제4장 회귀분석

회귀분석의 개요(the nature of regression analysis)

- 회귀분석(regression analysis)이란 하나의 종속변수와 하나 또는 2개 이상의 독립변수들 간의 관련성을 규명할 수 있는 수학적 모형을 측정된 변수들의 자료로부터 회귀식을 추정하는 통계적 방법임.
- 회귀분석은 본질적으로 인과관계가 있는 두 변수간의 함수식을 분석대상으로 하며, 다음과 같이 두 가지 측면에서 이용됨.
  1. 첫째, 관측된 두 변수의 값을 기초로 두 변수간의 함수관계가 성립하는지, 만약 함수관계가 성립한다면 어떤 특징을 갖는 함수관계(예 : 1차 선형관계)인지 이해하는데 이용됨.
  2. 둘째, 그 값이 알려진 독립변수를 기초로 종속변수의 값을 추정 또는 예측하는 데 이용됨.
- 결국 회귀분석은 "종속변수가 하나 이상의 독립변수에 어떻게 의존하고 있는가를 분석"하는 과정을 의미함.

통계적 관계의 함수식을 구하는 것이 목적임

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

예를 들어 원의 반지름을 r 이라 하고 원의 면적을 S 라 하자. 그러면 원의 면적과 반지름 사이에는 $S = \pi r^2$ 이라는 수학적 함수관계로 나타난다. 이러한 수학적 관계식은 오차의 개념이 없기 때문에 확정적 관계이다. 그런데 사회과학 현상이나 자연과학 현상에서 나타나는 변수들은 대부분 확정적 관계로 표현될 수 있는 경우는 거의 없다.

소득수준이 같은 근로자 가구라도 소비지출액은 각각 다르게 나타난다. 또한 동일한 등고선에 위치한 두 지역의 기온은 같은 온도로 나타나지 않는다. 뿐만 아니라 동일한 IQ 를 지닌 학생들이라 해서 성적이 같게 나타나지 않는다. 이와 같이 사회과학 현상이나 자연과학 현상은 변수 사이에 존재하는 관련성에 오차가 있어 대부분 통계적 관계로 나타난다. 그러므로 변수 사이에 존재하는 관계식을 밝힐 때에는 통계적 분석이 이루어져야 한다.

↑ 회귀선의 방정식은 통계적 관계를 기반삼여 도출하는 것이다.

만약 두 변수 X, Y 사이에 통계적 관계가 존재한다면 두 변수는 확률적으로 관련성이 있다는 뜻이다. 독립변수 X 값을 알면 종속변수 Y 값을 정확하게 알 수 있다는 뜻이 아니다. ← '관계성'의 측면.

상관계수는 회귀분석에서 매우 중요한 통계량이다. 왜냐하면 상관계수 값에 의하여 적합한 회귀함수가 도출되기 때문이다. 회귀분석에서 독립변수가 한 단위 변화함에 따라 종속변수에 미치는 영향력 크기를 회귀계수 (regression coefficient) 라 한다. 일반적으로 두 변수 사이에 상관관계가 거의 없을 때 회귀계수는 의미가 없게 된다.

→ '상관계수'를 통해 두 변수의 산점도가 직선형태를 띄고 있다는 것이 증명돼야 회귀분석이 가능하다.

회귀분석에서는 두 변수 또는 여러 변수 사이에 존재하는 관련성 정도를 나타내는 상관계수를 산출한다.

독립변수와 종속변수 사이에 관련성이 존재해야 회귀분석이 적용될 수 있다. 만약 두 변수 사이에 관련성이 거의 없는데 회귀분석을 적용하여 구한 회귀계수 추정치는 의미가 없게 된다. 그리고 표본회귀선의 적합성을 결정하는 결정계수 (determination coefficient) 는 상관계수로부터 구할 수 있다.

→ '상관계수'가 '회귀분석'의 기반이다.

두 변수의 연관성을 대략적으로 파악하는 단계.

⇒ 즉, 산점도를 먼저 그리고 상관계수를 분석한 후, 회귀분석을 실시한다.

