R 응용 및 계량경제분석

도서출판 신아문화사

서 문

Praise the Lord.

2014년 12월 1일 제주연구원장으로 취임하게 되어 학교를 휴직하고 3년을 가르치는 일에서 떠나게 되었는데 마지막 강의 시간에 학생들에게 다음과 같은 약속을 하였다.

1993년 제주대학교에 부임하여 계량경제학을 강의하면서 2시간 이론강의 및 1시간 실습을 지켜왔고, 그 동안 RATS, EViews, Gauss 등 다양한 통계패키지 및 언어를 가르쳐 왔는데 지금은 빅데이터의 시대이고 빅데이터 처리에 최적화되어 있는 R 언어를 공부해야 할 때이다. 따라서 본인이 3년의 임기를 마치고 학교로 복귀하여 계량경제학을 다시 가르치게 되는 2018년 1학기에는 R 언어로 실습을 하겠다고 하였다. 이 약속은 시대의 변화에 따른 것이기도 했지만 나를 채찍질하기 위한 나와의 약속이기도 하였다.

‘눈에 보이지 않으면 곧 잊혀진다(out of sight and out of mind)’는 말이 있듯이 3년 동안 연구원 업무에 집중하다 보니 항상 R 언어를 공부해야 한다는 생각은 있었지만 우선순위에서 밀려 공부를 하지 못했다.

2017년 12월 1일 학교로 다시 돌아와 보니 3년 전에 학생들과 했던 약속이 떠올랐고 3월 초에 개강을 앞두고 마음이 급해졌다. 2018년 새해 들어 R 언어를 집중적으로 공부하고 2018년 1학기 강의안을 마련하였다. 다소 부족했지만 2018년 1학기 강의를 마치고, 2학기에 강의안을 업데이트 하였고, 그 결과물이 『R 응용 및 계량경제분석』으로 발간되게 되었다.

본 교재에서 사용된 데이터는 본인의 개인 홈페이지(http://kanggc.iptime.org/data/)에서 다운받을 수 있고, 실습을 위한 코드는 http://kanggc.iptime.org/code/에서 다운받을 수 있다. R 언어의 유용한 점 중의 하나는 다양한 packages가 있어 이를 활용하면 다양한 분석이 가능하다는 것이다. 본 교재에서 활용되는 packages들을 한 번에 install할 수 있도록 install-packages.R을 만들어 놓았으므로 본 교재로 실습하기 전에 먼저 install-packages.R을 실행하기를 바란다.

지난 두 학기 동안 R 언어를 가르치면서 강조했던 단어가 R 언어의 열성적인 지지자를 뜻하는 R Enthusiast이었다. 학생들에게 그들의 열정을 쏟아 부어도 아깝지 않은 그 무엇을 소개하고 싶었는데 계량경제학과 관련해서는 그것이 바로 R 언어라고 몇 번을 강조하였다.

본 교재가 나오기까지 많은 도움이 있었다. 무엇보다도 2018년도 제주대학교 국립대학육성사업의 지원이 있었기에 본 교재가 출간될 수 있었다. 처음 대하는 R 언어를 포기하지 않고 꿋꿋하게 따라와 준 학생들에게도 감사를 드린다. 그리고 항상 기도로 격려를 해 주는 사랑하는 아내와 두 딸 셀라와 셀리에게도 감사의 말을 전한다.

2019년 1월

뉴욕에서 저자

목 차

제 1 장

계량모형의 추정

1. 계량모형 및 추정방법
2. 좋은 추정량
3. 추정방법

제1장 계량모형의 추정

1. 계량모형 및 추정방법

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.

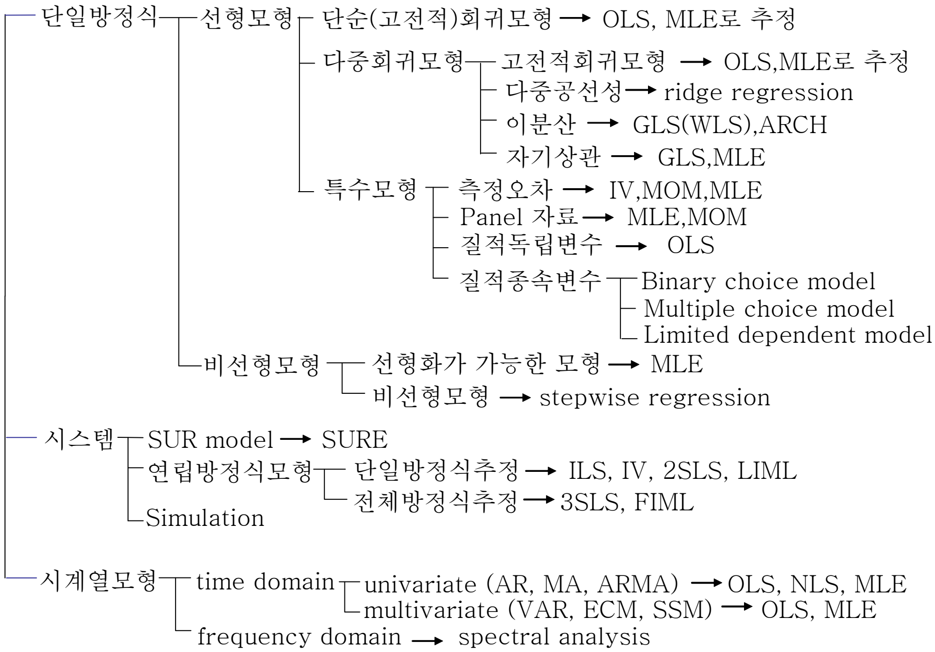
회귀모형을 이용한 회귀분석 중 단일방정식에 의한 회귀분석은 선형모형을 이용한 회귀분석과 비선형모형을 위한 회귀분석으로 나누어지며, 선형모형은 단순회귀모형, 다중회귀모형 및 특수모형을 이용한 회귀분석으로 구분된다. 이 책에서는 단순회귀모형, 다중회귀모형, 특수모형 중 질적 독립변수까지만 다루게 된다.

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 : 3SLS), 전정보최우추정법(Full Information Maximum Likelihood : FIML) 등이 이용된다.

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

이러한 계량모형의 분류 및 추정방법을 요약해 보면 <그림 1-1>과 같다.



<그림 1-1> 계량모형의 분류 및 추정방법

2. 좋은 추정량

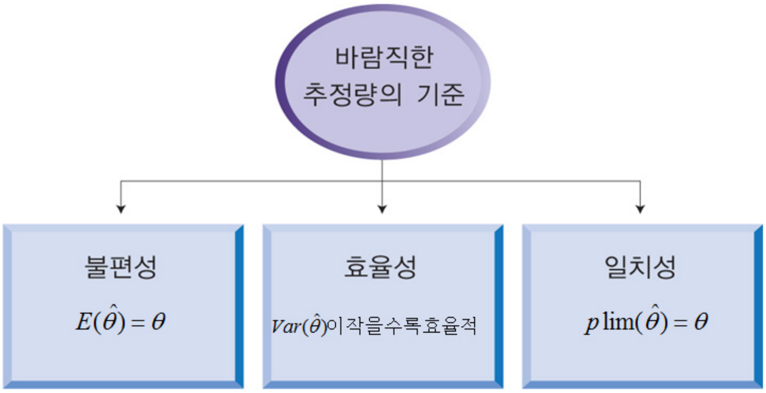
일반적으로 좋은 추정량이 갖추어야 할 특성으로는 <그림 1-2>에 나타나 있는 것과 같이 불변성(unbiasedness), 효율성(efficiency), 일치성(consistency) 등이 있다.

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 $\dfrac{var(\widehat{\theta_1})}{var(\widehat{\theta_2})}$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.

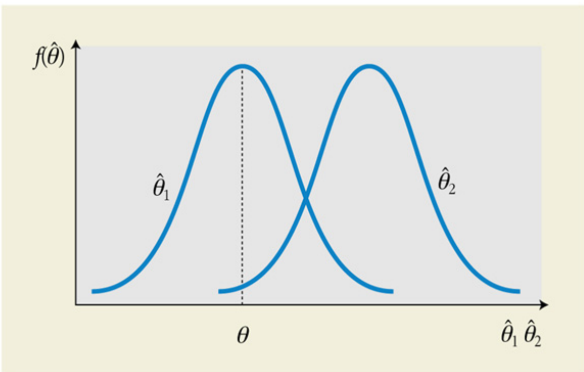


〈그림 1-2〉 추정량의 기준

(1) 불편성

추정량 $\hat{\theta}$의 기댓값이 모수 θ와 일치하면 즉, $E(\hat{\theta})=\theta$ 이면 $\hat{\theta}$은 모수 θ의 불편추정량이라고 한다. 추정량 $\hat{\theta}$의 기댓값과 모수 θ의 차이 즉, $E(\hat{\theta})-\theta$ 를 편의(bias)라고 하므로 편의가 0인 추정량을 불편추정량이라고 하는데 표본평균 $\overline{X}$ 및 표본분산 S^2은 불편추정량이다.

다음의 〈그림 1-3〉에서 $\hat{\theta}_1$은 θ의 불편추정량이지만 $\hat{\theta}_2$은 θ의 불편추정량이 아님을 알 수 있다.



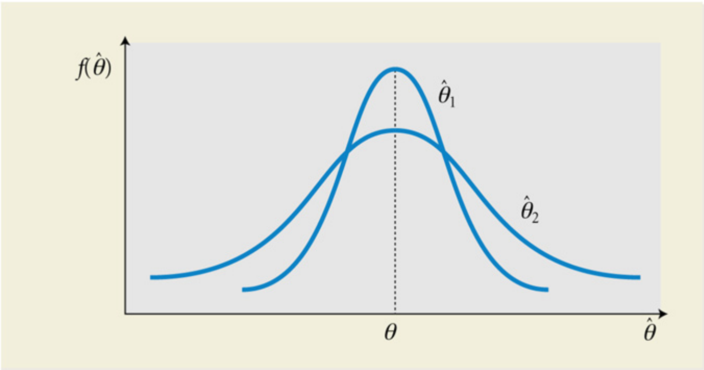
〈그림 1-3〉 추정량의 불편성

(2) 효율성

$\hat{\theta}_1$과 $\hat{\theta}_2$가 모두 불편추정량이라고 하자. 이때 $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ 이면, $\hat{\theta}_1$이 $\hat{\theta}_2$보다 더 효율적이라고 한다.

$\hat{\theta}_1$이 다른 모든 불편추정량보다 작은 분산을 가지면 $\hat{\theta}_1$은 가장 효율적인(most efficient) 혹은 최소분산(minimum variance) 불편추정량이 된다.

다음의 <그림 1-4>에서 $\hat{\theta}_1$과 $\hat{\theta}_2$는 모두 불편추정량이지만 $\hat{\theta}_1$이 $\hat{\theta}_2$보다 효율적인 추정량임을 알 수 있다.



<그림 1-4> 추정량의 효율성

(3) 일치성

표본크기가 무한히 증가할 때 추정량 $\hat{\theta}$이 모수 θ에 근접하는 것을 일치성(consistency)이라고 한다.

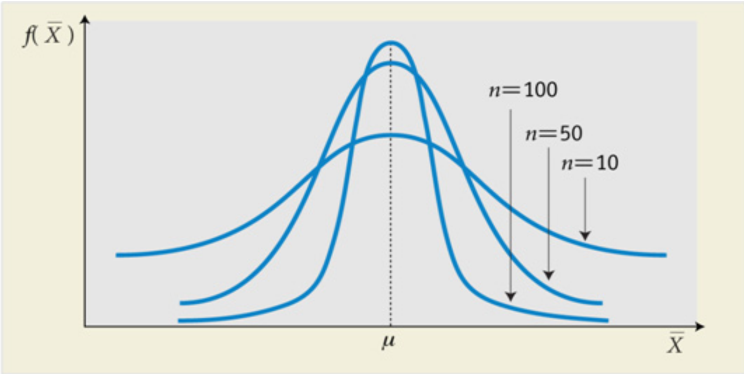
$n \rightarrow \infty$ 일 때 임의의 $\epsilon > 0$ 에 대해 다음이 성립하면 $\hat{\theta}$은 θ의 일치추정량이다.

$$P(\mid \hat{\theta} - \theta \mid \leq \epsilon) \rightarrow 1$$

이를 간단히 $plim\hat{\theta} = \theta$ 이라고 나타내며, θ를 $\hat{\theta}$의 확률극한(probability limit)이라고 한다.

다음의 <그림 1-5>에서 표본의 크기가 커짐에 따라 추정량의 분산이 작아짐으로

써 추정량 $\hat{\theta}$이 모수 θ에 근접함을 알 수 있다.



<그림 1-5> 추정량의 일치성

한편, 모집단에서 무수히 많은 표본을 추출하여 각 표본 추정량의 값을 계산했을 때 좋은 추정량이 되기 위해서는 추정값들의 확률분포가 모수를 중심으로 밀집되어야 할 것이다. 이 밀집성의 정도는 평균제곱오차(Mean Squared Error, MSE)로 측정할 수 있는데 평균제곱오차는 분산과 편의 제곱의 합이다.

$$MAE(\theta) = E[(\hat{\theta}) - \theta)^2] = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta)]^2$$

3. 추정방법

(1) 보통최소자승법

단순회귀모형에서 회귀계수의 추정방법은 보통최소자승법(Ordinary Least Squares: OLS), 최우추정법(Maximum Likelihood Estimation: MLE) 및 최소절대편차(Least Absolute Deviations: LAD) 추정법 등이 있다.

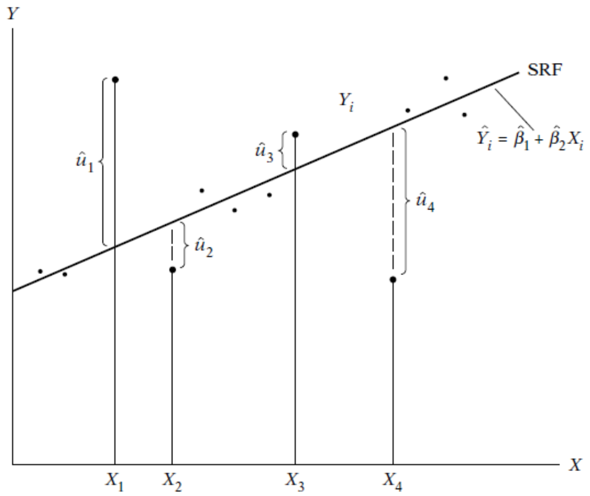
<그림 1-6>에서 보듯이 실제 관측치를 가장 잘 나타내도록(또는 추적하도록) 회귀선을 추정해야 하는데 이 때 실제 관측치와 회귀모형에 의한 추정치의 차이를 나타내는 잔차(residual)가 큰 역할을 한다.

실제 관측치를 잘 나타내는 표본회귀식을 구하기 위해서는 잔차의 합이 최소가 되도록 하는 것이 바람직하다. 그러나 잔차의 합이 0이 되는 식은 유일하지 않으므로 잔차의 제곱의 합이 최소가 되게 하는 회귀식을 구할 수 있는데 이러한 추정방법을 보통최소자승법이라고 한다.

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

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

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



<그림 1-6> 보통최소자승법

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

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$$

$\sum_{1=1}^{n} e_i^2$은 최소화해야 할 목적함수이며 선택변수가 $\widehat{\beta_0}$, $\widehat{\beta_1}$이므로 이 식을 선택함수에 대해 각각 1차 미분하면 다음의 두 식이 도출된다.

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \widehat{\beta_0}} = \frac{\partial \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2}{\partial \widehat{\beta_0}} = 2\sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)(-1) = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \widehat{\beta_1}} = \frac{\partial \sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2}{\partial \widehat{\beta_1}} = 2\sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)(-X_i) = 0$$

위 두 식으로부터 다음의 두 식을 구할 수 있는데 이 두 식을 정규방정식(normal equation)이라고 한다.

$$\sum_{i=1}^{n} Y_i = n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} X_i^2$$

정규방정식을 연립으로 풀면 다음과 같이 $\widehat{\beta_0}$, $\widehat{\beta_1}$을 구할 수 있다.

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X} \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X} \sum_{i=1}^{n} X_i} = \frac{cov(X, Y)}{var(X)}$$

(2) 최우추정법

보통최소자승법 외에 회귀계수를 추정하는 또 다른 방법으로는 최우추정법이 있다. 보통최소자승법을 이용한 회귀계수의 추정에서는 교란항의 분포에 대한 가정은 필요 없었는데 최우추정법을 사용하기 위해서는 교란항에 대한 가정이 필요하며 일

반적으로 정규분포를 가정한다.

단순회귀모형에서 $Y_i \sim NI(\beta_0 + \beta_1 X_i, \sigma^2)$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.

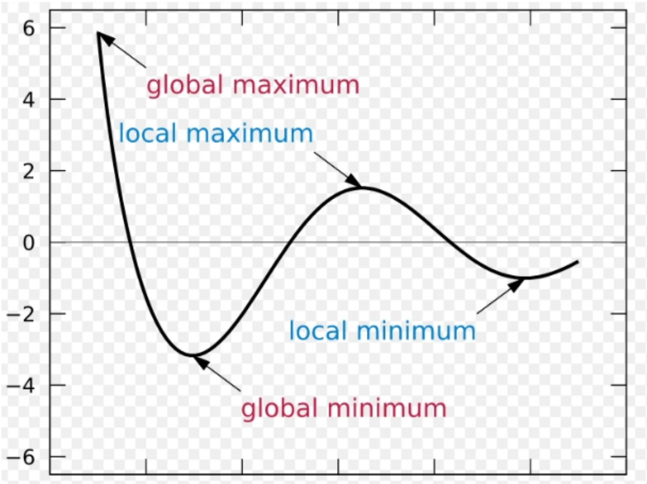
따라서 $Y_1, Y_2, ..., Y_n$ 의 결합확률밀도함수는 다음과 같이 나타낼 수 있다.

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

위 식에서 $Y_1, Y_2, ..., Y_n$ 이 알려져 있고, $\beta_0, \beta_1, \sigma^2$ 이 알려져 있지 않은 경우 이를 우도함수(likelihood function)라고 하고 $L(\beta_0, \beta_1, \sigma^2)$로 표기하므로 우도함수는 다음과 같다.

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

최우추정법이란 주어진 Y의 값들을 관측할 확률을 최대로 하는 모수 $(\beta_0, \beta_1, \sigma^2)$을 추정하는 방법이다. 즉, 우도함수의 최댓값을 찾는 방법이다. 최댓값에는 <그림 1-7>과 같이 전역적 최댓값(global maximum) 및 국소적 최댓값(local maximum)이 있기 때문에 전역적 최댓값을 빠르게 찾기 위해서는 초기 값(initial value)을 정하는 것이 중요하다. 최우추정법에서는 보통최소자승법에 의한 추정치를 초기 값으로 설정하는 것이 일반적이다.



<그림 1-7> 최댓값 및 최솟값

우도함수의 미분 값을 간단히 구하기 위해 동 식을 로그형태로 바꾸면 다음 식과 같게 된다.

$$lnL = -\frac{n}{2}ln\sigma^2 - \frac{1}{2}ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i-\beta_0-\beta_1X_i)^2}{\sigma^2}]$$

위 식을 $\beta_0, \beta_1, \sigma^2$에 대해 편미분하면 다음의 세 식을 얻게 된다.

$$\frac{\partial lnL}{\partial\beta_0} = -\frac{1}{\sigma^2}\sum(Y_i-\beta_0-\beta_1X_i)(-1)$$

$$\frac{\partial lnL}{\partial\beta_1} = -\frac{1}{\sigma^2}\sum(Y_i-\beta_0-\beta_1X_i)(-X_1)$$

$$\frac{\partial lnL}{\partial\sigma^2} = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum(Y_i-\beta_0-\beta_1X_i)^2$$

위 세 식의 최우추정량을 $\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\sigma^2}$로 나타내면 다음의 세 식과 같다.

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

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

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

위 세 식 중 처음 두 식으로부터 다음의 두 식을 얻게 되는데 이는 최소자승법에서의 정규방정식과 동일하다.

$$\sum_{i=1}^{n} Y_i = n\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \widehat{\beta_0}\sum_{i=1}^{n} X_i + \widehat{\beta_1}\sum_{i=1}^{n} X_i^2$$

한편, 위 세 식 중 세 번째 식으로부터 다음과 같은 분산에 대한 추정량을 얻을 수 있다. 최소자승법으로 구한 분산에 대한 추정량은 불편추정량이 되는 반면에 최우추정법에서 구한 분산에 대한 추정량은 점근적 불편추정량이 된다.

$$\begin{aligned}\widehat{\sigma^2} &= \frac{1}{n}\sum(Y_i - \widehat{\beta_0} - \widehat{\beta_1}X_i)^2 \\ &= \frac{1}{n}\sum e_i^2\end{aligned}$$

최우추정량은 다음과 같은 특징을 갖는다.

첫째, 회귀계수에 대한 최우추정량은 최소자승법에 의한 추정량과 동일하다. 회귀계수에 대한 최우추정량을 계산할 수 있는 두 식이 보통최소자승법에서의 정규방정식과 동일하므로 회귀계수에 대한 최우추정량과 OLS 추정량은 동일하게 된다.

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

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

(3) 최소절대편차 추정법

보통최소자승법과 유사하지만 다른 방법으로는 최소절대편차 추정법이 있다. 보통 최소자승법은 잔차의 제곱의 합인 $\sum_{1=1}^{n} e_i^2$이 최소가 되도록 하는 회귀계수(β_0, β_1)를 구하는 방법이다. 반면에 최소절대편차 추정법은 다음과 같이 잔차의 절댓값의 합(sum of absolute errors)을 최소화하는 (β_0, β_1)를 구하는 방법이다.

$$\sum_{i=1}^{n} \mid e_i \mid = \sum_{i=1}^{n} \mid Y_i - \widehat{Y}_i \mid = \sum_{i=1}^{n} \mid Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i \mid$$

보통최소자승법에서는 정규방정식의 해를 구하면 그것이 회귀계수를 추정하는 추정량이 되므로 자료만 있으면 추정량에 의해 추정치를 구할 수 있다. 즉, 회귀계수를 구하는 해석적인 해(analytical solution)가 존재한다. 그러나 최소절대편차 추정법은 해석적인 해를 가지고 있지 않기 때문에 해를 효율적으로 구하기가 쉽지 않다.

최소절대편차 추정법에서 해를 구하기 위해서는 선형계획(linear programming)과 같은 반복적 접근(iterative approach)이 필요하다. 따라서 다음과 가은 선형계획에서 이용하는 심플렉스 방법(simplex-based method)이 최소절대편차 추정량을 구하는데 이용된다.

$$\begin{array}{l} Min \quad \sum_{i=1}^{n} \mid \beta_0 - \beta_1 X_i - Y_i \mid \\ (\beta_0, \beta_1) \\ s.t \quad \beta_0 + \beta_1 X_i - Y_i \leq k \end{array}$$

최소절대편차 추정법은 다음과 같은 특징이 있다.

첫째, 추정결과가 보통최소자승법에 의한 추정결과보다 강건(robust)하다. 즉, 보통최소자승법은 잔차를 제곱하므로 이상치(outliers)에 영향을 많이 받는 반면에 최소절대편차 추정법은 잔차의 절댓값을 이용하므로 이상치에 영향을 덜 받게 된다. 이것은 보통최소자승법은 양이든 음이든 큰 잔차에는 큰 가중치를 부여하는 반면에 최소절대편차 추정법은 양이든 음이든 크기다 동일한 잔차에는 동일한 가중치를 부여한다는 의미이다, 따라서 이상치에 동일한 가중치를 부여하는 연구에는 최소절대편차 추정법이 활용될 수 있다.

둘째, 추정결과가 보통최소자승법에 의한 추정결과보다 불안정하다. 즉, 보통최소자승법은 유일한 해를 제공해 주지만 최소절대편차 추정법에서는 복수 해가 있을 수 있어 불안정한 추정결과를 줄 수 있다. 그 이유는 최소절대편차 추정법은 최소한 2개의 데이터 포인트를 반드시 통과하는 회귀식을 추정하기 때문이다

셋째, 교란항에 대한 분포가 이중지수분포(double exponential distribution)에 따르면, 최소절대편차 추정량과 최우추정법 추정량은 같게 된다.

4. 세 가지 추정방법의 비교

$Y_i = \beta_0 + \beta_1 X_i + u_i$ 의 모형에서 $\beta_0 = 1$, $\beta_1 = 0.5$ 의 값을 주고, X의 경우 평균이 4이고, 표준편차가 1인 정규분포를 가정하고, 교란항의 경우 평균이 0이고 표준편차가 2인 정규분표를 가정하여 표본의 크기가 500인 데이터를 생성하였다. 생성된 데이터로 단순회귀모형을 OLS, LAD, MLE로 추정하는 임의실험을 1000회 한 후 추정방법별로 추정된 회귀계수의 평균과 분산을 각각 계산하니 다음과 같았다.

첫째, 세 가지 추정방법 모두 회귀계수의 추정치의 평균은 참 값과 유사하게 나타나고 있으며, 분산의 크기는 세 가지 방법이 다르게 나왔다.

구분	OLS		LAD		MLE	
	평균	분산	평균	분산	평균	분산
$\hat{\beta}_0$	1.006349	0.146264	1.011647	0.223051	1.001818	0.143442
$\hat{\beta}_1$	0.4978783	0.008670	0.497052	0.013132	0.500369	0.008318

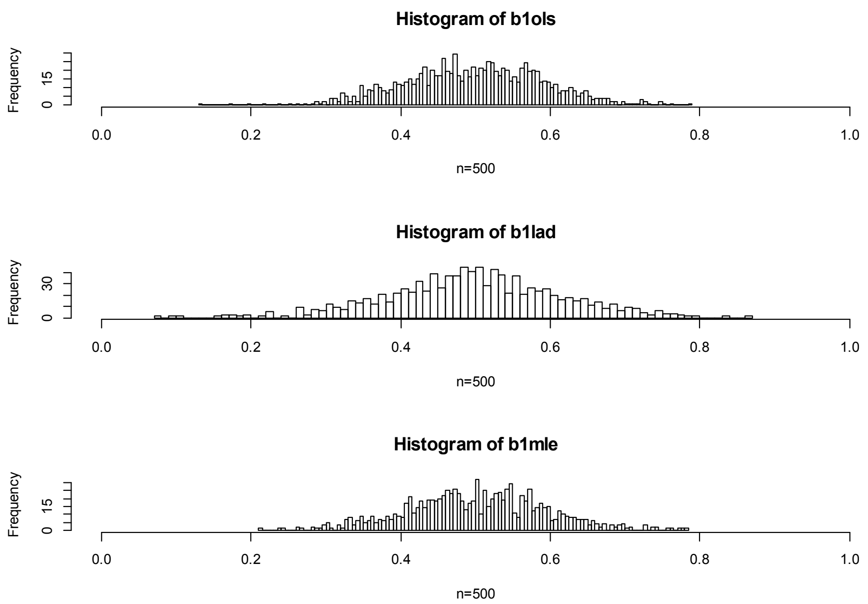
b2-ch1-1.R의 실행결과

```
(중략)

> hist(b1mle, breaks=100, xlim=c(0,1),xlab="n=500")

> mean(b0ols)
[1] 1.006349

> mean(b1ols)
[1] 0.4978783

> var(b0ols)
[1] 0.1462646

> var(b1ols)
[1] 0.008670201

> mean(b0lad)
[1] 1.011647

> mean(b1lad)
[1] 0.4970523

> var(b0lad)
[1] 0.2230516

> var(b1lad)
[1] 0.0131329

> mean(b0mle)
[1] 1.001818

> mean(b1mle)
[1] 0.5003692

> var(b0mle)
[1] 0.1434428

> var(b1mle)
[1] 0.008318494
```

둘째, 표본의 크기에 따라 β_1 회귀계수 추정치의 평균 및 분산을 살펴보면 세 가지 추정방법 모두 표본이 커짐에 따라 분산이 작아지고, 이에 따라 평균이 참 값인 0.5에 점근적으로 수렴함을 확인할 수 있어 세 가지 추정량 모두 일치추전량임을 알 수 있다.

셋째, 불편성과 분산의 크기를 동시에 고려한 평균제곱오차의 크기를 살펴보면, OLS 및 MLE가 LAD보다 작은 것으로 나타났고, OLS와 MLE는 거의 비슷한 것으로 나타났다.

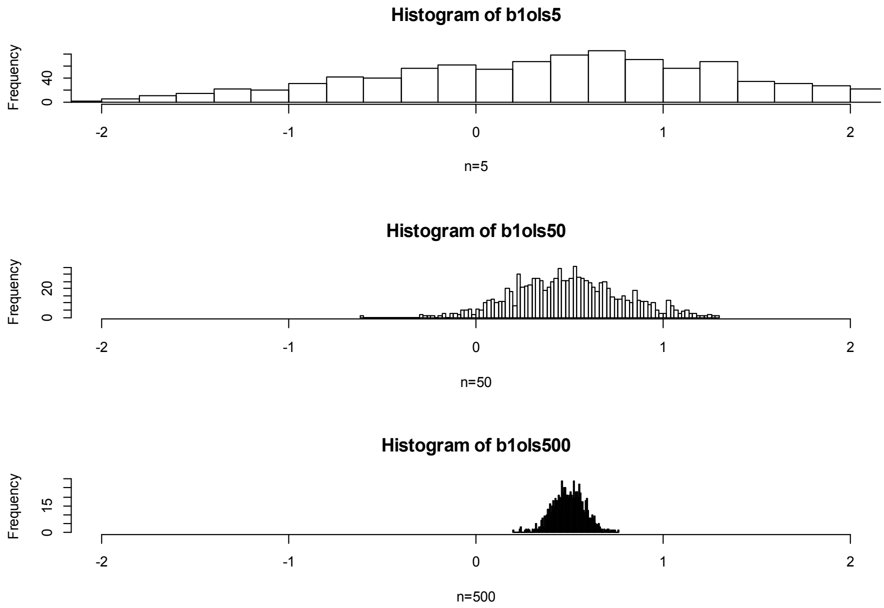
$\beta_1 = 0.5$	n=5	n=50	n=500
OLS(평균)	0.494661	0.49546	0.496774
OLS(분산)	2.051899	0.083464	0.007424
LAD(평균)	0.460898	0.475697	0.50153
LAD(분산)	2.848483	0.134219	0.012328
MLE(평균)	0.488962	0.498777	0.502674
MLE(분산)	1.795466	0.085551	0.008141
MSE(OLS)	2.051928	0.083485	0.007435
MSE(LAD)	2.850012	0.13481	0.012331
MSE(MLE)	1.795588	0.085552	0.008148

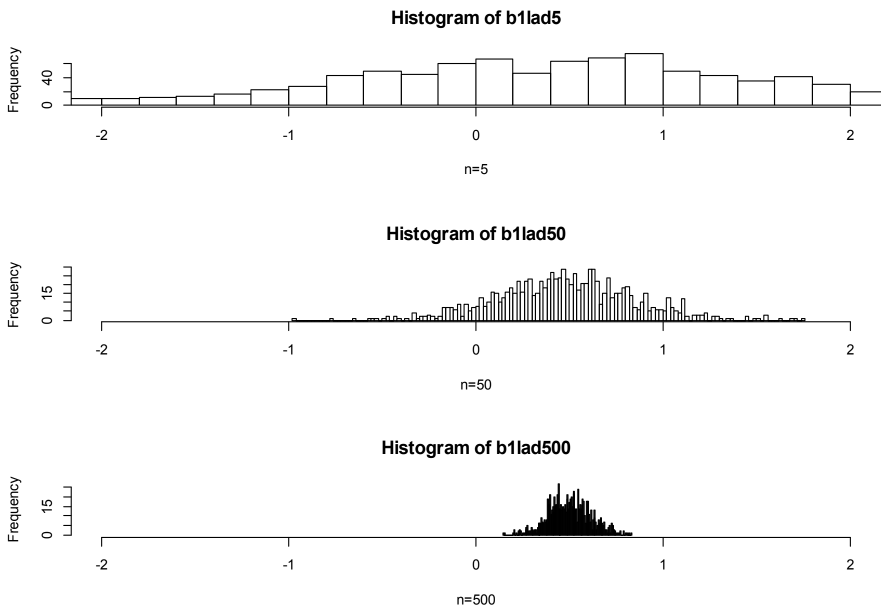


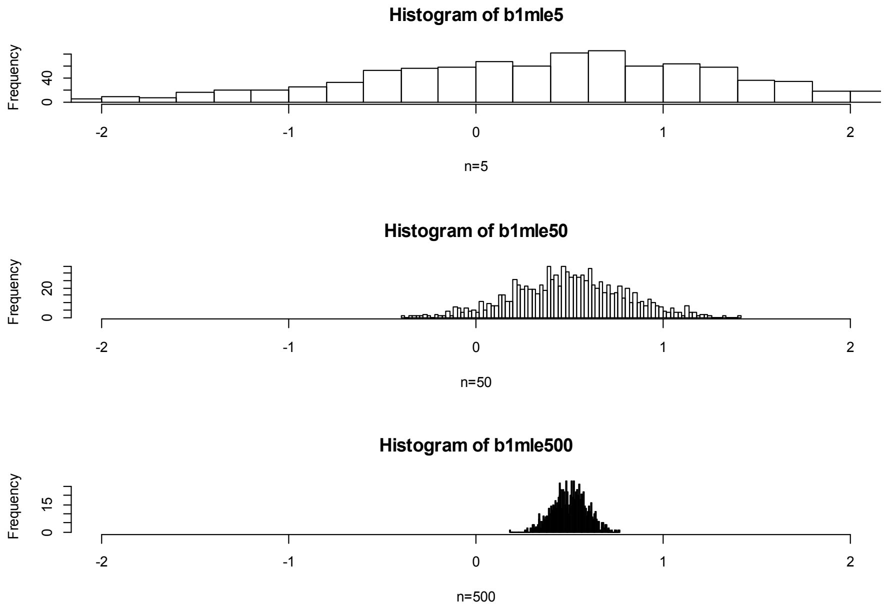
b2-ch1-2.R의 실행결과
(중략) > hist(b1mle5, breaks=100, xlim=c(-2,2),xlab="n=5") > hist(b1mle50, breaks=100, xlim=c(-2,2),xlab="n=50") > hist(b1mle500, breaks=100, xlim=c(-2,2),xlab="n=500") > mean(b0ols5) [1] 1.06303 > mean(b1ols5) [1] 0.494661 > mean(b0ols50) [1] 1.017586 > mean(b1ols50) [1] 0.4954604 > mean(b0ols500) [1] 1.011585 > mean(b1ols500) [1] 0.4967737 > var(b0ols5) [1] 34.67107 > var(b1ols5) [1] 2.051899 > var(b0ols50) [1] 1.439099 > var(b1ols50) [1] 0.08346432

```
> var(b0ols500)
[1] 0.1272816

> var(b1ols500)
[1] 0.007424318

> mean(b0lad5)
[1] 1.197464

> mean(b1lad5)
[1] 0.4608975

> mean(b0lad50)
[1] 1.108627

> mean(b1lad50)
[1] 0.4756969

> mean(b0lad500)
[1] 0.9900516

> mean(b1lad500)
[1] 0.5015301

> var(b0lad5)
[1] 45.6021

> var(b1lad5)
[1] 2.848483

> var(b0lad50)
[1] 2.351205

> var(b1lad50)
[1] 0.1342193

> var(b0lad500)
[1] 0.2074501
```

```
> var(b1lad500)
[1] 0.01232843

> mean(b0mle5)
[1] 1.064074

> mean(b1mle5)
[1] 0.4889621

> mean(b0mle50)
[1] 1.000705

> mean(b1mle50)
[1] 0.4987766

> mean(b0mle500)
[1] 0.991524

> mean(b1mle500)
[1] 0.5026738

> var(b0mle5)
[1] 30.548

> var(b1mle5)
[1] 1.795466

> var(b0mle50)
[1] 1.438708

> var(b1mle50)
[1] 0.08555079

> var(b0mle500)
[1] 0.1371804

> var(b1mle500)
[1] 0.008141261
```


Histogram of b1ols5
Frequency
0
40
-2
-1
0
1
2
n=5
Histogram of b1ols50
Frequency
0
20
-2
-1
0
1
2
n=50
Histogram of b1ols500
Frequency
0
15
-2
-1
0
1
2
n=500


Histogram of b1lad5
Frequency
0
40
-2
-1
0
1
2
n=5
Histogram of b1lad50
Frequency
0
15
-2
-1
0
1
2
n=50
Histogram of b1lad500
Frequency
0
15
-2
-1
0
1
2
n=500


Histogram of b1mle5
Frequency
0
40
-2
-1
0
1
2
n=5
Histogram of b1mle50
Frequency
0
20
-2
-1
0
1
2
n=50
Histogram of b1mle500
Frequency
0
15
-2
-1
0
1
2
n=500

제 2 장

단순회귀분석

제2장 단순회귀분석

1. 단순회귀모형

2-변수 모집단 회귀함수(PRF)를 단순회귀모형(simple regression model)이라고도 하는데 다음과 같이 나나낼 수 있다.

