241216 # The multicollinearity (다중공선성) Ref) https://brunch.co.kr/@gimmesilver/76

Definition : 독립변수 간에 (3개 이상) 강한 상관관계가 나타나서 회귀분석의 전제가정인 독립변수 간의 상관관계가 높으면 안된다 같은 의미이다 정

: Multicollinearity denotes when Independent variables in a linear regression equation are correlated

otherwise, collinearity denotes when two independent variables in a regression analysis are themselves correlated.

Ex $X_3 = X_2 + 2X_1$: X_3는 X_1과 X_2 간에 대해선 상관관계가 나타난 다중공선성 발생

다중공선성은 생각하지 마라

by gimmesilver · Jun 25, 2022

선형 회귀 관련 교재나 설명 자료를 보면 꼭 빠지지 않고 나오는 주제 중 하나가 다중공선성입니다. 다중공선성이란 회귀 모델의 독립변수들이 서로 강한 상관 관계를 갖고 있는 상태를 말합니다. 보통 이런 자료를 보면 선형 회귀 모델은 독립변수들이 서로 독립이어야 한다는 가정이 있는데 이 가정을 위배하기 때문에 문제가 된다고 설명하죠. 그런데 다중공선성이 있는 경우 회귀 분석에서 어떤 문제가 발생할 수 있는지 혹은 다중공선성이 있을 때 그럼 어떻게 해야하는지에 대한 설명은 대체로 미흡한 편입니다. 그렇다 보니 이에 대해 오해하거나 엉뚱한 조치를 취하는 경우를 종종 보곤 합니다.

이름태면, 간혹 데이터 분석 경진 대회 심사에 참여할 때면 아래와 같이 분석 내용을 발표하는 참가팀이 있습니다.

> 우리는 Y를 예측하기 위해 X1 ~ X100 의 피처를 추출했다. 예측 모델로는 랜덤 포레스트를 이용했는데 피처가 너무 많다고 생각해서 각 변수들 간의 상관 관계를 측정해 봤더니 상관 계수가 0.9 이상인 변수 쌍이 다수 발견되었다. 그래서 다중공선성 문제가 우려된다고 판단하여 이런 변수쌍 중에 하나만 남기는 방식으로 피처의 수를 줄였다.

위 작업에는 최소한 3가지 오류가 있습니다.

첫째, 모델링의 목적이 예측인 경우엔 다중공선성을 신경 쓸 필요가 없습니다. 다중공선성이 회귀 모델에 미치는 효과는 크게 두 가지 입니다.

- 회귀 모델에 어떤 변수가 포함되는지 여부에 따라 특정 변수의 회귀 계수값이 크게 변동할 가능성이 높습니다.
- 회귀 계수의 표준 오차가 커져서 통계적 유의성에 영향을 줍니다.

즉, 다중공선성은 독립변수의 회귀 계수 추정에 영향을 줄 뿐, 종속변수 예측에는 아무런 영향을 주지 않습니다.

둘째, 모델링 방식이 선형 회귀가 아니라면 다중공선성을 고려할 필요가 없습니다. 다중공선성은 선형 회귀 모델의 기본 가정을 위배하기 때문에 생기는 문제입니다. 그러니 랜덤 포레스트 같은 다른 기법을 사용한다면 다중공선성을 따질 필요는 없는 것이죠. 좀 더 엄밀히 말하면, 트리 모델의 경우 각 노드에 사용될 변수를 선택하는 기준에 영향을 줄 수는 있습니다. 하지만 이것 역시 모델의 해석(변수 중요도)에 영향을 줄 뿐 예측 결과 자체에는 영향을 주지 않습니다. 심지어 랜덤 포레스트는 트리 모델이지만 개별 트리는 피처의 일부만 사용하는 특징이 있기 때문에 이런 문제에 영향을 덜 받습니다.

셋째, 다중공선성은 '두' 변수 간의 상관계수만 측정해서는 정확히 파악할 수 없습니다. 다시 앞해, 다중공선성을 확인하겠다면서 pairwise scatter plot 만 보거나 상관 계수 매트릭스만 측정해서는 안됩니다. 만약 어떤 독립변수가 여러 독립변수들의 선형 결합 관계인 경우라면 변수 쌍의 선형계수는 낮더라도 다중공선성이 생길 수 있습니다. 예를 들어 만약 x3 ~ x1 - x2 인 경우라면, x3는 x1와 x2 각각에 대해선 선형 상관 계수가 낮을 수 있지만 (두 변수에 의해 설명되는 변수이기 때문에) 다중공선성이 발생합니다. 다중공선성을 확인하려면 상관계수가 아니라 '분산팽창요인(Variance Inflation Factor, VIF)'을 측정해야 합니다.

그렇다면 만약 모델링의 목적이 예측이 아니라 회귀 계수 추정인 경우엔 어떨까요? **설명 회귀 계수 추정을 위한 분석이라 하더라도 다중공선성은 반드시 없애야 하는 문제가 아닙니다.** 따라서 이런 경우에도 다중공선성을 이유로 무분별하게 변수를 제거해서는 안됩니다.

