제 12 강

확률적 설명변수

Random Explanatory Variables

강 외생성(strict exgoneity)

- 설명변수 x가 확률변수가 아니라는 가정
  - 실험실에서의 통제된 자료와 같이 x를 변화시킴에 있어서, y에 영향을 미치는 다른 모든 요인들(ε)은 x의 변화와 무관함을 보증
    - 이 경우 우리는 x의 변화가 y에 대해 미치는 영향의 크기를 $\frac{\Delta E(y_t \mid x_t)}{\Delta x_t} = \beta_2$ 에서 확인할 수 있음
- 설명변수 x가 확률변수라고 가정해도
  - x의 변화가 y에 영향을 미치는 다른 모든 요인들(ε)과 독립적이라면 동일한 결과를 얻음
  - $E(\varepsilon_t \mid x_1, \ldots, x_T) = 0$ 를 충족하는 경우 설명변수 x는 강 외생적(strictly exogenous)라고 함
  - 이는 설명변수 x가 어떻게 변화하던 무관하게, y에 영향을 미치는 다른 요인들(ε)이 평균적으로 y에 미치는 영향은 항상 0임
    - 따라서 x의 변화는 ε 의 변화와 독립적임을 의미함

참확률변수인 설명변수를 고려한 새로운 가정

강 외생성(strict exgoneity)하의 가정 1

1. $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$, $t=1,\cdots,T$
2. (x_t, y_t) 는 확률표본(각 쌍은 동일하고 독립적인 분포)
3. (주어진 x_t에 대해) $E(\varepsilon_t | x_t) = 0$ (강 외생성)
4. $var(\varepsilon_t | x_t) = \sigma^2 = var(y_t | x_t)$
5. x 는 반드시 두 개의 다른 값을 가진다.
6. (optional) $\varepsilon_t | x_t \sim N(0,\sigma^2)$

확률변수인 설명변수를 고려한 새로운 가정

강 외생성(strict exgoneity)하의 가정 2

1. $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$, $t=1,\cdots,T$
2. (주어진 $x=(x_1,\cdots,x_T)$에 대해) $E(\varepsilon_t | x) = 0$ (강 외생성)
3. $Var(\varepsilon_t | x) = \sigma^2 = var(y_t | x)$
4. $cov(\varepsilon_i, \varepsilon_j | x) = cov(y_i, y_j | x) = 0$
5. x 는 반드시 두 개의 다른 값을 가진다.
5. (optional) $\varepsilon_t | x_t \sim N(0,\sigma^2)$

새로운 가정하에서 최소제곱추정량의 성질

소표본 성질 (finite sample properties)

- X가 확률변수일 경우 최소제곱 추정량의 소표본 성질은 다음과 같이 요약될 수 있음
  - 최소제곱 추정량은 선형 불변 추정량임
  - 1-5의 가정들 하에서 최소제곱 추정량은 모수에 대한 최우수 선형불편추정량(BLUE)이며 통상적인 오차항의 분산에 대한 추정량(잔차제곱합/자유도)은 불편임
  - 추가로 6의 가정하에서 최소제곱 추정량들의 x를 조건부로 하는 분포는 정규분포를 하게 되며, 그들의 분산들은 통상적인 방법으로 추정됨. 따라서 통상적 구간추정이나 가설검정 절차 역시 유효(valid)함

새로운 가정하에서 최소제곱추정량의 성질

대표본(점근적) 성질 (large sample (asymptotic) properties)

- $T \to \infty$ 인 경우 혹은 충분히 큰 표본을 가지고 있는 경우 근사적으로, 최소제곱추정량의 확률분포는 어떻게 되는가?
  - 표본의 크기가 점차 커짐에 따라 최소제곱추정량은 그 모수로 확률적으로 수렴함
  - 이러한 성질을 갖는 추정량은 일치(*consistent*) 추정량이라 하며, 일치성은 최소제곱 추정량의 대표본 성질임
    - 일치성은 어떤 추정량이 실제로 사용될 수 있기 위해서 갖추어야 하는 최소한의 요건이기도 함

완화된 가정의 도입과 최소제곱추정량의 성질

외생성(exogeneity)

- 완화된 가정: 3*. $E(\varepsilon_t)=0$ and $\text{cov}(x_t,\varepsilon_t)=0$

  강외생성 $\Rightarrow \text{cov}(\varepsilon_t,x_t)=0$ & $E(\varepsilon_t)=0$

  $\text{cov}(\varepsilon_t,x_t)=0$ & $E(\varepsilon_t)=0 \not\Rightarrow$ 강외생성

- 이 가정은 다음을 의미함
  - 중요한 설명변수가 누락되지 않음 & 올바른 함수형태를 사용하고 있음
  - 오차항이 설명변수와 상관되도록 하는 요인들이 존재하지 않음

  $\text{cov}(x_t,\varepsilon_t)=0$ → x는 외생적이라고 함

완화된 가정의 도입과 최소제곱추정량의 성질

외생성(exogeneity)

