계량분석

Quadratic Function and Fractional Polynomials

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 29, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

선형성 가정

선형성 가정

우리가 지금까지 자연스럽게 언급해온 선형모형(linear model)은 사실 가정에 가깝다.

- 선형모형에서 "선형"은 이른바 선형성(linearity)이라는 가정을 의미한다.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- 우리는 비선형모형(nonlinear model)을 지금까지 다루지 않았다. 다음은 비선형모형의 예제들이다:

$$y = \beta_0 \cdot \beta_1 X_1 \cdot \beta_2 X_2 \cdot \ldots \cdot \beta_k X_k \cdot \epsilon$$
$$y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + + \beta_k X_k + \epsilon}$$

- 여기서 e는 자연상수(natural constant)를 의미한다. 오일러 상수(Euler constant)라고도 한다. 때때로 e^x 는 exp(x) 라고도 쓴다.

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} \approx 2.718\ldots$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

선형성 가정

선형성은 사실 수학적인 의미를 내포하고 있다.

- $y = \beta_0 + \beta_1 X_1 + \epsilon$에서 X_1과 y의 관계는 β_1의 기울기를 갖는 직선으로 나타낼 수 있다. 그러므로 선형성을 갖는다. **desmos**에서 실험해보자.
- $y = \beta_0 + X_1^{\beta_1} + \epsilon$에서는 더이상 X_1과 y의 관계를 직선으로 나타낼 수 없다. 즉 선형성을 갖지 않는다. 이에 관해서도 **desmos**에서 실험해보자.
- 데이터의 두 변수가 비선형적인 관계를 가지고 있을 때, 우리는 그 "관계를 흉내낼 수 있는 모형"을 구축할 수 있다.
- 생각을 거꾸로 하여 우리는 데이터를 생성하는 과정(data-generating process)으로써 모형의 역할/기능을 이해해 볼 수 있다. 그 맥락에서 선형성 가정은 "모형이 반드시 X_1과 y의 관계를 선형적으로 생성해야 한다"는 것을 뜻한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

선형성 가정

이렇게 생각하면 선형성 가정은 굉장히 답답하게 느껴질 수 있다.

- 현실에서 두 변수의 관계가 얼마든지 비선형적(nonlinear)일 수도 있기 때문이다. 무슨 예들이 있을까 먼저 고민해보자! 예제에 대해 충분히 고민을 해야 한다.
- 현실의 복잡다단하고 비선형적인 관계가 모형에서는 선형적으로 묘사되므로 굉장히 심한 제약을 가하고 있는 것이 아닐까?
- 하지만 간단한 대수적 조작을 통해 비선형적인 관계도 선형적인 관계로 바꾸어 묘사할 수 있는 수학적 트릭이 있다. 이것을 배우는 것이 오늘 수업의 목적이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 이차함수

이차항과 이차함수

상호작용항과 이차항은 사실 비슷한 원리를 공유한다.

- 선형모형에서 회귀계수는 철저하게 하나의 독립변수와 하나의 종속변수 사이의 관계만을 묘사한다. 다음과 같은 선형모형에서 "X_1와 y의 관계"는 ("X_2와 y의 관계"와 무관하게) β_1으로만 표현된다.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- 만일 X_2가 변화하면 이는 즉각 β_2에 반영되며 β_1에는 영향을 미치지 않는다.
- 하지만 현실에서는 X_2와 X_2가 상호작용하며 y에 영향을 미치는 경우가 많다. 사회학이론은 많은 부분에서 이런 가능성을 시사한다.
- 우리는 저번 주에 상호작용 효과(interaction effects)를 통해 이 문제에 대응할 수 있다는 것을 배웠다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 이차함수

- 우리는 "X_1과 X_2의 y에 대한 상호작용 효과"를 살펴보듯 "X_1과 X_1의 y에 대한 상호작용 효과"를 살펴볼 수 있다.
- 아래의 선형모형을 통해 "X_1과 X_2의 y에 대한 상호작용 효과"를 살펴볼 수 있다.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- 아래의 선형모형을 통해 "X_2과 X_2의 y에 대한 상호작용 효과"를 살펴볼 수 있다.

$$\begin{aligned} y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 X_2 + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \epsilon \end{aligned}$$

- 위의 X_2^2를 이차항(squared term) 또는 제곱항이라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 이차함수

이차항은 곡선형(curvilinear)의 관계를 묘사한다.

- $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2$ 와 같이 회귀모형을 설정하였다고 가정하고, β_3가 (+)인 경우와 (-)인 경우 그래프가 각각 어떻게 변화하는지 **desmos**에서 살펴보자.
- 이차항이 (+)인 이차함수(quadratic function)를 그래프로 그리면 X가 증가함에 따라 y가 감소하다가, 어느 순간 다시 y가 증가한다.
- 이차항이 (-)인 이차함수를 그래프로 그리면 X가 증가함에 따라 y가 증가하다가, 어느 순간 다시 y가 감소한다.
- 마찬가지로 이차항을 모형 안에 넣으면 모형은 이 같은 곡선(=일종의 비선형관계)으로 y와 X_2의 관계를 묘사할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 이차함수

eCampus에서 nations.dta를 다운받고 Stata에서 열어 이차항을 만들어보자(**[Stata 코드]** 참고).

