# Chapter 5: Time Series and Panel Data

Manuel V. Montesinos

Statistics Brush-Up Course
Competition and EPP Master Programs

Fall 2020

# *Time Series Data*

To create a **date variable**, we need to know what the date is for the **first observation** and the **frequency** of the data (daily, weekly, quarterly, annual, etc.).

Stata treats each time period as an integer. The integer records the number of time units that have passed from an agreed-upon base, which for Stata is 1960.

Suppose that we want to create a quarterly time series with 50 observations starting from the first quarter of 1960. For this, type:

```
set obs 50
generate date = q(1960q1) + _n-1
format date %tq
browse date
```

# *Time Series Data*

To generate the variable `date`, we have used the `q()` function, since we want the variable to refer to quarters. Other options are `y()` for yearly, `m()` for monthly, `w()` for weekly, and `td()` for daily.

Stata translates 1960q1 into integer equivalents: 0 is the equivalent to 1960q1, the second quarter is set to 1, and so on. Adding `_n-1` is done to increment the observations by one.

# *Time Series Data*

## *Setting data as time series*

Once you have set a date variable in a date format, you need to declare your data as time series in order to use time series operators. For that type `tsset date`.

If there are gaps in the time series, you can use the `tsfill` command to fill in the gap (use `tsset` before `tsfill`).

# *Time Series Data*

## *Time series operators*

There is a helpful set of operators for anayzing time series data:

- **Lags**: use the `L.` operator to generate variables with past values.
- **Forwards:** use the `F.` operator to generate forward or lead values of variables.
- **Differences:** use the `D.` (or `D1.`) operator to generate the first difference of a variable. Similarly, use `D2.` to take a second difference, and so on:

$$D1 = y_t - y_{t-1} \qquad D2 = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$$

- **Seasonal operators:** use the `S.` operator to generate seasonal differences (e.g. quarterly, yearly differences):

$$S1 = y_t - y_{t-1} \qquad S2 = y_t - y_{t-2}$$

# *Time Series Data*

## *Time series operators*

Note that a log difference is approximately the same as a percentage change, where the approximation is most accurate when the percentage change is small.

Hence, taking the first difference of the logarithm of a variable, we can obtain its percentage change:

```
generate lle = log(le)
generate lleD1 = D1.lle
list year le lleD1 in 1/5
```

# *Time Series Data*

## *Regression with time series*

In a regression, you can use the time series operators directly:

```
reg y x L1.x L2.x ...
reg y x L(1/2).x ...
```

The shortcut `L(1/2)` (or `F(1/2)`, `D(1/2)`, `S(1/2)`) defines a list of the first two lags (or leads, differences, seasonal differences) of `x`.

In our example: `reg le leL1 leL2`

## *Exercise*

Use the `ex_1.dta` dataset and:

1. Set it as a time series.
2. Generate log differences in consumption and investment.
3. Generate the exact percentage change in consumption and investment.
4. Check how good an approximation the log difference is to the exact percentage change.
5. Compute the difference of the difference and the lag-2 seasonal difference for consumption and compare these.
6. Run the following regressions:

$$
\begin{aligned}
inv_t = \beta_0 + \beta_1 cons_t + \beta_2 cons_{t-1} + \beta_3 income_t \\
+ \beta_4 income_{t-1} + \beta_5 income_{t-2} + u_t \qquad (1)
\end{aligned}
$$

$$\log(inv)_t = \beta_0 + \beta_1 \log(cons)_t + \beta_2 \log(income)_t + u_t \qquad (2)$$

$$inv_t = \beta_0 + \beta_1 \Delta cons_t + \beta_2 \Delta income_t + u_t \qquad (3)$$

# *Panel Data*

In **panel data**, the behavior of entities is observed across time. These entities can be individuals, countries, states, companies, etc.

Some panel data concepts:

- **Balanced panel**: same number of time observations ($T$) for all the entities ($N$).
- **Unbalanced panel**: entities can have a different number of observations ($T_i$).
- **Attrition**: drop-out process of entities from the panel, leading to an unbalanced panel.
- **Short panel**: large number of entities but few time observations for each.
- **Large panel**: large number of time observations for each entity.

# *Panel Data*

## *Setting as panel data*

To work with panel data, first we need to **declare** the dataset as a panel:

```
xtset panelvariable [timevariable]
```

Note that the time variable is **optional**. If the time variable is yearly, quarterly, etc., you should declare this in order to use Stata's time-series operators.

Let's use data from the **National Longitudinal Survey of Young Women**, a sample of women who were ages 14-24 in 1968:

- Every woman has a unique identifier (`idcode`), which works as panel variable for us.
- Each woman was surveyed once a year (`year` works as time variable for us).

# *Panel Data*

## *Describing panel data*

Use the `describe` command as usual to get a description of the dataset.

By using `xtdescribe` you will get panel specific information:

- 4,711 women with an `idcode` from 1 to 5,159.
- The `year` variable spans for 21 periods, but the maximum we can observe for any women in the dataset is 15.
- Distribution of $T_i$: about 50% of women are observed for 5 years or less. Only 5% of women are observed for 13 years or more.
- Finally, the pattern takes value 1 when there is an observation for that year. The largest fraction of women was observed in 1968.

# Panel Data

## Regression with panel data

For regression with panel data, use `xtreg, [fe | be | re]`:

- Use the **fixed-effects (FE)** estimator to remove the effect of time-invariant characteristics and assess the net effect of the predictors on the outcome variable.

$$y_{it} - \bar{y}_i = \left(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i\right)' \beta + \left(v_{it} - \bar{v}_{it}\right).$$

- Use the **between-effects (BE)** estimator to control for omitted variables that change over time but are constant between cases:

$$\bar{y}_i = \bar{\boldsymbol{x}}_i' \boldsymbol{\beta} + \bar{\eta}_i + \bar{v}_i.$$

- Use the **random-effects (RE)** estimator to work under the assumption that variation across entities is random and uncorrelated with the independent variables in the model.

# *Panel Data*

## *Data structure and reshaping the data*

Data can be organized in two ways:

- **Wide**: one record (row) per case. Observations on a variable for different time periods (or dates) held in different columns. Variable names identify time.
- **Long**: multiple records (rows) per case. Observations on a variable for different time periods held in different rows for each entity. The dataset's row identifier refers to time.

You can restructure your data from wide to long using the `reshape` command.

Here `i()` is required and specifies the variables whose unique values denote a logical observation (e.g. individuals); `j()` specifies the variable whose unique values denote a sub-observation (e.g. year).

# *Exercise*

Use the `ex_2.dta` dataset and:

1. Note that each variable has a name ending in the year for which the value is recorded. Reshape the dataset.

2. Declare the dataset as panel.

3. Run the following regression using the random-effects estimator and the fixed-effects estimator:

$$\log(wage)_{it} = \beta_0 + \beta_1 occcode_{it} + \beta_2 hours_{it} + \beta_3 tenure_{it} + \beta_4 age_{it} + u_{it}.$$

4. Reshape the data back to its original form.