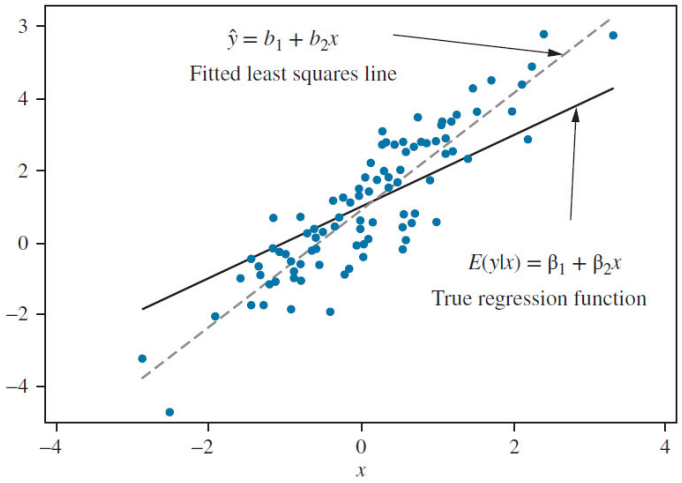
- 완화된 가정하에서, 최소제곱 추정량의 소표본 성질들은 더 이상 유효(valid)하지 않음
  - 더 이상 BLUE가 아님
- 하지만, 최소제곱 추정량의 대표본 성질은 여전히 유효함
  - 최소제곱추정량은 여전히 일치추정량임
  - 최소제곱추정량은 오차항의 조건부 분포가 정규분포든 아니든 간에 대표본 하에서 근사적으로 정규분포를 하게 됨
  - 대표본 하에서 구간추정과 가설검정은 여전히 유효함

내생적인 설명변수

내생성(endogeneity)

- 3* 이 충족될 수 없는 경우, 특히 $cov(x,\varepsilon)\neq 0$ 인 경우,
  - 최소제곱추정량은 비일치(inconsistent)추정량이며, 아무리 큰 표본 하에서도 모수로 수렴하지 않음
  - 통상적 가설검정이나 구간추정 절차 역시 유효(valid)하지 않음
- 설명변수들이 확률변수로 간주되는 경우, 최소제곱추정량이 적절한 것인가를 판단함에 있어서 설명변수들과 오차항간의 상관 여부가 핵심적인 문제임
  - $cov(x,\varepsilon)\neq 0$ 인 경우 설명변수는 내생적이라고 하며, 내생적 설명변수를 포함한 회귀식은 내생성의 문제가 있다고 함

내생성 하의 최소제곱 추정량의 비일치성



$$Cov(x_t,\varepsilon_t) > 0, \beta_2 > 0 \rightarrow \Delta y_t(+) = \beta_1 + \beta_2 \Delta x_t(+) + \Delta\varepsilon_t(+)$$

내생성 하의 최소제곱 추정량의 비일치성

$$y=\beta_1+\beta_2 x+e \;\Rightarrow\; E(y)=\beta_1+\beta_2 E(x) \;\because E(e)=0$$

$$\Rightarrow\; y-E(y)=\beta_2\left[x-E(x)\right]+e, \quad \text{양변에 } x-E(x) \text{ 를 곱함}$$

$$\Rightarrow\; \left[x-E(x)\right]\left[y-E(y)\right]=\beta_2\left[x-E(x)\right]^2+\left[x-E(x)\right]e$$

$$\Rightarrow\; E\left[x-E(x)\right]\left[y-E(y)\right]=\beta_2 E\left[x-E(x)\right]^2+E\left\{\left[x-E(x)\right]e\right\}$$

$$\Rightarrow\; \operatorname{cov}(x,y)=\beta_2 \operatorname{var}(x)+\operatorname{cov}(x,e)$$

$$\Rightarrow\; \beta_2=\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}-\frac{\operatorname{cov}(x,e)}{\operatorname{var}(x)}$$

내생성 하의 최소제곱 추정량의 비일치성

$$\operatorname{cov}(x,e)=0 \;\Rightarrow\; \beta_2=\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}$$

$$b_2=\frac{\sum(x_t-\bar{x})(y_t-\bar{y})/(T-1)}{\sum(x_t-\bar{x})^2/(T-1)}=\frac{\hat{\operatorname{cov}}(x,y)}{\hat{\operatorname{var}}(x)} \xrightarrow{p} \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}=\beta_2$$

$$\operatorname{cov}(x,e)\neq 0 \;\Rightarrow\; \beta_2=\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}-\frac{\operatorname{cov}(x,e)}{\operatorname{var}(x)}$$

$$b_2 \xrightarrow{p} \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}=\beta_2+\frac{\operatorname{cov}(x,e)}{\operatorname{var}(x)}\neq\beta_2$$

Bias: $\dfrac{\operatorname{cov}(x,e)}{\operatorname{var}(x)}=corr(x,e)\dfrac{\sigma_e}{\sigma_x}$

설명변수가 내생적인 경우

예: 설명변수의 누락(Omitted Variables)

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t \quad : \text{Correct Model } (\beta_3 \neq 0)$$

$$y_t = \beta_1 + \beta_2 x_{2t} + v_t \quad : \text{Incorrect Model}$$

단, $E(\varepsilon_t) = 0,\ Cov(x_{2t}, \varepsilon_t) = 0$

$$\Rightarrow v_t = \beta_3 x_{3t} + \varepsilon_t$$

$$\Rightarrow Cov(x_{2t}, v_t) = Cov(x_{2t}, \beta_3 x_{3t} + \varepsilon_t) = Cov(x_{2t}, \beta_3 x_{3t})$$

