계량분석
Variable Transformation

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 21, 2024

진행 순서

1. 가변수를 활용한 계단함수
2. 이차항과 상호작용항
3. 로그 변환
4. 연습문제

가변수를 활용한 계단함수

가변수를 활용한 계단함수

이론이 일차항인지 이차항인지 다차항인지 여부를 말해주어야 한다.

- 하지만 많은 경우 우리는 그렇게 강한 이론을 갖고 있지 않으므로, 먼저 데이터 자체를 통해 살펴볼 수 밖에 없다.
- 우리가 일차항이나 이차항 등 이미 다항 차수를 정해놓고 그에 따라 모형을 데이터에 적합(fit)시킨다면, 그것은 어떤 의미에서 회귀모형에 제약을 가하고 적합시키는 셈이다(Why?).
- 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.

가변수를 활용한 계단함수

Stata에서 집의 역사와 가격 사이의 관계를 계단함수로 살펴보자.

- KIELMC.DTA을 보면 두 개의 year가 있다.
- 집값 price와 집의 나이 age의 관계를 산포도와 선형적합선으로 살펴보자.
- price를 예측하는 회귀모형을 구축하기 위해 cbd, i.year, age를 독립변수로 투입하자.
- age와 집값의 선형적 관계는 어떠한가? 순서대로 이차항과 삼차항을 추가하여 회귀분석을 수행하자. 이들은 통계적으로 유의한가?
- xtile 명령어를 사용하여 age를 10개의 분위수(percentiles)로 나누어 범주형 변수로 재범주화하자.
- 기준범주를 제외한 모든 가변수를 한 번에 넣어 회귀분석을 수행하자.
- margins과 marginsplot을 통해 그래프를 만들어보자. age와 price의 관계는 어떠한가?

가변수를 활용한 계단함수

계단함수를 사용하는 방식은 사실 몇 가지 단점을 갖고 있다.

- 첫째로 수많은 가변수(dummy variables)를 회귀모형에 집어넣어야 하므로 그만큼 자유도(degree of freedom)를 소모해야 한다.
- 둘째로 원래 연속변수를 계단함수로 만든 과정에서 정보의 손실이 일어난다.
- 셋째로 사용하는 범주의 구간이 자의적이 된다는 점이다. 만일 하나 이상의 범주에 지나치게 적은 사례만이 들어간다면 그 회귀계수는 매우 불안정할 것이라고 예상할 수 있다(Why?).
- 하지만 논문/보고서 등에 실제로 사용하지 않더라도 일단 계단함수를 고려하여 관심변수와 종속변수 사이의 관계를 제약없이 살펴보는 것은 좋은 출발점이 된다.

이차항과 상호작용항

이차항과 상호작용항

이질적인 두 집단이 상이한 비선형관계를 갖고 있는지도 확인할 수 있다.

- 가령 (1) 노조원과 비노조원이라는 두 이질적 집단 X_1이 있고, (2) 직무경력 X_2의 임금 Y에 대한 효과는 비선형적이라고 하자.
- 노조원 여부에 따라 직무경력과 임금 사이의 비선형관계가 상이할 수 있지 않을까?
- 그렇다면 상호작용항과 이차항을 동시에 사용하여 이와 같은 특수한 이론을 검증해 볼 수 있다.
- 먼저 노조원과 직무경력의 상호작용항 X_1X_2은 다음과 같이 회귀식에서 사용한다.

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2 + \mu$$

- 직무경력의 이차항 X_2^2은 다음과 같이 모형 속에 사용한다.

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

이차항과 상호작용항

- 어렵게 생각하지 말고 두 식을 그냥 더하면 된다.

$$\begin{aligned} Y &= (\beta_0 + \beta_1 X_2 + \beta_2 X_2^2 + \epsilon) + (\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_1 X_2 + \mu) \\ &= (\beta_0 + \gamma_0) + \gamma_1 X_1 + \gamma_3 X_1 X_2 + (\beta_1 + \gamma_2) X_2 + \beta_2 X_2^2 + (\epsilon + \mu) \end{aligned}$$

- 꼼꼼히 들여다보면 이차항과 상호작용항이 모두 들어있다.
- 통합된 식에서 $(\beta_0 + \gamma_0)$ 은 상수가 되고, $(\epsilon + \mu)$ 은 오차항이 된다.

이차항과 상호작용항

가장 직관적인 상호작용항인 가변수와 연속변수의 결합을 살펴보자.

