계량분석
Variable Transformation

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 29, 2021

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

진행 순서

1. 더미변수를 활용한 계단함수
2. 이차항과 상호작용항
3. 로그 변환

더미변수를 활용한 계단함수

더미변수를 활용한 계단함수

논리적으로 말하자면 이론이 일차항인지 이차항인지 다차항인지 여부를 말해 주어야 한다.

- 하지만 많은 경우 우리는 그렇게 강한 이론을 갖고 있지 않으므로 먼저 데이터 자체를 통해 살펴보지 않을 수 없다.
- 우리가 일차항이나 이차항 등 이미 다항 차수를 정해놓고 그에 따라 모형을 데이터에 적합(fit)시킨다면 그것은 어떤 의미에서 제약을 가하는 셈이다. “우리의 모형은 이러하다” 하고 미리 전제로 하고 모형을 설정한 뒤, 현실에 맞추어 본 것이기 때문이다.
- 따라서 구체적인 관계를 미리 설정하지 않고 (비효율적이지만) 개방성이 높은 모형을 일단 데이터 자체에 그대로 맞추어 보는 것도 대안이 될 수 있다.
- 그 대안 중 하나가 계단함수(step function)를 사용하는 것이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

더미변수를 활용한 계단함수

eCampus에서 KIELMC.DTA를 다운받아 Stata에서 열자. 이 데이터를 통해 집의 역사와 가격 사이의 관계를 살펴보기로 한다(**[Stata 코드]** 참고).

- 데이터를 간단히 살펴보자. 두 개의 year가 있다.
- 집값과 age의 관계를 산포도와 선형적합선으로 살펴보자.
- 집값(price)을 예측하는 회귀모형을 구축하기 위해 cbd, i.year, age를 독립변수로 투입하자. 집의 역사(age)와 집값의 관계는 선형적으로 어떠한가? 순서대로 age 이차항과 age 삼차항을 추가적으로 투입하여 회귀분석을 수행하자. 이들 변수들은 통계적으로 유의한가?
- **xtile** 명령어를 사용하여 age를 10개의 분위수(percentiles)로 나누어 범주형 변수로 재코딩하자.
- 기준범주(reference category)를 제외한 모든 더미 변수를 한 번에 넣어 회귀분석을 수행하자. 더미변수의 의미를 해석해보자.
- **margins**과 **marginsplot**을 통해 그래프를 만들어보자. 집의 역사(age)와 집값의 관계는 어떠한가?

더미변수를 활용한 계단함수

계단함수를 사용하는 방식은 사실 몇 가지 단점을 갖고 있다.

- 첫째로 수많은 더미변수를 회귀모형에 집어넣어야 하므로 그만큼 자유도를 많이 낮추게 된다. 모형의 효율성이 낮아진다.
- 둘째로 사용하는 범주의 구간이 자의적이 된다는 점이다. 만일 하나 이상의 범주에 지나치게 적은 사례만이 들어간다면 그 회귀계수는 매우 불안정할 것이라고 예상할 수 있다.
- 셋째로 원래 연속변수를 계단함수로 만든 과정에서 정보의 손실이 일어난다.
- 하지만 논문/보고서 등에 실제로 사용하지 않더라도 일단 계단함수를 고려하여 관심변수와 종속변수 사이의 관계를 제약없이 살펴보는 것은 좋은 출발점이 된다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 상호작용항

이차항과 상호작용항

이차항과 상호작용항을 함께 모형에 투입하여 이질적인 두 집단이 상이한 비선형관계를 갖고 있는지 확인할 수 있다.

