계량경제학
이론과 실습

본 과제(결과물)는 교육과학기술부의 재원으로 한국연구재단의 지원을 받아 수행된 광역경제권 선도산업 인재양성사업의 연구결과입니다.

제주대학교 강기춘

목 차

제 1 장

계량경제학 개요

1.계량경제학 정의
2.계량경제학 목적 및 중요성
3.계량경제학 분류 및 접근방법
4.자료의 형태 및 통계 패키지
5.회귀분석의 분류

제1장 계량경제학 개요

1.계량경제학 정의

계량경제학은 영어로 Econometrics인데 이는 Economic Measurement라는 의미로 양으로 측정된 경제학이라고 할 수 있다. R.Frisch는 계량경제학은 경제이론, 경제수학, 통계학을 조화시킨 학문으로 이들 세 분야와는 엄격히 구분되는 경제학의 한 분야라고 정의하였으며, A.S.Goldberger는 계량경제학은 경제이론, 수학, 그리고 통계적 추론 등의 분석도구를 실제 현상을 분석하는데 응용하는 사회과학이라고 정의하였다. 따라서 계량경제학이란 경제이론(economic theory), 수학(mathematics) 및 통계학(statistics)의 도구를 이용하여 경제관계를 경험적으로 결정하는 경제학의 한 분야라고 할 수 있다.

계량경제학이 경험적으로 결정하는 경제관계의 유형은 다음과 같이 나눌 수 있다.

첫째는 행위관계(behavioral relations)이다. 행위관계를 나타내는 대표적인 예가 소비함수인데 소비함수는 주어진 소득 하에서 재화와 서비스에 대한 소비자들의 평균적인 지출행위를 나타낸다.

$$C=\alpha+\beta Y \quad (0<\beta<1)$$

둘째는 기술적 관계(technical relations)이다. 모든 경제관계가 행위관계만이 있는 것은 아니다. 생산함수는 행위관계를 나타낸다고 보기 보다는 기술적인 관계로 간주한다.

$$Q=AK^{\alpha}L^{\beta}$$

셋째는 정의관계(definitional relations)이다. 이 관계는 항등식(identity)이라고도 하는데 예를 들면 개방경제에서 지출국민소득이 소비, 투자, 정부지출 및 순수출의 합으로 정의된 것 같이 경제변수가 항등식이 반드시 성립하도록 정의가 된 관계를 말한다.

$$Y=C+I+G+X-M$$

넷째는 제도적 관계(institutional relations)이다. 이 관계는 경제제도에 의해 경제변수간의 관계가 설정이 된 것을 말하는데 다음의 식은 국민총생산의 20%가 조세총액이라고 할 경우 즉, 조세부담률이 20%일 경우를 나타낸다.

$$T = 0.2Y$$

다섯째는 균형관계(equilibrium relations)이다. 이 관계는 조정방정식(adjustment equation)이라고도 하는데 수요와 공급의 일치를 나타낸다.

$$Q_d = Q_s$$

한편, 이러한 경제관계를 나타내는 방법은 크게 두 가지가 있다. 그 하나는 시간의 변화를 고려하는 지의 여부에 따라 정태적 관계(static relation)와 동태적 관계(dynamic relation)가 있다. 또 다른 하나는 경제주체 또는 경제전체에 따라 미시적 관계(micro relation)와 거시적 관계(macro relation)가 있다.

경제관계는 유형에 따라 수학적으로 나타낼 수 있는데 수학적으로 표현한 경제관계를 경제모형(economic model)이라 한다. 경제모형은 크게 시계열모형(time series model)과 구조모형(structural model)이 있는데 위에서 언급한 경제모형을 구조모형이라 한다.

경제모형은 여러 가지 면에서 유용한데 첫째는 경제의 각 부문이 어떻게 상호작용을 하는 지를 명시적으로 나타내고 있어 경제에 대한 체계적인 이해가 가능하고 둘째는 경제모형이 잘 설정이 되었고 그러한 관계가 지속된다는 가정 하에 미래의 경제상황을 예측할 수 있다. 셋째는 경제환경 변화에 대한 경제변수의 움직임을 알 수 있기 때문에 바람직한 경제정책을 선택하는데도 도움이 된다.

2.계량경제학 목적 및 중요성

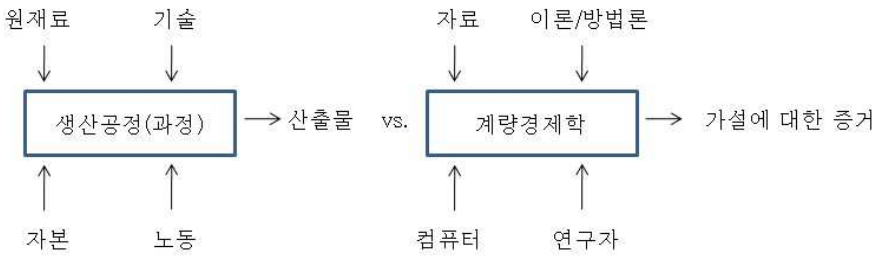
(1)계량경제학 목적

계량경제학의 목적은 경제통계자료의 특징 분석, 경제이론의 검증, 경제정책의 분석, 미래에 대한 예측, 실증분석방법론의 개발 등 다양하다. 첫째 경제통계자료의 특징 분석이란 통계자료의 기술통계량, 요인분석 등 경제변수의 여러 특성을 파악하는 것을 의미하며, 둘째 경제이론의 검증이란 경제학의 각 분야에서 개발된 경제이론이 현실경제를 설명하는데 적합한 자료인지를 실제자료를 이용하여 분석하는 즉, 경제이론의 현실설명력을 살펴보는 것을 말한다. 셋째 경제정책의 분석이란 한 경제변수의 변화가 다른 경제변수에 어떤 영향을 미치는가를 분석하는 것을 의미하며, 넷째 미래에 대한 예측이란 알려진 경제변수들 간의 관계를 이용하여 특정 경제변수의 미래 예측치를 추정하는 것을 의미하고, 다섯째 실증분석방법론의 개발은 경제학의 실증연구에 적합한 방법론을 개발하는 것을 의미한다.

(2)계량경제학 필요성

계량경제학의 중요성을 원재료, 자본, 노동, 기술의 생산요소를 사용하여 생산물을 만들어 내는 생산과정(production process)에 비유해 설명해 보면 <그림1-1>과 같다. 계량경제학은 생산과정의 원재료에 비유할 수 있는 자료(data), 자본에 비유할 수 있는 컴퓨터, 기술에 비유할 수 있는 이론 및 분석방법을 이용하여 연구자가 경제이론을 뒷받침 해 주는 증거를 찾는 학문이라고 할 수 있다. 생산과정이 생산이론의 기초가 되듯이 계량경제학은 이론의 실제증거를 찾아 이론의 성립여부를 판단해 주는 매우 중요한 분야라고 할 수 있다.

<그림 1-1> 생산과정과 비교한 계량경제학



3.계량경제학 분류 및 접근방법

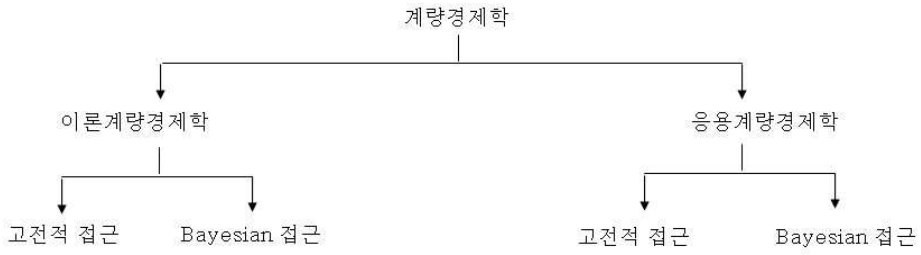
(1)계량경제학 분류

계량경제학은 크게 이론계량경제학(theoretical econometrics) 및 응용계량경제학(applied econometrics)을 구분되는데 이론계량경제학은 경제변수들 간의 관계를 나타내는 계량모형의 효율적인 추정방법 등을 개발하는 분야이고 응용계량경제학은 이론계량경제학에서 개발한 방법론을 생산, 소비, 투자, 수요, 공급 등 경제학의 여러 분야에 적용하여 분석하는 분야이다.

한편, 이론계량경제학이든 응용계량경제학이든 접근방법은 고전적 접근(classical approach) 및 베이지언 접근(Bayesian approach)을 활용하는데 고전적 접근방법은 주어진 모집단에서 생성된 표본을 확률표본으로 간주하고 분석하는 반면에 베이지언 접근방법은 관측된 표본을 생성시킨 모집단을 확률집단으로 간주하고 분석한다.

이에 따라 계량경제학 구분을 도식화 해 보면 <그림 1-2>와 같다.

<그림 1-2> 계량경제학의 분류



(2)계량경제학 접근방법

계량경제학은 다음과 같은 순서에 따라 단계적으로 접근한다.

첫째는 계량모형의 설정(Specification of Econometric Model)단계이다. 계량모형의 설정이란 경제이론과 일치되게 변수들을 정하고 그 변수들의 위치를 정하는 것인데 변수의 선정, 모수의 부호와 크기, 모형의 수학적 형태 등을 고려해야 한다. 예를 들면, 케인즈의 소비이론에 따르면 소비(C)는 소득(Y)의 함수이다. 이들 관계를 확정적으로 나타내는 수학모형(mathematical model)을 표시하면 식 (1.1)과 같다.

$$C = \alpha + \beta Y \quad (0<\beta<1) \tag{1.1}$$

그러나 현실적으로 볼 때 소비에 영향을 주는 변수들은 소득 이외에도 연령, 가족구성원의 수, 종교 등 많이 있다. 이러한 변수들의 영향은 정확하게 측정하기가 어렵기 때문에 모두 일정한 확률분포를 이루고 있는 교란항(disturbance term: U)에 포함시키면 식 (1.2)식과 같이 되는데 이를 계량모형(econometric model)이라고 한다.

$$C = \alpha + \beta Y + U \ (0<\beta<1) \tag{1.2}$$

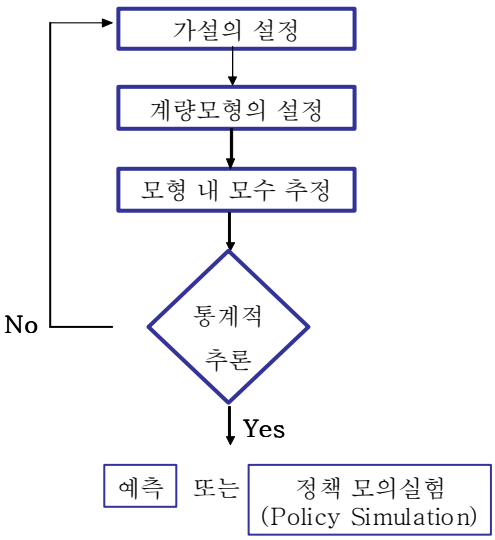
둘째는 추정(Estimation)의 단계이다. 추정이란 이용 가능한 자료로 모형 내에 있는 모수(parameter)의 추정량을 얻는 것을 말하는데 추정단계에서는 자료수집, 설명변수들 간의 상관관계 확인, 적절한 통계기법 등을 고려해야 한다.

셋째는 평가의 단계이다. 모형의 모수가 추정되었으면 추정결과가 통계적으로 유의한지(statistically significant)와 경제적으로 유의한 지(economically significant)를 살펴보아야 한다. 평가단계는 추정치에 대한 평가와 예측력에 대한 평가로 구분될 수 있는데 추정치 평가에서는 경제이론, 통계, 계량경제적 기준에 의해 평가하고, 예측력 평가에서는 예측치와 실제치의 차에 대한 유의성을 고려해야 한다.

이러한 계량경제학의 접근방법을 순서도(flow chart)로 나타내면 <그림 1-3>과 같은데 그림에서 가설의 설정 및 계량모형의 설정은 위에서 설명한 계량모형의 설정단계에 해당되고, 모형 내 모수 추정은 추정의 단계에 해당되며, 통계적 추론 및 예측(또는 정책 모의실험)은 평가의 단계에 해당된다.

케인즈의 소비이론을 계량경제학의 접근 순서도에 따라 설명해 보면 먼저 한계소비성향은 0과 1사이라고 하는 가설을 설정하고 위의 (1.2)식과 같은 계량모형을 설정한다. 다음으로 소득과 소비에 관한 자료를 모아 모형의 모수($\alpha, \beta, \sigma_u^2$)들의 추정치를 구하고 앞에서 설정한 가설을 검정한다. 통계적 추론에서 귀무가설이 기각되지 않으면 주어진 독립변수(소득)의 값으로 종속변수(소비)의 미래 값을 예측하고, 소비를 바람직한 수준으로 유지하기 위해서는 소득수준이 얼마가 되어야 하는 지 등 정책 모의실험을 시도한다.

<그림 1-3> 계량경제학 접근 순서도



4.자료의 형태 및 통계 패키지

계량경제분석에서 이용하는 자료의 형태에는 횡단면 자료(Cross-Sectional Data), 시계열 자료(Time Series data), 합동 횡단면 자료(Pooled Cross Sections), 패널자료(Panel or Longitudinal data) 등이 있다.

횡단면 자료란 일정시점에서 하나 이상의 변수에 대해 수집된 자료를 의미하는데 예를 들어 2007년 전국 16개 시도의 지역내총생산(GRDP)과 총소비를 조사한 자료는 횡단면 자료이다.

시계열 자료란 일별, 주별, 월별, 분기별, 연도별 등 시간에 걸쳐 수집한 자료를 의미하는데 예를 들어 연도별 제주지역의 지역내총생산과 총소비를 조사한 자료는 시계열 자료이다. 시계열 자료는 거시경제변수를 측정한 자료에서 주로 많이 발생한다.

합동 횡단면 자료란 횡단면 자료와 시계열자료가 결합된 자료를 의미하는데 예를 들어 1985년부터 2007년까지 전국 16개 시도의 지역내총생산과 총소비를 조사한 자료는 합동 횡단면 자료이다.

패널자료(Panel data) 또는 종적자료(Longitudinal data)는 동일한 횡단면 단위에 기준을 두고 시간의 흐름에 따라 수집한 자료를 의미하는데 예를 들어 특정 가구를 패널로 선택한 후에 매년 그 가구를 대상으로 교육비 지출을 조사한 자료는 패널자료 또는 종적자료이다.

계량경제학에서 통계 소프트웨어 활용은 반드시 필요하다. 현재 시중에는 WinRats-32(Regression Analysis for Time Series). GAUSS for Windows(Matrix programming language), EViews(Econometric Views), SAS(Statistical Analysis System, general econometrics and modelling), Stata(Statistics Data), SPSS(Statistical Package for the Social Sciences), Limdep(Limited Dependent model) 등 다양한 종류의 통계 소프트웨어가 출시되어 활용되고 있다.

WinRATS-32는 윈도우용 RATS(Regression Analysis for Time Series) 프로그램인데 RATS는 계량경제학과 관련된 분야를 주 대상으로 하는 통계 패키지로서 Doan과 Litterman에 의해 개발되었다. 특히, 최근 관심을 모으고 있는 ARIMA 및 VAR 등 시계열분석에 매우 강력한 기능을 가지고 있다.

GAUSS는 1984년 미국의 Aptech systems사에 의해 개발되어 출시된 이후 빠른 계산을 수행하거나 방대한 자료를 분석할 수 있는 프로그램으로 발전해 왔다. GAUSS는 계량경제학과 관련된 분야를 주 대상으로 하는 행렬기반의 고급수준의 프로그래밍 언어(matrix programming language)이기 때문에 초기에 배우기가 약간 까다로운 점이 있지만 행렬연산이 매우 간단하고 내장된 함수가 많으며 이 언어를 사용한 최신 응용프로그램들을 인터넷 등을 통해 얻을 수 있다는 장점이 있다.

EViews는 Quantitative Micro Software에 의해 개발되어 1994년 초에 보급되기 시작한 계량분석과 프로그래밍이 동시에 가능한 통계처리 소프트웨어로서 사용이 간편하고 그래픽 기능이 뛰어난 장점이 있으며 시계열분석에 많이 이용되고 있다.

SAS(Statistical Analysis System)는 1975년 미국 North Carolina 대학에서 개발된 프로그램으로 통계분석에 관한 패키지 프로그램 중 가장 적용 범위가 넓고 그 사용이 일반화되어 있다.

Stata는 Statistics와 Data의 합성어로 1980년대 중반 미국의 StataCorp이 개발한 통계 소프트웨어 패키지이다. Stata는 학술용 및 업무용으로 이용되고 있으며, 경제학, 사회학, 정치학 등 사회고학은 물론 의학 분야 등 자연과학에서도 많이 이용되고 있다. 또한 전 세계적으로 사용되는 프로그램이며 다양한 사용자 모임이 개최될 정도로 사용자층이 광범위하고 다양하다

이상에서 살펴본 바와 같이 소프트웨어마다 제 나름대로의 장단점을 가지고 있는데 최근에 경제학 분야에서 그 사용자가 많이 늘어나고 있는 통계 패키지는 이 책에서 활용하는 있는 EViews와 Stata이다.

5.회귀분석의 분류

계량경제학의 기본적인 분석방법인 회귀분석은 회귀모형을 이용한 분석과 시계열모형을 이용한 분석 등 크게 두 가지로 구분되며, 회귀모형이나 시계열모형 모두 단일방정식에 의한 분석과 연립방정식(또는 다변수모형)에 의한 분석으로 구분될 수 있다.

회귀모형을 이용한 회귀분석 중 단일방정식에 의한 회귀분석은 선형모형을 이용한 회귀분석과 비선형모형을 위한 회귀분석으로 나누어지며, 선형모형은 단순회귀모형, 다중회귀모형 및 특수모형을 이용한 회귀분석으로 구분된다.

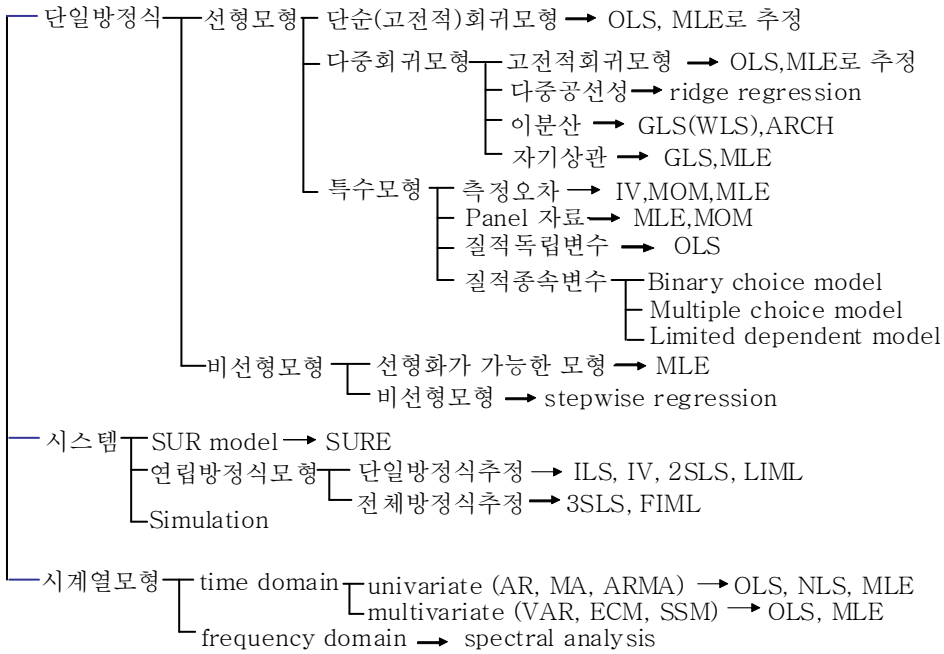
한편, 각 모형의 모수는 다양한 추정방법으로 추정되는데 고전적 회귀모형의 경우 보통최소자승법(Ordinary Least Squares : OLS) 또는 최우추정법(Maximum Likelihood Estimation : MLE)이 이용되며, 다중회귀모형에서 교란항에 대한 기본가정이 성립되지 않는 경우 즉 이분산(heteroscedasticity)이나 자기상관(autocorrelation)이 존재할 경우 일반화최소자승법(Generalized Least Squares : GLS)이 이용된다.

연립방정식모형의 경우 단일방정식 추정에는 간접최소자승법(Indirect Least Squares : ILS), 수단변수(Instrumental Variable : IV) 추정방법, 2단계최소자승법(Two-stages Least Squares : 2SLS), 제한정보최우추정법(Limited Information Maximum Likelihood : LIML) 등이 이용되고, 전체방정식 추정의 경우 3단계최소자승법(Three-stages Least Squares : 2SLS), 전정보최우추정법(Full Information Maximum Likelihood : LIML) 등이 이용된다.

시계열모형의 경우 보통최소자승법, 비선형최소자승법(Nonlinear Least Squares : NLS), 최우추정법 등이 이용된다.

이러한 회귀분석의 분류 및 추정방법을 요약해 보면 <그림 1-4>와 같다.

<그림 1-4> 회귀분석의 분류 및 추정방법



제 2 장

회귀분석과 회귀모형

1.회귀분석 개요

2.회귀모형의 개념

제2장 회귀분석과 회귀모형

1.회귀분석 개요

회귀(regression)란 용어는 Galton이 주장한 '보편적 회귀의 법칙(law of universal regression)'에서 최초로 사용되었다. 일반적으로 키가 큰 부모에게서 키가 큰 자녀들이 태어나고 키가 작은 부모에게서 키가 작은 자녀들이 태어나는 경향이 있지만 키가 큰 부모집단 자녀들의 평균 신장이 부모들의 평균 신장보다 작은 반면에 키가 작은 부모집단 자녀들의 평균 신장이 부모들의 평균 신장보다 커서 키가 크고 작은 자녀들이 모두 전체의 평균 신장을 향해 움직이거나 "회귀(regression)"하는 경향이 있다는 것이다.

그러나 현대적 의미의 회귀는 이와는 매우 다르다. 회귀분석이란 종속변수와 독립변수와의 의존관계를 분석하는 것을 말하며 이미 알려진 독립변수의 값으로 종속변수의 평균적인 값을 추정 또는 예측하는데 관심이 있다.

변수들 사이의 관계를 나타내는 데는 확정적(또는 함수적) 관계와 확률적(또는 통계적) 관계가 있다. 확정적 관계(deterministic relation)란 모형 내 변수들이 확률변수가 아니므로 독립변수가 주어지면 정확하게 종속변수의 값을 계산할 수 있는 관계를 의미하는데 물리학에서 말하는 Newton의 법칙이 확정적 관계를 나타내는 예이다. 확률적 관계(stochastic relation)란 모형 내 변수들이 모두 확률변수이므로 독립변수가 주어진다고 해서 정확하게 종속변수의 값을 계산할 수 없는 관계를 의미한다. 예를 들어 감귤의 수확량이 기온, 일조량, 강우량에 의존한다고 할 때 이 변수들만으로는 감귤수확량을 정확하게 예측할 수 없는데 그 이유는 이러한 독립변수들 외에 다른 변수들도 비록 정확하게 식별해 내지는 못할지라도 감귤수확량에 영향을 미칠 수 있기 때문이다.

회귀분석(regression analysis)과 밀접한 관계에 있지만 개념적으로 다른 것으로 상관분석(correlation analysis)이 있다. 첫째 회귀분석에서는 독립변수는 확정변수로 가정하고 종속변수는 확률변수로 가정하는 반면에 상관분석에서는 두 변수 모두 확률변수를 가정한다. 둘째 회귀분석에서는 독립변수의 주어진 값으로 종속변수의 평균값을 추정하고 예측하는 것이 목적인 반면에 상관분석에서는 두 변수 간의 선형성의 정도(degree of linearity)를 측정하는 것이 목적이다. 셋째 회귀분석에서는 영향을 주는 변수를 독립변수로 영향을 받는 변수를 종속변수로 구분하는 반면에 상관분석에서는 독립변수와 종속변수의 구분이 없다.

독립변수와 종속변수를 구분하는 회귀분석에서는 다양한 용어들이 사용되고 잇

는데 독립변수(Independent variable)는 설명변수(Explanatory variable), 예측변수(Predictor), 회귀변수(Regressor), 자극 또는 통제변수(Stimulus or control variable), 외생변수(Exogenous variable)라고도 하며 종속변수(Dependent variable)는 설명된 변수(Explained variable), 예측된 변수(Predictand), 피회귀변수(Regressand), 반응변수(Response variable), 내생변수(Endogenous variable)라고도 한다.

회귀분석의 종류에는 단순회귀분석(simple regression analysis)과 다중회귀분석(multiple regression analysis)이 있는데 단순회귀분석이란 독립 및 종속변수가 모두 하나인 회귀분석으로 2변수 회귀분석(two-variable regression analysis)이라고도 하며, 다중회귀분석은 하나의 종속변수와 둘 이상의 독립변수를 가진 회귀분석을 말한다.

2.회귀모형의 개념

제주도에는 전체 60개의 물산업 관련 기업체가 있으며(따라서 이를 모집단으로 볼 수 있음), 제주도 물산업 관련 기업의 연간매출액 Y와 연간 홍보비 지출액 X의 관계를 회귀모형을 이용하여 분석한다고 하자. 즉, 기업의 홍보비 지출액을 알고 있을 때 연간매출액의 모집단 평균수준을 예측한다고 하자.

<표 2-1> 홍보비 지출액과 연간매출액

X / Y	80	100	120	140	160	180	200	220	240	260
연간 매출액	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	-	88	-	113	125	140	-	160	189	185
	-	-	-	115	-	-	-	162	-	191
합계	325	462	445	707	678	750	685	1043	966	1211
$E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

전체 60개 기업을 비슷한 홍보비 지출액 수준에 따라 10개의 집단으로 나누어 각 홍보비 지출액(단위: 천만 원) 집단에 속한 기업의 연간매출액(단위: 억 원)을 조사한 결과 <표 2-1>과 같다고 하자.

이 표를 보는 방법은 다음과 같다. 예를 들어 홍보비로 100천만 원을 사용하는 기업은 6개가 있는데 연간매출액의 범위는 65억 원에서 88억 원에 이른다. 또한 홍보비로 220천만 원을 사용하는 기업은 7개가 있는데 연간매출액이 가장 많은 기업은 162억 원, 가장 작은 기업은 135억 원인 것으로 나타나고 있다. 즉, 이 표는 홍보비 지출이 고정되어 있을 때 그 지출수준에 해당되는 집단에서의 연간 매출액 분포를 보여 주고 있는데 이를 X가 주어져 있을 때 Y의 조건부 분포(conditional distribution)라고 한다.

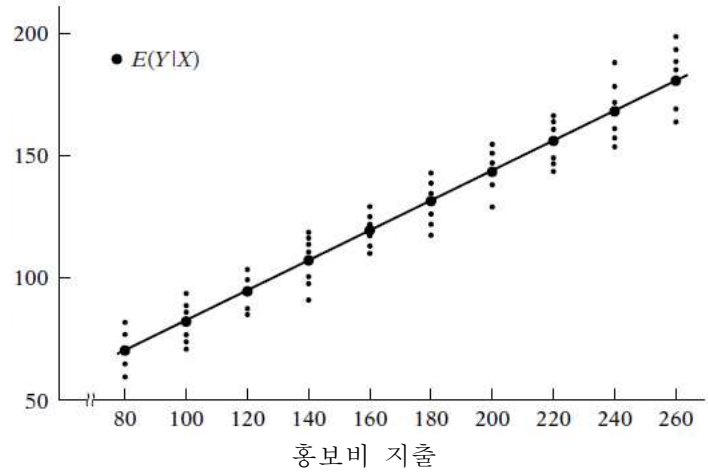
<표 2-2> <표 2-1> 자료에 대한 조건부 확률

X Y	80	100	120	140	160	180	200	220	240	260
연간매출액	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	-	1/6	-	1/7	1/6	1/6	-	1/7	1/6	1/7
	-	-	-	1/7	-	-	-	1/7	-	1/7
$E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

<표 2-2>의 조건부 확률을 이용하여 Y에 대한 조건부 평균(기댓값) 즉, X가 특정한 값 X_i를 취할 경우 Y의 기대치($E(Y \mid X)$)를 구할 수 있는데 Y 값에 조건부 확률을 곱한 후 그 값을 다 더하면 된다. 예를 들어 X가 100일 경우 조건부 평균은 65(1/6)+70(1/6)+74(1/6)+80(1/6)+85(1/6)+88(1/6)=77로 구해진다.

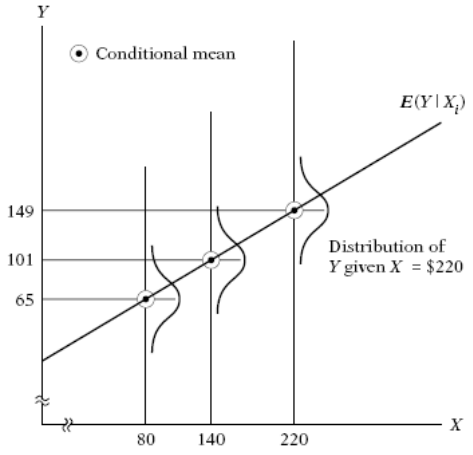
한편, <표 2>에서 구해진 조건부 평균과 <표 2-1>의 연간매출액을 산포도로 그려보면 <그림 2-1>과 같은데 개별기업의 연간매출액은 크고 작은 차이가 있으나 홍보비 지출이 증가할수록 연건매출액도 평균적으로 증가하고 있음을 나타내 주고 있다. X 값에 대한 Y의 조건부 평균이 <그림 2-1>에서 진한 점으로 표시되어 있는데 조건부 평균이 우상향의 기울기를 가진 직선 위에 놓여 있다.

<그림 2-1> 홍보비 지출수준에 대한 연간매출액의 조건부 분포



설명변수의 값이 주어졌을 때 종속변수의 조건부 평균(기댓값)을 이은 선을 모집단 회귀선(population regression line)이라고 하는데 <그림 2-2>는 <표 2-1>에 있는 자료의 모집단 회귀선을 나타내고 있다. 그림에서 보듯이 모집단 회귀선은 조건부 평균을 통과하여 지나고 있음을 알 수 있다.

<그림 2-2> 모집단 회귀선



<그림 2-2>를 통해 알 수 있듯이 각각의 조건부 평균은 $E(Y \mid X_i)$은 X_i의 함수임을 알 수 있는데 이를 모집단 회귀함수(population regression function)라고 한다. 만약 이 모집단 함수가 선형이라면 다음의 (2.1)식과 같이 나타낼 수 있다.

$$E(Y|X_i) = \beta_0 + \beta_1 X_i \tag{2.1}$$

홍보비 지출이 증가하면 연간매출액도 평균적으로 증가한다. 그러나 홍보비 지출이 증가함에 따라 개별 기업의 연간매출액이 반드시 증가하는 것은 아니다. 주어진 홍보비 지출 수준에서 개별기업의 연간매출액은 조건부 평균 주위에 모여 있기는 하지만 서로 다를 수 있다. 개별기업의 연간매출액과 조건부 평균의 차이를 편차(deviation) 또는 확률적 교란항(stochastic disturbances)이라고 하며 편차를 u_i라고 할 경우 다음의 (2.2)식과 같이 나타낼 수 있다.

$$u_i = Y_i - E(Y|X_i) \tag{2.2}$$

또는 $Y_i = E(Y|X_i) + u_i$

따라서 (2.1)식과 (2.2)식을 결합하면 모집단 회귀함수(population regression function: PRF)는 (2.3)식과 같이 나타낼 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2.3}$$

단, u_i는 확률적 교란항 또는 확률적 오차항(stochastic error term)이라 한다.

지금까지는 모집단에 대한 논의이다. 그러나 대부분 실제 계량분석에서 이용되는 자료는 표본이므로 표본 및 표본에 근거한 회귀함수(이를 표본 회귀함수라고 함)에 대한 논의가 필요하다.

모집단(<표 2-1> 자료)에 대한 정보는 없고 특정한 X에 대해서 Y 값을 무작위로 추출한 표본 <표 2-3>과 <표 2-4>만 알고 있다고 하자. 즉, 이 두 표본에서 Y 값은 X_i가 주어졌을 때 Y 값 중에서 무작위로 선택된 것이다.

표본을 이용하여 특정한 홍보비 지출액 X에 대한 연간매출액 Y를 예측할 수 있을까? 즉, 표본을 이용하여 모집단 회귀함수(PRF)를 추정할 수 있을까?

<표 2-3> <표 2-1>의 모집단에서 추출한 확률표본 1

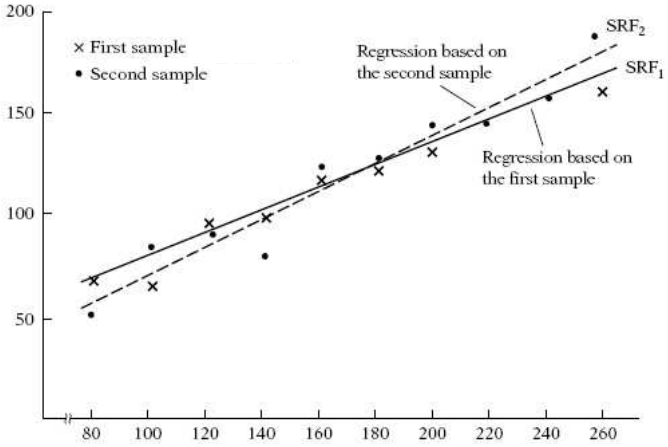
Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

<표 2-4> <표 2-1>의 모집단에서 추출한 확률표본 2

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

이러한 질문에 답하기 위해 두 확률표본의 산포도(scatter plot)를 그린 것이 <그림 2-3>에 나타나 있다. 산포도에는 추출된 표본을 잘 설명할 수 있는 두 회귀선을 나타냈는데 이를 표본 회귀선(sample regression line)이라고 하며 <표 2-3>에 있는 확률표본 1의 표본 회귀선은 SRF_1이고 <표 2-4>에 있는 확률표본 2의 표본 회귀선은 SRF_2이다.

<그림 2-3> 두 확률표본에 근거한 회귀선



두 표본 회귀선은 각각 주어진 표본을 가장 잘 설명할 수 있도록 나타낸 것인데 두 표본 회귀선 중 어느 것이 모집단 회귀선을 잘 나타내고 있는 지 확실하게 알 수 있는 방법은 없다.

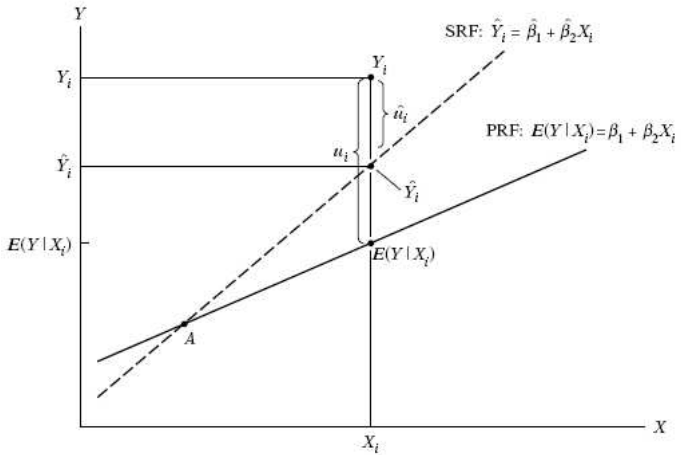
모집단에 대해 모집단 회귀선과 모집단 회귀함수가 정의되는 것처럼 표본에 대해서도 표본 회귀선과 표본 회귀함수(sample regression function: SRF)가 정의될 수 있으며 선형의 표본 회귀함수는 다음의 (2.4)식과 같이 나타낼 수 있다.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \qquad (2.4)$$

단, $\hat{\beta}_0, \hat{\beta}_1$은 β_0, β_1의 추정량을 나타내고, e_i는 잔차항(residual term)이라 하는데 확률적 교란항인 u_i와 유사하며 u_i의 추정량이라 할 수 있다.

<그림 2-5>는 모집단 회귀함수와 표본 회귀함수의 관계를 나타내고 있다. 회귀분석의 주목적은 표본 회귀함수(SRF)인 $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$를 이용하여 모집단 회귀함수(PRF)인 $Y_i = \beta_0 + \beta_1 X_i + u_i$를 추정하는 것인데 표본추출의 변동이 있기 때문에 PRF를 정확하게 추정할 수 없고 근사치를 추정할 뿐이다. 예를 들어 특정한 X_i에 대한 모집단의 조건부 평균은 $E(Y \mid X)$인데 표본을 이용하여 조건부 평균을 추정한 추정치는 $\hat{Y}_i$이므로 $\hat{Y}_i$는 $E(Y \mid X)$의 근사치이다.

<그림 2-5> 모집단 회귀함수와 표본 회귀함수



제 3 장
단순회귀분석

1.단순회귀모형의 가정
2.회귀계수의 추정
3.OLS 추정량의 특성
4.회귀계수의 분산 추정
5.결정계수
6.가설검정
7.예측
8.최우추정법
9.회귀모형의 응용
실습 : 단순회귀모형 추정

제3장 단순회귀분석

1.단순회귀모형의 가정

2-변수 모집단 회귀함수(PRF)를 단순회귀모형(simple regression model)이라고도 하는데 이를 다시 한 번 나타내면 (3.1)식과 같다.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad (3.1)$$

단, Y_i 는 종속변수, X_i는 독립변수, β_0, β_1는 회귀계수, u_i는 교란항 또는 교란항을 각각 나타낸다.

단순회귀모형을 추정하고 분석결과를 제대로 해석하기 위해서는 X 및 교란항에 대한 다음과 같은 가정이 필요하다.

(가정 1) 독립변수 X는 비확률변수이다.

이 가정은 표본을 반복해서 추출할 때 X는 고정된 것으로 해 놓고 Y를 추출한다는 것을 의미하므로 X는 확정변수이며 X는 오차 없이 측정된다(measurement without error). 예를 들어 제2장 <표 2-1>의 모집단에서 <표 2-3>이나 <표 2-4>와 같은 표본을 추출할 때 홍보비 지출을 80천만 원으로 고정한 다음 홍보비 지출이 80천만 원인 기업 중 연간매출액 자료를 추출하고, 또한 홍보비 지출은 100천만 원으로 고정한 다음 홍보비 지출이 100천만 원인 기업 중 연간매출액 자료를 추출한다는 의미이다.

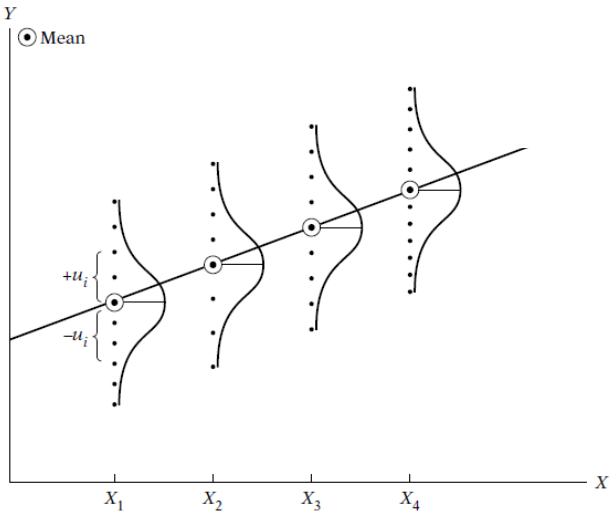
(가정 2) 교란항의 평균은 0이다.

$$E(u_i) = 0 (i = 1,2,3,..,n)$$

제2장 <그림 2-3>에서 보듯이 어떤 관측치는 회귀선 위에 위치하여 교란항의 값이 양수인 반면에 어떤 관측치는 회귀선 아래에 위치하여 교란항의 값이 음수가 되므로 이들의 평균은 0이 된다. 이 가정을 좀 더 정확하게 표현하면 $E(u_i \mid X_i) = 0$인데 이럴 경우 $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$가 된다. 이것은 모형에 명시

적으로 포함되지 않고 교란항에 묶여져 있는 변수들은 Y의 평균치에 체계적인 영향(systematic effect)을 미치지 않는다는 의미이다. 즉, 양의 u_i 값들과 음의 u_i 값들이 서로 상쇄되어 Y에 미치는 u_i의 평균효과가 0이라는 의미이다.

<그림 3-1> 교란항 u_i의 조건부 분포



(가정 3) 교란항은 모든 X_i에 대한 동일한 분산을 갖는다.

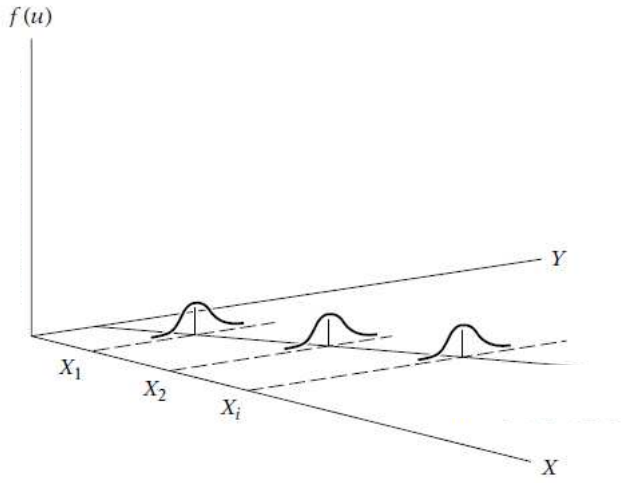
$$E(u_i^2) = \sigma_u^2$$

이 가정은 각 독립변수 X의 값에 있어서 종속변수 Y가 X의 평균을 중심으로 분포되어 있는 정도가 같다는 것을 의미하는데 이를 동분산(homoscedasticity)이라고 한다. 정규분포에 대한 가정이 없어도 추정에는 문제가 없으나 가설검정을 위해서는 분포에 대한 가정이 필요하다

(가정 4) 모든 u값은 서로 독립이다.

$$Cov(u_i, u_j) = E(u_i, u_j) = 0, (i \neq j)$$

<그림 3-1> 교란항의 동분산



이 가정은 교란항에 계열상관(serial correlation) 또는 자기상관(autocorrelation)이 없다는 것인데 이는 교란항의 변화에 체계적인 형태가 존재하지 않는다는 것을 의미한다.

(가정 5) u와 X는 서로 독립이다.

$$Cov(u_i, X_i) = E(u_i, X_i) = 0$$

이 가정은 교란항과 설명변수가 상관관계가 없다는 것을 의미하는데 이를 직교조건(orthogonality condition)이라 한다. 직교조건의 의미는 종속변수와 관계가 있는 모든 설명변수들이 제대로 모형에 포함되어 있어 회귀모형에서 X가 Y에 미치는 영향과 u가 Y에 미치는 영향이 서로 분리되어 있다는 것을 의미한다. 나중에 추정량의 성질에서 설명하겠지만 좋은 추정량 중 하나는 불편추정량(unbiased estimator)인데 만약에 종속변수와 관계가 있는 독립변수를 모형에 포함시키지 않으면(즉, 직교조건이 성립하지 않으면) 최소자승추정량(OLS)은 편의(bias)가 있는 추정량(biased estimator)이 된다.

단순회귀모형에서 기억해야 할 점은 첫째 X와 Y는 선형관계(linear relationship)에 있다는 것이다. 선형관계라고 할 때는 자료에 있어서의 선형관계 또는 회귀계수에 있어서의 선형관계 등이 있는데 여기서는 회귀계수에 있어서

의 선형을 의미한다. 둘째 단순회귀모형은 확정식이 아니고 확률식이라는 것이다. 단순회귀모형에 대한 가정 중 독립변수 X는 비확률변수이나 교란항이 확률변수이므로 단순회귀모형은 확률식이 된다. 셋째 교란항은 여러 가지 의미를 내포하고 있다. 먼저 교란항은 생략되거나 배제된 모든 변수를 대신한다. 모형 설정 시 간결한 모형(parsimonious of model)의 원칙에 따라 특정 독립변수가 반드시 모형에 포함되어야 할 이유가 없다면 모형에서 배제된 그러한 변수들을 교란항이 대신하게 할 수 있다. 다음으로 교란항은 아무리 노력해도 설명할 수 없는 인간 행동의 임의성을 반영하고 있다. 또한 교란항은 측정오차(measurement error)를 반영하고 있는데 측정오차란 X와 Y변수를 측정하는데 발생하는 오차로서 X와 Y변수가 관찰이 되지 않는 변수일 경우 관측 가능한 변수를 대리변수(proxy variable)로 사용하게 되는데 이 때 발생하는 오차를 측정오차라고 하며 이 경우 교란항이 측정오차를 나타낸다.

2.회귀계수의 추정

단순회귀모형에서 회귀계수의 추정방법은 보통최소자승법(Ordinary Least Squares: OLS)과 최우추정법(Maximum Likelihood Estimation: MLE)이 있다. <그림 3-2>에서 보듯이 실제 관측치를 가장 잘 나타내도록(또는 추적하도록) 표본회귀함수((3.3)식)를 결정하는 것이 중요한데 이 때 실제 관측치와 회귀모형에 의한 추정치의 차이를 나타내는 (3.4)식의 잔차(residual)가 큰 역할을 한다.

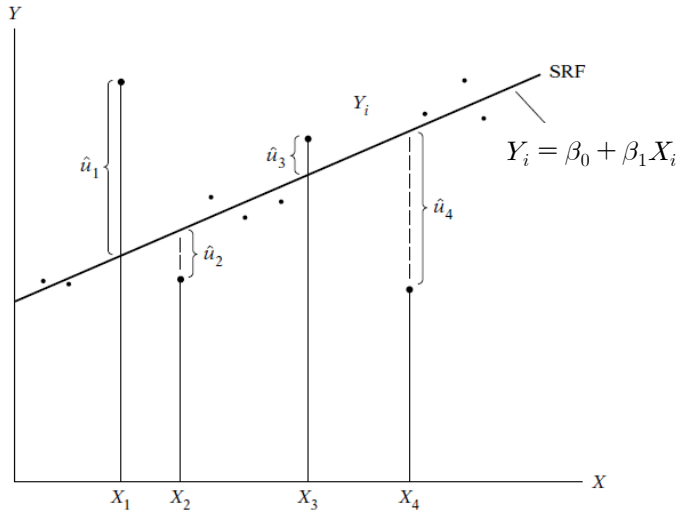
실제 관측치를 잘 나타내는 표본 회귀식을 구하기 위해서는 잔차의 합계가 최소가 되도록 하는 것이 바람직하다. 그러나 잔차의 합이 0이 되는 식은 유일하지 않으므로 잔차의 제곱의 합이 최소가 되게 하는 회귀식을 구하면 되는데 이러한 추정방법을 보통최소자승법(OLS)이라고 한다.

(회귀모형) $$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{3.2}$$

(추정된 회귀선) $$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i \tag{3.3}$$

(잔차) $$e_i = Y_i - \widehat{Y_i} \text{ 또는 } \widehat{u_i} = Y_i - \widehat{Y_i} \tag{3.4}$$

<그림 3-2> 보통최소자승법



보통최소자승법은 (3.5)식과 같은 잔차의 제곱의 합인 $\sum_{1=1}^{n} e_i^2$이 최소가 되도록 하는 회귀계수(β_0, β_1)를 구하는 방법이다.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 \qquad (3.5)$$

(3.5)식이 최소화해야 할 목적함수이며 선택변수가 $\widehat{\beta}_0, \widehat{\beta}_1$이므로 이 식을 선택함수에 대해 각각 1차 미분하면 (3.6)과 (3.7)식이 각각 도출된다.

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \widehat{\beta}_0} = \frac{\partial \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2}{\partial \widehat{\beta}_0} = 2\sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)(-1) = 0 \qquad (3.6)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \widehat{\beta}_1} = \frac{\partial \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2}{\partial \widehat{\beta}_1} = 2\sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)(-X_i) = 0 \qquad (3.7)$$

(3.6)식과 (3.7)식으로부터 다음의 (3.8)식과 (3.9)식을 구할 수 있는데 이를 정규방정식(normal equation)이 유도된다.

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \tag{3.8}$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \tag{3.9}$$

정규방정식을 연립으로 풀면 $\hat{\beta}_0, \hat{\beta}_1$을 구할 수 있는데 먼저 $\hat{\beta}_1$을 구하기 위하여 (3.8)식에 $\overline{X}$ 곱하면 (3.10)식을 구할 수 있다.

$$\overline{X}\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 \overline{X} + \hat{\beta}_1 \overline{X} \sum_{i=1}^{n} X_i \tag{3.10}$$

(3.9)식에서 (3.10)식을 빼주면 (3.11)식이 구해진다.

$$\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i\right) \tag{3.11}$$

(3.11)식을 $\hat{\beta}_1$에 대해 풀면 (3.12)식과 같게 된다.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{cov(X, Y)}{var(X)}$$

또는 $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{3.12}$$

단, $x_i = X_i - \overline{X}, y_i = Y_i - \overline{Y}$는 각각 편차를 나타낸다.

(3.12)식에서 알 수 있는 것은 기울기를 나타내는 회귀계수 $\hat{\beta}_1$은 독립변수의 분산에 대한 독립변수와 종속변수의 공분산의 비율이라는 것이다.

한편, 절편을 나타내는 $\hat{\beta}_0$을 구하기 위해 (3.8)식의 양변을 n으로 나누어 주면 (3.13)식이 되고 따라서 $\hat{z}_0$는 (3.14)식과 같게 된다.

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X} \tag{3.13}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{3.14}$$

따라서, 보통최소자승법으로 구한 $\hat{\beta}_0, \hat{\beta}_1$의 추정량은 다음과 같다.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X} \sum_{i=1}^{n} X_i}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

(참고 1) 합산연산자(Summation Operator)

기호 $X_1, X_2, \ldots, X_n$이 n개의 관측치를 나타낼 때, 이들의 총합 $X_1 + X_2 + \ldots + X_n$은 총합기호 $\sum$를 사용하여 나타내면 간단하다. 총합기호는 다음과 같은 특성을 가지고 있다.

① $\sum_{i=1}^{n} (X_i \pm Y_i) = \sum_{i=1}^{n} X_i \pm \sum_{i=1}^{n} Y_i$

② $\sum_{i=1}^{n} cX_i = c \sum_{i=1}^{n} X_i$

③ $\sum_{i=1}^{n} c = nc$

④ $\sum_{i=1}^{n} (X_i \pm c) = \sum_{i=1}^{n} X_i \pm nc$

⑤ $\sum_{i=1}^{n} (X_i \pm c)^2 = \sum_{i=1}^{n} X_i^2 \pm 2c \sum_{i=1}^{n} X_i + nc^2$

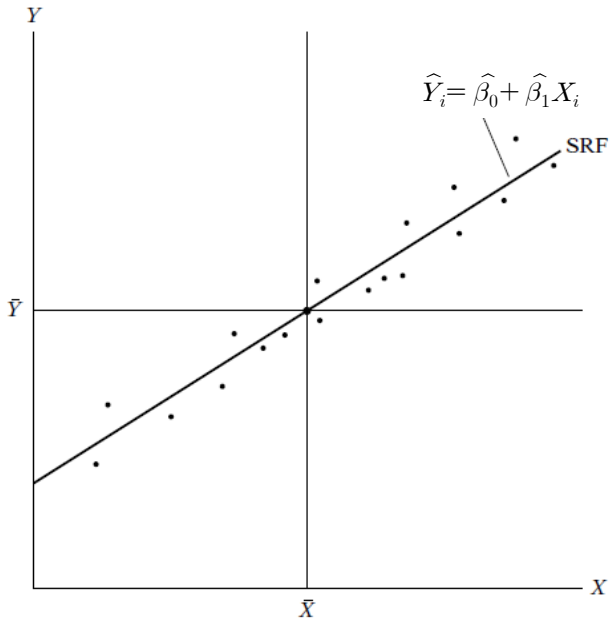
추정된 회귀선은 다음과 같은 특징을 가진다.

첫째, 최소자승법으로 구한 회귀선은 항상 X, Y의 표본평균점을 지난다. (3.13)식에서 $\overline{Y}=\hat{\beta}_0+\hat{\beta}_1\overline{X}$이므로 즉, 추정된 회귀식에 의하면 X가 $\overline{X}$일 때 Y는 $\overline{Y}$이므로 추정된 회귀식(SRF)은 <그림 3-3>과 같인 X와 Y의 평균점을 반드시 지난다. 또한 이 특징은 편차를 이용한 회귀선의 추정에도 시사점을 주는데 예를 들어 X와 Y의 편차를 각각 $x_i=X_i-\overline{X}, y_i=Y_i-\overline{Y}$라고 하고 편차를 이용하여 회귀모형을 설정하고 추정한다고 할 때 회귀모형은 (3.15)식과 같게 되고 보통최소자승법으로 추정한 회귀계수의 추정량은 (3.12)식과 같으며 추정된 회귀선은 (3.16)식과 같다.

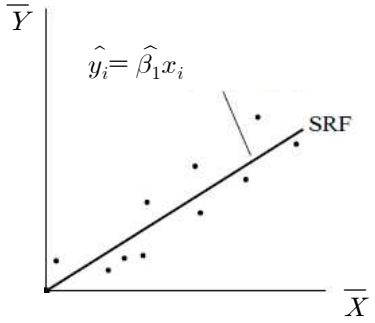
(회귀모형) $y_i=\beta_1 x_i+u_i$ (3.15)

(추정된 회귀선) $\hat{y_i}=\hat{\beta}_1 x_i$ (3.16)

<그림 3-3> SRF와 표본평균점



<그림 3-4> 편차 회귀선



둘째, Y의 추정치 $\hat{Y}_i$의 평균값은 실제 Y의 평균값과 같다. 즉, $\overline{\hat{Y}} = \overline{Y}$이다. 먼저 추정된 회귀선에 $\hat{\beta}_0$의 추정량을 대입하고 정리하면 (3.17)식과 같게 된다.

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ &= (\overline{Y} - \hat{\beta}_1 \overline{X}) + \hat{\beta}_1 X_i \\ &= \overline{Y} + \beta_1 (X_i - \overline{X})\end{aligned} \tag{3.17}$$

양변에 총합기호를 사용하면 (3.18)식과 같게 된다.

$$\sum \hat{Y}_i = n\overline{Y} + \beta_1 \sum (X_i - \overline{X}) \tag{3.18}$$

(3.18)식에서 $\sum (X_i - \overline{X}) = 0$이므로(왜냐하면 편차의 합은 항상 0이므로) (3.18)식의 양변을 n으로 나누어 주면 $\overline{\hat{Y}} = \overline{Y}$가 성립한다.

셋째, 잔차의 합 $\sum e_i$은 0이다. 즉, 잔차의 평균 $\overline{e}$은 0이다. (3.2)-(3.4)식을 이용하면 (3.19)식과 같이 나타낼 수 있다. 즉, 실제치는 추정치와 잔차의 합으로 나타낼 수 있다.

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \tag{3.19}$$

(3.19)식의 양변에 총합기호를 사용하면 (3.20)식과 같게 되고 이 식이 (3.8)식의 정규방정식과 동일하기 위해서는 $\sum e_i = 0$이 반드시 성립해야 한다.

$$\sum Y_i = n\hat{\beta_0} + \hat{\beta_1}\sum X_i + \sum e_i \tag{3.20}$$

이를 이용하면 표본회귀함수(SRF)를 편차형태(deviation form)로 나타낼 수 있다. (3.19)식에서 (3.14)식을 빼면 다음의 식과 같게 된다.

$$Y_i - \overline{Y} = \hat{\beta_1}(X_i - \overline{X}) + e_i$$

또는 $y_i = \beta_1 x_i + e_i$

$y_i = \beta_1 x_i + e_i$는 편차형태의 표본회귀함수이고 따라서 보통최소자승법으로 추정한 회귀식은 $\hat{y_i} = \hat{\beta_1} x_i$이 된다.

넷째, 잔차와 Y의 추정치는 상관관계가 없고, 잔차와 X도 상관관계가 없다. 즉, $\sum \hat{y_i} e_i = 0, \sum X_i e_i = 0$이다. (3.7)식의 정규방정식으로부터 $\sum X_i e_i = 0$임을 알 수 있고, 이를 이용하면 다음의 (3.21)식과 같이 $\sum x_i e_i = 0$임을 보일 수 있다.

$$\begin{aligned} \sum x_i e_i &= \sum (X_i - \overline{X}) e_i \\ &= \sum X_i e_i - \overline{X}\sum e_i \\ &= 0 \quad (\because \sum X_i e_i = 0, \sum e_i = 0) \end{aligned} \tag{3.21}$$

또한 이를 이용하면 (3.22)식과 같이 $\sum \hat{y_i} e_i = 0$ 임을 보일 수 있다.

$$\begin{aligned} \sum \hat{y_i} e_i &= \sum \hat{\beta_1} x_i e_i \\ &= \hat{\beta_1} \sum x_i e_i \\ &= 0 \end{aligned} \tag{3.22}$$

(예제 3-1)
다음의 홍보비 지출액 X(단위: 천만 원)와 연간매출액 Y(단위: 억 원)에 관한 자료를 이용하여 회귀계수와 회귀식을 구하고 해석하라.

X	2	3	4	5	6
Y	4	4	6	6	10

(풀이)
먼저 주어진 자료를 이용하여 OLS 추정량을 구하기 위한 기초계산을 하면 다음과 같다.

$$\sum X_i = 20,\ \sum Y_i = 30,\ \sum X_i Y_i = 134,\ \overline{X} = 4,\ \overline{Y} = 6,\ \sum X_i^2 = 90,\ \sum Y_i^2 = 204$$

위에서 계산한 값들을 OLS 추정량에 대입하면 다음과 같은 회귀계수를 구할 수 있다.

$$\therefore\ \widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{134 - 4\times 30}{90 - 4\times 20} = \frac{14}{10} = 1.4$$

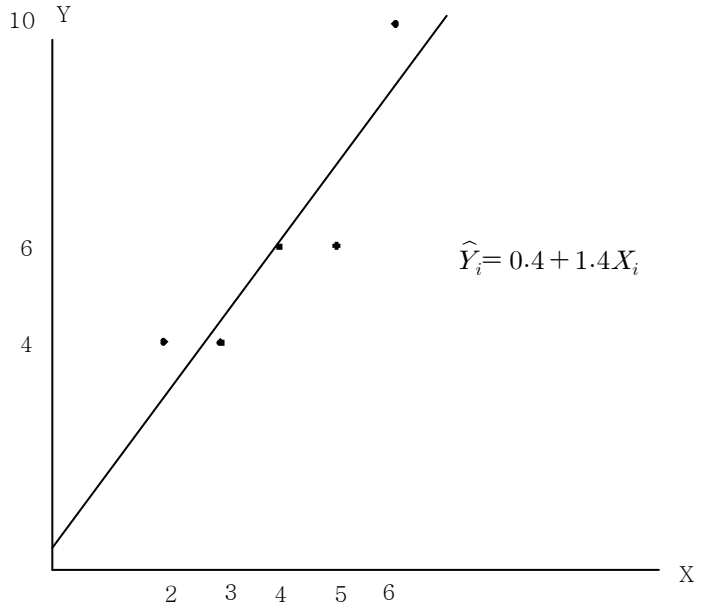
$$\therefore\ \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X} = 6 - (1.4)(4) = 0.4$$

따라서 추정된 회귀식은 다음과 같이 나타낼 수 있다.

$$\widehat{Y_i} = 0.4 + 1.4X_i$$

추정된 회귀식에 대한 해석은 다음과 같다. 추정된 회귀계수가 두 개 있지만 보통 기울기를 나타내는 $\widehat{\beta_1}$에 대한 해석에 관심을 가지고 있다. 이에 대한 해석은 독립변수인 X의 값이 1단위 증가함에 따라 종속변수 Y 값은 평균적으로 1.4단위 증가한다. 예를 들어 독립변수 X가 홍보비 지출액(단위: 천만 원)이고 종속변수가 연간매출액(단위: 억 원)이라면 홍보비 지출액이 천만 원 증가하면 연간매출액은 평균적으로 1.4억 원 증가한다고 할 수 있다.

<그림 3-5> 예제 2-1의 추정회귀선



3.OLS 추정량의 특성

보통최소자승법으로 추정한 OLS 추정량(회귀계수)의 특성에 대해 살펴보자. 일반적으로 좋은 추정량이 갖추어야 할 특성으로는 불변성(unbiasedness), 효율성(efficiency), 일치성(consistency) 등이 있다.

첫째, 불편성이란 추정량 $\hat{\theta}$의 기댓값이 추정 대상이 되는 모수 θ와 일치하는 경우 즉, $E(\widehat{\theta}) = \theta$이 성립하는 경우 $\hat{\theta}$은 θ의 불편추정량(unbiased estimator)이다.

둘째, 모수 θ에 대한 두 개의 불편추정량 $\widehat{\theta_1}, \widehat{\theta_2}$에 대해서 $\widehat{\theta_1}$의 분산이 $\widehat{\theta_2}$의 분산보다 작을 경우 $\widehat{\theta_1}$은 $\widehat{\theta_2}$보다 효율적인 추정량이라고 하고 두 추정량의 분산비율 즉 $\frac{var(\widehat{\theta_1})}{var(\widehat{\theta_2})}$을 상대적 효율성(relative efficiency)이라고 한다.

셋째, 확률변수 $\widehat{\theta_n}$을 n개의 자료로부터 도출된 모수 θ에 대한 추정량이라고 할 때, 자료의 수 n이 증가함에 따라 추정량 $\widehat{\theta_n}$이 모수 θ로부터 벗어날 확률이 점점 작아져 $\widehat{\theta_n}$의 확률극한(probability limit)이 추정 대상이 되는 모수 θ와 일치하는 경우 $\widehat{\theta_n}$은 θ에 대한 일치추정량(consistent estimator)이라고 하고 수식으로 $plim(\hat{\theta}) = \theta$ 또는 $\hat{\theta} \xrightarrow{p} \theta$로 나타낸다.

그러면 OLS 추정량인 $\widehat{\beta_0}, \widehat{\beta_1}$은 어떠한 추정량인가? 이에 대한 대답은 $\widehat{\beta_0}, \widehat{\beta_1}$은 최량선형불편추정량(Best Linear Unbiased Estimator: BLUE)이라는 것인데 회귀계수에 대한 OLS 추정량의 선형성, 불편성 및 최소분산에 대해 살펴보자.

(참고 2)기대연산자(Expectation Operator) 및 분산연산자(Variance Operator)

1.기댓값(평균)을 계산하는 기대연산자는 다음과 같은 특성을 가지고 있다.
① $E(c) = c$
② $E(X \pm c) = E(X) \pm c$
③ $E(cX) = cE(X)$
④ $E(a \pm bX) = a \pm bE(X)$

2.분산을 계산하는 분산연산자는 다음과 같은 특성을 가지고 있다.
① $var(c) = 0$
② $var(X) = E(X^2) - [E(X)]^2$
③ $var(cX) = c^2 var(X)$
④ $var(X \pm c) = var(X)$

(1)선형성

회귀계수의 선형성(linearity in Y)에 대해 살펴보기 위해 (3.12)식에서 다음의 (3.23)식을 도출한다.

$$\hat{\beta_1} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (Y_i - \overline{Y})}{\sum x_i^2} = \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\overline{Y}\sum x_i}{\sum x_i^2} = \sum w_i Y_i \quad (3.23)$$

단, $w_i = \frac{x_i}{\sum x_i^2}$ 이다.

또한 (3.23)식의 결과와 $\hat{\beta_0}$에 대한 추정량을 결합하면 (3.24)식과 같이 나타낼 수 있다.

$$\begin{aligned} \hat{\beta_0} &= \overline{Y} - \hat{\beta_1}\overline{X} \\ &= \overline{Y} - (\sum w_i Y_i)\overline{X} \\ &= \frac{1}{n}\sum Y_i - (\sum w_i Y_i)\overline{X} \\ &= \sum (\frac{1}{n} - w_i \overline{X}) Y_i \end{aligned} \quad (3.24)$$

(3.23)식과 (3.24)식에서 보여주고 있듯이 $\hat{\beta_0}, \hat{\beta_1}$은 Y에 대해서 선형이므로 $\hat{\beta_0}, \hat{\beta_1}$은 선형추정량(linear estimator)이다.

(2)불편성

회귀계수의 불편성(unbiasedness)에 대해 살펴보기 위해 먼저 다음과 같은 (3.25)식과 (3.26)식을 도출해 보자.

$$\sum w_i = \sum (\frac{x_i}{\sum x_i^2}) = \frac{\sum x_i}{\sum x_i^2} = 0 \quad (3.25)$$

$$\sum w_i X_i = \sum w_i = \sum w_i x_i + \overline{X}\sum w_i = \sum w_i x_i = \frac{\sum x_i^2}{\sum x_i^2} = 1 \quad (3.26)$$

$\hat{\beta_1}$의 불편성을 살펴보기 위해 (3.23)식으로부터 다음의 (3.27)식을 도출할 수 있다.

$$\begin{aligned}\hat{\beta}_1 &= \sum w_i Y_i \\ &= \sum w_i(\beta_0 + \beta_1 X_i + u_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i X_i + \sum w_i u_i \\ &= \beta_1 + \sum w_i u_i \quad (\because \sum w_i = 0, \sum w_i X_i = 1)\end{aligned} \tag{3.27}$$

(3.27)식에 기댓값을 계산하면 (3.28)식과 같게 되고 $\hat{\beta}_1$은 β_1의 불편추정량(unbiased estimator)이다.

$$\begin{aligned}E(\hat{\beta}_1) &= \beta_1 + E(\sum w_i u_i) \\ &= \beta_1 + \sum w_i E(u_i) \\ &= \hat{\beta}_1\end{aligned} \tag{3.28}$$

또한 $\hat{\beta}_0$의 불편성을 살펴보기 위해 (3.14)식으로부터 다음의 (3.29)식이 도출된다.

$$\begin{aligned}\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \\ &= \beta_0 + \beta_1 \overline{X} + \overline{u} - \hat{\beta}_1 \overline{X} \\ &= \beta_0 + \beta_1 \overline{X} + \overline{u} - (\beta_1 + \sum w_i u_i)\overline{X} \\ &= \beta_0 + \overline{u} - \overline{X} \sum w_i u_i\end{aligned} \tag{3.29}$$

(3.34)식에 기댓값을 계산하면 (3.30)식과 같게 되고 $\hat{\beta}_0$은 β_0의 불편추정량(unbiased estimator)이다.

$$\begin{aligned}E(\hat{\beta}_0) &= \beta_0 + E(\overline{u}) - \overline{X} \sum w_i E(u_i) \\ &= \hat{\beta}_0 \quad (\because E(\overline{u}) = 0, E(u_i = 0))\end{aligned} \tag{3.30}$$

(3)최소분산

회귀계수 $\hat{\beta}_1$의 최소분산(minimum variance)을 살펴보기 위해 위의 (3.23)식을

활용해 보자. 즉,

$$\widehat{\beta_1} = \sum w_i Y_i$$

단, $w_i = \frac{X_i}{\sum X_i^2}$ 이다.

다른 선형추정량 $\widetilde{\beta_1}$을 다음의 (3.31)식과 같다고 하자.

$$\widetilde{\beta_1} = \sum c_i Y_i \tag{3.31}$$

여기서 c_i는 w_i와 같을 수도 있고 같지 않을 수도 있다. 즉, $c_i = w_i$이면 $\widehat{\beta_1} = \widetilde{\beta_1}$이고, $c_i \neq w_i$이면 $\widehat{\beta_1} \neq \widetilde{\beta_1}$이다.

$$\begin{aligned} E(\widetilde{\beta_1}) &= \sum c_i E(Y_i) \\ &= \sum c_i E(\beta_0 + \beta_1 X_i + u_i) \\ &= \beta_0 \sum c_i + \beta_1 \sum c_i X_i \end{aligned}$$

따라서 $\widetilde{\beta_1}$가 불편추정량이 되기 위해서는 $\sum c_i = 0, \sum c_i X_i = 1$이 되어야 한다. 한편, $\widetilde{\beta_1}$의 분산을 계산해 보면 (3.32)식과 같게 된다.

$$\begin{aligned} Var(\widetilde{\beta_1}) &= Var(\sum c_i Y_i) \\ &= \sum c_i^2 Var(Y_i) \\ &= \sigma_u^2 \sum c_i^2 \\ &= \sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2} + \frac{x_i}{\sum x_i^2})^2 \\ &= \sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2})^2 + \sigma_u^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} + 2\sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2})(\frac{x_i}{\sum x_i^2}) \\ &= \sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2})^2 + \sigma_u^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} + 2\sigma_u^2 \frac{\sum c_i x_i}{\sum x_i^2} - 2\sigma_u^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} \end{aligned}$$

$$= \sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2})^2 + 2\sigma_u^2 \frac{1}{\sum x_i^2} - \sigma_u^2 \frac{1}{\sum x_i^2}$$
$$= \sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2})^2 + \sigma_u^2 \frac{1}{\sum x_i^2}$$
$$= \sigma_u^2 \sum (c_i - w_i)^2 + \sigma_u^2 \frac{1}{\sum x_i^2} \qquad (3.32)$$

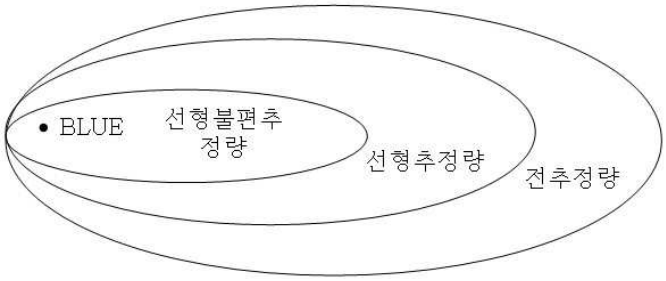
(3.32)식에서 $c_i = w$이면 $var(\widetilde{\beta_1}) = var(\widehat{\beta_1})$이고, $c_i \neq w_i$이면 $var(\widetilde{\beta_1}) \geq var(\widehat{\beta_1})$인데 그 이유는 $\sigma_u^2 \sum (c_i - \frac{x_i}{\sum x_i^2})^2 \geq 0$이기 때문이다. 따라서 $\widehat{\beta_1}$은 다른 어떤 선형불편추정량보다 분산이 작게 되는데 이러한 특성을 가진 OLS 추정량을 최량선형불편추정량(Best Linear Unbiased Estimator: BLUE)이라고 하며 이를 요약한 것이 Gauss-Markov 정리이다.

Gauss-Markov 정리

고전적 회귀모형이 있을 때 최소자승법으로 구한 추정량 $\widehat{\beta_0}$ 및 $\widehat{\beta_1}$는 선형불편추정량 중에서 최소의 분산을 갖는다. 즉, 최소자승법으로 구한 추정량은 BLUE(Best Linear Unbiased Estimator)이다

이를 집합개념으로 설명해 보면 <그림 3-6>과 같다. 모수를 추정하는 공식을 나타내는 추정량의 종류는 많이 있다. 이를 전추정량의 집합이라고 하자. 이들 추정량 중에는 선형추정량이 있고 비선형추정량이 있는데 그 중에서 선형추정량만을 선택하면 이들은 전추정량의 부분집합이 된다. 선형추정량 중에는 불편추정량이 있고 편의를 가진 추정량이 있는데 그 중에서 불편추정량만을 선택하면 선택된 선형불편추정량은 전추정량 및 선형추정량의 부분집합이 된다. 선형불편추정량 중에서 분산의 크기를 비교해 보면 그 중 가장 작은 분산을 가진 추정량이 있는데 이를 최량선형불편추정량(BLUE)라고 한다.

<그림 3-6> 추정량의 집합관계



4.회귀계수의 분산 추정

회귀계수에 대한 통계적 검정을 하기 위해서는 회귀계수의 분산을 추정해야 한다. 그러나 회귀계수의 분산 추정량은 교란항의 분산 추정량의 함수이기 때문에 회귀계수의 분산을 추정하기 위해서는 먼저 교란항의 분산인 σ_u^2을 추정한다.

교란항의 분산 σ_u^2에 대한 좋은 추정량(불편추정량)은 $\sum e_i^2$을 자유도인 n-2로 나눈 것 즉, $\widehat{\sigma_u^2}=\frac{\sum e_i^2}{n-2}$이다. 여기서 2는 단순회귀모형에서 상수항을 포함한 독립변수의 수이다. 이것은 단순회귀분석에서 성립하고 다중회귀분석에서는 $\widehat{\sigma_u^2}=\frac{\sum e_i^2}{n-k}$로서 k는 다중회귀모형에서 상수항을 포함한 독립변수의 수이다.

교란항의 분산 σ_u^2에 대한 추정량 $\widehat{\sigma_u^2}=\frac{\sum e_i^2}{n-2}$이 불편추정량이 되기 위해서는 $E(\widehat{\sigma_u^2})=E(\frac{\sum e_i^2}{n-2})=\sigma_u^2$이 되어야 하므로 $E(\sum e_i^2)=(n-2)\sigma_u^2$임을 다음과 같이 보여주면 된다.

먼저 (3.2)식 즉 $Y_i=\beta_0+\beta_1 X_i+u_i$의 평균을 구하면 $\overline{Y}=\beta_0+\beta_1\overline{X}+\overline{u}$이 되는데 (3.2)식과 (3.2)식의 평균과의 차이를 구하면 (3.33)식과 같다.

$$\begin{aligned} Y_i-\overline{Y} &= \beta_1(X_i-\overline{X})+u_i-\overline{u} \\ &= \beta_1 x_i+u_i-\overline{u} \end{aligned} \qquad (3.33)$$

<그림 3-7>에서 나타내고 있듯이 $e_i = y_i - \hat{y_i}$이므로 이는 다음의 (3.34)식과 같게 된다.

$$\begin{aligned} e_i &= y_i - \hat{y_i} \\ &= Y_i - \overline{Y} - \hat{\beta}_1 x_i \\ &= \beta_1 x_i + u_i - \overline{u} - \hat{\beta}_1 x_i \\ &= -(\hat{\beta}_1 - \beta_1)x_i + (u_i - \overline{u}) \end{aligned} \qquad (3.34)$$

(3.34)식의 양변을 제곱한 후 더하면 (3.35)식과 같게 된다.

$$\sum e_i^2 = (\hat{\beta}_1 - \beta_1)^2 \sum x_i^2 + \sum (u_i - \overline{u})^2 - 2(\hat{\beta}_1 - \beta_1)\sum x_i (u_i - \overline{u}) \qquad (3.35)$$

(3.35)식의 양변의 기댓값을 구하면 (3.36)과 같다.

$$E(\sum e_i^2) = \underbrace{E[(\hat{\beta}_1 - \beta_1)^2 \sum x_i^2]}_{①} + \underbrace{E[\sum (u_i - \overline{u})^2]}_{②} - 2\underbrace{E[(\hat{\beta}_1 - \beta_1)\sum x_i (u_i - \overline{u})]}_{③} \qquad (3.36)$$

(3.36)식을 ①-③의 세 부분으로 나누어 기댓값을 각각 구해 보면 (3.37)-(3.39)식과 같게 된다.

$$\begin{aligned} ① \rightarrow E[(\hat{\beta}_1 - \beta_1)^2 \sum x_i^2] &= \sum x_i^2 E(\hat{\beta}_1 - \beta_1)^2 \\ &= \sum x_i^2 var(\hat{\beta}_1) \\ &= \sigma_u^2 \end{aligned} \qquad (3.37)$$

$$\begin{aligned} ② \rightarrow E[\sum (u_i - \overline{u})^2] &= E[\sum u_i^2 - \frac{2}{n}(\sum u_i)^2 + n(\frac{\sum u_i}{n})^2] \\ &= E[\sum u_i^2 - \frac{1}{n}(\sum u_i)^2] \\ &= E(\sum u_i^2) - \frac{1}{n} E[(\sum u_i)^2] \end{aligned}$$

$$= E(\sum u_i^2) - \frac{1}{n}\sum E(u_i^2)$$
$$= n\sigma_u^2 - \frac{1}{n}n\sigma_u^2$$
$$= (n-1)\sigma_u^2 \qquad (3.38)$$

③ $\rightarrow E[(\hat{\beta}_1-\beta_1)\sum x_i(u_i-\bar{u})] = E[(\hat{\beta}_1-\beta_1)(\sum x_i u_i - \bar{u}\sum x_i)]$

$$= E[(\beta_1 + \frac{\sum x_i u_i}{\sum x_i^2} - \beta_1)(\sum x_i u_i - \bar{u}\sum x_i)]$$
$$= E[\frac{\sum x_i u_i}{\sum x_i^2}(\sum x_i u_i - \bar{u}\sum x_i)]$$
$$= E[\frac{(\sum x_i u_i)^2}{\sum x_i^2}]$$
$$= E[\frac{\sum x_i^2 u_i^2}{\sum x_i^2}]$$
$$= \sigma_u^2 \qquad (3.39)$$

따라서 (3-37)-(3.39)의 세 식을 더해 보면 (3.40)식과 같이 된다.

$$\therefore E(\sum e_i^2) = \sigma_u^2 + (n-1)\sigma_u^2 - 2\sigma_u^2 = (n-2)\sigma_u^2 \qquad (3.40)$$

한편, 잔차의 제곱의 합은 다음의 (3.41)식에 의해 구할 수 있다.

$$\sum e_i^2 = \sum(Y_i - \widehat{Y_i})^2$$
$$= \sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$
$$= \sum(Y_i - \overline{Y} + \hat{\beta}_1\overline{X} - \hat{\beta}_1 X_i)^2$$
$$= \sum[(Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})]^2$$
$$= \sum[(Y_i - \overline{Y})^2 - 2\hat{\beta}_1(X_i - \overline{X})(Y_i - \overline{Y}) + \hat{\beta}_1^2(X_i - \overline{X})^2]$$
$$= \sum(Y_i - \overline{Y})^2 - 2\hat{\beta}_1\sum(X_i - \overline{X})(Y_i - \overline{Y}) + \hat{\beta}_1^2\sum(X_i - \overline{X})^2$$

$$= \sum(Y_i - \overline{Y})^2 - 2\hat{\beta}_1 \sum(X_i - \overline{X})(Y_i - \overline{Y}) + \hat{\beta}_1 \sum(X_i - \overline{X})(Y_i - \overline{Y})$$
$$= \sum(Y_i - \overline{Y})^2 - \hat{\beta}_1 \sum(X_i - \overline{X})(Y_i - \overline{Y}) \quad (3.41)$$

교란항의 분산이 추정되면 단순회귀분석의 추정에서 마지막으로 해야 할 일은 회귀계수의 분산을 추정하는 것이다.

먼저 $\hat{\beta}_1$의 분산을 추정해 보자. (3.27)식에서 다음의 (3.42)식을 구할 수 있다.

$$\hat{\beta}_1 - \beta_1 = \sum w_i u_i \quad (3.42)$$

(3.42)식을 이용하여 (3.43)식의 회귀계수 $\hat{\beta}_1$의 분산을 구할 수 있다.

$$\begin{aligned}
Var(\hat{\beta}_1) &= E[(\hat{\beta}_1 - \beta_1)^2] \\
&= E[(\sum w_i u_i)^2] \\
&= E[(w_1 u_1 + w_2 u_2 + ... + w_n u_n)^2] \\
&= E(w_1^2 u_1^2 + w_2^2 u_2^2 + ... + w_n^2 u_n^2 + 2w_1 w_2 u_1 u_2 + ... + 2w_{n-1} w_n u_{n-1} u_n) \\
&= E(w_1^2 u_1^2 + w_2^2 u_2^2 + ... + w_n^2 u_n^2) \quad (\because E(u_i u_j) = 0 \ (\text{for } i \neq j) \\
&= w_1^2 E(u_1^2) + w_2^2 E(u_2^2) + ... + w_n^2 E(u_n^2) \\
&= w_1^2 \sigma_u^2 + w_2^2 \sigma_u^2 + ... + w_n^2 \sigma_u^2 \\
&= \sigma_u^2 \sum w_i^2 \\
&= \sigma_u^2 \sum (\frac{x_i}{\sum x_i^2})^2 \\
&= \sigma_u^2 \frac{1}{\sum x_i^2} \quad (3.43)
\end{aligned}$$

한편, (3.29)식을 이용하여 $\hat{\beta}_0$의 분산을 추정할 수 있다. (3.29)식에서 다음의 (3.44)식을 구할 수 있다.

$$\begin{aligned}\hat{\beta}_0-\beta_0 &= \bar{u}-\bar{X}\sum w_i u_i \\ &= \frac{1}{n}\sum u_i-\bar{X}\sum w_i u_i \\ &= \sum(\frac{1}{n}-\bar{X}w_i)u_i \qquad (3.44)\end{aligned}$$

(3.44)식을 이용하여 (3.45)식의 회귀계수 $\hat{\beta}_0$의 분산을 구할 수 있다.

$$\begin{aligned}Var(\hat{\beta}_0) &= E[(\hat{\beta}_0-\beta_0)^2] \\ &= E[(\sum(\frac{1}{n}-\bar{X}w_i)u_i)^2] \\ &= E[((\frac{1}{n}-\bar{X}w_1)u_1+...+(\frac{1}{n}-\bar{X}w_n)u_n)^2] \\ &= E[(\frac{1}{n}-\bar{X}w_1)^2u_1^2+...+(\frac{1}{n}-\bar{X}w_n)^2u_n^2] \quad (\because E(u_iu_j)=0\ ) \\ &= \sigma_u^2[(\frac{1}{n}-\bar{X}w_1)^2+...+(\frac{1}{n}-\bar{X}w_n)^2] \\ &= \sigma_u^2[(\frac{1}{n^2}-2\bar{X}\frac{w_1}{n}+\overline{X^2}w_1^2)+..+(\frac{1}{n^2}-2\bar{X}\frac{w_n}{n}+\overline{X^2}w_n^2)] \\ &= \sigma_u^2[\frac{n}{n^2}-2\frac{\bar{X}}{n}\sum w_i+\overline{X^2}\sum w_i^2] \\ &= \sigma_u^2[\frac{n}{n^2}+\overline{X^2}\sum w_i^2] \\ &= \sigma_u^2(\frac{1}{n}+\overline{X^2}\sum(\frac{x_i}{\sum x_i^2})^2) \\ &= \sigma_u^2(\frac{1}{n}+\overline{X^2}\frac{1}{\sum x_i^2})\end{aligned}$$

$$= \sigma_u^2 (\frac{\sum (X_i - \overline{X})^2 + n\overline{X^2}}{n\sum x_i^2})$$

$$= \sigma_u^2 (\frac{\sum X_i^2 - 2n\overline{X^2} + n\overline{X^2} + n\overline{X^2}}{n\sum x_i^2})$$

$$= \sigma_u^2 (\frac{\sum X_i^2}{n\sum x_i^2}) \qquad (3.45)$$

지금까지 회귀계수 $\hat{\beta_0}, \hat{\beta_1}$의 추정과 회귀계수 분산의 추정에 대해 살펴보았는데 결론적으로 회귀계수는 각각 다음의 분포를 따른다.

$$\hat{\beta_0} \sim \ (\beta_0, \sigma_u^2 \frac{\sum X_i^2}{n\sum x_i^2})$$

$$\hat{\beta_1} \sim \ (\beta_1, \sigma_u^2 \frac{1}{\sum x_i^2})$$

(예제 3-1)(계속)
다음의 홍보비 지출액 X(단위: 천만 원)와 연간매출액 Y(단위: 억 원)에 관한 자료를 이용하여 교란항의 분산 및 회귀계수의 분산을 구하라.