- 아래 식에서 노조원인 경우 $X_1 = 1$이라고 하자.

$$Y = (\beta_0 + \gamma_0) + \gamma_1 X_1 + (\beta_1 + \gamma_2) X_2 + \beta_2 X_2^2 + \gamma_3 X_1 X_2 + (\epsilon + \mu)$$

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

$$\begin{aligned}
Y &= (\beta_0 + \gamma_0) + \gamma_1 \cdot 1 + (\beta_1 + \gamma_2) X_2 + \beta_2 X_2^2 + \gamma_3 \cdot 1 \cdot X_2 + (\epsilon + \mu) \quad (\text{if } X_1 = 1) \\
&= (\beta_0 + \gamma_0 + \gamma_1) + (\beta_1 + \gamma_1 + \gamma_3) X_2 + \beta_2 X_2^2 + (\epsilon + \mu) \\
Y &= (\beta_0 + \gamma_0) + \gamma_1 \cdot 0 + (\beta_1 + \gamma_2) X_2 + \beta_2 X_2^2 + \gamma_3 \cdot 0 \cdot X_2 + (\epsilon + \mu) \quad (\text{if } X_1 = 0) \\
&= (\beta_0 + \gamma_0) + (\beta_1 + \gamma_2) X_2 + \beta_2 X_2^2 + (\epsilon + \mu)
\end{aligned}$$

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

이차항과 상호작용항

Stata에서 다시 연습해보자.

- price를 종속변수로, y81, cbd, y81×cbd의 상호작용항을 독립변수로 회귀모형에 투입하자.
- price를 종속변수로, y81, cbd, cbd 이차항을 독립변수로 회귀모형에 투입하자.
- 상호작용항과 이차항을 모두 회귀모형에 투입하자.
- 상호작용항과 이차항의 상호작용항까지도 회귀모형에 투입하자.
- 결과가 각각 어떻게 다른지 검토해보자.
- 특히 margins와 marginsplot를 사용하여 그 관계를 그래프로 나타내보자.

로그 변환

로그 변환

변수를 변환하여 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 원점수를 변환한다.
- 자연로그는 자연상수와 대척한다. 즉 $ln(e^x) = x$ 이다.
- 원점수를 로그로 바꾸면 큰 값이 상대적으로 작아진다. 이 원리를 desmos에서 연습해보자.
- 어떤 의미에서 이것도 비선형모형을 적합시키는 방법이 된다(Why?).
- 많은 사회현상에서 롱테일(long-tail)이 나타난다. 소수의 극단치가 꼭 나타나 그 분포가 다소 긴 꼬리를 갖는다.
- 이런 변수를 로그 변환하면 큰 값들이 상대적으로 작아지므로 분포가 온건하게 변화한다.

로그 변환

Stata에서 로그 변환을 연습해보자.

- 지금까지 내내 종속변수로 KIELMC.DTA에서 price를 사용했지만 막상 히스토그램을 보지 않았다(확인해보면 롱테일이 여기서도 나타난다).
- 로그 변환하여 새로운 변수 lprice를 만들고 그 변수의 히스토그램도 살펴보자.
- 로그 변환은 해당 변수의 이상점(outliers)을 보정하는 의미를 갖는다. 이상점은 어떻게 조정되었나?

로그 변환

사용하는 숫자형 변수들의 히스토그램을 그려보자.

- 히스토그램이 만일 롱테일을 가졌다면 한번씩 로그 변환을 해주어 모형에 투입해 주는 것이 바람직하다.
- 만일 로그 변환을 하건 하지 않건 회귀분석 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 하여, 이른바 임금방정식(wage equation)을 세울 때는 임금을 반드시 로그 변환한다. 이로 인해 해석에 유의해야 하는데 곧 다루게 된다.
- 이때 $ln(0) =$. 임에 주의해야 한다. 즉 0의 자연로그는 정의될 수 없으므로 영유아 표본에서 age같은 경우에는 제법 많은 결측치가 발생할 수도 있다(Why?).
- 이를 피하기 위해 일부러 $ln(x+1)$ 처럼 임의의 상수를 더해주기도 한다.

로그 변환

"아니, 멋대로 변수를 (로그) 변환해도 괜찮은걸까?"

