Principles of Econometrics (3e)

Ch. 10 확률 설명변수와 적률에 기초한 추정법

2013년 1학기

윤성민

10.0 서론

The assumptions of the simple linear regression are:

- SR1. $y_i = \beta_1 + \beta_2 x_i + e_i \quad i = 1, \ldots, N$
- SR2. $E(e_i) = 0$
- SR3. $\text{var}(e_i) = \sigma^2$
- SR4. $\text{cov}(e_i, e_j) = 0$
- SR5. The variable x_i is not random, and it must take at least two different values.
- SR6. (optional) $e_i \sim N(0, \sigma^2)$

✓ In this chapter, we relax the assumption that variable *x* is not random.

- **설명변수 x는 확률적인가 아닌가?**

- 경제학자가 사용하는 통계는 대부분 비실험적 성격의 통계

  $\Rightarrow$ x와 y의 값은 보통 동시에 알려짐

  $\Rightarrow$ 따라서 x와 y는 모두 확률변수라고 볼 수 있음

- 지금까지(1~9장) 설명변수가 확률적이 아니라고 생각한 이유

  - 통제된 실험에 의한 자료를 분석하는 경우에 타당한 가정임
  - OLS와 관련된 대수학을 간단하게 처리할 수 있음
  - x가 확률적이더라도 OLS를 사용할 수 있는 경우가 있음

▪ 10장의 주요 내용

- 이제부터는 설명변수가 확률적이라고 가정함

✓ 어떤 경우(혹은 조건하에서) 여전히 OLS를 사용해도 좋은가?

$\Rightarrow \quad E(e|x)=0$, 혹은 $\text{cov}(x,e)=0$

✓ 어떤 경우에 OLS는 적절한 추정법이 되지 못하는가?

$\Rightarrow \quad E(e|x)\neq 0$, 혹은 $\text{cov}(x,e)\neq 0$

✓ 설명변수가 확률적인 경우에 적절한 추정법은 무엇인가?

⇒ 적률추정법 **Method of Moments Estimation**
(MM 혹은 GMM)

10.1 x가 확률변수인 경우의 OLS

- 이 장에서는 가정을 다음과 같이 수정하기로 함

- A10.1 $y_i = \beta_1 + \beta_2 x_i + e_i$ 는 모집단에서 y와 x의 관계를 적절하게 나타냄(선형모형 가정의 타당성)
- A10.2 (x_i, y_i) 자료들은 무작위 표본추출(random sampling)을 통해 수집되었음. 즉, $(x_i, y_i) \sim iid$ (x_i는 확률변수라고 가정)
- A10.3 $E(e_i \mid x_i) = 0 \Leftrightarrow \mathrm{cov}(x_i, e_i) = 0$
- A10.4 x_i는 적어도 2개의 상이한 값을 가짐
- A10.5 $\mathrm{var}(e_i \mid x_i) = \sigma^2$
- A10.6 $e_i \mid x_i \sim N(0, \sigma^2)$

(1) OLS 추정량의 소표본 특성

- $E(e|x)=0$, 혹은 $\text{cov}(x,e)=0$ 이면

⇒ OLS를 적용해도 아무 문제 없음

즉, 설명변수가 확률변수이더라도

위 경우라면 OLS 추정량은 BLUE

- Gauss-Markov 정리에 의해 유한표본(finite sample) 혹은 소규모표본(small sample)에서도 성립함

- 자료가 무작위 표본추출법을 통하여 구해졌다면,

⇒ OLS 추정법은 표본의 크기와 무관하게 사용할 수 있음

(2) OLS 추정량의 점근적 특성: x가 확률적이 아닌 경우

(Asymptotic (Large) Sample Properties of the Least Squares Estimator)

- 대규모표본이면, 오차항이 정규분포하든 하지 않든 다음의 두 가지 특성이 존재함.

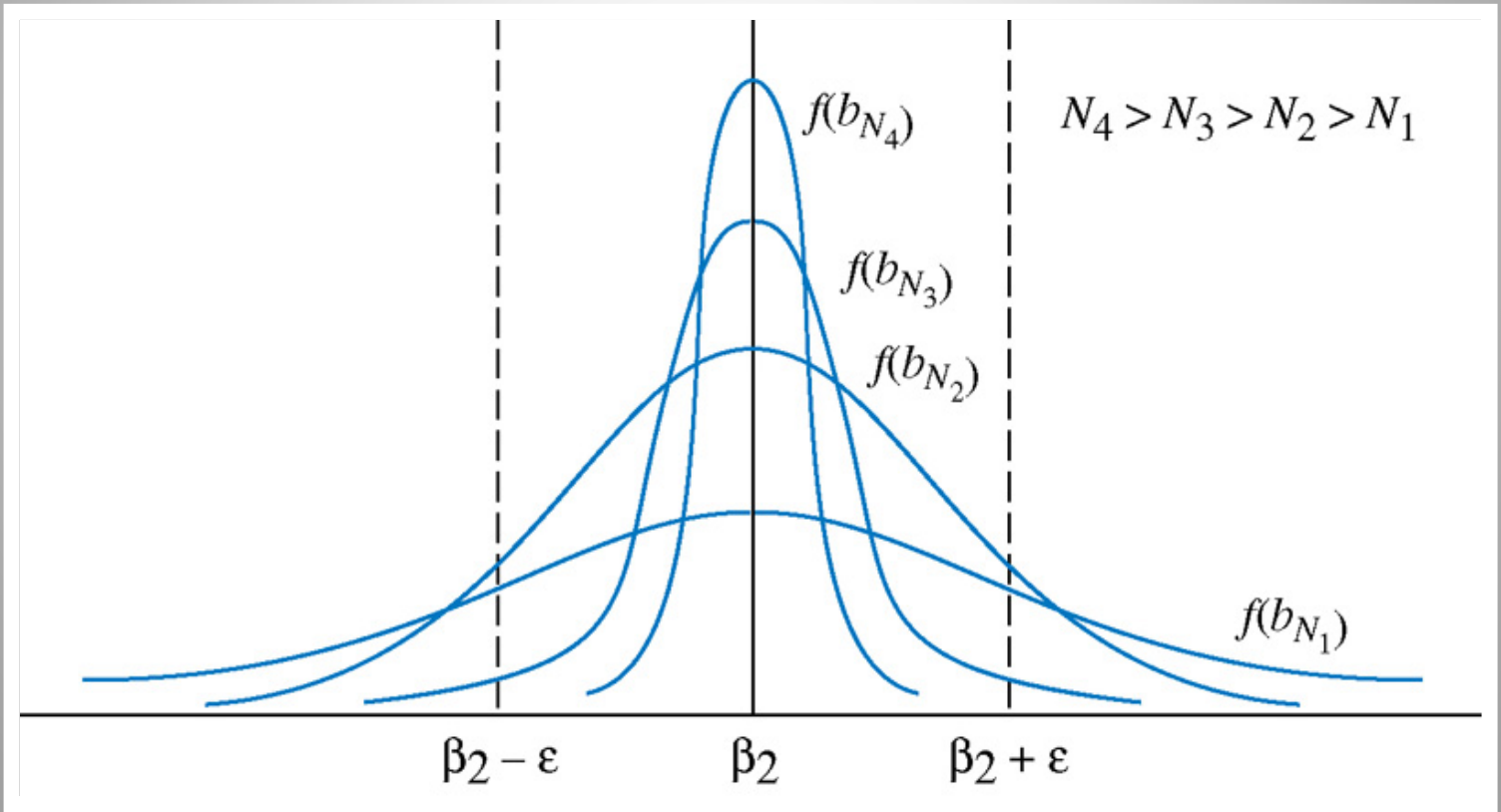
(a) OLS 추정량은 불편추정량이다.

