

# 패널자료 분석 전문과정 국외출장 결과보고

2010. 7.

# I. 출장 개요

## 1. 출장목적

- 패널조사 결과의 횡단면/종단면 자료를 분석하기 위한 STATA 프로그램 사용방법 및 결과 해석 능력 배양

## 2. 출장지

- 영국 사우햄프턴 대학 통계과학연구소  
(Southampton Statistical Sciences Research Institute)

## 3. 출장기간

- 2010. 6. 28. ~ 7. 4.

## 4. 출장자

- 표본과장 김규영

## 5. 프로그램 주요 일정 및 내용

- 1일
  - 종단면(longitudinal)자료 소개
  - stat 프로그램 소개 및 사용법 설명
  - 반복측정자료의 탐색
  - 종단면자료 탐색
  - 반복측정자료의 모델링 방법
- 2일
  - population average models 방법

- population average models의 적합방법 프로그래밍
- (random effect models)
- 확률효과모델 적합방법 프로그래밍
- 로지스틱 회귀모델(logistic regression models)
- 반복측정모델(repeated measures models) 프로그래밍
- 로지스틱 회귀모델 적합방법 프로그래밍

3일

- 반복측정자료에 대한 로지스틱 회귀 모델
- 반복측정자료에 대한 로지스틱 회귀 모델 적합방법 프로그래밍
- 가중치와 복합표본설계의 관리
- 가중치와 복합표본설계의 관리에 대한 프로그래밍
- 종합 복습

## 6. 참가자(상세내용은 별첨 참조)

- Prof. Alan Felstead, Cardiff University 외 24명

## 7. 강사

- Ann Berrigton, Southampton Statistical Sciences Research Institute
- Peter W. F. Smith, Southampton Statistical Sciences Research Institute
- Marcel Vieira, Federal University of Juiz de Fora

## II. 프로그램 내용

### 1. 종단면(longitudinal)자료 소개

#### 종단면 자료의 형태

- 반복적 횡단면(**repeated cross-section**) 자료 : 동일한 내용을 다른 표본에게서 반복적으로 조사하여 얻어지는 자료
- 회고적 횡단면(**retrospective cross-section**) 자료 : 과거부터의 자료도 응답자로 얻지만 한번에 횡단면으로 조사되어 얻어지는 자료
- 패널(**panel**)자료 : 동일한 응답자에게서 다른 시기에 동일한 또는 다른 질문을 통해 얻어지는 조사자료

#### 종단면 연구방법

- 코호트(**cohort**) 방법 : 특정 연령대로부터 얻어지는 표본이 계속 조사
- 반복측정 방법 : 동일한 개체로부터 여러 차례 측정이 이루어짐

### 2. 왜 종단면 자료를 수집하고 분석해야 하는 가?

#### 횡단면 자료는 어느 한 특정 시점에서의 사회의 snapshot을 제공

- 주로 macro-level에 관련된 자료

#### 종단면 자료는 micro-level 과정에 주로 관련되어 보다 풍부하고 상세한 사회적 과정과 구조적 내용을 볼 수 있다

### 3. 패널자료 분석시 고려사항

- 패널탈락
- wave 무응답 및 항목 무응답
- 패널 조건
- 복잡한 자료구조 : 개인과의 연결 문제 등
- 복잡한 조사설계
  - 층화, 집락 및 표본설계에 따른 가중치
- 시간에 따른 동일 개인에 대한 관찰치의 상관관계 처리

### 4. 프로그램 및 분석 사례자료

- 사례자료 : 영국 가구 패널조사(BHPS : British Household Panel Survey)
- 표본설계 및 패널관리
  - 1991년 약 5,500가구(10,000명의 개인) 사례자료 : 영국 가구 패널조사(BHPS : British Household
  - 표본설계 : 우편번호 주소화일을 활용한 층화집락추출
  - 새로운 가구로 나뉘어도 원표본에서 연간조사 실시
    - 새로운 가구도 조사 실시
  - 표본가구에 들어오는 새로운 가구원이 적격일 경우나 아이들이 16세에 도달하면 조사
  - 새로운 표본 확장 : 스코틀랜드, 웨일즈 및 NI
  - 조사내용 : 주내용과 순환모듈로 구성
  - 무응답 : 항목무응답과 단위무응답이 존재
    - 1차 웨이브 단위무응답은 횡단면 조사와 유사
    - 2차 이상의 웨이브 단위무응답은 사망, 이민 등으로 표본에서 탈락하는 경우(attrition)와 중간 무응답이 존재(장래에 다시 응답)

※ BHPS 응답율

웨이브	부적격율(%)	적격 응답율(%)	전 웨이브 상에서 적격 응답율(%)
2	1.4	87.7	87.7
3	2.9	81.5	79.1
4	4.3	79.9	74.8
5	5.6	76.8	70.6
6	6.9	77.3	68.7
7	8.4	76.0	66.7
8	9.5	74.1	64.7
9	10.5	72.1	62.4
10	12.0	70.4	60.0
11	12.8	68.4	59.3
12	13.7	66.6	57.1
13	14.8	64.9	55.1

5. 반복측정자료

□  $y_{ij}$  :  $t_{ij}$  시기의 응답

-  $t_{ij}$  시기의 측정되는  $p$  벡터의 공분산 :

$$\mathbf{x}_{ij} = \begin{pmatrix} x_{ij1} \\ x_{ij2} \\ \vdots \\ x_{ijp} \end{pmatrix}$$

- 개체(subjects) :  $i = 1, \dots, m$

- 개체  $i$ 에 대한 관찰치 :  $j = 1, \dots, n_i$

□ 상관관계 구조

- OLS 회귀모델시 잔차 :  $r_{ij} = y_{ij} - \hat{y}_{ij}$

- OLS 가정이 유효하다면 잔차들은 개인의 시간에 대해서는 상관없이 된다.

## □ 상관관계 구조의 형태

- 독립 : 동일한 개체의 관찰치 간에는 상관관계가 없다
  - $Cor(r_{ij}, r_{ik}) = \rho_{jk} = 0$
  - OLS 모델의 가정
- **exchangeable(uniform)** : 동일한 개체의 어느 관찰치 짝 간에는 동일한 잔차 상관관계가 있다
  - $Cor(r_{ij}, r_{ik}) = \rho_{jk} = \rho$
- **AR(1)(Exponential)** : 동일한 개체의 두 관찰치 간에 다음과 같은 잔차 상관관계를 갖을 경우
  - $\rho_{jk} = \rho^l$  여기서  $l$ 은 두 관찰치간의 시간길이 즉 lag임
  - 동일한 개체의 두 관찰치 간의 잔차 상관관계는 점차 감소 즉  $\rho, \rho^2, \rho^3, etc$
- **unstructured** : 동일한 개체의 두 관찰치 간의 잔차 상관관계에 아무런 제약이 없다
  - $-1 \leq \rho_{jk} \leq 1$

## □ 반복측정자료의 회귀모델

- $y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$
- $\epsilon_{ij}$  가 평균이 0 이고  $i$ 와  $j$ 에 대해 독립이라면 모델을 적합하기 위해 표준적 방법을 사용할 수 있다
- 모델에 더 많은 공변량을 추가하면

$$y_{ij} = \beta_0 + \underline{x}_{ij}^T \underline{\beta} + \epsilon_{ij} \text{ 가 된다.}$$

