제 1 장

계량경제학 입문

1.1 계량경제학을 왜 공부해야 하는가?

계량경제학은 경제학 분야에서만 중요한 것이 아니다. 이는 회계, 재무, 마케팅, 경영관리와 같은 경영학 분야에서도 사용되는 연구 수단이 되었다. 이는 또한 사회과학을 연구하는 사람들, 특히 역사학, 정치학, 사회학을 연구하는 사람들에 의해서도 사용되고 있다. 계량경제학은 임업과 같은 다양한 분야와 농업경제학에서 중요한 역할을 담당하고 있다. 계량경제학의 관심분야가 넓어진 이유는 부분적으로 경제학이 사업분석을 하는 데 기초가 되는 사회과학이기 때문이다. 경제학자가 이용하는 계량경제학에 기초한 연구방법은 다양한 사람들에게 유용하게 이용되고 있다.

계량경제학은 경제학자들을 교육시키는 데 특별한 역할을 담당하고 있다. 여러분들은 경제학을 배우는 학생으로서 '경제학자처럼 생각하는 방법'을 배우고 있다. 여러분들은 기회비용, 희소성, 비교우위와 같은 경제개념을 배우고 있다. 또한 수요 및 공급, 거시경제 행태, 국제무역과 관련된 경제모형에 관해 교육받고 있다. 위와 같은 학습을 통해 우리가 살고 있는 세상을 보다 잘 이해하게 될 것이다. 즉, 시장이 어떻게 운용되고 정부정책이 시장에 어떤 영향을 미치는지 알게 될 것이다.

경제학을 전공하거나 부전공하는 경우 독자들은 졸업 후 다양한 취업기회를 갖게 될 것이다. 민간기업에 취업하고자 하는 경우, 고용하는 사람들은 독자들이 "나를 위해 귀하는 무엇을 할 수 있는가?"에 대해 대답해 주기를 바랄 것이다. 전통적인 경제학 교과과정에 따라 대답하는 학생은 "나는 경제학자처럼 생각할 수 있다."라고 답할 것이다. 우리들은 이 대답이 호소력이 있다고 생각할 수도 있지만 경제학을 이해하지 못하는 고용인들에게는 만족스럽지 못할 수도 있다.

문제는 경제학과 학생으로 배운 것과 경제학자가 실제로 하는 것 사이에는 큰 차이가 있다는 데 있다. 극히 극소수의 경제학자들이 경제 이론만을 연구하면서 생활을 영위하고 있으며 이들은 일반적으로 대학에 취업하고 있다. 대부분의 경제학자들은 민간기업에 취업하거나 정부에 근무하거나 또는 대학에서 강의를 하면서 부분적으로 '경험적인' 경제분석에 종사하게 된다. 이를 통해 그들은 경제자료를 이용하여 경제적 관계를 규명하고 경제가설을 검정하여 경제적 결과를 예측하게 된다.

계량경제학을 배움으로써 '경제학과 학생'인 것과 '실질적인 경제학자'인 것 사이의 차이를 메울 수 있다. 계량경제학 소프트웨어 사용방법을 포함하여 이 책에서 배우게 될 계량경제학과 관련된 지식을 이용하여 독자들은 고용인의 질문에 다음과 같이 대답할 수 있을 것이다.

"나는 귀사 생산품의 판매액을 예측할 수 있다." "나는 경쟁으로 인해 단위당 1달러만큼 가격이 하락할 경우 귀사의 판매액에 미치는 영향을 평가할 수 있다." "나는 귀사의 새로운 광고

캠페인이 실제로 판매액을 증대시켰는지 여부를 검증할 수 있다." 이런 대답들은 경제학자처럼 생각하고 경제자료를 분석하는 귀하의 능력을 반영하는 것이므로 고용인의 귀에 감미로운 음악처럼 들릴 것이다. 고용인에게 유용한 정보를 제공할 수 있을 경우 귀하는 가치 있는 피고용인이 될 것이므로 바람직한 일자리를 얻을 수 있는 기회가 증대될 것이다.

반면에 대학원에 진학하여 공부를 계속하려는 경우라면 이 책이 정말로 가치가 있다는 사실을 깨닫게 될 것이다. 당신의 목표가 경제학, 재무관리, 회계학, 마케팅, 농업 경제학, 사회학, 정치학, 또는 임업학 분야에서 석사학위나 박사학위를 받는 것이라면 장래에 더욱 많이 계량경제학을 접하게 될 것이다. 대학원 과정은 매우 기술적이며 수학을 많이 사용하므로 나무를 공부하면서 자주 숲을 등한시하게 된다. 이 책을 공부함으로써 기술적인 것이 강조되는 과정으로 들어가기 전에 계량경제학의 내용에 관해 개관을 할 수 있고 이에 대한 '직관'을 살릴 수 있다.

1.2 계량경제학은 무엇을 가르치는가?

이 절에서는 계량경제학의 성격을 살펴볼 것이다. 계량경제학은 회계학, 사회학, 경제학 같은 독자들의 전공분야에서 중요한 변수들이 상호 어떻게 연계되는지에 관한 이론으로부터 출발한다. 경제학에서는 수학적인 함수개념을 이용하여 경제변수들 간의 관계를 나타낼 수 있다. 예를 들면, 소득과 소비의 관계를 다음과 같이 나타낼 수 있다.

$$\text{소비} = f(\text{소득})$$

위의 식에 따르면 소비 수준은 소득의 함수, 즉 $f(\cdot)$로 표현할 수 있다.

개별 상품, 예를 들면 현대 쏘나타에 대한 수요는 다음과 같이 나타낼 수 있다.

$$Q^d = f(P, P^s, P^c, INC)$$

위의 식에 따르면 현대 쏘나타에 대한 수요 Q^d는 현대 쏘나타의 가격 P, 대체재가 되는 자동차의 가격 P^s, 가솔린처럼 보완재가 되는 품목의 가격 P^c, 소득수준 INC의 함수, $f(P, P^s, P^c, INC)$가 된다.

쇠고기와 같은 농산품의 공급은 다음과 같이 나타낼 수 있다.

$$Q^s = f(P, P^c, P^f)$$

여기서 Q^s는 공급량, P는 쇠고기 가격, P^c는 경쟁관계에 있는 생산품의 가격(예를 들면, 돼지고기 가격), P^f는 생산과정에 사용되는 요소 또는 투입 물품의 가격(예를 들면, 옥수수 가격)을 나타낸다.

위의 식들은 경제변수들이 상호 간에 연계되는 방법을 보여 주도록 작성된 일반적인 경제모형이다. 이런 형태의 경제모형은 우리들이 경제분석을 가능하도록 하는 지표가 된다.

대부분의 경제적 결정이나 선택을 하는 데 있어 어떤 경제변수들이 상호 연계되어 있거나 연결관계의 방향을 아는 것만으로는 충분하지 않다. 이 밖에 우리는 관련된 변수들의 크기를 이해하여야만 한다. 즉, 한 변수의 변화가 다른 변수에 **얼마만큼**(how much) 영향을 미치는지 말할 수 있어야만 한다.

> **계량경제학**(econometrics)은 '얼마만큼'과 같은 종류의 물음에 답하기 위해 통계학적인 분석도구를 사용하여 경제학, 경영학, 사회과학과 관련된 이론과 자료를 어떻게 이용할 수 있는지에 관해 연구하는 분야이다.

1.2.1 실례

적절한 예를 들기 위해 중앙은행이 직면하게 되는 문제를 생각해 보자. 미국에서는 연방준비제도가 중앙은행에 해당하여 연방준비제도 이사회(FRB) 의장인 벤 버냉키(Ben Bernanke)가 이를 책임지고 있다. 가격 인상이 인플레이션율의 증가로 이어지는 시기에 연방준비제도 이사회는 경제의 성장률을 감속시킬지 여부를 결정하여야 한다. 연방준비제도 이사회는 은행들이 연방준비은행으로부터 차용할 경우 부과하는 이자율(할인율) 또는 은행 간에 하룻밤 동안 이루어지는 대출에 대해 부과되는 이자율(연방자금금리)을 인상함으로써 이를 달성할 수 있다. 이런 이자율이 인상될 경우 자본을 확대하기 위하여 자금을 구하려는 기업이나 자동차 또는 냉장고처럼 내구 소비재를 구입하려는 개인, 즉 잠재적인 투자자들이 직면하는 이자율이 인상된다. 소비자와 기업이 직면하는 이자율이 인상될 경우 내구재의 수요량이 감소하며 이는 총수요를 감소시켜 인플레이션율을 낮추게 된다. 이런 관계는 경제이론을 통해 알 수 있다.

연방준비제도 이사회 의장인 버냉키가 직면하는 실제문제는 "인플레이션을 낮추고 안정적으로 성장하는 경제 기조를 유지하기 위하여 얼마만큼 할인율을 인상하여야 하는가?"이다. 이 물음에 대한 대답은 이자율 인상과 투자 감소가 국민 총생산에 미치는 영향에 대한 기업과 개인의 반응도에 달려 있다. 주요한 탄력성 및 승수를 **모수**(parameter)라 한다. 경제 모수의 값은 알려져 있지 않으며 경제정책 수립 시 경제 자료의 표본을 이용하여 추정되어야 한다.

계량경제학은 주어진 자료를 이용하여 경제 모수를 어떻게 하면 가장 잘 추정할 수 있는지를 다룬다. 계량경제학을 발전시키고 올바르게 사용해야 한다. 왜냐하면 연방준비제도 이사회와 같은 정책입안자들이 사용하는 추정값이 잘못될 경우 이자율을 너무 대폭 또는 소폭으로 조절할 수 있으며 이는 우리 모두에게 중요한 결과를 초래할 수 있다.

정책입안자들은 매일 연방준비제도 이사회 의장인 버냉키가 직면하는 것과 유사한 '얼마만큼'이란 물음에 직면하게 된다. 이런 예는 다음과 같다.

- 시 위원회는 정규 경찰관을 거리에 배치하는 데 추가적으로 백만 달러를 사용할 경우 폭력범죄가 얼마만큼 감소할 것인지에 관해 심사숙고하고 있다.
- 미국 대통령 입후보자인 클린턴(Clinton)은 캘리포니아주에서 정치 광고를 하는 데 추가적으로 백만 달러를 사용할 경우 얼마나 많은 캘리포니아 유권자가 추가적으로 그를 지지할지 알고 싶어 한다.
- 어떤 지역의 피자헛 체인점의 소유주는 해당 지역 신문에 얼마만큼 광고를 하여야 하는지 결정해야 하며 이를 알아보기 위해 광고와 피자 판매 간의 관계를 추정해야만 한다.
- 미국 루이지애나 주립대학교 당국은 수업료가 학기당 100달러 인상될 경우 등록 학생수가 얼마나 감소하며 이로 인해 수업료 수입이 증가할 것인지 또는 감소할 것인지 추정해야만 한다.
- 미국의 프록터 앤드 갬블사의 최고경영자는 새로운 공장 및 시설에 얼마나 투자할지를 결정하려 할 때 자사 제품인 세탁용 세제 타이드에 대한 향후 10년 후의 수요량을 추정해야만 한다.
- 부동산 개발업자는 미국 루이지애나주 바톤루즈시의 남부지역에 인구 및 소득이 얼마나 증가할지를 예측해야만 새로운 쇼핑센터 건립 시 이윤이 남는지 여부를 알 수 있다.
- 귀하는 저축 중 얼마만큼을 주식형 펀드에 투자하고 화폐시장에 투자할지를 결정해야만 한다. 이를 위해 귀하는 향후의 경제활동 수준, 인플레이션율, 이자율을 예측해야만 한다.
- 샌프란시스코 베이 지역의 대중교통위원회는 통근열차(BART) 요금률이 인상될 경우, 버스, 통근열차를 이용하는 승객의 수에 어떤 영향을 미치는지 평가해야 한다.

위와 같은 '얼마만큼'에 대한 질문에 답하기 위해 정책입안자들은 실증적인 경제분석에 기초한 정보를 이용한다. 경제학자들은 이런 분석에서 문제가 되고 있는 변수들 간의 관계를 추론하기 위해 경제이론 및 경제논리를 사용한다. 중요한 근간이 되는 모수를 추정하고 예측하기 위해 관련 변수들에 대한 자료를 수집하고 경제적 방법론을 이용한다. 위의 예에서 정책입안자들은 상이한 방법을 이용하여 '추정값' 및 '예측값'을 구한다. 연방준비제도 이사회는 계량경제학적인 분석을 하기 위하여 많은 경제학자를 고용하고 있다. 프록터 앤드 갬블사의 최고

경영자는 판매액을 예측하기 위하여 계량경제학적인 분석을 담당할 상담역을 채용하려 한다. 여러분들은 주식 중개인으로부터 투자에 관한 조언을 받을 수 있으며 이들은 모회사에 근무하는 경제학자들이 제공하는 계량경제학적인 예측에 근거하고 있다. '얼마만큼'과 같은 물음에 대한 대답이 무엇에 근거를 하든지 대답할 수 있는 자료를 분석하는 계량경제학적인 방법을 경제학자가 이용하는 것은 바람직한 일이다.

다음 절에서는 모수를 경제모형에 어떻게 도입하고 경제모형을 계량경제모형으로 어떻게 전환시키는지를 살펴볼 것이다.

1.3 계량경제모형

계량경제모형은 무엇이며 어디에서 비롯되는 것일까? 이 절에서는 전반적인 견해와 독자들에게 익숙하지 않은 용어를 소개할 것이다. 진도를 더 나가기 전에 모든 용어들을 명확히 정의해 두어야 할 것이다. 먼저 계량경제모형에서는 경제적인 관계가 정확하지 않다는 점을 깨달아야만 할 것이다. 경제이론에 따르면 개인 또는 기업의 특정한 형태를 예측할 수 없으며 많은 개인 또는 기업의 평균적 또는 체계적 행태를 설명할 수 있을 뿐이다. 자동차 판매 대수를 검토할 경우 쏘나타의 실제 판매 대수는 위에서 말한 체계적 부분과 **무작위 오차**(rancom error)라고 부르는 무작위적이며 예측할 수 없는 부분, 즉 e로 구성된다. 따라서 현대 쏘나타의 판매 대수를 나타내는 **계량경제모형**(econometric model)은 다음과 같이 나타낼 수 있다.

$$Q^d = f(P,\ P^s,\ P^c,\ INC) + e$$

무작위 오차 e는 위의 모형에 포함되어 있지는 않지만 판매에 영향을 미치는 많은 요소를 의미하며 이는 경제활동에 내재된 본질적인 불확실성을 반영한다.

계량경제모형을 명시하기 위하여 경제변수들 간의 대수학적인 관계에 대해 언급할 필요가 있다. 예를 들어 여러분들은 경제원론 수업시간에 수요량이 가격에 대해 선형함수라는 점을 배웠을 것이다. 이 가정을 다른 변수들에게도 적용할 경우 수요의 체계적인 부분을 다음과 같이 나타낼 수 있다.

$$f(P,\ P^s,\ P^c,\ INC) = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC$$

이에 상응하는 계량경제모형은 다음과 같다.

$$Q^d = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC + e$$

함수 형태는 변수들 간의 관계에 관한 가설을 나타낸다. 특정 문제에서는 경제이론 및 자료에 부합되는 형태를 결정하는 데 노력을 기울이게 된다.

수요방정식, 공급방정식 또는 생산함수에 관계없이 모든 계량경제모형은 체계적인 부분과 관찰할 수 없는 무작위적인 부분으로 나누어진다. 체계적인 부분은 경제이론과 함수 형태에 관한 가정을 통해 설명할 수 있는 부분이다. 반면에 무작위적인 부분은 '교란' 부분으로 변수들 간의 관계를 이해하는 데 어려움을 주며 무작위 변수 e를 사용하여 나타낸다.

1.4 자료를 어떻게 구할 수 있는가?

자료를 어디서 구할 수 있는가? 경제학자들과 사회과학자들은 모든 변수에 관한 자료가 관찰될 뿐이지 통제된 실험을 통해 구할 수 없는 불확실한 세계에서 연구를 진행하게 된다. 이로 인해 경제모수에 관한 연구가 훨씬 더 어려워지게 된다. 경제적으로 중요한 물음에 답하기 위해 이런 자료들을 사용하는 절차에 대해 알아보는 것이 이 책의 주요한 목적이다. 다음 절에서 실험적 자료와 비실험적 자료의 특성을 명확히 알아볼 것이다.

1.4.1 실험적 자료

경제 관계를 나타내는 알지 못하는 모수에 관해 정보를 얻을 수 있는 한 가지 방법은 실험결과를 정리하거나 관찰하는 것이다. 자연과학 및 농업에서는 통제된 실험을 쉽게 가상해 볼 수 있다. 과학자들은 주요 통제 변수의 값을 정하고 나서 결과를 관찰할 수 있다. 유사한 토지 구획에 특정 품종의 밀을 심고 나서 각 토지 구획에 대해 비료 및 살충제의 양을 상이하게 사용한 후 추수시점에 각 구획에서 생산된 밀의 양을 관찰할 수 있다. N개 토지 구획에서 실험을 반복하면 N개 관찰값의 표본을 만들 수 있다. 이런 통제된 실험은 경영이나 사회과학 분야에서 거의 이루어지지 않는다. 실험적 자료의 주요한 특징은 설명변수들의 값이 반복적으로 이루어지는 실험에서 특정값으로 고정된다는 점이다.

경영적인 예로는 마케팅 연구에서 찾아볼 수 있다. 슈퍼마켓에서 이루어지는 특정 상품의 주당 판매에 관심이 있다고 가상하자. 한 상품의 판매가 이루어질 경우 이 상품은 계산대에서 스캔이 되면 계산서에 나타나게 될 가격과 총액이 기록된다. 이와 동시에 자료 기록이 이루어지며 각 시점에서 현재의 상점진열과 쿠폰 사용뿐만 아니라 해당 상품의 가격과 모든 경쟁 상점의 가격을 알 수 있다. 가격과 쇼핑환경은 상점관리를 통해 통제됨으로 이런 '실험'은 '통제' 변수의 값을 동일하게 하여 여러 날 또는 여러 주에 걸쳐 반복적으로 이루어질 수 있다.

1.4.2 비실험적 자료

비실험적 자료의 예로는 조사자료를 들 수 있다. 미국 루이지애나 주립대학교에 소재하는 공공정책 연구소는 고객에 대한 전화 및 우편조사를 시행하였다. 전화조사에서 대상 인원은 무작위적으로 뽑아 전화 통화를 하였다. 질문에 대한 대답은 기록되어 분석이 이루어졌다. 이런 환경에서 모든 변수에 대한 자료는 동시에 수집되며 그 값은 고정되지도 반복되지도 않는다. 이런 자료는 비실험적 자료이다.

수집된 자료는 몇 가지 형태를 가질 수 있다. 자료를 분석할 경우 수집된 형태에 따라 해결해야 할 방법론상의 문제를 갖게 된다. 자료는 다음과 같은 형태를 갖게 된다.

- **시계열 형태**–일정 기간에 걸쳐 수집된 자료. 예를 들면 1880년부터 2007년까지 미국 밀의 연도별 가격 또는 1980년부터 2007년까지 제너럴 일렉트릭사 주식의 일별가격을 들 수 있다.
- **횡단 형태**–특정시기에 표본 단위별로 수집된 자료, 예를 들면 2006년 캘리포니아주 군별 소득 또는 2006년 미국의 주별 고등학교 졸업률이 있다.
- **패널자료 형태**–일정 기간에 걸쳐 개별적인 미소한 단위별로 계속 수집된 자료. 예를 들면 미국 교육부는 몇 가지 조사를 계속적으로 시행하는데, 그중에는 동일한 학생이 8학년부터 20대 중반까지 어떤 변화를 겪는지 계속 조사하는 것이 있다. 이 데이터베이스는 학생의 특성과 성적뿐만 아니라 학생이 속한 가족, 학교, 기타의 사회경제적 특성을 기록한다. 이것들은 노동경제학, 가정경제학, 보건경제학, 교과교육과 관련된 연구에 풍부한 자료를 제공해 주고 있다.

자료는 다양한 단계별로 통합되어 수집될 수 있다.

- **미시**–개인, 가계, 기업처럼 개별적인 경제 의사결정 단위별로 수집한 자료
- **거시**–지방, 주, 또는 국가 수준에서 개인, 가계, 기업을 하나로 모으거나 통합하여 수집한 자료

수집된 자료는 유량 또는 저량이 될 수 있다.

- **유량**–2006년 마지막 분기 동안의 가솔린 소비량처럼 일정기간 동안 측정된 결과
- **저량**–2006년 4월 1일 현재 쉐브론사의 미국 소재 저유탱크에서 저장된 원유량 또는 2007년 7월 1일 현재 미국 웰스 파고(Wells Fargo)은행의 자산가치처럼 특정 시점에서 측정된 결과

수집된 자료는 양적 또는 질적일 수 있다.

- **양적인 자료**–숫자 또는 숫자를 변형시켜 표현할 수 있는 가격 또는 소득과 같은 자료. 예

를 들면 실질가격 또는 1인당 소득을 들 수 있다.

