<YDMS 2023-1> 3주차 과제

회귀

주하연

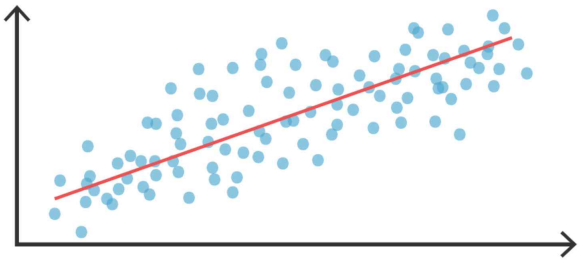
1. 회귀분석 개념

회귀분석(Regression)

하나나 그 이상의 독립변수의 종속변수에 대한 영향의 추정을 할 수 있는 통계기법을 말한다.

회귀분석은 데이터를 수집하고 모델을 적합시키고 모델을 평가하는 과정을 거치는데 이렇게 만들어진 모델은 '예측'에서 사용된다.

하나의 독립변수를 가진 회귀분석에서, 하나의 방정식은 독립변수와 종속변수의 결합분포를 보여주는 지점들의 분포구성을 통해 지나가는 하나의 선을 설명하고 있다.
이 방정식은 $Y_i=a+bX_i+e_i$라는 형태를 갖는다. Xi는 독립변수의 값을 말한다. a는 Y축을 지나가는 회귀선의 지점이며, b는 회귀선의 기울기이고, e_i는 회귀선 예측의 오차이다.



*회귀선 : 흩어진 점들에 가장 적합한 선이다. 회귀선에서 가장 유용한 값은 기울기 b이다. b는 종속변수에 독립변수가 미치는 영향을 나타낸다.

- 회귀분석을 통해 회귀선을 추정하는 방법 : 최소자승법(Ordinary Least Squares)

실제 관측값과 회귀선에 의한 예측값의 잔차(오차)의 제곱의 합을 최소화하는 회귀선을 찾는 방법

->예측값과 실제 관측값 사이의 오차가 최소화되는 최적의 회귀선을 얻을 수 있다.

- 회귀분석의 사용(목적)

  주요 목표 : 각 독립변수와 종속변수의 관계 결정

  * 예측 : 독립 변수의 값을 기반으로 종속 변수의 값을 예측하는 데 사용된다.

  ex) 공부 시간과 성적 간의 관계를 알고 있다면, 공부 시간을 독립 변수로 사용하여 학생의 시험 성적을 예측할 수 있다.

* 인과 관계 분석 : 독립 변수와 종속 변수 간의 인과 관계를 파악하는 데 사용된다.

ex) 광고 비용과 판매량 간의 관계를 분석하여 광고 비용이 판매량에 미치는 영향을 알 수 있다.

* 변수 간 상호작용 분석 : 독립 변수 간의 상호작용을 분석하는 데 사용된다. 독립 변수 간의 상호작용은 종속 변수에 대한 영향을 변경시킬 수 있다. -> 예측 모델의 정확성 향상

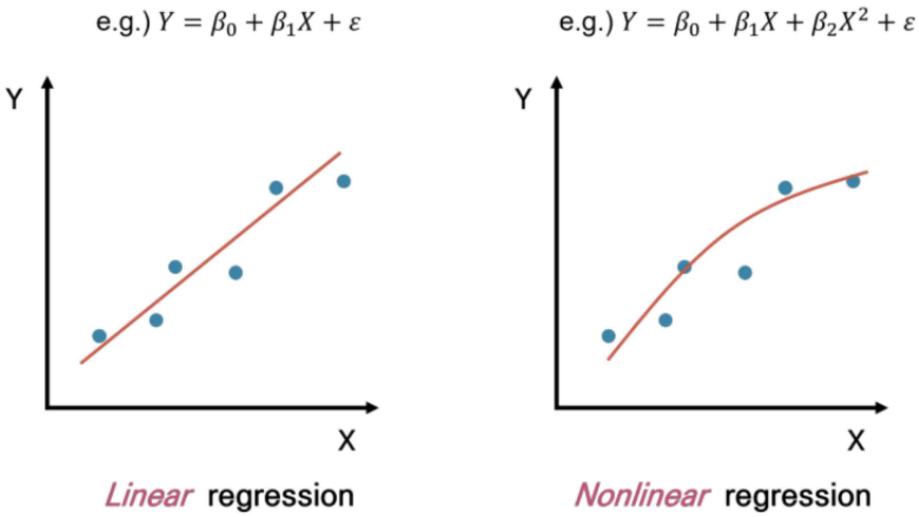
* 이상치 탐지 : 이상치를 탐지하는 데 사용될 수 있다. (이상치는 일반적인 패턴에서 벗어난 값으로, 데이터의 정확성과 모델의 성능을 저하시킬 수 있다.)