X	2	3	4	5	6
Y	4	4	6	6	10

(풀이)
먼저 주어진 자료를 이용하여 OLS 추정량을 구하기 위한 기초계산을 하면 다음과 같다.

$\sum X_i = 20$, $\sum Y_i = 30$, $\sum X_i Y_i = 134$, $\overline{X} = 4$, $\overline{Y} = 6$, $\sum X_i^2 = 90$, $\sum Y_i^2 = 204$
$\sum (X_i - \overline{X})^2 = \sum X_i^2 - \overline{X}\sum X_i = 90 - (4)(20) = 10$

$$\sum(Y_i-\overline{Y})^2=\sum Y_i^2-\overline{Y}\sum Y_i=204-(6)(30)=204-180=24$$

$$\sum(X_i-\overline{X})(Y_i-\overline{Y})=\sum X_iY_i-\overline{X}\sum Y_i=134-(4)(30)=14$$

$$\sum e_i^2=\sum(Y_i-\overline{Y})^2-\widehat{\beta_1}\sum(X_i-\overline{X})(Y_i-\overline{Y})=24-(1.4)(14)=4.4$$

$$\therefore\ \widehat{\sigma_u^2}=\frac{4.4}{3}=1.4667$$

$$var(\widehat{\beta_0})=\widehat{\sigma_u^2}\frac{\sum X_i^2}{n\sum(X_i-\overline{X})^2}=(1.4667)\frac{90}{(5)(10)}=2.64006$$

$$var(\widehat{\beta_1})=\widehat{\sigma_u^2}\frac{1}{\sum(X_i-\overline{X})^2}=(1.4667)\frac{1}{10}=0.14667$$

5.결정계수

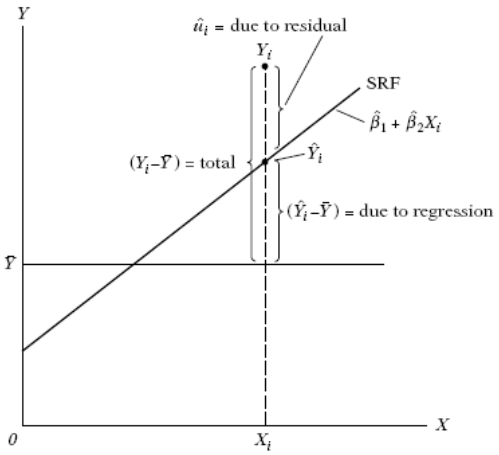
표본으로부터 추정한 회귀선이 변수의 표본관측에 얼마나 적합한 지를 측정하는 계수를 결정계수(coefficient of determination : R^2) 또는 설명력(explanatory power)이라 한다. 결정계수는 표본관측이 추정된 회귀식에 가까운 정도를 계수로 나타낸 것으로 종속변수의 전변동과 회귀변동의 비율로 측정한다. 따라서 결정계수의 값은 0과 1 사이로, 큰 값일수록 적합도 또는 설명력이 높다는 것을 의미하고 작은 값일수록 적합도 또는 설명력이 낮다는 것을 의미한다.

<그림 3-7>에서 보면 다음의 (3.46)식이 성립한다.

$$Y_i=\widehat{Y_i}+e_i$$

또는 $Y_i-\overline{Y}=\widehat{Y_i}-\overline{Y}+e_i$ (3.46)

<그림 3-7> Y의 변동



(3.46)식의 양변을 제곱한 후 정리하면 (3.47)식이 도출된다.

$$\begin{aligned}\sum(Y_i-\overline{Y})^2 &= \sum[(\widehat{Y}_i-\overline{Y})+e_i]^2 \\ &= \sum(\widehat{Y}_i-\overline{Y})^2+\sum e_i^2+2\sum(\widehat{Y}_i-\overline{Y})e_i \\ &= \sum(\widehat{Y}_i-\overline{Y})^2+\sum e_i^2 \qquad (3.47)\end{aligned}$$

$$\begin{aligned}(\because\ 2\sum(\widehat{Y}_i-\overline{Y})e_i &= 2\sum(\widehat{\beta}_0+\widehat{\beta}_1X_i-\overline{Y})e_i \\ &= 2\widehat{\beta}_0\sum e_i+2\widehat{\beta}_1\sum X_ie_i-2\overline{Y}\sum e_i \\ &= 2\widehat{\beta}_0\sum e_i+2\widehat{\beta}_1\sum X_ie_i-2\overline{Y}\sum e_i \\ &= 0)\end{aligned}$$

(3.47)식에서 $\sum(Y_i-\overline{Y})^2$을 총자승합(total sum of squares) 또는 총변동(Total variation), $\sum(\widehat{Y}_i-\overline{Y})^2$을 회귀자승합(regression sum of squares) 또는 회귀변동(regression variation), $\sum e_i^2$을 잔차자승합(error sum of squares) 또는 잔차변동(residual variation)이라고 하며, 동 식은 총변동은 회귀선의 추정으로 설명된 변동(explained variation)과 설명되지 않은 변동(unexplained variation)의 합 또는 모형에 의한 부분과 오차에 의한 부분의 합으로 분해될 수 있음을 나타내고 있다.

(3.41)식에서 $\sum e_i^2 = \sum(Y_i - \overline{Y})^2 - \hat{\beta}_1 \sum(X_i - \overline{X})(Y_i - \overline{Y})$인데 이를 (3.47)식과 비교해 보면 회귀변동은 $\hat{\beta}_1 \sum(X_i - \overline{X})(Y_i - \overline{Y}) = \hat{\beta}_1^2 \sum(X_i - \overline{X})^2$임을 알 수 있다.

따라서 결정계수는 전변동에 대한 회귀변동의 비율이라는 정의에 따라 (3.48)식과 같이 나타낼 수 있다.

$$R^2 = \frac{\text{회귀변동}}{\text{전변동}} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{\hat{\beta}_1^2 \sum(X_i - \overline{X})^2}{\sum(Y_i - \overline{Y})^2} \tag{3.48}$$

앞의 (예제 3-1)에서 결정계수를 구해 보면 $R^2 = \frac{\hat{\beta}_1^2 \sum(X_i - \overline{X})^2}{\sum(Y_i - \overline{Y})^2} = \frac{19.6}{24} = 0.817$이 된다.

6.가설검정

가설검정이란 모수에 대한 특정한 가설을 세워 놓고 표본에서 계산된 통계량을 기초로 하여 이 가설의 채택여부를 판단하는 것을 말한다. 가설은 경제이론 또는 경험적으로 도출될 수 있으며 귀무가설과 대립가설이 있다. 귀무가설(H_0)은 모집단의 모수가 어느 주어진 값과 같다는 가설이고 대립가설(H_1)은 귀무가설에 대한 대안이다.

회귀분석에서 회귀계수에 대한 추정도 중요하지만 추정된 회귀계수에 대한 유의성 검정도 매우 중요하다. 회귀모형의 설정에서 어떤 변수를 독립변수로 할 것인 지에 대한 고민은 경제이론에서 출발한다. 이론적으로 종속변수에 영향을 주는 변수를 독립변수로 선정해야 하는데 이를 독립변수의 경제적 유의성(economical significance)이라고 한다. 예를 들어 소비를 설명하는 단순회귀모형을 만든다고 할 때 여러 가지 변수들이 소비에 영향을 줄 수 있지만 소비이론에 따라 소득이 소비에 영향을 주는 가장 중요한 변수로 선정한다.

그러나 경제이론에 따라 선정된 즉 경제적 유의성을 가진 독립변수가 현실 경제에 있어서도 종속변수에 반드시 영향을 주는 것은 아니다. 실증분석 결과 실제로도 독립변수가 종속변수에 영향을 주는 것을 독립변수의 통계적 유의성(statistical significance)이라고 한다. 즉, 이론과 현실이 일치할 때는 종속변

수를 설명함에 있어 독립변수가 경제적으로 그리고 통계적으로 유의한(의미 있는) 변수가 되지만 이론과 현실이 일치하지 않을 때는 종속변수를 설명함에 있어 독립변수가 경제적으로는 유의한 변수가 되지만 통계적으로는 유의한 변수가 되지 못한다. 예를 들어 소비이론에 따를 경우 소비에 가장 중요한 변수는 소득이지만 현실에서는 소득이 소비에 영향을 주는 변수가 되지 못할 수도 있다.

따라서 독립변수의 통계적 유의성을 살펴보는 것은 매우 중요하고 의미 있는 작업인데 회귀계수가 독립변수와 종속변수의 관계를 나타내므로 회귀계수에 대한 검정을 통해 이 문제를 해결할 수 있다.

회귀계수에 대한 가설검정은 주어진 회귀모형으로부터 회귀계수에 대한 귀무가설을 설정한 후 유의성 검정을 한다. 귀무가설은 다양하게 설정할 수 있지만 일반적으로 회귀계수의 값이 0이라는 귀무가설을 설정하는데 이는 독립변수가 종속변수에 영향을 주지 못한다는 것이다.

(회귀모형) $Y_i = \beta_0 + \beta_1 X_i + u_i$

(가설) $H_0 : \beta_0 = 0 \quad H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

앞에서 회귀계수 및 회귀계수의 분산에 대한 추정을 통해 회귀계수가 각각 다음의 분포에 따름을 알 수 있었다.

$$\widehat{\beta_0} \sim \left(\beta_0, \sigma_u^2 \frac{\sum X_i^2}{n \sum x_i^2}\right)$$

$$\widehat{\beta_1} \sim \left(\beta_1, \sigma_u^2 \frac{1}{\sum x_i^2}\right)$$

따라서 회귀계수에 대한 다음의 t-통계량은 (n-2)의 자유도를 갖는 t-분포를 따른다.

$$t = \frac{\hat{\beta_0} - \beta_0}{\sqrt{Var(\hat{\beta_0})}} = \frac{\hat{\beta_0} - \beta_0}{se(\beta_0)} \sim t_{n-2}$$

$$t = \frac{\hat{\beta_1} - \beta_1}{\sqrt{Var(\hat{\beta_1})}} = \frac{\hat{\beta_1} - \beta_1}{se(\hat{\beta_1})} \sim t_{n-2}$$

주어진 유의수준 하에서 각각의 회귀계수에 대한 유의성 검정은 위의 t-통계량의 값이 임계값(critical value) $t_{(n-2,\frac{\alpha}{2})}$보다 크거나 $-t_{(n-2,\frac{\alpha}{2})}$보다 작으면 귀무가설을 기각하고, t-통계량의 값이 임계값인 $-t_{(n-2,\frac{\alpha}{2})}$과 $t_{(n-2,\frac{\alpha}{2})}$사이에 있으면 귀무가설을 허용한다(또는 기각하지 못한다).

가설검정의 접근법에는 유의성 검정 외에 신뢰구간을 이용하는 방법도 있다. 위의 t-통계량에서 $(1-\alpha)$의 신뢰수준으로 회귀계수에 대한 신뢰구간은 각각 다음과 같다.

β_0에 대한 $(1-\alpha)\times 100\%$ 신뢰구간 : $\hat{\beta_0} \pm t_{(n-2,\frac{\alpha}{2})} se(\hat{\beta_0})$

β_1에 대한 $(1-\alpha)\times 100\%$ 신뢰구간 : $\hat{\beta_1} \pm t_{(n-2,\frac{\alpha}{2})} se(\hat{\beta_1})$

(예제 3-1)(계속)
다음의 홍보비 지출액 X(단위: 천만 원)와 연간매출액 Y(단위: 억 원)에 관한 자료를 이용하여 홍보비 지출액의 통계적 유의성을 검정하고 홍보비 지출액에 대한 98% 신뢰구간을 구하라.

X	2	3	4	5	6
Y	4	4	6	6	10

회귀계수 및 회귀계수의 분산에 대한 추정결과 추정된 회귀식은 다음과 같이 쓸 수 있는데 각 회귀계수 밑에 표시된 괄호안의 값은 추정된 회귀계수의 표준편차(이를 표준오차라고 함)를 나타낸다.

$$\hat{Y} = 0.4 + 1.4X$$
$$= (1.62)(0.38)$$

먼저 홍보비 지출액이 연간매출액에 영향을 주지 않는다는 가설을 검정하기 위하여 다음과 같은 귀무가설 및 대립가설을 설정한다.

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

t-통계량을 계산해 보면 다음과 같이 3.655가 계산이 된다.

$$t = \frac{1.4 - 0}{0.383} = 3.655$$

다음으로 5% 유의수준($\alpha = 0.05$)에서 자유도가 3일 때 임계값은 3.182 즉, $t_{3,0.025} = 3.182$이다. 위에서 계산한 t-통계량의 값 3.655는 임계값 3.182보다 크므로 5% 유의수준 하에서 귀무가설을 기각한다. 따라서 홍보비 지출액은 연간매출액에 영향을 준다고 해석하고 $\hat{\beta_1}$는 통계적으로 유의하다(statistically significant)라고 한다.

한편, β_1에 대한 95% 신뢰구간은 다음과 같이 계산된다.

$$\hat{\beta_1} \pm t_{(n-2,\frac{\alpha}{2})} se(\hat{\beta_1}) = 1.4 \pm 3.182(0.383) = [0.181, 2.618]$$

이와 같이 계산된 신뢰구간의 의미는 여러 표본을 이용하여 β_1에 대한 추정을 계속하여 반복하면 β_1의 추정치 즉, $\hat{\beta_1}$들 중 95%가 0.181과 2.618 사이에 놓이게 된다는 것이다.

7.예측

모형설정→모형추정→가설검정→예측이라는 단순회귀분석의 절차에 따라 주어진 독립변수의 값에 대한 종속변수의 값을 구하는 예측(forecasting, prediction)을 살펴보기로 하자. 일반적으로 예측에는 주어진 X에 대해 하나의 Y

값을 구하는 점예측(point forecasting)과 점예측에 대한 신뢰구간을 구하는 구간예측(interval forecasting)이 있다.

회귀분석의 예측에는 독립변수의 값이 주어졌을 때 모집단회귀선 위의 한 점($E(Y|X_0)$)을 예측하는 평균예측(mean prediction)과 주어진 독립변수의 값에 대응하는 개별 Y의 값(Y_0)을 예측하는 개별예측(individual prediction)이 있다. 따라서 평균예측과 개별예측에서 각각의 점예측치 및 예측구간대를 구해보도록 하자.

(1)점예측

점예측에 대해 살펴보면 $X_i = X_0$일 때, $\widehat{Y_0}$은 평균예측 $E(Y|X_0)$ 및 개별예측 Y_0의 선형불편추정량이 되고 따라서 점예측치가 되는데 이를 살펴보도록 하자. (3.3)식과 제2장의 (2.1)식의 회귀모형에서 독립변수 X 가 특정한 값 X_0일 때 Y의 평균에 대한 예측치 및 개별 Y에 대한 예측치는 모두 $\widehat{Y_0}$이다. 만약에 $E(\widehat{Y_0}) - Y_0 = 0$이 성립하면 $\widehat{Y_0}$는 실제치 Y_0의 선형불편추정량이 되는데 다음의 (3.49)식이 나타내고 있듯이 $E(\widehat{Y_0}) - Y_0 = 0$이 성립하므로 즉, $\widehat{Y_0}$ 는 Y_0의 선형불편추정량이 되고 따라서 점예측치가 된다.

$$\begin{aligned} E[Y_0 - \widehat{Y_0}] &= E[\beta_0 + \beta_1 X_0 + u_0 - \widehat{\beta_0} - \widehat{\beta_1} X_0)] \\ &= \beta_0 + \beta_1 X_0 + E(u_0) - \beta_0 - \beta_1 X_0 = 0 \\ &= 0 \end{aligned} \tag{3.49}$$

(2)구간예측

구간예측을 위해서는 실제치인 Y_0와 예측치 $\widehat{Y_0}$의 차이인 예측오차의 분산(σ_ϵ^2)을 알아야 하는데 예측오차의 분산은 평균예측과 개별예측이 서로 다르다. 먼저 개별예측의 예측오차 분산을 계산해 보자. (3.3)식에서 X 가 X_0일 때 개별 Y에 대한 예측치는 $\widehat{Y_0}$이고 따라서 예측오차는 $Y_0 - \widehat{Y_0}$이다. 따라서 개별예측에 대한 예측오차의 분산은 다음의 (3.50)식과 같다.

$$
\begin{aligned}
\sigma_{\varepsilon}^{2} &= E[(Y_0 - \widehat{Y_0}) - E(Y_0 - \widehat{Y_0})]^2 \\
&= E[Y_0 - \widehat{Y_0}]^2 \\
&= E[(Y_0 - E(Y_0) + (E(Y_0) - \widehat{Y_0})]^2 \\
&= E[(Y_0 - E(Y_0)]^2 + E[(E(Y_0) - \widehat{Y_0})]^2 + 2E[(Y_0 - E(Y_0)(E(Y_0) - \widehat{Y_0})] \\
&= E(u_0)^2 + E[-(\widehat{\beta_0} - \beta_0) - (\widehat{\beta_1} - \beta_1)X_0]^2 + 2E(u_0)(\beta_0 + \beta_1 X_0 - \widehat{\beta_0} - \widehat{\beta_1} X_0) \\
&= E(u_0)^2 + E[(\widehat{\beta_0} - \beta_0)^2] + E[(\widehat{\beta_1} - \beta_1)^2 X_0^2] + 2X_0 E[(\widehat{\beta_0} - \beta_0)(\widehat{\beta_1} - \beta_1)] \\
&= \sigma_u^2 + var(\widehat{\beta_0}) + X_0^2 var(\widehat{\beta_1}) + 2X_0 cov(\widehat{\beta_0}, \widehat{\beta_1}) \\
&= \sigma_u^2 + \sigma_u^2 (\frac{1}{n} + \overline{X^2} \frac{1}{\sum x_i^2}) \; + \; X_0^2 \sigma_u^2 \frac{1}{\sum x_i^2} \; - \; 2X_0 \frac{\overline{X} \sigma_u^2}{\sum x_i^2} \\
&= \sigma_u^2 (1 + \frac{1}{n} + \frac{1}{\sum x_i^2} (\overline{X^2} + X_0^2 - 2X_0 \overline{X})) \\
&= \sigma_u^2 (1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}) \qquad (3.50)
\end{aligned}
$$

(참고)회귀계수의 공분산

$$
\begin{aligned}
cov(\widehat{\beta_0}, \widehat{\beta_1}) &= E[(\widehat{\beta_0} - \beta_0)(\widehat{\beta_1} - \beta_1)] \\
&= E[\sum (\frac{1}{n} - \overline{X} w_i) u_i (\sum w_i u_i)] \\
&= E[(\frac{1}{n} \sum u_i - \overline{X} \sum w_i u_i (\sum w_i u_i)] \\
&= E[-\overline{X} (\sum w_i u_i)^2] \\
&= \frac{-\overline{X} \sigma_u^2}{\sum x_i^2}
\end{aligned}
$$

다음으로 평균예측의 예측오차 분산을 계산해 보자. 제2장의 (2.1)식에서 X 가

X_0일 때 평균 Y에 대한 예측치는 $\widehat{Y_0}$이고 따라서 예측오차는 $E(Y_0)-\widehat{Y_0}$이다. 따라서 평균예측에 대한 예측오차의 분산은 다음의 (3.51)식과 같다.

$$
\begin{aligned}
\sigma_{\varepsilon}^{2} &= E[(E(Y_0)-\widehat{Y_0})-E(E(Y_0)-\widehat{Y_0})]^2 \\
&= E[E(Y_0)-\widehat{Y_0}]^2 \\
&= E[-(\widehat{\beta_0}-\beta_0)-(\widehat{\beta_1}-\beta_1)X_0]^2 \\
&= E[(\widehat{\beta_0}-\beta_0)^2]+E[(\widehat{\beta_1}-\beta_1)^2X_0^2]+2X_0E[(\widehat{\beta_0}-\beta_0)(\widehat{\beta_1}-\beta_1)] \\
&= var(\widehat{\beta_0})+X_0^2var(\widehat{\beta_1})+2X_0cov(\widehat{\beta_0},\widehat{\beta_1}) \\
&= \sigma_u^2(\frac{1}{n}+\overline{X^2}\frac{1}{\sum x_i^2}) \; + \; X_0^2\sigma_u^2\frac{1}{\sum x_i^2} \; - \; 2X_0\frac{\overline{X}\sigma_u^2}{\sum x_i^2} \\
&= \sigma_u^2(\frac{1}{n}+\frac{1}{\sum x_i^2}(\overline{X^2}+X_0^2-2X_0\overline{X})) \\
&= \sigma_u^2(\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2}) \qquad (3.51)
\end{aligned}
$$

(3.50)식 및 (3.51)식의 예측오차 분산을 구하는 다음의 식으로부터 몇 가지 사실을 관찰할 수 있다.

(개별예측의 예측오차 분산) $\sigma_u^2(1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2})$

(평균예측의 예측오차 분산) $\sigma_u^2(\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2})$

첫째, 표본의 크기(n)가 커질수록 예측오차의 분산이 작아진다. 즉, 관측자료가 많을수록 좋은 예측을 할 수 있다는 것이다.

둘째, 교란항의 분산(σ_u^2)이 커질수록 예측오차의 분산이 커진다. 즉, 원래 회귀모형에서 불확실성이 커서 교란항의 분산이 크면 예측이 어려울 수밖에 없다는 것이다.

셋째, 독립변수의 표본평균으로부터 멀어질수록 예측오차도 커진다. 즉, X_0가 $\overline{X}$로부터 멀어질수록 표본회귀함수의 예측력은 크게 감소한다는 것이다.

예측오차의 분산이 구해지면 예측구간(또는 예측대)을 구할 수 있는데 $(1-\alpha)\times$ 100% 예측구간은 (3.52)과 같다.

$$\hat{\beta}_0+\hat{\beta}_1 X_0 \pm t_{(n-2,\frac{\alpha}{2})}\sigma_\epsilon \qquad (3.52)$$

요약하면 개별예측이든 평균예측이든 점예측치는 $\hat{Y}_0$로 동일하며 예측구간은 예측오차의 분산의 크기에 의해 결정된다는 것이다. 위 (3.52)식으로부터 알 수 있고 또한 <그림 3-8>에서도 나타내고 있듯이 개별예측에 대한 예측구간이 평균예측에 대한 예측구간보다 더 넓다는 것이다. 즉, 표본회귀함수를 이용한 개별예측의 예측력은 평균예측에 대한 예측력보다 떨어진다는 것을 의미한다.

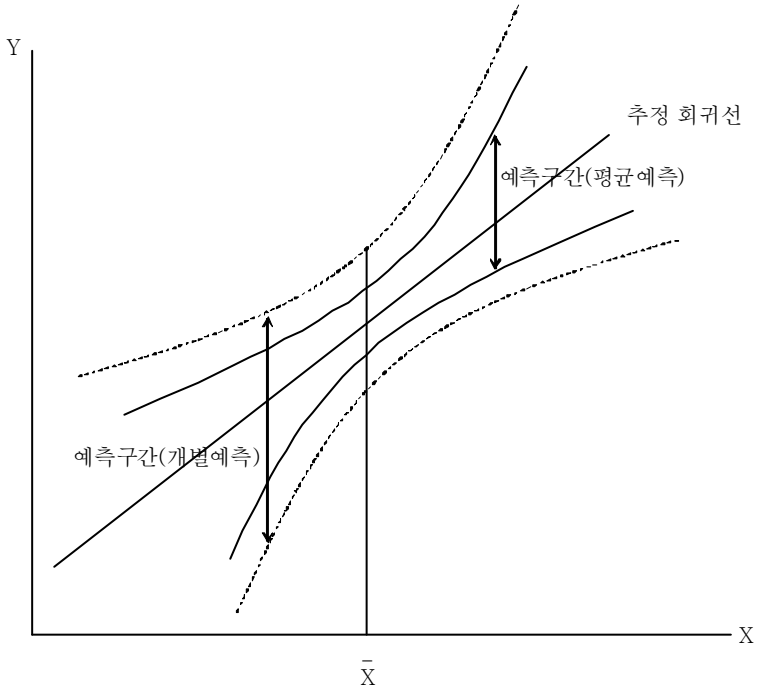
(예제 3-1)(계속)
다음의 홍보비 지출액 X(단위: 천만 원)와 연간매출액 Y(단위: 억 원)에 관한 자료를 이용하여 홍보비 지출액이 7(천만 원)일 경우 연간매출액의 점예측치 예측구간을 각각 구하라.

X	2	3	4	5	6
Y	4	4	6	6	10

추정된 회귀식 $\hat{Y}$= 0.4 + 1.4X에서 홍보비 지출액이 7(천만 원)일 때 즉, $X_0=7$일 때 연간매출액에 대한 개별예측치 및 평균예측치는 모두 $\hat{Y}$= 0.4 + 1.4X= 0.4+1.4(7) = 10.2(억 원)이다.

개별예측에 대한 예측구간을 구해보도록 하자. 먼저 예측오차의 분산을 구해보자. 앞의 계산에서 교란항의 분산 $\hat{\sigma}_u^2$은 1.4667, 독립변수의 평균 $\overline{X}$는 4, 표본의 크기 n은 5, $\sum x_i^2=10$이므로 이를 (3.50)식에 대입하면 3.08이 된다.

<그림 3-8> 예측구간



$$\sigma_u^2(1+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2})=(1.4667)(1+\frac{1}{5}+\frac{(7-4)^2}{10})=3.08$$

다음으로 개별예측에 대한 95% 예측구간을 구해보자. 점예측치가 10.2이고 예측오차의 분산이 3.08, $t_{3,0.025}=3.182$이므로 이를 (3.52)식에 대입하면 (3.53)식과 같이 예측구간을 구할 수 있다.

$$\widehat{\beta_0}+\widehat{\beta_1}X_0 \pm t_{(n-2,\frac{\alpha}{2})}\sigma_\epsilon = 10.2+(3.182)\sqrt{3.08}=[4.61,15.88] \qquad (3.53)$$

8.최우추정법

보통최소자승법 외에 회귀계수를 추정하는 또 다른 방법으로는 최우추정법

(Maximum Likelihood Estimation: MLE)이 있다. 보통최소자승법을 이용한 회귀계수의 추정에서는 교란항(교란항)의 분포에 대한 가정은 필요 없었는데 최우추정법을 사용하기 위해서는 교란항에 대한 가정이 필요하며 일반적으로 정규분포를 가정한다.

2변수 회귀모형 $y_i = \check{z}_0 + \check{z}_1 X_i + u_i$에서 $Y_i \sim NI(\beta_0 + \beta_1 X_i, \sigma^2)$라고 가정하자. 이 경우 $Y_1, Y_2, ..., Y_n$의 결합확률밀도함수는 다음과 같다.

$$f(Y_1, Y_2, .., Y_n \mid \beta_0 + \beta_1 X_i, \sigma^2) = f(Y_1 \mid \beta_0 + \beta_1 X_i, \sigma^2) f(Y_2 \mid \beta_0 + \beta_1 X_i, \sigma^2)$$
$$\cdots$$
$$f(Y_N \mid \beta_0 + \beta_1 X_i, \sigma^2)$$

단, $f(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e[-\frac{1}{2}\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}]$로서 정규분포의 확률밀도함수이다.

따라서 $Y_1, Y_2, ..., Y_n$의 결합확률밀도함수는 다음과 같이 나타낼 수 있다.

$$f(Y_1, Y_2, .., Y_n \mid \beta_0 + \beta_1 X_i, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e[-\frac{1}{2}\sum\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}] \quad (3.54)$$

(3.54)식에서 $Y_1, Y_2, ..., Y_n$이 알려져 있고, $\check{z}_0, \check{z}_1, \sigma^2$이 알려져 있지 않을 경우 이를 우도함수(likelihood function)라고 하고 $L(\check{z}_0, \check{z}_1, \sigma^2)$로 표기하므로 우도함수는 (3.55)식과 같다.

$$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} e[-\frac{1}{2}\sum\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}] \quad (3.55)$$

최우추정법이란 주어진 Y의 값들을 관측할 확률을 최대로 하는 모수$(\beta_0, \beta_1, \sigma^2)$을 추정하는 방법이다. 즉, 우도함수((3.55)식)의 최댓값을 찾는 방법이다.

(3.55)식의 미분값을 간단히 구하기 위해 동 식을 로그형태로 바꾸면 (3.56)식과 같게 된다.

$$lnL = -\frac{n}{2} ln\sigma^2 - \frac{1}{2} ln(2\pi) - \frac{1}{2}\sum \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}] \quad (3.56)$$

(3.56)식을 $\beta_0, \beta_1, \sigma^2$에 대해 편미분하면 다음의 (3.57)-(3.59)식과 같다

$$\frac{\partial lnL}{\partial \beta_0} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_0 - \beta_1 X_i)(-1) \quad (3.57)$$

$$\frac{\partial lnL}{\partial \beta_1} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_0 - \beta_1 X_i)(-X_1) \quad (3.58)$$

$$\frac{\partial lnL}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (3.59)$$

(3.57)-(3.59)식을 0으로 하고 최우추정량을 $\hat{\beta_0}, \hat{\beta_1}, \hat{\sigma^2}$로 나타내면 다음의 (3.60)-(3.62)식과 같다.

$$\frac{1}{\hat{\sigma^2}}\sum (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i) = 0 \quad (3.60)$$

$$\frac{1}{\hat{\sigma^2}}\sum (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i) X_1 = 0 \quad (3.61)$$

$$-\frac{n}{2\hat{\sigma^2}} + \frac{1}{2\hat{\sigma^4}}\sum (Y_i - \hat{\beta_0} - \hat{\beta_1} X_i)^2 = 0 \quad (3.62)$$

(3.60) 및 (3.61)식으로부터 다음의 (3.63) 및 (3.64)식을 얻게 되는데 이는 최소자승법에서의 정규방정식과 동일하다.

$$\sum_{i=1}^{n} Y_i = n\hat{\beta_0} + \hat{\beta_1}\sum_{i=1}^{n} X_i \quad (3.63)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \tag{3.64}$$

한편, (3.62)식으로부터 (3.65)식과 같은 분산에 대한 추정량을 얻을 수 있다. 최우추정법에서 분산에 대한 추정량은 점근적 불편추정량이 된다.

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n}\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ &= \frac{1}{n}\sum e_i^2 \end{aligned} \tag{3.65}$$

최우추정량은 다음과 같은 특징을 갖는다.

첫째, 회귀계수에 대한 최우추정량은 최소자승법에 의한 추정량과 동일하다. (3.63) 및 (3.64)식으로부터 회귀계수에 대한 취우추정량을 얻게 되는데 이 식이 보통최소자승법에서의 정규방정식과 동일하므로 회귀계수에 대한 최우추정량과 최소자승추정량은 동일하게 된다.

둘째, 교란항의 분산에 대한 추정량은 불편추정량은 되지 못하나 점근적으로 불편추정량이 된다.

셋째, 교란항의 분산에 대한 추정량은 N이 무한히 커질 때 참값에 접근하는 일치추정량이 된다.

9.회귀모형의 응용

(3.12) 및 (3.14)식에 의한 회귀계수 β_0, β_1의 추정치는 표본자료의 측정단위에 따라 달라진다. (예제 3-1)에서 주어진 자료로 추정된 회귀식 $\hat{Y}_i = 0.4 + 1.4X_i$에서 독립변수인 홍보비 지출액의 측정단위는 천만 원이고 종속변수인 연간매출액의 측정단위는 억 원이다. 따라서 홍보비 지출액이 천만 원 증가하면 연간매출액은 평균적으로 1.4억 원 증가한다고 할 수 있다.

만약에 홍보비 지출액의 측정단위는 변하지 않고 연간매출액의 측정단위가 천만 원일 경우 회귀계수의 추정치는 어떻게 될 것인가? 독립변수 또는 종속변수의 측정단위가 달라지면 회귀계수의 추정치 역시 달라진다. 그러나 표본자료의 측정단위를 고려한 β_0, β_1의 추정치와 해석은 달라지지 않는다. 만약에 측정단위에 따라 해석이 달라진다면 어떤 측정단위를 사용하는 것이 바람직한 지 판단해야 하

고 이에 대한 답을 도출할 수 없을 것이다. 또한 결정계수 R^2와 t-통계량의 값은 측정단위에 영향을 받지 않으므로 자료의 측정단위가 변하더라도 이들의 값은 변하지 않고 따라서 통계적 추론 역시 문제가 발생하지 않는다.

회귀분석에서 회귀함수의 형태를 선택하는 것도 매우 중요한데 경제변수 간의 함수형태는 경제이론이나 경험으로부터 결정된다. 가장 간편한 경제관계의 함수형태는 선형형태이나 경제관계가 항상 선형관계로 나타내어질 수 있는 것은 아니다. 또한 선형의 경우도 회귀계수에 있어서 선형인지 자료에 있어서 선형인지를 구분해야 하는데 본 장에서 선형회귀모형은 회귀계수에 있어서 선형인 경우를 가정하였다.

회귀분석에서 회귀함수의 형태를 선택하는 기준은 다음과 같다.

첫째, 경제이론에 근거한 함수형태를 선택한다. 예를 들어 생산함수를 추정한다고 하면 경제이론에 근거하여 Cobb-Douglas 생산함수를 선택할 수 있을 것이다.

둘째, 가능한 한 간단한 함수형태를 선택한다. 모형의 설명력에 큰 차이가 없다면 경제학의 효율성 원칙에 따라 간단한 함수형태를 선택하는데 이를 간결성 원칙(parsimonious principle)이라고 한다.

셋째, 예측력이 좋은 함수형태를 선택한다. 간결한 모형이 좋기는 하지만 무형의 현실 설명력 즉 예측력이 떨어진다면 모형의 유용성이 크게 떨어질 수밖에 없다.

일반적으로 계량경제분석에 널리 사용되는 모형에는 다음과 같은 것들이 있다.

(1)선형(linear)함수 모형

선형(linear)함수 모형은 (3.66)식과 같다.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{3.66}$$

추정된 회귀계수는 함수형태에 따라 해석이 달라지는데 선형함수 모형의 경우 기울기는 β_1이고 독립변수의 변화에 대한 종속변수의 변화를 나타내는 탄력성은 $\beta_1 \frac{X}{Y}$가 된다. 탄력성은 $\epsilon_{Y|X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX}$이고, 선형함수 모형을 전미분하면

$dY = \beta_1 dX$ 즉, 기울기 $\frac{dY}{dX} = \beta_1$이므로 선형함수 모형의 탄력성은$\beta_1 \frac{X}{Y}$ 이 된다.

(2)전대수(double log)함수 모형

(3.67)식과 같은 전대수(double log)함수 모형의 예는 Cobb-Douglas 생산함수에 로그를 취한 함수이다.

$$\ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + u_i \qquad (3.67)$$

전대수함수 모형의 경우 기울기는 $\beta_1 \frac{X}{Y}$이고 독립변수의 변화에 대한 종속변수의 변화를 나타내는 탄력성은 β_1이 된다. 예를 들어 $Y = \beta_0 X^{\beta_1}$와 같은 모형이 있을 경우 이모형에 로그를 취하면 $\ln Y = \ln \beta_0 + \beta_1 \ln X$와 같은 전대수함수 모형이 된다. 이를 전미분하면 $\frac{dY}{Y} = \frac{\beta_1}{X} dX$이므로 기울기는 $\frac{dY}{dX} = \beta_1 \frac{Y}{X}$이다. 한편, 탄력성은 다음과 같이 β_1이 되어 추정된 회귀계수가 탄력성이 된다.

$$\epsilon_{Y|X} = \frac{dY/Y}{dX/X} = \frac{X}{Y}\frac{dY}{dX} = (X/\beta_0 X^{\beta_1})(\beta_0 \beta_1 X^{\beta_1 - 1}) = \beta_1$$

(3)쌍곡선(reciprocal)함수 모형

(3.68)식과 같은 쌍곡선(reciprocal)함수 모형으로 물가상승률과 실업률의 역관계를 나타낸 주는 필립스 곡선이 좋은 예이다.

$$Y_i = \beta_0 + \beta_1 (\frac{1}{X_i}) + u_i \qquad (3.68)$$

쌍곡선함수 모형의 경우 기울기는 $-\beta_1 \frac{1}{X^2}$ 이고 독립변수의 변화에 대한 종속변수의 변화를 나타내는 탄력성은 $-\beta_1 \frac{1}{YX}$이 된다. 쌍곡선함수 모형을 전미분하

면 $dY=-\beta_1 X^{-2}dX$이므로 기울기 $\frac{dY}{dX}=-\beta_1\frac{1}{X^2}$이다. 한편, 탄력성은 $\epsilon_{Y|X}=\frac{dY/Y}{dX/X}=\frac{X}{Y}\frac{dY}{dX}=\frac{X}{Y}(-\beta_1 X^{-2})=-\beta_1\frac{1}{XY}$이다.

(4)반로그(semi-log)함수 모형

(3.69)식과 같은 반로그(semi-log)함수 모형으로 종속변수가 반로그인 모형과 독립변수가 반로그인 모형을 구분될 수 있다.

$$\ln Y_i=\beta_0+\beta_1 X_i+u_i$$

또는

$$Y_i=\beta_0+\beta_1\ln X_i+u_i \qquad (3.69)$$

반로그함수 모형 $\ln Y_i=\beta_0+\beta_1 X_i+u_i$의 경우 기울기는 $\beta_1 Y$이고 독립변수의 변화에 대한 종속변수의 변화를 나타내는 탄력성은 $\beta_1 X$가 된다. 반로그함수 모형을 전미분하면 $\frac{dY}{Y}=\beta_1 dX$이므로 기울기 $\frac{dY}{dX}=\beta_1 Y$이다. 한편, 탄력성은 $\epsilon_{Y|X}=\frac{dY/Y}{dX/X}=\frac{X}{Y}\frac{dY}{dX}=\frac{X}{Y}(-\beta_1 Y)=\beta_1 X$이 된다.

반로그함수 $Y_i=\beta_0+\beta_1\ln X_i+u_i$의 경우 기울기는 $\beta_1\frac{1}{X}$이고 독립변수의 변화에 대한 종속변수의 변화를 나타내는 탄력성은 $\beta_1\frac{1}{Y}$이 된다. 반로그함수 모형을 전미분하면 $dY=\frac{\beta_1}{X}dX$이므로 기울기 $\frac{dY}{dX}=\beta_1\frac{1}{X}$이다. 한편, 탄력성은 $\epsilon_{Y|X}=\frac{dY/Y}{dX/X}=\frac{X}{Y}\frac{dY}{dX}=\frac{X}{Y}(-\beta_1\frac{1}{X})=\beta_1\frac{1}{Y}$가 된다.

함수형태에 따른 기울기 및 탄력성에 대한 지금까지 논의를 요약하여 나타내면 <표 3-1>과 같다.

<표 3-1> 함수형태에 따른 기울기 및 탄력성

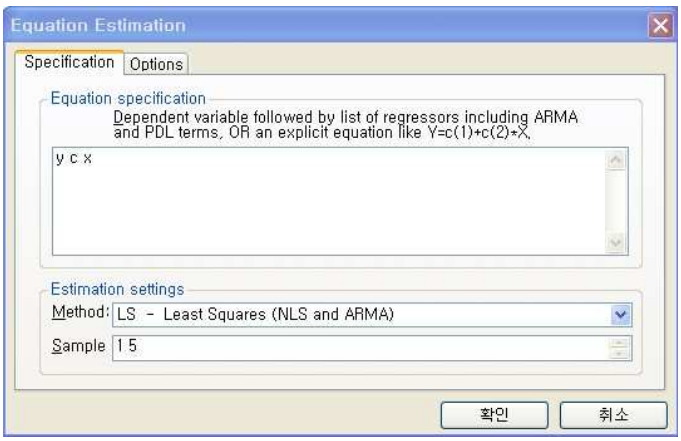
함수	형태	기울기($\frac{dY}{dX}$)	탄력성($\frac{X}{Y}\frac{dY}{dX}$)
선형 (linear)	$Y_i = \beta_0 + \beta_1 X_i + u_i$	β_1	$\beta_1 \frac{X}{Y}$
전대수 (log-log)	$\ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + u_i$	$\beta_1 \frac{X}{Y}$	β_1
쌍곡선 (reciprocal)	$Y_i = \beta_0 + \beta_1 (\frac{1}{X_i}) + u_i$	$-\beta_1 \frac{1}{X^2}$	$-\beta_1 \frac{1}{YX}$
반로그 (semi-log)	$\ln Y_i = \beta_0 + \beta_1 X_i + u_i$	$\beta_1 Y$	$\beta_1 X$
반로그 (semi-log)	$Y_i = \beta_0 + \beta_1 \ln X_i + u_i$	$\beta_1 \frac{1}{X}$	$\beta_1 \frac{1}{Y}$

실습 : 단순회귀모형 추정

(예제 3-1)의 자료를 EViews를 이용하여 ex3-1.wf1으로 저장한 후 홍보비 지출을 독립변수, 연간매출액을 종속변수로 하는 단순회귀모형을 추정한 후 여러 가지 결과를 살펴보도록 하자.

Quick-Estimate Equation을 선택하면 나오는 대화상자에서 <그림 3-A1>과 같이 입력을 한다. 여기서 y는 종속변수를 나타내고, c는 단순회귀모형의 상수항을 추정하기 위한 것이며, x는 독립변수를 나타낸다. 그림과 같이 입력한 후 확인을 클릭하면 <그림 3-A2>와 같은 추정결과를 나타내 준다.

<그림 3-A1> 단순회귀모형(예제 3-1) 추정회귀식

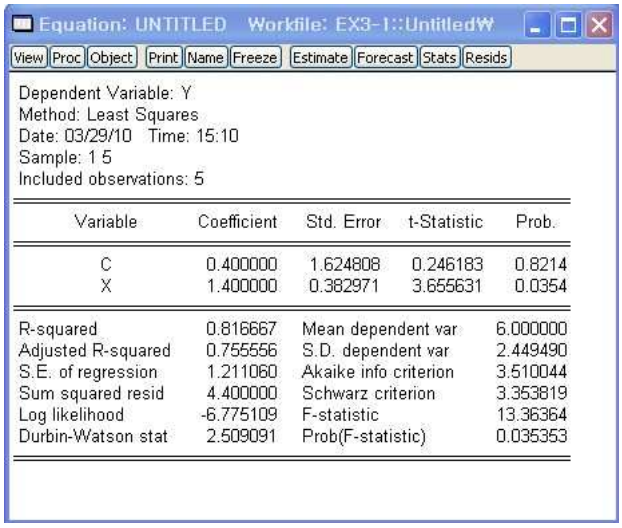


<그림 3-A2>의 추정결과에서 나타내는 바는 각각 다음과 같은데 이를 앞에서 계산한 값들과 비교해 보면 동일함을 알 수 있다.

ⓐ Coefficient : 회귀계수의 추정치로서 C는 상수항을 의미하고 X는 독립변수를 각각 나타낸다. 즉, $\widehat{\beta_0}=0.4, \widehat{\beta_1}=1.4$이다.

ⓑ Std. Error : 회귀계수의 표준오차로서 $\widehat{\beta_0}$의 표준오차는 1.624808이고, $\widehat{\beta_1}$의 표준오차는 0.382971이다.

<그림 3-A2> 단순회귀모형(예제 3-1) 추정결과


Equation: UNTITLED Workfile: EX3-1::Untitled

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 03/29/10 Time: 15:10
Sample: 1 5
Included observations: 5

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.400000	1.624808	0.246183	0.8214
X	1.400000	0.382971	3.655631	0.0354

R-squared	0.816667	Mean dependent var	6.000000
Adjusted R-squared	0.755556	S.D. dependent var	2.449490
S.E. of regression	1.211060	Akaike info criterion	3.510044
Sum squared resid	4.400000	Schwarz criterion	3.353819
Log likelihood	-6.775109	F-statistic	13.36364
Durbin-Watson stat	2.509091	Prob(F-statistic)	0.035353

ⓒ t-Statistic : 귀무가설 $H_0 : \beta_0 = 0$ 및 $H_0 : \beta_1 = 0$에 대한 t-통계치로서 β_0에 대한 귀무가설의 t-통계치는 0.246183이고, β_1에 대한 귀무가설의 t-통계치는 3.655631이다.

ⓓ Prob. : 유의확률로서 $\widehat{\beta_0}$의 유의확률은 0.8214이고, $\widehat{\beta_1}$의 유의확률은 0.0354이다. 예를 들어 $\widehat{\beta_1}$의 유의확률이 0.0354이라는 것은 β_1에 대한 귀무가설의 t-통계치는 3.655631보다 크거나 -0.3655631보다 작을 확률이 0.0354라는 의미이므로 5% 유의수준에서 가설검정을 할 경우 귀무가설을 기각한다. 즉, 이 확률이 주어진 유의수준보다 작을 경우 귀무가설을 기각한다.

ⓔ R-squared : 결정계수 R^2의 값이 0.816667이라는 것은 홍보비 지출로 연간매출액의 변화를 81.67% 설명할 수 있다는 것을 의미한다. 즉, 독립변수의 종속변수 설명력이 81.67%라는 것이다.

ⓕ S.E.of regression : 교란항 분산의 제곱근(표준편차)으로 $\widehat{\sigma_u} = 1.21106$이다.

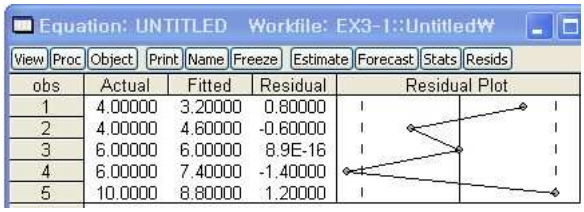
ⓖ Sum squared resid : 잔차의 제곱의 합으로 $\sum e_i^2 = 4.4$이다.

ⓗ Mean dependent var : 종속변수 Y의 평균으로 $\overline{Y} = 6$이다.

ⓘ S.S. dependent var : 종속변수 Y의 표준편차로 $S_Y = \sqrt{\frac{1}{4}\sum (Y_i - \overline{Y})^2} = 2.44949$이다.

<그림 3-A2>의 상태에서 View-Actual, Fitted, Residual/Actual, Fitted, Residual Table을 선택하면 <그림 3-A3>이 나타나는데 다음과 같은 의미를 가진다.

<그림 3-A3> 실제값, 추정치 및 잔차


Equation: UNTITLED Workfile: EX3-1::Untitled

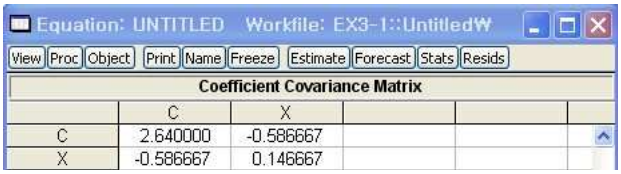
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

obs	Actual	Fitted	Residual	Residual Plot
1	4.00000	3.20000	0.80000	
2	4.00000	4.60000	-0.60000	
3	6.00000	6.00000	8.9E-16	
4	6.00000	7.40000	-1.40000	
5	10.0000	8.80000	1.20000	

ⓐ obs : 관측치(observation)을 나타내는 것으로 즉, X_i 및 Y_i에서 $i(i=1,2,3,4,5)$로 표시된다.
ⓑ Actual : 종속변수 Y의 실제값을 나타낸 것으로 $Y_1=4, Y_2=4,..,Y_5=10$이다.
ⓒ Fitted : 종속변수 Y의 추정치를 나타낸 것으로 $\widehat{Y_1}=3.2, \widehat{Y_2}=4.6,..,\widehat{Y_5}=8.8$이다.
ⓓ Residual : 종속변수 Y의 실제값과 그 추정치의 차이인 잔차를 나타낸 것으로 $e_1=0.8, e_2=-0.6,..,e_5=1.2$이며 잔차의 합($\sum e_i$)은 0이 됨을 보여주고 있다.
ⓔ Residual Plot : 잔차를 그린 것으로 양의 잔차가 2개, 음의 잔차가 2개, 0인 잔차가 1개임을 보여주고 있다.

<그림 3-A2>의 상태에서 View-Covariance Matrix를 선택하면 <그림 3-A4>와 같은 회귀계수의 분산-공분산행렬이 나타나는데 다음과 같은 의미를 가진다.

<그림 3-A4> 회귀계수의 공분산행렬


Equation: UNTITLED Workfile: EX3-1::Untitled

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Coefficient Covariance Matrix

	C	X
C	2.640000	-0.586667
X	-0.586667	0.146667

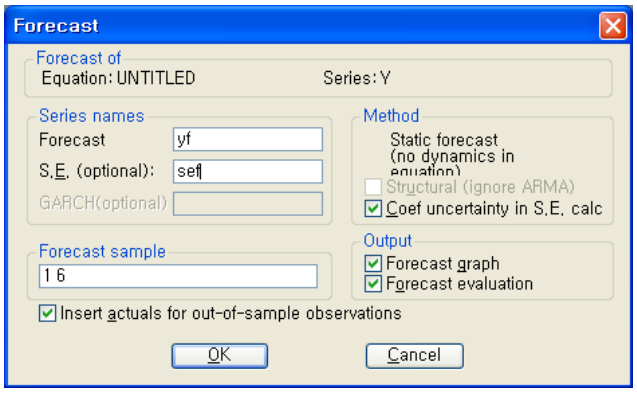
ⓐ 대각행렬에 있는 값들은 회귀계수의 분산으로서 $\widehat{\beta_0}$에 대한 분산은 2.64이고,

$\hat{\beta}_1$에 대한 분산은 0.146667이다.

ⓑ 비대각행렬에 있는 값은 회귀계수의 공분산으로 $\hat{\beta}_0$과 $\hat{\beta}_1$의 공분산은 -0.586667이다. 이는 $cob(\hat{\beta}_0,\hat{\beta}_1)=\frac{-\overline{X}\sigma_u^2}{\sum x_i^2}=-(4)(1.45\frac{57)}{10}=-0.5866$임을 통해 확인할 수 있다.

예측을 하기 위해서는 먼저 Proc-structure/resize current page를 선택하여 Data range를 6으로 하여 예측을 위한 공간을 하나 더 확보하고, 독립변수를 클릭하여 Edit+/-를 눌러 예측을 위한 독립변수의 값(여기서는 $X_0=7$)을 입력한다. 회귀모형을 추정하고 추정결과 창에서 Forecast를 클릭하면 <그림 3-A5>와 같은 예측 대화상자가 나타나는데 거기서 그림과 같이 입력하고 OK를 클릭하면 종속변수의 점예측치(yf) 및 개별예측치 예측오차의 표준오차(sef)가 계산된다.

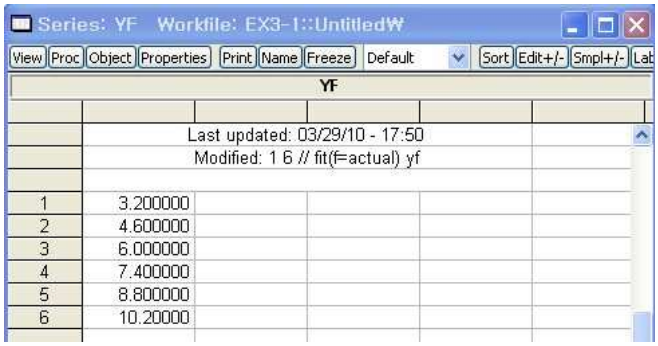
<그림 3-A5> 예측 대화상자



종속변수의 점예측치(yf)를 클릭해 보면 <그림 3-A6>이 나타나는데 여기서 예측치가 10.2임을 알 수 있다.

한편, 개별예측치 예측오차의 표준오차(sef)를 클릭해 보면 <그림 3-A7>과 같이 나타나 있는데 여기서 개별예측치 예측오차의 표준오차가 1.754993임을 확인할 수 있고, 이를 이용하면 개별예측에 대한 예측구간 [4.61,15.88]을 구할 수 있다.

<그림 3-A6> 종속변수 예측치



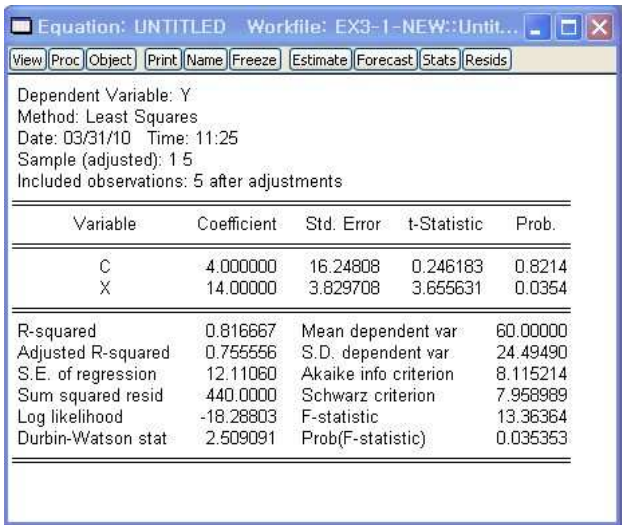
<그림 3-A7> 개별예측치 예측오차의 표준오차

(예제 3-1)에서 독립변수의 측정단위는 천만 원으로 변하지 않고 종속변수의 측정단위가 억 원에서 천만 원으로 변화되었을 경우의 회귀분석 결과가 <그림 3-A8>과 나타나 있다. 먼저 새로운 회귀계수는 14로서 이전(<그림 3-A2>)보다 10이 증가한 것으로 나타나고 있으나 측정단위가 천만 원이므로 홍보비 지출액이 천만 원 증가할 경우 연간매출액은 14천만 원 즉, 1.4억 원 증가한 것으로 나타나 측정단위를 고려하면 동일한 결과임을 알 수 있다. 또한 측정단위의 변화로 회귀계수가 변했으나 동 회귀계수의 표준오차도 변하여 t-통계량에는 변화가 없음을 확인할 수 있다.

한편, (예제 3-1)에서 독립변수 및 종속변수 모두 로그를 취한 전대수함수 모형의 회귀분석 결과가 <그림 3-A9>에 나타나 있다. 회귀계수 β_1에 대한 추정치 즉, 탄력성이 0.775546이다. 이는 홍보비 지출이 1% 증가할 때 연간매출액은

0.775546% 증가한다는 것을 의미하므로 연간매출액은 홍보비 지출에 비탄력적인 것으로 나타났다.

<그림 3-A8> 측정단위 변화에 따른 추정 결과


Equation: UNTITLED Workfile: EX3-1-NEW::Untit...

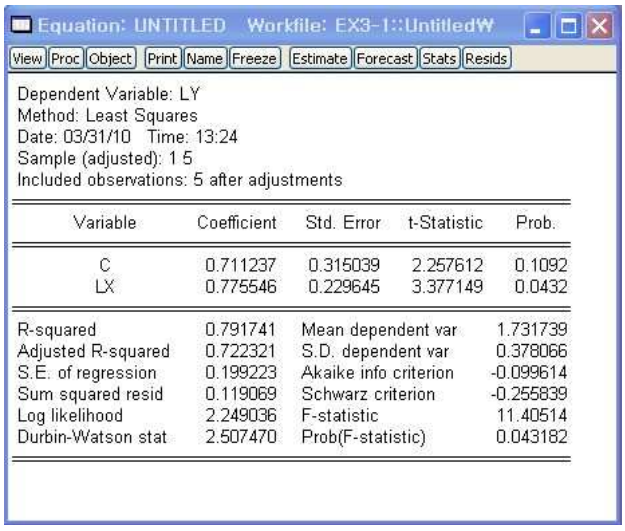
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 03/31/10 Time: 11:25
Sample (adjusted): 1 5
Included observations: 5 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.000000	16.24808	0.246183	0.8214
X	14.00000	3.829708	3.655631	0.0354

R-squared	0.816667	Mean dependent var	60.00000
Adjusted R-squared	0.755556	S.D. dependent var	24.49490
S.E. of regression	12.11060	Akaike info criterion	8.115214
Sum squared resid	440.0000	Schwarz criterion	7.958989
Log likelihood	-18.28803	F-statistic	13.36364
Durbin-Watson stat	2.509091	Prob(F-statistic)	0.035353

<그림 3-A9> 측정단위 변화에 따른 추정 결과


Equation: UNTITLED Workfile: EX3-1::UntitledW

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LY
Method: Least Squares
Date: 03/31/10 Time: 13:24
Sample (adjusted): 1 5
Included observations: 5 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.711237	0.315039	2.257612	0.1092
LX	0.775546	0.229645	3.377149	0.0432

R-squared	0.791741	Mean dependent var	1.731739
Adjusted R-squared	0.722321	S.D. dependent var	0.378066
S.E. of regression	0.199223	Akaike info criterion	-0.099614
Sum squared resid	0.119069	Schwarz criterion	-0.255839
Log likelihood	2.249036	F-statistic	11.40514
Durbin-Watson stat	2.507470	Prob(F-statistic)	0.043182

연습문제

1. 다음은 1998년부터 2002년까지 통화량증가율과 물가상승률의 연도별 자료이다. 이를 이용하여 다음 물음에 답하라.

연도	1998	1999	2000	2001	2002
통화량증가율(%)	2	3	4	5	6
물가상승률(%)	4	5	4	6	7

①"1998-2002년 기간 중 통화량이 1% 상승하면 물가는 1.4% 상승한다고 할 수 있다". 이 주장을 평가하라.

②“동 기간 중 통화량은 물가변동을 72% 설명한다”. 이 주장을 평가하라.

③ “위의 모형을 이용하면 물가상승률의 미래값을 예측하는데 유용하다”. 이를 평가하라.

2. 다음의 자료는 감귤의 가격과 수요량을 나타낸 것이다. 이를 이용하여 물음에 답하라.

수요량(Q)	2	3	4	5	6
가격(P)	10	6	6	4	4

① 두 변수의 상관계수를 추정하고 그 의미를 설명하라.

② 단순회귀모형의 회귀계수를 추정하라.

③ 교란항의 분산을 추정하라

④ 회귀계수의 분산을 추정하라

⑤ 결정계수를 구하라

⑥ 주어진 독립변수에 대한 추정치를 각각 구한 후 직교조건이 성립함을 보이고 그 경제적 의미를 설명하라.

⑦ "독립변수가 종속변수에 영향을 주지 않는다"는 귀무가설을 5% 유의수준 하에서 검정하라.

⑧ 독립변수의 값이 7이라고 할 때 점예측치 및 99% 예측구간을 구하라.

⑨ 감귤이 사치재인지 아니면 필수재인지를 밝혀라.

⑩ ①번에서 설정한 모형에서 독립변수와 종속변수를 바꾼 회귀모형의 회귀계수를 추정하라. 이를 통해서 내릴 수 있는 결론은 무엇인가.

3. 다음의 자료를 이용하여 물음에 답하라.

주당소득	80	100	120	140	160	180	200	220	240	260
주당소비	70	65	90	95	110	180	120	140	155	150

① 모형을 $Y_i = \beta_1 + \beta_2 X_i + u_i$로 설정하고 각각 무엇이 X변수, Y변수인지를 설정하고 그 이유를 설명하라.

② 최소자승법으로 회귀계수 $\widehat{\beta_1}$과 $\widehat{\beta_2}$을 추정하라.

4. 1970년부터 1992년까지의 23개의 자료를 이용하여 소득과 소비의 관계를 추정한 결과가 다음과 같을 때 다음의 물음에 답하라.

$$C_t = \underset{(0.02)}{0.49\, Y_t}, S^2 = 24.9$$

단, C_t는 t기의 소비, Y_t는 t기의 소득, 괄호 안은 표준오차를 각각 나타낸다.

① 한계소비성향이 0.54라는 다음의 가설을 5%와 1% 유의수준 하에서 각각 검정하라.

$H_0 : \beta = 0.54$
$H_1 : \beta \neq 0.54$

② 한계소비성향의 95%와 99%의 신뢰구간을 구하고 그 의미를 써라.

5. X_1, X_2, X_3의 3개의 확률표본자료를 이용하여 모평균(μ)을 추정하기 위한 추정량을 다음의 3가지 방법으로 구하였다. 물음에 답하라.

$$\overline{X} = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3$$

$$\hat{X} = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{2}X_3$$

$$\widetilde{X} = \frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3$$

① 3개의 추정량 중 불편추정량은?

② 3개의 추정량 중 가장 효율적인 추정량은?

6. 다음의 자료를 이용하여 물음에 답하라.

X	1	2	3	4	5
Y	2	5	3	8	7

TX	10	20	30	40	50
TY	20	50	30	80	70

① X와 Y의 공분산과 상관계수를 구하라

② TX와 TY의 공분산과 상관계수를 구하라

③ 단순회귀모형을 $Y_i = \beta_0 + \beta_1 X_i + u_i$로 설정하고 최소자승법으로 모형 내 모수를 모두 구하라.

④ 잔차의 합을 구하고 직교조건이 성립함을 보여라.

⑤ 회귀계수 β_1에 대한 95% 구간 추정치를 구하라.

⑥ 위 모형의 결정계수를 구하라

⑦ X는 Y에 영향을 준다고 볼 수 있나?

⑧ 단순회귀모형을 $X_i = \alpha_0 + \alpha_1 Y_i + e_i$로 설정하고 회귀계수를 구하라.

⑨ Y는 X에 영향을 준다고 볼 수 있나?

⑩ 단순회귀모형을 $TY_i = \gamma_0 + \gamma_1 TX_i + \epsilon_i$로 설정하고 최소자승법으로 회귀계수를 구하라.

7. 다음의 자료를 이용하여 물음에 답하라.

표본번호	자료 1군		자료 2군		자료 3군	
	소득(백원)	소비(원)	소득(백원)	소비(원)	소득(원)	소비(원)
1	4	40	-4	-22	40000	4000
2	6	60	-2	- 2	60000	6000
3	7	50	-1	-12	70000	5000
4	10	70	2	8	100000	7000
5	13	90	5	28	130000	9000

7-1 소득(단위: 백 원)을 독립변수, 소비(단위: 원)을 종속변수로 하는 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + u_i$을 설정하고 자료 1군을 이용하여 최소자승법으로 추정하

였다.

① 한계소비성향을 구하고 해석하라.

② 교란항의 분산을 구하라

③ 소득은 소비에 영향을 주지 않는다는 가설을 5% 유의수준 하에서 검정하라.

④ 한계소비성향의 95% 신뢰구간을 추정하라

⑤ 소득은 소비를 어느 정도 설명하고 있나?

⑥ 소득이 15백 원인 사람의 소비수준은 어느 정도 될까?

⑦ ⑥번 문제의 95% 예측대를 구하라

7-2. 소득을 독립변수, 소비를 종속변수로 하는 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 을 설정하고 자료 2군을 이용하여 최소자승법으로 추정하였을 경우 다음의 빈칸을 채워라

	계수 추정치	분산 추정치	$\widehat{\sigma_u^2}$	R^2
$\widehat{\beta_0}$				
$\widehat{\beta_1}$				

7-3. 소득을 독립변수, 소비를 종속변수로 하는 단순회귀모형 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 을 설정하고 자료 3군을 이용하여 최소자승법으로 추정하였을 경우 다음의 빈칸을 채워라(단, 소득의 단위는 백 원으로 하고 소비의 단위는 원으로 하라)

	계수 추정치	분산 추정치	$\widehat{\sigma_u^2}$	R^2
$\widehat{\beta_0}$		(필요 없음)		
$\widehat{\beta_1}$				

8. 다음은 우리나라 시도의 지역총생산(GRDP) 자료를 이용하여 분석한 결과를 몇 개의 지역에 대해서 요약표를 만든 것이다. 물음에 답하되 그 이유를 간단히 설명하라.

	서울	부산	경기	경남	경북	충북	전북	강원	제주
Mean	9.4	6.9	13.1	10.4	8.5	7.8	6.1	4.7	8.5
Var	16.8	18.0	19.8	23.5	9.9	17.5	13.9	9.1	50.4
Skewness	-0.14	0.09	-0.27	-0.66	0.39	-0.88	-0.19	0.34	1.23
Corr[1)]	0.82	0.95	0.94	0.49	0.83	-0.77	0.23	0.06	-0.16
$\widehat{\beta_1}$[2)]	1.25	1.50	1.55	1.52	0.57	-1.21	0.32	0.07	-0.42
Sig.	0.01	0.00	0.00	0.01	0.18	0.01	0.54	0.87	0.68
R^2	0.67	0.89	0.88	0.71	0.23	0.60	0.05	0.00	0.03

[1)] 전국GRDP와 각 지역GRDP의 상관계수이다.

[2)] 전국GRDP를 독립변수로 지역GRDP를 종속변수로 한 단순회귀모형의 회귀계수 모형: 지역$GRDP = \beta_0 + \beta_1$전국$GRDP + u$

① 연평균성장률이 가장 높은 지역은?

② GRDP증가율이 좌우대칭에 가장 가까운 분포를 하는 지역은?

③ GRDP증가율의 변동이 가장 큰 지역은?

④ 전국GRDP증가율이 상승할 때 GRDP증가율이 하락하는 지역은?

⑤ 전국GRDP증가율과 가장 밀접한 관계를 가지고 움직이는 지역은?

⑥ 전국GRDP증가율의 설명력이 가장 큰 지역은?

9. 다음은 2000년부터 2004년까지 소득증가율(income)과 소비증가율(consumption)의 연도별 자료이다. 이를 이용하여 다음 물음에 답하라.

연도	2000	2001	2002	2003	2004
소득증가율(%)	2	3	4	5	6
소비증가율(%)	3	1	3	2	4

① "2000-2004년 기간 중 소득이 1% 상승하면 소비는 0.3% 상승한다고 할 수 있다"라는 주장을 평가하라.

② "2005년도 소득증가율이 5%라고 할 때 2005년도의 소비증가율은 2.9%가 될 것이다"라는 주장을 평가하라.

10. 다음은 재정정책 및 금융정책의 유효성을 살펴보기 위하여 1990년부터 1996년까지 한국의 분기별 자료를 이용하여 단순회귀모형을 분석한 결과이다. 물음에 답하라.

모형 1: $\ln GDP_t = 0.08 + \underset{(0.21)}{0.68}\ln M2_t \quad R^2 = 0.86$

모형 2: $\ln GDP_t = 0.09 + \underset{(0.23)}{0.72}\ln M3_t \quad R^2 = 0.75$

모형 3: $\ln GDP_t = 1.25 + \underset{(0.63)}{1.25}\ln G_t \quad R^2 = 0.90$

(단, 괄호안의 값은 표준오차이다)

① 모형 1과 모형 3을 IS-LM의 그림을 이용하여 설명해 보라.

② 1990년대 한국의 경우 재정정책과 금융정책 중 무슨 정책이 유효하다고 볼 수 있나? 그 이유를 설명하라.

③ $\ln M2_t = 200, \ln M3_t = 220, \ln G_t = 80$이라고 할 때 1997년 한국의 GDP는 얼마가 되겠는가? 세 가지 모형 중 예측에 가장 적합한 모형을 이용하여 예측치를 구하라.

11. 다음은 각종 과일류 시장에서 조사한 10개의 자료로 가격과 수요량의 관계를 추정한 결과이다. 물음에 답하라.

모형 1(감귤시장): $\ln Q_O = 0.08 - 1.12 \ln PP_O \quad R^2 = 0.86$
(0.35)

모형 2(감귤시장): $\ln Q_O = 0.10 - 1.08 \ln CP_O \quad R^2 = 0.88$
(0.32)

모형 3(사과시장): $\ln Q_A = 0.09 - 0.72 \ln PP_A \quad R^2 = 0.75$
(0.23)

모형 4(사과시장): $\ln Q_A = 0.07 - 0.63 \ln CP_A \quad R^2 = 0.80$
(0.16)

모형 5(배시장) : $\ln Q_P = 1.25 - 1.25 \ln PP_P \quad R^2 = 0.90$
(0.63)

모형 6(배시장) : $\ln Q_P = 1.25 - 1.21 \ln CP_P \quad R^2 = 0.91$
(0.61)

(단, PP는 도매가격, CP는 소매가격을 각각 나타내고 괄호안의 값은 표준오차이다)

① 수요법칙이 성립하는 모형은? 그 이유를 설명하라.

② 수요의 가격탄력도가 가장 큰 품목은? 그 이유를 설명하라.

③ 감귤의 수요량을 예측하기 위해서는 도매가격과 소매가격 중 어떤 가격을 이용하는 것이 좋나? 그 이유를 설명하라.

제 4 장
다중회귀분석

1.모형
2.모형에 대한 가정
3.모형의 추정
4.추정량의 특성
5.회귀계수의 분산 추정
6.결정계수
7.가설검정
8.예측
실습 : 다중회귀모형 추정

제4장 다중회귀분석

제3장에서 살펴본 단순회귀분석은 계량분석의 기초가 되는 매우 중요한 부분이며 본 장에서 공부할 다중회귀분석은 단순회귀분석의 확장이므로 단순회귀분석에 대한 철저한 이해가 없이는 본 장을 이해하기가 어렵다.

단순회귀분석에서는 종속변수에 영향을 주는 독립변수가 하나 밖에 없는 것으로 가정하였는데 이러한 가정은 현실을 제대로 반영하지 못하고 있다. 예를 들어 어떤 가구의 소비수준에 영향을 미치는 변수들을 고려한다고 할 때 단순회귀분석에서는 소득만을 독립변수로 선택하였으나 현실에서는 소비에 영향을 주는 변수로는 소득 이외에도 많이 있을 수 있다. 가구 구성원의 수, 학력의 정도, 가구주의 결혼여부, 연령 등의 변수가 소득에 영향을 줄 수 있을 것이다.

만약에 이러한 변수들이 소비에 영향을 줄 가능성은 있지만 실제로 영향을 주지 않는다면 독립변수로 포함되어야 할 이유가 없고 따라서 소득만이 소비에 영향을 주는 단순회귀모형으로 충분할 것이다. 즉, 소득만이 소비를 설명한다고 하는 모형이 제대로 설정되었다고 할 수 있다. 그러나 이러한 변수들이 소비에 영향을 줄 가능성이 있을 뿐만 아니라 실제로 영향을 준다면 독립변수로 반드시 포함되어야 할 것이므로 소득만이 소비에 영향을 주는 단순회귀모형으로는 충분하지 않을 것이다. 즉, 소득만이 소비를 설명한다고 하는 모형은 제대로 설정된 모형이 아니라고 할 수 있다.

소득 외에 소비를 설명하는 변수가 있음에도 불구하고 이를 반영하지 않은 단순회귀모형을 설정하여 보통최소자승법으로 추정하면 추정량은 편의(bias)를 갖는 추정량 즉, 불편추정량이 되지 못하여 좋은 추정량이라고 할 수 없다. 따라서 소득을 설명하는 변수로 소비와 함께 이러한 변수들이 독립변수로 포함되어야 하는데 이와 같이 상수항을 제외한 독립변수의 수가 두 개 이상인 경우의 회귀분석을 다중회귀분석(multiple regression analysis)이라고 한다.

1.모형

상수항과 k-1개의 독립변수에 의해 설명되는 다중회귀모형은 일반적으로 (4.1)식과 같이 나타낸다.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i \ (i = 1,2,..,n) \qquad (4.1)$$

다중회귀모형에서 회귀계수는 다음과 같은 의미를 가진다. 먼저 상수항은 독립변수의 값이 0일 경우 종속변수의 기댓값을 나타내는데 독립변수의 값이 0일 경우는 별로 없으므로 관심의 대상이 되지 않는다. 독립변수와 종속변수의 관계를 나타내는 회귀계수 $\beta_1,\beta_2,...,\beta_k$를 부분회귀계수라고 하는데 예를 들어 β_3는 다른 독립변수들의 값이 변하지 않고 X_3의 값이 1단위 증가할 때 종속변수의 평균변화의 크기를 의미한다.

다중회귀분석의 일반모형을 행렬로 표현하면 (4.2)식과 같다.

$$Y = X\ B + U$$
nx1 nxk kx1 nx1

$$\text{단, } Y=\begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix},\ X=\begin{bmatrix} 1\ X_{21}\ X_{31}\cdots\ X_{k1} \\ 1\ X_{22}\ X_{32}\cdots\ X_{k2} \\ .\quad .\quad .\quad . \\ 1\ X_{2n}\ X_{3n}\cdots\ X_{kn} \end{bmatrix},\ B=\begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix},\ U=\begin{bmatrix} u_1 \\ u_2 \\ . \\ u_n \end{bmatrix} \qquad (4.2)$$

2.모형에 대한 가정

단순회귀모형을 추정하고 분석결과를 제대로 해석하기 위해서는 X 및 교란항에 대한 가정이 필요하였는데 다중회귀모형에서도 다음과 같은 가정이 필요하다.

(가정 1) 독립변수 X는 비확률변수이다.

X는 일정한 값을 가진 원소로 되어 있어 확률변수가 아니므로 X는 오차 없이 측정되며 full rank를 가진다(즉, 독립변수 간에 정확한 선형관계가 없다)는 의미이다.

(가정 2) 교란항의 평균은 0이다.
$$E(U)=0$$

(가정 3) 교란항은 모든 X_i에 대한 동일한 분산을 갖는다.
$$E(UU')=\sigma_u^2 I_n$$

이 가정을 행렬로 표현해 보면 (4.3)식과 같게 되는데 이 가정은 교란항의 분산이 일정하다는 것인데 이를 동분산(homoscedasticity)라고 한다. 또한 대각외 원소가 모두 0인데 이는 $E(u_i, u_j) = 0 (i \neq j)$를 의미하므로 이를 비자기상관(no autocorrelation)이라고 한다.

$$E(UU') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) \cdot\cdot E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) \cdot\cdot E(u_2u_k) \\ \cdot & \cdot \cdot\cdot \cdot \\ E(u_nu_1) & E(u_nu_2) \cdot\cdot E(u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_u^2 & 0 \cdot\cdot 0 \\ 0 & \sigma_u^2 \cdot\cdot 0 \\ \cdot & \cdot \cdot\cdot \cdot \\ 0 & 0 \cdot\cdot \sigma_u^2 \end{bmatrix}$$

$$= \sigma_u^2 I_n \qquad (4.3)$$

(가정 4) 교란항은 정규분포에 따른다.

$$U \sim N(0, \sigma_u^2 I)$$

이 가정은 최소자승추정량에 대한 확률분포 도출을 위해 필요하다.

3.모형의 추정

모형 내 모수의 추정은 잔차 제곱의 합을 최소화하는 모수를 구하는 보통최소자승법을 이용할 수 있다. 행렬로 표현한 다중회귀모형((4.2)식)에서 잔차는 (4.4)식과 같다.

$$e = Y - X\hat{B} \qquad (4.4)$$

(4.4)식을 이용하여 최소화 문제에서 목적함수인 잔차의 제곱의 합($e'e$)을 나타내면 (4.5)식과 같게 된다.

$$\begin{aligned} Min\ (e'e) &= (Y - X\hat{B})'(Y - X\hat{B}) \\ &= (Y' - \hat{B}'X')(Y - X\hat{B}) \quad (\because (A+B)' = A' + B', (AB)' = B'A') \\ &= (Y'Y - Y'X\hat{B} - \hat{B}'X'Y + \hat{B}'X'X\hat{B}) \\ &= (Y'Y - 2\hat{B}'X'Y + \hat{B}'X'X\hat{B}) \end{aligned} \tag{4.5}$$

(4.5)식을 최소가 되게 하는 추정량 $\hat{B}$을 구하는 것이 최소자승법인데 (4.5)식을 $\hat{B}$으로 편미분하면 (4.6)식과 같은 정규방정식이 도출된다.

$$\frac{\partial (e'e)}{\partial \hat{B}} = -2X'Y + 2X'X\hat{B} = 0 \quad (\because \frac{\partial (x'Ax)}{\partial x} = 2Ax)$$

$$X'X\hat{B} = X'Y \tag{4.6}$$

(4.6)식으로부터 (4.7)식과 같은 $\hat{B}$에 대한 OLS 추정량을 구할 수 있다.

$$\therefore\ \hat{B} = (X'X)^{-1}X'Y \tag{4.7}$$

(참고 1) 행렬의 미분

① 1차형식(linear form)

$$\frac{\partial (a'x)}{\partial x} = a$$

② 2차형식(quadratic form)

$$\frac{\partial (x'Ax)}{\partial x} = 2Ax \text{ 또는 } \frac{\partial (x'Ax)}{\partial x} = 2x'A$$

회귀모형에서 독립변수의 상대적인 중요성을 알고자 할 경우 독립변수 및 종속

변수를 표준화한 다음 표준화된 변수로 설정한 (4.8)식과 같은 회귀모형을 추정하면 되는데 이때 추정된 회귀계수를 베타회귀계수(Beta coefficients or standardized regression coefficients)라고 한다.

$$\frac{Y_i - \overline{Y}}{s_Y} = \beta_2^* \frac{X_{2i} - \overline{X_2}}{s_{X_2}} + \ldots + \beta_k^* \frac{X_{ki} - \overline{X_k}}{s_{X_k}} + u_i \qquad (4.8)$$

(4.8)식에 의해 추정된 베타회귀계수($\widehat{\beta^*}$)와 (4.1)식에 의해 추정된 회귀계수($\hat{\beta}$)는 (4.9)식과 같은 관계가 있다.

$$\widehat{\beta_j^*} = \hat{\beta}_j \frac{s_{X_j}}{s_{Y,}} (j = 2,3,...,k) \qquad (4.9)$$

한편, 추정된 베타회귀계수의 값에 대한 해석은 다음과 같다. 예를 들어 그 값이 0.7이면 독립변수가 표준편차 하나 크기(one standard deviation change in the independent variable)만큼 변할 때 종속변수의 표준편차는 0.7 크기만큼 변한다고 해석한다.

(예제 4-1)
다음의 홍보비 지출액 X_2(단위: 천만 원) 및 연구개발 지출액 X_3(단위: 천만 원)과 연간매출액 Y(단위: 십억 원)에 관한 자료를 이용하여 회귀계수와 회귀식을 구하고 해석하라.

Y	1	1	2	3
X_2	1	2	3	2
X_3	2	1	1	2

(풀이)

먼저 주어진 자료를 행렬로 표현하면 다음과 같다.

$$Y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 3 \end{bmatrix}, \quad X = \begin{bmatrix} 1\,1\,2 \\ 1\,2\,1 \\ 1\,3\,1 \\ 1\,2\,2 \end{bmatrix}$$

위 두 행렬을 이용하여 다음과 같은 행렬식, 역행렬 등을 구할 수 있다.

$$X'X=\begin{bmatrix}4 & 8 & 6\\ 8 & 18 & 11\\ 6 & 11 & 10\end{bmatrix},\quad |X'X|=4$$

$$(X'X)^{-1}=\begin{bmatrix}14.75 & -3.5 & -5\\ -3.5 & 1 & 1\\ -5 & 1 & 2\end{bmatrix},\quad X'Y=\begin{bmatrix}7\\ 15\\ 11\end{bmatrix}$$

(4.7)식에 의해 다중회귀모형에서 OLS 추정량은 다음과 같이 구할 수 있다.

$$\therefore \hat{B}=(X'X)^{-1}X'Y=\begin{bmatrix}-4.25\\ 1.5\\ 2\end{bmatrix}$$

따라서 추정된 회귀식은 $\hat{Y}=-4.25+1.5X_2+2X_3$이고 그 해석은 다음과 같다. 연구개발 지출액이 변화하지 않을 경우 홍보비 지출액이 천만 원 증가하면 연간 매출액은 평균 15억 원 증가하며, 홍보비 지출액이 변화하지 않을 경우 연구개발 지출액이 천만 원 증가하면 연간매출액은 평균 20억 원 증가함을 의미한다.

4.추정량의 특성

OLS 추정량 $\hat{B}$ 역시 최량선형불편추정량(Best Linear Unbiased Estimator: BLUE)이 되는데 회귀계수의 선형성, 불편성 및 최소분산의 특성에 대해 살펴보자.

(1)선형성

회귀계수의 선형성(linearity in Y)에 대해 살펴보면 (4.7)식의 $\hat{B}=(X'X)^{-1}X'Y$에서 X는 비확률변수이므로 $\hat{B}$은 Y에 대해서 선형이다. $\hat{B}=(X'X)^{-1}X'Y$에서 $(X'X)^{-1}X'$는 제3장 (3.23)식에 나타나 있는 $\frac{x_i}{\sum x_i^2}$의 행렬식 표현이다.