문제가 안되는 경우: $r_{23} = 0$

설명변수가 내생적인 경우

예: 변수오차(Errors in Variables)의 경우

- 오차가 있는 설명변수를 이용하여 추정하는 경우 이는 오차항과 상관되며, 따라서 최소제곱추정량은 비일치추정이 됨
  - 노동자의 임금이 노동자의 능력의 함수라 가정:

y_t = wages of the t'th worker

x_t^* = the ability of the t'th worker

$$y_t = \beta_1 + \beta_2 x_t^* + \varepsilon_t$$

  - 능력을 나타내는 변수(*로 표시)는 관측하기가 매우 어려움
  - 능력을 표준화된 시험성적(수능, SAT, TOEIC 등)으로 측정하고자 하며, 이를 x_t 로 표시.
  - 경우에 따라 이는 대용변수(**proxy variable**) **라 함**.

$$x_t = x_t^* + u_t$$

설명변수가 내생적인 경우

예: 변수오차(Errors in Variables)의 경우

- 여기서 u_t 는 확률적인 측정 오차이며 평균이 0이고 분산은 σ_u^2
- u_t 는 ε_t,와 상관되어 있지 않다고 가정함
- $x_t^* = x_t - u_t$ 를 추정식에 대입하면,

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_t^* + \varepsilon_t \\ &= \beta_1 + \beta_2 (x_t - u_t) + \varepsilon_t \\ &= \beta_1 + \beta_2 x_t + (\varepsilon_t - \beta_2 u_t) \\ &= \beta_1 + \beta_2 x_t + v_t \end{aligned}$$

- 이 추정식에서 설명변수 x_t 는 확률변수이며, 이 식에 최소제곱추정을 적용하게 되면 측정오차에 대한 가정으로부터 다음을 얻을 수 있으며, 앞서 설명한 바에 의하면 비일치추정을 하게 됨

$$\begin{aligned} \operatorname{cov}(x_t, v_t) &= E(x_t v_t) = E\left[(x_t^* + u_t)(\varepsilon_t - \beta_2 u_t)\right] \\ &= E(-\beta_2 u_t^2) = -\beta_2 \sigma_u^2 \neq 0 \end{aligned}$$

설명변수가 내생적인 경우

예: 동시성(Simultaneity)의 경우

- 설명변수와 종속변수가 모두 시스템 내에서 동시에 결정되는 내생변수인 경우, 설명변수는 오차항과 상관되는 확률변수이며 이 경우 최소제곱 추정은 적절하지 않음

$q_t = \beta_1 + \beta_2 p_t + \beta_3 y_t + \varepsilon_t$ 수요함수, y는 소득: 외생변수

$q_t = \alpha_1 + \alpha_2 p_t + \alpha_3 w_t + \nu_t$: 공급함수, w는 요소비용: 외생변수

q_t, p_t : 거래량과 가격은 모두 내생변수로서 시스템에서 결정됨

$$\Rightarrow p_t = \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2} + \frac{\beta_3 y_t}{\alpha_2 - \beta_2} - \frac{\alpha_3 w_t}{\alpha_2 - \beta_2} + \frac{\varepsilon_t - \nu_t}{\alpha_2 - \beta_2}$$

$$\Rightarrow \operatorname{cov}(p_t, \varepsilon_t) = E\left[\left(\frac{\varepsilon_t}{\alpha_2 - \beta_2}\right)\varepsilon_t\right] = E\left[\frac{\varepsilon_t^2}{\alpha_2 - \beta_2}\right] = \frac{\sigma^2}{\alpha_2 - \beta_2} \neq 0$$

- 회귀식의 설명변수가 내생적인 경우를 '내생성의 문제 혹은 동시성의 문제를 갖는다'라고 표현함

적률방법에 기초한 추정량

적률방법 (Method of moments)

- 어떤 확률변수의 k차 수학적 적률(k th mathematical moment)은 그 확률변수의 k승의 기대값임

$$E\left(Y^k\right)=\mu_k = k \text{ th mathematical moment of } Y$$

- 이 수학적 적률은 대응하는 k차 표본 적률(k th sample moment)에 의해 일치 추정될 수 있음

$$\hat{E}\left(Y^k\right)=\hat{\mu}_k = k\text{th sample moment of } Y$$
$$=\sum_{t=1}^{T} y_t^k \Big/ T$$

- 적률추정법은 m개의 모수들을 추정함에 있어서 m개의 수학적 적률들을 m개의 표본적률들에 등치시킴으로서 추정량을 구하는 방법임

$$E(Y)=\mu$$
$$\text{var}(Y)=\sigma^2=E(Y-\mu)^2=E\left(Y^2\right)-\mu^2$$

적률방법에 기초한 추정량

적률방법 (Method of moments)

- 두 개의 모수 μ 와 σ^2를 추정함에 있어서 두 개의 수학적 적률과두 개의 표본 적률을 등치시킴
- Y의 처음 두 수학적 적률과 그에 대응되는 표본 적률들:

$$E(Y)=\mu_1=\mu, \quad \hat{\mu}=\sum_{t=1}^{T} y_t \Big/ T$$
$$E\left(Y^2\right)=\mu_2, \quad \hat{\mu}_2=\sum_{t=1}^{T} y_t^2 \Big/ T$$