* 변수 중요도 평가 : 독립 변수의 중요도를 평가하는 데 사용된다. 각 독립 변수의 회귀계수를 통해 변수가 종속 변수에 얼마나 영향을 미치는지 평가할 수 있다.

- 회귀분석의 종류

선형 회귀 : 직선 모양

비선형 회귀 : 곡선 모양

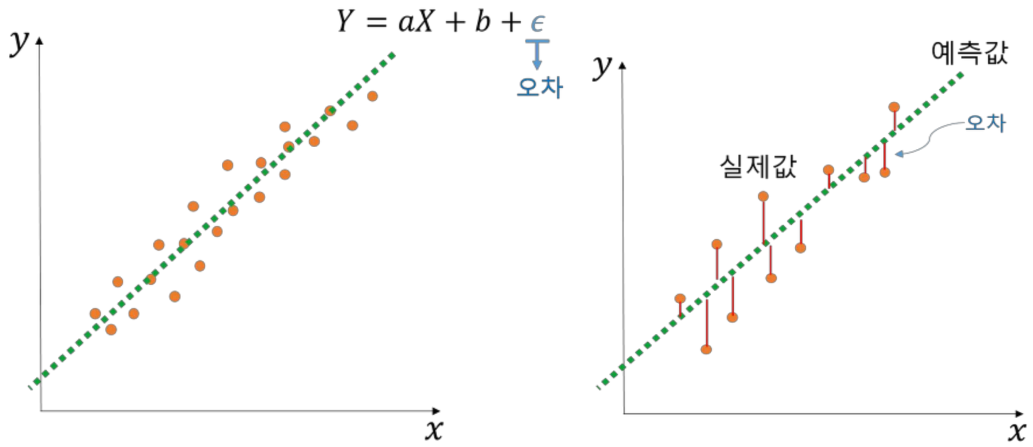


A. 단순 회귀 분석

하나의 종속변수와 하나의 독립변수 사이의 관계를 분석

독립변수와 종속변수가 선형적인 관련성이 있다는 전제 하에 변수들간의 관계를 선형 함수식으로 모형화하기 위한 분석방법

컴퓨터는 데이터와 일차 방정식 선 사이의 차이를 가장 적게하는 식을 계속 찾아 나가면서 최종적인 함수식을 찾는다.



i. 단순회귀분석의 기본 가정

1. 선형성 : 데이터가 직선적인 패턴을 따른다고 가정한다. 만약 비선형적인 패턴을 가지고 있다면, 추가적인 변환을 통해 선형성 가정을 충족시킬 필요가 있다.
2. 독립성 가정 : 독립 변수 간에 상관관계가 없어야한다.
3. 등분산성 가정 : 잔차들이 독립 변수 값과 관계없이 일정한 분산을 가져야 한다. 등분산성 가정이 위배될 경우, 예측 구간의 신뢰도가 왜곡될 수 있다.
4. 잔차의 정규성 가정 : 잔차들은 정규 분포를 따라야 한다. 이는 잔차들이 평균 0을 중심으로 대칭적으로 분포하며, 정규 분포를 따르는 것을 의미한다. 만약 잔차가 비정규 분포를 따른다면, 모델의 예측력과 신뢰도가 저하될 수 있다.

ii. 단순회귀분석 과정

1. 데이터 수집

   분석하고자 하는 독립변수와 종속변수에 대한 데이터를 수집한다.

2. 산점도 작성

   수집한 데이터를 이용하여 독립변수와 종속변수 간의 산점도를 작성한다. 산점도는 데이터의 분포를 시각화하고, 두 변수 간의 관계를 파악하는 데 도움을 준다

3. 회귀 모델 적합

적합한 회귀 모델을 선택하고 데이터에 모델을 적합시킨다. 선형 회귀 모델은 독립 변수와 종속 변수간의 선형관계를 가정하며, 최소제곱법을 사용하여 회귀 계수를 추정한다.

4. 회귀 계수 추정

   회귀 모델을 적합시키면 독립 변수의 회귀 계수를 추정할 수 있다. 회귀 계수는 독립 변수의 변화가 종속 변수에 어떤 영향을 미치는지를 나타낸다. 일반적으로 회귀분석 결과에서는 회귀계수와 함께 통계적 유의성을 평가하기 위한 p-value나 신뢰구간 등의 정보도 제공된다.

