

HW3: Stata Application of Panel Data Method

Korea University

Instructor: Do Won Kwak

31 March, 2021

Problem I: Construction of Panel Data

Background The data set for this exercise comes from the paper by Baltagi and Khanti-Akom (1990) "On Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators", published in Journal of Applied Econometrics, vol. 5, p. 401-406. They demonstrate efficient estimation for a returns to schooling example based on a panel of 595 individuals observed over the period 1976-1982 drawn from the Panel Study of Income Dynamics (PSID). Use the data in *psidw.csv* for the following questions.

- (i) Load *psidw.csv* data to Stata using import command.
- (ii) Identify the types of data. Is it panel data or cross-section data? How many individuals are in the sample? How many years of data were collected for each individual?
- (iii) Convert wide data to long form with *id* as cross-section and *year* as time.
- (iv) Run OLS the following equation with $t = 1$.
$$lwage_{it} = \beta_0 + \beta_1 educ_{it} + \beta_2 union_{it} + \beta_3 married_{it} + \beta_4 exper_{it} + \beta_5 exper_{it}^2 + \beta_6 black_{it} + \beta_7 female_{it} + u_{it} \quad (1)$$
- (v) Perform hypothesis test on $H_0 : \beta_5 = 0$. What do you test with this null hypothesis?
- (vi) Compare equation (1) with $lwage_{it} = \beta_0 + \beta_1 educ_{it} + u_{it}$. What are the terms for OVBs in estimating this simple regression?
- (vii) Define the dataset as panel data with *id* as cross-section and *year* as time. Our collected data start from 1976 to 1979. Replace year value to reflect this information.

$$lwage_{it} = \beta_0 + \beta_1 educ_{it} + \beta_2 union_{it} + \beta_3 married_{it} + \beta_4 exper_{it} + \beta_5 exper_{it}^2 + \beta_6 black_{it} + \beta_7 female_{it} + f_t + c_i + u_{it} \quad (2)$$

(viii) Estimate the equation (2) by the pooled OLS. Compare this estimation results to OLS estimates from cross-section data.

(ix) Estimate the equation (2) by the random effects and the fixed effects estimators.

(x) Perform the Hausman test and choose between the random effects and fixed effects estimators. Provide a justification for your choice.

(xi) Why can't we estimate the coefficients on *black* and *female* variables by the fixed effects estimators?

Problem II: Combining data sets

Background Combining data sets: In many empirical research projects, the raw data to be utilized are stored in a number of separate files: separate "waves" of panel data, timeseries data extracted from different databases, and the like. Stata only permits a single data set to be accessed at one time. How, then, do you work with multiple data sets? Several commands are available, including `append`, `merge`, and `joinby`. How, then, do you combine datasets in Stata? First of all, it is important to understand that at least one of the datasets to be combined must already have been saved in Stata format. Second, you should realize that each of Stata's commands for combining datasets provides a certain functionality, which should not be confused with that of other commands. The **`append`** command combines two Stata-format data sets that possess *variables* in common, adding observations to the existing variables. The same variables need not be present in both files, as long as a subset of the variables are common to the "master" and "using" data sets. It is important to note that "PRICE" and "price" are different variables in Stata, and one will not be appended to the other.

(i) Use two datasets – `data2a.dta` and `data2b.dta` and combine into one dataset using `append` command. How many individuals are in the sample?

Background Combining data sets: We now describe the *merge* command, which is Stata's basic tool for working with more than one dataset. The `merge` command takes a first argument indicating whether you are performing a one-to-one, many-to-one, one-to-many or many-to-many merge using specified key variables. It can also perform a one-to-one merge by observation. Like the `append` command, the `merge` works on a "master" dataset the current contents of memory and a single "using" dataset. One or more key variables are specified, and you need not sort either dataset prior to merging. The distinction between "master" and "using" is important. When the same variable is present in each of the files, Stata's default behavior is to hold the master data inviolate and discard the using dataset's copy of that variable. The rule for `merge`, then, is that if datasets are to be combined on one or more merge keys, they each must have one or more variables with a common name and datatype (string vs. numeric). In the example above, each dataset must have a variable named `id`. That variable can be numeric or string, but that characteristic of the merge key variables must match across the datasets to be merged. This is the simplest kind of merge: the one-to-one merge. Stata supports several other types of merges. But the key concept should be clear: the `merge` command combines datasets "horizontally", adding variables values to existing observations.

(ii) Use two datasets – `data1.dta` and `data2.dta` and combine into one dataset using `merge` command. How many variables are in the sample?

(iii) Use two datasets – `data12.dta` and `data3.dta` and combine into one dataset using `merge` command. How many variables are in the sample? How many individuals are in the sample?

(iv) Reshape the dataset to estimate the panel data model.

(v) Summarize the dataset using `panel data` command.

Problem III: Download a Trade Data and make the data to be estimated in Stata

Background (1) Register to make an account at WITs.

<https://wits.worldbank.org/WITS/WITS/Restricted/Login.aspx>

(2) Download **trade flows** and **tariff data** from **WITs** as shown in the class.

(3) Download FTA and Custom Union data from **Mario Larch's** homepage.

<https://www.ewf.uni-bayreuth.de/en/research/RTA-data/index.html>

(4) Register to make an account at CEPII. Download **gratvity** variables at CEPII.

http://www.cepii.fr/cepii/en/bdd_modele/bdd_modele.asp (Also download the manual of pdf file.)

