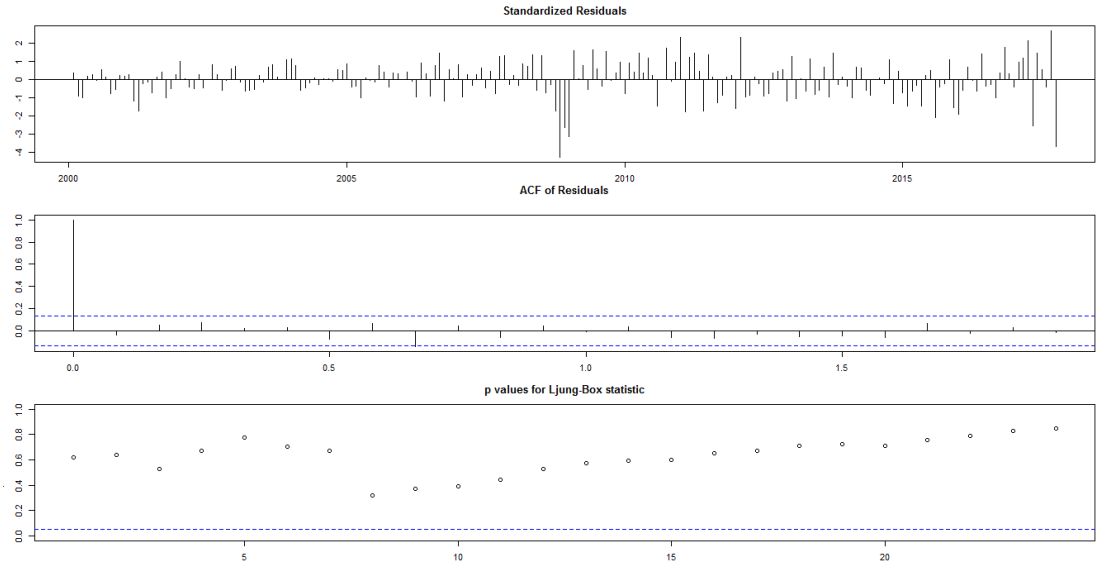
[표 1.1] ARIMA 모형에 따른 AIC 값

```r
##estimation
fit_a1 <- arima(dexport, order=c(2,0,3))
fit_a2 <- arima(dexport, order=c(2,0,2))

summary(fit_a1)
coeftest(fit_a1)
summary(fit_a2)
coeftest(fit_a2)
```

fit_a1은 RMSE가 1953745, MAE가 1466841이고 fit_a2는 RMSE는 1978072, MAE는 1505078로 fit_a1이 더 적합함을 확인할 수 있다.

```r
##validtion 시계열모형에 대한 적합도 검정
tsdiag(fit_a1, gof.lag=24) #24시점까지
```

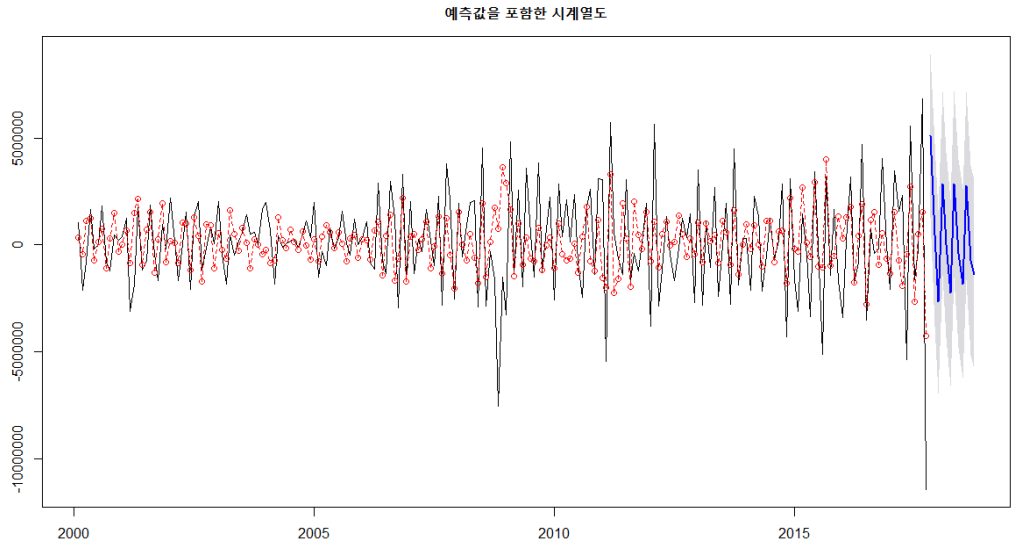
이제 시계열 모형에 대한 적합도를 검정해보면, [그림 1.6]에서 마지막 p-vavlue가 모두 유의하지 않아서 모형이 적합함을 알 수 있다.



