계량분석

Quadratic Function and Fractional Polynomials

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 21, 2024

진행 순서

선형성 가정

선형성 가정

지금까지 자연스럽게 언급해온 선형모형은 사실 가정에 가깝다.

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

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

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

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

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

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

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. 즉 선형성을 갖는다.
- $y = \beta_0 + X_1^{\beta_1} + \epsilon$에서는 더이상 X_1과 Y의 관계를 직선으로 나타낼 수 없다. 즉 선형성을 갖지 않는다.
- 이에 관해서 desmos에서 실험해보자.

선형성 가정

- 하지만 현실에서는 두 변수의 관계가 얼마든지 비선형적(nonlinear)일 수도 있다.
- 무슨 예들이 있을까 먼저 고민해보자! 예제에 대해 충분히 고민을 해야 한다.
- 현실의 복잡다단하고 비선형적인 관계가 모형에서는 선형적으로 묘사된다면 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.

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 관계는 (X_2와 Y의 관계와 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. 사회이론은 많은 부분에서 이런 가능성을 시사한다.
- 우리는 상호작용 효과(interaction effects)로 이 문제를 검정할 수 있다.

이차항과 이차함수

- 우리는 X_1과 X_2의 Y에 대한 상호작용 효과를 살펴보듯 X_1과 X_1의 Y에 대한 상호작용 효과를 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^2를 이차항(squared term)이라고 부른다.

이차항과 이차함수

이차항은 곡선형의 관계를 묘사한다.

- 다음의 회귀식을 설정하고 β_2가 (+)인 경우와 (-)인 경우 그래프가 각각 어떻게 변화하는지 desmos에서 살펴보자.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

- 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 일종의 비선형관계, 즉 곡선형(curvilinear)으로 Y와 X의 관계를 묘사할 수 있다.
- (상호작용항과 마찬가지로) 일차항 부분을 회귀식에서 생략하면 안된다.

이차항과 이차함수

Stata에서 이차항을 만들어보자.

- lifeexp.dta에서 각 변수가 무엇을 의미하는지 확인하자. 결측치(missing values)는 항목별 삭제(listwise deletion)하자.
- lexp를 종속변수로 하고 gnppc를 독립변수로 하여 회귀식을 추정하고 그 선형관계를 해석해보자.
- lexp와 gnppc 사이의 관계를 보여주는 산점도, 선형 적합선, 그리고 이차 적합선 (quadratic fitting line)을 그려보자.
- gnppc의 이차항을 만들자.
- 회귀식에 gnppc 이차항을 넣는 여러 방식을 연습해보자. 이때 반드시 이항 연산자 (binary operator) 사용법에 숙달되어야 한다.
- gnppc 이차항은 통계적으로 유의한가? 곧바로 해석할 수 있겠는가?

이차항과 이차함수

이차항을 해석할 때는 반드시 시각화를 사용해야 한다.

- 앞서 lexp와 gnppc의 관계를 살펴보았을때 일정 정도까지 gnppc가 증가하면 lexp가 증가하지만, 특정 경계를 넘어서면 lexp는 증가 추세를 멈추고 오히려 살짝 감소함을 확인하였다.
- 이 특정 경계를 수학적으로 다음과 같이 묘사할 수 있다(Why?).

$$\frac{\delta \text{lexp}}{\delta \text{gnppc}} = 0$$

- 함수의 미분값은 그 함수의 기울기를 의미한다(Why?).

이차항과 이차함수

- 이때 데이터에서 의해 관찰되는 gnppc의 범위(range)를 살펴보아야 한다!
- 특정 경계가 생각보다 아주 금방 나타날수도 있고, 반대로 결코 도달하지 않을 수도 있기 때문이다(Why?).
- 경계가 나타나는 시점에 따라, 이차항의 계수가 똑같은 양수(혹은 똑같은 음수)라도 그 함의가 완전히 달라질 수도 있다(Why?).
- 그러므로 반드시 그림을 그려야 하고, 그 도구인 margins와 marginsplot 사용법에 숙달되어야 한다.
- margins를 사용하기에 앞서 모형 추정에서는 반드시 연산자를 사용해야 하며 generate으로 이차항을 만들어선 안된다(Why?).

이차항과 이차함수

이차항의 유의성 검정은 무엇을 의미할까?

- 이차항에 대한 귀무가설을 통계적으로 유의하게 기각할 수 있다면, 모집단에서도 비선형관계가 있다고 할 수 있다.

$$H_0 : \beta_2 = 0 \qquad H_a : \beta_2 \neq 0$$

- 이차항을 넣었다면 이차항이 통계적으로 유의한가 여부가 중요할 뿐, 일차항 부분의 유의성 여부에는 주목할 필요가 없다(Why?).
- 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 의미만 해석하면 안된다(Why?).

