계량분석
Assumptions about the Error Term

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 14, 2024

진행 순서

1. 오차항의 가정: 조건부 영평균
2. 오차항의 가정: 등분산성
3. 오차항의 가정: 자기상관 없음
4. 오차항의 가정: 정규성
5. 회귀 가정에 대한 코멘트
6. 연습문제

오차항의 가정: 조건부 영평균

오차항의 가정: 조건부 영평균

조건부 영평균은 개념적으로 혼동스러우니 주의해야 한다.

- 먼저 조건부 영평균의 정의를 살펴보자. 조건부 기댓값(conditional expectation) 개념을 수학적으로 꼼꼼히 살펴보지 않고 넘어가기 때문에 혼란스러울 수 있다.
- 이것은 어떤 독립변수가 데이터 안에서 "특정 값"을 가진다고 전제하고($X = x$) 모형에서 구한 오차항의 평균값이 0이라는 가정이다.

$$E(\epsilon_i|X) = 0$$

- 가령 지도배부 수(map)로 이용객 수(rider)를 예측한다고 하자. 지도배부 수가 0일 때도, 100일 때도, 200일 때도, 추정 이후 모형에서 남은 오차항은 평균적으로 0이어야 한다.

오차항의 가정: 조건부 영평균

- 이 가정은 평균 독립성(mean independence)이라고도 불리운다(Allison 1998).
- 비확률적 독립변수(non-stochastic Xs) 또는 조건부 영평균 덕택에 다음의 식이 성립한다(Why?).

$$E(Y|X) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

오차항의 가정: 조건부 영평균

- 초보자는 조건부 영평균과 영평균(zero mean)은 다르다는 점을 종종 혼동한다.
- 영평균은 (가정이고 뭐고 상관없이) 정규방정식(normal equation)으로부터 당연히 도출된다(Why?).

$$E(\epsilon_i) = 0$$

- 게다가 그 실질적인 의미도 사뭇 다르다. 영평균은 무조건 평균(unconditional mean)에 관한 것이다.
- 교과서에 따라서는 영평균을 조건부 영평균 대신 써놓는 경우가 있다. 이는 비확률적 독립변수를 전제했기 때문이다. 이때는 자동적으로 $Cov(\epsilon, X) = 0$ 가정이 성립하므로 조건을 붙일 필요조차 없다(Why?).

오차항의 가정: 조건부 영평균

조건부 영평균은 상수를 회귀식 안에 넣으면 자동적으로 성립한다.

- 회귀식의 상수는 괜히 있는 것이 아니다. 설령 $E(\epsilon_i|X) = 0$가 성립하지 않더라도 그것을 성립시키기 위해 적절히 오차항에 더하기 빼기를 추가하는 상수항을 넣어주면 이 가정을 무조건 성립시킬 수 있다!
- 즉 $E(\epsilon_i|X) \neq 0$ 이더라도 $E(\epsilon_i'|X) = 0$ 이 되도록 γ를 더하고 빼주면 된다.

$$\begin{aligned} Y &= (\beta_0 + \gamma) + \beta_1 X_1 + ... + \beta_k X_k + (\epsilon - \gamma) \\ &= \beta_0' + \beta_1 X_1 + ... + \beta_k X_k + \epsilon' \end{aligned}$$

- 물론 이 경우 β_0 대신 β_0'를 얻었으므로 추정된 상수는 왜곡된다(biased).

오차항의 가정: 조건부 영평균

조건부 영평균은 특히 모형 설정에 민감하다.

- '왜곡되지 않은 상수 추정을 전제로 했을 때," 조건부 영평균 가정은 주로 모형설정 오류(model mis-specification)나 변수의 측정오류(measurement error) 등으로 인해 위협받는다.
- 만약 진정한 Y와 X의 관계를 잘 반영하지 않는 회귀모형을 데이터에 적합시켰다면 조건부 영평균이 성립하리라고 기대하기 어렵다.
- 그러나 완전모형 설정(perfect model specification)은 보통 불가능하므로 이 부분은 결국 이론과 기존문헌에 의존할 수 밖에 없다.

오차항의 가정: 조건부 영평균

조건부 영평등의 진단법은 사실 따로 없다.