[그림 1.6] 시계열 모형에대한 적합도 검정 그림

```
##forecasting
fore_a1 <- forecast(fit_a1, h=12, level=0.95) #95% 신뢰구간

par(mfrow=c(1,1))
plot(fore_a1, main="예측값을 포함한 시계열도")

#모델이 샘플데이터를 얼마나 잘 설명하는지
lines(fitted(fit_a1),col="red", lty=2, type="o", lwd=1, pch=1)
```

[그림 1.7]은 예측값을 포함한 시계열 플롯과 그 위에 모델이 샘플데이터를 설명해주는 선을 그린 그림이다.



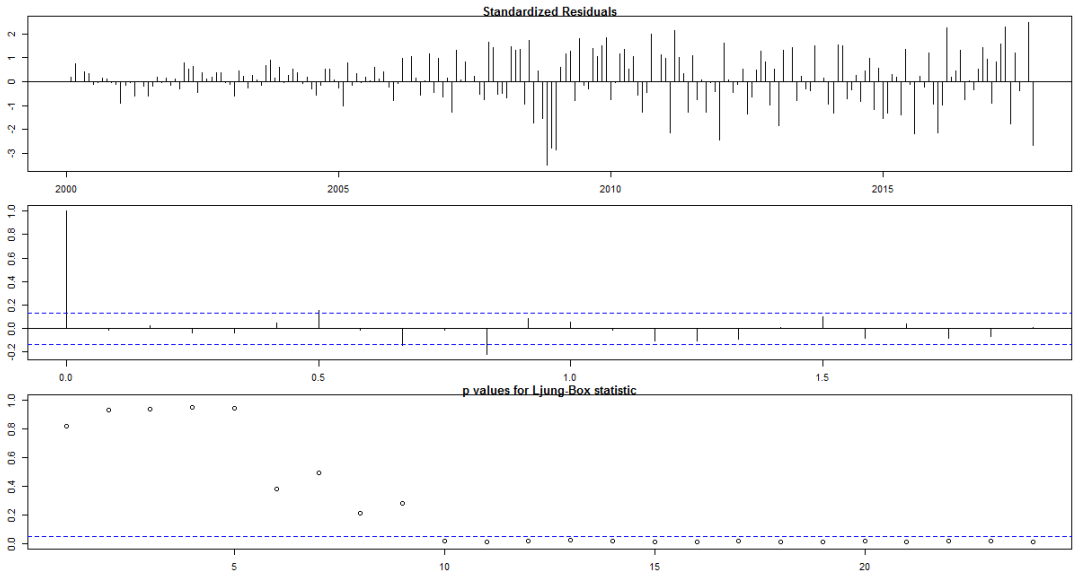
[그림 1.7] 예측값을 포함한 시계열도

다른 방법으로 auto.arima를 통해 자동으로 시계열모형을 찾아본다. 아래의 출력값을 보면 arima모형이 ARIMA(1,1,3)(0,0,2)[12]로 예측 됐음을 확인할 수 있다.

```
> auto.arima(export)
Series: export
ARIMA(1,1,3)(0,0,2)[12]

Coefficients:
          ar1     ma1      ma2     ma3    sma1    sma2
      -0.8223  0.3918  -0.3968  0.1689  0.3569  0.1546
s.e.   0.0552  0.0831   0.0670  0.0694  0.0853  0.0753

sigma^2 estimated as 5389591108830:  log likelihood=-3422
.47
AIC=6858.94   AICc=6859.49   BIC=6882.47
```

