02 CHAPTER

단순 선형회귀 모형

경제이론에 따르면 경제변수 간에는 여러 관계가 존재한다. 미시경제학에서는 물품의 수요량 및 공급량이 가격에 의존한다는 수요 및 공급 모형을 생각해 볼 수 있다. 또한 생산량이 사용되는 생산요소, 예를 들면 노동량의 함수로 설명하는 '생산 함수와 총생산물곡선'을 들 수 있다. 거시경제학에서는 경제의 총투자량이 이자율에 의존한다는 '투자 함수'와 총소비를 가처분 소득수준에 연계시키는 '소비 함수'를 생각해 볼 수 있다.

각 모형은 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(예 : 물품의 가격)가 변화할 경우 다른 변수(예 : 수요 또는 공급량)는 얼마만큼 변하는가? 또한 한 변수의 값을 아는 경우 다른 변수의 이에 상응하는 값을 예측하거나 예상할 수 있는가? 이 장에서는 **회귀 모형**(regression model)을 이용하여 위의 물음들에 답할 것이다. 모든 모형처럼 회귀 모형도 **가정**(assumption)에 기초하고 있다. 이 장에서는 이런 가정들을 분명히 해두고자 한다. 왜냐하면 이 가정들은 다음 장들에서 살펴볼 분석들이 적절해지도록 하는 조건이 되기 때문이다.

2.1 경제 모형

회귀 모형에 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 **확률변수**이다. 왜냐하면 가계를 뽑아서 위와 같은 질문을 하고, 이에 대답을 할 때까지는 그 값을 알 수 없기 때문이다.

유의사항

제0장에서 '확률변수'는 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 확률변수 자체를 나타낼 것이며 상황별로 이에 대한 해석을 명확히 할 것이다. ■

연속적 확률변수 y는 다양한 식료품 지출액을 구하게 될 확률을 나타내는 **확률밀도 함수**(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 10대가 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 확률밀도 함수는 지출이 모집단에서 어떻게 '분포되어' 있는지 알려주며, 이는 그림 2.1의 분포 중 하나와 같을 수 있다.

그림 2.1(a)의 확률분포는 가계소득에 대해 '조건부'이므로 실제로 조건부 확률밀도 함수이다. x가 주당 가계소득인 경우 조건부 확률밀도 함수는 $f(y|x = \$1{,}000)$이다. y의 **조건부 평균** 또는 **기댓값**은 $E(y|x = \$1{,}000) = \mu_{y|x}$이며, 이는 모집단의 1인당 평균 주당 식료품 지출액이 된다. y의 **조건부 분산**은 $\text{var}(y|x = \$1{,}000) = \sigma^2$이며, 이는 평균 $\mu_{y|x}$에 대한 가계지출 y의 퍼진 정도를 나타낸다. 모수 $\mu_{y|x}$와 σ^2을 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.

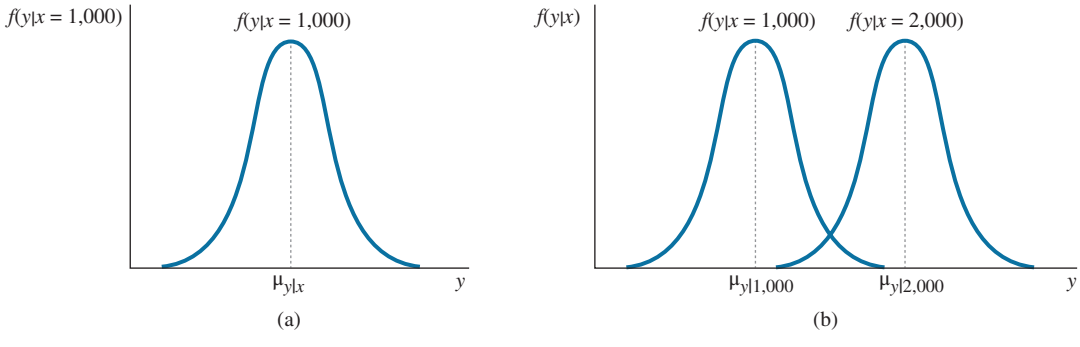

그림 2.1 (a) 소득이 $x = \$1{,}000$인 경우 식료품 지출액 y의 확률분포 $f(y|x = 1{,}000)$, (b) 소득이 $x = \$1{,}000$ 및 $x = \$2{,}000$인 경우 식료품 지출액 y의 확률분포

유의사항

확률변수의 기댓값을 '평균'값이라고 하는데, 이는 실제로 확률변수의 확률분포 중앙에 위치하는 모평균을 축약한 것이다. 이는 표본의 숫자값을 산술적으로 평균한 표본평균과 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 의존한다. 따라서 y를 **종속변수**(dependent variable)라 하고, x를 **독립변수**(independent variable) 또는 **설명변수**(explanatory variable)라고 한다. 계량경제학에서 우리는 실사회 지출액이 확률변수라는 사실을 알고 있으며, 자료를 사용하여 그 관계에 관해 알아보고자 한다.

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 지출 습관상 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 소비형태를 알아보는 데도 사용된다. 이런 소비형태는 거꾸로 해당 인구집단의 소비행위에 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 체인의 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 체인은 고소득 지역에 위치하고 다른 슈퍼마켓 체인은 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 **경제 모형**(economic model)을 수립하고 나서 수량적이거나 실증적인 경제분석의 기초가 될 **계량경제 모형**(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 확률밀도 함수를 살펴보았다. 각 밀도 함수 $f(y|x)$에 따르

면 지출이 평균값 $\mu_{y|x}$를 중심으로 분포하겠지만 소득수준이 높은 가계의 평균 지출액이 소득수준이 낮은 가계의 평균 지출액보다 크다.

자료를 활용하기 위해서 가계소득 및 식료품 지출액에 대한 자료를 어떻게 구하는지 설명하고 계량경제 분석을 수행할 계량경제 모형을 이제는 설정해야 한다.

2.2 계량경제 모형

앞 절에서의 경제논리가 주어진 경우 식료품 지출액과 소득 사이의 관계를 수량화하기 위해서는 그림 2.1의 생각을 계량경제 모형으로 진전시켜야 한다. 먼저 3인 가계는 매주 $80를 지출하고 받은 소득의 각 달러 중 10센트를 식료품에 또한 지출한다는 확고한 규칙을 갖고 있다고 가상하자. y = 주당 가계 식료품 지출액($), x = 주당 가계소득($)이라고 하자. 대수학적으로 나타내면 이 규칙은 $y = 80 + 0.10x$이다. 이 관계를 알고 있다면 가계소득이 $1,000인 어떤 주에 해당 가계는 식료품에 $180를 지출할 것이라고 계산할 수 있다. 주당 소득이 $100만큼 증가하여 $1,100가 되는 경우 식료품 지출액은 $190로 증가한다. 이것이 소득이 주어진 경우 식료품 지출액에 대한 **예측**(prediction)이다. 어떤 변수의 값이 주어진 경우 다른 변수의 값을 예측하는 것이 회귀분석의 주된 용도 중 하나이다.

