Chapter 13 Stationary and nonstatonary time series

1. 정상성(stationary=안정성)

* 직관적으로 정상성이 의미는 시계열의 수준과 분산에 체계적인 변화가 없고, 엄밀하게 주기적 변동이 없다는 것으로 미래는 확률적으로 과거와 동일하는 것을 뜻한다. 다음 세가지 조건을 약한 정상성 또는 공분산 정산성이라 하고 이들 조건을 모두 만족하는 시계열을 정상시계열(stationary time series)라고 한다.

첫째, 평균값은 시간 t에 관계없이 일정하다. 즉 E(Yt)=u이다.

둘째, 분산값은 시간 t에 관계없이 일정하다. 즉 Var(Yt)=시그마^2

셋째, 공분산은 시간 t에 의존하지 않고 오직 시차에만 의존한다. 즉 Cov(Yt, Yt+k)=감마k 시차 k만 의존, t와 관계없다.

* 유한개의 관찰된 시계열 자료로부터 모집단인 시계열 모형에 관한 통계적 추정 및 검정을 하기 위해서는 분석을 단순화시키는 어떤 가정이 필요한데, 이 중 가장 중요한 것이 정상성 가정이다. 만약 위 세가지 요건 중 하나라도 성립하지 않는 경우 비정상(=불안정, nonstationary) 시계열이라고 한다.

이때, 시계열의 평균이 일정하지 않는 경우에는 원계열을 차분해주므로 정상시계열이 되고, 시계열의 분산이 일정하지 않는 경우에는 원계열에 자연로그를 취하면 정상시계열이 된다.

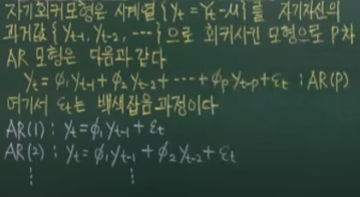
1. 중요한 확률 과정
2. 백색잡음 과정(white noise process)

* 시계열이 상호독립적이고 동등하게 분포하는 확률변수의 관측치일 때 백색잡음 과정이라고 한다.

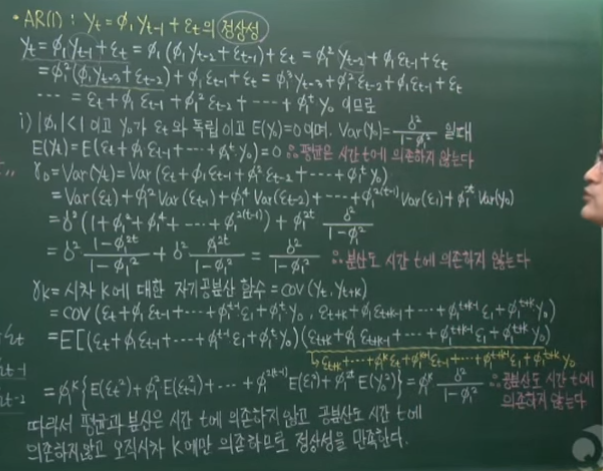


1. 자기회귀(AR)과정 (=자기회귀 (AR)모형)

* Yt에 평균은 뺀 yt(편차)를 확률 변수로 한 것이다.



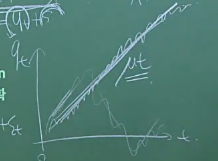
* 대분분의 경우 AR(1), AR(2)에서 끝납니다.
* AR(1)의 정상성은 분석해볼 수 있습니다. 정상시계열인지 아닌지



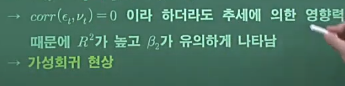
* 감마k는 yt와 시차가 k만큼 떨어진 yt+k의 (시차 k에 대한) 자기공분산 함수이다. Cov(yt, yt+k)
* 감마0는 자기자신끼리의 공분산은 분산을 의미한다.

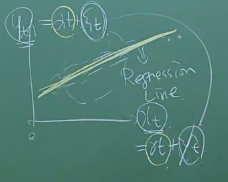
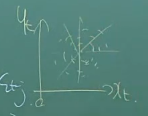
1. 위의 식의 과정을 통해 세가지 조건을 모두 만족하므로 정상성을 만족한다.
2. l파이l<1이고, y0=0일 때는 조건 2,3이 시간 t에 의존하게 된다. 하지만 t가 무한대로 가면, 즉 시계열 자료가 충분히 크면 l파이l<1이므로 파이 제곱t는 0으로 수렴하게 된다. 따라서 AR(1)은 점근적 정상성을 갖는다고 하며 일반적으로 이면 초기값 y0와 상관없이 AR(1)은 정상성을 갖는다고 한다.
3. 는 random walk process(확률보행과정)이라고 하고, 비정상 시계열이다.
4. 단위근 검정