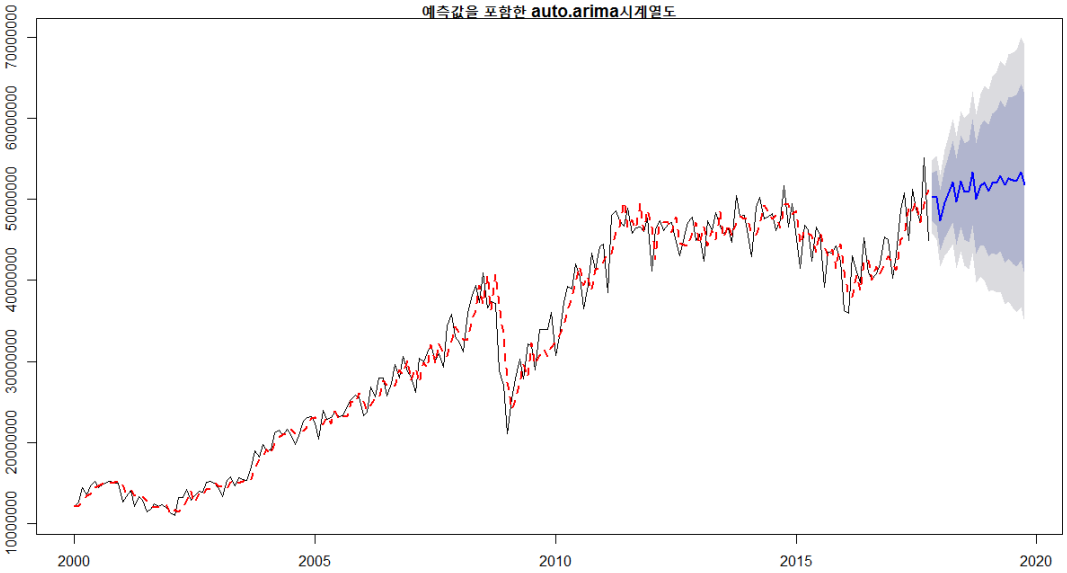
```
##estimation
fit_auto <- arima(export, order=c(1,1,3),
                  seasonal=list(order=c(0,0,2), period=12))
summary(fit_auto)
coeftest(fit_auto) #모수추정
```

위의 코드를 통해서 fit_auto의 모수 추정값들은 모두 유의하고 RMSE는 2283264, MAE는 1698720, AIC는 6858.94가 나온 것을 확인할 수 있었다.



[그림 1.8] auto.arima로 적합한 시계열 그림의 적합도 검정 그림

하지만 p-value값이 2010년 이후 유의해서 모형이 잘 설명해주지 못한다는 사실을 확인할 수 있었다. 이를 그림으로 그려보면 [그림 1.9]와 같다.



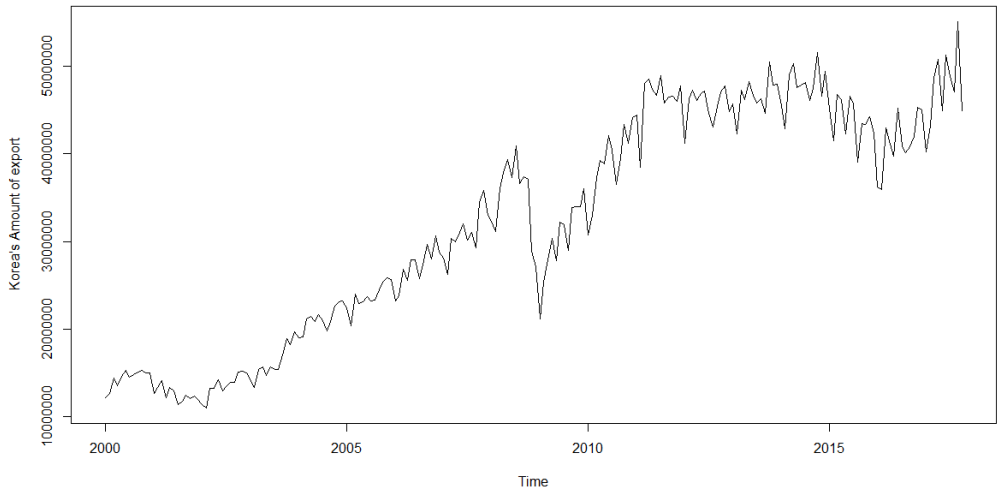
[그림 1.9] 예측값을 포함한 auto.arima의 시계열도

즉, auto.arima로 적합한 모형보다 계절성을 제거하고 차분한 arima모형이 더 적합함을 알 수 있다.

2. 평활법

```
######################
#######평활법########
######################
#가법계절지수 F=S_계절+T_추세+I_불규칙
#승법계절지수 F=STI

##계절지수평활법
plot(export, xlab="Time", ylab="Korea's Amount of export")
```

ARIMA 모형에서 불러온 export 데이터를 시계열그림으로 그리면 [그림 2.1]과 같다. 시계열 그림에서 주기, 선형적 추세, 폭이 커지는 것을 보아 승법 계절지수임을 추측할 수 있다.



