

# R commander를 활용한 회귀분석의 이해

## 머 리 말

정보화 기술(IT)의 발달로 누적되는 자료(data)의 양은 많아지고 그로부터 조사연구자에게 가치가 있는 정보를 캐내어 의사결정을 하거나 지식을 쌓게 하는 과학적인 수단인 통계적 방법에 대한 관심이 빅데이터 열풍과 함께 증가하고 있다.

회귀분석(regression analysis)은 부모의 키와 자녀의 키 사이 관계를 관찰한 결과 키는 무한정 커지거나 작아지는 것이 아니라 키 평균으로 돌아가려는 경향이 있다는 것을 발견하고 그 경향을 평균으로의 회귀라 부른데서 시작하여 이제는 여러 변수들간의 관계를 규명하고 모형화하여 추정이나 예측 등에 사용하는 통계적 방법으로 널리 사용되고 있다.

주어진 자료를 사용하여 회귀분석을 할 때 사용되는 계산도구였던 SAS나 SPSS 등과 같은 통계패키지들이 사용조건들을 점차 까다롭게 하여 소요되는 비용과 시간이 많아져 사용하기에 부담이 되고 있어 이에 대안으로 누구나 자유롭게 이용할 수 있고 프로그래밍 기능도 갖고 있어서 빅데이터 처리도구로도 인정받고 있는 통계패키지인 R을 사용하여 회귀분석을 시도하는 여러 저술들이 발간되었다. 이 책은 기본적으로 명령어 사용방식인 R을 사용한 저술이었던 'R을 활용한 회귀분석의 이해'를 좀 더 편리하게 메뉴 방식으로 사용할 수 있도록 제공되어 있는 라이브러리인 R commander를 활용하여 회귀분석의 논리를 이해하고 접근하기 어려워 보이는 R을 사용하는데 도움이 될 것을 바라면서 보완하여 서술하였으며 독자 여러분의 많은 조언을 통하여 개선될 수 있기를 기대합니다.

# 차례

1. 서론 .....	5
1.1 회귀분석의 개념 / 5	
1.2 기초적 통계이론과 R commander / 6	
1.3 회귀모형의 분류와 산점도 / 7	
2. 단순회귀분석 .....	12
2.1 상관분석 / 12	
2.2 단순회귀모형 / 13	
2.3 회귀계수의 추정 / 17	
2.4 분산분석(ANOVA) / 22	
2.5 원점을 지나는 회귀직선 / 26	
2.6 가중최소제곱법 / 28	
2.7 제곱합들의 기대값 계산 / 31	
3. 단순회귀모형에 대한 추론 .....	33
3.1 절편 $\beta_0$ , 기울기 $\beta_1$ 의 신뢰구간 / 33	
3.2 주어진 $x$ 에서 $E(Y x)=\mu_{y.x}$ 의 신뢰구간 / 36	
3.3 주어진 $x$ 에서 새로운 반응변수의 예측구간 / 38	
3.4 가설검정 /39	
3.5 적합결여검정 / 42	
4. 모형의 타당성과 잔차분석 .....	47
4.1 잔차분석 / 47	
4.2 오차의 자기상관 / 49	
4.3 모형의 변환 / 51	

5. 중회귀분석 .....	53
5.1 설명변수가 둘인 중회귀모형 / 53	
5.2 행렬의 기초 / 55	
5.3 이차형식의 분포 / 60	
5.4 중회귀모형의 추정 / 64	
5.5 추정량들의 성질 / 67	
5.6 분산분석 / 71	
5.7 회귀방정식의 정확도 / 74	
5.8 절편 없는 중회귀 모형 / 75	
5.9 회귀계수의 표준화 / 77	
5.10 예 / 79	
6. 중회귀모형에 관한 추론 .....	83
6.1 구간추정 / 83	
6.2 검정 / 90	
6.3 부분 F-검정과 축차 F-검정 / 92	
6.4 중회귀모형의 타당성 / 102	
7. 회귀모형의 진단과 처방 .....	106
7.1 다중공선성(multicollinearity) / 106	
7.2 주성분회귀와 능형회귀 / 110	
7.3 이상치 및 영향치(influential cases) / 116	
7.4 잔차그림 / 125	
8. 다항회귀 .....	126
8.1 독립변수가 하나인 다항회귀 모형 / 127	
8.2 독립변수가 둘인 다항회귀 모형 / 129	
8.3 2차 다항회귀에서 최고 또는 최적점 추정 / 130	

9. 가변수를 사용한 회귀분석 .....	133
9.1 하나의 질적변수 / 133	
9.2 상호작용항을 포함하는 모형 / 134	
9.3 구간별 선형회귀 / 136	
9.4 종속 가변수 / 138	
9.5 로지스틱 반응함수와 로지스틱 회귀분석 / 144	
10. 변수의 선택 .....	150
10.1 변수선택의 판정기준 / 150	
10.2 변수를 선택하는 방법 / 153	
부 록 .....	162
참고문헌 .....	169

## 1.1 회귀분석의 개념

회귀(regression)라는 단어를 사전에서 찾아보면 되돌아감으로 나와 있다. 회귀라는 단어가 학문적으로 사용된 것은 영국의 유전학자 Francis Galton(1822-1911)의 자료 분석적 연구에서부터인 것으로 알려져 있다. Galton은 처음에는 sweet pea에 대해서 나중에는 사람들에 대해서 부모의 키와 자녀의 키 사이 관계를 연구하기 위하여 조사하고 관찰한 결과 키는 무한정 커지거나 무한정 작아지는 것이 아니라 키 평균으로 돌아가려는 경향이 있다는 것을 발견하고 그 경향을 평균으로의 회귀(regression to mediocrity)라 하였다.

Galton이 수집한 자료를 통하여 부모 키와 자녀의 키 간에는 직선 관계가 있음을 발견하고 자녀의 키는 평균 키를 중심으로 회귀하려는 경향이 있음을 파악하는 경험적 연구를 통하여 회귀분석 개념을 도출하였다면, Karl Pearson(1903)은 회귀모형을 통한 수학적 전개를 하였다. Pearson은 1078 가족의 부자간 키를 인치 단위로 조사한 자료에서 아버지의 키(X)를 독립변수(independent variable)로 하고 아들의 키(Y)를 종속변수(dependent variable)로 하여 다음 선형함수 관계를 얻었다.

$$Y = 33.73 + 0.516X$$

이와 같이 회귀관계를 나타내는 직선을 회귀직선이라 부르고 회귀직선을 사용하여 분석하는 방법을 회귀분석(Regression Analysis)이라 하였다. 이와 같이 처음에는 회귀 분석이라는 용어는 키에 대한 회귀관계를 나타내는 분석을 뜻하였지만 이제는 일반적으로 변수들 간의 관계를 나타내는 분석을 의미한다. 예를 들어 곡물의 수확량이 투입된 비료의 양에 따라 어떻게 변하는지 설명하거나 예측 등을 할 때 또는 상품의 판매량이 투입된 판촉비에 따라 어떻게 변하는지 설명하거나 예측 등을 할 때 분석수단으로 사용될 수 있다. 즉, 회귀분석은 독립변수와 종속변수간의 관계를 추론하여 독립변수가 종속변수에 미치는 영향력을 알아보거나 독립변수의 변화에 따라 종속변수의 변화를 예측하기 위해서 사용하는 통계적 분석방법이다.

서로 연관된 변수들 중에서 다른 변수에 영향을 주는 변수를 독립변수라고 하며,

수학적 용어인 독립변수를 통계적 용어로 설명변수(explanatory variable)나 예측변수(predict variable) 또는 외생(exogenous)변수라 한다. 같은 맥락에서 독립변수에 의하여 영향을 받는 종속변수를 통계적인 용어로 반응변수(response variable) 또는 내생(endogenous)변수라 한다.

상관분석은 단순히 두 변수 사이의 관련 정도만을 분석하지만, 회귀분석은 두 변수 사이의 관계를 알아내고 이를 통해 한 변수로부터 다른 변수의 변화를 예측할 수 있는 통계적 분석방법이다. 두 변수에 대한 측정값이 실험에 의해 얻어진 것이라면 회귀분석이 인과관계에 대한 믿을 만한 정보를 준다고 생각할 수 있지만 관찰에 의한 것이라면 두 변수 외의 여러 변동요소들이 개입될 수 있기 때문에 해석에 주의를 기울여야 한다. 다음 절에서는 자료와 변수 등에 대한 기본 개념을 알아보자.

## 1.2 기초적 통계이론과 R commander

통계학은 연구 목적에 따라 수집한 자료를 분석하여 정보를 얻는 과학적 방법이다. 통계학(statistics)의 어원은 라틴어의 status(국가 또는 상태)라 하며, 원래 국가 또는 정치와 밀접한 관계가 있는 학문이었다. 고대의 통치자들은 국가의 재정 및 방위를 위하여 납세와 징병을 부과시켜야 했고, 이를 위해서 과세대장, 토지대장, 징병대장 등을 만들어 사용했으며, 이때부터 통계조사의 형태가 실시되었다고 볼 수 있다.

대부분의 학문과 마찬가지로 통계학도 발전과 변모를 하여 자료로부터 필요한 정보를 과학적으로 얻는 방법을 연구하여 지식사회를 여는 학문이 되었다.

### 1.2.1 변수와 자료

자료를 수집하기 전에 자료 분석 목적을 설정하고 자료를 통하여 어떤 정보를 얻을 것인지 명확하게 설정해야 한다. 키와 같이 통계처리의 대상이 되는 특성에 대한 자료는 변동성(variability)을 갖고 있으며 이러한 변동성을 갖는 특성(character)을 나타내는 기호를 변수(variable)라 한다. 이러한 변수의 변동성을 나타내는 측도의 하나가 산포의 측도인 분산(variance)이다.

통계처리는 변동이 있는 현상을 분석하는 것이 목적이며 변동이 큰 현상일수록 분석하여 얻어야 할 정보가 많다는 점을 이해하면 통계처리의 가장 기본적인 도구가 변수라는 것은 아무리 강조해도 지나침이 없다.

분석의 대상이 되는 자료는 대체로 행렬의 형태이며, 행은 개체(subject, case)에

해당되고 열은 변수(variable)로 구성되어 있다. 행의 각 원소(셀) 즉 각 개체에 대한 변수의 측정값을 관측치(observation)라 한다. 변수는 관심의 대상이 되는 개체의 특성(항목)을 나타낸다. 변수의 종류에 따라 적절한 통계 분석 방법이 다르므로 변수를 다음과 같이 분류할 필요가 있다.

### (1) 변수의 수학적 분류

이산형(discrete) : 취할 수 있는 값이 유한개이거나 셀 수 있는 무한개인 변수를 이산형 변수라 한다. 성별, 직업, 교통량, 나이 등이 여기에 해당한다.

연속형(continuous) : 취할 수 있는 값이 셀 수 없이 무한히 많은 변수를 연속형 변수라 한다. 즉 변수의 범위(range) 중 어떤 구간을 설정하더라도 측정치가 발생할 수 있는 경우로 키, 몸무게, IQ, 소득 등이 여기에 해당된다.

### (2) 변수의 통계적 분류

수치형 변수(metric, measurable, numerical, quantitative) : 실험 개체의 측정 가능한 특성을 나타내는 변수로 키, 몸무게, 평점, IQ, 교통량, 사망자 수 등이 그 예이다. 수치형 변수에는 연속형 변수인 계량형 변수와 이산형 변수인 계수형 변수가 있을 수 있다. 수치형 변수는 측정(measure)하는 척도(scale)에 따라 다음과 같이 분류하기도 한다.

- (i) 비(ratio)척도: 숫자 0 이 없는 것을 나타내는 물리적 의미를 갖는 변수
- (ii) 구간(interval)척도: 온도와 같이 숫자 0 이 섭씨일 때와 화씨일 때 나타내는 것이 다른 상대적인 의미를 가질 수 있는 변수

범주형 변수(non-metric, classified, categorical, qualitative) : 개체를 분류하기 위해 사용된 변수로 이산형만 가능하며 성별, 결혼여부 등이 그 예이고, 다음과 같이 분류하기도 한다.

- (i) 명목형(nominal)척도: 개체를 분류만 하는 변수로 성별, 결혼여부 등이 해당된다.
- (ii) 순서형(ordinal)척도: 순서를 갖는 변수로 소득수준(상, 중, 하), 리커트 (5점)척도 등이 해당된다. 소득과 같은 수치형 변수는 소득수준과 같이 순서형척도인 범주형 변수로 변환할 수 있다.

### (3) 변수의 시간에 따른 분류

데이터가 시간적 순서를 가지면 이를 시계열 자료(time series)라 하고, 그렇지 않



은 경우를 횡단면 자료(cross-section : 일정 시간에 한꺼번에 조사)라 한다. 분기별 경제 지표(환율, 수출량)나 기업의 연차별 자료(연도별 매출액), 주가 등이 시계열 자료에 해당된다.

#### (4) 변수의 인과 관계에 따른 분류

각 개체에 대해 여러 개의 변수가 측정될 경우 서로 독립적으로 변하지 않고 연관성을 갖고 변하는 경우가 많다. 특히 한 쪽은 영향을 주고 다른 쪽은 영향을 받아 변하는 비대칭적 관계를 인과(cause and effect)관계라고 하며 회귀분석 목적의 하나가 바로 이 인과관계를 잘 설명하는 통계모형을 적합시키는 것이다. 통계 모형의 인과 관계에서 원인이 되는(영향을 주는) 변수를 설명(explanatory)변수 혹은 독립(independent)변수라 하고 결과나 영향을 받는 변수를 종속(dependent)변수 혹은 반응(response)변수라 한다. 일반적으로 종속 변수는 Y, 설명 변수는 X 로 표시한다.

회귀분석은 기본적으로 계량형 변수들 간의 관계를 분석하며, 분산분석에서는 설명 변수가 범주형이며 요인(factor)이라 한다. 인과관계는 이론적, 경험적 타당성에 근거하여 데이터 수집 전에 설정되는 것이지 분석 후 설정되는 것은 아니다.

### 1.2.2 통계 소프트웨어 R과 R commander

수집된 자료로부터 정보를 얻는 통계처리를 할 때 필요한 계산에 사용할 통계 소프트웨어가 필수적이다. 프로그램이 단일 작업을 위하여 작성된 것이라면 소프트웨어는 다수 프로그램들이 하나의 목적을 위해 결합된 형태이다. 통계 소프트웨어가 개발되고 발전함에 따라 자료 정리 및 분석을 위한 시간이 줄어들어 보다 복잡하고 다양한 계산이 가능해지고 또한 GUI등 사용자가 편리하게 작업할 수 있는 도구와 환경이 주어지면서 통계 비전문가라도 손쉽게 통계 수치를 얻을 수 있게 되었다.

회귀분석이 가능한 통계 소프트웨어는 여러 가지이다. 예를 들면 사회과학 분야에서 주로 사용되는 SPSS(Statistical Package for Social Science), 경영 과학이나 QC 분야의 Minitab, 통계 그래픽에 강한 SYSTAT이나 STATGRAPHICS, 시뮬레이션과 수리적 계산에 편리한 S-plus 또는 R, 경제 자료 분석의 RATS, 종합적인 자료 분석 기능이 있는 SAS(Strategic Application System) 등이 있다. 또한 사무용 소프트웨어 EXCEL에도 기본적인 데이터 분석 기능과 add-on 기능을 이용한 보다 다양한 통계분석 기능이 포함되어 있다.

여기서는 누구나 자유롭게 사용할 수 있는 통계 소프트웨어 R을 이용하여 필요한 계산을 하고자 하며, 통계 소프트웨어 R을 처음 접하거나 사용하기에 어려움이 있는

경우 대부분의 기본적인 통계처리를 메뉴화 하여 R을 사용할 수 있도록 John Fox(2005)가 개발한 인터페이스인 R commander를 활용한다. R commander는 구성 요소의 하나인 스크립트 창에 해당 메뉴의 통계처리를 수행하는 R 문장을 보여주고 편집과 실행도 할 수 있도록 되어 있어 사용할수록 익숙해지고 편리하게 사용할 수 있다. R commander의 설치 및 사용법을 부록에서 간추려 설명하였으므로 참조하면 큰 어려움 없이 이해하고 사용할 수 있을 것이다.

### 1.2.3 자료 수집과 분석 방법

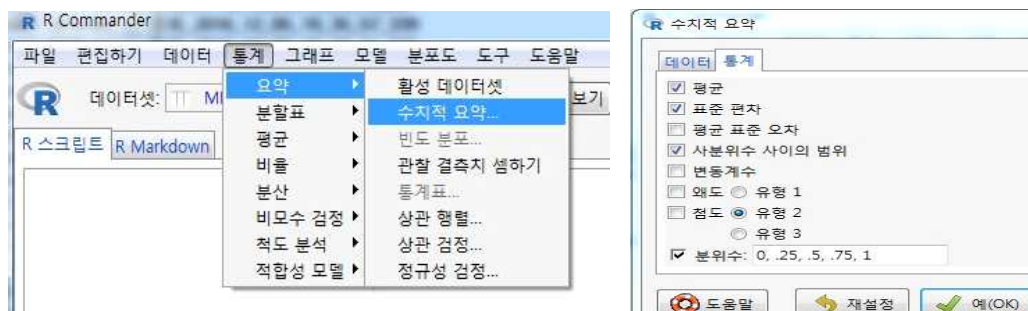
자료를 수집하는 방법으로 수집환경을 조사 연구자가 제어할 수 있는 ‘실험’과 그렇지 못한 ‘관찰’을 들 수 있다. 실험 자료에 의해 연구된 인과관계에 비하여 관찰 자료에 의해 연구된 인과관계는 잘못된 해석이나 결론을 도출할 위험이 크다는 것을 주의해야 한다.

조사연구 목적에 맞게 수집된 자료에서 필요한 정보를 얻기 위한 첫 단계는 도표와 요약 통계량을 통한 기술통계적인 분석이고 그로부터 얻은 정보를 기반으로 모형을 설정하고 모수에 대한 추정과 검정을 통한 추론통계적인 정보를 얻는다. 다음 1.3절의 [표 1.1]에 소개된 물질의 농도(conc)와 반응 속도(rate)에 대한 자료를 [부록 4.1]에서 소개한 방식으로 ‘MM1’이라는 이름의 데이터프레임인 R의 데이터셋으로 입력했을 때 R commander를 이용하여 처리할 수 있는 ‘통계’ 메뉴를 보자.

#### (1) 수치적 요약

메뉴 막대의 ‘통계’를 클릭하고 [‘통계’ -> ‘요약’ -> ‘수치적 요약’]을 선택하면 활성 데이터셋(여기서는 ‘MM1’)에 있는 변수들에 대해 요약된 정보가 [그림 1-1]과 같이 ‘출력물’ 창에 나타난다.

[그림 1.1]



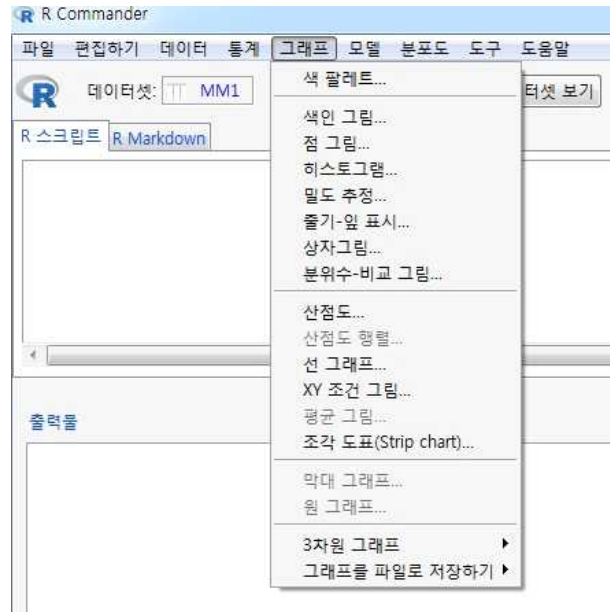
출력물											실행하기
<pre>&gt; numSummary(MM1[,c("conc", "rate")], statistics=c("mean", "sd", "IQR", + "quantiles"), quantiles=c(0,.25,.5,.75,1))</pre>											
	mean	sd	IQR	0%	25%	50%	75%	100%	n		
conc	44.21485	66.37880	33.87841	2.856829	6.89093	24.93581	40.76934	203.7842	8		
rate	65.38487	36.69939	66.61307	14.583420	29.69444	75.23542	96.30751	108.8837	8		

위치를 대표하는 값인 ‘평균’, 산포의 측도인 ‘표준 편차’, 집단화된 자료에서는 ‘평균 표준 오차’, 분포의 윤곽을 파악할 수 있는 사분위수와 최대값과 최소값으로 구성된 다섯 숫자 요약, 이상치의 영향을 줄인 산포의 기준으로 사용가능한 ‘사분위수 사이의 범위’, 평균 크기를 고려한 상대 표준편차인 ‘변동계수’, 대칭성과 밀집도에 대해 정규분포와 상대적인 비교를 할 때 사용되는 ‘왜도’와 ‘첨도’ 그리고 사용자가 정한 분위를 입력한 ‘분위수’ 등 수치적 요약된 정보를 알 수 있다.

## (2) 도표적 요약

메뉴 막대의 ‘그래프’를 클릭하면 활성 데이터셋(여기서는 ‘MM1’)에 있는 변수들에 대해 그릴 수 있는 다양한 형태의 그래프 메뉴가 [그림 1-2]와 같이 나타난다.

[그림 1.2]



‘색 팔레트...’에서 필요하다면 그래프의 색을 8개 색깔 중에서 설정하는 ‘색 팔레트

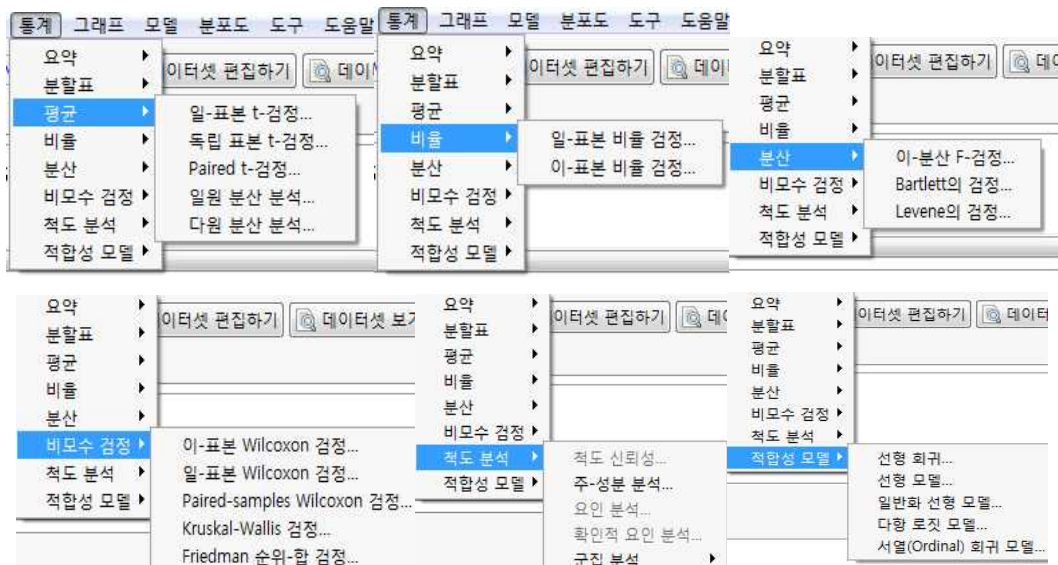
설정하기'를 하고 하나의 수치형 변수에 대한 일변량 분석을 할 때는 순서대로 그려주는 '색인 그림...', 자료가 많지 않을 때 '점 그림...', 수치형 변수를 범주화하여 분포를 보는 '히스토그램...', '밀도추정...', 정보를 잃지 않고도 히스토그램과 같이 분포를 보는 '줄기-잎 표시...', 다섯 숫자 요약을 나타낸 '상자그림...', 정규성이나 기타 분포의 적합성 검토에 사용될 수 있는 '분위수-비교 그림...' 등의 메뉴를 사용하며, 범주형 변수에 대해서 도수에 비례하는 높이를 나타내는 '막대 그래프...', 도수에 비례하는 면적을 나타내는 '원 그래프' 등의 메뉴를 사용한다.

하나 이상의 변수들에 대해 분석을 할 때 두 변수에 대한 상관관계를 직관적으로 보여주는 2차원 산점도인 '산점도...', 세 변수 이상에 대한 '산점도 행렬...', 시계열 자료를 그려주는 '선 그래프...', 두 변수에 대한 산점도를 다른 변수의 조건에 따라 그려주는 'XY 조건 그림...', 한 요인의 수준이나 2번째 요인과의 수준조합에서 평균을 나타내는 '평균그림...', 각 변수에 대한 일차원 산점도를 그려주는 '조각 도표(Strip chart)...', 3차원 산점도를 그려주는 '3차원 그래프' 등의 메뉴를 사용한다.

### (3) 통계적 추론

기술통계적인 초기 분석을 한 뒤 비교나 의사결정을 위한 통계적 추론 단계가 이어지고 R commander를 이용하여 처리 할 수 있는 통계적 추론 메뉴는 [그림 1-3]과 같다. 활성 데이터셋은 [부록 4.3]에서 설명한 방법을 사용하여 'car' 패키지의 데이터셋 'AMSSurvey'를 선택하였다.

[그림 1.3]



표본에 대한 자료를 구하고 그것을 이용하여 모수(모평균, 모비율, 모분산 등)를 추정하거나 모집단 또는 모수에 대한 가설 즉 통계적 가설을 검정하게 된다. 모수 추정에 사용되는 통계량을 추정량(estimator)라 하고, 가설검정에 사용되는 통계량을 검정 통계량(test statistic)이라 한다. 통계량들도 표본에 따라 변하는 변수이므로 표본분포(sampling distribution)를 갖는다.

모수를 추정하는 방법은 점추정(point estimation)과 구간추정(interval estimation)이 있다.

구간추정과 가설 검정을 위해서는 추정량과 검정통계량의 분포에 대한 정보가 필요하다. 분포를 알아야 확률을 이용하여 신뢰구간이나 유의확률 즉 확률값(p-value)등을 계산할 수 있다.

모집단의 분포를 모르는 경우, 크기가 작은 소표본에서 구한 검정통계량의 분포는 알 수 없으므로 분포무관하거나 비모수적(distribution free, nonparametric)인 방법을 사용하게 된다.

모집단 평균(‘일-표본 t-검정...’) 및 비율(‘일-표본 비율 검정...’, z-검정), 모분산( $\chi^2$ -검정)과 같은 하나의 변수들에 대한 분석을 하는 방법과 독립인 두 모집단 평균(‘독립 표본 t-검정...’) 및 비율 차이(‘이-표본 비율 검정...’, z-검정), 등분산성 또는 모분산 차이(‘이-분산 F-검정...’)와 같은 두 변수들에 대한 분석을 할 수 있다. 짝진 모평균 차이(‘Paired t-검정...’)는 짝 이룬 데이터의 차이를 구한 후 모집단 평균 차이에 대해 검정하면 된다. 한 요인의 수준별 모평균 비교(‘일원 분산 분석...’)와 여러 요인의 수준조합별 모평균 비교(‘다원 분산 분석...’)를 할 수 있다.

일반적으로 자료에 의하여 입증되어야 하는 가설을 대립가설( $H_1$ )로 놓고, 이미 알려져 있거나 자료에 의해 새로이 입증될 것이 없다는 가설을 귀무가설( $H_0$ )로 놓고 검정한다. 가설검정의 결과는 다음과 같이 정리한다.

판정 \ 실제	실제	
	$H_0$ 참	$H_0$ 거짓( $H_1$ 사실)
$H_0$ 채택	옳은 판정	제 2종 오류
$H_0$ 기각( $H_1$ 채택)	제 1종 오류	옳은 판정

제 1종 오류를 범할 확률의 최대허용치를 유의수준이라 하고,  $\alpha$ 로 나타내며 주로 5%, 1%, 10% 등 작은 수준으로 미리 정하고 검정한다. 검정통계량이 주어지면 정해진 유의수준  $\alpha$ 에 해당하는 기각역(rejection region)을 정하고 경계가 되는 기각치(critical value)와 비교하여 가설의 옳고 그름을 판정하는 것이 전통적인 검정방법이

다.

그러나 대부분의 통계소프트웨어에서는 표본(자료)에 의해 주어진 검정통계량의 값에 의해 유도한 제 1종 오류를 범할 확률의 추정값인 유의확률(significance probability)을 계산해주며, 계산된 유의확률이 정해진 유의수준  $\alpha$ 보다 작을 때 귀무가설을 기각하는 검정방법을 사용한다. 유의확률은 확률값(p-value)이라고도 부르며 귀무가설을 기각하기 위한 최소의 유의수준과 같은 값으로 생각할 수 있다.

### 1.3 회귀모형의 분류와 산점도

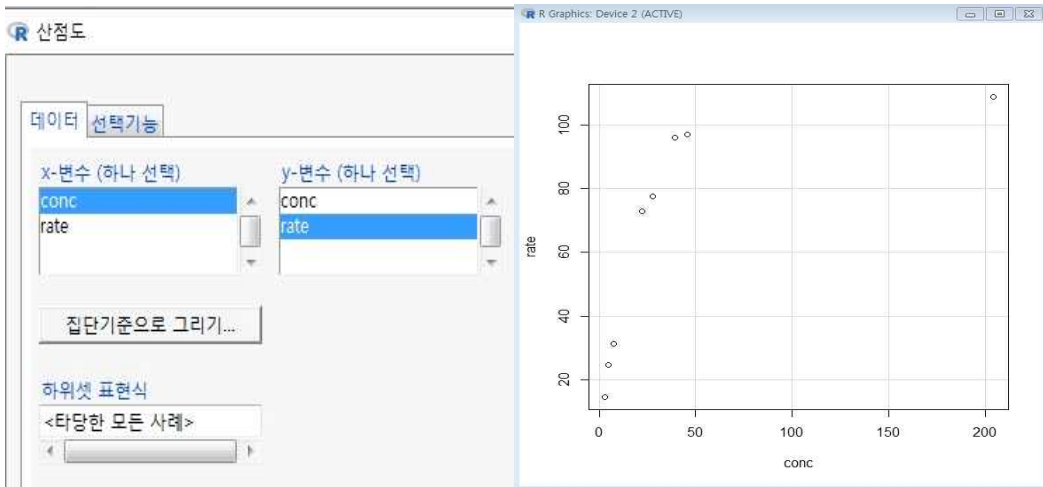
회귀분석은 종속변수와 함수관계인 독립변수의 개수가 하나인 경우는 단순회귀분석(Simple Regression Analysis), 독립변수의 개수가 둘 이상인 경우는 중회귀분석(Multiple Regression Analysis)이라 한다. 독립변수와 종속변수의 관계가 2차 이상의 다항식일 때 다항회귀분석(Polynomial Regression Analysis)이라 한다. 독립변수와 종속변수의 관계를 회귀계수들의 선형결합으로 나타낼 수 있을 때 선형회귀분석(Linear Regression Analysis)이라 하고, 그렇지 못할 때 비선형회귀분석(Nonlinear Regression Analysis)이라 한다.

두 변수간의 함수관계를 알기 위하여 우선 자료를 좌표평면위에 나타내는 산점도를 그려 보는 것이 필요하다. 예를 들어, Ritz and Streibig (2009)에서는 물질의 농도에 따른 화학반응의 속도를 측정한 [표 1.1] 자료에 대하여 산점도를 그려 보고 비선형인 Michaelis-Menten 함수와 그를 변형한 선형 회귀모형을 적합하는 방법을 비교 분석하였다.

[표 1.1] 물질의 농도에 따른 화학반응의 속도

농도(conc)	2.856829	5.005303	7.519473	22.101664	27.769976	39.198025	45.483269	203.784238
속도(rate)	14.58342	24.74123	31.34551	72.96985	77.50099	96.08794	96.96624	108.88374

산점도를 그리기 위하여 활성 데이터셋으로 ‘MM1’을 선택하고, 메뉴 막대의 ‘그래프’를 클릭하여 나타나는 부메뉴 중 ‘산점도’를 누르면 나오는 ‘산점도’ 창에서 ‘x-변수(하나 선택)’ 아래 칸에서 conc를 택하고, ‘y-변수 (하나 선택)’ 아래 칸에서 rate를 택한 다음 ‘선택기능’ 버튼을 눌러 ‘그림 선택기능’, ‘점 식별하기’, ‘그림 이름표와 점 정보’, ‘범례 위치’ 등을 지정하고 ‘예(OK)’를 누르면 별도의 ‘R Graphics’ 창에 [그림 1.4]와 같은 산점도 그림이 나타난다.



[그림 1.4] 물질의 농도에 따른 화학반응의 속도에 대한 산점도

이 산점도와 같이 나타나는 농도와 반응속도에 대한 자료를 적합하는 모형으로 잘 알려진 비선형인 Michaelis-Menten 함수 형태는 'rate'를  $y$ , 'conc'를  $x$ 로 나타낼 때 다음과 같다.

$$y = \frac{V_m x}{K + x}.$$

여기서,  $V_m$ 은 최대 반응속도를 나타내는 상수이고  $K$ 는 Michaelis 상수라 하며 자료로부터 추정해야 할 관심의 대상인 모수들이다. 약간의 산술적 조작을 하여 이 함수를 다음과 같이 Lineweaver-Burk 방정식 형태로 변형할 수 있다.

$$\frac{x}{y} = \frac{K}{V_m} + \frac{1}{V_m}x.$$

여기서, 종속변수를  $y_{trans} = \frac{x}{y}$ , 모수들을  $a = \frac{K}{V_m}$ 과  $b = \frac{1}{V_m}$ 로 변환하면 위 함수는  $y_{trans} = a + bx$ 와 같은 꼴의 선형모형이 되어 단순선형회귀분석을 적용할 수 있게 된다.

## 제 2 장

## 단순회귀분석

일반적으로 변수를 나타내는 기호는 대문자를 사용하지만 서로 관계를 갖고 있는 둘 또는 그 이상의 변수들 중에서 다른 변수에 영향을 주는 변수를 독립변수(independent variable)또는 설명변수(explanatory variable)라 하며 소문자  $x$ 로 표시하고, 독립변수에 의하여 영향을 받는 변수를 종속변수(dependent variable)또는 반응변수(response variable)라 하며 소문자  $y$ 로 표시하기로 한다.

통상적인 회귀분석에서 반응변수는 연속형이어야 하며 설명변수는 수치형이어야 하나 범주형일 경우도 더미 변수를 사용하는 등의 방법을 사용하여 적용할 수 있다. 반응변수가 계수형이거나 범주형일 경우 로지스틱 회귀분석이나 포아송 회귀분석과 같은 일반화 선형모형을 이용한 방법을 사용해야 적절한 정보를 얻을 수 있다.

### 2.1 상관분석

두 변수  $x$ 와  $y$  사이에 있어서 한쪽의 변화가 다른 쪽의 변화에 어떤 영향을 주는 경향이 있을 때  $x$ 와  $y$  사이에는 상관(correlation)이 있다고 한다. 산점도(scatter diagram)를 그려 보면 직관적으로 상관관계를 알 수 있음은 이미 설명한 바 있다. 두 변수 사이의 상관관계를 객관적으로 나타내기 위하여 산점도에 나타난 자료가 직선과 얼마나 가까워 보이는지를 말해주는 선형성의 측도인 상관계수(correlation coefficient)를 아래와 같이 정의하고, R commander 메뉴 막대의 ‘통계’를 클릭하여 [‘통계’ -> ‘요약’ -> ‘상관행렬...’ 또는 ‘상관 검정...’]을 선택하면 활성 데이터셋(여기서는 ‘MM1’)에 있는 변수들에 대해 상관계수들을 구하고 유의성을 검정할 수 있다.

(1) 모집단 상관계수(population coefficient of correlation)

$$\rho_{xy} = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sqrt{E[(x - \mu_x)^2]} \sqrt{E[(y - \mu_y)^2]}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(X, Y)}{\sqrt{V(X)} \sqrt{V(Y)}}$$



## (2) 표본상관계수(sample coefficient of correlation)

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)}}$$

이와 같이 정의된 (표본)상관계수의 성질은 다음과 같다.

- 1)  $r_{xy}$ 의 값은 항상  $-1 \leq r_{xy} \leq 1$  이다.
- 2)  $x$ 와  $y$  사이에 상관이 없으면  $r_{xy} = 0$ .
- 3)  $0 < r_{xy} \leq 1$  이면 양상관이라 한다. 즉  $x$ 가 증가할 때  $y$ 도 증가한다.
- 4)  $-1 \leq r_{xy} < 0$  이면 음상관이라 한다. 즉  $x$ 가 증가할 때  $y$ 는 감소한다.

[예2.1] [표1.1]의 자료에 대하여 R commander를 이용하여 상관계수를 구해 보자.

**R 상관 행렬**

변수 (두개 이상 선택)

conc  
rate

상관 유형

☒ Pearson product-moment

☐ Spearman 순위-순서

☐ Partial

사용할 관측치

☒ 모든 관측치

☐ 모든 쌍별(pairwise-complete) 관측치

☒ 쌍별(pairwise) 유의-값

**출력물**

```
> rcorr.adjust(MM1[,c("conc", "rate")],
, type="pearson", use="complete")
```

Pearson correlations:

	conc	rate
conc	1.0000	0.6688
rate	0.6688	1.0000

Number of observations: 8

Pairwise two-sided p-values:

	conc	rate
conc		0.0698
rate	0.0698	

Adjusted p-values (Holm's method)

	conc	rate
conc		0.0698
rate	0.0698	

**출력물**

```
> with(MM1, cor.test(conc, rate, alternative="two.sided", method="pearson"))
```

Pearson's product-moment correlation

data: conc and rate

t = 2.2033, df = 6, p-value = 0.06979

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.06793107 0.93350857

sample estimates:

cor

0.6687546

## 2.2 단순 회귀 모형

변수  $x$ 와  $y$  사이에 단순선형회귀모형을 적합 시키는 경우 통상적으로 다음과 같은 기본적인 가정을 한다.

- (1) 반응변수( $y$ )와 설명변수( $x$ ) 사이에 선형(직선)관계가 있다.
- (2) 설명변수( $x$ )의 주어진 값에서 반응변수( $y$ )는 평균이 주어진 값에 의존하나 분산은 주어진 값에 의존하지 않고 일정한 (조건부)분포를 갖는 (통상적으로 정규)확률변수이다.
- (3) 설명변수( $x$ )는 확률변수가 아닌 수학적 변수로 오차 없이 측정할 수 있다.
- (4) 오차항( $\epsilon$ )은 서로 독립이고 같은 분포를 따르는 (통상적으로 정규)확률변수이다. 따라서 오차항만을 확률변수로 수반하는 반응변수( $y$ )도 서로 독립인 확률변수이다.

이러한 가정 아래 단순회귀모형은 다음과 같이 나타낼 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad (2.1)$$

여기서

$y_i$ :  $i$  번째 측정된  $y$ 의 값

$\beta_0, \beta_1$ : 모집단 회귀계수이며  $\beta_0$ 는 절편(intercept)이 되고  $\beta_1$ 은 기울기(slope)로 설명변수의 주어진 값이 한 단위 증가할 때 반응변수의 증가량을 나타내며 미분계수가 된다.

$x_i$ :  $i$  번째 주어진 값

$\epsilon_i$ :  $i$  번째 측정된  $y_i$ 의 오차항으로 설명변수 또는 회귀직선에 의해 알아내지 못한 부분을 나타내고 서로 독립인 확률변수이다. 따라서  $Cov(\epsilon_i, \epsilon_j) = 0, i \neq j$  이다.

일반적으로 오차항의 분포는 독립이고  $\epsilon_i \sim N(0, \sigma^2)$ 라 가정한다. 즉, 오차항에 대하여 독립성(independent), 등분산성(homoscedasticity)과 정규성(normality)을 가정한다. 오차항에 대한 정규성 가정은 모수인 회귀계수들에 대한 통계적 추론을 하기 위하여 필요하다.

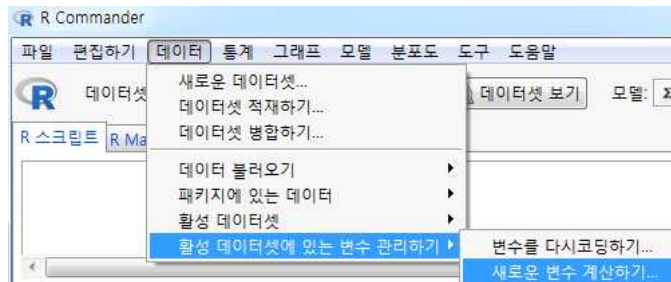
반응변수에 대한 평균과 분산은 다음과 같이 나타낼 수 있다.

$$\begin{aligned}
 E(y) &= E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) \\
 &= \beta_0 + \beta_1 x = \mu_{y|x} \\
 V(y) &= V(\beta_0 + \beta_1 x + \epsilon) = V(\epsilon) = \sigma^2
 \end{aligned}$$

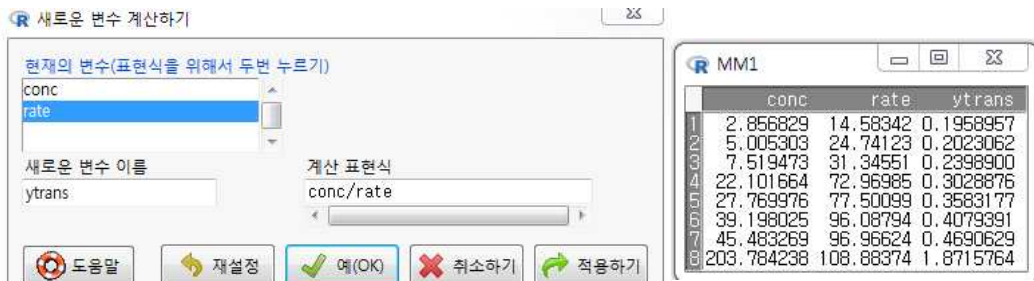
따라서,  $\epsilon \sim N(0, \sigma^2)$ 이면  $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ 이 된다.

회귀모형에서 가정한 ‘반응변수  $y$ 의 분산이  $x$ 의 값에 관계없이 일정함’을 등분산성(homoscedastic)이라 하며 그렇지 않은 경우를 이분산성(heteroscedastic)이라 한다.

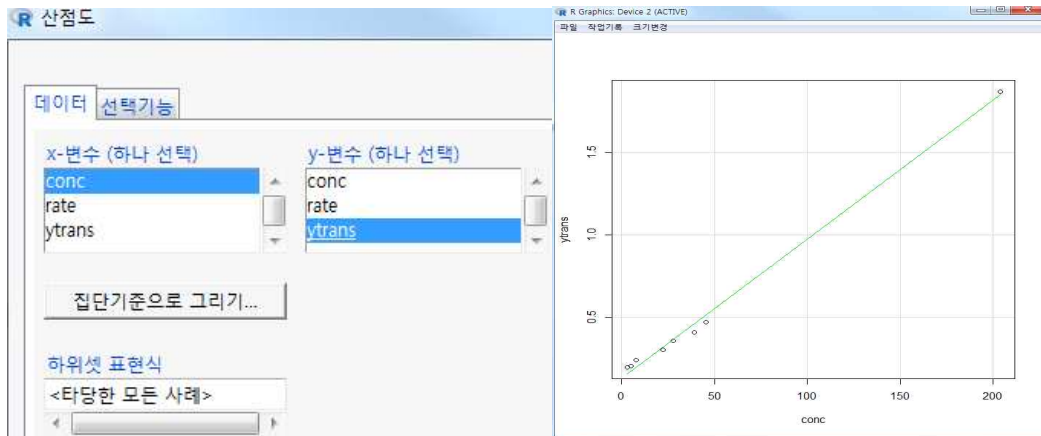
[예 2.2] [표 1.1]의 자료에 대하여 반응변수를  $y_{trans}=x/y$ 로 변환하여 유도한 선형회귀 모형에 대한 산점도와 적합 결과를 R commander를 사용하여 구해 보자.



메뉴 막대에서 ‘데이터’를 클릭하고 [‘데이터’ -> ‘활성 데이터셋에 있는 변수 관리하기’ -> ‘새로운 변수 계산하기...’]를 선택하면 ‘새로운 변수 계산하기’ 대화창이 나타나며 원하는 변수변환을 하면 된다.



메뉴막대에서 [‘그래프’ -> ‘산점도...’]를 선택하면 ‘산점도’ 대화창이 나타나며 원하는 형태의 산점도를 그릴 수 있다.



R 산점도

데이터 선택기능

x-변수 (하나 선택)  
 conc  
 rate  
 ytrans

y-변수 (하나 선택)  
 conc  
 rate  
 ytrans

집단기준으로 그리기...

하위셋 표현식  
 <타당한 모든 사례>

그림 선택기능

☐ x-변수 조금씩 움직이기  
☐ y-변수 조금씩 움직이기  
☐ x-축 로그  
☐ y-축 로그  
☐ 한계적인 상자그림  
☒ 최소-제곱 선  
☐ 평할선  
☐ Show spread

Span for smooth 50

☐ 농도 타원(들) 그리기

집중 수준들: .5, .9

점 식별하기

☐ 자동적으로  
☐ 마우스를 이용하여 쌍방향으로  
☒ 식별하지 마시오

식별할 점의 숫자 2

그림 이름표와 점 정보

x-축 이름표 <자동>

y-축 이름표 <자동>

그래프 제목 <자동>

문자 플롯팅 <자동>

Point(점) 크기 1.0

축 텍스트 크기 1.0

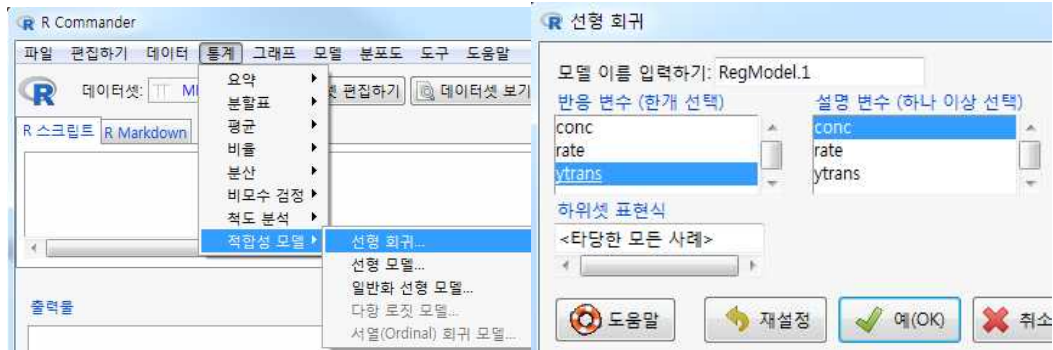
축-이름표 텍스트 크기 1.0

범례 위치

☒ Above plot  
☐ 왼쪽 위  
☐ 오른쪽 위  
☐ 왼쪽 아래  
☐ 오른쪽 아래

도움말 재설정 예(OK) 취소하기 적용하기

메뉴 막대에서 ['통계' -> '적합성 모델' -> '선형 회귀...']를 선택하면 '선형 회귀' 대화 창이 나타나며 '반응 변수 (한개 선택)'와 '설명 변수 (하나 이상 선택)'를 택하고 '예(OK)'를 누르면 '출력물' 창에서 lm() 함수를 사용하여 회귀분석한 결과를 볼 수 있다.



#### 출력물

```
> RegModel.1 <- lm(ytrans~conc, data=MM1)
> summary(RegModel.1)

Call:
lm(formula = ytrans ~ conc, data = MM1)

Residuals:
    Min       1Q   Median       3Q      Max
-0.055792 -0.024542  0.006248  0.029475  0.042963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1335963  0.0179213   7.455  3e-04 ***
conc        0.0084222  0.0002351  35.823 3.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04129 on 6 degrees of freedom
Multiple R-squared:  0.9953, Adjusted R-squared:  0.9946
F-statistic: 1283 on 1 and 6 DF, p-value: 3.155e-08
```

'출력물' 창에서 보면 적합된 선형모형의 절편(a)과 기울기(b)에 대한 최소제곱추정치가 각각 0.1335963과 0.0084222로 주어져 있고 각각에 대한 유의확률들이 모두 0.001보다 작아 매우 유의하다는 유의성코드 "\*\*\*"가 붙어 있다. 따라서 비선형함수에서의 모수인  $V_m$  과  $K$ 는 각각 다음과 같이 추정할 수 있다.

$$V_m = \frac{1}{b} = \frac{1}{0.0084222} = 118.7338, \quad K = V_m \times a = 118.7338 \times 0.1335963 = 15.8624.$$

주어진 결과들을 이용하는 것이 목적이라면 자료 분석 단계로 넘어 가겠지만 위와 같은 결과들이 어떻게 도출되었는지 알고 싶다면 이론적인 전개에 대한 이해가 필요하므로 알아보기로 한다.

## 2.3 회귀계수의 추정

$\beta_0$ 와  $\beta_1$ 이 모집단 회귀계수인 모수들(parameters)이므로 자료로부터  $\beta_0$ 와  $\beta_1$ 을 추정하여야 한다. 추정방법으로 최소제곱법(Least Squares method)와 최우추정법(Maximum Likelihood Estimation)이 있고, 추정량에 대한 기호는  $b_0, b_1$  또는  $\hat{\beta}_0, \hat{\beta}_1$  등을 사용한다. [예 2.1]의 ‘출력물’ 창의 ‘Coefficients:’ 줄에 추정값이 계산되어 있다.

모형 (2.1)을 벡터와 행렬을 사용하여 다음과 같이 나타낼 수 있으며 행렬의 성질에 대해서는 5장에서 다루기로 한다.

$$y = X\beta + \epsilon \quad (2.2)$$

여기서

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

### 2.3.1 최소 제곱법(Ordinary Least Squares estimation, OLS)

최소제곱법은 모수들을 추정할 때, 관측치가 모형에서 벗어난 정도를 나타내는 오차  $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ 가 가능한 작아지도록 하려는 방법이다. 오차의 절대값들의 합이 최소가 되도록 추정하는 방법도 생각할 수 있겠지만 통상적으로 오차의 제곱합(Sum of Squares)

$$Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

을 최소화시키는 추정량  $\widehat{\beta}_0, \widehat{\beta}_1$  즉  $b_0, b_1$ (최소제곱추정량)을 찾는 최소제곱법이 주로 사용된다.

최소제곱법은 관찰치  $y_i, i = 1, 2, \dots, n$ 와 모집단 회귀식  $\mu_{y|x} = \beta_0 + \beta_1 x$ 와의 차이인  $\epsilon_i$ (오차)의 제곱합을 최소로 하는 표본회귀식  $\hat{y}_i = b_0 + b_1 x_i$ 을 찾는 방법이며 오차제곱합의 최소값이 잔차제곱합과 같게 된다는 다음 조건을 만족시킨다.

$$\begin{aligned} \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \\ &= \sum_{i=1}^n \epsilon_i^2 \end{aligned}$$

### 2.3.2 정규방정식 (Normal equation)

최소제곱추정량을 구하기 위하여 오차제곱합  $Q$ 를  $\beta_0$ 와  $\beta_1$ 에 대해 편미분한다.

$Q = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$  이므로  $Q$ 를  $\beta_0$ 와  $\beta_1$ 에 대해 편미분하면

$$\begin{aligned} \frac{\delta Q}{\delta \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\delta Q}{\delta \beta_1} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i . \end{aligned}$$

이 결과를 0으로 놓고 정리한 연립방정식을 정규방정식이라 한다.

$$\begin{aligned} \sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i &= 0 \\ \sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 &= 0 \end{aligned}$$

또는

$$\begin{aligned} \beta_0 n + \beta_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned}$$

정규방정식을 행렬을 사용하여 나타내면

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{bmatrix}$$

이므로, 이 정규방정식의 해는

$$\begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{bmatrix} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{bmatrix}$$

과 같이 나타낼 수 있다. 여기서  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ ,  $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$  이라 하면 이 정규방정식의 해는 산술적 계산을 통해 다음과 같이 나타낼 수 있다.

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

이와 같은 정규방정식의 해  $b_0$ 와  $b_1$ 이 모수  $\beta_0, \beta_1$ 의 최소제곱추정량이고, 다음과 같이 R을 이용하여 계산할 수 있다.

```
> attach(MM1)          # 데이터셋 'MM1'의 변수를 작업공간으로 가져오기
> sxx=sum((x-mean(x))*(x-mean(x)))
> syy=sum((ytrans-mean(ytrans))*(ytrans-mean(ytrans)))
> sxy=sum((x-mean(x))*(ytrans-mean(ytrans)))
> c(sxx, syy, sxy)
[1] 0.085541334 0.005769046 0.022206754
> b1=sxy/sxx; b0=mean(y)-b1*mean(x)
> c(b0, b1)
[1] 0.004303145 0.259602617
```



### 2.3.3 최우추정법(Maximum Likelihood Estimation)

최우추정법은 모수가 미지인 경우 결합확률밀도함수 또는 확률밀도함수들의 곱인 우도함수(likelihood function)를 최대로 하는 추정량을 찾는 방법이다. 확률표본(random sample)  $\{X_1, X_2, \dots, X_n\}$ 이 모수가  $\theta$  인 확률밀도함수  $f(x; \theta)$ 를 갖는 모집단에서 추출되었다면 우도함수는  $L(\theta; x_1, x_2, \dots, x_n) = \prod f(x_i; \theta)$  이고 이 우도함수를 최대로 하는  $\tilde{\theta}$ 이 최우추정량(MLE, Maximum Likelihood Estimator)이다.

최우추정법을 사용하기 위해서는 분포를 알아야 해당하는 확률밀도함수를 곱하여 우도함수를 구할 수 있다. 회귀모형에 대한 기본가정에서 오차항의 분포가  $\epsilon_i \sim iid N(0, \sigma^2)$ 으로 주어지고 따라서 확률밀도함수는

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\epsilon_i^2}{2\sigma^2}\right\}$$

와 같이 주어지므로 우도함수  $L(\beta_0, \beta_1, \sigma^2 | \epsilon_1, \epsilon_2, \dots, \epsilon_n)$ 는 결합확률밀도함수

$$\begin{aligned} f(\epsilon_1, \epsilon_2, \dots, \epsilon_n; \beta_0, \beta_1, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum_{i=1}^n \epsilon_i^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\} \end{aligned}$$

와 같다.

즉, 우도함수는 결합확률밀도함수와 같은 형태의 함수인데 전자는 모수의 함수이고 후자는 확률변수의 함수인 것이 다른 점이다.

우도함수  $L(\beta_0, \beta_1, \sigma^2 | \epsilon_1, \epsilon_2, \dots, \epsilon_n)$ 를 최대로 하는  $\beta_0$ 와  $\beta_1$ 을 구하는 방법은 로그우도함수  $\ln L(\beta_0, \beta_1, \sigma^2 | \epsilon_1, \epsilon_2, \dots, \epsilon_n) \propto -\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = -Q$ 를 최대로 하는  $\beta_0$ 와  $\beta_1$ 을 구하는 방법과 동등하므로  $Q$ 를 최소로 하는 방법인 최소제곱법과 같아진다. 따라서, 회귀계수  $(\beta_0, \beta_1)$ 에 대한 최우추정량  $(\tilde{\beta}_0, \tilde{\beta}_1)$ 과 최소제곱추정량  $(b_0, b_1)$ 은 동일하게 구해진다.

### 2.3.4 적합된 회귀직선의 성질

최소제곱법으로 구한 회귀계수  $\beta_0, \beta_1$  의 최소제곱추정량  $b_0, b_1$  은 최량선형불편 추정량( best linear unbiased estimator : BLUE)이다. 즉 선형불편추정량들 중에서 분산이 가장 작아 안정적인 것이다. 이 최소제곱추정량을 사용하여 적합된 표본회귀식 (fitted model)은  $\hat{y} = b_0 + b_1x$  과 같이 나타내고, 주어진  $x_i$ 에서 실제값( $y_i$ )과 추정치( $\hat{y}_i$ )의 차이를 잔차(residual)라 하며  $\hat{e}_i = e_i = y_i - \hat{y}_i$  과 같이 나타낸다. [예 2.2]의 ‘출력물’ 창에서 “Residuals:” 줄에 잔차에 대한 정보가 주어져 있다. 이러한 잔차의 성질에는 다음과 같은 것들이 있다.

1) 잔차  $e_i = y_i - \hat{y}_i$  의 합은 0 이다. 즉  $\sum e_i = 0$  이다.