이차항과 이차함수

- 그러므로 이차항을 살펴볼 때는 (논문/보고서에서 사용할 것인가와는 별개로) 위계적으로 모형을 구축해보고 살펴볼 필요가 있다.
- lexp를 종속변수로 하고, gnppc와 gnppc 이차항을 단계적으로 독립변수로 투입하는 위계적 회귀모형을 차례로 구축해보자.
- 요약통계량에서 사례수, F, R^2, Adj. R^2 를 꼼꼼하게 살펴보자.

이차항과 이차함수

이차항과 일차항 사이에는 높은 상관관계가 있는 것이 보통이다.

- life와 life^2 사이의 상관계수를 살펴보자. 매우 높다(Why?).
- 둘 다 한꺼번에 회귀식에 투입하면 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.

$$Y = \beta_0 + \beta_1(X - \bar{X}) + \beta_2(X - \bar{X})^2 + \epsilon$$

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

이차항과 이차함수

평균중심화의 영향은 직접 확인해 볼 수 있다.

- 다음의 평균중심화된 변수로 회귀식을 세웠다고 하자.

$$E(Y|X) = \beta_0 + \beta_1(X - \bar{X}) + \beta_2(X - \bar{X})^2$$

- 이는 평균중심화하지 않은 회귀식과 다음의 관계를 가지고 있다(Why?).

$$E(Y|X) = (\beta_0 - \beta_1\bar{X} + \beta_2\bar{X}^2) + (\beta_1 - 2\beta_2\bar{X})X + \beta_2X^2$$

- 위 식에서 확인할 수 있듯, 평균중심화를 하더라도 (1) 이차항에는 전혀 차이가 없지만, (2) 상수와 1차항을 엉망으로 뒤바꾸어 놓을 뿐이다.
- 어떤 변수에 대한 평균중심화는 다른 변수의 계수와 표준오차에 아무런 영향도 주지 않는다(Why?).
- 결론적으로 말해, 이차항 회귀식에서 평균중심화는 별 소득이 없다.

다항식

다항식

좀 더 높은 차원의 항을 추가하여 더 많은 굴곡을 추가할 수 있다.

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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + ... + \beta_k X^k + \epsilon$$

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

$$Y = 1 + X$$
$$Y = 1 + X + X^2$$
$$Y = 1 + X + X^2 + X^3$$

- 차수가 하나 증가할 때마다 굴곡이 하나씩 더 추가된다.

다항식

- 자료에 기반하여 산점도를 그리고 고차 적합선(high-order polynomial fitting line)으로 요약할 수도 있다.
- 미리 차수를 제시하기 어렵다면, 매우 탄력적인 분수다항함수(fractional polynomial function)를 사용하여 그려볼 수 있다.
- Stata 뿐만 아니라 사실 엑셀에서도 비슷한 기능을 지원한다.

다항식

모형이 복잡할수록 일반적인 설명력은 오히려 감소한다.

- 복잡한 모형은 (그것이 잘 맞는) 특수한 상황에서는 매우 뛰어난 설명력을 발휘할 수 있겠지만, 일반적인 상황에서는 오히려 형편없는 설명력만을 가진다.
- 데이터 사이언스(data science)에서는 이를 두고 과적합(overfitting)이라고 표현한다.
- 단순한 모형은 특수한 상황에서야 보잘 것 없는 설명력만을 가질 수도 있겠지만, 일반적인 상황에서 괜찮은 설명력을 가질 수 있다.
- 결국 밸런스를 유지하는 것이 중요하다. 문제는 "어떻게" 이다.

다항식

"변수 추가에도 불구하고 설명력에 변화가 없다"는 귀무가설을 검증한다.

- 이 목적을 위해 왈드 검정(Wald test)을 사용할 수 있다.
- Stata에서는 test 명령어로 수행한다.
- 왈드 검정은 다음과 같은 가설 구조를 F 검정한다 이때, m은 다항 차수를 의미한다.

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

- 해당 회귀계수의 t 값의 제곱과 왈드 검정의 F 값은 동일하다(Why?).

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) 정도까지만 다항식을 만드는 것이 보통이다.

다항식

- 이 아이디어를 일반화한 James Ramsey의 Regression Equation Specification Error Test (RESET)도 있다.
- Stata에서는 ovtest 명령어로 수행할 수 있다.
- 만약 귀무가설을 기각하는데 실패했다면, (설령 그 변수가 통계적으로 유의하더라도) 그 변수의 추가가 충분히 의미있게 설명력을 더한다고는 볼 수 없을 것이다.
- 본래 RESET은 누락변수(omitted variable)이 없나 확인하는 검정 기법이다. 이를 이차항에 적용한 셈이다. 구체적인 내용은 교재를 참고하자.

연습문제

연습문제

연습 1. nations.dta에서 삼차항을 고려한 다음 회귀식을 추정하고 해석하시오.

- death와 life 사이의 관계를 보여주는 산점도와 고차 적합선을 그려보자.
- death를 종속변수로 하고 food, life, life 이차항, life 삼차항을 독립변수로 하여 회귀모형을 구축해보자.
- life 3차항을 해석해보자. 통계적으로 유의한지도 확인하자.
- 위계적 회귀모형을 구축하여 모형간 변화를 적절히 확인해보자.