회귀분석

#독립변수 #종속변수 #회귀분석 #선형회귀분석 #공선성
#OLS #결정계수 #잔차 #SSE #SSR #SST

회귀분석이란

독립변수와 종속변수의 데이터를 모아 살펴보면, 각각의 데이터마다 오차가 있긴 하나 전체적으로 봤을 때 어떤 경향성이 나타난다. 개별 데이터가 조금씩 경향을 벗어날 수는 있지만 평균적으로 이들은 그러한 경향에 돌아온다, 즉 회귀한다는 점에서 회귀분석의 어원을 알 수 있다.

장점

회귀분석은 데이터에서 변수들의 관계를 식으로 나타낼 수 있으며, 그 식이 얼마나 해당 데이터를 잘 설명하는지도 측정할 수 있다. 이를 통해 어떤 현상을 설명하거나, 새로운 데이터에 대해 종속변수의 값을 예측하는 등의 일이 가능하다. 따라서 거의 모든 데이터 활용 시나리오에서 회귀분석이 쓰이고 있다.

단점

수치적인 데이터의 값을 예측하거나 설명하는 것에는 회귀분석이 매우 절대적인 입지를 가지고 있지만, 데이터를 보고 분류하는 등 특정한 상황에 대한 확률을 측정하는 경우, 독립변수와 종속변수가 명확하게 주어지지 않는 경우, 데이터가 비정형인 경우 등에는 적용이 힘들다. 또한 실제 데이터가 회귀분석에 맞게 깔끔한 경향을 나타내는 경우도 드물며, 변수들의 관계로 오류가 발생할 수 있다.

선형회귀분석

종속변수 Y와 독립변수 X1, X2, X3, … 등이 있다고 하자. 선형회귀분석은 이들의 관계를 선형식, 즉 1차식으로 아래와 같이 표현한다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots$$

선형회귀는 해석이 쉽고, 계산이 빠르다는 점에서 장점이 있다. 선형관계가 아닌 데이터도 로그를 이용해 데이터를 변환하여 선형회귀를 적용할 수 있어 범용성 또한 높다.

선형회귀분석의 과정

변수 설정

특정한 종속변수의 분석에 사용할 독립변수들을 결정한다. 대개의 경우 독립변수가 많아질수록 예측력이 높아지지만, 모형이 너무 복잡해지거나 변수 사이의 공선성(Collinearity)로 인해 각 변수가 독립적이라는 회귀분석의 가정이 적절하지 않는 경우가 생길 수 있다. 따라서 적절한 변수를 설정하는 것이 중요하다.

또한 변수의 값을 변환시켜야 할 수도 있다. 로그를 이용해 선형관계로 만드는 것도 필요할 수 있으며, 특정한 변수가 값의 단위가 큰 경우, 전체 분석 과정에서 해당 변수의 영향력이 너무 커지기 때문에 이를 보정하기 위해 전체 변수들이 −1에서 1 등의 같은 스케일을 가지도록 조정해야 할 수도 있다. 이 때 정규분포를 활용한 표준화도 고려될 수 있다.

계수 결정

회귀분석의 오차가 최소화가 되도록 하는 계수를 찾는다. 오차를 계산하는 방법에는 다양한 것이 있으나 많은 경우 최소제곱법, 그 중에서도 OLS(Ordinary Least Squares)를 사용한다. 단순하면서 계산이 빨라 가장 많이 사용된다. 이는 측정값(실제 데이터)에서 함수값(회귀분석에 의한 예측값)을 뺀 것을 제곱하여 오차를 구하며, 이 오차의 합이 최소가 되는 계수를 찾아 구한다.

평가

회귀분석 이후 해당 데이터를 식이 잘 설명하는지 파악하기 위해 결정 계수 R^2의 값을 계산한다. 결정 계수는 0부터 1 사이의 값을 가지며 전체 데이터의 분포 중에서 회귀식이 설명할 수 있는 분포의 비율을 의미한다. 따라서 1일 경우 데이터가 완벽하게 회귀식의 그래프 위에 있는 것을, 0이면 완전히 맞지 않는 경우가 된다. 나아가 회귀식에 대해 F 검정, 계수에 대해 T 검정 등의 통계적 검정을 진행하고, 잔차(데이터와 예측값의 차이)가 정규분포를 따르는 지 확인한다.

