

# STATA 기초

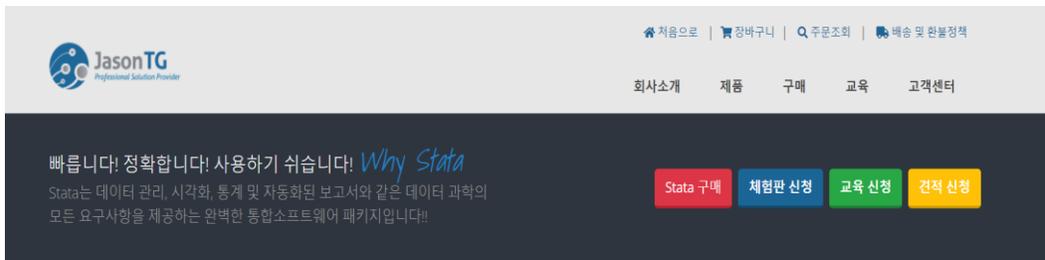
김기만, [kcskgm@gmail.com](mailto:kcskgm@gmail.com)

# STATA란?

- 1980년대 중반 미국의 STATCorp이 개발한 통계소프트웨어
- Statistics + Data의 합성어
- 업무, 학습용/ 경제학, 사회학, 정치학, 의학 등의 분석에 사용
- 미국, 유럽 등지에서 다양한 유저모임 개최
- 한국에도 STATA학회가 있음
- 인터넷에서 STATA관련 공부를 하려면 필요한 사이트
  - STATA faq's <http://www.stata.com/support/faqs/>
  - Resources for learning STATA
  - The STATA listserv <http://www.stata.com/statalist/>
  - UCLA's resources for learning STATA <http://www.ats.ucla.edu/stat/stata/>

# Stata체험 및 구독

- Stata는 상용 S/W, 이에 반하여 R은 freeware임
- Stata를 체험하기 위하여서는 한국총판사이트(<https://www.jat.co.kr/>)에 접속하여서 체험판신청을 하시면 됩니다.



- 체험판을 신청하면 2-3일후에 메일로 1달 정도 사용할 수 있는 라이선스 등이 옵니다
- 사용후 필요성이 있으면 별도로 구독서비스로 구독



⚠ 체험판 신청 후 발송까지 1~2 영업일이 소요됩니다.

성(영문) \*  이름(영문)\*

이메일(소속 이메일) \*  이메일(자주쓰는 이메일) \*

학교/기관명(영문) \*  학교/부서명(영문) \*

직업/학위과정 \*  연락처 \*

당사 홈페이지를 어떻게 알게 되었나요? \*

지인 소개  네이버 or 구글 검색  본사(stata.com)의 내용확인

기타

체험판을 요청하는 목적은? \*

# Stata체험 및 구독(결과화면)

The screenshot displays the Stata/MP 18.0 interface. The main window shows the Stata logo and the following text:

**18.0**  
*MP-Parallel Edition*

Statistics and Data Science  
Copyright 1985-2023 StataCorp LLC  
StataCorp  
4905 Lakeway Drive  
College Station, Texas 77845 USA  
800-STATA-PC <https://www.stata.com>  
979-696-4600 [stata@stata.com](mailto:stata@stata.com)

Stata license: Single-user 2-core , expiring 18 Apr 2024  
Serial number: 501809373297  
Licensed to: kim giman  
              korea ogpc

Notes:

1. Unicode is supported; see [help unicode\\_advice](#).
2. More than 2 billion observations are allowed; see [help obs\\_advice](#).
3. Maximum number of variables is set to 5,000 but can be increased; see [help set\\_maxvar](#).

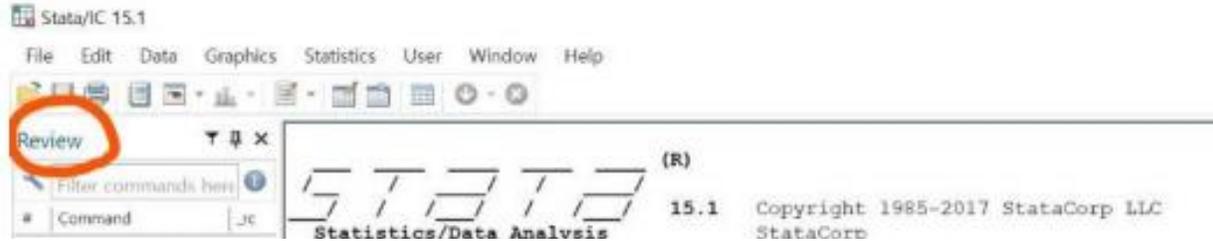
The interface also includes a command window on the left with the text "표시할 항목이 없습니다." (No items to display) and a right-hand panel with sections for "변수" (Variables) and "속성" (Properties). The "속성" section shows the following settings:

속성	값
변수	
이름	
라벨	
타입	
서식	
값라벨	
노트	
데이터	
프레임	default
파일명	
라벨	
노트	
변수	0
관측치	0
크기	0
메모리 사용량	64M
정렬기준	