회귀분석의 두 번째 주된 용도는 어떤 변수의 변화를 다른 변수의 변화 탓으로 돌리거나 연계시키는 것이다. 그것 때문에 'Δ'는 통상적인 대수학적 방법에서 '변화'를 나타낸다. 소득 변화 $100는 $\Delta x = 100$을 의미한다. 지출규칙 $y = 80 + 0.10x$로 인해 식료품 지출액의 변화는 $\Delta y = 0.10\Delta x = 0.10(100) = 10$이다. 소득 증가 $100는 식료품 지출액 증가 $10로 이어지거나, 식료품 지출액 증가 $10 발생의 원인이 된다. 기하학적으로 보면 이 규칙은 'y절편' 80 및 기울기 $\Delta y/\Delta y = 0.10$인 직선이다. 경제학자는 가계의 "식료품에 대한 한계지출성향이 0.10이다."라고 말할 수도 있다. 이것은 추가적인 소득 각 1달러에서 10센트가 식료품에 지출된다는 의미이다. 이를 달리 표현해서 경제학자가 사용하는 간단한 표현으로 나타내면 "소득의 식료품 지출액에 대한 한계효과는 0.10이다"라고 한다. 많은 경제 및 계량경제분석은 두 경제변수 사이의 인과관계를 측정하려 한다. 여기서 **인과관계**(causality)에 대한 언급, 즉 소득 변화가 식료품 지출액 변화로 이어진다는 말은 가계 지출액 규칙이 주어질 경우 매우 명백해진다. 하지만 반드시 이렇게 분명한 것은 아니다.

실제로는 많은 다른 요소들이 가계의 식료품 지출액에 영향을 미칠 수 있다. 예를 들면, 가계 구성원의 연령 및 성별, 신체적 크기, 육체근로자인지 또는 사무근로자인지 여부, 큰 경기 후에 모임이 있었는지 여부, 도시가계인지 또는 농촌가계인지 여부, 가계 구성원이 채식주의자이거나 팔레오 다이어트(원시인과 같은 식단을 유지하는 다이어트)를 하는 사람인지 여부를 생각할 수 있으며, 여기에는 물론 기호 및 선호("나는 정말로 송로 같은 것들을 좋아한다.") 그리고 충동구매("복숭아가 정말로 먹음직스럽게 보인다.")가 포함된다. 많은 요소가 영향을 미친다. 'e = 소득 이외에 식료품 지출액에 영향을 미치는 그 밖의 모든 다른 요소'라고 하자. 나아가 엄격하든 또는 그렇지 않든 간에 어떤 가계가 식료품 지출액 규칙을 갖고 있더라도 우리는 그것을 알 수 없다. 이런 현실을 감안하여 가계의 식료품 지출액 결정은 다음과 같은 식에 기초한다고 가정하자.

$$y = \beta_1 + \beta_2 x + e \tag{2.1}$$

y 및 x 이외에 식 (2.1)은 '80' 및 '0.10' 대신에 알지 못하는 2개의 모수 β_1 및 β_2, 그리고 **오차항**(error term) e를 포함한다. 여기서 오차항은 주당 가계 식료품 지출액에 영향을 미치는 그 밖의 모든 다른 요소를 나타낸다.

가계에 대해 실험을 이행하고 있다고 가상하자. 가계소득이 주당 $100만큼 증가하고 다른 조건이 동일하다고 가상하자. 다른 조건이 동일하다라든지 또는 그 밖의 모든 다른 요소는 같다라는 말은 경제원론 과정에서 폭넓게 논의된 '세터리스 패리버스(*ceteris paribus*)'란 가정이다. $\Delta x = 100$은 가계소득의 변화를 나타낸다고 하자. 가계의 식료품 지출액에 영향을 미치는 그 밖의 모든 다른 요소, e가 동일하다는 가정은 $\Delta e = 0$을 의미한다. 소득 변화가 미치는 영향은 $\Delta y = \beta_2 \Delta x + \Delta e = \beta_2 \Delta x = \beta_2 \times 100$이다. 주당 식료품 지출액의 변화 $\Delta y = \beta_2 \times 100$은 소득 변화에 의해 설명되거나 또는 소득 변화가 원인이 되어 발생하였다. 알지 못하는 모수 β_2는 소득의 식료품에 대한 한계지출성향을 의미하며, 식료품 지출에 사용된 소득 증가액의 비율이다. 이것은 '얼마나'란 질문, 즉 "그 밖의 모든 다른 요소는 동일하다고 보고 소득 변화가 주어진 경우 식료품 지출액은 얼마나 변화하는지"에 대해 알려준다.

앞 단락에서 언급한 실험은 실행할 수가 없다. 어떤 가계에 추가적으로 소득 $100를 줄 수는 있지만 그 밖의 모든 다른 요소를 일정하게 유지할 수는 없다. 소득 증가가 식료품 지출에 미치는 한계효과를 단순하게 계산한 $\Delta y = \beta_2 \times 100$이 가능하지 않다. 하지만 β_2를 추정하는 회귀분석을 활용하여 '얼마나'라는 물음에 답을 줄 수 있다. 회귀분석은 변수들 사이의 관계를 살펴보기 위해 자료를 사용하는 통계적 방법이다. **단순 선형회귀분석**(simple linear regression analysis)은 y변수와 x변수 사이의 관계를 검토한다. '단순'이란 말은 용이하기 때문이 아니라 단지 하나의 x변수가 있기 때문에 사용되었다. y변수를 종속변수, 결과변수, 설명변수, 왼쪽 변수, 피회귀변수라고 한다. 위의 예에서 종속변수는 y = 주당 가계 식료품 지출액이다. 변수 x = 주당 가계소득을 독립변수, 설명변수, 오른쪽 변수, 회귀변수라고 한다. 식 (2.1)은 단순 선형회귀 모형이다.

모든 모형은 현실을 추상화한 것이며, 모형을 이해하려면 가정이 필요하다. 이것은 회귀 모형에도 동일하게 적용된다. 단순 선형회귀 모형의 첫 번째 가정은 식 (2.1)이 고려하고 있는 모집단의 구성원들에게 준수된다는 것이다. 예를 들어, 모집단을 남부 호주처럼 일정 지역의 3인 가계라고 정의해 보자. 알지 못하는 β_1 및 β_2를 **모수**(population parameter)라고 하자. 우리는 행태 규칙 $y = \beta_1 + \beta_2 x + e$가 모집단의 모든 가계에 준수된다고 본다. 매주 식료품 지출액은 β_1에 소득의 일정 비율 β_2를 더하고 다른 요소 e를 더한 것이다.