대개 우리는 특정 한 두 개의 요인이 종속변수에 미치는 영향력을 추정하기 위해 회귀 분석을 합니다 (모델에 들어간 모든 변수의 회귀 계수를 정확히 측정하려는 시도는 비현실적인 목표이죠). 회귀 모델에 들어가는 여러 변수 중에 실제 관심 있는 변수를 제외한 나머지 변수들은 단지 관심 변수의 회귀 계수를 정확히 추정하기 위해 보조하는 통제 변수입니다. 따라서 이런 통제 변수들의 회귀 계수는 정확히 추정할 필요가 없습니다. 때문에 만약 다중공선성이 이런 통제 변수들 사이에서만 발생한다면 신경 쓸 필요가 없습니다. 참고로 통제 변수들 간에 다중공선성이 발생하는 것인지를 알려면 관심 변수는 제외하고 나머지 변수들만 이용해서 VIF를 측정해 보면 됩니다.

더 나아가 통제 변수와 관심 변수 사이에 상관성이 높다 하더라도 애당 통제 변수를 제거해야 할지는 VIF값만 보고 판단할 것이 아니라 도메인 지식을 이용해 변수 간의 관계를 확인해야 합니다. 왜냐하면 해당 통제 변수가 관심변수와 종속변수 양쪽에 인과 관계가 있는 교란 변수일 가능성도 있기 때문입니다. 이 변수가 교란 변수라면 설령 독립변수 간에 상관성이 높다 하더라도 잠부로 제거해서는 안되겠죠.

예를 들어 음주량과 흡연량이 암에 미치는 영향을 정량적으로 추정하기 위해 회귀 분석을 한다고 가정해보죠. 마침 내가 관측한 데이터들에서 음주량과 흡연량 간에 상관성이 매우 높다면 이 두 변수 중 하나를 제거하는 것이 옳은 선택일까요? 아마 그렇지 않을 겁니다. 왜냐하면 실제 음주량과 흡연량은 둘 다 암에 인과적 영향을 줄 가능성이 높기 때문이죠. 즉, 이 둘은 서로간에 교란 요인이 되며 심지어 상호 작용 효과가 있을 가능성도 높습니다. 따라서 이런 경우엔 계수의 신뢰도가 낮아지더라도 둘 중 하나를 제거해서는 안됩니다.

제가 생각하기에 다중공선성을 확인하는 목적과 처리 방법은 크게 아래와 같이 세 가지로 정리할 수 있겠습니다.

첫째, 내가 미처 고려하지 못한 데이터 간의 숨은 관계가 있는 것은 아닌지 확인하는 용도가 될 수 있습니다. 가령 특정 변수의 VIF가 지나치게 높다면 이게 혹시 내가 미처 생각하지 못한 collidor 나 mediator 일 가능성은 없는지 도메인 전문가나 동료 분석가와 같이 논의해 볼 수 있겠죠 (이에 해당한다고 판단되면 그 때 변수를 제거하면 됩니다). 혹은 상관 관계가 높은 변수들을 조합한 파생 변수를 만든 후 대체하는 방법도 생각할 수 있습니다. 이를테면, 게임 캐릭터의 여러 가지 스탯 정보를 이용한 회귀 모델을 만든다면, 이런 스탯들은 보통 서로 상관관계가 높을 겁니다. 그러면 관련성 높은 스탯 정보들을 합해서 '공격 관련 스탯' 이나 '방어 관련 스탯' 이라는 파생변수를 만든 후 기존변수를 대체할 수도 있을 겁니다.

둘째, 상관 관계를 완화할 수 있는 추가 데이터 확보가 필요한지 여부를 판단하는 용도가 될 수 있습니다. 가령, 위에 음주량과 흡연량이 암에 미치는 영향을 분석하는 사례를 생각해 보면, 지금 관측데이터에는 음주와 흡연을 같이 하거나 아예 둘 다 안 하는 사람들 데이터만 주로 있어서 생긴 문제일 테니, 술을 좋아하는 비흡연자와 술을 안 좋아하는 골초들의 데이터를 추가 확보하면 둘 간의 상관성을 낮출 수 있을 겁니다. 그리고 사실상 이런 데이터를 확보하는 것이 보다 정확한 분석을 위해선 더 적절하겠죠. 보통 학교에서는 고정된 데이터에 대해서 모델을 잘적합시키기 위해 이런 저런 조작을 가하는 방법만 해법으로 제시하는 경우가 많은데 실전에서는 데이터를 추가 확보하는 방법도 충분히 고려할 수 있는 선택지입니다.

셋째, 그날 현재 회귀 모델의 한계를 인정하는 방법도 있습니다. 사실상 대개의 경우 회귀 모델은 불완전합니다. 그러니 다중공선성이 있더라도 그냥 모델을 생성한 후 계수의 표준오차가 크다는 점을 과감 없이 보여주는 것이 인위적인 처리를 통해 통계적 유의성을 확보하려는 것보다 더 옳은 선택일 수 있습니다. 학교에서 이렇게 하면 눈문을 못 써서 졸업이 힘들 수 있지만 회사에서는 오히려 한계점을 명확히 제시하고 결과를 보여주는 것이 더 좋은 평가를 받을 수 있습니다 (아닐 수도 있구요).

정리하자면, 다중공선성은 많은 분석가들이 오해하고 있는 개념입니다. 선형 회귀 분석을 하는 경우를 제외하면 다중공선성 자체를 고려할 필요가 없으며, 설령 선형 회귀 분석을 하는 경우라 하더라도 대개의 경우 다중공선성이 있는 변수를 무분별하게 제거하는 것은 결코 좋은 해결책이 아닙니다.

통계 회귀분석 데이터분석

gimmesilver
.
구독자 1,294
+ 구독