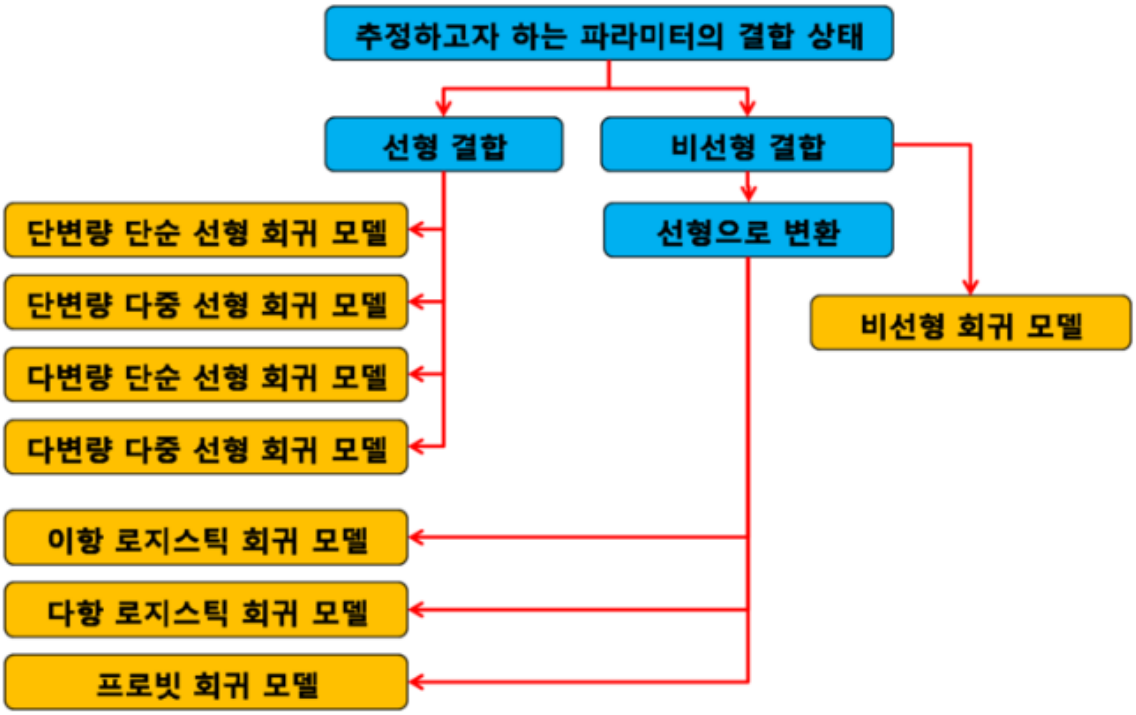
두 변수의 연관성을 직선의 함수로 나타내고 해당 함수를 통해 종속변수의 값을 예측하는 단계

회귀분석을 본격적으로 시행하기 전

'상관계수'를 통해서 두 변수 간의 관계를 먼저 파악해야 함!

변수 간의 관계를 대략적으로 파악해야 한다. 즉, 산점도에서 변수 간의 관계가 직선관계라고 판단되면 선형회귀분석을 할 수 있지만 곡선관계로 판단되면 선형회귀분석을 해서는 안 될 것이다.

X. 회귀 분석 모델의 종류


추정하고자 하는 파라미터의 결합 상태
선형 결합
비선형 결합
단변량 단순 선형 회귀 모델
단변량 다중 선형 회귀 모델
다변량 단순 선형 회귀 모델
다변량 다중 선형 회귀 모델
선형으로 변환
비선형 회귀 모델
이항 로지스틱 회귀 모델
다항 로지스틱 회귀 모델
프로빗 회귀 모델

$$Y = a_0 + a_1 X_1 + a_2 X_2$$

한편, 아래와 같은 식은 위의 두 식과는 다르게 독립 변수 X_1 뿐 아니라 X_1 의 제곱항이 들어 있는 함수입니다.

계수의 관점으로 보면, '선형결합'이 된 상태이지만, 독립변수의 관점으로 보면, '비선형 결합'이 된 상태이다.

$$Y = a_0 + a_1 X_1 + a_2 X_1^2$$

Y ← 관찰값(=실제값)

위의 세가지 식은 모두 **독립 변수의 계수(즉, 파라미터) 관점에서 보면 모두 선형 결합(linear combining) 이 된 상태** 입니다. 선형 결합이란 것은, 계수들이 곱셈이나 나눗셈, 로그, 삼각함수 등과 같은 것이 아니라 단순히 **덧셈과 뺄셈으로만 결합**되어 있는 것을 의미합니다.

만일 위의 세가지 식을 계수가 아닌 **독립 변수 자체의 관점에서 본다면**, 첫번째와 두번째 식은 선형 결합 이지만 **세번째 식은 선형 결합이 아닙니다.** 왜냐하면 식에 독립 변수 X_1 의 제곱항이 들어있기 때문입니다. 이러한 결합을 **비선형 결합(nonlinear combining)** 이라고 합니다.

선형 회귀 분석(linear regression analysis) 이라는 것은 위에서 든 예들처럼 **계수(coefficient)**들이 덧셈과 뺄셈만을 이용해 **선형으로 결합되어 있는 회귀 모델 분석**을 의미합니다. 독립 변수의 제곱항이 있는 세번째 회귀 모델을 **비선형 회귀 분석(nonlinear regression analysis)**으로 설명하는 경우도 종종 있습니다. 하지만, 회귀 모델에서는 계수의 결합 방식을 통해 선형이냐 비선형이냐를 결정하기 때문에, 엄밀한 의미에서는 세번째 수식도 선형 회귀 모델 함수 입니다.

핵심!!!