# STATA소개

## Variable 창

- 파일에 포함되어 있는 변수 이름 등을 보여줌
- 변수이름은 영어, 숫자, 한글 등 사용가능
- (될 수 있는 한 영어와 숫자로 사용)
- 변수이름 첫 글자를 숫자로 시작하면 안됨,
- 띄어쓰기 안됨
- 변수명 입력시 커서를 변수명 올려놓고 더블 클릭



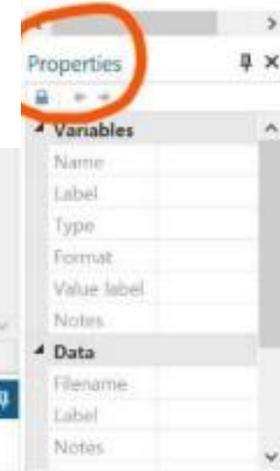
## Review창

- Command 창에서 입력한 명령문들이 모이는 장소
- 명령문의 실행 이력을 한눈에 볼 수 있음
- 명령문을 클릭하면 command창에 명령문이 그대로 입력
- 에러 메시지를 가진 명령문은 빨간색으로 표기
- 명령문 저장(do file)- 오른쪽 마우스 클릭 > save All or Save Selected



## Result 창

Stata명령문이 실행되어 나타나는 창  
결과를 복사하여 사용할 수 있음



## Command 창

- 명령어를 입력하는 창(대소문자 구분)
- 붙여넣기, 잘라내기 등도 가능



# Stata 출력 결과물을 활용하기 (XML을 이용)

- 회귀모형의 추정결과를 논문이나 보고서에서 바로 사용할 수 있는 표로 만들어주는 명령어는 Stata에서 매우 다양하게 제공하고 있음
- 웹페이지 작성에 유용하게 사용되는 xml 파일형식으로 추정결과를 저장해 주는 xml\_tab 명령어의 활용법이 있음
- xml\_tab을 사용하면 사용자는 출력 테이블의 요소에 다양한 형식을 적용할 수 있음
- xml\_tab은 결과를 e()에 저장하는 Stata 추정 명령 후에 계수, 표준 오류, t 및 p 값, 요약 통계 등의 형식화된 테이블을 생성 가능
- 단일 추정 또는 행렬의 결과를 테이블에 저장하고, 저장된 여러 추정 또는 행렬을 하나의 테이블로 결합하고, 여러 테이블을 XML 통합 문서의 다른 시트로 출력할 수 있음

## 예시)

```
.sysuse auto
.regress price rep78 length mpg
.estimates store reg1
.regress price rep78 length mpg turn if foreign==1
.estimates store reg2
.xml_tab reg1 reg2
```

reg price rep78 length mpg

Source	SS	df	MS			
Model	156427769	3	52142589.5	Number of obs	=	69
Residual	420369190	65	6467218.31	F(3, 65)	=	8.06
Total	576796959	68	8482308.22	Prob > F	=	0.0001
				R-squared	=	0.2712
				Adj R-squared	=	0.2376
				Root MSE	=	2543.1

	price	Coefficient	Std. err.	t	P> t	[95% conf. interval]
rep78	698.2603	341.0836	2.05	0.045	17.06941	1379.451
length	30.68091	22.83613	1.34	0.184	-14.926	76.28781
mpg	-178.1566	90.21653	-1.97	0.053	-358.3315	2.018202
_cons	1783.936	6011.455	0.30	0.768	-10221.77	13789.64

. regress price rep78 length mpg turn if foreign==1

Source	SS	df	MS			
Model	63621799.1	4	15905449.8	Number of obs	=	21
Residual	35033555.5	16	2189597.22	F(4, 16)	=	7.26
Total	98655354.6	20	4932767.73	Prob > F	=	0.0016
				R-squared	=	0.6449
				Adj R-squared	=	0.5561
				Root MSE	=	1479.7

	price	Coefficient	Std. err.	t	P> t	[95% conf. interval]
rep78	-342.1465	586.7731	-0.58	0.568	-1586.05	901.7569
length	134.488	41.75029	3.22	0.005	45.98132	222.9946
mpg	-20.38885	79.98716	-0.25	0.802	-189.954	149.1764
turn	37.72369	306.5909	0.12	0.904	-612.2201	687.6675
_cons	-15796.21	11483.62	-1.38	0.188	-40140.4	8547.986

reg1

reg2

	reg1		reg2	
	coef	se	coef	se
Repair record 1978	698.260**	341.084	-342.147	586.773
Length (in.)	30.681	22.836	134.488***	41.750
Mileage (mpg)	-178.157*	90.217	-20.389	79.987
Turn circle (ft.)			37.724	306.591
_cons	1,783.936	6,011.455	-15,796.207	11,483.623

note: .01 - \*\*\*; .05 - \*\*; .1 - \*;

# Stata출력 결과물을 활용하기(한글의 경우)

- Command창에서 명령어를 실행
- 결과 창에 실행결과가 나타남
- 실행결과를 블록으로 설정
- 오른쪽 마우스를 클릭하여 Copy table을 선택
- 한글 프로그램에 들어가 붙여넣기 실행
- 붙여넣기 한 부분을 블록으로 설정
- 한글프로그램 상단 메뉴에서 표 아이콘 클릭
- 표를 정리