-위에서 불안정한 확률과정을 배웠다. 단위근 검정에서 볼 확률과정은 주로 확률보행과정이다. 시계열이 추세가 있는 것처럼 나옵니다. 상수항이 있다면 오른 쪽 그림과 같이 상수에 의해 추세가 있듯이 나타난다. 또한 안정적인 확률과정은 외생적 쇼크가 있을 때 평균으로 회귀하는데, 확률보행과정은 외생적 충격이 있을 때 평균으로 회귀하지 않는다. 베타계수가 계속 1이기 때문이다.

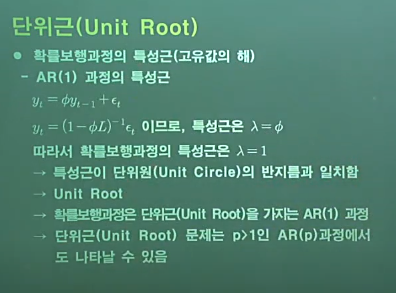
 

* 확률보행과정은 시계열의 안정성이 보장되지 않기 때문에 가성회귀라는 문제점이 있다. 가성회귀(Squrious regression) 실제로 관계가 없지만 관계가 있듯이 나타나는 현상이다. 사실 상관없는 것인데 그저 추세에 의해 상관있어 보이는 것이다. 추세 때문에 xt와 yt사이의 관계가 없더라도 회귀직선이 아래와 같이 나타날 수 있다. 그 뒤의 상수들인 et와vt의 코릴레이션이 0이면 R^2값이 0이다. B2는 알파t와 감마t에 관한 추세에 대한 영향력을 나타내는 인자다



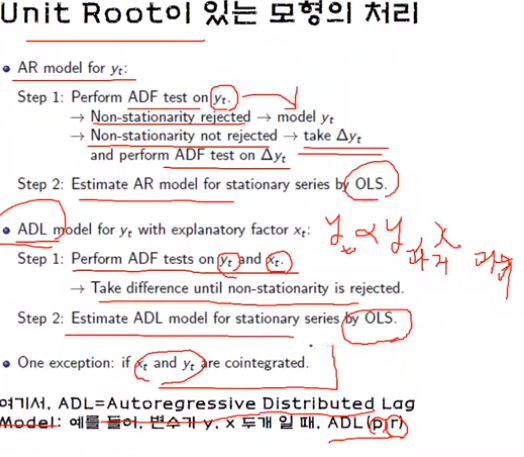
* 시계열 자료를 차분하면 회귀직선을 높으면, 가성회귀현상이 제거된다.



* 반지름이 1인 원 안에 특성근이 있어야 stationary한건데 특성근이 원주에 있을 때를 말하다. 고유치로 된 백터의 길이가 정확하게 1인 경우는 불안정하다.
* 단위근은 그래프를 보고 식별이 불가능하다. 따라서 단위근 여부에 대한 가설검정이 필요하다.

시스템이 statinary하기만 하면 OLS에서 tf테스트를 할 수 있다. 어떤 타임시리즈에서 yt+k의 계수 베타가 1이냐 1보다 작은 값이냐가 중요하다. 위에서 말함. 1이면 unit root Root가 있고, Stochastic Trend가 있다. Unit Root가 있으면 시계열 자료에 Stochastic Trend가 있다는 뜻이고, Non-stationary하다는 의미로 시계열자료를 처리하기가 힘들어진다. 따라서 Unit Root가 있는지를 먼저 판단하기 위해 Augmented Dicky Fuller(ADF) 검정을 한다. 만일 이 검정을 통과하지 못하면 본래의 시계열 자료를 차분(Differencing)한 후 다시 ADF 검정을 한다. 일단 ADF 검정을 통과하면 대개의 경우 OLS 회귀분석을 통해 AR 또는 ADL 모형의 계수 값들을 추정할 수 있다.

ADF는 unit root가 있는지 없는지 stochastic이 있는지 없는지 판단하는 모형이다. 귀무가설H0은 로우값=0인지 보는 것이고, 대립가설 H1은 로우값이 0보다 작은 것을 보는 것이다. t값이 -2.9보다 작으면 귀무가설을 기각이 되고, 대립가설이 채택된다. 이런경우 stationay한거다 unit root값이 없는 거다.