회귀분석 공식과 용어

표기

통계학에서 변수를 보면 머리에 뭘 쓰고 있는 경우가 있다. 뜻은 다음과 같다.

Y_i=i번째 변수의 값, $\overline{Y}$=평균, $\widehat{Y}$=예측값

잔차 (Residual)

종속변수의 예측치와 실제 관측치(데이터의 값)의 차이를 잔차라고 한다. 이 잔차의 제곱 합을 최소화시키는 계수를 찾는 것이 최소제곱법, 곧 OLS가 회귀식을 구하는 방법이다. Y = a + bX라는 회귀식이 있을 때 잔차는 다음과 같다.

$$r_i = \left|Y_i - \widehat{Y}_i\right| = |Y_i - a - bX_i|$$

SSE, SSR, SST

위의 잔차의 제곱 합을 구한 것을 SSE (Sum of Squared Error)라고 한다.

$$\mathrm{SSE} = \sum \left(Y_i - \widehat{Y}_i\right)^2$$

이는 실제 값과 예측된 값의 차이가 얼마나 되는지를 나타낸다. 한편 예측된 값이 실제 데이터의 평균과는 얼마나 차이가 나는지를 SSR (Sum of Squares due to Regression)이라 하고 아래와 같이 계산한다.

$$\mathrm{SSR} = \sum \left(\widehat{Y}_i - \overline{Y}\right)^2$$

SSE가 회귀식이 설명하지 못 하는 부분을 의미한다면, SSR은 회귀식이 설명할 수 있는 부분을 의미한다. 이에 따라 SSE를 때로는 RSS(Residual Sum of Squares), SSR을 ESS(Explained Sum of Squares)라고 하기도 한다.

SST는 실제 데이터가 평균으로부터 떨어진 정도를 의미하며 다음과 같다.

$$\mathrm{SST} = \sum (Y_i - \overline{Y})^2$$

결정 계수 (R squared)

OLS에서 SST = SSE + SSR가 성립한다. 이 때, 결정 계수는 다음과 같다.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

전체 데이터의 분포 (SST) 중, 회귀식이 설명할 수 있는 부분 (SSR)이 차지하는 비율이라고 이해할 수 있다. 따라서 이 값이 1이 될 경우 데이터의 모든 값을 회귀식으로 설명할 수 있는 것이다.

그래서 계수는 어찌 구하나요

결정 계수의 값을 올리려면 SSE를 최소화해야 한다. 식을 다시 쓰면

$$\mathrm{SSE} = \sum \left(Y_i - \widehat{Y}_i\right)^2 = \sum (Y_i - a - bX_i)^2$$

SSE를 b로 편미분했을 때 0이 나올 때 SSE가 최소가 된다. 이를 이용하여 열심히 미분을 하고 정리를 하면 b는 다음과 같이 구해진다.

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

남은 a는 다음과 같이 계산한다.

$$a = \overline{Y} - b\overline{X}$$

b의 식을 보면 아래는 분모는 X의 분산, 분자는 X와 Y의 공분산으로 놓고 N이 약분된 것이라는 걸 알 수 있다. 그러면 b가 양수고 큰 값이 나오는 경우는, X와 Y가 공분산이 양수, 즉 양의 상관관계가 있고, X가 Y에 비해 상대적으로 중심에 몰려있을 때라고 볼 수 있다.

너무 복잡해요

대개의 경우 선형회귀는 엑셀을 포함한 모든 데이터 분석 소프트웨어에서 기본으로 제공하며, 식 또한 간단하기 때문에 일반적인 프로그래밍 언어로도 만들어 쓰기 쉽다. 중요한 것은 의미이며 직접 계산을 하는 것은 필요하지 않다.