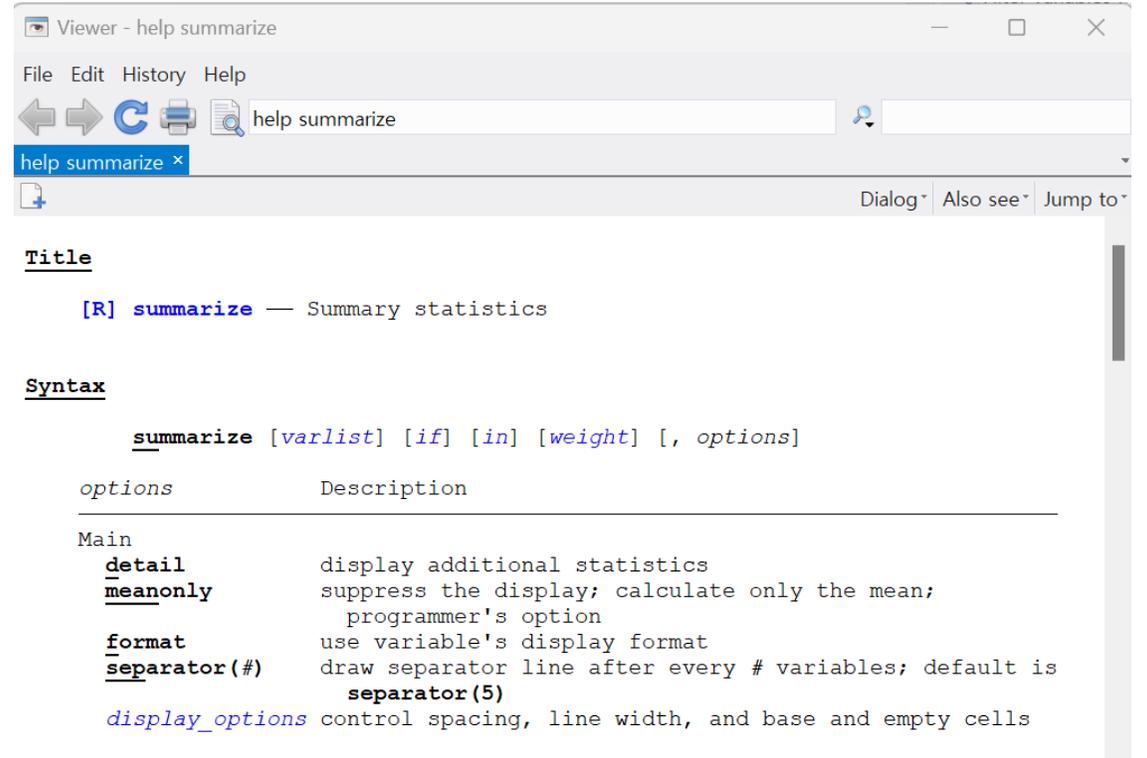
# Stata출력 결과물을 활용하기(MS word)

- Command창에서 명령어를 실행
- 결과 창에 실행결과가 나타남
- 실행결과를 블록으로 설정
- 오른쪽 마우스를 클릭하여 Copy table을 선택
- MS-word프로그램에서 붙여넣기
- 붙여넣기 한 표를 다시 블록설정
- Word상단 메뉴에서 표를 선택하고 '텍스트를 표로'를 선택
- 표정리

# Help 명령어 이용하기

- 알고 싶은 명령어 앞에 `help`를 입력하다.
- `help summarize`
- 명령어에 밑줄친 부분은 명령어의 줄임표현이 가능하다는 것임

예) `summarize => su`,  
`describe => desc`



The screenshot shows a window titled "Viewer - help summarize". The window contains the following text:

```
File Edit History Help
help summarize
help summarize x
Dialog Also see Jump to
```

**Title**

```
[R] summarize — Summary statistics
```

**Syntax**

```
summarize [varlist] [if] [in] [weight] [, options]
```

options	Description
<b>detail</b>	display additional statistics
<b>meanonly</b>	suppress the display; calculate only the mean; programmer's option
<b>format</b>	use variable's display format
<b>separator (#)</b>	draw separator line after every # variables; default is <b>separator (5)</b>
<b>display_options</b>	control spacing, line width, and base and empty cells

# Stata 작업폴더 변경

- 메인화면의 좌측 하단에서 현재 설정된 작업 폴더 경로를 확인합니다
- 작업 폴더는 상단 메뉴의 [파일 > 작업 폴더 변경], `cd` 명령문으로 변경할 수 있습니다
- 참고 1. Stata를 재실행하면 다시 기본 작업 폴더로 돌아갑니다