(2)불편성

회귀계수의 불편성(unbiasedness)에 대해 살펴보기 위해 (4.7)식으로부터 다음의 (4.10)식을 도출할 수 있다.

$$\begin{aligned}\hat{B} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(XB+U) \\ &= B+(X'X)^{-1}X'U \end{aligned} \qquad (4.10)$$

(4.10)식에 기댓값을 계산하면 (4.11)식과 같게 되고 $\hat{B}$은 B의 불편추정량(unbiased estimator)이다.

$$\begin{aligned}E(\hat{B}) &= B+(X'X)^{-1}X'E(U) \\ &= B \end{aligned} \qquad (4.11)$$

(3)최소분산

회귀계수 $\hat{B}$의 최소분산(minimum variance)을 살펴보기 위해 다른 선형추정량을 $\tilde{B}=[(X'X)^{-1}X'+C]Y$라 하고 이를 전개하면 (4.12)식이 된다.

$$\begin{aligned}\tilde{B} &= [(X'X)^{-1}X'+C]Y \\ &= [(X'X)^{-1}X'+C](XB+U) \\ &= B+CXB+[(X'X)^{-1}X'+C]U \end{aligned} \qquad (4.12)$$

(4.12)에 다음과 같이 기댓값을 구해보면 $\tilde{B}$가 불편추정량이 되기 위해서는 $CX=0$이 되어야 한다.

$$\begin{aligned}E(\tilde{B}) &= B+CXB \\ &= B \text{ if } CX=0 \end{aligned}$$

한편, $\tilde{B}$의 분산을 계산하기 위해 (4.12)식을 다음과 같이 정리한다.

$$\tilde{B}-B=[(X'X)^{-1}X'+C]U$$

이를 이용하여 $\tilde{B}$의 분산을 계산해 보면 (4.13)식과 같게 된다.

$$\begin{aligned} Var(\tilde{B}) &= E[(\tilde{B}-B)(\tilde{B}-B)'] \\ &= E[((X'X)^{-1}X'U+CU)((U'X(X'X)^{-1}+U'C')] \\ &= E[((X'X)^{-1}X'UU'X(X'X)^{-1})+((X'X)^{-1}X'UU'C') \\ &\quad +(CUU'X(X'X)^{-1})+CUU'C'] \\ &= \sigma_u^2[(X'X)^{-1}+CC'] \end{aligned} \tag{4.13}$$

다음에서 논의하겠지만 $Var(\hat{B})=\sigma_u^2(X'X)^{-1}$이므로 $Var(\tilde{B})$는 $Var(\hat{B})$보다 CC'(nonnegative definite)만큼 더 크다. 따라서 $\hat{B}$은 다른 어떤 선형불편추정량보다 분산이 작게 되는데 이러한 특성을 가진 OLS 추정량을 최량선형불편추정량(Best Linear Unbiased Estimator: BLUE)이라고 한다.

5.회귀계수의 분산 추정

회귀계수에 대한 통계적 검정을 하기 위해서는 회귀계수의 분산을 추정해야 한다. 그러나 회귀계수의 분산 추정량은 교란항의 분산 추정량의 함수이기 때문에 회귀계수의 분산을 추정하기 위해서는 먼저 교란항의 분산인 σ_u^2을 추정한다.

다중회귀분석에서 교란항의 분산 σ_u^2에 대한 좋은 추정량(불편추정량)은 $\hat{\sigma_u^2}=\frac{e'e}{n-k}$로서 k는 상수항을 포함한 독립변수의 수이다.

한편 $e'e$는 (4.5)식을 정리하면 얻어지는 (4.14)식에 이용하여 구할 수 있다.

$$\begin{aligned} e'e &= Y'Y-2\hat{B}'X'Y+\hat{B}'X'X\hat{B} \\ &= Y'Y-2\hat{B}'X'X\hat{B}+\hat{B}'X'X\hat{B} \\ &= Y'Y-\hat{B}'X'X\hat{B} \end{aligned}$$

또는 $Y'Y-\hat{B}'X'Y$ (4.14)

다음으로 회귀계수 $\hat{B}$의 분산을 추정해 보자. (4.7)식에서 다음의 (4.15)식을 구할 수 있다.

$$\begin{aligned}\hat{B} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(XB+U) \\ &= (X'X)^{-1}X'XB+(X'X)^{-1}X'U \\ &= B+(X'X)^{-1}X'U \end{aligned} \tag{4.15}$$

한편, $\hat{B}$의 분산을 계산하기 위해 (4.15)식을 다음과 같이 정리한다.

$$\hat{B}-B=(X'X)^{-1}X'U$$

이를 이용하여 $\hat{B}$의 분산을 계산해 보면 (4.16)식과 같게 된다.

$$\begin{aligned}Var(\widehat{B}) &= E[(\hat{B}-E(\widehat{B}))(\hat{B}-E(\widehat{B}))'] \\ &= E[(\hat{B}-B)(\hat{B}-B)'] \\ &= (X'X)^{-1}X'E(UU')X(X'X)^{-1}(\because (A^{-1})'=(A')^{-1}) \\ &= \sigma_u^2(X'X)^{-1} \end{aligned} \tag{4.16}$$

(4.16)식을 회귀계수의 분산-공분산행렬(variance-covariance matrix)이라고 하는데 이를 자세히 나타내면 다음과 같다.

$$Var(\widehat{B})=\begin{bmatrix} Var(\widehat{\beta_1}) & Cov(\widehat{\beta_1},\widehat{\beta_2}) & \cdots & Cov(\widehat{\beta_1},\widehat{\beta_k}) \\ Cov(\widehat{\beta_2},\widehat{\beta_1}) & Var(\widehat{\beta_2}) & \cdots & Cov(\widehat{\beta_2},\widehat{\beta_k}) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\widehat{\beta_k},\widehat{\beta_1}) & \vdots & \cdots & Var(\widehat{\beta_k}) \end{bmatrix} = \sigma_u^2(X'X)^{-1}$$

6.결정계수

제3장의 <그림 3-7>에서 보여 주고 있듯이 $Y = \hat{Y} + e$이다. 총변동인 $Y'Y$는 다음의 (4.17)식과 같이 나타낼 수 있다.

$$\begin{aligned} Y'Y &= (\hat{Y}+e)'(\hat{Y}+e) \\ &= (X\hat{B}+e)'(X\hat{B}+e) \\ &= (\hat{B}'X'+e')(X\hat{B}+e) \\ &= \hat{B}'X'X\hat{B}+2\hat{B}'X'e+e'e \\ &= \hat{B}'X'X\hat{B}+e'e\ (\because \hat{B}'X'e = \hat{B}'X'(Y-X\hat{B}) = \hat{B}'(X'Y-X'X\hat{B}) = 0) \end{aligned} \quad (4.17)$$

(4.17)식은 단순회귀분석과 마찬가지로 총변동 = 회귀변동 + 잔차변동의 관계가 성립함을 보여주고 있다.

결정계수의 정의에 따라 결정계수는 (4.18)식과 같이 구할 수 있다.

$$R^2 = \frac{\hat{B}'X'X\hat{B}}{Y'Y} \quad (\text{단, } \overline{Y} = 0 \text{일 때}) \quad (4.18)$$

만약에 $\overline{Y} \neq 0$이면 수정항(correction term) $n\overline{Y^2}$이 필요하다. 이때 결정계수는 다음의 (4.19)식과 같다.

$$R^2 = \frac{\hat{B}'X'X\hat{B} - n\overline{Y^2}}{Y'Y - n\overline{Y^2}} \quad (4.19)$$

결정계수는 설명변수의 설명력을 나타내므로 클수록 모형의 설명력이 높고, 모형의 적합도가 높다는 것을 의미하므로 좋다. 그러나 해석에 있어서 주의해야 할 사항이 있다. 설명변수가 비록 통계적으로 유의하지 않더라고 설명변수가 추가되기만 하면 결정계수는 높아진다. 따라서 설명변수의 통계적 유의성과 결정계수를 모두 고려해야 한다. 또한 종속변수의 자료형태가 다를 경우 결정계수를 비교할 수 없다. 예를 들어 비록 독립변수가 동일하다 하더라도 한 모형에서는 종속변수의 값을 그대로 사용하고 다른 모양에서는 종속변수의 로그치를 사용했다면 두 모형의 결정계수는 비교할 수 없다.

(예제 4-1)(계속)
다음의 홍보비 지출액 X_2(단위: 천만 원) 및 연구개발 지출액 X_3(단위: 천만 원)과 연간매출액 Y(단위: 십억 원)에 관한 자료를 이용하여 교란항의 분산, 회귀계수의 분산 및 결정계수를 구하라.

Y	1	1	2	3
X_2	1	2	3	2
X_3	2	1	1	2

(풀이)

$$e'e = Y'Y - \hat{B}'X'X\hat{B}$$

$$= [1\,1\,2\,3]\begin{bmatrix}1\\1\\2\\3\end{bmatrix} - [-4.25\;1.5\;2]\begin{bmatrix}4 & 8 & 6\\8 & 18 & 11\\6 & 11 & 10\end{bmatrix}\begin{bmatrix}-4.25\\1.5\\2\end{bmatrix}$$

$$= 15 - 14.75$$
$$= 0.25$$

$$\therefore \hat{\sigma_u^2} = \frac{0.25}{4-3} = 0.25$$

$$\therefore R^2 = \frac{\hat{B}'X'X\hat{B} - n\overline{Y^2}}{Y'Y - n\overline{Y^2}} = \frac{14.75 - 4(1.75)^2}{15 - 4(1.75)^2} = \frac{2.5}{2.75} = 0.909$$

$$\therefore Var(\hat{B}) = \hat{\sigma_u^2}(X'X)^{-1}$$

$$= (0.25)\begin{bmatrix}14.75 & -3.5 & -5\\-3.5 & 1 & 1\\-5 & 1 & 2\end{bmatrix}$$

$$= \begin{bmatrix}3.6875 & & \\ & 0.25 & \\ & & 0.5\end{bmatrix}$$

7.가설검정

회귀계수에 대한 가설검정은 주어진 회귀모형으로부터 회귀계수에 대한 귀무가설을 설정한 후 유의성 검정을 한다. 귀무가설은 다양하게 설정할 수 있지만 일반적으로 회귀계수의 값이 0이라는 귀무가설을 설정하는데 이는 독립변수가 종속변수에 영향을 주지 못한다는 것이다.

다중회귀분석에서 가설검정은 단순회귀분석에서의 가설검정을 확장한 것이기는 하지만 다양한 가설검정이 있다. 개별 독립변수의 통계적 유의성을 검정하는 단일가설검정(single hypothesis test), 여러 개 독립변수의 통계적 유의성을 동시에 검정하는 결합가설검정(joint hypothesis test), 모형 내 모든 독립변수의 통계적 유의성을 검정하는 전체검정(test of overall relationship) 등의 방법이 있다.

이러한 다양한 가설검정을 수행하기 위한 귀무가설을 행렬의 형태로 나타내면 (4.20)식과 같다.

$$H_0 : RB = r \tag{4.20}$$

단, R은 qxk의 주어진 행렬
r은 qx1의 주어진 벡터
q는 제약의 수($q \leq k$)

(1)단일 회귀계수 검정

하나의 회귀계수(예를 들어 i번째 독립변수의 통계적 유의성 검정)에 대한 귀무가설은 $H_0 : RB = r$으로 R과 r의 값 및 귀무가설은 각각 다음과 같다.

$R = [0\,0 \ldots 1\,0 \ldots 0]$과 $r = 0$

$H_0 : \beta_i = 0$ (q=1)

이상과 같은 일반적인 귀무가설 하에서 (4.21)식의 F-통계량은 분자의 자유도가 1이고 분모의 자유도가 n-k인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

$$F = \frac{\hat{\beta}_i^2}{\hat{\sigma}_u^2 a_{ii}} \sim F_{n-k}^1 \qquad (4.21)$$

단, a_{ii}는 $(X'X)^{-1}$에서 i번째 대각원소를 나타낸다.

(4.21)식의 F-통계량은 i번째 독립변수의 통계적 유의성을 검정하는 (4.22)식의 t-통계량을 제곱한 것과 같으므로 단일가설검정은 결국 단순회귀분석에서 개별 회귀계수에 대한 유의성 검정인 t-검정과 같은 것이라고 할 수 있다.

$$t = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}_u^2 a_{ii}}} \sim t_{n-1} \qquad (4.22)$$

단, a_{ii}는 $(X'X)^{-1}$에서 i번째 대각원소를 나타낸다.

따라서 개별 회귀계수에 대한 F-검정 또는 t-검정 결과 귀무가설을 기각할 수 없으면 그 독립변수는 종속변수에 통계적으로 유의한 변수가 아니므로 모형에서 빠져야 하고, 귀무가설이 기각되면 모형에 그대로 남아 있어야 한다.

(2)결합검정

여러 개의 회귀계수에 대한 결합검정에서 귀무가설은 $H_0 : RB = r$로 R과 r의 값 및 귀무가설은 각각 다음과 같다.

$$R = \begin{bmatrix} 0\,0\,1\,0\,0\,0 \ldots 0 \\ 0\,0\,0\,1\,0\,0 \ldots 0 \\ 0\,0\,0\,0\,1\,0 \ldots 0 \end{bmatrix} \text{과 } r = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ (q=3)

이상과 같은 일반적인 귀무가설 하에서 (4.23)식 및 (4.24)식의 F-통계량은 분자의 자유도가 q이고 분모의 자유도가 n-k인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/n-k} \sim F_{n-k}^q \quad (4.23)$$

단, 여기서 R_{ur}^2 및 R_{ur}^2 은 각각 제약이 가해지지 않은 모형과 제약이 가해진 모형에서의 결정계수이다.

$$F = \frac{(e'e_r - e'e_{ur})/q}{(e'e_{ur})/n-k} \sim F_{n-k}^q \quad (4.24)$$

단, $e'e_r$ 및 $e'e_{ur}$은 각각 제약이 가해진 모형과 제약이 가해지지 않은 모형에서 잔차의 제곱의 합이다.

따라서 여러 개의 회귀계수에 대한 F-검정 결과 귀무가설을 기각할 수 없으면 해당 독립변수들은 종속변수에 통계적으로 유의한 변수가 아니므로 모형에서 동시에 빠져야 하고, 귀무가설이 기각되면 해당 독립변수들을 모형에 그대로 남겨두는 것은 아니라 개별 회귀계수에 대한 검정을 통해서 각각 모형 내 존치 여부를 결정해야 한다.

예를 들어 다음의 (4.25)식과 같은 화폐수요함수를 나타내는 다중회귀모형이 있다고 하자.

$$m_t = \beta_1 + \beta_2 m_{t-1} + \beta_3 Y_t + \beta_4 rb_t + \beta_5 rc_t + u_t \quad (4.25)$$

단, m_t및 m_{t-1}은 각각 t기 및 t-1기의 화폐수요, Y_t는 국내총생산, rb_t는 제1금융권 금리, rc_t는 제2금융권 금리를 각각 나타낸다.

(4.25)식을 제약이 가해지지 않은 모형이라고 하고 이 식을 추정한 후 얻게 되는 결정계수를 R_{ur}^2 , 잔차의 제곱의 합을 $e'e_{ur}$라고 한다. (4.25)식의 화폐수요를 나타내는 다중회귀모형에서 먼저 우리는 경제적으로 유의한 것으로 판단하여 모형에 포함시킨 독립변수들이 통계적으로도 유의한 지 살펴보아야 한다. 그러기 위해서는 개별 회귀계수에 대한 통계적 유의성을 t-검정을 통해 판단할 수 있다.

또한 우리는 동 모형에 금리에 관한 변수가 2개 있으므로 각각의 금리가 화폐수요에 영향을 주는 지에 관심이 있을 뿐 아니라 제1금융권 금리 및 제2금융권 금리가 동시에 화폐수요에 영향을 주는 지의 여부에 대해서도 관심이 있다. 여러 개의 회귀계수에 대한 결합검정에서 귀무가설은 $H_0 : RB = r$로 이 경우 R과 r의 값 및 귀무가설은 각각 다음과 같다.

$$R=\begin{bmatrix}0\,0\,0\,1\,0\,0\,\ldots\,0\\0\,0\,0\,0\,1\,0\,\ldots\,0\end{bmatrix}\text{과 }\ r=\begin{bmatrix}0\\0\end{bmatrix}$$

H_0:β$_4$= β$_5$= 0 (q=2)

이상의 귀무가설 하에서 (4.25)식의 화폐수요함수는 (4.26)식과 같이 나타낼 수 있는데 이를 제약이 가해진 모형이라고 한다.

$$m_t=\beta_1^*+\beta_2^* m_{t-1}+\beta_3^* Y_t+u_t^* \qquad (4.26)$$

(4.26)식을 추정한 후 얻게 되는 결정계수를 R_r^2, 잔차의 제곱의 합을 $e'e_r$이라고 한다. (4.25)식 및 (4.26)식의 추정을 통해 구한 R_{ur}^2 및 R_r^2과 (4.23)식을 이용하면 다음의 (4.27)식은 분자의 자유도가 2이고 분모의 자유도가 n-k인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

$$F=\frac{(R_{ur}^2-R_r^2)/2}{(1-R_{ur}^2)/n-k}\sim F_{n-k}^2 \qquad (4.27)$$

가설검정 결과 귀무가설을 기각할 수 없으면 제1금융권 금리 및 제2금융권 금리가 동시에 화폐수요에 영향을 주지 않으므로 화폐수요함수에서 동시에 빠져야 하며, 귀무가설이 기각되면 제1금융권 금리 및 제2금융권 금리 각각에 대한 가설검정을 통해서 모형 내 존치 여부를 결정해야 한다.

(3)전체검정

모형 내 모든 회귀계수의 통계적 유의성을 동시에 검정하는 전체검정에서 귀무가설은 $H_0: RB=r$으로 R과 r의 값 및 귀무가설은 각각 다음과 같다.

$$R=\begin{bmatrix}0\,1\,0\,0\,\ldots\,0\\0\,0\,1\,0\,\ldots\,0\\0\,0\,0\,1\,\ldots\,0\\.\ .\ .\ .\ \ldots\ .\\0\,0\,0\,0\,\ldots\,1\end{bmatrix}\text{과 }\ r=\begin{bmatrix}0\\0\\0\\.\\0\end{bmatrix}$$

H_0: $\beta_2 = \beta_3 = \beta_4 = ... = \beta_k = 0$ (q=k-1)

이상과 같은 일반적인 귀무가설 하에서 (4.28)식의 F-통계량은 분자의 자유도가 k-1이고 분모의 자유도가 n-k인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

$$F = \frac{R^2/k-1}{(1-R^2)/n-k} \sim F_{n-k}^{k-1} \qquad (4.28)$$

따라서 전체검정 결과 귀무가설을 기각할 수 없으면 상수항을 제외한 모든 독립변수들은 종속변수에 통계적으로 유의한 변수가 아니므로 모형설정이 완전히 잘못된 것을 의미한다. 따라서 모형설정을 완전히 새롭게 해야 한다.

(예제 4-1)(계속)
다음의 홍보비 지출액 X_2(단위: 천만 원), 연구개발 지출액 X_3(단위: 천만 원)과 연간매출액 Y(단위: 십억 원)에 관한 자료를 이용하여 X_2 및 X_3이 각각 Y에 영향을 주는 지의 여부와 두 변수가 동시에 Y에 영향을 주는 지의 여부를 검정하라.

Y	1	1	2	3
X_2	1	2	3	2
X_3	2	1	1	2

(풀이)

먼저 개별 회귀계수의 통계적 유의성을 검정해 보자.

홍보비 지출액이 연간매출액에 영향을 주지 않는다는 가설을 검정하기 위하여 다음과 같은 귀무가설 및 대립가설을 설정한다.

$H_0 : \beta_2 = 0$
$H_1 : \beta_2 \neq 0$

t-통계량을 계산해 보면 다음과 같이 3.0으로 계산된다.

$$t = \frac{1.5 - 0}{0.5} = 3$$

5% 유의수준($\alpha = 0.05$)에서 자유도가 3일 때 임계값은 3.182 즉, $t_{3,0.025} = 3.182$이다. 위에서 계산한 t-통계량의 값 3.0은 임계값 3.182보다 작으므로 5% 유의수준 하에서 귀무가설을 기각하지 못한다. 따라서 홍보비 지출액은 연간매출액에 영향을 주지 못한다고 해석하고 $\hat{\beta}_2$는 통계적으로 유의하지 않다(statistically insignificant)라고 한다. 이 경우 홍보비 지출액의 독립변수는 모형에서 삭제하여야 한다.

연구개발 지출액이 연간매출액에 영향을 주지 않는다는 가설을 검정하기 위하여 위와 동일한 방법으로 귀무가설 및 대립가설을 설정하고, t-통계량을 계산해 보면 다음과 같이 2.828로 계산된다.

$$t = \frac{2.0 - 0}{0.7071} = 2.828$$

이와 같이 계산한 t-통계량의 값이 5% 유의수준에서의 임계값 3.182보다 작으므로 5% 유의수준 하에서 귀무가설을 기각하지 못한다. 따라서 연구개발 지출액은 연간매출액에 영향을 주지 못한다고 해석하고 $\hat{\beta}_3$는 통계적으로 유의하지 않다(statistically insignificant)라고 한다. 이 경우 연구개발 지출액의 독립변수는 모형에서 삭제하여야 한다.

다음으로 홍보비 지출액 및 연구개발 지출액이 동시에 연간매출액에 영향을 주지 않는 지를 검정해 보자. 홍보비 지출액 및 연구개발 지출액이 개별적으로 연간매출액에 영향을 주지 않는다고 해서 두 독립변수가 동시에 종속변수에 영향을 주지 않는다고 말할 수는 없고 통계적 검정을 통해서 판단해야 한다. 이 가설을 검정하기 위하여 다음과 같은 귀무가설 및 대립가설을 설정한다.

$H_0 : \beta_2 = \beta_3 = 0$
$H_1 : H_0$이 사실이 아니다

모형 내에 상수항을 제외한 독립변수가 두 개 있으므로 이는 전체검정에 해당되므로 (4.27)식에 따라 F-통계량을 계산해 보면

$$F = \frac{R^2/k-1}{(1-R^2)/n-k} = \frac{0.909/2}{(1-0.909)/1} = 5.0$$이다.

5% 유의수준 하에서 F-분포의 임계값은 $F_1^2 = 199.5$이고 F-통계량의 값은 5이므로 귀무가설을 기각하지 못한다. 즉, 홍보비 지출액 및 연구개발 지출액은 연간 매출액에 영향을 주지 못하는 변수이며 따라서 모형설정이 완전히 잘못된 것으로 해석할 수 있다. 이 경우에는 새로운 독립변수로 모형을 설정한 후 위의 절차를 다시 거쳐야 한다.

8.예측

일반적으로 예측에는 제3장의 <그림 3-8>에서 나타내는 바와 같이 주어진 X에 대해 하나의 Y값을 구하는 점예측(point forecasting)과 점예측에 대한 신뢰구간을 구하는 구간예측(interval forecasting)이 있다.

독립변수 X가 특정한 값 X_0일 때 개별 Y_0의 예측치는 (4.29)식과 같다.

$$\widehat{Y_0} = x_0^{'}\widehat{B} \tag{4.29}$$

구간예측을 위해서는 실제치인 Y_0와 예측치 $\widehat{Y_0}$의 차이인 예측오차의 분산(σ_ϵ^2)을 알아야 하며 제3장에서도 살펴보았듯이 예측오차의 분산은 평균예측과 개별예측이 서로 다르다.

평균예측의 경우 예측오차의 분산은 (4.30)식과 같고, 개별예측의 경우 예측오차의 분산은 (4.31)식과 같다.

$$Var(\widehat{Y_0} \mid x_0) = \widehat{\sigma_u^2} x_0'(X'X)^{-1}x_0 \tag{4.30}$$

$$Var(Y_0 \mid x_0) = \widehat{\sigma_u^2}(1 + x_0'(X'X)^{-1}x_0) \tag{4.31}$$

따라서 개별 Y_0에 대한 $(1-\alpha)\times 100\%$ 예측구간은 (4.32)식과 같다.

$$\widehat{Y_0} \pm t_{\frac{\alpha}{2},n-k} \sqrt{(\widehat{\sigma_u^2}[1 + x_o^{'}(X'X)^{-1}x_0]} \quad (4.32)$$

(예제 4-1)(계속)
다음의 홍보비 지출액 X_2(단위: 천만 원) 및 연구개발 지출액 X_3(단위: 천만 원)과 연간매출액 Y(단위: 십억 원)에 관한 자료로 다중회귀모형을 설정하였고 모형의 설정이 제대로 되었다고 하자. X_2가 3천만 원이고, X_3이 2천만 원일 경우 연간매출액(개별예측)에 대한 점예측치 및 95% 예측구간을 구하라.

Y	1	1	2	3
X_2	1	2	3	2
X_3	2	1	1	2

(풀이)

$X_2 = 3,\ X_3 = 2$일 때 점예측치는 다음과 같다.

$$\begin{aligned}\widehat{Y_0} &= x_0^{'}\hat{\beta} \\ &= [1\,3\,2]\begin{bmatrix}-4.25\\ 1.5\\ 2.0\end{bmatrix} \\ &= 4.25\end{aligned}$$

개별예측의 예측오차 분산은 다음과 같다.

$$\begin{aligned}Var(\widehat{Y_0}) &= \widehat{\sigma_u^2}[1 + x_0^{'}(X'X)^{-1}x_0] \\ &= (0.25)(1 + [1\,3\,2]\begin{bmatrix}14.75 & -3.5 & -5\\ -3.5 & 1 & 1\\ -5 & 1 & 2\end{bmatrix}\begin{bmatrix}1\\3\\2\end{bmatrix} \\ &= 0.9375\end{aligned}$$

따라서 개별 예측치 Y_0에 대한 95% 예측구간은 다음과 같다.

$$\widehat{Y_0} \pm t_{\frac{\alpha}{2},n-k}\sqrt{(\widehat{\sigma_u^2}[1+x_o^{'}(X'X)^{-1}x_0])} = 4.25 \pm (12.706)(0.968) = [-8.049, 16.549]$$

지금까지 살펴본 다중회귀분석에서의 여러 결과들 즉, 회귀계수 추정량, 교란항의 분산 추정량, 회귀계수의 분산 추정량, 잔차의 제곱의 합, 결정계수, 예측오차의 분산 등을 제3장의 단순회귀분석에서 결과들과 비교해 정리해 보면 <표 4-1>과 같은데 다중회귀분석에서의 결과들은 행렬로 표현되었을 뿐이지 그 결과는 단순회귀분석에서의 결과와 동일함을 확인할 수 있다.

<표 4-1> 단순회귀 및 행렬표현 다중회귀의 비교

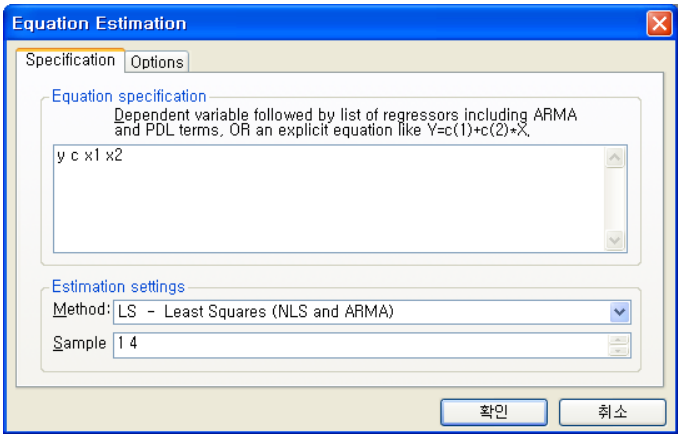
구분	단순회귀	다중회귀
$\hat{\beta}$	$\widehat{\beta_1} = \frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$	$\widehat{\mathrm{B}} = (X'X)^{-1}X'Y$
$\widehat{\sigma_u^2}$	$\frac{\sum e_i^2}{n-2}$	$\frac{e'e}{n-k}$
$Var(\widehat{\beta})$	$\widehat{\sigma_u^2}\frac{1}{\sum x_i^2}$	$\widehat{\sigma_u^2}(X'X)^{-1}$
$\sum e_i^2$	$\sum(Y_i-\overline{Y})^2 - \widehat{\beta_1}\sum(X_i-\overline{X})(Y_i-\overline{Y})$	$Y'Y - \widehat{B}'X'Y$
R^2	$\frac{\widehat{\beta_1^2}\sum(X_i-\overline{X})^2}{\sum(Y_i-\overline{Y})^2}$	$\frac{\widehat{B}'X'X\widehat{B} - n\overline{Y^2}}{Y'Y - n\overline{Y^2}}$
σ_ε^2 (평균예측)	$\widehat{\sigma_u^2}(\frac{1}{n} + \frac{(X_0-\overline{X})^2}{\sum x_i^2})$	$\widehat{\sigma_u^2}x_0'(X'X)^{-1}x_0$
σ_ε^2 (개별예측)	$\widehat{\sigma_u^2}(1 + \frac{1}{n} + \frac{(X_0-\overline{X})^2}{\sum x_i^2})$	$\widehat{\sigma_u^2}(1 + x_0'(X'X)^{-1}x_0)$

실습 : 다중회귀모형 추정

(예제 4-1)의 자료를 EViews를 이용하여 ex4-1.wf1로 저장한 후 홍보비 지출액 및 연구개발 지출액을 독립변수, 연간매출액을 종속변수로 하는 다중회귀모형을 추정한 후 여러 가지 결과를 살펴보도록 하자.

Quick-Estimate Equation을 선택하면 나오는 대화상자에서 <그림 4-A1>과 같이 입력을 한다. 여기서 y는 종속변수를 나타내고, c는 단순회귀모형의 상수항을 추정하기 위한 것이며, x1은 홍보비 지출액을 나타내는 독립변수, x2는 연구개발 지출액을 나타내는 독립변수를 나타낸다.

<그림 4-A1> 다중회귀모형(예제 4-1) 추정회귀식



그림과 같이 입력한 후 확인을 클릭하면 <그림 4-A2>와 같은 추정결과를 나타내 준다. 추정된 회귀식은 $\hat{Y}=-4.25+1.5X_2+2X_3$이고 그 해석은 다음과 같다.

연구개발 지출액이 변화하지 않을 경우 홍보비 지출액(X1)이 1단위 즉, 천만 원 증가하면 연간매출액은 평균 1.5단위 즉, 15억 원 증가하며, 홍보비 지출액이 변화하지 않을 경우 연구개발 지출액(X2)이 1단위 즉, 천만 원 증가하면 연간매출액은 평균 1단위 즉, 20억 원 증가함을 의미한다.

홍보비 지출액 및 연구개발 지출액 회귀계수의 표준오차는 각각 0.5, 0.707107이고 t-통계량은 각각 3, 2.828427이다. 이들 t-통계량에 대한 유의확률은 각각

0.2048, 0.2163으로서 두 독립변수 모두 5% 유의수준에서 통계적으로 유의한 변수는 아니다.

결정계수 R^2의 값이 0.909091이므로 홍보비 지출액 및 연구개발 지출액으로 연간매출액의 변화를 90.91% 설명할 수 있다. 제3장의 단순회귀분석에서 홍보비 지출로 연간매출액의 변화를 81.67% 설명할 수 있었던 것에 비해 설명력이 증가하였다. 여기서 주의할 점은 다중회귀모형에서 추가된 연구개발 지출액의 독립변수가 통계적으로 유의한 변수라면 결정계수가 높아지는 것은 당연하지만 통계적으로 유의한 변수가 아니라 하더라도 결정계수가 증가한다는 것이다. 즉, 개별 독립변수의 통계적 유의성에 관계없이 독립변수가 추가되면 결정계수는 커진다.

F-statistic이 5인데 이는 전체검정에 대한 F-통계량의 값이며, 이 F-통계량에 대한 유의확률이 0.301511이므로 모형의 설정이 완전히 잘못 되었다고 하는 귀무가설을 기각할 수 없다.

<그림 4-A2> 다중회귀모형(예제 4-1) 추정결과

Equation: EQ01 Workfile: EX4-1::Untitled

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 04/02/10 Time: 00:18
Sample: 1 4
Included observations: 4

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-4.250000	1.920286	-2.213211	0.2702
X1	1.500000	0.500000	3.000000	0.2048
X2	2.000000	0.707107	2.828427	0.2163

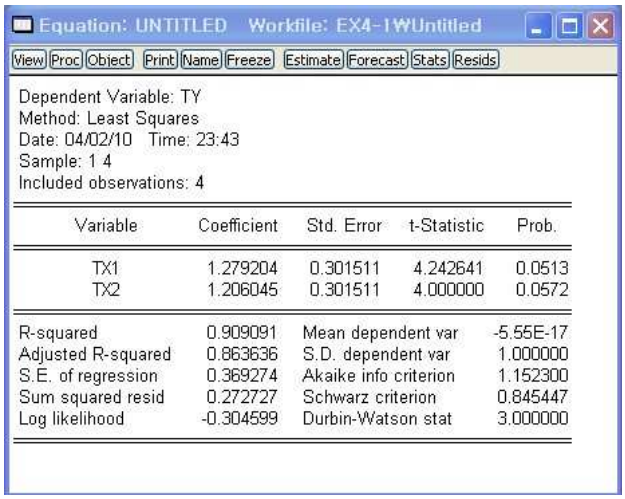
R-squared	0.909091	Mean dependent var	1.750000
Adjusted R-squared	0.727273	S.D. dependent var	0.957427
S.E. of regression	0.500000	Akaike info criterion	1.565288
Sum squared resid	0.250000	Schwarz criterion	1.105009
Log likelihood	-0.130577	F-statistic	5.000000
Durbin-Watson stat	3.000000	Prob(F-statistic)	0.301511

한편, 독립변수의 상대적인 중요성을 알아보기 위해 베타회귀계수를 구하고자 할 경우 먼저 독립변수 및 종속변수를 표준화한 다음 (4.8)식과 같이 상수항을 제외한 회귀모형을 추정해야 한다. 따라서 EViews의 명령어 입력창에서 다음을 입력하여 자료를 표준화한다.

```
genr ty=(y-@mean(y))/@stdev(y)
genr tx1=(x1-@mean(x1))/@stdev(x1)
genr tx2=(x2-@mean(x2))/@stdev(x2)
```

다음으로 표준화된 자료로 다중회귀모형을 추정하는데 즉, ty를 종속변수로 두고 tx1 및 tx2를 독립변수로 하는 회귀모형을 추정하면 <그림 4-A3>와 같은 추정결과를 나타내 준다.

<그림 4-A3> (예제 4-1)베타회귀계수 추정


Equation: UNTITLED Workfile: EX4-1₩Untitled

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: TY
Method: Least Squares
Date: 04/02/10 Time: 23:43
Sample: 1 4
Included observations: 4

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TX1	1.279204	0.301511	4.242641	0.0513
TX2	1.206045	0.301511	4.000000	0.0572

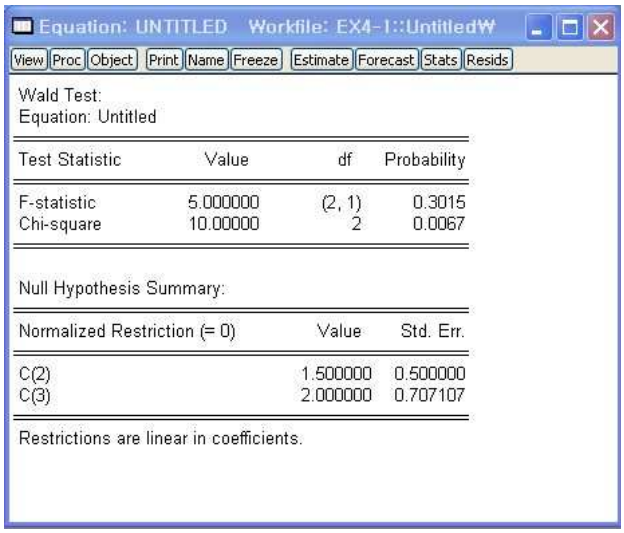
R-squared	0.909091	Mean dependent var	-5.55E-17
Adjusted R-squared	0.863636	S.D. dependent var	1.000000
S.E. of regression	0.369274	Akaike info criterion	1.152300
Sum squared resid	0.272727	Schwarz criterion	0.845447
Log likelihood	-0.304599	Durbin-Watson stat	3.000000

베타회귀계수($\widehat{\beta_2^*}=1.279$)는 회귀계수($\widehat{\beta_2}\frac{s_{X_1}}{s_Y}=(1.5)\frac{(0.81497)}{(0.957427)}=1.279$)임을 보여 줄 수 있으며 β_3에 대해서도 이와 같이 (4.9)식이 성립함을 보여줄 수 있다. 베타회귀계수를 살펴보면 홍보비 지출액의 회귀계수가 연구개발 지출액의 회귀계수보다 크므로 연간매출액은 연구개발보다는 홍보비 지출액에 더 큰 영향을 받는다고 할 수 있다.

홍보비 지출액 및 연구개발 지출액이 동시에 연간매출액에 영향을 주지 않는지를 검정해 보자. 이를 위해 <그림 4-A2>와 같이 다중회귀모형을 추정한 후 View-Coefficient Tests-Wald-Coefficient Restrictions를 클릭하면 나타나는 Wald Test 대화상자에서 C(2)=0, C(3)=0을 입력하고 OK를 클릭하면 <그림 4-A4>

와 같은 가설검정의 결과를 보여준다. 여기서 F-통계량을 보면 되는데 그 값이 5이고, 분자의 자유도가 2이고 분모의 자유도가 1인 F-분포에서 F-통계량 5의 유의확률은 0.3015이므로 홍보비 지출액 및 연구개발 지출액이 동시에 연간매출액에 영향을 주지 않는다는 귀무가설을 기각할 수 없다.

<그림 4-A4> 가설검정(결합검정) 결과


Equation: UNTITLED Workfile: EX4-1::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	5.000000	(2, 1)	0.3015
Chi-square	10.00000	2	0.0067

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(2)	1.500000	0.500000
C(3)	2.000000	0.707107

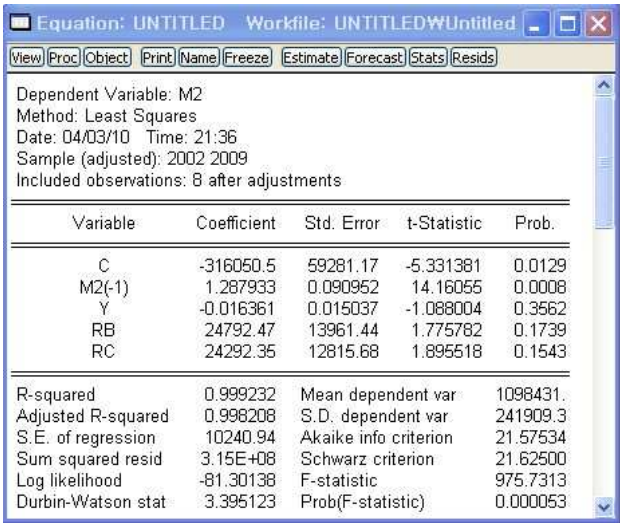
Restrictions are linear in coefficients.

(4.25)식의 화폐수요함수에서 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는다는 귀무가설을 검정하기 위하여 먼저 2001년부터 2009년까지 우리나라의 자료를 이용하여 (4.25)식을 추정한 결과가 <그림 4-A5>에 나타나 있다. 여기에 있는 결정계수는 제약이 가해지지 않은 모형의 결정계수(R_{ur}^2)로 0.999232이다.

한편, 귀무가설이 사실일 경우 즉 제약이 가해진 모형인 (4.26)식을 추정한 결과가 <그림 4-A6>에 나타나 있는데 여기에 있는 결정계수는 제약이 가해진 모형의 결정계수(R_r^2)로 0.987914이다.

(4.27)식을 이용하여 F-통계량을 구해보면 다음과 같이 22.105로서 5% 유의수준 하에서 분자의 자유도가 2이고 분모의 자유도가 3인 F-분포표의 임계치인 9.55보다 크므로 5% 유의수준 하에서 귀무가설을 기각한다.

<그림 4-A5> (4.25)식 추정결과



Equation: UNTITLED Workfile: UNTITLED₩Untitled

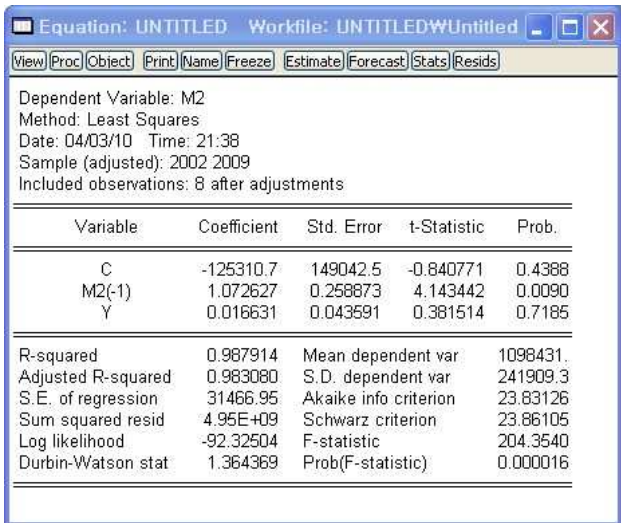
View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: M2
Method: Least Squares
Date: 04/03/10 Time: 21:36
Sample (adjusted): 2002 2009
Included observations: 8 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-316050.5	59281.17	-5.331381	0.0129
M2(-1)	1.287933	0.090952	14.16055	0.0008
Y	-0.016361	0.015037	-1.088004	0.3562
RB	24792.47	13961.44	1.775782	0.1739
RC	24292.35	12815.68	1.895518	0.1543

R-squared	0.999232	Mean dependent var	1098431.
Adjusted R-squared	0.998208	S.D. dependent var	241909.3
S.E. of regression	10240.94	Akaike info criterion	21.57534
Sum squared resid	3.15E+08	Schwarz criterion	21.62500
Log likelihood	-81.30138	F-statistic	975.7313
Durbin-Watson stat	3.395123	Prob(F-statistic)	0.000053

<그림 4-A6> (4.26)식 추정결과



Equation: UNTITLED Workfile: UNTITLED₩Untitled

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: M2
Method: Least Squares
Date: 04/03/10 Time: 21:38
Sample (adjusted): 2002 2009
Included observations: 8 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-125310.7	149042.5	-0.840771	0.4388
M2(-1)	1.072627	0.258873	4.143442	0.0090
Y	0.016631	0.043591	0.381514	0.7185

R-squared	0.987914	Mean dependent var	1098431.
Adjusted R-squared	0.983080	S.D. dependent var	241909.3
S.E. of regression	31466.95	Akaike info criterion	23.83126
Sum squared resid	4.95E+09	Schwarz criterion	23.86105
Log likelihood	-92.32504	F-statistic	204.3540
Durbin-Watson stat	1.364369	Prob(F-statistic)	0.000016

즉, 화폐수요함수에서 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는 것은 아니므로 금리에 대한 두 변수 중 하나는 반드

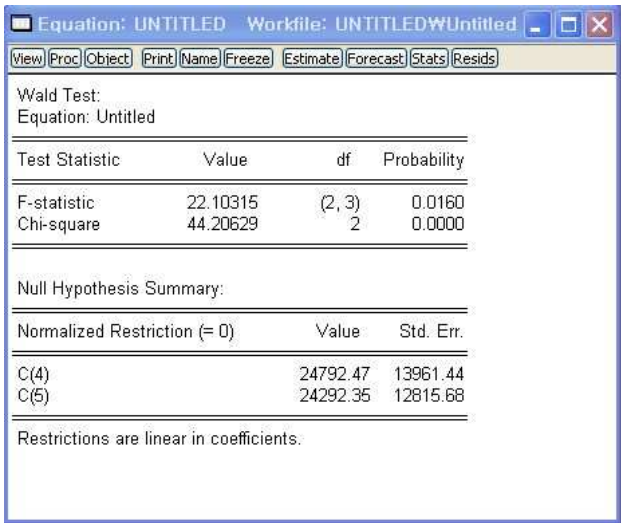
시 포함되어야 한다.

$$F = \frac{(R_{ur}^2 - R_r^2)/2}{(1-R_{ur}^2)/n-k} = \frac{(0.999232-0.987914)/2}{(1-0.999232)/3} = \frac{0.005659}{0.000256} = 22.105 \sim F_3^2$$

고 할 수 있다.

화폐수요함수에서 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는 지를 검정해 보자. 이를 위해 <그림 4-A5>와 같이 다중회귀모형을 추정한 후 View-Coefficient Tests-Wald-Coefficient Restrictions를 클릭하면 나타나는 Wald Test 대화상자에서 C(4)=0, C(5)=0을 입력하고 OK를 클릭하면 <그림 4-A7>와 같은 가설검정의 결과를 보여준다. 여기서 F-통계량을 보면 되는데 그 값이 22.10315이고, 분자의 자유도가 2이고 분모의 자유도가 3인 F-분포에서 F-통계량 22.10315의 유의확률은 0.016이므로 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는다는 귀무가설을 기각한다.

<그림 4-A7> 가설검정(결합검정) 결과



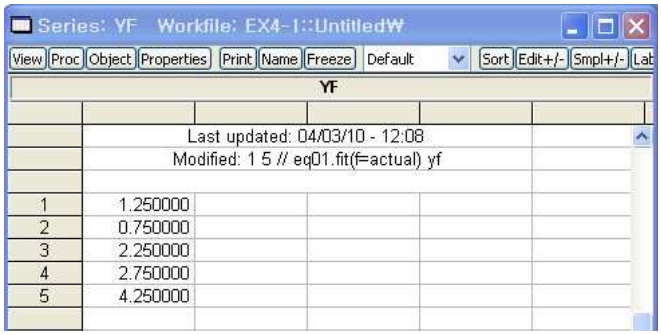
예측을 하기 위해서는 먼저 Proc-structure/resize current page를 선택하여

Data range를 5로 하여 예측을 위한 공간을 하나 더 확보하고, 독립변수를 각각 클릭하여 Edit+/-를 눌러 예측을 위한 독립변수의 값(여기서는 $X_2 = 3, X_3 = 2$)을 각각 입력한다. 회귀모형을 추정하고 추정결과 창에서 Forecast를 클릭하면 나타나는 예측 대화상자에서 S.E. (optional)에 sef를 입력하고 OK를 클릭하면 종속변수의 점예측치(yf) 및 개별예측치 예측오차의 표준오차(sef)가 계산된다.

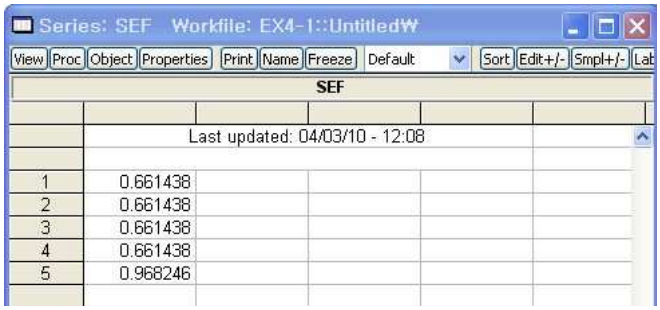
종속변수의 점예측치(yf)를 클릭해 보면 <그림 4-A8>과 같이 나타나는데 여기서 예측치가 4.25임을 알 수 있다.

한편, 개별예측치 예측오차의 표준오차(sef)를 클릭해 보면 <그림 4-A9>와 같이 나타나는데 여기서 개별예측치 예측오차의 표준오차가 0.968246임을 확인할 수 있고, 이를 이용하면 개별예측에 대한 예측구간 $[-8.049, 16.549]$을 구할 수 있다.

<그림 4-A8> 종속변수 예측치



<그림 4-A9> 개별예측치 예측오차의 표준오차



연습문제

1. 1996년부터 1999년까지 화폐수요(Y), 실질이자율(X_1),국민소득(X_2)에 관한 우리나라의 연도별자료를 이용하여 물음에 답하라.

연도	1996	1997	1998	1999
Y	1	1	2	3
X_1	-2	-1	-1	-2
X_2	1	2	3	2

① 다중회귀모형을 추정하고 추정된 회귀모형을 써라.

② 추정된 회귀계수의 부호는 화폐수요이론이 예측하는 바와 일치하나?

③ 이자율과 국민소득은 각각 통계적으로 유의한 설명변수인가?

④ 이자율과 국민소득은 화폐수요를 어느 정도 설명하고 있나?

⑤ 설명변수가 전혀 설명력을 갖지 못한다는 귀무가설을 5% 유의수준 하에서 검정하라.

⑥ 2000년의 실질이자율과 국민소득이 각각 2, 3이라고 할 때 2000년 화폐수요의 점예측치와 95% 예측구간을 각각 구하라.

2. 다중회귀모형을 $Y = XB + U$($Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{3i} + u_i$)로 설정하고 다음의 자료가 주어졌을 때 이를 이용하여 물음에 답하라.

Y	1	2	3	2
X_1	1	1	2	3
X_2	2	1	1	2

① 최소자승법으로 추정한 회귀계수를 각각 구하라.

② $\widehat{\sigma_u^2} = 0.1$일 때 회귀계수의 표준오차를 각각 구하라.

③ 다음의 가설을 5% 유의수준 하에서 검정하라.

$H_0 : \beta_2 = 0$
$H_1 : \beta_2 \neq 0$

④ $R^2 = 0.5$라고 할 때 다음의 가설을 5% 유의수준 하에서 검정하라.

$H_0 : \beta_2 = \beta_3 = 0$

3. 소비에 영향을 주는 요인을 소득과 물가상승률로 보고 회귀모형을 설정하였다. 다음과 같은 연도별 자료가 주어졌을 때 이를 이용하여 물음에 답하라.

	2006년	2007년	2008년	2009년
소비(단위:10만원)	1	2	1	3
소득(단위:100만원)	1	3	1	3
물가상승률(%)	2	2	3	1

① 최소자승법으로 추정한 회귀계수를 각각 구하라.

② 다른 요인은 변동이 없고 소득이 100만원 증가하면 소비는 얼마나 변하나?

③ 다른 요인은 변동이 없고 물가상승률이 1% point 증가하면 소비는 얼마나 변하나?

④ 교란항의 분산을 구하라.

⑤ 회귀계수의 분산을 모두 구하라.

⑥ 모형의 결정계수를 구하라.

⑦ 이자율은 화폐수요에 영향을 주지 않는다는 귀무가설을 5% 유의수준 하에서 검정하라.

⑧ 설명변수가 전혀 설명력을 갖지 못한다는 귀무가설을 5% 유의수준 하에서 검정하라

⑨ 2010년의 이자율과 국민소득이 각각 3, 2라고 할 때 2010년 화폐수요의 점예측치를 구하라.

⑩ 2010년의 이자율과 국민소득이 각각 3, 2라고 할 때 2010년 화폐수요의 95% 예측구간을 구하라.

⑪ 우리나라의 경우 화폐수요이론의 성립여부를 결정하라. 만약, 성립하지 않는다면 다음에는 어떤 식으로 접근해야 하는 지를 설명하라.

4. 다음은 2001년부터 2004년까지 물가상승률(inflation), 임금상승률(wage), 통화량증가율(money)의 연도별 자료이다. 이를 이용하여 다음 물음에 답하라.

연도	2001	2002	2003	2004
물가상승률(%)	1	1	2	3
임금상승률(%)	2	2	3	5
통화량증가율(%)	2	1	1	2

① "2001-2004년 기간 중 물가상승은 비용인상인플레이션으로 볼 수 있다"라는 주장을 평가하라.

② “2005년도 임금상승률이 6%, 통화량증가율이 3%라고 할 때 2005년도의 물가상

승률은 3.75%가 될 것이다"라는 주장을 평가하라.

5. 다음은 2005년부터 2008년까지 우리나라의 화폐수요, 이자율 및 국민소득의 연도별 자료이다. 이를 이용하여 다음 물음에 답하라.

연도	2005	2006	2007	2008
화폐수요(천조 원)	1	1	2	3
이자율(%)	2	3	2	1
국민소득(천조 원)	1	2	1	2

① 화폐수요이론에 따르면 화폐수요는 국민소득 및 이자율의 함수이다. 화폐수요이론의 성립여부를 우리나라의 자료를 이용하여 평가하라. 단, 통계적 유의성은 무시하라 즉, 통계적으로 유의하다고 가정하라.

② 화폐수요에 영향을 미치는 요인인 국민소득과 이자율 중 우리나라의 경우 어느 요인이 화폐수요에 더 큰 영향을 주는 지를 설명해 보라. 단, 통계적 유의성은 무시하라 즉, 통계적으로 유의하다고 가정하라.

③ 위의 두 문제에서 살펴 본 결론의 통계적 유의성에 대해 평가하라.

6. 다음은 여러 가지 통화지표를 이용하여 화폐수요함수를 추정한 것을 정리한 표이다.

모형	종속변수	상수항	시차변수	Y	RB	RCP	R^2	F검정유의수준
모형1	M1	-0.69	0.61	0.31	0.04 (0.73)	-0.19 (0.16)	0.979	0.213
모형2	M2	-2.09	0.43	0.62	-0.12 (0.03)	-0.02 (0.67)	0.998	0.001
모형3	M3	-2.98	0.61	0.65	-1.35 (0.19)	-1.32 (0.20)	0.988	0.001

* 괄호안의 숫자는 한계유의수준을 나타냄
* F검정은 은행이자율(RB)과 회사채수익률(RCP)이 동시에 화폐수요함수에 영향을 주지 않는다는 귀무가설을 검정한 것임

① 모형1과 모형2를 비교할 때 어느 모형이 적합한 가를 설명하라.

② 모형2와 모형3을 비교할 때 어느 모형이 적합한 가를 설명하라.

제 5 장

가변수

1.가변수모형
2.가변수모형의 유형
3.가변수의 응용
4.가변수를 이용한 연구 실례
실습 : 가변수모형 추정

제5장 가변수

1.가변수모형

계량분석에 사용되는 대부분의 자료는 정량적인 자료이다. 그러나 필요에 따라서는 정성적인 자료를 사용할 경우가 있는데 정성변수를 독립변수로 사용하는 경우와 정성변수를 종속변수로 사용하는 경우 두 가지가 있다. 본 장에서 다루고자 하는 내용은 정성변수를 독립변수로 사용하는 경우이며 정성변수를 종속변수로 사용하는 경우를 이산선택모형(discrete choice model)이라고 한다.

남녀 간에 임금격차가 존재하는 지 또는 학력별로 임금격차가 존재하는 지를 살펴보거나 특정 시점을 전후하여 경제구조에 어떤 변화가 있는 지를 살펴 볼 경우 독립변수에 정성변수를 포함시켜 분석할 수 있는데 독립변수에 정성변수인 가변수를 포함하여 분석하는 모형을 가변수모형이라고 한다.

한편, 소득, 나이, 학력, 성(gender), 지역, 종교 등의 개인의 특성(individual characteristics)이 그 개인의 경제적 선택이나 정치적 선택에 미치는 영향을 살펴보고자 할 경우 개인의 특성변수를 독립변수로 하고 경제적 또는 정치적 선택행위를 정성변수인 가변수로 하여 종속변수로 처리할 수 있는데 이 경우를 이산선택모형이라고 한다.

가변수(dummy variable)란 독립변수 중의 일부가 성질을 달리하는 질적인 자료로 되어 있을 경우 사용된다. 계량분석에서 질적인 면을 나타내 주는 변수이며 가변수를 사용하면 더욱 정확한 통계적 추론을 할 수 있다. 왜냐하면 질적인 면을 고려해야 함에도 불구하고 이를 고려하지 않을 경우 이는 모형 내 있어야 할 변수를 뺀 경우이고 이 경우 최소자승추정량은 불편성을 가지지 못한다. 따라서 가변수를 사용해야 할 경제적 유의성 및 통계적 유의성이 있다면 이를 모형에 포함시켜 계량분석을 해야만 정확한 통계적 추론을 할 수 있다.

개인적인 특성을 나타내는 정성변수를 범주(category)라고 하는데 가변수는 질적 범주를 구별하기 위해 사용되는 변수이므로 모형에서 어느 한 질적 변수의 2개의 질적 범주를 구분하기 위해서는 상수항을 포함한 모형에서 하나의 가변수를 사용하거나 상수항을 제외한 모형에서 2개의 가변수를 사용할 수 있는데 통상적으로 상수항을 포함한 모형에서 하나의 가변수를 사용한다. 한편, 두 개 이상의 질적 변수나 어느 한 질적 변수의 두 개 이상의 질적 범주를 모형이 포함하고 있을 경우 두 개 이상의 가변수가 필요하다. 일반적으로 한 개의 질적 변수가 k개의 질적 범주가 있다면 상수항을 포함한 모형에서 k-1개의 가변수를 사용한다.

예를 들어 IMF 경제위기를 전후하여 소비행태에 변화가 있었는 지를 살펴본다고 하자. 만약에 가변수를 고려하지 않는다면 (5.1)식의 소비모형을 보통최소자승법으로 추정하면 되는데 이를 보통회귀식이라고 하자.

$$C_t = \beta_0 + \beta_1 Y_t + u_t \tag{5.1}$$

단, C_t는 소비, Y_t는 소득을 나타낸다.

IMF 경제위기를 경험하면서 경제주체의 소비행태에 변화가 있었는 지를 살펴보기 위해 가변수를 (5.2)식과 같이 모형에 포함시킬 수 있다.

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + u_t \tag{5.2}$$

단, $D_t = \begin{cases} 1, & t \text{ 가 } IMF\text{위기 이후} \\ 0, & t \text{ 가 } IMF\text{위기} \end{cases}$

한편 근무연수에 따라 임금이 변하는데 그 외에 성별 및 인종별 임금격차가 존재하는 지를 살펴본다고 하자. 이 경우는 성(gender)과 인종이라고 하는 두 개의 질적 변수를 가지고 있다. 만약에 가변수를 고려하지 않는다면 (5.3)식의 임금결정모형을 보통최소자승법으로 추정하면 될 것이다.

$$Y_t = \beta_0 + \beta_1 X_t + u_t \tag{5.3}$$

단, Y_t는 임금, X_t는 근무연수를 나타낸다.

근무연수에 따른 임금결정 외에 성별 및 인종별 임금격차가 존재하는 지를 살펴보기 위해 가변수를 (5.4)식과 같이 모형에 포함시킬 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{2i} + u_i \tag{5.4}$$

단, $D_{1i} = \begin{cases} 1, & i\text{가 남자} \\ 0, & i\text{가 여자} \end{cases}$

$D_{2i} = \begin{cases} 1, & i\text{가 백인} \\ 0, & i\text{가 유색인} \end{cases}$

또 다른 예를 들어 보자. 한편 근무연수에 따라 임금이 변하는데 학력에 따라 임금격차가 존재하는 지를 살펴본다고 하자. 이 경우는 학력이라고 하는 질적 변

수는 두 개 이상의 범주를 포함할 수 있다. 만약에 가변수를 고려하지 않는다면 (5.5)식의 임금결정모형을 보통최소자승법으로 추정하면 된다.

$$Y_t = \beta_0 + \beta_1 X_t + u_t \qquad (5.5)$$

단, Y_t는 임금, X_t는 근무연수를 나타낸다.

근무연수에 따른 임금결정 외에 성별 및 인종별 임금격차가 존재하는 지를 살펴보기 위해 가변수를 (5.6)식과 같이 모형에 포함시킬 수 있다.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{Hi} + \beta_3 D_{Ci} + u_i \qquad (5.6)$$

단, $D_{Hi} = 0, D_{Ci} = 0$, i가 중졸이하
$D_{Hi} = 1, D_{Ci} = 0$, i가 고졸
$D_{Hi} = 0, D_{Ci} = 1$, i가 대졸이상

2.가변수모형의 유형

어느 한 변수의 두 개의 질적 범주를 구분하는 가변수모형의 경우도 평균의 구분, 기울기의 구분 또는 평균과 기울기의 구분 등 다양한 구분이 필요하다.

(1)절편(평균)의 변화를 나타내는 가변수모형

절편(평균)의 변화를 나타내는 가변수모형은 다음의 (5.7)식과 같이 나타낼 수 있으며 (5.2)식과 같은 모형이다.

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + u_t \qquad (5.7)$$

단, $D_t = \begin{cases} 1, & t \text{ 가 } IMF\text{위기 이후} \\ 0, & t \text{ 가 } IMF\text{위기} \end{cases}$

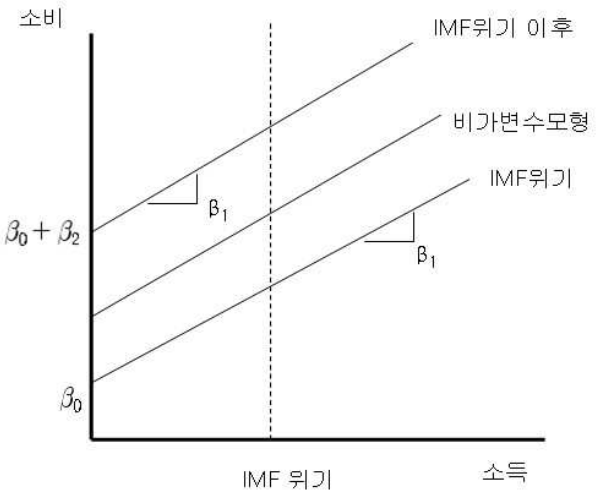
(5.7)식의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 (5.7)식의 추정은 다음의 (5.8) 및 (5.9)식을 개별적으로 추정한 결과와 동일하다.

(IMF위기) $\qquad C_t = \beta_0 + \beta_1 Y_t + u_t \qquad (5.8)$

(IMF위기 이후) $C_t = (\beta_0 + \beta_2) + \beta_1 Y_t + u_t$ (5.9)

가변수에 대한 가설검정에서 귀무가설은 $H_0 : \beta_2 = 0$으로 IMF위기와 IMF위기 이후의 소비수준에는 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, β_2가 통계적으로 유의하면) <그림 5-1>과 같이 IMF위기와 IMF위기 이후의 소비수준에 차이가 있다고 결론을 내린다. 따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 (5.7)식을 추정하면 이와 동일한 결과를 얻을 수 있다.

<그림 5-1> 절편의 변화를 나타내는 가변수



한편, 회귀계수에 대한 해석은 다음과 같다. $\hat{\beta}_1$은 한계소비성향으로 IMF위기의 한계소비성향과 IMF위기 이후의 한계소비성향과 같다. $\hat{\beta}_0$은 IMF위기의 절대소비수준을 나타내며, $\hat{\beta}_2$는 IMF위기 이후 소비수준과 IMF위기 소비수준의 차이를 나타낸다. 따라서 IMF위기 이후 소비수준은 $\hat{\beta}_0 + \hat{\beta}_2$이 된다.

(2)기울기(한계)의 변화를 나타내는 가변수모형

기울기(한계)의 변화를 나타내는 가변수모형은 다음의 (5.10)식과 같이 나타낼 수 있다.

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t Y_t + u_t \qquad (5.10)$$

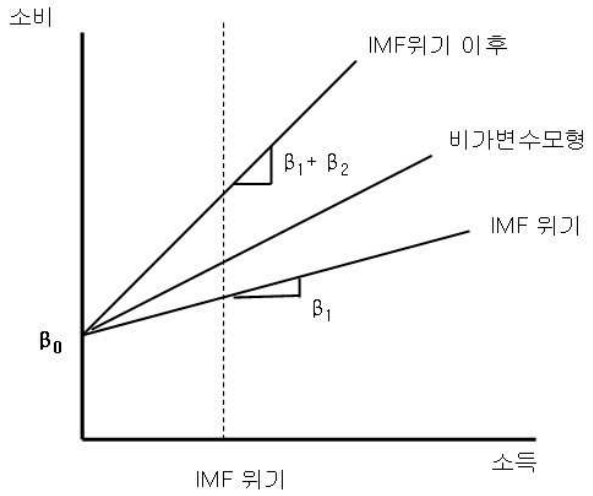
단, $D_t = \begin{cases} 1, & t \text{ 가 } IMF\text{위기 이후} \\ 0, & t \text{ 가 } IMF\text{위기} \end{cases}$

(5.10)식의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 (5.10)식의 추정은 다음의 (5.11) 및 (5.12)식을 개별적으로 추정한 결과와 동일하다.

(IMF위기) $C_t = \beta_0 + \beta_1 Y_t + u_t \qquad (5.11)$

(IMF위기 이후) $C_t = \beta_0 + (\beta_1 + \beta_2) Y_t + u_t \qquad (5.12)$

<그림 5-2> 기울기의 변화를 나타내는 가변수



가변수에 대한 가설검정에서 귀무가설은 $H_0 : \beta_2 = 0$으로 IMF위기와 IMF위기 이후의 한계소비성향에 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, β_2가 통계적으로 유의하면) <그림 5-2>와 같이 IMF위기와 IMF위기 이후의 한계소비성향

에 차이가 있다고 결론을 내린다. 따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 한다. 가변수를 포함한 (5.10)식을 추정하면 이와 동일한 결과를 얻을 수 있다.

한편, 회귀계수에 대한 해석은 다음과 같다. $\widehat{\beta_0}$은 절대소비수준으로 IMF위기의 절대소비수준과 IMF위기 이후의 절대소비수준은 같다. $\widehat{\beta_1}$은 IMF위기의 한계소비성향을 나타내며, $\widehat{\beta_2}$는 IMF위기 이후 한계소비성향과 IMF위기 한계소비성향의 차이를 나타낸다. 따라서 IMF위기 이후 한계소비성향의 값은 $\widehat{\beta_1}+\widehat{\beta_2}$이 된다.

(3)절편과 기울기의 동시변화를 나타내는 가변수모형

절편과 기울기의 동시변화를 나타내는 가변수모형은 다음의 (5.13)식과 같이 나타낼 수 있다.

$$C_t = \alpha_0 + \alpha_1 D_t + \beta_0 Y_t + \beta_1 D_t Y_t + u_t \qquad (5.13)$$

단, $D_t = \begin{cases} 1, & t \text{ 가 } IMF\text{위기 이후} \\ 0, & t \text{ 가 } IMF\text{위기} \end{cases}$

(5.13)식의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 (5.13)식의 추정은 다음의 (5.14) 및 (5.15)식을 개별적으로 추정한 결과와 동일하다.

(IMF위기) $\qquad C_t = \alpha_0 + \beta_0 Y_t + u_t \qquad (5.14)$

(IMF위기 이후) $\qquad C_t = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) Y_t + u_t \qquad (5.15)$

(5.13)식의 가변수모형에서는 3종류의 가설검정이 가능하다.

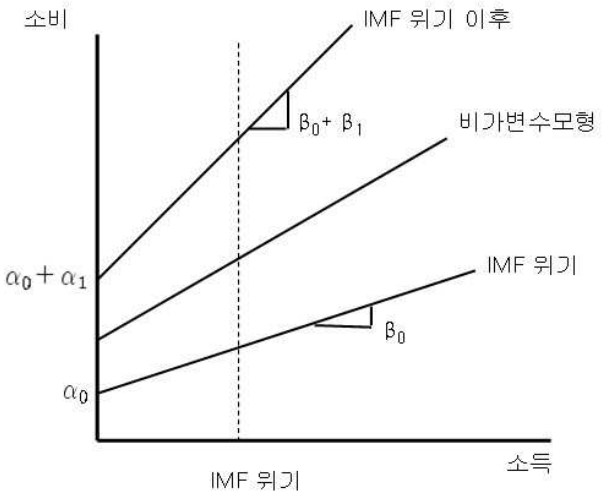
먼저, $H_0 : \alpha_1 = 0$으로 IMF위기와 IMF위기 이후의 소비수준에는 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, α_1이 통계적으로 유의하면) <그림 5-1>과 같이 IMF위기와 IMF위기 이후의 소비수준에 차이가 있다고 결론을 내린다. 따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 하는데 가변수를 포함한 (5.7)식을 추정하면 이와 동일한 결과를 얻을 수 있다.

다음으로 $H_0 : \beta_1 = 0$으로 IMF위기와 IMF위기 이후의 한계소비성향에 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, β_1이 통계적으로 유의하면) <그림 5-2>

와 같이 IMF위기와 IMF위기 이후의 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 한다. 가변수를 포함한 (5.10)식을 추정하면 이와 동일한 결과를 얻을 수 있다.

마지막으로, $H_0 : \alpha_1 = \beta_1 = 0$으로 IMF위기와 IMF위기 이후의 소비수준 및 한계소비성향에 차이가 없다는 귀무가설을 검정하고 결론을 내린다. 귀무가설을 기각하면(즉, α_1과 β_1이 동시에 통계적으로 유의하면) <그림 5-3>과 같이 IMF위기와 IMF위기 이후의 소비수준 및 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 한다. 가변수를 포함한 (5.13)식을 추정하면 이와 동일한 결과를 얻을 수 있다.

<그림 5-3> 절편과 기울기의 변화를 나타내는 가변수



한편, 회귀계수에 대한 해석은 다음과 같다. $\widehat{\alpha_0}$은 IMF위기의 절대소비수준을 나타내며, $\widehat{\alpha_1}$은 IMF위기 이후의 절대소비수준과 IMF위기의 절대소비수준의 차이를 나타낸다. 또한 $\widehat{\beta_0}$은 IMF위기의 한계소비성향을 나타내며, $\widehat{\beta_1}$은 IMF위기 이후의 한계소비성향과 IMF위기의 한계소비성향의 차이를 나타낸다.

3.가변수의 응용

시계열자료에서 나타나는 중요한 특징 중의 하나는 많은 시계열들이 계절성을 나타낼 수 있다는 것인데 시계열의 계절성 여부를 판단하고 처리하는 방법으로 가변수를 활용할 수 있다.

일반적으로 경제시계열은 기술향상 및 인구증가 등의 요인에 의해 체계적으로 변하는 부분인 추세요인(secular trend, T_t), 추세를 중심으로 순환변동의 현상을 보이는 순환요인(cyclical factor, C_t), 기후 및 사회적 관습 등에 의해 변하는 계절요인(seasonal factor, S_t), 기상재해 및 파업 등에 의해 변하는 불규칙요인(irregular factor, I_t) 등 네 가지 요인으로 구성되어 있다고 본다.

한편, 이러한 요인들은 (5.16)식과 같이 곱의 형태로 나타낸 승법모형(multiplicative model)과 (5.17)식과 같이 합의 형태로 나타낸 가법모형(additive model)으로 나타낼 수 있는데 가법모형은 계절성분의 진폭이 수준에 관계없이 일정할 때 적합한 모형이고, 승법모형은 계절성분의 진폭이 시계열 수준에 따라 달라질 때 적합한 모형으로 알려져 있으나 실제로는 승법모형 또는 로그 가법모형이 많이 사용된다.

$$Y_t = T_t \times C_t \times S_t \times I_t \tag{5.16}$$

$$\ln Y_t = \ln T_t + \ln C_t + \ln S_t + \ln I_t \tag{5.17}$$

경제시계열에 계절적 요인이 있을 경우 계절적 요인을 설명할 수 있는 계절적 변수가 필요한데 이러한 계절성을 설명하는 변수로 가변수를 이용할 수 있다. 예를 들어 계절적 요인이 있는 분기별 자료의 경우 가변수를 이용하여 계절변동을 고려한 회귀식(모형의 절편이 분기별로 다르고 기울기는 같은 모형이라고 가정할 경우)은 다음의 (5.18)식과 같게 된다. 분기별 자료이므로 1분기에서 4분기까지 범주가 4개 이므로 상수항을 포함한 모형에서 3개의 가변수를 사용한다는 것은 앞에서 설명한 바와 같다.

$$Y_t = \alpha_1 + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + \beta X_t + u_t \tag{5.18}$$

단, $D_{2t} = \begin{cases} 1, & t\text{가 } 2\text{분기이면} \\ 0, & t\text{가 다른 분기이면} \end{cases}$

$D_{3t} = \begin{cases} 1, & t\text{가 } 3\text{분기이면} \\ 0, & t\text{가 다른 분기이면} \end{cases}$

$$D_{4t} = \begin{cases} 1, & t\text{가 } 3\text{분기이면} \\ 0, & t\text{가 다른 분기이면} \end{cases}$$

각 분기별로 계절성이 뚜렷하여 계절성을 나타내는 가변수가 모두 통계적으로 유의할 경우 각 분기의 회귀식은 (5.19)-(5.22)식과 같다.

(1분기) $Y_t = \alpha_1 + \beta X_t + u_t$ (5.19)

(2분기) $Y_t = \alpha_1 + \alpha_2 + \beta X_t + u_t$ (5.20)

(3분기) $Y_t = \alpha_1 + \alpha_3 + \beta X_t + u_t$ (5.21)

(4분기) $Y_t = \alpha_1 + \alpha_4 + \beta X_t + u_t$ (5.22)

가변수가 모두 2,3,4분기의 절편은 1분기에 비해 각각 $\alpha 2, \alpha 3, \alpha 4$만큼 차이가 나서 독립변수 X와 무관하게 Y의 값에서 발생하는 계절성을 설명할 수 있다. 각 분기별 자료로 분기별로 Y와 X의 관계를 별도로 추정할 수도 있지만 가변수를 포함한 (5.18)식을 추정하면 이와 동일한 결과를 얻을 수 있다.

4.가변수를 이용한 연구 실례: 질적 가변수 및 양적 가변수의 이용

본 연구는 당초 국립대 및 사립대 평균발전기금에 차이가 있는 지를 살펴보기 위해 시작되었다. 두 집단의 평균차이의 검증을 통해 이를 확인 할 수 있고 가변수모형을 이용해 가변수의 통계적 유의성 검증을 통해 확인할 수 있다. 또한 이것에서 확장하여 대학유형 외에 기타 특성변수를 통제했을 경우에도 국립대 및 사립대 평균발전기금에 차이가 있는 지를 살펴보기로 하였다.

(1) 질적 가변수를 사용한 회귀모형

① 하나의 질적 가변수 이용

다음의 모형과 같이 대학의 발전기금을 설명함에 있어 국립대 및 사립대학 등 대학유형을 고려해 보자

$$y_i = \beta_0(1) + \beta_1 D_{1i} + u_i (i = 1,2,...N) \quad \text{(C1)}$$

단, y(money)는 발전기금, D_1(nadum)은 국립대를 나타내는 가변수, (D_2(pdum)를 사립대를 나타내는 가변수라고 하면 $D_1 + D_2 = 1$)

(C1)의 모형에서 β_0는 사립대학의 평균발전기금을 나타내고 β_1은 국립대와 사립대의 평균발전기금의 차이를 나타내는데 그 이유를 살펴보자

$D_1 + D_2 = 1$을 (C1)식의 상수항 1에 대입하면 (C1-1)식이 되고((C1-1)식은 (C1)식과 동일함) 이를 전개하면 (C1-2)식이 된다.

$$y = \beta_0(D_1 + D_2) + \beta_1 D_1 + u \quad \text{(C1-1)}$$

$$y = \beta_0 D_2 + (\beta_0 + \beta_1) D_1 + u \quad \text{(C1-2)}$$

따라서 사립대의 경우 $D_1 = 0, D_2 = 1$이므로 평균발전기금은 β_0가 되고 국립대의 경우 $D_1 = 1, D_2 = 0$이므로 평균발전기금은 $\beta_0 + \beta_1$이 된다. 따라서 β_1은 국립대와 사립대의 평균발전기금의 차이를 나타낸다.

추정은 국립대 및 사립대의 발전기금을 종속변수로 두고 (C1-1)식 또는 (C1-2)식으로 추정하면 된다.

(C1-1)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	655123.1	2023235	0.32	0.746	-3339117	4649364
_cons(상수항)	3737746	744193.4	5.02	0.000	2268570	5206921

(C1-2)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	4392869	1881397	2.33	0.021	678641.9	8107096
pdum	3737746	744193.4	5.02	0.000	2268570	5206921

또는 국립대 발전기금을 종속변수로 두고 (C1-3)식을 추정하고 사립대 발전기금을 종속변수로 두고 (C1-4)식을 추정하면 된다.

$$y_i = \beta_0 + \beta_1 D_{1i} + u_i (i = \text{국립대}) \quad \text{(C1-3)}$$

$$y_i = \beta_0 + \beta_1 D_{2i} + u_i (i = \text{사립대}) \quad \text{(C1-4)}$$

(C1-3)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	(dropped)					
_cons	4392869	2271487	1.93	0.066	-317906	9103644

(C1-4)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
pdum	(dropped)					
_cons	3737746	718073.7	5.21	0.000	2318584	5156907

② 두 개의 질적 가변수 이용

다음의 모형과 같이 대학의 발전기금을 설명함에 있어 국립대 및 사립대학 등 대학유형 외에 서울 및 지방 등 지역을 고려해 보자

$$y_i = \beta_0(1) + \beta_1 D_{1i} + \beta_2 L_{1i} + u_i (i = 1,2,...N) \quad \text{(C2)}$$

$$y_i = \beta_0(1) + \beta_1 D_{1i} + \beta_2 L_{1i} + \beta_3 D_{1i} L_{1i} + u_i (i = 1,2,...N) \quad \text{(C2-1)}$$

단, y는 발전기금, D_1은 국립대를 나타내는 가변수, L_1(sdum)은 서울소재 대학을 나타내는 가변수, (L_2(ldum)는 지방소재 대학을 나타내는 가변수라고 하면 $L_1 + L_2 = 1$)

(C2)의 모형에서 β_0는 지역을 통제했을 경우 지방소재 사립대학의 평균발전기금을 나타내고 β_1은 국립대와 사립대의 평균발전기금의 차이를 나타내는데 그 이유를 살펴보자

$D_1 + D_2 = 1$ 및 $L_1 + L_2 = 1$을 (C2)식의 1에 대입하면 (C2-2)식이 되고((C2-2)식은 (C2)식과 동일함) 이를 전개하면 (C2-3)식이 된다.

$$y = \beta_0(D_1 + D_2)(L_1 + L_2) + \beta_1 D_1 + \beta_2 L_1 + u \quad \text{(C2-2)}$$

$$y = \beta_0 D_1 L_1 + \beta_0 D_1 L_2 + \beta_0 D_2 L_1 + \beta_0 D_2 L_2 + \beta_1 D_1 + \beta_2 L_1 + u \quad \text{(C2-3)}$$

따라서 지방소재 사립대의 경우 평균발전기금은 β_0가 되고 지방소재 국립대의 경우 평균발전기금은 $\beta_0 + \beta_1$이 되며, 서울소재 사립대의 경우 평균발전기금은 $\beta_0 + \beta_2$가 되고 서울소재 국립대의 경우 평균발전기금은 $\beta_0 + \beta_1 + \beta_2$이 된다. 이에 따라 β_1은 지역과 관계없이 국립대와 사립대의 평균발전기금의 차이를 나타낸다. 따라서 국립대 및 사립대의 평균발전기금의 차이 여부는 β_1의 부호 및 통계적 유의성에 근거하여 검증하면 된다.

한편, 두 개의 질적 변수간의 상호작용효과를 고려한 (C2-1)의 모형에서는 지방소재 사립대의 경우 평균발전기금은 β_0가 되고 지방소재 국립대의 경우 평균발전기금은 $\beta_0 + \beta_1$이 되며, 서울소재 사립대의 경우 평균발전기금은 $\beta_0 + \beta_2$가 되고 서울소재 국립대의 경우 평균발전기금은 $\beta_0 + \beta_1 + \beta_2 + \beta_3$가 된다. 이에 따라 β_1은 지방소재 국립대와 사립대의 평균발전기금의 차이를 나타내며, $\beta_1 + \beta_3$는 서울소재 국립대 및 사립대의 평균발전기금의 차이를 나타낸다. 따라서 지방소재 국립대 및 사립대의 평균발전기금의 차이 여부는 β_1의 부호 및 통계적 유의성에 근거하여 검증하면 되고, 서울소재 국립대 및 사립대의 평균발전기금의 차이 여부는 $\beta_1 + \beta_3$의 부호 및 통계적 유의성에 근거하여 검증하면 된다.

추정은 국립대 및 사립대의 발전기금을 종속변수로 두고 (C2-1)식으로 추정하면 된다.((C2)식으로 추정하는 경우는 두 개 질적 가변수간의 상호작용효과(interaction effect)가 없다고 가정하는 경우임)

(C2-1)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	68148.1	1994593	0.03	0.973	-3869892	4006188
sdum	6102968	1624287	3.76	0.000	2896045	9309892
cross*	1.74e+07	6416072	2.71	0.008	4690131	3.00e+07
_cons	2284658	792570.9	2.88	0.004	719839.5	3849476

* 두 개 질적 가변수 간의 상호작용효과를 나타냄

또는 서울지역 국립대 발전기금을 종속변수로 두고 (C2-4)식을 추정하고, 서울지역 사립대 발전기금을 종속변수로 두고 (C2-5)식을 추정하고, 지방 국립대 발전기금을 종속변수로 두고 (C2-6)식을 추정하고, 지방 사립대 발전기금을 종속변수로 두고 (C2-7)식을 추정하면 된다.

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 L_{1i} + u_i (i = \text{서울지역 국립대}) \quad \text{(C2-4)}$$

$$y_i = \beta_0 + \beta_1 D_{2i} + \beta_2 L_{1i} + u_i (i = \text{서울지역 사립대}) \quad \text{(C2-5)}$$

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 L_{2i} + u_i (i = \text{지방 국립대}) \quad \text{(C2-6)}$$

$$y_i = \beta_0 + \beta_1 D_{2i} + \beta_2 L_{2i} + u_i (i = \text{지방 사립대}) \quad \text{(C2-7)}$$

(C2-4)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	(dropped)					
sdum	(dropped)					
_cons	2.58e+07	2.58e+07	1.00	0.500	-3.02e+08	3.54e+08

(C-5)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
pdum	(dropped)					
sdum	(dropped)					
_cons	8387626	2470566	3.40	0.002	3366832	1.34e+07

(C-6)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	(dropped)					
ldum	(dropped)					
_cons	2352806	809428.2	2.91	0.009	664368.4	4041244

(C2-7)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	(dropped)					
sdum	(dropped)					
_cons	2284658	475410.3	4.81	0.000	1342601	3226715

위의 예를 통해 볼 때 질적 가변수는 그 수에 관계없이 회귀모형의 절편의 차이를 가져오는 것으로 해석된다.

(2) 양적 변수를 사용한 회귀모형

① 양적 변수의 합이 1이 되는 경우의 회귀모형

다음의 모형과 같이 대학의 발전기금을 설명함에 있어 대학의 계열별 구성 비율을 고려해 보자

$$y_i = \beta_0(1) + \beta_1 r_{1i} + \beta_2 r_{2i} + \beta_3 r_{3i} + u_i (i = 1,2,...N) \quad \text{(C3)}$$

단, y는 발전기금, r_1(natural)은 자연과학비율, r_2(engin)는 공학비율, r_3(art)는 예체능비율, (r_4(human)가 인문사회비율을 나타낸다고 하면 $r_1 + r_2 + r_3 + r_4 = 1$)

(C3)의 모형에서 β_0는 인문사회계열비율의 변화에 대한 발전기금의 변화를 나타내고 β_1은 자연과학비율의 변화에 의한 발전기금의 변화의 차이를 나타내는데 그 이유를 살펴보자

$r_1 + r_2 + r_3 + r_4 = 1$을 (C3)식의 1에 대입하면 (C3-1)식이 되고((C3-1)식은 (C3)식과 동일함) 이를 전개하면 (C3-2)식이 된다.

$$y = \beta_0(r_1 + r_2 + r_3 + r_4) + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + u \quad \text{(C3-1)}$$

$$y = \beta_0 r_4 + (\beta_0 + \beta_1) r_1 + (\beta_0 + \beta_2) r_2 + (\beta_0 + \beta_3) r_3 + u \quad \text{(C3-2)}$$

따라서 인문사회계열의 비율이 1% 증가할 경우 발전기금은 β_0의 크기만큼 증가하고 자연과학비율이 1% 증가할 경우 발전기금은 $\beta_0+\beta_1$의 크기만큼 증가한다.