(b) OLS 추정량의 분산은 0으로 수렴한다.

(일치추정량, *consistent* estimator)

❖ 일치성 (consistency of estimator)

- 대규모표본에서 나타나는 특성, 점근적 특성
- 표본의 크기가 증가하면 추정치는 모수에 근접함을 의미



(3) OLS 추정량의 점근적 특성: x가 확률적인 경우

: $\text{cov}(x,e)=0$ 의 경우

(a) 대표본이면 OLS 추정량은 일치추정량이다

(추정치는 모수에 근접 혹은 수렴함)

(b) 대표본이면 OLS 추정량은 정규분포한다

(통상적인 구간추정, 가설검정 타당함)

(4) OLS 추정량의 점근적 특성: x가 확률적인 경우

: $\text{cov}(x,e) \neq 0$ 의 경우

- 대표본이더라도 OLS 추정량은 일치추정량이 아니다

  ⇒ OLS 추정치는 모수에 수렴하지 않음

  통상적인 구간추정이나 가설검정 타당하지 않음

➢ 요약

$\text{cov}(x,e) = 0$ 이면, 설명변수가 확률적이더라도 OLS 적용 가능

$\text{cov}(x,e) \neq 0$ 이면, OLS는 적절한 추정방법이 아님(대표본 경우에도)

cf. 검정방법: Hausman test (10.4)

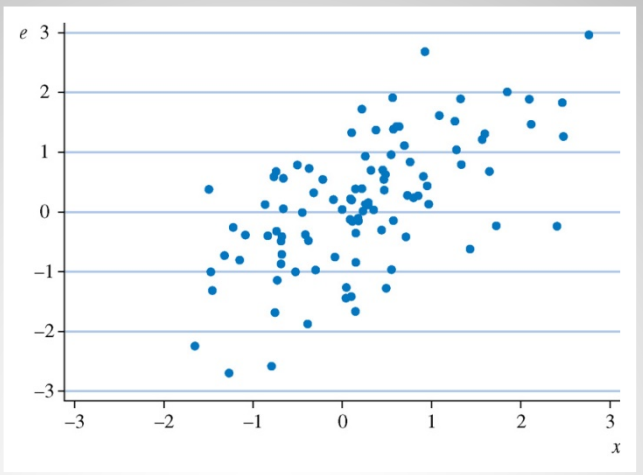
❖ $\mathrm{cov}(x,e)\neq 0$ 의 경우, OLS가 일치추정량이 아닌 이유

: Monte Carlo 모의분석 기법을 이용한 설명

- $\mathrm{cov}(x,e)>0$인 가상적인 x와 e 통계자료 생성 (상관계수 0.6)
  ⇒ 그림 10.2
- 모회귀선이 $E(y)=\beta_1+\beta_2 x=1+1\times x$ 라고 가정하자
  즉, 모수의 참값은 $\beta_1=1$ $\beta_2=1$ 이라고 가정
- 다음과 같이 인위적인 y값 생성 $y=E(y)+e=\beta_1+\beta_2 x+e=1+1\times x+e$
- 인위적으로 생성된 (y,x) 자료를 이용, OLS로 추정 ⇒ 그림 10.3

