1. 계량경제모형(회귀분석모형)은 통계적 또는 확률적 관계식이다.

회귀분석은 한 종속변수가 하나 이상의 독립변수에 의해 어떠한 영향을 받고 또한 어떠한 관계로 나타나는지 분석하는 기법이다. 다시 설명하면 종속변수가 독립변수에 의해 어떠한 통계적 관계식으로 나타나는지를 밝히는 것이 회귀분석의 주요 목적이다.

회귀분석에서 밝히고자 하는 통계적 관계식은 바로 계량모형을 의미하는데 수학적 관계식과는 매우 다르다. 여기서 y_i는 개인별 소비지출이며, x_i는 개인별 소득이다. 또한 β_1,β_2는 파라메타이며, e_i는 오차항이다.

통계적(확률적) 관계식 : $y_i = \beta_1 + \beta_2 * x_i + e_i$
수학적(확정적) 관계식 : $y_i = \beta_1 + \beta_2 * x_i$

변수 사이에 존재하는 관계는 위와 같이 크게 두 종류로 구분할 수 있다. 하나는 확정적 관계이고 다른 하나는 통계적 관계이다. 어떠한 오차항도 허용되지 않는 변수 사이에 관련성이 존재할 때 확정적 관계라 한다. 다시 설명하면 변수 사이에 관련성이 수학적 함수관계로 표현되면 확정적 관계이다.

그런데 사회과학 현상에서 나타나는 변수들간의 관계는 확정적 관계로 표현될 수 있는 경우는 거의 없다.

소득수준이 같은 근로자 가구라도 소비지출액은 각각 다르게 나타난다. 즉 사회과학 현상은 변수사이에 존재하는 관련성에 오차가 있어 대부분 통계적(확률적) 관계로 나타난다. 그러므로 변수 사이에 존재하는 관계식을 밝힐 때에는 통계적 분석이 이루어져야 한다.

만약 두 변수 x,y사이에 통계적 관계가 존재한다면, 두 변수는 확률적으로 관련성이 있다는 뜻이다. 독립변수 x값을 알면 종속변수 y값을 정확하게 알 수는 없지만, y의 평균값 또는 기대값은 알 수 있다는 뜻이다.

예를 들어 근로자 가구 소득을 독립변수 x로, 소비지출액을 종속변수 y라 하자. 일반적으로 x가 증가하면 y도 증가하고, x가 감소하면 y도 감소한다. 그렇다면 소득수준이 한 단위 변할 때 소비지출은 어느 정도 변할 것인가? 즉, 한계소비성향을 나타내는 β_2는 얼마일까?

그런데 모집단 전체 가구를 모두 조사하여 β_1,β_2를 구하기에는 시간과 비용이 너무 많이 든다. 따라서 사회과학에서는 모집단으로부터 표본을 추출하여 표본 자료에 나타난 통계량(추정량)을 바탕으로 전체 가구의 한계소비성향을 추정할 수밖에 없다.

2. 모집단 회귀모형과 표본회귀모형을 정확히 구분해야 한다.

2.1 모회귀모형(모집단 또는 진회귀모형)

전체 집단을 대상으로 종속변수와 독립변수 사이에 인과관계를 나타내는 계량경제모형을 모집단 회귀모형 또는 간단하게 모회귀 모형(population regression model,)이라 부른다.

모집단에서 종속변수 y는 확률변수이다. 그러면 모집단에서 독립변수 x_i에 대응되는 종속변수 y의 확률분포는 어떠한 모습으로 나타날까? 교과서 그림자료을 참고하라.

교과서 그림에서 모집단에서 독립변수 x_i에 대응되는 종속변수 y_i의 값은 무수히 많을 수 있으며, 정규분포를 한다.

이 경우 x_i에 대한 종속변수 y_i의 기댓값은 $E[y_i|x_i]$이다. 이때 조건부 확률 기댓값 $E[y_i|x_i]$는 x_i와 항상 일대 일 대응되므로 $E[y_i|x_i]=E[y_i]$으로 표시할 수 있다.