- 데이터를 살펴보아 각 변수가 무엇을 의미하는지 확인하자. listwise deletion하자.
- death를 종속변수로 하고 food, life를 독립변수로 하여 회귀모형을 구축해보자. death와 life 사이의 선형관계를 해석해보자.
- death와 life 사이의 관계를 보여주는 산포도와 선형적합선(linear fitting line)을 그려보자.
- death와 life 사이의 관계를 보여주는 산포도와 이차적합선(quadratic fitting line)을 그려보자.
- life의 이차항(=제곱항)을 만들자.
- death를 종속변수로 하고 food, life, life 이차항을 독립변수로 하는 회귀모형을 여러가지 방법으로 만들어보자. 이항 연산자(binary operator)를 사용해서도 넣어보자.
- life 이차항은 통계적으로 유의한가? 곧바로 해석할 수 있겠는가?

이차항과 이차함수

이차항이 통계적으로 유의하다면 모집단에서도 정말로 비선형 관계를 지지한다고 볼 수 있다([Stata **코드**] 참고).

- 이차항을 넣었다면 이차항이 통계적으로 유의한가 여부가 중요할 뿐, 일차항 부분의 유의성 여부에는 주목할 필요가 없다.
- 만일 이차항이 통계적으로 유의하지 않았다면 일차항이라도 통계적으로 유의한지 살펴볼 필요가 있다. 이 경우 이차항을 빼고 회귀모형을 다시 확인해야 한다.
- 이차항이 통계적으로 유의하지 않았는데, 일차항은 통계적으로 유의했다고 해서 곧장 일차항의 의미만을 해석해서는 안된다(Why?)
- 그러므로 이차항을 살펴볼 때는 (논문/보고서에서 사용할 것인가와는 별개로) 위계적으로 모형을 구축해보고 살펴볼 필요가 있다. death를 종속변수로 하고 life, life 이차항을 단계적으로 독립변수로 투입하는 위계적 회귀모형을 차례로 구축해보자. 단 (1) 모든 모델에 통제변수로 food를 투입하고, (2) 요약통계량에서 사례수, R^2, Adj. R^2는 반드시 보고하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항을 해석할 때는 좀 더 복잡한 접근법이 요구된다(**[Stata 코드]** 참고).

- 앞서 death와 life의 관계를 살펴보았을때 일정 정도까지 life가 증가하면 death는 감소하지만, "특정 경계"를 넘어서면 오히려 death는 증가함을 확인하였다.
- 수학적으로 $\delta death/\delta life = 0$ 이 되는 life 지점이 바로 "특정 경계"이다(Why?)
- 이때 데이터에서 의해 관찰되는 life의 범위(range)를 살펴보아야 한다! "특정 경계" 지점이 결코 오지 않을 수도 있기 때문이다.
- **margins**와 **marginsplot**를 사용하여 그 관계를 그래프로 나타내보자. 그래프를 png 파일로 저장하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 일차항 사이에는 높은 상관관계가 있는 것이 보통이다(**[Stata 코드]** 참고).

- life와 life 이차항 사이의 상관계수를 살펴보자.
- 나중에 자세히 살펴보겠지만 높은 다중공선성(multicollinearity) 문제를 일으키는 것이 아닐까 의심스러울 수 있다.
- 다음과 같이 평균중심화(mean centering)를 통해 상관계수를 인위적으로 낮출 수도 있다. 여기서 중심화(centering)란 결국 편차(deviation)를 의미한다.

$$y = \beta_0 + \beta_1 X_1 + \beta_2(X_2 - \bar{X}_2) + \beta_1 1(X_2 - \bar{X}_2)^2 + \epsilon$$

- 평균중심화된 이차항을 사용하더라도 이차항의 회귀계수, 표준오차, t 값, 유의확률 (p-value), R^2, Adj. R^2는 결국 똑같다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다항식

다항식

이차항을 넘어 좀 더 높은 차원의 항을 추가하여 더 많은 굴곡을 추가할 수 있다.