➢ 기울기가 체계적으로 과대평가, 대표본인 경우에도 과대평가
⇒ OLS 추정량은 일치추정량 아님

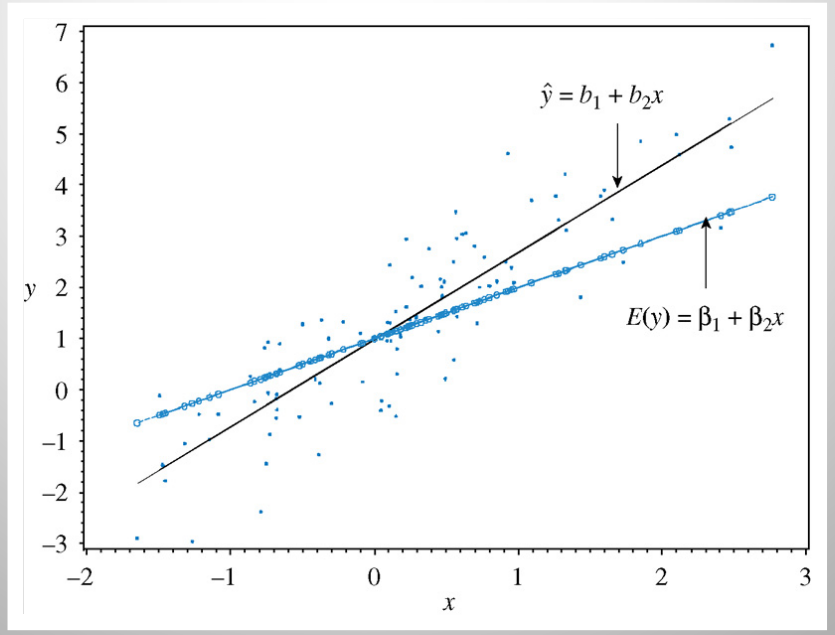
- $\mathrm{cov}(x,e)>0$인 가상적인 x 와 e 와 통계



$$y=1+1\times x+e$$

- OLS 추정

⇒ 체계적인 오류



- $\mathrm{cov}(x,e)>0$인 가상적인 x 와 e 와 통계 $y=1+1\times x+e$

Dependent Variable: Y
Method: Least Squares
Date: 09/19/12 Time: 10:43
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.978893	0.088281	11.08838	0.0000
X	1.703431	0.089950	18.93754	0.0000

R-squared	0.785385	Mean dependent var	1.386287
Adjusted R-squared	0.783195	S.D. dependent var	1.838819
S.E. of regression	0.856198	Akaike info criterion	2.547166
Sum squared resid	71.84128	Schwarz criterion	2.599270
Log likelihood	-125.3583	Hannan-Quinn criter.	2.568253
F-statistic	358.6306	Durbin-Watson stat	2.103601
Prob(F-statistic)	0.000000		

- OLS 추정 $\Rightarrow$ 체계적인 오류

- $\mathrm{cov}(x,e)>0$인 가상적인 x 와 e 와 통계 $y=1+1\times x+e$

Dependent Variable: Y
Method: Generalized Method of Moments
Date: 09/19/12 Time: 17:46
Sample: 1 100
Included observations: 100
Linear estimation with 1 weight update
Estimation weighting matrix: HAC (Bartlett kernel, Newey-West fixed bandwidth = 5.0000)
Standard errors & covariance computed using estimation weighting matrix
Instrument specification: Z1 Z2
Constant added to instrument list

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.166695	0.106393	10.96593	0.0000
X	1.061392	0.190278	5.578105	0.0000

R-squared	0.673462	Mean dependent var	1.386287
Adjusted R-squared	0.670130	S.D. dependent var	1.838819
S.E. of regression	1.056113	Sum squared resid	109.3068
Durbin-Watson stat	1.969202	J-statistic	3.522889
Instrument rank	3	Prob(J-statistic)	0.060527

- GMM 추정 ⇒ 체계적인 오류가 나타나지 않음

- $\mathrm{cov}(x,e)>0$인 가상적인 x 와 e 와 통계 $y=1+1\times x+e$

Dependent Variable: Y
Method: Two-Stage Least Squares
Date: 09/19/12 Time: 17:48
Sample: 1 100
Included observations: 100
Instrument specification: Z1 Z2
Constant added to instrument list

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.137591	0.116444	9.769431	0.0000
X	1.039872	0.194223	5.354022	0.0000

R-squared	0.666207	Mean dependent var	1.386287
Adjusted R-squared	0.662801	S.D. dependent var	1.838819
S.E. of regression	1.067780	Sum squared resid	111.7351
F-statistic	28.66555	Durbin-Watson stat	1.967390
Prob(F-statistic)	0.000001	Second-Stage SSR	302.0610
J-statistic	3.555022	Instrument rank	3
Prob(J-statistic)	0.059366		

- 2SLS 추정 ⇒ 체계적인 오류 나타나지 않음

10.2 x와 e가 상관된 경우 $\operatorname{cov}(x,e) \neq 0$

- 측정오차(Measurement Error) 혹은 변수오차(Errors-in-variables) 문제
- 누락변수(Omitted Variables)
- 연립방정식 편의(Simultaneous Equations Bias)
- 계열상관이 존재하는 시차종속변수 모형(Lagged Dependent Variable Models with Serial Correlation)

➢ 공통점

- 설명변수가 내생적으로 결정됨
- 내생성 문제(endogeneity problem)가 존재한다고 말함

(1) 측정오차 혹은 변수오차 문제

- 설명변수를 측정할 때 오차가 있는 경우

(예) 저축함수 추정할 때, 저축(y)은 항상소득에 의존한다고 가정

$$y_i = \beta_1 + \beta_2 x_i^* + v_i$$

- 항상소득(x^*)의 대리변수(proxy variable)로 현재소득(x) 사용

$$x_i = x_i^* + u_i$$

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i^* + v_i \\ &= \beta_1 + \beta_2 (x_i - u_i) + v_i \\ &= \beta_1 + \beta_2 x_i + (v_i - \beta_2 u_i) \\ &= \beta_1 + \beta_2 x_i + e_i \end{aligned}$$

$$\begin{aligned} \mathrm{cov}(x_i, e_i) &= E(x_i e_i) = E\left[(x_i^* + u_i)(v_i - \beta_2 u_i)\right] \\ &= E(-\beta_2 u_i^2) = -\beta_2 \sigma_u^2 \neq 0 \end{aligned}$$

