**회귀분석**

**회귀분석(Regression Analysis)**

**Theory**

하나 혹은 그 이상의 원인(독립변수)이 결과(종속변수)에 미치는 영향을 추정하여 식으로 표현하는 통계기법이다. 변수들 사이의 인과관계를 밣히고 모형을 적합하여 관심 있는 변수를 예측하거나 추론하는 분석방법입니다.

회귀분석은 종속변수의 오차를 최소화 시키는 직선을 찾는 방법이며, 오차의 최소제곱합(MSE)의 최소값을 갖게하는 독립변수들의 계수를 구하게됩니다.

선형회귀분석의 가정은 **독립변수와 종속변수 간의 선형성**, 오차의 등분산성, 오차의 독립성, 오차의 정규성, 극단값 검정 이다.

* 독립변수와 종속변수 간의 선형성 : 상관관계 시각화 (히트맵, 상관계수)
* 오차의 등분산성 : 오차 그래프
* 오차의 정규성 : 샤피로-윌크 검정, 정규성 검정

Normal Q-Q plot : Normal Q-Q plot은 표준화된 잔차의 확률도이다. 정규성 가정을 만족한다면 그래프의 점들은 45도의 직선을 이루는 형태를 띄어야 한다.

* 오차의 독립성 : 더빈-왓슨 검정(Durbin-Watson) - 0이면 양의 상관, 2면 독립, 4면 음의 상관을 갖는다. 보통 1.5 ~ 2.5의 값이 나오면 독립이라고 판단한다.
* 극단값: Cook’s Distance – Cook’s Distance는 극단값을 나타내는 지표로, x축은 관측값을 순서대로 나열한 것이며, y축은 해당 지점의 쿡의 거리를 나타낸다. 쿡의 거리는 한 관측치가 회귀모형에 미치는 영향을 나타내는 측동이며, 일반적으로 1이상일 경우 매우 큰 영향을 주는 관측값으로 간주한다.

**단순선형회귀분석(Simple Linear Regression)**

단순선형회귀는 하나의 독립변수가 종속변수에 미치는 영향을 추정할 수 있는 통계기법이다.

회귀식을 다음과 같이 **Y = a + bx + c** (절편, 계수, 오차) 표현하며, a와 b를 회귀계수라고 한다.

**단순선형회귀는 이 a와 b를 찾고, 통계적으로 유의미한지**를 파악하는 분석기법이다.

* **회귀분석시 검토사항**:

1. 회귀식의 회귀계수가 유의한가 :

**p-value**가 0.05 유의수준보다 작으면 해당 회귀 계수가 통계적으로 유의하다고 볼 수 있다. 또한, 회귀계수가 클수록 독립변수가 종속변수에 더욱 큰 영향을 준다.

1. 모형은 전체 데이터를 얼머나 설명할 수 있는가 : 결정계수 = SSR/SST

**결정계수(R^2)**을 확인하며, 0 ~ 1 사이의 값을 가진다. 추정된 회귀식이 전체 데이터셋에서 설명할 수 있는 데이터의 비율을 의미한다.

1. 회귀모형은 통계적으로 유의한가 :

회귀분석의 결과로 산출되는 F-통계량의 p-value가 0.05보다 작으면 해당 회귀식은 통계적으로 유의하다고 볼 수 있다.

1. 모형이 데이터에 적합한가 :

오차 그래프를 그리고, 회귀진단을 수행하여 판단.

**▷ 결측치, 이상치 -> 상관성 검증 -> 회귀분석 -> 회귀분석결과 분석 -> 회귀진단(오차)**

**다중선형회귀분석(Multiple Linear Regression)**

다중선형회귀분석은 두 개 이상의 독립변수가 종속변수에 미치는 영향을 추정하는 통계기법으로, 다음와 같이 여러 독립 변수가 사용된 회귀식을 생성한다. **Y= ax0 + bx1 + cx2**

* **다중선형회귀분석식 검토사항**

1. 데이터가 전제하는 가정을 만족하는가 :