- 저번주에 배운 상호작용항(=X_1X_2)은 다음과 같이 모형 속에 사용하였다.

$$y = \xi_0 + \xi_1 X_1 + \xi_2 X_2 + \xi_3 X_1 X_2 + \epsilon$$

- 오늘 배운 이차항(=X_2^2)은 다음과 같이 모형 속에 사용하였다.

$$y = \beta_0 + \beta_1 X_2 + \beta_2 X_2^2 + \epsilon$$

- 최소한 둘 이상의 이질적 집단(X_1)이 연구 대상 안에 포함되어 있다고 하자. 이 집단의 소속에 따라 어떤 독립변수(X)와 종속변수(y) 사이의 비선형관계가 상이할 수 있지 않을까? 두 항을 동시에 섞음으로써 이와 같은 특수한 이론을 검증해 볼 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 상호작용항

- 어렵게 생각하지 말고 두 식을 그냥 더하면 된다.

$$\begin{aligned} \mathrm{y} &= (\beta_0 + \beta_1 \mathrm{X}_2 + \beta_2 \mathrm{X}_2^2 + \epsilon) + (\xi_0 + \xi_1 \mathrm{X}_1 + \xi_2 \mathrm{X}_2 + \xi_3 \mathrm{X}_1 \mathrm{X}_2 + \mu) \\ &= (\beta_0 + \xi_0) + \xi_1 \mathrm{X}_1 + \xi_3 \mathrm{X}_1 \mathrm{X}_2 + (\beta_1 + \xi_2) \mathrm{X}_2 + \beta_2 \mathrm{X}_2^2 + (\epsilon + \mu) \end{aligned}$$

- 꼼꼼히 들여다보면 이차항과 상호작용항이 모두 들어있다.
- 통합된 식에서 $(\beta_0 + \xi_0)$ 은 상수가 되고, $(\epsilon + \mu)$ 은 오차항이 된다.

가장 직관적인 상호작용항인 더미변수와 연속변수의 결합을 살펴보자.

- 아래 식에서 만일 X_1가 0과 1의 더미변수라고 하자.

$$y = (\beta_0 + \xi_0) + \xi_1 X_1 + (\beta_1 + \xi_2)X_2 + \beta_2 X_2^2 + \xi_3 X_1 X_2 + (\epsilon + \mu)$$

- 각각 0과 1을 집어넣고 식을 단순화시켜 보면 각각 다음과 같다.

$$\begin{aligned} y &= (\beta_0 + \xi_0) + \xi_1 \cdot 1 + (\beta_1 + \xi_2)X_2 + \beta_2 X_2^2 + \xi_3 \cdot 1 \cdot X_2 + (\epsilon + \mu) \quad (\text{if } X_1 = 1) \\ &= (\beta_0 + \xi_0 + \xi_1) + (\beta_1 + \xi_1 + \xi_3)X_2 + \beta_2 X_2^2 + (\epsilon + \mu) \\ y &= (\beta_0 + \xi_0) + \xi_1 \cdot 0 + (\beta_1 + \xi_2)X_2 + \beta_2 X_2^2 + \xi_3 \cdot 0 \cdot X_2 + (\epsilon + \mu) \quad (\text{if } X_1 = 0) \\ &= (\beta_0 + \xi_0) + (\beta_1 + \xi_2)X_2 + \beta_2 X_2^2 + (\epsilon + \mu) \end{aligned}$$

- 꼼꼼히 들여다보면 더미변수에 따라 회귀계수가 조금씩 다름을 확인할 수 있다!

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

이차항과 상호작용항

연습을 위해 아까 사용하던 KIELMC.DTA 데이터를 다시 불러오자(**[Stata 코드]** 참고).

- 이차항과 상호작용항 통합식도 그래프를 그려가면서 해석해야 한다.
- 집값을 종속변수로, y81, cbd, y81 × cbd의 상호작용항을 독립변수로 회귀모형에 투입하자. **margins**와 **marginsplot**를 사용하여 그 관계를 그래프로 나타내보자. 그래프를 png 파일로 저장하자.
- 집값을 종속변수로, y81, cbd, cbd 이차항을 독립변수로 회귀모형에 투입하자. **margins**와 **marginsplot**를 사용하여 그 관계를 그래프로 나타내보자. 그래프를 png 파일로 저장하자.
- 상호작용항과 이차항을 모두 독립변수로 회귀모형에 투입하자. **margins**와 **marginsplot**를 사용하여 그 관계를 그래프로 나타내보자. 그래프를 png 파일로 저장하자. 앞서 얻은 결과와 어떻게 다른지 검토해보자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

이차항과 상호작용항

그 외의 상호작용항 유형에 대해서도 연습해야 한다.