- $\mathrm{cov}(x, e) \neq 0$ 따라서 측정오차 경우 OLS는 부적절한 추정방법임

- **설명변수 측정오차 문제의 사례 (**교과서 p. 381), 저축=f(현재소득)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22074	0.22074	0.22	0.6444
Error	48	49.09998	1.02292		
Corrected Total	49	49.32072			

Root MSE	1.01139	R-Square	0.0045
Dependent Mean	3.95066	Adj R-Sq	-0.0163
Coeff Var	25.60061		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.34277	0.85612	5.07	<.0001
x	1	-0.00519	0.01116	-0.46	0.6444

(2) 누락변수

- 누락된 변수가 포함된 설명변수와 상관된 경우에는 $\text{cov}(x, e) \neq 0$

(예) 임금함수

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + e_i$$

- $EDUC$ 는 교육받은 연수

- 누락변수 (Omitted Variables) : 경험, 능력, 열의(동기) 등등
  - 누락된 요인들의 영향은 오차항에 포함됨
  - 누락변수들은 교육연수와 밀접히 관련되어 있음

  $\Rightarrow \text{cov}(EDUC_i, e_i) \neq 0$

- OLS는 부적절 (불편추정량 아님, 일치추정량 아님)

(3) 연립방정식 편의

(예) 수요-공급모형

- 경쟁시장에서 균형가격(P_i)과 균형(수급)량(Q_i)은
  동시에 결정되므로 모두 내생변수임

- 이를 무시하고 (가격을 외생변수로 가정하고),
  아래의 단일방정식을 OLS로 추정하면, 내생성 문제가 발생함

$$Q_i = \beta_1 + \beta_2 P_i + e_i$$

- 내생성을 고려하면 $\text{cov}(P_i, e_i) \neq 0$

- OLS는 부적절 (불편추정량 아님, 일치추정량 아님)
  $\Rightarrow$ 연립방정식 편의(simultaneous equations bias)라고 함

(4) 계열상관이 존재하는 시차종속변수 모형

- 시차종속변수가 설명변수로 포함된 모형, 예를 들어

$$y_t = \beta_1 + \beta_2 y_{t-1} + \beta_3 x_t + e_t$$

- 이 모형의 오차항이 시계열적으로 상관되는 경우, 예를 들어

$$\text{AR(1) process: } e_t = \rho e_{t-1} + v_t$$

- 만약 $\rho \neq 0$ 이라면, $y_{t\text{-}1} \sim e_{t\text{-}1} \sim e_t$ 셋 모두 관련됨

$\Rightarrow \ \text{cov}(y_{t-1}, e_t) \neq 0$

$\Rightarrow$ OLS는 부적절 (불편추정량 아님, 일치추정량 아님)

10.3 적률방법에 기초한 추정량

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

- $\text{cov}(x, e) \neq 0$ 이면,
  OLS 추정량은 불편추정량도 일치추정량도 아님, 대안 필요함

- 적률방법 추정량

  $\text{cov}(x, e) = 0$ 이면 $\Rightarrow$ OLS 추정량과 동일해 짐

  $\text{cov}(x, e) \neq 0$ 이면 $\Rightarrow$ 수단변수 추정법, 2SLS (2단계 OLS)와 동일

▪ 적률 추정방법

- 모집단에서의 적률(moment)

- 확률변수 Y의 k번째 **모집단 적률**은 다음과 같이 정의됨

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

- 표본에서의 적률

- k번째 **표본적률**은 다음과 같이 정의됨

$$\widehat{E\left(Y^k\right)} = \hat{\mu}_k = k^{\text{th}} \text{ sample moment of } Y$$

$$= \sum y_i^k \Big/ N$$

- k번째 모집단 적률은 k번째 표본적률로 추정 가능 (일치추정량)

▪ 적률방법에서의 추정절차

- m번째 모집단 적률을 m번째 표본적률과 같다고 놓고 m개의 미지의 모수를 추정함

- 즉 $\mu(=\mu_1)$를 $\hat{\mu}(=\hat{\mu}_1)$, μ_2를 $\hat{\mu}_2$로, 이렇게 각각의 적률을 추정함

Population Moments $\Leftarrow$ Sample Moments

$$E(Y)=\mu_1=\mu \qquad \hat{\mu}=\sum y_i/N$$

$$E(Y^2)=\mu_2 \qquad \hat{\mu}_2=\sum y_i^2/N$$

▪ 적률방법에 의한 평균 및 분산의 추정방법

- 모집단 평균의 추정량

$$\hat{\mu} = \sum y_i / N = \bar{y}$$

- 모집단 분산의 추정량 (불편추정량과 약간 다르나, 일치추정량)

$$\tilde{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{\sum y_i^2 - N\bar{y}^2}{N} = \frac{\sum (y_i - \bar{y})^2}{N}$$

- **적률방법에 의한 회귀모형의 추정방법**

  : $\mathrm{cov}(x,e)=0$ 인 경우

$$y_t=\beta_1+\beta_2 x_t+e_t$$

- 적률조건(moment conditions)

$$E(e_i)=0 \quad \Rightarrow \quad E(y_i-\beta_1-\beta_2 x_i)=0$$

$$E(x_i e_i)=0 \quad \Rightarrow \quad E\left[x_i(y_i-\beta_1-\beta_2 x_i)\right]=0$$

- 두 개의 모집단 적률을 표본적률로 대체시키면,

  β_1,β_2 의 적률방법 추정량 b_1,b_2 를 아래와 같이 구할 수 있음

$$\frac{1}{N}\sum(y_i-b_1-b_2x_i)=0$$
$$\frac{1}{N}\sum x_i(y_i-b_1-b_2x_i)=0$$
$$\Rightarrow \quad b_2=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}, \quad b_1=\bar{y}-b_2\bar{x} \quad \Rightarrow \text{ OLS와 동일}$$