일반적으로 모집단은 크고 모집단의 모든 구성원을 검토하는 것이 불가능하기 때문에 (또는 불가능할 정도로 비용이 많이 소요되기 때문에) 통계학이란 학문 분야가 탄생하였다. 일정한 지역에서 비록 중간 규모의 도시라 하더라도 3인 가계의 모집단은 너무 커서 개별적으로 조사할 수 없다. 통계학적 및 계량경제학적 방법은 모집단으로부터의 자료표본을 검토하고 분석하는 것이다. 자료를 분석한 후에 **통계적 추론**(statistical inference)을 한다. 이것은 자료분석에 기초하여 내린 모집단에 관한 결론이거나 판단이다. 이런 추론은 자료가 수집된 특정 모집단에 관한 결론이다. 남부 호주에 소재하

는 가계들에 관한 자료는 미국 남부에 소재하는 가계들에 대해 추론을 하거나 결론을 내리는 데 유용할 수도 있고 유용하지 않을 수도 있다. 호주 멜버른시에 있는 가계는 미국 루이지애나주 뉴올리언스시에 있는 가계와 동일한 식료품 지출 형태를 갖는가? 이것은 흥미로운 연구주제가 될 수 있다. 만일 동일한 형태를 갖지 않는다면, 호주 자료표본으로부터 미국 뉴올리언스시에 관한 타당한 결론을 도출할 수 없다.

2.2.1 자료 생성 과정

자료표본과 이들 자료를 실제로 어떻게 구하는지는 뒤이은 추론을 하는 데 매우 중요하다. 자료표본을 수집하는 메커니즘은 매우 특정한 분야이며(예를 들면, 농학은 경제학과 다르다), 이 책의 범위를 벗어난다.[1] 가계 식료품 지출 예에서 모집단으로부터 **무작위**로 선택한 N개 자료 쌍으로 구성된 일정 시점에서의 표본(이것은 **횡단면 자료**가 된다)을 구할 수 있다고 가정하자. (y_i, x_i)는 i번째 자료 쌍을 나타내며 $i = 1, \cdots, N$이다. 변수 y_i 및 x_i는 관찰될 때까지 알지 못하기 때문에 **확률변수**라고 한다. 무작위로 선택된 가계의 경우 첫 번째 관찰값 쌍 (y_1, x_1)은 모든 다른 자료 쌍과 통계적으로 독립적이 된다. 각 관찰값 쌍 (y_i, x_i)는 모든 다른 자료 쌍 (y_j, x_j)과 **통계적으로 독립적**(statistically independent)이며, 여기서 $i \neq j$이다. 나아가 확률변수 y_i 및 x_i는 이들 값의 분포를 설명하는 결합 *pdf* $f(y_i, x_i)$를 갖는다고 가정한다. 우리는 종종(예를 들면, 이변량 정규분포와 같은) 결합분포의 정확한 성격을 알지 못하지만, 동일한 모집단에서 추출한 모든 쌍은 동일한 결합 *pdf*를 따른다고 가정한다. 따라서 자료 쌍들은 통계적으로 독립적일 뿐만 아니라 **동일하게 분포된다**(identically distributed)고 본다. 이것을 약자로 *i.i.d.* 또는 *iid*로 표기한다. *iid*인 자료 쌍들을 **무작위 표본**(random sample)이라고 한다.

행태 규칙 $y = \beta_1 + \beta_2 x + e$가 모집단의 모든 가계에 준수된다는 첫 번째 가정이 참인 경우, 각 자료 쌍 (y_i, x_i)에 대해 식 (2.1)을 다시 표현하면 다음과 같다.

$$y_i = \beta_1 + \beta_2 x_i + e_i, \; i = 1, \cdots, N \tag{2.1}$$

관찰할 수 있는 자료는 위의 식을 따른다고 가정하기 때문에, 이를 때때로 **자료 생성과정**(data generating process, DGP)이라고 한다.

2.2.2 무작위 오차 및 강 외생성

단순 회귀 모형 식 (2.1)의 두 번째 가정은 '그 밖의 모든 다른 요소' 항 e와 관련된다. 특정 가계를 선택해서 이들을 관찰하기 전까지는 변수 (y_i, x_i)가 어떤 값을 취하는지 알지 못하므로 이를 확률변수라고 한다. 오차항 e_i 또한 확률변수이다. 모든 사람의 기호와 선호가 다르다는 이유만으로 상이해진다면, 소득을 제외하고 식료품 지출액에 영향을 미치는 모든 다른 요소는 각 모집단 가계에 대해 상이하게 된다. 식료품 지출액 및 소득과 달리 **무작위 오차항**(random error term) e_i는 **관찰될 수 없다**. 케이크를 한 조각 먹는 것으로부터 도출된 경제적 '효용'을 직접 측정할 수 없다. 두 번째 회귀 가정은 i번째 가계의 식료품 지출액에 영향을 미치는 모든 다른 요소 집합체의 효과인 e_i의 값을 예측하기 위

1 예를 들면, 다음을 참조하시오. Paul S. Levy and Stanley Lemeshow (2008) *Sampling of Populations: Methods and Applications, 4th Edition*, Hoboken, NJ: John Wiley and Sons, Inc.

해서 x변수, 즉 소득이 사용될 수 없다는 것이다. i번째 가계에 대한 소득값 x_i가 주어진 경우, 무작위 오차 e_i에 대한 최선(최적)의 예측량은 조건부 기댓값 또는 조건부 평균, $E(e_i|x_i)$이다. e_i를 예측하기 위해 x_i가 사용될 수 없다는 가정은 $E(e_i|x_i)=0$이라고 말하는 것과 같다. 즉 가계소득이 주어진 경우 무작위 오차가 영이라고 예측하는 것보다 더 나은 일을 할 수 없다. 모든 다른 요소가 식료품 지출액에 미치는 영향은 매우 특수한 방법으로 평균이 영에 달하게 된다. 이것이 참일 수도 있고 아닐 수도 있는 다른 상황에 관해 뒤에서 논의할 것이다. 지금은 $E(e_i|x_i)=0$이 두 가지 의미를 갖는다는 점만을 생각하자. 첫 번째 의미는 $E(e_i|x_i)=0 \Longrightarrow E(e_i)=0$이다. 무작위 오차의 조건부 기댓값이 영이라면, 무작위 오차의 **무조건부 기댓값**(unconditional expectation)도 또한 영이 된다. 모집단에서 무작위 오차항으로 요약할 수 있는 모든 누락된 요소의 평균 효과는 영이 된다.