증명은 정규방정식으로부터 쉽게 유도할 수 있다. 즉

$$\sum e_i = \sum (y_i - \hat{y}_i) = \sum (y_i - b_0 - b_1x_i) = \sum y_i - nb_0 - b_1 \sum x_i = 0.$$

```
> mm.res=resid(RegModel.1)
```

```
> summary(mm.res)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.727e-03	-4.763e-05	1.363e-04	-6.389e-20	4.264e-04	8.333e-04

2) 잔차제곱합  $\sum e_i^2$  이 오차제곱합 중 최소가 됨은 최소제곱법의 정의상 당연하다.

3)  $\sum_{i=1}^n x_i e_i = 0$ 이다. 증명은 정규방정식으로부터 유도할 수 있다. 즉

$$\sum_{i=1}^n x_i e_i = \sum_{i=1}^n x_i (y_i - \hat{y}_i) = \sum_{i=1}^n x_i (y_i - b_0 - b_1x_i) = \sum_{i=1}^n x_i y_i - nb_0 - b_1 \sum_{i=1}^n x_i^2 = 0.$$

```
> sum(x*mm.res)
```

```
[1] 8.682088e-20
```

4)  $\sum_{i=1}^n \hat{y}_i e_i = 0$ 이다. 즉

$$\sum_{i=1}^n \hat{y}_i e_i = \sum_{i=1}^n (b_0 + b_1x_i) e_i = b_0 \sum_{i=1}^n e_i + b_1 \sum_{i=1}^n x_i e_i = 0.$$

```
> mm.pred=predict(RegModel.1)
> sum(mm.pred*mm.res)
[1] 1.905824e-21
```

## 2.4 분산분석(ANALYSIS Of VARIANCE, ANOVA)

앞에서 주어진 자료로부터 회귀모형을 적합시키는 방법을 알아보았으나 정확도 (precision)에 대한 평가를 하지는 않았다. 정확도를 평가하는 측도로 사용되는 분산 분석표에서의 F-검정, 결정계수, 추정량의 표준오차등을 알아보자.

### 2.4.1 분산분석

하나의 측정치  $y_i$ 와  $y_i$ 들의 평균  $\bar{y}$ 의 차이인 총편차 ( $y_i - \bar{y}$ )를 다음과 같이 회귀선에 의하여 설명되지 않는 편차인 잔차 ( $y_i - \hat{y}_i$ )와 회귀선에 의하여 설명되는 편차인 회귀편차 ( $\hat{y}_i - \bar{y}$ )의 합으로 나타낼 수 있다.

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

↪총편차            ↪잔차            ↪회귀편차

식의 양변을 제곱해서 더하면

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

이 되며, 오른쪽의 마지막 항은

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum e_i(\hat{y}_i - \bar{y}) = \sum e_i \hat{y}_i - \bar{y} \sum e_i = 0$$

이므로

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

이 된다. 여기서 총편차의 제곱합  $\sum (y_i - \bar{y})^2$ 은 총변동(total variation)이라 부르며, SST(Total Sum of Squares)로 나타낸다. 다른 기호로  $S_{yy} = \sum (y_i - \bar{y})^2$ 와 같이 나타내고 총제곱합이라 부르기도 한다. 따라서 총제곱합은 잔차의 제곱합 SSE(Sum of Squares due to residual Errors)와 회귀편차의 제곱합 SSR(Sum of Squares due to Regression)으로 분해됨을 알 수 있다.

총제곱합은 수정항을  $CT = n(\bar{y})^2$ 라는 기호로 나타낼 때, 다음과 같이 계산할 수 있다.

$$SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 = \sum y_i^2 - CT = S_{yy}$$

회귀직선에 의해 설명되는 변동인 회귀제곱합은  $b_0 = \bar{y} - b_1\bar{x}$ 임을 이용하여

$$\begin{aligned} SSR &= \sum (\hat{y}_i - \bar{y})^2 = \sum (b_0 + b_1x_i - \bar{y})^2 = \sum (\bar{y} - b_1\bar{x} + b_1x_i - \bar{y})^2 \\ &= \sum (b_1(x_i - \bar{x}))^2 = b_1^2 \sum (x_i - \bar{x})^2 = b_1^2 S_{xx} = \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

로 계산할 수 있으며, 회귀직선에 의해 설명 되지 않는 변동인 잔차제곱합은

$$SSE = \sum_{i=1}^n (y_i - \hat{y})^2 = \sum_{i=1}^n (y_i - b_0 - b_1x_i)^2$$

이지만  $SSE = SST - SSR$ 로 계산하는 것이 편리하다.

```
> sst=syy; ssr=sxy*sxy/sxx; sse=sst-ssr
> c(sst, ssr, sse)
[1] 5.769046e-03 5.764932e-03 4.114128e-06
```

## 2.4.2 자유도

통계학에서 자유도(degrees of freedom)는 제곱합과 관련된 카이제곱 분포의 모수로서 나타난다. 서로 독립인 k개 표준정규확률변수들의 제곱을 합한 통계량은 자유도 k인 카이제곱분포를 따른다는 사실로부터 자유도는 통계량에 포함된 서로 독립적으로 정보를 줄 수 있는 변수의 개수를 나타낸다고 유추할 수 있다. 제곱합은 변동의 크기를 측정하는 산포의 측도인 분산과 동등한 역할을 한다고 본다. 산포의 측도인 제곱

합을 구성하는 독립적인 편차의 개수를 그 제곱합의 자유도라 생각할 수 있다.

$SST$ 를 구성하는 총편차  $(y_i - \bar{y})$ 는  $n$ 개이지만, 모두 합하면 0이라는 제약조건이 하나 있어서 독립적인 편차의 개수인 자유도는  $(n-1)$ 이 된다. 총편차는  $n$ 개이지만 전체평균이라는 모수를 추정하느라 제약조건이 하나 붙었기 때문에 자유도가 1만큼 줄었다고 이해할 수 있다.

$SSR$ 을 구성하는 회귀편차  $(\hat{y}_i - \bar{y})$ 도  $n$ 개이지만,  $(\hat{y}_i - \bar{y}) = b_1(x_i - \bar{x})$ 이므로  $n$ 개의 회귀편차들 모두 하나의 확률변수  $b_1$ 에 각각 주어진 상수  $(\hat{x}_i - \bar{x})$ 만 곱한 것이어서 독립적인 편차는 한 개이고 따라서 자유도는 1이다.

$SSE$ 를 구성하는 잔차  $(y_i - \hat{y})$ 도  $n$ 개이지만, 모두 합하면 0이고  $x_i$ 를 곱해서 합해도 0이라는 두가지 제약조건이 있어서 자유도는  $(n-2)$ 가 된다. 잔차는  $n$ 개이지만 모수  $\beta_0, \beta_1$ 를 추정하느라 제약조건이 두 가지 붙었기 때문에 자유도가 2만큼 줄었다고 이해할 수 있다.

### 2.4.3 단순회귀의 분산분석표(ANOVA Table)

주어진 자료에 대해 적합된 회귀직선의 유의성(significance)을 측정하는 도구의 하나로 회귀제곱합  $SSR$ 이 잔차제곱합  $SSE$ 보다 얼마나 큰가를 나타내는 통계량  $F_0 = MSR/MSE$ 를 다음과 같은 분산분석표를 만들어 계산할 수 있다.

[표 2.1] 단순회귀의 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$	$F_\alpha$
회귀	SSR	1	MSR	MSR/MSE	$F_\alpha(1, n-2)$
잔차	SSE	$n-2$	MSE		
계	SST	$n-1$			

분산분석표에서 제곱평균은 제곱합을 자유도로 나눈 것이다. 회귀제곱평균 MSR과 잔차제곱평균 MSE의 기대값은 다음과 같다. 이에 대한 증명은 연습으로 다룬다.

$$E(MSR) = \sigma^2 + \beta_1^2 S_{xx}$$

$$E(MSE) = \sigma^2$$

따라서 MSE는  $\sigma^2$ 의 불편추정량이고, 분산분석표에서 F-검정통계량에 의해 검정하려는 귀무가설은  $H_0 : \beta_1 = 0$ 임을 알 수 있다. 분산분석표에서  $F_0 = \text{MSR}/\text{MSE}$ 의 값이 클수록  $\beta_1$ 이 0과 차이가 크다는 증거가 되어 귀무가설  $H_0 : \beta_1 = 0$ 을 기각하고 적합한 회귀직선이 유의하다고 판정할 수 있다. 유의수준  $\alpha$ 인 경우, 귀무가설  $H_0 : \beta_1 = 0$ 을 기각하는 기준은 ' $F_0 > F_\alpha(1, n-2)$ ', 또는 p-값 기준으로 ' $\text{Pr}(F) < \alpha$ '이다.

```
> anova(RegModel.1)
Analysis of Variance Table
Response: ytrans
      Df Sum Sq   Mean Sq    F value    Pr(>F)
x       1  0.0057649  0.0057649    7006.3 4.612e-09 ***
Residuals  5  0.0000041  0.0000008
```

```
> sse=sum(mm.res*mm.res)
[1] 4.114128e-06
> n=length(mm.res)
[1] 7
> mse=sse/(n-2)
[1] 8.228256e-07
```

## 2.4.4 결정계수

회귀제곱합 SSR이 잔차제곱합 SSE보다 얼마나 큰가를 나타내는 또 다른 통계량으로 총변동 중에서 회귀직선에 의하여 설명되는 비율

$$R^2 = \text{SSR}/\text{SST} = 1 - \text{SSE}/\text{SST}$$

을 생각할 수 있다.  $R^2$ 는 결정계수(coefficient of determination)라 부르며,  $0 \leq R^2 \leq 1$ 을 만족한다. 극단적으로 모든 자료들이 회귀직선 위에 놓인다면 잔차제곱합  $\text{SSE} = 0$ 이므로  $R^2 = 1$ 이다. 결정계수  $R^2 = 0$ 인 경우는 회귀제곱합  $\text{SSR} = \sum (\hat{y}_i - \bar{y})^2 = 0$ 이라는 것이고 따라서 모든  $\hat{y}_i - \bar{y} = 0$  즉  $\hat{y}_i = \bar{y}$ 이므로 적합한 회귀직선의 기울기가 0이다. 일반적으로  $R^2$ 의 값은 0과 1 사이에 있으며, 결정계수의 값이 1에 가까울수록 적합한 회귀직선의 유용성이 높다고 판단한다.

결정계수는 총변동 중에서 회귀직선에 의하여 설명되는 변동이 기여하는 비율이라는 뜻에서 회귀직선의 기여율이라 부르기도 한다.

```
> rsquare=ssr/sst; rsquare
[1] 0.9992869
```

### 2.4.5 추정량의 표준오차

MSE가  $\sigma^2$ 의 불편추정량이므로  $\sqrt{MSE}$ 를  $\sigma$ 의 추정량으로 사용할 수 있으며 (회귀)추정 표준오차(standard error of estimate)라 부르고  $S_{y.x}$ 로 나타내기도 한다. 즉

$$S_{y.x} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2}{n-2}}.$$

이 값이 작을수록 회귀직선의 정확도가 높다고 할 수 있다. [예 2.2] ‘출력물’ 창의 “Residual standard error:” 줄에  $S_{y.x}$  값이 계산되어 있다.

## 2.5 원점을 지나는 회귀직선

두 변수 사이의 인과관계가  $x$ 가 0일 때  $y$ 도 0이어야 하는 경우가 있다. 예를 들어  $x$ 가 상품 재고량일 때  $y$ 가 판매량인 경우를 생각할 수 있다.

이러한 경우 적절한 모형은 통상적인 회귀모형에서 절편을  $\beta_0 = 0$  으로 놓아 원점을 지나는  $y_i = \beta_1 x_i + \epsilon_i$ 와 같은 회귀모형이다. 원점을 지나는 회귀모형에서도 통상적인 회귀모형에서와 같은 선형성, 등분산성, 독립성 등의 가정을 한다.

원점을 지나는 회귀모형에서  $\beta_1$ 의 최소제곱 추정량을 구해보자. 오차제곱합은

$$Q = \sum (y_i - \beta_1 x_i)^2$$

와 같이 주어지며, 이를 최소로 하는  $\beta_1$ 의 해  $b_1$ 이  $\beta_1$ 의 최소제곱추정량이다.

원점을 지나는 회귀모형에 대한 정규방정식은

$$\sum x_i(y_i - \beta_1 x_i) = 0$$

이므로 해  $b_1$ 은  $S_{xy}^* = \sum x_i y_i$ ,  $S_{xx}^* = \sum x_i^2$ 로 놓으면 다음과 같이 구해진다.

$$b_1 = \sum x_i y_i / \sum x_i^2 \equiv S_{xy}^* / S_{xx}^*.$$

원점을 지나는 회귀모형에서 총변동은 다음과 같고 자유도는  $n$ 이다.

$$SST = \sum y_i^2 = S_{yy}^*.$$

회귀제곱합은 다음과 같고 자유도는 1이다.

$$SSR = \sum \hat{y}_i^2 = \sum (b_1 x_i)^2 = b_1^2 \sum x_i^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2} = b_1 S_{xy}^*.$$

잔차제곱합은 다음과 같고 자유도는  $(n-1)$ 이다.

$$SSE = \sum (y_i - \hat{y}_i)^2 = SST - SSR = S_{yy}^* - b_1 S_{xy}^*.$$

[표 2.2] 원점을 지나는 회귀모형의 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$	$F_\alpha$
회귀	SSR	1	MSR	MSR/MSE	$F(1, n-1; \alpha)$
잔차	SSE	$n-1$	MSE		
계	SST	$n$			

오차항의 분산  $\sigma^2$ 의 불편추정량은 다음과 같다.

$$\hat{\sigma}^2 = MSE = (S_{yy}^* - b_1 S_{xy}^*) / (n-1).$$

R에서 원점을 지나는 회귀분석을 하려면 다음과 같은 형태로 모형을 적합시킨다.



```
> RegModel.0 <- lm( y ~ 0 + x, data='데이터셋이름' )
```

원점을 지나는 회귀직선을 산점도에 나타내려면 다음과 같이 x와 y의 구간을 조정할 필요가 있을 것이다.

```
> plot( x, y, xlim=c(0, max(x)), ylim=c(0, max(y)) )
> abline( 0, RegModel.0$coefficients )
```

## 2.6 가중최소제곱법

통상적인 회귀모형에서 가정하는 등분산성,  $Var(\epsilon_i) = \sigma^2$  또는  $Var(y_i) = \sigma^2$ 이 만족되지 않는 경우가 있다.

예를 들어  $x_i$ 에 대응하는  $n_i$ 개의 등분산( $\sigma^2$ )인 반응변수( $z$ )들의 평균을 새 반응변수( $y$ )로 삼는 경우를 생각하자. 반응변수가  $y_i = \sum_{j=1}^{n_i} z_j / n_i$ 이고  $Var(z_j) = \sigma^2$ 라 할 때

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

인 회귀모형에 대해  $Var(y_i) = Var(\epsilon_i) = \sigma^2 / n_i$ 이 되어 등분산이 아님을 알 수 있다.

이 때  $y_i^* = \sqrt{n_i} y_i$ 로 변환하면  $Var(y_i^*) = n_i Var(y_i) = n_i \sigma^2 / n_i = \sigma^2$ 이 되어 등분산성을 만족한다. 만약  $x_i^* = \sqrt{n_i} x_i$ ,  $\epsilon_i^* = \sqrt{n_i} \epsilon_i$ 로 놓으면, 회귀모형은

$$y_i^* = \sqrt{n_i} \beta_0 + \beta_1 x_i^* + \epsilon_i^*$$

과 같이 나타낼 수 있고 등분산성을 갖는 변환된 오차의 제곱합은  $\sum (\epsilon_i^*)^2$ 이다. 따라서 최소제곱법은

$$\sum (\epsilon_i^*)^2 = \sum (y_i^* - \sqrt{n_i} \beta_0 - \beta_1 x_i^*)^2 = \sum n_i (y_i - \beta_0 - \beta_1 x_i)^2$$

을 최소로 하는  $\beta_0, \beta_1$ 을 찾는 방법이 된다.

일반적으로 오차항의 분산이  $Var(\epsilon_i) = \sigma_i^2 = \sigma^2/w_i$  와 같은 형태일 때,  $w_i (> 0)$ 를 가중치라 부르고 최소화하려는 오차제곱합은 등분산성을 갖도록 변환된 오차의 제곱합이다. 이러한 경우에 사용하는 최소제곱법을 가중최소제곱법(method of Weighted Least Squares, WLS)이라 한다. 즉 가중최소제곱법에서는

$$Q = \sum w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

을 최소로 하는  $\beta_0, \beta_1$ 을 찾는다. 가중최소제곱법의 정규방정식은

$$\begin{aligned}\hat{\beta}_0 \sum w_i + \hat{\beta}_1 \sum w_i x_i &= \sum w_i y_i \\ \hat{\beta}_0 \sum w_i x_i + \hat{\beta}_1 \sum w_i x_i^2 &= \sum w_i x_i y_i\end{aligned}$$

이다. 가중평균을 각각

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}, \quad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

라 하고,

$$S_{xyw} = \sum w_i (x_i - \bar{x}_w) y_i, \quad S_{xxw} = \sum w_i (x_i - \bar{x}_w)^2$$

으로 놓으면 정규방정식의 해는 다음과 같다.

$$\hat{\beta}_1 = \frac{S_{xyw}}{S_{xxw}}, \quad \hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w.$$

등분산 가정이 어긋나 가중최소제곱법을 사용하는 경우로 등분산인  $n_i$ 개 반응변수 (z)들의 평균을 새 반응변수(y)로 사용하는 회귀모형에 대해 가중치  $w_i = n_i$ 를 이용하는 경우 외에도 반응변수의 분산이 설명변수 x에 비례하여 커지는 경우나 반응변수의 분산에 대한 추정값이 주어지는 경우 등을 생각할 수 있다. 반응변수의 분산이 설명변수 x에 비례하여 커지는 경우에는 반응변수를 로그변환한 값을 새 반응변수로 사용하면 등분산성을 갖게 되기도 하지만 변환하지 않고 가중치  $w_i = 1/x_i$ , 또는  $w_i = 1/x_i^2$ 을 이용한 가중최소제곱법이 선호된다. 반응변수  $y_i$ 의 분산에 대한 추정값이  $S_i^2$ 으로 주어지는 경우에는 변동이 작은 반응변수에 대한 가중치가 크게 되도록  $w_i = 1/S_i^2$ 을 이용한 가중최소제곱법을 사용할 수 있다.

[예 2.3] 다음은 Faraway(2002)에서 다룬 물리 실험 자료이다. 설명변수 energy와 반응변수 crossx 사이에 선형관계가 있다고 가정하고, 반응변수의 표준편차는 sd로 알려져 있다. 이 자료에 의하면 등분산성이 의심된다.

```
> strongx
      momentum energy crossx sd
1           4    0.345   367 17
2           6    0.287   311  9
3           8    0.251   295  9
4          10    0.225   268  7
5          12    0.207   253  7
6          15    0.186   239  6
7          20    0.161   220  6
8          30    0.132   213  6
9          75    0.084   193  5
10         150    0.060   192  5
```

통상적인 회귀분석을 위한 R의 모형과 결과는 다음과 같다.

```
> oreg <- lm(crossx~energy, data=strongx)
> summary(oreg)
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	135.00	10.08	13.40	9.21e-07 ***
energy	619.71	47.68	13.00	1.16e-06 ***

Residual standard error: 12.69 on 8 degrees of freedom

Multiple R-Squared: 0.9548, Adjusted R-squared: 0.9491

F-statistic: 168.9 on 1 and 8 DF, p-value: 1.165e-06

가중회귀를 위하여 다음과 같이 가중치를 “weights=sd^2”로 하여 모형을 정하고 결과를 보면 결정계수가 커짐을 알 수 있다.

```
> wreg <- lm(crossx ~energy, strongx, weights=sd^2)
> summary(wreg)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	118.88	10.91	10.89	4.47e-06 ***
energy	704.01	39.56	17.80	1.02e-07 ***

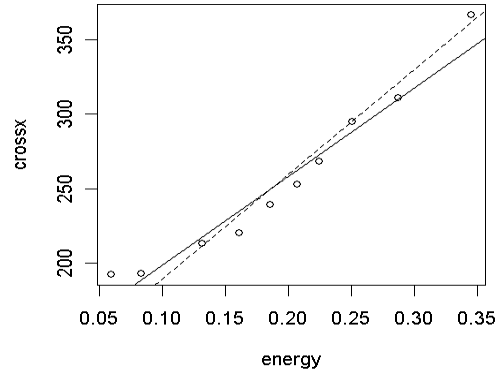
Residual standard error: 90.74 on 8 degrees of freedom

Multiple R-Squared: 0.9754, Adjusted R-squared: 0.9723

F-statistic: 316.8 on 1 and 8 DF, p-value: 1.017e-07

두 회귀적합을 비교하기 위하여 다음과 같이 그림을 그려 본다.

```
> plot(crossx ~ energy, data=strongx)
> abline(oreg)
> abline(wreg, lty=2)
```



[그림 2.1] 가중회귀적합선을 점선으로 나타낸 그림

## 2.7 제곱합들의 기대값 계산

앞에서 나온 제곱합들의 기대값 계산을 하면 다음과 같다.

$$\begin{aligned}
 E(SST) &= E\left(\sum (y_i - \bar{y})^2\right) \\
 &= E\sum (\beta_0 + \beta_1 x_i + \epsilon_i - \bar{y})^2 \\
 &= E\sum (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x} + \epsilon_i - \bar{\epsilon})^2 \\
 &= E\sum (\beta_1 (x_i - \bar{x}))^2 + E\left(\sum (\epsilon_i - \bar{\epsilon})^2\right) \\
 &= \beta_1^2 \sum (x_i - \bar{x})^2 + (n-1)\sigma^2
 \end{aligned}$$

$$\begin{aligned}
E(SSR) &= E(b_1^2 \sum (x_i - \bar{x})^2) \\
&= \sum (x_i - \bar{x})^2 \cdot E(b_1^2) \\
&= \sum (x_i - \bar{x})^2 \cdot [V(b_1) + \beta_1^2] \\
&= \sum (x_i - \bar{x})^2 \left[ \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \right] + \beta_1^2 \sum (x_i - \bar{x})^2 \\
&= \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2
\end{aligned}$$

$$\therefore E(MSR) = E(SSR/1) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2$$

$$SSE = SST - SSR \text{ 이므로 } E(SSE) = E(SST) - E(SSR) = (n-2)\sigma^2.$$

따라서  $E(MSE) = \sigma^2$ 이고, 평균제곱들의 기대값의 비는 다음과 같다.

$$\frac{E(MSR)}{E(MSE)} = \frac{\sigma^2 + \beta_1^2 \sum (x_i - \bar{x})^2}{\sigma^2} = 1 + \frac{\beta_1^2 \sum (x_i - \bar{x})^2}{\sigma^2}.$$

이로부터 검정통계량  $F_0 = \frac{MSR}{MSE}$ 가 클수록  $\beta_1$ 이 0과 멀리 떨어져 있다고 생각할 수 있을 것이다.

앞장에서 최소제곱법으로  $\beta_0, \beta_1$ 의 최량선형불편추정량(BLUE)인  $b_0, b_1$ 을 구하였다. 이렇게 구해진 추정량들은 통계량으로서 분산을 갖지만 가장 작다는 보장이 되어 있는 안정된 점추정량이므로 이들을 사용하여 관심의 대상인 모수들에 대한 원하는 신뢰수준의 구간추정이나 미리 정한 유의수준하의 검정을 하고 예측구간을 구한다.

구체적으로 절편  $\beta_0$ 와 기울기  $\beta_1$ 의 신뢰구간,  $x$ 가 주어졌을 때  $y$ 의 반응함수라고도 부르는 조건부 기대값인  $E(y|x) = \beta_0 + \beta_1 x$ 에 대한 구간추정 그리고  $x$ 가 주어졌을 때 새로운  $y$ 에 대한 예측구간을 구한다.

### 3.1 절편 $\beta_0$ , 기울기 $\beta_1$ 의 신뢰구간

기울기  $\beta_1$ 의 최소제곱추정량은 다음과 같이 구해졌다.

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

여기서 분자는

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i - \sum (x_i - \bar{x}) \bar{y} \\ &= \sum (x_i - \bar{x}) y_i - \bar{y} \sum (x_i - \bar{x}) = \sum (x_i - \bar{x}) y_i - \bar{y} \cdot 0 \\ &= \sum (x_i - \bar{x}) y_i \end{aligned}$$

와 같이 나타낼 수 있고,  $k_i = \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{(x_i - \bar{x})}{S_{xx}}$ 라 놓으면  $b_1 = \sum k_i y_i$ 이고

확률변수  $y_i$ 들의 계수  $k_i$ 에 의한 선형결합함수로서 통계량이므로 분포를 가지며 정규성 가정을 할 경우 기대값과 분산을 구하면 그 분포를 알 수 있다.

계수  $k_i$ 는  $\sum k_i = 0$  ,  $\sum k_i x_i = 1$  ,  $\sum k_i^2 = 1/S_{xx}$ 을 만족하고  
 $y_i$ 들은 독립성과 등분산성을 가정하므로

$$E(b_1) = \sum k_i E(y_i) = \sum k_i E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 \sum k_i + \beta_1 \sum k_i x_i = \beta_1 ,$$

$$Var(b_1) = Var(\sum k_i y_i) = \sum k_i^2 Var(y_i) = \sum k_i^2 \sigma^2 = \sigma^2 / S_{xx} .$$

따라서, 기울기  $\beta_1$ 의 최소제곱추정량  $b_1$ 의 분포는  $b_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$ 이다.

통상적인 방법으로 기울기  $\beta_1$ 의 신뢰수준  $100(1-\alpha)\%$  신뢰구간은 다음과 같이 나타낸다.

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 아는 경우

$$b_1 - z_{\alpha/2} \cdot S.E.(b_1) \leq \beta_1 \leq b_1 + z_{\alpha/2} \cdot S.E.(b_1)$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 모르는 경우

$$b_1 - t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(b_1) \leq \beta_1 \leq b_1 + t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(b_1)$$

여기서  $S.E.(b_1) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$  이고  $\widehat{S.E.}(b_1) = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$  이다.

오차의 분산  $\sigma^2$ 을 모르는 경우가 대부분이며, 기울기  $\beta_1$ 의 신뢰수준  $100(1-\alpha)\%$  신뢰구간은 다음과 같이 나타낼 수 있다.

$$b_1 - t_{\frac{\alpha}{2}}(n-2) \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}} \leq \beta_1 \leq b_1 + t_{\frac{\alpha}{2}}(n-2) \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$$

앞에서 다룬 데이터셋 ‘MM1’에 대하여 기울기에 대한 구간추정을 하면 다음과 같다.

```
> n=length(x); mse=sse/(n-2)
> c(b1-qt(0.975, n-2)*sqrt(mse/sxx), b1+ qt(0.975, n-2)*sqrt(mse/sxx))
```



```
[1] 0.2516301 0.2675752
```

다음과 같이 R의 신뢰구간을 구하는 함수 `confint()`를 이용하면 절편과 기울기에 대한 구간추정을 할 수 있다.

```
> confint(RegModel.1)
                2.5 %      97.5 %
(Intercept)  0.00314383  0.005462461
x            0.25163007  0.267575168
```

만약 신뢰수준 99%인 구간추정을 하고자 한다면 `confint(RegModel.1, level=0.99)`와 같이 신뢰수준을 지정하면 된다.

절편  $\beta_0$ 의 최소제곱추정량  $b_0 = \bar{y} - b_1 \bar{x}$ 의 식에서  $b_1 = \sum k_i y_i$ 를 대입하고 정리하면  $b_0 = \sum (\frac{1}{n} - \bar{x} k_i) y_i$ 로 나타낼 수 있다. 여기서  $l_i = \frac{1}{n} - \bar{x} k_i$ 로 놓으면  $b_0 = \sum l_i y_i$

이고 확률변수  $y_i$ 들의 계수  $l_i$ 에 의한 선형결합함수로서 통계량이므로 분포를 가지며 정규성 가정을 할 경우 기대값과 분산을 구하면 그 분포를 알 수 있다.

계수  $l_i$ 는  $\sum l_i = 1$ ,  $\sum l_i x_i = 0$ ,  $\sum l_i^2 = 1/n + \bar{x}^2/S_{xx}$ 을 만족하고  $y_i$ 들은 독립성과 등분산성을 가정하므로

$$E(b_0) = \sum l_i E(y_i) = \sum l_i E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 \sum l_i + \beta_1 \sum l_i x_i = \beta_0,$$

$$Var(b_0) = Var(\sum l_i y_i) = \sum l_i^2 Var(y_i) = \sum l_i^2 \sigma^2 = \sigma^2 (1/n + \bar{x}^2/S_{xx}).$$

따라서, 절편  $\beta_0$ 의 최소제곱추정량  $b_0$ 의 분포는  $b_0 \sim N(\beta_0, \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$ 이다.

절편  $\beta_0$ 의 신뢰수준  $100(1-\alpha)\%$  신뢰구간은 다음과 같이 나타낸다.

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 아는 경우

$$b_0 - z_{\alpha/2} \cdot S.E. (b_0) \leq \beta_0 \leq b_0 + z_{\alpha/2} \cdot S.E. (b_0)$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 모르는 경우

$$b_0 - t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(b_0) \leq \beta_0 \leq b_0 + t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(b_0)$$

여기서

$$S.E. (b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right), \quad \widehat{S.E.} (b_0) = \sqrt{MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)}$$

이므로 절편  $\beta_0$ 의 신뢰수준  $100(1-\alpha)\%$  신뢰구간은 오차의 분산  $\sigma^2$ 을 모르는 경우 다음과 같이 나타낼 수 있다.

$$b_0 - t_{\frac{\alpha}{2}}(n-2) \sqrt{MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)} \leq \beta_0 \leq b_0 + t_{\frac{\alpha}{2}}(n-2) \sqrt{MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)}$$

데이터셋 ‘MM1’에 대하여 절편에 대한 구간추정을 하면 다음과 같다.

```
> c(b0-qt(0.975, n-2)*sqrt(mse*(1/n+ mean(x)^2/sxx)),
+ b0+ qt(0.975, n-2)*sqrt(mse*(1/n+ mean(x)^2/sxx)))
[1] 0.003143830 0.005462461
```

절편  $\beta_0$ , 기울기  $\beta_1$ 에 대한 동시추론에 관한 내용은 박성현 (1999)을 참고하도록 한다.

### 3.2 주어진 $x$ 에서 $E(y|x) = \mu_{y \cdot x}$ 의 신뢰구간

주어진  $x = x_0$ 에서 조건부 기대값  $E(y|x_0) = \mu_{y \cdot x_0}$ 의 불편추정량으로

$$E(\widehat{y|x_0}) = \hat{\mu}_{y \cdot x_0} = b_0 + b_1 x_0 = \hat{y}_0$$

을 사용한다. 이 식에  $b_0 = \sum l_i y_i$ ,  $b_1 = \sum k_i y_i$ 를 대입하여 정리하면

$$\hat{y}_0 = b_0 + b_1 x_0 = \sum l_i y_i + \sum x_0 k_i y_i = \sum (l_i + x_0 k_i) y_i$$

와 같다. 여기서  $m_i = l_i + x_0 k_i = 1/n + (x_0 - \bar{x})k_i$ 로 놓으면  $\hat{y}_0 = \sum m_i y_i$ 이고 확률변수  $y_i$ 들의 계수  $m_i$ 에 의한 선형결합함수로서 통계량이므로 분포를 가지며 정규성 가정을 할 경우 기대값과 분산을 구하면 그 분포를 알 수 있다.

계수  $m_i$ 는  $\sum m_i = 1$ ,  $\sum m_i x_i = x_0$ ,  $\sum m_i^2 = 1/n + (x_0 - \bar{x})^2 / S_{xx}$ 을 만족하고  $y_i$ 들은 독립성과 등분산성을 가정하므로

$$E(\hat{y}_0) = \sum m_i E(y_i) = \sum m_i E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 \sum m_i + \beta_1 \sum m_i x_i = \beta_0 + \beta_1 x_0,$$

$$Var(\hat{y}_0) = Var(\sum m_i y_i) = \sum m_i^2 Var(y_i) = \sum m_i^2 \sigma^2 = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right).$$

따라서,  $E(y|x_0) = \mu_{y \cdot x_0}$ 의 추정량  $\hat{y}_0$ 의 분포는 다음과 같다.

$$\hat{y}_0 \sim N(\mu_{y \cdot x}, \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)).$$

회귀계수에 대한 신뢰구간을 구할 때와 같은 논리를 사용하여 조건부 기대값  $E(y|x_0) = \mu_{y \cdot x_0}$ 의 신뢰수준  $100(1-\alpha)\%$  신뢰구간은 다음과 같이 나타낼 수 있다.

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 아는 경우는

$$\hat{y}_0 - z_{\alpha/2} \cdot S.E.(\hat{y}_0) \leq \mu_{y \cdot x_0} \leq \hat{y}_0 + z_{\alpha/2} \cdot S.E.(\hat{y}_0)$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 모르는 경우는

$$\hat{y}_0 - t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(\hat{y}_0) \leq \mu_{y \cdot x_0} \leq \hat{y}_0 + t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(\hat{y}_0)$$

여기서

$$S.E.(\hat{y}_0) = \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) \cdot \sigma^2}, \quad \widehat{S.E.}(\hat{y}_0) = \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) \cdot MSE}.$$

데이터셋 'MM1'에 대하여 농도가 0.1로 주어질 때 평균반응함수에 대한 구간추정을 하면 다음과 같다.

```
> x0=0.1; y0hat=b0+ b1*x0
> se_y0hat=sqrt((1/n+ (x0-mean(x))^2/sxx)*mse)
> c(y0hat-qt(0.975, n-2)*se_y0hat, y0hat+ qt(0.975, n-2)*se_y0hat)
[1] 0.02938098 0.03114583
```

다음과 같이 R의 예측구간을 구하는 함수 predict()에서 interval = "confidence"로 지정하면 주어진 x에서의 y의 조건부평균에 대한 구간추정을 할 수 있다.

```
> x0 <- data.frame(x=0.1)
> predict(RegModel.1, x0, se.fit = TRUE, interval = "confidence", level = 0.95)
$fit
          fit          lwr          upr
[1,] 0.03026341 0.02938098 0.03114583
```

### 3.3 주어진 $x$ 에서 새로운 $y$ 의 예측구간

주어진  $x = x_0$ 에서  $E(\widehat{y}|x_0) = \hat{\mu}_{y \cdot x_0} = b_0 + b_1 x_0 = \hat{y}_{0,n}$ 로 나타낼 경우, 새로운  $y$ 의 값을 나타내는 확률변수  $y_{0,n}$ 에 대한 신뢰수준  $100(1-\alpha)\%$  예측구간을 생각해보자. 여기서  $y_{0,n}$ 은 (미지의 상수인) 모수가 아니고 주어진  $x_0$ 에서 예측의 대상인 새로운

$y$ 값을 나타내는 확률변수이기 때문에 신뢰구간이 아닌 예측구간이라는 용어를 사용하였다. 회귀모형에서의 독립성 가정에 의하여 새로운  $y$ 값인  $y_{0,n}$ 은 이미 뽑힌 크기  $n$ 인 표본들의 함수인  $\hat{y}_{0,n}$ 과 독립이라는 사실을 이용한다.

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 아는 경우는

$$\hat{y}_{0,n} - z_{\alpha/2} \cdot S.E.(\hat{y}_{0,n} - y_{0,n}) \leq y_{0,n} \leq \hat{y}_{0,n} + z_{\alpha/2} \cdot S.E.(\hat{y}_{0,n} - y_{0,n})$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 모르는 경우는

$$\hat{y}_{0,n} - t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(\hat{y}_{0,n} - y_{0,n}) \leq y_{0,n} \leq \hat{y}_{0,n} + t_{\frac{\alpha}{2}}(n-2) \cdot \widehat{S.E.}(\hat{y}_{0,n} - y_{0,n})$$

여기서

$$S.E.(\hat{y}_{0,n} - y_{0,n}) = \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) \cdot \sigma^2},$$

$$\widehat{S.E.}(\hat{y}_{0,n} - y_{0,n}) = \sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right) \cdot MSE}.$$

데이터셋 'MM1'에 대하여 농도가 0.1로 주어질 때 반응함수에 대한 예측구간을 구하면 다음과 같다.

```
> se_y0hat_y0=sqrt((1+ 1/n+ (x0-mean(x))^2/sxx)*mse)
> c(y0hat-qt(0.975, n-2)*se_y0hat_y0, y0hat+ qt(0.975, n-2)*se_y0hat_y0)
[1] 0.02777025 0.03275656
```

다음과 같이 R의 예측구간을 구하는 함수 predict()에서 interval = "prediction"으로 지정하면 주어진  $x$ 에서의  $y$ 에 대한 예측구간을 구할 수 있다.

```
> x0 <- data.frame(x=0.1)
> predict(RegModel.1, x0, se.fit = TRUE, interval = "prediction", level = 0.95)
$fit
      fit      lwr      upr
[1,] 0.03026341 0.02777025 0.03275656
```

### 3.4 가설검정

관심의 대상인 모수  $\theta$ 에 대하여 불편추정량이  $\hat{\theta}$ 이고 정규분포를 한다면 다음과 같이 검정통계량을 정의하여 검정하는 것이 통상적인 방법이다.

$$Z_0 = \frac{\hat{\theta} - \theta_0}{S.E.(\hat{\theta})}$$

여기서  $\theta_0$ 은 모수  $\theta$ 에 대하여 귀무가설에서 주장한 값이다. 즉  $H_0 : \theta = \theta_0$ 이다. 이렇게 정의된 검정통계량  $Z_0$ 는 귀무가설하에서 표준정규분포를 하므로 대립가설의 형태에 따라 양측인 경우  $z_{\alpha/2}$ , 단측인 경우  $z_{\alpha}$ 를 기준으로 검정하거나 유의확률 또는 확률값을 유의수준과 비교하여 검정한다.

실제로  $\hat{\theta}$ 의 표준오차  $S.E.(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$ 는 알려져 있지 않은 경우가 대부분이므로  $Var(\hat{\theta})$ 의 불편추정량  $\widehat{Var}(\hat{\theta})$ 을 구하여 표준오차의 추정량을  $\widehat{S.E.}(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}$ 와 같이 정의하여 다음 검정통계량을 정의하는데 사용한다.

$$t_0 = \frac{\hat{\theta} - \theta_0}{\widehat{S.E.}(\hat{\theta})}$$

이렇게 정의된 검정통계량  $t_0$ 는 t-분포를 하므로 주어진 자유도와 함께 대립가설의 형태에 따라 양측인 경우  $t_{\alpha/2}$ , 단측인 경우  $t_{\alpha}$ 를 기준으로 검정하거나 유의확률 또는 확률값을 유의수준과 비교하여 검정한다.

#### 3.4.1 절편 $\beta_0$ 에 대한 검정

절편  $\beta_0$ 에 대한 유의수준  $\alpha$ 인 가설검정은 다음과 같이 행한다.

$$H_0 : \beta_0 = \beta_{00} \quad V.S \quad H_1 : \begin{array}{l} \beta_0 \neq \beta_{00} \\ \beta_0 > \beta_{00} \\ \beta_0 < \beta_{00} \end{array}$$

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 아는 경우

$$Z_0 = \frac{b_0 - \beta_{00}}{S.E.(b_0)} \sim N(0,1) \quad , \quad S.E.(b_0) = \sqrt{\sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)},$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 모르는 경우

$$t_0 = \frac{b_0 - \beta_{00}}{\widehat{S.E.}(b_0)} \sim t(n-2) \quad , \quad \widehat{S.E.}(b_0) = \sqrt{MSE \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)}$$

### 3.4.2 기울기 $\beta_1$ 의 검정

기울기  $\beta_1$ 에 대한 유의수준  $\alpha$ 인 검정은 다음과 같이 행한다.

$$H_0 : \beta_1 = \beta_{10} \quad V.S \quad H_1 : \begin{array}{l} \beta_1 \neq \beta_{10} \\ \beta_1 > \beta_{10} \\ \beta_1 < \beta_{10} \end{array}$$

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 아는 경우

$$Z_0 = \frac{b_1 - \beta_{10}}{S.E.(b_1)} \sim N(0,1) \quad , \quad S.E.(b_1) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$  를 모르는 경우

$$t_0 = \frac{b_1 - \beta_{10}}{\widehat{S.E.}(b_1)} \sim t(n-2) \quad , \quad \widehat{S.E.}(b_1) = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}}$$

### 3.4.3 조건부 기대값 $E(y|x) = \mu_{y \cdot x}$ 에 대한 검정

주어진  $x$ 에서  $E(y|x) = \mu_{y \cdot x}$ 에 대한 유의수준  $\alpha$ 인 가설검정을 하자.

$$H_0 : \mu_{y \cdot x} = \mu_{0y \cdot x} \quad V.S \quad H_1 : \begin{array}{l} \mu_{y \cdot x} \neq \mu_{0y \cdot x} \\ \mu_{y \cdot x} > \mu_{0y \cdot x} \\ \mu_{y \cdot x} < \mu_{0y \cdot x} \end{array}$$

(1) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 아는 경우

$$Z_0 = \frac{\hat{y} - \mu_{0y \cdot x}}{S.E.(\hat{y})} \sim N(0,1) \quad , \quad S.E.(\hat{y}) = \sqrt{\sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)} ,$$

(2) 오차의 분산  $\sigma^2$  또는 표준편차  $\sigma$ 를 모르는 경우

$$t_0 = \frac{\hat{y} - \mu_{0y \cdot x}}{S.E.(\hat{y})} \sim t(n-2) \quad , \quad S.E.(\hat{y}) = \sqrt{MSE \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

## 3.5 적합결여검정

설명변수  $x$ 와 반응변수  $y$ 사이의 관계에 대하여 선형성을 가정하여 왔다. 실제로 두 변수 사이의 관계를 선형회귀모형  $y = \beta_0 + \beta_1 x + \epsilon$ 로 나타내는 것이 타당한가 검정하는 방법이 적합결여검정(Lack-of-Fit test)이다. 적합결여검정의 귀무가설과 대립가설은 다음과 같이 나타낸다.

$$H_0 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad \text{또는} \quad H_0 : E(y_i|x) = \beta_0 + \beta_1 x_i \quad .$$

$$H_1 : H_0 \text{에서 주장하는 모형이 아니다} \quad , \quad \text{또는} \quad H_1 : E(y_i|x) \neq \beta_0 + \beta_1 x_i \quad .$$

이러한 적합결여검정의 검정통계량은  $H_0$ 와  $H_1$ 하에서  $\sigma^2$ 에 대한 추정치에 근거하여 유도하는데  $H_1$ 하에서는 주어진 모형이 없으므로 모형과 무관한  $\sigma^2$ 에 대한 추정



방법이 필요하게 되고 이를 위해서는 주어진  $x$ 의 수준(level)에서 반응변수에 대한 독립적인 측정값이 여러 개 필요하다.  $x$ 의 적어도 하나의 수준에서 두 개 이상의  $y$ 에 대한 측정이 이루어져야하고 수준 수는 3개 이상이어야 한다.

주어진  $x$ 의 수준이  $x_1, x_2, \dots, x_k$ 이고, 각 수준에서  $n_1, n_2, \dots, n_k$ 개의 반복 측정치들이 다음과 같이 주어지고 전체 자료 수는  $\sum_{i=1}^k n_i = n$  개라 하자.

$$\begin{array}{l} x_1 \text{ 에서 } y_{11}, y_{12}, \dots, y_{1n_1} \\ x_2 \text{ 에서 } y_{21}, y_{22}, \dots, y_{2n_2} \\ \vdots \\ x_k \text{ 에서 } y_{k1}, y_{k2}, \dots, y_{kn_k} \end{array}$$

이  $n$ 개의 자료에 의하여 최소제곱법으로 구한 회귀직선을

$$\hat{y}_i = b_0 + b_1 x_i, \quad i = 1, 2, \dots, k$$

이라 하자.  $x$ 의 하나의 수준  $x_i$ 에서  $n_i$ 개의  $y$ 에 대한 측정치가 있었지만 회귀직선의 적합값은  $\hat{y}_i$  하나뿐이다.

$x$ 의 하나의 수준  $x_i$ 에서  $n_i$ 개의  $y$ 에 대한 평균을  $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ 로 나타내는 경우 이 회귀모형에 대한 잔차제곱합 SSE를 다음과 같이 두 가지 제곱합으로 분해하여 나타낼 수 있다.

$$\begin{aligned} SSE &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \hat{y}_i) \end{aligned}$$

에서

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \hat{y}_i) = \sum_{i=1}^k (\bar{y}_i - \hat{y}_i) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0$$

이므로

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2.$$

여기에서 첫 번째 항의 제곱합은 가정된 모형과 관계없이 오차에만 의존하며

$$SSPE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

로 나타내고, 순오차제곱합(Pure Error Sum of Squares)이라 부른다. 수준  $i$ 에서의 제곱합  $\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ 의 자유도가  $(n_i - 1)$ 이므로 독립성 가정하에서 순오차제곱합 SSPE의 자유도는  $\sum_{i=1}^k (n_i - 1) = (n - k)$ 이다.

$$SSLF = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^k n_i (\bar{y}_i - \hat{y}_i)^2$$

두 번째 제곱합은 가정한 모형이 적합하지 않을 때 큰 값을 갖는 적합결여의 측도이며

$$SSLF = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 = \sum_{i=1}^k n_i (\bar{y}_i - \hat{y}_i)^2$$

으로 나타내고, 적합결여제곱합(Lack of Fit Sum of Squares)이라 부른다. 적합결여제곱합은  $SSLF = SSE - SSPE$ 로 구할 수 있고, 잔차제곱합(SSE)의 자유도가  $(n - 2)$ 이므로 적합결여제곱합(SSLF)의 자유도는  $(n - 2) - (n - k) = (k - 2)$ 와 같이 구할 수 있다.

순오차제곱평균은  $MSPE = \frac{SSPE}{n - k}$ 와 같이 나타내고 가정한 모형이 적합하다는 귀무가설  $H_0$ 가 참인가와 관계 없이  $E(MSPE) = \sigma^2$ 이다. 즉, MSPE는  $\sigma^2$ 의 불편 추정량이다.

적합결여제곱평균은  $MSLF = \frac{SSLF}{k-2}$ 와 같이 나타낸다.

귀무가설  $H_0$ 가 참인 경우 분산의 불편추정량은  $\hat{\sigma}^2 = MSE$ 이므로 만약  $MSE \approx MSPE$ 이면  $H_0$ 가 참이라는 즉 가정한 선형모형이 적합하다는 증거로 생각할 수 있다. 따라서 가정한 선형모형이 적합한가를 검정하기 위한 통계량으로

$$F_0 = \frac{(SSE - SSPE) / (k - 2)}{MSPE} = \frac{MSLF}{MSPE}$$

을 사용하며,  $F_0 > F_{\alpha}(k-2, n-k)$ 이면 귀무가설을 기각하고 아니면 채택한다.

[표 3.1] 적합결여검정을 위한 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$
회귀	SSR	1	MSR	$F_0 = \frac{MSR}{MSE}$
잔차	SSE	n-2	MSE	
적합결여	SSLF	k-2	MSLF	$F_0 = \frac{MSLF}{MSPE}$
순오차	SSPE	n-k	MSPE	

예를 들기 위하여 Draper and Smith(1998)에 소개된 철의 함유율이 다른 13개의 Cu-Ni 합금 시료를 60일간 바다 물에 넣어 단위 규격당 부식에 의하여 줄어든 무게를 기록한 자료를 ‘corrosion’이라는 이름의 데이터셋으로 R에 입력하고 적합결여검정을 하여 보자.

```
> data(corrosion)
> corrosion
  Fe loss
1 0.01 127.6
2 0.48 124.0
```

```

3  0.71 110.8
4  0.95 103.9
5  1.19 101.5
6  0.01 130.1
7  0.48 122.0
8  1.44  92.3
9  0.71 113.1
10 1.96  83.7
11 0.01 128.0
12 1.44  91.4
13 1.96  86.2

```

```
> RegModel.1 <- lm(loss~Fe, data=corrosion)
```

```
> summary(RegModel.1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	129.787	1.403	92.52	< 2e-16 ***
Fe	-24.020	1.280	-18.77	1.06e-09 ***

Residual standard error: 3.058 on 11 degrees of freedom

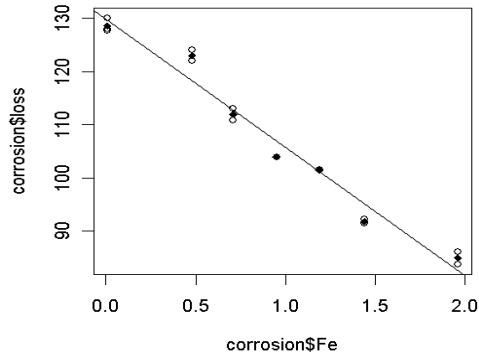
Multiple R-Squared: 0.9697, Adjusted R-squared: 0.967

F-statistic: 352.3 on 1 and 11 DF, p-value: 1.055e-09

이 자료에 대해 적합된 직선을 다음과 같이 그린다.

```
> plot(corrosion$Fe, corrosion$loss)
```

```
> abline(RegModel.1$coef)
```



[그림 3.1] 부식 자료에 대한 선형적합과 그룹별 평균 그림

이 그림에 설명변수 Fe를 범주화 하고 범주의 수준별로 주어지는 그룹에 대한 반응 변수 loss의 평균을 다이아몬드 기호로 나타낸 그림을 다음과 같은 방법으로 추가한다.

```
> RegModel.2 <- lm(loss~factor(Fe), data=corrosion)
> points(corrosion$Fe, RegModel.2$fit, pch=18)
```

적합결과 결정계수가 0.97이어서 회귀직선이 자료에 비교적 잘 적합된 것으로 보인다. 일반적으로 설명변수 x를 범주형으로 간주하여 모형을 적합시키면 범주의 수준 즉 각 x값마다 반응변수 y의 평균이 적합된다. 이러한 두 모형을 다음과 같이 통상적인 분산분석법으로 비교할 수 있다.

```
> anova(RegModel.1, RegModel.2)
```

Analysis of Variance Table

Model 1: loss ~ Fe

Model 2: loss ~ factor(Fe)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	11	102.850				
2	6	11.782	5	91.069	9.2756	0.008623 **

확률값이 매우 작아 적합결여가 있는 것으로 판단된다. 위 그림으로 보아서는 뚜렷하게 나타나지 않지만 직선이 아닌 다른 모형을 찾아야 할 것이다. 그러나 반복에 상

관이 있다면 순오차의 분산이 줄어 들 수 있기 때문에 우선 반복이 독립적인지 확인할 필요가 있다. 또한 측정되지 않은 제3의 변수가 적합결여를 초래했을 가능성도 고려해야 한다.

이러한 적합결여검정은 적합결여를 찾아내는데 좋은 방법이지만 귀무가설이 기각되지 않았다고 해서 모형이 참이라는 뜻은 아니라는 것을 유의해야 한다. 다만 자료가 적합결여를 찾아내기에 충분하지 못했을 수도 있기 때문에 귀무가설이 기각되지 않는 경우에 모형이 자료와 모순되지 않는다고 결론을 내리는 것이 합리적이다.

## 제 4 장

# 모형의 타당성과 잔차분석

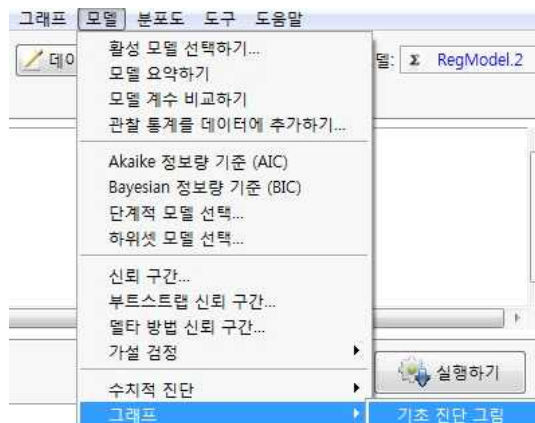
앞에서 알아본 단순회귀분석은 모형의 선형성, 오차항의 독립성, 등분산성, 정규성 가정을 하고 설명변수는 주어진다고 간주하였다. 그러나 실제로 이러한 가정이 충족되는지 검토되어야 한다.

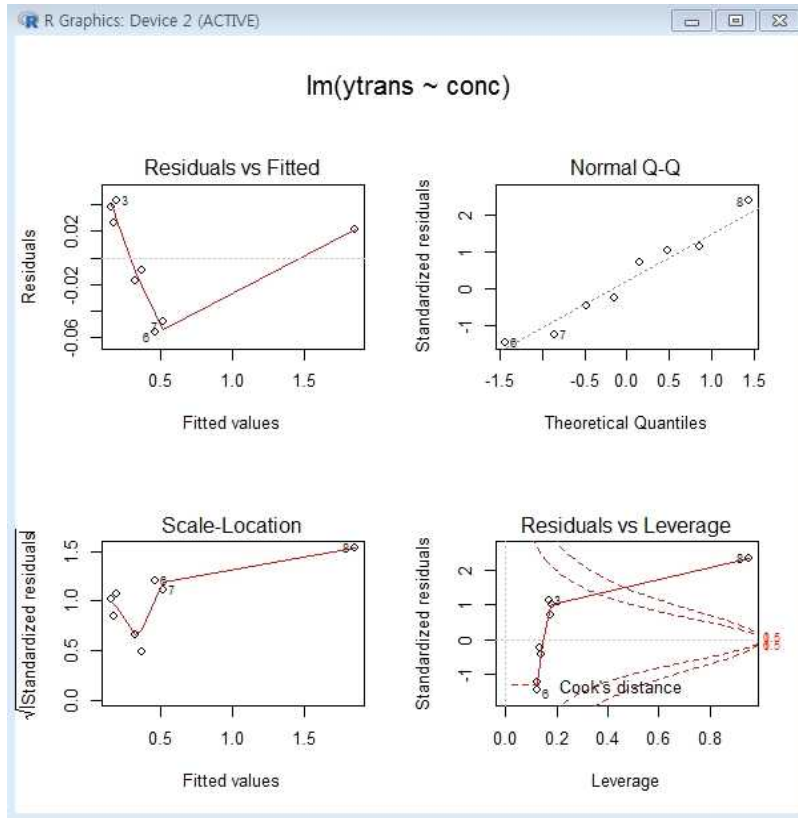
적합한 회귀직선이 타당하다는 선형성에 대해서는 앞에서 적합결여검정으로 판정하는 방법을 설명하였다. 이 장에서는 잔차의 검토를 통하여 타당성을 판정하는 방법들을 알아본다.

### 4.1 잔차분석

잔차는  $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1x_i$  로 구해지며, 오차  $\epsilon_i = y_i - E(y_i|x_i)$ 의 실현된 값으로 볼 수 있기 때문에 모형이나 오차항에 대한 가정이 충족되는지 검토하는데 기본적으로 사용된다.

앞에서 다뤘던 표1의 물질 농도에 따른 화학반응의 속도 자료인 ‘MM1’ 데이터셋에 대하여 잔차에 대한 기본적인 진단을 하는 그림을 R commander의 [‘모텔’ -> ‘그래프’ -> ‘기초 진단 그림’] 메뉴를 사용하여 다음과 같이 나타낼 수 있다.





[그림 4.1] 농도에 따른 반응속도 자료에 대한 잔차 그림

#### 4.1.1 회귀모형의 선형성

주어진 자료에 대한 산점도에 의해 선형성을 직관적으로 파악할 수 있고, 또한 잔차  $e_i$ 를  $\hat{y}_i$ 나  $x_i$ 에 대해서 그린 산점도가 0을 중심으로 랜덤하게 나타나는가로 파악할 수도 있다. [그림 4.1]의 왼쪽 위 ‘Residuals vs Fitted’ 그림에서는 랜덤하게 나타난다고 보기 어려우며 선형성이 의심된다.

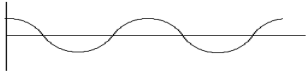
#### 4.1.2 오차항의 등분산성

잔차  $e_i$ 를  $\hat{y}_i$ 나  $x_i$ 에 대해 그린 산점도가 0을 중심으로 산포가 일정한지를 알 수 있다. [그림 4.1]의 왼쪽 위 ‘Residuals vs Fitted’ 그림이나 잔차를 표준화하고 절대값의 제곱근을 구해 양수값만 나타나도록 접어 올린 왼쪽 아래 ‘Scale-Location’ 그림을 보면 등분산성이 없다고 보이지는 않는다.