- **적률방법에 의한 회귀모형의 추정방법**

  : $\text{cov}(x,e) \neq 0$ 인 경우

- 두 번째 적률조건을 사용할 수 없게 됨, 왜냐하면 $E(x_t e_t) \neq 0$
- 그렇지만 두 번째 적률조건을 만족시키는 또 다른 z 가 있을 수 있음 (이 변수를 **수단변수(instrumental variable)**라 함)

$$E(z_i e_i) = 0 \quad \Rightarrow \quad E\left[z_i\left(y_i - \beta_1 - \beta_2 x_i\right)\right] = 0$$

- 그러면 다음 표본적률조건으로 β_1, β_2 의 추정량을 구할 수 있음

$$\begin{aligned} &\frac{1}{N}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0 \\ &\frac{1}{N}\sum z_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right) = 0 \end{aligned} \quad \Rightarrow \quad \begin{aligned} \hat{\beta}_2 &= \frac{N\sum z_i y_i - \sum z_i \sum y_i}{N\sum z_i x_i - \sum z_i \sum x_i} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\ \hat{\beta}_1 &= \bar{y} - \hat{\beta}_2 \bar{x} \end{aligned}$$

- 이 적률방법 추정량을 **수단변수 추정량**이라고 함

▪ 수단변수 추정량의 특성

- $\mathrm{cov}(z,e)=0$ 이면, 수단변수 추정량은 일치추정량임
- 대표본인 경우, 수단변수 추정량은 정규분포를 함

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

- 수단변수 추정량의 효율성을 높이기 위해서는

  x와 긴밀히 상관된 변수를 수단변수(z)로 선택해야 함

  (이유) r_{zx}^2 이 커질수록 수단변수 추정량의 분산이 작아짐
- 오차항 및 추정량의 분산은 다음과 같이 계산하면 됨

$$\hat{\sigma}_{IV}^2 = \frac{\sum\left(y_i-\hat{\beta}_1-\hat{\beta}_2 x_i\right)^2}{N-2} \qquad \mathrm{var}\left(\hat{\beta}_2\right) = \frac{\sigma^2}{r_{zx}^2\sum(x_i-\bar{x})^2} = \frac{\mathrm{var}(b_2)}{r_{zx}^2}$$

(예1) 수단변수 추정법을 이용한 저축함수 추정

- 저축 = f(항상소득)
- 앞에서 항상소득의 대리변수로 '현재소득'(x)을 이용하여 저축함수를 OLS로 추정하면, 측정오차 혹은 변수오차의 문제가 있음을 보았음

⇒ OLS 추정결과는 불편추정량도 일치추정량도 아님

$$\hat{S}_{OLS} = 4.3428 - 0.0052x$$
$$(\text{se}) \quad (0.856) \quad (0.001)$$

- 이제 항상소득의 수단변수로 '10년간의 평균소득'(z)을 이용해 보자.
  이 수단변수는 항상소득(장기평균소득)을 더 잘 대변함
  이 수단변수는 항상소득과도 상관되며, 현재소득과도 상관됨

$$\hat{S}_{OLS} = 0.9883 + 0.0392z$$
$$(\text{se}) \quad (1.524) \quad (0.020)$$

- **SAS program**

<OLS / Linear model>

```
proc reg ;
model  y = x ;
```

<OLS / Non-linear model>

```
proc model ;
parms b1 b2 ;
y = b1 + b2*x ;
fit y ;
```

<수단변수 추정법>

```
proc model ;
parms b1 b2 ;
exogenous z ;
y = b1 + b2*x ;
fit y / 2sls ;
```

*** instruments;**

- PROC REG, OLS, 저축=f(현재소득)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.22074	0.22074	0.22	0.6444
Error	48	49.09998	1.02292		
Corrected Total	49	49.32072			

Root MSE	1.01139	R-Square	0.0045
Dependent Mean	3.95066	Adj R-Sq	-0.0163
Coeff Var	25.60061		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.34277	0.85612	5.07	<.0001
Income	1	-0.00519	0.01116	-0.46	0.6444

- 수단변수 추정법: 2SLS

Nonlinear 2SLS Summary of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
Savings	2	48	65.2557	1.3595	1.1660	-0.3231	-0.3507

Nonlinear 2SLS Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b1	0.988267	1.5242	0.65	0.5198
b2	0.039176	0.0200	1.96	0.0564

- OLS, 저축=f(현재소득) savings.wfl

Dependent Variable: SAVINGS
Method: Least Squares
Date: 09/17/12 Time: 23:48
Sample: 1 50
Included observations: 50

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.342769	0.856116	5.072643	0.0000
INCOME	-0.005185	0.011163	-0.464539	0.6444

R-squared	0.004476	Mean dependent var	3.950660
Adjusted R-squared	-0.016264	S.D. dependent var	1.003267
S.E. of regression	1.011393	Akaike info criterion	2.899713
Sum squared resid	49.09998	Schwarz criterion	2.976194
Log likelihood	-70.49282	Hannan-Quinn criter.	2.928837
F-statistic	0.215796	Durbin-Watson stat	1.856169
Prob(F-statistic)	0.644362		

✓ 상식밖의 결과

- 수단변수 추정법: 2SLS

Dependent Variable: SAVINGS
Method: Two-Stage Least Squares
Date: 09/18/12 Time: 00:07
Sample: 1 50
Included observations: 50
Instrument specification: AVERAGE_INCOME
Constant added to instrument list

<Quick>-<Estimate Equation> 클릭,
<Method>에서 <TSLS> 선택,
윗칸: savings c income
아랫칸: average_income

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.988267	1.524164	0.648400	0.5198
INCOME	0.039176	0.020038	1.955094	0.0564

R-squared	-0.323089	Mean dependent var	3.950660
Adjusted R-squared	-0.350653	S.D. dependent var	1.003267
S.E. of regression	1.165973	Sum squared resid	65.25569
F-statistic	3.822391	Durbin-Watson stat	2.165285
Prob(F-statistic)	0.056408	Second-Stage SSR	44.12420
J-statistic	0.000000	Instrument rank	2

