Getting started with modelling

This workshop introduces you to basic time series data model specification and diagnostics in Stata. We expand on what you have learned so far with cross-sectional regression, and explore the unique statistical and theoretical challenges posed by analysing change over time.

2.1. Getting started with modelling (continued from session 1).

By lagging independent variables, we can introduce basic dynamics to try improve on the static model. Try comparing regressions on the original and differenced series for **`gininet`** and **`union`**:

```
reg gininet union
reg d.gininet d.union
```

What do you notice?

Now try specifying the model in lags of 1 and 2:

```
reg gininet union
reg gininet l.union
reg gininet l2.union
reg gininet l3.union
```

What can you conclude about the impact of unionisation on inequality? Is there a temporal effect? If so, how would you characterise it?

We can visualise this progressive lagged effect by plotting a cross-correlogram of the correlation between the dependent variable and various lags of the independent variable (note for this graph, the independent variable must come first – this is a rare exception to the general rule):

```
xcorr union gininet
```

When is the impact of unionisation strongest (hint: start from lag 0 and look back toward the negative values)?

2.2. Basic dynamic models for time series data

You may remember from our last session that there are a number of basic commands that cover regression specification, preliminary inspection, and diagnosis. There are also a number of different model specifications that cover different types of data-generating processes – these include the static model (without dynamics), finite distributed lag model (with lags of the independent variables), and the autoregressive distributed lag model (with lags of the dependent and independent variable to the right of the equation). In practice, your choice of model will be driven partly by statistical concerns (i.e. to address confounders such as autocorrelation and nonstationarity), but also by theoretical concerns (i.e. to properly model the real-world data generating process).

We talked quite a lot about the latter last week – should we expect unionisation to produce an instant impact on income inequality? Should tax reform instantly impact the incomes of the top 1%? Do fertility rates change immediately in response to rising affluence? When we think in these terms, we are thinking about the fit between our statistical model, and the data generating processes at work in the social world. As a result, you also need a good dose of theory and intuition to become a good longitudinal analyst. Nonetheless, there are certain principles we must adhere to as analysts in order to avoid confounding our models with serially correlated error terms, and nonstationary variables. The literature on these issues is quite dense (some examples are provided online for the brave amongst you), so we explore them with the needs of end-users in mind.

To recap, our suite of basic regression commands includes:

To create a cross-correlogram of lags of the independent variable with the dependent (remember to enter the independent variable first for this type of graph only): **`xcorr union gininet`**

To specify a basic bivariate OLS regression:

```
reg gininet union
```

To run a time-restricted model:

```
reg gininet union if year>1969 & year <1991
```

To generate a plot of residuals vs fitted values: **`rvfplot`**

To run parallel models, store and tabulate the model parameters and diagnostics:

```
reg gininet union estimates store model1 reg
labourshare union estimates store model2
estimates table model1 model2, star stats (N r2 r2_a)
```

These commands can be modified as you wish, using the **`if`**, **`in`**, and **`by`** qualifiers. If you want to do a basic calculation, you can type **`di`** before your sum:

```
di (2*2)+17
```

1.3. Diagnosing non-stationarity (unit root tests)

Referring back to our lecture slides on confounders of time series data, it is important that we test our variables for *stationarity*. The presence of stationarity may also be characterised in terms of the *order of integration* of our time series, or the presence of a *unit root* which will be referred to further on. These three terms are closely related, and we can test for stationarity by using a statistical test that takes non-stationarity as its null hypothesis. We will also examine another test which takes stationarity as its null hypothesis, and it is good practice to pair these tests together in order to arrive at a sounder conclusion. The procedure for stationarity testing is quite standard, and we can understand the relationship between these three key terms as follows.