### 4.1.3 오차항의 독립성

설명변수가 시간을 나타낼 때 잔차  $e_i$ 를  $x_i$ 에 대해서 그린 산점도가 어떤 일정한 주기를 보인다면 오차  $\varepsilon_i$ 들 사이에 종속관계가 예상된다.



### 4.1.4 오차항의 정규성

정규성을 검토하는 방법은 반응변수나 잔차의 히스토그램이 종모양인가를 보거나 정규확률('Normal Q-Q') 그림의 선형성을 보는 등과 같은 직관적인 방법과 정규분포의 성질을 이용하는 적합도검정과 같은 추론적인 방법이 있다. [그림 4.1]의 오른쪽 위 'Normal Q-Q' 그림을 보면 점들이 직선에서 약간 벗어났지만 정규성이 없다고 할 정도는 아닌 것으로 본다.

## 4.2 오차의 자기상관

주식 가격과 같이 시간에 따라 변하는 관측값들인 시계열 자료에는 상관이 존재하는 경우가 많다. 단순선형회귀모형에서 오차항  $\epsilon_i$ 들은 서로 독립이고 등분산이라 가정한다. 즉  $Var(\epsilon_i) = \sigma^2$ ,  $Cov(\epsilon_i, \epsilon_j) = 0$ 이라고 가정한다. 그러나 설명변수가 시간과 관련되어 있을 때 반응변수 또는 오차항  $\epsilon_i$ 들 사이에 독립성 가정을 만족하지 못하고 상관이 존재하는 경우가 있다. 즉  $Cov(\epsilon_i, \epsilon_j) \neq 0$ ,  $i \neq j$  일 수 있다.

단순선형회귀모형  $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ 에서 오차항  $\epsilon_t$ 들 사이에

$$\epsilon_t = \rho \epsilon_{t-1} + \delta_t \quad (4.1)$$

인 관계가 있고  $\rho \neq 0$ ,  $E(\delta_t) = 0$ ,  $Var(\delta_t) = \sigma_\delta^2$ ,  $Cov(\delta_t, \delta_{t'}) = 0$ ,  $t \neq t'$ 인 경우 오차항  $\epsilon_t$ 들 사이에는 일차자기상관(first-order autocorrelation)이 존재한다고 말하고

$\rho$  를 일차자기상관계수(first-order autocorrelation coefficient)라 부른다. 식 (4.1)을 전개하면

$$\begin{aligned}
 \epsilon_t &= \rho \epsilon_{t-1} + \delta_t \\
 &= \rho (\rho \epsilon_{t-2} + \delta_{t-1}) + \delta_t \\
 &= \delta_t + \rho \delta_{t-1} + \rho^2 \epsilon_{t-2} \\
 &= \delta_t + \rho \delta_{t-1} + \rho^2 (\rho \epsilon_{t-3} + \delta_{t-2}) \\
 &= \delta_t + \rho \delta_{t-1} + \rho^2 \delta_{t-2} + \rho^3 \epsilon_{t-3} \\
 &\vdots \\
 &= \sum_{j=0}^{\infty} \rho^j \delta_{t-j}
 \end{aligned}$$

와 같이 나타낼 수 있고,  $E(\delta_t) = 0$ ,  $Var(\delta_t) = \sigma_\delta^2$ ,  $Cov(\delta_t, \delta_{t'}) = 0$ ,  $t \neq t'$  이므로

$$\begin{aligned}
 E(\epsilon_t) &= E\left[\sum_{j=0}^{\infty} \rho^j \delta_{t-j}\right] = \sum_{j=0}^{\infty} \rho^j E(\delta_{t-j}) = 0 \\
 V(\epsilon_t) &= V\left[\sum_{j=0}^{\infty} \rho^j \delta_{t-j}\right] = \sum_{j=0}^{\infty} \rho^{2j} V(\delta_{t-j}) = \frac{\sigma_\delta^2}{1 - \rho^2} = \sigma_\epsilon^2
 \end{aligned}$$

임을 알 수 있다. 또한  $k = j+1$  로 놓으면

$$\begin{aligned}
 Cov(\epsilon_t, \epsilon_{t-1}) &= Cov\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j}, \sum_{j=0}^{\infty} \rho^j \delta_{t-1-j}\right) \\
 &= Cov\left(\sum_{j=0}^{\infty} \rho^j \delta_{t-j}, \sum_{k=1}^{\infty} \rho^{k-1} \delta_{t-k}\right) \\
 &= \rho Var\left(\sum_{k=1}^{\infty} \rho^{k-1} \delta_{t-k}\right) \\
 &= \rho Var(\epsilon_{t-1}) = \rho \sigma_\epsilon^2
 \end{aligned}$$

이고, 같은 방법에 의하여  $j \neq 0$ 에서

$$Cov(\epsilon_t, \epsilon_{t-j}) = \rho^{|j|} \frac{\sigma_\delta^2}{1 - \rho^2} = \rho^{|j|} \sigma_\epsilon^2$$

임을 보일 수 있다.

오차항들 사이에 자기상관이 있는 경우 통상적인 최소제곱법(Ordinary Least Squares method, OLS)에 의하여 얻어진 회귀계수는 불편추정량이지만 최소분산성을 갖지 못하며, 잔차제곱평균 MSE는 오차항의 분산  $\sigma^2$ 의 적절한 추정량이 아닐 수 있고 따라서  $t$ 분포나  $F$ 분포에 근거한 추론에 문제가 있을 수 있다.

일차자기상관계수  $\rho$ 가 존재하는지 검정하기 위하여, Durbin-Watson은 잔차를 이용한 d-통계량을 다음과 같이 정의하고 표본크기  $n$ 과 설명변수 개수에 의존하는 기준치인 상한( $d_U$ )과 하한( $d_L$ )을 계산하여 표로 만들었다. R commander에서는 메뉴 막대에서 [‘모델’ -> ‘수치적 진단’ -> ‘Durbin-Watson 자기상관 검정...’]을 선택한다.

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

먼저 양의 자기상관이 존재하는지 검정하는 절차를 알아보자.

(순서 1) 가설을 세운다.

$$H_0 : \rho = 0, \quad H_1 : \rho > 0$$

(순서 2) 검정통계량  $d$ 의 값을 구한다.

(순서 3) 표에서 해당하는 상한( $d_U$ )과 하한( $d_L$ )을 찾는다.

- 1)  $d < d_L$  이면  $H_0 : \rho = 0$  을 기각하고  $H_1$  을 채택한다.
- 2)  $d > d_U$  이면  $H_0 : \rho = 0$  을 기각하지 않는다.
- 3)  $d_L < d < d_U$  이면 불확정이다.

다음으로 음의 자기상관이 존재하는지 검정하는 절차를 알아보자.

(순서 1) 가설을 세운다.

$$H_0 : \rho = 0, \quad H_1 : \rho < 0$$

(순서 2) 검정통계량  $d$ 의 값을 구한다.

(순서 3) 표에서 해당하는 상한( $d_U$ )과 하한( $d_L$ )을 찾는다.

- 1)  $d > 4 - d_L$  이면  $H_0 : \rho = 0$  을 기각하고  $H_1$  을 채택한다.
- 2)  $d < 4 - d_U$  이면  $H_0 : \rho = 0$  을 기각하지 않는다.
- 3)  $4 - d_U < d < 4 - d_L$  이면 불확정이다.

### 4.3 모형의 변환

두 변수 사이에 가정할 수 있는 회귀모형으로 선형식이 적절하지 않은 경우가 있다. 비선형식을 회귀모형으로 사용하는 경우 통상적인 최소제곱법으로 모수를 추정하기 어려우며 사용하기 복잡하다. [표 1.1]의 자료에서와 같이 산점도가 선형으로 보이지 않더라도, 즉 비선형(nonlinear)으로 보일 때도 적절한 산술적 변환을 통해 다루기 쉬운 선형으로 바꾸어 분석할 수 있는 경우가 있다.

변수  $x$  와  $y$  사이에  $y = \beta_0 e^{\beta_1 x} \epsilon$  인 관계가 있다고 하자. 여기서,  $\beta_0, \beta_1$  은 모수이며  $\epsilon$  은 오차항이다. 관계식의 양변에 자연대수변환(natural logarithmic transformation)을 취하면

$$\ln y = \ln \beta_0 + \beta_1 x + \ln \epsilon$$

이 되며,  $\ln y = y', \ln \beta_0 = \beta_0', \ln \epsilon = \epsilon'$  으로 놓으면 다음과 같이 나타낼 수 있다.

$$y' = \beta_0' + \beta_1 x + \epsilon'$$

여기서  $\epsilon' \sim N(0, \sigma^2)$ , 즉  $\epsilon$  은 대수정규분포(log-normal distribution)를 가정하면, 새로운 모형  $y' = \beta_0' + \beta_1 x + \epsilon'$  은 단순선형회귀모형이므로 통상적인 최소제곱법으로 모수  $\beta_0', \beta_1$  에 대한 추정량  $b_0', b_1$  를 구할 수 있다. 변환 전 모수  $\beta_0$  에 대한 추정량  $b_0$  는  $\hat{\beta}_0' = \ln \hat{\beta}_0 = \ln \hat{\beta}_0' = b_0'$  인 관계로부터  $b_0 = \hat{\beta}_0 = e^{b_0'}$  으로 구한다.

모수  $\beta_0$  에 대한 신뢰수준  $100(1-\alpha)\%$  인 신뢰구간은

$$(b_0' - t(n-2; \alpha/2) \sqrt{S.E.(b_0')}, b_0' + t(n-2; \alpha/2) \sqrt{S.E.(b_0')})$$

인 사실로부터 다음과 같이 구할 수 있다.

$$(e^{b_0' - t(n-2; \alpha/2) \sqrt{S.E.(b_0')}}, e^{b_0' + t(n-2; \alpha/2) \sqrt{S.E.(b_0')}})$$

이와 같이 변환을 통한 선형화가 가능한 모형이 있으며, 보통 양수인 값을 갖는 변수들

에 적용되는 log변환, 분산을 안정화하기 위한 제곱근 변환, 역변환등이 많이 사용되는 변환이며 변수  $y$ 에 대하여  $y^k$  형태인 멱변환(power transformation)의 특정한 경우에 해당한다.

반응변수가 정규성이 의심되는 경우, Box와 Cox (1964)는 다음과 같이 멱변환과 유사한  $y^{(\lambda)}$  형태로 변환을 하면 근사적으로 정규분포를 가정할 수 있는 적절한 상수  $\lambda$ 를 찾는 방법을 제시하였다.

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

이와 같은 변환을 Box-Cox 변환이라 하며, 이렇게 변환된 반응변수를 사용한 회귀모형을 Box-Cox 변환모형이라 한다. R에서는 패키지 ‘MASS’의 boxcox() 함수를 사용하면 상수  $\lambda$ 의 추정값을 구할 수 있다.

정규성을 위해 변환을 하면 분산이 안정화되거나 반대의 경우도 나타나는 경우가 종종 있다. 멱변환을 사용하여 회귀모형을 적합하는 것에 대해 논란은 있으나 log변환 등은 실제로 쓰이는 경우가 많다. 정규성이나 등분산성과 같은 기본 가정들을 극복하기 위한 일반화 선형모형이나 비모수적 방법 등에 대해서는 강 & 유 (2016)를 참고하도록 한다.

앞에서는 설명변수가 하나인 회귀모형을 적합시키는 단순회귀분석을 설명하였다. 그러나 대부분의 실제 현상에서 반응변수와 관련된 설명변수는 두 개 이상인 경우가 일반적이며 적절한 설명변수들을 잘 선택하여 회귀모형을 적합시킬 경우 좀 더 정확한 분석과 예측 등을 할 수 있게 될 것이다.

예를 들어 앞에서 아들의 키를 설명하기 위하여 고려하였던 아버지의 키 외에 어머니의 키도 함께 고려할 수 있고 그 외에도 여러 변수들을 고려할 수 있을 것이다.

이처럼 반응변수의 변화를 설명하기 위하여 두 개 이상의 독립변수를 사용하는 선형회귀 모형을 중선형회귀모형(multiple linear regression model), 또는 중회귀모형(multiple regression model)이라 부른다.

## 5.1 설명변수가 둘인 중회귀모형

반응변수  $y$ 의 변동을 설명하여 주는 독립변수가  $x_1, x_2$  두개인 경우의 중회귀모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

여기에서  $\beta_0, \beta_1, \beta_2$ 는 모집단의 회귀계수(regression coefficient)로서 모수(parameter),  $x_{1i}, x_{2i}$ 는  $i$  번째 독립변수  $x_1$ 과  $x_2$ 의 주어진 값,  $\varepsilon_i$ 는  $i$  번째 측정된  $y_i$ 의 오차항으로

$$\varepsilon_i \sim N(0, \sigma^2), \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$$

을 만족한다고 가정한다.

모수  $\beta_0, \beta_1, \beta_2$ 의 최소제곱추정량을  $b_0, b_1, b_2$ 라 하면 적합된 회귀직선은

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

와 같이 나타낼 수 있다. 최소제곱추정량  $b_0$ ,  $b_1$ ,  $b_2$ 은 오차제곱합

$$Q = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

을 최소로 하는  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ 의 해이다. 정규방정식은

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) x_{1i} = 0$$

$$\frac{\partial Q}{\partial \beta_2} = -2 \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) x_{2i} = 0$$

를  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ 에 대해 정리한 연립방정식이다. 제곱합 기호

$$s_{ij} = \sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j), \quad s_{iy} = \sum_{k=1}^n (x_{ik} - \bar{x}_i)(y_k - \bar{y}_k)$$

를 사용하여 다음과 같이 정규방정식을 나타낼 수 있다.

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 \\ b_1 s_{11} + b_2 s_{12} &= s_{1y} \\ b_1 s_{12} + b_2 s_{22} &= s_{2y} \end{aligned}$$

설명변수의 개수가 세 개 이상이 되면 정규방정식은 네 개 이상의 식을 갖는 복잡한 연립방정식이 되며 행렬을 사용하면 편리하게 된다.

$b_1$ ,  $b_2$ 를 구하기 위하여 행렬을 사용하여 나타내면 다음과 같다.

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} s_{1y} \\ s_{2y} \end{pmatrix}$$

역행렬의 요소를 위첨자를 써서 나타내는 경우  $b_1$ ,  $b_2$ 는 다음과 같이 계산된다.

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}^{-1} \begin{pmatrix} s_{1y} \\ s_{2y} \end{pmatrix} = \begin{pmatrix} s^{11} & s^{12} \\ s^{12} & s^{22} \end{pmatrix} \begin{pmatrix} s_{y1} \\ s_{y2} \end{pmatrix}$$

즉

$$b_i = s^{i1} s_{y1} + s^{i2} s_{y2} = \sum_{j=1}^2 s^{ij} s_{yj}, \quad i=1, 2.$$

## 5.2 행렬의 기초

중회귀모형을 다루려면 행렬을 사용하는 것이 필수적이다.

### 5.2.1 행렬의 정의

$m$ 개의 원소(elements)  $a_{ij}$ , ( $i = 1, \dots, m; j = 1, \dots, n$ )를 다음과 같이 직사각형으로 나열한 것을  $m \times n$  행렬(matrix)이라 한다.

$$A = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}.$$

$k \times 1$  행렬  $\begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{k1} \end{pmatrix}$ 은 열(column)벡터(vector)라 하고, 통계학에서는 기본적으로 열 벡

터를 사용한다.  $1 \times k$  행렬  $(a_{11} \ a_{12} \ \cdots \ a_{1k})$ 은 행(row)벡터라 한다. 행렬이나 벡터가 아니면 스칼라(scalar)라 부른다.

(1) 정방행렬(square matrix): 행과 열의 수가 같은 행렬로  $n \times n$  정방행렬의  $n$ 을 이 행렬의 차수(order) 또는 차원이라 부른다.

(2) 전치행렬(transposed matrix): 행과 열의 위치를 바꾼 행렬로 행렬  $A$ 의 전치행렬은  $A'$ 으로 나타낸다.

$$\text{예 : } A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}, \quad A' = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$$

만약  $A = A'$ 이면  $A$ 를 대칭행렬이라 한다.

(3) 대각행렬(diagonal matrix) : 정방행렬에서 비대각원소가 모두 “0”인 행렬



$$\text{예 : } A = \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix}$$

(4) 단위행렬(identity matrix) : 대각행렬에서 대각원소가 모두 “1”이고 나머지는 모두 “0”인 행렬로 기호  $I$ 로 나타낸다.

## 5.2.2 행렬의 연산

(1) 행렬의 합과 차: 두 행렬  $A = (a_{ij})$  와  $B = (b_{ij})$ 의 합은  $A + B = (a_{ij} + b_{ij})$ 이다.

$$\text{예: } A \pm B = \begin{pmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} \\ a_{21} \pm b_{21} & a_{22} \pm b_{22} \end{pmatrix}$$

(2) 행렬의 스칼라 곱: 행렬  $A = (a_{ij})$ 에 임의의 상수  $\lambda$ 를 곱하면  $\lambda A = (\lambda a_{ij})$ 이다.

$$\text{예: } \lambda A = \lambda \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} \\ \lambda a_{21} & \lambda a_{22} \end{pmatrix}$$

(3) 행렬의 곱: 두 행렬  $A = (a_{ij})$ 와  $B = (b_{ij})$ 의 곱은  $AB = \left( \sum_{k=1}^n a_{ik} b_{kj} \right)$ 이다.

$$\text{예: } AB = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^3 a_{1k} b_{k1} & \sum_{k=1}^3 a_{1k} b_{k2} \\ \sum_{k=1}^3 a_{2k} b_{k1} & \sum_{k=1}^3 a_{2k} b_{k2} \end{pmatrix}.$$

일반적으로  $AB \neq BA$ 이다.

$$\text{예: } A = \begin{pmatrix} 1 & 2 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \text{ 이면 } AB = \begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 7 & 10 \end{pmatrix}.$$

성질:  $AI = IA = A$ ,  $(AB)C = A(BC)$ ,  $C(A+B) = CA + CB$ ,  $(AB)' = B'A'$ .

(4) 행렬식(determinant): 정방행렬  $A$ 의 행렬식은  $|A|$  또는  $\det(A)$ 로 나타낸다.  $n \times n$  행렬  $A$ 에서  $i$ 행과  $j$ 열을 뺀 나머지  $(n-1) \times (n-1)$  행렬의 행렬식을  $M_{ij}$ 로 나타내고  $a_{ij}$ 의 소행렬식(minor)이라 한다. 행렬  $A_{ij} = (-1)^{i+j} M_{ij}$ 를  $a_{ij}$ 의 여인자(cofactor)라 부르며  $A$ 의 행렬식은 다음과 같이 정의한다.

$$|A| = \sum_{j=1}^n a_{ij} A_{ij}, \quad i = 1, \dots, n.$$

$$\text{예: } |a_{11}| = a_{11}, \quad \left| \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \right| = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

성질: A, B가 정방행렬이면  $|AB| = |BA| = |A||B|$ ,  $|A| = |A'|$ ,  $|\lambda A| = \lambda^n |A|$

(5) 트레이스(trace): 정방행렬 A의 대각원소의 합이며  $\text{tr}(A)$ 로 나타낸다.

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \text{이면 } \text{tr}(A) = 1 + 5 + 9 = 15$$

트레이스의 성질에는 다음과 같은 것들이 있다.

$$\text{tr}(AB) = \text{tr}(BA), \text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB),$$

$$\text{tr}(A \pm B) = \text{tr}(A) \pm \text{tr}(B), \text{tr}(kA) = k\text{tr}(A)$$

### 5.2.3 특수한 행렬

(1) 역행렬(inverse matrix): 정방행렬 A에 대하여  $AA^{-1} = A^{-1}A = I$ 를 만족하는  $A^{-1}$ 가 존재하면  $A^{-1}$ 를 A의 역행렬이라 한다. 역행렬은 여인자와 소행렬식을 이용하여거나 소거법을 이용하여 구한다. 역행렬의 성질에는 다음과 같은 것들이 있다.

$$(A')^{-1} = (A^{-1})', (AB)^{-1} = B^{-1}A^{-1}, |A^{-1}| = \frac{1}{|A|}, (kA)^{-1} = \frac{1}{k}A^{-1}$$

(2) 직교행렬(orthogonal matrix): 정방행렬 A의 역행렬과 전치행렬이 같으면 A를 직교행렬이라 한다. 즉,  $A^{-1} = A'$ 인 A를 직교행렬이라 한다. 직교행렬(P)의 성질에는 다음과 같은 것들이 있다.

$$\text{tr}(P'BP) = \text{tr}(B), \det(P'BP) = \det(B), \det(P) = \pm 1$$

(3) 멱등행렬(idempotent matrix): 정방행렬 A가  $AA = A$ 를 만족하면 멱등행렬이라 한다.  $A = A'$ 인 멱등행렬을 대칭멱등행렬이라 한다. 앞으로 사용될 멱등행렬은 모두 대칭인 멱등행렬이다.

$$\text{예: } H = X(X'X)^{-1}X', I - H = I - X(X'X)^{-1}X'$$

### 5.2.4 선형독립과 행렬의 계수

(1) 선형독립: k개의  $n \times 1$  벡터  $c_1, \dots, c_k$ 중 한 벡터가 나머지 (k-1)개 벡터의 선형 결합으로 표시할 수 있으면 즉  $c_i = \lambda_1 c_1 + \dots + \lambda_{i-1} c_{i-1} + \lambda_{i+1} c_{i+1} + \dots + \lambda_k c_k$ 이

면 선형 종속이고 표시할 수 없으면 선형 독립이 된다.

(2) 행렬의 계수(rank): 행렬 A에서 선형독립인 열(또는 행)의 최대개수를 행렬의 계수라 하며 기호  $r(A)$ 로 나타내고, 다음과 같은 성질을 갖는다.

$$r(A) = r(A') = r(A'A) = r(AA')$$

(3) 정칙행렬(nonsingular matrix) : 정방행렬  $A_{n \times n}$ 에 대해서 완전계수  $r(A) = n$ 이면 행렬 A를 정칙행렬이라 하고  $r(A) < n$ 이면 비정칙(singular)행렬이라 한다. 정칙행렬 A의 행렬식은 0이 아니다. 즉,  $\det(A) \neq 0$ .

(4) 선형연립방정식을  $A_{m \times n}x_{n \times 1} = b_{m \times 1}$ 와 같이 나타낼 때  $m = n$ 이고 정칙이면 역행렬이 존재하고 유일한 해  $x = A^{-1}b$ 를 구할 수 있다. A가 비정칙행렬이면 해는 두 가지로 나누어서 생각해야 한다. 행렬 A의 오른쪽에 벡터 b를 첨가시킨 첨가행렬(augmented matrix)을  $B = (A:b)$ 라고 하자. 먼저  $r(A) = r(B)$ 이면 무한히 많은 해가 존재할 수 있다. 다음으로  $r(A) < r(B)$ 이면 해는 존재하지 않는다.

(5) 고유치와 고유벡터(eigen value, eigen vector): 행렬식  $|A_{p \times p} - \lambda I| = 0$ 은  $\lambda$ 의  $p$ 차 다항식이 된다. 이때  $\lambda$ 의 근을 고유치라 하고  $Ax = \lambda x$ 를 만족하는 벡터  $x$ 를 고유벡터라 한다. 고유치의 성질에는 다음과 같은 것들이 있다.

i) 대칭행렬 A의 고유치는 실수다.

$$\text{ii) } tr(A) = \sum_i \lambda_i, \quad tr(A'A) = \sum_i \lambda_i^2, \quad tr(A^{-1}) = \sum_i \lambda_i^{-1}, \quad \det(A) = \lambda_1 \lambda_2 \cdots \lambda_n$$

(6) 멱등행렬의 고유치는 “0” 또는 “1”이다.

(7)

주어진  $n$ 차원 정방행렬 A와 대각행렬 D에 대하여  $P^{-1}AP = D$ 인 정칙행렬 P가 존재하면 A는 대각화가능(diagonalizable)이라 한다. A가 대칭행렬이면 직교행렬에 의해 대각화가능하며, D의 대각요소들은 A의 고유치들이고 P의 열벡터들은 A의 고유벡터들이다.

### 5.2.5 이차형식(quadratic form)

종속변수  $y_i$ 들의  $n \times 1$  확률벡터를  $y$ 라 하고, 행렬  $A = (a_{ij})$ 를  $n \times n$  정방행렬이라 할 때

$$y' Ay = \sum_i^n \sum_{j=1}^n y_i a_{ij} y_j = \sum_i a_{ii} y_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_{ij} y_i y_j$$

을  $y$ 의 이차형식(quadratic form)이라 부른다.

(1) “0”이 아닌 모든 벡터  $y$ 에 대하여  $y' Ay > 0$ 을 만족하는 행렬  $A$ 를 양정치(positive definite)라 하고,  $y' Ay < 0$ 을 만족하는 행렬  $A$ 를 음정치(negative definite)라 한다. 양정치행렬  $A$ 의 고유치는 모두 양수이고, 음정치행렬  $A$ 의 고유치는 모두 음수이다.

(2) “0”이 아닌 모든 벡터  $y$ 에 대해서  $y' Ay \geq 0$ 을 만족하는 행렬  $A$ 는 양반정치(positive semidefinite)라 하고,  $y' Ay \leq 0$ 을 만족하는 행렬  $A$ 는 음반정치(negative semidefinite)라 한다.  $y' Ay = 0$ 을 만족하는 “0”이 아닌 벡터  $y$ 가 존재하는 경우이다. 양반정치행렬  $A$ 의 고유치는 양수이거나 “0”이고, 음반정치행렬  $A$ 의 고유치는 음수이거나 “0”이다.

위의 어떤 경우에도 속하지 않는 행렬은 비정치(indefinite)라 한다. 행렬  $A$ 가 양정치일 때 이차형식  $y' Ay$ 를 양정치 이차형식이라 부르며, 행렬  $A$ 의 속성에 따라 이차형식을 분류할 수 있다.

### 5.2.6 행렬의 분할(partition)

행렬  $A$ 와  $B$ 가 다음과 같이 분할되어 있다 하자.

$$A_{m \times n} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B_{m \times n} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

여기서 각각의 원소에 해당하는 것이 행렬이다. 덧셈과 곱셈은 행렬안의 행렬을 원소처럼 생각하여 한다. 예를 들어 행렬  $A$ 와  $B$ 의 합은 다음과 같다.

$$A + B = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{pmatrix}$$

정칙행렬  $A$ 의 역행렬은

$$A^{-1} = \begin{pmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \end{pmatrix}$$

과 같이 나타낼 수 있고,  $A_{11}$  또는  $A_{22}$ 가 정칙행렬인 경우

$$|A| = |A_{22}| |A_{11} - A_{12}A_{22}^{-1}A_{21}| = |A_{11}| |A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

이 된다.

### 5.2.7 행렬의 미분

두 개의  $k \times 1$  벡터  $a$ 와  $x$ 에 대해

$$z = \sum_{i=1}^k x_i a_i = (x_1, x_2, \dots, x_k) \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = x' a = a' x$$

의  $x$ 에 대한 편도함수(partial derivative)는 다음과 같다.

$$\frac{\partial z}{\partial x} = \begin{pmatrix} \frac{\partial z}{\partial x_1} \\ \vdots \\ \frac{\partial z}{\partial x_k} \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} = a.$$

이차형식  $z = y' A y$ 의  $y$ 에 대한 편도함수(partial derivative)는  $A$ 가 대칭이면

$$\frac{\partial z}{\partial y} = \frac{\partial y' A y}{\partial y} = 2 A y$$

이고,  $A$ 가 대칭이 아니면

$$\frac{\partial z}{\partial y} = \frac{\partial y' A y}{\partial y} = (A + A') y$$

이다.

## 5.3 이차 형식의 분포

분산분석표에서 구했던 제곱합  $SSR$ ,  $SSE$ ,  $SST$  등은 모두 종속변수의 실수값 함수

의 하나인 이차형식으로 표시할 수 있다. 제곱합을 이차형식으로 나타내면 자유도를 대수적으로 구할 수 있고 기대값과 분산등을 구하거나 분포를 구하여 이론을 전개할 때 매우 편리하다.

### 5.3.1 다변량정규분포(multivariate normal distribution)

확률벡터  $y' = (y_1, y_2, \dots, y_n)$ 가 기대값이  $\mu' = (\mu_1, \mu_2, \dots, \mu_n)$ 이고 분산공분산행렬이  $V$ 인 다변량정규분포를 한다면  $y \sim N(\mu, V)$ 로 나타내고, 확률밀도함수는 다음과 같다.

$$f(y_1, y_2, \dots, y_n) = \frac{e^{-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)}}{(2\pi)^{\frac{1}{2}n} |V|^{\frac{1}{2}}}.$$

특히  $y' = (x, y)$ 인 이변량정규분포의 확률밀도함수는 다음과 같다.

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}$$

여기에서  $\mu_x = E(X)$ ,  $\mu_y = E(Y)$ ,  $\sigma_x^2 = Var(X)$ ,  $\sigma_y^2 = Var(Y)$ 이고,  $\rho$ 는 모상관계수이다. 참고로  $x$ 가 주어졌을 때,  $Y$ 의 조건부기대값은 다음과 같다.

$$E(Y|x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x) = (\mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x) + \rho \frac{\sigma_y}{\sigma_x} x$$

따라서,  $\beta_0 = (\mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x)$ ,  $\beta_1 = \rho \frac{\sigma_y}{\sigma_x}$ 으로 놓으면 반응함수가  $E(Y|x) = \beta_0 + \beta_1 x$ 과 같은 회귀직선모형이 된다.

확률벡터  $y' = (y_1, y_2, \dots, y_n)$ 이  $N(0, I)$ 인 분포를 하면,  $y_i$ 들은 각각  $N(0, 1)$ 인 표준정규분포를 하고 서로 독립이므로  $y_i$ 들의 제곱합은 자유도  $n$ 인 카이제곱분포를 하며

$$y_1^2 + y_2^2 + \dots + y_n^2 \sim \chi^2(n)$$

와 같이 나타낸다. 즉, 이차형식  $y'y = y'Iy = \sum_{i=1}^n y_i^2 \sim \chi^2(n)$ 이다.

서로 독립인 두 이차형식  $Q_1, Q_2$ 가 각각  $Q_1 \sim \chi^2(n_1)$ ,  $Q_2 \sim \chi^2(n_2)$ 이면

$$\frac{Q_1/n_1}{Q_2/n_2} \sim F(n_1, n_2)$$

와 같이 분자의 자유도  $n_1$ , 분모의 자유도  $n_2$ 인 F-분포를 한다.

이차형식  $Q$ 가 자유도  $n$ 인 카이제곱분포를 하고  $y$ 는  $N(0, 1)$ 인 확률변수일 때 서로 독립이면

$$\frac{y}{Q/n} \sim t(n)$$

와 같이 자유도  $n$ 인 t-분포를 한다.

확률벡터  $y' = (y_1, y_2, \dots, y_n)$ 이  $N(\mu, I)$ 인 분포를 하면  $y_i$ 들은 각각  $N(\mu_i, 1)$ 인 정규 분포를 하고 서로 독립이고  $y_i$ 들의 제곱합은 비중심모수(noncentrality parameter)가  $\lambda = \mu' \mu / 2$ 이고 자유도  $n$ 인 비중심 카이제곱분포를 하며

$$y_1^2 + y_2^2 + \dots + y_n^2 \sim \chi^2(n, \lambda)$$

와 같이 나타낸다. 이차형식을 사용하여 나타내면 다음과 같다.

$$y' y = y' I y = \sum_{i=1}^n y_i^2 \sim \chi^2(n, \lambda), \quad \lambda = \frac{1}{2} \mu' \mu.$$

만약  $y \sim N(\mu, 1)$ 이면  $y^2 \sim \chi^2(1, \lambda)$ ,  $\lambda = \frac{1}{2} \mu^2$  이다.

따라서,  $y_i \sim iid N(\mu_i, 1)$ 이면  $\sum_{i=1}^n y_i^2 \sim \chi^2(n, \lambda)$ ,

$$\lambda = \frac{1}{2} \mu_1^2 + \frac{1}{2} \mu_2^2 + \dots + \frac{1}{2} \mu_n^2 = \frac{1}{2} (\mu_1^2 + \dots + \mu_n^2) = \frac{1}{2} \mu' \mu, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

추가로 다음 성질들을 소개한다.

- (1) 만약  $y' y \sim \chi^2(n, \lambda)$  일 때  $\lambda = 0$  이면  $y' y \sim \chi^2(n)$  이다.
- (2) 이차형식  $Q_1, Q_2, \dots, Q_k$ 가 서로 독립이고 각각 자유도가  $n_1, n_2, \dots, n_k$ 이며 비중심모수가  $\lambda_1, \lambda_2, \dots, \lambda_k$  이면, 즉  $Q_i \sim \chi^2(n_i, \lambda_i)$  이면

$$Q = \sum_{i=1}^k Q_i \sim \chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \lambda_i\right).$$

- (3) 이차형식  $Q_1 \sim \chi^2(n_1, \lambda)$ ,  $Q_2 \sim \chi^2(n_2)$  이고 서로 독립이면

$$\frac{Q_1/n_1}{Q_2/n_2} \sim F(n_1, n_2, \lambda)$$

는 비중심모수가  $\lambda$ 이고 분자의 자유도  $n_1$ , 분모의 자유도  $n_2$ 인 비중심  $F$ -분포 (*noncentral F-distribution*)를 한다고 말한다.

### 5.3.2 이차형식의 분포

이차형식의 분포에 대한 이론은 회귀분석의 이해에 중요한 역할을 한다. 자세한 내용은 박성현(1999)를 참고하고 기본적인 성질들만 알아보자.

[정리 5.1]

만약  $y \sim N(\mu, V)$  이면

$$1) E(y' A y) = \text{tr}(A V) + \mu' A \mu$$

$$2) \text{Cov}(y, y' A y) = 2 V A \mu$$

(증명)

$$1) E(y' A y) = E(\text{tr}(y' A y)) = E(\text{tr}(A y y')) = \text{tr}(E(A y y'))$$

$$= \text{tr}(A E(y y')) = \text{tr}(A (V + \mu \mu')) = \text{tr}(A V + A \mu \mu')$$

$$= \text{tr}(A V) + \text{tr}(\mu' A \mu) = \text{tr}(A V) + \mu' A \mu$$

$$2) \text{Cov}(y, y' A y) = E(y - \mu)(y' A y - \text{tr}(A V) - \mu' A \mu)$$

$$= E(y - \mu)[(y - \mu)' A (y - \mu) + \mu' A y + y' A \mu - \mu' A \mu - \mu' A \mu - \text{tr}(A V)]$$

$$= E(y - \mu)[(y - \mu)' A (y - \mu) + 2 y' A \mu - 2 \mu' A \mu - \text{tr}(A V)]$$

$$= E(y - \mu)[(y - \mu)' A (y - \mu) + 2(y - \mu)' A \mu - \text{tr}(A V)]$$

$$= E(y - \mu)(y - \mu)' A (y - \mu) + 2 E(y - \mu)(y - \mu)' A \mu - E(y - \mu) \text{tr}(A V)$$

$$(E(y - \mu)(y - \mu)' A (y - \mu) = 0, E(y - \mu) = 0)$$

$$= 2 V A \mu$$

[정리 5.2]

$$y \sim N(\mu, V) \text{ 이면 } \text{Var}(y' A y) = 2 \text{tr}(A V)^2 + 4 \mu' A V A \mu.$$

$$\text{만약, } \mu = 0 \text{ 이면 } \text{Var}(y' A y) = 2 \text{tr}(A V A V).$$



## [정리 5.3]

확률벡터  $y \sim N(\mu, V)$  일 때 이차형식이  $y'Ay \sim \chi^2(r(A), \frac{1}{n}\mu'A\mu)$  가 되기 위한 필요충분조건은  $AV$ 가 멱등행렬인 것이다. 여기서  $r(A)$ 는  $A$ 의 계수다. 위 정리 5.3의 특수한 경우로 다음 정리를 얻을 수 있다.

## [정리 5.4]

- 1) 확률벡터  $y \sim N(0, I)$  일 때 이차형식이  $y'Ay \sim \chi^2(p)$ 가 되기 위한 필요충분조건은  $A$ 가 계수  $p$ 인 멱등행렬인 것이다.
- 2) 확률벡터  $y \sim N(\mu, I\sigma^2)$  일 때  $y'y/\sigma^2 \sim \chi^2(n, \frac{1}{2}\mu'\mu/\sigma^2)$ 이다.
- 3) 확률벡터  $y \sim N(\mu, I)$  일 때 이차형식이  $y'Ay \sim \chi^2(p, \frac{1}{2}\mu'A\mu)$ 가 되기 위한 필요충분조건은  $A$ 가 계수  $p$ 인 멱등행렬인 것이다.

## [정리 5.5]

확률벡터  $y \sim N(\mu, V)$  일 때 두 이차형식  $y'Ay$ 와  $y'By$ 가 서로 독립이기 위한 필요충분조건은  $AVB = 0$  또는  $BVA = 0$ 이다.

## [정리 5.6]

확률벡터  $y \sim N(\mu, I)$  이면 행렬  $A_i$ ,  $i=1, 2, \dots, k$ 의 계수가 각각  $n_i$ 인  $k$ 개의 이차형식  $y'A_1y, y'A_2y, \dots, y'A_ky$ 는 다음의 세 조건 중 두 가지만 만족하면 서로 독립이며 각각  $\chi^2(n_i, \lambda_i)$ 인 분포를 한다. 여기서  $\lambda_i = \frac{1}{2}\mu'A_i\mu$ 이다.

- 1) 모든  $A_i$ 는 멱등행렬
- 2)  $\sum_{i=1}^k A_i$ 는 멱등행렬
- 3)  $A_i \cdot A_j = 0$ ,  $i \neq j$

## [정리 5.7] Cochran - Fisher 정리

확률벡터  $y \sim N(\mu, I_n)$  이고  $y'y = \sum_{i=1}^k y'A_iy$  이면  $y'A_iy \sim \chi^2(k_i)$ 이며 서로 독립이기 위한 필요충분조건은  $\sum k_i = n$ 이다. 여기서  $k_i = r(A_i)$ 이고  $n = r(I)$ 이다.

## 5.4 중회귀모형의 추정

설명변수가 두 개인 가장 기본적인 중회귀모형에서  $i$ 번째 측정치는

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = (1, x_{i1}, x_{i2}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \epsilon_i$$

와 같이 벡터를 사용하여 나타낼 수 있다. 전체  $n$ 개 자료에 대하여 행렬을 사용하여 나타내면 다음과 같다.

$$y = X\beta + \epsilon$$

여기서

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

이고, 오차항벡터의 기대값과 분산공분산행렬은 다음과 같다.

$$E(\epsilon) = E \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = \begin{pmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = 0$$

$$\begin{aligned} V(\epsilon) &= E(\epsilon\epsilon') \\ &= E \begin{pmatrix} \epsilon_1^2 & \epsilon_1\epsilon_2 & \cdots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & \cdots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \cdots & \epsilon_n^2 \end{pmatrix} \begin{pmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_n\epsilon_1) & E(\epsilon_n\epsilon_2) & \cdots & E(\epsilon_n^2) \end{pmatrix} \\ &= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \sigma^2 \cdot I \end{aligned}$$

따라서,  $y = X\beta + \epsilon$ ,  $E(\epsilon) = 0$ ,  $V(\epsilon) = \sigma^2 I$ 와 같이 벡터를 사용한 표현을 할 수 있다. 이러한 벡터를 사용한 표현은  $k$ 개의 설명변수가 있을 때도 요소의 개수만 다를 뿐 동일하다.

모수  $\beta$ 에 대한 최소제곱추정량은 다음 오차제곱합을 최소로 하는 해  $b$ 이다.

$$\begin{aligned} Q &= \sum_{i=1}^n \epsilon_i^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta) = y'y - \beta' X'y - y' X\beta + \beta' X' X\beta \\ &= y'y - 2\beta' X'y + \beta' X' X\beta. \end{aligned}$$

따라서,  $Q$ 를  $\beta$ 에 대하여 미분하여  $\frac{\partial Q}{\partial \beta} = -2X'y + 2X'X\beta = 0$ 으로부터  $X'Xb = X'y$

이 유도되고 이러한 모양의 방정식을 회귀모형에 대한 정규방정식이라 한다. 양변에  $(X'X)^{-1}$ 를 곱하여 최소제곱추정량  $b = (X'X)^{-1}X'y$ 를 얻는다.

단순회귀모형을 행렬과 벡터를 사용하여 오차제곱합을 표현하면 다음과 같다.

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^n \epsilon_i^2 = \epsilon' \epsilon = (y - X\beta)'(y - X\beta).$$

여기서,  $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ ,  $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ ,  $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ ,  $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$  라 하면

$\frac{\partial Q}{\partial \beta} = -2X'(y - X\beta) = 0$  이므로  $X'y - X'X\beta = 0$  이다. 따라서,  $X'X\beta = X'y$ 인 정규방정식의 해는 다음과 같다.

$$b = (X'X)^{-1}X'y = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}.$$

모수에 대한 최우추정량을 구하려면 오차에 대한 분포를 알아야 한다. 만약 오차항이 정규분포  $\epsilon \sim N(0, \sigma^2 I)$ 를 따른다면 우도함수는 다음과 같다.

$$\begin{aligned} f(\epsilon; \beta, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\sum \epsilon_i^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{\epsilon' \cdot \epsilon}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right) \end{aligned}$$

이 우도함수를 최대로 하는 모수를 좀 더 쉽게 찾기 위하여 자연로그를 취하면

$$L = \ln f(\epsilon; \beta, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)$$

와 같은 로그우도함수가 된다. 이 로그우도함수를 모수에 대해 미분하여 0으로 놓은

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= \frac{1}{\sigma^2} (X'y - X'X\beta) = 0 \\ \frac{\partial L}{\partial \sigma^2} &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{\sigma^4} (y - X\beta)'(y - X\beta) = 0 \end{aligned}$$

으로부터  $X'X\tilde{b} = X'y$  과  $\tilde{\sigma}^2 = \frac{1}{n} (y - X\tilde{b})'(y - X\tilde{b})$  이 얻어진다. 따라서 회귀계수에 대한 최우추정량은  $\tilde{b} = (X'X)^{-1}X'y$  이며 최소제곱추정량과 동일하지만, 분산에 대한 최우추정량  $\tilde{\sigma}^2$  은 MSE와 다름을 알 수 있다.

## 5.5 추정량들의 성질

최소제곱추정량  $b = (X'X)^{-1}X'y$  의 기대값을 구하면 다음과 같고 따라서 불편성을 가짐을 알 수 있다.

$$E(b) = E((X'X)^{-1}X'y) = (X'X)^{-1}X'E(y) = (X'X)^{-1}X'X\beta = \beta.$$

확률벡터  $y$  의 분산이  $V(y) = V(X\beta + \epsilon) = V(\epsilon) = I\sigma^2$  이므로 최소제곱추정량  $b$  의 분산은 다음과 같이 구할 수 있다.

$$\begin{aligned} V(b) &= V((X'X)^{-1}X'y) = (X'X)^{-1}X'V(y)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} = (X'X)^{-1}\sigma^2 \end{aligned}$$

[정리 5.1] Gauss-Markov 정리

중선형회귀모형  $y = X\beta + \epsilon$ ,  $\epsilon \sim N(0, \sigma^2 I)$  에서  $r(X) = k + 1$ ,  $E(\epsilon) = 0$  그리고  $\text{Var}(\epsilon) = \sigma^2 I$  이면  $\beta$  의 최소제곱추정량  $b = (X'X)^{-1}X'y$  는 최소분산불편추정량 (minimum variance linear unbiased estimator) 또는 최량선형불편추정량 (Best Linear Unbiased Estimator: BLUE)이다.

(증명)

## ① 선형성

$y$ 의 선형함수가 되어야 하므로  $b = A \cdot y$  형태가 되어야 한다. 여기서  $A$ 는 임의의 행렬로  $A = (X'X)^{-1}X' + B$ 로 놓을 수 있다. 왜냐하면 항상

$$A = (X'X)^{-1}X' + A - (X'X)^{-1}X' = (X'X)^{-1}X' + B$$

와 같이 놓을 수 있기 때문이다. 따라서  $b$ 는  $\beta$ 의 선형추정량이다.

## ② 불편성

$b$ 의 기대값은 다음과 같다.

$$E(b) = E[(X'X)^{-1}X'Y + BY] = (X'X)^{-1}X'X\beta + BE(Y) = \beta + B \cdot X\beta.$$

불편성을 만족하기 위하여 모든  $\beta$ 에 대해서  $\beta + B \cdot X\beta = \beta$ 가 되어야 하므로  $B \cdot X\beta = 0$ 이어야 한다.

$$\beta = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots$$

을 대입해 보면,  $BX = 0$ 이 된다.

$$\text{예) } B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$$

## ③ 최소 분산성

$b$ 의 분산은 다음과 같다.

$$\begin{aligned} V(b) &= V[(X'X)^{-1}X' + B]Y \\ &= [(X'X)^{-1}X' + B][(X'X)^{-1}X' + B]' \cdot \sigma^2 \\ &= [(X'X)^{-1}X'X(X'X)^{-1} + BX(X'X)^{-1} + (X'X)^{-1}X'B' + B'B] \cdot \sigma^2 \\ &= [(X'X)^{-1} + B'B] \cdot \sigma^2 \quad (\because BX = 0) \end{aligned}$$

여기서  $\text{var}(b_i)$ 가 최소가 되기 위해서는  $[(X'X)^{-1} + B'B]\sigma^2$ 가 최소가 되어야 한다. 따라서  $(B'B)_{ii}$ 가 최소가 되어야 한다.  $B'B$ 가 양반정치 행렬이므로

$(B'B)_{ii} \geq 0$  이고  $(B'B)_{ii} = 0$  이 최소이다. 즉  $B = 0$ 이어야 한다.

다음으로  $\sigma^2$ 의 불편추정량이 다음과 같음을 보이자.

$$\begin{aligned}\hat{\sigma}^2 &= MSE = \frac{SSE}{n-k-1} = \frac{Y'Y - b'X'Y}{n-k-1} \\ &= \frac{1}{n-k-1} \cdot Y'Y - Y'X(X'X)^{-1}X'Y \\ &= \frac{1}{n-k-1} \cdot Y'(I - X'(X'X)^{-1}X')Y\end{aligned}$$

기대값을 구하는 절차는 다음과 같다.

$$E(\hat{\sigma}^2) = E(MSE) = \frac{1}{n-k-1} E[Y'(I - X'(X'X)^{-1}X')Y]$$

여기서  $E(Y'AY) = \text{tr}(AV) + \mu'A\mu$  인 사실을 이용하여

$$V = \text{Var}(Y) = I \cdot \sigma^2, \quad \mu = E(Y) = X \cdot \beta$$

인 관계로부터

$$\mu'A\mu = \beta X'(I - X'(X'X)^{-1}X')X\beta = 0$$

이고

$$\begin{aligned}\text{tr}(AV) &= \text{tr}(I - X'(X'X)^{-1}X')\sigma^2 \\ &= [\text{tr}(I) - \text{tr}(X(X'X)^{-1}X')]\sigma^2 \\ &= [\text{tr}(I) - \text{tr}((X'X)^{-1}X'X)]\sigma^2 \\ &= [n - (k+1)]\sigma^2 \\ &= (n-k-1)\sigma^2\end{aligned}$$

이므로

$$E(\hat{\sigma}^2) = E(MSE) = \sigma^2.$$

한편 최우 추정량은

$$\tilde{\sigma}^2 = \frac{SSE}{n} = \frac{n-k-1}{n} \cdot \frac{SSE}{n-k-1} = \frac{n-k-1}{n} \cdot \sigma^2$$

이고, 기대값은 다음과 같다.

$$E(\widetilde{\sigma^2}) = \frac{n-k-1}{n} \cdot E(\widehat{\sigma^2}) = \frac{n-k-1}{n} \cdot \sigma^2$$

즉 최우추정량은 편의 추정량 (biased)이다.

### 5.5.1 제곱합과 이차형식

분산분석에 사용되었던 총제곱합, 회귀제곱합, 잔차제곱합 등은 이차형식으로 나타낼 수 있다. 총제곱합 SST를 이차형식으로 나타내면 다음과 같다.

$$\begin{aligned} SST &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = \sum y_i^2 - (\sum y_i)^2/n \\ &= y'y - \frac{1}{n} y'11'y = y'(I - \frac{11'}{n})y = y'(I - \frac{J}{n})y = y'Ay. \end{aligned}$$

여기에서

$$\begin{aligned} 1 &= \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, J = 11' = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (1 \ 1 \cdots 1) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}, \\ I &= \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, y'y = (y_1 \ y_2 \cdots y_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^n y_i^2 = y'Iy, \\ (\sum_{i=1}^n y_i)^2 &= (\sum_{i=1}^n y_i)(\sum_{i=1}^n y_i) = y'11'y = y'Jy. \end{aligned}$$

회귀제곱합 SSR을 이차형식으로 나타내보자.

잔차의 합  $\sum e_i = \sum (y_i - \hat{y}_i) = 0$  이므로  $\sum \hat{y}_i = \sum y_i$  임을 이용하면

$$\begin{aligned}
SSR &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^n \hat{y}_i^2 - 2 \sum_{i=1}^n \bar{y} \hat{y}_i + n \bar{y}^2 \\
&= \sum_{i=1}^n \hat{y}_i^2 - 2 \bar{y} \sum_{i=1}^n \hat{y}_i + n \bar{y}^2 \\
&= \sum_{i=1}^n \hat{y}_i^2 - n \bar{y}^2
\end{aligned}$$

이다.

마지막 식의 첫 항은 다음과 같이 나타낼 수 있다.

$$\begin{aligned}
\sum_{i=1}^n \hat{y}_i^2 &= \hat{y}' \hat{y} \\
&= (Xb)'(Xb) = b' X' X b \\
&= [(X' X)^{-1} X' y]' X' X [(X' X)^{-1} X' y] \\
&= y' X (X' X)^{-1} X' X (X' X)^{-1} X' y \\
&= y' X (X' X)^{-1} X' y.
\end{aligned}$$

따라서

$$\begin{aligned}
SSR &= y' X (X' X)^{-1} X' y - y' \frac{J}{n} y \\
&= y' [X(X' X)^{-1} X' - \frac{J}{n}] y = y' B y.
\end{aligned}$$

잔차제곱합 SSE를 이차형식으로 나타내보자.

잔차의 성질 중  $\sum e_i \hat{y}_i = 0$  임을 이용하면

$$\begin{aligned}
SSE &= \sum (y_i - \hat{y}_i)^2 \\
&= \sum y_i^2 - 2 \sum y_i \hat{y}_i + \sum \hat{y}_i^2 \\
&= \sum y_i^2 - 2 \sum (y_i - \hat{y}_i + \hat{y}_i) \hat{y}_i + \sum \hat{y}_i^2 \\
&= \sum y_i^2 - 2 \left( \sum e_i \hat{y}_i + \sum \hat{y}_i^2 \right) + \sum \hat{y}_i^2 \\
&= \sum y_i^2 - \sum \hat{y}_i^2.
\end{aligned}$$

으로 나타낼 수 있다.



또 다른 방법으로는 다음과 같이 나타낼 수 있다.

$$SSE = SST - SSR = \sum y_i^2 - n\bar{y}^2 - (\sum \hat{y}_i^2 - n\bar{y}^2) = \sum y_i^2 - \sum \hat{y}_i^2.$$

따라서

$$\begin{aligned} SSE &= y'y - y'(X'X)^{-1}X'y \\ &= y'[I - (X'X)^{-1}X']y = y'Cy \end{aligned}$$

## 5.6 분산분석

중회귀모형이 적합한지를 검정하기 위하여 다음과 같은 가설을 세운다.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0, \quad H_1 : \text{적어도 하나의 } i \text{에 대하여 } \beta_i \neq 0$$

만약 귀무가설  $H_0$ 가 기각되면 독립변수  $x_1, x_2, \dots, x_k$ 중 적어도 하나는 반응량의 변동에 영향을 끼친다는 것이다.

회귀모형에 대한 분산분석은 총변동( $SST$ )을 회귀식에 의해 설명되는 회귀변동( $SSR$ )과 회귀식에 의해 설명이 안 되는 변동인 잔차변동( $SSE$ )으로 분해하는데서 시작된다. 설명변수가  $k$ 개인 일반적인 경우 이들 변동 즉 제곱합들을 행렬을 사용하여 나타내보자.

$$SST = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2 = y'y - y' \frac{J}{n} y = y'(I - \frac{J}{n})y$$

$P_x = X(X'X)^{-1}X'$ 로 놓으면

$$\begin{aligned} SSR &= \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n(\bar{y})^2 \\ &= \hat{y}'\hat{y} - n\bar{y}^2 = b'X'Xb - y' \frac{J}{n} y \\ &= [y'X(X'X)^{-1}]X'X[(X'X)^{-1}X'y] - y' \frac{J}{n} y \\ &= y'X(X'X)^{-1}X'y - y' \frac{J}{n} y \\ &= y'(P_x - \frac{J}{n})y \end{aligned}$$

$$\begin{aligned}
 SSE &= SST - SSR \\
 &= y'(I - \frac{J}{n})y - y'(P_x - \frac{J}{n})y \\
 &= y'(I - P_x)y
 \end{aligned}$$

식  $SST = SSR + SSE$ 에서 다음과 같은 자유도 관계식을 유도할 수 있다.

$SST$ 의 자유도  $(n-1) = SSR$ 의 자유도  $(k) + SSE$ 의 자유도  $(n-k-1)$ .

[표 5.1] 중회귀모형의 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$	$F_\alpha$
회귀	SSR	k	MSR=SSR/k	MSR/MSE	$F_\alpha(k, n-k-1)$
잔차	SSE	n-k-1	MSE=SSE/n-k-1		
계	SST	n-1			

검정통계량  $F_0 = MSR/MSE$ 는 분자의 자유도가  $k$ 이고 분모의 자유도가  $n-k-1$ 인  $F(k, n-k-1)$ 분포를 따르며 다음과 같은 가설에 대한 검정을 한다.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0, \quad H_1 : \text{적어도 하나의 } i \text{에 대하여 } \beta_i \neq 0.$$

따라서,  $F_0 > F_\alpha(k, n-k-1)$ 이면  $H_0$ 를 기각하고, 아니면  $H_0$ 를 채택한다.

검정통계량  $F_0$ 에 의한 가설검정의 통계적 의미를 알아보자. 우선 제곱합  $SSE$ 와  $SSR$ 을 이차형식으로 나타내면  $SSE = y'(I - P_x)y$ ,  $SSR = y'(P_x - \frac{J}{n})y$  이었고, 각 이차형식의 행렬  $A = I - P_x$ 는 멱등행렬 즉  $(I - P_x)(I - P_x) = I - P_x$  이고,  $B = P_x - \frac{J}{n}$ 도 멱등행렬 즉  $(P_x - \frac{J}{n})(P_x - \frac{J}{n}) = P_x - \frac{J}{n}$ 이다.

