

회귀분석

울산의대 임상약리학과/마취통증의학과

노규정

Regression

**Department of Clinical Pharmacology and
Therapeutics/Anesthesiology and Pain Medicine
University of Ulsan College of Medicine**

Gyujeong Noh, M.D. & Ph.D.

목 차

목 차 3

책머리에 8

편집자 서문.....	9
집필진 소개.....	10
1. 편집자.....	10
2. 주 집필진.....	10
3. 수정 및 보완.....	10
4. 감수.....	10
회귀분석(Regression).....	11
1. 회귀분석 소개(Introduction: Regression Analysis).....	11
1.1. 회귀 모형(Regression Models).....	13
1.2. 회귀분석의 정식 사용(Formal Uses of Regression Analysis).....	16
1.3. 데이터베이스(Data Base)	18
2. 단순선형회귀모형(Simple Linear Regression Model)	19
2.1. 모형의 기술(Model Description)	19
2.2. 모형모수에 대한 가정과 해석(Assumptions and Interpretation of Model Parameters).....	20
2.3. 최소제곱공식(Least Squares Formulation)	24
2.4. 최대우도법(Maximum Likelihood Estimation)	35
2.5. 변이의 분할(Partitioning Total Variability)	38
2.6. 절편과 기울기의 가설검정(Test of Hypotheses on Slope and Intercept).....	44
2.7. 원점을 통과하는 단순 회귀, 고정 절편(Simple Regression through the Origin, Fixed Intercept)	52
2.8. 적합 모형의 질(Quality of Fitted Model).....	57
2.9. 평균반응과 예측구간에 대한 신뢰구간(Confidence Intervals on Mean Response and Prediction Intervals).....	64
2.10. 단순선형회귀에서 동시추론(Simultaneous Inference in Simple Linear Regression)	74
2.11. 잔차 조망(a Look at Residuals)	87
2.12. 확률변수 x 와 y (Both x and y Random)	97
3. 다중선형회귀모형(The Multiple Linear Regression Model).....	104
3.1. 모형 기술 및 가정(Model Description and Assumptions)	104
3.2. 일반 선형모형과 최소제곱과정(The General Linear Model and the Least	

Squares Procedure)	108
3.3. 이상적인 조건하에서 최소제곱추정량의 속성(Properties of Least Squares Estimators under Ideal Conditions).....	115
3.4. 다중선형회귀에서의 가설검정(Hypothesis Testing in Multiple Linear Regression)	121
3.5. 다중회귀에서 신뢰구간 및 예측구간(Confidence Interval and Prediction Intervals in Multiple Regression)	142
3.6. 반복측정자료(Data with Repeated Observation)	148
3.7. 다중회귀에서 동시추정(Simultaneous Inference in Multiple Regressions) ...	153
3.8. 다중회귀자료에서 다중공선성(Multicollinearity in Multiple Regression Data)	
158	
3.9. 품질적합, 품질예측, 모자행렬(Quality Fit, Quality Prediction, and the Hat Matrix) 169	
3.10. 범주형 또는 지표형 변수; 회귀모형과 ANOVA모형 (Categorical or Indicator Variables; Regression Model and ANOVA Model).....	173
4. 최선의 모형을 선택하기 위한 기준(Criteria for Choice of Best Model)	197
4.1. 모형간의 비교를 위한 표준기준(Standard Criteria for Comparing Models)	
199	
4.2. 모형 선택을 위한 상호 검증과 모형 성능의 결정(Cross Validation for Model Selection and Determination of Model Performance)	202
4.3. 예측에 대한 개념적 기준 - C_p 통계량(Conceptual Predictive Criteria - The C_p Statistic).....	219
4.4. 순차적 변수 선택 과정(Sequential Variable Selection Procedures)	231
4.5. 추가적인 설명과 모든 가능한 회귀들(Further Comments and All Possible Regressions).....	242
5. 잔차분석(Analysis of Residuals)	251
5.1. 잔차로부터 얻어지는 정보(Information retrieved from Residuals)	252
5.2. 잔차도(Plotting of Residuals)	255
5.3. 스튜던트화 잔차(Studentized Residuals)	263
5.4. 표준화 PRESS 잔차에 대한 관계(Relation to Standardized PRESS Residuals)	
269	
5.5. 이상치 탐지(Detection of Outliers)	270
5.6. 진단그림(Diagnostic Plots)	285
5.7. 정규잔차도(Normal Residual Plots)	299
5.8. 잔차분석에 대한 추가해설(Further Comments on Analysis of Residuals) ...	301
6. 영향력 진단(Influence Diagnostics)	302

6.1.	영향요인(Source of Influence).....	303
6.2.	진단: 잔차와 모자행렬 (Diagnostics: Residuals and the HAT Matrix).....	305
6.3.	영향의 정도를 결정하는 진단도구(Diagnostics that Determine Extent of Influence) 314	
6.4.	성능에 대한 영향 (Influence on Performance).....	327
6.5.	영향력이 큰 자료에서 우리는 무엇을 해야하는가?(What Do We Do with High Influence Points?).....	332
7.	비표준조건들, 가정위배와 변환(Nonstandard Conditions, Violations of Assumptions, and Transformations)	334
7.1.	이분산: 가중최소제곱(Heterogeneous Variance: Weighted Least Squares) ..	335
7.2.	상관오차와 관련된 문제들, 자기상관(Problem with Correlated Errors, Autocorrelation)	348
7.3.	적정과 예측을 증진시키는 변환(Transformations To Improve Prediction)	355
6.	다중공선성의 진단과 제거(Detecting and Combating Multicollinearity).....	441
8.1.	다중공선성 진단(Multicollinearity Diagnostics)	442
8.2.	분산 비율(Variance Proportions)	445
8.3.	다중공선성에 관한 추가 주제(Further Topics Concerning Multicollinearity) 454	
8.4.	다중공선성이 있는 경우 최소제곱법의 대안(Alternatives to Least Squares in Cases of Multicollinearity)	468
9.	비선형회귀(Nonlinear Regression)	504
9.1.	비선형 최소제곱(Nonlinear least Squares)	505
9.2.	최소제곱추정량의 특성(Properties of the Least Squares Estimators)	506
9.3.	추정량을 찾기 위한 가우스, 뉴튼방법(The Gauss-Newton Procedure for finding Estimates).....	508
9.4.	가우스, 뉴튼 방법의 다른 변형(Other Modification of the Gauss-Newton Procedure)	516
9.5.	특별한 비선형 모형(Some Special Classes of Nonlinear Models)	521
9.6.	비선형회귀분석에서 고려하여야 할 사항들(Further Considerations in Nonlinear Regression)	524
9.7.	자료를 변환하여 선형화하지 않는 이유는?(Why Not Transform Data to Linearize?)	530
10.	부록 A: 행렬대수의 몇 가지 특별한 개념들(Some Special Concepts in Matrix Algebra) 533	
10.1.	(A.1) 동시선형방정식의 해(Solutions to Simultaneous Linear Equations)...	533
10.2.	(A.2) 이차형식(The Quadratic Form).....	536

10.3.	(A.3) 고유값과 고유벡터(Eigenvalues and Eigenvectors).....	539
10.4.	(A.4) 분할 행렬의 역(The Inverse of a Partitioned Matrix).....	543
10.5.	(A.5) Sheerman-Morrison-Woodbury 정리(Sheerman-Morrison-Woodbury Theorem) 545	
11.	부록 B. 몇 가지 특별한 조작법(Some Special Manipulations)	546
11.1.	(B.1) 잔차평균제곱의 비편향성(Unbiasedness of the Residual Mean Square) 546	
11.2.	(B.2) 저설정된 모형에서 잔차제곱합과 평균제곱의 기대값(Expected Value of Residual Sum of Squares and Mean Square for an Underspecified Model)	548
11.3.	(B.3) 최대우도추정량(The Maximum Likelihood Estimator)	551
11.4.	(B.4) PRESS 통계량의 수식 전개(Development of the PRESS Statistic)554	
11.5.	(B.5) S_{-i} 의 계산(Computation of S_{-i}).....557	
11.6.	(B.6) 대응하는 모형오차에 대한 잔차의 우월성(Dominance of a Residual by the Corresponding Model Error)	559
11.7.	(B.7) 영향력 진단도구의 계산(Computation of Influence Diagnostics)560	
11.8.	(B.8) 비선형 모형에서 최대우도추정량 (Maximum Likelihood Estimator in the Nonlinear Model).....563	
11.9.	(B.9) 테일러 급수(Taylor Series)	564
11.10.	(B.10) C_k 통계량의 수식전개(Development of the C_k -Statistic)	565
12.	부록 C: 행렬을 이용한 선형회귀분석	569
12.1.	(C.1) 선형회귀분석에서의 기본적인 행렬	569
12.2.	(C.2) 무작위 벡터 혹은 행렬의 기대값	572
12.3.	(C.3) 무작위 벡터의 분산-공분산 행렬	573
12.4.	(C.4) 행렬식을 이용한 단순한 회귀모형	575
12.5.	(C.5) 최소자승법을 이용한 회귀분석 파라메터의 추정	577
12.6.	(C.6) 관찰 추정값	578
12.7.	(C.7) 잔차	579
12.8.	(C.8) 잔차의 분산-공분산 행렬	580
12.9.	(C.9) 회귀계수의 분산-공분산 행렬	581
	기초통계 요약(Summary of Basic Statistics).....	583
1.	변수(자료)의 유형	583
2.	종속변수 vs. 독립변수	583
3.	확률과 확률 분포	583
4.	표본 분포	589

5. 회귀분석 요약	601
6. 범주형 자료에 대한 확률분포들 (Distributions for categorical data)	610
7. 통계적 추론(Statistical inference for categorical data)	611
그리스 문자(Greek Letters)	617

책머리에

그간 “NONMEM”이라는 절실하고 소중한 지식을 전하기 위하여 수년 간 애쓰신 노규정 선생의 노고에 존경과 감사를 드립니다. 본서에서 전해주는 소중한 내용의 근본이, 통계학과 임상약리학에 바탕 하는 바, 이 분야에서는 방랑자격인 제가 감히 머리글을 올리는 무례함과 부족함을 너그러이 해아려 주시기 바랍니다.

본서를 번안 및 저술하신 노규정 선생은 임상약리학 주임교수 및 마취통증의학 교수로서, 두 가지 전공을 통하여, 임상의사로서, 한편으로는 연구자로서 힘든 일을 의욕적으로 실천하고 있는 분입니다. 2001년 서울아산병원에 오신 이후 저와는 늘 든든한 수술 팀의 동료이자, 다른 한편으로는, 비록 제가 전공하는 종양생물·유전학과는 다소 거리가 있어 보이지만, 실험실에서 진행하는 항암약제 반응성 표식자 연구에 대하여 언제든 고견을 주고 있습니다.

통계학은 예측성을 근간으로 현대의 거시적 산업구도로부터, 미세한 생명현상을 탐구하는데까지 필수적인 학문이지만, 복잡한 수리적 추론 및 전개를 접하는 순간 상당한 거리감을 느끼게 하는 것도 사실입니다. 실제로 사용자 입장에서는 단순한 수치와, 가설에 대한 검증 결과만을 요구하게 되는데, 이 경우 간편한 결과 획득은 가능할지 모르지만 결국 통계학의 근본인 적용성과 신뢰도에 심각한 오류를 야기할 수 있겠습니다. 이러한 관점에서 본서에서는 통계학적 요인과 구성요소의 전개과정에 관하여, 수리학적 배경을 자세하게 알려주고 객관적인 검증을 제시하며, 나아가서 모형개발을 가능하게 해 준 점이 돋보인다 하겠습니다. 즉, 통계처리의 단계별 기능과 의미를 상세하게 예제를 통하여 기술함으로써 객관적인 신뢰성을 부각시킨 점에서 실용성이 크다고 하겠습니다. 후반부에서는 21세기 생명과학의 근간이 되고 있는, 약제개발의 바탕을 구성하는 약력학과 약동학의 최적화 작업을 위하여 비선형혼합효과모형을 알기 쉽게 풀이하였습니다. 예를 들면 개인과 약제별로 각종 변수를 다양하게 규정하고, 적절하게 이를 적용함으로써 임상단계 진입에 필수적 관문인 약물의 혈장농도 혹은 약리효과의 개인차를 정확하게 검증하는 방법을 일러주고 있습니다. 본서는 이 방면에 입문하시거나 혹은 이미 종사하고 계시는 많은 연구자에게 도움이 되리라 확신하며, 특히 생명과학을 구성하는 정확성, 객관성 및 창의성을 제공하는 길잡이가 될 것입니다. 저자가 제게 일러 준 대로, 보다 실질적인 우리 모델의 예제를 향후 추가함으로써, 더욱 친근한 지침서로 발전할 것을 기대합니다. 다시 한번 그 길고 험한 여정을 쉽게 갈 수 있도록 해 주신 노규정 선생과 공동 집필, 번안 및 편집자 여러분의 노고에 깊은 감사의 마음을 전합니다.

2007년 중간

여름 서울아산병원 암센터 소장 김진천

편집자 서문

계량약물학(pharmacometrics)에 관련된 영어 책들을 보면 참 어렵다는 생각이 가장 먼저 들 것이다. 이 중에서도 특히 혼합효과모형(nonlinear mixed effects model)을 이용한 집단접근방법(population approach)은 완전히 이해한 다음 수행하기가 쉽지 않은 분야이다. 이 매뉴얼은 모형이라는 것을 들어 본 적이 없는 분들이 좌절하지 않고, 쉽게 이 분야로 진입하였으면 좋겠다는 마음에서 시작하였다. 그러나 선형회귀(linear regression) 분야는 너무 전문적인 내용들이 들어가 버렸고, 혼합효과 모형은 너무 쉬운 내용으로 구성되었다. 또한 약리적인 내용에는 많은 약동(pharmacokinetics), 약력(pharmacodynamics) 분야가 빠져 있다. 다 완성되고 난 후 배포하면 더 좋겠으나, 혼합효과모형을 처음 배우시는 분들을 생각하여 미완의 상태로 먼저 배포하고자 한다. 따라서 읽으시는 분들께서 구성의 조악함을 너무 탓하지 말아 주셨으면 한다.

이 매뉴얼의 구성은 4부로 되어 있다. 1부는 선형회귀 일반에 대한 내용이 실려 있다. 최소제곱추정절차(least squares estimation procedure) 및 최대우도추정절차(maximum likelihood estimation procedure)와 모형선택(model selection) 및 진단(diagnostics)에 관련된 내용이 주를 이룬다. 2부는 비이상적인 조건(non-ideal condition)에서 최소제곱추정절차의 대안으로 사용할 수 있는 모형과 비선형 회귀(nonlinear regression)에 관한 내용으로 구성되어 있다. 또한 모형과 관련된 행렬대수(matrix algebra)와 마지막에 기초통계에 대한 요약도 실었다. 3, 4부에서는 스탠포드 대학의 Dr. Shafer의 NONMEM workshop 매뉴얼을 번역하여 실었다. 혼합효과모형에 관하여 내가 본 어떤 책보다도 알기 쉽게 설명한 워크숍 매뉴얼이 Dr. Shafer가 만든 매뉴얼이다. 그것을 한글로 번역하여 배포할 수 있게 허락해 준 Dr. Shafer에게 무한한 감사를 드린다.

이 매뉴얼의 대상은 아마도 다양할 것이나, 임상약리학에 첫발을 내디딘 저년차 전공의, 임상에서 환자진료를 하면서 얻은 각종 약동, 약력 자료를 모형화하고자 하는 임상의사가 될 것으로 생각된다. 혹은 NONMEM®을 사용하여 자료를 분석하고자 하는 타분야의 연구자들에게도 도움이 될 것으로 믿는다. 나중에 매뉴얼의 구성이 더 탄탄해지면, 임상약리학 전공자도 볼 수 있게 되었으면 하는 바램이다.

모쪼록, 이 분야에서 큰 성취를 이루는 분들이 많아졌으면 좋겠고, 이 매뉴얼이 거기에 조그만 도움이 된다면 더 바랄 나위가 없을 것이다.

2007-06-28 오전 11시 49분

울산의대 서울아산병원 임상약리학과/마취통증의학과

노규정

집필진 소개

1. 편집자

노규정: 울산의대 서울아산병원 임상약리학과 과장 및 주임교수/마취통증의학과 교수

2. 주 집필진

노규정: 울산의대 서울아산병원 임상약리학과 과장 및 주임교수/마취통증의학과 교수

김순임: 순천향의대 마취통증의학과 교수

김계민: 인제의대 상계백병원 마취통증의학과 조교수

윤희석: 충남의대 마취통증의학과 조교수

이상석: 인제의대 상계백병원 마취통증의학과 조교수

신혜원: 고려의대 안암병원 마취통증의학과 부교수

배균섭: 울산의대 서울아산병원 임상약리학과 조교수

임형석: 울산의대 서울아산병원 임상약리학과 조교수

이은호: 울산의대 서울아산병원 마취통증의학과 임상조교수, 박사과정

이은경: 울산의대 임상약리학과 연구교수

윤성철: 울산의대 예방의학교실 연구교수

최병문: 국립의료원 마취통증의학과 전문의

3. 수정 및 보완

방지연: 울산의대 서울아산병원 마취통증의학과 임상강사, 박사과정

조상현: 울산의대 서울아산병원 임상약리학과 전공의, 예방의학 박사과정

김성훈: 서울아산병원 마취통증의학과 전공의

김종률: 울산의대 서울아산병원 임상약리학과 전공의, 석사과정

최상민: 울산의대 서울아산병원 임상약리학과 전공의, 석사과정

김운집: 울산의대 서울아산병원 임상약리학과 전공의, 석사과정

정진아: 울산의대 서울아산병원 임상약리학과 전공의, 석사과정

4. 감수

박희진: 서울대학교 통계학과 박사후과정

윤성철: 울산의대 예방의학교실 연구교수

이은경: 울산의대 임상약리학과 연구교수

회귀분석(Regression)

1. 회귀분석 소개(Introduction: Regression Analysis)

회귀분석(regression analysis)이라는 용어는 과학적인 체계(scientific systems)로 양(quantities) 사이의 관계에 대하여 추론(inference)을 내리는데 사용되는 통계기술의 집합(collection)을 말한다. 응용통계학분야에는 상품화된 자료분석 기술이 넘쳐 나며, 개개의 분석방법은 특정 문제를 해결하기 위하여 개발된다. 그러나 회귀분석(regression analysis)은 다양한 문제를 해결 가능하게 해주었고, 이 때문에 회귀분석에 대한 책도 많고 사용분야도 계속해서 확장되고 있다. 또한 회귀분석(regression analysis)이라는 이름 하에 새로운 부주제(subtopic)와 부분야(subarea)가 계속 증편되고 있다. 전문적인 통계학자들과 관련분야 과학자들의 상상(imagination)의 산물(product)인 분석 방법론(analytic methodology)은 오늘날 컴퓨터 소프트웨어 관리자들의 신속한 작업 덕분으로 쉽게 접근할 수 있게 되었다.

1885년에 Francis Galton 경은 한 연구에서 “회귀(regression)”이라는 용어를 처음으로 사용하였는데, 이는 자손들이 부모의 키에 따르지 않고 오히려 평균에 따르는 경향이 있다는 것을 증명하였다. 즉, 자손들이 “평균을 향하여 회귀(regression towards mediocrity)”한다는 것이다. 최소제곱법(method of least squares)은 19세기 초에 Carl Fredrich Gauss가 발견하였다고 하나, 논란의 여지가 있다. 확실한 것은, Adrien Marie Legendre가 1805년에 최소제곱법을 사용한 첫 번째 연구를 발표하였다. 회귀분석(regression analysis)과 최소제곱법은 1960년대 후반까지 거의 항상 같이 사용되었고 서로 궁합이 잘 맞는 것처럼 보였다. 그러나, 보통최소제곱법(ordinary least squares, OLS)은, 비이상적인 상황(nonideal situation)에서 부적절하며 개선의 여지가 있다는 것이 확실해졌다. 이에 대하여 수많은 논란이 있었으며, 이러한 논란은 OLS의 대안(alternatives) 개념(notion)을 모호하게 할 가능성이 있다. 그러나 다중공선성에 대한 편향추정법(biased estimation for combating multicollinearity)과 로버스트 회귀(robust regression)는 많은 자료분석가에 의하여 받아들여져서 회귀분석(regression analysis)의 전통 속에 자리를 잡게 되었다.

양 혹은 변수들(quantities, variables)이 서로에게 어떤 영향을 미치는가는 분석가가 흔히 품는 큰 의문이다. 자료를 생성하는 체계(system)는 화학적 또는 생물학적인 과정, 한 나라의 경제, 의학실험(medical experiment)의 환자군일 수 있으며 또는 장력 강도(tensile strength)를 연구하는 금속표본군일 수도 있다. 어떤 경우에는 타당변수(pertinent variable)가 확률변수(random variable)여서 결합확률분포(joint probability distribution)를 통하여 확률적인 의미(probability sense)를 가진다. 변수(variables)가 수학적인 양(mathematical quantities)일 수도 있으며, 이 경우 이들을 연결시키는 함수관계(functional relationship)가 존재한다는 가정(assumption)이 있다. 회귀분석(regression analysis)은, 변수(variables)가 측정되는 자료 세트(a set of data)에서 변수 간 관계에 대한 기전(mechanism)의 어떤 측면을 밝혀내도록

설계된다. 한 예로 회귀분석(regression analysis)이 자료로부터 어떤 종류의 정보를 얻을 수 있는지 설명해 보자. 우리가 실험에 참여하는 한 군의 사람들을 무작위로 골랐다고 가정하자. 각각의 사람마다 특정 거리를 달려야 한다. 우리는 개개인에 대한 다음과 같은 측정치를 만들 수 있다.

x_1 : 나이

x_2 : 체중

x_3 : 휴식 시 측정한 맥박수 (rest pulse)

x_4 : 뛰고 난 후 즉시 측정한 맥박수 (run pulse)

x_5 : 일정거리를 달리는데 걸리는 시간 (run time)

y : 산소소모율 (oxygen rate)

산소소모율 y 는 x_1, x_2, x_3, x_4 및 x_5 와는 약간 다른 역할을 한다. x 변수들은 y 를 결정하거나(determine), 예측하는데(predict) 사용되는 양들(quantities)이다. 그래서 x 변수들을 독립변수(independent variables) 혹은 회귀변수(regression variables)라고 한다. 자료가 주어지면, 분석가는 산소소모율에 영향을 미치는 각 회귀변수(regression variable)의 역할에 대하여 정보를 얻고 싶을 것이다. 만약 한 개 이상의 변수(variables)가 산소소모율에 미치는 영향이 무시할 만 하다면, 이것은 소중한 정보이다. 뿐만 아니라, 자료를 이용하여 변수들(variables) 사이에 존재하는 관계(관계가 존재한다고 가정할 때)를 추정(estimation)할 수 있다. 특정 기법(specific technique)을 개발하거나 토론하기에 앞서 통계적 모형의 개념(a notion of a statistical model)에 대해서 알아보도록 하자.

1.1. 회귀 모형(Regression Models)

다음 장에서는 통계적 모형(*statistical model*)의 사용과 개발에 대해서 주로 설명할 것이다. 회귀분석(regression analysis)에 사용되는 모든 과정(procedures)과 회귀분석에서 도출되는 결론은 회귀모형(regression model)의 가정(assumption)에 의하여 직, 간접적으로 좌우된다. 모형이란, 회귀분석(regression analysis)되는 자료의 생성 기전(mechanism)을 말한다. 회귀모형(regression model)은 대개 대수형태(algebraic form)로 나타낸다. 예를 들어, 산소소모율을 설명할 때 변수 간의 관계(relationship)를, 회귀변수(regressor variables)의 입장에서 선형(linear)인 구조식(structure)으로 잘 나타낼 수 있다고 가정한다면, 다음과 같은 모형이 적당할 것이다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \quad (1.1)$$

식(1.1)에서 $\beta_0, \beta_1, \dots, \beta_5$ 는 회귀계수(*regression coefficients*)라고 불리는 미지의 상수(unknown constants)이다. 회귀분석(regression analysis)의 과정은 이러한 계수(coefficients)에 대한 결론을 추론하는 것이다. x_4 (run pulse)의 증가로 인하여 산소소모율이 감소되는지, 증가되는지 결정하려면, 관련된 계수(coefficient)의 부호(sign)와 크기(magnitude)가 중요하다. ε 항은 모형이 정확하지 않을 경우를 설명하기 위하여 추가되며, 임의 변동(random disturbances) 또는 모형 오차(model error)를 기술한다. 이를 자료 세트(data set)에 적용할 때(1.1), ε 항은 모형에서 주어지는 항들로부터 동떨어진 개인의 모든 변동(variation)을 설명하는 도우미로 볼 수 있다. 2장에서는 회귀분석(regression analysis) 이론의 근거가 되는 가정, 흔히 ε 들에 주어져야만 하는 매우 중요한 가정(assumption)을 대략적으로 설명한다.

식(1.1)의 모형은 모수(parameter)인 β 에 대하여 선형이므로 선형모형(*linear models*)으로 분류된다. 모든 회귀(regression) 과정에는 자료 세트(set of data)에 모형을 적합(fit)하는 것이 포함된다. 자료 세트란 표본추출 된 다양한 실험단위들(experimental units, 예를 들어, 산소소모 상황에서 표본추출 된 개개인들)에서 변수(variables)에 대한 기록(readings)으로 정의된다. ‘자료에 적합한다’(*fit to a set of data*)는 말은 회귀계수(*regression coefficients*)와 적합된 회귀모형(*fitted regression model*)의 식(formulation)을 추정(estimation)하는 것이다. 여기서 적합된 회귀모형이란 통계적 추론(statistical inference)의 근거가 되는 경험적 장치(empirical device)이다. 적합의 질(quality of fit)에 대한 측도(measures)는 통계 분석의 기초를 형성하는 중요한 통계량(statistics)이다. 만약 가정된 모형(postulated model)이 자료를 만족스럽게 기술하지 않는다면 적합된 모형(fitted model)에서 얻어진 어떠한 결론도 믿을 수 없다.

이 책의 범위는 선형모형(*linear model*)에만 국한하지 않는다. 비선형모형(*nonlinear*

models)은 자연과학(natural sciences)이나 공학분야(engineering applications)에서 흔하게 볼 수 있다. 생화학자가 시간에 따르는 특정 세균의 성장을 기술하는데 다음과 같은 유형의 성장모형(growth model)을 가정한다고 하자.

$$y = \frac{\alpha}{1 + e^{\beta t}} + \varepsilon \quad (1.2)$$

특정 세균의 성장 y 는 시간 t 의 함수로 나타낸다. 여기에서 모수(parameters)인 α 와 β 는 자료로부터 추정된다. 선형회귀(linear regression)를 이용하여 해결하려던 동일한 문제들의 대부분을 종종 비선형회귀(nonlinear regression)로 다룰 수 있다. 그러나 비선형모형(nonlinear models)을 구축하는 컴퓨터인 측면은 그리 간단하기 않으며 특별한 방법을 필요로 한다(9장을 참조하시오).

적합된 회귀모형(fitted regression model)은, 자료를 설명해 주는 함수 관계(functional relationship)에 대한 추정값(estimate)이다. 가정된 모형(postulated model)의 종류는 자료에서도 출되는 회귀변수(regressor variable)들의 범위에 따라 결정된다. 예를 들어, 화학기술자는 모형이 곡선(curvature)이 되도록 해주는 항(term)이 있어야 하는 시스템을 알고 있을 수 있다. x_1 과 x_2 가 온도와 반응물의 농도(reactant concentration)를 나타내고, 반응 y 가 화학 반응의 단순 수율(simple yield)이라고 가정해 보자. 이 기술자의 목표는 아마도 다음의 두 가지일 것이다.

1. 자료 범위 밖에 있는 x_1 과 x_2 위치에서의 수율(yield)을 추정(estimation) 해줄 수 있는 적합된(fitted) 혹은 추정된(estimated) 회귀식(예측 공식, prediction equation),
2. 만족스러운 수율(yield)을 보이는 온도와 농도 조건을 발견하기 위하여, 자료의 범위 내에서의 관계(relationship) 연구.

모형 곡선(model curvature)을 나타내기 위하여, 다음과 같은 구조식을 사용한다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (1.3)$$

그러므로 회귀 변수(regressor variables)의 범위(range)는 모형의 종류에 영향을 미친다. x_1 과 x_2 의 범위가 좁으면 2차항(quadratic term)을 포함하지 않는 모형을 사용하는 것이 성공적일 것이다. 식(1.3)은 회귀 변수에는 거듭제곱(powers)과 2차수의 곱(products of order two)이 포함되지만, 계수(coefficients)가 선형이므로 선형모형(linear model)이다.

선형모형으로 자료 세트(data set)를 기술하는 거의 모든 회귀(regression) 분석에서 모형식(model formulation)이 자료의 관찰과정(data observational process)에서 지나치게 단순화된다. 특히 사회(social) 및 행동과학(behavioral science) 분야는, 자료를 얻은 시스템이 너무 복잡하여 정확한 구조(structure)를 가진 모형을 얻기가 매우 어렵다. 선형모형(linear model)은,

모형구축(model building)에 사용된 자료 범위 내에서 잘 작동하는 근사치(approximation)이다. 여기서 근사적으로 선형모형을 만드는 사람들을 비난하려는 의도는 없다. 관련분야 영역(subject matter field)이 잘 작동하는 이론을 제공하기에는 충분히 정교하지 않을 경우, 선형이면서 상식적인(common sense) 경험적 모형 접근법(empirical model approach)은, 특히 자료의 질이 합리적이라면, 훌륭한 정보를 제공할 수 있다.

1.2. 회귀분석의 정식 사용(Formal Uses of Regression Analysis)

이 책의 많은 부분이 특정 종류의 추론들(inferences)을 대표하는 범주(categories)에 할당되었으며, 이들 간의 차이를 반드시 기술하였다. 이는 추정 과정(estimation procedure)이나 심지어 모형의 채택까지도 연구 목적에 따라 달라질 수 있기 때문에 매우 중요하다. 이것은 방법론적인 측면에서 일부 사용자의 생각에 종종 반하는 것처럼 보인다. 경험이 없는 분석가에게는, 곁보기에 자료를 가장 잘 기술하는 모형만을 채택하여야 하는 것처럼 보일 수도 있다. 그러나 한 문제에 대해 만족스러운 해답을 주는 모형은 다른 문제를 해결하는데 있어서 반드시 성공적이지는 않다. 다소 중복되는 부분이 있긴 하지만, 회귀분석(regression analysis)의 목적을 다음의 서너 가지로 요약할 수 있다.

1. 예측(prediction)
2. 변수 선별(variable screening)
3. 모형 설정(model specification): 시스템 설명(system explanation)
4. 모수 추정(parameter estimation)

분석가는 분석을 시행할 때 목적이 무엇인지를 정확하게 아는 것이 중요하다. 처음에는 우선 목적을 고려하자. 단순히 모수 추정값(parameter estimates)만을 구하는 것이 목적이 아니다. 우리는 함수의 형태(functional form)가 예측(prediction)에 어떤 영향을 미치는가와 무관한 모형은 설정(model specification)하지 않는다. 모형에서 개별 회귀변수의 역할을 아주 정밀하게 알아내는 것 또한 중요하지 않다. 앞서의 화학 반응 예는 예측의 문제(prediction problem)이다. 따라서 반응수율(reaction yield)이 중요하며, 기술자는 그것을 적절하게 예측할 필요가 있다.

위에서 언급한 목적은 생각보다 많은 실제 적용례에서 적절하다. 모형의 식(model formulation)은 이차적인 것이며, 반응의 변동(variation)를 설명하는데 있어서 각 변수(variable)의 중요도를 알아내는 도구 정도로만 사용된다. 합리적인 정도의 반응 변동량(amount of variation)을 설명해주는 변수(variable)는 모형에 그대로 둔다. 역할이 적은 것으로 보이는 회귀변수는 제거한다. 이러한 과정은 더 상세한 연구나 모형 구축 과정(model-building process)에 앞서 종종 시행된다. 예를 들어, 담배 화학자(tobacco chemist)가 여러 가지 담배 제형(formulation)마다 향의 유형별 등급(taste-type rating)을 매기는 실험을 한다고 하자. 회귀 변수는 첨가물들의 농도이다. 적합된 모형으로부터 어떤 구성성분이 향에 영향을 미치는지 결정하기 위한 단일 목적만으로, 모형은 자료에 적합된다. 이런 역할을 하는 것으로 보이는 변수는 실험변수(experimental variables)로 계속 유지된다.

모형설정(model specification)은 그 자체로 설명된다. 분석가는 모형을 가정(postulation)하는데에 많은 신경을 써야 한다. 후보모형들(candidate models)은 종종 서로

경쟁적(competitive)이면서 함수형태(functional forms)도 다르다는 것을 모든 분석가들이 직, 간접적으로 알게 될 것이다. 개개의 함수형태는 회귀변수의 역할을 각각 다르게 정의한다. 모형이 선형적일 때 개별 회귀변수의 양상(complexion)이 자료 내에서 잘 정의되지 않는다면, 그 모형은 실패할 수 있다.

모수 추정(parameter estimation)은 종종 어떤 과학분야에서 회귀분석(regression analysis)을 시행하는 유일한 목적이 된다. 8장을 보면, 농업생산기능(agriculture production function)이 비용(expenditure)을 대변하는 일련의 회귀변수 세트(a set of input regressors)에 적합되는 자료 세트가 나온다. 비용의 종류별로 여섯 가지의 회귀변수와 한 개의 소나기 변수(rainfall variable)가 사용된다. 자료 수집 단위는 일년이며, 버지니아주에서 25년치 자료를 수집하였다. 이 25개의 자료 포인트(data points)은 선형모형으로 적합하며, 여기서 예측(prediction)과 변수 선별(variable screening)은 하나도 중요하지 않다. 그러나 회귀계수(regression coefficients)의 특정 범위(specific range)는 특별한 경제이론을 뒷받침해주며, 따라서 계수(coefficients)의 부호(sign)와 크기(magnitude)가 중요하다.

1.3. 데이터베이스(Data Base)

자료분석에서 상투적으로 “Garbage In-Garbage Out”이란 말을 많이 하는데, 이는 “자료가 좋아야 결과도 좋다”는 말로써 회귀모형(regression models) 구축에도 적용된다. 만약 자료가 변수의 경향을 반영하지 못한다면, 모형 구축이나 시스템에 대한 추론 도출에 성공할 수 없다. 자료와 변수 간에 약간의 연관성(association)을 보인다 하더라도, 이것만으로는 자료가 명확하게 검출 가능한 방식으로 그 연관성을 밝힌다고 말할 수 없다. 자료는 설계된 실험(designed experiment), 잘 개발된 조사(well developed survey), 시간 경과에 따른 자료수집(collection)과 제표(tabulation), 컴퓨터 모의실험(computer simulation) 등을 포함한 많은 출처로부터 얻을 수 있다. 표본의 크기(sample size)가 매우 중요하다는 것은 자명한 사실이다. 표본의 크기가 너무 작으면, 회귀 결과(regression results)로부터 오차(error)에 대한 적절한 척도(adequate measures of error)를 계산할 수 없으므로 모형 가정(model assumption)에 대한 검토를 할 수가 없다. 그러나 자료 세트의 크기가 유일한 고려 대상은 아니다. 자료 세트와 관련된 문제 중 많은 것들은 매우 명백하다. 예를 들어, 실험 단위(experimental unit)의 표본(sample)이, 모형화하려는 모집단(population)을 대표하지 않는다면, 우리는 폭넓은 관계(broad based relationship)를 보여주는 모형을 구축할 수 없다. 자료가 너무 특이적(too specific)이라면, 회귀분석(regression analysis)의 어떠한 목적에 대해서도 폭넓게 일반화(broad generalization) 될 수 없다. 때때로, 자료 내의 회귀변수 범위(ranges of regressor variables)로 인하여, 만들어진 모형과 도출된 결론이 자료 특이적(data specific)일 수 있다. 산소소모율에 대한 예에서, 실험에 참여하는 모든 사람이 건강상태가 좋은 운동선수라고 한다면, 변수(variables)의 범위와 추론(inferences)의 다른 특성들(characteristics) 모두가 이 모집단(population)에만 적용 가능하기 십상이다.

회귀분석(regression analysis)이 어려운 것은, 많은 경우 하나 이상의 가정 실패(failure of one or more assumptions) 때문이다. 특히 식(1.1)의 다중선형회귀모형(multiple linear regression model)은 회귀변수가 오차(error) 없이 측정되었다는 가정 하에서 분석되었다. 만약 회귀변수에 과도한 측정오차(measurement error)가 있다면, 회귀계수(regression coefficients)의 추정값(estimates)은 크게 영향을 받으므로, 예측(prediction), 변수선별(variable screening) 등과 같은 다른 추론들(inferences)이 불확실해진다.

아마도 회귀 자료 세트(regression data set)에서 발생 가능한 가장 심각한 문제는 모든 가능한 중요한 회귀변수들에 대하여 자료수집이 안된 경우일 것이다. 이러한 문제는 분석가가 적절한 회귀변수가 어떤 것인지 다 알지는 못할 경우에 발생한다. 심지어 전체 또는 대부분의 양(quantities)이 확인이 되었다 하더라도, 자료 수집 과정의 제한 때문에 이들이 측정되지 못할 수도 있다. 이런 일이 발생한다면 모형은 심하게 저설정(under-specification)될 수 밖에 없으므로, 회귀계수(regression coefficients)의 추정값(estimates)과 예측(prediction)이 나빠진다.

2. 단순선형회귀모형(Simple Linear Regression Model)

2.1. 모형의 기술(Model Description)

가장 간단한 회귀구조(regression structure)에 적용 가능한 모형은 단순선형회귀모형(simple linear regression model)이다. 여기서 단순(simple)하다는 것은 회귀변수(regressor variable, x)가 하나라는 것을 의미하며, 선형(linear)이라는 것은 x 를 기준으로 선형을 의미한다. 이러한 의미에서 단순선형회귀모형(simple linear regression model)을 표현하면 다음과 같다.

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (2.1)$$

여기에서 y 는 측정된 반응변수(response variable)이며, β_0 와 β_1 은 각각 절편(intercept)과 기울기(slope)이고 ε 는 모형의 오차(error)이다. 통상 관찰값의 쌍 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 이라는 자료 세트(data set)에서 식(2.1)의 모수(parameter) β_0 와 β_1 의 추정값(estimate)을 구하므로, 위 식은 아래와 같이 쓸 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (2.2)$$

2.3에서 최소제곱법(method of least squares)을 이용하여 β_0 와 β_1 을 추정(estimation)하는 것에 대하여 논의할 것이며, 최소제곱추정량(least squares estimator)이 어떤 조건하에서 이상적(ideal)인지 알아볼 것이다.

2.2. 모형모수에 대한 가정과 해석(Assumptions and Interpretation of Model Parameters)

회귀분석(regression analysis)에서 모형 식(model formulation)의 목적은 관측치들(observations)이 어떻게 생성(generation)되는가를 개념화하는데 있다. 또한, 이 식(formulation)을 이용하여 모수추정량(parameter estimator)에 대한 특성(property) 연구도 할 수 있다.

최소제곱법의 기본 가정(assumptions)은 중요하며, 반드시 알아 두어야 한다. 첫째로, x_i 는 비임의(nonrandom)이며 오차(error)가 거의 없는 관측치(observation)인 반면, ε_i 는 평균(mean)이 0이고 분산(variance)이 σ^2 으로 일정한 확률변수(random variable)로 가정하자(등분산가정, homogenous variance assumption). 모집단 평균(population mean)을 나타내는 기대 연산자(expectation operator)를 사용하면, $E(\varepsilon_i) = 0$ 이며 $\text{Var}(\varepsilon_i) = E[(\varepsilon_i - 0)^2] = E(\varepsilon_i^2) = \sigma^2$ 로 표현 가능하다. 또한, ε_i 는 관측치(observation)간에 서로 상관되어(correlated) 있지 않다고 가정한다. 물론 대부분의 실제 상황에서는 x_i 가 어느 정도는 임의변동(random variation)이 있을 수 있다. 그렇다 하더라도 x 의 측정 범위에 비하여 그것의 임의변동은 무시할만 하고 x_i 의 측정오차(measurement error) 또한 x 의 범위에 비하여 작다고 가정하자. x 와 y 모두 확률변수(random variable)인 경우는 나중에 다루기로 한다. 또한, x_i 의 측정오차(measurement error)가 미치는 효과는 7장에서 다루기로 한다.

식(2.2)의 단순선형회귀모형(simple linear regression model)은 x 의 특정값인 x_i 에서 y_i 분포의 평균으로 아래와 같이 표현가능하며,

$$E(y_i) = \beta_0 + \beta_1 x_i$$

이 때, 분포의 분산(variance of the distribution)은 σ^2 으로 x 의 수준(level)과는 독립적이다. 그러므로 식(2.2)에 의한 선형관계(linear relationship)는 평균반응(mean response)이 x 와 연관되어 있다는 것을 의미한다. 이런 이유로, 종종 다음과 같이 표기하기도 한다.

$$E(y|x) = \beta_0 + \beta_1 x$$

여기서 $E(y|x)$ 는 특정한 값 x 가 주어졌을 때 y 의 조건부평균(mean y conditional on a specific value of x)이라고 한다.

모형과 모형모수에 대한 추가적인 기술(Additional Description of Model and Model

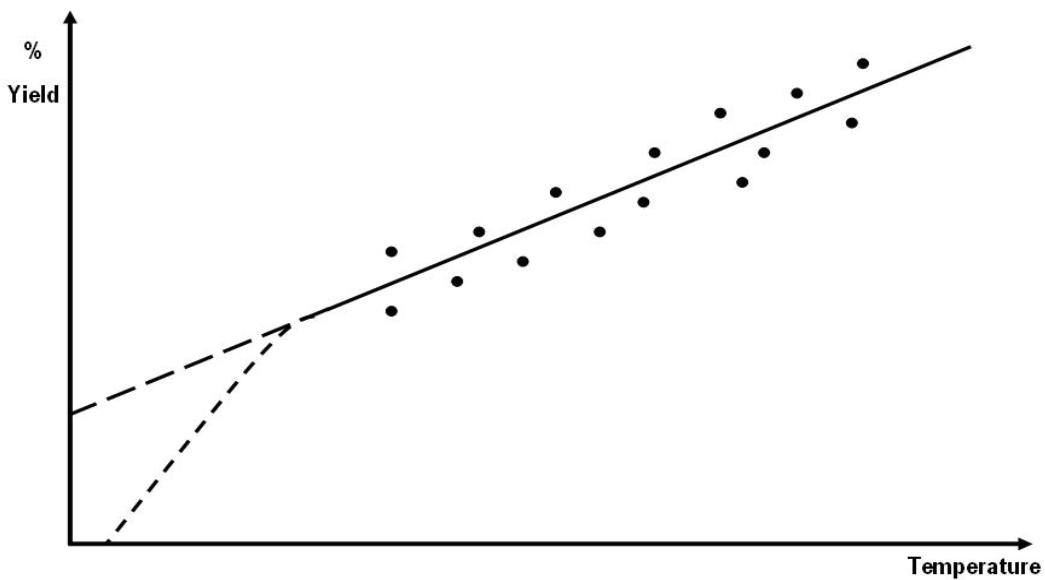
Parameters)

회귀분석에서 자료를 분석하는데 있어 중요한 것은 회귀모형의 절편(intercept)인 모수 β_0 와 회귀직선(regression line)의 기울기(slope)인 모수 β_1 을 추정하는 것이다. 이 모수들(parameters)을 회귀계수(regression coefficients)라고 하며, 이들의 추정값(estimates)은 전체 분석 과정에서 중요한 의미를 가진다. 기울기 β_1 은 x 의 단위 증가당 $E(y)$ 의 변화이다. 그러므로 β_0 와 β_1 의 값(그리고 추정값의 값)은 x 와 y 의 단위에 따라 달라지므로, 분석 시 추정값(estimates)을 해석할 때 이를 고려하여야 한다.

실험자들이 실험 자료로부터 얻어진 β_0 와 β_1 의 추정값(estimate)을 해석할 때에 혼란스러워질 수도 있다. 이 장의 초반에 지적하였듯이, 선형모형(linear model)은 자료가 어떻게 나왔는지를 설명하는 가장 간단한 경험적인 방법(empirical device)이다. 자료로부터 계산되는 모형의 특성들은 x 의 범위(the range of x)에 매우 의존적이다. 즉, 식(2.2)의 모형은 x 의 제한된 범위 내에서만 유효하다는 것을 전제로 한다. 이 범위 안에 0이 있다면 β_0 의 추정값(estimate)은 $x = 0$ 일 때의 y 의 평균으로 확실하게 해석(interpretation)된다. 그러나 자료의 범위가 원점(origin)으로부터 떨어져 있다면, β_0 는 해석(interpretation)의 여지가 별로 없는 회귀항(regression term)일 뿐이다. 종종 β_0 의 추정값(estimate)은 비현실적이거나 문제의 상황에서 도저히 받아들일 수 없는 수치로 판명될 수도 있다. 이 경우에 β_0 를 해석하는 것은 원래 적용하려던 범위를 벗어난 값에서 모형을 외삽(extrapolation)하는 것과 같다고 생각하면 될 것이다.

Fig 2.1에서 자료 세트(data set)는 온도(x)에 대한 화학 반응 y 의 수율(yield)을 측정한 것이다. 측정치(measurements)는 15개이다. 자료 범위 내의 온도에서는, 온도에 대하여 선형적인 모형이 매우 합리적인 것으로 보인다. 온도와 수율이 곡선관계(curvilinear relationship)를 보이는 점곡선(dotted curve)이 수율의 진짜 기대값(true expected value)이라고 하자. 실선을 온도가 0인 곳까지 연장시키면(dashed line), y 는 17%가 된다. 물론 이 수치가 합당한 β_0 의 추정값(estimate)일 수도 있다. 그러나 온도가 20°C 이하에서는 화학반응이 일어나지 않는다는 것을 미리 알고 있다면, 실제 곡선관계가 있다는 것을 몰랐다고 하더라도, 이 수치는 전혀 과학적인 의미를 가지지 못한다는 것을 알 수 있다. 따라서 0°C에서의 수율(yield)까지 확장시킬 필요가 없었던 것이다. 그러므로 이 경우 절편은 80 - 220°C 사이에서 일어나는 일들을 기술할 수 있는 기전(mechanism)의 일부일 뿐이다.

Figure 2.1 Reaction yield and temperature data



정규성 가정(Normal Theory Assumption)

어떠한 통계분석과정이라도 그에 적용되는 기본 가정(underlying assumptions)에 대해서는 반드시 알아두어야 할 필요가 있다. 회귀분석에 대한 기본가정을 완전히 이해하고 인식하기 위해서는 추정량(estimator)의 어떤 특성이 어떤 가정(assumptions)에 기초한 것인지를 분명하게 알아야 한다. 예를 들어, 추정량(estimator)의 편향(bias), 분산(variance), 공분산(covariance)의 특성들이 여기에서 기술되는 어떠한 가정들에 의존하는지 순차적으로 고찰해 볼 것이다. 후반부에서는 가설검증(hypothesis testing)에 대하여 고찰할 것이며, 이러한 추론 과정(inferential procedures)은 모형의 오차(ε_i)가 가우스 분포(Gaussian distribution)를 따른다는 정규성 이론 가정(normal theory assumption)에 기초한다.

대안모형 설정: 중심화모형(Alternative Model Formulations: Centered Model)

식(2.2)의 단순선형회귀모형(simple linear regression model)은 조작(manipulation)의 편이성과 좀 더 명확한 모수 해석(parameter interpretation)을 위해 다른 형태로 쓰여지기도 한다. 여러 가지 방식으로 쓰여진 선형회귀모형을 볼 수 있지만, 여기에서는 중심화 단순선형모형(centered simple linear model)을 소개하겠다..

단순선형회귀모형에서, $\bar{x} = \sum_{i=1}^n \frac{x_i}{n}$, 즉 회귀변수의 평균이라고 하자. 중심화(centering)라

는 것은 x_i 가 모형의 중심에 들어 간다는 것을 의미한다. 즉, 다음과 같이 표현된다.

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (2.3)$$

(2.3)의 모형은 x_i 가 원점으로 이동되었다는 것을 의미한다. 이 경우 β_0^* 는 중심화 회귀변수 (centered regressor), $(x_i - \bar{x})$ 가 0일 때의 $E(y)$ 이므로, 절편은 아래와 같이 된다.

$$\beta_0^* = \beta_0 + \beta_1 \bar{x}$$

이제는 중심화모형(centered model)과 비중심화모형(uncentered model)이 동일한지 판단하기가 쉬워진다. 2.3절에서는 β_0 와 β_1 의 추정량(estimator)을 얻기 위해 최소제곱법(method of least squares)을 이용할 것이다. 이 때, 중심화모형(centered model)을 사용하면 이 과정이 간단해진다. 이 책 전체에서 식 (2.2)와 (2.3)의 모형을 혼용할 것이다. 중심화 회귀변수(centered regressor)라는 용어는 8장에서 공선성 진단(collinearity diagnostics)을 고찰할 때 접하게 될 것이다.

2.3. 최소제곱공식(Least Squares Formulation)

최소제곱법(method of least squares)은 회귀분석에서 다른 추정방식(estimation procedure)보다 많이 사용되고 있으며 1970년대 이전에는 거의 유일하게 사용되었다. 물론 7장과 8장에서는 최소제곱법(least squares)을 대신 할 수 있는 중요한 다른 방법들을 고찰하게 될 것이다. 최소제곱법을 이용하여 잔차제곱합(residual sum of squares), $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 을 최소로 하는 b_0 와 b_1 의 추정량(estimator)인 b_0 와 b_1 을 얻을 수 있고, 이때 반응의 적합값(fitted value)인 \hat{y}_i 는 다음과 같다.

$$\hat{y}_i = b_0 + b_1 x_i$$

따라서, b_0 와 b_1 은 다음 식을 만족한다.

$$\frac{\partial}{\partial b_0} \left[\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \right] = 0 \quad (2.4)$$

$$\frac{\partial}{\partial b_1} \left[\sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \right] = 0 \quad (2.5)$$

잔차제곱합(residual sum of squares) $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 을 최소화시키는 것은 좋은 논리적 근거를 가지고 있으며, 이렇게 얻어진 b_0 와 b_1 값은 대부분의 경우에 좋은 특성을 가진다. 즉, 잔차제곱합(residual sum of squares)을 최소화시킴으로써 적합오차(errors in fit)인 잔차(residual)가 작아지면, 자료들이 적합된 회귀선(fitted regression line, $\hat{y} = b_0 + b_1 x$)에 가까워진다. Fig 2.2는 이를 설명하고 있으며, 여기서 수직 편차(vertical deviation)가 잔차(residual)를 의미한다. 최소제곱추정량(least squares estimators)의 수식전개는 식(2.3)의 중심화모형(centered model)을 사용하면 쉬워진다. 즉, 아래의 두 식을 생각하자.

$$\frac{\partial}{\partial b_0^*} \left\{ \sum_{i=1}^n [y_i - b_0^* - b_1(x_i - \bar{x})]^2 \right\}$$

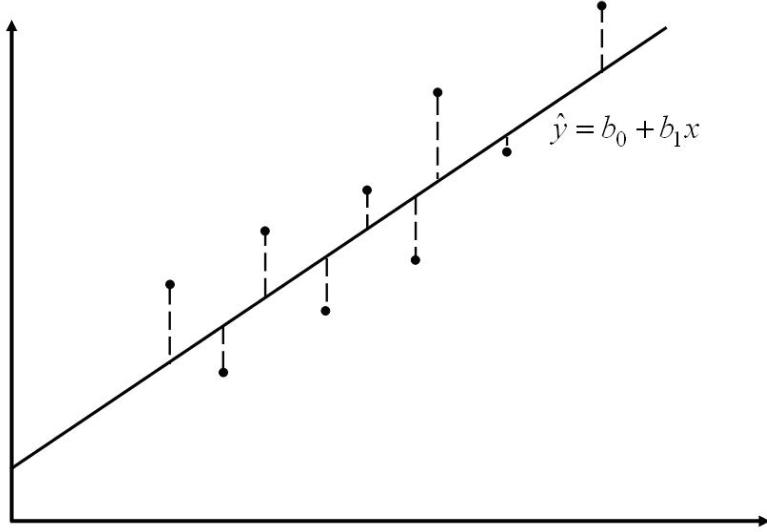
$$\frac{\partial}{\partial b_1} \left\{ \sum_{i=1}^n [y_i - b_0^* - b_1(x_i - \bar{x})]^2 \right\}$$

이 도함수(derivatives)들을 풀어보면, 다음과 같은 방정식을 얻을 수 있다.

$$nb_0 + b_1 \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n (x_i - \bar{x}) + b_1 \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n y_i (x_i - \bar{x})$$

Figure 2.2 Illustration of least squares residuals



$\sum_{i=1}^n (x_i - \bar{x}) = 0$ 이므로, 최소제곱 정규방정식(least squares normal equation)의 해는 다음과 같이

주어진다.

$b_1 = \frac{S_{xy}}{S_{xx}}$ $b_0^* = \bar{y}$	(2.6)
---	---

여기에서 $S_{xy} = \sum_{i=1}^n (x_i - \bar{x}) y_i$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ 이다.

이제, 비중심화모형(uncentered model)의 절편인 β_0 의 추정량(estimator)을 구하기 위해서 유념해야 할 것은

$$\beta_0 = \beta_0^* - \beta_1 \bar{x}$$

이고, 따라서

$$b_0 = \bar{y} - b_1 \bar{x}$$

이다. 따라서, 절편과 기울기 추정량(estimator)은 각각 아래와 같다.

$$b_0 = \bar{y} - b_1 \bar{x} \quad (2.7)$$

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad (2.8)$$

최소제곱법을 통해 식(2.7)과 (2.8)의 기울기와 절편의 추정값(estimates) 이외에 잔차(residuals), 즉, $y_i - \hat{y}_i$ 와 적합된 회귀선(fitted least squares line) $\hat{y} = b_0 + b_1 x$ 도 얻을 수 있다.

많은 경우, 도출된 결론은 추정된 기울기와 절편을 중심으로 전개되기 때문에 추정량(estimator)의 통계적 특성을 반드시 알고 있어야 한다.

추정량의 특성(Properties of the Estimators)

x_i 가 비임의(nonrandom)이고 $E(\varepsilon_i) = 0$ 이라는 가정 하에서 회귀계수 추정량들(estimators)이 불편추정량(unbiased estimators)이라는 것은 간단히 증명할 수 있다. 먼저 기울기 추정량 b_1 의 기대값은 아래와 같이 구할 수 있으며,

$$\begin{aligned} E\left(\frac{S_{xy}}{S_{xx}}\right) &= \sum_{i=1}^n \frac{(x_i - \bar{x})(E(y_i))}{S_{xx}} \\ &= \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i) \\ &= \frac{\beta_1 S_{xx}}{S_{xx}} \\ &= \beta_1 \end{aligned}$$

마찬가지로, 절편 추정량 b_0 의 기대값은 다음과 같다.

$$\begin{aligned}
E(b_0) &= E[\bar{y} - b_1 \bar{x}] \\
&= \frac{1}{n} E\left(\sum_{i=1}^n y_i\right) - \beta_1 \bar{x} \\
&= \frac{1}{n} \sum_{i=1}^n [\beta_0 + \beta_1 x_i] - \beta_1 \bar{x} \\
&= \beta_0
\end{aligned}$$

최소제곱추정량(least squares estimators)들의 분산(variance) 특성(properties)에 있어 주목해야 하는 것은 등분산가정(homogeneous variance assumption) 즉, $x = x_i$ 일 때 $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2$ 으로, 모형오차분산(model error variance)이 회귀변수(regressor variable)의 고정값들(fixed values)에 대해서는 일정하다는 것이다. 따라서 식(2.7)에서 b_1 의 분산(variance)을 다음과 같이 얻을 수 있다.

$$\text{Var}(b_1) = \frac{1}{S_{xx}^2} \sum_{i=1}^n \sigma^2 (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}} \quad (2.9)$$

절편의 경우, $\bar{y} = b_0^*$ 와 b_1 은 상관관계가 없다는 것을 쉽게 보일 수 있다. 먼저

$$\bar{y} = \beta_0^* + \bar{\varepsilon}$$

여기에서, $\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i$ 이다. 다음으로

$$\begin{aligned}
b_1 &= \sum_{i=1}^n \frac{(x_i - \bar{x}) [\beta_0^* + \beta_1 (x_i - \bar{x}) + \varepsilon_i]}{S_{xx}} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}
\end{aligned}$$

따라서,

$$\text{Cov}(\bar{y}, b_1) = \text{Cov}\left[\beta_0^* + \bar{\varepsilon}, \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) \varepsilon_i}{S_{xx}}\right]$$

을 구할 수 있다. 여기에 ε_i 는 무상관(uncorrelated)이며 평균이 0, 분산이 σ^2 이라는 가정을 이용하면,

$$\begin{aligned} Cov(\bar{y}, b_1) &= E[\beta_0^* + \bar{\varepsilon}] \left[\beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{S_{xx}} \right] - \beta_0^* \beta_1 \\ &= \beta_0^* \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})\sigma^2}{nS_{xx}} - \beta_0^* \beta_1 \\ &= 0 \end{aligned}$$

위의 사실을 이용하면 b_0 의 분산을 다음과 같이 얻을 수 있다.

$$Var(b_0) = Var\left(\frac{\sum_{i=1}^n y_i}{n}\right) + \bar{x}^2 Var(b_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad (2.10)$$

식(2.9)과 (2.10)의 분산으로 유도되는 전개과정에서 등분산가정(homogeneous variance assumption) 뿐만 아니라 ε_i 가 서로 무상관(uncorrelated)이라는 가정을 사용하였다. 또한, 이 결과는 x 가 비임의(nonrandom)라는 조건에 기초한다. 가설검증을 다루는 후반부에서, 기울기와 절편 모두의 추정표준오차(estimated standard error)가 필요하게 되며, 이때 (2.9)와 (2.10)이 매우 유용하게 사용된다.

최소제곱추정량(least squares estimator)의 비편향성(unbiasedness)과 분산 특성(variance property)은 다른 접근 방법을 통해서도 전개 가능하다. b_1 을 달리 표현하면 다음과 같다.

$$b_1 = \beta_1 + \sum_{i=1}^n d_i \varepsilon_i$$

여기에서 $d_i = \frac{x_i - \bar{x}}{S_{xx}}$ 이다.

위의 식에서 $E(\varepsilon_i) = 0$ 이므로 b_1 의 비편향성(unbiasedness)과 식(2.9)에서 b_1 의 분산을 쉽게 구할 수 있다. 마찬가지로, 절편의 경우에도 다음과 같이 표현할 수 있다.

$$b_0 = \beta_0 + \bar{\varepsilon} - \bar{x} \sum_{i=1}^n d_i \varepsilon_i$$

따라서, 식(2.10)의 비편향성(unbiasedness)과 분산도 쉽게 구할 수 있다.

예제 2.1 부부간 키의 성향 분석

Table 2.1은 부부간의 키의 성향을 알아보기 위해 최근 결혼한 사람들을 무작위로 선정하여 남편과 아내의 키를 조사한 것이다. 남편의 키를 독립변수(H)로 하고 아내의 키를 종속변수(W)로 하여 식(2.2)의 모형에 적용하여 적합한 결과 절편과 기울기의 추정값(estimate)(식(2.7)과 (2.8))은 아래와 같다.

$$\begin{aligned} b_0 &= 41.9302 \\ b_1 &= 0.6997 \end{aligned}$$

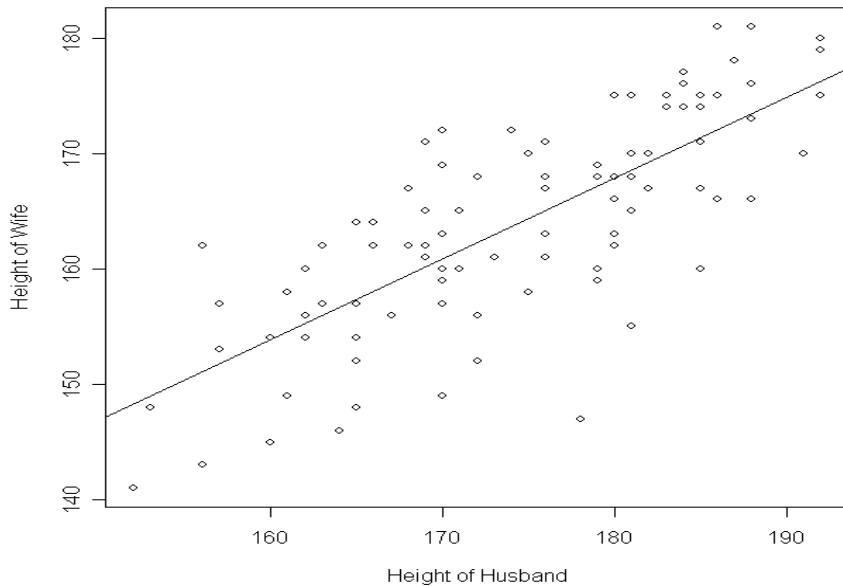
Fig 2.3은 자료와 최소제곱법에 의해 적합된 회귀선을 나타낸 것이다.

Table 2.1 남편(H)와 아내(W)의 키(Height) 자료(단위 : cm)

Row	H	W	Row	H	W	Row	H	W
1	186	175	33	180	166	65	181	175
2	180	168	34	188	181	66	170	169
3	160	154	35	153	148	67	161	149
4	186	166	36	179	169	68	188	176
5	163	162	37	175	170	69	181	165
6	172	152	38	165	157	70	156	143
7	192	179	39	156	162	71	161	158
8	170	163	40	185	174	72	152	141
9	174	172	41	172	168	73	179	160
10	191	170	42	166	162	74	170	149
11	182	170	43	179	159	75	170	160
12	178	147	44	181	155	76	165	148
13	181	165	45	176	171	77	165	154
14	168	162	46	170	159	78	169	171
15	162	154	47	165	164	79	171	165
16	188	166	48	183	175	80	192	175
17	168	167	49	162	156	81	176	161
18	183	174	50	192	180	82	168	162

19	188	173	51	185	167	83	169	162
20	166	164	52	163	157	84	184	176
21	180	163	53	185	167	85	171	160
22	176	163	54	170	157	86	161	158
23	185	171	55	176	168	87	185	175
24	169	161	56	176	167	88	184	174
25	182	167	57	160	145	89	179	168
26	162	160	58	167	156	90	184	177
27	169	165	59	157	153	91	175	158
28	176	167	60	180	162	92	173	161
29	180	175	61	172	156	93	164	146
30	157	157	62	184	174	94	181	168
31	170	172	63	185	160	95	187	178
32	186	181	64	165	152	96	181	170

그림 2.3 데이터와 적합된 회귀선



다음은 예제에서 사용한 R code이다.

```
#exmpla 2.1
data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
plot(H,W,xlab="Height of Husband",ylab="Height of Wife",main="그림 2.3 데이터와 적합된 회귀선")
abline(g,lty=1)

s_stat<-summary(g)
```

다른 형태의 추정량과의 비교(Comparison to Other Forms of Estimators)

지금까지 식 2.2의 단순선형회귀모형(simple linear regression model)에서 특정한 경우(specific case)에 최소제곱 추정량(least squares estimator)의 평균과 분산을 살펴보았다. 보다 일반적인 모형에서의 최소제곱추정량(least squares estimator)의 특성에 대해서는 설명하지 않겠다. 또한 이 장에서는 더 이상 최소제곱추정량(least squares estimator)이 다른 형태의 추정(other forms of estimation)과 어떻게 다른지에 관해서 설명하지 않을 것이다. 이 장에서 특별히 설명하고자 하는 것은 최소제곱계수의 분산(variance of the least squares coefficient)과 다른 비편향 형태의 추정(an alternative unbiased form of estimation)을 이용한 계수(coefficient)의 분산(variance)과

비교하는 것이다. 모형오차(model error)에 관해서 앞서 언급되었던 가정(assumption) 하에, 다양한 부류의 모형(a wide class of models)에서 최소제곱추정량(least squares estimator)은 최상의 특성(optimal properties)을 가진다. 이러한 최상의 특성(optimal properties)에 대해서 3장에서 좀더 일반적인 선형회귀모형(linear regression model)과 연관하여 설명할 것이고, 모형의 일반적인 특성(nature)을 좀더 이해할 때까지 설명을 미루도록 하겠다.

오차분산의 추정, 잔차의 자유도(Estimation of Error Variance, Residual Degrees of Freedom)

실제 상황에서는 오차분산(σ^2)의 추정값(estimate of the error variance)이 요구된다. 이 추정값(estimate)은 가설검정(hypothesis testing)을 위해 회귀계수 표준오차의 추정값(estimated standard errors of coefficient)을 계산하거나, 많은 경우 회귀모형(regression model) $\hat{y} = b_0 + b_1x$ 의 예측능력(prediction capability)이나 적합의 질(quality of fit)을 평가하는데 있어서 주요한 역할을하게 된다(4장 참조). 추정량(estimator)이 어떤 것인가를 직관적으로 보여주는 것은 어렵지 않다. 잔차(residuals), $y_i - \hat{y}_i$, 즉 관찰된 적합 오차(observed errors of fit)는 명백히 ε_i , 즉 관찰되지 않은 모형오차(model error)와 경험적으로 상반되는 개념이다. 그러므로 잔차(residual)의 표본 분산(sample variance)이 σ^2 의 추정량(estimator)을 제공해야 하는 것은 당연하다. 만일 적절한 분모 $n-2$ 로 잔차제곱합(residual sum of squares, $\sum_{i=1}^n (y_i - \hat{y}_i)^2$)을 나누면 불편추정량(unbiased estimator)이 얻어진다. 그래서 σ^2 의 추정량(estimator)을 다음과 같이 정의한다.

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$$

이 값 s^2 를 흔히 오차평균제곱(error mean square)이라고 한다. 강조해야 할 것은 모형이 정확하다는 가정하에 s^2 이 편향이 없다는 사실이다. 모형이 적절하지 않을 때의 s^2 에 대한 편향(bias)은 3장과 4장에서 논의될 것이다. 분모(denominator)는 종종 오차자유도 혹은 잔차자유도(error or residual degrees of freedom)라고 한다. 이 책에서 종종 자유도를 df로 표기한다. 이런 용어를 사용하는 이유는 다음 절에서 이 통계량(statistic)의 분포의 특성에 관하여 설명할 때 좀더 명확해질 것이다. 여기에서는 회귀모형에서 잔차의 자유도(residual degrees of freedom)는 자료의 수(the number of data points) n에서 추정된 모수의 수(the number of parameters estimated)를 뺀 것이라는 정도만 알아두면 될 것이다. (두 개의 모수는 기울기와 절편이다.) 예제 2.1에서 추정값(estimate) s^2 은 다음과 같다.

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{3303.3}{94} = 35.1$$

이제 분산을 추정할 때 σ^2 대신 모형오차분산(model error variance)의 추정값(estimate)을 사용할 수 있고 기울기와 절편의 표준오차(standard error)를 추정할 수 있다. 식(2.9) 와 (2.10) 을 활용하면,

$$\text{Estimated Standard Error of } b_0 = 10.66162$$

$$\text{Estimated Standard Error of } b_1 = 0.0616$$

이 표준오차(standard error)를 어떻게 사용하는지는 다음 절에서 설명될 것이다.

자료 분석을 공부하는 학생들에게 자유도(degrees of freedom)라는 개념은 종종 이해하기 어렵고 심지어 공포의 대상이 되기도 한다. 잔차의 자유도(residual degree of freedom)가 $n-2$ 가 되고, 이것은 2개의 모수(parameters)를 추정하기 위해 2 자유도(degrees of freedom)가 이미 사용되었기 때문이라는 사실을 설명했다. 비슷한 방식으로 s^2 이 잔차의 표본분산(sample variance of the residuals)임을 설명하였다. 잔차의 자유도는 항상 표본의 크기에서 잔차의 제약 개수(numbers of restriction)를 뺀 값으로 보여진다. 이 제약(restriction)은 모수 추정(parameter estimation)의 필요성으로 귀납된다. y_1, y_2, \dots, y_n 으로 표시되는 확률변수(random variables)를 가지는 $N(0, \sigma)$ 로부터 독립적이며 같은 분포를 따르도록(i.i.d) 추출(sampling)하는 경우, σ^2 를 추정하기 위한 자유도가 $n-1$ 이 된다는 것은 기초적인 통계를 배운 학생이라면 알 수 있을 것이다. 여기서 $\sum_{i=1}^n (y_i - \bar{y}) = 0$ 이라는 제약(restriction)에 한 개의 자유도가 쓰여지기

때문이며, 이 경우에 잔차는 $y_i - \bar{y}$ 이다. 단순선형회귀(simple linear regression)의 경우 잔차

$y_i - \hat{y}_i$ 상에 2개의 제약(restriction)이 있다. 이 제약들은 다음과 같으며,

$$\sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)x_i = 0$$

이 것은 식(2.4)와 (2.5)를 사용하여 쉽게 증명할 수 있다. 식(2.4)와 (2.5)는 각각 β_0 와 β_1 을 추정하는 과정에서 유도되었기 때문에 이 제약(restriction)이 모수 추정(parameter estimation)의

필요성과 쉽게 연관된다.

2.4. 최대우도법(Maximum Likelihood Estimation)

2.3절에서는 최소제곱법(method of least squares)에 대하여 설명하였고, 편향(bias)과 분산(variance)의 특성(property)에 대하여 알아보았다. 최소제곱법으로 구한 추정값의 편향(bias)과 분산(variance)의 특성을 알아보기 위해서 ε_i 가 무상관(uncorrelated)이며 평균이 0이고, 분산이 σ^2 이라는 가정이 필요하였다. 그러나 ε_i 의 분포의 형태(distribution type)에 대한 가정은 없었다. 이 장의 많은 부분에서는 정규성 가정(normal theory assumption)을 따르고 있고, 이러한 가정은 가설검정(test of hypothesis)이나 신뢰구간추정(confidence interval estimation)을 위한 추론(inference)에 사용된다. 만일 확률변수(random variable)(여기서는 ε_i)의 분포에 대해서 가정할 수 있다면, 좋은 점근적 특성(outstanding asymptotic property)을 가지면서, 직관적으로 와닿는 추정(estimation) 형태인 최대우도법(method of maximum likelihood)을 사용할 수 있다는 것을 기억하자. Roussas (1973)를 참조하자.

회귀계수의 추정(Estimation of Regression Coefficients)

ε_i 이 정규분포 $N(0, \sigma^2)$ 를 따른다고 가정하자. 결합밀도함수(joint density function) 혹은 우도함수(likelihood function)는 다음과 같이 주어진다.

$$L = \prod_{i=1}^n f(\varepsilon_i)$$

여기에서 $f(\varepsilon_i)$ 는 표본(sample)에서 i 번째 교란(disturbance) 또는 확률오차(random error)인 ε_i 의 밀도함수(density function)이다. 따라서 우도(likelihood)를 아래와 같이 표현할 수 있다..

$$\begin{aligned} L &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{\left(\frac{-1}{2\sigma^2}\right) \sum_{i=1}^n \varepsilon_i^2} \\ &= \frac{1}{(2\pi)^{n/2} \sigma^n} e^{\left(\frac{-1}{2\sigma^2}\right) \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2} \end{aligned}$$

우도함수(L)를 최대화하는 추정량(estimator) $\hat{\beta}_0, \hat{\beta}_1$ 을 구하는 방법은 우도함수의 지수부분인

$$\left[\frac{1}{\sigma^2} \right] \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

을 최소화하는 β_0, β_1 을 구하는 방법과 동일하다. 이 과정은 잔차제곱합(residual sum of squares)을 최소화하는 최소제곱법의 형태이다. 결과적으로 정규분포의 가정하에 회귀계수(regression coefficient) β_0, β_1 의 최대우도추정량(the maximum likelihood estimators)은 주어진 식(2.7)과 (2.8)의 최소제곱추정량(least squares estimators)과 동일하다. 반드시 명심해야 할 것은 만일 정규분포를 가정할 수 없다면 최소제곱추정량(least squares estimators)과 최대우도추정량(the maximum likelihood estimators)은 동일하지 않다는 점이다. 앞에서도 언급한 바와 같이 다중회귀(multiple regression)를 다루는 3장에서는 정규분포의 가정이 적절한 상황에서 최소제곱추정량(least squares estimators)의 특성에 대해 정식으로 다룰 것이며 정규분포를 가정할 수 없는 상황과 비교하여 볼 것이다.

오차분산의 추정(Estimation of Error Variance)

2.3 절에서 회귀계수(regression coefficient)의 최소제곱추정량(least square estimator)이 유도되었고 σ^2 의 중요한 추정량(estimator)인 s^2 에 대해 다루었다. σ^2 의 불편추정량(unbiased estimator)은 자료 분석에서 가장 흔히 사용되는 것이다. 그러나, 정규분포(normal theory case)하에서 이것이 곧 최대우도추정량(the maximum likelihood estimators)인가? β_0 와 β_1 을 b_0 와 b_1 으로 대체한 우도함수(likelihood function)는 다음과 같다.

$$L = \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} e^{\left(\frac{-1}{2\sigma^2}\right) \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

로그우도함수(log likelihood)를 σ^2 에 대하여 미분하면,

$$\frac{\partial \log L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

위의 도함수(derivative)를 0으로 놓으면 최대우도추정량(the maximum likelihood estimators) $\hat{\sigma}^2$ 을 얻을 수 있다. 따라서,

$$\hat{\sigma}^2 = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}$$

이 된다. 결과적으로 σ^2 의 최대우도추정량(the maximum likelihood estimators)은 편향된 추정량(biased estimator)이 되고 다음과 같이 쓸 수 있다.

$$\boxed{\hat{\sigma}^2 = s^2 \left(\frac{n-2}{n} \right)}$$

2.5. 변이의 분할(Partitioning Total Variability)

어떤 회귀문제(regression problem)에서든 반응(response)의 변동(variation)은 관측된다. 그래서, 적합모형(fitted model)은 이러한 반응의 변동(variation)을 설명(explanation)하는 것으로 볼 수도 있다. 물론 적합값 \hat{y}_i 과 y_i 는 가깝게 붙어있는 것이 중요하다. 만약 이것이 충족된다면, \bar{y} 주위의 \hat{y}_i 의 변동(variation)은 \bar{y} 주위의 y_i 의 변동(variation)과 매우 유사할 것이다. 다음의 사실을 주의해서 살펴보자(중심화모형(centered model)을 사용한다).

$$\begin{aligned}\bar{y} &= \sum \frac{\hat{y}_i}{n} \\ &= \frac{1}{n} \sum_{i=1}^n [b_0^* + b_1(x_i - \bar{x})] \\ &= \frac{1}{n} \sum_{i=1}^n [\bar{y} + b_1(x_i - \bar{x})] \\ &= \bar{y}\end{aligned}$$

결과적으로 변동(variation)의 원인(source)을 총제곱합(total sum of squares) 즉,

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

과 회귀제곱합(regression sum of squares), 즉,

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

으로 보는 것은 자연스러울 것이다. 전자는 관찰된 반응내의 변동(variation in observed response)을 측정한 것이고 후자는 회귀에 의해 “설명된” 변동(variation explained by the regression)을 표현한 것이다. 통계적 추론(statistical inference)의 많은 방법들이 총제곱합(total sum of squares)과 회귀제곱합(regression sum of squares)의 관계에 상당히 의존하고 있으며, 그 관계는 다음과 같다.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.11)$$

$$SS_{\text{Total}} = SS_{\text{Reg}} + SS_{\text{Res}}$$

여기서 SS_{Res} 는 잔차제곱합(residual sum of squares)이다. 식(2.11)은 다음과 같이 나타낼 수 있다.

$$(\text{Total variation in response}) = (\text{Variability explained by method}) + (\text{Variability unexplained})$$

절편을 포함하는 모형(모형(2.2))에서는 $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$ 이다. 결과적으로 변이(variability)는 회귀제곱합(regression sum of squares)과 잔차제곱합(residual sum of squares)이라는 중요한 두 가지 요소로 구성되며, 이들은 각각 회귀선에 의한 변동(variation due to the regression line)과 회귀선 주위의 변동(variation around the regression line)을 나타낸다. 따라서, 회귀제곱합(regression sum of squares)과 잔차제곱합(residual sum of squares)이라 부르는 개념을 명확하고 직관적으로 이해해야 한다. 분석에 있어서 바람직한 것은 되도록이면 SS_{Res} 보다 큰 SS_{Reg} 을 얻어내는 것이다. SS_{Reg} 는 x 변화에 의해 발생하는 y 변동을 말하는 것이다. 그러나, SS_{Res} 는 ε_i 에 의한 단순한 우연변동(chance variation)으로 볼 수도 있다. 식(2.11)의 중요한 항등식(identity)은 아래의 식으로부터 유도되었다.

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

이제 양쪽을 제곱해서 합을 하면 아래의 식을 구할 수 있다.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

이때 오른쪽 식의 마지막 항은 0이 된다. 이는 다음에서 쉽게 알 수 있다.

$$\begin{aligned} \sum_{i=1}^n (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) &= \sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) \\ &= \sum_{i=1}^n [b_0 + b_1(x_i - \bar{x})](y_i - \hat{y}_i) \\ &= b_1 \sum_{i=1}^n x_i (y_i - \hat{y}_i) \\ &= b_1 \sum_{i=1}^n x_i [(y_i - \bar{y}) - b_1(x_i - \bar{x})] \\ &= b_1 [S_{xy} - b_1 S_{xx}] \\ &= 0 \end{aligned}$$

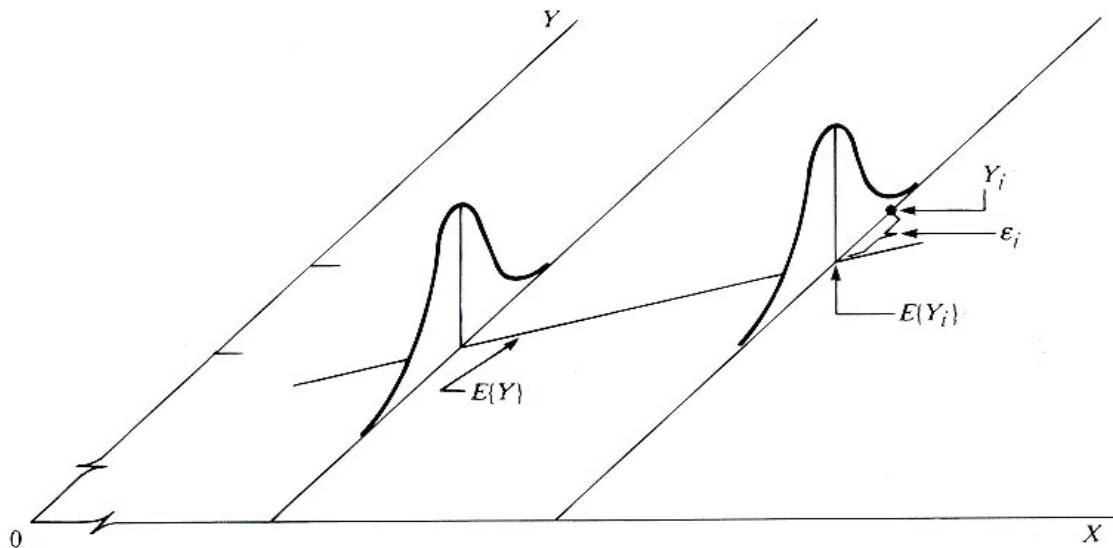
이제 회귀의 기울기(slope of the regression)에 대한 추론(inference)의 기초를 다지기 위해 정규분포 가정하에서 제곱합(sum of squares)의 분포특성(distributional property)에 대해 알아보자.

제곱합의 분포 특성(Distributional Properties of Sums of Squares)

최소제곱법(least squares procedure)의 장점 중의 하나는 SS_{Reg} 와 SS_{Res} 의 특성을 쉽게 공부하고 이해할 수 있는 기준(criteria)을 제공한다는 것이다. 이를 총변이(total variability)의 2가지 구성요소의 분포특성(distributional properties)을 이해하는 것은 가설검정을 위해 꼭 필요하다. 다시 한번 ε_i 가 정규 분포한다고 가정해보자(2.2 절에서 논의한 것에 덧붙여서). 간단하게 말하면 ε_i 는 독립적인(independent) $N(0, \sigma^2)$ 이다.

이제 독자들은 식(2.2)의 모형에서 정규분포와 그 외 다른 가정이 어떤 역할을 하는지 쉽게 이해할 수 있을 것이다. Fig 2.4 의 모형에서 x 에 대한 y 의 평균을 연결한 선과 그 선 주위의 산포(scatter)는 $E(y|x)$ 주위의 가우스 분포(Gaussian distribution)에 의해 생성된다. 각 분포의 분산이 동일한 것에 주목하자. 이는 등분산(homogeneous variance)을 가정하고 있다는 의미이다. 그림상의 점들은 자료 포인트(data point)를 의미한다. 평균들을 연결한 선이 “진정한” 회귀선(true regression line)이 될 것이다.

Figure 2.4 Simple linear regression model



회귀모형(regression model)으로 설명되는 변동이 실제하는(real) 것인지 혹은 변동(variation)이 단순한 우연변동(chance variation)인지를 판별하는 데 있어서 식(2.11)의 총제곱합(SS_{Total})의 분할(partitioning)이 유용하다. 여기서 ‘실재하는(real)’이라 함은 x 와 y

사이에 선형적인 관계가 실제로 존재한다는 것을 의미한다. 식(2.2)를 시작하면서 우리는 회귀선의 기울기 β_i 이 0이 아니면 y 는 x 에 대해 회귀(regression)한다고 정의하였다. Fig 2.5에서는 2개의 가설적 자료 세트(hypothetical data sets)를 보여주는데, 하나는 x 에 대한 y 의 회귀가 통계학적으로 유의해 보이고, 다른 하나는 변동이 우연변동(chance variation)이며 진정한 회귀(true regression)는 수평선인 것으로 보인다. Fig 2.5(a)에서 총제곱합(SS_{Total})의 상당 부분을 회귀제곱합(SS_{Reg})이 설명할 수 있음을 알 수 있다. 그러므로 적합선(fitted line) 주위에 약간의 임의변동(random variation)이 보이긴 하지만 회귀선이 양의 기울기(positive slope)를 보이고 x 가 증가하면 $E(y|x)$ 가 증가한다고 결론지을 수 있다. 반면에 그림 2.5(b)에서는 적합 회귀선(fitted regression line)이 수평이다. 그러므로 회귀제곱합(SS_{Reg})은 0이고 y 에 대한 전체 변동(entire variation)은 적합선 $\hat{y} = \bar{y}$ 주위의 오차변동(error variation)이라고 명확하게 말할 수 있다.

Figure 2.5

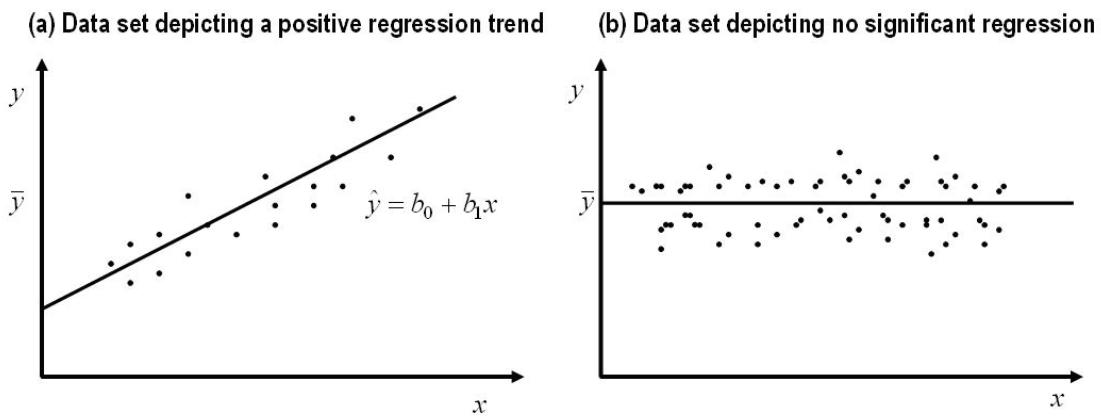


Fig 2.5에서 (a)에는 의미 있는 회귀가 있지만, (b)에서는 의미 있는 회귀가 없다는 것이 명확하다. 그러나, 많은 통계학적 가설 검정에서 회귀가 유의한지를 결정하는 것이 필요한데 이들 제곱합(sum of squares)의 분포 특성(distributional properties)이 도움이 되는 것은 분명하다.

자유도의 분할(Partitioning of Degrees of Freedom)

정규분포의 가정하에 잔차제곱합(SS_{Res})은 $\sigma^2 \chi_{n-2}^2$ 의 분포를 가진다. 이론적 정의를 위해 독자들은 Graybill(1976) 과 Searle(1971)를 참고하자. 회귀관계가 성립하지 않을 때 즉, $\beta_i = 0$ 이면, SS_{Reg}/σ^2 는 χ_1^2 이 된다. 결과적으로 잔차자유도(residual degrees of freedom)와 회귀자유도(regression degrees of freedom)는 같은 뜻이 된다. 이제 우리는 제곱합(sums of

squares)을 분할하였듯이 자유도(degrees of freedom)를 분할할 수 있다.

$$SS_{\text{Total}} = SS_{\text{Reg}} + SS_{\text{Res}} \quad (\text{sum of squares})$$

$$n - 1 = 1 + (n - 2) \quad (\text{degrees of freedom})$$

만약 독자가 중심화 형태(centered form)의 적합모형(fitted model), $\hat{y} = \bar{y} + b_1(x - \bar{x})$ 을 다루고 있다면, 자유도 분할의 개념은 매우 직관적이며 명확할 것이다. 만약 자료 포인트(data point)의 개수 n 을 이용 가능한 자유도의 총수(total number of available degrees of freedom)라 한다면 분할(partitioning)은 다음과 같이 된다.

Regression df	1	(Degree of freedom used for estimation of slope)
Residual df	$n - 2$	(Number of data points, with two df eliminated for estimation of two parameters in the model)
Total df	$n - 1$	(Number of data points, with one eliminated for estimation of mean \bar{y})

총자유도(total df)로 $n-1$ 을 사용하는 것은 총제곱합(total sum of squares)이 평균 \bar{y} 주위의 변동(variation)을 반영하고 있기 때문이다.

자유도의 개념을 포함하는 모든 자료분석도구(data analysis tool)에서는 자유도의 근원(source)은 반드시 χ^2 분포를 따르는 것으로 간주한다. 결과적으로 분할(partitioning)의 목적은 다음의 가설을 검정하기 위한 것이라고 생각할 수 있다.

$$H_0: \beta_I = 0$$

$$H_1: \beta_I \neq 0$$

이러한 맥락에서 자유도 분할(partitioning)의 이론적 근거를 다음과 같이 추가할 수 있다.

SS	df	Distribution
SS_{Reg}	1	$\sigma^2 \chi^2_1$ (under H_0)
SS_{Res}	$n - 2$	$\sigma^2 \chi^2_{n-2}$
SS_{Total}	$n - 1$	$\sigma^2 \chi^2_{n-1}$ (under H_0)

추가로 회귀제곱합(SS_{Reg})과 잔차제곱합(SS_{Res})은 서로 독립(independent)이다. 이론적 정의(theoretical justification)를 위해 다시 한번 독자들은 Graybill(1976)과 Searle(1971)를 참고하자. 이 결과들은 기울기에 관한 가설검정을 쉽게 수행할 수 있게 하고 그 쓰임을 잘 이해할 수 있도록 해준다. 아마도 더 중요한 것은 이 이론이 3장에서 논의될 다중회귀모형(multiple regression model)의 광범위한 추론(inferences)의 근간(foundation)이 된다는 것이다.

2.6. 절편과 기울기의 가설검정(Test of Hypotheses on Slope and Intercept)

이 시점에서 독자들은 각자의 영역에서 적합회귀선(fitted regression line)으로부터 어떤 종류의 정보를 추출할지 심각하게 생각해보아야 할 것이다. 그 가능성은 다음과 같이 요약될 수 있을 것이다.

1. 회귀변수(regressor variable)인 용량(dose) x 가 정말 반응(response) y 에 영향을 미치는가?
2. 자료(data)가 모형에 충분히 적합(adequate fit) 되는가?
3. 모형이 반응(response)을 만족스럽게 예측하는가? (내삽(interpolation) 혹은 외삽(extrapolation)을 통해)

첫번째의 경우 기울기 β_1 에 대한 가설 검정을 통해 해답을 얻을 수 있을 것이다. 전술한 바와 같이 가설은 종종 다음과 같이 주어진다.

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned} \tag{2.12}$$

만일 귀무가설(H_0)이 참이라면 다음과 같은 사실을 함축하고 있는데, 그것은 모형이 $E(y)=\beta_0$ 로 축소(reduce)되고 회귀변수(regressor variable)는 반응(response)에 영향을 미치지 않는다는 것이다.(적어도 식(2.1)에 포함된 것처럼 선형관계(linear relationship)를 보여주지는 못함). 가설이 기각되고 H_1 이 채택되면 x 가 선형양식(linear fashion)으로 반응(response)에 의미 있는 영향을 미친다고 결론지을 수 있다.

(2.12)에서 H_0 가설을 기각하는데 있어서 위험은 상존한다. 이런 이유로 검정 결과가 중요성을 가지지 못하는 경우가 종종 있다. x 가 의미있게 반응(response)에 영향을 미친다고 해도 회귀분석(regression analysis)에서 모형이 문제의 해결에 도움이 되지 못하는 경우도 있다. H_0 가설의 기각(rejection)은 단지 어떤 경향(trend)이 있다는 것을 감지한 것에 불과하다. 미리 예상되는(preconceived) 표준(standard)에 대한 회귀선의 적합의 질(quality of fit of the regression line)에 관한 정보는 제공하지 못한다. 또한 더 중요한 것은 그 모형이 얼마나 예측을 잘 할 수 있는가에 대해서는 아무 것도 알려주지 않는다는 것이다.

이미 예상하였듯이 (2.12) 에서 H_0 가설의 검정(test)은 회귀제곱합(SS_{Reg})과 잔차제곱합(SS_{Res})의 분포특성(distributional properties)을 이용한 것이다. 결론적으로 말하면 그것은 상대적인 크기(relative magnitude)에 달려있다. 이 검정(test)의 자세한 부분은 앞으로 다루어 질 것이다.

분산분석(Analysis of Variance)

2.5절에서 언급한 정규분포의 특성은 식(2.12)의 가설을 검정하기 쉽게 해준다. Table 2.2에서 분산분석(analysis of variance)의 요약된 계산을 통해 단순 F 검정(simple F-test)을 이용할 수 있다.

$$\frac{SS_{\text{Reg}}/1}{SS_{\text{Res}}/(n-2)} = \frac{MS_{\text{Reg}}}{s^2}$$

H_0 하에 상기 통계량(statistic)은 $\frac{\chi^2/1}{\chi^2/(n-2)}$ 이고, $\frac{MS_{\text{Reg}}}{s^2}$ 가 $F_{1,n-2}$ 를 따르므로

(2.12)에서의 가설(hypothesis)에 대한 검정 통계량(test statistics)이 될 수 있다.

Table 2.2는 단순선형회귀(simple linear regression)에서 분산(variance)을 분석하는 표준적인 방법이다. 물론 저변에 깔려있는 정보는 F 통계량(F-statistic)에서 나왔다. F 값(F-value)이 평균제곱(mean squares)의 비(ratio)라는 사실을 보면 어떤 원리를 알 수 있다(s^2 이 평균오차제곱(error mean square)을 의미함을 기억할 것). 여기에서 F 통계량(F-statistic)은 ‘모형으로 설명되는 분산(variance explained by the model)’을 ‘모형오차(model error) 혹은 실험오차(experimental error)에 기인한 분산(variance)’으로 나누어서 얻어지는 비(ratio)로 볼 수 있다. 결과적으로 $F = \frac{MS_{\text{Reg}}}{s^2}$ 의 값이 크면 대립가설 H_1 을 뒷받침하는 것이다. 예상되는 두

평균제곱(mean squares)의 값을 관찰함으로써 달리 볼 수도 있다. 우리는 $E(s^2) = \sigma^2$ 임을 알고 있고 이는(예제 2.5를 참조) 다음과 같은 결과로 이어진다.

$$E(MS_{\text{Reg}}) = \sigma^2 + \beta_1^2 S_{xx}$$

결과적으로 분산분석(analysis of variance)을 통해 유의한 기울기(significant slope)를 포착(detection)하는 것은 단순한 실험적 오차분산(experimental error variance) 이상의 통계적으로 유의한 $\beta_1^2 S_{xx}$ 값을 얻어 내는 것이다. 양의 성질을 가지는 $\beta_1^2 S_{xx}$ 는 한쪽 꼬리 F 검정(one-tailed F-test)에서 상부 꼬리(upper tail)를 사용하는 근거가 된다.

Table 2.2 분산분석표

요인(source)	제곱합(SS)	자유도(df)	평균제곱(MS)	F
회귀(Regression)	SS _{Reg}	1	SS _{Reg} /1	$F = \frac{MS_{Reg}}{s^2}$
오차(Residual)	SS _{Res}	$n - 2$	s^2	
전체(Total)	SS _{Total}	$n - 1$		

예제 2.2 부부간 키의 성향 분석

식(2.7)과 (2.8)을 이용하여 계산된 적합된 회귀식은 다음과 같고,

$$\hat{y} = 41.93015 + 0.69965x$$

평균제곱오차는 앞에서 살펴보았듯이 아래와 같이 구할 수 있다.

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = \frac{3303.3}{94} = 35.1$$

남편의 키와 아내의 키의 선형관계에 대한 가설검정, 즉

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

은 분산분석을 이용할 수 있다. Table 2.3는 분산분석 결과를 나타낸 것이다. 분산분석 결과 $F = MS_{Reg} / s^2 = 131.29$ 로 0.0001이하의 유의수준에서 유의하게 나타난다. 따라서 귀무가설을 기각할 수 있으므로(즉, 대립가설을 채택할 수 있으므로) 남편의 키와 아내의 키는 통계적으로 유의한 선형관계가 있다고 할 수 있다.

Table 2.3 분산분석표

요인(source)	제곱합(SS)	자유도(df)	평균제곱(MS)	F
회귀(Regression)	4613.7	1	4613.7	131.29
오차(Residual)	3303.3	94	35.1	
전체(Total)	7917	95		

다음은 위의 분석에 사용된 R-code이다

```
data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
ano<-anova(g)
```

t-검정과 신뢰구간 추정(The t-Tests and Confidence Interval Estimation)

t-검정을 사용하여 기울기와 절편에 대한 가설(hypothesis)을 각각 검정할 수 있다. 식(2.12)에 나와있는 양측 가설(two-sided hypothesis) 외에, 이 절에서 기술될 일반적인 공식(general formulation)을 이용하여 단측 가설(one-sided hypothesis)을 검정할 수 있다. 앞서 관찰하였듯이 가설 검정을 위해서 우리는 정규 이론 가정(normal theory assumption)을 해야 한다. 이 조건 하에서, 기울기는 y_i 의 선형 함수(linear function)이고 y_i 는 정규분포를 따르므로, 다음과 같이 기술할 수 있다.

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

분포들 간의 표준 관계(standard relationships)를 매우 직접적으로 응용하게 되면 다음과 같이 기술할 수 있다(Graybill (1976)을 보시오.).

$$\frac{(b_1 - \beta_1)}{s} \sqrt{S_{xx}} \sim t_{n-2}$$

여기서 t_{n-2} 는 자유도가 $n-2$ 인 스튜던트 t분포(student's t-distribution)이다.

따라서, $\beta_{1,0}$ 가 특정 상수(a specified constant)인 아래 가설을 검정하기를 원한다면,

$$H_0 : \beta_I = \beta_{I,0}$$

$$H_1 : \beta_I \neq \beta_{I,0}$$

다음의 통계량(statistic)을 이용하는 양쪽꼬리 t -검정(two-tailed t -test)을 해야 한다.

$$t = \frac{(b_1 - \beta_{I,0})}{S} \sqrt{S_{xx}} \quad (2.13)$$

$\beta_{I,0} = 0$ 인 특수한 경우에

$$t = \frac{b_1 \sqrt{S_{xx}}}{S} \quad (2.14)$$

이고, 따라서

$$t^2 = \frac{b_1^2 S_{xx}}{S^2} = \sum_{i=1}^n \frac{(\hat{y}_i - \bar{y})^2}{S^2} = \frac{SS_{\text{Reg}}}{S^2}$$

이것이 분산분석(analysis of variance) 접근법을 통하여 만들어진 F -통계량(F -statistic)이다.

만약 예를 들어 다음과 같은 단측 대립 가설(one-sided alternative hypothesis)에 관심이 있다면,

$$H_0 : \beta_I = \beta_{I,0}$$

$$H_1 : \beta_I > \beta_{I,0}$$

t_{n-2} 분포의 상부꼬리 임계역(upper tail critical region)을 사용하여 식(2.13)에 주어진 t -통계량(t -statistic)을 적용할 수 있다. 물론 가설이 다음과 같다면, 하부꼬리 임계역(lower tail critical region)을 사용한다.

$$H_0 : \beta_I = \beta_{I,0}$$

$$H_1 : \beta_I < \beta_{I,0}$$

절편을 검정하기 위해서, b_0 가 정규확률변수(normal random variables)의 선형 조합(linear combination)이라는 사실을 또 이용할 수 있으며, 따라서 다음을 관찰할 수 있게 된다.

$$b_0 \sim N\left[\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right]$$

그 결과로, 아래 가설을 검정하고자 한다면,

$$\begin{aligned} H_0 &: \beta_0 = \beta_{0,0} \\ H_1 &: \beta_0 \neq \beta_{0,0} \end{aligned}$$

이를 위해 적절한 검정 통계량(test statistic)은 아래와 같다.

$$t = \frac{b_0 - \beta_{0,0}}{s \sqrt{\left(\frac{1}{n}\right) + \left(\frac{\bar{x}^2}{S_{xx}}\right)}}$$

이것은 $H_0: \beta_0 = \beta_{0,0}$ 하에서 t_{n-2} 분포를 따른다. 기울기의 경우에서처럼, 만약 대립 가설(alternative hypothesis)이 단측(one-sided)이라면, 이에 해당하는 한쪽꼬리 t -검정(one-tailed t-test)이 적절하다.

이 시점에서 두 항목을 강조해야 할 필요가 있다. 기울기가 0이라는 가설을 검정하기 위해서는 식(2.14)에 있는 t -통계량(t-statistic)의 제곱이 분산분석법(analysis of variance)에서 생성되는 F -통계량(F-statistic)을 만들어낸다는 점을 주목하여야 한다. 이것은 양쪽꼬리 검정(two-tailed test)을 이용할 경우에는, 두 가지 방법이 같은 결론에 도달함을 의미한다. 또한, 식(2.14)의 t -통계량(t-statistic)이 다음의 형태를 지님을 주목하여야 한다.

$$t = \frac{\text{coefficient}}{\text{standard error of coefficient}}$$

여기에서 계수(coefficient)는 물론 b_1 이고, 표준오차(standard error)는 $\frac{s}{\sqrt{S_{xx}}}$ 이다. 후자는 물론,

b_1 의 표준편차(standard deviation)의 추정값(estimate)이다. 예제 2.3은 t -검정(t-test)의 용법의 실례를 보여준다.

예제 2.3 부부간 키의 성향 분석(계속)

Table 2.1의 자료를 다시 보자. 식(2.14)에 주어진 t -통계량(t-statistic)을 이용하여 가설검정을 해보자.

$$H_0 : \beta_I = 0$$

$$H_1 : \beta_I \neq 0$$

먼저, 회귀식의 기울기 추정값은 0.69965이고 이 추정값의 표준오차 $\frac{s}{\sqrt{S_{xx}}}$ 는 0.06106이다.

따라서, 자유도가 94인 t -통계량은 아래와 같이 주어진다.

$$t = \frac{0.69965}{0.06106} = 11.458$$

이것은 유의수준 0.0001미만에서 통계적으로 유의하다. 또한, 이 t -값은 같은 자료를 사용한 분산분석(analysis of variance)에서 얻을 수 있는 F 값($F=110.243$)의 제곱근과 동일하다. (2.12)에 제시된 가설을 검정하기 위한 t -검정은 이 절의 앞 절에서 약술된 분산분석 절차(procedure)와 동일하다.

자료 분석가가 기울기 β_I 이나 절편 β_0 에 대한 신뢰구간 추정값(confidence interval estimate)을 얻고자 하는 때가 종종 있다. 이를 위한 방법이 확립되어 있다. ε_i 에 대한 정규이론 가정(normal theory assumption) 하에서,

$$\frac{(b_I - \beta_I)\sqrt{S_{xx}}}{s} \sim t_{n-2}$$

따라서, 우리는 이것을 다음의 부등식으로 쓸 수 있다.

$$\Pr \left\{ -t_{\alpha/2, n-2} < \frac{(b_I - \beta_I)\sqrt{S_{xx}}}{s} < t_{\alpha/2, n-2} \right\} = 1 - \alpha$$

여기에서 $t_{\alpha/2, n-2}$ 는 $n-2$ 의 자유도를 가지는 t -분포의 상위(upper) $\alpha/2$ 퍼센트 지점(percent point)에 해당한다. 이제, $1-\alpha$ 의 확률로 β_I 이 아래에 기술된 구간 내에 있다고 얘기할 수 있다.

$b_I \pm t_{\alpha/2, n-2} \sqrt{\frac{s^2}{S_{xx}}}$

비슷한 방식으로, 아래의 통계량(statistic)의 표본 분포(sampling distribution)로부터

$$\frac{b_0 - \beta_0}{s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}}$$

β_0 에 대한 $100(1 - \alpha)\%$ 신뢰구간(confidence interval)이 다음과 같다고 기술할 수 있다.

$$b_0 \pm t_{\alpha/2, n-2} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

신뢰구간 추정시 흔히 경험하듯이 신뢰구간의 형태는 다음과 같다.

Point estimate \pm (Constant)(Standard error of point estimate)

예제 2.4 부부간 키의 성향 분석: 계수의 신뢰구간 추정값(Confidence Interval Estimates of Coefficients)

Table 2.1의 자료를 다시 이용해보면, 회귀선의 기울기 β_1 의 95% 신뢰구간 추정값(estimate)은 아래와 같이 얻을 수 있다.

$$0.69965 \pm (1.985523)0.06106 = (0.5784144, 0.820893)$$

마찬가지로 절편의 신뢰구간도 다음과 같이 구할 수 있다.

$$41.93015 \pm (1.985523)10.66162 = (20.7612518, 63.099055)$$

다음은 예제 2.3과 2.4에 사용된 R-code이다

```
#exmpla 2.1
data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
s_stat<-summary(g)
conf<-confint(g)
```

2.7. 원점을 통과하는 단순 회귀, 고정 절편(Simple Regression through the Origin, Fixed Intercept)

식(2.2)의 모형에서는 자료로부터 기울기(slope)와 절편(intercept)만 추정하면 된다. 하지만, 실제 현상에서는 절편이 특정한 값으로 고정된 경우가 많다. 사실, 많은 문제점들이 $\beta_0=0$ 라고 알려진 경우에 발생한다; 따라서 필요한 것은 원점을 통과하는 회귀선을 구하는 최소제곱 과정(least squares procedure)이다.

흥미롭게도, 많은 회귀분석 사용자들이, 그것이 모호한 값(dubious value)이라는 염려 때문에, 고정 절편 개념(fixed intercept notion)을 사용하기를 주저한다. 절편을 고정시킴으로써 우리는 최소제곱법(least squares)의 장점을 살리면서 기울기를 추정할 수 있게 된다. 모형에 도입된 추가 정보 즉, 알려진 β_0 값을 이용하여 최소제곱법(least squares)을 실행하기만 하면 된다.

자료 $(x_i, y_i), i = 1, 2, \dots, n$ 가 수집되어 있고, 일반적인 경우 즉 절편 β_0 를 알고 있는 경우를 다룬다고 가정해보자. 그리고 나면, 기울기의 추정량(estimator) b_1 은 다음을 만족해야 한다.

$$\frac{\partial}{\partial b_1} \left[\sum_{i=1}^n (y_i - \beta_0 - b_1 x_i)^2 \right] = 0 \quad (2.15)$$

이것은

$$\sum_{i=1}^n (y_i - \beta_0 - b_1 x_i) x_i = 0$$

가 되고, 따라서, 추정량(estimator) b_1 은

$$b_1 = \frac{\sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} \quad (2.16)$$

가 된다. 물론, 원점을 통과하는 회귀(regression through the origin)의 특수한 경우에 한하여 추정량(estimator) b_1 은 다음과 같이 변형된다.

$$b_1 = \frac{\sum y_i x_i}{\sum x_i^2} \quad (2.17)$$

원점 통과 회귀선의 기울기를 추론하는 방법은 2.2-2.6 절에서 주어진 내용을 엄밀하게 따른다. 예를 들어 무상관 오차(uncorrelated error)와 등분산(homogeneous variance)의 가정 하에서는

$$\text{var}(b_1) = \left(\frac{\sigma^2}{\sum_{i=1}^n x_i^2} \right)$$

이다. 그리고, β_1 에 대한 가설 검정을 위한 t -통계량(t -statistic)은 다음과 같다.

$$t = \frac{(b_1 - \beta_{1,0})}{s} \sqrt{\sum_{i=1}^n x_i^2} \quad (2.18)$$

여기에서,

$$s^2 = \sum_{i=1}^n \left[\frac{(y_i - \hat{y}_i)^2}{(n-1)} \right]$$

이고, 물론 i 번째 자료의 적합 반응(fitted response)을 나타내는 \hat{y}_i 는 $\hat{y}_i = b_1 x_i$ 로 얻을 수 있다. 만약 오차(errors)가 $N(0, \sigma^2)$ 을 따르고 독립적이면, (2.18)의 t -통계량(t -statistic)은 $H_0 : \beta_1 = \beta_{1,0}$ 하에서 t_{n-1} 을 따른다.

$\sum_{i=1}^n \left[\frac{(y_i - \hat{y}_i)^2}{\sigma^2} \right]$ 이 χ_{n-1}^2 변량(variate)이라는 사실로부터 잔차자유도(residual degrees of freedom)는 $n-1$ 임을 알 수 있다. 자유도가 n 이 아니고 $n-1$ 이라는 것은 직관적으로 납득이 되는데, 이는 추정해야 할 모수(parameter)가 단지 하나 즉, β_1 뿐이기 때문이다. 따라서 추정 시 단지 하나의 자유도만 사용된다. 만약 β_0 가 0이 아닌 특정한 값으로 알려져 있는 경우 기울기만 추정해야 된다면, 적합값(fitted value)은 $\hat{y}_i = \beta_0 + b_1 x_i$ 이며 (2.18)은 여전히 적용된다.

절편이 고정되어 있는 경우의 분산분석법(Analysis of Variance Approach for the Fixed Intercept Case)

절편이 고정되어 있는 경우에, 기울기를 검정할 목적으로 t -통계량(t -statistic)을 만들 때 앞서 기술된 방법을 사용하는 것은 매우 간단하다. 따라서, 가설 검정을 위해서 분산분석법(analysis of variance approach)은 흥미롭긴 하지만 필수적인 것은 아니다. 그러나, 절편이 고정되어 있는 경우에 있어서, β_0 를 추정해야 하는 일반적인 경우의 분산분석(analysis of variance)에 해당하는 것이 무엇인지 관찰하고 이해하는 것은 중요하다.

우선, (2.11)에서처럼 제곱의 합 $\sum_{i=1}^n (y_i - \bar{y})^2$ 을 분할하는 것은 더 이상 유효하지 않다는 점을 확실히 알아야 한다. 그러나, 이미 알고 있는 β_0 를 사용하여 다음의 항등식을 고려해보자.

$$\boxed{\sum_{i=1}^n (y_i - \beta_0)^2 = \sum_{i=1}^n (\hat{y}_i - \beta_0)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.19)$$

식(2.19)는 명백히 성립하는데 그 이유는 식(2.15)로부터 다음 식이 성립하기 때문이다.

$$\sum_{i=1}^n (\hat{y}_i - \beta_0)(y_i - \hat{y}_i) = b_1 \sum_{i=1}^n x_i (y_i - \hat{y}_i) = 0$$

여기에서, 우리는 \bar{y} 부근에서의 변동 보다는 β_0 부근에서의 변동(variation)을 분할(partition)하는 것을 다루고 있다. 다음의 구성 요소는

$$\sum_{i=1}^n (\hat{y}_i - \beta_0)^2 = b_1^2 \sum_{i=1}^n x_i^2$$

회귀제곱합(the regression sum of squares)에 해당하며, 정규이론 가정(normal theory assumption)과 함께, $\beta_1=0$ 라는 조건 하에서 $\sigma^2 \chi^2$ 을 따른다. 앞에서와 같은 방식으로, $\sum_{i=1}^n (\hat{y}_i - \beta_0)^2$ 과 SS_{Res} 는 독립적이고, $H_0 : \beta_1 = 0$ 하에서 아래와 같다.

$$\frac{b_1^2 \sum_{i=1}^n x_i^2}{S^2} \sim F_{1, n-1}$$

따라서, 재구성(reconstruction)에도 불구하고 변이 분할(partition of variability approach)은 여전히 타당하고, 반응(response)과 회귀변수(regressor variable) 간의 비선형 관계(non linear relationship)에 대한 가설을 검정하는데 있어서 상부 꼬리 F -검정(upper-tailed F -test)을 허용한다.

명백한 이유로, 분모(denominator) 자유도는 $n-2$ 에서 $n-1$ 로 바뀌고, 독자는 가설 값(hypothesized value)인 $\beta_{1,0}$ 가 0일 때, 위의 F -통계량(F-statistic)의 제곱근(square root)이 (2.18)의 t -통계량(t-statistic)임을 다시 관찰하게 될 것이다.

예제 2.5 열대어 자료

태평양 산호초에서 서식하는 열대어인 스쿠암피니스(squampinis)란 종의 암컷의 행태를 조사하기 위해 암컷과 수컷이 어울려 생활하는 10개의 활동군에서 3마리에서 9마리까지의 수컷을 제거하고 군에 남아 있는 암컷 중에서 성을 바꾼 수를 기록하였다. 이 경우 회귀분석을 통하여 수컷 한 마리당 몇 마리 꼴로 암컷이 성을 바꾸는지를 알아볼 수 있다. 수컷이 제거되지 않는 경우에는 성을 바꾸는 암컷이 없다고 가정한다면 원점을 지나는 회귀선이 타당할 것이다. 자료는 table 2.4에 나와있으며, fig 2.6은 최소제곱법을 이용하여 적합한 회귀선을 나타낸 것이다. 먼저 적합된 회귀모형은 다음과 같다.

$$\hat{y} = 1.01987x$$

즉, 식(2.17)에 있는 β_1 의 추정값 b_1 은 $b_1 = 1.01987$ 이다.

Table 2.5는 식(2.19)에서 $\beta_0=0$ 일 때, 분산분석 결과를 제시한 것이다. F -통계량은 추정된 회귀모형의 기울기가 0이 아님을(nonzero slope) 강하게 응호하는 정보를 제공한다.

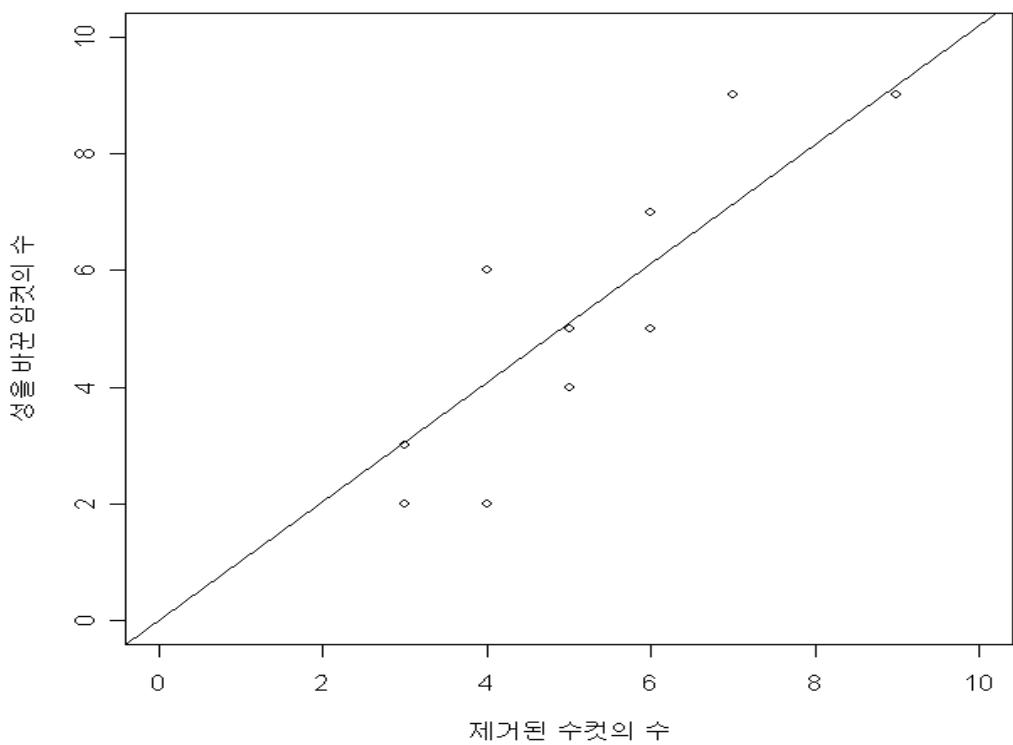
Table 2.4 열대어 자료

열대어 군	X (제거된 수컷의 수)	Y (성을 바꾼 암컷의 수)
1	3	2
2	3	3
3	4	2
4	4	6
5	5	4
6	5	5
7	6	5
8	6	7
9	7	9
10	9	9

Table 2.5 분산분석표

요인	제곱합	자유도	평균제곱	F
회귀	314.119	1	314.119	178.02
오차	15.881	9	1.765	
전체	330	10		

그림 2.6 데이터와 적합된 회귀선



다음은 예제를 분석하기 위해 사용된 R-code이다

```

data<-read.table("d:/data/ex2_5.R",header=TRUE)
g1<-lm(Y~X-1,data)
plot(Y~X,data,xlim=c(0,10),ylim=c(0,10),xlab="제거된 수컷의 수",ylab="성을 바꾼 암컷의 수",main="그림 2.6 데이터와 적합된 회귀선")
abline(g1,lty=1)
s_stat<-summary(g1)
ano<-anova(g1)
conf<-confint(g1)

```

2.8. 적합 모형의 질(Quality of Fitted Model)

독자는 2.6 절의 내용을 상기하여야 한다. 유의한 선형관계(significant linear relationship)가 존재하는지, 또는 특정 기울기(particular slope)가 합당한지 확인하기 위해서는 반드시 기울기에 대한 가설검정(hypothesis testing)이 필요하다. 또한, 적합 회귀모형(fitted regression)이 해결할 수 있는 다른 의문점들에 대해서도 언급하였다. 사실, 많은 회귀 문제(regression problem)에서 가설검정(hypothesis testing)은 분석자의 문제를 해결할 수 있는 추론(inference)의 형태는 아니다. 대개 성공적인 분석을 위해서는 정량적인 기준(quantitative criteria)을 적절히 선택하고 이를 통해 적합 모형(fitted model)의 질을 결정해야 한다. 여기에서 두 가지의 의문이 생긴다.

1. 자료(data)가 모형(model)에 충분히 적합(fit)한가?
2. 모형(model)이 반응(response)을 충분히 잘 예측(predict)하는가?

이 절에서 다루는 문제들은 다중회귀분석(multiple regression case)에까지 확대되겠지만 여기에서는 단순선형회귀분석(simple linear regression)의 경우에 대해서만 구체적으로 다루도록 하겠다. 그리고, 적합(fit)과 예측(prediction)을 구별(distinction)하도록 독자들을 계속 상기시킬 것이다.

결정계수(Coefficient of Determination, R^2)

결정계수(coefficient of determination)는, R^2 으로 표시하여 많이 사용되며, 간혹 회귀선의 적합(fit of the regression line)을 측정(measure)하는 것으로 오인되기도 한다. 만약 우리가 식(2.2)의 모형을 고려해 보면, 정의는 단순히 다음과 같다.

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.20)$$

(2.11)에 있는 기본적인 분할항등식(partition identity)으로부터, R^2 은 다음과 같이 바꿔 쓸 수 있다.

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{\text{Total}}}$$

모형적합자(model fitter)는 R^2 의 값을 상당히 자주 인용하는데, 아마도 이것은 해석(interpretation)의 용이함 때문일 것이다. (2.20)로부터 R^2 은 반응 자료에서 모형에 의해 설명되는 변동의 비율(the proportion of variation in the response data that is explained by the model)을 나타낸다는 것을 쉽게 알 수 있다. 명확히 $0 \leq R^2 \leq 1$ 이고, 자료에 대한 모형의 적합이 완전할 때, 즉, 모든 잔차(residual)가 0 일 때 R^2 은 1(upper bound) 이 된다. R^2 의 만족할 만한 수치(acceptable value)는 얼마인가? 이것은 대답하기 힘든 문제이고, 실제로, R^2 의 만족할 만한 수치(acceptable value)는 자료가 얻어진 과학 분야(scientific field)에 따라 다르다. 고도로 정밀한 장비(equipment)의 선형 보정(linear calibration)을 담당하는 화학자는 매우 높은 R^2 수치(아마도 0.99 이상)를 얻기를 강력히 원할 것이고, 반면에 인간 행동을 반영하는 자료를 다루는 행동과학자(behavioral scientist)는 R^2 의 수치가 0.70 정도로 높게 관찰되는 것을 다행으로 여길 것이다. 숙련된 모형적합자(experienced model fitter)라면 주어진 상황에서 R^2 의 값이 충분히 크다고 할 수 있을 때가 언제인지를 느낌으로 알 수 있을 것이다. 분명히, 어떤 과학 현상들은 다른 것 보다 훨씬 더 정확한 모형화(modeling)를 요한다.

비록 결정계수(coefficient of determination)가 해석하기 쉽고 통계학에서의 숙련도에 상관없이 대부분의 실험자들이 이해하기 쉽지만, 주의깊게 사용하지 않는다면 함정(pitfall)에 빠질 수 있다. 예를 들면, 이것은 후보모형(candidate model)들을 비교함에 있어서 위험한 기준(criterion)이 될 수 있다. 왜냐하면 모든 추가 모형항(additional model term)(예, (2.2)에 추가된 x_i 의 2차항)은 SS_{Res} 를 감소시킬 것이고(적어도 증가시키지는 않음) 이로인해 R^2 을 증가(적어도 감소시키지는 않음)시키기 때문이다(예제 2.6 참고). 이것은 다소 무분별한 과대적합(overfitting) (즉, 매우 많은 모형항(model term)의 포함)으로 인위적으로 R^2 가 높게 만들어 질 수 있음을 내포한다. R^2 이 증가했다고 해서 추가 모형항(additional model term)이 반드시 필요한 것은 아니다. 실제로, 예측(prediction)이 모형의 목적이라면, 더 복잡한 구조의 높은 R^2 을 갖는 모형이 단순한 구조의 모형보다도 우수한 예측식(superior prediction equation)을 반드시 나타내지는 않는다는 것을 이후에 알게 될 것이다. 따라서, 단순히 R^2 만을 이용한 모형선택과정(model selection process)은 바람직하지 않다. 과대적합(overfitting)의 개념과 예측능력(prediction capability)을 평가함에 있어서 그것의 역할에 대해서는 3장과 4장에서 상세하게 다룰 것이다.

회귀의 기울기(slope of regression)가 크거나 또는 회귀변수 자료(regressor data) x_1, x_2, \dots, x_n 의 흩어짐(spread)의 정도가 커져 결정계수(coefficient of determination)가 인위적으로 높게 나타날 수도 있다. 이것은 결정계수의 분자(numerator)와 분모(denominator)의 구조를 각각 살펴봄으로써 쉽게 알 수 있다. 우리는 이미 다음을 알고 있고,

$$E(SS_{\text{Reg}}) = \sigma^2 + \beta_1^2 S_{xx}$$

그리고 이는

$$E(SS_{\text{Res}}) = \sigma^2(n - 2)$$

결과적으로

$$\begin{aligned} E(SS_{\text{Total}}) &= E(SS_{\text{Reg}} + SS_{\text{Res}}) \\ &= \sigma^2(n - 1) + \beta_1^2 S_{xx} \end{aligned}$$

현재 우리는 $E(R^2)$ 을 단순히 기대값의 비(ratio of expected value)로 확인할 수는 없다. 그러나, 모수적 R^2 형태(parmetric R^2 -type quantity)은 다음과 같이 정의할 수 있다.

$$\mathfrak{R}^2 = \frac{\frac{1}{n} + \frac{\beta_1^2 S_{xx}}{n\sigma^2}}{\frac{n-1}{n} + \frac{\beta_1^2 S_{xx}}{n\sigma^2}} = \frac{n^{-1} + \omega}{1 - n^{-1} + \omega}$$

$\omega = \frac{\beta_1^2 S_{xx}}{n\sigma^2}$ 은 신호대잡음비(signal-to-noise ratio)로 가시화 될 수 있다. ω 가 크다는 것은 추정된 기울기의 표준오차에 대한 기울기의 비(the ratio of slope to the standard error of the estimated slope)가 크다는 것을 의미한다. 이것은 명확히 바람직한 상황(situation)이며, 양질의 회귀(high quality regression)를 확실히 도출해낼 것이다. 중간 정도 또는 큰 n 에 대해서 \mathfrak{R}^2 은 다음과 같이 단순화될 수 있다.

$$\mathfrak{R}^2 \cong \frac{\omega}{\omega+1}$$

고정된 n (fixed n)에 대하여, 만약 측정된 회귀(measured regressor)의 변동(variation)이 매우 크다면, 즉 S_{xx} 가 클 때, ω 가 커지는 것은 명확하다. 큰(large) S_{xx} 가 종종 바람직하다고는 해도((2.9) 참조), 이는 인위적으로 높은 결정계수(artificially high coefficient of determination)를 초래할 수 있다. 다시 말하자면, 잔차(residual)와 s^2 의 수치가 동일하게 낙관적인(optimism) 결과를 반영하지 못함에도 불구하고 R^2 통계량(R^2 -statistic)이 커질 수 있다는 점에서 인위적인 결과라는 것이다. 결과적으로, 적합의 질(quality of fit)이 단순히 R^2 에만 영향을 받는 것이 아니라, 모형(model)과 자료범위(data range)의 일반적 특성(nature)에도 중요한 영향을 받는다는 것이다.

절편이 0인 경우의 R^2 (R^2 in the Zero Intercept Case)

절편이 0이거나 특정한 값으로 고정된 모형의 경우 R^2 의 계산은 상당한 딜레마(dilemma)에 직면한다. 식(2.11)을 통해 R^2 은 아래의 식과 같이 해석할 수 있다.

$$R^2 = \frac{\text{Variation explained by regression}}{\text{Variation observed}}$$

여기에서 **변동(variation)**은 반응변수(response variable)에 관한 것이다. 원점을 지나는 회귀모형의 경우 식(2.11)은 다음과 같이 대체가능하다.

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

이것은 식(2.19)의 특수한 경우이다. 그러므로 절편이 없는 모형(no intercept model)에 대한 R^2 유사형(R^2 -analog)은 다음과 같이 될 수 있다.

$$R_{(0)}^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = \frac{SS_{\text{Reg}}}{SS_{\text{Total}}}$$

사실상, 어떤 컴퓨터소프트웨어 패키지에서는 $R^2_{(0)}$ 를 위에서 보는 바와 같이 계산한다. 그러나, 절편이 없는 모형에서의 $R^2_{(0)}$ 에 대한 표현(expression)과 절편이 있는 모형(intercept model) (2.20)에서의 R^2 의 표현(expression) 사이에는 큰 차이가 존재한다. 절편이 0인 모형(zero intercept model)에서는, 문자와 분모에 의해 묘사되는 변동(variation)이 0 주위의 산포(dispersion around zero)로 표현된다. 절편이 있는 모형(intercept model)에서는, R^2 통계량(R^2 -statistic)은 실제 자료 y 의 산포(dispersion)에 대한 회귀추정량 \hat{y} 의 산포(dispersion)의 비율로, 이는 평균 \bar{y} 주위에 있는 산포(dispersion)이다. 따라서, 절편이 없는 모형에서 총제곱합(total sums of squares)과 회귀제곱합(regression sums of squares)을 재정의(redefinition)하는 과정에서 $R^2_{(0)}$ 에 대한 대안(alternative)을 고려할 필요가 있다. $R^2_{(0)}$ 의 주된 난점(major difficulty)은 절편이 있는 모형(intercept model)과의 성능비교(performance comparison)에 사용할 수 없다는 사실이다. 비록 적합(혹은 예측)의 질이 우수하지 않아도 식(2.20)에서는 R^2 보다 $R^2_{(0)}$ 이 더 큰 경향(tendency)으로 나타날 것이다. 이것은 **수정되지 않은 제곱의 합(uncorrected sums of squares)**을 사용한 결과이다. 그러므로, 잔차제곱합(residual sum of squares)의 관점에서 거의 동등한 성능(performance)의 경우에 대해, $R^2_{(0)}$ 는 아마도 R^2 보다 상당히 클 것이다. 즉, 절편이

없는 모형(zero intercept model)의 경우 $R^2_{(0)}$ 은 너무 높게 나타나 비교에 사용하는 경우 잘못된 결론을 이끌어 낼 수 있다는 것이다. 비록 R^2 가 확실히 이러한 목적에 이상적인 통계량(ideal statistic)은 아니지만, 경쟁모형(competing model) 사이의 합리적인 비교를 위해서 절편이 없는 모형(zero intercept model)에서의 R^2 를 계산하는 대체방법(alternate ways)이 있다. 한가지 가능한 선택이 다음과 같이 주어진다

$$R^2_{(0)*} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

여기에서 물론,

$$\sum_{i=1}^n (y_i - \bar{y}_i)^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i x_i\right)^2}{\sum_{i=1}^n x_i^2}$$

비록 $R^2_{(0)*}$ 가 $R^2_{(0)}$ 보다 비교(comparison)를 위한 보다 합리적인 근거(basis)를 제공하지만, 통계량 $R^2_{(0)*}$ 은 $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 가 상대적으로 큰 경우에 음(negative)이 될 수 있다. 절편이 0인 모형(zero intercept model)의 상세한 논의는 Casella (Casella 1983)와 Hahn (Hahn 1977)에 나타나있다. 절편이 있는 모형(an intercept model)과 절편이 없는 모형(zero intercept model) 사이의 비교는 예제 2.12를 참고하라.

변동계수(Coefficient of Variation)

변동계수(coefficient of variation, CV)는 적합의 질(quality of fit)을 나타내고, 회귀선(regression line) 주위의 잡음(noise)의 퍼짐(spread) 정도를 나타내는 합리적인 기준(criterion)이다. CV는 다음과 같이 정의 된다.

$$CV = (s / \bar{y}) \times 100 \quad (2.21)$$

변동계수(CV)는 오차표준편차(error standard deviation)의 잔차추정값(residual estimate)이고, 평균반응값(mean response value)의 백분율(percent)로 측정된다. 오차표준편차(error standard

deviation)의 추정값(estimate)인 s 는 척도에 좌우될 수 있어서(it is not scale free) 적합의 질을 측정하기에는 만족스럽지 못하므로, 이러한 문제점이 없는 변동계수(CV)를 사용하게 되었다. 예를 들면, 오차표준편차(error standard deviation)로 s 의 값이 백만명당 14부분(14 parts per million)일 때 경험이 많지 않은 분석가는 이것을 알아내지(indetify) 못할 것이다. 그러나, 만약 s 의 이 수치가 5%의 CV를 초래한다면, 분석가는 s 에 의해 측정되는 회귀선 주위의 자연적 산포(natural dispersion)가 평균반응(average response measurement)의 단지 5 %라는 것을 알수 있게 된다.

예제 2.6 부부간 키의 성향 분석: 회귀분석

Table 2.1의 부부간 키의 성향분석 데이터를 다시 고려해보자. Table 2.6은 이들 자료를 R을 통해 분석한 출력결과이다. 분산분석, 추정된 회귀계수(절편, 기울기)와 이들의 표준오차, 그리고 t -통계량과 p -value를 주의 깊게 살펴볼 필요가 있다.

Table 2.6 부부간 키의 성향분석 출력결과

Analysis of Variance Table

Response: W

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
H	1	4613.7	4613.7	131.29 < 2.2e-16 ***	
Residuals	94	3303.3	35.1		
<hr/>					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Call:

lm(formula = W ~ H, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-19.4685	-3.9208	0.8301	3.9538	11.1287

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.93015	10.66162	3.933	0.000161 ***
H	0.69965	0.06106	11.458	< 2e-16 ***
<hr/>				

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

Residual standard error: 5.928 on 94 degrees of freedom

Multiple R-Squared: 0.5828, Adjusted R-squared: 0.5783

F-statistic: 131.3 on 1 and 94 DF, p-value: < 2.2e-16

2.9. 평균반응과 예측구간에 대한 신뢰구간(Confidence Intervals on Mean Response and Prediction Intervals)

아마도 대부분, 과학자 또는 공학자(engineer)는 적합한 회귀(fitted regression)를 사용하여 반응값(response value)을 예측하거나 평균반응(mean response)을 추정할 수 있을지에 대해서 매우 관심이 있다. 예측(prediction)은 자료분석에서 필수불가결하게 중요하므로, 이 부분에 대해서 3장과 4장에 걸쳐서 살펴볼 것이다. 많은 회귀상황(regression situation)에서 예측의 질(quality of prediction) (또는 질의 부족(lack of quality))을 적합의 질(quality of fit)로 규명하기란 그리 쉽지 않다. 여기서 짚고 넘어가야 할 개념(concept)은 경험없는 분석가가 이해하는 것 보다 훨씬 더 복잡하다. 이 시점에서 우리는 단순선형회귀모형(simple linear regression model)의 회귀식(regression equation)을 사용하여 사용자가 예측의 질(quality of prediction)이나 평균 반응을 추정하는 질(quality of estimation of mean response)을 확인할 수 있는 기준을 단지 제시하고자 한다.

우선, 우리는 통계량(statistic)의 본질(nature)을 명확히 해야 한다

$$\hat{y}(x_0) = b_0 + b_1 x_0$$

이것을 추정반응(estimated response)으로 간주할 수도 있으며, $x = x_0$ 일 때 평균 y 의 추정량(estimator)이다. $E(y | x_0) = \beta_0 + \beta_1 x_0$ 를 고려할 때 이것이 매우 명확해진다. 그러므로 우리는 $\hat{y}(x_0)$ 의 표준오차(standard error)를 x_0 라는 조건에서의 평균반응 추정량의 표준오차(standard error of the estimator of mean response conditional on x_0)로 해석한다. 물론, 표준오차(standard error)의 개념은 정밀도(precision) 또는 변동(variation)의 이미지(image)를 떠올리게 한다. 이러한 경우, 동일한 x 수준(x-level)에서 y 에 대하여 새로운 관측(observation)을 하는 방식으로 반복 회귀(repeated regression)가 실시된다면, 이는 x_0 에서 \hat{y} 의 변동(variation)을 반영할 것이다. $\text{Var } \hat{y}(x_0)$ 에 초점을 맞추고 다음을 주목하라

$$\begin{aligned} \text{Var } \hat{y}(x_0) &= \text{Var}[b_0 + b_1 x_0] \\ &= \text{Var}[\bar{y} + b_1(x_0 - \bar{x})] \end{aligned}$$

\bar{y} 와 b_1 은 독립적이기 때문에 (2.6절 참조)

$$\text{Var } \hat{y}(x_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \quad (2.22)$$

σ^2 을 추정량(estimator) s^2 으로 치환하면, 우리는 예측의 표준오차(standard error of prediction)를 다음과 같이 정의할 수 있다

$$s_{\hat{y}(x_0)} = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (2.23)$$

실제로, σ 는 s 로 대체되므로 (2.23)은 예측의 추정표준오차(estimated standard error)로 정의된다. 이러한 결과에 따라, 정규오차(normal error)의 조건 하에서, $\hat{y}(x_0)$ 는 정규분포를 따르고, $E(y|x_0)$ 에 대한 $100(1-\alpha) \%$ 신뢰구간(confidence interval)은 다음과 같이 표현할 수 있다

$$\hat{y}(x_0) \pm t_{\alpha/2, n-2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (2.24)$$

사실, (2.24)의 식은 신뢰구간(confidence interval)을 나타내는 것이고, $x = x_0$ 에서 신규반응관측(new response observation)에 대한 예측구간(prediction interval)과 혼동해서는 안된다. 후자는 $x = x_0$ 에서 관측치 y_0 가 실제로 속할 것으로 기대되는 범위(bounds)를 반영한다. 이러한 범위(bounds)는 다음 예제 뒤에 전개되고 묘사될 것이다.

Table 2.7 부부간 키의 성향 분석 결과

No	y_i	\hat{y}_i	$y_i - \hat{y}_i$	$s_{\hat{y}_i}$	UCL	LCL
1	175	172.0657	2.934252	0.935123	170.209	173.9225
2	168	167.8678	0.132174	0.697297	166.4833	169.2523
3	154	153.8748	0.125249	1.06346	151.7632	155.9863
4	166	172.0657	-6.06575	0.935123	170.209	173.9225
5	162	155.9737	6.026288	0.91874	154.1495	157.7979
6	152	162.2706	-10.2706	0.621429	161.0367	163.5045
7	179	176.2637	2.736329	1.237393	173.8068	178.7205
8	163	160.8713	2.128712	0.6601	159.5606	162.1819
9	172	163.6699	8.330097	0.605346	162.468	164.8718
10	170	175.564	-5.56402	1.184505	173.2122	177.9159
11	170	169.2671	0.732867	0.765379	167.7475	170.7868

12	147	166.4685	-19.4685	0.645344	165.1872	167.7499
13	165	168.5675	-3.56748	0.729579	167.1189	170.0161
14	162	159.472	2.528019	0.717718	158.0469	160.897
15	154	155.2741	-1.27406	0.96553	153.357	157.1911
16	166	173.4651	-7.46506	1.031272	171.4174	175.5127
17	167	159.472	7.528019	0.717718	158.0469	160.897
18	174	169.9668	4.033213	0.804228	168.37	171.5636
19	173	173.4651	-0.46506	1.031272	171.4174	175.5127
20	164	158.0727	5.927326	0.790148	156.5038	159.6415
21	163	167.8678	-4.86783	0.697297	166.4833	169.2523
22	163	165.0692	-2.06921	0.61363	163.8508	166.2876
23	171	171.3661	-0.36609	0.889442	169.6001	173.1321
24	161	160.1716	0.828365	0.686802	158.808	161.5353
25	167	169.2671	-2.26713	0.765379	167.7475	170.7868
26	160	155.2741	4.725941	0.96553	153.357	157.1911
27	165	160.1716	4.828365	0.686802	158.808	161.5353
28	167	165.0692	1.930789	0.61363	163.8508	166.2876
29	175	167.8678	7.132174	0.697297	166.4833	169.2523
30	157	151.7758	5.22421	1.218574	149.3563	154.1953
31	172	160.8713	11.12871	0.6601	159.5606	162.1819
32	181	172.0657	8.934252	0.935123	170.209	173.9225
33	166	167.8678	-1.86783	0.697297	166.4833	169.2523
34	181	173.4651	7.534944	1.031272	171.4174	175.5127
35	148	148.9772	-0.97718	1.43572	146.1265	151.8278
36	169	167.1682	1.831828	0.669042	165.8398	168.4966
37	170	164.3696	5.630443	0.606436	163.1655	165.5736
38	157	157.373	-0.37302	0.830738	155.7236	159.0225
39	162	151.0761	10.92386	1.271939	148.5507	153.6016
40	174	171.3661	2.633905	0.889442	169.6001	173.1321
41	168	162.2706	5.729404	0.621429	161.0367	163.5045
42	162	158.0727	3.927326	0.790148	156.5038	159.6415
43	159	167.1682	-8.16817	0.669042	165.8398	168.4966
44	155	168.5675	-13.5675	0.729579	167.1189	170.0161
45	171	165.0692	5.930789	0.61363	163.8508	166.2876
46	159	160.8713	-1.87129	0.6601	159.5606	162.1819
47	164	157.373	6.62698	0.830738	155.7236	159.0225

48	175	169.9668	5.033213	0.804228	168.37	171.5636
49	156	155.2741	0.725941	0.96553	153.357	157.1911
50	180	176.2637	3.736329	1.237393	173.8068	178.7205
51	167	171.3661	-4.36609	0.889442	169.6001	173.1321
52	157	155.9737	1.026288	0.91874	154.1495	157.7979
53	167	171.3661	-4.36609	0.889442	169.6001	173.1321
54	157	160.8713	-3.87129	0.6601	159.5606	162.1819
55	168	165.0692	2.930789	0.61363	163.8508	166.2876
56	167	165.0692	1.930789	0.61363	163.8508	166.2876
57	145	153.8748	-8.87475	1.06346	151.7632	155.9863
58	156	158.7723	-2.77233	0.752328	157.2786	160.2661
59	153	151.7758	1.22421	1.218574	149.3563	154.1953
60	162	167.8678	-5.86783	0.697297	166.4833	169.2523
61	156	162.2706	-6.2706	0.621429	161.0367	163.5045
62	174	170.6664	3.333559	0.845705	168.9873	172.3456
63	160	171.3661	-11.3661	0.889442	169.6001	173.1321
64	152	157.373	-5.37302	0.830738	155.7236	159.0225
65	175	168.5675	6.43252	0.729579	167.1189	170.0161
66	169	160.8713	8.128712	0.6601	159.5606	162.1819
67	149	154.5744	-5.5744	1.013838	152.5614	156.5874
68	176	173.4651	2.534944	1.031272	171.4174	175.5127
69	165	168.5675	-3.56748	0.729579	167.1189	170.0161
70	143	151.0761	-8.07614	1.271939	148.5507	153.6016
71	158	154.5744	3.425595	1.013838	152.5614	156.5874
72	141	148.2775	-7.27752	1.491317	145.3165	151.2386
73	160	167.1682	-7.16817	0.669042	165.8398	168.4966
74	149	160.8713	-11.8713	0.6601	159.5606	162.1819
75	160	160.8713	-0.87129	0.6601	159.5606	162.1819
76	148	157.373	-9.37302	0.830738	155.7236	159.0225
77	154	157.373	-3.37302	0.830738	155.7236	159.0225
78	171	160.1716	10.82837	0.686802	158.808	161.5353
79	165	161.5709	3.429058	0.638142	160.3039	162.838
80	175	176.2637	-1.26367	1.237393	173.8068	178.7205
81	161	165.0692	-4.06921	0.61363	163.8508	166.2876
82	162	159.472	2.528019	0.717718	158.0469	160.897
83	162	160.1716	1.828365	0.686802	158.808	161.5353

84	176	170.6664	5.333559	0.845705	168.9873	172.3456
85	160	161.5709	-1.57094	0.638142	160.3039	162.838
86	158	154.5744	3.425595	1.013838	152.5614	156.5874
87	175	171.3661	3.633905	0.889442	169.6001	173.1321
88	174	170.6664	3.333559	0.845705	168.9873	172.3456
89	168	167.1682	0.831828	0.669042	165.8398	168.4966
90	177	170.6664	6.333559	0.845705	168.9873	172.3456
91	158	164.3696	-6.36956	0.606436	163.1655	165.5736
92	161	162.9702	-1.97025	0.610394	161.7583	164.1822
93	146	156.6734	-10.6734	0.873714	154.9386	158.4081
94	168	168.5675	-0.56748	0.729579	167.1189	170.0161
95	178	172.7654	5.234598	0.982477	170.8147	174.7161
96	170	168.5675	1.43252	0.729579	167.1189	170.0161

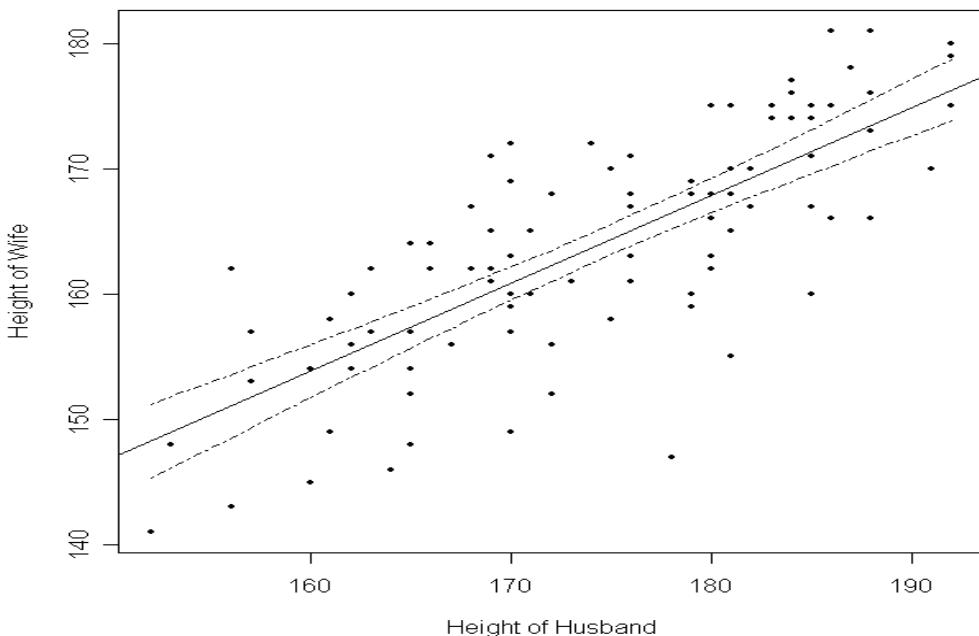
예제 2.1의 부부간 키 자료(accounting return)를 고찰해 보자. Table 2.7의 열은 96쌍의 각 y_i (아내의 키, height of wife), \hat{y}_i (적합된 아내의 키, fitted height of wife), $y_i - \hat{y}_i$ (잔차, residuals), $s_{\hat{y}_i}$ (예측의 표준오차, standard error of prediction), 그리고 y 에 대한 상위와 하위 95% 신뢰한계(upper and lower 95% confidence limits on y)를 나타낸다.

Table 2.7에서 결과의 해석 예는 다음과 같다. 14번째 부부의 아내의 실제 키는 162 cm이고 예측된(predicted) 아내의 키는 159.472 cm이다. 더 적절히 표현하면 회귀모형에 의해 적합된 아내의 키는 159.472이다. 예측의 표준오차는 0.717718 cm이다. 여기서 회귀변수의 수치가 실제 회귀모형을 적합하는 데 사용했기 때문에 적합된 값의 표준오차라는 표현이 더 정확할 것이다. 14번째 남편의 키에 대해 아내의 평균적인 키의 95% 신뢰구간은 158.0469 cm와 160.897 cm 사이이다.

여러 항목(item)들이 여기서는 가치가 없다. 우리는 적합값(적합된 아내의 키를 이야기 합니다)과 예측(prediction) 사이의 구분을 한다. 예측(prediction)과 대응된 표준오차(corresponding standard error) 그리고 신뢰한계(confidence limits)는 내삽(interpolation) 또는 외삽(extrapolation)이 요구되는 곳의 회귀값(regressor value)에 적용된다(예를 들어 회귀변수(regressor variable)의 수치가 자료세트에 포함되지 않은 곳). 적합값(fitted value)의 표준오차는 단순선형회귀 경우에는 확실히 중요하다. 이것은 내삽에 대한 회귀모형의 가능성에 다소간의 적응증(indication)을 준다. 그러나, 이것은 외삽의 영역에서는 모형의 성능(model's performance)을 나타내지 못한다. 예측(또는 적합)의 표준오차(stdandard error)는 대부분 컴퓨터 패키지들에서 회귀계산결과(regression computer printout)의 표준 영역(stdandard part)이 되어있다. 이것은 분석가에게 중요한 정보가 될 수 있다. 그러나, 식(2.23)에서 x_i 이외 다른 수치가 사용되지 않는다면, 예측(prediction), 확실하게 외삽(extrapolation)은 적절하게

평가되지 못한다. 예측의 표준오차는 모든 x_0 에 대해 일정하지는(constant) 않다. 그러므로 이것은 $s_{\hat{y}}$ (오차표준편차, error standard deviation)와 혼동되어서는 안 된다. $s_{\hat{y}(x_0)}$ 의 양(quantity)은 회귀선의 질(quality)이 예측된 지점에서 매우 잘 작동(function)함을 뜻한다. 예를 들면, 14번째 부부에 대한 $s_{\hat{y}}$ 는 0.717718 cm인 반면 35번째 부부에 대해서는 이 표준오차가 1.43572cm이다. 예측 성능(prediction performance)에서 이 비동일성(nonuniformity)은 2.11 절에서 더 논의되고 Figure 2.7에서 묘사되었다. 이는 부부간 키자료에 대한 평균반응(mean response)의 신뢰구간(confidence interval)을 반영한다.

그림 2.7 평균반응의 95% 신뢰구간



예측(prediction)이란 단어의 의미(implication)에 관하여 혼동하는 경우가 흔히 있다. 명백히 통계량 $\hat{y}(x_0)$ ($x = x_0$ 에서의 회귀선 위의 점)는 평균반응의 추정값(estimate)으로 사용되고, 또 예측값(predicted value)으로도 사용될 수 있는 양용성(dual purpose)이 있다. (2.23)에 있는, 예측의 표준오차(standard error of prediction)는 평균반응(mean response)에 대한 신뢰구간(confidence interval)을 얻는데 사용된다. 그러나, 이것은 미래의 단일관측(future single observation)에 대해서 어떤 형태의 추론(inference)을 하는 데에는 적절하지 않다. 고정된 $x = x_0$ 에서 평균반응(mean response)은 중요하지 않다고 가정하자. 그 보다는, x_0 에서의 단일반응관측치(single response observation)에 어떤 형태(type)의 범위(bound)가 있는지에 관심이 있다. $x = x_0$ 에서의 단일관측치(single observation)인 y_0 는 $\hat{y}(x_0)$ 와

무관하다(independent)는 것을 고려하라. 우리는 $y_0 - \hat{y}(x_0)$ 로 시작하여 y_0 에서의 예측구간(prediction interval)을 구할 수 있고, 다음을 고려하여 표준화(standardize)할 수 있다

$$\begin{aligned} Var(y_0 - \hat{y}(x_0)) &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

따라서, $E[y_0 - \hat{y}(x_0)] = 0$ 이므로, 정규이론가정(normal theory assumption) 조건 하에서 다음과 같이 나타낼 수 있다.

$$\frac{y_0 - \hat{y}(x_0)}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim N(0,1)$$

여기서 σ 는 s 로 대체할 수 있으며

$$\frac{y_0 - \hat{y}(x_0)}{s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2} \quad (2.25)$$

(2.25)로부터 확률경계(probability bound) 또는 예측구간(prediction interval)은 y_0 에 위치할 수 있다(즉, y_0 가 $(1 - \alpha)$ 의 고정확률(fixed probability)로 포함되는 구간). 이 예측구간(prediction interval)은 다음과 같이 주어진다

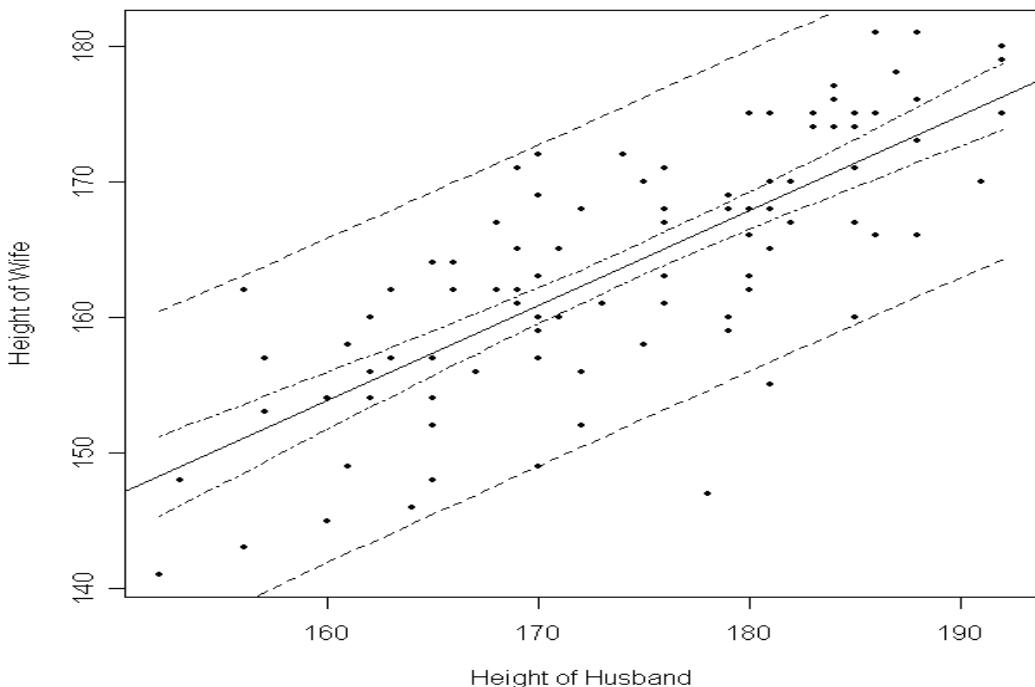
$$\hat{y}(x_0) \pm t_{\alpha/2, n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (2.26)$$

Fig 2.8은 신뢰구간(confidence interval)이 있는 동일한 플롯(plot) 상에 아내의 키에 대한 예측구간(prediction interval)을 보여준다. 물론 예측구간이 더 넓고, (2.26)에서 제곱근 내에서 큰 양(quantity)을 반영한다. 다음의 양(quantity)은

$$s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

예측구간(prediction interval)을 구하는데 표준오차와 같은 역할(a standard error type role)을 한다. 그러나, 이것은 예측된 값(predicted value) $\hat{y}(x_0)$ 의 표준오차(standard error)가 아니다. 그 보다는, 이것은 $y_0 - \hat{y}(x_0)$ 차이(difference)의 표준오차(standard error)이다.

그림 2.8 평균반응과 예측값의 95% 신뢰구간



종종 경험이 부족한 분석가는 fig 2.8에서 나타나는 영역 밖으로 떨어진 자료 포인트(data point)와 관련된 것을 해석하는데 곤혹스러워한다. 대역들(bands) 중 어느 하나의 바깥에 있는 점(a point outside one of the bands)을 의심스러운 관측치(suspect observation)의 신호(signal)로 해석하고 싶은 유혹을 느낄 것이다. 안쪽 대역들(Inner bands)는 모집단 평균반응(population mean response)에 대한 신뢰구간을 반영하고, 따라서 만약 관측치(observation)가 이 대역(band) 밖으로 떨어진다면, 그것에 대한 특별한 해석은 없다. 명백히, 관측치가 바깥쪽 대역들(outer bands)(i.e., those constructed for an individual observation)의 바깥쪽에 위치하는 경우는 더욱 흔하지 않고, 바깥쪽 대역들(outer bands)은 각각의 관측치(individual observation)가 적합모형(fitted model)에 적합된 정도(adquacy of fit)를 평가하는

지침(guideline)으로 사용하기에 더 적절한 것으로 보인다. 그러나, 일반적으로, 관측치(observation)가 예측구간(prediction interval)의 바깥에 위치한다고 해서 이 관측치를 자료세트(data set)에서 제거해도 되는 것은 아니다. 잔차(residual) 분석을 통한 바깥점(outlier)의 검출에 대해서는 5장에서 다룰 것이다.

모형의 예측능력을 변화시키는 요소(Factors Altering Prediction Capabilities of the Model)

어떤 요인들(factors)이 회귀모형(regression model)의 예측능력(prediction capability)에 영향을 미치는지 직관적으로 알 수 있다. 예를 들면, (2.22)에서 나머지 요인들이 동일하다는 가정 하에서, 표본크기(sample size)의 증가는 확실히 예측분산(prediction variance)을 감소시킨다.

또한, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ 양(quantity)이 증가하면 예측(prediction)이 향상되고, 이것은 회귀변수

(regressor variable)의 자료 범위(data range)에서 퍼짐(spread)이 커질수록 모형의 예측능력(prediction capability)이 향상된다는 것을 의미한다. 이것은 실험자가 x 값을 선택(control)할 수 있는 상황에서 중요한 고려사항이 될 수 있다. 주의할 점은 다음과 같다. 모형이 적절하게 설정(specify)되었다는 가정 하에서, 예측분산(prediction variance)은 예측의 질(quality of prediction)을 나타내는 적절한 측도(measure)이다. 만약 모형이 저설정(underspecify)되었다면 (즉, 너무 적은 모형항(model term)이 포함되었다면), 예측 반응(predicted response)에 편향(bias)이 생기게 되고, 이것은 어떻게 해서든지 고려해야만 하는 한 요인(a factor)이 되어 버린다. 만약 자료 분석가가 큰 S_{xx} 에 도달하기 위해 x 수준(x -level)에서 과도한 퍼짐(spread)을 사용한다면, 모형가정(model assumption)이 좋지 않게 왜곡될 위험에 처할 것이다. 예를 들면, x 의 범위가 좁은 경우 선형모형이 적절하겠지만 범위가 넓은 경우 곡선의 형태를 포함하는 모형이 적절할 것이다. 4장에서는 모형의 저설정(underspecification)과 과대설정(overspecification)을 논의할 때 편향(bias)에 대해 자세히 알아볼 것이다. 마지막으로, 직관적으로 알 수 있는 또 다른 결과는 x_0 (예측하고자 하는 지점)의 위치에 대한 것이다. 식(2.22)와 (2.23)은 x_0 가 \bar{x} (회귀변수의 평균)에 근접할 때 예측이 최상(best)이 될 것이라는 사실을 명확히 해준다. 다음에서 외삽(expolate)할 때 예측에 대해서 나쁜 결과가 얻어질 수 있음을 볼 수 있다. Fig 2.7과 2.8이 이것을 보여준다.

다음은 분석에 사용된 R-code이다

```
data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)

#table 2.8
pre<-predict(g,se=TRUE,interval="confidence")
```

```

res<-g$residuals
t2_8<-cbind(data,pre$fit[,1],res,pre$se.fit,pre$fit[,2],pre$fit[,3])
colnames(t2_8)<-c('H','W','fit_y','res','se_y','LCL','UCL')
n<-nrow(t2_8)

detach(data)
attach(t2_8)

#figure 2_7
plot(H,W,pch=20,xlab="Height of Husband",ylab="Height of Wife",main="그림 2.7 평균반응의 95% 신뢰구간")
abline(g)
lines(sort(H),LCL[order(H)],type="l",lty=4)
lines(sort(H),UCL[order(H)],type="l",lty=4)

detach(t2_8)

pre1<-predict(g,data,interval="prediction")
f2_8<-cbind(data,pre$fit[,2:3],pre1[,2:3])
colnames(f2_8)<-c('H','W','LCL','UCL','PLCL','PUCL')

attach(f2_8)
plot(H,W,pch=20,xlab="Height of Husband",ylab="Height of Wife",main="그림 2.8 평균반응과 예측값의 95% 신뢰구간")
abline(g)
lines(sort(H),LCL[order(H)],type="l",lty=4)
lines(sort(H),UCL[order(H)],type="l",lty=4)
lines(sort(H),PLCL[order(H)],type="l",lty=2)
lines(sort(H),PUCL[order(H)],type="l",lty=2)

```

2.10. 단순선형회귀에서 동시추론(Simultaneous Inference in Simple Linear Regression)

자료 분석가들이 여러 모수(parameters)에 대한 95% 신뢰구간 추정값(a confidence interval estimate)을 구할 때, 모든 모수들(all the parameter statements)에 대해서 신뢰 계수(confidence coefficient), 즉 신뢰도(degree of confidence)가 똑같이 95%라고 추정(presumption)하는 실수를 종종 범한다.

통계학 학생들은 통상적으로 그들의 학위 과정에서 꽤 일찍 결합신뢰영역(joint confidence regions)에 대해 배우게 된다. 사용자들은 모수들에 대한 결론을 도출해 내는 시도에서 이러한 중요한 개념을 너무 자주 무시한다. 예를 들면 회귀상황(regression setting)에서 β_0 와 β_1 에 대한 결합신뢰영역(joint confidence region) - 즉 β_0 와 β_1 둘 다 그 영역(region)에 들어가는 신뢰도가 95%인 영역-을 얻어내는(develop) 것이 흥미로울 수 있다.

이와 같은 관계에 있어서, 유사한 논란이 있을 수 있는데 예를 들면 fig 2.7에 있는 평균 반응의 신뢰구간 플롯(plots of confidence intervals on mean response)이 잘못 해석될 수도 있을 것이다. 그 그림은 포함된 96쌍의 부부중 아내의 키에 대한 상위 95% 신뢰구간 사이에 존재하는 연속적인 관계에 대해 나타내고 있다. 신뢰구간의 하한(Lower confidence bounds)에 대한 같은 플롯(plot)도 주어져 있다. 그러나, 사용자들은 95%의 확률(probability 0.95)로 모든 96개의 자료 위치(data locations)를 위한 구간(interval)이 동시에 아내의 키를 포함(cover)한다고 추론할 수는 없다. 자료 분석가들은 종종 fig 2.7에서 주어진 것과 같은 그림에서 진정한 회귀선(true regression line), $E(y|x) = \beta_0 + \beta_1 x_i$, 이 95%신뢰도를 가진 신뢰구간 내에 위치한다고 결론내기 위해 사용한다. 하지만, 이것은 절대로 사실이 아니다! 이번 절(section)에서 우리는 단순선형회귀(simple linear regression)에 있어서 동시신뢰영역(simultaneous confidence regions)에 대해 논의할 것이고 또한 적절한 곳에서 동시가설검정(simultaneous tests of hypotheses)에 대해 논의할 것이다. 이번 절(section)에 있는 것들은 3장에서 다중회귀(multiple regression)로 연결될 것임을 알아야 한다. 결론적으로 독자들은 다중선형회귀(multiple linear regression)라는 어쩌면 좀 더 복잡한 상황(setting)에서 표현된 것을 다시 보게 될 것이다.

β_0 와 β_1 의 결합신뢰영역(Joint Confidence Region on β_0 and β_1)

기울기(slope)와 절편(intercept)의 결합추정(joint estimation)은 응용분야(applications)에 있어서 종종 매우 중요하다. 신뢰영역 접근방법(confidence region approach)은 분석가로 하여금 점추정량(point estimators) b_0 와 b_1 의 정밀함(precision)을 알 수 있게 해준다. 반복되는 실험을 통해 100(1- α)% 신뢰영역을 구했을 때, 이 영역의 100(1- α)%는 β_0 와 β_1 모두를 포함한다.

2.6 절과 2.9 절에서 토의된 신뢰영역(confidence regions)의 경우에서와 같이, 정규성이론가

정(normal theory assumptions)이 필요하다. 선형모형이론(theory of linear model)으로부터(Searle, 1971 혹은 Graybill, 1976을 참고하시오),

$$[b_0 - \beta_0, b_1 - \beta_1] \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} b_0 - \beta_0 \\ b_1 - \beta_1 \end{bmatrix} \sim 2s^2 F_{2, n-2}$$

이러한 표현은 신뢰영역(confidence region)을 전개시키는데 사용 될 수 있다. 독자들은 위의 표현은 σ^2 를 제외한 $[b_0, b_1]$ 의 분산공분산행렬(variance-covariance matrix)의 역(inverse)이라는 사실을 알 수 있다. 부등식(inequality)이 확률 $1-\alpha$ (probability $1-\alpha$)로 유지되기에 신뢰영역(confidence region)을 전개하는데 이를 사용할 수 있다.

$$[b_0 - \beta_0, b_1 - \beta_1] \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} b_0 - \beta_0 \\ b_1 - \beta_1 \end{bmatrix} \leq 2s^2 F_{\alpha, 2, n-2}$$

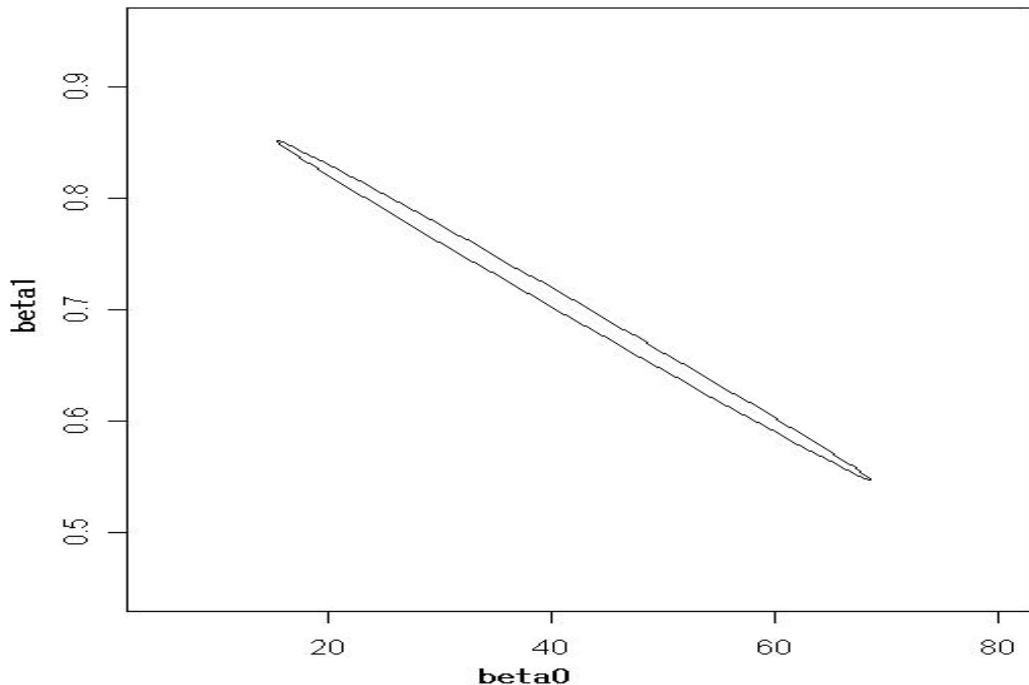
결과적으로 부등식(inequality)을 만족시키는 모든 β_0 와 β_1 조합들은 $100(1-\alpha)\%$ 신뢰영역(confidence region)의 안쪽(inside)에 위치한다. 물론 여기서 $F_{\alpha, 2, n-2}$ 는 F -분포(F -distribution)에서 상위 α 번째 백분위수(upper tail α th percentage point)이다.

β_0 와 β_1 에 대한 결합신뢰영역(joint confidence region)은 그래프 방식(graphical way)으로 사용하기 쉬운 타원형의 영역(elliptical region)을 형성한다. 타원형의 중심은 언제나 점(point) (b_0, b_1) 일 것이다. 명백하게 타원형을 이루는 방정식은 단지 부등식(inequality)의 좌변(left-hand side)과 우변(right-hand side)을 같다고 함으로써 유도된 것이다. 다음 예는 한 실례를 보여준다.

예제 2.7

Table 2.1의 부부간 키의 자료를 다시 고려해 보라. b_0 와 b_1 의 값은 각각 41.9302와 0.69970이다. 결합신뢰영역(joint confidence region)에 대한 식은 fig 2.9에서 나타난 타원형으로 산출해내는 것에 익숙해져 있다. 타원형 안의 어느 (β_0, β_1) 조합이라도 결합 95% 신뢰영역(joint 95% confidence region)안에 있다.

그림 2.9 공동신뢰영역



선형회귀선을 위한 신뢰대역(Confidence Band for the Linear Regression Line (Working-Hotelling Procedure))

단순선형회귀(simple linear regression)의 많은 적용에 있어서 전체회귀선(entire regression line)에 대한 신뢰영역(confidence region)을 가진다는 것은 흥미로운 것이다. 이것은 결합신뢰영역(joint confidence region)의 타원형그림(elliptical plot)에 있는 β_0, β_1 조합에 의해 산출된 모든 가능한 회귀(regression)들을 고려함으로써 발전되었다는 것을 독자들은 알아야 한다. 적합된 회귀선(fitted regression line) 주위의 쌍곡선 신뢰구간(hyperbolic confidence region)은 (2.24)의 것과 유사한 식을 사용함으로써 쉽게 산출될 수 있는데 그것은 다음과 같다.

$$\hat{y}(x_0) \pm \text{constant} \cdot s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

바꾸어 말하면 x 의 어느 위치(x_0)에서도 대역(band)의 폭(width)은 여전히 예측(prediction)의 표준오차(standard error)를 포함한다. 그러나, 여기에서의 적절한 상수(appropriate constant)는 t -분포의 백분위수(percentage point)가 아니다. $100(1-\alpha)\%$ 신뢰대역(confidence band)의 경계(boundary)는 다음과 같은 식에 의해 주어진다.

$$\hat{y}(x_0) \pm \sqrt{2 \cdot F_{\alpha/2, n-2} S^2} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

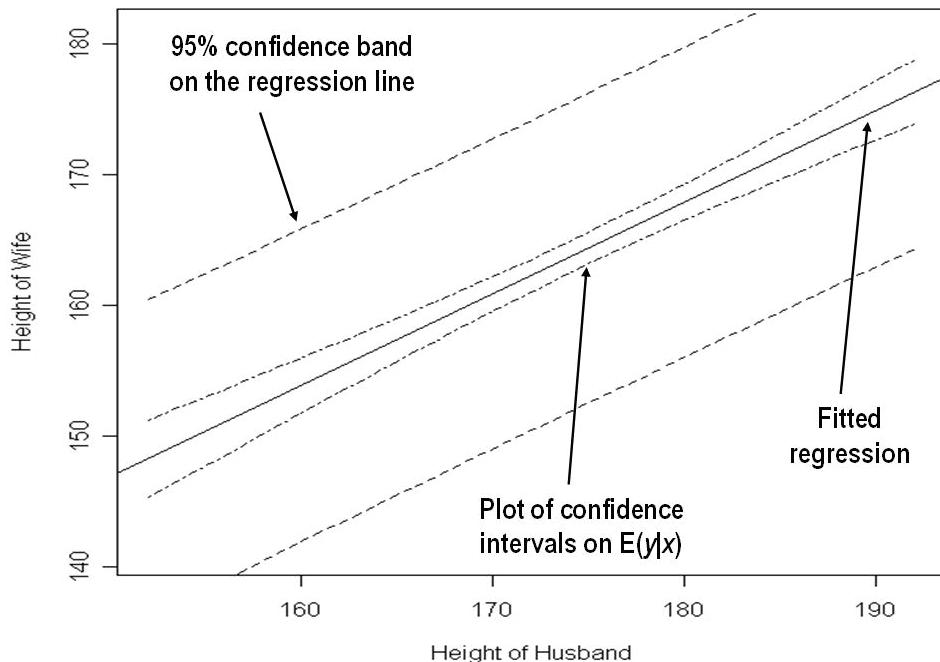
β_0 와 β_1 에 대한 결합신뢰영역(joint confidence region)의 경우처럼, 회귀선(regression line)에 대한 신뢰대역(confidence band)은 그것을 그래프로(graphically) 볼 수 있을 때 실험자(experimenter)에 의해 가장 잘 사용 될 수 있을 것이다. 부부간 키자료를 사용한 다음 예는 도움이 될 수 있을 것이다.

예제 2.8

Table 2.1의 부부간의 키 자료를 살펴보자. Fig 2.10은 “진정한” 회귀선(“true” regression line) $E(y|x) = \beta_0 + \beta_1 x$ 에 대한 적합된 회귀선(fitted regression line)과 95% 신뢰대역(confidence band)을 보여 준다. 또한 2.9절에서 논의된 $E(y|x)$ 에 있어서 신뢰구간(confidence interval)의 표준플롯(standard plot)도 보여준다. 회귀선(Regression line)에 있어서 대역(band)을 의미하는 Working-Hotelling approach가 조건부 신뢰구간(conditional confidence interval)의 플롯(plot)보다 얼마나 더 보수적(conservative)인지 확인할 수 있다..

