패널데이터 분석

김기만, kcskgm@gmail.com

패널분석이란?

- 패널 데이터는 일종의 종단적 데이터, 즉 서로 다른 시점에서 수집된 데이터
- 패널분석은 종단분석이라고 한다.
- 일반적인 분석을 횡단분석이라고 한다.
- 분석을 함에 있어 연도별로 변화하는 추이를 분석하는 것이다.
- 패널 로짓 또는 프로핏 회귀분석(panel logit or probit regression)을 제외하고는 숫자형으로 분석
- “기존 통계프로그램이 횡단면 또는 시계열에 초점을 맞춘 것에 비해 STATA는 초기 개발 단계에 서부터 패널데이터의 분석에 중점을 둔 통계 패키지”
- 패널 데이터(종단적 또는 단면 시계열 데이터라고도 함)는 개체(I)의 동작이 시간(T)에 걸쳐 관찰되는 데이터 세트입니다
- 패널분석은 시계열 데이터 및 횡단면 데이터와 달리 변수들 간의 동적관계를 분석

패널분석이란?

패널데이터는 종단적이고 시계열적인 데이터

종단적 데이터의 세 가지 주요 유형:

- **시계열 데이터**: 최소한 하나의 단위(작은 N)에 대한 많은 관측치(큰 T).예: 주가 동향, 전국 통계 집계
- **합동 단면적(POOLED)**: 서로 다른 기간의 동일한 모집단에서 추출된 여러 단위(대형 N)의 두 개 이상의 독립 표본:

  예)일반 사회 조사

  인도의 10년마다 실시되는 인구 조사
- **패널 데이터**: 여러 단위(대형 N)에 대한 두 개 이상의 관측치(소형 T)

  예) 가구 및 개인을 대상으로 한 패널조사(NSS EUS, CES)

  다양한 시점의 조직 및 기업에 대한 데이터(ASI, NSS)

  시간 경과에 따라 집계된 국가/지역 데이터(WDI,WEO,BOP)

패널데이터란

- (X_{it}, Y_{it}), i =1,...N; t =1,... T 이러한 실체는 주, 회사, 가족, 개인, 국가 등이 될 수 있습니다.

Entity	Year	Y	X1	X2	X3	.....
1	1998	#	#	#	#	.....
1	1999	#	#	#	#	.....
1	2000	#	#	#	#	.....
2	1998	#	#	#	#	.....
2	1999	#	#	#	#	.....
2	2000	#	#	#	#	.....
3	1998	#	#	#	#	.....
3	1999	#	#	#	#	.....
3	2000	#	#	#	#	.....
4	1999	#	#	#	#	.....
4	2000	#	#	#	#	.....

패널분석을 위한 기초

- 패널데이터는 데이터의 이질성으로 인한 생략변수 편향을 다룬다. 이는 우리가 관찰할 수 없거나, 사용할 수 없거나, 측정할 수 없지만 예측변수와 상관관계가 있는 변수를 제어함으로써 이를 수행합니다.
- 패널데이터를 기초로 분석은 pooled ols model, 고정효과모형(fixed effect model), 확률(임의)효과모형(random effect model)으로 분석을 한다.
- pooled ols model : 패널데이터 선형회귀 추정방법이다. 패널데이터를 횡단면으로 분석하는 것이다. 모형은 다음과 같다.
  - $y = \alpha + x_{it}\beta + \mu_{it}$, i = 1, 2, 3, 4,..... n, and t = 1,2,3, ... T
  - n은 패널개최의 수이고, T는 개최 i의 데이터의 기간이다.
  - Stata에서 pooled OLS분석 방법은 reg명령을 사용한다.
  - 기본 표준 오류는 오류가 주어진 t에 대해 i에 대해 독립적이라고 잘못 가정합니다.
  - 선형회귀 모형에서 오차항 ϵ_{it}에 자기상관이 존재해서는 안된다. 이를 검증하기 위해서는 xtserial을 사용한다.

패널분석을 위한 기초

- 고정효과모형(FIXED EFFECTS MODEL (FE))
  - 시간이 지나도 변하지 않지만 개체에 따라 달라지는 변수(문화적 요인, 회사 간 업무 관행의 차이 등) → 개체 고정 효과(일원고정효과 모형)
  - $y_{it} = \alpha_i + \beta x_{it} + \mu_{it} + e_{it}$ i = 1, 2, 3, 4,…… n, and t = 1,2,3, …. T
  - 시간이 지남에 따라 변경되지만 개체에 따라 변경되지 않는 변수(예: 국가 정책, 연방 규정, 국제 협약 등) → 시간 고정 효과(이원고정효과 모형)
  - $y_{it} = \alpha_i + \beta x_{it} + \delta_{it} + \mu_{it} + e_{it}$ i = 1, 2, 3, 4,…… n, and t = 1,2,3, …. T
  - 고정효과모형을 분석하는 명령어는 xtreg를 사용하며, 옵션으로 fe를 사용한다

    예시 : xtreg 종속변수 독립변수, fe(일원고정효과 모형)

    xi : xtreg 종속변수 독립변수, fe(이원고정효과 모형)
- 확률(임의)효과모형(random effect model(re)
  - 확률(임의)효과모형은 그룹간 정보와 그룹내 정보를 모두 사용
  - $cov(x_{it}, u_{it}) = 0$이라는 가정이 성립할 경우 고정효과 모형에 비해 추정이 더 효율적

패널분석을 위한 기초

- 확률(임의)효과모형을 분석하는 명령어는 xtreg를 사용하며, 옵션으로 re를 사용한다

  예시 : xtreg 종속변수 독립변수, re
- $y_{it} = \alpha_i + \beta x_{it} + \gamma Z_{it} + e_{it}$, i = 1, 2, 3, 4,..... n, and t = 1,2,3, ... T
- 고정 효과 모델과 달리 개체 간 변동이 무작위로 가정되고 모델에 포함된 예측 변수 또는 독립 변수와 상관 관계가 없다는 것
- 고정 효과와 확률 효과 사이의 중요한 차이점은 관찰되지 않은 개별 효과가 모델의 회귀 변수와 상관 관계가 있는 요소를 구현하는지 여부이며, 이러한 효과가 확률론적인지 아닌지는 아님 [GREEN, 2008, P.183]
- 엔터티 간의 차이가 종속 변수에 어느 정도 영향을 미치지만 예측 변수와 상관 관계가 없다고 믿을 만한 이유가 있는 경우 확률(임의)효과를 사용해야 한다.
- 확률(임의) 효과의 장점은 시간 불변 변수(예: 성별)를 포함할 수 있다는 것
- RE를 사용하면 모델에 사용된 샘플 이상의 추론을 일반화할 수 있음

패널분석에서 어떠한 MODEL을 선택할 것인지?

- **POOLED OLS VS FE MODEL = F-TEST**
- **FE MODEL VS RE MODEL = HAUSMAN TEST**
- **RE MODEL VS POOLED OLS = LM TEST**
- 결과를 기초로 MODEL을 선택한다.
- (1번 F-TEST) 고정효과 모형을 검증하여 P값이 통계적 유의성이 있는지 확인한다. 통계적 유의성에서 P값이 0.1보다 적으면 POOLED OLS가 아닌 고정효과 모형을 선택, 하단부분에 F TEST THAT ALL U_I=0: F(XXX,XXX)=XXX PROB > F=X.XXXX 이런식으로 나타납니다. 여기서 F-VALUE가 나타나는데, F값이 1, 5, 10% 유의수준에서 유의한 값이 산출되면, FE를 사용하는것이 POOLED OLS 보다 상대적으로 우수함을
- (2번 HAUSEMAN 검증) HAUSMAN 테스트를 개별 특성이 회귀 변수와 상관 관계가 있는지 여부를 테스트한다(GREEN, 2008, 9장 참조). 귀무 가설은 그렇지 않다는 것입니다(확률효과).
  - ✓ xtreg 종속변수 독립변수, fe
  - ✓ estimate store fe_model
  - ✓ xtreg 종속변수 독립변수, re
  - ✓ estimate store re_model
  - ✓ hausman fe_model re_model
  - ✓ 만약 결과 값의 P값이 통계적 유의수준 (0.01, 0.05, 0.1)안에 있으면 고정효과모형을 채택함. 해당 P값이 1, 5, 10% 통계적으로 유의하므로 귀무가설이 기각이됨. 따라서 확률효과모형의 추정량은 일치추정량이 아니며, 고정효과모형을 선택하는 것이 보다 적절하고, 귀무가설이 기각되지 않으면 확률(임의)효과 모형이 타당함을 의미

패널분석에서 어떠한 MODEL을 선택할 것인지?