앞에서  $y \sim N(\mu, V)$ 일 때 두 이차형식  $y'Ay$ ,  $y'By$ 가 서로 독립일 필요충분조건은  $AVB=0$  또는  $BVA=0$  라 하였고  $V = I\sigma^2$ 이므로 다음과 같이  $SSE$ 와  $SSR$ 은 서

로 독립임을 보일 수 있다.

$$(I - P_x) \cdot I\sigma^2 \cdot (P_x - \frac{J}{n}) = (I - P_x)(P_x - \frac{J}{n}) = P_x - P_x \cdot P_x - \frac{J}{n} - P_x \cdot \frac{J}{n} = 0$$

이차형식의 기대값은  $E(y' Ay) = \text{tr}(AV) + \mu' A \mu$  이고,  $(I - P_x)X = 0$  임을 사용하면

$$\begin{aligned} E(y'(I - P_x)y) &= \text{tr}((I - P_x) \cdot I\sigma^2) + \beta' X' (I - P_x) X \beta = \text{tr}(I - P_x) \sigma^2 \\ &= [\text{tr}(I) - \text{tr}(P_x)] \sigma^2 = [n - (k + 1)] \sigma^2 = (n - k - 1) \sigma^2 \end{aligned}$$

이므로

$$E(MSE) = \frac{1}{n - k - 1} E(y'(I - P_x)y) = \sigma^2.$$

마찬가지로

$$\begin{aligned} E\left(y' \left(P_x - \frac{J}{n}\right) y\right) &= \text{tr}\left(\left(P_x - \frac{J}{n}\right) I \sigma^2\right) + \beta' X' \left(P_x - \frac{J}{n}\right) X \beta \\ &= \left(\text{tr}(P_x) - \text{tr}\left(\frac{J}{n}\right)\right) \sigma^2 + \beta' X' \left(P_x - \frac{J}{n}\right) X \beta \\ &= (k + 1 - 1) \sigma^2 + \beta' X' \left(P_x - \frac{J}{n}\right) X \beta \\ &= k \cdot \sigma^2 + \beta' X' \left(P_x - \frac{J}{n}\right) X \beta \end{aligned}$$

이므로

$$E(MSR) = \frac{1}{k} E\left(y' \left(P_x - \frac{J}{n}\right) y\right) = \sigma^2 + \frac{1}{k} \beta' X' \left(P_x - \frac{J}{n}\right) X \beta.$$

따라서

$$\frac{E(MSR)}{E(MSE)} = \frac{\sigma^2 + \frac{1}{k} \beta' X' \left(P_x - \frac{J}{n}\right) X \beta}{\sigma^2} = 1 + \frac{1}{k \sigma^2} \beta' X' \left(P_x - \frac{J}{n}\right) X \beta.$$

만약  $\beta_1 = \beta_2 = \dots = \beta_k = 0$  이면  $\frac{E(MSR)}{E(MSE)} = 1$  이고, 아니면  $\frac{E(MSR)}{E(MSE)} > 1$  이 된다. 따

라서, 귀무가설  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$  의 검정통계량으로  $F_0 = \frac{MSR}{MSE}$  을 사용할 수

있다.

앞에서 확률벡터  $y \sim N(\mu, V)$  일 때, 이차형식  $y' Ay \sim \chi^{2'}(r(A), \frac{1}{2} \mu' A \mu)$  가 되기 위

한 필요충분조건은  $AV$ 가 멱등행렬, 즉  $\frac{\left(P_x - \frac{J}{n}\right)}{\sigma^2} \cdot I\sigma^2$ 이 멱등행렬인 것이므로

$$\frac{SSR}{\sigma^2} = \frac{y' \left(P_x - \frac{J}{n}\right) y}{\sigma^2} \sim \chi^2 \left(k, \frac{1}{2\sigma^2} \beta' X' \left(P_x - \frac{J}{n}\right) X \beta\right).$$

또는,  $X' P_x X = X' X = X' I X$  인 관계를 사용하여

$$\frac{SSR}{\sigma^2} = \frac{y' \left(P_x - \frac{J}{n}\right) y}{\sigma^2} \sim \chi^2 \left(k, \frac{1}{2\sigma^2} \beta' X' \left(I - \frac{J}{n}\right) X \beta\right)$$

또한,  $\frac{(I - P_x)}{\sigma^2} \cdot I\sigma^2$ 이 멱등행렬이고  $\frac{1}{2} \mu' (I - P_x) \mu = 0$ 임을 사용하면

$$\frac{SSE}{\sigma^2} = \frac{y' (I - P_x) y}{\sigma^2} \sim \chi^2 (n - k - 1).$$

따라서

$$F_0 = \frac{MSR}{MSE} = \frac{(SSR/\sigma^2)/k}{(SSE/\sigma^2)/(n-k-1)} \sim F(k, n-k-1, \lambda)$$

여기서, 비중심모수  $\lambda = \frac{1}{2\sigma^2} \beta' X' \left(I - \frac{J}{n}\right) X \beta$ 이다.

따라서,  $F_0 = \frac{MSR}{MSE}$ 는  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ 이 참일 때 분자의 자유도가  $k$ 이고

분모의 자유도가  $(n-k-1)$ 인 중심  $F$  분포를 따르며  $F_0 > F_\alpha(k, n-k-1)$ 이면  $H_0$ 를 기각하고 적합한 회귀모형이 유의하다고 결론 내린다.

## 5.7 회귀방정식의 정확도

적합된 회귀모형이 신뢰할만한 것인지, 주어진 자료들을 얼마나 잘 설명하고 있는지와 주어진 설명변수들의 값에서 평균반응값을 어느 정도 정확하게 예측할 수 있는

지를 평가하는 척도로 잔차제곱평균, 분산분석표의 F-검정통계량, 결정계수, 회귀계수 추정량의 분산, 종속변수의 추정량의 분산 등을 사용할 수 있다.

첫 번째로 앞의 분산분석에서 논의되었던 잔차제곱평균 MSE는 오차항의 분산에 대한 불편추정량이었고 MSE가 작으면 자료들이 적합한 회귀방정식 주위에 밀집되어 신뢰할 수 있다는 의미가 된다.

두 번째로 분산분석표의 F-검정통계량이 크다면 적합한 회귀모형이 유의하며 자료들을 잘 설명하고 있다는 의미를 갖는다. 유의성을 판단하는 기준으로 F-분포의 기각치  $F_{\alpha}(k, n-k-1)$ 를 사용할 수 있다.

세 번째로 결정계수(coefficient of determination,  $R^2$ )은 회귀모형에 의하여 설명되는 회귀변동 SSR이 총변동 SST에 비하여 어느 정도인가를 나타내는 척도로 다음과 같이 나타낸다.

$$R^2 = \frac{SSR}{SST}, \quad 0 \leq R^2 \leq 1$$

설명변수의 개수가 많을수록 회귀제곱합이 커지므로  $R^2$ 값도 따라서 커지는 성질을 갖기 때문에  $R^2$ 값이 크다고 회귀모형이 최적이라 말할 수는 없지만,  $R^2$ 값이 작은 모형은 적합하지 않다고 판단한다.

네 번째로 회귀계수  $\beta$ 의 최소제곱추정량  $b$ 는 불편성을 가자며 분산공분산행렬이

$$V(b) = (X'X)^{-1}\sigma^2$$

이다. 행렬  $(X'X)^{-1}$ 의 원소를  $c_{ii}(i, j=0, 1, \dots, k)$ 로 나타낼 때 특별히 관심 있는 설명변수가  $x_i$ 라 하면 추정량  $b_i$ 의 분산이 작기를 바라기 때문에  $c_{ii}$ 의 값이 작도록 계획행렬(design matrix)  $X$ 를 설계한다.

마지막으로 주어진 설명변수  $(x_1, x_2, \dots, x_k)$ 에서 평균반응값을 예측하는데 사용할 추정량은  $x = (1, x_1, \dots, x_k)$ ,  $b' = (b_0, b_1, \dots, b_k)$ 로 놓을 때

$$\hat{y} = b_0 + b_1x_{i1} + \dots + b_kx_k = (1, x_1, \dots, x_k) \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = x'b$$

이고  $\hat{y}$ 의 분산은 다음과 같으며 이 값이 작을수록 정확하다고 생각한다.

$$V(\hat{y}) = V(x'b) = x'V(b)x = x'(X'X)^{-1}x\sigma^2.$$

## 5.8 절편 없는 중회귀 모형

일반적으로 중회귀모형은 절편항(intercept term)을 포함하고 있으며 이를  $\beta_0$ 로 나타내고 최소제곱추정량은  $b_0$ 로 나타내었다. 만약 독립변수들의 값이 모두 0일 때 종속변수의 값도 0이어야 한다면 절편항은 포함될 필요가 없으며 중회귀모형은

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, \dots, n$$

이다. 계획행렬을

$$X = \begin{pmatrix} x_{11} & \cdots & x_{k1} \\ x_{12} & \cdots & x_{k2} \\ \vdots & & \vdots \\ x_{1n} & & x_{kn} \end{pmatrix}$$

으로 놓으면  $y = X\beta + \epsilon$ 으로 나타내지고 최소제곱추정량은  $\hat{\beta} = (X'X)^{-1}X'y$ 이다.

절편항이 없는 경우의 분산분석은 절편항이 있는 경우와 약간의 차이가 있다. 총변동은 다음과 같이 수정항  $CT = n\bar{y}^2$ 을 빼지 않은 형태이며, 자유도는 자료의 개수  $n$ 이다.

$$SST = \sum_{i=1}^n y_i^2 = y'y \quad .$$

회귀변동은 다음과 같고, 자유도는 독립변수의 개수  $k$ 이다.

$$SSR = \sum_{i=1}^n \hat{y}_i^2 = \hat{y}'\hat{y} = (Xb)'Xb = b'X'Xb = b'X'X(X'X)^{-1}X'y = b'X'y \quad .$$

잔차변동은  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ 이고, 자유도는  $n - k$ 이며 다음과 같다.

$$SSE = SST - SSR = y'y - b'X'y \quad .$$

이를 요약하는 절편이 없는 회귀의 분산분석표는 다음과 같다.

[표 5.2] 절편이 없는 회귀의 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$
회귀	$SSR = b'X'y$	k	$MSR = SSR/k$	$F_0 = MSR/MSE$
잔차	$SSE = y'y - b'X'y$	n-k	$MSE = SSE/(n-k)$	
계	$SST = y'y$	n		

검정통계량  $F_0 = MSR/MSE$ 는 분자의 자유도가  $k$ 이고 분모의 자유도가  $n-k$ 인  $F(k, n-k)$  분포를 사용하여 다음과 같은 가설에 대한 검정을 한다.

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0, H_1 : \text{적어도 하나의 } i \text{에 대하여 } \beta_i \neq 0.$$

만약,  $F_0 > F_{\alpha}(k, n-k)$ 이면  $H_0$ 를 기각하고, 아니면  $H_0$ 를 채택한다.

## 5.9 회귀계수의 표준화

어떤 설명변수에 대한 회귀계수의 추정치가 다른 설명변수에 대한 회귀계수의 추정치보다 크다고 해서 반응변수에 더 크게 영향을 준다고 판단하는 것은 회귀계수의 추정치가 각각의 독립변수의 단위에 의해 좌우된다는 점에서 잘못이므로 독립변수의 단위가 없도록 하는 변환이 의미가 있게 된다. 독립변수의 단위가 없도록 변환된 회귀 모형에서의 회귀계수를 표준화된 회귀계수(standardized regression coefficient) 또는 베타계수(beta-coefficient)라고 한다.

독립변수의 단위가 없도록 변환하는 방법으로는 일단 독립변수와 종속변수에서 각각의 평균을 빼서 편차형태로 만든 다음 이 편차형태의 독립변수와 종속변수를 각 변수의 표본표준편차로 나누는 단위 표준화법(unit normal scaling)을 생각할 수 있으나 그보다는 편차형태의 독립변수와 종속변수를 각 변수의 편차제곱합의 제곱근으로 나누어 변환된 모형에서의 계획행렬의 곱이 상관계수행렬이 되도록 해주는 단위 길이법(unit length scaling)을 알아보자.

중회귀모형을 다음과 같이 나타낼 때

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

독립변수  $x_{ij}$ 와 종속변수  $y_i$ 에서 각각의 평균  $\bar{x}_i, \bar{y}$ 을 빼서  $x_{ij} - \bar{x}_i, y_i - \bar{y}$  형태로 만들고

$$S_{ii} = \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2, \quad S_{yy} = \sum (y_i - \bar{y})^2$$

으로 독립변수와 종속변수를 나눈다. 이렇게 변환된 모형을 다음과 같이 나타내면

$$y_i^* = \beta_1^* Z_{1i} + \beta_2^* Z_{2i} + \cdots + \beta_k^* Z_{ki} + \epsilon_i^*$$

와 같이 절편이 없고, 표준화된 변수들은

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{S_{yy}}}, \quad Z_{1i} = \frac{x_{1i} - \bar{x}_1}{\sqrt{\sum (x_{1i} - \bar{x}_1)^2}}, \quad \cdots, \quad Z_{ki} = \frac{x_{ki} - \bar{x}_k}{\sqrt{\sum (x_{ki} - \bar{x}_k)^2}}$$

이고  $\beta_i^* = \beta_i \sqrt{\frac{S_{ii}}{S_{yy}}}$  인 관계가 있다. 이 모형에서 회귀계수의 추정량은

$$\hat{b}^* = \begin{pmatrix} \hat{b}_1^* \\ \hat{b}_2^* \\ \vdots \\ \hat{b}_k^* \end{pmatrix} = (Z'Z)^{-1} Z'y^*$$

이다. 여기서,

$$\begin{aligned} Z'Z &= \begin{pmatrix} \sum z_{1j}^2 & \sum z_{1j}z_{2j} & \cdots & \sum z_{1j}z_{kj} \\ \sum z_{2j}^2 & \cdots & \sum z_{2j}z_{kj} \\ \text{대칭} & & \ddots & \sum z_{kj}^2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ \vdots & & \ddots & & r_{k1} \\ & & & 1 \end{pmatrix} \end{aligned}$$

는 상관행렬이 된다.

표준화된 변수들에 대한 회귀계수의 추정량  $b^*$ 의 분산공분산행렬은 변환된 오차항의 분산공분산행렬이  $V(\epsilon^*) = \sigma_*^2$  라 할 때 다음과 같다.



$$V(b^*) = (Z'Z)^{-1} \sigma_*^2$$

설명변수가 두 개인 중회귀모형이라면

$$(Z'Z)^{-1} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}^{-1} = \frac{1}{1-r_{12}} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

이므로

$$V(b_1^*) = \frac{\sigma_*^2}{1-r_{12}} = V(b_2^*).$$

## 5.10 예

Faraway(2002)에 소개된 갈라파고스(galapagos) 군도의 거북이 종류 수에 관한 자료에 대한 중회귀분석을 해보자. 자료는 30개 섬들에 대한 7가지 변수들로 되어 있으며 R의 패키지 ‘faraway’에 포함되어 있다. 패키지 목록에 나열되어 있지 않을 경우 ‘R Console’의 ‘패키지들’ 메뉴 중 ‘Install package(s) from local file’에 의해 저장해 놓은 압축파일을 불러들인다. R commander 등을 이용하여 자료를 ‘gala’라는 이름의 데이터셋으로 읽어 들이고 다음과 같이 확인해 본다.

```
> gala
```

	Species Endemics		Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	23	25.09	346	0.6	0.6	1.84
Bartolome	31	21	1.24	109	0.6	26.3	572.33
...			(중략)				
Tortuga	16	8	1.24	186	6.8	50.9	17.95
Wolf	21	12	2.85	253	34.1	254.7	2.33

여기서 변수들이 나타내는 것은 다음과 같다.

Species: 섬에서 발견된 거북이 종류 수

Endemics: 고유종의 수

Elevation: 섬의 최고 높이(highest elevation) (m)

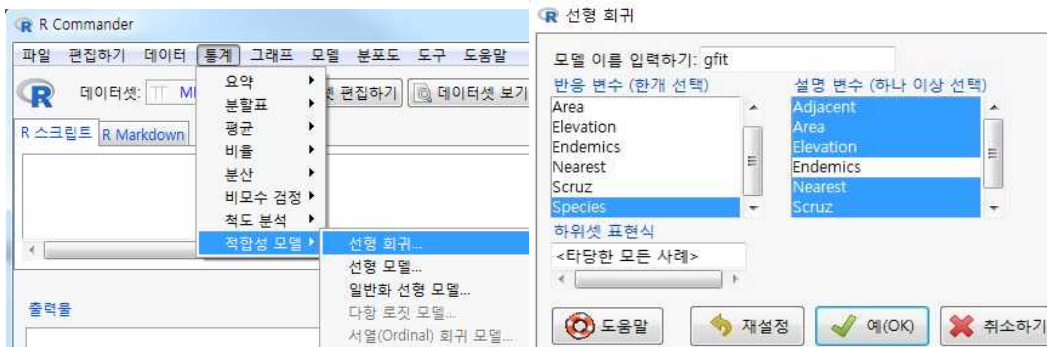
Nearest: 가장 가까운 섬으로부터의 거리 (km)

Scruz: 산타크루즈(Santa Cruz) 섬으로부터의 거리 (km)

Adjacent: 인접한 섬의 넓이 (km )

메뉴 막대에서 ['통계' -> '적합성 모델' -> '선형 회귀...']를 선택하면 '선형 회귀' 대화 창이 나타나며 '반응 변수 (한개 선택)'와 '설명 변수 (하나 이상 선택)'를 택하여 회귀분석을 하고 '출력물' 창에서 결과를 볼 수 있다.

'출력물' 창에 다음과 같이 lm() 명령문을 사용하여 중회귀모형을 적합시킨 결과가 나타난다.



```

> gfit <- lm(Species~Adjacent+Area+Elevation+Nearest+Scruz, data=gala)
> summary(gfit)

Call:
lm(formula = Species ~ Adjacent + Area + Elevation + Nearest +
    Scruz, data = gala)

Residuals:
    Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221   19.154198   0.369  0.715351
Adjacent    -0.074805    0.017700  -4.226  0.000297 ***
Area        -0.023938    0.022422  -1.068  0.296318
Elevation    0.319465    0.053663   5.953  3.82e-06 ***
Nearest      0.009144    1.054136   0.009  0.993151
Scruz       -0.240524    0.215402  -1.117  0.275208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
F-statistic: 15.7 on 5 and 24 DF, p-value: 6.838e-07

```

이 결과에는 여러 가지 유용한 값들이 나와 있고 다른 통계 패키지들과 거의 비슷한 결과를 보여 준다. R의 유용한 점의 하나는 관심이 있는 값을 직접 계산할 수 있다는 것이다. 물론 `lm()` 함수에 의해 계산된 값들만 이용한다면 모르지만 그렇지 않은 값을 계산할 때 필수적인 기능이다. 먼저 `X` 행렬을 다음과 같이 만들어 보자.

```
> x <- cbind(1, gala[, -c(1, 2)])
```

그리고 반응변수 `y`는 다음과 같이 나타내자.

```
> y <- gala$Species
```

여기서 `x`는 데이터프레임 형태이므로 행렬연산을 하려고 하면 오류가 발생한다. 따라서 다음과 같이 `x`를 행렬형태로 바꾼 다음 행렬연산을 한다.

```
> x <- as.matrix(x)
> t(x) %*% x
```

역행렬은 `solve()` 명령문을 사용하여 구할 수 있다.

```
> xtxi <- solve(t(x) %*% x)
> xtxi
```

역행렬  $(X'X)^{-1}$ 을 다른 방법으로 다음과 같이 구할 수 있다.

```
> gs <- summary(gfit)
> gs$cov.unscaled
```

명령문 `names()`를 사용하면 R 객체의 요소들을 볼 수 있다.

```
> names(gs)
> names(gfit)
```

특히, 적합(또는 예측) 값들과 잔차들은 다음과 같이 볼 수 있다.

```
> gfit$fit
```

```
> gfit$res
```

최소제곱추정량  $\hat{\beta}$ 은 다음과 같이 직접 계산할 수 있다.

```
> xtxi %*% t(x) %*% y
      [,1]
[1,] 7.068221
[2,] -0.023938
[3,] 0.319465
[4,] 0.009144
[5,] -0.240524
[6,] -0.074805
```

좀 더 효율적이고 안정적인 계산방식은 다음과 같다.

```
> solve(t(x) %*% x, t(x) %*% y)
      [,1]
[1,] 7.068221
[2,] -0.023938
[3,] 0.319465
[4,] 0.009144
[5,] -0.240524
[6,] -0.074805
```

오차분산의 제곱근  $\sigma$ 는 다음과 같이 추정할 수 있다.

```
> sqrt(sum(gfit$res^2)/(length(y)-6))
[1] 60.97519
```

또한 회귀계수 추정량들의 표준오차를 행렬의 대각요소를 구하는 함수 `diag()`를 이용하여 다음과 같이 구할 수 있다.

```
> sqrt(diag(xtxi))*60.975
      1      Area  Elevation  Nearest      Scruz  Adjacent
19.15413865  0.02242228  0.05366264  1.05413269  0.21540158  0.01770013
```

결정계수는 다음과 같이 구할 수 있다.

```
> 1-sum(gfit$res^2)/sum((y-mean(y))^2)
[1] 0.765847
```

F-통계량과 확률값은 다음과 같이 구할 수 있다.

```
> sst=sum((gala$Species-mean(gala$Species))^2)
> sse=sum(gfit$res^2)
> f0=((sst-sse)/5)/(sse/24)
> pval=1-pf(f0, 5, 24)
> f0
[1] 15.69941
> pval
[1] 6.837893e-07
```

## 제 6 장

## 중회귀모형에 관한 추론

앞 장에서 최소제곱추정량은  $b = (X'X)^{-1}X'y$  이고, 기대값 벡터는  $E(b) = \beta$ , 분산공분산행렬은  $V(b) = (X'X)^{-1}\sigma^2$  이 됨을 보았다.

이 장에서는 최소제곱추정량을 이용하여 중회귀모형에 대한 추론을 하는 방법을 알아 본다.

### 6.1 구간추정

#### 6.1.1 회귀모수( $\beta_i$ )의 구간추정

오차항이 서로 독립이고 정규분포  $N(0, \sigma^2)$ 를 한다면 반응변수  $y_i$ 도 정규분포를 한다. 최소제곱추정량  $b = (X'X)^{-1}X'y$ 는  $y_i$ 들의 선형결합이고, 기대값벡터  $E(b) = \beta$ , 분산공분산행렬  $V(b) = (X'X)^{-1}\sigma^2$  이므로 역시 다음과 같은 정규분포를 한다.

$$b \sim N(\beta, (X'X)^{-1}\sigma^2).$$

여기서  $C = (X'X)^{-1}$  로 놓고,  $i$ 번째 대각요소를  $c_{ii}$ 라 하자. 표준화된 확률변수

$$Z_i = \frac{b_i - \beta_i}{\sqrt{\text{Var}(b_i)}} = \frac{b_i - \beta_i}{\sqrt{c_{ii}\sigma^2}} \sim N(0, 1)$$

이고,

$$P\left[-z_{\alpha/2} \leq \frac{b_i - \beta_i}{\sqrt{c_{ii}\sigma^2}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

인 사실로부터  $\beta_i$ 의  $100(1 - \alpha)\%$  신뢰구간은 다음과 같이 구해진다.

$$b_i - z_{\alpha/2} \sqrt{c_{ii}\sigma^2} \leq \beta_i \leq b_i + z_{\alpha/2} \sqrt{c_{ii}\sigma^2}$$

만약  $\sigma^2$ 을 모르면 불편추정량인 MSE를 대신 사용하여  $\beta_i$ 의  $100(1-\alpha)\%$  신뢰구간은 다음과 같이 구해진다.

$$b_i - t(n-k-1, \frac{\alpha}{2}) \sqrt{c_{ii}\text{MSE}} \leq \beta_i \leq b_i + t(n-k-1, \frac{\alpha}{2}) \sqrt{c_{ii}\text{MSE}}$$

예를 들기 위하여 Faraway(2002)에 소개된 50개국의 경제 자료를 R commander 등을 이용하여 ‘savings’라는 이름의 데이터셋으로 입력하고 다음과 같이 확인한다.

```
> savings
```

	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
... (중략)					
Malaysia	4.71	47.20	0.66	242.69	5.08

여기서 반응변수 ‘sr’은 개인별 저축총액을 가처분 소득으로 나눈 값을 나타낸다. 설명변수 ‘dpi’는 일인당 가처분 소득을 나타내고, ‘ddpi’는 일인당 가처분 소득의 변화율을 나타낸다. 설명변수 ‘pop15’는 15세 미만인 인구비율을 그리고 ‘pop75’는 75세를 넘는 인구비율을 나타낸다. 이 설명변수들을 모두 사용하여 중회귀모형을 적합시킬 때 다음과 같이 ‘.’을 이용하여 편리하게 lm() 함수에 모형식을 지정하고 결과를 구할 수 있다.

```
> sfit <- lm(sr ~ ., savings)
```

```
> summary(sfit)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.5660865	7.3545161	3.884	0.000334 ***
pop15	-0.4611931	0.1446422	-3.189	0.002603 **
pop75	-1.6914977	1.0835989	-1.561	0.125530
dpi	-0.0003369	0.0009311	-0.362	0.719173
ddpi	0.4096949	0.1961971	2.088	0.042471 *

Residual standard error: 3.803 on 45 degrees of freedom  
Multiple R-Squared: 0.3385, Adjusted R-squared: 0.2797  
F-statistic: 5.756 on 4 and 45 DF, p-value: 0.0007904

이 결과에서 회귀계수추정량의 표준오차(Std. Error)는  $\sqrt{c_{ii}\text{MSE}}$  을 뜻하므로 만약 pop75에 대한 회귀모수의 95% 신뢰구간을 구한다면 다음과 같이 계산할 수 있다.

```
> ss <- summary(sfit)
> c(sfit$coef[3]-qt(0.975, length(savings$sr)-5)*ss$coef[3, 2],
+   sfit$coef[3]+qt(0.975, length(savings$sr)-5)*ss$coef[3, 2])
      pop75      pop75
-3.8739780  0.4909826
```

같은 방법으로 ddpi에 대한 회귀모수의 95% 신뢰구간을 구하면 다음과 같다.

```
> c(sfit$coef[5]-qt(0.975, length(savings$sr)-5)*ss$coef[5, 2],
+   sfit$coef[5]+qt(0.975, length(savings$sr)-5)*ss$coef[5, 2])
      ddpi      ddpi
0.01453363  0.80485623
```

이 신뢰구간은 상한이 하한의 50배가 넘는 정도로 넓다. 이는 일인당 가처분 소득의 변화율이 저축에 끼치는 영향을 정확하게 알기 어렵다는 것을 뜻한다.

신뢰구간은 양측검정과 대등한 점이 있다. 신뢰수준 95%인 신뢰구간은 유의수준 5%에서 기각되지 않는 모든 귀무가설을 포함한다. 따라서 ‘pop75’에 대한 신뢰구간은 유의수준 5%에서 기각되지 않는 귀무가설  $H_0 : \beta_{pop75} = 0$ 을 나타내는 0을 포함한다.

### 6.1.2 조건부기대값 $E(y|x_0)$ 의 구간추정

주어진 설명변수  $x_0' = (1, x_{10}, x_{20}, \dots, x_{k0})$ 에서 반응변수의 조건부 기대값  $E(y|x_0)$ 의 불편추정량을 구하면  $x_0'b$ 이다. 즉

$$E(x_0'b) = x_0'E(b) = x_0'\beta = (1 \ x_{10} \ \dots \ x_{k0}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = E(y|x_0)$$

$$= \beta_0 + \beta_1 x_{10} + \beta_2 x_{20} + \dots + \beta_k x_{k0} = x_0'\beta.$$



추정량  $x'b$ 의 분산은 다음과 같이 구할 수 있다.

$$V(x'b) = x'V(b)x = x'(X'X)^{-1}\sigma^2x = x'(X'X)^{-1}x\sigma^2.$$

추정량  $x'b = x'(X'X)^{-1}x'y$ 는  $y$ 의 선형결합이므로 다음과 같은 정규분포를 한다.

$$x'b \sim N(x'\beta, x'(X'X)^{-1}x\sigma^2).$$

표준화된 확률변수  $Z_0 = \frac{x'b - x'\beta}{\sqrt{x'(X'X)^{-1}x\sigma^2}} \sim N(0,1)$ 에 대해 다음이 만족된다.

$$P\left[-z_{\alpha/2} \leq \frac{x'b - x'\beta}{\sqrt{x'(X'X)^{-1}x\sigma^2}} \leq z_{\alpha/2}\right] = 1 - \alpha.$$

분산  $\sigma^2$ 을 알 때,  $E(y|x_0) = x_0'\beta$ 의  $100(1-\alpha)\%$  신뢰구간은  $\hat{y}_0 = x_0'b$ 이라 놓으면

$$\hat{y}_0 - z_{\alpha/2} \sqrt{x_0'(X'X)^{-1}x_0\sigma^2} \leq x_0'\beta \leq \hat{y}_0 + z_{\alpha/2} \sqrt{x_0'(X'X)^{-1}x_0\sigma^2}$$

분산  $\sigma^2$ 을 모를 때,

$$\hat{y}_0 - t_{\frac{\alpha}{2}}(n-k-1) \sqrt{x_0'(X'X)^{-1}x_0 \text{MSE}} \leq x_0'\beta \leq \hat{y}_0 + t_{\frac{\alpha}{2}}(n-k-1) \sqrt{x_0'(X'X)^{-1}x_0 \text{MSE}}$$

예를 들기 위하여 다시 갈라파고스 자료를 사용하여 회귀모형을 다음과 같이 적합시킨다.

```
> gfit <- lm(Species ~ Area+ Elevation+ Nearest+ Scruz+ Adjacent, data=gala)
```

어떤 섬에 대해 설명변수들의 값이 0.08, 93, 6.0, 12.0, 0.34일 때 거북이 종류의 수를 예측해보자. 실제로 같은 조건의 새로운 섬이 존재할 것 같지는 않아 연습으로 그치더라도 해볼만하다.

공식을 사용하여 다음과 같이 직접 예측값을 구할 수 있다.

```
> x0 <- c(1, 0.08, 93, 6.0, 12.0, 0.34)
> y0 <- sum(x0*gfit$coef)
```

```
> y0
[1] 33.91967
```

따라서 주어진 설명변수들 값에서 평균 종류 수는 34로 올림하여 추정할 수 있다. 이제 주어진 설명변수들 값에서 평균 종류 수에 대한 95% 신뢰구간을 구하는 방법을 알아보자.

먼저  $(X'X)^{-1}$ 를 계산할 필요가 있다.

```
> x <- cbind(1,gala[,3:7])
> x <- as.matrix(x)
> xtxi <- solve(t(x) %*% x)
```

평균반응값에 대한 신뢰구간의 폭을 계산하면 다음과 같다.

```
> bm <- sqrt(x0 %*% xtxi %*% x0) *qt(0.975,24)* 60.98
> bm
      [,1]
[1,] 32.88844
```

따라서 신뢰구간은 다음과 같이 구할 수 있다.

```
> c(y0-bm, y0+bm)
[1] 1.031231 66.808105
```

함수 predict()를 이용하면 다음과 같이 구간추정에 필요한 값들을 구할 수 있다.

```
> gm <- predict(gfit, data.frame(Area=0.08, Elevation=93, Nearest=6.0,
Scruz=12,
+ Adjacent=0.34), se=T)
> gm
$fit
[1] 33.91967

$se.fit
[1] 15.93385
```

이 결과를 이용하여 주어진 설명변수 값들에서 평균반응값에 대한 신뢰구간을 다음과 같이 구한다.

```
> c( gm$fit - gm$se.fit * qt(0.975, 24), gm$fit + gm$se.fit * qt(0.975, 24))
[1] 1.033826 66.805510
```

### 6.1.3 $q'\beta$ 의 구간추정

일반적인  $\beta_j$ 의 선형결합, 즉

$$q_0\beta_0 + q_1\beta_1 + q_2\beta_2 + \cdots + q_p\beta_p = (q_0, \cdots, q_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = q'\beta$$

의 구간추정에 대하여 알아보자.

우선  $q'\beta$ 의 불편추정량은  $q'b$ ,  $b = (X'X)^{-1}X'y$ 임을 다음과 같이 보일 수 있다.

$$E(q'b) = E(q'(X'X)^{-1}X'y) = q'(X'X)^{-1}X'E(y) = q'(X'X)^{-1}X'X\beta = q'\beta$$

또는

$$E(q'b) = q'E(b) = q'\beta.$$

불편추정량  $q'b$ 의 분산은 다음과 같다.

$$V(q'b) = q'V(b)q = q'(X'X)^{-1}\sigma^2q = q'(X'X)^{-1}q\sigma^2$$

따라서,  $q'b = q'(X'X)^{-1}X'y$ 로  $y$ 의 선형결합이므로 다음과 같은 정규분포를 한다.

$$q'b \sim N(q'\beta, q'(X'X)^{-1}q\sigma^2)$$

표준화된 확률변수  $Z_0 = \frac{q'b - q'\beta}{\sqrt{q'(X'X)^{-1}q \cdot \sigma^2}} \sim N(0,1)$ 에 대해 다음이 만족된다.

$$P\left[-z_{\alpha/2} \leq \frac{q'b - q'\beta}{\sqrt{q'(X'X)^{-1}q} \cdot \sigma} \leq z_{\alpha/2}\right] = 1 - \alpha$$

따라서  $q'b$ 의  $100(1-\alpha)\%$  신뢰구간은  $\sigma^2$ 을 알 때

$$q'b - z_{\alpha/2} \sqrt{q'(X'X)^{-1}q} \sigma \leq q'\beta \leq q'b + z_{\alpha/2} \sqrt{q'(X'X)^{-1}q} \sigma$$

이고,  $\sigma^2$ 을 모를 때

$$q'b - t_{\frac{\alpha}{2}}(n-k-1) \sqrt{q'(X'X)^{-1}q \text{MSE}} \leq q'\beta \leq q'b + t_{\frac{\alpha}{2}}(n-k-1) \sqrt{q'(X'X)^{-1}q \text{MSE}}$$

특히,  $q = (1, 0, \dots, 0)$  인 경우,  $q'\beta = (1, 0, \dots, 0) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \beta_0$ 이다.

따라서  $\beta_0$ 의  $100(1-\alpha)\%$  신뢰구간은  $\sigma^2$ 을 알 때

$$b_0 - z_{\alpha/2} \sqrt{(1, 0, \dots, 0)(X'X)^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \sigma^2} \leq \beta_0 \leq b_0 + z_{\alpha/2} \sqrt{(1, 0, \dots, 0)(X'X)^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \sigma^2}$$

즉,  $c_{00}$ 가  $(X'X)^{-1}$ 의 첫번째 원소를 나타내는 경우 다음과 같다.

$$b_0 - z_{\alpha/2} \sqrt{c_{00}\sigma^2} \leq \beta_0 \leq b_0 + z_{\alpha/2} \sqrt{c_{00}\sigma^2}.$$

#### 6.1.4 새로운 관찰치의 예측구간

새로운 관찰치( $y_0$ )의 예측구간을 구하기 위하여 예측치를  $\hat{y}_{0n} = x_0'b$ 로 놓으면

$$V(\hat{y}_{0n} - y_0) = V(\hat{y}_{0n}) + V(y_0) = x_0'(X'X)^{-1}x_0\sigma^2 + \sigma^2 = (1 + x_0'(X'X)^{-1}x_0)\sigma^2$$

따라서  $y_0$ 의  $100(1-\alpha)\%$  예측구간은  $\sigma^2$ 을 알 때

$$x_0' b - z_{\alpha/2} \sqrt{(1 + x_0' (X'X)^{-1} x_0) \sigma^2} \leq y_0 \leq x_0' b + z_{\alpha/2} \sqrt{(1 + x_0' (X'X)^{-1} x_0) \sigma^2}$$

이고  $\sigma^2$ 을 모를 때

$$x_0' b - t_{\frac{\alpha}{2}}(n-k-1) \sqrt{(1 + x_0' (X'X)^{-1} x_0) \text{MSE}} \leq y_0 \leq x_0' b + t_{\frac{\alpha}{2}}(n-k-1) \sqrt{(1 + x_0' (X'X)^{-1} x_0) \text{MSE}}$$

예로 앞에서 다룬 갈라파고스 자료에서 주어진 설명변수들의 값에서 거북이 종류 수에 대한 예측구간을 다음과 같이 구할 수 있다.

```
> bp <- sqrt(1+x0 %*% xtxi %*% x0) * qt(0.975, 24) * 60.98
> c(y0-bp, y0+bp)
[1] -96.16946 164.00879
```

또는 함수 predict()에서 interval="prediction"를 지정하여 예측구간을 구할 수 있다.

```
> gp <- predict(gfit, data.frame(Area=0.08, Elevation=93, Nearest=6.0,
+ Scrutz=12, Adjacent=0.34), interval="prediction", se=T)
> gp
$fit
      fit      lwr      upr
[1,] 33.91967 -96.1528 163.9921
```

거북이 종류 수는 음수가 될 수 없으므로 이러한 예측구간은 수정이 필요하다. 변수변환이나 포아송분포를 이용하는 방법을 고려할 수 있다.

## 6.2 검정

### 6.2.1 회귀계수에 관한 검정

회귀계수  $\beta_i$ 가 0인가 아닌가를 검정하기 위하여 다음과 같이 가설을 세운다.

$$H_0 : \beta_i = 0 \text{ v.s. } H_1 : \beta_i \neq 0$$

오차항의 분포가  $\epsilon_j \sim N(0, \sigma^2)$  이라면  $b \sim N(\beta, (X'X)^{-1}\sigma^2)$ 임을 알고 있다.

따라서  $b_i \sim N(\beta_i, c_{ii}\sigma^2)$ 이고  $c_{ii}$ 가  $(X'X)^{-1}$ 의  $i$ 번째 대각요소를 나타낼 때

$$\frac{b_i - \beta_i}{\sqrt{c_{ii}\sigma^2}} \sim N(0, 1)$$

이다. 만약  $\sigma^2$ 을 모르면 불편추정량 MSE를 사용하여  $t$ -분포를 이용한다.

$$\frac{b_i - \beta_i}{\sqrt{c_{ii}\text{MSE}}} \sim t(n - k - 1)$$

따라서 가설  $H_0 : \beta_i = 0 \text{ v.s. } H_1 : \beta_i \neq 0$ 의 검정 통계량은 다음과 같다.

1)  $\sigma^2$ 을 알 때

$$Z_0 = \frac{b_i}{\sqrt{c_{ii} \cdot \sigma^2}}$$

2)  $\sigma^2$ 을 모를 때

$$t_0 = \frac{b_i}{\sqrt{c_{ii} \cdot \text{MSE}}}$$

### 6.2.2 평균반응량 $E(y|x_0)$ 의 가설검정

$E(y_0) = x_0'\beta$ 의 불편 추정량은  $x'b$ 이며,  $x'b \sim N(x'\beta, x'(X'X)^{-1}x\sigma^2)$ 이다.

따라서  $H_0 : E(y_0) = \eta$ ,  $H_1 : E(y_0) \neq \eta$ 의 검정 통계량은

$$Z_0 = \frac{\hat{y}_0 - \eta}{\sqrt{x'(X'X)^{-1}x\sigma^2}} \quad (\sigma^2 : \text{알 때})$$

$$t_0 = \frac{\hat{y}_0 - \eta}{\sqrt{x'(X'X)^{-1}x\text{MSE}}} \quad (\sigma^2 : \text{모를 때})$$

이고 기각역과 판정기준은 다음과 같다.

$|Z_0| > z_{\alpha/2}$  이면  $H_0$  기각하고,  $|Z_0| \leq z_{\alpha/2}$  이면  $H_0$  채택.

$|t_0| > t(n-k-1, \frac{\alpha}{2})$  이면  $H_0$  기각하고,  $|t_0| \leq t(n-k-1, \frac{\alpha}{2})$  이면  $H_0$  채택.

### 6.2.3 회귀계수 $\beta_j$ 의 선형함수 $q'\beta$ 에 관한 검정

선형함수  $q'\beta$ 의 불편추정량은  $q'b$ 이고,  $q'b \sim N(q'\beta, q'(X'X)^{-1}q\sigma^2)$ 이다.

따라서

$$H_0 : q'\beta = C, H_1 : q'\beta \neq C$$

의 검정 통계량은

$$Z_0 = \frac{q'b - q'\beta}{\sqrt{q'(X'X)^{-1}q\sigma^2}} \quad (\sigma^2 : \text{알 때})$$

$$t_0 = \frac{q'b - q'\beta}{\sqrt{q'(X'X)^{-1}q\text{MSE}}} \quad (\sigma^2 : \text{모를 때})$$

이고 기각역과 판정기준은 다음과 같다.

$|Z_0| > z_{\alpha/2}$  이면  $H_0$  기각하고,  $|Z_0| \leq z_{\alpha/2}$  이면  $H_0$  채택.

$|t_0| > t(n-k-1, \frac{\alpha}{2})$  이면  $H_0$  기각하고,  $|t_0| \leq t(n-k-1, \frac{\alpha}{2})$  이면  $H_0$  채택.

## 6.3 부분 F-검정과 축차 F-검정

중회귀분석에서 어떤 특정한 변수들을 회귀모형에 포함시킬 것인지 판단하는 방법을 알아보자.

### 6.3.1 추가제곱합

어떤 변수 하나를 추가하는 경우  $SST = SSR + SSE$ 에서 총변동  $SST$ 는 일정하기

때문에 회귀제곱합  $SSR$ 이 커지면 잔차제곱합  $SSE$ 는 작아진다. 따라서 어떤 변수를 추가할 때  $SSR$ 의 증가량이 크면 이 변수를 추가하고, 증가량이 크지 않으면 추가하지 않는다. 이와 같이 변수를 추가할 때 증가하는 회귀제곱합의 양 즉  $SSR$ 의 증가량을 추가제곱합(extra sum of squares)이라고 부른다.

적합할 중회귀모형이 다음과 같다 하자.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

모수들에 대한 최소제곱추정량들이  $b_0, b_1, \dots, b_k, \hat{\sigma}^2 (=MSE)$ 라 할 때, 이 모형에 대한 회귀제곱합을  $SS(b_1, b_2, \dots, b_k | b_0)$ 로 나타내자. 즉

$$SS(b_1, b_2, \dots, b_k | b_0) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b' X' y - n(\bar{y})^2, df = k$$

여기서  $n(\bar{y})^2$ 은 수정항이라 부르며  $y = \beta_0 + \epsilon$ 과 같은 회귀모형에 대한 회귀제곱합에 해당한다. 이 회귀모형에서  $\beta_0$ 의 최소제곱추정량인  $\bar{y}$ 를  $b_0^*$ 라고 나타낼 때, 회귀제곱합은  $SS(b_0^*) = n(\bar{y})^2$ 로 표현한다. 따라서,  $SS(b_1, b_2, \dots, b_k | b_0)$ 는 설명변수  $x_1, x_2, \dots, x_k$ 를 추가할 때 증가하는 회귀제곱합의 양을 뜻하며 다음과 같이 나타낼 수 있다.

$$SS(b_1, b_2, \dots, b_k | b_0) = SS(b_0, b_1, b_2, \dots, b_k) - SS(b_0^*).$$

중회귀모형의  $k$ 개 변수 중에서  $q$ 개만 선택하고 편의상  $x_1, x_2, \dots, x_q$ 라 하자. 나머지  $k - q$ 개 변수를 중회귀모형에서 제외할 것인지 검토하는 방법을 알아보자. 선택한  $q$ 개만의 변수를 가진 모형을

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_q x_q + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

라 하고, 모수들에 대한 최소제곱추정량들을  $a_0, a_1, \dots, a_q$ 라 할 때 회귀제곱합은

$$SS(a_1, a_2, \dots, a_q | a_0) = a' X' y - n(\bar{y})^2, df = q$$

이다. 나머지  $k - q$ 개 변수를 모형에 추가함으로써 증가하는 회귀제곱합인 추가제곱합은

$$SS(b_1, b_2, \dots, b_k | b_0) - SS(a_1, a_2, \dots, a_q | a_0)$$



이고 편의상

$$SS(b_{q+1}, b_{q+2}, \dots, b_k | b_0, b_1, \dots, b_q)$$

로 나타낸다. 이러한 추가제곱합을 나타내는 또 다른 기호로 자주 사용되는

$$R(\beta_{q+1}, \beta_{q+2}, \dots, \beta_k | \beta_0, \beta_1, \dots, \beta_q)$$

와 같은 표현은 모수  $\beta_0, \beta_1, \dots, \beta_k$ 가 있는 모형에 대한  $SSR$ 에서  $\beta_0, \beta_1, \dots, \beta_q$ 만 있는 모형의  $SSR$ 을 뺀 것을 나타내며, 모수  $\beta_0, \beta_1, \dots, \beta_q$ 만 있는 모형에 대한  $SSE$ 에서  $\beta_0, \beta_1, \dots, \beta_k$ 가 있는 모형의  $SSE$ 를 빼도 같은 값이다.

이러한 추가제곱합은 모형에 있는  $k$ 개의 변수 중에서  $x_1, x_2, \dots, x_q$ 인  $q$ 개만 선택하고 나머지  $k - q$ 개 변수를 모두 중회귀모형에서 제외할 것인지를 나타내는 귀무가설

$$H_0 : \beta_{q+1} = \dots = \beta_k = 0$$

에 대한 다음과 같은 검정통계량에 사용된다.

$$F_0 = \frac{R(\beta_{q+1}, \beta_{q+2}, \dots, \beta_k | \beta_0, \beta_1, \dots, \beta_q) / (k - q)}{MSE}.$$

이 검정통계량의 값이  $F_{\alpha}(k - q, n - k - 1)$ 보다 작으면 귀무가설을 기각할 수 없어 변수  $x_{q+1}, \dots, x_k$ 를 모형에 포함시킬 필요가 없다고 결론을 내린다.

만약  $q=0$ 이면  $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ 이고 분산분석표의  $F$ -검정에 대한 귀무가설과 같게 되어 모형의 유의성을 검정하게 된다.

예를 들어 앞에서 다룬 ‘savings’ 자료에서 ‘p15’가 유의하지 않다는 귀무가설  $H_0 : \beta_1 = 0$ 을 검정하기 위한 추가제곱합과  $F$ -검정통계량을 계산하여 보자.

일단 ‘p15’를 포함한 모형의 잔차제곱합은 다음과 같은 방법으로 구하고 자유도는 45임을 안다.

```
> gf <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> ssef <- sum(gf$res^2)
```

```
> ssef
[1] 650.713
```

‘p15’를 뺀 모형의 잔차제곱합은 다음과 같은 방법으로 구하고 자유도는 46임을 안다.

```
> gr <- lm(sr ~ pop75 + dpi + ddp, data=savings)
> sser <- sum(gr$res^2)
> sser
[1] 797.7249
```

따라서 추가제곱합  $R(\beta_1|\beta_0, \beta_2, \beta_3, \beta_4)$ 과 F-검정통계량은 다음과 같이 구한다.

```
> R1_0234 <- sser-ssef
> f0 <- (R1_0234/(46-45))/(ssef/45)
> f0
[1] 10.16659
```

이 때 확률값은 다음과 같다.

```
> pval <- 1-pf(f0, 1, 45)
> pval
[1] 0.002603019
```

위와 같은 계산을 anova() 함수를 이용하여 다음과 같이 편리하게 할 수 있다.

```
> anova(gr, gf)
Analysis of Variance Table
Model 1: sr ~ pop75 + dpi + ddp
Model 2: sr ~ pop15 + pop75 + dpi + ddp
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     46  797.72
2     45  650.71  1    147.01 10.167 0.002603 **
```

### 6.3.2 완전모형과 축소모형

모든 설명변수  $x_1, x_2, \dots, x_k$ 를 포함하는, 즉  $(k+1)$ 개 회귀모수를 갖는 식

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i$$

을 완전모형(full model)이라 하고, 이 모형에 대한  $SSR$ 과  $SSE$ 를 구하고 괄호 속에 “ $F$ ”를 넣어  $SSR(F)$ ,  $SSE(F)$ 로 표시한다.

완전모형에 포함된  $k$ 개 보다 작은 수인  $q$ 개의 설명변수, 즉  $(q+1)$ 개 회귀모수만 포함하는 모형을 내포된(nested) 축소모형(reduced model)이라 하고, 이 모형에 대한  $SSR$ 과  $SSE$ 를 구하고 괄호 속에 “ $R$ ”를 넣어  $SSR(R)$ ,  $SSE(R)$ 로 표시한다.

완전모형과 축소모형에서의 회귀제곱합의 차이  $SSR(F) - SSR(R)$  또는 동등하게 잔차제곱의 차이  $SSE(R) - SSE(F)$ 가 크지 않으면 통계학의 기본 원리중 하나인 모수절약(parsimonious)의 원칙에 따라 축소모형을 선택하여 적합하기로 한다.

다음과 같이 분자에는 추가로 얻어진  $SSR$  즉 추가제곱합을 그 자유도로 나눈 값을 놓고, 분모에는 모든 변수를 포함했을 때, 즉 완전모형의  $MSE$ 를 놓아 검정통계량

$$F_0 = \frac{[SSR(F) - SSR(R)] / [df(F) - df(R)]}{MSE(F)}, \quad df(F) = k, \quad df(R) = q$$

을 만들면 축소모형이 옳은 경우  $F(k - q, n - k - 1)$ 인 분포를 한다.

예를 들어 앞에서 다룬 ‘savings’ 자료에서 청소년층과 노년층의 저축율이 유의한 차이가 없다는 귀무가설  $H_0 : \beta_1 = \beta_2$ 을 검정하기 위한 F-검정을 해보자. 이 경우 완전모형을

$$y = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 dpi + \beta_4 ddpi + \epsilon$$

로 나타내면, 축소모형은 다음과 같이 나타낼 수 있다.

$$y = \beta_0 + \beta_1 (pop15 + pop75) + \beta_3 dpi + \beta_4 ddpi + \epsilon$$

편리하게 검정하기 위하여 완전모형과 그에 내포된 축소모형에 대한 부분 F-검정을 할 수 있는 `anova()` 함수를 다음과 같이 이용한다.

```
> gf <- lm(sr ~ ., savings)
> gr <- lm(sr ~ I(pop15+ pop75)+ dpi+ ddpi, savings)
> anova(gr, gf)
Analysis of Variance Table
Model 1: sr ~ I(pop15 + pop75) + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     46 673.63
2     45 650.71  1      22.91 1.5847 0.2146
```

완전모형식에서 구두점 “.”은 데이터프레임에 있는 다른 모든 변수들을 간편하게 나타내는 기호이다. 축소모형식에서 함수 `I()`는 괄호 안의 요소가 각각 모형에 포함되도록 해석되는 것이 아니고 계산되는 것임을 지시한다. 확률값이 0.21이어서 귀무가설을 기각할 수 없고 이는 청소년층과 노년층을 구별하여 각각 다를 필요는 없다는 의미이다.

또 다른 예로 일인당 가처분 소득의 변화율을 나타내는 ‘ddpi’에 대한 회귀계수가 1이라는 귀무가설  $H_0 : \beta_4 = 1$ 을 검정하기 위한 F-검정을 해보자. 이 경우 축소모형은 다음과 같이 나타낼 수 있다.

$$y = \beta_0 + \beta_1 \text{pop15} + \beta_2 \text{pop75} + \beta_3 \text{dpi} + \text{ddpi} + \epsilon$$

여기서 ‘ddpi’ 항에는 회귀모수가 없음을 알 수 있다. 회귀식에서 이러한 항을 `offset`이라 부르며 다음과 같이 축소모형을 적합시킨 다음 `anova()` 함수를 사용하여 완전모형과 비교한다.

```
> gr <- lm(sr ~ pop15+ pop75+ dpi+ offset(ddpi), savings)
> anova(gr, gf)
Analysis of Variance Table
Model 1: sr ~ pop15 + pop75 + dpi + offset(ddpi)
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     46 781.61
2     45 650.71  1      130.90 9.0525 0.004286 **
```

확률값이 작아 귀무가설을 기각할 수 있다. 이러한 검정은 다음과 같이 t 통계량을 이용하여 간단하게 할 수도 있다.

```
> b4 <- 1; gs <- summary(gf)
> tstat <- (gf$coef[5]-b4)/gs$coef[5,2]
> tstat
-3.008735
> 2*pt(tstat,45)
0.004286094
```

확률값이 앞의 F-검정에서와 같다는 것을 알 수 있고, t 통계량 값을 제공하면 F-값과 같음을 확인 할 수 있다.

