
2012 호주 HILDA User Training 참가 결과

2012. 12.

I 개 요

1. 출장개요

- 교육명 : “HILDA User training”
 - 주 관 : Melbourne Institute of Applied Economic and Social Research (University of Melbourne)
 - 기 간 : ‘12. 11. 27. ~ 11. 29.
 - 장 소 : 호주 멜버른
- 출장자 : 복지통계과, 5급 정구현
조사연구실, 6급 심규호
- 출장기간 : ‘12. 11. 26. ~ 12. 1.

2. 출장목적

- 복지통계과는 2010년 가계금융복지조사를 시작으로 2012년 가계금융복지조사까지 2010년부터 계속적으로 가구 패널 조사를 실시하고 있음
- 통계개발원에서는 2012년 “가계금융조사를 이용한 무응답 조정 방법 연구”, “가계금융조사를 위한 종단, 횡단 가중값 작성 방법 연구” 등 가구패널 조사 관련 연구를 수행하였음
- 가구패널조사는 가구를 매년 추적하여 조사하는 패널조사로서 패널조사수행에 필요한 방법론과 조사 이후의 패널 이탈과 마모로 인한 가중값 연구 등의 방법론 연구가 많이 필요함
- 2차 wave 실시 후 정확한 패널분석을 위해 종단면 분석기법에 대한 연구 및 지식습득이 요구됨

- 호주의 HILDA(The Household, Income and Labour Dynamics in Australia) 패널 조사는 2001년에 시작된 조사로 패널조사의 대표적인 조사임
- HILDA User Training은 패널조사를 수행하는 기관 관계자나 학계 관련자를 위한 교육으로서 통계청 패널조사의 원활한 수행과 향후 패널 관련 연구에 중요한 도움이 될 것으로 판단됨

Ⅱ 주 요 내 용

1. HILDA 예제를 이용한 패널 자료 분석 기법

○ Day 1 : 패널 자료 분석 기초

- 패널 자료란 일반적으로 같은 개인(가구)에 대해 반복적으로 측정되는 관측값을 말함
- 종단 자료의 종류
 - 코호트 조사, 패널 조사, 비임시(non-temporal)조사 패널, 비-조사 패널(non-survey panels : aggregate panels)
- 장기 가구 패널(Long-term household panels)
- 왜 패널자료를 사용하는가?
 - 개인 간의 관찰되지 않은 효과를 분리해낼 수 있음
 - 일시적 순서화를 통해 원인의 추론이 강해질 수 있음
 - 횡단조사에서의 조사 이력을 이용하는 것보다 현재의 관측값을 반복하는 것이 더 신뢰성이 있음
 - “가난의 지속”, “불안정한 고용” 등의 어떠한 현상은 종단적으로 계승됨
- 그러나 패널자료는 다음과 같은 한계점을 가짐
 - 사람간의 변동은 일반적으로 개인에 대한 시간의 변동을 크게 넘어섬
 - 어떠한 중요한 변수들은 시간의 변동이 존재하지 않을 수 있음
 - 시간 변동은 측정 오차에 의해 부풀려질 수 있음
 - 패널 자료는 고정된 시간 구조(fixed timing structure)에서 얻어짐; 연속된 시간(continuous-time)에서의 생존 분석이 더 정보를 줄 수 있음
 - 아직까지도 패널자료에서 명확한 추론을 얻기 위해서는 강한 가정을 필요로 함

- 패널 자료의 기초 핸들링

- Wide Form : ID를 왼쪽에 웨이브를 변수로
- Long Form : ID와 웨이브를 별도의 변수로
- Matching(Merging), Aggregation, Appending, Ordering