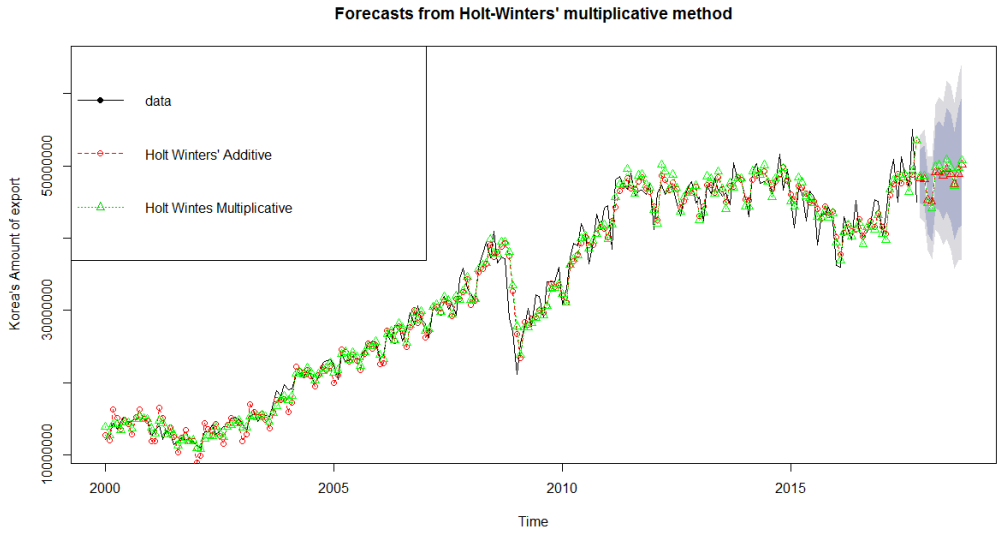
[그림 2.1] 한국의 수출금액 시계열 그림

```
fit1 <- hw(export, seasonal="additive", h=12) #가법
fit2 <- hw(export, seasonal="multiplicative", h=12) #승법
```

fit1은 가법계절지수를 적합 시키고 fit2는 승법계절지수를 적합 시켰다. [그림 2.2]를 보면 fit2의 시계열 예측도에 fit1, fit2의 적합결과를 덧그려보면 fit2가 기존의 데이터와 조금 더 적합함을 알 수 있다. [표 2.1]에서 승법계절지수를 사용한 fit2가 SSE, RMSE, MAE값이 더 낮으므로 fit2가 fit1보다 좋은 모형임을 확인할 수 있다.

	가법계절지수(fit1)	승법계절지수(fit2)
SSE	917822320278542	0.7530467
RMSE	2070963	2000916
MAE	1552079	1434172

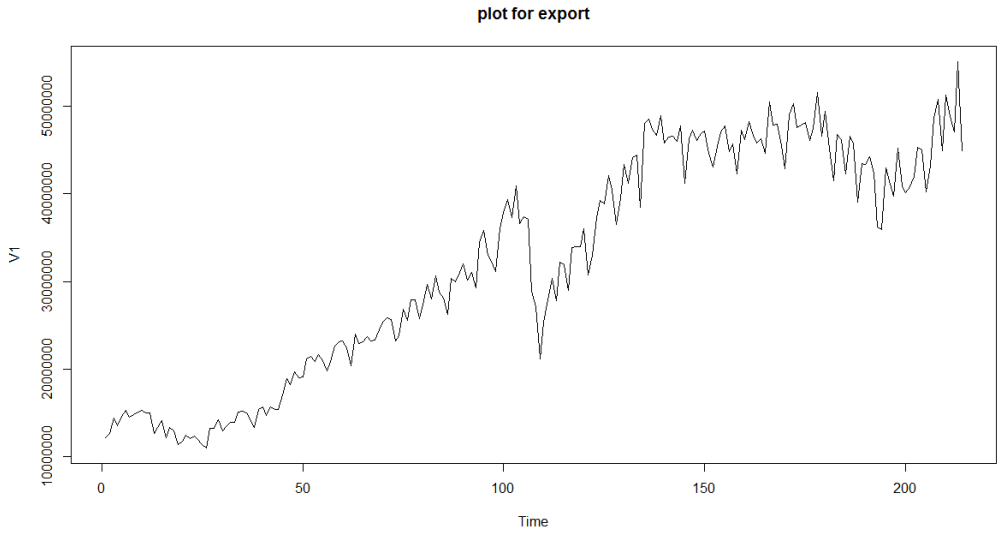
[표 2.1] 가법계절지수와 승법계절지수의 SSE, RMSE, MAE값



[그림 2.2] 계절지수평활법을 적용시킨 가법계절지수, 승법계절지수 시계열 적합도 그림

3. 인공신경망

한국수출금액 데이터 214개 중 14개는 모형평가, 200개는 모형구축에 사용하도록 하겠다.