```
> tstat^2
9.052483
```

완전모형에 있는  $k$ 개의 변수 중에서  $x_1, x_2, \dots, x_q$ 인  $q$ 개만 선택하고 나머지  $k-q$ 개 변수를 모두 중회귀모형에서 제외하는 귀무가설

$$H_0 : \beta_{q+1} = \dots = \beta_k = 0$$

을 적용한 축소모형에 대한 검정 방법은  $F_0 > F_\alpha(k-q, n-k-1)$  이면  $H_0$ 를 기각하고 아니면  $H_0$ 를 채택하는 것이다.

완전모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_q x_{qi} + \beta_{q+1} x_{q+1i} + \dots + \beta_k x_{ki} + \epsilon_i$$

에 대하여

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \beta_1 = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}, \beta_2 = \begin{pmatrix} \beta_{q+1} \\ \beta_{q+2} \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$X_1 = \begin{pmatrix} 1 & x_{11} & \cdots & x_{q1} \\ 1 & x_{12} & \cdots & x_{q2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{qn} \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_{q+11} & \cdots & x_{k1} \\ x_{q+12} & \cdots & x_{k2} \\ \vdots & & \vdots \\ x_{q+1n} & \cdots & x_{kn} \end{pmatrix}$$

으로 놓으면  $y = X_1\beta_1 + X_2\beta_2 + \epsilon$ 이다. 여기서

$$X_1\beta_1 + X_2\beta_2 = (X_1 \ X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = X\beta$$

이고, 최소제곱추정량은

$$\begin{aligned} \hat{\beta} = b &= (X'X)^{-1}X'y = \left( \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} (X_1 \ X_2) \right)^{-1} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} y \\ &= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \end{aligned}$$

이다. 만약  $X_1'X_2 = 0$  또는  $X_2'X_1 = 0$ 과 같이 직교하는 경우

$$\begin{aligned} \hat{\beta} &= \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} \\ &= \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix} = \begin{pmatrix} (X_1'X_1)^{-1}X_1'y \\ (X_2'X_2)^{-1}X_2'y \end{pmatrix} \end{aligned}$$

즉

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y, \quad \hat{\beta}_2 = (X_2'X_2)^{-1}X_2'y$$

가 된다. 마찬가지로 다음과 같은 축소모형

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_r x_{ri} + \epsilon_i$$

에 대하여

$$X_1 = \begin{pmatrix} 1 & x_{11} & \cdots & x_{q1} \\ 1 & x_{12} & \cdots & x_{q2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{qn} \end{pmatrix}$$

으로 놓고  $y = X_1\alpha + \epsilon$ 로 나타내면, 최소제곱추정량은

$$\hat{\alpha} = (X_1' X_1)^{-1} X_1' y$$

이다.

따라서 축소모형에서의 회귀계수에 대한 최소제곱추정량과 완전모형에서의 같은 회귀계수에 대한 최소제곱추정량은  $X_1' X_2 = 0$ ,  $X_2' X_1 = 0$  일 경우는 동일하지만 일반적으로는 그렇지 않다.

추가제곱합을 나타내는 기호  $R()$ 에 대해 행렬로 나타내면 다음과 같다.

$$\begin{aligned} R(\beta) &= R(\beta_0, \beta_1, \dots, \beta_k) = y' P_x y \\ &= y' X (X' X)^{-1} X' y = y' (X_1' X_2) \begin{pmatrix} X_1' X_1 & X_1' X_2 \\ X_2' X_1 & X_2' X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1' y \\ X_2' y \end{pmatrix} \end{aligned}$$

이때  $X_1' X_2 = 0$  이거나  $X_2' X_1 = 0$  이면

$$\begin{aligned} R(\beta) &= (y' X_1 \ y' X_2) \begin{pmatrix} (X_1' X_1)^{-1} & 0 \\ 0 & (X_2' X_2)^{-1} \end{pmatrix} \begin{pmatrix} X_1' y \\ X_2' y \end{pmatrix} \\ &= y' X_1 (X_1' X_1)^{-1} X_1' y + y' X_2 (X_2' X_2)^{-1} X_2' y \\ &= y' P_{x_1} y + y' P_{x_2} y \\ &= R(\beta_1) + R(\beta_2) \end{aligned}$$

따라서

$$R(\beta_1 | \beta_2) = R(\beta_1, \beta_2) - R(\beta_2) = R(\beta_1) + R(\beta_2) - R(\beta_2) = R(\beta_1)$$

즉  $X_1' X_2 = 0$  이거나  $X_2' X_1 = 0$ 와 같이 직교하는 경우

$$\begin{aligned} R(\beta) &= R(\beta_1) + R(\beta_2) \\ R(\beta_1 | \beta_2) &= R(\beta_1) \\ R(\beta_2 | \beta_1) &= R(\beta_2) \end{aligned}$$

분할을 여러 개로 했을 때도  $X_j' X_k = 0$ ,  $j \neq k$  와 같이 직교하는 경우 위의 성질은 성립한다.

완전모형에서의 회귀제곱합  $SSR(F)$ 와 축소모형에서의 회귀제곱합  $SSR(R)$ 을 행렬을 사용하여 나타내면

$$\begin{aligned} SSR(F) &= y' (P_x - \frac{J}{n}) y = y' P_x y - y' \frac{J}{n} y = y' X (X' X)^{-1} X' y - y' \frac{J}{n} y \\ SSR(R) &= y' (P_{x_1} - \frac{J}{n}) y = y' X_1 (X_1' X_1)^{-1} X_1' y - y' \frac{J}{n} y \end{aligned}$$

따라서

$$\begin{aligned} SSR(F) - SSR(R) &= y'X(X'X)^{-1}X'y - y'\frac{J}{n}y - (y'X_1(X_1'X_1)^{-1}X_1' - y'\frac{J}{n}y) \\ &= y'X'(X'X)^{-1}X'y - y'X_1(X_1'X_1)^{-1}X_1'y = y'(P_x - P_{x_1})y \end{aligned}$$

특히  $X_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$  즉  $q=0$ 인 경우,  $P_{x_1} = X_1(X_1'X_1)^{-1}X_1' = \frac{J}{n}$ 인 수정항의 행렬이다.

[표 6.1]  $H_0: \beta_2 = 0$ 를 검정하기 위한 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$
회귀	$SSR(F) = y'(P_x - \frac{J}{n})y$	k	$MSR(F)$	$F_0 = \frac{MSR}{MSE}$
$\beta_1$ 만의 회귀	$SSR(R) = y'(P_{x_1} - \frac{J}{n})y$	k-q	$MSR(R)$	
$\beta_2$ 추가	$Q = SSR(F) - SSR(R)$ $= y'(P_x - P_{x_1})y$	q	$Q/q$	$F_0 = \frac{Q/q}{MSE}$
오차	$SSE(F) = y'(I - P_x)y$	n-k-1	$MSE(F)$	
계	$SST = y'(I - \frac{J}{n})y$	n-1		

### 6.3.3 부분 F-검정

추가제곱합에서  $x_j$ 를 제외한 모든 설명변수들은 회귀모형에 포함되어 있을 때,  $x_j$ 도 포함시킬 것인가 검정하는 방법을 알아보자. 설명변수  $x_j$ 를 포함할 경우 추가제곱합은

$$R(\beta_i | \beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k)$$

으로 나타내고 자유도 1이다. 검정하려는 가설은

$$H_0: \beta_i = 0, H_1: \beta_i \neq 0$$

이고, 검정통계량은 다음과 같다.



$$F_0 = \frac{R(\beta_i | \beta_0, \beta_1, \dots, \beta_{i-1}, \beta_{i+1}, \dots, \beta_k)}{\text{MSE}}$$

이 검정통계량의 값이  $F_\alpha(1, n-k-1)$ 보다 크면 귀무가설을 기각하고 변수  $x_i$ 를 모형에 포함시키는 방법을 부분 F-검정(partial F-test)이라 부른다.

이러한 부분 F-검정은 중회귀모형에 포함된 설명변수 개수가 너무 많다고 생각될 때 중요하지 않다고 판단되는 설명변수를 제거하는 방법으로 활용된다.

즉 모든 설명변수들 각각에 대하여 부분 F-검정통계량을 계산하여 가장 작은  $F_0$  값에 대응하는 설명변수가 유의하면 제거할 설명변수는 없다고 판단하여 종료하고, 유의하지 않으면 제거하고 나머지 설명변수들만 포함한 새로운 중회귀모형에서 마찬가지로 부분 F-검정을 실시하는 절차를 수행하며 더 이상 제거할 필요가 있다고 판단되는 설명변수가 없을 때 종료한다.

### 6.3.4 축차 F-검정

회귀모형에 포함을 고려하는 대상인 설명변수가  $k$  개 있다고 하자. 중요하다고 판단되는 설명변수를 하나씩 모형에 추가하여 포함시키는 방법을 알아보자. 절편항만 있는 회귀모형에 설명변수  $x_i$ 를 포함할 경우 추가제곱합은  $R(\beta_i | \beta_0)$ 으로 나타내고 자유도 1이다. 이렇게 계산된  $k$  개의 추가제곱합 중 최대인 것에 대응하는 설명변수  $x_{\max}$ 에 대한 부분 F-검정통계량은 절편항과  $x_{\max}$ 를 포함한 회귀모형에서의 잔차제곱평균을  $MSE_{\max}$ 라 할 때 다음과 같다.

$$F_0 = \frac{R(\beta_{\max} | \beta_0)}{MSE_{\max}}$$

이 검정통계량의 값이  $F_\alpha(1, n-2)$ 보다 작으면 귀무가설을 기각할 수 없으므로 설명변수  $x_{\max}$ 가 유의하지 않다고 판단하여 종료하고, 크면 귀무가설을 기각하고 설명변수  $x_{\max}$ 를 모형에 포함시키는 부분 F-검정(partial F-test)을 실시한다. 이렇게 추가된 설명변수  $x_{\max}$ 를  $x_1$ 이라 놓을 때, 절편항과  $x_1$ 만 있는 회귀모형에 나머지 설명변수들 중 하나인  $x_i$ 를 포함할 경우 추가제곱합은  $R(\beta_i | \beta_0, \beta_1)$ 으로 나타내고 자유도 1이다. 이렇게 계산된  $k-1$  개의 추가제곱합 중 최대인 것에 대응하는 설명변수  $x_{\max}$ 에 대한 부분 F-검정통계량은 절편항과  $x_1, x_{\max}$ 를 포함한 회귀모형에서의 잔

차제곱평균을  $MSE_{\max}$  라 할 때 다음과 같다.

$$F_0 = \frac{R(\beta_{\max} | \beta_0, \beta_1)}{MSE_{\max}}$$

이 검정통계량의 값이  $F_{\alpha}(1, n-3)$ 보다 작으면 귀무가설을 기각할 수 없으므로 설명변수  $x_{\max}$ 가 유의하지 않다고 판단하여 종료하고, 크면 귀무가설을 기각하고 설명변수  $x_{\max}$ 를 모형에 포함시키는 부분 F-검정(partial F-test)을 실시한다. 이렇게 추가된 설명변수  $x_{\max}$ 를  $x_2$ 라 놓고 앞에서와 같은 절차를 수행하여 더 이상 추가할 설명변수가 없을 때 종료한다.

이와 같이 축차적(sequentially)으로 부분 F-검정을 수행하여 유의한 설명변수를 중요하다고 판단하여 회귀모형에 차례로 포함시키는 방법을 축차 F-검정(sequential F-test)이라고 부른다.

축차 F-검정에서는 차례로 수행되는 부분 F-검정의 통계량에서 완전모형이 차례로 바뀌기 때문에 잔차제곱평균  $MSE_{\max}$ 도 차례로 값이 달라진다는 것을 알 수 있다.

## 6.4 중회귀모형의 타당성

### 6.4.1 잔차의 검토

중회귀모형의 적합값  $\hat{y}$ 와 실제 자료  $y$ 와의 차이  $e = \hat{y} - y$ 를 잔차(residual)라고 부르며, 잔차를 검토하여 여러 가지 유용한 정보를 얻을 수 있다. 잔차를 행렬을 사용하여 나타내면 다음과 같다.

$$e = y - \hat{y} = y - Xb = y - X(X'X)^{-1}X'y = (I - X(X'X)^{-1}X')y = (I - P_x)y.$$

여기서,  $X(X'X)^{-1}X'$ 는 hat matrix라 부르며  $H$ 로 나타내기도 하고, 사영행렬(projection matrix)이라 하여  $P_x$ 로 나타내기도 한다. 잔차의 기대값과 분산은 다음과 같다.

$$E(e) = (I - P_x)E(y) = (I - P_x)X\beta = 0,$$

$$V(e) = (I - P_x)V(y)(I - P_x)' = (I - P_x)I(I - P_x)' \sigma^2 = (I - P_x)\sigma^2.$$

잔차의 성질로는 다음과 같은 것들이 있다.

i) 잔차의 합은 “0” 이다.

절편항이 있는 회귀모형에서는  $1'P_x = 1$ 이므로

$$\sum_{i=1}^n e_i = 1'e = 1'(I - P_x)y = (1' - 1'P_x)y = (1' - 1')y = 0$$

ii) 정규성 가정하에서  $\hat{y}_i$ 와 잔차  $e_i$ 는 서로 독립이다.

$$\text{Cov}(\hat{y}, e) = \text{Cov}(P_x y, (I - P_x)y) = P_x V(y) (I - P_x) = P_x (I - P_x)I \sigma^2 = 0$$

정규성 가정하에서는 공분산이 0이면 독립이다.

iii) 잔차  $e_j$ 의  $x_{ij}$ 에 의한 가중합은 “0” 이다. 즉

$$\sum_{j=1}^n x_{ij}e_j = 0, \quad i = 1, 2, \dots, k$$

이다.  $x_i' = (x_{i1}, x_{i2}, \dots, x_{in})$ ,  $e' = (e_1, e_2, \dots, e_n)$ 으로 놓으면

$$\sum_{j=1}^n x_{ij}e_j = x_i'e = x_i'(I - P_x)y = (x_i' - x_i'P_x)y = 0$$

이다. 따라서

$$X'e = X'(I - P_x)y = (X' - X'P_x)y = (X' - X')y = 0.$$

iv) 잔차들의  $\hat{y}_j$ 에 의한 가중합은 “0” 이다. 즉

$$\sum_{j=1}^n \hat{y}_j e_j = 0$$

이다.  $\hat{y} = P_x y$ 이고  $P_x P_x = P_x$ 이므로

$$\hat{y}'e = (P_x y)'(I - P_x)y = y'P_x'(I - P_x)y = y'P_x(I - P_x)y = 0.$$

### 6.4.2 적합결여검정 (lack of fit test)

변수들 사이의 관계를 나타내기 위해 중회귀모형이 적합한지 검정할 필요가 있다. 단 순선회귀분석에서와 마찬가지로 주어진 독립변수  $(x_1, x_2, \dots, x_k)$  에서  $y$ 의 반복 측정이 있는 경우에만 검정할 수 있다. 자료가 다음과 같이 주어졌다고 하자.

$$\begin{array}{l} (x_{11} \ x_{21} \ \cdots \ x_{k1}) \text{ 에서 } y_{11}, y_{12}, \dots, y_{1n_1} \\ (x_{12} \ x_{22} \ \cdots \ x_{k2}) \text{ 에서 } y_{21}, y_{22}, \dots, y_{2n_2} \\ \vdots \\ (x_{1p} \ x_{2p} \ \cdots \ x_{kp}) \text{ 에서 } y_{p1}, y_{p2}, \dots, y_{pn_p} \end{array}$$

적합할 중회귀모형은 다음과 같이 나타내자.

$$y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_{ij}, \quad i = 1, \dots, p, \quad j = 1, \dots, n_p$$

적합된 회귀모형에서의 잔차  $e_{ij}$  를 다음과 같이 분해하자.

$$e_{ij} = y_{ij} - \hat{y}_i = y_{ij} - \bar{y}_i + \bar{y}_i - \hat{y}_i$$

잔차제곱합은 다음과 같이 순오차제곱합 SSPE와 적합결여제곱합 SSLF로 분해할 수 있다.

$$\begin{aligned} SSE &= \sum_i \sum_j e_{ij}^2 = \sum_i \sum_j (y_{ij} - \hat{y}_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i + \bar{y}_i - \hat{y}_i)^2 \\ &= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j (\bar{y}_i - \hat{y}_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \hat{y}_i)^2. \end{aligned}$$

우측 첫 항인 순오차제곱합  $SSPE = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ 의 자유도는  $\sum_i (n_i - 1) = n - p$

이고, 순오차제곱평균  $MSPE = \frac{SSPE}{n - p}$ 는 모형에 관계없이 오차분산  $\sigma^2$ 의 불편 추정

량이 된다. 두 번째 항인 적합결여제곱합  $SSLF = \sum_i n_i (\bar{y}_i - \hat{y}_i)^2$ 의 자유도는 SSE의 자

유도에서 SSPE의 자유도를 빼서 구하면 다음과 같다.

$$df(SSLF) = df(SSE) - df(SSPE) = (n - k - 1) - (n - p) = p - k - 1.$$

중회귀모형이 적합한지 검정하기 위한 통계량  $F_L$ 의 값이  $F_{\alpha}(p-k-1, n-p)$ 보다 크면 귀무가설을 기각하고 중회귀모형이 적합하지 않다고 판정한다.

[표 6.2] 적합결여검정을 위한 분산분석표

요인	제곱합	자유도	제곱평균	$F_0$
회귀	SSR	k	MSR	
잔차	$SSE \begin{cases} SSLF \\ SSPE \end{cases}$	$\begin{matrix} p-k-1 \\ n-p \end{matrix}$	$\begin{matrix} MSLF \\ MSPE \end{matrix}$	$F_L = \frac{MSLF}{MSPE}$
계	SST	n-1		

## 7.1 다중공선성(multicollinearity)

설명변수들 사이에 선형관계가 존재할 때 공선성이 존재한다고 한다.  $k$  개 설명변수를 포함하는 중회귀모형의 행렬표현에서  $X$  행렬의  $k+1$  개 열벡터 사이에 공선성이 존재하면 일차독립이 아니고 일차종속인 관계가 존재한다는 뜻이고 적어도 하나의 변수가 다른 변수들의 선형결합이 되므로 독립변수로서의 역할을 할 수 없고 유일한 최소제곱추정량이 존재하지 않게 된다.

완전한 공선성은 아니더라도 공선성에 아주 가까운 관계가 존재할 때 다중공선성이 있다고 한다. 다중공선성은 회귀계수의 추정치가 불안정하여 의미 없게 만들고 추정 오차를 아주 크게 만들어 통계적 추론의 신뢰성을 약화시킬 뿐만 아니라 해석에도 문제가 있게 만든다.

다중공선성 문제를 해결하는 방법으로는 필요한 변수만 선택하는 변수선택법, 주성분회귀분석(principal regression), 능형회귀분석(ridge regression) 등이 있다.

다중공선성의 존재가 의심되는 몇 가지 현상은 다음과 같다.

- 1) 독립변수를 추가 또는 제거할 때 추정된 회귀계수가 크게 변한다.
- 2) 자료를 추가 또는 제거할 때 추정된 회귀계수가 크게 변한다.
- 3) 회귀계수의 부호가 이론과 다르다.
- 4) 회귀계수의 신뢰구간이 너무 넓다.
- 5) 중요하다고 생각되는 변수가 유의하지 않다.
- 6) 독립변수들 사이의 상관계수가  $\pm 1$ 에 가까운 것이 있다.

다중공선성의 존재를 판정하는 기준에는 분산확대인자(VIF, Variance Inflation Factor), 상태지수(condition index), 분산비율(variance proportion) 등이 있다.

### 7.1.1 분산확대인자(VIF, Variance Inflation Factor)

독립변수  $x_j$ 를 종속변수로 나머지 독립변수를 설명변수로 포함한 회귀모형에서의

결정계수를  $R_i^2$ 이라 할 때,  $x_i$ 에 대한 VIF는 다음과 같이 정의한다.

$$(\text{VIF})_i = (1 - R_i^2)^{-1}$$

다중공선성이 존재한다고 판정하는 기준은 다음과 같다.

1)  $\text{VIF}_i$ 들 중에서 가장 큰 것이 10을 넘는다. 즉  $\max R_i^2 > 0.9$  이다.

2)  $\overline{\text{VIF}} = \frac{1}{k} \sum_{i=1}^k (\text{VIF})_i$ 가 1 보다 훨씬 크다.

분산확대인자의 정의에서  $(1 - R_i^2)$ 을 독립변수  $x_i$ 의 허용치(tolerance)라고 부른다. 따라서  $k$  개의 허용치 중 가장 작은 것이 0.1 보다 작으면 다중공선성이 존재한다고 판정한다.

### 7.1.2 상태지수(condition index)

중회귀모형에서  $X'X$ 에 대한 고유치를  $\lambda_i, i=1, \dots, k$ 라 하고, 그 중 최대치를  $\lambda_{\max}$ 라 할 때  $i$  번째 상태지수는 다음과 같이 정의한다.

$$I_i = \sqrt{\lambda_{\max} / \lambda_i}$$

이러한 상태지수들 중 최대값이 10을 넘으면 다중공선성이 있다고 보고, 이 최대값이 30( 또는 100)이 넘으면 심각한 다중공선성이 있다고 본다.

### 7.1.3 분산비율(variance proportion)

중회귀모형에서  $X'X$ 에 대한  $k$  개의 고유치와 연관된  $k$  개의 고유벡터를 생각하자.  $k$  개 독립변수 각각에 대하여 회귀계수에 대한 추정량의 분산 중 각 고유벡터에 의해 설명되는 비율을 나타내는 측도가 분산비율이다.

최소의 고유치, 즉 최대의 상태지수를 갖는 고유벡터에 의해 설명되는 분산비율이 90% 이상인 독립변수가 두 개 이상일 때 다중공선성이 있다고 본다.

### 7.1.4 예

다중공선성이 있는 자료의 예로 Longley 데이터셋을 입력하여 분석하자.

```
> gl <- lm(Employed ~ ., longley)
```

```
> summary(gl)
```

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-3.482e+03	8.904e+02	-3.911	0.003560 **
GNP.deflator	1.506e-02	8.492e-02	0.177	0.863141
GNP	-3.582e-02	3.349e-02	-1.070	0.312681
Unemployed	-2.020e-02	4.884e-03	-4.136	0.002535 **
Armed.Forces	-1.033e-02	2.143e-03	-4.822	0.000944 ***
Population	-5.110e-02	2.261e-01	-0.226	0.826212
Year	1.829e+00	4.555e-01	4.016	0.003037 **

```
Residual standard error: 0.3049 on 9 degrees of freedom
```

```
Multiple R-Squared: 0.9955, Adjusted R-squared: 0.9925
```

```
F-statistic: 330.3 on 6 and 9 DF, p-value: 4.984e-10
```

반응변수는 고용된 인원수이었다. 확률값이 큰 세 변수도 모두 반응변수에 영향을 줄 것으로 예상되었던 것들이었는데 유의하지 않게 나타났다. 이유를 알기 위하여 우선 상관행렬을 구해보자.

```
> round(cor(longley[, -7]), 3)
```

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
GNP.deflator	1.000	0.992	0.621	0.465	0.979	0.991
GNP	0.992	1.000	0.604	0.446	0.991	0.995
Unemployed	0.621	0.604	1.000	-0.177	0.687	0.668
Armed.Forces	0.465	0.446	-0.177	1.000	0.364	0.417
Population	0.979	0.991	0.687	0.364	1.000	0.994
Year	0.991	0.995	0.668	0.417	0.994	1.000

몇 개의 상관계수가 크게 나타났다.  $X'X$ 의 고유치를 구해보자.

```
> x <- as.matrix(longley[, -7])
```

```
> e <- eigen(t(x) %*% x)
```

```
> e$val
```

```
[1] 6.665299e+07 2.090730e+05 1.053550e+05 1.803976e+04 2.455730e+01
2.015117e+00
```

```
> sqrt(e$val[1]/e$val)
```

```
[1] 1.00000 17.85504 25.15256 60.78472 1647.47771 5751.21560
```



고유치의 범위가 넓고 몇 개의 상태지수가 크다. 이는 하나만이 아니라 그보다 많은 선형결합이 문제의 원인이 되고 있음을 뜻한다. 이제 분산확대인자를 검토하자. 첫 번째 변수에 대하여 다음과 같이 구할 수 있다.

```
> summary(lm(x[, 1] ~ x[, -1]))$r.squared
[1] 0.9926217
> 1/(1-0.9926217)
[1] 135.5326
```

일차독립인 설명변수라면 VIF가 1이어야 하는데 너무 크게 나왔다. 모든 VIF를 한꺼번에 구하려면 vif()함수를 사용한다.

```
> vif(gl)
GNP.deflator      GNP  Unemployed Armed.Forces  Population      Year
135.53244      1788.51348      33.61889      3.58893    399.15102    758.98060
```

명백히 분산 확대가 일어났음을 알 수 있다. 예를 들어 GNP의 표준오차가 공선성이 없는 경우에 비하여  $\sqrt{178} \approx 42$ 배가량 커진 것을 볼 수 있다. 이 문제를 해결하는 한 방법은 변수 중 몇 개를 제외하는 것이다. 앞에서 구한 상관행렬을 보면 세 번째 변수와 네 번째 변수는 다른 변수들과 상관이 아주 강하지는 않기 때문에 제외하지 않기로 하고 다른 변수들만 제거 대상으로 삼는다.

```
> cor(x[, -c(3, 4)])
      GNP.deflator      GNP Population      Year
GNP.deflator  1.0000000 0.9915892 0.9791634 0.9911492
GNP           0.9915892 1.0000000 0.9910901 0.9952735
Population    0.9791634 0.9910901 1.0000000 0.9939528
Year          0.9911492 0.9952735 0.9939528 1.0000000
```

이 네 변수들은 서로 아주 강한 상관이 있어 어떤 것을 하나 택해도 다른 것들을 나타내는 역할을 할 수 있다. 임의로 Year를 택했다고 하자.

```
> summary(lm(Employed ~ Armed.Forces + Unemployed + Year, longley))
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.797e+03  6.864e+01 -26.183 5.89e-12 ***
```

```

Armed.Forces -7.723e-03  1.837e-03  -4.204  0.00122 **
Unemployed  -1.470e-02  1.671e-03  -8.793  1.41e-06 ***
Year          9.564e-01  3.553e-02  26.921  4.24e-12 ***

```

Residual standard error: 0.3321 on 12 degrees of freedom

Multiple R-Squared: 0.9928, Adjusted R-squared: 0.9911

F-statistic: 555.2 on 3 and 12 DF, p-value: 3.916e-13

여섯 개의 설명변수들 대신 세 개의 설명변수만 사용하여 적합했지만 결정계수 값에 별 차이가 없다는 것을 알 수 있다.

## 7.2 주성분회귀와 능형회귀

다중공선성이 있는 경우 문제를 야기하는 설명변수를 모형에서 제거하는 방법 외에 회귀모수를 추정하는 방법을 달리 하여 대처할 수 있다. 여기서는 주성분회귀와 능형회귀에 대해서 알아본다.

### 7.2.1 주성분회귀(Principal Component Regression)

변수들 사이에 공선성이 있는 경우 적절한 변환을 통하여 직교하도록 만드는 방법을 생각할 수 있다.

우선  $X'X$ 의 고유치를  $\lambda_1, \dots, \lambda_k$ 라 할 때, 다음과 같은 성질을 만족하는  $k \times k$  행렬  $P$ 를 생각하자.

$$Z = XP \text{ 이고 } Z'Z = \text{diag}(\lambda_1, \dots, \lambda_k), \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq 0.$$

여기서  $Z'Z = P'X'XP$  이므로  $P$ 의 열벡터가  $X'X$ 의 고유벡터이다.  $Z$ 의 열벡터는 주성분이라 불리며 서로 직교한다. 다중공선성이 존재한다면 고유치들 중에서 0에 가까운 것들이 있게 된다.

행렬  $X'X$ 의 고유치들 중에서 0에 가까운 것들을 제외한 나머지 고유치들의 대각행렬을  $\Lambda_1$ 이라 하고 대응하는 고유벡터 행렬을  $P_1$ 이라 하여 얻은 추정량  $b_k = P_1 \Lambda_1^{-1} P_1' X'y$ 를 사용하여 분석하는 방법을 주성분회귀라 한다.

주성분회귀 추정량은 통상적인 최소제곱 추정량보다 다중공선성이 심한 경우 우수하다고 알려져 있다. 그러나 주성분은 설명변수들의 선형결합으로 되어 있기 때문에 해석하기 곤란한 경우가 많다는 단점도 있다.

예로 들었던 ‘Longley’ 데이터셋을 분석해보자. 먼저  $X'X$ 의 고유치를 다음과 같이 구한다.

```
> data(longley)
> x <- as.matrix(longley[, -7])
> e <- eigen(t(x) %*% x)
> e$values
[1] 6.665299e+07 2.090730e+05 1.053550e+05 1.803976e+04 2.455730e+01
2.015117e+00
```

상대적으로 첫 번째 고유치가 크다. 대응하는 첫 번째 고유벡터를 생각해보자.

```
> dimnames(e$vectors) <- list(c("GNP def", "GNP", "Unem", "Armed", "Popn",
+ "Year"), paste("고유치", 1:6))
> round(e$vec, 3)
```

	고유치 1	고유치 2	고유치 3	고유치 4	고유치 5	고유치 6
GNP def	-0.050	0.070	-0.034	0.043	0.957	-0.273
GNP	-0.191	0.725	-0.343	0.554	-0.075	0.087
Unem	-0.157	0.622	0.564	-0.521	-0.007	0.011
Armed	-0.128	0.104	-0.746	-0.645	-0.012	0.000
Popn	-0.058	0.038	-0.011	0.036	-0.281	-0.956
Year	-0.957	-0.266	0.078	0.057	-0.015	0.053

첫 번째 고유벡터가 ‘year’에 크게 의존함을 알 수 있다. 다음과 같이 행렬  $X$ 를 검토해보면 변수들의 단위가 다르다는 것을 알 수 있다.

```
> x
```

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year
1947	83.0	234.289	235.6	159.0	107.608	1947
...(중략)						
1962	116.9	554.894	400.7	282.7	130.081	1962

단위를 통일하기 위하여 변수를 표준화할 수 있을 것이다. 이 때 주성분회귀는 상

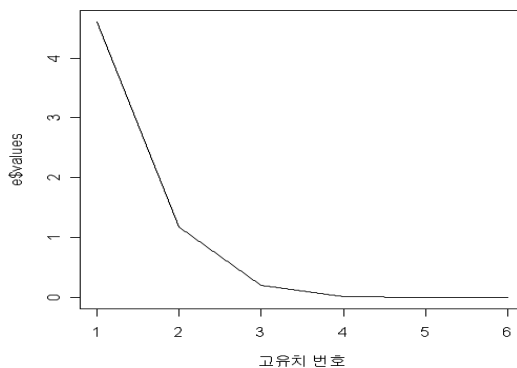
관계수 행렬에 대하여 하는 것과 같게 된다.

```
> e <- eigen(cor(x))
> e$values
[1] 4.6033770958 1.1753404993 0.2034253724 0.0149282587 0.0025520658
0.0003767081
>
> dimnames(e$vectors) <- list(c("GNP def", "GNP", "Unem", "Armed", "Popn",
+ "Year"), paste("고유치", 1:6))
> round(e$vec, 3)
```

	고유치 1	고유치 2	고유치 3	고유치 4	고유치 5	고유치 6
GNP def	0.462	0.058	0.149	0.793	-0.338	0.135
GNP	0.462	0.053	0.278	-0.122	0.150	-0.818
Unem	0.321	-0.596	-0.728	0.008	-0.009	-0.107
Armed	0.202	0.798	-0.562	-0.077	-0.024	-0.018
Popn	0.462	-0.046	0.196	-0.590	-0.549	0.312
Year	0.465	0.001	0.128	-0.052	0.750	0.450

주성분을 몇 개 사용할 것인가 판단하는 방법으로 많이 이용되는 스크리 그림 (scree plot)을 다음과 같이 그릴 수 있다.

```
> plot(e$values, type="l", xlab="고유치 번호")
```

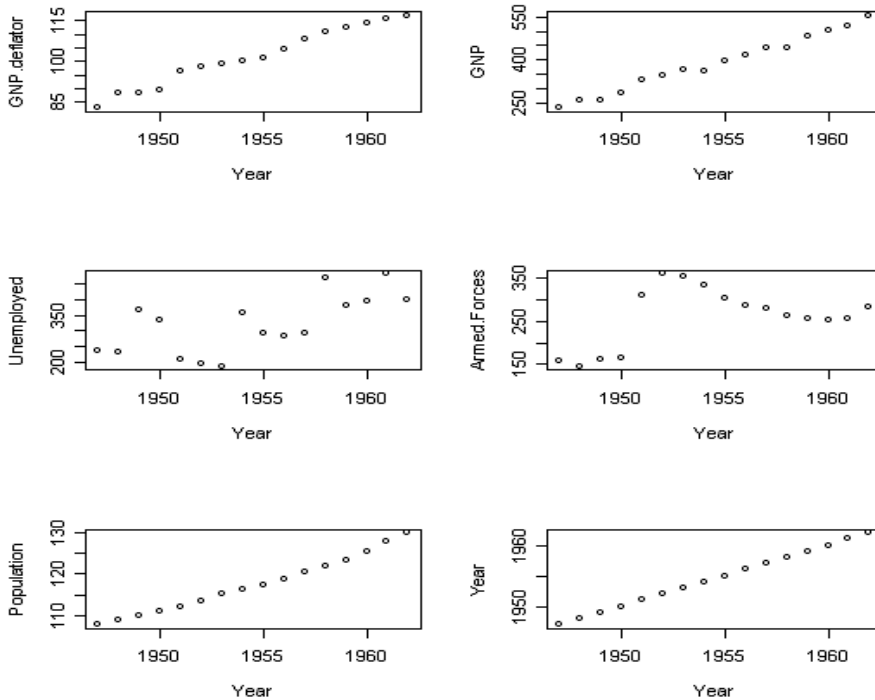


[그림 7.1] 주성분에 대한 스크리 그림

종종 스크리 그림은 눈에 띄게 꺾이는 점을 보이는데 그 점 이후의 고유치는 이전 고유치보다 매우 작아 무시할 수 있다고 판단한다. 위 그림에서는 3번에서 꺾인다고 보아 2개의 주성분만 생각해도 된다고 판단한다.

주성분회귀는 해석을 할 수 있는 경우에만 쓸모가 있다. 첫 번째 주성분은 표준화된 변수들 모두의 선형결합으로 보인다. 변수들 각각을 Year에 대해 그려 보자.

```
> for(i in 1:6) plot(longley[, 6], longley[, i], xlab="Year", ylab=names(longley)[i])
> par(mfrow=c(3, 2))
```



[그림 7.2] Longley 자료

이 그림으로부터 첫 번째 주성분은 연도(Year)에 따른 추세를 나타내고 두 번째 주성분은 Unemployed와 Armed.Forces의 차이를 나타낸다고 짐작한다. 따라서 이 두 성분을 이용하여 회귀분석을 하여 다음과 같은 결과를 얻는다.

```
> summary(lm(Employed ~ Year + I(Unemployed-Armed.Forces), longley))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.391e+03	7.889e+01	-17.638	1.84e-10 ***
Year	7.454e-01	4.037e-02	18.463	1.04e-10 ***
I(Unemployed - Armed.Forces)	-4.119e-03	1.525e-03	-2.701	0.0182 *

Residual standard error: 0.7178 on 13 degrees of freedom  
 Multiple R-Squared: 0.9638, Adjusted R-squared: 0.9582  
 F-statistic: 173 on 2 and 13 DF, p-value: 4.285e-10

이 예에서 사용된 방식이 전형적인 주성분회귀 사용법이다. 원 변수들의 선형결합으로 새 변수들을 만드는데 합당한 근거가 필요하다. 합리적인 근거가 있으면 간단하고 설명하기 좋은 모형을 얻을 수 있지만 항상 그럴 수 있는 것은 아니다.

## 7.2.2 능형회귀(Ridge Regression)

다중공선성이 문제가 되는 것은  $(X'X)^{-1}$ 항이 불안정하기 때문이다. 앞의 주성분회귀에서는 회귀계수에 대한 추정량이  $b_k = P_1 A_1^{-1} P_1' X' y$ 와 같이 유도되었다. 즉 회귀계수에 대한 통상적인 최소제곱추정량  $b = (X'X)^{-1} X' y$ 에서  $(X'X)^{-1}$  대신  $P_1 A_1^{-1} P_1'$ 를 사용하여 다중공선성 문제를 해결하려 한 것이다. 같은 맥락에서 Hoerl과 Kennard(1970)는 능형(ridge)상수가  $c$ 인 능형회귀추정량이라 불리는 다음과 같은 회귀계수에 대한 추정량을 제안하였다.

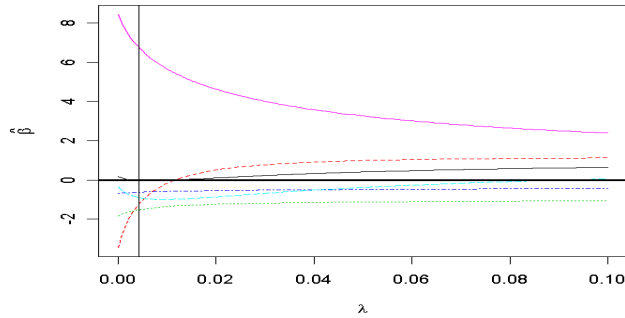
$$b(c) = (X'X + cI)^{-1} X' y, \quad 0 < c < 1$$

이 능형회귀추정량을 보면  $(X'X)^{-1}$  대신  $X'X$ 에 능형상수인  $c$ 배의 단위행렬을 더하여 역행렬을 구하는 방식이므로 안정된 계산을 할 수 있도록 하는 능형상수  $c$  값을 선택하는 방법이 관심의 대상이다.

능형상수  $c$  값을 선택하는 방법은 여러 가지 자동 결정법이 있지만 흔히 사용하는 방법은 능형회귀추정량  $b(c)$ 를 능형상수  $c$ 의 함수로 나타낸 그림인 능형추적도(ridge trace plot)에서  $b(c)$ 가 안정되는  $c$ 의 값들 중 최소값을 택하는 방법이다.

설명을 위하여 Longley 자료를 살펴보자. 능형회귀를 위한 함수 `lm.ridge()`는 MASS 패키지 안에 들어 있으며 능형상수를 `lambda()`로 지정한다.

```
> library(MASS)
> data(longley)
> gr <- lm.ridge(Employed ~ ., longley, lambda = seq(0, 0.1, 0.001))
> matplot(gr$lambda, t(gr$coef), type="l", xlab=expression(lambda),
+ ylab=expression(hat(beta)))
```



[그림 7.3] 능형추적도

위 능형추적도에서 수직선은 Hoerl-Kennard의 능형상수 값을 나타내고 가장 위의 곡선은 Year에 대한 계수를 나타낸다.

능형상수의 값을 자동으로 정하는 방법은 다음과 같다.

```
> select(gr)
modified HKB estimator is 0.004275357
modified L-W estimator is 0.03229531
smallest value of GCV at 0.003
```

능형회귀의 창안자인 Hoerl-Kennard의 능형상수 값이 그림에 나타나 있지만 좀 더 큰 능형상수 값인 0.03을 사용하고자 한다. 이 경우 능형회귀추정량은 다음과 같다.

```
> gr$coef[, gr$lam == 0.03]
GNP.deflator    GNP      Unemployed  Armed.Forces  Population    Year
0.2200496    0.7693585   -1.1894102   -0.5223393   -0.6861816   4.0064269
```

능형상수 값이 0인 경우는 통상적인 최소제곱추정량이며 다음과 같다.

```
> gr$coef[, 1]
GNP.deflator    GNP      Unemployed  Armed.Forces  Population    Year
0.1573796   -3.4471925   -1.8278860   -0.6962102   -0.3441972   8.4319716
```

통상적인 최소제곱추정량에서는 GNP가 고용인원수에 주는 영향이 상식과는 반대 방향으로 나오는데 비하여 능형추정량에서는 예상한 방향으로 영향을 주는 것으로 나온다.

회귀계수에 대한 능형추정량은 불편성을 갖지 못한다. 편의성은 바람직하지 않지만 다음과 같은 평균제곱합(mean squared error: MSE)기준에서는 다르게 생각할 수 있다.

$$\begin{aligned} \text{MSE}(b(c)) &= E[(b(c) - \beta)'(b(c) - \beta)] \\ &= E[(b(c) - E[b(c)])'(b(c) - E[b(c)])] + (E[b(c)] - \beta)'(E[b(c)] - \beta) . \end{aligned}$$

따라서 평균제곱합은 분산에 편의의 제곱을 더한 것으로 나타낼 수 있다. 적절한 능형상수  $c$ 의 값에 대해  $\text{MSE}(b(c))$ 가  $\text{MSE}(b)$ 보다 작은 경우가 있다.

능형추정량  $b(c)$ 는 편의된 추정량이며 최소잔차제곱합 기준으로는 최소제곱추정량  $b$ 보다 못하지만 다중공선성 문제를 해결하며 MSE 기준으로 우수하므로 종종 사용된다.

## 7.3 이상치 및 영향치(influential cases)

회귀분석에서 자료들(cases) 중에 다른 자료와 동떨어진 것이 있는 경우가 있다. 이와 같이 다른 자료와 동떨어진 자료를 이상치(outlier)라 부른다. 이상치들은 때때로 회귀선 추정에 영향을 주기 때문에 자료에 포함시킬지 아닐지를 결정해야 한다. 파악된 이상치가 착오에 의한 것이라면 분석에서 제외하면 되지만, 그렇지 않은 경우 오히려 중요한 정보를 줄 수 있기 때문에 제외하지 않고 이용할 수 있도록 하는 방법을 검토해야 한다.

자료를 독립변수  $x$ 의 관점에서 보았을 때 이상치인 경우가 있고 반응변수  $y$ 의 관점에서 보았을 때 이상치인 경우가 있으며 두 관점 모두에서 이상치로 보이는 경우가 있다. 회귀분석에서는 보통 독립변수를 제어할 수 있기 때문에  $x$ 의 관점에서 보았을 때 이상치 보다는  $y$ 의 관점에서 보았을 때 이상치가 주된 고려 대상이다.

### 7.3.1 독립변수 $x$ 의 관점에서 보았을 때 이상치의 식별 기준

회귀분석에서 헤트행렬(hat matrix)라 부르는  $H = X(X'X)^{-1}X'$ 의 대각 원소를 레버리지(leverage)라 한다.  $i$ 번째 대각원소를  $h_{ii}$ 로 나타내고,  $i$ 번째 관찰점의  $x$ 값과  $n$ 개의 모든 관찰점에서  $x$ 값들의 평균과의 거리에 대한 측도로 사용할 수 있다. 레버리지의 성질은 다음과 같다.



$$i) \sum_{i=1}^n h_{ii} = \text{tr}(X(X'X)^{-1}X') = \text{tr}((X'X)^{-1}X'X) = \text{tr}(I_{k+1}) = k+1$$

$$ii) 0 \leq h_{ii} \leq 1$$

따라서  $h_{ii}$ 의 평균은  $\bar{h} = \frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{k+1}{n}$  이고

$$h_{ii} > \frac{2(k+1)}{n}$$

이면  $x$ 의 관점에서 보았을 때 이상점이라 판별한다.

### 7.3.2 반응변수 $y$ 의 관점에서 보았을 때 이상치 식별 기준

잔차를 표준화하여 그 크기가 클 경우  $y_i$ 가 이상점이라고 판단하는 기준으로 다음과 같은 것들이 있다.

i) 스튜던트화 잔차(studentized residual)

잔차  $e_i$ 를 표준화하면  $\frac{e_i}{\sqrt{\text{Var}(e_i)}}$  이고, 잔차벡터를  $e$ 라 할 때

$$e = (I - P_x)y = (I - H)y, \quad H = X(X'X)^{-1}X'$$

이므로

$$V(e) = (I - P_x)V(y)(I - P_x)' \sigma^2 = (I - P_x)(I - P_x)' \sigma^2 = (I - P_x)\sigma^2 = (I - H)\sigma^2.$$

따라서  $\text{Var}(e_i) = (1 - h_{ii})\sigma^2$ 이고,  $\sigma^2$ 을 모르면 MSE로 추정하여 표준오차

$$\sqrt{\widehat{\text{Var}(e_i)}} = \sqrt{(1 - h_{ii})\text{MSE}}$$

를 구하고 스튜던트화 잔차

$$e_i^* = \frac{e_i}{\sqrt{(1 - h_{ii})\text{MSE}}}$$

를 정의한다. 만약  $|e_i^*|$ 가 상당히 크면 이상점이라 판단한다.

ii) 스튜던트화 제외 잔차 (studentized deleted residual)

회귀모형을  $i$ 번째 자료를 제외한  $(n-1)$ 개의 자료로 적합시키고, 적합된 값을  $\hat{y}_{-i}$ , 잔차를  $d_i$ 라 하자. 이 때  $d_i$ 는 제외잔차라 하고

$$\frac{d_i}{\sqrt{\text{Var}(d_i)}}$$

를 표준화 제외잔차라 한다. 제외잔차  $d_i$ 의 분산은 다음과 같다.

$$V(d_i) = V(y_i - \hat{y}_{-i}) = V(y_i) + V(\hat{y}_{-i}) = \sigma^2 + x_i'(X'(i)X(i))^{-1}x_i\sigma^2.$$

이때  $\sigma^2$  대신  $i$ 번째 관찰점을 제외시킨 MSE인  $MSE(i)$ 로 추정하여 스튜던트화 제외잔차

$$d_i^* = \frac{y_i - \hat{y}_{-i}}{\sqrt{MSE(i)(1 + x_i'(X'(i)X(i))^{-1}x_i)}} = \frac{e_i}{\sqrt{MSE(i)(1 - h_{ii})}}$$

를 정의하고,  $|d_i^*| \geq t(n-k-2, \frac{\alpha}{2})$  이면  $y_i$ 를 이상점이라 판정한다. 실제로 스튜던트화 제외잔차의 계산은

$$d_i^* = e_i \left[ \frac{n-k-2}{SSE(1-h_{ii}) - e_i^2} \right]^{\frac{1}{2}}$$

와 같이 하는 것이 자료를 제외하고 다시 적합하지 않아도 구할 수 있어 계산량이 줄어든다.

### 7.3.3 영향치의 식별

회귀모형을 적합시켜서 유의하면 예측모형으로 사용한다. 따라서 적합된 회귀모형은 안정성(stability)을 가져야 하며 어떤 특정 자료의 영향을 크게 받지 않아야 한다. 어떤 자료를 제외하고 적합한 모형이 크게 달라질 때 그 자료를 영향치(influential cases)라 한다.

영향치를 식별하기 위한 통계량들의 공통점은 모든 자료를 포함시켜서 얻은 분석결

과와  $i$ 번째 자료를 제외하고 얻은 분석결과의 차이를 재는 측도라는 것이다. 예를 들어 모든 자료를 포함시켜 얻은 회귀계수 추정량  $b$ 와  $i$ 번째 자료를 제외시켜 얻은  $b(i)$ 와의 차이를 비교하거나 또는 적합값  $\hat{y}_i$ 와  $i$ 번째 자료를 뺀 후의 적합값  $\hat{y}_{-i}$ 의 차이를 비교함으로써 이 차이가 크면 영향력 있는 관찰점이라 판단하는 것이다. 영향치를 식별하기 위하여 많이 사용되는 통계량들로는 Cook의 D, COVRATIO, DFFITS, DFBETAS 등이 있다.

i) Cook의 D-통계량(Cook's distance measure)

통계학자 Cook이 제안한 통계량으로  $i$ 번째 자료에 대한 D-통계량은 다음과 같이 정의한다.

$$D_i = \frac{(b - b(i))' X' X (b - b(i))}{(k+1)MSE} = \frac{(\hat{y} - \hat{y}(i))' (\hat{y} - \hat{y}(i))}{(K+1)MSE}$$

이때

$$D_i > F(k+1, n-k-1; \alpha)$$

이면 영향력 있는 관찰점이라 판단한다. Cook은 대략적으로

$$D_i > F(k+1, n-k-1; 0.05)$$

이면 영향치일 것으로 제안하였다. 실제로  $D_i$ 의 계산은 다음과 같은 스튜던트화 잔차와의 관계를 이용하면 자료를 제외하고 다시 적합하지 않아도 구할 수 있어 편리하다.

$$D_i = \frac{e_i^*}{(k+1)MSE} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

ii) COVRATIO

모든 자료를 포함시켜 얻은 회귀계수추정량의 분산공분산행렬의 행렬식과  $i$ 번째 자료를 제외하고 얻은 회귀계수추정량의 분산공분산행렬의 행렬식의 비율로 정의되는  $i$ 번째 자료에 대한 통계량 COVRATIO <sub>$i$</sub> 는 다음과 같다.

$$\text{COVRATIO}_i = \det[\text{MSE}(i)(X(i)'X(i))^{-1}] / \det[\text{MSE}(X'X)^{-1}]$$

통계량  $COVRATIO_i$ 가 1에서 멀리 떨어질수록  $i$  번째 자료가 강력한 영향치로 볼 수 있으며, Belsley, Kuh, Welsh(1980)은

$$|COVRATIO_i - 1| > 3(k+1)/n$$

인 경우  $i$  번째 자료를 영향치로 볼 것을 제안하였다.

iii) DFFITS 통계량

모든 자료를 포함시켜 얻은 회귀계수 추정량  $b$ 와  $i$ 번째 자료를 제외시켜 얻은  $b(i)$ 와의 차이를 가장 크게 하는 자료를 찾는  $DFFITS_i$  통계량은 다음과 같다.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{-i}}{\sqrt{MSE(i)h_{ii}}}$$

Belsley, Kuh, Welsh(1980)은

$$|DFFITS_i| > 2\sqrt{\frac{k+1}{n}}$$

인 경우  $i$  번째 자료를 영향치로 볼 것을 제안하였다.

iv) DFBETAS 통계량

자료가 각각의 회귀계수에 미치는 영향력을 나타내는 통계량으로  $i$ 번째 자료의  $j$ 번째 회귀계수  $b_j$ 에 대한 통계량  $DFBETAS_{j(i)}$ 는 다음과 같이 정의한다.

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{C_{jj}MSE(i)}}$$

여기서  $C_{jj}$ 는  $(X^T X)^{-1}$ 의  $j$ 번째 대각원소이다. 만약

$$|DFBETAS_{j(i)}| > \frac{2}{\sqrt{n}}$$

이면  $i$ 번째 자료를 영향치로 본다.

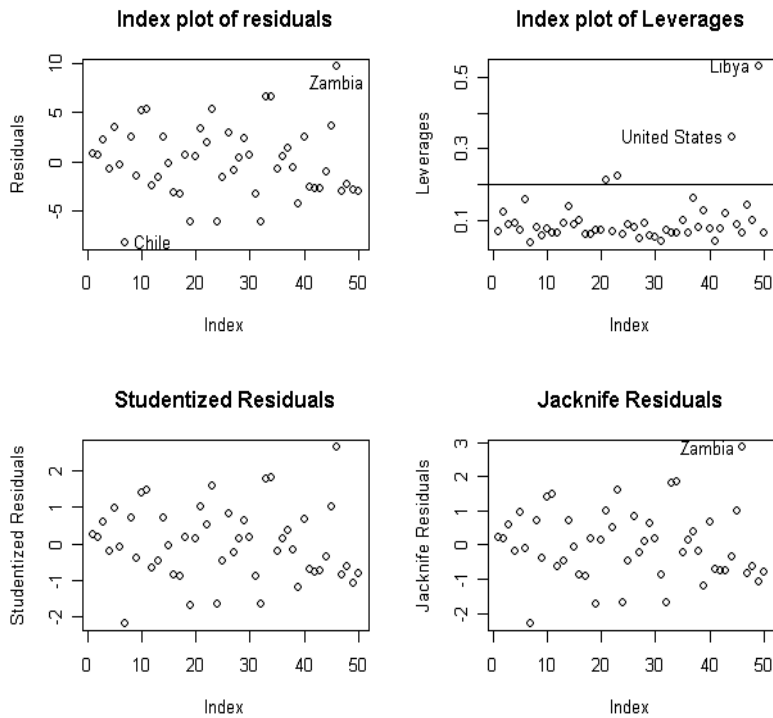
지금까지 알아본 Cook의 D,  $COVRATIO$ ,  $DFFITS$ ,  $DFBETAS$  등의 통계량은 하나의 자료만 제외하고 차이를 계산하는 방법들이었다. 따라서 두 개 이상의 자료가 개별적으로는 영향치로 식별되지 않지만 동시에 결합하면 영향을 주는 즉 결합영향력이

큰 경우 이러한 통계량들은 무의미하게 된다. Cook과 Weisberg(1982)는 이런 현상을 가면효과(masking effect)라 하였다.

### 7.3.4 예

예를 들기 위하여 앞에서 다룬 savings 자료에 대한 회귀진단을 해보자. 우선 자료에 회귀모형을 적합하고 잔차에 대한 그림을 그려 본다.

```
> gf <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
> plot(gf$res, ylab="Residuals", main="Index plot of residuals")
```



[그림 7.4] 잔차 그림

결과는 그림 7.4의 왼쪽 위 패널에 나타나 있으며, 가장 큰 잔차와 가장 작은 잔차에 대응하는 국가를 다음과 같이 식별할 수 있다.

```
> sort(gf$res)[c(1, 50)]
      Chile      Zambia
```

```
-8.242231  9.750914
```

가장 큰 잔차와 가장 작은 잔차에 대응하는 국가를 식별하는 또 다른 방법은 `identify()` 함수를 사용하는 것이다. 먼저 데이터프레임으로부터 행이름을 가져오는 함수 `row.names()`를 사용하여 국가이름을 요소로 하는 문자형 벡터를 만들어야 함을 유의하자.

```
> countries <- row.names(savings)
> identify(1:50, gf$res, countries)
```

식별을 원하는 점 근처에서 마우스 왼쪽 버튼을 클릭하면 국가이름이 나타나게 되며 중지하고 싶을 때는 마우스 오른쪽이나 중앙의 버튼을 클릭한다. 그림 7.4의 좌상에 이러한 방법으로 Chile와 Zambia가 식별되어 있다.

먼저 레버리지에 대해 알아보자. 다음과 같이 X-행렬을 `model.matrix()` 함수를 사용하여 도출하고 레버리지를 계산한 뒤 그림으로 나타낸다.

```
> x <- model.matrix(gf)
> lev <- hat(x)
> plot(lev, ylab="Leverages", main="Index plot of Leverages")
> abline(h=2*5/50)
> sum(lev)
[1] 5
```

레버리지의 합은 절편을 포함한 설명변수들의 개수( $p$ )인 5임을 확인하였고, 어떤 국가가 큰 레버리지를 갖는지 주먹구구식으로 판단하는 기준으로  $h=2p/n(=2*5/50=0.2)$ 를 횡선으로 그림 7.4의 오른쪽 위에 나타내었다. 이 기준을 넘는 국가들을 다음과 같이 식별할 수 있다.

```
> names(lev) <- countries
> lev[lev > 0.2]
```

Ireland	Japan	United States	Libya
0.2122363	0.2233099	0.3336880	0.5314568

함수 `names()`는 벡터 `lev`의 요소에 국가이름을 할당하여 식별하기 쉽게 해주는 역할을 하였다. 상호작용으로 이러한 식별을 하려면 다음과 같이 `identify()` 함수를 사용

한다.

```
> identify(1:50, lev, countries)
```

그림 7.4의 오른쪽 위에 Libya와 United States가 이 방법으로 식별되어 있다. 이제 스튜던트화 잔차를 구해보자.

```
> gs <- summary(gf)
> gs$sig
[1] 3.802669
> stud <- gf$res/(gs$sig*sqrt(1-lev))
> plot(stud, ylab="Studentized Residuals", main="Studentized Residuals")
```

그림 7.4의 왼쪽 아래에 나타나 있는 표준화 잔차가 좌상에 나타나 있는 원래의 잔차와 별 차이가 없어 보인다. 특별히 큰 레버리지가 있을 때만 가시적인 차이가 나타날 것이다.

다음으로 스튜던트화 제외잔차를 구해보자.

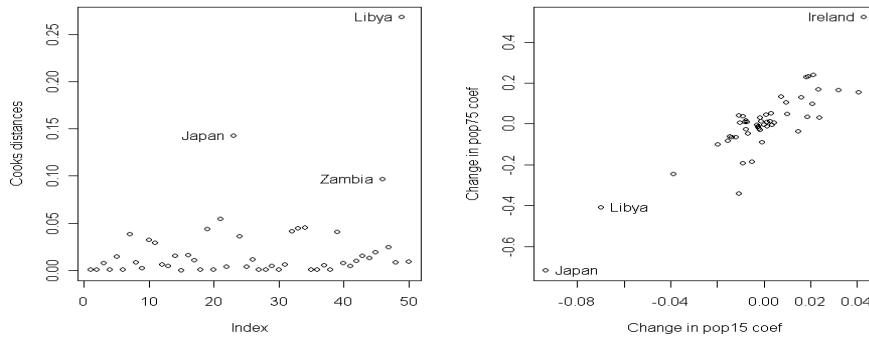
```
> jack <- rstudent(gf)
> plot(jack, ylab="Jackknife Residuals", main="Jackknife Residuals")
> jack[abs(jack)==max(abs(jack))]
```

Zambia  
2.853558

스튜던트화 제외잔차의 최대값이 2.85보다도 더 크게 나온다. 그림 7.5의 오른쪽 아래에 식별되어 있는 Zambia를 이상치로 판단할 것인가는 여러 가지를 고려하여 판단해야 한다.

다음으로 영향치를 찾는 방법을 알아본다. 영향치를 식별하기 위한 기준의 하나인 쿡 통계량이 그림 7.5의 왼쪽에 나타나 있으며 3개의 영향치를 식별하였다.

```
cook <- cooks.distance(gf)
plot(cook, ylab="Cooks distances")
identify(1:50, cook, countries)
```



[그림 7.5] 영향치 식별

가장 큰 쿡 통계량의 값을 갖는 관측치를 제외하고 적합시키면 다음과 같이 계산된다.

```
> gl <- lm(sr ~ pop15+ pop75+ dpi+ ddpi,savings,subset=(cook < max(cook)))
> summary(gl)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	24.5240460	8.2240263	2.982	0.00465 **
pop15	-0.3914401	0.1579095	-2.479	0.01708 *
pop75	-1.2808669	1.1451821	-1.118	0.26943
dpi	-0.0003189	0.0009293	-0.343	0.73312
ddpi	0.6102790	0.2687784	2.271	0.02812 *

Residual standard error: 3.795 on 44 degrees of freedom

Multiple R-Squared: 0.3554, Adjusted R-squared: 0.2968

F-statistic: 6.065 on 4 and 44 DF, p-value: 0.0005617

비교를 위하여 원래의 자료에 대한 적합 결과를 보자.

```
> summary(gf)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.5660865	7.3545161	3.884	0.000334 ***
pop15	-0.4611931	0.1446422	-3.189	0.002603 **
pop75	-1.6914977	1.0835989	-1.561	0.125530
dpi	-0.0003369	0.0009311	-0.362	0.719173
ddpi	0.4096949	0.1961971	2.088	0.042471 *



Residual standard error: 3.803 on 45 degrees of freedom  
 Multiple R-Squared: 0.3385, Adjusted R-squared: 0.2797  
 F-statistic: 5.756 on 4 and 45 DF, p-value: 0.0007904

변수 `ddpi`에 대한 회귀계수의 추정값이 50%가량 변화였다. 한 국가에 대한 자료가 추정값에 이렇게 큰 영향을 주는 것은 모형이나 추론의 안정성에 문제를 야기할 수 있다. 각 국가의 영향력을 이런 식으로 모두 조사하는 것은 비효율적이므로 다음과 같이 빠른 방법을 택한다.

```
> ginf <- lm.influence(gf)
> plot(ginf$coef[,2], ginf$coef[,3], xlab="Change in pop15 coef",
+      ylab="Change in pop75 coef")
> identify(ginf$coef[,2], ginf$coef[,3], countries)
```

## 7.4 잔차그림

이상치와 영향치는 개별 관측치 중 유별난 것이 있음을 나타낸다. 그러나 한편으로는 모형에 대한 가정도 검토할 필요가 있다. 우선 잔차를 적합값  $\hat{y}$ 에 대해 그려 본다. 이 그림이 회귀진단에서 가장 중요한 역할을 한다. 아무 문제가 없다면 잔차를 나타내는 세로축을 기준으로 일정한 분산을 보일 것이며 0을 기준으로 대칭인 산포를 할 것이다. 등분산이 아닌 경우 이분산성(heteroscedasticity)을 보일 것이다. 비선형성을 보여서 모형의 변환이 필요하다는 것을 나타낼 수도 있다.

