결정계수란?

선형회귀분석(linear regression analysis)에서 회귀직선의 적합도(goodness-of-fit)를 평가하거나 종속변수에 대한 설명변수들의 설명력을 알고자 할 때 결정계수(R sqaured, coefficient of determination))를 이용합니다.

결정계수는 설명변수의 변동량으로 설명되는 종속변수의 변동량을 의미하고, 식은 아래와 같습니다.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{\text{회귀선에 의해 설명되는 변동}}{\text{전체 변동}}$$

예를 들어 결정계수 R2=0.45 인 경우, 소아의 인지기능(종속변수) 변동은 언어학습시간(독립변수) 변동에 의해 45% 정도 설명된다고 해석할 수 있습니다.

결정계수는 피어슨 상관계수 (Pearson correlation coefficient)의 제곱과도 같습니다.위에서 언급한 것과 같이 상관계수는 -1~1 사이의 값을 갖기 때문에 결정계수는 0부터 1까지의 값을 갖게 됩니다.

수정된 결정계수란?

결정계수는 독립변수 개수가 많아질수록 그 값이 커지게 됩니다. 따라서 종속변수의 변동을 별로 설명해 주지 못하는 변수가 모형에 추가된다고 하더라도 결정계수값이 커질 수 있습니다.

이러한 문제를 보정한 것이 수정된 결정계수(adjusted coefficient of determination)입니다. 표본의 크기와 독립변수의 수를 고려하여 계산하게 되는데 그 식은 아래와 같습니다.

$$\text{adjusted } R^2 = 1 - \frac{n-1}{(n-p-1)(1-R^2)}$$

단순회귀분석을 하는 경우에는 일반 결정계수를 사용하면 되지만, 다중회귀분석을 수행하는 경우에는 수정된 결정계수를 함께 고려하는 것이 좋습니다.

결정계수(Coefficient of Determination,R2)

- 총변동중에서 회귀선에 의해 설명이 되는 변동이 차지하는 비율
- R2(R-Sq)의 범위는 0≤ R2 ≤ 1
- X와 Y간의 상관관계가 클수록 R2(R-Sq)의 값은 1에 가까와짐
- R2(R-Sq)의 값이 0에 가까워 질수록 회귀선은 쓸모가 없고, R2(R-Sq)의 값이 클수록 (R2≥0.65) 쓸모있는 회귀식이 된다

인자가 하나일때는 상관계수의 제곱값과 결정계수값이 같습니다.

수정결정계수

← 독립변수의 개수가 많아질 수록 '결정계수'가 높아진다. 이를 보완하기 위해 '수정결정계수'가 탄생하게됨.

- 결정계수는 상향편의 된 추정치 이므로 표본 결정계수의 값은 항상 모집단의 결정계수보다 클 수 밖에 없음. 따라서, 보다 정확한 추정치를 얻기 위해서는 수정결정계수를 사용해야 함.
- 수정결정계수의 값은 결정계수보다는 작고 때에 따라서는 음의 값도 나타날 수 있음
- 표본의 크기가 200개 이상일 때는 두 결정계수의 차이가 미미함.
- 표본이 200개 미만일 때는 반드시 수정결정계수를 보고서에 포함해야 함 (독립변수가 2개 이상이면 수정결정계수를 본다)