- 표현을 달리하자면 이차항을 일반화한 개념이 바로 다항식(polynomial equation) 내지 다항함수(polynomial function)이다.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \ldots + \beta_k X_k^k + \epsilon$$

- 아래와 같이 다항식을 만들면 곡선의 꼴이 어떻게 다른지 desmos에서 확인해보자.

$$y = 1 + x$$
$$y = 1 + x + x^2$$
$$y = 1 + x^2 + x^3$$

- 차수가 하나 증가할 때마다 굴곡이 하나씩 더 추가된다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

다항식

- 원하는 만큼의 다항을 추가하여 더 복잡한 곡선을 구성할 수 있다. 하지만 그렇게 복잡하게 만들수록 모형의 일반적인 설명력은 오히려 제약될 수 있다는 점에 유의해야 한다.
- 복잡한 모형은 (그것이 잘 맞는) 특수한 상황에서는 매우 뛰어난 설명력을 발휘할 수 있겠지만 일반적인 상황에서는 오히려 형편없는 설명력만을 가진다. 데이터 과학에서는 “지나치게 복잡한 모형이 과적합(overfitting) 문제를 갖는다”라고 표현한다.
- 단순한 모형은 특수한 상황에서야 보잘 것 없는 설명력만을 가질 수도 있겠지만 일반적인 상황에서 오히려 평범 이상의 설명력을 가질 수 있다.
- 결국 밸런스를 유지하는 것이 중요하다. 문제는 “어떻게” 이다.

다항식

"추가적인 변수를 사용함에도 불구하고 설명력을 그만큼 높이는가"를 통계적으로 검증해 볼 수 있다(**[Stata 코드]** 참고).

- 이 목적을 위해 사회통계학에서 가장 널리 쓰이는 방식은 다차항이 무의미한지 여부를 확인하는 왈드 검정(Wald test)이다. Stata에서는 **test** 명령어로 수행할 수 있다.
- 왈드 검정은 다음과 같은 가설 구조를 검정하기 위해 일원분산분석(one-way ANOVA)를 수행한다. 이때, m은 다항 차수를 의미한다.

$$H_0 : X^m = 0$$
$$H_a : X^m \neq 0$$

- 만약 영가설을 기각하는데 실패했다면 (설령 그 변수가 통계적으로 유의하더라도) 그 변수의 추가가 충분히 의미있게 설명력을 더한다고는 볼 수 없을 것이다.
- 당연히 해당 회귀계수의 t 값의 제곱과 왈드 검정의 F 값은 동일하다(Why?).
- 위 아이디어를 일반화한 James Ramsey의 Regression Equation Specification Error Test (RESET)도 있다. Stata에서는 **ovtest** 명령어로 수행할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다항식

- 또다른 방법은 R^2 의 증가분(increments)을 보는 것이다. 해당 다차항을 추가함으로서 R^2 가 크게 증가했다면 의미있게 설명력을 더한 것이지만 매우 작게 증가했다면 별 의미는 없다고 해석한다. 조정된 R^2 를 확인할 수도 있다.
- 몇몇 연구자들은 다시 R^2 의 증가분(increments)에 대한 유의성 검정을 시도하기도 하지만 대중적으로 보이지는 않는다.
- 일반적으로 이차항 내지 삼차항(cubic terms) 정도의 다항식을 만드는 것이 보통이다.
- 물론 이차항만을 넣을때도 위와 같은 방식으로 정당성을 테스트해 보는 것이 바람직하다.

다항식

이 회귀모형에서 이차항 내지 삼차항을 추정할 때 그보다 낮은 저차항을 빼놓고 회귀모형을 구축하지 않아야 한다.

- 예컨대 이차항을 포함한 모형을 추정할 때 일차항을 빼놓지 말 것!
- 사실 (예전) 수리사회학에서는 다음과 같이 일차항을 빼놓은 모형을 구축하기도 했다.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \beta_3 X_3^3$$

- 그러나 사회통계학에서는 이차항이나 삼차항은 결국 다항식의 일부이므로 반드시 저차항을 포함하여 모형을 만들어야 한다.

$$y = \beta_0 + \beta_1 X_1 + (\beta_2 X_2 + \beta_3 X_2^2) + (\beta_4 X_3 + \beta_5 X_3^2 + \beta_6 X_3^3) + \epsilon$$

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

다항식

다시 nations.dta를 Stata에서 열어 삼차항을 연습해보자(**[Stata 코드]** 참고).

- death와 life 사이의 관계를 보여주는 산포도와 고차적합선(high-order polynomial fitting line)을 그려보자. 미리 차수를 제시하기 어렵다면 매우 탄력적인 분수다항함수(fractional polynomial function)를 사용하여 그려볼 수 있다. 여기서는 삼차함수 적합선을 그려보자.
- death를 종속변수로 하고 food, life, life 이차항, life 삼차항을 독립변수로 하여 회귀모형을 구축해보자. 몇 가지 방법으로 삼차항을 선형모형 안에 넣을 수 있다. 이항 연산자를 사용하여 넣어보자.
- 3차항은 통계적으로 유의한지 확인하자. 해석하기 쉬울까?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

다항식

- life의 범위(range)를 다시 살펴보자. **margins**와 **marginsplot**를 사용하여 그 관계를 그래프로 나타내보자. 그래프를 png 파일로 저장하자.
- 위계적 회귀모형을 차례로 만들면서 변화하는 관계를 설명해보자. 단 (1) 모든 모델에 통제변수로 food를 투입하고, (2) 요약통계량에서 사례수, R^2, Adj. R^2는 반드시 보고하자.
- 삼차항을 넣는 것은 바람직한 것으로 판단되는가? **nestreg** 명령어를 연습해보자.