(예2) 임금함수 추정 – OLS 추정

- 기혼여성의 임금 = f(교육연수, 경험연수)

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$$

- 아래 OLS 추정결과는 문제가 있을 수 있음

$$\ln(WAGE) = -.5220 + .1075 \times EDUC + .0416 \times EXPER - .0008 \times EXPER^2$$

(se) (.1986) (.0141) (.0132) (.0004)

- 누락변수의 문제가 있을 가능성이 높음
  - "능력"은 임금에 영향을 미치지만 누락변수여서 오차항에 반영됨
  - "능력"이 있는 사람은 교육연수가 길 수 있음

$\Rightarrow \quad \mathrm{cov}(EDUC, e) \neq 0$

- OLS 추정결과는 불편추정량도 일치추정량도 아님

- OLS 추정결과 $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$

Dependent Variable: LNWAGE
Method: Least Squares
Date: 09/18/12 Time: 13:20
Sample: 1 753 IF WAGE>0
Included observations: 428

교육을 1년 더 받으면, 임금이 10.7% 증가한다는 의미,
추청치의 크기가 상식 밖으로 너무 큰 값으로 나타남

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.522041	0.198632	-2.628179	0.0089
EDUC	0.107490	0.014146	7.598332	0.0000
EXPER	0.041567	0.013175	3.154906	0.0017
EXPER2	-0.000811	0.000393	-2.062834	0.0397

R-squared	0.156820	Mean dependent var	1.190173
Adjusted R-squared	0.150854	S.D. dependent var	0.723198
S.E. of regression	0.666420	Akaike info criterion	2.035509
Sum squared resid	188.3051	Schwarz criterion	2.073445
Log likelihood	-431.5990	Hannan-Quinn criter.	2.050492
F-statistic	26.28615	Durbin-Watson stat	1.960988
Prob(F-statistic)	0.000000		

(예2) 임금함수 추정 - 수단변수 추정법을 이용

- 딸의 교육연수는 엄마의 교육연수와 밀접히 관련되어 있을 수 있음
- 이 관계를 고려하여 연립방정식 구성하여 2SLS로 추정하는 방법 있음

(1) 아래 식을 OLS로 추정

$$EDUC = 9.7751 + .0489 \times EXPER - .0013 \times EXPER^2 + .2677 \times MOTHEREDUC$$
$$\text{(se)} \quad (.4249) \ (.0417) \qquad (.0012) \qquad (.0311)$$

(2) 임금함수 추정에 $EDUC$ 대신 그것의 추정치 $\widehat{EDUC}$ 를 이용

$$\ln(WAGE) = .1982 + .0493 \times \widehat{EDUC} + .0449 \times EXPER - .0009 \times EXPER^2$$
$$\text{(se)} \quad (.4729) \ (.0374) \qquad (.0136) \qquad (.0004)$$

- $\text{cov}\left(\widehat{EDUC}, e\right) = 0$ 이므로 OLS로 추정 가능함
  - $\widehat{EDUC}$ 는 수단변수의 역할 (유의성 낮음, 불충분한 개선)

- 2SLS 추정결과 $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$

<Quick>-<Estimate Equation> 클릭,
<Method>에서 <TSLS> 선택,
윗칸: lnwage c educ exper exper2
아랫칸: exper exper2 mothereduc
Sample: 1 753 if wage>0

Dependent Variable: LNWAGE
Method: Two-Stage Least Squares
Date: 09/18/12 Time: 13:39
Sample: 1 753 IF WAGE>0
Included observations: 428
Instrument specification: EXPER EXPER2 MOTHEREDUC
Constant added to instrument list

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.198186	0.472877	0.419107	0.6754
EDUC	0.049263	0.037436	1.315924	0.1889
EXPER	0.044856	0.013577	3.303856	0.0010
EXPER2	-0.000922	0.000406	-2.268993	0.0238

R-squared	0.123130	Mean dependent var	1.190173
Adjusted R-squared	0.116926	S.D. dependent var	0.723198
S.E. of regression	0.679604	Sum squared resid	195.8291
F-statistic	7.347957	Durbin-Watson stat	1.941610
Prob(F-statistic)	0.000082	Second-Stage SSR	213.1462
J-statistic	0.000000	Instrument rank	4

▪ 수단변수가 과잉인 경우의 수단변수 추정법

- 단순회귀모형의 경우 수단변수는 하나만 필요함
  그러나 보통 그보다 더 많은 수단변수를 가지게 됨
- 예를 들어 w는 x 와는 상관되지만 e 와는 상관되지 않는
  또 하나의 수단변수라고 하면 세 개의 적률조건이 만들어 짐

$$E(e_t) = E(y_t - \beta_1 - \beta_2 x_t) = 0 \Rightarrow \sum \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right) = m_1 = 0$$

$$E(z_t e_t) = E\left[z_t (y_t - \beta_1 - \beta_2 x_t)\right] = 0 \Rightarrow \sum z_t \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right) = m_2 = 0$$

$$E(w_t e_t) = E\left[w_t (y_t - \beta_1 - \beta_2 x_t)\right] = 0 \Rightarrow \sum w_t \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_t\right) = m_3 = 0$$

- 두 개의 미지수 $\hat{\beta}_1$ $\hat{\beta}_2$ 에 대해 세 개의 관계식이 존재

- **수단변수가 과잉인 경우의 수단변수 추정법 (계속)**

- 이 경우는 다음의 추정절차가 최선의 방법으로 알려져 있음

(1) $x = a_0 + a_1 z + a_2 w$ 를 OLS로 추정하여, $\hat{x}$ 을 구함

(2) x 의 수단변수로 $\hat{x}$ 을 사용하여 회귀 모수를 OLS 추정함

- 두 단계의 OLS를 이용하여 모수를 추정하므로,
  수단변수 추정량을 2단계 최소제곱추정량
  (two-stage least squares estimators: 2SLS)이라고도 함