Figure 2.10 Fitted regression and 95% confidence band on the regression line for the height of couples data

그림 2.10 부부간 키 자료의 회귀선과 신뢰구간



예제에 사용된 R-code는 다음과 같다.

```
data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
pre<-predict(g,se=TRUE,interval="confidence")
pre1<-predict(g,data,interval="prediction")
f2_8<-cbind(data,pre$fit[,2:3],pre1[,2:3])
colnames(f2_8)<-c('H','W','LCL','UCL','PLCL','PUCL')
detach(data)
attach(f2_8)
plot(H,W,pch=20,xlab="Height of Husband",ylab="Height of Wife",main="그림 2.8 평균반응과 예측값의
95% 신뢰구간")
abline(g)
lines(sort(H),LCL[order(H)],type="l",lty=4)
lines(sort(H),UCL[order(H)],type="l",lty=4)
lines(sort(H),PLCL[order(H)],type="l",lty=2)
lines(sort(H),PUCL[order(H)],type="l",lty=2)
```

단순선형회귀의 동시 추론에 있어서 본페로니 접근법(Bonferroni Approach to Simultaneous Inference in Simple Linear Regression)

지금까지 β_0 와 β_1 에 대한 동시신뢰영역(simultaneous confidence region)과 회귀선(regression line)에 대한 신뢰대역(confidence band)을 다루고 있다. 부부간 키 자료의 예에서 살펴본 바와 같이, 두 방법(tools) 모두 플롯(plots)을 사용하여 신속하게 적용할 수 있다. 그러나 실제 상황에서 플롯(plot)을 이용하지 못할 수 있다. 종종 각 모수(parameter)에 대하여 개별적인 구간(separate intervals)을 사용하는 것이 편리한데, 물론, 이 구간들(intervals)의 결합 포함(joint coverage) 확률은 적어도 $100(1-\alpha)\%$ 이다. 또한, 플롯(plots)에 의존한다면, 다중회귀(multiple regression case)에서 회귀계수(regression coefficients)에 대한 결합신뢰영역(joint confidence region)이나 회귀관계(regression relationship)에 대한 신뢰영역(confidence region, 회귀[regression]에 대한 신뢰대역[confidence band]과 유사)을 다루기에 어려울 수도 있다. 따라서, 회귀계수(regression coefficients)에 대해 다음과 같은 신뢰구간(confidence interval)을 가지고

$$\begin{aligned} b_0 &\pm \text{constant} \cdot (\text{standard error of } b_0) \\ b_1 &\pm \text{constant} \cdot (\text{standard error of } b_1) \end{aligned}$$

계수(coefficients) β_0 와 β_1 이 동시에 구간(interval) 내에 있을 신뢰도(confidence)가 적어도 $100(1-\alpha)\%$ 가 되게끔 하는 상수(constant)를 얻는 것은 흥미롭다. 기대하는 바와 같이, 그것의 의미는, 만약 반복된 표본(samples)이 취해지고 반복된 회귀(regressions)에 의해서 다른 값의 b_0 와 b_1 이 얻어진다면, 적어도 그러한 b_0 와 b_1 조합(combinations)의 95%는 적절한 모수(parameter)를 포함(covering)하는 양 구간(both intervals)을 가질 것이라는 점이다.

두 모수(two parameters)에 대한 신뢰구간(confidence interval)을 구하기 위한 본페로니 절차(Bonferroni procedure)는 만약 $100(1-\alpha)\%$ 신뢰구간(confidence interval)이 각 모수(each parameter)에 적용(apply)된다면 다음 식이 성립한다는 사실에 토대를 두고 있다.

$$\boxed{\Pr[\bar{B}_0 \cap \bar{B}_1] \geq 1 - 2\alpha}$$

위 식은 단순선형회귀(simple linear regression)에 대해서 본페로니 부등식(Bonferroni inequality)을 적용한 것이다. 여기서 $\Pr[B_0]$ 은 β_0 에 대한 신뢰구간(confidence interval)이 참모수(true parameter)를 포함(cover)하지 않을 확률(probability)이고, $\Pr[B_1]$ 은 β_1 에 대한 신뢰구간(confidence interval)이 참모수(true parameter)를 포함(cover)하지 않을 확률(probability)이다. 그러므로, $\Pr[\bar{B}_0 \cap \bar{B}_1]$ 은 두 신뢰구간(two confidence intervals)이 그들의 각각의 모수를 동시에

포함(cover)할 확률이다. 이러한 형식의 본페로니 부등식(Bonferroni inequality)은 확률(probability)의 몇몇 기본 개념들로부터 전개될 수 있다. 그 전개 과정은 다음과 같다.

$$\Pr[B_0 \cup B_1] = \Pr[B_0] + \Pr[B_1] - \Pr[B_0 \cap B_1]$$

또한, 아래와 같은 사실로 인해

$$\Pr[\bar{B}_0 \cap \bar{B}_1] = 1 - \Pr[B_0 \cup B_1]$$

다음과 같이 얻을 수 있고

$$\Pr[\bar{B}_0 \cap \bar{B}_1] = 1 - (\Pr[B_0] + \Pr[B_1]) + \Pr[B_0 \cap B_1]$$

$\Pr[B_0 \cap B_1] \geq 0$ 이기 때문에 아래와 같이 얻을 수 있고

$$\boxed{\Pr[\bar{B}_0 \cap \bar{B}_1] \geq 1 - \Pr[B_0] - \Pr[B_1]}$$

$\Pr[B_0] = \Pr[B_1] = \alpha$ 이기 때문에 다음과 같다.

$$\Pr[\bar{B}_0 \cap \bar{B}_1] \geq 1 - 2\alpha$$

만약 β_0 와 β_1 의 신뢰구간(confidence intervals)에 대한 신뢰계수(confidence coefficients)가 모두 0.95, 즉 $1 - 0.05$ 라면, β_0 와 β_1 의 결합추정(joint estimation)에 대한 신뢰계수(confidence coefficient)는 적어도 0.90일 것이다. 따라서 그러한 구간 쌍(such pairs of intervals)의 적어도 90%는 두 모수(parameters)를 동시에 포함(cover)할 것이다.

여기서 독자들이 분명히 알아야 할 것은 결합신뢰계수(joint confidence coefficient)가 적어도 $1 - \alpha$ 가 되도록 하기 위한 구간(interval)을 산출하기 위해 어떻게 본페로니 부등식(Bonferroni inequality)을 이용하는가에 관한 것이다. 이것은 각각 신뢰계수(confidence coefficient) $1 - \alpha/2$ 를

가지는 β_0 와 β_1 에 대한 개별신뢰구간(individual confidence interval)을 계산(computing)함으로써 얻을 수 있다. β_0 와 β_1 에 대한 100(1-a)% 결합신뢰구간(joint confidence interval)은 다음과 같다.

$b_0 \pm t_{\alpha/4,n-2}$ (standard error of b_0)
$b_1 \pm t_{\alpha/4,n-2}$ (standard error of b_1)

예제 2.9 Bonferroni Confidence Bound for the Height of Couple Data

Table 2.7의 부부간 키 자료에 주의를 가져보라. 한 쌍의 95% 결합신뢰구간(joint confidence intervals)은 다음과 같이 주어질 수 있는데

$$b_0 \pm t_{0.0125,94}(10.66162)$$

$$b_1 \pm t_{0.0125,94}(0.06106)$$

그 값은

$$t_{0.0125,94} = 1.985523, b_0 = 41.9302, \text{ 그리고 } b_1 = 0.6997$$

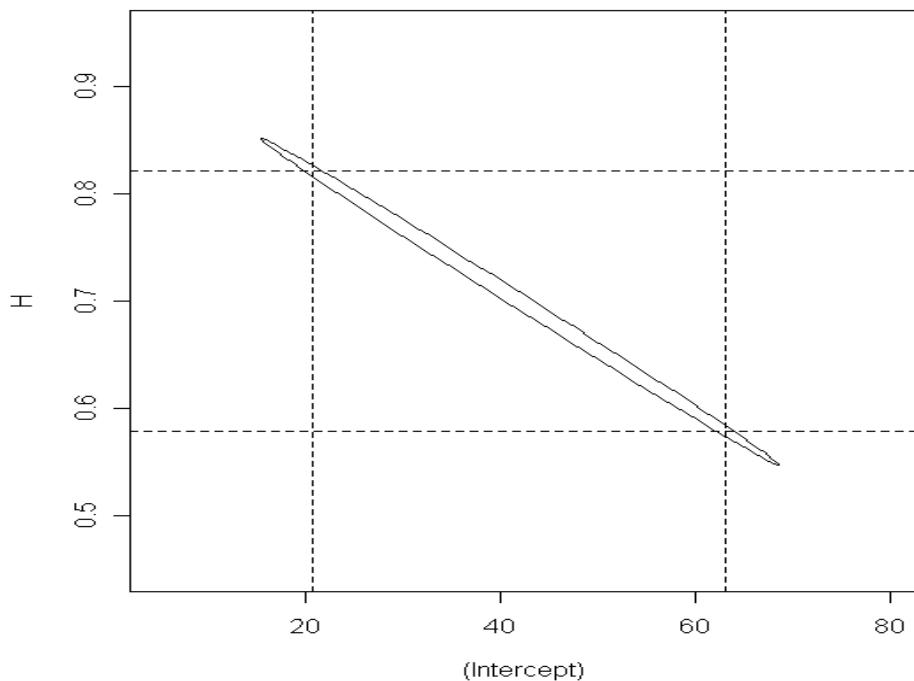
결과적으로 계산된 구간은

$$41.9302 \pm 21.16889$$

$$0.6997 \pm 0.12124$$

예제 2.7에서 주어진 β_0 와 β_1 에 대한 타원형결합신뢰영역(elliptical joint confidence region)과 부부간 키자료에 대한 본페로니 신뢰구간(Bonferroni confidence interval)간의 차이를 보이기 위해 fig 2.11를 보라. 본페로니 신뢰영역(Bonferroni confidence region)에 의해 포함(cover)되는 영역은 훨씬 더 크다는 것에 주목하라. 직사각형 영역(rectangular region)은 타원형 영역(elliptical region)에 비해 훨씬 덜 유효한 영역을 만들어 낸다. 독자들은 본페로니 접근법(Bonferroni approach)이 적어도 95%의 신뢰도(confidence)에 대한 영역(region)을 초래한다는 것을 염두에 두고 있어야 한다. 일반적으로 그것은 좀더 보수적(conservative) 결합신뢰영역(joint confidence region)을 산출할 것이다.

그림 2.11 공동신뢰영역



예제에 사용된 R-code는 다음과 같다.

```

data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
library(ellipse)
#figure 2.11
plot(ellipse(g),type="l",xlim=c(5,80),ylim=c(0.45,0.95),main="그림 2.11 공동신뢰영역")
abline(v=conf[1],lty=2)
abline(h=conf[2],lty=2)

```

예제 2.10 Bonferroni Confidence Intervals for Mean Response for height between couples

우리가 table 2.1에 있는 부부간 키 자료를 고려하고 있다고 가정해 보라. 부부간 키자료의 네 가지 다른 값에 대한 동시 추정 평균 키에 흥미가 있다고 가정하자. 이 값들은

$$x_1 = 160 \text{ cm}$$

$$x_2 = 165 \text{ cm}$$

$$x_3 = 170 \text{ cm}$$

$$x_4 = 175 \text{ cm}$$

평균 키의 점추정값(point estimates)과 95% 결합신뢰구간(joint 95% confidence intervals)은 다음과 같이 주어진다.

$$\hat{y}_i \pm t_{.05/8,94} s \sqrt{\frac{1}{96} + \frac{(x_i - \bar{x})^2}{S_{xx}}}, \quad i = 1,2,3,4$$

네 가지 모수(parameters)가 추정 되어지기 때문에 factor 8이 사용되는데, 즉 추정되는 네 가지 평균반응(mean responses)이 있다. $t_{.05/8,94}$ 값이 2.5467로 주어진다. 결과적으로 신뢰구간(confidence intervals)은 다음과 같다.

[152.3288, 155.4207]
 [155.8300, 158.9160]
 [159.3300, 162.4126]
 [162.8287, 165.9104]

예제에 사용된 R-code는 다음과 같다.

```
data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
x_temp<-c(160,165,170,175)
x0<-data.frame(x_temp)
colnames(x0)<-'H'
hat.y<-predict(g,x0)
tv<-abs(qt(0.05/8,94))
sxx<-sum(data[,1]^2)
m.x<-mean(data[,1])
lcl<-hat.y-tv*s_stat$sigma*sqrt(1/96+(x0-m.x)^2/sxx)
ucl<-hat.y+tv*s_stat$sigma*sqrt(1/96+(x0-m.x)^2/sxx)

conf_x0<-cbind(lcl,ucl)
```

평균반응을 위한 세가지 신뢰구간방법의 사용(Use of Three Confidence Interval Methods for Mean Response)

예제 2.10은 본페로니 신뢰구간(Bonferroni confidence interval)의 또 다른 예(illustration)를 보여준다. 명백하게 만약 자료 분석가가 추정값평균반응(estimate mean response)에 대해 적합 회귀(fitted regression)를 사용하고 신뢰구간(confidence interval)에 첨가하기(attach)를 원한다면 세 가지 방법이 사용될 수 있다: 2.9절에서 논의 되었던 $E(y|x)$ 에 대한 표준신뢰구간(standard confidence intervals)을 이용하는 방법과 회귀선에 대한 신뢰대역(confidence band)을 이용하는 방법 그리고 마지막으로 본페로니접근법이 있다. 결합신뢰구간(joint confidence interval)이 목적일 경우 첫 번째 방법은 부적절하다. 일반적으로 단지 소수의 예측(prediction)이 이루어진다면, 본페로니방법이 좀 더 효율적일 것이다. 양자의 경우에 있어 “제어된”(controlled) 신뢰계수(confidence coefficient)는 하한(lower bound)이다. (예제 2.14를 참조)

단순선형회귀에서 동시검정으로의 확장(Extension to Simultaneous Tests in Simple Linear Regression)

우리는 단순선형회귀에서 동시신뢰영역(simultaneous confidence region)에 대해서 설명하고, 전개하였다. 또한 기울기와 절편의 결합추정(joint estimation)과 평균반응(mean response)의 결합추정(joint estimation)에 초점을 맞추었다. 이전에 지적한 것처럼, 결합신뢰영역(joint confidence region)의 목적은 추정된 모든 모수(parameter)가 동시에 계산된 영역(computed region)안에 위치한다는 신뢰도(confidence)(가령 95%)를 조절(control)하는 것이다. 같은 개념을 여러 개의 모수에 대한 가설 검정(hypothesis testing)에도 사용할 수 있다. 가설 검정(hypothesis testing)에서 결합(joint)이라는 개념은 1종 오류(Type I error) 혹은 α error의 확률을 여러 개의 모수(multiple parameters)에 대해 동시에 고려한다는 것을 의미한다. 기울기(slope)와 절편(intercept)의 경우를 예로 든다면, 우리는 다음과 같은 것을 검정(test)해보고 싶을 것이다.

$$H_0 : \begin{cases} \beta_0 = \beta_{0,0} \\ \beta_1 = \beta_{1,0} \end{cases}$$

H_1 : 적어도 귀무가설 H_0 중의 하나는 참이 아니다.

이 가설을 0.05 수준에서 검정한다는 것은, 실제로 전체 가설(entire hypothesis)이 참일 때 H_0 중에서 적어도 하나의 가설(statement)이 거짓이라고 결론내릴 확률이 0.05라는 것을 의미한다.

모수(parameter)가 한 개일 경우, α 수준에서 귀무가설 $H_0 : \beta_1 = \beta_{1,0}$ 을 기각하는 것은 β_1 에 대한 100(1- α)% 신뢰구간 추정값(confidence interval estimate)을 산출하고 $\beta_{1,0}$ 가 다음의 신뢰구간 바깥(outside of the confidence interval)에 위치함을 보여주는 것과 같다는 점을 증명하기란

쉽다.

$$b_1 \pm t_{\alpha/2, n-2} \cdot (\text{standard error of } b_1)$$

물론 이 신뢰 구간을 다음과 같이 나타낼 수도 있다.

$$b_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{S_{xx}}}$$

절편에 대한 검정도 마찬가지이다. 그러므로, 양쪽 개별 가설(both individual hypothesis)을 검정한다면, 결합진술(joint statement)에 대한 1종 오류(Type I error)의 확률은 적어도 하나의 귀무값(null values, $\beta_{0,0}$ or $\beta_{1,0}$)이 모수에 대한 신뢰구간(confidence interval)의 바깥에 있을 확률이다. 이 확률은 다음과 같다.

$$1 - \Pr[\bar{B}_0 \cap \bar{B}_1] \leq 1 - (1 - 2\alpha) \leq 2\alpha$$

이와 같이, $\alpha = 0.05$ 수준(level)의 검정(test)이 두 모수(two parameters)에 대해 수행된다면, 1종 오류(Type I error)의 결합확률(joint probability)은 0.10의 상한(upper bound)을 가진다.

독자들은 이제 개개 모수(parameters)에 대한 검정(tests)과 신뢰구간추정값(confidence interval estimation) 사이에 존재하는 것과 같은 유형의 관계(relationship)가 동시검정(simultaneous tests)과 동시신뢰구간(simultaneous confidence intervals) 사이에 존재한다는 것을 짐작할 수 있을 것이다.

β_0 와 β_1 의 동시검정(Simultaneous Tests on β_0 and β_1)

β_0 와 β_1 에 대한 결합가설(joint hypothesis)의 기각(rejection)은 $\beta_0 = \beta_{0,0}$ 와 $\beta_1 = \beta_{1,0}$ 일 때 확률 α 로 발생한다. β_0 와 β_1 에 대한 타원형의 신뢰영역(elliptical confidence region)의 전개(development)로부터 다음을 알 수 있는데, 귀무 가설하에서 즉 $\beta_0 = \beta_{0,0}$ 와 $\beta_1 = \beta_{1,0}$ 일 때,

$$\left[b_0 - \beta_{0,0}, b_1 - \beta_{1,0} \right] \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} b_0 - \beta_{0,0} \\ b_1 - \beta_{1,0} \end{bmatrix} \sim 2s^2 F_{2,n-2}$$

검정(test)에 대한 적절한 검정절차(test procedure)(Graybill, 1976 참조)에서 아래식을 만족하면 H_0 를 기각(rejection)한다.

$$\left[b_0 - \beta_{0,0}, b_1 - \beta_{1,0} \right] \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \begin{bmatrix} b_0 - \beta_{0,0} \\ b_1 - \beta_{1,0} \end{bmatrix} \geq 2s^2 F_{\alpha,2,n-2}$$

이것은 $(\beta_{0,0}, \beta_{1,0})$ 의 조합(combination)이 β_0 와 β_1 의 $100(1-\alpha)\%$ 신뢰영역(confidence region) 바깥에 있을 때 아래의 결합가설(joint hypothesis)을 α 수준에서 기각하는 절차와 동일하다.

$$H_0 : \begin{cases} \beta_0 = \beta_{0,0} \\ \beta_1 = \beta_{1,0} \end{cases}$$

또, 타원형 신뢰영역(elliptical confidence region)의 플롯(plot)을 관찰함으로써 이러한 과정을 쉽게 수행할 수 있다.

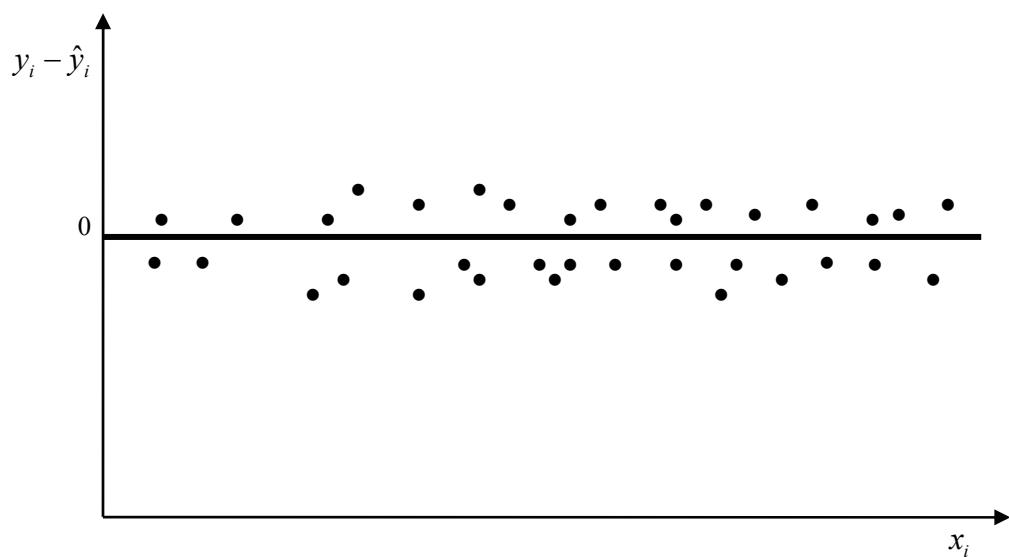
본페로니 신뢰영역(Bonferroni confidence region)은 또한 결합가설(joint hypothesis)을 검정하는 데 이용할 수 있다.. 단지 β_0 와 β_1 에 대한 개별 $100(1-\alpha/2)\%$ 신뢰구간을 산출하고 적어도 귀무값(null values)($\beta_{0,0}$ 또는 $\beta_{1,0}$)중 하나가 그 모수(parameter)에 대한 신뢰구간의 바깥에 위치하면 α 수준에서 H_0 를 기각(rejection)한다. 신뢰구간의 경우에서처럼, 본페로니 접근법(Bonferroni approach)을 사용하게 되면 보존의 대가(price of conservation)를 치러야 한다. 가설 검정(hypothesis testing)의 경우에서, 본페로니 신뢰영역 (Bonferroni confidence region)에 의해 포함(cover)되는 사각형의 지역(area)이 타원형 신뢰영역(elliptic confidence region) (Fig 2.11을 참조)에 의한 것 보다 훨씬 더 크기 때문에 β_0 와 β_1 의 더 많은 조합이 수용영역(acceptance region)에 포함된다; 이것은 기각(rejection)의 더 낮은 확률 혹은 검정(test)의 더 낮은 검정력(power)을 초래한다. 신뢰구간(confidence interval)의 크기를 결정하는 선택된 α 수준이 1종 오류(Type I error)의 참확률(true probability)의 상한(upper bound)이라는 점을 기억하길 바란다. β_0 와 β_1 의 동시검정(simultaneous testing)에 대한 실례(illustration)를 위해서 예제 2.15를 살펴보라.

2.11. 잔차 조망(a Look at Residuals)

이 절에서는 적합오차(errors of fit)로 불리는 보통잔차(ordinary residuals), 즉 $e_i = y_i - \hat{y}_i$ 로부터 얻을 수 있는 정보의 유형에 대한 예비(preliminary) 표현(presentation)을 다루게 된다. 세세한 것은 여기에서 언급하지 않을 것이고, 다중회귀(multiple regression)에서 나오는 개념들이 잔차의 통계적 특성(statistical properties of residuals)을 나타내 주기 때문에 잔차(residual)에 대해서는 5장에서 자세하게 언급할 것이다. 그렇지만 이 절에서 독자들은 잔차분석과 플로팅(plotting)에서 얻어지는 정보(infromation)에 대한 통찰력(insight)를 갖게 될 것이다.

잔차(residual)는 회귀분석의 조건(conditions)이 이상적인지 아닌지를 판단하는 양(quantities)으로 인식해야 한다. 이상적인 ε_i 에 대한 가정(assumption)을 생각해보자. $i=1, 2, \dots, n$ 에 대해서 $E(\varepsilon_i) = 0, E(\varepsilon_i)^2 = \sigma^2$ 이라는 가정의 중대한 위반(severe violation) 여부는 잔차를 분석함으로써 종종 확인할 수 있다. 독자들은 잔차를 모형의 오차(ε_i)에 가장 근접하게 필적하는(emulate) 양(quantity)으로 사용한다. 회귀변수(regressor variable) x 에 대한 잔차 플롯(plot of residual)의 이상적인(ideal) 형태(appearance)는 fig 2.12에서 나타난 바와 같이 0을 중심으로 임의값(random values)이 모여 있어야 한다. 최소제곱과정(least squares procedure)의 특성상 $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$ 이다.

Figure 2.12 단순선형회귀에서의 x_i 에 대한 $(y_i - \hat{y}_i)$ 의 이상적인 그림



모형이 오설정되면 무슨 일이 생길까? (What Happens if Model is Misspecified?)

비이상적인 조건(nonideal condition)을 어떻게 선택하는지를 더 잘 이해하기 위해서 가정(assumption) 중의 한가지가 위반(violation)되었을 때 잔차(residual)가 어떻게 작용(behavior)하는지를 알아보는 것이 도움이 될 수 있다. 자료 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 이 다음과 같은 모형에서 생성되었다고 가정해보자.

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2 + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (2.27)$$

여기에서 $i=1, 2, \dots, n$ 에 대해서 $E(\varepsilon_i) = 0$, $E(\varepsilon_i)^2 = \sigma^2$ 이다. 즉 (2.27)의 2차회귀모형 (quadratic regression model)과 관련된 조건은 이상적이다. 하지만, 분석가는 (2.27)의 모형을 적합시키지 않고 대신에 아래에 있는 단순선형회귀모형을 적합시켰다고 생각해보자.

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i^* \quad (i = 1, 2, \dots, n) \quad (2.28)$$

분석가에게 알려져 있지 않은 $\varepsilon_i^* = \varepsilon_i + \beta_2 x_i^2$ 즉, 모형에 있어서의 오차(error for the analyst's model)는 순전한 임의오차(truly random error)가 아니라 체계적 요소(systemic component)인 $E(\varepsilon_i^*) = \beta_2 x_i^2$ 도 가지고 있다. 이제 이것이 결과에 어떠한 영향을 주며 잔차(residual)로 어떻게 전달(transmit)되는 것일까? (2.28)의 부정확한(incorrect) 단순선형모형(simple linear model)의 잔차가 여전히 평균 0을 중심으로 분포하게 될 것인가? 적합된 단순선형모형을 생각해 보자.

$$\hat{y}_i = b_0^* + b_1(x_i - \bar{x})$$

최소제곱추정값(least squares estimates)은 다음과 같다.((2.7)식과 (2.8)식 참조)

$$b_0^* = \bar{y} = \frac{\sum_{i=1}^n [\beta_0^* + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2 + \varepsilon_i]}{n} = \beta_0^* + \beta_2 \bar{x}^2 + \bar{\varepsilon}$$

그리고

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) [\beta_0^* + \beta_1(x_i - \bar{x}) + \beta_2 x_i^2 + \varepsilon_i]}{S_{xx}}$$

$$= \beta_1 + \frac{n\beta_2(\bar{x}^3 - \bar{x}\bar{x}^2)}{S_{xx}} + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{S_{xx}}$$

이제 b_0^* 와 b_1 모두 편향된 추정량(biased estimator)이라는 것은 분명하다.

$$E(b_0^*) = \beta_0^* + \beta_2 \bar{x}^2 \quad (2.29)$$

$$E(b_1) = \beta_1 + \frac{n\beta_2(\bar{x}^3 - \bar{x}\bar{x}^2)}{S_{xx}} \quad (2.30)$$

이제 잔차(residual)는 어떻게 되는가? $y_i - \hat{y}_i$ 의 (여기에서 $\hat{y}_i = b_0^* + b_1(x_i - \bar{x})$ 이다) 성질(property)을 다룰 필요가 있다. 식(2.27), (2.29), (2.30)을 이용하면 다음과 같다.

$$\begin{aligned} E(y_i - \hat{y}_i) &= \beta_0^* + \beta_1(x_i - \bar{x}) \\ &\quad + \beta_2 x_i^2 - \left[\beta_0^* + \beta_2 \bar{x}^2 + \beta_1(x_i - \bar{x}) + \frac{(x_i - \bar{x})n\beta_2(\bar{x}^3 - \bar{x}\bar{x}^2)}{S_{xx}} \right] \\ &= \beta_2 \left[(x_i^2 - \bar{x}^2) + n(x_i - \bar{x}) \frac{(\bar{x}^3 - \bar{x}\bar{x}^2)}{S_{xx}} \right] \end{aligned}$$

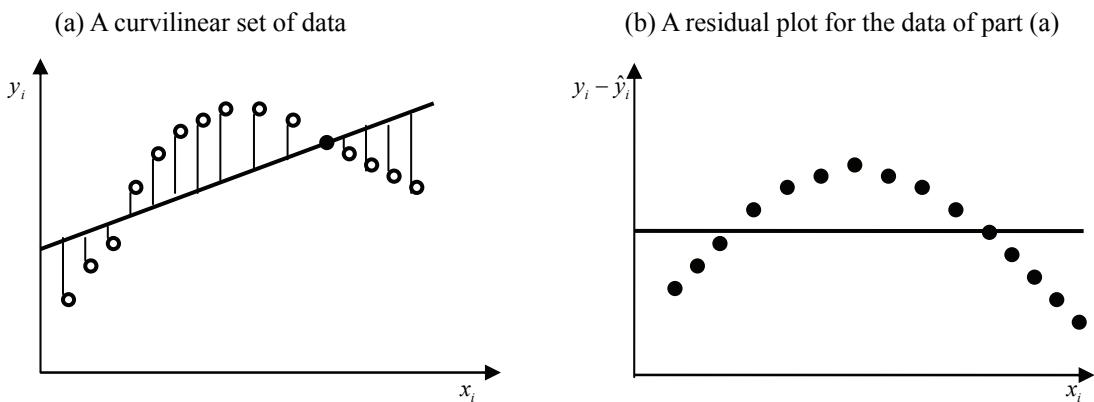
$$E(y_i - \hat{y}_i) = \beta_2(x_i^2 + rx_i + s) \quad (2.31)$$

여기에서 $r = -n(\bar{x}^3 - \bar{x}\bar{x}^2)/S_{xx}$, $s = -\left[\bar{x}^2 + \frac{n\bar{x}(\bar{x}^3 - \bar{x}\bar{x}^2)}{S_{xx}}\right]$ 이다.

$\beta_2 = 0$ 이 되는(즉, 적합된 모형이 정확한 경우) 특수한 경우에는 기대했던 바대로 잔차는 평균 0을 갖게 된다. 그렇지만 $\beta_2 \neq 0$ 이라면 i 번째 잔차의 기대값은 x_i 의 2차함수(quadratic function)가 될 것이다. 그 결과로 x 에 대한 잔차의 그림(plot)은 (2.31)식의 관계에서 확인할 수 있듯이 외연상 곡선형태(curvilinear type)를 띠게 된다. 또한 (2.29), (2.30)식을 통해 무시되었던 모형항(model term) $\beta_2 x_i^2$ 의 편향(bias)을 모형 계수(model coefficient) b_0 와 b_1

으로 전달함(transmit)을 알 수 있다. 잔차의 그림(plot)에서 볼 수 있는 형태를 fig 2.13에 제시하였다. Fig 2.13(a)의 수직편차(vertical deviation)는 fig 2.13(b)에서 잔차그림(residual plot) 형태로 전달(transmit)된다. 여기에서 볼 수 있듯이 잔차는 0을 중심으로 여전히 변동(fluctuate)하고 있지만 비임의(nonrandom) 형태인 것을 알 수 있다. 다른 형태의 잔차에 중점을 두고 모형의 오설정(model misspecification)을 찾아내기 위한 추가적인 방법들은 5장에서 살펴보기로 하겠다.

Figure 2.13



이분산의 검출에 있어서의 보조(Aid in Detection of Heterogeneous Variance)

잔차(residual)를 다루는 것은 이분산(heterogeneous variance)의 문제를 진단하는데 중요한 역할을 한다. 보통잔차(ordinary residual)의 그림이 유용할 수도 있지만, 5장에서 언급될 스튜던트화 잔차(studentized residuals)를 이용하는 것이 이 목적에는 더 선호된다. 따라서 다중회귀(multiple regression)의 형태일 경우에는 5장을 참조하여야 할 것이다.

정규성의 검출(정규확률그림, Detection of Normality, Normal Probability Plots)

이전 장에서 언급한 바와 같이 ε_i 가 정규성(normality)을 지닌다는 가정(assumption)은 단순선형회귀 방법의 일부 경우에서만 해당된다. 모수(parameter)에 대한 t -검정(t -test), F -검정(F -test), 신뢰한계(confidence limit)는 정규성을 따른다는 가정을 전제로 한다. 정규성을 따른다는 가정이 위배되는지를 알아보아야 하는 또 다른 이유로는 오차가 정규성을 지닐 때 최소제곱방법이 보다 바람직한 성질(properties)을 지니기 때문이다. 오차가 정규분포를 따르는지 그렇지 않은지의 여부에 따른 최소제곱과정의 성질에 대하여 3장에서 보다 상세하게 다를 것이고, 3장에서는 다중선형회귀(multiple linear regression)에 대해 일반화를 하게 될 것이다.

정규성을 알아내는 방법을 다루기에 앞서 정규성에서 약간 벗어난 것(minor departures)은 회귀결과(regression result)에 거의 영향을 미치지 않는다는 점을 알아야 한다. 또한 모형 오설정 (model misspecification)을 정규성에서 이탈(departure from normality)한 것으로 잘못 판단하는 경우도 있다. 적절한 모형을 완전히 찾아내지 못한 서툰 분석가의 경우, 실제로

예외(anomaly)가 오설정 모형(misspecified model)에서 생겨났음에도 불구하고 이 문제를 오차의 정규성 이탈에 의한 것으로 진단내릴 수도 있다. 아래의 전개과정이 나오듯이 비정규성 오차(nonnormal error)의 탐지(detection)는 $E(\varepsilon_i)=0$ 이라는 가정에 의존하며, 이것은 모형이 전체적으로 설정되지 않는다면(grossly unspecified) 유지되지 않는 조건(condition)이다.

다중선형회귀(multiple linear regression)에서 잔차의 세부성질(detailed property)에 따르면, ε_i 에 대한 표준가정(standard assumption)이 유지되더라도 잔차(residuals)는 서로 무관하지 않고(not uncorrelated), 공통분산(common variance)을 가지지도 않는다.(5장 참조) ε_i 에 대한 이상적인 조건이 유지될 때, 잔차(residuals) 간의 상관(correlation)은 심하지 않다. 그렇지만 일부 회귀자료세트(regression data set)에서 잔차는 그것의 분산과 많은 차이를 보이기도 한다. 많은 컴퓨터 플로팅 루틴(computer plotting routine)은 이것을 고려하지 않는다. 단순선형모형(simple linear model)의 경우를 생각해 보면

$$\begin{aligned} & \text{Var}(y_i - \hat{y}_i) \\ &= \text{Var } y_i + \text{Var } \hat{y}_i - 2\text{Cov}(y_i, \hat{y}_i) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) - 2\text{Cov}(y_i, \hat{y}_i + b_1(x_i - x)) \end{aligned}$$

이 식에서 공분산 항(covariance term)은 다음과 같이 쓰여진다.

$$\text{Cov}\left[y_i, \hat{y}_i + \frac{S_{xy}}{S_{xx}}(x_i - \bar{x})\right] = \frac{\sigma^2}{n} + \frac{\sigma^2(x_i - \bar{x})^2}{S_{xx}}$$

여기서 i 번째 잔차(residual)의 분산(variance)은 다음과 같이 표기할 수 있다.

$$\text{Var}(y_i - \hat{y}_i) = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right] \quad (2.32)$$

독자들은 $\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$ 부분을 $\text{Var } \hat{y}_i / \sigma^2$ 으로 인식해야 한다. 이는 2.9절에서 언급되었던

평균반응(mean response)의 신뢰구간(confidence interval)을 구하는 과정에 중요한 역할을 한다. 자료의 중심에 근접한 x 위치의 잔차는 중심에서 먼 위치에 있는 것보다 더 큰 분산(variance)을 갖는다. i 번째 잔차의 분산(variance)을 개념화하기 위해서는 동일한 x 값에 대해 반복해서 새로운 y 관측치(observation)를 얻음으로써 $x = x$ 에서의 잔차를 얻어 변이(variability)를 시각화할 수 있을 것이다.

잔차 사이에 분산이 동일하지 않으면(unequal variance), ε_i 에서의 비이상적인 조건(nonideal condition)(예를 들어 비정규성)을 탐지하는 방법으로 잔차를 이용하는데 몇 가지 어려움을

초래한다. 참모형(true model)의 오차(ε_i)에 더 잘 필적(emulate)하도록 잔차를 표준화(standardize)하는데 (2.32)식에서 배웠던 것을 궁극적으로 이용해야 한다.

비정규성을 검출하기 위해 무엇을 그려야 하니?(What Should be Plotted for Detection of Nonnormality?)

정규성으로부터 심각한 이탈을 발견해 내기 위해서 잔차의 정규확률그림(normal probability plot)을 그래프로 그려볼 수 있다. 이 그림(plot)은 평균이 μ 이고 분산이 σ^2 인 정규분포에서 독립적으로 표본을 추출해 낸다는 것을 전제로 하는 순서통계량(order statistics)의 통계 이론에 토대를 두고 있다. i 번째 가장 작은 값의 기대값(expected value)은 다음과 같다.

$$\mu + \sigma \mu_{(i)} \text{ 여기에서 } \mu_{(i)} \cong z \left(\frac{i - 0.375}{n + 0.25} \right)$$

여기에서 $z(w)$ 는 표준정규분포(standard normal distribution)의 $w(100)$ 백분위수를 의미한다. 즉 z 가 표준정규변량(standard normal variate)이라면 $\Pr(z \leq z(w)) = w$ 이다.

이와 같이 기대값($\mu_{(i)}$)에 대한 그림(plot)을 그릴 때 참모형 오차(true model error)(ε_i)는 ε_i 에서 $\mu=0$ 이기 때문에 원점을 지나는 직선에 합당하게 근접한 그림(plot)으로 나타난다. 물론 잔차는 분석가가 이용 가능한 양(quantities)이다. 정규확률그림(normal probability plot)은 잔차의 기대값에 대한 그림(a plot of residuals against their expected values)이다. 잔차가 알맞게 표준화된다면 기울기가 1이고 원점을 지나는 직선이 될 것이다. 두 가지 가능한 접근방식이 아래에 제시되었다.

(i) $z \left(\frac{i - 0.375}{n + 0.25} \right)$ 에 대하여 $\frac{(y_i - \hat{y}_i)}{s}$ 를 플로팅(plot)하라. 여기에서 $(y_i - \hat{y}_i)$ 는 i 번째 가장 작은 잔차(residual)이다.

(ii) 다음과 같이 표준화 형태(standardized form)의 잔차를 만들어라.

$$r_i = \frac{(y_i - \hat{y}_i)}{s_{(y_i - \hat{y}_i)}} \quad (2.33)$$

(2.32)식으로부터 잔차의 표준오차(standard error of residual)는 다음과 같다.

$$s_{(y_i - \hat{y}_i)} = s \sqrt{1 - \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]} \quad (2.34)$$

다시 $y_i - \hat{y}_i$ 는 i 번째 가장 작은 잔차를 의미한다.

두 경우에(in both cases), $z\left(\frac{i-0.375}{n+0.25}\right)$ 에 대하여 표준화된 형태의 잔차(standardized form of the residual)을 플로팅(plot)하라.

직선모양에서의 이탈(departure)은 정규성(normality)으로부터의 이탈(departure)을 의미한다. 물론, 일반적으로 잔차들(residuals)이 동일한 분산(equal variance)를 가지지 않으므로, 각각을 적절히 표준화(standardizing)하고 있다는 사실 하에 방법 (ii)를 이용한다.

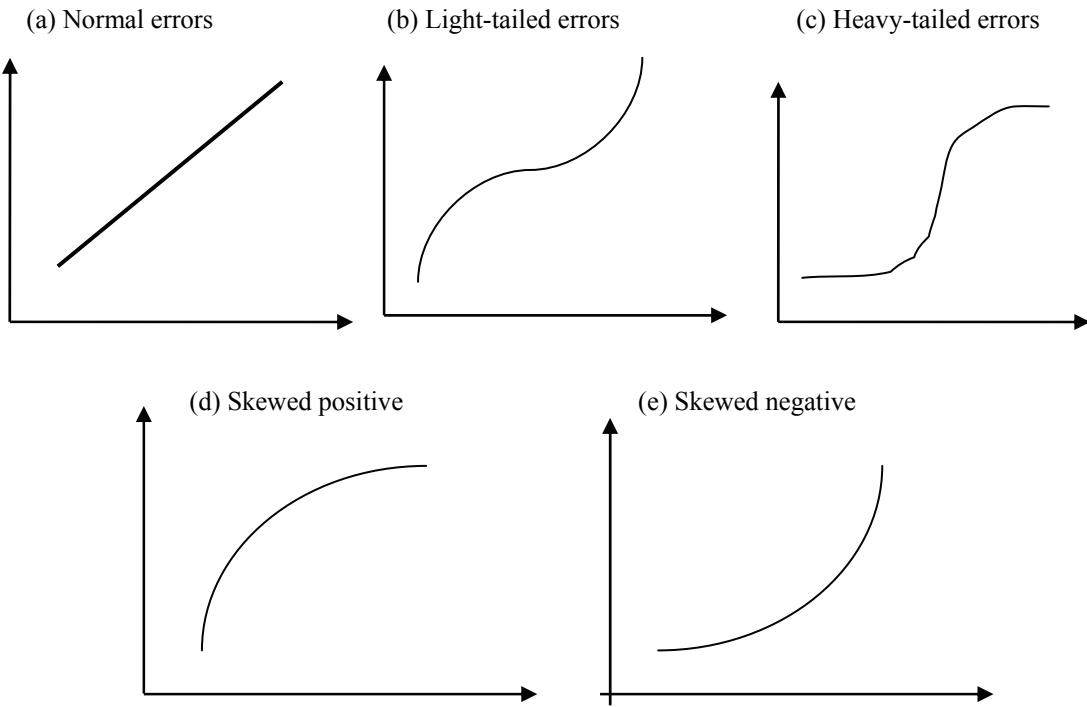
(2.33)식에서 주어진 잔차는 흔히 스튜던트화잔차(studentized residual)라고 불린다. 이것은 5장에서 다시 언급될 것이고 그 때 잔차를 이용하는 과정에 대해 좀 더 깊이 다루게 될 것이다. 예제 2.12에서 2.3절의 부부 키 자료에 대한 정규확률그림(normal probability plot)을 그리는데 두 방식(both methods)을 이용하였다. 값 $z\left(\frac{i-0.375}{n+0.25}\right)$ 는 i 번째 순서통계량(order statistic)

의 기대값(expected value)의 훌륭한 근사(approximation)를 제공할 것이다. 참값(true value)은 부록 C의 table C.5에 나와 있으며 종종 rankits라고 불린다. n 이 20일 때까지 표에 나와 있다. 상업화된 많은 소프트웨어 패키지를 이용하면 정규확률그림(normal probability plot)을 구할 수 있다.

순위스튜던트화잔차(ranked studentized residual)의 기대값에 대한 이상적인 그림(plot)은 기울기가 1.0이고 원점을 지나는 직선의 형태로 나타난다. 이론적으로 이러한 이상적인 그림(picture)에서의 편차(deviation)는 정규성(normality)으로부터의 이탈(departures)을 반영한다. 이 그림(plot)을 이용함에 있어 적절한 판단을 내리는 데에는 경험이 필요하다. 확실히 직선으로부터 임의(random) 편차(deviations)가 존재할 것이다. 진단적인 그림(plot)의 모든 경우에 있어서 표본크기(sample size)가 하나의 요인이 되며, 표본크기가 작은 경우에는 알 수 있는 것이 거의 없다.

정규그림(normal plot)에 근거하여 정규성으로부터 명백한 이탈(apparent departure)의 결과(outcome)를 정확히 지적하는 것은 매우 어렵다. 실제 자료 세트(data set)에서 등분산가정의 잘못(failure of the homogeneous variance assumption)이 실제로 정규성(normality)으로부터 편차(deviation)의 양상(appearance)을 나타낼지도 모른다. 잔차가 매우 큰 바깥점(outlier)의 경우 직선으로부터 유사한 이탈(similar departure)을 보일 수도 있다. 바깥점(outlier)에 관한 전반적인 내용은 5장에서 다루게 될 것이다. 그렇지만 직선으로부터 편차(deviation)가 있는 특정한 외양(specific appearance)에 대한 접근방식을 알고 있는 것은 유용하다. 이는 fig 2.14에 나타나 있는데, 그림(plot)에서 순위스튜던트화잔차가 Y축에 나타나 있다. Heavy-tailed error distribution은 정규분포(normal distribution) 보다 극단적인 관찰이 더 높은 빈도로 나타나고 있음을 반영한다. Light-tailed appearance의 경우에는 기대하는 것 보다 극단적인 관찰값이 거의 없음을 나타내고 있다. 이분산(heterogeneous variance)은 종종 light-tailed error distribution을 야기한다. 모형의 오설정 함수형태(model function form misspecification)는 명백히 치우친 오차분포(skewed error distribution)를 종종 나타낼 것이다.

Figure 2.14 순위잔차(ranked residual)의 전형적인 잔차 그림



예제 2.12 부부 키자료

Table 2.1의 자료를 고려해보자. 잔차에 대한 정규확률그림이 fig 2.15와 fig 2.16에 나타나 있다. fig 2.15에서는 방법 (i)을 따랐다. (즉 제곱근오차평균제곱(root error mean square, s)에 의해 표준화된 순위잔차의 그림) 여기에서 X 축(abscissa)은 $z\left(\frac{i - 0.375}{n + 0.25}\right)$ 이다.

Fig 2.16은 스튜던트화잔차를 가지고 방법 (ii)를 이용하였다. 두 직선 모두 이상적인 형태를 보인다고 판단할 수 있다.

예제에 사용된 R-code는 다음과 같다.

```

data<-read.table("d:/data/ex2_1.R",header=TRUE)
attach(data)
g<-lm(W~H)
s_stat<-summary(g)
res<-g$residuals
stand.res<-res/s_stat$sigma
ginf<-influence(g)
stud.res<-res/(s_stat$sigma*sqrt(1-ginf$hat))
qqnorm(stand.res,xlab="Z",ylab="표준화 잔차",main="그림 2.15 정규확률도",pch=20)

```

```

abline(0,1)

qqnorm(stud.res,xlab="Z",ylab="스튜던트화 잔차",main="그림 2.16 정규확률도",pch=20)
abline(0,1)

```

추가적인 언급(Additional Comments)

잔차그림(residual plot)과 잔차분석(residual analysis)에 대해 더 알고 싶으면 5장을 읽어보면 될 것이다. 다중회귀(multiple regression)의 내용에서 잔차분석에 대한 보다 많은 필요들이 전개될 것이다. 다중회귀변수(multiple regressor variable)의 경우에 가정의 위반(violation of assumption)을 밝히는 과정이 종종 더 어렵다. 비정규성(nonnormality)의 탐지(detection)는 종종 해석하기 어려운 조건이라는 점을 분명히 이해하여야 한다. 이 장 앞부분에서 언급했던 요점을 바꾸어 말하면 정규확률그림(normal probability plot)를 사용한다는 것은 $y_i - \bar{y}_i$ 가 ε_i 에 실제로 필적함을 의미하는 것이고, 여기서 ε_i 는 평균이 0이고 공통분산(common variance)을 가지며 독립이다. 많은 경우에 비정규성오차(nonnormal error)를 나타내는 조건(condition)은 단지 또 다른 가정의 위반을 드러낼 뿐이다. 5, 6장에서 잔차와 잔차그림(residual plot)을 더 잘 이해할 수 있는 보다 기술적인 과정에 대해 다루게 될 것이다. 잔차(residuals)는 진단(diagnostics)의 다른 유형을 구성하기 위해 광범위하게 이용된다.

고립된 바깥점 또는 잘못된 자료 포인트에서의 보조(Aid in Isolating Outliers or Erroneous Data Points)

확실히 잔차가 이상하게 높은 자료가 존재할 때 무엇인가가 잘못되었다거나 그 자료 포인트가 바깥에 위치한 관측치(바깥점, outlier)라는 의심이 들게 된다. 명확한 질문은 너무 크다는 것이 얼마나 크냐는 것이다. (“How large is too large?”) 이 문제는 중요하고 다소 복잡하다. 이 시점에서는 분석가가 어떤 임의의 기준(any arbitrary criterion)을 제공해서는 안된다고 단언하는 것으로 충분하다. 자세한 것은 5장에서 언급될 것이다.

높은 영향점의 검출에 있어서의 보조(Aid in Detection of High Influence Points)

6장에서 고영향 자료점(high influence data point)(즉, 적합 회귀에 많은 영향을 갖고 있는 값)의 의미에 대해 집중적으로 다룬 것이다. 분석가는 이 특수한 지점(point)이 어디에 있고 어떤 중요한 영향을 미치는지를 알 필요가 있다. 이러한 영향을 정량화하는 정보는 잔차로부터 얻을 수 있다.

그림 2.15 정규화률도

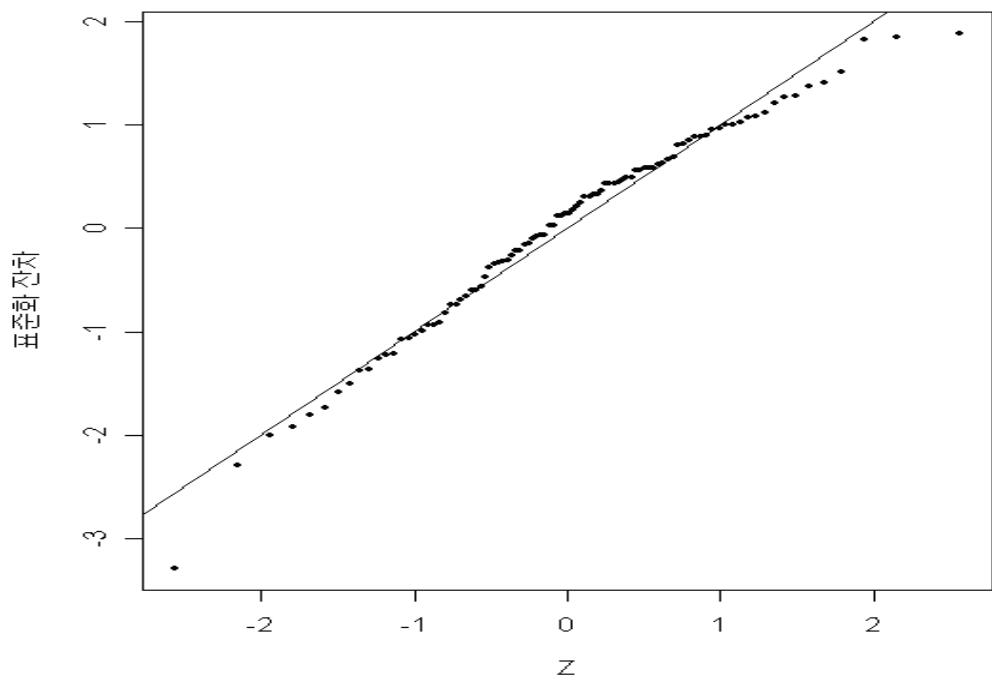
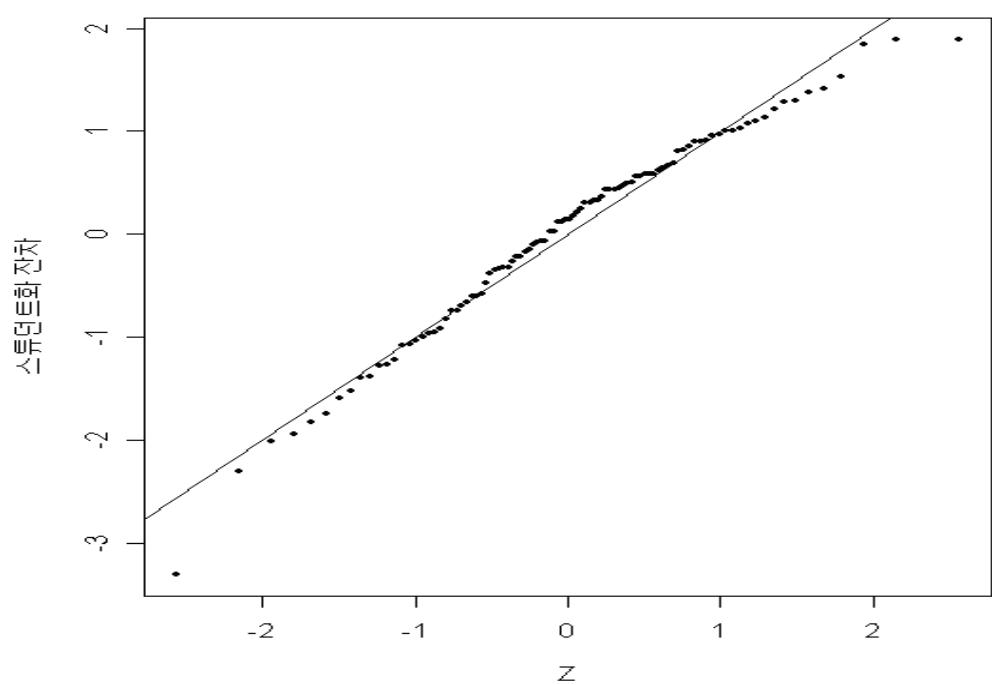


그림 2.16 정규화률도



2.12. 확률변수 x 와 y (Both x and y Random)

앞에서 전개된 내용의 대부분에서 기초적인 가정 중의 하나는 회귀변수(regressor variable)인 x 의 성질에 대해서 초점을 맞추었다. 우리는 그동안 단순선형회귀모형(simple linear regression model)에서 x_i 를 비임의(nonrandom)로 가정했다. 계수의 표준오차(standard errors of coefficients), 가설검정(test of hypothesis), 신뢰구간(confidence interval), 예측구간(prediction interval) 이 모든 것은 위의 가정(assumption)을 따른다. 그러나 실제적으로 두 가지 다른 상황이 매우 빈번히 발생한다.

1. 변수 x 와 y 는 모두 임의적(random)이며 결합밀도함수(joint density function)에서의 관측치(observation)이다.
2. 변수 x 는 무시할 수 없는 오차(nonnegligible error)를 가지고 측정된다.

몇 가지 방법(treatment)이 7장의 두가지 예제에서 주어질 것이다. 첫번째 예제에서는, x 가 임의(random)이지 않은 것으로 추정되는 경우(case)에서보다는 예측과 관련된 관심이 덜한 경우이다. x 와 y 가 결합하여 분포될(jointly distributed) 때, 자료로부터 알아야(learn) 할 본질적 요건(natural requirement)은 연관성의 구조(structure of the relationship)와 관련의 정도(degree of association)이다. 이것이 상관분석(correlation analysis)의 논제(topic)로 이끌어 줄 것이다.

상관계수, 상관분석(The Correlation Coefficient-Correlation Analysis)

두 임의변수(random variable)들간의 의존성의 구조(structure of dependency)가 중요하다는 실제 상황의 한 예로 성인 신체에서 특정한 뼈의 길이(length)(x)와 둘레(circumference)(y)간의 선형관계(linear association)의 강도(strength)를 결정하는 인류학적인 연구(anthropological study)를 고려해보자. n 개의 실험개체(성인들)가 임의적으로 선택되었고, 측정치(measurements) $(x_i, y_i) \quad i = 1, 2, \dots, n$ 는 결합밀도함수(joint density function) $f(x, y)$ 를 가지는 결합분포된(jointly distributed) 확률변수(random variables)의 실현(realization)이라고 할 수 있다.

조건부분포(conditional distribution) $f(y | x)$ 는 정규성(normal)을 가정하면 편리하며, 평균은

$$E(y | x) = \beta_0 + \beta_1 x \quad (2.35)$$

이고, 분산은

$$\sigma_{y|x}^2 = \sigma^2$$

이다. 여기에서 식(2.2)에 의해 정규이론가정(normal theory assumption)으로 가정된 것과

현재의 모형(present model)간에는 상당한 유사점(resemblance)이 보인다. 조건부 개념(*conditional sense*)상으로 그들이 비슷하더라도, 자료를 생산하는 가정된 과학 구조(*assumed scientific structure*)는 다르다. 식(2.2)을 사용할 때, 두 과학적 변수인 x 와 y 는 x 에 선형이라는 방식(*manner*)으로 수학적으로(*mathematically*) 관련(*relate*)되어 있다. ε_i 에 의한 임의성(*randomness*)은 근본적으로 정확하게 모형화할 수 없음을 나타내는 것이고 이를 모형오차(*model error*)라는 용어를 사용하여 표현한다. 현재의 경우(case)에서 식(2.35)의 선형관계는 단지 조건부 분포(*conditional distribution*) $f(y/x)$ 의 모수(parameter)로 보인다.

식(2.35)에서 조건부 기대(*conditional expectation*)에 대한 선형모형(*lineal model*)을 만들어내는 조건부 분포 $f(y/x)$ 는 이변량 정규분포(bivariate normal distribution) $f(y,x)$ 의 가정(*assumption*)의 산물(product)이다.

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

여기에서 $-\infty < x < \infty$, $-\infty < y < \infty$ 이다. $\mu_x, \mu_y, \sigma_x, \sigma_y$ 는 x 와 y 의 평균(mean)과 표준편차(standard deviation)이다. 모수(parameter) ρ 는 x 와 y 의 표준편차의 곱(product)에 대한 x 와 y 의 공분산(covariance)의 비(ratio)로 정의되는 상관계수(*correlation coefficient*)이다.

$$\rho = \frac{E(y-\mu_y)(x-\mu_x)}{\sigma_x\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \quad (2.36)$$

상관계수 ρ 의 해석(interpretation)은 다른 모수들과의 관계를 연구하면 쉽게 알 수 있다. 식(2.35)의 선형구조(*linear structure*)를 회귀선(*regression line*)이라고 한다. 조건부 분포(*conditional distribution*) $f(y/x)$ 를 알아내는 간단한(straightforward) 방법들을 사용함으로써(Graybill (1976)참고), 다음의 정규밀도(normal density)를 얻게 된다.

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\beta_0-\beta_1x}{\sigma}\right)^2\right]$$

여기에서,

$$\beta_1 = \frac{\sigma_y}{\sigma_x} \rho \quad (2.37)$$

$$\beta_0 = \mu_y - \mu_x \rho \frac{\sigma_y}{\sigma_x} \quad (2.38)$$

$$\sigma^2 = \sigma_y^2 (1 - \rho^2) \quad (2.39)$$

식(2.37), (2.38), (2.39)은 ρ 의 역할(role)을 보여주고, 여기에서 언급된 모형과 고정된 x (fixed x case)가 있는(2.2) 모형을 비교하는 데에 도움이 된다. 식(2.37)과 (2.39)로부터,

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_y^2} \quad (2.40)$$

$$\rho^2 = \beta_1^2 \frac{\sigma_x^2}{\sigma_y^2} \quad (2.41)$$

따라서, $\beta_1=0$ 일 때 $\rho=0$ 이다; 즉, 이는 선형회귀(linear regression)가 아니다. σ^2 이 조건부 분포의 분산(variance in a conditional distribution)이므로, $\sigma^2 \leq \sigma_y^2$ 이고, 따라서 $\rho^2 \leq 1$ 이다. 결과로서 다음과 같다.

$$-1 \leq \rho \leq 1$$

식(2.37)로부터, ρ 가 가지는 부호(sign)는 회귀선의 기울기인 β_1 의 부호이다. 만약 $\sigma^2 = 0$ 이라면 $\rho = \pm 1$ 일 수 있는데, 물론 이것은 y 와 x 간에 완전한(perfect) 선형관계(linear relationship)를 암시한다. 따라서 ρ 는 변수들(variables)간의 선형관계의 정도(degree of linear association)를 나타낸다.

모수추정과 가설검정 (Parameter Estimation and Hypothesis Testing)

상관분석(correlation analysis)이 ρ 의 추정(estimation)을 포함해야 한다는 것은 분명한 것이다. 상관력(strength of association)의 해석(interpretation)은 추정값이 1.0(혹은 -1.0)에 얼마나 가까운가에 달려있다. ρ 에 대한 가설의 검정은 또한 분석자 도구(analyst's arsenal of tool)의 한 부분이다.

ρ , β_0 , β_1 과 σ^2 의 추정량(estimator)은 현재의 사례(present case)와 이 장의 2.12절 이전에 소개되었던 고정된 x 경우(fixed x situation)를 더욱 비교(parallel)되게 한다. 우리는 먼저 ρ 의 추정(estimation)을 고려해보자. 일반적인 추정량(estimator)은 다음 식으로 주어지는 표본상관계수(sample correlation coefficient), r 이다.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (2.42)$$

β_l 에 대한 추정량(estimate)으로 최소제곱추정량(least squares estimator)을 사용하면, 아래와 같은 사실을 직관적으로 얻을 수 있다.

$$b_l = \frac{S_{xy}}{S_{xx}} \quad (2.43)$$

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{b_l S_{xy}}{S_{yy}} = \frac{SS_{\text{Re } g}}{S_{yy}} = 1 - \frac{SS_{\text{Re } s}}{S_{yy}} \quad (2.44)$$

식(2.44)의 결과는 2.8절에서 논의되었던 R^2 또는 결정계수(coefficient of determination)이다. 따라서 표본제곱상관계수(squared sample correlation coefficient)는 설명되는(explained) 변동(variation)의 비율(proportion)과 동등(equivalent)하다. 식(2.42)와 (2.43)은 회귀의 기울기인 두 가지 중요한 모수(parameter)인 ρ , β_l 의 추정량(estimate)을 나타내고 있다. 절편 β_0 는 다음과 같이 추정된다.

$$b_0 = \bar{y} - b_l \bar{x} \quad (2.45)$$

임의의 x 인 경우(random x case)와 고정된 x 의 경우(fixed x case)에서의 기울기와 절편의 추정량(estimate)이 같다는 것에 놀랄 이유가 없다. ρ , β_0 , β_l 의 추정값(estimate)뿐만 아니라, x 와 y 변수들간의 의미 있는 선형관계가 있는지 없는지에 대한 증거(evidence)를 평가(assess)하는 가설검정기전(hypothesis testing mechanism)에 자료분석가는 예민한 관심을 가질 수 있다. 따라서 아래의 가설이 관심의 대상이다.

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned} \quad (2.46)$$

여기에서 $\beta_l = 0$ 일 때 $\rho = 0$ 이므로, (2.46)의 H_0 에 대한 검정이 2.6절에서 논의되었던 β_l 에 대한 검정(test)과 유사할 것으로 예상할 수 있다. 다시 상기시키면, $H_0 : \rho = 0$ (Graybill (1976) 참조)라는 전제하에서

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

따라서

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2.47)$$

이것이 (2.46)의 H_0 를 검정하는데 적합한 통계량(statistic)이다. 좀더 자세히 살펴보면,

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\frac{S_{xy}}{\sqrt{S_{yy}S_{xx}}}\sqrt{n-2}}{\sqrt{1-\frac{S_{xy}^2}{S_{yy}S_{xx}}}} = \frac{b_1\sqrt{S_{xx}}}{\sqrt{\frac{S_{yy}-S_{xy}^2/S_{xx}}{n-2}}}$$

이것은 고정된 x 의 경우(fixed x case)에서 $H_0: \beta_l=0$ 를 검정하는 (2.14)의 t -통계량이다. (2.46)의 H_0 에 대한 양쪽꼬리검정(two-tailed t-test)은 (2.47)의 검정통계량(test statistic)을 이용하여 만들어진다. t -분포(t -distribution)의 자유도는 $n-2$ 이다. 만약 가설이 단측(one side)이면 상응하는 한쪽꼬리검정(one-tailed test)이 수행되어야 한다.

다음과 같은 좀더 일반적인 가설(more general hypothesis)을 검정하는 경우는

$$H_0: \rho = \rho_0$$

분석가는 통계량(statistic)의 근사 정규성(approximate normality)을 이용할 수 있다.

$$\frac{1}{2} \ln \frac{1+r}{1-r} \quad (2.48)$$

표본의 수가 많을 경우 ($n \geq 25$), (2.48)의 통계량은 대략 평균이 $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ 이고 분산이 $1/(n-3)$ 인 정규분포를 따른다. 따라서 검정(test)은 (2.48)의 통계량(아래식의 통계량)을 표준화하여 다음의 통계량을 만드는 것을 포함하고

$$z = \frac{\sqrt{n-3}}{2} \left[\ln \left(\frac{1+r}{1-r} \right) - \ln \left(\frac{1+\rho_0}{1-\rho_0} \right) \right] \quad (2.49)$$

표준정규분포(standard normal distribution)의 기각치(critical points)와 비교하는 것도 포함한다.

예제 2.13 체중과 최고 혈압 자료(Weight and systolic blood pressure data)

Table 2.8은 체중과 최고 혈압과의 관계를 연구하기 위해서 25~30살의 남성 26명을 임의로 선택하여 조사한 것이다. 체중과 최고 혈압 모두 확률변수로 결합확률분포는 정규분포를 따른다고 가정하자. 먼저, 두 변수간의 상관계수를 추정해보자.

자료로부터, 다음의 통계량을 계산할 수 있다.

$$S_{xx} = 880545$$

$$S_{xy} = 697076$$

$$S_{yy} = 555802$$

표본 상관 계수는 식(2.42)를 이용하여 계산되며, 결과는 다음과 같다.

$$r = 0.9964243$$

r 의 값은 체중과 최고 혈압간에 강한 선형 관계가 있다는 것을 보여준다. 이제, 식(2.47)에 주어진 t -통계량을 이용하여 가설검정을 해보자. 즉,

$$H_0 : \rho = 0$$

$$H_1 : \rho > 0$$

다음의 등식을 쓸 수 있다.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9964243\sqrt{24}}{\sqrt{1-0.9964243^2}} = 57.77521$$

t -통계량은 0.0001수준 이하에서도 매우 의미 있으며, 두 변수들간의 선형관계에 대한 강한 통계학적 증거를 제시한다.

Table 2.8 체중과 최고 혈압 자료

Subject	Weight(X)	Systolic BP(Y)	Subject	Weight(X)	Systolic BP(Y)
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159

다음은 예제에 사용된 R-code이다.

```
#example 2.13
data<-read.table("d:/data/ex2_13.R",header=TRUE)
sxx<-sum(data[,2]^2)
sxy<-sum(data[,2]*data[,3])
syy<-sum(data[,3]^2)
r<-sxy^2/(sxx*syy)
t_val<-(sqrt(r*(nrow(data)-2)))/(sqrt(1-r))
```

3. 다중선형회귀모형(The Multiple Linear Regression Model)

3.1. 모형 기술 및 가정(Model Description and Assumptions)

모형 구축(model building)을 하다 보면 하나 이상의 회귀변수(regressor variable)가 필요한 경우가 자주 있다. 이 경우 식 (2.1)과 결과 모형 (2.2)은 다중선형회귀(multiple linear regression) 상황(situation)까지 확장되어야 한다. 자료가 아래와 같이 발생하는 실험(experiment)을 생각해보자.

$$\begin{array}{ccccccc} y & x_1 & x_2 & \dots & \dots & x_k \\ y_1 & x_{11} & x_{21} & \dots & \dots & x_{k1} \\ y_2 & x_{12} & x_{22} & \dots & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ y_n & x_{1n} & x_{2n} & \dots & \dots & x_{kn} \end{array}$$

위 배열(array)의 각 행(row)은 자료(data)를 나타낸다. 자료에 의하여 정의되는 x 들의 영역에서 y_i 가 회귀변수(regressor variable)와 거의 선형적으로 관계가 있다면, 모형 공식(model formulation)은 다음과 같이 되는 것이 합리적일 것이다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1, 2, \dots, n; n \geq k + 1) \quad (3.1)$$

단순회귀(simple regression)와 마찬가지로, ε_i 는 모형의 오차(error)이며 평균이 0이고, 분산이 σ^2 이면서 관측값(observation)마다 무상관(uncorrelated)으로 가정한다. 또한 x_{ji} 는 비임의(nonrandom)이며 무시할 수 있을 정도로 오차가 작은(negligible error) 관측값으로 가정한다.

식 (3.1)의 다중선형회귀모형(multiple linear regression model)은 2장에서 논의되었던 단순선형회귀모형(simple linear regression model)과 상당히 다른 점이 있다. 우리는 더 이상 간단한 그림 혹은 그래프로 자료를 쉽게 표시할 수가 없다. 식 (3.1)의 β_j 의 의미는 좀 더 논의를 하여야 하며, 최소제곱 과정(least squares procedure)이 적용되는 것을 그림으로 표현하는 것은 더욱 복잡하고 더 많은 수식 전개를 요구한다. 흔히 통계 모형(statistical modeling)에서 선형(linear)이라고 하는 것이 무엇인지에 관하여 혼란이 있다. 다음에서 이 문제를 다루기로 한다.

선형 모형이란 무엇인가? (What is a Linear Model?)

독자들은 선형모형과 비선형모형의 차이를 이해할 필요가 있다.

선형모형이란 모수(parameter), 즉 회귀계수들(식 (3.1)의 β 들)이 선형이라는 것을 의미한다.

1장부터 8장까지 선형모형을 다루고 있으며, 9장에서는 비선형모형과 관련된 자료 및 예를 살펴볼 것이다. 이 두 형태의 모형의 예를 살펴봄으로써 독자들은 이 모형들을 더 잘 이해할 수 있게 될 것이다. y 와 x 간의 관련성에 대한 모형을 만들기 원하지만, 그 관련성이 가령 다항식(polynomial)을 따르는 경우를 생각해보자. 이것은 선형모형의 한 예이다. 예를 들어, x 의 이차방정식인 한 모형 (그러나 물론, β 들에 대해서는 선형인)이 다음과 같이 주어져 있다.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

선형 모형의 또 다른 예는 한 쌍의 회귀변수 사이의 상호작용을 포함하는 경우이다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

어떤 경우에는 회귀변수(regressor variable)를 변형시킬 필요가 있다. 예로써 x_1, x_2, x_3 3개의 회귀변수가 있는 경우를 생각해보자. 다음은 각각의 변수에 자연 로그 변환(natural log transformation)을 한 선형 모형이다.

$$y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 + \varepsilon$$

여기에 제시된 세가지 예 각각에서 회귀변수들이 변환되었지만, 모형은 모수(parameter)의 입장에서 보면 여전히 선형이다.

반응변수(response variable)인 y 도 종종 변환(transformation)된다. 예를 들어, y 에 대하여 로그 변환(log transformation)을 하고, x_1 과 x_2 에는 역수변환(reciprocal transformation)을 하게 되면, 그 결과는 선형 모형이며 다음과 같이 쓰어진다.

$$\ln y = \beta_0 + \beta_1 \left(\frac{1}{x_1}\right) + \beta_2 \left(\frac{1}{x_2}\right) + \varepsilon$$

정의한 바에 따르면, 여기에서 기술된 모형은 모두 선형 모형이다. 자료분석가들은 변환된

자료에 맞도록 선형 모형을 적합시킨다는(fitting) 점을 알아야 한다. 여기에서 변환(transformation)에 따르는 유용성이나 위험성에 대해 논의하고 싶지는 않다. 그 보다는 이 장에서 우리가 다룰 다중선형회귀 과정이 변수가 본래의 형태가 아닐 때에도 적용된다는 점을 명백히 하고자 한다. 변환(transformation)에 관한 심도 있는 논의는 7장에 나와 있다.

이 시점에서 독자들은 어떤 종류의 예가 비선형 모형(즉, 모수들이 일차함수가 아닌 모형)의 범주에 속하는지에 대하여 의문을 가질지 모르겠다. 반응 y 와 두개의 회귀변수 x_1 과 x_2 가 있다고 할 때, 다음은 비선형 모형의 두 예를 보여준다.

$$y = \beta_0 + \beta_1 x_1^{\gamma_1} + \beta_2 x_2^{\gamma_2} + \varepsilon$$

$$y = \frac{\beta_0}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2)}} + \varepsilon$$

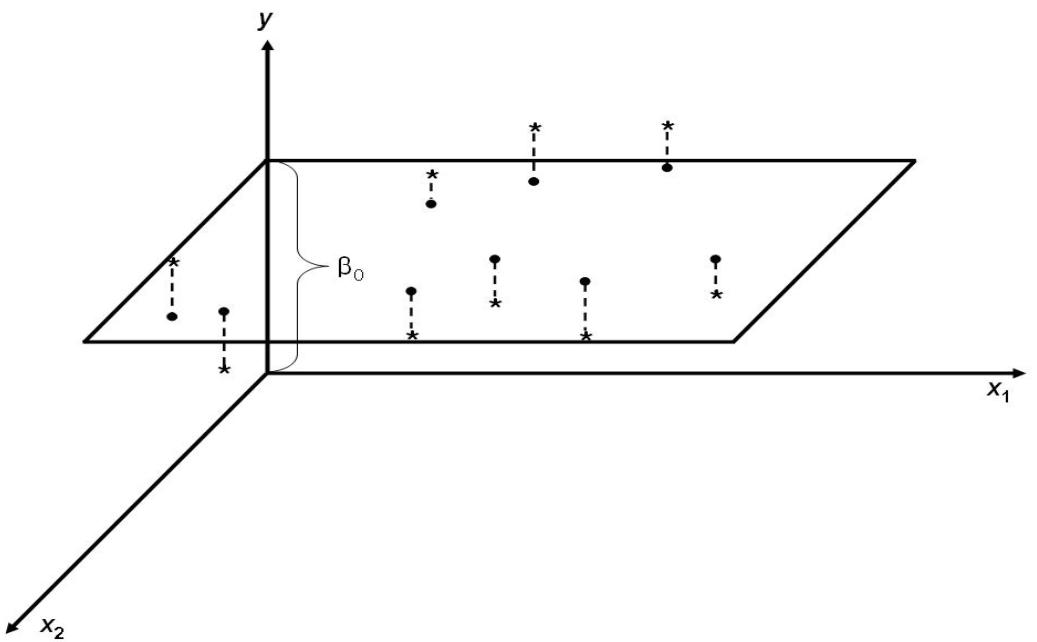
첫번째 모형은 5개의 모수(parameter)를 포함하고, 두번째 모형은 3개의 모수(parameter)를 포함하고 있다. 이 두 경우에서 모수(parameter)는 비선형(nonlinear) 방식으로 모형에 들어가 있다.

모형과 모형 모수의 해석(Interpretation of Model and Model Parameters)

회귀분석의 중요한 목적은 모형 식 (3.1)에 있는 β 들을 추정하는 것이다. β 들은 종종 편회귀계수(partial regression coefficient)라고 불린다. 가령 $\beta_1, \beta_2, \beta_3$ 를 모수(parameter)로 가지는 다중회귀(multiple regression)에서, 모수(parameter) β_1 은 나머지 다른 x 들이 일정하게 고정되어 있을 때 x_1 의 단위변화(unit change) 당 예상되는 반응의 변화로 해석될 수 있다. 나머지 β 들도 이와 유사하게 해석될 수 있다. 식 (2.1)의 단순선형회귀모형의 직선을 해석하는 것은 매우 간단하다. 다중선형회귀의 경우에, (3.1)의 해석은 이의 연장선 상에 있다.

자료와 참회귀(true regression) (회귀평면으로 표시됨)를 나타내는 fig 3.1을 살펴보자. ε_i 들은 회귀평면(regression plane)에 이르는 수직 거리이고, 회귀평면 상의 점들은 기대 반응(expected response)을 나타낸다. 회귀 계수(regression coefficients) β_1 과 β_2 는 각각 x_1 과 x_2 방향의 회귀평면의 기울기이다. 절편 β_0 또한 fig 3.1에 표시되어 있다. 명백히 회귀분석의 목적은 세 모수(parameter)들을 추정함으로써 모회귀평면(population regression plane)을 추정하는 것이다.

Figure 3.1 Multiple Linear Regression in Two Variables.(The notation * represents the data and • represents points on the regression plane.)



3.2. 일반 선형모형과 최소제곱과정(The General Linear Model and the Least Squares Procedure)

우리는 모수(parameter)를 추정하기 위해서 최소제곱법(least squares)을 또 다시 고려할 것이다. 식 (3.1)의 모형을 행렬 표기(matrix notation)로 나타내보자. 모형은 다음과 같이 나타낼 수 있다.

$$y = X\beta + \varepsilon \quad (3.2)$$

여기에서

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

이고, 벡터(vector) ε , 즉 모형오차(model errors) 열(column)은 아래와 같이 주어진다.

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

독자는 일반선형모형(general linear model)인 (3.2) 모형에 친숙해져야 한다. 행렬에 지나치게 몰두하지 않고서도 이 본문에서 논의되는 많은 개념들과 예들을 이해할 수는 있다. 그러나 행렬 조작(matrix manipulation)을 이해하지 못하면 회귀분석에서 수식 전개(development) 과정을 완전히 이해할 수 없다. 따라서 행렬을 적절히 다룰 수 없는 회귀분석 사용자들에게는, 다중공선성(multicollinearity), 편향추정(biased estimation), 로버스트 회귀(robust regression), 잔차 분석(residual analysis)과 같은 중요하고 시기적절한 논제들을 이해하기 어려울 것이다.

3.1절에서 언급된 가정에 근거하여, 오차벡터(error vector) ε 과 반응벡터(response vector) y 는 확률벡터들(random vectors)이다. X 행렬은 차원이 $n \times p$ 이고, 여기에서 $p = k + 1$ 은 모형 모수(parameter)의 총 개수이다. X 의 j 번째 열(column)은 회귀변수 x_j 의 측정값이 포함되는데, 이것의 오차는 매우 작다(negligible error). 이 행렬은 자료행렬(data matrix) 또는 모형행렬(model matrix)이다. 때로 이것은 설계행렬(design matrix)으로 불리기도 한다. 실제로,

X 행렬은 이 세 가지 개념 모두가 복합(combination)된 것이다. X 행렬은 회귀변수자료(regressor data)를 나타내므로, 회귀변수의 수준(level)이 마련되어 있는 경우에 X 행렬은 실험설계(experimental design)를 나타낸다. 그러나, X는 명백히 모형의 한 함수(a function of model)이다. 선형회귀(linear regression)와 결합(conjunction)하여 행렬을 사용하는데 친숙하지 않은 독자에게, X 행렬을 사용한 몇몇 예들이 도움이 될 것이다.

화학반응(chemical reaction)의 산물(yield)이 온도의 함수로 모형화되는 경우를 생각해보자. 자료는 다음과 같다.

y (yield, %)	x (temperature, °F)
77	160
79	165
82	170
83	175
85	180
86	185
87	190

단순선형회귀모형(simple linear regression model)을 가정해보자

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \varepsilon_i \quad (i = 1, 2, \dots, 7)$$

우리가 중심화모형(centered model)으로 나타내고 있음을 주목하라. 식 (3.2)로부터 우리는 아래와 같이 나타낼 수 있다.

$$\begin{bmatrix} 77 \\ 79 \\ 82 \\ 83 \\ 85 \\ 86 \\ 87 \end{bmatrix} = \begin{bmatrix} 1 & -15 \\ 1 & -10 \\ 1 & -5 \\ 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 15 \end{bmatrix} \begin{bmatrix} \beta_0^* \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

실험 계획(experimental plan)이 온도 뿐만 아니라 반응시간(reaction time)도 변화시키는 것이며, 실험 결과가 다음과 같다고 하자.

y (yield, %)	x_1 (temperature, °F)	x_2 (time, hr)
77	160	1
79	160	2
82	165	1
83	165	2
85	170	1
88	170	2
90	175	1
93	175	2

다중선형회귀모형(multiple linear regression model)(두개의 변수를 모두 중심화 모형으로 나타내었음)은 다음과 같다.

$$y_i = \beta_0^* + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \varepsilon_i$$

그리고, 식 (3.2)는 행렬형태(matrix form)로 다음과 같다.

$$\begin{bmatrix} 77 \\ 79 \\ 82 \\ 83 \\ 85 \\ 88 \\ 90 \\ 93 \end{bmatrix} = \begin{bmatrix} 1 & -7.5 & -0.5 \\ 1 & -7.5 & 0.5 \\ 1 & -2.5 & -0.5 \\ 1 & -2.5 & 0.5 \\ 1 & 2.5 & -0.5 \\ 1 & 2.5 & 0.5 \\ 1 & 7.5 & -0.5 \\ 1 & 7.5 & 0.5 \end{bmatrix} \begin{bmatrix} \beta_0^* \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix} \quad (3.3)$$

우리는 앞으로 나올 절에서 같은 예를 사용할 것이다. 앞으로 다중선형회귀(multiple linear regression)에서 모수추정(parameter estimation)을 위해서 최소제곱 과정(least square procedure)이 전개되는 것을 살펴볼 것이다.

최소제곱의 수식전개(Least Squares Development)

β 에 있어서 회귀계수(regression coefficient)에 대한 최소제곱추정량(least squares estimator) b 는 다음을 만족시키는 벡터이다.

$$\frac{\partial}{\partial b} [(y - Xb)'(y - Xb)] = 0$$

여기에서, $(y - Xb)'(y - Xb)$ 는 잔차제곱합(residual sum of squares)을 나타낸다. 미분(differentiation)을 시행하면 다음의 결과를 얻는다.

$$-2X'y + 2(X'X)b = 0$$

그리고, 이 일반적인 경우에 해당하는 최소제곱 정규방정식(least squares normal equation)은 아래와 같다.

$$(X'X)b = X'y$$

그 결과, X 가 완전열순위(full column rank)라고 가정한다면,

$$b = (X'X)^{-1}X'y \quad (3.4)$$

(3.4)의 $X'X$ 행렬은 $(k+1) \times (k+1)$ 인 대칭행렬(symmetric matrix)이고, 대각원소(diagonal elements)는 X 행렬의 열(column)에 있는 원소(element)들의 제곱합(sum of squares)이고, 대각 이외 부분의 원소들은 같은 열에 있는 원소들의 교차곱의 합(sums of cross product)이다.

$X'X$ 의 성질은 b 의 추정량(estimator)의 특성(property)에 중요한 역할을 하고, 종종 추정과정(estimation procedure)에서 보통최소제곱(ordinary least squares)의 성패를 좌우한다. 이것은 3.3절, 3.8절(다중공선성(multicollinearity)에 대한 토의)과 8장에서 더 자세히 다루어질 것이다.

식 (3.3)에 있는 생산량(yield), 온도(temperature), 시간(time) 자료를 생각해보라. β_0^* , β_1 과 β_2 의 최소제곱 정규방정식(least square normal equation)은

$$(X'X)b = X'y$$

$$\begin{bmatrix} 8 & 0 & 0 \\ 0 & 250 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} b_0^* \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 677 \\ 222.5 \\ 4.5 \end{bmatrix}$$

이고, 해답은 $b_0^* = 84.625$, $b_1 = 0.890$ 이며, $b_2 = 2.250$ 이다.

σ^2 의 추정(Estimation of σ^2)

다중회귀(multiple regression)에서 σ^2 의 좋은 추정값(good estimate)을 얻을 필요가 있다.

우리는 가설검정(hypothesis testing)을 통해서 변수 선별(variable screening)을 하거나 모형의 질(quality)을 평가할 때 그 추정값(estimate)을 사용한다.

이 추정량(estimator)을 논의하기에 앞서, 우리는 다중선형회귀(multiple linear regression)에서의 총 자유도(total degrees of freedom)의 분할(partition)이라는 측면에서 식 (2.11)을 다시 살펴보아야 한다. 다중 회귀의 경우에, 식 (2.11)은 여전히 성립한다. 그러나, 회귀제곱합(SS_{Reg})은 k 모형 항(k model terms)에 해당하는 변동(variation)을 설명한다. 따라서 총 자유도는 다음과 같이 분할된다.

$$n - 1 = k + (n - k - 1)$$

불편추정량(unbiased estimator)인 s^2 은 또 다시 변동(variation)을 나타낸다. 잔차(residual) 즉, 회귀 $\hat{y} = Xb$ 에 대한 변동(variation)에 있어서, 분모(denominator)는 이제 $n-p$ 가 되며 여기에서 p 는 추정된 모수(estimated parameter)의 개수이다. (3.2)에 있는 모형의 표기법(notation)에서 $p = k+1$ 이다. 따라서,

$$s^2 = \frac{(y - Xb)'(y - Xb)}{n - p} = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n - p} \quad (3.5)$$

selection)의 최종 판단 과정(definitive decision-making procedure)으로 받아들여서는 안된다.

최소제곱의 기하(Geometry of Least Squares)

벡터 기하(vector geometry)에 대한 사전 지식이 있는 사람들에게는 최소제곱과정(least squares procedure)의 기하(geometry)를 이해하는 것이 매우 단순할 것이다. 추가로, 다중선형회귀(multiple linear regression)에 해당하는 (2.11)의 제곱합 항등식(sum of squares identity)이 제시된다. $n=3$ 관측값($n=3$ observations)이고, $p=2$ 모수($p=2$ parameters)인 회귀 상황(regression situation)을 나타내고 있는 fig 3.2를 살펴보라. 3차원의 축을 가진 시스템이 y -관측공간(y -observation space)에 나타나 있다. y 벡터는 관측공간에 있는 관측벡터(observation vector)를 나타낸다. 그림의 2차원 평면은 추정공간(estimation space)이다. 추정공간(estimation space)이란 $X\hat{\beta}$ 형태(form)의 점(point)들을 포함하는 공간을 의미한다, 물론 여기에서 $\hat{\beta}$ 는 β 의 추정값(estimate)이다. y^* 점은 추정공간에서 임의의 후보 점(arbitrary candidate point)으로 간주되어야 한다. 이제, 추정공간(estimation space)의 어떤 점이 잔차제곱합을 최소로 만드는 \hat{y} 를 제공(produce)하는가? y^* 에서 y 에 이르는 거리의 제곱(squared distance)은 실제로 $(y - y^*)(y - y^*)'$ 이다. 또는 $y^* = X\hat{\beta}$ 라고 한다면, 거리의 제곱은 $(y - X\hat{\beta})(y - X\hat{\beta})'$ 가 된다. 그래서, 우리가 추정공간에 있는 점들 중에서 이 거리의 제곱을 최소로 만드는 점을 찾아내는 데에 최소제곱과정(least squares procedure)이 적용된다. 우리가 y 에서 추정공간(estimation space)으로 수직선을 그으면 \hat{y} 점이 될 것이라는 것은 명백하다. 이제 우리는 이 과정에서, 가장 작은 벡터 $y - \hat{y}$ 는 $X'(y - \hat{y}) = 0$ 이어야 한다는 것을 알고 있다. 추정공간(estimation space)에 있는 이 점에 대해서 $\hat{y} = Xb$ 라는 일반적인 표기를 사용하게 되면,

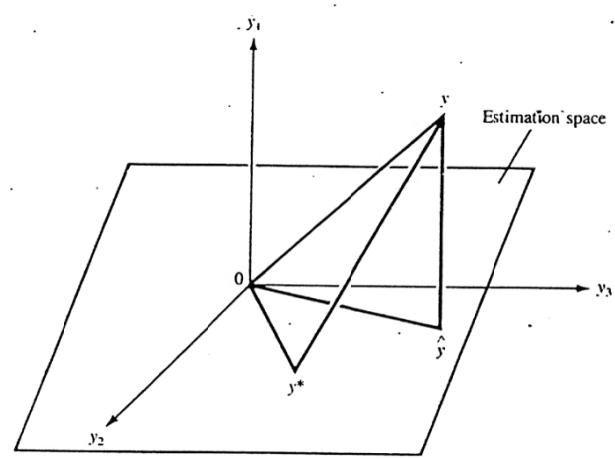
$$X'(y - Xb) = 0$$

이제, 상기 식으로부터 다음과 같이 b 를 얻을 수 있다는 것을 쉽게 알 수 있다.

$$(X'X)b = X'y$$

이것은 최소제곱추정량(least square estimator)을 얻는 해법을 알려주는 식 (3.4)의 정규방정식(normal equation)이다.

Figure 3.2 The geometry of the least squares procedures



3.3. 이상적인 조건하에서 최소제곱추정량의 속성(Properties of Least Squares Estimators under Ideal Conditions)

독자는 (3.1)에 있는 모형의 이상적인 조건을 상기해야 한다.

1. ε_i 의 평균값은 0이다 (모형의 함수 형태(model functional form)가 옳다).
2. ε_i 는 서로 관련되어 있지 않으며(uncorrelated), 공통분산(common variance)이 σ^2 이다 (등분산(homogeneous variance)).

β 들에 대한 가설 검정을 하기 위해서는 ε_i 가 정규분포를 한다는 가정이 선행되어야 한다.
b 벡터의 속성(properties)에 대하여 논의하기에 앞서 ε_i 가 정규분포를 한다는 가정에 대해 다시 살펴볼 것이다. 또, 방법에 있어서 무엇이 바람직한지 (혹은 바람직하지 않은지), 어떤 조건에서 이상적인 경우에 비해 성능(performance)이 떨어지는지에 대해 집중한다면, 왜 때로는 보통최소제곱(ordinary least squares)에서 벗어날(deviate) 필요가 있는지 알게 될 것이다..

모수 추정값들의 편향과 분산속성(Bias and Variance Properties of the Parameter Estimates)

$E(\varepsilon) = 0$ 라는 조건하에서, (3.4)의 b 는 β 에 대한 불편추정량(unbiased estimator)이다. 이것은 쉽게 증명될 수 있다. $E(y) = X\beta$ 이고, X 는 임의(random)가 아니므로

$$\begin{aligned} E(X'X)^{-1}X'y &= (X'X)^{-1}X'E(y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

추정값 벡터(vector of estimates) b 를 참계수 (true coefficient) 즉 β 들로 표현하는 것은 편리하고(convenient) 통찰력(insightful)이 있는 방법이다. 우리는 다음과 같이 나타낼 수 있다.

$$b = \beta + R\varepsilon$$

여기에서 $R = (X'X)^{-1}X$ 이다. R 행렬(matrix)은 6장에 있는 영향진단(influence diagnostics)에서 상대적으로 중요한 역할을 한다. 상기 표현식으로부터 b 의 비편향성(unbiasedness)이 즉각 뒤따른다는 점을 주목하라.

분산(variance), 즉 b 의 산포속성(dispersion properties)을 위해서, 우리는 분산공분산행렬(variance-covariance matrix) $E(b - \beta)(b - \beta)'$ 을 구한다. 3.3절의 시작 부분에 제시하였던 두번째 가정을 이용해야 한다.

b 의 분산공분산행렬(variance-covariance matrix)은 다음과 같이 주어진다.

$$Var(b) = \sigma^2(X'X)^{-1} \quad (3.6)$$

식 (3.6)이 뜻하는 바는, 계수의 분산(variance of coefficients)이 오차분산(error variance) σ^2 과는 별개로 $(X'X)^{-1}$ 의 대각원소(diagonal elements)에 나타난다는 것이다. 이와 유사하게, b 들 간의 공분산(covariance among the b 's)은 같은 행렬의 비대각 원소(off-diagonal elements)로써 나타난다. 조건 1과 2의 사례에서 ε_i 에 대하여 정규성(normality)을 가정하지 않고, 우리는 β 의 최소제곱추정량(least squares estimator)에 대한 매우 중요한 적정 성질(optimality property)을 언급한다. 그 결과가 Gauss-Markoff 이론이다.

(3.2)에 있는 모형의 경우, 만약 $E(\varepsilon) = 0$ 이고 $Var(\varepsilon) = \sigma^2 I$ (분산공분산행렬)라면, 추정량(estimators)은 선형불편추정량 부류(the class of linear unbiased estimator)의 최소분산(minimum variance)에 도달한다.

이 이론에 대한 증거로 Graybill (1976)을 참조하시요. 최소제곱추정량(least squares estimator)은 종종 BLUE (최량 선형 비편향 추정값, Best Linear Unbiased Estimators)로 언급된다. 여기에서 “최량(best)”은 최소분산(minimum of variance)의 의미로 사용된다. 선형모형에서 최소제곱추정량(least squares estimator)의 성질은 정규성 분포가 유지될 때 더 강한(stronger) 것으로 된다. (3.4)에서 b 의 성질(property)은 정규성 가정 하에서 다음과 같이 기술된다.

추정량의 성질에 대한 정상 오차의 효과(Effect of Normal Errors on Properties of Estimators)

식 (3.6)의 비편향(unbiasedness)이나 산포(dispersion)의 성질은 정규성 가정(normality assumption)에 의해 좌우되지 않는다. 그러나 최소제곱추정량과 추정의 다른 형태(alternative form of estimation) 사이에서 선택하고자 하는 실무자료분석가는, 오차(error)가 정규성(normal)을 따를 때의 최소제곱추정량(least squares estimator)의 성질(property)을 이해하여야 한다. 이 성질들은 다음 결과로 설명될 수 있다.

만약 오차(error)가 평균 0과 공통 분산(common variance)으로 정규성(normal)을 가지고 독립이라면, 모형 (3.2)에 있는 β 요소(element)의 최소제곱추정량은 모든 불편추정량의

부류에서 균일한 최소 분산(uniformly minimum variance in the class of all unbiased estimators (UMVU) 이 된다.

이 결과에 대한 증거를 얻고 싶다면 Graybill (1976)을 참조하라. BLUE 성질(property)의 경우에는 최소분산성질(minimum variance property)이 선형불편추정량(linear unbiased estimator)에 국한되지만, UMVU 경우에는 최소분산성질이 불편추정량(unbiased estimator)에만 국한된다. 선형추정량(linear estimator)에 의해 계수(coefficients)는 y 측정값(y observation)의 선형함수(linear function)가 된다. 바꾸어 말하면 선형추정량(linear estimator)은, 이 경우에는, R 이 $(k + 1) \times n$ 행렬인 Ry 의 한 형태이다. 명백하게 식 (3.4)의 최소제곱추정량은 다음과 같은 선형추정량(linear estimator)이다.

$$R = (X'X)^{-1}X'$$

정규성(normality) 하에서 UMVU 성질(property)을 가지는 것 외에, 최소제곱추정량(least square estimators) 또한 최대우도추정량(maximum likelihood estimators)이다. 우리는 제 2장의 단순선형회귀 사례에서 최대우도추정량(maximum likelihood estimators)을 전개하였다. 다중회귀(multiple regression)에 대한 최대우도 전개(maximum likelihood development)는 부록(Appendix B.3)에 있다.

이런 성질들(properties)을 주의깊게 살펴보면, 오차(errors)가 정규 분포를 따르지 않을 때 최소제곱추정량의 성능(performance) 면에서 개선될 여지가 더 크다는 것을 알 수 있다. 실제로 오차가 Gauss 분포로부터 상당히 편향(deviate)되어 있을 때, 회귀모형계수의 다른 추정량(other estimators of the regression model coefficients)인 로버스트 회귀추정량(robust regression estimator)의 성능이 보통최소제곱추정량(ordinary least square estimator) 보다 더 좋다. 로버스트 회귀(robust regression)는 제 7장에서 언급될 것이다.

예제 3.1 헬스클럽 자료(Health Club Data)

어느 회사의 헬스클럽에 등록한 30명을 대상으로 기록한 건강자료이다. 정의된 다음 변수들을 참조하여 물음에 답하여라.

X_1 : 몸무게(파운드)

X_2 : 분당정지맥박수

X_3 : 근력(들어올릴수있는최대의무게:파운드)

X_4 : 1/4마일시험주행속도(초)

Y : 1마일주행속도(초)

모형은 총 4개의 회귀변수가 있고 따라서 추정해야 할 모수는 5개가 된다. 모형행렬 X 는

다음과 같다.

$$X = \begin{bmatrix} & X_1 & X_2 & X_3 & X_4 \\ 1 & 217 & 67 & 260 & 91 \\ 2 & 141 & 52 & 190 & 66 \\ 3 & 152 & 58 & 203 & 68 \\ 4 & 153 & 56 & 183 & 70 \\ 5 & 180 & 66 & 170 & 77 \\ 6 & 193 & 71 & 178 & 82 \\ 7 & 162 & 65 & 160 & 74 \\ 8 & 212 & 66 & 220 & 77 \\ 9 & 138 & 70 & 180 & 62 \\ 10 & 147 & 54 & 150 & 75 \\ 11 & 197 & 76 & 228 & 88 \\ 12 & 165 & 59 & 188 & 70 \\ \dots & \dots & \dots & \dots & \dots \\ 23 & 257 & 64 & 313 & 96 \\ 24 & 236 & 72 & 225 & 84 \\ 25 & 149 & 57 & 173 & 68 \\ 26 & 161 & 57 & 173 & 65 \\ 27 & 198 & 59 & 220 & 62 \\ 28 & 245 & 70 & 218 & 69 \\ 29 & 141 & 63 & 193 & 60 \\ 30 & 177 & 53 & 183 & 75 \end{bmatrix}$$

X 행렬은 $\mathbf{X}'\mathbf{X}$ 와 $(\mathbf{X}'\mathbf{X})^{-1}$ 을 계산하는 데 사용된다. 회귀계수는 식 (3.4)으로부터 계산이 가능하다. 최소제곱법을 이용하여 구한 회귀식은 다음과 같다.

$$\hat{y} = -3.6186 + 1.2676x_1 - 0.5252x_2 - 0.5050x_3 + 3.9030x_4$$

잔차제곱합(residual sum of squares)은

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = 20551$$

이며 분산의 잔차추정값(residual estimate of variance), s^2 (잔차자유도(residual degrees of freedom)는 25)는 다음과 같다.

$$s^2 = \frac{20551}{25} = 822$$

Table 3.1 헬스클럽 자료(Health Club Data)

x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4	y
217	67	260	91	481	138	70	180	62	267
141	52	190	66	292	147	54	150	75	404
152	58	203	68	338	197	76	228	88	442
153	56	183	70	357	165	59	188	70	368
180	66	170	77	396	125	58	160	66	295
193	71	178	82	429	161	52	190	69	391
162	65	160	74	345	132	62	163	59	264
180	80	170	84	469	257	64	313	96	487
205	77	188	83	425	236	72	225	84	481
168	74	170	79	358	149	57	173	68	374
232	65	220	72	393	161	57	173	65	309
146	68	158	68	346	198	59	220	62	367
173	51	243	56	279	245	70	218	69	469
155	64	198	59	311	141	63	193	60	252
212	66	220	77	401	177	53	183	75	338

예제에 사용된 R-code는 다음과 같다.

```

data<-read.table("d:/data/ex3_1.R",header=TRUE)
attach(data)
g<-lm(y~x1+x2+x3+x4)
s_stat<-summary(g)
ano<-anova(g)

```

3.4. 다중선형회귀에서의 가설검정(Hypothesis Testing in Multiple Linear Regression)

많은 전통적인 과학 분야에서 모형구축연습(model-building exercise)의 주요 기능(primary function)은 회귀변수(regressor variable)가 반응 y 에 참으로 영향을 미치는지를 결정하는 것이다. 종종 완벽하게 연구되지 않은 새로운 체계(new system)에서 어떤 요인(factors)이 정말로 관계가 있는지를 결정하기 위해서 예비조사(preliminary investigation)가 필요하다. 그러한 상황을 다루기 위한 통계 방법들(total statistical arsenal)이 많이 있지만, 회귀분석(regression analysis)이 종종 선택된다. 그래서 어떤 회귀변수가 반응 y 에 유의한 변동(significant variation)을 일으키는지를 결정하는 것이 중요하다. 독자는 개별(individual) β_i 에 대한 가설검정(hypothesis testing)을 위해서 표준 과정(standard procedure)을 어떻게 적용하는지를 철저히 이해하여야 한다. 가설검정을 변수선별(variable screening)의 한 방법으로 사용할 때의 단점(drawback)에 대해서도 알아야 한다.

변이의 분할(Partitioning of Variability)

단순선형회귀(simple linear regression)의 경우에서 처럼, 총변이(total variability)를 잔차제곱합(residual sum of squares)과 회귀제곱합(regression sum of squares)과 관련지어 나타낸 기본 항등식(fundamental identity)은 다음과 같다.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.7)$$

여기에서 \hat{y}_i 는 $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$ 인 적합회귀식(fitted regression)이다. 이 결정계수(coefficient of determination)는 단순선형회귀에서와 동일하게 해석되어 다음과 같다.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{SS_{reg}}{SS_{Total}} \quad (3.8)$$

식 (3.7)에서 회귀제곱합(regression sum of squares)은 전통적 의미(traditional sense)로 단일회귀변수(single regression variable)의 가치(worth)를 평가하는데 도움을 주는 의미있는 부분(meaningful portions)으로 분할할 수 있다. SS_{reg} 는 다음과 같이 순차회귀제곱합(sequential regression sums of squares)으로 분할할 수 있다.

$$R(\beta_1, \beta_2, \dots, \beta_k | \beta_0) = R(\beta_1 | \beta_0) + R(\beta_2 | \beta_1, \beta_0) + R(\beta_3 | \beta_2, \beta_1, \beta_0) + \dots + R(\beta_k | \beta_{k-1}, \beta_{k-2}, \dots, \beta_2, \beta_1, \beta_0) \quad (3.9)$$

$R(\bullet | \bullet)$ 표기(notation)는 “존재하에서(in the presence of.)”를 의미하는 수직선(vertical line)으로 “...에 의하여 설명되는 회귀(regression explained by...)”를 의미한다. 한 예로 $R(\beta_2 | \beta_1, \beta_0)$ 는 단지 x_1 와 상수항(constant term)을 포함하는 모형에 회귀변수(regressor) x_2 가 더해질 때 회귀제곱합(regression sum of squares)의 증가이다. 결국 (3.9)의 구성요소(component)에 있어서 각 요소는 부분집합(a subset)을 포함하는 모형에서 특정 회귀변수(particular regressor variable)에 의해 설명되는 회귀(regression)에서 점진적인 증가(*incremental increase*)를 뜻한다. 식 (3.9)는 SS_{Reg} 를 “단일자유도”的 기여(“single degree of freedom” contributions)로 분할(a meaningful partitioning)하는 것을 뜻하며, 이것을 통해서 독자들은 상수항(constant term)에서 시작하여 순차적으로 각 회귀변수(regressor variable)가 더해지는 모형을 볼 수 있다. 모형에 각 회귀변수(regressor)를 추가함에 따라 회귀제곱합(regression sum of squares)은 증가하게 된다. SS_{Reg} 이 증가함에 따라 증가한 양만큼 잔차제곱합(residual sum of squares)이 감소함을 (3.7)을 통해 알 수 있다. 그러므로 (3.9)의 항은 회귀변수(regressor)의 순차적인 도입(introduction)으로 인해 감소되는 잔차제곱합의 개별감소(individual reduction)로도 볼 수 있다. 그래서 한 예를 들면 $R(\beta_3 | \beta_0, \beta_1, \beta_2)$ 은 β_3x_3 가 $\beta_0, \beta_1x_1, \beta_2x_2$ 를 포함하는 모형에 추가되면서 SS_{Reg} 의 증가 또는 SS_{Res} 의 감소로 나타난다.

식 (3.9)는 개별적인 회귀변수(regressor)의 가치(worth)를 평가하도록 도와준다. 만약 회귀변수(regressor)의 부분집합(subset)의 가치에 관한 정보를 원한다면 다른 종류의 분할이 더 유용하다. 예를 들어 $k = 4$ ($p = 5$)일 경우 회귀 x_3, x_4 의 가치를 생각해보고, 다음과 같이 기술해 보자.

$$R(\beta_1, \beta_2, \beta_3, \beta_4 | \beta_0) = R(\beta_1, \beta_2 | \beta_0) + R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2) \quad (3.10)$$

여기에서 우변에 있는 두 항은 두 자유도 분할(two degree of freedom partitions)을 의미한다. $R(\beta_1, \beta_2 | \beta_0)$ 는 β_1 과 β_2 의 추론(inference)에는 일반적으로는 이용되지 않는데, x_3 과 x_4 에 대해서 조정(adjustment) 되지 않기 때문이다. 반면에 $R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2)$ 는 총체적으로(collectively) x_3 과 x_4 의 중요성을 설명하는데 필수적이다.

단순선형(simple linear)의 경우(case) 회귀제곱합(regression sum of squares), $R(\beta_1, \dots, \beta_k | \beta_0) = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, 은 모형으로 설명되는 반응(response)에서의 총제곱합(total sum of squares)의 양(amount)을 정량한 것이다. 앞에서 설명되었던 경우에, SS_{Reg} 의 고립된 증가(isolating increase)는 어떤 변수(variable)나 부분집합(subset)이 SS_{Reg} 이나 R^2 로 정량화된 적합의 질(quality of fit)을 좌우하는지(responsible)를 결정하는데 이용될 수 있다. 다행히도 독자들은 식 (3.9)와 (3.10)으로부터 어느 정도의 일반성(generality)을 얻을 수 있다. (3.10)의 설명에서

만약 β_3, β_4 의 추론에 관심이 있다면 $R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2)$ 은 다음과 같이 계산될 수 있다.

$$R(\beta_3, \beta_4 | \beta_0, \beta_1, \beta_2) = R(\beta_3 | \beta_0, \beta_1, \beta_2) + R(\beta_4 | \beta_0, \beta_1, \beta_2, \beta_3)$$

다음에서 우리는 모수(parameter)의 부분집합(subset)에 관한 가정의 검정에 대하여 주로 다룰 것이다.

회귀모수의 부분집합 검정(*Tests on Subsets of the Regression Parameters*)

회귀제곱합(regression sum of squares)의 분할(partitions)은 (3.9)와 (3.10)에서 보여준 것 같이 개별 회귀계수(individual regression coefficient)와 회귀계수의 부분집합(subset of regression coefficient)에 대하여 전통적인 가설검정(traditional tests of hypothesis)을 하게 한다. 행렬표기(matrix notation)를 이용하면 이 검정을 꽤 간결하고 일반적으로 전개할 수 있다. β 와 X 가 다음과 같이 나누어졌다고 가정하자.

$$X = [X_1 : X_2] \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

여기에서 X_1 은 $n \times p_1$ ($p_1 + p_2 = p$)이고 β_1 은 X_1 의 열(column)과 관련된 p_1 스칼라(scalar) 요소를 포함한다. 그래서 (3.2)의 일반선형회귀모형(general linear regression model)을 다음과 같이 쓸 수 있다.

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (3.11)$$

β_1 이 미심쩍은 모수(questionable parameters)를 포함했다고 가정하자. 그래서 우리는 다음과 같은 검정을 하고자 한다.

$H_0 : \beta_1 = 0$	(3.12)
$H_1 : \beta_1 \neq 0$	

(3.11)의 모형에서 상수항인 β_0 는 β_1 이나 β_2 로 나타낼 수 있다. 예를 들어 만약 β 가 다음과 같고,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

다음의 가설을 검정(test)하고자 한다면,

$$H_0: \beta_1 = \beta_2 = 0$$

다음과 같이 분할하면 될 것이다.

$$\beta_1 = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \beta_2 = \begin{pmatrix} \beta_0 \\ \beta_3 \end{pmatrix}$$

그리고 X_1 과 X_2 의 열(column)이 그에 맞게 정리(arrange)될 것이다.

(3.12)에 해당하는 모수(parameter)의 부분집합(subset)에 대한 일반가설(general hypothesis)에서 적절한 검정(appropriate test)은 β_1 에 의해 설명되는 회귀제곱합(regression sum of squares), 또는 이와 동등한(equivalent), $X_2\beta_2$ 를 포함하는 모형에 $X_1\beta_1$ 을 첨가함으로써 감소되는 잔차제곱합(residual sum of squares)을 포함해야 한다. 초기의 전개로부터 검정(test)이 다음을 포함해야 함은 명백하다.

$$R(\beta_1 | \beta_2) = R(\beta_1, \beta_2) - R(\beta_2)$$

$R(\beta_1, \beta_2)$ 항은 상수를 포함한 모든 모형항에 의해서 설명되는 회귀이다. $R(\beta_2)$ 으로 표기되는 것은, 단지 $X_2\beta_2$ 를 포함하는 모형에 의해 설명되는 회귀제곱합(regression sum of squares)을 뜻한다. 추가제곱합 회귀(extra sum of squares regression), $R(\beta_1 | \beta_2)$,는 다음과 같이 기술된다(부록 A.2를 보시오).

$$\begin{aligned} R(\beta_1 | \beta_2) &= y'X(X'X)^{-1}X'y - y'X_2(X'_2X_2)^{-1}X'_2y \\ &= y' [X(X'X)^{-1}X' - X_2(X'_2X_2)^{-1}X'_2]y \end{aligned} \tag{3.13}$$

벡터 ε 에 대한 표준 가정(standard assumption) 하에서, H_0 에서 $R(\beta_1 | \beta_2)/\sigma^2$ 는 $\chi^2_{p_1}$ 이며 잔차제곱합(residual sum of squares)과 독립적이다. 그러므로 H_0 가정 하에서 $[R(\beta_1 | \beta_2)/p_1]/s^2$ 는 $F_{p_1, n-p}$ 이다. 그러므로, 다음의 검정통계량(test statistic)은

$$F = \frac{R(\beta_1 | \beta_2) / p_1}{s^2}$$

F-분포(*F*-distribution)를 사용하는 상부꼬리 한쪽꼬리검정 기준(upper tail one-tailed test criterion)을 제공하며, 이 기준에 따라 추가제곱합 회귀(extra sum of squares regression) 또는 감소된 잔차제곱합(residual sum of squares reduced)이 β_1 을 모형에 포함시키는 근거로 충분한지를 결정할 수 있다.

단순회귀계수의 검정, 부분 *F*와 *t*-검정(*Tests on Single Regression Coefficients. Partial F and t-Tests*)

위의 전개과정은 회귀모수(regression parameter)의 어떤 부분집합(any subset)을 검정(test)하는데에도 이용되는 근거(basis)를 제공한다. β_1 이 스칼라(scalar)인 경우, 즉 단순회귀계수(single regression coefficient)를 검정할 때 $p_1=1$ 이고 *F*-test가 $(1, n-p)$ 의 자유도일 때를 부분 *F*-검정(partial *F*-test)이라고 부른다. 양쪽꼬리 *t*-검정(two-tailed *t*-test)은 동일한 정보(information)를 제공할 수 있다. 3.3 절에서 지적한 대로 j 번째 회귀계수의 분산은 $\sigma^2 \cdot c_{jj}$ 이며 c_{jj} 는 $(X'X)^{-1}$ 의 j 번째 대각원소(diagonal element)이다.

$$\begin{aligned} H_0: \beta_j &= 0 \\ H_1: \beta_j &\neq 0 \end{aligned} \quad \left. \right\} (j = 0, 1, 2, \dots, k)$$

의 *t*-검정은 다음을 이용하여 완성할 수 있다.

$$t = \frac{b_j}{s\sqrt{c_{jj}}} \quad (3.14)$$

b_j 의 추정 표준편차(estimated standard deviation)인 $s\sqrt{c_{jj}}$ 는 b_j 의 표준오차(standard error)로 불린다. 위에서 언급한 이 검정은 부분 *t*-검정과 $t^2 = F$ 라고 불리는데 여기서 *F*는 동일한 가설(same hypothesis)에서의 부분 *F*-통계량(partial *F*-statistic)이다. 회귀계수에 대한 *t*-검정과 부분 *F*-검정 사이의 차이는 *t*-검정이 방향(direction)을 알려준다는 것이다. (3.14)에서 우리는 *t*-값(*t*-value)의 부호(sign)가 회귀계수(regression coefficient)의 부호에서 온다는 것을 알 수 있다. 그러나 유의한 부분 *F*(partial *F*)값은 계수의 방향(direction)을 나타내지 않는다.

그러므로 추가제곱합원리(extra sum of squares principle)는 최소제곱 추정과정(least squares estimation procedure), 즉 단일회귀계수(single regression coefficient)나 부분집합(subset)에 대한

가설(hypothesis)을 검정하는 기전(mechanism)에 주로 관심이 있는 사용자에게 도움이 될 것이다. SS_{Reg} 의 통계적 성질(statistical property)과 설명된 변동(variation explained)에 관해 알려진 것들을 활용(exploit)할 수 있다. 단순히 대입을 하면, 식 (3.12)의 귀무가설(H_0)을 기각하는 증거는 $X_2\beta_2$ 가 있는 모형에 $X_1\beta_1$ 를 추가(place)함으로써 발생하는 추가변동(extra variation)이 우연(chance)에 의해 발생하는 것 보다 더 큰 경우이다.

개별 모수에 대한 순차적 F-검정(Sequential F-tests on Individual Parameters)

예제 3.2에서 설명될 부분 F-검정(partial F-test)은 SS_{Reg} 에 더해지는 제곱합의 기여(contribution)로 형성(form)되지 않는다. 그것은 일반적으로 다음과 같다

$$R(\beta_1 | \beta_0, \beta_2, \dots, \beta_k) + R(\beta_2 | \beta_0, \beta_1, \beta_3, \dots, \beta_k) + \dots + R(\beta_k | \beta_0, \beta_1, \beta_2, \dots, \beta_{k-1}) \neq R(\beta_1, \beta_2, \dots, \beta_k | \beta_0)$$

그러므로 부분 F-검정에 이용되는 추가제곱합(extra sum of squares)은 SS_{Reg} 의 완전한 분할(partition)을 이루지는 않는다. 그 결과로, 일반적으로, 제곱합(sum of squares)은 그 자체로 독립적이지 않다. 그러나 검정(test)을 식 (3.9)의 우변(right-hand side)의 구성요소를 사용하여 독립적인 제곱합(independent sums of squares)으로부터 할 수 있다. 이러한 F-검정을 순차적 F-검정(sequential F-test)이라고 부른다. 순차적(sequential)과 부분(partial) F-검정(F-test)을 예제 3.2에서 설명하고 있다.

예제 3.2 오징어 자료

상어와 삼치가 먹은 오징어의 크기를 연구하기 위해 아래와 같은 변수를 선택하여 연구를 진행하였다. 회귀변수는 오징어의 부리나 입의 특징에 관계되는 것이며, 반응변수는 삼치가 먹은 오징어의 크기이다.

- x_1 : 부리의 길이(inch)
- x_2 : 날개의 길이(inch)
- x_3 : 부리- 새김눈(notch) 길이
- x_4 : 새김눈 - 날개 길이
- x_5 : 넓이(inch)
- y : 무게(pounds)

표본의 크기는 22개이며 자료는 표 3.2에 나타내었다. 모형은 다음과 같다.

$$y_I = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

먼저, 식(3.7)을 이용한 분산분석(analysis of variance)을 이용하여 분석결과를 살펴보자.

원인(source)	제곱합(SS)	자유도(df)	평균제곱(MS)	F
회귀(regression)	208.007	5	$\frac{SS_{\text{Reg}}}{5} = 41.6015$	$F = \frac{MS_{\text{Reg}}}{s^2} = 84.070$
오차(residual)	7.918	16	$S^2 = 0.4948$	
전체(total)	215.925	21		

회귀에 대한 자유도 5는 상수항을 제외한 항의 개수를 나타낸다. 따라서, 잔차에 대한 자유도는 $n-p = 22-6 = 16$ 이다. 위의 분산분석표를 이용하여 $R(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 | \beta_0)$ 에 대한 가설검정을 할 수 있다. 즉, 식 (3.7)에 있는 이 값을 회귀제곱합(자유도 5)이라고 부르며 아래의 가설을 검정하는데 사용된다.

$$H_0 : \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = 0$$

이것은 모수(parameter)의 부분집합(subset)에 대한 검정의 특별한 경우(case)로 볼 수 있다. F-통계량이 통계적으로 유의하다는 것은 ($F = 84.070$) 적어도 하나 이상의 회귀변수가 오징어의 무게에 영향을 미치고 있다는 것이다. 상기 분석의 간단한 부산물(by-product)로

$$R^2 = \frac{208.007}{215.925} = 0.9633$$

이것은 5개의 회귀변수를 이용한 모형이 오징어 무게의 변동을 96.33% 설명한다는 것을 의미한다. 순차(sequential) 제곱합과 부분(partial) 제곱합은 다음과 같다.

순차(sequential)	부분(partial)
$R(\beta_1 \beta_0) = 199.145$	$R(\beta_1 \beta_0, \beta_2, \beta_3, \beta_4, \beta_5) = 0.298731$
$R(\beta_2 \beta_0, \beta_1) = 0.126664$	$R(\beta_2 \beta_0, \beta_1, \beta_3, \beta_4, \beta_5) = 0.868762$
$R(\beta_3 \beta_0, \beta_1, \beta_2) = 4.119539$	$R(\beta_3 \beta_0, \beta_1, \beta_2, \beta_4, \beta_5) = 0.078273$
$R(\beta_4 \beta_0, \beta_1, \beta_2, \beta_3) = 0.263496$	$R(\beta_4 \beta_0, \beta_1, \beta_2, \beta_3, \beta_5) = 0.982690$
$R(\beta_5 \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = 4.352193$	$R(\beta_5 \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = 4.352193$

순차 제곱합(sequential sum of squares)을 더해가면 208.007이 되어, 모두 회귀변수(regressors)로 설명되는 회귀 제곱합(regression sum of squares)이 된다. 하지만 부분 제곱합(partial sum of squares)은 어떤 것을 더하더라도 특별한 의미는 없다. 다만, 부분 제곱합(partial sum of squares)은 개별적으로 사용되었을 때 하나의 회귀계수에 대한 가설검정을 할 수 있다. 반면, 순차 제곱합(sequential sum of squares)은 회귀계수의 부분집합에 대한 가설검정을 할 수 있다. 한 예로 우리가 다음을 검정하기를 원한다고 가정하자.

$$H_o : \begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix} = 0$$

만약 검정결과 H_0 를 기각할 수 없다면 넓이, 새김눈과 날개사이의 길이에 대한 계수는 0과 유의하게 다르다는 것을 동시에 발견할 수 없는 것이다. 이 예제에 적절한 제곱합 $R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3)$ 은 다음과 같이 구할 수 있다.

$$\begin{aligned} & R(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 | \beta_0) - R(\beta_1, \beta_2, \beta_3 | \beta_0) \\ &= R(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 | \beta_0) - [R(\beta_1 | \beta_0) + R(\beta_2 | \beta_1, \beta_0) + R(\beta_3 | \beta_0, \beta_1, \beta_2)] \\ &= 208.007 - (199.145 + 0.126664 + 4.119539) \\ &= 4.615689 \end{aligned}$$

다른 방법으로는 아래와 같이 쓸 수 있다.

$$\begin{aligned} & R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3) = \\ & R(\beta_4 | \beta_0, \beta_1, \beta_2, \beta_3) + R(\beta_5 | \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \\ &= 0.263496 + 4.352193 = 4.615689 \end{aligned}$$

따라서 검정통계량은 다음과 같다.

$$F = \frac{4.615689 / 2}{s^2} = \frac{2.307845}{0.494845} = 4.663683$$

여기서 분자와 분모의 자유도는 각각 2와 16이다. 이 가설은 0.025수준에서 기각된다.

부분 제곱합(partial sum of squares)은 F -통계량을 이용하여 개별적 회귀계수에 대한 가설검정을 할 수 있고 또 다른 방법으로 (3.14)에 주어진 것처럼 t -통계량을 이용하여 할 수도 있다. 대부분의 컴퓨터 패키지에서 회귀분석의 결과로 회귀계수의 추정값과 이들의 표준오차, 이는 (3.14)에 있는 t -통계량의 분모로 주어진다. 예를 들어 목록(listing)은 다음과

같다.

	Coefficient	Standard Error	t	Prob > t
b_0	-6.512215	0.933561	-6.976	0.0001
b_1	1.999413	2.573338	0.777	0.4485
b_2	-3.675096	2.773660	-1.325	0.2038
b_3	2.524486	6.347495	0.398	0.6961
b_4	5.158082	3.660283	1.409	0.1779
b_5	14.401162	4.855994	2.966	0.0091

마지막 열은 t -통계량이 어느 수준에서 유의한지를 나타내는 p -값이 주어져 있다. 여기에서는 b_1, b_2, b_3 그리고 b_4 가 유의하지 않다는 것을 나타낸다. 그러나 여기서 주의해야 할 것이 있다. t -검정은 다른 회귀변수들이 모형에 포함되었을 때 확실하게 타당성 있고 유효하지만 회귀변수들의 세트(set)를 동시에 이용하여 변수를 선별하고 모형을 구축해야 할 때에는 위험할 수 있다. 그러므로 부리넓이인 x_5 만 회귀모형에 남겨두는 것은 바람직한 결론은 아니다. 4장에서는 하나의 기준값을 사용하여 모형을 설정하는 방법에 대해 다룰 것이고 이러한 의미에서 추가적인 토론을 해 볼 것이다.

Table 3.2 오징어 자료

x_1	x_2	x_3	x_4	x_5	y
1.31	1.07	0.44	0.75	0.35	1.95
1.55	1.49	0.53	0.90	0.47	2.90
0.99	0.84	0.34	0.57	0.32	0.72
0.99	0.83	0.34	0.54	0.27	0.81
1.05	0.90	0.36	0.64	0.30	1.09
1.09	0.93	0.42	0.61	0.31	1.22
1.08	0.90	0.40	0.51	0.31	1.02
1.27	1.08	0.44	0.77	0.34	1.93
0.99	0.85	0.36	0.56	0.29	0.64
1.34	1.13	0.45	0.77	0.37	2.08
1.30	1.10	0.45	0.76	0.38	1.98
1.33	1.10	0.48	0.77	0.38	1.90
1.86	1.47	0.60	1.01	0.65	8.56
1.58	1.34	0.52	0.95	0.50	4.49
1.97	1.59	0.67	1.20	0.59	8.49

1.80	1.56	0.66	1.02	0.59	6.17
1.75	1.58	0.63	1.09	0.59	7.54
1.72	1.43	0.64	1.02	0.63	6.36
1.68	1.57	0.72	0.96	0.68	7.63
1.75	1.59	0.68	1.08	0.62	7.78
2.19	1.86	0.75	1.24	0.72	10.15
1.73	1.67	0.64	1.14	0.55	6.88

순차검정과 부분 검정에 대한 추가 논평(*Further Comments on Sequential and Partial Tests*)

많은 회귀분석 소프트웨어 패키지에는 순차 F -검정 (sequential F -statistics)과 부분 F -검정 (partial F -statistics)이 모두 포함되어 있다. 만약 변수선별과정(process of variable screening)을 통한 모형 구축(model building)에 관심이 있다면, 두 가지 유형의 검사 모두 매우 명백한 단점(drawback)을 가진다. 아마도 이러한 단점을 보여주는 가장 효과적인 방법은 각각이 무엇을 하기 위해 고안되었는지를 독자들에게 상기시키는 것일 것이다.

순차 F -검정(*Sequential F-test*)

이 검정은 선행하는 회귀 변수들(preceding regressor variables)을 포함하는 모형에서 하나의 변수가 기여(contribution)하는 정도에 대하여 알려준다. 이때 진입 순서(the order of entry)는 결과에 현저한 영향을 미칠 수 있다. 만약 회귀변수 4가 회귀변수 1, 2, 3이 있는데 조정(adjust)되었다면, SS_{Reg} 에 대한 회귀 변수 4의 기여 정도는 그것이 단지 회귀 변수 10이 있을 때에 조정(adjust)되었을 때와는 매우 다를 것이다. 다시 말해서, 한 회귀 변수의 적합성(appropriateness)은 종종 어떤 회귀변수(regressor variable)들이 그 모형에 함께 존재하는지에 의해 좌우된다. 그러므로, 순차 F -검정이 가상의 선발 순서(imaginative selection of order)와 조화롭게 사용되고, 그 과정이 많은 단계들에 의해서 실행되지 않는다면, 순차 F -검정에 의해서 모든 변수들을 포함하는 완전한 변수선별(full scale variable screening)이 효과적으로 이루어질 수 없다. 실제로 이것은 4장에서 다루어질 Stagewise regression Procedures (전진 선택(Forward Selection), 후진제거(Backward Elimination), 단계적 회귀(Stepwise regression))의 연산(operation) 논리(logic)를 제공한다.

부분 F -검정(*Partial F-test*)

부분 F -검정(parital F -test)은, 모든 회귀변수들이 관련되어 있는 모형에서 하나의 회귀변수의 중요성에 관한 정보를 제공한다. 물론 이 방법은 연구자가 특정 회귀변수의 역할에 대하여 관심이 있을 때 많은 정보를 제공할 수 있다. 그러나, 대규모 변수선별 과정(large scale variable screening procedure)에서, k 개의 회귀변수(k regressor variables)에 대한

k 개의 부분 F -검정(partial F -test)에서 얻은 정보만을 사용함으로써 의사결정(decision making)에 어려움을 종종 겪을 수 있다.

물론, 모든 회귀변수들의 존재 하에 한 변수가 통계적인 유의성을 가진다 해도, 부분집합(subset)에서도 이 변수가 중요한 공헌을 한다고 볼 수는 없다. 반대로, 유의성이 없는 변수도 특정 부분집합(subset)에서는 중요해질 수 있다. 회귀변수들 간의 상호관계(interrelationship)와 다중 연관성(multiple association) (다중공선성, multicollinearity) 때문에, 부분 F -검정(partial F -test)만을 사용하여 최적모형(best model)에 관한 결론을 도출하기란 매우 어렵다.

예를 들어 오징어 자료의 경우, 부리길이인 변수 x_1 은 모든 다른 변수들의 존재 하에서는 유의하지 않지만($t=0.78$, $p=0.4485$) 하나의 회귀변수로서의 x_1 은 $F = \frac{199.145}{16.780/20} = 237.360$ 으로 중요하게 보인다.

상기 검정통계량(test statistic)의 분자는 순차회귀제곱합(sequential regression sum of squares)이고, x_1 은 모형으로의 초기진입(initial entrance)을 나타낸다. s^2 은 x_1 이 유일한 회귀변수일 때의 회귀로부터 산출된다. 명백히 이것은 x_1 이 반응의 변동(variation in the response)의 상당 부분을 설명한다는 증거이다. 그러나, x_1 에 대한 t -검정 (또는 부분 F -검정)은 유일한 회귀변수로서의 x_1 (x_1 as a lone regressor)에 의해 설명되는 변동(variation)이 나머지 회귀변수들(the other regressors)에 기인함을 뜻한다.

대규모의 변수 선별(large-scale screening)에 있어, 부분 F -검정(partial F -test)과 순차적 F -검정(sequential F -test)의 역할은 4장에서 논의될 것이다.

일반선형가설 검정(Test of the General Linear Hypothesis)

각각의 회귀모수(regression parameter)와 모수의 부분집합(subset)에 대한 검정(test)은 융통성(flexibility)이 훨씬 큰, 보다 일반적인 틀 내에서 해결될 수 있다. 예를 들어, 두개의 독립적인 자료세트를 가지는 두개의 성장 모형(growth model)을 만드는 생물학자가 있다고 생각해보자. 또는 남성과 여성 각각에 대한 두 개의 회귀식을 만들고자 하는 사회학자가 있다고 하자. 각 경우에서, 회귀의 모수들이 두 개의 자료세트에 따라 현저하게 다른지 아닌지에 대해 사용자가 알고 있는 것은 매우 중요할 수 있다. 이러한 유형과 이외 다른 유형의 가설들을 일반선형가설(general linear hypothesis)을 사용하여 검정할 수 있다.

$$H_0 : C\beta = d$$

$$H_1 : C\beta \neq d$$

(3.15)

이 가설은 식 (3.12)에 있는 중요한 가설을 특수한 경우에 맞게 적용한 것이다. (3.15)에서

행렬 C 는 열(rank)이 r 이고 $r \leq p$ 인 $r \times p$ 행렬이다. 즉, 이 가설은 우리로 하여금 그 모수들에 대해서 $r \leq p$ 라는 넘치지 않고, 모순되지 않는 진술($r \leq p$ nonredundant and noncontradictory statements)을 가정하도록 허용한다.

일반 선형 가설의 실례(Illustrations of the General Linear Hypothesis)

(자료세트들의 회귀의 평행성과 동등성 검정: Test for Parallelism and Equality of Regressions Across Data Sets)

독립된 자료 세트(independent data set) 간에 독립적인 회귀식(independent regression equations)이나 개별 모수(individual parameter)를 비교하는 것의 중요성에 대해 언급하였다. 우리가 두개의 독립적인 자료 세트를 가지고 있다고 가정해보자.

$$\begin{array}{cc} y & x \\ y_1 & x_1 \\ y_2 & x_2 \\ \vdots & \vdots \\ y_{n_1} & x_{n_1} \end{array} \left. \right\} \text{Data set 1}$$

$$\begin{array}{cc} y_{n_1+1} & x_{n_1+1} \\ \vdots & \vdots \\ y_{n_1+n_2} & x_{n_1+n_2} \end{array} \left. \right\} \text{Data set 2}$$

그리고, 두 모형을 가정해보자.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \dots, n_1$$

$$y_i = \gamma_0 + \gamma_1 x_i + \varepsilon_i \quad i = n_1 + 1, \dots, n_1 + n_2$$

두 개의 회귀선(regression lines)의 기울기가 동일한지를 검정하기를 원한다고 생각해보자. 우리는 이 두개의 분리된 모형들을 하나의 모형(single model)으로 나타낼 수 있다. 즉,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_1} \\ \hline y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ 1 & \cdot & x_2 & \cdot \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{n_1} & 0 \\ \hline 0 & 1 & 0 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{n_1+n_2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \gamma_0 \\ \beta_1 \\ \gamma_1 \end{bmatrix} + \varepsilon$$

그 결과, 기울기의 동등성(equality)에 대한 가설은 다음과 같이 쉽게 나타낼 수 있다.

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \gamma_0 \\ \beta_1 \\ \gamma_1 \end{bmatrix} = 0$$

만약 전체 회귀선의 동등성(equality)을 검정할 필요가 있다면, 그 가설은 다음과 같이 될 수 있다.

$$H_0 : \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \gamma_0 \\ \beta_1 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

두개의 회귀선의 기울기가 특정 값 (예를 들어 2.0)으로 동등하다는 가설을 검정하고자 한다면, 그때에는 단순히 다음을 검정하면 된다.

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \gamma_0 \\ \beta_1 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

일반선형가설 검정 전개(Development of the Test of the General Linear Hypothesis)

앞서 나타내었듯이, 일반선형가설은 다중선형회귀모형의 모수(parameter)들에 대한 선형진술(linear statement)과 관련된 가설을 검정하도록 고안된 것이다. 그 가설을 검정하기 위해서는, 우리는 물론 ε 에 대한 정규이론가정(normal theory assumption)을 해야 한다. 즉, 우리는 $\varepsilon \sim N(0, \sigma^2 I_n)$ 를 가정한다. 이러한 조건 하에서는, $b \sim N(\beta, \sigma^2 (X'X)^{-1})$ 이 성립함을 3.3절의 내용으로부터 상기해보라. 여기에서 $\sigma^2(X'X)^{-1}$ 는 계수 벡터(coefficient vector) b 의 분산공분산행렬(variance-covariance matrix)이다. 검정을 받는 모수(parameter)벡터는 $C\beta$ 이고, 이것은 Cb 로부터 추정된다. Cb 의 원소(elements)는 정규분포를 하는 확률변수(random variables)의 선형 조합(linear combination)이다. 벡터 Cb 는 $C\beta$ 에 대해서 비편향(unbiased)이고, Cb 의 분산공분산행렬(variance-covariance matrix)은 $\sigma^2 C(X'X)^{-1}C'$ 이다. 따라서 우리는 다음과 같이 말할 수 있다.

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

$H_0: C\beta = d$ 하에서 선형 모형 이론(theory of linear models)은 다음을 나타낸다.

$$\frac{[Cb - d]'[C(X'X)^{-1}C']^{-1}[Cb - d]}{\sigma^2} \sim \chi_r^2$$

그리고,

$$\frac{[Cb - d]'[C(X'X)^{-1}C']^{-1}[Cb - d]}{rs^2} \sim F_{r, n-p}$$

Graybill (1976)의 저서를 참고하시오. 이 통계량(statistic)은 이 절에 있는 다른 F -통계량의 경우에서처럼 평균제곱의 비(ratio of mean squares)로 간주될 수 있다. 위의 식에서 분자는 Cb 의 분산공분산행렬(variance-covariance matrix)에 의해서 표준화된, Cb 와 d 간의 “거리의 제곱(squared distance)”을 의미함을 주목하여야 한다.

양(quantity) r 은 분자의 자유도이며, 그것은 모수(parameter)의 수 또는 검정을 받는 모수(parameter)들의 선형 조합의 수를 나타낸다. 가설 (3.15)을 검정하기 위한 과정은 다음의 통계량에 근거하여 귀무가설 H_0 를 기각하는 것이다.

$$F = \frac{[Cb - d]'[C(X'X)^{-1}C']^{-1}[Cb - d]}{rs^2} \quad (3.16)$$

H_0 의 기각은 상부꼬리 한쪽꼬리 F -검정(upper-tailed one-tailed F -test)에 근거하여 이루어진다. 그 개념은 이 절에서 토의된 다른 분산분석 F -검정(other analysis of variance F -test)에서와 유사하다. 사실 β 벡터 (식 (3.13)참조)의 부분집합(subset)에 대한 검정 통계량은 식 (3.16)의 특수한 경우에 해당한다.

예제 3.3 판매액과 광고비 자료

일반적으로 광고비의 투자가 많을수록 판매액이 증가할 것이라 생각할 수 있다. 이를 구체적으로 조사하기 위하여 같은 제품을 만드는 10개의 회사들에 대한 광고비와 판매액을 조사한 결과 table 3.3의 자료를 얻었다.

Table 3.3 판매액과 광고비 자료

	그룹	광고비	판매액
신문	1	6	42
	1	8	47
	1	9	50
	1	12	57

방송	2	4	39
	2	6	45
	2	8	50
	2	9	52
	2	10	55
	2	12	60

자료에서 1,2,3,4에 해당하는 회사는 신문을 광고 매체로 하였고, 나머지 회사는 방송을 광고 매체로 하였다. 광고매체에 따라 광고비의 판매액에 대한 효과에 차이가 있는지를 알아보기 위해 광고매체별 따라 단순선형회귀모형을 가정하고 절편과 기울기를 검정하였다. 즉, 다음의 두 모형을 가정할 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, 3, 4 \quad (\text{광고매체가 신문})$$

$$y_i = \gamma_0 + \gamma_1 x_i + \varepsilon_i \quad i = 5, 6, 7, 8, 9, 10 \quad (\text{광고매체가 방송})$$

여기서 y_i 는 판매액, x_i 는 i 번째 회사의 광고비이다.

관심이 있는 가설은 $H_0: \beta_1 = \gamma_1$ 이며, 대립가설은 2개의 기울기가 같지 않다는 것이다. X 행렬과 β 벡터는 아래와 같이 주어진다.

$$X = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ 1 & 0 & x_2 & 0 \\ 1 & 0 & x_3 & 0 \\ 1 & 0 & x_4 & 0 \\ 0 & 1 & 0 & x_5 \\ 0 & 1 & 0 & x_6 \\ 0 & 1 & 0 & x_7 \\ 0 & 1 & 0 & x_8 \\ 0 & 1 & 0 & x_9 \\ 0 & 1 & 0 & x_{10} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma_0 \\ \gamma_1 \end{bmatrix}$$

C 행렬은 2개의 기울기의 동등성(equality)을 검정하도록 고안되었으며, 그 결과는 다음과 같다.

$$C = [0 \quad 0 \quad 1 \quad -1] \quad d = 0$$

Table 3.3의 자료에 대한 2개의 단순 선형회귀모형에 적합한 최소 제곱은 다음과 같다.

$$\mathbf{b} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\gamma}_0 \\ \hat{\gamma}_1 \end{bmatrix} = \begin{bmatrix} 27.0667 \\ 29.0000 \\ 2.5067 \\ 2.5918 \end{bmatrix}$$

$(X' X)^{-1}$ 행렬은 아래에서 주어진 바와 같다.

$$(X'X)^{-1} = \begin{bmatrix} 4.3333 & 0 & -0.4667 & 0 \\ 0 & 1.8000 & 0 & -0.2000 \\ -0.4667 & 0 & 0.0533 & 0 \\ 0 & -0.2000 & 0 & 0.0245 \end{bmatrix} \quad (3.17)$$

잔차 평균제곱 (residual mean square), $s^2=0.11950$ 이다. 이제 기울기가 동일한지에 대한 가설을 검정하기 위해 식 (3.16)을 이용할 수 있다.

$$F = \frac{(Cb - d)'[C(X'X)^{-1}C']^{-1}(Cb - d)}{s^2} = 0.7797$$

자유도 1과 6일 때의 F 값은 기각역은 5.99로 기각역에 속하므로 그 결과 기울기가 같지 않다는 것은 증거가 없다. 두 회귀모형에서 기울기는 동등성 검정(equality test) 결과 귀무가설을 기각할만한 충분한 근거를 얻을 수 없었다. 이러한 결과는 기울기의 추정값을 통해서도 확인해 볼 수 있다. 하지만, 절편의 추정값을 보았을 때 차이가 있는 것으로 보이기 때문에 이 자료에 대한 심도 있는 분석이 필요하며, 이 장의 뒷편에서 보다 심도 있게 다루어질 것이다. 예제에 사용된 R-code는 다음과 같다.

```
x1<-c(1,1,1,0,0,0,0,0)
x2<-c(0,0,0,0,1,1,1,1,1)
x3<-c(6,8,9,12,0,0,0,0,0)
x4<-c(0,0,0,4,6,8,9,10,12)
x<-as.matrix(cbind(x1,x2,x3,x4))
y<-c(42,47,50,57,39,45,50,52,55,60)
c<-as.matrix(c(0,0,1,-1))
d<-0

inv<-solve(t(x)%*%x)
be<-inv%*%t(x)%*%y
h<-x%*%inv%*%t(x)
i<-diag(1,nrow=10)
```

```

sse<-t(y)%%*(i-h)%%*y
s_sq<-sse/6
s<-sqrt(s_sq)

f_inv<-solve(t(c)%%inv%%c)
f<-t(t(c)%%be)%%f_inv%%(t(c)%%be)/(s_sq)

dat<-data.frame(cbind(x,y))
g<-lm(y~x-1,data=dat)
ano<-anova(g)

```

일반 선형 가설과 추가제곱합 원리(*The General Linear Hypothesis and the Extra Sum of Squares Principle*)

일반선형가설에 앞서 가설검정 개발(hypothesis testing development)에 있어서, 우리는 추가제곱합 원리(extra sum of squares principle)에 대해서 상당한 노력을 쏟았다. 어떤 모형에서 회귀변수(또는 부분집합)의 유용성은 그 모형에서 나머지 회귀변수의 존재 하에 그 회귀변수(또는 부분집합)에 의해서 설명되는 변동(variation)(회귀 제곱합)의 분율(proportion)에 의존한다는 개념의 중심에는 매우 단순한 개념이 자리잡고 있다. 명백히, 유의하게 기여하는 것이 무엇인가에 대한 결정은 F -검정을 근거로 하여 이루어진다. 대신에, 설명되는 변동(variation explained)은 ‘축소된 잔차제곱합 (residual sum of squares reduced)’으로써 쉽게 간주될 수 있다. 예로써, 예제 3.2의 오징어 자료를 생각해보자. 아래 가설의 검정은

$$H_0 : \begin{bmatrix} \beta_4 \\ \beta_5 \end{bmatrix} = 0$$

다음을 이용하여 할 수 있다.

$$R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3) = R(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 | \beta_0) - R(\beta_1, \beta_2, \beta_3 | \beta_0) \quad (3.18)$$

이 식은 두 모형 즉, 다섯 개의 회귀변수를 모두 포함하는 ‘완전모형(full model)’과 검정 중인 회귀변수를 제외한 나머지 것만 포함하는 ‘축소모형(reduced model)’에 있어서의 SS_{Reg} 의 차이를 나타낸다. 독자들은 ‘축소 모형’을 귀무가설 H_0 의 제약(constraint)이 있을 때 존재하는 모형인 것으로 생각하여야 한다. SS_{Reg} 와 SS_{Res} 가 더해지면 SS_{Total} 이 되기 때문에, 식 (3.18)은 또한 ‘완전모형’과 ‘축소모형’ 간의 잔차제곱합의 차이(differences in the residual sum of

squares)로 간주될 수 있음은 명백하다. 다시 말해서, 오징어 자료에서, β_4 와 β_5 가 존재함으로써 증가되는 SS_{Reg} 의 양은 β_4 , β_5 를 모형에 포함시켰을 때에 감소되는 SS_{Res} 의 양과 같다. 따라서

$$R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3) = SS_{\text{Res, reduced}} - SS_{\text{Res, full}}$$

이며, 이것은 잔차제곱합(residual sum of squares)의 감소를 말한다.

우리는 이 식을 주목하였는데, 일반 선형 가설을 사용할 때에 ‘잔차제곱합’의 사용은 더 큰 유통성을 허용하기 때문이다. 이 시점에서 모수(parameter) 부분집합(subset)이 0이라는 것을 검정하기 위해서 ‘추가제곱합(extra sum of squares)’을 제공하는 식 (3.13)이 다음과 같이 쉽게 기술되어 왔음을 알아야 한다.

$$\begin{aligned} R(\beta_1 | \beta_2) &= (y' y - y' X_2 (X_2' X_2)^{-1} X_2' y) - (y' y - y' X (X' X)^{-1} X' y) \\ &= SS_{\text{Res, reduced}} - SS_{\text{Res, full}} \end{aligned}$$

완전모형(full model)과 축소모형(reduced model)을 비교하는 개념은 일반선형가설에도 확장된다. 우리는 식 (3.15)의 일반선형가설에 대한 검정 통계량이 (3.16)에 나온다는 것을 알고 있다. 선형모형의 이론[Graybrill (1976) 또는 Searle (1971) 참조]은 (3.15)의 가설에 대한 이론상의 결과를 제공하고, 이는 우리가 (3.16)의 분자를 추가제곱합 원리(extra sum of squares principle)에 의해서 산출할 수 있게끔 한다.

$$(Cb - d)' [C(X' X)^{-1} C']^{-1} (Cb - d) = SS_{\text{Res, reduced}} - SS_{\text{Res, full}} \quad (3.19)$$

우리는 식 (3.19)가 어떻게 도출되었는지에 대해 자세히 언급하지는 않을 것이다. ‘완전 모형(full model)’, 즉 모수(parameter)들에 대하여 아무런 제약이 없는 모형은 최소의 잔차제곱합(smallest residual sum of squares)을 가능하게 한다. 귀무가설(null hypothesis)은 모수(parameter)들에 제약을 두어서 잔차제곱합의 증가를 초래한다. H_0 하에서 SS_{Res} 의 증가가 통계적으로 유의하지 않을 때 귀무가설이 지지된다. 반대로, 대립가설(alternative hypothesis)은 모형에 H_0 를 유도하는 것이 SS_{Res} 의 유의한 증가를 초래할 때 지지된다. H_0 의 기각은 (3.16)에 나와 있는 검정 통계량의 값이 클 때이다. 예제 3.4는 예제 3.3의 상황에서 추가제곱합 원리(extra sum of squares principle)를 사용하는 것을 약술(outline)한다.

예제 3.4

Table 3.3의 자료를 다시 살펴보자. 회사관계자는 2 그룹의 기울기가 동일한지를 검정하는데에 관심이 있음을 상기하라. 우선, 각각의 기울기와 절편을 가지는 2개의 분리된 회귀식에 의해 묘사되는 ‘완전모형(full model)’을 고려해보자. X 행렬은 식 (3.17)와 같이 나타난다. 분산분석의 정보를 포함하여 분석 결과가 표 3.4에 나와 있다. ‘축소모형(reduced model)’은 가설 $H_0: \beta_1 = \gamma_1$ 에 관련되어 있다. 따라서, 우리는 축소모형(reduced model)을 다음과 같이 쓸 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, 3, 4 \quad (\text{광고매체가 신문})$$

$$y_i = \gamma_0 + \beta_1 x_i + \varepsilon_i \quad i = 5, 6, 7, 8, 9, 10 \quad (\text{광고매체가 방송})$$

그러면, 축소모형의 X행렬과 회귀계수 벡터는 아래와 같이 쓸 수 있다.

$$X = \begin{bmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 0 & x_3 \\ 1 & 0 & x_4 \\ 0 & 1 & x_5 \\ 0 & 1 & x_6 \\ 0 & 1 & x_7 \\ 0 & 1 & x_8 \\ 0 & 1 & x_9 \\ 0 & 1 & x_{10} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \gamma_0 \\ \beta_1 \end{bmatrix}$$

여기에서, 추정해야 할 회귀계수는 3개이다. 완전 모형(full model)은 4개의 모수(parameters)를 포함한다. Table 3.5는 축소 모형에서의 분석 결과이다.

Table 3.4와 table 3.5의 분석 결과는 일반선형가설을 사용하여 검정을 하는데 있어서 실용적으로 접근할 수 있게끔 한다. 컴퓨터 소프트웨어를 사용하기 때문에, 우리는 table 3.4와 3.5에서처럼 완전모형(full model)과 축소모형(reduced model)을 실행하기만 하면 된다. 그러면, 검정통계량은 다음과 같다.

$$F = \frac{(SS_{\text{Res, reduced}} - SS_{\text{Res, full}})}{\Delta df \cdot s^2}$$

여기에서 s^2 값은 완전모형(full model)에서의 오차평균제곱(error mean square)이고, Δdf 는 완전모형(full model)과 축소모형(reduced model)에서의 오차 자유도(error degrees of freedom)의 차이를 말한다. 이것은 귀무가설안의 회귀모수의 수 또는 회귀계수의 선형조합의 수를 나타낸다. (식 (3.16) 참조)

판매와 광고비 자료와 관련된 예와 두 개의 회귀선의 평행성 검정(test for parallelism)에 있어서, 잔차 제곱합(오차 제곱합)이 분산분석에 나타난다는 점을 유의하라. 따라서 우리는 다음을 얻게된다.

$$F = \frac{0.8105 - 0.7173}{1 \times 0.1195} = 0.7797$$

물론, 이것은 예제 3.3에서 얻어진 값과 같다. 다시, F-값은 유의하지 않다.

예제에 사용된 R-code는 아래와 같다.

```
x1<-c(1,1,1,1,0,0,0,0,0,0)
x2<-c(0,0,0,0,1,1,1,1,1,1)
x3<-c(6,8,9,12,4,6,8,9,10,12)

x_r<-as.matrix(cbind(x1,x2,x3))
y<-c(42,47,50,57,39,45,50,52,55,60)

dat1<-data.frame(cbind(x_r,y))
g1<-lm(y~x_r-1,data=dat1)
ano1<-anova(g1)

re_full<-ano$"Sum Sq"[2]
re_red<-ano1$"Sum Sq"[2]

f0<-(re_red-re_full)/s_sq
```

Table 3.4 완전모형의 분산분석표

	자유도	제곱합
회귀	4	25096.3
오차	6	0.7173
전체	10	25097

Table 3.5 기울기가 동일성을 가정한 축소모형의 분산분석표

	자유도	제곱합
회귀	3	25096.2
오차	7	0.8105
전체	10	25097

예제 3.4에서 잔차제곱합(residual sum of squares)의 차이를 이용한 검정통계량을 만들어보았다. 회귀제곱합(regression sum of squares)의 차이를 이용하게 되면 같은 결과를 얻게 되는지 확인하는 것은 쉽다. 물론 이것은 총 제곱합(total sum of squares)이 완전모형(full model)과 축소모형(reduced model)에서 동등하기 때문이다. 그러나, 일반 선형 가설의 모든 응용 과정에서 이것이 항상 같지는 않을 것이다. 단순선형회귀모형을 예로 들어 보면,

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

이것은 이 예에서는 완전모형(full model)으로 간주될 수 있다. 아래의 가설을 고려해보면,

$$\begin{aligned} H_0 &: \beta_1 = 2 \\ H_1 &: \beta_1 \neq 2 \end{aligned}$$

이제, 축소모형(reduced model)은 다음과 같이 주어진다.

$$z_i = \beta_0 + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

여기에서, $z_i = y_i - 2x_i$ 이다. 물론 여기에서, z_i 를 사용한 총 제곱합(total sum of squares)은 y_i 의 총 제곱합(total sum of squares)과 동일하지 않다.

3.5. 다중회귀에서 신뢰구간 및 예측구간(Confidence Interval and Prediction)

Intervals in Multiple Regression)

회귀변수(regressor variable)의 주어진 특정값 $x_{1,0}, x_{2,0}, \dots, x_{k,0}$ 에 대한 평균반응(mean response)의 신뢰구간(confidence interval)은 2장에서 설명한 단순선형회귀에서의 신뢰구간과 마찬가지로 중요한 역할을 한다. 또한 새로운 관측값(observation)에 대한 예측구간(prediction interval)도 중요하다. 이 구간은 회귀변수가 고정된 상황(fixed condition)에서 새로운 관측값의 확률적 경계(probabilistic bound)를 설정해준다.

경계(bound)를 설정하기 위해서는 모수(parameter) $Var(\hat{y}|x = x_0)$ 가 반드시 결정되어야 한다.

여기에서

$$x' = (1, x_1, x_2, \dots, x_k)$$

그리고 벡터는

$$x'_0 = (1, x_{1,0}, x_{2,0}, \dots, x_{k,0})$$

회귀변수에 대하여 특정값(specific value)을 부여한다. 다중회귀변수(multiple regressor)의 경우 $Var(\hat{y}|x = x_0)$ 는 다음과 같다.(부록 A.2를 참조)

$$Var(\hat{y}|x_0) = \sigma^2 x'_0 (X'X)^{-1} x_0 \quad (3.20)$$

$x=x_0$ 에서 정규오차(normal error)를 가정하면, $E(y|x = x_0)$ 에서 $100(1-\alpha)\%$ 신뢰구간(confidence bound)은 다음과 같이 주어진다.

$$\hat{y}(x_0) \pm t_{\alpha/2, n-p} s \sqrt{x'_0 (X'X)^{-1} x_0} \quad (3.21)$$

여기에서 $s \sqrt{x'_0 (X'X)^{-1} x_0} = s_{\hat{y}(x_0)}$ 은 일반회귀모형(general regression model)에서의 예측표준오차(standard error of prediction)가 된다. 그러므로, $x'_0 (X'X)^{-1} x_0$ 는 σ^2 과는 별개로 모든 회귀계수의 분산(variance)과 공분산(covariance)을 포함한다는 것을 쉽게 알 수 있다. 사실 이차형(quadratic form) $x'_0 (X'X)^{-1} x_0$ 는 이 교재의 많은 부분에서 중요한 역할을 한다. 따라서

독자들은 이 시점에서 $x_0'(XX)^{-1}x_0$ 는 $x=x_0$ 일때 예측분산(prediction variance)이라는 것(σ^2 와는 별개로)에 친숙해질 필요가 있다.

2장의 단순회귀(simple regression)에서 사용된 것과 비슷한 논법으로, $x=x_0$ 일 때 새로운 관측값에 대한 신뢰구간(confidence interval)은 다음과 같이 얻어진다.

$$\hat{y}(x_0) \pm t_{\alpha/2, n-p} s \sqrt{1 + x_0'(XX)^{-1}x_0} \quad (3.22)$$

예제 3.5 헬스클럽 자료(Health Club Data)

Table 3.1의 헬스클럽 자료를 다시 이용하여 보자. 먼저 추정된 회귀식은 아래와 같다.

$$\hat{y} = -3.6186 + 1.2676x_1 - 0.5252x_2 - 0.5050x_3 + 3.9030x_4$$

또한, R^2 은 0.8531이고, 분산의 잔차추정값(residual estimate of variance) s^2 는 822(초)이었고, 검정해야 할 가설은 다음과 같다.

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

Table 3.6 회귀분석 결과

	자유도	제곱합	평균제곱	F ₀	Pr > F ₀	R-Square
회귀	4	119360.5	119360.5	36.3	<0.0001	0.8531
오차	25	20551	822			
전체	29	139911.9				

	자유도	순차 제곱합	F ₀	Pr > F ₀	자유도	부분 제곱합	F ₀	Pr > F ₀
x_1	1	89117	108.4074	<0.0001	1	16051	19.526	0.000168
x_2	1	4680	5.6929	0.02493	1	304.6	0.3706	0.548194
x_3	1	3165	3.8497	0.06098	1	3466.8	4.2173	0.050614
x_4	1	22399	27.2478	<0.0001	1	22399	27.248	<0.0001

모수	추정값	T값	Pr > T	표준오차	잔차의 표준편차	반응치의 평균
절편	-3.6186	-0.064	0.949086	56.1027	26.62081	370.9333
b_1	1.2676	4.419	0.000168	0.2869		
b_2	-0.5252	-0.609	0.548194	0.8628		
b_3	-0.5050	-2.054	0.050614	0.2459		
b_4	3.9030	5.220	<0.0001	0.7477		

	실제값	예측값	잔차	예측값의 표준오차	평균반응 신뢰하한	평균반응 신뢰상한	예측값의 신뢰하한	예측값의 신뢰상한
1	481	460.1374	20.8626	14.31869	430.6475	489.6273	394.1331	526.1417
2	292	309.4516	-17.4516	10.18609	288.473	330.4302	246.7857	372.1175
3	338	321.4851	16.51493	8.619187	303.7335	339.2366	259.8245	383.1456
4	357	341.7096	15.29041	8.042051	325.1467	358.2725	280.3807	403.0385
5	396	404.57	-8.57001	8.617791	386.8213	422.3187	342.9103	466.2297
6	429	433.8981	-4.8981	9.527739	414.2754	453.5208	371.673	496.1232
7	345	375.619	-30.619	8.325441	358.4724	392.7656	314.1299	437.1081
8	469	424.5382	44.46179	13.46604	396.8044	452.272	359.2996	489.7768
9	425	444.8113	-19.8113	10.9852	422.1868	467.4357	381.5754	508.0471
10	358	392.9628	-34.9628	10.28069	371.7893	414.1363	330.2314	455.6941
11	393	426.246	-33.246	13.24899	398.9592	453.5328	361.1962	491.2958
12	346	331.3532	14.64679	9.034497	312.7463	349.9601	269.441	393.2654
13	279	284.7448	-5.74477	15.64013	252.5333	316.9562	217.4805	352.0091
14	311	289.5347	21.46527	11.5258	265.7969	313.2726	225.892	353.1774
15	401	419.8832	-18.8832	7.608759	404.2126	435.5537	358.7892	480.9771
16	267	285.6332	-18.6332	13.79409	257.2237	314.0426	220.1045	351.1618
17	404	371.335	32.665	14.28265	341.9193	400.7507	305.3639	437.3061
18	442	434.5096	7.490447	13.64996	406.3969	462.6222	369.109	499.9101
19	368	352.8205	15.17954	6.189472	340.073	365.5679	292.4102	413.2308
20	295	301.1689	-6.1689	9.680297	281.232	321.1058	238.844	363.4938
21	391	346.5133	44.48669	9.904422	326.1148	366.9118	284.0393	408.9874
22	264	279.1054	-15.1054	10.00262	258.5046	299.7061	216.565	341.6457
23	487	505.1672	-18.1672	20.54205	462.8601	547.4744	432.5256	577.8088
24	481	471.9511	9.048895	10.84961	449.6059	494.2963	408.8146	535.0876
25	374	333.358	40.64196	7.701165	317.4972	349.2189	272.215	394.5011

26	309	336.8606	-27.8606	8.33342	319.6976	354.0236	275.3669	398.3543
27	367	347.2676	19.73245	11.7157	323.1386	371.3965	283.4779	411.0572
28	469	429.4002	39.59981	17.84789	392.6418	466.1586	359.8438	498.9566
29	252	278.7412	-26.7412	11.58871	254.8738	302.6086	215.0501	342.4323
30	338	393.2235	-55.2235	12.59571	367.2821	419.1648	328.7265	457.7205

t-통계량(*t*-statistic)과 부분 *F*-통계량(partial *F*-statistic)은 귀무가설(null hypothesis)이 기각되지 않음을 보여준다. 이는 x_1 과 x_4 의 존재 하에서 x_2 는 사실상 필요하지 않다는 것을 의미한다.

예를 들어보면, 5 번째 관측값에 대하여 생각해보자. 실제 1 마일 주행속도는 396(초)이고, 적합값은 404.57(초)이다. 이 때 예측값의 표준오차(standard error)는(식 (3.21)을 참조) 8.617791(초)이고 1 마일 주행속도 평균의 95% 신뢰구간은 관찰치 5에서

$$[386.8213, 422.3187]$$

이고, 새로 관찰되는 1마일 주행속도의 95% 예측구간은 아래와 같다.

$$[342.9103, 466.2297]$$

예제에 사용된 R-code는 다음과 같다.

```

data<-read.table("d:/data/ex3_1.R",header=TRUE)
g0<-lm(y~1)
g1<-lm(y~x1)
g2<-lm(y~x1+x2)
g3<-lm(y~x1+x2+x3)

g234<-lm(y~x2+x3+x4)
g134<-lm(y~x1+x3+x4)
g124<-lm(y~x1+x2+x4)
g123<-lm(y~x1+x2+x3)

pre<-predict(g,se=TRUE,interval="confidence")
pre1<-predict(g,data,interval="prediction")
res<-g$residuals
t3_7<-cbind(data[,5],pre1[,1],res,pre$se.fit,pre$fit[,2:3],pre1[,2:3])

x_v<-c(150,55,170,60,200,60,180,70,250,65,190,80)

```

```

x0_tem<-matrix(x_v,nrow=4)
x0_temp<-t(x0_tem)
x0<-data.frame(x0_temp)
colnames(x0)<-c('x1','x2','x3','x4')
hat_y<-predict(g,x0,se=TRUE,interval="prediction")

t3_113<-cbind(x0,hat_y$fit[,1],hat_y$se.fit,hat_y$fit[,2:3])

```

추가 언급(Further Comments)

여기까지 회귀통계량(regression statistics) 혹은 개념을 설명하기 위하여 모든 예제들을 사용하였다. 위의 자료 세트(data set)에서 컴퓨터 출력물에서 알아내지 못할 수도 있는 몇 가지 질문을 해보자.

1. t -통계량(t-statistic)이나 부분 F -통계량(partial F-statistic)에 근거할 때, x_2 가 반응에 있어서 통계적으로 무의미한 일부 변동(a statistically insignificant portion of variation in the response)을 설명하는 것으로 보인다면, 이 때 $\beta_2 x_2$ 항을 모형에서 제거해야 하는가? 모형에서 $\beta_2 x_2$ 를 제거하는 것이 1마일 주행속도를 예측하는데 있어서 더 효과적인가?

우리는 4 장에서 이 문제를 다루어 볼 것이다.

2. 예측표준오차(standard error of prediction)의 진정한 용도는 무엇인가?

분명히, 평균반응(mean response)의 신뢰한계(confidence limits)를 계산하기 위해 예측표준오차(standard error of prediction)를 이용할 수 있다. 새로운 관측값에 대한 예측 구간(prediction interval)을 포함하는 이 결과들은 대부분의 회귀통계용 패키지(regression computer packages)에서 표준적이고, 또 확실히 도움이 된다. 그러나, 자료 포인트(data point)에서의 예측표준오차(standard error of prediction)가, 우리가 예측할 필요가 있는, 우도(likelihood)를 극대화하는 위치(location)를 나타내는 것은 아니라는 점을 반드시 명심해야 한다. 사실 그 위치(locations)에서의 반응 자료(response data)는 이미 가지고 있다. 더 관심을 끄는 것은 신뢰한계(confidence limit)와 회귀변수의 다른 조합에서의 예측표준오차(예측을 하기 위해 식[the equation]을 실제로 사용하고자 하는 위치[locations]에 나타내는)이다. 예를 들어, 헬스클럽자료에서, 데이터에 존재하지 않는 x 의 조합들을 암시하는 다음과 같은 정보를 고려해보자. 신뢰한계(confidence limit)와 예측표준오차(standard error of prediction)는 다음과 같다.

x_1	x_2	x_3	x_4	$S_{\hat{y}(x_0)}$	LCL	UCL
150	55	170	60	9.449	243.793	368.141
200	60	180	70	12.853	335.990	465.415
250	65	190	80	20.907	422.357	568.520

어떤 종류의 컴퓨터 패키지에서는 회귀를 구성하는 자료 포인트(data points)에서의 예측지향적 정보(prediction-oriented information) 뿐 아니라 이런 종류의 자료도 쉽게 얻을 수 있도록 해준다. 이 예제에서는 적합값(fitted value)으로서 \hat{y} 의 질(quality)이 참 예측변수(true predictor)로서의 가치보다 우월한 것으로 보일 수도 있다. 사실 선택된 3개의 위치(회귀변수가 약간의 외삽(extrapolation)을 나타내는)에서는 식의 예측능력(capability of prediction equation)이 실제로 감소한다. 이와 같은 정보는 회귀모형의 예측(prediction) 혹은 외삽능(extrapolation capability)을 정량화(quantify)하고 최선의 것을 선택하려는 우리의 노력에 상당한 영향을 미친다.

기초적인 다중선형회귀(multiple linear regression)에 관한 내용(material)이 마무리되는 시점에서 독자들은 몇 가지 자주 반복되는 점에 대하여 알아두어야 한다. 만약 후보 모형(candidate model)들 중에 한가지 모형을 선택하는 것이 중요하다면, 몇 가지 기준(criteria)에 대하여 살펴볼 필요가 있다. 확실히 부분 F 통계량(partial F -statistic)이 어떤 종류의 증거(evidence)를 제시해주기는 하지만 결코 단독으로 사용할 수는 없다. 어떤 기준(criteria)들은 이 절에서 언급된 예측표준오차(standard error of prediction)로부터 전개된 것이지만 반면에, 다른 것들은 직관적인 개념(intuitive concept)에 기초하고 있어서 무관한 자료(independent data)를 예측해 보는 과정을 통해서 후보 모형을 확인(validation)해야 한다. 훌륭한 분석가는 모형이 과학적으로 사리에 맞기 때문에 생겨났다는 것을 안다. 이상적으로는, 통계학자가 하나의 기준(criterion)에서 다음 기준(criterion)으로 기준을 달리해가면서 찾아나가는 토너먼트(tournament)에서 구조적으로 이치에 맞는 모형이 살아남는다. 종종 선택된 모형이 최고의 논리성(maximum sense)을 보이는 모형은 아니다. 확실히 이론적인 토대(theoretical underpinning)가 결여된 예측모형(prediction model)을 채용(adopt)할 수는 없다 하더라도, 이론적으로 정당한(justifiable) 모형을 관련 자료가 지지하지(support) 못할 수는 있다. 4장에서는 많은 부분을 할애하여 예측을 위한 최선의 모형(best model for prediction)을 선택하는 기준(criteria)에 대한 정보(information)와 예제(example)를 기술하였다.

3.6. 반복측정자료(Data with Repeated Observation)

생물학이나 물리학과 같은 분야에서는 종종 x -level에서의 실험(experiment)을 디자인(design)하는 것이 가능하다. 즉 수준(level)은 조절(control)될 수 있고, 각각의 x -조합(combination)에 대한 반응의 관측값(observation)을 반복 측정할 수 있다. 예를 들어 온도 x_1 과 반응시간 x_2 에 대한 반응률(reaction rate) y 와 연관된 회귀함수(regression function)에 관심이 있는 화학자는 각 실험단계(experimental point)에서 여러 번 관찰(multiple observation)하거나 여러 번 실험(multiple run)하는 방법을 통해 도움을 받을 것이다. 이런 종류의 실험에서 우리는 실험의 재연성(reproducibility) 혹은 실험오차분산(experimental error variance)을 나타내는 원천(source)으로부터 σ^2 의 추정값(estimate)을 얻는다. 그 후에 이런 변동(variation)을 일상적인 잔차변동(residual variation)으로부터 분리하여 모형에 독립적인 변동부분(model-independent portion of variation)을 알아낼 수 있으며, 이로써 모형의 타당성(validity)을 확인할 수 있는 것이다.

적합성결여의 개념(Concept of Lack of Fit)

적합성 결여(lack of fit)라는 용어는 순오차평균제곱(pure error mean square, 반복된 반응 관찰로 얻어진)과 비교하는데 사용되는 변동 요소(component of variation)를 나타내는 적합성결여평균제곱(lack of fit mean square)으로 모형의 적절성(model adequacy)을 점검하는 방식을 의미한다.

i 번째 실험조합(experimental combination)에서의 j 번째 반응을 y_{ij} 라고 해보자. 이때 $i = 1, 2, \dots, m; j = 1, 2, \dots, n_i$ 그리고, $\sum_{i=1}^m n_i = n$ 이다. 달리 말하면, 실험에서 $m \geq p$ 인 회귀변수(regressor variable)의 뚜렷한 조합(combination)이 있고, i 번째 조합에서 n_i 회의 실험이 시행되었다는 것이다. 이제 잔차제곱합(residual sum of squares)의 분할(partition)을 고려해보자.

$$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2 \quad (3.23)$$

여기서 \hat{y}_i 는 회귀변수의 i 번째 조합에서의 반응의 적합값(fitted value)이고, \bar{y}_i 는 i 번째 조합에서의 평균반응(average response)이다. 양(quantity) $\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ 은 이전에 서술한 바와 같이 반복된 관찰에 의한 변동(variation)을 측정하고, 적합모형(fitted model)이 정확(correct)한가 와는 상관없이 순오차평균제곱(pure error mean square)

$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 / \sum_{i=1}^m (n_i - 1)$ 와 비편향추정값(unbiasedly estimate) σ^2 을 측정한다. 양(quantity)

$\sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2$ 은 적합성부족제곱합(the lack of fit sum of squares)이라 하고 $m - p$ 만큼의

자유도를 가진다. 만일 모형이 정확하다면 \bar{y}_i 는 적합값 \hat{y}_i 에서 편향되지 않았다고 판단할 수 있다.

$$E(MS_{LOF}) = E\left[\sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2 / (m - p) \right] = \sigma^2$$

만일 모형이 저설정(underspecified)된다면, 즉 $E(y_i)$ 의 항에서 검정력(power)과 x_1, x_2, \dots 에서의 곱을 포함한다면 (MS_{LOF})는 σ^2 를 과대추정(overestimation)할 것이다. 사실 σ^2 가 과대추정된 양(quantity)은 참평균반응(true mean response) $E(y_i)$ 와 분석자가 참반응(true response)으로 추정한 양(quantity) 간의 차이를 측정한다. 예를 들어, $k = 1$ 이고 제시된 모형이

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

일 때 참값은

$$E(y_i) = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 \quad (3.24)$$

그래서

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^m n_i [\beta_{11} x_i^2]^2}{m - 2} \quad (3.25)$$

$E(MS_{LOF})$ 의 일반적인 전개에 관한 자세한 내용은 다음 절에서 다룰 것이다.

그러므로, 적합성결여(lack of fit)를 사용하여 모형의 적절성(model adequacy)을 평가하는 것은 다음과 같이 주어지고,

$$F = \frac{SS_{LOF}/(m-p)}{SS_{\text{Pure-error}}/(n-m)} = \frac{MS_{LOF}}{MS_{\text{Pure-error}}}$$

이것은 $m-p$ 와 $n-m$ 이 각각 분자와 분모의 자유도인 상부꼬리 한쪽꼬리 F -검정(upper-tailed one-tailed F -test)이다. F -통계량의 유의한 값(a significant value of F -statistic)은 모형 내에 위에서 언급한 항(terms of order)의 영향이 감지될 정도로 존재한다는 것을 뜻한다.

예제 3.6 열대어자료

Table 2.4의 열대어자료를 사용한 예제 2.5를 보자. 회귀변수의 한 위치에는 2번 관측된 자료이다. 먼저 자유도 식 (3.23)을 사용하면 순오차제곱합(pure error sum of squares)을 구할 수 있다.

$$SS_{\text{Pure-error}} = 11.000$$

순오차제곱합(pure error sum of squares)의 자유도는 4이다. 잔차제곱합(residual sum of squares)은 앞에서 살펴보았듯이 $SS_{\text{Res}}=15.881$ 이고, 자유도는 9이다. 따라서, 변동의 분할, 자유도와 F 검정값은 아래 그래프와 같다.

적합결여검정의 변동분해				
Source	SS	df	MS	F
Regression	314.119	1	314.119	
Lack of fit	4.881	5	0.976	0.355
Pure error	11.000	4	2.750	

위의 자료에서 F -통계량이 0.355이므로 제시된 회귀모형에 대한 적합결여는 유의하지 않다고 추론할 수 있다. 따라서, 원점을 지나는 단순회귀모형은 회귀변수와 반응변수의 관계를 충분히 설명하고 있다고 볼 수 있다.

예제에 사용된 R-code는 다음과 같다.

```
data<-read.table("d:/data/ex2_5.R",header=TRUE)
g<-lm(Y~X,data)
ga<-lm(Y~factor(X),data)
ano<-anova(g,ga)
```

부가적인 전개와 언급(Additional Development and Comments)

이 절에서는 행렬 전개(matrix development)를 사용하여 적합성 결여(lack of fit)에 대한

연구를 해보자. 4 장에서는 모형 오설정(model misspecification)에 따른 결과에 대해 공부할 것인데 이 때 여기에서 사용한 접근 및 결과가 많이 사용될 것이다. 만일 독자가 적절한 모형 선택에 대해 심각한 판단을 내려야 한다면 반드시 모형의 저설정(underspecification) 혹은 과대설정(overspecification)의 오류가 어떤 결과를 초래하는지 명확하게 이해하고 있어야 한다.