- $y_{ij} = \beta_0 + \underline{x}_{ij}^T \underline{\beta} + \epsilon_{ij}$  로 변형할 수 있다.  
 $= \beta_0 + \underline{x}_{ij}^T \underline{\beta} + u_i + \epsilon_{ij}$

여기서  $u_i$ 는  $y$ 에 영향을 미치는 측정되지 않은 개인 요소로 개체에 특정한 잔차임

## □ 고정 및 확률 절편 모델(fixed and random intercept model)

- 고정 효과 모델 :  $u_i$ 가 고정
  - 장점 :  $u_i$ 가  $x_{ij}$ 와 어떻게 관계하는가에 관한 가정이 불필요
  - 표준회귀모델 --> 추정시 OLS 방법 사용 가능
- 확률 효과 모델 :  $u_i$ 가 다음의 성질을 갖는 랜덤
  - 평균 : 0
  - $Var(u_i) = \sigma_u^2$
  - $Cor(u_i, \epsilon_{ij}) = 0$
  - 파라메타의 수가 적다. 고정효과에서는  $m$ 개의 파라메타  
-> 두 개의 파라메타로 축소( $\beta_0$ 와  $\sigma_u^2$ )
  - GLS
  - MLE(최대우도추정)
  - REML 추정

## □ Population average models

- 공변량  $x_{ij}$ 의 함수로  $y_{ij}$ 의 평균을 모델화
  - 잔차 상관관계 행렬에 대한 가정을 두고함
- 평균과 잔차 상관관계를 각각 모델할 수 있음
  - 잔차 상관관계 모델을 찾고, 평균 모델을 찾음
- MLE나 REML로 추정

## □ Transition models

- OLS를 사용한 회귀는 응답변수로 공변량외에 이전 시간에서의 값을 사용

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha y_{i(j-1)} + \epsilon_{ij}$$

여기서  $\alpha$ 는 개체내 상관관계(within-subject correlation)를 설명



## 6. 확률효과모델

□ 시간을 연속형 변수로 간주하고, 다른 어떤 공변량은 없다고 가정

-  $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \epsilon_{ij}$ , 여기서  $x_{ij}$  는 시간

-  $Var(u_i) = \sigma_u^2$

-  $Var(\epsilon_{ij}) = \sigma_\epsilon^2$

-  $Cor(u_i, \epsilon_{ij}) = 0$

- 다른 공변량을 추가 : 나이, 성별, 교육수준, 개체가 14세 이상시  
모의 경제활동여부, 시간  $t$ 에서의 경제활동

□ 확률 기울기(계수) 모델

-  $y_{ij} = \beta_0 + (\beta_1 + b_i)x_{ij} + u_i + \epsilon_{ij}$

•  $\beta_1$  : 평균 기울기

•  $b_i$  : 평균 기울기의 개체에 의존한 랜덤편차

•  $u_i$  : 개체에 의존한 랜덤 절편

□ Multi-process 모델

- 동시에 하나 이상의 상관 과정을 고려한 모델

$$y_{1ij} = \beta_{10} + \underline{x}_{ij}^T \underline{\beta}_1 + u_{1i} + \epsilon_{1ij}$$

$$y_{2ij} = \beta_{20} + \underline{x}_{ij}^T \underline{\beta}_2 + u_{2i} + \epsilon_{2ij}$$

□ 무응답과 탈락

- 무응답 체계

• MCAR(Missing completely at random) : 무응답 확률이 관찰치나 무응답 관찰치에 의존하지 않음

• MAR(Missing at random) : 무응답 확률이 관찰치에 의존하고, 무응답 관찰치에는 의존하지 않음

- NMAR(Not Missing at random) : 무응답 확률이 무응답 관찰치에 의존

## 7. 로지스틱 회귀모델

### □ 로지스틱 회귀모델

- $y_i$ 가 성공 확률  $p_i = P(y_i = 1)$ 를 갖는 이항분포(binary distribution)라 가정
- 성공에 대한 log-odds 모델  

$$\text{logit}(p_i) = \log[p_i/(1-p_i)] = \beta_0 + \underline{x}_i^T \underline{\beta}$$
- 파라메타는 MLE에 의해 추정

### □ odds ratio

그룹	응답		계
	성공	실패	
treated	30	70	100
control	20	80	100

- treated에 대한 성공 odds :  $30/70=0.43$
- control에 대한 성공 odds :  $20/80=0.25$
- odds ratio :  $0.43/0.25=1.72$

- 일반 모델 :  $OR = \frac{a/b}{c/d} = \frac{ad}{bc} \left( = \frac{a/c}{b/d} \right)$

a	b
c	d

- cross-product ratio라고도 불림

- 특성
  - $0 \leq OR \leq \infty$
  - $OR < 1$  ; 음의 association
  - $OR > 1$  ; 양의 association
  - $OR = 0$  ; no association

□ **multinomial** 로지스틱 회귀모델

- 이항 로지스틱 회귀모델의 일반화는 2이상의 범주를 갖는 응답으로 확장
- $R=3$ , 하나의 공변량  $x$ 에 대해 다음과 같은 2개의 logits을 갖는다

$$\text{logit}(p_1/p_3) = \beta_{01} + \beta_{11}x$$

$$\text{logit}(p_2/p_3) = \beta_{02} + \beta_{12}x$$

□ **순서화** 로지스틱 회귀모델

- $R > 2$  인 범주가 순서화되어 있다면  $R-1$  logits는 다음과 같은 형태의 싱글 logit으로 바꿀 수 있다

$$\text{logit}[P(y \leq j)] = \log\left[\frac{P(y \leq j)}{P(y > j)}\right] = \beta_{0j} - \underline{x}_i^T \underline{\beta}$$

여기서  $j = 1, \dots, R-1$ ,  $\beta_{01} \leq \beta_{02} \leq \dots \leq \beta_{0R-1}$

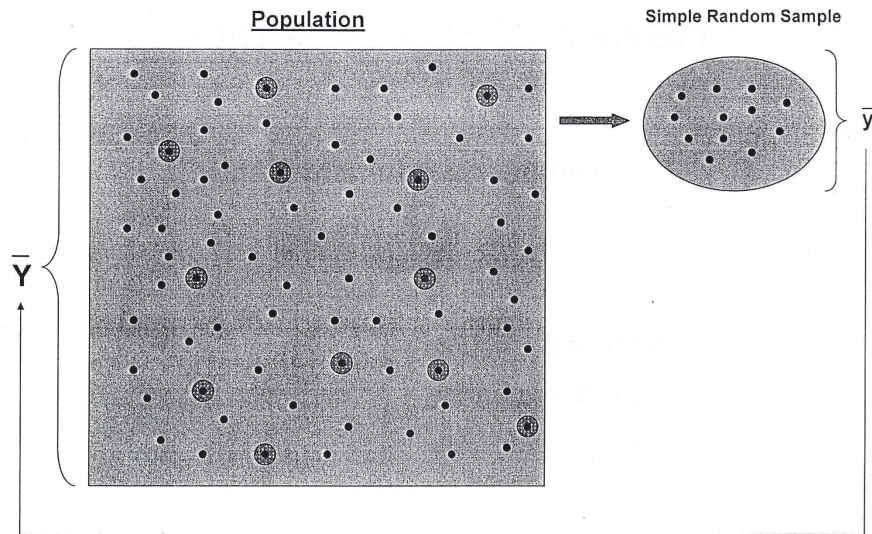
- propotional odds model 또는 cumulative odds model 로도 불림

