선형모델 가정에 대한 이해

왜 매수결정을 했는지에 대한 이유를 구체적으로 설명하기 유리한 모델은 Linear Model 입니다. 그 중 다변량 회귀모델 (Multivariate Linear Regression) 은 데이터분석을 배울 때, 가장 기초적으로 다루는 예측모델입니다. 예측하고자 하는 종속변수 Y (레이블 혹은 타겟 변수) 가 연속형이고, 이것을 설명 혹은 예측하는 독립변수 X (입력피쳐 혹은 입력변수) 들의 선형조합 Z 로 Fitting 을 하는 것인데, 충분한 이해없이 사용하면, 잘못된 결론을 내기 쉽습니다. 다변량 회귀분석 모델이 의미가 있을려면, 데이터가 상당히 강한 Assumptions 를 만족해야 합니다. 중요한 4 가지는 다음과 같습니다.

1. Normality - 에러(실제값 - 예측값)가 정규분포를 따라야한다. 사실 이건 Y 가 정규분포를 따라야 한다는 것과 크게 다르지 않습니다.
2. Weak Heteroscedasticity - 에러가 등분산성을 만족해야 한다. 즉 에러의 분산이 예측 값의 크기에 따라서 크게 변화하지 않아야 한다.
3. Linearity - 선형성. 이것은 추정된 베타값이 X 값의 크기에 따라서 변화하지 않아야 한다. 예를 들어, 소득을 추정하는데, 카드 사용량이 변수라면 카드 월 사용량이 백만원일 때 추정된 계수(coefficient) 가 50 이라면, 카드 사용량이 천 만원일때도 베타 계수가 50이여야 한다는 말입니다.
4. Weak Multicollinearity - (다중 공선성)이 크지 않아야 한다. 쉽게 이야기 하면 어떤 여러개의 X 가 Y 를 설명하는데 있어서 X 들이 같은 방향으로 움직이면 안 된다고 이해하면 될 것 같습니다. 다중 공선성이 큰 경우는 계수 값이 정확하지 않아서, 계수에 대한 해석이 불가능하게 됩니다. 아주 심한 경우는 다른 변수의 영향으로 양의 계수가 음의 계수로 바뀌게 됩니다.

위 가정 1 번과 2 번을 만족하지 않아도 Regression 을 할 수 있게 일반화 한 것이 일반화 회귀모형(Generalized Linear Model) 입니다. GLM 에서는 Y 가 갯수(count), 비율(proportion), 이진(0과 1) 등 같이 연속형 변수가 아니고 정규 분포를 따르지 않아도 선형모델을 구현할 수 있습니다. 물론 등분산성을 만족하지 않아도 됩니다. 대신에 Y 에 대한 명확한 분포 설정과 Y 에 대한 Link Function 필요합니다. 가장 많이 쓰이는 것이 Log Link 입니다. 이 부분을 쉽게 이해하기 위해서는 이렇게 생각하면 됩니다. X 의 선형조합 Z 는 음수의 값도 갖게 되는데, 비율이나, 갯수는 항상 양수입니다. 따라서 Y 에 Log 를 씌워서 음수를 갖게 할 수 있습니다. 반대로 EXP(a0 + a1*x1* + *a2*x2 ...) 로 항상 양수인 Y 를 Fitting 한다고 보시면 될 것 같습니다. 많이 다루는 로지스틱 회귀 모델은 Log(odds) 를 X 의 선형조합으로 Fitting 을 하는 일반화 선형모형의 한 예로 볼 수 있습니다. 데이터상으로는 Y 가 이항분포(Bernoulli 분포 혹은 0 과 1) 이므로 Link Function 가 Logit Link 즉, log(p/1-p) 로 하는 일반화 선형모형과 동일한 의미가 됩니다. Y 가 0 과 1 이므로 이것을 가장 잘 근사하게 따라갈 수 있는 변형은 Logit Link 인 것입니다. logit Link 를 풀면 Y = exp(z) / 1 + exp(z) 가 됩니다. 즉 Y 를 설명하기에 좋은 형태로 변경이 되는 것입니다. Y 가 개 수(count) 인 경우는 포아송 회귀분석 (Poisson regression) 입니다. 주어진 시간 혹은 범위에서 뽑은 count 샘플은 포아송 분포를 따라간다는 것이 알려져 있습니다. 예를 들면 인구 만명당 암 발생 환자 수 등이 예가 될 것 같습니다. 포아송 분포의 평균과 분산은 같습니다. 즉 평균이 증가하면 분산이 증가하는 분포입니다. 따라서 등분산성을 만족하지 않아도 Y 를 fitting 할 수 있습니다. 이 경우, Link 는 log 입니다. 즉, X 의 선형조합인 Z 에 Exponential 를 씌워서 양수가 되도록 합니다. Y 가 비율(Proportion) 인 경우도 있습니다. 그럼 비율은 어떤 분포일지 궁금합니다. 비율은 항상 0 과 1 사이 양수이므로 Link 함수는 log link 를 쓰면 될 것 같습니다. 일반적으로 비율은 분자의 특성에 따라 분포가 바뀔 수 있습니다. 위에 예시한 인구 만명 당 암환자의 비율은 GLM 으로 Fitting (Y ~ Normal 분포, Log link) 할 수 있습니다. 하지만 더 Fitting 을 잘 하려면 분자를 Y 로 하고 분모인 인구 수를 exposure 요인으로 처리하는 것입니다. 이 경우 당연히 Y 는 포아송분포가 됩니다. log(암 환수/인구 수) = Z(X 선형조합) 형태의 모델을 (암 환자수) = exp(Z)*(인구수) 이렇게 변경하는 것과 동일합니다. 그럼 여기서 인구수가 exposure 가 되고, 인구수를 고려하여 Z 에 계수값을 추정하게 됩니다. Proportion 을 Y 로 fitting 하는 것보다 훨씬 좋은 결과가 나옵니다.

마지막으로 3 번째 가정이 선형성을 만족하지 않아도 쓸 수 있는 Linear Model 이 있습니다. Generalized Additive Model (GAM) 인데, 이경우는 spline 함수를 이용하여 각 X 를 곡선으로 만들어 Y 와 fitting 합니다. 예를 들어 Y 가 그랜저를 살 확률이고, X 가 소득이라고 할 때, 소득이 증가

함에 따라 그랜저를 살 확률은 증가하다가 어느 순간 다시 감소할 것 입니다. 그럼 2 차원 곡선이 되는데요. 이런 경우도 소득을 spline 함수(곡선형태)로 만들면 Y 를 잘 Fitting 할 수 있습니다.

마지막 4 번째 가정은 선형모형의 구조상 피할 수 가 없습니다. 공선성을 일으키는 입력 변수를 빼거나, 주성분등으로 공선성을 완전히 제거해야 합니다. 기본적으로 Linear 모델이라는 것은 X 의 합으로 연결이 되어 있습니다. 따라서 Fitting 된 모델에서 X1 이 1 증가할 때, X2 도 1 증가하는 구조라면, X1 와 X2 의 계수의 추정은 해석하기 어렵게 됩니다. 하지만, 이런 구조이기 때문에 잘 fitting 된 선형모델에서는 X 변화에 따른 Y 의 변화를 이해할 수 있는 장점으로 작용합니다. 요즘 관심을 받고 있는 해석가능한 모델이 되는 것입니다.

By KHS