(C3)식에서 다른 모든 조건이 불변이고(ceteris paribus) 자연과학비율이 1% 증가한다는 가정은 4개 계열비율의 합이 1이므로 자동적으로 인문사회비율은 1% 감소함을 의미한다. 즉, 자연과학비율이 1% 증가하면 발전기금은 $\beta_0+\beta_1$만큼 증가하고, 인문사회비율이 1% 감소하면 발전기금은 β_0만큼 감소하므로 자연과학비율 변화에 의한 순효과(net effect)는 β_1이 되고 (C3)식에서와 동일한 크기가 된다. 따라서 β_1은 자연과학비율과 인문사회비율의 변화에 의한 발전기금의 변화의 차이를 나타낸다. 나머지 계열의 비율도 유사한 방법으로 해석할 수 있다.

추정은 국립대 및 사립대의 발전기금을 종속변수로 두고 (C3)식 또는 (C3-2)식으로 추정하면 된다.

(C3)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
natural	5022867	4518033	1.11	0.268	-3897345	1394308
engin	8886835	3588133	2.48	0.014	1802576	1597109
art	-2541798	4870493	-0.52	0.602	-1215789	7074298
_cons	1639119	1477551	1.11	0.269	-1278095	4556333

(C3-2)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
human	1639119	1477551	1.11	0.269	-1278095	4556333
natural	6661987	3837261	1.74	0.084	-914140.2	1.42e+07
engin	1.05e+07	2873669	3.66	0.000	4852305	1.62e+07
art	-902678.8	4169953	-0.22	0.829	-9135658	7330300

여기서 유의할 점은 (C3)식에서 β_0가 intercept이지만 양적 변수의 합이 1이 되는 변수를 설명변수로 사용할 경우 β_0는 slope의 의미를 가진다는 것이다.

② 양적 변수 및 양적 변수의 합이 1이 되는 경우를 동시에 사용한 회귀모형

다음의 모형과 같이 대학의 발전기금을 설명함에 있어 학생 수 및 대학의 계열별 구성 비율을 고려해 보자

$$y_i = \beta_0(1) + \beta_1 S_i + \beta_2 r_{1i} + \beta_3 r_{2i} + \beta_4 r_{3i} + u_i (i = 1,2,...N) \quad \text{(C4)}$$

단, y는 발전기금

r_1는 자연과학비율

r_2는 공학비율

r_3는 예체능비율(r_4가 인문사회비율을 나타낸다고 하면 $r_1 + r_2 + r_3 + r_4 = 1$)

S(student)는 학생 수

(C4)의 모형에서 β_0는 학생 수가 일정한 상태($\bar{S}$)에서 인문사회계열비율의 변화에 대한 발전기금의 변화를 나타내고 β_1은 자연과학비율의 변화에 의한 발전기금의 변화(또는 자연과학비율과 인문사회비율의 변화에 의한 발전기금의 변화의 차이)를 나타내는데 그 이유를 살펴보자

$r_1 + r_2 + r_3 + r_4 = 1$을 (C4)식의 1에 대입하면 (C4-1)식이 되고((C4-1)식은 (C4)식과 동일함) 이를 전개하면 (C4-2)식이 된다.

$$y = \beta_0(r_1 + r_2 + r_3 + r_4) + \beta_1 S + \beta_2 r_1 + \beta_3 r_2 + \beta_4 r_3 + u \quad \text{(C4-1)}$$

$$y = \beta_0 r_4 + \beta_1 S + (\beta_0 + \beta_2) r_1 + (\beta_0 + \beta_3) r_2 + (\beta_0 + \beta_4) r_3 + u \quad \text{(C4-2)}$$

따라서 학생 수가 일정한 상태($\bar{S}$)에서 인문사회계열의 비율이 1% 증가할 경우 발전기금은 β_0의 크기만큼 증가하고 자연과학비율이 1%가 증가할 경우 발전기금은 $\beta_0 + \beta_2$의 크기만큼 증가한다.

(C4)식에서 다른 모든 조건이 불변이고 자연과학비율이 1% 증가한다는 가정은 4개 계열비율의 합이 1이므로 자동적으로 인문사회비율은 1% 감소함을 의미한다. 즉, 자연과학비율이 1% 증가하면 발전기금은 $\beta_0 + \beta_2$만큼 증가하고, 인문사회비율이 1% 감소하면 발전기금은 β_0만큼 감소하므로 자연과학비율 변화에 의한 순효과(net effect)는 β_2가 되고 (C4)식에서와 동일한 크기가 된다. 따라서 β_2는 자연과학비율과 인문사회비율의 변화에 의한 발전기금의 변화의 차이를 나타낸다. 나머지 계열의 비율도 유사한 방법으로 해석할 수 있다.

추정은 국립대 및 사립대의 발전기금을 종속변수로 두고 (C4)식 또는 (C4-2)

식으로 추정하면 된다.

(C4)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	629.5751	60.53491	10.40	0.000	510.0522	749.0979
natural	-1338293	3574733	-0.37	0.709	-8396409	5719822
engin	1580271	2883996	0.55	0.584	-4114022	7274565
art	-3277634	3797437	-0.86	0.389	-1.08e+07	4220198
_cons	-1823170	1198964	-1.52	0.130	-4190460	544120.6

(C4-2)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	629.5751	60.53491	10.40	0.000	510.0522	749.0979
human	-1823170	1198964	-1.52	0.130	-4190460	544120.6
natural	-3161463	3136908	-1.01	0.315	-9355117	3032191
engin	-242898.2	2467886	-0.10	0.922	-5115606	4629809
art	-5100804	3275640	-1.56	0.121	-1.16e+07	1366770

한편, 학생 수의 변화에 의한 발전기금의 변화의 크기는 β_1이므로 위의 예를 통해 볼 때 양적 변수는 회귀모형의 기울기의 차이를 가져오는 것으로 해석된다.

(3) 질적 가변수 및 양적 변수를 동시에 사용한 회귀모형(공분산분석)

① 질적 가변수 및 양적 변수를 동시에 사용한 회귀모형

다음의 모형과 같이 대학의 발전기금을 설명함에 있어 학생 수 및 대학의 계열별 구성 비율을 고려해 보자

$$y_i = \beta_0(1) + \beta_1 D_{1i} + \beta_2 S_i + u_i (i = 1,2,...N) \quad \text{(C5)}$$

$$y_i = \beta_0(1) + \beta_1 D_{1i} + \beta_2 S_i + \beta_3 D_{1i} S_i + u_i (i = 1,2,...N) \quad \text{(C5-1)}$$

단, y는 발전기금, r_1는 자연과학비율, r_2는 공학비율, r_3는 예체능비율(r_4가 인

문사회비율을 나타낸다고 하면 $r_1+r_2+r_3+r_4=1$), S는 학생 수

(C5)의 모형에서 β_0는 학생 수가 0일 때 사립대의 발전기금을 나타내고 β_1은 학생 수가 동일한 경우($S=\overline{S}$) 국립대와 사립대의 발전기금의 차이를 나타내는데 그 이유를 살펴보자.

(C5)식에서 사립대의 경우 $D_1=0$을 대입하면 $y_i=\beta_0+\beta_2 S_i+u_i$가 되고 국립대의 경우 $D_1=1$을 대입하면 $y_i=(\beta_0+\beta_1)+\beta_2 S_i+u_i$가 된다. 국립대학 회귀식에서 사립대학 회귀식을 빼주면 β_1이 남고 S값의 수준에 관계없이 β_1은 평균발전기금의 차이를 나타낸다.

$D_1+D_2=1$을 (C5)식의 1에 대입하면 (C5-2)식이 되고((C5-2)식은 (C5)식과 동일함) 이를 전개하면 (C5-3)식이 된다.

$$y=\beta_0(D_1+D_2)+\beta_1 D_1+\beta_2 S+u \qquad \text{(C5-2)}$$

$$y=\beta_0 D_2+(\beta_0+\beta_1)D_1+\beta_2 S+u \qquad \text{(C5-3)}$$

예를 들어 학생 수가 국립 및 사립 모두 10000명일 경우 사립대의 발전기금은 $\beta_0+\beta_2$*10000이 되고, 국립대의 발전기금은 $\beta_0+\beta_1+\beta_2$*10000이 된다. 따라서 β_1은 학생 수가 동일한 경우 국립대와 사립대의 발전기금의 차이를 나타낸다. 따라서 국립대 및 사립대의 평균발전기금의 차이 여부는 β_1의 부호 및 통계적 유의성에 근거하여 검증하면 된다.

한편, 가변수 및 양적 변수간의 상호작용효과를 고려한 (C5-1)의 모형에서는 학생 수가 국립 및 사립 모두 10000명일 경우 사립대의 발전기금은 $\beta_0+\beta_2$*10000이 되고, 국립대의 발전기금은 $\beta_0+\beta_1+(\beta_2+\beta_3)$*10000이 된다. 따라서 β_1은 학생 수가 동일한 경우 국립대와 사립대의 발전기금의 수준의 차이를 나타내고 β_3는 학생 수가 동일한 경우 국립대와 사립대의 발전기금 증가의 차이를 나타낸다. 따라서 국립대 및 사립대의 발전기금의 수준차이 여부는 β_1의 부호 및 통계적 유의성에 근거하여 검증하고, 국립대 및 사립대의 발전기금의 증가차이 여부는 β_3의 부호 및 통계적 유의성에 근거하여 검증하며, 국립대 및 사립대의 발전기금의 차이 여부(수준 및 증가)는 β_1 및 β_3 부호 및 통계적 유의성(결합검증)에 근거하여 검증하면 된다.

추정은 국립대 및 사립대의 발전기금을 종속변수로 두고 (C5-1)식으로 추정하면 된다. ((C5)식으로 추정하는 경우는 학생 수의 변화에 의한 발전기금의 변화가 국립대 및 사립대 모두 동일한 것으로 가정하는 경우임)

(C5-1)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	-4907099	2857169	-1.72	0.088	-1.05e+07	733975.4
student	673.5438	64.1622	10.50	0.000	546.8647	800.2229
nstudent	46.41463	164.4621	0.28	0.778	-278.2924	371.1217
_cons	-1922025	775932.1	-2.48	0.014	-3453992	-390057

또는 국립대 발전기금을 종속변수로 두고 (C5-4)식을 추정하고 사립대 발전기금을 종속변수로 두고 (C5-5)식을 추정하면 된다.

$$y_i = \beta_0 + \beta_1 S_i + u_i (i = \text{국립대}) \quad \text{(C5-4)}$$

$$y_i = \beta_0 + \beta_1 S_i + u_i (i = \text{사립대}) \quad \text{(C5-5)}$$

(C5-4)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	673.5438	61.11802	11.02	0.000	552.7465	794.3411
_cons	-1922025	739118.1	-2.60	0.010	-3382862	-461187.7

(C5-5)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	719.9584	193.9041	3.71	0.001	316.7128	1123.204
_cons	-6829123	3521073	-1.94	0.066	-1.42e+07	493348.6

(C5-1)의 추정식을 이용하면 (C5-4) 및 (C5-5)의 추정식을 바로 계산할 수 있다.

② 질적 가변수, 양적 변수 및 양적 변수의 합이 1이 되는 경우를 동시에 사용한 회귀모형

다음의 모형과 같이 발전기금을 설명함에 있어 학생 수 및 계열별 학생비율을 통제변수로 사용한다고 하자

$$y_i = \beta_0(1) + \delta_1 D_{1i} + \gamma_1 S_i + \beta_1 r_{1i} + \beta_2 r_{2i} + \beta_3 r_{3i} + u_i (i = 1,2,...N) \quad \text{(C6)}$$

$$y_i = \beta_0(1) + \delta_1 D_{1i} + \gamma_1 S_i + \beta_1 r_{1i} + \beta_2 r_{2i} + \beta_3 r_{3i} + \gamma_2 D_{1i} S_i + \delta_2 D_{1i} r_{1i} + \delta_3 D_{1i} r_{2i} + \delta_4 D_{1i} r_{3i} + u_i \text{(C6-1)}$$
$(i = 1,2,...N)$

단, y는 발전기금, D_1은 국립대를 나타내는 가변수, S는 학생 수, r_1는 자연과학비율, r_2는 공학비율, r_3는 예체능비율, $D_1 S$(nstudent)는 D_1 및 S의 cross-term, $D_1 r_1$(nnatural)는 D_1 및 r_1의 cross-term, $D_1 r_2$(nengin)는 D_1 및 r_2의 cross-term, $D_1 r_3$(nart)는 D_1 및 r_3의 cross-term이다.

(C6) 또는 (C6-1)의 회귀모형을 추정하면 추정회귀계수를 얻을 수 있는데 각각의 회귀계수가 어떤 의미를 갖는 지를 살펴보자.

첫째, δ_1은 설명변수로 통제했을 경우 국립대 및 사립대의 평균발전기금의 차이를 나타내며 부호가 양이고 통계적으로 유의하면 국립대가 사립대보다 발전기금이 많다는 것을 의미한다.

둘째, 나머지 회귀계수의 의미를 살펴보기 위해 $D_1 + D_2 = 1$을 (C6-1)식의 1에 대입하고 아래첨자 i를 제거하면 (C6-2)식이 된다((C6)식은 설명변수 간 상호작용효과를 고려하지 않은 경우임)

$$y = \beta_0(D_1 + D_2) + \delta_1 D_1 + \gamma_1 S + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + \gamma_2 D_1 S + \delta_2 D_1 r_1 + \delta_3 D_1 r_2 + \delta_4 D_1 r_3 + u \quad \text{(C6-2)}$$

먼저 사립대학의 경우 $D_1 = 0, D_2 = 1$이 되므로 이 값을 (C6-2)식에 대입하고 정리하면 다음과 같게 된다.

$$y = \beta_0 + \gamma_1 S + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + u \quad \text{(C6-3)}$$

또한 $r_1 + r_2 + r_3 + r_4 = 1$이므로 이를 대입하면 (C6-4)식과 같다

$$y = \beta_0(r_1 + r_2 + r_3 + r_4) + \gamma_1 S + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + u \quad \text{(C6-4)}$$

(C6-4)식을 정리하면 (C6-5)식과 같게 된다.

$$y = \beta_0 r_4 + \gamma_1 S + (\beta_0 + \beta_1) r_1 + (\beta_0 + \beta_2) r_2 + (\beta_0 + \beta_3) r_3 + u \quad \text{(C6-5)}$$

(C6-5)식의 $\beta_0 r_4$항에서 β_0의 의미는 사립대학의 인문사회비율(r_4)이 1% 증가할 경우 발전기금이 β_0만큼 증가한다는 것이다. 또한 사립대학의 자연과학비율(r_1)이 1% 증가할 경우 발전기금은 $\beta_0 + \beta_1$만큼 증가하고 나머지 계열의 비율도 유사한 방법으로 해석하면 된다. 따라서 β_1는 사립대학 자연과학비율과 인문사회비율의 변화에 의한 발전기금의 변화의 차이를 나타낸다. 나머지 계열의 비율도 유사한 방법으로 해석할 수 있다.

한편, 국립대학의 경우 $D_1 = 1, D_2 = 0$이 되므로 이 값을 (C6-2)식에 대입하면 다음과 같게 된다.

$$y = (\beta_0 + \delta_1) + \gamma_1 S + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + \gamma_2 S + \delta_2 r_1 + \delta_3 r_2 + \delta_4 r_3 + u \quad \text{(C-6)}$$

또한 $r_1 + r_2 + r_3 + r_4 = 1$이므로 이를 대입하면 (C6-7)식과 같다

$$y = (\beta_0 + \delta_1)(r_1 + r_2 + r_3 + r_4) + \gamma_1 S + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + \gamma_2 S + \delta_2 r_1 + \delta_3 r_2 + \delta_4 r_3 + u \quad \text{(C6-7)}$$

(C6-7)식을 정리하면 (c6-8)식과 같게 된다.

$$y = (\beta_0 + \delta_1) r_4 + (\gamma_1 + \gamma_2) S + (\beta_1 + \delta_2 + \beta_0 + \delta_1) r_1 + (\beta_2 + \delta_3 + \beta_0 + \delta_1) r_2 + (\beta_3 + \delta_4 + \beta_0 + \delta_1) r_3 + u \quad \text{(C6-8)}$$

(C6-5)식 및 (C6-8)식을 비교해 보면 국립대와 사립대의 차이는 δ_1임을 알 수 있으며 (C6-5)식 및 (C6-8)식을 추정할 때 주의할 점은 상수항이 없는 모형으로 추정해야 한다는 것이다.

셋째, (C6)식에서 보면 다른 모든 조건이 불변이고 자연과학비율이 1% 증가할 경우(즉, 학생 수, 공학비율 및 예체능비율이 변하지 않고 자연과학비율이 1% 증가할 경우) 발전기금은 β_1만큼 증가하는 것으로 해석된다.

이를 (C6-5)식을 이용하여 자세하게 살펴보자. 다른 모든 조건이 불변이고 자

연과학비율이 1% 증가한다는 가정은 4개 계열비율의 합이 1이므로 자동적으로 인문사회비율은 1% 감소함을 의미한다. 자연과학비율이 1% 증가하면 발전기금은 $\beta_0+\beta_1$만큼 증가하고, 인문사회비율이 1% 감소하면 발전기금은 β_0만큼 감소하므로 순효과(net effect)는 β_1이 되고 (C6)식에서와 동일한 크기가 된다.

한편, (C6)식의 경우 사립대 및 국립대 모두 기울기의 변화는 동일한 것으로 가정하였으나 이 가정을 완화한 (C6-1)식의 경우 자연과학비율이 1% 증가할 경우(즉, 학생 수, 공학비율 및 예체능비율이 변하지 않고 자연과학비율이 1% 증가할 경우) 사립대의 발전기금은 β_1만큼 증가하는 반면에 국립대의 발전기금은 $\beta_1+\delta_2$만큼 증가하는 것으로 나타난다.

추정은 국립대 및 사립대의 발전기금을 종속변수로 두고 (C6-1)식으로 추정하면 된다. (C6)식으로 추정하는 경우는 학생 수 및 계열별 비율의 변화에 의한 발전기금의 변화가 국립대 및 사립대 모두 동일한 것으로 가정하는 경우임)

(C6-1)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
nadum	-1783626	9340136	-0.19	0.849	-2.02e+07	1.67e+07
student	648.2384	66.44527	9.76	0.000	517.0155	779.4612
natural	1172509	3602125	0.33	0.745	-5941334	8286351
engin	4597657	3258128	1.41	0.160	-1836825	1.10e+07
art	-5185945	4195060	-1.24	0.218	-1.35e+07	3098886
nstudent	205.17	190.9762	1.07	0.284	-171.9892	582.3291
nnatural	-2.49e+07	2.64e+07	-0.94	0.347	-7.71e+07	2.73e+07
nengin	-5258640	1.08e+07	-0.48	0.629	-2.67e+07	1.62e+07
nart	6912918	1.25e+07	0.56	0.580	-1.77e+07	3.15e+07
_cons	-2064334	1212820	-1.70	0.091	-4459535	330866.8

또는 국립대 발전기금을 종속변수로 두고 (C6-9)식을 추정하고 사립대 발전기금을 종속변수로 두고 (C6-10)식을 추정하면 된다.

$$y_i=\beta_0+\beta_1 S_i+\beta_2 r_{1i}+\beta_3 r_{2i}+\beta_4 r_{3i}+u_i(i=\text{국립대}) \quad \text{(C6-9)}$$

$$y_i=\beta_0+\beta_1 S_i+\beta_2 r_{1i}+\beta_3 r_{2i}+\beta_4 r_{3i}+u_i(i=\text{사립대}) \quad \text{(C6-10)}$$

(C6-9)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	853.4084	240.8241	3.54	0.002	347.4557	1359.361
natural	-2.37e+07	3.52e+07	-0.67	0.509	-9.77e+07	5.03e+07
engin	-660982.3	1.39e+07	-0.05	0.963	-2.99e+07	2.86e+07
art	1726973	1.58e+07	0.11	0.914	-3.14e+07	3.49e+07
_cons	-3847961	1.25e+07	-0.31	0.761	-3.00e+07	2.23e+07

(C6-10)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	648.2384	62.94544	10.30	0.000	523.8071	772.6696
natural	1172509	3412393	0.34	0.732	-5573148	7918165
engin	4597657	3086515	1.49	0.139	-1503800	1.07e+07
art	-5185945	3974097	-1.30	0.194	-1.30e+07	2670093
_cons	-2064334	1148938	-1.80	0.075	-4335569	206900

(C6-1)의 추정식을 이용하면 (C6-9) 및 (C6-10)의 추정식을 바로 계산할 수 있다.

한편, 사립대 발전기금을 종속변수로 두고 (C6-5)식을 추정하고, 국립대 발전기금을 종속변수로 두고 (C6-8)식을 추정하면 다음과 같다

(C6-5)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	648.2384	62.94545	10.30	0.000	523.8071	772.6696
human	-2064334	1148938	-1.80	0.075	-4335568	206900
natural	-891825.6	2997629	-0.30	0.767	-6817571	5033920
engin	2533323	2691977	0.94	0.348	-2788207	7854853
art	-7250280	3476717	-2.09	0.039	-1.41e+07	-377467.5

(C6-8)식 추정결과

money	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
student	853.4084	240.8241	3.54	0.002	347.4557	1359.361
human	-3847961	1.25e+07	-0.31	0.761	-3.00e+07	2.23e+07
natural	-2.76e+07	2.56e+07	-1.08	0.296	-8.14e+07	2.62e+07
engin	-4508943	6321742	-0.71	0.485	-1.78e+07	8772544
art	-2120988	9061093	-0.23	0.818	-2.12e+07	1.69e+07

위의 예에서 (C6-5)식의 추정결과는 (C6-10)의 추정결과와 일치하고, (C6-8)식의 추정결과는 (C6-9)식의 추정결과와 일치함을 알 수 있다.

본 연구는 당초 국립대 및 사립대 평균발전기금에 차이가 있는 지를 살펴보기 위해 시작되었다. 두 집단의 평균차이의 검증을 통해 이를 확인 할 수 있고 (C1)식에서 β_1의 통계적 유의성 검증을 통해 확인할 수 있다.

또한 이에서 확장하여 대학유형 외에 기타 특성변수를 통제했을 경우에도 국립대 및 사립대 평균발전기금에 차이가 있는 지를 살펴보기로 하였다.

첫째, 질적 가변수인 대학소재지역을 통제할 경우(대학유형 및 대학소재지역이 서로 독립이라고 가정할 경우) 국립대 및 사립대 평균발전기금에 차이 여부는 (C2)식에서 β_1의 통계적 유의성 검증을 통해 확인할 수 있다.

둘째, 양적 변수인 학생 수를 통제할 경우(대학유형 및 학생 수가 서로 독립이라고 가정할 경우) 학생 수가 동일 할 경우 국립대 및 사립대 평균발전기금에 차이 여부는 (C5)식에서 β_1의 통계적 유의성 검증을 통해 확인할 수 있다.

한편, 회귀모형에서 합이 1이 되는 양적 변수를 설명변수로 사용할 경우 해석에 유의할 필요가 있음을 보여 주었다. (C3)식에서 β_0는 절편처럼 보이지만 절편으로 해석해서는 안 되고 모형에서 제거된 변수의 기울기로 해석하여야 하며 (C3-2)식이 이를 보여 주고 있다.

또한 본 연구는 대학유형 외에 기타 특성변수를 통제했을 경우에도 국립대 및 사립대 평균발전기금에 차이에 있는 지를 살펴보기 위해 시작되었다.

기타 특성변수를 통제함에 있어 질적 가변수나 양적 변수를 통제변수로 사용하는 것은 본 연구의 당초 목적을 살펴보는데 아무런 문제가 없다. 그러나 (C6-1)식과 같이 합이 1이 되는 양적 변수를 통제(설명)변수로 사용할 경우 β_0는 절편처럼 보이지만 모형에서 제거된 변수의 기울기로 해석해야 한다.

사립대학을 나타내는 (C6-5)식에서 r_4의 기울기가 β_0이고 국립대학을 나타내는

(C6-8)식에서 r_4의 기울기가 $\beta_0+\delta_1$이므로 δ_1을 인문사회비율(r_4)의 변화에 따른 국립대 및 사립대학의 발전기금의 증가의 차이로 해석해야 하므로 국립대 및 사립대학의 평균발전기금의 차이(수준차이)를 보고자 했던 연구의 당초 목적을 달성할 수 없다.

따라서 본 연구의 시사점은 다음과 같이 정리할 수 있다

첫째, 통제변수가 없는 상태에서 국립대 및 사립대 평균발전기금에 차이가 있는 지를 살펴본 결과 즉, 두 집단의 평균차이의 검증 결과가 대학유형 외에 기타 특성변수를 통제했을 경우 어떻게 변화는 지를 살펴보기 위해서는 질적 가변수나 양적 변수를 통제변수로 사용해야 한다. 그래야만 통제변수를 사용하기 전의 결과에 대한 해석과 통제변수를 사용한 후의 결과에 대한 해석이 완전히 일치하게 된다.

둘째, 대학유형 외에 기타 특성변수를 통제했을 경우 국립대 및 사립대의 발전기금 변화에 차이가 있는 지를 살펴보기 위해서는 질적 가변수, 양적 변수뿐만 아니라 합이 1이 되는 양적변수를 사용할 수 있다. 그러나 이 경우 해석에 유의해야 한다. 당초 국립대 및 사립대 평균발전기금에 차이(수준 차이)가 있는 지를 살펴보는 것이 목적이지만 합이 1이 되는 양적 변수를 사용할 경우 국립대 및 사립대 발전기금 차이(증가속도에 의한 차이)를 살펴보는 것이 된다. 따라서 통제변수를 사용하기 전의 결과에 대한 해석과 통제변수를 사용한 후의 결과에 대한 해석이 완전히 일치하는 것이 아니게 된다.

실습 : 가변수모형 추정

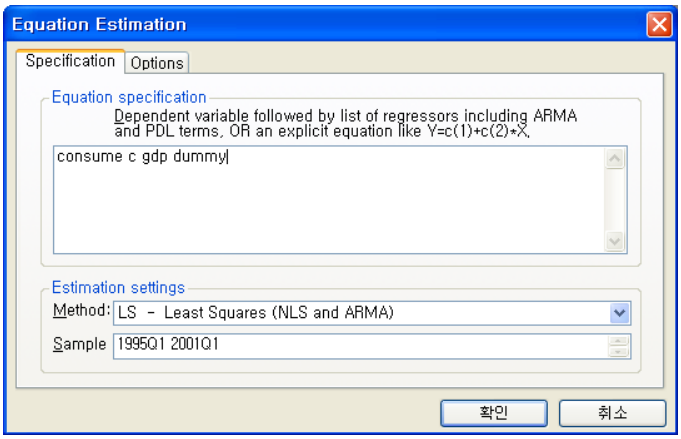
1997년 4/4분기에 발생한 외환위기로 인해 한국경제의 소비행태에 구조적인 변화가 있었는 지를 알아보고자 한다. 1995년 1/4분기부터 2001년 1/4분기까지 분기별 자료를 이용하여 소비함수를 추정하고 다음과 같이 가변수를 설정한다고 하자.

DUMMY = 0 (1995년 1/4분기 - 1997년 4/4분기: IMF 위기 이전)
DUMMY = 1 (1998년 1/4분기 - 2001년 1/4분기 : IMF 위기 이후)

가변수를 만들기 위하여 EViews의 명령어 입력창에서 다음을 입력하여 추세변수, 추세변수를 이용한 가변수 및 가변수와 독립변수를 곱한 새로운 변수를 만든다.

```
genr trend = @trend(1995:1)+1
genr dummy = (trend >=13)
genr dy = dummy*gdp
```

<그림 5-A1> 절편 변화 가변수 추정회귀식



절편의 변화를 나타내는 가변수모형을 추정하기 위하여 Quick-Estimate Equation을 선택하면 나오는 대화상자에서 <그림 5-A1>과 같이 입력을 한다. 여기서 consume은 종속변수인 소비를 나타내고, c는 회귀모형의 상수항을 추정하기 위한 것이며, gdp는 소득을 나타내는 독립변수, dummy는 가변수를 각각 나타낸다.

절편(평균)의 변화를 나타내는 가변수모형인 (5.7)식을 추정한 결과가 <그림 5-A2>에 나타나 있는데 추정된 회귀식은 $\hat{Y}= 13273.01 + 0.509220GDP - 3418.648$ 이다. 가변수는 통계적으로 유의한 것으로 나타나 IMF 위기 이후의 절대소비수준은 IMF 위기 이전의 절대소비수준보다 3418.648만큼 낮아진 것으로 해석할 수 있다.

<그림 5-A2> 절편 변화 가변수 추정 결과

Equation: UNTITLED Workfile: DUMMY::Untitled

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

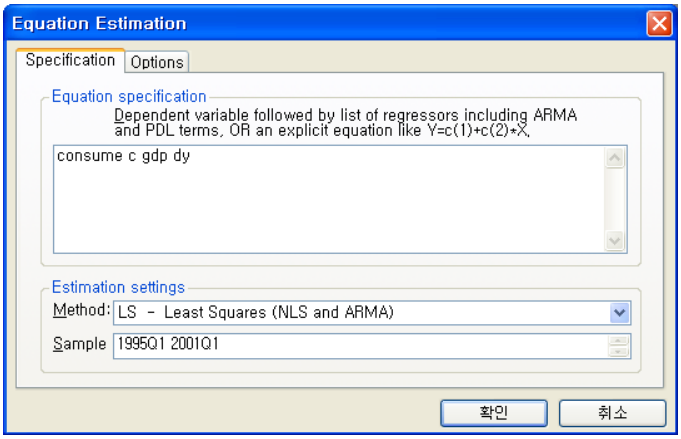
Dependent Variable: CONSUME
Method: Least Squares
Date: 04/05/10 Time: 20:13
Sample: 1995Q1 2001Q1
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13273.01	4046.739	3.279927	0.0034
GDP	0.509220	0.040143	12.68506	0.0000
DUMMY	-3418.648	697.9384	-4.898208	0.0001

R-squared	0.884052	Mean dependent var	65119.26
Adjusted R-squared	0.873511	S.D. dependent var	4070.576
S.E. of regression	1447.710	Akaike info criterion	17.50552
Sum squared resid	46108991	Schwarz criterion	17.65178
Log likelihood	-215.8190	F-statistic	83.87023
Durbin-Watson stat	1.761277	Prob(F-statistic)	0.000000

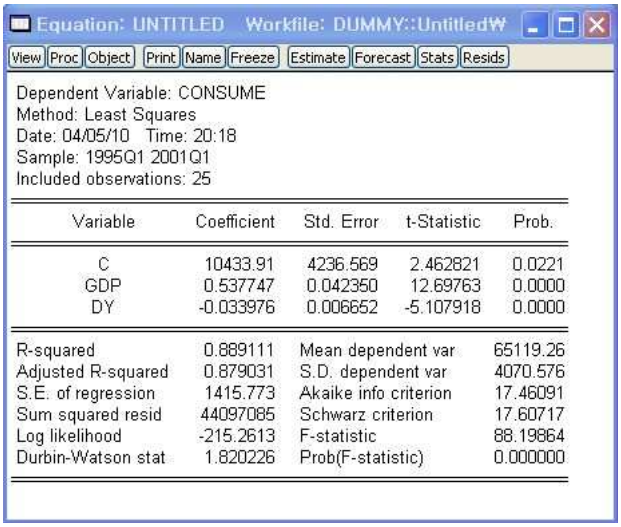
기울기의 변화를 나타내는 가변수모형을 추정하기 위하여 Quick-Estimate Equation을 선택하면 나오는 대화상자에서 <그림 5-A3>과 같이 입력을 한다. 여기서 consume은 종속변수인 소비를 나타내고, c는 회귀모형의 상수항을 추정하기 위한 것이며, gdp는 소득을 나타내는 독립변수, dy는 가변수와 gdp를 곱한 변수를 각각 나타낸다.

<그림 5-A3> 절편 변화 가변수 추정회귀식



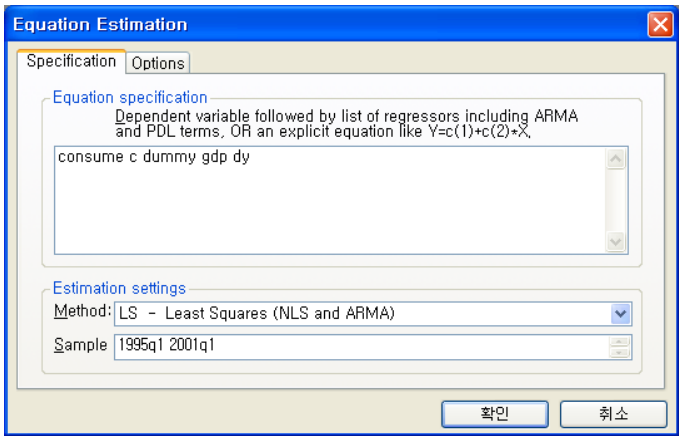
기울기의 변화를 나타내는 가변수모형인 (5.10)식을 추정한 결과가 <그림 5-A4>에 나타나 있는데 추정된 회귀식은 $\hat{Y}=10433.9+0.5377GDP-0.0339DY$이다. 가변수는 통계적으로 유의한 것으로 나타나 IMF 위기 이후의 한계소비성향은 IMF 위기 이전의 한계소비성향보다 0.033976만큼 낮아진 것으로 해석할 수 있다.

<그림 5-A4> 기울기 변화 가변수 추정 결과



절편 및 기울기의 변화를 나타내는 가변수모형을 추정하기 위하여 Quick-Estimate Equation을 선택하면 나오는 대화상자에서 <그림 5-A5>와 같이 입력을 한다. 여기서 consume은 종속변수인 소비를 나타내고, c는 회귀모형의 상수항을 추정하기 위한 것이며, dummy는 가변수를 나타내며, gdp는 소득을 나타내는 독립변수, dy는 가변수와 gdp를 곱한 변수를 각각 나타낸다.

<그림 5-A5> 절편 변화 가변수 추정회귀식



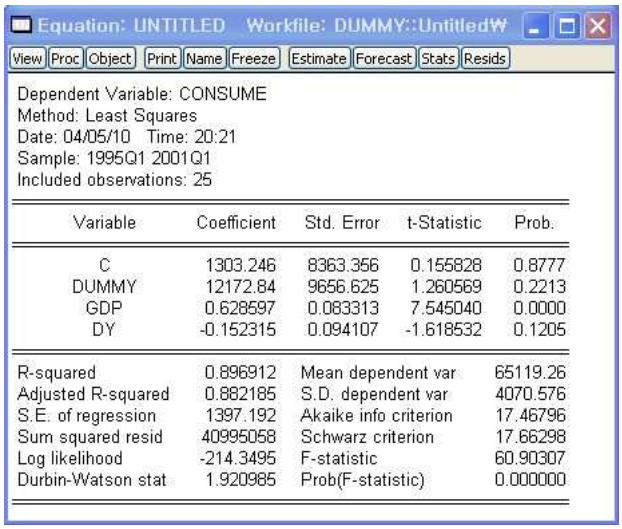
절편 및 기울기의 동시변화를 나타내는 가변수모형인 (5.13)식을 추정한 결과가 <그림 5-A6>에 나타나 있는데 추정된 회귀식은 $\hat{Y}= 1303.2 + 12172.8 + 0.628597GDP - 0.152315DY$이다. 그러나 절편의 변화를 나타내는 가변수 및 기울기의 변화를 나타내는 가변수 모두 통계적으로 유의하지 않은 것으로 나타나 이에 대한 추가적인 검정이 필요하다.

두 개의 가변수가 동시에 유의하지 않다는 귀무가설을 결합검정을 통해 살펴보면 <그림 5-A7>에 나타나 있는데 5% 유의수준에서 귀무가설을 기각할 수 없다. 따라서 절편이 다른 가변수모형 또는 기울기가 다른 가변수모형이 실제 데이터를 잘 나타내 주는 모형이라고 할 수 있다.

만약에 두 개의 가변수가 모두 통계적으로 유의하다면 IMF위기와 IMF위기 이후의 소비수준 및 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 IMF위기와 IMF위기 이후의 소비함수를 각각 추정해야 한다. IMF위기 이전의 소비함수를 추정한 것이 <그림 5-A8>에 나타나 있고, IMF위기 이후의 소비함수를 추

정한 것이 <그림 5-A9>에 나타나 있는데 이것은 (5.13)식의 추정결과인 <그림 5-A6>에서 도출할 수 있음을 확인할 수 있다.

<그림 5-A6> 절편 및 기울기 변화 가변수 추정 결과



Equation: UNTITLED Workfile: DUMMY::Untitled

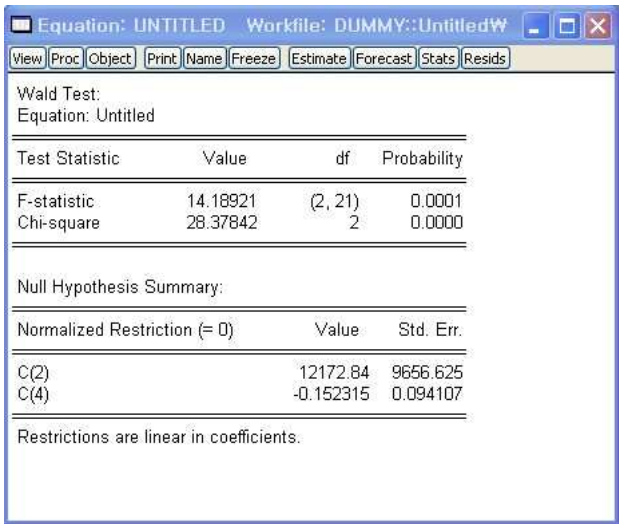
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUME
Method: Least Squares
Date: 04/05/10 Time: 20:21
Sample: 1995Q1 2001Q1
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1303.246	8363.356	0.155828	0.8777
DUMMY	12172.84	9656.625	1.260569	0.2213
GDP	0.628597	0.083313	7.545040	0.0000
DY	-0.152315	0.094107	-1.618532	0.1205

R-squared	0.896912	Mean dependent var	65119.26
Adjusted R-squared	0.882185	S.D. dependent var	4070.576
S.E. of regression	1397.192	Akaike info criterion	17.46796
Sum squared resid	40995058	Schwarz criterion	17.66298
Log likelihood	-214.3495	F-statistic	60.90307
Durbin-Watson stat	1.920985	Prob(F-statistic)	0.000000

<그림 5-A7> 가변수에 대한 결합검정 결과



Equation: UNTITLED Workfile: DUMMY::Untitled

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Wald Test:
Equation: Untitled

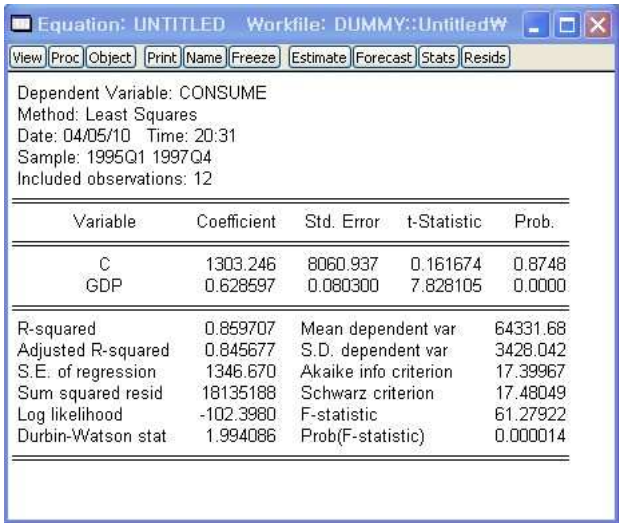
Test Statistic	Value	df	Probability
F-statistic	14.18921	(2, 21)	0.0001
Chi-square	28.37842	2	0.0000

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(2)	12172.84	9656.625
C(4)	-0.152315	0.094107

Restrictions are linear in coefficients.

<그림 5-A8> IMF위기 이전 소비함수 추정 결과



Equation: UNTITLED Workfile: DUMMY::Untitled

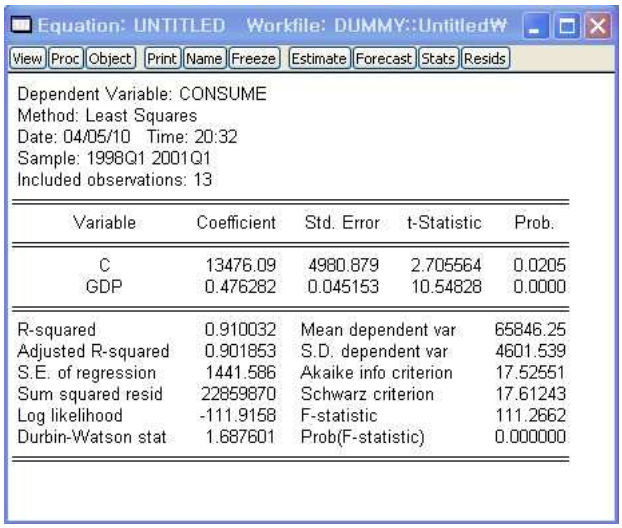
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUME
Method: Least Squares
Date: 04/05/10 Time: 20:31
Sample: 1995Q1 1997Q4
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1303.246	8060.937	0.161674	0.8748
GDP	0.628597	0.080300	7.828105	0.0000

R-squared	0.859707	Mean dependent var	64331.68
Adjusted R-squared	0.845677	S.D. dependent var	3428.042
S.E. of regression	1346.670	Akaike info criterion	17.39967
Sum squared resid	18135188	Schwarz criterion	17.48049
Log likelihood	-102.3980	F-statistic	61.27922
Durbin-Watson stat	1.994086	Prob(F-statistic)	0.000014

<그림 5-A9> IMF위기 이후 소비함수 추정 결과



Equation: UNTITLED Workfile: DUMMY::Untitled

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUME
Method: Least Squares
Date: 04/05/10 Time: 20:32
Sample: 1998Q1 2001Q1
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13476.09	4980.879	2.705564	0.0205
GDP	0.476282	0.045153	10.54828	0.0000

R-squared	0.910032	Mean dependent var	65846.25
Adjusted R-squared	0.901853	S.D. dependent var	4601.539
S.E. of regression	1441.586	Akaike info criterion	17.52551
Sum squared resid	22859870	Schwarz criterion	17.61243
Log likelihood	-111.9158	F-statistic	111.2662
Durbin-Watson stat	1.687601	Prob(F-statistic)	0.000000

연습문제

1. 다음과 같이 2개의 dummy변수를 사용한 추정의 결과가 있다. 단, 괄호안의 값은 t-통계량을 나타낸다.

$$Y_t = \beta_0 D_{1t} + \beta_1 D_{2t} + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad ⓐ$$

$$\widehat{Y_t} = \underset{(1.64)}{18.6} D_{1t} - \underset{(-0.81)}{9.5} D_{2t} + \underset{(5.11)}{0.027} X_{2t} + \underset{(13.77)}{0.16} X_{3t}$$

$$D_{1t} = \begin{cases} 1, & \text{2차대전시}(1939-1945\text{년}) \\ 0, & \text{그 외 연도} \end{cases}$$

$$D_{2t} = \begin{cases} 1, & \text{그 외 연도} \\ 0, & \text{2차대전시}(1939-1945\text{년}) \end{cases}$$

① 이 모형을 상수항과 하나의 dummy변수를 사용할 경우 추정계수가 어떻게 될 것인지를 ( )안을 채워라.

$$Y_t = \gamma_0 + \gamma_1 D_{1t} + \gamma_2 X_{2t} + \gamma_3 X_{3t} + u_t \qquad ⓑ$$

$$\widehat{Y_t} = (\quad) + \underset{(4.69)}{(\quad)} D_{1t} + \underset{(5.11)}{(\quad)} X_{2t} + \underset{(13.77)}{(\quad)} X_{3t}$$

② ⓐ회귀식에서의 β_0, β_1과 ⓑ회귀식에서의 γ_0, γ_1의 관계를 설명하라.

③ ⓑ회귀식에서 D_{1t}의 회귀계수에 대한 통계적 유의성을 5% 유의수준 하에서 검정하라.

④ 위에서 행한 통계적 검정결과의 경제적 의미를 설명하라.

2. 성별에 따른 임금차별이 있는 지를 살펴보고자 한다. 이 경우 다음과 같이 dummy변수를 이용하여 모형을 추정하면 되고 이 때 추정된 계수의 부호가 각각 $\widehat{\beta_1} > 0, \widehat{\beta_2} < 0$이면 임금차별이 있는 것으로 결론을 내릴 수 있다. 이 주장의 진위를 판단하라.

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + u_i$$

단, $D_{1i} = \begin{cases} 1, & \text{남자} \\ 0, & \text{여자} \end{cases}$

$D_{2i} = \begin{cases} 1, & \text{여자} \\ 0, & \text{남자} \end{cases}$

$X_i =$ 근속연수

3. 전쟁시와 평화시 소비수준에 차이가 있는 지를 살펴보기 위해 다음과 같이 dummy변수를 이용하여 모형 $C_t = \gamma_1 + \gamma_2 D_{1i} + \beta Y_t + u_t$ (단, $D_{1i} = 1$ 평화시, $D_{1i} = 0$ 전쟁시)를 추정하여 $\widehat{\gamma_1} = 0.27$, $\widehat{\gamma_2} = 1.01$, $\hat{\beta} = 0.6$을 얻었다.

① 전쟁시와 평화시의 소비수준과 한계소비성향을 각각 구하라.

② 전쟁시 소비수준과 평화시 소비수준이 차이가 있는 지를 알아보고자 한다. 귀무가설을 설정하고 검정절차를 설명하라.

제 6 장

자기상관

1.일반화최소자승법
2.자기상관 개념 및 유형
3.자기상관 검정
4.자기상관 해결책
실습 : 자기상관 추정

제6장 자기상관

1.일반화최소자승법

제4장에서 살펴본 (6.1)식의 행렬표현 다중회귀모형에서 기본적인 가정들이 성립하는 고전적 회귀모형의 경우 보통최소자승법으로 추정한 추정량은 최량선형불편추정량(BLUE)이 됨을 보인 바 있다.

$$Y = X\ B + U$$
nx1 nxk kx1 nx1

$$\text{단, } Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix},\ X = \begin{bmatrix} 1\ X_{21}\ X_{31} \cdots\ X_{k1} \\ 1\ X_{22}\ X_{32} \cdots\ X_{k2} \\ .\quad .\quad .\quad . \\ 1\ X_{2n}\ X_{3n} \cdots\ X_{kn} \end{bmatrix},\ B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix},\ U = \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_n \end{bmatrix} \tag{6.1}$$

여러 가지 기본적인 가정들 중 (6.2)식의 가정은 교란항의 동분산(homoscedasticity)과 비자기상관(no autocorrelation)을 가정한 것이며 이 경우 OLS로 추정한 추정량은 BLUE가 된다.

$$E(UU') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) \cdot\cdot E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) \cdot\cdot E(u_2u_k) \\ . \quad . \quad \cdot\cdot \quad . \\ E(u_nu_1)\ E(u_nu_2) \cdot\cdot & E(u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_u^2 & 0 \cdot\cdot 0 \\ 0 & \sigma_u^2 \cdot\cdot 0 \\ . \quad . \quad \cdot\cdot \quad . \\ 0 & 0 \cdot\cdot \sigma_u^2 \end{bmatrix}$$

$$= \sigma_u^2 I_n \tag{6.2}$$

그러나 현실적으로 (6.2)식의 가정이 성립하지 않는 경우가 발생한다. 즉, $E(UU') \neq \sigma_u^2 I_n$인 두 경우가 있는데 하나는 (6.3)과 같은 자기상관(1차 자기상관)이 있는 경우이고 다른 하나는 (6.4)식과 같은 이분산(heteroscedasticity)인 경우이다.

$$E(UU') = \sigma_u^2 \Omega = \frac{\sigma_\epsilon^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 \cdot\cdot \rho^{n-1} \\ \rho & 1 & \rho \cdot\cdot \rho^{n-2} \\ \cdot & \cdot & \cdot \cdot\cdot \cdot \\ \cdot & \cdot & \cdot \cdot\cdot \cdot \\ \rho^{n-1} & \rho^{n-2} & \cdot \cdot\cdot 1 \end{bmatrix} \neq \sigma_u^2 I_n \tag{6.3}$$

$$E(UU') = \sigma_u^2 \Omega = \sigma_u^2 \begin{bmatrix} k_1^2 & 0 & \cdot\cdot & 0 \\ 0 & k_2^2 & \cdot\cdot & 0 \\ \cdot & \cdot & \cdot\cdot & \cdot \\ 0 & 0 & \cdot\cdot & k_n^2 \end{bmatrix} \neq \sigma_u^2 I_n \tag{6.4}$$

위의 경우와 같이 자기상관이나 이분산이 있음에도 불구하고 보통최소자승법으로 추정하면 추정량은 편의는 없으나 효율적인 추정량이 되지 못한다. 제4장에서 살펴본 바와 같이 (6.1)식의 OLS 추정량은 (6.5)식이다.

$$\hat{B} = (X'X)^{-1} X'Y \tag{6.5}$$

(6.5)식으로부터 다음의 (6.6)식을 도출할 수 있다.

$$\begin{aligned} \hat{B} &= (X'X)^{-1} X'Y \\ &= (X'X)^{-1} X'(XB + U) \\ &= (X'X)^{-1} X'XB + (X'X)^{-1} X'U \\ &= B + (X'X)^{-1} X'U \end{aligned} \tag{6.6}$$

(6.6)식의 OLS 추정량의 성질을 살펴보기 위해 (6.6)식에 기댓값을 계산하면 (6.7)식과 같게 되고 $\hat{B}$은 B의 불편추정량(unbiased estimator)이다.

$$E(\hat{B}) = B + (X'X)^{-1} X'E(U)$$

$$= B \tag{6.7}$$

다음으로 $E(UU') \neq \sigma_u^2 I_n$인 경우 OLS 추정량의 분산을 구해 보면 (6.8)식과 같다.

$$\begin{aligned} Var(\widehat{B}) &= E[(\hat{B} - E(\hat{B})(\hat{B} - E(\hat{B}))'] \\ &= E[(\hat{B} - B)(\hat{B} - B)'] \\ &= (X'X)^{-1}X'E(UU')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma_u^2 \Omega X(X'X)^{-1} \\ &= \sigma_u^2 (X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned} \tag{6.8}$$

제4장에서 $E(UU') = \sigma_u^2 I_n$인 경우 분산은 $Var(\hat{B}) = \sigma_u^2 (X'X)^{-1}$으로 최소분산임을 살펴보았는데(이를 Gauss-Markov정리라고 함) $E(UU') \neq \sigma_u^2 I_n$인 경우 $Var(\widehat{B}) = \sigma_u^2 (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$으로 최소분산을 갖지 못한다.

즉, 자기상관이나 이분산이 있음에도 불구하고 보통최소자승법으로 회귀계수를 추정하면 편의는 없으나 효율적인 추정량을 얻지 못하고 통계적 추론에서도 문제가 발생한다.

한편, 고전적 회귀모형에서 교란항에 대한 기본가정이 충족되지 않을 경우 모형을 변환하여 최소자승법으로 추정하면 최량불편추정량을 얻을 수 있다. 이를 일반화최소자승법(Generalized Least Squares:GLS) 또는 Aitken의 최소자승법이라 한다.

GLS의 아이디어 및 절차는 다음과 같다.

첫째, $PP' = \Omega$라고 하면 $(P^{-1})'P^{-1} = \Omega^{-1}$이 된다. 여기서 P^{-1}를 L로 두면 아래의 명제에 의하여 $L'L = \Omega^{-1}$을 만족하는 비특이행렬 L이 존재한다.

(명제) 어떠한 대칭양정부호행렬(symmetric positive definite matrix) Ω에 대해서 $PP' = \Omega$를 충족시키는 비특이행렬(nonsingular matrix) P가 존재한다.

둘째, (6.9)식의 회귀모형이 고전적 회귀모형의 기본가정을 충족시키지 못하고 다음과 같다고 하자.

$$Y = XB + U \tag{6.9}$$

단, $E(UU') = \sigma^2\Omega$

이러한 경우 위의 회귀모형에 L을 곱하여 변환하면 다음의 (6.10)식 또는 (6.11)식과 같이 나타낼 수 있다.

$$LY = LXB + LU \tag{6.10}$$

또는

$$Y^* = X^*B + U^* \tag{6.11}$$

셋째, 변환된 모형((6.11)식)의 교란항 평균을 구해보면 (6.12)식과 같이 0이 된다.

$$\begin{aligned} E(U^*) &= E(LU) \\ &= LE(U) \\ &= 0 \end{aligned} \tag{6.12}$$

또한 교란항의 분산을 계산해 보면 (6.13)식과 같다.

$$\begin{aligned} E(U^*U^{*'}) &= Var(LU) \\ &= E(LUU'L') \\ &= \sigma^2 L\Omega L' \\ &= \sigma^2 LL^{-1}L'^{-1}L' (\because L^{-1}L'^{-1} = \Omega \leftarrow L'L = \Omega^{-1}) \\ &= \sigma^2 I_n \end{aligned} \tag{6.13}$$

(6.12)식 및 (6.13)식에서 보여주고 있듯이 변환된 모형의 교란항 평균은 0이고 분산은 자기상관이 되어 있지 않으며 동분산의 조건을 만족하므로 변환된 모형을 최소자승법으로 추정하면 BLUE를 얻을 수 있는데 이러한 방법을 일반화최소자승법이라고 한다.

넷째, 따라서 GLS 추정량은 (6.11)식의 변환된 모형을 OLS로 추정하면 구할 수

있는데 (6.14)식과 같다.

$$\begin{aligned}\widehat{B_G} &= (X^{*'}X^{*})^{-1}X^{*'}Y^{*} \\ &= (X'L'LX)^{-1}X'L'LY \\ &= (X^{-1}L^{-1}L^{-1'}X^{-1'})X'L'LY \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \end{aligned} \quad (6.14)$$

또한 교란항의 분산 및 회귀계수의 분산에 대한 추정량은 각각 (6.15)식 및 (6.16)식과 다음과 같다.

$$\widehat{\sigma_u^2} = \frac{e'\Omega^{-1}e}{n-k} \quad (6.15)$$

$$Var(\widehat{B_G}) = \widehat{\sigma_u^2}(X'\Omega^{-1}X)^{-1} \quad (6.16)$$

2.자기상관 개념 및 유형

경제시계열의 경우 과거 값이 현재에 영향을 주고 현재 값이 미래에 영향을 주어 추세나 순환현상을 보이는 경우가 많이 있다. 이러한 현상을 보이는 이유는 경제에 외부 충격(예상치 못한 변화)이 주어졌을 때 그 충격의 영향이 단기간에 끝나는 것이 아니고 장기적으로 지속되기 때문이다.

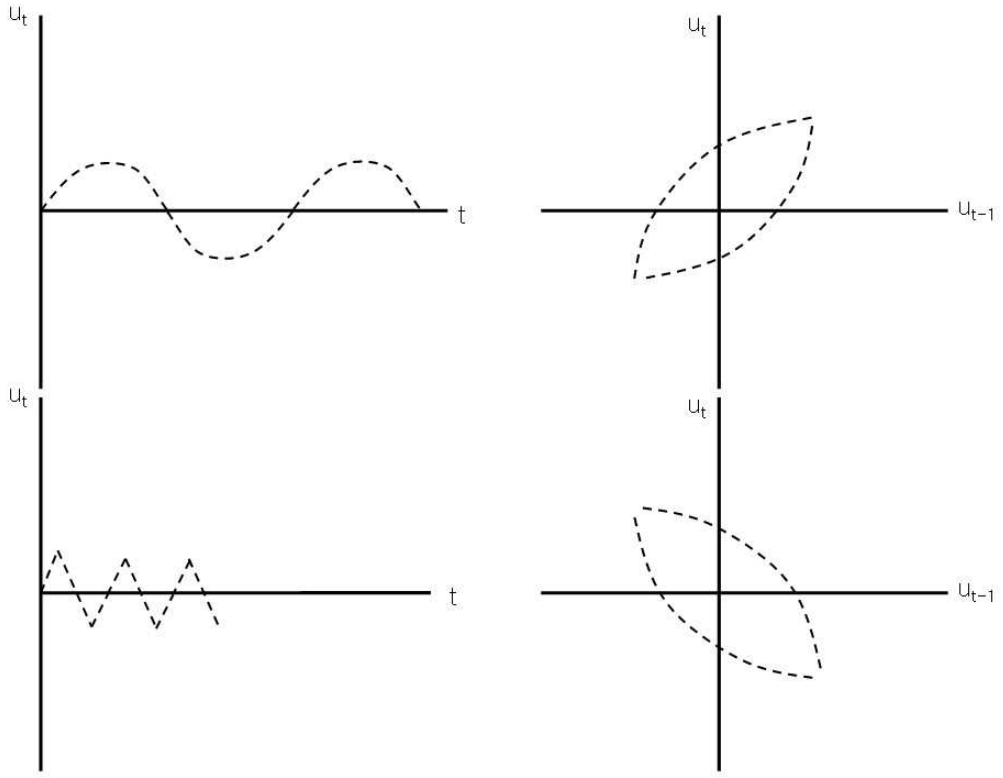
(6.17)식과 같이 시계열에 있어 t기의 값인 u_t(이를 예상치 못한 변화로 해석할 수 있음)가 t-1기의 값인 u_{t-1}와 상관관계가 있는 경우를 계열상관(serial correlation) 또는 자기상관(autocorrelation)이 있다고 한다. 특히 이 경우를 1차 자기상관(first-order autocorrelation)이라 한다.

$$Y_t = X_t B + u_t$$

$$u_t = \rho u_{t-1} + e_t \quad (6.17)$$

자기상관의 유형에는 양의 계열상관(positive autocorrelation) 및 음의 계열상관(negative autocorrelation)이 있으며 그 유형은 <그림 6-1>과 같다.

<그림 6-1> 양의 자기상관(상단그림) 및 음의 자기상관(하단그림)



3.자기상관 검정

시계열의 자기상관 여부를 탐지하는 방법으로는 회귀식으로부터 도출된 잔차를 그려 보아 자기상관 여부를 판단하는 방법인 잔차의 그래프 분석(residual plotting)과 통계적 검정을 통해 자기상관 여부를 판단하는 방법인 Durbin-Watson 검정방법, LM(Lagrange Multiplier) 검정방법 등이 있다.

자기상관을 탐지하는 가장 간단한 방법은 잔차에 상관관계가 나타나는 지를 그래프를 통해 살펴보는 것인데 잔차가 <그림 6-1>과 같은 경향을 보이면 자기상관이 있는 것으로 판단한다. 그러나 이 방법은 주관적이며 과학적이지 못하다고 할 수 있다.

자기상관 여부를 판단하는 또 다른 방법으로는 Durbin-Watson(DW) 검정방법이

있는데 이 방법은 교란항의 자기상관 여부를 검정하는데 가장 많이 사용되는 방법으로 다음의 (6.18)식과 같은 d-통계량을 이용한다.

$$
\begin{aligned}
d &= \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2} \\
&= \frac{\sum_{t=2}^{N} e_t^2 + \sum_{t=2}^{N} e_{t-1}^2 - 2\sum_{t=2}^{N} e_t e_{t-1}}{\sum_{t=1}^{N} e_t^2} \\
&\fallingdotseq 2(1-\hat{\rho}) \qquad (6.18)
\end{aligned}
$$

d-통계량을 이용한 자기상관 여부는 다음과 같이 판단한다.

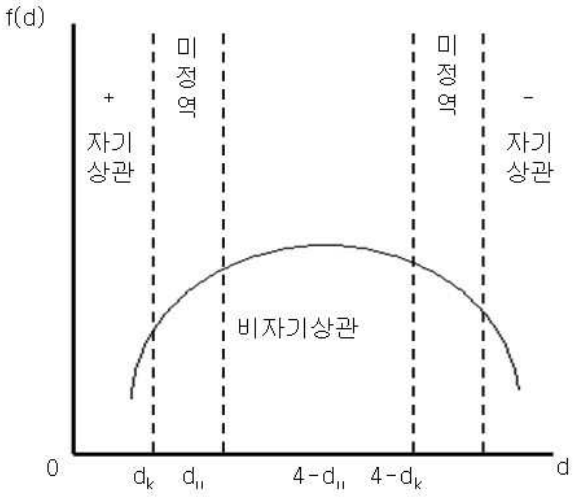
첫째, $\hat{\rho}=0$이면 $d \fallingdotseq 2$이므로 d-통계량의 값이 2에 가까우면 자기상관이 없는 것으로 판단한다.

둘째, $\hat{\rho}>0$이면 $d<2$이고, 특히 $\hat{\rho}$이 1로 접근할수록 d는 0으로 접근하므로 d-통계량의 값이 0과 2사이에 있으면 양(+)의 자기상관이 있는 것으로 판단한다.

셋째, $\hat{\rho}<0$이면 $d>2$이고, 특히 $\hat{\rho}$이 -1로 접근할수록 d는 4로 접근하므로 d-통계량의 값이 2와 4사이에 있으면 음(-)의 자기상관이 있는 것으로 판단한다.

한편, 자기상관 여부는 범위검정(bounds test) 방법을 이용하여 검정할 수도 있다. DW-검정(<부록 4>의 <표 5>)은 1차 자기상관을 검정하며, 회귀모형은 상수항을 반드시 포함하고 있어야 하고, 회귀모형의 설명변수에는 종속변수의 시차변수가 존재하지 않아야 한다. 즉, 자기회귀모형에는 DW-검정방법을 적용할 수 없으며 이때는 h통계량을 이용한다. 또한 이 방법은 분포를 이용한 검정방법이라는 장점은 있으나 자기상관 여부를 결정할 수 없는 영역(미정영역)이 있다는 단점이 있다.

<그림 6-2> 자기상관 검정



자기상관이 있음에도 불구하고 OLS로 추정하면 불편추정량은 얻을 수 있으나 효율적인 추정량은 얻지 못하고 따라서 정확한 통계적 추론이 어렵다는 것은 설명한 바 있다. 따라서 원자료(original data)를 적당한 방법으로 변환시켜 고전적 회귀모형의 기본가정에 맞도록 한 후 OLS로 추정하면 된다.

그러면 원자료를 어떻게 변환시킬 것인가? 이에 대한 단서를 찾기 위하여 (6,19)식과 같이 교란항이 1차 자기상관이 있을 경우 교란항의 분산-공분산행렬(variance-covariance matrix)은 어떻게 생겼는지를 살펴보자.

$$u_t = \rho u_{t-1} + e_t \qquad (6.19)$$

단, $E(e_t = 0)$. $E(e_t e_t') = \sigma_e^2 I_n$, $E(u_{t-1} e_t) = 0$이다.

(6.19)식을 1차 차분방정식(difference equation)이라 하는데 차분방정식을 푸는 방법은 여러 가지가 있으나 그 중 하나의 방법으로 연속적인 대입방법(successive substitution)이 있다.

(6.19)식을 연속적인 대입을 통해 차분방정식을 풀어보면 (6.20)식과 같다.

$$u_t = \rho u_{t-1} + e_t$$

$$
\begin{aligned}
&= \rho(\rho u_{t-2} + e_{t-1}) + e_t \\
&= e_t + \rho e_{t-1} + \rho^2 u_{t-2} \\
&= e_t + \rho e_{t-1} + \rho^2 (e_{t-2} + \rho u_{t-3}) \\
&\quad . \\
&= e_t + \rho e_{t-1} + \rho^2 e_{t-2} + \rho^3 e_{t-3} + \cdots \\
&= \sum_{i=0}^{\infty} \rho^i e_{t-i} \qquad (6.20)
\end{aligned}
$$

(6.20)식을 이용하여 u_t의 분산을 구해 보면 (6.21)식과 같다.

$$
\begin{aligned}
Var(u_t) &= \sigma_e^2 + \rho^2\sigma_e^2 + \rho^4\sigma_e^2 + \cdots \\
&= \sigma_e^2(1 + \rho^2 + \rho^4 + \cdots) \\
&= \frac{\sigma_e^2}{1-\rho^2} \\
&= \sigma_u^2 \qquad (6.21)
\end{aligned}
$$

또한 (6.20)식을 이용하여 u_t와 $u_{t-j}(j = 1,2,\ldots,s)$의 공분산을 구해 보면 각각 다음과 같다.

$$
\begin{aligned}
Cov(u_t u_{t-1}) &= \rho\sigma_e^2 + \rho^3\sigma_e^2 + \rho^5\sigma_e^2 + \cdots \\
&= \frac{\rho}{1-\rho^2}\sigma_e^2 \\
&= \rho\sigma_u^2
\end{aligned}
$$

$$
\begin{aligned}
Cov(u_t u_{t-2}) &= \rho^2\sigma_e^2 + \rho^4\sigma_e^2 + \rho^6\sigma_e^2 + \cdots \\
&= \frac{\rho^2}{1-\rho^2}\sigma_e^2 \\
&= \rho^2\sigma_u^2
\end{aligned}
$$

.

$$
\begin{aligned}
Cov(u_t u_{t-s}) &= \rho^s \sigma_e^2 + \rho^{(s+2)} \sigma_e^2 + \rho^{(s+4)} \sigma_e^2 + \ldots \\
&= \frac{\rho^s}{1-\rho^2} \sigma_e^2 \\
&= \rho^s \sigma_u^2
\end{aligned}
$$

따라서 교란항의 분산-공분산 행렬은 (6.22)식과 같다.

$$
\begin{aligned}
E(UU') &= \begin{bmatrix} E(u_1^2) & E(u_1 u_2) \ldots E(u_1 u_n) \\ E(u_2 u_1) & E(u_2^2) \ldots E(u_2 u_n) \\ \cdot & \cdot \ldots \cdot \\ \cdot & \cdot \ldots \cdot \\ E(u_n u_1) & E(u_n u_2) \ldots E(u_n^2) \end{bmatrix} \\
&= \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 \ldots \rho^{n-1} \\ \rho & 1 & \rho \ldots \rho^{n-2} \\ \cdot & \cdot & \cdot \ldots \cdot \\ \cdot & \cdot & \cdot \ldots \cdot \\ \rho^{n-1} & \rho^{n-2} & \ldots 1 \end{bmatrix} \\
&= \frac{\sigma_e^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 \ldots \rho^{n-1} \\ \rho & 1 & \rho \ldots \rho^{n-2} \\ \cdot & \cdot & \cdot \ldots \cdot \\ \cdot & \cdot & \cdot \ldots \cdot \\ \rho^{n-1} & \rho^{n-2} & \ldots 1 \end{bmatrix} \\
&= \sigma_e^2 \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} \ldots \frac{\rho^{n-1}}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \ldots \frac{\rho^{n-2}}{1-\rho^2} \\ \cdot & \cdot \ldots \cdot \\ \cdot & \cdot \ldots \cdot \\ \frac{\rho^{n-1}}{1-\rho^2} & \frac{\rho^{n-2}}{1-\rho^2} \ldots \frac{1}{1-\rho^2} \end{bmatrix} \\
&= \sigma_e^2 \Omega
\end{aligned}
$$

위의 Ω로부터 (6.23)식의 Ω 역행렬을 구할 수 있고 이는 $\Omega\Omega^{-1} = I$임을 통해 확인할 수 있다.

$$\Omega^{-1}=\begin{bmatrix} 1 & -\rho & 0 & . & . & 0 \\ -\rho & 1+\rho^2 & -\rho & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & 1+\rho^2 & -\rho \\ 0 & 0 & . & . & -\rho & 1 \end{bmatrix} \tag{6.23}$$

(6.23)식으로부터 (6.24)식의 L행렬을 구할 수 있고 이는 $L'L=\Omega^{-1}$임을 통해 확인할 수 있다.

$$L=\begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & . & . & 0 \\ -\rho & 1 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & 1 & 0 \\ 0 & 0 & . & . & -\rho & 1 \end{bmatrix} \tag{6.24}$$

L을 이용하여 원래자료를 변환시키면 종속변수는 다음의 (6.25)식과 같이 변환되는데 이를 Paris-Winstern 변환이라 한다.

$$y_1^*=\sqrt{1-\rho^2}\,y_1$$

$$y_t^*=y_t-\rho y_{t-1}(t\geqq 2) \tag{6.25}$$

L을 이용하여 독립변수 역시 다음의 (6.26)식과 같이 변환된다.

$$x_1^*=\sqrt{1-\rho^2}\,x_1$$

$$x_t^*=x_t-\rho x_{t-1}(t\geqq 2) \tag{6.26}$$

한편, 자기상관을 검정하는 또 다른 방법으로 LM(Lagrange Multiplier) 검정방법이 있다. 이 검정방법은 Durbin-Watson 검정방법의 한계점이나 단점에서 지적한 문제와 관계없이 사용할 수 있는 검정방법이라는 장점이 있으나 자료의 수가 많을 때 사용될 수 있다는 단점이 있다.

예를 들어, 다음의 (6.27)식에 대한 LM-검정은 다음과 같은 순서로 한다.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

$$u_t = \rho u_{t-1} + \epsilon_t \qquad (6.27)$$

제1단계로 (6.27)식의 모형을 추정한 후 다음의 (6.28)식과 같은 잔차를 구한다.

$$\hat{u_t} = Y_t - \hat{\beta_1} - \hat{\beta_2} X_{2t} - \hat{\beta_3} X_{3t} \qquad (6.28)$$

2단계로 (6.28)식에서 구한 잔차를 (6.27)식에 포함되어 있는 설명변수와 $\hat{u}_{t-1}$에 대해 다음의 (6.29)식과 회귀모형을 추정하는데 이를 보조회귀식(auxiliary regression)이라고 한다.

$$\hat{u_t} = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \rho \hat{u}_{t-1} + v_t \qquad (6.29)$$

이 경우 검정하고자 하는 귀무가설은 $H_o : \rho = 0$ 즉, 자기상관이 없다는 것이며 귀무가설 하에서 (6.30)식의 LM 검정통계량과 그 분포를 구할 수 있다.

$$LM = nR^2 \sim \chi_1^2 \qquad (6.30)$$

단, n은 관측치의 수, R^2는 보조회귀식에서의 결정계수이며 χ^2-분포의 자유도는 1이다.

지금까지 1차 자기상관의 경우를 예로 들었지만 이를 확장하면 LM 검정방법은 p차 자기상관의 검정에도 사용된다.

3.자기상관 해결책

자기상관이 있음에도 불구하고 OLS로 추정하면 불편추정량은 얻을 수 있으나 효율적인 추정량은 얻지 못하고 따라서 정확한 통계적 추론이 어렵다. 따라서 원자료를 적당한 방법으로 변환시켜 고전적 회귀모형의 기본가정에 맞도록 한 후 OLS로 추정한다.

(6.27)식의 모형을 예로 들어 원자료를 변환시키는 방법을 살펴보면 다음과 같

다.

먼저 (6.27)식을 한 기간 이전의 식으로 표현하면 (6.31)식과 같다.

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + u_{t-1} \tag{6.31}$$

(6.31)식의 양변에 ρ를 곱한 후 (6.27)식에서 빼면 다음의 (6.32)식과 같게 되는데 이를 Cochrane-Orcutt 변환(transformation)이라고 한다.

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{2t} - \rho X_{2t-1}) + \beta_3(X_{3t} - \rho X_{3t-1}) + \epsilon_t$$

$$\epsilon_t = u_t - \rho u_{t-1} \tag{6.32}$$

(6.32)식에서 $\beta_1^* = \beta_1(1-\rho)$라 표기하면 다음의 (6.33)식과 같다.

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \epsilon_t \tag{6.33}$$

단, $Y_t^* = Y_t - \rho Y_{t-1,}$ $X_{2t}^* = (X_{2t} - \rho X_{2t-1})$, $X_{3t}^* = (X_{3t} - \rho X_{3t-1})$이다.

(6.33)식의 추정방법은 ρ를 알고 있는 경우와 ρ를 모르는 경우에 따라 달라진다.

먼저 ρ의 값이 알려져 있을 때는 $Y_t^* = Y_t - \rho Y_{t-1,}$, $X_{2t}^* = (X_{2t} - \rho X_{2t-1})$, $X_{3t}^* = (X_{3t} - \rho X_{3t-1})$를 구한 후 (6.33)식을 OLS로 추정하면 효율적인 추정량을 얻을 수 있다.

다음으로 ρ의 값이 알려져 있지 않을 경우 ρ을 추정한 후 자료를 변환하고 OLS를 적용하는데 ρ값을 추정하는 방법에는 여러 가지가 있다.

(1) Durbin의 2단계 추정법

1단계로 (6.32)식을 다시 정리하면 (6.34)식이 되는데 이를 OLS로 추정하여 $\hat{\rho}$를 얻는다.

$$Y_t = \rho Y_{t-1} + \beta_1^* + \beta_2 X_{2t} - \beta_2 \rho X_{2t-1} + \beta_3 X_{3t} - \beta_3 \rho X_{3t-1} + \epsilon_t \tag{6.34}$$

2단계로 1단계에서 구한 $\hat{\rho}$으로 (6.33)식에 의해 $Y_t^*, X_{2t}^*, X_{3t}^*$를 구한 후 (6.33)식을 OLS로 추정하면 된다.

(2) Cochrane-Orcutt의 2단계 절차

1단계로 (6.31)식을 OLS로 추정하여 $\hat{u_t}$를 얻고 다음으로 $\hat{u_t} = \rho\hat{u}_{t-1} + e_t$의 회귀모형을 OLS로 추정하여 $\hat{\rho}$을 얻는다.

2단계로 이 $\hat{\rho}$을 이용하여 $Y_t^*, X_{2t}^*, X_{3t}^*$를 구한 후 (6.33)식을 OLS로 추정하면 된다.

(3) Cochrane-Orcutt의 반복절차

1단계로 (6.31)식을 OLS로 추정하여 잔차 $\hat{u_t}$를 구한다.

2단계로 1단계에서 추정된 잔차를 이용하여 다음 식을 OLS로 추정하여 $\hat{\rho}$을 구하는데 이를 ρ의 1단계 추정치라고 한다.

$$\hat{u_t} = \rho\hat{u}_{t-1} + e_t$$

3단계로 2단계에서 추정된 $\hat{\rho}$을 이용하여 $Y_t^*, X_{2t}^*, X_{3t}^*$를 구한 후 (6.33)식을 OLS로 추정하여 $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*$을 구한다.

4단계로 $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*$을 이용하여 다음의 새로운 잔차를 구한다.

$$\hat{u_t^*} = Y_t - \hat{\beta}_1^* - \hat{\beta}_2^* X_{2t} - \hat{\beta}_3^* X_{3t}$$

5단계로 새로운 잔차를 이용하여 다음 식을 OLS로 추정하여 새로운 1차 자기상관계수($\hat{\hat{\rho}}$)를 구하는데 이를 ρ의 2단계 추정치라고 한다.

$$\hat{u_t^*} = \rho\hat{u}_{t-1}^*$$

6단계는 위의 과정을 다음의 수렴조건이 충족될 때까지 반복한다.

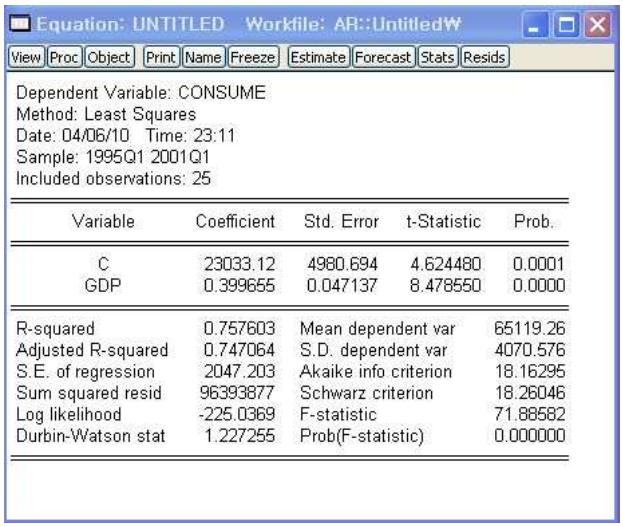
$$| \hat{\rho}_j - \widehat{\rho_{j+1}} | < 0.005$$

단, 0.005는 아주 작은 수의 예시로 이를 수렴한계치(convergence limit)라고 한다.

실습 : 자기상관 추정

1995년 1/4분기부터 2001년 1/4분기까지 한국의 GDP와 총소비(consume) 자료를 이용하여 소비함수를 추정한 결과가 <그림 6-A1>에 나타나 있다. DW-검정을 이용하여 1차 자기상관 여부를 판단해 보자. Durbin-Watson stat이 1.227255이고 n=25, 상수항을 제외한 독립변수의 수인 $k'=1$이다. 이를 <부록 4>의 <표 5>에서 찾아보면 $d_L=1.288$이고 $d_U=1.454$인데 Durbin-Watson stat 1.227255가 $d_L=1.288$보다 작으므로 <그림 6-2>에서 나타내고 있는 것처럼 DW-검정 결과 양(+)의 1차 자기상관이 있는 것으로 나타났다.

<그림 6-A1> 소비함수 추정 결과


Equation: UNTITLED Workfile: AR::UntitledW

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUME
Method: Least Squares
Date: 04/06/10 Time: 23:11
Sample: 1995Q1 2001Q1
Included observations: 25

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	23033.12	4980.694	4.624480	0.0001
GDP	0.399655	0.047137	8.478550	0.0000

R-squared	0.757603	Mean dependent var	65119.26
Adjusted R-squared	0.747064	S.D. dependent var	4070.576
S.E. of regression	2047.203	Akaike info criterion	18.16295
Sum squared resid	96393877	Schwarz criterion	18.26046
Log likelihood	-225.0369	F-statistic	71.88582
Durbin-Watson stat	1.227255	Prob(F-statistic)	0.000000

한편, LM-검정법을 이용하여 자기상관 여부를 판단하기 위해서는 <그림 6-A1>의 추정 결과 창에서 View-Residual Tests-Serial Correlation LM Test를 차례대로 클릭하면 <그림 6-A2>와 같은 결과를 나타내 준다. 여기서 Obs*R-squared 3.122012가 (6.30)식의 $LM=nR^2$ 값이며 Prob. Chi-Square(1) 0.077241은 χ_1^2의

유의확률이 0.077241이라는 의미이므로 10% 유의수준에서 1차 자기상관이 있는 것으로 결정한다.

<그림 6-A2> LM-검정 결과

Equation: UNTITLED Workfile: AR::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	3.139423	Prob. F(1,22)	0.090273
Obs*R-squared	3.122012	Prob. Chi-Square(1)	0.077241

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 04/06/10 Time: 23:20
Sample: 1995Q1 2001Q1
Included observations: 25
Presample missing value lagged residuals set to zero.

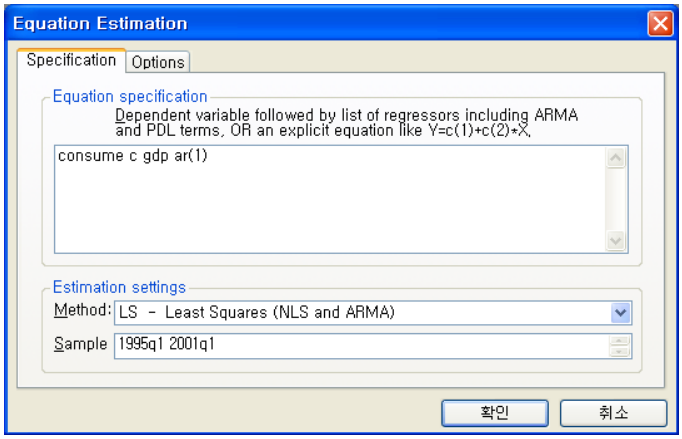
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	404.1083	4769.504	0.084728	0.9332
GDP	-0.004151	0.045148	-0.091950	0.9276
RESID(-1)	0.364122	0.205505	1.771842	0.0903

R-squared	0.124880	Mean dependent var	8.15E-12
Adjusted R-squared	0.045324	S.D. dependent var	2004.099
S.E. of regression	1958.155	Akaike info criterion	18.10956
Sum squared resid	84356164	Schwarz criterion	18.25582
Log likelihood	-223.3695	F-statistic	1.569711
Durbin-Watson stat	1.796393	Prob(F-statistic)	0.230537

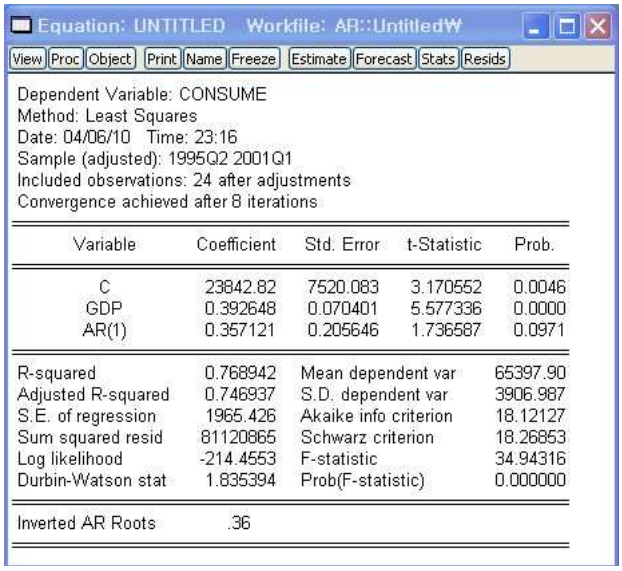
DW-검정과 LM-검정 모두 소비함수에 1차 자기상관이 있는 것으로 나타났으므로 회귀계수 및 1차 자기상관계수를 동시에 추정해야 한다. 이를 위해서 Quick-Estimate Equation을 선택하면 나오는 대화상자에서 <그림 6-A3>과 같이 입력을 한다. 여기서 consume은 종속변수인 소비를 나타내고, c는 회귀모형의 상수항을 추정하기 위한 것이며, gdp는 소득을 나타내는 독립변수, ar(1)은 1차 자기상관이 있음을 각각 나타낸다.

<그림 6-A4>는 1차 자기상관을 고려한 소비함수의 추정 결과를 나타내 주고 있는데 이것이 GLS 추정치이다. 이를 <그림 6-A1>의 추정 결과에 나타나 있는 OLS 추정치와 비교해 보면 그 값이 다름을 알 수 있다.