8. 가중치 및 복합표본설계

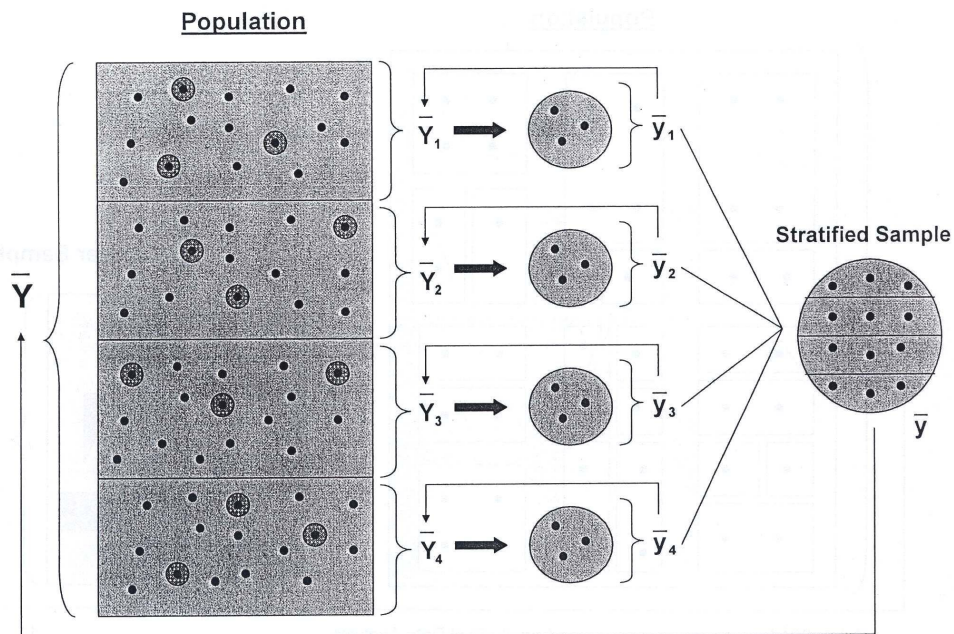
□ 표본설계

- 단순임의표본 vs 복합표본
  - 층화표본
  - 집락표본
  - 조사가중(survey weight)

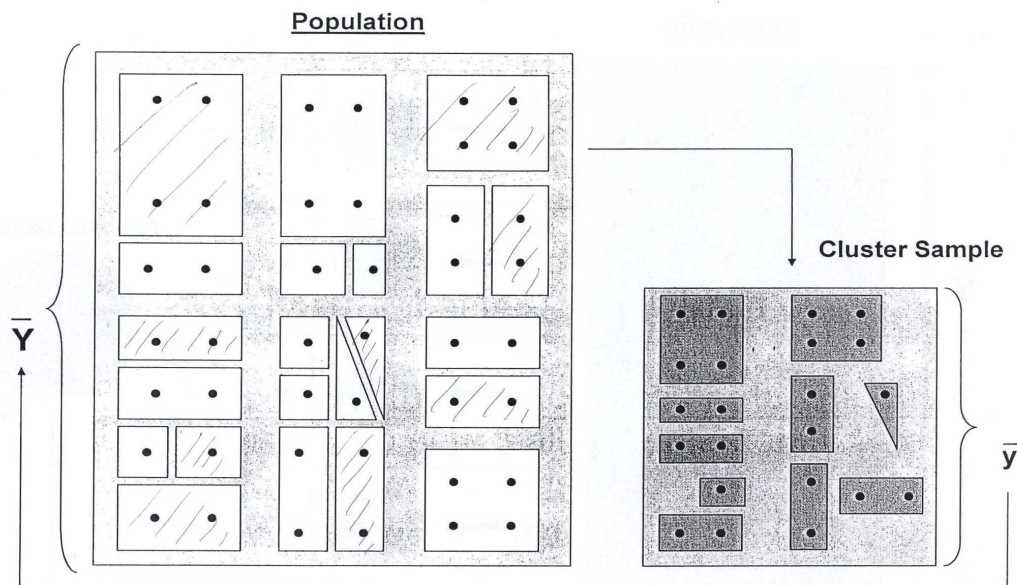
- 단순임의표본 : 모집단내 모든 단위가 동일한 추출확률을 갖음



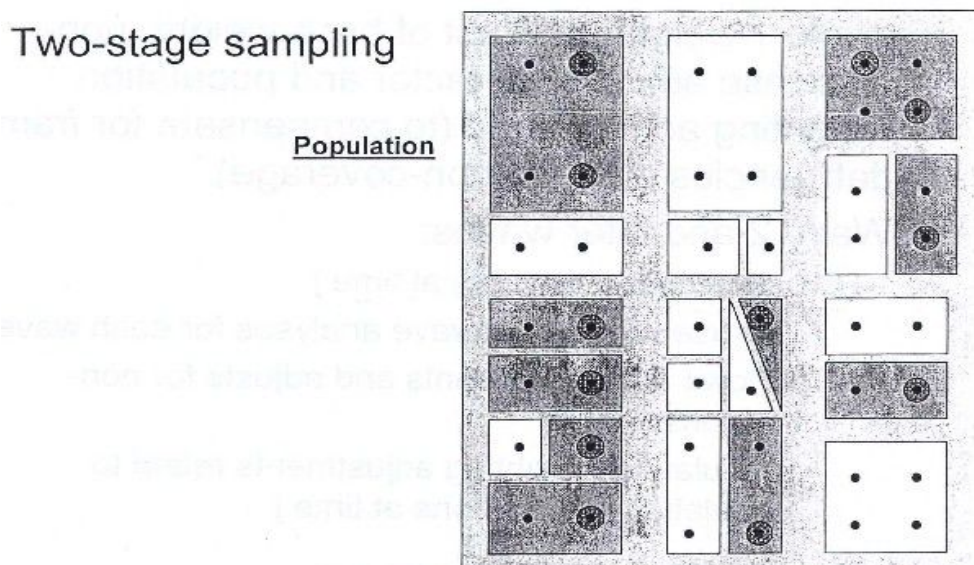
- 모집단이 넓게 흩어져 있을 때, 추출된 표본에 접근하는 데 극단적인 비용이 필요
  - 이용가능한 표본들이 많지 않음
  - 모집단이 동질성이 떨어지고 부그룹의 크기가 서로 다르면 정도가 떨어짐
- 층화표본 : 모집단을 알려진 특성에 의해 층으로 분할하고 각 층에서 독립적으로 표본을 추출



- 부그룹내가 전체 모집단보다 동질성을 보이면 표본오차는 감소하여 효율성이 증대
  - 층내 동질성이 크면 클수록 표본설계의 효율성은 증대
- 집락표본 : 집락은 하나 이상의 모집단 단위로 구성
- 집락내에는 이질성이 높을 수록 표본설계의 효율성은 증대
  - 집락간에는 동질성이 높을 수록 표본설계의 효율성은 증대



- 다단계 표본 : 2단계 표본에서는 1단계로 집락을 추출(PSU)  
 2단계로 집락내에서 단위를 추출(SSU 또는 USU)





□ 가중치

- 기본가중치 : 표본추출률의 역수
- 무응답 조정 가중치
- 보조정보를 활용한 사후가중치
- 종단면 자료 가중치
  - 웨이브 1 가중치 : 기본가중치 및 기타 조정에 의한 가중치
  - 웨이브 2 이상의 가중치
    - ..시간별 횡단면 가중치 : 각 웨이브별 분석에 필요한 가중치
    - ..종단면 가중치 : 웨이브 1과 웨이브 j간의 종단면 분석에 필요한 가중치