Combine Dataset Now let's combine these three set of data into one data file. We will combine the data one-by-one. When we combine two dataset, there is (i) the main dataset and (ii) the use data set. To combine this two dataset, we need common identifiers. Note that this is trade data between importer and exporter. And the information/variables between importer and exporter repeated over year, a panel data. Furthermore, we have information at the sector level. There are two things that are important. First, when we combine a more refined data (importer-exporter-year-product, e.g., tariff or trade flows data) with a broad concept data (importer-exporter-year e.g., FTA formation data), we must use the refined data as the main data and the broad concept data as the use data.

To combine data, we need to have common variables so that based on these common variables, we can combine two datasets. For example, in **WITs** data, identifiers for each observation is reporter-partner-product-tariffyear, and in RTA data from Mario Larch, identifiers for each observation is exporter-importer-year. We must make reporter-partner-tariffyear to be the same as exporter-importer-year so that we can combine two dataset. First, exporter-importer in RTA data is coded by iso3 code, while reporter-partner in WITs data is by world bank country number. Thus, please use iso3_code_r.dta and iso3_code_p.dta to link world bank country number and iso3 code. Note that in iso3_code_r.dta data, world bank country number is reporter and iso3 country code is iso3_r and in iso3_code_p.dta data, world bank country number is partner and iso3 country code is iso3_p.

(i) First, let's combine **WITs** data as a main data and iso3_code_r.dta data as a use data.

Ans) Use command merge m:1 reporter using iso3_code_r.dta
drop if _merge==2
drop _merge

(ii) Next, let's combine **WITs** data as a main data and iso3_code_p.dta data as a use data.

Ans) Use command merge m:1 partner using iso3_code_p.dta
drop if _merge==2
drop _merge
rename tariffyear year
save trade_tariff.dta, replace

(iii) Now, let's make RTA data to be combinable.

Ans) rename importer iso3_r
rename exporter iso3_p
save rta.dta, replace

(iv) Let's combine/merge trade flows and tariffs data with RTA data.

Ans) use trade_tariff.dta, clear
merge m:1 iso3_r iso3_p year using rta.dta
drop if _merge==2
drop _merge
save trade_tariff_rta.dta, replace

(v) Now, we want to combine gravity variables and income variables. Use Gravity_V202102.dta data. This data contains many unnecessary variables so we only keep necessary variables.

Ans) use Gravity_V202102.dta, clear
keep iso3_o iso3_d year distw contig gdp_o gdp_d gdp_o gdp_d pop_o pop_d comlang_off comcol
comrelig comleg_pretrans comleg_posttrans wto_o wto_d rta
keep if year>1995
rename iso3_o iso3_p
rename iso3_d iso3_r
save gravity.dta, replace

(vi) Now combine trade flows, tariffs, rta data with gravity and gdp data.

Ans) use trade_tariff_rta.dta, clear
merge m:1 iso3_r iso3_p year using gravity.dta
drop if _merge==2
drop _merge

(vii) Now estimate the following model and interpret $\hat{\beta}_1$:

$$\ln(\text{trade}) = \beta_0 + \beta_1 \text{RTA} + \beta_2 \ln(\text{GDP}_{\text{origin}}) + \beta_3 \ln(\text{GDP}_{\text{destination}}) + \beta_4 \ln(\text{population}_{\text{origin}}) \\ + \beta_5 \ln(\text{population}_{\text{destination}}) + \beta_6 \ln(\text{distance}_{\text{orig,destin}}) + \beta_7 \text{comlang} + \beta_8 \text{contig} + u$$

(viii)

Now estimate the following model with fixed effects interpret $\hat{\beta}_1$:

$$\ln(\text{trade})_{ijt} = \beta_0 + \beta_1 \text{RTA}_{ijt} + \beta_2 \ln(\text{GDP}_{\text{origin}}) + \beta_3 \ln(\text{GDP}_{\text{destination}}) + \beta_4 \ln(\text{population}_{\text{origin}}) \\ + \beta_5 \ln(\text{population}_{\text{destination}}) + \beta_6 \ln(\text{distance}_{\text{orig,destin}}) + \beta_7 \text{comlang} + \beta_8 \text{contig} + c_{ij} + f_t + u$$

(ix) Now estimate the following model, how does this estimation different from (vii)? Interpret $\hat{\beta}_1$ and γ . Perform a hypothesis test and also discuss violation of A4 due to confounding factors.

$$\ln(\text{trade}) = \beta_0 + \beta_1 \text{RTA} + \gamma \cdot \text{ltariff} + \beta_2 \ln(\text{GDP}_{\text{origin}}) + \beta_3 \ln(\text{GDP}_{\text{destination}}) + \beta_4 \ln(\text{population}_{\text{origin}}) \\ + \beta_5 \ln(\text{population}_{\text{destination}}) + \beta_6 \ln(\text{distance}_{\text{orig,destin}}) + \beta_7 \text{comlang} + \beta_8 \text{contig} + u$$

(x) Now estimate the following model, how does this estimation different from (viii)? Interpret $\hat{\beta}_1$ and γ . Perform a hypothesis test and also discuss violation of A4 due to confounding factors.

$$\ln(\text{trade})_{ijt} = \beta_0 + \beta_1 \text{RTA}_{ijt} + \gamma \cdot \text{ltariff} + \beta_2 \ln(\text{GDP}_{\text{origin}}) + \beta_3 \ln(\text{GDP}_{\text{destination}}) + \beta_4 \ln(\text{population}_{\text{origin}}) \\ + \beta_5 \ln(\text{population}_{\text{destination}}) + \beta_6 \ln(\text{distance}_{\text{orig,destin}}) + \beta_7 \text{comlang} + \beta_8 \text{contig} + c_{ij} + f_t + u$$