- 방금 전에는 더미변수 × 연속변수 유형을 연습했다. 하지만 더미변수 × 더미변수나 연속변수 × 연속변수 유형도 있으므로 짬을 내서 연습을 해야 한다.
- 지난 주의 강의안과 코드를 참고할 것.
- 실제 분석 상황에서는 연속변수를 범주화하여 일부러 더미변수 × 연속변수의 꼴로 바꾸는 경우도 많이 있다. 이것이 해석상 편리하기 때문이다.

로그 변환

로그 변환

아예 데이터 상에서 변수를 변환(transformation)함으로서 비선형관계에 대응할 수도 있다.

- 가장 널리 활용되는 변환 방식은 로그 변환(logarithimic transformation)이다. 이는 밑(base)을 자연상수(e)로 하는 로그인 자연로그(natural log)를 사용하여 원변수를 변환한다.
- 자연로그는 자연상수와 대칭한다. 즉 $\ln(e^x) = x$ 이다.
- 원점수를 로그로 바꾸면 큰 값이 상대적으로 작아진다. 이 원리를 **desmos**에서 연습해보자. 어떤 의미에서 이것도 곡선형 데이터에 모형을 적합시키는 방법이 된다.
- 생각보다 많은 사회 · 경제 · 정치 현상에서 롱테일(long-tail)이 나타난다. 극단치가 반드시 나타나 그 분포가 다소 긴 꼬리를 갖는다. 이런 변수를 로그 변환하면 큰 값들이 상대적으로 작아지므로 분포가 온건하게 변화한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

KIELMC.DTA 데이터에서 price의 히스토그램을 살펴보고 로그 변환하자 ([Stata 코드] 참고).

- 지금까지 내내 종속변수로 price를 사용했지만 막상 히스토그램을 보지 않았다. 이제 확인해보자. 롱테일이 여기서도 나타난다.
- 로그 변환하여 새로운 변수를 만들자. 그 변수의 히스토그램도 꼼꼼히 살펴보자. 극단치는 어떻게 조정되었나?
- 새로운 로그 집값을 예측하는 회귀모형을 구축하기 위해 cbd, i.year, age를 독립변수로 투입하자. age와 로그 집값의 관계는 선형적으로 어떠한가?
- 이번에는 age 이차항을 추가적으로 투입하고 그래프를 그려보자. 이차항은 통계적으로 유의한가? 이번에는 그 관계가 어떠한가?
- cbd, cbd 이차항, cbd $\times$ y81 상호작용항을 투입하고 회귀분석을 수행하자. 그래프를 그려보고 아까와 차이가 있는지 확인하자.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

로그 변환은 해당 변수의 극단치를 보정하여 비선형인 관계를 선형적인 것으로 근사한다(approximate).

- 정규분포하는 하나의 변수와 정규분포에 근접했지만 롱테일을 가진 다른 하나의 변수 사이의 적합성을 상상해보자. 이 적합선은 곡선형일 것이다(Why?)
- 만일 롱테일 변수의 극단성을 줄여줄 수 있다면 그 곡선은 다시 직선에 근접하게 될 것이다.
- 사용하는 숫자형 변수들의 히스토그램을 그려보고 만일 롱테일을 가졌다면 한번씩 로그 변환을 해주어 모형에 투입해 주는 것이 바람직하다. 만일 로그 변환을 하더라도 결과가 크게 달라지지 않으면 안심할 수 있다.
- 키나 체중과 같은 변수는 좀 더 정규분포에 가깝다. 반면 소득과 재정과 같은 변수는 특히 롱테일을 갖기 쉽다. 특히 임금(wage)을 종속변수로 하는 경우에는 이른바 임금방정식(wage equation)을 세울 때 임금을 반드시 로그 변환한다. 이로 인해 해석에 유의해야 하는데 곧 다루게 된다.