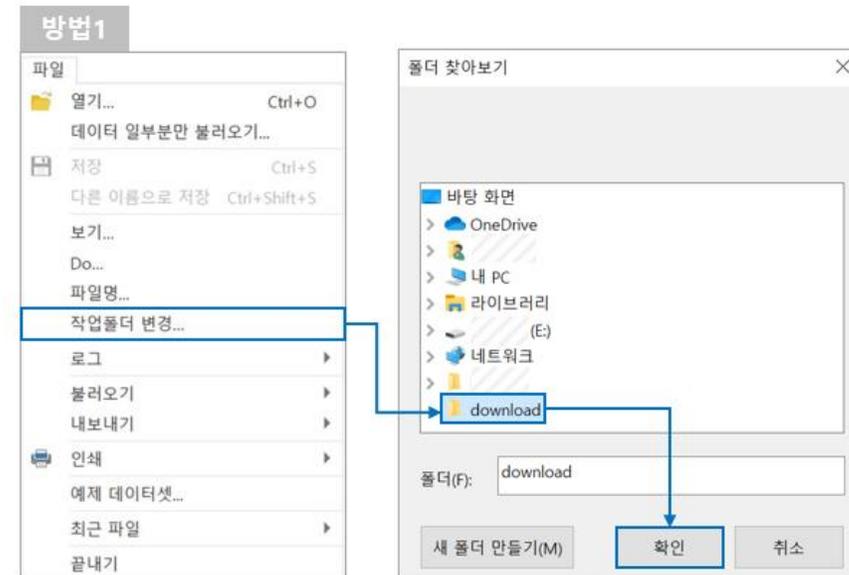
작업 폴더에 관한 아래와 같은 에러 메시지가 발생할 경우...

"cannot write in directory C:\Users\사용자이름\documents\STUP\_XXXX"

"cannot find Stata directory. the file "isstata.140" is missing"

(cannot write in directory라는 메시지에는 모두 적용)

Stata의 작업 폴더 경로의 한글을 영문으로 수정하여 다시 실행합니다.



# 자료가져 오기

- Stata형식의 파일인 \*.dta을 가져오는 것임
- 메뉴에서 열기를 선택한 후 해당 파일을 선택하거나 use명령어를 입력한 후 파일이름을 선택하면 된다.
  - ✓ Use hweight, clear
- 일부자료만을 사용하여 분석할 수 있다. 이러한 제한을 위하여 if, in이라는 명령어를 사용한다.
  - ✓ if 변수 관계연산자와 논리연산자를 사용할 수 있다
  - ✓ 관계연산자는 : ==, !=, >, <, >=, <=
  - ✓ 논리연산자는 : &, |, !(~)
  - ✓ 예) list mpg weight if mpg > 25,
  - ✓ List mpg weight if mpg >= 25 & mpg < 30
  - ✓ List mpg weight if (mpg > 25 | mpg < 10)
  - ✓ List mpg weight if foreign != 1
- ✓ If expression은 조건을 달아 자료의 일부만 보거나 가공하지만, in range는 범위를 제한한다(list mpg weight in 1/10, 1번부터 10번까지 차 무게값을 나열해 보라는 것임)

# 외부파일 가져오기(엑셀 등의 자료)

- Stata에서 데이터 편집기를 사용하여 데이터를 입력하고 편집할 수도 있으나 데이터를 편집하고 이를 가공하는 것은 엑셀 등이 유용
- 데이터를 스프레드시트 형식으로 되어 있다면 엑셀을 열어서 블록형태로 복사하고 붙여넣기 하는 방법도 있음
- 데이터를 stata에서 메뉴선택이나 import명령어를 이용하여 파일을 직접 불러올 수도 있음(예전에는 insheet라는 명령어를 사용하였으나 지금은 import로 대체됨)
- 메뉴에서 불러오는 것은 명령어를 프로그램한 형태이므로 이를 이해하는 것이 좋음
  - ✓ 예 : `import excel xxxx.xlsx, sheet("Sheet1") firstrow clear` (작업디렉토리에 있는 xxxx.xlsx파일의 데이터를 불러들이고 첫번째 행은 변수이름 이름이며, sheet1의 데이터를 불러들인다)
  - ✓ `import excel`에서 범위를 제한적으로 할 경우에는 `cellrange(A4:O45)` 이런 형식으로 , 뒤에 추가하면 된다.

# 외부파일 가져오기(CSV, txt 자료)

- csv화일을 가져오는 것으 엑셀이라는 명령어보다. 다양한 옵션을 사용한다. 명령어에 대한 help를 보면 알 수 있다.
- import delimited (see [D] import delimited)
  1. import delimited reads text-delimited files.
  2. The data can be tab-separated or comma-separated. A custom delimiter may also be specified.
  3. An observation must be on only one line.
  4. The first line of the file can optionally contain the names of the variables.

# 외부파일 가져오기(spss파일 자료)

- Spss로 작성한 자료를 stata로 분석할 경우 엑셀로 파일을 가져오는 것
- 명령어는 import spss 파일이름.sav
- 버전 16 이상의 IBM SPSS Statistics(.sav) 파일 또는 버전 21 이상의 압축된 IBM SPSS Statistics(.zsav) 파일 사용 가능
- import spss는 한 번에 최대 32,766개의 변수를 가져올 수 있음(Stata/BE에서는 최대 2,048개)
- SPSS 파일에 이보다 더 많은 변수가 포함되어 있는 경우 SPSS 파일을 여러 Stata 데이터세트로 나누어야 함
- 명령어로 입력하여도 되지만, 메뉴에서 선택하여도 됨