- Describing Panel Data

- (1) Describe patterns of panel data : Stata(xtdes)

```
. xtdes
pid: 10002251, 10004491, ..., 1.347e+08      n =      16082
wave: 1, 2, ..., 13                        T =       13

      Delta(wave) = 1; (13-1)+1 = 13
      (pid*wave uniquely identifies each observation)

Distribution of T_i:   min      5%      25%      50%      75%      95%      max
                     1        1        2        7       13       13       13

Freq.  Percent   Cum. | Pattern
-----+-----
    4648    28.90  28.90 | 1111111111111
     997     6.20  35.10 | 1.....
     646     4.02  39.12 | 11.....
     376     2.34  41.46 | .....1
     342     2.13  43.58 | 111.....
     327     2.03  45.62 | 1111.....
     261     1.62  47.24 | .....11
     254     1.58  48.82 | .1.....
     251     1.56  50.38 | .....111
    7980    49.62 100.00 | (other patterns)
-----+-----
    16082   100.00      | XXXXXXXXXXXXXXX
```

- (2) Between-and within-group variation

pid	wave	Wage	Grand mean	Ind. Mean	Within dev	Between dev	Total dev
10028005	1	9.302	9.841	10.948	-1.646	1.107	-0.539
10028005	2	10.444	9.841	10.948	-0.504	1.107	0.603
10028005	3	13.883	9.841	10.948	2.935	1.107	4.042
10028005	4	4.573	9.841	10.948	-6.375	1.107	-5.268
10028005	5	13.769	9.841	10.948	2.820	1.107	3.928
..	..	..	..	..	..	..	..
10028005	13	12.914	9.841	10.948	1.966	1.107	3.073
10060111	1	13.046	9.841	12.953	0.094	3.112	3.205
10060111	2	12.923	9.841	12.953	-0.030	3.112	3.081
10060111	3	13.453	9.841	12.953	0.500	3.112	3.612
10060111	4	13.505	9.841	12.953	0.553	3.112	3.664
10060111	5	12.418	9.841	12.953	-0.535	3.112	2.577

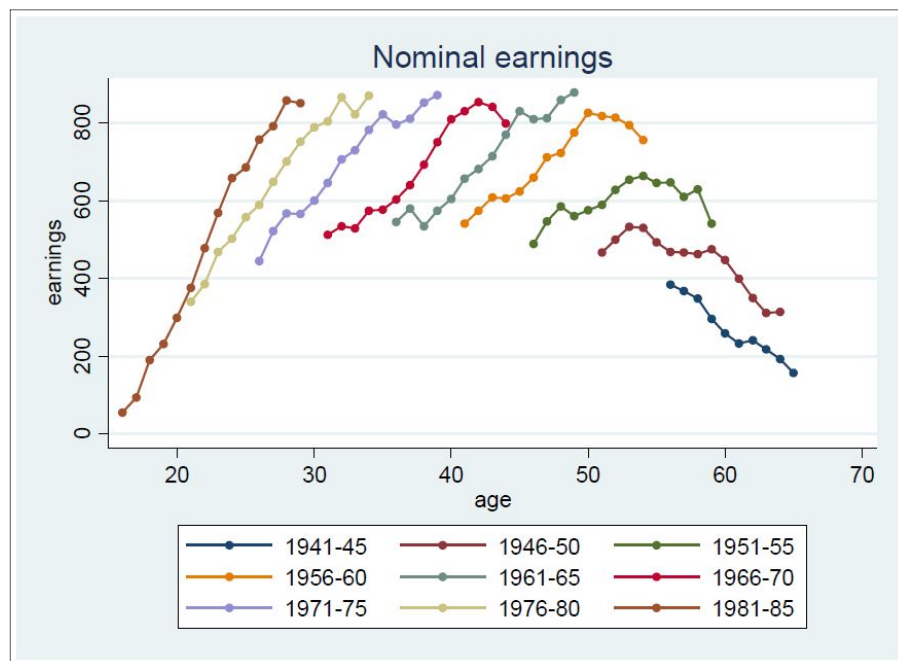
– (3) Transitions

```
. xttrans dlevec, freq
```

have you ever taken cannabis	have you ever taken cannabis				Total
	Yes	No	DK	DWTA	
Yes	728	111	0	1	840
	86.67	13.21	0.00	0.12	100.00
No	251	2,189	6	7	2,453
	10.23	89.24	0.24	0.29	100.00
DK	2	9	1	1	13
	15.38	69.23	7.69	7.69	100.00
DWTA	9	5	0	1	15
	60.00	33.33	0.00	6.67	100.00
Total	990	2,314	7	10	3,321
	29.81	69.68	0.21	0.30	100.00

• 13% of people who'd used cannabis before 2003 say they've never used before 2004!!