- (3번) LM 테스트는 확률(임의) 효과와 POOLED OLS MODDEL 중에서 결정하는 데 도움이 된다. LM 테스트의 귀무가설은 엔터티 간 분산이 0이라는 것입니다. 즉, 단위 간에 큰 차이가 없습니다(즉, 패널 효과 없음). STATA의 명령은 xttset0입니다. 확률(임의)효과 모형을 실행한 후 바로 명령을 실행하여야 한다.
  - 확률(임의)효과 모형에 대하여 검증을 한다. 검증을 하는 명령어는 다음과 같다.
  - ✓ qui xtreg 종속변수 독립변수, re
  - ✓ xttest0
  - ✓ 해석을 하면 P-value가 유의하다면, 즉 1, 5, 10%(0.01, 0.05, 0.1) 유의수준에서 유의한 값을 갖는다면 귀무가설을 기각하게 되고 확률(임의)효과 모형이 pooled OLS보다 상대적으로 우수하다는 통계결과를 보여줌. 반대로 통계적으로 유의한 값이 아니면, pooled OLS가 상대적으로 적절하다고 판단

=> 이상의 검증 절차를 거쳐서 고정효과 모형, 확률(임의)효과모형, pooled ols model을 선택

패널분석에서 어떠한 MODEL을 선택할 것인지?

- 어떠한 모델을 선택할 것인지는 검증을 통해선 선택하고 일반적으로
- 각 엔터티나 그룹의 개별 특성이 회귀 변수에 영향을 미친다는 것이 확실할 때마다 고정 효과를 사용
  - 예를 들어, 대부분의 국가에서 초과 근무에 대해 수집된 거시 경제 데이터입니다. 국가의 경제적 성과가 정부 유형, 정치적 환경, 문화적 특성, 공공 정책 유형 등 내부 특성에 의해 영향을 받을 수 있다고 믿을 만한 충분한 이유가 있을 수 있습니다.
- 확률(임의) 효과는 개별 특성이 회귀 변수에 영향을 미치지 않는다고 믿을 만한 이유가 있을 때마다(비상관) 사용된다.

데이터를 패널 데이터 형식으로 준비

- (X_{it}, Y_{it}), i =1,...N; t =1,... T 이러한 실체는 주, 회사, 가족, 개인, 국가 등이 될 수 있습니다.
- 패널 데이터를 분석하려면 변수는 열에 있어야 함, 행의 엔터티 및 시간 LONG 형이 되어야 함
- 기본적으로 우리는 WIDE형으로 바꾸어야 함
- STATA는 숫자를 열 이름으로만 사용하는 것은 부적절하다. 숫자 열 이름에 문자를 추가합니다. 예를 들어 연도로만 구성되어 있을 경우에는 연도 앞에 'x'를 추가하여야 한다.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
2	A			8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	B	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
4	C	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
5	D	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
6	E	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
7	F	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
8	G	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84

	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	x1995	x1996	x1997	x1998	x1999	x2000	x2001	x2002	x2003	x2004	x2005
2	A			8000.01	8212.90	7847.36	7702.89	7288.48	6430.98	6932.45	7486.24	8094.17
3	B	18268.01	18738.99	19360.46	20151.42	20715.54	20866.90	21364.02	21801.41	22404.59	22676.26	23039.43
4	C	21088.14	21608.14	21988.64	22739.28	23436.61	24194.85	24300.57	24411.48	24650.02	25076.01	25346.01
5	D	313.74	321.36	331.76	342.12	351.70	365.33	377.15	386.26	398.86	415.96	432.63
6	E	21123.66	21659.55	22299.13	22972.31	23613.87	24150.86	24788.69	25368.87	25885.48	26582.19	26890.73
7	F	29941.64	30703.73	31716.04	32671.27	33748.21	34599.47	34483.98	34669.47	35312.75	36450.55	37267.33
8	G	4891.60	5063.81	5328.88	5512.59	5647.06	5934.98	5864.12	5852.99	5872.29	6055.92	6162.84

RESHAPE WIDE TO LONG FORMAT

- Long형을 wide로 변환시키려면 ,

  => reshape wide gdp, i(countrycode) j(year)

CountryCode	Country Code
year	
gdp	
CountryName	Country Name

CountryCode	Country Code
gdp1960	1960 gdp
gdp1961	1961 gdp
gdp1962	1962 gdp
gdp1963	1963 gdp
gdp1964	1964 gdp
gdp1965	1965 gdp

	균형 패널	불균형 패널
시간갭이 없는 패널	①각 개체의 데이터 포괄기간이 동일하고 시간갭이 없음	③각 개체의 데이터 포괄기간이 다르지만 시간갭이 없음
시간갭이 있는 패널	②각 개체의 데이터 포괄기간이 동일하지만 시간갭이 존재	④각 개체의 데이터 포괄기간이 다르며 시간갭이 존재

패널은 균형패널과 불균형 패널로 구분된다.

균형 패널: 모든 개체가 항상 관찰된다.

불균형패널 : 일부 개체는 몇 년 동안 관찰되지 않는다.