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

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

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

첫째, 독립변수 X는 비확률변수이다. 이 가정은 표본을 반복해서 추출할 때 X는 고정된 것으로 해 놓고 Y를 추출한다는 것을 의미하므로 X는 확정변수이며 X는 오차 없이 측정된다는 의미이다.

둘째, 교란항의 평균은 0이다 즉, $E(u_i) = 0 (i = 1,2,3,..,n)$이다. <그림 1-1>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 명시적으로 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.

셋째, 교란항은 모든 X_i에 대한 동일한 분산을 갖는다. 즉, $E(u_i^2) = \sigma_u^2$이다. 이 가

정은 각 독립변수 X의 값에 있어서 종속변수 Y가 X의 평균을 중심으로 분포되어 있는 정도가 같다는 것을 의미하는데 이를 동분산(homoscedasticity)이라고 한다.

넷째, 모든 u값은 서로 독립이다 즉, $Cov(u_i, u_j) = E(u_i, u_j) = 0, (i \neq j)$이다. 이 가정은 교란항에 자기상관(autocorrelation)이 없다는 것인데 이는 교란항의 변화에 체계적인 형태가 존재하지 않는다는 것을 의미한다.

다섯째, u와 X는 서로 독립이다 즉, $Cov(u_i, X_i) = E(u_i, X_i) = 0$이다. 이 가정은 교란항과 설명변수가 상관관계가 없다는 것을 의미하는데 이를 직교조건(orthogonal 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.

2. 회귀계수의 추정

OLS에 의한 회귀계수의 추정량은 각각 다음과 같음을 앞에서 설명하였다.

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{cov(X, Y)}{var(X)}$$

이제 다음의 (예제 2-1)을 이용하여 회귀계수의 추정치를 계산해 보자.

(예제 2-1)
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.

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

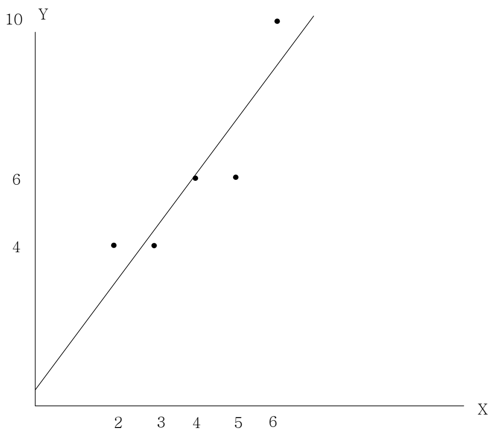
위에서 계산한 값들을 OLS 추정량에 대입하면 다음과 같은 회귀계수를 구할 수 있다.

$$\therefore\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{134 - 4\times 30}{90 - 4\times 20} = \frac{14}{10} = 1.4$$

$$\therefore\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 6 - (1.4)(4) = 0.4$$

따라서 추정된 회귀식은 다음과 같고 <그림 2-1>은 이를 나타내 주고 있다.

$$\hat{Y}_i = 0.4 + 1.4X_i$$



<그림 2-1> (예제 2-1)의 추정회귀선

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 한 단위 증가함에 따라 종속변수 Y 값은 평균적으로 1.4 단위 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.

3. 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 추정해야 한다.

교란항의 분산 σ_u^2에 대한 불편추정량은 $\sum e_i^2$을 자유도인 n-2로 나눈 것 즉, $\widehat{\sigma_u^2} = \dfrac{\sum e_i^2}{n-2}$이다. 여기서 2는 단순회귀모형에서 상수항을 포함한 독립변수의 수이다.

교란항의 분산이 추정되면 단순회귀모형의 추정에서 마지막으로 해야 할 일은 회귀계수의 분산을 추정하는 것인데 각각 다음과 같다.

$$Var(\widehat{\beta_1}) = \sigma_u^2 \frac{1}{\sum x_i^2}$$

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

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

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

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

(예제 2-1)(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 이용한 기초계산은 다음과 같다.

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

$$\sum(X_i - \overline{X})^2 = \sum X_i^2 - \overline{X}\sum X_i = 90 - (4)(20) = 10$$

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

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

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

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

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

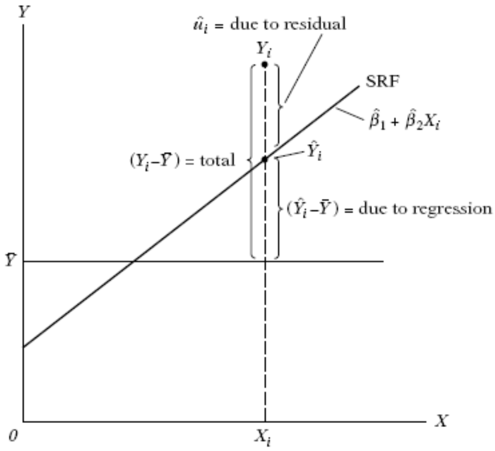
$$var(\widehat{\beta_1}) = \widehat{\sigma_u^2}\frac{1}{\sum(X_i - \overline{X})^2} = (1.4667)\frac{1}{10} = 0.14667$$

4. 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 2-2>에서 $\sum(Y_i - \overline{Y})^2$을 총자승합(total sum of squares) 또는 전변동(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 한다.



<그림 2-2> Y의 변동

전변동은 회귀선의 추정으로 설명된 변동(explained variation)과 설명되지 않은 변동(unexplained variation)의 합 또는 모형에 의한 부분과 오차에 의한 부분의 합으로 분해될 수 있다. 따라서 잔차변동은 전변동에서 회귀변동을 뺀 것으로 다음과 같이 나타낼 수 있다.

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

결정계수는 전변동에 대한 회귀변동의 비율이라는 정의에 따라 다음과 같이 나타낼 수 있다.

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

앞의 (예제 2-1)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.

5. 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 것인지에 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 경제

에 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 종속변수를 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.

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

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

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

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

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

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 : $\widehat{\beta_0}\pm t_{(n-2,\frac{\alpha}{2})}se(\widehat{\beta_0})$

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

(예제 2-1)(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.

$\widehat{Y} = 0.4 \ + \ 1.4X$

$\quad = (1.62)\ (0.38)$

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 $\widehat{\beta_1}$는 통계적으로 유의하다(statistically significant)라고 한다.

한편, β_1에 대한 95% 신뢰구간은 다음과 같이 계산된다.

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

이와 같이 계산된 신뢰구간의 의미는 여러 표본을 이용하여 β_1에 대한 추정을 계속하여 반복하면 β_1의 추정치 즉, $\widehat{\beta_1}$들 중 95%가 0.181과 2.618 사이에 놓이게 된다는 것이다. $\beta_1 = 0$ 이라는 귀무가설의 값이 β_1에 대한 95% 신뢰구간에 포함되지 않으므로 5% 유의수준에서 귀무가설을 기각한다.

6. 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($E(Y \mid X_0)$)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 예측구간을 구할 수 있다.

(1) 점예측

점예측에 대해 살펴보면 $X_i = X_0$일 때 $\widehat{Y_0} = \widehat{\beta_0} + \widehat{\beta_1} X_0$ 은 평균예측 $E(Y \mid X_0)$ 및 개별예측 Y_0의 점예측치가 된다.

(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.

개별예측의 예측오차 분산과 평균예측의 예측오차 분산은 각각 다음과 같다.

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

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

예측오차 분산을 구하는 위의 식으로부터 몇 가지 사실을 관찰할 수 있다.

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

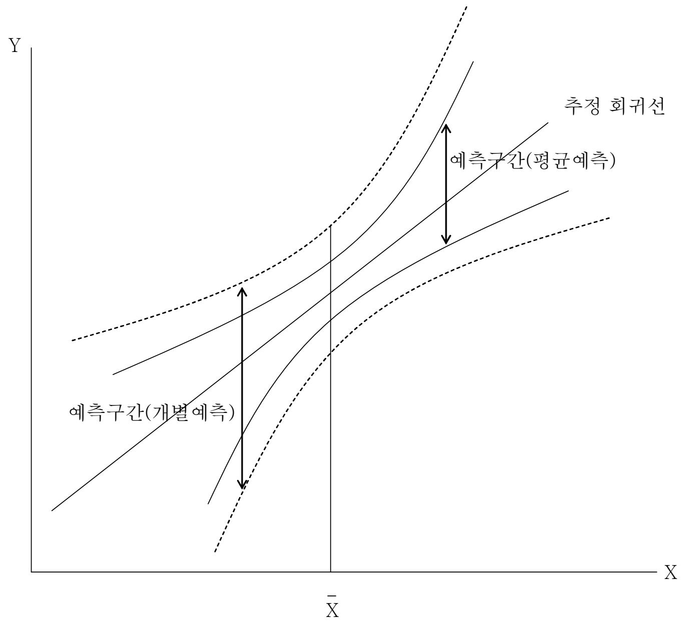
둘째, 교란항의 분산(σ_u^2)이 커질수록 예측오차의 분산이 커진다. 즉, 원래 회귀모형에서 불확실성이 커서 교란항의 분산이 크면 예측이 어려울 수밖에 없다는 것이다.

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

예측오차의 분산이 구해지면 예측구간을 구할 수 있는데 $(1-\alpha)\times$ 100% 예측구간은 다음과 같다. 개별예측이든 평균예측이든 점예측치는 $\widehat{Y_0}$로 동일하며 예측구간은 예측오차 분산의 크기에 의해 결정된다는 것을 알 수 있다.

$$\widehat{\beta_0}+\widehat{\beta_1}X_0 \pm t_{(n-2,\frac{\alpha}{2})}\sigma_\epsilon$$

<그림 2-3>에서 나타내고 있듯이 개별예측에 대한 예측구간이 평균예측에 대한 예측구간보다 더 넓다. 이는 표본회귀함수를 이용한 개별예측의 예측력은 평균예측의 예측력보다 떨어진다는 것을 의미한다.



<그림 2-3> 예측구간

(예제 2-1)(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 $\hat{Y}= 0.4 + 1.4X$에서 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 구하기 위해 예측오차의 분산을 구해보자. 앞의 계산에서 교란항의 분산 $\widehat{\sigma_u^2}$은 1.4667, 독립변수의 평균 $\overline{X}$는 4, 표본의 크기 n은 5, $\sum x_i^2 = 10$이므로 개별예측의 예측오차 분산은 3.08이 된다.

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

개별예측에 대한 95% 예측구간을 구해보자. 점예측치가 10.2이고 예측오차의 분산이 3.08, $t_{3,0.025} = 3.182$이므로 개별예측의 95% 예측구간을 다음과 같이 구할 수 있다.

$$\widehat{\beta_0}+\widehat{\beta_1}X_0 \pm t_{(n-2,\frac{\alpha}{2})}\sigma_\epsilon = 10.2 \pm (3.182)\sqrt{3.08} = [4.61, 15.78]$$

한편, 평균예측에 대한 예측오차의 분산과 평균예측에 대한 95% 예측구간은 각각 다음과 같이 계산된다.

$$\sigma_u^2(\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\sum x_i^2}) = (1.4667)(\frac{1}{5}+\frac{(7-4)^2}{10}) = 1.613$$

$$\widehat{\beta_0} + \widehat{\beta_1} X_0 \pm t_{(n-2,\frac{\alpha}{2})}\sigma_\epsilon = 10.2 \pm (3.182)\sqrt{1.613} = [6.158, 14.242]$$

지금까지 설명한 내용을 b2-ch2-1.R의 실행결과를 통해 확인할 수 있다.

첫째, 추정된 회귀계수 $\widehat{\beta_1} = 1.4$, $\widehat{\beta_0} = 0.4$ 이다.

둘째, 추정된 교란항의 분산 $\widehat{\sigma_u^2} = 1.46667$ 이다.

셋째, 추정된 회귀계수 $\widehat{\beta_1}$의 분산은 0.146667이다.

넷째, 결정계수 $R^2 = 0.8166667$ 이다.

다섯째, 귀무가설 $H_0 : \beta_1 = 0$ 라는 귀무가설을 검정하기 위해 계산된 검정통계량의 값은 3.655631로써 5% 유의수준에서 자유도가 3인 t-분포의 임계치는 3.182446보다 크므로 5% 유의수준에서 귀무가설을 기각한다.

여섯째, β_1에 대한 95% 신뢰구간은 [0.1812159, 2.618784]이고, 귀무가설의 값 0이 이 신뢰구간에 포함되지 않으므로 귀무가설을 기각함을 확인할 수 있다.

일곱째, X=7일 때, 개별예측치 및 평균예측치는 모두 10.2이다. 한편, 개별예측의 예측오차분산은 3.08이고 95% 예측구간은 [4.61, 15.78]이며, 평균예측의 예측오차분산은 1.613이고 95% 예측구간은 [6.158, 14.242]이다.

b2-ch2-1.R의 실행결과

```
> x<-c(2,3,4,5,6)

> y<-c(4,4,6,6,10)

> x;y
[1] 2 3 4 5 6
[1]  4  4  6  6 10

> (n<-length(x))
[1] 5

> (sumx<-sum(x))
[1] 20
```

```
> (sumy<-sum(y))
[1] 30

> (mx=mean(x))
[1] 4

> (my=mean(y))
[1] 6

> (xy<-x*y)
[1]  8 12 24 30 60

> (sumxy<-sum(xy))
[1] 134

> (sumxsq<-sum(x^2))
[1] 90

> (sumysq<-sum(y^2))
[1] 204

> beta1<-(sumxy-mx*sumy)/(sumxsq-mx*sumx)

> beta0<-my-beta1*mx

> beta0;beta1
[1] 0.4
[1] 1.4

> (b1hat<-cov(x,y)/var(x))
[1] 1.4

> (b0hat<-mean(y)-b1hat*mean(x))
[1] 0.4

> (dx<-x-mx)
[1] -2 -1  0  1  2
```

```
> (dy<-y-my)
[1] -2 -2  0  0  4

> (sumdxsq<-sum(dx^2))
[1] 10

> (sumdysq<-sum(dy^2))
[1] 24

> (sumdxdy<-sum(dx*dy))
[1] 14

> (ssr<-sumdysq-beta1*sumdxdy)
[1] 4.4

> (sigusq<-ssr/3)
[1] 1.466667

> (vbeta1<-sigusq/sumdxsq)
[1] 0.1466667

> (rsq<-(beta1^2*sumdxsq)/sumdysq)
[1] 0.8166667

> (t<-beta1/sqrt(vbeta1))
[1] 3.655631

> (pt(t,3))
[1] 0.9823236

> (tc<-qt(p=0.025, df=3, lower.tail=F))
[1] 3.182446

> (b1hat_lb<-b1hat-(tc)*sqrt(vbeta1))
[1] 0.1812159
```

```
> (b1hat_ub<-b1hat+(tc)*sqrt(vbeta1))
[1] 2.618784

> x0<-7

> (yhat<-beta0+beta1*x0)
[1] 10.2

> (sigesq_ind<-sigusq*(1+(1/n)+((x0-mx)^2/sumdxsq)))
[1] 3.08

> (sige_ind<-sqrt(sigesq_ind))
[1] 1.754993

> (yhat_ind_lb<-(yhat-(-tc)*sige_ind))
[1] 15.78517

> (yhat_ind_ub<-(yhat+(-tc)*sige_ind))
[1] 4.614829

> (sigesq_mean<-sigusq*((1/n)+((x0-mx)^2/sumdxsq)))
[1] 1.613333

> (sige_mean<-sqrt(sigesq_mean))
[1] 1.270171

> (yhat_mean_lb<-(yhat-(tc)*sige_mean))
[1] 6.15775

> (yhat_mean_ub<-(yhat+(tc)*sige_mean))
[1] 14.24225
```

한편, 선형모형의 회귀분석을 실행하는 lm 함수를 이용한 b2-ch2-2.R을 실행해 보면 위와 동일한 결과를 얻을 수 있다.

첫째, 회귀계수 추정치, 표준오차, 귀무가설에 대한 계산된 t-검정통계량의 값, 유의확률 및 유의수준을 보여주고 있다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.400      1.625   0.246   0.8214
x              1.400      0.383   3.656   0.0354 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

둘째, 교란항의 표준오차는 1.211, 결정계수는 0.8167임을 보여주고 있다.

```
Residual standard error: 1.211 on 3 degrees of freedom
Multiple R-squared:  0.8167,     Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

셋째, 추정 회귀계수에 대한 95% 신뢰구간을 보여주고 있는데, β_1에 대한 95% 신뢰구간은 [0.1812159, 2.618784]임을 확인할 수 있다.

```
                 2.5 %   97.5 %
(Intercept) -4.7708632 5.570863
x            0.1812159 2.618784
```

넷째, 점예측치는 10,2이고, 개별예측의 95% 예측구간은 [4.614829, 15.78517]임을 보여주고 있다.

```
   fit      lwr      upr
1 10.2 4.614829 15.78517
```

다섯째, 점예측치는 10,2이고, 평균예측의 95% 예측구간은 [6.15775, 14.24225]임을 보여주고 있다.

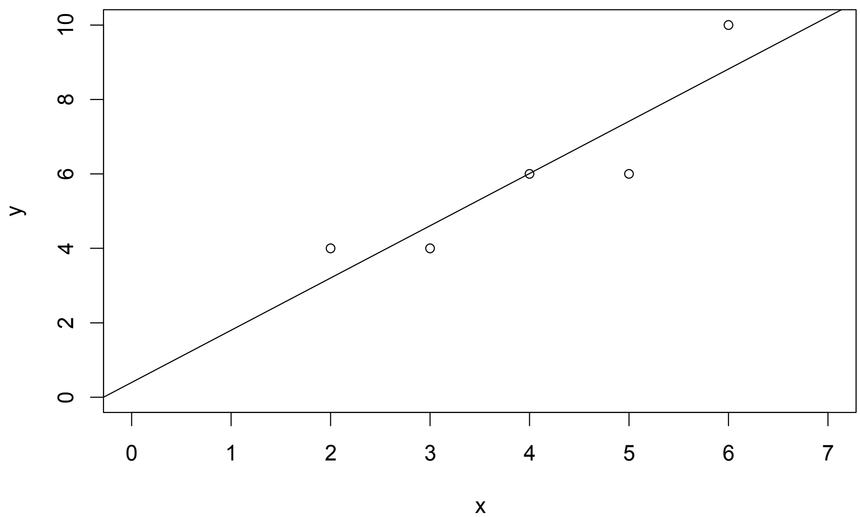
```
   fit     lwr      upr
1 10.2 6.15775 14.24225
```

b2-ch1-2.R의 실행결과

```
> x<-c(2,3,4,5,6)
> y<-c(4,4,6,6,10)
> x;y
[1] 2 3 4 5 6
[1]  4  4  6  6 10
>
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
        0.4          1.4

> ols<-lm(y~x)
> summary(ols)

Call:
lm(formula = y ~ x)

Residuals:
         1          2          3          4          5
 8.000e-01 -6.000e-01  7.404e-17 -1.400e+00  1.200e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.400      1.625   0.246   0.8214
x              1.400      0.383   3.656   0.0354 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.211 on 3 degrees of freedom
Multiple R-squared:  0.8167,     Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

```
>
> plot(y~x, ylim=c(0,10))
> abline(lm(y~x))
>
> confint(ols)
                 2.5 %    97.5 %
(Intercept) -4.7708632 5.570863
x            0.1812159 2.618784
>
> predict(lm(y~x))
  1   2   3   4   5
3.2 4.6 6.0 7.4 8.8
> new<-data.frame(x = 7)
> predict(lm(y~x), new, se.fit = TRUE)
$`fit`
   1
10.2

$se.fit
[1] 1.270171

$df
[1] 3

$residual.scale
[1] 1.21106

> pred.w.plim <- predict(lm(y ~ x), new, interval = "prediction")
> pred.w.plim
   fit      lwr      upr
1 10.2 4.614829 15.78517
> pred.w.clim<-predict(lm(y ~ x), new, se.fit=T, interval = "confidence")
> pred.w.clim
$`fit`
   fit     lwr      upr
1 10.2 6.15775 14.24225
```

```
$se.fit
[1] 1.270171

$df
[1] 3

$residual.scale
[1] 1.21106
```



7. 회귀모형의 형태

회귀분석에서 회귀함수의 형태를 선택하는 것도 매우 중요한데 경제변수 간의 함수형태는 경제이론이나 경험으로부터 결정된다. 가장 간편한 경제관계의 함수형태는 선형형태이나 경제관계를 항상 선형관계로 나타낼 수 있는 것은 아니다. 일반적으로 계량경제분석에 널리 사용되는 모형에는 다음과 같은 것들이 있다.

(1) 선형함수 모형

선형(linear)함수 모형은 다음과 같다.

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

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 $\epsilon_{Y \mid X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX}$이고, 선형함수 모형을 전미분하면 $dY = \beta_1 dX$ 즉, 기울기 $\frac{dY}{dX} = \beta_1$이므로 선형함수 모형의 탄력성은$\beta_1 \frac{X}{Y}$ 가 된다.

(2) 전대수함수 모형

전대수(double log)함수 모형은 다음과 같은데 대표적인 예는 Cobb-Douglas 생산함수에 로그를 취한 함수이다.

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

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) 쌍곡선함수 모형

쌍곡선(reciprocal)함수 모형은 다음과 같은데 물가상승률과 실업률의 역관계를 나타낸 주는 필립스 곡선이 좋은 예이다.

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

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

쌍곡선함수 모형을 전미분하면 $dY = -\beta_1 X^{-2}dX$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) 반로그함수 모형

반로그(semi-log)함수 모형으로 종속변수가 반로그함수인 모형과 독립변수가 반로그함수인 모형으로 구분될 수 있는데 다음과 같다.

(종속변수가 반로그함수인 모형) $\ln Y_i = \beta_0 + \beta_1 X_i + u_i$

(독립변수가 반로그함수인 모형) $Y_i = \beta_0 + \beta_1 \ln X_i + u_i$

종속변수가 반로그함수인 모형의 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.

독립변수가 반로그인함수인 모형의 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 2-1>과 같다.

<표 2-1> 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)	$\ln Y_i=\ln\beta_0+\beta_1\ln X_i+u_i$	$\beta_1\frac{Y}{X}$	β_1
쌍곡선 (reciprocal)	$Y_i=\beta_0+\beta_1(\frac{1}{X_i})+u_i$	$-\beta_1\frac{1}{X^2}$	$-\beta_1\frac{1}{XY}$
반로그 (semi-log)	$\ln Y_i=\beta_0+\beta_1 X_i+u_i$	$\beta_1 Y$	$\beta_1 X$
반로그 (semi-log)	$Y_i=\beta_0+\beta_1\ln X_i+u_i$	$\beta_1\frac{1}{X}$	$\beta_1\frac{1}{Y}$

(예제 2-1)을 <표 2-1>의 함수형태에 따라 추정한 b2-ch2-3.R을 실행해 보면 다음과 같은 결과를 얻을 수 있다.

첫째, 선형함수 모형을 추정하면 기울기는 1.4로써 독립변수 X가 한 단위 증가하면 종속변수 Y는 1.4 단위 증가하는 것으로 해석할 수 있다. 동 모형에서 탄력성은

$\beta_1 \frac{X}{Y}$인데 β_1의 추정치는 1.4이지만 X 및 Y의 어떤 자료로 탄력성을 구하는 지 불분명하므로 선형함수 모형은 기울기를 구하는데 적합하다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.400      1.625   0.246   0.8214
x              1.400      0.383   3.656   0.0354 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

둘째, 전대수함수 모형을 추정하면 탄력도는 0.7755로써 독립변수 X가 1% 상승하면 종속변수 Y는 0.7755% 상승하는 것으로 해석할 수 있다. 동 모형에서 기울기는 $\beta_1 \frac{Y}{X}$인데 β_1의 추정치는 0.7755이지만 X 및 Y의 어떤 자료로 기울기를 구하는 지 불분명하므로 전대수함수 모형은 탄력도를 구하는데 적합하다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7112     0.3150   2.258   0.1092
lx            0.7755     0.2296   3.377   0.0432 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b2-ch2-3.R의 실행결과

```
> x<-c(2,3,4,5,6)

> y<-c(4,4,6,6,10)

> x;y
[1] 2 3 4 5 6
```

```
[1]  4  4  6  6 10

> lx<-log(x)

> ly<-log(y)

> lx;ly
[1] 0.6931472 1.0986123 1.3862944 1.6094379 1.7917595
[1] 1.386294 1.386294 1.791759 1.791759 2.302585

> rx<-1/x

> rx
[1] 0.5000000 0.3333333 0.2500000 0.2000000 0.1666667

> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
        0.4          1.4


> ols1<-lm(y~x)

> summary(ols1)

Call:
lm(formula = y ~ x)

Residuals:
         1           2          3          4          5
 8.000e-01 -6.000e-01  7.404e-17 -1.400e+00  1.200e+00

Coefficients:
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.400      1.625   0.246   0.8214
x                1.400      0.383   3.656   0.0354 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.211 on 3 degrees of freedom
Multiple R-squared:  0.8167,     Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535


> lm(ly~lx)

Call:
lm(formula = ly ~ lx)

Coefficients:
(Intercept)           lx
     0.7112       0.7755


> ols2<-lm(ly~lx)

> summary(ols2)

Call:
lm(formula = ly ~ lx)

Residuals:
        1          2          3          4          5
 0.137490 -0.176966  0.005388 -0.167670  0.201757

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.7112     0.3150   2.258   0.1092
lx              0.7755     0.2296   3.377   0.0432 *
```

```
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1992 on 3 degrees of freedom
Multiple R-squared:  0.7917,       Adjusted R-squared:  0.7223
F-statistic: 11.41 on 1 and 3 DF,  p-value: 0.04318


> lm(y~rx)

Call:
lm(formula = y ~ rx)

Coefficients:
(Intercept)           rx
      10.09       -14.11


> ols3<-lm(y~rx)

> summary(ols3)

Call:
lm(formula = y ~ rx)

Residuals:
      1       2       3       4       5
 0.9624 -1.3887 -0.5643 -1.2696  2.2602

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.091      2.138   4.719    0.018 *
rx           -14.107      6.821  -2.068    0.130
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.816 on 3 degrees of freedom
Multiple R-squared:  0.5878,      Adjusted R-squared:  0.4504
F-statistic: 4.278 on 1 and 3 DF,  p-value: 0.1305


> lm(ly~x)

Call:
lm(formula = ly ~ x)

Coefficients:
(Intercept)            x
     0.8365       0.2238


> ols4<-lm(ly~x)

> summary(ols4)

Call:
lm(formula = ly ~ x)

Residuals:
       1        2        3        4        5
 0.10217 -0.12164  0.06002 -0.16378  0.12324

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8365     0.2062   4.057   0.0270 *
x             0.2238     0.0486   4.605   0.0193 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1537 on 3 degrees of freedom
Multiple R-squared:  0.8761,      Adjusted R-squared:  0.8348
```

```
F-statistic: 21.21 on 1 and 3 DF,  p-value: 0.01925


> lm(y~lx)

Call:
lm(formula = y ~ lx)

Coefficients:
(Intercept)           lx
    -0.2655       4.7615


> ols5<-lm(y~lx)

> summary(ols5)

Call:
lm(formula = y ~ lx)

Residuals:
      1       2       3       4       5
 0.9650 -0.9656 -0.3354 -1.3979  1.7339

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2655     2.4046  -0.110   0.9191
lx            4.7615     1.7528   2.716   0.0728 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.521 on 3 degrees of freedom
Multiple R-squared:  0.711,     Adjusted R-squared:  0.6146
F-statistic: 7.379 on 1 and 3 DF,  p-value: 0.07277
```

(예제 2-1)을 OLS, MLE, LAD로 추정한 b2-ch2-5.R을 실행해 보면 다음과 같은 결과를 얻을 수 있다. 그림에서 점선은 보통최소자승법에 의한 추정 회귀식이고, 파란 선은 최소절대편차 추정법에 의한 추정 회귀식을 나타낸다,

첫째, OLS에 의한 회귀계수의 추정치는 $\hat{\beta}_0 = 0.4, \hat{\beta}_1 = 1.4$ 이고, 교란항의 표준오차 추정치는 $\hat{\sigma} = 1.211$ 이다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.400      1.625   0.246   0.8214
x             1.400      0.383   3.656   0.0354 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.211 on 3 degrees of freedom
```

둘째, LAD에 의한 회귀계수의 추정치는 $\hat{\beta}_0 = 2.0, \hat{\beta}_1 = 1.0$ 인데 해가 유일하지 않다는 메시지가 나왔다.