- μ 와 σ^2 에 대한 적률방법추정량은 다음과 같이 주어짐

$$\hat{\mu}=\sum_{t=1}^{T} y_t \Big/ T=\bar{y}$$
$$\tilde{\sigma}^2=\hat{E}\left(Y^2\right)-\hat{\mu}^2=\frac{\sum_{t=1}^{T} y_t^2}{T}-\bar{y}^2=\frac{\sum_{t=1}^{T} y_t^2-T\bar{y}^2}{T}=\frac{\sum(y_t-\bar{y})^2}{T}$$

- 일반적으로 적률방법추정량은 대표본에서 일치추정량이나 유효성과 관련해서는 아무런 보장을 할 수 없음

적률방법에 기초한 추정량

적률방법 - 단순회귀모형

- 선형회귀모형에 있어서 ($y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$,), 보통 다음을 가정함

  $E(\varepsilon_t) = 0 \;\Rightarrow\; E(y_t - \beta_1 - \beta_2 x_t) = 0$: Moment Condition 1

- x_t 확률변수가 아니거나 확률변수라해도 ε_t와 상관되지 않는다면

  $E(x_t \varepsilon_t) = 0 \;\Rightarrow\; E[x_t(y_t - \beta_1 - \beta_2 x_t)] = 0$: Moment Condition 2

- β_1 과 β_2 에 대한 적률방법 추정량은 다음과 같이 도출됨

  $$\frac{1}{T}\sum(y_t - b_1 - b_2 x_t) = 0, \quad \frac{1}{T}\sum x_t(y_t - b_1 - b_2 x_t) = 0$$

  – 이 두 방정식은 최소제곱추정량을 도출했던 두 정규방정식과 동일하며, 따라서 그 해는 최소제곱추정량이 됨

적률방법에 기초한 추정량

적률방법 – 도구(수단)변수

- 최소제곱추정의 문제는 x 가 확률변수이고 오차항 ε 과 상관되어 있는 경우 발생함 $E(x_t \varepsilon_t) \neq 0$
- 하지만 다음의 적률조건을 만족하는 다른 변수 z_t (도구변수,**instrumental variable**) 가 있을 경우

  $E(z_t \varepsilon_t) = 0 \;\Rightarrow\; E[z_t(y_t - \beta_1 - \beta_2 x_t)] = 0$

- 이 경우 두 개의 적률조건을 이용하여 β_1 과 β_2 에 대한 추정량을 얻을 수 있음

  $$\frac{1}{T}\sum(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t) = 0, \quad \frac{1}{T}\sum z_t(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t) = 0$$

- 이들 방정식들을 풀면 적률방법추정량을 얻을 수 있는데, 이는 대개 도구(수단)변수추정량(**instrumental variable estimators**) 이라 함

$$\hat{\beta}_2 = \frac{\sum z_t \sum y_t - T\sum z_t y_t}{\sum z_t \sum x_t - T\sum z_t x_t} = \frac{\sum(z_t - \bar{z})(y_t - \bar{y})}{\sum(z_t - \bar{z})(x_t - \bar{x})}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

적률방법에 기초한 추정량

도구변수추정량의 성질

- 도구변수추정량은 대표본에서 그 모수로 (확률적으로) 수렴함, 즉 일치 추정량임
- 대표본에서 도구변수추정량은 근사적으로 정규분포를 함

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum (x_t - \bar{x})^2 r_{zx}^2}\right)$$

  - r_{zx}^2 는 도구변수 z 와 확률변수 설명변수 x간의 표본 상관계수의 제곱
  - 도구변수추정량의 유효성을 높이기 위해서는 도구변수가 문제가 되는 설명변수와 높은 상관을 가지길 원하게 됨
- 오차항의 분산은 다음의 추정량으로 일치추정할 수 있음

$$\hat{\sigma}_{IV}^2 = \frac{\sum \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right)^2}{T-2}$$

적률방법에 기초한 추정량

2단계최소제곱추정량

- 필요한 것 이상의 도구변수들을 가지게 될 수 있음
  - 예컨대, w 가 x 와 상관되어 있으나 ε 과는 상관되어 있지 않은 또 하나의 도구변수라면
- 이 경우, 두 도구변수를 최적으로 이용한 추정량을 다음과 같은 두 단계의 절차를 통해 얻을 수 있다는 것이 알려져 있음

  (1)문제가 되는 x를 상수항과 z 및 w에 회귀하고 예측치 $\hat{x}$ 들을 구함

  (2) 예측치 $\hat{x}$ 를 x에 대한 도구변수로 이용하여 추정량을 구함

$$\sum \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right) = 0$$
$$\sum \hat{x}_t \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right) = 0 \quad \Rightarrow$$
$$\hat{\beta}_2 = \frac{\sum \left(\hat{x}_t - \bar{\hat{x}}\right)(y_t - \bar{y})}{\sum \left(\hat{x}_t - \bar{\hat{x}}\right)(x_t - \bar{x})} = \frac{\sum (\hat{x}_t - \bar{x})(y_t - \bar{y})}{\sum (\hat{x}_t - \bar{x})(x_t - \bar{x})} = \frac{\sum (\hat{x}_t - \bar{x})(y_t - \bar{y})}{\sum (\hat{x}_t - \bar{x})^2}$$
$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \qquad \because \bar{\hat{x}} = \bar{x},\ (\hat{x} - \bar{x})(x - \bar{x}) = (\hat{x} - \bar{x})^2$$