- 질적인 자료– '어느 한쪽'을 선택하여 상황을 나타내는 자료, 예를 들면 소비자가 특정 상품을 구입하였는가 또는 하지 않았는가, 사람이 결혼을 하는가 또는 하지 않는가 등이 있다.

제17장에서는 인쇄물, 인터넷 전자매체를 통해 구할 수 있는 경제자료의 출처를 소개할 것이다.

1.5 통계적 추론

통계적 추론(statistical inference)이란 용어가 이 책에서 자주 등장할 것이다. 이는 자료의 표본을 분석하여 실제 세계에 대한 것을 '추론'하거나 알고자 한다는 의미이다. 통계적 추론을 하는 방법에는 다음과 같은 것들이 있다.

- 계량경제 방법을 이용하여 탄력성과 같은 경제모수의 추정
- 향후 10년 동안 미국의 2년제 대학에 등록할 학생수와 같은 경제적 결과의 예측
- 판매를 증가시키기 위해 상점에 진열하기보다 신문광고를 하는 것이 더 나은지에 관한 물음과 같은 경제적 가설의 검정

계량경제학은 위의 통계적 추론에 관한 모든 면을 포함하고 있으므로 이 책을 공부하게 되면 분석하고자 하는 자료의 특성에 따라 어떻게 적절히 추정, 예측, 검정하는지를 배우게 될 것이다.

1.6 연구 체계

실증적 경제연구는 일정한 형식을 따르므로 이 책 전반에 걸쳐 순서에 따른 절차를 강조할 것이다. 그 절차는 다음과 같다.

1. 연구는 모두 문제점이나 의문점에서 출발한다. 사고의 틀은 특정 주제에 관한 모든 것을 장기간에 걸쳐 검토한 후에 형성된다. 여러분들은 "영감이 99%의 피나는 노력에 의해 얻어진다"는 사실을 깨닫게 될 것이다. 이는 한 주제에 관해 장시간의 투자를 한 후에야 새롭고 흥미로운 의문점이 떠오른다는 의미이다. 다른 방법은 흥미로운 의문점을 찾기 위해 자연적인 호기심에 의존하는 것이다. 할 베어리언(Hal Varian)은 자신의 논문 "How to Build an Economic Model in Your Spare Time," *The American Economist*, 41(2), Fall

1997, pp. 3-10에서 학술적인 잡지 밖에서, 예를 들면 신문, 잡지 등에서 새로운 사고를 찾아보도록 권하고 있다. 그는 새로운 텔레비전 수상기를 구입하는 데서 비롯된 연구 주제와 관련지어 이를 논의하였다.

2. 경제이론은 문제에 관해 생각할 수 있는 방법을 제시해 준다. 즉, 관련된 경제변수들은 무엇이며 이들 간의 관계들은 어떤지에 관해 알려 준다. 최초의 질문이 주어지면 모든 연구계획은 경제모형을 수립하고 관심이 있는 의문점(가설)을 나열하는 데서 시작된다. 연구계획을 진행시키는 도중에 더 많은 의문점이 제기될 수도 있지만 초기에 연구의 동기가 되었던 의문점들을 나열하는 것이 좋다.
3. 활용할 수 있는 경제모형을 이용하여 계량경제모형을 설정할 수 있다. 함수 형태를 선택하고 오차항의 성질에 관해 가정을 수립해야 한다.
4. 표본자료를 수집하고 나서 초기의 가정과 자료 수집 방법을 고려하여 적절한 계량경제분석방법을 선택한다.
5. 통계적인 소프트웨어 패키지를 이용하여 미지의 모수를 추정하고 예측을 하며 가설을 검정한다.
6. 설정한 가정의 타당성을 점검하기 위하여 모형을 진단해야 한다. 예를 들어 오른편의 모든 설명변수가 모형과 관련이 있는지를 점검하고 올바른 함수 형태가 사용되었는지를 알아보아야 한다.
7. 실증분석 결과가 갖는 경제적 중요성과 의미를 평가해야 한다. 경제적 자원 분배 및 분산에 대해 갖는 의미와 선택한 정책이 갖는 의미를 알아본다. 더 많은 연구를 하거나 새롭고 더 나은 자료를 구할 경우 대답할 수 있는 질문에는 어떤 것이 있는지 검토한다.

앞으로 이 책은 이런 방향으로 전개될 것이다. 연구가 끝난 후에는 발견한 사실과 이용한 방법을 요약한 논문 또는 보고서를 작성해야 한다. 경제학 논문을 작성하는 지침은 제17장에서 소개할 것이다.

제 2 장

단순 선형회귀 모형

학습 목표 | 이 장에서 다룰 내용에 기초하여 다음과 같이 할 수 있어야 한다.

1. 추정량과 추정값의 차이를 설명하고 최소제곱 추정량은 확률변수이지만 최소제곱 추정값은 확률변수가 아닌 이유를 설명하시오.
2. 단순회귀 모형의 기울기 및 절편 모수에 대한 해석에 관해 논의하고 추정 방정식을 그래프로 나타내시오.
3. 관찰할 수 있는 변수 y를 체계적인 부분과 무작위적인 부분으로 분리하는 문제를 이론적으로 설명하고 도표를 통해 이를 나타내시오.
4. 단순 선형회귀 모형의 가정 각각에 대해 논의하고 설명하시오.
5. 자료의 산포도상에 선을 긋기 위하여 최소제곱원칙을 어떻게 사용할 수 있는지 설명하시오. 최소제곱 잔차와 종속변수의 최소제곱에 적합한 값을 정의하고 이를 도표로 나타낼 수 있어야 한다.
6. x에 대한 y의 탄력성을 정의하고 y 및 x가 어떤 방법으로든 변형되지 않는 경우 단순 선형회귀 모형상에서 이들을 산정한 값을 설명하시오. y 및 x 모두에 자연대수를 취하여 변형시킨 단순 선형회귀 모형에서 x에 대한 y의 탄력성을 어떻게 계산하는지 설명하시오.
7. "회귀모형 가정 SR1−SR5가 준수될 경우 최소제곱 추정량 b_2가 불편된다."는 말의 의미를 설명하시오. 특히 '불편'은 정확히 무엇을 의미하는가? 중요한 변수가 모형에서 빠지는 경우 b_2가 왜 편의되는지 설명하시오.
8. '표본추출 변동성'이란 문귀가 의미하는 바를 설명하시오.
9. σ^2, $\sum(x_i - \bar{x})^2$, N의 요소들이 미지의 모수 β_2를 추정하는 정확성에 어떤 영향을 미치는지 설명하시오.
10. 가우스−마코프 정리를 정의하고 설명하시오.

경제이론에 따르면 경제변수 간에는 여러 관계가 있다. 미시경제학에서는 물품의 수요량 및 공급량이 가격에 의존한다는 수요 및 공급모형을 생각해 볼 수 있다. 또한 생산량이 사용되는 생산요소, 예를 들면 노동량의 함수로 설명하는 '생산함수와 총 생산물 곡선'을 들 수 있다. 거시경제학에서는 경제의 총투자량이 이자율에 의존한다는 '투자함수'와 총소비를 가처분 소득 수준에 연계시키는 '소비함수'를 생각해 볼 수 있다.

각 모형들은 경제변수들 간의 관계를 포함한다. 이 장에서는 이런 관계들에 관해 알아보기 위해 경제자료의 표본을 어떻게 사용할 수 있는지 살펴보고자 한다. 경제학자로서 우리들은 다음과 같은 질문에 관심을 갖는다. 한 변수(예를 들면, 물품의 가격)가 변화할 경우 다른 변수(예를 들면, 수요 또는 공급량)는 얼마만큼 변하는가? 또한 한 변수의 값을 아는 경우 다른 변수의 이에 상응하는 값을 예측하거나 예상할 수 있는가? 이 장에서는 **회귀모형**(regression model)을 이용하여 위의 물음들에 답할 것이다. 모든 모형들처럼 회귀모형도 가정에 기초하고 있다. 이 장에서는 이런 가정들을 분명히 해 두고자 한다. 왜냐하면 이 가정들은 다음 장들에서 살펴볼 분석들이 적절해지도록 하는 조건이 되기 때문이다.

2.1 경제모형

회귀모형에 관한 개념을 도출하기 위해 단순하지만 중요한 경제적인 예를 들어 볼 것이다. 가계소득과 식료품에 대한 지출 사이의 관계를 고찰하는 데 관심이 있다고 가상해 보자. 특정 모집단으로부터 가계를 무작위적으로 추출하는 '실험'을 한다고 생각해 보자. 모집단은 특정 도시, 주, 지방, 또는 국가의 가계로 구성된다. 지금은 가계소득이 주당 $1,000인 가계에만 관심이 있다고 하자. 이 실험에서는 모집단으로부터 많은 가계를 무작위적으로 뽑아 이들과 설문조사를 하게 된다. 우리의 관심사는 식료품에 대한 해당 가계의 주당 지출액이므로 다음과 같은 질문을 할 것이다. "귀하의 가계는 지난주에 얼마나 많이 식료품에 지출을 하였습니까?" y라고 나타낼 주당 식료품 지출액은 확률변수이다. 왜냐하면 가계를 뽑아서 위와 같은 질문을 하고 이에 대답을 할 때까지는 그 값을 알 수 없기 때문이다.

> **유의사항**
>
> 제0장에서 확률변수는 대문자(Y)로 그의 값은 소문자(y)로 나타내어 확률변수와 이의 값을 구별하였다. 이런 구별은 받아들이기 어려울 정도로 복잡한 표시법이 될 수 있으므로 더 이상 이를 따르지 않을 것이다. y를 사용하여 확률변수의 값뿐만 아니라 확률

변수 자체를 나타낼 것이며 상황별로 이에 대한 해석을 명확히 할 것이다.

연속적 확률변수 y는 다양한 식료품 지출액을 구하게 될 확률을 나타내는 **확률밀도함수**(이를 요약해서 *pdf*라고 하자) $f(y)$를 갖는다. 식료품에 지출하는 1인당 총액은 여러 가지 이유로 인해 가계마다 명백히 다르다. 어떤 가계는 미식가가 좋아하는 음식을 많이 구입하고 다른 가계에는 십대가 같이 살며 또 다른 가계에는 노인이 있고 일부 가계는 채식주의자일 수도 있다. 위의 이런 요소들과 무작위적이며 충동적인 구매를 포함한 많은 다른 요소들로 인해 소득수준이 같음에도 불구하고 식료품에 대한 주당 지출액은 가계마다 다르다. 이 경우 확률밀도함수는 지출이 모집단에서 어떻게 '분포되어' 있는지 알려 주며 이는 그림 2.1의 분포 중 하나와 같을 수 있다.

그림 2.1a의 확률분포는 가계소득에 대해 '조건부'이므로 실제로 조건부 확률밀도함수이다. x = 주당 가계소득인 경우 조건부 확률밀도함수는 $f(y|x = \$1{,}000)$이다. y의 조건부 평균 또는 기댓값은 $E(y|x = \$1{,}000) = \mu_{y|x}$이며 이는 모집단의 1인당 평균 주당 식료품 지출액이 된다. y의 조건부 분산은 $\text{var}(y|x = \$1{,}000) = \sigma^2$이며 이는 평균 $\mu_{y|x}$에 대한 가계지출 y의 퍼진 정도를 나타낸다. 모수 $\mu_{y|x}$와 σ^2를 알 수 있는 경우 해당 모집단에 관한 가치 있는 정보를 알게 된다. 이런 모수들을 알고 조건부 분산 $f(y|x = \$1{,}000)$이 정규, 즉 $N(\mu_{y|x}, \sigma^2)$인 경우 정규분포의 특성을 이용하여 y가 특정 구간에 속할 확률을 계산할 수 있다. 즉, 주당 소득이 \$1,000인 경우 식료품에 대한 1인당 지출액이 \$50에서 \$75 사이인 가계인구의 비율을 계산할 수 있다.

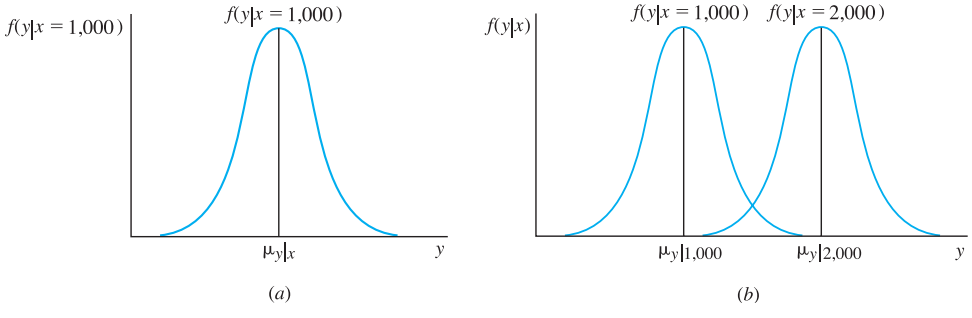


그림 2.1

(a) 소득이 x = \$1,000인 경우 식료품 지출액이 y의 확률분포 $f(y|x = 1{,}000)$, (b) 소득이 x = \$1,000 및 x = \$2,000인 경우 식료품 지출액 y의 확률분포

유의사항

확률변수의 기대 값을 '평균'값이라고 하는데 이는 실제로 확률변수의 확률분포 중앙에 위치하는 **모평균**을 축약한 것이다. 이는 표본의 숫자 값을 산술적으로 평균한 **표본평균**과 동일하지 않다. 이처럼 '평균'이란 용어가 두 가지 용도로 사용되는 차이점에 유의하자.

경제학자로서 우리는 보통 변수들 사이의 관계, 여기서는 y = 주당 식료품 지출액과 x = 주당 가계소득 사이의 관계를 연구하는 데 관심을 갖는다. 경제이론에 따르면 경제 재화에 대한 지출은 소득에 의존한다고 하지만 경제학 수업시간에 지출이 확률변수라는 사실은 아마도 언급하지 않았을 것이다. 우리의 문제는 자료를 사용하여 확률변수에 관한 정보를 얻는 일이다.

지출관계에 대한 계량경제학적 분석을 통해 다음과 같은 중요한 질문에 대답을 할 수 있다. 주당 소득이 $100만큼 증가한 경우 주당 평균 식료품 지출액은 얼마나 증대할 것인가? 그렇지 않으면 소득이 증가함에 따라 주당 식료품 지출액이 감소할 수 있는가? 주당 소득이 $2,000인 가계의 경우 주당 식료품 지출액이 얼마인지 예측할 수 있는가? 위의 물음에 대한 대답들은 정책입안자들에게 가치 있는 정보를 제공하게 된다.

> 1인당 식료품 지출액에 관한 정보를 이용하여 규모, 인종, 소득, 지리적 위치, 기타 사회경제적 및 인구통계적으로 상이한 가계들의 지출습관상의 유사점 및 차이점을 결정할 수 있다. 이런 정보는 시장의 현재상황, 상품 유통구조, 소비자 구매 습관, 소비자 생활 상황을 평가하는 데 가치 있는 자료가 된다. 이런 정보는 인구 통계 및 소득에 관한 예측치와 결합하여 소비 추세를 예상하는 데 사용될 수 있다. 또한 이는 예를 들면 노년층과 같이 특정 인구집단에 대한 일반적인 식료품 소비 형태를 알아보는 데도 사용된다. 이런 소비형태는 거꾸로 해당 인구집단의 소비행태에 적합한 물가지수를 개발하는 데 이용될 수 있다. [Blisard, Noel, Food Spending in American Households, 1997－1998, Electronic Report from the Economic Research Service, U.S. Department of Agriculture, Statistical Bulletin Number 972, June 2001]

예를 들어 우리가 슈퍼마켓 연쇄점의 관리인이며 장기계획 수립에 대해 책임을 지고 있다고 가상하자. 경제 예측에 따르면 해당 지역의 소득이 향후 몇 년에 걸쳐 증가할 것으로 보이는 경우 고객을 맞기 위해 시설을 확장할지 여부와 얼마나 확장할지를 결정해야만 한다. 또는 한 슈퍼마켓 연쇄점은 고소득 지역에 위치하고 다른 슈퍼마켓 연쇄점은 저소득 지역에 있다면 소득수준이 다른 경우의 식료품 지출액에 대한 예측은 해당 지역의 슈퍼마켓이 얼마나 커야 하는지를 결정하는 데 중요한 역할을 한다.

지출과 소득 사이의 관계를 조사하기 위해 우선 **경제모형**(economic model)을 수립하고 나서

수량적인 경제분석의 기초가 될 **계량경제모형**(econometric model)을 만들어야 한다. 식료품 지출에 대한 위의 예에서 경제이론에 따르면 조건부 평균 $E(y|x) = \mu_{y|x}$로 나타낸 식료품에 대한 주당 평균 가계 지출액은 가계소득 x에 의존한다. 소득수준이 다른 가계들을 고려할 경우 식료품에 대한 평균 지출액이 변할 것으로 기대된다. 그림 2.1b에서 주당 소득수준이 서로 다른 2개 소득, 즉 \$1,000 및 \$2,000의 식료품 지출액에 대한 확률밀도함수를 살펴보았다. 각 밀도함수 $f(y|x)$에 따르면 지출이 평균값 $\mu_{y|x}$를 중심으로 분포하겠지만 소득수준이 높은 가계의 평균 지출액이 소득수준이 낮은 가계의 평균 지출액보다 크다.

대부분의 경제학 교과서에서 소비를 소득에 연계시키는 소비 또는 지출함수는 선형함수로 나타내며 여기서도 동일한 가정을 할 것이다. 그러나 이는 단순히 가정일 뿐이라는 사실을 기억해야 한다. 그림 2.2에서 보는 것처럼 가계의 식료품 지출액에 관한 우리의 경제모형은 다음과 같다.

$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \tag{2.1}$$

(2.1)에 있는 조건부 평균 $E(y|x)$를 **단순 회귀함수**(simple regression function)라 한다. 이는 쉽기 때문이 아니라 식의 오른편에 단지 하나의 경제변수만이 있기 때문에 단순 회귀라 한다. 미지의 **회귀모수**(regression parameter) β_1 및 β_2는 각각 회귀함수의 절편 및 기울기이다. 위의 식료품 지출액 예에서 절편 β_1은 소득이 없는, 즉 x = \$0인 가계가 식료품에 대해 주당 평균적으로 지출하는 금액을 의미한다. 기울기 β_2는 주당 소득이 \$1 변화할 경우 $E(y|x)$의 변화를 의미하며 식료품에 대한 한계 지출성향이라 한다. 이를 대수학적으로 나타내면 다음과 같다.

$$\beta_2 = \frac{\Delta E(y|x)}{\Delta x} = \frac{dE(y|x)}{dx} \tag{2.2}$$

위에서 Δ는 '변화'를 나타내며 $dE(y|x)/dx$는 x에 대한 $E(y|x)$의 '도함수'를 나타낸다. 이 책에

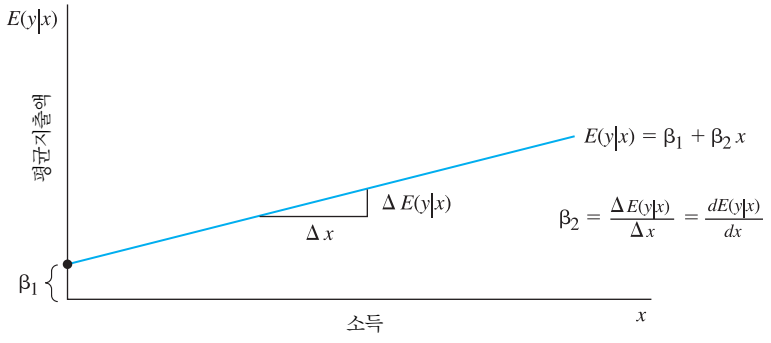


그림 2.2
경제모형 : 1인당 평균 식료품 지출액과 소득 사이의 선형관계

서는 대부분 도함수를 사용하지 않을 것이며 이 개념에 익숙하지 않거나 기억하지 못하는 경우 d를 Δ의 '형식화된' 변형이라 생각하고 진도를 계속 나가도 무방하다.

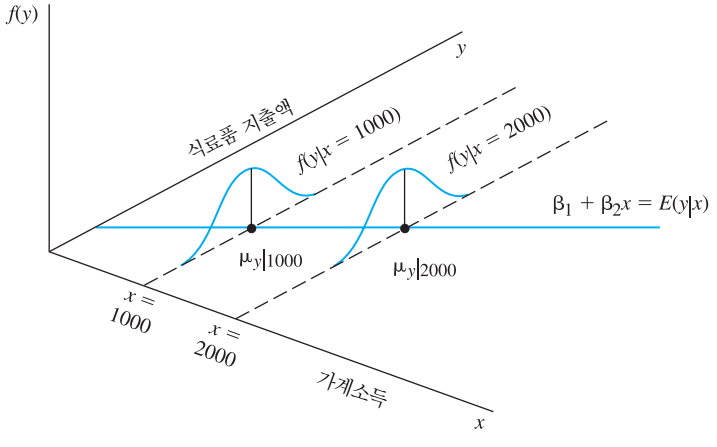
경제모형(2.1)은 이론에 기초하여 가계소득(x)과 식료품에 대한 가계의 평균 지출액 $E(y|x)$ 사이의 관계에 대한 논의를 요약하고 있다. 이 모형의 모수 β_1 및 β_2는 경제행태를 특징 짓고 경제적 결정을 내리는 데 기초가 되는 수량이다. 자료를 사용하기 위해서는 가계소득과 지출에 관한 자료를 어떻게 구하는지를 설명하고 계량경제 분석을 가능하게 하는 계량경제모형을 설정하여야 한다.