An *order of integration* refers to the number of times a time series must be differenced in order to render it stationary. Remember, differencing the data involves subtracting the contemporary value from the past ($\mathbf{X} - \mathbf{X}_{t-1}$). If our tests confirm *non-stationarity* (i.e. the presence of a *unit root*), and if after differencing the time series it then tests stationary, we say that the series is *integrated of order one*. A time series that requires two rounds of differencing would be *integrated of order two*. If the test suggests the time series is stationary, we say it is *integrated of order zero*. The test procedure for a given time series is thus reasonably straightforward:

1. Subject the time series to relevant unit root tests.
2. If non-stationarity is revealed, difference the data.
3. Repeat the test on the differenced data.
4. Continue to difference if the test still reveals non-stationarity, or stop if stationarity is achieved.

1.3.1. Conducting the Augmented Dickey-Fuller (ADF) test

For the ADF test, the null hypothesis is non-stationarity:

H_a: Variable is stationary
H_0: Variable is non-stationary

Interpreting the test output is a little different, since the distribution of the test statistic is non-normal (i.e. we cannot use the t-distribution to calculate tail probabilities in the usual manner). We can read off the results quite easily as Stata provides the relevant critical values for a given alpha level, as well as a p-value for our test statistic. As usual, we accept a level of .05 as sufficient evidence for rejection of the null hypothesis. The test is implemented as follows:

```
dfuller union, lags(1) regress
```

Including **`lags(1)`** accounts for autocorrelation in union, and **`regress`** gives us the full regression output from the ADF procedure. As we can see, the null hypothesis of non-stationarity cannot be rejected, therefore we repeat the procedure on the first-differenced data. There are two ways to do this. The long way is to create a new variable composed of contemporary values of union subtracted from past values ($\mathbf{X} - \mathbf{X}_{t-1}$). The easier way is to do as we did last week, and use Stata's difference operator to modify the original command:

```
dfuller d.union, lags(1)
```

As results show, the first-differenced series is also non-stationary, so we then difference the data again. We can do this as above, by augmenting the difference operator in the command:

```
dfuller d2.union, lags(1)
```

Since results now confirm stationarity after two rounds of differencing, we conclude that **union** is *integrated of order two*.

1.3.2. Conducting the KPSS test

It is good practice to repeat this procedure using a different kind of test. The KPSS tests takes stationarity as its null hypothesis, and is therefore a good counterpart to the ADF test. To think intuitively about this, we might say that in order to be deemed stationarity, we wish this test to 'fail' (i.e. failure to reject the null implies that the variable is stationary).

H_a: Variable is non-stationary
H_0: Variable is stationary

As with all such tests, it is possible that our variable is stationarity around a deterministic trend (i.e. the variable may fluctuate, but may show a consistent pattern of increase or decrease). By default, the KPSS test includes a trend in its calculation, so we need to tell Stata if we wish to exclude this. We can intuit this from a visual inspection of the time series graph before we decide, but we can also repeat the test with and without a trend, to check whether a trend is present. Evidence suggests that automatic bandwidth selection of lags, and quadratic spectral kernel are efficient for small samples (this is not important, but if you really want to punish yourselves, you are welcome to read the paper, Hobijn et al 1998, p.14). We implement the test as follows:

```
kpss union, notrend qs auto
```

The test statistic of 1.23 exceeds the critical 5% value of 0.463, thus we reject the null hypothesis: the variable is non-stationary. If we were feeling especially optimistic, we might look to the 10% critical value. This is increasingly commonplace, especially in macrocomparative research using large panels of countries – since the goal of such studies is often not inference as such (they assume they are working with a population of OECD countries for example), hypothesis thresholds tend to be much wider. As a result, it is not uncommon to see in many macrosociology journals a threshold of 10% used for evaluating parameter significance. As before, with the test implying autocorrelation, we repeat with the differenced variable:

```
kpss d.union, notrend qs auto kpss d2.union,
notrend qs auto
```

As you can see, the second-differenced variable is stationary, thus confirming our conclusion that the variable is integrated of order two. On the basis of these tests, union should be differenced twice (i.e. entered into the model as d2.union) before running the model. This should remove the issue of confounding due to non-stationarity, which as we have seen, could potentially inflate our R^2 leading to misguided model evaluation.