```
Coefficients:
            coefficients   lower bd        upper bd
(Intercept) 2.000000e+00 -1.797693e+308 1.797693e+308
x           1.000000e+00 -1.797693e+308 1.797693e+308

1: In rq.fit.br(x, y, tau = tau, ...) : Solution may be nonunique
2: In rq.fit.br(x, y, tau = tau, ...) : Solution may be nonunique
3: In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) :
  Solution may be nonunique
```

이를 좀 더 구체적으로 살펴보기 위하여 다음과 같은 LAD에 의한 두 개의 추정 회귀식을 얻는 b2-ch2-4.R을 실행해 보면 아래 그림에서 확인할 수 있듯이 최소한 두 개의 데이터 포인트를 지나는 다음과 같은 두 개의 추정 회귀식을 얻을 수 있다.

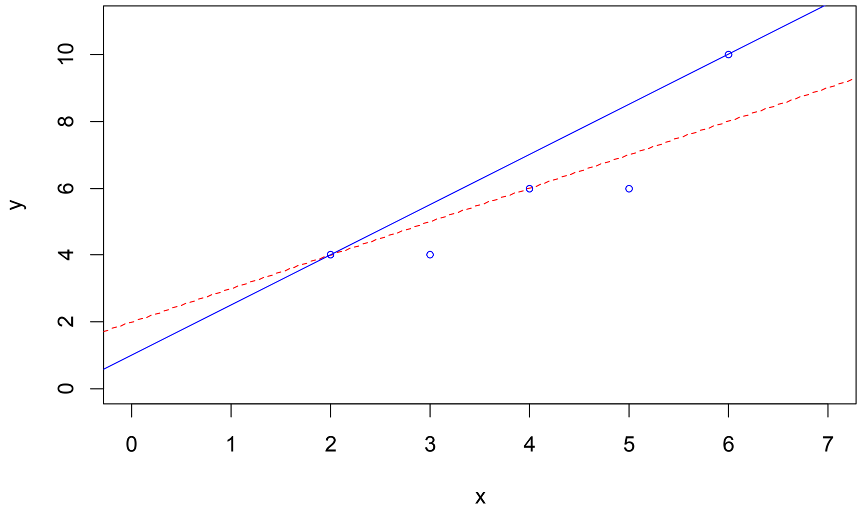
추정 회귀식 1 : $\hat{Y}=2+X$
추정 회귀식 2 : $\hat{Y}=1+1.5X$

이 두 회귀선이 교차하면서 발생하는 두 회귀선 사이의 영역에 최소절댓값 추정법의 다수 해가 존재한다.

b2-ch2-4.R의 실행결과

```
> library(openxlsx)

> library(bbmle)

> library(quantreg)

> x<-c(2,3,4,5,6)

> y<-c(4,4,6,6,10)

> # LAD estimator
> taus<-c(0.5, 0.7)

> rq(y~x)
Call:
rq(formula = y ~ x)

Coefficients:
(Intercept)          x
          2          1

Degrees of freedom: 5 total; 3 residual

> lad<-rq(y~x, tau=taus)

> summary(lad)

Call: rq(formula = y ~ x, tau = taus)
```

```
tau: [1] 0.5

Coefficients:
              coefficients    lower bd          upper bd
(Intercept)    2.000000e+00  -1.797693e+308   1.797693e+308
x              1.000000e+00  -1.797693e+308   1.797693e+308

Call: rq(formula = y ~ x, tau = taus)

tau: [1] 0.7

Coefficients:
              coefficients    lower bd          upper bd
(Intercept)    1.000000e+00  -1.797693e+308   1.797693e+308
x              1.500000e+00  -1.797693e+308   1.797693e+308

> plot(x,y,type = "n", cex = .8,xlim=c(0,7), ylim=c(0,11))

> points(x,y,cex = .8,col="blue",xlim=c(0,7), ylim=c(0,11))

> abline(rq(y~x, tau=0.5),col="red",lty = 2)

> abline(rq(y~x, tau=0.7), col="blue")
Warning messages:
1: In rq.fit.br(x, y, tau = tau, ...) : Solution may be nonunique
2: In rq.fit.br(x, y, tau = tau, ...) : Solution may be nonunique
3: In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) :
  Solution may be nonunique
4: In rq.fit.br(x, y, tau = tau, ci = TRUE, ...) :
  Solution may be nonunique
5: In rq.fit.br(x, y, tau = tau, ...) : Solution may be nonunique
```

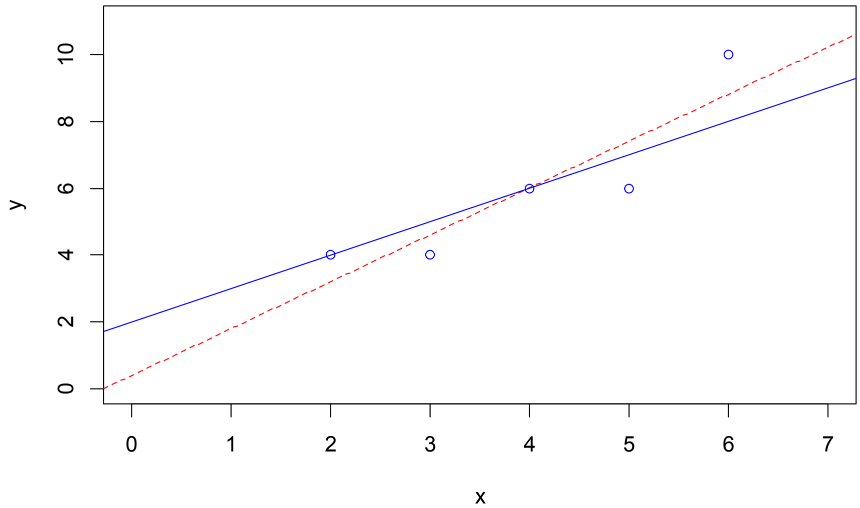


셋째, MLE에 의한 회귀계수의 추정치는 $\hat{\beta}_0 = 0.4, \hat{\beta}_1 = 1.4$ 이고, 교란항의 표준오차 추정치는 $\hat{\sigma} = 0.93808$ 이다. 회귀계수의 추정치는 OLS 및 MLE가 동일하지만 교란항의 표준편차 추정치는 다르게 나타나고 있다. 그 이유는 교란항의 분산을 구할 때, OLS의 경우 잔차의 제곱의 합을 자유도(여기서는 3)로 나누어 주는 반면에 MLE의 경우 잔차의 제곱의 합을 관측치 수(여기서는 5)로 나누어주기 때문이다.

```
Coefficients:
        Estimate Std. Error z value     Pr(z)
b0       0.40000    1.25857  0.3178  0.750621
b1       1.40000    0.29665  4.7194 2.365e-06 ***
sigma    0.93808    0.29665  3.1623  0.001565 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b2-ch2-5.R의 실행결과

```
library(openxlsx)
library(bbmle)
library(quantreg)

x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)

(n<-length(y))

# OLS estimator
lm(y~x)
ols<-lm(y~x)
summary(ols)
confint(ols)

# LAD estimator
#taus<-c(.05,.1,.25,.75,.9, 0.95)
rq(y~x)
lad<-rq(y~x)
summary(lad)

# MLE
fn<-function(b0, b1, sigma) {
  (n/2)*log(sigma^2) + 1/(2*sigma^2)*(sum((y-b0-b1*x)^2))
}
res<-mle2(fn,start=list(b0=0.5, b1=0.5, sigma=1))
summary(res)

plot(x,y,type = "n", cex=.8, xlim=c(0,7), ylim=c(0,11))
points(x,y,cex=.8,col="blue",xlim=c(0,7), ylim=c(0,11))
abline(lm(y~x),col="red",lty = 2)
abline(rq(y~x), col="blue")
```



한편, 가상적인 데이터를 이용하여 OLS, MLE, LAD로 추정한 b2-ch2-6.R을 실행해 보면 다음과 같은 결과를 얻을 수 있는데 우리가 관심을 가지고 있는 기울기의 경우 세 추정치 모두 거의 같고, 특히 OLS 및 MEL 추정치는 거의 같음을 보여주고 있고, 그림에서 보듯이 OLS와 LAD의 추정 회귀식도 유사하여 어 세 가지 추정방법의 유용성을 확인할 수 있다.

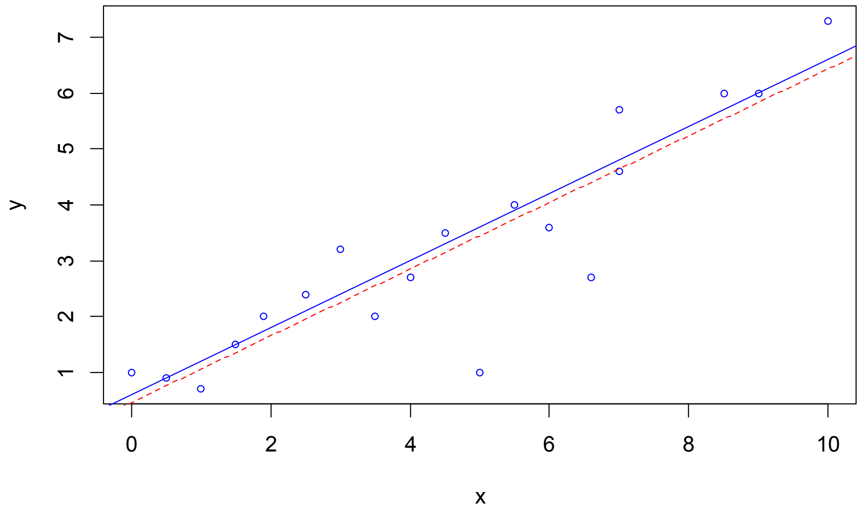
구분	OLS	LAD	MLE
$\hat{\beta}_0$	0.47172	0.6	0.471688
$\hat{\beta}_1$	0.59583	0.6	0.595838

b2-ch2-6.R의 실행결과

```
> library(openxlsx)

> library(bbmle)

> library(quantreg)

> x<-c(0,1.9,4,6,8.5,0.5,2.5,4.5,6.6,9,1,3,5,7,10,1.5,3.5,5.5,7)

> y<-c(1,2,2.7,3.6,6,0.9,2.4,3.5,2.7,6,0.7,3.2,1,5.7,7.3,1.5,2,4,4.6)

> (n<-length(y))
[1] 19

> # OLS estimator
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.4717       0.5958



> ols<-lm(y~x)

> summary(ols)

Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-2.4509 -0.2613  0.1658  0.4512  1.0575

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47172    0.38219   1.234    0.234
x            0.59583    0.07059   8.440 1.75e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8889 on 17 degrees of freedom
Multiple R-squared:  0.8073,    Adjusted R-squared:  0.796
F-statistic: 71.24 on 1 and 17 DF,  p-value: 1.746e-07


> confint(ols)
                 2.5 %    97.5 %
(Intercept) -0.3346268 1.2780661
x            0.4468945 0.7447682

> # LAD estimator
> rq(y~x)
Call:
rq(formula = y ~ x)

Coefficients:
(Intercept)           x
        0.6         0.6

Degrees of freedom: 19 total; 17 residual

> lad<-rq(y~x, model=T)

> summary(lad)

Call: rq(formula = y ~ x, model = T)

tau: [1] 0.5

Coefficients:
```

```
             coefficients lower bd upper bd
(Intercept) 0.60000       0.06835  0.95858
x            0.60000       0.51754  0.67755

> # MLE
> fn<-function(b0, b1, sigma) {
+    (n/2)*log(sigma^2)+1/(2*sigma^2)*(sum((y-b0-b1*x)^2))
+ }

> res<-mle2(fn,start=list(b0=0.5, b1=0.5, sigma=1))

> summary(res)
Maximum likelihood estimation

Call:
mle2(minuslogl = fn, start = list(b0 = 0.5, b1 = 0.5, sigma = 1))

Coefficients:
       Estimate Std. Error z value     Pr(z)
b0     0.471688   0.361513  1.3048      0.192
b1     0.595838   0.066774  8.9233 < 2.2e-16 ***
sigma  0.840795   0.136395  6.1644 7.074e-10 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-2 log L: 12.41055

> plot(x,y,type = "n", cex = .8)

> points(x,y,cex = .8,col="blue")

> abline(lm(y~x),col="red",lty = 2)

> abline(rq(y~x), col="blue")
```



8. OLS 추정량의 특성

좋은 추정량의 기준으로 볼 때 OLS 추정량인 $\widehat{\beta}_0$, $\widehat{\beta}_1$은 어떠한 추정량인가? 이에 대한 대답은 $\widehat{\beta}_0$, $\widehat{\beta}_1$은 선형성, 불편성 및 최소분산의 특성을 가지고 있다는 것이다. 이러한 특성을 가진 추정량을 최량선형불편추정량(Best Linear Unbiased Estimator: BLUE)이라고 하는데 OLS 추정량은 BLUE임을 보여주는 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

제 3 장

다중회귀분석

1. 다중회귀모형
2. 회귀계수의 추정
3. 회귀계수의 분산 추정
4. 결정계수
5. 가설검정
6. 예측

제3장 다중회귀분석

1. 다중회귀모형

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.

상수항과 k-1개의 독립변수에 의해 설명되는 일반적인 다중회귀모형은 다음과 같이 나타낸다.

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

다중회귀모형에서 회귀계수는 다음과 같은 의미를 가진다. 먼저 상수항 β_1은 독립변수의 값이 0일 경우 종속변수의 기댓값을 나타내는데 독립변수의 값이 0일 경우는 별로 없으므로 관심의 대상이 되지 않는다.

독립변수와 종속변수의 관계를 나타내는 회귀계수 β_2, β_3,..., β_k 를 부분회귀계수라고 하는데 예를 들어 β_3는 다른 독립변수들의 값이 변하지 않고 X_3의 값이 한 단위 증가할 때 종속변수의 평균변화의 크기를 의미한다.

다중회귀모형을 행렬로 표현하면 다음과 같다.

$$\underset{(n\times 1)}{Y} = \underset{(n\times k)}{X} \ \underset{(k\times 1)}{B} + \underset{(n\times 1)}{U}$$

단, $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ Y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 \ X_{21} \ X_{31} \cdots \ X_{k1} \\ 1 \ X_{22} \ X_{32} \cdots \ X_{k2} \\ . \quad . \quad . \quad . \\ 1 \ X_{2n} \ X_{3n} \cdots \ X_{kn} \end{bmatrix}$, $B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix}$, $U = \begin{bmatrix} u_1 \\ u_2 \\ . \\ u_n \end{bmatrix}$

다중회귀모형에 대한 가정은 단순회귀모형과 동일하다.

첫째, 독립변수 X는 비확률변수이다. 이 가정은 표본을 반복해서 추출할 때 X는 고정된 것으로 해 놓고 Y를 추출한다는 것을 의미하므로 X는 확정변수이며 X는 오차 없이 측정되며 full rank를 가진다(즉, 독립변수 간에 정확한 선형관계가 없다)는 의미이다.

둘째, 교란항의 평균은 0이다. 이에 대한 자세한 설명은 단순회귀모형에서 한 바 있다.

셋째, 교란항은 모든 X_i에 대한 동일한 분산을 가지고 자기상관이 없다 즉, $E(UU') = \sigma_u^2 I_n$이다. 이 가정을 행렬로 표현해 보면 다음과 같게 되는데 이 가정은 교란항의 분산이 일정하고(이를 동분산(homoscedasticity)라고 함), 또한 대각외원소가 모두 0(이를 비자기상관(no autocorrelation)이라고 함)이라는 것이다.

$$E(UU') = \begin{bmatrix} E(u_1^2) & E(u_1 u_2) & .. & E(u_1 u_n) \\ E(u_2 u_1) & E(u_2^2) & .. & E(u_2 u_k) \\ . & . & .. & . \\ E(u_n u_1) & E(u_n u_2) & .. & E(u_n^2) \end{bmatrix}$$

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

$$= \sigma_u^2 I_n$$

넷째, u와 X는 서로 독립이다. 이에 대한 자세한 설명은 단순회귀모형에서 한 바 있다.

2. 회귀계수의 추정

(1) 단일 회귀계수별 추정

단순회귀모형에서 기울기를 나타내는 회귀계수 β_1에 대한 OLS 추정량은 $\hat{\beta}_1 = \dfrac{Cov(X,Y)}{Var(X)}$ 임을 이미 알고 있다.

다중회귀모형에서 회귀계수 $\beta_k, (k=2,3,...,k)$ 에 대한 OLS 추정량은 다음과 같다.

$$\hat{\beta}_k = \frac{Cov(\widetilde{X}_k, Y)}{Var(\widetilde{X}_k)}$$

예를 들어, 다음과 같은 다중회귀모형을 가정해 보자.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

위 모형에서 회귀계수 β_2에 대한 OLS 추정량은 다음의 다중회귀모형에서 $\widetilde{X}_2$에 대한 회귀계수 추정량이 된다.

$$Y_i = \beta_1 + \beta_2 \widetilde{X_{2i}} + \beta_3 X_{3i} + u_i$$

$$\hat{\beta}_2 = \frac{Cov(\widetilde{X}_{2,} Y)}{Var(\widetilde{X}_2)}$$

단, $\widetilde{X}_{2i}$는 $X_{2i} = \alpha_1 + \alpha_2 X_{3i} + \widetilde{X}_{2i}$ 인데 즉, 독립변수를 X_3, 종속변수를 X_2로 하는 단순회귀모형의 잔차이다.

또한 회귀계수 $\hat{\beta}_3$에 대한 OLS 추정량은 다음의 다중회귀모형에서 $\widetilde{X}_3$에 대한 회귀계수 추정량이 된다.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 \widetilde{X_{3i}} + u_i$$

$$\hat{\beta}_3 = \frac{Cov(\widetilde{X}_{3,} Y)}{Var(\widetilde{X}_3)}$$

단, $\widetilde{X}_{3i}$는 $X_{3i} = \gamma_1 + \gamma_2 X_{2i} + \widetilde{X}_{3i}$ 인데 즉, 독립변수를 X_2, 종속변수를 X_3로 하는 단순회귀모형의 잔차이다.

(2) 베타회귀계수의 추정

회귀모형에서 독립변수의 상대적인 중요성을 알고자 할 경우 독립변수 및 종속변수를 표준화한 다음 표준화된 변수로 설정한 다음과 같은 회귀모형을 추정하면 되는데 이때 추정된 회귀계수를 베타회귀계수(Beta coefficients or standardized regression coefficients)라고 한다.

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

OLS로 추정된 베타회귀계수($\widehat{\beta^*}$)와 회귀계수($\hat{\beta}$)는 다음과 같은 관계가 있다.

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

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 3-1)
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 베타회귀계수를 구하고 해석하라.

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

다중회귀모형에서 회귀계수를 추정하고, 베타회귀계수를 추정하는 b2-ch3-1.R을 실행해 보면 다음과 같은 결과를 얻을 수 있다.

첫째, $\hat{\beta}_2$는 $Y_i = \beta_1 + \beta_2 \widetilde{X_{2i}} + \beta_3 X_{3i} + u_i$ 를 추정하여 구하므로 resx2의 회귀계수 $\hat{\beta}_2 = 1.5$ 임을 확인할 수 있고, $\hat{\beta}_3$는 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 \widetilde{X_{3i}} + u_i$ 를 추정하여 구하므로 resx3의 회귀계수 $\hat{\beta}_3 = 2.0$ 임을 확인할 수 있다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0000     0.7906   1.265    0.426
resx2         1.5000     0.5000   3.000    0.205
x3            0.5000     0.5000   1.000    0.500
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7500     0.7500   1.000    0.500
x2            0.5000     0.3536   1.414    0.392
resx3         2.0000     0.7071   2.828    0.216
```

둘째, 베타회귀계수를 구하면 $\widehat{\beta_2^*}=1.2792$, $\widehat{\beta_3^*}=1.2060$ 이므로 Y 변동은 X_3 변동보다 X_2 변동에 더 크게 반응하는 것을 알 수 있다. 베타회귀계수를 추정할 때 유의할 점은 상수항이 없는 회귀모형을 추정해야 한다는 것이다.

```
Coefficients:
     Estimate  Std. Error  t value  Pr(>|t|)
sx2    1.2792      0.3015    4.243    0.0513 .
sx3    1.2060      0.3015    4.000    0.0572 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

셋째, 베타회귀계수($\widehat{\beta^*}$)와 회귀계수($\hat{\beta}$)의 관계를 이용하여 추정된 베타회귀계수 $\widehat{\beta_2^*}=1.2792$, $\widehat{\beta_3^*}=1.2060$ 로 회귀계수를 계산해 보면 $\widehat{\beta_2}=1.5$, $\widehat{\beta_3}=2.0$ 이 됨을 확인할 수 있다.

b2-ch3-1.R의 실행결과

```
> x2<-c(1,2,3,2)

> x3<-c(2,1,1,2)

> y0<-c(1,1,2,3)

> n<-length(y)

> olsx2<-lm(x2~x3)

> resx2<-resid(olsx2)

> ols2<-lm(y0~resx2+x3)

> summary(ols2)
```

```
Call:
lm(formula = y0 ~ resx2 + x3)

Residuals:
    1      2      3      4
-0.25   0.25 -0.25   0.25

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0000     0.7906   1.265    0.426
resx2         1.5000     0.5000   3.000    0.205
x3            0.5000     0.5000   1.000    0.500

Residual standard error: 0.5 on 1 degrees of freedom
Multiple R-squared:  0.9091,     Adjusted R-squared:  0.7273
F-statistic:     5 on 2 and 1 DF,  p-value: 0.3015


> olsx3<-lm(x3~x2)

> resx3<-resid(olsx3)

> ols3<-lm(y0~x2+resx3)

> summary(ols3)

Call:
lm(formula = y0 ~ x2 + resx3)

Residuals:
    1      2      3      4
-0.25   0.25 -0.25   0.25

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7500     0.7500   1.000    0.500
x2            0.5000     0.3536   1.414    0.392
```

```
resxx3          2.0000      0.7071    2.828     0.216

Residual standard error: 0.5 on 1 degrees of freedom
Multiple R-squared:  0.9091,       Adjusted R-squared:  0.7273
F-statistic:      5 on 2 and 1 DF,  p-value: 0.3015


> y0bar<-mean(y0)

> x2bar<-mean(x2)

> x3bar<-mean(x3)

> y0sd<-sqrt(var(y0))

> x2sd<-sqrt(var(x2))

> x3sd<-sqrt(var(x3))

> sy0<-(y0-y0bar)/y0sd

> sx2<-(x2-x2bar)/x2sd

> sx3<-(x3-x3bar)/x3sd

> sols<-lm(sy0~sx2+sx3-1)

> summary(sols)

Call:
lm(formula = sy0 ~ sx2 + sx3 - 1)

Residuals:
      1         2         3         4
-0.2611   0.2611  -0.2611   0.2611

Coefficients:
```

```
    Estimate Std. Error t value Pr(>|t|)
sx2   1.2792     0.3015   4.243   0.0513 .
sx3   1.2060     0.3015   4.000   0.0572 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3693 on 2 degrees of freedom
Multiple R-squared:  0.9091,     Adjusted R-squared:  0.8182
F-statistic:    10 on 2 and 2 DF,  p-value: 0.09091


> y0sd
[1] 0.9574271

> x2sd
[1] 0.8164966

> x3sd
[1] 0.5773503

> beta2hat<-summary(sols)$coef[1,1]*(y0sd/x2sd)

> beta3hat<-summary(sols)$coef[2,1]*(y0sd/x3sd)

> beta2hat
[1] 1.5

> beta3hat
[1] 2
```

(3) 행렬을 이용한 추정

행렬로 표현한 다중회귀모형을 OLS로 추정하는 방법을 살펴보자. 행렬로 표현한 다중회귀모형에서 잔차는 다음과 같다.

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

보통최소자승법은 잔차의 제곱의 합을 최소화하므로 잔차의 제곱의 합 즉, $e'e$은 다음과 같게 된다.

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

위 식을 $\hat{B}$으로 편미분하여 0으로 두면 다음과 같은 정규방정식이 도출된다.

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

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

정규방정식의 해를 구하면 다음과 같이 $\hat{B}$에 대한 OLS 추정량을 구할 수 있다.

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

(예제 3-1) (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 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.

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

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

따라서, 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 연간매출액은 평균 15억 원 증가하며, 홍보비 지출액이 변화하지 않을 경우 연구개발 지출액이 천만 원 증가하면 연간매출액은 평균 20억 원 증가함을 의미한다.

3. 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 불편추정량은 $\widehat{\sigma_u^2} = \dfrac{e'e}{n-k}$이다. k는 상수항을 포함한 독립변수의 수이고, $e'e$는 다음과 같이 구할 수 있다.

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

한편, 회귀계수 $\hat{B}$의 분산은 다음 식과 같게 된다.

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

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

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

회귀계수의 분산-공분산행렬에서 대각행렬에 있는 원소들은 회귀계수의 분산이 되고, 비대각행렬에 있는 원소들은 회귀계수의 공분산이 된다.

4. 결정계수

제2장의 <그림 2-2>에서 보여 주고 있듯이 $Y=\hat{Y}+e$이다. 총변동인 $Y'Y$는 다음과 같이 나타낼 수 있다.

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

위 식은 단순회귀분석과 마찬가지로 총변동 = 회귀변동 + 잔차변동의 관계가 성립함을 보여주고 있다.

결정계수의 정의에 따라 결정계수는 다음과 같은데 이 식은 종속변수의 평균이 0일 경우에 활용될 수 있다.

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

일반적으로 종속변수의 평균이 0이 아니기 때문에 이 경우 즉, $\overline{Y}\neq 0$이면 수정항(correction term) $n\overline{Y}^2$이 필요하고, 이를 반영한 결정계수는 다음과 같다.

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

결정계수는 설명변수의 설명력을 나타내므로 클수록 모형의 설명력이 높고, 모형의 적합도가 높다는 것을 의미하지만 해석에 있어서 주의해야 할 사항이 있다.

첫째, 설명변수가 비록 통계적으로 유의하지 않더라고 설명변수가 추가되기만 하면 결정계수는 높아진다. 따라서 설명변수의 통계적 유의성과 결정계수를 모두 고려

해야 한다.

둘째, 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 3-1)(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)

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

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

$$= 15 - 14.75$$

$$= 0.25$$

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

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

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

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

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

5. 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 다음과 같다.

$$H_0 : RB = r$$

단, R은 $q \times k$의 주어진 행렬, r은 $q \times 1$의 주어진 벡터, q는 제약의 수($q \leq k$)를 각각 나타낸다.

(1) 단일검정

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

$$R=[0\,0 \ldots, 1, 0, \ldots, 0] \quad \text{과} \quad r=0$$

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

이상과 같은 일반적인 귀무가설 하에서 다음의 F-통계량은 분자의 자유도가 1이고 분모의 자유도가 $n-k$인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

$$F = \frac{\widehat{\beta_i^2}}{\widehat{\sigma_u^2} a_{ii}} \sim F^1_{n-k}$$

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

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

$$t = \frac{\widehat{\beta_i}}{\sqrt{\widehat{\sigma_u^2} a_{ii}}} \sim t_{n-k}$$

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.

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

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

이상과 같은 일반적인 귀무가설 하에서 다음의 F-통계량은 분자의 자유도가 q이고 분모의 자유도가 $n-k$인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

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

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

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 남겨 두는 것은 아니라 개별 회귀계수에 대한 검정을 통해서 개별 변수의 모형 내 존치 여부를 결정해야 한다.

예를 들어 다음과 같은 화폐수요함수를 나타내는 다중회귀모형이 있다고 하자.

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

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.

위 식을 제약이 가해지지 않은 모형이라고 하고 이 식을 추정한 후 얻게 되는 결정계수를 R_{ur}^2 이라고 한다. 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.

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

$$H_0 : \beta_4 = \beta_5 = 0 \quad (q = 2)$$

위 귀무가설 하에서 제약이 가해지지 않은 화폐수요함수는 다음과 같이 나타낼 수 있는데 이를 제약이 가해진 모형이라고 한다.

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

위 식을 추정한 후 얻게 되는 결정계수를 R_r^2이라고 한다.

제약이 가해지지 않은 모형 및 제약이 가해진 모형의 추정을 통해 구한 R_{ur}^2 및 R_r^2를 이용하여 계산한 F-통계량은 분자의 자유도가 2이고 분모의 자유도가 $n-k$인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

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

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 단일검정을 통해서 개별 변수의 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.

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

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

이상과 같은 일반적인 귀무가설 하에서 다음의 F-통계량은 분자의 자유도가 $k-1$이고 분모의 자유도가 $n-k$인 F-분포를 따르므로 F-분포표를 이용하면 의사결정을 할 수 있다.

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

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

(예제 2-1)(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 1일 때 임계값은 12.708 즉, $t_{1,0.025} = 12.708$ 이다. 위에서 계산한 t-통계량의 값 3.0은 임계값 12.708보다 작으므로 5% 유의수준 하에서 귀무가설을 기각하지 못한다. 따라서 홍보비 지출액은 연간매출액에 영향을 주지 못한다고 해석하고 $\widehat{\beta_2}$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 12.708보다 작으므로 5% 유의수준 하에서 귀무가설을 기각하지 못한다. 따라서 연구개발 지출액은 연간매출액에 영향을 주지 못한다고 해석하고 $\widehat{\beta_3}$는 통계적으로 유의하지 않다라고 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 F-통계량을 계산해 보면

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

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 연간매출액에 영향을 주지 못하는 변수이며 따라서 모형설정이 완전히 잘못된 것으로 해석할 수 있다. 이 경우에는 새로운 독립변수로 모형을 설정한 후 위의 절차를 다시 거쳐야 한다.

6. 예측

일반적으로 예측에는 주어진 X에 대해 하나의 Y값을 구하는 점예측과 점예측에 대한 신뢰구간을 구하는 구간예측이 있다.

독립변수 X가 특정한 값 x_0일 때 개별 Y_0의 예측치는 다음과 같다.

$$\widehat{Y_0} = x_0' \widehat{B}$$

구간예측을 위해서는 실제치인 Y_0와 예측치 $\widehat{Y_0}$의 차이인 예측오차의 분산(σ_ϵ^2)을

알아야 하며 예측오차의 분산은 평균예측과 개별예측이 서로 다르다.
평균예측 예측오차의 분산 및 개별예측 예측오차의 분산은 각각 다음과 같다.

(평균예측 예측오차의 분산) $Var(\widehat{Y_0} \mid x_0) = \widehat{\sigma_u^2} x_0{}'(X'X)^{-1}x_0$

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

따라서 평균예측 및 개별예측에 대한 $(1-\alpha)\times 100\%$ 예측구간은 각각 다음과 같다.

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

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

(예제 3-1)(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.

$$\widehat{Y_0} = x_0'\hat{\beta}$$

$$= [1\ 3\ 2]\begin{bmatrix} -4.25 \\ 1.5 \\ 2.0 \end{bmatrix}$$

$$= 4.25$$

평균예측의 예측오차 분산 및 개별예측의 예측오차 분산은 각각 다음과 같다.

(평균예측의 예측오차 분산) $\widehat{\sigma_u^2} x_0^{'}(X'X)^{-1}x_0$

$$= (0.25)[1\ 3\ 2]\begin{bmatrix} 14.75 & -3.5 & -5 \\ -3.5 & 1 & 1 \\ -5 & 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$

$$= 0.6875$$

(개별예측의 예측오차 분산) $\widehat{\sigma_u^2}[1 + x_0^{'}(X'X)^{-1}x_0]$

$$= (0.25)(1 + [1\ 3\ 2]\begin{bmatrix} 14.75 & -3.5 & -5 \\ -3.5 & 1 & 1 \\ -5 & 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix})$$

$$= 0.9375$$

따라서 평균예측의 95% 예측구간 및 개별예측의 95% 예측구간은 각각 다음과 같다.

(평균예측의 예측오차 분산) $\widehat{Y_0} \pm t_{\frac{\alpha}{2}, n-k}\sqrt{\widehat{\sigma_u^2} x_o^{'}(X'X)^{-1}x_0}$

$$= 4.25 \pm (12.706)(0.829) = [-6.285, 14.785]$$

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

$$= 4.25 \pm (12.706)(0.968) = [-8.049, 16.549]$$

지금까지 살펴본 다중회귀분석에서의 여러 결과들 즉, 회귀계수 추정량, 교란항의 분산 추정량, 회귀계수의 분산 추정량, 잔차의 제곱의 합, 결정계수, 예측오차의 분산 등을 제2장의 단순회귀분석에서 결과들과 비교해 정리해 보면 <표 3-1>과 같은데 다중회귀분석에서의 결과들은 행렬로 표현되었을 뿐이지 그 결과는 단순회귀분석에서의 결과와 동일함을 확인할 수 있다.

<표 3-1> 단순회귀 및 행렬표현 다중회귀의 비교

구분	단순회귀	다중회귀
$\hat{\beta}$	$\hat{\beta}_1 = \dfrac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X_i-\overline{X})^2} = \dfrac{\sum x_i y_i}{\sum x_i^2}$	$\hat{B} = (X'X)^{-1}X'Y$
$\hat{\sigma}_u^2$	$\dfrac{\sum e_i^2}{n-2}$	$\dfrac{e'e}{n-k}$
$Var(\hat{\beta})$	$\hat{\sigma}_u^2 \dfrac{1}{\sum x_i^2}$	$\hat{\sigma}_u^2 (X'X)^{-1}$
$\sum e_i^2$	$\sum(Y_i-\overline{Y})^2 - \hat{\beta}_1 \sum(X_i-\overline{X})(Y_i-\overline{Y})$	$Y'Y - \hat{B}'X'Y$
R^2	$\dfrac{\hat{\beta}_1^2 \sum(X_i-\overline{X})^2}{\sum(Y_i-\overline{Y})^2}$	$\dfrac{\hat{B}'X'X\hat{B} - n\overline{Y^2}}{Y'Y - n\overline{Y^2}}$
σ_ε^2 (평균예측)	$\hat{\sigma}_u^2 (\dfrac{1}{n} + \dfrac{(X_0-\overline{X})^2}{\sum x_i^2})$	$\hat{\sigma}_u^2 x_0'(X'X)^{-1}x_0$
σ_ε^2 (개별예측)	$\hat{\sigma}_u^2 (1 + \dfrac{1}{n} + \dfrac{(X_0-\overline{X})^2}{\sum x_i^2})$	$\hat{\sigma}_u^2 (1 + x_0'(X'X)^{-1}x_0)$

지금까지 행렬을 이용한 다중회귀모형에 대해 설명한 내용을 b2-ch3-2.R의 실행 결과를 통해 확인할 수 있다.

첫째, 추정된 회귀계수 $\hat{\beta}_1 = -4.25$, $\hat{\beta}_2 = 1.5$, $\hat{\beta}_3 = 2.0$ 이다.

둘째, 추정된 교란항의 분산 $\hat{\sigma}_u^2 = 0.25$ 이다.

셋째, 결정계수 $R^2 = 0.9090909$ 이다.