(예3) 임금함수 추정 - 수단변수 과잉 경우의 추정법

- 딸의 교육연수는 엄마의 교육연수와 아빠의 교육연수 모두와 밀접히 관련되어 있을 수 있음
- 이 관계를 고려하여 연립방정식 구성하여 2SLS로 추정하는 방법 있음

(1) 아래 식을 OLS로 추정, 추정결과는 <표 10.1>

$$EDUC = f(EXPER, EXPER2, MOTHEREDUC, FATHEREDUC)$$

(2) 임금함수 추정에 $EDUC$ 대신 그것의 추정치 $\widehat{EDUC}$ 을 이용

$$\widehat{\ln(WAGE)} = .0481 + .0614\widehat{EDUC} + .0442EXPER - .0009EXPER^2$$

(se) (.4003) (.0314) (.0134) (.0004)

- 앞의 추정결과와는 달리 $EDUC$ 추정치가 유의하게 나타남

- 2SLS 추정결과 $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER^2 + e$

Dependent Variable: LNWAGE
Method: Two-Stage Least Squares
Date: 09/19/12 Time: 10:01
Sample: 1 753 IF WAGE>0
Included observations: 428
Instrument specification: EXPER EXPER2 MOTHEREDUC FATHEREDUC
Constant added to instrument list

<Quick>-<Estimate Equation> 클릭,
<Method>에서 <TSLS> 선택,
윗칸: lnwage c educ exper exper2
아랫칸: exper exper2 mothereduc fathereduc
Sample: 1 753 if wage>0

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.048100	0.400328	0.120152	0.9044
EDUC	0.061397	0.031437	1.953024	0.0515
EXPER	0.044170	0.013432	3.288329	0.0011
EXPER2	-0.000899	0.000402	-2.237993	0.0257

R-squared	0.135708	Mean dependent var	1.190173
Adjusted R-squared	0.129593	S.D. dependent var	0.723198
S.E. of regression	0.674712	Sum squared resid	193.0200
F-statistic	8.140709	Durbin-Watson stat	1.945659
Prob(F-statistic)	0.000028	Second-Stage SSR	212.2096
J-statistic	0.374538	Instrument rank	5
Prob(J-statistic)	0.540541		

- **수단변수 추정법 – 일반적인 경우** $y = \beta_1 + \beta_2 x_2 + \cdots \beta_K x_K + e$

- 설명변수 중 일부가 내생변수(오차항과 상관된 변수)라고 하자

$$y = \overbrace{\beta_1 + \beta_2 x_2 + \cdots \beta_G x_G}^{G \text{ exogenous variables}} + \overbrace{\beta_{G+1} x_{G+1} + \cdots + \beta_K x_K}^{B \text{ endogenous variables}} + e$$

- L개의 수단변수 있다고 하자; $z_1, z_2, ..., z_L$

(1) 다음 식을 OLS로 추정,

$$x_{G+j} = \gamma_{1j} + \gamma_{2j} x_2 + \cdots + \gamma_{Gj} x_G + \theta_{1j} z_1 + \cdots + \theta_{Lj} z_L + v_j, \quad (j = 1, \ldots, B)$$

(2) 내생변수의 추정치 구함

$$\hat{x}_{G+j} = \hat{\gamma}_{1j} + \hat{\gamma}_{2j} x_2 + \cdots + \hat{\gamma}_{Gj} x_G + \hat{\theta}_{1j} z_1 + \cdots + \hat{\theta}_{Lj} z_L, \quad (j = 1, \ldots, B)$$

(3) 두 번째 단계의 OLS 추정

$$y = \beta_1 + \beta_2 x_2 + \cdots \beta_G x_G + \beta_{G+1} \hat{x}_{G+1} + \cdots + \beta_K \hat{x}_K + error$$

10.4 모형 설정에 대한 검정

- 설명변수가 오차항과 상관되어 있는지를 검정

  ⇒ 하우스만 검정(Hausman test)

  ⇒ OLS 이용할 것인지 수단변수법 이용할 것인지를 판단함

- 선택한 수단변수가 충분히 강한지(적절한지)를 검정

- 선택한 수단변수가 오차항과 비상관되었는지 여부에 대한 검정

(1) 설명변수와 오차항 사이의 상관관계 검정(내생성 검정)

- $H_0 : Cov(x,e) = 0$
  $H_1 : Cov(x,e) \neq 0$
  $\Rightarrow$ 하우스만 검정(Hausman test)

- 귀무가설이 참인 경우

  - OLS 추정량과 수단변수 추정량은 모두 일치추정량임
  - 대표본인 경우 동일해 짐, 즉 $q = \left(b_{ols} - \hat{\beta}_{IV}\right) \to 0$

$\Rightarrow$ OLS 사용하는 것이 적절함

- 귀무가설이 거짓인 경우

  - OLS 추정량은 일치추정량 아니고, 수단변수 추정량은 일치추정량임
  - 대표본인 경우에도 양자는 차이남, 즉 $q = \left(b_{ols} - \hat{\beta}_{IV}\right) \to c \neq 0$

$\Rightarrow$ 수단변수 추정량 사용하는 것이 적절함

- **Hausman test** <방법 1>

- $y_t = \beta_1 + \beta_2 x_t + e_t$ 에서 설명변수와 오차항의 상관 여부를 알고자 함
  수단변수는 z_1 , z_2 라고 하자

① OLS 이용하여 $x_t = a_0 + a_1 z_{t1} + a_2 z_{t2} + v_t$ 를 추정,
잔차를 계산 $\hat{v}_t = x_t - \hat{a}_0 + \hat{a}_1 z_{t1} + \hat{a}_2 z_{t2}$