2.2 계량경제모형

모형 $E(y|x) = \beta_1 + \beta_2 x$는 경제행태를 설명하고 있기는 하지만 현실과 다른 추상적 개념이다. 주당 소득이 $x = \$1{,}000$인 가계에 관해 무작위 표본을 추출할 경우 실제 지출액은 그림 2.1에서 보는 것처럼 평균 또는 평균값 $E(y|x = 1{,}000) = \mu_{y|x=1{,}000} = \beta_1 + \beta_2(1{,}000)$을 중심으로 분포되어 있음을 알 수 있다. 다양한 소득수준에 대해 가계 지출금액을 표본추출할 경우 표본값들은 평균값 $E(y|x) = \beta_1 + \beta_2 x$를 중심으로 체계적으로 분포될 것이라 생각된다. 그림 2.3에서는 각 소득수준에 대한 회귀선에 관해 그림 2.1과 같은 종모양의 형상을 보여 주고 있으며 이는 식료품 지출액의 확률밀도함수인 $f(y|x)$를 나타낸다.

그림 2.3에서 각 소득수준에 대한 평균 가계 지출액은 회귀함수 $E(y|x) = \beta_1 + \beta_2 x$로 나타낼 수 있음을 알 수 있다. 이 그림은 또한 식료품에 대한 가계 지출액의 금액이 각 소득수준에서 평균값 $E(y|x) = \beta_1 + \beta_2 x$를 중심으로 분산되어 있음을 가정하고 있다. 회귀함수는 가계

그림 2.3
2개의 소득수준에서 y에 대한 확률밀도함수



식료품 지출액에 대한 계량경제모형의 기초가 된다.

계량경제모형을 완벽하게 만들기 위해서는 몇 가지 가정을 추가하여야 한다.

유의사항

이 장과 이 책의 나머지 부분에서 **가정**(assumption)에 관해 많은 논의를 할 것이다. 가정은 "만일 …이라면, …할 것이다"라는 문구에서 "만일 …이라면"에 해당하는 부분이다. 그리고 중요한 점은 가정이 준수되지 않을 경우 도출한 결론도 준수될 수 없다는 것이다. 계량경제학적 분석을 하면서 직면하게 되는 문제 중 하나는 현실적인 가정을 수립하고 이들이 준수되는지를 점검하는 것이다.

그림 2.1a에서 자신의 평균에 대한 y값의 퍼진 정도를 $\text{var}(y|x = \$1{,}000) = \sigma^2$라고 하였다. 각 소득수준에 대해서도 퍼진 정도에 관해 이와 유사한 가정을 해야 한다. 기준이 되는 가정은 자신의 평균에 대한 y값의 퍼진 정도가 모든 소득수준 x에 대해 동일하다는 것이다. 즉 모든 x값에 대해 $\text{var}(y|x) = \sigma^2$가 되어야 한다. 그림 2.1b에서 보면 소득이 서로 다른 두 가계에 의한 식료품 지출의 확률밀도함수에서 평균은 서로 상이하지만 분산은 동일하다. 이 가정은 그림 2.3에서도 설명할 수 있다. 즉, 그림 2.1처럼 각 분산의 '퍼진 정도'를 동일하게 나타내고 있다.

분산이 일정하다는 가정, 즉 $\text{var}(y|x) = \sigma^2$은 각 소득수준 x에서 식료품 지출액 y가 평균값 $E(y|x) = \beta_1 + \beta_2 x$로부터 얼마나 멀리 떨어지는지 동일하게 불확실하며 그 불확실성은 소득 또는 그 밖의 어떤 것에도 의존하지 않는다는 의미이다. 이 조건을 만족시키는 자료를 **동분산적**(homoskedastic)이라고 한다. 이 가정이 위배되어서 소득의 모든 값 x에 대해 $\text{var}(y|x) \neq \sigma^2$인 경우 해당 자료를 **이분산적**(heteroskedastic)이라고 한다.

다음으로 표본이 무작위적이어야 한다. 이것은 자료를 수집할 경우 이들이 통계적으로 독립적이라는 의미이다. y_i 및 y_j가 무작위적으로 뽑은 두 가계의 지출액을 의미하는 경우 이들 (확률)변수 중 하나의 값을 알더라도 다른 변수가 어떤 값을 취할지에 대해 알 수 없다.

수학자들은 점점 더 약한 일련의 가정하에서 동일한 이론을 증명하기 위해 (약간 과장해서 말하면) 전 생애를 바친다. 이런 사고방식은 어느 정도 계량경제학자들에게도 적용된다. 따라서 계량경제모형들은 통계적인 독립성보다는 약하지만 앞으로의 몇몇 장에서 증명하고자 하는 것들을 증명하는 데는 충분한 가정을 할 것이다. y_i 및 y_j가 무작위적으로 뽑은 2개 가계의 지출액인 경우 이들의 공분산은 0이거나 또는 $\text{cov}(y_i, y_j) = 0$이라고 가정한다. 이는 통계적인 독립성보다 약한 가정이다. (왜냐하면 통계적인 독립성은 0의 공분산을 의미하지만 이의 역은 성립하지 않기 때문이다.) 0의 공분산은 y_i 및 y_j 사이의 체계적인 선형관계가 없음을 의미할 뿐

이다.

회귀분석을 하기 위해서는 변수 값 x에 관한 가정을 해야 한다. 회귀분석의 기본적인 사고는 한 변수 x의 변화가 다른 변수 y에 미치는 영향을 측정하는 것이며 이를 위해서는 x가 표본자료에서 몇 개의 다른 값, 우리가 든 예에서는 최소한 2개의 값을 가져야 한다. 표본에서 x의 모든 관찰값이 동일한 값, 예를 들면 $x = \$1{,}000$인 경우 회귀분석을 할 수 없다. 나아가 x값은 주어진 것이며 확률적이 아니라고 가정할 것이다. 도출한 모든 결과는 주어진 x값에 대해 조건부적이다. 이 가정에 대해서는 이후에 보다 자세히 논의할 것이다.

마지막으로 y값은 정규분포되는 것으로 가끔 가정한다. 이는 '종모양'의 곡선이 실제로 지능지수에서부터 옥수수 줄기의 길이, 호주 남자아이들의 출생 시 무게에 이르기까지 많은 유형적인 현상을 설명할 수 있다는 관찰에 기초하고 있다. 경제변수는 평균을 중심으로 정규분포된다고 가정하는 것이 가끔은 합리적이지만 지금은 '선택적'인 가정으로 볼 것이다. 왜냐하면 많은 경우에 이런 가정이 필요하지 않고 만일 할 경우 매우 강한 가정이 된다.

위에서 논의한 사항을 한데 모아서 **계량경제모형**(econometric model)을 정의할 것이며 이는 자료를 설명하는 일련의 가정이 된다. 다음의 요약은 이를 논리적으로 모아 놓은 것이다.

단순 선형회귀 모형에 관한 가정 – I

- x의 각 값에 대해 y의 평균값은 선형회귀로 나타낼 수 있다.

$$E(y|x) = \beta_1 + \beta_2 x$$

- x의 각 값에 대해 y의 값은 자신의 평균값을 중심으로 동일한 분산을 갖는 확률분포에 따라 분포된다.

$$\text{var}(y|x) = \sigma^2$$

- y의 표본값은 모두 비상관되며 0의 공분산을 갖는다. 이는 y값들 사이에 선형관계가 없음을 의미한다.

$$\text{cov}(y_i, y_j) = 0$$

  위의 가정은 y값이 모두 통계적으로 독립적이라고 할 경우 더욱 강한 가정이 된다.
- 변수 x는 확률적이지 않으며 최소한 2개의 상이한 값을 가져야 된다.
- (선택적) y값은 x의 각 값에 대해 자신의 평균을 중심으로 정규분포된다.

$$y \sim N\left[(\beta_1 + \beta_2 x), \sigma^2\right]$$

2.2.1 오차항의 도입

회귀모형에서 일반적으로 **종속변수**(dependent variable)라고 불리는 y의 측면에서 단순 선형회귀 모형에 대한 가정을 설명하는 것이 편리하다. 그러나 통계적인 목적을 위해서는 이 가정들을 다른 방법으로 설명하는 것도 유용하다.

회귀분석의 본질은 종속변수 y에 대한 관찰이 두 부분, 즉 체계적 요소와 무작위적 요소로 양분된다는 점이다. y의 체계적 요소는 자신의 평균 $E(y|x) = \beta_1 + \beta_2 x$이며 이는 수학적 기대이므로 무작위적일 수 없다. y의 무작위적 요소는 y와 이의 조건부 평균값 $E(y|x)$의 차이이다. 이를 **무작위 오차항**(random error term)이라 하며 다음과 같이 정의한다.

$$e = y - E(y|x) = y - \beta_1 - \beta_2 x \tag{2.3}$$

(2.3)을 재정리하면 다음과 같은 **단순 선형회귀 모형**(simple linear regression model)을 구할 수 있다.

$$y = \beta_1 + \beta_2 x + e \tag{2.4}$$

종속변수 y는 **독립변수**(independent variable) x와 함께 체계적으로 변화하는 요소와 무작위적 오차항 e로 설명할 수 있다.

식 (2.3)에 따르면 y와 오차항 e는 단지 $E(y|x) = \beta_1 + \beta_2 x$만큼만 차이가 나며 이는 무작위적이 아니다. y가 무작위적이므로 오차항 e도 무작위적이 된다. y에 관해 이미 가정을 하였으므로 오차항 e에 관한 특성을 식 (2.3)에서 직접 도출할 수 있다.

x가 주어진 경우 오차항의 기댓값은 다음과 같다.

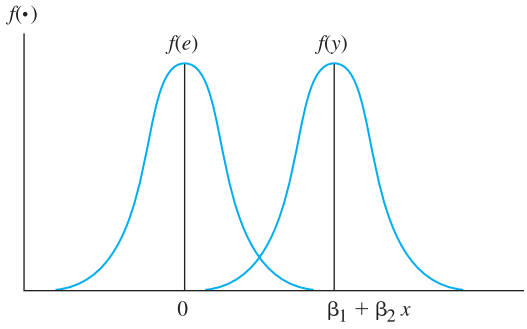
$$E(e|x) = E(y|x) - \beta_1 - \beta_2 x = 0$$

x가 주어진 경우 오차항의 평균값은 0이 된다.

y와 e는 상수(즉, 무작위적이지 않은 요소)만큼만 차이가 나므로 이들의 분산은 동일하고 σ^2이 되어야만 한다. 따라서 y 및 e의 확률밀도함수는 그림 2.4에서 보는 것처럼 위치를 빼고는 동일하다. 오차항의 확률밀도함수 $f(e)$의 중앙은 0으로 이는 기댓값, $E(e|x) = 0$이라는 점에 주목하자.

이제는 x가 확률적이 아니라는 단순화된 가정을 좀 더 논의할 수 있다. x가 확률적이 아니라는 가정은 그 값이 알려져 있다는 의미이다. 통계학에서 이와 같은 x값을 '반복된 표본에서 고정되어 있다'고 한다. 앞에서 살펴본 것처럼 통제된 실험을 할 경우 일련의 동일한 x값이 반복해서 사용될 수 있으며 y만이 확률적이 된다. 예를 들면 빅맥 가격의 변화가 근처 맥도날드에서

그림 2.4
e 및 y의 확률밀도함수



주당 판매되는 빅맥의 수에 어떤 영향을 미치는지 알고자 한다고 가정하자. 해당 점포의 소유주는 가격(x)을 설정하고 해당 주 동안에 판매되는 빅맥의 수(y)를 관찰할 수 있다. 그 다음 주에 가격을 변화시키고 판매에 관한 자료를 다시 수집할 수 있다. 이 경우 x = 빅맥의 가격은 확률적이 아니라 고정되어 있다.

경영학 및 경제학에서 x값이 고정되어 있는 경우는 드물다. 가계를 조사할 경우 1인당 식료품 지출액과 가계소득과 같은 변수에 관한 자료를 동시에 얻게 된다. 따라서 이런 경우 y 및 x는 둘 다 확률적이며 실제로 관찰될 때까지 알지 못한다. 하지만 x가 주어지며 확률적이 아니라고 가정을 하더라도 다음 장에서 논의하게 될 결과를 변화시키지는 않는다. 이런 가정을 함으로써 얻게 되는 추가적인 이점은 표기상의 단순성이다. x는 일정하며 비확률적이기 때문에 조건부를 의미하는 표기, 즉 "|"가 필요하지 않다. 따라서 $E(e|x) = 0$ 대신에 $E(e) = 0$이라고 나타낸다. x를 고정된 것으로 취급할 수 없는 중요한 경우가 있으며 이에 대해서는 제10장에서 논의할 것이다.

계량경제학에서는 회귀모형의 가정을 무작위 오차항 e의 측면에서 논의하는 것이 관례적이다. 앞으로 참조하는 데 이용하기 위해 가정들을 SR1−SR6라 명명할 것이며 여기서 'SR'은 '단순회귀(simple regression)'를 의미한다. x는 고정되며 확률적이 아니라고 취급하므로 이제부터 조건부 표기, 즉 $y|x$를 사용하지 않을 것이다.

단순 선형회귀 모형에 관한 가정−II

SR1. x의 각 값에 대해 y값은 다음과 같다.

$$y = \beta_1 + \beta_2 x + e$$

SR2. 무작위 오차 e의 기댓값은 다음과 같다.

$$E(e) = 0$$

왜냐하면 다음과 같이 가정하였기 때문이다.

$$E(y) = \beta_1 + \beta_2 x$$

SR3. 무작위 오차 e의 분산은 다음과 같다.

$$\text{var}(e) = \sigma^2 = \text{var}(y)$$

확률변수 y 및 e는 동일한 분산을 갖는다. 왜냐하면 이들은 단지 일정한 상수만큼 차이가 나기 때문이다.

SR4. 무작위 오차의 한 쌍인 e_i와 e_j의 공분산은 다음과 같다.

$$\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$$

무작위 오차 e가 통계적으로 독립적인 경우 종속변수 y의 값도 통계적으로 독립적이라고 할 경우 더욱 강한 가정이 된다.

SR5. 변수 x는 확률적이지 않으며 최소한 2개의 상이한 값을 가져야 한다.

SR6. (선택적) y값들이 정규분포될 경우 e값은 자신의 평균을 중심으로 정규분포되며 그 역도 성립한다.

$$e \sim N(0, \sigma^2)$$

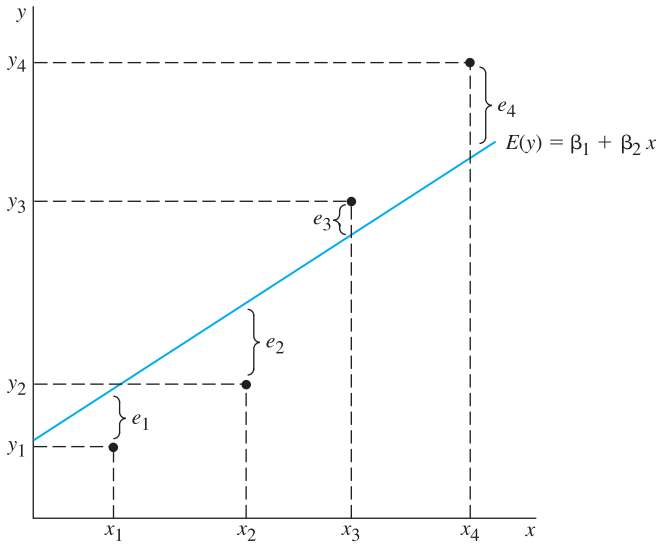
무작위 오차 e의 종속변수 y는 모두 확률변수이며 이 중 하나의 특성은 다른 하나의 특성에서 비롯된다. 그러나 이들 확률변수들 간에는 흥미로운 차이점이 있는데, 그것은 y가 '관찰할 수 있는' 반면에 e는 '관찰할 수 없는' 것이다. 회귀모수 β_1 및 β_2를 아는 경우 y값에 대해 $e = y - (\beta_1 + \beta_2 x)$를 계산할 수 있다. 이는 그림 2.5를 통해 알 수 있다. 회귀함수 $E(y) = \beta_1 + \beta_2 x$를 알 경우 y를 고정된 부분과 무작위 부분으로 분리할 수 있다 하지만 β_1과 β_2는 결코 알 수 없으므로 e를 계산하는 것은 불가능하다.

e에 관해 위와는 다소 다르게 생각해 보는 것도 필요하다. 무작위 오차 e는 소득 이외에 y에 영향을 미치는 모든 요소들을 나타내며 이로 인해 개별적인 관찰값 y가 평균값 $E(y) = \beta_1 + \beta_2 x$와 달라지게 된다. 식료품 지출액의 예에서 지출액 y와 이것의 평균 $E(y)$의 차이는 어떤 요인에서 비롯된 것일까? 이는 다음과 같이 설명할 수 있다.

1. 위의 모형에서 우리는 유일한 설명변수로 소득을 포함시켰다. 식료품 지출액에 영향을 미치는 다른 경제변수들은 오차항에 '함께 포함되어 있다.' 당연히 어느 경계모형에서나 모형

그림 2.5

y, *e*와 참인 회귀선 사이의 관계



내에 모든 중요하고 관련된 설명변수를 포함시키고자 하므로 오차항 e는 식료품에 대한 가계 지출액에 영향을 미치는 관찰할 수 없거나 또는 그리고 중요하지 않은 요소를 '내포하는 항'이 된다. 이 때문에 오차항은 x와 y 사이의 관계를 불분명하게 하는 교란적인 요소를 추가 시키게 된다.

2. 오차항 e는 발생할 수 있는 대략적인 근사오차를 설명할 뿐이다. 왜냐하면 가정하고 있는 선형함수 형태는 현실에 대해 단지 근사한 형태일 뿐이기 때문이다.
3. 오차항은 각 개인이 갖고 있는 무작위 행태적인 요소를 내포하고 있다. 한 개인의 식료품 지출에 영향을 미치는 모든 변수를 알고 있더라도 지출액을 완벽하게 예측하는 데는 충분하지 않을 수 있다. 예측할 수 없는 무작위 행태적인 요소가 또한 e에 포함될 수 있다.

어떤 중요한 요소를 빠뜨리거나 다른 심각한 **모형 설정 오차**(specification error)가 발생하는 경우 가정 SR2. $E(e) = 0$을 위반하게 되어 심각한 상황이 초래될 수 있다.

2.3 회귀모수의 추정

앞 절에서 소개한 경제 및 계량경제모형은 표본자료를 이용하여 절편 및 기울기의 모수인 β_1 및 β_2를 추정하는 데 기초가 된다. 예를 들어 무작위 표본 40개 가구로부터 얻은 가계 식료품 지출액과 주당 소득에 관한 자료를 검토해 보자. 대표적인 관찰값 및 통계적 요약은 표 2.1에

표 2.1 식료품 지출액과 소득에 관한 자료

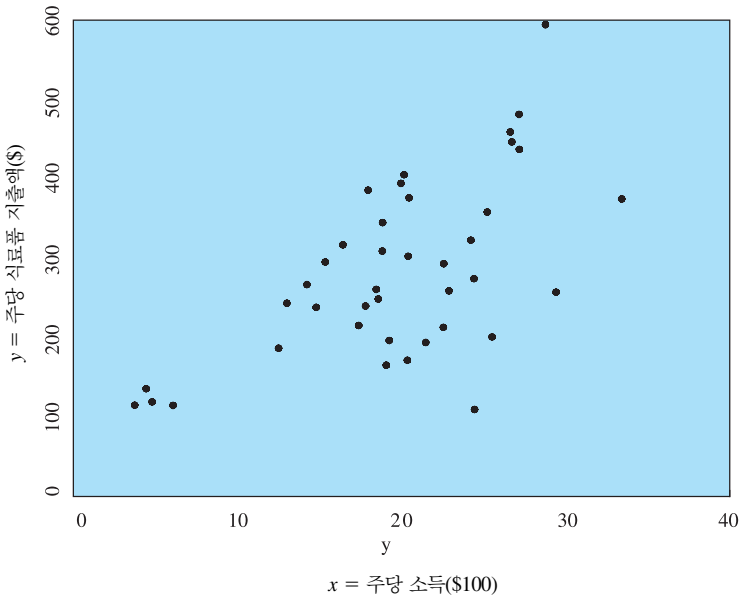
관찰(가계)	식료품 지출액($)	주당 소득($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	⋮	
39	257.95	29.40
40	375.73	33.40
	통계적 요약	
표본 평균	283.5735	19.6048
중앙값	264.4800	20.0300
최대값	587.6600	33.4000
최소값	109.7100	3.6900
표준 편차	112.7652	6.8478

있다. 단지 3인 가구만을 고려함으로써 가계 규모를 통제하였다. y값은 3인의 가구에 대한 주당 식료품 지출액을 달러로 나타낸 것이다. 소득은 달러로 나타내는 대신에 100달러 단위로 측정하였다. 왜냐하면 소득이 1달러 증가하더라도 식료품 지출액에 미치는 영향은 숫자로 나타낼 경우 미미하기 때문이다. 따라서 첫 번째 가계의 경우 관찰된 소득은 주당 369달러이며 주당 식료품 지출액은 115.22달러이다. 40번째 가계의 경우 주당 소득은 3,340달러이며 주당 식료품 지출액은 375.73달러이다.