<그림 6-A3> 1차 자기상관 추정 회귀식



<그림 6-A4> GLS 추정결과

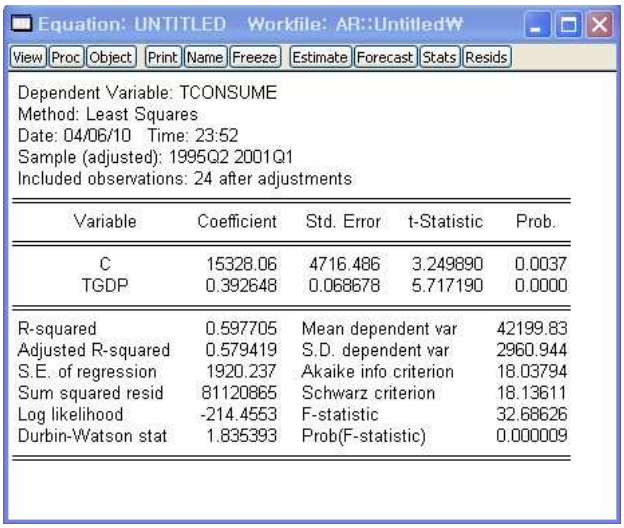


원래자료를 변환한 후 변환된 자료로 OLS로 추정하면 그 추정량이 GLS 추정량임을 설명한 바 있다. <그림 6-A4>에서 $\hat{\rho}=0.357121$이므로 이를 이용하여 다음과 같이 자료 변환을 한다.

```
genr tconsume = consume-0.357121*consume(-1)
genr tgdp = gdp - 0.357121*gdp(-1)
```

변환된 자료를 이용하여 소비함수를 OLS로 추정((6.33)식의 추정)한 결과가 <그림 6-A5>에 나타나 있는데 이 추정치는 변환된 자료의 OLS 추정치 또는 원래 자료의 GLS 추정치이다. 이를 <그림 6-A4>와 비교해 보면 한계소비성향이 동일함을 알 수 있다.

<그림 6-A5> 변환자료를 이용한 OLS 추정결과



연습문제

1. 다음의 모형을 각각 OLS로 추정하여 다음 표의 결과를 얻었다.

$$Y_{at} = \beta_0 + \beta_1 X_{2t} + \beta_2 X_{3t} + u_t \quad (1)$$

$$Y_{bt} = \beta_0 + \beta_1 X_{2t} + \beta_2 X_{3t} + u_t \quad (2)$$

	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	DW	n
(1)식	8.42	0.02	0.13	1.02	20
(2)식	-5.77	0.02	0.15	1.27	20

① 각 식에 대해 자기상관 여부에 대한 통계적 검정을 실시하라.

② ①식을 Cochrane-Orcutt의 반복적인 절차에 의해 추정한 결과가 다음과 같이 주어져 있을 때 Cochrane-Orcutt transformation에 의해 2단계로 추정할 경우의 회귀계수 값과 기타 값들을 해당되는 칸에 써 넣어라.

	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	DW	n
반복	15.239	0.0256	0.1379	1.81	15
2단계					

단, $\hat{\rho}$= 0.4668523

2. 회귀모형 에서 교란항에 대한 가정이 $E(UU') = \sigma_u^2 \Omega \neq \sigma_u^2 I_n$일 때 $L'L = \Omega^{-1}$을 만족시키는 L을 찾아 회귀모형을 $LY_t = LX_t B + LU_t$로 변환시키면 Lu의 분산은 $\sigma_u^2 I_n$가 되고 변환된 모형을 보통최소자승법(OLS)으로 추정하면 추정된 회귀계수는 $\widehat{B} = (X'X)^{-1}X'Y$가 된다. 이의 진위를 평가하라.

3. 회귀모형 $Y_t = X_t B + u_t$에서 교란항이 $u_t = \rho u_{t-1} + e_t$의 1차 자기상관이 있을

경우 u의 분산-공분산행렬인 $E(UU')$을 구하면 다음과 같다. 이의 진위를 평가하라.

$$E(UU') = \sigma_e^2 \begin{bmatrix} 1 & \rho & \rho^2 \cdot\cdot \rho^{n-1} \\ \rho & 1 & \rho \cdot\cdot \rho^{n-2} \\ \cdot & \cdot & \cdot\ \cdot\ \cdot\ \cdot \\ \cdot & \cdot & \cdot\ \cdot\ \cdot\ \cdot \\ \rho^{n-1} & \rho^{n-2} & \cdot\ \cdot\ \cdot\ 1 \end{bmatrix}$$

제 7 장

이분산

1.이분산 개념 및 유형
2.이분산 검정
3.이분산 해결책
실습 : 이분산 추정

제7장 이분산

1.이분산 개념 및 유형

제4장에서 살펴본 (7.1)식의 행렬표현 다중회귀모형에서 기본적인 가정들이 성립하는 고전적 회귀모형의 경우 보통최소자승법으로 추정한 추정량은 최량선형불편추정량(BLUE)이 됨을 보인 바 있다.

$$Y = X\ B + U$$
nx1 nxk kx1 nx1

$$\text{단, } Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix},\ X = \begin{bmatrix} 1\ X_{21}\ X_{31} \cdots\ X_{k1} \\ 1\ X_{22}\ X_{32} \cdots\ X_{k2} \\ .\quad .\quad .\quad . \\ 1\ X_{2n}\ X_{3n} \cdots\ X_{kn} \end{bmatrix},\ B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix},\ U = \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_n \end{bmatrix} \tag{7.1}$$

여러 가지 기본적인 가정들 중 (7.2)식의 가정은 교란항의 동분산(homoscedasticity)과 비자기상관(no autocorrelation)을 가정한 것이며 이 경우 OLS로 추정한 추정량은 BLUE가 된다.

$$E(UU') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) \cdot\cdot E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) \cdot\cdot E(u_2u_k) \\ . \quad . \quad \cdot\cdot \quad . \\ E(u_nu_1)\ E(u_nu_2) \cdot\cdot & E(u_n^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_u^2 & 0 \cdot\cdot 0 \\ 0 & \sigma_u^2 \cdot\cdot 0 \\ . \quad . \quad \cdot\cdot \quad . \\ 0 & 0 \cdot\cdot \sigma_u^2 \end{bmatrix}$$

$$= \sigma_u^2 I_n \tag{7.2}$$

그러나 현실적으로 (7.2)식의 가정이 성립하지 않는 경우가 발생한다. 즉, $E(UU') \neq \sigma_u^2 I_n$인 두 경우가 있는데 하나는 제6장의 (6.3)과 같은 자기상관(1차 자기상관)이 있는 경우이고 다른 하나는 (7.3)식과 같은 이분산(heteroscedasticity)인 경우이다.

$$E(UU') = \sigma_u^2 \Omega = \sigma_u^2 \begin{bmatrix} k_1^2 & 0 & . & . & 0 \\ 0 & k_2^2 & . & . & 0 \\ . & . & . & . & . \\ 0 & 0 & . & . & k_n^2 \end{bmatrix} \neq \sigma_u^2 I_n \qquad (7.3)$$

또는

$$E(UU') = E(u_i^2) = \sigma_{u_i}^2 \qquad (7.4)$$

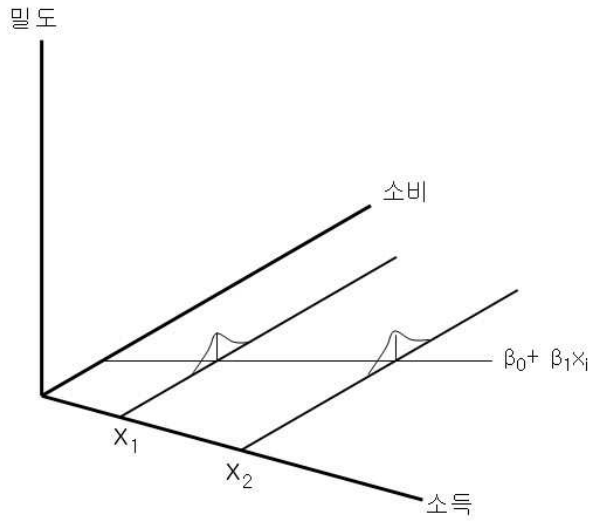
고전적 회귀모형에서 교란항에 대한 동분산 가정이 성립하지 않을 경우 OLS추정량은 편의는 없으나 효율적인 추정량이 되지 못한다. 즉, BLUE가 되지 못한다는 것을 제6장에서 살펴본 바 있다.

<그림 7-1>에서 보여주고 있듯이 교란항에 대한 동분산의 가정은 X의 모든 관측치에 대해 교란항의 확률분포가 일정하다는 것을 의미한다. 반면에 이분산이 있다는 것은 <그림 7-2>에서 나타내고 있는 것과 같이 X의 모든 관측치에 대해 교란항의 확률분포가 일정하지 않다는 것을 의미한다. 즉, X가 증가함에 따라 교란항의 분산이 커질 수도 있고 X가 증가함에 따라 교란항의 분산이 작아질 수도 있다. 일반적으로 거시자료보다 미시자료에, 시계열 자료보다 횡단면 자료에 이분산이 존재할 가능성이 크다. 왜냐하면 횡단면 자료는 다양한 규모의 가계나 기업 등에 관한 자료가 포함되어 있기 때문이다. 그러나 시계열 자료라고 해서 항상 동분산만 있는 것은 아니며 한 경제단위의 장기에 걸친 시계열 자료의 경우 이분산이 나타날 가능성이 있다.

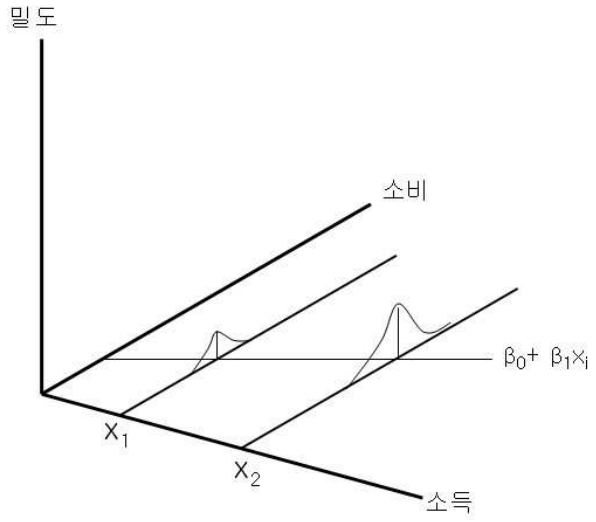
소득수준에 따른 소비수준의 변화를 예를 들면 동분산 가정의 경우 소득수준이 낮거나 높거나 관계없이 소비지출의 기복이 동일하다는 것이다. 그러나 이분산 가정의 경우 소득수준이 낮으면 소비지출에 기복이 작고 단순하여 교란항의 분포가 좁아져 분산이 작으나, 소득수준이 높으면 소비지출이 다양해져 교란항의 분포가 퍼져서 분산이 크게 된다는 것이다. 현실에서는 후자의 경우가 빈번히 발생하고 있으므로 교란항에 대한 동분산 가정보다는 이분산 가정이 더 현실과 부합

한다고 할 수 있다.

<그림 7-1> 교란항의 분포(동분산)



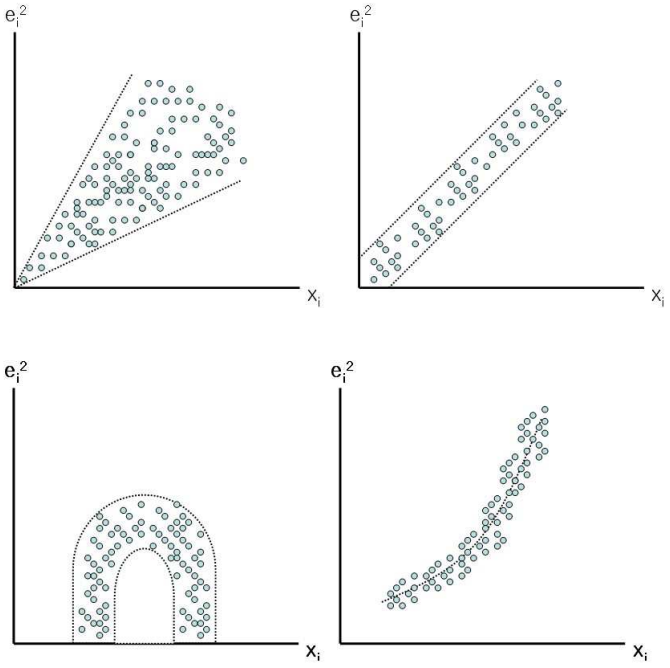
<그림 7-2> 교란항의 분포(이분산)



이분산 여부를 탐지하는 방법으로는 자기상관과 마찬가지로 회귀식으로부터 도출된 잔차를 시계열을 그려 이분산 여부를 판단하는 방법인 잔차의 그래프 분석(residual plotting)과 통계적 검정을 통해 자기상관 여부를 판단하는 방법이 있다.

잔차의 그래프 분석은 잔차의 제곱(잔차의 평균이 0이므로 잔차의 제곱은 교란항의 분산이 됨)과 종속변수의 추정치 $\hat{Y}$ 또는 독립변수 X_i를 그려 보고 <그림 7-3>과 같이 이들 사이에 일정한 관계가 발견되면 이분산이 있는 것으로 판단한다.

<그림 7-3> 이분산의 유형



2.이분산 검정

이분산이 있을 경우 일반적으로 교란항 분산의 크기는 설명변수 또는 종속변수

와 관련이 있다. 이러한 측면을 고려한 이분산 검정방법은 White 검정방법, Goldfeld-quant 검정방법, Glejser 검정방법 등 여러 가지 방법이 있다.

(1) White 검정

White 검정방법은 교란항의 분산과 모형에 포함된 설명변수 뿐만 아니라 이 변수들의 2차항 사이에 상관관계가 나타나는 지를 검정하는 방법이다. 예를 들어, 다음의 (7.5)식에 대한 White 검정은 다음과 같은 순서로 한다.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (7.5)$$

먼저, (7.5)식의 모형을 추정한 후 다음의 (7.6)식과 같이 잔차를 구한다.

$$\hat{u_i} = Y_i - \hat{\beta_1} - \hat{\beta_2} X_{2i} - \hat{\beta_3} X_{3i} \qquad (7.6)$$

다음으로, 잔차의 제곱을 이용하여 (7.7)식의 회귀모형을 추정하는데 이를 보조회귀식(auxiliary regression)이라고 한다.

$$\hat{u_i^2} = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i \qquad (7.7)$$

(7.7)의 보조회귀식에서 검정하고자 하는 귀무가설은 $H_0 : \alpha_2 = ... = \alpha_6 = 0$ 즉, 이분산이 없다는 것이며 귀무가설 하에서 다음의 LM 검정통계량과 그 분포를 구할 수 있다.

$$LM = nR^2 \sim \chi_q^2$$

단, n은 관측치의 수, R^2는 보조회귀식에서의 결정계수이며 χ^2-분포의 자유도는 q 즉, 제약식의 수이다.

(2) Goldfeld-Quant 검정

Goldfeld-Quant 검정방법은 교란항은 정규분포를 하고 자기상관이 없으며 교란항의 분산이 모형 내 독립변수의 값에 비례한다고 가정하는 경우에 적합한 검정

방법이다. 예를 들어 이분산이 다음 (7.8)식과 같은 관계에 있다고 하자.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\sigma_i^2 = \sigma^2 X_i^2 \qquad (7.8)$$

Goldfeld-Quant 검정방법은 다음과 같다.

1단계는 n개의 관측치를 독립변수의 크기에 따라 순서를 정한다.

2단계는 n개의 나열된 관측치 중 $\frac{n}{4}$정도의 가운데 관측치를 제외한다.

3단계는 값이 작은 쪽의 관측치(소분산집단)로 회귀식을 OLS로 추정하여 $\sum e_1^2$를 구하고 값이 큰 쪽의 관측치(대분산집단)로 회귀식을 OLS로 추정하여 $\sum e_2^2$를 구한다.

4단계는 다음의 통계량으로 F-검정을 실시한다.

$$F = \frac{\sum e_2^2}{\sum e_1^2} \sim F_{v_1, v_2}$$

단, $v_1 = v_2 = \frac{(n-c)}{2} - K$의 자유도, $c = \frac{n}{4}$(제외된 관측치), K=추정하는 모수이다.

(3)Glejser 검정방법

Glejser 검정방법은 교란항은 이분산과 밀접하게 관련된 것으로 생각되는 독립변수와 다음의 관계에 있다고 가정한다.

$$\sigma_i = \alpha + \beta X_i^r \qquad (7.9)$$

Glejser 검정방법은 다음과 같다.

1단계는 원래 모형을 OLS로 추정하여 잔차($\hat{u}$)를 구한다.

2단계는 잔차($\hat{u}$)의 절대치를 교란항의 분산과 가장 밀접한 관계에 있는 설명변

수에 대해 다음의 여러 형태를 추정한다.

$$| \hat{u} | = \alpha_0 + \alpha_1 X_i$$

$$| \hat{u} | = \alpha_0 + \alpha_1 X_i^{-1}$$

$$| \hat{u} | = \alpha_0 + \alpha_1 X_i^{1/2}$$

3단계는 위에서 가장 적합한 모형을 찾아(R^2나 F값) α_1이 0과 현저히 다르면 이분산이 있다고 본다.

3.이분산 해결책

이분산 문제를 해결하는 방법은 이분산의 구조에 따라 달라진다. 예를 들어, (7.5) 식에서 교란항의 분산이 다음과 같다고 하자.

$$E(u_i^2) = \sigma_i^2$$

첫째, σ_i^2의 값을 알고 있을 경우로써 (7.10)식과 같이 원래의 회귀식을 변환하여 이분산이 없는 모형으로 도출한다.

$$(\frac{Y_i}{\sigma_i}) = \beta_1 (\frac{1}{\sigma_i}) + \beta_2 (\frac{X_{2i}}{\sigma_i}) + \beta_3 (\frac{X_{3i}}{\sigma_i}) + (\frac{u_i}{\sigma_i}) \qquad (7.10)$$

또는

$$Y_i^* = \beta_1 (\frac{1}{\sigma_i}) + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + u_i^*$$

이와 같이 변환된 회귀모형(transformed regression model)에서 Y_i^*를 $(\frac{1}{\sigma_i})$와 X_{2i}^*, X_{3i}^*에 대해 OLS로 추정하면 BLUE를 얻을 수 있다. 이러한 추정방법은 일반최

소자승법(GLS)의 하나인데 이를 특별히 가중최소자승법(Weighted Least Squares: WLS)이라고 한다.

둘째, σ_i^2의 값을 모를 경우로써 σ_i^2의 추정치를 구하고 이를 가중치로 사용하면 된다. 즉, 먼저 (7.5)식 및 (7.7)식을 추정한 후 추정된 회귀계수를 이용하여 다음의 (7.11)식과 같이 잔차제곱에 대한 추정치(fitted value)를 구한다.

$$\widehat{\sigma_i^2} = \widehat{\alpha_1} + \widehat{\alpha_2} X_{2i} + \widehat{\alpha_3} X_{3i} + \widehat{\alpha_4} X_{2i}^2 + \widehat{\alpha_5} X_{3i}^2 + \widehat{\alpha_6} X_{2i} X_{3i} \qquad (7.11)$$

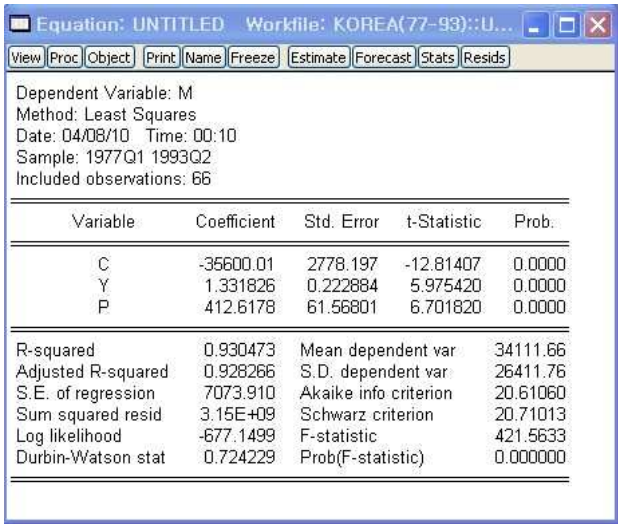
다음으로 $\widehat{\sigma_i}$ 를 (7.10)식의 σ_i 대신에 사용하는데 이러한 추정방법을 실행 가능한 일반최소자승법(Feasible Generalized Least Squares: FGLS)이라고 한다.

FGLS 추정치는 분산을 추정해야 하기 때문에 일반적으로 불편추정량은 되지 못하나 일치추정량이 되며 점근적으로 OLS 추정량보다 효율적인 추정량이 된다.

실습 : 이분산 추정

1977년 1/4분기부터 1997년 2/4분기까지의 총통화(M), GDP(Y), CPI(P)의 자료를 이용하여 총통화는 GDP와 CPI에 영향을 받는다는 다중회귀모형을 추정한 결과가 <그림 7-A1>에 나타나 있다.

<그림 7-A1> 다중회귀모형


Equation: UNTITLED Workfile: KOREA(77-93)::U...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: M
Method: Least Squares
Date: 04/08/10 Time: 00:10
Sample: 1977Q1 1993Q2
Included observations: 66

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-35600.01	2778.197	-12.81407	0.0000
Y	1.331826	0.222884	5.975420	0.0000
P	412.6178	61.56801	6.701820	0.0000

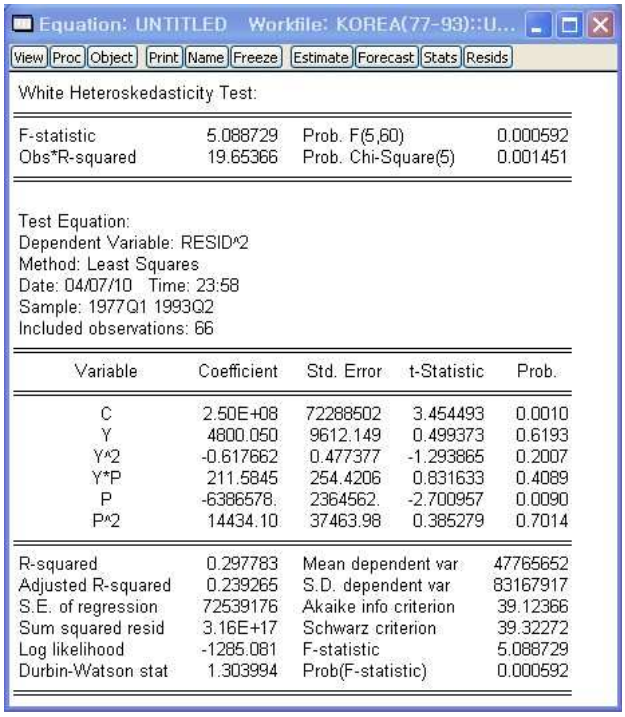
R-squared	0.930473	Mean dependent var	34111.66
Adjusted R-squared	0.928266	S.D. dependent var	26411.76
S.E. of regression	7073.910	Akaike info criterion	20.61060
Sum squared resid	3.15E+09	Schwarz criterion	20.71013
Log likelihood	-677.1499	F-statistic	421.5633
Durbin-Watson stat	0.724229	Prob(F-statistic)	0.000000

White의 검정방법을 이용하여 이분산 여부를 판단하기 위해서는 <그림 7-A1>의 추정 결과 창에서 View-Residual Tests-White Heteroscedasticity를 차례대로 클릭한다. 그리고 이 때 (7.7)식과 같이 교란항의 분산에 대한 보조회귀식의 모양을 지정해 주어야 하는데 (7.7)식의 보조회귀식 형태를 그대로 이용하기 위해서는 (cross term)을 이용하고, (7.7)식의 보조회귀식에서 $X_{2i}X_{3i}$항이 빠진 형태를 이용하기 위해서는 (no cross term)을 이용한다.

(cross term)을 이용한 White의 검정결과가 <그림 7-A2>에 나타나 있는데 여기서 Obs*R-squared 19.65366이 $LM = nR^2$ 값이며 Prob. Chi-Square(5) 0.001451은 χ_5^2의 유의확률이 0.001451이라는 의미이므로 5% 유의수준에서 이분산이 없다

는 귀무가설을 기각하여 이분산이 있는 것으로 결정한다.

<그림 7-A2> White 검정 결과



Equation: UNTITLED Workfile: KOREA(77-93)::U...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

White Heteroskedasticity Test:

F-statistic	5.088729	Prob. F(5,60)	0.000592
Obs*R-squared	19.65366	Prob. Chi-Square(5)	0.001451

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/07/10 Time: 23:58
Sample: 1977Q1 1993Q2
Included observations: 66

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.50E+08	72288502	3.454493	0.0010
Y	4800.050	9612.149	0.499373	0.6193
Y^2	-0.617662	0.477377	-1.293865	0.2007
Y*P	211.5845	254.4206	0.831633	0.4089
P	-6386578.	2364562.	-2.700957	0.0090
P^2	14434.10	37463.98	0.385279	0.7014

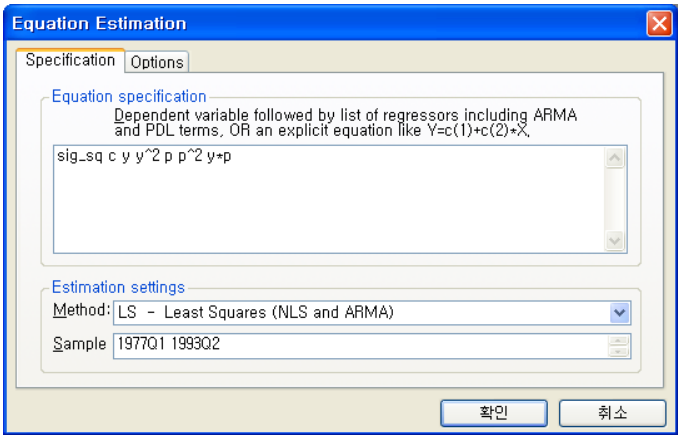
R-squared	0.297783	Mean dependent var	47765652
Adjusted R-squared	0.239265	S.D. dependent var	83167917
S.E. of regression	72539176	Akaike info criterion	39.12366
Sum squared resid	3.16E+17	Schwarz criterion	39.32272
Log likelihood	-1285.081	F-statistic	5.088729
Durbin-Watson stat	1.303994	Prob(F-statistic)	0.000592

이분산을 고려한 FGLS로 추정치를 구하기 위해서는 σ_i^2에 대한 추정치를 먼저 구해야 한다. 이를 위해 <그림 7-A1>의 다중회귀모형을 OLS로 추정한 후 잔차제곱(e_i^2)을 도출하기 위해 다음과 같이 입력한다.

```
genr sig_sq = resid^2
```

이 잔차제곱을 종속변수로 하는 보조회귀식을 추정하기 위하여 <그림 7-A3>과 같이 보조회귀식 추정식을 방정식 추정 대화상자에 입력하여 확인을 클릭하면 보조회귀식이 추정되고 또한 보조회귀식의 잔차가 만들어진다.

<그림 7-A3> 보조회귀식 추정식

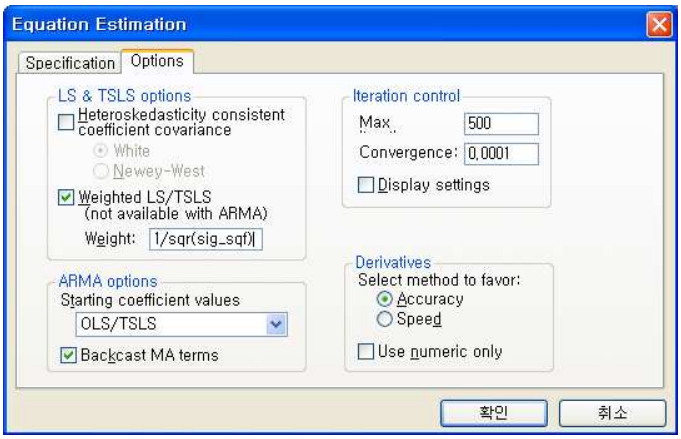


보조회귀식의 종속변수 추정치(sig_sqf)를 도출하기 위하여 다음을 입력한다.

```
genr sig_sqf = sig_sq-resid
```

<그림 7-A1>의 다중회귀모형을 추정하기 위하여 방정식 추정 대화상자의 Specification에 m c y p를 입력한 후 Options를 클릭하면 선택항목들이 있는데 거기서 <그림 7-A4>와 같이 입력한다.

<그림 7-A4> WLS(FGLS) 추정 옵션



<그림 7-A4>에서 Weighted LS/TSLS 밑에 있는 Weight 1/sqr(sig_sqf)는 자료 변환 시 $\frac{1}{\hat{\sigma}_i}$의 가중치를 사용하라는 의미이다.

WLS(FGLS) 추정 결과가 <그림 7-A5>에 나타나 있는데 이것이 WLS 추정치 또는 FGLS 추정치이다. 이 추정치를 <그림 7-A1>에 있는 OLS 추정치와 비교해 보면 다름을 알 수 있고 WLS 추정치의 표준오차가 OLS 추정치의 표준오차보다 작아 WLS 추정치가 효율적임을 보여주고 있다.

<그림 7-A4> WLS(FGLS) 추정 결과

Equation: UNTITLED Workfile: KOREA(77-93)::U...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: M
Method: Least Squares
Date: 04/08/10 Time: 00:08
Sample: 1977Q1 1993Q2
Included observations: 61
Weighting series: 1/SQR(SIG_SQF)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-34206.14	2844.585	-12.02500	0.0000
Y	1.250028	0.146945	8.506753	0.0000
P	384.5234	48.90808	7.862167	0.0000

Weighted Statistics

R-squared	0.926719	Mean dependent var	28758.10
Adjusted R-squared	0.924192	S.D. dependent var	15565.32
S.E. of regression	4931.886	Akaike info criterion	19.89276
Sum squared resid	1.41E+09	Schwarz criterion	19.99657
Log likelihood	-603.7292	F-statistic	366.7370
Durbin-Watson stat	0.848458	Prob(F-statistic)	0.000000

Unweighted Statistics

R-squared	0.910315	Mean dependent var	34454.60
Adjusted R-squared	0.907222	S.D. dependent var	27011.48
S.E. of regression	8227.545	Sum squared resid	3.93E+09
Durbin-Watson stat	0.421941		

연습문제

1. 회귀모형 $Y_t = \alpha + \beta X_t + u_t$에서 교란항의 분산이 $E(u_t^2) = \sigma_t^2 = \sigma^2 X_t$라고 가정하자. 이 회귀모형을 최소자승법(OLS)으로 추정하면 BLUE인 추정량을 얻지 못하므로 원자료를 X_t로 나누어 변환한 후 변환된 자료를 최소자승법으로 추정하면 되는데 이렇게 추정하는 방법을 가중최소자승법(WLS)이라 한다. 이를 평가하라.

제 8 장

시차분포모형

1.시차분포모형

2.시차분포모형 추정방법

3.인과성 검정

실습 : 시차분포모형 추정 및 인과성 검정

제8장 시차분포모형

1.시차분포모형

상수항과 k-1개의 독립변수에 의해 설명되는 (8.1)식과 같은 다중회귀모형이 있다고 하자.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i \ (i = 1,2,..,n) \tag{8.1}$$

이 모형에서는 모든 설명변수들의 현재 값이 종속변수의 현재 값에 영향을 주고 있는데 이러한 모형을 정태적 모형(static model)이라고 한다.

한편, (8.2)식과 같이 다중회귀모형에서 설명변수의 현재 값뿐만 아니라 과거(시차)값들이 종속변수의 현재 값에 영향을 줄 수도 있는데 이러한 모형을 동태적 모형(dynamic model)이라고 하며 특히, 이러한 동태모형을 시차분포모형(lag distributed model)이라고 한다.

$$y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_k X_{t-k} + u_t \tag{8.2}$$

또한 동태모형의 다른 형태로서 다중회귀모형에서 (8.3)식과 같이 종속변수의 하나 이상의 과거 값들이 설명변수로서 포함된 모형이 있는데 이를 자기회귀모형(autoregressive model) 또는 시계열모형(time series model)이라고 한다.

$$y_t = \alpha + \gamma_1 y_{t-1} + ... + \gamma_k y_{t-k} + u_t \tag{8.3}$$

(8.2)식의 시차분포모형에서 교란항에 대한 가정은 다중회귀모형에서 교란항에 대한 기본가정과 동일하며, 시차분포모형은 한 시점의 경제현상은 그 이전의 경제현상으로부터 영향을 받는 것으로 보고 있으며 이 때 시차의 수와 유형은 경험적으로 결정된다.

시차분포모형은 <표 8-1>과 같은 설명변수의 현재 값 및 과거 값들을 설명변수로 포함하고 있기 때문에 설명변수 간에 상관관계가 존재하는 다중공선성(multicollinearity)의 문제가 발생할 수 있다.

<표 8-1> 시차 값의 예시

관측치	X_t	X_{t-1}	X_{t-2}	X_{t-3}	.	.	X_{t-k}
1	X1	-	-	-	-	-	-
2	X2	X1	-	-	-	-	-
3	X3	X2	X1	-	-	-	-
4	X4	X3	X2	X1	-	-	-
.	.	.	.	.	.	.	-
k	Xk	.	.	.	.	.	Xk
.	.	Xk	.	.	.	.	.
	.	.	Xk	.	.	.	.
.	.	.	.	Xk	.	.	.
.	.	.	.	.	.	.	.
n	Xn	Xn-1	Xn-2	Xn-3	.	.	Xn-k

설명변수의 시차변수를 모형에 도입하면 독립변수(외생변수)의 변화가 종속변수(내생변수)에 미치는 영향(승수)을 시간의 흐름에 따라 파악할 수 있다. 따라서 시차분포모형에서 추정된 시차계수(lag coefficients)로부터 여러 가지 승수(multiplier)를 계산할 수 있다.

$$y_t = \alpha + 0.4X_t + 0.3X_{t-1} + 0.2X_{t-2} + u_t \tag{8.4}$$

예를 들어 시차분포모형이 (8.4)식과 같다고 할 때 $\frac{\partial y_t}{\partial X_t} = \beta_0 = 0.4$를 단기승수(short-run multiplier)라고 한다. 따라서 (8.4)식이 y_t는 소비, X_t는 소득인 소비함수라면 이 값은 단기 한계소비성향(Marginal Propensity to Consume: MPC)이 된다.

한편, 다음의 (8.5)식을 장기승수(long-run multiplier)라고 하고 소비함수에서 이 값은 장기 한계소비성향(MPC)이 된다.

$$\frac{\partial y_t}{\partial X_t} + \frac{\partial y_t}{\partial X_{t-1}} + \frac{\partial y_t}{\partial X_{t-2}} = \sum_{i=0}^{2} \beta_i = 0.4 + 0.3 + 0.2 = 0.9 \tag{8.5}$$

2.시차분포모형 추정방법

시차분포모형을 보통최소자승법(OLS)으로 추정할 경우 잘못된 통계적 추론을 가져올 수 있다. 왜냐하면 시차변수를 설명변수로 포함시키면 자유도가 작아지고(자유도는 n-k이므로), 자유도가 작아지면 귀무가설을 기각할 영역이 작아지므로 통계적으로 유의한 변수도 유의하지 않은 것으로 결론을 내릴 수 있게 된다. 따라서 시차분포모형은 OLS로 추정하지 않고 다른 방법으로 추정하는데 대표적인 추정방법으로는 Koyck과 Almon의 추정방법이 있다.

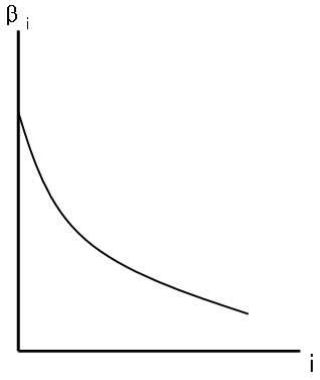
(1)Koyck의 추정방법

Koyck의 추정방법은 시차의 수를 가능한 줄이려는 의도에서 출발하였는데 (8.6)식과 같이 먼 과거로 거슬러 올라 갈수록 y에 대한 시차의 영향이 점점 작아진다는 가정을 도입하고 있으며 내생시차변수모형(endogenous lagged variable model)이라고 불린다.

$$\beta_k = \beta_0 \lambda^k, \quad k = 0,1,2,3,... \tag{8.6}$$

단, $0 < \lambda < 1$

<그림 8-2> 시차계수 구조(Almon)



(8.6)식의 가정을 (8.2)식의 시차분포모형에 대입하면 (8.7)식과 같다.

$$y_t = \alpha + \beta_0 X_t + (\beta_0\lambda)X_{t-1} + \ldots + (\beta_0\lambda^{k-1})X_{t-k+1} + (\beta_0\lambda^k)X_{t-k} + u_t \qquad (8.7)$$

(8.7)식의 모형에서 시차를 하나 뒤로 하면 (8.8)식과 같게 된다.

$$y_{t-1} = \alpha + \beta_0 X_{t-1} + (\beta_0\lambda)X_{t-2} + \ldots + (\beta_0\lambda^{k-1})X_{t-k} + u_{t-1} \qquad (8.8)$$

(8.8)의 모형에 λ를 곱하면 (8.9)식이 된다.

$$\lambda y_{t-1} = \lambda\alpha + (\beta_0\lambda)X_{t-1} + (\beta_0\lambda^2)X_{t-2} + \ldots + (\beta_0\lambda^k)X_{t-k} + \lambda u_{t-1} \qquad (8.9)$$

(8.7)식에서 (8.9)식을 빼면 (8.10)식과 같게 되고 이를 다시 정리하면 (8.11)식이 된다.

$$y_t - \lambda y_{t-1} = \alpha(1-\lambda) + \beta_0 X_t + u_t - \lambda u_{t-1} \qquad (8.10)$$

$$y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda y_{t-1} + v_t \quad (\text{단}, v_t = u_t - \lambda u_{t-1}) \qquad (8.11)$$

(8.11)식은 다중공선성 문제가 없으므로 OLS로 추정하면 $\hat{\alpha}, \hat{\beta_0}, \hat{\lambda}$을 구할 수 있으나(따라서 $\hat{\beta_k} = \hat{\beta_0}\hat{\lambda^k}$를 이용하여 $\hat{\beta_k}$구해지고) OLS추정량은 불편추정량 뿐만 아니라 효율적인 추정량도 되지 못한다. 그 이유는 (8.11)식의 교란항에 자기상관이 있고 또한 직교조건도 성립되지 않기 때문인데 이를 살펴보면 다음과 같다.

먼저 (8.2)식의 시차분포모형에서 u_t가 자기상관이 없다고 가정하더라도 (8.11)식의 v_t가 (8.12)식에서 보여주듯이 자기상관이 있어 OLS추정량은 효율적인 추정량이 되지 못한다.

$$\begin{aligned} E(v_t, v_{t-1}) &= E[(u_t - \lambda u_{t-1})(u_{t-1} - \lambda u_{t-2})] \\ &= -\lambda E(u_{t-1}^2) \\ &= -\lambda\sigma_u^2 \\ &\neq 0 \end{aligned} \qquad (8.12)$$

다음으로 (8.13)식에 보여주고 있듯이 V_t와 Y_{t-1}가 서로 독립이 아니므로(즉,

직교조건이 성립하지 않으므로) OLS추정량은 불편추정량이 되지 못한다.

$$\begin{aligned} E(v_t, y_{t-1}) &= E[(u_t - \lambda u_{t-1})(y_{t-1})] \\ &= -\lambda E(u_{t-1}^2) \\ &= -\lambda \sigma_u^2 \\ &\neq 0 \end{aligned} \tag{8.13}$$

이에 대한 해결책으로는 수단변수(instrumental variable: IV) 추정방법이 있다. (8.11)식의 교란항 v_t와 독립이고 설명변수 y_{t-1}과 상관관계가 있는 변수(이를 수단변수라 한다)를 이용한다. 여기서는 X_{t-1}이 수단변수로 이용될 수 있는데 그 이유는 X_{t-1}이 이러한 조건을 만족시키는 좋은 변수이기 때문이다. 따라서 (8.11)식에서 y_{t-1} 대신에 X_{t-1}을 사용하여 OLS로 추정한다.

이 경우 직교조건은 해결되어 불편(일치)추정량은 얻을 수 있으나 X_t와 X_{t-1} 사이에 다중공선성 문제가 발생할 가능성이 있으며 v_t가 자기상관이 되어 있으므로 효율적인 추정량을 얻지 못한다.

(2)Almon의 추정방법

Almon의 추정방법은 외생시차변수(exogenous lagged variable)의 모수를 추정한 후 시차모형의 모수를 계산하는 방법으로 다항식 시차(polynomial distributed lag: PDL) 모형이라고도 한다. 이 방법은 시차계수의 크기가 시차에 따라 2차 함수 또는 3차 함수 등 다항식의 구조를 가진다고 가정한다. 즉, 시차계수(β_i)가 다음의 (8.14)식과 같이 시차의 길이인 i의 적절한 차수의 다항식의 근사치로 계산될 수 있다고 가정한다.

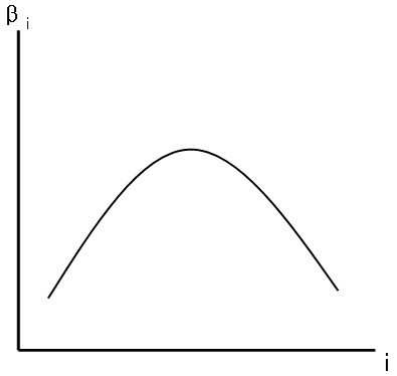
$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + \ldots + \alpha_k i^r \tag{8.14}$$

(8.14)식에서 r=2이면 2차 함수, r=3이면 3차 함수 등이 된다. 예를 들어 r=2인 경우 다음의 (8.15)식과 같은 2차 다항식이 된다.

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 \tag{8.15}$$

(8.15)식에 따른 시차구조를 그림으로 그려보면 <그림 8-2>와 같은데 이는 시차계수 β_i는 시차에 따라 <그림 8-2>와 같은 구조를 갖는다고 가정한 것과 같다

<그림 8-2> 시차계수 구조(Almon)



예를 들어 2차 다항식을 가정하는 경우 β는 α들과 다음의 관계에 있으므로 $\alpha_0, \alpha_1, \alpha_2$를 알면 β의 값들을 알 수 있다.

$$\beta_0 = \alpha_0$$

$$\beta_1 = \alpha_0 + \alpha_1 + \alpha_2$$

$$\beta_2 = \alpha_0 + 2\alpha_1 + 4\alpha_2$$

$$\vdots$$

$$\beta_k = \alpha_0 + k\alpha_1 + k^2\alpha_2$$

이를 (8.2)식에 대입하면 다음의 (8.16)식을 얻게 된다.

$$\begin{aligned} y_t &= \alpha + \alpha_0 X_t + (\alpha_0 + \alpha_1 + \alpha_2) X_{t-1} + \ldots + (\alpha_0 + k\alpha_1 + k^2\alpha_2) X_{t-k} + u_t \\ &= \alpha + \alpha_0 (X_t + X_{t-1} + \ldots + X_{t-k}) + \alpha_1 (X_{t-1} + 2X_{t-2} + \ldots + kX_{t-k}) \\ &\quad + \alpha_2 (X_{t-1} + 2^2 X_{t-2} + \ldots + k^2 X_{t-k}) + u_t \end{aligned}$$

$$= \alpha + \alpha_0 (\sum_{k=0}^{K} X_{t-k}) + \alpha_1 (\sum_{k=1}^{K} k X_{t-k}) + \alpha_2 (\sum_{k=1}^{K} k^2 X_{t-k}) + u_t$$
$$= \alpha + \alpha_0 Z_{1t} + \alpha_1 Z_{2t} + \alpha_2 Z_{3t} + u_t \quad (8.16)$$

단, $Z_{1t} \equiv \sum_{k=0}^{K} X_{t-k}$, $Z_{2t} \equiv \sum_{k=0}^{K} k X_{t-k}$, $Z_{3t} \equiv \sum_{k=0}^{K} k^2 X_{t-k}$이다.

(8.16)식을 OLS로 추정하여 $\hat{\alpha}, \hat{\alpha_0}, \hat{\alpha_1}, \hat{\alpha_2}$을 구하고 (8.15)식의 β와 α의 관계를 이용하여 $\hat{\beta_0}, \hat{\beta_1}, ..., \hat{\beta_k}$을 계산한다.

한편, Almon의 추정방법에서 다항식의 형태(r 의 수)와 시차수(k)의 결정은 여러 가지 조합을 시행하여 회귀계수의 t값과 모형의 R^2를 높이는 방향으로 한다.

3.인과성 검정

시차분포모형과 자기회귀모형을 결합하면 변수 간에 인과관계를 분석하는 인과성 검정(causality test)에 활용할 수 있다.

Granger는 인과성(causality)을 "만약 Y의 과거정보만을 가지고 Y를 예측할 때보다 X와 Y의 과거정보를 동시에 가지고 Y를 더욱 잘 예측할 수 있으면 X는 Y의 원인변수가 된다"라고 정의하였다.

예를 들어 생산 y_t와 통화량 m_t의 인과성 검정을 위해 (8.17) 및 (8.18)식을 고려해 보자.

$$y_t = c_1 + \sum_{i=1}^{k} \alpha_i m_{t-i} + \sum_{j=1}^{k} \beta_j y_{t-j} + u_{1t} \quad (8.17)$$

$$m_t = c_2 + \sum_{i=1}^{k} \gamma_i m_{t-i} + \sum_{j=1}^{k} \delta_j y_{t-j} + u_{2t} \quad (8.18)$$

(1)m에서 y로 일방적 인과관계

$H_0 : \alpha_i \neq 0 (f \forall i)$ and $\delta_j = 0 (f \forall j)$

위 귀무가설이 동시에 성립할 경우 즉, m은 y의 원인변수가 되고 y는 m의 원인

변수가 되지 못하므로 m에서 y로 일방적 인과관계(unidirectional causality)가 존재한다.

(2)y에서 m으로 일방적 인과관계

$$H_0 : \alpha_i = 0 (f \forall i) \quad \text{and} \quad \delta_j \neq 0 (f \forall j)$$

위 귀무가설이 동시에 성립할 경우 y는 m의 원인변수가 되고 m은 y의 원인변수가 되지 못하므로 y에서 m으로 일방적 인과관계(unidirectional causality)가 존재한다.

(3)m과 y의 쌍방적 인과관계

$$H_0 : \alpha_i \neq 0 (f \forall i) \quad \text{and} \quad \delta_j \neq 0 (f \forall j)$$

위 귀무가설이 성립할 경우 m은 y의 원인변수가 되고 y도 m의 원인변수가 되므로 m과 y는 쌍방적 인과관계(bilateral causality)가 존재한다.

(4)m과 y의 통계적 독립

$$H_0 : \alpha_i = 0 (f \forall i) \quad \text{and} \quad \delta_j = 0 (f \forall j)$$

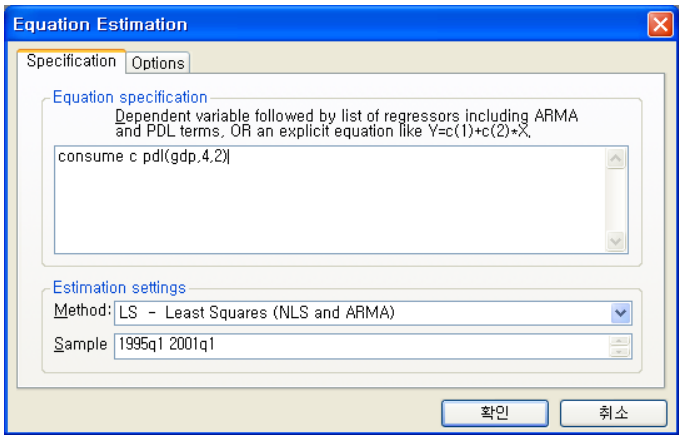
위 귀무가설이 성립할 경우 m은 y의 원인변수가 되지 못하고 y도 m의 원인변수가 되지 못하므로 m과 y는 통계적으로 독립(statistically independent)이다.

실습 : 시차분포모형 추정 및 인과성 검정

1995년 1/4분기부터 2001년 1/4분기까지 한국의 GDP와 총소비(consume) 자료를 이용하여 소비함수를 추정하고자 한다. 과거 4분기 동안의 소득이 현재의 소비에 영향을 주고, 시차계수의 구조가 2차 다항식이라고 가정하자.

이러한 시차분포모형을 Almon의 방법으로 추정하기 위하여 방정식 추정 대화상자에 <그림 8-A1>이 입력한다.

<그림 8-A1> 2차 다항식 시차분포모형 추정식



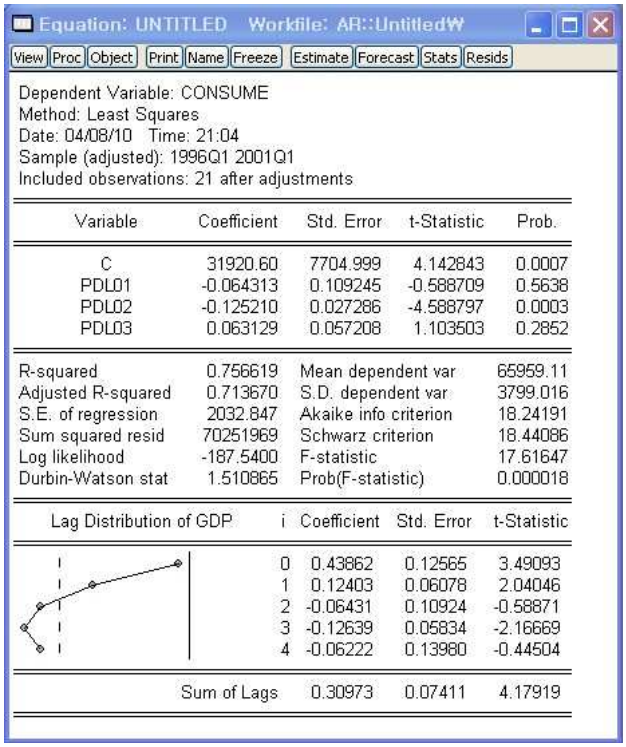
2차 다항식 시차분포모형의 추정결과가 <그림 8-A2>에 나타나 있는데 시차분포모형의 추정회귀식은 다음과 같다.

$$\hat{y_t} = 31920.6 + 0.438X_t + 0.124X_{t-1} - 0.064X_{t-2} - 0.126X_{t-3} - 0.062X_{t-4}$$

추정결과에 따르면 소득 수준의 변화는 1분기 후까지 소비의 움직임에 유의적인 영향을 미치다가 시간의 흐름에 따라 감소하는 것으로 나타났다.

한편, 단기 한계소비성향은 0.43862이고, 장기 한계소비성향은 $\sum_{i=0}^{4}\beta_i = 0.43862 + 0.12403 - 0.06431 - 0.12639 - 0.06222 = 0.3097$이다.

<그림 8-A2> 2차 다항식 시차분포모형 추정 결과



Equation: UNTITLED Workfile: AR::Untitled\W

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUME
Method: Least Squares
Date: 04/08/10 Time: 21:04
Sample (adjusted): 1996Q1 2001Q1
Included observations: 21 after adjustments

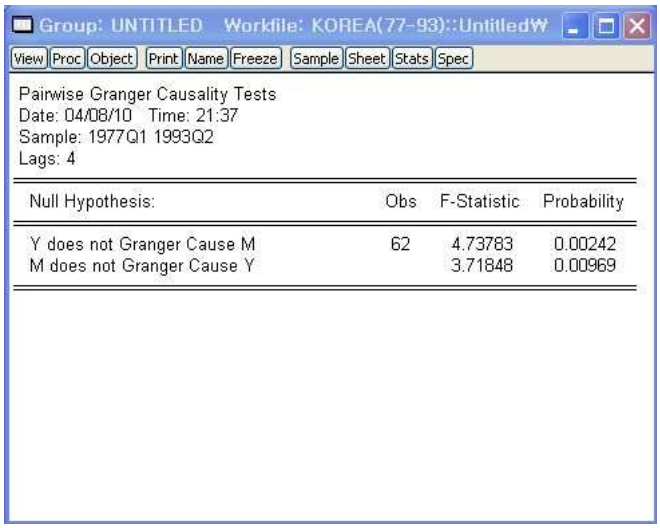
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	31920.60	7704.999	4.142843	0.0007
PDL01	-0.064313	0.109245	-0.588709	0.5638
PDL02	-0.125210	0.027286	-4.588797	0.0003
PDL03	0.063129	0.057208	1.103503	0.2852

R-squared	0.756619	Mean dependent var	65959.11
Adjusted R-squared	0.713670	S.D. dependent var	3799.016
S.E. of regression	2032.847	Akaike info criterion	18.24191
Sum squared resid	70251969	Schwarz criterion	18.44086
Log likelihood	-187.5400	F-statistic	17.61647
Durbin-Watson stat	1.510865	Prob(F-statistic)	0.000018

Lag Distribution of GDP	i	Coefficient	Std. Error	t-Statistic
	0	0.43862	0.12565	3.49093
	1	0.12403	0.06078	2.04046
	2	-0.06431	0.10924	-0.58871
	3	-0.12639	0.05834	-2.16669
	4	-0.06222	0.13980	-0.44504
	Sum of Lags	0.30973	0.07411	4.17919

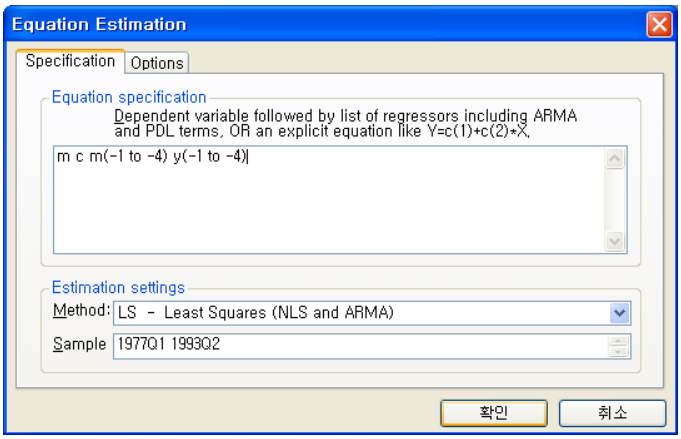
1977년 1/4분기부터 1997년 2/4분기까지의 총통화(M), GDP(Y)의 자료를 이용하여 Granger 인과성 검정을 해 보도록 하자. 먼저 M과 Y를 선택한 후 Open Group으로 두 변수를 동시에 불러 온다. View-Granger Causality를 클릭하면 나타나는 대화 상자에서 시차 수(lags to include)로 4를 입력하면 <그림 8-A3>과 같은 Granger 인과성 검정 결과를 준다. Y does not Granger cause M의 F-통계량이 4.73783이고 유의확률이 0.00242이므로 Y가 M의 원인변수가 되지 못한다는 귀무가설을 기각한다. M does not Granger cause Y의 F-통계량이 3.71848이고 유의확률이 0.00969이므로 M이 Y의 원인변수가 되지 못한다는 귀무가설을 기각한다. 따라서 M이 Y의 원인변수가 되고 Y도 M의 원인변수가 되므로 M과 Y는 쌍방적 인과관계(bilateral causality)가 존재한다고 할 수 있다.

<그림 8-A3> Granger 인과성 검정 결과



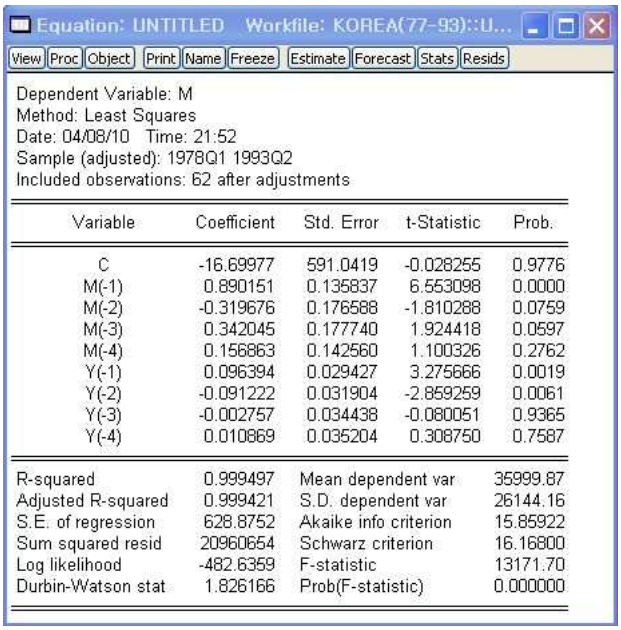
한편, 이러한 인과성 검정은 제4장의 가설점정에서 살펴보았던 결합검정을 통해서도 동일한 결론을 얻을 수 있다. 먼저 종속변수가 M_t이고 독립변수가 상수항, $M_{t-1}, M_{t-2}, M_{t-3}, M_{t-4}$, $Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}$인 총통화 회귀식을 추정하기 위하여 방정식 추정 대화상자에서 <그림 8-A4>와 같이 입력한다.

<그림 8-A4> 총통화 추정식(제약이 없는 모형)



총통화 회귀식의 추정 결과가 <그림 8-A5>에 나타나 있는데 여기에 있는 결정계수는 제약이 가해지지 않은 모형의 결정계수 즉 $R_{ur}^2 = 0.999497$이다.

<그림 8-A5> 총통화 회귀식 추정결과
(제약이 없는 모형)



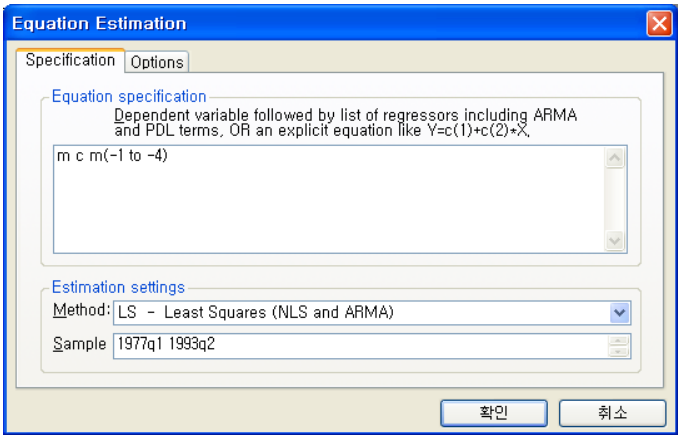
Y가 M의 원인변수가 아니라는 귀무가설을 검정해 보자. 이 제약을 가한 후 총통화 방정식을 추정하기 위하여 방정식 추정 대화상자에서 <그림 8-A6>과 같이 입력한다.

총통화 회귀식의 추정 결과가 <그림 8-A7>에 나타나 있는데 여기에 있는 결정계수는 제약을 가한 모형의 결정계수 즉 $R_r^2 = 0.999318$이다.

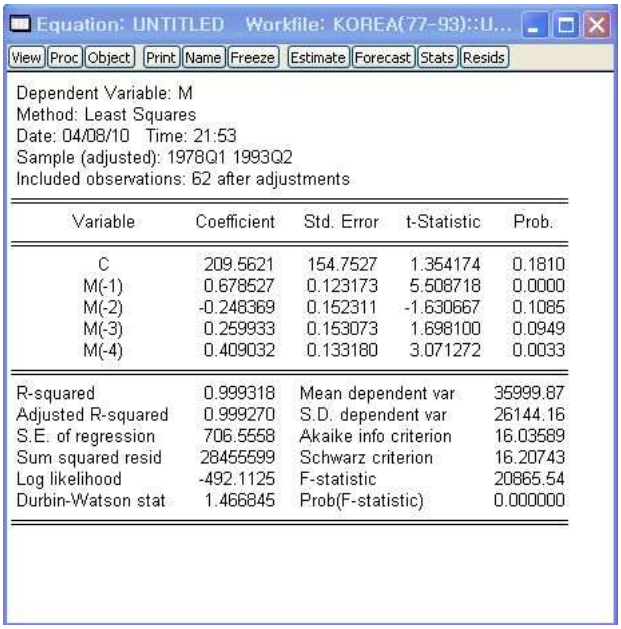
제4장의 (4.23)식에 의해 다음과 같이 F-통계량이 계산되는데 <그림 8-A3>에서 Y does not Granger cause M의 F-통계량과 일치함을 확인할 수 있다.

$$F = \frac{(R_{ur}^2 - R_r^2)/2}{(1-R_{ur}^2)/n-k} = \frac{(0.999497-0.999318)/4}{(1-0.999497)/53} = 4.7 \sim F_{53}^4$$

<그림 8-A6> 총통화 추정식(제약을 가한 모형)



<그림 8-A7> 총통화 회귀식 추정결과
(제약을 가한 모형)



연습문제

1. 현재소비(C_t)는 현재소득(Y_t)의 함수라는 단순한 소득-소비모형에서 현재소득과 과거 k기까지의 소득($Y_{t-1}, Y_{t-2} \ldots Y_{t-k}$)도 현재소비에 영향을 준다고 할 경우 다음 물음에 답하라.

① 경제분석에서 정태분석과 동태분석의 차이를 설명하라

② 위의 두 모형을 회귀모형으로 각각 나타내어라

③ 시차변수가 모형에 들어갈 수 있는 근거를 설명하라

④ 시차분포모형을 보통최소자승법으로 추정할 경우 발생할 수 있는 문제점은?

2. 시차분포모형 $y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + u_t$에서 β에 대한 2차 다항식 즉, $\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2$을 가정하여 추정한다고 할 때 그 추정방법은 먼저 $\hat{a_0}$, $\hat{a_1}$, $\hat{a_2}$을 추정하고 2차 다항식 관계를 이용하여 $\beta_0, \ldots, \beta_k$를 추정한다. 이때 시차수(k)에 따라서 추정해야 할 α(회귀계수)의 수도 달라진다. 이를 평가하라.

3. 시차분포모형 $y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + u_t$에서 β에 대한 2차 다항식 즉, $\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2$을 가정하여 Almon의 방법으로 변환하여 추정하니 $\hat{\alpha_0} = 0.66$, $\hat{\alpha_1} = 0.9$, $\hat{\alpha_2} = -0.43$을 얻었다.

① 이를 이용하여 $\beta_0, \ldots \beta_3$을 각각 추정하라.

② Y를 생산, X를 통화량이라고 할 때 통화량의 단기승수와 장기승수 값을 각각 구하라.

4. 생산과 통화 간에 Granger의 인과성 및 중립성 검정을 해 보고자 하는데 다음의 표는 추정결과를 요약해 놓은 것이다.(가설검정은 5% 유의수준 하에서 하라)

종속변수	상수항	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	m_{t-1}	m_{t-2}	m_{t-3}	m_{t-4}	R^2
y	0.12	0.81	0.35	-0.19	0.01	0.63	-0.75	-0.15	0.28	0.995507
y	-0.03	0.79	0.26	-0.09	0.04					0.994792
y	0.12	0.81	0.35	-0.19	0.01	0.60	-0.75	-0.12	0.27	0.99543
m	-0.16	-0.11	0.06	-0.12	0.21	1.19	0.02	-0.06	-0.17	0.999801
m	0.07					1.22	-0.001	-0.09	-0.12	0.999747
m	-0.16	-0.12	0.07	-0.13	0.18	1.19	0.12	-0.06	-0.17	0.999789
* 모든 회귀방정식은 1973년 2/4분기부터 1989년 4/4분기까지의 자료로 추정함 * 5% 유의수준 하에서 F분포의 임계치는 F(1,54)=4.02, F(4,54)=2.55임										

① 통화는 생산의 원인변수가 되지 않는다는 가설을 검정하고 결론의 경제적 의미를 설명하라.

② 생산은 통화의 원인변수가 되지 않는다는 가설을 검정하고 결론의 경제적 의미를 설명하라.

제 9 장
시계열분석

1.단일시계열모형
2.Box-Jenkins 접근방법
3.단위근
4.공적분
5.벡터자기회귀(VAR)모형
6.오차수정모형
실습 : AR 모형 추정 및 예측
단위근 및 공적분 검정
VAR 모형 및 ECM 모형

제9장 시계열분석

1.단일시계열모형

시계열(time series)이란 시간에 따라 특정 변수에 대한 관측치를 나열해 놓은 것으로 시계열에서 나타나는 중요한 특징은 관측치들이 독립적이지 않고 서로 관계가 있다는 것이다. 시계열모형(time series model)이란 어떤 변수가 과거 값들과 통계적으로 어떠한 관계에 있는 지를 수식으로 표시한 것을 말하는데 단일시계열모형(univariate time series model)이란 어떤 변수가 자기의 과거 값들과 통계적으로 어떠한 관계에 있는 지를 수식으로 표시한 것이다. 시계열분석(time series analysis)은 시계열이 가지고 있는 이러한 특징을 이용하여 관측치들의 변화들을 모형으로 표현한 후 경제이론을 검증하고, 정책을 분석하며, 관측치들의 미래 값을 예측하는 것을 말한다.

(1)확률과정

T개의 관측치(시계열자료) $y_{1},\ldots,y_{T}$가 있다고 하자. 이 관측치들은 확률변수 $Y_{1},\ldots,Y_{T}$의 실현된 값이고 확률변수 $Y_{1},\ldots,Y_{T}$는 무한계열의 일부분으로 볼 수 있는데 이 경우 무한한 계열을 확률과정(stochastic process)이라고 한다.

확률과정의 종류는 여러 가지가 있는데 대표적인 확률과정으로는 다음과 같은 것이 있다.

①백색잡음과정

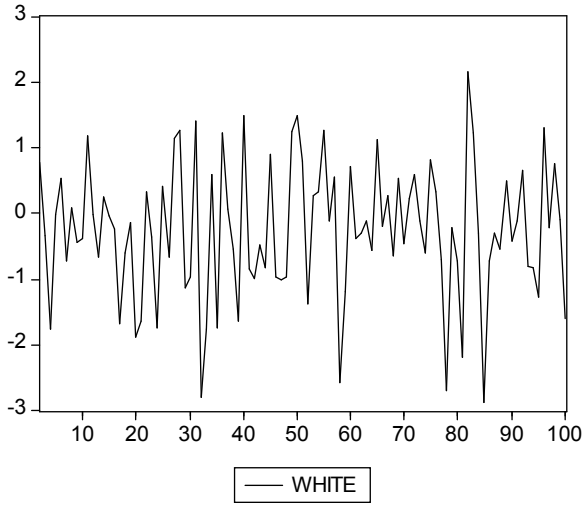
시계열이 상호 독립적이고 동등하게 분포하는 확률변수의 관측치일 경우 백색잡음과정(white noise process)이라고 한다. 다시 말해서 $e_{1},\ldots,e_{T}$이 있을 경우 다음이 성립하는 확률과정을 백색잡음과정이라고 하며 <그림 9-1>은 표준정규분포에서 가상적으로 생성된 백색잡음과정을 나타내고 있다.

$E(e_t)=0 \ \text{for} \ \forall \ t$ (zero mean)

$E(e_t e_s)=0 \ \text{for} \ \forall \ t \neq s$ (uncorrelated)

$E(e_1^2) = ...E(e_t^2) = \sigma_e^2 < \infty$ **(finite variance)**

<그림 9-1> 백색잡음과정



②자기회귀과정

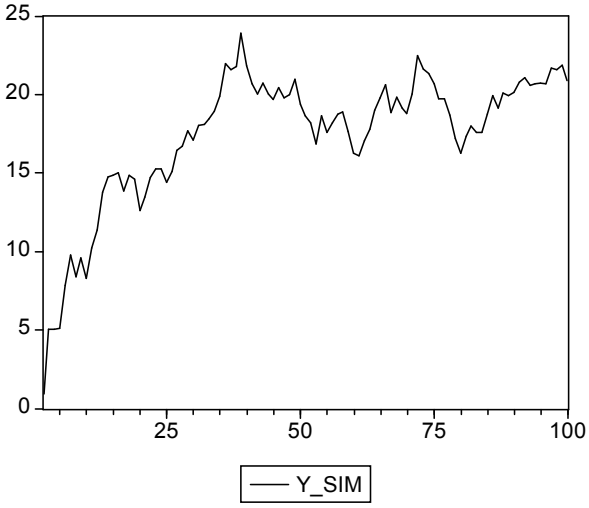
시계열이 자기의 과거 값들과 통계적인 관계를 가지고 있는 확률변수의 관측치일 경우 자기회귀과정**(autoregressive process)**이라고 하는데 **AR(1), AR(2), ..., AR(p)** 등이 있고 다음과 같이 나타낸다. 여기서 e_t는 백색잡음과정이라고 가정한다. **<그림 9-2-1>** 및 **<그림 9-2-2>**는 가상적으로 생성된 **AR(1)**의 확률과정을 나타내고 있고, **<그림 9-3-1>** 및 **<그림 9-3-2>**는 가상적으로 생성된 **AR(2)**의 확률과정을 나타내고 있다.

$y_t = \phi y_{t-1} + e_t$ **: AR(1)**

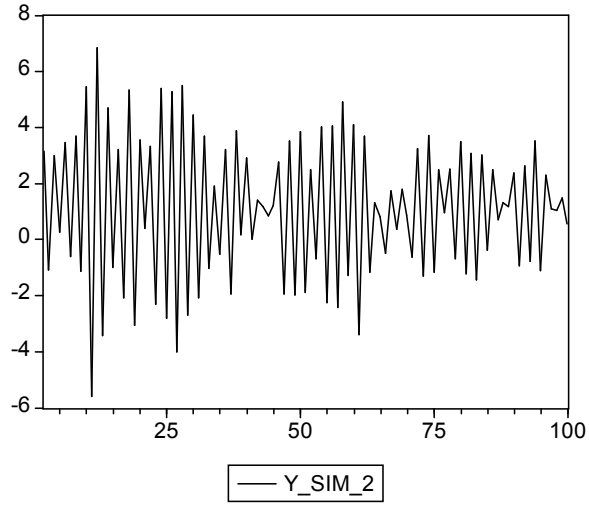
$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$ **: AR(2)**

$y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + e_t$ **: AR(p)**

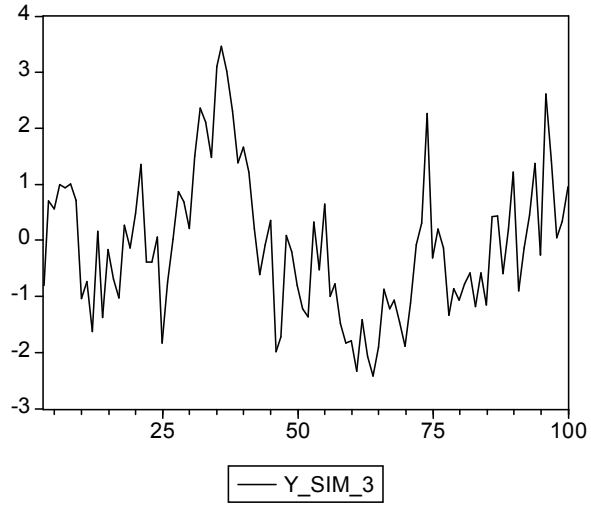
<그림 9-2-1> AR(1): $y_t = 2 + 0.9y_{t-1} + e_t$



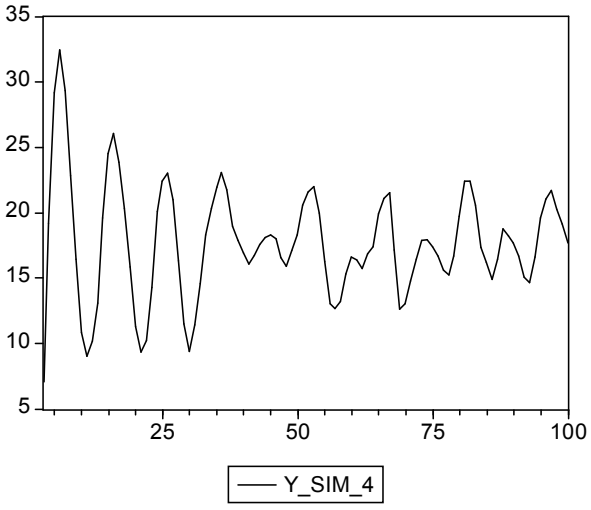
<그림 9-2-2> AR(1): $y_t = 2 - 0.9y_{t-1} + e_t$



<그림 9-3-1> AR(2): $y_t = 0.5y_{t-1} + 0.3y_{t-2} + e_t$



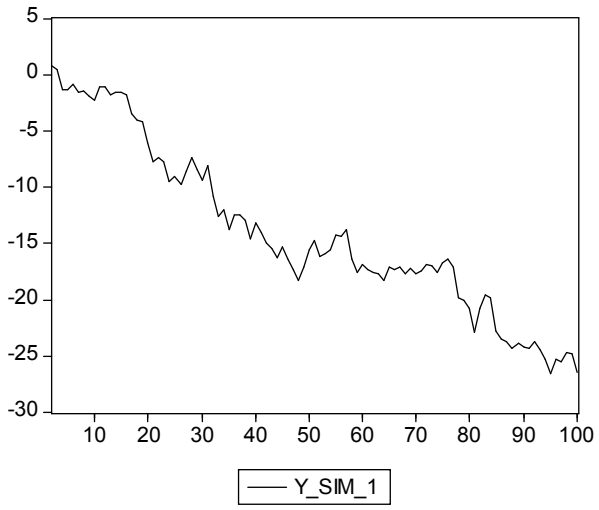
<그림 9-3-2> AR(2): $y_t = 7 + 1.4y_{t-1} - 0.8y_{t-2} + e_t$



③임의보행과정

AR(1)의 자기회귀모형에서 $\phi = 1$인 경우를 임의보행과정**(random walk process)**이라고 하며 여기서 e_t는 백색잡음과정이다. **<**그림 **9-4>**는 가상적으로 생성된 임의보행과정을 나타내고 있다.

<그림 9-4> Random Walk: $y_t = y_{t-1} + e_t$



④이동평균과정

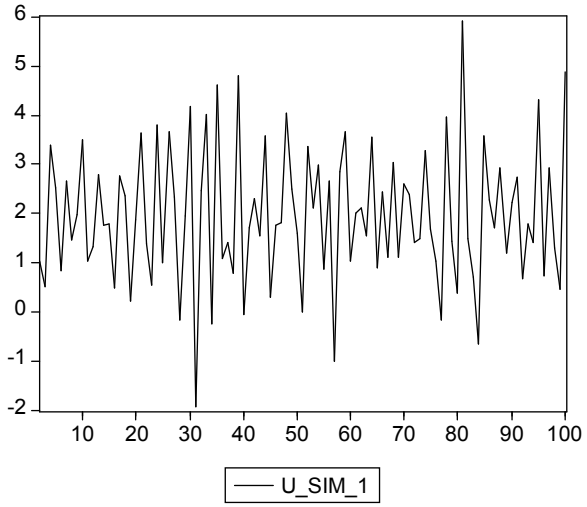
시계열이 교란항의 현재 값 및 과거 값들과 통계적인 관계를 가지고 있는 확률변수의 관측치일 경우 이동평균과정**(moving average process)**이라고 하는데 **MA(1), MA(2), ..., MA(q)** 등이 있고 다음과 같이 나타낸다. 여기서 e_t는 백색잡음과정이라고 가정한다. **<**그림 **9-5-1>** 및 **<**그림 **9-5-2>**는 가상적으로 생성된 **MA(1)** 및 **MA(2)**의 확률과정을 나타내고 있다.

$y_t = e_t - \theta e_{t-1}$: **MA(1)**

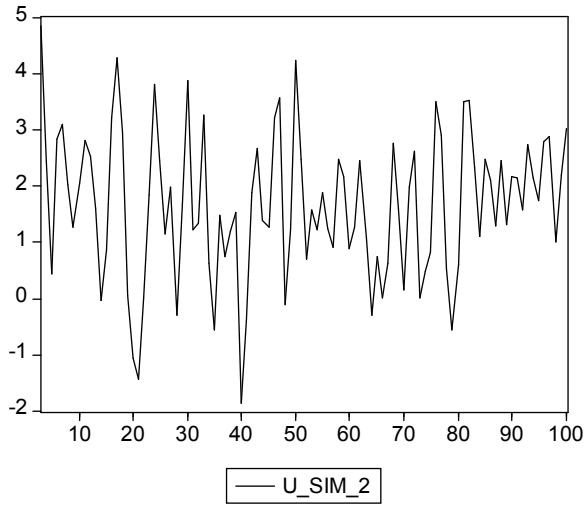
$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$: **MA(2)**

$y_t = e_t - \theta_1 e_{t-1} - .... - \theta_q e_{t-q}$: **MA(q)**

<그림 9-5-1> MA(1): $y_t = 2 + e_t - 0.8e_{t-1}$



<그림 9-5-2> MA(2): $y_t = 2 + e_t + 0.6e_{t-1} - 0.2e_{t-2}$



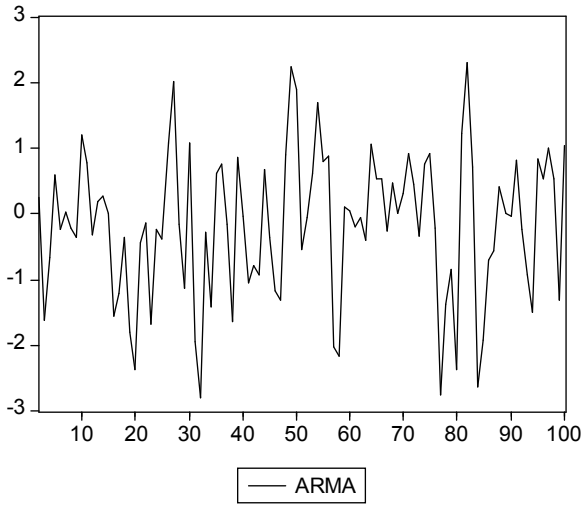
⑤자기회귀이동평균과정

자기회귀과정과 이동평균과정이 결합된 경우를 자기회귀이동평균과정**(ARMA process)**이라고 하고 특히, 수준변수가 아닌 차분변수가 **ARMA process**에 따를 경우를 **ARIMA process**라고 하고 다음과 같이 나타낸다. 여기서 e_t는 백색잡음과정이라고 가정한다. **<**그림 **9-6>**은 가상적으로 생성된 **ARMA(1,1)**의 확률과정을 나타내고 있다.

$$y_t - \phi_1 y_{t-1} + \ldots\ldots - \phi_p y_{t-p} = e_t + \theta_1 e_{t-1} \ldots\ldots + \theta_q e_{t-q} \quad : \textbf{ARMA(p,q)}$$

$$\Delta y_t - \phi_1 \Delta y_{t-1} + \ldots\ldots - \phi_p \Delta y_{t-p} = e_t + \theta_1 e_{t-1} \ldots\ldots + \theta_q e_{t-q} \quad : \textbf{ARIMA(p,d,q)}$$

<그림 9-6> ARMA(1,1): $y_t = 0.2 + 0.2 y_{t-1} + e_t + 0.6 e_{t-1}$



(2)시계열의 안정성

시계열분석에서 고려해야 할 중요한 사항은 시계열의 안정성**(stationarity)**인데 다음의 **3** 가지 조건을 만족하는 확률과정, X_t,를 안정적 시계열이라고 한다.

$$E(X_t) = E(X_{t+m}) = \mu_X$$

$$E[(X_t - \mu_X)^2] = E[(X_{t+m} - \mu_X)^2] = \sigma_X^2$$

$$E[(X_t - \mu_X)(X_{t+k} - \mu_X)] = E[(X_{t+m} - \mu_X)(X_{t+k+m} - \mu_X)] = \gamma_X$$

즉, 안정적 시계열은 시간 **t**에 관계없이 평균이 일정하고, 시계열의 분산이 **t**에 관계없이 일정하며, 두 시점에서 시계열의 공분산이 일정하며 그 값이 **k**에 관계없이 관측되는 시간에만 의존한다.

한편, 이상의 **3** 조건 중 하나가 성립하지 않으면 불안정 시계열**(non-stationary time series)**이라고 한다. 시계열의 평균이 일정하지 않을 경우를 **homogeneous non-stationarity**라고 하는데 이러한 경우는 원계열을 차분**(differencing)**을 해 주면 안정시계열이 된다. 만약 원계열 y_t의 수준이 시간에 따라 다를 경우는 **1**차 차분($\Delta y_t = y_t - y_{t-1}$)을 해 주고, 원계열 y_t의 수준과 기울기가 시간에 따라 다를 경우 다음과 같이 **2**차 차분을 해 준다.

$$\begin{aligned}\Delta^2 y_t &= \Delta y_t - \Delta y_{t-1}) \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2}\end{aligned}$$

시계열의 분산이 일정하지 않을 경우를 **non-stationary variance**라고 하는데 이러한 경우 원계열에 자연대수**(natural logarithm)**를 취하면 된다. 원계열에 자연대수를 취할 때 주의할 점은 원계열이 비율이나 증가율로 되어 있는 경우 자연대수를 취하지 않고 경제성장률, 물가상승률, 통화증가율과 같이 일반적으로 성장률(증가율)의 개념으로 생각될 수 있는 시계열은 자연대수를 취하고, 자연대수를 취한 후 차분여부를 결정해야 한다.

대표적인 불안정 시계열은 임의보행과정인데 그 이유를 살펴보자.

다음의 **(9.1)**식과 같은 임의보행과정을 **1**차 차분방정식**(difference equation)**이라고도 하는데 **1**차 차분방정식의 해를 구하는 방법 중 하나는 반복적인 대입을 통해 y_t를 교란항의 함수로만 표현하는 것이다.

$$y_t = y_{t-1} + e_t \qquad (9.1)$$

(9.1)식의 임의보행과정에서 반복적인 대입을 통해 **(9.2)**식을 구할 수 있다.

$$
\begin{aligned}
y_t &= y_{t-1} + e_t \\
&= y_{t-2} + e_{t-1} + e_t \\
&\quad \cdot \\
&\quad \cdot \\
y_t &= y_0 + \sum_{j=1}^{t} e_j
\end{aligned}
\qquad \textbf{(9.2)}
$$

만약에 y_0를 주어진 것으로 가정하면, 즉 $y_0 = 0$ 이라 하면 다음이 성립한다.

$$E(y_t) = 0$$

$$E(y_t^2) = E[(e_t + e_{t-1} \cdots + e_1)(e_t + e_{t-1} \cdots + e_1)] = t\sigma_e^2$$

$$E(y_t y_{t-1}) = E[(e_t + e_{t-1} \cdots + e_1)(e_{t-1} + e_{t-2} \cdots + e_1)] = (t-1)\sigma_e^2$$

임의보행과정의 평균은 0이지만 분산은 $t\sigma_e^2$으로 t에 의존하며, 공분산 역시 t에 의존하여 안정시계열의 조건을 충족시키지 못하는 불안정시계열이다.