또한 잔차를 설명변수  $x_i$ 에 대해 그려 보아야 한다. 이분산성이나 비선형성을 보는 외에 모형에 포함되어 있지 않은 설명변수의 경우 어떤 관계를 갖고 있다면 포함여부를 고려해야 한다. 예를 위하여 `savings` 자료를 다시 살펴보자.

```
> gf <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
```

먼저 잔차와 적합값 그림을 그리고 다음으로 잔차의 절대값과 적합값 그림을 그린다.

```
> plot(gf$fit, gf$res, xlab="Fitted", ylab="Residuals")
> abline(h=0)
> plot(gf$fit, abs(gf$res), xlab="Fitted", ylab="|Residuals|")
```

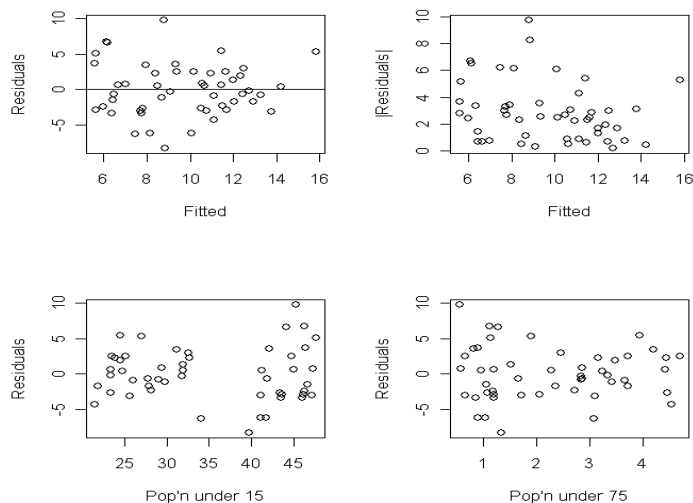
이 두 그림이 그림 7.6의 상단에 나타나 있는데 두 번째 것은 이분산성만을 검토하기 위한 그림이다. 즉 첫 번째 그림의 절반 아래를 접어 올려 이분산성을 좀 더 명백하게 탐지 할 수 있도록 한 것이다. 첫 번째 그림은 비선형성을 검토하는데 반드시 필요하다. 그림 7.의 하단에 나타나 있는 두 그림은 잔차와 설명변수에 대해 다음과 같이 그린 것이다.

```
> plot(savings$pop15, gf$res, xlab="Pop'n under 15", ylab="Residuals")
> plot(savings$pop75, gf$res, xlab="Pop'n under 75", ylab="Residuals")
```

그림 7.6의 하단 왼쪽 그림에서 적합값 35를 기준으로 잔차가 두 그룹으로 나뉘질 수 있음이 보인다. 두 그룹의 잔차에 대해 분산이 동일한지 검토해보자. 분산비가 등분산을 주장하는 귀무가설하에서 F분포를 한다는 것을 이용하여 검토하는 절차는 다음과 같다.

```
> var(gf$res[savings$pop15 > 35])/var(gf$res[savings$pop15 < 35])
[1] 2.785067
> table(savings$pop15 > 35)
FALSE TRUE
    27    23
> 1-pf(2.7851, 22, 26)
[1] 0.006787451
```

확률값이 작아 두 그룹의 잔차에 대해 분산이 유의한 차이가 있다고 판단한다.



[그림 7.6] 잔차그림

잔차그림을 보고 회귀진단을 하는 연습을 위해 가상 자료를 생성하고 그림으로 나타내는 방법을 생각해보자. 다음과 같이 반복을 위한 함수 `for()`를 사용하여 등분산, 강한 이분산, 약한 이분산, 비선형을 나타내는 네 가지 경우에 대한 자료를 생성하고 그림으로 나타낼 수 있다.

```
> par(mfrow=c(3, 3))  
> for(i in 1:9) plot(1:50, rnorm(50))  
> for(i in 1:9) plot(1:50, (1:50)*rnorm(50))  
> for(i in 1:9) plot(1:50, sqrt((1:50))*rnorm(50))  
> for(i in 1:9) plot(1:50, cos((1:50)*pi/25)+ rnorm(50))
```

독립변수와 종속변수 사이의 관계가 선형인 경우 단순선형회귀분석이나 중선형회귀 분석 등을 사용하였다. 그러나 독립변수와 종속변수 사이의 관계가 곡선 형태인 경우가 있다.

## 8.1 독립변수가 하나인 다항회귀

독립변수가 하나인 이차다항회귀모형

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

이 적합한 경우를 생각해보자. 행렬을 사용하여 나타내면

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

으로 놓을 때  $y = X\beta + \varepsilon$ 과 같은 형태가 되어 앞에서 다룬 선형모형에 대한 회귀분석과 모든 이론이 동일하게 된다. 따라서 최소제곱추정량은 다음과 같이 구해진다.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1} X'y.$$

결국 설명변수가 둘인 중회귀모형에서 두 번째 설명변수가 첫 번째 설명변수의 제곱인 특수한 경우라고 생각하는 것이다. 그러나 선형모형에서의 회귀계수는 설명변수가 한 단위 증가할 때 반응변수의 변화량에 대한 기대값을 뜻하지만 다항모형에서는 다르게 해석해야 하므로 주의해야 한다.

다항 회귀 모형에서 차수의 결정방법은 다음과 같다.

- 1) 산점도를 보고 다항모형에 대한 대략의 차수를 결정하거나,
- 2) 작은 차수에서 큰 차수까지 축차적으로 적합시키고 각각에서  $H_0: \beta_i = 0$ 에 대한 t-검정 또는 F-검정을 행하여 정하거나,
- 3) 충분히 큰 차수라 생각되는 차수를 정하고 높은 차수에서  $H_0: \beta_i = 0$ 를 부분 F-검정하여 채택되면 낮은 차수를 적합시킨다.

주의할 점은 고차항이 유의하면 저차항은 유의하지 않더라도 모형에서 제외하지 않는다는 것이다.

예를 들기 위하여 앞에서 다뤘던 savings 자료에서 변수 ddpi에 의한 다항회귀분석을 생각해보자. 우선 단순선형회귀모형을 적합한 결과는 다음과 같다.

```
> summary(lm(sr ~ ddpi, savings))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8830      1.0110   7.797 4.46e-10 ***
ddpi           0.4758      0.2146   2.217  0.0314 *

Residual standard error: 4.311 on 48 degrees of freedom
Multiple R-Squared:  0.0929,    Adjusted R-squared:  0.074
F-statistic: 4.916 on 1 and 48 DF,  p-value: 0.03139
```

변수 ddpi에 대한 확률값이 작아 유의성은 있으므로 이차항을 추가하기로 한다.

```
> summary(lm(sr ~ ddpi+I(ddpi^2), savings))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.13038    1.43472   3.576 0.000821 ***
ddpi         1.75752    0.53772   3.268 0.002026 **
I(ddpi^2)    -0.09299    0.03612  -2.574 0.013262 *

Residual standard error: 4.079 on 47 degrees of freedom
Multiple R-Squared:  0.205, Adjusted R-squared:  0.1711
F-statistic: 6.059 on 2 and 47 DF,  p-value: 0.004559
```

변수 ddpi^2에 대한 확률값이 작아 유의성이 있으므로 삼차항을 추가하기로 한다.

```
> summary(lm(sr ~ ddpi+ I(ddpi^2)+ I(ddpi^3), savings))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.145e+00	2.199e+00	2.340	0.0237 *
ddpi	1.746e+00	1.380e+00	1.265	0.2123
I(ddpi^2)	-9.097e-02	2.256e-01	-0.403	0.6886
I(ddpi^3)	-8.497e-05	9.374e-03	-0.009	0.9928

Residual standard error: 4.123 on 46 degrees of freedom

Multiple R-Squared: 0.205, Adjusted R-squared: 0.1531

F-statistic: 3.953 on 3 and 46 DF, p-value: 0.01369

변수  $ddpi^3$ 에 대한 확률값이 커서 유의성이 없으므로 이차항 모델을 택한다. 4차항 모형으로부터 최고차항의 유의성을 검정하는 방법으로 차수를 낮출 때도 이차항 모형을 택하게 될지 확인할 수 있을 것이다.

보다 저차항은 유의성을 판단할 필요가 없음을 보이기 위하여  $ddpi$ 에서 10을 뺀 변수를  $mddpi$ 로 하여 이차다항회귀모형을 적합하고 결과를 알아본다.

```
> savings <- data.frame(savings, mddpi=savings$ddpi-10)
```

```
> summary(lm(sr ~ mddpi+ I(mddpi^2), savings))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	13.40705	1.42401	9.415	2.16e-12 ***
mddpi	-0.10219	0.30274	-0.338	0.7372
I(mddpi^2)	-0.09299	0.03612	-2.574	0.0133 *

Residual standard error: 4.079 on 47 degrees of freedom

Multiple R-Squared: 0.205, Adjusted R-squared: 0.1711

F-statistic: 6.059 on 2 and 47 DF, p-value: 0.004559

이차항은 변함이 없으나 일차항은 유의하지 않게 되었다. 보다 저차항의 유의성은 변수변환에 따라 다른 결과를 보이기 때문에 신뢰할 수 없다.

## 8.2 독립변수가 둘인 다항회귀 모형

독립변수가 두 개인 다음과 같은 이차 다항회귀 모형을 생각하자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i .$$

여기에서  $\beta_1, \beta_2$ 는 선형효과 모수,  $\beta_{11}, \beta_{22}$ 는 제곱효과 모수,  $\beta_{12}$ 는 상호작용(교호작용)효과(interaction effect) 모수라고 한다.

다음과 같은 반응변수의 기대값은 반응표면(Response Surface)이라 부른다.

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 .$$

모든 이론은 독립변수가 하나인 경우와 동일하다. 즉  $x_1^2$ ,  $x_2^2$ ,  $x_1 x_2$  등을 서로 다른 독립변수로 생각한 중회귀모형으로 생각하면 된다. 중회귀모형을 행렬을 사용하여  $Y = X\beta + \epsilon$ 로 나타내자. 여기서

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{11}^2 & x_{21}^2 & x_{11}x_{21} \\ 1 & x_{12} & x_{22} & x_{12}^2 & x_{22}^2 & x_{12}x_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n}^2 & x_{2n}^2 & x_{1n}x_{2n} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{22} \\ \beta_{12} \end{pmatrix}$$

이고, 최소제곱추정량은 다음과 같다.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_{11} \\ \hat{\beta}_{22} \\ \hat{\beta}_{12} \end{pmatrix} = (X'X)^{-1} X'Y.$$

주의할 점은 교호작용이 있는지 반드시 검정할 필요가 있다는 것이다.

### 8.3 2차 다항회귀에서 최고 또는 최적점 추정

이차다항회귀모형에서 설명변수들이 어떤 값을 가질 때 평균반응량  $\hat{y}$  이 최대 또는 최소가 될 것인가 찾는 문제를 생각해보자. 이차다항회귀모형은 설명변수의 복합적인 작용에 따른 반응변수의 변화를 나타내는 반응표면의 최적조건을 찾는 통계적인 분석방법인 반응표면분석(response surface analysis)에서 바람직한 근사모형으로 사용된다.

예로 모형  $\hat{y} = b_0 + b_1x + b_2x^2$ 을 최대 또는 최소로 하는 조건을 찾아보자.

우선  $\frac{\partial \hat{y}}{\partial x} = b_1 + 2b_2x = 0$  일 때 최적이므로  $x = -\frac{b_1}{2b_2}$  가 최적조건이다.

따라서  $x = x_m = -\frac{b_1}{2b_2}$  일 때  $\hat{y}$ 은 최대 또는 최소가 된다.

만약  $b_2 > 0$  이면 최소이고,  $b_2 < 0$  이면 최대이다.

최적 반응값은  $\hat{y}_m = b_0 + b_1\left(-\frac{b_1}{2b_2}\right) + b_2\left(-\frac{b_1}{2b_2}\right)^2 = b_0 - \frac{b_1^2}{4b_2}$  이다.

다음으로 독립변수가 두 개인 아래와 같은 이차다항반응표면을 최대 또는 최소로 하는 조건을 찾아보자.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2$$

먼저

$$\frac{\partial \hat{y}}{\partial x_1} = b_1 + 2b_{11}x_1 + b_{12}x_2 = 0$$

$$\frac{\partial \hat{y}}{\partial x_2} = b_2 + 2b_{22}x_2 + b_{12}x_1 = 0$$

일 때 최적이므로, 최적 조건은 다음과 같다.

$$x_{1m} = \frac{2b_1b_{22} - b_2b_{12}}{b_{12}^2 - 4b_{11}b_{22}}, \quad x_{2m} = \frac{2b_2b_{11} - b_1b_{12}}{b_{12}^2 - 4b_{11}b_{22}}$$



이 점을  $\hat{y}$ 에 대입 하면 최적 반응값을 구할 수 있다.

행렬을 사용하여 반응변수에 대한 최적 조건을 찾는 방법을 알아보기 위하여 반응 표면함수  $\hat{y}$ 가 다음과 같은 이차다항회귀모형으로 주어졌다고 하자.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_{12} .$$

여기서

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, p = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, B = \begin{pmatrix} b_{11} & \frac{b_{12}}{2} \\ \frac{b_{12}}{2} & b_{22} \end{pmatrix}$$

로 놓으면

$$\begin{aligned} \hat{y} &= b_0 + (b_1 \ b_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1 \ x_2) \begin{pmatrix} b_{11} & \frac{b_{12}}{2} \\ \frac{b_{12}}{2} & b_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= b_0 + p'x + x'Bx \end{aligned}$$

이고,  $\frac{\partial \hat{y}}{\partial x} = 0$ 을 풀면  $p' + 2Bx = 0$ 이다. 따라서 최적조건은 다음과 같다.

$$x_0 = -\frac{1}{2}B^{-1}p.$$

이  $x_0$ 를 정상점(stationary point)이라 부르며 다음 세 가지 중 하나이다.

- 1)  $\hat{y}$ 이 최대가 되는  $x$ 의 점
- 2)  $\hat{y}$ 이 최소가 되는  $x$ 의 점
- 3)  $\hat{y}$ 이 최대도 최소도 아닌  $x$ 의 안부점(saddle point)

설명변수가 두 개인 경우에는 정상점이 위 세 가지 중 어디에 속하는지 등고선표(contour chart)를 그려서 파악한다. 그러나 실제로 정상점을 식에 의하여 구하는 것은 쉬우나 정상점 주위에서 반응표면의 모양이 어떤 형태인지 아는 것은 쉽지 않다. 이러한 문제를 다루는 통계적 분석방법을 정준분석(canonical analysis)이라 부르며 결론만 간단하게 나열하고 자세한 내용은 박성현 (1999)을 참고하도록 한다.

- 1)  $B$ 가 양정치(positive definite)이면, 즉  $B$ 의 고유치  $\lambda_i$ 들이 모두 양수이면 정상점  $x_0$ 에서의  $\hat{y}_0$ 은 최소값이고 이 때 정상점  $x_0$ 를 최소점이라 한다.
- 2)  $B$ 가 음정치(negative definite)이면, 즉  $B$ 의 고유치  $\lambda_i$ 들이 모두 음수이면 정상점  $x_0$ 에서의  $\hat{y}_0$ 은 최대값이고 이 때 정상점  $x_0$ 를 최대점이라 한다.
- 3) 만약  $B$ 가 비정치(indefinite)행렬이면, 즉  $B$ 의 고유치  $\lambda_i$ 들이 양수도 있고 음수도 있으면  $\hat{y}$ 은 정상점  $x_0$ 에서의  $\hat{y}_0$  보다 클 수도 있고 작을 수도 있어서 이 때 정상점  $x_0$ 를 안부점(saddle point)라 한다.

앞에서는 모든 변수들이 키나 속도 등을 나타내는 양적(quantitative)변수일 때 회귀분석을 하였다. 그러나 나타내는 것이 성별이나 신용상태 등과 같이 범주(category)를 나타내는 질적(qualitative)변수들을 고려해야 하는 경우도 있다.

질적변수는 범주의 수준이 두 가지인 경우 보통 0과 1로 나타내는 것이 편리하며, 이와 같이 0과 1인 값을 수치가 아닌 분류의 의미로 갖는 변수를 가변수(dummy variable)라 부르고, 두 가지 값만 취한다 하여 이진법변수(binary variable)라 부르기도 한다. 수학적 용어로는 지시변수(indicator variable)라 부르기도 한다.

## 9.1 하나의 질적변수

두개의 설명변수를 포함하는 중회귀모형에서 질적변수가 하나 포함된 경우를 생각해 보자. 예를 들어, 반응변수는 제품생산에 소요되는 시간이고 설명변수  $x_1$ 은 작업자의 근무기간일 때 설명변수  $x_2$ 는 제조방법을 나타내는 가변수로 A 방법인 경우 1, B 방법인 경우 0으로 나타낸다고 하자.

이 경우 설정할 수 있는 중회귀모형은 다음과 같다.

$$y_j = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i .$$

여기서  $x_{2i} = \begin{cases} 1 : A \text{ 방법} \\ 0 : B \text{ 방법} \end{cases}$ ,  $\epsilon_i \sim N(0, \sigma^2)$ ,  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ ,  $i \neq j$ .

이 모형의 반응함수는  $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ 이고, 가변수  $x_{2i}$ 의 값에 대응하는 범주의 수준에 따라 다음과 같이 나타낼 수 있다.

$$A \text{ 방법일 때, } E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 (1) = (\beta_0 + \beta_2) + \beta_1 x_1,$$

$$B \text{ 방법일 때, } E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 (0) = \beta_0 + \beta_1 x_1.$$

이 가변수의 추가는 기율기에는 영향을 주지 않고 절편에만 영향을 주었고 제품생산에 소요되는 시간에 대한 두 방법의 차이는  $\beta_2$ 로 나타난다.

이러한 모형에 대한 추정 및 검정은  $x_2$ 가 질적변수이지만 양적변수인 것처럼 취급하는 가변수이므로 통상적인 최소제곱법과 분산분석법을 그대로 적용하면 된다. 즉

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & 1 \\ 1 & x_{12} & 1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & 0 \end{pmatrix}$$

으로 놓고 최소제곱추정량을 구하면 다음과 같고 검정방법도 앞에서 다룬 바와 같다.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = (X'X)^{-1}X'y.$$

질적변수가 범주형으로 주어져 있을 경우 가변수로 만들어 주는 것이 필요하다.

## 9.2 상호작용항을 포함하는 모형

앞에서 다룬 모형에서는 설명변수  $x_1, x_2$  사이에 교호작용(interaction)을 고려하지 않았다. 만약 제조방법에 따라 근무기간의 효과가 다르다면  $x_1, x_2$  사이에 교호작용이 있다는 뜻이므로 상호작용항을 추가해야 한다. 상호작용항을 추가한 모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i.$$

여기서  $x_{2i}$ 는 가변수이고, 오차항은  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ 이다.

이 모형에서 교호작용은 두변수의 곱으로 표현하였다. 교호작용의 존재는 많은 경우 현실적으로 독립변수와 종속변수 사이의 의미를 파악하는데 어려움을 준다. 따라서 교호작용이 “0”인지 아닌지, 즉 “ $\beta_3 = 0$ ”을 검정할 필요가 있다. 두 독립변수의 곱

으로 표현된 교호작용이 무슨 의미인지 알아보자.

위의 교호작용이 있는 모형의 반응함수는  $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$  이고, 가변수  $x_2$ 의 값에 대응하는 범주의 수준에 따라 다음과 같이 나타낼 수 있다.

$$x_2 = 0 \text{ 이면, } E(y|x) = \beta_0 + \beta_1 x_1$$

$$x_2 = 1 \text{ 이면, } E(y|x) = \beta_0 + \beta_1 x_1 + \beta_3 x_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$$

따라서,  $\beta_2$ 는 절편의 차이를 나타내며,  $\beta_3$ 는 기울기의 차이를 나타낸다. 이러한 모형에 대한 추정 및 검정도  $x_2$ 가 질적변수이지만 양적변수인 것처럼 취급되는 가변수이므로 통상적인 최소제곱법과 분산분석법을 그대로 적용하면 된다. 즉

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \beta_3 \\ 1 & x_{11} & 1 & x_{11} \\ 1 & x_{12} & 1 & x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & 1 & x_{1n_1} \\ \vdots & \vdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & 0 & 0 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

으로 놓고 최소제곱추정량을 구하면 다음과 같고 검정방법도 앞에서 다룬 바와 같다.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{pmatrix} = (X'X)^{-1}X'y$$

만약 질적변수의 범주가 세 수준 이상인 경우 가변수를 여러 개 사용하여 회귀분석을 할 수 있다. 일반적으로 범주가  $q$  수준인 질적변수는  $q$ 개 가변수를 사용하여 나타내면  $X'X$  행렬이 비정칙이 되어 통상적인 최소제곱추정량을 구할 수 없는 문제가 있어,  $(q-1)$ 개 가변수를 사용하여 나타내어야 한다.

예를 들어 계절이라는 질적변수는 범주가 봄, 여름, 가을, 겨울의 네 수준이다. 그렇지만 계절이라는 질적변수를 나타내기 위하여 필요한 가변수는 봄이면 1 이고 아니면 0인  $x_1$ , 여름이면 1 이고 아니면 0인  $x_2$ , 가을이면 1 이고 아니면 0인  $x_3$ 인 세 개이다. 이 세 개의 가변수가 모두 0인 경우가 겨울을 나타낸다. 만약 겨울이면 1이고 아니면 0인 가변수  $x_4$ 를 추가하면  $x_1 + x_2 + x_3 + x_4 = 1$ 이 되어  $X$ 의 첫 번째 열인 절편항과 같게 되어  $X'X$ 가 완전계수행렬이 되지 못하여 역행렬을 구할 수 없는 문제가 있다.

### 9.3 구간별 선형 회귀

양적인 독립변수  $x$ 의 범위에 따라 종속변수  $y$ 의  $x$ 에 관한 회귀관계가 다르게 나타나는 경우 가변수를 사용하여 구간별 선형회귀(piecewise linear regression)모형을 적합시키는 방법을 사용할 수 있다.

예를 들어 제품 생산비용( $y$ )이 생산량( $x$ )에 비례하지만 어떤 일정 생산량( $x_p$ ) 기준으로 비례형태가 달라지는 경우가 있다. 이를 나타내는 구간별 선형회귀모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_{1j} + \beta_2 (x_{1j} - x_p) x_{2j} + \epsilon_j$$

여기서  $x_1$ 은 생산량을 나타내는 양적변수이고, 가변수는 다음과 같이 정한다.

$$x_{2j} = \begin{cases} 1, & x_{1j} > x_p \\ 0, & x_{1j} < x_p \end{cases}.$$

반응함수는  $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - x_p) x_2$ 와 같이 표현된다.

만약  $x_1 > x_p$  이면  $x_2 = 1$  이고

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - x_p) \cdot 1 = (\beta_0 - \beta_2 x_p) + (\beta_1 + \beta_2) x_1$$

만약  $x_1 < x_p$  이면  $x_2 = 0$  이고

$$E(y) = \beta_0 + \beta_1 x.$$

따라서  $x_1$ 이  $x_p$ 보다 클 경우와 작을 경우 각각에 대하여 절편과 기울기가 다른 두개의 직선을 얻는다. 따라서 두 개의 선형회귀직선을 적합시켜야 하는 것으로 보일 수 있으나 위와 같이 가변수를 이용하면 하나의 모형으로 적합시킬 수 있고 추정 및 검정의 모든 이론은 앞에서의 중회귀 분석과 같다. 즉,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & (x_{11} - x_p)x_{21} \\ 1 & x_{12} & (x_{12} - x_p)x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & (x_{1n} - x_p)x_{2n} \end{pmatrix}, \quad x_{2j} = \begin{cases} 1, & x_{1j} \geq x_p \\ 0, & x_{1j} < x_p \end{cases}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

으로 놓고 최소제곱추정량을 구하면 다음과 같고 검정방법도 앞에서 다룬바와 같다.

$$\hat{\beta} = (X'X)^{-1} X'Y, \quad V(\hat{\beta}) = (X'X)^{-1} \cdot \sigma^2$$

예를 들기 위하여 ‘savings’ 자료를 설명변수 ‘pop15’를 중심으로 분석한다. 먼저 변수 ‘pop15’가 35%보다 큰 경우와 작은 경우에 대해 각각 회귀모형을 적합하고 결과는 그림 9.1에서 실선으로 나타낸다.

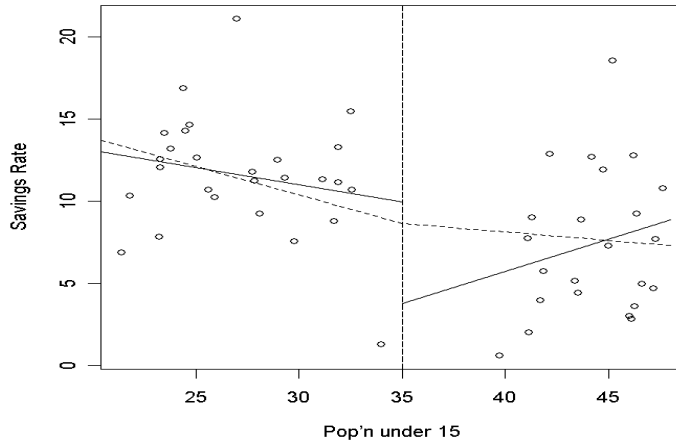
```
> g1 <- lm(sr ~ pop15, savings, subset=(pop15 < 35))
Coefficients:
(Intercept)      pop15
      17.2747      -0.2085
> g2 <- lm(sr ~ pop15, savings, subset=(pop15 > 35))
Coefficients:
(Intercept)      pop15
      -9.8702       0.3900
> plot(savings$pop15, savings$sr, xlab="Pop'n under 15", ylab="Savings Rate")
> abline(v=35, lty=5)
> segments(20, g1$coef[1]+g1$coef[2]*20, 35, g1$coef[1]+g1$coef[2]*35)
> segments(48, g2$coef[1]+g2$coef[2]*48, 35, g2$coef[1]+g2$coef[2]*35)
```

이렇게 두 부분에 대해 각각 적합된 회귀직선은 기준점에서 만나지 않으며 설명변수가 변함에 따라 적합이 연속적으로 이뤄져야 한다면 적절하지 않게 된다. 이때 다음과 같이 가변수를 이용한 회귀모형을 적합시키는 것이 하나의 해결 방법일 것이다. 결과는 [그림 9.1]에서 점선으로 나타낸다.

```
> d <- function(x) ifelse (x > 35, 1, 0)
> gd <- lm(sr ~ pop15+I((pop15-35)*d(pop15)), savings)
> gd
Coefficients:
              (Intercept)              pop15      I((pop15 - 35) * d(pop15))
                20.8092                -0.3471                  0.2430
> x <- seq(20, 48, by=1)
> py <- gd$coef[1]+gd$coef[2]*x+gd$coef[3]*(x-35)*d(x)
> lines(x, py, lty=2)
```

이 경우 가변수를 이용한 모형과 두 집단에 대하여 각각 적합한 모형 중 어떤 것이 나은가 생각해 보자. 두 집단의 경계인 중앙부에는 국가가 별로 없어 연속성을 꼭 필요로 하지 않으며 두 집단에 대한 적합한 직선의 기울기가 아주 달라 보인다고 주장

할 수 있다.



[그림 9.1] 구간별 선형회귀

## 9.4 종속가변수

종속변수  $y$ 가 범주의 수준이 생존과 사망과 같은 질적변수인 경우에도 가변수를 이용하여 로지스틱(logistic) 회귀분석과 같은 유용한 방법을 사용할 수 있다. 종속변수가 하나의 가변수로 나타내진다면 0 또는 1인 값을 취하므로 이진반응변수(binary response variable)라고 부른다.

### 9.4.1 반응함수의 의미와 문제점

단순회귀모형에서 종속변수가 가변수인 경우 모형은

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n, \quad y_i = 0 \text{ 또는 } 1$$

이 되며, 만약  $E(\epsilon_j) = 0$  이면 이 모형은  $E(y|x) = \beta_0 + \beta_1 x$  으로 표현되고,  $y_i = 0$  또는 1 이므로 변수  $y_i$ 는 다음과 같은 베르누이 분포를 따른다.



$$P(y_i = 1) = p_i, \quad P(y_i = 0) = 1 - p_i = q_i$$

따라서

$$E(y_i | x_i) = \beta_0 + \beta_1 x_i = 1 \times P(y_i = 1) + 0 \times P(y_i = 0) = p_i$$

이므로 독립변수가  $x_i$ 로 주어질 때,  $y_i$ 의 기대반응은 “ $y_i = 1$ ”인 확률을 나타낸다는 점을 유의해야 한다.

또한 종속변수가 0 또는 1만을 취하는 가변수인 경우 여러 문제점이 발생하게 된다. 특히 다음 세 가지 문제점은 통계적 추정과 검정에 어려움을 준다.

- 1) 오차항의 비정규성.
- 2) 오차항의 비등분산성.
- 3) 반응함수의 제약성.

첫 번째로 오차항  $\epsilon_j$ 의 비정규성은 다음처럼 확인한다. 모형  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  에서  $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$  로 나타낼 때  $y_i = 0$  또는 1 이므로

$$\epsilon_i = \begin{cases} 1 - \beta_0 - \beta_1 x_i, & y_i = 1 \\ -\beta_0 - \beta_1 x_i, & y_i = 0 \end{cases}$$

이 되어  $\epsilon_i$ 는 두 가지 값만 취하므로 정규분포가 아닌 것을 확인할 수 있으며 정규성 가정하에서 사용했던 통계적 추론방법들이 적절하지 않음을 알 수 있다.

두 번째로  $\epsilon_j$ 의 비등분산성은 다음과 같이 확인할 수 있다. 오차항의 분산은

$$V(\epsilon) = V(y_j) = E(y_j^2) - (E(y_j))^2$$

과 같고

$$E(y_j) = p_j, \quad E(y_j^2) = 1^2 \times P(y_j = 1) + 0^2 \times P(y_j = 0) = p_j$$

이므로

$$V(\epsilon_j) = p_j - p_j^2 = p_j(1 - p_j) = (\beta_0 + \beta_1 x_j)(1 - \beta_0 - \beta_1 x_j)$$

가 되어  $V(\epsilon_j)$ 는  $x_j$ 값에 따라서 이차함수 꼴로 변한다. 따라서 일반적인 최소제곱법(OLS)을 그대로 사용할 수 없고 등분산성을 갖도록 변환시킬 필요가 있다.

마지막으로 반응함수는 조건부확률이므로

$$0 \leq E(y|x) = \beta_0 + \beta_1 x = p \leq 1$$

이어야 한다. 따라서 주어진 자료  $x$ 의 범위 안에서  $0 \leq \beta_0 + \beta_1 x \leq 1$ 을 만족하도록 제약을 주어야 한다.

## 9.4.2 일반화 최소제곱법에 의한 모수추정

일반적인 중회귀모형  $y = X\beta + \epsilon$ 에서 분산공분산 행렬이 다음과 같다 하자.

$$V = V(\epsilon) = V \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} = \begin{pmatrix} V(\epsilon_1) & \text{Cov}(\epsilon_1, \epsilon_2) & \cdots & \text{Cov}(\epsilon_1, \epsilon_n) \\ \text{Cov}(\epsilon_1, \epsilon_2) & V(\epsilon_2) & \cdots & \text{Cov}(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\epsilon_1, \epsilon_n) & \text{Cov}(\epsilon_2, \epsilon_n) & \cdots & V(\epsilon_n) \end{pmatrix}$$

이러한 일반적인 중회귀모형  $y = X\beta + \epsilon$ ,  $V(\epsilon) = V$ 을 다음과 같이 변환하자.

$$V^{-\frac{1}{2}}y = V^{-\frac{1}{2}}X\beta + V^{-\frac{1}{2}}\epsilon.$$

여기서  $X^* = V^{-\frac{1}{2}}X$ ,  $y^* = V^{-\frac{1}{2}}y$ 이라 놓으면

$$y^* = X^*\beta^* + \epsilon^*, \quad V(\epsilon^*) = I$$

와 같이 변환되어 통상적인 중회귀모형이 되며, 최소제곱추정량은 다음과 같이 구할 수 있다.

$$\begin{aligned} \hat{\beta}^* &= (X^{*'}X^*)^{-1}X^{*'}y^* \\ &= (X'V^{-\frac{1}{2}}V^{-\frac{1}{2}}X)^{-1}(X'V^{-\frac{1}{2}}V^{-\frac{1}{2}}y) \\ &= (X'V^{-1}X)^{-1}(X'V^{-1}y). \end{aligned}$$

이와 같은 추정법을 일반화 최소제곱법(Generalized Least Squares method, GLS)라 부른다.

일반화 최소제곱법을 설명하기 위하여 R 패키지에 들어 있는 데이터셋 ‘Longley’의 회귀 자료를 살펴보자. 반응변수 ‘Employed’는 1947년에서 1962년까지 연도별로 고용된 사람 수를 나타내고, 설명변수로는 ‘GNP’와 ‘Population’만 고려하기로 한다.

```
> data(longley)
> gl <- lm(Employed ~ GNP + Population, data=longley)
> summary(gl, cor=T)
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	88.93880	13.78503	6.452	2.16e-05 ***
GNP	0.06317	0.01065	5.933	4.96e-05 ***
Population	-0.40974	0.15214	-2.693	0.0184 *

Residual standard error: 0.5459 on 13 degrees of freedom  
 Multiple R-Squared: 0.9791, Adjusted R-squared: 0.9758  
 F-statistic: 303.9 on 2 and 13 DF, p-value: 1.221e-11  
 Correlation of Coefficients:

	(Intercept)	GNP
GNP	0.98	
Population	-1.00	-0.99

이와 같은 시계열 자료는 자기상관이 있을 가능성이 있다. 오차항에 대해 다음과 같은 일차자기상관이 있다고 가정하고 자기상관계수를  $\rho$ 라고 하자.

$$\epsilon_{i+1} = \rho\epsilon_i + \delta_i.$$

여기서  $\delta_i \sim N(0, \tau^2)$ 인 분포를 가정한다. 자기상관계수  $\rho$ 를 다음과 같이 추정할 수 있다.

```
> cor(gl$res[-1], gl$res[-16])
[1] 0.3104092
```

이 같은 가정하에서는  $V = \rho^{|i-j|}$ 이다. 우선  $\rho$ 를 0.31041로 놓고 GLS 추정량을 구해보자.

```
> x <- model.matrix(gl)
> V <- diag(16)
> V <- 0.31041^abs(row(V)-col(V))
> Vi <- solve(V)
> xtxi <- solve(t(x) %*% Vi %*% x)
> beta <- xtxi %*% t(x) %*% Vi %*% longley$Employed
> beta
```

	[,1]
(Intercept)	94.8988949
GNP	0.0673895

```
Population -0.4742741
```

실제로  $\rho$  값을 알지 못하기 때문에 GLS 모형을 적합한 후 다시 추정해야 한다.

```
> res <- longley$Employed - x %*% beta
> cor(res[-1], res[-16])
[1] 0.3564162
```

자기상관계수  $\rho$ 를 0.3564162로 놓고 다시 모형을 적합하고, 또 다시  $\rho$ 를 추정하는 과정을 수렴할 때까지 반복한다.

이러한 GLS 적합을 하는 라이브러리가 nlme이고 다음과 같이 사용할 수 있다.

```
> library(nlme)
> gl <- gls(Employed ~ GNP + Population, correlation=corAR1(form= ~Year),
+ data=longley)
> summary(gl)
Correlation Structure: AR(1)
Formula: ~Year
Parameter estimate(s):
      Phi
0.6441692
```

```
Coefficients:
              Value Std.Error   t-value p-value
(Intercept)  101.85813  14.198932   7.173647  0.0000
GNP           0.07207   0.010606   6.795485  0.0000
Population   -0.54851   0.154130  -3.558778  0.0035
```

```
Residual standard error: 0.689207
```

```
Degrees of freedom: 16 total; 13 residual
```

자기상관계수  $\rho$ 가 0.64로 추정되었지만 구간추정을 다음과 같이 하여 자기상관계수  $\rho$ 가 0과 유의한 차이가 없음을 알 수 있다.

```
> intervals(gl)
Approximate 95% confidence intervals
```

```
Coefficients:
```

	lower	est.	upper
(Intercept)	71.18320461	101.85813306	132.5330615
GNP	0.04915865	0.07207088	0.0949831
Population	-0.88149053	-0.54851350	-0.2155365

Correlation structure:

	lower	est.	upper
Phi	-0.4444668	0.6441692	0.9646106

Residual standard error:

	lower	est.	upper
	0.2474787	0.6892070	1.9193820

오차항  $\epsilon_j$ 에 대한 정규성 가정은 할 수 없지만 독립성 가정은 할 수 있는 경우를 생각해보자. 특히  $\epsilon_j$ 가 서로 독립이고 베르누이분포를 따르는 경우

$$V = \begin{pmatrix} V(\epsilon_1) & 0 & \cdots & 0 \\ 0 & V(\epsilon_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V(\epsilon_n) \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n(1-p_n) \end{pmatrix}$$

가 된다. 여기서

$$V^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{p_1(1-p_1)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{p_2(1-p_2)}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{p_n(1-p_n)}} \end{pmatrix}$$

으로 놓으면

$$V(V^{-\frac{1}{2}} \cdot \epsilon) = V^{-\frac{1}{2}} \cdot V(\epsilon) \cdot V^{-\frac{1}{2}'} = V^{-\frac{1}{2}} \cdot V \cdot V^{-\frac{1}{2}} = \mathbf{I}.$$

따라서  $V^{-\frac{1}{2}}$  행렬을 곱해줌으로써 이분산성을 등분산성으로 만들어 줄 수 있다.

분산공분산행렬이 대각행렬인 경우 가중최소제곱법(Weighted Least Squares method, WLS)으로도 추정이 가능하며, 실제로  $p_i$ 들의 값을 모르기 때문에 다음과 같은 절차를 따르는 것이 일반적이다.

- 1) 통상적인 최소제곱법(OLS)에 의하여  $b_0, b_1$ 을 구한다.
- 2)  $p_j = \beta_0 + \beta_1 x_j$ 의 관계식에서  $p_j$ 를  $\hat{p}_j = b_0 + b_1 x_j$ 로 추정한다.
- 3)  $\hat{p}_j$ 를 이용하여 분산공분산행렬을 다음과 같이 추정한다.

$$\hat{V} = \begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & 0 & \cdots & 0 \\ 0 & \hat{p}_2(1-\hat{p}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{p}_n(1-\hat{p}_n) \end{pmatrix}$$

- 4) GLS 추정량은

$$b^* = (X' \hat{V}^{-1} X)^{-1} (X' \hat{V}^{-1} y)$$

에 의하여 구한다. 또한

$$\text{Var}(b^*) = (X' \hat{V}^{-1} X)^{-1}$$

을 구한다.

자료의 수가 많을 때는 중심극한정리에 의하여 정규성 가정하에서와 같은 방법으로 통계적 추론을 할 수도 있다.

## 9.5 로지스틱 반응함수와 로지스틱 회귀분석

종속변수( $y$ )가 이진반응변수인 경우에 두 변수  $x$ 와  $y$ 사이의 관계가 선형이 아니고 반응함수의 모양이 S자 형태의 곡선인 비선형(non-linear)일 때가 많다. 이 반응함수는  $x$ 가 증가함에 따라  $E(y|x)$ 의 값이 1로 서서히 수렴하는 성질을 보이며 반응함수의 제약성을 만족하게 된다.

예를 들어 종속변수  $y$ 가 생존을 나타내는 수준 A에서는 1, 사망을 나타내는 수준 B에서는 0인 값을 취하는 가변수라 하자. 생존(즉 A) 확률  $p = P(y = 1)$ 에 영향을 주는 독립변수를  $x$ 라 할 때,  $p$ 의 승산(오즈, odds)이라 불리는  $\frac{p}{1-p}$ 을 로그변환한 값

인 로짓(logit), 즉  $\ln(\frac{p}{1-p})$ 와의 관계가 선형회귀모형  $\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x$ 으로 주

어지면  $p$ 와  $x$  사이에는 로지스틱 회귀관계가 있다고 한다. 따라서  $p$ 와  $x$  사이에 로지스틱 회귀관계가 있다면, 승산은  $\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x)$ 와 같이 나타난다. 승산은

성공확률과 실패확률의 비이므로 승산이 1 또는 로짓이 0 이면 성공확률이 0.5가 되며 그 때의  $x$  값을 독성학에서는 LD50(Lethal Dose 50%)이라고 한다. 승산이 1 또는 로짓이 0보다 크다는 것은 성공확률이 실패확률보다 크다는 것을 의미한다.

로지스틱 회귀에서 절편( $\beta_0$ )의 의미는  $x=0$ 에서 승산의 로그값으로 해석할 수 있으나 주어진  $x$  범위 밖에서 발생할 수 있는 외삽의 위험은 유의해야 한다.

로지스틱 회귀에서 주어진  $x$ 와  $x+1$ 에서의 승산비(odds ratio)는 다음과 같다.

$$\left[ \frac{p(x+1)}{1-p(x+1)} \right] / \left[ \frac{p(x)}{1-p(x)} \right] = \exp(\beta_1).$$

따라서, 기울기( $\beta_1$ )의 의미는  $x$ 가 한 단위 증가할 때 승산비의 로그값으로 해석할 수 있으며 기울기가 양수라면 승산비가 1보다 크다는 것이므로  $x$ 가 증가할 때 승산 또는 성공확률이 높아짐을 뜻한다.

로지스틱 회귀관계를 포함한 좀 더 일반적인 선형모형들을 다루는 방법을 일반화 선형모형(Generalized Linear Models)이라 부른다. 일반화 선형모형은 반응변수가 정규분포일때뿐만 아니라 이항분포, 포아송분포, 감마분포 등을 포함하는 지수분포족(exponential family)일 때도 연결(link) 함수를 사용하여 설명변수들과 회귀관계를 나타낼 수 있도록 Nelder와 Weddurburn(1972)이 체계화한 방법이다. 로지스틱 회귀관계는 로짓이 선형모형이어서 연결 함수가 로짓인 일반화 선형모형의 하나로 볼 수 있고, R commander 메뉴 막대의 ‘통계’에서 ‘적합성 모델’의 ‘일반화 선형 모델...’을 사용할 때 ‘연결 함수’의 선택사항이기도 하다. ‘연결 함수’의 선택사항에는 ‘logit’ 외에 정규분포의 누적분포함수의 역함수인 프로빗(‘probit’)과 꼬리 부분 확률이 희소(rare)한 경우에 대안으로 사용할 수 있는 비대칭적인 변환을 하는 부로그로그(‘cloglog’, complementary log-log)가 있다. ‘cloglog’는 생존 자료 등과 관련하여 알려진 Gompertz 분포와 관계가 있어서 ‘Gompit’이라고도 한다.

로지스틱 회귀관계를  $p$ 에 대해 풀어 쓰면  $p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$ 이 되며, 이러한

$l(.) = \frac{e^{(.)}}{1+e^{(.)}}$ 와 같은 형태의 곡선을 로지스틱 곡선이라 부른다. 로지스틱 곡선의 특

징은 증가함수이고 곡선이 0과 1 사이에 있으며 S자 형태여서 성장모형(growth model)에 자주 사용되고 있다.

주어진  $x$ 에서의 조건부 기대값  $E(y|x) = p$ 이므로

$$\beta_0 + \beta_1 x \rightarrow 0 \text{ 일 때 } E(y|x) = \frac{e^0}{1+e^0} = \frac{1}{2},$$

$$\beta_0 + \beta_1 x \rightarrow \infty \text{ 일 때 } E(y|x) = \frac{e^\infty}{1+e^\infty} \rightarrow 0,$$

$$\beta_0 + \beta_1 x \rightarrow -\infty \text{ 일 때 } E(y|x) = \frac{e^{-\infty}}{1+e^{-\infty}} \rightarrow 1$$

이고, 따라서  $0 \leq E(y) \leq 1$ 인 제약조건이 만족된다.

이러한 로지스틱 곡선은 비선형이지만 쉽게 선형으로 만들 수 있다. 이를 위해  $p$ 에 대한 어떤 변환(transformation)  $f(p)$ 가  $x$ 에 대해 선형인, 즉  $f(p) = \beta_0 + \beta_1 x$ 을 만족하는  $p$ 의 함수  $f$ 를 찾는다. 우선  $f(p) = \beta_0 + \beta_1 x$ 에서  $\exp(f(p)) = \exp(\beta_0 + \beta_1 x)$

이다. 따라서  $\frac{\exp(f(p))}{1+\exp(f(p))} = \frac{\exp(\beta_0 + \beta_1 x)}{1+\exp(\beta_0 + \beta_1 x)} = p$ , 즉  $\frac{\exp(f(p))}{1+\exp(f(p))} = p$ 를 만족하

는  $f(p)$ 를 찾아야 한다. 식  $\exp(f(p)) = p + p \cdot \exp(f(p))$ 로부터  $(1-p)\exp(f(p)) = p$  이므로  $\exp(f(p)) = \frac{p}{1-p}$ , 즉  $f(p) = \ln \frac{p}{1-p}$ 가 되고 실수 전

체에서 값을 가질 수 있게 된다.

이러한 변환  $f(p)$ 를 로지스틱 변환(logistic transformation)이라 부르고  $p^* = \ln \frac{p}{1-p}$

를 로짓이라 하였다. 모형  $p^* = \beta_0 + \beta_1 x$ 에서  $p^*$ 는 알 수 없으며 따라서 추정해야 된다.

즉  $x_1$ 수준에서  $p_1^*$ ,  $x_2$ 수준에서  $p_2^*$ , ...,  $x_n$ 수준에서  $p_n^*$ 를 추정해야 한다. 결국 각  $x_i$ 수준에서 반복이 이루어져야 하며 이를 통하여  $p_i^*$ 를 추정해야 한다. 이 때  $p_i$ 의 추정량은

$\frac{R_i}{n_i}$ 이다. 여기서  $n_i$ 는  $i$ 번째 수준에서의 반복수이고  $R_i$ 는  $n_i$ 개의 측정치 중  $y=1$ 인



개수이다. 따라서  $\bar{p}_i = \frac{R_i}{n_i}$ 는  $i$ 번째 수준에서  $y=1$ 인 확률  $p_i$ 의 추정량이 되며, 로짓의

추정량은  $\bar{p}_i^* = \ln\left(\frac{\bar{p}_i}{1-\bar{p}_i}\right)$ 와 같이 구할 수 있다.

[정리] 변환된 모형  $\bar{p}_i^* = \beta_0 + \beta_1 x + \epsilon_i$  에서  $n_i$ 가 충분히 크면  $V(\bar{p}_i^*) \approx \frac{1}{n_i p_i (1-p_i)}$  이다.

(증명) 비율  $\frac{R_i}{n_i} = \bar{p}_i$  는  $n_i$ 가 커지면, 중심극한정리에 의하여  $\bar{p}_i \sim N\left(p_i, \frac{p_i(1-p_i)}{n_i}\right)$ 이

다. 따라서 분산안정화정리(delta-theorem, variance stabilizing theorem)를 사용하면

$$f(\bar{p}_i) \sim N(f(p_i), [f'(p_i)]^2 V(\bar{p}_i))$$

과 같은 근사적인 정규분포를 한다. 여기서,  $f(p_i) = p_i^* = \ln\left(\frac{p_i}{1-p_i}\right)$ 이므로

$$\begin{aligned} f'(p_i) &= \frac{\partial \ln\left(\frac{p_i}{1-p_i}\right)}{\partial p_i} = \frac{\partial (\ln p_i - \ln(1-p_i))}{\partial p_i} \\ &= \frac{1}{p_i} - \frac{-1}{1-p_i} = \frac{1}{p_i(1-p_i)} \end{aligned}$$

이며, 구하는 분산은 다음과 같다.

$$V(f(\bar{p}_i)) = V(\bar{p}_i^*) = \left(\frac{1}{p_i(1-p_i)}\right)^2 \cdot \frac{p_i(1-p_i)}{n_i} = \frac{1}{n_i p_i (1-p_i)}.$$

또한, 구해진 분산의 추정량은  $\widehat{Var}(\bar{p}_i^*) = \frac{1}{n_i \bar{p}_i (1-\bar{p}_i)}$  이다.

이렇게 변환된 모형  $\bar{p}_i^* = \beta_0 + \beta_1 x + \epsilon_i$  에서는 등분산성이 만족되지 않기 때문에 일반화 최소제곱법으로 구한 추정량은 다음과 같이 구해진다.

$$b^* = (X' \widehat{V}^{-1} X)^{-1} X' \widehat{V}^{-1} \bar{p}_i^*.$$

여기서 계획행렬  $X$ 와 분산공분산행렬의 추정량  $\widehat{V}$ 은

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_k \end{pmatrix}, \quad \hat{V} = \begin{pmatrix} \frac{1}{n_1 \bar{p}_1 (1 - \bar{p}_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2 \bar{p}_2 (1 - \bar{p}_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_k \bar{p}_k (1 - \bar{p}_k)} \end{pmatrix}$$

이며, 로지스틱 함수의 추정식은 다음과 같이 구한다.

$$\hat{p} = \frac{\exp(b_0^* + b_1^* x)}{1 + \exp(b_0^* + b_1^* x)}.$$

일반화 선형모형 방법에서 로지스틱 회귀모형의 모수 추정은 이와 같은 일반화 최소제곱법 대신 반복법을 사용하는 최우추정법으로 추정하며 R commander에서는 일반화 선형모형 방법을 채택하고 있다.

[예] Faraway(2002)에 소개된 snoring 자료에서 설명변수로 코고는 정도(snore)를 점수 (0, 2, 4, 5)로 나타내고, 종속변수를 심장질환(heartdisyes) 여부(있다, 없다)로 하여 로지스틱 회귀분석을 해보자.

우선 자료를 snoring이라는 데이터셋으로 입력한다.

```
> snoring <- data.frame( snore = c(0, 2, 4, 5), heartdisyes = c(24, 35, 21, 30),
n = c(1379, 638, 213, 254) )
> snoring
  snore heartdisyes    n
1     0           24 1379
2     2           35  638
3     4           21  213
4     5           30  254
```

심장질환을 가진 비율을 반응변수로 하고 n을 가중치로 하여 로지스틱회귀분석을 한다.

```
> snoring.lg <- glm(heartdisyes/n ~ snore, weights=n, family=binomial(),
data=snoring)
> snoring.lg
Coefficients:
```

```
(Intercept)      snore
      -3.8662      0.3973
Degrees of Freedom: 3 Total (i.e. Null); 2 Residual
Null Deviance:      65.9
Residual Deviance: 2.809      AIC: 27.06
```

심장질환 여부에 대한 도수를 반응변수로 하여 모델을 적합하려면 행렬로 데이터프레임에 추가한다.

```
> snoring$YN <- cbind(snoring$heartdisyes, snoring$n-snoring$heartdisyes)
> snoring
  snore heartdisyes   n YN.1 YN.2
1     0           24 1379   24 1355
2     2           35  638   35  603
3     4           21  213   21  192
4     5           30  254   30  224
```

행렬을 반응변수로 하여 가중치 없이 적합시켜도 같은 결과를 얻는다.

```
> snoring.lg <- glm(YN ~ snore, family=binomial(), data=snoring)
Coefficients:
(Intercept)      snore
      -3.8662      0.3973

Degrees of Freedom: 3 Total (i.e. Null); 2 Residual
Null Deviance:      65.9
Residual Deviance: 2.809      AIC: 27.06
```

따라서, 코고는 정도(snore)가  $x$ 일 때, 심장질환일 확률의 추정치를 구하면  $\hat{p}(x) = \frac{\exp(-3.8662 + 0.3973x)}{1 + \exp(-3.8662 + 0.3973x)}$  이다. 코고는 정도가 0, 즉  $x=0$ 일 때 심장질환

일 확률의 추정치를 구하면  $\hat{p}(0) = \frac{\exp(-3.8662)}{1 + \exp(-3.8662)} = 0.0205$  이다.  $x=0$ 일 때 절편

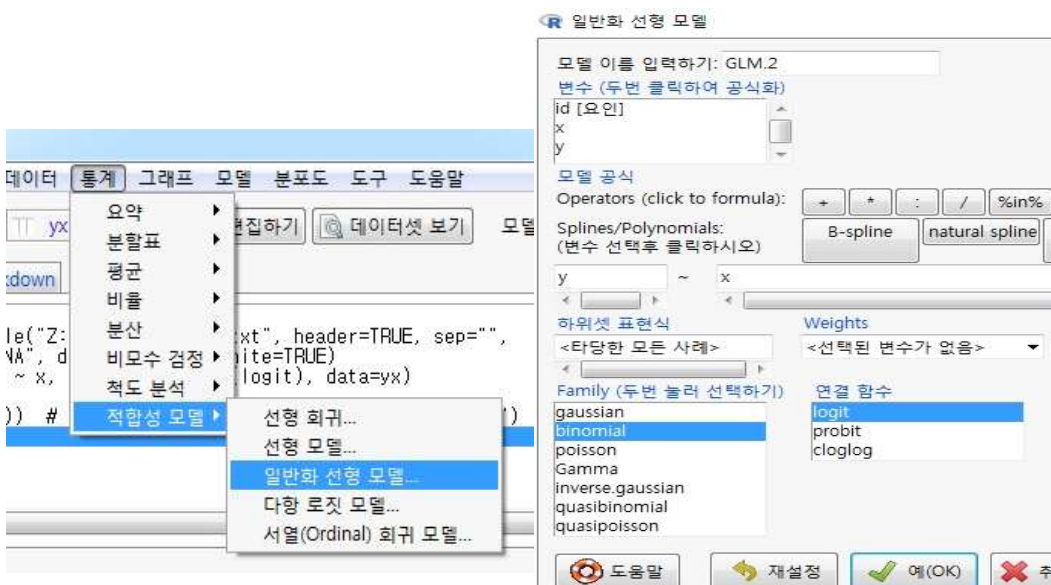
‘-3.8662’는 승산의 로그값이므로 지수함수로 역변환하여 승산을 계산하면 심장질환일 확률이 아닐 확률의  $\exp(-3.8662) = 0.0209$ 배로 추정된다. 코고는 정도가  $x$ 에서  $x+1$ 로 한 단위 증가할 때 승산비는 기울기 ‘0.3973’의 로그값이므로 코고는 정도가  $x$ 와  $x+1$ 에서의 심장질환일 확률과 아닐 확률의 승산비(odds ratio)를 추정하면

$\left[ \frac{\hat{p}(x+1)}{1-\hat{p}(x+1)} \right] / \left[ \frac{\hat{p}(x)}{1-\hat{p}(x)} \right] = \exp(0.3973) = 1.4878$ 이다. 따라서 코고는 정도가 1, 즉

$x=1$ 일 때의 승산은 코고는 정도가 0일 때와 비교하면 1.4878배가 되어, 심장질환일 확률이 아닐 확률의  $\exp(-3.8662) \times \exp(0.3973) = 0.0209 \times 1.4878 = 0.0311$ 배로 추정된다. 승산이 작아 심장질환일 확률이 아닐 확률에 비해 매우 작기는 하지만, 승산비가 1보다 크기 때문에 코고는 정도가 심해질수록 심장질환 확률이 높아짐을 알 수 있다.

이 예를 R commander를 이용하여 로지스틱 회귀분석을 하려면 종속변수가 이진 반응변수 형태로 자료가 주어져야하므로 그룹화되어 있어 이항분포인 변수에 대한 자료를 다음과 같은 문장을 수행하여 베르누이분포를 따르는 변수에 대한 자료로 수정하고, 부록에서와 같이 텍스트 파일로 만든 뒤 불러들여 'yx'라는 데이터셋을 만든 다음 ['통계' -> '일반화 선형 모델...']을 선택하면 나오는 '일반화 선형 모델' 창에서 '모델 이름 입력하기' 옆 칸에 원하는 이름을 정하여 입력하고, '모델 공식'에 y, x 변수를 선택한 다음 'Family'는 'binomial', '연결 함수'는 'logit'으로 지정한 뒤 '예(OK)'를 누르면 '출력물' 창에 앞에서 다뤘던 로지스틱 회귀분석 결과가 나타난다.