표 2.1에 있는 식료품 지출액 자료는 가정 SR1−SR5를 충족시킨다고 가정한다. 즉, 가계 식료품 지출액의 기댓값은 소득에 대해 선형함수라고 가정한다. y의 기댓값에 대한 이런 가정은 무작위 오차가 0의 값을 갖는다는 가정과 동일하며 중요한 요소를 빠뜨리지 않았다는 의미이다. 무작위 오차 e의 분산과 동일한 y의 분산은 일정하다고 가정한다. 이는 우리가 모든 관찰값에 대해 y와 x 사이의 관계에 관해 동일하게 확신하지 못한다는 의미이다. 상이한 가계의 y값은 서로 상관되지 않으며 이는 무작위 표본추출을 통해 얻은 자료일 경우 가능하다. x값은 실제로 무작위 표본추출 통해 구했지만 표본의 x값에 대해 조건부 분석을 할 것이다. 이로써 x값을 반복적인 표본에서 고정된 값을 갖는 것으로 취급할 것이다. 하지만 최종적으로 보면 이런 단순화 작업은 우리의 분석에 영향을 미치지 않는다.

가계 식료품 지출액의 표본 관찰값을 설명하는 데 필요한 이론이 주어진 경우 그 다음 문제는 식료품 지출액−소득 관계에 대한 미지의 절편 및 기울기 계수를 의미하는 미지의 모수 β_1 및 β_2를 추정하기 위해 표본 정보 y_i와 x_i를 어떻게 사용하느냐이다. 40개 자료의 점을 (y_i, x_i), $i = 1\cdots, N = 40$으로 나타내고 이를 도표로 나타내면 그림 2.6과 같은 **산포도**(scatter diagram)가 된다.

그림 2.6
식료품 지출액의 예에 관한 자료



유의사항

이 책에서는 편의상 횡단면 자료 관찰값에 대해서는 '*i*'라는 아랫첨자와 N이라는 표본 관찰값의 수를 사용할 것이다. 시계열 자료 관찰값에 대해서는 '*t*'라는 아랫첨자와 *T*라는 관찰값의 수를 사용할 것이다. 순수히 대수학적인 의미나 총체적인 상황하에서는 둘 중 하나를 사용할 수도 있다.

문제는 평균 지출선 $E(y) = \beta_1 + \beta_2 x$의 위치를 추정하는 것이다. 이는 평균적인 행태를 나타내므로 자료를 의미하는 모든 점의 중간 어느 지점에 위치할 것으로 판단된다. β_1 및 β_2를 추정하기 위해 자료들의 중간을 통과하는 적절한 선을 긋고 나서 자로 기울기 및 절편을 측정할 수도 있다. 이 방법의 문제점은 다른 사람은 서로 다른 선을 그리게 되어 형식적인 기준이 결여되므로 정확성을 평가하기가 곤란하다는 것이다. 다른 방법은 가장 적은 소득 $i = 1$로부터 가장 많은 소득 $i = 40$으로 선을 긋는 것이다. 이는 형식적인 규칙을 제시할 수는 있지만 매우 좋은 방법이라고는 할 수 없다. 왜냐하면 나머지 38개 관찰값의 정확한 위치에 관한 정보를 고려하지 않기 때문이다. 모든 자료가 갖고 있는 모든 정보를 이용할 수 있도록 규칙을 만드는 것이 더 나은 방법이라 할 수 있다.

2.3.1 최소제곱 원칙

β_1 및 β_2를 추정하기 위해 표본 관찰값을 어떻게 이용할지에 관해 설명해 주는 규칙 또는 공식이 필요하다. 여러 가지 규칙을 적용할 수 있지만 우리가 사용하고자 하는 규칙은 **최소제곱 원칙**(least squares principle)에 기초하고 있다. 이 원칙에 따르면 자료값들에 적합한 선을 구하기 위해 이 값들을 나타내는 각 점으로부터 선까지의 수직 거리를 제곱한 합이 가능한 작게 되도록 선을 그어야 한다. 거리를 제곱하는 이유는 양의 먼 거리가 음의 먼 거리에 의해 상쇄되는 것을 방지하기 위해서이다. 이 규칙은 자의적인 것 같지만 매우 효과적이며 자료의 중간을 통과하도록 선을 긋는 간단한 방법 중 하나이다. 최소제곱 원칙을 이용하여 자료에 가장 적합하게 그은 선과 절편 및 기울기는 β_1 및 β_2의 최소제곱 추정값인 b_1 및 b_2가 된다. 이 경우 적합한 선은 다음과 같다.

$$\hat{y}_i = b_1 + b_2 x_i \tag{2.5}$$

각 점으로부터 적합하게 그은 선까지의 수직거리가 **최소제곱 잔차**(least squares residuals)이며 다음과 같이 나타낼 수 있다.

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i \tag{2.6}$$

이 잔차는 그림 2.7(a)에서 찾아볼 수 있다.

이제는 위와 다른 어떤 다른 선을 자료에 맞추어 그었다고 가정하고 이를 다음과 같이 나타내자.

$$\hat{y}_i^* = b_1^* + b_2^* x_i$$

여기서 b_1^* 및 b_2^*는 위와는 다른 절편 및 기울기의 값을 나타낸다. 이 선의 잔차 $\hat{e}_i^* = y_i - \hat{y}_i^*$는 그림 2.7(b)에서 찾아볼 수 있다. 최소제곱 추정값 b_1 및 b_2는 이들을 제곱한 잔차의 합이 어떤 다른 선의 제곱한 잔차의 합보다 작다는 특성을 갖고 있다.

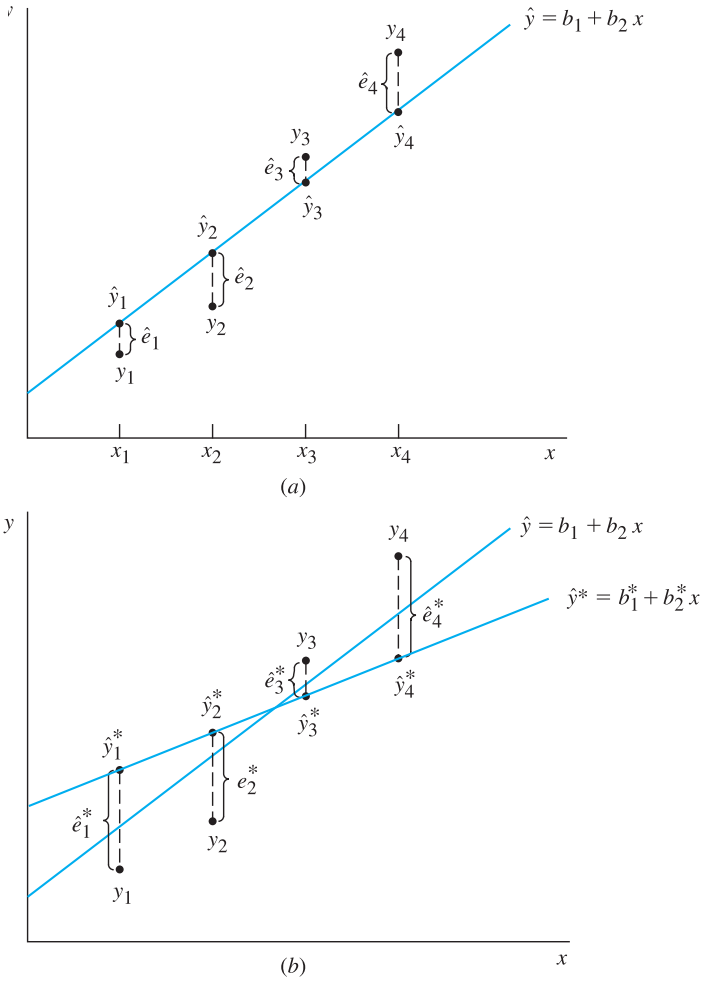
(2.6)의 최소제곱 잔차를 제곱한 합이 다음과 같다면

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2$$

그리고 다른 추정량에 기초한 제곱한 잔차의 합이 다음과 같다면

그림 2.7

(a) y, $\hat{e}$, 적절한 회귀선 사이의 관계, (b) 다른 적합한 선의 잔차



$$SSE^* = \sum_{i=1}^{N} \hat{e}_i^{*2} = \sum_{i=1}^{N} (y_i - \hat{y}_i^*)^2$$

다른 선이 자료들 사이로 어떻게 그려지든지 상관없이 다음과 같아진다.

$$SSE < SSE^*$$

최소제곱 원칙에 따르면 β_1 및 β_2의 추정값으로 b_1 및 b_2를 사용해야 한다. 왜냐하면 절편 및 기울기로서 이를 이용하는 선이 자료에 가장 잘 적합하기 때문이다.

이제 문제는 편리한 방법으로 b_1 및 b_2를 구하는 것이다. y와 x에 대한 표본 관찰값이 주어진 경우 '제곱을 합한' 함수를 최소화하는 미지의 모수 β_1 및 β_2값을 구하고자 한다.

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2$$

이는 쉽게 해결할 수 있는 계산문제로 자세한 내용은 이 장 끝의 부록 2A에 있다. 제곱한 잔차의 합을 극소화하는 β_1 및 β_2의 최소제곱 추정값의 공식은 다음과 같다.

최소제곱 추정량

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.7}$$

$$b_1 = \bar{y} - b_2\bar{x} \tag{2.8}$$

위에서 $\bar{y} = \sum y_i/N$ 및 $\bar{x} = \sum x_i/N$는 y 및 x에 대한 관찰값의 표본평균이다.

b_2에 대한 공식은 x_i가 모든 관찰값에 대해 동일한 값을 갖지 않는다고 가정해야 하는 이유를 설명해 주고 있다[SR5 참조]. 예를 들어 모든 관찰값에 대해 $x_i = 5$인 경우 (2.7)의 분자 및 분모가 0이 되어 b_2는 수학적으로 확정되지 않으며 존재하지 않는다.

표본값 y_i와 x_i를 (2.7) 및 (2.8)에 대입하면 절편 및 기울기 모수 β_1 및 β_2의 최소제곱 추정값을 구할 수 있다. b_1 및 b_2에 대한 공식은 완벽하게 일반적이며 표본값이 무엇이든지에 관계없이 사용될 수 있다. 이것은 매우 중요한 점이다. b_1 및 b_2에 대한 공식이 표본자료가 무엇이든

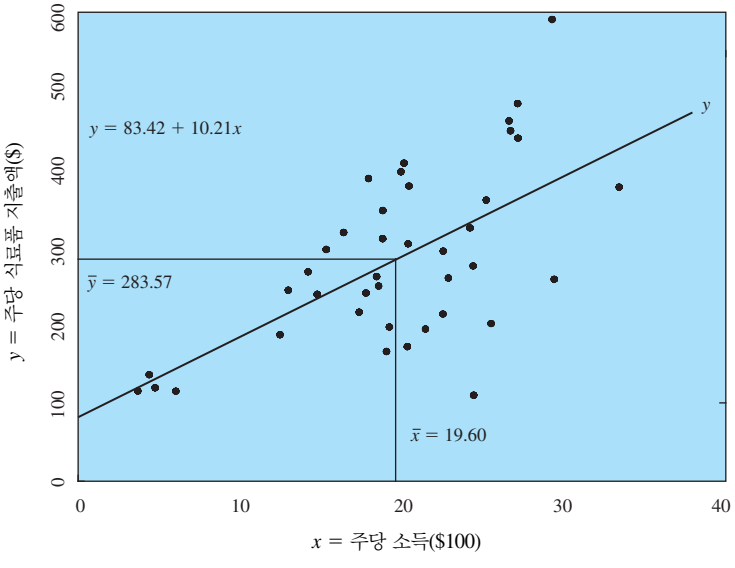


그림 2.8
적합하게 그은 회귀선

지에 관계없이 사용될 수 있는 경우 b_1 및 b_2는 확률변수가 된다. 실제 표본값을 공식에 대입시키면 확률변수의 관찰값인 수를 구할 수 있다. 이 두 경우를 구별하기 위하여 b_1 및 b_2에 대한 규칙 또는 일반적인 공식을 **최소제곱 추정량**(least squares estimator)이라 하며 특정 표본으로 공식을 이용하여 구한 수를 **최소제곱 추정값**(least squares estimate)이라 한다.

- 최소제곱 추정량은 일반적인 공식이며 확률변수이다.
- 최소제곱 추정값은 일반적인 공식을 관찰된 자료에 적용하여 구한 숫자이다.

추정량과 추정값의 구별은 이 책의 나머지 부분에서 다루게 될 모든 내용을 이해하기 위해 필수적으로 필요한 기본적인 개념이므로 명심해 두어야 한다.

2.3.2 식료품 지출액 함수의 추정값

최소제곱 추정량 (2.7) 및 (2.8)을 사용하여 표 2.1의 자료에 기초한 식료품 지출액 사례의 절편 및 기울기 모수인 β_1 및 β_2의 최소제곱 추정값을 구할 수 있다. (2.7)로부터 다음을 구할 수 있다.

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{18671.2684}{1828.7876} = 10.2096$$

(2.8)로부터는 다음을 구할 수 있다.

$$b_1 = \bar{y} - b_2\bar{x} = 283.5737 - (10.2096)(19.6048) = 83.4160$$

b_1 및 b_2의 값을 알리는 편리한 방법은 추정된 또는 적합한 회귀선으로 나타내는 것이다. 추정값을 적절히 반올림하면 다음과 같다.

$$\hat{y}_i = 83.42 + 10.21x_i$$

이 선을 그림 2.8에서 도표로 나타냈다. 이 선의 기울기는 10.21이며 수직축과 교차하는 절편은 83.42이다. 최소제곱에 적합하게 그은 선은 매우 정확한 방법으로 자료의 중간을 통과하게 된다. 왜냐하면 최소제곱 모수 추정값에 기초하여 그은 선의 특징 중 하나는 표본평균 $(\bar{x}, \bar{y}) = (19.6048, 283.5735)$으로 정의된 점을 통과하기 때문이다. 이는 (2.8)을 $\bar{y} = b_1 + b_2\bar{x}$로 나타냄으로써 직접적으로 알 수 있다. 이처럼 '평균값의 점'은 회귀분석에서 유용한 참조값이 된다.

2.3.3 추정값의 해석

최소제곱 추정값을 일단 구하면 고려하고 있는 경제모형의 테두리 내에서 이를 해석하게 된다. $b_2 = 10.21$인 값은 β_2의 추정값이다. 주당 가계소득 x는 \$100 단위로 측정되었음을 기억하자. 회귀선의 기울기 β_2는 주당 가계소득이 \$100 증가할 경우 식료품에 대한 가계의 증가액을 의미한다. 따라서 소득이 \$100 증가할 경우 식료품에 대한 주당 지출액은 대략 \$10.21만큼 증가할 것으로 추정된다. 어떤 지역의 소득 및 가구수의 변화에 대한 정보를 갖고 있는 슈퍼마켓 연쇄점은 소득이 \$100 증가할 때마다 주당 가구에 대해 \$10.21만큼 판매를 증대시킬 수 있다고 추정할 수 있으며 이는 장기적인 계획 수립에 매우 유용한 정보가 된다.

엄격히 말해 절편 추정값인 $b_1 = 83.42$는 소득이 0인 가계의 식료품에 대한 주당 지출액을 추정한 값이다. 대부분의 경제모형에서 추정된 절편값을 해석할 경우 매우 주의를 기울여야 한다. 문제는 그림 2.8의 식료품 지출액 자료에서 알 수 있듯이 $x = 0$인 근처에서는 보통 자료를 구하기가 어렵다는 점이다. 소득이 0인 근처에서 관찰값을 구할 수 없는 경우 추정된 관계가 그 경우에는 현실에 적합하지 않다고 할 수 있다. 따라서 추정된 모형에 따르면 소득이 0인 가계는 식료품에 주당 \$83.42를 지출한다고 하지만 이 추정값을 문자 그대로 해석하는 것은 위험할 수 있다. 이런 문제는 추정한 모든 경제모형에 적용된다.

2.3.3a 탄력성

소득 탄력성은 소득 변화에 대한 소비자 지출의 반응을 나타내는 유용한 방법이다. 변수 x에 대한 어떤 변수 y의 탄력성은 다음과 같이 정의된다.

$$\varepsilon = \frac{y\text{의 백분율 변화}}{x\text{의 백분율 변화}} = \frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x}\cdot\frac{x}{y}$$

(2.1)의 선형 경제모형에서 다음의 관계를 도출할 수 있다.

$$\beta_2 = \frac{\Delta E(y)}{\Delta x}$$

따라서 소득에 대한 '평균' 지출액의 탄력성은 다음과 같다.

$$\varepsilon = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x}\cdot\frac{x}{E(y)} = \beta_2\cdot\frac{x}{E(y)} \tag{2.9}$$

탄력성을 추정하기 위해 β_2를 $b_2 = 10.21$로 대체시켜야 한다. 선형모형에서 탄력성은 회귀선

상의 각 점에서 상이하므로 'x'와 '$E(y)$'를 어떤 것으로 대체시켜야만 한다. 탄력성은 가장 일반적으로 '평균값의 점' $(\bar{x}, \bar{y}) = (19.60, 283.57)$에서 계산되는데 이는 회귀선상의 대표적인 점이기 때문이다. 평균값의 점에서 소득 탄력성을 계산하면 다음과 같다.

$$\hat{\varepsilon} = b_2 \frac{\bar{x}}{\bar{y}} = 10.21 \times \frac{19.60}{283.57} = 0.71$$

위의 추정된 소득 탄력성은 다른 경우와 마찬가지로 해석된다. 즉, 주당 가계소득이 1% 변화하면 $(x, y) = (\bar{x}, \bar{y}) = (19.60, 283.57)$인 경우 식료품에 대한 주당 가계 지출액은 약 0.71% 증가한다. 추정된 소득 탄력성이 1보다 작기 때문에 식료품은 '사치품'이 아닌 '필수품'으로 분류되어 이는 '평균적인' 가계에 대해 기대할 수 있는 것과 일치한다.

2.3.3b 예측

추정된 식은 예측하거나 예상하는 데 또한 사용될 수 있다. 주당 소득이 $2,000인 가계의 주당 식료품 지출액을 예측하고자 한다고 가상하자. 위의 추정된 식에 $x = 20$을 대체시켜 예측하면 다음과 같은 결과를 얻을 수 있다.

$$\hat{y}_i = 83.42 + 10.21x_i = 83.42 + 10.21(20) = 287.61$$

위의 결과를 이용하여 주당 소득이 $2,000인 가계는 식료품에 주당 $287.61를 지출한다고 예측할 수 있다.

2.3.3c 컴퓨터를 이용한 분석 결과

서로 상이한 많은 컴퓨터 소프트웨어 패키지를 이용하여 최소제곱 추정값을 계산할 수 있다. 각각의 소프트웨어 패키지의 회귀분석 결과는 서로 상이한 것처럼 보이며 분석 결과를 설명하기 위해 상이한 용어를 사용한다. 이런 차이에도 불구하고 다양한 분석 결과는 동일한 기본적인 정보를 제공하고 있으며 독자들은 이런 정보가 어디에 위치하며 어떻게 해석하는지를 알 수 있어야 한다. 하지만 소프트웨어 패키지는 또한 여러분들이 의미를 쉽게 알 수 없는 다양한 숫자들을 제공함으로써 이런 문제들을 다소 복잡하게 만들고 있다. 예를 들면 식료품 지출액 자료를 기초로 EViews의 소프트웨어 패키지를 이용한 분석 결과가 그림 2.9에 있다.