모집단 자료를 이용하여 통계적 또는 확률적 회귀식을 표현할 경우 다음과 같다.

식(1) 모회귀기모형 : $y_i=\beta_1+\beta_2{}^*x_i+e_i$

식(2) 모회귀선(또는 진회귀선) : $E[y_i|x_i]=E[y_i]=E[\beta_1+\beta_2{}^*x_i+e_i]=\beta_1+\beta_2{}^*x_i$

식(2)의 모회귀선에서 β_1,β_2는 파라메타(또는 모수)이며, e_i는 오차항(또는 교란항)이다. 또한 모회귀선을 소득지출방정식이라고 가정할 경우 β_1은 절편이고 β_2는 한계소비성향을 의미하는 기울기이다.

모집단 자료를 이용한 모회귀모형(식1) 또는 모회귀선(식2)에 대한 특징을 알아보자.

첫째, 종속변수 y의 평균치는 독립변수 x에 관하여 선형함수이다.
종속변수의 평균치가 독립변수에 관하여 직선 형태인 선형관계로 나타난다. 즉, 독립변수 x값이 여러개 주어졌을 때 이에 대응되는 종속변수 y의 기대치를 연결하면 직선 형태로 나타나게 된다. 이러한 직선 형태의 선을 모회귀선 (population regeression line) 이라한다.

모회귀선은 직선 형태로 존재하지 않을 수도 있지만 회귀분석 이론을 전개하기 위해서 모회귀선은 직선 모양의 선형함수라 가정한다. 식(2)를 모집단이 갖는 회귀모형이므로 진회귀선(true regression line)이라고 한다.

둘째, 독립변수 x가 특정한 다른 값으로 주어진다 하더라도 이에 대응하는 종속변수 y의 분

산은 일정하다. 독립변수가 어떠한 값이라 할지라도 y의 확률분포 모양은 동일하다고 가정한다.

셋째, 종속변수 y는 확률변수로 정규분포한다. 또한 오차항 e_i도 정규분포를 한다.
회귀분석에서 독립변수는 주어진 값으로, 종속변수는 확률변수로 간주한다. 그러므로 종속변수는 모회귀선으로부터 오차가 존재할 수 있다.

식(1)에서 오차항 e_i는 모집단의 회귀선 주변에서 나타나기 때문에 교란항 (disturbance terms)이라고 부르기도 한다.

회귀분석에서 가장 주된 목적은 식(2)의 모회귀선에서 모수(또는 파라메타) β_1,β_2를 추정하는 것이다.

그러나 모집단 크기는 매우 커서 모집단 전체를 조사하여 모수 β_1와 β_2값을 구하는 것은 불가능하다. 그러므로 모집단으로부터 표본을 추출하여 표본의 회귀모형으로부터 모수 β_1와 β_2값을 추정할 수밖에 없다.

2.2. 표본회귀모형

사회과학에서는 거의 대부분이 모집단 자료를 이용하기 보다는 시간과 비용제약 등으로 표본을 이용한다. 따라서 표본 자료를 이용할 경우 식(1)과 식(2)의 표기도 명확히 구분해야 한다. 즉 모집단의 회귀선에 모수 β_1와 β_2를 구하는 대신에, 이에 대한 추정량 b_1,b_2를 표본회귀모형으로부터 구해야 한다.

모집단 일부 표본에 의해서 추정된 회귀선을 표본회귀선(sample regression line)이라 한다. 여기서 모회귀선과 표본회귀선은 명확하게 구분되어야 한다. 즉, 모수(=파라메타)와 계수(=파라메타 추정치)는 전혀 다르다.

모회귀선은 모집단이 갖는 유일한 회귀모형으로 존재한다. 하지만, 표본회귀선은 수 많은 방법으로 표본을 샘플링한 후 각각의 샘플링한 표본집단 관측치 자료로 부터 표본회귀선을 그릴 수 있게 된다.