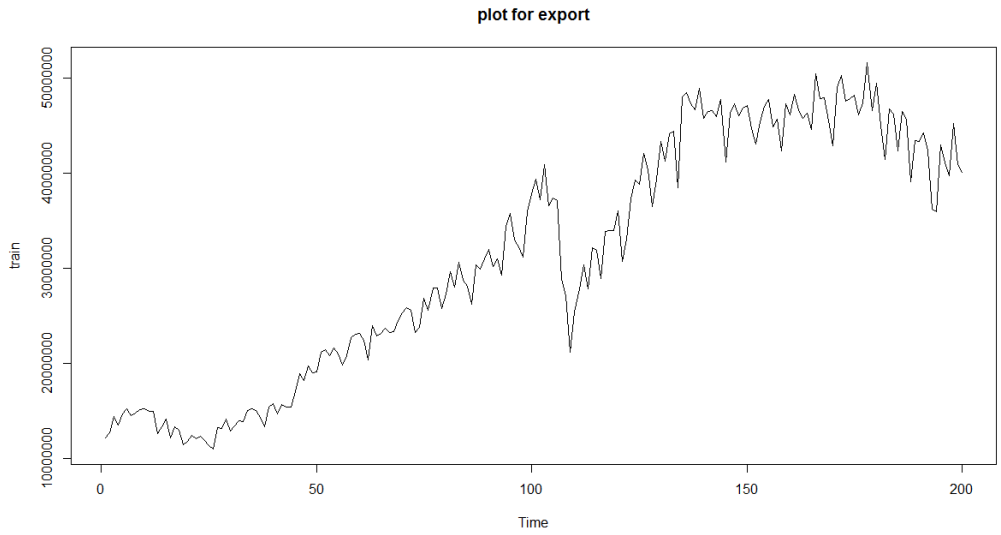


[그림 3.1] 한국수출금액 시계열 그림

[그림 3.1]에서 export 원자료의 시계열 그림을 확인할 수 있다.

```
#cross-validation
h<-14
train <- export[1:(length(export)-h)]
test <- export[((length(export)-h)+1):length(export)]
plot.ts(train, main="plot for export")
```

h=14로 지정하여 214개의 데이터로 이루어져 있는 export를 train이라는 변수를 사용해 모형구축에, test를 통해 모형평가에 사용하도록 지정한다. 모형구축을 위한 train의 시계열 그림은 [그림 3.2]로 확인할 수 있다.



[그림 3.2] train의 시계열 그림

```
##arima
auto.arima(train)
fit<-arima(train,order=c(1,1,2))
summary(fit)
coeftest(fit)
fore <- forecast(fit)
```

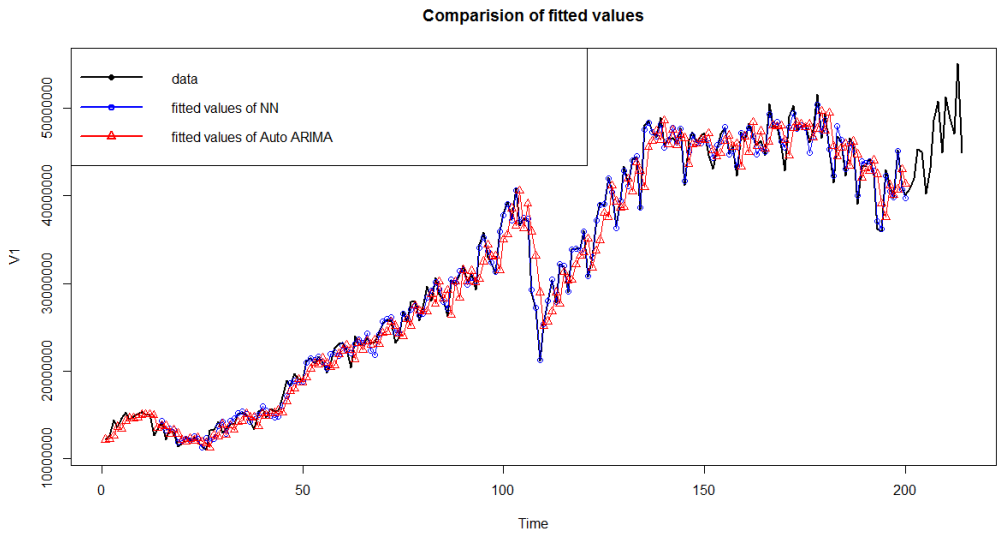
인공신경망과 arima모형과의 비교를 위해 train을 이용해 auto.arima를 적합한 결과 arima(1,1,2)를 적합 시킨다. summary 결과 RMSE는 2354278, MAE는 1716476임을 확인할 수 있었고 모든 모수가 유의함을 coeftest를 통해서 확인했다. 하지만 포토맨토우 검정을 통해서 p-value값이 6시점이후로 귀무가설을 기각, 모형이 적합하지 않다는 검정결과를 얻을 수 있었다. 하지만 이번 절에서는 ann과의 비교를 위해 사용하겠다.

```
##ann
nfit <- nnetar(train)
fcast <- forecast(nfit, h=h, level=0.95)
```