☞ 종단면 가중치 수 :  $2^j - 1$ 개

□ misspecification 효과

- 추정량의 분산을 추정할 때 표본설계의 특성을 무시하여 발생하는 효과

$$meff(\hat{\beta}_k; v_0(\hat{\beta}_k)) = \frac{V_{true}(\hat{\beta}_k)}{E_{true}(v_0(\hat{\beta}_k))}$$

$\hat{\beta}_k$  추정량

표본설계를 무시한 분산추정량

- 실제 표본설계에 의한 관심변수의 점 추정량의 분산을 특정 분산 추정량의 기대값으로 나눈 값
- 참 분산을 얼마나 과대추정 또는 과소추정 하는 지를 측정

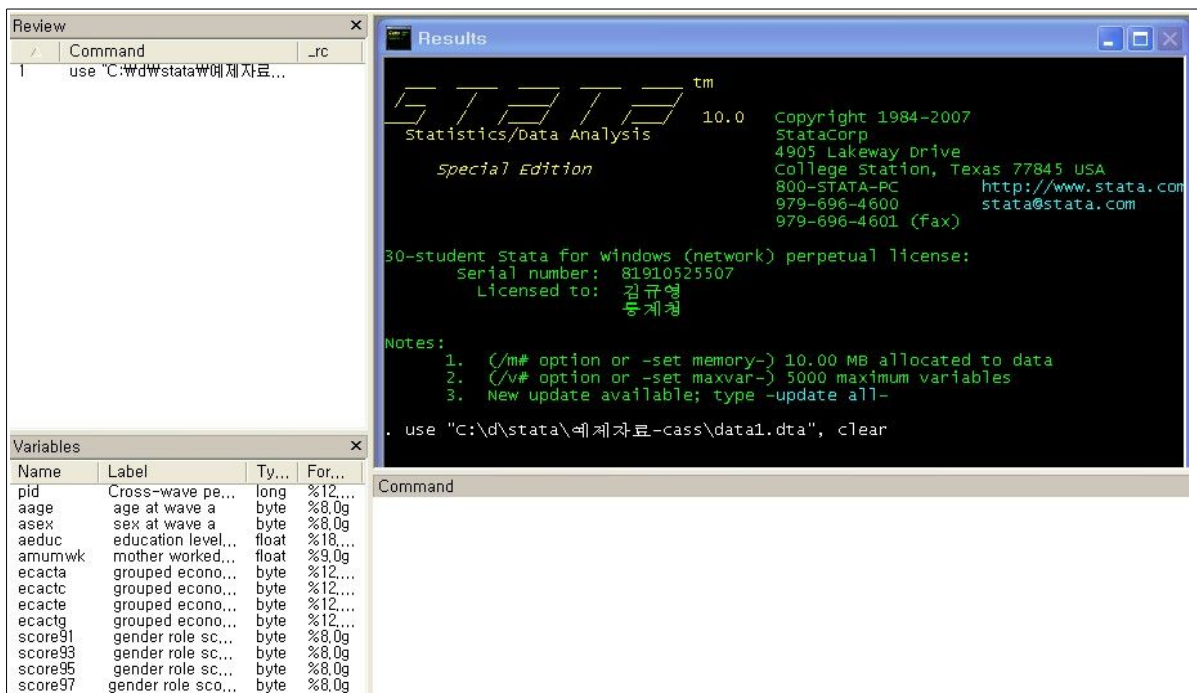
$meff(\hat{\beta}_k, \nu_0(\beta_k))$	의 미
$< 1$	$V_{true}(\hat{\beta}_k)$ 의 과대추정
$= 1$	$V_{true}(\hat{\beta}_k)$ 의 정확한 추정
$> 1$	$V_{true}(\hat{\beta}_k)$ 의 과소추정

- meff는 가중치와/또는 집락이  $\nu_0(\beta_k)$ 에서 무시되면 증대되고, meff는 층화가  $\nu_0(\beta_k)$ 에서 무시되면 감소된다
- 추정 :  $\widehat{meff}(\hat{\beta}_k, \nu_0(\hat{\beta}_k)) = \nu(\hat{\beta}_k) / \nu_0(\hat{\beta}_k)$
- meff는 부정확하게 설명된 분산추정량  $\nu_0(\hat{\beta}_k)$ 의 상대적 편향을 측정하는데 사용

## 9. stata 주요 명령문 및 결과

□ stata 화면 구성 : 4개의 부분으로 구성

- review, variable, results, command



The screenshot displays the Stata software interface with three main windows:

- Review Window:** Shows the command `use "C:\d\stata\예제자료\...`
- Variables Window:** Lists variables with their labels, types, and formats.
 

Name	Label	Ty...	For...
pid	Cross-wave pe...	long	%12...
aage	age at wave a	byte	%8.0g
asex	sex at wave a	byte	%8.0g
aeduc	education level...	float	%18...
amumwk	mother worked...	float	%9.0g
ecacta	grouped econo...	byte	%12...
ecactc	grouped econo...	byte	%12...
ecacte	grouped econo...	byte	%12...
ecactg	grouped econo...	byte	%12...
score91	gender role sc...	byte	%8.0g
score93	gender role sc...	byte	%8.0g
score95	gender role sc...	byte	%8.0g
score97	gender role sco...	byte	%8.0g
- Results Window:** Shows the Stata logo and version information (10.0), copyright (1984-2007), and license details. It also displays notes about memory allocation and available updates.

□ 기초통계량

- 연속변수의 집계 명령어 : sum 변수, detail

```
. sum aage,detail
```

age at wave a					
Percentiles		Smallest			
1%	16		16		
5%	16		16		
10%	17		16	Obs	1429
25%	20		16	Sum of Wgt.	1429
50%	25			Mean	26.55353
		Largest		Std. Dev.	8.485907
75%	31		58		
90%	38		59	Variance	72.01061
95%	43		59	Skewness	1.140139
99%	55		59	Kurtosis	4.372914

- 변수에 빈도분포표 생성 : tab 변수 [변수], row col

```
. tab aeduc asex, row col
```

Key			
<i>frequency</i>			
<i>row percentage</i>			
<i>column percentage</i>			
education level at wave a	sex at wave a		Total
	males	females	
others	369 53.63 47.92	319 46.37 48.41	688 100.00 48.15
high and A levels	401 54.12 52.08	340 45.88 51.59	741 100.00 51.85
Total	770 53.88 100.00	659 46.12 100.00	1,429 100.00 100.00

- wide form자료를 long form자료로 변환 : reshape

```
. reshape long score ecact, i(pid) j(year)
(note: j = 91 93 95 97)
```

Data	wide	->	long
> _____			
Number of obs.	1429	->	5716
Number of variables	13	->	8
j variable (4 values)		->	year
xij variables:			
	score91 score93 ... score97	->	score
	ecact91 ecact93 ... ecact97	->	ecact