# Stata시작하기 – Do파일

- Do file은 stata명령어를 기록한 텍스트 파일, 동 파일을 실행하면 동일한 결과를 얻을 수 있음
- 주석은 별표(\*)를 해당 줄의 앞에 표시하거나 이중사선(//)을 넣으면 됨
- /\* \*/은 주석으로 인식됨
- Do editor를 열고자 할 경우에는 메뉴에서 선택하여도 되고 CTRL + 9를 동시에 입력하여도 됨

myjob.do then might read

```
-----begin myjob.do-----  
* a sample analysis job  
version 13  
use http://www.stata-press.com/data/r13/census5  
/* obtain the summary statistics: */  
tabulate region  
summarize marriage_rate divorce_rate median_age if state!="Nevada"  
-----end myjob.do-----
```

or equivalently,

```
-----begin myjob.do-----  
// a sample analysis job  
version 13  
use http://www.stata-press.com/data/r13/census5  
// obtain the summary statistics:  
tabulate region  
summarize marriage_rate divorce_rate median_age if state!="Nevada"  
-----end myjob.do-----
```



- Do file을 실행하는 방법은 다양 : 윈도우 커맨드에서 입력, 메뉴에서 선택, Do file내에서 불러올 수 있음(windows 경우)
- "C:\Program Files\Stata18\Stata" /b do myjob or "C:\Program Files\Stata18\Stata" /s do myjob

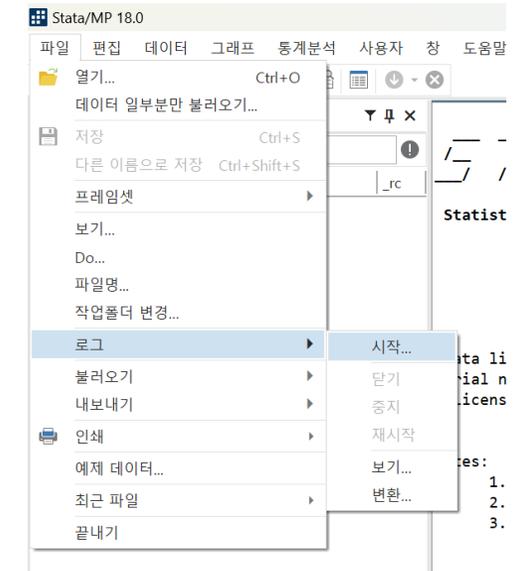
# Do file 작성하기

- Stata를 사용하면서 지속적으로 do file을 생성한다. 파일을 생성할 때에는 이름을 의미있게 작성하는 것이 좋다.
- 화일이름이 의미있게 작성되어도 시간이 흐르고 나면 무엇을 위한 do file인지 알 수가 없다. 그래서 주석을 붙이는 것을 이용하여 다음과 같이 정보를 만들어 주면 좋다. 나는 정보시스템을 개발한 경력이 있어서 그 내용을 담아서 do file을 작성한다.
- ```
/*-----*/
```
- ```
/* 화일생성일자 : 2024. 4. */
```
- ```
/* 화일생성이유 : 규제 비용분석을 위한 패널데이터 분석 */
```
- ```
/* 분석일자 : 2024.4 */
```
- ```
/* 원시자료 : panel_regulation.dta */
```
- ```
/*-----*/
```
- 분석단계별로 명령문에 대한 설명문을 작성한다. 데이터를 분석하는 과정을 단계별로 정리해 두면 추후에 do file만 보면 분석이 어떻게 이루어 졌는지 파악할 수 있다.
- ```
/*회귀분석을 실시, 종속변수 mpg, 독립변수 price -----*/
```

# Stata 시작하기 – log파일

- Stata가 그래프를 제외하고 결과 창으로 전송되는 모든 항목의 복사본을 생성
- 모든 출력을 저장하는 데 도움이 됨. do 파일과 같은 방식으로 저장하지는 않지만 명령도 유지
- 로그 파일을 생성하려면 "파일" -> "로그" -> "시작"으로 이동
- Stata의 기본값은 .smcl 확장자로 파일을 저장(stata에서 로그파일을 열수 있지만, 다른 프로그램에서 읽을 수 없음)
- 다른 프로그램에서 사용 가능한 다른 확장자는 .log
- .log 파일로 저장하려면 대화 상자의 "파일 형식" 메뉴에서 Stata Log 옵션을 선택하면 됨
- 로그 파일을 시작한 후에는 언제든지 일시 중지하고 나중에 다시 시작할 수 있음
- "파일" -> "로그" -> "일시 중지"(또는 "재개")로 이동하여 이 작업을 수행

```
. log using session
-----
name: <unnamed>
log: C:\example\session.smcl
log type: smcl
opened on: 17 Mar 2013, 12:35:08
. use http://www.stata-press.com/data/r13/census5
(1980 Census data by state)
. tabulate region [freq=pop]
+-----+-----+-----+-----+
| Census |      |      |      |
| region | Freq. | Percent | Cum. |
+-----+-----+-----+-----+
NE	49,135,283	21.75	21.75
N Cntrl	58,865,670	26.06	47.81
South	74,734,029	33.08	80.89
West	43,172,490	19.11	100.00
+-----+-----+-----+-----+			
Total	225,907,472	100.00	
+-----+-----+-----+-----+			
. summarize median_age			
+-----+-----+-----+-----+			
Variable	Obs	Mean	Std. Dev.
+-----+-----+-----+-----+			
median_age	50	29.54	1.693445
+-----+-----+-----+-----+
. log close
name: <unnamed>
log: C:\example\session.smcl
log type: smcl
closed on: 17 Mar 2013, 12:35:38
```