Exercise 1

Test the following variables for stationarity using the battery of tests supplied above:

`labourshare, fdi, co2emiss, healthspend`.

Tip: inspect the graph for each variable. Try running the KPSS test on **`co2emiss`** *with a trend included – what do you find?*

1.4. Diagnosing autocorrelation

We will presently discuss how the potential presence of error autocorrelation can be factored into our choice of model specification. First, it is worth reviewing a formal statistical test for the presence of error autocorrelation. Unlike variants of other classical tests such as the DurbinWatson, we will examine the Breusch-Godfrey, which carries less assumptions regarding model specification than others. This test is performed immediately after running the regression model. For illustration, we take our FDL model of inequality and unionisation which we can specify as:

```
reg gininet l.union
```

We can plot the regression residuals by entering:

```
predict e, resid
```

This generates a new variable e which contains the regression residuals, which is the difference between the observed and predicted value. We can conduct a basic visual inspection of the residual plot with the following command:

```
tsline e, yline(0)
```

What do you conclude from the above graph? Do we find positive residuals close together and negative residuals close together? A more formal inspection of this possibility is to see how closely past residual values are correlated with contemporary values. We can do this as follows:

```
regress e, l.e
```

Does this model suggest that proximate residuals are closely correlated? This is a useful first pass, but there is a more formal approach available to us through the Breusch-Godfrey test. The null hypothesis is no serial correlation (this is included in the test output). The second option **`small`** returns a test statistic with an F-distribution:

```
reg gininet union l.union estat
bgodfrey,     lags(1)     estat
bgodfrey, lags(1) small
```

In both cases, we reject the null hypothesis – it appears there is some degree of residual autocorrelation. What can we do about this? The next section considers our options in a little more detail, but for now, we could try including a lagged dependent variable to the right of the equation in order to absorb some of the serial correlation. Remember, we also need to account for nonstationarity so we difference our variables accordingly:

```
reg d.gininet l.gininet d.union l.union estat
bgodfrey, lags(1)
```

As we can see, although the test statistic has reduced substantially, results indicate the presence of error autocorrelation (however, both visual inspection of the plot, and the low predictive power of an auxiliary autoregression on the residuals suggest not to the same degree). In practice, due to the nature of the data-generating process it may be impossible to fully purge, however it is crucial that we (1) explore and control for its presence as best we can, (2) diagnose our model carefully, and (3) acknowledge its potential confounding influence when reporting our results.

1.5. Finite Distributed Lag (FDL) and Autoregressive Distributed Lag (ARDL) models

Although these terms sound nasty, they are quite easy to understand. Parameters on both types of model are estimated using OLS, so you can comfortably rely on the knowledge you have accumulated so far with regard to multiple regression. All these equations describe is the parameterisation of the models – how the variables are entered, in what order, and with what degree of lag. If you recall from our original slides, the basic static regression model with no dynamics can be written as:

$$Y_t = \alpha_0 + \beta_o X_t + \varepsilon_t$$

This model contains no lags of either the dependent or independent variable to the right of the equation, just contemporary values of the independent (X). An **FDL** model is similar to the above, but includes lags of the independent variable (X) as a separate explanatory variable:

$$Y_t = \alpha_0 + \beta_o X_t + \beta_1 X_{t-1} + \varepsilon_t$$

Finally, an ARDL or general model contains the same parameters as the FDL, but with an autoregressive term – this is simply a regression of the lag of the dependent variable on itself. Remember, we can think of this intuitively in terms of real-world social processes. Institutions are often slow to change in response to exogenous influences, thus past values of a particular variable are also likely to influence future values. Statistically, we understand the proximity of similarly signed error terms as autocorrelation. Think, for example, of unionisation levels in Denmark. Unionisation levels are high in Danish society, and employers and state are both incentivised to negotiate with unions to maintain industrial peace. As a result, unionization is likely to change slowly with a given time period, if at all. Because levels are so high, it is these high levels themselves that partly keep the process in place. This is sometimes referred to in the social science literature as 'institutional inertia' or 'path dependence'. We can capture this process by using an **ARDL** specification:

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_o X_t + \beta_1 X_{t-1} + \varepsilon_t$$

Specifying these models in Stata involves little more than figuring out what order the variables are entered into the model – we use the same set of commands as before, we just think a little more creatively (and of course statistically) about the appropriate specification for our data type. How can we think about model specification in intuitive real-world terms? It is worth referring to DeBoef and Keele's (2008) notes on interpretation:

> "An exogenous variable may have only short-term effects on the outcome variable. These may occur at any lag, but the effect does not persist into the future. The reaction of economic prospections to the machinations of politicians, for example, may be quite ephemeral – influencing evaluations today, but not tomorrow."

> "An exogenous variable may have both short and long-term effects. In this case, the changes in X_{t-I} affect Y_t, but that effect is distributed across several future time periods. Often this occurs because the adjustment process necessary to maintain long-run equilibrium is distributed over some number of time points. Levels of democracy may affect trade between nations both contemporaneously and into the future. These effects may be distributed across only a few or perhaps many future time periods. How many time periods is an empirical question that can be answered with our data."

We can give this a try with our inequality data. Shortly, we will take a look at a published study and replicate their findings by re-specifying their time series models. Remember, it is often difficult to identify a 'correct' model, but there are certainly better or worse justifications – keep this in mind when approaching your assessment for the class. You do not need to identify the 'right' model as such (by definition, there is no such thing), but you do need to provide a sound justification for your selection.

The good news (at this stage) is that you have already specified some of these models throughout the course.

The **static** model is entered as:

```
reg gininet union
```

The **FDL** model is entered as:

```
reg gininet union l.union
```

The **ARDL** model is entered as:

```
reg gininet l.gininet union l.union
```

Finally, we utilised an especially versatile specification above known as an Error-Correction Model (ECM). The statistical properties of this specification are well-documented, and it is a popular choice in the social sciences due to its ability to cope with non-stationary data, and the modelling

of cointegrated processes (see Keele and DeBoef 2004 for further information). The **ECM** may be written as follows:

$$\Delta Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_o \Delta X_t + \beta_1 X_{t-1} + \varepsilon_t$$

As you can see, the ECM involves a small modification to the ARDL specification – instead of modelling the outcome Y in levels, we instead model is in first differences. Our question thus becomes, 'what are the effects of the independent variables on *change* in the dependent variable'. We can interpret the parameters as follows:

$\boldsymbol{\beta_o \Delta X_t}$ captures the short-term instantaneous effects of changes in X on Y within a single period.

$\boldsymbol{\beta_1 X_{t-1}}$ captures the long-run effects of lags of X on future changes in Y.

However, changes in the independent variables will also provoke an 'equilibrium' response in the dependent. Y will thus adjust to changes in the independent variables, but the adjustment of Y is not instantaneous. It occurs at a rate dictated by $\boldsymbol{\alpha_1 Y_{t-1}}$.

We can explore this more intuitively with an example from our dataset modelling the relationship between unemployment and income inequality:

```
reg d.gininet l.gininet d.unemp l.unemp
```

From your regression output:

gininet L1 denotes $\boldsymbol{\alpha_1 Y_{t-1}}$, the adjustment rate of Y. **union D1** denotes $\boldsymbol{\beta_o \Delta X_t}$, the short-term effect of X on Y. **union L1** denotes $\boldsymbol{\beta_1 X_{t-1}}$, the long-run effect of X on Y.

We can thus conclude that unemployment appears to exert a non-significant short-term influence on Y, with a stronger long-term impact 0f .077, consistent with existing theory.