- i : 그룹식별자
- j : 그룹내 식별자

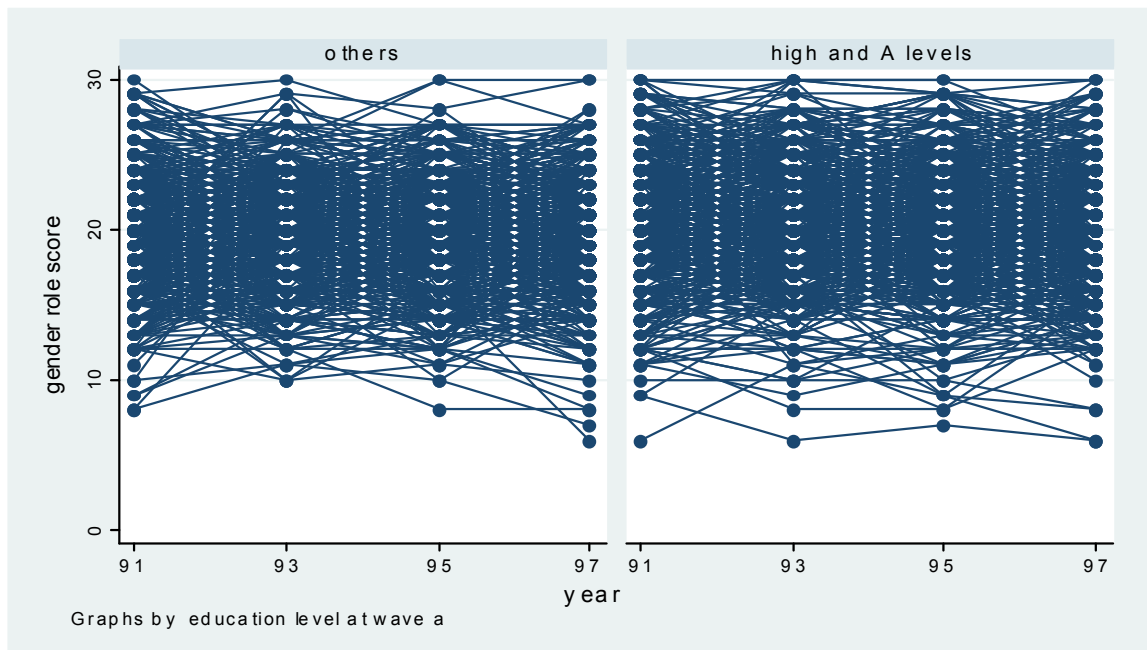
- 종단면 자료에 대한 평균 및 표준편차 등 분석 : **xt**

```
. xtsum score, i(pid)
```

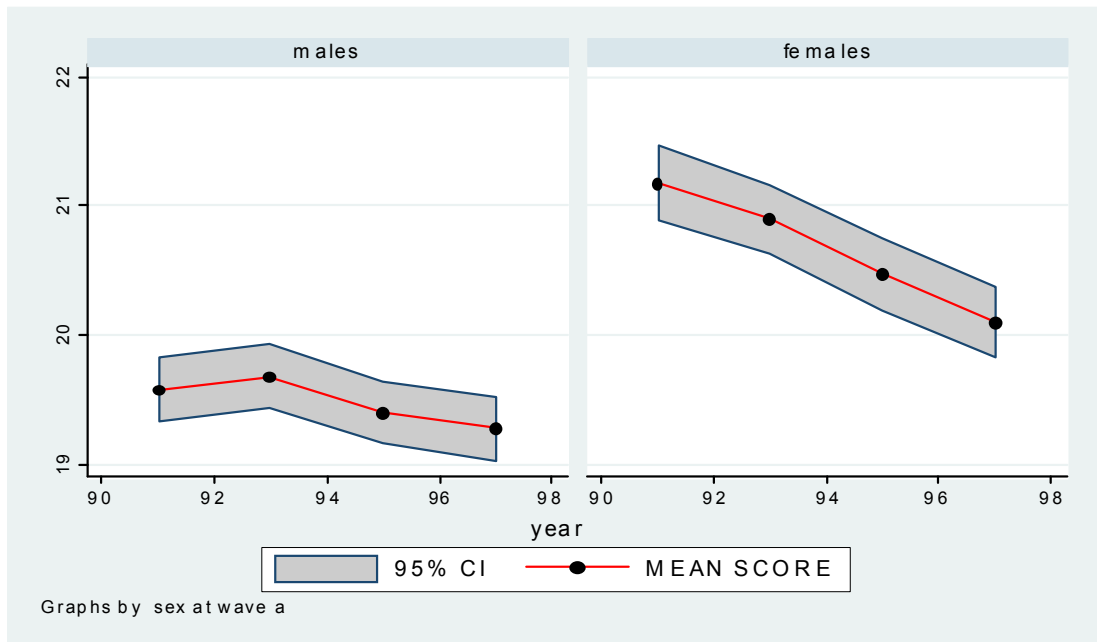
Variable		Mean	Std. Dev.	Min	Max	Observation
score	overall	20.02537	3.545611	6	30	N = 571
	between	2.987781		7	29.5	n = 142
	within	1.910291	10.77537	28.52537		T = 142

- overall : 전체 자료에 대한 통계량
- between : 개체 평균의 변이
- within : 각 개체 평균의 시간에 대한 변이

- 응답자의 profile를 plot : **twoway**



- 변수에 대한 평균과 표준오차의 plot : collapse -> gen -> twoway



- 종단면 자료에 대한 모델 적합 : **xtreg**

- population average model : pa
- exchangeable : default option
- independence(상관관계 없음) : corr(ind)
- ar(1) : corr(ar 1) t(time)

```
. xtreg score time, pa i(pid) corr(ind)
```

Iteration 1: tolerance = 1.24e-16

```
GEE population-averaged model
Group variable:          pid
Link:                    identity
Family:                  Gaussian
Correlation:             independent

Scale parameter:        12.50157

Pearson chi2(16):       71458.96
Dispersion (Pearson):   12.50157

Number of obs          = 5716
Number of groups       = 1429
Obs per group: min = 4
                   avg = 4.0
                   max = 4
Wald chi2(1)          = 30.91
Prob > chi2           = 0.0000

Deviance              = 71458.96
Dispersion            = 12.50157
```

score	Coef.	Std. Err.	z	P> z	[95% Conf. Inte	
time	-.1162701	.0209147	-5.56	0.000	-.1572621	-.0752781
_cons	20.37418	.0782555	260.35	0.000	20.2208	20.52756

- random effect model : re (GLS 사용)
- maximum-likelihood estimation : mle

- likelihood ratio 검증 절차

- xtreg 전체 : 전체 모델 추정
- estimates store full
- xtreg 부분 : 부분 모델 추정
- estimates store reduced
- lrtest full reduced

1 - 로지스틱 회귀모델 적합 : **logit**

```
. xi : logit prhlth i.lsex if year==0
i.lsex          _lsex_1-2          (naturally coded; _lsex_1 omitted)
```

```
Iteration 0:   log likelihood =3696.4963
Iteration 1:   log likelihood =3674.5218
Iteration 2:   log likelihood =3674.3559
Iteration 3:   log likelihood =3674.3559
```

```
Logistic regression                               Number of obs   =   11586
LR chi2(1)                                         =         44.28
Prob > chi2                                        =         0.0000
Pseudo R2                                         =         0.0060

Log likelihood =-3674.3559
```

prhlth	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_lsex_2	.4286796	.0653783	6.56	0.000	.3005405	.5568187
_cons	-2.483857	.05224	-47.55	0.000	-2.586245	-2.381468