5. 모델 평가

   적합한 회귀 모델의 성능을 평가한다. 이를 위해 잔차분석을 수행하거나 결정계수 등의 평가 지표를 사용할 수 있다. 잔차는 실제 종속 변수 값과 회귀 모델로 예측한 값 간의 차이를 의미하며, 잔차 분석을 통해 모델이 데이터를 얼마나 잘 설명하는지 확인할 수 있다.

iii. 단순회귀분석 장점

1. 간단하고 이해하기 쉬움 : 모델의 개념과 결과해석이 비교적 직관적이다.
2. 변수 중요도 평가 : 단순회귀분석은 독립변수의 중요도를 평가하는 데 사용된다. 독립변수의 회귀 계수를 통해 변수가 종속변수에 얼마나 영향을 미치는지 평가할 수 있다. 이를 통해 변수의 상대적인 영향력을 비교할 수 있고, 중요한 변수를 식별할 수 있다.
3. 기초 분석 도구로 활용 가능 : 복잡한 다변량 분석이나 머신러닝 기법을 적용하기 전에, 데이터의 기본적인 패턴과 변수 간의 관계를 파악하는 데 활용될 수 있다.

iv. 단순회귀분석 단점

가정의 제한성 : 단순회귀분석은 몇 가지 가정을 전제로 한다. 이러한 가정들은 데이터에 완벽하게 부합하지 않을 수 있으며, 가정이 위배될 경우 모델의 성능과 신뢰도가 저하될 수 있다.

B. 다중 회귀 분석

하나의 종속변수와 여러 독립변수 사이의 관계를 분석

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

i. 다중회귀분석 가정(1~4는 단순회귀분석가정과 동일)

1. 선형성
2. 오차항의 평균은 0이다.
3. 등분산성
4. 독립성 : 오차항은 서로 독립이고, 오차항은 각 독립변수와도 독립적이다.
5. 독립변수간에는 정확한 선형관계가 없다$\rho(X1i, X2i) \neq \pm 1$
   A. 한 독립변수가 다른 독립변수와 1 차함수관계에 있어서는 안된다
6. 관측된 값들의 수는 독립변수의 수보다 최소한 2 이상 커야 한다.
   A. 자유도가 1 이상 되기위해서 이다.

ii. 다중회귀분석 과정

1. 데이터 수집

   분석에 필요한 독립 변수와 종속 변수에 대한 데이터를 수집한다. 데이터는 관측된 값으로 이루어져야 하며, 독립 변수와 종속 변수 간의 연속적인 관계를 가지고 있어야 한다.

2. 변수 선택

   분석에 사용할 독립 변수들을 선택한다. 이 단계에서는 독립 변수들의 중요도를 평가하고, 다중공선성을 고려하여 최적의 변수 조합을 결정한다. 독립 변수 간의 상관관계를 평가하고 필요에 따라 변수 변환 또는 변수 선택 기법을 사용하여 변수 집합을 결정한다.

3. 회귀 모델 설정

   선택된 독립 변수들로 회귀 모델을 설정한다. 일반적으로 다중 회귀 모델은 독립 변수들의 선형 조합으로 종속 변수를 예측하는 방정식으로 표현된다. 예를 들어, 종속 변수 Y를 독립 변수 X1, X2, ..., Xn의 선형 조합으로 모델링할 수 있다.

4. 회귀 모델 적합

   설정된 회귀 모델을 데이터에 적합시킨다. 최소 제곱법 등의 방법을 사용하여 회귀 계수를 추정한다. 회귀 모델은 독립 변수들의 가중치와 상수항으로 구성된다. 회귀 계수 추정을 통해 각 독립 변수의 영향력과 방향성을 확인할 수 있다.

5. 모델 평가

   적합된 회귀 모델의 성능을 평가한다. 이를 위해 잔차(residual) 분석을 수행하여 모델이 데이터를 얼마나 잘 설명하는지 확인한다. 잔차 분석을 통해 모델의 적합성, 등분산성, 정규성 등을 평가할 수 있다. 또한, 결정계수(coefficient of determination) 등의 지표를 사용하여 모델의 설명력을 평가할 수 있다.