두 번째 의미는 $E(e_i|x_i)=0 \Longrightarrow \text{cov}(e_i, x_i)=0$이다. 무작위 오차의 조건부 기댓값이 영이라면, i번째 관찰값에 대한 무작위 오차는 그에 상응하는 관찰값 x_i와 영인 공분산과 영인 상관을 갖는다. 위의 예에서 i번째 가계에 대해 소득을 제외하고 식료품 지출액에 영향을 미치는 모든 요소를 나타내는 무작위 구성요소 e_i는 해당 가계의 소득과 상관되지 않는다. 이것이 참일 수 있다는 것을 어떻게 보여줄 수 있는지 의구심을 가질 수도 있다. 결국 e_i는 관찰할 수 없다. 대답은 그것이 매우 어려운 일이라는 것이다. 여러분은 모형에서 누락됐을지도 모를 어떤 것도 x_i와 상관되지 않는다는 점을 자신과 다른 사람에게 확신시켜야만 한다. 주요한 도구는 경제적 논리, 즉 여러분의 지적 실험(즉 사고력), 관련 주제에 관한 문헌을 읽고 동료나 동급생과의 토론 등을 꼽을 수 있다. 대부분의 경제 모형에서 절대적인 확신을 갖고 $E(e_i|x_i)=0$이 참이라고 입증할 수는 없다.

$E(e_i|x_i)=0$은 두 가지 의미를 갖는다는 사실에 주목하였다. 그 의미 중 한 개라도 참이 아닌 경우 $E(e_i|x_i)=0$은 참이 아니다. 즉 다음과 같다.

$$E(e_i|x_i)\neq 0,\ \text{(i) } E(e_i)\neq 0 \text{인 경우 또는 (ii) } \text{cov}(e_i, x_i)\neq 0 \text{인 경우}$$

첫 번째 경우로 무작위 오차 e_i의 모집단 평균이 영이 아니라면, $E(e_i|x_i)\neq 0$이다. 예를 들면, $E(e_i)=3$처럼 $E(e_i)\neq 0$인 경우에 대해서는 다음에 살펴볼 것이다. $E(e_i|x_i)=0$이 갖는 두 번째 의미는 $\text{cov}(e_i, x_i)=0$이라는 것이다. i번째 관찰값에 대한 무작위 오차는 설명변수에 대한 i번째 관

정리문제 2.1 외생성 가정이 준수되지 못하는 경우

무작위 표본자료를 사용하여 근로자 임금과 교육 연수 사이의 관계를 살펴보는 회귀 모형을 생각해 보자. 다음의 단순 회귀 모형에서 $WAGE_i$는 i번째 무작위로 뽑은 근로자의 시간당 임금이며 $EDUC_i$는 해당 근로자의 교육 연수이다 : $WAGE_i=\beta_1+\beta_2 EDUC_i+e_i$. 무작위 표본의 이런 쌍관계($WAGE_i$, $EDUC_i$)에서 iid가 준수된다고 가정한다. 이 모형에서 무작위 오차 e_i는 근로자의 임금에 영향을 미치는 $EDUC_i$ 이외의 모든 다른 요소의 영향을 설명한다. 재능, 지능, 인내심, 근면성 모두 근로자의 중요한 특성으로 임금에 영향을 미칠 가능성이 크다. e_i에 함께 딸려 들어간 위와 같은 요인 중 일부가 $EDUC_i$와 상관될 가능성이 있는가? 잠시 동안만 생각해 보아도 '그렇다'라고 답하게 된다. 교육수준이 높은 사람일수록 능력, 지능, 인내심, 근면성이 더 높다고 보는 것은 그럴듯하게 들린다. 따라서 $EDUC_i$는 위의 회귀식에서 내생적 변수이므로 강 외생성에 대한 가정은 준수되지 못하였다고 주장할 수 있다.

찰값과 영인 공분산과 상관을 갖는다. $\text{cov}(e_i, x_i) = 0$인 경우, 쌍 (y_i, x_i)가 iid라는 첫 번째 가정이 준수된다면 설명변수 x는 외생적이라고 한다. x가 외생적일 때 β_1 및 β_2를 추정하기 위해서 회귀분석이 성공적으로 사용될 수 있다. 더 약한 조건 $\text{cov}(e_i, x_i) = 0$, 즉 단순한 외생성과 더 강한 조건 $E(e_i|x_i) = 0$을 구별하기 위해서 $E(e_i|x_i) = 0$인 경우 x는 **강하게 외생적**(strictly exogenous)이라고 한다. $\text{cov}(e_i, x_i) \neq 0$인 경우 x는 내생적이라고 한다. x가 내생적일 때 통계적 추론을 하는 것은 더 어려워지며, 이따금 훨씬 더 어려워진다. 이 책의 나머지 부분에서 외생성 및 강 외생성에 관한 많은 것을 알아볼 것이다.

2.2.3 회귀 함수

강 외생성 가정의 중요성은 다음과 같다. 강 외생성 가정 $E(e_i|x_i) = 0$이 참이라면 x_i가 주어진 경우 y_i의 조건부 기댓값은 다음과 같다.

$$E(y_i|x_i) = \beta_1 + \beta_2 x_i + E(e_i|x_i) = \beta_1 + \beta_2 x_i, \quad i = 1, \cdots, N \tag{2.2}$$

식 (2.2)에서 조건부 기댓값 $E(y_i|x_i) = \beta_1 + \beta_2 x_i$를 **회귀 함수**(regression function) 또는 **모집단 회귀 함수**(population regression function)라고 한다. 모집단에서 x_i에 조건부인 i번째 관찰값에 대한 종속변수의 평균값은 $\beta_1 + \beta_2 x_i$로 구할 수 있다고 본다. 또한 x_i가 주어진 경우 무작위 오차의 평균은 영이고 x의 어떤 변화도 무작위 오차 e의 어떠한 상응하는 변화와도 상관되지 않는다는 의미에서, x의 변화, 즉 Δx가 주어진 경우 그 밖의 모든 것이 일정하다고 볼 경우 이에 따른 $E(y_i|x_i)$의 변화는 $\beta_2 \Delta x$가 된다고 본다. 이런 의미에서 x의 변화는 x_i가 주어진 경우 y의 기댓(모집단 평균)값의 변화로 이어지거나 또는 이를 일으키는 원인이 된다고 말할 수 있다.