- 다변량과 변수간에 상호작용을 감안한 로지스틱

```
. xi : logit prhlth i.lsex*i.lagecat i.tenure i.marstat if year==0
i.lsex          _ilsex_1-2          (naturally coded; _ilsex_1 omitted)
i.lagecat       _ilagecat_1-4       (naturally coded; _ilagecat_1 omitted)
i.lsex*i.lage~t _ilsexlag_#_#       (coded as above)
i.tenure         _itenure_1-3        (naturally coded; _itenure_1 omitted)
i.marstat       _imarstat_1-3       (naturally coded; _imarstat_1 omitted)
```

```
Iteration 0:  log likelihood = -3696.4963
Iteration 1:  log likelihood = -3543.8463
Iteration 2:  log likelihood = -3519.2054
Iteration 3:  log likelihood = -3518.9192
Iteration 4:  log likelihood = -3518.9189
```

```
Logistic regression              Number of obs = 11586
                                LR chi2(1)      = 355.15
                                Prob > chi2     = 0.0000
                                Pseudo R2       = 0.0480

Log likelihood = -3518.9189
```

prhlth	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_ilsex_2	.8001531	.2064913	3.87	0.000	.3954375	1.204869
_ilagecat_2	.8417341	.2035308	4.14	0.000	.442821	1.240647
_ilagecat_3	1.272338	.203099	6.26	0.000	.8742709	1.670404
_ilagecat_4	1.030592	.2331115	4.42	0.000	.573702	1.487482
_ilsexlag~2	-.4567494	.2341859	-1.95	0.051	-.9157453	.0022465
_ilsexlag~3	-.542264	.2329664	-2.33	0.020	-.9988698	-.0856582
_ilsexlag~4	-.4413684	.27067	-1.63	0.103	-.9718717	.089135
_itenure_2	1.058194	.0738955	14.32	0.000	.9133616	1.203027
_itenure_3	.4399116	.1292921	3.40	0.001	.1865039	.6933194
_imarstat_2	.2639955	.0863587	3.06	0.002	.0947356	.4332554
_imarstat_3	.0441424	.1051343	0.42	0.675	-.1619171	.2502019
_cons	-3.666146	.1896225	-19.33	0.000	-4.037799	-3.294493

· or 옵션 : odds ratio와 표본오차를 구할 때 사용

```
. xi : logit prhlth i.lsex if year==0, or
i.lsex          _ilsex_1-2          (naturally coded; _ilsex_1 omitted)
```

```
Iteration 0:  log likelihood = -3696.4963
Iteration 1:  log likelihood = -3674.5218
Iteration 2:  log likelihood = -3674.3559
Iteration 3:  log likelihood = -3674.3559
```

```
Logistic regression              Number of obs = 11586
                                LR chi2(1)      = 44.28
                                Prob > chi2     = 0.0000
                                Pseudo R2       = 0.0060

Log likelihood = -3674.3559
```

prhlth	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
_ilsex_2	1.535229	.1003707	6.56	0.000	1.350589	1.745112

1 - population average와 확률효과 로지스틱 회귀모델을 반복  
 이항 측도에 적합 : **xtlogit**

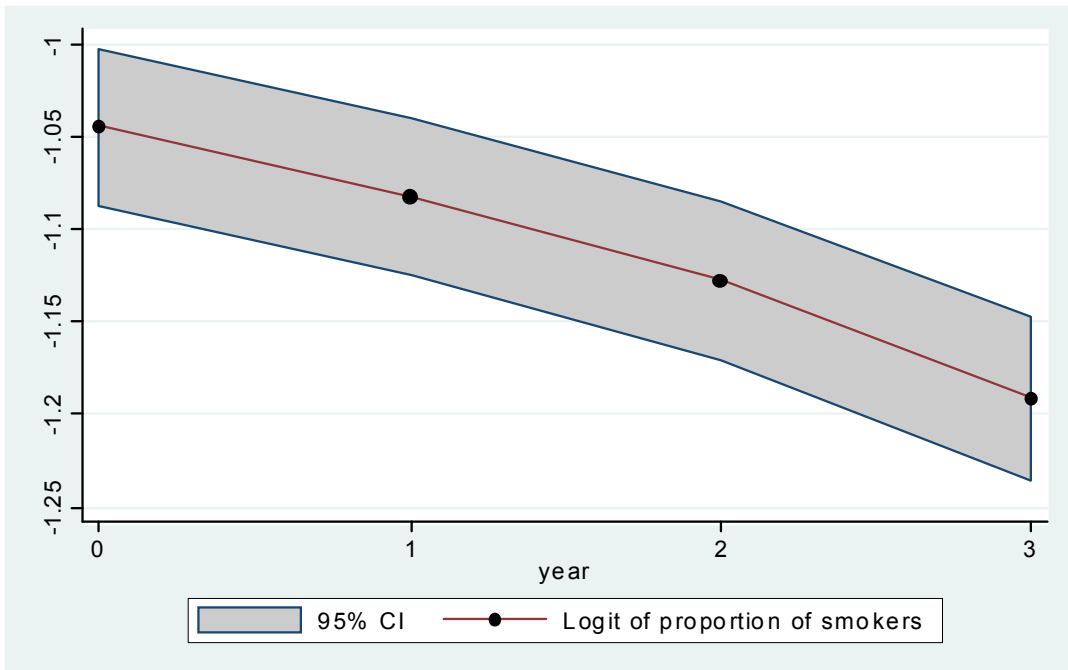
```
. xi: xtlogit smoker i.year, pa i(pid)
i.year          _Iyear_0-3          (naturally coded; _Iyear_0 omitted)
```

Iteration 1: tolerance =  $3.107e-13$

GEE population-averaged model  
 Group variable: **pid** Number of obs = **46344**  
 Link: **logit** Number of groups = **11586**  
 Family: **binomial** Obs per group: min = **4**  
 Correlation: **exchangeable** avg = **4.0**  
 max = **4**  
 Wald chi2(3) = **171.82**  
 Prob > chi2 = **0.0000**  
 Scale parameter: **1**

smoker	Coef.	Std. Err.	z	P> z	[95% Conf. Int]	
_Iyear_1	-.037544	.0115594	-3.25	0.001	-.0601999	-.0148881
_Iyear_2	-.0827898	.0116318	-7.12	0.000	-.1055877	-.0599918
_Iyear_3	-.1464061	.0117478	-12.46	0.000	-.1694314	-.1233808
_cons	-1.044336	.0211719	-49.33	0.000	-1.085833	-1.00284

- 시간에 따른 확률 로짓을 그래프화 절차
  - preserve : 결과를 저장
  - sort year : 자료를 연도별로 정렬
  - collapse (mean) smoker (sd) sdsmloker= smoker (count)  
 n= smoker, by (year) : 연도별로 평균과 표본오차 등을 생성
  - gen high=logit( smoker+2\* sdsmloker/sqrt(n))  
 gen low=logit( smoker-2\* sdsmloker/sqrt(n))  
 : 95% 신뢰수준하의 상하한 구간 계산
  - gen logitmloker=logit(smoker) : 특정 변수의 로짓값에 대한 변수 생성
  - twoway (rarea low high year, bfcolor(gs12)) (connected  
 logitmloker year, mcolor> r(black)), legend(order (1  
 "95% CI" 2 "Logit of proportion of smokers"))  
 : 그래프 생성



1 - 자료를 wide-form으로 변형 : **reshape**

• reshape wide score lrwght, i(pid) j(time) :

```
reshape wide score lrwght, i(pid) j(time)
(note: j = 0 2 4 6 8)
```