패널데이터 만들기

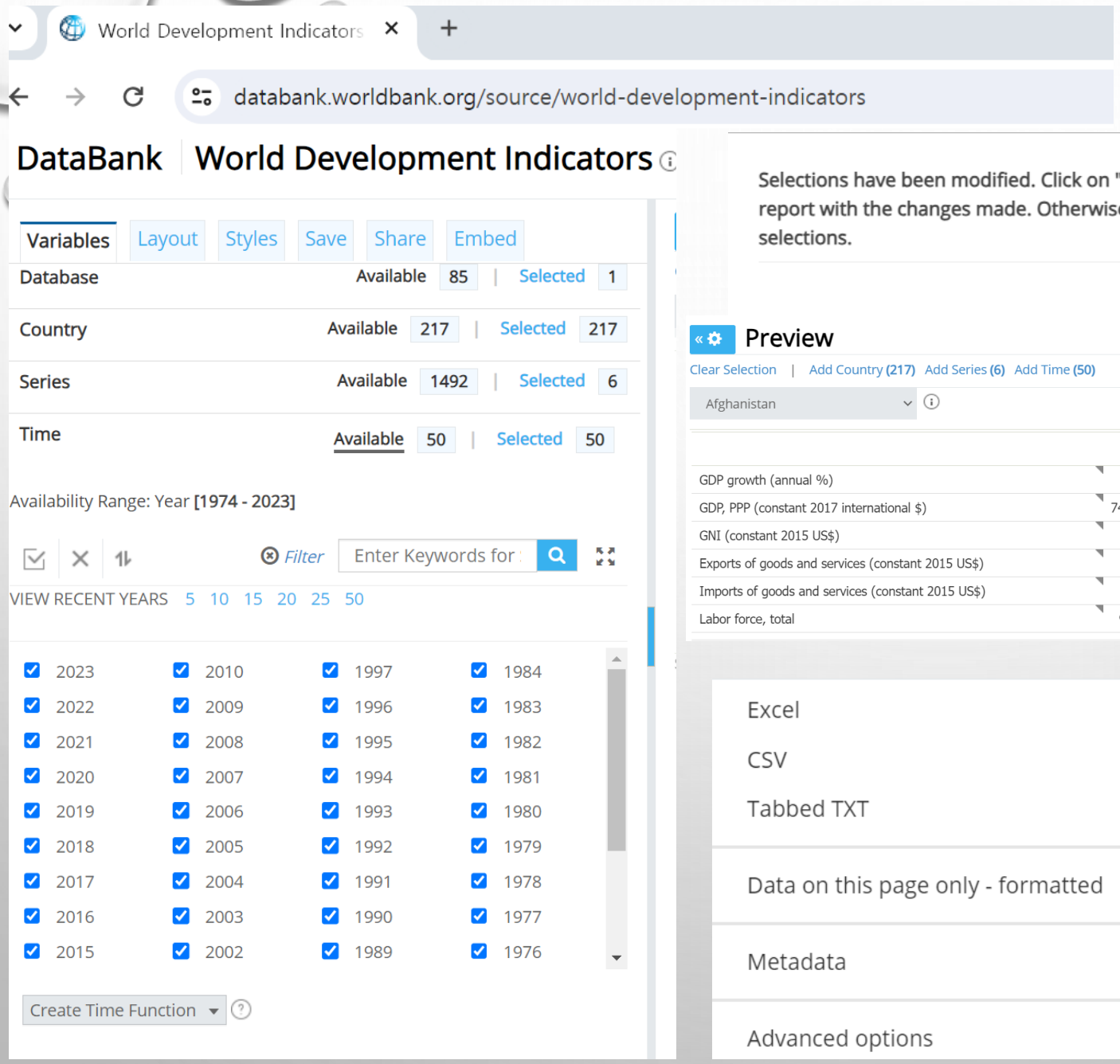
- 예제 자료는 세계은행 홈페이지에 있는 자료를 활용

The data used in the following slides was extracted from the World Development Indicators database:

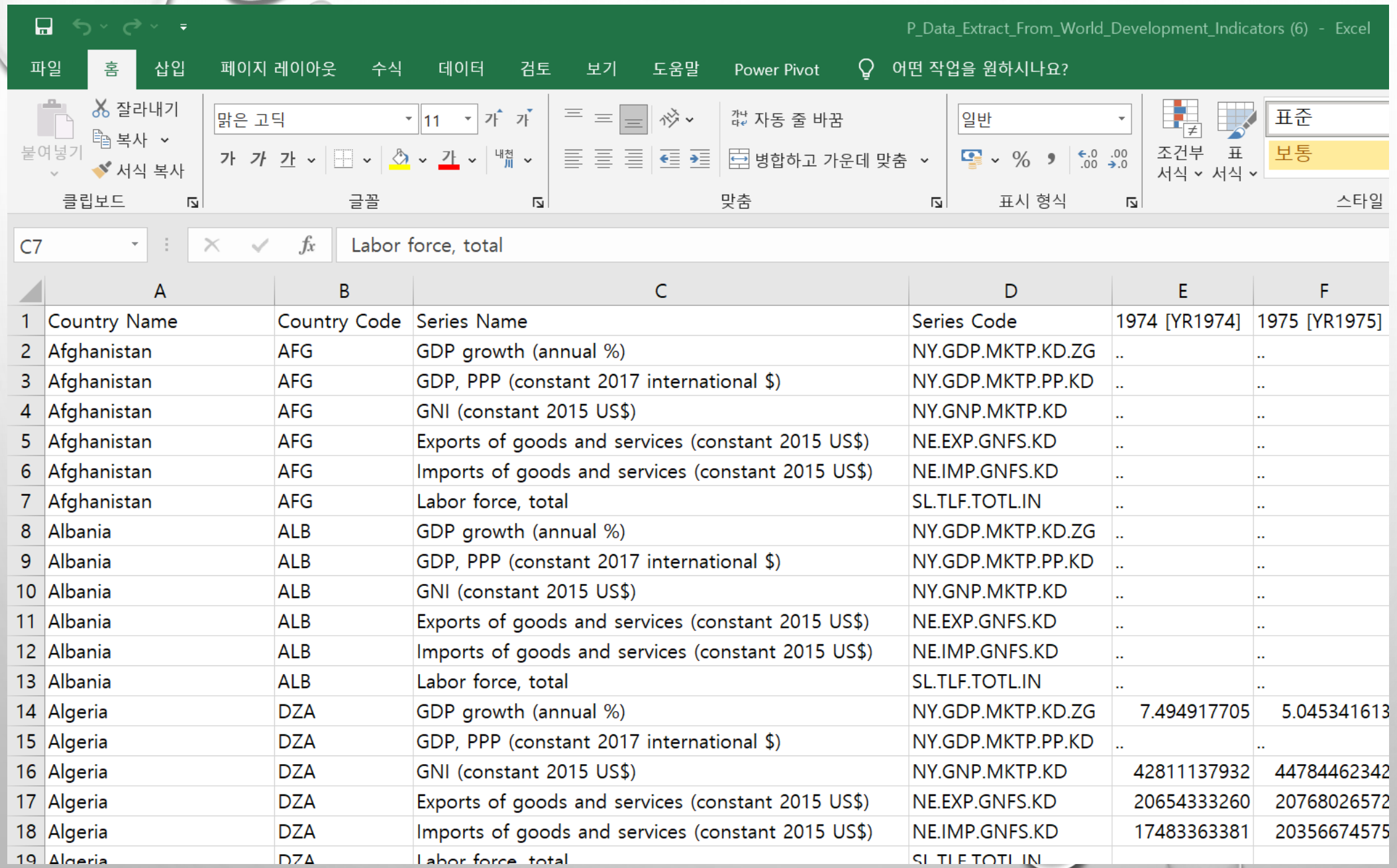
https://databank.worldbank.org/source/world-development-indicators

Selected variables since 2000, all countries only:

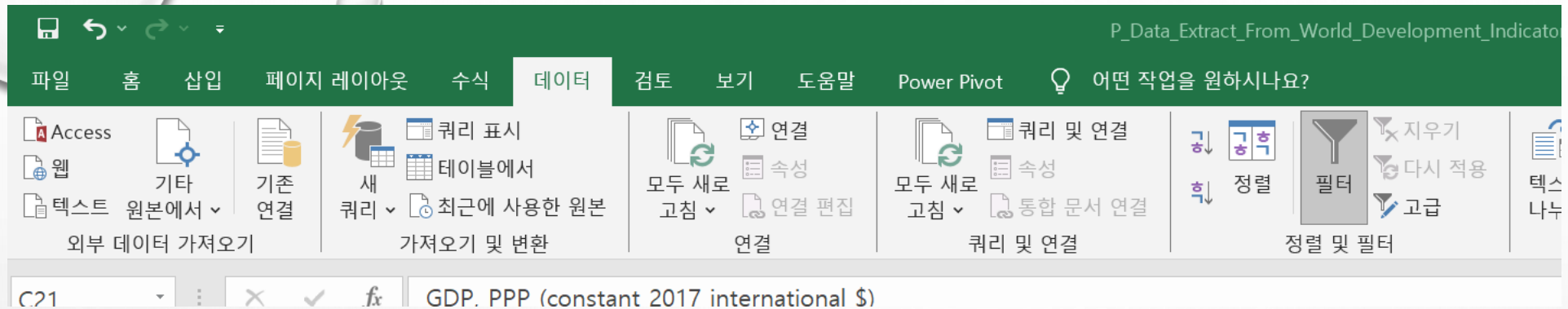
- GDP per capita (constant 2015 US$)
- Exports of goods and services (constant 2015 US$)
- Imports of goods and services (constant 2015 US$)
- Labor force, total



	2017	2018	2019	2020	2021	2022	2023
GDP growth (annual %)	2.65	1.19	3.91	-2.35	-20.74	..	..
GDP, PPP (constant 2017 international $)	74,711,922,906.	75,600,418,108.	78,557,606,648.	76,710,638,229.	60,801,742,189.	..	..
GNI (constant 2015 US$)	..	..	..	..	..	..	..
Exports of goods and services (constant 2015 US$)	..	..	..	..	..	..	..
Imports of goods and services (constant 2015 US$)	..	..	..	..	..	..	..
Labor force, total	9,254,593.00	9,242,721.00	9,220,323.00	9,105,733.00	9,356,574.00	8,803,873.00	8,920,521.00