- 결론적으로 그렇다! 왜냐하면 로그 변환에는 단조성(monotonicity)이 있기 때문이다.
- 단조성에는 두 가지 종류가 있다. 강한 단조성(strong monotonicity)은 "X가 증가(감소)하는 모든 상황에서 Y는 증가(감소)한다"는 원리를, 약한 단조성(weak monotonicity)은 "X가 증가(감소)모든 상황에서 Y는 최소한 감소하지 않는다"는 원리를 의미한다.
- 로그 변환할 때 실질적인 의미(예컨대 가격 액수나 주택의 연수같이 구체적인 의미)를 잃는 대신, 단조성 덕택에 여전히 "X가 증가(감소)할 때 Y는 증가(감소)한다"와 같은 해석을 유지할 수 있다(Why?).

로그 변환

로그 변환보다 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.
- 일반적으로 lnskew0 명령어를 사용하던 그냥 평범하게 로그 변환하던 구한 값은 매우 유사하다.
- Box-Cox 변환(Box-Cox transformation) 등도 있지만 우리 수업에서는 다루지 않는다.
- Stata에서는 boxcox와 bcskew0로 사용할 수 있다.

로그 변환

로그 변환은 회귀계수의 해석을 다르게 만든다.

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

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

- β_1을 일컫어 X의 Y에 대한 한계효과(marginal effect)라고 부른다.
- "X가 한 단위 변화할 때, Y가 얼마나 변화하는가"는 $\delta Y / \delta X$로 표현될 수 있고 그래프에서는 곧 기울기(slope)를 의미한다(Why?).

로그 변환

- $ln(Y) = \beta_0 + \beta_1 \cdot lnX$ 와 같이 로그-로그(log-log) 모형이 주어졌다고 하자. 즉 종속변수와 독립변수를 모두 로그 변환한 경우에 해당한다.
- 이 식을 Y에 대해 정리하면 $Y = e^{\beta_0 + \beta_1 \cdot lnX}$ 이고, Y를 X에 대해 편미분하면 다음과 같다.

$$\begin{aligned}\frac{\delta Y}{\delta X} &= \frac{\delta e^{\beta_0 + \beta_1 \cdot lnX}}{\delta(\beta_0 + \beta_1 \cdot lnX)} \cdot \frac{\delta(\beta_0 + \beta_1 \cdot lnX)}{\delta X} \\ &= e^{\beta_0 + \beta_1 \cdot lnX} \cdot \frac{\beta_1}{X} = Y\frac{\beta_1}{X}\end{aligned}$$

- 이를 다시 β_1에 대해 정리하면 다음과 같다.

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

로그 변환

- 보다 구체적으로 이 회귀계수는 이렇게 해석된다: "X가 1 퍼센트 변화할 때 Y는 β_1 퍼센트 변화한다."
- 다시 말해, 종속변수와 독립변수를 모두 로그 변환하면 회귀계수를 해석할 때 퍼센트 변화로 전환하여 해석할 수 있다.
- 로그-로그 모형은 미시경제학에서 탄력성(elasticity) 개념과 잇닿아있다.
- 이렇게 구한 회귀계수는 무단위적(unitless)이라는 중요한 특성을 가지며, 이로 인해 실질적인 유의성(substantial significance)을 확인하는데 중요한 수단이 된다!

로그 변환

KIELMC.DTA로 로그-로그 모형을 연습하자.

- (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 10퍼센트 증가함에 따라 집값은 6.96 퍼센트 증가한다."

로그 변환

한쪽만 로그 변환하면 해석은 어떻게 달라질까?

- 로그-선형(log-linear) 모형의 회귀계수는 이렇게 해석된다: "X가 한 단위 변화할 때 Y는 β_1 퍼센트 변화한다."
- 선형-로그(linear-log) 모형의 회귀계수는 이렇게 해석된다: "X가 1 퍼센트 변화할 때 Y는 β_1만큼 변화한다."
- 퍼센티지 계산에서 혼동하지 않도록 주의를 기울여야 한다. 원래 수식을 들여다보면 알 수 있지만 본래 비율(proportion)이였으므로 퍼센트로 계산할 때 100을 곱하거나 나누는 상황을 잘 구분해야 한다.

연습문제

연습문제

연습 2. KIELMC.DTA을 사용하여 다음에 답하시오.

- 첫번째 모형의 종속변수는 price이고, 독립변수는 age, age 이차항, lland, age×y81 상호작용항이다.
- 첫번째 모형의 종속변수는 lprice이고, 독립변수는 age, age 이차항, lland, age×y81 상호작용항이다.
- 첫번째 모형의 종속변수는 lprice이고, 독립변수는 age, age 이차항, land, age×y81 상호작용항이다.
- 세 회귀분석의 결과를 비교하여 꼼꼼하게 해석하시오.