– (4) Age and cohort.



○ Day 2 : 패널 회귀 분석과 일반화

- Modelling Approaches

- 1. 각각의 시간 기간(time period)에 대한 종단 집합에 대해서는
 - Between-group regression
 - The Structural Equations(SEM) approach - 각각의 기간에 대해 한 개의 수식(equation)
- 2. 각각의 개인(individual)에 대한 종단 집합에 대해서는
 - Within-group regression
 - 개인의 이질성(heterogeneity)을 이용한 동적 모형(Dynamic models)
 - Latent growth curve analysis
 - Trajectory Analysis
- 3. 개인과 시간의 변동을 모두 확인할 수 있는 종합적인 모형사용

- 관련 방법 : Latent growth cyurves

- 사회학, 심리학, 범죄학에서 많이 사용되지만 경제학에서는 많이 사용되지 않음.
- 상수와 계수가 전 개인에 걸쳐 랜덤하게 퍼짐을 가정
- 모든 개인이 같은 계수를 가진다고 가정하지 않는다는 장점을 가짐
- 묘사성이 떨어지며 완벽하지 않은 동적 모형을 가지는 단점이 있음

- 관련 방법 : SEM(Structural Equation modeling)

- 많은 분야에서 사용되나 서로 다른 용어를 사용함에서 차이가 있음
- 일반적인 범용적인 구조를 가진다는 장점이 있음.(예를 들면 어떤 특수한 경우에는 회귀 계수가 모든 기간에 걸쳐 같음)
- 어떻게 모수가 시간 걸쳐 존재하는지에 대한 이론적 배경이 없음
- 길거나(long) 아주 불균형한 패널(unbalanced panels)에서는 사용하기 어려움

- 관련 방법 : Multi-Level Models

- 일반적인 패널 자료를 Multiple Dimension(예. 가구 및 개인)에 적용시킴
- MLWin, MPlus, Lisrel, xtmixed(Stata)에서 분석이 가능함

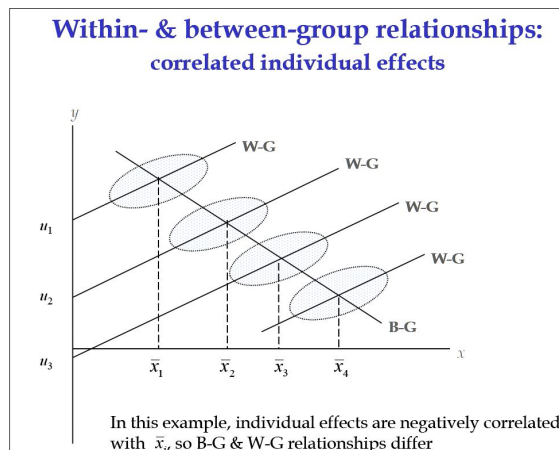
- Pooled regression

- 종속변수 y 에 영향을 미치는 설명변수, 예를 들면 x 와 z 가 서로 개인에 따라서 또는 시간에 따라서 상관이 없음을 가정
- 따라서 이 방법은 개인의 효과를 무시할 가능성이 있음
- Pooled regression은 적절하지 않은 조건에서는 불편추정량을 제공하나 잘못된 표준오차를 제공하는 비효율적인 결과를 줄 수 있으므로 좋은 방법은 아님