	A	B	C	D	E	F
1	Country Name	Country Code	Series Name	Series Code	1974 [YR1974]	1975 [YR1975]
2	Afghanistan	AFG	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	..	..
3	Afghanistan	AFG	GDP, PPP (constant 2017 international $)	NY.GDP.MKTP.PP.KD	..	..
4	Afghanistan	AFG	GNI (constant 2015 US$)	NY.GNP.MKTP.KD	..	..
5	Afghanistan	AFG	Exports of goods and services (constant 2015 US$)	NE.EXP.GNFS.KD	..	..
6	Afghanistan	AFG	Imports of goods and services (constant 2015 US$)	NE.IMP.GNFS.KD	..	..
7	Afghanistan	AFG	Labor force, total	SL.TLF.TOTL.IN	..	..
8	Albania	ALB	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	..	..
9	Albania	ALB	GDP, PPP (constant 2017 international $)	NY.GDP.MKTP.PP.KD	..	..
10	Albania	ALB	GNI (constant 2015 US$)	NY.GNP.MKTP.KD	..	..
11	Albania	ALB	Exports of goods and services (constant 2015 US$)	NE.EXP.GNFS.KD	..	..
12	Albania	ALB	Imports of goods and services (constant 2015 US$)	NE.IMP.GNFS.KD	..	..
13	Albania	ALB	Labor force, total	SL.TLF.TOTL.IN	..	..
14	Algeria	DZA	GDP growth (annual %)	NY.GDP.MKTP.KD.ZG	7.494917705	5.045341613
15	Algeria	DZA	GDP, PPP (constant 2017 international $)	NY.GDP.MKTP.PP.KD	..	..
16	Algeria	DZA	GNI (constant 2015 US$)	NY.GNP.MKTP.KD	42811137932	44784462342
17	Algeria	DZA	Exports of goods and services (constant 2015 US$)	NE.EXP.GNFS.KD	20654333260	20768026572
18	Algeria	DZA	Imports of goods and services (constant 2015 US$)	NE.IMP.GNFS.KD	17483363381	20356674575
19	Algeria	DZA	Labor force, total	SL.TLF.TOTL.IN		



① 해당 항목별로 필터를 설정

② 설정된 필터로 추출된 데이터를 드래그인 드롭하여 선택하고 복사

③ 새로운 시트를 선택하여 복사한 자료를 붙이기

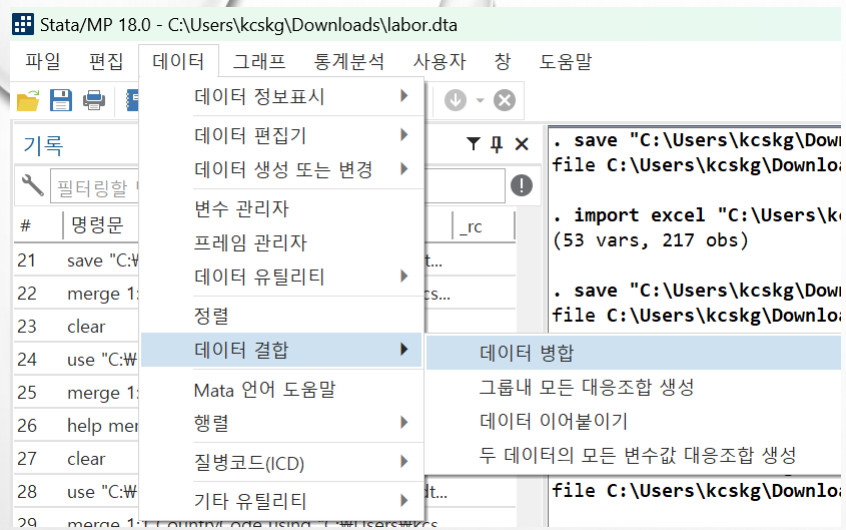
④ 각각의 항목을 선택하여 ②, ③ 을 반복 수행한다.

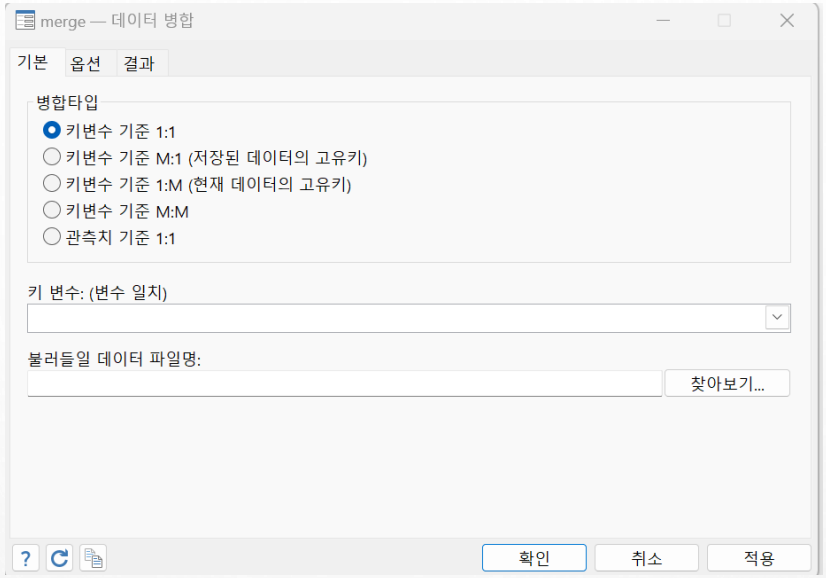
⑤ 열이름을 변경한다. 첫 글자가 stata에서 숫자를 사용할 수 없기에 영문+숫자로 변경을 한다. import1974로 바꾸고, 자동변환 기능을 이용하여 import2022까지 만들고 데이터가 없는 2023은 삭제

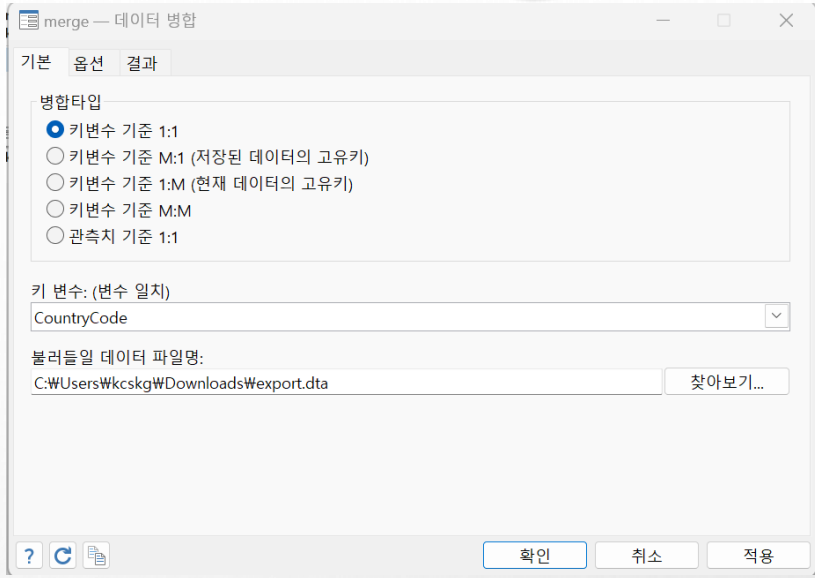
⑥ 국가별로 데이터를 만들기 위하여 키값인 국가이름과 국가코트를 복사하여서 별도의 시트로 정리하는 방법도 있고 이를 stata에 외부자료로 가져온 후 merge기능을 이용하여 합치기 함

⑦ stata에서 합치기 기능을 이용하여 합치기 하는 방법이 있도 있고 excel을 이용하여 합칠 수도 있다.