6. 모델 해석

   적합된 회귀 모델을 해석한다. 회귀 계수의 통계적 유의성과 방향성을 평가하여 각 독립 변수의 영향력을 해석한다.

iii. 다중회귀분석 장점

1. 추가적인 독립변수를 도입함으로써 오차항의 값을 줄일 수 있다.
2. 단순회귀분석의 단점을 극복 : 종속변수를 설명하는 독립변수가 두개일 때 단순회귀모형을 적용하면, 모형설정이 부정확할 뿐 아니라 종속변수에 대한 중요한 독립변수를 누락함으로써 계수 추정량에 대해 bias가 생길 수 있다. -> 다중회귀분석을 통해 bias를 제거할 수 있다.

iv. 다중회귀분석 문제점 -> **다중공산성**

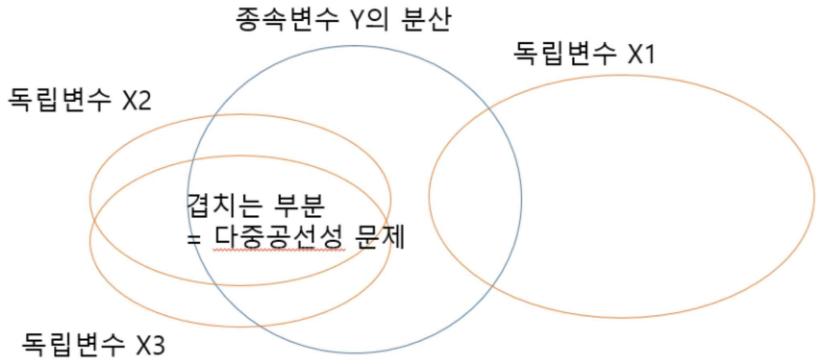
독립변수들 간에 강한 상관관계가 있는 경우 발생한다.

(독립변수의 수가 종속변수의 수에 비해 과도하게 많은 경우 상관관계가 높아질 가능성이 커진다.)

-> 변수의 상관관계가 높으면 어떤 변수에서 어떤 영향이 오는지 정확하게 판단할 수 없다. (변수의 중요성 판단의 어려움)

-> 회귀 계수가 불안정해지고 더 이상 해석할 수 없게 될수도 있다. (예측력의 감소)

-> 독립변수의 통계적 유의성을 왜곡



1. 다중공산성 진단법

   A. 상관계수 확인

      상관계수가 절댓값 0.7 이상인 변수들은 다중공산성의 가능성이 있고, 0.9를 넘는다면 다중공산성의 문제가 있다고 할 수 있다.

   B. 허용/공차(tolerance) 확인

      tolerance = 1-R^2 이다. (여기서 R = R-squared값)

      만약 R^2가 1이면 독립변수 간에 심각한 상관관계가 있다는 것을 의미하며 동시에 다중공산성이 심각하다는 것을 의미한다.

   C. 분산팽창지수(VIF : Variance Inflation Factor) 확인

      VIF = 1/tolerance = 1 / (1-R^2)이다. VIF가 클수록 다중공산성이 크다는 의미이고, 이 값이 3 이상이라면 다중공산성의 가능성이 있고, 10보다 크면 다중공산성의 문제가 있다고 할 수 있다.

   D. 회귀계수의 t-검정 및 p-value

      회귀계수의 t-검정에서 유의수준(일반적으로 0.05)보다 큰 p-value값을 가진 변수가 있다면, 해당 변수의 다중공산성 가능성이 낮다고 판단할 수 있다.

2. 다중공산성 해결

   A. 변수 제거

      가장 일반적인 방법이다.

   B. 변수 선택

      상관계수나 VIF를 기준으로 변수를 선택하거나 변수선택 알고리즘인 Stepwise, Lasso, Ridge 등을 활용할 수 있다.

C. 변수 변환(스케일링)

정규화, 표준화, 중심화(각각 데이터에서 평균을 뺀 값으로 변환), 로그화(자연로그나 일반로그를 씌운 값으로 변환)

D. 데이터 추가 수집

데이터의 추가 수집을 통해 상관관계가 낮은 변수를 추가할 수 있다.

E. PCA

2. 규제 회귀

다중회귀분석에서 다중공산성 문제를 해결하고 모델의 과적합을 줄이기 위해 사용되는 방법

선형회귀모델의 경우, 특성에 곱해지는 계수(또는 기울기)의 크기를 작게 만드는 것

비용 함수 목표 = $Min(RSS(W) + alpha * ||W||_2^2)$

여기서 RSS의 값을 최소화하면서 과적합되지 않는 alpha의 지점을 찾아야만 한다.
alpha를 0에서부터 지속적으로 증가시키면서 회귀계수(w)의 크기를 감소시킬 수 있다.
이렇게 alpha를 패널티로 부여해서 회귀계수의 크기를 줄이는 것을 규제라고 한다.