넓게 생각해보면 적합성 결여(lack of fit)라는 개념은 저설정(underspecification)을 다루는 것이고, 사실 적합모형이 적절한지에 대한 정보를 제공하는 가설 검정의 방식을 분석자에게 제공한다. 이것을 좀 더 발전시키면 실험자는 다음과 같은 모형을 설정할 수 있다.

$$y = X_1\beta_1 + \varepsilon^* \quad (3.26)$$

X_1 은 $n \times p$ 그리고, β_1 은 p 요소(element)를 가진다. 참모형(true model)은 다음과 같으며,

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

여기에서 X_2 는 $n \times (m - p)$ 이고 추정가능한 ‘곱(product)과 거듭제곱(power)으로 되어 있는 항(terms)’을 포함한다. 표본크기(sample size)와 자료(data)는 이용 가능하다. 결과적으로 $E(\varepsilon^*) \neq 0$. 만일 $X_2\beta_2$ 를 무시하면, β_1 의 최소제곱 추정값(least squares estimate)은 $b_1 = (X'_1 X_1)^{-1} X'_1 y$. 잔차제곱합은 $SS_{\text{Res}} = (y - X_1 b_1)'(y - X_1 b_1)$. 무상관 오차(uncorrelated error)와 공통오차분산(common error variance)의 가정(assumption) 하에서 (부록 B.2 참조),

$$E(SS_{\text{Res}}) = \sigma^2(n - p) + \beta'_2 [X'_2 X_2 - X'_2 X_1 (X'_1 X_1)^{-1} X'_1 X_2] \beta_2 \quad (3.27)$$

$\beta'_2 [X'_2 X_2 - X'_2 X_1 (X'_1 X_1)^{-1} X'_1 X_2] \beta_2$ 항은 β_2 의 이차형태(quadratic form)이고, 무시되었던 모수(ignored parameter)이다. 순오차제곱합(pure error sum of squares)의 기대값(expectation)은

$$E(SS_{\text{pure-error}}) = \sigma^2 \left[\sum_{i=1}^m (n_i - 1) \right]$$

$\sum n_i = n$, 총 표본크기(total sample size)이다. 그러므로 $E(MS_{\text{LOF}})$ 는 다음과 같다.

$$\begin{aligned}
E(MS_{LOF}) &= \frac{E(SS_{\text{Res}}) - E(SS_{\text{pure-error}})}{m-p} \\
&= \frac{\sigma^2(m-p) + \beta_2' [X_2' X_2 - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2] \beta_2}{m-p} \\
&= \sigma^2 + \frac{\beta_2' [X_2' X_2 - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2] \beta_2}{m-p}
\end{aligned} \tag{3.28}$$

그러므로 적합성결여(lack of fit)에 대한 F -검정은 뚜렷한 양의 값을 갖는 이차 식의 기여도(contribution of the positive definite quadratic form)를 탐지하기 위해 디자인되었다. (부록 A.2 참조)

$$\beta_2' [X_2' X_2 - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2] \beta_2$$

반면 β_2 는 적합모형(fitted model)으로 추정할 수 있는 범위를 넘어서는 회귀모수(regression parameter)의 벡터이다. 결과적으로 적합성 결여 검정(lack of fit test)의 민감도(sensitivity)는 이 이차식(quadratic form)의 중요도와 β_2 에 포함된 계수(coefficients) 및 요소(elements) X_1, X_2 의 크기와 직결된다.

3.7. 다중회귀에서 동시추정(Simultaneous Inference in Multiple Regressions)

2.10에서부터 단순선형회귀에서 동시추론(simultaneous inference)을 고려해 볼 필요를 느꼈다. 기울기와 절편에 대한 동시검정(simultaneous tests)뿐만 아니라 이 두 모수(parameter)에 대한 동시신뢰구간(simultaneous confidence region)도 전개되었다. 더불어 자료(data)의 위치(location)에서 평균 반응에 대한 동시 신뢰구간(simultaneous confidence intervals)도 고려되었다. 2 장과 이 장에서 얻어진 결과물(machinery)로 다중회귀까지 연장하는 것은 꽤 간단한 일이다. 먼저 각각의 모수(parameter)에 대한 추론(inference)을 다루는 것으로부터 시작해보자.

계수의 동시검정(Simultaneous Tests on Coefficients)

다음 가설을 가정해 보자.

$$\begin{aligned} H_0: \beta_0 &= \beta_{0,0} \\ \beta_1 &= \beta_{1,0} \\ &\vdots \\ \beta_k &= \beta_{k,0} \end{aligned}$$

식 (3.1)의 다중선형회귀모형(multiple linear regression model)의 구조는 3.4절의 일반선형가설을 채용하고 있을 뿐이다. 가설은 (행렬 형식으로)

$$\begin{aligned} H_0: \beta &= \beta_0 \\ H_1: \beta &\neq \beta_0 \end{aligned}$$

식 (3.15)의 특수한 경우의 가설이고 $C = I_p$ 그리고 $d = 0$. 식 (3.16)의 검정 통계량(test statistic)을 이용하면

$$F = \frac{(b - \beta_0)'(XX)(b - \beta_0)}{ps^2} \quad (3.29)$$

통상적인 정규이론(normal theory)을 가정할 때 귀무가설 H_0 하의 $F_{p,n-p}$ 이다. 그래서, 일반적인 상부꼬리 F -검정(upper tail F -test)이 사용되었다. 식 (3.12)의 가설을 이용하여 계수의 부분집합을 검정하는 것 또한 일반선형가설에서의 특별한 경우이다. 사실 식 (3.13)에서의 검정통계량(test statistic)은 $C = [I_p : 0]$ 일 때 (3.16)의 특별한 경우이다.

계수의 동시신뢰구역(Simultaneous Confidence Region on Coefficients)

β 의 요소(element)에 대한 동시검정(simultaneous testing)에 사용되는 방법론은 모수의 동시신뢰구역(simultaneous confidence region)을 계산할 수 있도록 해준다. 이것은 예제 (2.7)에서 묘사된 β_0 와 β_1 에 대한 타원형의 신뢰구역(elliptical confidence region)을 확장(extension)한 것이다. (3.29)의 결과로부터

$$\Pr \left[\frac{(b - \beta)'(X'X)(b - \beta)}{ps^2} \leq F_{\hat{\sigma}, p, n-p} \right] = 1 - \alpha$$

결과적으로 부등호를 충족하는 어떤 조합의 β 든지

$$\boxed{\frac{(b - \beta)'(X'X)(b - \beta)}{ps^2} \leq F_{\alpha, p, n-p}}$$

100(1 - α)% 신뢰구역(confidence region) 내에 들어간다.

회귀계수를 위한 본페로니신뢰구역(Bonferroni Confidence Region for Regression Coefficients)

2.10절에서 우리는 단순선형회귀의 신뢰한계(confidence limit)에 대한 본페로니 접근법(Bonferroni approach)에 대하여 논의하였다. 본페로니 신뢰한계(Bonferroni confidence limits)는 결합신뢰계수(joint confidence coefficients)의 경계(bound)를 조절(control)하는 방식(way)으로 직사각형 신뢰구역(rectangular confidence region)을 형성한다. 단순선형회귀의 경우와 동일하게 본페로니 신뢰 구간(Bonferroni confidence intervals)은 다음과 같은 형식이다.

$$b_j \pm t \cdot (\text{standard error of } b_j)$$

여기에서 t 상수는 t 분포의 적절한 백분위수지점(percentile point)을 뜻한다. 단순선형의 경우 결합신뢰구역의 경계(bound)는 $t_{\alpha/4}$ 백분위수지점(percentile point)을 이용하여 관리된다(적절한 자유도를 가지고). 간단히 p 가 결합신뢰구역(joint confidence region)에 포함된 모수의 개수라 하면 일반화는 $t_{\alpha/2p}$ 이다. 그러므로, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ 상에서 100(1 - α)% 신뢰구역(confidence region)은

$$b_0 \pm t_{\alpha/2p, n-p} \cdot (\text{standard error of } b_0)$$

$$b_1 \pm t_{\alpha/2, n-p} \cdot (\text{standard error of } b_1)$$

$$\vdots$$

$$b_k \pm t_{\alpha/2, n-p} \cdot (\text{standard error of } b_k)$$

$p = k + 1$ 이다. 물론 여기서 정확한 신뢰계수는 적어도 $1-\alpha$ 이다. 계수의 표준오차(the standard errors of the coefficients)는 식 (3.14)에 기술되어 있다. 어떤 경우 단순선형회귀에서와 같이 정확한 방법을 도표로(graphically) 사용하기 어렵다는 이유만으로 정확한 신뢰구역(confidence region)보다 본페로니 신뢰경계(Bonferroni confidence bound)가 더 도움이 되기도 한다. 결과적으로 본페로니 경계(Bonferroni bounds)가 다소 보수적(conservative)이지만 사용하기는 편리하다.

평균반응의 공동신뢰경계(Simultaneous Confidence Bounds on Mean Response)

2.10절에서 다중추정(multiple estimation) 혹은 예측(prediction)이 동시에 일어나는 단순선형회귀에서 평균반응에 대한 신뢰구간(confidence interval)을 결정하는 문제에 대하여 언급하였다. 결합신뢰계수(joint confidence coefficients)의 경계를 포함할 만큼 충분히 보존적(conservative)인 신뢰한계(confidence limits)를 다루는 것부터 시작해보자. 우리는 본페로니 접근법과 Working-Hotelling 접근법을 사용하였다. 독자들이 이미 3.5 절에서 접해본 바와 같이 본페로니 접근법은 다중선형회귀의 경우에도 멋지게 적용할 수 있음을 짐작할 수 있을 것이다. r 위치(location)에 추정하고자 하는 평균반응이 있다고 가정하자. 즉, r 개의 회귀변수조합(combination of the regressor variable)이 있다. 이들 조합을 x'_1, x'_2, \dots, x'_r 이라 하자. 모형 내 절편의 경우 x_j 의 첫번째 요소(initial element)는 1이 된다. 물론 이들은 자료 포인트(data point)로 표시되는 위치가 필요 없다. 단순선형회귀의 경우와 마찬가지로 (예제 2.10을 참조), 신뢰구간(confidence interval)의 동시적 특성으로 인하여 t 백분위수지점(t percentile point)은 증가하게 된다. 그러므로, r 평균반응(mean response)을 추정할 때 적절한 신뢰구간(confidence interval)의 형태는 다음과 같다.

$$\hat{y}(x_j) \pm t_{\alpha/2r, n-p} s_{\hat{y}(x_j)}$$

$\hat{y}(x_j)$ 는 위치 x_j 에서의 추정된 반응(estimated response)이고, $s_{\hat{y}(x_j)}$ 는 위치 x_j 에서의 예측표준오차(standard error of prediction)이다. 이제 식 (3.20)을 이용한다면 본페로니 접근법은 평균 반응값(mean response value)을 포함하는(cover) 다음의 구간(interval)을 적어도 $100(1-\alpha)\%$ 만큼은 확신할 수 있도록 해준다.

$$\hat{y}(x_j) \pm t_{\alpha/2r,n-p} \sqrt{x'_j (X'X)^{-1} x_j} \quad (j = 1, 2, \dots, r)$$

Working-Hotelling 접근법도 동시예측값(simultaneous predicted values)의 신뢰한계(confidence limits) 즉, 평균반응(mean response)의 접합신뢰한계(joint confidence limit)를 구하기 위해 사용될 수 있다. 물론 모형에서 모수(parameters)의 개수가 충분히 크고 r 값이 상대적으로 작다면 본래로니 접근법을 사용하는 것이 훨씬 효과적이다. Working-Hotelling 접근법을 사용하여 r 위치에서 접합신뢰구간(joint confidence intervals)을 구하는 방법은 다음과 같다.

$$\hat{y}(x_j) \pm \sqrt{pF_{a,p,n-p}} \sqrt{x'_j (X'X)^{-1} x_j} \quad (j = 1, 2, \dots, r)$$

예제 3.7 헬스클럽자료를 이용한 동시추정

Table 3.1의 자료를 이용하여 동시추정문제에 대해서 알아보고 평균반응의 일반적인 신뢰구간과 결합신뢰구간을 비교해보도록 하겠다. 예제 3.5에서 사용한 모형과 동일한 모형을 사용하면 일반적인 신뢰구간은 table 3.6에 제시된 평균반응 신뢰구간과 동일하다. 만약 연구자가 26~30으로 관찰된 조사자의 동시추정에 관심 있다고 가정하다. 결합신뢰구간을 구하는 데 필요한 표준오차의 예측값은 table 3.6에 제시되어있다. $r=5$ 일 때 계산된 최소 95%를 충족하는 신뢰한계와 접합신뢰한계의 적절한 t 백분위수지점은(표 C.2 참조) 다음과 같다.

$$t_{.025/5,26} = t_{.005,26} \approx 2.779$$

따라서, 26~30번째 조사자에 대한 95% 결합신뢰한계는

Location		
26	336.8606 ± 2.779 (8.33342):	[315.2365; 358.4848]
27	347.2676 ± 2.779 (11.7157):	[316.8668; 377.6683]
28	429.4002 ± 2.779 (17.84789):	[383.0872; 475.7132]
29	278.7412 ± 2.779 (11.58871):	[248.6700; 308.8124]
30	393.2235 ± 2.779 (12.59571):	[360.5392; 425.9077]

결합신뢰한계가 table 3.6에 예시된 95% 신뢰구간보다 얼마나 보수적인지 주의 깊게 살펴야 한다.

예제에 사용된 R-code는 다음과 같다.

```
data<-read.table("d:/data/ex3_1.R",header=TRUE)
attach(data)
g<-lm(y~x1+x2+x3+x4)
pre<-predict(g,se=TRUE,interval="confidence")
jol<-pre$fit[26:30,1]-abs(qt(0.005,260))*pre$se.fit[26:30]
jou<-pre$fit[26:30,1]+abs(qt(0.005,260))*pre$se.fit[26:30]
joint_in<-cbind(jol,jou)
```

3.8. 다중회귀자료에서 다중공선성(Multicollinearity in Multiple Regression)

Data)

회귀 분석(regression analysis)은 여러가지 이유 때문에 분석도구(analytic tool)로 이용 된다. 이전에 함수 관계(functional relationship)의 추정값(estimate)을 전개할 필요성을 언급하였는데, 이는 예측(prediction)에 이용할 수 있기 때문이다. 다른 한편으로, 회귀의 동기(motivation)는 특정한 회귀변수(regressor variables)들(즉, 회귀 계수의 추정값)에 대한 반응의 변화 비율(rate of change of response)을 추정하는 것일 수도 있다. 그러나, 회귀를 구하는 목적에도 불구하고 과학자가 모형화하려고 하는 표기(notation)를 자료(data)가 제공하지 못할 수 있어서 여러 문제가 발생할 수 있다. 실험상의 큰 오차 혹은 자료자체의 잡음(noise)은 좋은 적합(good fit)을 이루는 것을 방해할 뿐더러 질적통계적 추론(quality statistical inference)의 가능성(chance)도 허용하지 않는다. 우리는 모형오차분산(model error variance) σ^2 이 과도하게 큰 실험상황(experimental situation)의 결과로서 이러한 어려움을 구분(categorize)하고 있다.

또 다른 어려움은, 과거 20년 동안 상당한 연구가 이루어져 온 것인데, 다중공선성(multicollinearity)의 문제이다. **다중공선성(multicollinearity)**이란 이름은 많은 통계적 도구의 잠재적(potential)인 사용자들이 그것이 무엇인지를 실제로 이해하기 전부터 그 존재가 알려져 있었다. 그 용어 자체로 정의를 내리는데, 다중(multi)은 많음을 의미하고 공선성(collinear)은 선형종속(linear dependencies)을 의미한다. 다중공선성(multicollinearity)은 회귀변수들(regressor variables)에 있어서의 어떤 상태(condition)를 나타내는 것이다. 이번 장에서 우리는 다중공선성(multicollinearity)에 대해 완벽하게 다룬다기 보다는 단지 그것을 소개하고 독자들이 회귀결과(regression results)에 대한 그것의 영향을 이해하도록 할 것이다. 심각한 다중공선성(serious multicollinearity)의 발견방법(method of detection), 형태진단(formal diagnosis), 그것을 다루기 위한 과정(procedure of combating)은 8장에서 다를 것이다. 예를 들면, 보통의 최소제곱과정(least squares procedure)외의 다른 방법을 이용해 하하는 추정(estimation)의 편향된 형태(biased form)는 8장에서 상세하게 논할 것이다. 그러나 다중회귀(multiple regression)를 처음 접하는 독자를 위해 다중공선성(multicollinearity)이 얼마나 정확한 통계적 추론(precise statistical inference)을 불가능하게 하는지, 즉, 회귀 계수(regression coefficients)의 추정 혹은 아마도 예측(prediction)이 다중공선성 때문에 얼마나 악화되는지에 대해 집중하는 것은 중요할 것이다. 능선회귀(ridge regression) 혹은 주성분회귀(principal components regression)와 같이 그 문제에 대항하는 방법들은 다소 논란의 여지가 있고, 실제로 여러 통계 연구자들에 의해 종종 비판을 받는다. 그럼에도 불구하고, 독자들이 어떤 유용한 방법도 사용하지 못한다 하더라도, 다중공선성(multicollinearity)의 원천(source)과 효과에 대해 명확히 이해하는 것은 여전히 중요하다.

(3.2) 일반선형모형(general linear model)을 고려해 보라. 다중공선성(multicollinearity)의 원천(source)은 X의 열(columns)의 성질(nature)에 의해 초래된다. 우리가 X를 다음과 같이 분할한다고 가정해보면,

$$X = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_k \end{bmatrix} \quad (3.30)$$

각 열(column)은 특정회귀변수(particular regression variable)를 위한 측정치(measurements)로 나타난다. 모형에 상수항이 있다고 가정하고 회귀변수의 한 열로서 첫번째 열을 살펴보라. 또한 이해를 쉽게 하기 위해 회귀변수(regressor variables)가 중심화(centered)되었고, 기준화(scaled)되어 있다고 가정해 보자. 즉, 만약 자연단위(natural units)에서 회귀변수(regressor variable) x_i 에 대한 j 번째 측정치(measurement)가 x_{ij} 라면, x_{ij} 와 $x_{ij} - \bar{x}_i$ 로부터 $\bar{x}_i = \sum_{j=1}^n x_{ij} / n$ 을 빼고 S_i 즉 $S_i = \sqrt{\sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}$ 로 나누어 진다. 이 중심화와 기준화(centering and scaling)는

상관행렬(correlation matrix)로 $X'X$ 가 된다. 즉, 우리는 다음과 같은 대칭행렬(symmetric matrix)을 쓸 수 있는데

$$X'X = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & r_{12} & \dots & r_{1k} \\ 0 & r_{12} & 1 & \dots & r_{2k} \\ \vdots & & & & \vdots \\ \vdots & & & & r_{k-1,k} \\ 0 & & & & 1 \end{bmatrix}$$

여기에서 r_{ij} 은 회귀변수(regressor) x_i 와 x_j 사이의 단순상관계수(simple correlation coefficient)이다. 우리가 여기서 언급해야 하는 것은 X 의 첫 번째 열(column)에서 통상적인 하나의 열(column)이 제거될 수 있고 결과적으로 다음과 같은 형태를 취하는데

$$y = \beta_0 1 + X^* \beta \quad (3.31)$$

여기에서, 이런 형태에서 절편(intercept)과는 별도로 $\beta' = [\beta_1, \beta_2, \dots, \beta_k]$ 는 계수의 벡터(vector of coefficients)이고 X^* 는 중심화되고 기준화된 회귀변수(entered and scaled regressor variables)의 $n \times k$ 행렬(matrix)이다. 표기(Notation) 1은 하나의 n -vector를 표시하기 위해 사용된다.

따라서 $X^* X^*$ 는 $k \times k$ 상관행렬(correlation matrix)이다.

$$\mathbf{X}^* \mathbf{X}^* = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ & 1 & r_{23} & \cdots & r_{2k} \\ & & & \vdots & \\ & & & & r_{k-1,k} \\ & & & & 1 \end{bmatrix}$$

흥미롭게도 상관(correlation)이란 용어가 여기서는 약간 오칭(misnomer)이다. 우리는 회귀변수(regressors)가 확률변수(random variables)가 아니라고 생각하고 있다. 그럼에도 불구하고, 단순상관계수(simple correlation coefficient) r_{ij} 은 자료에 있어서 x_i 와 x_j 사이의 선형의존(linear dependency)을 나타낸다.

다중공선성이란? (What is Multicollinearity?)

일련의 회귀자료(regression data)에서 매우 바람직한 상태(condition)는 “상호간에 동시에 움직이는 것(moving with each other)”이 없는 회귀변수(regressor)를 가지는 것이다. 선형의존성에 가까운 경우(near linear dependencies)는 반응에 대한 각 회귀변수(each regressor)의 효과(impact)를 선별해(sort out)내기가 더 어렵다. 간단한 예로 이것을 설명할 수 있다.

두 개의 회귀변수(regressor variables)를 양쪽 다 포함하는 두 개의 약간 다른 경우(scenarios)를 생각해 보자. 회귀변수들(regressors) 사이에서 선형의존성에 가까운(near linear dependencies) 것이 분석가가 회귀계수(regression coefficients)를 추정하려고 할 때 이를 얼마나 심하게 방해하는지를 보여주는 것이 우리의 의도이다. 첫 번째로, $n=8$ 인 다음의 회귀자료(regressor data)를 가지고 있다고 가정해 보자.

x_1	10	10	10	10	15	15	15	15
x_2	10	10	15	15	10	10	15	15

다음은 동일한 표본크기($n=8$)와 동일한 회귀변수의 범위(ranges)를 갖는 또다른 자료를 고려해보자.

x_1	10.0	11.0	11.9	12.7	13.3	14.2	14.7	15.0
x_2	10.0	11.4	12.2	12.5	13.2	13.9	14.4	15.0

만약 자료에 $k = 2$ 선형회귀(linear regression)를 적용시킨다면, 자료 세트(data set)에 대한 회귀계수의 분산을 비교해보는 것은 흥미로울 것이다. 상관계수(correlation coefficients)를 계산하면 첫 번째 자료세트의 경우 $r_{12} = 0$ 이고 두 번째 자료세트의 경우는 $r_{12} = 0.99215$ 임을 알수 있다. 이와 같이 우리는 첫번째 자료세트에 대해 다음을 얻을 수 있고

$$X^* X^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (X^* X^*)^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

두 번째 자료세트에서는 다음을 얻을 수 있다.

$$(X^* X^*) = \begin{bmatrix} 1 & 0.99215 \\ 0.99215 & 1 \end{bmatrix} \quad (X^* X^*)^{-1} = \begin{bmatrix} 63.94 & -63.44 \\ -63.44 & 63.94 \end{bmatrix}$$

이와 같이 우리는 두 개의 자료세트에서 모수 추정값의 분산을 구할 수 있다. (식 (3.31)의 중심화되고 기준화된 모형(centered and scaled model)을 이용하였다)

Data set 1	Data set 2
$\frac{\text{Var } b_1}{\sigma^2} = \frac{\text{Var } b_2}{\sigma^2} = 1.0$	$\frac{\text{Var } b_1}{\sigma^2} = \frac{\text{Var } b_2}{\sigma^2} = 63.94$

공선성(collinearity)으로 인해 계수의 분산이 얼마나 과장(inflated)되었는지를 알 수 있을 것이다. 이러한 분산의 과장(inflation of variance)은 공선성(collinearity)으로 인해 손상(damage)이 있음을 강조(underscore)하는 쉽게 이해할 수 있는(accessible) 정보(information)이다. 첫 번째 자료세트에서 회귀변수들(regressors)이 직교(orthogonal) 상태에 있다고 일컫는데, 이런 직교(orthogonality) 상태(condition)는 회귀변수들을 조절(control)할 수 있는 능력(capability)이 있을 때의 상당히 바람직한 실험 디자인(experimental design)의 속성(property)이다. 두 번째 자료세트의 경우 회귀변수는 명백히 함께 이동(moving)하며, 이러한 선형에 가까운 의존성(near linear dependency)은 추정의 질(quality of estimation)에 있어 손해를 끼침(take its toll)은 명백하다. 독자들은 첫 번째 자료세트를 고려해야 하고, 그것의 직교회귀변수(orthogonal regressors)를 이상적인 경우(ideal case)로 인식해야 한다.

다중공선성(multicollinearity)은 X^* 의 열(column)인 x_j^* 간에 선형에 가까운 의존성(near linear dependencies)이 있을 때 생긴다. 즉, 상수(constant)의 세트(set)(모두 0이 아닌)가 다음과 같이 있다.

$$\sum_{j=1}^k c_j x_j^* \cong 0$$

\cong 라고 쓰는 이유는 만일 우변이 동일하게(identically) 0이라면 선형의존성(linear dependencies)이 정확(exact)하고 따라서 $(X^* X)^{-1}$ 가 존재하지 않기 때문이다. 물론 가까운 의존성(near

dependencies)은 실제 자료에서 존재할 수 있고, 흔히 다중공선성(multicollinearity)이라고 부르는 효과를 나타낼지도 모른다. 회귀계수(regression coefficient)는 회귀변수(regressor variable)에 대한 반응(response)의 변화율(a rate of change) 혹은 편도함수(partial derivative)임을 명심해야 한다. 회귀변수들이 동시에 움직이는 방식으로 x-자료(x-data)가 존재할 때, 확실한 변화율(rate of change)의 추정값(estimate)을 구하여야 하는 자료구조(data structure)에서는 최소제곱과정(least squares procedure)은 부적절하다.

다중공선성을 설명하기 위한 고유값과 고유벡터의 이용(Use of Eigenvalues and Eigenvectors to Explain Multicollinearity)

X^*X^* 행렬(상관형태(correlation form))을 고려한다고 가정해 보자. 우리는 직교행렬(orthogonal matrix)이 존재한다는 것을 알고 있다(see Graybill(1976))

$$V = [v_1, v_2, \dots, v_k]$$

이 경우에

$$V'(X^*X^*)V = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k) \quad (3.33)$$

λ_i 는 상관행렬(correlation matrix)의 고유값(eigenvalues)이다. 식 (3.33)에서 주어진 연산(operation)은 X^*X^* 의 고유분해(eigenvalue decomposition)라 불린다. V 의 열(column)들은 (X^*X^*) 의 고유값과 관련된 정규화된 고유벡터(normalized eigenvectors)이다. 여기에서의 목적을 위해, 벡터 v_j 의 i 번째 요소를 v_{ij} 로 표시할 필요가 있다. 만약 다중공선성(multicollinearity)이 존재한다면, 적어도 하나의 $\lambda_i \approx 0$ 이다. (부록 A.3. 참조) 이와 같이 우리는 j 의 적어도 한 값에 대해서는 다음과 같이 적을 수 있고

$$v_j'(X^*X^*)v_j \approx 0$$

적어도 하나의 고유벡터(eigenvector) V_j 를 의미하는 것으로 다음과 같이 쓸 수 있다.

$$\sum_{l=1}^k v_{lj} x_l^* \approx 0$$

상관행렬(correlation matrix)의 작은(small) 고유값의 개수는 (3.32)의 정의를 따르면 다중공선성(multicollinearities)의 개수와 관련이 있고, (3.32)에서 가중치(weight) c_j 는 고유벡터와 관련된 개별요소(individual elements)이다. 이러한 전개는 단순히 일시적인 흥미 이상의 것을 위해 이

번에 나타내었다. 그것은 8장의 전개과정에서 다중공선성(multicollinearity)을 진단하는데 중요한 도움이 될 것이다.

양 극단: 공선성과 직교성(Two Extremes: Collinearity and Orthogonality)

변수(variable)들이 서로 직교(orthogonal)인 상황을 생각해 보면 즉, 상관행렬 X^*X^* 가 항등행렬(identity matrix)인 것이다. 이런 이상적인 경우에서, σ^2 과 별개로, $Var(b_i)=1.0$ ($i = 1, 2, \dots, k$)이다(여기서, b_i 는 i 번째 중심화되고 기준화된 회귀변수의 계수로 간주한다). 이제 $k=2$ 인 다음과 같은 또 다른 가설적 경우(hypothetical case)를 고려해보자.

$$X^{*\prime}X^* = \begin{bmatrix} 1.0 & 0.975 \\ 0.975 & 1.0 \end{bmatrix}$$

그리고,

$$(X^{*\prime}X^*)^{-1} = \begin{bmatrix} 20.2532 & -19.747 \\ -19.747 & 20.2532 \end{bmatrix}$$

이와 같이, 다중공선성(multicollinearity)은 기준화되고 중심화된 모형(scaled and centered model)에서 회귀계수(regression coefficients)의 분산(variances)을 1.0에서 20.2532까지 팽창(inflation)시킨다. 두 개의 회귀변수(regressor variables)가 직교(orthogonal)일 때의 이상적인 경우($X^*X^*=I$)보다 무려 20배 증가시킨다. 우리는 이것을 변수에 있어서의 나쁜 요인(ill-conditioning 즉, 다중공선성)이 20.2532의 분산증폭요인(variance inflation factors)을 초래하였다고 말한다.

i 번째 회귀계수(regression coefficient)에 대한 분산증폭요인(variance inflation factor, VIF)은 다음과 같은 식으로 쓰여질 수 있는데,

$$VIF = \frac{1}{1 - R_i^2} \quad (3.34)$$

여기서 R_i^2 는 다른 회귀변수(regressor variables) x_j ($i \neq j$)에 대해 변수(variable) x_i 를 회귀(regress)할 때 산출되는 회귀의 다중결정계수(coefficient of multiple determinant)이다. 이와 같이 이러한 인위적인 회귀(artificial regression)에 있어서 다중상관(multiple correlation)이 높을수록, 계수(coefficient) b_i 의 추정값(estimate)의 정밀도(precision)는 낮아진다.

계수에 대한 다중공선성(multicollinearity)의 영향을 설명하는 두 번째 접근 방법은 다음에 대한 고찰이다.

$$E(b - \beta)'(b - \beta) \quad (3.35)$$

여기서 b 는 식(3.31)의 모형에서 β 의 최소제곱추정값(least squares estimates)의 벡터이다. 만약 모형이 정확하다면, (3.35)에서의 양(quantity)은 계수(coefficients)의 분산 합인 것은 명확하다. 또한 추정값벡터(estimate vector) b 와 참모수벡터(true parameter vector) β 사이의 기대제곱거리(expected squared distance)이다. 만약 우리가 한번 더 중심화되고 기준화된 회귀변수(centered and scaled regressors)에 대해 생각해 본다면(부록 A.3 참조) 다음과 같다.

$$\frac{E(b - \beta)'(b - \beta)}{\sigma^2} = \text{tr}(X^{*'} X^{*})^{-1} = \sum_{i=1}^k \frac{1}{\lambda_i} \quad (3.36)$$

여기서 $\lambda_1, \lambda_2, \dots, \lambda_k$ 은 식 (3.33)에서 나타난 고유값(eigenvalues)이다. 이와 같이 나쁜 요인(ill-conditioned)이나 거의 비정칙(near singular)인 $X^{*'} X$ 에 대해서, 적어도 고유값의 하나는 작을 것이고, $E(b - \beta)'(b - \beta)$ 는 클 것이다. 이상적인 경우, $\sum_{i=1}^k (1/\lambda_i) = k$ 이다. 식 (3.36)으로부터 이는 명확해지는데,

$$E(b - \beta)'(b - \beta) = E(b'b) - (\beta'\beta)$$

그래서

$$E(b'b) = \beta'\beta + \sigma^2 \sum_{i=1}^k \frac{1}{\lambda_i} \quad (3.37)$$

식 (3.37)은 너무 긴(too long) 회귀계수(regression coefficients)의 벡터를 산출하는 다중공선성(multicollinearity)의 경향(tendency)을 강조하고 있는데, 회귀계수가 길다는 것은 계수가 크기(magnitude)에 있어 너무 큰 경향(tendency)을 가진다는 것을 의미한다. 만약 λ_i 의 어느 것이 작고, 분명히 $b'b$ 가 $\beta'\beta$ 를 향하여 많이 편향(biased)되어 있다면 계수(coefficient)들이 클 것으로 기대될 것이다. b 그 자체가 비편향(unbiased)이라는 사실에도 불구하고 이것은 진실이다. 예를 들면, 공선상황(collinear situations)이 심할 때에는 0.0005의 값을 취한 고유값(eigenvalues)이 드물지 않을 것이다. 명백하게 이러한 상황 (3.37)으로부터, $\sum_{i=1}^k b_i^2$ 은 매우 편향되어 있고,

그 결과 계수(coefficients)중 몇몇이 크기(magnitude)에서 과대 평가되는 경향이 있다.

설명에서처럼, 3 개의 회귀변수(regressor)의 경우에서 다음과 같은 고유값(eigenvalues)이 있다고 가정해 보면,

$$\lambda_1=2.5 \quad \lambda_2=0.4999 \quad \lambda_3=0.0001$$

그러면 명확하게 심각한 의존성(dependency)이 있다. 자 그럼, $\sum_{i=1}^k \beta_i^2$ 의 추정(estimation)에 대한 효과는 무엇인가? (3.37)로부터

$$E\left(\sum_{i=1}^k b_i^2\right) = \sum_{i=1}^k \beta_i^2 + \sigma^2 \left(\frac{1}{2.5} + \frac{1}{0.4999} + \frac{1}{0.0001} \right)$$

이와 같이 $\sum_{i=1}^k b_i^2$ 는 $\sum_{i=1}^k \beta_i^2$ 를 평균적으로 대략(roughly) $10,000\sigma^2$ 정도까지 과대평가한다. 이 상적인 경우 즉, 공선성(collinearity)이 없을 때, $\sum_{i=1}^k b_i^2$ 의 편향(bias)의 크기(magnitude)는 $3\sigma^2$ 일 것이다.

불안정성과 잘못된 부호(Instability and Wrong Sign)

우리가 비록 여기서 다중공선성(multicollinearity)의 진단(diagnosis)에 대한 포괄적인 논의를 하지는 않을지라도 문제의 잠재적 심각성(potential severity)을 평가하기 위한 특정 양(quantity)을 알아보는 것은 필요하다. 예를 들면, 분산증폭요인(VIF), 상관행렬(correlation matrix), 상관행렬의 고유값(eigenvalues of the correlation matrix)을 살펴보아야 한다. 다중공선성(multicollinearity)의 한가지 효과는 회귀계수(regression coefficients)의 불안정성(instability)인데, 이는 회귀계수를 산출(generate)해내는 특정 자료세트(data set)에 상당부분 의존하고 있다는 것이다. 분석가는 y -관측값(y -observations)을 인위적으로 변화(altering) 혹은 교란(perturbing)시키는 방법을 쓰거나 계수(coefficients)의 상대적인 안정성(relative stability)을 점검해 봄으로써 이 불안정성을 간파(detect)할 수 있다. 또한, 분석가는 다중공선성(multicollinearity)이 심각한 문제라면, 회귀변수(regressor variables)의 여러 세트(set)중 하나를 제거 할 수 있고, 남은 계수들(coefficients)은 양(amount)이 많이 변할 수 있고, 심지어 부호(sign)까지 바뀔 수 있다. 명확하게, 이러한 불안정(instability)은 실험자에게 만족스럽지 않을 것이다. 결과적으로 원래(original)의 자료에 비해 합리성(reasonable)에서 별 차이가 없다면 작은 자료 교란(small data perturbations)이 회귀계수(regression coefficients)에서의 작은 변화들을 유발한다면 그것은 좋은 것일 것이다.

적합의 질과 예측의 영향(Effect on Quality of Fit and Prediction)

회귀변수(regressor variables)에서 다중공선성(multicollinearity)으로 인한 상황은 바람직하지 않은 “악영향(fallout)”을 일으켜 모형의 적합을 전개(spread)하지 못하게 한다. 사실, 잔차제곱 합(residual sum of squares)이나 그것에 기초를 둔 다른 통계량은 여전히 모형과 자료가 허용

(allow)하는 한 매력적인 것임을 최소제곱과정(the procedure of least squares)은 보장해준다. 또 한 독자는 회귀분석(regression analysis)에 있어서 잔차가 매우 작을 수 있더라도 계수의 추정이 잘 되지 않을 상황을 예견해야만 한다. 계수(coefficients)의 악화(deterioration)에 대한 설명은 $X'X$ 의 상황(conditioning of $X'X$)을 밝혀(trace)내는 논의(argument)에 의해 주어져왔다. 그러나, 적합의 관점에서 보면(from the standpoint of fit), 표준가정(standard assumptions)이 행해진 경우(ε_i 의 정규성이 포함된)에서 잔차제곱합(residual sum of squares)은 $\sigma^2 \chi_{n-p}^2$ 분포(distribution)를

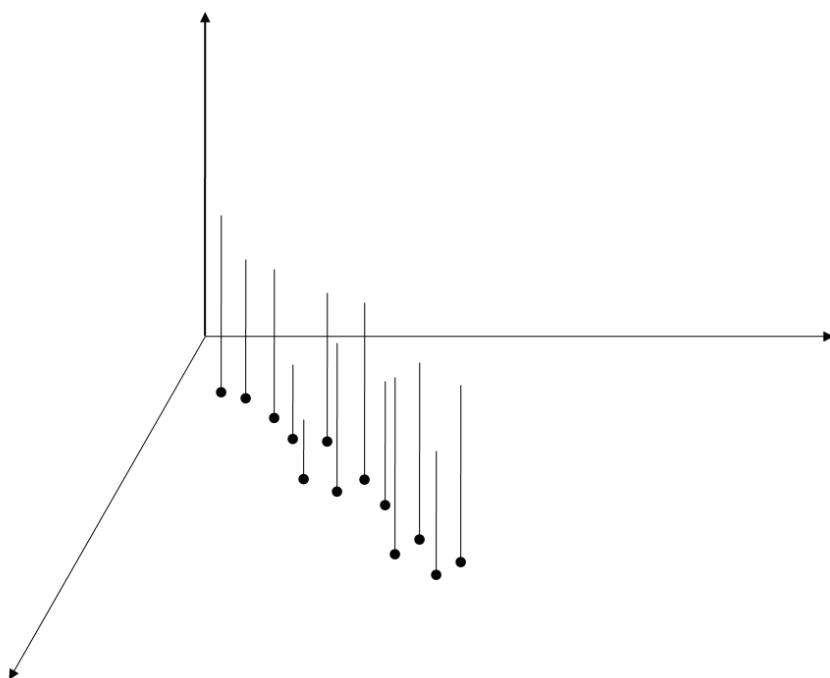
따르고, 이 분포는 $X'X$ 의 조건부(conditioning)에 의해서 영향을 받지 않는다. 결과적으로 적합의 전통적인 분석(traditional analysis)은 잠재적 다중공선성(multicollinearity)의 문제를 나타내지(signal) 않는다. 회귀자료(regression data)에 다중공선성(multicollinearity)이 만연할(infested) 때(회귀변수들간에 서로 다른 대리변수(proxy)를 초래하는) 잔차제곱표면합(residual sum of squares surface)은 지극히 평평하다(flat)는 것을 독자는 마음에 새겨 두어야 한다. 그래서, 잔차제곱합(residual sum of squares)이 매우 크게 변하는 계수공간(coefficient space)에서 큰 구역(large region)이 있고, 계수의 최소제곱값(least squares value)이 항상 최선의 선택은 아닐 것이다. 공선성(collinearity)은 반드시 진단 되어야 하며, 진단 과정(procedure)들은 8장에서 논의 할 것이다.

만약에 적합(fit)이 심각한 다중공선성(multicollinearity)에 의해 영향 받지 않는다는 것을 확신한다면 회귀 변수들(regressor variables)의 자료 조합(combination) 또는 그 근방에서 반응의 예측(prediction of response)이 상대적으로 덜 영향을 받을 것이라는 사실을 쉽게 받아들일 수 있다. 자료에서 기술된 다중공선성(multicollinearity)에 반하지(counter) 않는 x 조합의 자료범위(data range) 내의 지점(points)에서는 예측(prediction)은 여전히 좋을 것이다. 그러나, 자료 내에서의 관계(relationship)와 일치하지 않는 조합들(combinations)에서의 예측, 혹은 자료 범위 밖에서 외삽(extrapolation)하여 나타낸 지점(point)에서의 예측은 다중공선성(multicollinearity)에 의해 나쁜 영향을 받을 수 있을 것이다. 이 문제에 대한 보다 심도 깊은 내용은 4장과 8장에서 다룰 것이다. 이 시점에서, 독자들은 왜 다중공선성(multicollinearity)을 주의 깊게 진단하여야만 하고 때때로 편향 추정(biased estimation) 기법을 통해 이를 제거하기 위해 노력해야 하는지를 분명히 간파하여야 할 것이다. 독자들은 높은 공선자료(highly collinear data)의 보통최소제곱분석(ordinary least squares analysis)이 관련 정보를 아파 숨길 수 있다는 것을 이해할 것이다.

하나의 간단한 그림으로 나타낸 실례는 fig 3.3에 있는 소위 말하는 “(말뚝울타리(picket fence)” 표시이다. 이러한 실례는 명백한 공선성(collinearity)을 가진 두 회귀변수(regressor)를 보여준다. 말뚝에 있어서 울타리의 높이는 y -관측값(observations)을 나타낸다. 회귀모형(regression model)을 만드는 일은 말뚝울타리의 꼭대기에서 날개(plane)의 균형을 잡는 것에 비견될 수 있다. 만약 울타리 하나의 높이에서 약간의 변화가 나타난다면 그 날개는 매우 불안정해 지거나 매우 큰 변화를 초래하게 될 것이 명확하다. 이러한 불안정성은 말뚝울타리에 대해 수직 방향에 있는 예측에 있어서 매우 큰 변화를 초래한다. 마찬가지로, x_1 과 x_2 의

방향의 기울기(slopes)에 있어서도 큰 변화가 발생될 수 있다. 이 그림에서의 요점은 만약 공선성(collinearity)이 심각하다면, 적합값들(fitted values)은 안정적일지라도 y -관측값에 있어서 약간의 변화가 계수(coefficients)들에 있어서 큰 변화를 일으킨다는 것이다. 첨가하여, 만약 공선성(collinearity)에 의해 기술된 경로(path)와는 다른 방향으로 \hat{y} 를 사용할 필요가 있다면, 예측 결과들에 있어서 심각한 불안정성이 있음을 예상 할 수 있다.

FIGURE 3.3 Picket fence illustration of collinearity



예제 3.8 Annual Data on Advertising, Promotions, Sales Expenses, and Sales (Millions of Dollars)

어떤 회사에서 판매량(S)을 고려하고, 광고지출비(A), 판촉지출비(P), 판매비용 (SE), 이전 기간의 광고지출비(A_1), 이전기간의 판촉지출비(P_1) 모두를 통상적인 회귀변수로 하여 이들이 판매량에 주는 효과를 고려하고자 한다. 회귀변수들의 역할을 판단하기위한 목적으로 다중회귀가 시행되었다. 이 자료에 대한 추정회귀식은 아래와 같다.

$$\hat{y} = -14.1948 + 5.361A + 8.372P + 22.521SE + 3.855A_1 + 4.125P_1$$

또한, 추정회귀식의 표준오차의 추정값과 결정계수는 다음과 같다.

$$s = 1.742$$

$$R^2 = 0.9169$$

위의 결과를 볼 때 적합결과가 만족스럽게 나타나고 있다고 할 수 있다. 하지만 이 조사

를 시행한 회사에는 A, P, A_1, P_1의 합이 매 2년의 예산에서 5단위로 적절하게 고정되어야 한다는 규칙이 있었다. 즉, $A_t + P_t + A_{t-1} + P_{t-1} = 5$, 의 관계는 이 예제에서 다중공선성의 원인이 된다. 먼저, 회귀변수들간의 상관관계를 보여주는 상관행렬을 살펴보자.

$$\text{Correlation} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ x_1 & 1.000 & -0.357 & -0.129 & -0.140 & -0.496 \\ x_2 & & 1.000 & 0.063 & -0.316 & -0.293 \\ x_3 & & & 1.000 & -0.166 & 0.208 \\ x_4 & & & & 1.000 & -0.358 \\ x_5 & & & & & 1.000 \end{bmatrix}$$

다중공선성의 계량적인 척도인 분산팽창인자(variance inflation factors)를, 이것은 상관행렬(correlation matrix)의 역수의 대각(diagonal)인데, 구해보면 다음과 같다.

$$\begin{aligned} x_1: & \quad \text{VIF} = 36.94 \\ x_2: & \quad \text{VIF} = 33.47 \\ x_3: & \quad \text{VIF} = 1.08 \\ x_4: & \quad \text{VIF} = 25.92 \\ x_5: & \quad \text{VIF} = 43.52 \end{aligned}$$

여기에서 회귀 계수(regression coefficient) b_3 을 제외한 다른 변수에서 다중공선성이 전혀 없는 상황과 비교하였을 때 분산이 크게 나타나 추정이 좋지 못하다는 사실을 명확하게 알 수 있다. 만약 다중공선성을 감소시키면서 보통의 최소제곱법을 이용하여 추정하고자 한다면 회귀변수(regressor) 중 어느 것이든 하나를 제거하는 것이 가장 간단한 해결방법일 것이다. 하지만 이러한 변수제거가 다중공선성(multicollinearity)에 해결하는 최종적인 방법도 아니다. 여기에서의 의도는 다중공선성(multicollinearity)이 얼마나 회귀 계수의 분산(variances of regression coefficients)에 영향을 미칠 수 있는지를 설명하고 상대적으로 책의 초기에 독자로 하여금 다중공선성을 포함하고 있는 일련의 데이터를 접하도록 하는 것이다. 부가적인 설명과 도해는 8장에서 보여질 것인데, 공선성(collinearity)을 다루는 광범위한 자료들을 포함하고 있다. 게다가, 이 데이터는 4장과 8장에서 그 이상의 분석을 위한 대상(object)이 된다.

다음은 예제에서 사용한 R code이다.

```
ad<-read.table('c:/advertising.txt',header=T)
```

```

attach(ad)
fit<-lm(y~x1+x2+x3+x4+x5,ad)
summary(fit)
vif(fit)
anova(fit)

```

3.9. 품질적합, 품질예측, 모자행렬(Quality Fit, Quality Prediction, and the Hat Matrix)

이전에 이야기 하였던 동기(motivation)의 일부는 하나의 예측식(prediction equation) (즉 regressor variables 사이에 존재하는 기능적 관계에 잘 필적하는 함수)에 이르기 위한 연구자의 기초적인 욕구에서 기인한다. 예측(prediction)이 주요한 관심사일 때 이용 가능한 여러 후보군(candidate)으로부터 모형의 분리(separation)와 순위(ranking)를 고려하는 양(quantities)의 유형(type)에 대해서는 4장에서 자세히 다루게 될 것이다. 다시 강조하는데, 품질적합(quality fit)과 품질예측(quality prediction)이 반드시 일치하는 것은 아니다. 구축해 놓은 모형의 예측능력(prediction capability)에 관심이 있는 연구자라면 어떤 양(quantity)이 회귀분석(regression analysis)에서 강조되었는지를 열심히 살펴보아야 할 것이다.

예측의 표준오차와 모자행렬(Standard Error of Prediction and Hat Matrix)

3.5절의 내용을 떠올려보면, 평균반응(mean response)에 대한 신뢰한계(confidence bound)의 개념에 상당한 관심을 기울여야 함을 기억할 것이다. 이 신뢰한계(confidence bound)는 아래에 주어진 점(point) x'_0 에서의 \hat{y} 의 예측 또는 추정 표준편차(prediction or estimated standard deviation)의 표준오차(std error)에 토대를 두고 있으며,

$$x'_0 = [1, x_{1.0}, x_{2.0}, \dots, x_{k.0}]$$

(여기에서 initial element 인 1은 상수(a constant)로 모형을 수용한다(accommodate).)

x'_0 는 다음에서 얻어진다.

$$s_{\hat{y}(x_0)} = s\sqrt{x'_0(X'X)^{-1}x_0} \quad (3.38)$$

여기까지의 전개과정(development)을 살펴본다면, 예측의 질(quality of prediction)을 나타내는 개념적인 양(conceptual quantity)을 탐색(quest)하게 되면 결국 양(quatity) $h_{00} = x'_0(X'X)^{-1}x_0$ (σ^2 과는 별개로)이 예측분산(prediction variance)이라는 것으로 직접 귀결될 것이다. 새로운 y -자료를 취하고 동일한 표본크기(sample size)와 동일한 회귀변수의 조합(same combinations of the regressor variables)으로 회귀를 계산한다면, 이 양(quantity)은 예측(prediction)의 재현성(reproducibility)을 나타낸다. 이제 (3.38)식을 예측을 평가하는 하나의 적절한 기준(criterion)으로 고려하고 싶은 생각이 든다: 실제로, (3.38)은 single norm이나, 분석가가 이용할 수 있는 단일숫자(single number)를 제공하지 못한다는 점을 제외하고는 적절한 기준이 되기에 충분하다. 명백히 예측인자(predictor)로서 $\hat{y}(x_0)$ 의 질(quality)은 x -공간에서의 x_0 의 위치에 의존하고, h_{00} 는 회귀변수 공간(regressor space)에 따라 상당히 다를 수 있다. 적합된 모형(fitted model)이 잘 예측하는지 아닌지에 대한 의문에 직면하게 될 때 자료 분석가의 확실한 반박(retort)은 “어디에서(where)”라는 질문이다. 분석가가 어디에서 예측이 완성되었는지에 대한 통찰력을 가진다면 예측의 표준 오차(standard error of prediction)는 유용한 기준(criterion)이 될 것이다.

4장에서 예측을 더 잘 평가하는 것에 대해 다룰 것이고, 한 숫자로 예측 능력(prediction capability)을 요약하기 위한 시도로 single norms을 제시할 것이다. 독자들은 가능한 한 예측의 표준오차(standard error of prediction)를 모형이 균일한 질(uniform quality)로 반응을 예측하지 못한다는 사실을 설명해주는 결정 인자(decision maker)로 적용해야 한다. 예제 3.5의 헬스클럽 자료를 이용하는 과정에서 이 예를 볼 수 있다.

양(quantity) H 는 대각원소가 (diagonal element)가 2차형태(quadratic form)인 행렬(matrix)로 정의된다.

$$h_{ii} = x'_i(X'X)^{-1}x_i$$

여기에서 x_i 는 모형과 i 번째 자료 포인트의 위치를 반영한다. 비대각요소(off-diagonal element)는 $h_{ij} = x'_i(X'X)^{-1}x_j$, $i \neq j$ 이다. 여기에서 행렬 H 는 다음과 같이 표기할 수 있다.

$$H = X(X'X)^{-1}X' \quad (3.39)$$

위의 값(quantity)은 4, 5, 6, 7장에서의 전개과정에 많이 나올 것이고 흔히 “모자”행렬(HAT matrix)이라고 불린다. 이것은 대칭성이 있고, 멱등행렬(idempotent, $H^2=H$)이고, y 를 \hat{y} 로 변환시키는 $n \times n$ 행렬이다. 즉,

$$\hat{y} = Xb = X(X'X)^{-1}X'y = Hy$$

개다가 사영행렬(projection matrix)로서 H 는 일반선형모형이론(geneal linear models theory)에서 중요한 역할을 한다. 사실 간단한 조작에 의해 잔차(residuals)의 벡터(vector)를 다음과 같이 표현할 수 있다.

$$e = y - Xb = y - X(X'X)^{-1}X'y = [I - H]y \quad (3.40)$$

또한 제곱잔차합(residual sum of squares)은 다음과 같다.

$$e'e = y'[I - H]^2 y \quad (3.41)$$

$I - H$ 가 역등행렬(idempotent)이기 때문에(연습문제 3.12를 볼 것) (3.41)식은 다음과 같이 다시 쓸 수 있다.

$$e'e = y'[I - H]y \quad (3.42)$$

앞으로 나올 표현에서 이용 가능한 모자행렬의 몇 가지 특성이 있는데 여기에 소개하면 다음과 같다.

1. $\text{tr}(H)=p$, p 는 모형 모수(parameter)의 수
2. 일정한 상수 항을 갖고 있는 모형에서 $\frac{1}{n} \leq h_{ii} \leq 1.0$

이 두가지 성질에서 흥미로운 점을 관찰할 수 있는데, 특성 1은 다음 식이 성립함을 의미한다.

$$\sum_{i=1}^n \frac{\text{Var } \hat{y}(x_i)}{\sigma^2} = p \quad (3.43)$$

이것은, σ^2 과는 별개로, 자료점의 위치에 걸쳐 합해진 예측분산(prediction variance summed over the locations of the data points)이 모형 모수의 수(the number of model parameters)와 동일하다는 것을 보여주는 결과이다. 이 시점에서 독자들에게 이러한 결과가 함축하는 바가 명확할 수도 있고 그렇지 않을 수도 있지만, 모형을 개발하는 과정에서 (적어도 예측분산에 관한 한) 단순모형(simple models)의 선택에 어느 정도 신빙성을 제공해 줄 수는 있을 것이다.

특성 2는 아래의 식이 성립함을 뜻한다.

$$\frac{1}{n} \leq \frac{\text{Var } \hat{y}(x_i)}{\sigma^2} \leq 1$$

이 결과가 합당하다는 것을 직관적으로 알 수 있다. 이것은, 자료점의 위치에 있어서의 예측의 정밀함(precision in a prediction at the location of a data point)은 관측치에서의 오차분산(error variance in an observation) 보다 더 나쁘지 않음을 제시해준다. 즉, $Var \hat{y}(x_i) \leq \sigma^2$. 또한, 예측의 정밀함(precision in prediction)은 모든 관측치가 같은 위치(same location)에서 얻어졌을 때의 평균 반응의 예측치(precision of the average response)(σ^2/n)보다 더 낫지 않을 것이다.

이 절에서 다루는 내용이 연이어 나올 다음 장에서 중요한 기초자료(building block)가 될 것이다. 모자 행렬(HAT matrix)의 요소(elements)와 관련된 양(quantities)으로부터 많은 정보를 뽑아낼 수 있다.

3.10. 범주형 또는 지표형 변수; 회귀모형과 ANOVA모형 (Categorical or Indicator Variables; Regression Model and ANOVA Model)

회귀분석의 매우 중요한 적용(important application)은 보통의 전통적인 양적인 변수(traditional quantitative variable)뿐 아니라 질적인 변수(qualitative variable)도 포함하는 회귀변수 목록(a list of regressor variable)을 수반한다. 예를 들어, 화공학자가 온도(x_1)와 압력(x_2)의 함수로서 반응 y 의 반응량(yield of reaction)을 모형화 한다면, 두 개의 서로 다른 촉매를 이용하는 모형을 구축해야 할 것이고, 이 때 촉매(catalyst)가 범주형 변수(categorical variable)의 예(an example)가 된다. 만일 개인 집을 구입하는 가격 y 에 대한 모형이라고 가정해보면 관계 있는 회귀변수의 함수(function)로서 x_1 (생활공간의 평방피트), x_2 (대지의 에이커), x_3 (방의 개수), ... 를 생각해 볼 수 있다. 그러나, 미국의 4가지 뚜렷한 지질학적 위치는 집을 구입하는 가격에 영향을 미칠 수 있으므로 지질학적 위치는 범주형 변수로서의 자료에 포함될 자격이 충분히 있다. 종종 지표형(indicator) 또는 심지어 가변수(dummy variable)와 같은 항들(terms)은 보통의 양적인 변수(quantitative variables) 만큼이나 결정적인 모형 항(term)으로 나타나는 무엇인가를 서술할 때 이용한다. 그것들은 보통 장애변수(nuisance variable)로 나타나고, 가급적 계획되지 않는다. 그렇지만 세심한 연구자는 범주형 변수(categorical variable)의 미리 선택된 수준(prechosen levels)이 포함되도록 가정해 놓고 자료를 취하거나 실험과정을 계획할지도 모른다.

반응이 변수선별활동(variable screening exercise)을 통해 범주형 변수(categorical variable)에 의해 영향을 받을 것인지 아닌지를 결정하는 것은 중요하다. 확실히, 예측이 중요하다면 예측식 ($\hat{y}(x)$) 이 범주(category)에 의한 효과(effect)를 포함할 필요가 있다. 범주형 변수(categorical variable)를 보여주는 가장 효과적인 모형은 하나의 양적인 변수(quantitative variables)와 두개의 범주(categories)를 갖는 특수한 경우이다.

지표형변수(indicator variable)의 개념(concept)은 모형에서 변수의 역할(role of variable)을 설명하기 위해 지표함수(indicator function)를 이용하는 것을 필요로 한다. 이것은 표준연속 회귀변수(standard continuous regressor)와 지표형변수(indicator variable)를 구별하는 표기법(notation)을 사용할 것을 요구한다.

두 범주 내에서의 단일 범주형변수(Single Categorical Variable with Two Categories)

두 가지 범주(categories)가 있는 상황에서(예를 들어 이전의 예로 든 두 가지 촉매) 단일 양적인 회귀변수(single quantitative regressor variable)를 고려한다고 가정해 보자. 이는 부가적인 모형 항(additional model form)을 양산하게 된다. z 를 다음과 같이 정의하자.

첫 번째 범주 내에 있다면 $z=0$

두 번째 범주 내에 있다면 $z=1$

따라서 모형은 다음과 같이 성립된다.

$$y_i = \beta_0 + \beta_1 x_{li} + \beta_2 z_i + \varepsilon_i \quad (3.44)$$

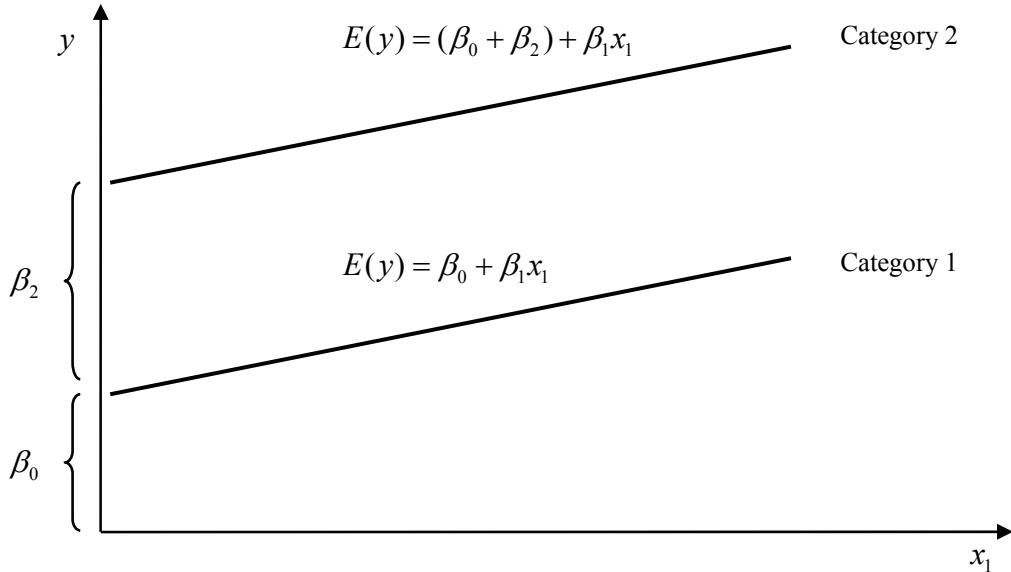
X 행렬로 표시하면 다음과 같다.

$$X = \begin{bmatrix} 1 & x_{11} & 0 \\ 1 & x_{12} & 0 \\ 1 & x_{13} & 0 \\ \vdots & \vdots & \vdots \\ & & 0 \\ \cdots & & \\ & & 1 \\ \vdots & \vdots & 1 \\ & & \vdots \\ 1 & x_{1n} & 1 \end{bmatrix} \quad \left. \begin{array}{l} \text{First category} \\ \text{Second category} \end{array} \right\}$$

그 결과로 첫 번째 범주에서는 $y_i = \beta_0 + \beta_1 x_{li} + \varepsilon_i$, 두 번째 범주에서는

$y_i = (\beta_0 + \beta_2) + \beta_1 x_{li} + \varepsilon_i$ 가 된다. 단일 범주형 변수(single categorical variable)는 범주간의 반응에 있어 일정한 상수 차이에 의한 절편의 단순한 이동(shift)을 놓게 된다. Fig 3.4에 그 예가 제시되었다.

Figure 3.4 단일 양적 변수, 두 수준을 지닌 하나의 범주형 변수



여기에서 살펴 본 특수한 경우는 실제로 특별하지만, 여기에서 이용되었던 중요한 모형가정(model assumption)은 보다 일반화된 경우에 대해서도 적용된다. 회귀계수(regression coefficient) 즉 양적인 회귀변수(quantitative regressor variable) 혹은 변수들(variables)에 대한 변화율(rate of change)은 모든 범주에서 동일하다.

하나의 범주(one category) 이상으로 이야기를 확대하기 전에 두 범주(two categories)에 할당되는 0과 1은 전적으로 임의(random)에 의해서 이루어 진다는 것을(실제로는 인위적이지만) 명심해야 한다. 만일 (0,1) 이외의 다른 값을 할당한다면 β_2 의 측정값은 달라질 것이다. 그렇지만 β_2 와 관련된 t -통계량이나 예측 $\hat{y}(x_0)$ 과 같은 중요한 추정식(inferential equation)은 범주형 변수(categorical variable)의 수준(level)을 할당하는 것과는 무관하다.

예제 3.9 당뇨병관련 쥐실험 자료

Table 3.7에는 정상적인 쥐와 당뇨병을 갖고 있는 쥐의 몸무게(body weight, 단위: gram)와 신장의 무게(kidney weight, 단위: mg)를 적어 놓았다. 실험에 사용된 정상적인 쥐와 당뇨병을 갖고 있는 쥐의 수는 각각 25마리와 9마리이다. 당뇨병을 갖고 있는 쥐는 두 순수혈통을 가진 쥐로부터 교차번식(cross-bred)되었음을 참고하기 바란다. 같은 몸무게라면 당뇨병을 갖고 있는 쥐의 신장 무게가 정상적인 쥐보다는 항상 더 무겁다는 가설을 검정한다든지, 당뇨병이 있는 쥐의 경우 몸무게가 늘어날수록 신장 무게가 몸무게에서 차지하는 비율이 더 커진다는 사실이 분석의 관심대상이다. 신장 무게는 일반적으로 몸무게에 비례한다고 알려져 있다.

신장 무게와 몸무게의 관계에 당뇨병이 어떠한 관련이 있는지를 알아보기 위하여 당뇨병을 갖고 있는 경우의 주는 1의 값을 갖고 정상적인 주는 0의 값을 갖는 범주형 변수 z 를 고려하여 보자. 모형은 다음 식과 같이 나타난다.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \quad i = 1, 2, \dots, 34$$

Table 3.8은 처리한 결과를 나타내었는데, 분산분석, 기본적인 통계량, 회귀계수에 대한 t -통계량, 잔차, 평균반응에 대한 신뢰구간이 나타나 있다.(특정한 표제는 용어의 일관성을 유지하기 위해 변경시켰다.)

Table 3.7 당뇨병관련 주실험 자료

당뇨병을 갖고 있는 주(z=0)		정상적인 주(z=1)			
몸무게(X)	신장무게(Y)	몸무게(X)	신장무게(Y)	몸무개(X)	신장무개(Y)
42	1030	34	810	37	780
44	1240	43	480	38	660
38	1150	35	680	32	750
52	1280	33	920	36	780
48	1240	34	650	32	670
46	1100	26	650	32	670
34	1040	30	650	38	700
44	1080	31	560	42	720
38	870	31	620	36	800
		27	740	44	830
		28	600	33	640
		27	640	38	800
		30	690		

Table 3.8 분석결과

Analysis of Variance Table					
Response: y					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	702566	702566	71.296	1.545e-09 ***
z	1	482541	482541	48.968	7.450e-08 ***
Residuals	31	305481	9854		

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	445.427	119.007	3.743	0.000742 ***
x	7.502	3.463	2.166	0.038108 *
z	347.258	49.625	6.998	7.45e-08 ***

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

Residual standard error: 99.27 on 31 degrees of freedom

Multiple R-Squared: 0.7951, Adjusted R-squared: 0.7818

F-statistic: 60.13 on 2 and 31 DF, p-value: 2.139e-11

Observed	Fitted	Residual	Lower 95% CL	Upper 95% CL
Value	Value			
1 1030	1107.7759	-77.775854	1039.9980	1175.5537
2 1240	1122.7802	117.219818	1054.8389	1190.7214
3 1150	1077.7672	72.232802	1001.9585	1153.5759
4 1280	1182.7975	97.202506	1089.5437	1276.0513
5 1240	1152.7888	87.211162	1076.2524	1229.3253
6 1100	1137.7845	-37.784510	1066.8103	1208.7588
7 1040	1047.7585	-7.758542	955.5810	1139.9360
8 1080	1122.7802	-42.780182	1054.8389	1190.7214
9 870	1077.7672	-207.767198	1001.9585	1153.5759
10 810	700.5003	109.499740	659.9995	741.0010
11 480	768.0197	-288.019736	691.9308	844.1087
12 680	708.0024	-28.002424	666.7450	749.2599
13 920	692.9981	227.001904	652.0319	733.9643
14 650	700.5003	-50.500260	659.9995	741.0010
15 650	640.4829	9.517052	571.6516	709.3143
16 650	670.4916	-20.491604	621.5966	719.3866
17 560	677.9938	-117.993768	632.6789	723.3086
18 620	677.9938	-57.993768	632.6789	723.3086
19 740	647.9851	92.014888	584.7292	711.2410

20	600	655.4873	-55.487276	597.4816	713.4930
21	640	647.9851	-7.985112	584.7292	711.2410
22	690	670.4916	19.508396	621.5966	719.3866
23	780	723.0068	56.993248	676.9060	769.1075
24	660	730.5089	-70.508916	680.6439	780.3739
25	750	685.4959	64.504068	642.8821	728.1097
26	780	715.5046	64.495412	672.3324	758.6767
27	670	685.4959	-15.495932	642.8821	728.1097
28	670	685.4959	-15.495932	642.8821	728.1097
29	700	730.5089	-30.508916	680.6439	780.3739
30	720	760.5176	-40.517572	690.3082	830.7269
31	800	715.5046	84.495412	672.3324	758.6767
32	830	775.5219	54.478100	693.3666	857.6772
33	640	692.9981	-52.998096	652.0319	733.9643
34	800	730.5089	69.491084	680.6439	780.3739

범주형 변수(즉, 당뇨여부)에 대한 회귀계수의 가설검정(test)과 부분 F -통계량(partial F -statistic)은 다음과 같다.

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

$$F = \frac{R(\beta_2 | \beta_0, \beta_1)}{s^2} = \frac{482541}{9854} = 48.968$$

검정결과 유의수준 0.001이하에서 β_2 는 0과 유의한 차이가 있다라는 것을 알 수 있다.

많은 자료 세트처럼, 여기에서도 대답하지 못하는 몇 가지 문제가 있다. 범주형 변수는 당뇨 여부에 따른 차이를 설명한다. 하지만, 어떤 모형을 선택해야 하는가? 모형의 목적이 당뇨병이 있는 쥐의 경우 몸무게가 늘어날수록 신장 무게가 몸무게에서 차지하는 비율이 더 커진다는 사실을 밝히고자 한다면 여기에서 채택된 모형이 가장 효과적인 것인가에 대한 의문은 해결해 주지 못한다. 이 장 앞에서 언급했듯이 모형 선택에 관해서는 4장에서 상세히 다를 것이다.

예제에 사용된 R-code는 아래와 같다.

```
data<-read.table("d:/data/ex3_9.R",header=TRUE)
g<-lm(y~x+factor(z),data)
g1<-lm(y~x,data)
g2<-lm(y~factor(z),data)
```

```

s_stat<-summary(g)

ano<-anova(g)

ano1<-anova(g1,g)

ano2<-anova(g2,g)

ano3<-anova(g2)

pre<-predict(g,interval="confidence")

res<-residuals(g)

t3_10<-cbind(data[,2],pre[,1],res,pre[,2:3])

colnames(t3_10)<-c('Observed Value','Fitted Value','Residual','Lower 95% CL','Upper 95% CL')

```

다중수준을 지닌 단일 범주형변수(One Categorical Variable with Multiple Levels)

예제 3.9에서 당뇨병 여부는 범주형 변수(categorical variable)이고 두 가지 수준(level)이 있다. 하나의 범주형 변수(categorical variable)에 여러 개의 수준(levels)을 가지는 경우(ℓ 수준, k 개의 정량적 회귀변수를 가진 모형)를 생각해 보면 다음과 같이 표기할 수 있다.

$$y = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \beta_{k+1} z_{1i} + \beta_{k+2} z_{2i} + \cdots + \beta_{k+\ell-1} z_{\ell-1,i} + e_i \quad (3.45)$$

여기에서 $z_1, z_2, \dots, z_{\ell-1}$ 값은 0 아니면 1이고, 의문시되는 자료 포인트(data point in question)가 그 범주 내에 어디에 있는지에 의해 결정된다. (3.45)식의 모형을 수용하는 X 행렬은 다음과 같다.

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_k & z_1 & z_2 & \cdots & z_{\ell-1} \\ 1 & x_{11} & x_{21} & \cdots & x_{k1} & 1 & 0 & 0 \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} & 1 & 0 & 0 \\ \vdots & x_{13} & x_{23} & \cdots & x_{k3} & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \cdots & \vdots & 1 & 0 & 0 \\ \dots & & & & & & & \\ 1 & \vdots & \vdots & \cdots & \vdots & 0 & 1 & 0 \\ 1 & \vdots & \vdots & \cdots & \vdots & 0 & 1 & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \cdots & \vdots & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \cdots & \vdots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \cdots & \vdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} & 0 & 0 & 0 \end{bmatrix} \quad (3.46)$$

Category level 1

Category level 2

Category level $\ell-1$

Category level ℓ

여기에서 변수 $z_1, z_2, \dots, z_{\ell-1}$ 의 $\ell-1$ 개의 계수들을 설명하는 X 행렬의 $\ell-1$ 개의 부가적인 열(additional column)이 있는데, 이는 보통회귀모형(ordinary regression model)에 단일 부가 효과(single additive effect)를 일으키며, 그 효과는 어떤 수준(level)이 적절한지에 따라 좌우된다. 예를 들어 계수가 $b=(X'X)^{-1}X'y$ 로 알려진 후에 두 번째 범주에서 예측된 반응은 다음과 같이 주어진다.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + b_{k+2} \quad (3.47)$$

범주 수준이 1로 정해져 있는 x_{k+2} 를 제외하고는, 모든 범주 수준(all category level)은 0으로 정해져 있다. 앞서 언급했듯이 범주형계수(categorical coefficients)는 상수항에서 양(positive) 또는 음(negative)으로 이동(shift)하는 것으로 시작화되고, 이 이동은 어떤 범주형 예측에 토대를 두었는가에 달려 있다. (3.47)의 예측에 있어 β_0 에서의 이동은 $\beta_0 + \beta_{k+2}$ 로 나타난다. 임의로 최종 범주 수준 ℓ 을 $z_1=z_2=\dots=z_{\ell-1}=0$ 인 값으로 할당시켰고, z_ℓ 을 제외하고는, $b_{k+2}=0$ 로 임의적인 할당이 이루어졌다. 0으로의 할당(assignment)은 어떤 계수에서도 주어질 수 있다. β_{k+j} ($j=1, 2, \dots, \ell$)을 추정할 수는 없을지라도 $\beta_0 + \beta_{k+j}$ 는 추정 가능하고, 새로운 절편"(따라서, 예측 반응)은 임의의 할당(arbitrary assignment)과 무관하다(independent).

예제 3.10

Table 3.9의 자료를 살펴보면 반응변수 y 는 석탄정화장치(coal cleansing system)에서 부유된 고체(suspended solid)의 양이다. 회귀변수 x_1 을 이용하여 정화탱크의 pH를 기록하였다. 정화공정에서는 세가지의 서로 다른 중합체(polymer)가 사용되었다. 다른 중요한 조건은 동일하게 유지시켰다. 분석은 다중회귀를 이용하였고, 여기에는 pH에 대한 선형모형항(linear model term)과 3개의 범주 또는 계급(class)를 갖는 중합체를 나타내는 범주형변수가 포함된다. 이 중합체 범주는 0과 1의 값을 갖는 두 가지 회귀변수로 모형화시켰다. 이 모형을 기술하면

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 z_{1i} + \beta_3 z_{2i} + \varepsilon_i \quad (i=1,2,\dots,18)$$

여기에서 x_{1i} 는 i 번째 자료에서의 pH를 의미하고, z_{1i} 와 z_{2i} 의 값은 중합체로 결정된다. 모형행렬은 다음과 같이 표시된다.

$$X = \begin{bmatrix} 1 & x_{1.1} & 1 & 0 \\ 1 & x_{1.2} & 1 & 0 \\ 1 & \vdots & 1 & 0 \\ \cdots & & & \\ 1 & \vdots & 0 & 1 \\ \cdots & & & \\ 1 & \vdots & 0 & 0 \\ 1 & x_{1.17} & 0 & 0 \\ 1 & x_{1.18} & 0 & 0 \end{bmatrix} \quad \left. \begin{array}{l} \text{Polymer 1} \\ \text{Polymer 2} \\ \text{Polymer 3} \end{array} \right\}$$

회귀분석 결과 추정회귀식은 다음과 같이 주어진다.

$$\hat{y} = -161.89733 + 54.2940x_{1i} + 89.9981z_{1i} + 27.1657z_{2i}$$

Table 3.9는 회귀분석 결과를 나타낸 것이고 pH는 확실히 정화과정에서 효과(effect)를 미치고 있으며, 중합체의 종류에 따른 효과도 있는 것으로 나타난다. 이 자료에 대한 더 많은 연구가 다음 절에서 나올 것이다.

Table 3.9 예제 3.10의 정화자료

x_1 (pH)	y (amount of suspended solids (mg/L))	Polymer
6.5	292	Polymer 1
6.9	329	Polymer 1
7.8	352	Polymer 1
8.4	378	Polymer 1
8.8	392	Polymer 1
9.2	410	Polymer 1
6.7	198	Polymer 2
6.9	227	Polymer 2
7.5	277	Polymer 2
7.9	297	Polymer 2
8.7	364	Polymer 2
9.2	375	Polymer 2
6.5	167	Polymer 3
7.0	225	Polymer 3
7.2	247	Polymer 3
7.6	268	Polymer 3
8.7	288	Polymer 3
9.2	342	Polymer 3

Figure 3.5 산점도

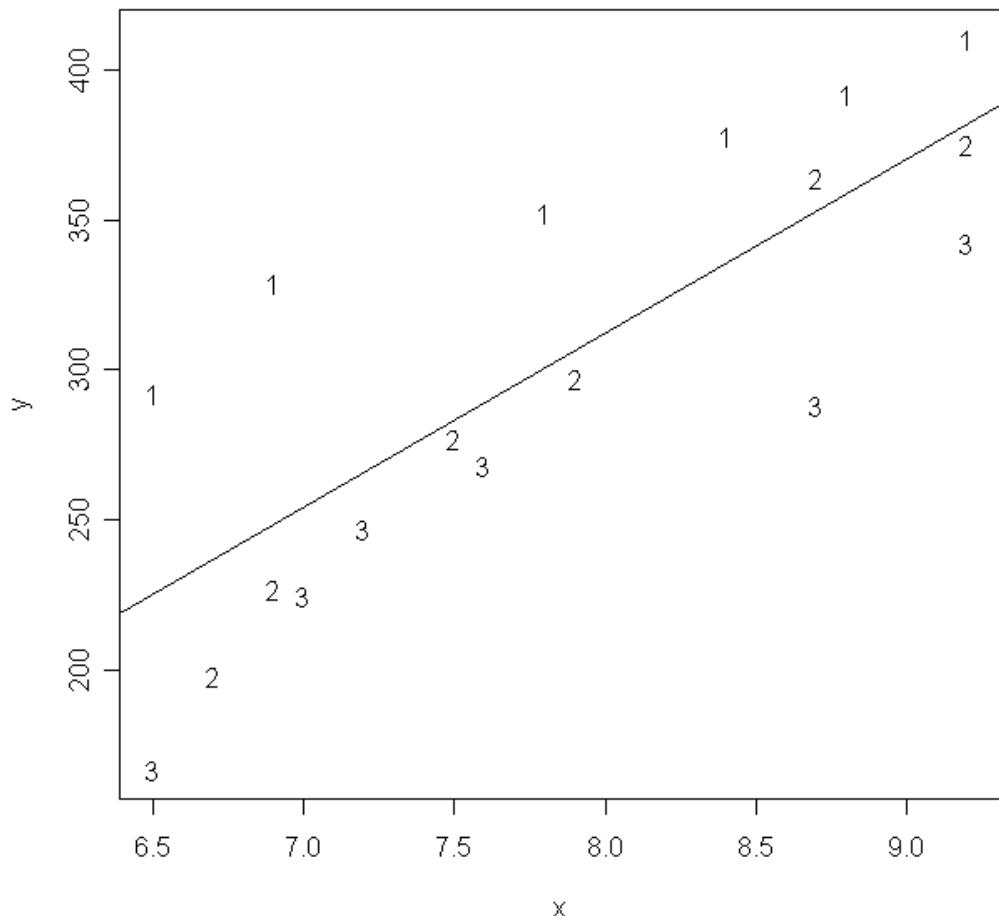


Table 3.10 분석결과

Analysis of Variance Table					
	Response: y	Df	Sum Sq	Mean Sq	F value
z1		1	29527	29527	81.3934 3.291e-07 ***
z2		1	3367	3367	9.2808 0.008711 **
x		1	47288	47288	130.3551 1.767e-08 ***
Residuals	14	5079	363		

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

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-161.897	37.433	-4.325	0.000699 ***
z1	89.998	11.052	8.143	1.11e-06 ***
z2	27.166	11.010	2.467	0.027127 *
x	54.294	4.755	11.417	1.77e-08 ***

Signif. codes:	0	'***'	0.001	'**'
	0.01	'*'	0.05	'.'
	0.1	' '	1	
Residual standard error:	19.05	on 14 degrees of freedom		
Multiple R-Squared:	0.9404,	Adjusted R-squared:	0.9277	
F-statistic:	73.68	on 3 and 14 DF,	p-value:	8.14e-09

계급간 회귀변수 효과의 동등성(Equality of Regressor Effects across Classes))

이 절의 초반부에서 우리는 양적인 회귀변수(quantitative regressor)의 효과가 범주형 변수(categorical variable)의 여러 수준(levels)이나 계급(classss)에 걸쳐(across) 동등하다(same)는 가정을 강조하였다. 이것은 fig 3. 4에 제시된 평행한 선들을 이용한 모형 묘사(parallel lines model description)에서 확실히 알 수 있다. 기술자, 생물학자, 물리 과학자 혹은 어떤 과학 연구자들에게 이 진술(statement)은 확실히 중요하다. 가법성 가정(additivity assumption)은 맹목적으로 컴퓨터에 자료(data)와 모형 명령어(model statement)를 입력(submit)하는 연구자에 의해 흔히 간과된다. 물론 연구자들이 가법성 가정을 하는 것이 적합한지 아닌지를 알지 못하는 경우도 흔하다. 명백하게 가정이 유지되지 않거나 어떤 의미에서 조사(investigate)되지 않는다면, 사용자는 전반적으로 잘못된 모형을 적합하여 타당하지 않은 해석을 하는 위험에 빠질 수도 있다. 여기에서 우리가 논의할 것은 가법성 가정(additivity assumption)을 조사하는 방법론(methodology)적인 것이다.

가법성 검정을 위한 일반 선형 가설의 사용(Using the General Linear Hypothesis to Test Additivity)

이제 fig 3.4에 그려져 있는 간단한 사례를 참조해보자. 이것은 2개의 수준(level)과 x_1 으로 표시되어 있는 하나의 양적회귀변수(quantitative regressor variable)를 가진 하나의 범주형 변수(categorical variable)이다. 식 3.44에서 알 수 있듯이, 각 범주에 대해서 x_1 의 공통 기울기 계수(common slope coefficient)를 가정(assume)하고 있다는 것은 명백하다. 그러나, 좀 더 일반적인 모형, 즉, x_1 에 대하여 2개의 다른(separate) 기울기를 가정하는 모형을 살펴보자. Class 1로부터 n_1 관측치가, class 2로부터 n_2 관측치가 취해지고, $n_1 + n_2 = n$ 이라고 가정할 때, 다음과

같이 기술할 수 있다.

$$\begin{aligned} E(y_i) &= \beta_0 + \beta_{11} x_{i1} & (i = 1, 2, \dots, n_1) \\ &= \beta_0 + \beta_{12} x_{i1} + \beta_2 z_i & (i = n_1 + 1, n_1 + 2, \dots, n) \end{aligned} \quad (3.48)$$

여기의 z_i 는 3.44의 모형에서처럼, 첫번째 범주에서는 0으로 두 번째 범주에서는 1로 여전히 정의된다. 따라서 우리는 본질적으로 다른 절편과 다른 기울기를 갖는 두 가지 회귀 모형을 갖게 된다. 분명하게 가법성에 대한 검정(test for additivity)은 다음을 검정하는 것이다.

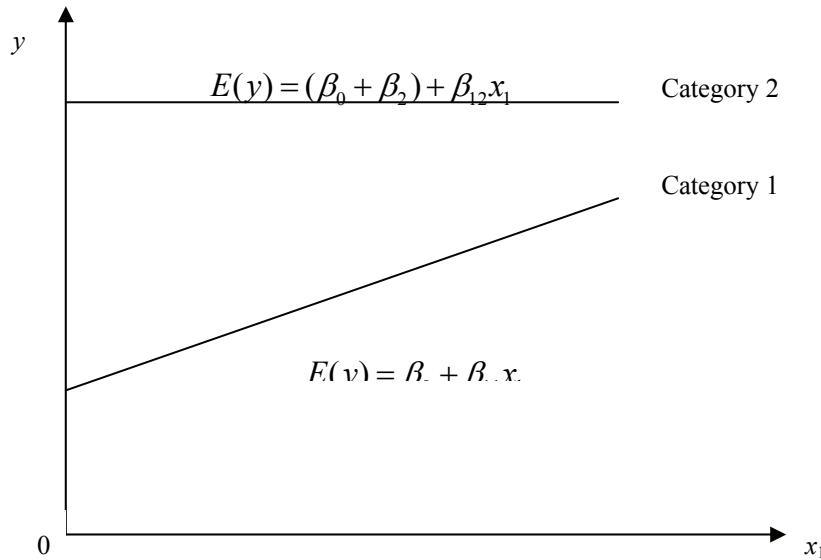
$$\begin{aligned} H_0 &: \beta_{11} = \beta_{12} \\ H_1 &: \beta_{11} \neq \beta_{12} \end{aligned}$$

즉, 두 회귀선의 기울기의 동등성(equality)을 검정하는 것이다. 이 가설의 검정은 3.4절에서 일반 선형 가설(general linear hypothesis)을 다를 때에 논의되었다. 어떤 상황에서도, 이 검정(test)은 각각의 양적 회귀변수(each quantitative regressor)의 회귀계수들(regression coefficients)의 동등성(equality), 혹은 아마도 부류(class)간의 양적 회귀변수의 부분집합(a subset of the quantitative regressors across classes)의 회귀계수들의 동등성(equality)을 살펴보기 위한 것이다. 분명히, 각각의 부류(class)에 대하여 다른(separate) 독립된 회귀모형을 취함으로써, 일반 선형 가설의 사용이 확실히 적용된다. 이것은 예제 3.10의 사례에서처럼 세가지 이상의 범주(categories)로까지 쉽게 확장된다.

가법성 검정을 위한 상호작용의 사용(Using Interaction to Test Additivity)

비록 일반선형가설(general linear hypothesis)이 가법성가정(additivity assumption)을 검정하는데 잘 적용된다 하더라도, 동등한 결과를 얻어내는 대체적인 접근(alternative approach)이 논의될 가치는 있다. Fig 3.4의 2개의 범주 상황에서 단일 양적 회귀변수(quantitative regressor)를 먼저 고려해보자. 사실 만약 회귀선의 기울기가 다르다면 모형은 fig 3.6에 묘사된 것과 같다. 이러한 것이 발생했을 때, 우리는 x_1 과 범주형 변수 z_2 사이에 상호작용(interaction)이 있다고 말한다. 상호작용(interaction)이란 가법성(additivity)으로부터 편차(deviation)를 의미한다. 따라서 비가법성(nonadditivity)의 조건(condition)은 상호작용(interaction)을 포함하는 모형으로 잘 모형화할 수 있다.

Figure 3.6 Interaction between the quantitative regressor and the categorical variable



다시 2개의 classes에 대해 범주형 변수(categorical variable) z_2 에 0과 1을 대입해보자. 우리는 다음과 같은 상호작용이 있는(따라서 기울기가 다른) 모형을 고려해볼 수 있다.

$$y_i = \beta_0 + \beta_{11}x_{1i} + \beta_2 z_i + \beta_{12}x_{1i}z_i + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (3.49)$$

이 상호작용 모형은 양적 회귀변수(quantitative regressor) x_1 의 다른 기울기(separate slopes)를 갖고 있는 (3.48)의 모형과 동등하다는 것을 독자들은 명백히 알아야 한다. 사실, (3.48)의 β_{12} 는 (3.49)식의 $\beta_{11} + \beta_{12}$ 이다. 행렬 X 로 나타내면,

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & x_{11} & 0 & 0 \\ 1 & x_{12} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n_1} & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1,n_1+1} & 1 & x_{1,n_1+1} \\ 1 & \vdots & 1 & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & 1 & x_{1,n} \end{array} \right] \left\{ \begin{array}{l} \text{First category} \\ \text{Second category} \end{array} \right\}$$

이 공식을 유도하면,

$$E(y_i) = \beta_0 + \beta_{11}x_{1i} \quad (\text{first category})$$

$$E(y_i) = \beta_0 + (\beta_{11} + \beta_{1z})x_{1i} + \beta_2 \quad (\text{second category})$$

여기에서 z_i 는 두 번째 범주(second category)에서 1의 값이 대입된다. 분명하게 만약 상호작용(interaction)이 없다면, 즉 $\beta_{1z}=0$ 이라면 모형은 가법성 모형(additive model) 혹은 (3.44)의 동일 기울기 모형(equal slope model)으로 축소될 것이다.

앞서 말한 것의 결론(upshot)은 가법성 가정(additive assumption)에 관련되어 있을 때, 상호작용 항(interaction term)들을 간단히 모형화하고 이러한 항들의 가설을 검정할 수 있다는 것이다. (3.49)의 모형을 둘러싼 개요(scenario)에서, 3.4 절에서 논의되었던 t -검정 혹은 부분 F -검정(partial F -test)을 사용한 다음의 검정은 2개의 기울기의 동일성을 검정하는 일반 선형 가설의 접근법과 동등하다는 것이다.

$$\begin{aligned} H_0 &: \beta_{1z} = 0 \\ H_1 &: \beta_{1z} \neq 0 \end{aligned}$$

F 검정으로 얻어진 일반선형가정(general linear hypothesis)을 회상(recall)해보자. 모형 (3.48)의 상황에서, F 에 대한 자유도는 분자(완전모형(full model)과 축소모형(reduced model)의 오차자유도(error df) 차이)는 1이 되고, 오차(총표본크기에서 추정된 모수의 수를 뺀 것)는 n_1+n_2-4 가 될 것이다. 단순히 $H_0: \beta_{1z}=0$ 를 검정하는 경우에, n_1+n_2-4 의 자유도를 가진 t -검정으로서 F -검정과 동등한 검정을 수행할 수 있다.

우리는 가법성 가정(additivity assumption)을 검정하기 위하여 상호작용의 사용 예를 연습해 보고자 한다.

예제 3.11

Table 3.9에 보여진 예제 3.10의 자료를 참조하라. 지시변수(indicator variable)인 중합체는 세 가지 수준(levles)의 범주를 가지고 있고, 단일 양적 회귀변수(single quantitative regressor)인 “시스템의 pH”가 있다.

Fig 3.7의 자료의 그림을 참조하라. 그림만으로는 가법성 가정(additivity assumption)이 만족되는지는 불명확하다. 아마도 세가지 중합체에 대해 공통 기울기 가정(common slope assumption)을 하는 것은 적당치 않을 것이다. 이것을 검정하기 위하여 우리는 다음의 모형을 적합시킨다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 z_{1i} + \beta_3 z_{2i} + \beta_{1z_1} x_{1i} z_{1i} + \beta_{1z_2} x_{1i} z_{2i} + \varepsilon_i \quad (i = 1, 2, \dots, 18) \quad (3.50)$$

여기에서 전과 마찬가지로, 양적 회귀변수(quantitative regressor) x_1 은 pH이고, z_1 과 z_2 는 세 가지 중합체를 나타내기 위해 사용된다. 만약 가법성 가정(additivity assumption)에 대해 검정하고 싶다면, 다음을 검정하면 된다.

$$H_0 : \beta_{1z_1} = \beta_{1z_2} = 0$$

이것은 3.4절에서 논의된 것처럼 F -검정을 사용하여 쉽게 완성된다. 만약 이 가정이 기각된다면 우리는 공통 기울기 가정(common slope assumption)이 유지되지 않는다고 결론지어야만 한다. 그리고나서 명백히, 이것이 뜻하는 바는 각각의 중합체에 대해 다른 회귀선(separate regression lines)을 적합시켜야 한다는 것이다. Table 3.11은 식 (3.50)의 상호작용 모형을 분석한 결과를 보여주고 있다. 가법성 가정(additivity assumption)을 검정하기 위해서, 다음의 내용이 필요하다.

$$R(\beta_1, \beta_2, \beta_3, \beta_{1z}, \beta_{1z_2} | \beta_0) - R(\beta_1, \beta_2, \beta_3 | \beta_0) = R(\beta_{1z_1}, \beta_{1z_2} | \beta_0, \beta_1, \beta_2, \beta_3)$$

이 2개의 회귀 제곱합(regression sum of squares)은 table 3.10와 3.11에서 찾아낸 것이다. 더불어, “Type 1”제곱합(sums of squares)이 순차제곱합(sequential sums of squares)이기 때문에 $R(\beta_{1z_1}, \beta_{1z_2} | \beta_0, \beta_1, \beta_2, \beta_3)$ 는 1747.0925+778.9528로 주어지는 것을 쉽게 알 수 있다. 따라서,

$$F_{2,12} = \frac{2526.0453 / 2}{212.7223} = 5.94$$

이 F 통계량은 0.016수준(level)에서 유의하며, 이는 상호작용 모형항들(interaction model terms)이 유의하다는 것을 의미한다. 다음과 같이 결론지어야 한다.

- (i) 예제 3.10의 모형적합은 적절하지 않다.
- (ii) pH의 효과는 3개의 다른 polymer에 대하여 각각 다르다.
- (iii) 모형적합을 위하여 3개의 polymer에 대하여 다른 회귀선(separate regression lines)을 적합시켜야 한다.

Figure 3.7 Plot of the cleansing data of Table 3.9

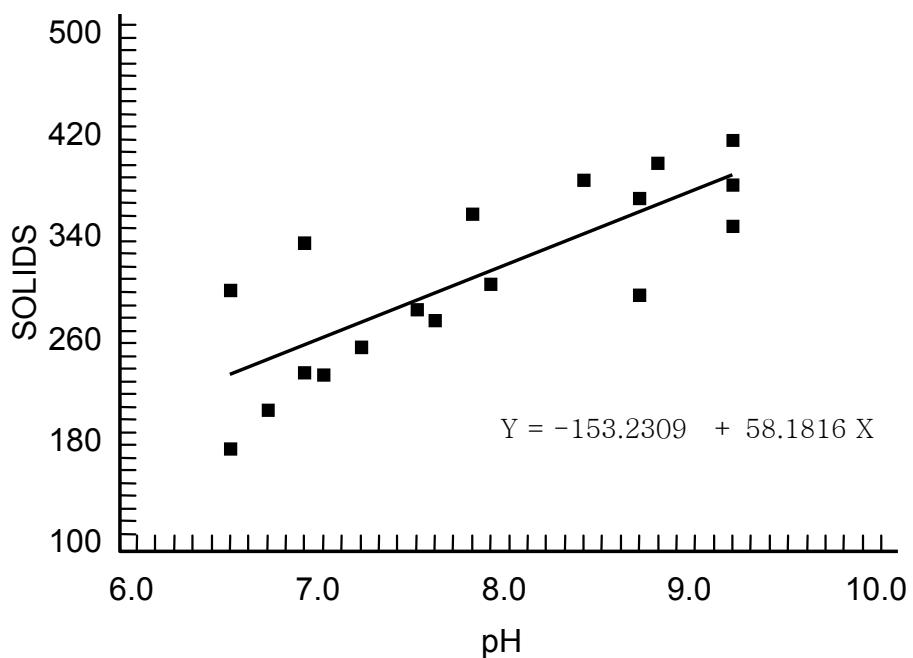


Table 3.11 모형(3.50)을 이용한 분석결과

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	54856	54856	257.8741	1.778e-09 ***
z1	1	23118	23118	108.6762	2.281e-07 ***
z2	1	2208	2208	10.3812	0.007328 **
x1:z1	1	1747	1747	8.2130	0.014199 *
x1:z2	1	779	779	3.6618	0.079823 .
Residuals	12	2553	213		

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

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-158.275	48.517	-3.262	0.0068 **
x	53.824	6.253	8.607	1.76e-06 ***
factor(z1)1	197.692	68.795	2.874	0.0140 *
factor(z2)1	-108.740	71.051	-1.530	0.1518
z1	NA	NA	NA	NA

```

z2          NA      NA      NA      NA
x:z1       -13.561   8.737  -1.552   0.1466
x:z2        17.394   9.090   1.914   0.0798 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 14.59 on 12 degrees of freedom
Multiple R-Squared: 0.9701,    Adjusted R-squared: 0.9576
F-statistic: 77.76 on 5 and 12 DF,  p-value: 1.016e-08

Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ z1
Model 3: y ~ z1 + z2
Model 4: y ~ z1 + z2 + x1
Model 5: y ~ z1 + z2 + x1 + x1 * z1
Model 6: y ~ z1 + z2 + x1 + x1 * z1 + x1 * z2

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     17 85260
2     16 55734  1    29527 138.8039 5.933e-08 ***
3     15 52367  1     3367  15.8270  0.001832 **
4     14  5079  1    47288 222.3005 4.164e-09 ***
5     13  3332  1     1747   8.2130  0.014199 *
6     12  2553  1      779   3.6618  0.079823 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

```

다중 범주형 변수들(Multiple Categorical Variables)

이전에 범주형 변수(categorical variable)를 이용할 때 계수(coefficients) 중의 한가지를 임의로 0(zero)으로 할당하는 것을 논의했었다. 만약 식 (3.46)의 X 가 0과 1의 추가적인 열(column)을 가지는 것으로 확대되고, 후자를 마지막 범주 수준(last categorical level)의 자료로 할당하였다면 흥미로운 결과가 나타날 것이다. 그 결과로, X 행렬은 열 순위(column rank)가 감소되었고, $X'X$ 는 비정칙(singular)이 된다. 따라서 어떤 독특한 추정량(estimator)도 찾을 수 없다. 그러므로 임의의 할당(arbitrary assignment)은 매우 실용적인 접근법이 된다.

다중범주형변수들(multiple categorical variables)은 단일변수(single variable)에서 논의된 것과 매우 유사한 방식으로 순응될(accommodate) 수 있다. 예를 들어서, 비료(fertilizer)의 질소(x_1), 인(x_2), 칼륨(x_3)의 영향과 담배 수확량(lb/acre) 산출에의 반응을 모형화하기 위한 담배 실험이 계획되었다고 가정해보자. 그러나, 실험군(experimental unit)들은 반드시 2개의 범주형으로 구분(divide) 되어야만 하는 방식이다. 첫째로, 세가지 종류의 비료가 있다. 담배 과학자는 또한 균등한 토양을 포함하는 2개의 농장에서 실험을 수행해야만 한다는 것을 알고 있다. 그러므로 두가지 범주형 변수가 있으며, 두 범주형 변수 중 하나는 세가지 수준이고, 다른 것은 두 가지 수준이다. 결과로서, 범주형변수를 묘사하는 X 행렬의 부분(portion)은 다음과 같다.

	<u>Fertilizer Farms</u>			<u>Fertilizer Farms</u>		
	z_1	z_2	z_3	z_1	z_2	z_3
Fertilizer 1	1 0 1 ⋮ ⋮ ⋮ 1 0 1 0 1 1	Farm 1		1 0 0 ⋮ ⋮ ⋮ 1 0 0 0 1 0	Farm 2	
Fertilizer 2	⋮ ⋮ ⋮ 0 1 1 0 1 1 0 0 1			⋮ ⋮ ⋮ 0 1 0 0 1 0 0 0 0		
Fertilizer 3	⋮ ⋮ ⋮ 0 0 1			⋮ ⋮ ⋮ 0 0 0		
	0 0 1			0 0 0		

이것은 물론 두가지 범주형 변수가 교차 분류된(cross classified) 즉, 각각의 수준이 매번 다른 수준(아마도 같은 횟수로서)으로 발생한다(each level occurs with every other level)는 가정이다. 실험적 디자인의 개념을 포함하지 않으면 다중범주형변수(multiple categorical variables)의 상세한 연구(in-depth study)는 어렵다.

이어지는 다음 절에서, 지시변수(indicator variable)의 개념과 특히 분산분석(analysis of variance)과 공분산분석모형(analysis of covariance models)과 같은 다른 형태의 모형들과의 관계성(relationship)을 간단히 논의할 것이다.

분산분석과 공분산모형분석의 관계(Relationship to Analysis of Variance and Analysis of Covariance Models)

예제 3.10의 상황에서 3개의 polymer가 소독탄(cleaning coal)으로 사용되었다고 가정해 보자. 그러나 그 자료들은 변수로서 pH를 포함하지 않고 있다고 처음에 가정하자. 앞에서와 같이,

각각의 polymer에 대해서 6번의 실험이 수행되었다. 자료는 다음과 같은 형식을 취할 것이다.

$$\text{Polymer 2} \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} & y_{16} \\ y_{21} & y_{22} & y_{23} & y_{24} & y_{25} & y_{26} \\ y_{31} & y_{32} & y_{33} & y_{34} & y_{35} & y_{36} \end{bmatrix}$$

이 실험을 위한 모형을 기술하는 다양한 방식이 있다. 만약 이것을 양적 회귀변수(quatitative regressor)가 없고 3개의 수준(level)에서 하나의 지시변수(indicator variable)를 가진 회귀환경(regression setting)으로 생각한다면, 다음과 같이 기술할 수 있다.

$$y_{ij} = \beta_0 + \alpha_i + \varepsilon_{ij} \quad i = 1, 2, 3 \quad j = 1, 2, \dots, 6$$

그러나, X 행렬과 β 벡터를 자세히 관찰하면 완전 열 순위(full column rank)가 아니다는 것을 알 수 있다.

$$\beta = \begin{bmatrix} \beta_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

$$X = \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 1 & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{array} \right] \quad (3.51)$$

Polymer 1

Polymer 2

Polymer 3

X 의 순위(rank)는 3이다. 첫 번째 열에 더해진 마지막 세 열(the last three columns)에 주목하라. 결과로서, XX' 는 비정칙(singular)이며, 우리는 이 모형이 과다모수화(overparameterized)되었다고 말한다. 이것은 모형이 더 이상 4개의 모수를 필요로 하지 않으며, 실제로 4개의 모수를 허용하지 않는다는 것을 뜻한다. 독자들은 이것을 표준 일원분산분석 모형(standard one-way analysis of variance model)으로 인지해야 한다. 과다모수화(overparameterization)에 대한 해법(solution)은 앞의 절들에서 행했던 범주형 변수들(categorical variables)을 모형화하는 것이다. 즉,

$$y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

여기에서 z_{1i} 는 첫 번째 polymer의 관측값(observation)으로서 1이고, 다른 것은 0이다. 변수 z_{2i} 는 두 번째 polymer의 관측값으로서 1이며 다른 것은 0이다. X 행렬이 다음과 같이 주어졌다.

$$X = \begin{bmatrix} z_1 & z_2 \\ 1 & 1 & 0 \\ 1 & \vdots & \vdots \\ \hline 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ 1 & \vdots & \vdots \\ 1 & \vdots & \vdots \\ 1 & \vdots & \vdots \\ \hline 1 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 1 & \vdots & \vdots \\ 1 & \vdots & \vdots \\ 1 & \vdots & \vdots \\ 1 & 0 & 0 \end{bmatrix} \quad (3.52)$$

Polymer 1

Polymer 2

Polymer 3

이 모형은 양적 회귀변수(quatitative regressor)인 pH를 제외하면, 예제 3.10에 사용된 것과 동일하다. (3.53)의 X 행렬은 $\alpha_3=0$ 이라는 임의의 제약(arbitrary constraint)에 의하여 완전 열 순위(full column rank, XX^T 는 정칙, nonsingular)가 된다. 모수(parameter)에 대한 선형 제약(linear constraints)의 이용은 분산모형(variance model)의 분석에서 일반적이다. 따라서, (3.51)에 의하여 기술된 상황, 즉 일원분산분석모형(one-way analysis of variance model)은 z_1 과 z_2 의 값으로 0과 1이 할당되는, 범주형 변수(categorical variables)에 대한 회귀모형(regression model)으로 볼 수 있다. 사실, 어떤 분산모형분석(analysis of variance model)도 완전 순위(full rank)가 되게 하고 X 행렬의 요소(elements)가 0과 1이 되도록 적절한 임의 제약들(proper arbitrary constraints)을 가하면 회귀모형(a regression model)으로 간주될 수 있다.

이제 4개의 모수모형(parameter model)인 polymer와 pH를 포함하는 예제 3.10의 자료로 되돌아가 보자. 이것은 물론 범주형 변수 회귀(categorical variance regression)의 전통적인 이용(classical use)이다. 세 번째 polymer와 관련된 회귀 계수(regression coefficient)가 0이 되도록 하는 임의 제약(arbitrary constraint) 때문에 그렇게 되어왔다. 이제 만약 이러한 제약을 사용하지 않는다면 모형은 다음과 같을 것이다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 z_{1i} + \beta_3 z_{2i} + \beta_4 z_{3i} + \varepsilon_i \quad (3.53)$$

여기에서 x_{1i} 는 pH이고, z_{1i}, z_{2i}, z_{3i} 는 0이거나 포함된 polymer에 의존하는 어떤 것이다. 다르게 말해서,

$$X = \begin{bmatrix} x_1 & z_1 & z_2 & z_3 \\ \hline 1 & x_{11} & 1 & 0 & 0 \\ 1 & x_{12} & 1 & 0 & 0 \\ 1 & \vdots & 1 & 0 & 0 \\ 1 & \vdots & 1 & 0 & 0 \\ 1 & \vdots & 1 & 0 & 0 \\ 1 & \vdots & 1 & 0 & 0 \\ \hline 1 & \vdots & 0 & 1 & 0 \\ 1 & \vdots & 0 & 1 & 0 \\ 1 & \vdots & 0 & 1 & 0 \\ 1 & \vdots & 0 & 1 & 0 \\ 1 & \vdots & 0 & 1 & 0 \\ \hline 1 & \vdots & 0 & 0 & 1 \\ 1 & \vdots & 0 & 0 & 1 \\ 1 & \vdots & 0 & 0 & 1 \\ 1 & \vdots & 0 & 0 & 1 \\ 1 & x_{1,17} & 0 & 0 & 1 \\ 1 & x_{1,18} & 0 & 0 & 1 \end{bmatrix} \quad \left. \right\} \text{Polymer 1} \quad \left. \right\} \text{Polymer 2} \quad \left. \right\} \text{Polymer 3}$$

이 사례의 X 행렬은 완전열순위 이하(less than full column rank)이다. 행렬 $X'X$ 는 비정칙(singular)이며 (3.53) 모형에서 계수 $\beta_2, \beta_3, \beta_4$ 는 추정가능하지 않다. 독자들은 이것이 부류변수(class variable)와 연속형회귀형변수 또는 공분산(continuous regression type variable or covariance)을 포함하는 공분산분석모형(analysis of covariance model)이라는 것을 인지해야 한다. 자연스러우면서(natural) 확실히 편리한(convenient) 모형(working model)은 $\beta_4=0$ 으로 설정하고 범주형변수회귀(categorical variable regression)로 돌아가는 것이다. 그러나 이러한 제약(constraint)은 유별난(unique) 것이 아니다. 사실 계수 $\beta_2, \beta_3, \beta_4$ 들 간의 선형 제약들 중의 어느 하나(any one of a large class of linear constraints)이면 충분할 것이다.

앞에서 기술된 내용은 선형모형을 공부하는 독자들이 범주형 변수회귀(categorical variable regression)가 ANOVA와 공분산분석모형(analysis of covariance model)과 관련되는 부분에 대해

서 더 잘 이해하도록 도와줄 것이다. 엄격히 말해서 분산분석 모형(anaylsis of variance model)과 공분산분석 모형(analysis of covariance model)은 범주형 회귀 모형(categorical regression model)으로 나타내어질 수 있다.

4. 최선의 모형을 선택하기 위한 기준(Criteria for Choice of Best Model)

표준적인 모형적합(standard model-fitting)을 위해서는 일련의 회귀변수(regressor variables)나 모형항(model terms)에 관한 데이터 처리(자연변수의 변환, transforms of the natural variables)에 능숙한 과학적인 연구자가 종종 필요하다. 딜레마(dilemma)는 모형에 어떤 항(terms)을 포함시킬지에 대한 불확실성에서 발생한다. 다중공선성(multicollinearity)이 존재한다면 각 변수(individual variables)의 중요성에 대하여 연구자가 선입견이나 편견을 가지는 경우, 모형에 어떤 항(terms)을 포함시킬지 결정하기는 더욱 어려워진다. 또한 사용되는 자료(data)의 질 혹은 양이 부적당하거나, 연구자가 채택된 모형(adopted model)으로 무엇을 수행할지 명확한 계획을 가지고 있지 않은 경우도 이러한 결정을 어렵게 할 수 있다.

자료를 기술하는 최상의 근사치(best approximation that describes the data)를 찾으려고 진정으로 노력하는 경우에야 비로소 우리가 만든 모형(postulated model)의 정확성을 추정할 수 있다. 분석가는 선형 모형(linear regression)이 단지 경험적 근사치(empirical approximation)에 불과하다는 사실을 반드시 명심해야 한다. 유능한 모형 개발자(model builder)라면, 많은 자료세트(many data sets)에서 효율면에서 거의 동일한 여러 모형이 적합(fit)될 수 있음을 결국 알게될 것이다. 그러므로 문제는 여러 후보 모형들(a pool of candidate models) 중에서 어느 모형을 선택하는가 하는 점이다.

모형 개발자는 일련의 후보 모형들(a set of candidate models) 중에서 하나를 제대로 선택하는 방법은 ‘그 모형에게 요구되는 바’에 의해서 좌우될 수 있음을 배워야 하며, 실험하는 사람의 목적을 충분히 고려해야 한다. 분석에 앞서 “이 모형으로 무엇을 할 것인가?”라는 질문을 반드시 던져야 한다. 아마도 모형 개발 임무(model-building assignment)는 다음의 목적 중 한가지 이상을 충족시켜야 할 것이다:

1. 자료가 수집된 시스템(system)에 관해 무언가를 알아내고자 할 때(Learn something about the system from which the data are taken). 이는 말하자면 경제학자에게는 계수의 “부호”(“sign” of a coefficient), 혹은 생물학자에 있어서 성장곡선의 기울기, 혹은 공정공학자(a processor engineer)나 담배 화학자(tobacco chemist)에게는 회귀 변수들에(regressor variables) 대한 최적의 반응 요건 같은 것에 지나지 않는다.

2. 어떤 회귀변수가 중요하고 어떤 것은 중요하지 않은지 알고자 할 때(Learn which regressors are important and which are not); 즉, 변수 선택(variable selection) 혹은 변수 선별(variable screening) 작업을 수행하고자 할 때. 예측(prediction)의 필요성과는 별도로 회귀(regression)는 종종 이런 목적으로 사용된다. 변수선별 작업(variable screening)은 종종 모형을 꼼꼼하게 찾는 과정의 전초전(prelude)이 된다. 불행하게도 자료가 다중공선성(multicollinearity)을 내포하는 경우 변수선별(variable screening) 과정에서 종종 문제를 겪게 된다. 명백히 1과 2는 관련이 있다.

3. 예측(Prediction): 후보 모형들(candidate models) 중에서 최상의 예측을 하는 것을

골라내는 것은 종종 매우 어려운 작업이다. 우리는 이 문제를 제 3장에서 언급하였다. 지나치게 복잡한 모형을 선택하고 싶은 유혹을 뿌리치기 어려운 때도 있다. 모형 선택(model selection)을 능숙하게 한다는 것이 실제로 옳은 모형(correct model)을 찾아내는 것을 뜻하는 것은 아니라는 점을 알아야 한다. 사실 옳은 모형(correct model)을 결코 찾지 못할 수도 있다.

우리는 관련 과학 분야의 전문가들의 견해를 무시할 수 없다. (통계학이 확실한 과학적 지식이나 추론을 대신하는 경우는 거의 없다.) 통계학적 절차는 우리를 결론으로 이끄는 매체(vehicle)이며, 과학적 논리가 그 길을 포장하고 있다. 그럼에도 불구하고, 훌륭한 과학자는 적절한 예측식(adequate prediction equation)을 도출하기 위해서는 자료가 뒷받침하는 바(what the data can support)를 고려하여 균형(balance)을 잡아야 함을 명심해야 한다. 자료의 부적절함과 임의 잡음(random noise) 때문에 참 구조(true structure)를 밝혀내지 못하는 때도 있다. 이러한 이유 때문에 경험있는 통계학자와 그 분야의 숙련자가 적절히 협력해야 하는 것이다.

여기까지는, 우리는 전통적인 최소제곱 모형 적합(traditional least squares model fitting)으로 분류되는 원리(fundamentals)에 초점을 맞추었다. 이 장에서는 먼저 예측능(prediction capabilities)의 측면에서 후보 모형들(candidate models)을 비교하는 데 사용되는, 상대적으로 현대적인 기준에 대해 알아볼 것이다. 그리고나서, 이 장의 후반부에서, 1960년대 후반과 1970년대 초반에 걸쳐 지대한 관심과 인기를 얻었던 순차적 변수 선택 알고리즘(sequential variable selection algorithms)에 대하여 논의하고 기술할 것이다.

4.1. 모형간의 비교를 위한 표준기준(Standard Criteria for Comparing Models)

이 절에서는 그 동안 개발되어 온 것들을 살펴보고, 이전에 논의되었던 몇몇 기준(criteria)이 예측(prediction)의 측면에서 어떻게 적용되는지 언급할 것이다.

1. 결정계수(Coefficient of Determination): R^2 . 예전에 언급한 바와 같이, 확실히 R^2 는 모형이 현재 자료(present data)에 적합(fit)되는 정도(capability)를 나타내는 측도(measure)이다. 어떤 새로운 회귀변수를 모형에 추가하여도 R^2 을 감소시키는 결과를 초래하지는 않는다. 비록 최선의 모형을 선택하는 법칙들(rules)과 알고리즘이 있기는 하지만(철저히 R^2 에 기초한), 이 통계량(statistic) 자체는 개념적으로 예측을 위한 것(prediction oriented)이 아니며, 예측 성능(prediction performance)에 기초를 두고 있다. 그러므로 일련의 후보 모형 중에서 최선의 예측 모형(best prediction model)을 고르는데 R^2 를 단독 기준으로 사용하는 것은 권장되지 않는다.

2. 오차분산의 추정값(Estimate of Error Variance): s^2 . 흔히 잔차평균제곱(residual mean square)으로 불리는 이 통계량(statistic)은 다중회귀(multiple regression)와 신뢰경계(confidence bounds) 등 가설검정(hypothesis testing)을 하는데 있어 대단히 중요한 역할을 한다. (제 3장 참조) 또한 예측(prediction)에 있어서 최선의 모형을 선택하고자 할 때 중요한 정보를 제공할 수 있다. 가령, p 개의 항을 가진 모형(p -term model)과 m 개의 항을 가진 모형(m -term model) ($m > p$)을 비교하는 경우를 생각해보면, $s^2(m)$ 이 $s^2(p)$ 보다 클 수 있다. 이는 아마도 m 개의 항을 가진 모형을 채택함으로써 감소되는 잔차 제곱합(residual sum of squares)의 양이 잔차 자유도(residual degrees of freedom)의 손실을 상쇄(counteract)하지 못하기 때문일 것이다. 논리적이고 확실히 간단한 방법은 제일 작은 s^2 값을 가지는 후보 모형(candidate model)을 선택하는 것이다.

3. 수정 R^2 (Adjusted R^2). 식 (3.8)에서와 같이 경쟁 모형(competing models)을 비교하기 위한 통계량(statistic)으로 R^2 을 사용하는 것은 매우 위험할 수 있다. 물론 어떤 새로운 모형 항(term)을 추가하더라도 R^2 는 증가한다(적어도 감소하지는 않는다). 이 장에서는 과적합(overfitting), 즉 과도하게 많은 모형 항(model term)을 포함시키는 것의 위험성에 대하여 배울 것이다. 많은 소프트웨어 패키지는 과적합(overfitting)을 방지하는, R^2 -유사 통계량(R^2 -like statistic)을 계산해낸다. 이를 참고함으로써 오차 자유도(error degrees of freedom)를 희생하면서 중요하지 않은 모형 항(model term)을 포함시키는 잘못을 피할 수 있다.

식 (3.7)과 (3.8)로부터 다음 식을 유도할 수 있다.

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{\text{Total}}}$$

수정 R^2 (adjusted R^2)은 위의 식에서 SS_{Res} 와 SS_{Total} 을 각각에 해당하는 평균제곱(corresponding

mean squares)으로 대체하여 얻는 통계량(statistic)이다. 다시 말하면, 수정 R^2 (adjusted R^2)은 $\overline{R^2}$ 로 표기(denote)할 수 있다.

$$\overline{R^2} = 1 - \frac{\frac{SS_{\text{Res}}}{(n-p)}}{\frac{SS_{\text{Total}}}{(n-1)}} = 1 - \frac{s^2(n-1)}{SS_{\text{Total}}}$$

여기서 s^2 는 모형의 잔차평균제곱(residual mean square)이다. 같은 자료세트(data set)에서 얻어진 후보 모형들(candidate models)을 비교하는데 있어서, 수정 R^2 (adjusted R^2)에 근거한 모형들의 순위(rank order)는 s^2 에 근거한 모형들의 순위(rank order)와 동일하다는 점을 주목해야 한다. 이 점은 명백한데, 모든 후보 모형에 대해서 SS_{Total} 이 동일하기 때문이다.

s^2 를 이용할 때의 더 나은 통찰력(insight)과 정당성(justification)을 얻기 위하여 3.6절의 전개를 다시 보기 바란다. 만일 모형이 저설정(underspecified) 상태라면, 예를 들어 실험자가 다음과 같은 모형을 적합하는데

$$y = X_1\beta_1 + \varepsilon^* \quad (p \text{ parameters})$$

실제 모형은 다음과 같다면

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (m \text{ parameters}; m > p)$$

이 때에, “축소(short)” 모형에 대한 잔차평균제곱(residual mean square)의 기대값(expected value), s_p^2 은 다음과 같이 주어진다(부록 B.2 참조).

$$E(s_p^2) = \sigma^2 + \frac{\beta_2' [\mathbf{X}_2' \mathbf{X}_2 - \mathbf{X}_2' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2] \beta_2}{n-p} \quad (4.1)$$

$n - p$ 는 항(term)이 p 개인 모형(p -term model)의 잔차 자유도(residual degrees of freedom)를 나타낸다. 독자들은 식 (4.1)이 적합성결여 검정(lack of fit test)이 어떤 것인지 기술하기 위해 3장에서 사용되었던 것을 기억할 것이다. 저설정 모형(underspecified model)에서 s^2 은 상향(upward) 편향(bias)되었다는 것을 의미하고 편향(bias)의 정도는 벡터(vector) β_2 로 지칭되는 빠트린 변수(omitted variable)의 계수(coefficients)에 의해 크게 좌우된다. 그러므로, 대체로

모형이 심하게 저설정(underspecified) 되어 있다면 s^2 이 부풀려질 것으로 예상된다. 만일 $m >$

p 이고 s_m^2 이 s_p^2 보다 큰 경우, 항(term)이 p 개인 모형(p -term model)의 저설정(underspecification)으로 인한 편향(bias)은 매우 작다고 할 수 있다; 또는 $\beta_2 = 0$ 이고 저설정(underspecification)이 전혀 없다고 할 수도 있다. 따라서, 후보 모형들(candidate models)의 잔차평균제곱값(residual mean square values)의 비교는, β_2 에서 제거된 모수(eliminated parameters)의 기여(contribution) 정도를 평가하는 경험적인 방법으로 간주될 수 있다. 미학적(esthetic) 관점에서 이 논거에 추가할 사항은 행렬(matrix) $\mathbf{X}_2' \mathbf{X}_2 - \mathbf{X}_2' \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{X}_2$ 는 β_2 의 최소제곱추정량(least squares estimator)의 분산-공분산 행렬(variance-covariance matrix: [σ^2 와는 별도로])의 역(inverse)행렬이라는 점이다. (부록 A.4 참조) 그러므로 남겨진 모수(left out parameters)로 인한 s^2 의 편향(bias)은 무시된 모수(ignored parameters)의 표준화된 형태(standardized form)라 할 수 있다.

위 두 가지 기준(criteria)은 계산하기 쉬우며 이 장 후반의 예제에서 다시 설명할 것이다. 다음 절에서는 덜 전통적이지만 후보 회귀모형(candidate regression model)이 실제로 실행(예측)하도록 하는 기준에 대하여 알아 볼 것이다. 우리는 “최선의 모형 선택(selection of best model)”이라는 말과 모형타당성 검사(model validation)라는 말을 혼용하였다. 모형타당성 검사(model validation)는 독립적인 자료(independent data)에 대하여 모형을 검사해보는 것 즉, 모형이 도출된 자료와 무관한 반응값(response values)을 예측하게끔 함으로써 각 후보 모형(candidate model)을 평가하는 것이다.

4.2. 모형 선택을 위한 상호 검증과 모형 성능의 결정(Cross Validation for Model Selection and Determination of Model Performance)

여러 후보 모형(candidate models) 중에서 반응을 가장 잘 예측하는 것 하나를 선택하는 것은 매우 중요하다. “예측” 이란 미래의 $x = x_0$ 에서 $\hat{y}(x_0)$ 에 대한 $E[y(x_0)]$ 를 추정하는 것이다. 일반적으로 보통잔차(ordinary residuals) 즉, $y_i - \hat{y}_i$ 는 회귀모형이 얼마나 잘 예측하는지 알려주지 못한다. 사실 최소제곱과정(least squares procedure)은 회귀함수(regression function)에서 잔차(residual)가 참 예측오차(true prediction error)보다 작은 특성을 갖도록 설계되었다; 반드시 명심해야 할 것은 \hat{y}_i 는 결코 y_i 로부터 독립적이지 않고 사실은 y_i 에 접근하게끔(drawn to it) 되어있다는 사실이다. 이 잔차(residuals)라는 것은 적합의 질(quality of fit)을 평가하는 측도(measure)이지 미래예측의 질(quality of future prediction)을 평가하는 측도는 아니다.

다음의 병원의 인력관리에 관한 예를 보자. 여기서 y 는 월평균 근무시간, x_1 은 일평균 환자수, x_2 는 월평균 X선 노출량, x_3 는 월평균 병상점유일수, x_4 는 단위면적당 적정근무자수, x_5 는 평균재원일수 일때, 자료는 아래와 같다.

Site	x_1	x_2	x_3	x_4	x_5	y
1	15.57	2463	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1603.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20106	3655.08	180.5	6.15	3503.93
11	96.00	13313	2912.00	60.9	5.88	3571.89
12	131.42	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34703	12446.33	169.4	10.78	11732.17
16	463.70	39204	14098.40	331.4	7.05	15414.94

17	510.22	86533	15524	371.6	6.35	18854.45
----	--------	-------	-------	-------	------	----------

자료에 대하여 완전모형(full model) 즉, 변수(variables) x_1, x_2, x_3, x_4 , 그리고 x_5 를 포함하는 모형을 적합(fitting)한다고 가정해보자, 최소제곱 회귀식(least squares regression)을 반복해보면 다음과 같다.

$$\hat{y} = 1962.948 - 15.8517x_1 + 0.05593x_2 + 1.58962x_3 - 4.21867x_4 - 394.314x_5$$

이 경우 반응(response)은 개월당 근무시간(man-hours per month)이 된다. 위의 적합 모형(fitted model)에서 잔차(residuals)는 다음과 같다.

Site	y_i (man-hours)	\hat{y}_i	$y_i - \hat{y}_i$
1	566.52	775.025	-208.505
2	696.82	740.670	-43.850
3	1033.15	1103.923	-70.773
4	1603.62	1240.496	363.124
5	1611.37	1564.422	46.948
6	1613.27	2151.272	-538.002
7	1854.17	1689.700	164.470
8	2160.55	1736.236	424.314
9	2305.58	2736.989	-431.409
10	3503.93	3681.853	-177.923
11	3571.89	3239.289	332.601
12	3741.40	4353.333	-611.933
13	4026.52	4257.088	-230.568
14	10343.81	8766.748	1576.061
15	11732.17	12237.027	-504.857
16	15414.94	15038.391	376.549
17	18854.45	19320.697	-466.247

자료세트(data set) 중에서 크기가 큰 것 몇 개를 살펴보자. 특히 15,16,17 site를 살펴보기로 한다. 예를 들어 site 17은 예측 반응(predicted response)이 19320.697이고 잔차(residual)는 -466.247이다. 여기서 이들 세 개의 잔차(residuals)가 예측능(quality of prediction)을 평가하는데 얼마나 효과가 있는지, 적합값(fitted values) \hat{y}_i 가 y_i 에 얼마나 접근해 있는지 볼 수

있다. Site 17을 제외시키고 최소제곱회귀(least squares regression)를 하게 되면 어떤 일이 발생할지 생각해보라. 잔차(residual) 즉, 적합오차(fitting error)와 $y_{17} - \hat{y}_{17,-17}$ 값(value) 간에 비교를 할 수 있으며, 여기서 $\hat{y}_{17,-17}$ 는 site 17을 회귀에 포함시키지 않고서 얻은 site 17의 예측값(predicted value)을 말한다. $y_{17} - \hat{y}_{17,-17}$ 는 회귀의 타당성(validation of the regression) 검사를 위해 site 17을 사용하는 경우에 얻게되는 가설적 예측 오차(hypothetical prediction error)를 나타낸다. 이 결과들은 site 15, 16에서도 같으며, 다음과 같다.

Site	Ordinary Residual	$y_i - \hat{y}_{i,-i}$
15	-504.857	-2510.842
16	376.549	2232.496
17	-466.247	-3675.121

그 결과는 다소 극적이다. 적어도 위의 세 sites의 경우에는, 무관한 site(independent sites)에서의 반응을 예측하는 면에서, 이 sites를 사용한 모형 적합(fit)이 분석가의 예측보다 훨씬 우수할 수 있다.

위의 예를 통해서 알 수 있듯이, 모형이 실행될 조건을 가장한(simulate) 일련의 잔차 (a set of residual), 즉, 다음과 같은 유형의 잔차(residual)를 만들어야 할 필요가 있다.

$$r_j = \hat{y}(x_j) - y(x_j) \quad (j = 1, 2, \dots, n^*) \quad (4.2)$$

이때 $y(x_j)$ 와 $\hat{y}(x_j)$ 는 독립적(independent)이다. 식 (4.2)의 오차(error)는 실제로 예측 오차(prediction error)이다. n^* 은 사용 가능한 오차의 개수(the number of these errors available)이다. 물론 이 오차들의 노름(norm), 예를 들어, $\sum_{j=1}^{n^*} |r_j|$ 혹은 $\sum_{j=1}^{n^*} (r_j)^2$ 은 비교목적(comparative purpose)으로 사용될 수 있다. 이제 분명히 알아보아야 할 것은 “이 잔차들(residuals)이 어떻게 개발되었는가?”이다.

자료 분할(Data Splitting)

식 (4.2)에서 주어진 이상적인 예측잔차(prediction residuals)를 만들어 내는 것은 다소 어렵다. 모형의 타당성 검사에서 될 수 있는 한 실제(realism)과 가까워지도록 해야 한다.

모형 개발단계(model-building stage)에서는 그 모형이 어떤 상황에서 작업을 수행하게 될지 알 수는 없다. 이러한 상황은 시간적 예측(forecasting in time), 내적 예측(internal prediction), 또는 심지어 외삽작업(extrapolation)이 될 수도 있다. 교차타당성 입증(cross validation)을 가능하게 하는 한가지 방법으로 자료분할(data splitting)이 있다. 즉, 자료를 2개의 부표본(subsample): 적합표본(a fitting sample)과 타당성 표본(a validation sample)으로 나누는 것이다. Montgomery 와 Peck(1982) 그리고 Snee(1977)의 저서를 참조하라. 자료를 다음과 같이 쓸 수 있다.

$$\left. \begin{array}{ccccc} y_1 & x_{11} & x_{21} & \cdots & x_{k1} \\ y_2 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{n_1} & x_{1,n_1} & x_{2,n_1} & \cdots & x_{k,n_1} \end{array} \right\} Fitting\ sample$$

$$\left. \begin{array}{ccccc} y_{n_1+1} & x_{1,n_1+1} & x_{2,n_1+1} & \cdots & x_{k,n_1+1} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{n_1+n_2} & x_{1,n_1+n_2} & x_{2,n_1+n_2} & \cdots & x_{k,n_1+n_2} \end{array} \right\} Validation\ sample$$

이제 적합표본(fitting sample)을 이용하여 적절한 후보 모형(candidate model)을 적합하고, 계수(coefficients)를 추정할 수 있다. 그리고 적합된 모형(fitted model)을 사용하여 타당성표본(validation sample)에서의 반응(response)을 추정할 수 있다. 쓸모있고 유익한(informative) 예측오차(prediction errors)를 사용하여, 후보 모형 중에서 최상의 것을 결정하는 기준이 되는 다음과 같은 노름(norm)을 만들 수 있다.

$$\sum_{j=n_1+1}^{n_1+n_2} (y_j - \hat{y}_j)^2 \quad \text{그리고}, \quad \sum_{j=n_1+1}^{n_1+n_2} |y_j - \hat{y}_j|$$

\hat{y}_j 는, 첫 번째 n_1 관측값(observations)으로부터의 적합모형(fitted model)을 사용하여 얻은, j 번째 자료점(data point)에서의 추정 반응(estimated response)이다.

비록 자료 분할(data splitting)에는 어려움이 따르지만, 매우 유용하다. 그 이용 목적에 따라 자료를 분할하는 방법을 결정해야 한다. 예를 들어, 자료가 시간 종속적(time dependent)이라면 가장 최근의 관측값들을 타당성표본(validation sample)에 할당함으로써 모형의 예측능력(forecasting capability)에 관한 소중한 정보를 얻는 것이 합당하다. y 의 추정이 꼭 필요한, 회귀 변수(regressor variables)의 특정 부분(region)이 있다고 가정해 보자.

이 잠재적으로 중요한 구간에 속해있는 자료를 타당성표본(validation sample)으로 하는 것이 합당할 것이다. 여기서 한가지 주의할 점은 비록 두 표본의 상대적인 크기에 대하여 보편적으로 받아들여지는 경험적인 지침(rule of thumb)은 없지만, 적합(fit)에 적절한 정도의 잔차자유도(residual degrees of freedom)를 확보하기 위해서는 충분한 수의 관측값이 적합표본(fitting sample)에 포함되어야 한다. 좀 더 관대한 지침은 $n \geq 2p + 20$ 이 되도록 하고 (p 는 모형 모수[parameter]의 개수, n 은 전체 표본의 크기[total sample size]임), 타당성(validation)과 적합(fitting)을 위해 표본을 거의 동일한 크기로 분할하는 것이다. 이것에 대해 좀더 자세히 알고 싶다면 Snee(1977)을 참조하기 바란다.

만일 자료 분할(data splitting)이 최선의 모형(best model) 선택이나 모형 안정성(model stability)과 일반적인 예측성능(general predictive performance)을 연구하기 위한 교차 타당성 검사(cross validation)를 위해 사용된다면, 최종 선택되는 모형은 모든 정보를 담고 있는 전체 자료세트를 이용하여 적합되어야 한다. 자료 분할(data splitting)의 목적은 회귀변수(regressor)의 부분집합(subset)을 연구하거나 함수 형태(functional forms)(그 모두는 자체의 특성이나 성질을 지님)를 알아보기 위함이지 전체 자료 집합에 기인하는 모수(parameter)를 최종 추정(final estimation)하기 위함은 아니다.

PRESS 통계량(PRESS Statistic)

모형 개발(model building)과 관련되는 많은 경우에 현실적으로 타당성 입증(validation)을 위해 자료를 분할하는 일은 드물다.(사실 많은 모형 개발자들은 자료 분할이 실제 이용되고 있을 때 조차도 그 개념을 거부한다.) 자료를 모으는데 드는 비용에 대한 실험자의 딜레마를 이해해야 한다. 확실히, 모든 분야에서 새로 수집된 자료로 자료 분할(data splitting)을 수행하거나 타당성 검사(validation)를 하는 것은 불가능하다.

타당성 검사(validation)에 이용될 수 있으며 자료 분할(data splitting)과 일맥상통하는 흥미롭고 매우 중요한 기준(criterion)은 PRESS 통계량(PRESS statistic)이다. 또 다시, 우리는 식 (4.2)에 있는 유형의 예측오차(prediction error)를 만들어 내는 데에 관심을 가져야 한다. 표본(sample)에서 첫번째 관측값을 제외한 자료 집합(data set)을 고려하자. 그리고 특정 후보 모형(candidate model)의 계수(coefficients)를 추정하기 위해 나머지 $n-1$ 개의 관측값을 사용한다. 그리고나서 다시, 첫번째 관측값을 포함시키고, 두 번째 관측값을 제외한 계수(coefficient)를 추정한다. 이렇게 한번에 한 개의 관측값을 제거하면 후보 모형은 결국 n 회의 적합(fit)을 거치게 된다. 제거된 반응(deleted response)은 매번 추정되고, 결국 n 개의 예측오차(prediction error) 혹은 PRESS 잔차(PRESS residuals) $y_i - \hat{y}_{i,-i}$ ($i = 1, 2, \dots, n$)를 얻게된다. 이 PRESS 잔차들(PRESS residuals)은 $\hat{y}_{i,-i}$ 가 y_i 와 독립적인(independent) 참 예측오차(true prediction errors)이다. 따라서, 이런 방식에서 관측값 y_i 는 모형 적합(fit)과 모형평가(model assessment)에 동시에 사용되지 않으며, 이것이 진정한 타당성 검사(true test of validation)이다. 예측값(prediction) $\hat{y}_{i,-i}$ 는 $x = x_i$ 에서 평가된 회귀함수(regression function)이지만,

y_i 는 계수(coefficients)를 구하는데 사용되지 않고 제외된다. 이를 다음과 같이 표기할 수 있으며,

$$\hat{y}_{i,-i} = \mathbf{x}'_i b_{-i} \quad (4.3)$$

여기에서 b_{-i} 는 i 번째 관측값을 사용하지 않고 얻은 계수집합(the set of coefficients)이다. 따라서 각 후보 모형(candidate model)은 n 개의 PRESS 잔차(PRESS residuals)를 가지며 PRESS(Prediction Sum of Squares)는 다음과 같이 정의된다.

$$\begin{aligned} \text{PRESS} &= \sum_{i=1}^n (y_i - \hat{y}_{i,-i})^2 \\ &= \sum_{i=1}^n (e_{i,-i})^2 \end{aligned} \quad (4.4)$$

그러므로, 최선의 모형(best model)을 선택하려면 PRESS가 제일 작은 모형을 선호해야 할 것이다. PRESS는 중요한데, 이것으로 부터 각각에 대한 적합표본(fitting sample)의 크기가 $n-1$ 인 n 개의 타당성검사(validation)의 형식(form)에 대한 정보를 얻을 수 있다. PRESS를 처음 접하는 독자는 그 계산의 복잡성 때문에 기준(criterion)으로 이용하는 데 저항을 느낄 것이다. 그러나 그 계산은 전혀 복잡하지 않으며 이 절의 후반부에서 논의될 것이다. PRESS 잔차는 예측 능력(prediction capability)을 반영하는 또 다른 R^2 -유사통계량(R^2 -like statistics)을 생성하는 데 사용될 수 있다. 이 통계량은 다음과 같이 구한다.

$$R_{\text{Pred}}^2 = 1 - \frac{\text{PRESS}}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.5)$$

PRESS 잔차의 효용(Utility of the PRESS Residuals)

개별 PRESS 잔차(individual PRESS residuals)는 식 (4.4)의 PRESS 통계량(PRESS statistic)의 계산과정에서의 역할과는 별도로 유용할 수 있다. 개별 PRESS 잔차(individual PRESS residuals)는 회귀의 안정성(stability of regression)에 대한 측도(measure)이고, 분석가로 하여금 어느 자료점(data points)이나 관측값(observations)이 회귀(regression)의 결과에 상당한 영향을 미치는지 알아낼 수 있도록 해준다. 예를 들어, 특정 뺑 포장지의 강도가 반응값(response)인 실험에서 특정 자료점(particular data point)의 보통 잔차(ordinary residual)가 17.75 gm/in^2 이지만 PRESS 잔차(PRESS residuals)는 850.92 gm/in^2 이라 가정하자. PRESS 잔차(PRESS residuals)와

보통 잔차(ordinary residuals)가 이렇게 크게 차이가 나는 경우, 문제가 되는 자료점(data point)은 회귀식 구축(construction of the regression)에 상당히 영향을 미치는 관측값이라는 것을 뜻한다. 이는 이 자료점(data point)을 사용할 경우 적합값(the fitted) \hat{y}_i 를 관측값(observation) y_i 쪽으로 강하게 끌어당기는(근접하게 된다는) 것을 의미한다. 이런 현상은 4.2 절 초반에서 다루었던 병원인력관리 중 3개의 분리된 site (isolated sites)의 경우에서도 볼 수 있다. 보통 잔차(ordinary residuals)와 예측오차(prediction errors) 즉, PRESS 잔차(PRESS residuals) 간의 확연한 구분은 이 세 sites가 매우 영향력 있다는 뜻이다.

PRESS 잔차(PRESS residuals)는 영향력이 큰 관측값(highly influence observations)을 구별해내는 유일한 방법이 아닐 뿐더러 항상 최선(the best)의 방법도 아니다. 제 6장 전체에서 자료점(data point)의 영향(influence)을 구별해내는 진단도구(diagnostics)에 대하여 자세히 설명할 것이다. 제 5장에서는 잔차(residual)에 대해 더 자세히 공부할 것이고, 영향력(influence)이라는 것에 대하여 다시 알아보고 PRESS 잔차(PRESS residuals)의 다른 용도에 대해서도 논의할 것이다.

PRESS의 계산(Computation of Press)

이전에도 언급했듯이 PRESS의 계산은 무척 간단하고 분석자가 반복된 회귀(repeated regression)를 수행하지 않아도 된다. PRESS 잔차(PRESS residuals)는 보통 잔차(ordinary residual)로부터 계산할 수 있다.

$$e_{i,-i} = \frac{y_i - \hat{y}_i}{1 - x_i'(X'X)^{-1}x_i} = \frac{e_i}{1 - h_{ii}} \quad (4.6)$$

따라서 PRESS 는 다음과 같다.

$$PRESS = \sum_{i=1}^n \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

식 (4.6)은 주목할 정도는 아니지만 흥미롭다. 만일 어떤 자료점(data point)이 제외되었을 때 회귀의 행태(behavior)를 결정하기 쉽다는 것은 타당성 검사(validation)를 위해서 뿐만 아니라 진단적 도구(diagnostic tool)로서도 매우 가치있다는 점을 알아야 한다. 적합된 반응값(fitted values of response), 회귀계수(regression coefficient), 분산-공분산 행렬(variance-covariance matrix) 등에 대해 식 (4.6)과 비슷한 공식들이 6장에서 매우 자세히 다루어 질 것이다. 이 모든 결과들의 기원(gensis)은 부록 A.5에 수록되어 있고 특히 식 (4.6)에

대해서는 부록 B.4에 실려있다.

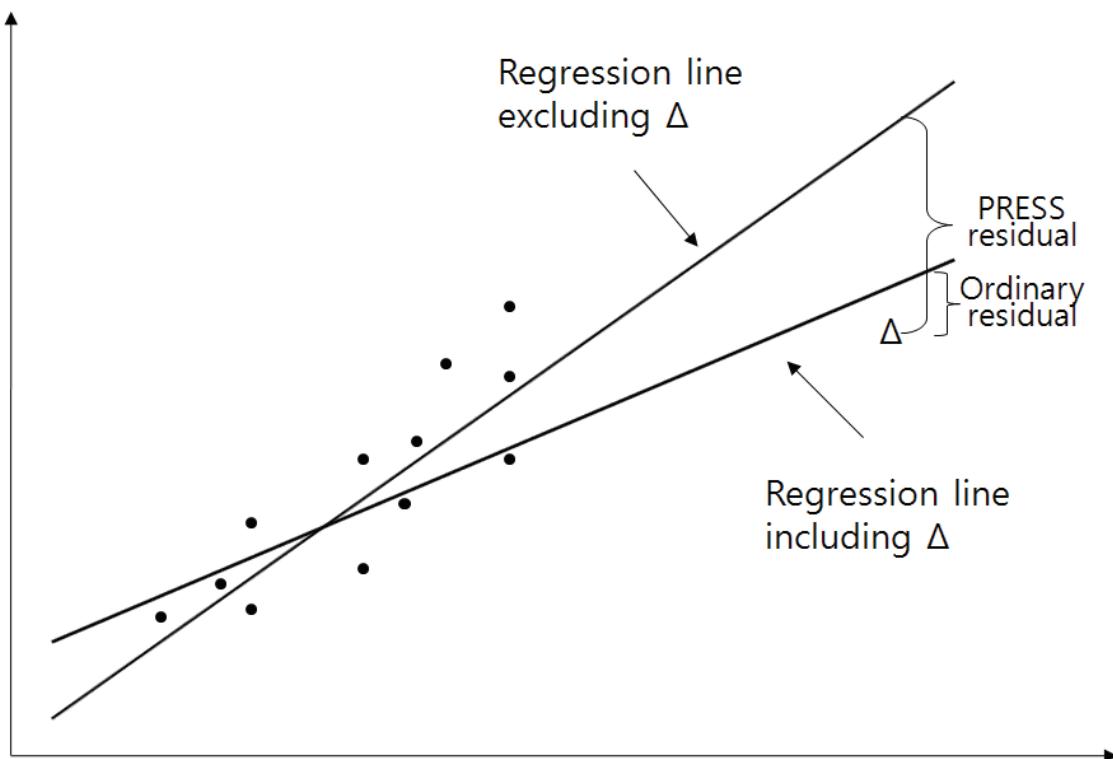
식 (4.6)에 관해서 주목할 점이 한가지 더 있다. h_{ii} 값은 모자행렬(HAT matrix)의 대각(diagonal)이고, σ^2 와 별개로 예측분산(prediction variance)을 나타낸다(3.9절 참조). 따라서, 기대하는 결과의 유형(type)이 걸으로 드러난다. 예측(prediction)이 잘 안되는 (h_{ii} 가 1(unity)에 가까운) 자료점(data points)은 PRESS 잔차(PRESS residual)가 보통 잔차(ordinary residual)와 차이가 많이 나는 자료점(data point)이다. 여기에서 물론 신뢰경계(confidence bounds)와 예측경계(prediction bounds)는 상대적으로 모호(loose)하다. 당연히 그 점은 잠재적으로 영향력이 높은 점(potentially high influence point)이다. 아마도 h_{ii} 가 표준화된 제곱거리 척도(standardized squared distance measure), 점 $x_{1i}, x_{2i}, \dots, x_{ki}$ 에서 점 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ 까지의 거리라고 하면 좀 더 이해가 빠를 것이다. 예를 들어, $k=1$ 인 경우, h_{ii} 를 다음과 같이 쓸 수 있다.(예제 4.7 참조)

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

어느 방향이던 x 가 극단으로 가는 자료점(data point)의 예측값(prediction)이 비교적 불량하고 PRESS 잔차(PRESS residual)가 크기에서 증가하는 경향이 있다.

일반적인 잔차(ordinary residual), PRESS 잔차(PRESS residual) 그리고 영향력에 대한 개념(notion of influence)을 이해하기 위해 아래의 그림이 유용하게 사용된다. Figure 4.1을 보자. 여기에서 회귀변수가 하나이다. 만일 하나의 관찰치(single observation), Δ ,가 제외되면 회귀식의 기울기(slope)와 절편(intercept)은 상당히 변한다. 이 관찰치는 모자행렬의 대각원소의 값이 상대적으로 큰 편에 속한다. 더구나 PRESS잔차가 일반적인 잔차(ordinary residual)보다 크기가 크다.

FIGURE 4.1 A k = 1 data set with a single influential data point



예제 4.1 Hald 자료

다음의 자료는 변수선택과 관련하여 가장 많이 분석에 사용되는 것으로 Hald(1960)에 의하여 조사된 것이다. 반응변수는 1 그램(gram)의 시멘트에서 발생하는 열(calories)이며, 설명변수는 시멘트의 네 가지 원료의 양을 나타내는 것으로 아래와 같다.

y : 시멘트 1그램당 발열량

x_1 : tricalcium aluminate

x_2 : tricalcium silicate

x_3 : tetracalcium alumino ferrite

x_4 : decalcium silicate

Table 4.1는 자료를 나타낸 것이다. 설명변수에 따라 다음 5가지 모형에 대해서 연구해 보기로 하겠다. 다음 table의 적절한 통계량이 모형들을 비교하는 데 사용될 수 있다.

모형	R^2	R^2_{Pred}	SS_{Res}	s^2	PRESS
x_4	0.6745	0.5603	883.8669	80.3515	1194.2182
x_1, x_2	0.9787	0.9654	57.9045	5.7904	93.8825

x_1, x_2, x_4	0.9823	0.9686	47.9727	5.3303	85.3511
x_1, x_2, x_3	0.9823	0.9669	48.1106	5.3456	90.0000
x_1, x_2, x_3, x_4	0.9824	0.9594	47.8636	5.9830	110.3466

이 정보에 기초해서 모형 (x_1, x_2, x_4) 가 시멘트의 발열량에 최선의 모형인 것으로 보인다.
이 변수의 특정 조합에서 최소의 PRESS와 s^2 가 얻어졌다.

Table 4.2에서 table 4.6은 각 모형의 결과값 중 특정 항목을 비교해 보기 위해 제시된 것이다.

Table 4.1 Hald 자료

y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Table 4.2 예제 4.1의 모형 1

Analysis of Variance Table

Response: Y

Df	Sum Sq	Mean Sq	F value	Pr(>F)							
X4	1	1831.90	1831.90	22.799 0.0005762 ***							
Residuals	11	883.87	80.35								

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

(Intercept) 117.5679 5.2622 22.342 1.62e-10 ***

X4 -0.7382 0.1546 -4.775 0.000576 ***

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

Residual standard error: 8.964 on 11 degrees of freedom

Multiple R-Squared: 0.6745, Adjusted R-squared: 0.645

F-statistic: 22.8 on 1 and 11 DF, p-value: 0.0005762

Obs Actual Predict Std error Lower 95% Upper 95% Residual

	Value	Predict	Mean	Mean	
1	78.5	73.27822	5.262207	61.69618	84.86026 5.2217773
2	74.3	79.18352	4.212891	69.91101	88.45603 -4.8835171
3	104.3	102.80470	2.927606	96.36108	109.24831 1.4953050
4	87.6	82.87433	3.617729	74.91176	90.83689 4.7256738
5	95.9	93.20859	2.529029	87.64224	98.77495 2.6914085
6	109.2	101.32837	2.776776	95.21673	107.44001 7.8716286
7	102.7	113.13896	4.466234	103.30885	122.96907 -10.4389603
8	72.5	85.08881	3.296251	77.83381	92.34381 -12.5888116
9	93.1	101.32837	2.776776	95.21673	107.44001 -8.2283714
10	115.9	98.37572	2.561891	92.73704	104.01441 17.5242758
11	83.8	92.47043	2.561891	86.83175	98.10911 -8.6704297
12	113.3	108.70999	3.731550	100.49690	116.92307 4.5900105
13	109.4	108.70999	3.731550	100.49690	116.92307 0.6900105

Table 4.3 예제 4.1의 모형 2

Analysis of Variance Table

Response: Y

Df	Sum Sq	Mean Sq	F value	Pr(>F)
----	--------	---------	---------	--------

X1 1 1450.08 1450.08 250.43 2.088e-08 ***

X2 1 1207.78 1207.78 208.58 5.029e-08 ***

Residuals 10 57.90 5.79

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	52.57735	2.28617	23.00	5.46e-10 ***		
X1	1.46831	0.12130	12.11	2.69e-07 ***		
X2	0.66225	0.04585	14.44	5.03e-08 ***		

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'		
Residual standard error:	2.406	on 10 degrees of freedom				
Multiple R-Squared:	0.9787,		Adjusted R-squared:	0.9744		
F-statistic:	229.5	on 2 and 10 DF,	p-value:	4.407e-09		
Obs	Actual	Predict	Std error	Lower 95%	Upper 95%	Residual
	Value	Predict	Mean	Mean		
1	78.5	80.07400	1.2060356	77.38679	82.76122	-1.5740019
2	74.3	73.25092	1.2314382	70.50710	75.99473	1.0490811
3	104.3	105.81474	0.8297558	103.96593	107.66355	-1.5147396
4	87.6	89.25848	1.1843598	86.61956	91.89740	-1.6584773
5	95.9	97.29251	0.6958245	95.74212	98.84291	-1.3925146
6	109.2	105.15249	0.8164554	103.33331	106.97167	4.0475109
7	102.7	104.00205	1.4473996	100.77704	107.22706	-1.3020510
8	72.5	74.57542	1.1817850	71.94224	77.20860	-2.0754199
9	93.1	91.27549	1.0185111	89.00610	93.54487	1.8245131
10	115.9	114.53754	1.7846157	110.56117	118.51391	1.3624574
11	83.8	80.53567	1.0322647	78.23565	82.83570	3.2643257
12	113.3	112.43724	1.0671170	110.05956	114.81493	0.8627555
13	109.4	112.29344	1.1136881	109.81199	114.77489	-2.8934397

Table 4.4 예제 4.1의 모형 3

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	272.0439	4.934e-08 ***
X2	1	1207.78	1207.78	226.5879	1.094e-07 ***
X4	1	9.93	9.93	1.8633	0.2054
Residuals	9	47.97	5.33		
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 '' 1

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)		
(Intercept)	71.6483	14.1424	5.066	0.000675 ***		
X1	1.4519	0.1170	12.410	5.78e-07 ***		
X2	0.4161	0.1856	2.242	0.051687 .		
X4	-0.2365	0.1733	-1.365	0.205395		

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'		
Residual standard error:	2.309	on 9 degrees of freedom				
Multiple R-Squared:	0.9823,	Adjusted R-squared:	0.9764			
F-statistic:	166.8	on 3 and 9 DF,	p-value:	3.323e-08		
Obs	Actual	Predict	Std error	Lower 95%	Upper 95%	Residual
	Value	Predict	Mean	Mean		
1	78.5	78.43831	1.6657869	74.67004	82.20659	0.0616864
2	74.3	72.86734	1.2144549	70.12005	75.61462	1.4326632
3	104.3	106.19097	0.8424665	104.28518	108.09676	-1.8909669
4	87.6	89.40164	1.1411571	86.82016	91.98311	-1.8016371
5	95.9	95.64375	1.3800907	92.52177	98.76574	0.2562468
6	109.2	105.30178	0.7909415	103.51254	107.09101	3.8982233
7	102.7	104.12867	1.3917943	100.98022	107.27713	-1.4286727
8	72.5	75.59188	1.3565163	72.52322	78.66053	-3.0918781
9	93.1	91.81823	1.0549978	89.43165	94.20480	1.2817747
10	115.9	115.54612	1.8648596	111.32751	119.76472	0.3538826
11	83.8	81.70227	1.3081671	78.74299	84.66155	2.0977319
12	113.3	112.24439	1.0335421	109.90635	114.58242	1.0556137
13	109.4	111.62467	1.1754906	108.96552	114.28381	-2.2246678

Table 4.5 예제 4.1의 모형 4

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	271.2642	4.996e-08 ***
X2	1	1207.78	1207.78	225.9385	1.108e-07 ***
X3	1	9.79	9.79	1.8321	0.2089
Residuals	9	48.11	5.35		

```

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.19363   3.91330 12.315 6.17e-07 ***
X1          1.69589   0.20458   8.290 1.66e-05 ***
X2          0.65691   0.04423 14.851 1.23e-07 ***
X3          0.25002   0.18471   1.354    0.209
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.312 on 9 degrees of freedom
Multiple R-Squared: 0.9823,      Adjusted R-squared: 0.9764
F-statistic: 166.3 on 3 and 9 DF,  p-value: 3.367e-08

Obs Actual Predict Std error Lower 95% Upper 95% Residual
              Value   Predict   Mean   Mean
1     78.5  78.64476 1.5677176 75.09833 82.19118 -0.1447580
2     74.3  72.69032 1.2535868 69.85451 75.52613  1.6096799
3    104.3 105.63580 0.8081341 103.80767 107.46393 -1.3358002
4     87.6  89.21293 1.1384568 86.63756 91.78830 -1.6129282
5     95.9  95.72454 1.3374885 92.69894 98.75015  0.1754552
6    109.2 105.22890 0.7864973 103.44972 107.00808  3.9710971
7    102.7 104.17256 1.3963875 101.01371 107.33141 -1.4725605
8     72.5  75.75427 1.4310280 72.51706 78.99148 -3.2542731
9     93.1  91.55913 1.0007938 89.29518 93.82309  1.5408650
10    115.9 115.68240 1.9119584 111.35725 120.00755  0.2176025
11    83.8  81.91652 1.4228285 78.69786 85.13519  1.8834754
12    113.3 112.45497 1.0253935 110.13537 114.77457  0.8450334
13    109.4 111.82289 1.1251105 109.27771 114.36807 -2.4228886

```

Table 4.6 예제 4.1의 모형 5

Analysis of Variance Table					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	242.3679	2.888e-07 ***
X2	1	1207.78	1207.78	201.8705	5.863e-07 ***
X3	1	9.79	9.79	1.6370	0.2366

X4	1	0.25	0.25	0.0413	0.8441	
Residuals	8	47.86	5.98			
Signif. codes:	0	'***'	0.001	'**'	0.01 '*' 0.05 '.' 0.1 '' 1	
Coefficients:						
		Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	62.4054	70.0710	0.891	0.3991		
X1	1.5511	0.7448	2.083	0.0708 .		
X2	0.5102	0.7238	0.705	0.5009		
X3	0.1019	0.7547	0.135	0.8959		
X4	-0.1441	0.7091	-0.203	0.8441		
Signif. codes:	0	'***'	0.001	'**'	0.01 '*' 0.05 '.' 0.1 '' 1	
Residual standard error:	2.446	on 8 degrees of freedom				
Multiple R-Squared:	0.9824,	Adjusted R-squared:	0.9736			
F-statistic:	111.5	on 4 and 8 DF,	p-value:	4.756e-07		
Obs	Actual	Predict	Std error	Lower 95%	Upper 95%	
				Residual		
	Value	Predict	Mean	Mean		
1	78.5	78.49524	1.8144777	74.31105	82.67943	0.004760418
2	74.3	72.78880	1.4120116	69.53269	76.04490	1.511200700
3	104.3	105.97094	1.8579076	101.68659	110.25528	-1.670937532
4	87.6	89.32710	1.3290552	86.26229	92.39191	-1.727100255
5	95.9	95.64924	1.4627074	92.27624	99.02225	0.250755562
6	109.2	105.27456	0.8618704	103.28708	107.26203	3.925442702
7	102.7	104.14867	1.4819588	100.73127	107.56607	-1.448669087
8	72.5	75.67499	1.5634174	72.06974	79.28024	-3.174988517
9	93.1	91.72165	1.3269571	88.66168	94.78162	1.378349477
10	115.9	115.61845	2.0470658	110.89791	120.33899	0.281547999
11	83.8	81.80902	1.5955553	78.12966	85.48837	1.990983571
12	113.3	112.32701	1.2543585	109.43446	115.21957	0.972989035
13	109.4	111.69433	1.3480154	108.58581	114.80286	-2.294334073

완전모형 (x_1, x_2, x_3, x_4)에 대해서는 x_3, x_4 의 회귀계수가 t 검정 상 통계적으로 유의하지 않다는 것을 주지해야 한다. 또한 모형 (x_1, x_2, x_3)과 (x_1, x_2, x_3, x_4)의 평균값 y 에 대한 신뢰한계(confidence limits)는 모형 (x_1, x_2, x_4)와는 달리 일반적으로 밀접하지(tight) 않다는 것도 알아야 한다. 이는 모형(x_1, x_2, x_4)가 적절하다는 더 나은 증거가 된다.

Table 4.7 모형 (x_1, x_2, x_3)와 (x_1, x_2, x_4)의 잔차와 Press 잔차

	(x_1, x_2)		(x_1, x_2, x_4)	
	잔차	Press	잔차	Press
1	-1.574	-2.1020	0.0617	0.1287
2	1.0491	1.4213	1.4327	1.9807
3	-1.5147	-1.7191	-1.891	-2.1814
4	-1.6585	-2.1887	-1.8016	-2.3841
5	-1.3925	-1.5196	0.2562	0.3987
6	4.0475	4.5741	3.8982	4.4166
7	-1.3021	-2.0402	-1.4287	-2.2443
8	-2.0754	-2.7351	-3.0919	-4.7220
9	1.8245	2.2227	1.2818	1.6201
10	1.3625	3.0278	0.3539	1.0182
11	3.2643	4.0005	2.0977	3.0897
12	0.8628	1.0740	1.0556	1.3202
13	-2.8934	-3.6821	-2.2247	-3.0032

흔히 예측오차(prediction errors)로 불리는(불행하게도 종종 삭제된 잔차[deleted residuals]라고도 불린다) PRESS 잔차(PRESS residuals)는 많은 회귀 관련 패키지(regression computer packages)의 기본적인 출력 사항에 포함된다. Table 4.7는 모형 (x_1, x_2) 과 (x_1, x_2, x_4) 의 보통 잔차(ordinary residuals)와 PRESS 잔차(PRESS residuals)를 나타낸 것이다. 두 모형간의 PRESS 잔차(PRESS residuals)를 비교해보면 두 모형의 성능(performance) 차이를 알 수 있다; 예를 들어 첫 번째 관측값에서, 모형 (x_1, x_2) 은 2.1020만큼 과소추정하는 반면 모형 (x_1, x_2, x_4) 은 같은 첫 번째 관측값에 대해 0.1287만큼 과다추정(overpredict)하고 있다. 사실 모형 (x_1, x_2, x_4) 는 모형 (x_1, x_2) 보다 7개의 관측값에서 Press 잔차값(PRESS residuals)이 더 나은 값을 가지는 것을 볼 수 있다.

예제에 사용된 R-code는 아래와 같다.

```
data<-read.table("d:/data/ex4_1.R",header=TRUE)

g<-lm(Y~X1+X2+X3+X4,data)
r.sq<-summary(g)$r.squared
adr.sq<-summary(g)$adj.r.squared
res<-residuals(g)
sse<-sum(res^2)
ginf<-influence(g)
sig.sq<-(summary(g)$sigma)^2
```

```

sst<-sum(anova(g)$"Sum Sq")
r.pre<-1-press/sst
pre.re<-res/(1-ginf$hat)
press<-sum(pre.re^2)
a.press<-sum(abs(pre.re))
stat.m<-round(c(a.press,press,sig.sq),4)
pre<-predict(g,se=TRUE,interval="confidence")
t4_2<-cbind(data[,1],pre$fit[,1],pre$se.fit,pre$fit[,2:3],res)
g1<-lm(Y~X1+X2,data)
g2<-lm(Y~X1+X2+X4,data)
res1<-residuals(g1)
res2<-residuals(g2)
ginf1<-influence(g1)
ginf2<-influence(g2)
pre.re1<-res1/(1-ginf1$hat)
pre.re2<-res2/(1-ginf2$hat)
t4_7<-cbind(res1,pre.re1,res2,pre.re2)

```

절대 PRESS 잔차(Absolute PRESS Residuals)

상대적으로 큰 PRESS는 하나 또는 몇몇의 큰 PRESS 잔차(PRESS residuals)로 인해 생길 수 있는데, 이것은 종종 문제가 되는 자료점들(data points)이 x 공간 내에서 극단적인(extreme) 값을 갖기 때문에 두드러진다. 명백히 그것들의 예측오차(prediction errors)를 무시할 수는 없다. 그러나, 몇몇의 큰 PRESS 잔차(large PRESS residuals)가 기준(criterion)에 영향을 크게 미치는 것을 피하기 위해 다음을 PRESS의 대안으로 사용할 수 있으며, 그 근거는 PRESS 잔차(PRESS residuals)를 제곱하지 않아도 되기 때문이다.

$$\sum_{i=1}^n |y_i - \hat{y}_{i,-i}| = \sum_{i=1}^n |e_{i,-i}|$$

분석가가 모형들을 판별하기 위해 PRESS를 사용하는 경우라면, 큰 예측오차(large prediction errors)의 위치를 구분(isolation)하기 위해 PRESS 잔차(PRESS residuals) 자체를 관찰해야 한다.

4.3. 예측에 대한 개념적 기준 - C_p 통계량(Conceptual Predictive Criteria - The C_p Statistic)

4.2 절에서 모형의 타당성 정보(model validation information)에 근거한 모형선택 기준(model selection criteria)을 다루었다. 여기에서는 또 다른 선택 기준으로, 예측에 근거한 선택 기준(prediction oriented criterion)을 소개하고자 한다. 이것은 건전한 개념적 토대를 가지고 있으며, 비록 실제 경험적 예측오차 공식(formulations of empirical prediction errors)에 기초하고 있지는 않지만 계산상 쉽게 얻을 수 있다. 물론 최상의 모형(best model)을 선택하고자 할 때 연구자는 여러 가지 판단 기준을 고려하고 싶어하는데, 특히 선택이 까다로울 때 더욱 그러하다.

새로운 기준을 제시하기에 앞서, 자료 분석가가 과소적합(underfit: 상대적으로 중요한 항을 모형에서 무시하거나 빠뜨리는 것) 또는 과대적합(overfit: 약간의 기여를 하거나 전혀 기여도가 없는 항을 모형에 포함시키는 것)을 할 때 직면하게 되는 점들을 살펴볼 것이다. 이 후에 나오는 결과는 ε_i 가 동질적 분산(homogeneous variance) σ^2 과 상관없다는 가정 하에서 최소제곱과정(procedure of least squares)에 의해 얻어진다.

과소 적합의 영향(Impact of Underfitting)

제 3장에서, 모형의 저설정(underspecification)과 그것이 추정량(estimator) s^2 에 미치는 영향에 관하여 다루었다. 식 (4.1)은 모형이 저설정(underspecification) 되는 경우에 발생할 수 있는 s^2 의 양의 편향(positive bias)을 보여준다. 같은 맥락에서, $E(y) = X_1\beta_1 + X_2\beta_2$ 라고 가정해보자. 여기서 β_1 과 β_2 는 각각 p 와 $m - p$ 개의 모수(parameters)를 가진다. 식 (3.26)의 경우처럼, 모형 $y = X_1\beta_1 + \varepsilon^*$ 는 β_1 에 대한 최소제곱 추정량(least squares estimator)

$$b_1 = (X_1'X_1)^{-1}X_1'y$$

으로 적합(fit)된다.

$$\begin{aligned} E(b_1) &= (X_1'X_1)^{-1}X_1'E(y) = (X_1'X_1)^{-1}X_1'[X_1\beta_1 + X_2\beta_2] \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 = \beta_1 + A\beta_2 \end{aligned} \tag{4.7}$$

여기에서 $A = (X_1'X_1)^{-1}X_1'X_2$ 는 별명행렬(alias matrix)로 불리고, $A\beta_2$ 는 저설정(underspecification)에 기인한 b_1 에서의 계수(coefficients)의 편향 벡터(vector of biases)를 나타낸다.

이제, 저설정(underspecification)이 예측반응(predicted response) $\hat{y}(x_{1,0})$ 에 미치는 영향(impact)을 살펴보자. 다음과 같이 정의한다.

$$\hat{y}(x_{1,0}) = x'_{1,0} \mathbf{b}_1 \quad (4.8)$$

여기에서, $x'_{1,0}$ 은 미래의 관심시점(future point of interest)에서 (또는 자료점 $x'_{1,i}$ 에서) 평가되는 항(terms)의 벡터(vector)이다. 기대값(expectation)을 살펴보면 다음과 같다.

$$E[\hat{y}(x_{1,0})] = x'_{1,0} [\beta_1 + A\beta_2]$$

점 $x'_0 = [x'_{1,0}, x'_{2,0}]$ (모든 적절한 모형항[relevant model terms]과 관련 있는)에서 평균 반응(mean response) (즉, 정말로 추정하고자 하는 것)은 $E[y(x_0)] = x'_{1,0}\beta_1 + x'_{2,0}\beta_2$ 이다. 따라서 x_0 에서의 예측의 편향(bias in the prediction)은 다음과 같다.

$$E[\hat{y}(x_{1,0})] - E[y(x_0)] = [x'_{1,0}A - x'_{2,0}]\beta_2 \quad (4.9)$$

따라서, 제곱편향(squared bias)은 이차 방정식(quadratic form)으로 표현될 수 있다.

$$\{\text{Bias}[\hat{y}(x_{1,0})]\}^2 = \beta'_2 [x'_{1,0}A - x'_{2,0}]' [x'_{1,0}A - x'_{2,0}] \beta_2 \quad (4.10)$$

식 (4.1)과 (4.10)으로부터, 저설정 모형(underspecified model)의 s^2 에 있어서의 편향(bias)은 그 자료점들(data points)에서의 예측의 제곱편향의 합(sum of the squared biases in the prediction)으로 표기될 수 있음이 명백해진다. 즉,

$$E(s_p^2) = \sigma^2 + \frac{1}{n-p} \sum_{i=1}^n \{\text{Bias } \hat{y}(x_i)\}^2 \quad (4.11)$$

여기에서, 과소적합된(underfitted) p 개의 항이 있는 모형 (p -term model)에 대한

오차평균제곱(error mean square)을 나타내기 위해 s_p^2 을 다시 사용한다.

저설정 모형(underspecified model)이 $\hat{y}(x_{1,0})$, 회귀계수(regression coefficients), 오차분산의 추정값(estimate of error variance)과 같은 중요한 양(quantities)에 고정된 편향(fixed bias)을 유발한다는 것은 명백하다. 식 (4.11)은 모형선택 기준(model selection criterion)을 개발하는데 있어서 특히 중요한데, 다음 절에서 다루어질 것이다. 우리가 과소적합(underfit) 하였을 때, 즉, 중요한 변수(variables)를 무시하였을 때, 우리는 무시된 변수들(ignored variable)에 의해 설명되는 변이(variation)를 잔차제곱합(residual sum of squares)에 묻어놓게 되며 따라서, 잔차평균제곱(residual mean square)을 부풀리게 된다. 이러한 부풀림(inflation)은 자료점(data points)에서의 예측 편향(bias in prediction)을 반영하는 것으로 간주할 수 있다.

예제 4.1과 모형 (x_2, x_3) 는 과소적합 모형(underfitted model)에서 s^2 에 있어서의 편향(bias)을 잘 보여준다. Table 4.3와 4.4에 있는 (x_1, x_2) 와 (x_1, x_2, x_4) 의 비교는 변수가 2개인 모형(two-variable model)이 과소적합 모형(underfitted model)이라는 것을 확실히 시사한다. 이것은 두 모형의 s^2 값을 비교해보면 드러난다. 다음으로 우리는 과설정(overspecification), 더 정확히 말하자면, 과대적합(overfitting)의 영향(impact)에 대하여 다룰 것이다.

과대 적합의 영향(Impact of Overfitting)

저설정 모형(underspecified model)이 중요 추정값(important estimated quantities)의 편향(bias)을 초래하는 반면, 과대적합(overfitting, 즉, 거의 또는 전혀 기여하지 않는 항을 모형에 포함)은 더 단순한 모형에 비해 분산(variances)이 더 커지는 결과를 낳는다. 중요한 분산(variance) 결과는 다음과 같이 요약될 수 있다.

1. 최소 제곱 추정량(least squares estimator) $b_0, b_1, b_2, \dots, b_k$ 와 함께 회귀변수(regressors) x_1, x_2, \dots, x_k 를 포함하는 모형에서, 만약 회귀변수(regressor variable) x_{k+1} 이 모형에 더해져서 새로운 회귀 계수(regresscofficients) $b_0^*, b_1^*, \dots, b_k^*, b_{k+1}^*$ 을 만들어 내면, 다음이 성립한다.

$$Var b_i^* \geq Var b_i \quad (i = 0, 1, 2, \dots, k) \quad (4.12)$$

2. 방금 기술된 상황에서 두개의 회귀 예측변수(regression predictors) $\hat{y}_1 = \sum_{i=0}^k b_i x_i$ 와 $\hat{y}_2 = \sum_{i=0}^{k+1} b_i^* x_i$ 를 살펴보자. 여기서 후자는 추가적인 회귀변수(additional regressor)가 기여하는 바를 포함하고 있다.

$$x'_0 = (1, x_{1,0}, x_{2,0}, \dots, x_{k,0}, x_{k+1,0}),$$

$$Var \hat{y}_1(x_0) \leq Var \hat{y}_2(x_0) \quad (4.13)$$

항목(items) 1과 2의 결과는 매우 중요하고, 과소적합(underfitting)과 과대적합(overfitting) 사이의 절충(tradeoff)을 설명하는데 도움이 된다. 너무 단순한 모형에서는 편향된 계수(biased coefficients) 및 편향된 예측(biased prediction)이 문제가 될 수 있고, 반면에 지나치게 복잡한 모형에서는 계수(coefficients) 및 예측값(prediction)의 분산(variance)이 커질 수 있다. 그러므로, 많은 경우에 적절한 모형(proper model)은 편향된 모형(a biased model)과 분산이 큰 모형(a model with heavy variance) 사이의 타협점(compromise)이 될 것이다. 주변변수(marginal variables)를 추가함으로써 발생하는 분산(variance)의 크기는 미심쩍은 회귀변수(questionable regressor)에 의해 유발된 다중공선성(multicollinearity)에 크게 의존한다.

1과 2에서의 결과를 증명하고자 한다면, Rao (1971)를 참고하시오,

적절한 타협과 맬로우즈 C_p (The Proper Compromise and Mallows' C_p)

과소적합(underfitting)과 과대적합(overfitting)에 관하여 토의해 보면 모형개발자(model builder)나 모형선택자(model selector)의 실제 임무가 무엇인지 명확히 알 수 있다. 우리는 후보들로부터 회귀변수(regressors)의 적절한 부분집합(proper subset) 또는 적절한 함수 형태(proper function form)를 선택함으로써 적절한 균형(balance)을 얻도록 해야 한다. 예를 들어, 예측(prediction)에 대한 합리적인 기준(reasonable criterion)은 예측값의 평균제곱오차(mean squared error of prediction)이다. 예로써 x_0 에서 특정 후보모형(candidate model)에 대한 예측값(prediction)인 $\hat{y}(x_0)$ 의 평균제곱오차(mean squared error)는 다음과 같이 주어진다.

$$MSE[\hat{y}(x_0)] = Var \hat{y}(x_0) + [E\hat{y}(x_0) - Ey(x_0)]^2 \quad (4.14)$$

물론 이것은, 같은 기준(criterion)에서 편향(bias)과 분산(variance)을 통합해 놓은 것이며, 실제로 다음의 최종 형태(definitive form)로 표시된다.

$$MSE[\hat{y}(x_0)] = E[\hat{y}(x_0) - Ey(x_0)]^2 \quad (4.15)$$

식 (4.15)는 개념적으로 이치에 맞지만, 그것이 정말로 사용될 수 있을까? 이전에 “과소적합의 영향(impact of underfitting)”으로 제시되었던 식을 살펴보자. 식 (4.9)을 사용하여 (무시된 모수 벡터[ignored parameter vector] β_2 를 가지는 저설정 모형[underspecified model]에

대해서) 다음과 같이 쓸 수 있다.

$$MSE[\hat{y}(x_{1,0})] = \sigma^2 x'_{1,0} (X'_1 X_1)^{-1} x_{1,0} + [(x'_{1,0} A - x'_{2,0}) \beta_2]^2 \quad (4.16)$$

우리는 식 (4.16)에 있는 기준(criterion)을 직접적으로 적용할 수는 없다. 벡터(vector) β_2 는 알려져 있지 않고, 예측(prediction)에 관한 여러 개념적 기준(conceptual criteria)이 그랬던 것처럼, 식 (4.16)의 기준(criterion)은 회귀변수 위치 벡터들(regressor location vectors)인 $x_{1,0}$ 와 $x_{2,0}$ 의 함수라는 문제가 있다.