적률방법에 기초한 추정량

2단계최소제곱추정량

- 이러한 적률방법에 기초한 도구변수추정량은 2단계최소제곱추정량 (two-stage least squares estimators), 이라고도 하는데 이는 같은 추정량이 두 번의 최소제곱추정을 통해서도 얻어질 수 있기 때문임
  - Stage 1: x를 상수항, z 및 w, 에 대해 회귀하여 $\hat{x}$ 를 얻음
  - Stage 2: 다음의 선형회귀식에 대해 최소제곱추정을 적용

$$y_t = \beta_1 + \beta_2 \hat{x}_t + error_t$$

,

- 근사적으로 추정량 $\hat{\beta}_2$ 의 분산은 대표본 에서다음과 같이 계산됨

$$\text{var}\left(\hat{\beta}_2\right) \to \frac{\sigma^2}{\sum(\hat{x}_t - \bar{x})^2} = \frac{\sigma^2}{\sum(x_t - \bar{x})^2 r_{\hat{x}x}^2}$$ (다음 페이지 참조)

- 오차항의 분산에 대한 추정량은 다음과 같이 원래 모형에서의 잔차로부터 계산되어야만 함

$$\hat{\sigma}_{IV}^2 = \frac{\sum\left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right)^2}{T-2}$$

적률방법에 기초한 추정량

2단계최소제곱추정량

- $\frac{\sigma^2}{\sum(x_t - \bar{x})^2 r_{\hat{x}x}^2} = \frac{\sigma^2}{\sum(\hat{x}_t - \bar{x})}$ 는 다음과 같이 보일 수 있음

$$\frac{1}{r_{\hat{x}x}^2} = \frac{\sum(x_t - \bar{x})^2 \sum\left(\hat{x}_t - \bar{\hat{x}}\right)^2}{\left[\sum\left(\hat{x}_t - \bar{\hat{x}}\right)(x_t - \bar{x})\right]^2} = \frac{\sum(\hat{x}_t - \bar{x})^2 \sum(\hat{x}_t - \bar{x})^2}{\left[\sum(\hat{x}_t - \bar{x})(x_t - \bar{x})\right]^2}, \quad \left(\because \bar{x} = \bar{\hat{x}}\right)$$

$$\frac{\sigma^2}{\sum(x_t - \bar{x})^2 r_{\hat{x}x}^2} = \frac{\sigma^2 \sum(x_t - \bar{x})^2 \sum(\hat{x}_t - \bar{x})^2}{\sum(x_t - \bar{x})^2 \left[\sum(x_t - \bar{x})(\hat{x}_t - \bar{x})\right]^2} = \frac{\sigma^2 \sum(\hat{x}_t - \bar{x})^2}{\left[\sum(x_t - \bar{x})(\hat{x}_t - \bar{x})\right]^2}$$

$$\sum(\hat{x}_t - \bar{x})(x_t - \bar{x}) = \sum(\hat{x}_t - \bar{x})(\hat{x}_t + \hat{v} - \bar{x})$$
$$= \sum(\hat{x}_t - \bar{x})^2, \qquad \left(\because \sum(\hat{x}_t - \bar{x})\hat{v} = 0\right)$$

$$\frac{\sigma^2 \sum(\hat{x}_t - \bar{x})^2}{\left[\sum(x_t - \bar{x})(\hat{x}_t - \bar{x})\right]^2} = \frac{\sigma^2 \sum(\hat{x}_t - \bar{x})^2}{\left[\sum(\hat{x}_t - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum(\hat{x}_t - \bar{x})}$$

약한 도구(weak instrument)의 문제

weak instrument problem

- 약한 도구변수(weak instrument)의 문제
  - 많은 경우, Cov(z,ε)=0 는 확실히 보장되지 않으며, 만약 Cov(z, ε)≠0 이라면, β_2 에 대한 도구변수추정도 비일치이며 그 편차(bias)는 다음과 같이 주어짐

$$\frac{\mathrm{cov}(z,e)}{\mathrm{cov}(z,x)} = corr(z,e)\frac{\sigma_e}{\sigma_x}\frac{1}{corr(z,x)}$$

  - 만약z가 Corr(z,x)이 상당히 작은 약한 도구변수(weak instrument)라면 β_2 에 대한 도구변수(2SLS)추청은 OLS추정보다 더 큰 편차(bias)를 가질 수 있음

약한 도구(weak instrument)의 문제

weak instrument problem

- 약한 도구변수(weak instrument)의 문제
  - 더욱이 도구변수추정량의 분산은 OLS추정량의 분산보다 크므로, 이 경우, 이중으로 문제가 될 수 있음.

$$\mathrm{var}(b_2^{LS}) = \frac{\sigma^2}{\sum(x_t-\bar{x})^2} < \mathrm{var}(\hat{\beta}_2^{IV}) = \frac{\sigma^2}{\sum(x_t-\bar{x})^2 r_{zx}^2}$$

  - 결국 좋은 도구변수를 찾는 것이 중요하며, 그 보다 더 좋은 것은 좋은 데이터를 가지고 분석하는 것임
  - 좋은 데이터라고 하면 설명변수의 외생성에 대해 설득력 있는 근거를 가지고 있는 데이터이며, 실험실에서 만들어진 것과 같은 데이터의 성질을 갖는 데이터일 것임
    Ex) 자연적 실험(Natural Experimant) 데이터 : 쌍둥이 데이터