```
> x=rep(snore, n)
> vy=as.vector(rbind(n-heartdisyes,heartdisyes))
> y=rep(rep(0:1,length(snore)),vy)
> yx=cbind(x,y)
```



```

출력물
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.86625    0.16621  -23.261  < 2e-16 ***
x            0.39734    0.05001   7.945  1.94e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 900.83  on 2483  degrees of freedom
Residual deviance: 837.73  on 2482  degrees of freedom
AIC: 841.73

Number of Fisher Scoring iterations: 6

> exp(coef(GLM.1)) # Exponentiated coefficients ("odds ratios")
(Intercept)      x
  0.02093677  1.48785669

```

일반적인 회귀모형의 적합성을 판단하기 위해서는 F-검정을 하였으나 로지스틱 회귀모형의 적합성은 이탈도(deviance)에 의해 판단한다. 이탈도는 관측값 개수만큼 모수를 갖는 포화(saturated) 모형의 우도와 현재 모형의 우도의 비의 로그값, 즉 포화 모형의 로그우도에서 현재 모형의 로그우도를 뺀 값의 2배로 정의되며 점근적으로 카이제곱분포를 사용하는 검정통계량이다. 이탈도가 클수록 현재 모형이 적합하지 않다는 증거가 된다. ‘출력물’ 창의 ‘Residual deviance:’ 옆에 나온 ‘837.73’이 현재 모형의 이탈도이며 자유도는 관측치 ‘2484’개에서 추정해야 할 모수 ‘2’개를 뺀 개수인 ‘2482’이다.  $\text{Residual deviance} < \chi^2_{0.05}(2482) = 2800.038$ 이므로 ‘현재 모형이 적합하다’는 귀무가설을 기각할 수 없다.

또한, 일반적인 회귀모형의 추가제곱합을 이용한 부분(partial) F-검정과 유사하게 부분 이탈도를 이용하여 완전모형과 축소모형에 대한  $\chi^2$ -검정을 할 수 있다. ‘출력물’ 창의 각 회귀계수에 대한 z값은 Wald 통계량과 동등하게 일반적인 회귀모형에서의 t-검정과 같이 개별적인 회귀계수의 유의성 검정을 한다. 일반적인 회귀모형에서 설명력의 측도로 사용하는 편차제곱합들의 함수인 결정계수 ‘ $R^2$ ’ 대신 우도의 함수인 Akaike 정보량 기준 ‘AIC’가 ‘출력물’ 창에 나타나 있으며, 모수의 추정은 ‘Fisher Scoring’ 방법으로 하였음도 나타나 있다. 로지스틱 회귀모형의 성능은 성공을 성공으로 예측할 수 있는 ‘민감도(sensitivity)’를 실패를 실패로 예측할 수 있는 ‘특이도(specificity)’에 대해 그런 ROC 곡선으로도 나타낼 수 있다.

일반적인 회귀모형에서와 마찬가지로 설명변수가 여러 개인 다중로지스틱 회귀모형을 설정할 수 있으며, 범주형 설명변수가 있으면 기준(reference) 수준을 정하여 가변수를 사용한다.

또한 10장에서 다룰 일반적인 회귀모형에서 사용하는 F-검정통계량을 이용한 변수선택법과 마찬가지로 이탈도를 이용한 변수선택법을 사용할 수 있다.

## 제 10 장

## 변수의 선택

반응변수에 영향을 줄 것으로 예상되는 설명변수의 수가 많을 경우 어떤 설명변수를 모형에 포함시킬 것인가를 선택하는 문제를 해결해야 한다. 변수를 선택하여 회귀모형을 설정할 때 고려할 원칙은 다음과 같다.

- 1) 가능한 많은 독립변수를 모형에 포함시켜 종속변수의 값을 정확하게 예측한다.
- 2) 너무 많은 개수의 설명변수가 관련되면 자료의 수집, 분석에 시간과 비용이 많이 들고 공선성과 같은 문제가 발생하여 안정성과 신뢰성이 떨어질 수 있기 때문에 가능한 모형에 포함될 독립변수의 개수를 줄인다.

위의 두 가지 원칙은 서로 이율배반적이므로 적절한 판정기준에 의하여 필요한 설명변수만 선택하여 최적회귀방정식을 찾으면 이해하기도 쉽고 예측도 잘할 수 있을 것이다. 이렇게 모형의 설정에 있어서 간단명료한 모형을 선호하는 원리를 모형간명화의 원칙(principle of parsimony)라 한다. 중요하지 않은 변수들까지 포함하는 완전모형보다는 필요한 변수들만 포함하는 축소모형이 더 바람직하다는 원칙이다.

### 10.1 변수선택의 판정기준

반응변수에 영향을 줄 것으로 예상되는 설명변수의 개수가  $k$ 개일 경우 일부인  $p$ 개의 설명변수만을 포함하는 회귀모형은 나타낼 수 있는 경우의 수가

$${}_kC_p = \frac{k!}{(k-p)!p!}, 0 \leq p \leq k$$

이고, 모든 가능한 회귀모형의 경우의 수는

$$\sum_{p=0}^k {}_kC_p = 2^k$$

이다. 만약  $k=10$ 이라면 모든 가능한 회귀모형은  $2^{10}=1024$ 개이다. 이렇게 많은 회귀모형 중에서 최적 모형을 선택하는 기준을 알아보자.

주로 사용되는 판정기준으로는 잔차제곱평균(Mean Squared Error), 결정계수(Coefficient of Determination), 수정결정계수(Adjusted Coefficient of Determination), 맬로우스(Mallows)의  $C_p$ , 아카이케(Akaike)의 정보기준(Information Criteria, AIC), 베이즈(Bayes) 정보기준(Information Criteria, BIC),  $PRESS_p$  등이 있다.

이러한 판정기준은 변수선택의 절대적이라기보다는 참고사항이라고 할 수 있다. 이들 기준들에 따라 적절한 것으로 판단되는 몇 개의 모형을 선정한 다음 잔차 및 영향력 분석, 다중공선성, 자료해석의 실제적인 의미 등을 종합적으로 고려하여 최종적인 모형을 선택하는 것이 바람직하다.

### 10.1.1 잔차제곱평균과 결정계수

절편항( $\beta_0$ )이 있고  $p$ 개의 설명변수를 포함하는 모형에서 잔차제곱평균  $MSE_p$ 는 회귀모형에 의하여 설명되지 않는 변동인 잔차제곱합  $SSE_p$ 와 다음 관계가 있다.

$$MSE_p = \frac{SSE_p}{n-p-1}$$

따라서  $SSE_p$ 를 작게 하는 것은  $MSE_p$ 를 작게 하는 것이다.  $SSE_p$ 는  $p$ 가 증가하면 언제나 감소한다. 그러나  $p$ 가 증가하면  $n-p-1$ 도 감소하므로  $MSE_p$ 는  $p$ 가 증가할 때 감소할 수도 있고 증가할 수도 있다. 즉  $MSE_p$ 가 가장 작게 되는  $p$  값을 찾을 수 있다. 이러한 최소  $MSE_p$  기준과 유사하게 최대수정결정계수 기준을 생각할 수 있다. 결정계수는

$$R_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST}$$

이므로  $SSE_p$ 를 작게 하는 것은  $R_p^2$ 를 크게 하는 것이다. 모든  $k$ 개를 설명변수로 포함한 완전모형에서  $p(\leq k)$ 개의 설명변수를 선택한 축소모형이 나타나는 경우의 수는

${}_kC_p$ 가지이고, 각 경우마다 결정계수  $R_p^2$ 을 계산하여 가장 큰 모형을 선택한다. 그러나  $SSE_p$ 는  $p$ 가 증가하면 언제나 감소하므로  $R_p^2$ 은 항상 증가하게 되어 증가가 둔화되는 경우의  $p$ 를 선택해야 한다. 이러한 최대결정계수 기준은 주관적이며 독립변수가 추가되는 것은 모형간명화 원칙에 반하는 것이어서 독립변수의 개수  $p$ 가 늘어나면 최적성이 감소되도록 결정계수를 수정할 필요가 있다.

### 10.1.2 수정결정계수(adjusted $R_p^2$ , $R_{ap}^2$ )

수정결정계수는

$$R_{ap}^2 = 1 - \frac{n-1}{n-p-1} (1 - R_p^2)$$

과 같이 정의한다. 수정결정계수를  $MSE_p$ 를 사용하여 나타내면

$$R_{ap}^2 = 1 - \frac{n-1}{n-p-1} (1 - R_p^2) = 1 - \frac{n-1}{n-p-1} (SSE_p / SST) = 1 - \frac{n-1}{SST} \cdot MSE_p$$

이므로  $MSE_p$ 를 최소로 하는  $p$ 는  $R_{ap}^2$ 를 최대로 한다.

### 10.1.3 $C_p$

맬로우즈(C. L. Mallows)가 제안한  $C_p$  통계량은 완전모형의  $MSE$ 를  $\hat{\sigma}^2$ 으로 나타낼 때

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2(p+1)$$

이다.  $C_p$ 도  $SSE_p$ 의 함수이나 모형선택을 위한 기준은 다음과 같다.

- (i)  $C_p$ 가  $p+1$ 에 가까운 모형이다.
- (ii)  $C_p$ 가 최소인 모형이다.

이 두 조건을 만족하는  $p$ 를 찾아 최적모형을 선택한다.  $C_p$ 를 종축으로 하고  $p$ 를 횡축으로 하여  $C_p = p+1$ 인 직선 주위에 해당하는 모형중  $C_p$ 가 최소인 것을 택하면 편리하다.

### 10.1.4 $AIC_p$ , $BIC_p$

일반적으로  $AIC_p = -2 \log \text{likelihood} + 2p$ 이고,  $BIC_p = -2 \log \text{likelihood} + p \log n$



이다. 선형회귀모형에서는 deviance로 불리는  $-2 \log \text{likelihood}$ 가  $n \log(SSE/n)$ 이다. 설명변수가 많이 포함된 모형이 SSE가 작지만 대신  $p$ 가 커지므로 절충이 필요하게 된다.  $BIC_p$ 가  $AIC_p$ 에 비하여  $p$ 에 대한 가중치가 크므로 좀 더 작은 모형을 택하게 하는 경향이 있다.

주어진 모형에 대하여 R이나 R commander에서  $AIC_p$ 값은 함수  $AIC(\text{모형})$ 를 사용하여 계산하고,  $BIC_p$ 값은 함수  $AIC(\text{모형}, k=\log(n))$ 로 계산할 수 있다. 도움말을 얻기 위하여  $\text{help}(AIC)$  등을 수행하면 다음과 같은 사용법과 인수에 대한 설명을 볼 수 있다.

사용법:  $AIC(\text{object}, \dots, k)$

인수(Arguments)

object:  $AIC_p$ 값을 계산하려는 모형

... : 선택사항인 추가적 모형 관련 객체

k : 모수 개수에 대한 벌점을 주기 위한 것으로 기본값은  $k = 2$ 이고  $AIC_k$  계산에 사용되며  $k = \log(n)$ 은  $BIC_k$  계산에 사용된다..

## 10.2 변수를 선택하는 방법

앞 절에서 모든 가능한 회귀모형 중 최적인 것을 선택하는 기준을 알아보았다. 그러나 모든 가능한 회귀모형을 특정기준에 의하여 비교하고 최적인 회귀모형을 선택하는 방법은 다음과 같은 문제를 갖고 있다.

첫째로 고려하는 독립변수의 개수가 많을 때 비교해야 할 회귀모형의 수가 너무 많아 비효율적이다.

둘째로 판정기준에 따라 선택되는 최적 모형이 다를 경우 어떤 한 모형을 최적인 것으로 결정할 수 없는 경우가 있다.

이런 문제점을 보완하기 위하여 자동적으로 하나의 모형을 선택하게 해주는 절차가 뒤로부터 제거(backward elimination)라 불리는 변수제거법과 앞으로부터 선택(forward selection)이라 불리는 변수선택법 그리고 단계별 회귀법(stepwise regression method)이 있다.

그러나 미리 정한 알고리즘에 의해 자동적으로 선택된 하나의 모형이 모든 모형 중

에서 최적이라는 보장은 없다는 것을 유의해야 한다. 그리고 알고리즘에 따라 자동적으로 선택된 모형이 서로 다를 수 있으며 이들은 여러 가지 판정기준에 의하여 비교 평가되어야 한다. 이들에 대하여 알아보자.

### 10.2.1 모든 가능한 회귀(all possible regression)

모든 가능한 변수들의 조합, 즉  $k$ 개의 독립변수를 고려하는 경우  $2^k$ 개 조합에 대한 회귀모형 중에서 판정기준인  $R_p^2$ ,  $R_{ap}^2$ ,  $MSE_p$ ,  $C_p$ ,  $AIC_p$ ,  $BIC_p$  등을 적용하여 최적인 모형을 찾는 방법으로 앞서와 같은 문제가 있다.

### 10.2.2 뒤로부터 제거하는 방법(Backward elimination method)

회귀모형에 기여도가 적은 설명변수를 가능한 제거하고 남는 변수들을 선택하는 방법으로 다음과 같은 절차를 따른다.

- 1) 모든 독립변수를 포함하는 회귀방정식을 적합시킨다.
- 2) 각 변수의 기여도를 측정하기 위하여 부분 F-검정통계량  $F_0$ 을 구한다.
- 3) 제일 작은  $F_0$ 값 또는 가장 큰 유의확률(p-value)을 갖는 변수가 미리 정한 유의수준  $\alpha$ 에서 유의하면 현재 모형을 최종 모형으로 정하고 변수선택을 종료하고, 유의하지 않으면 변수를 제거하고 다음 단계로 간다.
- 4) 앞에서 제거한 변수외의 모든 독립변수를 포함하는 회귀방정식을 적합시킨다. 그리고 위의 순서 2단계로 되돌아간다.

뒤로부터 제거하는 변수제거법에서 중요한 기준은 유의수준  $\alpha$ 이며, 통상적으로 사용되는 유의수준은 10%이다. 변수를 좀 더 많이 제거하려면 유의수준을 10%보다 낮게 정하면 될 것이다.

변수제거법에서는 최대로  $k + (k-1) + \cdots + 1 = k(k+1)/2$ 개의 회귀분석을 하면 되므로, 모든 가능한 회귀의  $2^k$ 개에 비교하면 작은 숫자이므로 바람직한 방법으로 보이나, 각 단계마다 한 번 제거된 변수는 절대로 선택되지 않아 최종 모형이 모든 모형 중에서 최적이라는 보장은 없다.

예를 들어 savings 자료에 대하여 유의수준 5%를 기준으로 가장 큰 확률값을 갖는

설명변수를 뒤로부터 제거하는 방법으로 최종 모형을 다음과 같이 택하였다.

```
> summary(gf)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.5660865	7.3545161	3.884	0.000334 ***
pop15	-0.4611931	0.1446422	-3.189	0.002603 **
pop75	-1.6914977	1.0835989	-1.561	0.125530
dpi	-0.0003369	0.0009311	-0.362	0.719173
ddpi	0.4096949	0.1961971	2.088	0.042471 *

Residual standard error: 3.803 on 45 degrees of freedom

Multiple R-Squared: 0.3385, Adjusted R-squared: 0.2797

F-statistic: 5.756 on 4 and 45 DF, p-value: 0.0007904

```
> gf <- update(gf, . ~ . - dpi)
```

```
> summary(gf)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.1247	7.1838	3.915	0.000297 ***
pop15	-0.4518	0.1409	-3.206	0.002452 **
pop75	-1.8354	0.9984	-1.838	0.072473 .
ddpi	0.4278	0.1879	2.277	0.027478 *

Residual standard error: 3.767 on 46 degrees of freedom

Multiple R-Squared: 0.3365, Adjusted R-squared: 0.2933

F-statistic: 7.778 on 3 and 46 DF, p-value: 0.0002646

```
> gf <- update(gf, . ~ . - pop75)
```

```
> summary(gf)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	15.59958	2.33439	6.682	2.48e-08 ***
pop15	-0.21638	0.06033	-3.586	0.000796 ***
ddpi	0.44283	0.19240	2.302	0.025837 *

Residual standard error: 3.861 on 47 degrees of freedom  
 Multiple R-Squared: 0.2878, Adjusted R-squared: 0.2575  
 F-statistic: 9.496 on 2 and 47 DF, p-value: 0.0003438

뒤로부터 제거방식으로 dpi와 pop75가 제거된 모형이 최종적으로 택해졌다.

참고로 주어진 모형에 대하여 R이나 R commander에서 함수 step(모형)을 사용하면  $AIC_p$ 를 기준으로 변수선택법을 수행한다. 도움말을 얻기 위하여 help(step) 등을 수행하면 다음과 같은 사용법과 인수에 대한 설명을 볼 수 있다.

```
ii) step(object, scope, scale = 0, direction = c("both", "backward", "forward"),
      trace = 1, keep = NULL, steps = 1000, k = 2, ...)
```

인수(Arguments)

object : 변수선택 초기 모형으로 사용되는 객체

scope : 변수선택에서 검토되는 모형들의 범위를 정의한다. 단일 공식이거나 upper와 lower 성분 공식 모두를 포함하는 리스트일 수 있다. lower 성분 공식의 우측은 모형에 항상 포함되며 검토되는 모형 공식의 우측은 upper 성분 공식에 포함된다. 만약 scope가 단일 공식이면 upper 성분 공식을 지정하는 것이고, lower 성분 공식은 없는 것이다. 만약 scope가 생략되면 초기 모형이 upper 성분 공식이 된다.

scale : used in the definition of the AIC statistic for selecting the models, currently only for lm, aov and glm models. The default value, 0, indicates the scale should be estimated: see extractAIC.

direction: 변수선택 방법을 정하는 인수로 "both", "backward", "forward" 중 하나를 지정하며 기본적으로 "both"가 지정되어 있다. 만약 scope 인수가 생략되면 "backward"가 기본적인 변수선택 방법이 된다.

trace : if positive, information is printed during the running of step. Larger values may give more detailed information.

keep : a filter function whose input is a fitted model object and the associated AIC statistic, and whose output is arbitrary. Typically keep will select a subset of the components of the object and return them. The default is not to keep anything.

steps : the maximum number of steps to be considered. The default is 1000 (essentially as many as required). It is typically used to stop the

process early.

k : the multiple of the number of degrees of freedom used for the penalty.

Only  $k = 2$  gives the genuine AIC:  $k = \log(n)$  is sometimes referred to as BIC or SBC.

... : any additional arguments to extract AIC.

예를 들어 자료 savings 데이터셋에 대하여 함수 `step()`을 사용하여 변수제거법에 의하여 최종 모델을 택하는 방법은 다음과 같다.

```
> gf <- lm(sr ~ pop15 + pop75 + dpi + ddpi, savings)
```

```
> step(gf)
```

```
Start: AIC= 138.3
```

```
sr ~ pop15 + pop75 + dpi + ddpi
```

	Df	Sum of Sq	RSS	AIC
- dpi	1	1.89	652.61	136.45
<none>			650.71	138.30
- pop75	1	35.24	685.95	138.94
- ddpi	1	63.05	713.77	140.93
- pop15	1	147.01	797.72	146.49

```
Step: AIC= 136.45
```

```
sr ~ pop15 + pop75 + ddpi
```

	Df	Sum of Sq	RSS	AIC
<none>			652.61	136.45
- pop75	1	47.95	700.55	137.99
- ddpi	1	73.56	726.17	139.79
- pop15	1	145.79	798.40	144.53

Coefficients:

(Intercept)	pop15	pop75	ddpi
28.1247	-0.4518	-1.8354	0.4278

함수 `step()`에 의해  $AIC_p$ 를 기준으로 기본적으로 수행되는 뒤로부터 제거방식으로 변수 dpi만 제거된 모형이 최종적으로 택해졌음을 알 수 있다. 앞에서 부분 F-검정을 사용하여 뒤로부터 제거방식으로 선택한 최종모형보다 pop75가 제거되지 않은 차이가 있다.

### 10.2.3 앞으로부터 선택하는 방법(forward selection method)

뒤로부터 제거하는 방법은 모든 가능한 독립변수를 포함하는 가장 큰 회귀모형으로부터 시작하여 중요하지 않다고 판단되는 변수를 하나씩 제거해 나가면서 더 이상 제거할 변수가 없다면 종료하고 남아 있는 변수들만 포함하는 회귀모형을 최종 모형으로 정하였다.

이와는 반대로 절편항만 포함하는 가장 작은 회귀모형으로부터 시작해서 기여도가 큰 설명변수를 하나씩 선택해 나가면서 더 이상 선택할 변수가 없다면 종료하고 이미 선택된 변수들만 포함하는 회귀모형을 최종 모형으로 정하는 방법이 앞으로부터 선택하는 방법이며 다음과 같은 절차를 따른다.

- 1) 설명변수  $k$  개의 각각에 대하여 절편항을 포함하는 단순회귀모형을 적합시키고 이 중에서  $R^2$  값이 가장 큰 것 또는 동등하게 F-검정통계량 값을 가장 크게 하는 기여도가 높은 설명변수를 찾는다. 이 변수가 유의하지 않다고 검정되면 절편항만 포함하는 회귀모형을 최종 모형으로 하여 종료하고, 유의하다고 검정되면 선택하여  $x_1$ 으로 나타내고 회귀모형에 포함하여 다음 단계로 넘어 간다.
- 2) 앞에서 선택한  $x_1$  외의 설명변수들 중 하나를 추가하여 독립변수가 두 개인 모형을 적합시킨다. 이 중에서 가장 큰 축차 F-검정통계량 값을 갖는 추가된 변수가 유의하지 않다고 검정되면 절편항과  $x_1$ 만 포함하는 회귀모형을 최종 모형으로 하여 종료하고, 유의하면 선택하여  $x_2$ 로 나타내고 회귀모형에 포함하여 다음 단계로 넘어 간다.
- 3) 앞에서 선택한  $x_1, x_2$  외의 설명변수들 중 하나를 추가하여 독립변수가 세 개인 모형을 적합시킨다. 이 중에서 가장 큰 축차 F-검정통계량 값을 갖는 추가된 변수가 유의하지 않다고 검정되면 절편항과  $x_1, x_2$ 만 포함하는 회귀모형을 최종 모형으로 하여 종료한다. 유의하면 선택하여  $x_3$ 로 나타내고 회귀모형에 포함하여 다음 단계로 넘어 간다. 이런 방식으로  $i$ 번 단계에 왔다면 선택된 설명변수는  $x_1, x_2, \dots, x_i$  이고 나머지 설명변수들 중 하나를 추가하여 독립변수가  $i+1$  개인 모형을 적합시킨다. 이 중에서 가장 큰 축차 F-검정통계량 값을 갖는 추가된 변수가 유의하지 않다고 검정되면 절편항과  $x_1, x_2, \dots, x_i$ 만 포함하는 회귀모형을 최종 모형

으로 하여 종료하고, 유의하면 선택하여  $x_{i+1}$ 로 나타내고 회귀모형에 포함하여 다음 단계로 넘어 가는 절차를 종료할 때까지 반복한다. 추가되는 변수  $x_j (i < j \leq k)$ 에 대한 측차 F-검정통계량의 값  $F_j$ 는 다음과 같다.

$$F_j = \frac{R(\beta_j | \beta_0, \dots, \beta_i)}{MSE_{i+1}}, \quad j = i+1, \dots, k$$

여기서  $MSE_{i+1}$ 은  $i+1$ 개의 설명변수들  $x_1, x_2, \dots, x_i, x_j$ 를 포함하는 모형의 잔차제곱평균을 나타낸다.

변수선택법에서도 중요한 기준은 유의수준  $\alpha$ 이다. 보통  $\alpha=0.5(50\%)$ 를 사용하며 좀 더 많은 변수를 선택하려면 유의수준을 50%보다 더 높게 정해야 한다. 변수선택법도 최대로  $k+(k-1)+\dots+1=k(k+1)/2$ 개의 회귀분석을 하면 되므로, 모든 가능한 회귀의  $2^k$ 개에 비교하면 작은 숫자이므로 바람직해 보이나, 각 단계마다 한 번 선택된 변수는 절대로 제거되지 않아 최종 모형이 모든 모형 중에서 최적이라는 보장은 없다.

주어진 모형에 대하여 R이나 R commander에서 함수 `step(모형, direction="forward")`를 사용하여  $AIC_p$ 를 기준으로 수행한다.

## 10.2.4 단계별 회귀방법(stepwise regression method)

변수선택법을 보완하기 위하여 변수제거법을 결합한 방법으로, 매 단계마다 선택과 제거를 반복하면서 중요한 변수를 찾아내는 방법이다. 단계별 회귀방법은 중요한 변수를 하나씩 추가하면서 이미 선택된 변수들 중 추가되는 변수 때문에 중요성을 상실하여 제거될 것이 있는지를 매 단계마다 검토하는 방법이다. 그러나 이 방법에 의해서 선택된 최종 모형도 모든 모형 중에서 최적이라는 보장은 없다.

단계별 회귀방법의 절차는 다음과 같다.

- 1) 변수선택법의 1단계와 같다.
- 2) 앞에서 선택한  $x_1$ 외의 설명변수들 중 하나를 추가하여 독립변수가 두 개인 모형을 적합시킨다. 이 중에서 가장 큰 측차 F-검정통계량 값을 갖는 추가된 변수가 유의하지 않다고 검정되면 절편항과  $x_1$ 만 포함하는 회귀모형을 최종 모형으로 하여 종료한다.

추가된 변수가 유의하면 선택하여  $x_2$ 로 나타내고 회귀모형에 포함한다. 이  $x_1, x_2$

를 포함하는 회귀모형에서 각 변수에 대한 부분 F-검정통계량 값  $F_0$ 를 구하여 제일 작은  $F_0$ 값 또는 가장 큰 유의확률(p-value)을 갖는 변수가 미리 정한 유의수준  $\alpha$ 에서 유의하면 다음 단계로 간다. 유의하지 않으면 변수를 제거하는데 만약 축차 F-검정에서 유의하여 추가된 변수인  $x_2$ 가 부분 F-검정에서 유의하지 않아 제거되는 경우에는 절편항과  $x_1$ 만 포함하는 회귀모형을 최종 모형으로 하여 종료하고 그렇지 않으면 다음 단계로 간다.

3) 앞에서 선택된 설명변수들 외의 한 설명변수를 추가하여 모형을 적합시킨다. 이 중에서 가장 큰 축차 F-검정통계량 값을 갖는 추가된 변수가 유의하지 않다고 검정되면 절편항과 앞에서 선택된 설명변수들만 포함하는 회귀모형을 최종 모형으로 하여 종료한다.

추가된 변수가 유의하면 선택하여 회귀모형에 포함하고 이 회귀모형에서 각 변수에 대한 부분 F-검정통계량 값  $F_0$ 를 구하여 제일 작은  $F_0$ 값 또는 가장 큰 유의확률(p-value)을 갖는 변수가 미리 정한 유의수준  $\alpha$ 에서 유의하면 다음 단계로 간다. 유의하지 않으면 변수를 제거하는데 만약 축차 F-검정에서 유의하여 추가된 변수가 부분 F-검정에서 유의하지 않아 제거되는 경우에는 절편항과 앞에서 선택된 설명변수들만 포함하는 회귀모형을 최종 모형으로 하여 종료하고 그렇지 않으면 다음 단계로 간다.

이런 방식으로 선택된 설명변수가  $x_1, x_2, \dots, x_i$  인 경우 나머지 설명변수들 중 하나를 추가하여 독립변수가  $i+1$  개인 모형을 적합시킨다.

이 중에서 가장 큰 축차 F-검정통계량 값을 갖는 추가된 변수가 유의하지 않다고 검정되면 절편항과  $x_1, x_2, \dots, x_i$ 만 포함하는 회귀모형을 최종 모형으로 하여 종료한다.

유의하면 선택하여  $x_{i+1}$ 로 나타내어 회귀모형에 포함하고 이 회귀모형에서 각 변수에 대한 부분 F-검정통계량 값  $F_0$ 를 구하여 제일 작은  $F_0$ 값 또는 가장 큰 유의확률(p-value)을 갖는 변수가 미리 정한 유의수준  $\alpha$ 에서 유의하면 다음 단계로 간다. 유의하지 않으면 변수를 제거하는데 만약 축차 F-검정에서 유의하여 추가된 변수가 부분 F-검정에서 유의하지 않아 제거되는 경우에는 절편항과 앞에서 선택된 설명변수들만 포함하는 회귀모형을 최종 모형으로 하여 종료하고 그렇지 않으면 다음 단계로 넘어 가는 절차를 종료할 때까지 반복한다.



단계별 회귀방법이 끝나는 경우는 다음 두 가지이다.

- (i) 모형 밖의 설명변수 중에서 모형 안으로 들어올 변수가 없는 경우, 즉 축차 F-검정통계량 값이 최대인 변수가 유의하지 않은 경우
- (ii) 모형에 추가되었던 변수가 즉시 모형에서 제거되는 경우

변수선택은 자체가 목적이 되는 것은 아니다. 목적은 자료에 있는 관계를 잘 설명하고 예측할 수 있는 모형을 정하는데 있다. 자동적인 변수선택법은 이런 목적에 부합된 결과를 준다는 보장은 없으며 단지 지침이 될 수 있는 것이다.

단계별 변수선택법은 가능한 모형들 중에서 제한된 탐색을 하고 반복되는 검정을 통해 모형을 비교하여 택하는 방식이다. 판정기준에 의한 방식이 좀 더 많은 모형들 중에서 보다 편리하게 탐색하고 비교할 수 있다는 점 때문에 선호된다.

## [부 록] R commander의 설치와 사용

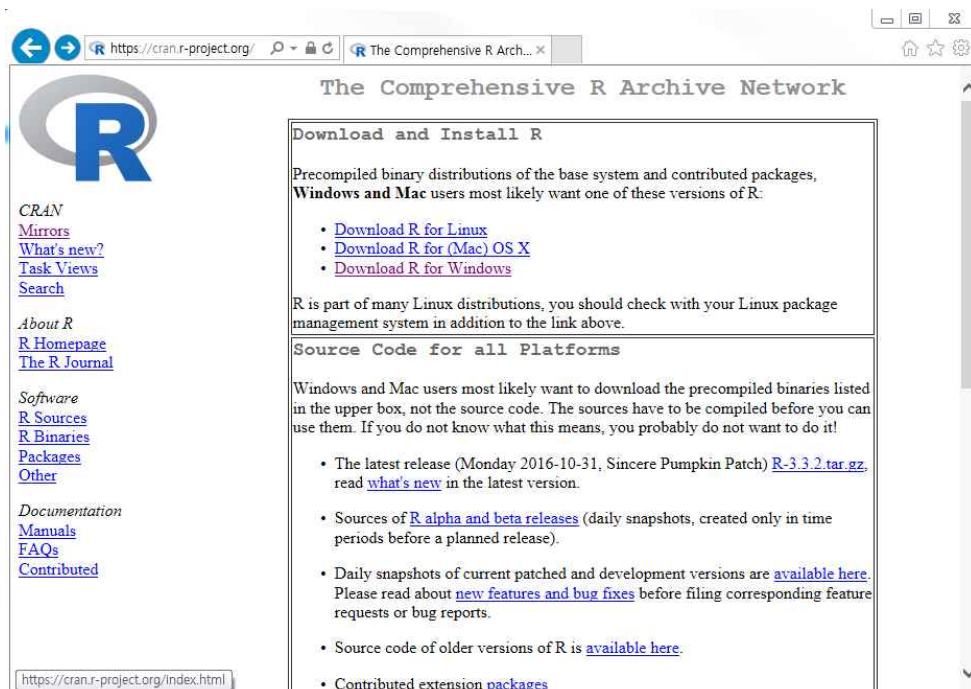
R은 지금은 Lucent Technologies로 바뀐 예전의 AT&T사 Bell 연구소의 John Chambers 등에 의해 개발된 S 언어와 비슷한 통계계산과 그래픽을 위한 언어이다. R은 S와 몇 가지 차이점은 있지만 많은 경우 S 언어로 작성한 프로그램이 R에서도 수정 없이 실행된다. R에 대한 추가적인 정보는 <http://www.r-project.org/> 에서 찾을 수 있다.


R commander(Rcmdr)는 R을 사용하여 기본적인 통계분석을 하기 쉽도록 John Fox 등이 tcltk 패키지에 기초하여 개발한 GUI(Graphic User Interface)이다. R commander에 대한 추가적인 정보는 <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/> 에서 찾을 수 있다.

### 1. R의 설치와 실행

윈도우스 운영체제하의 R 버전을 [그림 A-1]에 나타나 있는 웹사이트 <http://cran.r-project.org/> 에서 찾은 다음 설치하는 법을 알아보도록 한다.

[그림 A-1]

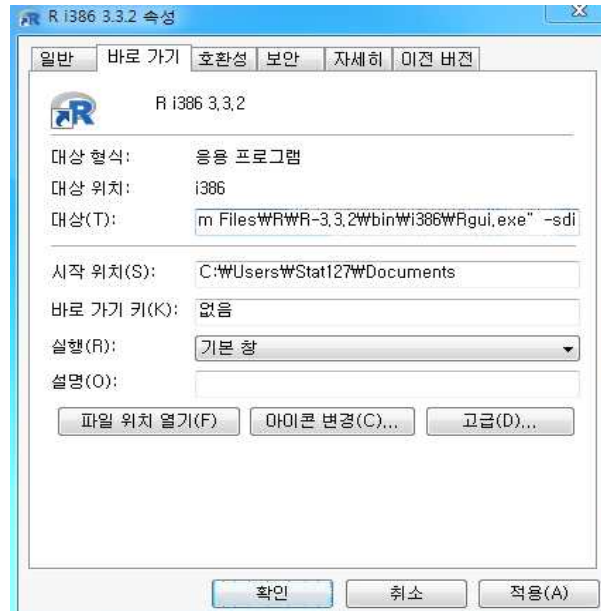


- (i) 위 주소에서 'Download R for Windows'를 선택하여 클릭하면 나타나는 화면에서 'base'를 누르면 최신 버전의 R을 설치하기 위한 'R-3.3.2-win.exe'과 같은 실행파일을 나타내는 화면이 나온다. 파일을 클릭하여 선택하면 저장 여부를 묻는 대화창이 열린다.
- (ii) '저장'을 클릭하여 자신의 컴퓨터에 위 프로그램파일을 저장한다.
- (iii) 자신의 컴퓨터에 저장된 설치 프로그램파일을 더블 클릭하여 설치 과정을 시작한다.
- (iv) 설치 프로그램이 제공하는 설치마법사에 따라 진행한다. 설치 과정 중에 구성 요소 설치를 묻는 창에서 'Message translations'를 선택해야 메뉴 표시가 한글로 나온다. 설치가 끝나면 자신의 컴퓨터 바탕화면에 R 아이콘 이 나타난다.

메뉴 표시를 한글에서 영문으로 변경하고 싶으면 R 아이콘을 우클릭하여 '속성'을 택하고 ['속성' -> '바로 가기' -> '대상']으로 가서 '대상(T)' 옆의 상자에 입력되어 있는 경로명 "C:\WProgram Files\WRWR-3.3.2\Wbin\Wi386\WRgui.exe"의 오른쪽에 한 칸을 띄어 추가로 'LANGUAGE=en'을 입력하고 '확인'을 누른 다음 다시 시작한다. 또는 R이 설치되어 있는 장소 중 "C:\WProgram Files\WRWR-3.3.2\Wetc"에 있는 'Rconsole'이라는 파일을 메모장 등에서 열어 보고 'Language for messages' 줄 아래에 'language = en'을 입력하고 저장해도 된다. Rconsole 파일의 수정을 막아 놓아서 안 될 경우 다음 절차에 따라 해제하도록 한다. Rconsole 파일을 우클릭하고, '속성'의 두 번째 메뉴에 있는 '보안' 탭을 누르면 나타나는 '그룹 또는 사용자 이름' 아래의 상자 밑의 '편집'을 클릭하여 'Rconsole의 사용 권한' 창을 띄운다. 가장 아래 'User (사용자이름\WUsers)'를 선택하면 나타나는 'Users의 사용 권한'의 '수정' 부분에 '허용'을 클릭하고 '확인'을 누르고 종료 후 재시작하면 될 것이다. R commander는 영문으로 개발되어 한글 메뉴를 사용할 때 오류가 종종 발생하며 그럴 경우 메뉴 표시를 영문으로 변경하면 오류가 없이 실행될 수 있다.

R 프로그램은 기본적으로 다중문서 인터페이스인 MDI(multiple document interface)로 설치되는데 이 환경에서 R commander를 실행하면 작업 창이 사라지는 등의 불안정성이 보고되어 단일문서 인터페이스인 SDI(single document interface)로 설정을 변경하여 사용하도록 권장되고 있다. 설정을 변경하여 사용하려면 바탕화면의 R 아이콘을 복사하여 하나 더 만든 뒤 복사된 아이콘을 우클릭하여 '속성'을 선택한다. ['속성' -> '바로 가기' -> '대상']으로 가서 [그림 A-2]에서와 같이 '대상(T)' 옆의 상자에 입력되어 있는 "C:\WProgram Files\WRWR-3.3.2\Wbin\Wi386\WRgui.exe"의 오른쪽에 한 칸을 띄고 '-sdi'를 추가하여 입력한 뒤 '확인'을 누른다. 원래 아이콘과 구별이 필요하면 복사된 아이콘을 우클릭하여 원하는 이름으로 변경한다.

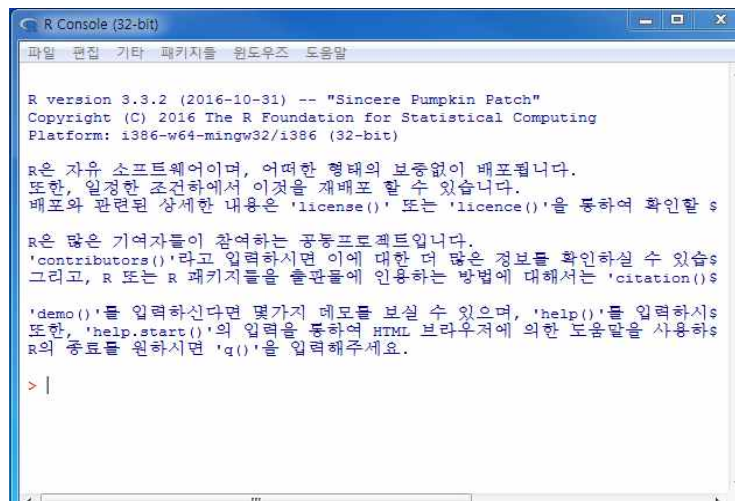
[그림 A-2]



## 1.1 R의 실행

R을 사용하기 위하여 R 아이콘을 클릭하면 [그림 A-3]과 같은 R 콘솔이 열린다.

[그림 A-3]



R이 시작되면 R 콘솔 창에 프롬프트라 불리는 기호 ‘>’가 나타난다. ‘>’ 다음에 명령문을 입력하고 ‘Enter’를 누르면 실행결과가 나타난다. 명령문이 완성되지 않은 상태에서 ‘Enter’를 누르면 연결 기호 ‘+’가 나타나며 이 상태에서 벗어나려면 ‘Esc’를 누른다. 이러한 대화형(interactive mode) 처리방식 외에 일정량의 명령문들을 묶어 일괄(batch mode) 처리하는 방식도 사용할 수 있다. R에서 일괄 처리방식을 위해 명령문들을 묶기 위해서 내부의 R 편집기 창이나 ‘source()’ 함수를 이용하거나 외부의 Windows batch 파일을 사용하는 방법이 있다.

R 콘솔의 메뉴 막대에서 [‘파일’ -> ‘새로운 스크립트’]를 선택할 때 나타나는 R 편집기 창에서 입력한 명령문들을 실행하려면 해당 명령문들을 마우스로 선택하여 ‘Ctrl+R’ 키를 누르면 되고, 한 줄만 실행하려 할 때는 그 줄에 커서를 놓고 ‘Ctrl+R’ 키를 눌러도 되며 입력한 명령문들을 모두 실행하려면 R 편집기 창의 메뉴 막대에서 [‘편집’ -> ‘전부실행’]을 눌러도 된다. R 편집기 창을 사용하여 입력한 명령문들로 만들어진 스크립트 파일을 저장하려면 메뉴 막대에서 [‘파일’ -> ‘저장’] 또는 [‘파일’ -> ‘다른 이름으로 저장’]을 선택하고 경로와 파일 이름을 지정하여 주면 된다.

함수 ‘source()’는 일반적인 문서 편집기를 사용하여 입력한 명령문들로 만들어진 스크립트 파일의 이름을 저장된 경로와 함께 지정하면 이 스크립트 파일에 입력된 명령문들을 모두 실행해주지만 각 명령문의 처리 결과를 R 콘솔 창에 나타내려면 함수 ‘print()’를 사용해야 하며, 스크립트 파일의 내용과 실행 결과를 모두 R 콘솔 창에 나타내려면 옵션 ‘echo=TRUE’를 추가해야 한다.

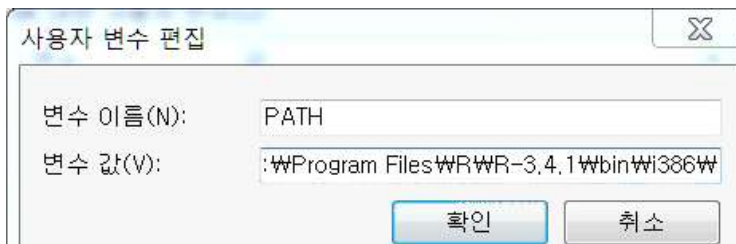
R에서 명령문을 실행할 때 콘솔 창이나 그래픽스 창에 나타나는 결과를 저장할 수 있다. 콘솔 창에 나타나는 결과를 저장하고자 할 때 사용할 수 있는 함수로 ‘sink()’가 있다. 저장할 파일의 이름(예를 들어 ‘out\_r.txt’)을 저장할 장소의 경로와 함께 지정하고 ‘sink(out\_r.txt)’를 실행한 다음 수행하려는 명령문들을 실행하면 결과들이 콘솔 창 대신 파일 ‘out\_r.txt’에 다시 ‘sink()’를 실행할 때까지 저장된다. R에서 명령문을 실행할 때 나타나는 결과는 객체가 되며 함수 ‘cat()’는 객체에 할당된 값을 화면에 나타내는 기능을 가진다. 화면에 나타내는 대신 파일로 저장하려면 저장할 파일의 이름(예를 들어 ‘out\_c.txt’)을 저장할 장소의 경로와 함께 선택사항 ‘file’로 지정한다. 함수 ‘cat()’를 계속 사용하여 ‘out\_c.txt’에 저장하려면 두 번째 ‘cat()’를 사용할 때부터 선택사항 ‘append=TRUE’를 사용해야 한다.

R에서 유의할 점은 대문자와 소문자를 구별한다는 것과 ‘#’ 기호 이후의 명령문은 처리하지 않기 때문에 주석을 달 때 사용할 수 있다는 것이다.

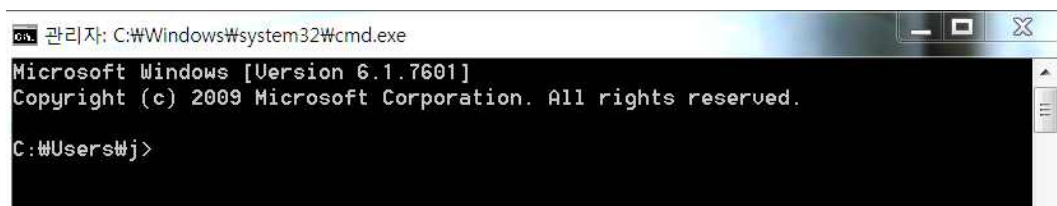
## 1.2 일괄처리

앞에서 알아본 R 내부에서 스크립트 파일을 실행시키는 방법이 주로 사용되지만, 시뮬레이션을 수행하는 경우 등과 같이 R 콘솔창을 열지 않고 외부에서 일괄처리하는 방식으로 실행시키는 것이 편리할 때가 있으며 그 절차는 다음과 같다.

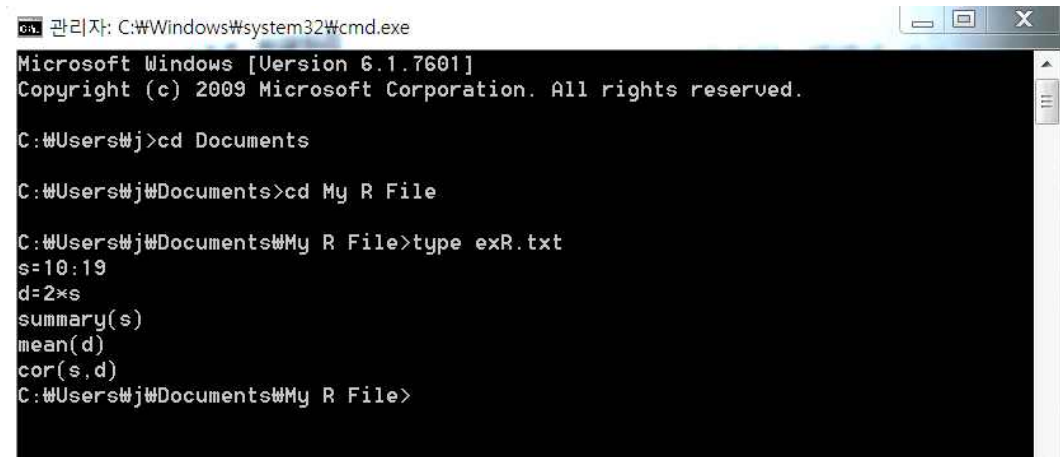
- i) R 실행을 위한 'R.exe'가 있는 폴더를 기존 경로(path)에 추가한다. 윈도우7에서 경로는 다음과 같이 추가할 수 있다. 윈도우의 '시작' 버튼을 클릭하고 '제어판', '시스템 및 보안', '시스템'을 차례로 선택한 뒤 '고급 시스템 설정'을 클릭하여 '시스템 속성' 창을 연다. '고급' 탭을 선택하고 하단에 있는 '환경변수'를 클릭하여 '환경변수' 창이 열리면 변수 'PATH'를 선택하고 '편집'을 클릭하여 '사용자 변수 편집' 창을 열고 '변수 값'에 'R.exe'가 있는 폴더를 기존 경로의 연결기호인 세미콜론(;) 뒤에 빈칸 없이 추가한다.



- 2) 윈도우 명령처리기 창을 활성화시킨다. 방법은 윈도우에서 '시작' 버튼을 클릭하면 나타나는 '프로그램 및 파일 검색' 창에 'cmd'를 입력하고 'Enter' 키를 누르면 된다.



- 3) 실행할 스크립트 파일이 있는 폴더로 이동한다. 예를 들어 실행할 스크립트 파일이 'C:\\Users\\Me\\Documents\\My R file'에 있는 'exR.txt'라 하면 명령어 'cd'를 사용하여 이동하고, 파일의 내용은 명령어 'type'으로 확인할 수 있다.



```

관리자: C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Wj>cd Documents

C:\Users\Wj\Documents>cd My R File

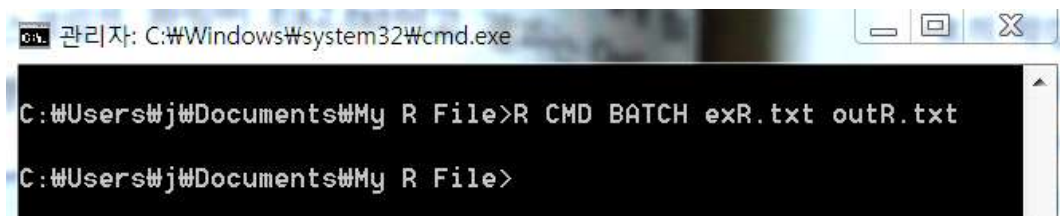
C:\Users\Wj\Documents\My R File>type exR.txt
s=10:19
d=2*s
summary(s)
mean(d)
cor(s,d)
C:\Users\Wj\Documents\My R File>

```

4) R을 일괄처리방식으로 실행하는 기본 방식은 다음과 같다.

R CMD BATCH infile outfile

여기서 infile은 실행할 스크립트 파일을 나타내고, outfile은 실행 결과가 저장되는 파일을 나타낸다. 스크립트 파일 'exR.txt'의 실행 결과를 outR.txt에 저장하려면 다음과 같이 실행한다.



```

관리자: C:\Windows\system32\cmd.exe

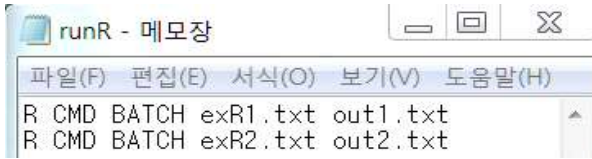
C:\Users\Wj\Documents\My R File>R CMD BATCH exR.txt outR.txt

C:\Users\Wj\Documents\My R File>

```

5) 여러 개의 스크립트 파일을 순차적으로 실행하려면 다음과 같이 한다.

가) 실행 명령문을 확장자 '.bat'인 파일에 차례로 입력한다. 예를 들어 'runR.bat' 파일에 아래와 같이 명령문을 입력한다.



나) 명령처리기 창에서 파일 'runR.bat'를 실행한다. 확장자 '.bat'를 생략하고 'runR'만 입력하고 'Enter' 키를 눌러도 된다.



## 2. R commander의 설치와 실행

R 콘솔에서 다음과 같이 R commander 설치를 위한 명령문을 입력한 다음 'Enter'를 누를 때 나타나는 'CRAN mirror'라는 창의 리스트에서 'Korea'를 선택하면 되는데 첫 리스트에 보이지 않을 경우는 '(HTTP mirrors)'를 선택하면 확장된 리스트에 보이게 되고 'OK'를 누르면 필요한 패키지들을 설치하느라 시간이 좀 걸려 완료된다.

```
install.packages("Rcmdr", dependencies=TRUE)
```

이렇게 해서 설치가 되지 않으면 <http://cran.r-project.org/> 에서 필요한 R 패키지들을 내려 받아 설치할 수 있겠으나 권하지 않는다.

R commander를 시작하기 위하여 R 아이콘을 더블 클릭하여 나타나는 R 콘솔에서 다음 명령문을 입력한다.

```
> library(Rcmdr)
```

이 때, R commander창이 [그림 A-4]처럼 나타나고 사용할 수 있게 된다. R을 시작할 때 마다 위 명령문을 입력하지 않아도 자동으로 R commander창이 나타날 수 있게 하려면 R이 설치되어 있는 장소 중 "C:\Program Files\WRWR-3.3.2\etc"에



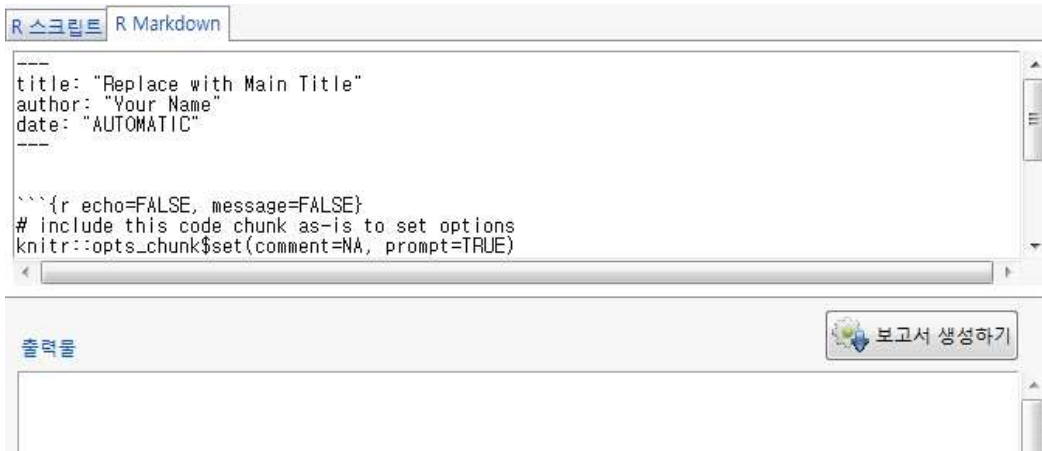
있는 Rprofile.site 파일을 열어 다음 문장을 추가 입력한 후 저장하고 R을 다시 시작하면 된다.

```
local({
old <- getOption("defaultPackages")
options(defaultPackages=c(old, "Rcmdr"))
})
```

제대로 안 되는 경우는 Rprofile.site 파일의 수정을 막아 놓아서일 수 있으므로 앞에서와 같이 다음 절차에 따라 해제하도록 한다. Rprofile.site 파일을 우클릭하고, ‘속성’의 두 번째 메뉴에 있는 ‘보안’ 탭을 누르면 나타나는 ‘그룹 또는 사용자 이름’ 아래의 상자 밑의 ‘편집’을 클릭하여 ‘Rprofile.site의 사용 권한’ 창을 띄워 가장 아래 ‘User (사용자이름\Users)’를 선택하면 나타나는 ‘Users의 사용 권한’의 ‘수정’ 부분에 ‘허용’을 클릭하고 ‘확인’을 누르고 종료 후 재시작한다.

[그림 A-4]





### 3. R commander의 구성

R commander는 [그림 A-4]에서와 같이 ‘스크립트’, ‘출력물’, ‘알림글’ 순의 세 가지 창과 여러 가지 메뉴 또는 버튼 등으로 구성되어 있다. 가장 위에 메뉴 막대가 있고 바로 아래의 도구막대에 현재 사용 중인 활성 데이터셋을 나타내는 (상자)칸과 ‘데이터셋 편집하기’와 ‘데이터셋 보기’ 버튼 그리고 활성 모델을 나타내는 칸이 있다. 그 아래에 ‘스크립트’ 창이 있는데 메뉴 막대에 있는 메뉴를 선택하여 실행할 때 생성되는 R 프로그램 문장을 보여주며, R 문장을 직접 입력할 수도 있는 곳이다. ‘스크립트’ 창 아래에 있는 ‘출력물’ 창에는 메뉴를 선택하여 실행할 때 수행되는 R 문장이 빨간색으로 실행 결과는 파란색으로 나타난다. 또한 R 문장을 ‘스크립트’ 창에 직접 입력한 경우 실행하고자 하는 부분을 마우스 드래그나 키보드의 쉬프트 키와 화살표 등을 이용하여 선택한 뒤 ‘실행하기’ 버튼을 누르거나 키보드의 ‘Ctrl’과 ‘R’을 동시에 누르면 그 결과가 ‘출력물’ 창에 나타난다. ‘알림글’ 창에는 오류, 경고, 주석 등의 메시지가 주어질 수 있다.

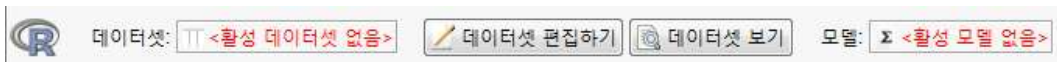
R commander 창의 위쪽에 있는 메뉴 막대는 다음과 같은 메뉴를 갖고 있으며 각 메뉴의 역할을 간단히 알아보도록 한다.

파일 편집하기 데이터 통계 그래프 모델 분포도 도구 도움말

- 1) 파일: 스크립트 파일을 열거나 저장하고 결과 창 그리고 작업 영역을 저장하는 메뉴 항목들을 포함한다. ‘작업 디렉토리 바꾸기’는 함수 ‘setwd()’를 수행하며 작업자가 사용하는 폴더를 지정할 수 있도록 한다. 작업을 여러 개 한다면 각 작업마다 미리 분석 결과를 저장할 곳을 다르게 지정하는 것이 편리할 것이므로 항상 새 작업을 시작할 때마다 가장 먼저 ‘작업 디렉토리 바꾸기’를 수행할 것을 권장한다. 현재의 작업 디렉토리는 함수 ‘getwd()’를 실행하면 알 수 있다. ‘스크립트 파일 열기...’는 저장되어 있는 R 프로그램 등의 스크립트를 불러 온다. ‘스크립트 파일 저장...’은 함수 ‘source()’를 수행하며 작성한 스크립트를 저장하는 부메뉴이다. 메뉴 방식으로 수행된 R 프로그램의 스크립트를 저장하여 재활용할 수 있도록 해주며 익숙해지면 커맨드 형식의 R 프로그램을 수행할 때 도움이 된다. 저장하려는 스크립트의 프로그램에 대하여 ‘#’로 시작하는 문장을 첨가하여 설명해두는 것이 재활용할 때 좋다. ‘출력물 저장하기...’는 ‘출력물’ 창의 결과물을 저장하는 부메뉴이다. 결과물 양이 많을 때 외부파일로 저장하여 확인하는 것이 편리하다. ‘R 작업 공간 저장하기...’는 함수 ‘save.image()’를 수행하며 작업자가 다룬 여러 데이터셋을 확장자가 RData인 하나의 파일로 작업공간에 저장하는 부메뉴이다. 함수 ‘load()’를 사용하여 저장된 작업공간을 불러올 수 있다.
- 2) 편집하기: ‘스크립트’ 창이나 ‘출력물’ 창에서 편집하기 위한 자르기, 복사, 붙이기 등과 같은 메뉴 항목들을 포함하며, ‘스크립트’ 창이나 ‘출력물’ 창에서 우클릭 하여도 편집 메뉴가 나타난다.
- 3) 데이터: 자료를 읽고 다루기 위한 메뉴 항목들을 포함하는 부메뉴들이 있다. 다음 절에서 추가로 설명한다.
- 4) 통계: 여러 가지 기본적인 통계분석들에 대한 메뉴 항목들을 포함하는 부메뉴들이 있다.
- 5) 그래프: 간단한 통계 그래프를 생성하는 메뉴 항목들을 포함한다.
- 6) 모델: 모형에 대한 수치적 요약, 신뢰구간, 가설검정, 그래프 등을 얻기 위한 메뉴 항목들을 포함한다.
- 7) 분포도: 이산형이나 연속형인 표준적인 통계분포의 확률, 분위수, 그래프 등을 얻기 위한 메뉴 항목들을 포함한다. 통계표를 대신하여 사용될 수 있다.