A. 라쏘(Lasso) 회귀

L1규제를 적용한 회귀 모형이다. 이는 W의 절댓값에 대해서 alpha로 패널티를 부여한 방식이다. 불필요한 변수의 계수를 0으로 만들어 변수선택 효과를 얻을 수 있다.

변수 선택과 변수 제거를 자동으로 수행하며, 특정 변수들만이 모델에 영향을 준다. 따라서 변수의 중요도를 알 수 있다.

이는 변수선택이 필요한 경우나, 특정 변수들의 중요도 파악이 중요한 경우에 쓴다.

B. 릿지(Ridge) 회귀

L2규제를 적용한 회귀모델이다. 이는 회귀계수의 제곱을 규제하는데, 변수들의 계수를 0에 가깝게 축소하지만 완전히 0으로 만들지는 않는다.

다중공산성을 줄이는 효과가 있으며, 변수들 사이의 상관관계를 고려하여 계수를 조절한다.

이는 변수들 간의 중요도가 비슷한 경우나, 모든 변수를 모델에 포함시키는 것이 필요한 경우에 쓴다.

C. 엘라스틱넷(ElasticnNet) 회귀

L1규제와 L2규제를 결합한 회귀이다.

$$RSS(W) + alpha2 * ||W||_2^2 + alpha1 * ||W||_1$$

위의 함수를 최소화하는 W를 찾는 것이다.

라쏘회귀의 alpha값에 따라 회귀계수의 값이 급격히 변동할 수 있는 점을 완화하기 위해, L2규제를 라쏘 회귀에 추가한 것이다.

그러나 L1규제와 L2규제가 결합되었으므로 시간이 상대적으로 오래걸린다.

위의 세가지 규제 방법 중 어떤 것이 가장 좋은지는 상황에 따라 다르다. 각각의 알고리즘에서 하이퍼파라미터를 변경해 가면서 최적의 예측 성능을 찾아내야 한다.

3. 회귀성능 평가지표

A. 결정계수

결정계수는 R-제곱(R2)라고도 하는데, 독립변수에서 종속변수의 분산 비율을 나타내는 통계측정값이다. 0과 1사이의 값을 가지며, 1에 가까울수록 모델이 종속변수의 변동성을 잘 설명하고 있는 것으로 해석된다.

$$R^2 = 1 - (SSres/SStot)$$

- SSres = 잔차의 제곱합(y의 예측 값과 실제 값 간의 차이)
- SStot = 총 제곱합(y의 각 값과 y의 평균 사이의 차이)
- R^2 = 0에서 1까지의 값

B. 평균제곱오차 MSE

오차의 크기를 나타내는 지표이다.

오차의 제곱을 평균으로 나눈 것이다. MSE가 0에 가까울수록 추측한 값이 원본에 가까운 것이기 때문에 정확도가 높다고 할 수 있다.

예측값과 실체값 차이의 면적의 평균이라고 할 수 있다.

C. 평균절대오차 MAE

예측값과 셀제값의 절댓값 오차를 평균하여 반환한 값이다. MSE와 마찬가지로 0에 가까울수록 좋은 모델이다.

인간이 보기에 직관적으로 차이를 알 수 있고 MSE, RMSE에 비해 오차값이 이상치의 영향을 상대적으로 적게 받는다. (오차를 제곱하지 않기 때문에)

D. 평균 제곱근 오차 RMSE

MSe값에 루트를 씌운 값이다.

지표 자체가 직관적이며 예측변수와 단위가 같다.

E. RMSLE(Root Mean Squared Log Error)

로그의 값들을 통해 차이들을 확인

로그 정규화 등을 통한 데이터의 분포를 조정하는 전처리를 거친 뒤에, 해당 산출물에 대한 결과물도 정규화한 값들을 기준으로 모델의 성능을 검증하는데 활용

4. **본 데이터에는 주택 가격에 영향을 미칠 것이라 판단되는 독립 변수들이 있습니다. 주택 가격에 대하여 다중 선형 회귀분석을 적합시켜 어떠한 변수들이 주택 가격에 어떻게 영향을 미치는지알아보시오. (설명 모형)**

https://colab.research.google.com/drive/1CAUgVw8oIIxAaOr9x1ap_WckAIzJ9qUU?usp=sharing

디스커션 : 이 과정을 진행하면서 상관관계랑 같다는 생각이 들었습니다. 이 설명모형과정을 상관계수로 대체가 가능한지 궁금합니다.

5. 데이터를 분할하여 주택 가격을 예측하는 회귀모델을 구축하고 회귀성능 평가지표를 통해 예측성능을 평가하시오. (예측 모형)