다중회귀모형에서의 2SLS 추정

- $y=\beta_1+\beta_2 x_2+\cdots+\beta_K x_K+\varepsilon$ 에서 x_K만 내생적인고 나머지 설명변수들은 외생적
  - 이 외생적 설명변수들은 모형 내 도구(included instruments)의 역할을 함
  - 외부의 도구(external instruments)들이 $z_1,\cdots z_L$로 L개가 주어짐
  - 1단계 : 내생적 설명변수를 종속변수로 모형 내 모든 외생적 설명변수와 외부의 도구변수를 독립변수로 하는 회귀식을 LS추정

$$x_K=\alpha_1+\alpha_2 x_2+\cdots+\alpha_{K-1}x_{K-1}+\gamma_1 z_1+\cdots+\gamma_L z_L+\nu_K$$

  - 2단계: 1단계로부터 얻어진 적합값으로 x_K를 대체후 LS추정

$$y=\beta_1+\beta_2 x_2+\cdots+\beta_K \hat{x}_K+\varepsilon^*$$

다중회귀모형에서의 2SLS 추정

- $y=\beta_1+\beta_2 x_2+\cdots+\beta_G x_G+\beta_{G+1}x_{G+1}+\cdots+\beta_K x_K+\varepsilon$ 에서 $x_2,...,x_G$ 는 외생적 설명변수이고 나머지 B=K-G개의 설명변수는 내생적인 경우 (보다 일반적인 경우)
  - 외부의 도구(external instruments)들은 $z_1,\cdots z_L$로 L개가 주어짐
  - 일치추정을 위한 필요조건은 L≥B임

    L=B: 정확 식별 (exact identification)
    L>B: 과도 식별 (over identification)
    L<B: 과소 식별 (under identification)

  - 1단계: B개의 내생적 설명변수를 각각 종속변수로 하고, G개의 외생적 설명변수들과 L개의 외부의 도구변수들을 독립변수로하는 B개의 식을 LS추정함
  - 2단계; 1단계 추정으로부터 얻어지는 B개의 내생적 설명변수에 대한 적합값으로 원래의 추정식의 내생적 설명변수들을 대체하여 LS 추정함

설명변수와 오차항 사이의 상관검정

$H_0 : Cov(x,\varepsilon) = 0 \quad H_1 : Cov(x,\varepsilon) \neq 0$

- 검정의 아이디어는 최소제곱추정량과 도구변수추정량을 비교하는 것임
  - 귀무가설이 참이라면 두 가지 추정량 모두 일치 추정량이 따라서,
    $$q = \left(b_{ols} - \hat{\beta}_{IV}\right) \to 0$$
  - 귀무가설이 참이 아니라면 최소제곱추정량은 일치추정량이 아니며, 도구변수추정량은 일치추정량이므로
    $$q = \left(b_{ols} - \hat{\beta}_{IV}\right) \to c \neq 0$$
- 이러한 원리에 바탕을 둔 검정방법들을 통상 하우스만 검정(Hausman Test)이라고 함
  - 쉽게 생각할 수 있는 것은 두 가치 추정치의 차이의 제곱을 적절히 가중하여 합한 값에 기반한 검정이지며, 통상 카이스퀘어 검정을 하게 됨

설명변수와 오차항 사이의 상관검정

데이비슨-매키논 (Davidson and MacKinnon) 검정

- 오차항과 상관되어 있다고 생각되는 각 변수마다 최소한 하나의 도구변수가 필요함
- z_{t1} 와 z_{t2} 가 x에 대한 도구변수들이라고 하면,

  $x_t = a_0 + a_1 z_{t1} + a_2 z_{t2} + v_t$ 를 최소제곱에 의해 추정하고 잔차를 얻음

  $$\Rightarrow \hat{v}_t = x_t - \hat{a}_0 + \hat{a}_1 z_{t1} + \hat{a}_2 z_{t2}$$

  - 만약 하나 이상의 설명변수가 의심이 된다면, 이러한 추정과정을 이용가능한 도구변수들을 사용하여 각 변수에 대해 반복함
  - 이렇게 얻어진 잔차들을 다음과 같이 회귀식에 설명변수로 포함시킴
    $$y_t = \beta_1 + \beta_2 x_t + \delta \hat{v}_t + \varepsilon_t$$
  - 이 인위적 회귀식(artificial regression)을 최소제곱에 의해 추정하고 통상의 t 검정을 이용해 유의성 검정을 수행함

    $H_0 : \delta = 0$ (no correlation between x and ε), $H_1 : \delta \neq 0$ (correlation between x and ε)
  - 하나 이상의 변수가 의심이 될 경우에는 포함된 잔차들의 모수들에 대한 결합적 유의성을 F검정을 통해 검정함

도구변수의 힘(strength) 검정

- 도구변수들의 힘은 결국 내생적 설명변수(들)의 (외생적)변화를 얼마나 (추가적으로) 잘 설명하는가에 의해서 결정됨
- 여러 개의 도구변수가 이용 가능한 경우 약한 도구의 문제를 피하기 위해 중요한 것은 도구변수들의 전체적 설명력임
- 모형 내에 내생적 설명변수가 1개인 경우 : 1단계 최소제곱추정을 바탕으로 한 F검정을 통해 이루어질 수 있음
  - $x_2,\cdots,x_{K-1}$ 가 내부도구, x_K 가 내생적 설명변수, $z_1,\cdots,z_L$: 도구변수
  - $x_K = \alpha_1 + \alpha_2 x_2 + \cdots + \alpha_{K-1}x_{K-1} + \gamma_1 z_1 + \cdots + \gamma_L z_L + \nu_K$: 1단계 최소제곱추정