로그 변환

KIELMC.DTA를 살펴보자(**[Stata 코드]** 참고).

- 히스토그램을 통해 어떤 변수들이 상대적으로 롱테일을 가지고 있는지 확인해보자. 이것들을 로그 변환해보고 회귀분석에서 원변수 대신 넣는 것을 고려해보자.
- 예를 들어 age를 로그 변환한다면 $\ln(0) = .$ 임에 주목하자. 즉 0의 자연로그는 정의될 수 없으므로 제법 많은 결측치가 발생하게 되는데 이를 피하기 위해 일부러 $\ln(x+1)$ 처럼 임의의 상수를 더해주기도 한다.
- "그런데 그래도 괜찮을까? 아니, 애시당초 멋대로 독립변수를 (로그) 변환해도 괜찮은걸까?"

로그 변환

- 결론적으로 그렇다. 왜냐하면 우리는 로그 변환을 하더라도 단조성 (monotonicity)를 여전히 기대할 수 있고 설령 똑같은 상수를 더해주어도 (정의상) 단조성은 유지되기 때문이다.
- 강한 단조성(strong monotonicity)은 "X가 증가(감소)할 때 y는 증가(감소)한다"는 원리를, 약한 단조성(weak monotonicity)은 "X가 증가(감소)할 때 최소한 y는 감소하지 않는다"는 원리를 의미한다.
- 그러므로 로그 변환할 때 실질적인 의미(예컨대 가격 액수나 주택의 연수같이 구체적인 의미)를 잃는 대신 여전히 "X가 증가(감소)할 때 y는 증가(감소)한다" 와 같은 해석은 유지할 수 있게 된다.

로그 변환

로그 변환을 넘어 좀 더 복잡한 변환 원리도 있다.

- 가령 Stata에서 **lnskew0** 명령어는 왜도(skewness)를 0에 가깝게 해주는 $\ln(X-k)$ 변환을 자동적으로 수행해준다. 여기서 k는 (아까처럼 임의로 사용한 1이 아닌) "왜도를 최소화해주는" 어떤 상수이다.
- 다만 일반적으로 **lnskew0** 명령어를 사용하던 그냥 평범하게 로그 변환하던 구한 값은 매우 유사하다.
- 좀 더 수준 높은 변환 원리는 Box-Cox 변환(Box-Cox Transformation)이지만 우리 수업에서는 다루지 않는다. Stata에서는 **boxcox**와 **bcskew0**로 사용할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

로그 변환의 진가는 해석을 명확하게 해주는 것에 있다.

- $y = \beta_0 + \beta_1 X$ 와 같이 선형모형이 주어졌을 때, y를 X_1에 대해 편미분($\delta y/\delta X$)하면 그것은 "X가 한 단위 변화할 때, y가 얼마나 변화하는지"를 보여준다(Why?).

$$\frac{\delta y}{\delta X} = \beta_1$$

- 이를 일컨어 X의 y에 대한 한계효과(marginal effect)라고 부른다.
- 이 식 $\delta y/\delta X$이 보여주듯 일반적으로 한계효과는 곧 회귀계수(β_1) 그 자체이다.
- "X가 한 단위 변화할 때, y가 얼마나 변화하는가"는 $\frac{\delta y}{\delta X}$로 표현될 수 있고 그래프에서는 곧 기울기(slope)를 의미한다(Why?).