EViews 분석 결과에 따르면 모수 추정값은 'Coefficient' 열에 있으며 상수항(추정값 b_1)은 '*C*' 그리고 '*INCOME*' (추정값 b_2)라 명명되어 있다. 컴퓨터 소프트웨어 프로그램은 일반적으로 추정값을 컴퓨터 프로그램에 명명된 변수의 명칭(여기서 변수를 *INCOME*라고 명명하였다)과 '상수'에 대한 약자로 나타낸다. 현재 수준에서 여러분들이 알 수 있는 숫자로는 'Sum squared resid'라 명명된 $SSE = \sum \hat{e}_i^2 = 304505.2$와 'Mean dependent var'이라 명명된 y의 표본

그림 2.9
EViews를 이용한 회귀분석 결과

Dependent Variable: *FOOD_EXP*
Method: Least Squares
Sample: 1 40
Included observations: 40

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

R-squared	0.385002	Mean dependent var	283.5735
Adjusted R-squared	0.368818	S.D. dependent var	112.6752
S.E. of regression	89.51700	Akaike info criterion	11.87544
Sum squared resid	304505.2	Schwarz criterion	11.95988
Log likelihood	−235.5088	Hannan-Quinn criter	11.90597
F-statistic	23.78884	Durbin-Watson stat	1.893880
Prob(F-statistic)	0.000019		

평균 $\bar{y} = \sum y_i/N = 283.5735$가 있다.

앞으로 계속해서 계산 결과의 나머지 부분에 대해서도 논의를 할 것이며 여러분들은 결국 이들 결과가 제시하는 나머지 정보도 이해할 수 있게 될 것이다.

2.3.4 다른 경제모형

식료품에 대한 가계 지출액과 소득 간의 관계를 예로 들어 단순회귀분석을 소개하였다. 단순회귀 모형은 경제학, 경영학, 사회과학에서 여러 관계의 모수를 추정하는 데 사용된다. 회귀분석을 적용하는 일은 흥미로우면서도 유용하며 예를 들면 다음과 같다.

- 전기 기사의 시간당 임금이 5% 인상될 경우 새 집의 가격은 얼마나 인상되는가?
- 담배세가 $1 인상될 경우 미국 루이지애나주에서는 추가 수입이 얼마나 발생하는가?
- 중앙은행이 이자율을 0.5% 인상할 경우 6개월 내에 소비자 차용은 얼마나 감소하는가? 인상이 이루어진 이후의 달에는 실업률이 어떻게 변화하는가?
- 2008년에 유치원 교육 프로그램에 대한 기금을 증대시킬 경우 2020년 고등학교 졸업률에는 어떤 영향을 미치는가? 2012년과 그 이후에 청소년 범죄율에는 어떤 영향을 미치는가?

회귀분석은 사회과학 및 자연과학의 거의 모든 분야뿐만 아니라 경제학 및 재무관리 분야에도 적용된다. 한 변수가 변화할 경우 다른 변수에 얼마나 영향을 미치는지 알고자 한다면 위와 같은 회귀분석을 이용할 수 있다.

단순 선형회귀 모형은 언뜻 보는 것보다 훨씬 더 유연한 형태를 갖는다. 왜냐하면 변수 y와

x는 대수, 제곱, 세제곱, 기본 경제변수의 역수를 포함하여 변형이 이루어질 수 있기 때문이다. 따라서 단순 선형회귀 모형은 변수들 간의 비선형 관계에 사용될 수 있다. 이 다소 이해하기 곤란한 용어는 '선형회귀'의 선형이란 용어는 실제로 모수가 어떤 방법으로든 변형되지 않는다는 것을 의미한다는 데서 비롯됐다고 할 수 있다, 선형회귀 모형에서 모수는 예를 들면 $\ln(\beta_2)$ 또는 $\beta_1 \cdot \beta_2$ 또는 $\beta_2^{\beta_1}$처럼 승수형태를 갖거나 변형되어서는 안 된다.

2.4 최소제곱 추정량에 대한 평가

식료품 지출액 자료를 이용하고 최소제곱 공식 (2.7) 및 (2.8)에 기초하여 회귀모형 $y_i = \beta_1 + \beta_2 x_i + e_i$의 모수를 추정하였다. 최소제곱 추정값으로 $b_1 = 83.42$ 및 $b_2 = 10.21$을 구했다. "이 추정값은 얼마나 타당한가?"와 같은 질문을 할 수 있지만 이는 잘못된 질문이라고 할 수 있다. 이 물음에 답을 할 수가 없다. 왜냐하면 모집단의 모수 β_1 또는 β_2의 참값을 알 수 없으므로 $b_1 = 83.42$ 및 $b_2 = 10.21$이 참값에 얼마나 근접하였는지를 말할 수 없다. 최소제곱 추정값은 모수의 참값에 근접할 수도 있고 근접하지 않을 수도 있으며 이를 알 수 없다.

추정값의 특성을 알아보기보다는 한 발자국 뒤로 물러서서 최소제곱 추정 절차의 성격에 대해 살펴볼 것이다. 그 이유는 다음과 같다. 또 다른 표본조사 40개 가구를 선택하여 다른 표본자료를 작성할 경우 처음 표본과 동일한 소득을 갖는 가구를 조심스럽게 선택하더라도 상이한 추정값 b_1 및 b_2를 얻게 된다. 이런 **표본추출 변동**(sampling variation)은 불가피하게 된다. 가계의 식료품 지출액, y_i, $i = 1, \cdots, 40$은 확률변수이므로 표본이 상이할 경우 추정값도 상이하게 된다. 이들의 값은 표본이 추출될 때까지 알지 못한다. 따라서 추정절차의 관점에서 볼 때 b_1 및 b_2는 확률변수 y에 의존하게 되므로 이들도 역시 확률변수가 된다. 이런 경우 b_1 및 b_2를 최소제곱 추정량이라 한다.

표본추출 특성(sampling properties)이라 불리는 추정량 b_1 및 b_2의 특성을 알아보고 다음과 같은 의문점들을 살펴볼 것이다.

1. 최소제곱 추정량 b_1 및 b_2가 확률변수라면 이들의 평균, 분산, 공분산, 확률분포는 무엇인가?
2. 최소제곱 원칙은 β_1 및 β_2의 추정값을 구하기 위해 자료를 이용하는 한 가지 방법일 뿐이다. 최소제곱을 이용한 추정량은 사용할 수 다른 규칙 및 이를 이용한 추정량과는 어떻게 비교할 수 있는가? 예를 들면 β_2에 근접한 추정값을 구할 수 있는 확률이 더 높은 다른 추정량이 있는가?

위의 물음에 대한 대답은 본질적으로 가정 SR1−SR5가 충족되었는지에 의존한다. 이 책의 뒷

부분에서 이런 가정들이 특정 모형에서 준수되는지 여부를 어떻게 점검하고 1개 이상의 가정이 준수되지 않는 경우 어떻게 해야 되는지를 알아볼 것이다.

유의사항

다음 절들에서 최소제곱 추정량의 특성에 관해 살펴볼 것이다. 중요한 결과에 관한 증명은 이 장 뒷부분에 있는 부록에 수록되어 있다. 회귀모형에서 이런 개념들을 검토하기에 앞서 보다 단순한 내용의 테두리 내에서 이를 살펴보는 것이 여러 가지 면에서 볼 때 더 타당하다.

2.4.1 추정량 b_2

공식 (2.7) 및 (2.8)을 사용하여 최소제곱 추정값 b_1 및 b_2를 계산해 보자. 하지만 이는 추정량의 이론적 특성을 검토하는 데 그렇게 적합하지 않다. 이 절에서는 분석을 용이하게 하기 위해서 b_2의 공식을 재작성할 것이다. (2.7)에서 b_2는 다음과 같다.

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

자료값에서 표본평균을 감했기 때문에 위의 식은 추정량을 평균으로부터의 편차로 나타내었다고 본다. 가정 SR1과 약간의 대수를 이용하여 b_2를 다음과 같이 **선형 추정량**(linear estimator)으로 나타낼 수 있다.

$$b_2 = \sum_{i=1}^{N} w_i y_i \tag{2.10}$$

여기서 다음과 같다.

$$w_i = \frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} \tag{2.11}$$

위에서 w_i항은 확률적이 아닌 x_i에만 의존하므로 w_i도 역시 확률적이지 않다. (2.10)처럼 y_i의 가중평균인 추정량을 선형 추정량이라 한다. 그리고 나서 대수학을 이용하면(부록 2D 참조) 이론적으로 편리한 방법으로 b_2를 다음과 같이 나타낼 수 있다.

$$b_2 = \beta_2 + \sum w_i e_i \tag{2.12}$$

여기서 e_i는 선형회귀 모형 $y_i = \beta_1 + \beta_2 x_i + e_i$의 오차항이다. 이 식은 우리가 알지 못하는 β_2 및 관찰할 수 없는 e_i에 의존하기 때문에 계산을 하는 데 유용하지는 않다. 하지만 최소제곱 추정량의 표본추출 특성을 이해하는 데 (2.12)는 매우 유용하다.

2.4.2 b_1 및 b_2의 기댓값

추정량 b_2는 표본이 수집될 때까지 그 값을 알 수 없으므로 확률변수이다. 우리는 지금 모형에 대한 가정이 준수될 경우 $E(b_2) = \beta_2$, 즉 b_2의 기댓값이 모수의 참값인 β_2와 동일해진다는 사실을 살펴보고자 한다. 모수 추정량의 기댓값이 모수의 참값과 동일한 경우 그 추정량은 **불편**(unbiased)되었다고 한다. $E(b_2) = \beta_2$이므로 최소제곱 추정량 b_2는 β_2의 불편 추정량이 된다. $E(b_2) = \beta_2$이므로 최소제곱 추정량 b_2는 β_2의 불편 추정량이 된다. 불편성에 대한 직관적인 의미는 수학적인 기댓값을 구하기 위해 반복적으로 시행된 표본추출에 대한 해석에서 찾아볼 수 있다. 크기가 N인 많은 표본이 수집되고 각 표본에 대해 b_2에 관한 공식을 이용하여 β_2를 추정할 경우 가정이 모두 준수된다면 모든 표본으로부터 구한 추정값 b_2의 평균값은 β_2가 된다.

선형회귀 모형의 가정들이 하는 역할을 설명할 수 있도록 위의 결과가 참이라는 사실을 보여 줄 것이다. (2.12)에서 어떤 부분이 확률적인가? 모수 β_2는 확률적이 아니다. 우리가 추정하고자 하는 것은 모집단의 모수이다. 가정 SR5가 준수될 경우 x_i는 확률적이지 않다. w_i는 x_i의 값에만 의존하므로 역시 확률적이지 않다. (2.12)에서 유일한 확률적 요소는 무작위 오차항 e_i이다. 합의 기댓값은 기댓값의 합이라는 사실을 이용하여 b_2의 기댓값을 다음과 같이 구할 수 있다.

$$\begin{aligned} E(b_2) &= E(\beta_2 + \sum w_i e_i) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \cdots + w_N e_N) \\ &= E(\beta_2) + E(w_1 e_1) + E(w_2 e_2) + \cdots + E(w_N e_N) \\ &= E(\beta_2) + \sum E(w_i e_i) \\ &= \beta_2 + \sum w_i E(e_i) = \beta_2 \end{aligned} \tag{2.13}$$

위의 식 (2.13)의 마지막 줄에서 우리는 다음과 같은 두 가지 가정을 이용하였다. 첫째, w_i는 확률적이 아니며 상수는 기댓값 밖으로 내보낼 수 있기 때문에 $E(w_i e_i) = w_i E(e_i)$가 된다. 둘째, $E(e_i) = 0$이라는 가정에 의존하였다. $E(e_i) \neq 0$인 경우 $E(b_2) \neq \beta_2$가 되며 b_2는 β_2의 편향된 추정량이 된다. e_i은 경제모형에서는 빠졌지만 y_i에 영향을 주는 요인들을 포함하고 있다는 사실을 기억하자. 중요한 것을 포함시키지 않을 경우 $E(e_i) \neq 0$ 및 $E(b_2) \neq \beta_2$가 된다. 따라서 관련된 모든 설명변수를 포함하고 있다는 의미에서 올바른 계량경제모형을 설정하는 것이 최소제

곱 추정량이 불편되기 위한 필수요소이다.

추정량 b_2의 불편성은 중요한 표본추출의 특성이다. 모집단으로부터 표본추출을 반복적으로 시행할 경우 최소제곱 추정량은 평균적으로 '적절한' 값을 갖게 되며 이것이 바로 추정량의 바람직한 특성이다. 이런 통계적 특성만이 b_2가 β_2의 적절한 추정량이라는 것을 의미하지는 않지만 논의의 한 부분을 구성한다. 불편성은 동일한 모집단에서 추출한 많은 자료표본에 달려 있다. b_2가 불편이라는 점은 단지 한 번의 표본추출에서 발생할 수 있는 사실에 관해 어떤 의미도 갖지 않는다. 개별적인 추정값(숫자인) b_2는 β_2에 근접할 수도 큰 차이가 날 수도 있다. β_2를 알 수 없기 때문에 표본이 하나인 경우 추정값이 β_2에 근접하는지 여부를 알 수 없다. 따라서 추정값 $b_2 = 10.21$은 β_2에 근접할 수도 있고 하지 않을 수도 있다.

모형에 대한 가정이 준수될 경우 β_1의 최소제곱 추정량 b_1은 불편 추정량이 되며 $E(b_1) = \beta_1$이 된다.

2.4.3 반복적인 표본추출

약간 다른 방법으로 불편 추정법을 설명하기 위하여 표 2.2는 표 2.1의 가계와 동일한 소득을 갖는 동일한 모집단으로부터 크기가 $N = 40$인 10개의 무작위 표본을 이용하여 식료품 지출액 모형의 최소제곱 추정값을 보여 주고 있다. 표본에 따라 최소제곱 모수 추정값이 변한다는 점에 유의하자. 이런 표본추출에 따른 변동은 각 표본에 40개의 상이한 가구가 포함되어 있다는 단순한 사실에서 기인하며 이들 주당 식료품 지출액은 무작위적으로 변한다.

불편성의 특성은 동일한 모집단으로부터 같은 크기의 많은 표본을 추출할 경우 b_1 및 b_2의 대략적인 평균값이라는 의미이다. 이 10개 표본에서 b_1의 평균값은 $\bar{b}_1 = 78.74$이며 b_2의 평균값은 $\bar{b}_2 = 9.68$이다. 많은 표본에서 추정값의 평균을 취할 경우 해당 평균값은 모수의 참값인

표 2.2 10개 표본에서 구한 추정값

표본	b_1	b_2
1	131.69	6.48
2	57.25	10.88
3	103.91	8.14
4	46.50	11.90
5	84.23	9.29
6	26.63	13.55
7	64.21	10.93
8	79.66	9.76
9	97.30	8.05
10	95.96	7.77

β_1 및 β_2에 근접할 것이다. 불편성은 어떤 표본의 추정값이 모수의 참값에 근접한다는 것을 의미하지 않으므로 한 추정값이 불편되었다고는 할 수 없으며 최소제곱 추정 절차(또는 최소제곱 추정량)가 불편되었다고만 할 수 있다.

2.4.4 b_1 및 b_2의 분산과 공분산

표 2.2는 β_1 및 β_2의 최소제곱 추정값이 표본에 따라 변화한다는 사실을 보여 주고 있다. 이런 변동성을 이해하는 것이 추정량의 신뢰성과 표본추출의 정확성을 평가하는 데 중요한 요소가 된다. 이제 추정량 b_1 및 b_2의 분산과 공분산을 구해 보자. 분산과 공분산에 대한 의미를 알아보기 전에 왜 이들을 살펴보는 것이 중요한지 생각해 보자. 확률변수 b_2의 분산은 확률변수의 값과 그의 평균 사이의 차이를 제곱한 것의 평균이며 우리는 $E(b_2) = \beta_2$이라는 사실을 알고 있다. b_2의 분산은 다음과 같이 정의된다.

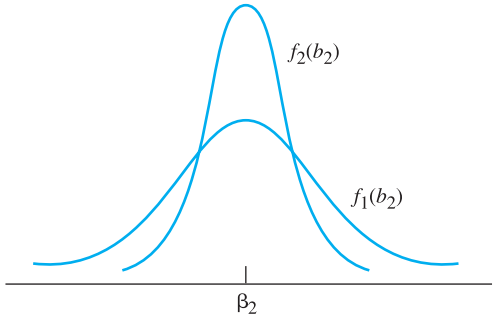
$$\text{var}(b_2) = E[b_2 - E(b_2)]^2$$

상기의 분산은 b_2의 확률분포가 퍼진 정도를 측정한다. 그림 2.10은 평균값은 같지만 분산이 상이한 b_2의 두 가지 확률분포인 $f_1(b_2)$와 $f_2(b_2)$를 보여 주고 있다.

확률밀도함수 $f_2(b_2)$는 확률밀도함수 $f_1(b_2)$보다 분산이 더 작다. 추정량의 정확성에 관심을 갖고 있고 선택을 할 수 있다면 b_2가 $f_1(b_2)$보다는 $f_2(b_2)$인 확률분포를 갖도록 선호하게 된다. 분산이 $f_2(b_2)$인 경우 확률은 모수의 참값인 β_2 주위에 더 집중되며 이는 β_2에 근접한 추정값을 얻을 확률이 더 높아진다는 의미이다. β_2에 근접한 추정값을 얻는 것이 목표라는 사실을 기억해야 한다.

추정량의 분산은 추정값이 표본에 따라 얼마나 변화하는지를 알려 준다는 의미에서 추정량의 정확성을 측정한다. 결과적으로 추정량의 **표본추출 분산**(sampling variance) 또는 **표본추출 정확성**(sampling precision)에 대해 자주 언급하게 된다. 추정량의 분산이 적을수록 해당 추정

그림 2.10
b_2에 대한 두 가지 확률밀도함수



량의 표본추출 정확성은 커지게 된다. 어떤 추정량의 표본추출 분산이 다른 추정량의 표본추출 분산보다 작은 경우 해당 추정량이 다른 추정량보다 더 정확하다.

이제는 b_1 및 b_2의 분산과 공분산에 대해 논의해 보자. 이 장 뒷부분에 있는 부록 2E에서는 최소제곱 추정량 b_2의 분산을 도출할 것이다. 회귀모형에 대한 가정 SR1–SR5가 준수되는 경우 (SR6가 반드시 준수될 필요는 없음) b_1 및 b_2의 분산과 공분산은 다음과 같다.

$$\text{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N\sum (x_i - \bar{x})^2} \right] \tag{2.14}$$

$$\text{var}(b_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{2.15}$$

$$\text{cov}(b_1, b_2) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right] \tag{2.16}$$

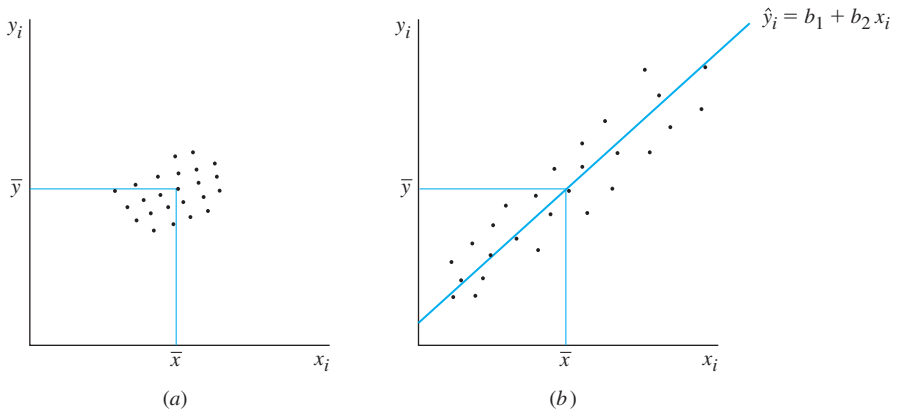
이 절 앞부분에서 불편 추정량인 경우 분산이 작은 것이 큰 것보다 낫다고 하였다. (2.14)–(2.16)에 있는 분산과 공분산에 영향을 미치는 요인들을 생각해 보자.

1. 무작위 오차항의 분산인 σ^2은 위에 있는 세 가지 공식 모두에 포함되어 있다. 이는 기댓값 $E(y)$에 대해 y값이 퍼진 정도를 반영한다. 분산 σ^2이 클수록 퍼진 정도가 더 커지며 y값이 평균 $E(y)$로부터 떨어지는 위치의 불확실성이 확대된다. β_1 및 β_2에 대한 정보가 덜 정확할수록 σ^2는 더욱 커진다. 그림 2.3에서 분산이 확률분포 $f(y|x)$의 퍼진 정도를 반영한다는 사실을 알았다. 분산항 σ^2이 클수록 통계모형의 불확실성이 커지며 최소제곱 추정량의 분산과 공분산이 증가한다.
2. 표본평균에 대한 x값의 차이를 제곱하여 합산한 $\sum(x_i - \bar{x})^2$이 분산과 공분산 모두에 포함되어 있다. 이는 독립 또는 설명변수 x의 표본값들이 평균으로부터 얼마나 벗어나 있는지를 측정한다. 변수들이 벗어나 있을수록 제곱의 합은 커지며 덜 벗어나 있을수록 제곱의 합이 작아진다. 제곱의 합, 즉 $\sum(x_i - \bar{x})^2$이 커질수록 최소제곱 추정량의 분산이 작아지며 미지의 모수를 더 정확하게 추정할 수 있다. 그림 2.11을 참조하면 이를 직관적으로 이해할 수 있다.