넷째, 추정된 회귀계수의 분산은 $Var(\hat{\beta}_1) = 3.6875$, $Var(\hat{\beta}_2) = 0.25$, $Var(\hat{\beta}_3) = 0.5$ 이다.

다섯째, $X_2 = 3$, $X_3 = 2$ 일 때, 개별예측치 및 평균예측치는 모두 4.25이다. 한편, 개별예측의 예측오차 분산은 0.9375이고 95% 예측구간은 [-8.05273, 16.55273]이며, 평균예측의 예측오차 분산은 0.6875이고 95% 예측구간은 [-6.285428, 14.785428]이다.

b2-ch3-2.R의 실행결과

```
> x2<-c(1,2,3,2)

> x3<-c(2,1,1,2)

> y0<-c(1,1,2,3)

> xls<-matrix(c(1,2,3,2,2,1,1,2), nrow=4, ncol=2)

> x<-matrix(c(1,1,1,1,1,2,3,2,2,1,1,2), nrow=4, ncol=3)

> y<-matrix(c(1,1,2,3), nrow=4)

> n<-length(y)

> x
     [,1] [,2] [,3]
[1,]    1    1    2
[2,]    1    2    1
[3,]    1    3    1
[4,]    1    2    2

> xpx=t(x)%*%x

> xpx
     [,1] [,2] [,3]
[1,]    4    8    6
[2,]    8   18   11
[3,]    6   11   10

> xpxinv=solve(xpx)

> xpxinv
      [,1] [,2] [,3]
[1,] 14.75 -3.5   -5
[2,] -3.50  1.0    1
[3,] -5.00  1.0    2
```

```
> xpy=t(x)%*%y

> xpy
     [,1]
[1,]    7
[2,]   15
[3,]   11

> beta<-xpxinv%*%xpy

> beta
      [,1]
[1,] -4.25
[2,]  1.50
[3,]  2.00

> ypy=t(y)%*%y

> ypy
     [,1]
[1,]   15

> bpxpxb=t(beta)%*%t(x)%*%x%*%beta

> bpxpxb
      [,1]
[1,] 14.75

> epe=ypy-bpxpxb

> epe
     [,1]
[1,] 0.25

> sigusq=epe/(n-3)
```

```
> sigusq
     [,1]
[1,] 0.25

> rsq<-(bpxpxb-n*mean(y)^2)/(ypy-n*mean(y)^2)

> rsq
          [,1]
[1,] 0.9090909

> varcov<-0.25*xpxinv

> varcov
        [,1]   [,2]  [,3]
[1,]  3.6875 -0.875 -1.25
[2,] -0.8750  0.250  0.25
[3,] -1.2500  0.250  0.50

> se<-sqrt(diag(varcov))

> se
[1] 1.9202864 0.5000000 0.7071068

> x0<-matrix(c(1,3,2))

> yhat<-t(x0)%*%beta

> x0
     [,1]
[1,]    1
[2,]    3
[3,]    2

> yhat
     [,1]
[1,] 4.25
```

```
> sigisq<-0.25*(1+t(x0)%*%xpxinv%*%x0)

> sigisq
       [,1]
[1,] 0.9375

> sigi<-sqrt(sigisq)

> sigi
          [,1]
[1,] 0.9682458

> yhat+c(-1,1)*qt(.975, 1)*sigi
[1] -8.05273 16.55273

> sigesq<-0.25*(t(x0)%*%xpxinv%*%x0)

> sigesq
       [,1]
[1,] 0.6875

> sige<-sqrt(sigesq)

> sige
          [,1]
[1,] 0.8291562

> yhat+c(-1,1)*qt(.975, 1)*sige
[1] -6.285428 14.785428
```

한편, 선형모형의 회귀분석을 실행하는 lm 함수 및 다른 함수를 이용한 b2-ch3-3.R을 실행해 보면 위와 동일한 결과를 얻을 수 있다.

첫째, 다중회귀모형의 회귀계수, 표준오차, t-값, 확률 등을 보여주고 있으며, 교란항의 표준오차(Residual standard erroe)가 0.5, R^2가 0.9091임을 보여주고 있다.

둘째, 이 결과를 이용하여 가설검정도 할 수 있는데 먼저 개별 회귀계수에 대한 검정(단일검정)의 회귀계수별로 t-값 및 확률을 보면 된다. X_2의 추정치는 1.5, 표준오차는 0.5, t-값은 3.0, 확률은 0.205인데 확률이 0.05보다 크므로 5% 유의수준에서 귀무가설 $H_0 : \beta_2 = 0$ 를 허용한다. 또한 X_3의 추정치는 2.0, 표준오차는 0.7071, t-값은 2.828, 확률은 0.216인데 확률이 0.05보다 크므로 5% 유의수준에서 귀무가설 $H_0 : \beta_3 = 0$ 를 허용한다. 상수항을 포함한 독립변수가 3개인 다중회귀모형에서는 결합검정이 곧 전체검정과 동일하다. 전체검정의 귀무가설 $H_0 : \beta_2 = \beta_3 = 0 (q = 2)$ 에 대한 계산된 F-통계량의 값이 5로써 분자의 자유도가 2, 분모의 자유도가 1인 F-분포에 따르는데, p-value가 0.3015로써 0.05보다 크므로 5% 유의수준에서 귀무가설을 허용한다.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.2500     1.9203  -2.213    0.270
x2            1.5000     0.5000   3.000    0.205
x3            2.0000     0.7071   2.828    0.216

Residual standard error: 0.5 on 1 degrees of freedom
Multiple R-squared:  0.9091,     Adjusted R-squared:  0.7273
F-statistic:     5 on 2 and 1 DF,  p-value: 0.3015
```

셋째, 점예측치는 개별예측치 및 평균예측치는 모두 4.25이다. 한편, 평균예측의 95% 예측구간은 [-6.285428, 14.78543]이고, 개별예측의 95% 예측구간은 [-8.05273, 16.55273]임을 확인할 수 있다.

```
> pred.w.clim
   fit       lwr      upr
1 4.25 -6.285428 14.78543

> pred.w.plim
   fit      lwr      upr
1 4.25 -8.05273 16.55273
```

b2-ch3-3.R의 실행결과

```
> x2<-c(1,2,3,2)

> x3<-c(2,1,1,2)

> y<-c(1,1,2,3)

> ols<-lm(y~x2+x3)

> summary(ols)

Call:
lm(formula = y ~ x2 + x3)

Residuals:
    1     2     3     4
-0.25  0.25 -0.25  0.25

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.2500     1.9203  -2.213    0.270
x2            1.5000     0.5000   3.000    0.205
x3            2.0000     0.7071   2.828    0.216

Residual standard error: 0.5 on 1 degrees of freedom
Multiple R-squared:  0.9091,	Adjusted R-squared:  0.7273
F-statistic:     5 on 2 and 1 DF,  p-value: 0.3015


> confint(ols)
                 2.5 %    97.5 %
(Intercept) -28.649553 20.149553
x2           -4.853102  7.853102
x3           -6.984644 10.984644

> resid(ols)
    1     2     3     4
```

```
-0.25   0.25  -0.25   0.25

> predict(lm(y~x2+x3))
   1     2     3     4
1.25  0.75  2.25  2.75

> new<-data.frame(x2=3, x3=2)

> predict(lm(y~x2+x3), new, se.fit = TRUE)
$`fit`
   1
4.25

$se.fit
[1] 0.8291562

$df
[1] 1

$residual.scale
[1] 0.5


> pred.w.clim<-predict(lm(y~x2+x3), new, interval = "confidence")

> pred.w.clim
   fit       lwr        upr
1  4.25  -6.285428  14.78543

> pred.w.plim<-predict(lm(y~x2+x3), new, interval = "prediction")

> pred.w.plim
   fit      lwr       upr
1  4.25  -8.05273  16.55273
```

(예제 3-2 : 결합검정)
다음의 화폐수요함수에서 제1금융권금리 및 제2금융권금리가 동시에 화폐수요에 영향을 주지 않는다는 귀무가설을 검정하라.

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

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.

먼저, 화폐수요함수를 추정하는 위 모형은 종속변수 m_t의 시차변수인 m_{t-1}가 독립변수로 포함되어 있는 동태모형이므로 추정에 앞서 데이터의 범위를 명확하게 해주어야 한다. m_t의 시차변수인 m_{t-1}는 m_t에 대해 한 단위 시간을 지연(lag)시킨 것이므로 m_t 데이터 수가 2001년부터 2009년까지 9개이면 m_{t-1}의 데이터 수는 1개가 줄어든 8개가 된다. 또한 화폐수요함수를 추정할 때는 종속변수 및 모든 독립변수의 데이터 수가 동일해야 하므로 추정에서는 2002년부터 2009년까지 8개의 데이터 포인트를 활용한다.

연도	m_t	Y_t	rb_t	rc_t	m_{t-1}
2001	739418.2	6514153	7.70	5.68	-
2002	824174.4	7205390	6.70	5.78	739418.2
2003	888951.1	7671137	6.24	4.55	824174.4
2004	929726.9	8268927	5.90	4.11	888951.1
2005	994139.6	8652409	5.59	4.27	929726.9
2006	1076775.9	9087438	5.99	4.83	994139.6
2007	1196994.4	9750130	6.55	5.23	1076775.9
2008	1367834.7	10264518	7.17	5.27	1196994.4
2009	1508848.5	10630591	5.65	4.04	1367834.7

이상의 작업은 다음의 명령어를 통해 수행할 수 있다.

```
> nm2<-m2[2:n]

> ny<-y[2:n]

> nrb<-rb[2:n]

> nrc<-rc[2:n]

> lagm2<-m2[1:n-1]
```

화폐수요함수에서 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는다는 귀무가설을 검정하기 위하여 b2-ch3-4.R을 실행해 보면 다음과 같은 결과를 확인할 수 있다.

첫째, 2002년부터 2009년까지 자료를 이용하여 제약이 가해지지 않은 다음의 모형을 추정하면 결정계수(R_{ur}^2)는 0.9992319이다.

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

한편, 귀무가설이 사실일 경우 즉 제약이 가해진 다음의 모형을 추정하면 결정계수(R_r^2)는 0.9879142이다.

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

```
> ur.lm<-lm(nm2~ny+lagm2+nrb+nrc)

> ur.r2<-summary(ur.lm)$r.squared

> ur.r2
[1] 0.9992319

> r.lm<-lm(nm2~ny+lagm2)
```

```
> r.r2<-summary(r.lm)$r.squared

> r.r2
[1] 0.9879142
```

둘째, 계산된 F-통계량은 22.10315이고 분자의 자유도가 2이고 분모의 자유도가 3인 F-분포에서 F-통계량 22.10315의 유의확률은 0.016022072이므로 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는다는 귀무가설을 기각한다. 또는 22.10315는 분자의 자유도가 2이고 분모의 자유도가 3인 F-분포표의 임계치인 9.55보다 크므로 5% 유의수준 하에서 귀무가설을 기각한다. 즉, 화폐수요함수에서 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는 것은 아니므로 금리에 대한 두 변수 중 하나는 반드시 포함되어야 한다.

$$F = \frac{(R_{ur}^2 - R_r^2)/2}{(1 - R_{ur}^2)/n-k} = \frac{(0.9992319 - 0.9879142)/2}{(1-0.9992319)/3} = 22.10315 \sim F_3^2$$

```
> Fstat1<-((ur.r2-r.r2)/2)/((1-ur.r2)/(m-k))

> Fstat1
[1] 22.10315

> pval<-1-pf(Fstat1, 2, m-k)

> pval
[1] 0.01602072
```

셋째, linearHypothesis 함수를 이용하여 화폐수요함수에서 제1금융권 금리(rb) 및 제2금융권 금리(rc)가 동시에 화폐수요에 영향을 주지 않는 지를 검정해 보면 계산된 F-통계량의 값은 22.103이고 p-value 0.01602로써 위와 동일한 결과를 주고 있음을 확인할 수 있다.

```
> jointHo<-c("nrb","nrc")

> linearHypothesis(ur.lm, jointHo)
Linear hypothesis test

Hypothesis:
nrb = 0
nrc = 0

Model 1: restricted model
Model 2: nm2 ~ ny + lagm2 + nrb + nrc

  Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
1      5 4950843985
2      3  314630331  2 4636213653 22.103 0.01602 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b2-ch3-4.R의 실행결과

```
> library(car)

> library(stargazer)

> sample1<-("http://kanggc.iptime.org/book/data/exer.txt")

> sample1_dat<-read.delim(sample1,header=T)

> sample1_dat
        m2         y   rb   rc
1  739418.2  6514153 7.70 5.68
2  824174.4  7205390 6.70 5.78
3  888951.1  7671137 6.24 4.55
4  929726.9  8268927 5.90 4.11
5  994139.6  8652409 5.59 4.27
```

```
6 1076775.9   9087438 5.99 4.83
7 1196994.4   9750130 6.55 5.23
8 1367834.7 10264518 7.17 5.27
9 1508848.5 10630591 5.65 4.04

> m2<-ts(sample1_dat$m2, start=c(2001), frequency=1)

> y<-ts(sample1_dat$y, start=c(2001), frequency=1)

> rb<-ts(sample1_dat$rb, start=c(2001), frequency=1)

> rc<-ts(sample1_dat$rc, start=c(2001), frequency=1)

> n<-length(m2)

> k=5

> nm2<-m2[2:n]

> ny<-y[2:n]

> nrb<-rb[2:n]

> nrc<-rc[2:n]

> lagm2<-m2[1:n-1]

> m<-length(nm2)

> ur.lm<-lm(nm2~ny+lagm2+nrb+nrc)

> ur.r2<-summary(ur.lm)$r.squared

> ur.r2
[1] 0.9992319

> r.lm<-lm(nm2~ny+lagm2)
```

```
> r.r2<-summary(r.lm)$r.squared

> r.r2
[1] 0.9879142

> stargazer(ur.lm, r.lm, type="text")

================================================
                      Dependent variable:

-----------------------------------------------
                              nm2
                      (1)                (2)
-------------------------------------------------------
ny                   -0.016              0.017
                     (0.015)            (0.044)

lagm2                 1.288***           1.073***
                     (0.091)            (0.259)

nrb                 24,792.470
                   (13,961.440)

nrc                 24,292.350
                   (12,815.680)

Constant          -316,050.500**     -125,310.700
                   (59,281.170)      (149,042.500)

-------------------------------------------------------
Observations            8                  8
R2                    0.999              0.988
Adjusted R2           0.998              0.983
Residual Std. Error  10,240.940 (df = 3)   31,466.950 (df = 5)
F Statistic        975.731*** (df = 4; 3) 204.354*** (df = 2; 5)
================================================
```

```
Note:                                    *p<0.1; **p<0.05; ***p<0.01

> Fstat1<-((ur.r2-r.r2)/2)/((1-ur.r2)/(m-k))

> Fstat1
[1] 22.10315

> pval<-1-pf(Fstat1, 2, m-k)

> pval
[1] 0.01602072

> jointHo<-c("nrb","nrc")

> linearHypothesis(ur.lm, jointHo)
Linear hypothesis test

Hypothesis:
nrb = 0
nrc = 0

Model 1: restricted model
Model 2: nm2 ~ ny + lagm2 + nrb + nrc

  Res.Df        RSS Df  Sum of Sq      F   Pr(>F)
1      5 4950843985
2      3  314630331  2 4636213653 22.103 0.01602 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

제 4 장

가변수

1. 가변수모형
2. 가변수모형의 유형

제4장 가변수

1. 가변수모형

계량분석에 사용되는 대부분의 자료는 정량적인 자료이다. 그러나 필요에 따라서는 정성적인 자료를 사용할 경우가 있는데 정성변수를 독립변수로 사용하는 경우와 정성변수를 종속변수로 사용하는 경우가 있다. 본 장에서 다루고자 하는 가변수모형은 정성변수를 독립변수로 사용하는 경우이다.

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 있다.

가변수(dummy variable)란 독립변수 중의 일부가 성질을 달리하는 질적인 자료로 되어 있을 경우 사용된다. 계량분석에서 질적인 면을 나타내 주는 변수인 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 OLS 추정량은 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 다음과 같이 소비모형을 보통최소자승법으로 추정하면 되는데 이를 보통회귀식이라고 하자.

$$C_t = \beta_0 + \beta_1 Y_t + u_t$$

단, C_t는 소비, Y_t는 소득을 나타낸다.

IMF 경제위기를 경험하면서 경제주체의 소비행태에 변화가 있었는 지를 살펴보기 위해 다음과 같이 가변수를 모형에 포함시킬 수 있다.

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + u_t$$

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

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 다음 식의 임금결정모형을 보통최소자승법으로 추정하면 될 것이다.

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

단, Y_t는 임금, X_t는 근무연수를 나타낸다.

근무연수에 따른 임금결정 외에 성별 및 인종별 임금격차가 존재하는 지를 살펴보기 위해 다음과 같이 가변수를 모형에 포함시킬 수 있다.

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

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 임금 격차가 존재하는 지를 살펴보기 위해 다음과 같이 가변수를 모형에 포함시킬 수 있다.

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

단, $D_{Hi} = 0, D_{Ci} = 0,\ i$가 중졸 이하

$D_{Hi} = 1, D_{Ci} = 0,\ i$가 고졸

$D_{Hi} = 0, D_{Ci} = 1,\ i$가 대졸 이상

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 가변수모형(가변수모형 1)은 다음과 같이 나타낼 수 있다.

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t + u_t$$

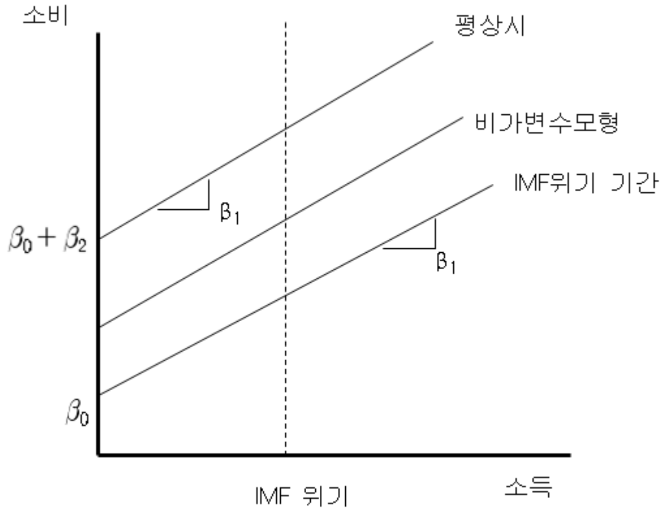
단, $D_t = \begin{cases} 0, & t \text{ 가 } IMF\text{위기 기간} \\ 1, & t \text{ 가 평상시} \end{cases}$

위의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 가변수모형의 추정은 다음의 평상시 소비함수 및 IMF위기 기간 소비함수를 개별적으로 추정한 결과와 동일하다.

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

(평상시) $C_t = (\beta_0 + \beta_2) + \beta_1 Y_t + u_t$

가변수에 대한 가설검정에서 귀무가설은 $H_0 : \beta_2 = 0$으로 평상시와 IMF위기 기간의 소비수준에는 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, β_2가 통계적으로 유의하면) <그림 4-1>과 같이 평상시와 IMF위기 기간의 소비수준에 차이가 있다고 결론을 내린다. 따라서 이 경우는 평상시와 IMF위기 기간의 소비함수를 각각 추정해야 하는데 가변수를 포함한 모형을 추정하면 이와 동일한 결과를 얻을 수 있다.



<그림 4-1> 절편의 변화를 나타내는 가변수

한편, 회귀계수에 대한 해석은 다음과 같다. $\widehat{\beta_1}$은 한계소비성향으로 평상시의 한계소비성향과 IMF위기 기간의 한계소비성향과 같다. $\widehat{\beta_0}$은 IMF위기 기간 절대소비수준을 나타내며, $\widehat{\beta_2}$는 평상시 소비수준과 IMF위기 기간 소비수준의 차이를 나타낸다. 따라서 평상시 소비수준은 $\widehat{\beta_0} + \widehat{\beta_2}$이 된다.

(2) 기울기(한계)의 변화를 나타내는 가변수모형

기울기(한계)의 변화를 나타내는 가변수모형(가변수모형 2)은 다음과 같이 나타낼 수 있다.

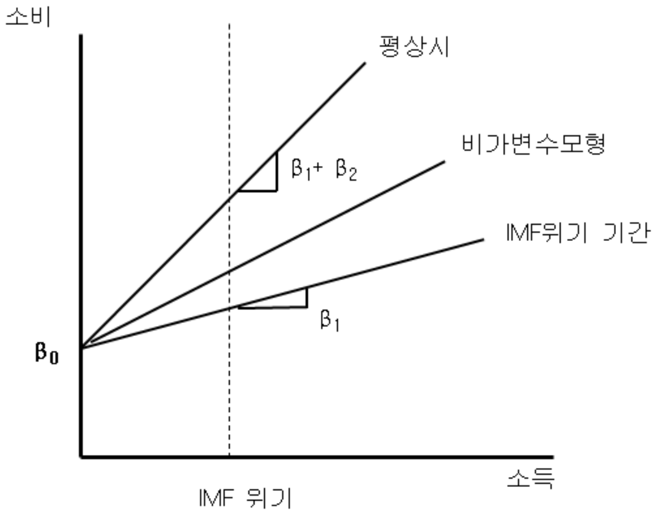
$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 D_t Y_t + u_t$$

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

위식의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 가변수모형의 추정은 다음의 평상시 소비함수 및 IMF위기 기간 소비함수를 개별적으로 추정한 결과와 동일하다.

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

(평상시) $C_t = \beta_0 + (\beta_1 + \beta_2) Y_t + u_t$



<그림 4-2> 기울기의 변화를 나타내는 가변수

가변수에 대한 가설검정에서 귀무가설은 $H_0 : \beta_2 = 0$으로 평상시와 IMF위기 기간의 한계소비성향에 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, β_2가 통계적으로 유의하면) <그림 4-2>와 같이 평상시와 IMF위기 기간의 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 평상시와 IMF위기 기간의 소비함수를 각각 추정해야 한다. 가변수를 포함한 가변수모형을 추정하면 이와 동일한 결과를 얻을 수 있다.

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

(3) 절편과 기울기의 동시변화를 나타내는 가변수모형

절편과 기울기의 동시변화를 나타내는 가변수모형(가변수모형 3)은 다음과 같이 나타낼 수 있다.

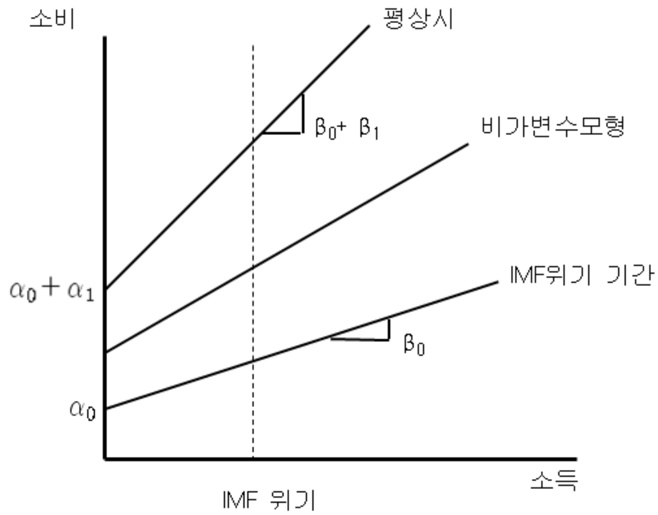
$$C_t = \alpha_0 + \alpha_1 D_t + \beta_0 Y_t + \beta_1 D_t Y_t + u_t$$

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

위식의 가변수모형을 추정한 후 가변수가 통계적으로 유의할 경우 가변수모형의 추정은 다음의 평상시 소비함수 및 IMF위기 기간 소비함수를 개별적으로 추정한 결과와 동일하다.

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

(평상시) $C_t = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1) Y_t + u_t$



<그림 4-3> 절편과 기울기의 변화를 나타내는 가변수

위식의 가변수모형에서는 3종류의 가설검정이 가능하다.

첫째, $H_0:\alpha_1=0$으로 평상시와 IMF위기 기간의 소비수준에는 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, α_1이 통계적으로 유의하면) 평상시와 IMF위기 기간의 소비수준에 차이가 있다고 결론을 내린다. 따라서 이 경우는 평상시와 IMF위기 기간의 소비함수를 각각 추정해야 하는데 가변수모형을 추정하면 이와 동일한 결과를 얻을 수 있다.

둘째, $H_0:\beta_1=0$으로 평상시와 IMF위기 기간의 한계소비성향에 차이가 없다는 가설이다. 귀무가설을 기각하면(즉, β_1이 통계적으로 유의하면) 평상시와 IMF위기 기간의 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 평상시와 IMF위기 기간의 소비함수를 각각 추정해야 한다. 가변수모형을 추정하면 이와 동일한 결과를 얻을 수 있다.

셋째, $H_0:\alpha_1=\beta_1=0$으로 평상시와 IMF위기 기간의 소비수준 및 한계소비성향에 차이가 없다는 귀무가설을 검정하고 결론을 내린다. 귀무가설을 기각하면(즉, α_1과 β_1이 동시에 통계적으로 유의하면) <그림 4-3>과 같이 평상시와 IMF위기 기간의 소비수준 및 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 평상시와 IMF위기 기간의 소비함수를 각각 추정해야 한다. 가변수모형을 추정하면 이와 동일한 결과를 얻을 수 있다.

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

지금까지 살펴본 3가지 유형의 가변수모형에 대한 가설검정의 결과에 따라 결정되는 모형에 대한 논의를 요약해 보면 <표 4-1>과 같다.

<표 4-1> 가설검정 결과에 따른 결정모형의 요약

구분	1단계 (2개 가변수에 대한 단일검정)	2단계 (2개 가변수에 대한 결합검정)	3단계 (2개 회귀식의 가변수에 대한 단일검정)	4단계
case 1	$\alpha_1 = 0$ 기각 & $\beta_1 = 0$ 허용	가변수모형 1		
case 2	$\alpha_1 = 0$ 허용 & $\beta_1 = 0$ 기각	가변수모형 2		
case 3	$\alpha_1 = 0$ 기각 & $\beta_1 = 0$ 기각	가변수모형 3		
case 4	$\alpha_1 = 0$ 허용 & $\beta_1 = 0$ 허용	$\alpha_1 = \beta_1 = 0$ 허용	가변수가 없는 모형	
case 5	$\alpha_1 = 0$ 허용 & $\beta_1 = 0$ 허용	$\alpha_1 = \beta_1 = 0$ 기각	$\alpha_1 = 0$ 기각 & $\beta_1 = 0$ 허용	가변수모형 1
case 6	$\alpha_1 = 0$ 허용 & $\beta_1 = 0$ 허용	$\alpha_1 = \beta_1 = 0$ 기각	$\alpha_1 = 0$ 허용 & $\beta_1 = 0$ 기각	가변수모형 2
case 7	$\alpha_1 = 0$ 허용 & $\beta_1 = 0$ 허용	$\alpha_1 = \beta_1 = 0$ 기각	$\alpha_1 = 0$ 기가 & $\beta_1 = 0$ 기각	가변수모형 1, 가변수모형 2 중 결정계수가 큰 모형

(예제 4-1)
1997년 4/4분기에 발생한 외환위기로 인해 한국경제의 소비행태에 구조적인 변화가 있었는 지를 알아보고자 한다. 1995년 1/4분기부터 2001년 1/4분기까지 분기별 자료를 이용하여 소비함수를 추정하는데 다음과 같이 가변수를 설정한다고 하자.

D = 1 (1995년 1/4분기 - 1997년 4/4분기: 평상시)
D = 0 (1998년 1/4분기 - 2001년 1/4분기 : IMF 위기 기간)

(예제 4-1)에서 지금까지 설명한 3가지 가변수모형의 추정과 가설검정을 수행하는 b2-ch4-1.R을 실행한 결과를 단계별로 설명해 보자.

첫째, 가변수를 만들기 위하여 명령어를 이용하여 추세변수, 추세변수를 이용한 가변수 및 가변수와 독립변수를 곱한 새로운 변수를 만든다.

```
> n=length(c)

> tr=1:n

> d.log<-tr <=  12

> d<-as.numeric(d.log)

> dy<-d*y
```

이를 실행하면 Y, C, D, DY의 자료구조는 다음과 같게 된다.

분기	Y	C	D	DY
1995:1	93061.99	58431.8	1	93062.99
1995:2	93613.55	59455.3	1	93614.55
1995:3	94788.82	61165.7	1	94789.82
1995:4	95894.36	63787.3	1	95895.36
1996:1	99094.99	62845.6	1	99095.99
1996:2	99903.92	63957.8	1	99904.92
1996:3	100951.85	64929.8	1	100952.85
1996:4	102886.05	68684.3	1	102887.05
1997:1	103604.5	65445.2	1	103605.5
1997:2	106036.68	66575.1	1	106037.68
1997:3	106458.19	68103.8	1	106459.19
1997:4	106926.21	68598.5	1	106927.21
1998:1	99153.69	58780.6	0	99153.69
1998:2	97622.45	58902.6	0	97622.45
1998:3	97887.44	60811.6	0	97887.44

분기	Y	C	D	DY
1998:4	100077.5	63193.2	0	100077.5
1999:1	104465.5	62768.3	0	104465.5
1999:2	108484.4	64474.1	0	108484.4
1999:3	110529.11	67192.6	0	110529.11
1999:4	114271.82	70045.3	0	114271.82
2000:1	116666.09	68711.6	0	116666.09
2000:2	118947.45	69481.3	0	118947.45
2000:3	120696.31	70508.4	0	120696.31
2000:4	119995.03	72155.9	0	119995.03
2001:1	120633.92	68975.7	0	120633.92

둘째, 위의 자료구조에서 C를 종속변수, D 및 Y를 독립변수로 설정한 가변수모형 1의 추정 및 가설검정 결과는 다음과 같다.

추정된 회귀식은 $\hat{C} = 9854 + 3419D + 0.5092Y$ 이다. 가변수는 통계적으로 유의한 것으로 나타나 평상시 절대소비수준은 IMF 위기 기간 절대소비수준보다 3419만큼 높은 것으로 해석할 수 있다.

```
Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  9.854e+03   4.432e+03    2.223    0.0368 *
d            3.419e+03   6.979e+02    4.898  6.74e-05 ***
y            5.092e-01   4.014e-02   12.685  1.36e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

셋째, 위의 자료구조에서 C를 종속변수, Y 및 DY를 독립변수로 설정한 가변수모형 2의 추정 및 가설검정 결과는 다음과 같다.

추정된 회귀식은 $\hat{C} = 10433 + 0.5038Y + 0.003398DY$ 이다. 가변수는 통계적으로 유의한 것으로 나타나 평상시 한계소비성향은 IMF 위기 기간 한계소비성향보다 0.003398만큼 높은 것으로 해석할 수 있다.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.043e+04  4.237e+03   2.463   0.0221 *
y           5.038e-01  3.845e-02  13.103 7.21e-12 ***
dy          3.398e-02  6.652e-03   5.108 4.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

넷째, 위의 자료구조에서 C를 종속변수, D, Y 및 DY를 독립변수로 설정한 가변수모형 3의 추정 및 가설검정 결과는 다음과 같다.