독립변수와 종속변수 간 선형성, 오차의 독립성/등분산성/정규성 등을 만족하는지 확인

1. 모형 내 회귀계수가 유의한가 :

t검정통계량 유의확률

1. 모형은 데이터를 얼마나 설명할 수 있는가 :

결정계수

1. 회귀모형은 통계적으로 유의한가:

f검정통계량 유의확률

1. 모형이 데이터에 적합한가 :

오차 그래프를 그리고, 회귀진단 수행

* **다중공선성(Multicolinearity) 검중**

다중공선성은 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제이다. 이러한 다중공선성의 문제가 존재하면 정확한 회귀계수의 추정이 곤란하다. 따라서 독립변수들 간 상관관계가 있는지를 파악한 후, 다중공선성 문제가 있는 변수를 제거해주거나 처리를 해주어야 한다.

1. 독립변수들 간의 **상관계수**를 구하여 상관성 파악
2. **허용오차**(0~1사이 값)를 구했을 떄, 0.1이하이면 다중공선성 문제가 심각하다고 할 수 있다. 허용오차란 한 독립변수의 분산 중 다른 독립변수들에 의해서 설명되지 않은 부분을 의미하므로, 그 값이 작을수록 공선성은 높다고 볼 수 있다.
3. **분산팽창요인(VIF : Variance Inflation Factor)**은 허용오차의 역수로 그 값이 클수록 독립변수 간의 상관성이 높다. 일반적으로 VIF가 10이상일 경우 공선성 문제가 심각하다고 할 수있다.

사실 회귀계수가 통계적으로 유의미하다면 VIF가 크더라도 특별히 대처할 필요는 없다. 하지만,

그렇지 않다면 변수들을 더하거나 빼서 새로운 변수를 만들 수 있다. 혹은 변수를 제거하는 방법이 있는데, 이는 자료의 다양성을 해치고 분석하려던 가설이나 이론에 영향을 미칠 수 있기 때문에 가급적이면 자제하는 것이 좋다. 기준은 VIF가 10이상인 특성 그 다음은 그냥 높은 순

**▷ 결측치, 이상치 -> 상관성, 다중공선성 검증 -> 회귀분석 -> 회귀분석결과 분석 -> 회귀진단(오차)**

**최적회귀방정식의선택**

모형 내 설명변수의 수가 증가할수록 데이터 관리에는 많은 노력이 요구된다. 따라서 상황에 따라 종속변수에 영향을 미치는 유의미한 독립변수들을 선택하여 최적의 회귀방정식을 도출하는 과정이 필요하다.

일반적으로 변수를 선택할 때는 F-통계량이나 AIC와 같은 특정 기준을 근거로 변수를 제거하거나 선택한다.

F-통계량의 유의확률이 유의수준보다 큰 변수는 통계적으로 유의하지 않으므로 제거해야하고, AIC와 같은 벌점화 기준을 가장 낮게 만드는 변수 조합을 선택해야 한다.

모형의 복잡도에 따라 벌점을 주는 방식으로 AIC와 BIC방법이 주로 사용된다. 후보 모형에 대해 AIC 혹은 BIC를 계산했을 때 그 값이 최소가 되는 모형을 선택한다. 일반적으로 AIC를 사용하는 것이 보편적이다.

* **단계적 변수 선택(Stepwise Variable Selection)**

1. 전진선택법(Forward Selection)

**절편만 있는 상수모형에서 시작하여** 중요하다고 여겨지는 설명변수부터 차례대로 모형에 추가하는 방법

1. 후진제더법(Backward Elimination)

**모든 독립변수를 포함한 모형에서 출발하여** 종속변수에 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때까지 반복한다.

1. 단계적 방법(Stepwise Method)

전진선택법에 의해 변수를 추가하면서 새롭게 추가된 변수에 의해 기존 변수의 중요도가 약화되면 해당 변수를 제거한다. 이처럼 단계별로 변수의 추가 or 제거를 수행하며 기준 통계치를 가장 많이 개선시킬 때까지 반복한다.

**▷ 결측치, 이상치 -> 상관성, 다중공선성 검증 -> 회귀분석 -> 단계적선택법 -> 회귀진단(오차)**