# Stata 주요 명령어(1)

- use : Stata의 데이터 형식인 .dta 파일을 불러오는 명령어
  - ✓ use 명령어 뒤에 (파일경로가 포함된)파일명\*을 입력합니다. 만약 인터넷 상에 파일이 있다면 url을 입력합니다.
  - ✓ \* 파일명만 입력하면, 현재 설정된 작업 폴더가 파일 경로가 - 예: C:\Users\Documents
  - ✓ use "파일명", clear use "https://www.stata-press.com/data/r16/nlswork", clear
  - ✓ 명령문 뒤에 clear 옵션을 붙여주면 기존 데이터가 있을 경우 기존 파일을 삭제하고 use 명령문의 파일로 교체합니다. 파일명(또는 파일 경로)를 큰 따옴표("")로 묶어 입력하면, 한글이나 공백이 포함돼도 에러가 발생하지 않습니다
- su (=summarize) : 명령어 su 통해서는 변수들의 관측수, 평균, 표준편차 등에 관한 요약통계량 제공
  - 변수들의 관측수(Obs), 평균(Mean), 표준편차(Std. Dev.) 최소값(Min), 최대값(Max) 에 관한 정보를 알 수 있음
- des (describe) : 변수들의 특성, 타입, 레이블 등을 표시 명령어
  - 데이터파일의 디렉터리 위치, 관측치개수, 변수 개수, 변수들의 저장형태, 디스플레이 포맷, 레이블 등의 기본정보를 제공해 변수특성을 확인
- table : 범주형 변수와 관련된 요약 통계량을 제공하는 명령어
  - table 변수명
- tabulate(=tab) : 범주형 변수의 빈도표를 만드는 가장 기본적인 명령어로, 명령어 gen 을 옵션으로 더미변수로 만들수 있음
  - tab foreign, gen(foreign\_) : 변수 foreign에 대해 2개의 더미변수 foreign\_1, foreign\_2 를 추가로 생성함.

# Stata 주요 명령어(2)

- **변수의 형태를 변환하는 명령어 encode, decode**
- 명령어 encode : 문자변수 → 숫자변수로 변환하는 명령어
- Command 창에 encode make, gen(make\_) 라고 입력하면, make\_ 라는 새로운 숫자변수를 만들고, 문자변수인 make 의 값들을 make\_ 라는 새 변수행에 숫자변수로 변환하라는 뜻
- 명령어 decode : 숫자변수 → 문자변수로 변환하는 명령어 (숫자변수중 value label이 붙은 경우에만 decode 사용 가능)
- 명령어 encode와 달리 명령어 decode는 모든 숫자변수를 문자변수로 변환 할 수 있는 것은 아니고, **숫자변수 중 Value label 이 있는 숫자변수만 문자변수로 변환**
- Command 창에 decode foreign, gen(foreign\_1) 라고 입력하면, foreign\_1 라는 새로운 문자변수를 만들고, 숫자변수인 foreign 의 값들을 foreign\_1 라는 새 변수행에 문자변수로 변환하라는 뜻

# Stata 주요 명령어(3)

- **데이터를 정렬하는 명령어 sort, gsort, bysort**
- 명령어 `sort` : 변수의 값에 줄을 세우는 명령어로 `sort`는 오름차순으로만 정렬이 되고 숫자, 알파벳, 한글 모두 적용됨
- 명령어 `gsort` : 변수의 값을 오름차순, 내림차순(변수앞에 "-" 붙임)으로 정렬시키는 명령어
  - Command 창에 `gsort -price` 라고 입력하면, 변수 `price` 값들을 내림차순으로 정렬해달라는 뜻
- 명령어 `bysort` : `by` 다음에 나오는 변수를 동일한 값을 가진 것끼리 정렬해 그룹화 한 다음 : (콜론) 다음의 명령어를 실행하라는 뜻.
  - 어떤 기준을 의미하는 prefix 인 "by"와 "sort" 가 결합한 명령어로 붙여서 `bysort`로 사용한다.
  - `by foreign sort: su price =bysort foreign: su price`
  - Command창에 입력한 `by foreign sort: su price`는 변수 `foreign`에 대해서 오름차순으로 정렬을 하여 그룹화 한 후, 그룹의 가격에 대한 요약 통계량을 보여달라는 뜻입니다.

# Stata 주요 명령어(4)

## • 새로운 변수를 만드는 명령어 :

### gen

- `gen gap=flife-mlife`
- `gen ln_pop=log(pop)` → 인구변수에 log 취한 값을 새로운 변수 `ln_pop`로 만드는 명령어
- 특정 값을 갖는 변수 만들기
- `gen var1=2`
- `gen var2=_n`
- `gen var3=_n-1`
- → list 로 확인 가능
- `gen begin_n1=begin[_n-1]`

- (KOSPI 주가지수 주가 상승폭 만들기)
- `gen begin_n1=begin[_n-1]`
- ※ 꺾쇠괄호 `[n]`은 해당 변수의 `n`번째 관측치를 말한다.
- `gen diff_begin=begin-begin[_n-1]` → 주가 상승폭
- `gen return=100*(diff_begin)/begin[_n-1]` → 수익률
- `gen return=100*(begin-begin[_n-1])/begin[_n-1]` → 수익률
- `gen` 을 이용하여 특정 조건을 만족하는 더미변수 만들기
- `gen var4=return>0`
- `tab var4`
- 결측치를 포함하여 계산하므로 이를 빼려면 다음과 같이 코딩
- `drop var4`
- `gen var4=return>0 if return~.`
- `tab var4`