(3)자기상관함수

자기상관함수**(autocorrelation function: ACF)**는 관측치들 사이에 어느 정도의 상호의존성이 있는 지를 측정하고 안정시계열의 여부와 모형의 식별에 유용한 판단 기준이 되는데 **(9.3)**식과 같다. 이 식에서 ρ_k는 기준 관측치와 k기만큼 떨어진 관측치의 상관도를 나타낸다.

$$\rho_k = \frac{Cov(y_t y_{t+k})}{Var(y_t)} = \frac{\sum_{t=1}^{T-k}(y_t - \overline{y})(y_{t+k} - \overline{y})}{\sum_{t=1}^{T}(y_t - \overline{y})^2} \equiv \frac{\gamma_k}{\gamma_0} \quad \textbf{(9.3)}$$

예를 들어 다음의 **(9.4)**식과 같은 **AR(1)** 확률과정의 **ACF(**이를 이론적 **ACF**라고 함**)**를 살펴보자.

$$y_t = \phi y_{t-1} + e_t,\ e_t \sim (0, \sigma_e^2) \quad \textbf{(9.4)}$$

(9.4)식 역시 **1**차 차분방정식이므로 반복적인 대입을 통해 y_t를 교란항의 함수로만 표현해 보면**(**$y_0 = 0$로 가정**)** **(9.5)**식과 같다.

$$\begin{aligned} y_t &= \phi y_{t-1} + e_t \\ &= \phi^2 y_{t-2} + \phi e_{t-1} + e_t \\ &\quad \cdot \\ &= \sum_{j=0}^{\infty} \phi^j e_{t-j} \end{aligned} \quad \textbf{(9.5)}$$

(9.5)식을 이용하여 y_t의 평균, 분산 및 공분산을 구해보면 다음과 같다.

$$E(y_t) = 0$$

$$\begin{aligned} Var(y_t) &= \sigma_e^2 + \phi^2\sigma_e^2 + \phi^4\sigma_e^4 + \cdots\cdots \\ &= \frac{\sigma_e^2}{1-\phi^2} \\ &\equiv \gamma_0 \end{aligned}$$

$$Cov(y_t y_{t-1}) = E(y_t y_{t-1})$$

$$= E[(e_t + \phi e_{t-1} + ...)(e_{t-1} + \phi e_{t-2} + ...)]$$
$$= \phi\sigma_e^2 + \phi^3\sigma_e^2 + ...$$
$$= \frac{\phi}{1-\phi^2}\sigma_e^2$$
$$\equiv \gamma_1$$

$$Cov(y_t y_{t-k}) = \frac{\phi^k}{1-\phi^2}\sigma_e^2$$
$$\equiv \gamma_k$$

따라서 **(9.3)**식의 정의에 의해 다음과 같이 $\rho_0,\rho_1,...,\rho_k$를 구할 수 있다.

$$\rho_0 \equiv \frac{\gamma_0}{\gamma_0} = 1$$

$$\rho_1 \equiv \frac{\gamma_1}{\gamma_0} = \phi$$

.

$$\rho_k \equiv \frac{\gamma_k}{\gamma_0} = \phi^k$$

이러한 **ACF**를 그래프로 나타낸 것을 상관도(**correlogram**)라 하며 이를 이용하여 시계열의 안정성여부를 판단할 수 있다. 즉, 시계열이 안정적이면($\phi \to 0$) **ACF**는 급격하게 **0**을 향해 감소하고, 시계열이 불안정적이면($\phi \to 1$) **ACF**는 **0**을 향해 천천히 감소한다.

한편, 다음의 **(9.6)**식과 같은 **MA(1)** 확률과정의 **ACF** 살펴보자.

$$y_t = e_t - \theta e_{t-1},\ e_t \sim (0,\sigma_e^2) \qquad \textbf{(9.6)}$$

(9.6)식을 이용하여 y_t의 평균, 분산 및 공분산을 구해보면 다음과 같다.

$$E(y_t) = 0$$

$$\begin{aligned} Var(y_t) &= \sigma_e^2 + \theta^2\sigma_e^2 \\ &= (1+\theta^2)\sigma_e^2 \\ &\equiv \gamma_0 \end{aligned}$$

$$\begin{aligned} Cov(y_t y_{t-1}) &= E(y_t y_{t-1}) \\ &= E[(e_t - \theta e_{t-1})(e_{t-1} - \theta e_{t-2})] \\ &= -\theta\sigma_e^2 \\ &\equiv \gamma_1 \end{aligned}$$

$$\begin{aligned} Cov(y_t y_{t-2}) &= E(y_t y_{t-2}) \\ &= E[(e_t - \theta e_{t-1})(e_{t-2} - \theta e_{t-3})] \\ &= 0 \\ &\equiv \gamma_2 \end{aligned}$$

따라서 **(9.3)**식의 정의에 의해 다음과 같이 $\rho_0, \rho_1, \ldots, \rho_k$를 구할 수 있다.

$$\rho_1 \equiv \frac{\gamma_1}{\gamma_0} = \frac{-\theta}{1+\theta^2}$$

$$\rho_2 \equiv \frac{\gamma_2}{\gamma_0} = 0$$

$$\vdots$$

$$\rho_k = 0$$

(4)편자기상관함수

편자기상관함수(partial autocorrelation function: PACF)는 k 이외의 모든 시차를 갖는 관측치($y_{t-1}, y_{t-2}, \ldots, y_{t-k+1}$)들로부터의 영향력을 배제한 가운데 특정한 두 관측치, y_t 와 y_{t-k}가 얼마나 관련이 있는 지를 나타내는 척도로 회귀계수가 편자기상관함수가 된다.

$$y_t = \mu + \phi_{11} y_{t-1} + e_t \rightarrow \hat{\phi}_{11}$$

$$y_t = \mu + \phi_{11} y_{t-1} + \phi_{22} y_{t-2} + e_t \rightarrow \hat{\phi}_{22}$$

$$y_t = \mu + \phi_{11} y_{t-1} + \phi_{22} y_{t-2} + \phi_{33} y_{t-3} + e_t \rightarrow \hat{\phi}_{33}$$

따라서 AR(1)에서는 $\hat{\phi}_{22} = \hat{\phi}_{33} = \ldots = \hat{\phi}_{kk} = 0$이 된다.

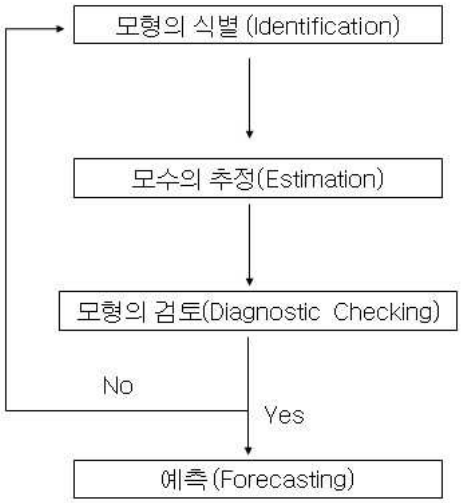
2.Box-Jenkins 접근방법

시계열분석 기법을 이용하여 시계열 자료의 동태적임 움직임을 파악하기 위하여 시계열 모형을 설정하고, 모형을 추정하며, 모형의 적합성을 검토하고, 예측하는 과정을 거쳐야 하는데 이를 박스-젠킨스(Box-Jenkins) 접근방법이라고 하며 이를 도식화시키면 <그림 9-7>과 같다.

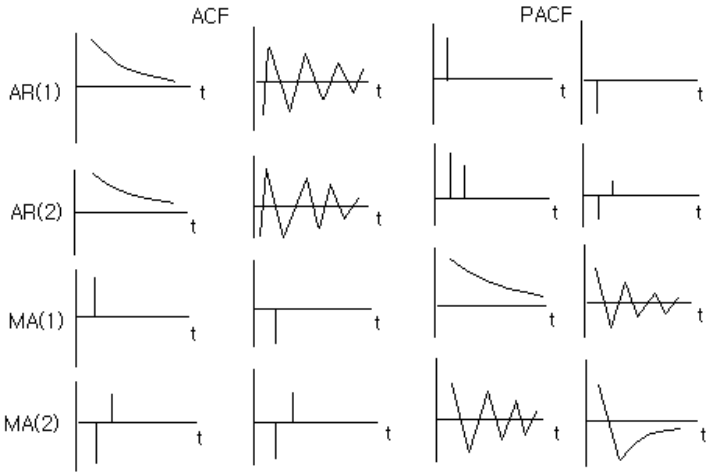
(1)모형의 식별

박스-젠킨스는 시계열 자료는 AR(p) 모형, MA(q) 모형, ARMA(p,q) 모형, ARIMA(p,d,q) 모형 등으로 모형화할 수 있다고 보았으며 시계열분석 기법인 자기상관함수(ACF) 및 편자기상관함수(PACF)로 <그림 9-8>과 같이 모형의 식별이 가능하다고 하였다. 먼저, ACF가 점진적으로 감소하면 불안정시계열이므로 원계열을 차분하여 안정시계열로 만들어 줄 수 있다. 다음으로 ACF가 0을 향해 감소하고 PACF는 1-2개 정도만 통계적으로 유의하면 AR 모형이고, PACF가 0을 향해 감소하고 ACF는 1-2개 정도만 통계적으로 유의하면 MA 모형이며, ACF와 PACF가 모두 0을 향해 감소하면 ARMA 모형이다.

<그림 9-7> 박스-젠킨스 접근방법



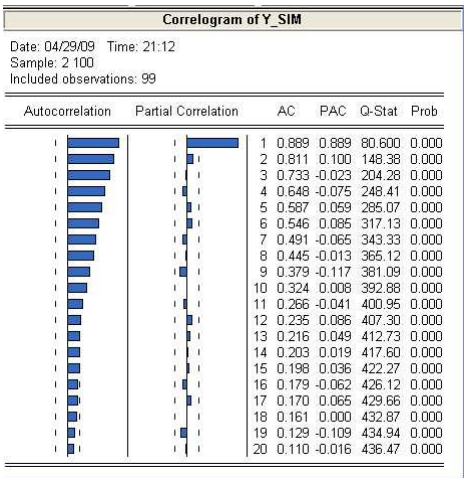
<그림 9-8> ACF와 PACF를 이용한 모형식별



<그림 9-2-1>에 있는 AR(1) 모형의 상관도(correlogram)를 그려보면 <그림 9-9-1>과 같고, <그림 9-2-2>에 있는 AR(1)의 상관도(correlogram)를 그려보면 <그림 9-9-2>와 같다. 이를 통해 확인할 수 있듯이 ACF는 점차적으로(또는 순환

하면서) 0으로 감소하고 있고 PACF는 시차 1개만이 통계적으로 유의하다.

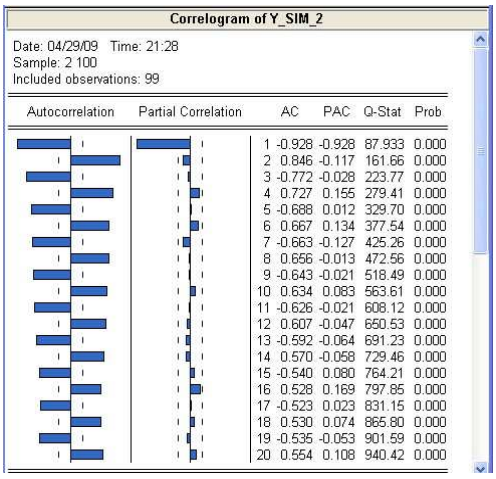
<그림 9-9-1> AR(1) 모형의 상관도


Correlogram of Y_SIM

Date: 04/29/09 Time: 21:12
Sample: 2 100
Included observations: 99

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.889	0.889	80.600	0.000
		2	0.811	0.100	148.38	0.000
		3	0.733	-0.023	204.28	0.000
		4	0.648	-0.075	248.41	0.000
		5	0.587	0.059	285.07	0.000
		6	0.546	0.085	317.13	0.000
		7	0.491	-0.065	343.33	0.000
		8	0.445	-0.013	365.12	0.000
		9	0.379	-0.117	381.09	0.000
		10	0.324	0.008	392.88	0.000
		11	0.266	-0.041	400.95	0.000
		12	0.235	0.086	407.30	0.000
		13	0.216	0.049	412.73	0.000
		14	0.203	0.019	417.60	0.000
		15	0.198	0.036	422.27	0.000
		16	0.179	-0.062	426.12	0.000
		17	0.170	0.065	429.66	0.000
		18	0.161	0.000	432.87	0.000
		19	0.129	-0.109	434.94	0.000
		20	0.110	-0.016	436.47	0.000

<그림 9-9-2> AR(1) 모형의 상관도


Correlogram of Y_SIM_2

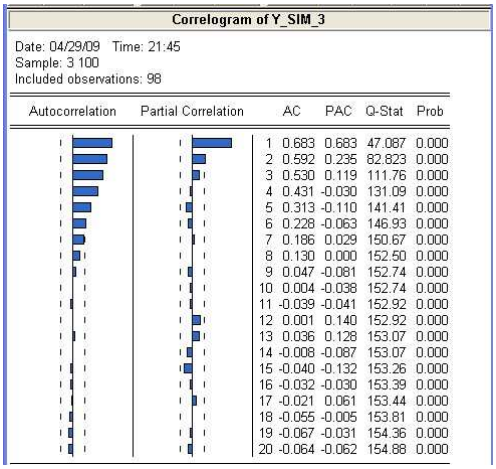
Date: 04/29/09 Time: 21:28
Sample: 2 100
Included observations: 99

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.928	-0.928	87.933	0.000
		2	0.846	-0.117	161.66	0.000
		3	-0.772	-0.028	223.77	0.000
		4	0.727	0.155	279.41	0.000
		5	-0.688	0.012	329.70	0.000
		6	0.667	0.134	377.54	0.000
		7	-0.663	-0.127	425.26	0.000
		8	0.656	-0.013	472.56	0.000
		9	-0.643	-0.021	518.49	0.000
		10	0.634	0.083	563.61	0.000
		11	-0.626	-0.021	608.12	0.000
		12	0.607	-0.047	650.53	0.000
		13	-0.592	-0.064	691.23	0.000
		14	0.570	-0.058	729.46	0.000
		15	-0.540	0.080	764.21	0.000
		16	0.528	0.169	797.85	0.000
		17	-0.523	0.023	831.15	0.000
		18	0.530	0.074	865.80	0.000
		19	-0.535	-0.053	901.59	0.000
		20	0.554	0.108	940.42	0.000

<그림 9-3-1>에 있는 AR(2) 모형의 상관도(correlogram)를 그려보면 <그림 9-10-1>과 같고, <그림 9-3-2>에 있는 AR(1) 모형의 상관도(correlogram)를 그려보면 <그림 9-10-2>와 같다. 이를 통해 확인할 수 있듯이 ACF는 점차적으로(또는

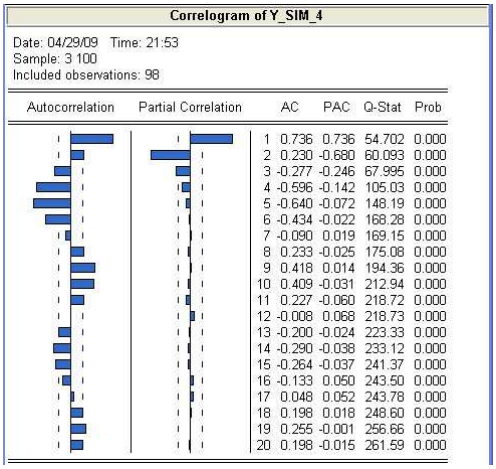
순환하면서) 0으로 감소하고, PACF는 시차 2개만이 통계적으로 유의하다.

<그림 9-10-1> AR(2) 모형의 상관도



	AC	PAC	Q-Stat	Prob
1	0.683	0.683	47.087	0.000
2	0.592	0.235	82.823	0.000
3	0.530	0.119	111.76	0.000
4	0.431	-0.030	131.09	0.000
5	0.313	-0.110	141.41	0.000
6	0.228	-0.063	146.93	0.000
7	0.186	0.029	150.67	0.000
8	0.130	0.000	152.50	0.000
9	0.047	-0.081	152.74	0.000
10	0.004	-0.038	152.74	0.000
11	-0.039	-0.041	152.92	0.000
12	0.001	0.140	152.92	0.000
13	0.036	0.128	153.07	0.000
14	-0.008	-0.087	153.07	0.000
15	-0.040	-0.132	153.26	0.000
16	-0.032	-0.030	153.39	0.000
17	-0.021	0.061	153.44	0.000
18	-0.055	-0.005	153.81	0.000
19	-0.067	-0.031	154.36	0.000
20	-0.064	-0.062	154.88	0.000

<그림 9-10-2> AR(2) 모형의 상관도

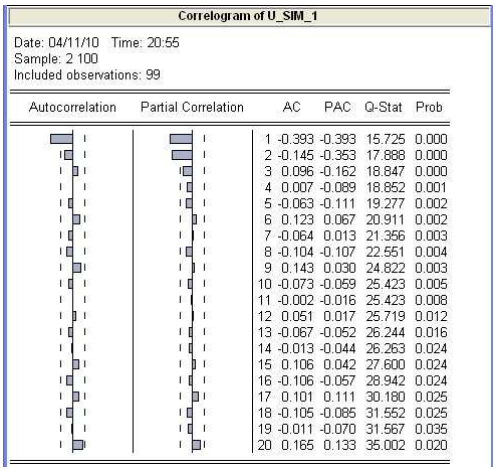


	AC	PAC	Q-Stat	Prob
1	0.736	0.736	54.702	0.000
2	0.230	-0.680	60.093	0.000
3	-0.277	-0.246	67.995	0.000
4	-0.596	-0.142	105.03	0.000
5	-0.640	-0.072	148.19	0.000
6	-0.434	-0.022	168.28	0.000
7	-0.090	0.019	169.15	0.000
8	0.233	-0.025	175.08	0.000
9	0.418	0.014	194.36	0.000
10	0.409	-0.031	212.94	0.000
11	0.227	-0.060	218.72	0.000
12	-0.008	0.068	218.73	0.000
13	-0.200	-0.024	223.33	0.000
14	-0.290	-0.038	233.12	0.000
15	-0.264	-0.037	241.37	0.000
16	-0.133	0.050	243.50	0.000
17	0.048	0.052	243.78	0.000
18	0.198	0.018	248.60	0.000
19	0.255	-0.001	256.66	0.000
20	0.198	-0.015	261.59	0.000

<그림 9-5-1>에 있는 MA(1) 모형의 상관도(correlogram)를 그려보면 <그림 9-11-1>과 같고, <그림 9-5-2>에 있는 MA(2) 모형의 상관도(correlogram)를 그려보면 <그림 9-11-2>와 같다. 이를 통해 확인할 수 있듯이 PACF는 점차적으로(또

는 순환하면서) 0으로 감소하고, ACF는 시차 1개 또는 2개만이 통계적으로 유의하다.

<그림 9-11-1> MA(1) 모형의 상관도

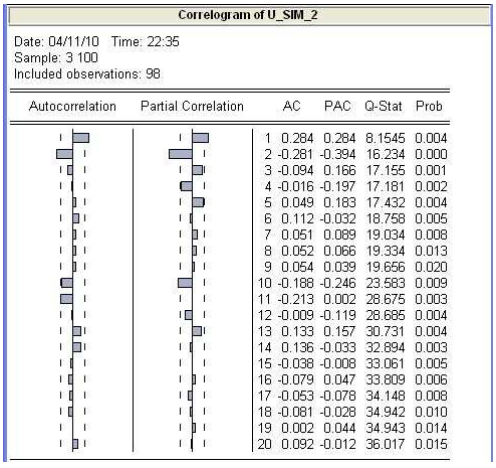


Correlogram of U_SIM_1

Date: 04/11/10 Time: 20:55
Sample: 2 100
Included observations: 99

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.393	-0.393	15.725	0.000
		2	-0.145	-0.353	17.888	0.000
		3	0.096	-0.162	18.847	0.000
		4	0.007	-0.089	18.852	0.001
		5	-0.063	-0.111	19.277	0.002
		6	0.123	0.067	20.911	0.002
		7	-0.064	0.013	21.356	0.003
		8	-0.104	-0.107	22.551	0.004
		9	0.143	0.030	24.822	0.003
		10	-0.073	-0.059	25.423	0.005
		11	-0.002	-0.016	25.423	0.008
		12	0.051	0.017	25.719	0.012
		13	-0.067	-0.052	26.244	0.016
		14	-0.013	-0.044	26.263	0.024
		15	0.106	0.042	27.600	0.024
		16	-0.106	-0.057	28.942	0.024
		17	0.101	0.111	30.180	0.025
		18	-0.105	-0.085	31.552	0.025
		19	-0.011	-0.070	31.567	0.035
		20	0.165	0.133	35.002	0.020

<그림 9-11-2> MA(2) 모형의 상관도



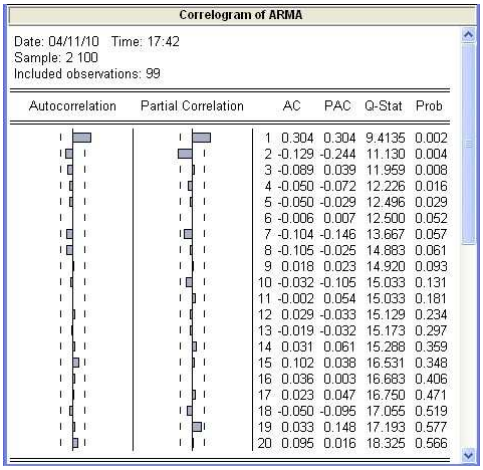
Correlogram of U_SIM_2

Date: 04/11/10 Time: 22:35
Sample: 3 100
Included observations: 98

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.284	0.284	8.1545	0.004
		2	-0.281	-0.394	16.234	0.000
		3	-0.094	0.166	17.155	0.001
		4	-0.016	-0.197	17.181	0.002
		5	0.049	0.183	17.432	0.004
		6	0.112	-0.032	18.758	0.005
		7	0.051	0.089	19.034	0.008
		8	0.052	0.066	19.334	0.013
		9	0.054	0.039	19.656	0.020
		10	-0.188	-0.246	23.583	0.009
		11	-0.213	0.002	28.675	0.003
		12	-0.009	-0.119	28.685	0.004
		13	0.133	0.157	30.731	0.004
		14	0.136	-0.033	32.894	0.003
		15	-0.038	-0.008	33.061	0.005
		16	-0.079	0.047	33.809	0.006
		17	-0.053	-0.078	34.148	0.008
		18	-0.081	-0.028	34.942	0.010
		19	0.002	0.044	34.943	0.014
		20	0.092	-0.012	36.017	0.015

<그림 9-6>에 있는 ARMA(1,1) 모형의 상관도(correlogram)를 그려보면 <그림 9-12>와 같은데 ACF와 PACF는 시차 1개만이 통계적으로 유의하다.

<그림 9-12> ARMA(1,1) 모형의 상관도



ACF 및 **PACF**를 이용한 모형식별을 요약하여 정리하면 **<표 9-1>**과 같다.

<표 9-1> 시계열모형의 **ACF** 및 **PACF**

모형	**ACF**	**PACF**
AR	-점차적으로 0으로 수렴함	-시차 **p**까지만 0과 현저히 다르고 이후 0으로 갑자기 이동 -마지막 스파이크의 시차길이는 **AR**의 차수와 동일
MA	-시차 **q**까지만 0과 현저히 다르고 이후 0으로 갑자기 이동 -마지막 스파이크의 시차길이는 **MA**의 차수와 동일	-점차적으로 0으로 수렴함
ARMA	-시차 **q**까지는 0과 현저히 다르고, 그 이후는 **AR(p)**처럼 점차적으로 소멸	-시차 **p**까지는 0과 현저히 다르고, 그 이후는 **MA(q)**처럼 점차적으로 소멸
white noise	-모든 시차에 대해 0이다	-모든 시차에 대해 0이다

(2)모수의 추정

일반적으로 다음과 같은 **AR(p)**모형은 최소자승법**(OLS)**으로 추정한다. 또한 교란항이 정규분포를 한다고 가정하면 최우추정법을 이용할 수도 있다.

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + e_t$$

다음과 같은 **MA(q)**모형은 비선형최소자승법**(Nonlinear Least Squares)**으로 추정하거나 최우추정법을 이용할 수도 있다.

$$y_t = \mu + e_t + \theta_1 e_{t-1} \dots + \theta_q e_{t-q}$$

(3)모형의 검토

모형설정이 제대로 되었으면 잔차항은 백색잡음확률과정을 따를 것이다. 따라서 잔차의 표본자기상관함수를 계산해 봄으로써 모형의 적합도를 살펴보거나 통계적 검정을 통해서 판단한다. 대표적인 통계적 검정으로는 다음과 같은 통계량을 이용한 **Ljung-Box**검정방법이 있다.

$$Q = T(T+2)\sum_{k=1}^{K} \frac{1}{T-k}\gamma_k^2 \sim \chi^2_{K-p-q}$$

단, $K = Min(\frac{T}{2}, T^{1/3})$

(4)모형의 예측

추정된 **AR(p)**모형을 이용하면 미래의 값에 대한 예측치를 도출할 수 있다. **AR(1)**의 경우를 예를 들어 **n**시점까지의 정보를 바탕으로 **k**기 이후 즉, **(n+k)**기에 대한 예측치를 구하는 방법은 다음과 같다.

$$y_t = \mu + \phi y_{t-1} + e_t$$

위 식에서 **(n+1)**기의 y 즉, y_{n+1}은 다음과 같이 나타낼 수 있다.

$$y_{n+1} = \mu + \phi y_n + e_{n+1}$$

따라서 **n**시점까지의 정보를 이용하여 **(n+1)**기에 대한 예측치를 구하면 다음과 같다.

$$\hat{y}_{n+1} = \hat{\mu} + \hat{\phi} y_n$$

유사한 방법으로 **n**시점에서의 **(n+2)**기에 대한 예측치는 다음과 같다.

$$\begin{aligned}\hat{y}_{n+2} &= \hat{\mu} + \hat{\phi} y_{n+1} \\ &= \hat{\mu} + \hat{\phi}(\hat{\mu} + \hat{\phi} y_n) \\ &= (\hat{\phi} + 1)\hat{\mu} + \hat{\phi^2} y_n\end{aligned}$$

일반적으로 **n**시점에서 **h**기 이후 즉, **(n+h)**기에 대한 예측치는 다음과 같다.

$$\hat{y}_{n+h} = (\hat{\phi}^{h-1} + \hat{\phi}^{h-2} + \ldots + 1)\hat{\mu} + \hat{\phi}^h y_n$$

3.단위근

시계열분석에서는 안정적인 시계열자료를 이용해야 하므로 자료의 안정성(stationarity) 여부를 먼저 살펴보아야 한다. 최근까지 시계열자료는 확정적 추세(deterministic trend)를 중심으로 하는 추세안정(trend stationary)계열이므로 회귀방법으로 확정적 추세를 제거한 계열은 안정적인 것으로 보고 사용하여 왔다. 그러나 1980대 들어와 대부분의 거시시계열들이 확률적 추세(stochastic trend)를 가지고 있는 차분안정(difference stationary)계열이므로 차분(differencing)을 함으로써 안정적인 시계열을 얻어 사용하고 있다.

(9.1)식의 **AR(1)**모형에서 $\rho = 1$인 경우 단위근**(unit root)**을 가졌다고 하고 단위근을 가진 시계열은 임의보행과정**(random walk)**의 불안정한 시계열이 된다.

$$y_t = \rho y_{t-1} + e_t, \quad e_t \sim i.i.d \tag{9.7}$$

일반적으로 확률적인 추세를 가진 시계열을 단위근을 가졌다고 말하는데 단위근을 가진 불안정 시계열을 회귀분석에 그대로 이용하면 표본수가 증가함에 따라 회귀계수의 t-값도 증가하여 상관관계가 없는 변수 사이에도 마치 강한 상관관계가 있는 것으로 나타나는 가성회귀(spurious regression)의 문제가 발생하게 된다. 즉, (9.1)식에서 $\rho=1$이라는 귀무가설 하에서 계산된 **t-**통계량은 큰 자료에서 조차 **t-**분포를 따르지 않는다**.** 또한 경제시계열이 확률적 추세를 가지고 있다는 것은 충격(innovation)이 확률적 추세를 가지고 있다는 것인데 이러한 충격들은 그 영향이 지속적인 것이 특징이다.

시계열이 단위근을 가지고 있는지를 검정하는 방법은 여러 가지가 있는데 대표적인 방법으로는 Dickey-Fuller(DF) 검정방법과 Augmented Dickey-Fuller(ADF) 검정방법이 있다.

(1)Dickey-Fuller(DF) 검정방법

DF 검정방법은 최소자승법으로 (9.8)-(9.10)식의 회귀방정식을 추정하고 y_{t-1}의 회귀계수가 0인지 즉, $\rho=1$인지를 검정하는 t-통계량을 부록 <표 6>에 있는 분포표의 임계치와 비교하여 검정한다.

$$\Delta y_t=(\rho-1)y_{t-1}+e_t,\ \ e_t\sim i.i.d \tag{9.8}$$

$$\Delta y_t=\alpha+(\rho-1)y_{t-1}+e_t,\ \ e_t\sim i.i.d \tag{9.9}$$

$$\Delta y_t=\alpha+\beta t+(\rho-1)y_{t-1}+e_t,\ \ e_t\sim i.i.d \tag{9.10}$$

이때 (9.8)식의 t-통계량은 부록 <표 8>의 $\hat{\tau}$의 분포표를 이용하고, (9.9)식의 t-통계량은 부록 <표 8>의 $\hat{\tau_\mu}$의 분포표를 이용하며, (9.10)식의 t-통계량은 부록 <표 8>의 $\hat{\tau_\tau}$의 분포표를 각각 이용한다.

(2)Augmented Dickey-Fuller(ADF) 검정방법

(9.8)-(9.10)식에서 교란항이 i.i.d의 가정을 충족시키지 못할 경우 (9.11)-(9.13)식과 같이 y_t의 차분변수의 시차변수를 설명변수로 포함시켜 $\rho=1$이

라는 귀무가설을 검정한다.

$$\Delta y_t = (\rho - 1) y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + e_t, \quad e_t \sim i.i.d \tag{9.11}$$

$$\Delta y_t = \alpha + (\rho - 1) y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + e_t, \quad e_t \sim i.i.d \tag{9.12}$$

$$\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + e_t, \quad e_t \sim i.i.d \tag{9.13}$$

이때 (9.11)식의 t-통계량은 부록 <표 8>의 $\hat{\tau}$의 분포표를 이용하고, (9.12)식의 t-통계량은 부록 <표 8>의 $\hat{\tau_\mu}$의 분포표를 이용하며, (9.13)식의 t-통계량은 부록 <표 8>의 $\hat{\tau_\tau}$의 분포표를 각각 이용한다.

4.공적분

개별적으로는 단위근을 갖는 불안정한 시계열이지만 그들 사이에 안정적인 시계열을 생성하는 선형결합이 존재할 경우 이들 사이의 선형결합 관계를 공적분 관계라고 한다.

다음의 (9.14)식 및 (9.15)식의 간단한 이변량 방정식 체계를 이용하여 공적분의 개념을 설명해 볼 수 있다.

$$x_{1t} = \gamma_0 + \gamma_1 x_{2t} + u_{1t} \tag{9.14}$$

$$x_{2t} = \delta_0 + x_{2t-1} + u_{2t} \tag{9.15}$$

(9.15)식에서 $\Delta x_{2t} = \delta_0 + u_{2t}$이므로 x_{2t}는 단위근을 가진 시계열 즉, $x_{2t} \sim I(1)$이다.

(9.14)식을 차분하면 (9.16)식과 같게 된다.

$$\begin{aligned}\Delta x_{1t} &= \gamma_1 \Delta x_{2t} + \Delta u_{1t} \\ &= \gamma_1(\delta_0 + u_{2t}) + \Delta u_{1t} \\ &= \gamma_1\delta_0 + \gamma_1 u_{2t} + u_{1t} - u_{1t-1}\end{aligned} \tag{9.16}$$

(9.16)식은 (9.17)식과 같은 일반적인 IMA(1,1)과정으로 표현할 수 있으므로 x_{1t} 역시 단위근을 가진 시계열 즉$x_{1t} \sim I(1)$이다.

$$\Delta x_{1t} = \delta^* + v_t + \theta v_{t-1} \tag{9.17}$$

한편, (9.14)식에서 $x_{1t} - \gamma_0 - \gamma_1 x_{2t} = u_{1t}$이고 $u_{1t} \sim I(0)$이므로 즉, 각각 단위근을 갖는 두 시계열 x_{1t}와 x_{2t}의 선형결합이 안정적인 시계열 u_{1t}를 생성하므로 x_{1t}와 x_{2t}는 공적분관계에 있다고 한다.

다음의 공적분시스템을 시뮬레이션 해보면 <그림 9-13>과 같다

$x_{1t} = 4.5 + x_{2t} + u_{1t}$ (공적분 관계)

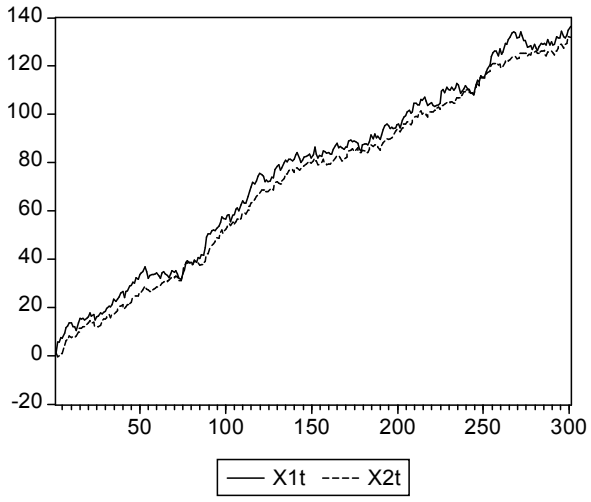
$x_{2t} = 0.5 + x_{2t-1} + u_{2t}, u_{2t} \sim N(0,1)$ (임의보행과정)

$u_{1t} = 0.9u_{1t-1} + v_t, v_t \sim N(0,1)$ (AR(1)오차)

그림에서 보여주고 있듯이 x_{1t}와 x_{2t}는 각각 우상향하는 확률추세를 갖는 시계열이므로 물론 정확한 것은 통계적 검정을 통해서 확인할 수 있겠지만 단위근을 가질 가능성이 매우 높다고 할 수 있다. 또한 두 시계열이 아주 유사한 움직임을 보이고 있으므로 두 시계열이 어떤 관계(공적분 관계)를 가질 가능성이 매우 높다.

변수 간에 존재하는 공적분 관계의 경제적 의미는 다음과 같다. 변수들 사이에 공적분 관계가 있다는 것은 변수들이 장기적으로 안정적인 균형관계(long-run equilibrium relation)가 있다는 것이다. 따라서 한 변수가 어떤 이유에서 공적분 관계에 있는 다른 변수와 안정적인 균형관계가 깨질 경우 이 상태가 장기간 지속되

<그림 9-13> 공적분 시계열의 시뮬레이션



지 않고 반드시 이전의 안정적인 관계로 회귀한다는 것이다.

공적분 검정(cointegration test)방법 역시 여러 가지가 있는데 Engle and Yoo의 검정방법과 Johansen의 검정방법이 대표적인 검정방법이다.

(1)Engle and Yoo 검정방법

Engle and Yoo의 검정방법은 각각 단위근을 가진 두 변수 y_t와 x_t에 대해 제1단계로 (9.18)식(이를 공적분 관계식이라고 함)을 최소자승법으로 추정한 후 제2단계로 (9.18)식의 잔차를 이용하여 (9.19)식으로 단위근 검정을 하는데 <표 9-2>에 주어진 검정통계량의 임계치에 근거하여 단위근 존재 여부를 판단한다. 단위근 검정결과 단위근이 존재하면 두 변수 y_t와 x_t 사이에는 공적분 관계가 존재하지 않으며, 단위근 검정결과 단위근이 존재하지 않으면 두 변수 y_t와 x_t 사이에는 공적분 관계가 존재한다.

$$y_t = \alpha + \beta x_t + \epsilon_t \tag{9.18}$$

$$\Delta\epsilon_t = \mu + (\rho - 1)\epsilon_t + \sum_{j=1}^{4} \gamma_j \Delta\epsilon_{t-j} + e_t \tag{9.19}$$

<표 9-2> 공적분 검정통계량의 임계치(ADF검정통계량인 경우)

모형 내 변수	표본크기	유의 수준	
		0.05	0.10
2	50	-3.29	-2.90
	100	-3.17	-2.91
	200	-3.25	-2.98
3	50	-3.75	-3.36
	100	-3.62	-3.32
	200	-3.78	-3.51
4	50	-3.98	-3.67
	100	-4.02	-3.71
	200	-4.13	-3.83
5	50	-4.15	-3.85
	100	-4.36	-4.06
	200	-4.43	-4.14

Engle and Yoo의 검정방법은 사용하기에 간편한 점이 있으나 여러 개의 공적분 관계가 있을 경우에는 이용될 수 없다. 예를 들면, 4변수로 모형이 구성된 경우 최대한 3개의 서로 독립인 공적분 벡터가 존재할 수 있는데 Engle and Yoo의 방법처럼 한 변수를 나머지 변수의 현재 값에 회귀시킨 후 그 잔차에 대해 단위근 검정을 실시하는 방법은 복수의 공적분 벡터가 존재할 경우 분산이 최소가 되는 1개의 공적분 벡터 선형결합만이 추정되기 때문에 적합하지 않게 된다. 또한 공적분 관계식에서 종속변수와 독립변수의 선정이 애매할 수 있다는 단점이 있다.

(2)Johansen 검정방법

각각 단위근을 가진 p개의 변수로 구성된 모형에서는 최대한 p-1개의 서로 독립적인 공적분 관계가 있을 수 있기 때문에 서로 독립적인 공적분 관계의 수가 몇 개인가를 검정하는 것 또한 중요하다.

최근에 개발된 공적분 검정으로 Johansen의 방법이 있는데 이 방법은 최우추정방법(maximum likelihood estimation)을 이용하여 공적분 관계를 추정하고 우도비검정(likelihood ratio test)을 바탕으로 공적분 관계식의 수를 결정할 수 있

는데 이 방법을 요한은 공적분 검정이라고 한다.

이 검정방법은 Dickey-Fuller의 단위근 검정을 다변량 경우로 확장한 것으로 볼 수 있다. 즉, DF검정에서 AR(1)과정인 단일시계열 y_t를 다음과 같이 나타낼 때 나타낼 때 $(\rho-1)=0$이면 y_t는 단위근을 갖는다.

$$\Delta y_t = (\rho-1)y_{t-1} + \epsilon_t$$

이와 유사하게 n개의 시계열로 구성된 벡터 x_t가 시차가 1인 벡터자기회귀모형 즉, VAR(1)일 때 이를 다음과 같이 표현하는 경우 Λ의 위수(rank)가 공적분 관계식의 수(r)가 되며 Λ의 위수에 따라 다음의 3가지 경우로 구분된다.

$$\Delta x_t = (A_t - I)x_{t-1} + v_t = \Lambda x_{t-1} + v_t$$

첫째, Λ의 위수(rank)가 n이면 x_t의 각 수준변수들이 안정적임을 뜻하므로 수준변수로써 VAR모형을 설정할 수 있다.

둘째, Λ의 위수(rank)가 0이면 r=0인 경우로서 이는 각 수준변수들이 불안정적인 시계열임과 동시에 수준변수 간에 장기적인 공적분 관계가 존재하지 않으므로 차분변수로 구성된 VAR모형을 설정하여야 한다.

셋째, Λ의 위수(rank)가 r(0<r<n)이면 각 수준변수들이 불안정적인 시계열이면서 수준변수 간에 r개의 공적분 관계가 존재하므로 오차수정모형으로 설정하여 추정해야 한다.

5.벡터자기회귀(Vector AutoRegressive: VAR) 모형

벡터자기회귀**(VAR)** 모형은 선험적인 경제이론을 배제한 상태에서 자료분석으로부터 경제시계열들 간의 관계에서 나타나는 특징적인 현상을 도출하고자 시도된 다변량시계열모형**(multivariate time series model)**이다.

전통적인 **Keynesian** 거시계량모형은 특정방정식의 식별을 위해 일부 변수들만 모형에 포함시키고 나머지 변수들은 모형에 포함하지 않는 소위 '**0**의 제약'을 가해야 한다. 예를 들어 다음과 같은 화폐수요함수를 식별하기 위해 소득**(y)**, 이자율**(i)**, 기대인플레이션(π^e), 화폐수요의 과거시차(m_{t-1})만을 설명변수로 포함시

키고 나머지 변수들은 모형에서 제외시킨다.

(화폐수요함수) $m_t = \alpha + \beta_0 y_t + \beta_1 i_t + \beta_2 \pi_t^e + \beta_3 m_{t-1} + e_t$

VAR 모형은 **Keynesian** 거시계량모형의 식별문제 부적합성에 대한 인식에서 출발하였으며 다음과 같은 특징이 있다.

첫째, 원칙적으로 이용가능한 모든 정보를 포함시킨다. 모형 내 모든 변수를 동일방정식에 동시에 포함시켜 특정변수의 변화를 자신은 물론 다른 모든 변수의 과거 값과 현재의 충격에 의해 설명되는 것으로 해석한다.

둘째, 기본적으로 변수선정에 제한을 두지 않으나 실제적으로는 문제가 되는 대표적인 변수들만을 선정하여 그들 간의 관계를 분석한다. 즉, 추정상의 편의를 위해 변수의 수와 시차수를 제약한다.

셋째, 변수선정에 있어 특별한 원칙은 없으며 특정 목적에 따라 달라질 수 있다.

넷째, 모형자체가 경제이론에서 출발한 것이 아니기 때문에(이를 **atheoretical macroeconomic model**이라고 함) 추정결과는 의미가 없다.

다섯째, 모형은 구조모형의 **reduced form** 모형과 비슷하며 모형의 추정방법은 **OLS** 또는 **Seemingly Unrelated Regression Estimation(SURE)**을 이용한다.

(1)VAR 모형

3변수와 **k**개의 시차를 갖는 **VAR** 모형은 다음의 **(9.20)**식과 같은 시차방정식체계로 구성된다.

$$\begin{bmatrix} X_{1t} \\ X_{2t} \\ X_{3t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + [A_1] \begin{bmatrix} X_{1t-1} \\ X_{2t-1} \\ X_{3t-1} \end{bmatrix} + \ldots + [A_k] \begin{bmatrix} X_{1t-k} \\ X_{2t-k} \\ X_{3t-k} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix} \qquad (9.20)$$

(9.20)식을 행렬 형태로 표현하면 다음의 **(9.21)**식과 같게 된다.

$$X_t = \mu + A_1 X_{t-1} + \ldots + A_k X_{t-k} + u_t$$

또는

$$X_t = \mu + A(L)X_t + u_t \qquad (9.21)$$

(9.21)식에서 A(L)은 시차연산자(lag operator) L의 다항식 행렬이다. 예를 들어 3변수가 2개의 시차를 갖는다고 할 때 A(L) 행렬은 다음의 (9.22)식과 같게 된다.

$$A(L) = \begin{bmatrix} a_{11}L + a_{12}L^2 & a_{13}L + a_{14}L^2 & a_{15}L + a_{16}L^2 \\ a_{21}L + a_{22}L^2 & a_{23}L + a_{24}L^2 & a_{25}L + a_{26}L^2 \\ a_{31}L + a_{32}L^2 & a_{33}L + a_{34}L^2 & a_{35}L + a_{36}L^2 \end{bmatrix} \qquad (9.22)$$

(2)VAR 모형의 분석도구

VAR 모형을 이용한 실증분석은 (9.21)식의 VAR 모형을 추정한 후 인과성 검정(causality test). 충격반응함수(impulse response function), 예측오차 분산분해(forecasting error variance decomposition) 등의 분석방법을 이용한다.

①인과성 검정

Granger는 인과성(causality)을 "만약 Y의 과거정보만을 가지고 Y를 예측할 때보다 X와 Y의 과거정보를 동시에 가지고 Y를 더욱 잘 예측할 수 있으면 X는 Y의 원인변수가 된다"라고 정의하였다.

예를 들어 생산(Y) 방정식에 포함된 통화량(X)의 과거 값들이 현재 Y에 영향을 주지 못하여 그 계수 값이 전부 0이 된다면 통화량은 물가의 원인변수가 되지 못한다고 한다.

인과성 검정에 대한 자세한 내용은 제8장 시차분포모형에서 인과성 검정을 참고하면 된다.

②충격반응함수

VAR 모형에서 도출되는 충격반응함수(impulse response function)란 (9.21)식

의 모형으로부터 도출된 이동평균모형으로 경제에 예상치 못한 변화(충격)가 주어졌을 때 모형내의 모든 변수들이 시간이 흐름에 따라 어떻게 각 충격에 반응하는 가를 나타내 주는 것이다.

(9.21)식은 다음의 (9.23)식과 같이 다시 나타낼 수 있다.

$$B(L)X_t = \mu + u_t \qquad (9.23)$$

단, $B(L) = I - A(L)$이고 I는 3×3 단위행렬(identity matrix)이다.

(9.23)식이 안정적이면 (9.24)식과 같은 이동평균모형(Moving Average Representation)으로 나타낼 수 있다.

$$X_t = \mu^* + C(L)u_t \qquad (9.24)$$

단, $\mu^* = B(L)^{-1}\mu$, $C(L) = B(L)^{-1}$이다.

(9.24)식에서 $C(L)$은 각 변수의 충격인 u_{1t}, u_{2t}, u_{3t}가 t기 및 무한미래에 걸쳐 각 변수에 미치는 영향을 나타내는 계수항의 행렬로 다음의 (9.25)식과 같다.

$$C(L) = \begin{bmatrix} C_{11}L & C_{12}L & C_{13}L \\ C_{21}L & C_{22}L & C_{23}L \\ C_{31}L & C_{32}L & C_{33}L \end{bmatrix} \qquad (9.25)$$

(9.25)식에서 $C(L)$의 각 원소는 시차연산자 L의 다항식으로 구성되어 있다. 예를 들어 $C_{12}(L) = c_{12}^0 L^0 + c_{12}^1 L^1 + c_{12} L^2 + \ldots$인데 이는 모형 내 두 번째 변수의 t기 충격이 첫 번째 변수에 미치는 t기(c_{12}^0), t+1기(c_{12}^1) 및 무한미래까지의 영향을 나타낸다.

그런데 (9.25)식의 $C(L)$을 그대로 이용하여 충격(원인)과 반응(결과)의 관계로 해석하면 문제가 생긴다. 왜냐하면 (9.21)식에서 도출된 잔차항이 서로 독립이 아니기 때문에 즉, u_t의 공분산행렬이 대각행렬(diagonal matrix)이 아니기 때문에 한 변수에서 충격이 발생하면 이로 인해 다른 변수들이 변하고 이것이 다시 환류(feedback)되어 처음 충격이 시작된 변수에 다시 영향을 주게 된다.

따라서 각 충격이 서로 독립적인 충격이 되도록 u_t를 (9.26)식과 같이 변환시

키면 되는데 이를 u_t의 공분산행렬을 직교화(orthogonalization)한다고 한다.

$$\epsilon_t = D(L)^{-1} u_t \qquad (9.26)$$

이러한 직교화 과정을 거치면 한 변수가 미치는 충격만을 분리하여 그 영향을 독립적으로 파악할 수 있게 되는데 그 이유는 (9.27)식을 만족시키는 즉, ϵ_t의 공분산행렬이 항등행렬이 되게 하는 $D(L)$을 찾으면 되기 때문이다.

$$E(\epsilon_t \epsilon_t') = D(L)^{-1} u_t u_t' D(L)^{-1\prime} = I \qquad (9.27)$$

이러한 독립적인 충격들을 보장해 줄 수 있는 행렬 $D(L)$은 $D(L)D(L)' = V$ (단, V는 교란항 u_t의 공분산행렬 즉, $E(u_t u_t') = V$)를 만족시키는 어떤 행렬이 가능하다. 즉, $D(L)D(L)' = V$를 만족시키는 행렬 $D(L)$는 유일하지 않은데 VAR 모형에서 흔히 사용하는 방법이 촐레스키분해(Choleski factorization)로 이 방법은 (9.26)식 및 (9.27)식을 만족시키는 하삼각행렬(lower triangular matrix)을 찾는다.

이렇게 구해진 하삼각행렬(이를 Choleski factor라고 함)을 이용하여 (9.24)식을 (9.28)식과 같이 변형시킬 수 있다.

$$X_t = F(L)\epsilon_t \qquad (9.28)$$

단, $F(L) = C(L)D(L)$, $\epsilon_t = D(L)^{-1} u_t$이며 상수항은 제외하였다.

(9.28)식에서 $F(L)$의 각 원소는 시차연산자 L의 다항식으로 구성되어 있으므로 $F(L)$이 나타내고 있는 충격반응관계는 한 변수의 충격이 다른 변수들에게 미치는 독립적인 영향 즉, 다른 변수들로부터의 환류되는 영향을 배제한 영향만을 나타낸다. 따라서 $F(L)$에서 얻어지는 충격반응관계는 특정변수에 대해 $D(L)$ 행렬에 나타나 있는 표준편차 크기에 해당하는 변화(충격)가 다른 변수에 미치는 독립적인 영향을 나타낸다.

③예측오차의 분산분해

예측오차의 분산분해(forecasting error variance decompositions)란 한 변수

의 변화를 설명함에 있어 모형 내 각 충격이 설명하는 비율로 표시한 것이다. 따라서 예측오차의 분산분해를 이용하면 한 변수의 변화를 설명함에 있어 모형 내 각 충격의 상대적 중요도를 측정할 수 있다. 즉, 예측오차의 분해란 한 변수의 변화에 관한 예측오차를 각 변수들에 의해서 발생하는 비율로 분할하는 것으로 이를 이용하여 한 변수의 변화를 설명함에 있어 모형 내 각 충격의 상대적 중요도를 측정할 수 있다.

(9.28)식을 모형 내 첫 번째 변수인 x_{1t}에 대해 풀어 쓰면 다음의 **(9.29)**식과 같다.

$$x_{1t} = F_{11}(L)\epsilon_{1t} + F_{12}(L)\epsilon_{2t} + F_{13}(L)\epsilon_{3t} \qquad \textbf{(9.29)}$$

(9.28)식에서 $F(L)$의 각 원소는 시차연산자 **L**의 다항식으로 구성되어 있으므로 예를 들어 $F_{11}(L) = h_{11}^0 L^0 + h_{11}^1 L^1 + h_{11} L^2 + \ldots$.이다.

x_1의 **t+k**기의 값, 즉 **k**기 이후의 값인 x_{1t+k}은 다음의 **(9.30)**식과 같다.

$$\begin{aligned} x_{1t+k} &= F_{11}(L)\epsilon_{1t+k} + F_{12}(L)\epsilon_{2t+k} + F_{13}(L)\epsilon_{3t+k} \\ &= f_{11}^0\epsilon_{1t+k} + f_{11}^1\epsilon_{1t+k-1} + f_{11}^2\epsilon_{1t+k-2} + \ldots + f_{11}^k\epsilon_{1t} + f_{11}^{k+1}\epsilon_{1t-1} + \ldots \\ &\quad + f_{12}^0\epsilon_{2t+k} + f_{12}^1\epsilon_{2t+k-1} + f_{12}^2\epsilon_{2t+k-2} + \ldots + f_{12}^k\epsilon_{2t} + f_{12}^{k+1}\epsilon_{2t-1} + \ldots \\ &\quad + f_{13}^0\epsilon_{3t+k} + f_{13}^1\epsilon_{3t+k-1} + f_{13}^2\epsilon_{3t+k-2} + \ldots + f_{13}^k\epsilon_{3t} + f_{13}^{k+1}\epsilon_{3t-1} + \ldots \end{aligned} \qquad \textbf{(9.30)}$$

(9.30)식으로부터 x_{1t+k}에 대한 예측치는 다음의 **(9.31)**식과 같다.

$$\begin{aligned} E_t(x_{1t+k} \mid x_{1t}, x_{1t-1}, \ldots) = {} & f_{11}^{k+1}\epsilon_{1t-1} + f_{11}^{k+2}\epsilon_{1t-2} + \ldots \\ & + f_{12}^{k+1}\epsilon_{2t-1} + f_{12}^{k+2}\epsilon_{2t-2} + \ldots \\ & + f_{13}^{k+1}\epsilon_{3t-1} + f_{13}^{k+2}\epsilon_{3t-2} + \ldots \end{aligned} \qquad \textbf{(9.31)}$$

(9.30)식 및 **(9.31)**식으로부터 x_{1t+k}에 대한 예측오차는 다음의 **(9.32)**식과 같이 도출된다.

$$x_{1t+k} - E_t(x_{1t+k} \mid x_{1t}, x_{1t-1}, ..) = f_{11}^0\epsilon_{1t+k} + f_{11}^1\epsilon_{1t+k-1} + f_{11}^2\epsilon_{1t+k-2} + .. + f_{11}^k\epsilon_{1t}$$

$$+f_{12}^{0}\epsilon_{2t+k}+f_{12}^{1}\epsilon_{2t+k-1}+f_{12}^{2}\epsilon_{2t+k-2}+..+f_{12}^{k}\epsilon_{2t}$$
$$+f_{13}^{0}\epsilon_{3t+k}+f_{13}^{1}\epsilon_{3t+k-1}+f_{13}^{2}\epsilon_{3t+k-2}+..+f_{13}^{k}\epsilon_{3t} \quad (9.32)$$

따라서 x_{1t+k}에 대한 예측오차의 분산은 **(9.33)**식과 같다.

$$\sum_{k=0}^{k}\sum_{j=1}^{3}(h_{1j}^{k})^2\sigma_j^2=\sigma_1^2((f_{11}^{0})^2+(f_{11}^{1})^2+(f_{11}^{2})^2+(f_{11}^{k})^2)$$
$$+\sigma_2^2((f_{12}^{0})^2+(f_{12}^{1})^2+(f_{12}^{2})^2+(f_{12}^{k})^2)$$
$$+\sigma_3^2((f_{13}^{0})^2+(f_{13}^{1})^2+(f_{13}^{2})^2+(f_{13}^{k})^2) \quad (9.33)$$

(9.33)식으로부터 x_{1t+k}의 예측오차의 분산 중에서 **j**번째 변수가 차지하는 비율은 **(9.34)**식과 같이 계산된다.

$$\frac{\sum_{k=0}^{k}(h_{1j}^{k})^2\sigma_j^2}{\sum_{k=0}^{k}\sum_{j=1}^{3}(h_{1j}^{k})^2\sigma_j^2}\times 100 \quad (9.34)$$

(9.34)식을 각 변수에 대해 계산함으로써 특정변수의 미래를 예측하는데 있어서 모형 내 각 변수의 중요도를 평가할 수 있다.

6.오차수정모형(Error Correction Model: ECM)

만약에 X_t와 Y_t가 공적분 관계가 있다면 이들 간에는 장기적인 관계가 존재하므로 이를 모형에 포함할 필요가 있다. 변수들 간에 공적분 관계가 존재할 경우 차분변수로 구성된 **VAR** 모형은 잘못 설정된 모형이 된다. 이 경우 **VAR** 모형 내에 변수들 간의 장기적인 관계를 포함해야 되는데 이를 오차수정모형**(Error Correction Model: ECM)**이라고 한다.

(9.21)식과 같은 **VAR(k)** 모형은 **(9.23)**식과 같이 변형하여 다시 나타낼 수 있는데 이를 오차수정모형이라고 한다.

$$\Delta X_t = \Gamma_1{}'\Delta X_{t-1} + ... + \Gamma_{k-1}{}'\Delta X_{t-k+1} + \Pi' X_{t-1} + \mu' + u_t \quad (9.23)$$

단, $\Gamma_i' \equiv -(A_{i+1} + A_{i+2} + ... + A_k)(i = 1,2,...,k-1)$,
$\Pi' \equiv (I_n - A_1 - .. - A_k) = A(1)$이다.

이 때 n개의 시계열로 구성된 X_t 간에 r개의 공적분 관계가 존재한다는 것은 Π행렬을 다음의 (9.24)식과 같이 다시 쓸 수 있음을 의미한다.

$$\Pi = \alpha\beta' \quad (9.24)$$

여기서 α와 β는 각각 nxr행렬로서 α는 오차수정행렬이고 β는 r개의 공적분 벡터로 구성된 행렬이다. 그리고 α는 모형내의 모든 n개의 식에 대하여 각각의 공적분 벡터에 부여되는 조정계수(adjustment coefficient)행렬로서 이는 각 추정된 균형관계식으로 조정되는 평균속도를 의미하는바 계수가 크면 조정이 신속히 이루어지는 것을 뜻한다.

오차수정모형의 해석은 다음과 같다. 경제주체들이 경제변수들의 움직임을 볼 때 두 가지에 관심을 가진다고 생각할 수 있다. 첫째는 경제변수들이 어느 수준인가 하는 것이고 둘째는 경제변수들이 어떠한 증가율로 변화하고 있는가하는 것이다. 이렇게 볼 때 오차수정모형의 단기동학을 나타내는 $\Gamma(i = 1,2,..k-1)$는 증가율의 변화를, 데이터의 장기균형관계를 나타내는 ΠΠ는 변수들의 수준을 포착해 주고 있다.

만약에 x_t와 y_t가 각각 단위근을 가지고 있고 x_t와 y_t가 공적분 관계가 있다면 즉, $z_t = y_t - \beta x_t$가 안정적인 시계열이라면 x_t와 y_t는 Granger표기정리(Granger representation theorem)에 의해 다음과 같은 오차수정모형으로 표시된다.

$$\Delta x_t = \mu_1 + \sum_{i=1}^{q} \alpha_i \Delta x_{t-i} + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \pi_1 z_{t-1} + \epsilon_{1t}$$

$$\Delta y_t = \mu_2 + \sum_{i=1}^{q} \delta_i \Delta x_{t-i} + \sum_{i=1}^{p} \gamma_i \Delta y_{t-i} + \pi_2 z_{t-1} + \epsilon_{2t}$$

실습 : AR 모형 추정 및 예측, 단위근 및 공적분 검정, VAR 모형 및 ECM 모형

1.EViews에서 다음을 입력하면 다음의 **AR(1)** 모형을 이용하여 가상적인 시계열을 생성할 수 있는데 **<그림 9-2-1>**이 이를 나타내고 있고, 시계열을 이용하여 **ACF** 및 **PACF**를 계산해 보면 **<그림 9-9-1>**와 같다.

$$y_t = 2 + 0.9y_{t-1} + e_t$$

```
smpl @first @first
series y_sim_1 = 0
smpl @first+1 @last
series y_sim_1 = 2+0.9*y_sim_1(-1)+nrnd
```

2.EViews에서 다음을 입력하면 다음의 **AR(1)** 모형을 이용하여 가상적인 시계열을 생성할 수 있는데 **<그림 9-2-2>**가 이를 나타내고 있고, 시계열을 이용하여 **ACF** 및 **PACF**를 계산해 보면 **<그림 9-9-2>**와 같다.

$$y_t = 2 - 0.9y_{t-1} + e_t$$

```
smpl @first @first
series y_sim_2 = 0
smpl @first+1 @last
series y_sim_2 = 2-0.9*y_sim_2(-1)+nrnd
```

3.EViews에서 다음을 입력하면 다음의 **AR(2)** 모형을 이용하여 가상적인 시계열을 생성할 수 있는데 **<그림 9-3-1>**이 이를 나타내고 있고, 시계열을 이용하여 **ACF** 및 **PACF**를 계산해 보면 **<그림 9-10-1>**과 같다.

$$y_t = 0.5y_{t-1} + 0.3y_{t-2} + e_t$$

```
smpl @first @first+1
series y_sim_3 = 0
smpl @first+2 @last
series y_sim_3 = 0.5*y_sim_3(-1)+0.3*y_sim_3(-2)+nrnd
```

4.EViews에서 다음을 입력하면 다음의 AR(2) 모형을 이용하여 가상적인 시계열을 생성할 수 있는데 <그림 9-3-2>가 이를 나타내고 있고, 시계열을 이용하여 ACF 및 PACF를 계산해 보면 <그림 9-10-2>와 같다.

$$y_t = 7 + 1.4y_{t-1} - 0.8y_{t-2} + e_t$$

```
smpl @first @first+1
series y_sim_4 = 0
smpl @first+2 @last
series y_sim_4 = 7+1.4*y_sim_4(-1)-0.8*y_sim_4(-2)+nrnd
```

5.EViews에서 다음을 입력하면 다음의 MA(1) 모형을 이용하여 가상적인 시계열을 생성할 수 있는데 <그림 9-5-1>이 이를 나타내고 있고, 시계열을 이용하여 ACF 및 PACF를 계산해 보면 <그림 9-11-1>과 같다.

$$y_t = 2 + e_t + 0.8e_{t-1}$$

```
smpl @first @first
series u_sim_1 = 0
smpl @first @last
series error_sim = nrnd
smpl @first+1 @last
series u_sim_1 = 2+error_sim-0.8*error_sim(-1)
```

6.EViews에서 다음을 입력하면 다음의 MA(2) 모형을 이용하여 가상적인 시계열을

생성할 수 있는데 <그림 9-5-2>가 이를 나타내고 있고, 시계열을 이용하여 ACF 및 PACF를 계산해 보면 <그림 9-11-2>와 같다.

$$y_t = 2 + e_t + 0.6e_{t-1} - 0.3e_{t-2}$$

```
smpl @first @first
series u_sim_1 = 0
smpl @first @last
series error_sim = nrnd
smpl @first+2 @last
series u_sim_2 = 2+error_sim+0.6*error_sim(-1)-0.2*error_sim(-2)
```

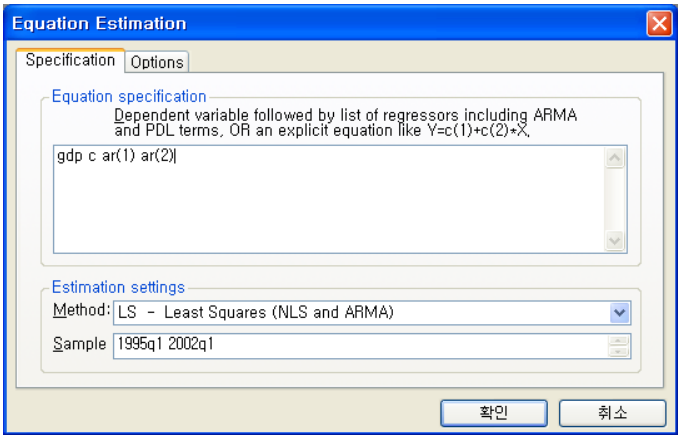
7.EViews에서 다음을 입력하면 다음의 ARMA(1,1) 모형을 이용하여 가상적인 시계열을 생성할 수 있는데 <그림 9-6>이 이를 나타내고 있고, 시계열을 이용하여 ACF 및 PACF를 계산해 보면 <그림 9-12>와 같다.

$$y_t = 0.2 + 0.2y_{t-1} + e_t + 0.6e_{t-1}$$

```
workfile corr_samples u 100
rndseed 12311 2
series e = nrnd
smpl @first @first
series arma = e
smpl @first+1 @last
series arma = 0.2+0.2*arma(-1)+e+0.3*e(-1)
```

1995년 1/4분기부터 2001년 1/4분기까지 한국의 GDP와 총소비(consume) 자료를 이용하여 GDP를 AR 모형으로 가정하고 모형을 추정하고 예측해 보고자 한다. GDP가 AR(2) 모형이라고 가정하고 방정식 추정 대화상자에서 <그림 9-A1>과 같이 입력하여 확인을 클릭하면 <그림 9-A2>와 같은 AR(2) 모형의 추정결과를 나타내 준다.

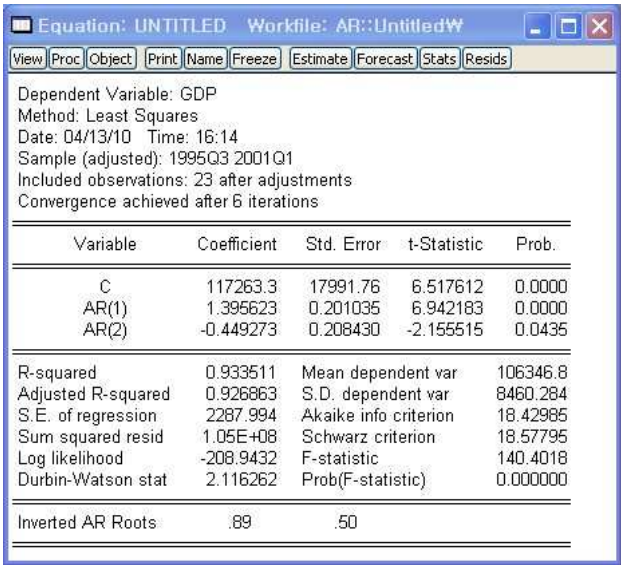
<그림 9-A1> AR(2) 모형 추정식



AR(2) 모형의 추정회귀식은 다음과 같다.

$$\hat{y_t} = 117263.3 + 1.395623 y_{t-1} - 0.449273 y_{t-2}$$

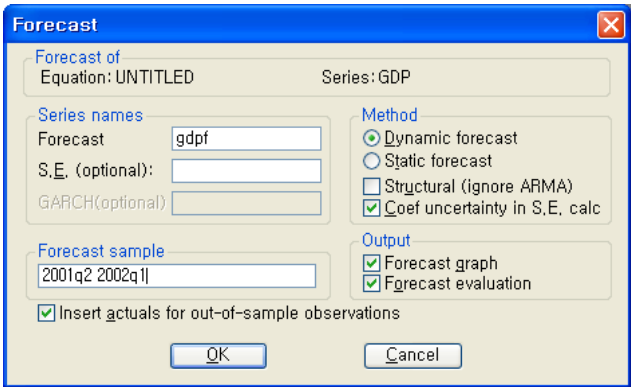
<그림 9-A2> AR(2) 추정 결과



예측을 하기 위해서는 예측기간을 설정해 주어야 하는데 예를 들어 2001년 2/4분기부터 2002년 1/4분까지 예측하고자 하면 workfile window에서 Proc-structure/resize current page를 선택한 후 2002q1까지 연장시킨다.

추정 결과 창에서 **Forecast**를 클릭하면 나타나는 예측 대화상자에서 **Forecast sample**에 **<그림 9-A3>**과 같이 입력하고 **OK**를 클릭하면 **<그림 9-A4>**와 같은 예측치를 계산해 준다.

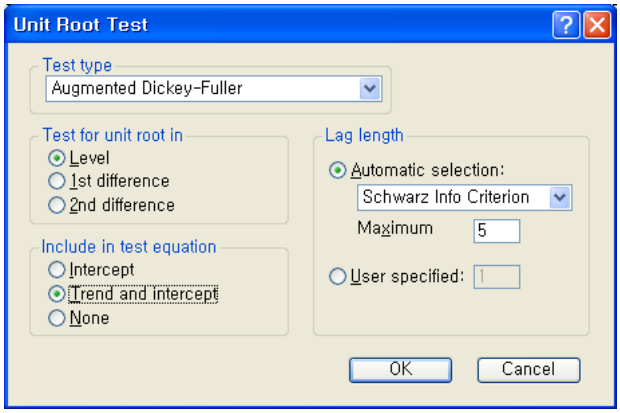
<그림 9-A3> 예측치 변수 및 예측기간 설정



<그림 9-A4> 예측치

단위근 검정을 하기 위하여 **LGDP(GDP의 로그치)**를 선택하고 **View-Unit Root Test**를 선택하면 나타나는 단위근 검정 대화상자에서 **<그림 9-A5>**와 같이 입력하고 **OK**를 클릭하면 **<그림 9-A6>**과 같은 **ADF** 단위근 검정 결과가 나타난다.

<그림 9-A5> ADF 단위근 검정



<그림 9-A6>의 단위근 검정 결과는 **(9.13)**식에서 **p=1**인 경우(시차의 수는 자동으로 선택하도록 하였음)로 $\rho = 1$을 검정하는 **t-**통계량의 값이 **-2.191656**이고 **5%** 유의수준에서의 임계치가 **-3.622033**이므로 단위근이 있는 것으로 나타난다.

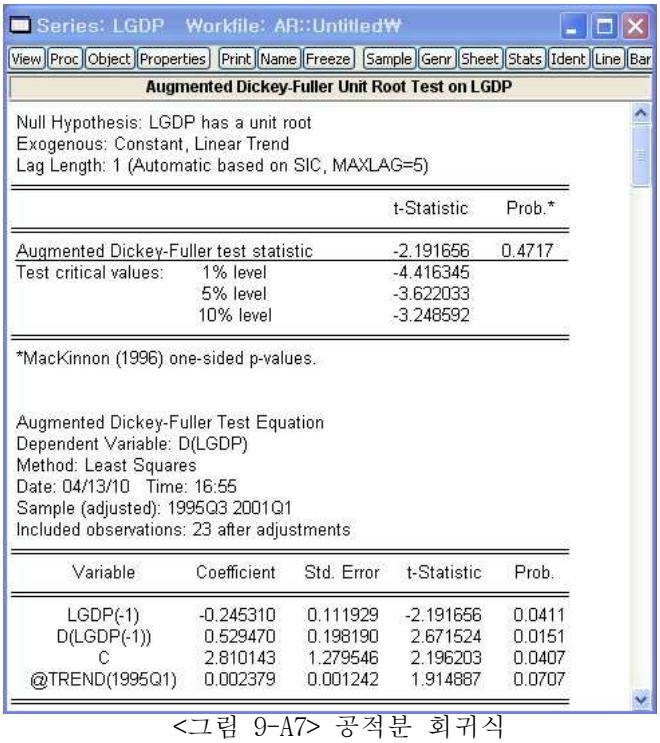
한편, **LCONSUMEP(CONSUME**의 로그치**)**에 대한 단위근 검정에서 역시 단위근이 있는 것으로 나타났다. 따라서 두 변수 간에 공적분 관계가 있는 지를 검정해 보기로 하자.

먼저 **(9.18)**식과 같은 공적분 회귀식을 추정하기 위하여 **<그림 9-A7>**과 같이 입력하고 추정하면 잔차항이 계산된다. 이 잔차항을 다음과 같이 변환시켜 **residyc**를 얻는다.

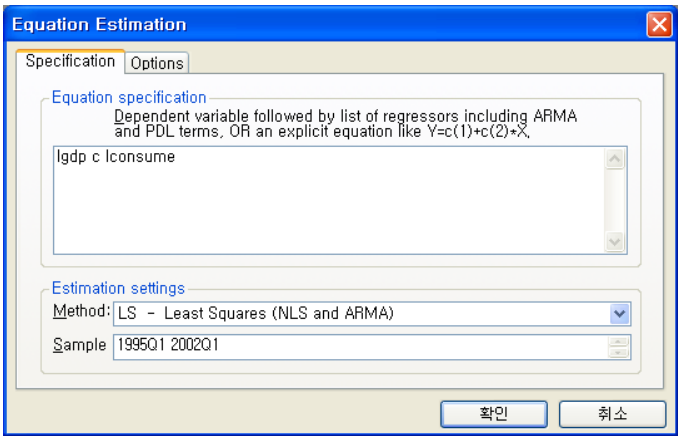
```
genr resyc = resid
```

(9.19)식과 같이 **resiyc**에 대해 단위근 검정을 실시하면 **<그림 9-A8>**과 같은 검정 결과를 주는데 **residyc**가 단위근이 있는 것으로 나타나 **GDP**와 **CONSUME** 간에는 공적분 관계가 없는 것으로 해석한다. 즉, **GDP**와 **CONSUME** 간에는 장기적인 균형관계가 존재하지 않는 것으로 나타났다.

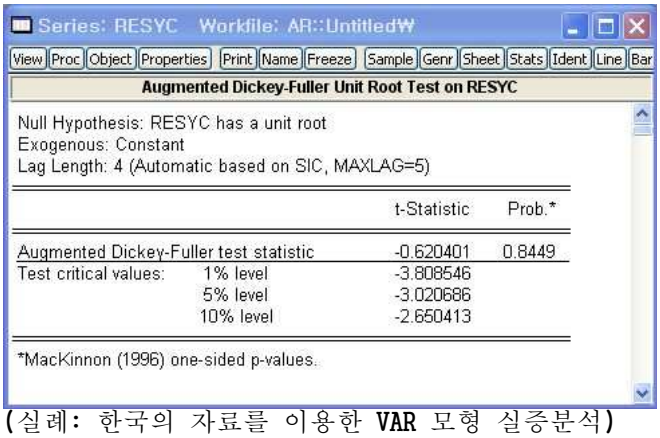
<그림 9-A6> ADF 단위근 검정 결과



<그림 9-A7> 공적분 회귀식



<그림 9-A8> 잔차항의 ADF 단위근 검정 결과



(실례: 한국의 자료를 이용한 VAR 모형 실증분석)

-변수: 실질GNP(y), GNP Deflator(p), 총통화(M)

-자료: 차분변수(전기대비증가율)

-추정기간: 1977:1 - 1993:2

-시차=2

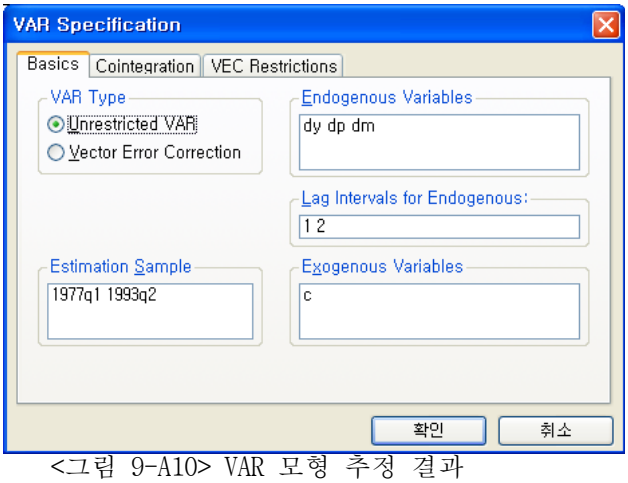
-모형: VAR(2)

$$\begin{bmatrix} y_t \\ p_t \\ m_t \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} A_1 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ p_{t-1} \\ m_{t-1} \end{bmatrix} + \begin{bmatrix} A_2 \end{bmatrix} \begin{bmatrix} y_{t-2} \\ p_{t-2} \\ m_{t-2} \end{bmatrix} + \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \end{bmatrix}$$

먼저, 단위근 검정결과 모든 변수가 단위근을 가지고 있는 것으로 나타나 차분변수 즉 전기대비증가율을 **VAR** 모형에 이용하였다.

다음으로, 공적분 검정결과 세 변수 사이에는 공적분 관계가 없는 것으로 나타나 **<그림 9-A9>**와 같이 세 변수로 구성된 시차**=2**의 **VAR(2)** 모형을 설정하였다. 확인을 클릭하면 **VAR(2)** 모형을 추정한 결과를 **<그림 9-A10>**와 같이 나타내 준다.

<그림 9-A9> VAR 모형 설정



<그림 9-A10> VAR 모형 추정 결과

Vector Autoregression Estimates
Date: 04/15/10 Time: 10:55
Sample (adjusted): 1977Q4 1993Q2
Included observations: 63 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	DY	DP	DM
DY(-1)	-0.815781	0.036908	0.064983
	(0.13264)	(0.02375)	(0.01391)
	[-6.15038]	[1.55386]	[4.67037]
DY(-2)	-0.335892	0.031414	-0.017608
	(0.18448)	(0.03304)	(0.01935)
	[-1.82077]	[0.95089]	[-0.90992]
DP(-1)	-0.516077	-0.012978	0.205594
	(0.72876)	(0.13050)	(0.07645)
	[-0.70816]	[-0.09945]	[2.68939]
DP(-2)	1.456372	-0.028864	0.066629
	(0.75009)	(0.13433)	(0.07868)
	[1.94159]	[-0.21488]	[0.84679]
DM(-1)	0.874104	0.320425	0.518781
	(1.32129)	(0.23661)	(0.13860)
	[0.66155]	[1.35421]	[3.74294]
DM(-2)	-3.851660	0.484703	-0.230548
	(1.09347)	(0.19582)	(0.11471)
	[-3.52241]	[2.47529]	[-2.00992]
C	0.166439	-0.017097	0.026562
	(0.05453)	(0.00976)	(0.00572)
	[3.05247]	[-1.75092]	[4.64393]

즉, 추정된 **VAR(2)** 모형은 각각 다음과 같다.

$$\Delta y_t = 0.166 - 0.815\Delta y_{t-1} - 0.335\Delta y_{t-2} - 0.516\Delta p_{t-1} + 1.456\Delta p_{t-2} + 0.874\Delta m_{t-1} - 3.851\Delta M_{t-2}$$

$$\Delta p_t = -0.017 + 0.036\Delta y_{t-1} + 0.031\Delta y_{t-2} - 0.012\Delta p_{t-1} - 0.028\Delta p_{t-2} + 0.032\Delta m_{t-1} + 0.484\Delta M_{t-2}$$

$$\Delta m_t = 0.026 + 0.064\Delta y_{t-1} - 0.017\Delta y_{t-2} + 0.205\Delta p_{t-1} + 0.066\Delta p_{t-2} + 0.518\Delta m_{t-1} - 0.23\Delta M_{t-2}$$

VAR(2) 모형을 추정한 후 **View-Residuals-Covariance Matrix**를 선택하면 <그림 9-A11>과 같은 잔차항의 공분산행렬을 나타내 준다. 이것이 교란항의 공분산행렬 V이며 각 충격들은 서로 독립적인 충격들이 아니다.

<그림 9-A11> 잔차항의 공분산행렬

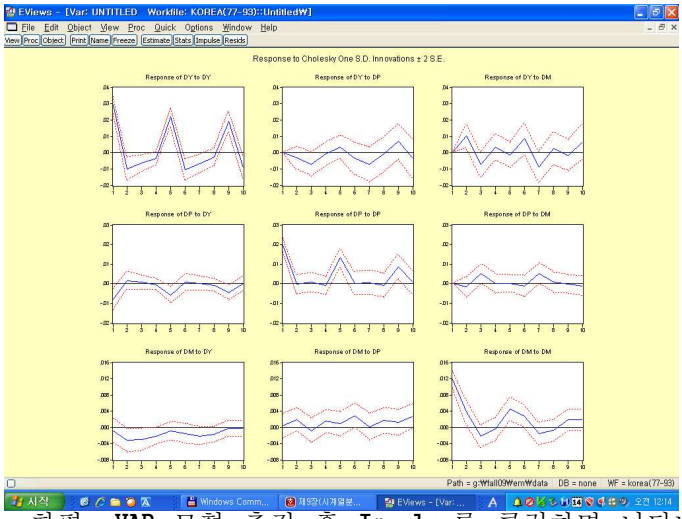
<그림 9-All>의 공분산행렬로부터 다음과 같은 하삼각행렬(촐레스키 요인)을 찾을 수 있다.

$$D = \begin{bmatrix} 0.153957 & 0 & 0 \\ -0.006208 & 0.026862 & 0 \\ 0.005343 & 0.001207 & 0.015193 \end{bmatrix}$$

위의 하삼각행렬은 아래에서 논의할 충격반응함수에서 각 충격에 대한 모형 내 변수의 t기의 반응을 나타내며 이 촐레스키 요인을 통해 $DD' = V$와 $D^{-1}VD^{-1\prime} = I$가 성립함을 확인할 수 있다.

VAR 모형 추정 후 **Impulse**를 클릭하면 나타나는 대화상자에서 **Multiple Graphs**를 선택하여 확인을 누르면 **<그림 9-Al2>**와 같이 모형 내 각 충격에 대한 세 변수의 충격반응함수를 그림으로 보여준다.

<그림 9-A12> 충격반응함수(그림)



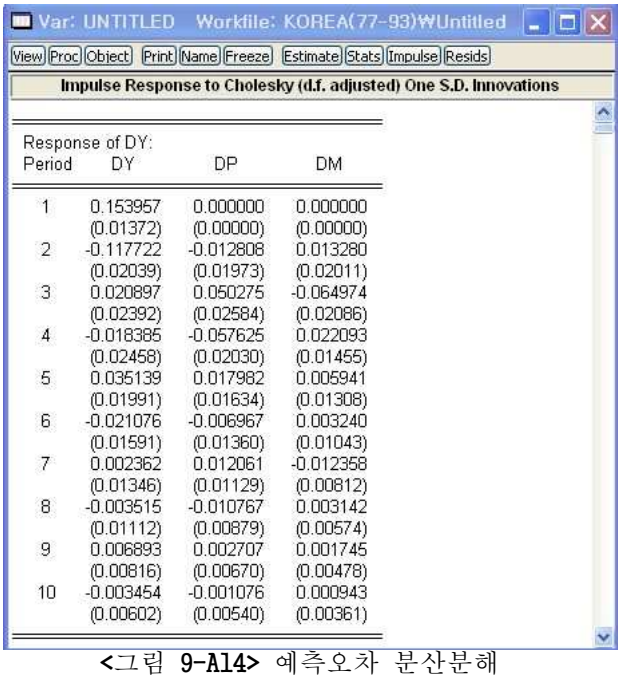
한편, VAR 모형 추정 후 Impulse를 클릭하면 나타나는 대화상자에서 Table을 선택하여 확인을 누르면 <그림 9-A13>과 같이 모형 내 각 충격(생산충격, 물가충격, 통화충격)에 대한 Y(여기서는 전기대비 총생산(GDP)증가율을 나타냄)의 반응을 나타내 준다. <그림 9-A13>은 총생산 충격에 대한 모형 내 변수의 시간 흐름에 따른 반응만을 나타내고 있으나 실제로는 각 충격에 대한 모형 내 변수의 충격반응을 모두 나타내 주고 있다.