식 (2.2)의 회귀함수는 그림 2.2에서 y절편 $\beta_1 = E(y_i|x_i = 0)$과 다음과 같은 기울기를 갖는 직선으로 나타낼 수 있다.

$$\beta_2 = \frac{\Delta E(y_i|x_i)}{\Delta x_i} = \frac{dE(y_i|x_i)}{dx_i} \tag{2.3}$$

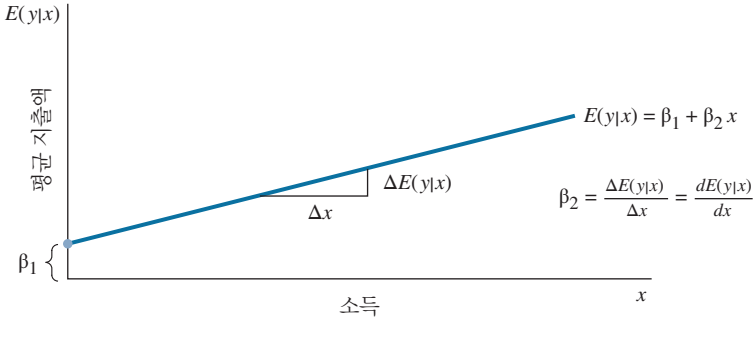

그림 2.2 경제 모형 : 1인당 평균 식료품 지출액과 소득 사이의 선형관계

정리문제 2.2 가계 식료품 지출액 모형에서의 강 외생성

강 외생성 가정이 의미하는 바는 i번째 가계의 소득이 주어진 경우 i번째 가계의 식료품 지출액에 영향을 미치는 그 밖의 모두의 평균이 영이라는 것이다. 이런 가능성을 검정하는 한 가지 방법은 "i번째 가계의 소득을 사용하여 e_i의 값을 예측할 수 있는가?"라는 질문과 상통한다. 여기서 e_i는 식료품 지출액에 영향을 미치는 소득 이외의 모든 요인의 결합된 영향을 의미한다. 위의 질문에 대한 답변이 긍정적인 경우 강 외생성에 대한 가정은 준수되지 못한다. 부정적인 경우 $E(e_i|x_i)=0$은 타당한 가정이 될 수 있다. 이와 같다면 식 (2.1)은 인과관계를 나타내는 모형으로 해석될 수 있으며, β_2는 그 밖의 다른 것이 일정하다고 할 경우 식 (2.3)에서 보는 것처럼 가계 식료품 기대 (평균) 지출액에 소득이 미치는 한계효과로 해석될 수 있다. $E(e_i|x_i)\neq 0$인 경우 x_i를 사용하여 영이 아닌 e_i값을 예측할 수 있으며, 이는 다시 y_i값에 영향을 미치게 된다. 이 경우 β_2는 소득 변화가 미치는 모든 영향을 나타낼 수 없으며 해당 모형이 인과관계를 보여준다고 할 수 없다.

여기서 Δ는 '변화'를 나타내며 $dE(y|x)/dx$는 x에 대한 $E(y|x)$의 '도함수'를 의미한다. 이 책에서는 대부분 도함수를 사용하지 않을 것이며, 이 개념에 익숙하지 않거나 기억하지 못하는 경우 'd'를 형식화된 변형이라고 생각하고 진도를 계속 나가도 무방하다.

강 외생성 가정에 따른 또 다른 중요한 결과는 계량경제 모형에서 종속변수를 두 가지 구성요소, 즉 독립변수값이 변화함에 따라 체계적으로 변하는 구성요소와 또 다른 구성요소인 무작위적인 '잡음'으로 분해할 수 있다는 점이다. 즉 계량경제 모형 $y_i=\beta_1+\beta_2 x_i+e_i$는 두 개 부분, $E(y_i|x_i)=\beta_1+\beta_2 x_i$ 및 무작위 오차 e_i로 나누어질 수 있으며 다음과 같아진다.

$$y_i=\beta_1+\beta_2 x_i+e_i=E(y_i|x_i)+e_i$$

종속변수 y_i의 값은 설명변수값의 변함에 따른 조건부 평균 $E(y_i|x_i)=\beta_1+\beta_2 x_i$의 변동에서 기인하여 체계적으로 변화한다. 그리고 종속변수 y_i의 값은 e_i에서 기인하여 무작위적으로 변화한다. e 및 y의 조건부 pdf는 그림 2.3에서 보는 것처럼 위치를 제외하고 동일하다. 주당 소득이 $x=\$1{,}000$인 가계에 대해 식료품 지출액의 두 값 y_1 및 y_2가 조건부 평균과 관련하여 그림 2.4에 있다. 기호 및 선호, 그 밖의 모든 것의 변동으로 인해서 가계별로 식료품 지출액상의 변동이 있게 된다. 일부 가계는 동일한 소득을 갖는 가계들의 평균값보다 더 많이 지출하며, 또 다른 일부는 더 적게 지출한다. β_1 및 β_2를 알고 있다면 조건부 평균 지출액 $E(y_i|x=1{,}000)=\beta_1+\beta_2(1{,}000)$을 계산할 수 있으며, 또한 무작위 오차 e_1 및 e_2의 값도 계산할 수 있다. β_1 및 β_2를 결코 알지 못하므로 e_1 및 e_2도 결코 계산할 수 없다. 하지만 우리가 가정하는 것은 소득 x의 각 수준에서 무작위 오차로 나타내는 모든 것의 평균값은 영이라고 한다는 점이다.

2.2.4 무작위 오차 변동

무작위 오차항의 조건부 기댓값은 영, 즉 $E(e_i|x_i)=0$이라고 가정하였다. 무작위 오차항에 대해 우리는 그것의 조건부 평균값 또는 기댓값 그리고 분산에 관심을 갖는다. 이상적으로 말하면 무작위 오차의 조건부 분산(conditional variance)은 일정하여 다음과 같다.

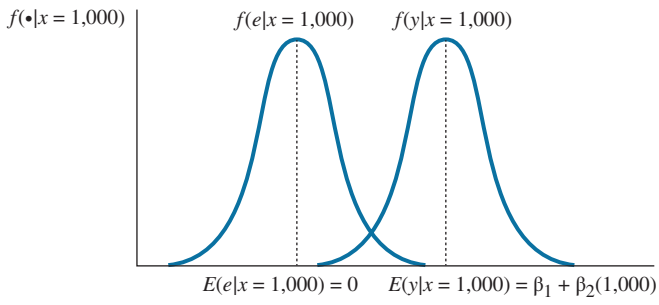

그림 2.3 e 및 y에 대한 조건부 확률밀도 함수

$$\text{var}(e_i|x_i) = \sigma^2 \tag{2.4}$$

이것은 **동분산**(homoskedasticity 또는 homoscedasticity) 가정이다. 각 x_i에서 무작위 오차 구성요소의 변동은 동일하다. 모집단 관계 $y_i = \beta_1 + \beta_2 x_i + e_i$를 가정하면 종속변수의 조건부 분산은 다음과 같다.

$$\text{var}(y_i|x_i) = (\beta_1 + \beta_2 x_i + e_i|x_i) = \text{var}(e_i|x_i) = \sigma^2$$

x_i에 대한 조건부로 마치 그것이 알려져서 무작위적이지 않은 것처럼 취급하기 때문에 위와 같이 단순화가 이루어진다. x_i가 주어진 경우 구성요소 $\beta_1 + \beta_2 x_i$가 무작위적이지 않아서 앞에서 살펴본 분산 규칙이 적용된다.[2]

pdf $f(y|x = 1{,}000)$ 및 $f(y|x = 2{,}000)$이 동일한 분산 σ^2을 갖는 그림 2.1(b)에서 이것은 명시적인 가정이었다. 강 외생성이 준수되는 경우 그림 2.2에서 보는 것처럼 회귀함수는 $E(y_i|x_i) = \beta_1 + \beta_2 x_i$가 된다. 조건부 분산 $f(y|x = 1{,}000)$ 및 $f(y|x = 2{,}000)$은 그림 2.5에서 조건부 평균 함수를 따라 위치한다.