    $H_0 : \gamma_1 = \ldots = \gamma_L = 0,\ H_1 : \text{o.w.}$

- Staiger and Stock (1997) 에 따르면 weak instrument의 문제를 피하기 위해서는 통상적인 유의수준 하에서의 판단만으로는 부족하며 F>10의 기준을 제시함
- 내생적 설명변수가 2개 이상인 경우에는, 좀 더 복잡한 방법이 요구됨 (Stock and Yogo (2005), beyond scope of this course!

과다식별 상황에서의 도구변수 외생성 검정

외생성 검정 (Sargan's test)

- 과다 식별의 경우 (L>B)에 한해 도구변수가 외생적인지 여부에 대한 검정의 수행이 가능함 (Sargan's test)
  - 즉 적어도 B개의 도구변수는 valid하다는 전제하에서 L-B개의 잉여 도구변수에 대한 validity를 검정하는 것
  - H_0 : L-B개의 잉여 도구 변수가 모두 외생적, H_1 : o.w.
- 우선 모든 L개의 도구변수 $z_1,\cdots z_L$를 이용하여 도구변수추정을 수행하고 그로부터 잔차 $\hat{\varepsilon}_{IV}$ 를 얻음
  - 이 잔차를 L개의 도구변수에 대해 회귀함 (auxilary regression)

    $$\hat{\varepsilon}_{IV} = \delta_0 + \delta_1 z_1 + \ldots + \delta_L z_L + \nu$$
  - 위 식에서 B개의 도구변수가 vailid하다는 전제하에 L개의 모수 중 B개의 모수는 0이며, 따라서 귀무가설은 나머지 L-B개의 모수가 0이라는 것이 됨
  - 위 식에서 얻어지는 R^2값에 관측치의 수를 곱하면 귀무가설 하에서 이는 자유도가 L-B인 카이스퀘어 분포를 함 (LM검정)
  - 우측검정을 통해 검정을 수행
  - 귀무가설이 기각 될 경우 어떤 도구변수의 validity가 문제되는지에 대해서는 별도의 분석을 수행해야 함

도구변수 추정의 예

교육에 대한 수익(returns to education) 추정

- 교육에 대한 투자는 얼마 만큼의 수익을 주는가?
  - 교육에 대한 투자의 수익률은 많은 나라들에 있어서 상당한 재정을 교육부문에 투자하고 있다는 점을 감안할 때 매우 중요한 지표임
    - 교육에 대해 돌아가는 충분히 수익률이 높다면, 사회적 자원을 교육부문에 추가로 투자하는 정책들이 뒷받침 될 수 있음
    - 교육에 대해 돌아가는 수익률이 하락한다면, 원인 진단을 통해, 보다 생산적인 교육이 이루어질 수 있도록 만드는 정책들이 뒷받침 될 수 있음
  - 뿐만 아니라 교육에 대한 높은 수익률은 개인들이 더 많은 교육을 받고자 하는 유인을 높임
    - 그만큼, 추가적 교육을 통해 유용한 것을 배우는 것을 의미함
  - 하지만, 이러한 교육의 수익을 정량적으로 측정하는 것은 쉽지 않음.

도구변수 추정의 예

교육에 대한 수익(returns to education) 추정

- 교육에 대한 수익 측정의 어려움
  - 단순한 접근 : $소득= \beta_0+\beta_1(교육)+u$
    - 측정오차(measurement error)? -소득, 교육에 대한 적절한 측정이 필요
    - 부족한 설명변수? - 경력 등 통제 변수들의 추가가 필요
    - 잘못된 함수관계? - 적절한 함수관계에 대한 검토가 필요
    - 내생성? - 교육 수준은 과연 소득에 영향을 미치는 여타 요인들(u)과 무관한가? → No!
  - 핵심 이슈 : 내생성
    - 내생성이 가장 극복하기 어려운 이슈이자 핵심이슈임
    - 교육과 오차항에 포함되어 있는 여러 가지 누락된 요인들(타고난 재능, 가정 환경 등등) 간에 상관관계가 존재함
      - 이러한 문제는 통제변수들을 추가하거나 패널자료를 사용하는 것만으로는 해결되지는 않을 수 있음.

도구변수 추정의 예

Angrist and Krueger, 1991

- 추정식
  - $Ln(\text{주당 소득}) = \beta_0+\beta_1(\text{교육년수})+\beta_2 x_2+\cdots+\beta_K x_K+u$
    - $x_2,\cdots, x_K$: 기타 통제변수들 (연령, 인종을 비롯한 여러 가지 더미변수들)
    - β_1: 1년의 추가 교육이 주당 소득을 몇 % 증가시키는지를 측정함 (교육에 대한 수익률)
    - 자료: 1980년도의 미국 인구총조사(population census)
- 도구변수: 태어난 분기(Quarter)
  - Angrist and Krueger (1991)은 미국 대부분 주에서 실시되었던 의무교육제도의 특별한 측면에 주목하고, 사람들의 태어난 분기를 도구변수로 사용
    - 미국 대부분 주에 있어서의 의무교육제도는 만 6세가 되는 해에 학교에 입학을 해야 함.