$MSE \hat{y}(x_0)$ 의 위치 의존성(location dependency) 문제를 극복하기 위해서, 자료점(data points)의 적합값(fitted value) $\hat{y}(x_i)$ 의 평균제곱오차(mean squared error)의 합을 생각해보자.

$$\sum_{i=1}^n \frac{[MSE \hat{y}(x_i)]}{\sigma^2} = \sum_{i=1}^n \frac{\{[Var \hat{y}(x_i)] + [Bias \hat{y}(x_i)]^2\}}{\sigma^2} \quad (4.17)$$

(4.17)에 있는 양(quantity)은 표준화된 전체 오차(standardized total error)로 간주될 수 있다. 이것이 후보모형(candidate model)의 외삽(extrapolation)이나 내삽(interpolation) 능력을 반영하지 않는다는 것은 분명하다. 그렇지만, 편향(bias)과 분산(variance)간에 실행가능한 균형(workable balance)을 이루는 식 (4.17)의 양(quantity)에 관한 추정값(estimate)을 얻을 수는 있다. (4.17)의 구성 요소들을 살펴보자. 완전히 일반화하기 위해서, 의문시 되는 후보모형(candidate model)이 p 개의 모수(parameters)를 포함하며, “참” 모형(true model)이 벡터(vector) β_2 에 의해 묘사되는 $m-p$ 개의 추가적인 모수(additional parameters)를 포함하고 있다고 가정해보자. 따라서 i 번째 자료(data point)에서의 예측값(prediction)을 다음과 같이 표시할 것이다.

$$\hat{y}(x_i) = x'_{1i} b_1$$

여기에서, $b_1 = (X'_1 X_1)^{-1} X'_1 y$ 이다. 그러므로, 우리 기준(criterion)의 분산 부분(variance portion)을 다음과 같이 쓸 수 있다.

$$\sum_{i=1}^n \frac{Var \hat{y}(x_{1i})}{\sigma^2} = \sum_{i=1}^n x'_{1i} (X'_1 X_1)^{-1} x_{1i} \quad (4.18)$$

식 (4.18)을 변형시켜 계산하기 매우 편리한 다음의 형태를 얻을 수 있다.

$$\begin{aligned}
 \sum_{i=1}^n x'_{li} (X'_1 X_1)^{-1} x_{li} &= \sum_{i=1}^n \text{tr } x'_{li} (X'_1 X_1)^{-1} x_{li} \\
 &= \sum_{i=1}^n \text{tr } x_{li} x'_{li} (X'_1 X_1)^{-1} \\
 &= \text{tr} \sum_{i=1}^n x_{li} x'_{li} (X'_1 X_1)^{-1}
 \end{aligned} \tag{4.19}$$

(4.19)를 자세히 살펴보면 $\sum_{i=1}^n x_{li} x'_{li}$ 는 사실 $X'_1 X_1$ 행렬이다. 그러므로, 표준화된 전체 평균 제곱오차(standardized total mean squared error)의 분산 부분(variance portion)은 다음과 같다.

$$\sum_{i=1}^n \frac{\text{Var } \hat{y}(x_{li})}{\sigma^2} = \text{tr } I_p = p$$

이것은 다소 이상한 결과로 보일 수도 있다. 자료 위치(data locations)에 걸쳐 합해진 예측분산(prediction variance)은 σ^2 과 별개로 모수의 개수(number of parameters)와 같다. 독자는 이 전개(development)를 추가적인 문제 즉 모자행렬(HAT matrix)의 성질(property) 1 (3.9 절에 있는)의 증명(proof)을 푸는 과정으로 간주해야 한다. 양(quantity) $\sum_{i=1}^n x'_{li} (X'_1 X_1)^{-1} x_{li}$ 은 의문시되는 후보 모형(candidate model)에 대한 모자행렬 (HAT matrix)의 대각합(trace)이다.

식 (4.17)에 있는 기준(criterion)의 편향 부문(bias portion)으로 넘어가서, 우리는 $\sum_{i=1}^n [\text{Bias } \hat{y}(x_{li})]^2$ 을 살펴볼 필요가 있으며, 이 양(quantity)은 추정 가능하다. 식 (4.11)을 살펴보자. 우리는 저설정 모형(underspecified model)의 잔차분산(residual variance)의 추정값(estimate) s^2 이 다음의 양(quantity) 만큼 편향되어(biased) 있다는 것을 배웠다.

$$\sum_{i=1}^n \frac{[\text{Bias } \hat{y}(x_i)]^2}{n-p}$$

결과로써, 만약 σ^2 이 알려져 있다면, 식 (4.17)의 양(quantity)의 추정값(estimate) (그 추정값은 C_p 통계량으로 불린다)은 아래와 같다.

$$C_p = p + \frac{(s^2 - \sigma^2)(n - p)}{\sigma^2} \quad (4.20)$$

식 (4.20)을 표현하는 여러 다른 방법이 있다. 어떤 이는 그것을 SS_{Res} 로 표현하는 것을 선호한다. 하여튼, 그것은 ‘분산 + 편향’(‘variance + bias’)을 표현하고, 만약 σ^2 의 독립적인 추정값(estimate), 즉 $\hat{\sigma}^2$ 를 얻을 수 있다면, C_p 통계량(C_p statistic)은 모형들을 식별하는 기준(criterion)으로 매우 유용해질 수 있다. P 개의 모수(parameters)를 가지는 회귀모형(regression model)의 C_p 를 다음과 같이 나타낼 수 있다.

$$C_p = p + \frac{(s^2 - \hat{\sigma}^2)(n - p)}{\hat{\sigma}^2} \quad (4.21)$$

우리는 C_p 값이 최소인 후보모형(candidate model with the smallest C_p value)을 선호한다.

어떤 모형의 C_p 값을 판정하는 합리적인 노름(norm)은 $C_p = p$, 즉, 모형이 추정된 편향(estimated bias)을 포함하지 않음을 시사하는 값이다. 즉, \hat{y} 에 있어서의 모든 오차(error)는 분산(variance)이고, 모형이 저설정(underdspecification)되어 있지 않은 경우이다. 물론, 추정값(estimate) $\hat{\sigma}^2$ 의 미심쩍은 성질 때문에 종종 분명한 해석이 어렵다. 많은 실제 상황에서, 완전모형(full model) 또는 가장 완전한 모형(most complete model)에 대한 잔차평균제곱(residual mean square)이 이 추정값(estimate)으로 사용된다. 완전한 모형(complete model)의 잔차평균제곱(residual mean square)이 후보모형들(candidate models) 중에서 최소(the smallest)의 σ^2 추정값(estimate of σ^2)일 필요는 없기 때문에, 몇몇 후보모형들(candidate models)에서 식 (4.12)이 $C_p < p$ 를 초래할 가능성이 꽤 있다.

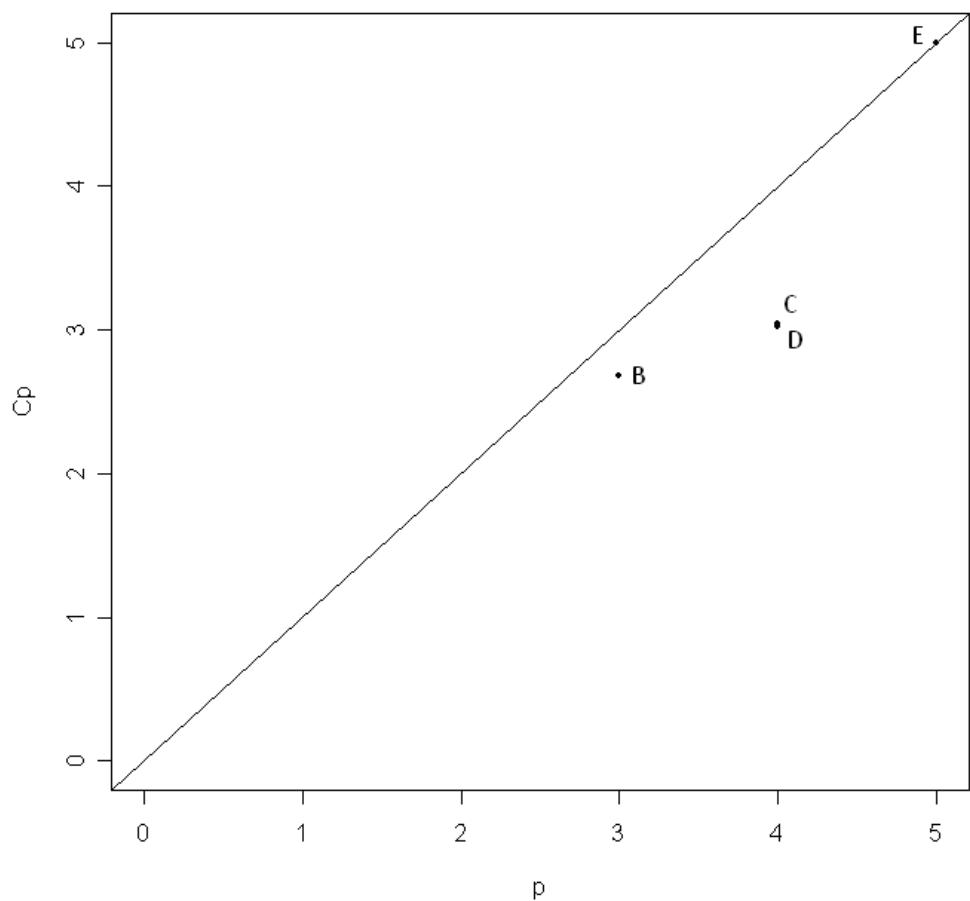
종종 다양한 후보모형들(candidate models)에 대한 C_p 값이 노름(norm)을 나타내는 $C_p = p$ 직선으로 플롯(plot)에 그려질 수 있다. 심하게 편향된 모형(heavily biased model)에서, p 보다 훨씬 큰 C_p 값이 발생한다.

Table 4.8은 모형에 따른 C_p 통계량(C_p statistic)을 나타낸 것이다. 모형 A가 가장 바람직하지 않는 것으로 보인다. 모형 B, C, D가 합당한 후보모형(reasonable candidates)으로 판단되어지며 이 중 모형 B가 모수(parameter)의 값과 가장 가까우므로 가장 합당한 모형이라 할 수 있다.

Figure 4.2는 전형적인 C_p 플롯을 나타낸 것이다. 단, A의 경우 C_p 값이 다른 모형들과 너무 차이나서 나타내지 않았다. B, C와 D 모두 분산선(variance line) 아래에 존재한다. 가장 아래에 있는 모형 D는 가장 나쁜 성능을 보이며, 모형 C와 거의 같이 나타나 후보모형으로 B모형이 적합하게 보인다.

Figure 4.1 C_p against p plot.

Cp against p plot



이 절에서 우리는 C_p 통계량(C_p statistic)이 모형 판별의 목적으로 사용되는 예를 살펴볼 것이다. 이 예에서는 table 4.1에 있는 Hald 자료를 고려한다. 후보모형(candidate model)의 s^2 값에 독립적인 추정값(estimate) $\hat{\sigma}^2$ 을 사용하는 것은 항상 바람직하다. 불행히도 그러한 추정값이 항상 이용 가능하지는 않다. 따라서, 자료분석가들은 실제로 $\hat{\sigma}^2$ 로서 완전한 모형(complete model)의 잔차평균제곱(residual mean square)을 흔히 사용한다. Hald 자료의 경우 $\hat{\sigma}^2 = 5.983$ 이다. 모형 (x_1, x_2) ($p = 3$)에서 C_p 는 다음과 같다.

$$C_p = 3 + \frac{(5.7904 - 5.983)(13 - 3)}{5.983} = 2.6782$$

분명히, 이 값은 3.0보다는 작으므로 편향된 모형(biased model)으로 보인다. 유사한 계산에 의해서 모형 (x_1, x_2, x_4) 에 있어서 $C_p = 3.0182$ 이고 $p = 4$ 이다. 물론, 완전 모형(full model)에서 C_p 는 5.0으로 고정되어 있다. 결과를 보면, C_p 통계량(C_p statistic)의 관점에서 가장 선호되는 모형은 모형 (x_1, x_2) 이다. PRESS 통계량(PRESS statistics)의 사용은 모형 (x_1, x_2, x_4) 의 다른 모형을 선호하는 것을 볼 수 있다.

Table 4.8 모형별 Cp값 비교

모형	p	Cp
A : x_4	2	138.7308
B : x_1, x_2	3	2.6782
C : x_1, x_2, x_4	4	3.0182
D : x_1, x_2, x_3	4	3.0413
E : x_1, x_2, x_3, x_4	5	5

예제에 사용된 R-code는 아래와 같다.

```

data<-read.table("c:/data/ex4_1.R",header=TRUE)
g1<-lm(Y~X4,data)
g2<-lm(Y~X1+X2,data)
g3<-lm(Y~X1+X2+X4,data)
g4<-lm(Y~X1+X2+X3,data)
g5<-lm(Y~X1+X2+X3+X4,data)
s1<-summary(g1)$sigma^2
s2<-summary(g2)$sigma^2
s3<-summary(g3)$sigma^2
s4<-summary(g4)$sigma^2

```

```

s5<-summary(g5)$sigma^2
ss<-c(s1,s2,s3,s4,s5)
n<-length(data[,1])
p<-c(2,3,4,4,5)
cp<-p+(ss-s5)*(n-p)/s5

```

과대적합된 모형을 위한 잔차평균제곱의 특성(*Properties of Residual Mean Square for an Overfitted Model*)

우리는 식 (4.11)에서 적합된 모형이 편향된 경우에 즉, 분석가(analyst)가 과소적합(underfitting) 할 때 잔차평균제곱(residual mean square)이 상향(upward) 편향된다는 것을 배웠다. 모형 판단자(model discriminator)로서 s^2 의 역할에 대해서 더 많이 알기 위해서, 과대적합된 모형(overfitted model)을 사용할 때 그것의 특성을 살펴보는 것이 도움이 된다. 아래의 모형을 가정해보자.

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (m \text{ parameters}) \quad (4.22)$$

여기에서 사실 $\beta_2 = 0$ 이고, 따라서 m 모수 모형(m parameter model)은 과대적합된 모형(overfitted model)이다.

$$X = [X_1 : X_2]$$

과대적합된 모형(overfitted model)에 대한 잔차평균제곱(residual mean square)은 다음과 같이 주어진다.

$$s_m^2 = \frac{y'[I - X(X'X)^{-1}X']y}{n - m}$$

이것의 기대값(expected value)은 아래와 같다.

$$\begin{aligned}
E(s_m^2) &= \frac{1}{n-m} E[y'[I - X(X'X)^{-1}X']y] \\
&= \frac{1}{n-m} \{\sigma^2 tr[I - X(X'X)^{-1}X'] + [E(y)]'[I - X(X'X)^{-1}X'][E(y)]\} \\
&= \frac{1}{n-m} \{\sigma^2(n-m) + \beta_1' X_1'[I - X(X'X)^{-1}X'] X_1 \beta_1\}
\end{aligned} \tag{4.23}$$

이러한 전개의 자세한 내용은 부록 B.2에 나와 있는 것과 같다. 식 (4.23)으로부터, 과대적합된 모형(overfitted model)의 잔차평균제곱(residual mean square)의 기대값(expected value)은 다음과 같다.

$$E(s_m^2) = \sigma^2 + \frac{1}{n-m} \beta_1' X_1'[I - X(X'X)^{-1}X'] X_1 \beta_1$$

$$X'[I - X(X'X)^{-1}X'] = 0, \quad X_1'[I - X(X'X)^{-1}X'] = 0 \quad \text{이므로, 따라서}$$

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

연구자가 과대적합(overfit)한다 하여도, 즉, 0인 모형 항(model terms)을 포함한다 하여도, 잔차평균제곱(residual mean square)은 σ^2 에 대해 편향되지 않는다. 그러나, 추정량(estimator)은 “정확한” 모형(correct model)의 적합(fitting)에서 계산된 오차평균제곱(error mean square)보다도 더 작은 자유도(degrees of freedom)를 가진다.

정규이론(normal theory)의 경우 다음과 같이 된다.

$$\frac{s_m^2(n-m)}{\sigma^2} \sim \chi_{n-m}^2$$

Graybill (1976)을 참고하면, 그 결과로

$$s_m^2 \sim \frac{\sigma^2 \chi_{n-m}^2}{n-m}$$

이고

$$Var s_m^2 = \frac{\sigma^4}{(n-m)^2} Var \chi_{n-m}^2$$

χ_v^2 확률변수(random variable)의 분산(variance)은 $2v$ 이고, 여기에서 v 는 자유도(degrees of freedom)이다. 따라서

$$Var(s_m^2) = \frac{2\sigma^4}{n-m}$$

결과적으로, 감소하는 오차자유도(error degrees of freedom) (과대적합된 모형의 경우에 $n-m$)는 s^2 의 분산(variance)이 더 커지도록 한다. 따라서, 과대적합된 모형(overfitted model)에서 σ^2 의 추정량(estimator)의 분산(variance)은 정확한 모형(correct model)으로부터 계산된 σ^2 의 추정량(estimator)의 분산(variance) 보다 더 크다. 이것으로부터 우리는 경쟁하는 모형들 간의 식별자(discriminator)로서 그리고 C_p 통계량(C_p statistic)의 중요한 요소(component)로서 잔차평균제곱(residual mean square), s^2 의 유용성을 알 수 있다.

4.4. 순차적 변수 선택 과정(Sequential Variable Selection Procedures)

1960년대 초반에 컴퓨터 영역은 급성장하였고, 이러한 시대 상황 속에서 순차적 F -검정(sequential F -test) – 가능성이 있는 수많은 변수들(variables)중에서 회귀 변수들(regressor variables)의 합당한 부분집합(reasonable subset)에 효율적으로 도달하기 위해 만들어진 과정 -에 기초한 여러 가지 체계적인 과정이 탄생하였다. 그 과정은 상대적으로 작은 수의 부분집합 회귀들(subset regressions)이 컴퓨터에서 실제로 실행되도록 고안되었다.

이러한 순차적 알고리듬(sequential algorithms)은 필요에 의해서 효율적인 계산이 되도록 발전하였고, 또한 2^k 개의 모든 부분집합 모형(subset models)에서 정보를 모으는 것이 힘들었던 시절로부터 개발되어 왔음을 분석하는 유념해야 한다. 최근에는 모든 부분집합들(all subsets)에서 적어도 어느 정도의 중요한 결과를 얻기 위한 신속 계산 알고리듬(computationally swift algorithms)에 근거한 많은 소프트웨어 패키지가 이용 가능하다. (이후에 이러한 개발의 일부가 인용될 것이다.) 이 사실은 다중공선성(multicollinearity)이 종종 혼란을 야기한다는 점과 함께, 드문 경우를 제외하고는 순차적 알고리듬(sequential algorithms)이 실용적이지 않다는 것을 시사한다. 그럼에도 불구하고, 그것은 최소 제곱 회귀(least squares regression)의 중요한 부분이고, 많은 분석가들에 의해서 사용된다 (아마도 그것이 대부분의 표준 회귀 컴퓨터 패키지에 나와있기 때문에). 이러한 순차적 알고리듬(sequential algorithms)은 다양한데, 3개의 일반적인 유형을 서술하였다.

전진 선택(Forward Selection)

연속적이고 체계적으로, 각 방법은 F 검정(F -tests)에 근거한 기존 모형으로부터 회귀 변수들(regressor variables)을 추가하거나 혹은 삭제한다. 전진 선택(forward selection)의 경우에, 초기 모형(initial model)은 단지 상수 항 만을 포함한다. 그 과정은 어떤 한 회귀변수(single regressor)의 최대 R^2 를 만드는 변수(variable)를 포함하도록 선택한다. 이 회귀변수(regressor)를 x_1 이라고 하자. 두 번째의 회귀변수(regressor)로, x_1 의 존재하에 R^2 의 최대 증가를 일으키는 것이 선택된다. 이 회귀변수(regressor)를 x_2 라고 하자. 이것은 최대의 부분 F (partial F)를 (x_1 과 한 조를 이루는 변수를 포함하는 모형이라는 의미에서 부분적인) 가지는 회귀변수(regressor)를 선택하는 것과 동등하다는 점을 주목하라. 따라서, 이 가설의 경우의 2단계에서

$$F = \frac{R(x_2|x_1)}{s^2(x_1, x_2)}$$

이것은 두 번째 단계에서 최대의 F 이다. $s^2(x_1, x_2)$ 표기는 회귀변수(regressors) x_1, x_2 와 관련된

모형의 잔차평균제곱(residual mean square)이다. 우리는 일반적으로 불완전한 모형(incomplete model) 즉, 그 단계에 존재하는 모형으로부터의 분산(variance)의 잔차 추정값(residual estimate)을 계산한다.

상기 과정은 어떤 단계에서 포함시킬 후보 회귀변수(candidate regressor)가 미리 선택된 F_{IN} (preselected F_{IN})을 초과하지 않을 때까지 (또는, 물론, 모든 회귀변수 들이 포함될 때까지) 계속된다.

단계적 회귀 (Stepwise Regression)

단계적 회귀(stepwise regression)는 각 단계의 선택(selection)시 모형에 현존하는 모든 회귀변수들(regressors)이 부분 F -검정(partial F -test)을 통해 평가된다는 점에서, 전진회귀(forward selection)의 중요한 변형이라 할 수 있다.

미리 선택된(preselected) F_{OUT} 의 임계값(critical value)이 사용된다. 따라서, 각 단계에서 한 회귀변수(regressor)가 유입될 수 있고, 다른 것이 제거될 수 있다. 이것의 논리적 근거는 매우 명백하다. 다중공선성(multicollinearity)은 그 과정의 초기 단계에서 중요한 후보였던 회귀변수(regressor)를 별 가치가 없는 회귀변수(regressor)로 만들 수 있다. 따라서, 유입(entry) 후 모든 단계에서, 한 변수는 계속 실행되거나 또는 제거되어야만 한다. 이 과정은 F_{IN} 에 근거하여 더 이상의 추가적인 회귀변수(regressor)가 유입될 수 없고, 모형에 있는 어떤 회귀변수(regressor)도 F_{OUT} 에 근거하여 제거될 수 없을 때 중지된다.

후진 제거(Backward Elimination)

위의 두 과정은 모형 내에 회귀변수(regressors)가 없는 상태에서 시작되는 데에 비해, 후진제거(backward elimination)는 모든 회귀변수(all regressors)들이 존재하는 상태에서 시작하여 한번에 하나씩 제거해 나간다. 최초로 제거되는 것은 R^2 에 있어 최소의 감소_smallest decrease를 초래하는 회귀변수(regressor)이다(즉, 최소의 부분 F -통계량[partial F -statistic]을 초래하는 회귀 변수). 그 과정은 제거될 후보 회귀변수가 미리 선택된 F_{OUT} 을 초과하는 부분 F 값(partial F value)을 가질 때까지 계속된다.

다음 절은 같은 자료로 위의 세가지 순차적 알고리듬(sequential algorithms)을 적용한 것을 보여주는, 주석(annotated)이 달린 컴퓨터 출력물이다.

예제 4.3 Hald자료

앞의 Hald자료에서 MSE와 PRESS 잔차에 의한 최적 모형(best model)과 C_p 값에 의한 최적 모형이 다른 것을 볼 수 있었다. 따라서 다시 한번 최적의 모형을 찾기 위해서 전진선택법(forward selection), 후진 제거법(backward elimination), 단계적 회귀법(stepwise regression)을 이용하여 모형 수립과정을 실행해보기로 한다.

Table 4.9, 4.10과 4.11은 SAS 프로그램을 사용하여 각각 전진 선택(forward selection), 후진 제거(backward elimination), 단계적 회귀(stepwise regression)를 시행하여 얻은 컴퓨터 출력물을 보여준다. 적절한 통계량과 함께 전진 선택(forward selection)과 단계적 회귀(stepwise regression)에 의해 제시된 모형은 다음과 같다.

$$\hat{y} = 71.6483 + 1.4519x_1 + 0.4161x_2 - 0.2365x_4$$

$$s^2 = 5.3303$$

$$R^2 = 0.9823$$

$$C_p = 3.0182$$

반면, 후진 제거(backward elimination), 에 의해 제시된 모형은 다음과 같다.

$$\hat{y} = 52.5774 + 1.4683x_1 + 0.6623x_2$$

$$s^2 = 5.7904$$

$$R^2 = 0.9787$$

$$C_p = 2.6782$$

Table 4.9 전진선택법 절차

Step 1

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X4	1	1831.90	1831.90	22.799	0.0005762 ***
Residuals	11	883.87	80.35		

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	117.5679	5.2622	22.342	1.62e-10 ***
X4	-0.7382	0.1546	-4.775	0.000576 ***

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

Residual standard error: 8.964 on 11 degrees of freedom
 Multiple R-Squared: 0.6745, Adjusted R-squared: 0.645
 F-statistic: 22.8 on 1 and 11 DF, p-value: 0.0005762

Step 2

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	193.96	7.120e-08 ***
X4	1	1190.92	1190.92	159.30	1.815e-07 ***
Residuals	10	74.76	7.48		

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	103.09738	2.12398	48.54	3.32e-13 ***
X1	1.43996	0.13842	10.40	1.11e-06 ***
X4	-0.61395	0.04864	-12.62	1.81e-07 ***

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

Residual standard error: 2.734 on 10 degrees of freedom

Multiple R-Squared: 0.9725, Adjusted R-squared: 0.967

F-statistic: 176.6 on 2 and 10 DF, p-value: 1.581e-08

Step 3

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	272.0439	4.934e-08 ***
X2	1	1207.78	1207.78	226.5879	1.094e-07 ***
X4	1	9.93	9.93	1.8633	0.2054
Residuals	9	47.97	5.33		

```

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.6483    14.1424   5.066 0.000675 ***
X1          1.4519     0.1170  12.410 5.78e-07 ***
X2          0.4161     0.1856   2.242 0.051687 .
X4         -0.2365     0.1733  -1.365 0.205395
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-Squared: 0.9823,      Adjusted R-squared: 0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08

```

Table 4.10 후진제거법 절차

```

Step 1
Analysis of Variance Table

Response: Y
            Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 1450.08 1450.08 242.3679 2.888e-07 ***
X2          1 1207.78 1207.78 201.8705 5.863e-07 ***
X3          1     9.79     9.79   1.6370    0.2366
X4          1     0.25     0.25   0.0413    0.8441
Residuals   8    47.86    5.98
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.4054    70.0710   0.891   0.3991
X1          1.5511     0.7448   2.083   0.0708 .
X2          0.5102     0.7238   0.705   0.5009
X3          0.1019     0.7547   0.135   0.8959
X4         -0.1441     0.7091  -0.203   0.8441
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

```

Residual standard error: 2.446 on 8 degrees of freedom
 Multiple R-Squared: 0.9824, Adjusted R-squared: 0.9736
 F-statistic: 111.5 on 4 and 8 DF, p-value: 4.756e-07

Step 2

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	272.0439	4.934e-08 ***
X2	1	1207.78	1207.78	226.5879	1.094e-07 ***
X4	1	9.93	9.93	1.8633	0.2054
Residuals	9	47.97	5.33		
<hr/>					

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	71.6483	14.1424	5.066	0.000675 ***
X1	1.4519	0.1170	12.410	5.78e-07 ***
X2	0.4161	0.1856	2.242	0.051687 .
X4	-0.2365	0.1733	-1.365	0.205395
<hr/>				

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

Residual standard error: 2.309 on 9 degrees of freedom

Multiple R-Squared: 0.9823, Adjusted R-squared: 0.9764
 F-statistic: 166.8 on 3 and 9 DF, p-value: 3.323e-08

Step 3

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	250.43	2.088e-08 ***
X2	1	1207.78	1207.78	208.58	5.029e-08 ***

```
Residuals 10    57.90    5.79
```

```
---
```

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

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

(Intercept)	52.57735	2.28617	23.00	5.46e-10 ***
-------------	----------	---------	-------	--------------

X1	1.46831	0.12130	12.11	2.69e-07 ***
----	---------	---------	-------	--------------

X2	0.66225	0.04585	14.44	5.03e-08 ***
----	---------	---------	-------	--------------

```
---
```

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

```
Residual standard error: 2.406 on 10 degrees of freedom
```

```
Multiple R-Squared: 0.9787,      Adjusted R-squared: 0.9744
```

```
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

이 예에서 사용된 유의수준(significance levels)은 유입(entry) 시 0.500이고, 제거(removal) 시 0.100이다. 대부분의 사용 가능한 컴퓨터 알고리듬에서 분석가는 이러한 모수들(parameters)을 통제한다. 일반적으로 권장되는 것은 제거(removal)를 위해서 0.01, 0.05 또는 0.10처럼 꽤 전통적인 값들을 사용하는 것이다. 그러나, 유입(entry)을 위해서는 상당히 높은 값, 가령 0.25-0.50을 사용하도록 제안한다. 우리는 여기에서 이러한 권장 사항에 대한 자세한 논리적 근거를 보여주지는 않을 것이다. 그러나, 독자는 어느 전진 절차(forward procedure)(즉, 전진선택[forward selection] 또는 단계적[stepwise])에서도 F-임계값(F critical value)이 초기 단계에서 어떤 확률적 의미(probabilistic sense)에서도 엄밀하게 적절하지 않음을 알아야 한다. 모형의 과소지정(underspecification)이 s^2 에 미치는 영향(impact)에 관하여 4.3 절에서 논의한 것을 기억해보라. 초기 단계에서 만들어진 모형들은 종종 과소지정되며(underspecified), s^2 이 심하게 과추정(overestimation)될 수 있는 결과를 보인다. 그리고 나서, 그러한 경향은 갑자기 멈출 것이고("stop short"), 아마도 F-통계량(F-statistic)의 꺼짐(deflation)으로 인해 중요 변수들(variables)이 유입되도록 하지는 않을 것이다. 알고리듬이 여러 모형에 실제로 적용될 때 분석가가 회귀변수(regressor variables)의 역할에 대해 더 많은 것을 배우게 되므로, 알고리듬이 조기에 종결되도록 하는 것은 비생산적이다. 분명히 이러한 문제는 후진 제거(backward elimination)의 경우에 상당히 덜 심각하다. 유의 수준(significance level) 문제에 관한 수학적 세부 내용은 Pope and Webster (1972)에서 찾을 수 있다.

Table 4.11 단계별 회귀 절차

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
X4	1	1831.90	1831.90	22.799	0.0005762 ***						
Residuals	11	883.87	80.35								

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	117.5679	5.2622	22.342	1.62e-10 ***							
X4	-0.7382	0.1546	-4.775	0.000576 ***							

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

Residual standard error: 8.964 on 11 degrees of freedom
 Multiple R-Squared: 0.6745, Adjusted R-squared: 0.645
 F-statistic: 22.8 on 1 and 11 DF, p-value: 0.0005762

Step 2

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
X1	1	1450.08	1450.08	193.96	7.120e-08 ***						
X4	1	1190.92	1190.92	159.30	1.815e-07 ***						
Residuals	10	74.76	7.48								

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	103.09738	2.12398	48.54	3.32e-13 ***							
X1	1.43996	0.13842	10.40	1.11e-06 ***							
X4	-0.61395	0.04864	-12.62	1.81e-07 ***							

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

Residual standard error: 2.734 on 10 degrees of freedom
 Multiple R-Squared: 0.9725, Adjusted R-squared: 0.967
 F-statistic: 176.6 on 2 and 10 DF, p-value: 1.581e-08

Step 3

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	272.0439	4.934e-08 ***
X2	1	1207.78	1207.78	226.5879	1.094e-07 ***
X4	1	9.93	9.93	1.8633	0.2054
Residuals	9	47.97	5.33		

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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	71.6483	14.1424	5.066	0.000675 ***
X1	1.4519	0.1170	12.410	5.78e-07 ***
X2	0.4161	0.1856	2.242	0.051687 .
X4	-0.2365	0.1733	-1.365	0.205395

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

Residual standard error: 2.309 on 9 degrees of freedom

Multiple R-Squared: 0.9823, Adjusted R-squared: 0.9764

F-statistic: 166.8 on 3 and 9 DF, p-value: 3.323e-08

Step 4

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1450.08	1450.08	250.43	2.088e-08 ***
X2	1	1207.78	1207.78	208.58	5.029e-08 ***
Residuals	10	57.90	5.79		

```

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

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
X1          1.46831    0.12130   12.11 2.69e-07 ***
X2          0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-Squared: 0.9787,      Adjusted R-squared: 0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09

```

자료 분석가는 순차적 모형개발 알고리듬(sequential model-building algorithms)을 하나의 최종 모형(final model)을 생성하는 블랙 박스로 간주하기보다는, 사용자로 하여금 여러 모형들이 실행되는 것을 관찰하게끔 하는 훈련과정으로 간주해야 할 것이다. 많은 회귀 변수들(regressors) 사이에 공선성(collinearity)이 존재하는 자료 집합(data set)에 있어 순차적 절차(sequential procedure)는 부분적으로 비효율적일 수 있다. 어떤 특정 모형개발 전략이 이론적 민감도(theoretical sensibility), 예측능(prediction performance) 등의 표준적인 처리 면에서 최상의 변수 부분집합(best variable subset)을 초래한다는 보증은 없다는 것을 명심해야 한다. 최종 결과는 선택된 순차적 방법(sequential method)과 F_{IN} , F_{OUT} 값들에 매우 많이 의존한다.

예제에 사용된 R-code는 다음과 같다.

```

#forward
g1<-lm(Y~X4,data)
g2<-lm(Y~X1+X4,data)
g3<-lm(Y~X1+X2+X4,data)

#backward
g1<-lm(Y~X1+X2+X3+X4,data)
g2<-lm(Y~X1+X2+X4,data)
g3<-lm(Y~X1+X2,data)

#stepwise
g1<-lm(Y~X4,data)

```

```
g2<-lm(Y~X1+X4,data)
g3<-lm(Y~X1+X2+X4,data)
g4<-lm(Y~X1+X2,data)
```

4.5. 추가적인 설명과 모든 가능한 회귀들(Further Comments and All Possible Regressions)

여기에서 토의되는 순차적 알고리듬(sequential algorithms)은 분석가가 모든 가능한 회귀(regressions)에 관한 정보를 얻지 못하거나, 비정상적으로 많은 개수의 회귀변수들(regressor variables)로 인해 모든 가능한 회귀들을 해 볼 수 없는 경우에 실제로 이용할 수 있는 방법이다. 그럼에도 불구하고, 이 알고리듬의 한가지 약점은 MAXR은 별도로 하더라도, 개발 초기단계에서 강도(strength)를 고려해야 한다는 것이다. 이 알고리듬은 많은 부분집합 모형(subset models)에 대한 결과를 보여주지 않으면서 해답(one solution)을 주도록 디자인되었다. 독자들은, 참된 최적 모형(true best model)은 어떤 과정(procedures)을 거치더라도 살아남을 것이라고 절대로 확신할 수 없다는 것을 기억해야만 한다; 사실, 최적 모형(best model)이 계산적으로 만들어질 것이라는 확실성도 없다. 따라서, 가능하다면, 사용자는 최종 선택을 위해서 또 다른 기법에 의지해야 한다.

경험적인 모형 개발을 다루는 과학자라면 결정을 내리기(decision-making) 위해서 가능한 많은 정보를 알아야 한다고 우리는 생각한다. 완전한 정보(full information)는 가능한 모든 회귀(regressions)를 사용해야만 얻을 수 있다. 모든 가능한 회귀(regressions)로부터의 결과를 만들어내는 알고리듬을 알고 싶다면 Furnival (1971), Furnival과 Wilson (1974), 또는 Schatzoff 등(1968)의 저서를 참조하라. 특히 BMDP와 SAS 같은 컴퓨터 통계 패키지를 공부해야 한다. 일반적으로 사용자들은 모든 부분집합 모형들(subset models)의 회귀(regression)에 관련된 필수 정보를 모두 얻을 수 있다고 기대해서는 안된다. 합리적인 원리는, 모든 가능한 회귀(all possible regressions)에 대하여 컴퓨터를 실행시켜서 후보모형(candidate models)의 수를 감소시키기에 충분한 정보를 얻어내는 것이다. 그 후, 더 완전한 정보, 잔차(residuals)에 관한 정보, 때로는 자료 분할(data splitting), 이상치(outlier)와 지렛대(leverage) 계산 등을 통해서 이 후보모형들을 심층 조사할 수 있다(5, 6, 7장을 참고하라). 모든 가능한 회귀 정보(regression information)를 만들어 내는 매체(vehicle)에 대한 예가 IML (interactive matrix language)로 쓰여진 SAS (1987) 프로그램이다. 이 프로그램은 모든 가능한 모형에 대한 s^2 , R^2 , C_p , PRESS, 절대 PRESS 잔차의 합(sum of the absolute PRESS residuals)을 생성해 낸다.

예제 4.4 Hald 자료

예제 4.1에 설명을 위해서 사용되었던 Hald 자료를 다시 고려해 보자. 가능한 추가 연구를 위해서 강조된 모형들을 상기해보자. 여기서 우리는 모든 가능한 조합을 위해서 R^2 , C_p , s^2 , PRESS, 절대 PRESS 잔차의 합(sum of the absolute PRESS residuals)을 생성하고자 선택하였다. Table 4.12은 그 결과를 나타낸 것이다(모든 조합들[combinations]은 절편[intercept]을 가진다). Table 4.12에서 앞에서 강조되었던 모형 외에 추가적인 모형들도 알 수 있다. 만약 절대 PRESS 잔차합(sum of the absolute PRESS residuals)을 고려한다면 모형 (x_1, x_2, x_3) 이 가장 작은 값을 제공한다. 반면, PRESS 통계량(PRESS statistic)과 s^2 값을 고려한다면 모형 (x_1, x_2, x_4) 가

가장 작은 값을 제공하고 C_p 통계량(C_p statistic)을 고려한다면 모형 (x_1, x_2)가 가장 최선의 모형으로 나타난다. 즉 어떠한 통계량을 고려하느냐에 따라 최선의 모형은 달라질 수 있다. 따라서, 자료 분석가는 기준(criteria)을 잘 알고 있어야 하며, 관련 전문가나 과학자들에게 각각의 변수(variable)의 상대적인 중요성에 관한 경험적 의견(educated opinion)에 대해 자문을 구해야 한다. 이 연구에 관해서, 대략 3개의 모형을 자세히 살펴보아야 하고, 잔차(residuals) 혹은 다른 진단(diagnostics)에 관한 연구도 함께 살펴보아야 한다. 또한, 최소 제곱(least squares)의 대용물(an alternative)로서 편향된 추정(biased estimation)을 고려하고, 다중공선성(multicollinearity)이 있는지 더 세심하게 진단하여야 한다. (8장을 참고하시오)

Table 4.12 모든 가능한 회귀결과

변수수	$\sum_{i=1}^n e_{i,-i} $	R^2	C_p	PRESS	s^2	해당변수
1	105.6427	0.6745	138.7308	1194.2182	80.3515	x_4
1	99.4746	0.6663	142.4864	1202.0868	82.3942	x_2
1	130.2304	0.5339	202.5488	1699.6116	115.0624	x_1
1	165.0845	0.2859	315.1543	2616.3639	176.3091	x_3
2	32.3072	0.9787	2.6782	93.8825	5.7904	x_1, x_2
2	34.1035	0.9725	5.4959	121.2244	7.4762	x_1, x_4
2	49.9340	0.9353	22.3731	294.0139	17.5738	x_3, x_4
2	73.5898	0.8470	62.4377	701.7432	41.5443	x_2, x_3
2	110.4507	0.6801	138.2259	1461.8142	86.8880	x_2, x_4
2	153.1886	0.5482	198.0947	2218.1183	122.7072	x_1, x_3
3	28.5078	0.9823	3.0182	85.3511	5.3303	x_1, x_2, x_4
3	28.3890	0.9823	3.0413	90.0000	5.3456	x_1, x_2, x_3
3	30.9805	0.9813	3.4968	94.5371	5.6485	x_1, x_3, x_4
3	36.8836	0.9728	7.3375	146.8527	8.2016	x_2, x_3, x_4
4	32.1804	0.9824	5.0000	110.3466	5.9830	x_1, x_2, x_3, x_4

독자들은 이 설명을 실제 사례 연구(true case study)로 간주해서는 안 된다. 최선의 모형(best model)에 대한 어떤 최종적인 선택(final selection)을 하기 전에, 5, 6, 7, 8장에서 다른 진단적 정보(diagnostic information)를 얻어야 할 것이다.

예제에 사용된 R-code는 다음과 같다.

```
g1.4<-lm(Y~X4,data)
g1.2<-lm(Y~X2,data)
g1.1<-lm(Y~X1,data)
g1.3<-lm(Y~X3,data)
```

```

g2.12<-lm(Y~X1+X2,data)
g2.14<-lm(Y~X1+X4,data)
g2.34<-lm(Y~X3+X4,data)
g2.23<-lm(Y~X2+X3,data)
g2.24<-lm(Y~X2+X4,data)
g2.13<-lm(Y~X1+X3,data)

g3.124<-lm(Y~X1+X2+X4,data)
g3.123<-lm(Y~X1+X2+X3,data)
g3.134<-lm(Y~X1+X2+X4,data)
g3.234<-lm(Y~X1+X2+X4,data)

g4<-lm(Y~X1+X2+X3+X4,data)

#leverage
g1.4.hat<-influence(g1.4)$hat
g1.2.hat<-influence(g1.2)$hat
g1.1.hat<-influence(g1.1)$hat
g1.3.hat<-influence(g1.3)$hat

g2.12.hat<-influence(g2.12)$hat
g2.14.hat<-influence(g2.14)$hat
g2.34.hat<-influence(g2.34)$hat
g2.23.hat<-influence(g2.23)$hat
g2.24.hat<-influence(g2.24)$hat
g2.13.hat<-influence(g2.13)$hat

g3.124.hat<-influence(g3.124)$hat
g3.123.hat<-influence(g3.123)$hat
g3.134.hat<-influence(g3.134)$hat
g3.234.hat<-influence(g3.234)$hat

g4.hat<-influence(g4)$hat

#residual
g1.4.res<-residuals(g1.4)

```

```

g1.2.res<-residuals(g1.2)
g1.1.res<-residuals(g1.1)
g1.3.res<-residuals(g1.3)

g2.12.res<-residuals(g2.12)
g2.14.res<-residuals(g2.14)
g2.34.res<-residuals(g2.34)
g2.23.res<-residuals(g2.23)
g2.24.res<-residuals(g2.24)
g2.13.res<-residuals(g2.13)

g3.124.res<-residuals(g3.124)
g3.123.res<-residuals(g3.123)
g3.134.res<-residuals(g3.134)
g3.234.res<-residuals(g3.234)

g4.res<-residuals(g4)

#press
g1.4.pre<-sum(g1.4.res/(1-g1.4.hat))
g1.4.pre<-sum(g1.4.res/(1-g1.4.hat))
g1.4.pre<-sum(g1.4.res/(1-g1.4.hat))
g1.4.pre<-sum(g1.4.res/(1-g1.4.hat))

g2.12.pre<-sum(g2.12.res/(1-g2.12.hat))
g2.14.pre<-sum(g2.14.res/(1-g2.14.hat))
g2.34.pre<-sum(g2.34.res/(1-g2.34.hat))
g2.23.pre<-sum(g2.23.res/(1-g2.23.hat))
g2.24.pre<-sum(g2.24.res/(1-g2.24.hat))
g2.13.pre<-sum(g2.13.res/(1-g2.13.hat))

g3.124.pre<-sum(g3.124.res/(1-g3.124.hat))
g3.123.pre<-sum(g3.123.res/(1-g3.123.hat))
g3.134.pre<-sum(g3.134.res/(1-g3.134.hat))
g3.234.pre<-sum(g3.234.res/(1-g3.234.hat))

```

```

g4.pre<-sum(g4.res/(1-g4.hat))

#abs press

g1.4.absp<-sum(abs(g1.4.res/(1-g1.4.hat)))
g1.2.absp<-sum(abs(g1.2.res/(1-g1.2.hat)))
g1.1.absp<-sum(abs(g1.1.res/(1-g1.1.hat)))
g1.3.absp<-sum(abs(g1.3.res/(1-g1.3.hat)))

g2.12.absp<-sum(abs(g2.12.res/(1-g2.12.hat)))
g2.14.absp<-sum(abs(g2.14.res/(1-g2.14.hat)))
g2.34.absp<-sum(abs(g2.34.res/(1-g2.34.hat)))
g2.23.absp<-sum(abs(g2.23.res/(1-g2.23.hat)))
g2.24.absp<-sum(abs(g2.24.res/(1-g2.24.hat)))
g2.13.absp<-sum(abs(g2.13.res/(1-g2.13.hat)))

g3.124.absp<-sum(abs(g3.124.res/(1-g3.124.hat)))
g3.123.absp<-sum(abs(g3.123.res/(1-g3.123.hat)))
g3.134.absp<-sum(abs(g3.134.res/(1-g3.134.hat)))
g3.234.absp<-sum(abs(g3.234.res/(1-g3.234.hat)))

g4.absp<-sum(abs(g4.res/(1-g4.hat)))

#R-squared

g1.4.rs<-summary(g1.4)$r.squared
g1.2.rs<-summary(g1.2)$r.squared
g1.1.rs<-summary(g1.1)$r.squared
g1.3.rs<-summary(g1.3)$r.squared

g2.12.rs<-summary(g2.12)$r.squared
g2.14.rs<-summary(g2.14)$r.squared
g2.34.rs<-summary(g2.34)$r.squared
g2.23.rs<-summary(g2.23)$r.squared
g2.24.rs<-summary(g2.24)$r.squared
g2.13.rs<-summary(g2.13)$r.squared

g3.124.rs<-summary(g3.124)$r.squared

```

```

g3.123.rs<-summary(g3.123)$r.squared
g3.134.rs<-summary(g3.134)$r.squared
g3.234.rs<-summary(g3.234)$r.squared

g4.rs<-summary(g4)$r.squared

#sigma
g1.4.sig<-summary(g1.4)$sigma^2
g1.2.sig<-summary(g1.2)$sigma^2
g1.1.sig<-summary(g1.1)$sigma^2
g1.3.sig<-summary(g1.3)$sigma^2

g2.12.sig<-summary(g2.12)$sigma^2
g2.14.sig<-summary(g2.14)$sigma^2
g2.34.sig<-summary(g2.34)$sigma^2
g2.23.sig<-summary(g2.23)$sigma^2
g2.24.sig<-summary(g2.24)$sigma^2
g2.13.sig<-summary(g2.13)$sigma^2

g3.124.sig<-summary(g3.124)$sigma^2
g3.123.sig<-summary(g3.123)$sigma^2
g3.134.sig<-summary(g3.134)$sigma^2
g3.234.sig<-summary(g3.234)$sigma^2

g4.sig<-summary(g4)$sigma^2

#Cp
n<-length(data[,1])
p<-c(2,2,2,2,3,3,3,3,3,4,4,4,4,5)
ss<-
c(g1.4.sig,g1.2.sig,g1.1.sig,g1.3.sig,g2.12.sig,g2.14.sig,g2.34.sig,g2.23.sig,g2.24.sig,g2.13.sig,g3.124.sig,g3.1
23.sig,g3.134.sig,g3.234.sig,g4.sig)
cp<-p+(ss-g4.sig)*(n-p)/g4.sig

```

예제 4.5 Hald 자료

예제 4.4에 계속 이어서 15가지의 모형의 결과를 살펴보도록 하자. 각 모형에 대해서

계산된 s^2 , R^2 , C_p , PRESS, $\sum_{i=1}^n |e_{i,-i}|$ 는 이미 앞의 예제에서 제시하였기 때문에 다시 표시하지는 않을 것이다. Table 4.13은 15개 모형에서 각 관측값에 대한 각각의 잔차들($y - \hat{y}$),

뿐만 아니라 예측 오차의 제곱합(sum of squares of the prediction error)과 절대 예측 오차의 합(sum of the absolute prediction error)을 나타내었다.

그 결과 앞의 결과와 마찬가지로 모형 (x_1, x_2, x_3) , (x_1, x_2, x_4) 그리고 (x_1, x_2) 를 지지하는 결과를 얻을 수 있다. 예제 4.5에 설명된, 모든 가능한 회귀에 의해 생산된 정보는 예측 방정식으로서 최선을 수행하는 모형의 선택이 필요할 때 도움이 될 것이다. s^2 , C_p , PRESS, 절대 PRESS 잔차의 합(sum of absolute PRESS residuals)과 같은 통계량(statistic)은 그런 관점에서 유용하지 못할 것이다. 그러나, 잔차(residuals) 분석, 공선성 진단(collinearity diagnostics), 변환(transformations), 영향력 관찰(detection of influential observations) 그리고 이어지는 장에서 논의될 다른 주제 등의 개념을 고려하지 않고서는 모형의 최종선택을 해서는 안된다.

Table 4.12 Prediction residuals and prediction performance criteria for the Hald data

	(x_1, x_2, x_3, x_4)	(x_1, x_2, x_4)	(x_1, x_2, x_3)	(x_1, x_3, x_4)	(x_2, x_3, x_4)
1	0.0048	0.0617	-0.1448	0.0617	0.0617
2	1.5112	1.4327	1.6097	1.4327	1.4327
3	-1.6709	-1.891	-1.3358	-1.891	-1.891
4	-1.7271	-1.8016	-1.6129	-1.8016	-1.8016
5	0.2508	0.2562	0.1755	0.2562	0.2562
6	3.9254	3.8982	3.9711	3.8982	3.8982
7	-1.4487	-1.4287	-1.4726	-1.4287	-1.4287
8	-3.175	-3.0919	-3.2543	-3.0919	-3.0919
9	1.3783	1.2818	1.5409	1.2818	1.2818
10	0.2815	0.3539	0.2176	0.3539	0.3539
11	1.991	2.0977	1.8835	2.0977	2.0977
12	0.973	1.0556	0.845	1.0556	1.0556
13	-2.2943	-2.2247	-2.4229	-2.2247	-2.2247
$\sum_{i=1}^n (y - \hat{y})^2$		47.86364	47.97273	48.11061	47.97273
$\sum_{i=1}^n y - \hat{y} $		20.63206	20.87565	20.48642	20.87565
	(x_1, x_2)	(x_1, x_4)	(x_3, x_4)	(x_2, x_3)	(x_2, x_4)
1	-1.574	2.1601	-2.1073	-6.5389	3.6729

2	1.0491	1.6882	-1.3054	-3.8574	-5.1153
3	-1.5147	-2.3579	-2.8916	-0.662	1.8681
4	-1.6585	-2.4811	-0.0274	0.9212	5.2781
5	-1.3925	2.9834	-4.2715	-8.1535	0.652
6	4.0475	3.7701	4.6575	5.9777	7.9929
7	-1.3021	-1.0335	-3.8373	-4.1565	-10.7927
8	-2.0754	-5.0234	-0.5033	-0.0614	-11.1927
9	1.8245	0.6297	-0.6439	-0.3155	-7.7962
10	1.3625	-1.4737	8.2566	13.4864	19.0079
11	3.2643	0.1371	4.7506	5.665	-7.2602
12	0.8628	1.7305	1.5115	2.033	4.1035
13	-2.8934	-0.7295	-3.5884	-4.338	-0.4183
$\sum_{i=1}^n (y - \hat{y})^2$	57.90448	74.76211	175.73800	415.44273	868.88013
$\sum_{i=1}^n y - \hat{y} $	24.82129	26.19820	38.35218	56.16665	85.15083

	(x_1, x_3)	(x_4)	(x_2)	(x_1)	(x_3)
1	-13.0031	5.2218	0.5591	-16.0606	-24.168
2	-7.7785	-4.8835	-6.0083	-9.0481	-17.0659
3	2.5581	1.4953	2.6853	2.2644	4.1436
4	-14.1419	4.7257	5.7134	-14.4356	-12.5564
5	4.3969	2.6914	-2.5582	1.3394	-6.768
6	6.9636	7.8716	8.3745	7.1644	10.2994
7	15.0076	-10.439	-10.7515	15.6144	13.8456
8	-13.0398	-12.5888	-9.3866	-10.8481	-10.0755
9	7.2256	-8.2284	-6.9364	7.8832	5.5014
10	-6.9887	17.5243	21.3875	-4.823	10.7205
11	-2.2342	-8.6704	-5.1887	0.4519	2.4803
12	11.0636	4.59	3.7941	11.2644	14.3994
13	9.9706	0.69	-1.6842	9.2332	9.2436
$\sum_{i=1}^n (y - \hat{y})^2$	1227.07206	883.86692	906.33634	1265.68675	1939.40047
$\sum_{i=1}^n y - \hat{y} $	114.37232	89.62018	85.02767	110.43075	141.26750

예제에 사용된 R-code는 다음과 같다.

```
resi<-cbind(g4.res, g3.124.res, g3.123.res, g3.134.res, g3.234.res, g2.12.res, g2.14.res, g2.34.res, g2.23.res, g2.24.res,  
g2.13.res, g1.4.res, g1.2.res, g1.1.res, g1.3.res)  
resi_ab<-abs(resi)  
sresi_ab<-colSums(resi_ab)  
  
resi_sq<-resi^2  
sresi_sq<-colSums(resi_sq)
```

5. 잔차분석(Analysis of Residuals)

2, 3, 4장의 여러 곳에서, 잔차(residuals)의 개념을 소개하고 수식 전개에 포함하고 있었다. 잔차(residuals)는 회귀분석(regression analysis) 결과의 주 요소로서 제시되고 있으며, 가정된 모형(가정 자체도 포함)과 관찰된 자료 간에 얼마나 차이가 나는지 알아내고 평가하는데 흔히 사용된다. 많은 상용 회귀분석 프로그램에 잔차(residuals) 계산은 이미 기본으로 포함되어 있다. 따라서 회귀분석을 하는 사람이 잔차분석(analysis of residuals)에 이용 가능한 도구들과, 이들로부터 어떤 정보를 얻을 수 있는지 아는 것이 중요하다.

2.12 절에서는 단순선형모형(simple linear regression)에서의 잔차 사용(the use of residuals)을 강조하였는데, 모형 오지정(model misspecification)과 오차의 정규성 가정(assumption of normal errors)에 대한 위배를 찾아 내는데 초점을 맞추었다. 독자는 이 장에서 다루어질 자세한 내용에 앞서 이것을 숙지하여야 한다. 이 장에서, 모형 오지정(model misspecification)은 다중선형회귀(multiple linear regression)를 다루는 부분에서 다시 나오게 될 것이다. 등분산가정(homogeneous variance assumption)의 위배(violation)를 밝히는 진단용 그림(diagnostic plots)도 다루고 예를 들 것이다. 2장에서 잔차(residuals)의 표준화된 형태인 스튜던트화 잔차(studentized residual)를 다루었는데, 여기에서는 스튜던트화 잔차(studentized residual)의 사용에 대하여 자세히 고찰할 것이다. 이상치(outlier)를 찾아내기 위한 진단도구(diagnostics)로 스튜던트화 잔차(studentized residual)를 사용하는데, 이에 대해 언급하고 그림으로도 보여줄 것이다. 통계학자들은 회귀분석에 들어가기 전에 실험 자료를 변환(transformation)하는 문제에 대하여 최근 매우 관심을 쏟아 왔다. 자료를 분석하는데 참계량(true “metric”)을 결정하는데 이러한 자료변환(transformations)의 필요성을 진단해 주는 방법들을 매우 유용하게 사용할 수 있다. 본장에서는 이 목적으로 사용되는 몇 가지의 진단용 그림(diagnostic type plots)에 관하여 고찰하기로 한다. 이러한 그림(plots)은 회귀 분석(regression)에서 각 변수들(variables)의 역할을 그림으로 보여줄 수 있다. 부가적으로, 이러한 특별한 잔차 그림(residual plots)은 고영향력 관측값(high influence observations)의 존재를 밝힐 수도 있다. 이 장은 고영향력 관측값(high influence observations)의 진단에만 할애된 6장과 매우 밀접하게 묶여 있다.

5.1. 잔차로부터 얻어지는 정보(Information retrieved from Residuals)

이상적인 조건 하에서 잔차(residuals)의 특성을 검토하고 나면 이들의 가치를 좀 더 이해 할 수 있을 것이다. 3.1에서 ε_i 에 대한 가정(assumptions), 즉 i 번째 잔차(residual)인 e_i 는 평균이 0이고 분산(variance)이 다음과 같이 주어진다는 것을 기억하자.

$$Var(e_i) = \sigma^2(1 - h_{ii}) \quad (5.1)$$

h_{ii} 는 모자행렬(HAT matrix)의 i 번째 대각원소(HAT diagonal)이다. 부가적으로 i 번째와 j 번째 잔차 ($i \neq j$)는 다음과 같은 공분산(covariance)을 가진다.

$$Cov(e_i, e_j) = -h_{ij} \cdot \sigma^2 \quad (5.2)$$

여기에서 h_{ij} 는 모자행렬(HAT matrix), $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 의 (i, j) 번째 원소(element)이다. 이러한 중요한 결과들은 앞으로 나올 5.3 절에 있는 결과를 살펴보면 쉽게 입증할 수 있다. 식 (5.3) 으로부터 잔차벡터(vector of residuals), e 는 다음과 같이 쓸 수 있다.

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} = [\mathbf{I} - \mathbf{H}]\mathbf{y}$$

이제 잔차(residuals)의 분산공분산행렬(variance-covariance matrix)을 알려진 규칙을 토대로 정할 수 있다. 즉,

$$\text{Var}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})(\sigma^2 \mathbf{I}_n)(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{I} - \mathbf{H})^2$$

$(\mathbf{I} - \mathbf{H})$ 는 멱등행렬(idempotent, 제곱을 해도 그 값이 변하지 않는 행렬. 단위행렬과 영행렬은 모두 멱등행렬)이므로 $(\mathbf{I} - \mathbf{H})^2 = (\mathbf{I} - \mathbf{H})$, 따라서

$$\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

그러므로 (5.1)과 (5.2)의 결과가 도출된다.

(5.1)은 식 (2.32)의 단순선형모형(simple linear regression)에서 주어지는 잔차(residual)의 분산(variance)을 일반화한 것이다. 회귀변수(regressor variable)가 한 개인 경우 다음은 모자행렬(H)의 i 번째 대각원소(i th HAT diagonal)이다.

$$\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

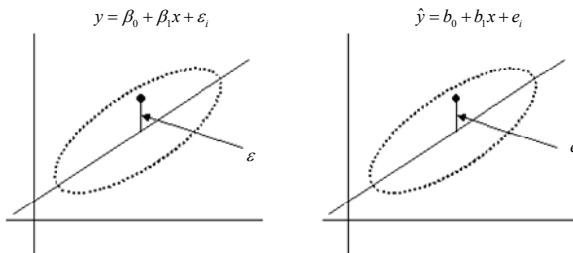
회귀변수가 하나인 경우에(single regressor), 중심화회귀모형(centered regression model)을 사용하여 $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 의 i 번째 대각(diagonal)을 도출함으로써 이를 입증해보라. 식 (2.32)는 쉽게 입증이 될 것이다.

잔차(residuals)를 연구하는 목적은 가정 위배(violations of assumptions)를 찾아내기 위한 것이다. 모형의 가정(model assumptions)은 개념상의 오차(error)인 ε_i 의 행태(behavior)에 대한 우리의 믿음을 알려주는 것이다.¹

물론, 많은 회귀분석기법(regression technology)의 타당성(validity)은 이러한 가정에 기초를 두고 있다. 잔차(residuals)는 오차(ε_i)의 특성을 반영하는데 유용하다. 만약 잔차(residuals)가 공통분산(common variance), σ^2 과 서로 독립으로(uncorrelated) 가정되어 있는 오차(ε_i)처럼 행동한다면, 회귀분석은 매우 단순화될 수 있을 것이다. 그러나 불행하게도 이러한 이상적인 경우는 식 (5.1)과 (5.2)에서 볼 수 있듯이, 일반적으로 흔하지 않다. (5.1)은 직관적으로 보기에도 합리적인 중요한 결과를 알려준다. 잔차(residual)는 그것과 연관된 자료가 중심 근처에 있을 때보다 자료 중심(data center)으로부터 멀리 떨어져 있을 때(즉, 1 근처의 h_{ii} , 0에 대한 좀 더 정밀한 추정량(estimator)이다. 5.3 절의 모자대각(HAT diagonals)에 대한 고찰을 읽어 보기 바란다. 따라서 우리는 최소제곱회귀식(least square regression equation)이 안에 있는 자료(an interior point)보다 멀리 떨어져 있는 자료(a remote point)를 더 잘 적합(fit)할 것이라고 말할 수 있다. Figure 4.1을 보라.

잔차(residuals)의 분산공분산 구조(variance-covariance structure)가 비교적 복잡하긴 해도, 잔

¹ 잔차(residual)와 오차(error)



위의 2개의 그래프에서 왼쪽에 있는 모형은 모집단의 모수식을 표현한 것이다. 즉, 우리가 궁극적으로 알고자 하는 실제의 식인 것이다. 이 경우, 모든 자료를 하나의 회귀식으로 100% 설명할 수 없다. 그래서 생각해 낸 것이 바로 오차(error)라고 하는 것으로, 이 값은 회귀식의 값과 실제값과의 차이를 말한다. 여기에서는 어떤 하나의 점과 회귀식과의 차이를 표현한 입실론(epsilon)이 바로 오차이다. 이에 비해서 잔차(residual)라고 하는 것은 표본의 회귀식에 나온 값이다. 표본에서도 마찬가지로 회귀식을 구할 수 있다. 그러나, 그 회귀식은 모집단의 실제 회귀식과는 차이가 있을 수 있다. 이때에 모집단의 회귀식과 마찬가지로 표본의 회귀식에서도 잔차라는 것을 생각할 수 있으며, 같은 아이디어에 의해 구해지게 된다. 그러나, 오차는 모수의 개념이므로 표본에서는 오차라는 용어대신 통계량의 개념을 갖는 잔차(residual)라는 용어로 대신 부르게 된다. 결국, 오차와 잔차는 같은 개념이지만 모집단의 값인가, 표본의 값인가에 따라 서로 달리 부르게 되는 것이다.

차는 유용한 많은 정보를 내포하고 있다. 물론, 분석 시 잔차의 특성 중에서 이상적인 경우 외의 편차(deviation from the ideal)를 설명하여야 한다. 부가적으로, 잔차 분석(residual analysis)의 많은 기술들이 본질적으로는 진단도구(diagnostic)로 비춰지지, 제대로 된 통계적 추론(statistical inference)의 한 부분으로 인식되지 않는다는 것이 중요하다. 우리는 앞으로도 자주 이에 대하여 독자들의 기억을 상기시킬 것이다. 2, 3장에서 우리는 가정에 대하여 고찰하였다. 어떠한 경우에도 가정 위배(violations of assumptions)를 무시하거나 부정하려고 해서는 안 된다. 그러나 이러한 가정 위배(violations of assumptions)를 다루는 여러 가지 방법이 있으며, 이를 검출(detection)하는 것부터 시작하여야 한다. 잔차(residuals)는 이러한 가정 위배를 검출하는 기초가 된다. 가정 위배를 발견할 수 있는 경우는 모형 저설정(model underspecification), 등분산 가정(homogeneous variance assumption)으로부터의 이탈, 의심스러운 자료(suspect data point)의 존재, 모형 오차(model errors)의 정규성(normality)으로부터의 이탈, 따로 떨어져 있는 고영향력 자료(isolated high influence data points) 등이다.

5.2. 잔차도(Plotting of Residuals)

여러 개의 회귀변수(multiple regressor variables)가 있는 경우, 적합된 값 \hat{y}_i 에 대한 보통잔차(ordinary residuals)를 단순하게 그리면 모형 저설정(model underspecification)이나 등분산가정(homogeneous variance assumption)으로부터의 편차(deviation)를 아는데 도움이 된다. 이상적인 상황에서의 잔차그림(residual plot)의 전형적인 모습이 fig. 5.1이다. 이 그림은 어떠한 경향도 보이지 않으면서 0 주위로 무작위로 모여 있는 양상을 보여주고 있다.

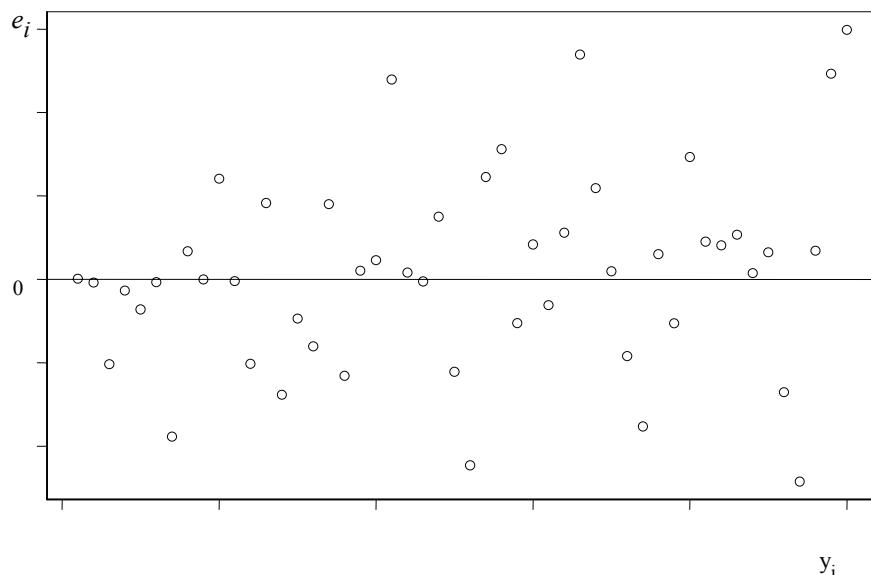
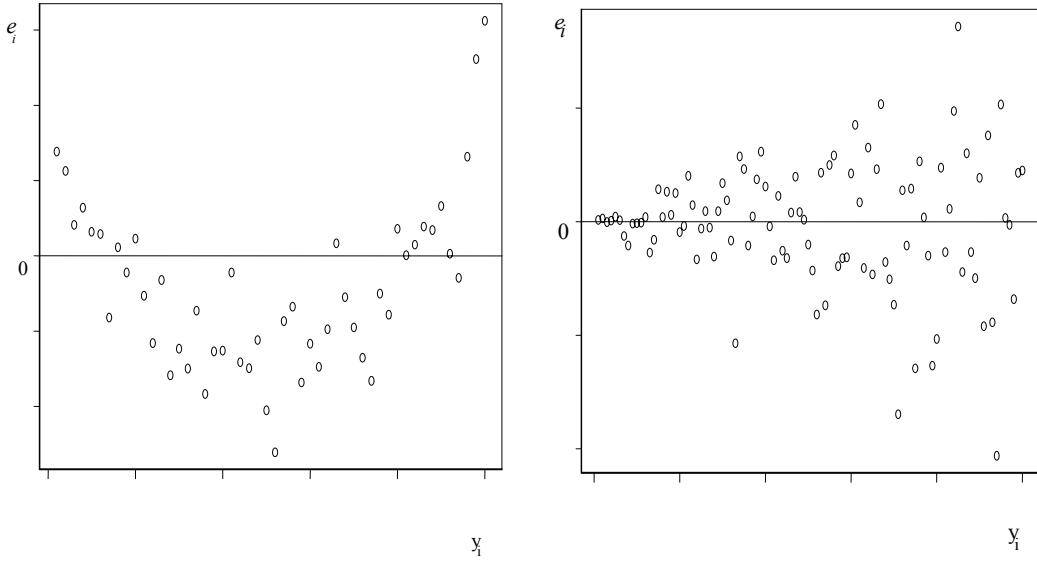


Figure 5.1 Ideal residual plot

Fig. 5.2는 모형 저설정(model underspecification) (fig. 5.2 (a)) 혹은 이분산(heterogeneous variance, fig. 5.2 (b))의 증거를 보여주는 잔차 양상이다. Fig. 5.2 (b)의 상황은 합리적으로 분명해 보인다. 깔때기 효과(funnel effect)는 반응변수(response variable)가 커질수록 잔차의 편차(deviation)도 0으로부터 커지는 것을 말한다. 그러므로 이 경우 오차 분산(error variance)이 일정하지 않고 측정된 반응이 클수록 증가한다. Fig. 5.2 (a)에서 보이는 잔차의 계통적 경향(systematic trend)은 일반적으로 모형의 항(term), 아마도 회귀변수들(regressor variables) 중의 하나에서 이차항(quadratic term)이 하나 빠진 것을 의미한다.



(a) Model should involve curvature

(b) Heterogeneous variance

Figure 5.2 Residual plots indicating violation of assumptions

예제 5.1 흑체리나무

표 5.1의 자료는 쓰러진 31개의 흑체리나무를 조사하여 나무의 높이(단위: 피트), 지표면 4.5피트에서의 나무의 지름(x_1 , 단위: 인치) 및 부피(x_2 , 단위: 입방피트)를 기록하여 얻은 것이다. 이러한 자료를 근거로 성장속도에 따르는 나무의 부피를 예측하는 것이 분석의 목적이다.

아래와 같은 모형을 이용한 단순선형회귀적합을 생각해보자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad (i = 1, 2, \dots, 31)$$

최소제곱법으로 적합하면 다음과 같은 회귀식을 얻을 수 있다.

$$\hat{y} = -57.99 + 4.71x_1 + 0.34x_2$$

이때 $R^2 = 0.948$, $s = 3.882$. 표 5.2는 관측된 흑체리나무의 부피 y , 적합된 흑체리나무의 부피 \hat{y} 그리고 잔차를 보여 준다.

Table 5.1 흑체리나무 자료

순서	$x_1(\text{지름})$	$x_2(\text{높이})$	$y(\text{부피})$
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7
7	11	66	15.6
8	11	75	18.2
9	11.1	80	22.6
10	11.2	75	19.9
11	11.3	79	24.2
12	11.4	76	21
13	11.4	76	21.4
14	11.7	69	21.3
15	12	75	19.1
16	12.9	74	22.2
17	12.9	85	33.8
18	13.3	86	27.4
19	13.7	71	25.7
20	13.8	64	24.9
21	14	78	34.5
22	14.2	80	31.7
23	14.5	74	36.3
24	16	72	38.3
25	16.3	77	42.6
26	17.3	81	55.4
27	17.5	82	55.7
28	17.9	80	58.3
29	18	80	51.5
30	18	80	51
31	20.6	87	77

Table 5.2 적합값과 잔차

순서	y (부피)	\hat{y} (Fitted Population)	$e_i = y_i - \hat{y}_i$
1	10.3	4.837660	5.46234035
2	10.3	4.553852	5.74614837
3	10.2	4.816981	5.38301873
4	16.4	15.874115	0.52588477
5	18.8	19.869008	-1.06900844
6	19.7	21.018327	-1.31832696
7	15.6	16.192688	-0.59268807
8	18.2	19.245949	-1.04594918
9	22.6	21.413021	1.18697860
10	19.9	20.187581	-0.28758128
11	24.2	22.015402	2.18459773
12	21	21.468465	-0.46846462
13	21.4	21.468465	-0.06846462
14	21.3	20.506154	0.79384587
15	19.1	23.954110	-4.85410969
16	22.2	27.852203	-5.65220290
17	33.8	31.583966	2.21603352
18	27.4	33.806482	-6.40648192
19	25.7	30.600978	-4.90097760
20	24.9	28.697035	-3.79703501
21	34.5	34.388184	0.11181561
22	31.7	36.008319	-4.30831896
23	36.3	35.385260	0.91474029
24	38.3	41.768998	-3.46899800
25	42.6	44.877702	-2.27770232
26	55.4	50.942868	4.45713224
27	55.7	52.223751	3.47624891
28	58.3	53.428513	4.87148717
29	51.5	53.899329	-2.39932888
30	51	53.899329	-2.89932888
31	77	68.515305	8.48469518

다음은 table 5.1의 회귀분석 결과이다.

Call:
lm(formula = y ~ x1 + x2)
Residuals:
Min 1Q Median 3Q Max -6.4065 -2.6493 -0.2876 2.2003 8.4847
Coefficients:
Estimate Std. Error t value Pr(> t) (Intercept) -57.9877 8.6382 -6.713 2.75e-07 *** x1 4.7082 0.2643 17.816 < 2e-16 *** x2 0.3393 0.1302 2.607 0.0145 * ---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948, Adjusted R-squared: 0.9442
F-statistic: 255 on 2 and 28 DF, p-value: < 2.2e-16

이 자료는 잔차의 집합(a set of residuals)으로 보아 회귀변수(regressor variable)가 곡선의 방식으로 들어가야 함을 보여주는 예이다. 사실 잔차(residuals)를 한번 보기만 해도 일련의 양수와 음수가 교차하면서 나타나는 다소 이상한 패턴을 알 수 있다. Fig. 5.3은 \hat{y} 값에 대한 잔차의 그림(plot)이다.

Fig. 5.3의 그림(plot)에서 보면, 모형 설정(model specification)에 대하여 심각한 의문이 생기는 것은 자명하다. 이 그림(plot)은 fig. 5.2 (a)와 유사하다. 이 경우 이차항(quadratic term)을 하나 넣음으로써 좀 더 이상적인 모습에 가까운 잔차그림(residual plot)을 얻을 수 있음을 예상할 수 있다.

자료에 적합된 이차회귀모형(quadratic regression model)의 결과는 다음과 같다.

$$\hat{y} = -9.92041 - 2.88508x_1 + 0.26862x^2 + 0.37639x_2$$

이 때 $R^2 = 0.9771$, $s = 2.625$. R^2 와 제곱근 잔차평균제곱(root residual mean square)으로 보아, 이차항(quadratic term)을 넣어 곡률(curvature)을 도입함으로써 모형을 향상시켰음을 알 수 있

다. Fig. 5.4의, 향상된 모형에 대한 잔차그림(residual plot)은 fig. 5.1의 고전적인 잔차그림(classic residual plot)에 좀 더 가까워졌다.

Figure 5.3 Plot for the linear model fit of Table 5.1

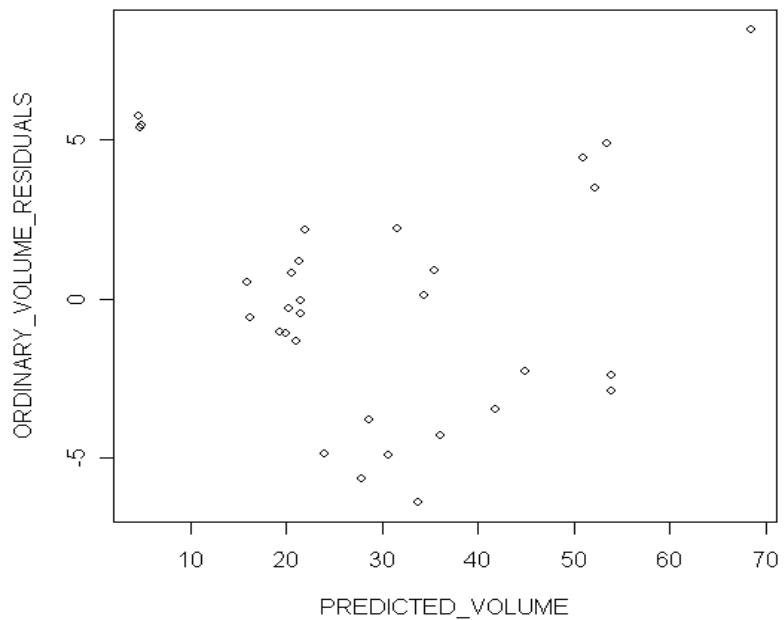
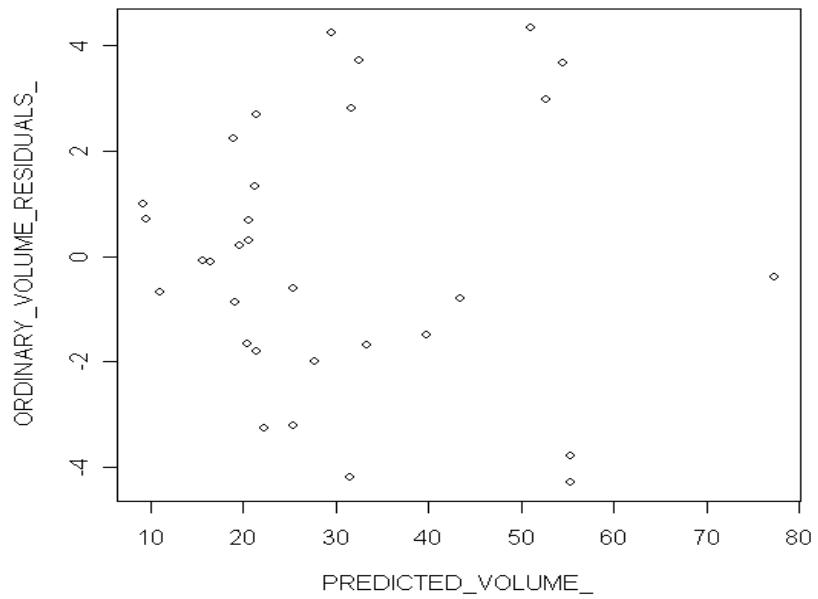


Figure 5.4 Plot for the quadratic model fit of Table 5.1.



```

Call:
lm(formula = y ~ x1 + I(x1^2) + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2928 -1.6693 -0.1018  1.7851  4.3489 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -9.92041   10.07911  -0.984 0.333729  
x1          -2.88508   1.30985  -2.203 0.036343 *  
I(x1^2)      0.26862   0.04590   5.852 3.13e-06 *** 
x2          0.37639   0.08823   4.266 0.000218 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,      Adjusted R-squared: 0.9745 
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16

```

이것은, 원통과 원추의 중간 정도로 볼 수 있는 나무의 부피 계산을 위해서는 지름의 제곱항이 필요하다는 사실과 부합된다.

$$(원통부피 = \pi \times \text{지름}^2 \times \text{높이})/4, 원추부피 = \pi \times \text{지름}^2 \times \text{높이}/12).$$

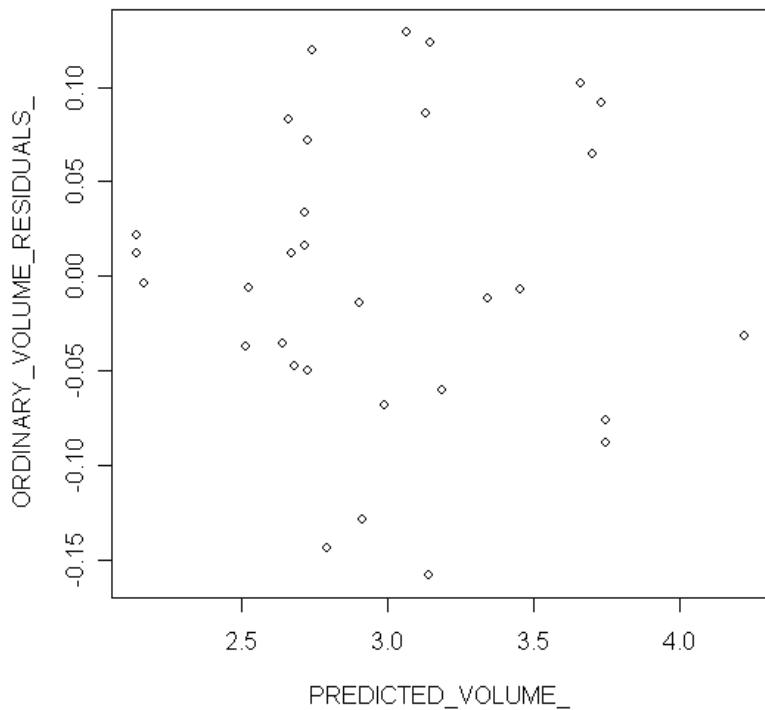
Fig. 5.3의 잔차산점도에 비하여 이상적인 잔차산점도에 가까워졌지만 나무의 부피가 커질수록 잔차가 커지는 경향이 보여 여전히 더 연구를 진행할 필요가 남아 있다.

반응변수 y 가 부피임을 감안하여 반응변수(response variable)에 1/3제곱근을 취한 다음과 같은 모형에 적합시켜보도록 하겠다.

$$Y^{1/3} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \varepsilon$$

이 모형의 R^2 와 제곱근 잔차평균제곱(root residual mean square)은 각각 0.9777, 0.08251이며, 잔차평균제곱이 이전 모형에 비하여 개선되었다. 잔차그림은 fig. 5.5와 같다. 이 잔차그림은 이상적인 fig. 5.1과 매우 유사하다. 여기서 요점은 모형 오설정(model misspecification)이 분명할 경우 잔차그림으로부터 눈으로 이것을 확인할 수 있다는 것이다. 물론 잔차그림에서 알 수 있는 것에 대하여 모형 평가를 하여야 한다. 분명한 것은 모형이 좋아졌다는 것이다.

Figure 5.5 Plot for quadratic and transformation on y model of Table 5.1



다음은 예제 5.1에서 사용한 R code 이다.

```

tree<-read.table('c:/tree.txt',header=T)
attach(tree)
fit<-lm(y~x1+x2)
summary(fit)

PREDICTED_VOLUME<-fit$fitted.values
ORDINARY_VOLUME_RESIDUALS<-fit$residuals
plot(ORDINARY_VOLUME_RESIDUALS~PREDICTED_VOLUME)

fit1<-lm(y~x1+I(x1^2)+x2)
summary(fit1)

PREDICTED_VOLUME_<-fit1$fitted.values
ORDINARY_VOLUME_RESIDUALS_<-fit1$residuals
plot(ORDINARY_VOLUME_RESIDUALS_~PREDICTED_VOLUME_)

fit2<-lm(I(y^0.33)~x1++I(x1^2)+x2)
summary(fit2)

PREDICTED_VOLUME_<-fit2$fitted.values
ORDINARY_VOLUME_RESIDUALS_<-fit2$residuals
plot(ORDINARY_VOLUME_RESIDUALS_~PREDICTED_VOLUME_)

```

5.3. 스튜던트화 잔차(Studentized Residuals)

앞서 기술된 것과 같은 그림(plots)이 진단 도구(diagnostic tool)로서 매우 유용할 수는 있으나, 확률잡음(random noise)은 흔히 자료세트(data set)으로부터 나온 그림들을 애매하게 한다. 확실히 fig. 5.1 혹은 5.2의 전형적인 그림이 어느 정도의 규칙성(regularity)를 가지고 나오리라고 예상할 수는 없다. 잔차그림(residual plots)을 능숙하게 판독하고, 정보를 추출해 내는데 경험이 중요하다.

많은 경우, 보통 잔차그림(plots of the ordinary residuals)은 모형이나 가정(assumption)에서의 어려움을 밝히는데 적절한 그림이 아니다. 이상적으로는, 그려질 양(quantities)은 공통분산(common variance)과 서로 독립(uncorrelated)이어서 ε_i 를 아주 유사하게 모방(emulate)할 수 있어야 한다. 우리는 잔차(residuals)가 일반적으로는 무상관(uncorrelated)으로 알고 있다. 식 (5.1)로부터 자료 간에 모자행렬의 대각원소(HAT diagonals)간의 변이(variation)가 크면 잔차(residuals)의 분산(variances)이 큰 차를 보일 것이라는 것은 분명하다. 1(unity) 근처의 모자대각(HAT diagonal)은 자료 중심(data center)으로부터 멀리 떨어져 있는 자료를 정의한다는 것을 우리는 알고 있다. 회귀 적합(regression fit)이 이러한 자료에서 좋다는 것, 즉 잔차가 작다는 것은 모형의 부적절성(model inadequacy)을 완전히 가릴 수 있다. 결과적으로 가정으로부터의 이탈은, 흔히 같은 정밀도(precision)를 가지고 있는 잔차를 가지고 연구해 보면 가장 잘 알 수 있다. $\frac{e_i}{\sigma\sqrt{1-h_{ii}}}$ 형태의 스튜던트화 잔차(studentized residual)는 평균이 0이고 분산이 1(unit)이다. 따라서 아래에 있는, σ 대신에 s 가 치환된 스튜던트화 잔차(studentized residual)는 잔차진단도구(residual diagnostics)로 매우 유용하다.

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad (5.3)$$

스튜던트화 잔차(studentized residual)는 척도에 무관하며(scale-free), t 분포와 비슷하다(그렇다고 정확히 t 분포를 따르는 것은 아니다). 이들의 크기를 재기 위한 허술한 척도(crude yardstick)를 개발하는 것이 보통잔차(ordinary residual)의 경우보다 더 간단하다. 요약하면, 스튜던트화 잔차(studentized residual)가 보통잔차(ordinary residual)보다 좋은 점은 스튜던트화하면 h_{ii} 로 측정되는 회귀변수 공간(regressor space) 내에서의 자료의 위치(location)에 의한 효과를 제거한다는 것이다. 스튜던트화 잔차(studentized residual)는 많은 통계프로그램에서 회귀분석 출력 결과의 표준 부분(standard part)이 되고 있다.

예제 5.2 Education expenditure data

Table 5.3의 자료는 50개 주(state)에서 수집되었으며, 변수는 다음과 같다.

y : Per capita expenditure on education projected for 1975

x_1 : Per capita income in 1973

x_2 : Number of residents per thousand under 18 years of age in 1974

x_3 : Number of residents per thousand living on urban areas in 1970

다중회귀(multiple regression)를 하여 다음과 같은 결과를 얻었다

$$\hat{y} = -556.6 - 0.07239x_1 + 1.552x_2 - 0.004269x_3$$

$$R^2 = 0.5913, s = 40.47.$$

Table 5.4는 잔차(residuals), 모자행렬 대각원소의 수치들(HAT diagonal values), 그리고 스튜던트화 잔차(studentized residual)를 보여주고 있다. Fig. 5.6은 \hat{y}_i 에 대한 스튜던트화 잔차(studentized residual)의 그림을 보여 주고 있다. 교육에 대한 지출이 증가할수록 스튜던트화 잔차(studentized residual)가 커짐을 알 수 있다. 이 정보로부터 등분산가정(homogeneous variance assumption)이 깨졌음을 추측할 수 있다. 49번째 관측값의 경우, 보통잔차(ordinary residual)는 99.242505이고 스튜던트 잔차(studentized residual)는 3.28255765으로, 이상적인 조건 하에서는 이 잔차는 회귀변수 공간(regressor space)에서의 위치로 보아, 0으로부터 비정상적으로 멀리 떨어져 있다는 것을 말한다. 또한, 모자대각(HAT diagonal)은 0.44190992이며 이는 이 자료가 자료중심(data center)으로부터 멀리 떨어져 있음을 알려 준다. Fig. 5.6에서, 지출이 커질수록 r_i 는 더 커지며 지출이 커진다면 이 모형이 맞지 않을 가능성이 있다는 것을 의미 한다. 이 경우 많은 사람들이 등분산가정(homogeneous variance assumption)이 맞지 않는 것으로 해석하고 싶을 것이다.

Table 5.3 Education expenditure data

Case	x_1	x_2	x_3	y
1	3944	325	508	235
2	4578	323	564	231
3	4011	328	322	270
4	5233	305	846	261
5	4780	303	871	300
6	5889	307	774	317
7	5663	301	856	387
8	5759	310	889	285
9	4894	300	715	300
10	5012	324	753	221
11	4908	329	649	264

12	5753	320	830	308
13	5439	337	738	379
14	4634	328	659	342
15	4921	330	664	378
16	4869	318	572	232
17	4672	309	701	231
18	4782	333	443	246
19	4296	330	446	230
20	4827	318	615	268
21	5057	304	661	337
22	5540	328	722	344
23	5331	323	766	330
24	4715	317	631	261
25	3828	310	390	214
26	4120	321	450	245
27	3817	342	476	233
28	4243	339	603	250
29	4647	287	805	243
30	3967	325	523	216
31	3946	315	588	212
32	3724	332	584	208
33	3448	358	445	215
34	3680	320	500	221
35	3825	355	661	244
36	4189	306	680	234
37	4336	335	797	269
38	4418	335	534	302
39	4323	344	541	268
40	4813	331	605	323
41	5046	324	785	304
42	3764	366	698	317
43	4504	340	796	332
44	4005	378	804	315
45	5560	330	809	291
46	4989	313	726	312
47	4697	305	671	316

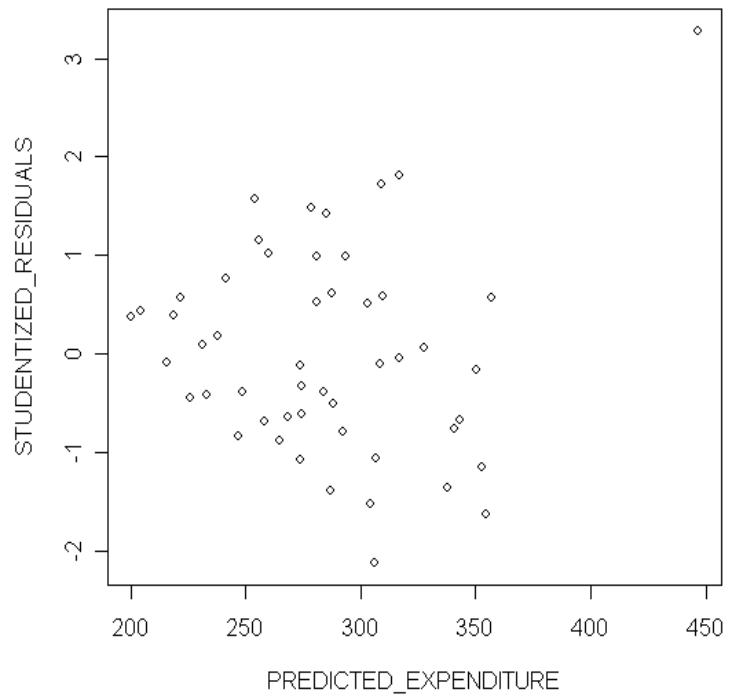
48	5438	307	909	332
49	5613	386	484	546
50	5309	333	831	311

Table 5.4 Residuals and HAT diagonal values for data of Table 5.2

Site	y_i	\hat{y}_i	Residual (e_i)	Standardized residual (r_i)	HAT Diagonal (h_{ii})
1	235	231.1686	3.831363	0.09733808	0.05403612
2	231	273.7177	-42.717746	-1.07306955	0.03240301
3	270	241.4687	28.531341	0.76022279	0.14000446
4	261	291.9893	-30.989271	-0.79197796	0.06517545
5	300	255.9879	44.012108	1.14710936	0.10118919
6	317	342.8855	-25.885512	-0.67494201	0.10192174
7	387	316.8640	70.135956	1.81318985	0.08645809
8	285	337.6406	-52.640645	-1.36266135	0.08882850
9	300	260.2496	39.750376	1.01092154	0.05598006
10	221	305.8782	-84.878174	-2.12901703	0.02956293
11	264	306.5544	-42.554355	-1.06575157	0.02655894
12	308	352.9788	-44.978752	-1.15807341	0.07896484
13	379	357.0274	21.972556	0.56248852	0.06831861
14	342	285.1260	56.873964	1.41990320	0.02041279
15	378	308.9834	69.016617	1.72824364	0.02628695
16	232	286.9874	-54.987446	-1.39139458	0.04641356
17	231	258.2083	-27.208343	-0.68549471	0.03810080
18	246	304.5214	-58.521448	-1.52825084	0.10468794
19	230	264.6732	-34.673218	-0.88717004	0.06737260
20	268	283.7637	-15.763694	-0.39576015	0.03130860
21	337	278.4872	58.512825	1.48652602	0.05400098
22	344	350.4382	-6.438174	-0.16465208	0.06647781
23	330	327.3615	2.638465	0.06661982	0.04230028
24	261	274.0362	-13.036180	-0.32651580	0.02674665
25	214	199.9949	14.005128	0.37174840	0.13341819
26	245	237.9478	7.052162	0.18084487	0.07153445
27	233	248.4972	-15.497239	-0.39609152	0.06534539
28	250	274.1350	-24.135047	-0.60699352	0.03470114
29	243	221.8095	21.190469	0.56231841	0.13293805

30	216	232.7695	-16.769464	-0.42522191	0.05040008
31	212	215.4513	-3.451339	-0.08816394	0.06432075
32	208	225.7838	-17.783804	-0.45568483	0.07006248
33	215	246.7523	-31.752273	-0.83719017	0.12171596
34	221	204.3328	16.667204	0.43005390	0.08290715
35	244	268.4633	-24.463256	-0.64255930	0.11501543
36	234	218.6797	15.320274	0.39338992	0.07398042
37	269	273.8305	-4.830468	-0.12493688	0.08729431
38	302	280.8888	21.111178	0.53156816	0.03696826
39	268	287.9508	-19.950825	-0.50432705	0.04449978
40	323	302.9697	20.030302	0.50269453	0.03060617
41	304	308.2027	-4.202665	-0.10578809	0.03637149
42	317	280.9624	36.037603	0.98511653	0.18290950
43	332	293.7557	38.244259	0.98479129	0.07917376
44	315	316.5794	-1.579391	-0.04645103	0.29413383
45	291	354.6186	-63.618583	-1.62968163	0.06954360
46	312	287.2560	24.744023	0.62115283	0.03110376
47	316	253.9378	62.062171	1.56936766	0.04514348
48	332	309.6634	22.336582	0.57737113	0.08618789
49	546	446.7575	99.242505	3.28255765	0.44190992
50	311	341.0121	-30.012115	-0.76664796	0.06430387

Figure 5.6 Plot of studentized residuals against predicted expenditure data of Table 5.2



다음은 예제 5.2에서 사용한 R code 이다.

```

education<-read.table('c:/education.txt',header=T)
attach(education)
fit<-lm(y~x1+x2+x3)
summary(fit)
PREDICTED_EXPENDITURE<-fit$fitted.values
PREDICTED_EXPENDITURE
fit$residuals
inf.m<-influence.measures(fit)
inf.m$infmat
r<-fit$residuals
h<-inf.m$infmat[,8]
STUDENTIZED_RESIDUALS<-r/(40.47*sqrt(1-h))
STUDENTIZED_RESIDUALS
plot(STUDENTIZED_RESIDUALS~PREDICTED_EXPENDITURE)

```

5.4. 표준화 PRESS 잔차에 대한 관계(Relation to Standardized PRESS Residuals)

표준화잔차(standardized residual)(혹은 스튜던트화 잔차studentized residual)와 표준화 PRESS 잔차(standardized PRESS residual)와의 관계는 흥미롭다. 지금까지 우리는 4장에서와 같이 모형을 선택하는 상황에서 PRESS 잔차(PRESS residual)를 사용하였다. 그러나 잔차분석(residual analysis)의 진단적 측면에서의 이들(PRESS 잔차)의 역할은 개념적인 관점이나 분석적인 관점 모두에서 중요하다. 식 (4.6)으로부터 i 번째 PRESS 잔차(PRESS residual)는 다음과 같이 주어짐을 상기해 보라.

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}$$

결과적으로, 식 (5.1)로부터, i 번째 PRESS 잔차(PRESS residual)의 분산(variance)은 다음과 같아 기술된다.

$$\text{Var } e_{i,-i} = \frac{1}{(1 - h_{ii})^2} [\sigma^2 (1 - h_{ii})] = \frac{\sigma^2}{1 - h_{ii}} \quad (5.4)$$

따라서 우리는 PRESS 잔차(PRESS residual)를 표준화할 수 있고, 그 결과로 다음을 얻게된다.

$$\frac{e_{i,-i}}{\sigma_{e_{i,-i}}} = \frac{\frac{e_i}{1 - h_{ii}}}{\sqrt{\frac{\sigma^2}{1 - h_{ii}}}} = \frac{e_i}{\sigma \sqrt{1 - h_{ii}}} \quad (5.5)$$

이것은 표준화 보통잔차(standardized ordinary residual)와 동일하다. 이것은 미심쩍은 자료점(suspect data points)를 검출하는 도구로 사용할 때 몇 가지 흥미로운 통찰력을 제공한다. 이 중요한 주제는 다음 절에서 설명하도록 하겠다.

5.5. 이상치 탐지(Detection of Outliers)

회귀분석가들에게 있어서, 전체 자료의 경향(trend)에 맞게 적합되지 않는 개별 자료들(individual data points)을 어떻게 처리하는가에 관한 것 보다 더 큰 딜레마는 없다. 여러 가지 이유로 인해 이러한 미심쩍은 자료들(suspect points)이 생겨나지만, 대부분은 가정 실패(assumption failure)의 범주(category)로 분류된다. 결과적으로, 문제를 잔차의 분석(analysis of residuals)으로 보는 것이 자연스럽고 편리하다. 적절한 분석도구(analytical device)를 완전하게 이해하기 위해서는, 독자들은 어떤 모형 위배(model violation)들이 의심스러운 자료(suspicious data point)를 만들어 낼 수 있는지 알아야 한다. 다음과 같은 것들이 가능하다.

1. i 번째 지점에서 모형이 붕괴(breakdown)되고, 위치이동(location shift), 즉 $E(\varepsilon_i) = \Delta_i \neq 0$ 을 발생시킨다. 이것을 가리켜, 평균이동 이상치 모형(mean shift outlier model)이라고 한다.
2. i 번째 지점에서 모형이 붕괴되고 $\text{Var}(\varepsilon_i)$ 가 다른 자료 위치(other data locations)에서의 오차 분산(error variance)보다 크다.

자료를 치우치지 않게 적합하도록 제안된 모형(proposed model fit to the balance of the data)을 따르지 않는 것처럼 보이는 관측값들에 대하여 흔히 달리 설명하기도 한다. 물론, 문제가 되는 점(the point in question)에서 이동(shift(bias))이 없거나, 오차분산(error variance)이 정말로 안정되어 있을 수도 있으나, 커다란 임의장애(random disturbance)가 우연히 생길 수도 있다. 위의 1, 2는 회귀변수(regressor variables)의 특정 위치(specific location)의 자료에서 를 것으로 예상되는 잔차(residual)에서 나타난다. 잔차의 크기는 진단(diagnostic)에서 중요한 부분이나, 모자대각(HAT diagonal)에 의하여 정량화되는, 그 지점의 위치(location of the point)도 고려되어야 한다. 이상치 검출(outlier detection)이 자연스럽게 발전한 통계량(statistic)이 PRESS 잔차(PRESS residual)이다.

$$e_{i,-i} = y_i - \hat{y}_{i,-i}$$

i 번째 지점과 연관된 모형이동(model shift), Δ_i 가 있다면, i 번째 PRESS 잔차(i th PRESS residual)의 평균(mean)은 다음과 같이 주어진다.

$$E(y_i - \hat{y}_{i,-i}) = \Delta_i$$

그러나, $e_{i,-i}$ 가 0으로부터 이탈되는 것이 단순한 우연이 아니라는 것을 결정하려면, PRESS 잔차(PRESS residual)를 표준화할 필요가 있다. 결과적으로, 진단측도(diagnostic measure)는 식 (5.5)와 같이 주어진다.

$$\frac{e_{i,-i}}{\sigma_{e_{i,-i}}} = \frac{e_i}{\sigma \sqrt{1 - h_{ii}}}$$

따라서 잔차(residual)와 PRESS 잔차(PRESS residual)는 적절하게 표준화된 경우에 이상치 검출(detection of outliers)에 똑같은 진단도구(diagnostic)가 된다. 그러므로, 실제로는 σ 가 s 로 치환된 스튜던트화 잔차(studentized residual)(식 5.3)가 이상치 검출(outlier detection)을 하는데 사용되는 현실적인 진단도구(diagnostic)가 된다. 간단히 말해서, 표준적이고 이상적인 조건에서 예상할 수 있는 것보다 잔차(residual) 혹은 PRESS 잔차(PRESS residual)가 0으로부터 더 많이 떨어져 있는지 결정하기 위하여 스튜던트화 잔차(studentized residual)를 사용한다. 어떤 경우에는, 식 (5.5)의 σ 를 대신할 수 있는 새로운 추정량(alternative estimator)을 약간 더 선호할 수도 있다. 이것은 다음에서 제시된다.

내외부적 스튜던트화: 이상치 진단도구로서의 R 스튜던트 통계량(Internal and External Studentization: The R-Student Statistic as an Outlier Diagnostic)

식 (5.3)의 r_i 스튜던트화 잔차(studentized residual)는 이상치 검출(outlier detection)에 대한 진단도구(diagnostic tool)이다. 그러나, 이 통계량(statistic)을 잘 살펴보면, 스튜던트화(studentization)에 σ 를 대신할 수 있는 새로운 추정값(alternative estimate)을 사용할 수 있음을 알 수 있다. 만약 이상치(outlier)가 이전 페이지의 항목(item) 1의 경우처럼 모형 붕괴(model breakdown)로부터 발생하였다면, 즉 평균이동 이상치(mean shift outlier)라면 우리는 식 (3.5)에 기술된 오차평균제곱(error mean square)인 s^2 이 위쪽으로 편향(upward bias)되는 것을 알 수 있다. 모형 오설정(model misspecification) 상황에서 s^2 의 편향(bias)에 관하여 4.3절에서 개념적으로 관찰하였으므로 독자는 이를 쉽게 알 수 있을 것이다. 이러한 평균이동이상(mean shift anomaly)은 단순한 모형 오설정(model misspecification)이다. 새로운 추정량(alternate estimator)은 i 번째 관측값을 이용하지 않고 계산된 제곱근 잔차평균제곱(root residual mean square)이다. PRESS 통계량(statistics)의 전개에서 4장에서 사용된 것과 비슷한 계산과정을 사용할 수 있다(부록 B.5 참조). 원하는 추정값(estimate)인 s_{-i} 는 다음과 같이 주어진다.

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - \frac{e_i^2}{1-h_{ii}}}{n-p-1}} \quad (5.6)$$

i 번째 관측값을 사용하지 않고 계산된 잔차제곱합(residual sum of squares)이 전체 자료(all data)를 이용하여 계산한 잔차제곱합(residual sum of squares)과 양(quantity) $\frac{e_i^2}{1-h_{ii}}$ 만큼 다르다

는 것이 흥미롭다.

식 (5.6)의 추정값은 σ 대신 사용되어 흔히 R 스튜던트로 불리는 외부적 스튜던트화 잔차(externally studentized residual)를 계산하는데 이용하며, R 스튜던트는 다음과 같이 주어진다.

$$t_i = \frac{y_i - \hat{y}_i}{s_{-i} \sqrt{1 - h_{ii}}}$$

(5.7)

r_i 와 t_i 의 두 진단도구(diagnostics) 간의 차이를 알아야 하며, 많은 경우 r_i 와 t_i 의 수치적인 차는 그리 크지 않다. 이들 간의 차이는 i 번째 관측값이 결과에 미치는 영향력(influence)에 따라 달라진다. 식 (5.6)은 모자행렬의 대각원소(HAT diagonal) 값이 1(unity)에 가깝고 비교적 잔차가 클 경우 s 와 s_{-i} 이 상당한 차이가 나고 따라서 r_i 와 t_i 도 차이가 나게 된다(개별 자료의 영향력에 대한 것은 6장에서 다룬다).

흔히 R 스튜던트 통계량(R-student statistic)은 부조화스러운(discordant) 자료가 있을 경우 더 예민한데, 즉 더 커진다. 정규성(normality)을 포함한 표준적인 가정(standard assumption)을 ε_i 에 대하여 할 경우, 평균이동(mean shift)이 없다는 가정 하에서 하나의 t_i 값은 t_{n-p+1} 자유도를 따른다. 따라서 student's t 분포로부터의 임계점(critical points)을 사용하고자 할 때, r_i 보다는 t_i 가 더 적합하다.

R 스튜던트 통계량(R-student statistic)을 사용하면, 가설검정(hypothesis testing)을 통하여 이상치(outliers)를 논리 형식에 맞게 검출(detection)할 수 있다. 위치이동 이상치(location shift outlier)의 경우, 가설(hypothesis)은 다음과 같이 주어진다.

$$H_0: \Delta_i = 0$$

$$H_1: \Delta_i \neq 0$$

i 번째 지점에서 오차분산(error variance)의 증가를 검출하려고 시도할 경우, 즉 5.5절의 시작에서 기술된 모형 붕괴(model breakdown)의 경우도, R 스튜던트가 적절하다. 이 경우, i 번째 자료에서 $\text{Var}(\varepsilon_i) = \sigma^2 + \sigma_i^2$ (σ_i^2 은 그 지점에서의 이상(anomaly) 때문에 오차분산(error variance)이 증가하는 것을 나타낸다)라고 가정함(postulating)으로써 모형을 보완한다. 가설은 다음과 같다.

$$H_0: \sigma_i^2 = 0$$

$$H_1: \sigma_i^2 \neq 0$$

또한 R 스튜던트 통계량(R-student statistic)이 적용된다. 어떤 이상치 모형(outlier model)이건 양쪽꼬리 t 검정 (two-tailed t -test)이 적절하다.

예제 5.3 아세틸렌자료(The Acetylene Data)

예제 3.10에서 제시된 것과 유사한 실험을 수행하여 3개의 정량적인 요인들이 *n*-헵탄(heptane)이 아세틸렌으로 전환되는 비율에 미치는 영향에 대하여 예비적인 통찰력을 얻게 되었다. 반응변수(response variable)는 *n*-헵탄(heptane)이 아세틸렌으로 전환되는 비율로, 단위는 %이며, 영향을 미칠 것으로 고려되는 요인들은 다음과 같다.

x_1 : Reactor Temperature ($^{\circ}\text{C}$)

x_2 : Ratio of H_2 to *n*-Heptane(mole ratio)

x_3 : Contact Time

세 요소 모두 실험 과정에서 조절되었다. 실험의 순서는 무작위적이었다. 아래의 자료를 이용하여 추정한 회귀식은

$$\hat{y} = -121.26962 + 0.12685x_1 + 0.34816x_2 - 19.02170x_3$$

$$R^2 = 0.9198, s^2 = 14.19.$$

자료분석에 앞서서, 자료 16로부터의 결과가 잘못되었을 수도 있다는 걱정이 있었다. 필요한 대로 상황이 항상 일정하지는 않다. 따라서 기술자들은 이 정보를 삭제하는 것이 더 낫다고 느꼈다. 그러나 처음의 분석은 자료에 관한 어림짐작을 지지하느냐 그렇지 않느냐에 관한 통찰력을 얻기 위하여 자료를 제거하지 않은 상태에서 진행되었다. Table 5.4는 16개의 관측값마다 반응(response), 적합된 반응(fitted response), 잔차(residuals), 모자대각 수치(HAT diagonal values), *R* 스튜던트 값을 보여 주고 있다.

Experiment	x_1	x_2	x_3	y
1	1300	7.5	0.0120	49.0
2	1300	9.0	0.0120	50.2
3	1300	11.0	0.0115	50.5
4	1300	13.5	0.0130	48.5
5	1300	17.0	0.0135	47.5
6	1300	23.0	0.0120	44.5
7	1200	5.3	0.0100	28.0
8	1200	7.5	0.0380	31.5
9	1200	11.0	0.0320	34.5
10	1200	13.5	0.0260	35.0
11	1200	17.0	0.0340	38.0

12	1200	23.0	0.0410	38.5
13	1100	5.3	0.0840	15.0
14	1100	7.5	0.0980	17.0
15	1100	11.0	0.0920	20.5
16	1100	7.0	0.0860	29.5

16번째 잔차(residual)는 6.947569%임을 주목하라. 이것은 이 자료세트(data set)에서 가장 큰 잔차다. 또한, 6번째 잔차 또한, -6.919752%로 다른 자료보다 큰 값을 가진다. R 스튜던트는 이 잔차가 0과 유의하게 다른가를 결정하는데 사용된다. 자료 16의 t 는 2.720784이며, 자료 6의 t 는 -2.915938이다. 이는 0.05수준 미만에서 유의하다. 다른 자료들의 R 스튜던트 값들은 이보다 작다. 결과는 이 관측값에 관한 의심에 대한 확증을 주는 것으로 보인다. 따라서 관측치 6,16을 제거하는 것이 회귀 분석을 하는데 있어서 합리적일 것이다.

Table 5.4 Results for the coal-cleansing example

y_i	\hat{y}_i	Residual (e_i)	HAT Diagonal (h_{ii})	R-student (t_i)
49.0	46.02331	2.9766900	0.2391913	0.8986299
50.2	46.54555	3.6544537	0.2040324	1.0964791
50.5	47.25137	3.2486277	0.1694015	0.9417694
48.5	48.09323	0.4067663	0.1664511	0.1133011
47.5	49.30227	-1.802274	0.1986807	-0.517906
44.5	51.41975	-6.919752	0.3550444	-2.915938
28.0	32.03937	-4.03937	0.2000349	-1.223441
31.5	32.84336	-1.343360	0.1625811	-0.375481
34.5	34.17604	0.3239578	0.2106031	0.092706
35.0	35.16057	-0.160566	0.3583251	-0.050950
38.0	36.22694	1.773055	0.2004158	0.509875
38.5	38.18274	0.317262	0.3218056	0.097956
15.0	18.51703	-3.517030	0.2558554	-1.090818
17.0	19.01667	-2.016673	0.3966297	-0.673310
20.5	20.34935	0.150645	0.2655907	0.044681
29.5	22.55243	6.947569	0.2953572	2.720784

다음은 예제 5.3에서 사용한 R code이다.

```

Acetylene<-read.table('c:/Acetylene.txt',header=T)
attach(Acetylene)
fit<-lm(y~x1+x2+x3)
summary(fit)
anova(fit)
fit$fitted.values
fit$residuals
inf.m<-influence.measures(fit)
inf.m$infmat
rstudent<-rstudent(fit)
rstudent

```

R 스튜던트에 대한 엄격한 척도(Critical yardstick for R-student)

우리는 이미 식 (5.7)에서 주어진 R 스튜던트가, 적절한 이상치 가설(outlier hypothesis) 하에서는 t_{n-p-1} 를 따른다는 것을 알았다. 따라서 미리 의심이 가는 한 개의 관측값이 있는 경우, t 검정 (t -test)과 그에 상응하는 기각값(critical values)은 적절하다. 그러나 분석가가 모든 관측값들을 동시에 철저하게 조사한다면, 전형적인 t 검정(formal t -test)을 사용하면 안 된다. 실제로 미리 확신이 없는 상태에서 가장 큰 t 값을 보이는 관측값을 검정한다는 것은 n 번의 t 검정(t -test)이 수행되고 있다는 것을 의미한다. $\alpha = 0.05$ 수준에서 이러한 검정을 하는 것은 옳지 못하다. $n = 50$ 의 표본크기(sample size)에서 가장 큰 R 스튜던트 값이 표준 $t_{0.05}$ 기각값(standard $t_{0.05}$ critical value)보다 클 확률은 0.9에 가깝다. 따라서 형식적으로 따로따로 수행된 것처럼 취급되는 다중 검정(multiple test)은 전적으로 부적절하다. 보수적이긴 하나, 유용한 R 스튜던트 통계량(R-student statistic)에 대한 기각값(critical values)은 본페로니 부등식(Bonferroni inequality)을 사용하여 계산할 수 있다. Miller(1965)와 Cook과 Weisberg(1980)를 참조하라. n 번의 검정에 대하여 α 라는 전체적인 유의수준(overall significance level)을 원한다고 가정하자.

t_{n-p-1} 분포의 $\frac{\alpha}{n} \times 100\%$ 지점을 사용하면 α 보다 크지 않은 유의수준(significance level)이 얻어진다. 이 검정은 분명히 분석 전에 의심스러운 자료(suspect point)가 없는 상황에서 안내 역할을 한다. 부록 C의 table C.4는 이러한 이상치 검정(outlier test)을 하는데 필요한 기각값(critical values)을 제공한다.

예제 5.4 Education expenditure data

예제 5.2의 자료를 생각해보자. 이상치 분석(outlier analysis)이 중요하고, 분석 전에 의심스러운 자료는 없는 것으로 가정하자. 모든 관측값에 대하여 동시에 유의성 검정(significance test)을 하여 어떤 자료가 이상치(outlier)인지 결정할 필요가 있다. 다음은 50개 자료 개개의 잔차(residual), R 스튜던트 값(R-student value), 그리고 모자행렬의 대각원소(HAT diagonal) 값이다.

Site	e_i	t_i (R-student)	h_{ii}
1	3.831363	0.09627917	0.05403612
2	-42.717746	-1.07482298	0.03240301
3	28.531341	0.75664279	0.14000446
4	-30.989271	-0.78867641	0.06517545
5	44.012108	1.15109502	0.10118919
6	-25.885512	-0.67086048	0.10192174
7	70.135956	1.86100940	0.08645809
8	-52.640645	-1.37574843	0.08882850
9	39.750376	1.01111465	0.05598006

10	-84.878174	-2.21772443	0.02956293
11	-42.554355	-1.06730684	0.02655894
12	-44.978752	-1.16242592	0.07896484
13	21.972556	0.55823496	0.06831861
14	56.873964	1.43613306	0.02041279
15	69.016617	1.76760969	0.02628695
16	-54.987446	-1.40601829	0.04641356
17	-27.208343	-0.68145676	0.03810080
18	-58.521448	-1.55136204	0.10468794
19	-34.673218	-0.88503178	0.06737260
20	-15.763694	-0.39208246	0.03130860
21	58.512825	1.50683754	0.05400098
22	-6.438174	-0.16289210	0.06647781
23	2.638465	0.06589147	0.04230028
24	-13.036180	-0.32330529	0.02674665
25	14.005128	0.36821984	0.13341819
26	7.052162	0.17892269	0.07153445
27	-15.497239	-0.39241188	0.06534539
28	-24.135047	-0.60274683	0.03470114
29	21.190469	0.55806497	0.13293805
30	-16.769464	-0.42138161	0.05040008
31	-3.451339	-0.08720321	0.06432075
32	-17.783804	-0.45170169	0.07006248
33	-31.752273	-0.83437761	0.12171596
34	16.667204	0.42618917	0.08290715
35	-24.463256	-0.63837469	0.11501543
36	15.320274	0.38972631	0.07398042
37	-4.830468	-0.12358596	0.08729431
38	21.111178	0.52735322	0.03696826
39	-19.950825	-0.50017379	0.04449978
40	20.030302	0.49854576	0.03060617
41	-4.202665	-0.10463920	0.03637149
42	36.037603	0.98474108	0.18290950
43	38.244259	0.98440897	0.07917376
44	-1.579391	-0.04594205	0.29413383
45	-63.618583	-1.66042752	0.06954360

46	24.744023	0.61692464	0.03110376
47	62.062171	1.59542895	0.04514348
48	22.336582	0.57311144	0.08618789
49	99.242505	3.70992238	0.44190992
50	-30.012115	-0.76312010	0.06430387

50개의 잔차(residuals)를 검정하고 있다면, R 스튜던트 (3개의 모형 매개변수)에 대한 적절한 0.05 수준의 기각값(critical values)은 3.51이다(Table C.4, 부록 C 참조). 따라서 이상치(outlier)로 분류할 수 있는 유일한 자료는 49번째 주(state) 이다. 이 경우 R 스튜던트 값은 3.70992238이다.

다음은 예제 5.4에서 사용한 R code 이다.

```
education<-read.table('c:/education.txt',header=T)
attach(education)
fit<-lm(y~x1+x2+x3)
inf.m<-influence.measures(fit)
fit$residuals
rstudent<-rstudent(fit)
rstudent
inf.m$infmat[,8]
```

이상치 검정: 진단 혹은 형식적인 통계적 추론?(Outlier Tests: Diagnostic or Formal Statistical Inference?)

회귀분석가들이 적합의 질(quality of fit)을 높이기 위하여 자료를 제거하고 싶은 유혹을 이기지 못하는 경우가 종종 있다. 분석 결과 그것이 이상치(outlier)인 것으로 확인되더라도, 의심스러운 자료(suspect point)를 제거해야 한다고 믿을 이유는 흔히 없다. 제거할 것이냐 말 것이냐의 선택은 통계학자 단독으로 결정하여서는 안되며, 자료를 모은 사람과 같이 결정해야 한다. 불완전한 경험적 모형 개발(empirical model building)의 목적으로 회귀 자료(regression data)를 모으고 있는 한, 분석가가 생각하고 있는 모형을 만족시키지 못하는 관측값들이 있기 마련이다.

동일한 조건 하의 관측값들과 비교하였을 때, 천 번에 한번 발생하는 관측값은 나머지 자료의 경향을 단지 따르지 않는 관측값보다 자료를 모은 사람에게 다른 의미를 가지며 모형화 결함(flaw in modeling)의 원인일 수도 있다. 후자의 경우, 이러한 지점을 찾아 내면, 모르고 지나갔을 수도 있는 중요한 정보를 제공해준다. 모형화에서의 실수는 후향적으로 이해하

기가 훨씬 쉽다. 이상치 진단(outlier diagnosis)을 포함한 분석 이후에 모형을 부득이하게 변경할 수도 있다. 예를 들어, 범주형 변수(categorical variable)를 포함하거나(3장), 특정 변수에 대하여 변환(transformation)을 하는 것(7장)은 이러한 연습으로부터 유래되는 필연적인 전략이다.

많은 경우, 이상치 진단(outlier diagnosis)은 그 이상치(outlier)가 발생한 부분(region)의 추가 실험으로 이어진다. 분석가들은 이상치(outlier)가 있다는 것은 관측값이 특정 모형(specified model)과 맞지 않는다는 것만을 의미함을 명심할 필요가 있다. 이제 물론 만약 이상한 관측값(curious observation)이 연구 중인 시스템 외적인 것, 즉 더러운 시험관, 키 편치 오류, 혹은 계산 오류 등의 결과라면, 그것은 없애야 한다. 통계적 계산으로는 이러한 이상(anomaly)이 이들 범주에 들어가는 것인지 결정하는 것이 불가능하므로, 정상적으로는 관측값 삭제(deletion)는 통계적 분석 외의 고려에 의하여 결정된다.

이상치에 대한 회귀분석(outlier regression analysis)의 최근의 경향은 이 장의 앞 부분에 기술된 형식적 가설 검정의 틀(formal hypothesis testing frameworks)을 덜 강조하는 편이다. 대신에, 이상치 검정 통계량(outlier test statistics)을, 추가적인 조사가 필요한 자료 위치(data locations)가 어디인지 알려주는 진단적인 성격을 가진 것으로 생각하여야 한다. 여기서의 의미는 엄격한 기각값(tight critical values)은 중요하지 않다는 것이다. 단지 대강의 경계값(cutoff value)만이 필요하다. 예를 들어, 예제 5.4의 Education expenditure data 자료와 R 스튜던트 값으로 보면, 정식으로 이상치로 탐지된 곳은 49번뿐이었다. 사용된 방법은 모든 관찰치를 동시에 검증하는 보수적인 본페로니 형식의 검증(conservative Bonferroni type test)이었다. 그러나, 다른 형태의 R 스튜던트 값들로 부터, 이상적인 편차는 더 복잡하고, 단순히 49번째 자료만 이상치라고 간단하게 처리할 수는 없다. 한 개만 유의수준 0.05에서 유의한 것으로 판명되기는 하였으나, 2개의 자료에서 R 스튜던트 값이 2.0이 넘었다는 것을 무시하면 안 된다. 더불어서 R 스튜던트 값이 큰 것은 모두 비교적 큰 곳이었다는 것도 주목할 만하다. 분석가로 하여금 모형화의 실수를 찾아낼 수 있도록 도와줄 수 있는 것은 R 스튜던트 값(혹은 예제 5.2에서의 스튜던트화 잔차)의 이러한 진단적 관점이다. 또한 독자는 이상치의 처리는 6장에 제시된 영향력 진단(influence diagnostics)에 관한 정보와 매우 많이 관련되어 있다는 것을 고려하여야 하며, 이 둘을 적절하게 같이 혼합하여야 한다.

이상치 검출에 지표변수의 이용(Use of Indicator Variables in Outlier Detection)

5.5절까지, 통계적 추론(statistical inference)을 통하여 혹은 진단정보(diagnostic information)를 사용하여 이상치(outlier)를 찾아내는 것에 집중하였다. R 스튜던트 통계량(R 스튜던트 statistic)은 한 개의 미리 표시된 관측값(prelabeled observation)의 경우나, 보수적인 본페로니 접근 방식(conservative Bonferroni approach)을 통하여 관측값 여러 개 혹은 전부를 동시에 검정하는 경우에 사용할 수 있다. 한 개 이상의 미리 표시된 관측값(prelabeled observation)을 검정할 경우 지표변수(indicator variable)를 사용할 수도 있다. 지표변수(indicator variable)와 이상치

(outlier)를 이야기하기 전에 이 방법을 이용할 사용자에게 주의를 주어야 한다. 지표변수(indicator variable)를 이용하여 이상치(outlier)를 모형화하는 것이 나쁠 수 있다. 한 개의 혹은 한 그룹의 이상치(outlier)가 추측적으로(a priori) 표시되는 것이 중요하다. 회귀분석을 공부하는 많은 학생들이 지시변수(indicator variable)를, 3장에 기술된 것과 같이 모형의 위치이동(model shift in location)을 나타내는 것뿐이라고 생각한다. 그러나 이것은 평균이동 이상치 모형(mean shift outlier model)에 대한 우리의 정의와 같다는 것을 이해하여야 한다. 즉, i 번째 자료가 이상치(outlier)라고 하는 것은 i 번째 지점에서 회귀모형의 절편(intercept)이 이동되었다는 것을 나타낸다. 따라서 이 소절(subsection)의 내용은 한 개 이상의 자료가 의심스러운 경우에 대한, 정식 R 스튜던트 검정 통계량(formal R-student test statistic)을 알맞게 확장하는데 쓸 수 있다. 그러나, 많은 회귀분석도구와 마찬가지로 이를 남용할 가능성도 상당하다.

$r \geq 1$ 인 관측값 y_i 의 집합이 이상치(outlier)로 의심되는 경우를 생각해보자. 따라서 행렬로 표기된 모형을 생각한다면, 관측값 벡터(observation vector)는

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_r \\ \vdots \\ y_{n-r} \end{bmatrix} \quad \text{그리고} \quad \begin{bmatrix} y_1 \\ \vdots \\ y_r \\ \vdots \\ y_{n-r} \end{bmatrix} = X\beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_r \\ \vdots \\ \varepsilon_{n-r} \end{bmatrix}$$

이고, 다음을 검정하고 싶다.

$$H_0: \Delta_1 = 0 \quad (5.8)$$

여기에서 $E(\varepsilon_i) = \Delta_1$. 식 (5.8)에 대한 대립가설(alternative hypothesis)은 이상치(outlier)로 의심되는 것이 어떤 것이냐에 따라서 다른 형태를 취할 수도 있다. 다음과 같이 느껴질 수도 있다.

- 1) 모든 r 개의 가능한 이상치(outlier)들이 하나의 공통적인 평균이동(common mean shift)을 가지기 쉽다.
- 2) 모든 r 개의 이상치(outlier)들이 평균이동(mean shift)에 대하여 각기 다른 값을 가지기 쉽다.
- 3) 1)과 2)사이의 절충(compromise)

관측값 집합이 의심스럽고, 어려움의 원인(source)이 개별 관측값에 대하여 동일할 경우, 하나의 공통적인 평균이동(common mean shift)이 합리적인 대립가설(alternative hypothesis)이 될 것이다. 그러나, 가능한 이상치(outlier)들의 집합이 겉보기에 오차(error)의 크기가 다르다면 대립가설(alternative hypothesis)은 2)에 기술된 것과 비슷한 것이 사용되어야 한다. 어떤 경우는 겉보기에 이상치(outlier)로 보이는 것들이 그룹으로 나뉘어질 수 있고 따라서 그룹 내의 모든 평균이동(mean shift)들이 비슷하다고 가정할 수도 있다.

가능한 이상치 집합에 대한 공통 평균이동(Common Mean Shift for a Set of Possible Outliers)

1)의 경우를 먼저 생각해보자. 관측값 y_i 에 대하여 r 이라고 표시된 그룹이 있고, $E(\varepsilon_i) = \Delta_1 = j\gamma$ (γ 는 0이 아닌 상수이며 j 는 r 행들 중의 하나)라고 쓸 수 있다. 가설(hypothesis)은 다음과 같아진다.

$$\begin{aligned} H_0 : \gamma &= 0 \\ H_1 : \gamma &\neq 0 \end{aligned} \quad (5.9)$$

이 가설의 검정(test)은 하나는 r 로 표시된 관측값을 포함하고 다른 하나는 $n - r$ 개의 영향을 받지 않는 관측값을 포함하는 두 개의 범주(categories)가 있는 지표 변수(indicator variable)로 확대된 회귀모형(regression model)을 적합함(fitting)으로써 이루어진다. 식 (5.9)의 가설은 지표 변수(indicator variable)에 대한 t 통계량(t statistic)이 유의한가 아닌가에 따라서 기각되거나 받아들여진다.

왜 지표 변수(indicator variable)에 대한 t 검정(t -test)이 식 (5.9)의 이상치(outlier) 가설에 대한 검정인지 이해하는 것은 매우 간단하다. y_i 에서, 가능한 이상치(outlier) 집합에서 공통 평균이동(common mean shift)이 있는 상황에 대한 회귀모형(regression model)은 다음과 같이 주어진다.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_2 \end{bmatrix} = X\beta + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \gamma + \varepsilon \quad (5.10)$$

오차, ε 벡터는 표준적인 $N(0, \sigma^2 I)$ 구조를 가진다. 3장의 수식 전개로부터, 이 모형이 두 개

의 범주(catagories)를 가지는 지표변수(indicator variable)를 추가한 모형과 동일하고, 평균이동(mean shift)의 유의성(significance)에 관한 정보가 식 (5.9)의 가설에 대한 t 검정(t -test)으로부터 얻어진다는 것은 명확해진다.

공통 이동이 없는 이상치 모형(Outlier Model with No Common Shift)

표시된 이상치들(labeled outliers)이 그룹으로 나뉠 수 있는 가설적 이상치 모형(hypothetical outlier model)은 실제로는 매우 드물다. 이상치의 가능성 있는 것(possible outliers)으로 표시할 수 있는 관측값들을 흔히 볼 수 있다 하더라도, 이들을 그룹화하기는 어렵다. 예를 들어, 분석 전에 추측컨데 의심스러워 보이는 5개의 관측값이 있고 이들을 그룹화할 이유가 없다면, 5개 범주에 대한 지표 변수(indicator variable)를 가지는 모형을 통하여 이상치들의 유의성(significance of outliers)을 동시에 검증할 수 있다. 예로서,

First outlier	1	0	0	0	0
Second outlier	0	1	0	0	0
Third outlier	0	0	1	0	0
Fourth outlier	0	0	0	1	0
Fifth outlier	0	0	0	0	1
	0	0	0	0	0

	0	0	0	0	0

예제 5.5

예제 5.2의 education 자료를 생각해보자. 앞의 분석결과 3, 25, 29, 42, 44, 49와 같은 관측값이 모든 관측값을 포함하여 적합된 모형 내에서는 사실상 이상치(outlier)인가를 검정하여 결정하는 것이 중요하다. 이것은 5개의 범주를 가진 지표변수(indicator variable)를 사용하여 각 범주마다 이상치(outlier) 위치에 1을 주고 나머지 위치는 0을 주는 모형에 적합이 되어야 한다.

Table 5.5는 5개의 관측값이 평균이동 이상치(eman shift outlier)로 모형화된 결과를 보여주고 있다. 이 모형은 각 자료마다 각각의 독립적인 이동(separate shift)을 허용한다. 49번째 관측값을 나타내는 D6이 이 이상치(outlier) 모형에서 유의하나, 나머지는 유의하지 않다. 이는 예제 5.4의 R 스튜던트 계산에 비추어 보면 그리 놀랄만한 것은 아니다. 각각의 R 스튜던트 값은, 예제 5.5에서처럼 4개의 이상치(outlier)를 모형화하는 것과는 반대로, 한 개의 자료가 한 개의 이상치(outlier)로 모형화되는 모형으로부터 추론 가능한 것으로 간주할 수 있다. 독자는

이 예제의 분석이 이러한 자료의 최종 분석이 아니라는 것을 이해하여야 한다. 여기서는 단지 이상치(outlier) 모형을 설명하기 위하여 사용하였다.

Table 5.5

Call :				
Lm(formula = y ~ x1 + x2 + x3 + D1 + D2 + D3 + D4 + D5 + D6,				
Data = edu)				
Residuals:				
Min	1Q	Median	3Q	Max
-7.945e+01	-1.852e+01	-1.943e-15	1.499e+01	8.452e+01
Coefficients:				
	Estimate	Std.Error	t value	Pr(> t)
(Intercept)	-188.7685	171.62900	-1.100	0.277974
x1	0.04999	0.01356	3.686	0.000677 ***
x2	0.58244	0.44882	1.297	0.201970
x3	0.06644	0.06279	1.058	0.296323
D1	45.89763	40.13366	1.144	0.259580
D2	5.00785	40.20153	0.125	0.901489
D3	-21.11263	40.85919	-0.517	0.608199
D4	58.13859	42.64121	1.363	0.180370
D5	30.06220	46.75135	0.643	0.523878
D6	197.28443	51.57915	3.825	0.000449 ***
Signif. Codes: 0'***' 0.001'**' 0.01'*' 0.05'.' 0.1'"1				
Residual standard error: 36.33 on 40 degrees of freedom				
Multiple R-Squared: 0.7137, Adjusted R-squared: 0.6493				
F-statistic: 11.08 on 9 and 40 DF, P-value: 2.006e-08				

다음은 위 예제에 대한 R code이다.

```
education<-read.table('c:/work/data/education.txt',header=T)
attach(education)
dummy = matrix(c(0,0,1,rep(0,47), rep(0, 24), 1, rep(0, 25), rep(0, 28), 1, rep(0, 21),
                rep(0, 41), 1, rep(0, 8), rep(0, 43), 1, rep(0, 6), rep(0, 48), 1, 0), ncol=6,
                dimnames=list(NULL,c("D1", "D2", "D3", "D4", "D5", "D6")))
edu <- cbind(education, dummy)
fit <- lm(y~x1+x2+x3+D1+D2+D3+D4+D5+D6, data=edu)
```

```
summary(fit)
```

5.6. 진단그림(Diagnostic Plots)

5.2 절에서 잔차그리기(plotting of residuals)에 관하여 논의를 하였다. 이러한 그림들은 모형 가정(model assumption)이 합리적인가를 보기 위하여 사용한다. 회귀변수(regressor variable)가 한 개인 경우, x, y 의 산점도는 두 변수의 관계를 분명하게 보여주고, 모형에 곡선(curvature)이 필요한지 혹은 어떤 한 자료점이 의심스러운지(suspicious) 혹은 매우 영향을 많이 미치는지(highly influential) 결정하는데 도움이 된다. 다중회귀(multiple regression)의 경우 비슷한 그림들은 그리 정보가 많지 않다. 즉, 하나의 y 에 대하여 k 개의 x 가 있는 그림은 x 끼리의 상호 의존성(interdependency), 즉 다중공선성(multicolliearity)이 고려되지 않았기 때문에 개별 회귀 변수(individual regressor variable)의 역할을 진짜로 밝히기는 어렵다. 개별 회귀변수에 대한 잔차의 그림(plots of residual)으로부터 정보를 얻고자 하는 경우도 비슷한 약점이 존재한다. 진단그림(diagnostic plots)은 적합된 모형(fitted model)에서 개별 회귀변수의 영향을 그림으로 보여주는데 사용할 수 있다. 더불어서, 어떤 회귀변수가 변환(transform)되어야 할지 알려주는 그림(plot)도 있다. 이 두 가지 형태의 그림(plot)은 모두 잔차(residual)를 사용한다. 그러나 이러한 형태의 잔차(residual)는 회귀분석을 공부하는 학생들이 사용하는데 익숙한 것들은 아니다.

3장에서 소개된 행렬(matrix notation)로 표현된 선형회귀모형(linear regression model)을 생각해보자. 즉,

$$y = X\beta + \varepsilon$$

여기서 X 는 $n \times p$ 차원(dimension)이다.

이제 회귀변수 x_j 의 역할을 그림으로 결정하고 싶다고 하자. 우리는 X 를 다음과 같이 분할함(partitioning)으로써 복잡한 절차(machinery)를 전개하고자 한다.

$$X = [x_j : X_{-j}] \quad (5.11)$$

여기에서 x_j 는 X 의 첫 열(column)에 위치되는 x_j 의 측정값 열이다. 행렬 X_{-j} 는 x_j 를 제외한 모든 회귀변수(regressor variables)들의 측정값을 포함한다. 따라서 X_{-j} 는 $n \times (p - 1)$ 의 차원을 가진다. X_{-j} 내의 변수들에 대하여 y 를 회귀분석한다고 가정하자. 그러면 잔차(residuals)의 벡터(vector)는 다음과 같다.

$$y - X_{-j} (X'_{-j} X_{-j})^{-1} X'_{-j} y = e_{y/X_{-j}} \quad (5.12)$$

추가적으로, X_j 내의 변수들(variables)에 대하여 x_j 를 회귀분석하고 그에 해당하는 잔차(residuals)의 벡터(vector)는 다음과 같다.

$$x_j - X_{-j} (X'_{-j} X_{-j})^{-1} X'_{-j} x_j = e_{x_j/X_{-j}} \quad (5.13)$$

식 (5.12)와 (5.13)의 잔차 형태와 잔차의 표준적인 세트 $e = y - Xb = e_{y/X}$ 가 k 개의 모든 회귀변수를 포함하는 회귀분석에서, 미리 선택된 회귀변수 x_j 의 역할을 밝힐 수 있는 잔차그림(residual plots)의 형태를 만드는데 사용할 것이다. 이 세 가지 잔차의 세트가 어떻게 모자행렬(HAT matrices)로 쓰여지는지를 보면 흥미롭고 시사하는 바가 크다. 우리는 이미 $e = e_{y/X} = (I - H)y$ 라는 것을 알고 있다. 물론 여기서 $H = X(X'X)^{-1}X'$ 이다. 식 (5.12)와 (5.13)의 잔차의 세트는 $e_{y/X_{-j}} = (I - H_{-j})y$ 이고 $e_{x_j/X_{-j}} = (I - H_{-j})x_j$ 이다. 여기서 $H_{-j} = X_{-j}(X'_{-j} X_{-j})^{-1} X'_{-j}$ 이다.

회귀변수, x_j 에 대하여 필요한 진단정보(diagnostic information)의 형태를 제공할 수 있는 잔차그림(residual plots)의 형태는 다음의 네 가지이다.

- 1) 잔차 대 예측변수 그림 (Residual against predictor plots)
- 2) 편회귀 그림, 추가변수 그림 (Partial regression plots, added variable plots)
- 3) 성분잔차합 그림, 편잔차 그림 (Component plus residual plots, partial residual plots)
- 4) 덧편잔차 그림(augmented partial residual plots)

우리는 네 가지 모두 수식 전개를 해 갈 것이며 장단점에 대하여 논의해 보겠다.

잔차 대 예측변수 그림(Residual against Predictor Plots)

잔차 대 예측변수그림(residual against predictor plots)은 매우 쉬우며 사실 2장과 5.2 절에서 이의 특별한 형태를 이미 고찰한 바 있다. 이것은 단순히 $e_{y/X}$ against x_j 의 산점도(scatter plot)일 뿐이다. 이것은 모형 오설정(model misspecification)의 진단을 예제 5.1의 흑체리나무 자료를 이용하여 그림으로 설명한 그림 형태이다. 거기와 2장에서 지적한 대로 그림은 기울기가

0이고 임의 산점(random scatter)을 보여야 한다. 회귀변수(regressor) x_j 에 대한 모형 오설정(model misspecification)은 식별할 수 있는 추세(discernible trend)로 나타난다. 이러한 형태의 그림은 회귀식에서 모형항(model term) $b_j x_j$ 의 영향(impact)를 나타내는데에 효과적이지 않다.

편회귀그림, 추가변수그림(Partial Regression Plots, Added Variable Plots)

편회귀그림(partial regression plots)은 SAS를 포함한 많은 컴퓨터 패키지에 포함되어 있으며 다음과 같은 그림을 제공한다.

$$e_{y|X_{-j}} \text{ against } e_{x_j|X_{-j}}$$

여기에서 $e_{y|X_{-j}}$ 는 x_j 를 제외한 모든 회귀변수에 대한 y 의 선형의존성에서 얻어지는 잔차집합이다. 비슷한 방식으로, $e_{x_j|X_{-j}}$ 는 x_j 를 제외한 다른 모든 회귀변수에 대한 x_j 의 선형의존성에서 얻어지는 잔차집합이다. 그러므로 우리는 X_j 내의 회귀변수에 대하여 조절된(adjusted) y 와 비슷하게 조절된(adjusted) x_j 의 그림을 얻게 된다.

편회귀그림(partial regression plots)의 최소제곱기울기(least squares slope)는 완전회귀모형(complete regression model)의 회귀계수(regression coefficient)인 b_j 이다. 좀 더 이해를 돋기 위하여, 다음과 같은 완전모형(complete model)을 생각해보자.

$$y = X_{-j}\beta_{-j} + x_j\beta_j + \varepsilon = X\beta + \varepsilon$$

이제 $I - H_{-j}$ 로 양변을 곱하면 다음을 얻는다.

$$(I - H_{-j})y = (I - H_{-j})X_{-j}\beta_{-j} + (I - H_{-j})x_j\beta_j + (I - H_{-j})\varepsilon$$

$$(I - H_{-j})X_{-j} = 0 \text{ 이므로,}$$

$$e_{y|X_{-j}} = \beta_j e_{x_j|X_{-j}} + \varepsilon^*$$

여기서 $\varepsilon^* = (I - H_{-j})\varepsilon$.

이 수식 전개는 편회귀그림(partial regression plots)이 β_j 의 기울기를 가진다는 것을 의미한다. 그러므로, 우리가 지적한 대로, 최소제곱기울기(least squares slope)는 β_j 의 비편향 추정량(unbiased estimator)인 b 이다.

통계를 배우는 학생들은 위의 수식 전개로부터 다중회귀분석(multiple regression)의 개별 회귀계수(coefficients)에 대한 추론은, 종속변수(dependent variables)와 독립변수(independent variables)가 특별한 형태의 잔차(residuals)이며, 이 잔차는 회귀변수들이 조화를 이루도록 y 와 x_j 가 이미 조절(adjust)되어 있는 단순선형모형(simple linear regression, 원점을 통과)을 다루는 정도로 범위가 좁아질 수 있다는 것을 알아야 한다.

같은 직선 상에서 더 많은 정보를 얻을 수 있음을 알 수 있게 해 주는 사례가 있다. 2장에서 배운 것으로부터(식 (2.16)), 원점을 통과하는 이 단순회귀(simple regression)의 회귀계수(coefficient) b_j 는 다음과 같이 쓰여질 수 있다.

$$b_j = \frac{(e_{y|X_{-j}})'(e_{x_j|X_{-j}})}{(e_{x_j|X_{-j}})'(e_{x_j|X_{-j}})}$$

이것은 우연히도, 모든 회귀변수들(regressors)을 포함하는 다중선형회귀(multiple linear regression)에서의 β_j 의 최소제곱추정량(least squares estimator)과 동일하다.

그림(plot)의 사용예로 돌아가서, x_j 가 선형적으로 회귀분석에 들어간다면, 편회귀그림(partial regression plots)은 원점을 통과하는 선형관계를 보여야 한다. 이 그림(plot) 주변으로의 변이(variation)는 모형 항 $b_j x_j$ 의 강도(strength)를 반영한다. 또한, 이 그림(plot)은 어떤 단일 자료가 회귀분석 결과에 불균형적으로 영향(influence)을 미쳤는지 그림으로 밝혀줄 수도 있다. 특히 이 그림(plot)은 어떤 회귀계수(regression coefficients)가 이러한 영향력 있는 자료점(influential data points)에 의하여 영향을 가장 많이 받고 있는지도 결정할 수 있다. 다음의 보기가 그 설명이다.

예제 5.6 영업사원 능력 평가 자료

영업사원 50명의 잠재적인 영업 능력을 측정한 네 가지 평가 시험 점수(창의력, 단순추론,

복합추론, 계량능력)와 100을 기준으로 최근 6개월간의 영업 수익성을 평가한 영업 수익성 평가지수(y)이다. 구체적인 자료의 내용은 다음과 같다.

사원	x_1	x_2	x_3	x_4	y	사원	x_1	x_2	x_3	x_4	y
1	9	12	9	20	96.0	26	12	16	11	39	118.5
2	10	15	12	32	107.8	27	8	12	9	26	100.3
3	10	17	13	31	108.3	28	9	12	9	25	99.5
4	16	17	11	34	112.5	29	12	17	12	32	109.8
5	11	12	11	32	105.3	30	13	10	8	34	107.0
6	10	12	7	15	95.3	31	5	14	13	30	104.3
7	16	19	12	39	121.0	32	9	9	7	16	92.5
8	8	10	13	17	90.8	33	14	16	12	39	118.0
9	7	15	11	27	103.3	34	18	15	10	43	119.5
10	18	16	8	18	99.5	35	18	17	10	42	121.5
11	9	17	13	32	109.3	36	15	19	12	41	122.3
12	7	16	11	24	101.8	37	17	20	10	32	113.8
13	14	12	12	36	109.5	38	10	15	11	14	96.0
14	7	10	10	15	91.8	39	13	14	12	29	103.8
15	10	14	11	21	97.5	40	18	20	15	51	122.0
16	14	18	11	39	120.5	41	10	18	8	31	111.8
17	8	10	11	34	105.5	42	7	9	5	16	93.5
18	11	14	11	35	110.8	43	17	17	11	27	105.3
19	5	11	11	42	115.0	44	12	15	12	37	114.0
20	10	15	7	23	102.0	45	10	16	11	49	120.0
21	9	17	11	44	121.0	46	13	11	8	10	92.8
22	10	12	11	19	94.5	47	8	13	14	47	115.5
23	13	12	4	28	99.8	48	14	20	12	37	119.0
24	13	15	6	23	102.5	49	1	5	9	15	87.3
25	18	13	12	37	112.0	50	8	8	8	9	89.8

자료에 대하여 최소제곱에 의한 적합을 하면 다음과 같은 다중선형회귀모형(multiple linear regression model)을 얻는다.

$$\hat{y} = 73.15526 + 0.14245x_1 + 0.84501x_2 - 0.27220x_3 + 0.76269x_4$$

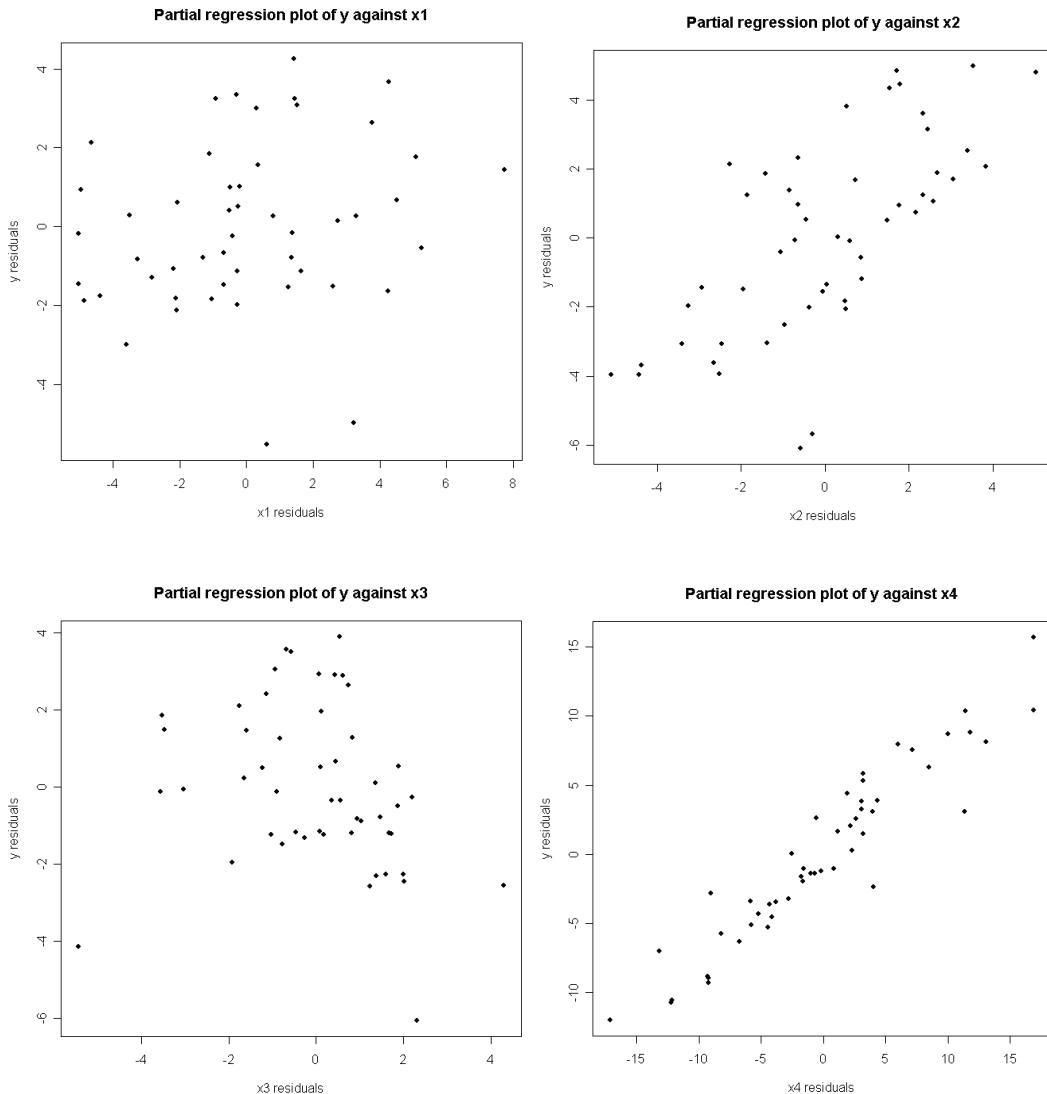
$$R^2 = 0.9591, s = 2.138$$

잔차집합(set of residuals)은 다음과 같다.

y_i	\hat{y}_i	잔차 ($e_i = y_i - \hat{y}_i$)	y_i	\hat{y}_i	잔차 ($e_i = y_i - \hat{y}_i$)
96.0	97.38142	-1.38141940	118.5	115.13554	3.36445939
107.8	108.39459	-0.59458594	100.3	101.81512	-1.51511841
108.3	109.04971	-0.74970764	99.5	101.19488	-1.69487771
112.5	112.73689	-0.23689432	109.8	110.36950	-0.56950418
105.3	106.27422	-0.97421540	107.0	107.21109	-0.21109319
95.3	94.25482	1.04518206	104.3	105.03974	-0.73973668
121.0	117.96817	3.03183400	92.5	92.34003	0.15996582
90.8	92.17207	-1.37206536	118.0	115.14824	2.85176039
103.3	104.42598	-1.12597767	119.5	118.46821	1.03179213
99.5	100.79033	-1.29033033	121.5	119.39553	2.10446749
109.3	109.66995	-0.36994833	122.3	119.35110	2.94890164
101.8	102.98291	-1.18291083	113.8	114.16119	-0.36118936
109.5	109.48013	0.01986799	96.0	94.93834	1.06166105
91.8	91.32084	0.47916010	103.8	105.68886	-1.88885571
97.5	99.43217	-1.93217244	122.0	127.43377	-5.43376721
120.5	117.11046	3.38954115	111.8	111.25573	0.54426949
105.5	105.68223	-0.18222952	93.5	92.59954	0.90046187
110.8	110.25231	0.54769331	105.3	107.54050	-2.24050365
115.0	112.20142	2.79858193	114.0	112.49295	1.50705381
102.0	102.89138	-0.89137569	120.0	122.47756	-2.47755531
121.0	119.36665	1.63334582	92.8	89.75150	3.04849857
94.5	96.21677	-1.71677281	115.5	117.31564	-1.81563676
99.8	105.41377	-5.61377129	119.0	117.00289	1.99711111
102.5	103.59093	-1.09093154	87.3	86.51330	0.78670372
112.0	111.65764	0.34236430	89.8	85.74153	4.05846953

각 회귀계수에 대한 t 통계량(t -statistic)은 1.403, 6.408, -1.618, 19.312였다. 이것은 x_1 (창의력), x_3 (복합추론)가 영업수익성에 미치는 영향이 적음을 나타낸다. 이것은 편회귀(partial

regression) 잔차도표를 통해서도 확인할 수 있다. 즉, 각 변수가 회귀분석 내에서 어떤 역할을 하는지 그림으로 통찰력을 좀 더 얻기 위하여 편회귀그림(partial regression plot)을 그려보았다.



x_1 (창의력)과 x_3 (복합추론)의 경우는 그림의 기울기가 0에 가까워서 이들 변수를 모형에 포함시키는 것이 의미가 없어 보인다. 반면에, x_2 (단순추론)과 x_4 (계량능력)의 경우는 뚜렷한 직선형태를 보이고 있어 반응변수(response variable)와 유의한 선형관계를 가지고 있음을 알 수 있다. 이는 회귀계수(regression coefficient)의 유의성 검증결과와 동일함을 확인할 수 있다.

다음은 위 예제에 대한 R code이다.

```
y <- c(96.0, 107.8, 108.3, 112.5, 105.3, 95.3, 121.0, 90.8, 103.3, 99.5, 109.3, 101.8, 109.5, 91.8,
97.5, 120.5, 105.5, 110.8, 115.0, 102.0, 121.0, 94.5, 99.8, 102.5, 112.0, 118.5, 100.3, 99.5,
109.8, 107.0, 104.3, 92.5, 118.0, 119.5, 121.5, 122.3, 113.8, 96.0, 103.8, 122.0, 111.8, 93.5,
```

```

105.3, 114.0, 120.0, 92.8, 115.5, 119.0, 87.3, 89.8)
x1 <- c(9, 10, 10, 16, 11, 10, 16, 8, 7, 18, 9, 7, 14, 7, 10, 14, 8, 11, 5, 10, 9, 10, 13, 13, 18, 12, 8, 9, 12,
13, 5, 9, 14, 18, 18, 15, 17, 10, 13, 18, 10, 7, 17, 12, 10, 13, 8, 14, 1, 8)
x2 <- c(12, 15, 17, 17, 12, 12, 19, 10, 15, 16, 17, 16, 12, 10, 14, 18, 10, 14, 11, 15, 17, 12, 12, 15, 13,
16, 12, 12, 17, 10, 14, 9, 16, 15, 17, 19, 20, 15, 14, 20, 18, 9, 17, 15, 16, 11, 13, 20, 5, 8)
x3 <- c(9, 12, 13, 11, 11, 7, 12, 13, 11, 8, 13, 11, 12, 10, 11, 11, 11, 11, 7, 11, 11, 4, 6, 12, 11, 9, 9,
12, 8, 13, 7, 12, 10, 10, 12, 10, 11, 12, 15, 8, 5, 11, 12, 11, 8, 14, 12, 9, 8)
x4 <- c(20, 32, 31, 34, 32, 15, 39, 17, 27, 18, 32, 24, 36, 15, 21, 39, 34, 35, 42, 23, 44, 19, 28, 23, 37,
39, 26, 25, 32, 34, 30, 16, 39, 43, 42, 41, 32, 14, 29, 51, 31, 16, 27, 37, 49, 10, 47, 37, 15, 9)
sale <- data.frame(y, x1, x2, x3, x4)
fit <- lm(y~x1+x2+x3+x4, sale)
summary(fit)
fit$fitt
fit$resi
d <- residuals(lm(y~x2+x3+x4, sale))
m <- residuals(lm(x1~x2+x3+x4, sale))
plot(m, d, main='Partial regression plot of y against x1',
     xlab='x1 residuals', ylab='y residuals', type="p", pch=19)
d1 <- residuals(lm(y~x1+x3+x4, sale))
m1 <- residuals(lm(x2~x1+x3+x4, sale))
plot(m1, d1, main='Partial regression plot of y against x2',
     xlab='x2 residuals', ylab='y residuals', type="p", pch=19)
d2 <- residuals(lm(y~x1+x2+x4, sale))
m2 <- residuals(lm(x3~x1+x2+x4, sale))
plot(m2, d2, main='Partial regression plot of y against x3',
     xlab='x3 residuals', ylab='y residuals', type="p", pch=19)
d3 <- residuals(lm(y~x1+x2+x3, sale))
m3 <- residuals(lm(x4~x1+x2+x3, sale))
plot(m3, d3, main='Partial regression plot of y against x4',
     xlab='x4 residuals', ylab='y residuals', type="p", pch=19)

```

일부 경험 많은 회귀분석가들은 편회귀그림(partial regression plot)의 가장 중요한 용도는 그림에 있다는 것을 안다. 모형 개발 과정(model-building procedure)에 대한 보고서를 읽는 사람들 중에는 분명히 비전문가인 사람들이 있다. 하지만 이들은 회귀계수(regression coefficients), 표준오차(standard errors), t 통계량(t -statistic) 등을 이해 혹은 식별하기는 어려우나, 각각의 회귀변수(regressor) 역할을 알아야 할 필요가 있다. 편회귀그림(partial regression plot)은 이러한

정보를 간단한 형식으로 제공한다. 편회귀그림(partial regression plot)은 일반인들이 다중회귀(multiple regression)의 기전을 볼 수 있도록 2차원 그림으로 표현해주는 유일한 도구이다.

편회귀그림(partial regression plot)은 Mosteller와 Tukey (1977)에 의하여 도입되었다. 하나 혹은 그 이상의 회귀변수를 변환할 필요가 있는지 알려 주는데 있어서 편회귀그림(partial regression plot)이 효과적이지 못한 자료세트도 있다. 때로는 성분잔차합그림(component-plus-residual plots)과 덧편잔차그림(augmented partial residual plots)이 더 효과적일 수 있다.

성분잔차합그림(Component-plus-residual Plots)

성분잔차합그림(component-plus-residual plots)은 다음 식에 의하여 그릴 수 있다.

$$e_{y|X} + x_j b_j \text{ against } x_j$$

성분잔차합(component-plus-residual)이라는 말은 이 그림(plot)의 세로축에 그려지는 것을 기술하는 것이다. Larsen과 McCleary (1972)는 이 그림을 사용할 것을 권고한 바 있다. 편회귀그림(partial regression plot)의 경우처럼, 산점도(scatter plot)의 최소제곱 기울기(least squares slope)는 b_j 이다. 이 둘을 구분 짓는 것은 가로축이 연구 중인 회귀변수(the regressor under study)가 측정되는 계량(metric)인 x_j 라는 것이다. 몇몇의 경우에는 성분잔차합그림(component-plus-residual plots)이 회귀변수 x_j 의 비선형성(nonlinearity)을 찾아내는데 더 효과적일 때도 있다.

예제 5.7 영업사원의 능력과 수익성 자료

예제 5.6의 자료를 생각해보자. Fig. 5.11은 x_1 에 대한 성분잔차합그림(component-plus-residual plots)이다. 이 그림은 x_1 과 x_3 이 특정형태 없이 다중회귀(multiple regression)에 들어가는 것을 의미한다. 즉, 설명변수(explanatory variable)가 반응변수(response variable)와 특정한 관계가 없음을 의미한다. 이를 fig. 5.7의 편회귀그림(partial regression plot)과 비교해보라. Fig. 5.12의 x_2 와 x_4 에 대한 그림은 두 설명변수(explanatory variable)가 반응변수(response variable)와 선형적인 관계로 모형에 포함되었음을 의미한다.

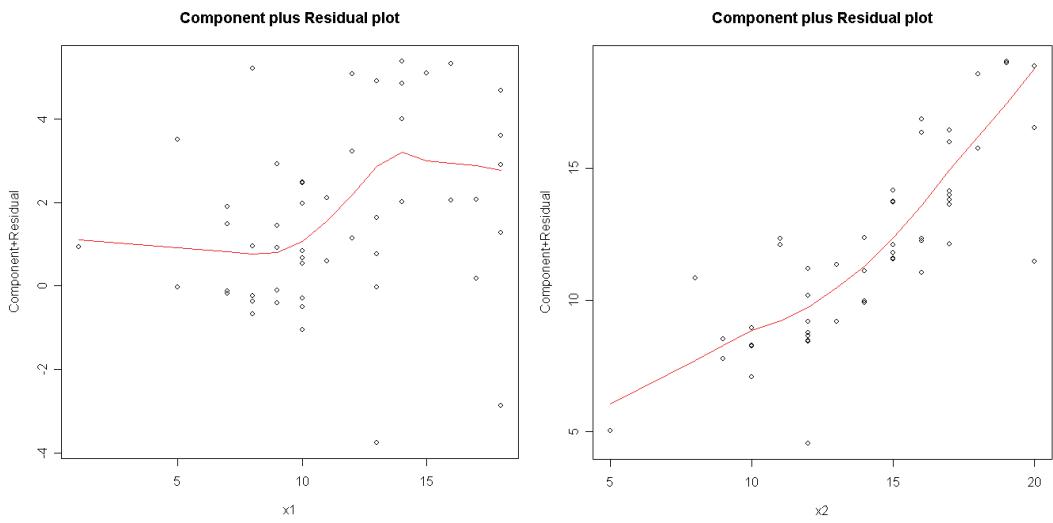


Figure 5.11

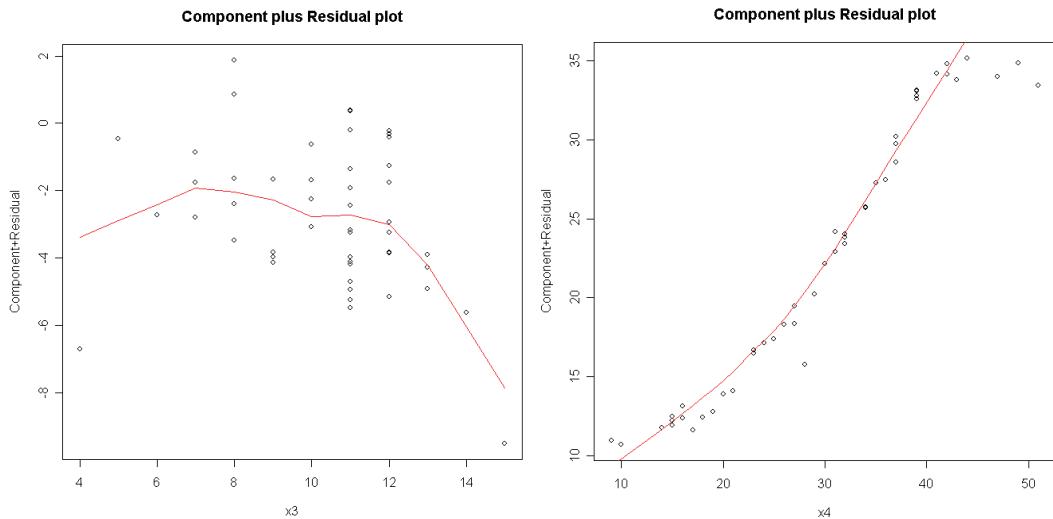


Figure 5.12

다음은 위 예제에 대한 R code이다.

```
component.residual(fit, 1, xlab = "x1", ylab = "Component+Residual")
title(main = "Component plus Residual plot")

component.residual(fit, 2, xlab = "x2", ylab = "Component+Residual")
title(main = "Component plus Residual plot")

component.residual(fit, 3, xlab = "x3", ylab = "Component+Residual")
title(main = "Component plus Residual plot")

component.residual(fit, 4, xlab = "x4", ylab = "Component+Residual")
title(main = "Component plus Residual plot")
```

덧편잔차그림(Augmented Partial Residual Plots)

5.6 절의 네 가지 진단용 그림(diagnostic plots) 중 1), 2), 3)은 특정 회귀변수(regressor)의 비

선형성(nonlinearities)을 밝히기 위하여 고안된 것들이다. 어떤 것이 가장 효과적인가는 어디에 특정하게 적용하는가에 달려있다. Mallows (1986)는 3인 편잔차그림(partial residual plot)의 변형을 도입한 바 있다. 이 그림(plot)은 비선형항(nonlinear terms)이 있을 경우, 즉 회귀변수를 변환(transform)할 필요가 있을 경우 더 민감한 것으로 보인다. 앞서 지적하였듯이, 편회귀그림(partial regression plot)은 가로축이 잔차(residual)이지 회귀변수(regressor) x_j 그 자체는 아니므로 불리한 면이 있다. 편잔차그림(partial residual plot)의 경우, Mallows는 회귀변수(regressor), x_j 의 비선형적인 면(nonlinear contribution)이 자주 모형의 다른 회귀변수(regressors)와 상관관계가 있을 수도 있다고 추측하였다. 다른 회귀변수와의 이러한 공선성(collinearity) 때문에 완전모형(complete model)의 잔차 ($e_{y|X}$ 의 잔차)에 비선형성(nonlinearity)이 나타나지 않는다. 앞의 1), 2), 3)의 방법에서는, x_j 의 비선형성(nonlinearity)을 알아내는데 잔차 그림에 서의 전체 혹은 일부가 곡선의 형태로 나타나는 것에만 의존한 것은 분명하다. 그러나 결측항(missing terms)이나 변수 변환(transformation)의 필요성을 판단하는 데 있어 가장 효과적인 방법은 모형 그 자체에서 찾아야 한다. 그러나 Mallows는 회귀변수 x_j 에 대해 많은 변환(transformation)을 해서 과도한 공선성(collinearity)을 발생시키기 보다는 2차항(quadratic term)을 이용한 덧편잔차그림(augmented partial residual plot)을 적합 모형과 그림에서 이용하기를 제안했다. 여기서 이차항(quadratic term)은 적합된 모형이나 그림 모두에 사용된다. 따라서 분석가는 변수 x_j 에 선형적인 항(linear term)과 이차항(quadratic term)이 같이 들어 있는 모형을 적합하는 것이다. 그러면, 그림은 다음과 같다.

$$e_{y|X,x_j^2} + x_j b_j + x_j^2 b_{jj} \text{ against } x_j$$

여기에서 b_j 는 선형항(linear term)의 계수이고, b_{jj} 는 이차항(quadratic term)의 회귀계수이며, $e_{y|X,x_j^2}$ 는 적합된 모형(fitted model)의 잔차(residual)이다.

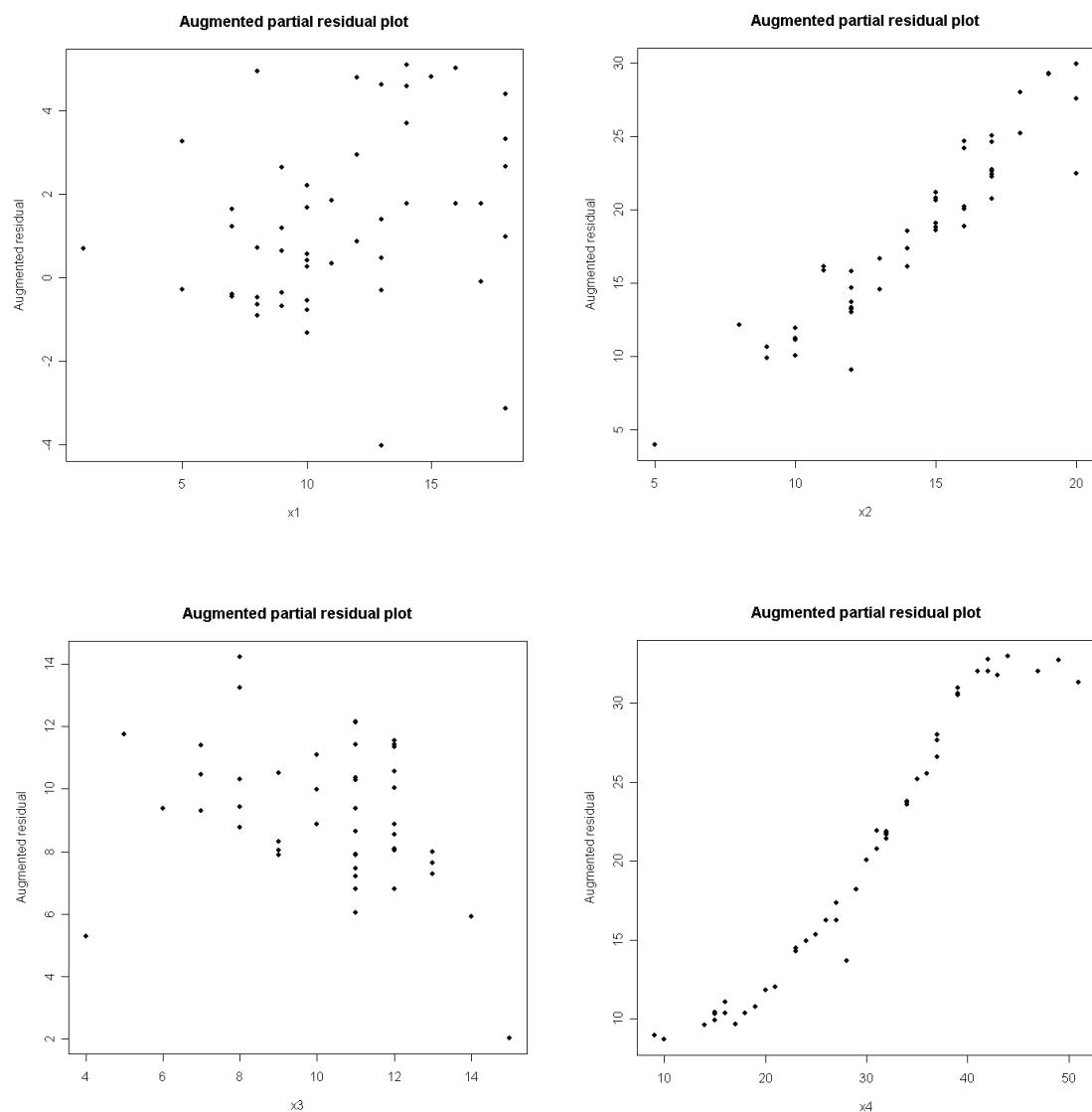
$$y = X\beta + \beta_j x_j + \beta_{jj} x_j^2 + \varepsilon$$

덧편잔차그림(augmented partial residual plot)이 오로지 이차항(quadratic term)의 도입 필요성을 탐지(detect)하기 위한 의도로 나온 것은 아니다. 이차항(quadratic term)을 도입하는 이유는

x_j 의 가능성 있는 변환(potential transform)에 대하여 간소화된 형태의 근사값(truncated approximation)을 보이기 위한 것이다. 잔차를 보아서 이 이차항(quadratic term)이 필요하다고 보이면, x_j 의 변환(transformation)이 모형 향상에 도움이 되겠다고 믿을 만한 근거가 된다.

예제 5.8

예제 5.6의 자료를 다시 한번 생각해보자. 각 인자에 선형항(linear term) 뿐만 아니라 이차항(quadratic term)을 포함하는 모형을 덧편잔차그림(augmented partial residual plot)으로 나타내었다. 덧편잔차(agmented partial residuals)는 적합된 모형(fitted model)으로부터 얻었으며, 각 인자에 대한 그림을 통해 인자의 유의성이나 변환(transformation) 필요에 대하여 좀 더 명확히 알 수 있다. 아래 그림을 보면 편잔차그림(partial residual plot)의 경우와 마찬가지로, 덧편잔차그림(augmented partial residual plot)에서도 x_1 과 x_3 에 대해서는 의미가 없음을, x_2 과 x_4 에 대해서는 의미가 있음을 알 수 있다.



다음은 위 예제에 대한 R code이다.