② 인위적인 회귀식 $y_t = \beta_1 + \beta_2 x_t + \delta \hat{v}_t + e_t$ 를 OLS로 추정

③ t-검정 이용하여 다음과 같은 유의성 검정 실시

$H_0 : \delta = 0 \left(\text{no correlation between } x \text{ and } e\right) \Rightarrow z_1,\ z_2$ 불필요 $\Rightarrow OLS$로 충분

$H_1 : \delta \neq 0 \left(\text{correlation between } x \text{ and } e\right) \Rightarrow x \sim y \sim e \Rightarrow \text{cov}(x, e) \neq 0$

✓ OLS 추정량의 성과와 수단변수 추정량의 성과를 비교하는 검정

- **Hausman 검정 결과** / 앞에서 Monte Carlo simulation에서 사용한 자료

Dependent Variable: Y
Method: Least Squares
Date: 09/19/12 Time: 18:38
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.137591	0.079746	14.26510	0.0000
X	1.039872	0.133013	7.817819	0.0000
VHAT	0.995728	0.162939	6.111053	0.0000

R-squared	0.845043	Mean dependent var	1.386287
Adjusted R-squared	0.841848	S.D. dependent var	1.838819
S.E. of regression	0.731268	Akaike info criterion	2.241466
Sum squared resid	51.87097	Schwarz criterion	2.319621
Log likelihood	-109.0733	Hannan-Quinn criter.	2.273097
F-statistic	264.4899	Durbin-Watson stat	2.024781
Prob(F-statistic)	0.000000		

- 귀무가설 기각, 즉 설명변수와 오차항은 상관됨

  따라서 수단변수 추정법을 사용하여야 함

- **Hausman 검정 결과(2)** / 앞에서 Monte Carlo simulation에서 사용한 자료

- OLS 결과

Dependent Variable: Y

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.978893	0.088281	11.08838	0.0000
X	1.703431	0.089950	18.93754	0.0000

- 2SLS 결과

Instrument specification: Z1 Z2

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.137591	0.116444	9.769431	0.0000
X	1.039872	0.194223	5.354022	0.0000

Endogeneity Test
Specification: Y C X
Instrument specification: C Z1 Z2
Endogenous variables to treat as exogenous: X

<TSLS> 추정 후,
<View>–<IV Diagnostics and Tests/>–
<Regressor Endogeneity Test> 클릭

	Value	df	Probability
Difference in J-stats	27.24186	1	0.0000

- **Hausman 검정 결과(3)** / 앞의 저축함수 추정에서 사용한 자료

- OLS 결과

Dependent Variable: SAVINGS

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.342769	0.856116	5.072643	0.0000
INCOME	-0.005185	0.011163	-0.464539	0.6444

- 2SLS 결과

Instrument specification: C AVERAGE_INCOME

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.988267	1.322096	0.747500	0.4584
INCOME	0.039176	0.017381	2.253907	0.0288

Endogeneity Test
Specification: SAVINGS C INCOME
Instrument specification: C AVERAGE_INCOME
Endogenous variables to treat as exogenous: INCOME

<TSLS> 추정 후,
<View>-<IV Diagnostics and Tests/>-<Regressor Endogeneity Test> 클릭

	Value	df	Probability
Difference in J-stats	11.08634	1	0.0009

- **Hausman test** <SAS에서의 검정방법> : 저축함수 경우

- OLS 추정량과 수단변수 추정량을 각각 추정하여 비교

<SAS program>

```
proc model data=savings ;          * Automatic Hausman test;
parms b1 b2 ;                      * parms;
exogenous z ;                      * instrument;
y = b1 + b2*x ;                    * model;
fit y / ols 2sls hausman ;         * Hausman comparing OLS-2sls;
```

<Hausman's Specification Test Results>

Comparing	To	DF	Statistic	Pr > ChiSq
OLS	2SLS	2	7.11	0.0286

- 귀무가설 기각, 즉 설명변수와 오차항은 상관됨

  따라서 수단변수 추정법을 사용하여야 함

(2) 약한 수단변수에 대한 검정

- 선택한 변수가 약한 수단변수인 경우라면,
  수단변수 추정량은 큰 편의 및 큰 표준오차를 가질 수 있음
- 다음과 같은 모형을 생각해보자

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_G x_G + \beta_{G+1} x_{G+1} + e$$

- x_2, x_3, ..., x_G는 외생변수, x_{G+1}은 내생변수, z_1은 수단변수

$$x_{G+1} = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_G x_G + \theta_1 z_1 + v$$

- z_1이 강한 수단변수인지 약한 수단변수인지를 검정하는 방법
- $H_0 : \theta_1 = 0$ 검정, t-값이 3.3보다 작으면, 약한 수단변수로 판단

- **Test for weak instruments** / Monte Carlo simulation에서 사용한 자료

Dependent Variable: X
Method: Least Squares
Date: 09/19/12 Time: 18:32
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.194732	0.079499	2.449486	0.0161
Z1	0.569978	0.088785	6.419747	0.0000
Z2	0.206786	0.077161	2.679940	0.0087

R-squared	0.333594	Mean dependent var		0.239161
Adjusted R-squared	0.319854	S.D. dependent var	z_1	0.956655 z_2
S.E. of regression	0.788963	Akaike info criterion		2.393346
Sum squared resid	60.37887	Schwarz criterion		2.471501
Log likelihood	-116.6673	Hannan-Quinn criter.		2.424977
F-statistic	24.27844	Durbin-Watson stat		1.888121
Prob(F-statistic)	0.000000			

- z_1 은 유의(강한 수단변수), z_2 의 t 값은 3.3보다 작음(약한 수단변수)
- 두 추정치가 동시에 0이라는 귀무가설(결합가설)에 대해 F-검정, 기각

<과제>

10.5 / data는 WILEY 교과서 홈페이지에 있음

http://principlesofeconometrics.com/poe3/poe3.htm