예를 들어 독립적인 통화충격(여기서는 총통화의 전기대비 증가율)이 총생산증가율에 미치는 영향을 보면 다음과 같다.

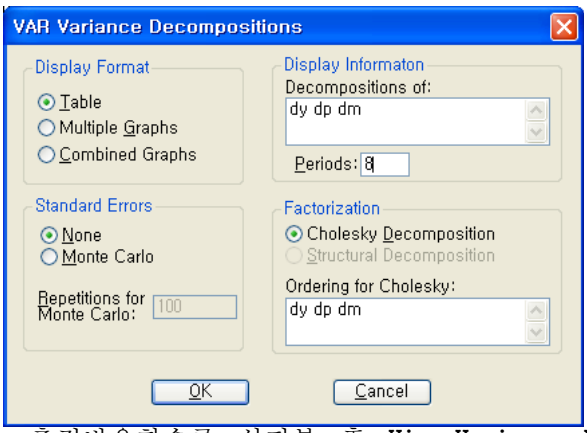
$$\Delta y_t = 0\epsilon_{3t} + 0.013\epsilon_{3t-1} - 0.064\epsilon_{3t-2} + \cdots.$$

위 충격반응함수에 대한 해석은 다음과 같다. 위의 D행렬을 보면 통화충격의 표준편차는 0.015193이므로 이는 1.52%에 해당한다. 따라서 통화량이 예상 외로 1.52% 증가하면 1기 후의 GDP증가율은 예상보다 1.3% 증가하고, 2기 후에는 6.4% 감소하는 것으로 해석하면 된다.

<그림 9-A13> 충격반응함수(표)



<그림 9-A14> 예측오차 분산분해

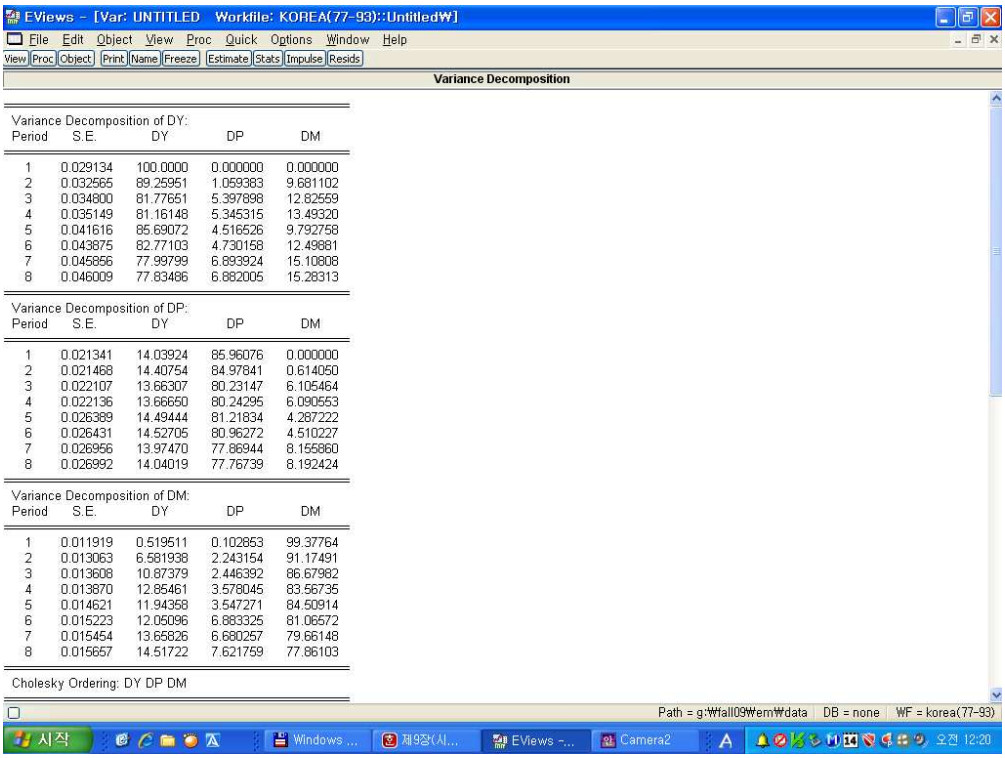


충격반응함수를 살펴본 후 **View-Variance Decompositions**를 클릭하면 **<그림 9-A14>**와 같이 예측오차 분산분해 대화상자가 나타는데 **OK**를 클릭하면 **<그림 9-A15>**와 같은 결과를 보여준다.

예를 들어 총생산증가율의 **8**분기 시차에 해당하는 **77.83, 6.88, 15.28**의 의미

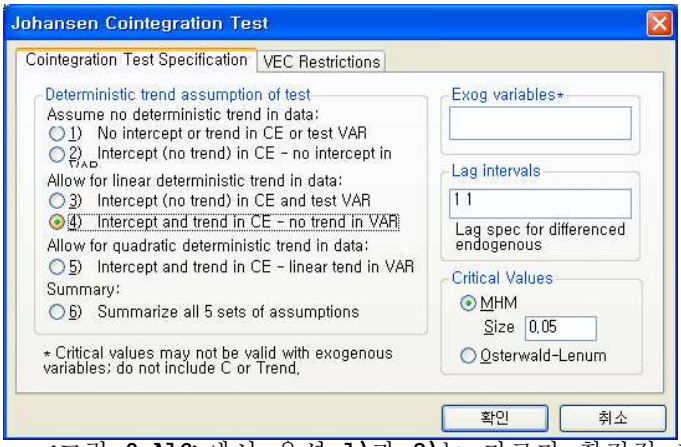
는 앞으로 8분기(2년) 후의 총생산증가율을 3변수 VAR 모형으로 예측했을 때 발생하는 예측오차의 분산을 100%라고 하면, 그 중 77.83%는 총생산 자체의 알 수 없는 내재적인 변화에 의해 발생하고 6.88%는 물가 충격에 의해 발생하며 15.28%는 총통화의 예상치 못한 충격 때문에 발생하는 것으로 해석한다.

<그림 9-A15> 예측오차 분산분해 결과



제주지역 농림어업 GRDP(ljl), 전국의 농림어업 GRDP(lkl), 제주지역 GRDP(ljt) 간에 공적분 존재 여부를 요한슨 검정을 통해 살펴보기 위해 세 변수를 선정한 후 View-Conintegration Test를 선택하면 <그림 9-A16>과 같은 공적분 검정 대화상자가 나타난다.

<그림 9-A16> 요한슨 공적분 검정



<그림 9-A16>에서 옵션 1)과 2)는 자료가 확정적 추세가 없는 경우의 검정방법이며, 옵션 3)과 4)는 자료가 선형의 확정적 추세가 있는 경우의 검정방법이고, 옵션 5)는 자료가 2차의 확정적 추세가 있는 경우의 검정방법이다.

적절한 옵션을 선택한 후 확인을 누르면 <그림 9-A17>과 같은 공적분 검정 결과를 나타내 준다. trace-통계량과 max-통계량 모두 5% 유의수준에서 하나의 공적분 관계가 존재하고 있음을 보여주고 있다.

따라서 ECM 모형을 이용한 실증분석을 해야 하므로 <그림 9-A18>과 같이 세 변수로 구성된 시차=1의 ECM 모형을 설정한 후 모형 추정, 충격반응함수, 예측오차 분산분해 순서로 실증분석을 하면 된다.

ECM 모형의 추정 결과는 <그림 9-A18>과 같은데 그 해석은 다음과 같다.

먼저, 세 변수 간의 공적분 회귀식은 다음과 같다.

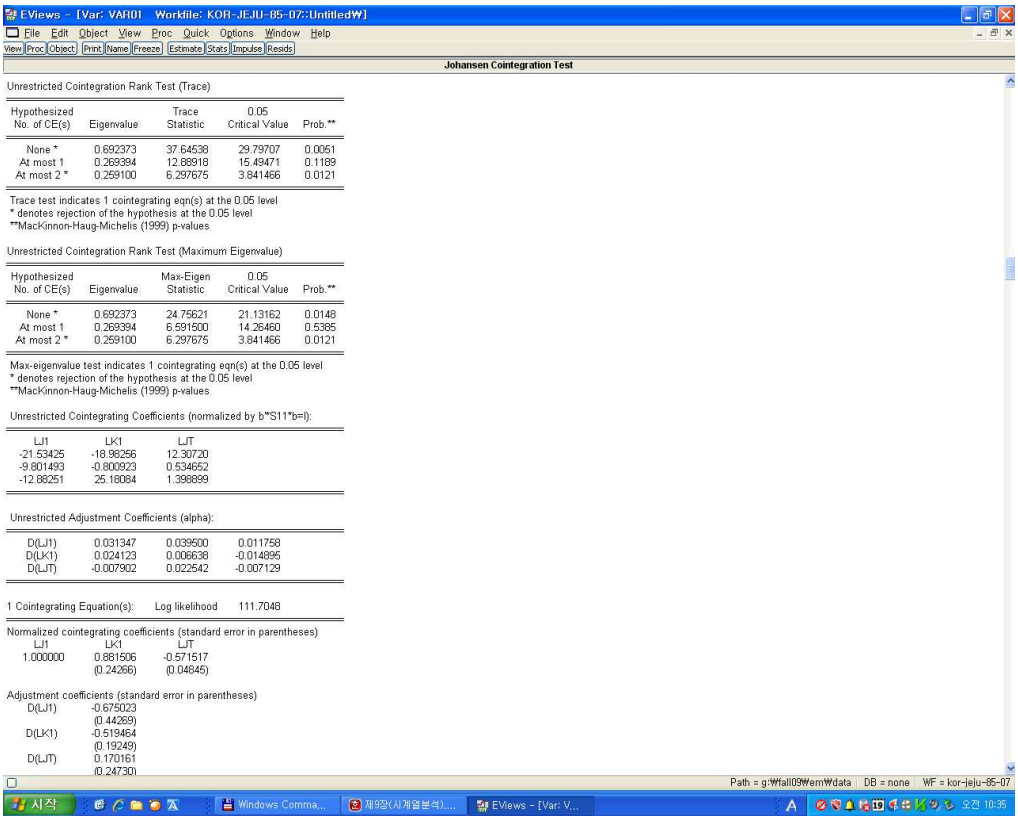
$$lj1_t = 19.985 - 0.8815lk1_t + 0.5715ljt_t$$

다음으로, 조정계수에 대한 해석은 다음과 같다. t-1기의 제주 농림어업 GRDP의 실제치가 제주 농림어업 GRDP, 전국 농림어업 GRDP 및 제주 GRDP 간에 존재하는 장기균형관계 균형치에서 벗어날 경우 t-1기 또는 t기에 67.5%가 제주 농림어업 GRDP의 변화에 반영된다는 것을 의미하는데 조정계수가 통계적으로 유의할 경우는 다음 기(t기)에 조정되고 유의하지 않을 경우는 같은 기(t-1기)에 조정된다.

ECM 모형 추정 후 충격반응함수, 예측오차 분산분해 등은 VAR 모형에서 살펴

본 바와 동일하게 수행하면 된다.

<그림 9-A17> 요한슨 공적분 검정 결과



EViews - [Var: VAR01 Workfile: KOR-JEJU-85-07::Untitled\]

File Edit Object View Proc Quick Options Window Help

View Proc Object Print Name Freeze Estimate Stats Impulse Resids

Johansen Cointegration Test

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.692373	37.64538	29.79707	0.0051
At most 1	0.269394	12.88918	15.49471	0.1189
At most 2 *	0.259100	6.297675	3.841466	0.0121

Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None *	0.692373	24.75621	21.13162	0.0148
At most 1	0.269394	6.591500	14.26460	0.5385
At most 2 *	0.259100	6.297675	3.841466	0.0121

Max-eigenvalue test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegrating Coefficients (normalized by b'*S11*b=I):

LJ1	LK1	LJT
-21.53425	-18.98256	12.30720
-9.801493	-0.800923	0.534652
-12.88251	25.18084	1.398899

Unrestricted Adjustment Coefficients (alpha):

D(LJ1)	0.031347	0.039500	0.011758
D(LK1)	0.024123	0.006638	-0.014895
D(LJT)	-0.007902	0.022542	-0.007129

1 Cointegrating Equation(s): Log likelihood 111.7048

Normalized cointegrating coefficients (standard error in parentheses)

LJ1	LK1	LJT
1.000000	0.881506	-0.571517
	(0.24266)	(0.04845)

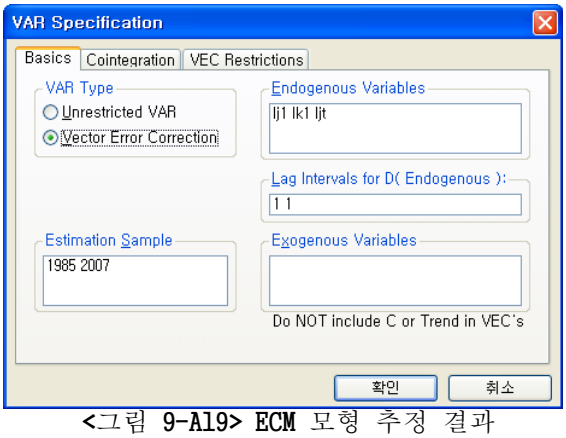
Adjustment coefficients (standard error in parentheses)

D(LJ1)	-0.675023
	(0.44269)
D(LK1)	-0.519464
	(0.19249)
D(LJT)	0.170161
	(0.24730)

Path = g:\fall09\em\data DB = none WF = kor-jeju-85-07

시작 Windows Comma... 제9장(시계열분석).... EViews - [Var: V... 오전 10:35

<그림 9-A18> ECM 모형 설정



<그림 9-A19> ECM 모형 추정 결과

Cointegrating Eq:	CointEq1
LJ1(-1)	1.000000
LK1(-1)	0.881506 (0.24266) [3.63264]
LJT(-1)	-0.571517 (0.04845) [-11.7968]
C	-19.98540

Error Correction:	D(LJ1)	D(LK1)	D(LJT)
CointEq1	-0.675023 (0.44269) [-1.52483]	-0.519464 (0.19249) [-2.69868]	0.170161 (0.24730) [0.68807]
D(LJ1(-1))	-0.450775 (0.31613) [-1.42593]	0.156213 (0.13746) [1.13644]	-0.415696 (0.17660) [-2.35387]
D(LK1(-1))	0.898611 (0.48462) [1.85427]	-0.093113 (0.21072) [-0.44188]	0.212099 (0.27073) [0.78344]
D(LJT(-1))	0.256285 (0.51021) [0.50231]	0.075768 (0.22185) [0.34153]	0.484171 (0.28502) [1.69870]
C	0.013339 (0.03313) [0.40266]	4.99E-05 (0.01440) [0.00347]	0.034219 (0.01851) [1.84909]

연습문제

1.ex9.wfl 자료는 1987:1-2000:4까지 한국의 분기별 자료로 실질총생산(Y:단위 10억원), 총통화(M:단위 10억원), 물가(P:소비자물가지수), 3년만기 회사채수익률(r)이다. 이를 이용하여 다음 물음에 답하되 다음 사항을 유의하라.

* log수준변수를 이용하여 분석할 때는 상수항 및 시간추세를 포함하라
* 전기대비증가율(근사치)변수를 이용하여 분석할 때는 상수항만 포함하라
* 단위근검정은 상수항 및 시간추세를 포함한 ADF로 하고 시차는 4로 하라
* Johansen 공적분검정에서 옵션 3을 선택하라
* VAR모형 이용시 시차는 2로 하고 ECM모형 이용시 시차는 1로 하라
* 모든 검정의 유의수준은 5%로 하라

①Long and Plosser는 어느 특정 시점에서의 충격이 대부분의 거시경제시계열에 지속적으로 영향을 준다는 사실을 밝힌바 있다. 한국의 경우에도 이러한 특징적인 사실들을 발견할 수 있는 지를 분석하여 그 결과를 설명하라.

②경제학자들은 “통화량의 증대는 단기적으로는 실질국민소득에 영향을 줄 수 있다”라는데 대부분 동의하고 있다. 이러한 주장이 한국경제에도 성립하는 지를 분석하여 설명하고 그 근거를 제시하라.

③경제학자들은 “통화량의 증대는 장기적으로는 실질국민소득에 영향을 줄 수 없다”라는데 대부분 동의하고 있다. 이러한 주장이 한국경제에도 성립하는 지를 분석하여 설명하고 그 근거를 제시하라.

④실질총생산, 물가 및 통화량 변수로 구성된 3변수 시계열모형을 이용하여 VAR모형 및 ECM모형 중 어느 모형이 적합 한 지 결정하되 그 근거를 제시하고 다음의 빈 칸을 태워 넣어라.

한국의 경우 통화량이 예상외로 ( )% 증가하면 실질총생산은 1기 후에 예상보다 ( )% ( )하고, 4기 후에는 예상보다 ( ) ( )한다.

제 10 장

패널분석

1.패널자료
2.패널모형
3.패널모형 추정
실습 : STATA 활용

제10장 패널분석

1.패널자료

일반적으로 경제분석에 사용되는 자료의 형태는 횡단면 자료(cross-section data), 시계열 자료(time series data), 패널자료(panel data) 등이 있다. 횡단면 자료란 특정 시점에서 개인, 회사, 지역 등 여러 대상을 관찰하여 기록한 자료이며, 시계열 자료란 특정 대상을 시간의 경과에 따라 반복적으로 측정한 자료이고, 패널자료란 개인, 회사, 지역 등 특정 대상을 시간의 경과에 따라 반복적으로 측정한 자료로서 횡단면 자료와 시계열 자료를 하나로 통합한 자료를 말한다.

패널자료는 pooled data(pooling of time series and cross-sectional observations), combination of time series and cross-section data, micropanel data, longitudinal data(a study over time of a variable or group of subjects)라고도 불리는 것처럼 횡단면 자료의 정보와 시계열 자료의 정보를 모두 이용할 수 있기 때문에 즉, 횡단면 분석 또는 시계열 분석 만으로 파악할 수 없는 추가적인 정보를 얻을 수 있기 때문에 연구자들이 실증분석에서 많이 이용하고 있다.

independently pooled cross section across time과 패널자료는 차이가 있는데 전자의 경우 서로 다른 시점에서 임의로 추출된(sampling randomly) 자료로서 이 자료의 특징은 다른 조건이 일정할 때 교란항의 상관관계가 없으므로 동일한 분포(identical distribution)가 될 가능성이 낮다는 것이며 이 자료를 사용하는 또 다른 이유는 표본의 크기를 늘림으로써 더 정확한 추정량을 얻고 검정통계량의 검정력을 높일 수 있기 때문이다. 후자의 경우 특정 시점에서 조사대상은 임의로 추출되지만 그 다음 조사시점에서는 동일한 조사대상에 대한 자료를 수집한다.

패널자료 또는 패널자료를 이용한 패널 분석은 다음과 같은 장점이 있다.

첫째, 개별적인 특이성(individual heterogeneity)을 통제할 수 있다. 개별적인 특이성을 통제하지 못할 경우 횡단면 분석이나 시계열 분석은 왜곡된 결과를 줄 수 있는데 패널 분석은 횡단면 분석이나 시계열 분석에서는 불가능한 개별특성효과(individual effect)와 시간특성효과(time effect)를 모두 통제할 수 있다.

둘째, 연구자에게 다양한 정보를 제공해 주며 다중공선성(multi-collinearity)

문제를 줄일 수 있다. 또한 관측치 수가 크기 때문에 자유도가 높아져 추정치의 효율성이 높아지고, 변동성(variability)을 제공해 주어 분석이 용이하다.

셋째, 조정의 동태성(dynamics of adjustment)을 가능하게 해 주어 횡단면 분석에서는 포착하기 힘든 다양한 변화의 포착이 가능하다.

넷째, 패널자료는 횡단면 분석이나 시계열 분석에서 포착하기 힘든 효과를 잘 측정할 수 있다.

다섯째, 패널자료는 횡단면 자료나 시계열 자료에 비해 복잡한 행태적 모형 구축 및 검증을 가능하게 해 준다.

여섯째, 패널자료는 개인, 기업, 정부 등 미시적 단위에서 수집되는 자료에서 발생하는 편의(bias)를 통제하게 해 준다.

2.패널모형

패널모형이란 패널자료가 가지고 있는 다양하고 풍부한 정보들을 가장 효율적으로 추출해 내는 분석기법으로서 시계열 분석과 횡단면 분석을 동시에 수행하는 계량분석모형을 말한다.

회귀모형을 설정할 때 모든 변수를 독립(설명)변수로 포함시킬 수 없으나 종속변수에 영향을 미침에도 불구하고 독립변수로 포함시키지 않은 변수(누락변수)가 있을 경우 추정된 모형의 통계적 추론은 위험할 수 있다. 즉, 이 경우 보통최소자승법(OLS)으로 추정하면 편의가 있는 추정량(biased estimator)을 얻게 된다. 그러나 패널모형을 이용하면 이러한 누락변수(omitted variable)에 대한 한계를 극복할 수 있다.

패널모형은 다음과 같은 장점이 있다.

첫째, 패널모형은 시계열 분석에서 발생하는 추정오차와 횡단면 분석에서 발생하는 추정오차를 통제할 수 있어서 횡단면 분석이나 시계열 분석에 비해 현실을 더 잘 분석할 수 있다.

둘째, 횡단면 분석이나 시계열 분석에서는 통제가 불가능한 관찰되지 않은 누락변수(unobservable omitted variable)에 대한 처리를 해 주기 때문에 제반 변수들에 대한 통제가 불가능한 사회과학연구에 매우 유용한 분석모형이다. 관찰되지 않은 누락변수를 제어하기 위해서 교란항에 대해서 개인(individual) 간에는 다르나 시간변동이 없는 변수, 시간변화에 따라 변동하나 개인 간에 차이가 없는 변수, 개인 간에도 차이가 있고 시간변화에 따라서도 변동하는 확률적 교란항으로 구분하여 다룬다.

패널자료를 이용한 회귀모형의 기본모형은 다음의 (10.1)식과 같이 나타낼 수 있다.

$$y_{it} = \alpha + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it}$$

또는

$$y_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k X_{kit} + \epsilon_{it}, (\text{단 } \alpha_i = \alpha + \mu_i) \quad (10.1)$$

패널모형은 교란항(u_{it})의 구조에 대한 가정에 따라 One-way Error Component Model과 Two-way Error Component Model로 나누어지는데 이를 오차구성모형(ECM: Error Component Model)이라고 한다.

One-way Error Component Model은 (10.2)식과 같다.

$$u_{it} = \mu_i + \epsilon_{it} \quad (10.2)$$

단, μ_i는 관측되지 않은 개별특성효과(unobservable individual effect)로 individual time-invariant variable이다.

한편, Two-way Error Component Model은 (10.3)식과 같다.

$$u_{it} = \mu_i + \lambda_t + \epsilon_{it} \quad (10.3)$$

단, λ_t는 관측되지 않은 시간 효과(unobservable time effect)로 time(or period) individual-invariant variable, ϵ_{it}는 횡단관측치에 의한 영향과 시계열에 의한 영향을 혼합한 확률적 교란항으로 individual time-varying variable이다.

또한 오차구성모형은 각각의 교란항을 고정된 상수로 볼 것인가 또는 확률변수로 볼 것인가에 따라 고정효과모형(FEM: Fixed Effect Model)과 확률효과모형(REM: Random Effect Model)로 구분된다.

Fixed Effect Model은 (10.4)식과 같다.

$$y_{it} = \alpha + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it}$$

$$u_{it} = \mu_i + \epsilon_{it}, \epsilon_{it} \sim iid(0, \sigma_\epsilon^2)$$

또는

$$y_{it} = \alpha_i + \sum_{k=1}^{K} \beta_k X_{kit} + \epsilon_{it}, (\text{단 } \alpha_i = \alpha + \mu_i) \quad (10.4)$$

단, μ_i는 관측되지 않은 횡단효과로서 상수로 취급하며, 교란항에 대한 가정은 다음과 같다.

$$E(\epsilon_{it}) = 0$$

$$E(\epsilon_{it}\epsilon_{js}) = \begin{cases} \sigma_\epsilon^2 (i = j, t = s) \\ 0 (otherwise) \end{cases}$$

$$cov(X_{it}, \mu_i) \neq 0$$

한편, Random Effect Model은 (10.5)식과 같다.

$$y_{it} = \alpha + \sum_{k=1}^{K} \beta_k X_{kit} + \mu_i + \epsilon_{it} \quad (10.5)$$

단, μ_i는 관측되지 않은 횡단효과로서 확률변수로 취급하며, 교란항에 대한 가정은 다음과 같다.

$$E(\epsilon_{it}) = 0$$

$$E(\epsilon_{it}\epsilon_{js}) = \begin{cases} \sigma_\epsilon^2 (i = j, t = s) \\ 0 (otherwise) \end{cases}$$

$$E(\mu_i \mu_j) = \begin{cases} \sigma_\mu^2 (i = j) \\ 0 (i \neq j) \end{cases}$$

$E(\epsilon_{it}u_j)=0(\forall i,t,j)$

$cov(X_{it},\mu_i)=0$

지금까지의 이상의 논의를 정리해 보면 패널모형은 <표 10-1>과 같이 구분된다.

<표 10-1> 패널모형의 분류

구분	Fixed Effect Model	Random Effect Model
One-way ECM	-시간의 흐름에 따라 변하지 않으며 관측되지 않는 특정한 변수가 횡단면 관측치에 잠재해 있음	-시간의 흐름에 따라 고정되지 않고 확률적으로 변하는 관측되지 않는 특정한 변수가 횡단면 관측에 잠재해 있음
Two-way ECM	-시간의 흐름에 따라 변하지 않으며 관측되지 않는 특정한 변수가 횡단면 관측치 및 시계열 관측치에 잠재해 있음	-개별특성 및 시간의 흐름에 따라 고정되지 않고 확률적으로 변하는 관측되지 않는 특정한 변수가 횡단면 관측치 및 시계열 관측치에 잠재해 있음

Fixed Effect Model의 장점은 개별특성효과를 구분하여 계수를 추정하는 것인데 이렇게 하기 위해 더미변수를 생성하는 과정에서 자유도를 잃게 되고 이로 인해 계수값 추정이 상대적으로 정확성을 잃게 될 수 있다.

Random Effect Model의 장점은 계수값 추정에 정확성이 떨어지는 위험이 작다는 장점이 있으나 개별특성효과가 독립변수와 상관관계가 없어야 한다는 다소 엄격한 가정을 하고 있다는 단점이 있다.

다양한 패널모형 중 어느 모형을 선택하는 것이 합리적인가 하는 문제는 다음과 같이 정리될 수 있다.

첫째, 시간불변의 개별특성효과(μ_i)가 독립변수들과 상관관계가 있으면 Fixed Effect Model을 선택하고, 상관관계가 없을 경우 Random Effect Model을 선택한다.

둘째, 실증분석에서 두 모형 중 어느 것이 더욱 적합한지에 대한 검정은 Hausman Specification Test를 활용할 수 있다.

셋째, Random Effect Model이 유효한 경우라도 Fixed Effect Model에 의한 추

정량은 일치추정량이기 때문에 시간불변의 개별특성효과(μ_i)와 독립변수와의 상관관계에 대한 확실한 정보가 없을 경우 Fixed Effect Model을 선호하는 경향이 있다.

3.패널모형 추정

패널자료(Pooled data)를 이용한 회귀모형(panel data regression model)은 다음과 같은 4가지가 있음

original formulation(pooling OLS): $y_{it} = \alpha + \beta' X_{it} + \epsilon_{it}$ (10.6)

deviation from the group means(within-groups): $y_{it} - \overline{y_{i.}} = \beta'(X_{it} - \overline{X_{i.}}) + \epsilon_{it} - \overline{\epsilon_{i.}}$ (10.7)

group means(between groups): $\overline{y_{i.}} = \alpha + \beta' \overline{X_{i.}} + \overline{\epsilon_{i.}}$ (10.8)

first-difference : $\Delta y_i = \beta' \Delta X_i + \Delta \epsilon_i$ (10.9)

첫째, (10.6) 모형의 경우 NT개의 자료로 OLS로 추정한다.

둘째, (10.7) 모형의 경우 종속변수와 설명변수들은 특정 횡단면 관측치에 대하여 그 시계열 상의 평균을 구해 각각의 값에서 빼 준 형태의 새로운 변수로 전환되는데 이와 같이 방정식 내부에서 일어난 변환을 내부변환(within transformation)이라고 한다. 설명변수가 시간에 따라 변하지 않는 변수(예를 들어, 성, 인종)라면 within transformation에 의해 모형에서 사라져 버린다. within transformation에 의한 추정방법을 최소자승더미변수(LSDV: Least Square Dummy Variable)모형에 의한 추정방법이라고도 하는데 그 이유는 아래에서 논의하고 있는 모형④에서 α_i를 마치 횡단더미변수로 취급해 추정하여도 동일한 결과를 얻을 수 있기 때문이다. 그러나 N이 클 경우 더미변수를 사용하는데 현실적인 어려움이 있을 수 있다.

셋째, (10.8) 모형의 경우 자료는 N개(group means)임에 유의해야 하며 individual heterogeneity(u_i)가 원래 모형에 있을 경우 여기서도 그대로 모형에 남아 있게 된다.

넷째, (10-9) 모형의 경우 시간의 변화에 따른 종속변수의 변화를 살펴봄으로

써 시간의 흐름에 따라 변하지 않으나 횡단면 관측치에 잠재해 있는 누락변수의 효과를 제거할 수 있다.

(10.6)-(10.9)식의 4가지 모형 모두 보통최소자승법(OLS)으로 추정할 수 있으며 이 경우 효율적인 추정량(efficient estimator)은 아니라 하더라도 일치추정량(consistent estimator)은 구할 수 있다.

서로 다른 시간에 측정된 independently pooled cross section 자료를 사용할 경우 time dummy를 이용하여 시간에 따라 절편이 달라지는 모형을 설정하거나 절편 및 기울기가 달라지는 모형을 설정한 후 추정하면 된다(실습 case ①-1).

또한 서로 다른 시간에 측정된 independently pooled cross section 자료를 사용할 경우 위의 경우와 같이 time dummy를 사용해도 되고 또한 특정 정책이나 사건의 발생을 기준으로 두 기간을 이전시기 및 이후시기(before and after)로 구분한 후 특정 정책이나 사건의 영향(impact)을 살펴볼 수 있다.

이 경우 연구자가 관심이 있는 treatment group과 이의 comparison group으로 구분한 후 이후시기(after)에서 treatment group과 comparison group의 평균의 차이를 구하고 또한 이전시기(before)에서 treatment group과 comparison group의 평균의 차이를 구한 후 이 두 개 평균차이의 차이를 구하면 그것이 특정 정책이나 사건의 영향이 되는데 이렇게 구하는 방법을 DID(difference-in-differences) 추정량이라고 한다(case ①-2).

이를 살펴보기 위해서 다음의 (10.9)와 같이 모형을 설정할 수 있는데 여기서 δ_1은 정책이나 사건의 영향을 측정한다.

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + other\ factors \qquad (10.9)$$

	before	after	after-before
control	β_0	$\beta_0 + \delta_0$	δ_0
treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
treatment-control	β_1	$\beta_1 + \delta_1$	δ_1

여기서 정책이나 사건의 영향을 측정하는 δ_1은 다음의 (10.10)식과 같이 추정되므로 DID estimator 또는 average treatment effect라고도 한다.

$$\hat{\delta_1} = (\overline{y}_{2,T} - \overline{y}_{2,C}) - (\overline{y}_{1,T} - \overline{y}_{1,C}) \qquad (10.10)$$

(1)Two-period 분석

서로 다른 두 기간(연속 기간이든 떨어져 있는 기간이든 상관없음)에 걸쳐 측정 된 자료를 사용하여 특정 정책이나 사건의 발생을 기준으로 두 기간을 이전시기 및 이후시기(before and after)로 구분한 후 특정 정책이나 사건이 종속변수의 변화에 미치는 영향을 살펴 볼 경우 관측되지 않은 영향이 시간에는 불변이더라도 실제로 개인별 또는 개별기업별로 다를 수 있음에도 불구하고 이를 무시함으로써 누락변수의 문제가 발생할 수 있는데 이를 해결할 수 있는 방법은 두 기간의 관측치의 차이를 구하는 것이며 이를 first-differenced(FD) estimator라고 한다(case ②-1). two periods인 경우 FD estimator와 Fixed Effect Model의 추정결과가 동일하다.

이를 살펴보기 위해서 종속변수에 영향을 미치는 관측되지 않은 요인(unobserved factors)을 두 가지로 구분하여 다음의 (10.11)식과 같이 모형을 설정할 수 있다.

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 X_{it} + \alpha_i + u_{it} \qquad (10.11)$$

(10.11)식에서 α_i는 시간에는 불변하지만 개인별 또는 개별기업별 종속변수에는 영향을 주는 관측되지 않은 요인으로 unobserved effect, fixed effect, unobserved heterogeneity 등으로 불리며, u_{it}는 시간에 따라 변하면서 개인별 또는 개별기업별 종속변수에 영향을 주는 관측되지 않은 요인으로 idiosyncratic error 또는 time-varying error라고 불린다. 이 모형을 pooled OLS로 추정할 경우 불편추정량 및 일치추정량을 얻을 수 없는데 그 이유는 α_i가 X_{it}와 상관관계가 있을 가능성이 매우 높기 때문이다.

위 모형에 의해 이후시기(t=2) 및 이전시기(t=1)로 구분한 후 그 차분을 구하면 다음의 (10.12)식과 같다.

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 X_{i2} + \alpha_i + u_{i2} (t = 2)$$

$$y_{i1} = \beta_0 + \beta_1 X_{i1} + \alpha_i + u_{i1} (t = 1)$$

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (X_{i2} - X_{i1}) + (u_{i2} - u_{i1})$$

또는

$$\Delta y_i = \delta_0 + \beta_1 \Delta X_i + \Delta u_i \qquad (10.12)$$

(10.12)식을 first-differenced equation이라 하고 이를 OLS로 추정하여 얻은 추정량을 first-differenced(FD) estimator이라고 하며 first-difference로 인해 α_i가 사라져 버리므로 consistent estimator가 된다.

한편 기간이 두 기간 이상일 경우 위와 유사한 방법을 사용할 수 있다(case ②-2). 이를 살펴보기 위해서 예를 들어 T=3의 경우를 가정해 보면 다음의 (10.13)식과 같은 모형을 설정할 수 있다.

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 X_{it1} + \ldots + \beta_k X_{itk} + \alpha_i + u_{it}, (T = 1,2,3) \qquad (10.13)$$

(10.13)식의 경우 T=3에서 T=1을 빼주고, T=3에서 T=2를 빼주면 다음의 (10.14)식과 같은 two-period로 전환된다.

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta X_{it1} + \ldots + \beta_k \Delta X_{itk} + \Delta u_{it}, (T = 2,3) \qquad (10.14)$$

위 모형은 두 개의 차분된 시간더미가 있는 모형으로 상수항이 없으며 t=2일 경우 $\Delta d2_t = 1, \Delta d3_t = 0$이고 t=3일 경우 $\Delta d2_t = -1, \Delta d3_t = 1$임에 유의해야 한다.

또한 위 모형의 경우 R^2의 계산이 불편하므로 다음과 같이 상수항을 포함한 모형으로 변화시킬 수 있으며 이를 pooled OLS로 추정하면 FD 추정량을 구할 수 있다.

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta X_{it1} + \ldots + \beta_k \Delta X_{itk} + \Delta u_{it}, (T = 2,3)$$

위 모형에서 $cov(\Delta u_{i2}, \Delta u_{i3}) \neq 0$ 즉, 자기상관의 문제를 해결해야 하며, FD 추정방법은 시간에 따라 변하지 않는 변수(예를 들면 면적 등)는 추정할 수 없다는 점과 차분으로 인한 정보의 손실이 있다는 단점이 있다.

(2)Fixed Effect Model

Fixed Effect Model은 가장 일반적인 모형(모형①)에서부터 단순한 모형(모형⑤)에 이르기까지 5가지로 분류할 수 있다.

<모형①> 회귀계수가 횡단 관측치 및 시간변화에 따라 모두 변하는 경우

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_{kit} X_{kit} + u_{it}, i = 1,2,...,N,\ t = i,2,...,T$$

단, i는 횡단면 상의 개별 관측치, t는 시간, N은 횡단관측치의 총합, T는 시계열기간을 나타낸다.

<모형②> 회귀계수가 횡단관측치에 따라 변하나 시간변화에는 일정한 경우

$$y_{it} = \alpha_{i} + \sum_{k=1}^{K} \beta_{ki} X_{kit} + u_{it}, i = 1,2,...,N,\ t = i,2,...,T$$

<모형③> 기울기계수(β)는 일정하나 절편(α)은 횡단 관측치 및 시간에 따라 변하는 경우

$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} \beta_{k} X_{kit} + u_{it}, i = 1,2,...,N,\ t = i,2,...,T$$

<모형④> 기울기계수(β)는 일정하나 절편(α)이 횡단관측치에 따라서만 변하는 경우

$$y_{it} = \alpha_{i} + \sum_{k=1}^{K} \beta_{k} X_{kit} + u_{it}, i = 1,2,...,N,\ t = i,2,...,T$$

<모형⑤> 모든 추정치의 값이 횡단 관측치 또는 시간변화에 관계없이 일정한 경우

$$y_{it} = \alpha + \sum_{k=1}^{K} \beta_{k} X_{kit} + u_{it}, i = 1,2,...,N,\ t = i,2,...,T$$

고정효과모형(FEM)이나 최소자승더미변수(LSDV)모형은 사용하기 편리하다는 장점이 있으나 다음 몇 가지 사항을 항상 염두에 두어야 한다.

첫째, 더미변수의 수가 너무 많을 경우 자유도의 고갈문제가 발생할 수 있다.

둘째, 설명변수의 수가 많을 경우 다중공선성의 문제가 항상 발생할 수 있으며 이 경우 좋은 추정량을 얻기에 어려울 수 있다.

셋째, FEM에 시간에 따라 변하지 않는 변수(성별, 인종 등)를 설명변수로 포함시킬 수 있으나 LSDV 접근방법으로는 그러한 변수가 종속변수에 미치는 영향을 식별하지 못할 수 있다.

넷째, 교란항(u_{it})에 대한 고전적인 가정들이 성립하지 않을 수 도 있는데 이러한 문제들은 Random Effect Model으로 해결될 수도 있음

(3)Random Effect Model

확률효과모형(REM)은 다음의 (10.15) 모형에서 v_i가 확률변수이며 $cov(X_{it},v_i)=0$을 가정한 경우이다.

$$y_{it}=\alpha+\sum_{k=1}^{K}\beta_k X_{kit}+v_i+\epsilon_{it} \tag{10.15}$$

FEM에서는 v_i를 회귀함수를 이동시키는 상수로 가정하였으나 REM에서는 cross-sectional units 자료를 모집단에서 추출된 표본으로 간주함으로써 v_i를 확률변수로 가정한다.

또한 FEM에서는 v_i가 독립변수와 상관관계가 있다고 간주하므로 v_i를 제거하기 위한 방법을 이용하였지만 REM에서는 v_i가 독립변수와 상관관계가 없다고 가정하므로 v_i를 제거하는 방법을 이용하면 비효율적인 추정량을 얻게 된다.

한편, REM에서 composite error term ($w_i=v_i+\epsilon_{it}$)이 자기상관이 있으므로(즉, $corr(w_{it},w_{is})=\dfrac{\sigma_v^2}{\sigma_v^2+\sigma_\epsilon^2}(t\neq s)$) pooled OLS로는 좋은 추정량을 얻을 수 없다. 따라서 확률효과모형의 추정방법은 다음의 변환된 모형을 pooled OLS로 추정하는 방법(이를 GLS 추정량 또는 random effects estimator라고 함)과 동일하며 이 경우 consistent and efficient estimator를 얻을 수 있다.

$$(y-\theta\overline{y_i})=\beta_0(1-\theta)+\beta_1(X_1-\theta\overline{X_{i1}})+....+\beta_k(X_k-\theta\overline{X_{ik}})+((1-\theta)v_i+(\epsilon_{it}-\theta\overline{\epsilon_i}))$$

단, $\theta=1-\sqrt{\frac{\sigma_\epsilon^2}{T\sigma_v^2+\sigma_\epsilon^2}}, 0\le\theta\le 1$이다

$\theta=1$이면 ($\sigma_\epsilon^2=0$) FEM이 되므로 θ가 1에 가까울수록 그룹의 특성을 고려하는 것이 매우 중요하며 따라서 REM은 FEM보다 제약이 덜한 모형이라고 할 수 있다.

$\theta=0$이면 ($\sigma_v^2=0$) pooled OLS가 되므로 θ가 0에 가까울수록 그룹의 특성을 고려할 필요성이 사라진다.

위의 모형에서 볼 수 있듯이 REM의 장점은 시간에 따라 변하지 않는 설명변수라도 모형에 포함될 수 있다는 것인데(FEM에서는 시간에 따라 변하지 않는 설명변수를 사용할 수 없음) 그 이유는 REM에서는 관측되지 않은 요인이 모든 설명변수(시간에 따라 변하든 변하지 않든)와 상관관계가 없다고 가정하기 때문이다.

결론적으로 pooled OLS, FEM 및 REM을 비교해 보면 REM estimator가 pooled OLS estimator보다 일반적으로 더 효율적이며, Hausman 검정을 통해 FEM 또는 REM 중 적합한 모형을 선택할 수 있으나 aggregate data를 이용한 정책 분석에서는 FEM이 REM보다 더 선호된다.

실습 : STATA 활용

<case ①-1 : time dummy>

1.fertile1.dta 자료를 이용하여 Wooldridge(2006)의 p.451에 있는 Table 13-1을 추정하면 다음과 같은 결과를 얻을 수 있다.

```
. use "C:\kanggc\FERTIL1.dta", clear
. reg kids  educ age agesq black east northcen west farm othrural town smcity y74 y76 y78 y80 y82 y84
```

Source	SS	df	MS		
Model	399.610888	17	23.5065228	Number of obs =	1129
Residual	2685.89841	1111	2.41755033	F(17, 1111) =	9.72
				Prob > F =	0.0000
				R-squared =	0.1295
				Adj R-squared =	0.1162
Total	3085.5093	1128	2.73538059	Root MSE =	1.5548

kids	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.1284268	.0183486	-7.00	0.000	-.1644286	-.092425
age	.5321346	.1383863	3.85	0.000	.2606065	.8036626
agesq	-.005804	.0015643	-3.71	0.000	-.0088733	-.0027347
black	1.075658	.1735356	6.20	0.000	.7351631	1.416152
east	.217324	.1327878	1.64	0.102	-.0432192	.4778672
northcen	.363114	.1208969	3.00	0.003	.125902	.6003261
west	.1976032	.1669134	1.18	0.237	-.1298978	.5251041
farm	-.0525575	.14719	-0.36	0.721	-.3413592	.2362443
othrural	-.1628537	.175442	-0.93	0.353	-.5070887	.1813814
town	.0843532	.124531	0.68	0.498	-.1599893	.3286957
smcity	.2118791	.160296	1.32	0.187	-.1026379	.5263961
y74	.2681825	.172716	1.55	0.121	-.0707039	.6070689
y76	-.0973795	.1790456	-0.54	0.587	-.448685	.2539261
y78	-.0686665	.1816837	-0.38	0.706	-.4251483	.2878154
y80	-.0713053	.1827707	-0.39	0.697	-.42992	.2873093
y82	-.5224842	.1724361	-3.03	0.003	-.8608214	-.184147
y84	-.5451661	.1745162	-3.12	0.002	-.8875846	-.2027477
_cons	-7.742457	3.051767	-2.54	0.011	-13.73033	-1.754579

추정된 2차 함수 모형에서 fertility의 전환점은 약 46세인 것으로 추정되었는데 그 이유는 $y = ax - bx^2$의 2차 함수에서 전환점은 $\frac{a}{2b}$가 되고 추정모형에서 a(=age)가 0.5321346이고 b(=age2)가 0.005804이므로 $\frac{a}{2b} = \frac{0.5321346}{2 \times 0.005804} = 45.86$이기 때문이다.

2.CPS78_85.dta 자료를 이용하여 Wooldridge(2006)의 p.452에 있는 (13.1)식을 추정하면 p.453에 있는 (13.2)식과 동일한 결과를 얻을 수 있다.

```
. use "C:\kanggc\CPS78_85.dta", clear

. reg  lwage y85 educ y85educ exper expersq union female y85fem
```

Source	SS	df	MS	Number of obs =	1084
Model	135.992074	8	16.9990092	F(8, 1075) =	99.80
Residual	183.099094	1075	.170324738	Prob > F =	0.0000
				R-squared =	0.4262
Total	319.091167	1083	.29463635	Adj R-squared =	0.4219
				Root MSE =	.4127

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y85	.1178062	.1237817	0.95	0.341	-.125075	.3606874
educ	.0747209	.0066764	11.19	0.000	.0616206	.0878212
y85educ	.0184605	.0093542	1.97	0.049	.000106	.036815
exper	.0295843	.0035673	8.29	0.000	.0225846	.036584
expersq	-.0003994	.0000775	-5.15	0.000	-.0005516	-.0002473
union	.2021319	.0302945	6.67	0.000	.1426888	.2615749
female	-.3167086	.0366215	-8.65	0.000	-.3885663	-.244851
y85fem	.085052	.051309	1.66	0.098	-.0156251	.185729
_cons	.4589329	.0934485	4.91	0.000	.2755707	.642295

교육으로 인한 수입의 변화(return to education)는 1978년 7.5%였으나 1985년에는 1978년 보다 1.85%p 높은 9.3%인 것으로 나타났으며, 또한 다른 모든 조건이 동일할 때 1978년의 경우 여성이 남성보다 31.7% 임금이 낮은 것으로 나타났다. 여기서 31.7%는 근사치이고 실제치는 $(e^{-0.317}-1)\times 100 = 27.2\%$인데 그 이유는 종속변수가 log형태이고 설명변수가 수준변수일 때 설명변수 1단위 변화에 따른 종속변수의 변화는 $(e^{\beta\times\Delta x}-1)\times 100\%$로 계산되기 때문이다.

<case ①-2 : difference in differences>

3.KIELMC.dta 자료를 이용하여 Wooldridge(2006)의 p.455에 있는 (13.4)식 및 (13.5)식을 추정한 후 treatment group 및 comparison group의 특정 사건 이전 및 이후 평균을 구하고 p.456의 (13.6)식에 의해 DID 추정량을 구하면 다음과 같은 결과를 얻을 수 있다.

```
. reg  rprice  nearinc if year==1981
```

Source	SS	df	MS	Number of obs =	142
Model	2.7059e+10	1	2.7059e+10	F(1, 140) =	27.73
Residual	1.3661e+11	140	975815069	Prob > F =	0.0000
				R-squared =	0.1653
Total	1.6367e+11	141	1.1608e+09	Adj R-squared =	0.1594
				Root MSE =	31238

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97	-19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43	107422.6

```
. reg  rprice  nearinc if year==1978

      Source |       SS       df       MS              Number of obs =     179
-------------+------------------------------           F(  1,   177) =   15.74
       Model |  1.3636e+10     1  1.3636e+10           Prob > F      =  0.0001
    Residual |  1.5332e+11   177   866239953           R-squared     =  0.0817
-------------+------------------------------           Adj R-squared =  0.0765
       Total |  1.6696e+11   178   937979126           Root MSE      =   29432

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nearinc |  -18824.37   4744.594    -3.97   0.000    -28187.62   -9461.117
       _cons |   82517.23   2653.79     31.09   0.000     77280.09    87754.37
------------------------------------------------------------------------------
```

	before(1978년)	after(1981년)	after-before
control(=fr)	82,517.23	101,307.5	18,790.27
treatment(=nr)	63,692.86	70,619.23	6,926.37
treatment-control	-18,824.37	-30,688.27	**-11,863.9**

한편, 위의 방법은 time dummy를 이용하여 추정할 수 도 있는데 KIELMC.dta 자료를 이용하여 Wooldridge(2006)의 p.456에 있는 (13.7)식을 추정하면 p.457에 있는 Table 13.2의 column (1)과 같은 결과를 얻을 수 있으며 위의 추정방법과 동일한 결과를 얻을 수 있다.

```
. xi: reg  rprice  i.year*nearinc
i.year            _Iyear_1978-1981    (naturally coded; _Iyear_1978 omitted)
i.year*nearinc    _IyeaXnea_#         (coded as above)

      Source |       SS       df       MS              Number of obs =     321
-------------+------------------------------           F(  3,   317) =   22.25
       Model |  6.1055e+10     3  2.0352e+10           Prob > F      =  0.0000
    Residual |  2.8994e+11   317   914632749           R-squared     =  0.1739
-------------+------------------------------           Adj R-squared =  0.1661
       Total |  3.5099e+11   320  1.0969e+09           Root MSE      =   30243

------------------------------------------------------------------------------
      rprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Iyear_1981 |   18790.29   4050.065     4.64   0.000     10821.88    26758.69
     nearinc |  -18824.37   4875.322    -3.86   0.000    -28416.45   -9232.293
_IyeaXn~1981 |   -11863.9   7456.646    -1.59   0.113    -26534.67    2806.866
       _cons |   82517.23   2726.91     30.26   0.000     77152.1     87882.36
------------------------------------------------------------------------------
```

위의 두 가지 추정방법 모두 -11863.9의 값은 두 지역 주택평균가격 차이(differences)에 있어서 두 기간의 차이(difference)를 의미하므로 difference in differences(DID) estimator라고 한다.

-11863.9의 값은 소각장에 가까이 위치하든 멀리 위치하든 다른 요인에 의한 주택가격의 변화가 동일하다는 가정 하에 새로운 소각로의 건설에 따른 주택가격의 변화를 의미한다.

<case ②-1 : first difference of two-period>

4.JTRAIN.dta 자료를 이용하여 Wooldridge(2006)의 p.468에 있는 (13.23)식에 의해 FD 추정량을 구하면 다음과 같은 결과를 얻을 수 있다.

```
. tsset fcode year
       panel variable:  fcode (strongly balanced)
        time variable:  year, 1987 to 1989
                delta:  1 unit

. reg D.scrap D.grant if year==1988
```

Source	SS	df	MS	Number of obs =	54
				F(1, 52) =	1.17
Model	6.73345587	1	6.73345587	Prob > F =	0.2837
Residual	298.400031	52	5.73846213	R-squared =	0.0221
				Adj R-squared =	0.0033
Total	305.133487	53	5.7572356	Root MSE =	2.3955

D.scrap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grant						
D1.	-.7394436	.6826276	-1.08	0.284	-2.109236	.6303488
_cons	-.5637143	.4049149	-1.39	0.170	-1.376235	.2488069

```
. tsset fcode year
       panel variable:  fcode (strongly balanced)
        time variable:  year, 1987 to 1989
                delta:  1 unit

. reg D.lscrap D.grant if year==1988
```

Source	SS	df	MS	Number of obs =	54
				F(1, 52) =	3.74
Model	1.23795555	1	1.23795555	Prob > F =	0.0585
Residual	17.1971834	52	.330715065	R-squared =	0.0672
				Adj R-squared =	0.0492
Total	18.4351389	53	.34783281	Root MSE =	.57508

D.lscrap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
grant						
D1.	-.3170579	.1638751	-1.93	0.058	-.6458974	.0117816
_cons	-.0574357	.097206	-0.59	0.557	-.2524938	.1376223

<case ②-2 : first difference of multi-periods, ie. T≥3>

5.EZUNEM.dta 자료를 이용하여 Wooldridge(2006)의 p.473에 있는 (13.32)식에 의해 FD 추정량을 구하면 다음과 같은 결과를 얻을 수 있다.

```
. use "C:\kanggc\EZUNEM.dta", clear

. reg  guclms d82 d83 d84 d85 d86 d87 d88 cez
```

Source	SS	df	MS		
Model	12.8826331	8	1.61032914	Number of obs =	176
Residual	7.79583789	167	.046681664	F(8, 167) =	34.50
				Prob > F =	0.0000
				R-squared =	0.6230
				Adj R-squared =	0.6049
Total	20.678471	175	.118162691	Root MSE =	.21606

guclms	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
d82	.7787595	.0651444	11.95	0.000	.6501469	.9073721
d83	-.0331192	.0651444	-0.51	0.612	-.1617318	.0954934
d84	-.0171382	.0685455	-0.25	0.803	-.1524655	.1181891
d85	.323081	.0666774	4.85	0.000	.1914417	.4547202
d86	.292154	.0651444	4.48	0.000	.1635413	.4207666
d87	.0539481	.0651444	0.83	0.409	-.0746645	.1825607
d88	-.0170526	.0651444	-0.26	0.794	-.1456652	.1115601
cez	-.1818775	.0781862	-2.33	0.021	-.3362382	-.0275169
_cons	-.3216319	.046064	-6.98	0.000	-.4125748	-.230689

<모형⑤>

6.Gujarati(2004)의 p.639에 있는 자료(grunfeld35-54.dta)를 이용하여 위의 모형⑤(p.640의 (16.2.1)식)을 추정해 보면 p.641의 (16.3.1)식과 같은 결과를 얻을 수 있다.

$$y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

단, y는 총투자(gross investment), X_2는 기업가치(value of firm), X_3는 총자본량(stock of plant and equipment)이다.

이 경우 4개 회사의 개별특성(individual heterogeneity)이 전혀 고려되지 않는다.

```
. use "C:\kanggc\grunfeld35-54.dta", clear

. reg y x2 x3
```

Source	SS	df	MS		
Model	4849457.37	2	2424728.69	Number of obs =	80
Residual	1560689.67	77	20268.697	F(2, 77) =	119.63
				Prob > F =	0.0000
				R-squared =	0.7565
				Adj R-squared =	0.7502
Total	6410147.04	79	81141.1018	Root MSE =	142.37

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.1100955	.0137297	8.02	0.000	.0827563	.1374348
x3	.3033932	.0492957	6.15	0.000	.2052328	.4015535
_cons	-63.30413	29.6142	-2.14	0.036	-122.2735	-4.334734

<모형④>

7.Gujarati(2004)의 p.639에 있는 자료(grunfeld35-54.dta)를 이용하여 위의 모형④(p.642의 (16.3.2)식)을 추정해 볼 수 있다.

$$y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

이를 Fixed Effect Model이라고 하는 이유는 기업별로 절편은 서로 다르지만 기울기는 시간에 따라 변하지 않고(time invariant) 모든 기업에 동일하기 때문에 붙여진 이름이다.

한편, 기업별 특성(firms heterogeneity) 또는 firm effect(경영방식의 차이 또는 경영자의 재능 등이 그 예임)가 있는 (16.3.2.)식은 가변수를 이용하여 다음과 같이 표현할 수 있으며 이 식을 추정해 보면 p.643의 (16.3.4)식과 같은 결과를 얻을 수 있음

$$y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{41} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

Fixed Effect를 추정하기 위해 더미변수를 사용하기 때문에 이를 최소자승더미변수(LSDV: Least Square Dummy Variable)모형이라고도 한다. LSDV모형은 covariance모형이라도도 하고 X_2와 X_3을 covariates라고 한다.

(참고) 먼저 데이터를 불러 들여 기업별 더미변수를 만든 후 회귀모형을 추정한다.

```
. use "C:₩kanggc₩grunfeld35-54.dta", clear
. tab id, gen(id_dum)
. reg  y id_dum2 id_dum3 id_dum4 x2 x3
```

Source	SS	df	MS	Number of obs =	80
Model	5990684.14	5	1198136.83	F(5, 74) =	211.37
Residual	419462.898	74	5668.41754	Prob > F =	0.0000
				R-squared =	0.9346
				Adj R-squared =	0.9301
Total	6410147.04	79	81141.1018	Root MSE =	75.289

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
id_dum2	161.5722	46.45639	3.48	0.001	69.00583	254.1386
id_dum3	339.6328	23.98633	14.16	0.000	291.839	387.4266
id_dum4	186.5665	31.50681	5.92	0.000	123.7879	249.3452
x2	.1079481	.0175089	6.17	0.000	.0730608	.1428354
x3	.3461617	.0266645	12.98	0.000	.2930315	.3992918
_cons	-245.7924	35.81112	-6.86	0.000	-317.1476	-174.4371

또한 firm effect와 동일한 방법으로 time effect(기술변화, 정부규제 또는 조세정책의 변화, 전쟁이나 분쟁과 같은 외부환경의 변화 등이 그 예임)를 가변수를 이용해 다음과 같이 표현하고(p.644의 (16.3.6)식) 추정하면 다음과 같은 결과를 얻을 수 있다.

$$y_{it} = \lambda_0 + \lambda_1 Dum35 + \lambda_2 Dum36 + ... + \lambda_{19} Dum53 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

```
. char year [omit] 1954

. xi: reg y i.year x2 x3
i.year            _Iyear_1935-1954    (naturally coded; _Iyear_1954 omitted)
```

Source	SS	df	MS	Number of obs =	80
Model	4938658.06	21	235174.193	F(21, 58) =	9.27
Residual	1471488.98	58	25370.4997	Prob > F =	0.0000
				R-squared =	0.7704
				Adj R-squared =	0.6873
Total	6410147.04	79	81141.1018	Root MSE =	159.28

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Iyear_1935	21.02495	129.7675	0.16	0.872	-238.7329	280.7828
_Iyear_1936	-14.03424	133.1953	-0.11	0.916	-280.6535	252.5851
_Iyear_1937	-58.41449	135.1823	-0.43	0.667	-329.0113	212.1823
_Iyear_1938	-48.66584	126.5326	-0.38	0.702	-301.9483	204.6167
_Iyear_1939	-96.80262	128.0009	-0.76	0.453	-353.0243	159.4191
_Iyear_1940	-36.10807	129.116	-0.28	0.781	-294.5618	222.3457
_Iyear_1941	21.16647	127.7266	0.17	0.869	-234.5062	276.8391
_Iyear_1942	30.29937	124.1757	0.24	0.808	-218.2653	278.864
_Iyear_1943	-12.76907	124.8506	-0.10	0.919	-262.6847	237.1466
_Iyear_1944	-17.82782	125.6498	-0.14	0.888	-269.3431	233.6875
_Iyear_1945	-38.96994	126.769	-0.31	0.760	-292.7257	214.7858
_Iyear_1946	26.90836	125.7493	0.21	0.831	-224.8062	278.623
_Iyear_1947	27.81252	119.3515	0.23	0.817	-211.0954	266.7204
_Iyear_1948	26.05879	117.3271	0.22	0.825	-208.7969	260.9145
_Iyear_1949	-28.65514	116.2965	-0.25	0.806	-261.4479	204.1376
_Iyear_1950	-20.3938	115.8973	-0.18	0.861	-252.3874	211.5998
_Iyear_1951	.9819593	116.2576	0.01	0.993	-231.733	233.6969
_Iyear_1952	21.96068	114.6821	0.19	0.849	-207.6005	251.5219
_Iyear_1953	39.43192	113.441	0.35	0.729	-187.6448	266.5087
x2	.1159174	.01817	6.38	0.000	.0795462	.1522886
x3	.2696593	.0833411	3.24	0.002	.1028339	.4364847
_cons	-56.33982	99.75287	-0.56	0.574	-256.0169	143.3373

<모형③>

8.Gujarati(2004)의 p.639에 있는 자료(grunfeld35-54.dta)를 이용하여 위의 모형③(p.644의 (16.3.7)식)을 추정하면 다음과 같은 결과를 얻을 수 있다.

$$y = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \lambda_0 + \lambda_1 Dum35 + \lambda_2 Dum36 + ... + \lambda_{19} Dum53 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

```
. xi: reg y id_dum2 id_dum3 id_dum4 i.year x2 x3
i.year            _Iyear_1935-1954    (naturally coded; _Iyear_1954 omitted)
```

Source	SS	df	MS	Number of obs =	80
				F(24, 55) =	42.58
Model	6082748.11	24	253447.838	Prob > F =	0.0000
Residual	327398.928	55	5952.70778	R-squared =	0.9489
				Adj R-squared =	0.9266
Total	6410147.04	79	81141.1018	Root MSE =	77.154

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
id_dum2	105.2457	67.68668	1.55	0.126	-30.40141	240.8929
id_dum3	341.1008	24.8116	13.75	0.000	291.3773	390.8244
id_dum4	220.324	41.16781	5.35	0.000	137.8219	302.8262
_Iyear_1935	134.3636	72.00957	1.87	0.067	-9.946798	278.674
_Iyear_1936	87.3297	66.46131	1.31	0.194	-45.86174	220.5211
_Iyear_1937	29.58592	66.10801	0.45	0.656	-102.8975	162.0693
_Iyear_1938	48.12215	66.82507	0.72	0.474	-85.79829	182.0426
_Iyear_1939	-7.886558	63.9173	-0.12	0.902	-135.9797	120.2066
_Iyear_1940	51.85022	63.78342	0.81	0.420	-75.9746	179.6751
_Iyear_1941	107.7241	63.45228	1.70	0.095	-19.43706	234.8854
_Iyear_1942	118.3592	64.92635	1.82	0.074	-11.75613	248.4745
_Iyear_1943	71.99879	63.47345	1.13	0.262	-55.20486	199.2024
_Iyear_1944	68.47821	63.65855	1.08	0.287	-59.09638	196.0528
_Iyear_1945	44.28525	62.83216	0.70	0.484	-81.63321	170.2037
_Iyear_1946	104.1942	61.82815	1.69	0.098	-19.71216	228.1006
_Iyear_1947	100.1916	62.87224	1.59	0.117	-25.80715	226.1904
_Iyear_1948	92.30316	63.16828	1.46	0.150	-34.2889	218.8952
_Iyear_1949	31.20783	62.10854	0.50	0.617	-93.26046	155.6761
_Iyear_1950	35.80472	61.1856	0.59	0.561	-86.81396	158.4234
_Iyear_1951	47.8422	57.53413	0.83	0.409	-67.45878	163.1432
_Iyear_1952	57.96435	56.39349	1.03	0.309	-55.05073	170.9794
_Iyear_1953	54.01372	54.99911	0.98	0.330	-56.20695	164.2344
x2	.129307	.0274237	4.72	0.000	.0743487	.1842653
x3	.3672492	.0416591	8.82	0.000	.2837625	.450736
_cons	-359.5819	82.63602	-4.35	0.000	-525.1882	-193.9756

<모형②>

9.Gujarati(2004)의 p.639에 있는 자료(grunfeld35-54.dta)를 이용하여 위의 모형②(p.645의 (16.3.8)식)을 추정하면 p.645의 Table 16.2와 같은 결과를 얻을 수 있다.

$$y = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 (D_{2i} X_{2it}) + \gamma_2 (D_{2i} X_{3it}) + .. + \gamma_5 (D_{4i} X_{2it}) + \gamma_6 (D_{4i} X_{3it}) + u_{it}$$

```
. char id [omit] 1

. xi: reg y i.id*x2 i.id*x3
i.id              _Iid_1-4          (naturally coded; _Iid_1 omitted)
i.id*x2           _IidXx2_#         (coded as above)
i.id*x3           _IidXx3_#         (coded as above)

      Source |       SS       df       MS              Number of obs =      80
-------------+------------------------------           F( 11,    68) =  120.38
       Model |  6097055.58     11  554277.78            Prob > F      =  0.0000
    Residual |   313091.46     68  4604.28617           R-squared     =  0.9512
-------------+------------------------------           Adj R-squared =  0.9433
       Total |  6410147.04     79  81141.1018           Root MSE      =  67.855

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _Iid_2 |  (dropped)
       _Iid_3 |  (dropped)
       _Iid_4 |  (dropped)
          x2 |   .0265512   .0378814     0.70   0.486    -.0490399    .1021423
  _IidXx2_2 |   .0926585   .0424167     2.18   0.032     .0080173    .1772996
  _IidXx2_3 |   .1448793   .0646504     2.24   0.028     .0158714    .2738871
  _IidXx2_4 |   .0265042   .1111402     0.24   0.812    -.1952726    .2482809
       _Iid_2 |  -139.5103   109.2808    -1.28   0.206    -357.5768    78.55615
       _Iid_3 |  -40.12173   129.2343    -0.31   0.757    -298.0047    217.7613
       _Iid_4 |   9.375908   93.11719     0.10   0.920    -176.4365    195.1884
          x3 |   .1516939    .062553     2.43   0.018     .0268713    .2765164
  _IidXx3_2 |   .2198306   .0682914     3.22   0.002     .0835573     .356104
  _IidXx3_3 |   .2570148   .1204774     2.13   0.037     .0166058    .4974238
  _IidXx3_4 |  -.0600001   .3785974    -0.16   0.875    -.8154795    .6954793
       _cons |  -9.956308   76.3518    -0.13   0.897     -162.314    142.4013
------------------------------------------------------------------------------
```

10.다음의 예는 Green(2000)의 교재에 있는 TABLE 14.1(p.566)을 추정하기 위한 Do파일인데 이를 실행하면 다음과 같은 결과를 얻을 수 있다.

```
/* DO FILE TO ESTIMATE THE TABLE 14-1 in GREEN(2000)'s TEXTBOOK*/

set more 1
#delimit ;
log using c:₩kanggc₩tablea14-1.log, replace;
use c:₩kanggc₩tablea14_1.dta, clear;

/*Estimation of No effects*/
reg lc lq lpf lf;
by id, sort : egen float lc_mean = mean(lc);
by id, sort : egen float lq_mean = mean(lq);
by id, sort : egen float lpf_mean = mean(lpf);
by id, sort : egen float lf_mean = mean(lf);

/*Estimation of Group means(Between Effects Estimation)*/
reg  lc_mean lq_mean lpf_mean lf_mean if time==1;
gen dlc=lc-lc_mean;
gen dlq=lq-lq_mean;
gen dlpf=lpf-lpf_mean;
gen dlf=lf-lf_mean;

/*Estimation of Firm effects(with-in groups)*/
reg  dlc dlq dlpf dlf, nocons;
tab id, gen(id_dum);

/*Estimation of Firm effects(LSDV)*/
reg  lc lq lpf lf id_dum1 id_dum2 id_dum3 id_dum4 id_dum5 id_dum6, nocons;
xi i.time, noomit;

/*Estimation of Time effects*/
reg  lc lq lpf lf _I*, nocons;

/*Estimation of Firm and Time effects*/
reg  lc lq lpf lf id_dum1 id_dum2 id_dum3 id_dum4 id_dum5 id_dum6  _I*;
tab time, gen(t_dum);
reg  lc lq lpf lf id_dum2 id_dum3 id_dum4 id_dum5 id_dum6 t_dum2 t_dum3 t_dum4
t_dum5 t_dum6 t_dum7 t_dum8 t_dum9 t_dum10 t_dum11 t_dum12 t_dum13 t_dum14
t_dum15;
log close;
exit;
```

```
. /*Estimation of No effects*/
>
> reg lc lq lpf lf;
```

Source	SS	df	MS	Number of obs =	90
Model	112.705452	3	37.5684839	F(3, 86) =	2419.34
Residual	1.33544153	86	.01552839	Prob > F =	0.0000
				R-squared =	0.9883
				Adj R-squared =	0.9879
Total	114.040893	89	1.28135835	Root MSE =	.12461

lc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq	.8827385	.0132545	66.60	0.000	.8563895	.9090876
lpf	.453977	.0203042	22.36	0.000	.4136136	.4943404
lf	-1.62751	.345302	-4.71	0.000	-2.313948	-.9410727
_cons	9.516923	.2292445	41.51	0.000	9.0612	9.972645

```
. /*Estimation of Group means(Between Effects Estimation)*/
>
> reg  lc_mean lq_mean lpf_mean lf_mean if time==1;
```

Source	SS	df	MS	Number of obs =	6
Model	4.94698124	3	1.64899375	F(3, 2) =	104.12
Residual	.031675926	2	.015837963	Prob > F =	0.0095
				R-squared =	0.9936
				Adj R-squared =	0.9841
Total	4.97865717	5	.995731433	Root MSE =	.12585

lc_mean	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq_mean	.7824568	.1087646	7.19	0.019	.3144803	1.250433
lpf_mean	-5.523904	4.478718	-1.23	0.343	-24.79427	13.74647
lf_mean	-1.751072	2.743167	-0.64	0.589	-13.55397	10.05182
_cons	85.8081	56.48199	1.52	0.268	-157.2143	328.8305

또는 패널자료를 만드는 명령어인 xtset과 패널회귀분석을 위한 명령어인 xtreg를 이용하면 다음과 같은 추정결과를 얻을 수 있으며 결과는 위와 동일하다.

```
. xtset id time
       panel variable:  id (strongly balanced)
        time variable:  time, 1 to 15
                delta:  1 unit

. xtreg lc lq lpf lf, be

Between regression (regression on group means)  Number of obs      =        90
Group variable: id                              Number of groups   =         6

R-sq:  within  = 0.8808                         Obs per group: min =        15
       between = 0.9936                                        avg =      15.0
       overall = 0.1371                                        max =        15

                                                F(3,2)             =    104.12
sd(u_i + avg(e_i.))=  .1258491                  Prob > F           =    0.0095
```

lc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq	.7824552	.1087663	7.19	0.019	.3144715	1.250439
lpf	-5.523978	4.478802	-1.23	0.343	-24.79471	13.74675
lf	-1.751016	2.74319	-0.64	0.589	-13.55401	10.05198
_cons	85.80901	56.48302	1.52	0.268	-157.2178	328.8358

```
. /*Estimation of Firm effects(with-in groups)*/
> reg  dlc dlq dlpf dlf, nocons;
```

Source	SS	df	MS	Number of obs =	90
				F(3, 87) =	3871.82
Model	39.0683861	3	13.0227954	Prob > F =	0.0000
Residual	.292622861	87	.003363481	R-squared =	0.9926
				Adj R-squared =	0.9923
Total	39.361009	90	.437344544	Root MSE =	.058

dlc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dlq	.9192846	.028841	31.87	0.000	.86196	.9766092
dlpf	.4174918	.0146657	28.47	0.000	.3883422	.4466414
dlf	-1.070396	.1946109	-5.50	0.000	-1.457206	-.6835858

```
. /*Estimation of Firm effects(LSDV)*/
>
> reg  lc lq lpf lf id_dum1 id_dum2 id_dum3 id_dum4 id_dum5 id_dum6, nocons;
```

Source	SS	df	MS	Number of obs =	90
				F(9, 81) =	.
Model	16191.3043	9	1799.03381	Prob > F =	0.0000
Residual	.292622872	81	.003612628	R-squared =	1.0000
				Adj R-squared =	1.0000
Total	16191.5969	90	179.906633	Root MSE =	.06011

lc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq	.9192846	.0298901	30.76	0.000	.8598126	.9787565
lpf	.4174918	.0151991	27.47	0.000	.3872503	.4477333
lf	-1.070396	.20169	-5.31	0.000	-1.471696	-.6690963
id_dum1	9.705942	.193124	50.26	0.000	9.321686	10.0902
id_dum2	9.664706	.198982	48.57	0.000	9.268794	10.06062
id_dum3	9.497021	.2249584	42.22	0.000	9.049424	9.944618
id_dum4	9.890498	.2417635	40.91	0.000	9.409464	10.37153
id_dum5	9.729997	.2609421	37.29	0.000	9.210804	10.24919
id_dum6	9.793004	.2636622	37.14	0.000	9.268399	10.31761

또는 xtset과 xtreg를 이용하면 다음과 같은 추정결과를 얻을 수 있으며 결과는 절편을 제외하면 위와 동일하다.

```
. xtset id time
       panel variable:  id (strongly balanced)
        time variable:  time, 1 to 15
                delta:  1 unit

. xtreg lc lq lpf lf, fe

Fixed-effects (within) regression               Number of obs      =        90
Group variable: id                              Number of groups   =         6

R-sq:  within  = 0.9926                         Obs per group: min =        15
       between = 0.9856                                        avg =      15.0
       overall = 0.9873                                        max =        15

                                                F(3,81)            =   3604.80
corr(u_i, Xb)  = -0.3475                        Prob > F           =    0.0000

------------------------------------------------------------------------------
          lc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          lq |   .9192846   .0298901    30.76   0.000     .8598126    .9787565
         lpf |   .4174918   .0151991    27.47   0.000     .3872503    .4477333
          lf |  -1.070396     .20169    -5.31   0.000    -1.471696   -.6690963
       _cons |   9.713528    .229641    42.30   0.000     9.256614    10.17044
-------------+----------------------------------------------------------------
     sigma_u |  .1320775
     sigma_e |  .06010514
         rho |  .82843653   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(5, 81) =    57.73                 Prob > F = 0.0000
```

그러면 개별회사의 절편은 어떻게 구하나? 회귀계수를 추정한 후 개별 절편은 다음과 같은 식에 의해 구한다.

$$\widehat{\alpha_i} = \overline{y_i} - \overline{X_i}'\widehat{\beta}, i = 1,2,...,N$$

따라서 위 식에 의해 6개 회사의 개별 절편은 먼저 다음과 같이 6개의 절편의 차이(합이 0이 됨에 유의할 것)를 구한 다음 절편의 평균(9.713528)에서 절편의 차이를 각각 더해 주면 두 번째 위에 있는 결과(Estimation of Firm effects)와 동일한 결과를 얻을 수 있다.

```
. collapse (mean) lc lq lpf lf, by(id)

. gen a = lc - (_b[_cons] + _b[lq]*lq+_b[lpf]*lpf+_b[lf]*lf)

. l a
```

	a
1.	-.007586
2.	-.0488218
3.	-.2165074
4.	.1769701
5.	.0164692
6.	.0794759

```
. gen aa = a+9.713528

. l aa
```

	aa
1.	9.705942
2.	9.664706
3.	9.497021
4.	9.890498
5.	9.729998
6.	9.793004

```
. /*Estimation of Time effects*/
>
> reg  lc lq lpf lf _I*, nocons;
```

Source	SS	df	MS	Number of obs =	90
				F(18, 72) =	59513.52
Model	16190.5087	18	899.472708	Prob > F =	0.0000
Residual	1.08819022	72	.015113753	R-squared =	0.9999
				Adj R-squared =	0.9999
Total	16191.5969	90	179.906633	Root MSE =	.12294

lc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq	.8677268	.0154082	56.32	0.000	.8370111	.8984424
lpf	-.4844835	.3641085	-1.33	0.188	-1.210321	.2413535
lf	-1.954404	.4423777	-4.42	0.000	-2.836268	-1.07254
_Itime_1	20.4958	4.209528	4.87	0.000	12.10426	28.88735
_Itime_2	20.57804	4.221526	4.87	0.000	12.16258	28.9935
_Itime_3	20.65573	4.224177	4.89	0.000	12.23499	29.07647
_Itime_4	20.74076	4.24575	4.89	0.000	12.27701	29.20451
_Itime_5	21.19983	4.440331	4.77	0.000	12.34819	30.05147
_Itime_6	21.41162	4.538621	4.72	0.000	12.36404	30.4592
_Itime_7	21.50335	4.571397	4.70	0.000	12.39044	30.61626
_Itime_8	21.65403	4.622886	4.68	0.000	12.43847	30.86958
_Itime_9	21.82957	4.656906	4.69	0.000	12.5462	31.11294
_Itime_10	22.1138	4.792648	4.61	0.000	12.55983	31.66777
_Itime_11	22.46533	4.949909	4.54	0.000	12.59786	32.33279
_Itime_12	22.65134	5.008592	4.52	0.000	12.66689	32.63578
_Itime_13	22.61656	4.986139	4.54	0.000	12.67687	32.55624
_Itime_14	22.55223	4.955942	4.55	0.000	12.67274	32.43172
_Itime_15	22.53677	4.940532	4.56	0.000	12.68799	32.38554

```
. /*Estimation of Firm and Time effects*/
>
> reg  lc lq lpf lf id_dum1 id_dum2 id_dum3 id_dum4 id_dum5 id_dum6  _I*;
```

Source	SS	df	MS	Number of obs =	90
Model	113.864044	22	5.17563838	F(22, 67) =	1960.82
Residual	.176848775	67	.002639534	Prob > F =	0.0000
				R-squared =	0.9984
				Adj R-squared =	0.9979
Total	114.040893	89	1.28135835	Root MSE =	.05138

lc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq	.8172487	.031851	25.66	0.000	.7536739	.8808235
lpf	.16861	.163478	1.03	0.306	-.1576935	.4949135
lf	-.8828142	.2617373	-3.37	0.001	-1.405244	-.3603843
id_dum1	.1742825	.0861201	2.02	0.047	.0023861	.346179
id_dum2	.1114508	.0779551	1.43	0.157	-.0441482	.2670499
id_dum3	-.143511	.0518934	-2.77	0.007	-.2470907	-.0399313
id_dum4	.1802087	.0321443	5.61	0.000	.1160484	.2443691
id_dum5	-.0466942	.0224688	-2.08	0.042	-.0915422	-.0018463
id_dum6	(dropped)					
_Itime_1	(dropped)					
_Itime_2	.0547017	.0302618	1.81	0.075	-.0057012	.1151046
_Itime_3	.0973352	.0326845	2.98	0.004	.0320967	.1625737
_Itime_4	.1509845	.0370793	4.07	0.000	.0769739	.2249952
_Itime_5	.2200954	.1117987	1.97	0.053	-.0030557	.4432464
_Itime_6	.2659341	.1539895	1.73	0.089	-.0414302	.5732983
_Itime_7	.2971599	.1690435	1.76	0.083	-.0402522	.6345721
_Itime_8	.3532919	.1923205	1.84	0.071	-.0305814	.7371653
_Itime_9	.421245	.2097703	2.01	0.049	.0025418	.8399482
_Itime_10	.4657525	.2707517	1.72	0.090	-.0746701	1.006175
_Itime_11	.581335	.339615	1.71	0.092	-.0965395	1.259209
_Itime_12	.6594972	.3662185	1.80	0.076	-.0714779	1.390472
_Itime_13	.6754036	.3565651	1.89	0.063	-.0363033	1.387111
_Itime_14	.6744931	.3440841	1.96	0.054	-.0123016	1.361288
_Itime_15	.6931382	.3378385	2.05	0.044	.0188098	1.367467
_cons	12.2469	1.885399	6.50	0.000	8.48363	16.01018

또한 첫 번째 회사와 첫 번째 시간을 base로 한 가변수 회귀모형을 추정하면 STATA에서는 다음과 같은 결과를 준다.

```
. reg  lc lq lpf lf id_dum2 id_dum3 id_dum4 id_dum5 id_dum6 t_dum2 t_dum3 t_dum4 t_dum5 t_dum6 t_dum7 t_dum8 t_dum9 t_dum
> 10 t_dum11 t_dum12 t_dum13 t_dum14 t_dum15
```

Source	SS	df	MS		
Model	113.864044	22	5.17563838	Number of obs =	90
Residual	.176848775	67	.002639534	F(22, 67) =	1960.82
				Prob > F =	0.0000
				R-squared =	0.9984
				Adj R-squared =	0.9979
Total	114.040893	89	1.28135835	Root MSE =	.05138

lc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lq	.8172487	.031851	25.66	0.000	.7536739	.8808235
lpf	.16861	.163478	1.03	0.306	-.1576935	.4949135
lf	-.8828142	.2617373	-3.37	0.001	-1.405244	-.3603843
id_dum2	-.0628317	.0235099	-2.67	0.009	-.1097576	-.0159058
id_dum3	-.3177935	.0417073	-7.62	0.000	-.4010416	-.2345454
id_dum4	.0059262	.0610688	0.10	0.923	-.1159676	.1278199
id_dum5	-.2209768	.0812562	-2.72	0.008	-.3831649	-.0587887
id_dum6	-.1742825	.0861201	-2.02	0.047	-.346179	-.0023861
t_dum2	.0547017	.0302618	1.81	0.075	-.0057012	.1151046
t_dum3	.0973352	.0326845	2.98	0.004	.0320967	.1625737
t_dum4	.1509845	.0370793	4.07	0.000	.0769739	.2249952
t_dum5	.2200954	.1117987	1.97	0.053	-.0030557	.4432464
t_dum6	.2659341	.1539895	1.73	0.089	-.0414302	.5732983
t_dum7	.2971599	.1690435	1.76	0.083	-.0402522	.6345721
t_dum8	.3532919	.1923205	1.84	0.071	-.0305814	.7371653
t_dum9	.421245	.2097703	2.01	0.049	.0025418	.8399482
t_dum10	.4657525	.2707517	1.72	0.090	-.0746701	1.006175
t_dum11	.581335	.339615	1.71	0.092	-.0965395	1.259209
t_dum12	.6594972	.3662185	1.80	0.076	-.0714779	1.390472
t_dum13	.6754036	.3565651	1.89	0.063	-.0363033	1.387111
t_dum14	.6744931	.3440841	1.96	0.054	-.0123016	1.361288
t_dum15	.6931382	.3378385	2.05	0.044	.0188098	1.367467
_cons	12.42119	1.894991	6.55	0.000	8.638768	16.2036

한편, EViews에서는 다음과 같은 결과를 주는데 어떻게 STATA 결과에서 EViews가 주는 것과 동일한 결과를 얻을 수 있는 지 살펴보기로 하자.

Variable	Coefficient
C	12.66687
LOG(Q_?)	0.817249
LOG(PF_?)	0.168611
LF_?	-0.882812
Fixed Effects (Cross)	
1--C	0.128326
2--C	0.065495
3--C	-0.189467
4--C	0.134253
5--C	-0.092650
6--C	-0.045956
Fixed Effects (Period)	
2001--C	-0.374023
2002--C	-0.319322
2003--C	-0.276689

2004--C	-0.223039
2005--C	-0.153929
2006--C	-0.108090
2007--C	-0.076864
2008--C	-0.020733
2009--C	0.047220
2010--C	0.091728
2011--C	0.207310
2012--C	0.285472
2013--C	0.301378
2014--C	0.300468
2015--C	0.319113

먼저, STATA와 같이 첫 번째 회사와 첫 번째 시간을 base로 한 가변수 회귀모형을 다음과 같이 설정하고 추정하면 20개의 추정치를 얻을 수 있다.

$$y_{it} = \_cons + id_2F_2 + id_3F_3 + ... + id_6F_6 + t_2T_2 + t_3T_3 + ... + t_{15}T_{15} + u_{it} \quad ⓐ$$

다음으로, EViews가 주는 결과와 동일한 모형을 다음과 같이 설정하면 22개의 추정해야 할 회귀계수가 있다.

$$y_{it} = C + 1_cF_1 + 2_cF_2 + ... + 6_cF_6 + 2001_cT_1 + 2002_cT_2 + ... + 2015_cT_{15} + u_{it} \quad ⓑ$$

단, ⓑ모형에서 $1_c + 2_c + ... + 6_c = 0$이고 $2001_c + 2002_c + ... + 2015_c = 0$이므로 2개의 제약이 생겨 ⓐ모형의 추정계수를 이용하여 ⓑ모형의 회귀계수를 계산해 낼 수 있다.