```
xx1 <- x1*x1
fit1 <- lm(y~x1+x2+x3+x4+xx1)
ay1 <- fit1$resi+x1*coef(fit1)['x1']+xx1*coef(fit1)['xx1']
plot(x1, ay1, main='Augmented partial residual plot', xlab='x1', ylab='Augmented residual', type="p", pch=19)
xx2 <- x2*x2
fit2 <- lm(y~x1+x2+x3+x4+xx2)
ay2 <- fit2$resi+x2*coef(fit1)['x2']+xx2*coef(fit2)['xx2']
plot(x2, ay2, main='Augmented partial residual plot', xlab='x2', ylab='Augmented residual', type="p", pch=19)
xx3 <- x3*x3
fit3 <- lm(y~x1+x2+x3+x4+xx3)
ay3 <- fit3$resi+x3*coef(fit3)['x3']+xx3*coef(fit3)['xx3']
plot(x3, ay3, main='Augmented partial residual plot', xlab='x3', ylab='Augmented residual', type="p", pch=19)
xx4 <- x4*x4
fit4 <- lm(y~x1+x2+x3+x4+xx4)
ay4 <- fit4$resi+x4*coef(fit4)['x4']+xx4*coef(fit4)['xx4']
plot(x4, ay4, main='Augmented partial residual plot', xlab='x4', ylab='Augmented residual', type="p", pch=19)
```

진단그림의 성능에 대한 요약(Summary of Performance of Diagnostic Plots)

여기서 논의된 진단 그림(diagnostic plot)은 적합된 회귀(fitted regression)에 대해 그래프를 통해 여러가지 측면을 부각시키기 위해 사용되었다. 자세히 언급하자면, 변환(transform)의 필요성, 회귀변수 각각의 상대적 중요도 그리고 명백하게 나타나는 영향력이 큰 관측값 (highly influential observations)이 그것이다. 6장에서 고영향력 관측값 (highly influential observations)을 찾는 방법에 대한 논의가 있겠지만, 지금까지의 예제들로 보아, 어떤 관측값들이 적합된 모형(fitted model)의 개별 변수의 역할을 기술하는데 어떤 식으로 우세한 영향을 미치는지 그림(plot)을 이용하여 실제로 밝힐 수 있다는 것은 분명하다. 모형을 변환하지 않더라도, 이러한 그림들은 어떤 변수가 중요하며 어떤 관측값이 영향력이 높은지 보여줄 수 있다.

비선형성 검출(Detection of Nonlinearities)

덧편잔차그림(augmented partial residual plot)은 회귀변수의 비선형성(nonlinearity)을 알아내는데 아마도 가장 좋다. 많은 경우에, 편잔차그림(partial residual plots)이 편회귀그림(partial regression plots)보다 더 효과적일 수 있다. 예측변수에 대한 잔차그림(residual against predictor plots) 역시 변환(transformation)의 필요성을 밝힐 수 있다.

변수의 중요성에 대한 시각적 검출(Visual Detection of Importance of Variables)

편회귀그림(partial regression plot)은 기울기(slope)가 다중회귀의 회귀계수(multiple regression coefficient)인 직선 주위의 개별 자료의 산점(scatter)을 보여준다. 이 산점 특성(scatter property)은 b_j 주변의 표본 추출의 변동(sampling variation)을 그려주고, b_j 의 표준오차(standard error)를 그림으로 보여준다. 따라서 편회귀그림(partial regression plot)은 개별 회귀변수가 다중선형 회귀(multiple linear regression)에 들어 갈 때, 이들의 상대적 중요도를 훌륭하게 요약해준다.

고영향력 관측값의 검출(Detection of high Influential Observations)

편회귀그림(partial regression plot)은 영향력 있는 관측값(influential observations)을 찾아내는데 아마도 가장 좋을 것이다. 이 그림은 다음 장에서 다루어질 영향력 진단(influence diagnostics)에 훌륭한 보조적 그림이다.

5.7. 정규잔차도(Normal Residual Plots)

잔차는 모형 오차(model errors)에 대한 정규성 가정(normality assumption)에 대한 타당성(validity)에 관한 정보를 줄 수 있다. 진단 과정(diagnostic procedure)을 자세하게 공부할 준비가 되어 있는 독자라면 누구나, 이러한 정규성(normality)으로부터의 이탈을 찾아내는 것이 왜 중요한지 먼저 공부하여야 한다. 많은 사람들이 직접적으로 혹은 간접적으로 정규성(normality)에 대한 점검의 필요를 알고 있다. 그러나 왜 이것이 중요할까? 2장과 3장에서 가정(assumptions)에 대해서 강조했다는 것을 기억하자. 가설검정(hypothesis testing)과 신뢰구간 추정(confidence interval estimation)이 타당하기 위해서는 ε_i 의 정규성(normality)이 필요하다.

모형을 비교하는데 사용되는 예측 기준(prediction criteria; PRESS, C_p , 등)은 정규성 가정(normality assumption)을 요구하지 않는다. R^2 나 S^2 등의 모형적합성능 기준(model fitting performance criteria)의 유용성은 모형 오차(model errors)가 가우스 분포(Gaussian distribution)든 아니든 그대로 유지된다. 정규성 가정(normality assumption)은 최소제곱추정(least squares estimation)을 수행하는데도 필요하지 않다. 그러나 최소제곱추정량(least squares estimators)은 ε_i 가 비정규성일 때 보다 정규성을 따를 때 성능(performance)이 더 좋다.

따라서 비정규성 조건(nonnatural situation) 하에서 사용되고 있는 최소제곱추정량(least squares estimators)이 반드시 최적의 추정량(optimal estimator)은 아니라는 것과 좀 더 개선할 여지가 있을 수 있다는 것을 이해하여야 한다. 비가우스 분포 오차(non-Gaussian errors)가 발생할 경우, 최소제곱(least square)의 대안도 있다. 이것은 7장에서 로버스트 과정(robust procedures)이라는 명칭 하에 논의할 것이다.

2장에서 우리는 모형 오차(model errors)의 정규성(normality)으로부터의 이탈을 찾아내기 위하여 잔차(residuals)를 그리는 방법을 개발하였다. 기술적으로 같은 방법이 다중선형회귀(multiple linear regression)에도 적용되며, 이는 식 (5.3)에 의한 순위스튜던트화 잔차(ranked studentized residuals)를 그들의 예상치(expected value)에 대하여 그려보는 것이다. 이상적인 그림은 기울기가 1이면서 원점을 통과하는 것이다. 직선으로부터 많이 벗어나면 ε_i 의 정규성 가정(normality assumption)에 대한 위배가 있다는 것을 의미한다. 기술적으로 자세한 부분을 여기에서 반복하지는 않겠다. 그러나, 모형 오차(model errors)에 대한 정보를 전달하는데 있어서 잔차(residual)가 얼마나 불충분한지 다음에서 간략하게 다루겠다.

정규 직선(normal straight line)에서 벗어나는 것은 비정규 오차(nonnatural errors) 때문이기보다는 모형 오설정(model misspecification)에서 기인한다. 정규오차그림(normal error plots)의 타당성(validity)은 정확한 모형 설정(model specification)에 달려 있다. 더불어서, 표본 크기가 작다면, 편차가 존재할지라도 정규성(normality)으로부터의 편차(deviation)를 찾아내기가 매우 어렵다. 잔차(residuals)는 그 구조가, 특정 e_i 가 그에 해당하는 특정 ε_i 에 관한 이상적인 정보를 주지 못하게 되어 있다. i 번째 잔차인 e_i 가, 단지 ε_i 가 아니라 모든 ε 의 함수라는 것을 설명하기는 간단하다. 잔차의 벡터(vector of residuals)를 생각해보자.

$$\begin{aligned}
y - Xb &= [I - H]y \\
&= [I - H][X\beta + \varepsilon] \\
&= \varepsilon + X\beta - HX\beta - H\varepsilon \\
&= \varepsilon - H\varepsilon
\end{aligned}$$

h_{ij} 를 모자행렬(HAT matrix) H 의 (i, j) 요소(element)를 의미한다고 하자. 그러면 e_i 는 다음과 같다.

$$e_i = \varepsilon_i - \sum_{j=1}^n h_{ij}\varepsilon_j \quad (5.14)$$

식 (5.14)로부터 한 개의 특정 잔차 e_j 는 단지 ε_j 의 함수가 아니라 모든 오차(all errors)의 선형결합(linear combination)임을 알 수 있다. 표본이 작을 경우, $h_{ij}(i \neq j)$ 를 무시할 수는 없다. n 이 커질수록 식 (5.14)의 e_i 는 모형 오차(model error), ε_i 에 의하여 지배된다(부록 B.6). 표본 크기가 작거나 중등도이면, 개별 ε_i 가 정규분포를 하지 않을지라도 식 (5.14)의 우변(right-hand side)이 거의 정규분포를 따르는 경향이 있다(Gnadesikan (1977)). 따라서 표본크기가 작을 경우, 정규성으로부터의 이탈을 성공적으로 찾아내기가 쉽지 않다.

5.8. 잔차분석에 대한 추가해설(Further Comments on Analysis of Residuals)

앞에서는 사용자의 인내와 주의를 강조하였다. 독자가 이러한 도구들을 갑자기 많이 사용해 볼 기회가 많지는 않을 것이다. 5.2 절의 잔차그림(residual plots)의 경우, 이전에는 분명하지 않던 현상을 밝혀 내는 그림을 얻기 전에, 사용자들은 많은 자료집합에 대하여 많은 그림을 연구해야 할지도 모른다. 그림은 매우 정보량이 풍부할 수 있다. 그러나, 이러한 그림이 항상 극적인 결과를 만들어 내지는 않는다. 긍정적인 결과(positive results)가 나오는 검정결과들도 있다. 그러나 우리는 극단적인 보기들을 의도적으로 사용하였고 따라서 결과가 깨끗하고 빈틈이 없어 보였다. 사용자는 인내심이 강해야 하고 잔차 분석(residual analyses)을 계속 사용해 보아야 한다. 여기에 경험을 대체할 만한 것은 없다. 논의된 모든 분석과 그림은 사용 회귀분석 프로그램 패키지에 다 있으며, 여기서는 이들을 어떻게 사용할 것인가에 관한 기본적인 정보만을 제공하였다.

6장과 7장의 소재의 많은 부분은 이 장의 논의와 연관되어 있다. 앞에서도 말했지만, 6장은 영향력진단(influence diagnostics)을 다루며, 7장은 비이상적 조건(nonideal conditions)을 다룬 것이다. 후자에는 가정(assumptions)과 변환(transformations)의 상대적 중요성도 포함된다. 요점은 가정이 위배되었을 경우 분석하는 사람이 어디로 돌아서야 하는지에 있다. 6장은 분석하는 사람으로 하여금 이상치(outliers)와 고도 지렛대자료(high leverage data points)에 의한 영향(influence)의 정도를 결정하도록 해준다. 다시 한번 강조하지만, 잔차(residuals)는 매우 중요한 역할을 한다.

6. 영향력 진단(Influence Diagnostics)

5장에서는 가정의 위배 가능성(possible violation of assumptions)에 대하여 알아보기 위해서 잔차(residuals)를 공부하는 데에 주력하였다. 제시된 기법 중의 이상치 분석(outlier analysis)은 말하자면 의심스러운 자료값(suspect data points)을 부각시키기 위한 것이다. 분석가, 연구원, 자료를 기록하는 사람 또는 자료를 얻는 과정에서 관여하는 누구라도 이상치를 탐지(detect) 할 수 있다. 염려되는 것은, 잘못된 관측값(erroneous observation)이 회귀 결과(regression results)에 부당한 영향(an undue amount of influence) 즉, 역효과를 초래하는 것이다. 그러나, 자료(data points)가 심각한 모형 부족(model deficiency)을 보인다 하더라도, 그것으로부터 많은 것을 배울 수도 있다.

5장의 이상치 진단(outlier diagnostics)에서 제시된 상황은 y 축 방향(y -direction)의 오류(errors)였다. 즉, 측정된 반응(measured response)에서 이상(anomalies)을 초래하는 모형 이동(model shift)이었다. 물론 그것의 증상은, 우연히 발생한 것으로 간주하기에는 너무 큰 잔차(residual)였다. 그러한 관측값(observation)으로부터 도출한 회귀통계량(regression statistics)의 영향은 잘못된 측정 반응(errant measured response)쪽으로 회귀결과를 이동시킨다. 분석가는 이러한 관측값들을 찾아낼 수 있어야 하고, 그것이 예측값(predicted values), 추정 회귀계수(estimated regression coefficients), 성능 기준(performance criteria) 등에 미치는 영향(influence)의 정도를 규명할 수 있어야 한다.

이러한 영향(influence)를 규명하고, 그 정도를 평가하기 위한 방법은 표준적인 회귀 컴퓨터 패키지(standard regression computer package)에 포함되어 있다. 이 장에서 제시될 진단도구(diagnostic material)는, 최소제곱과정(least squares procedure)과 조화를 이루면서, 모형(model)을 개발하고 동시에 개발된 모형을 비판하는 기준(criteria)을 보여줄 것이다.

6.1. 영향요인(Source of Influence)

어떤 형태의 자료세트(data sets)에서는 하나의 관측값(a single observation)이나 작은 부분집합(a small subset)을 이용하여 거의 완벽히 회귀 계수(regression coefficients)를 결정할 수 있다. 이 경우에 대다수의 자료는 거의 영향(impact)을 미치지 않을 수도 있다. 무엇이 어떤 회귀자료(a regression data points)를 영향력이 큰 관측값(high influence observation)이 되도록 하는 것일까? 우선, 앞에서 언급한 대로 이상치(outlier)가 한 요인일 수 있다. 그러나, 영향력이 큰 관측값(high influence observations) 모두가 이상치(outlier) 때문인 것은 아니다. 실제로, 영향력이 큰 관측값(high influence observations) 전부가 y 축 방향에 있어서의 오류(errors in the y -direction) 때문에 발생하는 것은 아니다. 영향력있는 관측값(influential observation) 중 어떤 것은 이치에 맞고, 자료세트(data sets)의 매우 중요한 부분일 수도 있다. 하나의 관측값이 x 축 방향(x -direction)에서 극단적(extreme)일 때 영향력(influence)이 발생할 수도 있다. 즉, 그것이 적절한 관측값(proper observation)이고 모형의 오류(model fallacy)에 대한 증거를 반드시 보여주지 않는다 하더라도, x 축 자료의 중심(data centroid)으로부터 매우 크게 동떨어져 있는 경우가 이에 해당된다. 하나의 회귀변수(regressor)에 의한 예로, fig. 6.1(a)를 살펴보자. 단일 고지렛대 관측값(single high leverage observation)은 그 자체로 회귀의 기울기(slope)를 거의 완전히 결정짓는다. 그러나, 모형 오류(model fallacy)라는 관점에서, 이 관측값이 이상치(outlier)라는 증거는 없다. 반면에, fig. 6.1(b)를 살펴보면, 나머지 자료들의 경향(trend)에서 벗어난, 단일 관측값(single observation)은 아마도 모형 이동(model shift)이나 이질적 분산(heterogeneous variance)에 의해 생겼을 것이다. 비록 절편(intercept)이 관측값에 의해 매우 큰 영향을 받더라도, 추정 기울기(estimated slope)에 대한 그것의 영향(impact)은 fig. 6.1(a)에 있는 동떨어진 관측값(isolated observation)과 비교할 때 미미할 것이다.

Figure 6.1

(a) Single influential observation remote from center (b) Single observation with error in y -direction

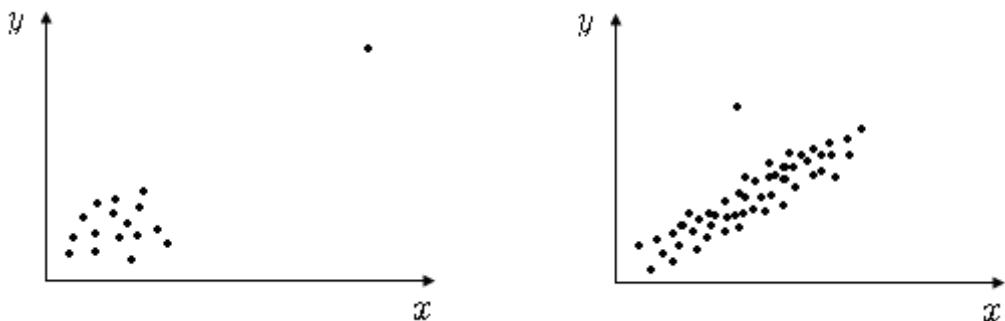


Fig. 6.1은 정도의 차이는 있겠지만 실제로 일어날 수 있는 예를 보여준다. 양자 모두 교육상 유용한 것이기 때문에 양자를 분간하는 것이 중요하다. Fig. 6.1(a)의 경우에서, 빈 공간(the

gap)을 채우는 추가적인 자료(additional data)가 있으면 도움이 될 것이다. 보충적인 정보가 없는 경우에는, 모든 가능한 방법을 동원하여 단일 고영향 관측값(single high influence observation)을 매우 조심스럽게 점검해야 한다. 전체적인 기술을 하는데 일부 정보(one piece of information)만을 이용하는 것은 아주 위험하다. Fig. 6.1(b)의 경우에, 5장에서 언급된 이상치 분석(outlier analysis)에 관한 내용이 적용된다. Fig. 6.1(b)의 이상치(outlier)에 대해서, 증상(symptom)과 진단 정보(diagnostic information)는 잔차(residual), 서로 마주 본(vis-a-vis) R -스튜던트 통계량(R -student statistic)에 있다. Fig. 6.1(a)에서, 잔차(residual)는 상대적으로 작으리라는 것을 예상할 수 있다. 이 경우에 진단 과정(diagnostic process)은 잔차(residual)와 모자 대각값(HAT diagonal value)으로 시작한다.

6.2. 진단: 잔차와 모자행렬 (Diagnostics: Residuals and the HAT Matrix)

여기에서, 개별 자료들 중 어떤 것이 회귀(regression)에 비정상적으로 큰 영향(disproportionate influence)을 미치는지를 결정하는 진단 정보(diagnostic information)에 초점을 맞추고자 한다. 우선, X 행렬(matrix)을 상기해 보자.

$$X = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix}$$

그리고, 반응 벡터(vector of responses)는

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

여기에서 i 번째 자료(i th data points)는 $[x'_i : y_i]$ 로 주어진다. 대개 진단(diagnostics)은 잔차(residuals) $e_i = y_i - \hat{y}_i$ 와 모자행렬(HAT matrix) $H = X(X'X)^{-1}X'$ 의 요소(elements)로부터 전개된다. 아래에 있는, R -스튜던트 값(R -student value)을 이용한 잔차의 표준화(standardization)는 (5장 참조) y 축 방향의 오차(errors in the y -direction)로 영향(influence)을 일으키는 자료들(data points)을 검출하는데 사용되는 적당한 진단도구(natural diagnostic)이다.

$$t_i = \frac{e_i}{s_{-i} \sqrt{1 - h_{ii}}} \quad (6.1)$$

5장에서는 가설 검정(hypothesis-testing framework)에서 이상치 검정통계량(outlier test statistic)으로서의 R -스튜던트(R -student)를 강조하였다. 그것의 진단적 가치(diagnostic value)에 대해서 초점을 맞추어보자.

R -스튜던트 값(R -student value)이 크다는 것은 i 번째 자료에서 비정상적으로 큰 적합 오차(fitting error)가 발생했다는 것을 의미하는 신호(signal)이다. R -스튜던트(R -student)는 그 관측값의 영향(observation's influence)의 크기(extent)를 평가하는 것은 아니며, 또한 어떤 통계량(statistics)이 영향을 받는지 밝혀주지도 않는다. 두번째 유형의 신호, 즉, 어떤 자료가 고지 렛대(high leverage)를 일으키는지에 관한 진단적 정보(diagnostic information)를 제공하는 것으로 모자대각(HAT diagonal)이 있다. (5.3절 참조)

$$h_{ii} = x_i' (X' X)^{-1} x_i \quad (6.2)$$

모자 대각(HAT diagonal)은 점 x_i 로부터 x 값들의 중심값(data center)인 \bar{x} 까지의 표준화된 거리(standardized distance)를 측정할 수 있도록 하기 때문에, 지렛대(leverage)의 측도(measure)가 된다. 모자 대각(HAT diagonal)은 x 값들의 극단(extreme)에 있는 관측값들(observations)을 강조할 것이다. 독자들은 h_{ii} 가 y 값들과 관련되어 있지 않음을 주목해야 한다. 따라서, 그것은 성능기준(performance criteria) 뿐 아니라 적어도 한 개의 회귀계수(regression coefficient)에 부당한 영향(undue influence)을 미칠 수 있는 관측값들을 드러내는 위험신호(red flag)로 의도되었다.

한 점이 큰 모자대각(HAT diagonal)을 가지지만, 동시에 나머지 자료에 의한 모형의 경향(trend)을 매우 잘 따른다면, 이 점은 회귀계수들(regression coefficients)에게 부당한 영향(undue influences)을 미치지는 않을 것이다. 예를 들어, fig. 6.1 (a)를 fig. 6.2와 비교하여 보자. Fig. 6.2에서는, fig. 6.1(a)에서와 달리, 멀리 떨어져 있는 관측값이 전체적인 경향(trend)을 좌우하지 않는다는 것이 명백하다. Fig. 6.2에서처럼 멀리 떨어져 있는 관측값은 기울기(slope)와 절편(intercept)에 미미한 영향을 미칠 것이다. 그러나, 그것은 기울기와 절편 모두의 표준오차(standard error)에 대하여 매우 긍정적인 효과(positive effect)를 가져올 것이다. 따라서 fig. 6.2에 있는 자료는 이미 확립된 회귀(regression)를 보강(reinforcement)하고, 따라서 적합 모형(fitted model)의 성능(performance)을 향상시키게 된다.

Figure 6.2 Large HAT diagonal but not influential observation



여기에는 그림들은 물론, 하나 이상의 회귀변수를 가진 경우에 접하게 되는 상황을 매우 단순화시킨 것이다. 그러나, R -스튜던트 값들(R -student values)과 h_{ii} 값들은 여전히, 비정상적인 영향(disproportionate influence)을 미칠 가능성성이 있는 자료들을 확인하는 적절한 진단도

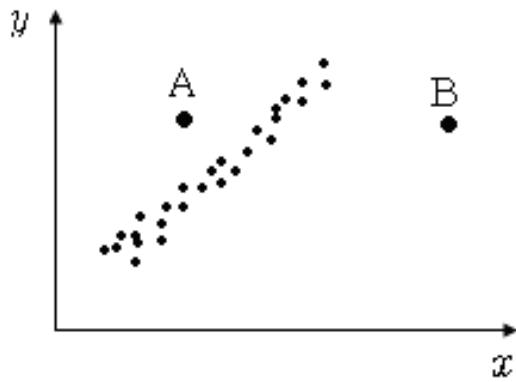
구(diagnostics)이다.

모자대각(HAT diagonal(leverage))과 문제시 되는 지점(point)에서의 모형 적합의 특성(잔차)의 영향(influence)을 파악할 수 있다. 명백히 모자대각(HAT diagonal)이 큰 관측값은 잔차가 작을 것이다. 고지렛대 점(high leverage point)에 의한 영향(influence)의 정도는 그 점이 나머지 자료들로 만들어진 모형을 얼마나 잘 따르는가의 함수가 될 것이다. 물론, 큰 모자대각(HAT diagonal)과 상대적으로 큰 잔차(residual)를 가지는 자료점(a data point)이 가장 크게 영향을 줄 것이다.

영향력이 큰 관측값을 탐지해내는 것이 왜 중요한가?(Why is It Important to Detect Influential Observations?)

지금까지 이 장에서 다룬 내용이 5장의 내용, 즉, 이상치 탐지(outlier detection)와 매우 많이 관련되었음을 알았다. 지렛대(leverage)와 이상치(outlying observation)라는 개념은 모두 관측값에 있어서 비정상적인 상태(unusual condition)를 기술한다. 고지렛대 관측값(high leverage observation) x_i (즉, 회귀 변수와 모형 항의 측정에 전적으로 관련되어 있는)는 나머지 자료로부터 멀리 떨어져 있다. 이상치(outlier)는 문제의 모형(the model in question)에 잘 맞지 않는 관측값이다. 모형 식(model formulation)이 바뀐다면, 어떤 자료(data points)가 고지렛대 점(high leverage points)이고 어떤 것이 이상치(outliers)인지가 변할 수 있다. 명백히, 모형 식(model formulation)이 변한다면, 어떤 관측값이 큰 영향력(influence)을 갖는지도 변할 수 있다. 모든 고지렛대 관측(high leverage observation)이 영향을 미치고, 모든 이상치(outlier)가 영향을 미치는가? 분명히, 어떤 고지렛대 관측값(high leverage observation)이 모형의 적합을 나쁘게 하는 관측값이라면, 자료 집합에서 그것의 존재는 회귀의 어떤 면에 큰 영향(high influence)을 미칠 것이다. 그러나, fig. 6.3을 살펴보자. 여기에서 점 B가 영향을 미치는 관측값이라는 것은 명백하다. 만약 그것을 제거한다면, 회귀선의 기울기는 극적으로 변할 것이다. 반면에, 점 A가 미치는 영향은 그리 크지 않다. 그런데도, 분명히 A는 이상치(outlier)인 것처럼 보인다. 점 B와 fig. 6.2에서 동떨어진 점(remote point)은 고지렛대 관측값(high leverage observations)이다. 그러나, 회귀의 기울기와 절편에 미치는 영향이라는 측면에서 보면, 점 B는 회귀의 기울기와 절편에 영향을 미치는 반면, fig. 6.2의 동떨어진 점(remote point)은 그렇지 않다. 그렇다면, 점 B를 이상치(outlier)라고 할 수 있을까? 자료세트(data set)에 점 B를 포함하여 얻은 적합선(fitted line)은 점 B에서 상당히 작은 잔차(residual)를 초래할 것이라는 것을 주목하라.

Figure 6.3 Point B is clearly influential



요약해보면 다음과 같다:

- 이상치(outlier)가 반드시 영향(influence)을 미치는 것은 아니다. (지렛대에 의존함)
- 고지렛대 관측값(high leverage observations)이 반드시 영향을 미치는 것은 아니다. (fig. 6.2를 주목할 것)
- 영향을 미치는 관측값들(influential observations)이 반드시 이상치(outliers)인 것은 아니다. (fig. 6.1 (a)를 주목할 것)

독자들은 영향력 진단(influence diagnostics)의 유용성을 명백히 알아야 한다. 곧 예를 통해서 살펴보겠지만, 다소 이상한 결과, 즉 결과를 설명하기 어려운 경우를 관측하게 된다. 과학적으로 말이 안되는 음의 계수(negative coefficient)가 있을 수 있다. 중요한 회귀변수(regressor)가 통계학적으로 무의미하게 나올 수도 있고, 또는 과학적으로는 이해가 되는 모형이 형편없는 성능(poor performance)을 가질 수도 있다. 회귀에서 이처럼 이치에 맞지 않는 경우는 하나 또는 몇몇의 관측값이 미치는 영향에 기인할 수 있다. “고영향(high impact)” 관측값들을 탐지해냄으로써 문제의 근본 원인을 종종 밝혀낼 수 있다.

모자 대각과 R-스튜던트: 얼마나 큰가?(HAT Diagonal, R-Student: How Large?)

잠재적인 고영향 관측값들(potentially high influence observations)을 발견해내기 위해서는 먼저, 모자대각(HAT diagonal)이 크거나 또는 R-스튜던트 값(R-student value)이 크거나 아니면 두 가지가 모두 큰 자료들을 주목해야 한다. 5장에서, R-스튜던트(R-student)의 크기가 어느 정도 클 때 특히 주의를 기울여야 하는지에 관한 지침을 살펴보았다. 그러나, 그 대부분은 가설 검정의 맥락에서 표현되었다. 여기에서는 심화된 조사를 요하는 자료를 분별해 내거나, 회귀 변수들에서 심화된 실험적 탐구(further experimental exploration)를 필요로 하는 영역(areas)을 구별하고 있다. 그 결과, 분석가는 엄격한 척도값(strict yardstick value)에 집착해서는 안된다. 물론, ± 2 지침(± 2 guidelines)은 잔차(residual)가 0으로부터 2 추정표준오차(two estimated standard error) 만큼 벗어나 있음을 의미하며, 이것은 그 점을 심도깊은 진단분석(further diagnostic analysis)를 요하는 범주(category)에 포함시켜야 함을 분명히 뜻한다.

모자대각(HAT diagonal)의 경우에, 우리는 다음의 사실을 이용해야 한다.

$$\sum_{i=1}^n h_{ii} = p \quad (6.3)$$

여기에서 p 는 모형 모수(model parameters)의 개수이다(식 3.43 참조). 결과로서, h_{ii} 의 평균 즉, $\frac{p}{n}$ 는 노름(norm)을 제공한다. 확실히, $h_{ii} > \frac{2p}{n}$ 인 어떤 h_{ii} 는 그 결과에 큰 영향력 (strong influence)을 행사할 잠재력(potential)이 있다. p 와 n 의 상대적인 크기를 고려하여야 한다. 유감스럽게도 $\frac{2p}{n} > 1$ 이고, 따라서 모자대각(HAT diagonal)이 $\frac{2p}{n}$ 를 초과할 수 없는 자료세트가 있을 수 있다. 물론, 이런 경우에는, 이 지침(guideline)이 적용되지 않는다. 예제 6.1에서, 하나의 관측값이 얼마나 중요한 회귀결과를 초래하고, 결론에 극적으로 영향을 미치는지 볼 수 있다. 그러므로, 이러한 결과들을 왜 규명되어야 하는지는 명백해진다.

예제 6.1 원반 던지기 자료

중학교 2학년 학생 중에서 15명을 임의로 추출하여 각 학생의 악력(kg), 신장(cm), 체중(kg)과 원반 던지기에서 던진 거리(m)를 측정하여 다음의 데이터를 얻었다.

X_1 : 악력

X_2 : 신장

X_3 : 체중

Y : 원반던지기 거리

Student order	X_1	X_2	X_3	Y
1	28	146	34	22
2	46	169	57	36
3	39	160	48	24
4	25	156	38	22
5	34	161	47	27
6	29	168	50	29
7	38	154	54	26
8	23	153	40	23

9	42	160	62	31
10	27	152	39	24
11	35	155	46	23
12	39	154	54	27
13	38	157	57	31
14	32	162	53	25
15	25	142	32	23

Table 5.1에 제시되어 있는 $n=31$ 관측값들에 대한 흑체리나무 자료를 사용하였던 예제 5.1을 생각해보자. 그 예제에서는 특정 가정(specific assumptions)에 위배되는 것을 강조하기 위하여 스튜던트화 잔차(studentized residuals)의 활용을 보여주고자 하였다. 특히, 등분산 가정(homogeneous variance assumption)이 흑체리나무 부피에 대한 스튜던트화 잔차(studentized residuals)의 본질을 토대로 연구되었다. 영향력 진단 연구(influence diagnostic study)의 초기 단계로, 잔차(residual), 모자대각(HAT diagonal) 값들, 그리고 R -스튜던트 값들(R -student values)을 원반던지기 자료 예제를 통하여 다시 고려해 보자.

Student order	e_i	h_{ii}	t_i (R-student)
1	0.3611	0.2976	0.1625
2	3.9291	0.5533	3.1012
3	-3.9976	0.2583	-2.0981
4	-1.2448	0.2146	-0.5365
5	-0.0368	0.1339	-0.0149
6	1.3981	0.4641	0.7387
7	-1.5197	0.2077	-0.6568
8	0.4212	0.2732	0.1864
9	0.6491	0.3179	0.2973
10	0.9117	0.1317	0.3710
11	-3.0871	0.1029	-1.3320
12	-0.7211	0.2099	-0.3070
13	2.5924	0.2332	1.1916
14	-2.5547	0.2260	-1.1656
15	2.8992	0.3753	1.5359

우리는 *R*-스튜던트 값들(*R*-student values)이 큰, 하나 또는 그 이상의 관측값들을 가지고 영향력을 결정하는 것이 유익하다는 점을 알게 될지도 모른다. 우선, 2번과 3번 학생은 매우 영향력(influence)이 큰 것으로 보인다. 2번 학생과 3번 학생의 *R*-스튜던트 값(*R*-student values)의 절대값은 모두 크기가 2.0을 초과하고, 특히 2번 학생의 경우 모자 대각(HAT diagonal)이 $\frac{2p}{n} = \frac{8}{15}$ 의 지침(guideline)을 약간 초과하여 강한 영향력(influence)을 행사할 가능성이 있을 것이라 판단된다. *R*-스튜던트 값(*R*-student values)의 절대값이 가장 큰 2번 학생이 결과에 미치는 영향의 정도를 보기 위해서 계수와 추정 표준오차(estimated standard errors)와 R^2 그리고 2번 학생의 자료가 있을 때와 없을 때의 s 를 포함하는 결과를 table 6.1에 제시하였다.

이 두 결과의 차이는 아주 크다. 변수들의 역할에 관한 결론이 크게 바뀌게 됨을 볼 수 있다. 자료 2를 없애기 전에는 모든 설명변수에 대한 계수가 양의 값을 가지며, $H_0: \beta = 0$ 에 대한 *t*-값으로 각각 1.093, 1.300, 0.750의 값을 가진다. 따라서 *t*-통계량(*t*-statistic)에 근거하면 계수(coefficients) b_1, b_2, b_3 는 통계학적으로 무의미하다. 자료 2를 제거한 후에는 계수 b_1, b_2 는 음의 값을 바꿔지만, *t*-값은 -0.763, -0.253으로 여전히 통계학적으로 무의미하다. 계수 b_3 는 양의 값을 남아 있지만 *t*-값이 2.544으로 증가하여, *t*-통계량에 근거할 때 계수 b_3 가 유의하다는 결론을 내릴 수 있게 된다.

Table 6.1 Results with and without Student 2

With Student 2	Without Student 2
$R^2 = 0.6913$	$R^2 = 0.7137$
$s = 2.532$	$s = 1.896$
$b_0 = -13.2173$	$b_0 = 16.70499$
$b_1 = 0.2014$	$b_1 = -0.13365$
$b_2 = 0.1710$	$b_2 = -0.02986$
$b_3 = 0.1249$	$b_3 = 0.38057$
$t_{b0} = -0.751$	$t_{b0} = 1.023$
$t_{b1} = 1.093$	$t_{b1} = -0.763$
$t_{b2} = 1.300$	$t_{b2} = -0.253$
$t_{b3} = 0.750$	$t_{b3} = 2.544$

예제 6.1은 하나의 관측값 때문에 잘 수행된 관측 연구 또는 실험 노력이 혼란스러워지고, 불확실성(uncertainty)으로 가득 차 버릴 수 있다는 것을 보여준다. 우리는 예제 5.3에서 아세틸렌 자료를 통하여 특정 관측값(6,16번째 관측값)을 이상치(outlier)로 보는 증거를 보았다. 그러나, 그것이 결과에 미치는 영향을 측정하지는 않았다. 그러나 예제 6.1에서 우리는, 여러

회귀변수들(regressors)의 역할에 관한 결론이 한 개의 관측값(두 번째 학생)에 의해 크게 영향을 받는다는 것을 알게 되었다. 5장에서 스튜던트화 잔차(studentized residuals)의 플롯(plots)에 근거하여 등분산 가정(homogeneous variance assumption)을 살펴보았다. 어떤 가정이 위배되었는지 질문해 볼 수도 있겠으나, 작용(activity)이 큰 것들의 잔차(residual)가 작용이 더 작은 것들의 잔차(residual)보다 더 크다는 것은 의심할 여지가 없다. 이러한 관측값에 대한 원인으로 여러 가지를 생각할 수 있다. 분석가는 잘못된 관측, 혹은 잘못된 모형으로 인하여 잘못된 결론을 내리기 보다는, 이에 대해서 좀더 조사해 볼 필요가 있다.

예제 6.1에서처럼 이상치를 제거하고 다시 계산하고 싶은 충동에 얹매여서는 안된다. 진단도구(diagnostics)에 대해 공부한 후에는 영향을 받는 것이 무엇인지, 각 관측값에 의해 어느 정도로 영향을 받는지 결정할 수 있는데, 이것은 단지 하나의 회귀를 계산하는 정도의 노력으로 가능하다.

다음은 예제 6.1에서 사용한 R code 이다.

```
circle<-read.table('c:/circle.txt',header=T)
attach(circle)
fit<-lm(y~x1+x2+x3)
rstudent<-rstudent(fit)
inf.m<-influence.measures(fit)
fit$residuals
inf.m$infmat
rstudent

circle1<-circle[-2,]
attach(circle1)
fit_<-lm(y~x1+x2+x3)
summary(fit)
summary(fit_)
```

6.3. 영향의 정도를 결정하는 진단도구(Diagnostics that Determine Extent of Influence)

R-스튜던트(*R*-student) (또는 스튜던트화 잔차)와 모자대각(HAT diagonal) 값들은 개별 관측값 중 어떤 것이 과도한 영향(influence)을 미칠 가능성이 있는지를 보여준다. 영향의 정도를 판정하기 위해서 개념상 유의한 일련의 통계량(statistics)을 사용할 수 있다. 표준적인 회귀 소프트웨어(standard regression computer software)를 사용하면 i 번째 자료를 나머지 자료로부터 제외할 경우에 특정 핵심통계량(certain key statistics)에 어떤 변화가 생기는지를 관측할 수 있다. 또한 어떤 통계량(statistic)이 영향을 받으며, 어떤 정도로 영향을 받는지 쉽게 진단할 수 있다. 예제 6.1에서 보았듯이, 영향력이 큰 관측값을 따로 떼어 놓는다면, 결론이 심하게 바뀔 수 있다. 이 절에서는 쉽게 이용할 수 있는 진단도구(diagnostics)의 개요를 다룰 것이며, x_i 가 있는 경우를 x_i 가 없는 경우와 대조해 볼 것이다. 우리는 i 번째 점을 물리적으로 (physically) 제거하지 않고도 진단도구(diagnostics)를 계산할 수 있다. (예제 6.1에서는 실례를 들기 위해서 두 번째 관측값을 제거하였다.) 즉, 각 관측값이 자료세트로부터 삭제되었을 때 발생하는 변화에 대한 정보를 얻을 수 있다.

우리가 얻을 수 있는 풍부한 정보는 분석가에 의해 지혜롭고, 제한적으로 사용될 때 이로울 것이다. 이 진단도구(diagnostics)는 비용이 들지 않으며, 급속히 현대 회귀분석(modern regression analysis)의 주 요소(major component)가 되었다. 계산의 용이성은 PRESS 잔차(PRESS residuals)(4장 참조)가 얻어진 단순성(simplicity)과 매우 밀접한 관련이 있다. 아래에 나와있는 i 번째 PRESS 잔차(PRESS residuals)의 식이 부록 B.4에서 전개된 것임을 상기하라.

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}}$$

이 장에서 논의할 진단도구(diagnostics)의 전개에 관심이 있다면, 부록 B.7을 참고하시오.

적합값에 대한 영향(Influence on the Fitted Value, DFFITS)

우리는 관측값 i 가 예측값(predicted value) 또는 적합값(fitted value) \hat{y}_i 에 무슨 영향을 미치는지 알게 될 것이다. 적절한 진단도구(diagnostic)는 아래와 같다.

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i} \sqrt{h_{ii}}} \quad (6.4)$$

접두사 “DF”는 x_i 가 있는 경우와 x_i 가 없는 경우의 결과의 차이(difference)를 의미한다.

이 경우에, 그것은 적합값(fitted value) \hat{y}_i 와 예측값(predicted value) $\hat{y}_{i,-i}$ (즉, i 번째 점 없이 얻어진 회귀에서 x_i 에서의 예측 반응)의 차이이다. $\text{Var } \hat{y}_i = \sigma^2 h_{ii}$ 이므로, (6.4)에서 분모는 단지 표준화(standardization)되었음을 의미한다. “- i ”는 i 번째 관측값이 계산에 포함되지 않았음을 의미한다. 따라서 i 번째 점에 대한 $(\text{DFFITS})_i$ 의 값은, i 번째 점이 자료세트에서 제거된다면 적합값(fitted value) \hat{y}_i 가 변화시키는 추정 표준오차의 수(the number of estimated standard errors)이다. $(\text{DFFITS})_i$ 에 해당하는 값을 계산하는 과정은 흥미롭고 유익하다. $(\text{DFFITS})_i$ 는 다음과 같이 계산된다.

$$\begin{aligned} (\text{DFFITS})_i &= \left[\frac{e_i}{s_{-i} \sqrt{1 - h_{ii}}} \right] \left[\frac{h_{ii}}{1 - h_{ii}} \right]^{1/2} \\ &= (R - \text{student})_i \left[\frac{h_{ii}}{1 - h_{ii}} \right]^{1/2} \end{aligned} \quad (6.5)$$

(자세한 사항은 부록 B.7을 참조하시오.) (6.5)에 있는 각 항은 하나의 회귀로부터 계산되었다. (s_{-i} 가 식 (5.6)을 사용하여 계산되었음을 상기하라.) DFFITS는 지렛대 측도 (leverage measure) $[h_{ii}/(1 - h_{ii})]^{1/2}$ 에 따라 확대되거나 축소된, 본질적으로는 R-스튜던트 값(R-student value)임을 주목하라. 자료가 이상치(larger R-student in magnitude)이거나 고지렛대 점(high leverage point: h_{ii} 가 1.0에 근접)이라면, DFFITS는 클 것이다. 그러나 만약 $h_{ii} \approx 0$ 이면, R-스튜던트(R-student)의 효과는 분명히 줄어들 것이다. 반면에, 0에 가까운 잔차(near-zero residual)가 비정상적으로 작은 $(R - \text{student})_i$ 를 초래한다면, 극도의 고지렛대(high leverage)를 가진 점의 $(\text{DFFITS})_i$ 는 작을 것이다. 따라서, 예상대로 식 (6.5)에 있는 진단도구(diagnostic)는 지렛대(covariates)와 y 축 방향의 오차(errors in the y-direction)의 영향에 의해 결정된다.

회귀 계수에 대한 영향(Influence on the Regression Coefficients)

각 회귀계수(regression coefficient)에 대해서, 영향진단도구(influence diagnostics)는 하나의 통계량(statistic)을 제공하며, 이것은 i 번째 관측값을 제외하였을 때 계수(coefficients)가 변화시키는 표준오차의 수(number of standard errors)를 알려준다. 이 통계량은 다음과 같이 정의된다.

$$(DFBETAS)_{j,i} = \frac{b_j - b_{j,-i}}{s_{-i} \sqrt{c_{jj}}} \quad (6.6)$$

여기에서 c_{jj} 는 $(X'X)^{-1}$ 의 j 번째 대각원소(diagonal element)이다. 결과로써, 분모는 j 번째 회귀계수(regression coefficient)인 b_j 의 표준 오차의 추정값(estimate of the standard error)이다.

통계량(statistic) $b_{j,-i}$ 는 i 번째 관측값을 사용하지 않고 계산한 j 번째 회귀계수(regression coefficient)를 말한다. $(DFBETAS)_{j,i}$ 의 값이 큰 것은 i 번째 관측값이 j 번째 회귀계수(regression coefficient)에 대해 꽤 큰 영향을 미친다는 것을 나타낸다. $(DFBETAS)_{j,i}$ 의 부호(sign) 또한 의미가 있다. 예를 들어, 음의 계수(negative coefficient)가 무의미하고 해석이 불가능한 가운데, j 번째 회귀계수(regression coefficient) b_j 가 음의 값인 경우를 생각해보자. 식 (6.6)에서, $(DFBETAS)_{j,i}$ 가 음의 값이고 상대적으로 크다면, 음의 계수(negative coefficient)가 i 번째 관측값에 기인했을 가능성이 있음을 알 수 있다. 이러한 상황은 i 번째 관측값에 대해서 가능한 한 많이 점검해야 할 필요가 있음을 분명히 보여준다. 아마도, 계수(coefficient)의 부호가 잘못된 경우는 하나의 잘못된 관측값 또는 관측 범위 내에서 모형의 오류(model fallacy) 때문일 수도 있다.

$(DFFITS)_i$ 에서처럼, $(DFBETAS)_{j,i}$ 의 계산은 간단하지는 않지만 매우 흥미롭다. $p \times n$ 행렬을 고려해보자.

$$R = (X'X)^{-1} X'$$

여기서 (q,s) 원소는 $r_{q,s}$ 로 표시한다. R 행렬(matrix)의 원소들(elements)이 중요한 역할을 한다는 것이 판명되었다. 실제로, R 의 j 번째 행(row)의 n 원소들(elements)에 의해서 n 관측값들이 계수(coefficient) b_j 에 가하는 지렛대(leverage)를 생성하는 것으로 간주할 수도 있다. R 의 j 번째 행(row) r'_j 로 나타낸다고 가정해보자. 그러면,

$$\begin{aligned}
 (DFBETAS)_{j,i} &= \frac{r_{j,i}}{\sqrt{r_j' r_j}} \frac{e_i}{s_{-i}(1-h_{ii})} \\
 &= \frac{r_{j,i}}{\sqrt{r_j' r_j}} \frac{1}{\sqrt{1-h_{ii}}} (R-student)_i
 \end{aligned} \tag{6.7}$$

다시, 이 진단도구(diagnostics)는 지렛대 측도(leverage measures)와 y 축 방향에서의 오차 영향(impact of errors in the y -direction)의 조합(combination)을 나타낸다. $r_{j,i}/\sqrt{r_j' r_j}$ 값은 i 번째 관측값이 j 번째 계수(coefficient)에 미치는 영향(impact)의 정규화 측도(normalized measure)이다. 이 값은 i 번째 관측값의 통상적인 지렛대 측도(usual leverage measure)인 h_{ii} 에 의해 부풀려진다. 예상된 바와 같이, $(R-student)_i$ 또한 같은 역할을 한다. 부록 B.7은 상기 결과가 유도되는 세부 내용을 보여준다.

우리는 어떤 관측값들이 특정 회귀계수(regression coefficient)에 영향을 미치는지를 확인하기 위하여 $(DFBETAS)_{j,i}$ 를 사용한다. 결과로써, 분석가는 회귀계수(regression coefficients)에 대한 영향(influence)을 평가하는데 있어서 $n \times p$ 통계량(statistic)을 관찰해야만 한다. 추가적으로, 각각의 자료로부터 얻은 다수의 정보를 하나의 숫자로 조합하여 계수 집합(coefficient set)에 대한 영향력(influence)을 측정하는 진단 통계량이 있다. 그 통계량(statistic)을 Cook's Distance 또는 Cook's D라고 하며, 스칼라 양(scalar quantity)으로 주어진다.

$$D_i = \frac{(b - b_{-i})'(X'X)(b - b_{-i})}{ps^2} \tag{6.8}$$

Cook's Distance 측도(measure)는 꽤 표준적인 영향 측도(standard influence measure)가 되어왔으며, 특정 상업화된 컴퓨터 패키지에 이것이 들어 있다. 그러나 통계량(statistics)이나 행렬 대수(matrix algebra)에 익숙치 않은 분석가들이 이해하기에는 다소 어렵다. Cook's D의 의미를 이해하기 위하여, 사용자는 먼저 아래의 벡터(vector)를 고려해야 한다.

$$d_i = b - b_{-i}$$

여기에서 b 는 계수(coefficients)의 벡터이고, b_{-i} 는 i 번째 관측값을 제외하였을 때의 계수(coefficients)의 벡터이다. 이제, i 번째 관측값이 b 의 계수들(coefficients)에 미치는 영향(influence)의 복합 측도(composite measure)를 d_i 를 표준화(standardize)한 하나의 스칼라 양

(scalar quantity)으로 표현해야 한다. 적절한 표준화는 d_i 가 분산-공분산 행렬의 역(inverse of the variance-covariance matrix)에 의해 표준화된 이차 형태(quadratic form)로 만들어내는 것이다.

$Var(b) = \sigma^2 (X'X)^{-1}$ 이므로, 식 (6.8)에 있는 양(quantity)은 d_i 의 표준화된 형태(standardized version)가 된다. 실제로, Cook's Distance는 거리 측도로, 최소제곱 계수(least squares coefficients) b 와 b_{-i} 벡터 간의 표준화된 거리를 나타낸다. $X'X$ 는 양의 정부호 행렬(positive definite matrix)이므로, Cook's Distance는 양의 값(positive quantity)이다. 큰 값의 D_i 는 i 번째 관측값이 계수 집합(the set of coefficients)에 부당한 영향(undue influence)을 미친다는 것을 의미한다. 어떤 특정 계수(coefficient)가 영향을 받는지 결정하기 위해서 $(DFBETAS)_{j,i}$ 에 주의를 기울여야만 한다.

다시, $(DFFITS)_i$ 와 $(DFBETAS)_{j,i}$ 처럼, D_i 는 잔차(residual) 및 자료 지렛대 측도(data point leverage measures)와 관련되어 있다. 사실, D_i 는 하나의 회귀(부록 B.7 참조)로부터 다음과 같이 매우 간단하게 계산된다.

$$D_i = \left(\frac{e_i^2}{(1-h_{ii})^2} \right) \left(\frac{h_{ii}}{s^2 p} \right) = \left(\frac{r_i^2}{p} \right) \left(\frac{h_{ii}}{1-h_{ii}} \right) \quad (6.9)$$

여기에서 r_i 는 i 번째 스튜던트화 잔차(studentized residual)이다. (실제로, 만약 s_{-i} 가 D_i 의 분모에 있는 s 대신 사용된다면, r_i 는 R -student 값인 t_i 로 대체된다.) 이전처럼, i 번째 점에서 적합도가 나쁘거나 (r_i^2 가 큼) 또는 고지렛대 (h_{ii} 가 1.0에 근접)일 때 혹은 두 경우가 모두 발생할 때, D_i 는 커진다. 또한 D_i 와 식 (6.5)의 $(DFFITS)_i$ 사이에 유사성이 있음을 주목하라. 여기에서, 적합값(fitted value)에 강한 영향을 미치는 관측값은 또한 적어도 하나의 회귀계수(regression coefficient)에 대해도 큰 영향(influence)을 미치게 될 것임을 알 수 있다.

DFFITS, DFBETAS와 Cook's D의 판단기준(How Large is Large on DFFITS, DFBETAS, and Cook's D?)

영향 진단(influence diagnostics)의 목적은 어떤 자료들(data points)이 가장 결정적으로 작용하는지 알아내는 것이다. 여기에서 실질적인 문제에 봉착할 수 밖에 없는데, 그것은 어느 정도의 값을 가져야 이러한 방법들이 하나의 신호(signal)로 활용될 수 있는가 하는 점이다. 예를 들어, “중간 정도의 영향(moderate influence)”과 “심한 영향(heavy influence)” 등을 규정하는 DFFITS의 수준(level)이 있는가?

회귀 진단(regression diagnostics)에 대한 훌륭한 교과서에서, Belsley, Kuh와 Welsch (1980)는

“절사(cutoffs)” 또는 수준(level)에 대해서 논의하였는데, 이것들은 자료를 영향력에 따라 분류하는 방법을 제공한다. 이 절사(cutoffs)의 사용에 대한 보다 상세한 이론적 설명을 위해서 그들의 저서를 참조하여야겠지만, 이 절의 후반부에서 절사(cutoffs)에 대해 다룰 것이다. 식(6.4)와 (6.6)으로부터, 두 개의 통계량(statistics) 사이의 차이를 계산하고, 적절한 표준화(standardization)를 시행함으로써 진단도구(diagnostics) DFBETAS와 DFFITS가 나타난다. 앞서 언급하였던 것처럼, 이러한 두 방법은 t -통계량과 유사하다. 그러나, \hat{y}_i 와 $\hat{y}_{i,-i}$ 는 종속 확률변수들(dependent random variables)이고, t -분포(t -distribution)는 적절한 형식적 척도(appropriate formal yardstick)를 제공하지 않는다. 물론 표준 오차(standard error)의 개념에 익숙한 분석가라면 $(DFBETAS)_{j,i}$ 가 2.0을 초과하면 그 자료의 영향(influence)은 의심할 여지가 없다는 것을 알 것이다. 그러나, 왜 2.0이라는 값이 한 점이 영향력(influence)이 있는 것으로 분류하는 최소 요구 조건(minimal requirement)으로 간주되는가? 자료세트(data sets)가 큰 경우 (예를 들어 $n > 100$)에, DFFITS나 DFBETAS 값이 2.0인 경우는 극도로 드물다. 그러나, n 이 크다 할지라도, 하나의 관측값이 어떤 형태의 조사(inspection)를 정당화하기에 충분한 정도의 영향(influence)을 주는 경우를 접하는 것이 드물지는 않다.

어느 수준(levels)의 진단도구(diagnostics)를 결정적인 것으로 간주해야 하는지에 관한 학계의 견해는 다양하다. 다음은 세 가지의 대조적인 견해이다.

1. 분석가는 표본의 크기가 작거나 보통(moderate)인 경우에, 만약에 제거된다면 회귀 결과가 상당히 변화할 것이라고 생각되는 어떤 관측값을 찾기를 바랄 것이다. 이 경우에 DFFITS와 DFBETAS에 대하여 대략 ± 2 의 척도(yardstick)를 사용하는 것이 합리적이다.
2. 표본 크기에 상관없이, 영향을 미치는 관측값(influential observations)을 찾아내는 것을 척도(yardstick)로 생각할 수 있다. 수학적인 척도(yardstick)는 분석가에게 어떤 자료점들(data points)이 이상적인 것(the ideal)에서 상당히 왜곡되었는가에 대한 느낌(impression)을 줄 수 있으며, 이때 이상적인 것(the ideal)이란 모든 자료들이 균등한 영향(uniform influence)을 미치는 상황으로 생각하면 된다. 이 척도(yardstick)는 표본 크기의 함수(a function of sample size)가 될 것이다.
3. 분석가는 절사(cutoff)나 기각값(critical value)에 대한 공식(formula)을 필요로 하지 않는다. 기준(criteria)이 의미하는 것이 무엇인지를 알고, 자료가 만들어진 체계(system)에 대한 어느 정도의 지식이 있다면, 경험에 근거한 척도값(yardstick value)에 종종 도달할 수 있을 것이다.

첫 번째 경우에서 사용된 특정 척도(yardstick)는 표본 크기(sample size)와 무관하다. 두 번째 경우는 절사값(formal cutoff values)은 표본의 크기(sample size)를 고려하여 제시되어야 한다고 제안한다. 만약 1의 견해를 선택한다면, ± 2 가 지침(guideline)이 될 수 있다. 그러나, 예

를 들어 DFBETAS에서 ± 2 의 의미를 알고 있는 사용자는 2.0 미만이라는 보다 엄격한 척도값(yardstick value)을 도입하고 싶어할 것이다.

만약 척도(yardsticks)가 필요하다면 그것이 n 에 의존적이어야 한다는 것은 의심할 여지가 없다. 그러나, n 에 근거하여 모든 경우에 적절하게 사용될 수 있는 기각값(critical values)을 만 들어낸다는 것은 매우 어렵다. Belsley, Kuh와 Welsch (1980)는 $(DFBETAS)_{j,i}$ 의 절사(cutoff)

로 $\frac{2}{\sqrt{n}}$ 를, $(DFFITS)_i$ 의 절사 (cutoff)로는 $2\sqrt{\frac{p}{n}}$ 를 제시하였다. 이것과 관련된 자세한 내용 및 가정(assumptions)을 알고 싶다면 그들의 저서를 참고하기 바란다. 물론, 이러한 값들이 분석가들이 점검하고 싶어하는 것보다 더 많은 자료점들(data points)을 노출시키는 경우도 많을 것이다. 그러나, 이러한 값들은 큰 표본(large samples)에서 이치에 맞는 경향이 있다.

위에 있는 3에 주어진 견해를 도입하기 위해서 분석가는 필요한 경험을 쌓아야만 한다. 비록 절사값(cutoff)을 얻기 위한 시도는 훌륭하다 하여도, 이러한 진단 통계량(diagnostic statistics)의 목적이 유의성 검정(significance testing)이 아니라는 점을 항상 인지하고 있어야 한다. 그러므로, 요구되는 자연 기각값(natural critical value)이라는 것은 없다. 게다가, 하나의 관측값이 회귀(regression)에 미치는 영향을 결정하기 위해 필요한 정보, 예를 들어, 특정 $(DFBETAS)_{j,i}$ 나 $(DFFITS)_i$ 과 같은 충분한 정보를 사용할 수 있어야 한다. 예를 들어,

$(DFFITS)_i = 1.70$ 이면, 이것은 그 자료점(data point)이 제거되었을 때 \hat{y}_i 가 어떻게 변화하는지를 결정하기 위한 반응(response)의 실제 단위(actual units)로 (예측값의 표준오차를 사용함으로써) 해석될 수 있다. 게다가 $(DFBETAS)_{j,i}$ 값이 가령 1.4라면, 분석가는 (계수의 추정 표준오차를 사용함으로써) 그 문제에서(in the uits of the problem) 계수(coefficient)가 어느 정도로 변화되었는지를 결정할 수 있다. 따라서, 한 개의 관측값이 한 계수(coefficient)의 유의성(significance)이나 무의미성(insignificance)에 책임이 있는지, 또는 해석이 불가능한 부호(sign)를 얻은 계수(coefficient)에 대해 책임이 있는지에 대한 평가가 필요하다. 분석가는 Belsley, Kuh와 Welsch가 제시한 지침(guidelines)을 무시해서는 안된다. 그러나 관측값이 적합값(fitted value)이나 계수(coefficient)에 어떠한 직접적인 영향을 미치는지 알기 위해 문제의 맥락에 맞게 어느 정도 노력해야 한다.

식 (6.8)에서 주어진 Cook's D는 자유도(degrees of freedom) p 와 $n-p$ 를 가지는 F -유사통계량(F -like statistic)이다. 그러나 t -분포(t -distribution)가 DFFITS와 DFBETAS에 적절하지 않은 것처럼, F -분포(F -distribution)에 근거한 기각척도(critical yardstick)도 적절하지 않다. 계수들(coefficients)에 대한 전반적인 영향(influence)을 평가하는데 있어서, 사용자는 Cook's D의 특

정 값을 다음과 같이 해석한다. Cook's D의 값이 대략 $F_{p,n-p}$ 분포(distribution)의 50% 지점과 같다고 가정해보자. i 번째 점의 삭제는 계수들의 벡터(vector of coefficients)를 신뢰영역(confidence region)의 중앙(center)으로부터 50% 공동영역(confidence ellipsoid)으로 이동시킨다고 말할 수 있다. 그러나, 많은 분석가들은 이러한 해석을 싫어할 것이고, DFBETAS 값에 더 크게 의존할 것이다.

예제 6.2 원반 던지기 자료

다시 예제 6.1의 원반던지기 자료를 고려하자. 예제 6.1에서, 우리는 자료 2가 적합도의 질(the quality of fit) 뿐 아니라 회귀계수들(regression coefficients)에 대해 강한 영향(strong influence)을 미친다는 것을 배웠다. 우리는 예를 들기 위하여 실제로 자료 2를 제거하였다. 비록 우리가 자료 2에 초점을 두었지만, 자료 3 또한 R -스튜던트에 근거하여 영향력이 있을 것이라 추측하였다. Table 6.2는 DFFITS, Cook's D와 DFBETAS 값들을 보여주는데, 15명의 학생 자료들 중 하나도 제거하지 않고서 이 모두가 미치는 영향을 보여준다.

Table 6.2 Cook's D, DFFITS, and DFBETAS

학생번호	R-Student	h_{ii}	Standard error of Prediction	DFFITS	Cook's D
1	0.1625	0.2976	1.3811	0.1058	0.0031
2	3.1012	0.5533	1.8831	3.4515	1.6700
3	-2.0981	0.2583	1.2867	-1.2383	0.2928
4	-0.5365	0.2146	1.1728	-0.2805	0.0210
5	-0.0149	0.1339	0.9264	-0.0059	0.0000
6	0.7387	0.4641	1.7247	0.6875	0.1232
7	-0.6568	0.2077	1.1540	-0.3364	0.0298
8	0.1864	0.2732	1.3232	0.1143	0.0036
9	0.2973	0.3179	1.4273	0.2030	0.0112
10	0.3710	0.1317	0.9188	0.1445	0.0057
11	-1.3320	0.1029	0.8122	-0.4512	0.0475
12	-0.3070	0.2099	1.1599	-0.1582	0.0068
13	1.1916	0.2332	1.2227	0.6572	0.1040
14	-1.1656	0.2260	1.2035	-0.6298	0.0960
15	1.5359	0.3753	1.5509	1.1906	0.3154

학생번호	DFBETAS	DFBETAS	DFBETAS	DFBETAS
------	---------	---------	---------	---------

	Intercept	b_1	b_2	b_3
1	0.0404	0.0428	-0.0250	-0.0540
2	-2.2699	2.4288	2.0391	-2.0479
3	0.5280	-0.9941	-0.5298	0.9257
4	0.0850	0.0762	-0.1297	0.0767
5	0.0038	-0.0017	-0.0040	0.0030
6	-0.4403	-0.3633	0.4557	0.0968
7	-0.2092	0.01592	0.2365	-0.1786
8	0.02148	-0.0870	-0.0077	0.0447
9	0.06160	-0.0347	-0.0905	0.1335
10	0.02718	-0.0252	-0.0030	-0.0245
11	-0.0384	-0.0252	0.0206	0.1997
12	-0.0936	-0.0177	0.1069	-0.0639
13	0.3023	-0.2086	-0.3703	0.4825
14	0.1032	0.4554	-0.0864	-0.3869
15	0.7562	0.1439	-0.5795	-0.2440

학생번호 2의 영향은 매우 명백하다. 사실, 2.4288이라는 $(DFBETAS)_{1,2}$ 값이 예이다. 만약 자료 2가 제거되면, 회귀계수(regression coefficient) b_1 은 2.4288 추정 표준오차(estimated standard error) 만큼 감소한다. 계수 b_2 , b_3 에 미치는 영향 또한 각각 2.0391, -2.0479인 DFBETAS값들로 나타내었다. 따라서, 자료 2의 존재가 모형에서 세 개의 계수들의 역할에 대한 결론뿐 아니라 그들의 부호도 바꾼다는 점은 너무 분명하여, 자료 2를 제거하고 회귀를 실제로 시행해 볼 필요도 없다. 자료 2의 영향(influence)을 더 깊이 설명하자면, Cook's D 값이 1.6700이고, DFFITS가 3.4515이라는 것을 주목하면 되겠다. 후자는 자료 2를 포함하는 것은 예측 반응(m)이 3.4515추정 표준오차(estimated standard error)만큼 증가한다는 것을 의미 한다. 분석가가 이것을 문제의 반응 단위로 변환하는데 관심이 있는 경우에, 자료 2에서의 추정된 예측값의 표준오차(estimated standard error of prediction)가 필요하다. 보통 진단도구를 표시하는 어떤 회귀 패키지(regression package)에서도 이 표준오차(standard error)를 얻을 수 있다. Table 6.2에서 자료 2에 대한 적절한 값은 1.8831 m 이다. (이것은 물론 s_{-i} 대신 s 를 사용한 추정치이다.) s_{-i} 를 사용하여 계산된 예측값의 표준오차(standard error of prediction)는 다음과 같다.

$$1.8831 \left(\frac{s_{-i}}{s} \right) = (1.8831) \left(\frac{1.896}{2.532} \right) = 1.410094$$

자료 2의 존재는 그 위치에서의 예측 반응(predicted response)을 $(3.4515)(1.410094)=4.8669 m$

만큼 증가시킨다. 이처럼 큰 변화는 두 번째 학생이 결과에 미치는 강한 효과를 강조한다. 물론, 결과를 확증하기 위해서 또는, 두 번째 학생의 존재 하에서 모형이 그렇게 크게 변화하는 비통계적 이유가 있는지를 밝혀내기 위해서 최선의 노력을 다해야 한다.

자료 3은 또한 흥미로운 예시를 제공한다. R-스튜던트 값 -2.0981은 이 점에서 상대적으로 나쁜 적합(poor fit)으로 생긴 큰 잔차(residual)를 나타낸다. 그러나, 그것이 정말로 영향을 미치는 것인가? 작은 모자대각(HAT diagonal) ($h_{ii} = 0.2583$)을 보일 경우 그 영향은 경미할 것으로 예상된다. 그리고 실제로 진단도구(diagnostics)를 살펴보면, 상대적으로 가벼운 영향(relatively light influence)을 보인다.

이 예제에서 영향 진단 통계량(influence diagnostic statistic)은 몇몇 자료는 조사할 필요가 있음을 시사한다. 진단통계량(diagnostic statistic)은 나머지 점들 (즉 영향력이 없는 점들)의 자료를 생성하는 모형들에서 2, 3번 학생보다 우세하지 않다는 것을 보여준다. 이것은, 더 복잡한 모형이 더 나은 설명을 할 수 있다는 개념을 확실히 지지하는 것이다. 더 많은 학생을 위한 독립된 (separate) 모형의 가능성을 공부하기 위해서 추가적인 학생 자료가 필요하다.

다음은 예제 6.2에서 사용한 R code이다.

```
circle<-read.table('c:/circle.txt',header=T)
attach(circle)
fit<-lm(y~x1+x2+x3)
rstudent<-rstudent(fit)
inf.m<-influence.measures(fit)
rstudent
inf.m$infmat
pre<-predict(fit,se=TRUE)
pre$se.fit
```

예제 6.3 토양침식자료

토양의 침식과 관련된 몇몇 변수들이 침식에 미치는 영향을 알아보기 위해 다음과 같은 실험을 하였다. 경사진 농경지에 위치한 일 평방피트짜리 11개의 면적을 실험대상으로 하여 20분 동안 2인치의 인공비를 뿌린 후에 침식된 토양을 조사하였다. 변수와 자료는 다음과 같다.

Y : SL (1 에이커당 파운드)

X_1 : SG (땅의 경사도)

X_2 : LOBS (토양의 침식이 가장 심한 곳의 깊이(단위:인치))

X_3 : PGC (흙으로 덮여진 부분의 비율(단위:퍼센트))

SL	SG	LOBS	PGC
271	0.43	1.95	0.34
35.6	0.47	5.13	0.32
31.4	0.44	3.98	0.29
37.8	0.48	6.25	0.30
40.2	0.48	7.12	0.25
39.8	0.49	6.50	0.26
55.5	0.53	10.67	0.10
43.6	0.50	7.08	0.16
52.1	0.55	9.88	0.19
43.8	0.51	8.72	0.18
35.7	0.48	4.96	0.28

7번과 10번 관측값의 *R*-스튜던트 값(*R*-student value)을 보면 각각 3.0976, -3.8509으로 그 절대값이 2.0을 초과 한다. 즉, *R*-스튜던트 통계량(*R*-student statistic)은 그 관측값이 자료세트(data set)로부터 제거되어야 함을 알려준다. 그러나, 바깥에 위치한 관측값들(outlying observations)을 제거하기 전에, 그것이 결과에 미치는 영향(influence)을 평가하는 것이 흥미로울 것이다. 11개의 모든 관측값의 영향(influence)을 결정하기 위한 필수 진단도구(essential diagnostics)는 다음과 같다.

Experiment	e_i	<i>R</i> -Student	h_{ii}	DFFITS	Cook's D
1	0.8163	0.7089	0.4643	0.6600	0.117
2	0.7872	0.5689	0.2476	0.3263	0.029
3	-0.0171	-0.0131	0.3625	-0.0099	0.000
4	-0.0103	-0.0075	0.2991	-0.0049	0.000
5	-0.1619	-0.1211	0.3316	-0.0853	0.002
6	-0.1295	-0.0842	0.1178	-0.0308	0.000
7	2.1514	3.0976	0.5326	3.3064	1.227
8	-0.3974	-0.3577	0.5301	-0.3800	0.041
9	0.5879	0.6068	0.6285	0.7893	0.171
10	-3.0495	-3.8509	0.1884	-1.8555	0.289
11	-0.5771	-0.4268	0.2976	-0.2778	0.022

Experiment	DFBETAS		DFBETAS		DFBETAS
	Intercept	b_1	b_2	b_3	
1	-0.0029	0.1092	-0.3195	-0.1966	

2	-0.0598	0.0024	0.0770	0.2303
3	-0.0072	0.0062	-0.0026	0.0021
4	-0.0001	0.0013	-0.0028	-0.0040
5	-0.0549	0.0662	-0.0717	-0.0354
6	0.0107	-0.0081	0.0013	-0.0113
7	1.5430	-1.2807	0.9770	-1.3176
8	0.0242	-0.1291	0.2503	0.3319
9	-0.6145	0.5344	-0.2148	0.2230
10	-0.6541	0.6481	-0.7098	0.1962
11	0.1808	-0.2075	0.2100	0.0423

자료 7에서 R -스튜던트 값(R -student value) 3.0976는 모자 대각값(HAT diagonal value) 0.5326을 수반한다. 자료 10에서 R -스튜던트 값(R -student value) -3.8509는 모자 대각값(HAT diagonal value) 0.1884를 수반한다.

자료 10은 작은 모자 대각값(HAT diagonal) ($h_{ii} = 0.1884$)을 갖는 것으로 보아 그 영향력이 경

미할 것으로 판단된다. 자료 7에서 모자 대각값은 0.5326로 $\frac{p}{n} \approx 0.36$ 의 평균 이상이지만,

$\frac{2p}{n} \approx 0.72$ 의 척도 보다는 크지 않다. 그러나 b_1, b_2, b_3 와 관련된 DFFITS와 DFBETAS는 다

른 관측값에 해당하는 진단도구보다도 더 크다. 분명히 DFBETAS에 대한 척도값 $\frac{2}{\sqrt{n}}$ 와

DFFITS에 대한 $2\sqrt{\frac{p}{n}}$ 과의 비교는 그 영향의 정도가 상당하다는 것을 시사할 것이다. 회귀

진단(regression statistics)을 사용함으로써 회귀를 실제로 다시 실행하지 않고도 이상치(outlier)의 영향(impact)을 평가할 수 있다.

다음은 예제 6.3에서 사용한 R code이다.

```
soil<-read.table('c:/soil.txt',header=T)
attach(soil)
fit<-lm(SL~SG+LOBS+PGC)
rstudent<-rstudent(fit)
inf.m<-influence.measures(fit)
summary(fit)
fit$residuals
inf.m$infmat
rstudent
```


6.4. 성능에 대한 영향 (Influence on Performance)

진단도구 $(DFFITS)_i$ 와 $(DFBETAS)_{j,i}$ 는 자료세트(data set)내에서 결과에 큰 영향을 미치는 관측값들을 강조한다. 이러한 진단값들(the values of these diagnostics)은 영향(influence)을 반영하지만, 그러한 영향(influence)이 회귀방정식(regression equation)의 일부분에 대해 더 나은 성능(performance)을 제공하는지에 대해서는 알려주지 않는다. 비정상적으로 큰 $(DFBETAS)_{j,i}$ 값은 단지 i 번째 관측값이 j 번째 회귀계수(regression coefficient)인 b_j 를 묘사하는데 있어 두드러지는지를 나타낼 뿐이다. 그 값은 i 번째 관측값의 존재 여부가 계수의 추정(the estimation of the coefficient)을 상당히 확실하게 하는지 아닌지에 대해서는 초점을 두지 않는다.

계수들(coefficients)의 분산-공분산(variance-covariance) 특성에 대한 편리한 스칼라 측도(scalar measure)를 제공하는 하나의 통계량(a single staticstic)으로 회귀계수(regression coefficients)의 일반화 분산(generalized variance: GV)이 있으며, 다음과 같다.

$$GV = |Var b| = |(X'X)^{-1} \sigma^2| \quad (6.10)$$

(6.10)에 있는 표현식은 다소 단순하고 흥미롭게 해석될 수 있다. Graybill (1976)을 참고하라. 분석가는 회귀계수 집합(set of regressions)의 질 추정(quality estimation)을 위해서 작은 값의 행렬식(determinant)을 얻으려고 애쓴다. σ^2 과 별개로 일반화 분산(generalized variance)은 X 행렬의 조건부 함수(a function of the conditioning of the X matrix)이다. 일반화 분산(generalized variance)을 얻을 때 i 번째 관측값의 역할을 알아내기 위하여, 우리는 i 가 있는 속성(property)에 대한 i 가 없는 속성(property)의 비(COVRATIO라 함)를 정의한다. 추정치 s_{-i}^2 과 s^2 은 각각 분자와 분모의 σ^2 대신 사용된다 그 결과는 다음과 같다.

$$(COVRATIO)_i = \frac{|(X'_{-i} X_{-i})^{-1} s_{-i}^2|}{|(X'X)^{-1} s^2|} \quad (6.11)$$

여기에서 X_{-i} 는 i 번째 관측값이 제거된 $(n-1) \times p$ 자료 행렬(data matrix)을 의미한다. (6.11)에 있는 COVRATIO에는 표준오차 형태의 조정(standard error type scaling)이 없다. 그렇지만, COVRATIO가 1.0을 초과하는 경우 i 번째 점(point)이 성능을 어느 정도 향상(improvement)시킨다는 점은 명백하다. 즉, i 번째 점을 제거한 경우에 비하여 i 번째 자료를 포함하고 얻은 회귀계수(coefficient)의 추정 일반화 분산(estimated generalized variance)이 감소한다.

COVRATIO가 1.0 미만인 값은 i 번째 자료를 포함하는 것이 추정 일반화 분산(estimated generalized variance)을 증가시킨다는 것을 나타낸다. (6.12)로부터 무엇이 $(COVRATIO)_i$ 가 1.0 을 초과하게 하는지 쉽게 알 수 있다.

$$(COVRATIO)_i = \frac{(s_{-i})^{2p}}{s^{2p}} \left(\frac{1}{1-h_{ii}} \right) \quad (6.12)$$

(6.12)의 전개가 궁금하면, Belsley, Kuh와 Welsch (1980)를 참고하시오. (6.12)에 있는 항 $(1-h_{ii})^{-1}$ 은 $|(X'X)^{-1}|$ 에 대한 $|(X'_{-i}X_{-i})^{-1}|$ 의 비(ratio)이다. 강한 지렛대(strong leverage, $h_{ii} \geq 1.0$)는 COVRATIO가 커지도록 한다. 이것은 예상된 결과인데, 왜냐하면, 극단(extreme)에 있는 점은 그것이 y 축 방향으로 동떨어진 관측값(outlying observation)이 아니라면 나머지 자료 집합에 의해 정해진 경향(trend)을 뚜렷하게 할 것이기 때문이다. i 번째 관측값이 실제 이상치(outlier)라면, $\frac{s_{-i}^{2p}}{s^{2p}}$ 는 1.0보다 훨씬 작아질 것이다. (식 (5.6)을 보시오.) 따라서, 우리가 논의한 모든 진단도구에서처럼, 지렛대 (거리 x_i 는 자료의 중앙으로부터 떨어진 거리임)와 점 x_i 에서의 적합 오차(the error in fit)가 함께 작용하여 진단 결과(diagnostic result)를 만들어낸다. COVRATIO_i의 경우에, x_i 에서의 높은 지렛대(hight leverage)와 작은 잔차(small residual)는 회귀계수(regression coefficients)의 분산 특성(dispersion property)을 강화한다.

이전의 진단도구(diagnostics)에서처럼, COVRATIO에 대한 척도(yardstick)를 정하기란 매우 어렵다. 이러한 문제점은, 대부분의 사용자들에게 있어서 일반화 분산(generalized variance)에 대한 영향(influence)의 개념(notion)이 적합 반응(fitted response)이나 추정 회귀계수(estimated regression coefficient)의 경우에서 보다 더 모호하여 더욱 복잡해진다. Belsley, Kuh와 Welsch (1980)는 규모가 큰 표본에서 사용하기에 적격인 개략적인 척도(rough yardstick)를 제시한다.

본질적으로, 그 척도(yardstick)는, 만약 $(COVRATIO)_i > 1 + \frac{3p}{n}$ 이거나

$(COVRATIO)_i < 1 - \frac{3p}{n}$ 라면 i 번째 자료가 일반화 분산(generalized variance)에 비정상적인 영향(influence)을 미치고 있음을 시사한다. 이 지침(guideline)의 하한(lower bound)은 오직 $n > 3p$ 일 때에만 적용됨을 기억해야 한다.

예제 6.4 토양침식자료

예제 6.3에서 살펴본 토양침식자료를 계속 살펴보도록 하자. Table 6.3은 중요한 진단 정보를 제공한다. 우리는 처음으로 COVRATIO를 포함하였다. 이 예는 교육상 매우 유익하다. 간추린 해설이 다음에 제시되어 있다.

가장 큰 지렛대(leverage)는 관측값 9 ($h_{ii} = 0.6285$)와 연관되어 있다. 모자 대각(HAT diagonal) 값은 허술한 척도(crude guideline)인 $\frac{2p}{n} \approx 0.72$ 보다 작지만, $\frac{p}{n} \approx 0.36$ 을 초과한다. 그러나 R -스튜던트(R -student)값은 작은 것은($t_9 = 0.6285$) 그 점에서 적합도가 좋음(good fit)을 의미한다. Cook's D (0.171) 와 DFFITS (0.7893)는 자료 9에 의한 영향이 본질적으로 없다는 것을 반영한다. 그러나, 분석가는 그 점을 불활성(inert)인 것으로 간주해서는 안되는데, 자료 9의 부재 시에 회귀를 강화시키기 때문이다. 이러한 결론은 (COVRATIO)₉에 반영되어 있는데, 그 것의 값은 3.9299이다. 따라서 관측값 9는 그것이 회귀의 속성을 강화한다는 점에서 매우 중요하다. 자료 8에 대해서도 유사한 결론을 도출할 수 있다. 자료 7,10은 영향을 주는 관측값들이다. 점 7에 대한 R -스튜던트(R -student) 값은 커서 적합성에 있어서 꽤 큰 오차가 있음을 시사한다. 이것은 높은 지렛대(high leverage) ($h_{ii} = 0.5326$)와 결합되어 그 점이 세 개의 계수들에게 강한 영향을 미치도록 한다. 자료 10은 -3.8509로 큰 R -스튜던트(R -student) 값을 가지지만 지렛대(leverage)가 작아 ($h_{ii} = 0.3068$) 적합성에 있어서는 꽤 큰 오차가 있으나, 그 영향력은 경미할 것으로 판단된다. 자료 7과 10 모두 추정 일반화 분산(estimated generalized variance)이 증가하도록 ($\text{COVRATION} < 1$) 한다.

Table 6.3 Influence diagnostics

Observation	e_i	t_i (R-Student)	h_{ii}	COVRATIO	Cook's D	DFFITS
1	0.8163	0.7089	0.4643	2.5070	0.117	0.6600
2	0.7872	0.5689	0.2476	1.9956	0.029	0.3263
3	-0.0171	-0.0131	0.3625	2.9058	0.000	-0.0099
4	-0.0103	-0.0075	0.2991	2.6429	0.000	-0.0049
5	-0.1619	-0.1211	0.3316	2.7446	0.002	-0.0853
6	-0.1295	-0.0842	0.1178	2.0901	0.000	-0.0308
7	2.1514	3.0976	0.5326	0.0868	1.227	3.3064
8	-0.3974	-0.3577	0.5301	3.6236	0.041	-0.3800
9	0.5879	0.6068	0.6285	3.9299	0.171	0.7893
10	-3.0495	-3.8509	0.1884	0.0157	0.289	-1.8555
11	-0.5771	-0.4268	0.2976	2.3401	0.022	-0.2778

Observation	DFBETAS	DFBETAS	DFBETAS	DFBETAS
	Intercept	b_1	b_2	b_3
1	-0.0029	0.1092	-0.3195	-0.1966
2	-0.0598	0.0024	0.0770	0.2303
3	-0.0072	0.0062	-0.0026	0.0021
4	-0.0001	0.0013	-0.0028	-0.0040
5	-0.0549	0.0662	-0.0717	-0.0354
6	0.0107	-0.0081	0.0013	-0.0113
7	1.5430	-1.2807	0.9770	-1.3176
8	0.0242	-0.1291	0.2503	0.3319
9	-0.6145	0.5344	-0.2148	0.2230
10	-0.6541	0.6481	-0.7098	0.1962
11	0.1808	-0.2075	0.2100	0.0423

다음은 예제 6.4에서 사용한 R code이다.

```

soil<-read.table('c:/soil.txt',header=T)
attach(soil)
fit<-lm(SL~SG+LOBS+PGC)
rstudent<-rstudent(fit)
inf.m<-influence.measures(fit)
summary(fit)
fit$residuals
inf.m$infmat
rstudent

```

6.5. 영향력이 큰 자료에서 우리는 무엇을 해야하는가?(What Do We Do with High Influence Points?)

분석가는 사용 가능한 진단도구(diagnostics)가 일련의 독립적인 영향 측도(a set of independent influence measures)를 대표하지 않는다는 점을 숙지해야 한다. 예를 들어, 만약 Cook's D_i 가 비정상적으로 높은 결과를 보이고, 적어도 하나의 $(DFBETAS)_{j,i}$ 가 특정 회귀계수 (regression coefficient)에 강한 영향(strong influence)을 미친다는 것을 보게 된다면, 확신을 가질 수 있을 것이다. 이와 똑같은 조건이 $(DFFITS)_i$ 와 $(COVRATIO)_i$ 에 대한 영향(impact)을 반영한다. 따라서, 진단도구(diagnostics)를 통해 얻어진 정보에는 중복되는 부분이 많다. 실제로, 경험이 많은 분석가는 R -스튜던트(R -student)와 모자 대각(HAT-diagonal)값만 보고도 영향(influence)을 주는 자료들을 빨리 알아낼 수 있을 것이다. 그러나, 분석가는 모든 사용 가능한 진단도구(available diagnostics)를 살펴보아야 한다. 예를 들어, 만약 결론이 계수들(coefficients)의 해석에 강하게 의존하여 도출되었다면, DFBETAS는 추가적인 중요한 정보를 제공할 수 있을 것이다.

진단도구(diagnostics)는 어떤 자료들의 정확성(accuracy)을 재검하기 위한 자원이 충분하다면, 영향력(influence)이 있는 자료들은 철저히 조사해야 함을 알리는 신호(signs)를 제공하기 위해 고안되었다. 이것은 하나의 관측값으로 인하여 예상치 못한 결과가 발생하였을 때 특히 중요하다. 영향력이 큰 관측값을 제거하여야 할 필요가 있었던 적이 있는가? 고영향 자료(high influence observation)에 대한 분석가의 태도는 이상치(outlier) (그 자체가 상당한 영향을 미치기도 하는)에 대한 태도보다 더 엄격하지는 않다. 만약 고영향(high influence) 자료를 재평가 하는 중에 중대한 문제를 발견하면, 그 자료의 존재 여부에 대해 생각해보아야 한다. 그러나, 재평가를 통해 영향력 있는 자료가 타당한 관측값(valid observation)임이 입증된다면, 그것을 제거할 합당한 이유가 없다. 어떤 경우에는 고영향(high influential) 관측값이 가장 중요한 정보의 집합일 수 있다. 그것이 가정된 모형(postulated model)을 일차적으로 지지하기도 한다. 반면에, 그것이 가정된 모형(postulated model)에 반대되는 증거일 수 있지만 전체적인 분석에 중대한 것일 수도 있다. 이제, 자료세트에서 가장 바람직한 지렛대 상태(leverage condition)는 균일한 분포의 지렛대(uniform distribution of leverage)를 얻는 것이다. 이것은 모든 모자 대각(HAT diagonal)이 $\frac{p}{n}$ 의 값을 취하고, 지렛대(leverage)로부터 나온 잠재적 영향력(potential influence)이 자료들 사이에 균등하게 나누어질 때 가능하다. 그러나, 지렛대의 불균등 배분(uneven allotment of leverage) (따라서 영향의 불균등 배분)은 많은 형태의 자료 집합에서 심심찮게 발생한다. 이러한 조건의 존재가 회귀를 구할 수 없다는(regression cannot be salvaged) 것을 자동적으로 의미하는 것은 아니다. 영향을 미치는 점들(influential points)이 모형 개발로부터 분석가를 영구히 편향시키는 이유가 되어서는 안된다. 요컨대, 진단도구(diagnostics)는 계산하는 큰 노력 없이, 중요한 결과를 나타낸다. 그러나, 진단에서 나온 정보는 종종 분석가로 하여금 더 많은 것을 조사하게끔 한다. 따라서, 효과적인 모형 개발

(building an effective model)이라는 궁극적인 목표에 도달하는 길이 어느 정도 변경될지도 모른다.

회귀영향 진단도구(regression influence diagnostics)의 사용자는 보통 회귀분석(ordinary regression analysis)의 항목(context)에 대한 정보를 가져야 한다. 전통적인 가설 검정(traditional hypothesis testing) 및 모형 적합(model fitting)으로 사용 가능한 모형을 개발하는 데에 실패하는 경우가 흔히 발생하므로, 진단 과정(diagnostic procedure)을 개발 과정에 추가하는 것은 환영할 만하다. 모형이 적합함에도 전통적인 의미에서 기각되었을 때, 표준적인 과정(변수선별(variable screening), 가설검정(hypothesis testing) 등)으로는 왜 모형이 실패하는지에 관하여 적절히 규명할 수 없다. 또한 적절한 모형(suitable model)이 무엇인지에 관하여 힌트를 줄 수도 없다. 마찬가지로, 전통적인 과정으로 어떤 모형을 폐기할 수 없다면, 그 모형이 완전한 검사를 거쳤다는 확신도 없다. 진단도구(diagnostics)는 이러한 공백을 채우기 위해 고안되었다. 전통적인 분석가들은 모형 성능(model performance)을 진단도구(diagnostics)를 통해 하는 것처럼 자세히 살펴보지 않는다.

5장 뿐만 아니라 7장에 있는 내용, 특히 변환(transformation)을 다룬 내용과도 매우 명확한 관련이 있다. 이상치(outlier)의 존재나 고영향(high influence) 자료의 존재는 변환(transformation)을 요하는 신호가 될 수 있다. 만약 진단도구 정보(diagnostic information)를 4장에서 제시된 모형개발(model building) 단계의 진전(progression)으로 연관짓지 못한다면, 그 것은 우리의 잘못(oversight)이다. 잔차(residuals), PRESS 잔차(PRESS residuals), 예측 표준오차(standard errors of prediction) 등에 초점을 두고서 모형후보(candidates)들을 보다 자세히 살펴볼 수 있을 것이다. 연산 명령(ordering of operations)에 있어서, 초기에 모든 자료 집합 (모든 회귀변수)에 대하여 영향 진단(influence diagnostics)을 관측하는 것에 관한 논쟁이 있다. 이러한 검사(inspection)가 결정에 영향을 주게 되면, 모형 개발 과정(model-building process)을 상당히 변화시킬 수도 있다. 예제를 통해, 하나의 관측에 의해 변수(variables)의 역할이 어떻게 바뀔 수 있는지를 이미 살펴보았다.

주의하여야 할 것이 있다. 많은 소프트웨어 패키지는 한 개의 관측값을 제거한 결과를 제공한다. 다중 관측 진단도구(multiple observation diagnostics)로의 손쉬운 접근을 위해 수학적 계산을 사용할 수 있다. 이 장의 내용은 경험이 있는 사용자에게도 다소 어려운 내용이었다. 컴퓨터 출력물의 분량으로 인해 낙담하거나, 기진맥진해지지 않기 위해서는, 이 연장선상에 있는 내용으로 천천히 접근해야 할 것이다.

7. 비표준조건들, 가정위배와 변환(Nonstandard Conditions, Violations of Assumptions, and Transformations)

3장과 4장에서는 다중선형회귀모형을 위한 기본적인 통계적 추론(statistical inference)에 대하여 다루었다. 또한 가능한 모든 모형 중 최적의 모형을 알아내는 기준과 방법에 대해서도 살펴보았다. 이 모든 내용의 기본은 회귀계수(coefficiten)의 추정법으로 최소제곱법(method of least squares)을 사용하는 데 있다. 최소제곱법을 이용하기 위해 필요한 가정(assumption)은 다음과 같다.

1. $E(\varepsilon_i) = 0$
2. ε_i 는 서로 독립이며 분산은 σ^2 으로 등분산이다.

추가로, ε_i 의 정규성(normality)은 불편추정량(unbiased estimator) 중 최소분산성질을 설명하는 데 필요하다. 5장에서는 이런 가정들이 위배(violation)되었는지를 탐지(detection)하는 방법으로 잔차(residuals)을 이용하여 그래프를 그리거나 분석하는 방법이 제시되었다.

6장은 5장의 연장으로 볼 수 있다. 자료에서 잔차의 성질은 영향력 진단을 위한 측도로 볼 수 있다. 이러한 의미에서 5장과 6장은 어떤 또는 모든 모형에서 반드시 고려되어야 하는 모형 또는 자료 진단에 대해서 다루었다. 분석의 측면에서 모형 진단의 부분은 모형의 오지정(misspecification), 이분산(heterogeneous variance), 오차의 비정규성(nonnormal error), 이상점(outlier)의 존재 등 가정에 위배되는 것을 찾기 위한 것이다.

이 장에서는 5장에서 개발한 진단방법을 통해 탐지된 몇몇의 비정상적인 조건하에서 모형설정(building)의 타협점을 찾아가는 것에 주목한다.

이 장에서 고려하는 비정상적인(nonstandard) 조건은 다음과 같다.

1. 이분산(heterogeneous)
2. 자료변환(data transformation)이 필요한 모형 오지정(model misspecification)
3. 오차의 상관성(correlated error)
4. 이상점(outlier)이 있는 자료와 오차의 비정규성(nonnormal)
5. 비정규성(nonnormal) 모형오차를 가지는 회귀자료
6. 측정오차(measurement error)가 있는 회귀변수들(regressor variables)

이 중 가장 강조되는 것은 자료변환(transformation)이다. 분석가는 반응변수(response variable)나 회귀변수(regressor variable)에 어떤 값(metric)을 사용하는 것이 가장 적절한가를 아는 것이 중요하며, 동시에 결정은 해당 문제의 과학자에 의해 결정을 내릴 수 있어야 하고 내려져야 한다. 그러나 적절한 변환을 나타낼 수 있게 고안된 진단정보를 필요로 하는

상황들이 있다. 독자는 5장에서 자료변환에 필요한 다양한 종류의 그림을 다루었다는 것을 상기할 필요가 있다. 그러나, 어떤 변환방법을 사용해야하는지를 제시해주진 않았다.

이 장에서는 또한, 정규성 가정의 위배에 대해 광범위하게 토론한다. 정규성의 부재에 강건한(robust)한 추정 방법(estimation method)에 대해 설명하고 예시된다. 이러한 강건한(robust)한 추정방법은 정규성의 가정에 벗어나거나 자료의 이상점(outlier)에 민감하지 않다. 추가로 반응변수의 특수한 유형과 관계된 정규성 이탈에 대해서도 다룬다. 생물학, 건강학 그리고 물리학 등이 많은 응용분야에서 반응변수는 이산형이다. 즉, 0 또는 1과 같은 이산적 자료거나 1, 2, 3,...와 같은 셀수 있는 형태이다. 다른 응용분야에서 모형의 오차가 연속적이나 정규분포 이외의 다른 분포 가정이 더 적절할 수 있다. 이 장에서 비정규분포의 오차를 일반화 선형모형의 틀에서 알아보기로 한다. 이항 회귀와 포아송(Poisson) 회귀 모형이 여기에 포함된다.

자기 분야에 적용에 대해서 관심이 있는 독자는 5장에 있는 내용들과 자연스럽게 연관되어야 한다. 여기에서 다루는 방법은 5장에 있는 방법들이 특정한 가정을 위배했을 때 기본적인 최소제곱합에 대한 대체방법으로 고려될 수 있다.

실제 적용(real applications)에서 가정의 위배는 예외보다는 통례이다. 경험이 있는 독자는 이러한 사실을 이미 알고 있다. 분석하는 입장에서는 가정 위배(violation)이 너무 애매하거나 정상적인 절차에서 중요하지 않을 때를 알아야 한다. 종종 분석가 중 가정 위배에 너무 과잉 반응하는 경우가 있다. 자료를 변환(transformation)하거나 자료에 가중치(weighting)를 부여하거나 명백한 이상점(outlier)을 제거하는 조작을 과잉으로 할 수 있다. 적절한 곳에서 가정의 위배를 조정하거나 해결하는 것은 분석의 절차에서 중요한 부분이다. 그러나, 무의식적으로 조정하거나 생각없이 해서는 안된다. 인내하고 약간의 심사숙고를 한다면 이러한 방법은 우리의 목적을 달성할 수 있게 해준다.

7.1. 이분산: 가중최소제곱(Heterogeneous Variance: Weighted Least Squares)

등분산 가정은 실제상황에서 종종 위배된다. 과학적인 측정 시 회귀변수(regressor variable) 또는 반응변수(response variable)의 값이 커질수록 적합 모형이나 경향에 대한 변동도 커지게 된다. 5장에서 잔차진단(residual diagnostics)에서 이러한 사실을 볼 수 있었다. 이론적인 측면에서 최소제곱(the least squares)법을 이분산 경우에 적용시키기 위한 변환은 단순하다. 하지만, 솔직히 말하면, 실제적으로 무엇을 해야하는지는 항상 명확하지 않다. 그러나, 최소제곱법이 등분산(homogeneous) 가정에 상당히 민감하기 때문에 이 문제는 아주 중요한 의미를 가진다(Seber (1977)). 이분산(heterogeneous variance) 처리를 위해 가장 먼저 고려되는 방법은 가중최소제곱 방법이다.

일반적인 선형모형은 다음과 같다.

$$y = X\beta + \varepsilon$$

위의 모형에서 최소제곱추정량은 다음과 같이 주어진다.

$$\hat{\beta} = (X'X)^{-1} X'y \quad (7.1)$$

추정량 (7.1)은 이상적인(ideal) 상태(모든 가정을 만족하는 상태)에서는 상당히 적절하다. 그러나 $\text{Var}(\varepsilon) = \sigma^2 I_n$ 이라는 가정을 완화하여 대신 양정치(positive definite) 행렬 V (Appendix A.2 참조)에 대해서 아래와 같이 가정해보자.

$$\text{Var}(\varepsilon) = V \quad (7.2)$$

식 (7.2)에서 최소제곱추정에 대한 좀 더 일반화 접근을 할 수 있다. 행렬 V 는 분산-공분산 행렬이며, 이것을 사용함으로써 오차의 등분산가정의 이탈 뿐만아니라, 오차간의 공분산도 고려할 수 있다. 우리는 다음을 고려할 수 있다.

$$V = \text{diag}[\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2] \quad (7.3)$$

이 경우는 관찰값간의 오차의 분산은 다르고 서로 독립(uncorrelated errors with error variances)을 가정한 것이다. 독자는 서로 다른 분산을 가지는 개별 관측치, y_i ,를 고려한 이분산의 영향으로 이해할 수 있다. 그 결과, 추정결과의 정확도 차이를 허용하는 하나의 방법으로 관측치에 그 차이만큼 가중하는 적절한 추정량 $\hat{\beta}$ 를 고려하는 것이 합리적이다. 여기에서 주어진 조건은 모든 관찰치를 이용한 모형에 관련된 정보에서 얻어진 것이나, 단편적인 정보에서는 동일한 정확성을 가지지는 못한다.

오차 ε_i 가 식 (7.2)에 주어진 분산-공분산 구조를 가질 때, 일반적인 선형모형에서 β 의 적절한 추정량(appropriate estimator)은 식 (7.1)의 최소제곱추정량(ordinary least squares estimators)이 아니라, 아래의 일반화 최소제곱추정량(generalized least squares estimator)이다.

$$\hat{\beta}^* = (X'V^{-1}X)^{-1}X'V^{-1}y \quad (7.4)$$

다음은 일반화 최소제곱추정량의 중요한 특징이다.

1. 추정량 $\hat{\beta}^*$ 는 불편추정량이다. 즉 $E(\hat{\beta}^*) = \beta$.
2. $\hat{\beta}^*$ 는 ε 의 정규성 조건, $\varepsilon \sim N(0, V)$, 하에서 최대우도 추정량(maximum likelihood estimator)이다.

3. 추정량 β^* 은 $\varepsilon \sim N(0, V)$ 하에서 최소분산을 가지는 불편추정량(unbiased estimator)이다.
4. 정규성의 가정을 완화한다면, 더 일반화된 가우스-마르코프 정리(Gauss-Markoff Theorem)을 적용한다; 즉 추정량 β^* 은 모든 선형 불편추정량 중 최소분산을 가진다.

특징 1-4에서, 식 (7.4)에 있는 추정량은 조건이 이상적일 때 최소제곱추정량과 동일한 성질을 가진다. 즉, $V = \sigma^2 I$ 일 때, β^* 은 식 (7.1)에 있는 \mathbf{b} 가 된다.

β^* 에 의해 최소화되는 것(What is Being Minimized by β^*)

일반화 최소제곱 명칭은 β^* 가 어떤 제곱합을 최소화하나는 것이다. 최소화된 함수는 V 에 의존하며 다음과 같다. (연습문제 7.2 참조)

$$SS_{\text{Res}, V} = (y - X\beta^*)'V^{-1}(y - X\beta^*) \quad (7.5)$$

최대우도 방법에 관심이 있는 독자들은 식 (7.5)의 결과가 $\varepsilon \sim N(0, V)$ 이라는 조건 하에서 우도를 최대화한 것과 동일하다는 사실을 알 수 있을 것이다

가중최소제곱 (Weighted Least Squares)

앞에서 언급한 것처럼 모형의 오차가 서로 독립(uncorrelated)이지만 등분산(homogenous) 가정을 만족하지 않는다고 하자. 다시 말하면 n 개 자료에서 오차분산은 $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ 이다. V 행렬은 다음과 같다.

$$V = \text{diag}[\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2] \quad (7.6)$$

이분산(heterogeneous variance) 상황을 도표로 설명을 한 것이 그림 7.1이다. 예를 들어, 다음과 같은 회귀모형을 가장하자.

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

여기서 ε_{ij} 는 서로 독립이고 평균이 0이지만 이분산(unequal variances)이다. 즉 다음과 같다.

$$\text{Var}(\varepsilon_{ij}) = \sigma_i^2$$

Simple linear regression with heterogeneous variance

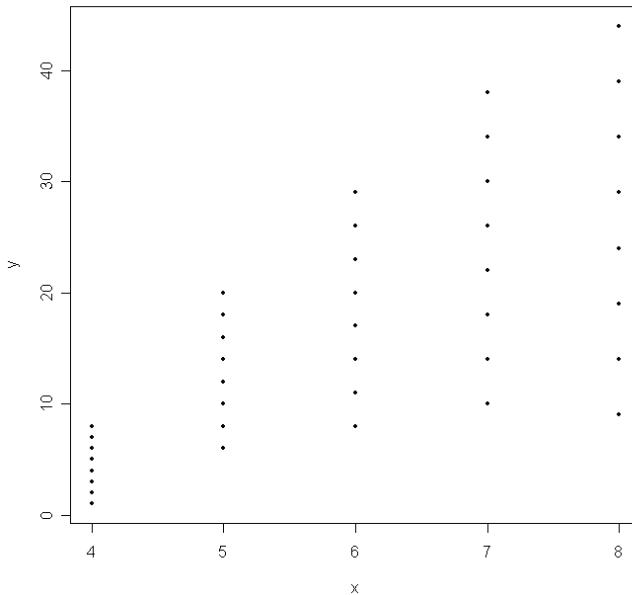


그림 7.1에서 x 가 증가함에 따라 변동(variability)도 증가한다. x 가 증가하면, 분산도 증가하여 기울기(slope)와 절편(intercept)를 추정하는데 필요 지식은 감소하게 된다. 반대로 x 가 작을 수치 값일 때, 오차 분산도 작다. 즉, 실험의 재현성이 높고, 정보량도 많아진다. 이런 경우에서 β 의 일반화 최소제곱 추정량이 아래의 식을 최소화시키는 것을 밝히는 것이 용이하다(식 7.5에서)

$$SS_{\text{Res}(\text{weighted})} = \sum_{i=1}^n w_i (y_i - \hat{y}_i)^2 \quad (7.7)$$

여기에서 $w_i = 1/\sigma_i^2$ 이다. 이것은 최소제곱추정량(least square estimator)할 수 있다. 다만, 각각의 잔차(residual)에 오차의 표준편차의 역수가 가중된 것이다. 적절한 가중 최소제곱 추정량은 식 (7.4)에 나타나있고, V는 식 (7.6)에 의해 주어진다.

식 (7.7)의 최소화 기준(the criterion of minimization)은 논리적으로 최소제곱법을 확장한 것이다. 이것의 의미는 큰 오차분산을 가진 자료 위치에서 잔차는 그 값 그대로 고려하는 것이 아니다. 모형이 더 뽀족하거나 정확한 자료의 위치에서 잔차제곱합(residual sum of squares)의 기여가 가장 커야 한다.

대부분의 상용화된 회귀분석 패키지에는 가중회귀분석을 포함하고 있다. 사용자는 종종 가중치(weight)을 넣지 않아도 된다. 그러나 여기에서 언급한 대로 가중치(weight)가 오차분산에 따라 변하지 않는다면 추정량은 정확하게 일반화 최소제곱 추정량은 아니다. 사실상 이 장의 후반부에 언급되는 로버스트(robust) 회귀방법은 분석도구로 가중회귀를 개발하였다. 그러나 정규성, 독립성 그리고 이분산이 존재하는 경우 가중치 $w_i = 1/\sigma_i^2$ 를

사용하는 것이 가중회귀에서는 가장 최적이다.

실질적인 어려움(Pratical Difficulties)

만약 분석가가 이분산(heterogeneous variace)이 문제가 된다면 가중 회귀가 고려되어야 한다. 그러나 많은 경우에 실질적인 어려움이 존재한다. 어디에서 σ_i 를 구하냐는 것이다.

그림 7.1에서 표본분산(sample varinace)이 개별 오차분산을 추정하기 위해서 사용될 수 있다. 따라서, 가중치가 추정될 수 있다. 그러나 회귀변수(regresssor variable)가 실험계획법에서 처럼 구해지지 않는 상황에서 측정치 y 는 각각의 회귀변수의 조합에서 하나의 관측치만 구해진다.; 이런 경우 σ_i 추정량은 유용하지 않다. 그 결과 최적의 추정된 가중치로부터 구한 가중최소제곱량은 그렇게 실용적이지 않다. 가중치의 추정은 본질적으로 각각의 회귀변수의 조합에서 반복적으로 실행되는 상황에 제한될 수 밖에 없다. 이러한 반복적인 실행에서는 오차 분산의 단순 추정량을 계산할 수 있다. 그러나 분석가는 제한된 정보를 기초로 한다면 σ_i 대신에 추정량을 사용하는 것에 주의해야 한다. 가중회귀에서 잘못 추정된 가중치(poorly estimated weight)를 적용하는 것은 가중치를 적용하지 않는 경우보다 더 불만족스러운 결과를 가져온다. 추정된 일반화최소제곱 (추정 가중치, estimating weights) 사용을 위한 지침서에서는 표본(sample)의 크기가 대략 9라면 추정된 가중치를 사용할 수 없다(Deaton et al.(1983) 참조). 이것은 경험(rule of thumb)에 의한 것이 아니라 안내서이다. 많은 실험계획들은 각각의 실험점에서 9번 이상의 중복된 실험은 고려하지 않기 때문에, 가중치가 무시되는 경우가 많다. 가중치의 추정과 사용에 대한 더 많은 정보를 원한다면 Williams (1967)과 Deaton 등(1983)을 참고하기 바란다.

일반화 제곱추정량의 분산-공분산 성질(Variance-Covariance Properties of the Generalized Least Squares Estimator)

가중 최소제곱추정량은 일반화 최소제곱 추정량의 한 종류이다. 전자는 이분산에서 최적불편추정량(optimal unbiased estimator) β 를 구하기 위한 것이다. 그러나 이전 절에서 언급하였듯이 최적성(optimality)은 σ_i , 즉 가중치 w_i 를 알 때 가지게 된다. 만약 추정값으로 w_i 로 바꾸면 가중치(weight)는 확률변수(random variable)가 되고 추정량 β 의 성질은 매우 복잡해진다. 그러나 선형회귀를 배우는 학생에게는 V를 알고 있는 경우 일반화 최소제곱 추정량(generalized least squares estimator)의 성질을 아는 것은 중요하다.(또는 가중회귀 특별 사례의 성질을 이해하는 것이 중요하다).

앞에서 언급한 대로 식 (7.4)의 추정량 β^* 는 불편추정량이다. 분산-공분산 행렬은 아래와 같이 쉽게 계산된다.

$$\begin{aligned}\text{Var}(\beta^*) &= \text{Var}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \mathbf{X}'\mathbf{V}^{-1}[\text{Var } \mathbf{y}] \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\end{aligned}$$

여기에서 $\text{Var } \mathbf{y} = \mathbf{V}$ 이므로

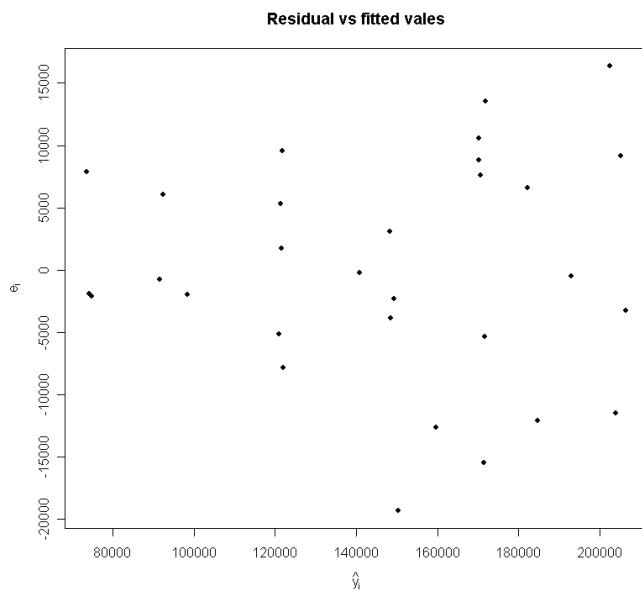
$$\text{Var}(\boldsymbol{\beta}^*) = \text{Var}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \quad (7.8)$$

만약 $\mathbf{V} = \sigma^2 \mathbf{I}$ 라면 식 (7.8)은 $(\mathbf{X}'\mathbf{X})^{-1} \sigma^2$ 이 되고 이는 최소제곱추정량(ordinary least square)의 분산-공분산 행렬이다.

예제 7.1은 등분간가정이 위배되었을 때 자료 분석을 보여준다. 여기에서 가중회귀는 자료로부터 추정된 가중치를 사용한다.

예제 7.1 음식 판매 수익과 광고 비용의 연관성(Restaurant Food Sales Data)

30개 음식점에 대한 월평균 음식 판매 수익과 연간 광고 비용에 대한 자료이다. 이 때 주 관심은 두 변수의 연관성이다. 또한, 선형회귀 모형에서 일반화 최소제곱량으로 구한 추정식 $\hat{y} = 49,440 + 8.048x$ 의 적합여부이다. 최소제곱에서의 잔차를 추정량에 대해서 도식해보면 다음과 같다.



여기서 일반화 최소제곱 추정에서 가정한 등분산에 위배됨을 볼 수 있다. 즉, 이 자료에서 일반화 최소제곱량은 적합하지 않다. 이런 이분산성 문제를 해결하기 위해서는 가중치에 대해서 알아야 한다.

자료에 대해서 다시 살펴보면, 광고 비용 x 에 대해서 유사한 값으로 묶으면 이것을 반복 자료로 볼 수 있다. 그리고 이렇게 그룹화한 x 에 대한 y 의 추정 분산(s_y^2)을 구할 수 있다. 표를 보면 그룹한 x 에 대한 평균 \bar{x} 에 대한 y 의 분산은 비례한다. 이를 도식화하면 \bar{x} 에 대한 s_y^2 는 대략 선형의 증가형태를 보인다. 이를 최소 제곱 적합하면 다음과 같다.

$$s_y^2 = -9,226,002 + 7,782\bar{x}$$

Obs. i	Income, y_i	Advertising Expense, x_i	\bar{x}	\bar{y}	Weights, w_i
1	81,464	3,000			6.21771E-08
2	72,661	3,150	3,078.333	26,794,616	5.79507 E-08
3	72,344	3,085			5.97094 E-08
4	90,743	5,225			2.98667 E-08
5	98,588	5,350	5,287.5	30,722,013	2.90195 E-08
6	96,507	6,090			2.48471 E-08
7	126,574	8,925			1.60217 E-08
8	114,133	9,015			1.58431 E-08
9	115,814	8,885	8,955.0	52,803,695	1.61024 E-08
10	123,181	8,950			1.59717 E-08
11	131,434	9,000			1.58726 E-08
12	140,564	11,345			1.22942 E-08
13	151,352	12,275			1.12852 E-08
14	146,926	12,400	12,171	59,646,475	1.11621 E-08
15	130,963	12,525			1.10416 E-08
16	144,630	12,310			1.12505 E-08
17	147,041	13,700	13,700.0		1.00246 E-08
18	179,021	15,000			9.09750 E-08
19	166,200	15,175			8.98563 E-08
20	180,732	14,995			9.10074 E-08
21	178,187	15,050			9.06525 E-08
22	185,304	15,200	15,095.0	120,571,061	8.96988 E-08
23	155,931	15,150			9.00144 E-08
24	172,579	16,800			8.06478 E-08
25	188,851	16,500	16,650.0	132,388,992	8.22031 E-08
26	192,482	17,830			7.57282 E-08
27	203,112	19,500			6.89136 E-08
28	192,482	19,200			7.00460 E-08
29	218,715	19,000	19,262.5	138,856,871	7.08218 E-08
30	214,317	19,350			6.94752 E-08

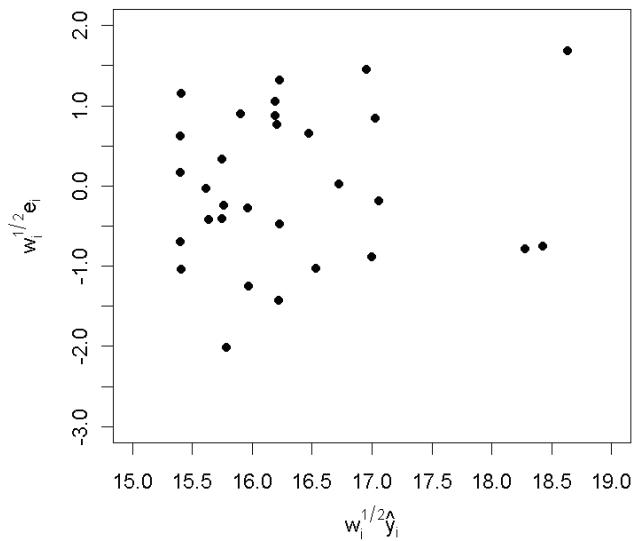
이 식에 각 x_i 를 각 y_i 의 분산 추정치에 대응시킨다면 적합값의 역순을 가중치 w_i 로 보는

것이 합당하다. 이 각 관측치에 대한 가중치는 표에서 볼 수 있다.

이러한 가중치를 이용해서 가중 최소 제곱 추정을 해보면 $\hat{y} = 50,974.567 + 7.922x$ 추정식을 구할 수 있다.

여기서 가중 최소 제곱추정에 의한 적합으로 어떻게 달라졌는지 알아보기 위해 잔차도를 그려보면 다음과 같다.

Weighted residuals vs weighted fitted values



이는 일반화 최소 제곱 추정에 의한 방법보다 향상되었음을 알 수 있다. 즉, 잔차의 등분산이 어느정도 조정되었다. 이를 볼 때 가중 최소 제곱 추정은 이분산성 문제를 어느 정도 해결할 수 있다.

다음은 위의 예제에 사용된 R code이다.

```
ta5.9 <- read.table('c:/chapter 7/table5.9.txt',header=T)
attach(ta5.9)
fit <- lm(income~a.expense)
summary(fit)
plot(fit$fitt,fit$resi,xlab=expression(hat(y)[i]),ylab=expression(e[i]),
     main='Residual vs fitted vales', type="p", pch=19)

temp <- cbind(mean(a.expense[1:3]),var(income[1:3]))
temp <- rbind(temp,cbind(mean(a.expense[4:5]),var(income[4:5])))
temp <- rbind(temp,cbind(mean(a.expense[7:11]),var(income[7:11])))
temp <- rbind(temp,cbind(mean(a.expense[12:16]),var(income[12:16])))
temp <- rbind(temp,cbind(mean(a.expense[18:23]),var(income[18:23])))
```

```

temp <- rbind(temp,cbind(mean(a.expense[24:25]),var(income[24:25])))
temp <- rbind(temp,cbind(mean(a.expense[27:30]),var(income[27:30])))
temp
fit1 <- lm(temp[,2]~temp[,1])
summary(fit1)

w <- c(1/as.matrix(data.frame(1,a.expense))%*%c(fit1$coef))
fit2 <- lm(income~a.expense,weight=w, data=ta5.9)
summary(fit2)

plot(fit$fitt*sqrt(w),fit$resi*sqrt(w),main='Weighted residuals vs weighted fitted values',
     xlab=expression(w[i]^{1/2}*hat(y)[i]),ylab=expression(w[i]^{1/2}*e[i]), typ="p", pch=19,
     lab=c(10, 10, 20), xlim=c(15, 19), ylim=c(-3, 2))

```

가중회귀에 사용되는 더 많은 기법들(Further Techniques in Weighted Regression)

7장 1절에서 등분산(homogeneous variance) 가정을 할 수 없는 경우, 식 (7.4)에 의해 구한 가중회귀추정량(weighted regression estimator)을 사용하는 것이 타당하다는 사실을 알 수 있었다. 예제 7.1에서는 가중회귀식을 구하기 위한 계산을 설명하였다. 이제부터 일반적인 최소제곱추정법에 사용된 기본적인 가설검정과 계산과정을 무엇인지 그리고 이것을 가중회귀에 적용하기 위해서는 무엇을 조정해야하는지를 알아보도록 하자. 표면적인 어려움은 모형 오차와 관계된 가정된 분산-공분산구조와 관련되어 있다. 3장과 4장에서 언급된 기본적인 가설검정들, C_p , PRESS의 계산 등은 식 (7.1)의 추정량이나 $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ 의 형태를 가지는 분산-공분산에 따라 결정된다. 그 결과로 가중 회귀 사례에 대한 확장은 자연스럽게 V행렬에 관계되고 실제로 V와 잘 추정된 가중치에 의해 의존한다. 여기서 우리는 가중 회귀의 과정에 대하여 좀 더 알아보기로 하자. 물론 가중치는 $w_i = 1/\sigma_i^2$ 를 안다고 가정한다.

표준오차와 가설검정(Standard Errors and Tests)

우리가 V를 안다면, 확률변수는 아니다, 가중회귀계수(weighted regression coefficient)의 표준 오차(standard error)는 꽤 쉽게 계산할 수 있다. 회귀계수의 분산은 다음과 같다.

$$c_{jj} = j\text{th diagonal element of } (X'V^{-1}X)^{-1}$$

따라서, 표준오차는 다음과 같이 계산된다.

$$\sigma_{\beta_j^*} = \sqrt{c_{jj}}$$

모형계수(model coefficient)의 유의성 검정(test of significance)은 이 통계량을 이용하여 쉽게 만들 수 있다.

$$t = \frac{\beta_j^*}{\sqrt{c_{jj}}}$$

여기서 β_j^* 는 가중회귀에서 얻은 x_j 의 계수이다. 해당 가설은 다음과 같다.

$$\begin{aligned} H_0 : \beta_j^* &= 0 \\ H_1 : \beta_j^* &\neq 0 \end{aligned}$$

물론 여기에서 ‘ t ’ 표기를 사용하였으나, 만약 V 를 안다면 표준오차는 추정량이 아니다. 그 결과 적절한 가설검정은 표준정규분포(standard normal distribution)의 기각역(critical region)을 가지는 양측검정(two-tailed test)이다.

비가중최소제곱(unweighted least squares)의 경우에서처럼 제곱합은 변동분해를 설명하는데 유용하다. 또한 비가중 사례와 같이 설명된 변동(variation explained)이나 회귀식에 의한 제곱합을 가능하게 해준다. 행렬 표기에서 제곱의 합의 항등식은 다음과 같다.

$$y'V^{-1}y = \beta^{*'}X'V^{-1}y + (y - X\beta^*)'V^{-1}(y - X\beta^*)$$

또는,

$$\text{Total weighted SS} = \text{Weighted regression SS} + \text{Weighted residual SS}$$

식 (7.7)에 있는 가중제곱합(weighted sum of squares)처럼 쉽게 가중잔차제곱합(weighted residual sum of squares)을 알 수 있다.

$$(y - X\beta^*)' V^{-1} (y - X\beta^*)$$

가중제곱합(weighted residual sum of squares)은 모형의 가정이 약간 변한다고 하더라도 분산 추정에 중요한 역할을 한다. 오차분산에 대해 가정하기 보다는 대각행렬(diagonal matrix) V 를 있다고 가정하자.

$$\text{Var}(\boldsymbol{\epsilon}) = V\sigma^2$$

다른 말로 표현하면 상수 σ^2 와 상관없이 분산을 있다고 하자. 회귀계수(regression coefficient)에 대한 추정과정은 이 절에서 이미 언급한 가중최소제곱(weighted least squares)과 동일하다. σ^2 를 모른다는 것은 가중치가 변하지 않는다는 것이다. 즉, 가중치는 V^{-1} 의 대각원소으로 남아있다는 것이다. 이 경우 가중잔차제곱합(weighted residual sum of squares)이 σ^2 의 추정량으로 사용된다. 추정량은 다음과 같다.

$$s^2 = \sum_{i=1}^n \frac{w_i(y_i - \hat{y}_i)^2}{(n-p)}$$

여기에서 p 는 모형 변수의 수이다.

예제 7.2 음식 판매 수익과 광고 비용의 연관성(Restaurant Food Sales Data)

예제 7.1의 음식 판매 수익과 광고 비용의 연관성 자료를 보자. 예제에서 가중치를 사용하여서 회귀 계수의 분산-공분산 행렬을 구하면 다음과 같다.

$$\begin{aligned} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} &= \begin{pmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i x_{1i} \\ \text{symm.} & \sum_{i=1}^n w_i x_{1i}^2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 5.164446E(-07) & 0.004343582 \\ 0.004343582 & 50.62447531 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 6793976.2470 & -580.2394 \\ -580.2394 & 0.0693 \end{pmatrix} \end{aligned}$$

$\mathbf{X}' \mathbf{V}^{-1} \mathbf{y}$ 의 값은 아래와 같이 주어진다.

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \begin{bmatrix} 0.060578 \\ 621.451975 \end{bmatrix}$$

그 결과 식(7.4)로부터 구한 추정량과 식(7.8)로부터 구한 표준오차는 다음과 같다.

$$\beta_0^* = 50974.567 \quad s_{\beta_0^*} = 2606.526$$

$$\beta_1^* = 7.922 \quad s_{\beta_1^*} = 0.263$$

그리고 회귀계수에 대한 유의성 검정을 통해 0과 다른지를 결정할 수 있다. 먼저, β_0^* 에 대해 알아보자.

$$t = \frac{50974.567}{2606.526} = 19.557$$

그리고 β_1^* 에 대하여

$$t = \frac{7.922}{0.263} = 30.092$$

회귀계수의 가설검정 결과로부터, 모든 회귀계수는 0으로부터 상당히 다르다는 결론을 유추해 낼 수 있다. 광고 비용은 음식 판매수익에 영향을 미친다.

다음은 위 예제에 대한 R-code이다

```
x <- cbind(intercept=1, a.expense)
ivarc <- t(x) %*% (w*x)
varc <- solve(ivarc)
obs <- t(x) %*% (w*income)
coef <- varc %*% obs
c <- sqrt(diag(varc))
t <- coef / c
```

분산 안정화 변환(Transformations to Stabilize Variance)

회귀분석 적용시 많은 분야에서 등분산 가정이 만족되지 않는 것은 자연스러우며 충분히 예상가능하다. 문제의 원인은 오차분산이 평균 $E(y)$ 와 독립이지 않다는 것이다. 앞에서 언급하였듯이 회귀변수(regressor)나 반응(response) 값이 커질수록 회귀선 주위의 분산도 점점 더 커진다.

분석가가 $E(y)$ 와 오차분산의 정확한 함수관계를 알 수는 없다. 그러나 잔차나 반복된 자료의 분산을 통해 구조식에 대해 알 수 있는 경우도 있다. 오차분산이 평균 반응에 따라 어떻게 변화하는지를 안다면 오차분산을 안정화시키는 반응변수의 변환을 제시할 수 있다. 분석가는 회귀식을 구하기 위해 변환된 y (transformed y)를 이용하기도 한다.

분산 안정화 변환(variance-stabilizing transformation)으로 로그변환(the natural log transformation, $\ln y$)이 종종 사용된다. 이것은 오차 표준편차(standard deviation)가 $E(y)$ 에 비례할 때 적합한 변환이다. 다른 것으로는 역 변환으로 즉 $1/y$ 이 있다. 이 변환은 오차의 표준편차가 평균 반응의 2차방정식(quadratic function)으로 나타나는 경우에 적합하다. 즉 σ 는 $[E(y)]^2$ 에 비례하는 경우이다.

자료변환의 목적은 등분산(homogeneous variance) 가정을 더 적합하게 만드는데 있다. 어떤 변환은 매우 효과적일 수 있다. 그러나 하나의 가정이 위배된 경우 이를 바로 잡기 위한 자료변환이 다른 가정을 위배될 수 있게 한다는 것을 인식해야 한다. 오차분산을 안정화시키기 위한 변환은 모형의 함수형태에 관한 가정을 변화시킨다. 적합과 예측의 질을 악화시키는 오차 분산의 안정화 변환은 피해야 한다. 만약 자료를 변환한 경우, 예를 들어 자연 log로 변환한 경우, $\ln y$ 가 아닌 y 의 적합과 예측의 변환효과가 무엇인지 주의깊게 고려해야 한다.

7.2. 상관오차와 관련된 문제들, 자기상관(Problem with Correlated Errors, Autocorrelation)

7장 1절에서 모형 오차의 분산-공분산상태가 동일하지 않을 때 회귀계수의 추정을 가능하게 하는 일반적인 방법에 대하여 소개하였다. 이분산(heterogenous variance)문제가 있는 경우 가중최소제곱(weighted least squares)의 사용을 강조하였다. 한편 식 (7.4)에 주어진 추정된 회귀계수를 가진 일반화 최소제곱(generalized least squares)의 방법은 만약 분산-공분산의 구조를 알고 행렬 V 에 대한 정확한 추정량을 알고 있을 때 사용할 수 있는 일반적인 방법이다. 그러나 V 는 알 수 없는 경우가 많다.

시간에 걸쳐 결과에 자연스러운 효과가 있을 때 특정한 형태의 회귀분석방법이 적용된다. 관계된 자료는 종종 시계열 자료(time series data)라 불린다. 회귀분석의 자료가 시간에 의존적일 때 오차가 서로 독립이라는 가정해서는 안 된다. 다소, 오차는 계열상관(serial correlation)이 있다고 가정할 수 있다. 다른 말로 $E(\epsilon_i, \epsilon_{i+j}) \neq 0$ 이며, 이러한 오차들을 자기상관(autocorrelated)이라고 말한다.

시간의 효과에 기인한 자기상관의 개념은 상당히 이해하기 쉽다. 과학자가 상품의 가격과 수요에 관련된 모형을 구축하는데 관심이 있다고 하자. 시간에 따라 상품의 가격과 수요의 평균값과 같은 자료는 얻을 수 있고 대신 시간에 의해 영향을 받거나 상품의 수요와 관련된 변수는 모형에 포함하지 않는다고 가정하다. 이런 경우 계절 영향이나 소비자의 인구 크기와 같은 변수들은 시간변수와 같은 것을 나타낸다. 그 결과 자기상관(autocorrelation)은 모형 오지정(misspecification)의 문제로 종종 볼 수 있다. 단순선형회귀모형을 가정하자.

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

모형의 오차는 다음과 같이 쓸 수 있다.

$$z_j = \epsilon_i + \sum_j \beta_j w_{ji}$$

여기에서 ϵ_i 는 더 일반적인(conventional) 모형오차(model error)이고 $\sum_j \beta_j w_{ji}$ 는 모형에 포함되지 않은 시간에 관계되는 회귀변수를 나타낸다. 명백하게 z_i 는 잠재(hidden)변수들에 의존하며 다른 시간간에 상관되어 있다.

자기상관의 영향과 탐지(Impact and Detection of Autocorrelation)

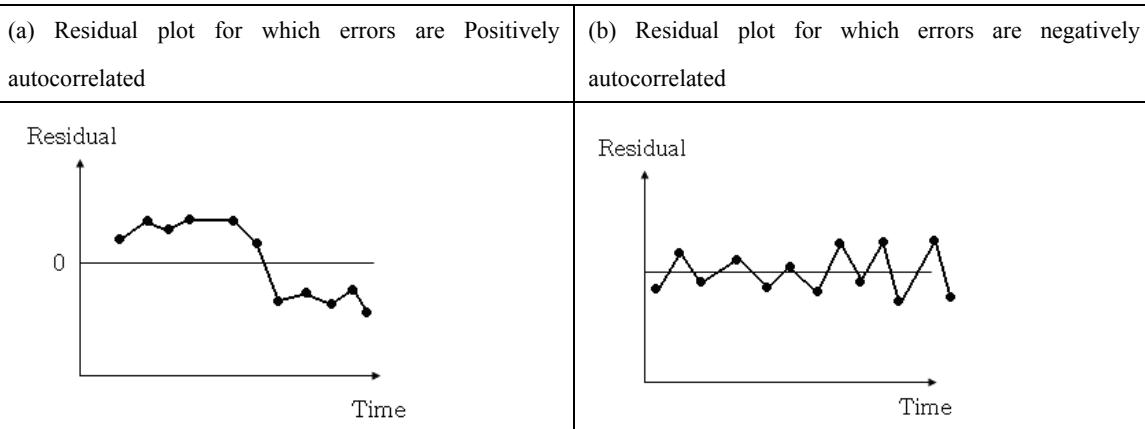
물론, 자기상관(autocorrelated)오차가 존재한다는 것은 오차들의 분산-공분산 행렬이 $\sigma^2 I$ 가 아니며 기본가정이 이상적이지 않다는 것을 의미한다. 자기상관은 오차분산의 추정을

어렵게 하므로 가설검정과 신뢰구간 추정도 어렵게 한다. 양(positive)의 상관된 오차가 존재하는 경우 σ^2 의 추정량은 과소추정(underestimate)되므로 회귀계수에 대한 t 통계량을 크게 하고 회귀계수의 신뢰구간은 짧게 추정하는 문제를 발생시킨다.

오차의 자기상관을 탐지(detection)하는 방법은 진단도표(diagnostic plot)나 정규검정(formal test)을 통해서 가능하다. 잔차의 단순도표도 도움될 수 있다. 만약 최소제곱분석(standard least squares analysis)에서 얻은 잔차(residuals)를 시간에 따라 도표를 그리면 동일한 부호를 가진 잔차들이 많아 나타나 함께 뭉칠 수 있는 경향이 있다면 오차는 양(positive)의 자기상관이 있다는 것을 나타낸다. 반대부호의 잔차로 빠르게 변한다면 음(negative)의 자기상관을 나타낸다. 그림 7.2는 이를 설명한다.

자기상관을 탐지하기 위한 정규검정(formal test)이 있다. 아마도 가장 잘 알려진 방법은 더빈-왓슨(Durbin-Watson, D-W) 검정이다. 이것은 많은 상용화된 회귀분석패키지에서 볼 수 있다. 더빈-왓슨검정은 양성의 자기상관을 알아내기 위해 고안되었으며, 널리 사용됨에도 불구하고 가정된 상관구조의 종류에 제한되어 있어서 동등한 시간 구간에 대해서 우리는 다음과 같이 가정할 수 있다.

Figure 7.2



$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad (7.9)$$

여기에서 ε_t 는 시간 t 일때 오차이고, ρ 는 자기상관(autocorrelation)이며 u_t 는 $N(0, \sigma_u^2)$ 을 가정한 확률잡음(random disturbance)이다. 물론 모수(parameter) ρ 는 1.0을 넘을 수 없다. 식 (7.9)의 오차구조는 1차 자기회귀 오차구조 (first-order autoregressive error)를 나타낸다. 모형 오차가 (7.9)를 통해서 서로 관련되어 있으므로 주어진 오차인 ε_t 의 성질에 좀 더 초점을 맞추는 것이 효과적이다. 만약 연속적으로 (7.9)를 사용하면 다음과 같다.

$$\varepsilon_t = \sum_{j=0}^{\infty} \rho^j u_{t-j} \quad (7.10)$$

따라서, 모형오차(model error)는 (7.9)의 독립정규분포(independent normally distributed) 오차의 선형조합이다. ε_t 에 대한 큰 영향을 미치는 u 값은 시간상 가장 최근에 일어난 잡음(disturbance)이다. 다음과 같이 쓸 수 있다.

$$E(\varepsilon_t) = 0$$

$$\text{Var}(\varepsilon_t) = \sigma_u^2 \sum_{j=0}^{\infty} \rho^{2j}$$

식 $\sum \rho^{2j}$ 는 $1/(1-\rho^2)$ 로 쓸 수 있다. 따라서,

$$\text{Var } \varepsilon_t = \sigma_u^2 \left(\frac{1}{1-\rho^2} \right) \quad (7.11)$$

어렵지 않게 비슷한 형태로 보일 수 있다.

$$\text{Cov}(\varepsilon_t + \varepsilon_{t+j}) = \rho^{|j|} \sigma_u^2 \left(\frac{1}{1-\rho^2} \right) \quad (7.12)$$

식 (7.9)에 구해진 자기상관 구조는 $\rho = 0$ 이 아니면 모형 오차 사이에 자기상관을 얻게 된다. 추가로 예측한 대로 자기상관은 시간이 가까울수록 더 심해진다.

더빈-왓슨검정의 가설은 다음과 같다.

$$H_0 : \rho = 0$$

$$H_1 : \rho > 0$$

위의 통계량은 인접한(adjacent) 잔차사이의 차(difference)에 어느 정도 기초했다는 것을 알 수 있다. 만약 오차가 양(positive)의 상관을 나타낸다면 인접한 잔차는 숫자적으로 비슷할 것이라고 예측할 수 있다. 더빈-왓슨 통계량은 다음과 같다.

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_i^2} \quad (7.13)$$

여기에서 물론 e_t 는 최소제곱 분석으로부터 얻은 보통 잔차(ordinary residual)이다. 명백하게, $\rho > 0$ 이라면 d 의 분자(numerator)는 작다고 예측할 수 있다. 따라서, d 가 작은 수치이면 H_0 는 기각된다. 더빈(Durbin)과 왓슨(Watson) (1951)에 의해 구해진 범위(bounds)는 다양한 n 과 α 에 대해 표C. 6에 표기되어 있다. 이러한 범위(bounds)는 d_L 과 d_U 로 표현된다. 더빈-왓슨 검정은 다음과 같다.

만약 $d < d_L$ 이면 귀무가설 $H_0: \rho = 0$ 은 기각

만약 $d > d_U$ 이면 귀무가설 $H_0: \rho = 0$ 은 기각되지 않는다.

만약 $d_L < d < d_U$ 이면 검정결과를 보류한다.

예제 7.3 탄산음료 집중판매 자료(Soft Drink Concentrate Sales)

탄산음료 회사는 지역별 연간 광고료에 따른 집중 판매량에 대해서 예측하고 싶어한다. 다음의 예제는 20년간의 연간 탄산음료 판매량을 나타낸다. 이 때 선형관계가 있다면 일반화 최소제곱 추정을 할 수 있고, 추정식은 $\hat{y} = 1608.51 + 20.09x$ 이다. 아래의 표는 일반화 최소 제곱 추정에 의해 잔차를 구한 것이다.

Year	Annual Regional Concentrate Sales, y	Annual Advertising Expenditures(\$\times 1000), x
1960	3083	75
1961	3149	78
1962	3218	80
1963	3239	82
1964	3295	84
1965	3374	88
1966	3475	93
1967	3569	97
1968	3597	99
1969	3725	104
1970	3794	109
1971	3959	115
1972	4043	120

1973	4194	127
1974	4318	135
1975	4493	144
1976	4683	153
1977	4850	161
1978	5005	170
1979	5236	182

이와 같은 시간에 따른 자료는 자기 상관이 나타나는 경우가 많다. 이는 시간에 대한 잔차도를 그려서 확인할 수 있다. 또한, 아래의 더빈-왓슨(Durbin-Watson) 검정을 통해서도 가능하다. 위의 자료에 대한 더빈-왓슨(Durbin-Watson) 검정량을 구해보면,

$$d = \frac{\sum_{i=2}^{17} (e_i - e_{i-1})^2}{\sum_{i=1}^{17} e_i^2} = 1.08$$

이다. 이때 신뢰수준을 0.05에서 부록 C.7를 보면 자료수가 20일 때 계수 d의 하한은 1.20, 상한은 1.41임을 찾을 수 있다. 구해진 검정량 1.08이 범위에 포함되지 않으므로 자기 상관이 발생했다고 할 수 있다.

위의 예제에 대한 R code는 다음과 같다.

```
x <- c(75, 78, 80, 82, 84, 88, 93, 97, 99, 104, 109, 115, 120, 127, 135, 144, 153, 161, 170, 182)
y <- c(3083, 3149, 3218, 3239, 3295, 3374, 3475, 3569, 3597, 3725, 3794, 3959, 4043, 4194, 4318, 4493,
4683, 4850, 5005, 5236)
z <- c(825000, 830445, 838750, 842940, 846315, 852240, 860760, 865925, 871640, 877745, 886520, 894500,
900400, 904005, 908525, 912160, 917630, 922220, 925910, 929610)
fit <- lm(y~x+z)
fit$resi
dwtest(y~x+z)
```

시간 의존적인(time-dependent) 자료의 자기상관오차(autocorrelated error) 를 아는 것은 종종 단순히 잔차 그림들(plots)를 관찰함으로써 알 수 있다. 예를 들어, 그림 5.3, 5.4, 5.5의 잔차그림(plots)에서 이차항이 추가된다면 잔차는 전형적인(orthodox) 형태가 될 것이라고 추측(suspicion)할 수 있다. 인구조사에서 시간, 즉 연도(year)는 주요한 회귀변수(regressor variable)이다. 따라서, 어떤 형태의 자기상관을 생각하는 것은 자연스러운 것이다.

일반적으로 자료가 시간 의존적이라면, 잔차를 조사해야 하고 반응의 변동성(variability)을 설명할 수 있는 숨겨진 회귀변수(regressor variable)를 찾아 자기상관을 제거해야 한다. 시간이 회귀변수가 아닐 때에도 자기상관이 나타날 수 있다는 것은 강조되어야 하는 것이다. 어떠한 시간 의존적인 회귀자료의 집합이라도 자기상관오차를 보인다. 더빈-왓슨(Durbin-Watson) 통계량은 이것을 알아내는데 도움이 된다.

만약 자기상관 오차를 알아 낸다면 자기상관을 제거하는데 사용될 수 있는 몇가지 방법이 있다. 물론 초기에 해당 분야 과학자들은 추가적인 회귀변수나 변수를 찾아야 했다. 이 책에서 자기상관을 다루는 방법들에 대하여 면밀하게 공부하지는 않는다. 이에 대한 자세한 토론과 개발을 위한다면 Box와 Jenkins(1976)가 쓴 책을 참고하기 바란다.

자기상관을 다루는 방법들(Methods for Handling Autocorrelation)

자기상관을 제거하는 과정은 작용변수의 시차변수(lagged values)를 사용하여 모형을 재구성하는 것이다. 시간 의존적인 자료에서 자기상관이 있다면 반응변수 y_t 는 인접시간의 수치인 y_{t-1} 에 의존한다는 것을 경험적으로 알고 있다. 따라서, 고려된 적합모형의 형태는 다음과 같다.

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 y_{t-1} + \varepsilon_t$$

4장에서 언급되었던 잔차나 성능기준값(performance criteria)에 대한 재검토는 이러한 모형의 가치를 평가하는데 사용될 수 있다. 더빈-왓슨(Durbin-Watson) 통계량은 시차변수를 포함한 모형에 사용되어선 안 된다.

자기상관을 포함한 모형에서 모수(parameter)를 추정하는 적합한 방법은 Cochrane과 Orcutt (1949)에 의해서 개발되었다. 이 방법은 최소제곱(standard least squares) 가정(assumption)이 가능하도록 반응변수(response variable)과 회귀변수들(regressor variables)를 변환하는 것에 기초를 하고 있다. 식 (7.9)의 자기회귀 구조를 만족하는 오차를 가진 단순선형회귀방법을 고려해 보자.

$$y_t = \beta_0 + \beta_1 x_1 + \varepsilon_t \quad (t = 1, 2, \dots, n)$$

변환된 반응변수를 고려해보자.

$y_t^* = y_t - \rho y_{t-1}$

(7.14)

식(7.9)와 식(7.14)를 이용하면 아래의 식을 얻을 수 있다.

$$\begin{aligned}
 y_t^* &= \beta_0 + \beta_1 x_t + \varepsilon_t - \rho(\beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}) \\
 &= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + (\varepsilon_t - \varepsilon_{t-1}), \quad (t=1,2,\dots,n) \\
 &= \beta_0^* + \beta_1^* x_t^* + u_t
 \end{aligned} \tag{7.15}$$

모형 (7.15)의 오차는 식 (7.9)의 자기회귀 구조의 오차이고 평균이 0이고 분산이 σ_u^2 인 정규분포이고 독립이다. 그 결과, 모수 ρ 를 안다면 최소제곱(standard least square) 방법을 사용할 수 있다. 물론 ρ 를 안다는 가정은 실용적이진 않다. ρ 의 점 추정값은 β_0 , β_1 의 추정량을 구하는데 사용될 수 있다. 식 (7.9)의 1차 자기회귀 구조에서 ρ 의 단순 추정량을 쉽게 얻게 된다. 원점을 지나는 단순선형회귀로 식 (7.9)를 보면 다음과 같은 추정량을 얻을 수 있다.

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2} \tag{7.16}$$

여기에서 e_t 는 x_t 에 대한 y_t 의 최소제곱회귀(ordinary least square regression)의 잔차(residual)이다. 그러므로 β_0 , β_1 에 대한 반복(iteration) 추정과정은 다음과 같다.

- (i) 최소제곱(ordinary least square)법으로 적합한 후 $\hat{\rho}$ 를 얻는다.
- (ii) y_t^* 와 x_t^* 의 형태로 y_t^* 에 대한 x_t^* 의 최소제곱회귀(ordinary least squares regression)를 한 후 β_0^* 와 β_1^* 를 찾아낸다.
- (iii) 절편(intercept)과 기울기(slope)의 추정량은 $\beta_0^*/(1-\rho)$ 와 β_1^* 이다.

추정과정의 추가적인 반복이 필요한 경우 식 (7.5)에서 모형의 잔차에 대한 더빈-왓슨(Durbin-Watson) 통계량을 기준값으로 사용할 수 있다. 두 번째 과정에서 더빈-왓슨 통계량으로 판단할 때 여전히 양의 자기상관이 명백히 존재한다면 계속적인 작업이 필요하다. 또한, 두 번째 과정은 모형 (7.15)의 잔차를 사용하여 식 (7.16)에서 $\hat{\rho}$ 을 계산하는데 관계된다. 이 과정에서 새로운 y_t^* 와 x_t^* 를 구하고 따라서 새로운 β_0^* 와 β_1^* 를 구할 수 있다.

7.3. 적정과 예측을 증진시키는 변환(Transformations To Improve Prediction)

자료의 변환은 때때로 더 좋은 적합결과와 더 좋은 예측모형을 얻기 위한 효과적인 대안이 된다. 자료의 변환이나 재표현(reexpression)은 더 효과적인(reasonable) 가정(assumption)을 만들기 위해 사용된다. 따라서, 이 장에서는 회귀자료에 대한 자기상관을 제거하고 분산을 안정화시키기 위해 자료의 변환을 사용한다.

다음 절에서 대안모형형식(alternative model form) 또는, 변환(transformation)을 결정하는 과정에 대하여 토론을 하고 이러한 변환이 좋은 이유를 증명하기로 한다. 몇몇의 경우에는 특정한 종류의 변환이 필요한 이유가 명백하다. 단순회귀경우 자료 도표에서 곡선형(curvilinear)의 모양을 보일 때이다.

단순회귀 사례에서의 변환(Transformation in the Case of a Single Regressor)

회귀변수가 하나인 경우를 생각해보자. 자료가 곡선으로 나타날 때 모형구조(model structure)를 변환할 필요가 있다. 물론 아래와 같은 직선회귀모형(ordinary straight line regression model)에서의 이탈(deviation)은 잔차도표나 자료의 단순 도표로부터 얻을 수 있다.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

만약 도표가 특별한 곡선형의 경향을 보이면 이 경향에 맞게 모형을 변환시킬 수 있다.

포물선(Parabola)

우리가 고려해야 할 첫번째 대안 모형은 단순한 변환이라기 보단 이차항(quadratic)을 모형에 추가하는 것이다. 모형은 다음과 같이 주어진다.

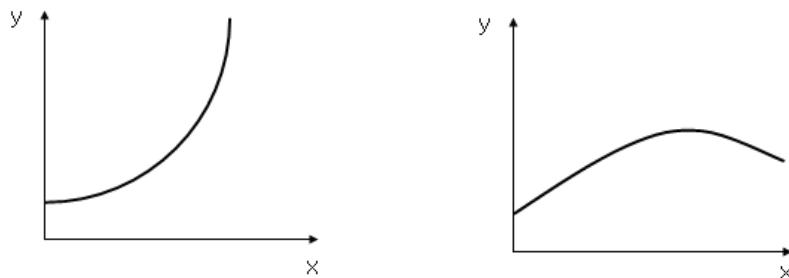
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

포물선(parabolic) 모형에 의해 생성된 자료로 만든 도표의 성격은 계수 β_0 , β_1 , β_2 의 부호가 크기(magnitudes)에 따라 결정된다. 두 개의 전형적인 예가 그림 7.3에 있다.

FIGURE 7.3

(a) Parabola for β_0, β_1 , and $\beta_2 > 0$

(b) Parabola for $\beta_0 > 0, \beta_1 > 0, \beta_2 < 0$



쌍곡선(Hyperbola, Inverse Transformation on y and x)

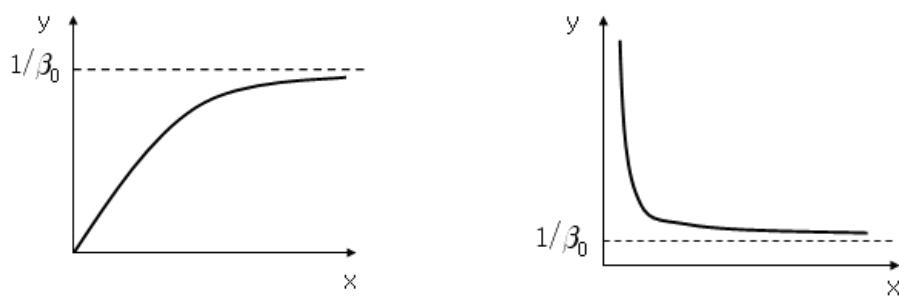
생물학, 경제학, 어떤 다른 분야에서의 적용은 쌍곡선 함수를 사용해야 하며 이것은 반응변수 y 와 회귀변수 x 둘 다 변환하여 구할 수 있다. 쌍곡선의 사용을 나타내는 전형적인 도표가 그림 7.4에 있다. 쌍곡선의 함수형태는 모형 계수에 비선형(nonlinear)이다. 이 식은 $y = x/(\alpha + \beta x)$ 로 주어진다. 이 선형화된 식은 두 변수의 역 변환, $1/x$ 에 대한 $1/y$ 의 회귀에 관계되어 있고 다음과 같은 모형구조를 가진다. (관찰치는 $x_i, y_i, i = 1, 2, \dots, n$).

$$\frac{1}{y_i} = \beta_0 + \beta_1 \left(\frac{1}{x_i} \right) + \varepsilon_i$$

FIGURE 7.4

(a) Hyperbola with negative curvature

(b) Hyperbola with positive curvature



여기에서 $\beta_0 = \beta$ 와 $\beta_1 = a$ 라는 것은 명백하다. 분석가에게 관심을 끄는 점근선(asymptote)은 그림 7.4에 점선으로 나타난다. 음(negative)과 양(positive)의 곡선은 각각 $\beta_1 > 0$ 와 $\beta_1 < 0$ 일 때 나타난다.

지수함수; y 에 대한 자연 로그 변환(Exponential Function; Natural Log Transformation on y)

7장 1절에서 어떤 상황에서 이분산(heterogeneous variance)문제에 직면했을 때 유용한 변환으로 반응변수(response variable)에 대한 로그변환(log transformation)에 대해 알아보았다. 이 변환은 또한 자료의 모양이 어떤 특정한 타입의 곡선을 나타낼 때, 타당한(reasonable) 모형 가정을 만드는데 유용할 수 있다. 만약 모양이 직선이 아니고 다소 그림 7.5의 각각 부분을 묘사한 것처럼 구조는 $y = \alpha e^{\beta x}$ 의 형태일 수 있다. 따라서 적합한 모형의 형태는 다음과 같다.

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

여기에서 물론 $\beta_0 = \ln \alpha$ 이고 $\beta_1 = \beta$ 이다. 그림 7.5(a)는 $\beta > 0$ 인 상황을 설명하고 그림 7.5(b) $\beta < 0$ 인 사례를 보여준다.

멱함수, y 와 x 에 대한 자연 로그 변환(Power Functions, Natural Log Transformations on y and x)

때때로, x 에 대한 y 의 도표에서 $y = ax^\beta$ 의 형태의 곡선을 보인다. 이 경우 두 변수를 로그변환(log transformation)함으로써 선형함수(linear function)로 바꿀 수 있다. 따라서, 적합된 모형은 다음과 같다.

$$\ln(y_i) = \beta_0 + \beta_1 (\ln x_i) + \varepsilon_i$$

여기서 회귀계수 β_0 , β_1 은 최소제곱과정(ordinary least square procedure)에 의해서 추정된다. 실제 멱함수의 도표 모양은 상수 β 의 부호와 크기에 의해 결정된다.

FIGURE 7.5

(a) Exponential function: use in y transform (b) Exponential function; use in y transform

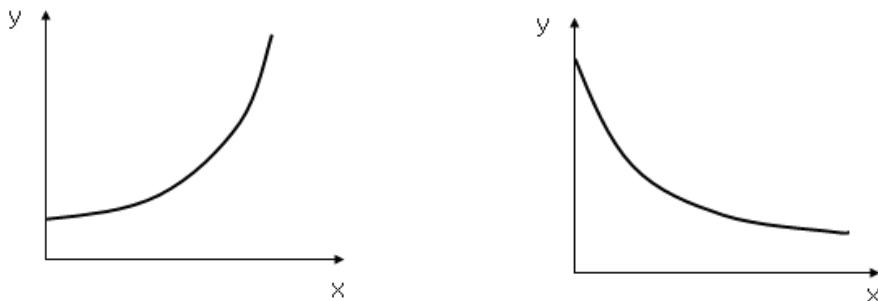
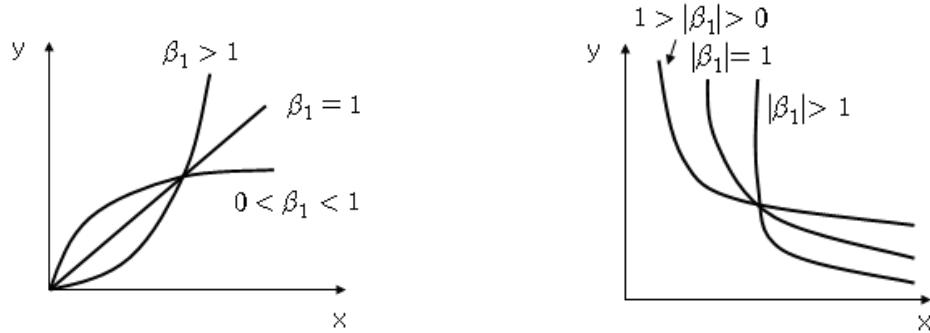


FIGURE 7.6

- (a) Power function:
use natural log transformation on y and x
- (b) Power function: negative: β_1
use natural log transformation on y and x



역지수, y에 대한 자연 로그 변환: x에 대한 역변환(Inverse Exponential, Natural Log Transformation on y: Inverse Transformation on x)

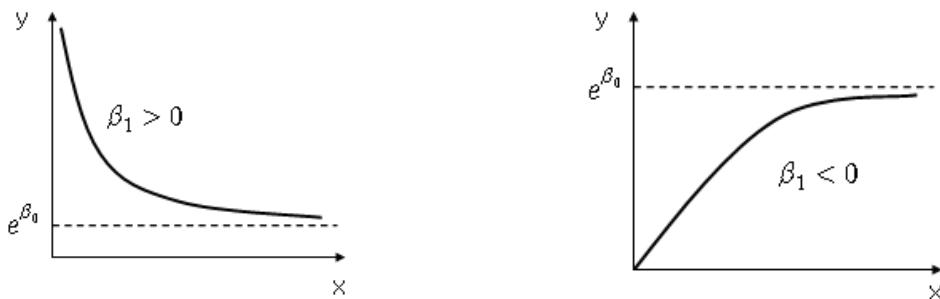
자연에는 지수(exponential)의 형태인 많은 과학 현상이 있으나 반대로 초기에 언급된 것에 맞아 떨어지지 않는 것도 있다. 종종 x대신에 x의 역에 지수적으로 비례한다. 이는 그림 7.5에 제시된 것과 유사하지 않은 형태를 보인다. 함수의 형태는 $y = \alpha e^{\beta/x}$ 이다. 변환 후 적합된 모형은 다음과 같다.

$$\ln(y_i) = \beta_0 + \beta_1 \left(\frac{1}{x_i} \right) + \varepsilon_i$$

그림 7.7은 x에 대한 y의 모양을 나타낸다.

FIGURE 7.7

- (a) Inverse exponential:
use natural log transformation
on y and inverse transformation on x
- (b) Inverse exponential:
use natural log transformation
on y and inverse transformation on x



이 절의 목적은 직선형의 단순 선형회귀보다 적합이나 예측이 좋은 다른 모형을 제공하는데 있다. 단순회귀 변수의 경우에서 만약 도표의 곡선이 모양에서 여기에 묘사된 것과 유사하다면 좋은 결과를 얻기 위해서 어떤 변환을 적용할지를 결정해야한다. 모형이 향상되었다는 것은 (i) 잔차의 형태가 더 좋아야 하며 (ii) 더 잘 적합되어야 하고 더 좋은 예측 통계량들을 제공해야 한다. (ii)의 경우 분석가는 원래의 y 의 단위에서 잔차를 구해야 하고 따라서, 오차평균제곱(error mean square)을 구할 수 있다. (잔차 자유도로 나눈 잔차제곱합). 추가로 PRESS 잔차는 원래의 변환되지 않은 단위에서 y 변수를 구할 수 있으므로 PRESS 통계량은 계산할 수 있다. 사용자는 변환된 모형의 타당성을 평가할 수 있는 특정한 기준값을 가지고 있어야 한다. 예제 7.4를 통해 설명하도록 하겠다.

변환시 모형구조에 생기는 변화(What Happens to Model Structure Under Transformation)

변환(transformation)은 통계자료분석의 중요한 부분이다. 여기에는 사용자가 변환의 목적을 잘 이해할 수 있도록 정형화(formalize)해보도록 하겠다. 그러나 자료분석가는 자료를 변환(transformation)하겠다면 먼저 모형의 전체적인 구조(total structure)가 무엇인지를 알아야 한다. 변환은 자료가 새로운 모형에 더 가까울 것이라는 생각으로 모형 기술(model statement)을 펜이나 컴퓨터 프로그램상에서 바꾼다는 것으로 생각할 수 있다. 하지만, 좀 더 복잡하며 분석가의 주의를 요하는 것이 있다. 예제를 통해서 가장 잘 설명할 수 있을 것이다. x 에 대한 y 의 도표는 그림 7.6(a)과 유사한 곡선형태를 나타낸다고 가정하자. $y = \alpha x^\beta$ 형태의 멱함수(power function)를 이용하여 자료를 생성하여 아래 형태의 모형에 적합하였다.

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i \quad (7.17)$$

사실, 적합결과를 나타내는 모든 지수(indication)를 향상되었다는 것에 초점을 맞출 수 있다. 그러나, 어디가 변환되었는지? 분석가가 한 것은 무엇인지? 실제 멱함수(true power function)은 $y = \alpha x^\beta$ 에서 변환이 이루어졌고, 만약 모형에서 오차를 무시한다면(물론 이렇게 할 수는 없다.), 멱함수에 자연 로그(natural log)를 취함으로 $\ln y = \beta_0 + \beta_1 \ln x$ 을 유도할 수 있다. 그러나 만약 가법 오차항(additive error terms)을 두 관계식에 포함시켰다면, 멱함수(power function)을 로그변환(log transformation)하더라도 식 (7.17)이 되진 않는다. 식 (7.17)의 모형이 유효하기 위해서 어떤 가정이 필요한지를 명확히 알아야 한다.(등분산가정하에서)

아래의 모형을 이용하여 식 (7.17)의 모형을 얻을 수 있다.

$$y = \alpha x^\beta (1 + \varepsilon^*) = \alpha x^\beta + \varepsilon^{**} \quad (7.18)$$

여기에서 ε^* 은 $E(\varepsilon^*) = 0$ 이고 등분산이라는 일반적인 특성들(usual properties)을 가지는 것으로 가정한다. 이러한 가정하에서 $\text{Var } \varepsilon^{**} = \sigma^2 [E(y)]^2$ 이며 이것은 $E(y)$ 에 따라 변한다. 이런 경우 승법오차구조(multiplicative error structure)가 있다고 말한다. 이러한 조건하에서 식 (7.18)

양변에 자연로그(natural log) 변환을 하면 다음과 같다.

$$\ln y = \ln \alpha + \beta \ln x + \ln(1 + \varepsilon^*) \quad (7.19)$$

식 (7.19)의 모형을 면밀히 살펴보면, 선형회귀(linear regression) 형태이다. 그러나 결과적으로 오차항(error term)은 무엇에 관한 것인가? $E[\ln(1 + \varepsilon^*)] = \alpha^*$ 라고 가장하면 다음과 같이 쓸 수 있다.

$$\ln(y) = (\ln \alpha + \alpha^*) + \beta \ln x + \varepsilon \quad (7.20)$$

여기에서 $E(\varepsilon) = 0$ 이고 $\text{Var}(\varepsilon) = \sigma^2$ 이므로 등분산(homogenous variance)이므로 $E(y)$ 에 따라 변하지 않는다. 식 (7.20)의 모형이 (7.17)의 형태이다. 결과적으로 승법오차(multiplicative error)을 가정한 식 (7.18) 모형을 변환하여 (7.17)의 얻을 수 있다.

(7.18)의 모형이 비현실적이지만 가법오차구조(additive error structure)라고 가정하자.

$$y = \alpha x^\beta + \varepsilon^* \quad (7.21)$$

여기서 ε^* 는 앞에서 처럼 기본적인 가정(standard assumption)을 만족한다. 자연로그(natural log)를 취하면 기본오차가정(standard error assumption)을 가지는 식 (7.17)처럼 되지 않는다. (7.21)을 다음과 같이 썼다고 가정하자.

$$y = \alpha x^\beta \left\{ 1 + \frac{\varepsilon^*}{E(y)} \right\}$$

여기에서 물론 $E(y) = \alpha x^\beta$ 이다. 로그변환(log transformation)을 적용하기 위해서 다음과 같이 쓸 수 있다.

$$y = \alpha x^\beta (1 + \theta)$$

여기서 $\theta = \varepsilon^*/E(y)$ 이고 로그변환(log transformation)하면 다음과 같다.

$$\ln y = \ln \alpha + \beta \ln x + \ln(1 + \theta) \quad (7.22)$$

이 경우, $\ln(1 + \theta)$ 항은 식 (7.17)의 모형 오차의 역할을 한다. 그러나 $1 + \theta$ 의 분산은 $E(y)$ 에 따라 변하므로 표준 최소제곱가정(standart least squares assumption)을 만족하지 않는다. 이러한

전개의 결과 자료의 변환이 오차구조를 변한시키다. 변환(transformation)이 구조적인 장점을 가질 수는 있지만, 등분산(homogeneous variance)나 정규성가정(normality assumption)을 위배를 가져올 수 있다. 그러므로 가정 위배(assumption violation)를 제거하느냐와 이차적인 위배를 유발하느냐의 결정해야하는 상황이 생길 수 있다. 예를 들어 정규성의 경우 식 (7.21)에 있는 ε^* 가 정규분포를 따른다면 식 (7.22)에 있는 $\ln(1 + \theta)$ 는 정규분포일 수 없다.

여기에서 소개된 변환(transformation)의 단점은 무분별한 사용에 있다. 그러나 변환을 사용할 때 스튜던트화 잔차(studentized residual)을 주의깊게 조사해야한다. 특히 (7.21)에 주어진 모형구조와 같이 계수에 대해 비선형(nonlinear)이고 모형오차(model error)가 가법적인(additive) 경우 비선형 회귀모형(nonlinear regression model)에 속한다는 것을 알아야 한다. 최소제곱에 의한 비선형 모형의 모수추정은 더 상세히 9장에 소개된다. 곡선을 포함한 자료를 다루는 경우 비선형 회귀에 대하여 익혀야 한다. 특정한 경우 자료를 변환시키지 않는 것보다 차라리 가법모형(additive model)을 가정하고 비선형회귀를 하는 것이 더 낫다.

예제 7.4 세균 소멸(Bacteria Deaths due to X-ray radiation)

주어진 자료는 시간에 변화를 주어 해저 박테리아 배양 접시를 X선에 노출시켰을 때, 살아있는 박테리아 수를 측정한 것이다. 학설에 의하면, 각 박테리아에는 단일 생명 중심부가 있고, 이는 박테리아가 활동력이 저하되거나 죽기 전에 광선을 노출되어야 한다고 한다. 특정 박테리아는 이러한 형태를 가지지 않아 살아남을 수 있다고 말하고 있다. 이 학설이 옳다면, 로그 변환을 통한 생존 박테리아 수는 노출 시간의 길이에 대해 직선의 형태로 나타날 것이다.

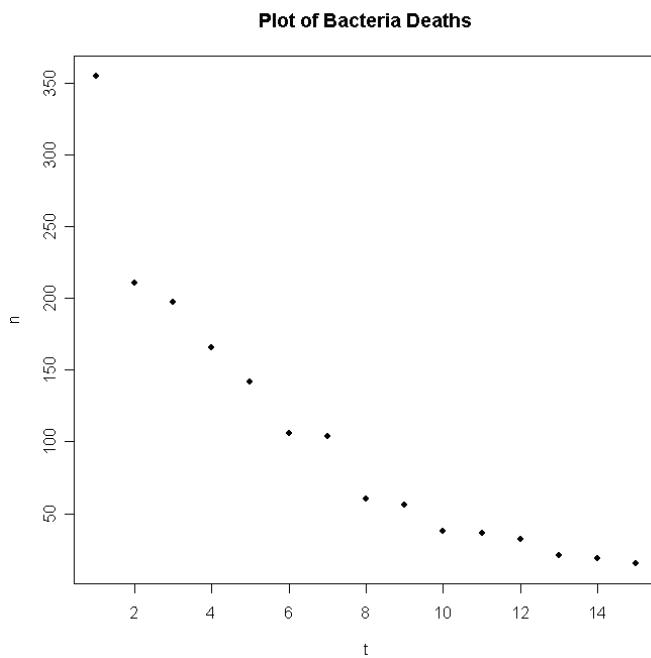
	Y(N)	Time
1	355	1
2	211	2
3	197	3
4	166	4
5	142	5
6	106	6
7	104	7
8	60	8
9	56	9
10	38	10
11	36	11
12	32	12
13	21	13
14	19	14

해당 자료를 산점도로 나타내면 다음과 같다.

단순선형모형(Simple Linear Model)

우리가 예측하듯이 단순 선형 회귀 모형으로 만족스러운 결과를 얻지 못한다. 그러나 이 식은 변환 모형과 비교하여 기초나 출발점으로 사용될 수 있다. 최소제곱단순회귀모형은 다음과 같다.

$$\hat{y} = 259.58 - 19.46t$$



다음의 성능(performance) 기준으로

$$R^2 = 0.8234$$

$$s = 41.83$$

$$\text{PRESS} = 35910.85$$

$$\sum_{i=1}^n |y_i - \hat{y}_{i_n-i}| = 513.1847$$

잔차와 PRESS 잔차는 아래의 표와 같다.

곡선이 있을 경우에 단순 선형 모형의 부적절성은 잔차의 패턴을 반영한다. 회귀변수는 τ 가 낮은 수치일 때 반응을 과대평가하고 범위의 중간 부분에서는 과소평가되며 위 부분에서는 과대평가된다. 다음 절에서는 자료가 곡선일 경우를 고려한 변환이 설명되고 비교된다.

y	\hat{y}	$e_i = y_i - \hat{y}_i$	$y_i - \hat{y}_{i,-i} = e_{i,-i}$
355	240.116667	114.883333	151.494505
211	220.652381	-9.652381	-11.994083
197	201.188095	-4.188095	-4.961918
166	181.723810	-15.723810	-17.945652
142	162.259524	-20.259524	-22.480845
106	142.795238	-36.795238	-40.036269
104	123.330952	-19.330952	-20.791293
60	103.866667	-43.866667	-47.000000
56	84.402381	-28.402381	-30.548015
38	64.938095	-26.938095	-29.310881
36	45.473810	-9.473810	-10.512550
32	26.009524	5.990476	6.836957
21	6.545238	14.454762	17.125529
19	-12.919048	31.919048	39.662722
15	-32.383333	47.383333	62.483516

역지수함수(Inverse Exponential Function)

역지수함수(inverse exponential function)를 이용하여 생존 박테리아수를 대한 모형화를 해보자. 결과적으로 다음과 같은 모형을 얻었다.

$$\hat{\ln}(y) = 3.5568 + 3.0239\left(\frac{1}{t}\right)$$

결정계수 값, 상기 모형에서 설명된 $\ln y_i$ 의 변동부분은 0.5726이다. 그러나 선형회귀모형으로 성능의 비교는 원래 단위에서 비교해야지 로그변환(log transformation)에서의 값으로 비교해서는 안된다. 사실상 PRESS 통계량과 s^2 는 변환된 y 를 원래 단위로 변환함으로써 잔차와 PRESS 통계량을 구할 수 있다. 적정화된 수치는 적합한 변환된 모형으로부터 $\ln y$ 자료의 역대수(antilog)를 취함으로써 얻을 수 있다. 결과는 아래의 표에 주어져 있다. 잔차로부터

y	$\ln y$	$\hat{y} = \text{antilog}(\ln y)$	$\text{residual} = y_i - \hat{y}_i$
-----	---------	-----------------------------------	-------------------------------------

(Natural Units)			
355	6.580719	721.05784	-366.057842
211	5.068773	158.97918	52.020824
197	4.564791	96.04254	100.957464
166	4.312800	74.64923	91.350774
142	4.161606	64.17447	77.825525
106	4.060809	58.02124	47.978761
104	3.988812	53.99069	50.009306
60	3.934814	51.15262	8.847385
56	3.892815	49.04877	6.951229
38	3.859216	47.42817	-9.428168
36	3.831726	46.14213	-10.142129
32	3.808818	45.09711	-13.097108
21	3.789434	44.23137	-23.231367
19	3.772819	43.50254	-24.502543
15	3.758420	42.88062	-27.880617

$$s(\text{natural}) = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-2}} = 113.8517$$

여기에서 잔차는 원래단위이다. 적절한 PRESS 잔차는 변환 모형의 PRESS 잔차로부터 얻을 수 있다. 물론 후자의 것은 기준 회귀 산출로부터 일상적으로 얻는다. 우리는 $\ln(y_{i,-i})$ 의 역대수(antilog)와 그리고 $y_i\text{-antilog}(y_{i,-i})$ 로 i 번째 PRSS 잔차를 계산할 수 있다. 이러한 PRSS 잔차의 제곱합은 자연단위로 PRESS 통계이다. 결과는 다음과 같다.

y_i	$\ln y_{i_n-i}$	$\text{antilog}(\ln y_{i_n-i})$	$y_i - \text{antilog}(y_{i_n-i})$
355	9.141176	9331.73395	-8976.733955
211	5.015456	150.72487	60.275131
197	4.501030	90.10986	106.890139
166	4.254816	70.44387	95.556126
142	4.104390	60.60578	81.394216
106	4.015323	55.44119	50.558805
104	3.936482	51.23805	52.761949
60	3.921392	50.47064	9.529360
56	3.881135	48.47921	7.520789
38	3.879547	48.40230	-10.402304
36	3.855297	47.24265	-11.242651
32	3.842389	46.63673	-14.636733
21	3.864247	47.66735	-26.667345
19	3.857934	47.36737	-28.367375
15	3.868534	47.87217	-32.872168

이 결과 PRESS는 287545.5이고 절대 PRESS 잔차의 합으로

$$\sum_{i=1}^n |y_i - \text{antilog}(\ln y_{i_n-i})| = 9565.41$$

s와 PRESS 지식을 근거로 하여 역기하모형보다 단순 선형모형이 더 바람직한 것은 명백하다. 자연단위로의 성능 범주의 가치는 다른 전이된 모형과 비교할 근거가 될 것이다.

쌍곡선(Hyperbola)

쌍곡선 모형을 이용하여 생존 박테리아 모형에 적용하면 다음과 같다.

$$\left(\hat{\sqrt{y_i}} \right) = 0.031748 - 0.043603 \left(\sqrt{t} \right)$$

변환된 $(1/y)$ 의 결정계수는 0.2918으로 주어진다.

다시 원래 단위 s^2 는 잔차로 부터 계산할 수 있다. 결과는 다음과 같다.

y_i	$(1/y_i)^{-1}$	$y_i - (1/y_i)^{-1}$
355	84.34950	439.349503487
211	100.54054	110.459462891
197	58.09407	138.905929279
166	47.96837	118.031634216
142	43.42683	98.573170922
106	40.84853	65.151468803
104	39.18670	64.813295306
60	38.02644	21.973557643
56	37.17045	18.829551760
38	36.51291	1.487090748
36	35.99198	0.008020624
32	35.56909	-3.569092346
21	35.21895	-14.218950213
19	34.92427	-15.924269218
15	34.67284	-19.672839146

$$s(\text{natural}) = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-2}} = 140.9618$$

결과로 얻은 자연단위의 잔차 평균 제곱은 역기하모형에서 사용되었던 접근방법과 유사하게 쌍곡선으로 계산된다. PRESS 잔차는 다음에 적힌 마지막 항목에 기록되어 있다.

y_i	$(1/y_i)_{-i}$	$[(1/y_i)_{-i}]$	$y_i - [(1/y_i)^{-1}] = e_{i_n-i}$
355	-0.06487237	-15.41488	370.414883941
211	0.01092692	91.51709	119.482911693
197	0.01829069	54.67263	142.327374759
166	0.02192253	45.61517	120.384830260
142	0.02417879	41.35857	100.641433866
106	0.02561640	39.03749	66.962514443
104	0.02678830	37.32973	66.670266031
60	0.02710776	36.88980	23.110196513
56	0.02770028	36.10071	19.899288145
38	0.02748589	36.38230	1.617700680
36	0.02778456	35.99122	0.008782107
32	0.02780747	35.96156	-3.961560568
21	0.02646299	37.78862	-16.788624021
19	0.02616762	38.21517	-19.215166824
15	0.02487561	40.20002	-25.200023241

여기에서 $\text{PRESS} = 207525.00$ 이고 $\sum_{i=1}^n \left| y_i - [(1/y_i)_{-i}]^{-1} \right|^2 = 1096.686$ 이다.

포물선(Parabola)

2차항을 포함한 포물선 모형을 생존 박테리아 자료에 적용하였을 때 최소제곱회귀 모형은 다음과 같다.

$$\hat{y} = 350.6066 - 51.5910t + 2.00979t^2$$

여기에서 $R^2 = 0.95250$ 이고 $s = 22.580$ 이다. 잔차와 PRESS 잔차는 다음과 같다.

y_i	\hat{y}_i	$y_i - \hat{y}_i$	$y_i - \hat{y}_{i-i}$
-------	-------------	-------------------	-----------------------

355	301.02353	53.9764706	100.8351648
211	255.45630	-44.4563025	-60.7382319
197	213.90491	-16.9049127	-20.2617959
166	176.36936	-10.3693601	-11.8579243
142	142.84964	-0.8496445	-0.9671106
106	113.34577	-7.3457660	-8.4735665
104	87.85772	16.1422754	18.8996443
60	66.38552	-6.3855204	-7.5223881
56	48.92915	7.0708468	8.2786650
38	35.48862	2.5113769	2.8969503
36	26.06393	9.9360698	11.3097638
32	28.52353	11.3449257	12.9735364
21	19.26206	1.7379444	2.0830557
19	21.88487	-2.8848739	-3.9414466
15	20.65507	-13.5235294	-25.2637363

여기에서 $\text{PRESS} = 15925.820$ 이고 $\sum_{i=1}^n |y_i - [(1/y_i)_{-i}]^{-1}| = 296.303$ 이다.

PRESS의 수치, 절대 PRESS 잔차의 합, 근잔차(root residuals) 평균 제곱을 기준으로 할 때, 쌍곡성모형과 포물선모형이 단순선형회귀보다 우수하다. 예측범주는 원래 변환항에서 반응을 예측하는 변환 모형의 능력을 기초로 하였다. 포물선 모형은 모든 다른 모형들에 더 좋은 것처럼 보인다 만약 우리가 변환항에서 회귀에만 계산을 국한시킨다면 포물선은 결정계수가 0.9525이며 단순선형회귀의 결정계수보다 더 실제적으로 증가한다. 그러나, 잔차의 문제성은 여전히 모든 모형에 존재한다. 심지어 포물선의 경우에서 과소평가가 계속적인 과대평가 후에 나타나며 다시 과소평가가 나타난다.

모형비교 통계량

	s	PRESS	절대 PRESS 잔차
단순회귀모형	41.83	35910.85	513.18
역기하모형	113.85	287545.50	9565.41
쌍곡선모형	140.96	207525.00	1096.69
포물선모형	22.58	15925.82	269.30

다음의 내용은 위의 예제에 대한 R code이다.