- 표본을 회귀분석하고 추정한 $\hat{\epsilon}$은 이미 조건부 영평등을 가정하고 구한 것이다. 그러므로 표본의 잔차항을 열심히 살펴본다고 해서 가정의 성립 여부를 검증할 수 없다.
- Stata에서 비선형관계를 선형모형으로 적합시킨 다음, 잔차(residuals)를 계산하고 이를 독립변수 X에 대해 산포도와 적합선을 그려보자.
- 이러한 그림을 RVP 도표(Residuals-Versus-Predictor plot)라고 부른다.
- $E(\epsilon_i|X_i) = 0$가 성립하고 그렸으므로 상식적으로 RVP 도표는 "모든 X에 대해" $Y = 0$에 계속해서 수평선이 된다.
- 선형모형에서 그린 RVP 도표와 비선형모형에서 그린 RVP 도표를 비교해보자.

오차항의 가정: 등분산성

오차항의 가정: 등분산성

등분산성은 오차항의 분산이 X에 따라 변화하지 않는다는 가정이다.

- 이 가정의 의미를 그림으로 이해하자. 독립변수가 어떤 값이든 오차항이 일정하게 퍼져있음을 의미한다.

$$Cov(\epsilon_i, \epsilon_i) = Var(\epsilon_i|X) = \sigma^2$$

- 이 가정이 위배되는 가장 흔한 원인으로 두 가지가 지목된다:

  (1) 변수에 이상점(outliers)이 있거나 분산이 커서 들쭉날쭉한 값이 끼어있는 경우,
  (2) 모형에 반드시 있어야 하는 변수가 빠져 불완전한 모형 설정(model mis-specification)이 이루어진 경우.

- 등분산성이 위배되는 상황을 이분산성(heteroscedasticity)이라고 부른다.

오차항의 가정: 등분산성

RVP 도표나 RVF 도표를 그려보는 것이 가장 기본적인 진단법이다.

- Stata에서 시뮬레이트된 자료를 활용하여 일부러 불완전한 회귀모형을 추정해보자.
- 추정된 잔차항과 독립변수와의 관계를 살펴보면 독립변수가 변화할 때 추정된 잔차항의 분산도 변화하는 현상을 확인할 수 있다.
- 다음에는 완전한 모형을 만들어 회귀분석을 해보자.
- 이번엔 독립변수가 변화하더라도 추정된 잔차항의 분산은 일정함을 확인할 수 있다.

오차항의 가정: 등분산성

- 좀 더 복잡한 방법은 Breusch-Pagan 검정이 있다.
- 귀무가설은 등분산성이고 대립가설은 이분산성이다.

$$H_0 : Var(\epsilon) = \sigma^2$$
$$H_a : Var(\epsilon) \neq \sigma^2$$

- 검정통계량은 χ^2 이다(Why?).
- 대부분의 χ^2 검정이 그러하듯 이 검정도 표본 수에 지나치게 민감하다는 단점이 있다.

오차항의 가정: 등분산성

이분산성 상황에 대해서도 여러가지 대응책이 있다.

- 만약 이상점이 존재하는 상황이라면 이것을 제거하고 다시 회귀분석을 수행하면 된다.
- 변수의 분산이 크다면 로그 변환을 시도해 볼 수 있다. 이런 식의 로그 변환은 특별히 분산안정화 변환(variance-stabilizing transformation)이라고도 불리운다.
- 일반적인 표준오차 대신 강건표준오차(robust standard error)를 보고한다. Stata 에서 이것은 regress 명령어 뒤에 robust 옵션을 붙여서 손쉽게 얻어낼 수 있다.

오차항의 가정: 등분산성

- 위의 모든 방법으로도 이분산성 문제가 심각하게 남아있는 경우 최후의 수단으로 가중최소자승(Weighted Least Square; WLS)을 사용한 회귀분석을 수행할 수 있다.
- 이는 보통최소자승(OLS) 대신 이분산을 야기하는 구조 그 자체를 모형 속에 고려하는 방법으로 계산 논리는 흥미롭지만, 현실적으로 이분산 구조를 모형 속에 제대로 구현하지 못할 경우 더 큰 문제를 야기할 수도 있다.
- 등분산성 가정이 위배되더라도 추정값의 표준오차(standard errors)에 큰 차이가 없으면 너무 걱정하지 않아도 된다.
- 그러므로 다양한 방법으로 모형을 추정해 보고 나란히 보고하여 얼마나 추정이 강건한가를 주장할 수 있다(온라인 부록도 물론 활용할 수 있다).