Unit Root Test(단위근검정)는 Stationary한지를 알아보는 Test이다. Non-stationary time series를 바로 분석하면 많은 문제가 있는데, 대표적으로 spurious regression의 문제가 있다. 따라서 times series가 non-stationary할 경우 spurious regression이 있는 변수간 유의미한 회귀관계가 일어나는 것을 막기 위해 time series의 stationary를 판단하는 테스트. 이게 DF검정과 ADF검정의 기본이다.

교수님 수업시작

미래 수요예측의 정량적 기법으로 시계열 분석법이 있다.

시계열이란 한 사건 또는 여러 사건에 대하여 시간의 흐름에 따라 일정한 간격으로 이들을 관찰하여 기록한 자료를 말한다. 시계열 자료란 시간과 더불어 관측된 자료로 이는 종단면 자료에 해당하며 횡단면 자료는 고정된 시간에서 측정된 자료를 의미하며 측정시간이 고정되어 있는 반면 여러 개의 변수로 구성된다. 시계열분석에 의한 수요예측기법은 과거의 자료(매출액, 수요량)를 토대로 그 추세나 경향을 분석하여 미래의 수요를 예측하는 방법으로 과거의 추세나 사이클이나 계절적인 변동이 일정할 때 사용할 수 있는 기법이다.

시계열 분석법으로 수요예측을 하려면 먼저 4가지 변동요인을 알아야 한다.

1. 추세변동은 장기 변동요인으로 수요가 일정하게 증가하거나 감소하는 형태이다.
2. 계절변동은 1년이나 월간 주간 단위로 계절요인으로 수요량이나 판매량이 주기적으로 되풀이 되는 변동이다.
3. 계절변동은 계절요인에 따라 주간 단위로 발생하는 시계열의 단기 변동요인이다.
4. 불규칙변동이다. 돌발적인 사고나 파업이나 악천후 등 단기간에 일어나는 변동이다.

(Stata랑 EVIEWS는 R과 달리 돈주고 사야햄 stata들어가서 통계분석->시계열->ARIMA 및 ARFIMA모형)

Conitegration(공적분)

* 단기적으로 다를 수도 있지만, 장기적으로 보았을 때는 서로 일정한 관계가 있다는 것.
* 쉬운 예시로는 술취한 사람이 반려견을 끈에 묶어서 돌아다니는 것과 같이 사람과 반려견은 아무런 관계없이 서로 움직이지만, 서로 멀어지지 않고 결국엔 집으로 가는 것과 비슷하다.
* 원래 단위근 검정에서 I(1)이면 믿을 수 없지만 공적분이 존재한다면 믿을 수 있습니다.

Engle and Granger(1987) Cointegration

* 두 시계열 x,y가 공적분이라고 가정했을 때, y를 종속변수, x를 독립변수로 회귀분석을 하여 만들어지는 잔차 zt에 대해서는 다음과 같은 식이 성립한다. 이러한 식을 ECM식이라고 한다. 여기서 et는 회귀 분석으로 만들어진 ECM형식의 잔차이다.
* 여기서 zt(yt와 xt를 회귀 분석한 결과)와 et(회귀 분석으로 ECM형식의 잔차)가 정상 시계열이면 두 시계열이 공적분인지 안ㄹ 수 있다.
* 해당 공적분에서 귀무가설과 대립가설은 아래와 같다.

귀무가설 : 공적분 관계가 존재하지 않는다.

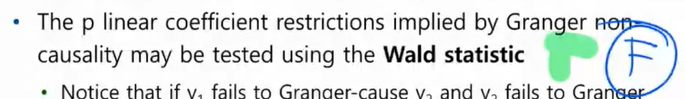
대립가설 : 공적분 관계가 존재한다.

* 즉, p-value를 5%라고 지정했을 때, p-value 값이 0.5이하이면, 귀무가설을 기각할 수 있기 때문에 대립가설이 채택이 되고 그 말은 공적분 관계가 존재한다라고 말할 수 있다.

Chapter 14. Cointegration and error correction models

* 사하라 사이즈와 한국인의 GDP는 둘다 상관관계가 있다고 나올것이다. B1이 굉장히 high하게 나오고, R^2값다 high하게 나올것이다.
* 다시말한다. 단위근검정을 통해 정상 시계열인지 확인을 하고, 비정상시계열이면 공적분을 하고요 공적분없으면 차분해야한다 델타 y와 델타 x에 대해서 하는데, 석유가격수준에 대한 정부는 잃어버린다. change에만 대해 이야기 할 수 있다.

GRANGER CAUSALITY TEST (CONT.)