```

n <- c(355, 211, 197, 166, 142, 106, 104, 60, 56, 38, 36, 32, 21, 19, 15)
t <- c(seq(1, 15, 1))
plot(t, n, type="p", pch=19, main='Plot of Bacteria Deaths')
fit <- lm(n~t)
summary(fit)
inf <- influence(fit)
pre <- sum((fit$resi / (1-inf$hat))^2)
spre.res <- sum(abs(fit$resi/(1-inf$hat)))
pre.res <- fit$resi/(1-inf$hat)
fit$fitt
fit$resi
pre.res

lny <- log(n)
x1 <- 1/t
fit1 <- lm(lny~x1)
summary(fit1)
fit1$fitt ## fitted value(log units) ##
exp(fit1$fitt) ## fitted value(natural units) ##
resi <- n-exp(fit1$fitt) ## residual(natural units)
s <- sqrt((sum((resi)^2))/13) ## residual standard error(natural) ##
inf <- influence(fit1)
pre <- lny - (fit1$resi / (1-inf$hat)) ## press statistic(log units) ##
exp(pre) ## press statistic(fitted valud) ##
n-exp(pre) ## press statistic(natural units) ##
sum(abs(n-exp(pre))) ## absolute press residuals ##
sum((n-exp(pre))^2)

```

```

y1 <- 1/n
fit2 <- lm(y1~x1)
summary(fit2)
1/(fit2$fitt) ## fitted value(natural units) ##
resi <- n-1/(fit2$fitt) ## residual(natural units)
s <- sqrt((sum((resi)^2))/13) ## residual standard error(natural) ##
inf <- influence(fit2)
pre <- y1 - (fit2$resi / (1-inf$hat)) ## press statistic(log units) ##
1/pre ## press statistic(fitted valud) ##
pre.res <- n-(1/pre) ## press statistic(natural units) ##
sum((pre.res)^2) ## press statistics ##
sum(abs(pre.res)) ## press residual ##

```



```

t2 <- t^2
fit3 <- lm(n~t+t2)
summary(fit3)
fit3$fitt
fit3$resi
inf <- influence(fit3)
pre.res <- fit3$resi / (1-inf$hat)
sum((pre.res)^2)
sum(abs(pre.res))

```

다중회귀모형에서 변환(Transformations On Multiple Regressors)

자료를 변환하는 필요성은 단순회귀사례에서 꽤 명백하다. 왜냐하면 이차원 도표에서 종종 곡선의 형태가 나타나기 때문이다. 다중회귀에서는 공선성(collinearity)은 이차원 도표를 왜곡할 수 있기 때문에 곡선형태를 탐지하는 것은 상대적으로 어렵다. 물론 5장에서 부분 지렛값 도표(partial leverage plot)를 통해 회귀변수(regressor)에 대한 변환이 필요하다는 것을 알 수 있었다. 다음에 올 내용은 다중회귀에서 모형 수정(alteration)에 관해서 제안할 것이다. 이러한 수정은 적합이나 예측을 향상시키는데 도움이 된다.

모형의 상호작용(Interaction in the Model)

모형을 향상시키는 가장 단순한 방법은 회귀변수사이에 교호작용(interaction)을 고려하는 것이다. 3장의 지시변수들에 대한 내용에서 상호작용의 의미(implication)를 상기할 필요가 있다. 교호작용(interaction)과 3장과 8장에서의 변수들간의 다중공선성(multicollinearity)과

운동에는 주의해야한다. x_1 과 x_2 사이의 교호작용(interaction)은 변수 x_1 에 의한 반응변수의 변화율(rate of change)이 변수 x_2 의 수준(level)에 의존한다는 것이다. 다른 말로 하면 회귀변수들의 가법성(additivity)으로 설명할 수 없는 편의(deviation)가 일반적인 다중선형회귀모형(multiple linear regression model)에 영향을 미친다는 것이다. 예를 들어, 회귀변수가 3개 있는 경우 전체 3개의 교호작용(interaction)이 있고 모형의 형태를 다음과 같다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \quad (7.23)$$

식 (7.23)은 일차 회귀항(first-order terms)과 3개의 선형 교호작용항(interaction)이 포함되어 있다. 앞에서 설명했던 교호작용(interaction)이 모형에 어떻게 적용되는가를 확인하기 위해서는 변화률을 고려하여보자.

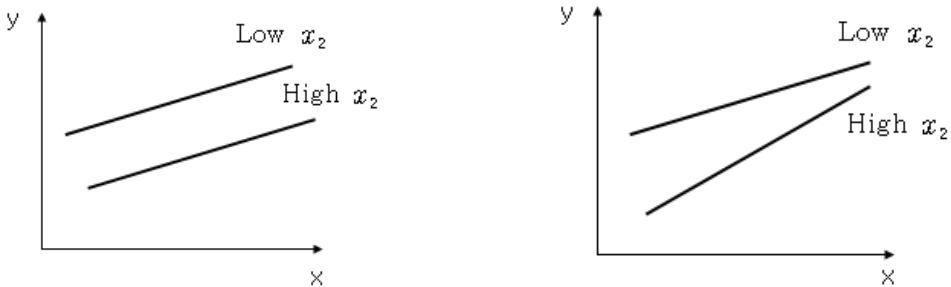
$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_{12} x_2 + \beta_{13} x_3$$

x_1 방향의 회귀기울기(slope of the regression)는 x_2, x_3 에 의존한다. 그림 7.9는 회귀변수가 2개 있는 교호작용(interaction)으로 그림으로 나타낸 것이다.

FIGURE 7.9

(a) No interaction x_1 and x_2

(b) Interaction between x_1 and x_2



회귀변수에 대한 변환(Power Transformations on the Regressor Variables (Box-Tidwell Procedure))

이 절의 초기에 단순회귀사례에서의 변환을 논하였다. 자료 도표(plot of data)의 형태를 통해 변환의 종류를 결정할 수 있지만, 항상 단순하게 답을 주진 않는다. 도표의 전형적인 형태가 자료의 잡음으로 인해 흐려지며 다양한 종류의 비선형 도표와 몇몇은 유사한 형태를 보인다. 다중회귀의 경우에는 다중공선성은 이차원도표(two dimension plot)가 개별 회귀변수의 실질적인 의미를 나타내는 것을 방해하여 부분 도표(partial plot)에서는 변환이 필요하다고 하지만 이차원도표에서는 변환이 필요없다고 나타내지도 한다. 그 결과 자료분석시

활용 가능한 추가적인 기술이 필요하다.

Box와 Tidwell (1962)은 다음과 같은 모형 형태에서 멱지수들(exponents) $\alpha_1, \alpha_2, \alpha_3, \alpha_k$ 를 추정하는 과정을 제안하였다.

$$y = \beta_0 + \beta_1 w_1 + \cdots + \beta_k w_k + \varepsilon \quad (7.24)$$

여기에서

$$w_j = \begin{cases} x_j^{\alpha_j} & \text{if } \alpha_j \neq 0 \\ \ln(x_j) & \text{if } \alpha_j = 0 \end{cases}$$

이 방법은 한 개 이상의 회귀변수에 대한 멱지수들(exponents)을 적용한다. Box와 Tidwell의 방법은 다중선형회귀 소프트웨어에서 제공하고 있으므로 쉽게 사용할 수 있다.

이 방법은 식 (7.24) 모형 함수의 테일러 급수 전개(Taylor Series expansion)(별책 B. 9 참조)를 기초로 한다. 여기에서는 $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_k)$ 를 초기값으로 가정하자. 보통은 1.0의 값을 사용하다. 값 $\alpha_0 = (\alpha_{1,0}, \alpha_{2,0}, \dots, \alpha_{k,0})$ 근처에서 테일러 급수 전개(Taylor Series expansion)하면 다음과 같다.

$$\begin{aligned} E(y) \cong & [f(\alpha_1, \alpha_2, \dots, \alpha_k)]_{\alpha=\alpha_0} + (\alpha_1 - \alpha_{1,0}) \left[\frac{\partial f}{\partial \alpha_1} \right]_{\alpha=\alpha_0} \\ & + (\alpha_2 - \alpha_{2,0}) \left[\frac{\partial f}{\partial \alpha_2} \right]_{\alpha=\alpha_0} + \cdots + (\alpha_k - \alpha_{k,0}) \left[\frac{\partial f}{\partial \alpha_k} \right]_{\alpha=\alpha_0} \end{aligned}$$

여기에서 $[f(\alpha_1, \alpha_2, \dots, \alpha_k)]_{\alpha=\alpha_0}$ 는 단지

$$\beta_0 + \beta_1 x_1^{\alpha_{1,0}} + \beta_2 x_2^{\alpha_{2,0}} + \cdots + \beta_k x_k^{\alpha_{k,0}}$$

초기값을 모두 1.0을 사용하면, 다음과 같이 얻을 수 있다.

$$\begin{aligned} E(y) \cong & \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + (\alpha_1 - 1)\beta_1 x_1 \ln x_1 \\ & + (\alpha_2 - 1)\beta_2 x_2 \ln x_2 + \cdots + (\alpha_k - 1)\beta_k x_k \ln x_k \end{aligned} \quad (7.25)$$

따라서, (7.25)의 모형은 다음과 같이 표현할 수 있다.

$$E(y) \equiv \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_k z_k \quad (7.26)$$

여기에서

$$\left. \begin{array}{l} \gamma_j = (\alpha_j - 1)\beta_j \\ z_j = x_j \ln x_j \end{array} \right\} \quad j=1,2,\dots,k$$

α_j 를 추정하는 단계적 절차는 다음과 같다.

1. 아래의 모형으로 다중선형회귀분석을 실시한다.

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

추정된 모수를 b_0, b_1, \dots, b_k 로 쓴다.

2. $x_1, x_2, \dots, x_k, z_1, z_2, \dots, z_k$ 에 대해 y 를 회귀분석한다. z 들의 계수 $\gamma_1, \gamma_2, \dots, \gamma_k$ 를 추정한 후 추정된 값을 $\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_k$ 로 쓴다.
3. $\alpha_1, \alpha_2, \dots, \alpha_k$ 를 아래의 식을 이용하여 추정한다.

$$\hat{\alpha}_j = \frac{\hat{\gamma}_j}{b_j} + 1 \quad (j=1,2,\dots,k) \quad (7.27)$$

식 (7.27)에 주어진 결과는 α_j 에 대한 갱신 추정량(updated estimate)으로 볼 수 있다. 종종 한번 계산으로 충분할 때도 있지만, 잔차제곱합(residual sum of squares)를 주의깊에 살펴보아야 한다. 두 번째 단계에서는 다음과 같다.

1. $w_1^* = x_1^{\hat{\alpha}_1}, w_2^* = x_2^{\hat{\alpha}_2}, \dots, w_k^* = x_k^{\hat{\alpha}_k}$ 를 이용하여 아래의 모형을 적합하여

$$E(y) = \beta_0 + \beta_1 w_1^* + \cdots + \beta_k w_k^*$$

추정량 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ 를 얻는다.

2. $z_1^* = w_1^* \ln w_1^*, z_2^* = w_2^* \ln w_2^*, \dots, z_k^* = w_k^* \ln w_k^*$ 을 정의하라.

3. $w_1^*, w_2^*, \dots, w_k^*, z_1^*, z_2^*, \dots, z_k^*$ 에 대해 y 를 회귀분석한 후 $z_1^*, z_2^*, \dots, z_k^*$ 의 새로운 계수를 $\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_k$ 로 한다.

4. 갱신된 $\hat{\alpha}_j$ 의 값을 아래의 식으로 계산한다.

$$\hat{\alpha}_j = \left(\frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1 \right) \quad (7.28)$$

본질적으로 2차 반복식은 처음 식과 같은 과정에 따라 진행된다. 그러나, x_j 은 x_j^{qj} 로

대치되며, 여기서 $\hat{\alpha}_j$ 는 이전의 반복으로부터 구한 값이다. $(x_j^{\alpha_j})$ 의 면지수의 추정량은 $(\hat{\gamma}_j / \hat{\beta}_j) + 1$ 로 나타난다. 이것으로 식 (7.28)의 결과를 얻을 수 있다.

사용자는 여기에서 테일러 급수 전개(Taylor Series expansion)의 역할에 대하여 이해해야 한다. 식 (7.26)의 형태는 원래의 모형(linear in the x 's)에 z 와 관련된 부분을 더한 것이다. 더해진 부분은 x 들의 선형성(linearity)으로 설명할 수 없는 부분을 고려한 것이다. 그 결과, 추정값 γ_j 가 크면 변환이 필요하다는 것을 의미하고 값이 0에 가까우면 $(\alpha_j - 1) \approx 0$ 으로 변환이 필요하지 않다는 것을 의미한다. ($\alpha_j \approx 2$)이면 2차 형태, ($\alpha_j \approx 0$)이면 자연로그(natural log), ($\alpha_j \approx -1$)이면 역수 변환 등이 필요하다는 것을 의미한다.

예제 7.5 풍차 자료(Windmill Data)

Observation Number, i	Wind Velocity (mph), x_i	DC Output y_i	Observation Number, i	Wind Velocity (mph), x_i	DC Output y_i
1	5.00	1.582	13	4.60	1.562
2	6.00	1.822	14	5.80	1.737
3	3.40	1.057	15	7.40	2.088
4	2.70	0.500	16	3.60	1.137
5	10.00	2.236	17	7.85	2.179
6	9.70	2.386	18	8.80	2.112
7	9.55	2.294	19	7.00	1.800
8	3.05	0.558	20	5.45	1.501
9	8.15	2.166	21	9.10	2.303
10	6.20	1.866	22	10.20	2.310
11	2.90	0.653	23	4.10	1.194
12	6.35	1.930	24	3.95	1.144
			25	2.45	0.123

전력을 얻기 위해 풍차를 사용하게 되는데, 이때 바람의 속도와 전력을 관계를 알고자 한다. 자료의 산점도를 살펴보면 선형관계가 이루어지지 않음을 알수 있다. 이는 변수변환을 해야 함을 나타내고 이것을 Box-Tidwell 과정을 통해 이 자료집합에 적용하며 이는 여기에서 보여준다. y 가 전력량이고 회귀변수는 바람 속도이다. x 에 대한 적절한 변환을 하기 위해서 우리는 모형을 고려해 보자.

$$y = \beta_0 + \beta_1 w + \varepsilon$$

여기에서 $w = x^\alpha$ 이고 추정될 수 있다. 이 절에서 초기에 언급된 계산을 따라서 회귀를

적용할 수 있다.

$$\hat{y} = b_o + b_I x$$

$b_0 = 0.1309, b_I = 0.2411, R^2 = 0.87450$ 이며 $s = 0.2361$ 이다. 다음에 x 에 대한 y 의 회귀를 하고 $z = x \ln x$ 로 적합하면 결과는 다음과 같다.

$$\hat{y} = -2.4168 + 1.5344x - 0.4626 z$$

그리고 (7.27)에 의해서 α 의 처음 추정량은 다음과 같다.

$$\hat{\alpha} = \frac{-0.4626}{0.2411} + 1.0 = -0.92$$

두번째 단계는 $w^* = x^{0.92}$ 에 대한 회귀 y 에 대해서 이루어진다. 결과는 다음에 대해서 주어진다.

$$\hat{y} = 3.1039 - 6.6784w^*$$

회귀는 w^* 과 $z^* = w^* \ln w^*$ 에 대한 y 의 적정 후에 결과를 가지고

$$\hat{y} = 3.2409 - 6.445w^* + 0.5994z^*$$

이 결과들로부터 다음 단계에서 $\hat{\alpha}$ 의 계산은 식 (7.28)에 의해 주어지고

$$\hat{\alpha} = \left[\frac{0.5994}{-6.6784} + 1 \right] [-0.92] = -0.84$$

계속 반복하면 값이 일정해진다. 이 값을 이용해서 회귀분석을 실시한 결과가 Box와 Tidwell방법에 의한 변수변환으로 구한 추정 회귀식이 된다.

다음의 내용은 위의 예제에 대한 R code이다.

```
x <- c(5.00, 6.00, 3.40, 2.70, 10.00, 9.70, 9.55, 3.05, 8.15, 6.20,  
      2.90, 6.35, 4.60, 5.80, 7.40, 3.60, 7.85, 8.80, 7.00, 5.45,  
      9.10, 10.20, 4.10, 3.95, 2.45)  
  
y <- c(1.582, 1.822, 1.057, 0.500, 2.236, 2.386, 2.294, 0.558, 2.166, 1.866,
```

```

0.653, 1.930, 1.562, 1.737, 2.088, 1.137, 2.179, 2.112, 1.800, 1.501,
2.303, 2.310, 1.194, 1.144, 0.123)

wind <- data.frame(x, y)

fit <- lm(y~x, data=wind)

summary(fit)

z <- x*log(x)

wind <- data.frame(x, z, y)

fit1 <- lm(y~x+z, data=wind1)

summary(fit1)

w <- x^(-0.92)

wind2 <- data.frame(w, y)

fit2 <- lm(y~w, data=wind2)

summary(fit2)

z1 <- w*log(w)

wind3 <- data.frame(w, z1, y)

fit3 <- lm(y~w+z1, data=wind3)

summary(fit3)

box.tidwell(y~x, data=wind, verbose=T)

```

반응에 대한 BOS-COX 변환(Box-Cox Transformation On The Response)

많은 경우 다중 멱지수(multiple exponents)을 추정하는 Box와 Tidwell 방법보다는 반응변수에 대한 멱변환(power transformation) 방법을 사용한다. 이 방법은 하나의 모수만 추정할 필요가 있으며, 하나의 중요한 변환의 틀(an important family of transformation)로 접근가능하다. Box-Cox(1964) 변환은 y^λ 를 사용한다. 여기에서, λ 는 자료로부터 추정한다. 아래에서 주어진 w 를 반응변수(response variable)로 정의한다.

$$w_j = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln y & \text{if } \lambda = 0 \end{cases}$$

Box-Cox 과정의 목적은 동시에 모형에 있는 λ , 모수 $\beta_0, \beta_1, \dots, \beta_k$ 를 추정하는 것이다.

$$w_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (7.29)$$

반응변수에 대한 멱변환은 많은 이점을 가지고 있다. 이것은 반응변수와 회귀변수 사이에 존재하는 다양한 관계를 다룰 수 있다. 이것은 회귀변수의 변환을 배제하는 것은 아니다.

예로 회귀변수의 제곱(squares)항이나 교차(cross)항이 식 (7.29)의 모형에 포함될 수 있다. 만약 간단히 생각하면 λ 가 범위[-2, 2]에 있을 때, Box-Cox 과정은 음 또는 양 제곱근(negative or positive square root)($\lambda=1/2, \lambda=-1/2$), 제곱(square)($\lambda=2$), 로그(log)($\lambda=0$), 역제곱(inverse square)($\lambda=-2$) 그리고 -2와 2사이에 다른 분수식 멱(fractional power)변환을 적용가능하게 한다.

식 (7.29)의 모형 적합성은 물론 β 들과 중요한 모수 λ 의 동시추정에 관계있다. $\lambda = 0$ 일 때의 적절한 변환으로 $\ln y$ 이 사용되는 것은 λ 이 0에 가까워짐에 따라 $(y^\lambda - 1)/\lambda$ 의 극한이 $\ln y$ 에 가까워진다는 결과이고, λ 에 대해서 연속이라는 것을 의미한다. 그림을 중심으로한 많은 추정방법은 λ 와 회귀계수를 동시추정하도록 되어있다. 또한, 반응에 대한 적절한 계량(proper metric)을 결정하려고 할 때 이러한 도표는 분석가에게 유용한 진단적 정보를 제공한다. 추정과정을 고려되는 중요한 방법은 최대우도법(maximum likelihood)이다.

β 와 λ 를 추정하는 최대우도((Maximum Likelihood Method for Estimating β 's and λ))

최대우도추정 과정이 2장에서 소개되었다. 식 (7.29) 모형에 대한 우도함수(likelihood function)를 가지고 시작해보자. 여기에서 오차(error)는 정규분포(normal distribution)을 따르고 등분산 σ^2 을 가지고 서로독립이다. 적절한 우도함수는 다음과 같다.

$$L(\beta, \lambda, \sigma^2) = \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} e^{(-1/2\sigma^2)(\mathbf{w}-\mathbf{X}\beta)(\mathbf{w}-\mathbf{X}\beta)} \cdot J(\lambda, \mathbf{y}) \quad (7.30)$$

이것은 2장과 별책 B에 언급된 우도함수(likelihood function)이다. 그러나 이 우도함수에는 추가 모수 λ 와 아래의 자코비안항(Jacobian)이 들어있다.

$$J(\lambda, \mathbf{y}) = \prod_{i=1}^n \frac{\partial w_i}{\partial y_i}$$

우도는 선형모형의 형태에 기초한다.

$$\mathbf{w} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

여기에서 \mathbf{w} 는 식 (7.29)에 있는 반응변수로 $n \times 1$ 벡터(vector)이다.

여기에서의 면변환은 자료에서 추론된 λ 값으로 식 (7.29)에 적용한 것이다. λ 값에 따라 변하는 반응변수의 척도(scale)와 우도(likelihood)의 관계를 살펴볼 필요는 있다. 예를 들어, $\lambda=0$ (반응변수는 $\ln y$)일 때 변환된 반응변수의 값과 $\lambda=2$ (반응변수는 y^2)일 때는 상당히 다를 수 있다. 자코비안(Jacobian) $J(\lambda, y)$ 이 척도의 변화(change of scale)를 설명한다. 이러한 조정은 λ 의 선택에 따라 우도값이 변하기 때문에 필요하다. 변환에 따른 자코비안(Jacobian)은 다음과 같다.

$$J(\lambda, y) = \prod_{i=1}^n \frac{\partial w_i}{\partial y_i} = \prod_{i=1}^n y_i^{\lambda-1}$$

변환족(family of transformations)에서 다양한 형태를 평가하는데 있어 척도화된 우도는 모형의 약간씩 다른 형태를 고려하여 쉽게 얻을 수 있다. 처음에 표준형에서 변환을 다시 써보자. 변환은 다음과 같이 재정의할 수 있다.

$$\begin{aligned} z_i &= \frac{w_i}{\{(\lambda, y)\}^{1/n}} = \frac{w_i}{\left\{ \left(\prod_{i=1}^n y_i \right)^{1/n} \right\}^{i-1}} \\ &= \frac{y_i^\lambda - 1}{\lambda(\bar{y})^{\lambda-1}} \quad (\lambda \neq 0) \\ &= \frac{\ln y_i}{(\bar{y})^{\lambda-1}} \quad (\lambda = 0) \end{aligned} \quad (7.31)$$

양 $y_i = \left(\prod_{i=1}^n y_i \right)^{1/n}$, y 의 기하평균(geometric mean)으로 변환된 반응변수(w)의 표준화를 제공해 준다. z 를 반응변수로 할 때, 로그우도의 최대값은 다음과 같다.

$$\max[\ln L(\lambda, \beta)] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (\mathbf{z} - \mathbf{X}\hat{\beta})'(\mathbf{z} - \mathbf{X}\hat{\beta})$$

여기에서 $\hat{\sigma}^2$ 는 σ^2 의 최대우도추정량(maximum likelihood estimator)이다. 3장에서 고정된 λ 에 대해서, $\hat{\sigma}^2$ 는 $\hat{\sigma}^2 = \text{SS}_{\text{Res}}/n$ 으로 주어진다는 것을 알았다. 따라서, 로그우도함수의 최대값은 다음과 같이 쓸 수 있다.

$$\max[\ln L(\beta, \lambda)] = -\frac{n}{2} \ln \hat{\sigma}^2 \quad (7.32)$$

λ 에 대해 최대화하는 것은 $\hat{\sigma}^2$ 를 최소화하는 것과 동일하다. 따라서, 잔차제곱합(residual sum of squares)을 최소화하는 것이다. 이것으로 변환을 위한 λ 를 결정하는 단계적인 과정을 이끌어낸다.

- (i) -2와 2사이를 격자(grid)로 짤라 λ 를 선택한다.
- (ii) 각각의 λ 에 대하여 다음 모형을 적합한다.

$$\mathbf{z} = \mathbf{X}\beta + \varepsilon \quad (7.33)$$

여기에서 \mathbf{z} 은 (7.31)에 정의되어 있다.

(iii) λ 에 대하여 (7.32)에 있는 $\max[\ln L(\beta, \lambda)]$ 의 그림을 그린다.

(iv) $\max[\ln L(\beta, \lambda)]$ 을 가장 크게 하는 값으로 추정량 $\hat{\lambda}$ 을 선택한다.

이 과정을 통해서 선택된 $\lambda = \hat{\lambda}$ 으로 반응변수 y 이 변환을 해야한다. 만약 λ 의 최적값이 1.0과 유의하게 다르다면 반응변수의 변환을 통해 더 좋은 적합결과를 얻을 수 있다는 걸 의미한다. 구간 [-2.0, 2.0]에서 10–20값을 사용하면 목적에 맞는 결과를 얻기에 충분하다. 분명히 과학적 현상을 묘사하는 모형을 얻을 수 있는 값이 이 구간에 존재한다. 만약 λ 가 2에 가까우면 반응변수의 제곱이 합리적인 접근방법이다; 만약 $\lambda \approx -1$ 이면 역수(reciprocal) 변환이 적용된다. 만약 0.15이면 경험적으로 반응변수로 $y^{0.15}$ 를 사용하는 것 보다 로그변환(log transformation)을 사용하는 것이 더 적절할 수 있다. 한 가지 명심해야하는 것은 식 (7.23)보다는 잔차제곱합(residual sum of squares)을 그림에 사용할 수 있다는 것이다. λ 으로 선택된 값은 잔차제곱합(residual sum of squares)을 최소화하는 것이다. 만약 추정된 λ 가 어떤 특정한 값과 유의하게 다른가를 결정하기 위해서는 가설검정이나 신뢰구간의 방법이 필요하다.

λ 에 대한 신뢰구간(Confidence Interval on λ)

변환에 필요한 중요한 판단기준인 적절한 λ 은 λ 에 대한 신뢰구간으로부터 얻을 수 있다. 점근적 $100(1-\alpha)\%$ 신뢰구간(Atkinson(1982), Draper and Smith(1981) 참조)은 우도비(likelihood ratio) 이론을 기초로 한다. 임의의 값 $\lambda = \lambda_0$ 이라고 하자. 값 λ_0 가 $100(1-\alpha)\%$ 신뢰구간에 포함되느냐 되지 않느냐는 아래의 차이에 의존한다.

$$\max[\ln L(\lambda, \beta)] - \max[\ln L(\beta, \lambda_0)] = -\frac{n}{2} \ln \hat{\sigma}^2(\hat{\lambda}) - \left[-\frac{n}{2} \ln \sigma^2(\lambda_0) \right] \quad (7.34)$$

여기에서 $\hat{\sigma}^2(\hat{\lambda})$ 는 앞에서 처럼 식 (7.33)의 모형에 대한 오차분산의 최대우도추정량으로 이는 λ 는 $\hat{\lambda}$ 로 정하였고, Box-Cox 과정에서 SS_{Res} 를 최소화로 하는 값이다. 앞에서 언급한 것과 같이 이것은 λ 의 최대우도추정량을 제공하고 그 결과 λ 의 모든 값에서 우도의 최대값을 만든다. $\hat{\sigma}^2(\hat{\lambda})$ 는 식 (7.33)과 동일한 모형에서 다음과 같이 계산된다.

$$\hat{\sigma}^2(\lambda_0) = \frac{SS_{Res}(\lambda_0)}{n}$$

그러나 z 의 변환에 대한 모형은 $\lambda=\lambda_0$ 에 의해서 제약된다. 식(7.34)의 원쪽항의 값(quantity)은 양수이다. 따라서,

$$-\frac{n}{2} \ln \hat{\sigma}^2(\hat{\lambda}) - \left[-\frac{n}{2} \ln \sigma^2(\lambda_0) \right] \geq 0$$

제약식이 있는 경우, 더 적은 모수가 추정된다. 변환의 모수로 λ_0 의 합리성을 나타내는 증거는 이 차이에 놓여 있다. 이러한 접근법을 우도비 접근법(likelihood ratio approach)이라 하며 점근분포(asymptotic distribution)는 카이제곱(chi-square)분포와 연관되어 있다. $100(1-\alpha)\%$ 신뢰구간 내에 포함될 λ_0 의 값은 다음을 만족한다.

$$-\frac{n}{2} \ln \hat{\sigma}^2(\hat{\lambda}) - \left[-\frac{n}{2} \ln \sigma^2(\lambda_0) \right] \leq \frac{1}{2} \chi_{\alpha,1}^2 \quad (7.35)$$

여기에서 $\chi_{\alpha,1}^2$ 는 자유도(degree of freedom) 1을 갖는 카이제곱분포에서 상위 α 백분위수(upper α percentile)이다. 카이제곱분포의 백분위수값은 표 C. 6.에 있다.

λ 에 대한 합리적인 값을 그래프로 보여주는 것이 도움이 되는데 λ 에 대한 $(-n/2)\ln\sigma^2(\lambda)$ 의 도표로부터 얻을 수 있다. 아래의 부등식으로부터 근사신뢰구간을 얻을 수 있다.

$$-\frac{n}{2} \ln \hat{\sigma}^2(\hat{\lambda}) \geq -\frac{n}{2} \ln \hat{\sigma}^2(\hat{\lambda}) - \frac{1}{2} \chi_{\alpha,1}^2$$

설명은 그림 7.10에 주어진다. 근사 $100(1-\alpha)\%$ 신뢰구간을 나타내는 수직선이 보이고 타당한 변환모수(transformation parameter)의 값이 나타난다.

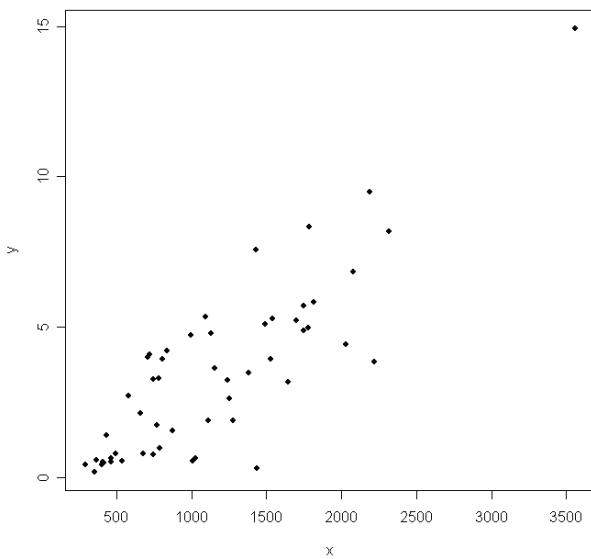
예제 7.6 Box-Cox 변환 : 전력효용자료(Box-Cox Transformation for Electronic Utility Data)

전력 효용은 한달동안의 에너지 총 사용량과 시간당 최대 에너지 요구량의 관계를 통해서 알 수 있다.

Customer	x (KWH)	y (KW)	Customer	x (KWH)	y (KW)
1	679	0.79	27	837	4.20
2	292	0.44	28	1748	4.88
3	1012	0.56	29	1381	3.48

4	493	0.79	30	1428	7.58
5	582	2.70	31	1255	2.63
6	1156	3.64	32	1777	4.99
7	997	4.73	33	370	0.59
8	2189	9.50	34	2316	8.19
9	1097	5.34	35	1130	4.79
10	2078	6.85	36	463	0.51
11	1818	5.84	37	770	1.74
12	1700	5.21	38	724	4.10
13	747	3.25	39	808	3.94
14	2030	4.43	40	790	0.96
15	1643	3.16	41	783	3.29
16	414	0.50	42	406	0.44
17	354	0.17	43	1242	3.24
18	1276	1.88	44	658	2.14
19	745	0.77	45	1746	5.71
20	435	1.39	46	468	0.64
21	540	0.56	47	1114	1.90
22	874	1.56	48	413	0.51
23	1543	5.28	49	1787	8.33
24	1029	0.64	50	3560	14.94
25	710	4.00	51	1495	5.11
26	1434	0.31	52	2221	3.85
			53	1526	3.93

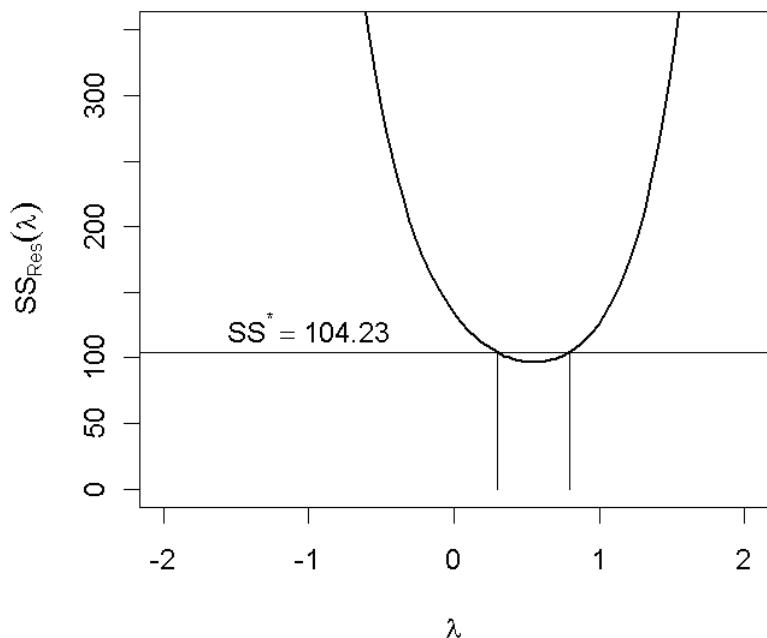
이 자료의 산점도를 그려보면 회귀 변수(시간당 최대 에너지 요구량)가 증가함에 따라 오차 분산도 증가하는 이분산성을 가지는 것을 알 수 있다.



이것은 일반적인 선형회귀로는 정확한 추정치를 얻기 어렵고, 적절한 변환을 통해 변수변환후에 회귀분석해야 함을 의미한다. Box-Cox 과정은 변환이 필요한지를 알아내기 위해서 사용된다. 식(7.33)의 모형은 식(7.31)에 정의된 반응 z 과 함께 사용되었다. -2에서 2까지의 λ 값이 사용되었는데, λ 에 대한 $(-n/2)\ln \sigma^2(\lambda)$ 의 도표가 만들어진다. 도표는 그림7.11에 있다. 독자는 $(-n/2)\ln \sigma^2(\lambda)$ 를 극대화하는 λ 값이 또한 SS_{Res} 를 최소화 한다는 것을 명심해야 한다.

λ	$SS_{\text{Res}}(\lambda)$
-2	34,101.0381
-1	986.0423
-0.5	291.5834
0	134.0940
0.125	118.1982
0.25	107.2057
0.375	100.2561
0.5	96.9495
0.625	97.2889
0.75	101.6869
1	126.8660
2	1275.5555

Plot of residual sum of square vs lambda



다음의 내용은 위의 예제에 대한 R code이다.

```
y <- c(0.79, 0.44, 0.56, 0.79, 2.70, 3.64, 4.73, 9.50, 5.34, 6.85,
      5.84, 5.21, 3.25, 4.43, 3.16, 0.50, 0.17, 1.88, 0.77, 1.39,
      0.56, 1.56, 5.28, 0.64, 4.00, 0.31, 4.20, 4.88, 3.48, 7.58,
      2.63, 4.99, 0.59, 8.19, 4.79, 0.51, 1.74, 4.10, 3.94, 0.96,
      3.29, 0.44, 3.24, 2.14, 5.71, 0.64, 1.90, 0.51, 8.33, 14.94,
      5.11, 3.85, 3.93)
x <- c(679, 292, 1012, 493, 582, 1156, 997, 2189, 1097, 2078,
      1818, 1700, 747, 2030, 1643, 414, 354, 1276, 745, 435,
      540, 874, 1543, 1029, 710, 1434, 837, 1748, 1381, 1428,
      1255, 1777, 370, 2316, 1130, 463, 770, 724, 808, 790,
      783, 406, 1242, 658, 1746, 468, 1114, 413, 1787, 3560,
      1495, 2221, 1526)
ta5.2 <- data.frame(x, y)
plot(y~x, data=ta5.2, type="p", pch=19)

ssres<-function(lambda){
  x<-ta5.2$x
  y<-ta5.2$y
  box.cox<-function(lambda,y){
```

```

y.dot<-exp(1/(length(y))*sum(log(y)))
if(lambda==0) y.dot*log(y) else (y^lambda-1)/(lambda*(y.dot^(lambda-1)))
}
anova(lm(box.cox(lambda,y)~x))[2,2]
}

lambda<-c(-2,-1,-.5,0,.125,.25,.375,.5,.625,.75,1,2)
SS.res<-c()
for(i in 1:length(lambda)) SS.res[i]<-ssres(lambda[i])
cbind(lambda,SS.res)

remove(SS.res)
lambda<-seq(-2,2,length=100)
SS.res<-c()
for(i in 1:length(lambda)){
SS.res[i]<-ssres(lambda[i])
}

par(oma=c(0,0,1,0),mfrow=c(1,1),cex=1.5,pch=16)
plot(lambda,SS.res,type='l',ylim=c(0,350),main='Plot of residual sum of square vs lambda',lwd=2,
      xlab=expression(lambda),ylab=expression(SS[Res](lambda)))
(SS.star<-ssres(.5)*exp(qchisq(.05,1,lower.tail = F)/length(y)))
abline(h=SS.star)

tmp.f<-function(lambda) abs(ssres(lambda)-SS.star)
lambda.l<-optimize(tmp.f,interval = c(0,.5))$minimum
lambda.r<-optimize(tmp.f,interval = c(.5,1))$minimum
lines(c(lambda.l,lambda.l),c(0,SS.star))
lines(c(lambda.r,lambda.r),c(0,SS.star))
text(-1,SS.star+20,expression(SS^{**}==104.23))

```

7.4. 이진형 반응변수를 가지는 회귀분석(Regression with a Binary Response)

많은 실제 상황에서 반응변수가 기본적으로 이진이다. 즉, 두 개의 확률로 결과가 나타나며, 이들의 값으로 0 또는 1로 할당될 수 있다. 모형설정의 동기는 연속형 반응변수이 사례와 동일하다. 이진 반응변수에 회귀변수 x_1, x_2, \dots, x_k 의 역할을 결정할 필요가 있으며, 부가적으로 적합(predic) 또는 x_1, x_2, \dots, x_k 의 특정한 조합(combination)에서 두 개의 특별한 반응 결과 중의 하나인 확률을 추정할 필요가 있다. 기계학, 생물학 그리고 건강과학을 포함한 많은 분야에서 적용할 수 있다. 최근 몇 년 동안 무기연구(weapon research)에서 이런 종류의 모형이 광범위하게 사용되고 있다. 여기에서는 제시되는 것은 실험단위(experimental unit)에 대한 반응이 두 개의 범주(category) 중에 하나이거나 범주(category)가 회귀변수(regressor variable)의 수준(level)에 의존하는 어떠한 상황에서도 적용할 수 있다. 다음으로 수치 예제(numerical example)가 제시될 것이다.

반응변수(response variable)가 연속적(continuous)하지 않은 경우 분석의 어려움이 발생하는 것은 자연스러운 것이기 때문에 놀랄 필요없이 이런 상황에 주의해야한다. 문제점을 하나 하나 생각해보기 전에 이진반응모형이 의미하는 바가 무엇인지 해석(explain)할 필요가 있다. n 개의 회귀자료에서 다음의 모형을 고려하자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad \begin{cases} (i=1,2,\dots,n) \\ y_i = \{0,1\} \end{cases} \quad (7.36)$$

만약 일반적으로 $E(\varepsilon_i) = 0$ 을 가정하면, 다음과 같다.

$$E(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \quad (7.37)$$

지금부터 $y=1$ 일때 $x_{1i}, x_{2i}, \dots, x_{ki}$ 에서 관측치의 모집단의 비율로 $E(y_i) = P_i$ 를 볼 수 있다. 다른 말로,

$$\begin{cases} P_i = \text{Prob}(y_i = 1) \\ Q_i = 1 - P_i = \text{Prob}(y_i = 0) \end{cases} \quad (i=1,2,\dots,n) \quad (7.38)$$

따라서, n 개의 서로 다른 점에서 n 개의 확률 P_1, P_2, \dots, P_n 있으며 이것은 베르누이(Bernoulli) 분포의 모수이다.

만약 (7.36)에 있는 모형의 오차항(error term)을 고려한다면 ε_i 는 두 값만 가능하므로 연속형(continuous)일 수는 없다. 즉,

$$\varepsilon_i = y_i - [\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}]$$

$1 - P_i = Q_i$ 또는 $0 - P_i = -P_i$ 이 가능하다. 그 결과 모형오차에 정규성의 가정을 할 수 없고, 두번째 가정인 등분산(homogeneous variance)을 명백히 위반한다. $E(\varepsilon_i) = 0$ 이므로,

$$\begin{aligned}\text{Var}(\varepsilon_i) &= E(\varepsilon_i^2) \\ &= (Q_i)^2 \text{Prob}[y_i = 1] + (-P_i)^2 \text{Prob}[y_i = 0] \\ &= Q_i^2 P_i + P_i^2 Q_i \\ &= Q_i P_i (Q_i + P_i) \\ &= P_i Q_i\end{aligned}$$

P_i 는 회귀변수의 수준(level)에 따라 변하기 때문에 오차분산이 등분산이지 않다. (7.39)로부터

$$\text{Var}(\varepsilon_i) = (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}) \times (1 - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \dots - \beta_k x_{ki})$$

따라서, 최소제곱법을 사용하는 데 다음과 같은 2가지 어려움이 따른다.

1. ε_i 의 분포는 이산적이라 정규분포가 아니다.
2. 오차분산이 등분산(homogeneous variance)이 아니다.

두 가정을 만족하지 않는 것은 반응변수가 이진적(binary)이기 때문에 자연스럽다고 할 수 있다. 해결책(solution) 중 하나로 가중최소제곱법(weighted least squares)을 사용하여 식 (7.39)의 장점을 취하는 것이다.

7장 1절에서, 이분산(heterogeneous)의 경우 모수를 추정하기 위해서 가중최소제곱법(weighted least squares)에 대해서 알아보았다. 식 (7.39)는 i 번째 자료에서 오차분산이다. 그러나, P_i 와 $\text{Var}(\varepsilon_i)$ ($i = 1, 2, \dots, n$)는 현실적으로 알 수 있으므로, 최선은 자료로부터 추정하는 것이다. 계산과정은 다음과 같다.

1. 최소제곱법(ordinary least squares)을 이용하여 회귀함수를 추정한다.
2. 적합값(fitted value) $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ 을 이용하여 가중치(weight)를 추정한다. 즉.

$$\hat{w}_i = \frac{1}{\hat{\sigma}_{ei}^2} = \frac{1}{\hat{y}_i(1-\hat{y}_i)} \quad (i=1,2,\dots,n)$$

여기에서 $\hat{y}_i(1-\hat{y}_i)$ 는 i 번째 자료에서 오차분산(error variance) P_iQ_i 의 추정량을 나타낸다.

3. 식 (7.4)를 사용하여 모수를 재추정한다. 즉,

$$\beta^* = (X'V^{-1}X)^{-1}X'V^{-1}y$$

여기에서

$$V = \text{diag}[\hat{y}_1(1-\hat{y}_1), \hat{y}_2(1-\hat{y}_2), \dots, \hat{y}_n(1-\hat{y}_n)]$$

경우에 따라 2번째 과정을 반복적으로 사용할 수 있다. 즉, 2 단계에서 새로운 가중치(weight)을 얻기 위해 3 단계에서 얻은 추정값을 이용한다.

가중최소제곱(weighted least squares)이 반응변수가 이진적(binary)일 때 타당한 접근방법이지만, 더 많이(more popular) 사용하는 방법은 로지스틱 함수(logistic function)를 통해 회귀변수(regressor variable)에 대한 반응변수(response)의 평균(mean)을 모형화하는 것이다. 중요한 개념은 다음 부절(subsection)에서 알아보기로 한다.

로지스틱 회귀(Logistic Regression)

성공확률은 회귀변수들(regressor variables) x_1, x_2, \dots, x_k 의 함수로 가정한다. 다음과 같은 형태로 자료가 취한다고 가정하면,

$$\begin{matrix} n_1 & r_1 & x_{11} & x_{21} & \cdots & x_{k1} \\ n_2 & r_2 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ n_s & r_s & x_{1,s} & x_{2,s} & \cdots & x_{k,s} \end{matrix} \left. \right\} (s > k)$$

여기에서 r_1, r_2, \dots, r_s 는 n_1, n_2, \dots, n_s 번 시도 시 성공횟수를 나타낸다. s 개의 시행에서 회귀변수에 변화가 있다. 회귀변수는 다양한 약의 조합 용량(dosage of a combination of drugs)이라고 할 수 있으며, 성공은 특정 질병의 치료를 나타낸다. 공학(engineering)에서 적용하는 경우 회귀변수는 실패에 영향을 받는 적재형 변수(load type variable)이라 할 수 있다. i 번째 적재에 n_i 개의 시제품이 시험되었을 때 r_i 는 실패(또는 실패가 아닌) 수를 나타낸다. 뿐만 아니라, 사람이나 사람의 활동이 두개의 범주 중 한 범주로 속하는 사회과학에서 적용되는 것을 종종 볼 수 있다. 예를 들어, 반응변수는 개개인에 따른 광고쿠폰의 상품교환 여부이고, 회귀변수는 가격인하인 마케팅 자료에서 모형을 만들 수 있다. n_i ($i = 1, 2, \dots, s$)개의 표본에서 r_i 는 교환 받아간 수이다. 이 연구의 목적은 상품교환과 가격인하의 관계를 알아보기 위한 것이다.

반응변수가 이진적(binary)인 경우 가장 흔히 사용되고 유용한 모형은 로지스틱

회귀모형(logistic regression model)으로 다음과 같다.

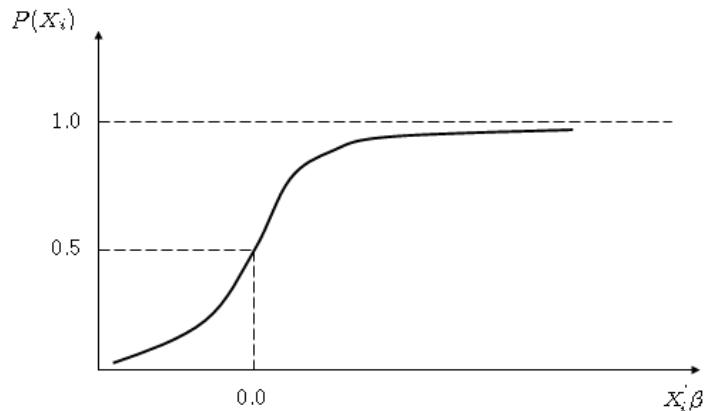
$$P(\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i^\top \beta}} \quad (i = 1, 2, \dots, s) \quad (7.40)$$

이 모형은 회귀변수과 발생 확률(probability of occurrence) $P(\mathbf{x}_i)$ 과 관계된다. 여기에서 다음의 양(quantity)은 다중선행회귀에 관계되는

$$\mathbf{x}_i^\top \beta = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$$

i 의 주어진 값에서 $n_i > 1$ 개의 관측치가 발생하면 r_i 개의 성공이 나타난다. 물론, 로지스틱 함수는 0과 1사이의 값을 가지며 그림 7.12에 나타나듯이 S-모양을 나타낸다.

FIGURE 7.12 A plot depicting the logistic function



우리는 결과적으로 식 (7.40) 모형에서 모수의 최대우도추정을 고려할 것이다. 그러나 로지스틱 모형(logistic model)의 선형화를 위해서 로짓변환(logit trasformation)의 모수의 최대우도추정을 고려하는 것이 타당할 것이다. 즉 (7.40)으로부터, 다음과 같이 쓸 수 있다.

$$\ln \left[\frac{P(\mathbf{x}_i)}{(1 - P(\mathbf{x}_i))} \right] = \mathbf{x}_i^\top \beta = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} \quad (7.41)$$

지금 각 조합에서의 정보는 확률값 $P(\mathbf{x}_i)$ 을 추정 가능하도록 한다. 예상했듯이 추정량은 표본에서의 성공률로 주어진다. 즉,

$$\hat{P}(\mathbf{x}_i) = \frac{r_i}{n_i} \quad (i=1,2,\dots,s)$$

그래서, (7.41) 모형에 위의 추정량을 넣고 아래에 대해 회귀분석을 실시한다.

$$\ln\left(\frac{\hat{P}(\mathbf{x}_i)}{1 - \hat{P}(\mathbf{x}_i)}\right) \text{ 대한 } x_1, x_2, \dots, x_k$$

따라서, 가정한 모형은 다음과 같다.

$$\boxed{\ln\left(\frac{\hat{P}(x_i)}{1 - \hat{P}(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (i=1,2,\dots,s)} \quad (7.42)$$

변환된 로지스틱 회귀모형(transformed logistic regression model) (7.42)는 최소제곱모형(least squares model)의 적용을 고려할 수 있다. 그러나 이전의 절에서처럼 오차분산(error variance)에 대하여 주의해야한다. 고정된 \mathbf{x}_i 의 조합에서

$$\text{Var} \ln\left(\frac{\hat{P}(\mathbf{x}_i)}{1 - \hat{P}(\mathbf{x}_i)}\right) \approx \frac{1}{n_i P(\mathbf{x}_i)(1 - P(\mathbf{x}_i))}$$

그 결과, 가중회귀모형은 고려해야할 접근법의 하나이다. 분산의 역수가 가중치(weight)이기 때문에 i 번째 자료에서 추정된 가중치는 다음과 같이 주어진다.

$$\boxed{w_i = n_i [\ln(\hat{P}(\mathbf{x}_i))(1 - \hat{P}(\mathbf{x}_i))] \quad (i=1,2,\dots,s)} \quad (7.43)$$

따라서, 회귀변수의 추정치은 아래에 주어진 값을 대각원소(diagonal element)로 하는 대각행렬(diagonal matrix) V 을 가지는 식 (7.4)에 의해 주어진다.

$$\frac{1}{n_i (\hat{P}(\mathbf{x}_i))(1 - \hat{P}(\mathbf{x}_i))}$$

로지스틱 회귀모형(logistic regression model)을 적합하는 데 앞에서 설명한 가중최소제곱법을 적용하는 것은 개개의 \mathbf{x}_i 의 값이 작지 않은 때 합리적이다. 하지만 두

가지 제약조건이 있다. 첫번째는, 이번 장 앞에서 지적하였듯이 가중회귀는 가중치(weight)가 상대적으로 적은 정보에 의해서 추정될 때는 피해야 한다. 두번째로는 아래의 분산은 단지 근사값에 불과하고 n_i 가 충분히 클 때 추정값이 가장 정확하다.

$$\ln\left(\frac{\hat{P}(\mathbf{x}_i)}{1 - \hat{P}(\mathbf{x}_i)}\right)$$

일단 최소제곱법(least squares)을 추정한 후, 예측(prediction)이나 두 결과 중 하나의 확률을 추정(estimation)하기 위해 로지스틱 회귀분석(logistic regression)을 사용하는 것은 매우 간단하다. 특별히 $x=x_0$ 에서 P_o (성공확률)을 추정하기 원한다면, 다음과 같이 계산하면 된다.

$$\hat{P}(\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i \beta^*}}$$

여기에서 β^* 은 식 (7.42)의 모형으로부터 가중최소제곱추정량(vector of weighted least squares estimator)이다.

예제 7.7 할인 판매 실적 자료

어느 구두방에서 불경기 기간에 매상고를 올리기 위하여 할인 판매를 실시하였다. 첫 주에 정가의 5%를 할인하였더니 찾아온 400명의 손님 중에서 64명이 구두를 샀고, 두 번째 주일에 10% 할인하였더니 찾아온 400명의 손님 중에서 102명이 구두를 샀다. 표는 5주간에 걸쳐 점차적으로 할인율을 높여 판매한 결과이다.

주, j	할인율(%), x_j	방문한손님수, n_j	구입한 손님수, r_j	구입손님비율, \bar{p}_j
1	5	400	64	0.160
2	10	400	102	0.255
3	15	400	140	0.350
4	20	400	206	0.515
5	30	400	296	0.740

로지스틱 회귀모형은 가정해보자. n_i 의 상대적으로 큰 값을 가지고 식 (7.42)의 모형을 고려할 수 있다. 따라서, 다음과 같은 모형을 가진다.

$$\ln\left(\frac{\hat{P}(x_i)}{1 - \hat{P}(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \quad (i=1,2,\dots,5)$$

가중회귀는 식 (7.43)에 주어진 가중치를 사용함으로써 적절해진다. 변환반응(transformed response), 할인율, 계산된 가중치(computed weight)는 다음과 같다.

$\ln(\hat{P}(x_i)/(1 - \hat{P}(x_i)))$	할인율(%) (x_i)	가중치(weight) (w_i)
-1.65822808	5	53.76
-1.07212067	10	75.99
-0.61903921	15	91.00
0.06001801	20	99.91
1.04596856	30	76.96

추정된 가중최소제곱회귀식은 다음과 같다.

$$\ln\left(\frac{\hat{P}(x_i)}{1 - \hat{P}(x_i)}\right) = -2.18602 + 0.10858x$$

그 결과 로지스틱 회귀함수는 다음과 같이 주어진다.

$$\hat{P}(x) = \frac{1}{1 + e^{-[-2.18602 + 0.10858x]}}$$

여기에서 \hat{P} 는 주어진 할인율 x_i 에서 판매률로 해석된다.

다음은 위의 예제에 대한 R code이다.

```

x <- c(5, 10, 15, 20, 30)
n <- c(rep(400, 5))
r <- c(64, 102, 140, 206, 296)
p <- c(0.160, 0.255, 0.350, 0.515, 0.740)
hp <- r/n
w <- n*hp*(1-hp)
p1 <- hp / (1-hp)
lnp <- log(p1)

```

```

fit <- lm(lnp~x)
summary(fit)

```

그룹자료의 최대 우도 추정(Maximum Likelihood Estimation for Grouped Data)

앞에서 지적했듯이, 로지스틱회귀분석(logistic regression)은 많은 과학분야에서 다루어지고 있다. 이러한 관심은 많은 소프트웨어 패키지(software package)가 로지스틱회귀분석(logistic regression)을 옵션(option)으로 포함하고 있다는 사실에서 알 수 있다. 예를 들어, SAS (1983)은 이 절의 뒷부분에서 언급될 PROC LOGIST를 제공한다. 더 일반적인 패키지인 GLIM(Baker와 Nelder(1978) 참조) 또한 로지스틱 회귀분석에 사용될 수 있다. (7.40)의 모형에서 회귀계수의 최대우도추정(maximum likelihood estimation)은 폐쇄형(closed form) 답(solution)을 주지 않지만, 계수를 구하기 위한 반복과정(iterative procedure)은 상대적으로 쉽다.

다시 한번 n_i 개체 또는 i 번째 그룹과 관계된 관찰치들을 가지고 식 (7.40)의 모형을 고려해보자. 반응변수는 이산형(성공 또는 실패)이고 i 번째 그룹에서 우도(likelihood)는 확률의 곱(product of probabilities)이거나 n_i 개의 표본에서 r_i 개의 성공이 발생할 결합확률(joint probability)이다. 그 결과 i 번째 그룹에서 우도(likelihood)는 다음과 같다.

$$[P(\mathbf{x}_i)]^{r_i} [1 - P(\mathbf{x}_i)]^{n_i - r_i} = \left[\frac{1}{1 + e^{-\mathbf{x}_i \beta}} \right]^{r_i} \left[\frac{e^{-\mathbf{x}_i \beta}}{1 + e^{-\mathbf{x}_i \beta}} \right]^{n_i - r_i}$$

$e^{-\mathbf{x}_i \beta}$ 를 η_i 로 두면, 전체 표본에서의 우도(likelihood)는 다음과 같다.

$$L(\beta, \mathbf{x}_i) = \prod_{i=1}^s \left[\frac{1}{1 + \eta_i} \right]^{r_i} \left[\frac{\eta_i}{1 + \eta_i} \right]^{n_i - r_i} \quad (7.44)$$

보통과 같이 β 의 모수에 대한 최대우도추정량(maximum likelihood estimates)을 찾기 위해 우도(likelihood)에 로그(log)를 취한다. 아래와 같다.

$$\ln L(\beta, \mathbf{x}_i) = \sum_{i=1}^s \{ r_i [-\ln(1 + \eta_i)] + [n_i - r_i] [\ln \eta_i - \ln(1 + \eta_i)] \}$$

단순화하면, 다음과 같다.

$$\begin{aligned}\ln L(\beta, \mathbf{x}_i) &= \sum_{i=1}^s \left\{ \eta_i [\ln \eta_i - \ln(1 + \eta_i)] - r_i [\ln \eta_i] \right\} \\ &= \sum_{i=1}^s \left\{ n_i \left[-\mathbf{x}_i' \beta - \ln(1 + e^{-\mathbf{x}_i' \beta}) \right] - r_i [\ln \eta_i] \right\}\end{aligned}$$

모수 β 의 추정하기 위해서 β 에 관한 도함수를 0으로 놓고 푼다.

$$\frac{\partial \ln L(\beta, \mathbf{x}_i)}{\partial \beta} = -\sum_{i=1}^s n_i \left[1 - \frac{e^{-\mathbf{x}_i' \beta}}{(1 + e^{-\mathbf{x}_i' \beta})} \right] \mathbf{x}_i + \sum_{i=1}^s r_i \mathbf{x}_i = 0$$

그 결과, 최대우도추정량은 아래의 식의 해답으로 $\hat{\beta}$ 주어진다.

$$\sum_{i=1}^s n_i \left[1 - \frac{e^{-\mathbf{x}_i' \hat{\beta}}}{(1 + e^{-\mathbf{x}_i' \hat{\beta}})} \right] \mathbf{x}_i = \sum_{i=1}^s r_i \mathbf{x}_i \quad (7.45)$$

방정식 (7.45), 만약 $\mathbf{x}_i' \beta$ 에 p 개의 모수가 포함되어 있다면 p 개 방정식은 β 에 관해 선형이 아니며 직접 풀 수 없다. 그러나 추정량을 얻기 위한 많은 반복방법(iterative method)이 있고 따라서 회귀변수의 함수로써 확률을 추정할 수 있다. 식 (7.45)는 자료가 그룹으로 되어있을 때 풀 수 있다. 즉, 예제 (7.7) 경우처럼 회귀변수의 각 조합에서, 몇몇 개체나 실험단위가 관측된 경우이다. 그러나 계획된 실험에서 하지 않는 관찰연구(observational study)에서는 자료를 그룹으로 얻을 수 없다. 결과적으로 이런 두 가지 형태의 예제를 가지고 설명해야하며, 그룹으로 관측되지 않은 자료의 우도(likelihood)를 보여주는 것도 중요하다.

비그룹 경우 우도함수(The Likelihood Function for the ungrouped Case)

비그룹 자료의 경우, 관찰치 n 개 중 n_1 개는 성공으로 $n - n_1$ 개는 실패로 가정하자. 그러나 필연적으로 회귀변수가 동일한 값을 가지는 실험단위의 그룹은 필연적으로 없다. 처음 n_1 개의 관찰치를 성공으로 하면, 우도함수(likelihood function)은 다음과 같다.

$$\begin{aligned}L(\beta, \mathbf{x}_i) &= \prod_{i=1}^n \left(\frac{1}{1 + \eta_i} \right) \prod_{i=n+1}^n \left[1 - \left(\frac{1}{1 + \eta_i} \right) \right] \\ &= \prod_{i=1}^n \left(\frac{1}{1 + \eta_i} \right) \prod_{i=n+1}^n \left(\frac{\eta_i}{1 + \eta_i} \right) \\ &= \prod_{i=1}^n \left(\frac{1}{1 + e^{-\mathbf{x}_i' \beta}} \right) \prod_{i=n+1}^n \left(\frac{e^{-\mathbf{x}_i' \beta}}{1 + e^{-\mathbf{x}_i' \beta}} \right)\end{aligned}$$

위의 우도함수는 간단한 대수(algebra)를 이용하면 아래와 같이 간단한 형태로 쓸 수 있다.

$$L(\beta, \mathbf{x}_i) = \frac{\prod_{i=1}^n (e^{\mathbf{x}_i \cdot \beta})}{\prod_{i=1}^n (1 + e^{\mathbf{x}_i \cdot \beta})} \quad (7.46)$$

다시 식 (7.46)을 β 에 대해 최대화하는 것은 그룹 자료의 경우 β 에 대해 비선형함수를 최대화하는 것과 관련되어 있다.

계수의 추론에 대한 기본으로의 우도비(Likelihood Ratio as a Basis for Inference on Coefficients)

몇몇의 상용화 패키지(commercial package)는 로지스틱 회귀모형에 대한 β 의 추정량 뿐만 아니라 로지스틱 회귀모형에서 변수의 역할(role of variables)에 관한 통계적 추론(inference)에 대한 정보를 얻을 수 있다. 추가로, 회귀계수의 점근적 분산-공분산 행렬의 추정치도 얻을 수 있다. 이것으로 회귀계수의 표준오차를 사용할 수 있게 해준다. 이 절에서는 그룹 자료와 비그룹 자료의 두 가지 예제를 가지고 이런 방법에 대해서 알아보기로 한다.

로지스틱 모형의 적합결여 측도(measure of lack of fit)는 최대우도추정의 특징을 이용하여 결정할 수 있다. 앞 절의 Box-Cox 변환 과정에서 λ 의 신뢰구간을 구하기 위해서 변환 모수(transformation parameter) λ 의 최대우도추정을 사용하였다. 이 신뢰구간은 우도비 기준(likelihood ratio criterion)을 기초로 한다. 유사한 개념이 최대우도(maximum likelihood)를 사용하는 경우 적합성 결여(lack of fit)와 가설검정(test of hypothesis)을 가능하도록 한다. 이러한 분석의 목적으로 비그룹 자료의 경우를 고려해야 한다. 로지스틱 회귀모형 적합의 질적측도(measure of quality of logistic regression fit)는 식 (7.46)의 최대우도(maximum likelihood)와 비그룹 경우 성공-실패 자료를 완벽하게 적합(perfect fit)할 때의 최대값과 비교하기 위한 것이다. 예측한대로 복합한 모형을 적합 할수록 우도(likelihood)는 커진다. 이것은 2, 3와 4장에서 일반적인 회귀분석(standard regression context)에서 SS_{Res} 가 최소화되는 것과 동일하다. 가장 복잡한 모형, 즉 모수가 가장 많이 포함하는 경우, 가장 큰 우도는 간단히 다음과 같다.

$$y_i = P_i + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

여기서 y_i 는 반응변수의 관찰치로 기본으로 0 또는 1(실패 또는 성공)이고 P_i 는 성공확률이다. 그 결과, 회귀변수는 무시되고 예상한 대로 n 개의 모수 P_1, P_2, \dots, P_n 을 추정해야하고 우도(likelihood)는 다음과 같다.

$$L(\mathbf{P}) = \prod_{i=1}^n P_i^{y_i} (1 - P_i)^{1-y_i}$$

여기에서 \mathbf{P} 에 의해서 성공의 확률 벡터로 둔다. 이 모형은 본질적으로 회귀변수(regressor)와 확률의 추정(estimate of probability)은 어떠한 관련도 없다. 우도함수(likelihood function)에 로그(log)를 취하면 아래와 같은 식을 얻을 수 있다.

$$\ln L(\mathbf{P}) = \sum_{i=1}^n [y_i \ln P_i + (1 - y_i) \ln(1 - P_i)]$$

(7.47)

비록 식 (7.47)은 약간 까다로우나, 만약 특정한 확률모수 P_i 를 고려한다면 우도에 대한 기여를 자세히 조사하면 다음과 같음을 알 수 있다.

$$y_i \ln P_i + (1 - y_i) \ln(1 - P_i)$$

그리고 만약 $y_i = 0$ 이라면 로그우도(log likelihood)는 $P_i = 0$ 일 때 가장 크고, 만약 $y_i = 1$ (성공)이라면 $P_i = 1.0$ 일 때 최대화된다. 그 결과 최대우도추정량은 다음과 같이 예측할 수 있다. 즉,

$$\begin{aligned}\hat{P}_i &= 1.0 && \text{if } y_i = 1.0 \\ &= 0 && \text{if } y_i = 0\end{aligned}$$

이 모형을 잔차(residual) $y_i - \hat{y}_i$ 가 0일 경우인 과대적합모형(overfitted model)으로 볼 수 있다.

이 모형은 만약 회귀변수라 어떤 영향도 주지 못하는 의미없는 모양인 반면, 회귀변수가 전체적으로 효과적인지 결정하는 완벽한 적합(perfect fit)의 기본을 제공한다. 따라서, 적합결여검정(lack of fit test)은 로지스틱 회귀모형이 완전적합모형(perfect fit model)과 유의한 차이가 있는지를 제공할 것이다.

지금 가장 큰 우도를 가지는 두개의 우도가 있고 다음과 같다.

$$L(\bar{\mathbf{P}}) = \prod_{i=1}^n (y_i)^{y_i} (1 - y_i)^{1-y_i}$$

(7.48)

그리고

$$L(\hat{\beta}) = \frac{\prod_{i=1}^n (e^{x_i \beta})}{\prod_{i=1}^n (1 + e^{x_i \beta})} \quad (7.49)$$

여기에서 $\hat{\beta}$ 는 계산된 반복과정(computerized iterative routine)을 통해 식 (7.49)를 최대화한 최대우도추정량(maximum likelihood estimator)이다. 여기서, 우도비(likelihood ratio)의 로그(log)는 로지스틱 모형에 대한 적합도(goodness of fit)의 중요한 측도(measure)이다.

$$\lambda(\beta) = -2 \ln \left[\frac{L(\hat{\beta})}{L(\bar{P})} \right] \quad (7.50)$$

위의 통계량(statistic)을 우도비 통계량(likelihood ratio statistic)이라 불리고 이것은 현재의 로지스틱모형(logistic model)이 더 완벽한 모형인 $y_i = P(x_i) + \varepsilon_i$ 에 의해서 적합된 결과와 비교하여 자료의 특징을 잘 나타내는지를 결정하는 데 사용된다. 만약 $L(\hat{\beta})/L(\bar{P})$ 의 비가 충분히 1.0에 가깝다면 $L(\hat{\beta})$ 가 $L(\bar{P})$ 보다 유의적으로 작지 않기 때문이고, 현재의 로지스틱 모형이 유의적인 적합결여(lack of fit)가 없는 것이다. 식 (7.50)의 검정통계량(test statistic) $\lambda(\beta)$ 는 적합된 로지스틱 회귀(fitted logistic regression)와 관련된 편차(deviance)라 불린다. 만약 아래와 같은 가정을 구상한다면,

$$H_0 : P(x_i) = \frac{1}{1 + e^{-x_i \beta}}$$

적절한 검정의 정보는 $\lambda(\beta)$ 에 있다. 사실상 H_0 하에 $\lambda(\beta)$ 는 $n-k-1$ 의 자유도를 가지는 카이제곱(χ^2)분포를 따르는 확률변수이다. 따라서, $\lambda(\beta)$ 가 충분히 0에 가까워 적합결여가 유의하지 않으면 H_0 는 기각되지 않는다. 그 결과 카이제곱분포의 상위 꼬리(upper tail)에 해당되는 값을 가지면 H_0 는 기각된다. 즉

Reject H_0 at the level of significant α if $\lambda(\beta) > \chi^2_{\alpha, n-k-1}$

식 (7.50)을 연구하고, 그것과 Box-Cox과정에서 λ 의 신뢰구간을 만들어내는 것과 연결하는 것을 그려보자.

둘 다 아래와 같은 일반적인 형태를 보인다.

In $L(\text{unrestricted}) - \ln L(\text{restricted})$

여기에서 무제약(unrestricted)의 의미는 모든 모수를 최대우도추정량을 대치하는 것이다.(또는, 제약된 (the restricted) 경우보다 더 많은 모수에 대하여). 식 (7.50)에 사용된 “2”는 Box-Cox 과정에서 카이제곱분포의 백분위수와 대치된다는 것에 주목할 필요가 있다.

그룹 자료의 경우 개념은 동일하다. 그러나 가장 완전한 모형(most complete model)에서 최대우도추정량(maximum likelihood estimator)은 $\hat{P}_i = r_i / n_i$ 이고 따라서,

$$L(\hat{P}) = \prod_{i=1}^s \left(\frac{r_i}{n_i} \right)^{r_i} \left(\frac{n_i - r_i}{n_i} \right)^{(n_i - r_i)}$$

그리고

$$L(\hat{\beta}) = \prod_{i=1}^s \left[\frac{1}{1 + e^{-\mathbf{x}_i^\top \beta}} \right]^{r_i} \left[\frac{e^{-\mathbf{x}_i^\top \beta}}{1 + e^{-\mathbf{x}_i^\top \beta}} \right]^{n_i - r_i}$$

여기에서 $\hat{\beta}$ 는 그룹 경우의 최대우도추정량(maximum likelihood estimator)을 포함한다. 통계량 $\lambda(\beta)$ 는 자유도 $s-k-1$ 갖는 카이제곱(χ^2) 분포를 따른다.

앞에서 봤듯이, 로지스틱 모형은 편차(deviance)가 충분히 크다면 부적절하는 것을 알 수 있다. 이런 의미에서, 편차(deviance)는 일반적인 회귀모형(ordinary regression)의 잔차제곱합(residual sum of squares)과 다른 역할을 하는 것은 아니다(Exercise 7.9. 참조). 이러한 편차를 사용하면 다른 형태의 가설검정에 확장할 수 있다. 사실상 편차의 의미는 보통 회귀분석의 경우(ordinary regression case)에서 잔차제곱합(residual sum of squares)과 동일하다. 어떤 회귀변수나 회귀변수이 부분집합에 대한 추론은 각 회귀변수가 얼마나 많이 편차의 감소에 영향을 미치는지를 결정함으로써 계산될 수 있다. 예를 들어, k 개의 회귀변수가 있는 경우 j 번째 회귀변수 x_j 의 기여도를 생각해보자. 다음과 같이 정의한다.

$$\lambda(\beta_j | \beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k) = \lambda(\beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k) - \lambda(\beta)$$

이 식의 용어(notation)는 3장의 $R(\cdot)$ 용어와 많이 일치한다(consistent). $\lambda(\beta)$ 는 모든 모수를 포함한 모형에서 계산된 편차(deviance)이고, 반면 $\lambda(\beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k)$ 는 $\beta_j x_j$ 이 포함되지 않은 모형에서 계산된 편차(deviance)이다. $\beta_j x_j$ 에 의해 감소된 편차의 양은, 다른 회귀변수에 의해 조정된, 다음과 같이 계산된다.

$$\lambda(\beta_j | \beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k) = -2 \ln \left[\frac{L(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_k)}{L(\hat{\beta})} \right] \quad (7.51)$$

식 (7.51)의 분자 우도(numerator likelihood)는 $\beta_j x_j$ 가 제거된 최대우도이다. 이 중요한 결과를 증명하기 위해서 아래를 고려할 필요가 있다.

$$\begin{aligned} \lambda(\beta_j | \beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k) &= -2 \ln \left[\frac{L(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_k)}{L(\hat{\beta})} \right] + 2 \ln \left[\frac{L(\hat{\beta})}{L(\bar{\beta})} \right] \\ &= -2 \ln \left[\frac{L(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \dots, \hat{\beta}_k)}{L(\bar{\beta})} \right] \end{aligned}$$

식 (7.51)의 좌측 편은 편차(deviance)의 차이를 표현한 것으로 볼 수 있고, 이것은 설명된 변동의 차이로 볼 수 있다. 설명된 변동은 완전묘형과 축소된 모형간의 $2\ln L$ 의 차이로 표현한 것이다. 로그 우도비 통계량(log likelihood ratio statistic)은

$$\lambda(\beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k)$$

β_j 에 대한 적절한 가정하에 자유도 1을 가지는 근사적 카이제곱(χ^2)분포를 따른다. 이것은 $\lambda(\beta)$ 의 개념과 일치한다. 즉, 자유도는 로그 우도비(log likelihood)의 분모와 분자에서 추정되는 모수의 개수의 차이이다.

위의 공식은 개개의 회귀계수에 대한 부분 카이제곱(χ^2)검정이 가능하게 한다.
즉, 검정을 위해서

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

만약 $\lambda(\beta_j | \beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k)$ 이 적절한 카이제곱분포의 백분위수를 초과하면 H_0 를 기각한다.

즉, 만약 $\lambda(\beta_j | \beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k) > \chi^2_{\alpha,1}$ 이면 H_0 를 기각한다.

동일한 카이제곱(χ^2)과정으로 회귀계수의 부분집합에 대한 검정을 할 수 있다. 다시,

회귀계수의 부분집합 β_1 의 중요성은, 여기에서 $\beta' = [\beta'_1, \beta'_2]$, β_2 가 있는 모형에서 β_1 가 모형에 들어갈 때 얼마나 편차가 감소하는가에 의존한다. β_1 이 $r < k + 1$ 개의 모수를 포함한다고 가정하자. 아래의 가설에서

$$\begin{aligned} H_0 : \beta_j &= 0 \\ H_1 : \beta_j &\neq 0 \end{aligned}$$

만약 $\lambda(\beta_1 | \beta_2) > \chi^2_{r,\alpha}$ 이면, 가설의 H_0 를 기각한다.

여기에서 물론

$$\begin{aligned} \lambda(\beta_1 | \beta_2) &= \lambda(\beta_2) - \lambda(\beta) \\ &= -2 \ln \left[\frac{L(\hat{\beta}_2)}{L(\hat{\beta})} \right] \\ &= 2 \ln L(\hat{\beta}) - 2 \ln L(\hat{\beta}_2) \end{aligned}$$

우도비 과정의 용어(likelihood ratio procedure)는 로지스틱 회귀모형에서 중요한 모형 항을 결정하는 데 큰 도움이 될 수 있다. 사실, 4장에서 다룬 단계적(stepwise) 변수선택 형태의 과정을 아주 많은 회귀변수가 있는 상황에서 사용할 수 있다. (SAS (1983) 참조)

계수의 표준오차(Standard Errors of Coefficients)

점근이론(Asymptotic theory)을 사용하여 회귀계수의 근사 분산과 공분산(approximate variance and covariance)을 알 수 있다. 분산과 공분산의 추정치는 로그우도(log likelihood)를 2번 미분한 행렬을 통해 얻을 수 있다. 이것은 벡터(vector) β 에서 개별 모수에 대한 가설검정 시 계수의 표준오차(standard error of coefficient)로 사용된다. 따라서, 일반적인 회귀모형에서의 t 검정과 동일한 형태를 보인다. 일반적인 선형회귀(standard linear regression)를 사용하지 않으므로 $X'X$ 행렬이 중요한 역할을 하지 않고 대신 아래의 도함수(the derivative)로부터 C 행렬을 만든다.

$$c_{ii} = \frac{-\partial^2 \ln L(\hat{\beta})}{\partial \hat{\beta}_i^2} \quad (i = 1, 2, \dots, k+1)$$

그리고

$$c_{ij} = \frac{-\partial^2 \ln L(\hat{\beta})}{\partial \hat{\beta}_i \partial \hat{\beta}_j} \quad (i \neq j)$$

회귀계수의 분산-공분산행렬의 추정치(estimate of variance-covariance)는 행렬 C^{-1} 이다. 그 결과,

아래의 가설검정을 위한 근사적 카이제곱(χ^2) 통계량(statistic)은 다음과 같다.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

$$\chi^2 = \frac{b_j^2}{c_{jj}} \quad (j = 0, 1, \dots, k)$$

여기에서 c^{jj} 는 C^{-1} 의 대각원소(diagonal elements)이다. 계수의 표준오차는 c^{jj} 의 제곱근(square roots)이다. 그러므로 개별 회귀계수의 검정에는 두 가지 방법이 있으며, 하나는 편차(deviance)를 이용하는 것이고 다른 하나는 근사 분산-공분산 행렬(approximate variance-covariance matrix)를 이용하는 것이다. 둘 다 관련된 분포는 자유도가 1인 카이제곱분포이다. 그러나 수치결과가 동일하지는 않다. 점근적으로는 둘 다 옳다(correct). 추정된 분산-공분산 행렬에 관한 점근이론에 관한 더 많은 정보는 McGullagh와 Nelder (1983)을 참고하라.

로지스틱 모형의 성능 측도(Measures of Performance of the Logistic Model)

이전 부절(section)에서 알아본 가설검정에 덧붙여서, 최대우도법(maximum likelihood)을 이용해 모수추정을 한 경우 적당한 수의 성능 측도를 고려하여 로지스틱 회귀모형을 설정해야한다. 쉬우면서 R^2 과 유사한 통계량이 있으면 로그우도(log likelihood)를 고려함으로써 확실히 사용 가능할 것이다. 만약 로지스틱 회귀모형(logistic regression model)이 모수(parameter) $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ 을 가졌다면, 상수 β_0 가 있을 때 x_1, x_2, \dots, x_k 에 의해서 설명되는 회귀 제곱합은 다음의 자유도 k 인 편차(deviance)와 동일하다.

$$\begin{aligned} \lambda(\beta_1, \beta_2, \dots, \beta_k | \beta_0) &= \lambda(\beta) - \lambda(\beta_1, \beta_2, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k) \\ &= 2 \ln L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k) - 2 \ln L(\hat{\beta}_0) \end{aligned}$$

총 제곱합(total sum of squares)과 동일한 것은 상수항 β_0 만 포함한 로지스틱 모형(logistic model)을 적합 시 얻을 수 있는 편차(deviance)이다. 그 결과, 타당한 설명된 χ^2 의 비율(proportion)은 다음과 같다.

$$R^2 = \frac{\lambda(\beta_1, \beta_2, \dots, \beta_K | \beta_0)}{\lambda(\beta_0)}$$
(7.52)

수정된(adjusted) R^2 의 형태는 C_p 통계량과 유사한 통계량의 근(roots)으로 구해진다. 이 값은

만약 회귀변수를 추가하였을 때 편차(deviance)가 충분히 감소하지 않거나 설명된 χ^2 이 증가하지 않는 경우 감소하는 경향을 보인다. 이 통계량(statistic)은 다음과 같다.

$$\boxed{\bar{R}^2 = \frac{\lambda(\beta_1, \beta_2, \dots, \beta_k | \beta_0) - 2p}{\lambda(\beta_0)}} \quad (7.53)$$

여기에서 $p=k+1$ 는 모형 모수(model parameter)의 수이다. 여기에서 만약 $\lambda(\beta_1, \beta_2, \dots, \beta_k | \beta_0)$ 가 $2p$ 보다 작다면 \bar{R}^2 은 0에 고정된다. 만약 변수가 모형에 추가되어 $\lambda(\beta_1, \beta_2, \dots, \beta_k | \beta_0)$ 가 증가하면 \bar{R}^2 는 감소하고, C_p 통계량과 유사한 통계량은 증가한다(4장 참조).

다른 다소 독창적인 성능측도(performance measure)와 심지어 이상점(outliers과 영향력 진단(influence diagnostics)도 있다. 그러나, 로지스틱 회귀(logistic regression) 추론에 관한 더 자세한 내용을 알기 원한다면 Pregibon (1981), Atkinson (1982), Cox (1970)와 Walker and Duncan (1967)를 참조하라.

두가지 예제가 나오고 여기에서 연구된 많은 방법에 대해 설명할 것이다. 첫번째 예제는 그룹 자료이고, 두번째 예제는 관찰 연구에 관련된 것이다. 여기에서 최대우도(maximum likelihood)를 사용하고 로지스틱 모형(logistic model)을 위한 적합결여검정(lack of fit test)을 설명하도록 한다.

예제 7.8 진폐증 자료(Pneumoconiosis Data)

표7.4에 있는 자료를 고려해보자. 이 자료는 진폐증을 앓고 있는 광부들에 대해 해마다 발생하는 진폐증 환자에 대해 심각한 경우의 수를 조사한 것이다. 여기에서 흥미분야인 반응변수 y 는 심각한 증상에 대한 비율이다. 이때 반응확률모형은 이항변수이기 때문에 자료에는 로지스틱 회귀모형을 적합해야한다. 여기서 자료는 그룹화되어 있기 때문에 최대 우도법으로 계수를 추정해야 한다. 이때 추정 결과는 표에 나타나 있다. 표에서 회귀 계수는 모두 0과 확연하게 다르다는 것을 알 수 있다. R^2 유사 통계량은 다음과 같이 계산될 수 있다.

$$\begin{aligned} R^2 &= \frac{\lambda(\beta_1 | \beta_0)}{\lambda(\beta_0)} = 1 - \frac{\lambda(\beta_0, \beta_1)}{\lambda(\beta_0)} \\ &= 1 - \frac{6.0508}{56.9028} = 0.8937 \end{aligned}$$

지금 R^2 과 유사한 통계량은 모형이 타당하다는 것을 의미하는 것처럼 보이고,

적합결여검정에서는

$$\begin{aligned}\lambda(\beta_0, \beta_1) &= -2 \ln \left[\frac{L(\beta_0, \beta_1)}{L(\hat{p})} \right] \quad (8-2 = 6 \text{ df}) \\ &= 6.0508\end{aligned}$$

상당히 유의하다 ($P < 0.001$). 그러므로 우리는 회귀함수에서 모형의 선형부분의 변화로부터 도움을 받을 수 있다. 다른 것으로 고려해보자.

$$P(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \ln x_i)}}$$

표 7.6은 분석결과를 보여준다.

TABLE 7.4 진폐증 자료(Pneumoconiosis Data)

Number of Years Of Exposure	Number of Severe Cases	Total Number Of Miners	Proportion of Severe Cases, y
5.8	0	98	0
15.0	1	54	0.0185
21.5	3	43	0.0698
27.5	8	48	0.1667
33.5	9	51	0.1765
39.5	8	38	0.2105
46.0	10	28	0.3571
51.5	5	11	0.4545

TABLE 7.5 최대우도추정량을 이용한 진폐증자료 분석결과
(Analysis of Pneumoconiosis Data using maximum likelihood estimation)

추정치(estimate)	표준오차(Standard Error)	χ^2	P
$\beta_0 = -4.79648$	0.56859	-8.436	< 2e-16
$\beta_1 = 0.09346$	0.01543	6.059	1.37e-09
Model $\chi^2 = \lambda(\beta_0) - \lambda(\beta_0, \beta_1)$ $= 50.852$ (1 df)		$\lambda(\beta_0, \beta_1) = 6.0508$ $\lambda(\beta_0) = 56.9028$	(6 df) (7 df)

TABLE 7.6 로그변환 후 분석결과

추정치(Estimate)	표준오차(Standard Error)	χ^2	P
$\beta_0 = -10.932$	1.896	-5.766	8.14e-09
$\beta_1 = 2.693$	0.534	5.044	4.56e-07
Model $\chi^2 = \lambda(\beta_0) - \lambda(\beta_0, \beta_1)$		$\lambda(\beta_0, \beta_1) = 2.1475$	(6 df)
$= 54.7553$ (1 df)		$\lambda(\beta_0) = 56.9028$	(7 df)

표 7.6에 있는 계수 β_1 는 $\ln x$ 의 계수이다. 이 경우에 설명된 분산의 부분은

$$R^2 = 1 - \frac{2.1475}{56.9028} = 0.9622602$$

단지 노출 년도수(x)에 대한 단순 선형회귀로 적합하는 것보다 높으며 이 때 2.1475의 추가 편차(6 df)는 중요치 않다. 대부분 편차가 작을수록 로그변환을 포함한 모형은 선호한다. 더 많은 증거를 두 개의 모형을 사용한 적합한 값의 표로부터 관찰할 수 있다.

Linear Load		Log Load	
Observed	Fitted	Observed	Fitted
0	0.01400281	0	0.002030434
1	0.03246665	1	0.025627428
3	0.05802922	3	0.064860114
8	0.09741786	8	0.118626259
9	0.15902872	9	0.186348085
8	0.24886118	8	0.263055390
10	0.37820213	10	0.349822753
5	0.50421483	5	0.421745159

로그 모형이 추정부분에서 더 좋다는 것은 명백하다. 로지스틱 모형에서 기하부분에서 보이는 경쟁 모형을 비교하였을 때 계산된 편차가 더 유익하다.

다음은 위의 예제에 대한 R code이다.

```

x <- c(5.8, 15.0, 21.5, 27.5, 33.5, 39.5, 46.0, 51.5)
f <- c(0, 1, 3, 8, 9, 8, 10, 5)
n <- c(98, 54, 43, 48, 51, 38, 28, 11)
nf<-n-f
y<-cbind(f,nf)

```

```

dat7_8<-data.frame(x,y)
fit <- glm(y~x,data=dat7_8,family = binomial(logit))
summary(fit)
lnx <- log(x)
dat7_8_1 <- data.frame(lnx, y)
fit1 <- glm(y~lnx, data=dat7_8_1, family=binomial(logit))
summary(fit1)

```

다음 예에서 자료가 그룹되지 않았을 경우에 대한 설명을 한다. 사실상 사용자는 회귀변수를 조절할 수 있는 능력이 없다.

예제 7.9 자가 소유 자료(Home Ownership Status Data)

자료는 표 7.7에 있다. 로지스틱 회귀는 모수를 추정하기 위해서 사용되는 최대우도방법을 가지고 자료를 적합화 시킨다. 모형은 다음과 같다.

$$P(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

여기에서 x_i 는 수입이 y 는 자가 소유 상태이다. 표 7.8은 자료를 분석결과로 적합결여검정, 모형에 대한 χ^2 통계 및 회귀변수들에 대한 χ^2 통계이다. 이 목표는 로지스틱 모형의 적절성을 결정하고 회귀변수의 유의성 검정을 하는 것이다. 이전의 예에서처럼, R 프로그램이 사용되었다. 모형 편차, 즉 적합 검정 결여를 유발하는 통계는 χ^2 값으로 22.4350이고 자유도는 18이다. P-값은 1.25이다. 그러므로 로지스틱 모형은 상당히 적당해 보인다.

TABLE 7.7 자가 소유 자료(home Ownership Status Data)

Household	Income	Home Ownership Status	Household	Income	Home Ownership Status
1	38,000	0	11	38,700	1
2	51,200	1	12	40,100	0
3	39,600	0	13	49,500	1
4	43,400	1	14	38,000	0
5	47,700	0	15	42,000	1
6	53,000	0	16	54,000	1
7	41,500	1	17	51,700	1
8	40,800	0	18	39,400	0
9	45,400	1	19	40,900	0

10	52,400	1	20	52,800	1
----	--------	---	----	--------	---

TABLE 7.8 자가 소유 자료(home Ownership Status Data) 분석결과

Coefficient	Standard Error	χ^2	P
$\beta_0 = -8.7395139$	4.4394326	-1.969	0.0490
$\beta_1 = 0.0002009$	0.0001006	1.998	0.0458
Model $\chi^2 = \lambda(\beta_0) - \lambda(\beta_0, \beta_1)$			
$= 5.091 \quad (1 \text{ df})$			
$P < 0.0001$			
$\lambda(\beta_0, \beta_1) = 22.435 \quad (18 \text{ df})$			

아래의 내용은 위 예제에 대한 R code이다.

```
x <- c(38000, 51200, 39600, 43400, 47700, 53000, 41500, 40800, 45400, 52400,
      38700, 40100, 49500, 38000, 42000, 54000, 51700, 39400, 40900, 52800)

y <- c(0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1)

dat7_9<- data.frame(y,x)

fit <- glm(y~x,data=dat7_9,family = binomial(logit))

summary(fit)
```

이 다음 절에서는 이산형 반응변수를 가진 회귀모형에 대한 용어를 좀 더 고려할 것이고, 사실상 반응변수가 포아송 분포를 따를 때 회귀모형에 대한 분석방법을 알아보도록 하겠다.

7.5. 이산형 반응변수를 가진 모형개발, 포아송 회귀(Further Development in Models with a Discrete Response, Poisson Regression)

이전 절의 전개는 반응변수가 이진적(binary)인 경우에 관계되었다. 응용분야는 의학, 경제학, 기계학, 생물학 등 풍부하다. 그러나 많은 응용분야에서 반응변수는 이산적(discrete)이지만 반드시 이진적(binary)이지는 않다. 갯수(number of count)가 하나이상의 회귀변수(regressor)의 함수라 가정하자. 예를 들어, 생물학적 실험에서 특별한 식물에 있는 곤충들의 수는 식물의 어떤 생물학적 특성에 대한 함수이다. 금속학분야에서 결함(거품 등)의 개수를 모형하는 것이 중요하다. 많은 응용분야는 의학과 공업분야이다. 이 절에서는 발생(incidents)의 평균 개수를 포아송 분포의 모수로 가정할 것이다. 따라서, 이 포아송 평균은 회귀 변수의 함수가 된다.

포아송 분포(The Poisson Distribution)

포아송 회귀는 물론 포아송 분포의 사용을 기초로 한다. 포아송 분포 모형은 사건(event or incidents) y 에 대한 확률을 모형하는데 이는 확률을 가진 포아송 과정에 의해 다음과 같다.

$$p(y; \mu) = \frac{e^{-\mu} \mu^y}{y!} \quad (y = 0, 1, 2, \dots)$$

포아송 분포의 평균은 μ 이다. 모수(parameter) μ 는 어떤 특정 단위, 시간, 거리, 면적, 용량 등에 매우 의존한다(dependent). 이 분포는 선택된 기본기간(시간)(basic period)로 동안 발생하는 상대적으로 드문 사건을 모형하는데 사용된다. 예를 들어, 만약 μ 가 단위시간당(per unit time) 평균 발생(incidents)이고 t 가 관심을 가지고 있는 시간이라면, 평균 y 는 μt 이다. 그러므로 t 시간동안 사건 y 의 확률은 다음과 같다.

$$p(y; \mu) = \frac{e^{-\mu t} (\mu t)^y}{y!}$$

y 사건들에 대한 확률과 사건의 평균 개수 μt 를 이렇게 표현하는 것은 단위 시간당 사건의 평균 개수가 일정하다는 가정에서 이루어 진다. 그러나 사건들의 평균 개수는 자료 수집 과정에서 변화하는 회귀변수의 정도에 따라 결정된다.

포아송 회귀 모형(Model for Poisson Regression)

7장 4절에서 로지스틱 회귀(logistic regression)의 경우와 같이 포아송 회귀 모형은 본질적으로 회귀변수(regressor)의 함수로 이산형 분포(로지스틱의 경우 이항분포이다.)의 평균을

나타낸다. 자료의 형태는 다음과 같다고 가정하자.

$$\begin{array}{cccccc} y_1 & x_{11} & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{1n} & x_{2n} & \cdots & x_{kn} \end{array}$$

모형은 아래와 같다.

$$y_i = \mu_i + \varepsilon_i \quad (i = 1, 2, \dots, n)$$

지금 μ_i 는 시간 t_i 동안 사건(incident)의 평균 횟수이다. 포아송 분포를 사용하고 μ_i 는 자료의 측정시 마다 독립적으로 변하지는 않는다고 가정한다. k 개의 회귀변수의 함수로써 μ_i 를 모형화하면 포아송분포는 다음과 같다.

$$p(y_i; \beta) = \frac{e^{-t_i[\mu(\mathbf{x}_i, \beta)]} [t_i \mu(\mathbf{x}_i, \beta)]^{y^i}}{y_i!} \quad (y = 0, 1, 2, \dots) \quad (7.54)$$

여기에서 $\mu(\mathbf{x}_i, \beta)$ 는 포아송 평균(Poisson mean)이며 일반적인 포아송(standard Poisson) 경우 μ 로 대치할 수 있다. 벡터 β 는 추정될 모수의 집합을 나타낸다. 함수 $\mu(\mathbf{x}_i, \beta)$ 는 사용자가 선택할 수 있다. 이 함수는 항상 비음수(nonnegative)의 형태여야 한다. 물론 $e^{\mathbf{x}_i \beta}$ 가 후보이며 여기에서 $\mathbf{x}_i \beta$ 는 선형함수이다. 또 다른 함수로 $\ln \mathbf{x}_i \beta$ 이며 여기서 $\mathbf{x}_i \beta > 1$ 이다. 물론 다른 함수로 선형함수 그 자체 즉, $\mathbf{x}_i \beta$ 도 가능하다. 여기서 $\mathbf{x}_i \beta > 0$ 이다. 회귀변수와 분포평균과의 관계를 나타내는 이런 함수를 연결함수(link function)이라 부른다. 예를 들어, 로지스틱 회귀의 경우, 연결함수는 식 (7.40)에 보여진 로지스틱(logistic)함수이다. 따라서, 아래와 같이 포아송 평균에 대한 모형을 세울 수 있다.

$$\mu_i = t_i \mu(\mathbf{x}_i, \beta) \quad (i = 1, 2, \dots, n) \quad (7.55)$$

확률변수 y_i 는 평균 μ_i 이며 포아송 분포를 적용하므로 y_i 의 분산 또한 μ_i 이다. 물론, 자료에 따라 변한다. 다음 절에서 연결 함수의 역할에 대하여 자세히 알아보도록 하겠다.

포아송 회귀에서 최대우도추정(Maximum Likelihood Estimation in Poisson Regression)

포아송 회귀는 로지스틱 회귀와 유사한 구조라는 것을 알게 될 것이다. 평균은 회귀변수 집합의 함수로 모형화된다. 먼저, β 구성요소를 추정하기 위한 우도함수(likelihood function)와 우도방정식(likelihood equation)을 세울 필요가 있다. 식 (7.54)에서 표현되었듯이 포아송 분포에서 우도(likelihood)는 다음과 같다.

$$\begin{aligned}
 L(\mathbf{y}, \boldsymbol{\beta}) &= \prod_{i=1}^n p(y_i, \boldsymbol{\beta}) \\
 &= \prod_{i=1}^n \left\{ \frac{[t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})]^{y_i} e^{-[t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})]} }{y_i!} \right\} \\
 &= \frac{\left\{ \prod_{i=1}^n [t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})]^{y_i} \right\} e^{-\sum_{i=1}^n t_i \mu(\mathbf{x}_i, \boldsymbol{\beta})}}{\prod_{i=1}^n y_i!}
 \end{aligned} \tag{7.56}$$

일단 $\mu(\mathbf{x}_i, \boldsymbol{\beta})$ 의 함수형태가 선택되면, 반복방법(iterative technique)을 사용하여 (7.56)을 최대화하면 회귀계수 β 에 대한 최대우도추정량(maximum likelihood estimator)을 구할 수 있다. 이러한 최대화는 GLIM이나 CATMAX와 같은 패키지를 통해 구할 수 있다. 이 과정은 로지스틱 회귀와 포아송 회귀에서 최대우도추정량을 구하기 위해 추천되는 방법으로 반복재가중최소제곱(iteratively reweighted least squares(IRWLS))이라고 불린다. IRWLS는 이 장의 뒷부분에서 로버스트(robust)회귀와 함께 살펴보도록 하겠다. 실제로 NLIN이라고 불리는 SAS 프로시저(procedure)는, 비선형회귀를 다룬 9장에서 설명하도록 한다., IRWLS를 사용해서 로지스틱이나 포아송 회귀에 사용할 수 있다.

물론, 정식으로 최대우도추정량을 구하는 방법은 아래의 식을 풀어야 한다.

$$\frac{\partial \ln L(\mathbf{y}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

여기서,

$$\ln L(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^n y_i \ln t_i \mu(\mathbf{x}_i, \boldsymbol{\beta}) - \sum_{i=1}^n t_i \mu(\mathbf{x}_i, \boldsymbol{\beta}) - \sum_{i=1}^n \ln(y_i)$$

따라서, 우도방정식은 다음과 같다.

$$\sum_{i=1}^n \left[\frac{y_i}{\mu(\mathbf{x}_i, \hat{\beta})} - t_i \right] \left[\frac{\partial \mu(\mathbf{x}_i, \hat{\beta})}{\partial \beta} \right] = \mathbf{0} \quad (7.57)$$

식 (7.57)을 $\hat{\beta}$ 에 대해 풀어야 한다. 주의해야하는 점은 일반적인 회귀(standard regression)와 달리 비록 $\mu(\mathbf{x}_i, \beta) = \mathbf{x}_i^\top \beta$ 가 선형함수이더라도 우도방정식(likelihood equation)은 비선형이다. 또한, 식 (7.57)의 공식에서 $\partial[\mu(\mathbf{x}_i, \beta)]/\partial \beta$ 는 열벡터(column vector)라는 것도 주목해야 한다.

포아송 회귀의 중요한 결과는 무엇인가?(What are the Important Results in Poisson Regression?)

2장과 3장에서 일반적인 선형회귀(ordinary linear regression)에서 모형오차(model error)에 관한 의미있는 정규성(normality) 가정이 있다는 것을 배웠다. 그리고 최대우도추정(maximum likelihood estimation)은 최소제곱추정(least squares estimation)과 동일하다. 회귀계수 추정 후 아래의 적합 모형(fitted model)을 평균반응(mean response)을 추정하기 위해 사용할 수 있다.

$$(\hat{E}y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

로지스틱 회귀(logistic regression)의 경우와 같이, 포아송 회귀는 평균반응이 특정한 이산형분포(particular discrete distribution)일 때 사용하고 반면, 일반적인 회귀모형(standard regression model)은 평균이 정규분포를 따를 때 사용한다. 그러나 3가지 모든 경우에 평균 y 에 대해 회귀변수(regressor variable)를 모형화한다. 로지스틱 회귀에서 모형은 식 (7.40)와 같이 주어지고 반면, 포아송의 경우 식 (7.55)로 주어진다. 따라서, 포아송 경우 추정 후 평균은 다음과 같이 얻을 수 있다.

$$\hat{\mu}_i = t_i \mu(\mathbf{x}_i, \hat{\beta})$$

예를 들어, 만약 $\mu(\mathbf{x}_i, \beta)$ 로 $e^{\mathbf{x}_i^\top \beta}$ 가 선택되면, 기간 t 와 회귀변수 x 에 대해 평균반응은 다음과 같다.

$$\hat{\mu}_i = t_i e^{\mathbf{x}_i^\top \hat{\beta}}$$

이것은 기간 t 와 회귀 x 에 주어진 회귀변수의 집합에 대해 사건(incident)의 평균 개수를 추정치를 준다. 부가적으로 y 사건(incident)에 대한 확률을 추정하길 원할 수 있다. 물론 이것은 포아송 분포를 사용하거나 식 (7.54)를 가지고 잔차를 계산하여서 구할 수 있다.

모형의 설정(model building), 계수에 대한 추론(inferences on coefficients) 등에서 7.4절의

로지스틱 모형의 경우에 소개되었던 것처럼 우도비(likelihood ratio)와 편차의 개념(concept of deviance)을 여전히 사용할 수 있다. 식 (7.56)을 통해 포아송 사례에 $L(\mathbf{y}, \boldsymbol{\beta})$ 와 $\ln L(\mathbf{y}, \boldsymbol{\beta})$ 를 통찰할 수 있게 해준다. 다음으로 포아송 회귀에서 추론의 과정을 설명할 수 있는 예제를 선택하였다.

예제 7.10 항공기 사고 자료(The Aircraft Damage Data)

다음 자료는 베트남 전쟁동안 미국 항공기의 사고율을 조사한 것이다.

Observation	y	x_1	x_2	x_3
1	0	0	4	91.5
2	1	0	4	84.0
3	0	0	4	76.5
4	0	0	5	69.0
5	0	0	5	61.5
6	0	0	5	80.0
7	1	0	6	72.5
8	0	0	6	65.0
9	0	0	6	57.5
10	2	0	7	50.0
11	1	0	7	103.0
12	1	0	7	95.5
13	1	0	8	88.0
14	1	0	8	80.5
15	2	0	8	73.0
16	3	1	7	116.1
17	1	1	7	100.6
18	1	1	7	85.0
19	1	1	10	69.4
20	2	1	10	53.9
21	0	1	10	112.3
22	1	1	12	96.7
23	1	1	12	81.1
24	2	1	12	65.6
25	5	1	8	50.0
26	1	1	8	120.0
27	1	1	8	104.4
28	5	1	14	88.9
29	5	1	14	73.7
30	7	1	14	57.8

자료에 사용한 변수는 다음과 같다.

x_1 : 항공기 종류

x_2 : 폭탄 적재 중량

x_3 : 총 비행월수

y : 손상 발생 횟수

포아송 회귀 모형로 자료를 적합하였다. 연결함수(link function)로 $\ln \mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ 을 사용하였고 최대우도추정량들(maximum likelihood estimators)을 식 (7.57)을 사용하여 구했다. 우리는 $t = 1$ 이라고 놓았을 때 하나의 항공기로 생각해야 한다. 포아송 회귀분석의 목표는 변수들이 실제로 항공기 손상 위치에 영향을 미치는지를 결정하는 것이었다. 완전("full") 모형은 다음과 같이 쓸 수 있다.

$$\hat{\mu} = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}$$

모형 변수들의 어떤 부집합들이 포함되는지에 대해 새롭게 알기 위해서 우리는 몇몇 부분집합 모형들의 모형 편차를 아래의 표에 나타내었다.

모형	편차(Deviance)	모형	편차(Deviance)
$x_1 x_2 x_3$	25.95	x_1	38.28
$x_1 x_2$	28.63	x_2	29.21
$x_1 x_3$	32.19	x_3	50.54
$x_2 x_3$	27.22		

하나의 변수가 사용된 최적의 모형은 x_2 와 두개의 변수가 사용된 최적의 모형은 x_2 와 x_3 을 사용한 경우이다. x_1 은 완전 모형에서 유의하지 않다. 왜냐하면,

$$\begin{aligned}\lambda(\beta_1 | \beta_0, \beta_2, \beta_3) &= \lambda(\beta_0, \beta_2, \beta_3) - \lambda(\beta_0, \beta_1, \beta_2, \beta_3) \\ &= 27.22 - 25.95 = 1.27\end{aligned}$$

그 결과 타당한 모형은 x_2 을 포함한다. 다음은 1개의 모수 모형에서 각 모수의 추정량들, 표준오차들, χ^2 -수치를 구할수 있다. 표준오차는 점근 표준 오차이며 방법론은 절 7.4에 있는 로지스틱 회귀 경우에서 토론된 것과 같다.

	추정치	표준오차	χ^2	유의확률
β_0	-1.70097	0.50685	-3.356	0.000791
β_2	0.23112	0.04677	4.942	7.72e-07

그 결과 최종모형은 다음과 같다.

$$\hat{\mu} = e^{[-1.70097 + 0.23112x_2]}$$

여기에서 물론 $\hat{\mu}$ 는 항공기의 평균 손상횟수. 따라서, 적재 중량이 증가함에 따라 항공기의 평균 손상횟수가 증가한다.

아래는 위 예제에 대한 R code이다.

```
y <- c(0, 1, rep(0, 4), 1, 0, 0, 2, rep(1, 4), 2, 3, 1, 1, 1, 2, 0, 1, 1, 2, 5, 1, 1, 5, 5, 7)
x1 <- c(rep(0, 15), rep(1, 15))
x2 <- c(rep(c(4, 5, 6, 7, 8, 7, 10, 12, 8, 14), c(3, 3, 3, 3, 3, 3, 3, 3, 3)))
x3 <- c(91.5, 84.0, 76.5, 69.0, 61.5, 80.0, 72.5, 65.0, 57.5, 50.0,
       103.0, 95.5, 88.0, 80.5, 73.0, 116.1, 100.6, 85.0, 69.4, 53.9, 112.3,
       96.7, 81.1, 65.6, 50.0, 120.0, 104.4, 88.9, 73.7, 57.8)
dat7_10<- data.frame(y,x1,x2,x3)
fit <- glm(y~x1+x2+x3,data=dat7_10,family = poisson(log))
fit1 <- glm(y~x1+x2,data=dat7_10,family = poisson(log))
fit2 <- glm(y~x1+x3,data=dat7_10,family = poisson(log))
fit3 <- glm(y~x2+x3,data=dat7_10,family = poisson(log))
fit4 <- glm(y~x1,data=dat7_10,family = poisson(log))
fit5 <- glm(y~x2,data=dat7_10,family = poisson(log))
fit6 <- glm(y~x3,data=dat7_10,family = poisson(log))
```

7.6. 일반화 선형모형(Generalized Linear Models)

앞의 두 개의 절에서 반응변수가 이산형일 때 회귀분석의 과정과 사용에 대해서 설명하였다. 이것은 이 장에서 다룰게 될 모든 내용의 중요한 개념적 출발이라는 데 주목할 만하다. 7장 이전에 많은 비율은 선형모형에서 가정으로써는 함축적으로는 정규오차(normal error)의 전제(presumption)를 사용하였다. 2장과 3장에서는 선형회귀모형(linear regression model)에서 오차가 정규분포를 따를 때 최소제곱추정(least squares estimation)이 최대우도추정량(maximum likelihood estimator)이라는 것을 강조하였다. 또한, 3장에서 최소제곱추정량(least squares estimator)의 성질은 오차가 비정규분포(nonnorma)일 때보다 정규분포(normal)일 때(uniformly minimum variance unbiased) 더 좋다(stronger)는 것을 상기할 필요가 있다.

다른 명백한 비정규 오차 상황들(Other Obvious Nonnormal Error Situations)

확실히 이 전 두개의 절에서 전개된 이항과 포아송 회귀 상황은 보통 최소제곱이 적용되지 않는 경우이다. 그러나 확실히 다른 것들도 있다. 생물학분야(biological applications)에서 개수를 세는 현상(phenomena)이 전염성이 있다면 사실상 반응의 분산은 평균보다 확실히 크다. 이런 경우 음이항분포(negative binomial distribution)가 적용된다. 반응변수가 연속적이지만 정규오차 가정(Gaussian error assumption)이 이치에 맞지 않는 상황도 있다. 신뢰성 이론(reliability theory)에서 신뢰성장모형(reliability growth models)은 회귀모형의 중요한 응용분야인데 이것은 시스템(system)의 평균 신뢰성이나 합금(alloy)의 파괴강도는 공정 회귀변수(processing regressor variable)의 함수이다. 이 경우에 오차는 지수(exponential), 감마(gamma), 심지어 와이블(Weibull) 확률변수일 수 있다. 통계학자는 과학적 과정을 자료의 근사 분포에 분류할 수 있기를 원한다. 그러나 최근에서야 이러한 지식과 모형 적합의 개념(notion)간에 흥미있는 연결조직을 개발하여왔다. 이것은 많은 응용분야(예를 들어, 신뢰성 성장)에서 회귀변수의 함수를 계산하기 위해 확률이나 모수가 필요하기 때문에 중요하다. 포아송과 로지스틱 회귀는 더 일반적인 선형모형에서 특별한 경우이다.

일반화 선형모형의 형태(Form of the Generalized Linear Model(GLM))

선형모형 그리고 회귀의 중요한 단일한 접근방법(unified approach)은 Nelder과 Wedderburn (1972)에 의해서 소개되었다. 강조하는 것은 선형회귀모형을 적합시에 오차의 분포가 일반적이라는 것이다. 오차가 정규분포를 따르는 건 특별한 경우이다. 포아송과 이항분포도 이 일반적인 분류에 속한다. 일반적으로 다음에 의하여 모형화되는 실험을 생각해보자.

- (i) 관찰치 y_i ($i = 1, 2, \dots, n$)는 독립적으로 분포하고 평균은 $E(y_i) = \mu_i$ 이다. 자료는 회귀구조로 얻어진다. 즉, 회귀변수의 집합 $x_{1i}, x_{2i}, \dots, x_{ki}$ 또한 관찰된다.
- (ii) 밀도함수(또는 이산형의 경우 확률함수)는 다음과 같은 일반적인 형태를 가진

지수족(exponential family)이다.

$$f(y_i, \phi, \theta_i) = \exp\{r(\phi)[y_i \theta_i - g(\theta_i)] + h(\phi, y_i)\} \quad (7.58)$$

여기에서 ϕ 는 단순히 장애(nuisance parameter)이다

- (iii) 모수 θ_i 는 자료마다 변할 수 있다. 이것은 선형함수 $\mathbf{x}_i' \boldsymbol{\beta}$ 를 통해 밀도함수와 회귀변수를 연결하는 모수(parameter)이다.
- (iv) $\mathbf{x}_i' \boldsymbol{\beta}$ 은 $\mu_i = E(y_i)$ 의 선형 예측치(linear predictor)이다.
- (v) μ_i 와 $\mathbf{x}_i' \boldsymbol{\beta}$ 사이의 연결함수(link function)은 모수 θ_i 를 통해서 이루어지며 다음과 같다.

$$\mathbf{x}_i' \boldsymbol{\beta} = s(\mu_i) \quad (7.59)$$

θ_i 가 반드시 평균 μ_i 일 필요는 없다. 그러나, 회귀변수를 가져오고 선형모형을 식으로 체계화하기 위해서 θ_i 와 μ_i 는 관계가 있어야 한다.

- (vi) 평균 μ_i 는 반드시 선형형태(linear fashion)의 회귀변수와 연결될 필요는 없다. 비선형인 경우가 더 흔하다. 그러나, $E(y_i)$ 는 $\mathbf{x}_i' \boldsymbol{\beta}$ 의 단조 미분가능한(monotonic and differentiable) 함수이다.
- (vii) 분산 $\text{Var}(y_i)$ 이 등분산(homogeneous)하다는 가정은 없다. 그러나 y_i 의 분산은 오직 μ_i 를 통해 x 에 따라 변한다고 가정한다.

위에 언급된 7개의 항목들은 일반선형모형의 기본 원칙들이다. 중요한 것은 중요한 모수(important parameter) θ_i 는 분포 평균과 분산 즉, $E(y_i)$ 와 y_i 의 분산을 묶는다는 것이다. 평균과 y_i 의 분산은 자료 간 회귀변수에 따라 변할 수 있다는 것을 명심해야 한다.

y_i 의 평균과 분산(The Mean and Variance of y_i)

식 (7.58)의 지수족(exponential family)을 고려해보자. McCullagh와 Nelder(1983)으로부터 다음을 알 수 있다.

$$E(y_i) = u_i = \frac{\partial g(\theta_i)}{\partial \theta_i}$$

$$\text{Var}(y_i) = \sigma_i^2 = \frac{\partial^2 g(\theta_i)/\partial \theta_i^2}{r(\phi)}$$

그 결과, 밀도함수의 $g(\theta_i)$ 부분(proportion)과 연결함수(link function)을 통해 선형식 $\mathbf{x}_i'\boldsymbol{\beta}$ 의 함수로 평균 μ_i 를 모형화할 수 있다. 다음 예제가 도움이 될 수 있다. y_i 의 분포가 정규분포라고 가정하자. 식 (7.58)의 지수족(exponential family)에서 밀도함수는 다음과 같이 쓸 수 있다.

$$f(y_i, \mu_i, \sigma) = \exp\left\{ \sigma^{-2} \left[\mu_i y_i - \frac{1}{2} \mu_i^2 - \frac{1}{2} y_i^2 \right] + \ln \frac{1}{\sqrt{2\pi}\sigma} \right\}$$

이 경우에

$$\begin{aligned} \theta_i &= \mu_i \\ g(\theta_i) &= \frac{1}{2} \mu_i^2 \\ r(\phi) &= \sigma^{-2} \\ h(y_i, \phi) &= -\frac{1}{2} y_i^2 + \ln \frac{1}{\sqrt{2\pi}\sigma} \end{aligned}$$

따라서, 이미 알고 있듯이 평균과 분산은 다음과 같다.

$$\begin{aligned} \frac{\partial g(\theta_i)}{\partial \theta_i} &= \mu_i \\ \frac{\partial^2 g(\theta_i)/\partial \theta_i^2}{r(\phi)} &= \sigma^2 \quad (\text{모든 } i \text{에 대해서 일정하다.}) \end{aligned}$$

그러므로 예측했듯이, 일반선형모형 틀(framework)내에서 등분산(homogeneous variance)가정과 오차의 정규성 가정은 매우 일치한다. 정규분포의 경우 식 (7.59)의 연결함수(link function)는 항등연결(the unity link)이다. 즉,

$$\mathbf{x}_i'\boldsymbol{\beta} = \mu_i$$

따라서 일반적인 다중선형모형을 얻을 수 있다.

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (7.60)$$

여기에서 $\text{Var}(y_i) = \sigma^2$ 이다. 또한 2와 3장으로부터 $\boldsymbol{\beta}$ 의 최대우도추정량(maximum likelihood

estimator)는 위의 모형을 이용한 최소제곱추정량(least squares estimator)라는 것을 알고 있다.

GLM의 틀에서 포아송과 로지스틱 회귀(Poisson and Logistic Regression in the Framework of GLM)

7.5절에서 언급한 포아송 회귀를 고려해보자. 시간 t_i 에서 사건의 수 y_i 의 확률함수는 식 (7.54)로 주어진다. 즉,

$$p(y_i, \beta) = \frac{e^{-t_i} [\mu(\mathbf{x}_i, \beta)]^{y_i}}{y_i!}$$

이 확률함수가 지수족에 속하는 것은 쉽게 알 수 있다. 왜냐하면

$$p(y_i, \beta) = \exp\{-t_i[\mu(\mathbf{x}_i, \beta)] + y_i \ln t_i \mu(\mathbf{x}_i, \beta) - \ln(y_i)!\}$$

따라서, 다음을 얻을 수 있다.

$$\begin{aligned} r(\phi) &= 1.0 \\ \theta_i &= \ln t_i [\mu(\mathbf{x}_i, \beta)] \\ g(\theta_i) &= t_i \mu(\mathbf{x}_i, \beta) = e^{\theta_i} \\ h(y_i, \phi) &= -\ln y_i! \end{aligned}$$

결과로

$$\begin{aligned} \mu_i &= \frac{\partial g(\theta_i)}{\partial \theta_i} = e^{\theta_i} = t_i \mu(\mathbf{x}_i, \beta) \\ Var(y_i) &= \frac{\partial^2 g(\theta_i)}{\partial \theta_i^2} = e^{\theta_i} = t_i \mu(\mathbf{x}_i, \beta) \end{aligned}$$

7.5절을 상기하면 연결함수(link function) $\mathbf{x}_i^\top \beta = \ln(\mu_i/t_i)$ 를 통해 포아송 평균 $t_i \mu(\mathbf{x}_i, \beta)$ 와 회귀변수를 연결하여 다음 식을 얻는다.

$$\mu_i = t_i e^{\mathbf{x}_i^\top \beta}$$

그 결과로, 일단 포아송 회귀를 적합하기 위한 모형을 아래와 같이 쓸 수 있다.

$$y_i = t_i e^{\mathbf{x}_i \cdot \boldsymbol{\beta}} + \varepsilon_i \quad (7.61)$$

그러나 $\text{Var}(y_i) = t_i e^{\mathbf{x}_i \cdot \boldsymbol{\beta}}$ 는 관찰치마다 다르기 때문에 일반적인 최소제곱법(standard least squares)을 적용하지 않는다. 7.5절의 전개와 설명으로부터 포아송 회귀모형에서 $\boldsymbol{\beta}$ 의 추정을 위해서 최대우도추정방법(maximum likelihood estimation method)을 추천하고 있다는 것을 알고 있다.

GLM의 틀(framework)은 본질적으로 회귀변수의 선형함수, $\mathbf{x}_i \cdot \boldsymbol{\beta}$, 와 포아송 분포의 확률구조를 연결하는 것을 가지고 있다. 정규분포와 포아송 분포의 대비가 흥미롭다. 오차가 정규분포인 경우 식 (7.60)의 선형모형을 얻을 수 있다. 포아송 경우 연결함수로써 매우 타당한 $\mathbf{x}_i \cdot \boldsymbol{\beta} = \ln(\mu_i / t_i)$ 를 통해서 식 (7.61)의 비선형모형을 얻을 수 있다. 정규분포의 경우 항등 연결함수 $\mathbf{x}_i \cdot \boldsymbol{\beta} = \mu_i$ 로 선형모형을 얻었다. 추가로 정규성은 $\text{Var}(y_i)$ 가 등분산이고, 일반적인 최소제곱법(standard least squares)이 계수의 최대우도추정량을 구하는 것을 보았다. 분명히 이 책의 처음 5개의 장은 이러한 종류의 모형을 사용하였다.

7.4절에서 다룬 로지스틱 회귀는 GLM에 속한다. 확률함수는 지수족에 속하는 하나이다. 그룹화된 자료를 고려해보자. 즉, 회귀변수의 집합 \mathbf{x}_i 에 관측치 n_i 가 얻어지고 성공 r_i 가 측정된다. 여기에서 성공 r_i 가 얻어질 확률은 이항분포에 의해 주어진다. 여기에서 $r_i = y_i$ 이며 반응변수로 확률변수이다. 성공 r_i 의 확률은 다음과 같이 주어진다(Walpole와 Myers(1989)) 참조).

$$\begin{aligned} p(r_i, \boldsymbol{\beta}, \mathbf{x}_i) &= \binom{n_i}{r_i} [P(\mathbf{x}_i)]^{r_i} [1 - P(\mathbf{x}_i)]^{n_i - r_i} \quad (r_i = 0, 1, 2, \dots, n) \\ &= \exp \left[\ln \binom{n_i}{r_i} + r_i \ln P(\mathbf{x}_i) + (n_i - r_i) \ln [1 - P(\mathbf{x}_i)] \right] \end{aligned}$$

앞에서 처럼, $P(\mathbf{x}_i)$ 는 회귀변수의 i 번째 측정시에 성공확률이다. 이 경우, 식 (7.58)로부터 다음을 얻을 수 있다.

$$\begin{aligned} \theta_i &= \ln P(\mathbf{x}_i) - \ln [1 - P(\mathbf{x}_i)] \\ g(\theta_i) &= -n_i \ln [1 - P(\mathbf{x}_i)] = -n_i \ln \left[\frac{e^{-\theta_i}}{1 + e^{-\theta_i}} \right] \\ r(\phi) &= 1 \\ h(y_i, \phi) &= -\ln \binom{n_i}{r_i} \end{aligned}$$

위에서 얻은 정보로부터

$$\mu_i = \frac{\partial g(\theta_i)}{\partial \theta_i} = n_i P(\mathbf{x}_i)$$

그리고

$$Var(y_i) = \frac{\partial^2 g(\theta_i)}{\partial \theta_i^2} = n_i P(\mathbf{x}_i)[1 - P(\mathbf{x}_i)]$$

7.4절에서 로지스틱모형(logistic model)을 소개하였다. 이는 y_i 의 평균, 즉 $n_i P(\mathbf{x}_i)$ 과 회귀변수에 관계이다. 이것은 반드시 로지스틱 연결(logistic link)이라는 연결함수를 정의해야한다. 다시 관찰치가 성공의 개수 r_i 인 자료집합에서 로지스틱 관계는

$$n_i P(\mathbf{x}_i) = \frac{n_i}{1 + e^{-\mathbf{x}_i \beta}}$$

또는 $P(\mathbf{x}_i)$ 가 단일 관찰치(single observation)의 평균일 경우, 로지스틱 모형은 다음과 같이 쓸 수 있다.

$$P(\mathbf{x}_i) = \frac{1}{1 + e^{-\mathbf{x}_i \beta}}$$

(7.62)

식 (7.62)는 측정된 반응변수가 이진적(binary)인 경우 로지스틱 회귀를 나타낸다. 정식 모형(formal model)은 다음과 같다.

$$y_i = \frac{1}{1 + e^{-\mathbf{x}_i \beta}} + \varepsilon_i$$

(7.63)

다시 포아송회귀의 경우 식 (7.63)의 모형은 계수에 비선형이다. 또한,

$$Var(y_i) = n_i P(\mathbf{x}_i)[1 - P(\mathbf{x}_i)]$$

위의 식에 주어진 것 처럼 분산이 회귀변수에 따라 변하기 때문에 최소제곱법(least squares procedures)은 적절하지 않다. 7.4절에 사용된 최대우도방법이 로지스틱회귀과 관련있다.

연결함수는 어디에서 오는가? (Where Does the Link Function Come from?)

앞에서 언급한 것과 같이 연결함수(link function)는 $\mu_i = E(y_i)$ 와 선형함수(linear function) $\mathbf{x}_i \boldsymbol{\beta}$ 에 대한 관련된 모형을 결정한다. 확실히 회귀결과는 연결함수(link function)의 선택에 따라 결정된다. 여기에서 정규분포(normal), 이항분포(binomial), 포아송(Poisson) 분포에 대하여 각각의 상황에 맞게 타당한 것을 제안하였다. 사실, 단조, 미분가능(monotonic, differentiable)함수면 어떤 것이라도 사용될 수 있다. 하지만, 특별히 선호되는 함수는 있다. 이항분포의 경우, 확실히 로지스틱 함수(logistic function)가 이에 해당되고, 포아송분포의 경우 μ_i 에 양의 값을 주는 함수가 필요하다. 앞에서 언급한 세 가지 경우, 아래의 관계에서 유도된 연결함수(link function)이다.

$$\mathbf{x}_i \boldsymbol{\beta} = \theta_i \quad (7.64)$$

여기에서 θ_i 는 식 (7.58)에서 언급된 것이다(연습문제 7.10 참조). 이러한 종류의 연결함수를 정준연결함수(canonical link function)라 부른다. 이것은 회귀변수의 선형함수(linear function), $\mathbf{x}_i \boldsymbol{\beta}$ 가 중요한 위치모수(location parameter)의 단조함수에 관한 모형을 만들고자 할 때 사용된다. 이러한 경우의 모형은 종종 간단하고 해석가능하다(simple and interpretable).

일반화선형모형(GLM) 분석에서는 정규분포의 경우와 별개로 일반적인 최소제곱방법을 사용할 수 없다는 것에 주의해야 한다. 다음 절에서는, 일반화 선형모형의 분석에 대해서 알아보고 7.4절과 7.5절의 내용에서 이를 전개하겠다.

GLM의 분석(Analysis of the GLM)

정규분포와 별개로, 연결함수(link function)에 의해 모형은 비선형(nonlinear)이고, 오차 분산은 등분산(homogeneous)가 아니다. 7.4절과 7.5절에서는 이항분포와 포아송분포의 경우 $\boldsymbol{\beta}$ 의 추정을 위한 최대우도추정(maximum likelihood estimation)을 사용해야 한다고 강조하였다. 최대우도과정(maximum likelihood procedure)은 분포가 지수족(exponential family)에 속하는 모든 분포에서 일반적으로 사용할 수 있다. GLM 형태의 모형을 위한 최대우도추정의 일반적인 용어(general notion)를 얘기하기 전에, 확률구조(probability structure)와 평균과 회귀변수와 관계되는 적절한 정준연결함수(canonical link function) 그리고 지수족(exponential) 분포의 분산을 제공하는 표 7.10을 살펴보자.

정준연결함수(canonical link function)가 실제 고려되는 연결함수 중의 하나라는 점을 분명히 알아야 한다. 또한, 표 7.10의 분산은 최대우도분석(maximum likelihood analysis)시 상당히 도움이 된다.

Table 7.10. Distributions, variances, and link functions for some distributions in the exponential family (continued)

	$r(\phi)$	Canonical Link function	Mean	Variance	Model (Canonical Link)	θ_i	$g(\theta_i)$
Normal :							
	$f(y_i, \mu_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}((y_i - \mu_i)/\sigma)^2}$, $\sigma > 0$	$x'_i \beta = \mu_i$	μ_i	σ^2 (constant)	$E(y_i) = x'_i \beta$	μ_i	$\frac{\theta_i^2}{2}$
Poisson :							
	$f(y_i, \mu_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$, $y_i = 0, 1, 2, \dots; \mu_i > 0$	$x'_i \beta = \ln \mu_i$	μ_i	μ_i	$E(y_i) = x'_i \beta$	$\ln(\mu_i)$	e^{θ_i}
Binomial (Logistic): Grouped Data							
422	$f(y_i, n_i, P(\mathbf{x}_i)) = \binom{n_i}{y_i} P(\mathbf{x}_i)^{y_i} (1 - P(\mathbf{x}_i))^{n_i - y_i}$, $y_i = 0, 1, 2, \dots; 0 \leq P(\mathbf{x}_i) \leq 1$	$x'_i \beta = \ln \left[\frac{\mu_i}{n_i P(\mathbf{x}_i)} \right]$	$n_i P(\mathbf{x}_i)$	$\mu_i (1 - P(\mathbf{x}_i))$	$E(y_i) = \frac{\mu_i}{1 + e^{-x'_i \beta}}$	$\ln \frac{1 - P(\mathbf{x}_i)}{P(\mathbf{x}_i)}$	$-\ln \left[\frac{e^{-\theta_i}}{1 + e^{-\theta_i}} \right]$
Gamma:							
	$f(y_i, r, \lambda_i) = \frac{\lambda_i^r}{\Gamma(r)} e^{-\lambda_i y_i} y_i^{r-1}$, $y_i, \lambda_i, r > 0$	$x'_i \beta = \frac{1}{\lambda_i}$	$\frac{r}{\lambda_i}$	$\frac{\mu_i^2}{r}$	$E(y_i) = \frac{1}{x'_i \beta}$	$\frac{\lambda_i}{r}$	$-\ln \left(\frac{1}{\theta_i} \right)$
Geometric :							
	$f(y_i, P(\mathbf{x}_i)) = P(\mathbf{x}_i) (1 - P(\mathbf{x}_i))^{y_i}$, $y_i = 0, 1, 2, \dots; 0 \leq P(\mathbf{x}_i) \leq 1$	$x'_i \beta = \ln \left[\frac{\mu_i}{1 - \mu_i} \right]$	1	$\mu_i (1 - \mu_i)$	$E(y_i) = \frac{1 - P(\mathbf{x}_i)}{P(\mathbf{x}_i)}$	$\ln(1 - P(\mathbf{x}_i))$	$-\ln(1 - e^{\theta_i})$
Negative Binomial :							
	$f(y_i, P(\mathbf{x}_i), \alpha) = \binom{y_i + \alpha - 1}{\alpha - 1} P(\mathbf{x}_i)^\alpha (1 - P(\mathbf{x}_i))^{y_i}$, $y_i = 0, 1, 2, \dots; \alpha > 0; 0 \leq P(\mathbf{x}_i) \leq 1$	$x'_i \beta = -\ln \left[\frac{\alpha + \mu_i}{\mu_i} \right]$	1	$\mu_i \left(1 + \frac{\mu_i}{\alpha} \right)$	$E(y_i) = \frac{\alpha}{e^{\theta_i} - 1}$	$\ln(1 - P(\mathbf{x}_i))$	$-\alpha \ln(1 - e^{\theta_i})$

7.4절과 7.5절에서 우도함수 개요(outlining the likelihood function)에 대해 상당히 상세하게 살펴보았다. 또한, 소프트웨어 패키지를 통해 해답을 얻을 수 있다는 것을 지적하였다. 이 과정은 반복된다. 이 절에서 GLM 분포족(GLM family of distribution)에 대한 최대우도의 일반적인 성질(general nature)에 대해 알아보고자 한다. 만약 식 (7.58)에 있는 밀도함수(density function)를 사용한 후 선형함수(linear function) $\mathbf{x}_i'\boldsymbol{\beta}$ 에 적절한 연결함수(link function)로 사용한다면, $\boldsymbol{\beta}$ 의 추정량을 구하기 위한 식들은 다음과 같다(Nelder와 Wedderburn (1972) 참조).

$$\boxed{\mathbf{X}'\Delta\mathbf{e} = \mathbf{0}} \quad (7.65)$$

여기에서

$$\mathbf{e} = \begin{bmatrix} y_1 - \hat{\mu}_1 \\ y_2 - \hat{\mu}_2 \\ \vdots \\ y_n - \hat{\mu}_n \end{bmatrix}$$

는 $\hat{\mu}_i$ 에 $\hat{\boldsymbol{\beta}}$ 를 포함한 잔차(residual) 벡터(vector)이다. 행렬 Δ 는 연결함수(link function)와 반영한 대각행렬(diagonal matrix)이다. 연결 함수는 $\mathbf{x}_i'\boldsymbol{\beta}$ 와 μ_i 를 연결하고, $\hat{\theta}_i$ 와 $\mathbf{x}_i'\boldsymbol{\beta}$ 를 연결한다. Δ 의 대각선 원소(diagonal element) δ_i ($0, 1, 2, \dots, n$)는 다음과 같다.

$$\delta_i = \frac{\partial \theta_i}{\partial \eta_i}$$

여기에서 $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ 이다. 정준연결함수(canonical link function) $\theta_i = \eta_i$ 를 사용하면, 이 경우 Δ 단위행렬(identity matrix)이 된다. 식 (7.65)의 전개와 관련되어있지만, 여기에서는 언급하지 않겠다. 정준연결함수 사용시 주목해야되는 것은, 아래의 식이 반드시 풀린다는 것이다.

$$\boxed{\mathbf{X}'\mathbf{e} = \mathbf{0}} \quad (7.66)$$

또한, 오차가 정규분포인 경우를 고려함으로써 식 (7.65)의 중요성을 알아챌 수 있다. 정준연결함수를 사용하고 오차가 정규분포를 따르는 경우, $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ (모형이 선형)이고,

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \quad \text{그리고 } (\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

이것은 추정량 $\boldsymbol{\beta}$ 에 대한 정규방정식(normal equation)이다.

식 (7.65) 또는 (7.66)에 대한 답이 보이지 않지만, 표면적으로 만만치 않다(formidable). 그러나(정규분포, 정준연결함수를 제외한 경우) 잔차 자체는 모형의 비선형성을 반영한다.

벡터 $\hat{\beta}$ 는 잔차에 나타난다. 뉴턴-랩슨(Newton-Raphson)방법이 사용될 수 있다(McCullagh and Nelder(1983) 참조). 연결함수 고려시, 정준연결함수(canonical link function)이외의 것을 사용할 상황이 있을 수 있다. 예를 들어, 오차가 정규분포를 따르는 경우, $\mu_i = \ln(\mathbf{x}_i' \beta)$ 를 가정하길 원한다고 하자. 정규분포의 경우,

$$\theta_i = \mu_i$$

따라서,

$$\frac{\partial \theta_i}{\partial (\mathbf{x}_i' \beta)} = \frac{1}{\mathbf{x}_i' \beta}$$

그래서, 식 (7.65)에 있는 행렬 Δ 는 β_1 의 최대우도추정(maximum likelihood estimation)에 대한 답(solution)에서 대각원소로 $1/\mathbf{x}_i' \beta$ 를 포함한다. 추가로 물론 모형은 다음과 같다.

$$y_i = \ln(\mathbf{x}_i' \beta) + \varepsilon_i$$

그 결과적으로 아래의 식을 가진 식 (7.65)에 의해서 회귀계수의 추정량을 구할 수 있다.

$$e_r = y_i - \ln(\mathbf{x}_i' \hat{\beta})$$

따라서, 여기에서 언급된 어떠한 분포라도, 정준연결함수가 필요한 것은 아니다. 표 7.10에 있는 정보를 기본으로, 행렬 Δ 를 구성할 수 있고 식 (7.65)는 우도방정식(likelihood equation)에 대한 해결을 위해 사용할 수 있다.

GLM의 추가 분석, 편차의 사용(Further Analysis of GLM, Use of Deviance)

7.4절과 7.5절에서 포아성모형과 로지스틱 모형을 위해 적합된 모형의 회귀계수에 대한 유의성 검정(significance tests)과 적합결여검정(testing for lack of fit)의 방법을 설명하기 위해 편차(deviance)라는 용어를 도입하였다. 편차(deviance)라는 용어는 식 (7.58)의 지수족(exponential family)에 해당되는 분포에 일반적으로 적용된다. 사실, 식 (7.58)의 일반적인 형태가 주어지면 지수족의 우도(the likelihood of exponential family)를 $L(\beta)$ 로 표시하고, 연결함수(link function)를 통해서 모수(parameter) β 는 최대우도추정량(maximum likelihood estimator) $\hat{\beta}$ 로 대치된다. 앞에서 지적했듯이, 적합된 모형의 최대우도(maximum likelihood) $L(\hat{\beta})$ 와 완벽하게 적합되었을 때(the fit is perfect) 계산된 우도, 즉, 추정량 $\hat{\mu}_i = y_i$ 에 의해서 최대화된 $L(\hat{\mu})$, 두 개의 우도가 있다. 편차는 적합결여의 측도로 다음과 같이 정의된다.

$$\lambda(\beta) = -2[L(\hat{\beta}) - L(\hat{\mu})] \quad (7.67)$$

이것은 적합된 모형이 옳다는 가정하에 자유도(degree of freedom)가 $n-k-1$ 인 근사 카이제곱분포(asymtotically chi-squared distribution)이다. 여기서 $k+1$ 은 β 에 있는 모수의 개수이다. 만약 $\lambda(\beta) > \chi^2_{\alpha, n-k-1}$ 이라면, 모형의 적합결여가 나타난다. 부가적으로 계층적 구조(hierarchical structure)는 모수의 부분집합이나 개개의 가설검정에 사용된다. 이것은 7.4절과 7.5절에 있는 포아송과 기하모형에 대해 나타냈던 것과 유사하다. GLM를 다루는 데 있어서 더 많은 경험이 필요한 독자는 연습문제 7.11, 7.12, 7.13과 7.14를 참고로 하시오. 수치적인 연습은 연습문제 7.15에 있다.

7.7. 정규성 가정의 실패; 이상점의 존재(Failure of Normality Assumption: Presence of Outliers)

이 절에서는 조건이 이상적이지 않은 경우 사용할 수 있는 방법을 지속적으로 조사한다. 처음에는 등분산 가정이 만족되지 않을 때, 대안으로 가중최소제곱방법에 대해 살펴보았고, 모형에 이분산과 곡선의 형태가 나타날 때 이의 해결을 위해 변환방법을 다루었다. 또한, 이산형 반응변수와 좀 더 일반적인 선형모형을 다룰 수 있는 방법을 소개하였다. 이 절에서는 정규성 가정을 재조사하고 다음과 같은 상황하에서 회귀관계를 추정하기 위한 다른방법을 고려해보자.

1. 비정상 오차들
2. 자료집합에서 분산 자료나 바깥점

5장에서 오차에 대한 정규성으로부터 심각하게 이탈한 편차를 알아낼 수 있는 잔차를 계산하는 방법에 대해 알아보았다. 또한, 정규성 가정이 희생될 때보다 정규성 가정하에서 구한 최소제곱추정량이 더 좋은 성질들을 가지고 있음을 상기할 필요가 있다(모든 선형불편추정량 중 최소분산을 가짐(minimum variance of all linear unbiased estimator)). 이것은 오차가 확실히 정규분포가 아닐 때, 최소제곱법의 대안이 더 좋을 수 있고, 정규성 이탈에 저항적(resistant)이라는 것을 의미한다.

회귀분석의 많은 응용분야에서, 자료집합에서 이상점(outlier)의 효과에 저항적인 접근방법이 필요하다. 5장에서는, 이상점(outliers)의 탐지에 대해 논의하였고, 6장에서는, 이상점의 영향력(influence of outlier)을 진단(diagnostic)하는 방법에 대해서 알아보았다. 물론 잔차(residual)의 값으로부터 이상점(outlier)이 존재한다는 사실을 알 수 있다. 아주 자연스럽게도 최소제곱법은 이상점(outlier)이 회귀결과에 서로다른 영향력을 줄 수 있게한다. 이것은 최소제곱 기준이 잔차의 제곱합을 최소화하는 것이기 때문에 놀라운 것이 아니다. 그래서, 최소제곱추정(least squares estimation)이 훌륭하고 직관적으로 타당한 추정방법인 반면에, 이상점(outlier)가 존재하거나 오차가 비정규분포인 비이상적인 상황에서는 성능에 문제가 발생한다.

최소제곱의 효과를 감소시키는 특정한 비정규분포는 두꺼운 꼬리(heavy tails)를 가진 것들이 있다. 물론 실제적으로 평균이동의 존재와 두꺼운 꼬리를 갖는 분포를 구별할 수는 없다. 그러나, 최소제곱분포를 사용하는 경우 그 마지막 결과는 동일하다. 즉, 많이 벗어난 자료점으로 회귀분석결과가 치우쳐 결과에 큰 영향있다. 이러한 비이상적인 조건(nonideal condition)에 저항적이거나 민감하지 않는(insensitive) 것이 로버스트(robust) 방법이다. 회귀분석의 경우, 잔차에 초점이 맞추어져 있고, 유용한 로버스트(robust) 방법은 잔차가 높은 값을 가지는 자료점의 영향을 실질적으로 감소시키는 것이다.

영향력함수(The Influence Function)

개개인의 자료값이 어떻게 회귀분석 결과에 영향을 미치는지를 측정하는 값은 영향력함수(influence function)에 의해서 주어진다. 이것은 특히 최소제곱(least squares)의 경우 두드러진다. 영향력함수를 설명하기 위해 적합함수로 시작해보자.

$$y_i = \mathbf{x}_i' \mathbf{b} \quad (7.68)$$

여기에서 \mathbf{b} 는 최소제곱추정량의 벡터이다. 최소제곱 방법은 $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 를 최소화하는 \mathbf{b} 를 결정하는 것이다. 이것은 다음식의 답을 \mathbf{b} 로 하는 것이다.

$$\sum_{i=1}^n e_i \mathbf{x}_i = \mathbf{0}$$

여기에서 $e_i = y_i - \hat{y}_i$ (7.69)

식 (7.69)는 값이 큰 잔차를 가진 자료의 영향에 대한 좋은 설명을 제공한다. 추정의 좀 더 일반적인 형태를 다음의 식이 해(solution)로 가정하다.

$$\sum_{i=1}^n \psi\left[\frac{e_i}{\sigma}\right] \mathbf{x}_i = \mathbf{0} \quad (7.70)$$

최소제곱과정(식 7.69)은 그리고 식(7.70)의 특별한 형태(special case)이다. 이 함수 $\psi(\cdot)$ 는 영향력함수(influence function)라고 부른다. 식 (7.69)로부터, 최소제곱 경우, i 번째 자료값에 의한 영향은 잔차 e_i 와 비례한다. 다른 말로 하면, 영향력함수(influence function)는 잔차제곱합의 최소화(minimization of residual sum of squares) 결과인 e_i 에 선형이다. 이상점(outlier)에 저항적인 방법은 영향하에서 큰 잔차를 가진 자료점수를 허락하지 않도록 영향함수 $\psi(\cdot)$ 를 선택함으로써 식 (7.70)으로부터 공식화할 수 있다.

M-추정치(The M-estimator)

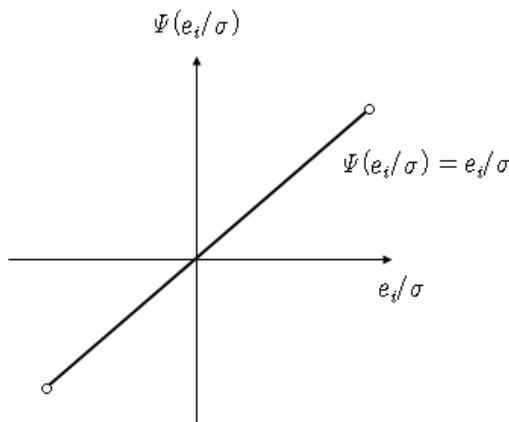
선형회귀모형에서 계수 추정은 물론 최소제곱추정의 경우는 식 (7.69)과 식 (7.70)의 일반적인 경우에 있는 $y_i - \hat{y}_i$ 에서 볼 수 있다. (7.70)에 있는 회귀계수의 해를 M-추정량(M-estimator)으로 부른다. 영향력함수(influence function) $\psi(\cdot)$ 는 일반적으로 적합시 큰 오차를 포함한 자료의 영향을 줄이기 위한 방법으로 선택된다. 이 점에서 $\psi(e_i) = e_i$ 를 가진

최소제곱추정량이 이상점(outlier)에 로버스트(robust)하지 않다는 것을 명백히 해둘 필요는 있다. 다른 것과 비교할 목적으로 그림 7.13에 있는 최소제곱을 위한 영향력함수(influence function) 그림을 제공하였다.

유계인 영향력함수를 사용하는 것이 더 합리적인 접근일 수 있다는 것을 쉽게 볼 수 있다. Huber(1973)에 의해서 제안된 영향력함수가 직관적으로 매력적이다. Huber 함수는 다음과 같다.

$$\begin{aligned}\psi(e_i^*) &= e_i^* & |e_i^*| \leq r \\ &= r & e_i^* > r \\ &= -r & e_i^* < -r\end{aligned}$$

FIGURE 7.13 최소제곱추정을 위한 영향력함수



여기에서 $e_i^* = e_i/\sigma$ 이다. 그림 7.14이 여기에 해당된다.

Huber 함수는 수준(level) r 에서 나타나는 이상의 잔차에 의해 발생되는 어떠한 영향도 허락하지 않는 것이다. R 의 합리적인 값은 1, 1.5 또는 아마도 2.0일 수 있다. 함축된 의미는, $r=1$ 인 경우, σ 를 초과하는 잔차가 σ 값을 가지는 잔차보다 더 큰 영향을 미치지 않는다는 것이다.

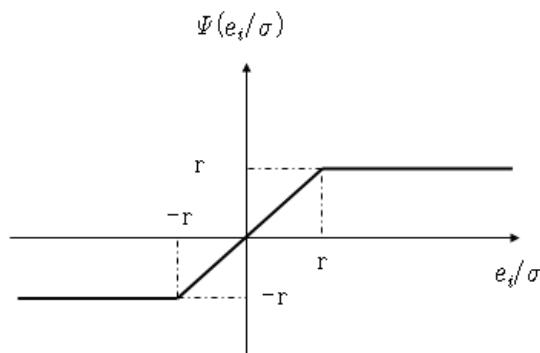
영향력함수(influence function)의 명칭은 회귀분석 사용자에게 큰 잔차를 가진 자료의 영향을 감소시키는 관점으로 추정량을 구하도록 해준다. 근본적으로, 그림 7.14에 묘사된 $\psi(\cdot)$ 를 가지는 식 (7.70)은 회귀계수를 구하기 위한 것이다. 회귀계수는 $\psi(e_i/\sigma)$ 의 e_i 에 연관되었다는 것을 명심해야 한다. 다음 부절에서, 회귀계수를 구하기 위해 식 (7.70)의 답을 찾기 위한 회귀계산에 대해 다루도록 한다.

M-추정치의 계산, 반복가중최소제곱(Computation of the M-estimator, Iteratively Reweighted Least Squares)

(7.70)에 있는 회귀계수의 해는 M-추정치(M-estimator)로 이것은 이상점(outlier)으로 보이는 자료값을 줄이기 위해 선택된 영향력함수(influence function) $\psi(\bullet)$ 를 사용한다. 비록 다른

것을 선택할 수 있지만 Huber의 영향력 함수(influence function) ψ 를 선택하는 것은 합리적으로 보인다. (7.70)에서 풀어야 하는 식들은 비선형(nonlinear)이다. 그러므로 반복적(iterative)으로 풀어야 한다. 또한 실제적으로 σ 는 $\hat{\sigma}$ 로 대치되어야 한다. 이 추정량은 다양하게 선택가능하다. 그러나, 로버스트(robustness) 의미와 일치해야하고 척도(scale)에 로버스트(robust)한 추정량을 선택해야한다.; 합리적인 선택은 다음과 같다.

FIGURE 7.14 Huber의 영향력함수 그림



$$\hat{\sigma} = 1.5 \text{med}|e_i| \quad (i = 1, 2, \dots, n) \quad (7.71)$$

여기에서 $\text{med}|e_i|$ 는 절대 잔차(absolute residual)의 중앙값이다. σ 의 추정량에 대한 상세한 내용은 Welsch (1975)를 참조하라.

M-추정치(M-estimator)를 구하는 많은 유용한 컴퓨터 프로그램이 있다. 자료에 가중치를 부여하는 반복적인 가중최소제곱을 통해 어떻게 반복적으로 해에 접근할 수 있는지가 매우 흥미롭다. 식 (7.70)은($\hat{\sigma}$ 는 σ 를 대신한다.) 다음과 같이 쓸수 있다.

$$\sum_{i=1}^n \frac{\psi\left(\frac{e_i^*}{\hat{\sigma}}\right)}{\left(\frac{e_i^*}{\hat{\sigma}}\right)} \left(\frac{e_i^*}{\hat{\sigma}}\right) \mathbf{x}_i = 0 \quad (7.72)$$

여기에서 e_i^* 은 i 번째 척도무관한 잔차고 $e_i^* = e_i / \hat{\sigma}$ 이다. 식 (7.72)은 다음과 같은 형태이다.

$$\sum_{i=1}^n w_i e_i^* \mathbf{x}_i = 0 \quad (7.73)$$

여기에서 $w_i = \psi(e_i^*)/(e_i^*)$ 이다. 지금 식 (7.73)은 $\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2$ 를 최소화하는 해다. 즉, 가중최소제곱(weighted least squares)이다. 그러므로 가중회귀는 M-추정치를 계산하는 방법으로 사용될 수 있다. 가중은 잔차와 계수에 의존한다. 그 결과, 반복과정(Iterative procedure)이 필요하다. 이러한 반복과정은 아래와 같다.

1. 초기 추정량 \mathbf{b}_0 의 벡터를 구하고 그것들로부터 잔차 $e_{i,0}$ 을 구하라.
2. 초기 잔차들로부터 $\hat{\sigma}_o$ 와 초기 가중 $w_i = \psi(e_i^*)/(e_i^*)$ 을 계산하라.
3. 가중최소제곱을 이용하여 새로운 로버스트(robust) 모수추정량을 얻는다.

$$\mathbf{b}_{RO} = (\mathbf{X}' \mathbf{W}_0 \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}_0 \mathbf{y}$$

여기에서 \mathbf{W}_0 은 가중치 $w_{i,0}$ 를 i 번째 대각원소로 가진 대각행렬이다.

4. 3단계에서 얻은 모수추정을 1단계에 있는 \mathbf{b}_0 의 역할을 하여, 새로운 잔차, $\hat{\sigma}$ 의 새로운 값과 새로운 가중치를 얻는다.
5. 3단계로 돌아가라.

이 방법은 수렴(convergence)할 때까지 계속된다. 이 과정을 반복재가중최소제곱(Iterative reweighted least squares(IRWLS))이라 부른다. 이것은 영향함수 $\psi(\cdot)$ 의 선택에 의존한다. Huber 함수의 경우 전환상수(turning parameter) r 을 선택해야 한다. 확실히 만약 r 이 크다면 예를 들어 $r = 3$ 이라면 로버스트(robust) 추정량은 잔차의 분포에 따라 의존하지만 최소제곱처럼 할 수 있다. 전환상수(turning parameter)에 대한 타당하고 합리적인 값은 $r = 1.0$ 또는 1.5 이다. 이 방법은 또한, 회귀계수의 초기값 \mathbf{b}_0 이 필요하고 이 초기값에서 $\hat{\sigma}_o$ 이 계산된다. 명백하게 초기값을 구하는 가장 단순한 방법은 최소제곱법이다. 그러나 초기값의 선택의 다른방법은 절대 잔차의 합을 최소화하는(즉 $\sum_{i=1}^n |y_i - \hat{y}_i|$ 의 최소화) 회귀계수를 사용하는 것이다. 이 추정문제에 관한 컴퓨터 알고리즘은 아주 흔하다. 예를 들어, SAS에서는 이러한 프로시저(procedure)를 포함한다.

M-추정(M-estimation)의 접근방법은 특히 Huber's $\psi(\cdot)$ 함수와 같은 영향함수의 강조를 통해서 이루어진다. 그러나 어떤 기준을 사용하는 것이 바람직하다. 즉, Huber 함수를 최소화하는 것이 무엇이냐는 것이다. 영향함수가 기준함수(criterion function)를 미분한 결과라는 것을 알았을 것이다. Huber의 M-추정치의 경우 다음을 최소화하는 \mathbf{b} 를 선택하였다.

$$\sum_{i=1}^n \rho(e_i^*) = \sum_{i=1}^n \rho\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}}\right)$$

여기에서

$$\begin{aligned}\rho(e_i^*) &= \frac{e_i^{*2}}{2} & |e_i^*| \leq r \\ &= r|e_i^*| - \frac{1}{2(r^2)} & |e_i^*| > r\end{aligned}$$

따라서, Huber의 $\psi(\cdot)$ 함수를 통한 M-추정(M-estimation)은 일반적인 최소제곱(ordinary least squares)과 같이 작은 잔차의 제곱과 관련되어 있지만, 큰 잔차의 영향을 줄이는 방법을 다룬다.

예제 7.11 트럭 배달 시스템 자료(The Delivery Time Data)

소프트 드링크 판매회사의 트럭배달 시스템에 대한 평가를 하기 위해 배달시간 y 를 반응변수로 하고 배달상자의 수를 x_1 , 배달거리를 x_2 로 한 자료이다.

Observation Number	Delivery Time(min) y	Number of Cases x_1	Distance(feet) x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	20	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132

19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

이에 대한 최소제곱 회귀선은 다음과 같다.

$$\hat{y} = 2.341231 + 1.615907x_1 + 0.014385x_2$$

여기서 $R^2 = 95.96$ 이고 계수의 추정치도 모두 유의하므로 적합에서 특별한 점을 파악할 수 없지만, 이에 대한 잔차를 그려보면 정규성에 대한 가정이 의문시된다. 오차는 꼬리부분에서 많이 나타나며 특히 9, 20번째 자료에서 두드러진다.

Huber's 영향함수와 반복재가중최소제곱(IRWLS)의 계산은 계수의 집합을 만들어내는데 사용된다. 값 $r = 2.0$ 은 M-추정치를 계산하는데 선택된다. 우리는 식(7.71)로부터 $\hat{\sigma}_o$ 의 값을 사용하고 초기 추정량과 초기 잔차를 알기 위해서 보통 최소제곱을 적용한다. 표 7.11은 초기 IRWLS 연습시 적용되는 초기 잔차와 가중을 제공한다.

최소제곱추정계산에 대한 절대 잔차의 합은 57.099분이다. 척도모수(scale parameter)의 초기 로버스트(robust) 추정은 $\hat{\sigma}_o = 1.629783$ 이다. 가중 최소제곱은 새로운 계수와 새로운 잔차를 구하기 위해서 적용된다. 새로운 추정량들과 새로운 잔차는 표 7.12에서 볼 수 있다. 명칭 $b_{j,Ro}$ 는 로버스트(robust) 추정을 말한다.

절대 잔차의 합은 첫번째 IRWLS 단계후에 55.46029이다. 이 과정은 새로운 추정량인 각 단계에서 계산된다. 최종 반복의 결과들은 표 7.13에 있다. 최종 로버스트(robust) 회귀계수를 구한 가중치와 최종 잔차의 집합이 포함되어있다.

최종 반복에서 절대잔차의 합은 54.18899로 이것은 보통 최소제곱에 의해서 제공되는 적합은 전반적인 질보다 더 나음을 나타낸다. 표 7.14는 각 반복에서 OLS에 대한 회귀계수의 수치와 절대잔차의 합을 보여준다.

TABLE 7.11 트럭 배달 시스템 자료에서 초기 잔차와 회귀계수의 초기값

Site	$e_i = y_i - \hat{y}_i$	$w_{i,o}$	Site	$e_i = y_i - \hat{y}_i$	$w_{i,o}$
1	-5.0280843	1.0	14	1.0675359	1.0
2	1.1463854	1.0	15	0.6712018	1.0
3	-0.0497937	1.0	16	-0.6629284	1.0
4	4.9243539	1.0	17	0.4363603	1.0

5	-0.4443983	1.0	18	3.4486213	1.0
6	-0.2895743	1.0	19	1.7931935	1.0
7	0.8446235	1.0	20	-5.7879699	1.0
8	1.1566049	1.0	21	-2.6141789	1.0
9	7.4197062	1.0	22	-3.6865279	1.0
10	2.3764129	1.0	23	-4.6075679	1.0
11	2.2374930	1.0	24	-4.5728535	1.0
12	-0.5930409	1.0	25	-0.2125839	1.0
13	1.0270093	1.0			

TABLE 7.12 IRWLS 첫번째 과정에서 회귀계수가 잔차

Site	$e_i = y_i - \hat{y}_i$
1	-5.1461536
2	0.7392174
3	-0.4476047
4	4.5719162
5	-0.6600842
6	-0.4255778
7	0.3632314
8	1.0112443
9	8.8816912
10	2.1305589
11	2.7202259
12	-0.5410711
13	0.6882172
14	0.8761782
15	0.6756930
16	-0.5672146
17	0.2245732
18	3.2971787
19	1.3716780
20	-5.2331963
21	-2.5680573
22	-2.5378139

Robust Regression
Coefficients
$b_{0,Ro} = 2.96249413$
$b_{1,Ro} = 1.55026040$
$b_{2,Ro} = 0.01430685$

23	-4.6029207
24	-4.6194278
25	-0.5595634

IRWLS 마지막 단계 결과

Site	$e_i = y_i - \hat{y}_i$	w _i	
1	-5.08344667	0.5378440	Robust Regression Coefficients $b_{0,Ro} = 3.38198302$ $b_{1,Ro} = 1.49998755$ $b_{2,Ro} = 0.01407420$
2	0.52173080	1.0000000	
3	-0.63717293	1.0000000	
4	4.37213094	0.6254116	
5	-0.74303801	1.0000000	
6	-0.41638117	1.0000000	
7	0.06988012	1.0000000	
8	0.99252256	1.0000000	
9	10.31006155	0.2653721	
10	2.10318954	1.0000000	
11	3.26516802	0.8379887	
12	-0.40781109	1.0000000	
13	0.52914633	1.0000000	
14	0.86581227	1.0000000	
15	0.81288838	1.0000000	
16	-0.30343606	1.0000000	
17	0.15325210	1.0000000	
18	3.26030999	0.8386894	
19	1.11138320	1.0000000	
20	-4.61890375	0.5918309	
21	-2.45224626	1.0000000	
22	-1.46175966	1.0000000	
23	-4.46526002	0.6123489	
24	-4.48899906	0.6090076	
25	-0.74306290	1.0000000	

TABLE 7.14 IRWLS 9번째 단계에서 회귀계수 및 절대잔차합

Step	$b_{0,Ro}$	$b_{1,Ro}$	$b_{2,Ro}$	$\sum_{i=1}^n y_i - \hat{y}_i $
OLS	2.34123115	1.61590721	0.01438483	57.099
1	2.96249413	1.55026040	0.01430685	55.46029
2	3.23874610	1.50358360	0.01454736	54.5849
3	3.33285284	1.49214617	0.01446407	54.28425
4	3.35171046	1.49985340	0.01420023	54.27019
5	3.37008781	1.49857664	0.01415745	54.21469
6	3.37559485	1.49983271	0.01410408	54.20561
7	3.37957894	1.49975529	0.01408975	54.19445
8	3.38107287	1.49997392	0.01407825	54.1914
9	3.38198302	1.49998755	0.01407420	54.18899

다음은 위 예제에 대한 R code이다.

```

x1<-c(7,3,3,4,6,7,2,7,30,5,16,10,4,6,9,10,6,7,3,17,10,26,9,8,4)
x2<-c(560,220,340,80,150,330,110,210,1460,605,688,215,255,
      462,448,776,200,132,36,770,140,810,450,635,150)
y<-c(16.68,11.50,12.03,14.88,13.75,18.11,8.00,17.83,79.24,21.50,40.33,21.00,13.50,
      19.75,24.00,29.00,15.35,19.00,9.50,35.10,17.90,52.32,18.75,19.83,10.75)
x<-cbind(x1,x2)
delivery<-data.frame(x,y)

res<-matrix(rep(0,25*50),nrow=25)
fitv<-matrix(rep(0,25*50),nrow=25)
wt<-matrix(rep(1,25*51),nrow=25)
bet<-matrix(rep(0,3*50),nrow=3)
sig<-rep(0,50)
res_sta<-matrix(rep(1,25*50),nrow=25)
psi<-matrix(rep(1,25*50),nrow=25)

hub<-function(dat,r){
  n<-length(dat)

```

```

p<-rep(0,n)
for(i in 1:n){
  if(dat[i] > r) p[i]<-r
  else if(dat[i] < -r) p[i]<-r
  else p[i]<-dat[i]
}
return(p)
}

for(i in 1:50){
  fit<-lm(y~x,data=delivery,weights=wt[,i])
  bet[,i]<-fit$coefficients
  res[,i]<-residuals(fit)
  fitv[,i]<-y-residuals(fit)
  sig[i]<-median(abs(res[,i]-median(res[,i])))/0.6745
  res_sta[,i]<-res[,i]/sig[i]
  psi[,i]<-hub(res_sta[,i],2)
  wt[,i+1]<-psi[,i]/res_sta[,i] #update weight
}

```

로버스트 회귀에서 유계영향은 잔차를 조정한다는 것이다. 조정은 자료값에 의해 추론된 지렛값을 고려한다. 효과적으로 유계영향방법의 예로는 로버스트 회귀에서 이전에 언급하였듯이 잔차를 스튜던트화 잔차로 대치하는 것이다. 다른 가능성은 PRESS 잔차나 심지어 6장에서 언급된 DFFITS 영향진단을 사용하는 것이다. 분명히 만약 h_{ij} 가 1.0에 가까워지면 스튜던트화 잔차는

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

$e_i^* = e_i / \hat{\sigma}$ 보다 더 큰 값으로 나타날 것이고 이것은 여기에서 언급된 일반적인 로버스트 회귀(standard robust regression) 하에서 더 작은 가중치가 할당된다. 그 결과, 유계 영향회귀는 이상점에 더 노출된 값으로 $e_i / \hat{\sigma}$ 를 대치한다. 유계 영향회귀에 대한 더 많은 토론이 Krasker와and Welsch (1979)에서 볼 수 있다. 추가로 독자는 이 장의 끝부분에 있는 연습 7.7을 살펴보기바란다.

로버스트(Robust) 회귀에 관한 더 많은 언급들(Further Comments Concerning Robust Regression)

M-추정(M-estimation)의 수렴(convergence) 성질은 초기값에 따라 결정된다. 추가로 몇몇 영향력함수의 경우 수렴은 더 불확실하다. Hampel (1974), Andrews (1974), Beaton와 Tukey (1974) 그리고 Birch (1980)을 보시오.

많은 경우에서, 분석가는 평균절대편차를 수렴의 기준으로 사용하는 것으로 만족한다. 일단 로버스트(robust) 과정이 시작되면, 평균 절대 잔차는 감소할 것이다(몇몇 관찰치의 영향이 줄어든다고 가정하면). 예제 6.6에서처럼, 반복과정이 진행되는 동안 $\sum|e_i|$ 를 지켜보아야 한다.

앞에서 언급한 것처럼, 로버스트(robust) 회귀의 목적은 이상점(outlier)가 존재하는 경우 최소제곱의 대안을 제공하는 것이다. 또한 로버스트(robust) 회귀를 이상점(outlier) 진단 방법으로 볼 수 있다. 분명히 1보다 적은 가중을 받는 어떠한 점은 의심할 수 있다. 그러나 방법의 주요 기능은 추정(estimation)이지 진단(diagnosis)이 아니다. 오차가 뚜거운 꼬리를 가지는 분포를 따르는 경우 회귀계수는 최소제곱법의 회귀계수에 비해 우수하다. 5장에서, 상세히 이상점 진단(diagnosis of outlier)에 관해서 알아보았다. 이상점(outlier)의 탐지가 유익하고 유용하다고 증명된 많은 상황이 있다. 심지어 이상점이 결과에 역효과를 미치더라도 분석가는 이상점을 완전히 제거하려고 하진 않는다. 로버스트(Robust) 회귀는 진단적인 측면이 있고 비이상적인 상태에서 최소제곱의 그것보다 성질면에서 더 우수한 계수 추정량을 제공한다. 이것을 로버스트(robust) 회귀의 불필요한 설명으로 치부해서는 안된다. 다른 영향함수 뿐만 아니라 유계 영향회귀에 대한 상세한 내용을 알고 싶다면 예로 Krasker and Welsch (1982)를 보시오.

7.8. 측정오차가 있는 회귀변수(Measurement Errors in the Regression Variable)

이 장에서 정규성(normality), 등분산(homogeneous variance), 모형지정(model specification)를 포함하는 모형위배(violation)를 다루었다. 이 절에서는 추가로 하나의 모형 위배의 효과에 대해서 간략하게 다루도록 한다.

2장의 초반부에 최소제곱의 언급된 성질이 회귀변수가 확률변수가 아니라는 가정에 의존한다고 강조하였다. 2장의 후반부에서는 x 와 y 가 확률변수로 이변량 정규분포라는 가정을 동반하는 선형모형에 대해서 다루었다. 그러나 회귀변수에 측정오차(measurement error)의 영향에 관한 것은 거의 없거나 전무하다. 이 장에서 이 주제를 철저하게 다루려는 의도는 없다. 목적은 왜 이 문제가 최소제곱계산에 어려움을 발생하는지를 인식하는 것이다. 이에 대한 중요한 방법은 Seber (1977), Davis와 Hutton (1975), Berson (1950) 그리고 Mandansky (1959)를 보시오.

일반적인 다중선형모형을 고려해보자.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad (i=1,2,\dots,n)$$

전통적인 고정된 x 모형에서, x_{ij} 는 확률변수가 아니다. 그러나 각 회귀변수가 오차가 있게 관찰되었다고 가정하자. 그 결과, 분석가는 다음을 관찰한다.

$$u_{ji} = x_{ji} + \delta_{ji} \quad (7.74)$$

δ_{ji} 는 확률 측정오차(random measurement error)이다. 다음과 같은 타당한 가정이라 볼 수 있다.

$$E(\delta_{ji}) = 0$$

$$Var(\delta_{ji}) = \sigma_j^2$$

그리고 δ_{ji} 는 각각 서로독립이고, ε_i 과도 독립이다. 그 결과 회귀모형은 다음과 같이 쓸 수 있다.

$$\begin{aligned}
y_i &= \beta_0 + \sum_{j=1}^k \beta_j (u_{ji} - \delta_{ji}) + \varepsilon_i \\
&= \beta_o + \sum_{j=1}^k \beta_j u_{ji} + \left(\varepsilon_i - \sum_{j=1}^k \beta_j \delta_{ji} \right) \\
&= \beta_o + \sum_{j=1}^k \beta_j u_{ji} + \xi_i
\end{aligned} \tag{7.75}$$

식 (7.75)는 관찰할 수 있는 회귀변수 즉, u_{ji} 에서 모형을 나타낸다. 새로운 모형 오차는 다음과 같고 기대값은 0이다.

$$\xi_i = \varepsilon_i - \sum_{j=1}^k \beta_j \delta_{ji}$$

ξ_i 의 분산은 다음과 같고 이것은 쉽게 보일 수 있다.

$$Var(\xi_i) = \sigma^2 + \sum_{j=1}^k \beta_j^2 \sigma_j^2 \tag{7.76}$$

(7.76)로부터 회귀변수에서 측정오차의 분산은 모형 오차로 전달되고, 그것에 의해 모형오차분산은 크게 된다. 오차분산 $Var(\xi_i)$ 는 만약 측정오차가 없을 때보다 더 크게된다.

분명히 증가(inflation)값은 σ_j^2 의 값, 개개인의 오차 분산에 크게 좌우된다.

모형오차분산의 증가(inflation of model error variance)는 측정오차(measurement error)가 존재 시에 반드시 직면해야 되는 유일한 어려움은 아니다. 관찰된 값, u_{ji} 는 회귀변수로 사용되어지고 식 (7.74)에 따라 확률변수이다. ξ_i 와 특정한 u_{ji} 에 대한 자세히 살펴보면 다음을 얻을 수 있다.

$$\begin{aligned}
 Cov(\xi_i, u_{ji}) &= E(\xi_i)(u_{ji} - \bar{x}_{ji}) \\
 &= E\left(\varepsilon_i - \sum_{r=1}^k \beta_r \delta_{ri}\right) (\delta_{ji}) \\
 &= -\beta_j \sigma_j^2
 \end{aligned} \tag{7.77}$$

그래서, 모형 오차 ξ_i 는 각 회귀변수와 상관(correlated)되어 있다. 따라서, 회귀변수의 임의적 성질(random nature)과 모형오차와 회귀변수사이의 공분산이 0이 아니므로, 3장에서 인용된 Gauss-Markoff 이론이 더이상 적용되지 않는다. 사실상, 3장에서 적용한 최소제곱추정량의 불편성(unbiased)을 보이기 위한 방법이 더 이상 적용될 수 없다는 것을 인식해야한다. 사실, 모수의 최소제곱추정량은 편의추정량이다. 예상했던대로, 회귀계수에서 측정오차의 효과는 j 번째 측정오차분산(measurement error variance) σ_j^2 이 j 번째 회귀변수(regressor)에서 생기는 변동(variation) $\sum_{i=1}^n (u_{ji} - \bar{u}_j)^2$ 과 비교해서 작다면 무시할 수 있다. 회귀계수의 편의와 이 문제의 심각성에 관한 자세한 내용을 원한다면 Davis와 Hutton (1975) 그리고 Seber (1977)을 보시오.

6. 다중공선성의 진단과 제거(Detecting and Combating Multicollinearity)

3장에서, 회귀변수들 간에 다중공선성(multicollinearity) 조건(condition)이 있을 때 모형구축의 효과와 위험성에 대하여 논의하였다. 다중공선성은, 회귀변수들이 독립적이지 않고 과잉 정보(redundant information)를 보일 때 존재한다. 따라서 변수들의 개별적 역할을 밝히려는 어떠한 시도도 대단히 애매해진다.

최소제곱회귀방법론(least squares regression methodology)의 잠재적 사용자들이 다중공선성의 효과를 이해하는 것뿐만 아니라, 진단하는 법을 배우는 것도 필수적이다. 공선성을 제거하기 위하여 설계된 대체 추정방법들(alternative estimation procedures)은 편향추정기법(biased estimation techniques)의 범주에 들어간다. 이러한 방법들은 보통최소제곱(ordinary least square)으로부터의 편차(deviation)를 기술해준다. 특정 회귀계수의 부호(sign)나 크기(magnitude)로부터 추론을 도출하려는 경제학자들은, 다중공선성이 있을 경우, 최소제곱계수들이 나쁘게 추정될 수 있다는 사실을 알아야 한다. 선형예측방정식(linear prediction equation)을 개발하는데 관심이 있는 기술자들은, 모형이 자료를 꽤 잘 적합한다 하더라도, 다중공선성으로 인하여 예측의 질이 나빠질 수 있다는 것에 주의하여야 한다.

8.1. 다중공선성 진단(Multicollinearity Diagnostics)

이 장에서, 수식전개의 대부분이 다중공선성 진단도구(multicollinearity diagnostics) 역할을 하는 양들(quantities)을 생성할 것이다. 이들을 이용하여 다중공선성 문제(multicollinearity problem)의 정도(extent)를 평가할 수 있다. 여기서는 예제들을 통하여, 이러한 진단법들을 도식화하고, 심각한 다중공선성을 발견하는데 도움을 주는 다른 방법들을 알아보고자 한다. 다음은 정식 진단도구들(formal diagnostic tools)이다.

회귀변수들 간의 단순상관(Simple Correlations Among the Regressor Variables)

분석가들은 회귀변수들의 상관행렬(correlation matrix) $X^* X^*$ 을 일반적으로 접하게 된다. 여기서 X^* 행렬의 열들(columns)은, 3.8절에서 논의된 바와 같이 중심척도화 된다. 물론 이 숫자들은 짹을 이룬 형태의 상관(pairwise type correlations)을 의미한다. 그러나 이름이 암시하듯이, 다중공선성은 여러 회귀변수들 간의 연관성(association)을 포함하고 있다는 것을 분명하게 알아야 한다. 따라서, 단순상관(simple correlations)만으로는 다중공선성의 정도를 항상 알 수는 없다. 많은 분석가들이 심한 다중공선성이 있다고 주장할 수 있는 지침값(guideline values)을 끊임없이 찾고 있다. 단순 상관에 절대적인 지침값이 있어서 1:1 연관성(one on one association)이 있는지 찾을 수 있으면 좋겠으나, 실제 명확한 지침값이 없기 때문에 다중공선성의 성질(nature) 혹은 정도(extent)를 단순상관으로 항상 알 수는 없다.

분산팽창요인(Variance Inflation Factors)

VIFs(분산팽창요인)는 이상적일 때 즉, 상관행렬(correlation matrix)이 항등행렬(identity matrix)일 때보다 각각의 회귀계수가 더 팽창(inflation)하는 것을 의미한다. i 번째 계수의 분산팽창요인은 식(3.34)로 정의된다. 이것은 다중상관(multiple association)의 개념을 포함하므로 알기 쉽다. 만약 식(3.34)의 R_i^2 가 1 (unity)에 가까운 값을 갖는다면, $(VIF)_i$ 는 매우 커질 것이다. 이는 i 번째 회귀변수가 다른 회귀변수와 강한 선형관계가 있을 때 발생할 것이다. VIFs는 단순상관값들보다 다중공선성 색출(detection)에 훨씬 생산적이다. VIFs는, 어떤 계수들이 얼마나 부정적으로 영향을 받고 있는지 알 수 있게 해준다. 수치값에 대하여 절대적인 주먹구구식 지침(rule of thumb)은 없으나, 만약 VIFs가 10을 넘어서면, 최소한 뭔가 있다고 할 만하다. 이때는 변수삭제(variable deletion)를 하거나 최소제곱추정 이외의, 다중공선성을 해결해 줄 수 있는 대체추정법을 생각해 보아야 한다.

XX 의 고유값의 체계(System of Eigenvalues of XX)

우리는 회귀자료 세트에 존재하는 상관행렬의 고유값(eigenvalues)과 고유벡터(eigenvectors)가 다중공선성에 중요한 역할을 한다는 것을 3장에서 알았다. 최소

고유값(sallest eigenvalue)이 0에 가까운 정도(nearness to zero)는 선형의존성(linear dependency)이 얼마나 강한가에 대한 측도이다. 한편 다중공선성에서, 연관되어 있는 정규화 고유벡터(normalized eigenvector)의 원소들(elements)은 대응하는 회귀변수들에 대한 가중치(weights)를 나타낸다(3.8절 참고). 물론 변수들이 직교시스템(orthogonal system)을 정의하여 분석가들에게 기준(norm)을 제공한다면, 고유값들은 모두 1 (unity)이 될 것이다. 추가적으로, 고유값의 범위(spectrum)는 또 다른 진단도구를 제공한다. 다중공선성은 최대 고유값과 최소 고유값의 비(ratio)로 측정될 수 있다. 즉, 다음과 같은 양(quantity)을 생각할 수 있다.

$$\phi = \frac{\lambda_{\max}}{\lambda_{\min}}$$

이것은 상관행렬(correlation matrix)의 조건수(condition number)라고 부른다. ϕ 의 값이 크면, 심각한 다중공선성을 의미한다. 지나치게 큰 조건수는 회귀계수들이 불안정하다는 증거이다 (3.8절 참고). 즉, 회귀자료가 조금만 변해도, 회귀계수들은 크게 변한다. 고유값과 고유값의 범위(spectrum)에 대한 수치적 주먹구구식 지침(numerical rule of thumb) 역시 분석가들의 관심사이다. 상관행렬의 조건수(condition number)가 1,000을 넘어서게 되면, 다중공선성을 생각해 보아야 한다. 개별 고유값(individual eigenvalues)에 관하여, 0에 가까운 고유값의 개수는 회귀변수들 간의, 탐지된 공선성(collinearity)의 개수를 의미한다. 물론, 믿을만한 주먹구구식 지침이 항상 믿을만하지는 않다. 따라서 역치값(threshold value, 이 값보다 작은 고유값은 심각한 공선성을 의미한다)을 정하기는 어렵다. 사실, 고유값 λ_j 보다는 고유값의 비(ratios of eigenvalues) $\phi_j = \lambda_{\max} / \lambda_j$ 가 의존성(dependency)의 영향(impact)을 진단하는데 좀 더 믿을만하다.

진단에 관한 추가적인 설명들(Futher Comments concerning Diagnostics)

다중공선성을 진단하려면 여러가지를 고려하여야 하며, 이들 모두가 회귀분석 결과의 중요한 부분들이다. 수학을 잘 못하는 분석가는 이러한 진단에 고유값 시스템(system of eigenvalues)을 잘 사용하지 않으려 할 것이나, 그것이 유용하다는 것은 부인할 수 없을 것이다. 진단도구들(diagnostic tools)은 공선성의 정도(severity of collinearity)를 측정하고, 최소제곱 이외의 대체추정법을 시도하여야 하는지 결정할 수 있도록 설계되었다. 그러나 다중공선성에 대처하기 위하여 사용하는 편향추정기법(biasd estimation techniques)이 예측이나 추정을 더 잘 할 수 있는지는, 해보기 전까지는 확신할 수 없다(8.4 절 참고). 그러므로 대체추정법의 성공 가능성을 보여주는 도구로서 진단도구를 볼 것이 아니라, 보통최소제곱(ordinary least squares)의 비효율성(inefficiency)을 보여주는 도구로 보아야 한다.

어떤 유형의 척도화(scaling)를 사용하였느냐에 따라서 진단도구의 성질이 흔히 달라진다.

3장에서, $X'X$ 는 상관형태(correlation form)이고, 이 상관(correlation)은 각각의 회귀변수들을 중심화(centering), 척도화(scaling)하는 과정에 의한 것으로 가정하였다. 따라서 $X^* X^*$ 라는 표기는 상수항(constant term)이 제거된 $k \times k$ 행렬이다(식3.31 참고). 이것 대신에, 각각의 회귀계수를 척도화만 하여 구한 $(k+1) \times (k+1)$ 행렬 $X'X$ 를 이용하여 진단도구를 계산할 수 있다. 일반적으로, 이 경우는 고유값들(eigenvalues)이 중심화만 한 경우와 똑같지는 않을 것이다. 다중공선성 진단도구(multicollinearity diagnostics)를 계산할 때, 자료분석컴퓨터 패키지는 이 점에서 다양하다. 각각의 회귀변수를 중심화하면 진단도구의 상수항은 제거된다. 이와 달리, 중심척도화시키면, 진단의 객체(objects)가 진짜로 표준화된 변수들(truly standardized variables)의 계수(coefficients)가 될 수 있다. 어떤 경우에는 X 행렬의 열들(columns)은 단위 길이(unit length) 즉, 열의 원소들(elements)의 제곱합(sum of squares)이 1.0이 되는 것이 중요하다. 따라서 중심화되지 않은 회귀변수들에서, i 번째 회귀변수의 j 번째 표기(reading)는 다음과 같다.

$$\frac{x_{ij}}{\sqrt{\sum_{j=1}^n x_{ij}^2}} \quad (8.1)$$

때때로 분석은 어떤 소프트웨어 패키지를 쓸 수 있느냐에 달려있을 수도 있다. 그러나, 어떤 경우는 자료를 중심화하여야 하나, 다른 경우는 이로 인하여 진단이 잘못 될 수도 있다. 8.3절에서 이 문제에 대하여 좀 더 시간을 할애하려고 한다. 공선성 진단 결과들(collinearity diagnostic results)은 자료의 중심화 여부에 따라 달라지는데, 진단도구들을 충분히 이해하고 나면, 이 문제를 더 잘 알게 될 것이다.

8.2. 분산 비율(Variance Proportions)

8.1절에서 소개된 진단도구는 선형의존성(linear dependency)의 정도와, 각 회귀계수의 분산이 이상적인 경우보다 얼마나 더 팽창되었는지 보여주기 위하여 고안된 것이다. 다중공선성이 심하면(serious multicollinearity) 하나의 회귀계수에만 영향을 미치는 것이 아니다. 사실, $X'X$ 혹은 X^*X^* 에 작은 고유값(small eigenvalue)이 하나라도 있다면, 모든 회귀계수가 한꺼번에 혹은 개별적으로 악영향을 받을 수 있음을 뜻한다. 각 계수마다 분산의 어느 정도 비율(proportion)이 각각의 의존성(dependency) 때문에 발생하는지 판단하는 것이 흔히 관심사이다. 모형절편(model intercept)은 회귀자료(regressor data)가 중심화되지 않는 한 영향을 받으며, 이런 유형의 분석방법으로 평가될 수 있을 것이다. 척도화되었으나 중심화는 되지 않은 $X'X$ 의 고유값 분해(eigenvalue decomposition)에 대하여 생각해보자.

$$V'(X'X)V = \begin{bmatrix} \lambda_0 & & 0 \\ & \lambda_1 & \\ 0 & & \ddots \\ & & & \lambda_k \end{bmatrix} \quad (8.2)$$

분산공분산행렬 $(X'X)^{-1}$ 은 다음과 같다.

$$(X'X)^{-1} = [v_0 v_1 \cdots v_k] \begin{bmatrix} 1/\lambda_0 & & 0 \\ & 1/\lambda_1 & \\ 0 & & \ddots \\ & & & 1/\lambda_k \end{bmatrix} \begin{bmatrix} v'_0 \\ v'_1 \\ \vdots \\ v'_k \end{bmatrix} \quad (8.3)$$

절편(intercept)이 포함되었을 때, $X'X$ 의 차원(dimension)이 증가하기 때문에 한 개의 추가적인 고유값을 고려하기 위하여 $\lambda_0, \lambda_1, \dots, \lambda_k$ 라는 표기를 사용한다. 계수들의 분산들은 (σ^2 은 별도로 하고) $(X'X)^{-1}$ 의 중심 대각선(main diagonals) 상에 있다. λ_j 와 연관된 고유벡터(eigenvector)의 i 번째 원소(element)를 v_{ij} 로 표시하고, 고유벡터들이 행렬 V 의 열들(columns)이라면, 식(8.3)으로부터 다음과 같이 쓸 수 있다.

$$c_{ii} = \sum_{r=0}^k \frac{v_{ir}^2}{\lambda_r} \quad (8.4)$$

여기서, $c_{ii} = \text{Var}(b_i)/\sigma^2$ 이다. 식(8.4)로부터, 작은 고유값이 모든 분산들에 어느 정도 영향을 미치는 것을 설명하기는 쉽다. 그 정도(extent)를 정량화하기 위하여, 다음을 정의할 수 있다.

$$P_{ji} = \frac{\nu_j^2 / \lambda_j}{c_{ii}} \quad (8.5)$$

P_{ji} 는, 고유값 λ_j 로 특성화되는 공선성 때문에 발생하는 b_j 의 분산의 비율(proportion of the variance)이다. 분산비율(variance proportions)은 선형의존성(linear dependency)의 효과를 평가하는데 있어서 다른 진단법들을 훌륭하게 보완한다. 수치적 예제들이 이러한 분산비율의 유용성을 정말로 잘 설명해주며, 다음은 이에 대한 질적인 기술이다:

높은 분산비율을 가진 회귀계수(최소한 2개 이상) 부분세트가 작은 고유값(a small eigenvalue, 심각한 선형의존성)을 동반하면, 그 부분세트 내 회귀계수들을 포함하는 의존성(dependency)를 의미하는 것이며, 이러한 의존성은 부분세트 내 계수들의 추정정밀도(precision of estimation)을 떨어뜨린다.

따라서 전체 진단은 여러가지 양들(quantities)를 함께 고려하여야 한다. 즉, 특정 의존성의 심각도(seriousness)를 평가하는데는 고유값(eigenvalue) 혹은 고유값의 비(ratio), 어떤 변수가 어느 정도로 의존성에 포함되는지 밝히는데는 분산비율(variance proportions), 개별 계수들의 손상을 판단하는데는 VIFs를 고려하여야 한다.

예제 8.1 Hald Data

Table 4.1의 Hald 자료를 생각해 보자. 3장에서 처음 주어지고 여기에서 반복되는 분산팽창요인(variance inflation factors)은 심각한 다중공선성을 의미한다. Table 8.1은 Hald 자료의 고유값들(eigenvalues)과 분산분해비율(variance decomposition proportions)이다. 분산비율은 중심화 되지 않은 자료(uncentered data)로 계산하였다. 따라서 5개의 고유값이 보고되어 있다.

고유값 중 하나는 심각한 의존성을 나타낸다. 가장 작은 고유값인 0.00006613815 ($\phi_j = 66289.2990$)는 회귀변수 x_1, x_2, x_3, x_4 의 계수들과 절편(intercept)에 심한 손상을 주는 의존성을 반영한다. 명백히, 이 의존성은 모든 회귀계수를 포함하는 것이다. 두 번째로 작은 고유값(0.0376383)의 영향(impact)은 고유값 비(eigenvalue ratio)로 볼 때 낮은 수치이다.

$$\frac{4.119699}{0.0376383} = 109.4550 < 1,000$$

이는 두 번째로 작은 고유값(second smallest eigenvalue)으로 특성화되는 의존성에는 모든 회귀계수들이 영향을 받지 않는다고 해석할 수 있음을 의미한다.

의존성의 중요도(degree of importance)는 최소 고유값(smallest eigenvalue)과 연관된 의존성의 정도(degree of dependency)와 일치하지는 않는다. 그러나 이 예제에서는 모든 회귀계수들이 최소 고유값에 의해서만 크게 영향을 받는 것을 볼 수 있고, 특히 절편의 영향은 100%이다. 절편의 추정 표준오차(estimated standard error)는 70.0710인데, 이는 정밀도가 형편없다는 표시임을 알아야 한다. 물론, 이 문제나 이 문제와 유사한 많은 문제들에서 상수항의 추정(estimation of the constant term)은 중요하지 않다는 것은 옳은 말인 것 같다. 회귀변수들(regressor variables)의 원점(origin)에서의 반응(response)은 본질적으로 거의 무의미하다.

정리해보면, 변수들 사이의 의존성은 모든 설명변수들의 계수와 절편 추정의 효율을 떨어뜨린다. 즉, 설명변수에 대한 계수의 추정은 의존성에 크게 영향을 받게 된다.

b_1, b_2, b_3, b_4 와 관련된 VIF 가 38.49621, 254.42317, 46.86839, 282.51286 임을 주목하기 바란다.

Table 8.1 Collinearity diagnostics for the Hald data

Coefficient	Parameter		Estimated		VIF (Regressors Centered)	
	Estimate	Standard Error	Estimate	Standard Error		
b_0	62.4054	70.0710				
b_1	1.5511	0.7448			38.49621	
b_2	0.5102	0.7238			254.42317	
b_3	0.1019	0.7547			46.86839	
b_4	-0.1441	0.7091			282.51286	
Variance proportions (regressors scaled, not centered)						
Eigenvalue	ϕ_j	b_0	b_1	b_2	b_3	b_4
4.119699	1.000	0.000	0.000	0.000	0.000	0.000
0.5538943	7.4377	0.000	0.010	0.000	0.003	0.000
0.288702	14.2697	0.000	0.001	0.000	0.002	0.002
0.0376383	109.4550	0.000	0.057	0.003	0.046	0.001
0.00006613815	62289.2990	1.000	0.932	0.997	0.950	0.997
Condition Number = 62289.2990						

다음은 예제 8.1에서 사용한 R code 이다.