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

- $\ln(y) = \beta_0 + \beta_1 \cdot \ln(X)$ 와 같이 로그-로그(log-log) 모형이 주어졌다고 하자. 즉 종속변수와 독립변수를 모두 로그 변환한 경우에 해당한다.
- 이 식을 y에 대해 정리하면 $y = e^{\beta_0 + \beta_1 \cdot \ln(X)}$ 이고 y를 X에 대해 편미분하면 이는 다음과 같다.

$$\frac{\delta y}{\delta X} = \frac{\delta e^{\beta_0 + \beta_1 \cdot \ln(X)}}{\delta(\beta_0 + \beta_1 \cdot \ln(X))} \cdot \frac{\delta(\beta_0 + \beta_1 \cdot \ln(X))}{\delta X}$$
$$= e^{\beta_0 + \beta_1 \cdot \ln(X)} \cdot \frac{\beta_1}{X} = y\frac{\beta_1}{X}$$

- 이를 다시 β_1 에 대해 정리하면 다음과 같다.

$$\beta_1 = \frac{\delta y}{\delta X} \cdot \frac{X}{y} = \frac{\left(\frac{\delta y}{y}\right)}{\left(\frac{\delta X}{X}\right)}$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

- 이렇게 구한 회귀계수는 무단위적(unitless)이라는 중요한 특성을 갖는다. 이로 인해 해석이 매우 용이해지므로 실질적인 유의성(substantial significance)을 확인하는데 중요한 수단이 될 수 있다.
- 보다 구체적으로 이 회귀계수는 이렇게 해석된다: “X가 1 퍼센트 변화할 때 y는 몇 퍼센트 변화하는가”(Why?).
- 다시 말해, 종속변수와 독립변수를 모두 로그 변환하면 회귀계수를 해석할때 퍼센트 변화로 전환하여 해석할 수 있다.
- 로그-로그 모형은 미시경제학에서 탄력성(elasticity) 개념과 잇닿아있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

KIELMC.DTA 데이터로 로그-로그 모형을 연습하자(**[Stata 코드]** 참고).

- (price를 로그 변환한) lprice를 종속변수로, age, i.y81, cbd 그리고 (area를 로그 변환한) larea를 독립변수로 한 회귀모형을 만들자. larea는 통계적으로 유의한 변수인가? 어떻게 해석할 것인가?
- "집의 면적이 1퍼센트 증가함에 따라 집값은 .696 퍼센트 증가한다" 혹은 "집의 면적이 10퍼센트 증가함에 따라 집값은 6.96 퍼센트 증가한다" 정도로 해석할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

수리적 증명을 생략하지만 로그-선형(log-linear) 모형이나 선형-로그 (linear-log) 모형도 있다([Stata 코드] 참고).

- 로그-선형 모형의 회귀계수는 이렇게 언어로 해석된다: "X가 한 단위 변화할 때 y는 몇 퍼센트 변화하는가."
- KIELMC.DTA 데이터에서 lprice를 종속변수로, age, i.y81, cbd, area를 독립변수로 한 로그-선형 모형을 만들자.
- "집의 면적이 한 단위(sqft.) 증가함에 따라 집값은 .034 퍼센트 증가한다" 혹은 "집의 면적이 백 단위(sqft.) 증가함에 따라 집값은 3.4 퍼센트 증가한다" 정도로 해석할 수 있다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

로그 변환

- 선형-로그 모형의 회귀계수는 이렇게 언어로 해석된다: “X가 1 퍼센트 변화할 때 y는 얼마만큼 변화하는가.”
- KIELMC.DTA 데이터에서 price를 종속변수로, age, i.y81, cbd, larea를 독립변수로 한 선형-로그 모형을 만들자.
- “집의 면적이 1퍼센트 증가함에 따라 집값은 692.73 달러만큼 증가한다” 정도로 해석할 수 있다.
- 특히 퍼센티지 계산이 몹시 혼동스러울 수 있으므로 주의를 기울여야 한다. 원래 수식을 들여다보면 알 수 있지만 본래 비율(proportion)이였으므로 퍼센트(percentage)로 계산할 때 100을 곱하거나 나누는 상황을 잘 구분해야 한다.