- 8) 도구: Rcmdr 패키지와 관련되지 않은 다른 패키지를 불러오거나 R commander의 선택사항을 정하는 메뉴 항목들을 포함한다. 각 분석에서 필요로 하는 패키지를 선택하여 설치해서 사용할 수 있다. ‘선택기능’에서 R commander의 창의 크기, 색 등을 지정한다.
- 9) 도움말: R commander에 대한 정보를 얻을 수 있는 메뉴 항목들을 포함한다. 각 R commander의 대화 상자는 도움말 버튼을 갖는다.

메뉴 막대 아래의 도구막대에는 다음과 같은 4개의 버튼이 있으며 각 버튼의 역할을 알아보자.



- 1) 첫 번째 ‘데이터셋’ 버튼은 사용 중인 활성 데이터셋의 이름을 나타낸다. 처음에는 활성 데이터셋이 없다. 이 버튼을 누르면 ‘데이터셋 선택하기’ 창이 뜨며 ‘데이터셋 (하나 선택)’ 아래 나타나는 현재 메모리에 있는 데이터셋들 중에서 고를 수 있게 된다.
- 2) ‘데이터셋 편집’ 버튼은 사용 중인 데이터셋을 수정하기 위한 자료 에디터를 열게 해준다. 현재 메모리에 있는 데이터셋들 중에서 고르기 위하여 이 버튼을 사용할 수 있다.
- 3) ‘데이터셋 보기’ 버튼은 데이터셋을 검토할 수 있게 해준다. ‘데이터셋 보기’를 클릭하여 띄운 현재 사용 중인 데이터셋 창은 다른 작업이 수행되는 동안에도 열려 있을 수 있다.
- 4) ‘모델’ 버튼은 사용 중인 선형회귀모형이나 일반화선형모형 등과 같은 통계모형의 이름을 나타낸다. 현재 메모리에 있는 서로 다른 모형들 중에서 고르기 위하여 이 버튼을 사용할 수 있다.

이 도구막대 아래쪽에 ‘R 스크립트’ 탭과 ‘R Markdown’ 탭이 있다. 두 탭 아래에 ‘스크립트’ 창이 있으며, 메뉴 방식의 GUI에 의해 생성된 명령문들이 이 창 안으로 누적하여 복사된다. ‘스크립트’ 창의 문장을 편집하거나 직접 R 명령문들을 이 창 안으로 입력할 수도 있다. ‘스크립트’ 창 우측 하단의 ‘실행하기’ 버튼을 누르면 커서가

포함된 줄의 문장이 실행된다. ‘R 스크립트’ 탭 우측의 ‘R Markdown’ 탭은 문서에 R 결과물이나 그래프 등을 복사하여 붙이던 일을 간단한 마크업 언어 Markdown으로 작업 결과를 편집할 수 있는 동적 문서로 생성하여 보고서로 출력할 수 있게 해준다. 만약 Pandoc과 같은 보조 소프트웨어가 설치되어 있다면 ‘R Markdown’ 문서를 출력형식 RTF, DOCX, HTML, PDF 등으로 변환하는 창에서 선택할 수 있겠지만 그렇지 않을 경우 HTML로 변환된다.

‘스크립트’ 창 아래쪽이 ‘출력물’ 창이다. 이 창에 수행되는 명령문은 빨간색으로 나타나고 실행 결과는 짙은 파란색으로 나타난다.

R commander 창의 하단에 메시지 ‘알림글’ 창이 있다. 오류 메시지는 빨간색이고, 경고 메시지는 초록색이며 다른 메시지들은 짙은 파란색으로 나타난다.

#### 4. R commander를 통한 자료 입력과 저장

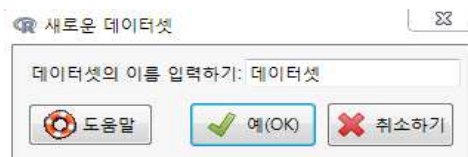
통계적 자료분석을 위하여 R commander를 통하여 R에서 사용할 수 있는 데이터 프레임과 같이 직사각형 자료형태인 데이터셋으로 입력하는 방법은 다음과 같다. 데이터 프레임은 행렬과 같은 직사각형이나 행렬이 같은 자료형태로만 구성된 반면 각 열들이 서로 다른 형태로도 구성된다는 점이 다르다. 파일로 저장하는 곳의 경로명은 영문으로 하는 것이 오류 발생 가능성을 줄여 준다.

- (i) 메뉴 막대에서 ‘데이터’를 클릭하고 [‘데이터’ -> ‘새로운 데이터셋...’]을 선택하여 나타나는 편집기에서 자료를 직접 입력한다. 이 방법은 아주 작은 규모의 단순한 구조의 데이터셋에 대해서는 사용할만하다.
- (ii) 아스키 파일과 같은 보통 문장으로 되어 있는 파일로부터 자료를 선택하거나 Excel, SPSS, SAS, Minitab, STATA와 같은 다른 통계처리 패키지에 들어 있는 파일로부터 자료를 선택하여 입력한다.
- (iii) R 패키지에 포함되어 있는 데이터셋은 이름을 알고 있으면 이름을 입력하거나 아니면 대화상자에 나열되어 있는 데이터셋을 선택하여 입력한다.

## 4.1 자료 편집기(Data Editor) 이용

R commander에서는 Excel과 같이 스프레드시트에 자료를 직접 입력하여 데이터셋을 만들어 사용할 수 있다. 예를 들어 두 변수 농도(conc), 속도(rate)에 대한 8개 관측치로 구성된 [표 1.1]의 자료를 입력하고 파일로 저장한 뒤 다시 불러들여 보자.

우선 R commander 메뉴의 ['데이터' -> '새로운 데이터셋...']을 선택하면 다음과 같은 '새로운 데이터셋' 창이 열린다.



'데이터셋의 이름 입력하기' 옆의 칸에 새 데이터셋에 대한 이름, 예를 들어 'MM1'을 입력하고 '예(OK)'를 누르면 다음과 같이 '데이터 편집기'가 나타난다.



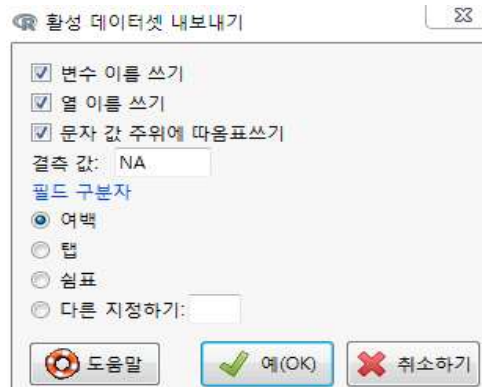
'데이터 편집기'의 '행(row) 추가하기'와 '열(column) 추가하기' 버튼을 눌러 관측치 또는 레코드(표본, 케이스 등)을 나타내는 행은 8개, 변수 또는 항목을 나타내는 열은 2개를 만든 뒤 1열부터 자료를 입력하고 상단의 이름 v1을 클릭하여 변수 이름을 'conc'로 바꾼다. 마찬가지로 2열에 자료를 입력하고 상단 이름 v2를 'rate'로 바꾼 다음, 현재 작업공간에 저장하고 '데이터 편집기'를 끝내기 위하여 ['파일' -> 'Exit and Save']를 선택한다. 작업자가 입력한 데이터셋과 R이 인식하고 있는 데이터셋이 같은지 확인하기 위하여 '데이터셋 보기' 버튼을 누른다. 확인된 자료가 R commander에서 사용 가능한 활성 데이터셋이 된다.

입력된 자료를 저장하는 방법으로 각 데이터셋을 각각 하나의 파일로 저장하는 데이터셋방식, 그리고 여러 데이터셋을 하나의 파일로 저장하는 작업공간방식을 생각할 수 있다.

먼저 데이터셋방식으로 현재 사용 중인 활성 데이터셋을 파일로 저장하기 위해서는 [‘데이터’ -> ‘활성 데이터셋’ -> ‘활성 데이터셋 저장하기...’]를 선택하고, 저장할 위치를 묻는 ‘다른 이름으로 저장’ 대화상자가 나타나면 원하는 위치의 폴더에 저장할 파일의 이름을 정하여 입력하고 ‘저장’을 누른다. 저장된 파일을 다시 불러오려면 메뉴 막대에서 [데이터 -> 데이터셋 탑재하기]를 선택하여 불러올 파일의 위치를 묻는 ‘열기’ 대화상자가 나타날 때 저장된 파일의 이름을 지정해주면 ‘출력물’ 창에 ‘load’ 문장이 나타나고 ‘알림글’ 창에 로드된 데이터셋에 대한 주석이 나타난다. 불러오고자 한 파일이 맞는지 확인하고 싶으면 ‘데이터셋 보기’ 버튼을 누른다.

만약 현재 사용 중인 활성 데이터셋을 다른 소프트웨어에서 사용할 수 있게 하기 위하여 지정된 필드 구분자에 의한 텍스트 형식의 파일로 외부로 내보내려면 메뉴 막대에서 [‘데이터’ -> ‘활성 데이터셋’ -> ‘활성 데이터셋 내보내기...’]를 선택한다. [그림 A-5]와 같은 ‘활성 데이터셋 내보내기’ 대화상자가 나타난다.

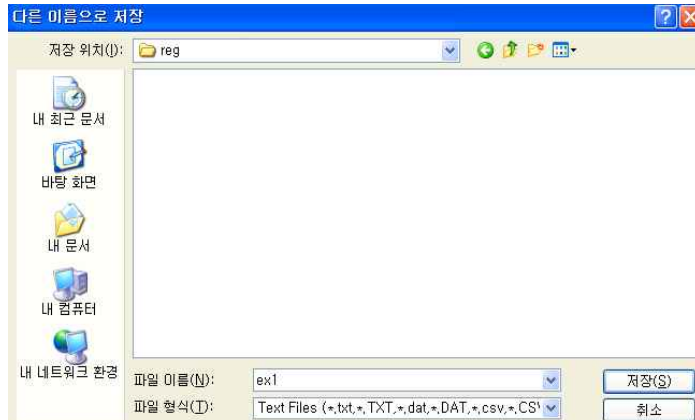
[그림 A-5]



자료 형태에 대한 선택항목들을 택한 뒤 ‘예(OK)’를 클릭하면 [그림 A-6]의 ‘다른 이름으로 저장’ 대화상자가 나타난다. ‘열 이름 쓰기’ 항목을 선택하면 저장되는 파일의 첫 열에 관측치 번호를 입력하는데 그 열을 나타내는 변수 이름이 지정되어 있지 않아 다시 불러 오려면 변수 이름을 열의 상단에 추가하는 수정이 필요할 수 있음을 유의해야 한다. 이 항목의 ‘열’은 ‘행’의 오류로 보인다. ‘필드 구분자’에서 ‘쉼표’를 선택하면 csv(comma seperated volume) 형식의 파일로 저장되므로 불러오기 등을 할 때 파일형식에 유의하여 사용해야 한다. 만약 저장할 파일 이름을 ‘MM1’이 아닌 ‘ex1’과 같은 다른 이름으로 바꾸려면 원하는 이름을 입력하고 원하는 장소의 경로를

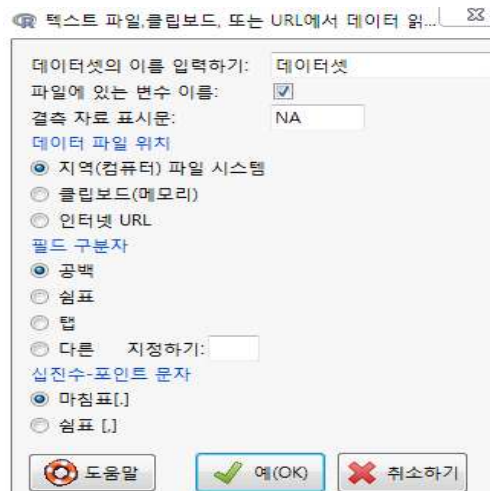
찾아가서 ‘저장’을 클릭한다. 저장 장소의 폴더 이름은 한글보다는 영문으로 되어 있는 곳이 오류가 발생할 가능성을 줄인다.

[그림 A-6]



텍스트 형식의 파일로 저장된 자료를 다시 불러오려면 [‘데이터’ -> ‘데이터 불러오기’ -> ‘텍스트 파일, 클립보드, 또는 URL에서...’]를 선택하면 [그림 A-7]과 같은 ‘텍스트 파일, 클립보드, 또는 URL에서 데이터 읽...’ 대화상자가 나타난다.

[그림 A-7]



불러올 파일의 자료로 만들어질 데이터셋의 이름과 선택사항들을 지정하고 ‘예 (OK)’를 클릭하면 ‘열기’ 대화상자가 나타나고, 불러올 파일을 포함하는 장소를 찾아



가서 그 파일의 이름을 클릭하고 ‘열기’를 누르면 사용 가능하게 된다.

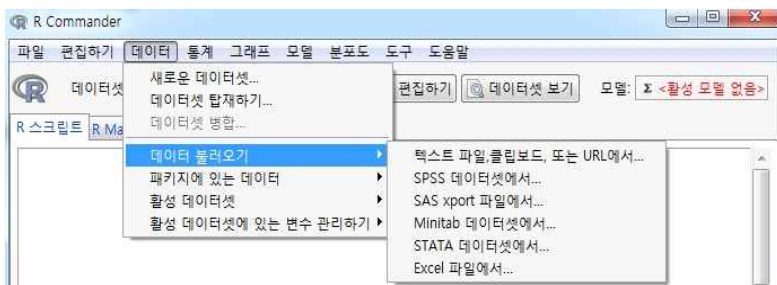
한편 작업공간방식으로 저장하기 위해서는 메뉴 막대에서 [‘파일’ -> ‘R 작업공간 저장하기’]를 선택하여 ‘다른 이름으로 저장’ 대화상자가 나타나면 원하는 장소의 폴더에 파일 이름을 정하여 입력한 뒤 ‘저장’을 누르면 된다. 저장해 놓으면 R commander를 끝냈다가 다시 시작하여도 사용할 수 있어 편리하다. R commander를 끝내기 위해 [‘파일’ -> ‘끝내기’]를 선택하거나 끝내기 단추를 누르면 스크립트와 출력 파일 등의 저장여부를 묻는 창이 열리니 필요에 따라 선택한다.

R commander를 끝냈다가 다시 시작하는 경우 저장된 데이터셋을 불러오려면 메뉴 막대에서 [‘데이터’ -> ‘데이터셋 적재하기...’]를 클릭하면 나타나는 ‘열기’ 창에서 저장된 장소로 찾아가 해당 파일을 택하고 ‘열기’를 누르면 된다. 작업공간방식으로 저장했던 경우에는 한 파일에 여러 개 데이터셋이 들어 있다는 오류 메시지가 ‘알림 글’에 나올 수 있는데 메뉴 막대 아래 ‘데이터셋’ 옆의 칸을 클릭하여 ‘데이터셋 선택하기’ 창에 저장당시의 모든 데이터셋들이 목록에 나타나면 필요한 데이터셋을 선택하여 사용한다. 확인을 하고 싶으면 메뉴 막대에서 [‘데이터’ -> ‘활성 데이터셋’ -> ‘활성 데이터셋 선택하기...’]를 클릭하면 선택 대상인 데이터셋의 목록이 나타난다.

## 4.2 데이터 불러오기 이용

데이터셋을 만들기 위하여 불러올 수 있는 외부 파일의 형식은 막대 메뉴의 [‘데이터’ -> ‘데이터 불러오기’]를 선택할 때 나타나는 [그림 A-8]의 부메뉴에서와 같이 텍스트 형식의 파일, SPSS 데이터셋, SAS xport 파일, Minitab 데이터셋, STATA 데이터셋 그리고 Excel 파일 등이다. 실제로 텍스트 형식의 파일 외에는 R commander에서 사용하려 할 때 여러 문제가 발생할 수 있어 해당 소프트웨어에서 텍스트 형식으로 변환해서 저장한 파일을 불러오기 하는 방법을 추천한다. 필요하다면 R에서 데이터베이스 관련 라이브러리 ‘RODBC’ 등을 사용하는 방법을 택할 수도 있을 것이다.

[그림 A-8]



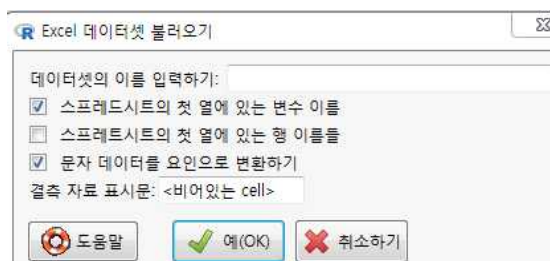
## 1) 텍스트 형식의 파일 불러오기

이 경우는 앞에서 ‘활성 데이터셋 내보내기...’에 의해 텍스트 형식으로 저장된 파일을 다시 불러오는 절차를 설명할 때 다룬 바 있다. 즉 [‘데이터’ -> ‘데이터 불러오기’ -> ‘텍스트 파일, 클립보드, 또는 URL에서...’]를 선택하여 [그림 A-7]과 같은 ‘텍스트 파일, 클립보드, 또는 URL에서 데이터 읽...’ 대화상자가 나타나면 불러올 파일의 자료로 만들어질 데이터셋의 이름을 지정하고 파일의 형식에 맞게 선택사항들을 정하고 ‘예(OK)’를 클릭한다. 데이터셋의 이름을 지정하지 않으면 디폴트인 ‘데이터셋’이란 이름으로 중복되는 등 혼란이 발생하게 될 것이다. 파일의 확장자가 텍스트 형식일 때는 필드 구분자를 ‘공백’과 ‘탭’ 중 어느 것을 사용했는지에 따라 선택하고, csv 형식일 때는 ‘쉼표’를 선택한다. ‘열기’ 대화상자가 나타나면, 불러올 파일이 있는 장소를 찾아가서 그 파일의 이름을 클릭하고 ‘열기’를 누르면 활성 데이터셋으로 사용 가능하게 된다.

## 2) Excel 파일 불러오기

[‘데이터’ -> ‘데이터 불러오기’ -> ‘Excel 파일에서...’]를 선택하여 [그림 A-9]와 같은 ‘Excel 데이터셋 불러오기’ 대화상자가 나타나면 불러올 파일의 자료로 만들어질 데이터셋의 이름을 지정하고 행이나 열 이름이 있다면 형식에 맞게 선택사항들을 정하고 ‘예(OK)’를 클릭한다. ‘열기’ 대화상자가 나타나면 불러올 파일이 있는 장소를 찾아가서 그 파일의 이름을 클릭하고 ‘열기’를 누르면 ‘표 선택하기’ 창이 나오므로 필요한 Sheet를 선택하면 활성 데이터셋으로 사용 가능하게 된다. 이때도 마찬가지로 파일의 이름이나 위치가 영문으로 되어 있어야 오류 발생 가능성이 적다.

[그림 A-9]



간혹 설치된 윈도우스와 R 버전이 32비트 또는 64비트인가에 따라 사용하는 라이브러리가 달라 오류가 날 수 있다. 필요할 경우 ‘XLConnect’ 라이브러리를 설치하기

위하여 `install.packages("XLConnect")`를 실행한다. 만약 JAVA 버전이 달라 'XLConnectJars' 라이브러리 오류가 나면 32비트와 64비트 중에서 맞는 JAVA 버전을 <http://www.java.com/en/download/manual.jsp>에서 찾아 설치하면 해결될 것이다.

또한 메뉴가 한글일 때 오류가 발생하는 경우도 있으며 그럴 때는 R Console에서 ['편집' -> 'GUI 설정...']을 선택하면 나타나는 'Rgui 구성 편집기' 창의 메뉴들 중에서 'Language for menus and messages'에 'en'을 입력하여 영어로 변경하면 해결되기도 한다.

기타 다른 소프트웨어의 파일이나 데이터셋을 불러오기 하는 것도 비슷한 절차를 거치겠으나 소프트웨어 버전이나 한글 충돌 등에 따른 여러 가지 오류가 발생할 가능성이 많아 안정적이지 못하다. 이러한 문제점을 가진 기존의 데이터 불러들이기 라이브러리인 'foreign'을 보완하는 'memisc', 'sas7bdat', 'readstata13', 'Hmisc' 등의 라이브러리가 제공되고 있어 `install.packages(c("foreign", "sas7bdat", "memisc", "readstata13", "Hmisc"))`와 같이 설치하여 사용한다. 기존 'foreign'으로도 안정적으로 불러오기가 되는 것은 Stata 버전 12 이하인 파일들인 것으로 알려져 있다.

특히 `read.spss` 관련 오류가 많이 보고되고 있으며, 그럴 때는 다음과 같이 'memisc' 라이브러리를 이용하여 'sample.sav'라는 SPSS 데이터셋을 'spss2rdata'라는 데이터셋으로 만든 다음 데이터프레임으로 변환하여 사용하는 방법을 택한다.

```
library(memisc)
spss2rdata <- as.data.set(spss.system.file("sample.sav"))
spss2rdata <- as.data.frame(spss2rdata)
```

좀 더 호환성이 좋다는 `por` 형식의 'sample.por'라는 SPSS 파일을 불러오기 해서 'por2rdata'라는 데이터셋으로 만드는 방법은 다음과 같다.

```
library(memisc)
xpt2rdata <- as.data.set(spss.portable.file("sample.por"))
xpt2rdata <- as.data.frame(xpt2rdata)
```

긴 이름 형식의 'sample.sas7bdat'라는 SAS 파일을 불러오기 해서 'sas2rdata'라는 데이터셋으로 만드는 방법은 다음과 같다.

```
library(sas7bdat)
```

```
sas2rdata <- read.sas7bdat("sample.sas7bdat")
```

좀 더 호환성이 좋다는 xpt 형식인 'sample.xpt'라는 SAS Transport(전송)파일을 불러오기 해서 'xpt2rdata'라는 데이터셋으로 만드는 방법은 다음과 같다.

```
library(Hmisc)
xpt2rdata <- sasxport.get("sample.xpt")
```

버전 13 이상인 STATA 파일 'sample.dta'를 불러오기 해서 'stata2rdata'라는 데이터셋으로 만드는 방법은 다음과 같다.

```
library(sas7bdat)
stata2rdata <- read.dta("sample.dta")
```

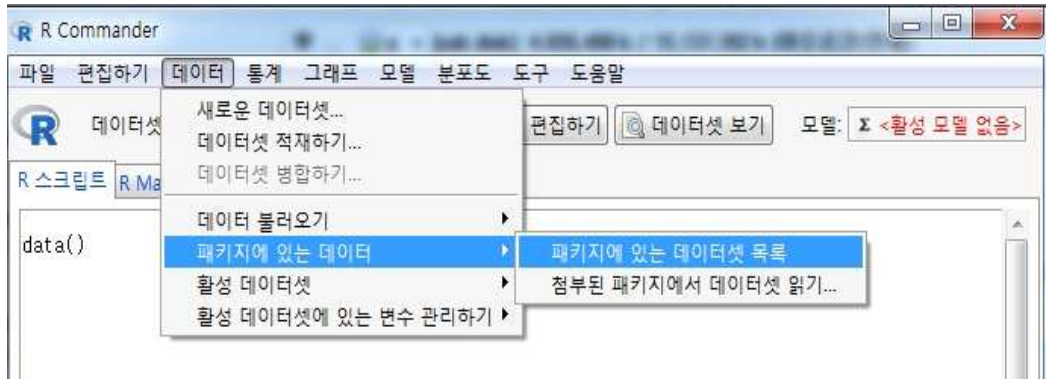
또한, dBase 형식인 파일 'sample.dbf'을 불러오기 해서 'dbf2rdata'라는 데이터셋으로 만드는 방법은 다음과 같다.

```
library(foreign)
dbf2rdata <- read.dbf("sample.dbf")
```

## 4.3 패키지에 있는 데이터

메뉴 막대의 '데이터'를 선택하고 [그림 A-10]에서와 같이 ['데이터' -> '패키지에 있는 데이터' -> '패키지에 있는 데이터셋 목록']을 클릭하면 R에 내장되어 있는 패키지들의 안에 들어 있는 데이터셋들이 간단한 설명과 함께 나열된 'R data sets' 창이 나타난다. 패키지는 사용자들이 작성한 모듈로 R 함수, 데이터 및 컴파일된 코드 등을 모은 것이며 다양한 분야의 분석도구를 제공하고 있다. 패키지는 사용자 컴퓨터의 R 프로그램의 'library' 폴더에 저장되어 있고 설치할 때 자동으로 저장되어 작동되는 것으로는 'base, datasets, graphics, grid, methods, stats, utils'등이 있으며 'base' 패키지라 한다. 자동으로 저장되기는 하지만 사용하려면 R 세션으로 불러들여야 하는 것으로는 'MASS, foreign, lattice'등이 있으며 'recommended' 패키지라 한다. 자동 저장되지 않은 다른 패키지들은 개별적으로 저장하여 R 세션으로 불러들여야 사용할 수 있으며 개별 사용자들이 직접 작성한 것이어서 간혹 오류가 있는 것도 있을 수 있다는 점을 유의해야 한다.

[그림 A-10]



‘R data sets’ 창 내의 가장 위에는 R commander를 개발한 John Fox가 회귀분석에 관련하여 저술한 책 ‘An R Companion to Applied Regression’에서 사용한 패키지 ‘car: Companion to Applied Regression’안의 데이터셋에 대한 설명이 있고, 그 아래에 여러 다른 패키지들과 그 안의 데이터셋들에 대한 설명이 있다. [‘데이터’ -> ‘패키지에 있는 데이터’ -> ‘첨부된 패키지에서 데이터셋 읽기...’]를 선택할 때 나타나는 ‘패키지로부터 데이터 읽기’ 창의 ‘패키지’와 ‘데이터셋’ 항목들을 지정하고, 만약 선택된 데이터셋의 이름을 변경하기 원하면 ‘데이터셋의 이름 변경하기’ 옆의 상자에 원하는 이름을 입력하여 필요한 자료들을 불러들일 수 있다. 다음 [그림 A-11]은 ‘car’ 패키지의 207개 국가에 대한 국내총생산(gdp)과 영아사망률(infant.mortality)에 관한 자료인 데이터셋 ‘UN’을 불러와 활성 데이터셋으로 사용하는 예를 보여 준다.

[그림 A-11]



## 4.4 기타 데이터 입력 방법

메뉴에 의한 방법 외에 R의 함수

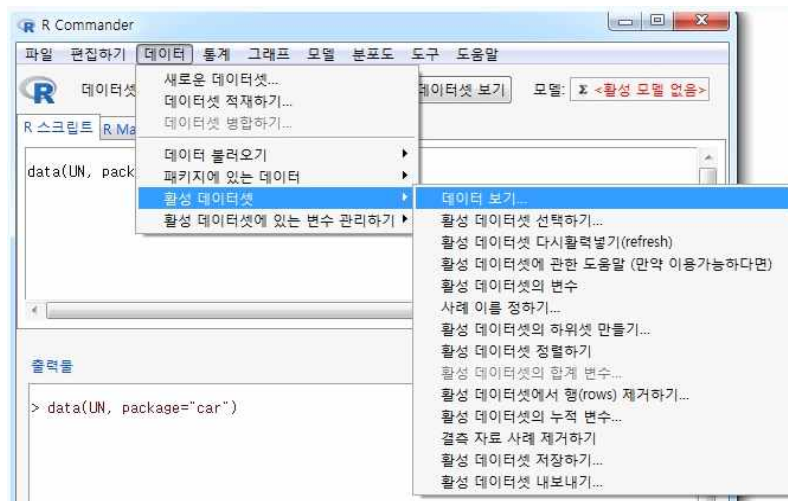
## 5. 활성 데이터셋과 변수

앞의 4절에서는 자료의 입력이나 불러오기와 데이터셋 만들기에 대해 다뤘으며 이 절에서는 만들어진 활성 데이터셋과 변수들을 관리하는 방법에 대해 알아본다.

### 5.1 활성 데이터셋의 관리

메뉴 막대의 ‘데이터’를 클릭하고 [‘데이터’ -> ‘활성 데이터셋’]을 선택하면 활성 데이터셋을 관리하기 위한 부메뉴들이 [그림 A-12]에서와 같이 나타난다.

[그림 A-12]



‘데이터 보기...’ 부메뉴에 의해 각 활성 데이터셋의 변수들과 관측값을 볼 수 있고, 현재 활성 데이터셋이 여러 개이면 ‘활성 데이터셋 선택하기...’를 사용하여 선택한다.

‘활성 데이터셋 다시활력넣기(refresh)’는 선택한 활성 데이터셋에 변경이 있을 때 클릭하여 사용한다.

‘활성 데이터셋의 변수’를 선택하면 현재 사용 중인 변수들의 이름(여기서는

"infant.mortality", "gdp")이 ‘출력물’ 창에 나타나며 ‘사례 이름 정하기...’는 행 이름을 나타내는 변수가 있을 때 사용한다.

‘활성 데이터셋의 하위셋 만들기...’를 선택하면 하위셋을 만들기 위한 다음과 같은 ‘데이터셋 하위설정하기’ 창이 열린다.



하위셋에 포함할 변수들을 정하고, 포함하려는 사례들을 선택하기 위하여 ‘하위셋 표현식’ 아래의 칸에 필요한 조건(‘==’, ‘!=’, ‘&’, ‘|’, ‘<’, ‘>’, ‘<=’, ‘>=’, ‘is.na(“변수”)’, ‘!is.na(“변수”)’ 등)을 입력해야 한다. 여기서는 ‘gdp > 1000’을 예로 사용한다. ‘새로운 데이터셋 이름’ 아래 칸에 원하는 이름(여기서는 ‘UN\_a’)을 입력하고 ‘예(OK)’를 누르면 전체 207개 국가 중 조건에 맞는 123개 국가에 대한 ‘UN\_a’라는 하위 데이터셋이 만들어진다.

‘활성 데이터셋 정렬하기’를 선택하여 선택사항에 따라 활성 데이터셋을 정렬할 수 있다.

‘활성 데이터셋의 합계 변수...’를 선택하면 나오는 ‘관찰치 합계하기’ 창에서 ‘합계 데이터셋 이름’ 옆 칸에 원하는 이름(여기서는 ‘AggregatedData1’)을 정하여 입력하고, ‘합계할 변수(하나 이상 선택)’ 아래 나열된 변수들 중 집단화하여 계산하려는 것들(여기서는 "infant.lmortality")을 고르고 ‘Aggregated by(하나 이상 선택)’ 아래 나열된 변수들 중 묶는 기준이 되는 요인들(여기서는 "gdp\_f")을 선택한 뒤 계산하기 원하는 ‘통계량’ 사양(여기서는 ‘평균’)을 지정한 뒤 ‘예(OK)’를 누르면 지정한 이름의 합계 데이터셋(여기서는 ‘AggregatedData1’)이 만들어진다.



‘활성 데이터셋에서 행(rows) 제거하기’ 창에서 제거할 사례들의 색인(행의 번호)을 지정하거나 또는 문자로 된 행(row) 이름은 인용부호와 함께 지정하고 ‘새로운 데이터셋 이름’을 입력하여 새로운 데이터셋을 만들 수 있다.

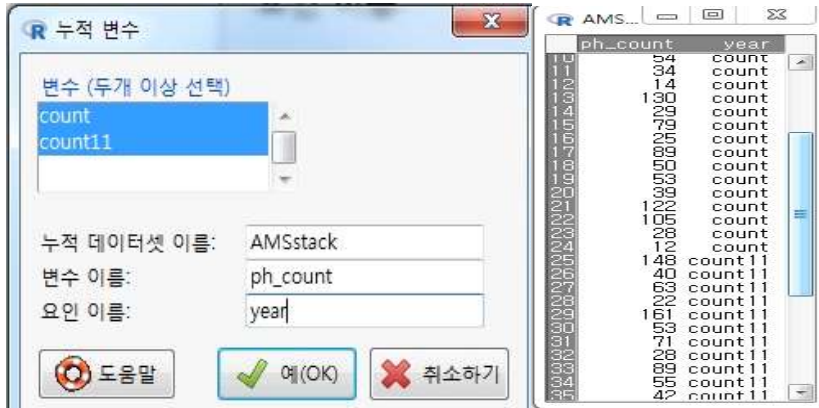


‘활성 데이터셋의 누적변수’를 선택하면 ‘누적 변수’ 창이 뜬다. ‘변수 (두개 이상 선택)’ 아래 나열된 변수들 중 선택된 것의 열들을 누적하여 길어진 하나의 열 변수로 변환하게 되며 ‘변수이름’ 칸에 새 변수의 이름을 지정하고, ‘요인 이름’ 칸에는 누적된 변수들의 이름이 범주가 되는 새로운 변수 즉 요인의 이름을 지정한 다음 ‘누적 데이터셋 이름’ 칸에 원하는 이름을 입력하면 새 변수와 요인의 두 열로 이뤄지게 되는 누적 데이터셋이 만들어진다. 이 기능은 요인을 사용하는 분산분석 모형이나 다른 여러 선형모형을 사용하여 통계처리할 때 유용하게 활용될 수 있다. 참고로 새로운 변수의 이름은 R에서 유효한 대소문자, 숫자, 언더스코어 등만으로 지어야 하며 숫자가 앞에 나올 수는 없다. [그림 A-13]은 ‘car’ 패키지의 미국 수학회에서 발표한 기관별 2008-09(count)과 2011-12(count11)년도 수학분야 박사 취득자 수에 대한 자료인 데이터셋 ‘AMSSurvey’를 불러와 24행인 두 변수 "count"와 "count11"을 누적하여 48행인 새 변수 "ph\_count"와 요인 "year"라는 변수들로 이뤄진 누적 데이터



셋 ‘AMSstack’으로 만드는 예를 보여 준다.

[그림 A-13]



‘결측 자료 사례 제거하기’를 선택하면 ‘결측 데이터 제거하기’ 창이 뜬다. ‘모든 변수 포함하기’ 항목을 택하면 각 변수에 대해 결측이 있는 사례들을 모두 제거하고 특정 변수의 결측만 문제라면 ‘변수 (하나 이상 선택)’ 아래 목록에서 지정하고 ‘새로운 데이터셋 이름’ 아래 칸에 원하는 이름을 입력하여 새 데이터셋을 만든다.



‘활성 데이터셋 저장하기’를 선택하면 현재 사용 중인 활성 데이터셋을 원하는 장소에 저장할 수 있는 대화창이 열린다.

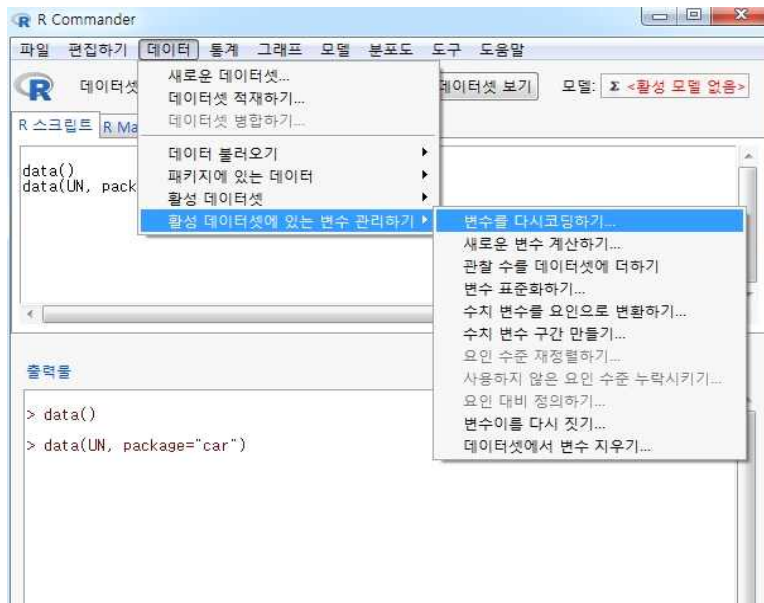
‘활성 데이터셋 내보내기’를 선택하면 현재 사용 중인 활성 데이터셋을 외부에서 사용할 수 있게 하기 위하여 ‘필드 구분자’를 지정하여 텍스트 파일 형식으로 원하는 장소에 저장할 수 있도록 하는 대화창이 열린다. ‘행 이름 쓰기’ 항목을 택하면 행번호들로 이뤄진 하나의 열이 만들어지는데 그 열의 변수명이 지정이 되어 있지 않아서

외부에서 사용하려 할 때 수정하지 않으면 오류가 발생할 수 있어 주의해야 한다.

## 5.2 활성데이터셋에 있는 변수 관리하기

메뉴 막대의 ‘데이터’를 클릭하고 [‘데이터’ -> ‘활성 데이터셋에 있는 변수 관리하기’]를 선택하면 활성 데이터셋에 있는 변수들을 관리하기 위한 부메뉴들이 [그림 A-14]에서와 같이 나타난다.

[그림 A-14]



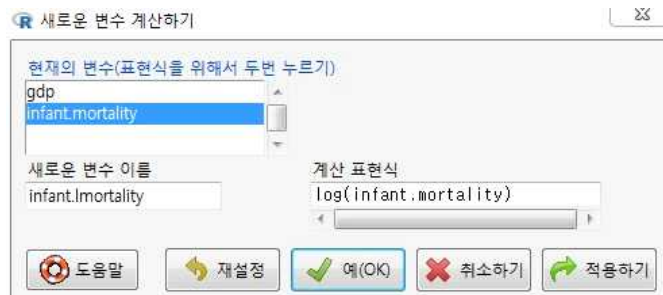
‘변수를 다시코딩하기...’는 보통 계량형 변수나 요인들을 값을 묶거나 수준들을 묶어서 다른 요인으로 다시 코딩을 할 목적으로 사용하지만 다른 수치형 변수로 만들 수도 있기는 하다. 이 메뉴를 클릭하면 ‘변수 다시코딩하기’ 창이 뜨며 ‘다시코딩할 변수(하나 이상 선택)’ 아래 칸에 나열되어 있는 활성 데이터셋의 변수들 중 코딩하려는 것(여기서는 "gdp")을 선택한다. ‘다중적 재코딩을 위한 새로운 변수 이름 또는 접미사’ 옆 칸에 새로 만들 변수의 이름(여기서는 "gdp\_c")을 지정하고 ‘다시 코딩하기’ 지시문 입력하기 아래 칸에 코딩 식들을 입력한다. 한 줄에 여러 식을 쓸 때는 각 식 뒤에 세미콜론(;)을 넣어야 한다. ‘(각각의) 새로운 변수를 요인으로 만들기’를 선택 해제하면 다시 코딩된 변수가 수치형이 된다. [그림 A-15]는 "gdp"가 결측치(NA)인 경우는 NA, 1000 이하는 1, 10000 이하는 2, 그 외는 3 으로 다시 코딩한 결과를 보인다. ‘(각각의) 새로운 변수를 요인으로 만들기’를 선택 해제하여 다시 코딩된 변수

(여기서는 "gdp\_n")가 수치형이 되도록 하려면 '다시 코딩하기' 지시문 입력하기' 아래 칸에 결측치(NA)인 경우는 NA=NA와 같이 문자임을 나타내는 인용부호를 빼야 한다.

[그림 A-15]



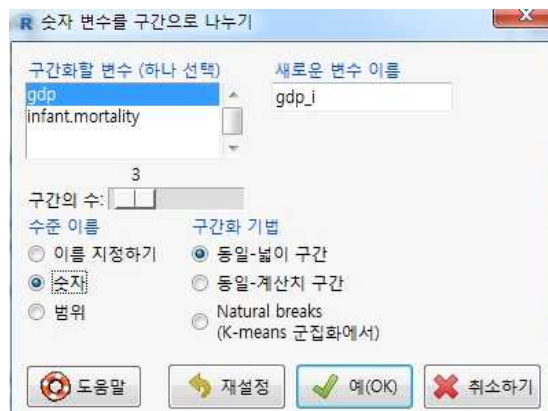
‘새로운 변수 계산하기...’를 클릭하면 나오는 ‘새로운 변수 계산하기’ 창에서 ‘현재의 변수(표현식을 위해서 두 번 누르기)’ 아래 칸에 나열된 변수들 중 변환하고자 하는 것(여기서는 "infant.mortality")을 선택하고 ‘새로운 변수 이름’ 아래 칸에 원하는 새 변수의 이름(여기서는 "infant.lmortality")을 정하여 입력한 뒤 ‘계산 표현식’ 아래 칸에 수식(여기서는 "log(infant.mortality)")을 입력한 다음 ‘예(OK)’를 누르면 변수 "infant.mortality"를 로그변환하여 계산한 새 변수 "infant.lmortality"가 만들어진다.



‘관찰 수를 데이터셋에 더하기...’를 클릭하면 현재 사용 중인 활성 데이터셋(여기서는 ‘UN’)에 관측치 번호를 나타내는 새 변수 "Obsnumber"의 열이 추가되었음을 보여주는 창이 뜬다.

‘변수 표준화하기...’를 클릭하면 나오는 ‘변수 표준화하기’ 창에서 ‘변수(하나 이상 선택)’ 아래 칸에 나열된 변수들 중 표준화하고자 하는 것(여기서는 "gdp", "infant.mortality")을 선택하고 ‘예(OK)’를 누르면 표준화된 새 변수들(여기서는 "Z.gdp", "Z.infant.mortality")이 현재 사용 중인 활성 데이터셋(여기서는 ‘UN’)에 추가되었음이 ‘출력물’ 창에 보여지고 ‘알림글’ 창에 주석이 나타난다.

‘수치 변수 구간 만들기...’를 클릭하면 나오는 ‘숫자 변수를 구간으로 나누기’ 창에서 ‘구간화할 변수(하나 선택)’ 아래 칸에 나열된 변수들 중 구간으로 나누고자 하는 것(여기서는 "gdp")을 선택하고 ‘새로운 변수 이름’ 아래 칸에 원하는 새 변수의 이름(여기서는 "gdp\_i")을 정하여 입력한다. 다음으로 ‘구간의 수’ 슬라이더를 움직여 만들고자 하는 구간의 개수(여기서는 3)를 택한 뒤 ‘구간화 기법’과 ‘수준 이름’의 선택사항(여기서는 ‘동일-넓이 구간’, ‘숫자’)들을 정하여 ‘예(OK)’를 누르면 현재 사용 중인 활성 데이터셋(여기서는 ‘UN’)에 구간화된 새 변수(여기서는 "gdp\_i")가 추가되었음을 보여주는 창이 뜬다. ‘수준 이름’의 선택사항은 수준의 이름을 원하는 이름을 지정하여 입력하기, 구간 번호인 숫자로 나타내기, 구간 범위로 나타내기와 같이 세가지이다. ‘구간화 기법’의 선택사항은 구간의 넓이를 같게 하기, 구간내 관측치 개수를 같게 하기, 구간 내 분산은 줄이고 구간 간 분산은 최대화 하기(Natural breaks)와 같이 세가지이다.



‘수치 변수를 요인으로 변환하기...’를 클릭하면 나오는 ‘수치 변수를 요인으로 변환하기’ 창에서 ‘변수(하나 이상 선택)’ 아래 칸에 나열된 변수들 중 요인으로 변환하고자 하는 것(여기서는 "gdp\_n")을 선택하고 ‘다중 변수를 위한 새로운 변수 이름 또는 접미사’ 옆의 칸에 변환된 요인의 이름(여기서는 "gdp\_f")을 정하여 입력한 다음 ‘요인 수준’의 선택사항(여기서는 ‘수준 이름 제공하기’)을 정하고 ‘예(OK)’를 누르면 ‘Level Names for gdp\_f’ 창이 나오며 숫자 값에 대한 수준 이름(여기서는 ‘L’, ‘M’,

‘H’)을 정하여 입력하면 변환된 요인(여기서는 "gdp\_fr")이 현재 사용 중인 활성 데이터셋(여기서는 ‘UN’)에 추가되며 ‘데이터셋 보기’로 확인할 수 있다.

‘요인수준 재정렬하기...’를 클릭하면 나오는 ‘요인 수준 다시 정렬하기’ 창에서 ‘요인(하나 선택)’ 아래 칸에 나열된 변수들 중 요인 수준 다시 정렬하고자 하는 것(여기서는 "gdp\_fr")을 선택하고 ‘요인 이름’ 아래 칸에 정렬된 새 요인의 이름(여기서는 "gdp\_fr")을 정하여 입력한 다음 ‘순서화된 요인 만들기’의 선택사항을 정하고 ‘예(OK)’를 누르면 ‘수준 다시 정렬하기’ 창이 나오며 ‘예전 수준’을 원하는 ‘새로운 순서’로(여기서는 역순인 ‘3’, ‘2’, ‘1’)을 정하여 입력하면 역순으로 정렬된 요인(여기서는 "gdp\_fr")이 현재 사용 중인 활성 데이터셋(여기서는 ‘UN’)에 추가되며 ‘데이터셋 보기’로 확인할 수 있다.

‘사용하지 않은 요인 수준 누락시키기...’를 클릭하면 나오는 창에서 ‘모든 요인’을 택하거나 또는 ‘수준을 누락시킬 요인(하나 이상 선택)’ 아래 칸에 나열된 변수들 중 수준을 누락시키고자 하는 것을 선택하고 ‘예(OK)’를 누르면 확인하는 창이 나오며 ‘예(OK)’를 누르면 선택된 요인의 사용하지 않은 요인 수준이 없어진다. 하위 데이터셋이 만들어질 때 사용하지 않은 요인 수준이 발생할 수 있으며 이 메뉴를 적용하여 누락시킬 수 있다.

‘요인 대비 정의하기...’를 클릭하면 나오는 ‘요인 대비 설정하기’ 창에서 ‘요인(하나 선택)’ 아래 칸에 나열된 변수들 중 대비를 정의하려는 것을 선택하고 ‘예(OK)’를 누르면 선택한 변수에 대하여 원하는 형식의 대비가 정의된 것을 알 수 있는 문장이 ‘출력물’ 창에 나타난다. 이 메뉴는 회귀분석이나 분산분석과 같이 선형모형을 사용하는 경우 범주형 변수나 요인을 더미변수를 이용하여 나타내어 사용하지만 대신하여 이 더미변수들의 선형결합인 대비로 코딩하여 이용할 수 있도록 한다.

‘변수 이름 다시 짓기...’를 클릭하면 나오는 ‘변수 이름 다시 짓기’ 창에서 ‘변수(하나 이상 선택)’ 아래 칸에 나열된 변수들 중 이름을 변경하려는 것을 선택하고 ‘예(OK)’를 누른다. 선택한 변수에 대한 ‘새로운 이름’ 아래 칸에 원하는 이름을 입력한 뒤 ‘예(OK)’를 누르면 ‘출력물’ 창에 변수 이름을 변경하는 문장이 나타나는 것을 볼 수 있다.

‘데이터셋에서 변수 지우기...’를 클릭하면 나오는 ‘변수 삭제하기’ 창에서 ‘삭제할 변수(하나 이상 선택)’ 아래 칸에 나열된 변수들 중 삭제하려는 것을 선택하고 ‘예(OK)’를 누르면 삭제를 확인하는 창이 뜨고 ‘예(OK)’를 누르면 ‘출력물’ 창에 변수를 삭제했음을 알 수 있는 문장이 나타난다.

## 6. 수치적 요약과 그래프의 생성

앞에서는 정보를 얻기 위한 원료인 자료를 다루는 방법들에 대한 메뉴를 알아보았

으며, 여기서는 정제된 자료로부터 필요한 정보를 얻는 통계적 방법에 관한 메뉴 중 ‘요약’에 대해서 간단히 알아보자.

메뉴 막대의 ‘통계’를 클릭하고 [‘통계’ -> ‘요약’ -> ‘활성 데이터셋’]을 선택하면 활성 데이터셋(여기서는 ‘UN’과 ‘AMSSurvey’)에 있는 변수들에 대해 요약된 정보가 [그림 A-16]과 같이 ‘출력물’ 창에 나타난다. 수치형 변수에 대해서는 평균과 결측치 개수와 더불어 사분위수와 최대값, 최소값으로 구성된 다섯 숫자 요약이 나타나고, 범주형 변수에 대해서는 도수와 결측치 개수가 나타난다.

[그림 A-16]

The figure consists of two screenshots of the R Commander interface. The top screenshot shows the '통계' (Statistics) menu open, with '요약' (Summary) selected, and the '활성 데이터셋' (Active Dataset) submenu also open. The '출력물' (Output) window displays the summary for the 'UN' dataset, showing statistics for 'infant.mortality' and 'gdp'. The bottom screenshot shows the same menu path, but with 'AMSSurvey' as the active dataset. The '출력물' window displays the summary for 'AMSSurvey', showing statistics for 'type', 'sex', 'citizen', and 'count'.

**Top Screenshot: Summary for UN dataset**

```
> summary(UN)
infant.mortality      gdp
Min.      : 2.00      Min.      : 36
1st Qu.   : 12.00     1st Qu.   : 442
Median    : 30.00     Median    : 1779
Mean      : 43.48     Mean      : 6262
3rd Qu.   : 66.00     3rd Qu.   : 7272
Max.      : 169.00    Max.      : 42416
NA's      : 6         NA's      : 10
```

**Bottom Screenshot: Summary for AMSSurvey dataset**

```
> summary(AMSSurvey)
      type sex citizen count count11
I(Pr):4  Female:12 Non-US:12 Min. : 12.00 Min. : 17.0
I(Pu):4  Male :12  US :12   1st Qu.: 31.25 1st Qu.: 29.5
II :4                               Median: 48.50 Median: 55.5
III :4                               Mean : 59.58 Mean : 68.5
IV :4                               3rd Qu.: 87.50 3rd Qu.: 95.5
Va :4                               Max. : 132.00 Max. : 161.0
```

‘수치적 요약...’을 선택하면 수치형 변수들에 대한 표준편차와 변동계수 등 좀 더 다양한 정보를 얻을 수 있고, 범주형 변수들에 대해서는 ‘빈도 분포...’에서 백분율이 주어진 도수분포표를 ‘출력물’ 창에 나타내며 ‘카이-제곱 적합성 검정(오직 하나의 변수)’을 선택할 수 있다.

‘관찰 결측치 셈하기’를 선택하면 활성 데이터셋의 모든 변수들에 대한 결측치 개수를 ‘출력물’ 창에 나타낸다.

‘통계표’에서는 ‘요인(하나 이상 선택)’에 의해 집단화된 ‘반응 변수(하나 이상 선택)’에 대하여 선택한 통계량을 계산하여 ‘출력물’ 창에 나타낸다.

‘상관 행렬’에서 ‘변수(두개 이상 선택)’에서 택한 변수들에 대한 세가지 ‘상관 유형’ 중 택한 상관계수값들의 행렬이 ‘출력물’ 창에 나타난다.

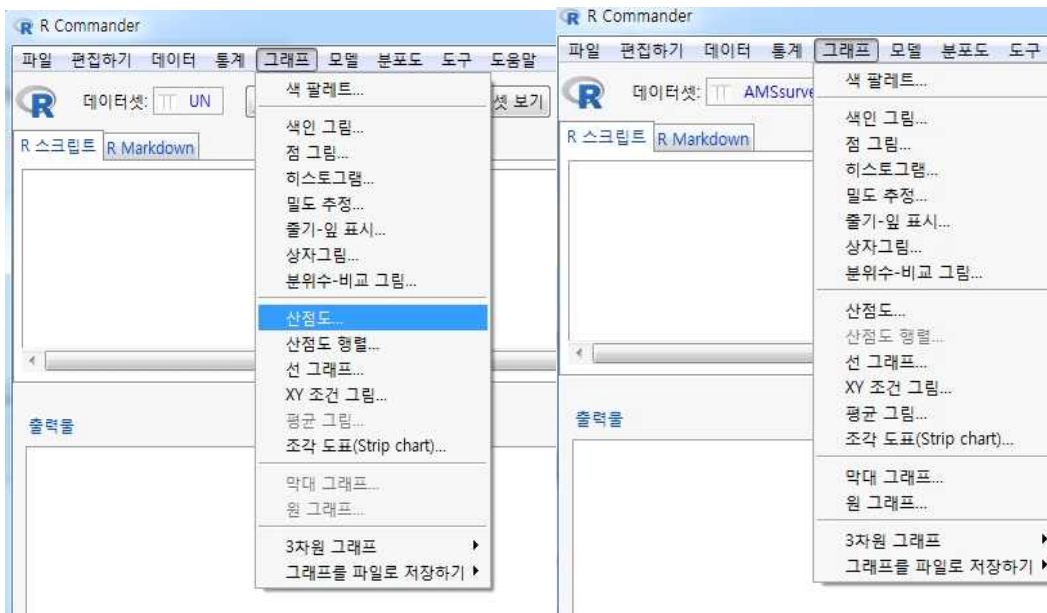
‘상관 검정’에서는 ‘변수(두개 만 선택)’에서 택한 두 변수에 대한 세가지 ‘상관 유

형’ 중 택한 상관계수값에 의해 지정한 형태의 대립가설에 대한 검정과 구간추정을 수행한 결과를 ‘출력물’ 창에 나타낸다.

‘정규성 검정...’에서는 선택한 ‘변수(하나 선택)’에 대한 ‘Shapiro-Wilk’ 등 여섯가지 ‘정규성 검정’ 방법 중 택한 결과를 ‘출력물’ 창에 나타낸다. 정규성이 의심되는 경우 변수변환이나 비모수적 방법 등을 사용한다.

도표를 활용하면 직관적으로 빠르게 정보를 얻을 수 있으므로 메뉴 막대의 ‘그래프’를 클릭하여 활성 데이터셋의 변수들에 대한 [그림 A-17]에 나타나는 히스토그램이나 산점도 등 다양한 형태의 그래프를 선택하여 별도의 ‘R Graphics’ 창에 보일 수 있다. 가장 최근의 그림만 나타나므로 이전 그림을 보려면 페이지 업 키와 페이지 다운 키를 사용해야 한다. [‘그래프를 파일로 저장하기’ -> ‘bitmap(으)로...’]에 의해 PNG, JPEG 그래픽 파일 유형으로 저장하거나 [‘그래프를 파일로 저장하기’ -> ‘PDF/Postscript/EPS(으)로...’]에 의해 PDF, Postscript, Encapsulated Postscript 그래픽 파일 유형으로 저장할 수 있고, [‘그래프를 파일로 저장하기’ -> ‘3차원 RGL 그래픽...’]에 의해 ‘3차원 그래프’에서 그려진 그림이 있을 때 마우스로 적절하게 움직여 조정한 뒤 파일로 저장한다.

[그림 A-17]





## [참고문헌]

- [1] 강근석, 유현조 (2016). R을 활용한 선형회귀분석, 교우사
- [2] 김정일 (2008). R 활용 회귀분석의 이해, 강호
- [3] 박성현 (1999). 회귀분석, 민영사
- [4] 최경화, 하미나 (2013). R-Commander를 이용한 통계분석, 단국대학교출판부
- [5] Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), Regression diagnostics: Identifying influential data and sources of collinearity, John Wiley & Sons, New York.
- [6] Box, G.E.P., and Cox, D.R. (1964). An Analysis of Transformations, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 26, No. 2. (1964), pp. 211-252.
- [7] Cook, R.D., and Weisberg, S. (1982). Residuals and Influence in Regression, Chapman & Hall
- [8] Draper, N.R. and Smith, H. (1998). Applied Regression Analysis, Wiley, New York
- [9] Faraway, J.J. (2002). Practical Regression and Anova using R, <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>
- [10] Fox, J. (2005). The R Commander: "A Basic-Statistics Graphical User Interface to R", Journal of Statistical Software, 14, Issue 9
- [11] Hoerl, A.E., and Kennard, R.W. (1970). "Ridge regression : Biased estimation for non-orthogonal problems", Technometrics, 12, pp. 55-67
- [12] Nelder, J. A. and Wedderburn R. W. M. (1972). "Generalized Linear Models", Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3, pp. 370-384
- [13] Ritz, C. and Streibig J. C. (2008). Nonlinear Regression with R, Springer, New York