```
hald<-read.table('c:/hald.txt',header=T)
attach(hald)
fit<-lm(y~x1+x2+x3+x4,hald)
summary(fit)
vif(fit)
x<-as.matrix(hald[,-1])
x_<-c(1,1,1,1,1,1,1,1,1,1,1,1)
x_1<-cbind(x_,x)
x1<-scale(x_1,scale=TRUE,center=FALSE)
e<-eigen(t(x1)%*%(x1)/12)
e$values
condition<-coldiag(hald[,-1],scale=TRUE,center=FALSE,add.intercept=TRUE)
condition
condindx<-condition$condindx^2
condindx
```

예제 8.2 Annual Data on Advertising, Promotions, Sales Expenses, and Sales (Millions of Dollars)

두 번째 진단례는 종속변수(dependent variable)로 판매량(S), 광고지출비(A), 판촉지출비(P), 판매비용(SE), 이전 기간의 광고지출비(A_1), 이전 기간의 판촉지출비(P_1) 모두를 통상적인 회귀변수로 하여, 이들이 판매량에 주는 효과를 보고자 한다. 회귀변수들의 역할을 판단하기 위한 목적으로 다중회귀(multiple regression)를 시행하였다. 이 조사를 시행한 회사에는 A, P, A_1, P_1의 합이 매 2년의 예산에서 5단위로 적절하게 고정되어야 한다는 규칙이 있었다.

즉, $A_t + P_t + A_{t-1} + P_{t-1} = 5$ 의 관계는 이 예제에서 다중공선성의 원인이 된다.

Table 8.3에서 공선성 진단(collinearity diagnostics)을 포함하는 회귀정보(regression information)가 주어져 있다. 진단은 상관행렬(correlation matrix), 상관행렬의 고유값(eigenvalues of the correlation matrix), 분산팽창요인(variance inflation factors), 분산분해비율(variance decomposition proportions)을 포함하고 있다. 중심화 자료(centered data)와 비중심화 자료(non-centered data) 모두를 고유값 분석(eigenvalue analysis)에 포함시켰다.

다중공선성 진단은 심각한 선형의존성을 보여준다. 이 결과들을 논의하기 전에, 회귀변수들(regressor variables)의 역할에 대한 분석가의 지식을 보자. 광고지출비, 판촉지출비, 판매비용, 이전 기간의 광고지출비, 이전 기간의 판촉지출비의 회귀계수들은 양수일 것으로 판단되며, 이전 기간의 광고지출비와 판촉지출비의 회귀계수는 현재의 광고지출비와

판촉지출비의 회귀계수의 중요성보다 작을 것이라고 생각된다.

자료에 대한 모형의 적합은 좋음을 알 수 있다(adjusted $R^2 = 0.8909$). 그러나 회귀계수들의 해석에 관한 문제는 중요한 일이고, 광고지출비(A), 판촉지출비(P), 이전 기간의 광고지출비(A_1), 이전 기간의 판촉지출비(P_1)의 분산팽창인자(variance inflation factors)가 25.9에서 43.52 사이의 큰 값을 갖는다. 이는 4개의 변수 사이에 강한 선형의존성(linear dependencies)이 있음을 나타내며, 판매비용(SE)의 VIF는 1.04로 이 변수가 다른 4개의 변수와 연관성이 존재하지 않음을 나타낸다.

분산비율을 보면 각 회귀계수의 분산이 상태지표(condition number)가 최대가 되는 최소고유값 0.007271377(중심화 자료의 결과)을 가지는 경우 0.985, 0.983, 0.012, 0.973, 0.989의 비율로 설명되므로, 네 설명변수 A, P, A_1, P_1의 공선성이 존재한다고 판단된다.(표 8.3 참고)

비중심화 자료의 고유값들과 분산비율들의 목록이 나열되어 있다. 이것은 중심화 자료와 비교하여 결과가 어떤지 알아보는데 도움이 된다. 고유값들이 중심화 자료의 대응 부분들과 다른점에 주목하자. 또한 비중심화 자료의 경우에는 상수항이, 심각한 고유값(serious eigenvalues)으로 기술되는 공선성의 한 부분에 포함된다는 것을 주목하자.

가장 심각한 의존성은 중심화와 비중심화 분석 모두에서 A, P, A_1, P_1를 포함하는 것으로 보인다. 그러나 이 상황에서는, 비중심화 자료의 진단이 중심화 자료의 진단과 상충될 수 있다. 이 경우 한가지 근본적인 차이는 상수항의 포함 여부이다. 중심화를 하면, 상수 추정값이 나머지 계수 추정값들(estimated coefficients)에 대하여 직교(orthogonal)가 가능해진다. 상수항을 포함하는 것이 중요한지 아닌지는, 절편(intercept)에 대한 분석가의 관심에 달려있다. 이 경우 절편보다는 계수들이 중요하다. 따라서, 중점은 중심화 자료의 진단(centered diagnostics)에 두어야 한다. 다음 절에서 중심화와 척도화에 관해 더 논의할 것이다.

Table 8.2 Annual Data on Advertising, Promotions, Sales Expenses, and Sales (Millions of Dollars)

A(x1)	P(x2)	SE(x3)	A_1(x4)	P_1(x5)	S(y)
1.98786	1	0.3	2.01722	0	20.11371
1.94418	0	0.3	1.98786	1	15.10439
2.19954	0.8	0.35	1.94418	0	18.68375
2.00107	0	0.35	2.19954	0.8	16.05173
1.69292	1.3	0.3	2.00107	0	21.30101
1.74334	0.3	0.32	1.69292	1.3	17.85004
2.06907	1	0.31	1.74331	0.3	18.87558
1.01709	1	0.41	2.06907	1	21.26599
2.01906	0.9	0.45	1.01709	1	20.48473
1.06139	1	0.45	2.01906	0.9	20.54032
1.45999	1.5	0.5	1.06139	1	26.18441
1.87511	0	0.6	1.45999	1.5	21.71606

2.27109	0.8	0.65	1.87511	0	28.69595
1.11191	1	0.65	2.27109	0.8	25.8372
1.77407	1.2	0.65	1.11191	1	29.31987
0.95878	1	0.65	1.77407	1.2	24.19041
1.9893	1	0.62	0.95878	1	26.58966
1.97111	0	0.6	1.9893	1	22.24466
2.26603	0.7	0.6	1.97111	0	24.79941
1.98346	0.1	0.61	2.26603	0.7	21.19105
2.10054	1	0.6	1.98346	0.1	26.03441
1.06815	1	0.58	2.10054	1	27.39304

TABLE 8.3 Multicollinearity diagnostics and regression analysis

Regression		VIF			
	Coefficients	(Data centered)	Standard Error	t	Prob > t
b_0	-14.194		18.715	-0.758	0.4592
b_1	5.361	36.941513	4.028	1.331	0.2019
b_2	8.372	33.473514	3.586	2.334	0.0329
b_3	22.521	1.075962	2.142	10.512	1.36e-08
b_4	3.855	25.915651	3.578	1.077	0.2973
b_5	4.125	43.520965	3.895	1.059	0.3053

<i>Standard regression analysis</i>					
Source	SS	df	MS	F	
Regression	307.5716	5	61.5143	35.30	
Residual	27.879	16	1.742		
Total	335.4503	21			

$R^2 = 0.9169,$
$Adjusted R^2 = 0.8909$

<i>Correlation matrix</i>					
	x_1	x_2	x_3	x_4	x_5
x_1	1.000	-0.357	-0.129	-0.140	-0.496
x_2	-0.357	1.000	0.063	-0.316	-0.296
x_3	-0.129	0.063	1.000	-0.166	0.208
x_4	-0.140	-0.316	-0.166	1.000	-0.358
x_5	-0.496	-0.296	0.208	-0.358	1.000

TABLE 8.3 Multicollinearity diagnostics and regression analysis (continued)

<i>Eigenvalues of $\mathbf{X}'\mathbf{X}$ (regressors scaled but not centered) and variance decomposition proportions</i>					
Eigenvalues	ϕ	b_0	b_1	b_2	b_3
5.2810014467	1.000	0.000	0.000	0.000	0.002
0.3797635420	13.90602536	0.000	0.000	0.007	0.000
0.2272296116	23.2708153	0.000	0.001	0.016	0.000
0.0600516997	87.94091545	0.000	0.005	0.000	0.291
0.0517834509	101.9824163	0.000	0.008	0.003	0.703
0.0001702492	31019.24383	1.000	0.985	0.973	0.003

Eigenvalues	b_4	b_5
5.2810014467	0.000	0.000
0.3797635420	0.000	0.012
0.2272296116	0.001	0.005
0.0600516997	0.016	0.001
0.0517834509	0.002	0.005
0.0001702492	0.980	0.977

Condition Number = 31019.24383

TABLE 8.3 Multicollinearity diagnostics and regression analysis

<i>Eigenvalues of $\mathbf{X}'\mathbf{X}$ (regressors scaled and centered)</i> <i>and variance decomposition proportions</i>					
Eigenvalues	ϕ	b_0	b_1	b_2	b_3
1.700954740	1	0.000	0.005	0.001	0.083
1.288206906	1.320404923	0.000	0.000	0.016	0.000
1.144651635	1.486002106	0.000	0.011	0.001	0.038
1.000000000	1.700954740	1.000	0.000	0.000	0.000
0.858915342	1.980352029	0.000	0.000	0.000	0.867
0.007271377	233.9247073	0.000	0.985	0.983	0.012

Eigenvalues	b_4	b_5
1.700954740	0.004	0.005
1.288206906	0.002	0.004
1.144651635	0.016	0.000
1.000000000	0.000	0.000
0.858915342	0.005	0.002
0.007271377	0.973	0.989

Condition Number = 233.9247073

고유값(eigenvalues), 고유벡터(eigenvectors), 변수팽창요인(variance inflation factors), 분산분해비율(variance decomposition proportions) 등을 사용하였기 때문에, 일부 분석가들은 이러한 복잡한 도구모음(set of tool)이 익숙하지 않을 것이다. 그러나 분석도구(analytical tools)를 사용하는 사람들이 모든 진단자원(diagnostic resource)을 최대한 이용하도록 하는 것이 중요하다. 특정 응용분야에서 진단을 해본 경험이 쌓이면, 이러한 방법(devices)들에 더 익숙해질 것이다.

다음은 예제 8.2에서 사용한 R code이다.

```
ad<-read.table('c:/advertising.txt',header=T)
attach(ad)
fit<-lm(y~x1+x2+x3+x4+x5,ad)
summary(fit)
vif(fit)
anova(fit)
x<-as.matrix(ad[,-6])
newx<-scale(x,scale=TRUE,center=TRUE)
```

```
round(cor(newx),3)
x_<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
x_1<-cbind(x_,x)
x1<-scale(x_1,scale=TRUE,center=FALSE)
e<-eigen(t(x1)%*%(x1)/21)
e$values
condition<-coldiag(ad[,-6],scale=TRUE,center=FALSE,add.intercept=TRUE)
condition
condindx<-condition$condindx^2
condindx
e_<-eigen(t(newx)%*%(newx)/21)
e_$values
condition_<-coldiag(ad[,-6],scale=TRUE,center=TRUE,add.intercept=TRUE)
condition_
condindx_<-condition_$condindx^2
condindx_
```

8.3. 다중공선성에 관한 추가 주제(Further Topics Concerning Multicollinearity)

예측(Prediction)

모형구축과정(model-building task)에서 다중공선성(multicollinearity)이 심각한 문제가 되는 상황에 직면한 자료분석가의 딜레마를 생각해보자.

지금까지, 심각한 다중공선성(multicollinearity)이 있으면 최소제곱추정값(least squares estimate)이 얼마나 나빠지는지(deterioration) 판단하는데 사용 가능한 진단정보(diagnostic information)를 전달하고자 하였다. 다중공선성(multicollinearity)이 계수들(coefficients)에 미치는 영향(influence)을 평가하기란 간단하다. 분산팽창요인(variance inflation factor)이나 고유값체계(eigensystem)를 검사(inspection)하면, OLS 추정값(estimate)이 불안정한지, 분산(variance)이 과도하게 큰지 파악할 수 있다. 그러나 3, 4장에서 살펴본 바와 같이 예측값(prediction)에 대한 영향(impact)을 파악하는데 진단도구(diagnostics)를 쉽게 이용할 수는 없다. 이상적인 단일예측기준(single prediction norm)을 분리해내기도 어렵고, 공선성(collinearity) 때문에 예측(prediction)이 손상되었다고 말할 수 있기도 어렵다. 그 이유는 예측의 질(quality of fit)을 나타내는 $\hat{y}(x_0)$ 는, x_0 가 회귀공간(regressor space) 내에 있는 위치에 따라 달라지기 때문이다. 이 말은 다중공선성(multicollinearity)이 있을 경우 더욱 중요하다. 최소제곱모형(least squares model)의 적합이 만족스럽다면, 예측이 잘 맞는 x 의 영역(region)이 있겠지만, 심각한 다중공선성(multicollinearity)이 있을 경우에는 예측이 잘 맞지 않는 영역이 있기 쉽상이다. 또한, 다중공선성이 있는 경우 외삽(extrapolation)할 때 매우 신중해야 한다. 따라서 모형구축 과정(model-building procedure)에서 예측의 질(quality of prediction)을 어느 정도 알 수 있게 해주어서, OLS를 사용할지 결정하는데 영향을 주는 진단도구(diagnostics)는 예측값 분산(prediction variance) 혹은 예측값의 표준오차(standard error of prediction)이다. 그러나 불행하게도, 분석가는 진단도구(diagnostics)를 계산(computation)하는 과정에서 특정 x_0 를 결정한 후 사용하여야 한다.

$$x_0 = \begin{bmatrix} 1 \\ x_{1,0} \\ x_{2,0} \\ \vdots \\ x_{k,0} \end{bmatrix}$$

위의 벡터는 예측(prediction)과 외삽(extrapolation)이 필요할 때, 모형(model)과 k 개의 회귀변수 수준(level) 모두를 나타낸다는 것을 상기하라. 그 다음에, 진단도구(diagnostic)를 이용하여 다음을 계산한다.

$$\frac{Var \hat{y}(x_0)}{\sigma^2} = x_0' (X'X)^{-1} x_0$$

3장에서 논의한 신뢰구간(confidence interval)에서의 예측값 분산(prediction variance)과 4장에서 언급한 C_p 통계량(statistic)의 역할(role)을 독자들은 상기하여야 한다. x_0 의 초기값인 1.0은 모형의 상수항(constant term)에 해당한다. $(X'X)^{-1}$ 에 관한 한, 회귀자료(regressor data)는 중심화(centering)와 척도화(scaling)가 모두 될 수도 있고, 단순히 척도화만 되거나 혹은 모두 안될 수도 있다. 회귀변수(regressor variable)가 척도(scale)나 위치변화(location change)를 하여도 예측값 분산(prediction variance)은 변하지 않는다. 회귀자료(regressor data)가 척도화(scaled)는 되었지만 중심화(centered)가 되지 않았다고 가정하면, $(X'X)$ 는 $(k+1) \times (k+1)$ 행렬이 될 것이다. 물론, $x_{1,0}, x_{2,0}, \dots, x_{k,0}$ 값을 유사하게 척도화하여야 한다. 이러한 접근방식(approach)에서, 분석가가 변동(variation)을 구할 때 다음 식을 이용하는데, 그 이유는 이 식이 신뢰구간범위(confidence bound) 해석에 이용될 수 있어 다소 편안하게 느껴지기 때문이다.

$$s_{\hat{y}(x_0)} = s \sqrt{x_0' (X'X)^{-1} x_0}$$

일부 소프트웨어 회귀 패키지(software regression package)는 자료세트(data set)의 일부가 아닌 회귀변수조합(regressor combination)들에 대한 $x_0' (X'X)^{-1} x_0$ 의 계산을 쉽게 할 수 있도록 해준다. 확실히, 이러한 양(quantity)에 관심이 있는 철저한 연구자는 가능한 한 많은 x_0 가 사용될 수 있는 진단도구(diagnostic)를 관찰함으로써 좀 더 많은 정보를 얻을 수 있을 것이다. 예측을 할 때 이러한 진단도구(diagnostic)를 이용하면 “예” 혹은 “아니오”的 대답을 얻을 수 있는 것이 아니라, 예측값이 나쁜 영역(region)을 분리해 내는(isolate) 과정에 이용할 수 있다. 자료 포인트 x_0 가, 자료에 의하여 특성화되는 다중공선성과 밀접하게 결합되어 있거나(consistent) 일치한다면(in line), 예측값(prediction)은 심각하게 영향을 받지 않을 것임을 직관적으로 알 수 있다. 따라서 적합의 질이 좋다는 가정하에서 적합값(fitted value) $\hat{y}(x_0)$ ($i = 1, 2, \dots, n$)는 일반적으로 나쁜 추정값(estimate)은 아닐 것이다. 자료에서 경험하였던 것과 유사한 선형 의존성(linear dependency)을 보이는, 자료 범위(data range) 내의 어떠한 조합, x_0 에서도 타당하게 좋은 예측값(prediction)이 나올 것이다.

Fig. 3.3은 이러한 상황(situation)을 설명하고 있다. 말뚝울타리(picket fence) 위에 누워서 균형을 유지하고 있는 평면이, 말뚝울타리에 수직방향(perpendicular)으로 위치할 때에는 안정성(stability)이 매우 나빠질 것이다. 따라서 말뚝울타리 위에 누워 있을 때, \hat{y} 와 이에 대응하는 적합값 $\hat{y}(x_i)$ 의 정밀도(precision)가 가장 좋을 것으로 기대할 수 있다. 반면에

x_0 가 수직방향으로 울타리에서 타당한 거리만큼 떨어져 있다면, $\hat{y}(x_0)$ 의 분산(variance)은 커질 것이다.

결과적으로, 만약 모집단 다중공선성(population multicollinearity)이 존재하고, 이것이 자료를 잘 묘사한다면, 다중공선성(multicollinearity)은 자료 영역(data region)내의 x_0 에서 예측값(prediction)에 심각한 영향을 주지는 않을 것이다. 그러나 독자들은 이로부터 지나친 위안을 얻어서는 안된다. 실제 문제(real-life problem)에서는 집단 다중공선성(population multicollinearity)이 거의 잘 정의되지 않고 있다. 많은 경우, 자료 내에 의존성(dependencies in the data)이 있다면 단지 특정 자료세트(data set)의 특성(characteristics) 또는 “개성(personality)”으로 드러날 것이다. 미래 예측값(future prediction)에 대한 x_0 가 자료와 동일한 양상(complexion) 즉, 마치 자료 내에 있는 것처럼 x_0 들이 함께 움직이지는 않는다면, x_0 에서의 예측값은 심하게 손상(damage) 받을 수 있다. 다음의 예제 8.3은 좋은 설명이 되는 실제의 보기이다.

자료 포인트 x_0 에서 예측을 한다고 가정하자. 중심대각(main diagonal)에 고유 값(eigenvalues), $X'X$ 를 가지는 대각행렬(diagonal matrix)을 Λ 라고 표기하자(식(8.2) 참고). 그렇다면 다음과 같은 식을 얻을 수 있는데,

$$\frac{Var \hat{y}(x_0)}{\sigma^2} = x_0' V \Lambda^{-1} V' x_0 = z_0' \Lambda^{-1} z_0 = \sum_{i=0}^k \frac{z_{0,i}^2}{\lambda_i}$$

여기에서 $z_0 = V' x_0$ 이다. 만약 다중공선성(multicollinearity)이 심각한 문제라면, 최소한 하나의 $\lambda_i \approx 0$ 는 잠재적으로(potentially) 분산(variance)이 커질 것이다. 그러나 독자들은 3장에서 만약 $\lambda_i \approx 0$ 이면, $Xv_i \approx 0$ 는, λ_i 에 의하여 특성화되는 근사선형의존성(near linear dependency)을 나타냄을 기억하여야 한다. $z_{0,i}$ 가 z_0 에 대응하는 원소(corresponding element)라고 가정해보면 이 원소는 $z_{0,i} = x_0' v_i$ 이다. 따라서 만약 X 의 근사선형의존성(near linear dependency)이 x_0 의 원소들(elements)의 특징이라면, $z_{0,i} \approx 0$ 이고 $\frac{1}{\lambda_i}$ 은 매우 큰 값(value)을 갖는다. 그러나 $x_0' v_i$ 가 0에 근접하지 않는다면(not near zero) 즉, 표본 다중공선성(sample multicollinearity)과 밀접하게 결합되어(consistent) 있지 않다면, 예측값 분산(prediction variance)은 커질 것이다. 후자의 조건은 Fig. 3.3에서 x_0 가 말뚝울타리(picket fence) 위에 있지 않거나 혹은 근처에 있지 않을 경우의 상황(situation)을 기술한다.

예제 8.3 토양자료

실제의 보기(real-life example)를 이용하여 다중공선성이 모형의 예측값(prediction)을 얼마나 나쁘게 하는지 살펴보자. 독자들은 회귀변수 수준(regressor level)의 위치 x_0 에 따라서 예측능력(prediction capability)이 감소할 수 있다는 것을 알게 되었다.

만약 x_0 가 “다중공선성의 주류(mainstream of multicollinearity)”에서 다소 떨어져 있는 지점이라면 예측값(prediction)은 나빠질 수 있다. Table 8.4의 자료들은 토양으로부터 증발되는 수분의 양을 나타내고 있다. 증발되는 수분의 양을 결정할 것으로 예상되는 x_1 (토양 내 최고온도), x_2 (토양 내 최저온도), x_3 (토양 내 평균온도), x_4 (최고기온), x_5 (최저기온), x_6 (평균기온)을 사용하여 회귀분석을 실시하였다. 우리는 $x_1, x_2, x_3, x_4, x_5, x_6$ 을 가지는 회귀식을 구하기 위하여 최소제곱법(least square procedure)을 이용하였다. 결과는 다음과 같다.

$$\hat{y} = -80.8871 + 1.7080x_1 - 1.5630x_2 - 0.1088x_3 + 0.7651x_4 - 0.4853x_5 + 0.3409x_6$$

여기서 $R^2 = 0.81680$ 이며 $s = 7.1140$ 이다. 또한, Table 8.4의 모자대각값(HAT diagonal values)들과 예측값(prediction)의 표준오차(standard error of prediction)들을 주목하기 바란다.

Table 8.5에 이러한 자료들의 회귀 진단값(regression diagnostics)을 표시하였다. 고유값의 스펙트럼(spectrum of eigenvalue)으로 보면, 심각한 문제(serious problem)로 인식될 만큼 충분히 크지는 않은, 다음과 같은 조건수(condition number)가 산출되기는 하나,

$$\phi = \frac{6.986511}{0.000032298} = 216314.044$$

분산팽창요인(VIF)을 보면 모든 회귀계수들 사이의 다중공선성을 의심할 수 있다.

b_1, b_2 와 관련된 분산팽창요인(VIF)의 많은 부분이 회귀변수 x_1, x_2 가 관련된 의존성(dependency)이 원인인 된다는 것을, 분산비율(variance proportion)을 보면 알 수 있다.

모형의 예측능력(predictive capability)에 대하여 약간의 통찰(insight)을 얻기 위하여, $\frac{Var \hat{y}(x_0)}{\sigma^2} = x'_0(X'X)^{-1}x_0$ 의 값을 추가된 자료로부터 계산하였다. 또한, 예측값의

표준오차(standard error of prediction) 값 $s\sqrt{x'_0(X'X)^{-1}x_0}$ 를 계산하였고, 자료세트(data set) 내 대응되는 증발된 수분량들과 비교할 수 있었다. 추가된 자료에 대한 이러한 결과들은 Table 8.6에 나타나 있다. 추가된 6개의 자료들은 이 자료가 추가되기 전에 회귀분석을 했던 자료들과 유사하다. 이러한 자료들 중 둘, 혹은 세 곳은 외삽(extrapolation)이 되지 않은 지점을 나타내고, 예측값의 표준오차(standard error of prediction)들은 이전 19개의 자료세트(data set)에서의 표준오차보다 상당히 부풀려져 있다(swelled). 20, 21, 24에 얻어진

증발된 수분량은 외삽(extrapolation)을 통하여 얻어진 것이다. 또한 토양 내 최고온도(x_1)와 토양 내 최저온도(x_2) 간에는 자료상에서 공선성이 존재하지 않는 것을 알 수 있다. 따라서, 예측능력(prediction capability)은 상대적으로 어마어마하게 큰, 예측값의 표준오차(standard error of prediction) 값으로 정량화(quatified)된다.

한가지 매우 명백한 결론에 이르게 된다. 다중공선성(multicollinearity)이 모형의 예측능력(prediction capability)을 감소시킬 수 있고 실제로 감소시킨다는 것을 예측값의 분산(prediction variance)(σ^2 은 별개로 하고)과 예측값의 표준오차(standard error of prediction)로 알 수 있다. 여기에 사용된 것과 같은 예제들은 드문 것이 아니다. 이전에 권고하였던 부분을 여기에서 한번 더 강조하는 것이다. 특히 다중공선성(multicollinearity)이 문제가 될 때, 예측값의 표준오차(standard error of prediction)는 예측능력(prediction capability)을 추정하는 지침(guide)으로 사용되어야만 한다.

Table 8.4 토양 자료

Day	y	x_1	x_2	x_3	x_4	x_5	x_6	$h_{ii} = \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$	$S_{\hat{y}(x)}$
1	30	84	65	147	85	59	151	0.434	4.685723
2	34	84	65	149	86	61	159	0.378	4.372292
3	33	79	66	142	83	64	152	0.133	2.596062
4	26	81	67	147	83	65	158	0.195	3.143116
5	41	84	68	167	88	69	180	0.546	5.256850
6	4	74	66	131	77	67	147	0.438	4.710231
7	5	73	66	131	78	69	159	0.480	4.930646
8	20	75	67	134	84	68	159	0.637	5.678655
9	31	84	68	161	89	71	195	0.483	4.943153
10	38	86	72	169	91	76	206	0.308	3.947486
11	43	88	73	178	91	76	208	0.304	3.923308
12	47	90	74	187	94	76	2211	0.296	3.871540
13	45	88	72	171	94	75	211	0.453	4.788242
14	45	88	72	171	92	70	201	0.542	5.235195
15	11	81	69	154	87	68	167	0.254	3.588718
16	10	79	68	149	83	68	162	0.133	2.596465
17	30	84	69	160	87	66	173	0.253	3.574934
18	29	84	70	160	87	68	177	0.193	3.122964
19	23	84	70	168	88	70	169	0.540	5.225466

Table 8.5 Collinearity diagnostics for data of Table 8.4

	Coefficient	VIF				
	b_1	48.38798				
	b_2	11.74082				
	b_3	33.81712				
	b_4	14.10343				
	b_5	29.68558				
	b_6	31.88631				
Eigenvalue	b_0	b_1	b_2	b_4	b_5	b_6
6.986511	0.000	0.000	0.000	0.000	0.000	0.000
0.009227	0.000	0.001	0.004	0.000	0.000	0.023
0.0032360	0.004	0.001	0.018	0.003	0.021	0.010
0.00074578	0.004	0.006	0.160	0.025	0.027	0.156
0.00014532	0.067	0.001	0.001	0.963	0.012	0.139
0.00010188	0.001	0.968	0.080	0.008	0.142	0.012
0.000032298	0.924	0.024	0.738	0.000	0.797	0.660

Table 8.6 Standard errors of prediction for six additional day

Day	x_1	x_2	x_3	x_4	x_5	x_6	$x^T_0(X'X)^{-1}x_0$	$S_{\hat{y}(x)}$
20	77	67	147	83	66	170	2.0815348	10.263741
21	87	67	166	92	67	196	1.0610546	7.327954
22	89	69	171	92	72	199	0.9991577	7.111003
23	89	72	180	94	72	204	0.3417002	4.158498
24	93	72	186	92	73	201	1.3688929	8.323358
25	93	74	188	93	72	206	0.7054046	5.974932

다음은 예제 8.3에서 사용한 R code이다.

```

soil<-read.table('c:/msoil.txt',header=T)
fit<-lm(y~x1+x2+x3+x4+x5+x6,soil)
summary(fit)
inf.m<-influence.measures(fit)
inf.m$infmat
s<-7.114*(inf.m$infmat[,11])^0.5
vif(fit)
x<-as.matrix(soil[,-7])

```

```

x_<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
x_1<-cbind(x_,x)
x1<-scale(x_1,scale=TRUE,center=FALSE)
e<-eigen(1/18*t(x1)%%x1)
e$values
condition<-colldiag(soil[,-7],scale=TRUE,center=FALSE,add.intercept=TRUE)
condition
condindx<-condition$condindx^2
condindx
x20<-c(1,77,67,147,83,66,170)
x21<-c(1,87,67,166,92,67,196)
x22<-c(1,89,69,171,92,72,199)
x23<-c(1,89,72,180,94,72,204)
x24<-c(1,93,72,186,92,73,201)
x25<-c(1,93,74,188,93,72,206)
h20<-t(x20)%%solve(crossprod(x_1))%%x20
h21<-t(x21)%%solve(crossprod(x_1))%%x21
h22<-t(x22)%%solve(crossprod(x_1))%%x22
h23<-t(x23)%%solve(crossprod(x_1))%%x23
h24<-t(x24)%%solve(crossprod(x_1))%%x24
h25<-t(x25)%%solve(crossprod(x_1))%%x25
hat<-rbind(h20,h21,h22,h23,h24,h25)
s20<-7.114 *sqrt(h20)
s21<-7.114 *sqrt(h21)
s22<-7.114 *sqrt(h22)
s23<-7.114 *sqrt(h23)
s24<-7.114 *sqrt(h24)
s25<-7.114 *sqrt(h25)
s_<-rbind(s20,s21,s22,s23,s24,s25)
result<-cbind(hat,s_)
result

```

자료의 중심화와 척도화에 대한 추가 설명(*More on Centering and Scaling Data*)

표준다중회귀모형(standard multiple regression model)은 다음과 같다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (8.6)$$

중심화(centering)와 척도화(scaling)의 과정을 통하여 대체식(alternative formulation)을 얻을 수 있다.

$$y = \beta'_0 + \beta'_1 \left[\frac{x_1 - \bar{x}_1}{S_1} \right] + \beta'_2 \left[\frac{x_2 - \bar{x}_2}{S_2} \right] + \cdots + \beta'_k \left[\frac{x_k - \bar{x}_k}{S_k} \right] + \varepsilon \quad (8.7)$$

3장에서 논의된 바와 같이, 이렇게 다시 쓰거나(rewriting) 재모수화(reparameterization)하면 확실한 이점이 있기 때문에 오랜 기간 동안 적합한 절차로 인식되고 있다. 엄밀히 말해서, 두 모형은 동등(equivalent)하다. 중심척도화 모형(centered and scaled model)에서 절편 β_0 는 β'_0 로 나타낼 수 있다.

$$\beta'_0 = \beta_0 + \beta'_1 \frac{\bar{x}_1}{S_1} + \beta'_2 \frac{\bar{x}_2}{S_2} + \cdots + \beta'_k \frac{\bar{x}_k}{S_k} \quad (8.8)$$

공선성(collinearity)을 평가하는 것과는 별개로, 중심화(centering)와 척도화(scaling)의 가장 잘 알려진 이점(advantage)은, 컴퓨터의 기억장치(storage) 혹은 정밀도(precision)가 낮을 때의 어려움들을 없애준다는 것이다. 그러나, 여기에서 논의될 요점은 다음과 같다. (a) 중심화(centering)와 척도화(scaling)는 전통적인 회귀분석(classical regression analysis)에서 어떤 역할을 하는가? (b) 중심화(centering)의 공선성 진단도구(collinearity of diagnostic)에 영향을 주는 것은 무엇인가?

표준회귀분석(Standard Regression Analysis)

식(8.7)의 중심척도화 모형(centered and scaled model)을 적합하기 위하여 자료세트(data set)를 이용하면, 우리는 식(8.6)의 모형에서 추정된 계수(estimated coefficients)를 얻을 수 있다. 중심화(centered)되고 척도화(scaled)된 모형의 i 번째 추정계수(estimated coefficient)를 S_i 으로 나눔으로써 (8.6)의 원래 변수(natural variables)의 계수들을 얻을 수 있다. 따라서 상수항(constant term) β_0 는 계산에 의해서 추정값(estimate), b'_0 로부터 추정된다.

$$b_0 = b'_0 - \frac{b'_1 \bar{x}_1}{S_1} - \frac{b'_2 \bar{x}_2}{S_2} - \cdots - \frac{b'_k \bar{x}_k}{S_k}$$

여기에서 b'_i 는, 식(8.7)의 중심척도화 모형(scaled and centered model)에서 얻은 추정값(estimate)이다. 게다가, $b'_0 = \bar{y}$ 이며, 이는 y 의 평균이다. 따라서 분석하는 사람은 분석에 사용된 모형이 어떤 것이든지 항상 하나의 모형식(one model formulation)에서 다른

것으로 이동할 수 있다. t 검정(t -test), 계수의 표준오차(standard error of coefficient) 등의 결과보고에 관한 한, 어떤 모형이 가장 의미가 있을까? 첫째, 예측된 반응(predicted response)과 관련된 \hat{y} , 잔차(residual), s^2 , PRESS, C_p , $s_{\hat{y}}$ 와 같은 일부 통계량(statistic)은 식(8.6)과 (8.7)의 모형에서 동일하다. 따라서 적합(fit)을 단순히 예측(prediction)만을 목적으로 한다면, 모형식(model formulation)은 (8.6) 혹은 (8.7)이 될 수도 있다. 상수항은 별개로 하고 회귀계수(regression coefficient)들에 대한 t 검정은 두 식에서 같을 것이다. i 번째 회귀변수(regressor)의 계수(coefficient)와 표준오차(standard error)는 척도화 상수(scale constant)인 S_i 인자(factor)에 의하여 편위될(deviated) 것이다. 따라서 S_i 는 비(ratio) 즉, t 통계량(t -statistic)에서 소거된다(cancel).

그러나, 회귀계수(regression coefficient)의 해석이나 개별적 회귀변수(individual regressor variables)의 역할에 관한 정보를 추출(extract)해 내는 것을 분석가의 목적이라고 가정해보자. 계수들(coefficients) 혹은 변화율(rate of change)이 두 가지 모형식(model formulation)에서 동일하게 해석(interpretation)되지 않을 것이라는 것은 명확하다. 결과적으로, 모형식을 선택하는 것은 종종 실제상황(real-life situation)의 유형(type)에 좌우된다. 이것을 다음 예제에서 살펴보자.

예제 8.4 Designed Experiment

어떤 공정에서 생산되는 제품의 강도(kg/cm^2)가 공정과정 중의 온도(x_1)와 압력(x_2)의 영향을 많이 받는 것으로 알려져 있다. 따라서 두 개의 회귀변수(regressor variable)인 온도(x_1)와 압력(x_2)이 반응(y) 즉, 제품의 강도에 미치는 영향에 관한 실험을 한다고 가정하자. 연구자는 처음에 실험하기에 알맞은 온도와 압력의 범위를 알 것이다. 예를 들어서, 실험이 다음과 같이 시행된다고 하자.

$x_1 (\text{ }^\circ\text{C})$	$x_2 (\text{psi})$
190	75
190	85
210	75
210	85

연구자는 의도적으로 직사각형의 실험영역(rectangular experimental region)을 구성하였는데, 이는 회귀변수들의 함수(function of the regressor variable)를 연구해야 하는 영역이기 때문이다. 관심이 단지 이 영역에서의 예측(prediction)에만 있다고 한다면, 원래의 선형모형(natural linear model) 혹은 중심척도화 선형모형(centralized and scaled linear model)이 적절(appropriate) 할 것이다. 그리고 사실 예측값의 결과(prediction results)는 두 가지 경우에서 동일하다. 그러나, 추정된 상수인 절편(intercept)과 두 기울기 계수(slope coefficients)로부터 어떤 해석을 이끌어

내고 싶다면, 선택(selection)을 해야 한다. 이 경우 화학자는, 20도의 변화는 대략 10 psi의 압력변화와 중요도 면에서 동일하다는 예비 판단(preliminary judgment)을 하였다. 결과적으로 우리는 모형의 정의(model definition)가 변화량(change)을 반영할 것이라고 기대할 수 있다. 그리고 의미 있는 회귀계수(regression coefficients)는 10 psi의 압력변화가 20도의 온도변화에 동등하도록 정의한 표준화된 변수들(standardized variables)에 토대를 두어야 한다. 다음은 중심척도화의 결과이다.

Variable 1	Variable 2
- 1/2	- 1/2
-1/2	1/2
1/2	-1/2
1/2	1/2

또는, \bar{x}_1, \bar{x}_2 주변으로 중심화(centering)하고, 온도(x_1)에 대해서 10도, 압력(x_2)에 대해서 5 psi로 척도화(scaling)한 표준 방식(standard manner)으로 쓰면 다음과 같다.

Variable 1	Variable 2
-1	-1
-1	1
1	-1
1	1

이제 계수들(coefficients)을 진짜로 해석할 수 있게 되었다. 기울기 b'_1 과 b'_2 는 표준화된 온도와 압력의 단위 변화(standardized unit change) 당 강도의 변화이다. 이러한 수들(numbers)은 실제로 의미가 있다. 또한 원래 변수 모형(natural variable model)에서의 절편 β_0 주위로 전개(evolve)할 이유는 없을 것이며, 0도와 0 psi에서의 추정강도(estimated strength)에는 아무도 관심을 갖지 않을 것이다. 원래 변수들(natural variables)의 참 원점(true origin)은 흥미롭지도 않고, 실제로 중요하지도 않지만 대신에 중심척도화 모형(centered and scaled model)에서 절편 β'_0 는 상당히 흥미를 끌 수 있다. 추정값 $b'_0 = \bar{y}$ 는 직사각형영역(rectangular region)의 중심(center)에서 추정된 반응(estimated response)을 나타내며, 표준화된 변수(standardized variable)에서의 원점(0,0)은 매우 중요할 가능성이 있다.

예제 8.4의 중심척도화 모형(scaled and centered model)에서의 결과들은 원래 변수모형(natural variable model)에서 보다 해석하기 쉽게 되어 있다. 그러나 원래 변수들(natural variables)의 계수들이 유일하게 해석 가능한 계수이고, 원래 변수들(natural variables)을 정의하여야 상수항(constant term)을 이해(make sense)할 수 있는 특수한 상황(예를

들어 $x_1=0, x_2=0, \dots, x_k=0$ 일 때의 반응값(value of the response)이 원래 원점(natural origin)인 경우)이 존재한다. 사회과학에서 특히 경제학에서 이러한 계수의 “구조적인 해석가능성(structural interpretability)”이 중요할 수 있다. 또한 회귀변수평균(regression average) $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ 이 원점(origin)으로 적절하지 않을 수도 있다. 이러한 수준들(levels)에서는 특별하다거나 또는 과학적으로 흥미로운 것이 없다. 만약 새로운 기관(site)에서 표본을 얻는다면, 이 평균들은 변할 것이다. 한편으로 예제 8.4에서는 수준들(levels)을 조절하여 구조적으로(structurally, 혹은 화학적으로) 의미있는(sensible) 중심(center)과 이전 고찰로 알게 된 척도(scaling)를 얻을 수 있었다.

중심화와 공선성진단(Centering and Collinearity Diagnostics)

표준회귀분석(standard regression analysis)에서 중심화(centering)와 척도화(scaling)의 역할에 관하여 앞서 언급한 것은 매우 기초적인 것으로 독자 여러분은 받아들여야 할 것이다. 그러나 어떤 개별적인 실제상황(individual practical situation)에서는, 인용된(quoted) 회귀식(regression equation)을 표준화된 회귀변수(standardized regressor)의 원래 변수(natural variable) 형태로 만들어야 하는지는 명확하지 않다. 분석하려는 자료의 특성을 잘 아는 과학자들의 지식과 경험이 도움이 될 것이다.

명확하지는 않을지라도, 표준화(standardization)의 역할이 공선성진단(collinearity diagnostics)에서 논점(issue)이다. 특히 공선성을 진단하기 전에 자료가 중심화(centering)되어야 하는지 아닌지에 대한 문제는 다소 논란(controversy)이 있다. 따라서 다음의 것에서, $i = 1, 2, \dots, k$ 에서 $\sum_{j=1}^n x_{ij}^2 = 1.0$ 이 되도록 자료가 척도화(scaled)되어 있다고 가정해보자. 두 가지 극단적인 견해(extreme view)는 “자료가 공선성을 진단하기 전에 중심화(centered)되어야 한다”는 것과 “자료가 공선성을 진단하기 전에 중심화(centered)되어서는 안 된다”는 것이다.

회귀변수들(regressor variables)이 중심화(centered)되었을 때 나오는 상수항, $b'_0 = \bar{y}$ 는 공선성에 의하여 조금도 손상(damaged) 받거나 영향을 받지 않음은 분명하다. 이것은 $X'X$ 행렬로 잘 설명할 수 있다. 중심척도화된 자료(centered and scaled data)에서 $X^{*'}X^*$ 는 회귀변수(regressor variable)들 간의 상관행렬(correlation matrix)이다.

$$(X'X) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & X^{*'}X^* & \\ 0 & & & \end{bmatrix}$$

따라서 $b'_0 = \bar{y}$ 이며, 이는 표준화된 회귀변수(standardized regressor)의 회귀계수(regression coefficient)들에 대하여 독립적이다. 한편으로 공선성이 심각하다면, 중심화되지 않은 모형(noncentered model)의 절편(intercept)은 심하게 악영향을 받을 수 있는데, 이것은 단순히식(8.8)에서 β_0 가 $\beta'_1, \beta'_2, \dots, \beta'_k$ 의 선형함수(linear function)이고, 이 선형함수 중의 일부가 심각한 의존성(dependency or dependencies)에 의하여 악영향을 받았다는 것으로 설명이 가능하다. 만약 자료가 중심화 된다면, 공선성 진단(collinearity diagnostics)은 상수항(constant term)에 대한 영향을 반영하지 않을 것이다. 중심화는 상수를 다른 회귀변수들(regressors)에 대하여 직교(orthogonal)하도록 바꾸기 때문에, 분산분해비율(variance decomposition proportion)로 보아, 상수(constant)에 대한 어떠한 손상(damage)도 발견할 수 없다. 따라서 단일(single) 또는 다중변수(multiple variable)에서 “상수항(constant term)을 가지는 공선성”을 발견할 수는 없다. “상수항(constant term)을 갖는 공선성”이라는 용어는 다소 혼란스럽지만 실제 각각의 회귀변수(individual regressor)들의 선형 조합(linear combination)이 상수(constant)이거나 단일회귀변수(single regressor)가 상수(constant)임을 암시한다.

중심화(centering)는 고유값들(eigenvalues)과 고유값들의 비(ratio)에 심대한 효과를 나타낼 수 있다. 앞에서 지적한 바와 같이, 이러한 고유값들(eigenvalues)의 비들(ratios)은, 회귀자료(regressor data)의 작은 변화(changes) 혹은 혼란(perturbation)에 대한 회귀결과의 민감도(sensitivity of regression results)를 판단하기 위하여 고안되었다. 그러나 중심화된 자료세트의 결과가 동일한 자료세트(data set)의 중심화되지 않은 자료세트로부터의 결과와 반대인 경우, $\phi_j = \lambda_{\max} / \lambda_j$ 비율의 해석에 어려움이 따를 것이다. 중심화를 하면,

중심화되지 않았을 때와는 다른 x 의 단위 변화(unit change)가 정의된다. 그 결과, “작은 변화에 대한 민감도(sensitivity to small change)”라는 개념은 두 모형에서 다를 것이다.

우리는 회귀 중심화(regression centering)가 표준회귀분석(standard regression analysis)과 공선성진단(collinearity diagnostic)에 미치는 영향에 대해서 논의하였다. 비록 그것들이 별도로 처리되어도, 유사점은 있다. 절편(intercept, 회귀변수가 0일 때의 추정된 반응)을 추정할 필요가 별로 없거나, 특정 상황에서 회귀변수평균(regressor average)이 원래의 원점(natural origin) 혹은 그 근처 위치를 의미한다면, 회귀변수 중심화는 매우 합리적이다. 진단도구(diagnostics)의 경우, 절편추정 효율(efficiency, 원래 변수에서)이나 공선성에서의 절편의 역할을 연구할 필요가 없다면, 중심화는 허용될 뿐만 아니라 필요할 것이다. 후자의 경우, 회귀변수의 범위(range)가 적용례(application)에 비추어 전형적(typical)이고 합리적이라면, 상수항을 가진 단일 회귀변수(single regressor)의 공선성은 중요성을 잃게 될 것이다.

예제 8.5 계획적인 공선성 제거

Experiment	x_1	x_2	x_3	y
1	1.5	6	1315	243
2	1.5	6	1315	261
3	1.5	9	1890	244
4	1.5	9	1890	285
5	2	7.5	1575	202
6	2	7.5	1575	180
7	2	7.5	1575	183
8	2	7.5	1575	207
9	2.5	9	1315	216
10	2.5	9	1315	160
11	2.5	6	1890	104
12	2.5	6	1890	110

위는 잘 설계된 실험자료이다. 실험배열(experimental array)은 4개의 추가적인 반복(replicate)을 가진 2^3 계승(factorial)의 1/2 부분(fraction)이다. 이 디자인의 한가지 장점은 공선성(collinearity)의 예정된 제거(planned elimination)에 있다. 이것은 중심척도화된 자료에서 계산한 상관행렬(correlation matrix), 분산팽창요인(variance inflation factor), 상관 행렬의 고유값(eigen value of correlation matrix), 조건수(condition number) 등을 관찰하면 쉽게 알 수 있다. Table 8.7에서는 중심화(centered)되지 않은 동일 자료세트(data set)의 공선성 진단도구(collinearity diagnostics)와 함께 공선성 정보(collinearity information)를 제시하고 있다. 자료가 중심화되었을 때, 공선성이 없으며 모든 것이 이상적으로 보인다는 것을 주목할 필요가 있다. 반면에 자료가 척도화는 되었으나 중심화가 되지 않았을 경우 진단도구(diagnostics)는 다른 이야기(story)를 하게 된다. 모든 계수(coefficient)들의 분산팽창요인(variance inflation factor)이 전부 상대적으로 크다. 한 개의 고유값(eigenvalue)은 공선성의 어려움을 나타낸다. 심지어 $X'X$ 행렬(이 경우에 상관행렬이 아님)도 다소 비관적인 상황(gloomy picture)을 그려내는 것으로 볼 수 있다.

이제, 이러한 자료세트(data set)의 경우에서 공선성을 해석하는데 어떤 진단도구(diagnostics)가 사용되어야 할 것인가? 중심화되지 않은 자료에 대한 다소 침울한(depressing) 진단도구(diagnostics)는, 엄밀히 말해서 소위 “상수를 가진 공선성(collinearity with the constant)”이다. 이것은 분산 비율(variance proportion) 속에 표시된다. 그러나, 중심화되지 않은 모형 절편의 추정값, 즉 원래 변수가 (0,0,0)일 때 추정된 반응(estimated response)은 전적으로 중요하지 않다. 중심(center), $\bar{x}_1 = 2.0$, $\bar{x}_2 = 7.5$, $\bar{x}_3 = 1593.33$ 이 실험영역에서 중요한 중심(centroid)이다. 이 경우, 자료의 중심화, 즉 진단도구(diagnostics)에서 절편의 역할을 제거하기는 필수 불가결한 것이다. 따라서

중심척도화된 진단도구(centered and scaled diagnostics)가 도입되어야만 한다.

TABLE 8.7 Collinearity diagnostics for coal data of Example 5.3

<i>Centered and scaled</i>				
Coefficient	VIF			
b_1	1.0			
b_2	1.0			
b_3	1.0			
$X^{*'} X^* = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 1.0 & 0 \\ 0 & 0 & 1.0 \end{bmatrix}$	(Correlation matrix)			
Eigenvalues = 1.0, 1.0, 1.0				
Condition Number = 1.0				
<i>Scaled, not centered</i>				
Coefficient	VIF			
b_0	108.431			
b_1	25.0			
b_2	38.5			
b_3	46.93			
$X^{*'} X^* = \begin{bmatrix} 1 & 0.9798 & 0.9869 & 0.9893 \\ & 1 & 0.9670 & 0.9693 \\ & & 1 & 0.9764 \\ & & & 1 \end{bmatrix}$				
Variance proportions				
Eigenvalue	b_0	b_1	b_2	b_3
3.9344	0.0006	0.0025	0.0016	0.0014
0.03526	0.0035	0.8096	0.1392	0.0508
0.02360	0.0019	0.0166	0.5213	0.4620
0.006763	0.9940	0.1713	0.3379	0.4859
Condition Number = 581.77				

8.4. 다중공선성이 있는 경우 최소제곱법의 대안(Alternatives to Least Squares in Cases of Multicollinearity)

다중공선성(multicollinearity)을 극복하기 위하여 고안된 많은 추정방법(estimation procedure)과 모형의 불안정성(instability)을 제거하고 회귀계수의 분산(variance of regression coefficient)을 줄이기 위하여 개발된 방법들이 있다. 이런 방법들을 사용하는데 있어서 어느 정도 논쟁거리(controversy)가 있기 때문에, 자료 분석가는 자신이 어느 때든지 이 방법들을 사용할 수 있는 백지위임장(carte blanche)을 소유하고 있다고 느껴서는 안 된다. 과학적 변수들(scientific variables)이 종속반응(dependent response)에 미치는 영향(influence)이 원래 증복되는 구조(structure)를 관련과학분야(subject matter field)에서 다룰 때, 이 방법들은 여전히 자료분석가의 레퍼터리(repertoire)에서 중요한 부분(valuable part)이 될 수 있다.

진단도구를 사용하여 다중공선성(multicollinearity)이 문제라고 이미 판단한 경우, 최소제곱법(least squares)을 대안(alternative)으로 삼지 않고, 먼저 다중공선성을 제거하려고 시도하면, 종종 상당한 이득(substantioal benefit)을 얻을 수 있다. k 개의 회귀변수가 있는 경우, 다중공선성이 있다는 것은 실제 모형구축에는 k 개 보다 더 적은 수의 변수가 관여한다는 것을 암시한다. 다시 말하면, k 개의 회귀변수로 모형화를 해야 하는 것에 대하여 정당한 근거가 될 수 있는 충분한 정보(information)가 회귀자료(regressor data)에는 없다. 결과적으로 분석가는 하나 이상의 회귀변수를 제거함으로써, 종종 다중공선성을 제거할 수 있거나 혹은 확실하게 줄여줄 수 있다. 진단도구(diagnostics) 혹은 모형화될 현상에 관한 지식이 어떤 변수를 제거하는 것이 좋은지를 제시해 주겠지만, 분명히 여기에는 거래조건(trade off)이 있다. 다중회귀에서 만약 변수 x_1 , x_2 가 밀접하게 상관되어(correlated) 있고, 변수 x_2 가 모형에서 제거된다면 일부(아마 거의 대부분)의 다중공선성은 제거될 것이다. 그러나 분석가는 모형 적합의 질(quality of fit of the model)이 심하게 손상 받지는 않았는지에 대해서 알아야 한다. 변수를 제거함으로써 모형의 예측능력이 향상되는 것이, PRESS, 절대 PRESS 잔차합(sum of the absolute PRESS residuals), 예측값의 표준오차(standard error of prediction) 등과 같은 통계량(statistics)에 반영되어야 한다. 추가적으로, 남아 있는 계수들의 분산팽창요인(variance inflation factor)에 집중하여야 한다.

표준최소제곱추정법(standard least squares estimation)이 아직 남아있지만, 다중공선성을 줄여주는 대안(alternative)으로 회귀변수들(regressor variables)을 변환(transformation)시키는 방법이 있다. 이렇게 회귀변수들을 변형시키면 모든 회귀변수들의 정보량 informational content)의 일부를 유지하면서 회귀변수 시스템(regressor system)의 차원(dimensionality)을 줄일 수 있다. 예를 들어, 두 개의 회귀변수 x_1 , x_2 가 있고, 이들이 고도로 상관되어 있을 때(highly correlated), 두 개의 변수를 더한 x_1+x_2 변수로 재정의(redefining)하거나 또는 두 변수를 나누어 비의 형태로 변형을 하게 되면, 만족할 만한 결과가 나올지도 모른다. 그러나 문제의 상황(context)에 맞지 않는, 회귀변수 함수들로 변환할 때에는 주의하여야 한다. 예를 들어서,

다른 단위로 측정된 회귀변수들을 더할 때는 어쩔 수 없을 때에만 하여야 한다.

예제 8.6 Hald Data

Table 4.1의 Hald data를 살펴보자. 그 자료들을 3장과 4장에서 설명 목적으로 이용하였다. 또한 Table 8.1에 보여진 진단도구(diagnostics)는 심각한 공선성(collinearity)을 보이고 있다. 4장에서 공선성 때문에 모형을 선택할 때 방해를 받았다는 것을 알아야 한다. 네 가지 회귀변수의 분산팽창요인(VIF)은 모두 큰 값을 가지며, 특히 두 개의 계수(b_2 , b_4)는 250~300의 분산팽창요인을 갖는다.

좋은 예측력(predictability)을 보이면서 공선성이 거의 없거나 혹은 약간 있도록 회귀변수들의 부분세트(subset)를 선택할 수 있는지 판단하는 것은 꽤 흥미로울 것이다. 몇몇 모형을 예제 4.5에서 살펴보았다. 이제 실제로 더 나은 예측 모형을 만들기 위한 변수 제거(variable deletion)가 공선성을 상당하게 줄여주는가를 판단하기 위하여 분산팽창요인(variance inflation factor)을 관찰할 것이다. 다음은 좋은 예측 성질(prediction property)을 가진 것으로 보이는 부분세트 모형들(subset models) 중 몇 가지에 대한 공선성 개요(synopsis of collinearity)이다. 비교의 편의를 위하여 완전모형(full model)을 포함시켰다.

x_1, x_2, x_3, x_4		
Coefficient	VIF	
b_1	38.49621	
b_2	254.42317	Regression $\hat{y} = 62.4054 + 1.5511x_1 + 0.5102x_2$
b_3	46.86839	+ 0.1019x ₃ - 0.1441x ₄
b_4	282.51286	$R^2 = 0.9824$
x_1, x_2		
Coefficient	VIF	
b_1	1.055129	Regression $\hat{y} = 52.57735 + 1.46831 x_1 + 0.66225 x_2$
b_2	1.055129	$R^2 = 0.9787$
x_3, x_4		
Coefficient	VIF	
b_3	1.000873	Regression $\hat{y} = 131.28241 - 1.19985 x_3 - 0.72460 x_4$
b_4	1.000873	$R^2 = 0.9353$
x_2, x_4		
Coefficient	VIF	
b_2	18.74113	Regression $\hat{y} = 94.1601 + 0.3109 x_2 - 0.4569 x_4$
b_4	18.74113	$R^2 = 0.6801$
x_1, x_4		
Coefficient	VIF	

b_1	1.064105	Regression $\hat{y} = 103.09738 + 1.43996x_1 - 0.61395x_4$
b_4	1.064105	$R^2 = 0.9725$
x_1, x_2, x_3		
Coefficient	VIF	
b_1	3.251068	Regression $\hat{y} = 48.19363 + 1.69589x_1 + 0.65691x_2 + 0.25002x_3$
b_3	1.063575	
b_5	3.142125	$R^2 = 0.9823$
x_1, x_2, x_4		
Coefficient	VIF	
b_1	1.066330	Regression $\hat{y} = 71.6483 + 1.4519x_1 + 0.4161x_2 - 0.2365x_4$
b_2	18.780309	
b_4	18.940077	$R^2 = 0.9823$
x_1, x_3, x_4		
Coefficient	VIF	
b_1	3.678168	Regression $\hat{y} = 111.68441 + 1.05185x_1 - 0.410044x_3 - 0.64280x_4$
b_3	3.459601	
b_4	1.181000	$R^2 = 0.9813$

(x_2, x_4) , (x_1, x_2, x_4) 모형은, 모든 변수들을 고려한 경우보다 VIF 값이 많이 감소하였으나, x_2 과 x_4 사이의 강한 상관(correlation) 때문에 여전히 VIF가 100이상의 큰 값을 가진다. 그러나 여기에서 연구된 나머지 후보들은, 모든 변수들을 고려한 경우보다 공선성이 충분히 감소하였으며 따라서 후보로 남게 된다.

R^2 와 VIF로 보아, 모형 (x_1, x_2) 는 좋은 모형으로 판단된다.

다음은 예제 8.6에서 사용한 R code이다.

```
data<-read.table('c:/hald.txt',header=T)
attach(data)
fit<-lm(y~x1+x2+x3+x4,data)
summary(fit)
vif(fit)
fit1<-lm(y~x1+x2,data)
summary(fit1)
vif(fit1)
fit2<-lm(y~x3+x4,data)
summary(fit2)
vif(fit2)
```

```

fit3<-lm(y~x2+x4,data)
summary(fit3)
vif(fit3)

fit4<-lm(y~x1+x4,data)
summary(fit4)
vif(fit4)

fit5<-lm(y~x1+x2+x3,data)
summary(fit5)
vif(fit5)

fit6<-lm(y~x1+x2+x4,data)
summary(fit6)
vif(fit6)

fit7<-lm(y~x1+x3+x4,data)
summary(fit7)
vif(fit7)

```

능형 회귀(Ridge Regression)

능형회귀는 논란이 있는 하지만, 다중공선성(multicollinearity)을 제거하는 좀 더 인기있는 추정방법(estimation procedure)이다. 여기와 다음절에서 논의되는 방법이 편향추정기법(biased estimation techniques)의 범주에 포함된다. 편향추정기법의 개념은 다음과 같다: 보통최소제곱(ordinary least squares)으로 불편추정값(unbiased estimate)를 구할 수 있고, 모든 선형 불편추정량들(linear unbiased estimators)은 최소 분산(minimum variance)을 가지나, 추정량(estimators)의 분산(variance)은 상한선(upper bound)이 없으며 다중공선성이 있을 경우 분산(variances)은 매우 커진다. 결과적으로, 다중공선성이 있을 경우, 보통최소제곱(ordinary least squares)을 사용하면 불편 속성(unbiased property)은 얻을 수 있으나 이에 대하여 큰 대가를 치러야 한다고 생각할 수 있다. 편향추정(biased estimation)은 회귀계수들(regression coefficients)의 안정성(stability)을 증가시키고 분산(variance)을 상당하게 감소시키기 위하여 사용된다. 편향추정으로 계수들(coefficients)이 편향(biased) 되기는 하나, 성공했을 경우, 분산(variance) 감소의 크기(magnitude)가 추정량(estimator) 편향(bias)의 크기보다 더 크다.

능형회귀를 사용하는데는 여러가지 이유가 있다. 아마도 가장 매력적인 것은 어떤 기법이 X^*X^* (또는 $X'X$)의 고유값(eigenvalues)이 회귀결과에 미치는 효과를 감소(reduction) 또는 “둔화”(“dampening”)시키는지 보여주는 것이다.

이 장 앞부분과 3장의 전개에서, 고유값(eigenvalues)이 작으면 분산(variances)이 커진다는 것을 알았다. 예를 들어서, 상관행렬(correlation matrix)이 다음과 같이 주어진, 이변수 시스템(two-variable system)에서,

$$(X^* X^*) = \begin{bmatrix} 1.0 & 0.999 \\ 0.999 & 1.0 \end{bmatrix}$$

고유값은 다음과 같다.

$$\lambda_1 = 1.999 \quad \lambda_2 = 0.001$$

이로 인하여, 다음과 같은 역(inverse) $X^* X^*$ 의 대각요소들(diagonals) 즉, 분산증폭요인들(variance inflation factors, VIFs)을 가진 회귀계수의 추정(estimate of the regression coefficients)이 매우 비효율적이 될 것이라는 것은 분명하다.

$$\begin{aligned} (VIF)_1 &= 500.25 \\ (VIF)_2 &= 500.25 \end{aligned}$$

이 2×2 경우의 상관행렬(correlation matrix)의 성질은, 상관행렬이 다음과 같이 주어지는 직교 사례(orthogonal case)의 $(X^* X^*)$ 가 가지는 바람직한 성질과 전혀 다르기 때문에,

$$(X^* X^*) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

고유값(eigenvalue) λ_2 는 작고 분산팽창요인(variance inflation factor)은 크다는 것을 알 수 있다.

다시 말해, 공선성(collinearity)이 있는 경우에는 위와 같이 대각요소들(diagonals)이 우세하지는 않다. 일반적으로 대각요소가 우세하지 않은 경우에는(nondominance of the diagonals) 적어도 하나의 고유값이 작아진다. 따라서, $(X^* X^*)$ 이 좀더 직교 사례(orthogonal case)처럼 되게 하려면 무엇을 하여야 하는가? 고유값(eigenvalues)을 커게, 행렬(matrix)의 행렬식(determinant)을 작게, 따라서 역행렬의 원소들(the elements of the inverse)을 작게 하려면 무엇을 해야 하는가?

행렬 $(X^* X^*)$ 를 행렬 $(X^* X^* + kI)$ 로 대체한다고 가정하자. 여기서, k 는 작은 양수값(positive quantity)이다. 앞의 설명은 $(X^* X^* + kI)$ 행렬에서 $k = 0$ 인 경우이다. 따라서 상관행렬(correlation matrix)을 다음의 행렬로 대체할 수 있다.

$$(X^* X^* + kI) = \begin{bmatrix} 1.1 & 0.999 \\ 0.999 & 1.1 \end{bmatrix}$$

이것이 고유값들과 역행렬의 원소들(inverse elements)에 대해서 무엇을 하는가? 행렬 V 는 $(X^* X^*)$ 를 대각화(diagnonalize)하기 때문에, $(X^* X^* + kI)$ 역시 대각화 한다. 따라서, 아래와 같다.

$$V'(X^* X^* + kI)V = \begin{bmatrix} \lambda_1 + k & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 + k & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \lambda_k + k \end{bmatrix}$$

그러므로, 새로운 행렬 $(X^* X^* + kI)$ 의 고유값들(eigenvalues)은 $\lambda_i + k$ ($i = 1, 2, \dots, k$) 이다. 중심 대각선(main diagonal)에 k 를 더하면 λ_i 는 $\lambda_i + k$ 로 대체된다. 따라서, 이전의 설명에서 고유값들은 2.099와 0.101로 대체된다. 추가적으로, 역행렬(inverse matrix)은 다음과 같다.

$$\begin{bmatrix} 1.1 & 0.999 \\ 0.999 & 1.1 \end{bmatrix}^{-1} = \begin{bmatrix} 5.1887 & -4.71229 \\ -4.171229 & 5.1887 \end{bmatrix}$$

결과적으로 $\lambda_2 = 0.001$ 의 손상효과(damaging effect)는 사라졌고, 큰 VIFs를 포함하여 역행렬(inverse)의 큰 원소들(large elements) 때문에 유발된 장애들이 완화되었다.

앞에서 언급한 것이 능형 회귀(ridge regression)에 대하여 완전한 정당성(total justification)을 부여한다고 간주해서는 안된다. 즉, 나쁜 상관행렬(poorly conditioned correlation matrix)의 중심대각(main diagonal)에 작은 양의 상수(positive constant)를 단순히 더하기만 했을 뿐인데도 어떤 중요한 변화가 생기는가에 초점을 맞추어야 한다. 다음 페이지에서, 여기에 관련된 수식 전개를, 우리가 일반적으로 능형회귀 추정량(ridge regression estimator)이라고 부르는 추정량(estimator)과 연결시킬 것이다.

계수 β 의 능형회귀 추정량(ridge regression estimator)은 아래의 (8.9)식들의 시스템에서 b_R 에 대하여 풀면 얻을 수 있다.

$$(X'X + kI)b_R = X'y \quad (8.9)$$

여기서 $k \geq 0$ 은 종종 축소모수(shrinkage parameter)라고 한다. 해는 다음에 의하여 주어진다.

$$b_R = (X'X + kI)^{-1} X'y \quad (8.10)$$

축소모수(shrinkage parameter) k 를 선택하는 다양한 방법들이 있다. $(X'X)$ 는 회귀변수들(regressor variables)을 척도화(scaling) 혹은 중심척도화(centering and scaling) 한 것으로 보아야 한다. 중심화의 경우, $X'X$ 는 상관행렬(correlation matrix) X^*X^* 가 된다. 식(8.10)의 능형추정량(ridge estimator)의 성질을 연구해보면, k 는 추정량의 분산(variance of the estimators)을 알맞게 해주는 역할을 한다. 식(3.37)의 식, $\sum_i \frac{Var b_i}{\sigma^2}$ 은 작은 고유값이 최소제곱계수의 분산(variances of the least squares coefficients)에 미치는 영향을 가장 극적으로 설명한 것이 아닌가 한다. 능형회귀 추정량(ridge regression estimator)의 경우, 이것과 동등한 특성(equivalent property)은 다음 식에 의하여 주어진다.

$$\sum_i \frac{Var b_{i,R}}{\sigma^2} = \sum_i \frac{\lambda_i}{(\lambda_i + k)^2} \quad (8.11)$$

예를 들어서, $\lambda_1 = 2.985, \lambda_2 = 0.01, \lambda_3 = 0.005$ 인 세 개의 회귀변수들(regressor variables)이 있다면, 최소제곱추정(least square estimation)은 다음과 같다.

$$\begin{aligned} \frac{\sum_{i=1}^3 Var b_i}{\sigma^2} &= \sum_{i=1}^3 \frac{1}{\lambda_i} \\ &= .3350 + 100 + 200 \\ &= 300.3350 \end{aligned}$$

$k = 0.10$ 인 능형회귀(ridge regression)의 경우, 분산의 합(sum of the variances)은 다음과 같다.

$$\sum_{i=1}^3 \frac{\lambda_i}{(\lambda_i + k)^2} \cong 2.3$$

다중공선성(multicollinearity)이 심할 때, 즉, 적어도 한 개가 근사 0 고유값(near zero eigenvalue)일 때, 분산이 크게 개선되고, 따라서 계수의 안정성(coefficient stability)도 경험할

수 있다. 식(8.11)은 이 절 앞부분의 요점 즉, 능형회귀(ridge regression)의 k 는 공선성(collinearity)의 결과인 작은 고유값들(eigenvalues)의 손상효과(damaging impact)를 완화한다는 것을 강조하고 있다.

회귀계수의 제곱편향합(sum of the squared biases of the regression coefficients), $\sum_{i=1}^k (\text{Bias } b_{i,R})^2 = \sum_{i=1}^k [E(b_{i,R}) - \beta_i]^2$ 에 관한 수식을 관찰해 보면, $k > 0$ 을 선택함으로써

생기는 편향(bias)을 가장 잘 정량화 할 수 있다. 이 수식은 다음과 같다(Hoerl and Kennard, 1970a).

$$\sum_{i=1}^k [E(b_{i,R}) - \beta_i]^2 = k^2 \beta' [X'X + kI]^{-2} \beta \quad (8.12)$$

그러므로 식(8.12)에 주어진 편향 항(bias term)보다 분산감소(variance reduction)가 더 큰 값이 되도록 k 를 선택하면, 능형회귀가 성공적일 것으로 기대할 수 있다. 분석가는 무엇이 편향(bias)인지 알지 못할 것이므로, 이것이 행해질 것이라는 보장은 없다. (8.11)의 분산기여(variance contribution)에 관하여 면밀하게 연구해보면, 공선성(collinearity)이 심각할 때, 즉 적어도 하나의 $\lambda_i \approx 0$ 일 때 분산감소(variance reduction)라는 의미에서, 좀 더 나은 개선의 기회가 있음을 알 수 있다. 식(8.12)는 β 에 모르는 계수들(unknown coefficients)을 포함하고 있다; 따라서 편향(bias)을 계산하려는 어떠한 시도도 오도(misleading)될 수 있다. 물론 k 를 선택하는 것은 분석가의 몫이나, 추정값들(estimates)이 개선되도록 모수 값(parameter value)을 선택하여야 한다. 여기서 개선이라는 말은 흔히 추정값들(estimates)이 좀 더 안정적이거나 예측(prediction)이 향상되는 것을 의미한다. 독자들은 문헌에 k 를 선택하는 방법들이 많이 나온다는 것을 알아야 한다. 여기서는 몇가지만 골라서 설명하기로 한다. 능형회귀분석(ridge regression)을 하기 전에, 사용자는 능형회귀가 그 영역에서 충분한 장점이 있는지 판단하여야 한다.

앞서 말한 것에서, $\sum_{i=1}^p$ 또는 $\sum_{i=0}^k$ 과는 달리, 합의 표기법(summation notation)으로서 $\sum_{i=1}^k$ 를 사용하였음을 주목해보자. 여기서는 표기법(notation)의 어떠한 모순(inconsistency)도 의도한 바가 아니다. 상수항(constant term), β_0 가 모수 집합(parameter set)의 부분일 때, 여전히 k 개의 회귀변수들(regressor variables, $p = k + 1$)을 가정한다. 엄밀히 말해서, 어떤 합(summation)이 적용될지는 자료가 중심척도화되었는지, 또는 단순히 척도화만 되었는지에 따라 달라진다. 만약 중심화가 되었다면, 식(8.7)의 결과 항, β'_0 은

공선성(collineality)에 영향을 받지 않으며, β'_0 의 최소제곱추정값(least squares estimate)은 단순히 $\bar{y} = \sum_{i=1}^n y_i / n$ 이다. 그러므로, 공선성에 의한 손상(damage of collineality)과 능형회귀의 완화 영향(moderating influence)에 대한 평가는, 식(8.11) 계수들의 분산합(sum of variances of the coefficients)에 대한 식으로 가장 잘 예시된다. 따라서 중심화된 자료(centered data)에서 능형회귀를 시행할 때, (8.9)와 (8.10)에 포함된 행렬은 $(X^* X^* + kI)$ 이며, 이는 단지 중심대각(main diagonal)에 축소모수(shrinkage parameter) k 가 추가된 $k \times k$ 상관행렬(correlation matrix)에 불과하다. 따라서 식(8.7)의 표기법에서 $b'_{i,R}$ 로 명명되어야 하는 능형회귀 추정량(ridge regression estimator)은, 중심척도화된 변수들의 계수들(coefficients of centered and scaled variables)이며, β'_0 의 추정값(estimate)은 \bar{y} 이다. 모형(8.6)에서 원래 변수들의 계수들(coefficients of natural variables)은 다음과 같이 계산된다.

$$b_{i,R} = \frac{b'_{i,R}}{S_i} \quad (i = 1, 2, \dots, k)$$

모형(8.6)의 상수항(constant term)은 다음과 같이 추정된다(8.8 참조).

$$b_{0,R} = b'_{0,R} - \frac{b'_{1,R}\bar{x}_1}{S_1} - \frac{b'_{2,R}\bar{x}_2}{S_2} - \dots - \frac{b'_{k,R}\bar{x}_k}{S_k}$$

***k*의 선택(Choice of *k*)**

능형추적법(ridge trace)은 축소모수(shrinkage parameter)를 선택하는데 매우 실용적인 방법이다(Hoerl and Kennard, 1970b). 모든 계수들(coefficients)이 안정될 때까지 계속 k 를 증가시킨다. 매우 흔하게, k 에 대한 계수들(coefficients)의 그림은 대각합(trace)을 생생히 그려내고, 분석가가 적절한 k 값을 판단하도록 도와준다. 안정성(stability)이 회귀 계수들(regression coefficients)의 수렴(converge)을 의미하는 것이 아님을 강조한다. k 값이 영(zero)에 가까우면, 다중공선성(multicollinearity)은 계수들에 급격한 변화를 초래할 것이다. 이러한 빠른 변화는, 계수의 분산(coefficient variance)이 팽창될 것으로 예상되는 k 간격(interval) 안에서 발생한다. k 가 커짐에 따라서 분산(variance)은 줄어들고, 계수들(coefficients)은 좀 더 안정된다. k 값은 계수들이 더 이상 급격하게 변화하지 않는 지점에서 선택된다.

능형추적법(ridge trace procedure)을 사용할 경우, 표준화 즉, 중심척도화된 회귀변수(coefficients of standardized, i.e., centered and scaled, regressors)의 계수들의 그림을 매우

자주 관찰하게 된다. 그러나, 가끔은 원래 변수(natural variables)의 계수들을 관찰함으로써 계수의 안정성(stability) 또는 해석력(interpretability)에 대하여 더 잘 알 수 있다. 이에 대한 사례가 예제 8.7이다.

예제 8.7 체내 지방자료

Table 8.8는 19명에 대한 체내 지방 자료를 나타낸 것으로 삼두근 두께(x_1), 허벅지 둘레(x_2), 팔의 중간부위 둘레(x_3)의 세 가지 설명변수와 체내 지방량의 반응변수로 구성되어 있다. 설명변수들 중 x_1 과 x_2 의 상관계수는 0.92로 매우 높다. 이 자료에 능형회귀를 적용해 보자. k 를 찾기 위하여 0에서 시작하여 0.001씩 증가시켜 나가면 k 값 0.02 근처에서 회귀계수 값들이 수렴하기 시작한다. 따라서, $k = 0.02$ 로 놓고 능형회귀를 해보면 다음과 같다.

$$\bar{y} = 3.618 + 0.956x_1 + 0.054x_2 - 0.378x_3$$

한편, 원래 자료를 능형회귀가 아닌 다중회귀모형에 적합시키면, 다음과 같다.

$$\hat{y} = 124.6 + 4.628x_1 - 3.086x_2 - 2.306x_3$$

이 경우 회귀계수 추정값의 표준편차들이 매우 크다. 즉, 추정값의 값이 매우 불안함을 알 수 있다.

표 8.8 체내 지방자료

삼두근 두께(x_1)	허벅지 둘레(x_2)	팔의 중간부위 둘레(x_3)	체내 지방량(y)
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4

30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

Table 8.9 체내 지방자료의 능형회귀결과

k	$b_{0,R}$	$b_{1,R}$	$b_{2,R}$	$b_{3,R}$	SS_{Res}
0	124.6	4.628	-3.086	-2.306	96.907
0.01	7.056	1.060	-0.035	-0.433	105.247
0.02	3.618	0.956	0.054	-0.378	105.742
0.03	2.427	0.920	0.085	-0.359	105.917
0.04	1.823	0.901	0.101	-0.349	106.007
0.05	1.457	0.890	0.110	-0.343	106.061
0.06	1.212	0.882	0.117	-0.339	106.097
0.07	1.037	0.877	0.121	-0.337	106.123
0.08	0.905	0.873	0.125	-0.334	106.143
0.09	0.802	0.870	0.128	-0.333	106.159
0.10	0.720	0.867	0.130	-0.332	106.171

예제에 사용된 R-code는 다음과 같다.

```

data<-read.table("c:/data/ex8_7.R",header=TRUE)
library(MASS)
x.m<-as.matrix(data[,1:3])
one<-rep(1,19)
x<-as.matrix(cbind(one,x.m))
y<-as.matrix(data[,4])
iden<-diag(1,4)
be1<-solve(t(x)%*%x)%*%t(x)%*%y
bes<-matrix(rep(0,55),nrow=5)
for(i in 1:11){
  k<-(i/100

```

```

be<-solve((t(x)%*%x+k*iden))%*%t(x)%*%y
bes[1:4,i]<-be
bes[5,i]<-sum((y-x%*%be)^2)}

```

능형추적(ridge trace) 혹은 안정성(stability)의 긍정적 측면의 대부분은 아마도 실용적(practical)이고 자료의존적(data-dependent)이라는 사실일 것이다. 능형추적(ridge trace) 혹은 안정성(stability)은 개념적인 기준(conceptual criteria) 보다는 안정성이 실제로 의미하는 것이 무엇인지에 대한 분석가 자신의 개념에 기반을 둔 것이다. 물론, 후자는 단순히 안정성이 다소 주관적이기 때문에, 가치 없는 절차(non-virtue of the procedure)로서 해석될 수 있다. 모수(parameter)를 좀 정확한 방법으로 결정하는 경우보다, k 를 선택하는 것이 어느 정도는 더 임의적일(arbitrary) 수 있다.

k 의 선택에 대한 예측 기준의 사용(Use of Prediction Criteria for Choice of k)

k 를 선택하는 많은 방법들은, 계수 추정(estimation of the coefficients)이 확실히 향상되는, k 의 간격(interval) 내 수치를 선택하도록 설계된다. 따라서 예측 성능(prediction performance)을 보다 직접적으로 보여주는 기준(criteria)으로 k 를 선택하는 것이 더 중요하다. 여러 기준이 있으며, 모두 여기에서 논의하고 설명할 것이다. 4장에서 논의되었던, 동일 분산편향 형태 교환(the same variance-bias type trade-off)을 기초로 하는 C_p 유사 통계량(C_p -like statistic)을 먼저 생각해보자. 통계량은 식(4.20)에서 주어진 것과 매우 유사하다.

$$C_k = \frac{SS_{Res,k}}{\hat{\sigma}^2} - n + 2 + 2\text{tr}[H_k] \quad (8.13)$$

여기에서 $H_k = [X^*(X^{*'}X + kI)^{-1}X^*]$, $SS_{Res,k}$ 는 능형회귀의 잔차제곱합(residual sum of squares), $\text{tr}[H_k]$ 는 H_k 의 대각합(trace)이다.

식(8.13)에서, X^* 와 $X^{*'}X^*$ 는 상수항(constant term)을 반영하지 않는다. 즉, k 개의 회귀변수(regressor variables)가 있을 때, X^* 는 $n \times k$ 이고, $X^{*'}X^*$ 는 k 개의 회귀변수들(regressor variables)간의 상관행렬(correlation matrix)이다.

H_k 가 보통최소제곱(ordinary least squares)의 HAT 행렬과 같은 역할을 수행한다는 점을 주목하라. C_p 통계량(C_p statistic)에 대한 식(4.20)에서 p 를 $1 + \text{tr}H_k$ 로 치환하면, 결과는 식(8.13)과 동일하다. 통계량 $\hat{\sigma}^2$ 은 OLS 추정의 잔차평균제곱(residual mean square)으로부터 온 것이다. 식(8.13)의 자세한 전개는 매우 흥미로우며, 부록 B.10에 나와있다.

C_k 통계량을 사용할 때는, C_k 를 최소화시키는 k 값과 함께, k 에 대한 C_k 의 그림을 그리는 절차가 있다. 예제 8.8을 살펴보라.

k 를 선택할 때 예측(prediction)이 매우 중요한 기준일 수 있기 때문에, 어떤 의미에서 교차 타당성(cross validation)을 보는 것은 당연한 것처럼 보인다. 다음과 같은 형태의 PRESS 유사 통계량(PRESS-like statistic)을 생각해볼 수 있다.

$$PR(Ridge) = \sum_{i=1}^n \left[\frac{e_{i,k}}{1 - \frac{1}{n} - h_{ii,k}} \right]^2 \quad (8.14)$$

여기서 $e_{i,k}$ 는 특정 k 값에 대한 i 번째 잔차(residual)를 나타내고, $h_{ii,k}$ 는 H_k 의 i 번째 대각원소(diagonal element)이다. 다시 말하면, k 값은 PR (Ridge)를 최소화 하도록 선택된다. k 에 대한 PR (Ridge)의 그림은 정보량이 풍부하다. 독자들은 식(8.14)와 4장의 OLS를 위한 PRESS 통계량 간의 유사점을 알아야 한다. 모자대각(HAT diagonal)을 $(1/n) + h_{ii,k}$ 로 대체해보라. 이것은 행렬 H_k 에서 온 것이며, 능형회귀의 HAT 행렬과 다소 유사한 역할을 하고 있다. 이제 실용적인 이유로 인하여, 식(8.14)의 식을 항상 사용할 수는 없다. 중심 척도화(center and scale)를 할 경우, 자료포인트(data point) 하나를 제외하면(set aside) 중심척도화 상수(center and scale constants)가 변화되고, 이로 인하여 모든 회귀변수들의 관측값들(regressor observations)이 변화된다. 결과적으로, 식(8.14)는 한번에 하나씩 관측값들(observations)을 삭제하고, 매번 처음부터 다시 능형회귀를 계산하여야만 얻을 수 있는 참 PRESS (true PRESS)의 근사(approximation)일 뿐이다.

식(8.14)를 매우 편안한 마음으로 사용할 수 있는 상황들이 있다. 그러나, 이들 이외의 다른 상황에서는 오도(misleading)가 있을 수 있다. 다음의 상황에서 PR (Ridge)를 사용하면 매우 만족스럽다:

1. 표본 크기(sample size)는 작지 않아야 한다.
2. 표본 자료(sample data)에 높은 지렛대 관측값(leverage observations)이 없어야 한다. 즉, 모자 대각(HAT diagonals)이 크지 않아야 한다.

1과 2에 대한 실제 숫자적인 지침(numerical guideline)은 보여주기 어렵다. 예제 8.8에서 설명할 것이다.

자료세트의 크기가 중간 정도 혹은 작거나, 높은 지렛대 관측값(leverage observations)이

있다 하더라도, 식(8.14)를 사용 가능하게 해주는 “제거 공식(deletion formula)”을 써지 않고도, 여전히 RESS 기준을 사용할 수 있다. 사실, 표본 크기가 중간 정도 혹은 작은 경우에, 한번에 하나씩 자료포인트를 컴퓨터로 제거하고 그때마다 반복적으로 능형회귀를 수행하는 것은 비교적 가능한 접근방법이다. SAS PROC IML 프로그램(1987)은 이러한 회귀(regression)를 수행하고, k 에 대한 PRESS의 그림을 그려준다. 물론 여기서 실제로 계산을 하면,

$$PRESS,k = \sum_{i=1}^n e_{i,-i,k}^2$$

여기에서 $e_{i,-i,k}$ 는 능형회귀에서 i 번째 PRESS 잔차(residual)이다. $e_{i,-i,k}$ 는 $\hat{y}_{i,-i,k}$ 를 포함한다. 특정 k 에 대한 i 번째 지점에서 예측(prediction)을 하기 위하여, i 번째 지점을 제거하고, 중심척도화된 회귀변수 자료(regressor data)를 재계산하고, 능형 회귀를 재계산하여 $\hat{y}_{i,-i,k}$ 를 계산하게 된다.

세 번째 기준은 교차타당성 접근방법(cross validation approach)이라기 보다는 예측 접근방법(prediction approach)을 대표하는데, 일반화 교차타당성(generalized cross validation, GCV)이라고 하며, 다음과 같다.

$$GCV = \frac{\sum_{i=1}^n e_{i,k}^2}{\{n - [1 + \text{tr}(H_k)]\}^2} = \frac{SS_{\text{Res},k}}{\{n - [1 + \text{tr}(H_k)]\}^2} \quad (8.15)$$

$1 + \text{tr}(H_k)$ 의 값 “1”은 상수항(constant term)의 역할이 H_k 에 포함되어 있지 않다는 사실을 설명해준다. GCV 이면의 원리에 관한 좀더 자세한 것은 Wahba et al. (1979)을 참고하기 바란다. 이 통계량은 PRESS와 동일한 예측 성향의 기준(norm)이다. (8.15)의 GCV와 (8.14)의 통계량 사이의 유사점을 알아보는 것은 쉬운 일이다. 다시 말하면, GCV를 최소화 시키는 k 를 선택하는 것이 적절한 절차이다. k 에 대한 단순 그림(simple plotting)만으로도 충분히 높은 설명력을 가질 수 있다.

k 를 선택하는 예측 방법에 관한 추가적인 설명들, H_k 의 역할(Further Comments Concerning the Prediction Methods for Choosing k , Role of H_k)

이제, 독자들은 C_k 와 OLS에서 이에 대응하는 C_p 기준(criterion) 사이의 관계를 이해하여야 한다. 부가적으로, 본문에서 PR (Ridge)이 OLS의 PRESS 기준(PRESS criterion)과 얼마나

유사한지를 보여줄 것이다. 이들을 구별해주는 양(quantity)이 H_k 행렬이다. 우리는 앞서서 이 행렬이 능형회귀(ridge regression)의 HAT 행렬에 대응한다고 지적한 바 있다. 실제로, 우리는 양(quantity) $\text{tr}(H_k)$ 를 회귀 자유도(regression degree of freedom)와 개념적으로 매우 유사한 것으로 간주할 수 있다. OLS 사례를 상기해보자.

$$\text{tr}[X^*(X^*X^*)^{-1}X^{*'}] = k$$

여기에서 k 는 회귀변수의 수(number of regressors), 즉 회귀자유도(regression degree of freedom)를 의미한다. 따라서, C_k 기준에서 H_k 의 역할은 매우 단순하다. 식(4.20)에서 C_p 통계량의 p 는 $1 + \text{tr}(H_k)$ 로 대체된다. 여기에서 후자는 상수항(constant term)에 유효 회귀자유도(effective regression degrees of freedom)를 더한 것, 또는 다른 방법으로는, 능형회귀(ridge regression)로 여분(redundancy)을 제거한 후 상수항(constant term)에 회귀변수의 유효 갯수(effective number of regressors)를 더한 것이다. 물론, 이것은 $n - (1 + \text{tr}(H_k))$ 과 동등한 유효 오차자유도(effective error degrees of freedom)를 만들어낸다. 이 유효 자유도(effective degrees of freedom) 개념은 k 를 선택하는 다른 간단한 방법에서도 이용할 것이다.

PR (Ridge), C_k , GCV에서 $\text{tr}(H_k)$ 의 역할을 살펴보자. 우리는 다음 사실을 알고 있다.

$$\begin{aligned}\text{tr}(H_k) &= \text{tr}[X^*(X^*X^* + kI)^{-1}X^{*'}] \\ &= \text{tr}[X^*X^*(X^{*'}X^* + kI)^{-1}] \\ &= \sum_{i=1}^k \frac{\lambda_i}{(\lambda_i + k)}\end{aligned}$$

$\sum_{i=1}^k \text{Var } b_{i,R}$ 와 같이, 만약 심각한 공선성(collinearity)이 있다면, $\text{tr}(H_k)$ 는 감소하다가 이어서 평탄하게 될 것(leveling off)이다. C_k 와 GCV의 경우, $\text{tr}(H_k)$ 는 기준(criterion) 감소 후 증가를 유발하며, 이는 필연적으로 $SS_{\text{Res},k}$ 를 커지게 한다. 따라서

이러한 두 기준은, H_k 의 작용(behavior)으로 발현되는 분산감소(variance reduction)와 유도되는 편향(bias) 간의 상호교환(trade-off)의 결과이다. 후자는 $SS_{Res,k}$ 를 결국 팽창시킨다.

앞서 논의에서, 개별적 계수들에 대한 고려는 덜 하고, 예측변수(predictor) \hat{y} 에 더 집중하자는 원칙, 즉 예측(prediction)을 강조하는 능형회귀 절차들에 초점을 맞추었다. 분명히 능형추적법(ridge trace method)은 계수들의 안정성(stability) 자체에 관심을 둔다. 또한 계수의 질 추정(quality estimation)을 주요 테마로 하는 축소모수(shrinkage parameter)를 선택하는데 사용되는 다른 방법들이 있다. 어떤 응용영역에서는, 하나 이상의 회귀계수들(regression coefficients)의 부호(sign)와 크기(magnitude)로 시스템을 해석하는 것이 매우 중요하다. 그러나 최소제곱법(least squares)은 회귀계수들의 부호와 양이, 거의 혹은 전적으로 의미가 없으므로 많은 경우 최소제곱법을 거부하고 싶은 유혹을 받을 수 있다. 그러므로 원래의 접근방법(natural approach)은 계수가 합리적인 값이 되도록 능형추적(ridge trace)으로 k 값을 선택하는 것이다. 이 방법은 오로지 예측기준(prediction criterion)에만 의지하여 k 를 결정하는 원칙과는 다른 것이다. 이 방법들의 대부분에서, 계수들의 안정성(stability)과 추정(estimation)이 중요하다. k 를 선택하는 다양한 방법들에 대하여 더라고 싶다면 Draper et al. (1979)과 Hocking (1976)을 참고하라.

k의 비확률적인 선택과 자유도 추적(Nonstochastic Choices of k and the df-Trace)

반응 자료(response data)의 함수(function)가 아닌 축소모수(shrinkage parameter) 값을 사용하는 것이 확실히 유리하다. 능형회귀 추정량(ridge regression estimator)의 분산과 표준오차를 다루는 이전의 수식 전개에서, 우리는 상수 k 가 확률변수(random variable)가 아니라고 가정하였다. 그러나, 이전에 언급된 모든 k 선택 방법들은 y 자료(y-data)를 이용하고 있다. 결과적으로, 능형추적(ridge trace), C_k , PR (Ridge), GCV는 k 를 선택하는 확률적 방법(stochastic method)이다. 즉, k 의 선택이 확률변수(random variable)가 된다. k 를 회귀자료(regressor data)만의 함수로서 선택하는 경우가 있다. 따라서 k 의 선택은 공선성(collinearity) 자체의 특성에 의하여 결정된다. 이 경우 k 는 확률변수가 아니다. 결과적으로, 분산팽창요인(variance inflation factor), 즉 행렬의 대각원소들(diagonal elements of the matrix)은 계수들의 실제 분산을 반영하는데 사용될 수 있다.

$$\frac{Var b_R}{\sigma^2} = (X^{*'} X^* + kI)^{-1} (X^{*'} X^*) (X^{*'} X^* + kI)^{-1}$$

따라서 회귀계수들의 표준오차는 대각원소의 제곱근(square root of diagonal elements)-오차평균제곱의 근(root error mean square, $\sqrt{\text{error mean square}}$)을 곱한-으로 주어질 수

있다. k 의 확률적인 선택(stochastic choice)에서는, 앞의 식은 단지 능형추정량들의 분산공분산 행렬(variance-covariance matrix of the ridge estimators)의 근사(approximation)이며, 특정한 경우에는, 근사의 질에 관해서는 거의 이야기 할 수 없다.

k 의 비확률적 선택방법(nonstochastic choice) 중 단순한 한가지는 모든 계수들의 분산팽창인자들(variance inflation factors)이 충분히 감소할 때까지 k 를 증가시키는 것이다. 예를 들어서, 모든 VIFs의 값이 10.0이하가 되도록 하고 싶다. 분산팽창인자는 y 관측값(y-observations)에 의존하지 않기 때문에, k 의 선택에 대한 이 결정과정(decision procedure)은 비확률적(non-stochastic)이다.

좀더 매력적으로 보이는 두 번째 방법은 자유도 추적 기준(df-trace criterion)으로 불린다(Tripp, 1983). 이 기준은 H_k 행렬에 중심을 두고 있다. 독자는 k 를 선택하는데, C_k , PR(Ridge), GCV 기준에서 H_k 가 얼마나 중요한 역할을 하는지 기억하여야 한다. 여기서 제시되는 기준은 다음에 기초를 두고 있다.

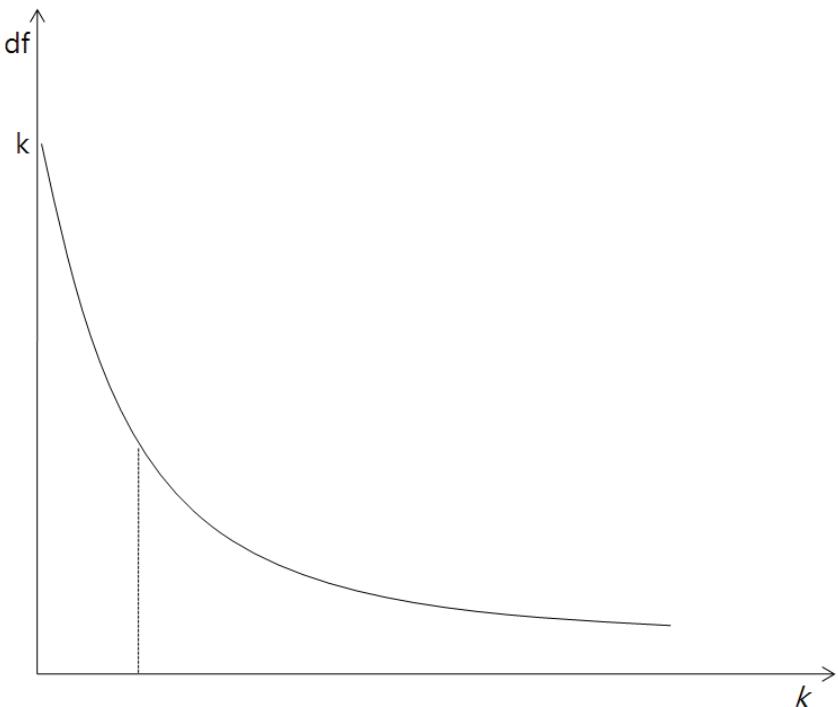
$$df = \text{tr}(H_k)$$

“df”는 자유도(degree of freedom) 즉, 앞서 설명되었던 유효 회귀자유도(effective regression degrees of freedom)를 의미한다. 이 방법은 자유도(df)가 안정되는 k 를 선택한다는 관점과 함께 k 에 대하여 df의 그림을 그린다. 이 방법은 모든 계수들이 안정되는 k 를 선택하는 능형 추적(ridge trace)과 매우 유사하다.

자유도 추적기준(df-trace criterion)은 유효 회귀자유도(effective regression degrees of freedom) 혹은 자료행렬의 유효 순위(effective rank of the data matrix)가 안정 또는 결정(settling) 될 때까지 k 의 증가를 허용하는 견고한 접근방법이다. 자유도(df)는 공선성(collinearity)의 구조(structure)에만 토대를 두고, k 를 선택하는 비확률적인 절차(non-stochastic procedure)임을 명심하여야 한다. 이 방법은 예측 기반의 기준(prediction based norm) 즉, C_k 와 GCV와는 결코 분리될 수 없다.

자유도 추적의 전형적인 그림은 Fig. 8.1에 있다. 예제 8.8은 자유도 추적을 실례를 들어서 설명한다.

Fig. 8.1 A typical df-trace and choice of k



예제 8.8 체내 지방자료

Table 8.8의 체내 지방자료를 다시 고려해보자. 최적의 k 를 결정하기 위하여, 앞에서 다룬 통계량 PRESS, C_k , df-trace를 k 의 각각의 값에서 계산하여 결과들을 Table 8.9에 정리하였으며, 시각적인 판단을 위하여 그림을 그렸다.

Table 8.10 k 결정을 위한 통계량

k	$b_{0,R}$	$b_{1,R}$	$b_{2,R}$	$b_{3,R}$	PRESS	C_k	$tr(H_k)$
0.00	124.6	4.628	-3.086	-2.306	172.017	4.000	3.000
0.01	7.056	1.06	-0.035	-0.433	162.362	4.515	2.612
0.02	3.618	0.956	0.054	-0.378	162.510	4.248	2.440
0.03	2.427	0.92	0.085	-0.359	162.560	4.081	2.343
0.04	1.823	0.901	0.101	-0.349	162.583	3.969	2.28
0.05	1.457	0.89	0.11	-0.343	162.593	3.890	2.236
0.06	1.212	0.882	0.117	-0.339	162.598	3.830	2.204
0.07	1.037	0.877	0.121	-0.337	162.599	3.784	2.179
0.08	0.905	0.873	0.125	-0.334	162.598	3.747	2.159
0.09	0.802	0.87	0.128	-0.333	162.595	3.717	2.142

0.10	0.72	0.867	0.13	-0.332	162.591	3.691	2.129
0.11	0.652	0.865	0.132	-0.33	162.587	3.669	2.117
0.12	0.596	0.863	0.133	-0.33	162.582	3.65	2.107
0.13	0.548	0.862	0.134	-0.329	162.576	3.634	2.098
0.14	0.507	0.86	0.135	-0.328	162.570	3.619	2.09
0.15	0.472	0.859	0.136	-0.327	162.564	3.606	2.083
0.16	0.441	0.858	0.137	-0.327	162.557	3.595	2.077
0.17	0.413	0.857	0.138	-0.327	162.551	3.584	2.071
0.18	0.389	0.856	0.139	-0.326	162.544	3.574	2.066
0.19	0.367	0.856	0.139	-0.326	162.537	3.565	2.062
0.20	0.348	0.855	0.140	-0.325	162.53	3.557	2.057

이 예제에서 능형회귀는 회귀변수 x_2 의 계수를 음수에서 양수로 변화시켰다. PRESS 개념(concept)을 사용하여 얻어진 계수 추정값(coefficient estimates)을 다른 예측 기준(prediction criteria)과 비교하기 위하여, 그림을 그렸으며 여기에 C_k 가 같은 방식으로 그려져 있다(Fig. 8.3). PRESS와 C_k 가 같은 결과를 주지 않는다는 사실은 또한 흥미롭다. 전에 언급한 바와 같이, 이것은 높은 지렛대 관측값(high leverage observations)이 우세할 때 예상되는 것으로서, 이러한 관측값들은 실제 이 자료세트에 존재한다.

자유도 추적기준(df-trace criterion)을 같은 자료에 적용하였다. Table 8.9와 Fig. 8.4는 0부터 0.2까지의 k 에서 일부 수치 결과들을 나타낸다. $k = 0$ 와 $k = 0.05$ 에서 $tr(H_k)$ 의 감소가 얼마나 되는지 주목하라.

체내 지방자료의 사례에서 k 를 선택하는 본 예제에서, 원리(philosophies)가 약간만 달라도 k 를 선택하는 방법이 같지 않을 것이라는 것을 보게 된다. PRESS 기준(PRESS criterion)은 계수들을 가장 덜 보수적으로 편향추정(biased estimation)하여 k 값이 가장 크며, 반면에 자유도 추적(df-trace)은 k 값이 가장 작다. 예측지향적이나 $tr(H_k)$ 의 영향을 많이 받는, C_k 기준(C_k criterion)을 사용하면 k 값은 중간 정도이다. PRESS로 구한 값과 다른 기준으로 구한 값 간의 차이는, 높은 지렛대 관측값(high leverage observations)이 자료에 있을 때 가장 두드러질 것이다. 따라서 분석가는 회귀의 목적을 분명히 알고 있어야 한다. 계수의 해석이 중요하다면 능형 추적이나 자유도 추적방법이 중요하다. 또한 다른 방법들도 연구되어야 한다. 예측이 중요하다면, GCV, C_k , PRESS가 시도되어야 한다.

축소모수인 k 를 선택하는 다양한 방법 모두를 소개하지는 않았다. 다 이야기하자면 책 한권을 가득 채우게 될 것이다. 우리는 독자에게 대표적으로 뚜렷이 구분되는 원리를 가진

몇 가지 방법들을 소개하려고 하였다. 사용된 방법이 특정 자료세트에서 가능한 최상의 것인지에 대한 보증은 결코 할 수 없음을 강조한다. 다중공선성 진단(multicollinearity diagnostics)으로 문제가 있음이 판단되고, 단순히 변수를 제거하는 방법(simple variable deletion)으로는 해결되지 않는 경우, 능형회귀의 하나 이상의 방법을 사용하고 싶을 수도 있다. 그러므로 목표는 어떤 방법 혹은 방법들을 주목하여야 하는지 판단하는데 필요한 경험을 획득하는 것이다. 이처럼 다양한 방법들을 하나의 컬렉션으로 볼 수도 있다. 이 컬렉션의 개별 방법마다 여러가지 다양한 세트의 편향 추정값들(biased estimates)을 구할 수 있다. 만약 진단을 하여 공선성에 의한 손상을 확인한 경우, 선택한 모든 방법들이 OLS 보다 더 나은 것일 수 있다는 사실에 사용자는 어느 정도 편안함을 얻는다. 확실히, 다중공선성을 가진 모형적합 문제에 능형회귀를 사용하면 안되지만, 능형회귀는 모형 구축가가 반드시 알고 있어야 할 테크닉 중 하나임에는 틀림없다.

Fig. 8.2 Plot of PRESS against k

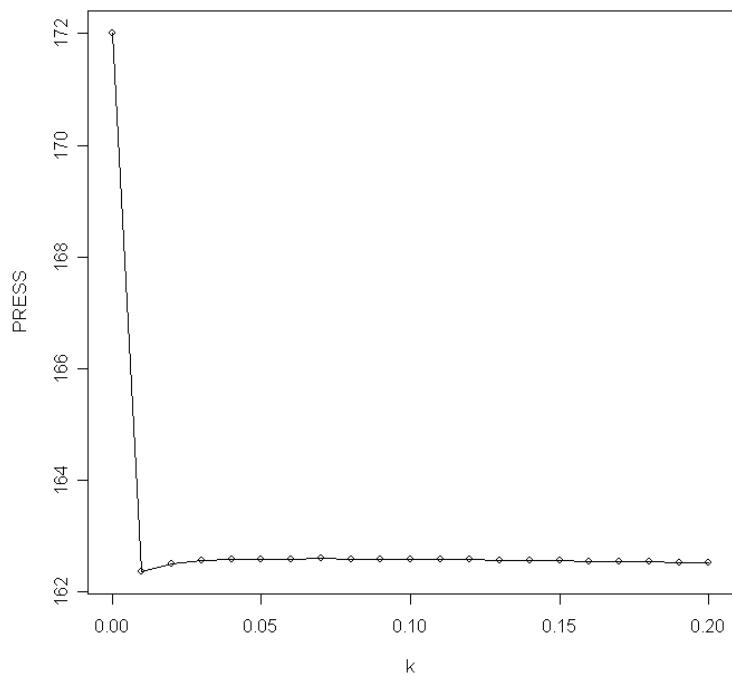


Fig. 8.3 Plot of C_k against k

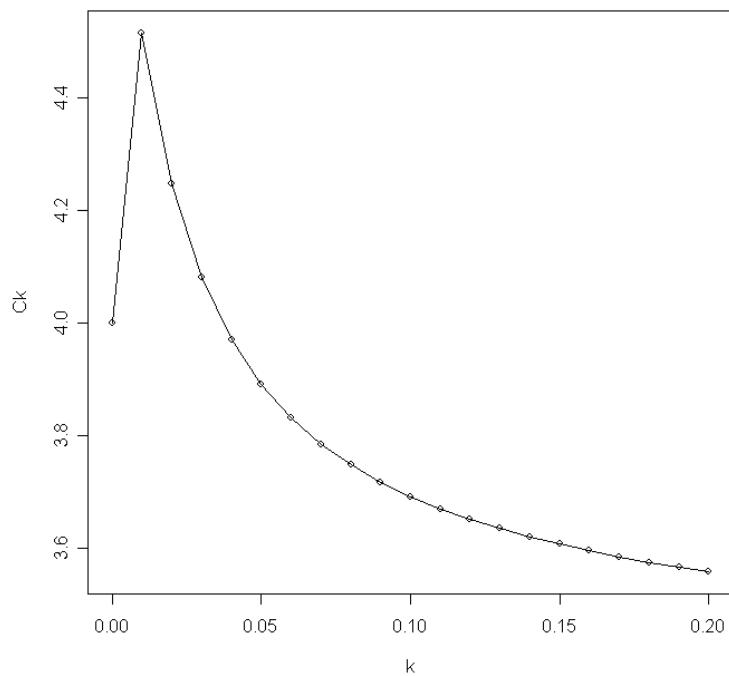
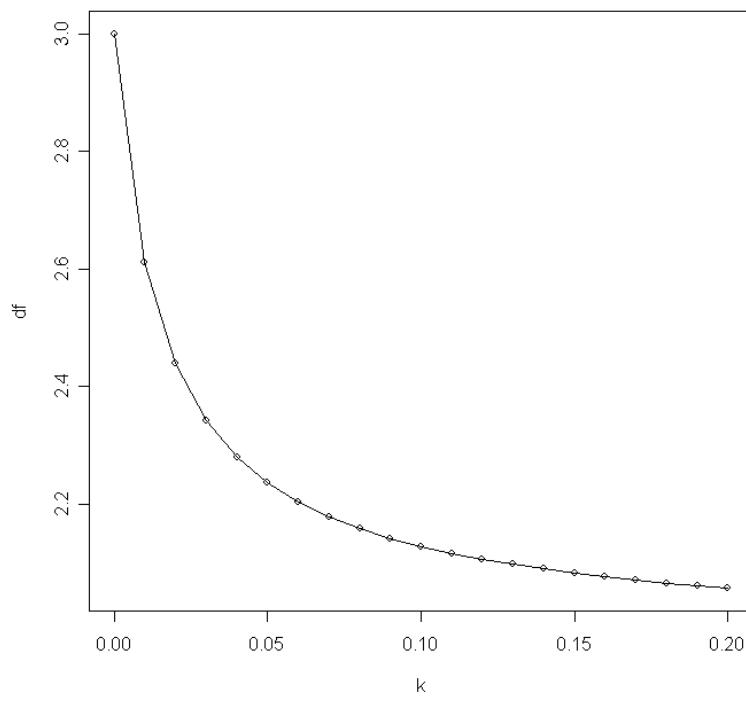


Fig. 8.4 Plot of df-trace



예제에 사용된 R-code는 다음과 같다.

```
data<-read.table("c:/data/ex8_7.R",header=TRUE)
x.m<-as.matrix(data[,1:3])
one<-rep(1,19)
x<-as.matrix(cbind(one,x.m))
y<-as.matrix(data[,4])
xx1<-data[,1]
xx2<-data[,2]
xx3<-data[,3]
x1<-(xx1-mean(xx1))/sd(xx1)
x2<-(xx2-mean(xx2))/sd(xx2)
x3<-(xx3-mean(xx3))/sd(xx3)
x.s<-cbind(x1,x2,x3)
iden<-diag(1,4)
iden1<-diag(1,3)
result<-matrix(rep(0,21*8),nrow=8)
g<-lm(y~.,data=data)
sig<-summary(g)$sigma^2

for(i in 1:21){
  k<-(i-1)/100
  be<-solve((t(x)%*%x+k*iden))%*%t(x)%*%y
  h<-x%*%solve((t(x)%*%x+k*iden))%*%t(x)
  h_d<-diag(h)
  h.s<-x.s%*%solve((t(x.s)%*%x.s+k*iden1))%*%t(x.s)
  h.s_d<-diag(h.s)
  result[1,i]<-k
  result[2:5,i]<-be
  res<-sum((y-x%*%be)^2)
  d<-sum(h.s_d)
  c<-res/sig-19+2+d
  press<-sum(((y-x%*%be)/(1-h_d))^2)
  result[6,i]<-press
  result[7,i]<-c
  result[8,i]<-d }
```

```

plot(result[1,],result[6,],xlab="k",ylab="PRESS",type="p")
lines(result[1,],result[6,])
plot(result[1,],result[7,],xlab="k",ylab="Ck",type="p")
lines(result[1,],result[7,])
plot(result[1,],result[8,],xlab="k",ylab="df",type="p")
lines(result[1,],result[8,])

```

주성분 회귀(Principal Components Regression)

주성분회귀는 다중공선성을 제거하는 또 하나의 편향추정기술(biased estimation technique)이다. 이 방법을 이용하여, 상관행렬(correlation matrix)의 주성분(principal components)으로 불리는 한 세트의 인공변수(a set of artificial variables)에 대하여 최소제곱추정(least squares estimation)을 수행할 것이다. 분석의 본질에 기초하여, 분산을 상당하게 감소시키는 몇 개의 주성분을 제거한다. 이 방법은 능형회귀의 원리와 다소 다르나, 비슷한 점은 편향추정값(biased estimate)을 구해준다는 것이다. 성공적으로 사용되었을 때, 이 방법은 추정(estimation)과 예측(prediction)이 OLS보다 우월하다. 주성분은 서로에 대하여 직교(orthogonal)하므로, 특정 분산량(specific amount of variance)이 서로에 기인하는 것으로 하기가 매우 쉽다.

X^*X^* (상관 형태, correlation form)의 고유값 $\lambda_1, \lambda_2, \dots, \lambda_k$ 과 연관된 정규화 고유벡터(normalized eigenvectors)의 행렬을 생각해보자. V 는 직교행렬(orthogonal matrix)이므로 $VV'=\mathbf{I}$ 이다. 그러므로 원래의 회귀모형을 다음의 형태로 작성할 수 있다.

$$y = \beta_0 1 + X^* V V' \beta + \varepsilon \quad (8.16)$$

$$y = \beta_0 1 + Z \alpha + \varepsilon \quad (8.17)$$

여기에서 $Z = X^*V$ 이고 $\alpha = V'\beta$ 이다. Z 는 $n \times k$ 행렬이며, α 는 새로운 계수들 $\alpha_1, \alpha_2, \dots, \alpha_k$ 의 $k \times 1$ 벡터이다. Z 의 행들(columns) 즉, 전형원소(typical element) z_{ij} 는 k 개의 새로운 변수들 즉, 주성분들(principal components)을 대표하는 것으로 볼 수 있다. 주성분들이 서로에게 직교인 것은 쉽게 알 수 있다.

$$\begin{aligned}
Z'Z &= (X^*V)'(X^*V) \\
&= V'X^*X^*V \\
&= \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)
\end{aligned} \quad (8.18)$$

따라서, 식(8.17)의 모형으로 z 에 대해서 회귀를 수행하면, 계수들의 분산(variances) 즉, $(Z'Z)^{-1}$ 의 대각 원소들(σ^2 은 별도로 하고)은 고유값(eigenvalues)의 역수(reciprocals)이다. 즉,

$$\frac{Var(\hat{\alpha}_j)}{\sigma^2} = \frac{1}{\lambda_j} \quad (j=1,2,\dots,k) \quad (8.19)$$

$\hat{\alpha}$ 들은 최소제곱추정량이라는 것을 주목하라. 만약 모든 주성분들이 회귀모형에 남아 있다면, 변환에 의하여 수행된 모든 것들은 본질적으로 회귀변수들의 회전(rotation)이다. 비록 새로운 변수들이 직교(orthogonal) 할지라도, 변환 전과 같은 크기(magnitude)의 분산($X'X$ 의 나쁜 조건 때문에)을 가지게 된다. 어떤 의미로는, 전체 분산(total variance)이 단지 재분포될(redistributed) 뿐이라고 볼 수 있다. 만약 다중공선성이 심각하다면, 적어도 하나의 작은 고유값(small eigenvalue)이 있을 것이다. 작은 고유값과 연관된, 적어도 하나 혹은 한 개의 주성분(principal component)을 제거하는 것은 모형의 전체 분산(total variance)을 상당히 감소시킬 것이며, 분명히 향상된 예측 방정식을 만들어 낼 것이다.

주성분은 무엇인가? (What are Principal Components?)

주성분 z 들을 만들어 내는 회귀변수들의 회전(rotation)은, 새로운 계수들의 세트 즉, α 들을 허용한다. 이 α 들은 추정량 $\hat{\alpha}_j$ 의 분산이 직접적으로 특정 선형의존성(linear dependency) 때문이라고 할 수 있게 정의된다. 이것은 식(8.19)로부터 명백하다. Fig. 8.7은 주성분회귀(principal components regression)에서 z 들이 실제로 무엇인지 더 잘 이해할 수 있도록 해 줄 것이다.

분명하게, 자료는 x_1 과 x_2 사이의 강한 연관성(association)을 보여주고 있다. 이제 이 의존성(dependency)이 β_1 과 β_2 모두의 추정값(estimates)에 영향력을 미치게 될 것이다. 그러나 z 좌표계(z-coordinate system)을 생각해보자.

$$Z = X^*V$$

는 다음과 같이 주어지는 Z 의 특정 열(column) 하나를 가진다.

$$z_j = X^*v_j \quad (8.20)$$

z_j 원소들(elements)은 z_j 축에서 측정된 자료이며, 여기에서는 $j = 1, 20$ 이다. z_j 값의 변동(variation) 다음과 같이 주어진다.

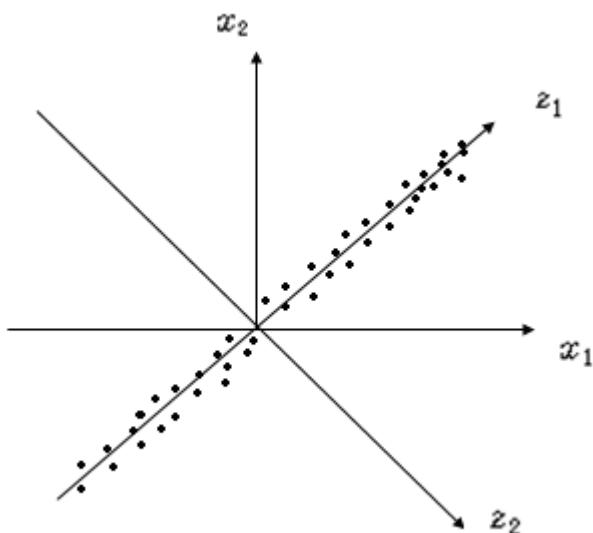
$$z'_j z_j = v'_j (X^* X) = \lambda_j \quad (j=1,2)$$

따라서 이 경우에서 주성분(principal components)에 대한 회귀(regression)는 다음을 포함한다.

$$Z'Z = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (8.21)$$

Fig. 8.5에서 λ_1 은 큰 고유값(large eigenvalue)이며 λ_2 는 작은 고유값(small eigenvalue)이다. 계수 $\hat{\alpha}_2$ 의 분산(variance)은 $\sigma^2 \frac{1}{\lambda_2}$ 로 주어진다. z_2 방향의 작은 변동(variation)은 $\hat{\alpha}_2$ 의 큰 분산(large variance)의 원인이 된다. 자료의 변동은 주로 z_1 축을 따라서 놓여있다. 따라서 큰 λ_1 은 $Var(\hat{\alpha}_1)$ 이 x_1 과 x_2 사이의 의존성에 의하여 상대적으로 영향을 받지 않도록 해준다. 앞의 설명에서, z 들로 변환(transformation)하면, 작은 고유값(small eigenvalues)이 특징인 선형의존성들이, 작은 수의 회귀계수에 예리하게 집중되도록 해준다. 이것은 어떤 z 들을 제거하여야 하는지에 대한 판단을 쉽게 해준다. Fig. 8.7의 사례에서 명백하게, 자료는 z_1 방향으로 가정되며, 따라서 주성분 z_2 는 본질적으로 회귀에 아무런 기여를 하지 않는다. 따라서 z_2 는 제거될 것이며, $\hat{\alpha}_2$ 에 의한 분산은 감소하게 된다. 그러므로 주성분회귀는 주성분들에 대한 회귀에서 변수를 선별(screening)하는 것에 지나지 않는다.

Fig. 8.5 Principal components for k=2



얼마나 많은 주성분들이 제거되는가? (How Many Principal Components Are Eliminated?)

주성분(pc) 회귀의 원리는 일반적으로 최소제곱 변수선별(least squares variable screening) 원리와 아주 많이 닮아 있다. 최소제곱추정(least squares estimation)이 성분들(components)에 대해서 수행되는 것이다. 만약 하나의 성분이 제거되면, 원래 변수들(original variables)인 x 들의 계수 추정량은 편향된다(biased). 분산의 원인이 되는 성분들을 제거하였기 때문에, 최소제곱 변수선별(least squares variable screening)과 마찬가지로 분산은 감소한다. 제거하여야 하는 성분(components)이 얼마나 많은지(만약 있다면) 판단하기가 어렵다. 기본적으로는 보통형태의 최소제곱기준(ordinary type of least squares criteria), s^2 , **PRESS**, 절대 **PRESS** 잔차합(sum of absolute PRESS residuals), C_p , $E(y)$ 에 대한 신뢰구간의 폭(width of the confidence interval on $E(y)$) 등이 기준이 될 수 있다. 예제 8.10은 이에 관한 설명을 담고 있다.

원래 변수로의 변환(Transformation Back to Original Variables)

주성분회귀(principal components regression)에 대하여 반대하는 사람들을 흔히 볼 수 있는데 이는 주성분들(principal components)에 대한 인위성(artificiality) 때문이다. 주성분회귀를 성공적으로 수행하였다면, 원래 변수(original variables)에서의 모형은 향상될 것으로, 의심할 여지없이 예상할 수 있다. 물론 s^2 , **PRESS** 등의 계산 통계량(computed statistics)을 원래의 표준화된 변수들(original standardized variables)로 변환된 모형에 적용해 볼 수 있다. 예를 들어서, k 개의 변수와 따라서 k 개의 주성분이 있는 경우, r ($< k$)개의 성분들(components)이 제거되었다고 가정해보자. 식(8.17)에서 모든 성분들이 남아있으면, 우리는 $\alpha = V'\beta$ 로 쓸 수 있다. 따라서,

$$\beta = V\alpha \quad (8.22)$$

마지막 r 개의 성분을 제거한다면, 모든 k 개의 모수들(parameters)에 대한 회귀계수들(regression coefficients)의 최소제곱추정량들(least squares estimators)은 다음과 같이 주어진다(주성분을 삭제하였다고 해서 원래의 회귀변수가 삭제되는 것은 아니다).

$$b_{pc} = \begin{bmatrix} b_{1,pc} \\ b_{2,pc} \\ \vdots \\ b_{k,pc} \end{bmatrix} = [v_1 v_2 \cdots v_{k-r}] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{k-r} \end{bmatrix} \quad (8.23)$$

따라서 r 개의 주성분들을 제거하였다는 것은 r 개의 고유벡터들(eigenvectors) 그리고 α 들 중에서 r 개를 제거하는 것과 동등한 것이다. x 들은 중심척도화되어(centered and scale) 변환된 모형(transformed model)의 상수항(constant term)은 \bar{y} 가 된다는 것을 가정한다.

주성분 계수들의 편향(Bias in Principal Components Coefficients)

r 개의 주성분은 제거되고 s 개의 성분은 유지되어 $s + r = k$ 인, 주성분 절차(principal components procedure)를 생각해보자. 또한 $X^* X^*$ 의 정규화 고유벡터(normalized eigenvectors)의 행렬 $V = [v_1 v_2 \dots v_k]$ 가 다음과 같이 분할된다고 가정해보자.

$$V = [V_r \ : \ V_s]$$

그리고 이와 유사하게, 행렬 Λ 을 $X^* X^*$ 의 고유값들(eigenvalues)의 대각행렬(diagonal matrix)로 가정해보자. 행렬 Λ 는 다음과 같이 분할된다.

$$\Lambda = \begin{bmatrix} \Lambda_r & 0 \\ 0 & \Lambda_s \end{bmatrix}$$

여기서 Λ_r 과 Λ_s 는 대각행렬이며, Λ_r 는 제거된 성분들과 연관된 고유값들을 가지고 있다. $V'(X^* X^*)V = Z'Z = \Lambda$ 이므로, α 들의 최소제곱추정량은 다음과 같이 쓸 수 있다.

$$\hat{\alpha} = (Z'Z)^{-1} Z'y = \Lambda^{-1} V' X^{*'} y \quad (8.24)$$

이것은 남아있는 α 들의 추정량이 다음으로 주어짐을 암시한다.

$$\hat{\alpha}_s = \Lambda_s^{-1} V_s^{-1} X^{*'} y$$

앞에서 지적한 바와 같이, 우리는 주성분회귀(principal components regression)를 주성분들(principal components)에 대한 표준최소제곱 모형구축(standard least squares model-building)으로 간주하여야 한다. 주성분들은 직교하기(orthogonal) 때문에, 우리는 $\hat{\alpha}_s$ 가 α_s 의 불편 추정량(unbiased estimator)임을 보이기 위하여 식(4.7)을 사용할 수 있다. 이제 b_{pc} 에 대한 식(8.23)을 생각해보자. 다음과 같이 쓸 수 있다.

$$b_{pc} = V_s \hat{\alpha}_s \quad (8.25)$$

따라서,

$$E(b_{pc}) = V_s \alpha_s = V_s V_s' \beta$$

$$VV' = I = V_r V'_r + V_s V'_s \text{ 이므로,}$$

$$\begin{aligned} E(b_{pc}) &= [I - V_r V'_r] \beta \\ &= \beta - V_r V' \beta \\ &= \beta - V_r \alpha_r \end{aligned}$$

따라서 p 개의 회귀계수의 추정량은 양(quantity) $V_r \alpha_r$ 만큼 편향되며(biased), 이때 α_r 은 제거되었던 주성분의 벡터이다.

주성분 계수들의 분산(Variance in Principal Components Coefficients)

주성분(principal components)을 제거하면, b_{pc} 의 회귀계수들(regression coefficients)의 분산(variances)이 감소된다. 분산 감소폭은 능형회귀와 마찬가지로 다중공선성의 정도에 따라서 달라진다. 분산 감소(variance reduction)가 무엇인지 분석적으로 판단하는 것은 상대적으로 쉬운 일이다. 만약 모든 성분들이 남아있다면, b_{pc} 는 보통최소제곱(ordinary least squares)으로 환원된다. 따라서,

$$\boxed{\begin{aligned} \frac{Var b}{\sigma^2} &= (X^* X^*)^{-1} \\ &= V A^{-1} V' \\ &= V_r A_r^{-1} V'_r + V_s A_s^{-1} V'_s \end{aligned}} \quad (8.26)$$

식(8.25)에서, $Var(\hat{\alpha}_s) = \Lambda_s^{-1}$ 이라는 사실을 이용하면, 다음과 같은 분산공분산행렬을 가지게 된다.

$$\boxed{\frac{Var b_{pc}}{\sigma^2} = V_s \Lambda_s^{-1} V'_s} \quad (8.27)$$

OLS 추정량과 주성분 추정량에 대한 분산공분산행렬의 차이는 양(quantity) $V_r \Lambda_r^{-1} V'_r$ 이다. 이 행렬의 대각원소들(diagonal elements)은, 제거된 주성분들과 연관된 고유값들(eigenvalues)의 역수(reciprocals)의 가중합(weighted sums)이다. 따라서 무시된 주성분들(ignored principal components)이 작은 고유값들(small eigenvalues)과 연관된다면,

상당한 분산 감소(variance reduction)를 예상할 수 있다.

예제 8.10 주성분 예제(Principal Components Example)

다음 자료는 17개 병원 의사들의 연 근무시간에 영향을 미칠 것으로 예상되는 다섯 개의 설명변수를 포함하고 있다.

y : 연 근무시간

x_1 : 하루 평균 환자수

x_2 : 월간 X-ray 촬영 횟수

x_3 : 월간 이용 병상수

x_4 : 해당 지역의 병원이용가능 인구/1000

x_5 : 평균재원일

x_1	x_2	x_3	x_4	x_5	y
15.57	2463	472.92	18	4.45	566.52
44.02	2048	1339.75	9.5	6.92	696.82
20.42	3940	620.25	12.8	4.28	1033.15
18.74	6505	568.33	36.7	3.9	1603.62
49.2	5723	1497.6	35.7	5.5	1611.37
44.92	11520	1365.83	24	4.6	1613.27
55.48	5779	1687	43.3	5.62	1854.17
59.28	5969	1639.92	46.7	5.15	2160.55
94.39	8461	2872.33	78.7	6.18	2305.58
128.02	20106	3655.08	180.5	6.15	3503.93
96	13313	2912	60.9	5.88	3571.89
131.42	10771	3921	103.7	4.88	3741.4
127.21	15543	3865.67	126.8	5.5	4026.52
252.9	36194	7684.1	157.7	7	10343.81
409.2	34703	12446.3	169.4	10.78	11732.17
463.7	39204	14098.4	331.4	7.05	15414.94
510.22	86533	15524	371.6	6.35	18854.45

최소제곱회귀(least squares regression)로 다음의 예측 방정식(prediction equation)을 얻었다.

$$\hat{y} = 1962.8886 - 15.8614x_1 + 0.0559x_2 + 1.5899x_3 - 4.21832x_4 - 394.3009x_5$$

여기에서, $SS_{\text{Res}} = 4535004$, $s^2 = 412273$, **PRESS** = 32194998, $R^2 = 0.99080$ 이다.

다음은 다중공선성 진단이다(분산비율과 고유값들).

Eigenvalue	Portion Intercept	Portion b_1	Portion b_2	Portion b_3	Portion b_4	Portion b_5
5.2013	0.001	0.000	0.002	0.000	0.001	0.000
0.6666	0.014	0.000	0.011	0.000	0.002	0.006
0.0791	0.026	0.000	0.379	0.000	0.012	0.018
0.0447	0.009	0.000	0.464	0.000	0.294	0.017
0.0082	0.805	0.000	0.142	0.001	0.254	0.757
0.0000	0.146	0.999	0.003	0.999	0.438	0.200

Condition Number = 182597.7

$\mathbf{X}'\mathbf{X}$ 의 이 고유값들은 척도화만 되고(scaled) 중심화되지 않은(uncentered) \mathbf{X} 행렬에 대한 것이다. 계수 b_1, b_2, b_3, b_4, b_5 에 대한 분산팽창요인은 각각 9597.063264, 7.940596, 8932.631828, 23.292392, 4.279938이다.

설명변수들을 각각 표준화시켜서 얻은 X^* 에서 구한 상관행렬(correlation matrix)은 다음과 같다.

$$X^* X^* = \begin{bmatrix} 1.0000000 & 0.9073795 & 0.9999040 & 0.9356913 & 0.6711974 \\ 0.9073795 & 1.0000000 & 0.9071495 & 0.9104688 & 0.4466496 \\ 0.9999040 & 0.9071495 & 1.0000000 & 0.9331684 & 0.6711088 \\ 0.9356913 & 0.9104688 & 0.9331684 & 1.0000000 & 0.4628609 \\ 0.6711974 & 0.4466496 & 0.6711088 & 0.4628609 & 1.0000000 \end{bmatrix}$$

상관행렬(중심척도화된 경우의 고유값들이 척도화만 된 경우와 어떻게 다른지 주목하라)의 고유값들은 $\lambda_1=4.197117e+00, \lambda_2=6.674841e-01, \lambda_3=9.463322e-02, \lambda_4=4.071166e-02, \lambda_5=5.397137e-05$ 로 주어지며, 고유벡터의 행렬은 다음과 같다:

$$V = \begin{bmatrix} 0.4852857 & -0.0020282794 & -0.1662322 & 0.4681915 & 0.719484037 \\ 0.4532353 & -0.3356046632 & 0.8042394 & -0.1874697 & 0.001157632 \\ 0.4849767 & -0.0008557357 & -0.1539606 & 0.5092612 & -0.694079268 \\ 0.4609693 & -0.3107995079 & -0.5371982 & -0.6338609 & -0.023438351 \\ 0.3337371 & 0.8892515559 & 0.1152392 & -0.2907321 & -0.006781832 \end{bmatrix}$$

최대 분산팽창인자가 9597.063264이므로 다중공선성 문제가 존재함을 알 수 있다. 분산비율을 보면 각 회귀계수의 분산이 최소 고유값에 의하여 각각 0.999, 0.003, 0.999, 0.438, 0.200의 비율로 설명되므로, 두 설명변수 x_1 과 x_3 사이에 공선성이 존재한다고 판단된다. 상관행렬의 값들로 판단하더라도 설명변수들 사이에, 특히 x_1, x_2, x_3, x_4 간에 강한 선형종속관계가 있음을 짐작할 수 있다. 이 중에서 x_1 과 x_3 사이의 표본상관계수는 0.999로 매우 강한 상관관계가 있다. 만약 주성분회귀(principal components regression)가 성공적으로 수행된다면, 주성분의 하나 혹은 그 이상의 제거가 포함될 것이다.

주성분들의 \mathbf{Z} 행렬은 $\mathbf{Z} = \mathbf{X}^* \mathbf{V}$ 로 구할 수 있으며, 여기서 \mathbf{X}^* 는 중심척도화(centred and scaled) 된다. \mathbf{X}^* 는 17×5 행렬이다. 주성분들은 다음 행렬의 열(columns)이다.

$$\mathbf{Z} = \begin{pmatrix} z_1 & z_2 & z_3 & z_4 & z_5 \\ -1.8117 & -0.30609 & 0.003793 & 0.120048 & -0.00148 \\ -1.16504 & 1.111003 & 0.153527 & -0.10704 & -0.00574 \\ -1.80908 & -0.40993 & 0.063498 & 0.19816 & 0.001283 \\ -1.74265 & -0.73249 & 0.017235 & 0.094701 & -0.0023 \\ -1.24284 & 0.180371 & 0.048451 & -0.00119 & -0.00434 \\ -1.38486 & -0.38253 & 0.268858 & 0.155498 & 0.001887 \\ -1.14627 & 0.22486 & 0.009053 & -0.03041 & -0.00523 \\ -1.21993 & -0.05181 & -0.03732 & 0.040373 & 0.019687 \\ -0.58559 & 0.394313 & -0.10235 & -0.12853 & -0.009 \\ 0.269541 & -0.09984 & -0.23024 & -0.64433 & 0.009181 \\ -0.61267 & 0.200594 & 0.144881 & -0.0029 & -0.00201 \\ -0.48828 & -0.44453 & -0.30515 & 0.159428 & 0.008359 \\ -0.17553 & -0.23818 & -0.18855 & -0.15002 & -0.01003 \\ 1.468502 & 0.186947 & 0.297791 & -0.02696 & -0.00063 \\ 3.22479 & 2.29596 & 0.147436 & 0.1724 & 0.005231 \\ 3.554094 & -0.33631 & -0.86803 & 0.196054 & -0.00394 \\ 4.867499 & -1.59233 & 0.577123 & -0.04529 & -0.00091 \end{pmatrix}$$

\mathbf{Z} 의 다섯 번째 열은 최소 고유값과 연관된 주성분이다. z_{5i} 의 작은 변동에 주목하자.

주성분회귀는 z_5 를 제거하고 남은 주성분들에 대한 반응변수 y 의 회귀이다. 이 회귀는 다음과 같은 z 들의 계수들을 제공한다:

$$\begin{aligned}\hat{\alpha}_1 &= 0.47924758 \\ \hat{\alpha}_2 &= -0.14653063 \\ \hat{\alpha}_3 &= 0.06353581 \\ \hat{\alpha}_4 &= 0.54385134\end{aligned}$$

3장과 4장에서 논의된 바와 같이, 보통최소제곱절차(ordinary least squares procedures)를 회귀에 적용하여 잔차제곱합(residual sum of squares)과 **PRESS** 통계량(statistic)을 계산한다. 결과는 다음과 같다. 이것은 주성분 자료와 표준화 종속변수와의 회귀적합 결과이다.

Before Deletion	After Deletion
$SS_{Res} = 0.1467$	$SS_{Res} = 0.1481$
$s^2 = 0.0133$	$s^2 = 0.0123$
$PRESS = 1.041251$	$PRESS = 0.9490778$

성분의 수를 줄인 후 회귀를 하면 잔차제곱합은 약간 증가되나, **PRESS** 통계량과 s^2 은 감소된다. 모형의 개선은 다음의 **PRESS** 잔차로 설명된다.

observation	Before Deletion	After Deletion
	PRESS residual	PRESS residual
1	−0.043659981	−0.041292847
2	−0.010749222	−0.000536739
3	−0.015676345	−0.017691983
4	0.079097023	0.082113641
5	0.009465535	0.015421063
6	−0.119457724	−0.121869126
7	0.03346904	0.039750528
8	0.163586527	0.055385267
9	−0.097553373	−0.074075414
10	−0.189088243	−0.164690723
11	0.065493702	0.068009367
12	−0.149175181	−0.147718821
13	−0.054507317	−0.032382017
14	0.335169789	0.335957375
15	−0.451549646	−0.41932868
16	0.403281512	0.39191776
17	−0.660931308	−0.646661367

식(8.23)은 중심척도화 된 회귀변수들의 계수 추정값을 결정하기 위하여 사용할 수 있다. 회귀의 상수항(constant term)은 \bar{y} 이다. 다음은 8.3절에서 기술된 것과 같이, 원래 변수들(natural variables)의 계수(coefficients)로 변환(transformation)한다. 결과는 다음과 같다.

$$\begin{aligned}
 b_{1,pc} &= 0.4769341 \\
 b_{2,pc} &= 0.2155306 \\
 b_{3,pc} &= 0.4997297 \\
 b_{4,pc} &= -0.1123973 \\
 b_{5,pc} &= -0.1211531
 \end{aligned}$$

이상으로 다중회귀분석에서 독립변수들 사이에 다중공선성이 심각할 때 사용할 수 있는 주성분 회귀분석에 대하여 알아보았으며, 주성분의 삭제는 OLS regression보다 다소 우월한 회귀를 만들어내는 것으로 보인다.

다음은 예제에서 사용한 R-code이다.

```

data<-read.table('c:/manpower.txt',header=T)
attach(data)
gdata<-lm(y~x1+x2+x3+x4+x5,data)
summary(gdata)
anova(gdata)
inf<- influence(gdata)
pre <- (gdata$resi/(1-inf$hat))
sum(pre^2)
x<-as.matrix(data[-6])
x_<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
x_1<-cbind(x_,x)
x1<-scale(x_1,scale=TRUE,center=FALSE)
t(x1)%*%t(x1)
e<-eigen(t(x1)%*%t(x1)/16)
colldiag(data[-6],scale=TRUE,center=FALSE,add.intercept=TRUE)
vif(gdata)

y<-data[,6]
x<-as.matrix(data[,-6])
n<-length(y)
standx<-scale(x,scale=TRUE,center=TRUE) # 표준화 x matrix
standy<-scale(y,scale=TRUE,center=TRUE) # 표준화 y matrix
cor(standx)
e<-eigen(cor(standx))

```

```

z<-standx%*%e$vectors      #Z matrix of principal component
z

# 가장 작은 고유치와 관련된 주성분을 제거하고
alpha<-solve(t(z[,-5])%*%(z[,-5]))%*%t(z[,-5])%*%standy
alpha

data<-data.frame(z1=z[,1],z2=z[,2],z3=z[,3],z4=z[,4],z5=z[,5],y=standy)
g_1<-lm(standy~z1+z2+z3+z4,data)
summary(g_1)
anova(g_1)

inf<- influence(g_1)
afterpre1 <- (g_1$resi/(1-inf$hat))
sum(afterpre1^2) ## press

# 가장 작은 고유치와 관련된 주성분을 제거하지 않고
alphaall<-solve(t(z)%*%z)%*%t(z)%*%standy
alphaall

data<-data.frame(z1=z[,1],z2=z[,2],z3=z[,3],z4=z[,4],z5=z[,5],y=standy)
g_2<-lm(standy~z1+z2+z3+z4+z5,data)
summary(g_2)
anova(g_2)

inf<- influence(g_2)
beforepre <- (g_2$resi/(1-inf$hat))
sum(beforepre ^2) ## press

#press residual 비교
result<-cbind(beforepre,afterpre1)
result

# 가장 작은 고유치와 관련된 주성분을 제거하고 주성분회귀 후 beta.
ev<-cbind(e$vectors[,1],e$vectors[,2],e$vectors[,3],e$vectors[,4])
coe<-ev%*%alpha
coe

```

주성분들을 제거하는 적절한 순서는? (What Is the Appropriate Order of Elimination of Principal Components?)

예제 8.10을 통해서, 주성분들의 제거 전략(strategy of elimination of principal components)은 최소 고유값(smallest eigenvalue)과 연관된 주성분을 제거하는 것으로 시작되어야 한다는 것을 알았다. 이 방법은 대개 적절하며, 이에 대한 논리적 근거는, 최소 고유값과 연관된 주성분이 가장 중요하지 않은 성분인 동시에 가장 큰 분산량(the largest amount of variance)이 이 주성분 때문에 발생한다는 것이다. 그러나, 좀 더 수용가능한 전략은 마치 표준변수선별(standard variable screening) 문제인 것처럼(실제로도 그렇다), 주성분 감축(principal component reduction)을 다루는 것이다. 회귀변수(regressors)로서의 주성분들은 직교(orthogonal)이며, 주성분을 제거하는 순서를 알려주는 통계량 세트는 다음의 t 통계량(t -statistics)이다.

$$t = \frac{\hat{\alpha}_j}{s_{\hat{x}_j}} = \frac{\hat{\alpha}_j \sqrt{\lambda_j}}{s}$$

다시 말해서, t 값을 순서대로 배열하고(rank ordered), 최소 t 값을 가지는 성분부터 제거한다. 많은 경우 t 값의 배열 순서(rank order)는 고유값들의 내림차순이 될 것이지만, 반드시 그렇지 는 않다.

9. 비선형회귀(Nonlinear Regression)

전술한 모든 장에서, 모형 설명에 관한 주요한 가정은 구조가 모형계수(model coefficients)에 선형적(linear)이라는 것이다. 물리학, 화학, 공학, 생물학 등 많은 분야의 실험은 덜 경험적(empirical)이고, 더 이론적인(theoretically based) 비선형모형(nonlinear model)이 필요하다.

비선형모형에는 많은 예들이 있다. 다음과 같이 몇 개의 실례를 들어보자.

$$y = \alpha e^{\beta x} + \varepsilon \quad (9.1)$$

$$y = \frac{\alpha}{1 + \exp(-(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k))} + \varepsilon \quad (9.2)$$

$$y = \alpha + \beta_1 x_1^{\gamma_1} \beta_2 x_2^{\gamma_2} + \dots + \beta_k x_k^{\gamma_k} + \varepsilon \quad (9.3)$$

$$y = \alpha \exp[-\beta_1 e^{-\beta_2 x}] + \varepsilon \quad (9.4)$$

식(9.1)-(9.4)의 모든 모형에서 모수(parameter) 중 적어도 하나는 모형에 비선형적인 방식으로 들어가 있는 것을 주목하라. 선형모형과 마찬가지로 가장 먼저 할 일은 모수(parameters)를 추정하는 것이다. 또 다시 최소제곱방법(least squares technique)이 광범위하게 사용된다.

9.1. 비선형 최소제곱(Nonlinear least Squares)

비선형모형을 위한 최소제곱추정량(least squares estimators)을 전개(development)할 때, 선형모형에서는 볼 수 없었던 복잡한 문제에 직면하게 된다. 이는 식(9.1)의 지수회귀모형(exponential regression model)의 경우를 보면 쉽게 알 수 있다. 자료세트 (y_i, x_i) ($i = 1, 2, \dots, n$)가 주어지면 α 와 β 의 추정량(estimators)은 아래의 식을 최소화함으로써 구해진다.

$$SS_{\text{Res}} = \sum_{i=1}^n (y_i - \alpha e^{\beta x_i})^2 \quad (9.5)$$

(9.5)를 α 와 β 에 대하여 미분하고 각 도함수(derivative)를 0로 놓으면 다음의 식을 얻을 수 있다.

$$\sum_{i=1}^n (y_i - \alpha e^{\hat{\beta} x_i}) \left(-e^{\hat{\beta} x_i} \right) = 0 \quad (9.6)$$

$$\sum_{i=1}^n (y_i - \alpha e^{\hat{\beta} x_i}) \left(-\hat{\alpha} e^{\hat{\beta} x_i} \cdot x_i \right) = 0 \quad (9.7)$$

(3.4)의 최소제곱추정식(least squares estimating equation)과는 다르게, 식(9.6)과 (9.7)은 모두 추정량 $\hat{\alpha}$ 와 $\hat{\beta}$ 에 대하여 비선형이다. 따라서 기초행렬대수(elementary matrix algebra)로는 추정량을 계산할 수 없다. 즉, 반복과정(iterative process)을 사용하여야 한다.

비선형모형 상황에서 잔차제곱합(residual sum of squares)을 최소화하기 위한 알고리즘이 많은 문헌에 소개되어 있다. 또한, 최소한 하나 이상의 비선형 추정방법을 가지고 있는 회귀분석 컴퓨터 패키지도 많이 있다. 다음 절에서 이 방법의 세부사항을 기술할 것이다. 그러나, 비선형 추정을 시작하기 전에, 비선형 추정량(nonlinear estimators)의 성질에 대하여 알려진 것과 알려지지 않은 것이 무엇인지 알아야 한다.

9.2. 최소제곱추정량의 특성(Properties of the Least Squares Estimators)

3장에서 일반선형모형(general linear model)을 위한 최소제곱추정량(least squares estimators)의 성능(performance)에 관한 상세한 내용을 기술하였다. 7장에서는, 비이상적인 조건이어서 가정이 맞지 않는 경우에 적용하는 방법들을 논의할 때, 역시 이러한 성능 특성을 살펴보았다. ε_i 가 정규분포를 하고, 독립적이면서 등분산(homogeneous variance)일 때, 추정량들은 모든 불편 추정량들(unbiased estimators)의 최소분산(minimum variance)을 가진다. 만약 정규분포 가정이 완화된다면, 추정량들은 모든 선형 불편추정량들(linear unbiased estimators)의 최소분산을 가진다.

이러한 특성들의 기초는 식(3.2) 모형에서 모수(parameter)에 대한 선형성(linearity)이다. 불행하게도 비선형모형의 최소제곱추정량은 이러한 특성을 가지지 않는다. 다음과 같이 일반적 공식(general formulation)으로 모형을 기술할 수 있다.

$$y_i = f(x_i, \theta) + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (9.8)$$

여기서 θ 는 p 개의 모수들(parameters)을 포함하는 벡터이고 $n > p$ 이다. 물론 f 는 $\hat{\theta} = [\theta_1, \theta_2, \dots, \theta_p]$ 에 대하여 비선형이라고 가정한다. 아래의 식을 최소화하는 벡터 $\hat{\theta}$ 를 θ 의 추정량이라고 가정하자.

$$SS_{\text{Res}} = \sum_{i=1}^n [y_i - f(x_i, \hat{\theta})]^2 \quad (9.9)$$

또한 ε_i 는 정규분포이고, 독립적이면서 평균은 0이고 공통분산 σ^2 를 가진다고 가정하자. $\hat{\theta}$ 은 θ 의 최대우도추정량(maximum likelihood estimator)이다. 그러나 표본의 크기가 큰 경우를 제외하고는, 추정량의 특성에 대하여 어떤 일반적인 설명도 할 수가 없다. 달리 말하면 유일한 특성은 점근적 특성들(asymptotic properties)이다. $\hat{\theta}$ 의 추정량들은 일반적으로 불편추정량이 아니지만, 극한에서는 불편, 최소분산추정량(unbiased and minimum variance estimators)이다. 즉, 불편성(unbiasedness)과 최소분산특성(minimum variance properties)은 표본크기가 커지면서 도달할 수 있다. 따라서 특정 비선형모형과 특정 표본크기에서 추정량의 성질에 관하여 정확하게 설명할 수 있는 것이 없다. 모수에 대한 근사 신뢰구간(approximate confidence interval)을 구하고, t 통계량(t-statistic)을 만드는 데 점근적

분산공분산(asymptotic variance-covariance) 결과들을 사용할 수는 있다. 상세한 내용과 설명은 9.3절에서 하겠다.

9.3절과 9.4절에서 최소제곱추정값(least squares estimates)을 찾아주는 계산 방법에 대하여 상세하게 기술할 것이다. 9.1절의 예제에서 비선형회귀는, 선형회귀처럼 추정값을 직접 계산할 수 없다는 것을 알았다. 그러나 많은 상용 소프트웨어패키지에 이러한 계산 알고리즘이 들어가 있다. 이들은 기법과 철학에서 다소 차이가 있으나, 다양한 영역에서 성공적으로 사용되고 있다.

9.3. 추정량을 찾기 위한 가우스, 뉴튼방법(The Gauss-Newton Procedure for finding Estimates)

비선형모형의 최소제곱추정량 $\hat{\theta}$ 를 찾기 위한 소프트웨어 계산 알고리즘들 중 가장 많이 사용하는 방법은 가우스, 뉴튼 방법(Gauss-Newton Procedure)이다. Bard (1974), Draper와 Smith (1981), Kennedy와 Gentle (1980) 그리고 Bates와 Watts (1988)을 참조할 것.

또한 수정 가우스, 뉴튼방법들(modified Gauss-Newton methods)도 많이 있으며, 이 중에서 몇 개는 이 장에서 설명할 것이다. 기본적으로 이 방법(procedure)은 반복적(iterative)이고, 모수의 초기추정값(starting value estimates)이 필요하다. 이러한 추정값을 벡터 $\theta'_0 = (\theta'_{1,0}, \theta'_{2,0}, \dots, \theta'_{p,0})$ 로 표시하겠다. 식(9.9)에서 잔차제곱합(residual sum of square)을 최소화하는 θ 를 얻기 위하여, 먼저 $\theta = \theta_0$ 근처에서 (9.8)의 비선형 함수를 테일러 급수 전개(Taylor series expansion)하고 선형항들(linear terms)만 남긴다. 따라서

$$\begin{aligned} f(x_i, \theta) &\equiv f(x_i, \theta_0) + (\theta_1 - \theta_{1,0}) \left[\frac{\partial f(x_i, \theta)}{\partial \theta_1} \right]_{\theta=\theta_0} + (\theta_2 - \theta_{2,0}) \left[\frac{\partial f(x_i, \theta)}{\partial \theta_2} \right]_{\theta=\theta_0} \\ &+ \dots + (\theta_p - \theta_{p,0}) \left[\frac{\partial f(x_i, \theta)}{\partial \theta_p} \right]_{\theta=\theta_0} \quad (i=1,2,\dots,n) \end{aligned} \quad (9.10)$$

식(9.10)은 식(9.8)의 비선형 형태(nonlinear form), $f(x_i, \theta)$ 를 기본적으로 선형화(linearization)시킨 것이다. 식(9.10)은 초기값 주변에서의 선형근사(linear approximation)로도 볼 수 있다.

(9.10)에서 선형화를 잘 살펴보면 다음과 같은 형태하는 것을 알 수 있다.

$$f(x_i, \theta) - f(x_i, \theta_0) \equiv \gamma_1 w_{1i} + \gamma_2 w_{2i} + \dots + \gamma_p w_{pi} \quad (i = 1, 2, \dots, n) \quad (9.11)$$

여기에서

$$w_{ji} = \left[\frac{\partial f(x_i, \theta)}{\partial \theta_j} \right]_{\theta=\theta_0}$$

은 j 번째 모수(θ_j)에 대한 비선형함수 $f(x_i, \theta)$ 의 도함수(derivative)를 나타낸다. 여기서

도함수는 모든 초기값에서 계산된다. 그리고

$$\gamma_j = \theta_j - \theta_{j,0}$$

식(9.11)의 좌변을 잔차 $y_i - f(x_i, \theta)$ 로 생각할 수 있는데, 여기서 모수들(parameters)은 초기값으로 대치된다. w_{ji} 는 이미 알고 있고, 선형회귀의 회귀변수(regressor variables) 역할을 하는 반면에, γ 는 모수값(parameter value)과 초기값(starting value)간의 차이(difference)이며, 회귀계수(regression coefficient) 역할을 한다. 결과적으로, 가우스, 뉴튼방법은 아래와 같은 선형회귀구조(linear regression structure)를 만든다.

$$y_i - f(x_i, \theta_0) = \gamma_1 w_{1i} + \gamma_2 w_{2i} + \dots + \gamma_p w_{pi} + \varepsilon_i \quad (i = 1, 2, \dots, n) \quad (9.12)$$

가우스, 뉴튼방법은 기본적으로 식(9.12)의 모형에 대한 선형회귀분석(linear regression analysis)이다. 물론, (9.12) 모형은 한번에 직접적인 연산(operation)으로 답을 구할 수 없다. γ 를 추정하여 θ 의 추정값을 구하는데, θ 는 초기추정값 θ_0 보다 개선(improvement)된 값으로 볼 수 있다. 이 연산(operation)은 θ_0 근처로부터 개선된 θ 의 추정값 근처로 초점을 바꾸어 준다. 이런식으로 연산의 초점이 계속 바뀐다. 정확한 반복 과정(iterative process)은 다음과 같다:

- 1) 모형 (9.12)에서 $\gamma_1, \gamma_2, \dots, \gamma_p$ 는 선형최소제곱(linear least squares)으로 추정한다.

첫번째 반복추정값(first iteration estimates)을 $\gamma_{1,1}, \gamma_{2,1}, \dots, \gamma_{p,1}$ 로 표기할 것이다.

- 2) $\hat{\theta}_{j,1} = \theta_{j,0} + \hat{\gamma}_{j,1} \quad (j = 1, 2, \dots, p)$ 를 계산한다. 이것이 최종추정값(final estimates)은 아니다. 대신에 $\hat{\theta}_{1,1}, \hat{\theta}_{2,1}, \dots, \hat{\theta}_{p,1}$ 는 첫번째 반복값(first iteration values)을 의미한다.

- 3) 단계 2에서 구한 $\hat{\theta}$ 값으로 모형 (9.12)의 초기값을 대치한다.

- 4) 단계 1로 돌아가 $\gamma_{1,2}, \gamma_{2,2}, \dots, \gamma_{p,2}$ 를 계산하고, $\hat{\theta}_{1,2}, \hat{\theta}_{2,2}, \dots, \hat{\theta}_{p,2}$ 을 구한다.

- 5) 수렴(convergence)될 때까지 이 과정을 계속한다. 수렴이란 r 번의 반복(r iterations) 후에 잔차제곱합(residual sum of squares)과 모수 추정값(parameter estimates)이 더 이상 변하지 않는 것을 의미한다.

단계 2의 이전 반복(previous iteration)에서 구한 추정값(estimated)에 각 반복(iteration)마다 $\hat{\gamma}_j$ 를 더한다. 즉, $\hat{\gamma}_j$ 는 증분(increments)이다. 수렴에 도달하면, 이 증분(increments) 즉, $\hat{\gamma}_j$ 는 무시해도 좋을 만큼 작아진다. 따라서 가우스, 뉴튼 방법은 현 추정값들(current estimates)에서 계산한 도함수들(derivatives)인 w 들에 대하여, 잔차들(residuals)을 회귀시키는 선형회귀의 연속이다. 각 단계마다 추정값 $\hat{\theta}_j$ 들이 갱신(updated)된다. 최소잔차제곱합(minimum residual sum of squares)은 추정값들(estimated)이 수렴되었을 때 얻어진다. 정식으로 말하자면, s 번째 반복(s th iteration)의 추정값 벡터(estimate vector) $\hat{\theta}_s = (\hat{\theta}_{1,s}, \hat{\theta}_{2,s}, \dots, \hat{\theta}_{p,s})$ 는 다음 식에 의하여 ($s-1$) 번째 반복의 추정값 벡터와 관계지울 수 있다.

$$\boxed{\hat{\theta}_s = \hat{\theta}_{s-1} + (W'_{s-1} W_{s-1})^{-1} W'_{s-1} [y - f(\hat{\theta}_{s-1})]} \quad (9.13)$$

여기서 W_{s-1} 은 (i, j) 원소(element)가 $[\partial f(x_i, \theta) / \partial \theta_j]_{\theta=\theta_{s-1}}$ 인 $n \times p$ 행렬이다. $[y - f(\hat{\theta}_{s-1})]$ 는 잔차벡터(vector of residuals)이며, 여기에서 함수 즉, $f(x_i, \theta_{s-1})$ 를 포함하고 있는 n 차원 벡터는, $\hat{\theta}_{s-1}$ 에서 계산된다.

가우스, 뉴튼 비선형추정으로 구한 계수들은, 이 계수들의 점근적 분산공분산행렬(asymptotic variance-covariance matrix)에 기초하여 추론한다. $\hat{\theta}$ 의 점근적 분산공분산행렬의 추정값은 다음과 같이 주어진다.

$$\boxed{\hat{Var}(\hat{\theta}) = s^2 (W' W)^{-1}} \quad (9.14)$$

여기서 W 는 앞에서 설명한 편미분 행렬(matrix of partial derivatives)이며 마지막 반복(final iteration)에서 얻어지는 최소제곱추정값(least squares estimates)에 대하여 계산된다. 그리고 s^2 은 익숙한 잔차평균제곱(residual mean square)이며 다음과 같다.

$$s^2 = \sum_{i=1}^n \frac{[y_i - f(x_i, \hat{\theta})]^2}{n-p}$$

식(9.14)의 결과는 예제 9.1과 9.2에서 설명하겠다.

가우스, 뉴턴 방법의 어려움(Difficulties with the Gauss-Newton Procedure)

사용자들은 여기에서 설명한 가우스, 뉴튼 방법을 사용하는데 어려움을 경험할 수 있다. 앞에서 γ 들로 기술하였던 증분 변화량(incremental change)이, 어떤 경우 추정이 좋지 않을 수 있다는 사실에서 단점들이 발생한다. 이것의 결과는 많은 반복(iteration)이 필요하고 따라서 수렴이 매우 늦어진다는 것이다. 어떤 경우에는, $\hat{\gamma}$ 의 부호가 잘못되어, 가우스, 뉴튼 방법 자체가 잘못된 방향으로 진행할 수도 있다. 정말로, 잔차제곱합(residual sum of squares)이 계속 증가하면서 수렴이 되지 않기도 한다.

수정 가우스, 뉴튼방법은 원래의 가우스, 뉴튼방법의 여러 어려움을 완화시키기 위하여 고안된 것들이다. 비선형 회귀를 하는 대부분의 상용 회귀 컴퓨터패키지들은 가우스, 뉴튼 방법뿐만 아니라 수정 가우스, 뉴튼방법들도 제공한다. 수정 가우스, 뉴튼방법들 중 어떤 것은 회귀계수들의 계산된 증분변화량(calculated incremental change)을 필요할 경우, 직접 줄이도록 고안된 것도 있다. 어떤 것은 이러한 증분의 구조(structure of increment)가 미묘하지만 주요한 방식으로 변한다.

분수 증분의 사용(Use of Fractional Increments)

보통 가우스, 뉴튼방법(ordinary Gauss-Newton procedure)은, 반복 시마다 $\hat{\gamma}_j$ 의 크기를 줄일 수 있도록 즉, 도약크기(jump size)를 허용함으로써 향상시킬 수 있다. 목표는 특정 반복(specific iteration)의 계산된 증분(computed increment)이 SS_{Res} 를 증가시키는 것을 허용하지 않는 것이다. 이를 위하여, 다음과 같은 합리적인 전략을 소개한다.

- 1) $(W'_{s-1} W_{s-1})^{-1} W'_{s-1} [y - f(\hat{\theta}_{s-1})] = \hat{\gamma}_{s-1}$ 을 사용하여 s 번째 반복($s = 1, 2, \dots$)을 위한 표준 가우스, 뉴튼 증분 벡터(standard Gauss-Newton increment vector)를 계산한다.
- 2) 가우스, 뉴튼 방법이 제시하는 대로 $\hat{\theta}_s = \hat{\theta}_{s-1} + \hat{\gamma}_{s-1}$ 를 계산한다.
- 3) 만약 $SS_{Res,s} < SS_{Res,s-1}$ 이면, $\hat{\theta}_s$ 를 이용하여 다음 반복을 계속한다.
- 4) 만약 $SS_{Res,s} > SS_{Res,s-1}$ 이면, 단계 2로 되돌아가서, 증분벡터(increment vector)로 $\frac{\hat{\gamma}_{s-1}}{2}$ 을 사용한다.

한번의 반복 동안 필요하다면 10번이라도 등분(halving)을 한다. 일반적으로, 증분이 약

10번정도 등분되고 SS_{Res} 가 더 이상 감소되지 않으면 반복과정을 멈춘다.

분수 증분(fractional increments)에 기초한 방법론은 정확하게 고전적인 가우스, 뉴튼 선형화(classical Gauss-Newton linearization)의 전략적 변형(strategic modification)이지만, **가우스, 뉴튼(Gauss-Newton)**이라는 말은 등분 기전(halving mechanism)이 있는 기본 가우스, 뉴튼 방법(basic Gauss-Newton procedure)을 뜻하는 용어로서 사용된다. 어떤 컴퓨터 패키지는 가우스, 뉴튼 방법에서 자동적으로 등분과정(halving procedure)을 실시한다. 이 변형은 가우스, 뉴튼 추정방법(Gauss-Newton estimation procedure)을 개선시키나, 사용자들은 이것이 사용되었는지 아닌지를 알아야 한다.

예제 9.1 여성의 연령별 자녀 양육 비율 자료

다음은 1945년도에 적어도 한 자녀를 가진 여성의 비율을 연령별로 나타낸 Table이다.

연령대	비율
15	0
20	17
25	60
30	82
35	88
40	90

자료를 살펴보면 급격한 증가 현상을 보이다가 일정 지점에서는 유지되는 경향을 볼 수 있다. 즉, 선형이 아닌 로그 그래프와 같은 모양의 비선형모형이 요구된다. 성장모형은 다음과 같은 형이 시도된다.

$$y_i = \alpha \exp(-\beta \exp(-\gamma x_i)) + \varepsilon_i \quad (i = 1, 2, 3, 4, 5, 6)$$

시작값은 모형의 성격을 고려하여 선택할 수 있다. 가우스, 뉴튼 방법은 5번의 반복(iteration)으로 수렴(convergence)한다. Table 9.1.0에 반복 시마다 모수 추정값을 기술하였다.

$\hat{\alpha} = 90.402$, $\hat{\beta} = 468.059$, $\hat{\gamma} = 0.282$, 잔차제곱합(residual sum of squares)은 0.11840에서 수렴되었다.

$\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ 의 근사 표준오차(approximate standard error)를 계산하는데 사용되는 오차평균제곱(error mean square)은 다음과 같이 얻어진다.

$$s^2 = \frac{SS_{\text{Res}}}{n-2} = 0.1184/4 = 0.0296$$

물론 모수에 대한 검정과 신뢰구간에 관한 추론의 많은 것은 식(9.14)에 의하여 결정된 점근적 분산공분산행렬(asymptotic variance-covariance matrix)에서 끌어낸다. 여기서 모수 추정값(parameter estimates)에서 도함수(derivatives)의 W 행렬을 계산하여야 하고 $s^2(W'W)^{-1}$ 를 계산한다. 이 결과는 다음과 같다.

$$s^2(W'W)^{-1} = \begin{bmatrix} 0.0258145092 & -2.12609674 & -2.260396e-04 \\ -2.1260967423 & 508.34042600 & 4.965316e-02 \\ -0.0002260396 & 0.04965316 & 4.921422e-06 \end{bmatrix}$$

$\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ 의 점근적 표준오차(asymptotic standard error)는, $s^2(W'W)^{-1}$ 의 대각원소(diagonal elements)의 제곱근을 계산하여 나온, $S_{\hat{\alpha}} = 0.1607$, $S_{\hat{\beta}} = 22.5464$, $S_{\hat{\gamma}} = 0.0022$ 이다. 모수의 점근적 신뢰구간들(asymptotic confidence intervals)은 대부분 비선형 상용소프트웨어 패키지(nonlinear commercial software package)의 기본적인 부분이다. 이들은 t -분포(t -distribution)에 기초를 두고 있다. 이 예에서, 잔차 자유도(residual degrees of freedom)는 3이고, 95% 신뢰구간은 $\hat{\alpha} \pm t_{0.025, 3} S_{\hat{\alpha}}$ 와 $\hat{\beta} \pm t_{0.025, 3} S_{\hat{\beta}}$ 와 $\hat{\gamma} \pm t_{0.025, 3} S_{\hat{\gamma}}$ 로 이루어진다.

수치적 결과는 다음과 같이 얻어진다

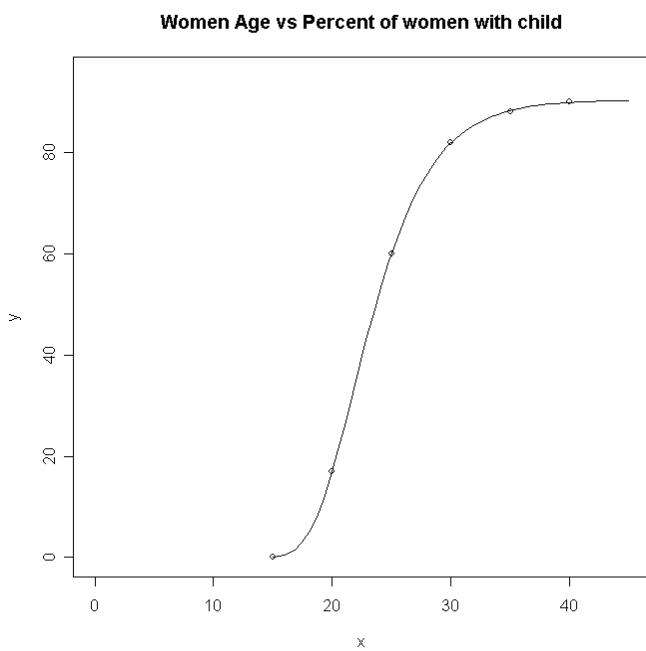
	Upper	Lower
α	90.9133	89.8907
β	539.8016	396.3164
γ	0.289	0.275

Table 9.1 Information from the Five Gauss-Newton Iterations

Iteration	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	SS_{Res}

0	89.00	942.00	0.31	17.54198
1	89.693560	589.818745	0.294899	4.928705
2	90.3980752	443.0846528	0.2808512	1.970093
3	90.4282929	465.8806783	0.2815535	0.1213901
4	90.4253236	468.0539854	0.2817029	0.1184228
5	90.4253417	468.0574737	0.2817027	0.1184227

Fig. 9.1 성장모형



다음의 Table은 적합값과 잔차를 나타내고 있다.

Women's Age	Women's Year of Birth (%)	\hat{y}	$y - \hat{y}$
15	0	0.09659105	-0.09659105
20	17	16.97364895	0.02635105
25	60	60.06930290	-0.06930290
30	82	81.81946462	0.18053538
35	88	88.24099222	-0.24099222
40	90	89.88631063	0.11368937

아래는 위 예제에 대한 R code이다.