- Least-squares dummy variables(LSDV) regression

- 어떠한 패널 개인 i 와 연관된 time t 에 대해 1을 주는 더미 변수를 가정
- 'within-group' 변환을 통해 각각의 변수에서 더미 변수의 효과를 제거
- 더미변수가 제거된 y 에 대해 설명변수를 회귀분석함

- Between-group regression



- 'Random effects' GLS & ML estimation

- 일반적으로 패널의 개인은 모집단으로부터 랜덤하게 표본으로 뽑히기 때문에 "random effect"라는 것은 종언부언하게 되는 것임
- 따라서 이러한 "random effect"효과를 정확하게 정의하고 추정하는 것이 필요

- Hausman test

- "Uncorrelated effects"를 검정하기 위한 방법

$$H_0: E(u_i | \mathbf{z}_i, \mathbf{X}_i) = 0$$

It is important to test H_0 . There are various equivalent ways of doing so, including:

(1) Hausman test: is the difference $\hat{\beta}_W - \hat{\beta}_{GLS}$ large?

(2) Between-within comparison: is $\hat{\beta}_W - \hat{\beta}_B$ large?

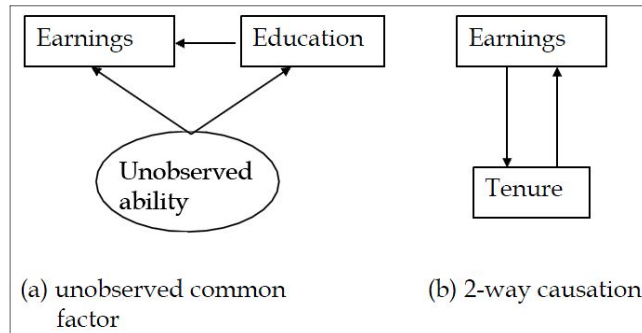
(3) Mundlak approach: estimate the model

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\gamma} + \eta_i + \varepsilon_{it}$$

by GLS and test $H_0: \boldsymbol{\gamma} = \mathbf{0}$.

- Endogeneity : a serious issue

- 다음과 같은 종류의 내생성을 가질 수 있음



- 내생성의 두종류 사례

- 소득모델

$$y_{it} = \alpha_1 Educ_i + \alpha_2 Female_i + \beta_1 Age_{it} + \beta_2 Tenure_{it} + u_i + \epsilon_{it}$$

- Two-way causation : 근속기간은 높은 소득과 관련

Tenure model : $Tenure_{it} = \gamma y_{it} + v_{it}$

- Unobserved common factors : 교육은 u_i (능력)과 관련

$$Educ_i = \delta u_i + other\ vars$$

- 내생성의 종류에 따라 다음의 방법을 적용

Strategy for dealing with endogeneity		
Type of endogeneity	Consequences	Method
2-way causation (e.g. tenure → wage & wage → tenure)	$\text{Cov}(x, u) \neq 0$ $\text{Cov}(x, \varepsilon) \neq 0$	Within-group IV (w-g to eliminate u_i and IV to deal with covariance with ε)
Common unobserved factor which persists over time (e.g. ability → wage, ability → education & education → wage)	$\text{Cov}(x, u) \neq 0$ $\text{Cov}(x, \varepsilon) = 0$	Within-group regression (eliminates u_i) and Hausman-Taylor to estimate coefficients of z_i
Common unobserved factor which does not persist over time (e.g. unobserved plant closure → wage & job change → tenure)	$\text{Cov}(x, u) = 0$ $\text{Cov}(x, \varepsilon) \neq 0$	Random-effects IV, using as IVs variables which are correlated with risk of job loss but not wages; no need to use within-group, since u_i isn't correlated with x
None	$\text{Cov}(x, u) = 0$ $\text{Cov}(x, \varepsilon) = 0$	GLS random effects regression

- The Instrumental variables principle

- 일반적인 종단 회귀분석모형에서는 설명변수와 오차항의 공분산이 "0"이 아니며 y에 대한 x의 OLS(ordinary least square) 회귀결과가 편향됨
- 다음과 같은 성질을 만족시키는 q 변수를 가정