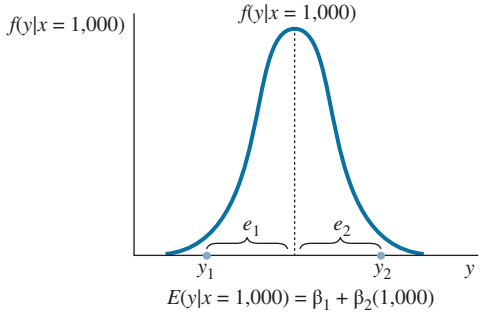

그림 2.4 무작위 오차

2 a 및 b를 상수라고 하면 다음과 같다.
$\text{var}(aX + b) = a^2 \text{var}(X)$

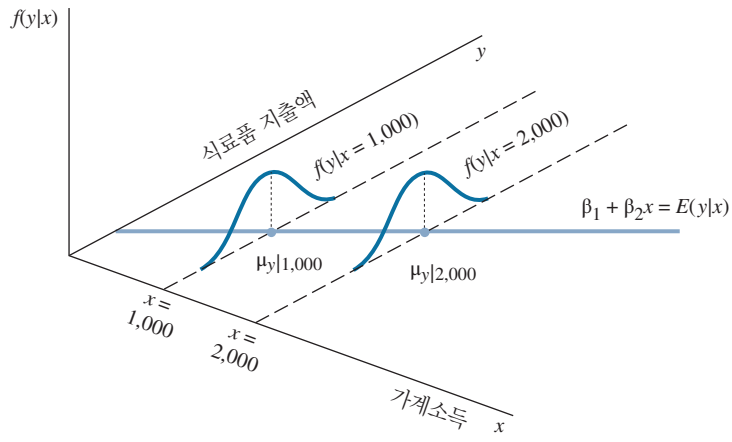

그림 2.5 2개의 소득수준에서 식료품 지출액 y에 대한 조건부 확률밀도 함수

가계 지출액의 예에서 사고의 틀은 다음과 같다. 가계소득 x의 특정 수준에 대해서 가계의 식료품 지출액 규모는 조건부 평균에 관해 무작위로 변동하며, 이는 각 x에서 무작위 오차 e의 평균값이 영이라는 가정에서 기인된다. 따라서 각 소득수준에서 가계의 식료품 지출액은 회귀 함수를 중심으로 변동한다. 조건부 동분산 가정은 각 소득수준에서 평균에 관한 식료품 지출액의 변동이 같다는 의미이다. 이것은 소득 각각의 모든 수준에서 식료품 지출액이 평균값 $E(y_i|x_i) = \beta_1 + \beta_2 x_i$에서 얼마나 멀리 떨어져 위치할지에 관해 동등하게 불확실하다. 나아가 이 불확실성은 소득 또는 그 밖의 어떤 것에 의존하지 않는다. 이 가정이 위배되어 $\text{var}(e_i|x_i) \neq \sigma^2$인 경우 무작위 오차는 **이분산적**(heteroskedastic)이라고 한다.

2.2.5 X의 변동

회귀분석의 목적 중 하나는 $\beta_2 = \Delta E(y_i|x_i)$를 추정하는 것이다. x의 변화가 미치는 효과를 추정하기 위해서 자료표본이 사용될 수 있기를 희망한다면, 표본에서 설명변수 x의 상이한 값들을 관측하여야만 한다. 직관적으로 볼 때 오직 소득이 $1,000인 가계들만의 자료를 수집할 경우 소득 변화가 식료품 지출액의 평균값에 미치는 효과를 측정할 수 없다. 기초 기하학에서 선분을 결정하기 위해서는 2개 점을 취한다는 사실을 기억하자. 실제로 x의 상이한 값들이 많아질수록 이들은 더 큰 변동을 보이며, 회귀분석도 더 나아질 것이라는 사실을 앞으로 알게 될 것이다.

2.2.6 오차 정규성

그림 2.1과 관련하여 이루어진 논의에서 소득이 주어진 경우 식료품 지출액이 정규분포한다는 가정을 명시적으로 하였다. 그림 2.3~2.5에서는 고전적인 종 모양의 곡선을 그려서 조건부적으로 정규분포하는 오차 및 종속변수에 대한 가정을 묵시적으로 하였다. 회귀분석을 시행하기 위해서 무작위 오

차가 조건부적으로 정규분포할 필요는 전혀 없다. 하지만 제3장에서 살펴볼 것처럼 표본이 작을 때 각 x값이 주어진 경우 무작위 오차 및 종속 변수가 정규분포하는 것이 통계적 추론을 하는 데 유리하다. 약간의 인터넷 탐색을 통해 알 수 있듯이 정규분포는 오래되고 흥미로운 유래를 갖고 있다.[3] 회귀 오차가 정규분포한다고 가정하는 논거 중 하나는 이것이 많은 상이한 요소들의 수집체를 나타낸다는 데 있다. **중심극한정리**(Central Limit Theorem)에 따르면 대체적으로 많은 무작위 요소들의 수집체는 정규분포하는 추세를 갖는다고 한다. 식료품 지출액 모형의 틀 내에서 무작위 오차가 기호 및 선호를 반영한다고 생각할 경우 각 소득수준에서의 무작위 오차가 정규분포한다고 보는 것은 매우 그럴듯하다. 조건부 정규오차 가정을 할 경우 $e_i|x_i \sim N(0, \sigma^2)$이라고 표기하고, 또한 $y_i|x_i \sim N(\beta_1+\beta_2 x_i, \sigma^2)$이라고도 나타낸다. 이렇게 할 경우 매우 강한 가정이 되며, 언급했던 것처럼 엄격히 말하면 필요한 가정은 아니다. 따라서 우리는 이것을 **선택적** 가정이라고 한다.

2.2.7 외생성 가정 일반화하기

지금까지 우리는 자료 쌍 (y_i, x_i)가 무작위 표본에서 추출되었고 *iid*라고 가정하였다. 설명변수의 표본값들이 상관된다면 어떤 일이 발생하는가? 그리고 그것은 어떻게 발생할 수 있는가?