⑧ 217개의 데이터를 동일한 순서로 정렬한 후 copy & paste기능을 하여 간단하게 붙일 수 있음 다만, 결측치가 있을 경우에는 엑셀의 vlookup함수를 이용하여야 함

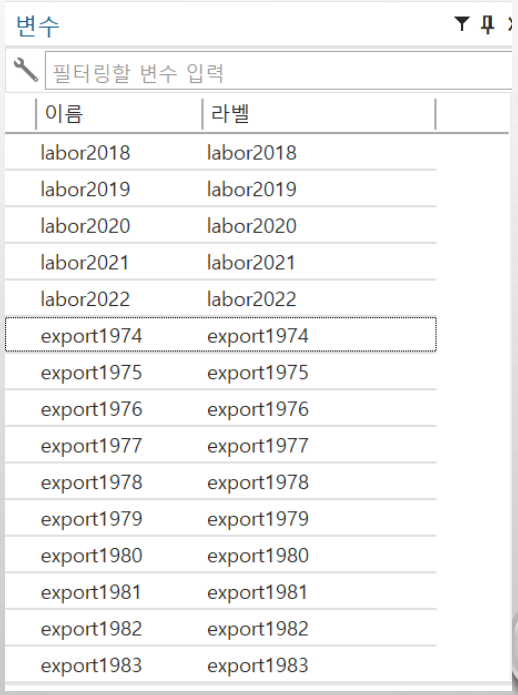







```
. merge 1:1 CountryCode using "C:\Users\kcskg\Downloads\export.dta"

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                               217  (_merge==3)
    -----------------------------------------
```

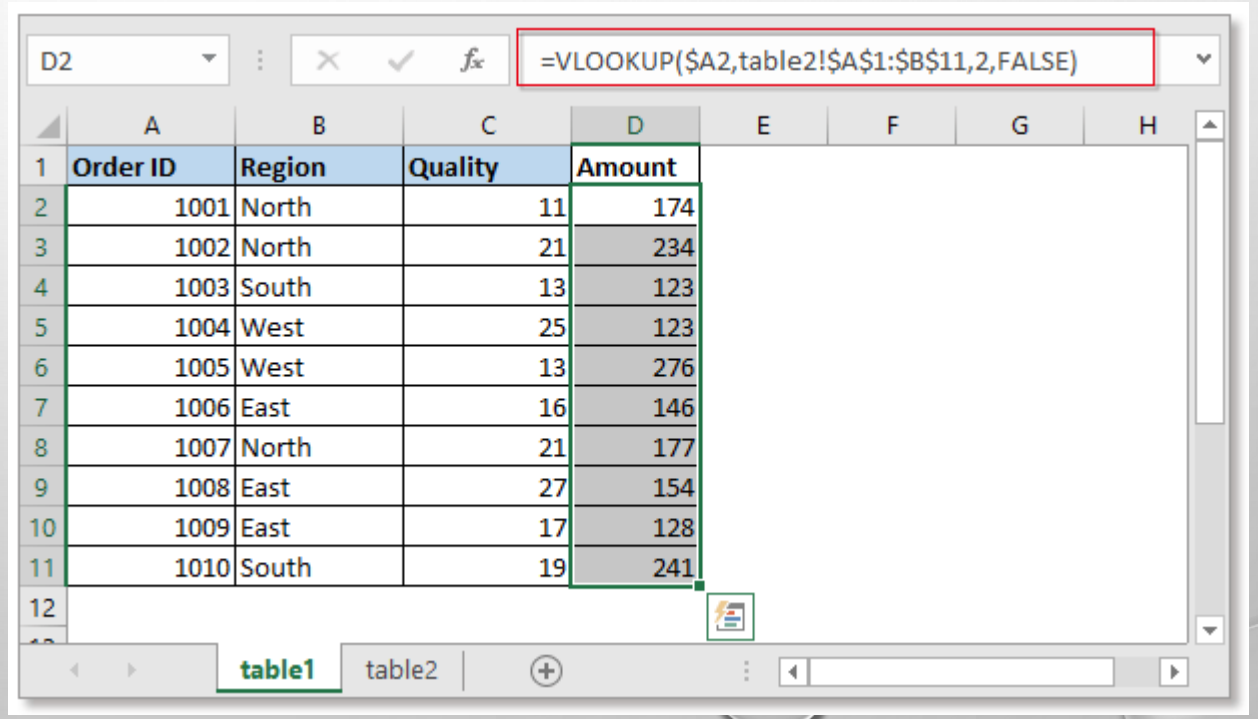


EXCEL로 데이터 합치기

- VLOOKUP함수를 사용하여 데이터를 합칠 수도 있지만, 데이터합치기 쿼리를 이용하여 합칠 수도 있음
- 데이터 합치기 쿼리를 이용하는 방법이다.
  ① 합치고자 하는 SHEET의 데이터를 블록으로 선택한 후 오른쪽 마우스를 클릭하여 표범위에서 데이터 가져오기를 선택한다.
  ② 표만들기 창이 나타나면 확인을 클릭한다.
  ③ 닫기 로드라는 아이콘을 클릭한다.
  ④ 합치고자 하는 표를 계속만든다.
  ⑤ 합치고자 하는 표가 완료되면 통합문서 쿼리창에서 해다 표를 선택하 후 오른쪽마우스를 클릭하면 메뉴가 나타나는데 병합을 클릭한다.
  ⑥ 합치고자 하는 표를 선택한다.
  ⑦ 연결하고자 하는 열이름을 선택한다. 위의 테이블과 아래의 테이블을 선택한 후, 확인을 클릭한다.
  ⑧ 새로 나타난 테이블의 마지막 열을 보면 표이름이 나타난다. 이를 클릭하면 표의 열이름이 나타난다. 나타난 테이블명에서 같은 이름이 있는 열이름을 선택하여 제거를 한다.
  ⑨ 아래에 있는 옵션을 해제한 후 확인을 클릭한다.
  ⑩ 닫기 및 로드라는 아이콘을 선택한다. 병합된 테이블을 확인할 수 있다.
  ⑪ 합치기를 계속할 경우 ⑦-⑩을 반복한다.

EXCEL로 데이터 합치기

- vlookup함수를 사용하여 데이터를 합칠 수도 있습니다.
- 사용법은 다음과 같습니다.
- 다음은 조회 테이블을 기반으로 기본 테이블을 빠르게 업데이트 할 수있는 VLOOKUP 함수입니다.
- 업데이트 된 데이터를 채우려는 셀에 아래 수식을 입력하십시오.
- =vlookup($A2,TABLE2!$A$1:$B$11,2,FALSE)
- 위의 공식에서 A2는 기본 테이블에있는 조회 값이고 TABLE! A1 : B11은 값을 조회하려는 테이블 범위이고 2는 반환하려는 테이블의 두 번째 열을 나타냅니다.



	A	B	C	D
1	Order ID	Region	Quality	Amount
2	1001	North	11	174
3	1002	North	21	234
4	1003	South	13	123
5	1004	West	25	123
6	1005	West	13	276
7	1006	East	16	146
8	1007	North	21	177
9	1008	East	27	154
10	1009	East	17	128
11	1010	South	19	241
12				

RESHAPE WIDE TO LONG FORMAT

- reshape명령어는 wide형 또는 long형으로 변화를 할 수 있다.
- 세계은행 자료를 보면 알 수 있는데 gdp자료를 패널자료로 만드는 예시

① 먼저 세계은행에서 관련자료를 다운로드 한다.(관련자료를 확인한다)

② stata에서 import명령문으로 불러들인다.

```
import excel "gdp_data.xlsx", sheet("Sheet1") cellrange(A4:BL270) firstrow
```

③ 불러들인 자료를 확인하다.

④ 초기에는 wide형으로 구성되어 있다. 이를 long형으로 바꾸기 위하여서는 다음과 같이 한다

```
. reshape long gdp, i( CountryCode ) j(year)
```

CountryName	Country Name
CountryCode	Country Code
gdp1960	gdp1960
gdp1961	gdp1961
gdp1962	gdp1962
gdp1963	gdp1963
gdp1964	gdp1964
gdp1965	gdp1965
gdp1966	gdp1966
gdp1967	gdp1967

CountryCode	Country Code
year	
CountryName	Country Name
gdp	

	Country...	year	CountryName	gdp
28	ABW	1987	Aruba	7886.8899
29	ABW	1988	Aruba	9769.5842
30	ABW	1989	Aruba	11395.978
31	ABW	1990	Aruba	12305.388
32	ABW	1991	Aruba	13494.685
33	ABW	1992	Aruba	14048.347
34	ABW	1993	Aruba	14942.275

CountryName	Country...	gdp1960	gdp1961	gdp1962	gdp1963	gdp1964
Aruba	ABW	.	.	.	.	.
Africa Eastern and Southern	AFE	162.72633	162.55597	172.27102	199.78492	180.22877
Afghanistan	AFG	59.773234	59.8609	58.458009	78.706429	82.095307
Africa Western and Central	AFW	107.93072	113.08006	118.82946	123.44109	131.85242
Angola	AGO	.	.	.	.	.