Data	long	->	wide
Number of obs.	6700	->	1340
Number of variables	4	->	11
j variable (5 values)	time	->	(dropped)
xij variables:	score	->	score0 score2 ... score8
	lrwght	->	lrwght0 lrwght2 ... lrwght8



· 가중치 사용

```
. xtreg score time ageg2-ageg4 qualif2-qualif5 [pweight=lrwght], pa i(p
```

```
Iteration 1: tolerance.27023509
Iteration 2: tolerance.00518949
Iteration 3: tolerance.00008099
Iteration 4: tolerance1=260e-06
Iteration 5: tolerance1=960e-08
```

```
GEE population-averaged model
Group variable:          pid
Link:                    identity
Family:                  Gaussian
Correlation:            exchangeable
Scale parameter:        13.561
Number of obs           6790
Number of groups        = 1340
Obs per group: min     = 5
                    avg = 5.0
                    max = 5
Wald chi2(8)            = 104.08
Prob > chi2              = 0.0000
```

(Std. Err. adjusted for clustering or

score	Coef.	Semi-robust Std. Err.	z	P> z	[95% Conf. Inter	
time	-.0323522	.016217	-1.99	0.046	-.0641369	-.0005675
ageg2	-.9255226	.19552	-4.73	0.000	-1.308735	-.5423105
ageg3	-1.21556	.2248062	-5.41	0.000	-1.656172	-.7749481
ageg4	-1.393913	.2516362	-5.54	0.000	-1.88711	-.9007149
qualif2	-.5326149	.2450542	-2.17	0.030	-1.012912	-.0523175
qualif3	-.6950972	.2555777	-2.72	0.007	-1.19602	-.1941741
qualif4	-.6125285	.2580184	-2.37	0.018	-1.118235	-.1068217
qualif5	-1.589552	.2697552	-5.89	0.000	-2.118262	-1.060841
_cons	21.99361	.3090711	71.16	0.000	21.38784	22.59938



### III. 기 타

#### 1. 참가자 명단

## Longitudinal Data Analysis

Number of Participants:		25	
Title	Name	Surname	Employer
Mr	Adrian	Antoniang	University College London
Miss	Jennifer	Baird	University of Southampton
Mrs	Nicola	Caloon	Queen's University Belfast
Ms	Emma	Calvert	University of Southampton
Ms	Charlotte	Chuang	Department for Work and Pensions
Mr	Christopher	Chayes	University of Liverpool
Prof	Alan	Edrinal	Cardiff University
Dr	Deborah	Fardon	University of Southampton
Ms	Frida	Geyse Rajoo	University of Southampton
Dr	Pollum	Gore	Lancaster University
Ms	Yudith	Gudlan	University of Bristol
Ms	Munshi	Hag	University of Birmingham
Ms	Jaclyn	Harris	Queen's University Belfast
Miss	Jan	Jamison	The Royal Veterinary College (RVC)
Dr	Azad	Khan	The University of Queensland
Dr	Kyoyoung	Kim	Statistics Korea
Dr	Gloria	Langat	University of Southampton
Dr	David	Lee	The University of Manchester
Dr	Pau	Mona Moyn	University of St. Andrews
Mr	James	Prior	Keele University
Mr	Robert	Sandres	University of Stirling
Ms	John	Sandhu	Audit Commission
Miss	Tania	Smith	University of Edinburgh
Mr	Steven	Wynne	Manchester Metropolitan University
Ms	Hannah	Zagil	University of Edinburgh

## 2. 프로그램 일정

# Longitudinal Data Analysis 30 June to 2 July 2010

## Programme

### Course Presenters

Ann Berrington, University of Southampton

Peter Smith, University of Southampton

Marcel Vieira, Federal University of Juiz de Fora

### Day 1 – Wednesday, 30 June

09.30-10.00	Registration and Coffee
10.00-10.15	Welcome and Overview of the Course
10.15-11.15	<u>Session 1</u> : Introduction to Longitudinal Data
11.15-11.30	Tea/Coffee
11.30-12.45	<u>Session 2</u> : Computing Workshop One: Introduction to Stata
12.45-14.00	Lunch
14.00-14.45	<u>Session 3</u> : Exploring Repeated Measures Data
14.45-15.45	<u>Session 4</u> : Computing Workshop Two: Exploring Longitudinal Data
15.45-16.00	Tea/Coffee
16.00-17.00	<u>Session 5</u> : Approaches to Modelling Repeated Measures Data
17.00-17.30	Optional Computing Time

**Day 2 – Thursday, 1 July**

09.30-10.15	<u>Session 6:</u> Population Average (Marginal) Models
10.15-11.00	<u>Session 7:</u> Computing Workshop Three: Fitting Population Average Models
11.00-11.15	Tea/Coffee
11.15-12.15	<u>Session 8:</u> Random Effects Models
12.15-13.00	<u>Session 9:</u> Computing Workshop Four: Fitting Random Intercept Models
13.00-14.15	Lunch
14.15-15.15	<u>Session 10:</u> Revision of Logistic Regression Models
15.15-15.45	<u>Session 11:</u> Exploring Repeated Binary Measures Data
15.45-16.00	Tea/Coffee
16.00-17.00	<u>Session 12:</u> Computing Workshop Five: Fitting Logistic Regression Models
17.00-17.30	Optional Computing Time

**Day 3 – Friday, 2 July**

09.30-11.00	<u>Session 13:</u> Logistic Regression Models for Repeated Measures Data
11.00-11.15	Tea/Coffee
11.15-12.30	<u>Session 14:</u> Computing Workshop Six: Fitting Logistic Regression Models for Repeated Measures Data
12.30-13.45	Lunch
13.45-15.15	<u>Session 15:</u> Handling Weights and Complex Survey Designs
15.15-15.30	Tea/Coffee
15.30-16.30	<u>Session 16:</u> Computing Workshop Seven: Handling Weights and Complex Survey Designs
16.30-17.00	<u>Session 17:</u> Review

### 3. CASS의 통계 전문 과정

#### CASS (Courses in Applied Social Surveys) Short Course Programme 2010/11

**Southampton**  
UNIVERSITY OF  
Southampton Statistical  
Science Research Institute

**The Psychology of Survey Response**  
*Prof Roger Tourangeau*  
Medical Research Council, London  
10-12 May 2010

**Structural Equation Modelling**  
*Prof Patrick Sturgis*  
University of Southampton  
13-15 October 2010

**Designing Effective Web Surveys**  
*Mick P. Couper*  
Medical Research Council, London  
26-28 January 2011

For further details or to register for a course online, visit the CASS website at:

[www.southampton.ac.uk/cass](http://www.southampton.ac.uk/cass)

Alternatively, please contact:

**Essentials of Survey Design and Implementation**  
*Dr Pamela Campanelli*  
Medical Research Council, London  
19-21 May 2010

**Survey Data Analysis I: Introducing Descriptive and Inferential Statistics**  
*Dr Gabriele Durrant*  
University of Southampton  
9-11 November 2010

**Questionnaire Design**  
*Dr Pamela Campanelli*  
University of Manchester  
10-11 March 2011

CASS Admin Asst  
S3RI, Rm 2101  
Building 58, University of Southampton,  
Southampton,  
SO17 1BJ

Tel: +44 (0)23 8059 5376

Fax: +44 (0)23 8059 5753

Email: [cass@southampton.ac.uk](mailto:cass@southampton.ac.uk)

As places are limited and these courses are popular it is advisable to register as early as possible.