도구변수 추정의 예

Angrist and Krueger, 1991

- 이는 4분기에 태어난 아이들((예를 들어, 12월 31일 생)은 학교를 5.75세에 들어가는 반면에, 1분기에 태어난 아이들(예를 들어 1월 1일 생)은 6.75세에 들어감을 의미함
- 반면에 의무교육제도는 만 16세 생일까지 학생들이 의무적으로 학교를 다닐 것을 요구함
- 따라서, 서로 다른 분기에 태어난 사람들은 **평균적으로** 다른 교육 기간을 가지게 됨. 그리고 이러한 교육 기간의 차이는 외생적으로 결정되는 것으로 볼 수 있음
- 4분기에 태어난 사람들과 1분기에 태어난 사람들 간의 평균적 소득 수준의 차이에 영향을 줄 다른 요인들이 있는가? 지적능력?, 가정환경? - Not likely

도구변수 추정의 예

Angrist and Krueger, 1991

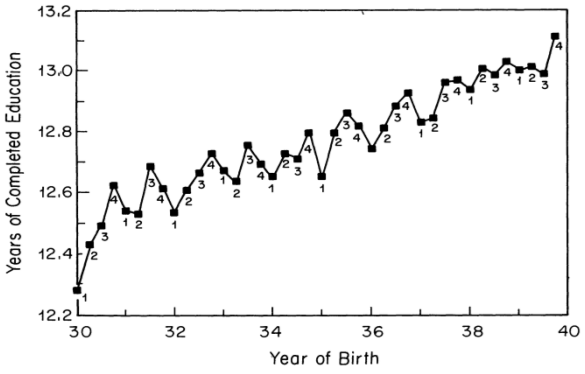
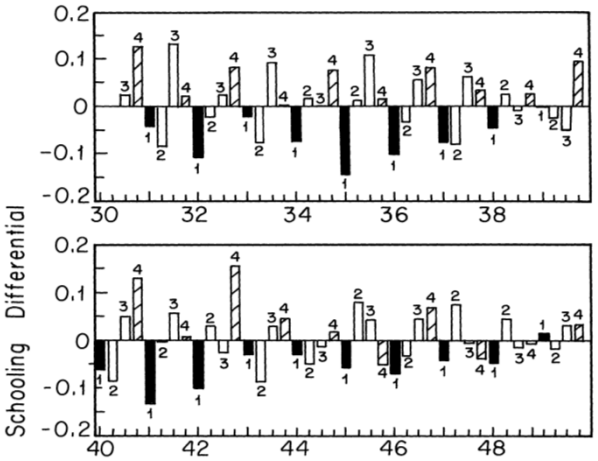


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

- 위 그림을 보면 예상한 바와 같이 같은 해에 태어나서 같은 해에 학교를 들어갔다고 해도, 1분기 태생이 4분기 태생에 비해 평균교육기간이 짧은 것을 확인할 수 있음

도구변수 추정의 예

Angrist and Krueger, 1991



- 위 그림은 앞서 그림에서 관측된 차이를 각 년도의 평균 대비 차이로 표시한 것임 – 1분기와 4분기 태생 간의 교육년수의 차이를 보다 확연히 볼 수 있음

도구변수 추정의 예

Angrist and Krueger, 1991

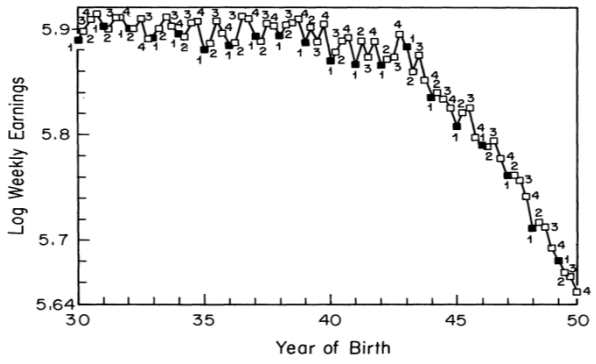


FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

- 위 그림은 태어난 분기에 따른 ln(주당소득)의 변화를 보여주고 있음
  - 이러한 차이는 태어난 분기에 따른 교육 수준의 외생적인 차이가 소득의 차이를 설명하는 것으로 볼 수 있음

도구변수 추정의 예

Angrist and Krueger, 1991

TABLE IV
OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1920–1929: 1970 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0802	0.0769	0.0802	0.1310	0.0701	0.0669	0.0701	0.1007
	(0.0004)	(0.0150)	(0.0004)	(0.0334)	(0.0004)	(0.0151)	(0.0004)	(0.0334)
Race (1 = black)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
SMSA (1 = center city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
Married (1 = married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region of residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
Age-squared	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
χ^2 [dof]	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

- 결과
  - 1분기에 태어난 사람들은 그 이후 분기에 태어난 사람들과 비교할 때, 평균적으로 교육년수가 약 0.1년 짧음.
  - 1분기에 태어난 사람들은 그 이후 분기에 태어난 사람들과 비교할 때 약 1% 소득이 적음
  - 1년의 추가 교육은 10%의 수익률로 추정됨(← (식 (8)의 추정 결과)