유의수준 95%로 인공신경망에 train을 적합 시킨다. Press로 prediction error sum of square 값을 확인해보니

press.ari	44860036924695.9
press.nn	37392855262836

다음과 같은 값으로 인공신경망으로 적합 시킨 모델이 더 좋은 모델이라는 것을 알 수 있다.



[그림 3.3] 인공신경망과 auto.arima의 적합값 비교 그림

[그림 3.3]은 auto.arima와 ann의 적합값 비교 그림으로 press 값으로 확인한 바와 같이 ann이 좀 더 적합함을 확인할 수 있다.

Ⅲ 결론

본문에서 시행한 3가지 모형적합방법에 대해서 아래의 표 SSE, RMSE, MAE로 비교를 해보자. 이 모든 통계량은 작을수록 더 좋은 모형임을 나타내주는 지표이다. ANN은 RMSE, MAE를 구할 수 없어 공통적으로 비교할 수 있는 SSE로 비교를 해보자. 우선 평활법이 압도적으로 SSE가 작다. 따라서 가장 적합한 모형은 평활법임을 알 수 있다. 그러나 ARIMA와 RMSE, MAE가 크게 차이 나지 않음을 보아 세가지 모형 모두 적합하나, 그 중 평활법이 가장 오차의 제곱합이 작음을 확인할 수 있다. 즉, 평활법 > ANN > ARIMA 순서대로 적합함을 알 수 있다. ANN은 단기예측에 좋은 모형이기 때문에 단기적으로 봤을 때는 ANN이 더 적합할 수도 있다.

이번 분석을 실행하면서 수출금액에 대해서 기존의 수출금액 자료만으로 예측하기는 쉬운 일이 아니라는 것을 느꼈다. 수출금액은 우리나라만의 문제가 아닌 전 세계적인 사건이나 사고에 큰 영향을 받는 데이터이기 때문일 것이다. 원자료 시계열 그림만 살펴봐도 지속적으로 증가하던 수출금액이 2010년 전에 큰 폭으로 감소하였음을 볼 수 있다. 이에 시계열 자료의 분석은 기존 데이터 자료와 연관성이 있는 지표들을 잘 살펴보아야 한다고 생각하게 되었다. 현재 우리나라의 수출품의 대부분이 반도체와 IT 부문인 것을 감안하여 앞으로의 우리나라의 반도체의 입지가 큰 영향을 줄 수 있다고 생각한다.

	SSE	RMSE	MAE
ARIMA	813046239124376	1953745	1466841
평활법	0.7530467	2000916	1434172
ANN	64908225859626		

참고문헌

- 책

[1] 이성덕, "SAS와 R을 이용한 시계열자료분석", 자유아카데미]

- 인터넷 사이트

[2] 관세청, "수출입 무역통계", https://unipass.customs.go.kr:38030/ets/index.do,

[3] 공공데이터포털, "수출입 무역통계, "https://www.data.go.kr/information/1000560/monthData.do,

[4]티스토리, http://untitledtblog.tistory.com/]

- 사전

[5] 네이버 지식백과, 수출액 [Amount of export], (통계표준용어, 통계청)

- 논문

[6] 조나래, (2015), "패널 1차 자기회귀모형의 동질성 검정 통계량 비교"