모회귀선과 구분하기 위해 표본회귀모형 또는 표본회귀선은 식(3)~(4)와 같이 표현된다.

식(3) 표본회귀모형 : $y_i = b_1 + b_2 * x_i + \hat{e_i}$

식(4) 표본회귀선 : $\hat{y_i} = b_1 + b_2 * x_i$

식(3)의 표본회귀모형에서 b_1,b_2는 파라메타 추청치(또는 계수)이며, $\hat{e_i}$는 잔차항(residual term)이다.

표본회귀모형(식3)과 표본회귀선(식4)에 대한 특징을 알아보자.

첫째, $\hat{y_i}$은 표본회귀선 상에 있는 추정된 종속변수 값이다. y_i와 $\hat{y_i}$의 차이는 잔차 $\hat{e_i}$이다.
둘째, b_1,b_2는 표본회귀계수로서 모회귀 파라메타 β_1,β_2에 대한 추정량이다.
셋째, 표본회귀선은 수 많은 방법으로 표본을 샘플링한 후 각각의 샘플링한 표본집단 관측치 자료로 부터 표본회귀선을 그릴 수 있다.

그러면 특정 집단을 대상으로 개인별 소득과 소비지출 표본 자료를 조사하여 수집하였을 경우 b_1,b_2으로 나타낼 수 있는 표본회귀선은 어떻게 구할 수 있을까?

3. 표본회귀모형과 표본회귀선에서 표본회귀계수(b_1,b_2)를 어떻게 구해야 할까? 최소제곱의 원칙(=최소제곱추정량, 통상적인 최소자승 추정량)을 산출하는 방법을 알아보자.

- 최소제곱원칙을 이용한 표본회귀선의 절편 및 기울기, 즉 통상적인 최소제곱추정량(b_1,b_2) 산출식은 다음과 같다. 단, 특정 표본집단의 관측치 x_i, y_i는 표본 관측치이며, 이들의 표본평균은 $\bar{x}=\sum x_i/n$, $\bar{y}=\sum y_i/n$로 나타낼 수 있다.

$$b_1=\bar{y}-b_2\bar{x} \ , \ b_2=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

3.1 최소제곱원칙 또는 통상적인 최소자승추정법(Ordinary Least Squares Estimation; OLS) 추정량 산출 방법 이해하기

- 최소제곱의 원칙 또는 OLS 추정법은 표본회귀선으로 부터 벗어나 잔차의 제곱을 합계한 값을 가장 작게 하는 계수(b_1,b_2)를 산출하는 방법이다.
- 잔차를 제곱하는 이유는 잔차의 값이 양 또는 음이 되어, 이를 합산할 시에 0의 값이 나될 수 있기 때문이다.

3.2 최소제곱 추정량 도출 과정

- 표본회귀모형에서 잔차 $\hat{e_i}$는 식 (5)와 같다.

식(5) $\hat{e_i}$ = $Y_i-\hat{y_i}=y_i-b_1-b_2x_i$

- 식(5)의 잔차를 제곱 해서 합계하면 식(6)~(7)과 같다. 식(7)에서 b_1,b_2의 값은 미지의 계수 값이다. 또한 y_i및 x_i는 주어진 표본 관찰 값이며, n은 표본 관측치 개수이다.

식(6) $\sum_{i=1}^{n}\hat{e_i^2}$ =. $S(b_1,b_2))$

식(7) $S(b_1,b_2))$ = $\sum_{i=1}^{n}(y_i-b_1-b_2x_i)^2$

3.3. 최소제곱 추정량(미지의 표본회귀계수 b_1,b_2) 산출 방법

- 식(7)은 '잔차 제곱의 합계'를 나타내는 식이다. 따라서 이를 극소화하는 b_1과 b_2의 값을 구하기 위해서는 식(7)를 b_1 또는 b_2로 미분하여 식(8)~식(9)를 유도한다.