Validity: $\text{cov}(q_i, \varepsilon_i) = 0$

Relevance: $\text{cov}(q_i, x_i) \neq 0$

- 아래의 추정값을 구하여 q변수를 해석

$$\hat{\beta}_{IV} = \frac{\text{sample cov}(q, y)}{\text{sample cov}(q, x)} = \frac{\sum_{i=1}^n (q_i - \bar{q})(y_i - \bar{y})}{\sum_{i=1}^n (q_i - \bar{q})(x_i - \bar{x})}$$

○ Day 3 : 이분형 응답 모형과 패널 이탈

- Binary models

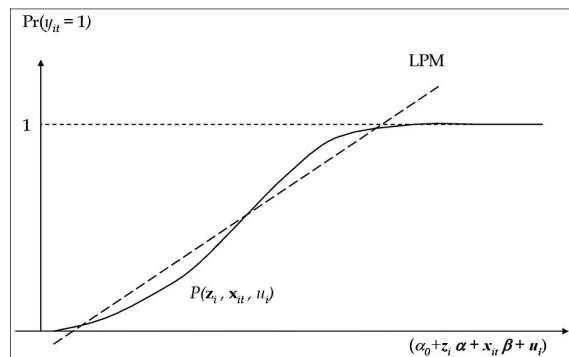
- 다음과 같은 이분형 모형을 가정

- $y_i = (0, 0, 0, 0, 1, 1, 1, 0, 1, 1)$ is a monthly panel observation
- 0 indicates unemployment, 1 indicates employment

- 이 경우 다음과 같은 선형 확률 모형이 됨(LPM)

$$y_{it} = \alpha_0 + \mathbf{z}_i \boldsymbol{\alpha} + \mathbf{x}_{it} \boldsymbol{\beta} + u_i + \varepsilon_{it}$$

- 그러나 이 모형은 $[0,1]$ 구간 밖으로 넘어갈 가능성이 있음



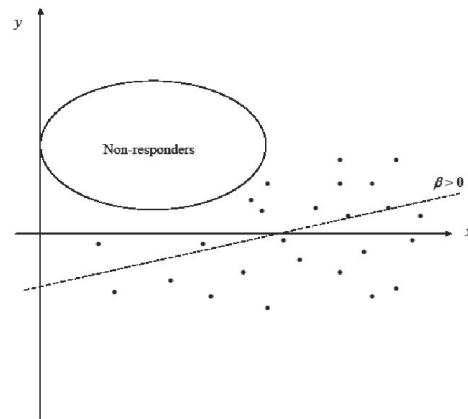
- 이러한 문제점에도 불구하고 LPM모형은 추정이 간단하고 적용이 쉬우나 좋은 방법은 아님
- 따라서 비선형 모형을 가정하게 됨
- Latent regression models같은 비선형 모형을 사용