패널데이터 만들기

- 세계은행 데이터를 엑셀자료로 다운로드하여 PC의 임의의 디렉토리에 저장을 한다
- 기본적으로 세계은행 자료는 GDP, EXPORT, IMPORT, LABOR로 별도자료로 저장을 한다.
- EXCEL자료를 국가별로 그리고 년도별로 GDP, EXPORT, IMPORT, LABOR로 구성을 한다.
- import명령어로 stata에 excel자료를 가져온다.
- stata화일로 저장을 한다. panel_example.dta
- Reshape명령어를 사용하여 wide형을 long형으로 수정

reshape gdp import export labor, i(country_code), j(year)

CountryName	Country Name
CountryCode	Country Code
labor1998	labor1998
labor1999	labor1999
labor2000	labor2000
labor2001	labor2001
labor2002	labor2002
labor2003	labor2003
labor2004	labor2004
labor2005	labor2005

```
. reshape long labor import export gdp, i( CountryCode) j(year)
(j = 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022)

Data                               Wide   ->   Long
-----------------------------------------------------------------------------
Number of observations              217   ->   5,425
Number of variables                 103   ->   8
j variable (25 values)                    ->   year
xij variables:
      labor1998 labor1999 ... labor2022   ->   labor
   import1998 import1999 ... import2022   ->   import
   export1998 export1999 ... export2022   ->   export
           gdp1998 gdp1999 ... gdp2022    ->   gdp
-----------------------------------------------------------------------------
```

CountryCode	Country Code
year	
CountryName	Country Name
labor	
import	
export	
gdp	

패널데이터 만들기

- 데이터가 패널로 설정되면 일련의 xt 명령을 사용할 수 있습니다그것을 분석하기 위해. 자세한 내용을 보려면 help xt을 입력하세요.
- 시각화 명령은 xtline입니다.
- xtline gdp
- 독립변수가 몇% 변할때 종속변수가 몇%변하는지를 변화량일 비율로 보고자 할 때 독립변수와 종속변수에 모두 자연로그(Ln)를 취해줍니다.
- 이를 위하여 gen이라는 명령어를 사용합니다.
  - ✓ gen ln_gdp = ln(gdp)
  - ✓ gen ln_import = ln(import)
  - ✓ gen ln_export = ln(export)
- 결측치가 있는 데이터를 정리한다.

패널데이터 만들기

- 패널분석을 위하여서는 패널데이터를 사전적으로 정의하여야 한다.
- tsset id year, 또는 xtset id year 등으로 정의한다.
- 이를 실행하지 않을 경우 “must specify panelvar; use xtset”에러 메시지가 나타난다.
- tsset CountryCode year이라고 정의하면 에러 메시지 “string variables not allowed in varlist; CountryCode is a string variable” 구분을 위한 id 및 시간 등의 값은 숫자로 하여야 함
- 이를 하기 위하여서는 encode CountryCode, gen(country)라는 명령어를 이용하여 숫자형으로 변환시킨 후 tsset 또는 xtset 명령어를 사용하여야 함

```
. xtset country1 year

Panel variable: country1 (strongly balanced)
 Time variable: year, 1998 to 2022
         Delta: 1 unit
```

- 고정효과 모형을 이용하여 1998년부터 2022년까지 전세계 국가들의 수출과 수입이 gdp의 영향도를 분석하는 명령어는 xtreg gdp import export, fe
- 고정효과모형을 실행하기 위하여는 옵션으로 fe 사용
- 명령어를 실행한 후 “no observations”라는 에러 메시지가 뜨면, 변수가 숫자형으로 변환시켜야 한다.

패널회귀분석의 절차(1)

- 패널회귀분석은 가정의 타당성 평가, 모델 사양 확인, 적합도 검사, 패널 회귀 모델의 전반적인 성능 평가가 포함된다.

- (1단계) **패널 데이터 구조 확인**:
  - ✓ 데이터세트가 횡단면 및 시계열 차원을 모두 포함하는 패널 데이터로 올바르게 구성되었는지 확인
- (2단계) **기술통계 분석**
  - ✓ 'summarize', 'describe' 또는 xtsum과 같은 명령을 사용하여 패널 데이터세트의 변수에 대한 요약 통계를 검사
- (3단계) **패널회귀 분석 실시**
  - ✓ 데이터의 성격과 모델에 따라 **xtreg**, **xtregar**, or **xtmixed**
- (4단계) **모델확인**
  - ✓ 고정 효과와 무작위 효과 모델 중에서 선택하기 위한 Hausman test(hausman 명령) 또는 무작위 효과에 대한 Breusch-Pagan LM test(xttest2 명령)와 같은 진단 테스트를 검토하여 모델 사양의 적절성을 평가
- (5단계) **자기상관 검정**
  - ✓ xtserial 명령이나 기타 적절한 테스트를 사용하여 잔차의 계열 상관(자기상관)을 검정
- (6단계) **이분산성 검정**
  - xttest3 또는 xttest0과 같은 명령을 사용하여 잔차에 이분산성이 있는지 조사

패널회귀분석의 절차(2)

- (7단계) **개별효과 평가**
  - ✓ XTTEST0 명령을 사용하여 개별 효과(고정 효과)가 있는지 테스트
- (8단계) **적합도 평가**
  - ✓ R-제곱, 수정된 R-제곱 또는 우도 비율 테스트와 같은 적절한 통계를 사용하여 패널 회귀 모델의 전반적인 적합도를 평가
- (9단계) **민감도 분석 수행:**
  - ✓ 결과의 견고성을 평가하기 위해 다양한 사양이나 데이터 하위 집합을 사용하여 모델을 재추정하여 민감도 분석을 수행
- (10단계) **다중공선성을 확인**
  - ✓ VIF(분산 팽창 요인) 또는 상관 행렬과 같은 진단 테스트를 사용하여 독립 변수 간의 다중 공선성의 존재를 조사
- (11단계) **잔여물 검사:**
  - ✓ 패널 회귀 모델의 잔차를 검사하여 잔차 플롯이나 진단 테스트를 사용하여 패턴, 이상치 또는 가정 위반을 확인
- (12단계) **결과해석**
  - ✓ 가정의 타당성, 계수의 유의성, 모형의 전체적인 적합도를 고려하여 패널 회귀 분석 결과를 해석

고정효과모형을 이용한 분석(fixed effect model)

- 기본데이터 처리(sum명령어를 이용하여 pooled data처리)

```
. sum gdpN trade laborN
```

Variable	Obs	Mean	Std. dev.	Min	Max
gdpN	3,599	2555.949	1448.267	3	5064
trade	3,599	3759.334	1821.368	37	7490
laborN	3,599	2386.156	1318.12	2	4673

- 패널형 기본데이터 처리(xtsum 명령어를 이용)

```
. xtsum gdpN trade laborN
```

Variable		Mean	Std. dev.	Min	Max	Observations
gdpN	overall	2555.949	1448.267	3	5064	N = 3599
	between		1137.472	217	4669	n = 169
	within		940.7814	-1639.851	6757.189	T-bar = 21.2959
trade	overall	3759.334	1821.368	37	7490	N = 3599
	between		1200.311	1064	6958	n = 169
	within		1444.643	-2100.226	8800.214	T-bar = 21.2959
laborN	overall	2386.156	1318.12	2	4673	N = 3599
	between		1110.125	232	4475.84	n = 169
	within		765.4386	-1369.604	6395.156	T-bar = 21.2959