식(8) $\frac{\partial S}{\partial b_1} = 2nb_1 - 2\sum y_i + 2(\sum x_i)b_2$

식(9) $\frac{\partial S}{\partial b_2} = 2(\sum x_i^2)b_2 - 2\sum x_i y_i + 2(\sum x_i)b_1$

- 식(8)과 식(9)의 값이 0이 되는 최적해, 즉 잔차제곱의 합계를 최소화시키는 미지의 계수 (b_1,b_2)는 식(10)~(11)을 풀면 구해진다.

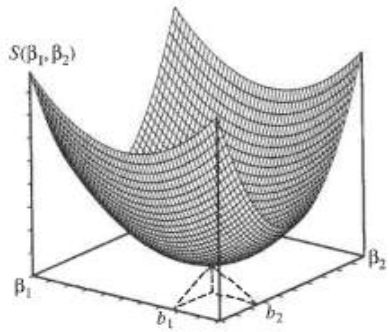
식(10) $2[\sum y_i - nb_1 - (\sum x_i)b_2] = 0$

식(11) $2[\sum x_i y_i - (\sum x_i)b_1 - (\sum x_i^2)b_2] = 0$

- 위에 두 방정식을 정리하면 식(12)~(13)과 같다.

식(12) $nb_1 + (\sum x_i)b_2 = \sum y_i$

식(13) $(\sum x_i)b_1 + (\sum x_i^2)b_2 = \sum x_i y_i$



- 식(12) 양변에 $\sum x_i$를 곱하고, 식(13) 양변에 n를 곱한후 두 방정식을 연립해서 풀면 식(14)와 같이 b_2가 구해진다.

식(14) $b_2 = \frac{n\sum x_i y_i - \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - \frac{\sum x_i y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$

- 이어서 식(15)~(16)을 이용하면 식(14)는 최종적으로 식(17)과 같다.

식(15) $\sum(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum X_i + n\bar{x}^2 = \sum x_i^2 - 2\bar{x} * (n\frac{1}{n}\sum x_i) + n\bar{x}^2$

$= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

식(16) $\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$

식(17) $b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

한편 식(12)의 양변을 n을 나누면 식 (18)과 같이 b_1가 구해진다.

식(18) $b_1 = \frac{\sum y_i - (\sum x_i)b_2}{n} = \bar{y} - b_2 * \bar{x}$

응용계량경제 4주차 퀴즈 문제

1. 모회귀선과 표본회귀선에 대한 질문이다.
   두 회귀선의 차이를 비교해서 수식을 이용하여 설명하시오.

2. 계량경제모형에서 통계적 또는 확률적 관계식을 쓰는 이유는? 이 경우 오차항 또는 잔차항에 대한 기본 가정을 설명하시오. 이때 오차항과 잔차항의 기본가정은 같은가?

3. 다음 주장에 대해 옳고 그름을 논하시오.
   3.1. 진회귀선은 오직 하나뿐이다.
   3.2. 표본회귀선은 무수히 존재할 수 있다.
   3.3. 표본회귀선에서 계수 b_2는 평균과 분산을 갖는 통계분포를 한다.
   3.4 표본회귀모형에서 종속변수 실제 값과 추정값의 차이는 오차(e_i)이다.

4. 다음과 같이 개인별 연간 소득과 소비지출액이 있다. 다음 질문에 답하시오.
   4.1 종속변수, 독립변수, 그리고 인과관계를 나타내는 표본회귀모형을 제시하시오.
   4.2 한계소비성향을 나타내는 표본회귀계수 b_2는 얼마인가?
   4.3 소득이 전혀 없는 경우에 소비지출액(b_1)은 얼마인가?
   4.4 소득이 10,000만원이면 소비지출액은 어느 정도 될까?

개인구분(i)	소득(x_i)	소비지출액(y_i)
1.	1000만원	500만원
2.	2000만원	500만원
3	3000만원	1000만원
4	4000만원	1000만원
5	5000만원	1000만원

5.. 만일 표본을 5명만 한정해서 문제 4에서 같이 한계소비성향을 추정했을 경우 과연 이 추정 값을 신뢰할 수 있는가에 대해 논하시오.