   그림 2.11의 (b)를 보면 x값들이 x축을 따라 넓게 퍼져 있는 반면 (a)는 자료들이 '뭉쳐 있음'을 알 수 있다. 손으로 자료에 적합한 선을 그으려는 경우 어떤 형태의 자료를 선호하겠는가? (b)의 경우에 최소제곱 추정선을 긋는 작업이 더 용이함을 명백히 알 수 있다. 왜냐하면 자료들이 x축을 따라 분포되어 있기 때문이다.
3. 표본 크기 N이 클수록 최소제곱 추정량의 분산과 공분산이 작아지므로 표본자료가 적은

그림 2.11

설명변수 x의 변동이 추정의 정확성에 미치는 영향 : (a) x의 변동이 적고 정확성이 낮은 경우, (b) x의 변동이 많고 정확성이 높은 경우



경우보다 많은 경우가 더 낫다. 각각의 합은 표본 크기 N을 포함하고 있으므로 분산과 공분산 모두에 N이 관련됨을 알 수 있다. 즉 $\text{var}(b_1)$에는 N이 명시적으로 포함되어 있으며 다른 경우에는 합을 구성하는 각 항이 양이거나 0이므로 (x값이 표본평균과 같은 경우 0이 됨) N이 증가할수록 제곱의 합인 $\sum(x_i - \bar{x})^2$가 점점 커진다. 결과적으로 제곱의 합이 분산과 공분산의 분모에 포함되어 있으므로 N이 증가할수록 $\text{var}(b_2)$와 $\text{cov}(b_1, b_2)$는 작아진다. N이 증가할수록 $\text{var}(b_1)$의 분자와 분모에 있는 합은 모두 커지지만 서로를 상쇄하게 되어 분모에 있는 N이 주요한 항으로서 남게 된다. 따라서 N이 증가함에 따라 $\text{var}(b_1)$은 작아지게 된다.

4. $\text{var}(b_1)$에는 $\sum x_i^2$항이 있는데 이 항이 커질수록 최소제곱 추정량 b_1의 분산이 커진다. 그 이유는 무엇일까? 절편의 모수 β_1은 $x = 0$일 때 y의 기댓값이다. 자료가 $x = 0$에서 멀리 떨어져 있을수록 식품 지출액의 예에서 살펴본 것처럼 β_1을 해석하는 일이 어려워지며 β_1을 정확하게 추정하는 일도 어려워진다. $\sum x_i^2$항은 원점, 즉 $x = 0$으로부터 자료와의 거리를 측정한다. x값이 0에 근접할 경우 $\sum x_i^2$항은 작아지며 이는 $\text{var}(b_1)$을 감소시킨다. 그러나 x값이 음수 또는 양수에 관계없이 크기가 커질 경우 $\sum x_i^2$항이 커지며 $\text{var}(b_1)$이 커진다.
5. x값의 표본평균이 $\text{cov}(b_1, b_2)$에 포함되어 있다. 표본평균 x의 크기가 커질수록 공분산이 증가하며 공분산은 x와 반대되는 부호를 갖는다. 이 논리는 그림 2.11을 통해 살펴볼 수 있다. (b)에서 최소제곱에 적합하게 그은 선은 평균점을 통과하여야 한다. 해당 자료에 적합한 선을 그은 경우 추정된 기울기 b_2를 증가시키는 경우를 생각해 보자. 그은 선은 평균점을 통과하여야 하므로 선과 수직축이 만나는 점을 낮추어야만 하며 이는 절편의 추정값 b_1이 감소해야 된다는 의미이다. 따라서 표본평균이 그림 2.11에서 보는 것처럼 양인 경우 기울기와 절편의 최소제곱 추정량 간 공분산은 음이 된다.

2.5 가우스— 마코프 정리

지금까지 최소제곱 추정량 b_1 및 b_2에 대해 알고 있는 것은 무엇인가?

- 추정량은 완벽하게 일반적이다. 공식 (2.7) 및 (2.8)을 사용하여 자료가 무엇이든지 단순 선형회귀 모형에서 미지의 모수 β_1 및 β_2를 추정할 수 있다. 따라서 이런 관점에서 보면 최소제곱 추정량 b_1 및 b_2는 확률변수이다.
- (2.10)에서 정의한 것처럼 최소제곱 추정량은 선형 추정량이다. b_1 및 b_2 둘 다 y_i값의 가중평균으로 나타낼 수 있다.
- 가정 SR1－SR5가 준수될 경우 최소제곱 추정량은 불편된다. 이는 $E(b_1) = \beta_1$ 및 $E(b_2) = \beta_2$라는 의미이다.
- b_1 및 b_2의 분산과 공분산에 관한 식을 알고 있다. 나아가 불편 추정량인 경우 분산이 작을수록 더 좋다고 하였다. 이는 모수의 참값에 근접한 추정값을 얻을 확률이 더 높아진다는 의미이다.

이제는 그 유명한 가우스-마코프(Gauss-Markov) 정리에 대해 논의해 보자. 이 장 뒷부분에 있는 부록 2F는 이 정리에 대한 증명을 하고 있다.

가우스-마코프 정리

선형회귀 모형에 관한 가정 SR1－SR5하에서 추정량 b_1 및 b_2는 β_1 및 β_2의 모든 선형 및 불편 추정량 중에서 최소의 분산을 가지며 β_1 및 β_2의 **최우수 선형 불편 추정량**(Best Linear Unbiased Estimators : BLUE)이다.

가우스-마코프 정리가 의미하는 것과 그렇지 않은 것을 정리해 보자.

1. 추정량 b_1 및 b_2는 선형 및 불편한 유사한 추정량들과 비교하여 '최우수'하다는 의미이지 모든 가능한 추정량 중에서 최우수하다는 의미는 아니다.
2. 추정량 b_1 및 b_2는 같은 부류 내에서 분산이 최소이므로 최우수하다고 본다. 2개의 선형 및 불편 추정량을 비교할 경우 보다 작은 분산을 갖는 추정량을 언제나 사용하길 원한다. 왜냐하면 추정에 관한 이런 규칙에 따를 경우 모수의 참값에 근접한 추정값을 구할 수 있는 확률을 높여 주기 때문이다.
3. 가우스-마코프 정리가 지켜지기 위해서는 가정 SR1－SR5가 준수되어야만 한다. 이 가정들 중 어느 하나가 준수되지 않을 경우 b_1 및 b_2는 β_1 및 β_2의 최우수 선형 불편 추정량

이 되지 못한다.

4. 가우스–마코프 정리는 정규성 가정(가정 SR6)에 의존하지 않는다.
5. 단순 선형회귀 모형에서 선형 및 불편 추정량을 사용하고자 하는 경우 더 이상의 탐색을 할 필요가 없다. 추정량 b_1 및 b_2가 사용하고자 하는 바로 그것들이다. 이것이 바로 우리가 왜 이 추정량을 학습하는지 이유가 되며(틀린 추정 규칙을 배울 이유가 없지 않은가?) 이 추정량들이 경제학뿐만 아니라 모든 다른 사회과학 및 자연과학에서 폭넓게 사용되는 이유이다.
6. 가우스–마코프 정리는 최소제곱 추정량에 적용되지만 단 하나의 표본에 기초한 최소제곱 추정값에는 적용되지 않는다.

2.6 최소제곱 추정량의 확률분포

지금까지 살펴본 최소제곱 추정량의 특성은 정규성 가정인 SR6에 의존하지 않았다. 무작위 오차 e_i가 평균 0 및 분산 σ^2를 갖는 정규분포를 한다고 추가적으로 가정할 경우 최소제곱 추정량의 확률도 정규분포하게 된다. 이 결론은 두 가지 단계를 거쳐 추론할 수 있다. 첫째, 가정 SR1에 기초하여 e_i가 정규분포하는 경우 y_i도 역시 정규분포하게 된다. 둘째, 최소제곱 추정량은 형태가 $b_2 = \sum w_i y_i$인 선형 추정량이며 정규 확률변수의 합계는 정규분포한다. 결과적으로 오차항에 대한 가정 SR6인 정규성 가정을 할 경우 최소제곱 추정량은 정규분포한다.

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N\sum(x_i - \bar{x})^2}\right) \tag{2.17}$$

$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right) \tag{2.18}$$

제3장에서 살펴볼 것처럼 최소제곱 추정량의 정규성은 통계적 추론을 하는 많은 경우에 있어 상당히 중요한 의미를 갖는다.

오차가 정규분포하지 않는 경우 어떤 일이 발생하는가? 최소제곱 추정량의 확률분포에 관해 언급할 수 있는가? 이 물음에 대한 대답은 때때로 할 수 있다고 본다.

중심극한 정리(central limit theorem)

가정 SR1−SR5가 준수되고 표본 크기 N이 **충분히 큰 대표본**(sufficiently large)인 경우 최소제곱 추정량은 (2.17) 및 (2.18)에서 살펴본 것처럼 정규분포에 근접한 분포를 갖게 된다.

문제는 "충분히 큰 대표본은 얼마나 커야 하는가?"이며 이에 대해 특정 숫자를 제시할 수는 없다. 이와 같이 모호하고 불만족스러운 대답을 해야 하는 이유는 "얼마나 커야 하는가?"라는 물음에 대한 대답이 여러 가지 요소에 따라 달라질 수 있기 때문이다. 예를 들면 이에 대한 대답을 무작위 오차의 분산이 어떤 모양(예를 들면, 완만한가? 대칭적인가? 비대칭적인가?)이며 x_i 값이 무엇인지 등에 따라 달라질 수 있다. 단순회귀 모형에서 혹자는 $N = 30$이면 충분히 크다고 하겠지만 이 책의 저자들은 이 물음에 대해 다소 보수적인 입장을 취하여 $N = 50$이 합리적이라고 본다. 그러나 중요한 점은 이런 숫자들이 대략적인 규칙이며 '충분히 큰 대표본'이라는 의미는 문제에 따라 달라질 수 있다. 그럼에도 불구하고 회귀모형에서는 좋든 싫든 간에 '대표본' 또는 '점근적인' 결과에 대해 논의가 자주 이루어지고 있다.

2.7 오차항의 분산에 대한 추정

무작위 오차항의 분산인 σ^2은 추정되어야 할 단순 선형회귀 모형의 미지의 모수이다. 무작위 오차 e_i의 분산은 다음과 같다.

$$\text{var}(e_i) = \sigma^2 = E[e_i - E(e_i)]^2 = E(e_i^2)$$

위의 관계가 성립되기 위해서는 가정 $E(e_i) = 0$이 준수되어야 한다. '기대'는 평균값이므로 σ^2은 제곱한 오차의 평균이라고 추정할 수 있다.

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{N}$$

위의 식은 무작위 오차항 e_i를 관찰할 수 없으므로 불행하게도 유용하게 사용할 수 없다! 그러나 무작위 오차 자체는 알려져 있지 않지만 이와 유사한 것, 즉 최소제곱 잔차를 구할 수는 있다. 무작위 오차는 다음과 같다는 점에 유의하자.

$$e_i = y_i - \beta_1 - \beta_2 x_i$$

(2.6)에서 최소제곱 잔차는 미지의 모수를 최소제곱 추정값으로 대체하여 구할 수 있다.

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

다음과 같은 결과를 얻기 위해서 무작위 오차 e_i를 이와 유사한 최소제곱 잔차로 대체시키는 일은 합리적인 것처럼 보인다.

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N}$$

위의 추정량은 대규모 표본에서는 만족스러운 결과를 얻을 수 있을지 모르지만 σ^2의 편의된 추정량이다. 하지만 간단한 수정을 통해 불편 추정량을 구할 수 있으며 이는 다음과 같다.

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} \tag{2.19}$$

분모에서 뺀 '2'는 모형에 포함된 회귀모수(β_1, β_2)의 수를 의미하며 이를 통해 추정량 $\hat{\sigma}^2$는 불편하게 되어 $E(\hat{\sigma}^2) = \sigma^2$이 성립된다.

2.7.1 최소제곱 추정량에 대한 분산 및 공분산의 추정

오차 분산의 불편 추정량을 구하기 위해 최소제곱 추정량 b_1 및 b_2의 분산과 이들의 공분산을 추정할 수 있다. (2.14)−(2.16)에 있는 미지의 오차 분산 σ^2를 $\hat{\sigma}^2$로 대체시키면 다음과 같은 결과를 얻을 수 있다.

$$\widehat{\text{var}(b_1)} = \hat{\sigma}^2 \left[\frac{\sum x_i^2}{N\sum (x_i - \bar{x})^2} \right] \tag{2.20}$$

$$\widehat{\text{var}(b_2)} = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \tag{2.21}$$

$$\widehat{\text{cov}(b_1, b_2)} = \hat{\sigma}^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right] \tag{2.22}$$

추정된 분산의 제곱근은 b_1 및 b_2의 '표준오차'이다. 이 값은 가설 검정과 신뢰구간을 구하는

데 사용된다. 이를 $\text{se}(b_1)$ 및 $\text{se}(b_2)$로 나타내면 다음과 같다.

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}}(b_1)} \tag{2.23}$$

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2)} \tag{2.24}$$

2.7.2 식료품 지출액 자료에 관한 계산

식료품 지출액 자료를 이용하여 위에서 다룬 내용을 계산해 보자. 식료품 지출액 모형에 있는 모수의 최소제곱 추정값은 그림 2.9에 있다. 우선 (2.6)의 최소제곱 잔차를 계산하고 나서 이를 이용하여 (2.19)의 오차 분산의 추정값을 계산할 것이다. 표 2.3은 표 2.1에 있는 처음 다섯 가구의 최소제곱 잔차를 보여 주고 있다.

식료품 지출액 자료에 대해 적합한 최소제곱 회귀선은 $\hat{y} = 83.42 + 10.21x$라고 추정하였다. 각 관찰값에 대한 최소제곱 잔차 $\hat{e}_i = y_i - \hat{y}_i$를 계산해 보자. $N = 40$인 관찰값 모두에 대한 잔차를 이용하여 오차 분산을 추정하면 다음과 같다.

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{304505.2}{38} = 8013.29$$

위의 식에서 분자 304505.2는 제곱한 최소제곱 잔차의 합이며 그림 2.9에서 'Sum squared resid'라고 명명되어 있다. 분모는 표본 관찰값의 수, $N = 40$에서 추정된 회귀모수의 수, 2를 감한 것이다. $N - 2 = 38$을 보통 '자유도'라고 하는데 그 이유는 제3장에서 살펴볼 것이다. 그림 2.9에는 $\hat{\sigma}^2$의 값이 포함되어 있지 않다. 대신에 EViews 소프트웨어는 'standard error of the regression'을 의미하는 'S.E. of regression'라고 명명된 수식 $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{8013.29} = 89.517$을 포함하고 있다.

특별히 지시하지 않으면 소프트웨어는 일반적으로 추정된 분산 및 공분산을 제공하지 않는

표 2.3 최소제곱 잔차

x	y	$\hat{y}$	$\hat{e} = y - \hat{y}$
3.69	115.22	121.09	−5.87
4.39	135.98	128.24	7.74
4.75	119.34	131.91	−12.57
6.03	114.96	144.98	−30.02
12.47	187.05	210.73	−23.68

다. 하지만 모든 소프트웨어 패키지는 자동적으로 표준오차를 제공한다. 예를 들면 그림 2.9에 있는 EViews의 분석 결과에는 'Std. Error'라고 명명된 행에 $\text{se}(b_1) = 43.41$ 그리고 $\text{se}(b_2) = 2.09$가 포함되어 있다. 'S.D. dependent var'라고 기재된 칸에는 y의 표본 표준편차, 즉 $\sqrt{\sum(y_i - \bar{y})^2/(N-1)} = 112.6752$가 있다.

회귀에 대한 추정된 분산 및 공분산의 전체값은 사용하는 소프트웨어에 따라 간단한 컴퓨터 명령어나 선택사항을 포함시킴으로서 구할 수 있다. 이들은 직사각형의 배열식이나 행렬식으로 나타내지며 대각선에는 분산, '대각선 밖'에는 공분산이 각각 위치한다. 이를 나타내면 다음과 같다.

$$\begin{bmatrix} \widehat{\text{var}(b_1)} & \widehat{\text{cov}(b_1, b_2)} \\ \widehat{\text{cov}(b_1, b_2)} & \widehat{\text{var}(b_2)} \end{bmatrix}$$

식료품 지출액 자료에 대한 최소제곱 추정량의 추정된 공분산 행렬식은 다음과 같다.

	C	*INCOME*
C	1884.442	−85.90316
INCOME	−85.90316	4.381752

여기서 *C*는 '상수항'을 의미하며 회귀식의 추정된 절편항의 모수 또는 b_1이다. 또한 소프트웨어는 추정된 기울기 b_2와 관련된 행에 변수 *INCOME*을 보여 주고 있다. 따라서 다음과 같다.

$$\widehat{\text{var}(b_1)} = 1884.442, \quad \widehat{\text{var}(b_2)} = 4.381752, \quad \widehat{\text{cov}(b_1, b_2)} = -85.90316$$

표준오차는 다음과 같다.

$$\text{se}(b_1) = \sqrt{\widehat{\text{var}(b_1)}} = \sqrt{1884.442} = 43.410$$

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}(b_2)}} = \sqrt{4.381752} = 2.093$$

제3장에서 이 값들을 폭넓게 이용할 것이다.

주요 용어

가정
가우스–마코프 정리
계량경제모형
경제모형
단순선형 회귀함수
동분산
독립변수
모형 설정 오차
무작위 오차항
반복적인 표본추출
불편 추정량
산포도
선형 추정량
이분산
예측
자유도
점근적
종속변수
최우수 선형 불편 추정량(BLUE)
최소제곱 추정값
최소제곱 추정량
최소제곱 원칙
최소제곱 잔차
탄력성
편의된 추정량
평균과의 편차
표본추출 정확성
표본추출 특성
회귀 모형
회귀모수

연습문제

문제

2.1 다음과 같은 5개 관찰값을 생각해 보자. 계산기만을 사용하여 이 문제에 대한 모든 물음에 답을 해야 한다.

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
3	5				
2	2				
1	3				
−1	2				
0	−2				
$\Sigma x_i =$	$\Sigma y_i =$	$\Sigma(x_i-\bar{x}) =$	$\Sigma(x_i-\bar{x})^2 =$	$\Sigma(y_i-\bar{y}) =$	$\Sigma(x_i-\bar{x})(y_i-\bar{y}) =$

(a) 이 표의 빈칸을 채우시오. 마지막 열에 합을 기입하시오. 표본 평균 $\bar{x}$ 및 $\bar{y}$는 무엇인가?

(b) (2.7) 및 (2.8)을 이용하여 b_1 및 b_2를 계산하고 그 의미를 설명하시오.

(c) $\sum_{i=1}^{5} x_i^2$, $\sum_{i=1}^{5} x_i y_i$를 계산하시오. 이 값을 이용하여 다음을 보이시오.

$$\Sigma(x_i-\bar{x})^2 = \Sigma x_i^2 - N\bar{x}^2$$

$$\Sigma(x_i-\bar{x})(y_i-\bar{y}) = \Sigma x_i y_i - N\bar{x}\bar{y}$$

(d) (b)의 최소제곱 추정값을 이용하여 y의 적합한 값을 계산하고, 다음 표의 빈 부분을 채우시오. 마지막 열에 합을 기입하시오.

(e) 그래프상에 자료의 점들을 표시하고 적합한 회귀선 $\hat{y}_i = b_1 + b_2 x_i$를 그리시오.

(f) (e)의 그래프상에 평균 점 $(\bar{x}, \bar{y})$를 그리시오. 적합하게 그은 선이 이 점을 통과하는가? 만일 그렇지 않다면 그래프상에 다시 정확하게 그려 보시오.

(g) 이 숫자 값에 대해 $\bar{y} = b_1 + b_2\bar{x}$를 보이시오.

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
3	5				
2	2				
1	3				
−1	2				
0	−2				
$\Sigma x_i =$	$\Sigma y_i =$	$\Sigma \hat{y}_i =$	$\Sigma \hat{e}_i =$	$\Sigma \hat{e}_i^2 =$	$\Sigma x_i\hat{e}_i =$

(h) 이 숫자 값에 대해 $\bar{\hat{y}} = \bar{y}$를 보이시오. 여기서 $\bar{\hat{y}} = \Sigma\hat{y}_i/N$이다.

(i) $\hat{\sigma}^2$를 계산하시오.

(j) $\widehat{\text{var}(b_2)}$를 계산하시오.

2.2 어떤 가계의 주당 소득은 $1,000이다. 이런 소득을 갖고 있는 가계의 식료품에 대한 평균 주당 지출액은 $E(y|x = \$1{,}000) = \mu_{y|x=\$1{,}000} = \$125$이며 이 지출액에 대한 분산은 $\text{var}(y|x = \$1{,}000) = \sigma^2_{y|x=\$1000} = 49$이다.