ⓐ모형 및 ⓑ모형은 다음의 관계가 성립한다.

(for i=1, t=1) $C + 1_c + 2001_c = \_cons$ ①

(for i=2, t=2) $C + 2_c + 2002_c = \_cons + id_2 + t_2$ ②

(for i=3, t=3) $C + 3_c + 2003_c = \_cons + id_3 + t_3$ ③

(for i=4, t=4) $C + 4_c + 2004_c = \_cons + id_4 + t_4$ ④

(for i=5, t=5) $C + 5_c + 2005_c = \_cons + id_5 + t_5$ ⑤

(for i=6, t=6) $C + 6_c + 2006_c = \_cons + id_6 + t_6$ ⑥

(for i=1, t=2) $C + 1_c + 2002_c = \_cons + t_2$ ⑦

(for i=1, t=3) $C+1_c+2003_c=\_cons+t_3$ ⑧

(for i=1, t=4) $C+1_c+2004_c=\_cons+t_4$ ⑨

(for i=1, t=5) $C+1_c+2005_c=\_cons+t_5$ ⑩

(for i=1, t=6) $C+1_c+2006_c=\_cons+t_6$ ⑪

(for i=1, t=7) $C+1_c+2007_c=\_cons+t_7$ ⑫

(for i=1, t=8) $C+1_c+2008_c=\_cons+t_8$ ⑬

(for i=1, t=9) $C+1_c+2009_c=\_cons+t_9$ ⑭

(for i=1, t=10) $C+1_c+2010_c=\_cons+t_{10}$ ⑮

(for i=1, t=11) $C+1_c+2011_c=\_cons+t_{11}$ ⑯

(for i=1, t=12) $C+1_c+2012_c=\_cons+t_{12}$ ⑰

(for i=1, t=13) $C+1_c+2013_c=\_cons+t_{13}$ ⑱

(for i=1, t=14) $C+1_c+2014_c=\_cons+t_{14}$ ⑲

(for i=1, t=15) $C+1_c+2015_c=\_cons+t_{15}$ ⑳

$1_c=-(2_c+...+6_c)$ 및 $2001_c=-(2002_c+...+2015_c)$을 ①식에 대입하여 정리하고, $1_c=-(2_c+...+6_c)$을 ⑦-⑳에 대입하여 정리한다.

$C;2_c,3_c,..,6_c,2002_c,2003_c,2015_c$의 값을 구해야 하므로 행렬표현을 이용하여 연립방정식을 풀면 된다.

$AX=H$에서 $X=A^{-1}H$를 이용하여 $X=(C,2_c,3_c,..,6_c,2002_c,2003_c,2015_c)'$를 구한다.

①-⑳식에 의해 다음의 표와 같은 A행렬이 도출됨

(i,t)	c	2_c	3_c	4_c	5_c	6_c	2002_c	2003_c	2004_c	2005_c	2006_c	2007_c	2008_c	2009_c	2010_c	2011_c	2012_c	2013_c	2014_c	2015_c
(1,1)	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
(2,2)	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
(3,3)	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
(4,4)	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
(5,5)	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
(6,6)	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
(1,2)	1	-1	-1	-1	-1	-1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,3)	1	-1	-1	-1	-1	-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
(1,4)	1	-1	-1	-1	-1	-1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
(1,5)	1	-1	-1	-1	-1	-1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
(1,6)	1	-1	-1	-1	-1	-1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
(1,7)	1	-1	-1	-1	-1	-1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
(1,8)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	1	0	0	0	0	0	0	0

(1,9)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
(1,10)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
(1,11)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
(1,12)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
(1,13)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
(1,14)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
(1,15)	1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	1

또한 ⓐ모형의 추정계수와 ①-⑳식에 의해 다음의 표와 같은 H행렬을 얻을 수 있다.

12.42119
12.41306
12.20073
12.5781
12.42031
12.51284
12.47589
12.51853
12.57217
12.64129
12.68712
12.71835
12.77448
12.84244
12.88694
13.00253
13.08069
13.09659
13.09568
13.11433

$X = A^{-1}H$을 계산하면 다음의 표와 같은 결과를 얻을 수 있다.

C=	12.66689
2_c	0.065495
3_c=	-0.18947
4_c=	0.134253

5_c=	-0.09265
6_c=	-0.04596
2002_c=	-0.31932
2003_c=	-0.27669
2004_c=	-0.22304
2005_c=	-0.15393
2006_c=	-0.10809
2007_c=	-0.07686
2008_c=	-0.02073
2009_c=	0.047221
2010_c=	0.091728
2011_c=	0.207311
2012_c=	0.285473
2013_c=	0.301379
2014_c=	0.300469
2015_c=	0.319114

마지막으로 $1_c = -(2_c + \ldots + 6_c) = 0.128326$ 및 $2001_c = -(2002_c + \ldots + 2015_c) = -0.37402$을 구할 수 있다.

REM을 추정하기 위해서는 FEM을 이용하여 $\widehat{\sigma_\epsilon^2}$, $\widehat{\sigma_v^2}$을 먼저 추정한 후 $\hat{\theta}$의 값을 계산해야 하는데 Green(2000)의 교재에 있는 TABLE 14.1(p.566) 및 TABLE 14.22(p.574)를 살펴보자.

$\widehat{\sigma_\epsilon^2}$는 최소자승더미변수(LSDV)모형의 residual variance estimator로서 0.003612628이고(p.254의 Estimation of Fixed effects(LSDV)의 Residual MS 값), group means모형의 residual variance estimator인 0.015837963(p.253의 Estimation of Group means(Between Effects Estimation의 Residual MS 값)는 $\frac{\widehat{\sigma_\epsilon^2}}{T} + \widehat{\sigma_v^2}$의 값이므로 $\widehat{\sigma_v^2} = 0.015837963 - \frac{\widehat{\sigma_\epsilon^2}}{T} = 0.015837963 - \frac{0.003612628}{15} = 0.0155973$이고,

$\hat{\theta} = 1 - \sqrt{\frac{\widehat{\sigma_\epsilon^2}}{T\widehat{\sigma_v^2} + \widehat{\sigma_\epsilon^2}}} = 1 - \sqrt{\frac{0.003612628}{15 \times 0.015837963}} = 0.87669$이 된다.

$\hat{\theta} = 0.87669$를 이용하여 종속변수 및 독립변수를 위 모형과 같이 변환

(transformation)한 후 변환된 모형을 pooled OLS로 추정하면 다음과 같은 FGLS(Feasible Generalized Least Squares) 추정량을 얻을 수 있다.
상수항을 제외한 회귀계수는 그대로 해석하면 되고 상수항의 GLS 추정량은 pooled OLS 추정량을 $1-\hat{\theta}$으로 나누어 구하면 된다.

Green(2000)의 교재에 있는 TABLE 14.2(p.574)의 Random effects모형을 추정하기 위해 pooled OLS로 추정하면 다음과 같은 결과(FGLS 추정량)를 얻는다.

```
. gen rlc=lc-0.87669*lc_mean

. gen rlq=lq-0.87669*lq_mean

. gen rlpf=lpf-0.87669*lpf_mean

. gen rlf=lf-0.87669*lf_mean

. reg  rlc rlq rlpf rlf
```

Source	SS	df	MS	Number of obs =	90
Model	40.1849632	3	13.3949877	F(3, 86) =	3697.12
Residual	.311585117	86	.003623083	Prob > F =	0.0000
				R-squared =	0.9923
				Adj R-squared =	0.9920
Total	40.4965483	89	.455017397	Root MSE =	.06019

rlc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rlq	.9066815	.0256252	35.38	0.000	.8557403	.9576226
rlpf	.4227781	.0140248	30.15	0.000	.3948977	.4506585
rlf	-1.064499	.2000699	-5.32	0.000	-1.462225	-.6667734
_cons	1.187218	.0259155	45.81	0.000	1.1357	1.238737

```
. di 1.187218/(1-0.87669)
9.6279134
```

또는 xtset과 xtreg를 이용하면 다음과 같은 추정결과를 얻을 수 있으며 이 결과는 절편을 제외하면 위와 동일하다.

```
. xtreg  lc lq lpf lf, re
```

```
Random-effects GLS regression                   Number of obs      =        90
Group variable: id                              Number of groups   =         6

R-sq:  within  = 0.9925                         Obs per group: min =        15
       between = 0.9856                                        avg =      15.0
       overall = 0.9876                                        max =        15

Random effects u_i ~ Gaussian                   Wald chi2(3)       =  11091.33
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
```

lc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lq	.9066805	.025625	35.38	0.000	.8564565	.9569045
lpf	.4227784	.0140248	30.15	0.000	.3952904	.4502665
lf	-1.064499	.2000703	-5.32	0.000	-1.456629	-.672368
_cons	9.627909	.210164	45.81	0.000	9.215995	10.03982
sigma_u	.12488859					
sigma_e	.06010514					
rho	.81193816	(fraction of variance due to u_i)				

여기서 sigma_u=.12488859는 group means모형의 $\sqrt{\widehat{\sigma_v^2}}$ 이고 sigma_e=.06010514는 LSDV 모형의 $\sqrt{\widehat{\sigma_\epsilon^2}}$ 이다.

참 고 문 헌

곽상경(2003), 『계량경제학』, 제5판, 다산출판사.

김기화(1990), 『경기순환이론』, 다산출판사.

김명직·장국현(2002), 『금융시계열분석』, 제2판, 홍문사.

남준우·이한식(2005), 『계량경제학 이론과 EViews·Excel 활용』, 홍문사.

서진교, 패널자료 분석방법, 『농촌경제』 제24권 제2호, 2001.

민인식·최필선(2008), 『STATA 기초적 이해와 활용』, 한국STATA학회.

민인식·최필선(2009), 『STATA 기초통계와 회귀분석』, 한국STATA학회.

송일호·정우수(2002), 『SAS와 EVIEWS를 이용한 계량경제 실증분석』, 삼영사.

안충영 외 역(1998), 『기초계량경제학』, 제3판, 진영사.

Baltagi, B.(1995), *Econometric Analysis of Panel Data*, John Wiley & Sons, Inc.

EViews user's Guide. Version 5.1, QMS(2005).

Green, W.(2000), *Econometric Analysis*, Fourth Edition, Prentice Hall.

Gujarati. D.(2004). *Basic Econometrics*, Fourth Edition, McGraw-Hill.

Hamilton, L.(2006), *Statistics with Stata*, Thomson.

Hill, R., W. Griffiths and G. Judge(2001), *Undergraduate Econometrics*, Second Edition, Wiley.

Stock, J. and M. Watson(2007), *Introduction to Econometrics*, Second Edition, Pearson Education, Inc.

Wooldridge, J.(2006), *Introductory Econometrics A Modern Approach*, Thomson.

부록 1

EViews 사용설명

1.EViews 기본사용
2.기본분석
3.회귀분석

부록 1 EViews 사용설명

1.EViews 기본사용

(1)EViews란?

1994년 초에 보급되기 시작한 윈도우형 EViews(Econometric Views)는 계량분석과 프로그래밍이 동시에 가능한 소프트웨어로써 빠른 속도로 이용자의 층과 범위를 확대하고 있다.

EViews는 윈도우에 DOS에서 사용하는 명령어를 직접 입력할 수 있다는 의미에서 DOS형 Micro TSP 7.0과 정확히 동일한 기능을 수행한다. 일단 익숙하게 되면 EViews는 DOS형 Micro TSP에 비해 훨씬 융통성 있고 편리하다.

본 사용설명서는 EViews version 5.1에 관한 것이다. 자세한 사용설명은 EViews-User's Guide-를 이용하거나 Help를 이용하면 된다.

(2)EViews로 무엇을 하나?

①자료처리
-시계열자료를 입력·확장하고 수정한다.
-공식에 의해 새로운 자료를 생성한다.
-자료를 이용하여 여러 종류의 그림을 그린다.
-기술통계량(상관계수, 공분산, 자기상관, 교차상관, 히스토그램)을 계산한다.
-여러 형태의 자료를 불러 들여오거나 내보낸다.

②추정과 예측
-보통최소자승법
-비선형최소자승법
-방정식시스템
-벡터자기회귀모형
-회귀분석에 근거한 예측
-연립방정식모형

(3)EViews 시작하기

EViews아이콘을 클릭한 후 새로운 workfile을 만들거나 기존의 workfile을 open할 수 있다.

①새로운 workfile 만들기

주메뉴 바에서 File/New/workfile을 선택하면 자료의 종류(frequency)를 물어보는데 자료의 종류, 시작 및 종료를 입력하고 OK를 누르면 된다.

②workfile의 열기

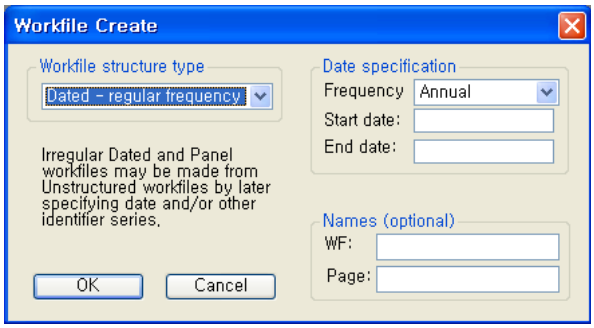
주메뉴 바에서 File/Open을 선택한 후 open하고자 하는 workfile을 선택한 후 OK를 클릭하면 된다.

(4)EViews용 Data set 만들기

①EViews에서 직접 자료를 입력하기

주메뉴 바에서 File/New/workfile을 선택하면 <그림 1>과 같은 대화상자가 나타나는데 여기서 자료의 종류(frequency)를 선택하고 <표 1>과 같은 방법으로 시작 일자와 종료 일자를 입력한다.

<그림 1> Workfile Range

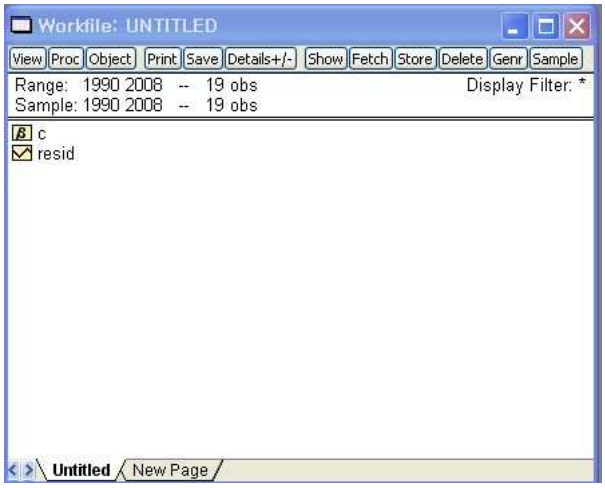


<표 1> 자료의 종류 및 입력방법

Frequency	Start date	End date	입력방법
monthly	Jan. 1950	Dec. 1994	1950:01 1994:12
quarterly	1st Q. 1950	3rd Q. 1994	1950:1 1995:3
annual	1950	1994	1950 1994
daily	Mar. 10, 1950	Dec. 7, 1994	3:10:1950 12:7:94
undated	30 observations		1 30

예를 들어 (연습 1)에 나타난 것과 같은 자료를 입력하기 위해서는 Annual을 선택하고 Start date에 1990, End date에 2008을 입력하고 OK를 클릭하면 <그림 2>와 같은 화면이 나타난다.

<그림 2> workfile 화면

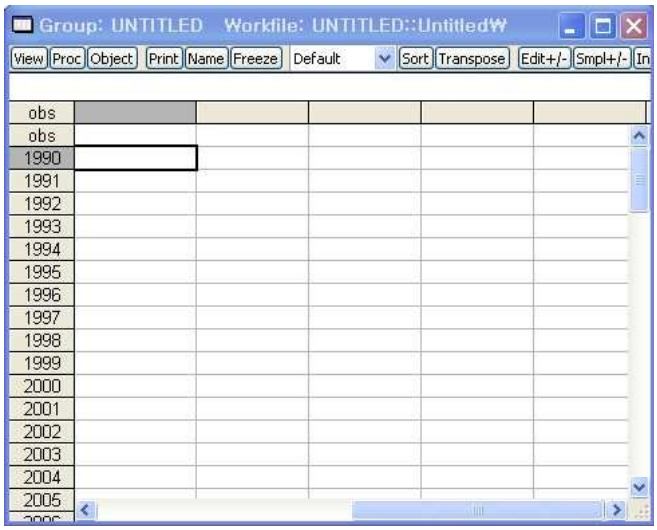


Quick-Empty Group(Edit Series)를 선택하면 나타나는 화면에서 오른쪽 스크롤바를 제일 위로 스크롤하면 다음의 <그림 3>과 같이 된다.

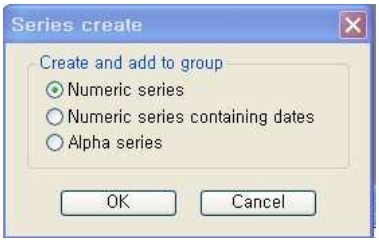
두 번째 obs 우편에 있는 셀을 클릭하여 계열의 이름(예를 들면 consumption)을 입력하고 Enter를 누르면 <그림 4>와 같은 대화상자가 나타나는데 OK를 누르면 <그림 5>와 같은 자료입력 창이 나타난다.

시계열 데이터를 순서대로 입력하거나 해당연도의 우편에 있는 셀을 클릭하여 데이터를 입력한 후 창을 종료하고 OK를 누르면 하나의 계열이 완성된다.
동일한 방법으로 원하는 계열을 모두 입력하면 된다.
자료 입력을 마치고 File/Save as를 선택하여 파일로 저장하면 되는데 이 때 확장자는 wf1이 된다.

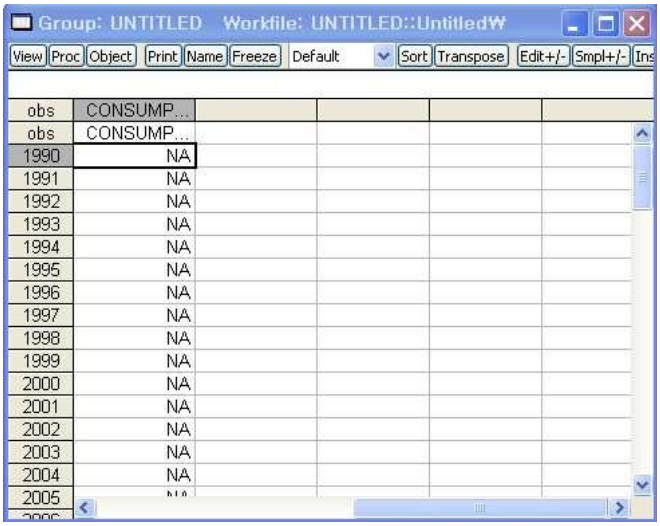
<그림 3> 계열 만들기의 준비



<그림 4> 계열 만들기



<그림 5> 계열의 입력

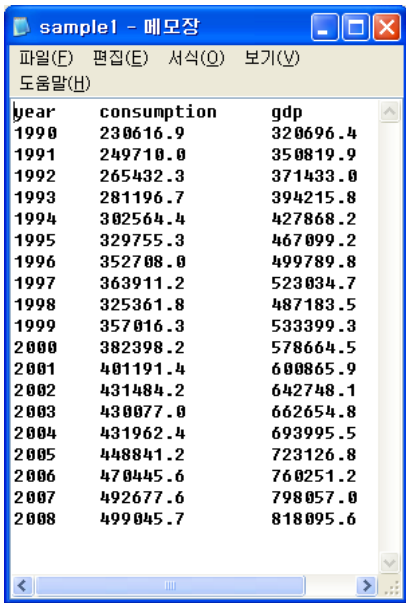


(연습1)다음과 같은 자료를 입력하고 이를 sample1.wf1으로 저장해 보라.

②외부에서 작성된 자료를 불러 들여오기

예를 들어 다음의 <그림 6>와 같이 외부에서 작성된 ASCII-TEXT 파일을 EViews 내로 불러 들여오기 위해서는 먼저 workfile window의 메뉴에서 File/New/workfile을 선택하여 <그림 1>과 같은 대화상자가 나타나면 자료의 종류와 시작 일자와 종료 일자를 입력한다.

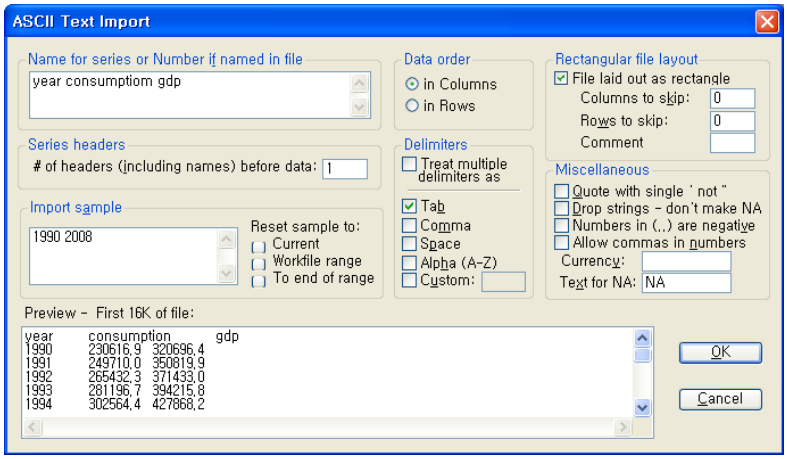
<그림 6> ASCII-TEXT 파일



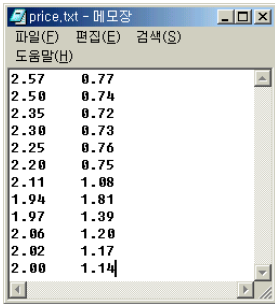
다음으로 File/Import/Read Text-Lotus-Excel을 선택한 후 불러 올 파일을 지정하면 다음의 <그림 7>과 같은 화면이 나타나는데 여기서 자료의 계열 이름을 입력하고 OK를 클릭하면 된다.

(주의)ASCII-TEXT 파일을 Import할 때 주의할 점은 ASCII-TEXT 파일의 자료형식인데 자료가 콤마가 없는 형태로 입력되어야 한다.

<그림 7> ASCII TEXT Import 화면



(연습2)다음을 Eviews로 불러 들여 sample2.wf1로 저장해 보라.

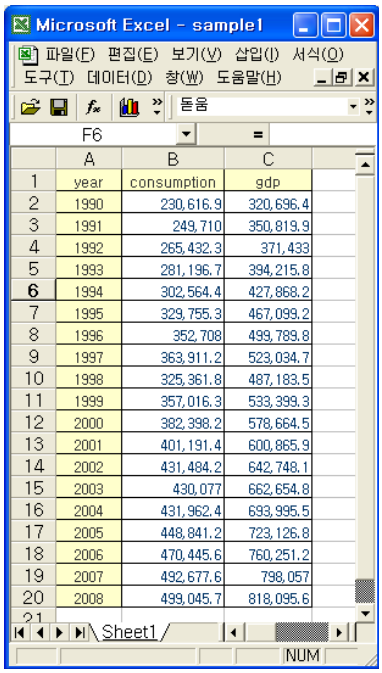


또한 다음의 <그림 8>과 같이 Excel에서 작성된 파일을 EViews에서 불러 들여 올 수 있는데 그 방법은 위와 유사하다.

먼저 workfile window의 메뉴에서 File/New/workfile을 선택하여 <그림 1>과 같은 대화상자가 나타나면 자료의 종류와 시작 일자와 종료 일자를 입력한다.

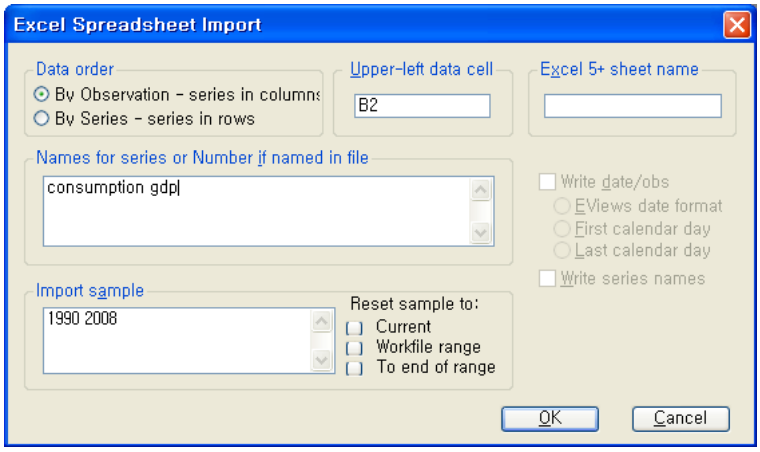
다음으로 File-Import-Read Text-Lotus-Excel을 선택한 후 불러 올 파일을 지정하면 다음의 <그림 9>와 같은 화면이 나타나는데 여기서 자료의 계열 이름을 입력하고 OK를 클릭하면 된다.

<그림 8> Excel 파일



	A	B	C
1	year	consumption	gdp
2	1990	230,616.9	320,696.4
3	1991	249,710	350,819.9
4	1992	265,432.3	371,433
5	1993	281,196.7	394,215.8
6	1994	302,564.4	427,868.2
7	1995	329,755.3	467,099.2
8	1996	352,708	499,789.8
9	1997	363,911.2	523,034.7
10	1998	325,361.8	487,183.5
11	1999	357,016.3	533,399.3
12	2000	382,398.2	578,664.5
13	2001	401,191.4	600,865.9
14	2002	431,484.2	642,748.1
15	2003	430,077	662,654.8
16	2004	431,962.4	693,995.5
17	2005	448,841.2	723,126.8
18	2006	470,445.6	760,251.2
19	2007	492,677.6	798,057
20	2008	499,045.7	818,095.6

<그림 9> Excel 파일 불러오기



(주의)Excel 파일을 import할 때 주의할 점은 <그림 9>의 Upper-left data cell에 import할 최초 자료의 주소를 정확하게 지정해 주어야 한다.

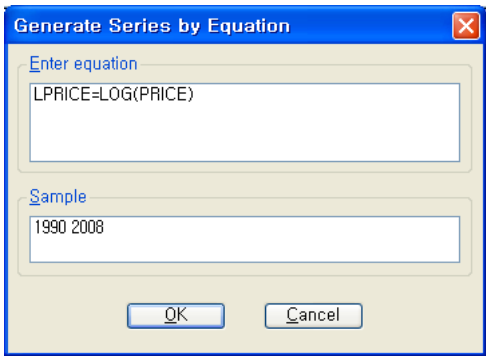
③EViews용 자료파일을 여러 가지 형태의 자료로 내 보내기

File/Export를 선택한 후 유형과 파일명을 지정하면 된다.

(5)새로운 계열 만들기

기존의 계열을 이용하여 새로운 계열을 만들 수 있다. 예를 들어 LPRICE=LOG(PRICE)는 PRICE의 자연대수의 로그치를 계산하여 LPRICE를 만든다.
새로운 계열을 만들기 위해서는 workfile window의 tool bar에서 Quick/Generate Series를 선택한 후 <그림 10>과 같은 Equation dialog box에 수식을 타이프하면 된다.

<그림 10>계열 만들기 대화상자



(예)
DM2 = M2-M2(-1)
⇒ 차분변수 만들기(difference between M2(t)-M2(t-1))

LOW= INCOME<1000 or EDUC<6
⇒LOW라는 변수는 INCOME이 1000보다 작거나 EDUC가 6보다 작으면 1의 값을 취하고, 그 외 경우이면 0의 값을 취하는 가변수이다.

2.기본분석

(1)그림 그리기

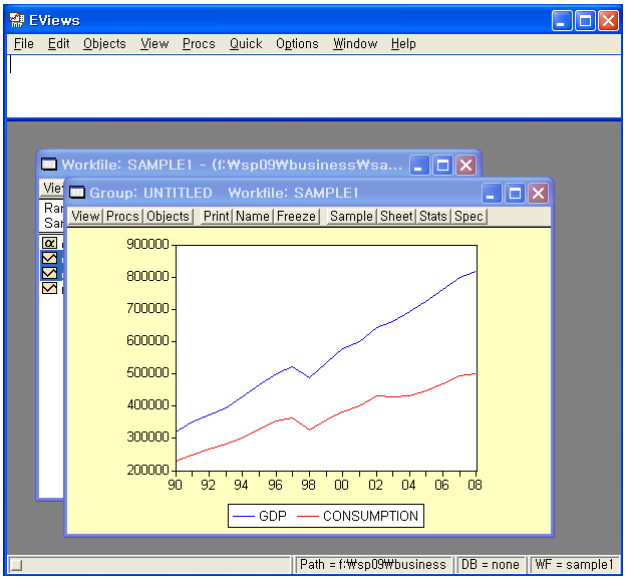
①workfile의 window에 있는 하나의 계열

workfile의 window에 있는 여러 계열 중 기술통계량을 구하고자 하는 하나의 계열을 선택하여 마우스로 두 번 누르면 자료가 있는 윈도우(이를 series window라 한다)가 나타난다. series window의 tool bar에 있는 view를 click한 후 Line graph 또는 Bar graph를 그릴 수 있다.

②workfile의 group window에서 여러 계열

여러 계열(sample1.wf1의 GDP 및 Consumption)을 선택하여 마우스 오른 쪽을 클릭하여 Open/as Group으로 그룹을 만든 후 그룹 내 계열들을 동시에 그림으로 그릴 수 있다. view/graph를 선택하고 적당한 그림의 종류를 선택하면 다음과 같은 <그림 11>이 완성된다.

<그림 11> 계열 그리기



(2)기술통계량의 계산

①workfile의 window에 있는 하나의 계열

workfile의 window에 있는 여러 계열 중 기술통계량을 구하고자 하는 하나의 계열을 선택하여 마우스로 두 번 누르면 자료가 있는 윈도우가 나타난다.
series window의 tool bar에 있는 view를 click하여 Descriptive Statistics/Histogram and Stat를 선택하면 히스토그램을 그리고 기초통계량을 계산해 준다.

(참고)EViews에서 만들어진 output이나 graph를 흔글에 붙여 넣기 위해서는 그림을 한 번 클릭하여 선택한 후 Edit/Copy를 누르면 copy to clipboard로 클립보드에 그림을 임시로 저장하거나 copy to disk file로 Windows Metafile 형식(wmf)의 그림 파일로 저장할 수 있다.

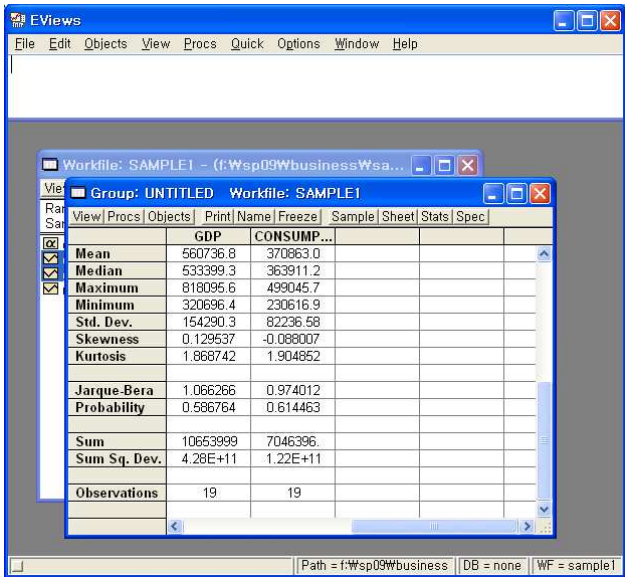
②workfile의 group window에서 여러 계열

여러 계열을 선택하여 마우스 오른 쪽을 클릭하여 Open/as Group으로 그룹을 만든 후 그룹 내 계열들의 기술통계량을 계산할 수 있다.
먼저 view를 선택하고 다음으로 Descriptive Stats/Common Sample를 선택하면 다음의 <그림 12>와 같이 평균, 중위수, 최대치, 최소치, 표준편차, 왜도, 첨도 등 여러 가지 통계량을 얻을 수 있다.

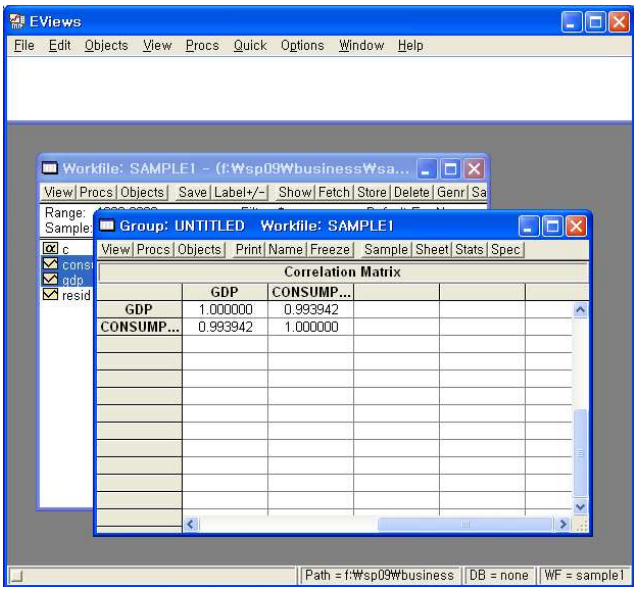
(3) 상관계수의 계산

여러 계열을 선택하여 마우스 오른쪽을 클릭하여 Open/as Group으로 그룹을 만든 후 그룹 내 계열들의 상관계수(또는 공분산)를 계산할 수 있다.
먼저 view를 선택하고 다음으로 correlations(또는 covariances)를 선택하면 다음의 <그림 13>과 같은 상관계수를 얻을 수 있다.

<그림 12> 기술통계량 계산



<그림 13> 상관계수의 계산

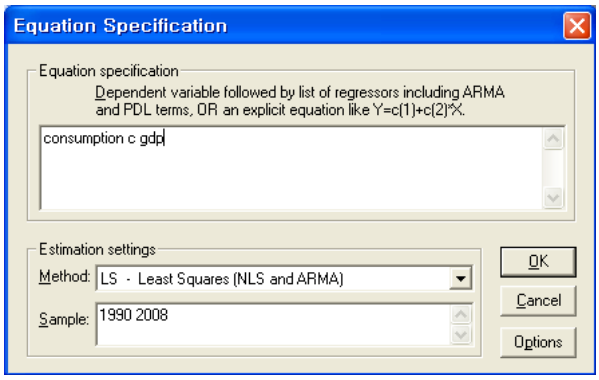


3.회귀분석

(1)단순회귀분석

단순회귀모형을 추정하기 위해서는 주메뉴에서 Quick/Estimate equation을 선택하면 다음의 <그림 14>와 같은 Equation Specification 대화상자가 나타난다.

<그림 14> Equation Specification 대화상자



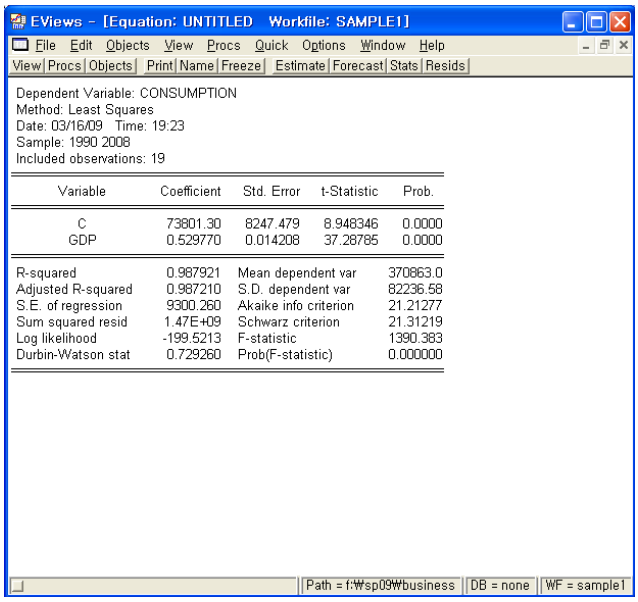
예를 들어 소득과 소비의 관계를 나타내는 소비함수를 추정한다고 하자. 여기서 종속변수(소비:consumption), 상수항(C) 및 독립변수(소득: gdp)를 입력하고 OK를 누르면 <그림 15>와 같은 결과를 준다.

(연습3)1970년 1/4분기부터 2008년 4/4분기까지 국내총생산자료를 나타내는 gdpq70-08.xls를 이용하여 다음을 수행하라

-gdpq70-08.xls에 있는 gdp계열을 EViews로 불러들여라
-gdp계열을 계절조정을 하라(단, 계절조정은 Census X-12로 하고 X-11 method는 additive를 선택하라)
-계절 조정된 gdp(gdp_sa)를 로그치로 변환한 lgdp_sa계열을 만들고 그림을 그려보라
-lgdp_sa를 종속변수로 하고 시간추세(time trend)를 독립변수로 하는 단순회귀모형 즉, y_{t}= + žt + i_{t}을 추정한 후(이때 시간추세는 @trend로 표시하면 됨)

분기평균증가율 및 연평균증가율을 계산하고, 실제자료의 분기평균 및 연평균증가율과 비교해 보라
-이상의 작업을 sample3.wf1로 저장하라

<그림 15> 단순회귀분석 출력결과

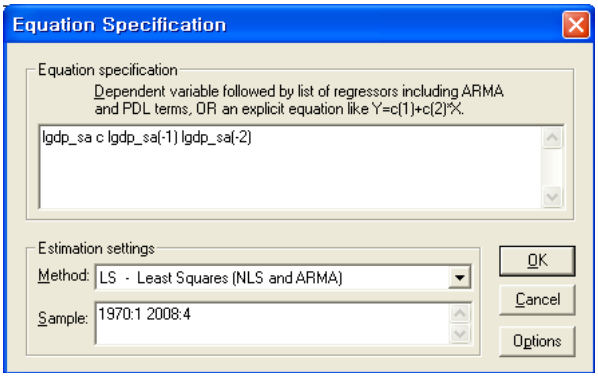


(2)다중회귀분석

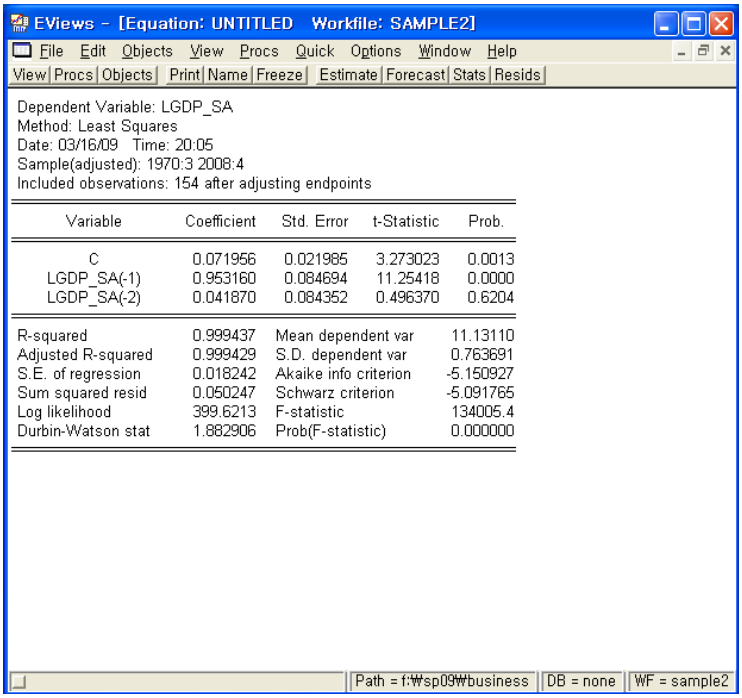
다중회귀모형을 추정하기 위해서는 주메뉴에서 Quick/Estimate equation을 선택하면 Equation Specification 대화상자가 나타난다.

예를 들어 계절 조정된 gdp가 전년도 gdp 및 전전년도 gdp에 영향을 받는다고 하는 다중회귀모형을 추정한다고 하자. 다음의 <그림 16>와 같이 종속변수(lgdp_sa), 상수항(C) 및 2개의 독립변수 즉, 전년도 gdp(lgdp_sa(-1)) 및 전전년도 gdp(lgdp_sa(-2))를 입력하고 OK를 누르면 <그림 17>과 같은 결과를 준다.

<그림 16> 다중회귀모형 추정식

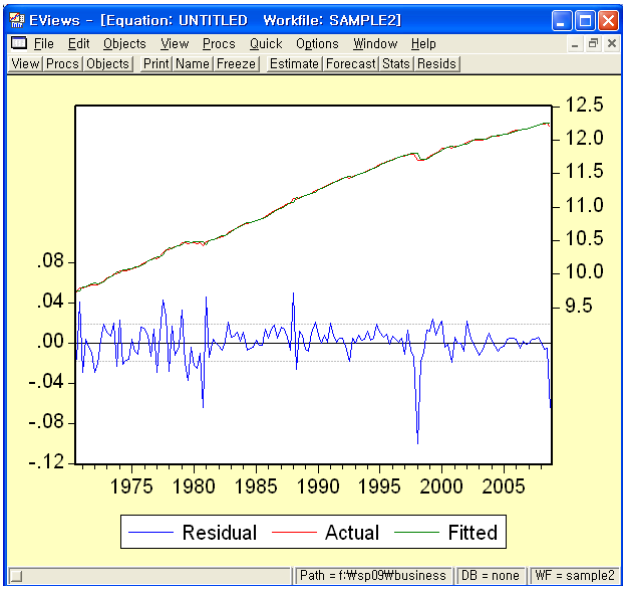


<그림 17> 다중회귀분석 출력 결과



(참고) 회귀분석 출력 후 View를 누르면 Representations, Estimation Output, Actual, fitted, residual, Covariance Matrix, and several tests 등 여러 가지를 얻을 수 있다. 다음의 <그림 18>은 Actual, Fitted, Residual Graph를 실행시킨 결과이다.

<그림 18> 다중회귀분석 출력결과



(3)예측(Forecasting)

equation tool bar에서 Procs/Forecast를 클릭하면 시계열의 추정치를 구할 수 있다.

예를 들어 sample2.wf1의 자료를 이용하여 2002년 및 2003년의 수요량을 예측할 경우 다음과 같은 순서로 하면 된다.

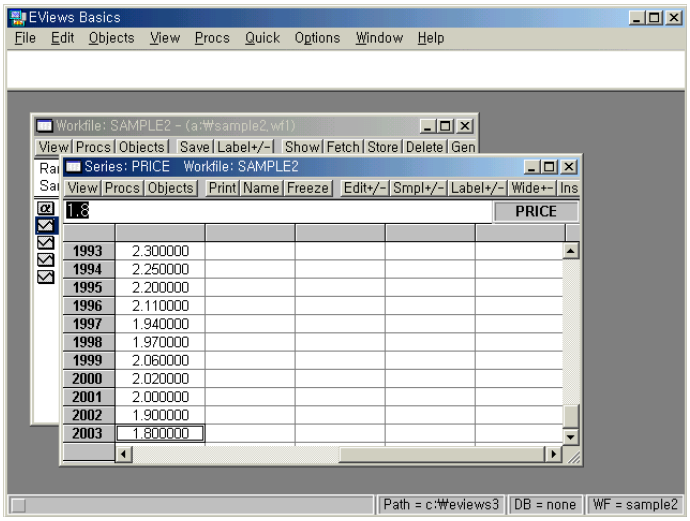
첫째, workfile window에서 Proc/structure-resize current page를 선택한 후 1990 2003을 입력한다.

둘째, 다음의 <그림 19>와 같이 독립변수의 값을 먼저 주어야 한다(여기서는 2002부터 2003년까지)

셋째, Quick/Estimate equation을 선택하여 단순회귀모형을 설정하고 추정기간

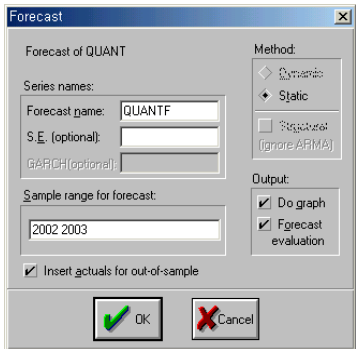
은 1990 2001로 하여 추정한다.

<그림 19> 독립변수 값 부여

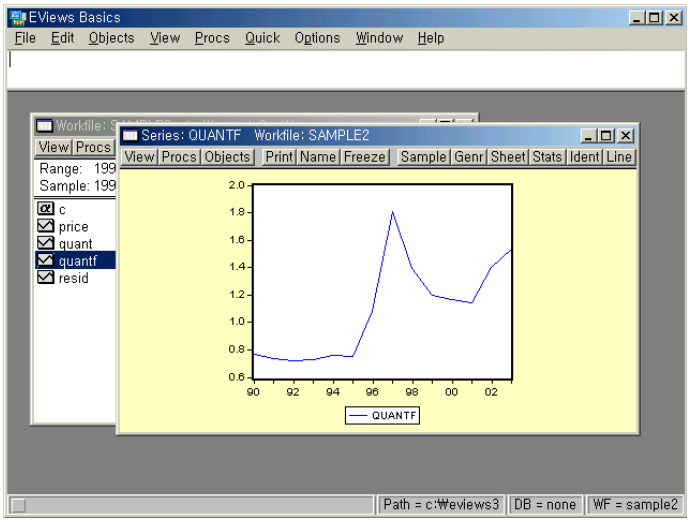


넷째, Forecast를 선택하면 다음의 <그림 20>과 같은 화면이 나타나는데 sample range for forecast에 2002 2003을 입력한다. OK를 클릭하면 수요량의 예측치는 QUANTF라는 계열로 저장된다.(S.E.란에 SEF를 입력하면 예측치의 표준오차가 계산됨)

<그림 20> Forecast 대화상자



다섯째, sample을 1990 2003으로 하고 수요량의 실제치 및 예측치를 그려 보면 다음과 같다.



여섯째, 다시 sample을 원위치(1990 2001)로 한다.

부록 2

Stata 사용설명

1.Stata 기본사용

2.기본분석

3.회귀분석

부록 2 Stata 사용설명

1.Stata 기본사용

(1)Stata란?

Stata는 Statistics와 Data의 합성어로 1980년대 중반 미국의 StataCorp가 개발한 통계 소프트웨어 패키지이다. Stata는 경제학을 비롯한 사회과학은 물론, 의학 등 자연과학에서도 널리 이용되고 있다.

Stata는 SAS 및 SPSS보다 20여년 늦게 개발되었으나 이들 통계 패키지를 점점 대체해 나가고 있는데 그 이유는 Stata가 다음과 같은 장점을 가지고 있기 때문이다.

첫째, Stata는 통계분석은 물론 데이터 관리(data management)와 그래픽에서 탁월한 기능을 가지고 있다.

둘째, Stata는 광범위한 통계분석이 가능하여 매우 다양한 분야의 사용자들의 요구에 부합된다.

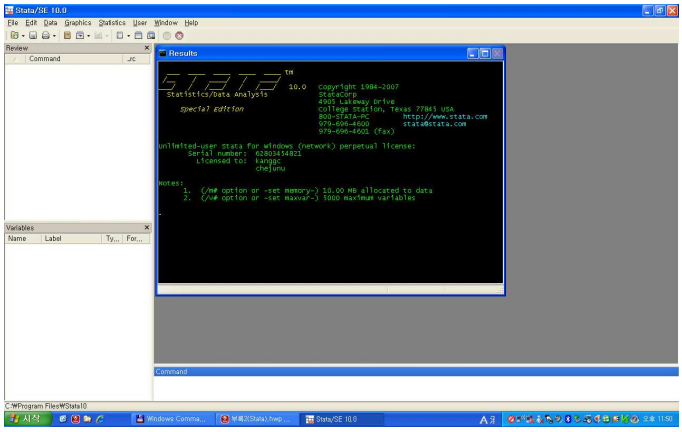
셋째, Stata는 인터넷과 상호작용을 통해 다양한 부가적인 기능을 발휘한다.

(2)Stata 시작하기

Stata아이콘을 클릭하면 <그림 1>과 같이 Results 창, Review 창, Variables 창, Command 창 등 4개의 창이 나타난다.

- Command 창 : 실행하고자 하는 명령어를 입력하는 창
- Review 창 : Command 창에서 입력한 명령문들이 모이는 장소이며 이 창에 있는 명령문을 한 번 클릭하면 command 창에 해당 명령문이 그대로 입력된다.
- Variables 창 : 데이터 파일에 포함되어 있는 변수이름, 변수의 레이블, 데이터의 포맷 형태 등을 보여 주는 창
- Results 창 : Stata 명령문이 실행되어 결과가 제시되는 창으로 이 창의 내용은 텍스트 형식이므로 복사하여 다fms 문서작성 프로그램에서 붙여넣기가 가능하다.

<그림 1> Stata 4개의 창



(3)명령어 실행방법

Stata에서 명령어를 실행시키는 방법에는 명령어를 입력하는 방법과 대화창(dialog box)을 이용하는 방법 등 2가지가 있다.

①명령어를 입력하는 방법

Command 창에서 명령어를 직접 입력하여 엔터를 쳐서 실행한다. Command 창에서 Stata 프로그램 내부에 있는 데이터 파일인 auto.dta를 불러들이는 명령어인 sysuse를 이용하여 Command 창에서 다음과 같이 입력하고 엔터를 치면 <그림 1>이 <그림 2>와 같이 변한다.

·sysuse auto

②대화창을 이용하는 방법

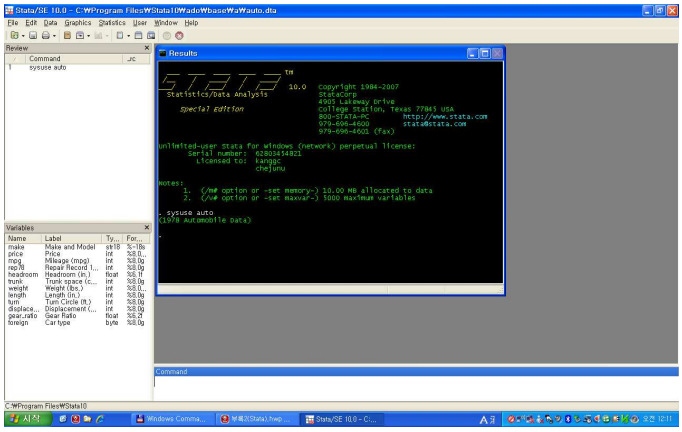
Stata는 많은 종류의 명령어가 있기 때문에 이를 모두 외워서 활용하는데 한계가 있다. 따라서 대화창을 이용하면 Stata의 다양한 명령어를 배울 수 있고 특히 명령어를 알고 있지만 명령어의 문법을 모를 경우 이 방법이 매우 유용하다.

요약통계량을 구하는 명령어인 summarize를 대화창을 사용하는 명령어인 db와

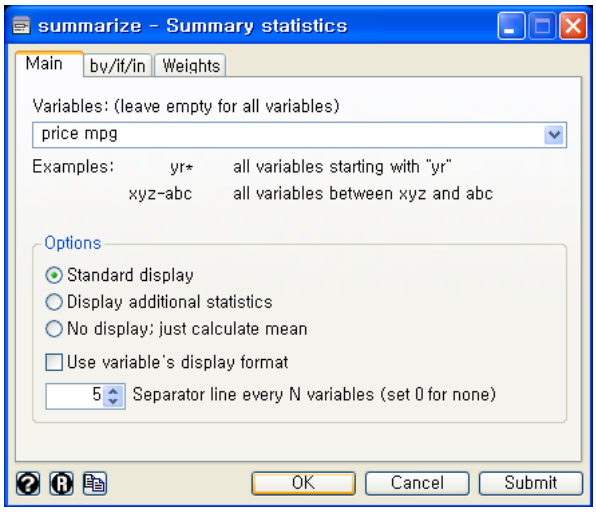
함께 다음과 같이 입력하면 <그림 3>과 같은 summarize 대화창이 나타나는데 그 대화창에서 요약통계량을 구하고자 하는 변수를 선택한 후 Ok를 누르면 <그림 4>와 같은 summarize 결과가 나타난다.

```
db summarize
```

<그림 2> Stata 작업창의 변화



<그림 3> summarize 대화창



<그림 4> summarize 결과

```
. db summarize
. summarize price mpg
```

Variable	Obs	Mean	Std. Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41

한편, <그림 3>의 summarize 대화창의 Options에서 Display additional statistics를 선택하고 OK를 클릭하면 <그림 5>와 같이 백분위수(percentiles), 중앙값(50%), 분산(variance), 왜도(skewness), 첨도(kurtosis) 등 다양한 요약 통계량을 얻을 수 있다.

<그림 5> summarize 결과(상세결과)

```
. summarize price, detail
```

Price

Percentiles		Smallest		
1%	3291	3291		
5%	3748	3299		
10%	3895	3667	Obs	74
25%	4195	3748	Sum of wgt.	74
50%	5006.5		Mean	6165.257
		Largest	Std. Dev.	2949.496
75%	6342	13466		
90%	11385	13594	Variance	8699526
95%	13466	14500	Skewness	1.653434
99%	15906	15906	Kurtosis	4.819188

(4)Stata용 Data set 만들기

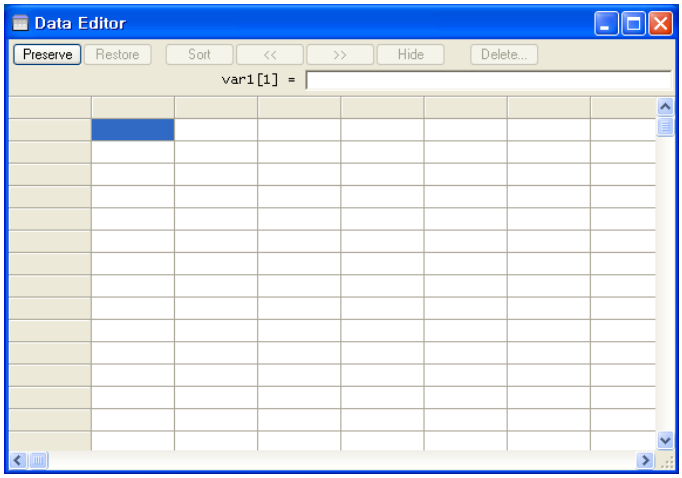
①Stata에서 직접 자료를 입력하기

Stata에서 숫자나 문자를 직접 입력하여 데이터 파일을 만드는 방법에는 Data Editor 창을 이용하는 방법과 input 명령어를 이용하는 방법이 있다.

Data Editor를 이용하는 방법으로는 주메뉴 바에서 Data-Data editor를 선택하거나 Command 창에서 edit 명령어를 입력하면 <그림 6>과 같은 Data Editor 창이 만들어지는데 엑셀에서 데이터를 입력하는 방법과 유사하게 데이터를 입력하면 된다.

한편, 프로그래밍을 할 경우에는 Data Editor 창을 이용한 데이터 입력방식을 사용할 수 없으므로 이 경우 input 명령어를 활용할 수 있다. 예를 들어 연도별-지역별 GRDP를 보여주는 데이터는 Command 창에서 <그림 7>과 같은 명령어를 입력하면 된다.

<그림 6> Data Editor 창



<그림 7> input을 이용한 데이터 입력

```
. input year str10 region grdp

        year      region       grdp
  1. 2005 서울 1500
  2. 2005 부산 1200
  3. 2005 제주 1300
  4. end
```

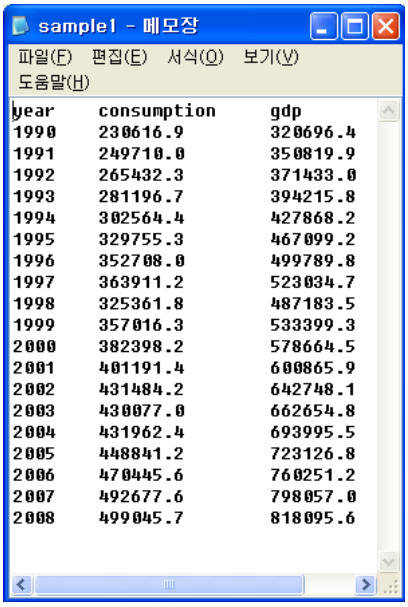
②외부에서 작성된 자료를 불러 들여오기

Stata 데이터 파일의 확장자는 dta이다. 주메뉴에서 File-Open을 이용하여 Stata 데이터 파일을 불러들일 수 있다.

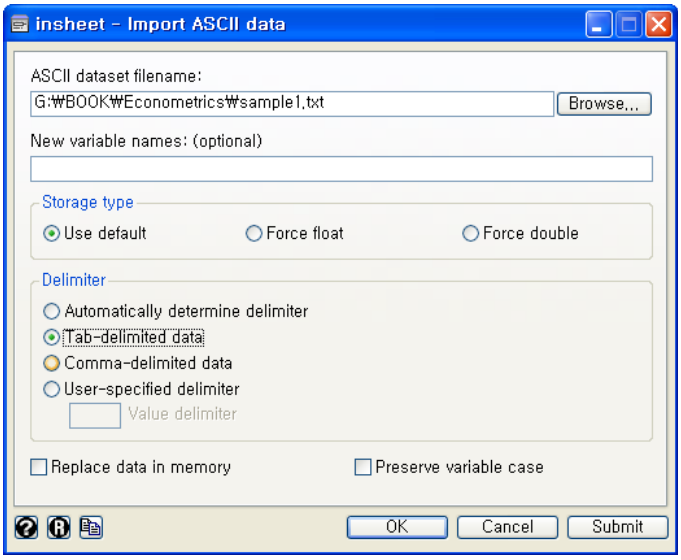
한편, <그림 8>과 같이 외부에서 작성된 ASCII-TEXT 파일을 Stata 내로 불러 들여오기 위해서는 먼저 주메뉴에서 File-import-ASCII data created by a spreadsheet를 선택하여 <그림 9>와 같은 대화창이 나타나면 ASCII-TEXT 파일이 있는 위치와 적당한 Delimiter를 선택해 주고 OK를 클릭하면 <그림 8>의 ASCII-TEXT 파일을 Stata 내로 불러 들여온다.

이렇게 Stata 내로 불러 들여온 파일은 File-save를 선택한 후 적절한 디렉토리와 파일명(예: sample1.dta)을 지정해 주면 Stata 데이터 파일로 저장된다.

<그림 8> ASCII-TEXT 파일



<그림 9> import 대화창



(5)데이터관리 기본 명령어

Stata의 장점 중의 하나는 데이터관리가 매우 편하다는 점이다.

① describe(줄여서 d만 입력해도 됨)

메모리로 불러들인 데이터 파일이나 특정 디렉토리에 저장되어 있는 데이터 파일의 전반적인 속성 즉 디렉토리 위치, 관측치 개수, 변수 개수 등의 기본적인 정보와 각 변수들의 저장형태, 디스플레이 포맷, 레이블 등 기술적인 정보들을 보여준다.

```
Contains data
  obs:            19
 vars:             3
 size:           266 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
year            int    %8.0g
consumption     float  %9.0g
gdp             float  %9.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

② list

데이터 파일의 모든 값들을 Results 창에 그대로 보여준다.

	year	consum~n	gdp
1.	1990	230616.9	320696.4
2.	1991	249710	350819.9
3.	1992	265432.3	371433
4.	1993	281196.7	394215.8
5.	1994	302564.4	427868.2
6.	1995	329755.3	467099.2
7.	1996	352708	499789.8
8.	1997	363911.2	523034.7
9.	1998	325361.8	487183.5
10.	1999	357016.3	533399.3
11.	2000	382398.2	578664.5
12.	2001	401191.4	600865.9
13.	2002	431484.2	642748.1
14.	2003	430077	662654.8
15.	2004	431962.4	693995.5
16.	2005	448841.2	723126.8
17.	2006	470445.6	760251.2
18.	2007	492677.6	798057
19.	2008	499045.7	818095.6

③ rename

이미 주어진 변수이름을 변경하고자 할 때 사용하는 명령어이다.

·rename consumption consume

④ sort

데이터를 정렬하는 명령어로 여러 변수가 있을 경우 어떤 변수를 기준으로 정렬할 것인 지를 결정해 주어야 한다.

·sysuse auto
·sort price
·list make price in 1/10

	make	price
1.	Merc. Zephyr	3,291
2.	Chev. Chevette	3,299
3.	Chev. Monza	3,667
4.	Toyota Corolla	3,748
5.	Subaru	3,798
6.	AMC Spirit	3,799
7.	Merc. Bobcat	3,829
8.	Renault Le Car	3,895
9.	Chev. Nova	3,955
10.	Dodge Colt	3,984

(참고) bysort 명령어는 by와 sort가 결합된 것으로 by는 Stata에서 prefix의 한 종류인데 어떤 기준을 의미한다.

⑤ summarize

어떤 변수의 요약 통계량(평균, 표준편차 등)을 계산하게 하는 명령어이다.

·summarize price

Variable	Obs	Mean	Std. Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906

⑥ drop

데이터 파일의 여러 변수들 중 특정 변수를 삭제하기 위하여 사용하는 명령어이다.

```
sysuse auto
drop price mpg
```

또는 모든 변수를 삭제할 수도 있다.

```
sysuse auto
drop _all
```

⑦ generate(줄여서 gen)

기존의 변수를 이용하여 새로운 변수로 만들게 하는 명령어이다.

```
use "<해당 디렉토리>sample1.dta", clear
gen ln_consume = log(consumption)
```

	ln_con~e	consum~n
1.	12.34851	230616.9
2.	12.42806	249710
3.	12.48911	265432.3
4.	12.54681	281196.7
5.	12.62005	302564.4
6.	12.70611	329755.3
7.	12.7734	352708
8.	12.80467	363911.2
9.	12.69269	325361.8
10.	12.78554	357016.3
11.	12.85422	382398.2
12.	12.90219	401191.4
13.	12.97499	431484.2
14.	12.97172	430077
15.	12.97609	431962.4
16.	13.01442	448841.2
17.	13.06144	470445.6
18.	13.10761	492677.6
19.	13.12045	499045.7

(참고)drop으로 어떤 변수들을 삭제했다고 해서 데이터 파일에서 해당 변수들이 없어지는 것은 아니고 메모리에서만 사라진다. 따라서 데이터 파일 자체를 변경하기 위해서는 File-Save를 통해 다시 저장해 주어야 한다.

Stata에서는 프로그램 내부에서 사용하고 있는 변수인 시스템변수가 있는데 예를 들면 -n은 어떤 관측치에 이르기까지 몇 개의 관측치가 있는 지를 나타내 주는 변수이다.

```
use "<해당 디렉토리>sample1.dta", clear
gen var2 = _n
list var2
```

	var2
1.	1
2.	2
3.	3
4.	4
5.	5
6.	6
7.	7
8.	8
9.	9
10.	10
11.	11
12.	12
13.	13
14.	14
15.	15
16.	16
17.	17
18.	18
19.	19

⑧ replace

기존 변수의 값을 새로운 값으로 대체하는 명령어이다.

```
replace gdp = gdp/1000
```

명령어 generate와 replace를 결합하면 범주형 변수를 만드는데 유용하다.

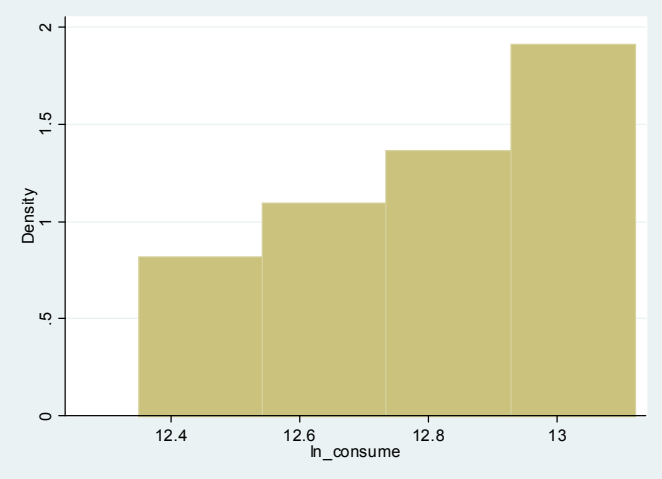
2.기본분석

(1)그림 그리기

①히스토그램

히스토그램을 그리는 명령어이다.

. histogram ln_consume



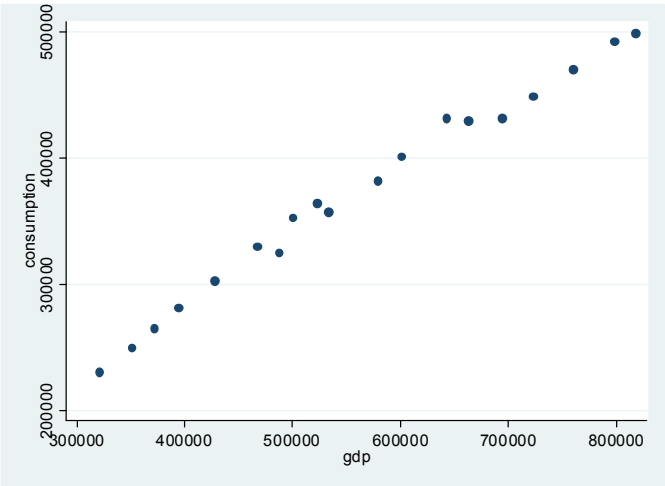
②산포도

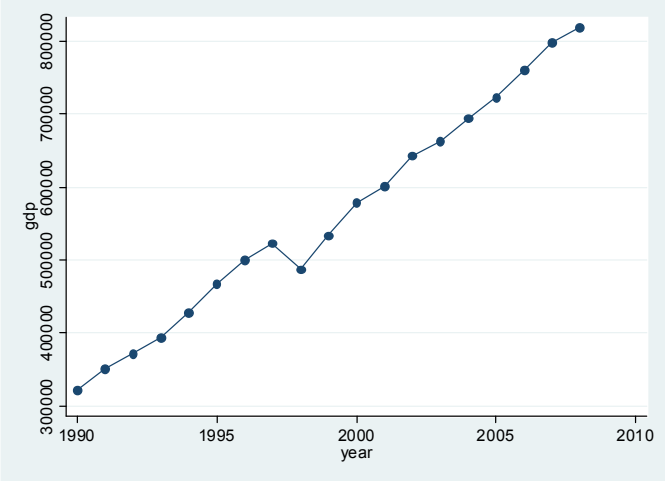
두 변수의 모든 관측치들을 XY 평면에 표시하는 명령어이다.

. twoway(scatter consumption gdp)

③선그래프

. twoway(connected gdp year)





(2) 상관계수의 계산

두 변수 간의 선형관계의 정도를 측정한다.

```
. corr(consumption gdp)
```

```
. corr consumption gdp
(obs=19)
```

	consum~n	gdp
consumption	1.0000	
gdp	0.9939	1.0000

3.회귀분석

(1)단순회귀분석

소득(gdp)과 소비(consumption)의 관계를 나타내는 소비함수를 추정한다고 하자.

```
. reg consumption gdp
```

Source	SS	df	MS	Number of obs =	19
				F(1, 17) =	1390.38
Model	1.2026e+11	1	1.2026e+11	Prob > F =	0.0000
Residual	1.4704e+09	17	86494821.6	R-squared =	0.9879
				Adj R-squared =	0.9872
Total	1.2173e+11	18	6.7629e+09	Root MSE =	9300.3

consumption	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdp	.5297702	.0142076	37.29	0.000	.4997948	.5597456
_cons	73801.32	8247.479	8.95	0.000	56400.66	91201.98

Stata에서는 명령의 실행결과가 스칼라, 매크로, 행렬 등 다양한 형태로 저장된다. 이 중 스칼라 형태로만 저장되는 경우를 r-class라고 하고 return list라는 명령문을 입력하면 저장된 r-class를 확인할 수 있다.

한편, 모형을 추정한 후 도출되는 결과들이 스칼라, 매크로, 행렬, 함수 등의 형태로 저장되는 경우를 e-class(estimation class)라고 하며 ereturn list라는 명령문을 입력하면 저장된 e-class를 확인할 수 있다.

```
. reg consumption gdp
. ereturn list
```

```
scalars:
                  e(N) =  19
               e(df_m) =  1
               e(df_r) =  17
                  e(F) =  1390.383419106143
                 e(r2) =  .9879208467506339
               e(rmse) =  9300.259224737087
                e(mss) =  120260965856.9591
                e(rss) =  1470411968.004223
               e(r2_a) =  .9872103083242005
                 e(ll) =  -199.5213420925717
               e(ll_0) =  -241.475946842314

macros:
            e(cmdline) : "regress consumption gdp"
              e(title) : "Linear regression"
                e(vce) : "ols"
            e(depvar) : "consumption"
                e(cmd) : "regress"
         e(properties) : "b V"
            e(predict) : "regres_p"
              e(model) : "ols"
          e(estat_cmd) : "regress_estat"

matrices:
                  e(b) :  1 x 2
                  e(V) :  2 x 2
```

(2)다중회귀분석

자동차 가격(price)이 가스 마일리지(mpg)와 정비기록(rep78)에 영향을 받는다는 다중회귀모형을 추정한다고 하자.

```
. sysuse auto
. reg price mpg rep78
```

Source	SS	df	MS		Number of obs =	69
					F(2, 66) =	11.06
Model	144754063	2	72377031.7		Prob > F =	0.0001
Residual	432042896	66	6546104.48		R-squared =	0.2510
					Adj R-squared =	0.2283
Total	576796959	68	8482308.22		Root MSE =	2558.5

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-271.6425	57.77115	-4.70	0.000	-386.9864	-156.2987
rep78	666.9568	342.3559	1.95	0.056	-16.5789	1350.492
_cons	9657.754	1346.54	7.17	0.000	6969.3	12346.21

부록 3

주요통계표

<표 1> 정규분포표

<표 2> t-분포표

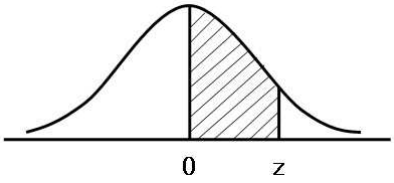
<표 3> χ^2-분포표

<표 4> F-분포표

<표 5> Durbin-watson

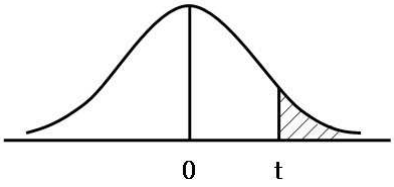
<표 6> Dickey-Fuller r-검정치 분포표

<표 1> 정규분포표



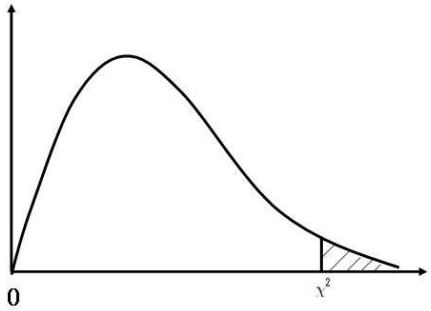
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3435	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4985	0.4985	0.4985	0.4986	0.4986
3.0	0.4986	0.4987	0.4987	0.4988	0.4988	0.4984	0.4989	0.4989	0.4990	0.4990

<표 2> t-분포표



v \ p	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.923
3	0.1638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.103
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.816
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

<표 3> χ^2-분포표



자유도	P=0.99	0.98	0.95	0.90	0.80	0.20	0.10	0.05	0.02	0.01
1	0.000157	0.000628	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.412	6.635
2	0.0201	0.0404	0.103	0.211	0.446	3.219	4.605	5.991	7.824	9.210
3	0.115	0.185	0.352	0.584	1.005	4.642	6.251	7.815	9.837	11.341
4	0.297	0.429	0.711	1.064	1.649	5.989	7.779	9.488	11.668	13.277
5	0.554	0.752	1.145	1.610	2.343	7.289	9.236	11.070	13.388	15.086
6	0.872	1.134	1.635	2.204	3.070	8.558	10.645	12.592	15.033	16.812
7	1.239	1.564	2.167	2.833	3.822	9.803	12.017	14.067	16.622	18.475
8	1.646	2.032	2.733	3.490	4.594	11.030	13.362	15.507	18.168	20.090
9	2.088	2.532	3.325	4.168	5.380	12.242	14.684	16.919	19.679	21.666
10	2.558	3.059	3.940	4.865	6.179	13.442	15.987	18.307	21.161	23.209
11	3.053	3.609	4.575	5.578	5.989	14.631	17.275	19.675	22.618	24.725
12	3.571	4.178	5.226	6.034	7.807	15.812	18.549	21.026	24.054	26.217
13	4.017	4.765	5.892	7.042	8.634	16.985	19.812	22.362	25.472	27.688
14	4.660	5.368	6.517	7.790	9.467	18.151	21.064	23.685	26.873	29.141
15	5.229	5.985	7.261	8.547	10.307	19.311	22.307	24.996	28.259	30.578
16	5.812	6.614	7.962	9.312	11.152	20.465	23.542	26.296	29.633	32.000
17	6.408	7.255	8.672	10.085	12.002	21.615	24.769	27.587	30.995	33.409
18	7.015	7.906	9.390	10.865	12.857	22.760	25.989	28.869	32.346	34.805
19	7.633	8.567	10.117	11.651	13.716	23.900	27.204	30.144	33.687	36.191
20	8.260	9.237	10.851	12.443	14.578	25.038	28.412	31.410	35.020	37.566
21	8.897	9.915	11.591	13.240	15.445	26.171	29.615	32.671	36.343	38.932
22	9.542	10.600	12.338	14.041	16.314	27.301	30.813	33.924	37.659	40.289
23	10.196	11.293	13.091	14.848	17.187	26.429	32.007	35.172	38.968	41.638
24	10.856	11.992	13.848	15.659	18.062	29.553	33.196	36.415	40.270	42.980
25	11.524	12.697	14.611	16.473	18.940	30.675	34.382	37.652	41.566	44.314
26	12.198	13.409	15.379	17.292	19.820	31.795	36.563	38.885	42.856	45.642
27	12.879	14.125	16.151	18.114	20.703	32.912	36.741	40.113	44.141	46.963
28	13.565	14.847	16.928	18.939	21.588	34.027	37.916	41.337	45.419	48.278
29	14.256	15.574	17.708	19.768	22.475	36.139	39.087	42.557	46.693	49.588
30	14.953	16.306	18.493	30.599	23.364	36.250	40.256	43.773	47.962	50.892

<표 4> F-분포표(5% 유의수준)

v2 \ v1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.09	241.14	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.76	8.74	8.70	8.66	8.64	8.62	8.59	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.48	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.37
6	5.99	4.74	7.35	4.12	3.94	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.23
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.23
8	5.32	4346	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.84	2.83	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.80	2.65	2.60	2.53	2.47	2.39	2.35	2.31	2.27	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.95	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.00

v_1: 분자의 자유도 v_2: 분모의 자유도.

<표 4(계속)> F-분포표(1% 유의수준)

v2 \ v1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	∞
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.87	6106.31	6157.28	6208.73	6234.63	6260.65	6286.78	6365.86
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.50
3	34.12	30.82	29.46	82.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	360	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.01
40	7.31	5.18	4.31	3.83	3.51	3.20	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.00

v_1: 분자의 자유도 v_2: 분모의 자유도.

<표 5> Durbin-Watson (5% 유의수준)

n	k'=1		k'=2		k'=3		k'=4		k'=5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.410	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774
90	1.636	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.778	1.665	1.802
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820

k'= 상수항을 제외한 설명변수의 수

n	k'=6		k'=7		k'=8		k'=9		k'=10	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.447	2.472	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	0.603	2.257	0.502	2.461	0.407	2.667	0.321	2.873	0.244	3.073
19	0.649	2.206	0.459	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	0.692	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	0.732	2.124	0.637	2.290	0.547	2.460	0.461	2.633	0.380	2.806
22	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.734
23	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	0.837	2.035	0.751	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	0.868	2.012	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	0.951	1.958	0.874	2.071	0.798	2.188	0.723	2.309	0.650	2.431
29	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.682	2.396
30	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.795	2.281
34	1.080	1.891	1.015	1.979	0.950	2.069	0.885	2.162	0.821	2.257
35	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.114	1.877	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.198
38	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.945	2.149
45	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.002	1.038	2.088
50	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.433	1.802	1.401	1.837	1.369	1.873	1.337	1.910	1.305	1.948
75	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

<표 5 (계속)> Durbin-Watson (1% 유의수준)

n	k'=1		k'=2		k'=3		k'=4		k'=5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.811	1.070	0.700	1.252	0.591	1.464	0.488	1.704	0.391	1.967
16	0.844	1.086	0.737	1.252	0.633	1.446	0.532	1.663	0.437	1.900
17	0.874	1.102	0.772	1.255	0.672	1.432	0.574	1.630	0.480	1.847
18	0.902	1.118	0.805	1.259	0.708	1.422	0.613	1.604	0.522	1.803
19	0.928	1.132	0.835	1.265	0.742	1.415	0.650	1.584	0.561	1.767
20	0.952	1.147	0.863	1.271	0.773	1.411	0.685	1.567	0.598	1.737
21	0.975	1.161	0.890	1.277	0.803	1.408	0.718	1.554	0.633	1.712
22	0.997	1.174	0.914	1.284	0.831	1.407	0.748	1.543	0.667	1.691
23	1.018	1.187	0.938	1.291	0.858	1.407	0.777	1.534	0.698	1.673
24	1.037	1.199	0.960	1.298	0.882	1.407	0.805	1.528	0.728	1.658
25	1.055	1.211	0.981	1.305	0.906	1.409	0.831	1.523	0.756	1.645
26	1.072	1.222	1.001	1.312	0.928	1.411	0.855	1.518	0.783	1.635
27	1.089	1.233	1.019	1.319	0.949	1.413	0.878	1.515	0.808	1.626
28	1.104	1.244	1.037	1.325	0.969	1.415	0.900	1.513	0.832	1.618
29	1.119	1.254	1.054	1.332	0.988	1.418	0.921	1.512	0.855	1.611
30	1.133	1.263	1.070	1.339	1.006	1.421	0.941	1.511	0.877	1.606
31	1.147	1.273	1.085	1.345	1.023	1.425	0.960	1.510	0.897	1.601
32	1.160	1.282	1.100	1.352	1.040	1.428	0.979	1.510	0.917	1.597
33	1.172	1.291	1.114	1.358	1.055	1.432	0.996	1.510	0.936	1.594
34	1.184	1.299	1.128	1.364	1.070	1.435	1.012	1.511	0.954	1.591
35	1.195	1.307	1.140	1.370	1.085	1.439	1.028	1.512	0.971	1.589
36	1.206	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.988	1.588
37	1.217	1.323	1.165	1.382	1.112	1.446	1.058	1.514	1.004	1.586
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.585
39	1.237	1.337	1.187	1.393	1.137	1.453	1.085	1.517	1.034	1.584
40	1.246	1.344	1.198	1.398	1.148	1.457	1.098	1.518	1.048	1.584
45	1.288	1.276	1.245	1.423	1.201	1.474	1.156	1.528	1.111	1.584
50	1.324	1.403	1.285	1.446	1.245	1.491	1.205	1.538	1.164	1.587
55	1.356	1.427	1.320	1.466	1.284	1.506	1.247	1.548	1.209	1.592
60	1.383	1.449	1.350	1.484	1.317	1.520	1.283	1.558	1.249	1.598
65	1.407	1.468	1.377	1.500	1.346	1.534	1.315	1.568	1.283	1.604
70	1.429	1.485	1.400	1.515	1.372	1.546	1.343	1.578	1.313	1.611
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.587	1.340	1.617
80	1.466	1.515	1.441	1.541	1.416	1.568	1.390	1.595	1.364	1.624
85	1.482	1.528	1.458	1.553	1.435	1.578	1.411	1.603	1.386	1.630
90	1.496	1.540	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.642
100	1.522	1.562	1.503	1.583	1.482	1.604	1.462	1.625	1.441	1.647
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725

k'=상수항을 제외한 설명변수의 수

n	k'=6		k'=7		k'=8		k'=9		k'=10	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.679	0.160	2.925
19	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.714
21	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	0.682	1.766	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.789	1.956
45	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805	0.955	1.864
55	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771	1.072	1.817
65	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761	1.120	1.802
70	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.313	1.646	1.284	1.682	1.256	1.716	1.227	1.746	1.199	1.785
80	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.228	1.769
95	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768	1.571	1.779

<표 6> Dickey-Fuller t-검정치 분포표

n	Probability of a Smaller Value							
	0.01	0.025	0.05	0.10	0.90	0.95	0.975	0.99
				$\hat{\tau}$				
25	-2.66	-2.26	-1.95	-1.60	0.92	1.33	1.70	2.16
50	-2.62	-2.25	-1.95	-1.61	0.91	1.31	1.66	2.08
100	-2.60	-2.24	-1.95	-1.61	0.90	1.29	1.64	2.03
250	-2.58	-2.23	-1.95	-1.62	0.89	1.28	1.63	2.01
500	-2.58	-2.23	-1.95	-1.62	0.89	1.28	1.62	2.00
∞	-2.58	-2.23	-1.95	-1.62	0.89	1.28	1.62	2.00
				$\hat{\tau}_{\mu}$				
25	-3.75	-3.33	-3.00	-2.63	-0.37	0.00	0.34	0.72
50	-3.58	-3.22	-2.93	-2.60	-0.40	-0.03	0.29	0.66
100	-3.51	-3.17	-2.89	-2.58	-0.42	-0.05	0.26	0.63
250	-3.46	-3.14	-2.88	-2.57	-0.42	-0.06	0.24	0.62
500	-3.44	-3.13	-2.87	-2.57	-0.43	-0.07	0.24	0.61
∞	-3.43	-3.12	-2.86	-2.57	-0.44	-0.07	0.23	0.60
				$\hat{\tau}_{\tau}$				
25	-4.38	-3.95	-3.60	-3.24	-1.14	-0.80	-0.50	-0.15
50	-4.15	-9.80	-3.50	-3.18	-1.19	-0.87	-0.58	-0.24
100	-4.04	-3.73	-3.45	-3.15	-1.22	-0.90	-0.62	-0.28
250	-3.99	-3.69	-3.43	-3.13	-1.23	-0.92	-0.64	-0.31
500	-3.98	-3.68	-3.42	-3.13	-1.24	-0.93	-0.65	-0.32
∞	-3.96	-3.66	-3.41	-3.12	-1.25	-0.94	-0.66	-0.33