금융 또는 거시경제 시계열 자료를 사용할 때 독립성 결여문제가 자연적으로 발생한다. 신규주택 착공건수 y_t 및 현재의 30년 만기 주택담보대출 고정금리, x_t에 대한 월간 보고서를 관찰하고 있다고 가상하자. 그리고 모형 $y_t = \beta_1 + \beta_2 x_t + e_t$를 가정해 보자. 자료 (y_t, x_t)는 거시경제 시계열 자료로 설명될 수 있다. 일정한 시점에서 (예컨대 가계, 기업, 사람, 국가들처럼) 많은 단위들에 대한 관찰값을 갖는 횡단면 자료와는 매우 다르게, 시계열 자료의 경우 많은 변수들에 대해 시간이 흐름에 따른 관찰값을 갖게 된다. 시계열 자료를 나타내기 위해서는 아래첨자 't'를 사용하고, 표본크기를 나타내기 위해서는 T를 사용하는 것이 관례적이다. 자료 쌍 (y_t, x_t), $t = 1, \cdots, T$에서 y_t 및 x_t 둘 모두 관찰될 때까지 그 값을 알지 못하기 때문에 확률적이다. 나아가 각 자료 시리즈는 시간을 가로질러 건너서 상관되기 쉽다. 예를 들면, 월간 주택담보대출 고정금리는 서서히 변동하기 쉽기 때문에 t기의 금리는 $t-1$기의 금리와 상관된다. 쌍 (y_t, x_t)가 확률분포로부터의 무작위 *iid* 뽑기를 나타낸다고 하는 가정은 현실적이지 못하다. 이 경우 외생성 가정을 생각할 때 x_t와 e_t 사이에 존재할지 모를 상관뿐만 아니라 e_t와 설명변수의 각 다른 값, 즉 x_s, $s = 1, 2, \cdots, T$ 사이에 존재할지도 모를 상관에 대해 관심을 가져야 한다. x_s가 x_t와 상관될 경우 x_s(예를 들면, 어떤 달의 주택담보대출 금리)가 y_t(예를 들면, 다음 달의 주택 착공건수)에 영향을 미칠 수 있다. 식 $y_t = \beta_1 + \beta_2 x_t + e_t$에 포함되는 것은 x_s가 아니라 x_t이기 때문에 x_s의 효과는 e_t에 포함되며, 이것은 $E(e_t|x_s) \neq 0$을 의미한다. e_t값을 예측하는 데 도움을 주기 위해서 x_s를 사용할 수 있다. 쌍 (y_t, x_t)가 독립적이라고 가정할 때 이 가능성은 배제된다. 즉 쌍 (y_t, x_t)의 독립성 그리고 가정 $E(e_t|x_t) = 0$이 모든 $s = 1, 2, \cdots, T$에 대해 $E(e_t|x_s) = 0$을 의미한다.

x의 값들이 상관된 모형들로 강 외생성 가정을 연장하기 위해서 모든 $(t, s) = 1, 2, \cdots, T$에 대해 $E(e_t|x_s) = 0$이라고 가정할 필요가 있다. 이것은 설명변수의 어떤 값을 사용하더라도 t기의 무작위 오차 e_t를 예측할 수 없다는 것을 의미한다. 또는 이전의 표시법으로 나타내면, 모든 $(i, j) = 1, 2, \cdots, N$

3 예를 들면, 다음과 같다. Stephen M. Stigler (1990) *The History of Statistics: The Measurement of Uncertainty, Reprint Edition*, Belknap Press, 73–76.

에 대해 $E(e_i|x_j)=0$이 된다. 보다 편리한 형태로 이 가정을 나타내기 위해 기호 $\mathbf{x}=(x_1, x_2, \cdots, x_N)$을 도입하여 보자. 즉 설명변수에 대한 모든 표본 관찰값들을 나타내기 위해서 $\mathbf{x}$를 사용하자. 그러면 강 외생성 가정을 나타내는 보다 일반적인 방법은 $E(e_i|\mathbf{x})=0,\ i=1, 2, \cdots, N$이 된다. 이런 가정에 기초하여 $i=1, 2, \cdots, N$에 대해 $E(y_i|\mathbf{x})=\beta_1+\beta_2 x_i$라고도 나타낼 수 있다. 이 가정은 다른 형태의 자료틀 내에서 이 장 및 제9장에서 추가적으로 논의할 것이다. 가정 $E(e_i|\mathbf{x})=0,\ i=1, 2, \cdots, N$은 $E(e_i|x_i)=0$ 그리고 쌍 $(y_i|x_i)$가 독립적이라고 가정하는 것보다 더 약한 가정이다. 이것을 통해 우리는 x에 대한 상이한 관찰값들이 독립적이라는 경우에 대해서뿐만 아니라 이들이 상관될 수 있는 경우에 대해서도 많은 결과를 도출할 수 있다.

2.2.8 오차상관

한 가계에 대한 오차(e_i) 또는 한 기간에 대한 오차(e_t)와 다른 가계에 대한 설명변수(x_j) 또는 다른 기간에 대한 설명변수(x_s) 사이에 존재할 수도 있는 상관 이외에 무작위 오차항들 사이에 상관이 존재하는 것도 가능하다.

어떤 시점에서 수집된 가계, 개인, 기업들에 대한 자료, 즉 횡단면 자료하에서 **공간적으로** 연계된 개체들에 대한 무작위 오차 사이에 통계적 독립성이 결여될 수 있다. 즉 동일한 지역에 거주하는 두 명(또는 그 이상)의 개인에 대한 관찰값을 수집한다고 가상하자. 특정 지역에 거주하는 사람들 사이에 유사성이 존재한다는 점은 매우 그럴듯해 보인다. 인근에 있는 가정들이 동질적이라면 이웃들은 소득도 유사할 것으로 기대할 수 있다. 일부 교외지역은 녹색 공간과 어린아이들을 위한 학교로 인해 인기가 있는데, 이것은 가계들이 연령 및 관심사항 면에서 유사한 구성원들로 구성될 수 있다는 것을 의미한다. 우리는 오차에 공간적인 요소 s를 추가하고, i번째 및 j번째 가계에 대한 무작위 오차 $e_i(s)$ 및 $e_j(s)$가 공통의 위치로 인해서 상관된다고 말할 수도 있다. 보다 큰 자료표본 내에서는 공간적인 요소로 인해 상관된 오차를 갖는 관찰값들의 **군집**이 있을 수 있다.

시계열과 관련해서는 미국 걸프만과 루이지애나주 뉴올리언스시를 황폐화시켰던 허리케인 카트리나의 경우를 생각해 보자. 이 허리케인으로 인한 충격은 발생되고 나서 사라져버린 것이 아니었다. 이 거대한 무작위적인 사건의 충격은 2005년 8월 동안 주택 및 금융시장에 영향을 미쳤으며, 그러고 나서 또한 9월, 10월 등등으로 현재까지 영향을 미치고 있다. 따라서 모집단 관계 $y_t=\beta_1+\beta_2 x_t+e_t$에서 무작위 오차는 시간이 흐름에 따라 상관되므로, $\text{cov}(e_t, e_{t+1})\neq 0$, $\text{cov}(e_t, e_{t+2})\neq 0$ 등등이 된다. 이것을 계량경제학에서 **계열상관**(serial correlation) 또는 **자기상관**(autocorrelation)이라고 한다.