* 인과관계를 추측할 수 있게 해주는 단순한 방법 중 하나이다.
* 
* 이것도 R로할수있다.
* 그래인저 인과관계를 확인하기 위해서는 시차(lag 또는 지연)를 파라미터로 넣어 주어야 한다. 이 각은 사람에 의해 주어져야 하며 적당한 값을 찾기는 매우 어려우므로 여러 시차를 반복을 통해서 실험해 보거나 경험적인 판단에 의존하여 적절한 구간을 실행해 보고 논리적으로 적당한 값을 찾아서 사용해야 한다. 이 검정법은 입력한 시차에 해당하는 지연데이터만 사용해서 선형 회귀를 수행하는 것이 아닌 lags가 N이라면 1부터 N까지의 지연에 해당하는 모든 데이터를 전부 사용한다는 것을 주의해야 한다. 정상성 - Stationary 두 변수는 모두 정상성(stationary)을 가지는 것을 전제로 하며 정상성을 확보하지 않은 상태에서 데이터를 입력하면 오해석할 여지가 많은 결과가 나온다는 점을 매우 주의한다. 이는 대부분의 시계열 분석에서 전제되는 조건중 하나이다. 이 때문에 이 테스는 여러가지 데이터에 일반화해서 자동하화기 어렵다.
* 시 계열 변수 A, 시계열 변수 B가 있으며 이 두 변수는 충분한 양(대략 30여개 이상의 선형회귀가 작동 가능한 데이터)의 요소들을 가지고 있는 연속형 변수이다. 즉 시간에 따라 순서를 가진 2개의 집계 데이터 또는 관측 데이터가 있고 이들이 가진 시간축은 같은 것이다. 이 두 변수를 시계열 또는 변수라고 하고 여기서는 편의상 변수라고 부르기로 하겠다. A의 과거 데이터의 집합 즉 A의 시차(lags) B의 고거 데이터의 집합 즉 B의 시차(lags) 위와 같은 데이터가 있을 때 다음과 같이 결과에 대한 조건이 통계적으로 유의미하면 A가 B에 대해 Granger Causality하다고 표현하며, 이는 인과관계가 있을 여지가 있다고 볼 수 있다. 하지만 명확하게 인과관계가 있다는 의미가 아니므로 매우 주의한다. A lags + B lags로 B의 데이터를 선형회귀한 것의 예측력 > B lags로만 B의 데이터를 선형회귀한 것의 예측력
* 사용 과정 및 방법 변수A → 변수B에 대해 결정한 lags를 입력하고 Granger Causality test를 수행한다. 유의 수준을 0.05(5%)로 설정하였다면 검정값(p-value)가 0.05이하이면 귀무가설을 기각할 수 있다. 귀무가설은 “Granger Causality를 따르지 않는다” 이다. 변수B → 변수A에 대한 것을 위와 동일한 방법으로 수행한다. lags바꾸어가며 반복해서 테스트한다. 이때 lags는 경험적으로 잘 알려진 것을 사용하거나 논리적으로 적당한 수준까지 입력한다. 또는 적절한 Lags를 찾는 방법을 활용한다.
* 결과 해석 : 위의 사용과정을 거쳐서 시차(lag)가 정해진 경우에는 경우의 수를 보면 총 4개의 결과가 나올 수 있다.

변수A → 변수B = Granger Causality 성립, 변수B → 변수A = Granger Causality 성립하지 않음

이 경우는 변수A가 변수B에 선행한다고 볼 수 있다. 즉 변수A가 변수A의 인과요인이 될 가능성이 높다.

변수A → 변수B = Granger Causality 성립하지 않음, 변수B → 변수A = Granger Causality 성립

이 경우는 변수B가 변수A에 선행한다고 볼 수 있다. 즉 변수B가 변수A의 인과요인이 될 가능성이 높다.

변수A → 변수B = Granger Causality 성립, 변수B → 변수A = Granger Causality 성립

쌍방으로 Granger Causality가 성립하는 경우로 이 경우는 제3의 외부변수(Exogenous Variable)가 영향을 공통으로 주었을 가능성이 높다. 이 경우 제3의 외부변수를 알아내던가 포기하던가 해야 한다. VAR모형(사실 Granger Causality도 VAR모형중 하나이다)을 사용해야 할 수 있다.

변수A → 변수B = Granger Causality 성립하지 않음, 변수B → 변수A = Granger Causality 성립하지 않음

두 변수가 서로 인과영향을 주지 않는다고 볼 수도 있지만 단언하지는 못한다. ARIMA모형으로 추가 확인이 가능한 것으로 알려져 있다. 주의 할 점은 위의 결과가 입력값으로 주는 시차에 따라서 달라질 수 있으므로 시차에 따른 해석을 달리해야 하는 문제가 있다. 즉 해석에 있어서 사람의 경험과 판단이 개입되어야 한다.