추정된 회귀식은 $\hat{C} = 13476 - 12172D + 0.4763Y + 0.1523DY$ 이다. 두 개의 가변수는 개별적으로는 통계적으로 유의하지 않은 것으로 나타나 이에 대한 추가적인 가설검정이 필요하다.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.348e+04  4.827e+03   2.792   0.0109 *
d           -1.217e+04  9.657e+03  -1.261   0.2213
y            4.763e-01  4.376e-02  10.883 4.31e-10 ***
dy           1.523e-01  9.411e-02   1.619   0.1205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

다섯째, 두 개의 가변수는 동시에 통계적으로 유의하지 않다는 귀무가설에 대한 결합검정의 결과 계산된 F-통계량 14.189의 p-value가 0.00012625로써 5% 유의수준에서 귀무가설을 기각한다. 따라서 가변수모형 1 또는 가변수모형 2가 실제 데이터를 잘 나타내 주는 모형이라고 할 수 있는데 이 경우가 <표 4-1>에서 case 7에 해당된다. 가변수모형 1의 R^2는 0.884이고, 가변수모형 2의 R^2는 0.889이므로 가변수모형 2를 이용하는 것이 바람직하다.

```
> jointHo<-c("d","dy")

> linearHypothesis(m3.lm, jointHo)
Linear hypothesis test

Hypothesis:
d = 0
dy = 0

Model 1: restricted model
Model 2: c ~ d + y + dy

  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     23 96393877
2     21 40995058  2  55398818 14.189 0.0001262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

여섯째, 만약에 두 개의 가변수가 모두 통계적으로 유의하다면 평상시와 IMF위기 기간의 소비수준 및 한계소비성향에 차이가 있다고 결론을 내린다. 따라서 이 경우는 평상시와 IMF위기 기간의 소비함수를 각각 추정해야 한다.

평상시 소비함수를 추정하면 $\hat{C}=1303.246+0.629Y$ 이고, IMF위기 기간 소비함수를 추정하면 $\hat{C}=13476.09+0.476Y$ 인데, 이것은 다음과 같이 추정된 가변수모형 3에서 도출할 수 있음을 확인할 수 있다.

$$\hat{C}=13476-12172D+0.4763Y+0.1523DY$$

```
Regression Results of separate period
================================================
                        Dependent variable:

------------------------------------------
                                  c
                      (1)                (2)
```

```
--------------------------------------------------------------
y                        0.629***              0.476***
                         (0.080)               (0.045)

Constant                 1,303.246             13,476.090**
                         (8,060.937)           (4,980.879)

--------------------------------------------------------------
Observations             12                    13
R2                       0.860                 0.910
Adjusted R2              0.846                 0.902
Residual Std. Error  1,346.670 (df = 10)    1,441.586 (df = 11)
F Statistic        61.279*** (df = 1; 10) 111.266*** (df = 1; 11)
==============================================
Note:                              *p<0.1; **p<0.05; ***p<0.01
```

b2-ch4-1.R의 실행결과

```
> library(car)

> library(stargazer)

> data<-read.table("http://kanggc.iptime.org/book/data/dummy.txt", header=T)

> y<-ts(data$GDP, start=c(1995,1), frequency=4)

> c<-ts(data$CONSUME, start=c(1995,1), frequency=4)

> n=length(c)

> tr=1:n

> d.log<-tr <=  12

> d<-as.numeric(d.log)
```

```
> dy<-d*y

> y;c;d;dy
          Qtr1      Qtr2      Qtr3      Qtr4
1995  93061.99  93613.55  94788.82  95894.36
1996  99094.99  99903.92 100951.85 102886.05
1997 103604.50 106036.68 106458.19 106926.21
1998  99153.69  97622.45  97887.44 100077.50
1999 104465.50 108484.40 110529.11 114271.82
2000 116666.09 118947.45 120696.31 119995.03
2001 120633.92
        Qtr1    Qtr2    Qtr3    Qtr4
1995 58431.8 59455.3 61165.7 63787.3
1996 62845.6 63957.8 64929.8 68684.3
1997 65445.2 66575.1 68103.8 68598.5
1998 58780.6 58902.6 60811.6 63193.2
1999 62768.3 64474.1 67192.6 70045.3
2000 68711.6 69481.3 70508.4 72155.9
2001 68975.7
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
          Qtr1      Qtr2      Qtr3      Qtr4
1995  93061.99  93613.55  94788.82  95894.36
1996  99094.99  99903.92 100951.85 102886.05
1997 103604.50 106036.68 106458.19 106926.21
1998      0.00      0.00      0.00      0.00
1999      0.00      0.00      0.00      0.00
2000      0.00      0.00      0.00      0.00
2001      0.00

> m1.lm<-lm(c~d+y)

> summary(m1.lm)

Call:
lm(formula = c ~ d + y)

Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2307.8   -806.9   -375.7   1054.6   3019.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.854e+03  4.432e+03   2.223   0.0368 *
d           3.419e+03  6.979e+02   4.898 6.74e-05 ***
y           5.092e-01  4.014e-02  12.685 1.36e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1448 on 22 degrees of freedom
Multiple R-squared:  0.8841,     Adjusted R-squared:  0.8735
F-statistic: 83.87 on 2 and 22 DF,  p-value: 5.092e-11


> m2.lm<-lm(c~y+dy)

> summary(m2.lm)

Call:
lm(formula = c ~ y + dy)

Residuals:
    Min       1Q   Median       3Q      Max
-2230.1   -874.9   -292.3   1064.8   2923.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.043e+04  4.237e+03   2.463   0.0221 *
y           5.038e-01  3.845e-02  13.103 7.21e-12 ***
dy          3.398e-02  6.652e-03   5.108 4.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1416 on 22 degrees of freedom
Multiple R-squared:  0.8891,     Adjusted R-squared:  0.879
```

```
F-statistic:  88.2 on 2 and 22 DF,  p-value: 3.117e-11


> m3.lm<-lm(c~d+y+dy)

> summary(m3.lm)

Call:
lm(formula = c ~ d + y + dy)

Residuals:
    Min       1Q   Median       3Q      Max
-1956.2   -748.5   -330.4    713.5   2707.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.348e+04  4.827e+03   2.792   0.0109 *
d           -1.217e+04  9.657e+03  -1.261   0.2213
y            4.763e-01  4.376e-02  10.883 4.31e-10 ***
dy           1.523e-01  9.411e-02   1.619   0.1205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1397 on 21 degrees of freedom
Multiple R-squared:  0.8969,     Adjusted R-squared:  0.8822
F-statistic:  60.9 on 3 and 21 DF,  p-value: 1.568e-10


> stargazer(m1.lm, m2.lm, m3.lm, type="text", title="Regression Results of using Dummy Variable")

Regression Results of using Dummy Variable
============================================================
                              Dependent variable:

--------------------------------------------------------
                                      c
                        (1)          (2)          (3)
------------------------------------------------------------
```

```
d                        3,418.648***                              -12,172.840
                          (697.938)                                (9,656.625)

y                          0.509***            0.504***             0.476***
                           (0.040)             (0.038)              (0.044)

dy                                             0.034***              0.152
                                               (0.007)              (0.094)

Constant                 9,854.362**         10,433.910**         13,476.090**
                         (4,432.226)         (4,236.569)          (4,827.492)

--------------------------------------------------------------------------------
Observations                 25                  25                   25
R2                          0.884               0.889                0.897
Adjusted R2                 0.874               0.879                0.882
Residual Std. Error  1,447.710 (df = 22)   1,415.773 (df = 22)   1,397.192 (df = 21)
F Statistic          83.870*** (df = 2; 22) 88.199*** (df = 2; 22) 60.903*** (df = 3; 21)
==========================================================
Note:                                              *p<0.1; **p<0.05; ***p<0.01

> jointHo<-c("d","dy")

> linearHypothesis(m3.lm, jointHo)
Linear hypothesis test

Hypothesis:
d = 0
dy = 0

Model 1: restricted model
Model 2: c ~ d + y + dy

  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     23 96393877
2     21 40995058  2  55398818 14.189 0.0001262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> m4.lm<-lm(c~y, data=data, subset=(d==1))
```

```
> summary(m4.lm)

Call:
lm(formula = c ~ y, data = data, subset = (d == 1))

Residuals:
    Min      1Q  Median      3Q     Max
-1382.5  -807.2  -131.8   196.0  2707.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1303.2462  8060.9371   0.162    0.875
y              0.6286     0.0803   7.828 1.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1347 on 10 degrees of freedom
Multiple R-squared:  0.8597,     Adjusted R-squared:  0.8457
F-statistic: 61.28 on 1 and 10 DF,  p-value: 1.424e-05


> m5.lm<-lm(c~y, data=data, subset=(d==0))

> summary(m5.lm)

Call:
lm(formula = c ~ y, data = data, subset = (d == 0))

Residuals:
    Min      1Q  Median      3Q     Max
-1956.2  -671.1  -453.2  1073.5  2143.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.348e+04  4.981e+03   2.706   0.0205 *
y           4.763e-01  4.515e-02  10.548  4.33e-07 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1442 on 11 degrees of freedom
Multiple R-squared:   0.91,        Adjusted R-squared:  0.9019
F-statistic: 111.3 on 1 and 11 DF,  p-value: 4.326e-07


> stargazer(m4.lm, m5.lm, type="text", title="Regression Results of separate
period")

Regression Results of separate period
==================================================
                              Dependent variable:

---------------------------------------
                                       c
                          (1)                     (2)
--------------------------------------------------------------
y                       0.629***                0.476***
                        (0.080)                 (0.045)

Constant               1,303.246              13,476.090**
                      (8,060.937)             (4,980.879)

--------------------------------------------------------------
Observations              12                      13
R2                       0.860                   0.910
Adjusted R2              0.846                   0.902
Residual Std. Error  1,346.670 (df = 10)     1,441.586 (df = 11)
F Statistic        61.279*** (df = 1; 10) 111.266*** (df = 1; 11)
==================================================
Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

(예제 4-2)
임금결정모형에서 근무연수(연령)에 따라 임금이 변하는데 성별 및 학력에 따라 임금 격차가 존재하는 지를 살펴보고자 한다. 85명의 개인별 연령(age), 교육정도(ed), 성별(gender), 월급여(income : 단위 백만 원)에 관한 자료를 이용하여 임금결정함수를 추정하는데 다음과 같이 가변수를 설정한다고 하자.

(임금결정함수) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1i} + \beta_3 D_{Hi} + \beta_4 D_{Ci} + u_i$
(성별 가변수) $D_{1i} = 0$(i가 여자), $D_{1i} = 1$(i가 남자)
(학력 가변수) $D_{Hi} = 0, D_{Ci} = 0$, i가 중졸 이하
$D_{Hi} = 1, D_{Ci} = 0$, i가 고졸
$D_{Hi} = 0, D_{Ci} = 1$, i가 대졸 이상

(예제 4-2)에서 설명한 임금결정함수 추정을 수행하는 b2-ch4-2.R을 실행한 결과를 이용하여 설명해 보자.

첫째, 교육정도에 근거하여 가변수를 만드는 명령에 따라 생성된 D_H 및 D_C 를 살펴보면 다음과 같다.

```
> ed;high;college
 [1] 3 0 5 3 2 2 3 3 3 6 3 2 3 3 3 3 4 3 3 5 1 2 5 3 3 5 3 4 5
[30] 3 3 5 4 4 2 3 4 3 5 4 5 5 3 3 5 3 3 3 5 5 5 3 3 5 3 2 1 5
[59] 5 3 3 5 3 3 3 3 1 5 5 3 2 3 4 5 5 5 3 3 3 3 3 5 3 3 3
 [1] 1 0 0 1 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 0 0 0 0 1 1 0 1 0 0
[30] 1 1 0 0 0 0 1 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0
[59] 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 0 0 0 1 1 1 1 1 0 1 1 1
 [1] 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 1
[30] 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 0 0 1 1 1 0 0 1 0 0 0 1
[59] 1 0 0 1 0 0 0 0 0 1 1 0 0 0 1 1 1 1 0 0 0 0 0 1 0 0 0
```

둘째, 추정된 임금결정함수는 다음과 같다. 연령에 대한 회귀계수의 추정치는 0.05198이므로 성별 및 학력에 관계없이 근무연수(연령)가 1년 증가하면 월급여는 0.05198백만원 즉, 5.198만원 증가하는 것으로 나타났다.

$$\hat{Y} = -1.36335 + 0.05198age + 0.65852gender + 0.38935D_H + 0.98185D_C$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.36335    0.59573  -2.289 0.024745 *
age          0.05198    0.01388   3.744 0.000340 ***
gender       0.65832    0.20868   3.155 0.002264 **
high         0.38935    0.23919   1.628 0.107500
college      0.98185    0.24109   4.073 0.000109 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

셋째, 월급여의 차이를 살펴보기 위해 추정결과를 정리해 보면 다음의 표와 같고 이에 대한 해석(통계적 유의성은 고려하지 않는다고 가정)은 다음과 같다.

성별은 나타내는 가변수의 값이 0, 학력을 나타내는 가변수의 값이 0인 경우를 비교의 기준으로 삼는 것이 일반적인데 이를 reference(비교기준)라고 한다. (예제 4-2)에서는 중졸 여자가 reference가 된다.

연령을 통제한 상태에서 즉, 동일한 연령에서 중졸 남자는 중졸 여자에 비해 월급여는 평균 0.705백만원 즉, 7.05십만 원 높은 것으로 나타났다. 또한 고졸은 남녀에 관계없이 중졸에 비해 평균 0.389백만원 즉, 3.89십만 원 높은 것으로 나타났다. 한편, 대졸 이상은 남녀에 관계없이 중졸에 비해 평균 0.982백만원 즉, 9.82십만 원 높은 것으로 나타났고, 남녀에 관계없이 고졸에 비해 평균 0.593백만원 즉, 5.93십만 원 높은 것으로 나타났다.

구분	여자	남자
중졸 이하	-1.363(**reference**)	(-1.363+0.658)=-0.705
고졸	(-1,363+0.389)	(-0.705+0.389)
대졸 이상	(-1,363+0.982)	(-0.705+0.982)

b2-ch4-2.R의 실행결과
> library(stargazer) > data<-read.table("http://kanggc.iptime.org/book/data/income.txt", header=T)

```
> age<-data$age

> ed<- data$ed

> gender<-data$gender

> income<-data$income

> age;ed;gender;income
 [1] 23 38 31 23 35 43 39 36 44 39 34 38 31 36 34 41 36 33 36
[20] 28 44 34 32 36 44 36 36 39 38 32 31 42 30 25 42 35 23 25
[39] 35 32 30 41 32 28 30 32 27 33 41 40 39 33 31 45 27 39 41
[58] 38 39 29 41 45 34 28 38 32 37 38 28 33 41 33 39 35 41 29
[77] 37 40 35 29 30 32 24 33 32
 [1] 3 0 5 3 2 2 3 3 3 6 3 2 3 3 3 3 4 3 3 5 1 2 5 3 3 5 3 4 5
[30] 3 3 5 4 4 2 3 4 3 5 4 5 5 3 3 5 3 3 3 5 5 5 3 3 5 3 2 1 5
[59] 5 3 3 5 3 3 3 3 1 5 5 3 2 3 4 5 5 5 3 3 3 3 3 5 3 3 3
 [1] 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 0 1
[30] 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1
[59] 1 0 0 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
 [1] 0.5500002 0.6500002 2.6872395 0.7749998 0.8000000
 [6] 1.6576906 2.3999993 1.4000003 0.5700002 2.3950005
[11] 1.1737602 1.1999999 1.1245044 1.2599996 1.2249995
[16] 1.5000005 1.3499998 1.7355553 1.8799994 2.4620010
[21] 2.1399992 1.2400001 1.9141898 2.0139993 0.6599999
[26] 1.1616698 0.8000000 0.2200000 3.0319999 1.4289993
[31] 1.8196996 2.7128383 1.4763096 1.4640000 0.9373950
[36] 1.1428995 0.8250003 0.2849999 0.9200000 1.0327751
[41] 1.1811298 3.5106789 1.5515143 1.0886000 1.8506195
[46] 1.5000005 0.6000000 0.7020002 1.7999996 2.7567500
[51] 1.7499994 1.4900005 2.9757986 3.2949811 1.7999996
[56] 1.0699998 0.9170002 1.1698195 3.9499995 1.5000005
[61] 0.9220005 3.1166803 1.4283893 1.6117498 1.4866800
[66] 1.4765547 0.4500001 3.6245257 0.8000000 1.1310003
[71] 1.0040004 0.6000000 1.7969997 2.4669990 3.2282250
[76] 2.6814118 1.5799994 2.2599992 1.0800004 1.3762996
[81] 1.7999996 1.6014804 0.6599999 1.3103502 0.6099999
```

```
> high<-ifelse(data$ed==3, 1, 0)

> college<-ifelse(data$ed>3, 1, 0)

> ed;high;college
 [1] 3 0 5 3 2 2 3 3 3 6 3 2 3 3 3 3 4 3 3 5 1 2 5 3 3 5 3 4 5
[30] 3 3 5 4 4 2 3 4 3 5 4 5 5 3 3 5 3 3 3 5 5 5 3 3 5 3 2 1 5
[59] 5 3 3 5 3 3 3 3 1 5 5 3 2 3 4 5 5 5 3 3 3 3 3 5 3 3 3
 [1] 1 0 0 1 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 0 0 0 0 1 1 0 1 0 0
[30] 1 1 0 0 0 0 1 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 1 0 1 0 0 0
[59] 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 0 0 0 1 1 1 1 1 0 1 1 1
 [1] 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 1
[30] 0 0 1 1 1 0 0 1 0 1 1 1 1 0 0 1 0 0 0 1 1 1 0 0 1 0 0 0 1
[59] 1 0 0 1 0 0 0 0 0 1 1 0 0 0 1 1 1 1 0 0 0 0 0 1 0 0 0

> m1.lm<-lm(income~age+gender+high+college)

> summary(m1.lm)

Call:
lm(formula = income ~ age + gender + high + college)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42566 -0.42209 -0.02595  0.40079  1.68014

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.36335    0.59573  -2.289 0.024745 *
age          0.05198    0.01388   3.744 0.000340 ***
gender       0.65832    0.20868   3.155 0.002264 **
high         0.38935    0.23919   1.628 0.107500
college      0.98185    0.24109   4.073 0.000109 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6458 on 80 degrees of freedom
Multiple R-squared:  0.4086,	Adjusted R-squared:  0.3791
F-statistic: 13.82 on 4 and 80 DF,  p-value: 1.299e-08


> stargazer(m1.lm, type="text", title="Regression Results of using Dummy
Variable")

Regression Results of using Dummy Variable
================================================
                        Dependent variable:
                    ---------------------------
                              income
-----------------------------------------------
age                          0.052***
                             (0.014)

gender                       0.658***
                             (0.209)

high                          0.389
                             (0.239)

college                      0.982***
                             (0.241)

Constant                     -1.363**
                             (0.596)

-----------------------------------------------
Observations                    85
R2                             0.409
Adjusted R2                    0.379
Residual Std. Error       0.646 (df = 80)
F Statistic            13.820*** (df = 4; 80)
================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

제 5 장

자기상관

1. 일반화최소자승법
2. 자기상관 개념 및 유형
3. 자기상관 검정
4. 자기상관 추정

제5장 자기상관

1. 일반화최소자승법

단순회귀모형 및 다중회귀모형에서 기본적인 가정들이 성립하는 고전적 회귀모형의 경우 보통최소자승법(OLS)으로 추정한 추정량은 최량선형불편추정량(BLUE)이 됨을 설명한 바 있다.

여러 가지 기본적인 가정들 중 다음의 가정은 교란항의 동분산(homoscedasticity)과 비자기상관(no autocorrelation)을 가정한 것이며 이 경우 OLS로 추정한 추정량은 BLUE가 된다.

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

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

$$= \sigma_u^2 I_n$$

그러나 현실적으로 위의 가정이 성립하지 않는 경우가 발생한다. 즉, 교란항의 분산-공분산 행렬 $E(UU') \neq \sigma_u^2 I_n$인 두 경우가 있는데 하나는 자기상관(1차 자기상관)이 있는 경우이고 다른 하나는 이분산인 경우이다.

$$(\text{자기상관})\quad E(UU')=\sigma_u^2\Omega=\sigma_e^2\begin{bmatrix}\frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & \cdot\cdot & \frac{\rho^{n-1}}{1-\rho^2}\\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & \cdot\cdot & \frac{\rho^{n-2}}{1-\rho^2}\\ \cdot & \cdot & \cdot\cdot & \cdot\\ \cdot & \cdot & \cdot\cdot & \cdot\\ \frac{\rho^{n-1}}{1-\rho^2} & \frac{\rho^{n-2}}{1-\rho^2} & \cdot\cdot & \frac{1}{1-\rho^2}\end{bmatrix}\neq\sigma_u^2 I_n$$

$$(\text{이분산})\qquad E(UU')=\sigma_u^2\Omega=\sigma_u^2\begin{bmatrix}k_1^2 & 0 & \cdot\cdot & 0\\ 0 & k_2^2 & \cdot\cdot & 0\\ \cdot & \cdot & \cdot\cdot\cdot & \cdot\\ 0 & 0 & \cdot\cdot & k_n^2\end{bmatrix}\neq\sigma_u^2 I_n$$

위의 경우와 같이 자기상관이나 이분산이 있음에도 불구하고 OLS로 추정하면 추정량은 편의는 없으나 효율적인 추정량이 되지 못한다.

회귀모형의 OLS 추정량은 교란항의 분산-공분산 행렬에 대한 가정과 관계없이 다음과 같으므로 $\hat{B}$은 B의 불편추정량(unbiased estimator)이다.

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

그러나 $\hat{B}$의 분산은 다음과 같으므로 교란항의 분산-공분산 행렬에 대한 가정에 따라 $\hat{B}$의 분산은 달라지게 된다.

$$\begin{aligned}Var(\hat{B})&=E[(\hat{B}-E(\hat{B})(\hat{B}-E(\hat{B}))']\\&=E[(\hat{B}-B)(\hat{B}-B)']\\&=(X'X)^{-1}X'E(UU')X(X'X)^{-1}\end{aligned}$$

위 식에서 고전적 회귀모형의 경우 교란항의 분산-공분산 행렬에 대한 가정이 $E(UU')=\sigma_u^2 I_n$ 이므로 OLS 추정량 $\hat{B}$의 분산은 다음과 같다.

$$\begin{aligned}Var(\hat{B})&=(X'X)^{-1}X'E(UU')X(X'X)^{-1}\\&=\sigma_u^2(X'X)^{-1}X'X(X'X)^{-1}\\&=\sigma_u^2(X'X)^{-1}\end{aligned}$$

한편, 자기상관이나 이분산이 있을 경우 분산-공분산 행렬에 대한 가정이 $E(UU')=\sigma_u^2\Omega$ 이므로 OLS 추정량 $\hat{B}$의 분산은 다음과 같다.

$$\begin{aligned} Var(\widehat{B}) &= (X'X)^{-1}X'E(UU')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma_u^2\Omega X(X'X)^{-1} \\ &= \sigma_u^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned}$$

Gauss-Markov정리에 OLS 추정량 $\hat{B}$의 분산 $Var(\hat{B})=\sigma_u^2(X'X)^{-1}$이 최소분산이므로 분산-공분산 행렬에 대한 가정이 $E(UU')=\sigma_u^2\Omega$ 인 경우의 OLS 추정량 $\hat{B}$의 분산 $Var(\widehat{B})=\sigma_u^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$은 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 있는데 이를 일반화 최소자승법(Generalized Least Squares : GLS) 이라고 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, 다음의 회귀모형이 고전적 회귀모형의 기본가정을 충족시키지 못하고 다음과 같다고 하자.

$$Y=XB+U$$

단, $E(UU')=\sigma^2\Omega$

이 경우 위의 회귀모형에 L을 곱하여 변환하면 다음과 같이 변환된 모형을 구할 수 있다.

$$LY = LXB + LU$$

또는

$$Y^* = X^* B + U^*$$

셋째, 변환된 모형에서 교란항 평균을 구해보면 다음과 같이 0이 된다.

$$\begin{aligned} E(U^*) &= E(LU) \\ &= LE(U) \\ &= 0 \end{aligned}$$

또한 교란항의 분산을 계산해 보면 다음과 같다.

$$\begin{aligned} E(U^* U^{*\prime}) &= E(LUU'L') \\ &= \sigma^2 L\Omega L' \\ &= \sigma^2 LL^{-1}L'^{-1}L' \, (\because L^{-1}L'^{-1} = \Omega \leftarrow L'L = \Omega^{-1}) \\ &= \sigma^2 I_n \end{aligned}$$

변환된 모형의 교란항 평균은 0이고 분산은 자기상관이 되어 있지 않으며 동분산의 조건을 만족하므로 변환된 모형을 최소자승법으로 추정하면 BLUE를 얻을 수 있는데 이러한 방법을 GLS라고 한다.

넷째, GLS 추정량은 변환된 모형을 OLS로 추정하면 구할 수 있는데 다음과 같다.

$$
\begin{aligned}
\widehat{B_G} &= (X^{*'}X^{*})^{-1}X^{*'}Y^{*} \\
&= (X'L'LX)^{-1}X'L'LY \\
&= (X^{-1}L^{-1}L^{-1'}X^{-1'})X'L'LY \\
&= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y
\end{aligned}
$$

또한 교란항의 분산 및 회귀계수의 분산에 대한 추정량은 각각 다음과 같다.

$$\widehat{\sigma_u^2} = \frac{e'\Omega^{-1}e}{n-k}$$

$$Var(\widehat{B_G}) = \widehat{\sigma_u^2}(X'\Omega^{-1}X)^{-1}$$

2. 자기상관의 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.

다음과 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 자기상관(autocorrelation)이 있다고 한다. 특히 이 경우를 1차 자기상관(first-order autocorrelation)이라 한다.

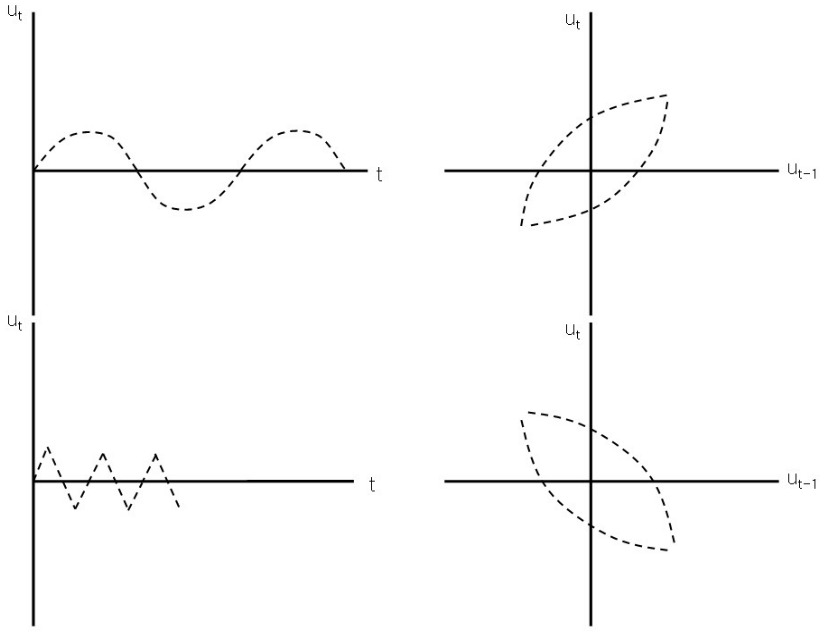
$$Y_t = X_tB + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

자기상관의 유형에는 양의 자기상관(positive autocorrelation) 및 음의 자기상관(negative autocorrelation)이 있으며 그 유형은 <그림 5-1>과 같다.

회귀모형 추정에서 구한 잔차의 시계열 그래프가 <그림 5-1>의 상단 왼쪽과 같은 모양을 보이거나 잔차와 잔차 시차변수의 산포도가 <그림 5-1>의 상단 오른쪽과 같은 모양을 보이면 양의 자기상관이 있는 경우이다.

한편, 회귀모형 추정에서 구한 잔차의 시계열 그래프가 <그림 5-1>의 하단 왼쪽과 같은 모양을 보이거나 잔차와 잔차 시차변수의 산포도가 <그림 5-1>의 하단 오른쪽과 같은 모양을 보이면 음의 자기상관이 있는 경우이다.



<그림 5-1> 양의 자기상관 및 음의 자기상관

3. 자기상관 검정

시계열의 자기상관 여부를 탐지하는 방법으로는 <그림 5-1>과 같이 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 5-1>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 다음과 같은 d-통계량을 이용한다.

$$
\begin{aligned}
d &= \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2} \\
&= \frac{\sum_{t=2}^{N} e_t^2 + \sum_{t=2}^{N} e_{t-1}^2 - 2\sum_{t=2}^{N} e_t e_{t-1}}{\sum_{t=1}^{N} e_t^2} \\
&\fallingdotseq 2(1-\hat{\rho})
\end{aligned}
$$

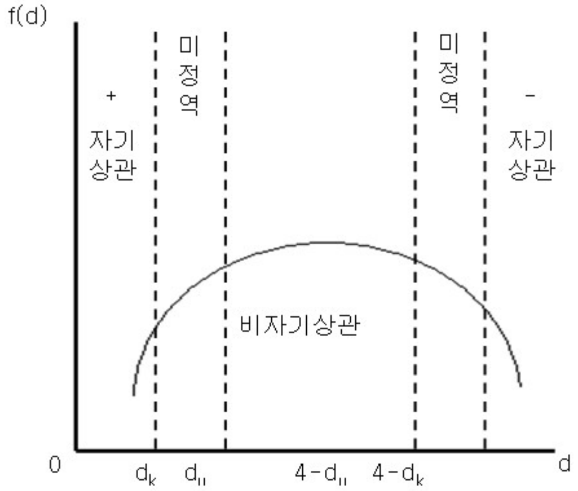
d-통계량을 이용한 자기상관 여부는 다음과 같이 판단한다.

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-검정은 1차 자기상관을 검정하며, 회귀모형은 상수항을 반드시 포함하고 있어야 하고, 회귀모형의 설명변수에는 종속변수의 시차변수가 존재하지 않아야 한다. 즉, 자기회귀모형에는 DW-검정방법을 적용할 수 없다. 또한 이 방법은 분포를 이용한 검정방법이라는 장점은 있으나 <그림 5-2>에 나타나 있듯이 자기상관 여부를 결정할 수 없는 영역(미정역)이 있다는 단점이 있다.



<그림 5-2> 자기상관 검정

자기상관이 있음에도 불구하고 OLS로 추정하면 불편추정량은 얻을 수 있으나 효율적인 추정량은 얻지 못하기 때문에 원자료(original data)를 적당한 방법으로 변환시켜 고전적 회귀모형의 기본가정에 맞도록 한 후 OLS로 추정하면 된다.

그러면 원자료를 어떻게 변환시킬 것인가? 다음과 같이 교란항이 1차 자기상관이 있을 경우 교란항의 분산-공분산행렬은 어떻게 생겼는지를 살펴보자.

$$Y_t = X_t B + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

단, $E(e_t = 0)$. $E(e_t e_t{}') = \sigma_e^2 I_n$, $E(u_{t-1} e_t) = 0$이다.

교란항이 1차 자기상관이 있을 경우 교란항의 분산-공분산 행렬은 다음과 같다.

$$E(UU') = \begin{bmatrix} E(u_1^2) & E(u_1u_2) & .. & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & .. & E(u_2u_n) \\ . & . & .. & . \\ . & . & .. & . \\ E(u_nu_1) & E(u_nu_2) & .. & E(u_n^2) \end{bmatrix}$$

$$= \sigma_e^2 \begin{bmatrix} \frac{1}{1-\rho^2} & \frac{\rho}{1-\rho^2} & .. & \frac{\rho^{n-1}}{1-\rho^2} \\ \frac{\rho}{1-\rho^2} & \frac{1}{1-\rho^2} & .. & \frac{\rho^{n-2}}{1-\rho^2} \\ . & . & .. & . \\ . & . & .. & . \\ \frac{\rho^{n-1}}{1-\rho^2} & \frac{\rho^{n-2}}{1-\rho^2} & .. & \frac{1}{1-\rho^2} \end{bmatrix}$$

$$= \sigma_e^2 \Omega$$

위의 Ω로부터 다음의 Ω의 역행렬인 Ω^{-1}을 구할 수 있고 이는 $\Omega\Omega^{-1} = I$ 임을 통해 확인할 수 있다.

$$\Omega^{-1} = \begin{bmatrix} 1 & -\rho & 0 & . & . & 0 \\ -\rho & 1+\rho^2 & -\rho & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & 1+\rho^2 & -\rho \\ 0 & 0 & . & . & -\rho & 1 \end{bmatrix}$$

위의 Ω^{-1}로부터 다음의 L행렬을 구할 수 있고 이는 $L'L = \Omega^{-1}$임을 통해 확인할 수 있다.

$$L = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & . & . & 0 \\ -\rho & 1 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ . & . & . & . & 1 & 0 \\ 0 & 0 & . & . & -\rho & 1 \end{bmatrix}$$

L행렬을 이용하여 원래자료를 변환시키면 종속변수는 다음과 같이 변환되는데 이를 Paris-Winstern 변환이라 한다.

$$y_1^* = \sqrt{1-\rho^2}\, y_1$$
$$y_t^* = y_t - \rho y_{t-1} (t \geqq 2)$$

$$x_1^* = \sqrt{1-\rho^2}\, x_1$$
$$x_t^* = x_t - \rho x_{t-1} (t \geqq 2)$$

한편, 자기상관을 검정하는 또 다른 방법으로 LM 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 1차 자기상관을 가정한 회귀모형에 대한 LM-검정은 다음과 같은 순서로 한다.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

제1단계로 위의 회귀모형을 추정한 후 다음과 같은 잔차를 구한다.

$$\widehat{u_t} = Y_t - \widehat{\beta_1} - \widehat{\beta_2} X_{2t} - \widehat{\beta_3} X_{3t}$$

2단계로 위 식에서 구한 잔차를 1차 자기상관을 가정한 회귀모형에 포함되어 있는 설명변수와 $\hat{u}_{t-1}$에 대해 다음의 회귀모형을 추정하는데 이를 보조회귀식(auxiliary regression)이라고 한다.

$$\widehat{u_t} = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \rho \hat{u}_{t-1} + v_t$$

이 경우 검정하고자 하는 귀무가설은 $H_o : \rho = 0$ 즉, 자기상관이 없다는 것이며 귀무가설 하에서 χ_1^2에 따르는 다음의 LM 검정통계량을 이용하여 의사결정을 할 수 있다.

$$LM = nR^2 \sim \chi^2_1$$

단, n은 관측치의 수, R^2는 보조회귀식에서의 결정계수이며 χ^2-분포의 자유도는 1이다.

4. 자기상관 추정

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.

다음의 1차 자기상관을 가정한 회귀모형 Ⓐ 식에서 원자료를 변환시키는 방법을 살펴보면 다음과 같다.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad Ⓐ$$

$$u_t = \rho u_{t-1} + e_t$$

먼저 Ⓐ 식을 한 기간 이전의 식으로 표현하면 다음과 같다.

$$Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + u_{t-1}$$

위 식의 양변에 ρ를 곱한 후 Ⓐ 식에서 빼면 다음과 같게 되는데 이를 Cochrane-Orcutt 변환(transformation)이라고 한다.

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{2t} - \rho X_{2t-1}) + \beta_3(X_{3t} - \rho X_{3t-1}) + e_t \qquad Ⓑ$$

$$e_t = u_t - \rho u_{t-1}$$

위 $\beta_1^* = \beta_1(1-\rho)$라 표기하면 다음과 같다.

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \epsilon_t \qquad \text{Ⓒ}$$

단, $Y_t^* = Y_t - \rho Y_{t-1},\ X_{2t}^* = (X_{2t} - \rho X_{2t-1}),\quad X_{3t}^* = (X_{3t} - \rho X_{3t-1})$이다.

Ⓒ 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 OLS로 추정하면 효율적인 추정량을 얻을 수 있는데 이를 GLS 추정량이라고 한다.

다음으로 ρ의 값이 알려져 있지 않을 경우 ρ을 추정한 후 자료를 변환하고 OLS를 적용하는데 ρ값을 추정하는 방법에는 여러 가지가 있다.

(1) Durbin의 2단계 추정법

1단계는 Ⓑ 식을 다시 정리하면 다음과 같은데 이를 OLS로 추정하여 $\hat{\rho}$를 얻는다.

$$Y_t = \rho Y_{t-1} + \beta_1^* + \beta_2 X_{2t} - \beta_2 \rho X_{2t-1} + \beta_3 X_{3t} - \beta_3 \rho X_{3t-1} + e_t$$

2단계는 1단계에서 구한 $\hat{\rho}$으로 $Y_t^*, X_{2t}^*, X_{3t}^*$를 구한 후 Ⓒ 식을 OLS로 추정하면 되는데 이렇게 구한 추정량을 GLS 추정량이라고 하고 BLUE가 된다.

(2) Cochrane-Orcutt의 반복절차

1단계는 Ⓐ 식을 OLS로 추정하여 잔차 $\widehat{u_t}$를 구한다.

2단계는 1단계에서 추정된 잔차를 이용하여 다음 식을 OLS로 추정하여 $\hat{\rho}$을 구하는데 이를 ρ의 1단계 추정치라고 한다.

$$\widehat{u_t} = \rho \hat{u}_{t-1} + e_t$$

3단계는 2단계에서 추정된 $\hat{\rho}$을 이용하여 $Y_t^*, X_{2t}^*, X_{3t}^*$를 구한 후 Ⓒ 식을 OLS로 추정하여 $\widehat{\beta_1^*}, \widehat{\beta_2^*}, \widehat{\beta_3^*}$을 구한다.

4단계는 $\widehat{\beta_1^*}, \widehat{\beta_2^*}, \widehat{\beta_3^*}$을 이용하여 다음의 새로운 잔차를 구한다.

$$\widehat{u_t^*} = Y_t - \widehat{\beta_1^*} - \widehat{\beta_2^*} X_{2t} - \widehat{\beta_3^*} X_{3t}$$

5단계는 새로운 잔차를 이용하여 다음 식을 OLS로 추정하여 새로운 1차 자기상관계수($\hat{\hat{\rho}}$)를 구하는데 이를 ρ의 2단계 추정치라고 한다.

$$\widehat{u_t^*} = \rho \hat{u}_{t-1}^*$$

6단계는 위의 과정을 다음의 수렴조건이 충족될 때까지 반복한다.

$$|\ \hat{\rho}_j - \hat{\rho}_{j+1}\ | < 0.005$$

단, 0.005는 아주 작은 수의 예시로 이를 수렴한계치(convergence limit)라고 한다.

(예제 5) 1995년 1/4분기부터 2001년 1/4분기까지 한국의 GDP와 총소비(consume) 자료를 이용하여 소비함수를 추정하기 위해 자기상관 여부를 검정하라.

(예제 5)에서 자기상관 여부를 판단하기 위해 b2-ch5-1.R을 실행한 결과를 설명해 보자.

첫째, 소비함수를 OLS로 추정한 후 구한 잔차($\hat{u}_t$)의 시계열 그래프 및 잔차($\hat{u}_t$)와 잔차 시차변수($\hat{u}_{t-1}$)의 산포도를 그려보면 자기상관이 존재할 가능성이 있음을 보여주고 있으나 그래프는 객관적인 판단이 될 수 없음을 알 수 있다

둘째, DW-검정을 이용하여 1차 자기상관 여부를 판단해 보자. (예제 5)는 n=25, 상수항을 제외한 독립변수의 수인 $k'=1$이다. 이를 <부록 2>의 <표 7>에서 찾아보면 $d_L=1.288$이고, $d_U=1.454$인데 DW 통계량 1.2273이 $d_L=1.288$보다 작으므로 <그림 5-2>에서 나타내고 있는 것처럼 DW-검정 결과 양(+)의 1차 자기상관이 있는 것으로 나타났다. 한편, DW 통계량 1.2273의 p-value가 0.01178로 0.05보다 작으므로 5% 유의수준에서 귀무가설 $H_0: \rho=0$ 즉, 자기상관이 없다는 귀무가설을 기각하고, 대립가설 $H_1: \rho>0$ 즉, 양(+)의 1차 자기상관이 있다는 대립가설을 허용한다.