회귀분석의 출발점은 오차상관이 존재하지 않는다고 가정하는 것이다. 시계열 모형에서는 $t\neq s$인 경우 $\text{cov}(e_t, e_s|\mathbf{x})=0$이라고 가정함으로써 시작하였으며, 횡단면 자료에서는 $i\neq j$인 경우 $\text{cov}(e_i, e_j|\mathbf{x})=0$이라고 가정함으로써 시작하였다. 제9장에서는 이런 가정들이 준수되지 않을 경우 대처하는 방법에 관해 논의할 것이다.

2.2.9 가정들 요약하기

매우 일반적인 방법으로 단순회귀 모형의 출발점이 되는 가정들을 요약하고자 한다. 요약에서는 아래첨자 i 및 j를 사용하였지만 가정들은 일반적이며 시계열 자료에 동등하게 적용된다. 이 가정들

이 준수될 경우 회귀분석은 알지 못하는 모집단 모수 β_1 및 β_2를 성공적으로 추정할 수 있으며, $\beta_2 = \Delta E(y_i|x_i)/\Delta x_i = dE(y_i|x_i)/dx_i$가 인과효과를 측정한다고 주장할 수 있다. 자료생성과정(DGP)에 관해 이런 강한 가정들을 한 회귀분석 및 계량경제에 관해 학습을 시작할 수 있다. 장래에 참조하기 위해서 이 가정들을 SR1~SR6이라고 명명하며, 여기서 SR은 'simple regression', 즉 단순회귀를 나타낸다.

계량경제학은 이런 가정들이 **준수될 수 없는** 자료 및 모형을 처리하는 데 대부분의 노력을 쏟고 있다. 이런 경우 β_1 및 β_2를 추정하고, 가설을 검정하며, 결과를 예측하는 데 사용되는 통상적인 방법을 수정하게 된다. 제2장 및 제3장에서는 이런 강한 가정이나 유사한 가정하에서의 단순회귀 모형을 학습할 것이다. 제4장에서는 모형화 문제와 진단검정을 소개할 것이다. 제5장에서는 한 개를 초과하는 설명변수를 갖는 **다중회귀분석**(multiple regression analysis)으로 모형을 확장할 것이다. 제6장에서는 다중회귀 모형에 관한 모형화 문제를 다룰 것이며, 제8장부터는 SR1~SR6이 어떻게 해서든 위배되는 상황들에 관해 살펴볼 것이다.

단순 선형회귀 모형에 관한 가정

SR1 : 계량경제 모형 모집단에서 수집한 모든 자료 쌍 (y_i, x_i)는 다음 관계를 충족한다.

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \cdots, N$$

SR2 : 강 외생성 무작위 오차 e_i의 조건부 기댓값은 영이다. $\mathbf{x} = (x_1, x_2, \cdots, x_N)$이라면 다음과 같다.

$$E(e_i|\mathbf{x}) = 0$$

강 외생성이 준수되는 경우 모집단 회귀함수는 다음과 같다.

$$E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i, \quad i = 1, \cdots, N$$

그리고

$$y_i = E(y_i|\mathbf{x}) + e_i, \; i = 1, \cdots, N$$

SR3 : 조건부 동분산 무작위 오차의 조건부 분산은 일정하다.

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

SR4 : 조건부적으로 상관되지 않은 오차 무작위 오차 e_i 및 e_j의 조건부 공분산은 영이다.

$$\text{cov}(e_i, e_j|\mathbf{x}) = 0, \quad i \neq j\text{인 경우}$$

SR5 : 설명변수는 변화해야만 한다 자료표본에서 x_i는 적어도 2개의 상이한 값을 취해야만 한다.

SR6 : 오차 정규성(선택적) 무작위 오차의 조건부 분포는 정규분포한다.

$$e_i|\mathbf{x} \sim \text{N}(0, \sigma^2)$$

무작위 오차 e와 종속변수 y는 모두 확률변수이며, 이 중 한 변수의 특성은 다른 변수의 특성에서 비롯된다. 그러나 이들 확률변수 간에는 흥미로운 차이점이 있는데, 그것은 y가 '관찰할 수 있는' 반면에 e는 '관찰할 수 없는' 것이라는 점이다. **회귀모수**(regression parameter) β_1 및 β_2를 아는 경우, y_i 및 x_i값에 대해 $e_i = y_i - (\beta_1 + \beta_2 x_i)$를 계산할 수 있다. 이는 그림 2.4를 통해 알 수 있다. 회귀 함수 $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$를 알 경우 y_i를 고정된 부분과 무작위 부분으로 분리할 수 있다. 하지만 β_1과 β_2는 결코 알 수 없으므로 e_i를 계산하는 것은 불가능하다.

오차항 e에 관해 위와는 다소 다르게 생각해 보는 것도 필요하다. 무작위 오차 e는 x 이외에 y에 영향을 미치는 모든 요소, 즉 우리가 그 밖의 모든 다른 요소라고 했던 것을 나타낸다. 이로 인해 개별적인 관찰값 y_i가 조건부 평균값 $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$와 달라지게 된다. 식료품 지출액의 예에서 1인당 지출액 y_i와 이것의 조건부 평균 $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$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 경제 모형에서나 모형 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 추가시키게 된다.
2. 오차항 e는 발생할 수 있는 대략적인 근사오차를 설명할 뿐이다. 왜냐하면 가정하고 있는 선형 함수 형태는 현실에 대해 단지 근사한 형태일 뿐이기 때문이다.

정리문제 2.3 식료품 지출액 모형에 관한 자료

앞에서 살펴본 경제 모형 및 계량경제 모형은 표본자료를 사용하여 절편 및 기울기 모수 β_1 및 β_2를 추정하는 기초가 된다. 예를 들어 설명하기 위해서 무작위 표본인 40개 가계의 식료품 지출액 및 주당 소득에 관한 자료를 검토해 보자. 표 2.1은 관찰과 요약된 통계량들을 보여주고 있다. 가계 규모를 통제하기 위해서 3인 가계만을 고려하였다. y값은 3인 가계의 주당 식료품 지출액을 달러로 측정한 것이다. 소득이 1달러 증가하더라도 식료품 지출액에는 숫자상으로 매우 작은 영향만을 미치기 때문에 소득을 달러 단위로 측정하는 대신에 $100 단위로 측정하였다. 따라서 첫 번째 가계의 경우 주당 소득은 $369이고, 주당 식료품 지출액은 $115.22이다. 40번째 가계의 경우는 주당 소득이 $3,340이고 주당 식료품 지출은 $375.73이다.

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
요약된 통계량		
표본평균	283.5735	19.6048
중앙값	264.4800	20.0300
최댓값	587.6600	33.4000
최솟값	109.7100	3.6900
표준편차	112.6752	6.8478