- Attrition and sample selection

- Incomplete Panels는 다음과 같은 문제를 가짐

- 추정값이 편향됨
- 무응답 패턴이 랜덤이 아님
- “sub-panel”이 랜덤이 아님으로 인해 매 웨이브에 따라 모집단을 대표하지 않게 됨

Illustration of non-response bias



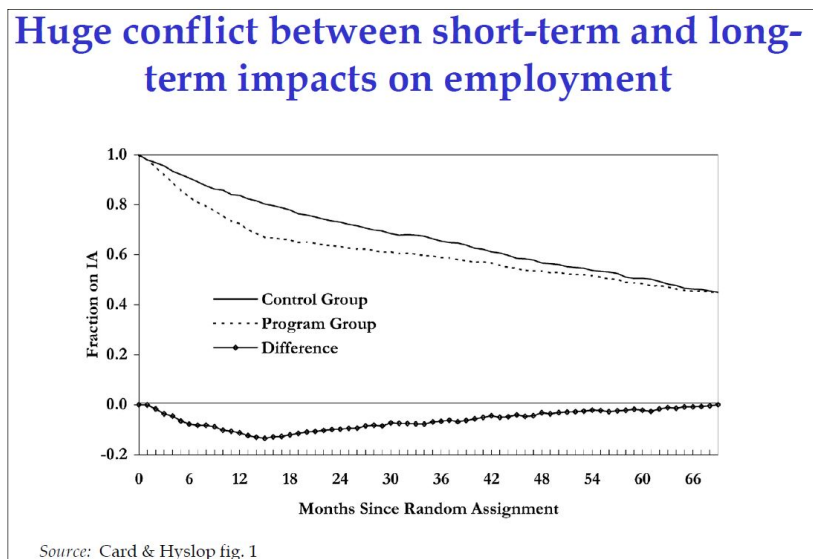
Spurious apparent relationship between y and x ($\beta > 0$)

- “unbiased panel”과 “biased sub-panel”의 “Random Effect”와 “Fixed Effect” 추정값의 “*Hausman 비교 검정*”을 통해 패널 사이의 유사성을 검정함

2. 패널 자료의 동적 모델링

○ Introduction

- 조사 기반의 패널 데이터는 다음의 두 가지 관점에서 볼 수 있음
 - 시간 T에서의 자료의 모습(반복 측정값)
 - 종단 자료는 독립이 아님
 - 종종 모수가 시간의 흐름에 따라 변동이 생김
 - T만큼의 길이를 가지는 n개의 모습
 - 개인의 자료는 독립적으로 추출됨, 따라서 통계적으로 독립
 - T가 작은 경우 개인 대 개인의 시계열 분석은 가능하지 않음
- 기존연구에서 살펴보면 패널은 다음의 두 가지 방법에 사용됨
 - 관측되지 않은 개인의 효과를 조정하면서 반복된 개별 자료를 분석
=> *fixed effect/within-group regression*
 - 개인이 시간의 흐름에 따른 환경의 변화를 조정해 가면서 발견되는 결과로서의 시간의 흐름을 분석
=> *Dynamic modeling*
- 왜 Dynamics를 걱정해야 하는가?
 - Canadian SSP(self sufficiency project)의 경험
 - 마감일(deadline effect)에 의해 dynamics 효과가 발생



○ Dynamic linear models

Dynamic regression: models implying long-term adjustment

Example of an AR(1) process with explanatory covariates:

$$y_t = \rho y_{t-1} + \beta x_t + \varepsilon_t, \quad -1 < \rho < 1$$

Distributed lag form:

$$y_t = \rho^t y_0 + \beta [x_t + \rho x_{t-1} + \dots + \rho^{t-1} x_1] + [\varepsilon_t + \rho \varepsilon_{t-1} + \dots + \rho^{t-1} \varepsilon_1]$$

[initial condition] [MA(t-1) in x] [MA(t-1) in ε]

○ What happens if you ignore dynamics?

- 변동성을 무시한 모형과 실제 모형에서의 추정값과 Bias가 발생
- 변동성을 무시한 모형에서의 Simulation된 예제 : 차이 발생

Simulated example of dynamic misspecification

• Example: a "true" wage model

$$\ln(\text{wage})_{it} = 0.4 + 0.5 \ln(\text{wage})_{it-1} + 0.02 \text{ age}_{it} + u_i + \varepsilon_{it}$$

$u_i \sim N(0, 0.2^2)$; $\varepsilon_{it} \sim N(0, 0.3^2)$; $\text{age}_{it} \sim \text{uniform}(20, 55)$; $n = 1000$; $T = 10$

• True long-run age-earnings impact is $\gamma / (1 - \rho) = 0.040$

• Random-effects regression of $\ln(\text{wage})_{it}$ on age_{it} gives downward bias:

Random-effects GLS regression		Number of obs	=	10000
Group variable: id		Number of groups	=	1000
R-sq: within = 0.5295		Obs per group: min	=	10
between = 0.6753		avg	=	10.0
overall = 0.0382		max	=	10
corr(u_i, X) = 0 (assumed)		Wald chi2(1)	=	23.22
		Prob > chi2	=	0.0000

	lnw	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
age		.0076083	.0015789	4.82	0.000	.0045137 .0107029
_cons		2.918111	.06888	42.37	0.000	2.783108 3.053113
sigma_u		.23807547				
sigma_e		.86466913				
rho		.07046827				(fraction of variance due to u_i)

○ Dynamic panel data regression models

- 기존의 모형은 dynamics를 반영하는데 한계가 있기 때문에 다음과 같은 모형을 제안
- Arellano-bond gmm(1991)
- Anderson-hsiao
- 모형의 추정 결과 기존의 모형들이 과대추정되는 것을 알 수 있음

Model	Estimated coefficient
FE regression	0.226 (0.007)
Pooled OLS	0.827 (0.008)
RE regression	0.667 (0.005)
Anderson-Hsiao	0.136 (0.028)
Arellano-Bond	0.132 (0.021)

- 현재에도 계속적으로 수정된 모형이 제안되어 오고 있음

Ⅲ 시 사 점

□ Panel 자료 분석에 대한 깊이 있는 이론적 배경을 공유하고 실제 HILDA 패널 자료를 이용한 분석 결과를 공유

- 호주 HILDA 에서는 패널자료 구축과 패널자료 분석을 위한 통계적 방법론 교육을 매년 개최하여 각국의 패널 관련 업무 담당자와 학계의 관심 있는 사람들에게 교육의 기회를 제공하고 있음
- 패널자료의 분석은 패널자료를 구축하는 기관에서는 큰 관심의 주제는 아니고 분석은 계량경제학이나 사회학을 하는 분석자의 몫이긴 하나, 패널자료를 분석해 봄으로서 패널자료를 이용한 통계가 어떠한 특성과 장점을 가지는지에 대해서는 분석해 볼 필요가 있다고 판단됨
- 그러나 패널자료의 분석은 기존의 통계적 방법론 이외에 깊이 있는 통계적 추론에 대해 알아야 하고 시계열이나 계량경제학 등의 많은 기법이 필요함
- 패널자료의 축적이 진행될수록 이러한 패널 자료의 특성과 통계적 의미를 정확하게 파악해야 통계청에서 발표하는 패널 자료의 분석 통계도 오해가 없을 것으로 판단됨
- 향후 패널 자료의 분석에 지속적인 관심을 가지고 분석을 시도하면서 전문가를 양성하여 패널 자료에 대한 구축과 분석에 대한 노하우를 축적할 필요가 있을 것으로 판단됨

- 아울러 통계청 패널조사 발전을 위해 관련기관 및 전문가와 네트워크를 형성하여 지속적인 교류를 통한 지식 공유 및 분석기법 강화 필요

IV 기 타 사 항

□ 자료제공 웹사이트

- <http://www.melbourneinstitute.com/hilda/training/past.html>

□ 관련문헌

- 계량경제
 - Arellano, M.(2003). *Panel Data Econometrics*, Oxford University Press,
 - Baltagi, B.H.(2005). *Econometric Analysis of Panel Data*, Wiley,
 - Hisao, C.(2003). *Analysis of panel Data(2nd ed.)*, Cambridge University press
 - Wooldridge, J.M.(2002), *Econometric Analysis of Cross section and Panel Data*, MIT Press
- 사회학
 - Halaby, C.N.(2004). Panel Data models in sociological research : Theory into practice, *American Review of Sociology* 30, 507-544
- STATA
 - Cameron, A.C. & Trivedi, P. K.(2009) *Microeconometrics Using Stata*, Stata Press
 - Rabe-Hesketh, S. & Skrondal, A.(2008) *Multilevel and Longitudinal Modeling Using Stata*, S&A., Stata press
 - StataCorp(2009), *Stata Statistical Software : Release 11 : XT reference Manual*, Stata Corporation