FIXED EFFECTS REGRESSION USING XTREG, FE

$$y_{it} = \alpha_i + \beta x_{it} + \mu_{it} + e_{it}$$

고정효과모형 옵션

이분산성을 제어하는 옵션

```
. xtreg ln_gdp ln_trade ln_labor, fe robust

Fixed-effects (within) regression               Number of obs      =       5,425
Group variable: country1                        Number of groups   =         217

R-squared:                                      Obs per group:
     Within  = 0.0580                                         min =          25
     Between = 0.1293                                         avg =        25.0
     Overall = 0.1037                                         max =          25

                                                F(2, 216)          =       10.13
corr(u_i, Xb) = 0.0392                          Prob > F           =      0.0001

                                    (Std. err. adjusted for 217 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
      ln_gdp | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   ln_trade  |   .1891074   .0468957     4.03   0.000     .0966758    .2815391
   ln_labor  |  -.0365564   .0383671    -0.95   0.342    -.1121781    .0390654
       _cons |   6.169313   .4089458    15.09   0.000     5.363277    6.975348
-------------+----------------------------------------------------------------
     sigma_u |   1.585357
     sigma_e |  1.2457059
         rho |  .61827066   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

데이터 수량
패널수

이 숫자가 0.05 미만이면 모델이 정상입니다.
이건 F – 모델의 모든 계수가
다음과 같은지 확인하기 위해 테스트

고정효과모형의 u_i는
회귀계수와 상관관계

0.05보다 낮은 값은 $null$을 거부하고
예측 변수가
결과에 유의미한 영향을 미친다는
결론을 내립니다(유의성 95%).

예측 변수가 1% 이상 증가할 때
시간이 지나면 출력(y)이
β%(탄성) 변경

클래스 내 상관관계(rho)는 출력의
분산 중 엔터티 간의 차이로
설명되는 정도를 보여줍니다.
이 예에서는 62%입니다.

$$rho = \frac{(sigma\_u)^2}{(sigma\_u)^2 + (sigma\_e)^2}$$

$sigma_u$ = 그룹내 잔차의 표준편차
Sigma_e = 잔차의 표준편차(전체오류항)

ENTITY AND TIME FIXED EFFECTS REGRESSION USING XTREG, FE

$$y_{it} = \alpha_i + \beta x_{it} + \delta_{it} + \mu_{it} + e_{it}$$

Time fixed effect

이분산성을 제어하는 옵션

```
. xtreg ln_gdp ln_trade ln_labor i.year, fe robust

Fixed-effects (within) regression               Number of obs      =       5,425
Group variable: country1                        Number of groups   =         217

R-squared:                                      Obs per group:
     Within  = 0.0702                                         min =          25
     Between = 0.1267                                         avg =        25.0
     Overall = 0.1054                                         max =          25

                                                F(26, 216)         =        2.83
corr(u_i, Xb) = 0.0674                          Prob > F           =      0.0000

                                    (Std. err. adjusted for 217 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
      ln_gdp |      Coef.   Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .1667699   .0475094     3.51   0.001     .0731286    .2604112
    ln_labor |  -.0349788   .0383007    -0.91   0.362    -.1104697    .0405122
             |
        year |
       1999  |   .0014681   .0491489     0.03   0.976    -.0954046    .0983409
       2000  |   .1978942   .0861808     2.30   0.023     .0280311    .3677572
       2001  |   .1830662   .0961262     1.90   0.058    -.0063993    .3725317
       2002  |   .3800648   .1191906     3.19   0.002     .1451393    .6149903
             |
       2021  |   .4211957   .1726042     2.44   0.015     .0809914      .7614
       2022  |   .0944115   .1860153     0.51   0.612     -.272226    .4610489
             |
       _cons |   5.932084   .4303323    13.78   0.000     5.083896    6.780272
-------------+----------------------------------------------------------------
     sigma_u |   1.591502
     sigma_e |  1.2404443
         rho |  .62208729   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

데이터 수량
패널수

이 숫자가 0.05 미만이면 모델이 정상입니다.
이건 F – 모델의 모든 계수가
다음과 같은지 확인하기 위해 테스트

0.05보다 낮은 값은 $null$을 거부하고
예측 변수가
결과에 유의미한 영향을 미친다는
결론을 내립니다(유의성 95%).

고정효과모형의 u_i는
회귀계수와 상관관계

예측 변수가 1% 이상 증가할 때
시간이 지나면 출력(y)이
β%(탄성) 변경

클래스 내 상관관계(rho)는 출력의
분산 중 엔터티 간의 차이로
설명되는 정도를 보여줍니다.
이 예에서는 62%입니다.

FIXED EFFECTS REGRESSION USING XTREG, FE (WITH LAGS ON PREDICTORS)

베타 계수는 예측 변수가 시간 경과에 따라 한 단위(1년 전 -"L1.")일 때 출력(y)의 변화를 나타냅니다.
이 예에서 모든 변수는 로그변환되었으며 해석은 다음과 같습니다.
예측 변수가 시간이 지남에 따라 1% 증가하면(1년 전 –"L1.") 출력(y)은 β%(탄력성)로 변경됩니다.

```
. xtreg ln_gdp l1.ln_trade l1.ln_labor , fe robust

Fixed-effects (within) regression               Number of obs      =       5,208
Group variable: country1                        Number of groups   =         217

R-squared:                                      Obs per group:
     Within  = 0.0293                                         min =          24
     Between = 0.0905                                         avg =        24.0
     Overall = 0.0681                                         max =          24

                                                F(2, 216)          =        7.69
corr(u_i, Xb) = 0.0809                          Prob > F           =      0.0006

                                    (Std. err. adjusted for 217 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
      ln_gdp | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |
         L1. |   .1307492   .0399075     3.28   0.001     .0520912    .2094072
             |
    ln_labor |
         L1. |  -.052238    .0387202    -1.35   0.179    -.1285559    .0240799
             |
       _cons |   6.627643   .3763705    17.61   0.000     5.885814    7.369472
-------------+----------------------------------------------------------------
     sigma_u |  1.6159341
     sigma_e |   1.227792
         rho |  .63399441   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

확률(임의)효과 모형(random effect model)

- sampling of sampling의 맥락에서 쓰인다.
- 예를 들어 보자. 어떤 한 집단(모집단의 크기가 무한하다고 하자)의 키 평균(모평균)을 구하기 위해 사람의 키를 임의로 표집(sampling)하여 키 평균을 추정할 수 있다.
- 하지만 키를 측정하는 도구가 매우 조잡해서 측정된 키의 표준오차가 10 cm 라고 가정해보자. 이때 측정에 따른 오차가 서로 독립이라면, 측정의 정확성을 높이는 것은 여러번 측정하는 것이다
- 확률효과는 엔터티의 오류 항이 시불변 변수가 설명 변수 역할을 할 수 있도록 하는 예측 변수와 상관 관계가 없다고 가정한 상태에서 분석하는 것이다.
- 분석을 위한 명령은 xtreg 종속변수 독립변수, re가 기본이다.

확률(임의)효과 모형(random effect model)

$$y_{it} = \alpha_i + \beta x_{it} + \gamma Z_{it} + e_{it}$$

확률(임의)효과모형 옵션

```
. xtreg ln_gdp ln_trade ln_labor, re robust

Random-effects GLS regression                   Number of obs     =       5,425
Group variable: country1                        Number of groups  =         217

R-squared:                                      Obs per group:
     Within  = 0.0574                                         min =          25
     Between = 0.1616                                         avg =        25.0
     Overall = 0.1242                                         max =          25

                                                Wald chi2(2)      =       19.66
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =      0.0001

                                (Std. err. adjusted for 217 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
      ln_gdp |      Coef.   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .1917432   .0433495     4.42   0.000     .1067798    .2767067
    ln_labor |   .0125553   .0313813     0.40   0.689     -.048951    .0740615
       _cons |   5.838727   .3890126    15.01   0.000     5.076276    6.601178
-------------+----------------------------------------------------------------
     sigma_u |  1.5160347
     sigma_e |  1.2457059
         rho |   .59695453   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

데이터 수량
패널수

확률(임의)효과모형의 u_i는
회귀계수와 상관관계

이 숫자가 0.05 미만이면 모델이 정상입니다.
이건 F – 모델의 모든 계수가
다음과 같은지 확인하기 위해 테스트

0.05보다 낮은 값은 $null$을 거부하고
예측 변수가
결과에 유의미한 영향을 미친다는
결론을 내립니다(유의성 95%).