```
        Durbin-Watson test

data:  ols.res
DW = 1.2273, p-value = 0.01178
alternative hypothesis: true autocorrelation is greater than 0
```

셋째, LM-검정법을 이용하여 자기상관 여부를 판단하기 위해서 계산한 LM 통계량 $LM=nR^2$의 값이 3.122012이고, χ^2_1-분포에서 계산된 LM 통계량의 유의확률이 0.07724137이므로 10% 유의수준에서 1차 자기상관이 있는 것으로 결정한다.

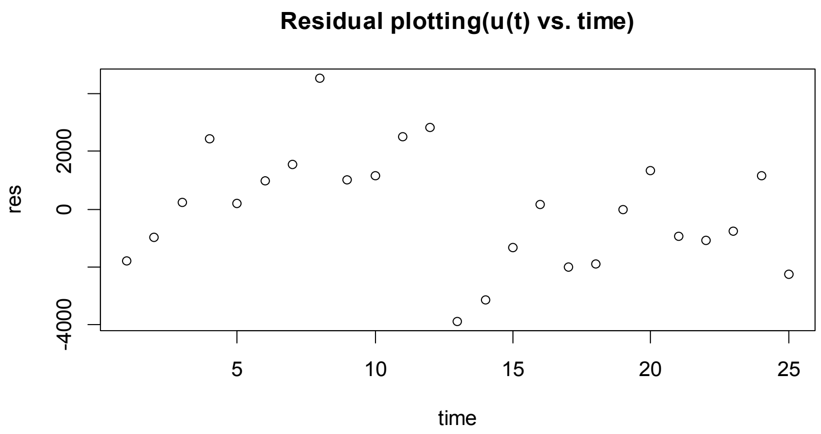
```
> lm<-n*summary(res.aux)$r.squared

> lm
[1] 3.122012

> (pchisq(lm, df=1, lower.tail=F))
[1] 0.07724137
```

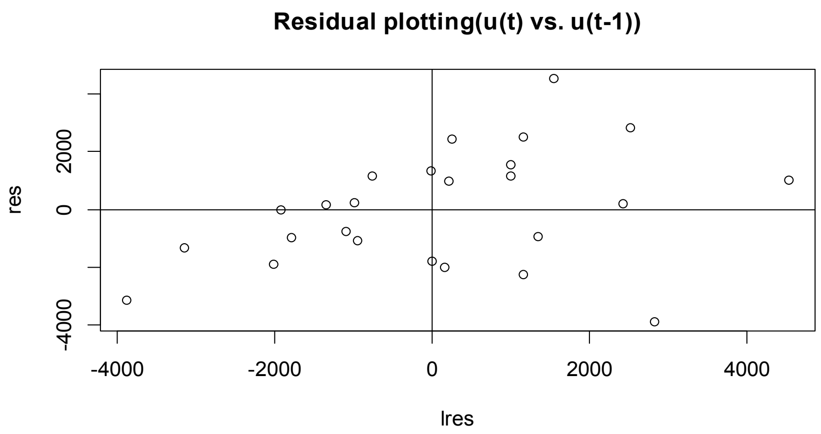
b2-ch5-1.R의 실행결과

```
> library(stargazer)

> library(lmtest)

> library(orcutt)

> sample1<-("http://kanggc.iptime.org/book/data/ar.txt")

> sample1_dat<-read.delim(sample1,header=T)

> consume<-ts(sample1_dat$consume, start=c(1995.1), frequency=4)

> gdp<-ts(sample1_dat$gdp, start=c(1995.1), frequency=4)

> ols.res<-lm(consume~gdp)

> summary(ols.res)

Call:
lm(formula = consume ~ gdp)

Residuals:
    Min      1Q  Median      3Q     Max
-3879.8 -1342.8   163.6  1166.1  4532.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.303e+04  4.981e+03   4.624 0.000119 ***
gdp         3.997e-01  4.714e-02   8.479 1.56e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2047 on 23 degrees of freedom
Multiple R-squared:  0.7576,    Adjusted R-squared:  0.7471
F-statistic: 71.89 on 1 and 23 DF,  p-value: 1.561e-08
```

```
> res<-resid(ols.res)

> lres<-append(res[1:24], 0, after=0)

> res.t<-ts(res)

> lres.t<-ts(append(res.t[1:24], 0, after=0))

> plot(res,main="Residual plotting(u(t) vs. time)",xlab='time')

> plot(lres,res,main="Residual plotting(u(t) vs. u(t-1))")

> abline(h=0, v=0)

> #DW test
> dwtest(ols.res)

        Durbin-Watson test

data:  ols.res
DW = 1.2273, p-value = 0.01178
alternative hypothesis: true autocorrelation is greater than 0


> #LM test
> n<-length(res)

> lres.t<-ts(append(res.t[1:24], 0, after=0))

> (res.aux<-lm(res.t~gdp+lres.t))

Call:
lm(formula = res.t ~ gdp + lres.t)

Coefficients:
(Intercept)          gdp        lres.t
```

```
 404.108292      -0.004151       0.364122


> lm<-n*summary(res.aux)$r.squared

> lm
[1] 3.122012

> (pchisq(lm, df=1, lower.tail=F))
[1] 0.07724137
```





DW-검정과 LM-검정 결과 모두 소비함수에 1차 자기상관이 있는 것으로 나타났으므로 회귀계수 및 1차 자기상관계수를 동시에 추정하는 b2-ch5-2.R을 실행한 결과를 이용하여 설명해 보자.

첫째, OLS로 추정한 한계소비성향은 0.3997이다.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.303e+04  4.981e+03   4.624 0.000119 ***
gdp         3.997e-01  4.714e-02   8.479 1.56e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

둘째, Cochrane-Orcutt의 반복절차에 의해 추정된 한계소비성향은 0.392648로써 이것이 GLS 추정치이며 OLS 추정치와 비교해 보면 그 값이 다름을 알 수 있다. 한편, 5번의 반복에 의해 최종적으로 추정된 1차 자기상관계수는 0.357123이다.

셋째, 원래 모형인 Ⓐ 식의 DW 통계량은 1.22725이고 p-value는 0.01178로 0.05보다 작으므로 5% 유의수준에서 자기상관이 없다는 귀무가설을 기각하지만, Cochrane-Orcutt 변환 모형인 Ⓒ 식의 DW 통계량은 1.8354이고 p-value는 0.2657로 0.05보다 크므로 5% 유의수준에서 자기상관이 없다는 귀무가설을 허용하여 자기상관 문제가 해결되었음을 확인할 수 있다.

```
 number of interaction: 5
 rho 0.357123

Durbin-Watson statistic
(original):      1.22725 , p-value: 1.178e-02
(transformed): 1.83540 , p-value: 2.657e-01

 coefficients:
 (Intercept)          gdp
23842.831304      0.392648
```

넷째, 원래자료를 변환한 후 변환된 자료로 OLS로 추정하면 그 추정량이 GLS 추정량임을 설명한 바 있다. $\hat{\rho}=0.357123$ 이므로 이를 이용하여 다음과 같이 자료 변환을 하고, 변환된 자료로 소비함수를 OLS로 추정하면 이 추정치는 변환된 자료의 OLS 추정치 또는 원래 자료의 GLS 추정치인데 두 한계소비성향이 동일함을 알 수 있다.

```
tconsume<-consume[2:25]-0.357123*consume[1:24]
tgdp<-gdp[2:25]-0.357123*gdp[1:24]
```

```
===============================================
                      Dependent variable:
                  -----------------------------
                     consume        tconsume
                       (1)             (2)
-----------------------------------------------
gdp                  0.393***
                     (0.069)

tgdp                                0.393***
                                    (0.069)

Constant          23,842.830***   15,328.010***
                   (7,336.524)     (4,716.484)

-----------------------------------------------
Observations            25              24
R2                                    0.598
Adjusted R2                           0.579
Residual Std. Error            1,920.237 (df = 22)
F Statistic                  32.686*** (df = 1; 22)
===============================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

b2-ch5-2.R의 실행결과

```
> library(stargazer)

> library(lmtest)

> library(orcutt)

> sample1<-("http://kanggc.iptime.org/book/data/ar.txt")

> sample1_dat<-read.delim(sample1,header=T)

> consume<-ts(sample1_dat$consume, start=c(1995.1), frequency=4)

> gdp<-ts(sample1_dat$gdp, start=c(1995.1), frequency=4)

> ols.res<-lm(consume~gdp)

> summary(ols.res)

Call:
lm(formula = consume ~ gdp)

Residuals:
    Min      1Q  Median      3Q     Max 
-3879.8 -1342.8   163.6  1166.1  4532.2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.303e+04  4.981e+03   4.624 0.000119 ***
gdp         3.997e-01  4.714e-02   8.479 1.56e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2047 on 23 degrees of freedom
Multiple R-squared:  0.7576,	Adjusted R-squared:  0.7471 
F-statistic: 71.89 on 1 and 23 DF,  p-value: 1.561e-08
```

```
> res<-resid(ols.res)

> res.t<-ts(res)

> #Cochrane-Orcutt estimation
> coch.res<-cochrane.orcutt(ols.res)

> coch.res
Cochrane-orcutt estimation for first order autocorrelation

Call:
lm(formula = consume ~ gdp)

 number of interaction: 5
 rho 0.357123

Durbin-Watson statistic
(original):      1.22725 , p-value: 1.178e-02
(transformed): 1.83540 , p-value: 2.657e-01

 coefficients:
 (Intercept)           gdp
23842.831304      0.392648

> #GLS estimation
> tconsume<-consume[2:25]-0.357123*consume[1:24]

> tgdp<-gdp[2:25]-0.357123*gdp[1:24]

> gls.res<-lm(tconsume~tgdp)

> summary(gls.res)

Call:
lm(formula = tconsume ~ tgdp)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4984.4  -852.1   107.4  1186.4  3926.2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.533e+04  4.716e+03   3.250  0.00367 ** 
tgdp        3.926e-01  6.868e-02   5.717 9.46e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1920 on 22 degrees of freedom
Multiple R-squared:  0.5977,	Adjusted R-squared:  0.5794 
F-statistic: 32.69 on 1 and 22 DF,  p-value: 9.464e-06


> stargazer(coch.res, gls.res, type="text")

=============================================
                     Dependent variable:     
                 ----------------------------
                    consume         tconsume 
                      (1)             (2)    
---------------------------------------------
gdp                 0.393***                 
                    (0.069)                  

tgdp                               0.393***  
                                   (0.069)   

Constant         23,842.830***  15,328.010***
                  (7,336.524)    (4,716.484) 

---------------------------------------------
Observations          25              24     
R2                                  0.598    
Adjusted R2                         0.579    
```

```
Residual Std. Error              1,920.237 (df = 22)
F Statistic                      32.686*** (df = 1; 22)
==========================================================
Note:                        *p<0.1; **p<0.05; ***p<0.01
```

제 6 장

이분산

1. 이분산 개념 및 유형
2. 이분산 검정
3. 이분산 추정

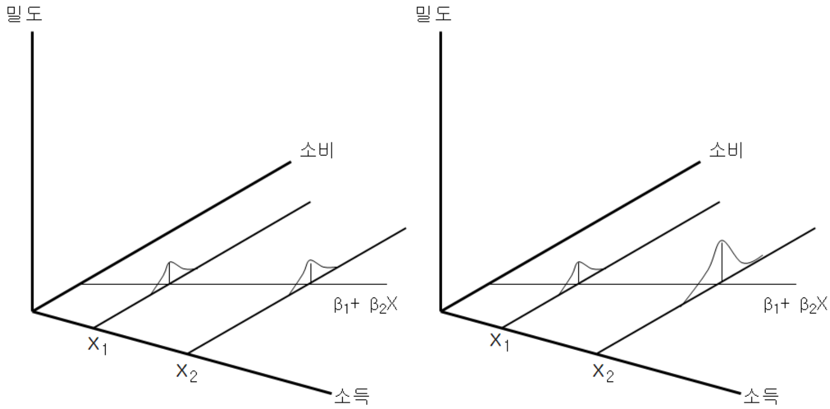
제6장 이분산

1. 이분산 개념 및 유형

고전적 회귀모형에서는 교란항에 대한 동분산을 가정하고 있는데 이것은 <그림 6-1>의 왼쪽 그림과 같이 X의 모든 관측치에 대해 교란항의 확률분포가 일정하다는 것을 의미한다. 반면에 이분산이 있다는 것은 <그림 6-1>의 오른쪽 그림과 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.

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 부합한다고 할 수 있다.

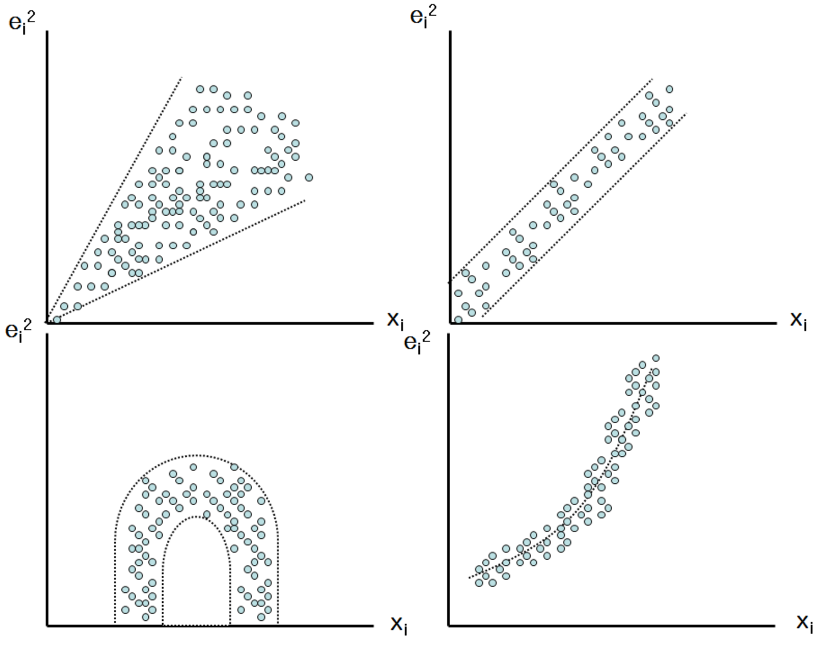
일반적으로 거시자료보다 미시자료에, 시계열 자료보다 횡단면 자료에 이분산이 존재할 가능성이 크다. 왜냐하면 횡단면 자료는 다양한 규모의 가계나 기업 등에 관한 자료가 포함되어 있기 때문이다. 그러나 시계열 자료라고 해서 항상 동분산만 있는 것은 아니며 한 경제단위의 장기에 걸친 시계열 자료의 경우 이분산이 나타날 가능성이 있다.



〈그림 6-1〉 교란항의 분포(동분산 및 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 독립변수 X_i를 그려 보고 〈그림 6-2〉와 같이 이들 사이에 일정한 관계가 발견되면 이분산이 있는 것으로 판단한다.



〈그림 6-2〉 이분산의 유형

2. 이분산 검정

이분산이 있을 경우 일반적으로 교란항 분산의 크기는 설명변수와 관련이 있다. 이러한 측면을 고려한 이분산 검정방법은 White 검정방법, Goldfeld-Quant 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, 다음 식에 대한 White 검정은 다음과 같은 순서로 한다.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

먼저, 위 모형을 추정한 후 다음과 같이 잔차를 구한다.

$$\widehat{u_i} = Y_i - \widehat{\beta_1} - \widehat{\beta_2} X_{2i} - \widehat{\beta_3} X_{3i}$$

다음으로, 잔차의 제곱을 이용하여 다음 회귀모형을 추정하는데 이를 보조회귀식(auxiliary regression)이라고 한다.

$$\widehat{u_i^2} = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + e_i$$

위 보조회귀식에서 검정하고자 하는 귀무가설은 $H_0 : \alpha_2 = ... = \alpha_6 = 0$ 즉, 이분산이 없다는 것이며 귀무가설 하에서 χ_q^2에 따르는 LM 검정통계량을 이용하여 의사결정을 할 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 검정방법이다. 예를 들어 이분산이 다음과 같은 관계에 있다고 하자.

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

$$\sigma_i^2 = \sigma^2 X_i^2$$

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(소분산 집단)로 회귀식을 OLS로 추정하여 $\sum e_1^2$를 구하고 값이 큰 쪽의 관측치(대분산 집단)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. 이분산 추정

이분산 문제를 해결하는 방법은 이분산의 구조에 따라 달라진다. 예를 들어, 회귀모형과 교란항의 분산이 다음과 같다고 하자.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$E(u_i^2) = \sigma^2 X_{2i}$$

원래의 회귀식을 변환하여 이분산이 없는 모형을 도출한다.

$$\frac{Y_i}{\sqrt{X_{2i}}} = \beta_1 \frac{1}{\sqrt{X_{2i}}} + \beta_2 \sqrt{X_{2i}} + \beta_3 \frac{X_{3i}}{\sqrt{X_{2i}}} + \frac{u_i}{\sqrt{X_{2i}}}$$

또는

$$Y_i^* = \beta_1 \frac{1}{\sqrt{X_{2i}}} + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + u_i^*$$

이와 같이 변환된 회귀모형(transformed regression model)에서 Y_i^*를 $\frac{1}{\sqrt{X_i}}$, X_{2i}^*, X_{3i}^*에 대해 OLS로 추정하면 효율적인 추정량을 얻을 수 있는데 이를 GLS 추정량이라고 한다. 이러한 추정방법은 일반최소자승법인데 이를 특별히 가중최소자승법(Weighted Least Squares: WLS)이라고 한다.

(예제 6)
1977년 1/4분기부터 1997년 2/4분기까지의 총통화(m2), 국민총생산(gnp), gnp 디플레이터(def)의 자료를 이용하여 총통화는 gnp와 def에 영향을 받는다는 화폐수요함수를 추정하기 위해 이분산 여부를 검정하라.

(예제 6)에서 이분산 여부를 판단하기 위해 b2-ch6-1.R을 실행한 결과를 설명해 보자.

첫째, 화폐수요함수를 OLS로 추정한 후 구한 잔차($\hat{u}_t$)의 시계열 그래프를 이분산이 존재할 가능성이 있음을 보여주고 있으나 그래프는 객관적인 판단이 될 수 없음을 알 수 있다

둘째, White 검정법을 이용하여 자기상관 여부를 판단하기 위해서 계산한 LM 통계량 $LM = nR^2$의 값이 19.65366이고, χ^2_5-분포에서 계산된 LM 통계량의 유의확률이 0.0014512이므로 5% 유의수준에서 이분산이 없다는 귀무가설을 기각하여 이분산이 있는 것으로 결정한다.

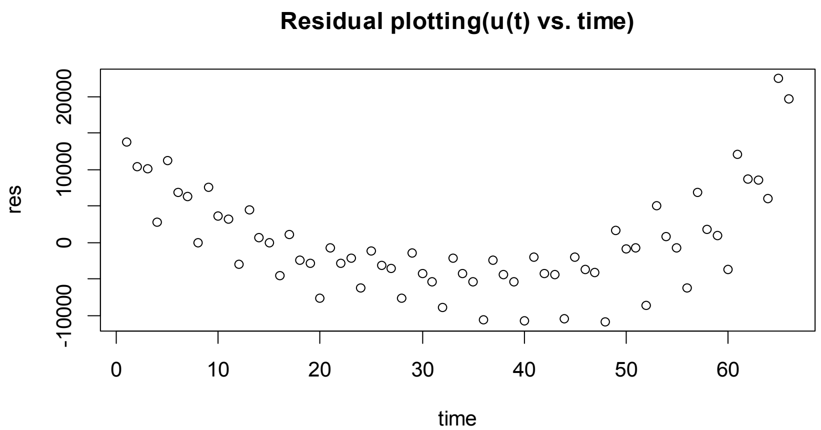
```
> lm<-n*summary(res.aux)$r.squared

> lm
[1] 19.65366

> (pchisq(lm, df=5, lower.tail=F))
[1] 0.0014512
```

b2-ch6-1.R의 실행결과

```
> library(stargazer)

> sample1<-("http://kanggc.iptime.org/book/data/korea(77-93)-1.txt")

> sample1_dat<-read.delim(sample1,header=T)

> gnp<-ts(sample1_dat$gnp, start=c(1977,1), frequency=4)

> def<-ts(sample1_dat$def, start=c(1977,1), frequency=4)

> m2<-ts(sample1_dat$m2, start=c(1977,1), frequency=4)

> ols.res<-lm(m2~gnp+def)
```

```
> summary(ols.res)

Call:
lm(formula = m2 ~ gnp + def)

Residuals:
    Min      1Q  Median      3Q     Max
-10838   -4271   -1723    3469   22428

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.560e+04  2.778e+03 -12.814  < 2e-16 ***
gnp          1.332e+00  2.229e-01   5.975 1.17e-07 ***
def          4.126e+02  6.157e+01   6.702 6.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7074 on 63 degrees of freedom
Multiple R-squared:  0.9305,     Adjusted R-squared:  0.9283
F-statistic: 421.6 on 2 and 63 DF,  p-value: < 2.2e-16


> res<-resid(ols.res)

> res.t<-ts(res)

> plot(res,main="Residual plotting(u(t) vs. time)",xlab='time')

> #LM test
> n<-length(res)

> res.t.sq<-res.t^2

> gnp.sq<-gnp^2

> def.sq<-def^2
```

```
> (res.aux<-lm(res.t.sq~gnp+def+gnp.sq+def.sq+gnp*def))

Call:
lm(formula = res.t.sq ~ gnp + def + gnp.sq + def.sq + gnp * def)

Coefficients:
(Intercept)         gnp          def       gnp.sq
  2.497e+08   4.800e+03   -6.387e+06   -6.177e-01
     def.sq     gnp:def
  1.443e+04   2.116e+02


> lm<-n*summary(res.aux)$r.squared

> lm
[1] 19.65366

> (pchisq(lm, df=5, lower.tail=F))
[1] 0.0014512
```



White 검정결과 화폐수요함수에 이분산이 있는 것으로 나타났으므로 이분산을 고려하여 회귀계수를 추정하는 b2-ch6-2.R을 실행한 결과를 이용하여 설명해 보자.

첫째, OLS로 추정한 회귀계수는 $\hat{\beta}_2 = 1.332, \hat{\beta}_3 = 412.6$ 인 반면에 GLS로 추정한

회귀계수는 $\hat{\beta}_2 = 1.532, \hat{\beta}_3 = 298.8$ 이다.

둘째, OLS로 추정한 회귀계수의 표준오차는 $se(\hat{\beta}_2) = 0.2229, se(\hat{\beta}_3) = 61.57$ 인 반면에 GLS로 추정한 회귀계수의 표준오차는 $se(\hat{\beta}_2) = 0.1979, se(\hat{\beta}_3) = 50.93$ 으로 GLS 추정량이 효율적인 추정량임을 확인할 수 있다.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.560e+04  2.778e+03 -12.814  < 2e-16 ***
gnp          1.332e+00  2.229e-01   5.975 1.17e-07 ***
def          4.126e+02  6.157e+01   6.702 6.59e-09 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> nm2<-m2/sqrt(gnp)
> nc<-1/sqrt(gnp)
> ngnp<-sqrt(gnp)
> ndef<-def/sqrt(gnp)

> wls.lm<-lm(nm2~nc+ngnp+ndef-1)

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
nc   -2.881e+04  2.248e+03 -12.816  < 2e-16 ***
ngnp  1.532e+00  1.979e-01   7.738 1.02e-10 ***
ndef  2.988e+02  5.093e+01   5.867 1.79e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

셋째, 데이터를 변환하지 않고, 변환에 사용된 독립변수를 지정해 주고(여기서는 w0<-1/gnp로 지정해 줌) OLS로 추정하면 추정된 회귀계수는 GLS 추정치가 된다.

```
> w0<-1/gnp

> gls.lm<-lm(m2~gnp+def, weight=w0)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.881e+04  2.248e+03 -12.816  < 2e-16 ***
gnp          1.532e+00  1.979e-01   7.738 1.02e-10 ***
def          2.988e+02  5.093e+01   5.867 1.79e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b2-ch6-2.R의 실행결과

```
> library(stargazer)

> sample1<-("http://kanggc.iptime.org/book/data/korea(77-93)-1.txt")

> sample1_dat<-read.delim(sample1,header=T)

> gnp<-ts(sample1_dat$gnp, start=c(1977,1), frequency=4)

> def<-ts(sample1_dat$def, start=c(1977,1), frequency=4)

> m2<-ts(sample1_dat$m2, start=c(1977,1), frequency=4)

> ols.res<-lm(m2~gnp+def)

> summary(ols.res)

Call:
lm(formula = m2 ~ gnp + def)

Residuals:
   Min     1Q  Median     3Q    Max
```

```
-10838  -4271  -1723   3469  22428

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.560e+04  2.778e+03 -12.814  < 2e-16 ***
gnp          1.332e+00  2.229e-01   5.975 1.17e-07 ***
def          4.126e+02  6.157e+01   6.702 6.59e-09 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7074 on 63 degrees of freedom
Multiple R-squared:  0.9305,     Adjusted R-squared:  0.9283
F-statistic: 421.6 on 2 and 63 DF,  p-value: < 2.2e-16


> res<-resid(ols.res)

> res.t<-ts(res)

> nm2<-m2/sqrt(gnp)

> nc<-1/sqrt(gnp)

> ngnp<-sqrt(gnp)

> ndef<-def/sqrt(gnp)

> wls.lm<-lm(nm2~nc+ngnp+ndef-1)

> summary(wls.lm)

Call:
lm(formula = nm2 ~ nc + ngnp + ndef - 1)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-69.004 -26.602  -7.132  23.311 141.640

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
nc    -2.881e+04  2.248e+03 -12.816  < 2e-16 ***
ngnp   1.532e+00  1.979e-01   7.738 1.02e-10 ***
ndef   2.988e+02  5.093e+01   5.867 1.79e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.47 on 63 degrees of freedom
Multiple R-squared:  0.9659,     Adjusted R-squared:  0.9643
F-statistic: 595.5 on 3 and 63 DF,  p-value: < 2.2e-16


> w0<-1/gnp

> gls.lm<-lm(m2~gnp+def, weight=w0)

> summary(gls.lm)

Call:
lm(formula = m2 ~ gnp + def, weights = w0)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-69.004 -26.602  -7.132  23.311 141.640

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.881e+04  2.248e+03 -12.816  < 2e-16 ***
gnp          1.532e+00  1.979e-01   7.738 1.02e-10 ***
def          2.988e+02  5.093e+01   5.867 1.79e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 46.47 on 63 degrees of freedom
Multiple R-squared:  0.9228,	Adjusted R-squared:  0.9203 
F-statistic: 376.4 on 2 and 63 DF,  p-value: < 2.2e-16


> stargazer(ols.res, wls.lm, type="text")

==========================================================
                              Dependent variable:
                         ---------------------------------
                              m2               nm2
                              (1)              (2)
----------------------------------------------------------
gnp                        1.332***
                           (0.223)

def                       412.618***
                           (61.568)

nc                                         -28,814.480***
                                            (2,248.348)

ngnp                                          1.532***
                                              (0.198)

ndef                                         298.800***
                                              (50.930)

Constant                 -35,600.010***
                          (2,778.197)

----------------------------------------------------------
Observations                  66                 66
R2                           0.930              0.966
Adjusted R2                  0.928              0.964
Residual Std. Error (df = 63) 7,073.910         46.468
F Statistic           421.563*** (df = 2; 63) 595.525*** (df = 3; 63)
==========================================================
Note:                            *p<0.1; **p<0.05; ***p<0.01
```

```
> stargazer(wls.lm, gls.lm, type="text")

==================================================================
                              Dependent variable:

------------------------------------------------
                         nm2                    m2
                         (1)                    (2)
------------------------------------------------------------------
nc                  -28,814.480***
                     (2,248.348)

ngnp                   1.532***
                       (0.198)

ndef                  298.800***
                      (50.930)

gnp                                          1.532***
                                             (0.198)

def                                         298.800***
                                            (50.930)

Constant                                 -28,814.480***
                                          (2,248.348)