(a) 주당 식료품 지출액은 정규분포한다고 가정하고서 위와 같은 소득을 갖는 가계가 일주일 동안 식료품에 대한 지출이 $110에서 $140 사이일 확률을 구하시오. 귀하 해법을 그래프를 그려 설명하시오.

(b) 주당 식료품 지출액의 분산이 $\text{var}(y|x = \$1{,}000) = \sigma^2_{y|x=\$1{,}000} = 81$인 경우 (a)에서의 확률을 구하시오.

2.3 x와 y의 다음과 같은 관찰값을 그래프 용지에 도표로 나타내시오.

x	1	2	3	4	5	6
y	4	6	7	7	9	11

(a) 자를 이용하여 자료에 적합한 선을 긋고 그 선의 기울기 및 절편을 구하시오.

(b) 식 (2.7)과 (2.8)을 이용하고 계산기를 사용하여 기울기 및 절편의 최소제곱 추정값을 계산하시오. 이 선을 도표에 그리시오.

(c) 표본평균 $\bar{y} = \Sigma y_i/N$과 $\bar{x} = \Sigma x_i/N$을 구하시오. $x = \bar{x}$에서의 y의 예측값을 구하여 도표에 나타내시오. 예측값을 통하여 무엇을 관찰할 수 있는가?

(d) (b)의 최소제곱 추정값을 이용하여 최소제곱 잔차 $\hat{e}_i$를 계산하고 이들의 합을 구하시오.

(e) $\Sigma x_i\hat{e}_i$를 계산하시오.

해답

(b) $\Sigma x_i = 21 \quad \Sigma y_i = 44 \quad \Sigma(x_i - \bar{x})(y_i - \bar{y}) = 22$
$\Sigma(x_i - \bar{x})^2 = 17.5 \quad b_2 = 1.257 \quad b_1 = 2.9333$

(c) $\bar{y} = 7.333 \quad \bar{x} = 3.5$ $x = \bar{x}$에서 y의 예측값은 $\bar{y}$이다.

(d) $\hat{e}_1 = -0.19048 \quad \hat{e}_2 = 0.55238 \quad \hat{e}_3 = 0.29524$
$\hat{e}_4 = -0.96190 \quad \hat{e}_5 = -0.21905 \quad \hat{e}_6 = 0.52381$
$\Sigma\hat{e}_i = 0$

(e) $\Sigma x_i\hat{e}_i = 0$

2.4 단순 선형회귀 모형은 $y = \beta_1 + \beta_2 x + e$이라고 정의한다. 그러나 $\beta_1 = 0$이라는 사실을 알고 있다고 가정하자.

(a) $\beta_1 = 0$인 경우 선형회귀 모형은 대수적으로 어떠한가?

(b) $\beta_1 = 0$인 경우 선형회귀 모형은 도표상에서 어떠한가?

(c) $\beta_1 = 0$인 경우 최소제곱의 '제곱을 합한' 함수는 $S(\beta_2) = \sum_{i=1}^{N}(y_i - \beta_2 x_i)^2$이 된다.

x	1	2	3	4	5	6
y	4	6	7	7	9	11

위의 자료를 이용하여 대략적인 최소값을 알아볼 수 있도록 β_2값들에 대한 제곱을 합한 함수값을 도표로 나타내시오. $S(\beta_2)$를 최소화하는 β_2값은 어떤 의미를 갖는

가?[요령 : 괄호 안에 있는 항을 제곱하고 합산을 하여 대수적으로 $S(\beta_2) = \sum_{i=1}^{N}(y_i - \beta_2 x_i)^2$과 같이 정리할 수 경우 귀하는 간단히 계산할 수 있을 것이다.]

(d) 미분법을 이용하여 이 모형에 있는 β_2의 최소제곱 추정값에 대한 공식이 $b_2 = \sum x_i y_i / \sum x_i^2$라는 사실을 보이시오. 이를 이용하여 b_2를 계산하고 이 값을 기하학적으로 구한 값과 비교하시오.

(e) (d)의 공식을 갖고 구한 추정값을 이용하여 적합한 (추정된) 회귀함수를 도표로 그리시오. 도표상에 점$(\bar{x}, \bar{y})$를 나타내시오. 귀하는 무엇을 관찰할 수 있는가?

(f) (d)의 공식을 갖고 구한 추정값을 이용하여 최소제곱 잔차 $\hat{e}_i = y_i - b_2 x_i$를 구하시오. 이들의 합을 구하시오.

(g) $\sum x_i \hat{e}_i$를 계산하시오.

2.5 한 중소기업체는 자신들의 주당 광고비가 주당 $600로 증가할 경우의 주당 매출액을 예측하기 위해 경영진단사를 고용하였다. 경영진단사는 지난 6개월 동안 해당기업의 주당 광고비 지출액과 이에 따른 주당 매출액을 기록하였다. 경영진단사는 자신의 보고서에 다음과 같이 기술하였다. "지난 6개월 동안 주당 평균 광고비는 $450이었으며 주당 평균 매출액은 $7,500이었다. 단순 선형회귀분석 결과에 기초하여 주당 광고비가 $600일 경우 매출액은 $8,500가 될 것으로 예상된다."

(a) 위와 같은 예측을 하기 위해 경영진단사가 사용한 추정된 단순회귀는 무엇인가?

(b) 추정된 회귀선을 도표로 나타내시오. 도표상에 주당 평균값을 나타내시오.

2.6 루이지애나 주립대학교 미식축구 경기에서 청량음료를 판매하는 업자는 경기할 때 기온이 상승하면 더 많은 청량음료를 판매할 수 있다는 사실을 알게 되었다. 5년 동안 치루어진 32개 초청경기에 기초하여 청량음료 판매와 온도의 관계를 다음과 같이 추정하였다. $\hat{y} = -240 + 6x$이며 여기서 y = 판매한 청량음료의 개수이며, x = 화씨로 측정한 온도이다.

(a) 추정한 기울기 및 절편의 의미를 해석하시오. 추정값이 의미가 있는가? 왜 그런가? 또는 왜 그렇지 않은가?

(b) 게임할 때 온도가 화씨 80도가 될 것이라고 예상되는 날 얼마나 많은 청량음료를 판매할 것으로 예측할 수 있는가?

(c) 온도가 몇 도 아래로 떨어질 경우 청량음료를 판매할 수 없다고 예측할 수 있는가?

(d) 추정된 회귀선을 도표로 나타내시오.

해답

(a) 절편 추정값 $b_1 = -240$은 온도가 0°F일 때 판매되는 청량음료의 수에 대한 추정값이다. −240개의 청량음료가 판매된다는 것은 불가능하므로 이 추정값은 의미 있는 것으로 받아들여지지 말아야 한다. 기울기 추정값 $b_2 = 6$은 온도가 1°F씩 증가할 때 판매되므로 청량음료의 증가에 대한 추정값이다. 이 추정값은 의미가 있다. 온도가 상승함에 따라 판매되는 청량음료의 수도 증가할 것으로 기대된다.

(b) 예측되는 판매 청량음료의 수는 $\hat{y} = -240 + 6 \times 80 = 240$이다.

(c) 청량음료가 판매되지 않을 경우 $y = 0$ 그리고 $0 = -240 + 6 \times x$, 즉 $x = 40$이다. 따라서 40°F 아래가 되면 청량음료가 판매되지 않을 것으로 예측된다.

2.7 미국 50개 주와 콜럼비아 특별구의 자료 $N = 51$개 관찰값에 기초한 단순회귀분석의 결과가 있다고 하자.

(a) 추정된 오차 분산이 $\hat{\sigma}^2 = 2.04672$이다. 제곱한 최소제곱 잔차의 합은 얼마인가?

(b) b_2의 추정된 분산은 0.00098이다. b_2의 표준오차는 얼마이며, $\sum(x_i - \bar{x})^2$의 값은 얼마인가?

(c) 종속변수 y_i = 18세 이상 남성의 주별 평균소득(천$)이며, x_i = 고등학교를 졸업한 18세 이상 남성의 백분율이다. $b_2 = 0.18$인 경우 이것이 의미하는 바를

설명하시오.

(d) $\bar{x} = 69.139$이고 $\bar{y} = 15.187$이라면 절편 모수의 추정값은 얼마인가?

(e) (b)와 (d)에 대한 대답을 알고 있는 경우 $\sum x_i^2$은 얼마인가?

(f) 아르칸사스주의 경우 $y_i = 12.274$, $x_i = 58.3$이다. 아르칸사스주의 최소제곱 잔차를 계산하시오. [요령 : (c)와 (d)에서 얻은 정보를 이용하시오.]

2.8 스터프(E. Z. Stuff) 교수는 최소제곱 추정량이 많은 문제점이 있다고 결론을 내렸다. 두 점은 선을 그을 수 있다는 사실에 유의하여 스터프 박사는 크기가 N인 표본의 두 점을 골라서 이들을 연결하는 선을 긋고 이 선의 기울기를 단순회귀 모형 b_2의 *EZ*추정량이라고 명명하였다. 대수학적으로 두 점이 (y_1, x_1)과 (y_2, x_2)인 경우 *EZ*추정 규칙은 다음과 같다.

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1}$$

단순회귀 모형의 모든 가정이 준수된다고 가정하고 다음 물음에 답하시오.

(a) b_{EZ}가 '선형' 추정량이라는 점을 보이시오.

(b) b_{EZ}가 불편 추정량이라는 점을 보이시오.

(c) b_{EZ}의 분산을 구하시오.

(d) b_{EZ}의 확률분포를 구하시오.

(e) 스터프 교수에게 자신이 고안한 추정량이 최소제곱 추정량만큼 좋지 않다는 사실을 확산시키시오. 증명을 할 필요는 없다.

컴퓨터를 이용한 문제

2.9 흥미로우면서도 유용한 경제개념 중 하나는 '학습곡선'이다. 이 개념은 자동차 산업이나 작업이 반복적으로 이루어지는 경우와 같은 조립라인 생산과정에서 발생하는 현상이다. 노동자들은 경험을 통해 배우게 되며 자신들의 업무를 수행하면서 점점 더 효율적이 된다. 이는 최종재를 생산하는 데 더 적은 시간과 노동비용이 투입된다는 의미이다. 이 개념은 시간 $t(UNITCOST_t)$에서의 단위당 비용을 포함되지 않은 시간 $t(CUMPROD_t)$까지의 누적 상품생산과 연계시키는 경제모형의 기초가 된다. 이 변수들 간의 관계는 자주 다음과 같이 표현된다.

$$UNITCOST_t = UNITCOST_1 \times CUMPROD_t^{\varepsilon}$$

여기서 $UNITCOST_1$는 최초 생산단위에 대한 단위당 생산비용이며 ε는 누적 생산에 대한 단위비용의 (음일 것으로 기대되는) 탄력성이다. 이런 변수들 간의 비선형관계는 양편에 대수를 취하여 다음과 같은 선형관계로 전환시킬수 있다.

$$\ln(UNITCOST_t) = \ln(UNITCOST_1) + \varepsilon \ln(CUMPROD_t)$$

$$= \beta_1 + \beta_2 \ln(CUMPROD_t)$$

모형을 좀 더 일반적인 형태로 변형시키기 위해 $\ln(UNITCOST_1)$과 ε의 이름을 다시 부칠 것이다. 언스트 번트(Ernst Berndt)는 이 책보다 수준이 높고 내용이 알찬 『The Practice of Econometrics: Classic and Contemporary』(Addison and Wesley, 1991)의 저자이다. 번트는 상기서 85페이지에서 도료의 침전제로 사용되는 이산화티타늄이라 불리는 물질의 생산과정에서 얻은 학습의 예를 들고 있다. 그는 1955년부터 1970년까지 듀퐁사로부터 얻은 생산 및 단위 비용에 관한 자료를 제공하였고 관련 자료는 파일 *learn.dat*에 있다.

(a) 컴퓨터 소프트웨어를 사용하여 *CUMPROD*에 대한 *UNITCOST*의 도표 및 $\ln(CUMPROD)$에 대한 $\ln(UNITCOST)$의 도표를 그리시오.

(b) β_1 및 β_2의 최소제곱 추정값 b_1 및 b_2를 구하고 이들의 경제적 의미를 설명하시오. 구한 값이 이치에 맞는가? (a)의 도표상에 손으로 또는 소프트웨어를 이용하여 적합한 회귀선을 그리시오.

(c) 최소제곱 추정량의 추정된 분산 및 공분산을 구하시오.

(d) $\hat{\sigma}^2$을 구하시오.

(e) 누적된 생산이 $CUMPROD_0 = 2000$인 경우 단위 생산비용을 예측하시오.

해답

(b) $b_1 = 6.0191$ 및 $b_2 = -0.3857$. 첫 번째 단위를 생산하는 비용은 다음과 같다. $\widehat{UNITCOST}_1 = \exp(b_1) = \exp(6.0191) = 411.208$. 추정값 $b_2 = -0.3857$에 따르면 누적생산이 1% 증가할 경우 비용이 0.386%만큼 감소한다. 숫자 값들은 의미가 있다.

(c) $\widehat{\text{var}(b_1)} = 0.075553$ $\widehat{\text{var}(b_2)} = 0.001297$

$\widehat{\text{cov}(b_1, b_2)} = -0.009888$

(d) $\hat{\sigma}^2 = 0.049930^2 = 0.002493$

(e) $\widehat{\ln(UNITCOST_0)} = 6.0191 - 0.0385\ln(2000) = 3.0874526$

$\widehat{UNITCOST_0} = \exp(3.0874526) = 21.921$

2.10 자본자산 가격결정 모형(CAPM)은 재무 분야에서 중요한 모형으로 주식 수익률의 변화를 공식적으로 거래되는 모든 주식으로 구성된 포트폴리오, 즉 시장 포트폴리오 수익률의 함수로서 설명한다. 일반적으로 투자에 대한 수익률은 위험이 없는 자산에 대한 수익인 기회비용에 대해 측정된다. 이 둘 사이의 차이가 위험한 투자에 대한 보상이 될 수도 있고 손실이 될 수도 있기 때문에 위험할증이라 한다.

CAPM에 따르면 주식 j에 대한 위험할증은 시장 포트폴리오에 대한 위험할증에 비례하며 다음과 같이 나타낼 수 있다.

$$r_j - r_f = \beta_j(r_m - r_f)$$

여기서 r_j 및 r_f는 각각 주식 j에 대한 수익률과 위험이 없는 자산에 대한 수익률을 나타낸다. r_m은 시장 포트폴리오에 대한 수익률을 의미하며 β_j는 j번째 주식의 베타값을 의미한다. 주식의 베타는 해당 주식의 변동을 나타내므로 투자자에게 중요한 의미를 갖으며 전체 주식시장의 변동에 대한 주식 j 수익의 민감도를 측정한다. 따라서 베타값이 1보다 작은 경우 해당 주식의 변동이 시장의 변동보다 작기 때문에 '방어적'이 되는 반면에 베타값이 1보다 큰 경우 '공격적 주식'이 된다. 투자자들은 보통 주식을 매입하기 전에 해당 주식의 베타 추정값을 알고자 하며 이 경우 위에서 살펴본 CAPM 모형은 '경제모형'이 된다. 경제모형에 (이론에 의하면 0이 되어야 하는) 절편과 오차항을 포함시키면 다음과 같은 '계량경제모형'을 구할 수 있다.

$$r_j - r_f = \alpha_j + \beta_j(r_m - r_f) + e$$

(a) 위의 계량경제모형이 왜 이 장에서 살펴본 단순회귀모형인지 설명하시오.

(b) 자료 파일 *capm2.dat*에는 6개 기업(마이크로 소프트, GE, GM, IBM, 디즈니, 모빌-엑손)의 월간 수익, 시장 포트폴리오에 대한 수익률(*MKT*), 위험이 없는 자산에 대한 수익률(*RKFREE*)이 있다. 1995년 1월부터 2004년 12월까지 120개 관찰값이 있다. 각 기업에 대한 CAPM 모형을 추정하고 추정된 베타 값에 대해 논의하시오. 어느 기업이 가장 공격적인 것처럼 보이는가? 어느 기업이 가장 방어적인 것처럼 보이는가?

(c) 재무이론에 따르면 절편의 모수인 α_j는 0이 되어야만 한다. 귀하의 추정값에 의하면 이는 타당한 것처럼 보이는가? 마이크로 소프트 주식에 대해 자료의 산포도를 따라 적합한 회귀선을 그리시오.

(d) $\alpha_j = 0$이라는 가정하에서 모형을 추정하시오. 주식의 베타 추정값이 크게 변화하는가?

2.11 파일 *br2.dat*에는 2005년 중반에 미국 루이지애나주 바톤 루즈시에서 판매된 1080개 주택에 관한 자료가 포함되어 있다. 이 자료에는 주택 판매가격, 제곱 피트

로 나타낸 주택 규모, 주택의 연령, 수영용 풀장의 존재 여부, 벽난로의 존재 여부, 선창가에 위치하는지 여부가 포함되어 있다. 또한 부동산업자의 설명 속에는 *style*이라고 명명된 변수가 있다. 파일 *br2.def*에는 변수에 대한 설명이 있다.

(a) 표본에 있는 모든 주택에 관해서 주택 규모에 대한 주택 가격을 도표로 나타내시오. 전통형태의 주택에 관해서 이와 같은 도표를 그리시오.

(b) 표본의 모든 가구에 대해서 회귀모형 $PRICE = \beta_1 + \beta_2 SQFT + e$을 추정하시오. 적합한 선을 그려 보시오.

(c) 전통형태의 주택에 대해서만 (b)와 같은 회귀모형을 추정하시오. 추정값의 의미를 설명하시오. 이 추정값을 (b)의 추정값과 비교해 보면 어떤 것처럼 보이는가?

(d) (b) 및 (c)에서 추정한 각각의 회귀식에 대해 최소제곱 잔차를 계산하고 *SQFT*에 대해 이 잔차를 도표로 나타내시오. 가정 중 어느 것이 위배된 것처럼 보이는가?

2.12 파일 *stockton2.dat*에는 2005년 중반 동안 미국 캘리포니아주 스톡톤시에서 판매된 880개 주택에 관한 자료가 포함되어 있다. 변수들에 관한 설명은 파일 *stockton2.def*에 있다.

(a) 표본의 모든 주택에 관해 주택 규모에 대한 주택 가격을 도표로 나타내시오.

(b) 표본의 모든 주택에 대해 회귀모형 $PRICE = \beta_1 + \beta_2 SQFT + e$을 추정하시오. 추정값이 갖는 의미를 설명하시오. 적합한 선을 그리시오.

(c) 판매 당시에 사람이 거주하지 않고 비어 있던 주택에 대해서만 (b)의 회귀모형을 추정하시오. 판매 당시에 (비어 있지 않고) 사람이 거주하고 있었던 주택에 대해서 동일한 회귀모형을 추정하시오. 이 추정값들을 서로 비교해 보면 어떤 것처럼 보이는가?

(d) (c)에서 추정한 각각의 회귀식에 대해 최소제곱 잔차를 계산하고 *SQFT*에 대해 이 잔차를 도표로 나타내시오. 가정 중 어느 것이 위배된 것처럼 보이는가?

(e) 거주면적이 2,000 제곱피트인 주택의 가격을 예측하시오.

해답

(b) $\widehat{PRICE} = -18{,}386 + 81.389SQFT$. 계수 $b_2 = 81.389$에 따르면 주택 규모가 제곱피트 증가할 때마다 주택 가격이 약 \$81 증가한다는 의미이다. 말로 표현하면 절편 $b_1 = -\$18{,}386$은 면적이 0 제곱피트인 주택의 가격은 $-\$18{,}386$이 된다는 의미이다. 따라서 이 모형은 면적이 0 제곱피트인 경우 의미 있는 것으로 받아들여지지 않는다.

(c) 빈 주택의 경우 : $\widehat{PRICE} = -4793 + 69.908SQFT$.
사람이 거주하는 주택의 경우 : $\widehat{PRICE} = -27{,}169 + 89.259SQFT$.
1 제곱피트 증가할 때마다 한계비용(가격의 변화)은 빈 주택의 경우보다 사람이 거주하는 주택의 경우에 더 높다. 제곱피트당 평균가격은 주택크기가 1156 제곱피트보다 작은 경우 사람이 거주하는 주택보다 빈주택의 경우 더 높다. 1156 제곱피트보다 더 큰 주택의 경우 제곱피트당 가격은 사람이 거주하는 주택에서 더 높다.

(d) 잔차의 크기는 주택 규모가 더 큰 주택에서 더 커지는 경향이 있으며 이는 SR3 $\text{var}(e|x_i) = \sigma^2$(동분산 가정)이 위배된 것처럼 보인다.

2.13 사람들은 신규주택 건설 및 판매가 주택담보 장기대출의 이자율에 의존한다고 생각한다. 이자율이 높을 경우 더 적은 수의 사람들이 신규주택을 구입하는데 필요한 자금을 차용하게 된다. 주택 건설업자들도 이런 사실을 알고 있기 때문에 주택담보 장기대출의 이자율이 높을 경우 신규 주택을 건설하는데 덜 적극적이다. 이는 직관적인 추론이므로 다음과 같은 질문을 해 보자.