```
x <- c(15, 20, 25, 30, 35, 40)
y <- c( 0, 17, 60, 82, 88, 90)
fit <- nls(y~A*exp(-B*exp(-C*x)), start=list(A=89, B=942, C=0.31), trace=T)
predict(fit)
residuals(fit)
vcov(fit)
plot(x, y, ylim=c(0, 95), xlim=c(0, 45), main='Women Age vs Percent of women with child')
A <- 90.425334
B <- 468.05914
C <- 0.28170291
x1 <- seq(from=15, to=45, length=50)
y1 <- A*exp(-B*exp(-C*x1))
lines(x1, y1, type="l")
```

9.4. 가우스, 뉴튼 방법의 다른 변형(Other Modification of the Gauss-Newton Procedure)

가우스, 뉴튼 방법의 어려움은 대부분 선형함수와 아주 다르게 작용하는 함수 때문이다. 더 상세한 내용은 Ratkowsky (1983)을 참조할 것. 또한 시작값의 질(the quality of starting values)이 중요한 역할을 한다. 이미 설명한 등분 과정(halving procedure)이 확실히 방법을 향상시켰으나, 가우스, 뉴튼 방법의 또 다른 중요한 변형 방법들이 있다. Hartley (1961)는 수렴이 보장되어, 보통의 가우스, 뉴튼 방법(ordinary Gauss-Newton Procedure)에서 경험하는 어려움으로 고생하지 않아도 되는 변형을 제안하였다. 이것은 각 반복(iteration)마다 SS_{Res} 의 성질에 따라 증분 크기(increment size)를 다시 줄인다. 두번째로, 가장 대중적인 변형 방법은 Marquardt (1963)에 의하여 개발되었다. Marquardt 방법(procedure)은 증분 변화의 벡터를 계산하기 위한 식(9.13)의 수정에 기초한다. 이 방법에서, s 번째 반복을 위한 증분의 벡터의 구조(structure of the vector of increments)는 아래 식의 해 $\hat{\gamma}_s$ 에 의해 얻어진다.

$$(W_s W_s + \lambda I_p) \hat{\gamma}_s = W_s' [y - f(\hat{\theta}_s)] \quad (\lambda > 0) \quad (9.15)$$

Marquardt의 논문에 있는 정보는, 식(9.15)에서 주어진 상수 λ 를 사용하면 왜 수렴이 향상되는지에 대하여 기하학적인 통찰력(geometric insight)을 제공해준다. 선형회귀에서 다중공선성(multicollinearity)과 편향된 추정(biased estimation)에 대하여 잘 알고 있는 독자는 식(9.15)과 회귀계수의 능형회귀추정량(ridge regression estimator) 사이의 유사성을 이해하여야 한다. Marquardt 방법은 각 반복마다 증분 변화의 추정량(estimators of the incremental changes)을 개선시키기 위한 능형회귀 접근방식(ridge regression approach)으로 볼 수 있다. 독자들은 왜 능선회귀가 합리적인가에 대해 이해하여야 한다. 가우스, 뉴튼 방법은 보통 최소제곱(ordinary least squares)의 결과로서 증분 변화를 계산한다; 왜냐하면 회귀변수(regressor variables)들은 같은 함수의 도함수이고, 능형회귀는 피할 수 없는 다중공선성의 자연스런 해법이기 때문이다. 따라서, 식(9.15)에서 λ 값은 기본적으로 능선회귀에서 k 값이다. W_s 행렬의 열(columns)을 척도화하는 것(scaling)은 $W_s W_s$ 의 대각(diagonal)에 unity elements를 만들기 위해서이다.

λ 의 계산(Computation of λ)

λ 의 계산 과정은 당연히 상당한 주의력을 요구한다. 이것은 컴퓨터 소프트 패키지에

따라 차이가 있을 수 있다. 그러나, 수렴을 확실히 하기 위해서는 다음과 같은 부등식을 만족해야만 한다.

$$SS_{Res}(\hat{\theta}_{s+1}) \leq SS_{Res}(\hat{\theta}_s) \quad (s = 0, 1, 2, \dots) \quad (9.16)$$

Marquardt는 반복(iteration) 시마다 SS_{Res} 를 감소시키는 값을 찾기 위한 시행착오적 평가방법(trial and error evaluation)을 제안하였다. 보통 가우스, 뉴튼 방법(ordinary Gauss-Newton Procedure)으로 만족스럽게 수렴할만한 상황이면 작은 λ 값을 사용하여야 한다. 식(9.16)을 만족시키는 때에만 비교적 큰 λ 값을 사용하여야 한다.

물론 반복할 때마다 λ 를 계산하는 유일한 알고리즘이나 방법은 없다. 실행할 수 있는 한 방법은, SAS에서 사용되는 NLIN 프로시저로 $\lambda = 10^{-8}$ 로 시작하는 것이다. 매 반복 시마다 일련의 시행착오적 계산(trial and error computation)을, 식(9.16)이 만족될 때까지 λ 를 10씩 곱하면서 시행한다. 또한 이 과정은 식(9.16)이 만족되는 한, 매 반복 시마다 10 단위로 감소하게 된다. 이러한 방법은 반복(iteration) 시마다 SS_{Res} 가 개선되도록 함과 동시에 λ 는 가능한 작게 유지시켜준다.

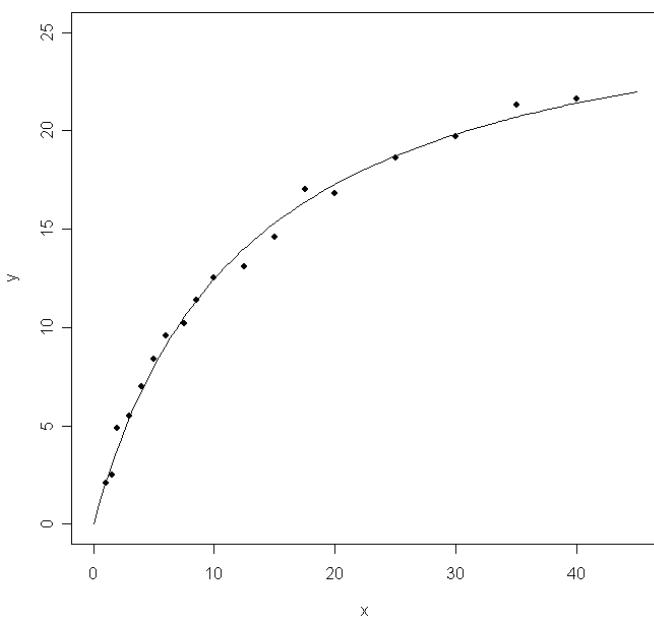
예제 9.2 식물의 성장에 관한 자료

동식물들의 성장에 관한 연구에서 성장치(Y)와 시간(X)의 관계를 나타내는 회귀식들 중의 하나는 다음과 같이 주어진다.

$$Y = \frac{\theta_1 X}{\theta_2 + X}$$

이 예제에서, 우리는 θ_1 과 θ_2 를 추정하기 위한 모형을 세우고, 산점도를 그리면 모수의 값을 정확히 추정하지 않더라도 대략적인 값을 알 수 있다. 이를테면, 산점도에서 최대 성장치를 나타내는 θ_1 은 22정도의 값이 되고 θ_2 는 이것의 반인 9정도의 값을 가진다고 할 수 있다. 그러므로 비선형 회귀분석에서는 각 모형에 포함된 모수들의 의미를 정확하게 먼저 파악하는 것이 모형의 해석과 분석에 필수적이다.

Growth of plant



Observation	X	Y
1	1	2.1
2	1.5	2.5
3	2	4.9
4	3	5.5
5	4	7.0
6	5	8.4
7	6	9.6
8	7.5	10.2
9	8.5	11.4
10	10	12.5
11	12.5	13.1
12	15	14.6
13	17.5	17.0
14	20	16.8
15	25	18.6
16	30	19.7
17	35	21.3
18	40	21.6

수렴은 네 번째 반복에서 얻어졌다. 선택된 시작값은 $\theta_1 = 22$, $\theta_2 = 9$ 였다. 반복 시마다

정보는 다음과 같다.

Iteration	θ_1	θ_2	SS_{Res}
0	22	9	54.47659
1	27.54322	12.23545	4.710087
2	28.12124	12.55969	4.302396
3	28.13652	12.57386	4.302271
4	28.13703	12.57443	4.302271

여기서 초기값은 Michaelis-Menten 식의 선형화(linearizing the Michaelis-Menten Equation)에 의해 쉽게 결정할 수 있었다. 예를 들면, 만약 x 에 대한 $1/y$ 으로 위치를 정한다면, 합리적인 θ_1 는 역기울기이다. 또한 θ_2 는 그 기울기의 절편과 θ_1 의 곱으로 계산할 수 있다.

점근적 분산공분산행렬(asymptotic variance-covariance matrix)는 다음과 같다.

$$s^2(W'W)^{-1} = \begin{bmatrix} 0.5299517 & 0.5202807 \\ 0.5202807 & 0.5822480 \end{bmatrix}$$

여기서 $s^2 = 0.2688422$ 이다

이 행렬에서 $\hat{\theta}_1$ 와 $\hat{\theta}_2$ 의 표준오차 추정값들은 $S_{\hat{\theta}_1} = 0.728$, $S_{\hat{\theta}_2} = 0.7631$ 인 것을 쉽게 확인할 수 있다.

아래는 위 예제에 대한 R code이다.

```

x <- c(1, 1.5, 2, 3, 4, 5, 6, 7.5, 8.5, 10,
      12.5, 15, 17.5, 20, 25, 30, 35, 40)
y <- c(2.1, 2.5, 4.9, 5.5, 7.0, 8.4, 9.6,
      10.2, 11.4, 12.5, 13.1, 14.6, 17.0,
      16.8, 18.6, 19.7, 21.3, 21.6)
fit <- nls(y~A*x/(B+x), start=list(A=22, B=9), trace=T)
predict(fit)
residuals(fit)
vcov(fit)
plot(x, y, ylim=c(0, 25), xlim=c(0, 45), main='Growth of plant', type="p", pch=19)
lines(fit)

```

```
A <- 28.137  
B <- 12.574  
x1 <- seq(from=0, to=45, length=100)  
y1 <- A*x1/(B+x1)  
lines(x1, y1, type="l")
```

9.5. 특별한 비선형 모형(Some Special Classes of Nonlinear Models)

비선형 모형화(nonlinear modeling)가 잘 적용되지 않는 분야들이 있다. 예제 9.1과 9.2에서처럼, 이론이나 지식으로 모형을 제시할 수 있는 분야는 대부분 성공적이다. 비선형 모형 적용의 선구자들은, 인자들(factors)이 서로 관계를 갖는 방법을 수학식(mathematical formulae)으로 나타내는 과학분야의 분석가들이었다. 이런 분야는 더 경험적인 선형모델(more empirical linear model)에 더 이상 의지할 필요가 없다. 우리는 특정 비선형모형을 합리적인 이해없이 적용하면 안된다고 강조한다. 많은 분석가들이 비선형모형 개발에 쉽게 성공하지 못하는데, 비선형모형에 대한 이해가 없는 상태에서 단지 자료에 적합한 선형모델을 찾을 수가 없어서 비선형모형을 적용하기 때문이다.

많은 비선형 모형은 특정 상황에 맞게 고안된 범주들로 분류된다. 각각의 범주마다, 여러 분야에서 분석가들이 성공적으로 사용한 많은 모형들이 있다. 다음의 부절에서, 우리는 몇가지 모형 범주들을 설명하여, 독자들이 적절히 적용할 수 있는 식견을 갖도록 할 것이다. 가장 잘 알려진 범주는 성장 모형(growth model)이며, 이것은 어떤 독립변수(시간 등)가 증가함에 따라 어떤 것이 어떻게 성장하는지를 설명하는데 사용하는 모형이다. 이것을 적용하는 표준 분야는 생물과 식물이 시간에 따라서 성장하는 생물학, 삼림학, 동물학 등이다. 경제, 인적자원 적용(manpower application), 신뢰성 공학(reliability engineering)까지도 성장모델을 적용할 수 있는 분야이다. 또한 성장모형은 화학요법 연구(chemotherapy research)에서도 주목을 받고 있다. Cater et al. (1983)을 참조할 것.

로지스틱 성장모형(The Logistic Growth Model)

로지스틱 회귀(logistic regression)는 (0, 1) 반응을 다루는 방법에 초점을 둔 것으로 7장에서 충분히 논의하였다. 로지스틱 성장(logistic growth)의 경우, 측정할 수 있는 양(measurable quantity) y 는 아래의 모형에 의하여 어떤 양 x 에 따라서 변한다.

$$y = \frac{\alpha}{1 + \beta \exp(-kx)} + \varepsilon \quad (9.17)$$

(9.17) 성장모형의 형태가 7장에서 설명한 이진형 반응(binary type response)를 다루는 모형과 비슷한 점을 주목하라. 로지스틱 모형에서, 모수들은 특별한 의미를 지닌다. $x = 0$ 이면 $y = \alpha / (1 + \beta)$ 가 되며, 따라서 이 양은 시간이 0 일 때, y 의 수준으로 생각할 수 있다. 한편, 중요한 모수 α 는 극한성장(limiting growth)으로 x 가 커질수록 y 가 접근해가는 값이다. β 와 k 값은 꼭 양수이어야 하며, 그래야만 로지스틱 함수를 해석할 수 있다. (9.17)의 로지스틱에서 x 에 대한 y 의 그림은 S자형태이다. Fig. 9.2와 9.3은 여러가지 k 와 β 에 대한 로지스틱 함수의 모양을 그림으로 설명한 것이다.

로지스틱 모형을 적용할 때 어떤 경우는, 지수(exponent)의 kx 부분이 좀 더 일반적인 선형 구조인 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$ 혹은 단일회귀변수 구조(single regressor structure)의 다항식(polynomial)으로 대체된다(Carter et al., 1983 참조).

Gompertz 성장모형(The Gompertz Growth Model)

Gompertz 모형은 성장 상황들(growth situations)에 적용하는 다른 S자형태 함수형식(S-shaped functional form)이다. 이 모형의 형식은 다음과 같다.

$$y = \alpha \exp[-\beta e^{-kx}] + \varepsilon \quad (9.18)$$

Gompertz 모형의 형식이 이중지수(double exponential)라는 것에 주목하라. 그리고 모수 α 는 한계성장(limiting growth)이다. $x=0$ 이면 $y = \alpha e^{-\beta}$ 이다.

Richards 성장모형(The Richards Growth Model)

Richards 모형(Richards(1959)를 참조하라)은 로지스틱 모형에 추가 모수가 적용된 단순한 변형이다. 함수 형식은 다음과 같다

$$y = \frac{\alpha}{[1 + \beta e^{-kx}]^{1/\delta}} + \varepsilon \quad (9.19)$$

Weibull 성장모형(The Weibull Growth Model)

Weibull 성장모형 또한 자료세트가 성장기전(growth mechanism)을 포함하고 있을 때 사용할 수 있다. 함수 형식(functional form)은 다음과 같다.

$$y = \alpha - \beta \exp[-\gamma x^\delta] + \varepsilon \quad (9.20)$$

이 경우에 $x = 0$ 일 때의 성장은 $\alpha - \beta$ 이고, 반면에 $x \rightarrow \infty$ 일 때 성장은 극대치(maximum)인 $y = \alpha$ 에 접근한다.

(9.17)-(9.20)에 있는 네 개의 모형이 독자들에게는 충분한 수의 함수형식이기는 하나, 이들이 성장모형의 전부는 아니다. 다른 형식과 더 많은 토론에 대해서는 Ratkowsky

(1983)과 Draper와 Smith (1981)을 참조하라.

Mitcherlich 법칙(The Mitcherlich Law)

우리의 관심을 성장모형에서 조금 바꾸어 농업, 화학, 공학에 유용한 모형 군으로 옮겨보자. 여기서 관심인 y 는 농작물 수확이나 화학반응과 같은 어떤 기전의 수율(yield of some mechanism)이고, x 는 수율을 증가시키는 힘(impetus)을 제공하는 회귀변수(regressor variable)로서 비료를 주는 빈도(fertilizer application rate), 시간 등일 수 있다. 전체적인 상황은 성장 상황과 확실히 유사하다. 대부분 Mitcherlich 법칙으로 알려진 모형에서 발전되었다. Phillips and Campbell (1968)을 참조하라. 모형은 다음과 같다.

$$y = \alpha - \beta\gamma^x + \varepsilon \quad (9.21)$$

모형 (9.21)은 또한 성장 상황에서도 사용된다. 로지스틱이나 Gompertz 모형에서와 같은 변곡점(point of inflection, 이 점은 $\partial^2 y / \partial x^2$ 가 0이 된다)이 없다: 따라서 이것은 S 형식이 아니다. (9.21)의 모형을 여러 가지 형태로 재모수화(reparameterization)시킨 것들이 많으며, 전부 같은 이름을 사용한다. 이 중 몇 개는 다음과 같다:

$$y = \alpha - \beta e^{-\gamma x} \quad (9.22)$$

$$y = \alpha - e^{-(\beta+\gamma x)} \quad (9.23)$$

$$y = \alpha [1 - e^{-\beta x}] \quad (9.24)$$

$$y = e^\alpha - \beta\gamma^x \quad (9.25)$$

식(9.24)의 모형은 예제 9.1에서 설명하였다. 이것은 또한 자주 MacArthur-Wilson 성장 방정식(MacArthur-Wilson Growth equation)으로 불린다. 우리 목적이 비회귀모형의 종류를 자세하게 소개하는 것은 아니므로, 자세한 것은 Ratkowsky (1983)를 참조하라.

9.6. 비선형회귀분석에서 고려하여야 할 사항들(Further Considerations in Nonlinear Regression)

우리는 이 절에서 비선형회귀분석을 할 때 몇 가지 곤혹스러운 문제에 대하여 논할 것이다. 시작값(starting values), 표준통계추론(standard statistical inference), 선형변환(transformations to linearize)을 다루는데 있어서 골치 아픈 사항들이 있다.

모수의 초기값(Stating Values on Parameters)

비선형회귀 방법론은, 벡터 θ_0 의 값들인 초기추정값(initial estimates)이 있다는 것을 전제로 한다. 만약 θ_0 가 최소제곱추정값(the least squares estimates)인 $\hat{\theta}$ 와 비슷하면, 수렴(convergence)의 어려움은 최소화될 것은 명백하다. 변형되지 않은 가우스, 뉴튼 방법(unmodified Gauss-Newton procedure)의 예에서, 적절한 시작값(good starting value)이 중요하다. 변형된 방법들(즉, reduced increments, Hartley Modification, Marquardt Modification)은 대개 시작값(starting values)에 대하여 덜 민감하다. 그러나 이 변형된 방법에서도 적절한 시작값(starting values)의 필요성은 무시될 수가 없다. 시작값(starting values)이 부적절하면, SS_{Res} 함수의 국소최소(local minimum)로 수렴(convergence)될 수도 있다. 실제로 분석가가 잘못된 값("wrong value")으로 수렴(convergence)된다는 것을 인식하지 못할 수도 있다.

어떤 경우 숙련된 분석가는 시작값(starting values) 역할을 할 수 있는 모수의 근사값(approximate parameter values)를 알 수 있다. 성장 함수의 경우, 모수(parameters)나 모수비(ratios of parameters)를 명확하게 해석할 수 있고, 만약 시스템과 모형에 대하여 많이 알고 있다면 시작값(starting values)을 찾을 수도 있다.

많은 경우, 자료세트에서 합리적인 시작값(starting values)을 결정할 수 있다. 이것의 간단한 예는 식(9.1)의 지수모형(exponential model)이다. 모형에서 결정론적인 부분(deterministic portion)인 $y = \alpha e^{\beta x}$ 을 생각해보자. 다음과 같이 자연로그(nature logs)를 취함으로써 선형화된 형식(linearized form)을 만들 수 있다.

$$\ln y = \ln \alpha + \beta x \quad (9.26)$$

결과적으로, α 의 시작값(starting values)으로 절편의 역대수(antilog of the intercept)를 사용하고, β 의 시작값(starting values)으로 회귀 기울기(the slope of the regression)를 사용하여, x 에 대한 $\ln y$ 의 선형회귀를 쉽게 수행할 수 있다.

다른 모형에서도 비슷한 방법을 사용할 수 있다. 예를 들면, 로지스틱 모형인 식(9.17)의 경우, 자료에서 y_{max} 에 근접한 값을 α 의 시작값(starting values)인 α_0 로 사용할 수 있다.

그러면 x 에 대한 $\ln((\alpha_0/y) - 1)$ 의 단순회귀는 그 회귀의 역대수(antilog)를 β_0 로 사용하고, 기울기는 $-k_0$ 로 하여 분석할 수 있다.

만약 선형회귀의 실행시간이 조금 길 경우, $\ln((\alpha_0/y) - 1)$ 을 x 에 대하여 그림을 간단하게 그려보면, 기울기(slope)와 절편값(intercept)을 어림잡을 수가 있고, 이에 따라 시작값(starting values)을 찾을 수가 있다.

Gomoertz 성장모형(growth model)인 식(9.18)의 예에서, 초기추정값(initial estimates)를 상당히 쉽게 얻을 수 있는데, 그 과정은 식을 약간 재배열(rearranging)하고 나면 명확하게 알 수 있다. y_{\max} 에 근접한 값으로 시작값(initial values) α_0 를 만든다. 모형의 결정론적인 부분(deterministic portion)을 조작하면 다음과 같이 된다.

$$\ln \left[-\ln \frac{y}{\alpha_0} \right] = \ln \beta - kx \ln$$

따라서, x 에 대한 $\ln[-\ln(y/\alpha_0)]$ 의 선형회귀(또는 그림)는, β_0 를 절편의 역대수(the antilog of the intercept)로, $-k_0$ 를 기울기(slope)로 하여 수행할 수 있다. 다른 비선형모형의 경우도 유사한 방법들을 사용하여 시작값(starting values)을 구할 수 있다. 때때로 비선형 구조의 복잡성 때문에 시작값을 찾기가 어렵다. 사용자가 제시하는 범위에서 시작값(starting values)을 계산해주는 격자망 탐색루틴(grid search routines)을 사용하는 컴퓨터 소프트웨어 알고리즘들이 있다(SAS User's Guide: Statistics (1985)를 참조).

표준통계추론(Standard Statistical Inference)

비선형회귀는 정확한 통계적 추론(statistical inferences)을 가능케 하는 고급품(luxury) 즉, 모형모수(model parameters)의 검정(test)과 신뢰구간(confidence interval), 평균반응(mean response)과 예측구간(prediction intervals)에 대한 신뢰구간 등을 가지고 있지 않다고 이미 지적한 바 있다. 유한 표본(finitie samples)에서는 모수 추정값(parameter estimates)의 점근적 분산공분산행렬(asymptotic variance-covariance matrix)을 이용하여 근사값(approximations)을 구할 수는 있다. 예제 9.1과 9.2는 이 행렬의 역할을 설명한다. 많은 분석가들은 비선형회귀분석을 하면서 실망을 하는데, 이는 선형회귀에서 당연하게 가능했던 몇몇

과정들 중 일부가 비선형회귀에서는 이용할 수 없거나 의심스럽기 때문이다.

사용자들은 비선형회귀에서, 표준기법(standard techniques) 중 어떤 것이 이용가능한지 알아야 한다. 우리는 모두 추정값(parameter estimate)의 표준오차(standard errors)를 얻는데 $(X'X)^{-1}$ 행렬을 대신해서 $(W'W)^{-1}$ 행렬을 사용해왔다. 엄밀히 말하면, W 행렬은 우리가 이용할 수 있는 것이긴 하나, 아래 행렬 W 의 추정값(estimate)이며,

$$W = \begin{bmatrix} \frac{\partial f(x_1, \theta)}{\partial \theta_1} & \frac{\partial f(x_1, \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(x_1, \theta)}{\partial \theta_p} \\ \frac{\partial f(x_2, \theta)}{\partial \theta_1} & \frac{\partial f(x_2, \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(x_2, \theta)}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(x_n, \theta)}{\partial \theta_1} & \frac{\partial f(x_n, \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(x_n, \theta)}{\partial \theta_p} \end{bmatrix}$$

도함수들(derivatives)은 참 모수(true parameters)에서 값을 구한다. 비선형모형의 최소제곱추정량(least squares estimators)인 $\hat{\theta}$ 에 관하여 알려져 있는 것을 형식화하기(formalize) 위하여 다음과 같이 쓸 수 있는데, 만약 식(9.8)에서 모형오차(model errors)의 벡터(vector)인 ε 이, 평균 0, 분산공분산행렬(variance-covariance matrix) $\sigma^2 I$ 인 정규분포(normally distributed)를 따른다면,

$\hat{\theta}$ 은 평균이 0이고 분산공분산행렬(variance-covariance matrix)이 $\sigma^2 (W'W)^{-1}$ 이며 근사적으로(approximately) 정규분포한다.

실제로 $\hat{\theta}$ 는 ε 가 정규분포가 아닌 경우에도 점근적으로 정규분포(asymptotically normal) 한다. 이 이론에 관한 상세한 것은 gallant (1987)를 참조하라. 따라서 (9.14)의 $(W'W)^{-1}$ 는 추정된 점근적 분산공분산행렬(estimated asymptotic variance-covariance matrix)이다. 추정값(estimates)은 σ^2 대신에 s^2 을, W 대신에 W 를 사용하여 구할 수 있다. 잔차제곱합(residual sum of squares)에도 이와 유사한 점근적 설명(asymptotic statements)을 할 수 있다. 통계학 학생들은 선형회귀에서 $\frac{SS_{\text{Res}}}{\sigma^2}$ 가 $n - p$ 의 자유도($n - p$ degrees of freedom)로 χ^2 분포하는 것을 안다. 비선형회귀에서,

$SS_{\text{Res}} / \sigma^2$ 는 $n - p$ 자유도로 근사적 χ^2 분포를 하며, 최소제곱추정량(least squares estimator) $\hat{\theta}$ 이다 그리고 SS_{Res} 는 독립(independent)이다.

앞서 말한 정보를 이용하여, 선형회귀에서 사용하는 표준방법론(standard methodology)과 매우 유사한 근사방법(approximate procedure)을 사용할 수 있다. 예를 들어, 다음과 같은 식으로 구한 계수(coefficients)에 대하여 근사적 t 검정(approximate t-test)을 할 수 있다(비록 t 통계량이 표준통계패키지에 자주 보이지는 않지만).

$$t = \frac{\hat{\theta}_j - \theta_{j,0}}{s_{\hat{\theta}_j}} \quad (9.27)$$

여기서 예제 9.1와 같이 $\hat{\theta}_j$ 는 θ_j 의 최소제곱추정량(least squares estimator)이고, $s_{\hat{\theta}_j}$ 는 $\hat{\theta}_j$ 의 표준오차(standard error)이며, $\theta_{j,0}$ 는 다음과 같은 가설 하에서 θ_j 의 값이다

$$\begin{aligned} H_0 : \theta_j &= \theta_{j,0} \\ H_1 : \theta_j &\neq \theta_{j,0} \end{aligned}$$

식(9.27)에 있는 통계량은 H_0 하에서 근사적으로 t_{n-p-1} 이다.

확실히, 한쪽 또는 양쪽꼬리 가설(one- or two-tailed hypotheses)이 수용될 수 있다.

$\hat{\theta}$ 의 비선형 함수; 신뢰 한계와 예측값 한계(Nonlinear Functions in $\hat{\theta}$; Confidences Limits and Prediction Limits)

대개 비선형회귀 사용자들은, $\hat{\theta}$ 요소(elements of $\hat{\theta}$)에 대하여 비선형인, 모수 추정값 함수(functions of parameter estimates)에 더 관심이 많다. 확실하게 적용할 수 있는 것들은 평균반응(mean response)에 대한 신뢰구간(confidence intervals)과 새로운 관찰 반응(new observed response)에 대한 예측한계(prediction limits)이다. 선형회귀를 위한 이러한 방법들은 2장과 3장에서 자세히 설명하였다. 실제로 이런 종류의 계산에서 W 행렬의 역할은 우리가

예측한 대로이다. 실제로 계산과정은 선형회귀와 모든 점에서 매우 유사하다. 어떤 비선형 함수 $g(\theta)$ 의 개념(notion)을 생각해보자. 자, 추정값 $\hat{g}(\hat{\theta})$ 에 대하여,

$\hat{g}(\hat{\theta})$ 는 근사적으로 정규분포 하며(approximately normal), 근사평균(approximate mean) $O/g(\hat{\theta})$, 근사분산(approximate variance) $O/\sigma^2 u'(W'W)^{-1} u$ 이다. 여기에서 벡터 $u' = [\partial g(\theta)/\partial \theta_1, \partial g(\theta)/\partial \theta_2, \dots, \partial g(\theta)/\partial \theta_p]$ 이다.

독자들은 평균반응(mean response)의 신뢰한계(confidence limits)에 이 결과를 적용하는 것을 마음에 그려 보아야 한다. 여기서 회귀변수의 임의의 위치 x_0 에서 예측하려고 한다면, 도함수의 벡터(vector of derivatives)인 u' 는 다음과 같다.

$$w'_0 = \left[\frac{\partial f(\theta, x_0)}{\partial \theta_1}, \frac{\partial f(\theta, x_0)}{\partial \theta_2}, \dots, \frac{\partial f(\theta, x_0)}{\partial \theta_p} \right]$$

사용자는 w_0 의 모수(parameter)를 추정값(estimate)으로 대치하여야 한다. 따라서, 예측된 반응(predicted response) $f(x_0, \hat{\theta})$ 의 근사표준오차(approximate standard error)에 대한 추정값(estimate)을 구할 수 있다.

$$S_{f(x_0, \theta)} = S w'_0 (W'W)^{-1} w_0$$

여기서 w_0 는 모수를 추정값(estimate)으로 대치하고, x_0 위치에서 구한 도함수(derivatives)의 벡터이다. 물론 W는 모수 추정값(estimate)을 포함한다. 다중선형회귀(multiple linear regression)의 예측값(predicted value)에 대한 표준오차(standard error)와 유사한 구조임을 눈여겨 보라. 따라서 임의의 위치 x_0 에서 평균반응(mean response)의 근사 $100(1-\alpha)\%$ 신뢰구간(approximate $100(1-\alpha)\%$ confidence intervals)은 다음과 같다.

$$f(x_0, \hat{\theta}) \pm t_{\alpha/2 \cdot n-p-1} s \sqrt{w'_0 (WW)^{-1} w_0}$$

물론 x_0 에서 새로운 관찰값(observation)에 대한 근사 $100(1-\alpha)\%$ 예측한계(approximate 100(1- α)% prediction limits)는 다음과 같다.

$$f(x_0, \hat{\theta}) \pm t_{\alpha/2 \cdot n-p-1} s \sqrt{1 + w'_0 (WW)^{-1} w_0}$$

분석가가 예측을 하기 위하여 비선형회귀함수(nonlinear regression function)를 사용할 때, 신뢰한계(confidence limit)와 예측한계(prediction limit)가 중요하다. 이들은 적합된 비선형모형(fitted nonlinear models)끼리 성능(performance)을 비교하는데 사용될 수도 있다. 만약 예측이 중요하다면, 신뢰한계(confidence limit)가 좁게(tighter) 나오는 모형을 고려해야만 한다. 물론 경쟁 모형(competing models)을 비교할 때는 오차평균제곱(error mean square)인 s^2 를 비교하여야 한다.

잔차(residuals)의 일반적 성질(general nature) 또한 여러 개의 경쟁 비선형 모형들 중에서 가장 적합한 모형을 결정할 때 중요한 정보를 줄 수 있다. 비록 선형모형에 사용되는 상용컴퓨터패키지로는 쉽게 접근하기 어려우나, 6장에서 설명된 것들과 유사한 진단방법(diagnostics)을 비선형모형에서도 만들어 낼 수 있다. 가우스, 뉴튼방법(Gauss-Newton procedures) 혹은 그 변형에서, 모수 추정값에 대한 해(solutions)에서 구해진 WW 행렬은 $X'X$ 의 역할을 한다. 따라서 HAT 대각값(HAT-diagonal values), 스튜던트화 잔차(studentized residuals), 그리고 이들로부터 유래된 모든 진단법들(diagnostics)을 계산할 수 있다.

9.7. 자료를 변환하여 선형화하지 않는 이유는?(Why Not Transform Data to Linearize?)

자료분석가들은 추정과정(estimation procedure)을 단순화하기 위하여, 원래 비선형인 모형에 변환(transformation)을 함으로써 선형성(linearity)을 유도한다. 이것은 합리적인 접근방법으로 실제로 예제 9.1과 9.2에서 시작값(starting values)을 정하는 방법에서 이미 언급하였다. 9.6절에서는 먼저 선형화(linearizing)를 하고 나서 시작값(starting values)을 구하는 방법을 알아 보았다. 그러나 다음과 같은 질문을 할 수 있는데: “왜 선형화를 하지 않으며, 왜 선형모형으로 모수를 추정하지 않을까? 이 질문에 답하려면 꼭 알아야 할 개념들을 흔히 잘 모르고 있다. 전형적인 예가 식(9.1)의 지수모형형태(exponential model form)이다. 식(9.26)에서 했던 것처럼, 자연로그(natural logarithms)를 취함으로써 모형의 결정론적인 부분(deterministic part)을 선형화한다. 다음의 대안 모형(alternative model)을 생각해보자.

$$\ln y = \ln \alpha + \beta x_i + \varepsilon_i \quad (9.28)$$

식(9.28)은 비선형 접근방법의 복잡성을 확실하게 제거하며, x 에 대하여 $\ln y$ 를 직접적으로 회귀시킴으로써 단순선형회귀(simple linear regression)를 적용할 수 있다. 예가 이것만 있는 것은 아니다. 변환(transformation)을 이용하여 모형을 선형화 혹은 단순화하는 예는 많이 있다. 다른 예로 Michaelis-Menten 모형으로 다음과 같다.

$$y_i = \frac{V_{\max}}{k + x_i} + \varepsilon_i \quad (9.29)$$

상기식에 $1/y$ 을 적용하여 추정(estimation)을 단순화할 수 있다. 따라서 우리는 다음과 같은 형태의 모형을 가정할 수도 있다.

$$\frac{1}{y_i} = \frac{k}{V_{\max}} + \left(\frac{1}{V_{\max}} \right) x_i + \varepsilon_i \quad (9.30)$$

그러면 우리는 x 에 대하여 $1/y$ 를 회귀시킴으로써 단순선형회귀를 수행할 수 있고, 이에 따라서 모수 V_{\max} 와 k 를 추정할 수 있다.

(9.28) 혹은 (9.30)의 선형화된 형태(linearized form)에 최소제곱법(least squares procedure)을 적용하면 상응하는 비선형 유사형태와 동일한 모수 추정값(estimate)을 구할 수 없다. 비선형 모형인 (9.1) 또는 (9.29)에서, 최소제곱은 y 에 대한 잔차제곱합(sum of squares of residuals)의 최소화(minimization)를 의미한다. 그러나 모형 형식(9.28)과 (9.30)에서는 y 의 변환(transform) 즉, (9.28)에서는 $\ln y$, (9.30)에서는 $1/y$ 에 대한 잔차 제곱합을 최소화시켰다. 어느 것이 맞는 것인가? 선형화가 전적으로 틀린 것일까? 해답은 모형가설(model assumption)이 옳은가에 달려 있다. 두 가지 독립적 모형과 모형철학이 있다. 식(9.1)의 모형은 오차구조(error structure)가 가법적(additive)이다. 만약 이 가정이 합리적이라면, 단순히 로그를 취한다고 식(9.28)의 모형이 되지는 않는다. 따라서, (9.28)은 지수모형(exponential model)에 로그를 취하여 승법오차구조(multiplicative error structure)를 가지게 되므로, (9.1)과 (9.28)는 다르다. (9.29)와 (9.30)에서도 이와 유사한 구별점(distinction)을 볼 수 있다.

우리는 모형오차(model errors)인 ε_i 에 대하여, 정규분포하고(normal), 독립적이며(independent), 평균 0, 공통 분산(common variance) σ^2 을 가진다는 가정을 자주 한다. 만약 이러한 오차가정(error assumptions)을 (9.29)에서 한다면, 전혀 다른 오차구조, 따라서 전혀 다른 오차가정이 필요한 (9.30)으로 변환하는 것은 별로 말이 되지 않는다. 따라서 “잘 맞는 모형오차(well-behaved model errors)가 원래의 비선형 모형에 수반된다고 가정하여야 할까 혹은 변환된 모형에 수반된다고 가정하여야 할까”로 쟁점(issue)이 넘어가게 된다.

변환된 모형의 결과로 얻어진 모수의 성질에서 또 다른 어려움이 발생한다. 선형모형으로 변환한 거의 모든 경우에서, 모수 추정량(estimate)의 성질 때문에 문제점이 발생한다. 그 예를 (9.29)와 (9.30)에서 볼 수 있다. (9.30)으로 변환하면 선형회귀를 시행할 수 있다. 그러나 (9.30)에서 모형 오차에 대한 일반적인 가정이 맞다 하더라도, k 와 V_{\max} 는 아니지만, $1/V_{\max}$

와 k/V_{\max} 의 모든 선형 불편추정량(linear unbiased estimators)의 최소분산(minimum variance)은 부적절한 특성(displeasing property)을 보인다. 따라서 추정량(estimate)의 성질에 관한 한, 선형화를 함으로써 얻을 수 있는 이점은 흔히 명확하지 않다.

앞에서 언급한 것과 같이, 중요한 논쟁거리가 여전히 남아 있다. 비선형 모형을 선형화(linearization)할 경우 모형이 동등하지 않다. 실은 오차구조가 다른 두 개의 모형이 생기는 것이다. 선형변환을 할 경우, 그에 따른 결과를 알아야 한다: (1) 불합리한 오차구조(unreasonable error structure), (2) 모수 추정량의 부적절한 성질. 관련된 토론은 7장 3절에 계속된다.

오차 항(error term)에 두 가지 상반되는 가정을 하여도 타당하게 안정적인 모형도 존재한다. 즉, 비선형 추정량(estimate)이, 변환하고 나서 구한 추정량들과 유사한 특성을 보이는 경우이다. 예를 들어, 로지스틱은 이런 안정된 특징을 보인다고 알려져 있다. 경쟁

모형 간의 비교를 알고 싶으면, Ratkowsky (1983)를 참조하라.

10. 부록 A: 행렬대수의 몇 가지 특별한 개념들(Some Special Concepts in Matrix Algebra)

이 부록에서는 행렬대수(matrix algebra)의 몇 가지 기본적인 개념을 논하고자 한다. 광범위하게 사용되고 있는 정의(definition)와 정리(theorem)가 제시될 것이다. 선형 방정식의 해를 구하는데 필요한 행렬 조작(matrix manipulation)부터 시작할 것이다.

10.1. (A.1) 동시선형방정식의 해(Solutions to Simultaneous Linear Equations)

3.2절에서 다음과 같은 형태의 일반선형모형(general linear model)을 소개하였다.

$$y = X\beta + \varepsilon$$

식(3.4)에서 주어지는 최소제곱 정규방정식(least squares normal equation)의 해를 구하려면 연립 선형방정식(simultaneous linear equations)의 계(systems)를 풀어야 한다. 다음과 같은 선형방정식 세트를 가정해보자.

$$\begin{aligned} a_{11}b_1 + a_{12}b_2 + \cdots + a_{1p}b_p &= g_1 \\ a_{21}b_1 + a_{22}b_2 + \cdots + a_{2p}b_p &= g_2 \\ \vdots &\quad \vdots \quad \vdots \quad \vdots \\ a_{p1}b_1 + a_{p2}b_2 + \cdots + a_{pp}b_p &= g_p \end{aligned}$$

우리는 위의 방정식 계(system of equations)를 다음과 같이 다시 쓸 수 있다.

$$\boxed{Ab = g} \quad (\text{A.1})$$

여기서

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \quad g = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_p \end{bmatrix}$$

행렬 A 가 비특이행렬(nonsingular matrix)일 때,² 즉 $p \times p$ 행렬 A^{-1} 이 존재할 때, 연립방정식세트의 유일 해 b 가 존재한다.

$$AA^{-1} = A^{-1}A = I_p$$

I 는 크기가 p 인 단위행렬(identity matrix)이다.³ 따라서 식(A.1)의 해는 양변을 A^{-1} 로 premultiply(행렬 앞에 곱함, 행렬 뒤에 곱하는 것은 postmultiply)하여 구할 수 있다. 즉,

$$A^{-1}Ab = A^{-1}g$$

$$Ib = A^{-1}g$$

$$b = A^{-1}g$$

행렬 A^{-1} 을 행렬 A 의 역행렬(inverse matrix)이라 하며, $AA^{-1} = A^{-1}A = I_p$ 이다. 이런 성질을 만족하는 A^{-1} 은 임의의 정사각행렬 A 에 대해 존재하지 않거나, 단 하나만 존재한다. 이런 행렬을 역행렬이라 하고 A^{-1} 로 표기한다. 역행렬이 존재하지 않는 행렬을 특이행렬(singular matrix) 혹은 비가역행렬(degenerate matrix)이라고 한다. 3장의 회귀분석 적용(regression application)에서, 해는 식(3.4) b 의 최소제곱추정량(least squares estimators)에 의하여 주어진다.

$$b = (X'X)^{-1}X'y$$

행렬 $X'X$ 는 식(A.1)의 행렬 A 와 같은 역할을 한다. $X'X$ 행렬은 제곱(squares)의 합과 교차곱(cross products)의 합을 포함하고 있는 대칭행렬(symmetric matrix)이다.⁴

$X'X$ 의 대각원소(diagonal elements)는 제곱(squares)의 합이며, 비대각원소(off-diagonal elements)는 교차곱(cross products)의 합이다. 예를 들어,

² 선형대수학에서 가역행렬(invertible matrix), 또는 비특이행렬(non-singular matrix)은 역행렬(inverse matrix)을 갖는 $p \times p$ 행렬을 가리킨다.

³ 선형대수학에서 크기 p 인 단위행렬(單位行列)은 주대각선(원쪽 위에서 오른쪽 아래로 가는 대각선)이 전부 1이고 나머지 원소는 0을 값으로 갖는 $p \times p$ 정사각행렬이다. 크기가 p 인 단위행렬은 보통 I_p 으로 표기하지만, 그 크기가 문맥상 자명하게 유추 가능한 경우 생략하여 I 로 쓰기도 한다. I_p 의 가장 중요한 성질로는 다음의 것이 있다. $AI_p = A$ 이고 $I_pB = B$ 이다.

⁴ 선형대수학에서 대칭행렬이란 어떤 행렬(A)과 그것의 전치행렬(A^T)이 같은 행렬을 말한다. 즉, $A = A^T$. 이것은 A 가 정방행렬(square matrix, 행과 열의 수가 같은 행렬)이라는 것을 뜻한다. 대칭행렬은 주대각선을 중심으로 대칭이라는 의미이다. 즉, $A = (a_{ij})$ 라고 하면 $a_{ij} = a_{ji}$ 이다. 전치행렬(轉置行列)이란 열을 행으로, 행을 열로 바꾼 것이다. 정사각 행렬의 경우에는 행렬의 원쪽 위에서 오른쪽 아래를 가르는 주대각선을 기준으로 대칭되는 원소끼리 바꿔치기하여 전치행렬을 얻을 수 있다. A 행렬의 전치행렬은 $A^T, {}^t A, A', A^T$ 등으로 나타낸다. $m \times n$ 행렬 A 의 전치행렬은 $n \times m$ 행렬이 된다. $1 \leq i \leq n, 1 \leq j \leq m$ 일 때 A^T 는 $A^T[i, j] = A[j, i]$ 라고 정의할 수 있다.

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ 1 & x_{13} & x_{23} & x_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{bmatrix}$$

이라면

$$XX' = \begin{bmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{3i} \\ \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{1i}x_{3i} \\ \sum_{i=1}^n x_{2i}^2 & \sum_{i=1}^n x_{2i}x_{3i} & \sum_{i=1}^n x_{3i}^2 \\ Symmetric & & \end{bmatrix}$$

기초 행렬대수에 친숙하지 않은 독자들을 위하여, A.3절은 $X'X$ 의 본질과 $X'X$ 의 특성들 때문에 심각한 공선성(collinearity) 문제가 어떻게 발생하는지에 대하여 좀 더 깊이 알 수 있도록 해줄 것이다.

10.2. (A.2) 이차형식(The Quadratic Form)

이차형식(quadratic form)은 회귀(regression)에서 중요한 행렬조작 중의 하나이다. 개념은 매우 단순하다. 행벡터(column vector) z (n 요소)와 전형원소(典型元素, typical element) a_{ij} 를 가진 $n \times n$ 행렬 A 를 생각해보자. 그러면 다음의 스칼라 양(scalar quantity)

$$z' A z = \sum_{i=1}^n a_{ii} z_i^2 + 2 \sum_{i=1}^n \sum_{\substack{j=1 \\ i < j}}^n a_{ij} z_i z_j$$

은 행렬 A 를 가진 z 의 이차형식이라고 한다.

양의 정부호 이차형식(positive definite quadratic form)이란 0이 아닌 모든 z 에 대하여 $z' A z > 0$ 인 경우를 말한다. 양의 준정부호 행렬(positive semi-definite matrix)이란 모든 z 에 대하여 $z' A z \geq 0$ 이나, 0이 아닌 몇몇 z 에 대하여 $z' A z = 0$ 인 경우를 말한다.

본문에서 이차형식을 많이 적용하였다. 3.4절에서는 회귀계수의 부분세트(subset)에 관한 가설을 검증할 경우 사용되는 “extra sum of squares” 원리에 중점을 두었다. 식(3.13)은 두 번째 부분세트(subset) $X_2 \beta_2$ 가 있는 상태에서 $X_1 \beta_1$ 항의 부분세트(subset)로 설명되는 회귀제곱합(regression sum of squares)을 표현하고 있다. 설명된 회귀는 벡터 y 와 행렬 $X(X'X)^{-1} X' - X_2(X_2'X_2)^{-1} X_2'$ 의 이차형식에 의하여 주어진다.

식(3.13)의 이차형식의 생성은 완전모형(full model)에 대한 회귀 제곱합(regression sum of squares)의 관점에서 볼 수 있다. 다음의 일반선형모형(general linear model)을 생각해보자.

$$y = X\beta + \varepsilon$$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

그리고 $n > k + 1$.

회귀 제곱합(regression sum of squares) $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 은 3장에서 토의된다. 이러한 양(quantity)은 k 자유도 제곱합(degrees of freedom sum of squares)인데, 이것은 회귀변수(regressors)들 x_1, x_2, \dots, x_k 에 의해 설명되는 변동(variation)에 대하여 설명한다. 회귀

변수와 상수항(constant term)에 의하여 설명되는 $k+1$ 회귀 제곱합은 다음과 같이 주어진다.

$$\sum_{i=1}^n \hat{y}_i^2 = (Xb)'(Xb)$$

즉,

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

따라서, X_2 에 회귀변수들(regressors)이 있는 상태에서 X_1 의 회귀변수들(regressors)로 설명되는 회귀(regression)을 가진 식(3.13)은 완전모형(full model, $y = X\beta + \varepsilon$)과 축소모형(X_2 에만 회귀변수들이 있음) 간의 회귀 제곱합(regression sum of squares) 차이에 의하여 주어진다.

이와 같이 $R(\beta_1|\beta_2)$ 는 두 개의 상응하는 이차 형태(quadratic form) 간의 차이다: 즉,

$$\begin{aligned}R(\beta_1|\beta_2) &= y'X(X'X)^{-1}X'y - y'X_2(X_2'X_2)^{-1}X_2'y \\ &= y'[X(X'X)^{-1}X' - X_2(X_2'X_2)^{-1}X_2']y\end{aligned}$$

이차 형태(quadratic form)를 적용한 또 다른 중요한 예는, 본문의 여러 군데에서 나타나는 표기법(notation)인, 예측값 분산(prediction variance)의 계산에서 볼 수 있다. 첫 번째는 3.5절에서 볼 수 있다. 식(3.20)은 위치(location) x_0 에서 y 의 적합값(fitted value) 혹은 예측값(predicted value)의 분산에 대한 것이다. σ^2 예측값에 대한 분산(variance of this prediction)은 행렬 $(X'X)^{-1}$ 을 가진 벡터 x_0 에서 이차형태(quadratic form)인 $\sigma^2 \cdot x_0'(X'X)^{-1}x_0$ 으로 주어진다. 이러한 특이한 이차 형태를 잘 보면 다음과 같음을 알 수 있다.

$$\frac{Var(\hat{y}|x_0)}{\sigma^2} = Var b_0 + \sum_{j=1}^k x_{j,0}^2 Var b_j + 2 \sum_{j=0}^k \sum_{l=0}^k x_{j,0} x_{l,0} Cov(b_j, b_l)$$

여기서 $x_{0,0}$ 는 1 (unity)이다. 위와 같은, 예측값 분산(prediction variance)의 확장된 표기법은 $(XX)^{-1}$ 을 통하여, 모든 계수들(coefficients)의 분산(variances)과 공분산(covariances)을 고려하고 있음을 쉽게 알 수 있다.

HAT diagonal은 이차 형태(quadratic form)를 사용한 명백한 예이면서 예측값 분산(prediction variance)의 다소 특별한 경우인데, 3.9절에서 소개되어 본문 전체에서 다루고 있다. 만약 문제가 되는 위치(location) x_0 가 우연히 자료 포인트(data point)의 하나라면, x'_0 는 X 행렬(matrix)의 열(row)인 x'_i 가 된다. 그러면 이차 형태(quadratic form) $x'_i(XX)^{-1}x_i$ 는

식(3.39)에서 소개되는 HAT 행렬(matrix) $X(X'X)^{-1}X'$ 의 대각(diagonal)이다.

10.3. (A.3) 고유값과 고유벡터(Eigenvalues and Eigenvectors)

고유값(eigenvalues)의 개념은 3장과 8장의 다중공선성(multicollinearity)에 대한 논의에서 두드러진다. 고유값벡터(eigenvalue-vector) 개념은 3.8절에서 처음으로 소개되었다. 여기서 우리는 정의를 내리고, 고유값(eigenvalues)이 다중공선성(multicollinearity)의 정도(extent)를 진단하고 정량화하는데 있어서 왜 그렇게 중요한 역할을 하는지에 대하여 얼마간의 통찰력을 제공한다.

$k \times k$ 대칭행렬(symmetric matrix) A 를 고려해 보자. 행렬 A 의 고유값(eigenvalues) $\lambda_1, \lambda_2, \dots, \lambda_k$ 은 행렬식(determinantal equation)에 대한 해(solutions)로 주어진다.

$$|A - \lambda I| = 0 \quad (\text{A.2})$$

다음 식의 해 v_i 에 의하여 정의되는 고유벡터(eigenvector)는 i 번째 고유값(eigenvalues)인 λ_i 와 연관된다.

$$(A - \lambda_i I)v_i = 0$$

A 가 대칭일 때, 고유값(eigenvalues)은 모두 실수(real)가 될 것이다.

통계 이론에서 고유값(eigenvalues)과 고유벡터(eigenvectors)를 적용한 예들은 많지만, 우리는 공선성(collinearity)에 집중할 것이다. 3.8절에서 지적한 것처럼 행들(columns)이, 연관된 정규화 고유벡터(the associate normalized eigenvectors)들로 이루어진 행렬 V 는 A 를 대각화(diagonalize)하는데 사용될 수 있다. 즉,

$$V'AV = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k) \quad (\text{A.3})$$

추가적으로, V 는 직교행렬(orthogonal matrix)이다; 즉, $V'V = VV' = I_p$ 이다. 현재, 우리의 관심은 3장에서 기술된, $A = (X^* X^*)$, 상관행렬(correlation matrix)의 경우이다. 만약 상관행렬(correlation matrix)이 대각(diagonal)이라면, 즉 회귀변수들(regressor variables)사이에 선형적인 연관성(linear association)이 전혀 없다면, 모든 고유값(eigenvalues)들은 1 (unity)일 것이다. 이것은 이상적인 경우이다. 만약 상관행렬(correlation matrix)이 거의 특이(near-

singular)하다면 즉, X^* 의 행들(columns) 중에 적어도 하나의 근접 종속성(near dependency)이 있다면, $X^{*\top}X^*$ 의 행렬식(determinant)은 거의 0에 수렴할 것이다. 고유값(eigenvalues)은 행렬식(determinant)과 중요한 관계를 가지고 있다; $A = (X^{*\top}X^*)$ 와 식(A.3)을 고려해 보라.

V 는 직교(orthogonal)이기 때문에, $|V| = 1.0$, $|V'| = 1.0$, 그리고

$$\begin{aligned} |X^{*\top}X^*| &= |V| \prod_{i=1}^k \lambda_i |V'| \\ &= \prod_{i=1}^k \lambda_i \end{aligned} \quad (\text{A.4})$$

다른 중요한 결과는 상관행렬(correlation matrix)의 고유값(eigenvalues) 합(sum)을 행렬의 차원(dimension)과 관계 지운다. 행렬대수에서 잘 알려진 결과는, A와 B가 곱(multiplication) AB 및 BA와 일치하는 행렬일 경우, $\text{tr}(AB) = \text{tr}(BA)$ 이다(Graybill (1976) 참조). 따라서,

$$\text{tr}(V'AV) = \text{tr}(VV'A)$$

자, $VV' = I$ 므로

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

그러나, 식(A.3)으로부터

$$\text{tr}(A) = \sum_{i=1}^k \lambda_i$$

그러므로 A가 상관행렬(correlation matrix)이라면

$$\sum_{i=1}^k \lambda_i = k \quad (\text{A.5})$$

여기서 k 는 상관행렬의 차원(dimension)이다.

위에서 말한 것으로부터 만약 상관행렬이 다중공선성(multicollinearity) 때문에 거의 특이적(near-singular)이 될 경우, 최소한 하나의 고유값(eigenvalue)은 0에 근접한다는 것이 분명하다. 이는 공선성(collinearity)이 없으면 모든 고유값이 1.0이 되는 이상적인 경우와 비교된다.

상관행렬에서 0에 가까운 고유값들(near zero eigenvalues)이, 회귀상황(regression

situation)에서 심각한 문제를 어떻게 발생시키는지 밝혀주는 또 다른 수식전개(development)가 있다. 3.8절의 식(3.35)와 (3.36)을 생각해보자. 식(3.36)에서 근접 0 고유값(near zero eigenvalue)이 하나라도 있으면, 회귀계수(regression coefficients) 분산들(variances)의 합(sum)이 바람직하지 않게 된다는 것은 명백하다. 식(3.36)은, A가 다시 상관행렬(correlation matrix) $X^* X^*$ 인 식(A.3)을 살펴보면 쉽게 증명된다.

$$X^* X^* = V \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots \\ 0 & & \lambda_k \end{bmatrix} V'$$

직교행렬(orthogonal matrix) V 에서, V 의 역행렬(inverse)은 V 의 전치행렬(transpose)과 동등(equal)하다. 다음을 구하기 위하여 양변에 역행렬을 취한다.

$$(X^* X^*)^{-1} = V \begin{bmatrix} 1/\lambda_1 & & 0 \\ & 1/\lambda_2 & \\ & & \ddots \\ 0 & & 1/\lambda_k \end{bmatrix} V'$$

회귀계수 분산의 합(sum of the variances of regression coefficients)은 $(X^* X^*)^{-1}$ 의 대각합(trace)이다(σ^2 은 별도로 하고).

$$\begin{aligned} tr(X^* X^*)^{-1} &= tr V' V \begin{bmatrix} 1/\lambda_1 & & 0 \\ & 1/\lambda_2 & \\ & & \ddots \\ 0 & & 1/\lambda_k \end{bmatrix} \\ &= tr \begin{bmatrix} 1/\lambda_1 & & 0 \\ & 1/\lambda_2 & \\ & & \ddots \\ 0 & & 1/\lambda_k \end{bmatrix} \\ &= \sum_{i=1}^k \frac{1}{\lambda_i} \end{aligned}$$

이것이 식(3.36)의 결과이다.

10.4. (A.4) 분할 행렬의 역(The Inverse of a Partitioned Matrix)

4.1절에서, 저설정된(underspecified) 모형에서 오차평균제곱(error mean square)의 편향(bias)에 대하여 언급하였다. 4장의 많은 것들은, 저설정된 모형 설정(underspecified modeling setting)이 다음과 같은 모형 가정에 의하여 구성될 수 있다는 것을 전제로 하고 전개된 것이다.

$$y = X_1\beta_1 + \varepsilon^* \quad (\text{A.6})$$

반면 참 모형(true model)은

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (\text{A.7})$$

인데 m 개의 모수가 있으며, 이때 $m > p$ 이다. 여기서,

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_2 \end{bmatrix}$$

저설정된(underspecified) 모형의 오차평균제곱(error mean square)에 대한 기대값(expected value)은 식(4.1)로 주어진다. 본문에서 지적한 것처럼, 이 편향(bias)은 다음 식으로 주어지는데,

$$\frac{1}{n-p} \beta_2' [X_2' X_2 - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2] \beta_2$$

이것은 무시된 계수들(ignored coefficients) 즉, β_2 의 계수들의 표준화된 형태(standardized form)이다. 이것의 의미는, b_2 (만약 식(A.7)의 완전한 모형을 적합했다면)의 분산공분산행렬(variance-covariance matrix)이 다음과 같이 주어진다는 것이다.

$$Var(b_2) = \sigma^2 (X_2' X_2 - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2)^{-1}$$

이것은 분할행렬(partitioned matrix)의 역(inverse)을 살펴보면 쉽게 입증할 수 있다.

$$X'X = \begin{bmatrix} X'_1 X_1 & X'_1 X_2 \\ X'_2 X_1 & X'_2 X_2 \end{bmatrix} \quad (\text{A.8})$$

자, $X'_1 X_1$ 은 $p \times p$ 의 정방대칭행렬(square symmetric matrix)이고, $X'_2 X_2$ 는 $(m-p) \times (m-p)$ 차원(dimension)의 대칭행렬이다. 3장에서 우리는 b 의 분산공분산행렬(variance-covariance matrix)이 다음과 같이 주어진다는 것을 기억할 수 있다.

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

따라서 σ^2 은 별도로 하고, b 의 분산공분산행렬(variance-covariance matrix)은 (A.8) 역행렬의 상단 좌측 모서리(upper left-hand corner)에 있는 $(p \times p)$ 정방행렬(square matrix)이고, $\text{Var}(b_2)$ 는 같은 역행렬의 우측하단(bottom right-hand corner)에 있는 정방대칭행렬(square symmetric matrix)이다. Graybill (1976)에 XX' 의 역행렬의 분할들(partitions)이 소개되어 있다. 즉, 다음과 같이 주어지는데,

$$(X'X)^{-1} = \begin{bmatrix} (C_{11})^{-1} & -(X'_1 X_1)^{-1} X'_1 X_2 C_{22}^{-1} \\ -C_{22}^{-1} X'_2 X_1 (X'_1 X_1)^{-1} & (C_{22})^{-1} \end{bmatrix} \quad (\text{A.9})$$

여기서

$$\begin{aligned} C_{11} &= (X'_1 X_1 - X'_1 X_2 (X'_2 X_2)^{-1} X'_2 X_1) \\ C_{22} &= (X'_2 X_2 - X'_2 X_1 (X'_1 X_1)^{-1} X'_1 X_2) \end{aligned} \quad (\text{A.10})$$

$(X'X)(XX)^{-1} = I$ 임을 증명해보라.

10.5. (A.5) Sheerman-Morrison-Woodbury 정리(Sheerman-Morrison-Woodbury Theorem)

이 절의 결과는, 4장의 PRESS 통계량뿐만 아니라 6장의 현대적 단일자료포인트 진단법(modern *single data point* diagnostics)의 기초를 이룬다. 이것은 본질적으로, i 번째 자료 포인트가 제거되거나(removed) *제외*(set aside) 경우, 중요한 회귀통계량(regression statistics)을 계산하기 쉽게 해준다. 여기서 매우 일반적인 형식(general form)의 기본 결과(fundamental result)를 얻을 것이다. 부록 B에서, 이 결과는 어떤 중요한 진단도구(diagnostic tools)의 전개를 설명하는데 사용된다.

$p \times p$ 이며, 행벡터 z 가 p 차원(p-dimensional column vector)인 정방정칙행렬(square nonsingular matrix) A 를 살펴보자. 여기서 A 는 $X'X$ 행렬이다. 벡터 z' 는 X 행렬의 i 번째 열이다. 이와 같이 $(A - zz')$ 는 i 번째 자료 포인트(data point)가 포함되지 않은 $X'X$ 행렬이 된다. 정리(theorem)는 다음과 같다(Rao (1973) 참조).

$$(A - zz')^{-1} = A^{-1} + \frac{A^{-1}zz'A^{-1}}{1 - z'A^{-1}z} \quad (\text{A.11})$$

우변에 $A - zz'$ 를 곱하면 단위행렬(identity matrix)이 된다는 것만 보여주면 이 결과는 증명될 수 있다.

$$\begin{aligned} & \left[A^{-1} + \frac{A^{-1}zz'A^{-1}}{1 - z'A^{-1}z} \right] [A - zz'] \\ &= I + \frac{A^{-1}zz'}{1 - z'A^{-1}z} - A^{-1}zz' - \frac{A^{-1}zz'}{1 - z'A^{-1}z} A^{-1}zz' \\ &= I + \frac{A^{-1}zz' - A^{-1}zz'(1 - z'A^{-1}z) - A^{-1}z(z'A^{-1}z)z'}{1 - z'A^{-1}z} \\ &= I + \frac{A^{-1}zz' - A^{-1}zz' + A^{-1}zz'(z'A^{-1}z) - A^{-1}zz'(z'A^{-1}z)}{1 - z'A^{-1}z} \\ &= I \end{aligned}$$

11. 부록 B. 몇 가지 특별한 조작법(Some Special Manipulations)

여기서는, 본문에 나오는 몇 가지 중요한 통계학적 결과와 개념에 관한 수식 전개(development)를 자세하게 살펴 보도록 하겠다. 행렬대수(matrix algebra)에 숙달된 독자들은 잘 따라올 수 있을 것이다.

11.1. (B.1) 잔차평균제곱의 비편향성(Unbiasedness of the Residual Mean Square)

여기서 우리는 모수(parameter) σ^2 에 대한 추정량(estimator) s^2 의 비편향성(unbiasedness)을 증명한다. 3.2절에서 식으로 나타낸 잔차평균제곱(residual mean square)으로 시작한다.

$$s^2 = \frac{(y - Xb)'(y - Xb)}{n - p} \quad (\text{B.1})$$

최소제곱정규방정식(least squares normal equations) $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$ 을 이용하여,

$$s^2 = \frac{\mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y}}{n - p}$$

위 방정식의 분자(numerator)는 총제곱합(total sum of squares), $\mathbf{y}'\mathbf{y}$ 와 회귀제곱합(regression sum of squares) 간의 차이로 볼 수 있다.

$$\mathbf{b}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

따라서 식(B.1)은 아래의 식이 된다.

$$s^2 = \frac{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{n - p}$$

분자(numerator)의 잔차제곱합(residual sum of squares)은 다음과 같이 \mathbf{y} 에 대한 이차형태(quadratic form)로 만들 수 있다.

$$s^2 = \frac{\mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}}{n - p} \quad (\text{B.2})$$

이러한 관점에서, 정리(theorem, Graybill (1976) 참조)를 이차형태(quadratic form)의 기대값(expected value)에 적용한다. 평균(mean) $E(y) = \mu$, 분산공분산행렬(variance-covariance matrix) $Var(y) = \sigma^2 I$ 인 확률벡터(random vector) y 가 주어진다면,

$$E(y' A y) = \sigma^2 \text{tr}(A) + \mu' A \mu \quad (\text{B.3})$$

이 정리를 식(B.2)의 이차형태의 기대값을 알아내기 위하여 사용할 수 있다.

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

여기서 $\text{tr } X(X'X)^{-1}X' = \text{tr}(X'X)(X'X)^{-1} = p$ 라는 것을 이용하였다. 이제 아래와 같이 쓸 수 있는데,

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

그러므로 s^2 는 σ^2 에 대한 불편추정량(unbiased estimator)이다.

11.2. (B.2) 저설정된 모형에서 잔차제곱합과 평균제곱의 기대값(Expected Value of Residual Sum of Squares and Mean Square for an Underspecified Model)

3.6절과 4.1절에서, 적합된 모형이 저설정되었을 때 s^2 에 발생하는 편향(bias)에 초점을 맞추었다. 즉, 다음과 같은 불충분한 모형(short model)으로 적합하고,

$$y = X_1\beta_1 + \varepsilon^* \quad (p\text{개의 모수})$$

실제 모형은 아래와 같다.

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (m\text{개의 모수})$$

여기서 물론 $m > p$ 이다. 다음은 부정확한 모형(incorrect model) 즉, p 항 모형(p -term model),에 대한 잔차평균제곱(residual mean square)이 m 항 모형(m -term model)의 추가 모수(extra parameters) β_2 에 의하여 얼마나 팽창(inflation)되었는지 결정하려고 한다. p 항 모형에 대한 $E(s^2)$ 를 얻기 위하여, B.1절에서 사용된 것과 동일한 결과 즉, 이차형태(quadratic form)의 기대값(expected value)을 이용한다. B.1절에서, 확률벡터(random vector) y 가 평균 μ , 분산공분산행렬(variance-covariance matrix) $\sigma^2 I$ 이고, $y'Ay$ 가 y 의 이차형태일 때,

$$E(y'Ay) = \sigma^2 \text{tr}(A) + \mu' A \mu$$

라는 것을 기억하라.

현재 상황에서, 과소 적합된 모형(underfitted model)에 대한 오차평균제곱(error mean square)은 다음과 같이 주어진다.

$$\frac{1}{n-p} y' [I - X_1(X_1'X_1)^{-1} X_1'] y$$

위의 식이 완전모형(full model)에 대한 잔차평균제곱(residual mean square) 형태와 얼마나 닮았는지 주목하라. 행렬 X_1 은 행렬 X 를 대체한다. 기대 연산자(expectation operator)를 적용하면,

$$\boxed{E\{y' [I - X_1(X_1'X_1)^{-1}X_1']y\} = \sigma^2 tr[I - X_1(X_1'X_1)^{-1}X_1'] + E(y)' [I - X_1(X_1'X_1)^{-1}X_1'] E(y)} \quad (\text{B.4})$$

그렇다면,

$$tr[I - X_1(X_1'X_1)^{-1}X_1'] = n - p$$

이면서,

$$tr[X_1(X_1'X_1)^{-1}X_1'] = tr[(X_1'X_1) \times (X_1'X_1)^{-1}] = p$$

그 다음에, 식(B.4) 우변의 두 번째 항을 구해보자. 참 모형(true model)에서 $E(y) = X_1\beta_1 + X_2\beta_2$ 이므로, 식(B.4)는 다음과 같이 됨을 안다.

$$\begin{aligned} E\{y' [I - X_1(X_1'X_1)^{-1}X_1']y\} &= \sigma^2(n - p) + (X_1\beta_1 + X_2\beta_2)' \\ &\quad \times [I - X_1(X_1'X_1)^{-1}X_1'] (X_1\beta_1 + X_2\beta_2) \end{aligned}$$

이제, 우변의 이차 형태(quadratic form)는 다음과 같이 간단하게 줄일 수 있는데,

$$(X_2\beta_2)' [I - X_1(X_1'X_1)^{-1}X_1'] (X_2\beta_2)$$

이것은,

$$\beta_2' [X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2] \beta_2$$

와 같이 표현될 수도 있다.

결론적으로 우리는 다음과 같은 식을 얻게 되고,

$$E(SS_{\text{Res}}) = \sigma^2(n - p) + \beta_2' [X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2] \beta_2$$

따라서,

$$E(s_p^2) = \sigma^2 + \frac{1}{n-p} \beta_2' [X_2' X_2 - X_2' X_1 (X_1' X_1)^{-1} X_1' X_2] \beta_2$$

이것은 식(3.27)과 (4.1)에 제시된 결과이다.

11.3. (B.3) 최대우도추정량(The Maximum Likelihood Estimator)

3.3절에서 오차가 정규분포한다는 조건에서 최소제곱법(least squares procedure)에 대하여 상당히 강조하였다. 선형회귀모형(linear regression model)의 경우, 최소제곱추정량(least squares estimator)의 성능(performance)은, 오차(errors)가 정규분포(normally distributed)하지 않는 비이상적인 상황(nonideal situation)보다는 정규분포하는 경우에(Gaussian errors) 더 낫다고 결론 내렸다. 또한 정규성(normality) 가정 하에서는 최소제곱추정량(least squares estimator)이 회귀계수(regression coefficients)의 벡터 β 의 최대우도추정량(maximum likelihood estimator)이라고 언급하였다. 최대우도(maximum likelihood)에 대한 기타 참고문헌은 7장과 9장에 소개하였다.

최대우도법(maximum likelihood procedure)의 특성(properties)에 대하여 수식을 전개하는 것은 본문의 범위를 넘어선다. 자세한 것은 Kendall and Stuart (1973)이나 Graybill (1976)을 보라. 그러나, 가우스밀도함수(Gaussian density function)와 우도 개념(notion of likelihood)에 익숙한 학생을 위하여, 정규오차(normal errors)의 경우 최소제곱추정량(least squares estimator)이 최대우도추정량(maximum likelihood estimator)이라는 것을 쉽게 증명할 수 있다.

식(3.2)의 일반선형모형(general linear model)을 고려해보라. 즉,

$$y = X\beta + \varepsilon$$

$\varepsilon \sim N(0, \sigma^2 I)$ 이라는 가정을 한다. ε 에 대한 정규오차 밀도함수(normal error density function)는 다음과 같다:

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2} \varepsilon_i^2\right\} \quad (-\infty < \varepsilon_i < \infty) \quad (i = 1, 2, \dots, n) \quad (\text{B.5})$$

$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 의 결합밀도(joint density)에 의하여 주어지는 우도(likelihood)는

$\prod_{i=1}^n f(\varepsilon_i)$ 이다. 식(B.5)로부터 우도(likelihood)는 다음과 같이 주어진다.

$$\prod_{i=1}^n f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left\{-\frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta)\right\}$$

우도(likelihood)의 자연로그(natural log)로 작업하는 것이 편리하다. 따라서, 아래의 식을

최대로 하는 \mathbf{b} 를 찾는다.

$$\ln \prod_{i=1}^n f(\varepsilon_i) = -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \quad (\text{B.6})$$

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

위의 항은 잔차제곱합(residual sum of squares)인데, 이것이 최소화될 때, 로그우도(log likelihood)가 최대화된다. 따라서 정규오차(normal error)일 경우 β 의 최대우도추정량(maximum likelihood estimator)은

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

으로 주어지는 최소제곱추정량(least squares estimator)과 동등하다.

7장에서, 일반화최소제곱(generalized least squares)의 개념을 소개하였다. 오차분산(error variances)이 동일(equal)하지 않다는 것을 안다면, 가중최소제곱(weighted least squares)을 고려하여야 한다. 여기서 강조할 것은 가중회귀(weighted regression)이다. 오차(errors)의 분산공분산행렬(variance-covariance matrix), \mathbf{V} 가 알려져 있다면, 식(7.4)에서 주어진 일반화최소제곱추정량(generalized least squares estimator), β^* 또한 최대우도추정량(maximum likelihood estimator)이라고 말할 수 있다. 물론, 가중최소제곱(weighted least squares)의 경우,

$$\mathbf{V} = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$$

최소제곱추정량(least squares estimator)으로 β^* 를 수식 전개할 경우, 보통최소제곱(ordinary least squares)과 거의 모든 면에서 유사하다. 더 일반적인 예에 대한 우도(likelihood)는 다음과 같이 주어진다. Graybill (1976) 혹은 Seber (1977)를 참조하라.

$$f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\beta)' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta) \right\} \quad \begin{array}{l} (-\infty < \varepsilon_i < \infty) \\ (i = 1, 2, \dots, n) \end{array}$$

여기서 결합확률밀도함수(joint probability density function)를 나타내기 위하여 $f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ 라는 표기법을 사용하였다. $f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ 를 최대화시키면

지수부분(exponential portion)을 최소화시키는 것이다. 따라서, β 에 대한 최대우도추정량(maximum likelihood estimator)은 β^* 이다. 식(7.5)에서 β^* 에 대하여 $SS_{\text{Res},V}$ 가 최소화된다. 결과적으로 추정량은

$$\beta^* = (X'V^{-1}X)^{-1}X'V^{-1}y$$

이 되는데, 이것은 식(7.4)에서 주어진 일반화최소제곱추정량(generalized least squares estimator)이다.

11.4. (B.4) PRESS 통계량의 수식 전개(Development of the PRESS Statistic)

식(4.6)에 주어진 i 번째 PRESS 잔차(residual)에 대한 식(expression)은 i 번째 자료포인트를 실제로 제거하지 않고도 $y_i - \hat{y}_{i,-i}$ 를 계산한다. A.5 절에서 전개된 Sherman-Morrison-Woodbury 정리(Theorem)를 사용하여 방정식(equation)을 증명할 수 있다.

$X'X$ 가 A 의 역할을 하고, x'_i (X 의 i 번째 열)가 z' 의 역할을 한다고 가정해 보자. 그러면, $(X'X - x_i x'_i)$ 는 i 번째 자료포인트가 제외되어(set aside)있을 때의 $X'X$ 행렬이다.

행렬 $(X'X - x_i x'_i)$ 는 아래의 식에 의하여 감소된(reduced) $X'X$ 행렬이다.

$$x_i x'_i = \begin{bmatrix} 1 & x_{1i} & x_{2i} & \cdots & x_{ki} \\ x_{1i}^2 & x_{1i}x_{2i} & \cdots & x_{1i}x_{ki} \\ & x_{2i}^2 & \cdots & x_{2i}x_{ki} \\ & & \ddots & \vdots \\ & & & x_{ki}^2 \end{bmatrix}$$

따라서,

$$(X'X - x_i x'_i) = \begin{bmatrix} n-1 & \sum_{j \neq i} x_{1j} & \sum_{j \neq i} x_{2j} & \cdots & \sum_{j \neq i} x_{kj} \\ \sum_{j \neq i} x_{1j}^2 & \sum_{j \neq i} x_{1j}x_{2j} & \cdots & \sum_{j \neq i} x_{1j}x_{kj} \\ & \sum_{j \neq i} x_{2j}^2 & \cdots & \sum_{j \neq i} x_{2j}x_{kj} \\ & & \ddots & \vdots \\ & & & \sum_{j \neq i} x_{kj}^2 \end{bmatrix}$$

이것은 i 번째 자료포인트를 사용하지 않는 $X'X$ 행렬이다. 사용되는 표기법(notation)은 $(X'_{-i} X_{-i})$ 이다. 식(A.11)로부터,

$$(X'_{-i} X_{-i})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} x_i x'_i (X'X)^{-1}}{1 - h_{ii}} \quad (\text{B.7})$$

식(B.7)은 6장에서 사용된 진단기준(diagnostic criteria)에 대한 수식 전개의 기초가 된다. PRESS 잔차(residual)는 다음에 의하여 주어진다.

$$e_{i,-i} = y_i - \mathbf{x}'_i \mathbf{b}_{-i}$$

여기서, \mathbf{b}_{-i} 는 i 번째 자료포인트를 제외하고(set aside) 계산된 계수(coefficients)의 벡터이다.

(B.7)로부터,

$$\begin{aligned} e_{i,-i} &= y_i - \mathbf{x}'_i \left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}} \right] \mathbf{X}'_{-i} \mathbf{y}_{-i} \\ &= y_i - \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{-i} \mathbf{y}_{-i} - \frac{h_{ii} \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{-i} \mathbf{y}_{-i}}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i - (1-h_{ii})\mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{-i} \mathbf{y}_{-i} - h_{ii} \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{-i} \mathbf{y}_{-i}}{1-h_{ii}} \\ &= \frac{(1-h_{ii})y_i - \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{-i} \mathbf{y}_{-i}}{1-h_{ii}} \end{aligned}$$

$\mathbf{X}'_{-i} \mathbf{y}_{-i} + \mathbf{x}'_i y_i = \mathbf{X}' \mathbf{y}$. 그러므로 우리는 i 번째 PRESS 잔차(residual)를 다음과 같이 쓸 수 있다.

$$e_{i,-i} = \frac{(1-h_{ii})y_i - \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}' \mathbf{y} - \mathbf{x}'_i y_i)}{1-h_{ii}}$$

$$\mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'' \mathbf{y} = \hat{y}_i . \text{ 따라서}$$

$$\begin{aligned} e_{i,-i} &= \frac{(1-h_{ii})y_i - \hat{y}_i + h_{ii}y_i}{1-h_{ii}} \\ &= \frac{y_i - \hat{y}_i}{1-h_{ii}} \\ &= \frac{e_i}{1-h_{ii}} \end{aligned}$$

11.5. (B.5) s_{-i} 의 계산(Computation of s_{-i})

식(5.6)에서, i 번째 관찰값(observation)을 제외하고 계산된 잔차표준편차(residual standard deviation)의 추정값(estimate)에 대하여 식(expression)이 주어졌다. 통계량(statistic) s_{-i} 는 R 스튜던트 통계량(R-student statistic)의 구성(formation)에 사용된다. s_{-i} 계산은, 행렬 X 에서 한 개의 열(row)을 제거한 후의 중요한 통계량을 결정하기 위하여 Sherman-Morrison-Woodbury 정리를 사용한 또 다른 예이다.

$(X'_{-i} X_{-i})^{-1}$ 을 $(X' X)^{-1}$ 와 관계 지우는 식(B.7)로 시작한다. 양변에 $X'y - x_i y_i$ 를 곱하면, 아래의 식을 얻는다.

$$b_{-i} = b - (X' X)^{-1} x_i y_i + \frac{(X' X)^{-1} x_i x_i' (X' X)^{-1} (X'y - x_i y_i)}{1 - h_{ii}}$$

항(terms)을 모으고 단순화시키면, 매우 유용한 방정식이 나온다.

$$b - b_{-i} = \frac{(X' X)^{-1} x_i e_i}{1 - h_{ii}} \quad (B.8)$$

자, 우리는 아래와 같이 쓸 수 있다.

$$(n - p - 1)s_{-i}^2 = \sum_{j \neq i} (y_j - x_j' b_{-i})^2 \quad (B.9)$$

따라서 다음과 같다.

$$b_{-i} = b - \frac{(X' X)^{-1} x_i e_i}{1 - h_{ii}}$$

아래와 같이 쓸 수 있다.

$$\begin{aligned}\sum_{j \neq i} (y_j - \mathbf{x}'_j \mathbf{b}_{-i})^2 &= \sum_{j=1}^n \left(y_j - \mathbf{x}'_j \mathbf{b} + \frac{\mathbf{x}'_j (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i e_i}{1-h_{ii}} \right)^2 - \left((y_i - \mathbf{x}'_i \mathbf{b} + \frac{h_{ii} e_i}{1-h_{ii}}) \right)^2 \\ &= \sum_{j=1}^n \left(e_j + \frac{h_{ij} e_i}{1-h_{ii}} \right)^2 - \frac{e_i^2}{(1-h_{ii})^2}\end{aligned}$$

이제 우리는 HAT 행렬과 잔차(residuals) 사이의 흥미로운 관계를 이용할 수 있다. 아래의 항(term)을 확장함으로써,

$$\sum_j \left(e_j + \frac{h_{ij} e_i}{1-h_{ii}} \right)^2$$

우리는 다음과 같은 것을 얻는다.

$$\sum_{j=1}^n e_j^2 + \frac{2e_i}{1-h_{ii}} \sum_{j=1}^n e_j h_{ij} + \frac{e_i^2}{(1-h_{ii})^2} \sum_{j=1}^n h_{ij}^2$$

$\mathbf{H}\mathbf{y} = \hat{\mathbf{H}}\mathbf{y}$ 이기 때문에, $\sum_{j=1}^n e_j h_{ij} = 0$ 이다. 추가적으로, $\mathbf{H}^2 = \mathbf{H}$ 이기 때문에, $\sum_{j=1}^n h_{ij}^2 = h_{ii}$ 이다. 따라서, $\sum_{j \neq i} (y_j - \mathbf{x}'_j \mathbf{b}_{-i})^2$ 은 다음과 같이 쓸 수 있다.

$$\begin{aligned}(n-p-1)s_{-i}^2 &= \sum_{j=1}^n e_j^2 + \frac{h_{ii} e_i^2}{(1-h_{ii})^2} - \frac{e_i^2}{(1-h_{ii})^2} \\ &= \sum_{j=1}^n e_j^2 - \frac{e_i^2}{1-h_{ii}} \\ &= (n-p)s^2 - \frac{e_i^2}{1-h_{ii}}\end{aligned}$$

마지막으로, 우리는 식(5.6)을 얻는다. 즉,

$$s_{-i} = \sqrt{\frac{(n-p)s^2 - [e_i^2/(1-h_{ii})]}{n-p-1}}$$

11.6. (B.6) 대응하는 모형오차에 대한 잔차의 우월성(Dominance of a Residual by the Corresponding Model Error)

5.7절에서, 우리는 ε_i 의 정규성(normality)과 관련된 증거를 밝히기 위하여 잔차(residuals)를 공부하였다. e_i 는 단지 대응하는 ε_i 의 함수(function)가 아니라, 오히려 모형오차 모두(all of the model errors)의 선형조합(linear combination)이라는 것을 설명하는데 식(5.14)을 이용하였다. 식(5.14)를 반복하면 다음의 식을 얻을 수 있다.

$$e_i = \varepsilon_i - \sum_{j=1}^n h_{ij} \varepsilon_j$$

모수의 수(number of parameters), p 는 일정한 반면, 표본크기(sample size), n 이 크게 증가한다면, h_{ij} 는 0이 되는 쪽으로 가는 경향이 있고 따라서 ε_i 가 우월하게 된다. 추가적으로, $\sum_{j=1}^n h_{ij} \varepsilon_j$ 의 분산(variance)은 $\sigma^2 (\sum_{j=1}^n h_{ij}^2) = \sigma^2 h_{ii}$ 이다. 만약 $n >> p$ 라면, h_{ij} 는 1.0보다 작아질 것이다. 따라서 $Var(e_i) \approx \sigma^2$ 이다. 결과적으로, 이런 상황의 $\sum_{j=1}^n h_{ij} \varepsilon_j$ 부분은 무시할 수 있게 되고, 따라서 e_i 의 정보는 본질적으로 ε_i 의 것이라는 것을 추측할 수 있다. 그렇다면, 모형에서 모수의 수(number of parameters)보다 표본크기(sample size)가 훨씬 클 때는 정규성을 더욱 정확하게 평가할 수 있다는 것은 확실하다.

11.7. (B.7) 영향력 진단도구의 계산(Computation of Influence Diagnostics)

이 절에서는, 6장에서 논의된 영향력 진단도구(influence diagnostic tools)에 대한 계산을 쉽게 해주는 세부적인 것들에 대하여 논의할 것이다. 통계량(statistics) DFFITS와 DFBETAS 그리고 Cook's D의 수식을 전개할 것이다. 각각의 경우, Sherman-Morrison-Woodbury 정리에서 이미 확인된 결과들을 사용한다.

DFFITS

식(6.4)에 주어진 **DFFITS**의 최종적인 형태(definitive form)를 살펴보자. 독자는 Sherman-Morrison-Woodbury 정리로부터 끌어낸 식(B.8)의 매우 유용한 공식을 재점검하여야 한다. 식(B.8)은 $\mathbf{b} - \mathbf{b}_{-i}$ 에 대한 식(expression)이며, 여기서 \mathbf{b}_{-i} 는 i 번째 자료포인트를 제외하고 얻어진 최소제곱추정량(least squares estimators)의 벡터이다. 다음과 같은 식,

$$\mathbf{b} - \mathbf{b}_{-i} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i e_i}{1 - h_{ii}}$$

여기서, 만약 양측에 \mathbf{x}'_i 를 단순하게 곱하면, 다음의 식을 얻는다.

$$\hat{y}_i - \hat{y}_{i,-i} = \frac{h_{ii}e_i}{1 - h_{ii}}$$

이제 **DFFITS**를 얻기 위하여, $\hat{y}_i - \hat{y}_{i,-i}$ 를 표준화시켜 보자.

$$\frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} = \left[\frac{e_i h_{ii}}{1 - h_{ii}} \right] \frac{1}{s_{-i}\sqrt{h_{ii}}} = \frac{e_i}{s_{-i}\sqrt{1 - h_{ii}}} \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

따라서 적합값(fitted values)에 대한 표준화된 차이(standardized difference)는 식(6.5)에서 주어진 것처럼 다음과 같다.

$$(DFFITS)_i = (R - \text{student})_i \left[\frac{h_{ii}}{1 - h_{ii}} \right]^{1/2}$$

COOK'S D

Cook's D 는 식(B.8)의 $\mathbf{b} - \mathbf{b}_{-i}$ 에 대한 식으로부터 쉽게 전개된다. 식(6.8)에 주어진 최종적인 형태(definitive form)는 다음과 같은 양(quantity)이다.

$$D_i = \frac{(\mathbf{b} - \mathbf{b}_{-i})' (\mathbf{X}' \mathbf{X}) (\mathbf{b} - \mathbf{b}_{-i})}{ps^2}$$

식(B.8)을 사용하여, 다음의 식을 얻는다.

$$D_i = \frac{x'_i (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{X})^{-1} x_i e_i^2}{(1 - h_{ii})^2 ps^2} = \left(\frac{e_i^2}{(1 - h_{ii})^2} \right) \left(\frac{h_{ii}}{ps^2} \right)$$

마지막으로, 식(6.9)에 주어진 것과 같은 다음의 식을 얻는다.

$$D_i = \left(\frac{r_i^2}{p} \right) \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

DFBETAS

식(6.6)에 주어진 최종적인 형태(definitive form)와 함께, DFBETAS 진단도구(diagnostic)는 식(B.8)에 주어진 $\mathbf{b} - \mathbf{b}_{-i}$ 의 j 번째 원소(element)이다. 표준화하기 전에,

$$b_j - b_{j,-j} = \frac{r_{j,i} e_i}{1 - h_{ii}}$$
(B.10)

식(B.10)을 $s_{-i} \sqrt{c_{jj}}$ 로 나누어 표준화한다. 여기서 c_{jj} 는 $(\mathbf{X}' \mathbf{X})^{-1}$ 의 j 번째 대각원소(diagonal element)이다. c_{jj} 요소는 $\mathbf{R}' \mathbf{R} = (\mathbf{X}' \mathbf{X})^{-1}$ 이기 때문에 단지 스칼라(scalar) 값 $r_j' r_j$ 이다. 따라서, 식(6.7)에 주어진 다음의 식을 얻는다.

$$\frac{b_j - b_{j,-i}}{s_{-i}\sqrt{c_{jj}}} = \left(\frac{r_{j,i}}{\sqrt{r'_j r_j}} \right) \left(\frac{e_i}{s_{-i}(1-h_{ii})} \right)$$

11.8. (B.8) 비선형 모형에서 최대우도추정량 (Maximum Likelihood Estimator in the Nonlinear Model)

비선형모형의 최소제곱추정량(least squares estimator)은 식(9.9)에 주어진 SS_{Res} 의 최소화(minimization)를 포함한다. 만약 모형오차(model errors)가 정규분포하며(normal), 독립이면서(independent), 공통분산(common variance), σ^2 을 가질 경우, 추정량(estimator)은 최대우도추정량(maximum likelihood estimator)이 된다. 부록 B.3에서 논의된, 선형모형(linear model)에 대한 최대우도(maximum likelihood)의 개념을 기억해보라. 이 개념은 비선형 쪽으로 쉽게 이전될 수 있다. 식(9.8)의 모형 표기(model notation)를 살펴보면, 우도(likelihood), 혹은 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 의 결합밀도(joint density)는 아래의 식으로 주어진다.

$$\prod_{i=1}^n f(\varepsilon_i) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - f(x_i, \theta)]^2 \right\}$$

따라서 최대우도추정량(maximum likelihood estimator)은 $\prod_{i=1}^n f(\varepsilon_i)$ 를 최대화시키는 θ 에 대한 $\hat{\theta}$ 값이다. 선형모형의 경우에서처럼, 우도(likelihood)는 지수(exponent)가 최소화될 때 최대화된다. 그러므로 $\sum_{i=1}^n [y_i - f(x_i, \hat{\theta})]^2$ 의 최소화와 더불어, 최소제곱추정량(least squares estimator)은 최대우도추정량(maximum likelihood estimator)이기도 하다.

11.9. (B.9) 테일러 급수(Taylor Series)

테일러 급수근사(Taylor series approximation)는 가우스, 뉴튼 비선형 추정법(Gauss-Newton nonlinear estimation procedure)의 기초로서 9.3절에 소개되어 있다. 식(9.10)에서 약술한 것처럼 Taylor 급수전개는 포함된 모수(parameters)에 대하여 선형이므로 자연스러운 메카니즘(natural mechanism)이다. 따라서 가우스, 뉴튼법은 $f(x, \theta)$ 의 선형화된 형태(linearized version)에서 계수(coefficients)를 추정(estimation)하며, 반복과정(iterative procedure)마다 이 추정값(estimate)을 계속 갱신한다.

Taylor 급수전개는 근사함수(approximating functions)에 대한 표준분석도구(standard analytical device)이다. 이 경우는 선형 항(linear terms) 이후 부분이 잘리기 때문에 다소 특별하다. 더 일반적으로 이야기하자면, 함수 $y = f(z_1, z_2, \dots, z_p)$ 가 있다고 가정하자.

$z = z_0$ 즉, $(z_1, z_2, \dots, z_p) = (z_{1,0}, z_{2,0}, \dots, z_{p,0})$ 근처에서 국소적 근사(local approximation)를 나타내는, Taylor 급수 전개는 아래의 식으로 주어진다.

$$\boxed{f(z) = f(z_0) + \sum_{i=1}^p (z_i - z_{i,0}) \left[\frac{\partial f}{\partial z_i} \right]_{z=z_0} + \sum_{i=1}^p \frac{(z_i - z_{i,0})^2}{2!} \left[\frac{\partial^2 f}{\partial z_i^2} \right]_{z=z_0} + \sum_{i < j} (z_i - z_{i,0})(z_j - z_{j,0}) \left[\frac{\partial^2 f}{\partial z_i \partial z_j} \right]_{z=z_0} + \dots} \quad (\text{B.11})$$

간단한 예는, $z=0$ 주변에서 e^z 의 전개(expansion)이다. 다음의 식과 같다.

$$\begin{aligned} e^z &= e^0 + e^0 z + \frac{e^0 z^2}{2!} + \frac{e^0 z^3}{3!} + \dots \\ &= 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots \end{aligned}$$

가우스, 뉴턴식으로 적용할 경우, θ 들은 식(B.11)의 z 들이며, 수식전개(expansion)는 선형항(linear terms) 이후부터 잘린다(truncated). 수식은 모수(parameters)의 시작값 $(\theta_{1,0}, \theta_{2,0}, \dots, \theta_{p,0})$ 주변에서부터 전개(expansion)된다.

11.10. (B.10) C_k 통계량의 수식전개(Development of the C_k -Statistic)

C_k 통계량은 8.4절에서 논의되었다. 전개(expansion)는 식(8.13)과 같이 주어진다. C_P

통계량(statistic)의 경우에서처럼, C_k 의 개념(intention)은 다음과 같은 양(quantity)을 추정하는 것이다.

$$\sum_{i=1}^n \text{Var } \hat{y}_{i,R} + \sum_{i=1}^n [\text{Bias } \hat{y}_{i,R}]^2 \quad (\text{B.12})$$

여기서 $\hat{y}_{i,R}$ 은, 능형회귀(ridge regression)를 사용하여, x_i 에서의 예측값(prediction) 혹은 적합값(fitted value)이다. 즉,

$$\hat{y}_{i,R} = x_i' b_R$$

중심화되고 척도화된 회귀변수(centered and scaled regressors)의 계수들(coefficients)인 추정량(estimators) $b'_{0,R}, b'_{1,R}, \dots, b'_{k,R}$ 은 아래에 대한 해로서 공식화될(formulated) 수 있다.

$$\begin{bmatrix} n & 0 & 0 & \cdots & 0 \\ 0 & & & & \\ 0 & (X^* X^* + kI) & & & \\ \vdots & & & & \\ 0 & & & & \end{bmatrix} \begin{bmatrix} b'_{0,R} \\ b'_{1,R} \\ \vdots \\ b'_{k,R} \end{bmatrix} = X' y \quad (\text{B.13})$$

여기서

$$X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad (X^*)$$

중심화되고 척도화된 x_i 자료포인트에서, 예측값 $\hat{y}_{i,R}$ 의 분산은 다음과 같이 주어지는데

$$\frac{\text{Var } \hat{y}_{i,R}}{\sigma^2} = x_i' \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & (X^{*'} X^* + kI) & & \\ \vdots & & & \\ 0 & & & \end{bmatrix}^{-1} (X' X) \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & (X^{*'} X^* + kI) & & \\ \vdots & & & \\ 0 & & & \end{bmatrix}^{-1} x_i$$

우리는 이제 $(\sum \text{Var } \hat{y}_{i,R} / \sigma^2)$ 을 다음과 같이 쓸 수 있다:

$$\begin{aligned} \frac{\sum_{i=1}^n \text{Var } \hat{y}_{i,R}}{\sigma^2} &= \text{tr} \left[X \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & (X^{*'} X^* + kI) & & \\ \vdots & & & \\ 0 & & & \end{bmatrix}^{-1} (X' X) \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & (X^{*'} X^* + kI) & & \\ \vdots & & & \\ 0 & & & \end{bmatrix}^{-1} X' \right] \\ &= \text{tr} \left[X \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & (X^{*'} X^* + kI) & & \\ \vdots & & & \\ 0 & & & \end{bmatrix}^{-1} X' \right]^2 \\ &= \text{tr}[A_k]^2 \end{aligned} \quad (\text{B.14})$$

자 이제, $\sum_{i=1}^n [(\text{Bias } \hat{y}_{i,R})^2 / \sigma^2]$ 을 살펴보자. C_P 통계량의 경우에서처럼, 편향 부분(bias portion)은 잔차제곱합(residual sum of squares)으로부터 추정된다. 우리가 8장에서 배웠던 것은 능형회귀(ridge regression)의 경우 계수들(coefficients)이 편향된다는(biased) 것이다. 적합값(fitted value)의 편향(bias), $\hat{y}_{i,R}$ 은 잔차제곱합(residual sum of squares)을 팽창시킨다. 우리는 다음과 같이 쓸 수 있다.

$$SS_{\text{Res},k} = (\mathbf{y} - \mathbf{X}\mathbf{b}_R)'(\mathbf{y} - \mathbf{X}\mathbf{b}_R) = \mathbf{y}'[\mathbf{I} - \mathbf{A}_k]^2 \mathbf{y}$$

이제 이차형태(quadratic form)의 기대값(expected value)에 대하여 식(B.3)에서 주어진 정리theorem)를 사용하여 아래의 식을 얻을 수 있다.

$$\begin{aligned} E\{\mathbf{y}'[\mathbf{I} - \mathbf{A}_k]^2 \mathbf{y}\} &= \sigma^2 \text{tr}[\mathbf{I} - \mathbf{A}_k]^2 + (\mathbf{X}\beta)'[\mathbf{I} - \mathbf{A}_k]^2 (\mathbf{X}\beta) \\ &= \sigma^2 \text{tr}[\mathbf{I} - \mathbf{A}_k]^2 + \beta' \mathbf{X}'[\mathbf{I} - \mathbf{A}_k]^2 \mathbf{X}\beta \end{aligned}$$

식(B.12)의 **총 오차(total error)**의 편향 부분(bias portion)은 다음과 같이 주어진다.

$$\begin{aligned} \sum_{i=1}^n (\text{Bias } \hat{y}_{i,R})^2 &= E(\mathbf{X}\beta - \mathbf{X}\mathbf{b}_R)' E(\mathbf{X}\beta - \mathbf{X}\mathbf{b}_R) \\ &= [\mathbf{X}\beta - \mathbf{X}E(\mathbf{b}_R)]' [\mathbf{X}\beta - \mathbf{X}E(\mathbf{b}_R)] \\ &= [\mathbf{X}\beta - \mathbf{A}_k \mathbf{X}\beta]' [\mathbf{X}\beta - \mathbf{A}_k \mathbf{X}\beta] \\ &= (\mathbf{X}\beta \beta)[\mathbf{I} - \mathbf{A}_k]^2 (\mathbf{X}\beta) \end{aligned}$$

따라서 $\beta' \mathbf{X}'[\mathbf{I} - \mathbf{A}_k]^2 \mathbf{X}\beta$, $\hat{y}_{i,R}$ 의 제곱편향합(sum of squared biases)에 대한 불편추정량(unbiased estimator)은 다음과 같이 주어진다.

$$\sum_{i=1}^n (\text{Bias } \hat{y}_{i,R})^2 = SS_{\text{Res},k} - \sigma^2 \text{tr}(\mathbf{I} - \mathbf{A}_k)^2$$

따라서 식(B.12)에서 식의 추정량은 다음과 같이 주어진다.

$$\begin{aligned} C_k &= \text{tr}(\mathbf{A}_k)^2 + \frac{SS_{\text{Res},k}}{\sigma^2} - \text{tr}(\mathbf{I} - \mathbf{A}_k)^2 \\ &= \frac{SS_{\text{Res},k}}{\sigma^2} - n + 2\text{tr}(\mathbf{A}_k) \end{aligned}$$

\mathbf{A}_k 와 행렬(matrix)의 정의(definition)로부터

$$H_k = X^* (X^{*\top} X^* + kI)^{-1} X^{*\top}$$

$$\text{tr}(A_k) = \text{tr}(H_k) + 1$$

그러므로, σ^2 대신에 OLS의 $\hat{\sigma}^2$ 을 사용하면,

$$C_k = \frac{SS_{\text{Res},k}}{\hat{\sigma}^2} - n + 2 + 2\text{tr}(H_k)$$

12. 부록 C: 행렬을 이용한 선형회귀분석

12.1. (C.1) 선형회귀분석에서의 기본적인 행렬

선형회귀분석에서 기본적인 행렬은 결과변수로 n 개의 관찰값을 가지는 $n \times 1$ \mathbf{Y} 행렬이다.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad (1)$$

위 행렬의 전치행렬(transpose matrix) \mathbf{Y}^T 는 열벡터이다.

$$\mathbf{Y}' = [Y_1 \quad Y_2 \quad \cdots \quad Y_n] \quad (2)$$

단순한 형태의 선형회귀분석에서의 또 하나의 기본적인 행렬은 아래와 같은 $n \times 2$ \mathbf{X} 행렬이다.

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad (3)$$

행렬 \mathbf{X} 는 1로 이루어진 열 하나와 독립변수 x 에 대한 관측값으로 이루어져 있다. 행렬 \mathbf{X} 의 전치행렬은 아래 식과 같다.

$$X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \quad (4)$$

회귀모형, $y_i = E\{Y_i\} + \varepsilon_i$ ($i = 1, \dots, n$) 은 행렬의 형태로 다음과 같이 표현된다.

$$E\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_1\} \\ E\{Y_2\} \\ \vdots \\ E\{Y_n\} \end{bmatrix} \quad (5)$$

오차의 경우는:

$$\underset{n \times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (6)$$

위로부터 회귀분석 모형은 행렬로 다음과 같이 표현할 수 있다.

$$\underset{n \times 1}{\mathbf{Y}} = E\{\underset{n \times 1}{\mathbf{Y}}\} + \underset{n \times 1}{\boldsymbol{\varepsilon}}$$

선형회귀 분석에서 종종 필요한 것이 $\mathbf{Y}^T \mathbf{Y}$ 이다.

$$\underset{1 \times 1}{\mathbf{Y}^T \mathbf{Y}} = [Y_1 \ Y_2 \ \dots \ Y_n] \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = [Y_1^2 + Y_2^2 + \dots + Y_n^2] = [\sum Y_i^2] \quad (7)$$

여기서 알 수 있듯이 $\mathbf{Y}^T \mathbf{Y}$ 는 1×1 행렬 즉 스칼라이다.

$\mathbf{X}^T \mathbf{X}$ 행렬도 또한 필요하며, 이는 2×2 행렬이다:

$$\underset{2 \times 2}{\mathbf{X}^T \mathbf{X}} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \quad (8)$$

$\mathbf{X}^T \mathbf{Y}$ 행렬도 또한 사용되는데 이는 2×1 행렬이다:

$$\underset{2 \times 1}{\mathbf{X}^T \mathbf{Y}} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \quad (9)$$

행렬계수(rank)는 행렬에서 선형적으로 비의존적인 행 혹은 열의 최대 갯수이다.

$r \times r$ 정방행렬의 역행렬(inverse matrix)은 행렬의 계수가 r 일 때 존재한다. 이러한 행렬은 nonsingular하다고 한다. $r \times r$ 정방행렬의 계수가 r 보다 작을 경우 singular 하다고 하며 이 경우 역행렬이 존재하지 않고 행렬식(determinant)도 0이다.

회귀분석에서 중요한 역행렬 중의 하나가 $\mathbf{X}^T \mathbf{X}$ 의 역행렬 $(\mathbf{X}^T \mathbf{X})^{-1}$ 이다.

$$\mathbf{X}^T \mathbf{X}_{2 \times 2} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{X}_{2 \times 2} \text{의 행렬식 } D = n \sum X_i^2 - (\sum X_i)(\sum X_i) = n \left[\sum X_i^2 - \frac{(\sum X_i)^2}{n} \right] = n \sum (X_i - \bar{X})^2$$

따라서

$$(\mathbf{X}^T \mathbf{X}_{2 \times 2})^{-1} = \begin{bmatrix} \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2} & \frac{-\sum X_i}{n \sum (X_i - \bar{X})^2} \\ \frac{-\sum X_i}{n \sum (X_i - \bar{X})^2} & \frac{n}{n \sum (X_i - \bar{X})^2} \end{bmatrix} \quad (10)$$

$\sum X_i = n\bar{X}$, $\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$ 이므로, 식 (10)을 단순화하면,

$$(\mathbf{X}^T \mathbf{X}_{2 \times 2})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum (X_i - \bar{X})^2} & \frac{1}{\sum (X_i - \bar{X})^2} \end{bmatrix} \quad (11)$$

12.2. (C.2) 무작위 벡터 혹은 행렬의 기대값

일반적으로 무작위 변수를 포함하고 있는 무작위 벡터 (random vector) \mathbf{Y} 의 기대값은:

$$\mathbb{E}\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_i\} \end{bmatrix}_{n \times 1} \quad (n=1, \dots, n)$$

(12)

$n \times p$ 차원의 무작위 벡터 \mathbf{Y} 의 기대값은,

$$\mathbb{E}\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_{ij}\} \end{bmatrix}_{n \times p} \quad (i=1, \dots, n; j=1, \dots, p) \quad (13)$$

$n=3$ 의 선형회귀분석의 경우 각기 기대값이 0인 $\varepsilon_1, \varepsilon_2, \varepsilon_3$ 의 3개의 오차가 있다.

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}_{3 \times 1} \quad (14) \quad \mathbb{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}_{3 \times 1} \quad (15) \quad (\begin{bmatrix} E\{\varepsilon_1\} \\ E\{\varepsilon_2\} \\ E\{\varepsilon_3\} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \text{이므로})$$

12.3. (C.3) 무작위 벡터의 분산-공분산 행렬

Y_1, Y_2, Y_3 세개의 관측값을 가지는 무작위 벡터 \mathbf{Y} 의 예를 다시 생각하면, 세 개의 무작위변수의 분산, $\sigma^2\{Y_i\}$ 과 변수들간의 공분산, $\sigma\{Y_i, Y_j\}$ 은 \mathbf{Y} 의 분산-공분산 행렬에 함께 표시되며 그 모양은 아래와 같다.

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} & \sigma\{Y_1, Y_3\} \\ \sigma\{Y_2, Y_1\} & \sigma^2\{Y_2\} & \sigma\{Y_2, Y_3\} \\ \sigma\{Y_3, Y_1\} & \sigma\{Y_3, Y_2\} & \sigma^2\{Y_3\} \end{bmatrix} \quad (14)$$

이로부터 아래와 같은 관계가 있음을 알 수 있다.

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = E\left\{ [\mathbf{Y} - E\{\mathbf{Y}\}] [\mathbf{Y} - E\{\mathbf{Y}\}]^T \right\} \quad (15)$$

분산-공분산 행렬을 $n \times 1$ 무작위 벡터 \mathbf{Y} 에 일반화 하면

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\}_{n \times n} = \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} & \cdots & \sigma\{Y_1, Y_n\} \\ \sigma\{Y_2, Y_1\} & \sigma^2\{Y_2\} & \cdots & \sigma\{Y_2, Y_n\} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\{Y_n, Y_1\} & \sigma\{Y_n, Y_2\} & \cdots & \sigma^2\{Y_n\} \end{bmatrix} \quad (16)$$

다시 이전의 예로 돌아가서, 3개의 오차가 일정한 분산을 가지고($\sigma^2\{\varepsilon_i\} = \sigma^2$), 서로

독립적($\sigma\{\varepsilon_i, \varepsilon_j\} = 0 ; i \neq j$)이라고 가정하면, 예에서의 무작위 벡터 ε 의 분산-공분산 행렬은 아래와 같다.

$$\boldsymbol{\sigma}^2\{\varepsilon\}_{3 \times 3} = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

이는 다시 $\boldsymbol{\sigma}^2\{\varepsilon\}_{3 \times 3} = \sigma^2 \mathbf{I}_{3 \times 3}$ 로 좀더 간단히 표현할 수 있다 (\mathbf{I} 는 3×3 단위행렬)

몇 가지 기본적 정리 (theorem)

선형회귀에서 종종 무작위 벡터 \mathbf{Y} 의 앞에 상수로만 이루어진 상수 행렬 \mathbf{A} 를 곱함으로써 얻어지는 무작위 벡터 \mathbf{W} 를 접하게 된다.

$$\mathbf{W} = \mathbf{AY} \quad (17)$$

이 경우 아래와 같은 기본적 정리가 사용된다.

$$E\{\mathbf{A}\} = \mathbf{A} \quad (18)$$

$$E\{\mathbf{W}\} = E\{\mathbf{AY}\} = \mathbf{AE}\{\mathbf{Y}\} \quad (19)$$

$$\sigma^2\{\mathbf{W}\} = \sigma^2\{\mathbf{AY}\} = \mathbf{A}\sigma^2\{\mathbf{Y}\}\mathbf{A}' \quad (20)$$

($\sigma^2\{\mathbf{Y}\}$: \mathbf{Y} 의 분산-공분산 행렬)

12.4. (C.4) 행렬식을 이용한 단순한 회귀모형

앞서 정의한 \mathbf{Y} 행렬((1)식)과 \mathbf{X} 행렬 ((3)식) 그리고 $\boldsymbol{\varepsilon}$ 벡터 ((6)식)을 이용하고 회귀계수(regression coefficient)로 이루어진 $\boldsymbol{\beta}$ 벡터를 정의하여 아래와 같이 회귀모형을 나타낼 수 있다.

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{n \times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \boldsymbol{\beta}_{2 \times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon}_{n \times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

(21)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix}$$

이므로 이는

다음과 같이 표시할 수 있다.

$$\mathbf{Y}_{n \times 1} = \mathbf{X}_{n \times 2} \boldsymbol{\beta}_{2 \times 1} + \boldsymbol{\varepsilon}_{n \times 1} \quad (22)$$

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \quad \text{이므로} \quad E\{\mathbf{Y}\}_{n \times 1} = \mathbf{X} \boldsymbol{\beta}_{2 \times 1}$$

(23)

따라서 $\mathbf{X} \boldsymbol{\beta}$ 벡터는 Y_i 에 대한 기대값이다.

오차에 관해서 회귀모형에서는 $E\{\varepsilon_i\} = 0$, $\sigma^2\{\varepsilon_i\} = \sigma^2$ 이라 가정하고 ε_i 는 독립적인 정규분포를 따르는 무작위 변수로 가정한다. $E\{\varepsilon_i\} = 0$ 를 행렬로 표현하면,

$$E\{\boldsymbol{\varepsilon}\}_{n \times 1} = 0 \quad (24)$$

오차가 일정한 분산 σ^2 을 가지고 모든 공분산이 0 ($\sigma\{\varepsilon_i, \varepsilon_j\}, i \neq j$; ε_i 는 독립적이라 가정하므로)인 경우 오차들은 분산-공분산 행렬에서 다음과 같이 표시된다.

$$\underset{n \times n}{\boldsymbol{\sigma}^2 \{ \boldsymbol{\varepsilon} \}} = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

(25)

식 (25)는 스칼라 행렬이므로,

$$\underset{n \times n}{\boldsymbol{\sigma}^2 \{ \boldsymbol{\varepsilon} \}} = \sigma^2 \mathbf{I} \quad (26)$$

따라서 정규분포 오차를 가정한 회귀 모형은 행렬식으로 다음과 같이 나타낼 수 있다.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (27)$$

($\boldsymbol{\varepsilon}$ 는 서로 독립적이면서 정규분포를 따르는 무작위변수로 구성된 벡터로서 $E\{\boldsymbol{\varepsilon}\} = \mathbf{0}$

이고 $\boldsymbol{\sigma}^2 \{ \boldsymbol{\varepsilon} \} = \sigma^2 \mathbf{I}$)

12.5. (C.5) 최소자승법을 이용한 회귀분석 파라메터의 추정

기본적인 선형회귀식 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ 으로부터,

$$nb_0 + b_1 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \quad b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i$$

(28)

(b_0 과 b_1 은 각각 β_0 과 β_1 의 점추정치)

이를 행렬로 표시하면

$$\begin{matrix} \mathbf{X}' \\ 2 \times 2 \end{matrix} \begin{matrix} \mathbf{X} \\ 2 \times 1 \end{matrix} \begin{matrix} \mathbf{b} \\ 2 \times 1 \end{matrix} = \begin{matrix} \mathbf{X}' \\ 2 \times 1 \end{matrix} \begin{matrix} \mathbf{Y} \\ 2 \times 1 \end{matrix}$$

(29)

(\mathbf{b} 는 회귀계수로 이루어진 벡터)

$$\begin{matrix} \mathbf{b} \\ 2 \times 1 \end{matrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

(30)

식 (29)에서 회귀계수 추정값을 행렬의 형태로 구하려면 식 (29)의 양변의 앞에 $X'X$ 의 역행렬(존재한다고 가정)을 곱해준다.

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X}\mathbf{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

$$(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} = \mathbf{I}, \quad \mathbf{I}\mathbf{\beta} = \mathbf{\beta} \text{ 이므로,}$$

$$\begin{matrix} \mathbf{\beta} \\ 2 \times 1 \end{matrix} = (\mathbf{X}'\mathbf{X})^{-1} \begin{matrix} \mathbf{X}' \\ 2 \times 1 \end{matrix} \begin{matrix} \mathbf{Y} \\ 2 \times 1 \end{matrix}$$

(31)

12.6. (C.6) 관찰 추정값

모델에 의한 추정치 \hat{Y} 로 이루어진 벡터를 $\hat{\mathbf{Y}}$ 로 표시하면,

$$\begin{array}{c} \hat{\mathbf{Y}} \\ \hline n \times 1 \end{array} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \quad (32)$$

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix} \text{이므로}$$

$$\begin{array}{c} \hat{\mathbf{Y}} \\ \hline n \times 1 \end{array} = \begin{array}{c} \mathbf{Y} \\ \hline n \times 2 \end{array} \begin{array}{c} \mathbf{b} \\ \hline 2 \times 1 \end{array} \quad (33)$$

$\hat{\mathbf{Y}}$ 는 또한 \mathbf{b} 에 대한 표현식을 이용하여 다음과 같이 표현될 수 있다.

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

혹은 마찬가지로,

$$\begin{array}{c} \hat{\mathbf{Y}} \\ \hline n \times 1 \end{array} = \begin{array}{c} \mathbf{H} \mathbf{Y} \\ \hline n \times n \quad n \times 1 \end{array} \quad (34)$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (34a)$$

식 (34)로부터 추정값 \hat{Y}_i 가 관측값 \mathbf{Y}_i 와 상관계수로 이루어진 행렬 \mathbf{H} 와 선형적인 관계로 이 표현됨을 알 수 있다. $n \times n$ 행렬 \mathbf{H} 는 hat 행렬로 불리우며, 회귀분석의 진단에 중요한 역할을 하며 대칭행렬이며 아래 식 (35)와 같이 등멱(idempotency)의 특성을 지닌다.

$$\mathbf{HH} = \mathbf{H} \quad (35)$$

12.7. (C.7) 잔차

잔차 벡터 $e_i = Y_i - \hat{Y}_i$ 를 \mathbf{e} 로 표시하면,

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n \times 1} \quad (36)$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} \quad (37)$$

12.8. (C.8) 잔차의 분산-공분산 행렬

잔차(e_i)도 \mathbf{H} 행렬을 이용해 아래와 같이 표현된다.

$$\mathbf{e} = \mathbf{Y} - \overset{\wedge}{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathbf{e} = \underset{n \times 1}{(\mathbf{I} - \mathbf{H})} \underset{n \times n}{\mathbf{Y}}$$

(38)

$(\mathbf{I} - \mathbf{H})$ 행렬도 \mathbf{H} 행렬과 마찬가지로 대칭적이고 등멱의 특성을 지닌다.

\mathbf{e} 의 분산-공분산 행렬은,

$$\sigma^2 \{ \mathbf{e} \} = (\mathbf{I} - \mathbf{H}) \sigma^2 \{ \mathbf{Y} \} (\mathbf{I} - \mathbf{H})^T \text{ 로 } \mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \text{ 으로부터 유도된다.}$$

$$\sigma^2 \{ \mathbf{Y} \} = \sigma^2 \{ \mathbf{e} \} = \sigma^2 \mathbf{I} \text{ 이고 } (\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})^T ((\mathbf{I} - \mathbf{H})) \text{ 가 대칭행렬} \text{이므로,}$$

$$\sigma^2 \{ \mathbf{e} \} = (\mathbf{I} - \mathbf{H}) \sigma^2 \mathbf{I} (\mathbf{I} - \mathbf{H}) = \sigma^2 (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})$$

또한 $(\mathbf{I} - \mathbf{H})$ 등멱의 특성을 지니므로.

$$\sigma^2 \{ \mathbf{e} \} = \sigma^2 (\mathbf{I} - \mathbf{H}) \quad (39)$$

σ^2 을 알 수 없는 경우가 많으므로 이에 대한 추정치인 Mean Squared Error (MSE)를 이용하여 아래와 같이 추정한다.

$$\mathbf{s}^2 \{ \mathbf{e} \} = MSE(\mathbf{I} - \mathbf{H}) \quad (40)$$

12.9. (C.9) 회귀계수의 분산-공분산 행렬

\mathbf{b} 의 분산-공분산 행렬을 구하여면 앞서의 $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}\mathbf{Y}$ 관계를 이용한다.

$$\text{여기서 } \mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

기본정리식 (20)에 의해, $\sigma^2\{\mathbf{b}\} = \mathbf{A}\sigma^2\{\mathbf{Y}\}\mathbf{A}'$ 로 유도된다.

$(\mathbf{X}'\mathbf{X})^{-1}$ 은 대칭행렬이므로,

$$\mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

따라서,

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

$$(\sigma^2\{\mathbf{Y}\} = \sigma^2\mathbf{I})$$

따라서 \mathbf{b} 의 분산-공분산 행렬은,

$$\sigma^2_{2 \times 2}\{\mathbf{b}\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} \end{bmatrix}$$

(41)

혹은 식 (11)을 이용하여

$$\sigma^2_{2 \times 2}\{\mathbf{b}\} = \begin{bmatrix} \frac{\sigma^2}{n} + \frac{\sigma^2\bar{X}^2}{\sum(X_i - \bar{X})^2} & \frac{-\bar{X}^2\sigma^2}{\sum(X_i - \bar{X})^2} \\ \frac{-\bar{X}^2\sigma^2}{\sum(X_i - \bar{X})^2} & \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \end{bmatrix}$$

(42)

로 나타낼 수 있다.

위식의 σ^2 대신 MSE 를 사용하면 \mathbf{b} 의 분산-공분산 행렬 추정값, $s^2\{\mathbf{b}\}$ 를 구할 수 있다.

$$\mathbf{s}^2 \{ \mathbf{b} \}_{2 \times 2} = MSE \left(\mathbf{X}' \mathbf{X} \right)^{-1} = \begin{bmatrix} \frac{MSE}{n} + \frac{MSE \bar{X}^2}{\sum (X_i - \bar{X})^2} & \frac{-\bar{X}^2 MSE}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X}^2 MSE}{\sum (X_i - \bar{X})^2} & \frac{MSE}{\sum (X_i - \bar{X})^2} \end{bmatrix} \quad (43)$$

기초통계 요약(Summary of Basic Statistics)

1. 변수(자료)의 유형

- ① 범주형(분류형) 변수(categorical variables) : 명목 또는 순서자료
 - (i) 명목 변수(nominal var.) ... 성별{남 0, 여 1}
 - (ii) 순서 변수(ordinal var.) ... 학력{중졸이하 1, 고졸 2, 대졸이상 3}
 - ⇒ 질적 속성을 가짐
- ② 측정형(연속형) 변수(measurement variables) : 구간 또는 비례척도의 자료
 - (i) 구간 척도(interval scale) ... IQ, 온도
 - (ii) 비례 척도(ratio scale) ... 키, 몸무게
 - ⇒ 질적 속성을 가짐
- ③ 계수형(이산형) 변수(counting variables) : 나이, 형제수, 출생아수 ...
 - ⇒ 질적 속성을 가짐

2. 종속변수 vs. 독립변수

- 반응변수, 결과변수, Y 변수 등 : 확률적 변수
- 설명변수, 예측변수, X 변수 등 : 수학적 변수

3. 확률과 확률 분포

3-1 확률의 정의

용어 정의

- ① 표본공간(sample space) : 통계적 실험에서 모든 가능한 결과의 집합
- ② 사건(event) : 관심이 있는 실험 결과의 집합(표본공간의 부분집합)
- ③ 근원사건 : 한 개의 원소로 된 사건

예

주사위 1 개 $S = \{ 1, 2, 3, 4, 5, 6 \}$

A : 짹수 $A = \{ 2, 4, 6 \}$

동전 1 개 $S = \{H, T\}$

동전 2 개 $S = \{(H,H), (H,T), (T,H), (T,T)\}$

사건의 연산

- ① 합사건 $A \cup B$
- ② 곱사건 $A \cap B$
- ③ 배반사건 $A \cup B = \emptyset$ 일 때 A 와 B 는 서로 배반

확률의 고전적 정의

표본공간의 각 근원사건이 일어날 가능성의 같은 경우에, 사건 A 의 확률은

$$P(A) = \frac{A \text{의 원소의 개수}}{\text{표본공간의 원소의 개수}}$$

성질 S : 표본공간, A : 사건

- ① $P(S) = 1, 0 \leq P(A) \leq 1$
- ② $A \subset B$ 이면 $P(A) \leq P(B)$
- ③ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

3-2 조건부 확률

조건부 확률 : A 하에서 B 가 일어날 조건부 확률 $P(B | A) = \frac{P(A \cap B)}{P(A)}$

예 주사위 1 개를 던질 때 A : 짝수, $B = \{6\}$ 이면 $P(B | A) = \frac{1}{3}$

$\Rightarrow A$ 를 새로운 표본 공간으로 간주하고, A 내에서 B 가 일어날 확률

곱셈법칙 $P(A \cap B) = P(A | B) \cdot P(B) = P(B | A) \cdot P(A)$

전확률공식

S 의 한 분할 : A_1, A_2, \dots, A_n

즉, $A_i \cap A_j = \emptyset, i \neq j$

$A_1 \cup A_2 \cup \dots \cup A_n = S$

$$P(B) = P(B \mid A_1)P(A_1) + \cdots + P(B \mid A_n)P(A_n)$$

독립사건 A 와 B 가 서로 독립이면

$$\textcircled{1} \quad P(A \cap B) = P(A)P(B)$$

$$\textcircled{2} \quad P(A \mid B) = P(A), P(B \mid A) = P(B)$$

3-3 확률변수와 확률분포

확률변수 표본공간에서 정의된 실수함수

예 동전 2 개, $X = \text{Table}$ 면의 개수

X 의 확률분포

x	0	1	2	합
$p(x) = P(X = x)$	1/4	1/2	1/4	1

확률변수의 종류 : 이산형, 연속형

확률분포 : 확률변수의 수 값들에 확률을 대응시켜주는 관계

① 이산형 : 높이의 합 = $P(a \leq X \leq b)$

② 연속형 : 면적 = $P(a \leq X \leq b)$

성질

[이산형]

[연속형]

$$\textcircled{1} \quad 0 \leq p(x) \leq 1 \quad p(x) \geq 0$$

$$\textcircled{2} \quad \sum_x p(x) = 1 \quad \int_{-\infty}^{\infty} p(x) dx = 1$$

$$\textcircled{3} \quad P(a \leq X \leq b) = \sum_{a \leq x \leq b} p(x) \quad P(a \leq X \leq b) = \int_a^b p(x) dx$$

예제

X 의 확률밀도함수가

$$p(x) = \begin{cases} bx(1-x), & 0 \leq x \leq 1 \\ 0 & o.w. \end{cases}$$

일 때, b 와 $P(0 \leq X \leq 3/4)$ 의 값은?

$$\textcircled{1} \int_0^1 bx(1-x)dx = b \left(\frac{x^2}{2} - \frac{x^3}{3} \right) \Big|_0^1 = b \cdot \frac{1}{6} = 1 \text{ 따라서 } b = 6$$

$$\textcircled{2} \int_0^{3/4} 6x(1-x)dx = 6 \left(\frac{x^2}{2} - \frac{x^3}{3} \right) \Big|_0^{3/4} = 6 \cdot \left(\frac{9}{32} - \frac{9}{64} \right) = \frac{27}{32}$$

3-4 기대값과 그의 성질

1) X 의 평균(mean) 또는 기대값(expected value)

$$E(x) = \begin{cases} \sum_x xp(x) \\ \int_{-\infty}^{\infty} xp(x)dx \end{cases}$$

기호 $E(X), \mu, \mu_X$

$g(X)$ 을 기대값

$$E(g(X)) = \begin{cases} \sum_x g(x)p(x) \\ \int_{-\infty}^{\infty} g(x)p(x)dx \end{cases}$$

성질

$$\textcircled{1} E(aX + b) = aE(X) + b$$

$$\textcircled{2} E(X_1 + X_2) = E(X_1) + E(X_2)$$

2) 분산과 표준편차

① X 의 분산(variance)

$$Var(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 p(x) \\ \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx \end{cases}$$

기호 $Var(X), \sigma^2, \sigma_X^2$

간편계산법 $Var(X) = E(X^2) - \mu^2$

② X 의 표준편차(standard deviation)

$$sd(X) = \sqrt{Var(X)}$$

기호 $sd(X), \sigma, \sigma_X$

성질 $Var(aX + b) = a^2 Var(X)$

예

$E(X) = \mu, Var(X) = \sigma^2$ 일 때 $Z = \frac{X - \mu}{\sigma}$ (Z : 표준화된 확률변수)로 정의하면

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{E(X) - \mu}{\sigma} = 0 \text{ 으로 정의하면}$$

$$Var(Z) = Var\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(X) = 1, \quad sd(X) = 1$$

3-5 두 확률변수의 결합분포

예 (X, Y) : 이변량 확률변수

- ① X : 키, Y : 몸무게
- ② X : 영어성적, Y : 수학성적

결합확률밀도함수 : 두 개 이상의 확률변수가 동시에 여러 가지 값들에 확률을 대응시켜 주는 관계

- ① 이산형 : $p(x, y) = P(X = x, Y = y)$
- ② 연속형 : $p(x, y)$ 의 그래프

예제 주사위 1개

X : (1,2) - 100 원, (3,4) - 200 원, (5,6) - 300 원

Y : 짹수 - 100 원, 훌수 -(눈의 수) \times 100 원

① X 와 Y 의 결합분포 $p(x, y)$

$x \backslash y$	100	300	500	X 의 주변분포
100	2/6	0	0	2/6
200	1/6	1/6	0	2/6
300	1/6	0	1/6	2/6
Y 의 주변분포	4/6	1/6	1/6	1

② $Z = X + Y$ 의 확률분포

z	200	300	400	500	800
$P(z)$	2/6	1/6	1/6	1/6	1/6

정의 $p(x, y) = p_1(x)p_2(y)$ 일 때 X 와 Y 는 서로 독립이라 한다.

예

① 예제의 X 와 Y 는 독립이 아님

② 주사위 2 개

X = 첫 번째 주사위의 눈의 수

Y = 두 번째 주사위의 눈의 수

$\Rightarrow x=1, \dots, 6, y=1, \dots, 6$ 의 각각에 대하여

$$p(x, y) = \frac{1}{36}, p_1(x) = \frac{1}{6}, p_2(y) = \frac{1}{6}$$

$$p(x, y) = p_1(x)p_2(y)$$

3-6 공분산과 상관계수

X 와 Y 의 공분산(Covariance)

확률변수 X 와 Y 가 같이 변하는 정도의 측도

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\text{간편계산법} : Cov(X, Y) = E[XY] - \mu_1\mu_2$$

X 와 Y 의 상관계수(Correlation Coefficient)

공분산은 각 확률변수가 취하는 값의 단위에 의존하므로, 이러한 단위에 대한 의존도를 없애주기 위한 것.

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

성질 $-1 \leq \rho \leq 1$

예

①

$x \backslash y$	1	3
1	1/4	1/4
3	1/4	1/4

$$\mu_X = \mu_Y = 2$$

$$E(XY) = 1 \cdot 1 \cdot \frac{1}{4} + 1 \cdot 3 \cdot \frac{1}{4} + 3 \cdot 1 \cdot \frac{1}{4} + 3 \cdot 3 \cdot \frac{1}{4}$$

$$Cov(X, Y) = 4 - 2 \cdot 2 = 0$$

$$Corr(X, Y) = 0$$

②

$x \backslash y$	1	3
1	1/8	3/8
3	3/8	1/8

$$\mu_X = \mu_Y = 2$$

$$E(XY) = 1 \cdot 1 \cdot \frac{1}{8} + 1 \cdot 3 \cdot \frac{8}{8} + 3 \cdot 1 \cdot \frac{3}{8} + 3 \cdot 3 \cdot \frac{1}{8} = \frac{7}{2}$$

$$Cov(X, Y) = \frac{7}{2} - 2 \cdot 2 = -\frac{1}{2}$$

$$\sigma_X^2 = E(X^2) - \mu_X^2 = (1^2 \cdot \frac{1}{2} + 3^2 \cdot \frac{1}{2}) - 2^2 = 1$$

$$\sigma_Y^2 = 1$$

$$\rho = Corr(X, Y) = \frac{-1/2}{\sqrt{1}\sqrt{1}} = -\frac{1}{2}$$

성질

$$\textcircled{1} \quad Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$$

② X 와 Y 가 서로 독립이면

$$\text{(i)} \quad E(XY) = E(X)E(Y)$$

$$\text{(ii)} \quad Cov(X, Y) = 0, \quad Corr(X, Y) = 0$$

$$\text{(iii)} \quad Var(X + Y) = Var(X) + Var(Y)$$

참고

$$\textcircled{1} \quad X$$
와 Y 가 서로 독립 $\Rightarrow \rho = 0$

② 공분산과 상관계수는 두 확률변수의 직선적인 관계의 정도를 측정한다.

4. 표본 분포

4-1 베르누이분포와 정규분포

(1) 베르누이 분포

베르누이 시행의 표본공간 : $S = \{s, f\}$

: 어떤 실험의 결과를 오직 두 가지 중의 하나로 생각하는 시행

베르누이 분포

$$P(X=1) = p \quad (\text{성공확률})$$

$$P(X=0) = 1-p \quad (\text{실패확률})$$

x	0	1
$P(x)$	$1-p$	p

기호 $X \sim B(1, p)$

성질

$$\textcircled{1} \quad E(X) = 0 \cdot (1-p) + 1 \cdot p = p$$

$$\textcircled{2} \quad Var(X) = E(X^2) - \{E(X)\}^2 = p(1-p) = pq$$

(2) 정규분포(normal distribution) - 가우스 분포

평균 μ , 표준편차 σ 인 정규분포의 밀도함수

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty,$$

기호 $X \sim N(\mu, \sigma^2)$

정리 $X \sim N(\mu, \sigma^2)$ 일때,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1) : \text{표준정규분포}$$

표준정규분포 : $Z = -3.49 \sim 3.49$ 에 대하여 $P\{Z \leq z\}$ 값

기호 $Z_a : P\{Z \geq z\} = \alpha$ 인 값

확률의 계산

$$\textcircled{1} \quad P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$

$$\textcircled{2} \quad P(Z \geq a) = 1 - P(Z < a)$$

정규분포의 확률계산

$X \sim N(\mu, \sigma^2)$ 일때,

$$\begin{aligned} P(a \leq Z \leq b) &= P\left\{\frac{a-\mu}{\sigma} \leq \frac{Z-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right\} \\ &= P\left\{\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right\} \end{aligned}$$

정리

$$X_1 \sim N(\mu_1, \sigma_1^2) \quad X_2 \sim N(\mu_2, \sigma_2^2)$$

X_1 과 X_2 는 서로 독립

$$a_1 X_1 + a_2 X_2 \sim N(a_1 \mu_1 + a_2 \mu_2, a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2)$$

4-2 표본분포

용어

- ① 모수 (parameter) - 모집단의 특성을 결정하는 상수
- ② 통계량(statistic) - 표본으로부터 계산 가능한 표본의 특성값
- ③ 추정량 - 모수의 추정을 위한 통계량 예) $\bar{X}, S^2, \hat{p}, \dots$
- ④ 표본분포 - 통계량의 확률분포

ex) 표본분포의 예

- 1. {찬성, 반대, 찬성, 찬성, 반대} 와 같은 모집단에서 크기 3인 표본을 단순랜덤비복원추출로 뽑아 모비율 $p(=0.6)$ 를 추정하는 문제.

- 2. 표본의 종류와 그 확률

가능한 표본	확률	표본비율의 값	오차
찬성, 찬성, 찬성	1/10	1	0.4
찬성, 찬성, 반대	6/10	2/3	0.067
찬성, 반대, 반대	3/10	1/3	-0.267

- 3. 사실들

- ① 모비율을 정확히 추정할 확률은 0
- ② 추정값의 오차의 절대값이 0.05 이하일 확률은 0
- ③ 추정값의 오차의 절대값이 0.1 이하일 확률은 0.6
- ④ 추정값의 오차의 절대값이 0.3 이하일 확률은 0.9

- 4. 표본비율

- ① 표본결과에 따라 하나의 수 값을 대응시키는 확률변수
- ② 표본으로부터 계산한 표본의 특성이므로 통계량
- ③ 가운데 두 열은 통계량의 확률분포
- ④ 표본분포(추정량의 확률분포)는 추정의 정확도를 나타내는 중요한 도구

단순랜덤추출법

: N 개의 원소가 있는 모집단에서 n 개의 표본을 추출할 때,

$\binom{N}{n}$ 개의 모든 경우들이 같은 확률로 추출되도록 하는 추출법

랜덤표본 또는 임의표본 (random sample)

- (1) (유한모집단) 단순랜덤 비복원추출로 뽑은 표본
- (2) (무한모집단) 서로 독립이며 같은 분포를 갖는 확률변수의 집합

예. 전구 생산 공정에서 n 개를 랜덤하게 추출

X_1 : 첫 번째 전구의 수명

⋮

X_n : n 번째 전구의 수명

관측전의 $\{X_1, X_2, \dots, X_n\}$: 랜덤표본

관측된 $\{x_1, x_2, \dots, x_n\}$: 데이터

4-3 초기하와 이항분포

- (1) 초기하분포(hypergeometric distribution)

모집단의 크기: N

속성 A 의 크기 : D

$X =$ 크기 n 인 랜덤표본에서 속성 A 를 갖는 것의 개수

분포함수

$$P(x) = \frac{\binom{D}{n} \binom{N-D}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \dots, \min\{n, D\}$$

성질

$$\textcircled{1} \quad E(X) = np$$

$$\textcircled{2} \quad Var(X) = np(1-p) \frac{N-n}{N-1}$$

(2) 이항분포(binomial distribution)

X_1, X_2, \dots, X_n : 서로 독립이고 모수 p 인 베르누이 확률변수

$$X = X_1 + X_2 + \dots + X_n$$

$$\text{기호} \quad X \sim B(n, p)$$

분포함수

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

누적분포함수 부록 C Table 2.

$$p(X \leq c) = \sum_{x=0}^c \binom{n}{x} p^x (1-p)^{n-x}$$

성질 $X \sim B(n, p)$ 일 때,

$$\textcircled{1} \quad E(X) = np$$

$$\textcircled{2} \quad Var(X) = np(1-p) = npq$$

초기하의 이항근사

$$X \sim \text{초기하분포}, N \rightarrow \infty, \frac{D}{N} \rightarrow p$$

$$\Rightarrow X \sim B(n, p)$$

4-4 표본평균의 분포

X_1, X_2, \dots, X_n : 평균 μ , 분산 σ^2 인 랜덤표본 (무한모집단을 가정)

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} : \text{표본평균}$$

① \bar{X} 의 기대값(평균)

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \dots + X_n)\right) = \mu$$

② \bar{X} 의 분산

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}(X_1 + \dots + X_n)\right) \\ &= \frac{1}{n^2} \{Var(X_1) + \dots + Var(X_n)\} = \frac{\sigma^2}{n} \end{aligned}$$

③ \bar{X} 의 표준편차 (표준오차)

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

용어 표준오차 (standard error) - 추정량의 표준편차

정리 $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ 에서의 랜덤표본

$$\Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\text{또는 } Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

중심극한정리: X_1, X_2, \dots, X_n 평균 μ , 분산 σ^2 인 랜덤표본

$\Rightarrow n$ 이 충분히 클 때

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\text{또는 } Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

이항분포의 정규근사

X_1, X_2, \dots, X_n : 베르누이 $B(1,p)$ 에서의 랜덤표본
(평균 p , 분산 $p(1-p)$)

$$X = \sum_{i=1}^n X_i \sim B(n, p)$$

$$\hat{p} = \frac{X}{n} : 표본비율$$

n 이 충분히 클 때,

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$$

참고 $np > 5$ 이고 $n(1-p) > 5$ 일 때 정규근사가 안전

연속성 수정

연속확률분포를 이용하여 이산확률분포의 확률을 근사시킬 때 근사의 정밀도를 높이는데 이용

$$P(a \leq X \leq b) \approx EP\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

예) $X \sim B(15, 0.4)$ 일 때, $P(7 \leq X \leq 10) = ?$

$$\begin{aligned} ① \quad P(7 \leq X \leq 10) &= P(X \leq 10) - P(X \leq 6) \\ &= 0.991 - 0.610 = 0.381 \end{aligned}$$

② 정규근사 : $np = 6$, $np(1-p) = 3.6$

$$\begin{aligned} P\left(\frac{7 - 6}{\sqrt{3.6}} \leq Z \leq \frac{10 - 6}{\sqrt{3.6}}\right) &= P(0.53 \leq Z \leq 2.11) \\ &= 0.9826 - 0.7019 = 0.2807 \end{aligned}$$

③ 연속성 수정에 의한 정규근사

$$P\left(\frac{6.5-6}{\sqrt{3.6}} \leq Z \leq \frac{10.5-6}{\sqrt{3.6}}\right) = P(0.26 \leq Z \leq 2.37) \\ = 0.9911 - 0.6026 = 0.3885$$

참고 \sqrt{npq} 의 값이 클 때는 연속성 수정이 필요치 않음.

4-5 χ^2, t, F 분포

(1) χ^2 분포

정의

Z_1, Z_2, \dots, Z_k : 서로 독립인 $N(0, 1)$ 확률변수

$$V = Z_1^2 + \dots + Z_k^2$$

$\Rightarrow V \sim$ 자유도가 k 인 χ^2 분포

기호

$$\textcircled{1} \quad V \sim \chi^2(k)$$

$$\textcircled{2} \quad \chi_\alpha^2(k) : P\{V \geq \chi_\alpha^2(k)\} = \alpha \text{ 인 점.}$$

χ^2 분포 Table : k 와 α 에 $\chi^2(k)$ 대한 값

정리 (χ^2 의 가법성)

$$V_1 \sim \chi^2(k_1), \quad V_2 \sim \chi^2(k_2)$$

V_1 과 V_2 는 독립

$$\Rightarrow V_1 + V_2 \sim \chi^2(k_1 + k_2)$$

정리 (표본분산의 분포)

X_1, X_2, \dots, X_n : $N(\mu, \sigma^2)$ 에서의 랜덤표본

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

증명 (idea)

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

참고 제곱합의 자유도 = 제곱이된 잔차의 개수 - 선형조건의 개수

예 $\sum_{i=1}^n (X_i - \bar{X})^2$ 의 자유도 :

제곱이된 잔차 : n 개

선형조건 $\sum_{i=1}^n (X_i - \bar{X}) = 0$: 1 개

\Rightarrow 자유도 : $n-1$ 개

정리 (합동표본분산의 분포)

$X_1, X_2, \dots, X_n : N(\mu_1, \sigma^2)$ 에서의 랜덤표본

$Y_1, Y_2, \dots, Y_n : N(\mu_2, \sigma^2)$ 에서의 랜덤표본

(서로 독립, σ^2 : 미지의 공통분산)

$$S_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad S_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

$$S_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

$$\Rightarrow \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

참고 합동표본분산(pooled sample variance)

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

(2) t 분포

idea $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

σ 대신 S 대입 : $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t$ 분포

정의

$$Z \sim N(0, 1), V \sim \chi^2(k)$$

Z 와 V 는 독립

$$\Rightarrow T = \frac{Z}{\sqrt{V/k}} \sim \text{자유도 } k \text{ 인 } t \text{ 분포}$$

기호

① $T \sim t(k)$

② $t_\alpha(k)$: $P\{T \geq t_\alpha(k)\} = \alpha$ 인 점.

성질

① $T = 0$ 에 관하여 대칭

② $k \rightarrow \infty$ 일 때 $T \sim Z$

t 분포 Table : k 와 α 에 $t_\alpha(k)$ 값

참고 $t_\alpha(\infty) = Z_\alpha$

예 $t_{0.025}(5) = 2.571, t_{0.025}(20) = 2.086, t_{0.025}(\infty) = 1.960 = Z_{0.025}$

정리 (studentized \bar{X} 의 분포) - μ 에 관한 추론에 사용

X_1, X_2, \dots, X_n : $N(\mu, \sigma^2)$ 에서의 랜덤표본

$$\Rightarrow T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

정리 (studentized $\bar{X} - \bar{Y}$ 의 분포) - $(\mu_1 - \mu_2)$ 에 관한 추론에 사용

X_1, X_2, \dots, X_n : $N(\mu_1, \sigma^2)$ 에서의 랜덤표본

Y_1, Y_2, \dots, Y_n : $N(\mu_2, \sigma^2)$ 에서의 랜덤표본

(서로 독립, σ^2 : 미지의 공통분산)

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

$$\Rightarrow T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2)$$

참고

① 모집단이 정규분포가 아니라도, n 이 크면

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

② n 이 충분히 크면

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1)$$

(3) F 분포

(두 분산의 비교, 분산분석 등에 사용)

정의

$$V_1 \sim \chi^2(k_1), \quad V_2 \sim \chi^2(k_2)$$

V_1 과 V_2 는 독립

$$\Rightarrow F = \frac{V_1/k_1}{V_2/k_2} \sim \text{자유도 } (k_1, k_2) \text{인 } F \text{ 분포}$$

기호

① $F \sim F(k_1, k_2)$

② $F_\alpha(k_1, k_2) : P\{F \geq F_\alpha(k_1, k_2)\} = \alpha$ 일 점.

F 분포 Table : (k_1, k_2) 에 대한 $F_\alpha(k_1, k_2)$ 값

성질

① $F \sim F(k_1, k_2)$ 일 때, $1/F \sim F(k_2, k_1)$

② $F_{1-\alpha}(k_1, k_2) = 1/F_\alpha(k_2, k_1)$

정리 (표본분산의 비의 분포)

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

참고 (F 와 t 의 관계)

$$T \sim t(k) \text{ 일 때 , } T^2 = \left(\frac{Z}{\sqrt{\chi^2/k}} \right)^2 = \frac{Z^2/1}{\chi^2/k} \sim F(1, k)$$

4-6 Q-Q plot (quantile quantile plot)

(; 분위수 대조도 ; 분위수대 분위수 그림)

Q-Q plot

(정규분포의 분위수, 자료분포의 분위수)의 산점도

정규분포의 분위수 : $\binom{i}{n}$ 분위수 또는 $\frac{i - \frac{3}{8}}{n + \frac{1}{4}}$ 분위수 ($Z_{(i)}$ 로 표시)

자료의 분위수 : i 번째로 큰 자료 $x_{(i)}$

성질

① 정규분포에 가까우면 직선의 모양

② 짧은 꼬리 (얇은 꼬리)

긴 꼬리 (두터운 꼬리)

왼쪽으로 긴 꼬리

오른쪽으로 긴 꼬리

5. 회귀분석 요약

1. 상관분석

두 변수 사이의 연관성을 분석, 표본상관계수를 이용하여 모상관계수에 대해 추론

$$- X \text{와 } Y \text{의 상관계수} : \rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

$$- \text{표본상관계수} : r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

(1) 기호 및 간편 계산법

$$S_{(xx)} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

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

$$S_{(xy)} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$r = \frac{S_{(xy)}}{\sqrt{S_{(xx)} S_{(yy)}}}$$

(2) 성질

$$-1 \leq r \leq 1$$

어떤 직선 주위에 밀집되어 나타날수록 -1 또는 1에 가깝게 주어진다.

(3) H_0 : $\rho = 0$ 의 검정(이변량 정규모집단의 경우)

$$\text{검정통계량} : T = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \quad (H_0 \text{에서 } t(n-2) \text{임을 이용})$$

검정법 : t -검정

2. 단순회귀분석의 모형과 적합

- 두 변수 사이의 함수관계를 분석
- 한 변수값으로부터 다른 변수의 값에 대한 예측
- 단순회귀분석 : 직선관계를 모형으로 분석
- 중회귀분석 : 두 개 이상의 변수가 한 변수에 영향을 줄 때 분석

(1) 모형

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, n$$

x_1, x_2, \dots, x_n : 설명변수(독립변수)

α, β : 회귀모수(모회귀계수)

Y_1, Y_2, \dots, Y_n : 반응변수(종속변수)

e_1, e_2, \dots, e_n : 서로 독립인 $(0, \sigma^2)$ 확률변수(오차항): 선형성, 등분산성, 독립성

(2) 모회귀직선

$$E(Y | x) = \alpha + \beta x$$

(3) 관측값

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

2.1 최소제곱추정값

$Q = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$ 을 최소로 하는 $\hat{\alpha}, \hat{\beta}$ 을 각각 α, β 의 최소제곱추정량(least squares estimator) 이라 한다.

(1) 추정량의 유도

$$Q = \sum_{i=1}^n [y_i - (\alpha + \beta x_i)]^2$$

$$\frac{\partial Q}{\partial \alpha} = 2 \sum (y_i - \alpha - \beta x_i)(-1) = 0 \Rightarrow \sum y_i = n\alpha + \beta \sum x_i$$

$$\frac{\partial Q}{\partial \beta} = 2 \sum (y_i - \alpha - \beta x_i)(-x_i) = 0 \Rightarrow \sum x_i y_i = \alpha \sum x_i + \beta \sum x_i^2$$

(2) 정규방정식(normal equation)

$$\begin{cases} n\alpha + \beta \sum x_i = \sum y_i \\ \alpha \sum x_i + \beta \sum x_i^2 = \sum x_i y_i \end{cases}$$

(3) 최소제곱추정량

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{(xy)}}{S_{(xx)}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

(4) 최소제곱 회귀직선

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

2.2 평균제곱오차

(1) 잔차

$$\hat{e}_i = \gamma_i = y_i - \hat{y}_i$$

(2) 잔차제곱합(residual sum of squares, error sum of squares)

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

(3) 평균제곱오차 (mean squared error)

$$MSE = \frac{SSE}{n-2} \quad (n-2 : SSE \text{의 자유도})$$

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

$$\therefore \hat{\sigma}^2 = MSE$$

(4) 제곱합의 분해

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

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

총제곱합 잔차제곱합 회귀제곱합

$$SST \qquad \qquad SSE \qquad \qquad SSR$$

$$n-1 \qquad \qquad n-2 \qquad \qquad 1 \quad (\text{자유도})$$

* SST (Total sum of squares)

* SSR (Regression sum of squares)

(5) 결정계수 (coefficient of determination)

$$r^2 = \frac{SSR}{SST} \quad (\text{뜻}) \text{ 총변동 가운데 회귀직선으로 설명되는 변동의 비율}$$

(6) 제곱합 계산법

$$SST = S_{(yy)}$$

$$SSR = \frac{[S_{(xy)}]^2}{S_{(xx)}}$$

$$SSE = SST - SSR$$

2.3. 단순회귀분석에서의 추론

오차항의 분포 가정이 있으면 모회귀계수에 관한 구간추정이나 검정과 같은 추론이 가능하므로 오차항이 정규분포라는 가정을 한다.

$$\text{모형} : Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, n$$

$e_i \sim$ 서로 독립인 $N(0, \sigma^2)$ 확률변수

2.3.1. 모형에 관한 추론 (모회귀계수에 대한 검정)

(1) 회귀직선모형에 대한 가설 $H_0: \beta=0, H_1: \beta \neq 0$

$$(2) \text{ 검정통계량} : F = \frac{SSR/1}{SSE/(n-2)}$$

귀무가설하에서 $F \sim F(1, n-2)$ 이므로 관측값이 f 이면 유의확률은 $P=P(F \geq f)$

(3) 기각역 : $F \geq F_\alpha(1, n-2)$

(4) 회귀직선의 유의성 검정을 위한 분산분석 Table

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	SSR	1	$MSR=SSR/1$	$F=MSR/MSE$	$P(F \geq f)$
잔차	SSE	$n-2$	$MSE=SSE/(n-2)$		
계	SST	$n-1$			

*MSR (Regression mean squares)

(5) 예제 8.5

2.3.2 β 에 관한 추론

(1) β 의 추정량

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i - \sum (x_i - \bar{x})\bar{y}}{\sum (x_i - \bar{x})^2} = \sum \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right) y_i$$

(2) $\hat{\beta}$ 의 기대값과 분산

$$E(\hat{\beta}) = \frac{\sum (x_i - \bar{x})(\alpha + \beta x_i)}{\sum (x_i - \bar{x})^2} = \beta \frac{\sum (x_i - \bar{x})x_i}{\sum (x_i - \bar{x})^2} = \beta$$

$$Var(\hat{\beta}) = \sum \frac{(x_i - \bar{x})^2}{[\sum (x_i - \bar{x})^2]^2} \sigma^2 = \frac{\sigma^2}{S_{(xx)}}$$

(3) $\hat{\beta}$ 의 분포

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{(xx)}}\right)$$

(4) $\hat{\beta}$ 의 표준오차의 추정량

$$SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{S_{(xx)}}}, \quad \hat{\sigma} = \sqrt{MSE}$$

(5) 표준화된 $\hat{\beta}$ 의 분포

$$T = \frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{S_{(xx)}}} \sim t(n-2)$$

① β 의 100(1- α)% 신뢰구간

$$\hat{\beta} \pm t_{\alpha/2}(n-2) \frac{\hat{\sigma}}{\sqrt{S_{(xx)}}}$$

② $H_0 : \beta = \beta_0$ 의 검정

$$\text{검정통계량} : T = \frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{S_{(xx)}}}$$

(6) 예제 8.6

(7) 참고

$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ 를 검정할 때

$$T^2 \sim F(1, n-2) \text{이며 } T \equiv F$$

2.3.3 평균반응 $E(Y|x) = \alpha + \beta x$ 에 관한 추론

(1) $E(Y|x) = \alpha + \beta x$ 의 추정량

$$\hat{Y} \equiv E(0Y|x) = \hat{\alpha} + \hat{\beta}x$$

(2) \hat{Y} 의 기대값과 분산

$$\hat{Y} = \alpha + \beta x$$

$$Var(\hat{Y}) = \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{(xx)}} \right] \sigma^2$$

(3) 표준화된 \hat{Y} 의 분포

$$T = \frac{(\hat{\alpha} + \hat{\beta}x) - (\alpha + \beta x)}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{(xx)}} \right]}} \sim t(n-2) \text{ 단 } \hat{\sigma}^2 = MSE$$

① $\alpha + \beta x$ 의 $100(1-\alpha)\%$ 의 신뢰구간

$$(\hat{\alpha} + \hat{\beta}x) \pm t_{\alpha/2}(n-2)\hat{\sigma} \sqrt{\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{(xx)}} \right]}$$

관심대상인 설명변수의 값을 중심으로 퍼져있는 설명변수의 값에서 반응변수를 관측하는 것이 바람직하다.

② $H_0: \alpha + \beta x = \mu_0$ 의 검정

$$\text{검정통계량} : T = \frac{(\hat{\alpha} + \hat{\beta}x) - \mu_0}{\hat{\sigma} \sqrt{\left[\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{(xx)}} \right]}}$$

(4) 예제 8.7

2.3.4. α 에 관한 추론

(1) α 의 추정량

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \sim N\left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{(xx)}} \right) \right)$$

$\hat{\alpha} + \hat{\beta}x$ 에서 $x=0$ 인 경우로 생각

(2) 표준화된 $\hat{\alpha}$ 의 분포

$$T = \frac{\hat{\alpha} - \alpha}{\hat{\sigma} \sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{(xx)}} \right]}} \sim t(n-2)$$

(3) 예제 8.8

2.4. 단순회귀분석에서의 잔차분석

(1) 모형

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \dots, n$$

e_1, e_2, \dots, e_n : 서로 독립인 $N(0, \sigma^2)$ 인 확률변수

(2) 가정

① 선형성 : $E(Y|X) = \alpha + \beta x$

② 독립성 : e_1, e_2, \dots, e_n 은 서로 독립

③ 등분산성 : e_1, e_2, \dots, e_n 의 분산은 모두 σ^2

④ 정규성 : $e_i \sim N(0, \sigma^2)$

(3) 잔차분석 (analysis of residual)

① 모형의 타당성 검토 - 가정에 대한 검토 - 잔차의 검토 - 잔차분석

$$\frac{e_i}{sd(e_i)} \sim i.i.d. N(0,1) \text{ 를 이용한다.}$$

② 잔차 : $\hat{e}_i = y_i - \hat{y}_i$

$$③ \text{잔차의 분산} : Var(\hat{e}_i) = \sigma^2(1-h_i), \quad h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{(xx)}}$$

$$④ \text{스튜던트화 잔차} : \hat{e}_{st,i} = \frac{\hat{e}_i}{sd(\hat{e}_i)} = \sqrt{\frac{y_i - \hat{y}_i}{MSE \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{(xx)}} \right)}}$$

⑤ 잔차도(residual plot) : $(x_1, \hat{e}_{st,1}), \dots, (x_n, \hat{e}_{st,n})$ 의 plotting

(4) 잔차도의 관찰

- ① 대략 0에 관하여 대칭적으로 나타나고
- ② 설명변수의 값에 따른 잔차의 산포가 크게 다르지 않고
- ③ 점들이 특정한 형식을 가지고 나타남이 없으며
- ④ 거의 모든 점이 ± 2 의 범위내에 나타나야 한다.

(5) 단순회귀 적용의 순서

- ① 산점도 검토 - 직선 관계 확인
- ② 최소제곱법에 의한 적합 및 잔차분석
- ③ 모든 조건이 만족되면 추론문제 및 예측모형 설정 시행

(6) 가정이 어긋난 경우 - 그림 8.10

(7) 예제 8.9

3. 중회귀분석

3.1. 중회귀모형

(1) 설명변수가 k 개 있는 중회귀모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + e_i \quad i = 1, 2, \dots, n$$

단, $\beta_0, \beta_1, \dots, \beta_k$: 회귀모수

$x_{1i}, x_{2i}, \dots, x_{ki}$: 설명변수(독립변수)

y_1, y_2, \dots, y_n : 반응변수(종속변수)

e_1, e_2, \dots, e_n : 서로 독립인 $N(0, \sigma^2)$ 확률변수

(2) 행렬 형식

$$y = X\beta + e$$

단, X 는 i 번째 행이 $(1, x_{1i}, x_{2i}, \dots, x_{ki})$ 인 $n \times (k+1)$ 행렬

β 는 $(\beta_0, \beta_1, \dots, \beta_k)'$ 로 된 $(k+1)$ 열벡터

e 은 $(e_1, e_2, \dots, e_n)'$ 으로 된 n 열벡터

y 는 $(y_1, y_2, \dots, y_n)'$ 으로 된 n 열벡터

이와 같은 행렬을 사용하면 β 의 최소제곱추정량 및 적합된 모형은 다음과 같다.

(3) 최소제곱추정량

$$\hat{\beta} = (X'X)^{-1}X'y \quad (X': X의 전치행렬)$$

$$\hat{y} = X\hat{\beta}$$

(4) 분산분석 Table

요인	제곱합	자유도	평균제곱	F 값	p-값
회귀	SSR	k	MSR	$f = MSR/MSE$	$P\{F \geq f\}$
잔차	SSE	$n-k-1$	MSE		
계	SST	$n-1$			

(5) 결정계수

$$R^2 = \frac{SSR}{SST}$$

(6) 잔차분석

스튜던트와 잔차를 구하는 방법은 설명변수가 하나인 회귀분석과 같으나 잔차도를 그릴 때 \hat{y}_i 를 가로축으로 하여 나타낸다.

6. 범주형 자료에 대한 확률분포들 (Distributions for categorical data)

- 이항 분포 (binomial distribution)

$$Y_i \sim \text{iid} \text{Bernoulli}(\pi), \quad i = 1, 2, \dots, n$$

Let $Y = \sum_{i=1}^n Y_i$, then $Y \sim B(n, \pi)$

- $\mu = E(Y) = n\pi, \quad \sigma^2 = \text{Var}(Y) = n\pi(1-\pi),$
- skewness : $E(Y - \mu)^3 / \sigma^3 = (1 - 2\pi) / \sqrt{n\pi(1-\pi)}$

- 다항 분포 (multinomial distribution)

: Each of n independent, identical trials can have outcome any of c categories. Let n_j be the number of trials having outcome in category j , then (n_1, n_2, \dots, n_c) have the multinomial distribution.

- Notation: $(n_1, n_2, \dots, n_c) \sim \text{multi}(\pi_1, \pi_2, \dots, \pi_c),$
Where $\pi_j = P(\text{a trial has outcome } \in \text{category } j)$
- Note that $\sum_j n_j = n, \quad \sum_j \pi_j = 1$
- pdf : $p(n_1, n_2, \dots, n_c) = \left(\frac{n!}{n_1! n_2! \dots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \dots \pi_c^{n_c}$
- $E(n_j) = n\pi_j, \quad \text{Var}(n_j) = n\pi_j(1-\pi_j), \quad \text{Cov}(n_j, n_k) = -n\pi_j\pi_k \quad (\exists j \neq k)$

- 포아송 분포 (Poisson distribution)

예) y : 일일 교통사고 사망자수

- Poisson pdf with mean $\mu (>0)$: $P(y) = \frac{e^{-y}\mu^y}{y!}, \quad y=0,1,\dots.$

\Leftarrow Poisson process of 0|항분포의 근사

- $E(Y) = \mu, \quad Var(Y) = \mu, \quad E(Y - \mu)^3 / \sigma^3 = 1/\sqrt{\mu}$

- 과대 산포 (overdispersion)

일일 교통사고 사망자수, 해조류 개수

- 포아송 분포와 다항 분포의 관계

- $Y_i \sim \text{Poisson}(\mu_i), \quad i = 1, \dots, c$

Let $n = \sum_i Y_i$, then $n \sim \text{Poisson}(\sum_i \mu_i)$

($\Leftarrow c=2$ 일 때 직접 유도해 봄.)

- $P(Y_1 = n_1, Y_2 = n_2, \dots, Y_c = n_c \mid \sum_i Y_i = n) = \frac{n!}{n_1! n_2! \dots n_c!} \pi_1 \pi_2 \dots \pi_c$

where $\pi_j = \mu_j / \sum_i \mu_i$

7. 통계적 추론(Statistical inference for categorical data)

- Likelihood

- Likelihood function and maximum likelihood Estimation

: Given the data, for a chosen probability distribution the likelihood function is the probability of those data, treated as a function of the unknown parameter.

- Notation: β : an unknown parameter or parameter vector,

$\ell(\beta)$: the likelihood function,

$L(\beta)$: the log-likelihood function, i.e $L(\beta) = \log \ell(\beta)$,

$\hat{\beta}$: maximum likelihood estimate(MLE),

$cov(\hat{\beta})$: the asymptotic covariance matrix of $\hat{\beta}$,

$I(\beta) = -E\left(\frac{\partial^2 L(\beta)}{\partial \beta' \partial \beta}\right)$: the information matrix.

- Under regularity conditions, $cov(\hat{\beta}) = I(\beta)^{-1}$

- Ex) $Y \sim B(n, \pi)$

cf) the kernel of the likelihood.

- Wald-Likelihood-Score Test

- Null hypothesis $H_0 : \beta = \beta_0$

- Wald test statistic

$$z = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

$$W = (\hat{\beta} - \beta_0)' cov(\hat{\beta})^{-1} (\hat{\beta} - \beta_0) \approx^{H_0} \chi^2(df), \text{ where } df = rank(cov(\hat{\beta}))$$

- Likelihood ratio statistic

ℓ_o : the maximized value of the likelihood function under H_0

ℓ_I : the maximized value generally (i.e., under $H_0 \cup H_I$)

$$\Lambda = \frac{\ell_0}{\ell_1} .$$

Then the likelihood-ratio test statistic

$$-2 \log(\Lambda) = -2 \log(\ell_0 / \ell_1) \approx {}^{H_0} \chi^2(df),$$

where the df equal, the difference in the dimensions of the parameter space under $H_0 \cup H_I$ and under H_0 .

- Score test statistic

$u(\beta) = \partial L(\beta) / \partial \beta$: the score function.

The Chi-squared form score test statistic

$$\frac{[u(\beta_0)]^2}{I(\beta_0)} \approx {}^{H_0} \chi^2(df)$$

- 신뢰구간 (Constructing confidence intervals)

- A 95% CI for β is the set of β_0 for which the test of $H_0 : \beta = \beta_0$ has a p -value exceeding 0.05
- The Wald confidence interval : the set of β_0 for which $|\hat{\beta} - \beta_0| / SE(\hat{\beta}) < z_{\alpha/2}$
- The likelihood-ratio confidence interval:

the set of β_0 for which $-2[L(\beta_0) - L(\hat{\beta})] < \chi^2_{\alpha}(1)$

- Binomial parameters

An observation y from $B(n, \pi)$.

$$H_0 : \pi = \pi_0, \quad MLE \hat{\pi} = y/n..$$

- The Wald test statistic and CI :

$$\text{Test statistic : } Z_w = \frac{\hat{\pi} - \pi_0}{SE(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$$

$$100(1-\alpha)\% \text{ CI : } \hat{\pi} \pm z_{\alpha/2} \sqrt{\hat{\pi}(1-\hat{\pi})/n}$$

- The score test statistic and CI

$$\text{Test statistic : } Z_s = \frac{u(\pi_0)}{I(\pi_0)^{1/2}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

$$(\because u(\pi_0) = \frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}, \quad I(\pi_0) = \frac{n}{\pi_0(1-\pi_0)})$$

$$100(1-\alpha)\% \text{ CI : } |\hat{\pi} - \pi_0| / \sqrt{\pi_0(1-\pi_0)/n} < z_{\alpha/2}$$

cf) The score statistic give better result of inference than the Wald statistic.

- The likelihood-ratio test statistic and CI

$$\text{Test statistic : } -2 \log \frac{\ell_0}{\ell_1} = 2 \left(y \log \frac{y}{n\pi_0} + (n-y) \log \frac{n-y}{n-n\pi_0} \right)$$

$$100(1-\alpha)\% \text{ CI : } -2 \log \frac{\ell_0}{\ell_1} < \chi^2_{\alpha}(1)$$

- Example (Proportion of vegetarians)

$$n=25, y=0$$

$$\text{The 95\% Wald interval : } \hat{\pi} \pm z_{\alpha/2} \sqrt{\hat{\pi}(1-\hat{\pi})/n} \Rightarrow (0,0)$$

$$\text{The 95\% score interval : } (0, 0.133)$$

$$\text{The 95\% likelihood-ratio interval : } (0, 0.074)$$

$$-2 \log \frac{\ell_0}{\ell_1} = -50 \log(1-\pi_0) < \chi^2_{0.05}(1) = 1.96^2 = 3.84$$

- Multinomial parameters

$$(n_1, n_2, \dots, n_c) \sim multi(\pi_1, \pi_2, \dots, \pi_c),$$

The kernel of the likelihood : $\pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}, \quad \sum_j n_j = n, \quad \sum_j \pi_j = 1.$

Log-likelihood: $L(\pi) = \sum_j n_j \log \pi_j$

MLE of π_j : $\hat{\pi}_j = \frac{n_j}{n}, \quad j = 1, \dots, c$

- **Pearson chi-square statistic**

$$H_0 : \pi_j = \pi_{j0}, \quad j = 1, \dots, c, \quad \sum_j \pi_{j0} = 1$$

$$\chi^2 = \sum_n \frac{(n_j - \mu_j)^2}{\mu_j},$$

where $\mu_j = n \pi_{j0}$ expected frequency.

Let χ^2 be the observed value of χ^2

$$p\text{-value} = P(\chi^2 \geq \chi_0^2),$$

Note $\chi^2 \approx^{H_0} \chi^2(c-1)$

- **Example** (Testing Mendel's Theories)

$$H_0 : \pi_{10} = 0.75, \pi_{20} = 0.25$$

Observation $n=8023, n_1=6022$ (yellow), $n_2=2001$ (green)

Expected frequencies

$$\mu_1 = n \pi_{10} = 8023 \times 0.75 = 6017.25, \quad \mu_2 = n \pi_{20} = 8023 \times 0.25 = 2005.75$$

In 1936, R.A. Fisher summarized independent Mendel's 84 results.

$$\Rightarrow \chi^2 = 42 (df = 84), p\text{-value} = 0.9996$$

- **Likelihood-ratio chi-squared**

$$H_0 : \pi_j = \pi_{j0}, \quad j = 1, \dots, c, \quad \sum_j \pi_{j0} = 1$$

$$\text{Likelihood ratio } \Lambda = \frac{\ell_0}{\ell_1} = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}}$$

Likelihood-ratio statistic

$$G^2 = -2 \log \Lambda = -2 \sum_j n_j \log(n_j/n\pi_{j0}) \approx \chi^2(df), \text{ where}$$

$df = \text{dimension of parameter space under } H_0 \cup H_1$

- dimension of the parameter space under H_0

그리스 문자(Greek Letters)

이미지	텍스트	이름과 설명
A α	A α	알파(ALPHA): 그리스문자의 첫번째 글자이다.
B β	B β	베타(BETA)
Γ γ	Γ γ	감마(GAMMA)
Δ δ	Δ δ	델타(DELTA)
E ε,ε	E ϵ	엡실론(EPSILON): 입실론 소문자 2번째형태는 "집합원소" 기호로 많이 사용된다.
Z ζ	Z ζ	제타(ZETA)
H η	H η	에타(ETA)
Θ θ	Θ θ	쎄타(THETA)
I ι	I ι	이오타(IOTA)
K κ	K κ	카파(KAPPA)
Λ λ	Λ λ	람다(LAMBDA)
M μ	M μ	谬(MU)
N ν	N ν	뉴(NU)
Ξ ξ	Ξ ξ	크사이(XI)
O ο	O \circ	오미크론(OMICRON) : 알파벳의 'o'와 비슷해서 거의 안 쓴다.

Π π	$\Pi \pi$	파이(PI) : 파이의 소문자는 보통 원의 직경에 대한 비율로 많이 쓴다. 파이의 대문자는 "곱하기"의 기호로 많이 쓴다.
P ρ,ϱ	$P \rho$	로우(RHO)
Σ σ,ζ	$\Sigma \sigma$	시그마(SIGMA): 시그마의 대문자는 "더하기"의 기호로 많이 쓴다.
T τ	$T \tau$	타우(TAU)
Υ υ	$\Upsilon \upsilon$	웁실론(UPSILON)
Φ φ,ϕ	$\Phi \phi$	화이(PHI): 소문자 2개는 바꿔서 많이 사용된다.
X χ	$X \chi$	카이(CHI)
Ψ ψ	$\Psi \psi$	프사이(PSI)
Ω ω,ϖ	$\Omega \omega$	오메가(OMEGA): 그리스문자의 마지막 글자이다.