오차항의 가정: 자기상관 없음

오차항의 가정: 자기상관 없음

자료 안에서 오차항 사이에는 상관관계가 없어야 한다.

- 데이터에 대해 회귀모형을 적합시킨 뒤, 임의의 i번째 사례에 대해 추정된 오차항 e_i와 (그와는 다른) j번째 사례에 대해 추정된 오차항 e_j 사이에 공분산(covariance)이 0이라는 가정이다.

$$Cov(\epsilon_i, \epsilon_j) = 0$$

- 개별 사례들의 성격은 오로지 모형 속에 추가된 독립변수로 공통점을 완전히 설명할 수 있었고, (설명하지 못하고 남은) 오차항 사이에는 평균적으로 아무런 공통점도 남아있지 않아야 한다.

오차항의 가정: 자기상관 없음

등분산성 가정과 자기상관 없음 가정은 사실 밀접하게 연관되어 있다.

- 오차항의 공분산행렬(variance-covariance matrix)은 다음과 같이 가정된다.

$$Cov(e_i, e_j) = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

- 대각행렬 부분이 바로 등분산성 가정이고 그 나머지 부분은 자기상관 없음 가정에 의해 채워진다(Why?).

$$Cov(\epsilon_i, \epsilon_i) = Var(\epsilon_i|X) = \sigma^2$$
$$Cov(\epsilon_i, \epsilon_j) = 0$$

오차항의 가정: 자기상관 없음

- 시공간적 자기상관(autocorrelation)이 존재할 때 특히 이 가정이 위배된다.
- (분석단위가 개인인 경우) 사회학이나 사회역학(social epidemiology)의 세계관에서 이 가정은 특히 문제가 된다.
- 시계열적 자기회귀(time-series autoregression)는 현재의 오차항 e_t이 과거의 오차항 e_{t-1}과 자기상관을 갖는 현상을 지칭한다.
- "오늘의 나를 가장 잘 설명하는 변수는 과거의 나"라는 평범한 진리는 자기상관 없음 가정을 위배하고 있다.

오차항의 가정: 자기상관 없음

- 공간적 자기회귀(spatial autoregression)는 특정 지역의 오차항이 그 근린의 오차항과 자기상관을 갖는 현상을 지칭한다.
- "가까운 것은 먼 것보다 중요하다(Near things are more related than distant things)"는 토블러의 지리학 제1법칙(Tobler's First Law of Geography)을 떠올릴 것.
- 뒤집어 말하자면, 이렇게 시계열 자료(time-series data)나 공간자료(spatial data)가 주어진 경우 보통최소자승(OLS) 회귀분석으로는 적절한 분석이 불가능한 까닭을 여기서 알 수 있다.

오차항의 가정: 자기상관 없음

자기상관 없음의 진단법은 지금 고민할 필요가 없다.

- 자기상관이 나타나는 상황 자체가 이미 수준 높은 고급사회통계학의 영역이다.
- 예컨대 흔히 시계열분석에서는 Durbin-Watson 통계량(statistic)부터 공부하기 시작하고, 공간회귀분석에서는 Moran's I부터 공부하기 시작하는데 이미 우리 수업의 단계를 훨씬 넘어선 것이다.
- 데이터의 수집 과정과 성격 자체를 들여다보고 자기상관이 존재할 것인지를 개념적으로 판단하자.

오차항의 가정: 자기상관 없음

자기상관이 존재하는 경우에도 대응책은 마련되어 있다.

- 가장 단순한 방법은 애시당초 순수한 임의표집(random sampling)으로 자기상관이 존재하지 않도록 하는 것이다. 이것은 여러가지 의미에서 현실적으로 어려운 경우가 많다(Why?).
- 다음으로 일반적인 표준오차 대신 강건표준오차(robust standard error) 혹은 군집표준오차(clustered standard error)를 보고한다. Stata에서는 regress 명령어 뒤에 robust 또는 cluster(·) 옵션을 사용한다.
- 또다른 방법은 자기상관의 구조를 직접 모델링하여 구현하는 것이다. 시계열분석(time-series analysis)나 공간회귀분석(spatial regression analysis), 소셜네트워크 회귀모형(social network regression modeling) 등은 이런 맥락에서 발전하였다.

오차항의 가정: 정규성

오차항의 가정: 정규성

필수적인 가정은 아니지만 가설검정 절차를 편리하게 만들어준다.