# Stata 주요 명령어(5)

- `gen p_k=price*rep78`
- (=price 변수와 rep78변수를 곱한 값을 p\_k변수를 만들어라)
- `gen new1_age=age`
- `replace new1_age=0 if age>30 & age==.`
- (=age변수를 동일하게 new1\_age로 만들고, age가 30초과일 경우와 age가 결측치이면 새롭게 만드는 new1\_age를 0으로 바꿔줘라)
  
- `gen new_age=0`
- `replace new_age=1 if age>=30`
- (=new\_age변수를 0으로 만들고, age가 30이상인 경우에만 new\_age를 1으로 바꿔줘라)
  
- `by year, sort : egen float rankfee = rank(fee)`
- (=fee변수를 랭킹을 먹이는데, 그 변수를 rankfee라는 새 변수로 지정을 하는데, 이때 year 기준으로 해줘라)
  
- `by year, sort : egen float avgfee = mean(fee)`
- (=fee변수의 평균을 시키는데, 그 변수를 avgfee라는 새 변수로 지정을 하는데, 이때 year 기준으로 해줘라)
  
- `sum price if rep78>=2 & rep78==., detail`
- (= price의 기술통계량을 보는데, rep78이 2이상인 경우와 rep78이 결측치는 경우만 보여주고, detail를 보여줘)

# Stata 주요 명령어(6)

- recode 변수 (. = 0) : 해당 '변수'의 값에서 결측(.)을 0으로 변환하라
- replace 변수 = . if 기준변수 == 0 : 만약 '기준 변수'의 값이 0인 경우 해당 행의 '변수'의 값을 .으로 변화시켜라
- drop : 해당 변수 삭제
- count if missing(기존변수) : 기존 변수내에 missing 값이 몇개인지 나타냄
- drop if missing(기존변수) : 기존 변수에서 관측값이 missing인 경우 제거
- keep if 기존변수 == 0 : 기존 변수 내에서 관측값이 0인 변수만 남기고 나머지 관측값들은 제거.
- keep if 기존변수 == "문자변수" : 문자 변수를 명령어에 사용하는 경우에는 ""를 사용함
- sysuse auto.dta, clear : stata에 기본 내장된 1978년 자동차 회사의 데이터를 불러옴
- summarize, 줄여서 sum : 변수에 대한 기초통계값 제시
- sum: 전체 변수에 대한 기초통계값 제시. 관측치, 평균, 표준편차, 최소값, 최대값
- sum price if foreign == 1 : foreign 변수 중 1의 값을 갖는 관측치에 대한 기초통계
- Rename → 변수 이름 변경하기
  - sysuse auto, clear
  - rename price price1000
  - rename weight 자동차무게

# Stata 주요 명령어(7)

- Stata 프로그램을 사용과정에서 필요한 명령어가 책이나 인터넷 등에서 설명이 되고 있는데 실행이 되지 않는 경우 사용하는 명령어는 'ssc install + 명령어', 'findit + 명령어'이다
- ssc install + 명령어 예시로 hireg를 설치해본다.

```
. ssc install hireg
checking hireg consistency and verifying not already installed...
installing into c:\ado\plus\...
installation complete.
```

- findit hireg 명령어 창에 입력하면 별도의 창이 나타나고 설치하고자 하는 창이 나타난다. 해당목록을 선택하면 다음과 설치를 위한 부분이 나타난다. 이를 선택하면 설치가 된다

---

search for **hireg** (manual: [R] search)

---

Search of official help files, FAQs, Examples, SJs, and STBs

Web resources from Stata and other users

(contacting <http://www.stata.com>)

1 package found (Stata Journal listed first)

-----

[hireg from http://fmwww.bc.edu/RePEc/bocode/h](http://fmwww.bc.edu/RePEc/bocode/h)

'HIREG': module for hierarchial regression / The hireg command conducts hierarchical regressions. Users / enter blocks of independent variables which are added to the / model in successive steps. R-squared change is reported at each / step along with a summary table at the end. All options

---

package **hireg** from <http://fmwww.bc.edu/RePEc/bocode/h>

---

**TITLE**

'HIREG': module for hierarchial regression

**DESCRIPTION/AUTHOR(S)**

The hireg command conducts hierarchical regressions. Users enter blocks of independent variables which are added to the model in successive steps. R-squared change is reported at each step along with a summary table at the end. All options available on the regress command may also be used with hireg.