고정효과, 확률(임의)효과, POOLED OLS ?

- (1단계) 고정효과 모형이 적절한지 검증 => xtreg ln_gdps ln_trade ln_labor, fe

```
. xtreg ln_gdp ln_trade ln_labor, fe

Fixed-effects (within) regression               Number of obs      =       5,425
Group variable: country1                        Number of groups   =         217

R-squared:                                      Obs per group:
     Within  = 0.0580                                         min =          25
     Between = 0.1293                                         avg =        25.0
     Overall = 0.1037                                         max =          25

                                                F(2, 5206)         =      160.28
corr(u_i, Xb) = 0.0392                          Prob > F           =      0.0000

------------------------------------------------------------------------------
      ln_gdp | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ln_trade |   .1891074   .0106217    17.80   0.000     .1682845    .2099304
    ln_labor |  -.0365564   .0268979    -1.36   0.174    -.0892876    .0161748
       _cons |   6.169313   .1858197    33.20   0.000     5.805028    6.533597
-------------+----------------------------------------------------------------
     sigma_u |   1.585357
     sigma_e |  1.2457059
         rho |   .61827066   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(216, 5206) = 37.86                     Prob > F = 0.0000
```

Prob > F = 0.0000으로
고정효과모형을 채택
P값이 통계적 유의성을
확보한 것으로 해석
Pool ols보다 고정모형 우세

고정효과, 확률(임의)효과, POOLED OLS ?

- (2단계) HAUSMAN검정을 실시하여 고정효과 모형과 확률(임의)효과 모형을 비교 검정

```
. qui xtreg ln_gdp ln_trade ln_labor, fe

. estimate store fe_model

. qui xtreg ln_gdp ln_trade ln_labor, re

. estimate store re_model
```

```
.  hausman fe_model re_model, sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |   fe_model     re_model       Difference       Std. err.
-------------+------------------------------------------------------------
    ln_trade |    .1891074     .1917432      -.0026358        .0031168
    ln_labor |   -.0365564     .0125553      -.0491117        .0151829
------------------------------------------------------------------------------
                           b = Consistent under H0 and Ha; obtained from xtreg.
            B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

    chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =  10.94
Prob > chi2 = 0.0042
```

명령어를 사용할 때 고정효과 모형을 저장한 것을 앞에 확률(임의)효과 모형을 저장한 것을 뒤에 기술

Prob > F = 0.0042로 고정효과모형을 채택
P값이 통계적 유의성을 확보한 것으로 해석
고정효과와 확률(임의)효과 비교 고정효과 우세

고정효과, 확률(임의)효과, POOLED OLS ?

- (3단계) LM 테스트는 확률(임의) 효과와 POOLED OLS MODDEL 중에서 우선순위 결정

```
. qui  xtreg  ln_gdp ln_trade ln_labor, re

.
.
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects

        ln_gdp[country1,t] = Xb + u[country1] + e[country1,t]

        Estimated results:
                         |       Var      SD = sqrt(Var)
                 ln_gdp  |   4.449113        2.109292
                      e  |   1.551783        1.245706
                      u  |   2.298361        1.516035

        Test: Var(u) = 0
                                  chibar2(01) = 22820.12
                                Prob > chibar2 =   0.0000
```

Prob > chibar2 = 0.0000로 확률(임의)효과모형을 채택 P값이 통계적 유의성을 확보한 것으로 해석 **확률효과와 pooled OLS비교 확률효과 우세**

고정효과, 확률(임의)효과, POOLED OLS ?

- 고정효과 모형, 확률(임의)효과 모형, POOLED OLS모형을 그대로 표출시키고 논문에 사용하면서 이를 해석할 수 있게 처리

```
. estimate table fe_model re_model ols_model, b(%9.4f) eq(1) star(0.01 0.05 0.1)
```

Variable	fe_model	re_model	ols_model
ln_trade	0.1891***	0.1917***	0.1738***
ln_labor	-0.0366	0.0126	0.1178***
_cons	6.1693***	5.8387***	5.2668***

Legend: * p<.1; ** p<.05; *** p<.01

결론적으로 pool ols모델보다는 확률(임의)효과 모형이 우선이 되고, 확률(임의)효과 모형보다는 고정효과 모형이 우선이 되어 **고정효과 모형으로 분석하는 것이 적절**

샘플로 사용한 데이터로 활용한 예시

- 자료 출처 : https://data.mendeley.com/datasets/vhh9cg2wzt/3

- THIS PANEL DATASET PRESENTS INFORMATION ON THE IMPACT OF DEMOCRACY AND POLITICAL STABILITY ON ECONOMIC GROWTH IN 15 MENA COUNTRIES FOR THE PERIOD 1983-2022. THE DATA ARE COLLECTED FROM FIVE DIFFERENT SOURCES; THE WORLD BANK DEVELOPMENT INDICATORS (WDI), THE WORLD BANK GOVERNANCE INDICATORS (WGI), THE PENN WORLD TABLE (PWT), POLITY5 FROM THE INTEGRATED NETWORK FOR SOCIETAL CONFLICT RESEARCH (INSCR), AND THE VARIETIES OF DEMOCRACY (V-DEM). THE DATASET INCLUDES TEN VARIABLES RELATED TO ECONOMIC GROWTH, DEMOCRACY, AND POLITICAL STABILITY. DATA ANALYSIS WAS PERFORMED USING STATISTICAL METHODS SUCH AS R IN ORDER TO ENSURE DATA RELIABILITY THROUGH IMPUTING MISSING DATA; HENCE, ENABLING FUTURE RESEARCHERS TO EXPLORE THE IMPACT OF POLITICAL FACTORS ON GROWTH IN VARIOUS CONTEXTS. THE DATA ARE PRESENTED IN TWO SHEETS, BEFORE AND AFTER THE IMPUTATION FOR MISSING VALUES. THE POTENTIAL REUSE OF THIS DATASET LIES IN THE ABILITY TO EXAMINE THE IMPACT OF DIFFERENT POLITICAL FACTORS ON ECONOMIC GROWTH IN THE REGION.

인터넷에서 회귀분석으로 많이 사용되는 보스턴의 집값에 대한 회귀분석 데이터 자료입니다.
데이터 자료
https://drive.google.com/file/d/1uWrgQj4r4FPoTocFqY5r_wJfIUhBsAGk/view?usp=drive_link

데이터에 대한 설명자료
https://drive.google.com/file/d/11LVj6lE446SOF082DtnTGgjXn9ybdSlM/view?usp=drive_link

무역규모, 민주화 지수 등에 따른 패널자료 분석자료
https://docs.google.com/spreadsheets/d/1h2ShTA5TP-ofu8NM44w7Kwu1sRg8SfY_/edit?usp=drive_link&ouid=117660642290208191364&rtpof=true&sd=true

무역규모 등에 따른 GDP영향 등을 분석하는 패널자료
https://drive.google.com/file/d/1VVAkaUDDTygGnsBmcH-jCmje-tPutY0U/view?usp=drive_link

Stata 기초자료 pdf
https://drive.google.com/file/d/16fUOWFNJnoAvQTIzeY49fmwpmg0tF1_L/view?usp=drive_link

회귀분석 pdf
https://drive.google.com/file/d/1bI6pfnXONMPbpolSmEYpsbulo4L8d2LS/view?usp=drive_link

패널분석 pdf

https://drive.google.com/file/d/1X_v2Vx8PE6GXrS_aRiQeU6mVhqKk4H2r/view?usp=drive_link

참고문헌

- oscar torres-reyna(2007), panel data analysis fixed and random effects using stata, http://www.princeton.edu/~otorres/
- 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