- 오차항은 정규분포한다는 것이 이 가정의 전부이다.

$$\epsilon_i \sim N(0, \sigma^2)$$

- 조건부 영평균과 등분산성을 결합한다고 해도 자동적으로 이 가정이 성립하는 것은 아니다(Why?). 두 가정은 정규분포에 대해 어떠한 전제도 하지 않는다!
- 이 정규성 가정이 워낙 포괄적(overarching)인 까닭에 논문에서는 이것 하나만 쓰고 "가정이 성립한다고 전제한다"라고 요약할 때가 많다.

오차항의 가정: 정규성

- 학부과정에서 배운대로 소표본인 경우 오차항이 정규분포해야만 회귀계수의 t 검정을 위해 t 분포를, 모형 적합도 검정을 위해 F 분포를 믿고 사용할 수 있다.
- 만약 이 가정이 없었다면, t 통계량이나 F 통계량의 논리는 더이상 성립하지 않으므로 신뢰구간(confidence interval)과 유의성 검정 결과를 신뢰할 수 없게 된다.
- 한편 사례가 충분히 크다면(대표본) 이 가정은 불필요하다. 표본이 커지면 t 분포는 자연스럽게 Z 분포를 따라가고, F 분포의 근본이 되는 χ^2 분포 역시 대표본에서는 정규분포에 근접하기 때문이다.

오차항의 가정: 정규성

정규성 가정은 보통 그림을 통해 진단한다.

- 표준화된 정규확률도표(standardized normal probability plot) 혹은 PP (Percent-Percent) 도표를 그린다.
- 이 그림의 Y 축은 이론적 정규분포의 누적표준분포, 즉 기대누적확률(expected cumulative probability)를 의미하고, X 축은 관찰된 데이터의 누적표준분포, 즉 관찰누적확률(observed cumulative probability)를 의미한다.
- 두 값들이 정확히 일치하면 그림은 45도 선을 따라가기 마련이지만, 만일 그렇지 않다면 정규성 가정이 아무래도 위배되고 있다고 의심할 수 있다.
- Shapiro-Wilk 정규성 검정(test of normality)라는 기법도 있다.
- 이들 기법은 사실 t, χ^2, F 분포 등을 사용하기 앞서 정규성 여부를 확인하기 위해서도 가끔씩 쓰이곤 한다. 하지만 대표본이라면 이 가정에 그렇게까지 집착할 필요는 없어 보인다.

회귀 가정에 대한 코멘트

회귀 가정에 대한 코멘트

가정이 위배되면 더이상 보통최소자승이 최적이라고 보장할 수 없다.

- 보통 연구자들은 다음과 같이 (모든 가정을 충족하는) 고전적인 회귀모형을 짧게 정의한다.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}$$

$$\mathbf{e} \sim N(0, \sigma^2 \mathbf{I})$$

- 볼드(bold)로 두껍게 표시된 기호는 이것들이 행렬임을 의미한다.

회귀 가정에 대한 코멘트

회귀분석의 가정은 고급통계학의 관문과도 같다.

- $1 + 1 = 2$ 를 배우는 것은 금방이지만 왜 이러한 식이 성립하는지 공리(axioms)나 가정을 공부하는 것은 어렵다.
- 회귀분석의 가정을 공부할 때는 일단 각 가정이 무엇을 의미하는가를 살펴보고 이를 직관적으로 파악하는데 중점을 두자. 처음 공부할 때는 증명은 일단 생략하고 결론이 무엇인가를 파악하자.
- 물론 사회통계학 자체를 전공한다면 수학 공부를 게을리해서는 안된다. 반드시 교재를 읽고 꼼꼼히 공부해야 한다.

연습문제

회귀 가정에 대한 코멘트

연습 1. hprice1.dta에서 price를 종속변수로, bdrms, lotsize, sqrft, colonial를 독립변수로 하는 회귀모형을 자료에 적합하시오. 회귀분석의 가정을 설명하고 각각의 성립 여부를 평가하시오.

회귀 가정에 대한 코멘트

연습 2. bwght.dta에서 bwght를 종속변수로, cigtax, cigprice, fatheduc, motheduc, parity, male, white, cigs, faminc를 독립변수로 하는 회귀모형을 자료에 적합하시오. 회귀분석의 가정을 설명하고 각각의 성립 여부를 평가하시오.