**KW:** regression

**KW:** hierarchical regression

**Requires:** Stata version 8.0

**Author:** Paul H. Bern, Princeton University  
**Support:** email [pbern@princeton.edu](mailto:pbern@princeton.edu)

**Distribution-Date:** 20050723

**INSTALLATION FILES**

[hireg.ado](#)

[hireg.hlp](#)

([click here to install](#))

# 자료를 붙이기와 합치기

- 데이터를 사용하면서 기존의 자료를 사용하다가 추가적으로 자료를 확인하였을 경우 동일한 자료 구조가 있는 경우 append 명령어를 사용
  - use auto1999
  - Append using auto2009
- 동일한 키값을 갖고 있는 파일을 수집하여 합칠 경우에는 merge 명령어를 사용한다. 우리가 회귀분석을 할 때 영향을 분석하는 경우 A라는 결과 B라는 결과는 같은 대상에서 결과가 다르기 때문에 이러한 자료가 생성이 된다.
  - ✓ 먼저, 데이터를 기준변수인 키값에 따라 정렬을 하여야 한다..
  - ✓ 다음으로 merge 명령어를 실행하면 됨
  - ✓ Use auto1, clear
  - ✓ Sort it
  - ✓ Merge id using auto2, sort
- 이러한 데이터를 stata에서 append, merge 명령어를 쓰기에는 10만건 이상의 많은 데이터가 있을 경우에는 유용하지만, 그 이하에서는 엑셀을 이용하여 자료를 정리하는 것이 유용.(vlookup 함수를 사용하면 불규칙하게 정리된 파일도 쉽게 merge할 수 있고 데이터를 문제를 쉽게 찾을 수 있음)
- 데이터를 정리하고 활용하는 것에 대하여 엑셀을 등을 유용하게 사용하는 것도 stata, spss 등을 보다 효율적으로 사용할 수 있음

# Stata 결측치 처리

- 기본적으로 결측치는 "."로 표시하고 매우 큰 양수(very large positive number)로 정의.
- 결측치는 총 27가지로 영어 알파벳과 섞어서 표기되며 다음과 같은 크기를 가진다.
- 모든 관측치  $< . < .a < .b < \dots < .z$
- 결측치 간에 "." 이 가장 작고 ".z"가 가장 큼. age라는 변수가 60을 초과하는 관측치를 골라내고자 할 때 `list if age > 60` 이라고 실행하면 이 리스트는 결측치도 포함하기 때문에 이를 제외하려면 다음과 같이 실행해야함
- `list if age > 60 & age < .`
- 혹은 age 변수가 결측치인 경우만 골라내고자 하면 다음과 같이 진행
- `list if age <=.`

# 기본데이터를 이용한 stata이용하기

```
. log using "C:\work\stata\log\example.smcl"
```

```
name: <unnamed>
log: C:\work\stata\log\example.smcl
log type: smcl
opened on: 16 Mar 2024, 19:36:40
```

```
. sysuse auto,clear
(1978 automobile data)
```

```
. sum
```

| Variable     | Obs | Mean     | Std. dev. | Min  | Max   |
|--------------|-----|----------|-----------|------|-------|
| make         | 0   |          |           |      |       |
| price        | 74  | 6165.257 | 2949.496  | 3291 | 15906 |
| mpg          | 74  | 21.2973  | 5.785503  | 12   | 41    |
| rep78        | 69  | 3.405797 | .9899323  | 1    | 5     |
| headroom     | 74  | 2.993243 | .8459948  | 1.5  | 5     |
| trunk        | 74  | 13.75676 | 4.277404  | 5    | 23    |
| weight       | 74  | 3019.459 | 777.1936  | 1760 | 4840  |
| length       | 74  | 187.9324 | 22.26634  | 142  | 233   |
| turn         | 74  | 39.64865 | 4.399354  | 31   | 51    |
| displacement | 74  | 197.2973 | 91.83722  | 79   | 425   |
| gear_ratio   | 74  | 3.014865 | .4562871  | 2.19 | 3.89  |
| foreign      | 74  | .2972973 | .4601885  | 0    | 1     |

```
. pwcorr price headroom mpg displacement, star(.01) sig
```

|              | price    | headroom | mpg      | displacement |
|--------------|----------|----------|----------|--------------|
| price        | 1.0000   |          |          |              |
| headroom     | 0.1145   | 1.0000   |          |              |
|              | 0.3313   |          |          |              |
| mpg          | -0.4686* | -0.4138* | 1.0000   |              |
|              | 0.0000   | 0.0002   |          |              |
| displacement | 0.4949*  | 0.4745*  | -0.7056* | 1.0000       |
|              | 0.0000   | 0.0000   | 0.0000   |              |

가장 많이 사용되는 기술통계분석  
sum으로 하면 자동으로 모든 데이터의 평균, 표준편차  
최대값, 최소값을 나타낸다.

가장 많이 사용되는 상관분석  
회귀분석 등에서 기본적으로 사용되며,  
pwcorr 변수 변수 변수, stat(0.01) sig 이렇게 해주면  
\*가 붙여집니다.

# 참고문헌

- oscar torres-reyna(2007), panel data analysis fixed and random effects using stata, <http://www.princeton.edu/~otorres/>
- 민인식·최필선(2019), stata 기초적 이해와 활용, 지필미디어
- 민인식·최필선(2019), 패널데이터 분석 stata, 지필미디어
- 민인식·최필선(2019),고급통계분석 stata, 지필미디어
- 정성호(2018), stata 더 친해지기, 박영사
- 인터넷 자료
  - <https://m.blog.naver.com/gustncjstk1?categoryNo=10&tab=1>
  - <https://graduationplease.tistory.com/70>