"주택담보 장기대출의 이자율이 1% 인상될 경우 주

택건설은 얼마나 감소하는가?" 30년 만기 주택담보 대출의 고정 이자율, 주택 착공 건수 (1000채), 주택 판매 건수(1000채)는 파일 *house_starts.dat*에 있다. 1990년 1월부터 2005년 4월까지 184개의 월간 관찰값이 수록되어 있다.

(a) 시간에 대해 위의 자료 각각을 도표로 나타내시오.

(b) 30년 만기 주택담보 대출의 고정이자율(*FIXED_RATE*)에 대해 주택 착공 건수(*STARTS*)를 도표로 나타내시오.

(c) *FIXED_RATE*에 대한 *STARTS*의 단순회귀모형을 추정하시오. 회귀가 자료에 얼마나 적합한지에 대해 언급하고 분석 결과가 갖는 의미에 대해 설명하시오. (b)에서 구한 도표에 적합한 회귀선을 그리시오.

(d) *FIXED_RATE*에 대해 주택 판매 건수(*SOLD*)를 도표로 나타내시오.

(e) *FIXED_RATE*에 대한 *SOLD*의 단순회귀 모형을 추정하시오. 회귀가 자료에 얼마나 적합한지에 대해 언급하고 분석 결과가 같은 의미에 대해 설명하시오. (d)에서 구한 도표에 적합한 회귀선을 그리시오.

(f) 30년 만기 주택담보 대출의 고정이자율이 6%인 경우 월간 주택 착공건수를 예측하시오.

2.14 레이 페어(Ray C. Fair) 교수는 오랫동안 미국의 대선 결과를 설명하고 예측하는 모형을 만들고 갱신하였다. 이에 관해서는 다음과 같은 그의 웹사이트와 논문을 참조해 보자. http://fairmodel.econ.yale.edu/vote2008/index2.htm(웹사이트) "A Vote Equation for the 2004 Election"(논문). 이 모형의 기본적인 전제는 양당(민주당 및 공화당)에 대한 일반 투표중 집권당(선거 당시의 집권당)이 차지하는 몫은 예를 들면 집권당이 얼마나 오랫동안 집권을 하고 있는지 그리고 현직 대통령이 재선을 노리고 있는지와 같이 경제 및 정치와 관련된 많은 변수들에 의해 영향을 받는다는 것이다. 페어 교수의 1880년부터 2000년까지의 선거에 대한 31개 관찰값은 파일 *fair.dat*에 있다. 종속변수는 *VOTE* = 집권당이 획득한 일반 투표의 백분율 몫이다. 설명변수로는 *GROWTH* = 선거 당해 연도의 처음 3분기 동안의 1인당 실질 GDP의 (연간) 성장률을 생각해 볼 수 있다. 경제상황이 좋아서 성장률이 높은 경우 집권당이 선거에서 승리할 확률이 더 높아질 것이라고 예상할 수 있다.

(a) *GROWTH*에 대한 *VOTE*의 도표를 그리시오. 양의 관계가 있는 것처럼 보이는가?

(b) 1880년부터 2000년까지의 모든 자료에 기초하고 최소제곱법을 이용하여 회귀모형 $VOTE = \beta_1 + \beta_2 GROWTH + e$을 추정하시오. 추정결과를 작성하고 논의하시오. (a)에서 구한 산포도상에 적합한 선을 손으로 그리시오.

(c) 경제 전반적인 인플레이션은 선거에서 집권당에게 불리하게 작용할 수 있다. 변수 *INFLATION*은 행정부의 처음 15분기 동안의 물가 상승을 의미한다. *INFLATION*에 대한 *VOTE*를 도표로 나타내시오. 추정결과를 작성하고 논의하시오.

해답

(b) $\widehat{VOTE} = 51.939 + 0.660 GROWTH$

(c) $\widehat{VOTE} = 53.496 - 0.445 INFLATION$

2.15 교육은 임금률에 얼마나 영향을 미치는가? 자료 파일 *cps_small.dat*에는 미국의 1997년 Current Population Survey(CPS)에서 구한

시간당 임금률, 교육, 기타 다른 변수들에 관한 1000개 관찰값이 있다.

(a) 변수 *WAGE* 및 *EDUC*에 관한 요약된 통계량과 막대그래프를 구하시오.

(b) 선형 회귀식 $WAGE = \beta_1 + \beta_2 EDUC + e$을 추정하고 추정결과에 대해 논의하시오.

(c) 최소제곱 잔차를 계산하고 이를 *EDUC*에 대해 도표로 나타내시오. 어떤 일정한 형태를 띠고 있는가? 가정 SR1−SR5 가 준수될 경우 최소제곱 잔차에는 어떤 일정한 형태가 존재하게 되는가?

(d) 남성, 여성, 흑인, 백인에 대해 각각 별개의 회귀식을 추정하시오. 추정결과를 비교하시오.

부록 2A 최소제곱 추정값의 도출

y 및 x에 대한 표본 관찰값이 주어진 경우 다음과 같은 '제곱의 합' 함수를 극소화하는 미지의 모수 β_1 및 β_2의 값을 구하고자 한다.

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2 \tag{2A.1}$$

점(y_i, x_i)는 관찰할 수 있으므로 '제곱의 합' 함수 S는 미지의 모수 β_1 및 β_2에만 의존한다. 미지의 모수 β_1 및 β_2에 대해 이차방정식인 이 함수는 그림 2A.1에서 보는 것처럼 '사발 모양의 표면'을 갖고 있다.

미분법과 '편미분법'에 익숙한 사람들은 β_1 및 β_2에 관한 S의 편도함수가 다음과 같다는 사실을 입증할 수 있을 것이다.

$$\frac{\partial S}{\partial \beta_1} = 2N\beta_1 - 2\sum y_i + 2(\sum x_i)\beta_2$$

$$\frac{\partial S}{\partial \beta_2} = 2(\sum x_i^2)\beta_2 - 2\sum x_i y_i + 2(\sum x_i)\beta_1 \tag{2A.2}$$

위의 도함수들은 축방향으로 향한 사발 모양 표면의 기울기를 나타내는 식이다. 직관적으로 '사발의 밑바닥'은 축방향으로 향한 사발의 기울기, 즉 $\partial S/\partial \beta_1$과 $\partial S/\partial \beta_2$가 0이 되는 곳이다.

대수적으로 점 (b_1, b_2)를 구하기 위해서 (2A.2)를 0이라 놓고 β_1 및 β_2를 각각 b_1 및 b_2로 대체시키면 다음의 결과를 얻을 수 있다.

$$2[\sum y_i - Nb_1 - (\sum x_i)b_2] = 0$$

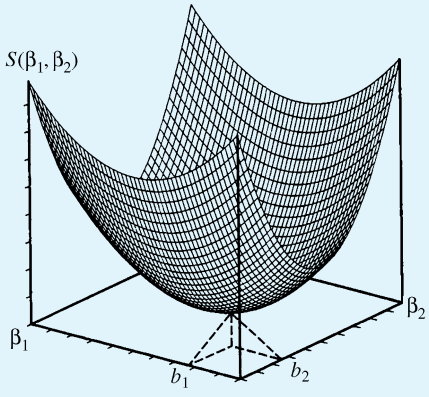


그림 2A.1
제곱함수의 합과 최소값 b_1 및 b_2

$$2[\Sigma x_i y_i - (\Sigma x_i)b_1 - (\Sigma x_i^2)b_2] = 0$$

위의 식을 재정리하면 정규 방정식이라고 일반적으로 알려진 다음의 두 식을 구할 수 있다.

$$Nb_1 + (\Sigma x_i)b_2 = \Sigma y_i \tag{2A.3}$$

$$(\Sigma x_i)b_1 + (\Sigma x_i^2)b_2 = \Sigma x_i y_i \tag{2A.4}$$

위의 두 식은 2개의 미지수 b_1 및 b_2에 대해 2개의 선형 방정식으로 구성된다. 이 두 선형 방정식을 b_1 및 b_2에 대해 풀면 최소제곱 추정값을 구할 수 있다. b_2에 대해 풀기 위해서는 첫 번째 식(2A.3)에 Σx_i를 곱하고 두 번째 식 (2A.4)에 N을 곱하여 두 번째 식으로부터 첫 번째 식을 빼서 왼쪽 편에 b_2만을 남겨 놓으면 된다.

$$b_2 = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma x_i^2 - (\Sigma x_i)^2} \tag{2A.5}$$

b_1에 대해 풀기 위해서는 (2A.3)의 양쪽 편을 N으로 나누어 재정리해야 한다.

부록 2B 평균으로부터의 편차로 나타낸 b_2의 형태

b_2에 대한 공식을 (2.7)로 전환하기 위해 제일 먼저 필요한 것은 합산 부호를 포함한 몇 가지 기교를 사용하는 일이다. 첫 번째로 필요한 사실은 다음과 같다.

$$\begin{aligned}\Sigma(x_i - \bar{x})^2 &= \Sigma x_i^2 - 2\bar{x}\Sigma x_i + N\bar{x}^2 = \Sigma x_i^2 - 2\bar{x}\left(N\frac{1}{N}\Sigma x_i\right) + N\bar{x}^2 \\ &= \Sigma x_i^2 - 2N\bar{x}^2 + N\bar{x}^2 = \Sigma x_i^2 - N\bar{x}^2\end{aligned} \tag{2B.1}$$

앞으로 $\Sigma(x_i - \bar{x})^2$를 계산하려는 경우 위에서 살펴본 편리한 공식인 $\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - N\bar{x}^2$를 사

용하는 것이 훨씬 계산이 용이하다. 그러면 다음과 같다.

$$\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - N\bar{x}^2 = \Sigma x_i^2 - \bar{x}\Sigma x_i = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N} \tag{2B.2}$$

위의 결과를 얻기 위해서 $\bar{x} = \Sigma x_i/N$ 따라서 $\Sigma x_i = N\bar{x}$라는 사실을 이용하였다. 두 번째 유용한 사실은 첫 번째와 유사한 것으로 다음과 같다.

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - N\bar{x}\bar{y} = \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{N} \tag{2B.3}$$

위의 결과도 유사한 방법으로 증명할 수 있다.

(2A.5)에 있는 b_2의 분자 및 분모를 N으로 나눌 경우 (2B.1)−(2B.3)을 이용하여 b_2를 평균으로부터 벗어난 편차의 형태인 다음과 같이 재정리할 수 있다.

$$b_2 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$$

b_2에 대한 위의 공식은 앞으로 몇몇 다른 장에서 반복해서 사용할 것이므로 기억을 해 두는 것이 바람직하다.

부록 2C b_2는 선형 추정량이다

(2.10)을 도출하기 위해서는 합산에 관한 다른 특성을 이용하여 이를 더 간략히 정리해야 한다. 평균에 대한 변수의 합은 0이며 다음과 같이 나타낼 수 있다.

$$\Sigma(x_i - \bar{x}) = 0$$

이를 이용하여 b_2에 대한 공식을 다음과 같이 변형시킬 수 있다.

$$\begin{aligned} b_2 &= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{\Sigma(x_i - \bar{x})y_i - \bar{y}\Sigma(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2} \\ &= \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2} = \Sigma\left[\frac{(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}\right] y_i = \Sigma w_i y_i \end{aligned}$$

여기서 w_i는 (2.11)에서 살펴본 것처럼 상수이다.

부록 2D b_2를 이론적인 식으로 나타내기

(2.12)를 도출하기 위해 (2.10)의 y_i를 $y_i = \beta_1 + \beta_2 x_i + e_i$로 대체하여 간략히 해 보자.

$$\begin{aligned} b_2 &= \sum w_i y_i = \sum w_i(\beta_1 + \beta_2 x_i + e_i) \\ &= \beta_1 \sum w_i + \beta_2 \sum w_i x_i + \sum w_i e_i \\ &= \beta_2 + \sum w_i e_i \end{aligned}$$

위의 식을 간략히 하기 위해 또 다른 두 가지 합산에 관한 기교를 이용해야 한다. 첫째, $\sum w_i = 0$으로 이를 통해 $\beta_1 \sum w_i$항을 제거할 수 있다. 둘째, $\sum w_i x_i = 1$로 $\beta_2 \sum w_i x_i = \beta_2$가 된다. 이로 인해 (2.10)을 (2.12)로 간략히 정리할 수 있다. 다음과 같은 이유로 인해 $\sum w_i = 0$가 된다.

$$\sum w_i = \sum \left[\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{\sum (x_i - \bar{x})^2} \sum (x_i - \bar{x}) = 0$$

위의 식의 마지막 단계에서 $\sum (x_i - \bar{x}) = 0$을 이용하였다. $\sum w_i x_i = 1$라는 사실을 보여 주기 위해 $\sum (x_i - \bar{x}) = 0$을 다시 한 번 이용해 보자. $\sum (x_i - \bar{x})^2$을 달리 표현하면 다음과 같다.

$$\begin{aligned} \sum (x_i - \bar{x})^2 &= \sum (x_i - \bar{x})(x_i - \bar{x}) \\ &= \sum (x_i - \bar{x}) x_i - \bar{x} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x}) x_i \end{aligned}$$

따라서 다음과 같아진다.

$$\sum w_i x_i = \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x}) x_i} = 1$$

부록 2E b_2의 분산 도출

출발점은 (2.12) $b_2 = \beta_2 + \sum w_i e_i$이다. 최소제곱 추정량은 분산이 다음과 같은 확률변수이다.

$$\text{var}(b_2) = E[b_2 - E(b_2)]^2$$

(2.12)로 대체시키고 최소제곱 추정량의 불편성 $E(b_2) = \beta_2$을 이용하면 다음과 같다.

$$\begin{aligned} \text{var}(b_2) &= E(\beta_2 + \sum w_i e_i - \beta_2)^2 \\ &= E\left(\sum w_i e_i\right)^2 \\ &= E\left(\sum w_i^2 e_i^2 + 2 \sum\sum_{i \neq j} w_i w_j e_i e_j\right) \qquad \text{(괄호항의 제곱)} \end{aligned}$$

$$= \sum w_i^2 E(e_i^2) + 2\sum\sum_{i \neq j} w_i w_j E(e_i e_j) \quad (w_i\text{가 확률적이 아니기 때문이다.})$$
$$= \sigma^2 \sum w_i^2$$
$$= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

위의 식에서 마지막 줄의 바로 전 단계는 다음과 같은 두 가지 가정을 이용하였다. 첫째, $\sigma^2 = \text{var}(e_i) = E[e_i - E(e_i)]^2 = E(e_i - 0)^2 = E(e_i^2)$이다. 둘째, $\text{cov}(e_i, e_j) = E[(e_i - E(e_i))(e_j - E(e_j))] = E(e_i e_j) = 0$이다. 마지막 단계는 다음과 같은 사실을 이용하였다.

$$\sum w_i^2 = \sum \left[\frac{(x_i - \bar{x})^2}{\left\{ \sum (x_i - \bar{x})^2 \right\}^2} \right] = \frac{\sum (x_i - \bar{x})^2}{\left\{ \sum (x_i - \bar{x})^2 \right\}^2} = \frac{1}{\sum (x_i - \bar{x})^2}$$

다른 방법으로는 합의 분산을 구하는 규칙을 이용할 수 있다. X 및 Y가 확률변수이고 a 및 b가 상수인 경우 다음과 같다.

$$\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \, \text{cov}(X, Y)$$

$b_2 = \beta_2 + \sum w_i e_i$에 기초하면 다음과 같다.

$$\text{var}(b_2) = \text{var}(\beta_2 + \sum w_i e_i) = \text{var}(\sum w_i e_i) \quad (\beta_2\text{는 상수이기 때문이다.})$$
$$= \sum w_i^2 \text{var}(e_i) + \sum\sum_{i \neq j} w_i w_j \, \text{cov}(e_i, e_j) \quad (\text{분산을 구하는 규칙에 따른 것이다.})$$
$$= \sum w_i^2 \text{var}(e_i) \quad (\text{cov}(ce_i, e_j) = 0\text{을 이용한 것이다.})$$
$$= \sigma^2 \sum w_i^2 \quad (\text{var}(e_i) = \sigma^2\text{을 이용한 것이다.})$$
$$= \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

b_2에 대한 분산식의 도출은 가정 SR3 및 SR4에 의존한다는 점을 주목하자. $\text{cov}(e_i, e_j) \neq 0$인 경우 이중합산의 모든 항을 생략할 수 없다. 모든 관찰값에 대해 $\text{var}(e_i) \neq \sigma^2$인 경우 σ^2를 합산 부호 밖으로 분해할 수 없다. 이 가정 중 어느 하나가 준수되지 못할 경우 $\text{var}(b_2)$는 (2.15)로 나타낼 수 없으며 다른 형태를 갖는다. 이는 b_1의 분산 및 공분산에도 적용된다.

부록 2F 가우스–마코프 정리의 증명

β_2의 최소제곱 추정량 b_2에 대해 가우스–마코프 정리를 증명할 것이며 이를 위해 선형 및 불편 추정량들 중에서 추정량 b_2가 가장 작은 분산을 갖는다는 점을 보여 줄 것이다. $b_2^* = \sum k_i y_i$(여기서 k_i는 상수)는 β_2의 다른 선형 추정량이라 하자. 최소제곱 추정량 b_2와 비교를 좀 더 쉽게 하기 위하여 $k_i = w_i + c_i$라 하자. 여기서 c_i는 다른 상수이며 w_i는 (2.11)에 주어져 있다. 이는 편법이기는 하지만 선택한 모든 k_i에 대해 c_i를 구할 수 있으므로 정당하다고 볼 수 있다. 새로운 추정량에 대해 y_i를 대체시키고 w_i에 대한 특성(부록 2D)을 이용하여 단순화시키면 다음과 같다.

$$\begin{aligned} b_2^* &= \sum k_i y_i = \sum (w_i + c_i) y_i = \sum (w_i + c_i)(\beta_1 + \beta_2 x_i + e_i) \\ &= \sum (w_i + c_i)\beta_1 + \sum (w_i + c_i)\beta_2 x_i + \sum (w_i + c_i) e_i \\ &= \beta_1 \sum w_i + \beta_1 \sum c_i + \beta_2 \sum w_i x_i + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \\ &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) e_i \end{aligned} \tag{2F.1}$$

위에서 $\sum w_i = 0$ 및 $\sum w_i x_i = 1$이다.
(2F.1)의 마지막 줄에 대해 수학적인 기댓값을 취하고 나서 기댓값의 특성과 가정을 이용하면 다음과 같다.

$$\begin{aligned} E(b_2^*) &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i + \sum (w_i + c_i) E(e_i) \\ &= \beta_1 \sum c_i + \beta_2 + \beta_2 \sum c_i x_i \end{aligned} \tag{2F.2}$$

선형 추정량 $b_2^* = \sum k_i y_i$이 불편되기 위해서는 다음과 같아야만 한다.

$$\sum c_i = 0 \quad \text{및} \quad \sum c_i x_i = 0 \tag{2F.3}$$

$b_2^* = \sum k_i y_i$이 선형 및 불편 추정량이 되기 위해서는 이 조건들이 준수되어야만 한다. 따라서 조건 (2F.3)가 준수된다고 가정하고 이를 이용하여 (2F.1)을 다음과 같이 단순화시킬 수 있다.

$$b_2^* = \sum k_i y_i = \beta_2 + \sum (w_i + c_i) e_i \tag{2F.4}$$

부록 2E의 단계를 밟고 다음과 같은 추가적인 사실을 이용하여 선형 불편 추정량의 분산을 구할 수 있게 되었다.

$$\sum c_i w_i = \sum \left[\frac{c_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] = \frac{1}{\sum (x_i - \bar{x})^2} \sum c_i x_i - \frac{\bar{x}}{\sum (x_i - \bar{x})^2} \sum c_i = 0$$

분산의 특성을 이용하여 다음과 같이 구할 수 있다.

$$\begin{aligned}\text{var}(b_2^*) &= \text{var}[\beta_2 + \Sigma(w_i + c_i)e_i] = \Sigma(w_i + c_i)^2\text{var}(e_i) \\ &= \sigma^2\Sigma(w_i + c_i)^2 = \sigma^2\Sigma w_i^2 + \sigma^2\Sigma c_i^2 \\ &= \text{var}(b_2) + \sigma^2\Sigma c_i^2 \\ &\geq \text{var}(b_2)\end{aligned}$$

마지막 줄은 $\Sigma c_i^2 \geq 0$ 라는 사실에서 비롯되며 선형 및 불편 추정량 b_2^*들에 대해 이들 각각의 추정량은 최소제곱 추정량 b_2의 분산보다 더 크거나 같은 분산을 갖는다고 할 수 있다. $\text{var}(b_2^*) = \text{var}(b_2)$인 유일한 경우는 모두 $c_i = 0$가 성립되는 때이며 이 경우 $b_2^* = b_2$가 된다. 따라서 b_2보다 나은 β_2의 다른 선형 불편 추정량은 존재하지 않으며 이로써 가우스–마코프 정리를 증명하였다.