**Course Fees:**

£30 per day for UK registered students

£60 per day for staff from UK academic institutions, ESRC funded researchers and UK registered charitable organisations

£220 per day for all other participants

Changes to the above may occur - please check the CASS website for updates.

**Applied Multilevel Modelling**  
*Dr Ian Brunton-Smith*  
University of Southampton  
16-18 June 2010

**Questionnaire Design**  
*Dr Pamela Campanelli*  
University of Edinburgh  
18-19 November 2010

**Essentials of Survey Design and Implementation**  
*Dr Pamela Campanelli*  
University of Cardiff  
30 March-1 April 2011

**Longitudinal Data Analysis**  
*Prof Peter Smith, Dr Ann Berrington and Dr Marcel Vieira*  
University of Southampton  
30 June-2 July 2010

**Paradata in Survey Research**  
*Dr Frauke Kreuter*  
National Council for Voluntary Organisations, London  
7-8 December 2010

**Regression Methods**  
*Dr Denise Silva*  
University of Southampton  
12-14 April 2011

**Regression Methods**  
*Dr Denise Silva*  
University of Southampton  
29 September-1 October 2010

**Survey Data Analysis II: Introduction to Linear Regression Modelling**  
*Dr Gabriele Durrant*  
University of Southampton  
18-20 January 2011

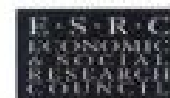
**Survey Data Analysis I: Introducing Descriptive and Inferential Statistics**  
*Dr Gabriele Durrant*  
University of Southampton  
15-17 May 2011

**Introduction to Survey Sampling and Estimation**  
*Dr Pedro Silva*  
University of Southampton  
4-6 October 2010

**Longitudinal Data Analysis**  
*Prof Peter Smith, Dr Ann Berrington and Dr Marcel Vieira*  
University of Southampton  
(dates tbc)

ESRC National Centre for

**Research Methods**



#### 4. 종단면 자료분석을 위한 통계패키지 비교

	Stata (10)		MLwiN	aML	SPSS
	xt	GLAMM			
Random intercept	✓	✓	✓	✓	✓
Random coefficient	✓	✓	✓	✓	✓
3+ levels	✓	✓	✓	✓	✓
Survey weights	×	✓	✓	×	×
Categ. responses	✓	✓	✓	✓	✓
Multi-process	×	✓	✓	✓	×

	Stata v10 (xt)	Sudaan
Survey weights	✓	✓
Clustering	×	✓
Stratification	×	✓
Categ. responses	✓	✓

STATA v.11 seems to have the capability to adjust the model coefficient standard errors for clustering.

SPSS v.15 and onwards include some facilities for GEE within the Advanced Models add-on.

## 5. 참고 자료

### General Texts

Diggle, P. J., Heagerty, P., Liang, K-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data. Second Edition*. Oxford: Oxford University Press. ✓

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004) *Applied Longitudinal Analysis*. Hoboken, New Jersey: Wiley.

Hand, D. and Crowder, M. (1996) *Practical Longitudinal Data Analysis*. London: Chapman and Hall.

Lindsey, J. K. (1999) *Models for Repeated Measurements*. Oxford: Oxford University Press.

Rose, D., ed. (2000) *Researching Social and Economic Change: the Uses of Household Panel Studies*. London: Routledge.

Singer, J. D. and Willett, J. B. (2003) *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford: Oxford University Press. Chapter 2 Exploring Longitudinal Data on Change.

Tatsk, J. W. R. (2003) *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge: Cambridge University Press.

### Using Stata

Rabe-Hesketh, S. and Everitt, B. S. (2007) *Handbook of Statistical Analysis Using STATA. (Fourth edition)*. Boca Raton, FL: Chapman & Hall. ✓

Rabe-Hesketh, S. and Skrondal, A. (2005) *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press. ✓

### Random Effects Models

Bryk, A. S. and Raudenbush, S. W. (1987) Application of hierarchical linear models to assessing change. *Psychological Bulletin* 101, 147-158.

Goldstein, H. (2003) *Multilevel Statistical Models. (Third Edition)*. London: Arnold.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling. Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall.

Snijders, T. A. B. and Bosker, R. J. (1999) *Multilevel Analysis*. London: Sage.

### Missing Data

Diggle, P. J., Heagerty, P., Liang, K-Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data. Second Edition*. Oxford: Oxford University Press.

Diggle, P. J., Farewell, D. and Henderson, R. (2007). Longitudinal data with dropout: objectives, assumptions and a proposal (with Discussion). *Applied Statistics*, 56, 499-550.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis With Missing Data, Second Edition*. Hoboken, N.J: John Wiley & Sons.

Schafer J. L., Graham J. W. (2002) Missing data: our view of the state of the art. *Psychological Methods*, 7, 147–177.

### **Logistic Regression Models**

Agresti, A. (2002) *Categorical Data Analysis, Second Edition*. New York: Wiley.

Agresti, A. (2007) *An Introduction to Categorical Data Analysis, Second Edition*. New York: Wiley.

Hosmer, D. W. and Lemeshow, S. (2000) *Applied logistic Regression, Second Edition*. New York: Wiley.

### **Other Non-linear Models**

Berrington, A., Hu, Y., Ramirez-Ducouing, K. and Smith, P. W. F. (2005) Multilevel modelling of repeated ordinal measures: an application to attitudes to divorce. Southampton, UK, Southampton Statistical Sciences Research Institute, 24pp. SSRi Applications and Policy Working Papers, A05/10.

Crouchley, R. and Davies, R. B. (1999) A comparison of population average and random-effect models for the analysis of longitudinal count data with base-line information. *Journal of the Royal Statistical Society, Series A*, 162, 331-347.

Long, J. S. and Freese, J. (2006) *Regression Models for Categorical Dependent Variables using Stata, Second Edition*. College Stations, Texas: Stata Press.

McCullagh, P. (2007) Proportional odds model. In *Encyclopedia of Biostatistics, Second Edition*. Wiley.

### **Multiprocess Models**

Lillard, L. and Waite, L. (1983) A joint model of marital childbearing and marital disruption. *Demography*, 30: 653-681.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling, Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman & Hall.

### **Survey Sampling and Weighting**

Lehtonen, R. and Pahkinen, E. J. (1996) *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: John Wiley & Sons.

### **A few substantive examples**

#### Using BHPS

Wiggins, R. D., Schofield, P., Sacker, A., Head, J. and Bartley, M. (2004) Social position and minor psychiatric morbidity over time in the British Household Panel Survey. *Journal of Epidemiology and Community Health*, 58, 779–787.

#### Dealing with problem of "stayers"

Griffiths, P., Brown, J. and Smith, P. (2004) A comparison of univariate and multivariate models for repeated measures of use of antenatal care in Uttar Pradesh. *Journal of the Royal Statistical Society*, 167, 597-611.

Yang, M., Goldstein, H. and Heath, A. (2000) Multilevel models for repeated binary outcomes: attitudes and vote over the electoral cycle. *Journal of Royal Statistical Society, A*, 163, 49-62.

#### Comparison of random effects and population average approach

Carière, I & Bouer, J. (2002) Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC Medical Research Methodology*, 2, 15-25.

Hu et al. (1988) Comparison of Population-Averaged and Sub-Specific Approaches for Analyzing Repeated Binary Outcomes. *American Journal of Epidemiology*, 147, 694-703.