------------------------------------------------------------------
Observations                  66                   66
R2                           0.966                0.923
Adjusted R2                  0.964                0.920
Residual Std. Error (df = 63)   46.468            46.468
F Statistic             595.525*** (df = 3; 63) 376.420*** (df = 2; 63)
==================================================================
========================================
Note:                              *p<0.1; **p<0.05; ***p<0.01
```

제 7 장

시차분포모형

1. 시차분포모형
2. 시차분포모형 추정

제7장 시차분포모형

1. 시차분포모형

상수항과 k-1개의 독립변수에 의해 설명되는 다음과 다중회귀모형이 있다고 하자.

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

이 모형에서는 모든 설명변수들의 현재 값이 종속변수의 현재 값에 영향을 주고 있는데 이러한 모형을 정태적 모형(static model)이라고 한다.

한편, 다음과 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.

$$y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_k X_{t-k} + u_t \qquad Ⓐ$$

위의 시차분포모형에서 교란항에 대한 가정은 다중회귀모형에서 교란항에 대한 기본가정과 동일하며, 시차분포모형은 한 시점의 경제현상은 그 이전의 경제현상으로부터 영향을 받는 것으로 보고 있는데 시차의 수와 유형은 경험적으로 결정된다.

시차분포모형은 <표 7-1>과 같은 설명변수의 현재 값 및 과거 값들을 설명변수로 포함하고 있기 때문에 설명변수 간에 상관관계가 존재하는 다중공선성(multicollinearity)의 문제가 발생할 수 있다.

<표 7-1> 시차 값의 예시

관측치	X_t	X_{t-1}	X_{t-2}	X_{t-3}	.	.	X_{t-k}
1	X_1	–	–	–	–	–	–
2	X_2	X_1	–	–	–	–	–
3	X_3	X_2	X_1	–	–	–	–
4	X_1	X_3	X_2	X_1	–	–	–
.	.	.	.	.	.	.	–
k	X_k	.	.	.	.	.	–
.	.	X_k	.	.	.	.	X_1
.	.	.	X_k	.	.	.	.
.	.	.	.	X_k	.	.	.
.	.	.	.	.	.	.	.
n	X_n	X_{n-1}	X_{n-2}	X_{n-3}	.	.	X_{n-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.

예를 들어, 시차분포모형이 다음과 같다고 할 때 $\frac{\partial y_t}{\partial X_t} = \beta_0 = 0.4$를 단기승수(short-run multiplier)라고 한다. 따라서 이 시차분포모형이 y_t는 소비, X_t는 소득인 소비함수라면 이 값은 단기 한계소비성향(Marginal Propensity to Consume: MPC)이 된다.

$$y_t = \alpha + 0.4X_t + 0.3X_{t-1} + 0.2X_{t-2} + u_t$$

한편, 다음을 장기승수(long-run multiplier)라고 하고 소비함수에서 이 값은 장기 한계소비성향(MPC)이 된다.

$$\frac{\partial y_t}{\partial X_t} + \frac{\partial y_t}{\partial X_{t-1}} + \frac{\partial y_t}{\partial X_{t-2}} = \sum_{i=0}^{2} \beta_i = 0.4 + 0.3 + 0.2 = 0.9$$

2. 시차분포모형 추정

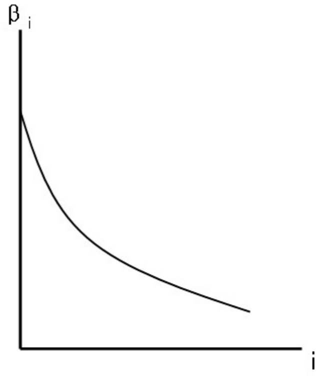
시차분포모형을 보통최소자승법(OLS)으로 추정할 경우 잘못된 통계적 추론을 가져올 수 있다. 왜냐하면 시차변수를 설명변수로 포함시키면 자유도가 작아지고(자유도는 n-k이므로), 자유도가 작아지면 귀무가설을 기각할 영역이 작아지므로 통계적으로 유의한 변수도 유의하지 않은 것으로 결론을 내릴 수 있게 된다. 따라서 시차분포모형은 OLS로 추정하지 않고 다른 방법으로 추정하는데 대표적인 추정방법으로는 Koyck과 Almon의 추정방법이 있다.

(1) Koyck의 추정방법

Koyck의 추정방법은 시차의 수를 가능한 줄이려는 의도에서 출발하였는데 <그림 7-1>과 같이 과거로 거슬러 올라 갈수록 y에 대한 시차의 영향이 점점 작아진다는 가정을 도입하고 있으므로 내생시차변수모형(endogenous lagged variable model)이라고 불린다.

$$\beta_k = \beta_0 \lambda^k, \quad k = 0,1,2,3,\ldots$$

단, $0 < \lambda < 1$



<그림 7-1> 시차계수 구조(Almon)

시차계수에 대한 가정을 시차분포모형에 대입하면 다음과 같다.

$$y_t = \alpha + \beta_0 X_t + (\beta_0 \lambda) X_{t-1} + \ldots + (\beta_0 \lambda^{k-1}) X_{t-k+1} + (\beta_0 \lambda^k) X_{t-k} + u_t \quad Ⓑ$$

위의 모형에서 시차를 하나 뒤로 하면 다음과 같게 된다.

$$y_{t-1} = \alpha + \beta_0 X_{t-1} + (\beta_0 \lambda) X_{t-2} + \ldots + (\beta_0 \lambda^{k-1}) X_{t-k} + u_{t-1}$$

위의 모형에 λ를 곱하면 다음과 같다.

$$\lambda y_{t-1} = \lambda\alpha + (\beta_0 \lambda) X_{t-1} + (\beta_0 \lambda^2) X_{t-2} + \ldots + (\beta_0 \lambda^k) X_{t-k} + \lambda u_{t-1} \quad Ⓒ$$

Ⓑ식에서 Ⓒ식을 빼고 정리하면 다음과 같게 된다.

$$y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda y_{t-1} + v_t \quad (\text{단, } v_t = u_t - \lambda u_{t-1}) \quad Ⓓ$$

Ⓓ식은 다중공선성 문제가 없으므로 OLS로 추정하면 $\hat{\alpha}$, $\hat{\beta}_0$, $\hat{\lambda}$ 을 구할 수 있고 $\widehat{\beta_k} = \widehat{\beta_0}\widehat{\lambda^k}$를 이용하여 $\widehat{\beta_k}$를 구한다. 이 때 OLS추정량은 불편추정량 뿐만 아니라 효율적인 추정량도 되지 못하는데 그 이유는 교란항에 자기상관이 있고 또한 직교조건도 성립되지 않기 때문이다.

이에 대한 해결책으로는 Ⓓ식의 교란항 v_t와 독립이고 설명변수 y_{t-1}과 상관관계가 있는 변수를 이용하는데 이를 수단변수(Instrumental Variable: IV)라고 한다. 여기서는 X_{t-1}이 수단변수로 이용될 수 있는데 그 이유는 X_{t-1}이 이러한 조건을 만족시키는 좋은 변수이기 때문이다.

(2) Almon의 추정방법

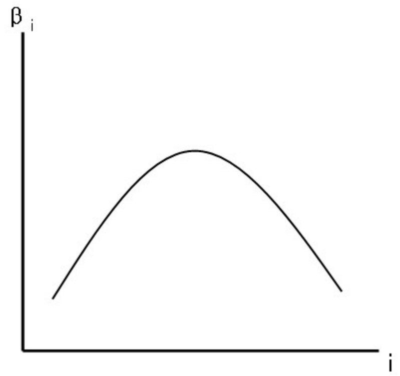
Almon의 추정방법은 외생시차변수(exogenous lagged variable)의 모수를 추정한 후 시차분포모형의 모수를 계산하는 방법으로 다항식 시차(polynomial distributed lag: PDL) 모형이라고도 한다. 이 방법은 시차계수의 크기가 시차에 따라 2차 함수 또는 3차 함수 등 다항식의 구조를 가진다고 가정한다. 즉, 시차계수(β_i)가 다음과 같이 시차의 길이인 i의 적절한 차수의 다항식의 근사치로 계산될 수 있다고 가정한다.

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2 + ... + \alpha_k i^r$$

위 식에서 r=2이면 2차 함수, r=3이면 3차 함수 등이 되는데 예를 들어 r=2인 경우 다음과 같은 2차 다항식이 된다.

$$\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2$$

위 식에 따른 시차구조를 그림으로 그려보면 <그림 7-2>와 같은데 이는 시차계수 β_i는 시차에 따라 <그림 7-2>와 같은 구조를 갖는다고 가정한 것과 같다



<그림 7-2> 시차계수 구조(Almon)

예를 들어 2차 다항식을 가정하는 경우 β는 α들과 다음의 관계에 있으므로 α_0, α_1, α_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 Ⓐ식에 대입하면 다음 식을 얻게 된다.

$$\begin{aligned}
y_t &= \alpha + \alpha_0 X_t + (\alpha_0 + \alpha_1 + \alpha_2) X_{t-1} + \ldots + (\alpha_0 + k\alpha_1 + k^2\alpha_2) X_{t-k} + u_t \\
&= \alpha + \alpha_0 (X_t + X_{t-1} + \ldots + X_{t-k}) + \alpha_1 (X_{t-1} + 2X_{t-2} + \ldots + kX_{t-k}) \\
&\quad + \alpha_2 (X_{t-1} + 2^2 X_{t-2} + \ldots + k^2 X_{t-k}) + u_t \\
&= \alpha + \alpha_0 (\sum_{k=0}^{K} X_{t-k}) + \alpha_1 (\sum_{k=1}^{K} kX_{t-k}) + \alpha_2 (\sum_{k=1}^{K} k^2 X_{t-k}) + u_t \\
&= \alpha + \alpha_0 Z_{1t} + \alpha_1 Z_{2t} + \alpha_2 Z_{3t} + u_t \qquad Ⓔ
\end{aligned}$$

단, $Z_{1t} \equiv \sum_{k=0}^{K} X_{t-k}$, $Z_{2t} \equiv \sum_{k=0}^{K} kX_{t-k}$, $Z_{3t} \equiv \sum_{k=0}^{K} k^2 X_{t-k}$이다.

위 식을 OLS로 추정하여 $\hat{\alpha}$, $\widehat{\alpha_0}$, $\widehat{\alpha_1}$, $\widehat{\alpha_2}$ 을 구하고 $\beta_i = \alpha_0 + \alpha_1 i + \alpha_2 i^2$의 관계를 이용하여 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ 를 계산한다.

(예제 7) 1995년 1/4분기부터 2001년 1/4분기까지 한국의 GDP와 총소비(consume) 자료를 이용하여 시차분포모형의 소비함수를 추정하고자 한다. 현재 소득 및 과거 4분기의 소득이 현재의 소비에 영향을 준다고 가정하자. Almon 추정에서는 시차계수의 구조가 2차 다항식이라고 가정하자.

(예제 7)에서 Ⓓ식의 Koyck의 추정방법을 이용한 b2-ch7-1.R을 실행한 결과를 설명해 보자. IV 추정방법을 이용하여 소비함수를 추정한 결과는 다음과 같다.

$$\hat{y}_t = 1.992 \times 10^4 + 0.347 X_t + 0.2168 y_{t-1}$$

그리고 추정된 Ⓓ식에서 $1.992 \times 10^4 = \hat{\alpha}(1-\hat{\lambda})$, $0.347 = \hat{\beta}_0$, $0.2168 = \hat{\lambda}$ 의 관계가 있으므로 $\hat{\alpha} = 30497.44$, $\hat{\beta}_0 = 0.2168204$, $\hat{\lambda} = 0.3469805$ 이 됨을 확인할 수 있다.

```
Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 1.992e+04  7.180e+03    2.774    0.0114 *
Y.1         3.470e-01  2.105e-01    1.649    0.1141
X.t         2.168e-01  1.028e-01    2.110    0.0471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2144 on 21 degrees of freedom
Multiple R-Squared: 0.7252,      Adjusted R-squared: 0.699
Wald test: 24.15 on 2 and 21 DF,  p-value: 3.599e-06

                             alpha      beta       phi
Geometric coefficients:  30497.44 0.2168204 0.3469805
```

b2-ch7-1.R의 실행결과

```
> library(dLagM)

> sample1<-("http://kanggc.iptime.org/book/data/ar.txt")

> sample1_dat<-read.delim(sample1,header=T)

> sample1_dat
   consume       gdp
1  58431.8  93061.99
2  59455.3  93613.55
3  61165.7  94788.82
4  63787.3  95894.36
5  62845.6  99094.99
6  63957.8  99903.92
7  64929.8 100951.85
```

```
8   68684.3 102886.05
9   65445.2 103604.50
10  66575.1 106036.68
11  68103.8 106458.19
12  68598.5 106926.21
13  58780.6  99153.69
14  58902.6  97622.45
15  60811.6  97887.44
16  63193.2 100077.50
17  62768.3 104465.50
18  64474.1 108484.40
19  67192.6 110529.11
20  70045.3 114271.82
21  68711.6 116666.09
22  69481.3 118947.45
23  70508.4 120696.31
24  72155.9 119995.03
25  68975.7 120633.92

> sample1.ts<-ts(sample1_dat, start=c(1995,1), end=c(2001,1), frequency=4)

> cons<-sample1.ts[,1]

> gdp<-sample1.ts[,2]

> lcons1<-cons[1:24]

> lgdp1<-gdp[1:24]

> fm.res <- ivreg(cons[2:25] ~ lcons1+gdp[2:25] | lcons1+lgdp1)

> summary(fm.res)

Call:
ivreg(formula = cons[2:25] ~ lcons1 + gdp[2:25] | lcons1 + lgdp1)

Residuals:
```

```
    Min        1Q   Median        3Q       Max
-6435.7   -775.4    191.6   1159.9   3931.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.992e+04  7.180e+03    2.774    0.0114 *
lcons1       3.470e-01  2.105e-01    1.649    0.1141
gdp[2:25]     2.168e-01  1.028e-01    2.110    0.0471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2144 on 21 degrees of freedom
Multiple R-Squared: 0.7252,       Adjusted R-squared: 0.699
Wald test: 24.15 on 2 and 21 DF,  p-value: 3.599e-06


> data(sample1_dat)

> koyck.res<-koyckDlm(x = sample1_dat$gdp,y = sample1_dat$consume)

> summary(koyck.res)

Call:
"Y ~ (Intercept) + Y.1 + X.t"

Residuals:
    Min        1Q   Median        3Q       Max
-6435.7   -775.4    191.6   1159.9   3931.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.992e+04  7.180e+03    2.774    0.0114 *
Y.1          3.470e-01  2.105e-01    1.649    0.1141
X.t          2.168e-01  1.028e-01    2.110    0.0471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2144 on 21 degrees of freedom
Multiple R-Squared: 0.7252,      Adjusted R-squared: 0.699
Wald test: 24.15 on 2 and 21 DF,  p-value: 3.599e-06

                              alpha      beta       phi
Geometric coefficients:  30497.44 0.2168204 0.3469805
```

(예제 7)에서 Almon의 추정방법을 이용한 b2-ch7-2.R을 실행한 결과를 설명해 보자.

첫째, Ⓔ식을 OLS로 추정하여 $\hat{\alpha}$, $\widehat{\alpha_0}$, $\widehat{\alpha_1}$, $\widehat{\alpha_2}$ 을 구하면 다음과 같다. 즉, $\hat{\alpha}=3.192\times 10^4$, $\hat{\alpha}_0=0.4386$, $\hat{\alpha}_1=-0.3777$, $\hat{\alpha}_2=0.006313$ 이다.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.192e+04  7.705e+03   4.143 0.000681 ***
z.t0         4.386e-01  1.257e-01   3.491 0.002799 **
z.t1        -3.777e-01  2.273e-01  -1.662 0.114871
z.t2         6.313e-02  5.721e-02   1.104 0.285185
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

둘째, $\beta_i=\alpha_0+\alpha_1 i+\alpha_2 i^2$의 관계를 이용하여 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ 를 계산하면 다음과 같다. 즉, $\hat{\beta}_0=0.439$, $\hat{\beta}_1=0.124$, $\hat{\beta}_2=-0.0643$, $\hat{\beta}_3=-0.126$, $\hat{\beta}_4=-0.0622$ 이다. 따라서 시차분포모형의 추정회귀식은 다음과 같다.

$$\widehat{y_t}=31920+0.439X_t+0.124X_{t-1}-0.0643X_{t-2}-0.126X_{t-3}-0.0622X_{t-4}$$

```
Estimates and t-tests for beta coefficients:
        Estimate Std. Error t value P(>|t|)
beta.0    0.4390     0.1260   3.490  0.0028
beta.1    0.1240     0.0608   2.040  0.0572
beta.2   -0.0643     0.1090  -0.589  0.5640
beta.3   -0.1260     0.0583  -2.170  0.0448
beta.4   -0.0622     0.1400  -0.445  0.6620
```

셋째, 시차분포모형의 추정결과에 따르면 소득 수준의 변화는 1분기 후까지 소비의 움직임에 유의적인 영향을 미치다가 시간의 흐름에 따라 감소하는 것으로 나타났다. 한편, 단기 한계소비성향은 0.439이고, 장기 한계소비성향은 다음과 같이 계산하므로 0.3015가 된다.

$$\sum_{i=0}^{4}\beta_i = 0.439 + 0.124 - 0.0643 - 0.126 - 0.0622 = 0.3097 = 0.3105$$

b2-ch7-2.R의 실행결과

```
> library(dLagM)

> library(dynlm)

> rm(list=ls())

> sample1<-("http://kanggc.iptime.org/book/data/ar.txt")

> sample1_dat<-read.delim(sample1,header=T)

> sample1.ts<-ts(sample1_dat, start=c(1995,1), end=c(2001,1), frequency=4)

> cons<-sample1.ts[,1]

> gdp<-sample1.ts[,2]
```

```
> lgdp1<-gdp[4:24]

> lgdp2<-gdp[3:23]

> lgdp3<-gdp[2:22]

> lgdp4<-gdp[1:21]

> #summary(lm(cons[5:25]~gdp[5:25]+lgdp1+lgdp2+lgdp3+lgdp4))
>
#summary(lm(cons[5:25]~gdp[5:25]+gdp[4:24]+gdp[3:23]+gdp[2:22]+gdp[1:
21]))
> #summary(d .... [TRUNCATED]

Time series regression with "ts" data:
Start = 1996(1), End = 2001(1)

Call:
dynlm(formula = cons ~ L(gdp, 0:4))

Residuals:
    Min      1Q  Median      3Q     Max
-2992.0  -967.9  -256.1   794.8  3618.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.241e+04  8.068e+03   4.017  0.00112 **
L(gdp, 0:4)0   5.670e-01  2.149e-01   2.639  0.01860 *
L(gdp, 0:4)1  -1.094e-01  3.599e-01  -0.304  0.76536
L(gdp, 0:4)2  -2.122e-02  3.563e-01  -0.060  0.95331
L(gdp, 0:4)3   5.347e-02  3.693e-01   0.145  0.88681
L(gdp, 0:4)4  -1.858e-01  2.358e-01  -0.788  0.44296
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2122 on 15 degrees of freedom
```

```
Multiple R-squared:  0.766,     Adjusted R-squared:  0.688
F-statistic: 9.821 on 5 and 15 DF,  p-value: 0.0002556


> z1<-gdp[5:25]+lgdp1+lgdp2+lgdp3+lgdp4

> z2<-lgdp1+2*lgdp2+3*lgdp3+4*lgdp4

> z3<-lgdp1+4*lgdp2+9*lgdp3+16*lgdp4

> (summary(lm.res<-lm(cons[5:25]~z1+z2+z3)))

Call:
lm(formula = cons[5:25] ~ z1 + z2 + z3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3198.0 -1205.6  -101.2   729.5  4031.0 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.192e+04  7.705e+03   4.143 0.000681 ***
z1           4.386e-01  1.257e-01   3.491 0.002799 ** 
z2          -3.777e-01  2.273e-01  -1.662 0.114871    
z3           6.313e-02  5.721e-02   1.104 0.285185    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2033 on 17 degrees of freedom
Multiple R-squared:  0.7566,    Adjusted R-squared:  0.7137
F-statistic: 17.62 on 3 and 17 DF,  p-value: 1.844e-05


> beta0<-lm.res$coeff[2]

> beta1<-lm.res$coeff[2]+lm.res$coeff[3]+lm.res$coeff[4]
```

```
> beta2<-lm.res$coeff[2]+2*lm.res$coeff[3]+4*lm.res$coeff[4]

> beta3<-lm.res$coeff[2]+3*lm.res$coeff[3]+9*lm.res$coeff[4]

> beta4<-lm.res$coeff[2]+4*lm.res$coeff[3]+16*lm.res$coeff[4]

> (coeff_fdl<-rbind(lm.res$coeff[1],beta0, beta1, beta2, beta3, beta4))
        (Intercept)
       3.192060e+04
beta0  4.386243e-01
beta1  1.240260e-01
beta2 -6.431342e-02
beta3 -1.263940e-01
beta4 -6.221560e-02

> data(sample1_dat)

> almon.res<-polyDlm(x = sample1_dat$gdp,y = sample1_dat$consume, q=4,
k=2, show.beta=T)
Estimates and t-tests for beta coefficients:
       Estimate Std. Error t value P(>|t|)
beta.0   0.4390     0.1260   3.490  0.0028
beta.1   0.1240     0.0608   2.040  0.0572
beta.2  -0.0643     0.1090  -0.589  0.5640
beta.3  -0.1260     0.0583  -2.170  0.0448
beta.4  -0.0622     0.1400  -0.445  0.6620

> summary(almon.res)

Call:
"Y ~ (Intercept) + X.t"

Residuals:
    Min      1Q  Median     3Q    Max
-3198.0 -1205.6  -101.2  729.5 4031.0

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.192e+04  7.705e+03   4.143 0.000681 ***
z.t0         4.386e-01  1.257e-01   3.491 0.002799 **
z.t1        -3.777e-01  2.273e-01  -1.662 0.114871
z.t2         6.313e-02  5.721e-02   1.104 0.285185
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2033 on 17 degrees of freedom
Multiple R-squared:  0.7566,	Adjusted R-squared:  0.7137
F-statistic: 17.62 on 3 and 17 DF,  p-value: 1.844e-05
```

참 고 문 헌

강기춘(2010), 『계량경제학:이론과 실습』, 온누리.

김영우(2017), 『Do it! 쉽게 배우는 R 데이터 분석』, 이지스퍼브리싱.

남준우 · 이한식(2005), 『계량경제학 이론과 EViews·Excel 활용』, 홍문사.

박범조(2013), 『응용 계량경제학 : R 활용』, 스그마프레스.

이용구 · 김삼용(2016), 『통계학의 이해 : EXCEL 실습 -제8판-』, 율곡출판사.

이재길(2017), 『R 프로그램에 기반한 시계열 자료 분석』, 황소걸음 아카데미.

폴 티터 지음 · 이제원 옮김(2012), 『R Cookbook : 데이터 분석과 통계 그래픽스를 위한 실전 예제』, 인사이트.

한치록(2017), 『계량경제학 강의』, 박영사.

Cowpertwait, P.S.P and A.V. Metcalfe(2009), *Introductory Time Series with R*, Springer.

Gujarati. D.(2004). *Basic Econometrics*, Fourth Edition, McGraw-Hill.

Heiss, F.(2016), *Using R for Introductory Econometrics*, John Wiley & Sons,Inc.

Hill, R., W. Griffiths and G. Judge(2001), *Undergraduate Econometrics*, Second Edition, Wiley.

Wooldridge, J.(2006), *Introductory Econometrics A Modern Approach*, Thomson.

부록 1

R 코드

제1장 계량모형의 추정

b2-ch1-1.R

```
library(quantreg)
library(bbmle)

par(mfrow=c(3,1))

set.seed(123455)
n<-500; r<-1000
b0<-1; b1<-0.5; su<-2


b0ols<-numeric(r)
b1ols<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  bols<-coefficients(lm(y~x))
  b0ols[j]<-bols["(Intercept)"]
  b1ols[j]<-bols["x"]
  }

hist(b1ols, breaks=100, xlim=c(0,1), xlab="n=500")


b0lad<-numeric(r)
b1lad<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
```

```
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  blad<-coefficients(rq(y~x))
  b0lad[j]<-blad["(Intercept)"]
  b1lad[j]<-blad["x"]
}

hist(b1lad, breaks=100,xlim=c(0,1),xlab="n=500")



b0mle<-numeric(r)
b1mle<-numeric(r)

for(j in 1:r) {

  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

fn<-function(beta0, beta1, sigma) {
  (n/2)*log(sigma^2)+1/(2*sigma^2)*(sum((y-beta0-beta1*x)^2))
  }
res<-mle2(fn,start=list(beta0=0.8, beta1=0.3, sigma=1.5))
summary(res)
str(summary(res))

b0mle[j]<-summary(res)@coef["beta0","Estimate"]
b1mle[j]<-summary(res)@coef["beta1","Estimate"]
}

hist(b1mle, breaks=100, xlim=c(0,1),xlab="n=500")

mean(b0ols)
mean(b1ols)

var(b0ols)
```

```
var(b1ols)

mean(b0lad)
mean(b1lad)

var(b0lad)
var(b1lad)

mean(b0mle)
mean(b1mle)

var(b0mle)
var(b1mle)
```

b2-ch1-2.R

```
library(quantreg)
library(bbmle)

par(mfrow=c(3,1))

set.seed(123456)
n<-5; r<-1000
b0<-1; b1<-0.5; su<-2


b0ols5<-numeric(r)
b1ols5<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  bols5<-coefficients(lm(y~x))
  b0ols5[j]<-bols5["(Intercept)"]
```

```
  b1ols5[j]<-bols5["x"]
  }

n<-50; r<-1000
b0<-1; b1<-0.5; su<-2


b0ols50<-numeric(r)
b1ols50<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  bols50<-coefficients(lm(y~x))
  b0ols50[j]<-bols50["(Intercept)"]
  b1ols50[j]<-bols50["x"]
}

n<-500; r<-1000
b0<-1; b1<-0.5; su<-2


b0ols500<-numeric(r)
b1ols500<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  bols500<-coefficients(lm(y~x))
  b0ols500[j]<-bols500["(Intercept)"]
  b1ols500[j]<-bols500["x"]
}

hist(b1ols5, breaks=100, xlim=c(-2,2), xlab="n=5")
```

```
hist(b1ols50, breaks=100, xlim=c(-2,2), xlab="n=50")
hist(b1ols500, breaks=100, xlim=c(-2,2), xlab="n=500")


n<-5; r<-1000
b0<-1; b1<-0.5; su<-2


b0lad5<-numeric(r)
b1lad5<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  blad5<-coefficients(rq(y~x))
  b0lad5[j]<-blad5["(Intercept)"]
  b1lad5[j]<-blad5["x"]
}


n<-50; r<-1000
b0<-1; b1<-0.5; su<-2


b0lad50<-numeric(r)
b1lad50<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  blad50<-coefficients(rq(y~x))
  b0lad50[j]<-blad50["(Intercept)"]
  b1lad50[j]<-blad50["x"]
}
```

```
n<-500; r<-1000
b0<-1; b1<-0.5; su<-2


b0lad500<-numeric(r)
b1lad500<-numeric(r)

for(j in 1:r) {
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  blad500<-coefficients(rq(y~x))
  b0lad500[j]<-blad500["(Intercept)"]
  b1lad500[j]<-blad500["x"]
}

hist(b1lad5, breaks=100,xlim=c(-2,2),xlab="n=5")
hist(b1lad50, breaks=100,xlim=c(-2,2),xlab="n=50")
hist(b1lad500, breaks=100,xlim=c(-2,2),xlab="n=500")


n<-5; r<-1000
b0<-1; b1<-0.5; su<-2

b0mle5<-numeric(r)
b1mle5<-numeric(r)

for(j in 1:r) {

  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

fn<-function(beta0, beta1, sigma) {
  (n/2)*log(sigma^2)+1/(2*sigma^2)*(sum((y-beta0-beta1*x)^2))
  }
```

```
res<-mle2(fn,start=list(beta0=0.8, beta1=0.3, sigma=1.5))
summary(res)
str(summary(res))

b0mle5[j]<-summary(res)@coef["beta0","Estimate"]
b1mle5[j]<-summary(res)@coef["beta1","Estimate"]
}

n<-50; r<-1000
b0<-1; b1<-0.5; su<-2

b0mle50<-numeric(r)
b1mle50<-numeric(r)

for(j in 1:r) {

  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  fn<-function(beta0, beta1, sigma) {
    (n/2)*log(sigma^2)+1/(2*sigma^2)*(sum((y-beta0-beta1*x)^2))
  }
  res<-mle2(fn,start=list(beta0=0.8, beta1=0.3, sigma=1.5))
  summary(res)
  str(summary(res))

  b0mle50[j]<-summary(res)@coef["beta0","Estimate"]
  b1mle50[j]<-summary(res)@coef["beta1","Estimate"]
}

n<-500; r<-1000
b0<-1; b1<-0.5; su<-2

b0mle500<-numeric(r)
b1mle500<-numeric(r)

for(j in 1:r) {
```

```
  x<-rnorm(n,4,1)
  u<-rnorm(n,0,su)
  y<-b0+b1*x+u

  fn<-function(beta0, beta1, sigma) {
    (n/2)*log(sigma^2)+1/(2*sigma^2)*(sum((y-beta0-beta1*x)^2))
  }
  res<-mle2(fn,start=list(beta0=0.8, beta1=0.3, sigma=1.5))
  summary(res)
  str(summary(res))

  b0mle500[j]<-summary(res)@coef["beta0","Estimate"]
  b1mle500[j]<-summary(res)@coef["beta1","Estimate"]
}

hist(b1mle5, breaks=100, xlim=c(-2,2),xlab="n=5")
hist(b1mle50, breaks=100, xlim=c(-2,2),xlab="n=50")
hist(b1mle500, breaks=100, xlim=c(-2,2),xlab="n=500")

mean(b0ols5)
mean(b1ols5)

mean(b0ols50)
mean(b1ols50)

mean(b0ols500)
mean(b1ols500)

var(b0ols5)
var(b1ols5)

var(b0ols50)
var(b1ols50)

var(b0ols500)
var(b1ols500)
```

```
mean(b0lad5)
mean(b1lad5)

mean(b0lad50)
mean(b1lad50)

mean(b0lad500)
mean(b1lad500)

var(b0lad5)
var(b1lad5)

var(b0lad50)
var(b1lad50)

var(b0lad500)
var(b1lad500)

mean(b0mle5)
mean(b1mle5)

mean(b0mle50)
mean(b1mle50)

mean(b0mle500)
mean(b1mle500)

var(b0mle5)
var(b1mle5)

var(b0mle50)
var(b1mle50)

var(b0mle500)
var(b1mle500)
```

제2장 단순회귀분석

b2-ch2-1.R

```
x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)
x;y

(n<-length(x))

(sumx<-sum(x))
(sumy<-sum(y))
(mx=mean(x))
(my=mean(y))
(xy<-x*y)
(sumxy<-sum(xy))
(sumxsq<-sum(x^2))
(sumysq<-sum(y^2))

beta1<-(sumxy-mx*sumy)/(sumxsq-mx*sumx)
beta0<-my-beta1*mx
beta0;beta1

(b1hat<-cov(x,y)/var(x))
(b0hat<-mean(y)-b1hat*mean(x))

(dx<-x-mx)
(dy<-y-my)
(sumdxsq<-sum(dx^2))
(sumdysq<-sum(dy^2))
(sumdxdy<-sum(dx*dy))
(ssr<-sumdysq-beta1*sumdxdy)
(sigusq<-ssr/3)
```

```
(vbeta1<-sigusq/sumdxsq)

(rsq<-(beta1^2*sumdxsq)/sumdysq)

(t<-beta1/sqrt(vbeta1))
(pt(t,3))

(tc<-qt(p=0.025, df=3, lower.tail=F))

(b1hat_lb<-b1hat-(tc)*sqrt(vbeta1))
(b1hat_ub<-b1hat+(tc)*sqrt(vbeta1))

x0<-7
(yhat<-beta0+beta1*x0)

(sigesq_ind<-sigusq*(1+(1/n)+((x0-mx)^2/sumdxsq)))
(sige_ind<-sqrt(sigesq_ind))
(yhat_ind_lb<-(yhat-(-tc)*sige_ind))
(yhat_ind_ub<-(yhat+(-tc)*sige_ind))


(sigesq_mean<-sigusq*((1/n)+((x0-mx)^2/sumdxsq)))
(sige_mean<-sqrt(sigesq_mean))
(yhat_mean_lb<-(yhat-(tc)*sige_mean))
(yhat_mean_ub<-(yhat+(tc)*sige_mean))
```

b2-ch2-2.R

```
x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)
x;y

lm(y~x)
ols<-lm(y~x)
summary(ols)
```

```
plot(y~x, ylim=c(0,10))
abline(lm(y~x))

confint(ols)

predict(lm(y~x))
new<-data.frame(x = 7)
predict(lm(y~x), new, se.fit = TRUE)
pred.w.plim <- predict(lm(y ~ x), new, interval = "prediction")
pred.w.plim
pred.w.clim<-predict(lm(y ~ x), new, se.fit=T, interval = "confidence")
pred.w.clim
```

b2-ch2-3.R

```
x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)
x;y

lx<-log(x)
ly<-log(y)
lx;ly

rx<-1/x
rx

lm(y~x)
ols1<-lm(y~x)
summary(ols1)

lm(ly~lx)
ols2<-lm(ly~lx)
summary(ols2)

lm(y~rx)
ols3<-lm(y~rx)
```

```
summary(ols3)

lm(ly~x)
ols4<-lm(ly~x)
summary(ols4)

lm(y~lx)
ols5<-lm(y~lx)
summary(ols5)
```

b2-ch2-4.R

```
library(openxlsx)
library(bbmle)
library(quantreg)

x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)


# LAD estimator
taus<-c(0.5, 0.7)
rq(y~x)
lad<-rq(y~x, tau=taus)
summary(lad)

plot(x,y,type = "n", cex = .8,xlim=c(0,7), ylim=c(0,11))
points(x,y,cex = .8,col="blue",xlim=c(0,7), ylim=c(0,11))
abline(rq(y~x, tau=0.5),col="red",lty = 2)
abline(rq(y~x, tau=0.7), col="blue")
```

b2-ch2-5.R

```
library(openxlsx)
library(bbmle)
library(quantreg)

x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)

(n<-length(y))

# OLS estimator
lm(y~x)
ols<-lm(y~x)
summary(ols)
confint(ols)

# LAD estimator
#taus<-c(.05,.1,.25,.75,.9, 0.95)
rq(y~x)
lad<-rq(y~x)
summary(lad)

# MLE
fn<-function(b0, b1, sigma) {
  (n/2)*log(sigma^2) + 1/(2*sigma^2)*(sum((y-b0-b1*x)^2))
}
res<-mle2(fn,start=list(b0=0.5, b1=0.5, sigma=1))
summary(res)

plot(x,y,type = "n", cex=.8, xlim=c(0,7), ylim=c(0,11))
points(x,y,cex=.8,col="blue",xlim=c(0,7), yiim=c(0,11))
abline(lm(y~x),col="red",lty = 2)
abline(rq(y~x), col="blue")
```

b2-ch2-6.R

```
library(openxlsx)
library(bbmle)
library(quantreg)

x<-c(0,1.9,4,6,8.5,0.5,2.5,4.5,6.6,9,1,3,5,7,10,1.5,3.5,5.5,7)
y<-c(1,2,2.7,3.6,6,0.9,2.4,3.5,2.7,6,0.7,3.2,1,5.7,7.3,1.5,2,4,4.6)

(n<-length(y))

# OLS estimator
lm(y~x)
ols<-lm(y~x)
summary(ols)
confint(ols)

# LAD estimator
rq(y~x)
lad<-rq(y~x, model=T)
summary(lad)

# MLE
fn<-function(b0, b1, sigma) {
  (n/2)*log(sigma^2)+1/(2*sigma^2)*(sum((y-b0-b1*x)^2))
}
res<-mle2(fn,start=list(b0=0.5, b1=0.5, sigma=1))
summary(res)

plot(x,y,type = "n", cex = .8)
points(x,y,cex = .8,col="blue")
abline(lm(y~x),col="red",lty = 2)
abline(rq(y~x), col="blue")
```

제3장 다중회귀분석

b2-ch3-1.R

```
x2<-c(1,2,3,2)
x3<-c(2,1,1,2)
y0<-c(1,1,2,3)
n<-length(y)

olsx2<-lm(x2~x3)
resx2<-resid(olsx2)
ols2<-lm(y0~resx2+x3)
summary(ols2)

olsx3<-lm(x3~x2)
resx3<-resid(olsx3)
ols3<-lm(y0~x2+resx3)
summary(ols3)

y0bar<-mean(y0)
x2bar<-mean(x2)
x3bar<-mean(x3)
y0sd<-sqrt(var(y0))
x2sd<-sqrt(var(x2))
x3sd<-sqrt(var(x3))
sy0<-(y0-y0bar)/y0sd
sx2<-(x2-x2bar)/x2sd
sx3<-(x3-x3bar)/x3sd
sols<-lm(sy0~sx2+sx3-1)
summary(sols)

y0sd
x2sd
x3sd
```

```
beta2hat<-summary(sols)$coef[1,1]*(y0sd/x2sd)
beta3hat<-summary(sols)$coef[2,1]*(y0sd/x3sd)
beta2hat
beta3hat
```

b2-ch3-2.R

```
x2<-c(1,2,3,2)
x3<-c(2,1,1,2)
y0<-c(1,1,2,3)
xls<-matrix(c(1,2,3,2,2,1,1,2), nrow=4, ncol=2)
x<-matrix(c(1,1,1,1,1,2,3,2,2,1,1,2), nrow=4, ncol=3)
y<-matrix(c(1,1,2,3), nrow=4)
n<-length(y)

x
xpx=t(x)%*%x
xpx
xpxinv=solve(xpx)
xpxinv
xpy=t(x)%*%y
xpy
beta<-xpxinv%*%xpy
beta

ypy=t(y)%*%y
ypy
bpxpxb=t(beta)%*%t(x)%*%x%*%beta
bpxpxb
epe=ypy-bpxpxb
epe
sigusq=epe/(n-3)
sigusq

rsq<-(bpxpxb-n*mean(y)^2)/(ypy-n*mean(y)^2)
```

```
rsq

varcov<-0.25*xpxinv
varcov
se<-sqrt(diag(varcov))
se

x0<-matrix(c(1,3,2))
yhat<-t(x0)%*%beta
x0
yhat

sigisq<-0.25*(1+t(x0)%*%xpxinv%*%x0)
sigisq
sigi<-sqrt(sigisq)
sigi
yhat+c(-1,1)*qt(.975, 1)*sigi

sigesq<-0.25*(t(x0)%*%xpxinv%*%x0)
sigesq
sige<-sqrt(sigesq)
sige
yhat+c(-1,1)*qt(.975, 1)*sige
```

b2-ch3-3.R

```
x2<-c(1,2,3,2)
x3<-c(2,1,1,2)
y<-c(1,1,2,3)

ols<-lm(y~x2+x3)
summary(ols)
confint(ols)
resid(ols)

predict(lm(y~x2+x3))
```

```
new<-data.frame(x2=3, x3=2)
predict(lm(y~x2+x3), new, se.fit = TRUE)
pred.w.clim<-predict(lm(y~x2+x3), new, interval = "confidence")
pred.w.clim

pred.w.plim<-predict(lm(y~x2+x3), new, interval = "prediction")
pred.w.plim
```

b2-ch3-4.R

```
library(car)
library(stargazer)

sample1<-("http://kanggc.iptime.org/book/data/exer.txt")
sample1_dat<-read.delim(sample1,header=T)
sample1_dat

m2<-ts(sample1_dat$m2, start=c(2001), frequency=1)
y<-ts(sample1_dat$y, start=c(2001), frequency=1)
rb<-ts(sample1_dat$rb, start=c(2001), frequency=1)
rc<-ts(sample1_dat$rc, start=c(2001), frequency=1)

n<-length(m2)
k=5

nm2<-m2[2:n]
ny<-y[2:n]
nrb<-rb[2:n]
nrc<-rc[2:n]
lagm2<-m2[1:n-1]
m<-length(nm2)

ur.lm<-lm(nm2~ny+lagm2+nrb+nrc)
ur.r2<-summary(ur.lm)$r.squared
ur.r2
```

```
r.lm<-lm(nm2~ny+lagm2)
r.r2<-summary(r.lm)$r.squared
r.r2

stargazer(ur.lm, r.lm, type="text")

Fstat1<-((ur.r2-r.r2)/2)/((1-ur.r2)/(m-k))

Fstat1

pval<-1-pf(Fstat1, 2, m-k)
pval

jointHo<-c("nrb","nrc")
linearHypothesis(ur.lm, jointHo)
```

제4장 가변수

b2-ch4-1.R

```
library(car)
library(stargazer)

data<-read.table("http://kanggc.iptime.org/book/data/dummy.txt", header=T)

y<-ts(data$GDP, start=c(1995,1), frequency=4)
c<-ts(data$CONSUME, start=c(1995,1), frequency=4)
n=length(c)

tr=1:n
d.log<-tr <= 12
d<-as.numeric(d.log)
```

```
dy<-d*y

y;c;d;dy

m1.lm<-lm(c~d+y)
summary(m1.lm)

m2.lm<-lm(c~y+dy)
summary(m2.lm)

m3.lm<-lm(c~d+y+dy)
summary(m3.lm)

stargazer(m1.lm, m2.lm, m3.lm, type="text", title="Regression Results of using Dummy Variable")

jointHo<-c("d","dy")
linearHypothesis(m3.lm, jointHo)

m4.lm<-lm(c~y, data=data, subset=(d==1))
summary(m4.lm)

m5.lm<-lm(c~y, data=data, subset=(d==0))
summary(m5.lm)

stargazer(m4.lm, m5.lm, type="text", title="Regression Results of separate period")
```

bi-ch4-2.R

```
library(stargazer)

data<-read.table("http://kanggc.iptime.org/book/data/income.txt", header=T)

age<-data$age
ed<- data$ed
```

```
gender<-data$gender
income<-data$income

age;ed;gender;income

high<-ifelse(data$ed==3, 1, 0)
college<-ifelse(data$ed>3, 1, 0)

ed;high;college

m1.lm<-lm(income~age+gender+high+college)
summary(m1.lm)

stargazer(m1.lm, type="text", title="Regression Results of using Dummy Variable")
```

제5장 자기상관

b2-ch5-1.R

```
library(stargazer)
library(lmtest)
library(orcutt)

sample1<-("http://kanggc.iptime.org/book/data/ar.txt")
sample1_dat<-read.delim(sample1,header=T)

consume<-ts(sample1_dat$consume, start=c(1995.1), frequency=4)
gdp<-ts(sample1_dat$gdp, start=c(1995.1), frequency=4)

ols.res<-lm(consume~gdp)
summary(ols.res)
```

```
res<-resid(ols.res)
lres<-append(res[1:24], 0, after=0)

res.t<-ts(res)
lres.t<-ts(append(res.t[1:24], 0, after=0))

plot(res,main="Residual plotting(u(t) vs. time)",xlab='time')
plot(lres,res,main="Residual plotting(u(t) vs. u(t-1))")
abline(h=0, v=0)

#DW test
dwtest(ols.res)

#LM test
n<-length(res)
lres.t<-ts(append(res.t[1:24], 0, after=0))
(res.aux<-lm(res.t~gdp+lres.t))
lm<-n*summary(res.aux)$r.squared
lm
(pchisq(lm, df=1, lower.tail=F))
```

b2-ch5-2.R

```
library(stargazer)
library(lmtest)
library(orcutt)

sample1<-("http://kanggc.iptime.org/book/data/ar.txt")
sample1_dat<-read.delim(sample1,header=T)

consume<-ts(sample1_dat$consume, start=c(1995.1), frequency=4)
gdp<-ts(sample1_dat$gdp, start=c(1995.1), frequency=4)

ols.res<-lm(consume~gdp)
summary(ols.res)
```

```
res<-resid(ols.res)
res.t<-ts(res)

#Cochrane-Orcutt estimation
coch.res<-cochrane.orcutt(ols.res)
coch.res

#GLS estimation
tconsume<-consume[2:25]-0.357123*consume[1:24]
tgdp<-gdp[2:25]-0.357123*gdp[1:24]
gls.res<-lm(tconsume~tgdp)
summary(gls.res)

stargazer(coch.res, gls.res, type="text")
```

제6장 이분산

b2-ch6-1.R

```
library(stargazer)

sample1<-("http://kanggc.iptime.org/book/data/korea(77-93)-1.txt")
sample1_dat<-read.delim(sample1,header=T)

gnp<-ts(sample1_dat$gnp, start=c(1977,1), frequency=4)
def<-ts(sample1_dat$def, start=c(1977,1), frequency=4)
m2<-ts(sample1_dat$m2, start=c(1977,1), frequency=4)

ols.res<-lm(m2~gnp+def)
summary(ols.res)
res<-resid(ols.res)
res.t<-ts(res)
```

```
plot(res,main="Residual plotting(u(t) vs. time)",xlab='time')

#LM test
n<-length(res)
res.t.sq<-res.t^2
gnp.sq<-gnp^2
def.sq<-def^2
(res.aux<-lm(res.t.sq~gnp+def+gnp.sq+def.sq+gnp*def))
lm<-n*summary(res.aux)$r.squared
lm
(pchisq(lm, df=5, lower.tail=F))
```

b1-ch6-2.R

```
library(stargazer)

sample1<-("http://kanggc.iptime.org/book/data/korea(77-93)-1.txt")
sample1_dat<-read.delim(sample1,header=T)

gnp<-ts(sample1_dat$gnp, start=c(1977,1), frequency=4)
def<-ts(sample1_dat$def, start=c(1977,1), frequency=4)
m2<-ts(sample1_dat$m2, start=c(1977,1), frequency=4)

ols.res<-lm(m2~gnp+def)
summary(ols.res)
res<-resid(ols.res)
res.t<-ts(res)

nm2<-m2/sqrt(gnp)
nc<-1/sqrt(gnp)
ngnp<-sqrt(gnp)
ndef<-def/sqrt(gnp)

wls.lm<-lm(nm2~nc+ngnp+ndef-1)
summary(wls.lm)
```

```
w0<-1/gnp

gls.lm<-lm(m2~gnp+def, weight=w0)
summary(gls.lm)

stargazer(ols.res, wls.lm, type="text")
stargazer(wls.lm, gls.lm, type="text")
```

제7장 시차분포모형

b2-ch7-1.R

```
library(dLagM)

sample1<-("http://kanggc.iptime.org/book/data/ar.txt")
sample1_dat<-read.delim(sample1,header=T)
sample1_dat
sample1.ts<-ts(sample1_dat, start=c(1995,1), end=c(2001,1), frequency=4)

cons<-sample1.ts[,1]
gdp<-sample1.ts[,2]

lcons1<-cons[1:24]
lgdp1<-gdp[1:24]

fm.res <- ivreg(cons[2:25] ~ lcons1+gdp[2:25] | lcons1+lgdp1)
summary(fm.res)

data(sample1_dat)
koyck.res<-koyckDlm(x = sample1_dat$gdp,y = sample1_dat$consume)
summary(koyck.res)
```

b2-ch7-2.R

```
library(dLagM)
library(dynlm)

rm(list=ls())

sample1<-("http://kanggc.iptime.org/book/data/ar.txt")
sample1_dat<-read.delim(sample1,header=T)
sample1.ts<-ts(sample1_dat, start=c(1995,1), end=c(2001,1), frequency=4)

cons<-sample1.ts[,1]
gdp<-sample1.ts[,2]

lgdp1<-gdp[4:24]
lgdp2<-gdp[3:23]
lgdp3<-gdp[2:22]
lgdp4<-gdp[1:21]

#summary(lm(cons[5:25]~gdp[5:25]+lgdp1+lgdp2+lgdp3+lgdp4))
#summary(lm(cons[5:25]~gdp[5:25]+gdp[4:24]+gdp[3:23]+gdp[2:22]+gdp[1:
21]))
#summary(dynlm(cons~gdp+L(gdp, 1)+L(gdp, 2)+L(gdp, 3)+L(gdp, 4)))
summary(dynlm(cons~L(gdp, 0:4)))

z1<-gdp[5:25]+lgdp1+lgdp2+lgdp3+lgdp4
z2<-lgdp1+2*lgdp2+3*lgdp3+4*lgdp4
z3<-lgdp1+4*lgdp2+9*lgdp3+16*lgdp4

(summary(lm.res<-lm(cons[5:25]~z1+z2+z3)))

beta0<-lm.res$coeff[2]
beta1<-lm.res$coeff[2]+lm.res$coeff[3]+lm.res$coeff[4]
beta2<-lm.res$coeff[2]+2*lm.res$coeff[3]+4*lm.res$coeff[4]
beta3<-lm.res$coeff[2]+3*lm.res$coeff[3]+9*lm.res$coeff[4]
beta4<-lm.res$coeff[2]+4*lm.res$coeff[3]+16*lm.res$coeff[4]

(coeff_fdl<-rbind(lm.res$coeff[1],beta0, beta1, beta2, beta3, beta4))
```

```
data(sample1_dat)
almon.res<-polyDlm(x = sample1_dat$gdp,y = sample1_dat$consume, q=4,
k=2, show.beta=T)
summary(almon.res)
```

부록 2

주요통계표

<표 1> 누적이항확률분포표

$$P(X \leq a) = \sum_{x=0}^{a} P(x)$$

(a) n=5

a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	.951	.774	.590	.328	.168	.078	.031	.010	.002	.000	.000	.000	.000
1	.999	.977	.919	.737	.528	.337	.187	.087	.031	.007	.000	.000	.000
2	1.00	.999	.991	.942	.837	.683	.500	.317	.163	.058	.009	.001	.000
3	1.00	1.00	1.00	.993	.969	.914	.813	.663	.472	.263	.081	.023	.001
4	1.00	1.00	1.00	1.00	.998	.990	.969	.922	.832	.672	.410	.226	.049

(b) n=10

a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	.904	.599	.349	.107	.028	.006	.001	.000	.000	.000	.000	.000	.000
1	.996	.914	.736	.376	.149	.046	.011	.002	.000	.000	.000	.000	.000
2	1.00	.988	.930	.678	.383	.167	.055	.012	.002	.000	.000	.000	.000
3	1.00	.999	.987	.879	.650	.382	.172	.055	.011	.001	.000	.000	.000
4	1.00	1.00	.998	.967	.850	.633	.377	.166	.047	.006	.000	.000	.000
5	1.00	1.00	1.00	.994	.953	.834	.623	.367	.150	.033	.002	.000	.000
6	1.00	1.00	1.00	.999	.989	.945	.828	.618	.350	.121	.013	.001	.000
7	1.00	1.00	1.00	1.00	.998	.988	.945	.833	.617	.322	.070	.012	.000
8	1.00	1.00	1.00	1.00	1.00	.998	.989	.954	.851	.624	.264	.086	.004
9	1.00	1.00	1.00	1.00	1.00	1.00	.999	.994	.972	.893	.651	.401	.096

(c) n=15

a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	.860	.463	.206	.035	.005	.000	000	.000	.000	.000	.000	.000	.000
1	.990	.829	.549	.167	.035	.005	.000	.000	.000	.000	.000	.000	.000
2	1.00	.964	.816	.398	.127	.027	.004	.000	.000	.000	.000	.000	.000
3	1.00	.995	.944	.648	.297	.091	.018	.002	.000	.000	.000	.000	.000
4	1.00	.999	.987	.836	.514	.217	.059	.009	.001	.000	.000	.000	.000
5	1.00	1.00	.998	.939	.722	.403	.151	.034	.004	.000	.000	.000	.000
6	1.00	1.00	1.00	.982	.869	.610	.304	.095	.015	.001	.000	.000	.000
7	1.00	1.00	1.00	.996	.950	.787	.500	.213	.050	.004	.000	.000	.000
8	1.00	1.00	1.00	.999	.985	.905	.696	.390	.131	.018	.000	.000	.000
9	1.00	1.00	1.00	1.00	.996	.966	.849	.597	.278	.061	.002	.002	.000
10	1.00	1.00	1.00	1.00	.999	.991	.941	.783	.485	.164	.013	.013	.000
11	1.00	1.00	1.00	1.00	1.00	.998	.982	.909	.703	.352	.056	.056	.000
12	1.00	1.00	1.00	1.00	1.00	1.00	.996	.973	.873	.602	.184	.184	.000
13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.995	.965	.833	.451	.451	.010
14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.995	.965	.794	.794	.140

(d) n = 20

a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	.818	.358	.122	.012	.001	.000	.000	.000	.000	.000	.000	.000	.000
1	.983	.736	.392	.069	.008	.001	.000	.000	.000	.000	.000	.000	.000
2	.999	.925	.677	.206	.035	.004	.000	.000	.000	.000	.000	.000	.000
3	1.00	.984	.867	.411	.107	.016	.001	.000	.000	.000	.000	.000	.000
4	1.00	.997	.957	.630	.238	.051	.006	.000	.000	.000	.000	.000	.000
5	1.00	1.00	.989	.804	.416	.126	.021	.002	.000	.000	.000	.000	.000
6	1.00	1.00	.998	.913	.608	.250	.058	.006	.000	.000	.000	.000	.000
7	1.00	1.00	1.00	.968	.772	.416	.132	.021	.001	.000	.000	.000	.000
8	1.00	1.00	1.00	.990	.887	.596	.252	.057	.005	.000	.000	.000	.000
9	1.00	1.00	1.00	.997	.952	.755	.412	.128	.017	.001	.000	.000	.000
10	1.00	1.00	1.00	.999	.983	.872	.588	.245	.048	.003	.000	.000	.000
11	1.00	1.00	1.00	1.00	.995	.943	.748	.404	.113	.010	.000	.000	.000
12	1.00	1.00	1.00	1.00	.999	.979	.868	.584	.228	.032	.000	.000	.000
13	1.00	1.00	1.00	1.00	1.00	.994	.942	.750	.392	.087	.002	.000	.000
14	1.00	1.00	1.00	1.00	1.00	.998	.979	.874	.584	.196	.011	.000	.000
15	1.00	1.00	1.00	1.00	1.00	1.00	.994	.949	.762	.370	.043	.003	.000
16	1.00	1.00	1.00	1.00	1.00	1.00	.999	.984	.893	.589	.133	.016	.000
17	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.996	.965	.794	.323	.075	.001
18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.992	.931	.608	.264	.017
19	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.999	.988	.878	.642	.182

(e) n = 25

a	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	.778	.277	.072	.004	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.974	.642	.271	.027	.002	.000	.000	.000	.000	.000	.000	.000	.000
2	.998	.873	.537	.098	.009	.000	.000	.000	.000	.000	.000	.000	.000
3	1.00	.966	.764	.234	.033	.002	.000	.000	.000	.000	.000	.000	.000
4	1.00	.993	.902	.421	.090	.009	.000	.000	.000	.000	.000	.000	.000
5	1.00	.999	.967	.617	.193	.029	.002	.000	.000	.000	.000	.000	.000
6	1.00	1.00	.991	.780	.341	.074	.007	.000	.000	.000	.000	.000	.000
7	1.00	1.00	.998	.891	.512	.154	.022	.001	.000	.000	.000	.000	.000
8	1.00	1.00	1.00	.953	.677	.274	.054	.004	.000	.000	.000	.000	.000
9	1.00	1.00	1.00	.983	.811	.425	.115	.013	.000	.000	.000	.000	.000
10	1.00	1.00	1.00	.994	.902	.586	.212	.034	.002	.000	.000	.000	.000
11	1.00	1.00	1.00	.998	.956	.732	.345	.078	.006	.000	.000	.000	.000
12	1.00	1.00	1.00	1.00	.983	.846	.500	.154	.017	.000	.000	.000	.000
13	1.00	1.00	1.00	1.00	.994	.922	.655	.268	.044	.002	.000	.000	.000
14	1.00	1.00	1.00	1.00	.998	.966	.788	.414	.098	.006	.000	.000	.000
15	1.00	1.00	1.00	1.00	1.00	.987	.885	.575	.189	.017	.000	.000	.000
16	1.00	1.00	1.00	1.00	1.00	.996	.946	.726	.323	.047	.000	.000	.000
17	1.00	1.00	1.00	1.00	1.00	.999	.978	.846	.488	.109	.002	.000	.000
18	1.00	1.00	1.00	1.00	1.00	1.00	.993	.926	.659	.220	.009	.000	.000
19	1.00	1.00	1.00	1.00	1.00	1.00	.998	.971	.807	.383	.033	.001	.000
20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.991	.910	.579	.098	.007	.000
21	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.998	.967	.766	.236	.034	.000
22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.991	.902	.463	.127	.002
23	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.998	.973	.729	.358	.026
24	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.996	.928	.723	.222

<표 2> 누적포아송확률분포표

$$P(X \le c) = \sum_{k=0}^{c} \frac{\mu^k e^{-\mu}}{k!}, \mu : \text{기댓값}$$

c / μ	0.02	0.04	0.06	0.08	0.10	0.15	0.20	0.25	0.30	0.35	0.40
0	0.980	0.961	0.942	0.923	0.905	0.861	0.819	0.779	0.741	0.705	0.670
1	1.000	0.999	0.998	0.997	0.995	0.990	0.982	0.974	0.963	0.951	0.938
2	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996	0.994	0.992
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

c / μ	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0	0.638	0.607	0.577	0.549	0.522	0.497	0.472	0.449	0.427	0.407	0.387
1	0.925	0.910	0.894	0.878	0.861	0.844	0.827	0.809	0.791	0.772	0.754
2	0.989	0.986	0.982	0.977	0.972	0.966	0.959	0.953	0.945	0.937	0.929
3	0.999	0.998	0.998	0.997	0.996	0.994	0.993	0.991	0.989	0.987	0.984
4	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.998	0.998	0.997
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

c / μ	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.368	0.333	0.301	0.273	0.247	0.223	0.202	0.183	0.165	0.150	0.135
1	0.736	0.699	0.663	0.627	0.592	0.558	0.525	0.493	0.463	0.434	0.406
2	0.920	0.900	0.879	0.857	0.833	0.809	0.783	0.757	0.731	0.704	0.677
3	0.981	0.974	0.966	0.957	0.946	0.934	0.921	0.907	0.891	0.875	0.857
4	0.996	0.995	0.992	0.989	0.986	0.981	0.976	0.970	0.964	0.956	0.947
5	0.999	0.999	0.998	0.998	0.997	0.996	0.994	0.992	0.990	0.987	0.983
6	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.997	0.995
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

c / μ	2.2	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.2
0	0.111	0.091	0.074	0.061	0.050	0.041	0.033	0.027	0.022	0.050	0.015
1	0.355	0.308	0.267	0.231	0.199	0.171	0.147	0.126	0.107	0.092	0.078
2	0.623	0.570	0.518	0.469	0.423	0.380	0.340	0.303	0.269	0.238	0.210
3	0.819	0.779	0.736	0.692	0.647	0.603	0.558	0.515	0.473	0.433	0.395
4	0.928	0.904	0.877	0.848	0.815	0.781	0.744	0.706	0.668	0.629	0.590
5	0.975	0.964	0.951	0.935	0.916	0.895	0.871	0.844	0.816	0.785	0.753
6	0.993	0.988	0.983	0.976	0.966	0.955	0.942	0.927	0.909	0.889	0.867
7	0.998	0.997	0.995	0.992	0.988	0.983	0.977	0.969	0.960	0.949	0.936
8	1.000	0.999	0.999	0.998	0.996	0.994	0.992	0.988	0.984	0.979	0.972
9	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.996	0.994	0.992	0.989
10	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.996
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

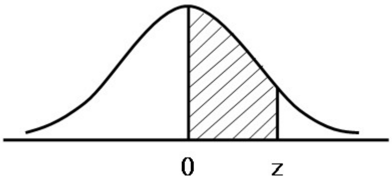
μ \ c	4.4	4.6	4.8	5.0	5.2	5.4	5.6	5.8	6.0	6.2	6.4
0	0.012	0.010	0.009	0.007	0.006	0.005	0.004	0.003	0.002	0.002	0.002
1	0.066	0.056	0.048	0.040	0.034	0.029	0.024	0.021	0.017	0.015	0.012
2	0.185	0.163	0.143	0.125	0.109	0.095	0.082	0.072	0.062	0.054	0.046
3	0.359	0.326	0.294	0.265	0.238	0.213	0.191	0.170	0.151	0.134	0.119
4	0.551	0.513	0.476	0.440	0.406	0.373	0.342	0.313	0.285	0.259	0.235
5	0.720	0.686	0.651	0.616	0.581	0.546	0.512	0.478	0.446	0.414	0.384
6	0.844	0.818	0.791	0.762	0.732	0.702	0.670	0.638	0.606	0.574	0.542
7	0.921	0.905	0.887	0.867	0.845	0.822	0.797	0.771	0.744	0.716	0.687
8	0.964	0.955	0.944	0.932	0.918	0.903	0.886	0.867	0.847	0.826	0.803
9	0.985	0.980	0.975	0.968	0.960	0.951	0.941	0.929	0.916	0.902	0.886
10	0.994	0.992	0.990	0.986	0.982	0.977	0.972	0.965	0.957	0.949	0.939
11	0.998	0.997	0.996	0.995	0.993	0.990	0.988	0.984	0.980	0.975	0.969
12	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.993	0.991	0.989	0.986
13	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.996	0.995	0.994
14	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

μ \ c	6.6	6.8	7.0	7.2	7.4	7.6	7.8	8.0	8.5	9.0	9.5
0	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
1	0.010	0.009	0.007	0.006	0.005	0.004	0.004	0.003	0.002	0.001	0.001
2	0.040	0.034	0.030	0.025	0.022	0.019	0.016	0.014	0.009	0.006	0.004
3	0.105	0.093	0.082	0.072	0.063	0.055	0.048	0.042	0.030	0.021	0.015
4	0.213	0.192	0.173	0.156	0.140	0.125	0.112	0.100	0.074	0.055	0.040
5	0.355	0.327	0.301	0.276	0.253	0.231	0.210	0.191	0.150	0.116	0.089
6	0.511	0.480	0.450	0.420	0.392	0.365	0.338	0.313	0.256	0.207	0.165
7	0.658	0.628	0.599	0.569	0.539	0.510	0.481	0.453	0.386	0.324	0.269
8	0.780	0.755	0.729	0.703	0.676	0.648	0.620	0.593	0.523	0.456	0.392
9	0.869	0.850	0.830	0.810	0.788	0.765	0.741	0.717	0.653	0.587	0.522
10	0.927	0.915	0.901	0.887	0.871	0.854	0.835	0.816	0.763	0.706	0.645
11	0.963	0.955	0.947	0.937	0.926	0.915	0.902	0.888	0.849	0.803	0.752
12	0.982	0.978	0.973	0.967	0.961	0.954	0.945	0.936	0.909	0.876	0.836
13	0.992	0.990	0.987	0.984	0.980	0.976	0.971	0.966	0.949	0.926	0.898
14	0.997	0.996	0.994	0.993	0.991	0.989	0.986	0.983	0.973	0.959	0.940
15	0.999	0.998	0.998	0.997	0.996	0.995	0.993	0.992	0.986	0.978	0.967
16	0.999	0.999	0.999	0.999	0.998	0.998	0.997	0.996	0.993	0.989	0.982
17	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.998	0.997	0.995	0.991
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.996
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

μ \ c	10.0	10.5	11.0	11.5	12.0	12.5	13.0	13.5	14.0	14.5	15.0
0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.003	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
3	0.010	0.007	0.005	0.003	0.002	0.002	0.001	0.001	0.000	0.000	0.000
4	0.029	0.021	0.015	0.011	0.008	0.005	0.004	0.003	0.002	0.001	0.001
5	0.067	0.050	0.038	0.028	0.020	0.015	0.011	0.008	0.006	0.004	0.003
6	0.130	0.102	0.079	0.060	0.046	0.035	0.026	0.019	0.014	0.010	0.008
7	0.220	0.179	0.143	0.114	0.090	0.070	0.054	0.041	0.032	0.024	0.018
8	0.333	0.279	0.232	0.191	0.155	0.125	0.100	0.079	0.062	0.048	0.037
9	0.458	0.397	0.341	0.289	0.242	0.201	0.166	0.135	0.109	0.088	0.070
10	0.583	0.521	0.460	0.402	0.347	0.297	0.252	0.211	0.176	0.145	0.118
11	0.697	0.639	0.579	0.520	0.462	0.406	0.353	0.304	0.260	0.220	0.185
12	0.792	0.742	0.689	0.633	0.576	0.519	0.463	0.409	0.358	0.311	0.268
13	0.864	0.825	0.781	0.733	0.682	0.628	0.573	0.518	0.464	0.413	0.363
14	0.917	0.888	0.854	0.815	0.772	0.725	0.675	0.623	0.570	0.518	0.466
15	0.951	0.932	0.907	0.878	0.844	0.806	0.764	0.718	0.669	0.619	0.568
16	0.973	0.960	0.944	0.924	0.899	0.869	0.835	0.798	0.756	0.711	0.664
17	0.986	0.978	0.968	0.954	0.937	0.916	0.890	0.861	0.827	0.790	0.749
18	0.993	0.988	0.982	0.974	0.963	0.948	0.930	0.908	0.883	0.853	0.819
19	0.997	0.994	0.991	0.986	0.979	0.969	0.957	0.942	0.923	0.901	0.875
20	0.998	0.997	0.995	0.992	0.988	0.983	0.975	0.965	0.952	0.936	0.917
21	0.999	0.999	0.998	0.996	0.994	0.991	0.986	0.980	0.971	0.960	0.947
22	1.000	0.999	0.999	0.998	0.997	0.995	0.992	0.989	0.983	0.976	0.967
23	1.000	1.000	1.000	0.999	0.999	0.998	0.996	0.994	0.991	0.986	0.981
24	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.995	0.992	0.989
25	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997	0.996	0.994
26	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.997
27	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998
28	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
29	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

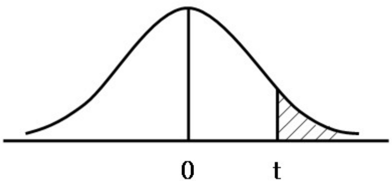
<표 3> 표준정규분포표

$$P(0 \leqq Z \leqq z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2 dx}$$



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

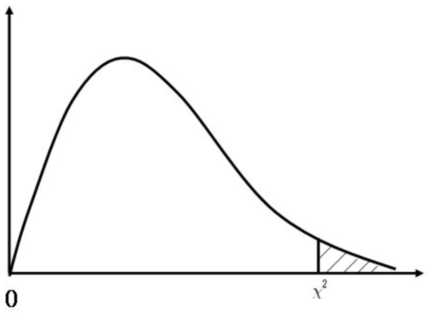
〈표 4〉 t-분포표



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 5> χ^2-분포표

$P(X \geqq \chi^2_\alpha) = \alpha$



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 6> F-분포표(5% 유의수준)

v1 \ v2	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

v1: 분자의 자유도 v2: 분모의 자유도.

<표 6(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

v1: 분자의 자유도 v2: 분모의 자유도.

<표 7> 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 7 (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 8〉 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
				$\widehat{\tau_\mu}$				
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
				$\widehat{\tau_\tau}$				
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

R 응용 및 계량경제분석

인 쇄 ‖ 2019년 1월 25일
발 행 ‖ 2019년 1월 25일

지은이 ‖ 강 기 춘

펴낸곳 ‖ 도서출판 신아문화사
찍은곳 ‖ 일신옵셋인쇄사

ISBN ‖ 978-89-97074-85-3 93320

본 교재는 2018년도 제주대학교 국립대학 육성사업에 의해 지원받았음