**Stationarity**

A variable is stationary if (the distribution is independent of time) (weak stationarity):

1. Constant mean (mean is independent of time)
2. Constant variance
3. Covariances independent of time (ACF)

Why worry if all the variables are stationary?

If y or x or both is/are nonstationary then e = y – a – bx is nonstationary

We assume = stationary.

Results in a lot of “false positives” – appear statistically significant when the variables may be unrelated.

**Must convert all variables into stationary form first** (before estimating models, testing, etc.)

How to convert a variable, x?

**Difference**: If x is already stationary, difference zero. Known as integrated of order zero, I(0).

If x is stationary after differencing once, it is integrated of order one, I(1).

If x is stationary after differencing twice, it is integrated of order two, I(2).

**Determining order of integration**

1. Visual inspection (look at time plots of, )
2. SDs (look at SD of, )
3. Autocorrelation functions (ACFs)
4. ADF test (use with caution!)
5. Think about the DGP (the data generating process).

Difference (don’t forget to take logs first if possible!) all the variables. Then build

**A dynamic regression model**

A static regression model:

A dynamic regression model:

Include lags of ALL variables, including the dependent variable

In R:

r1 = dynlm(dxrt~L(dxrt,1:12)+L(dut,0:12)+L(dpdt,0:12),start=c(1971,1), end=c(2018,7))

AIC and BIC need the SAME number of obs. exactly to compare models!

where

**Build a dynamic regression model:**

Start with a “general” specification (“the kitchen sink model”)

As large a number of lags for each variable as is reasonable/possible.

Narrow down to a “best” specification (best choice of lags lengths).

How to choose lag lengths?

1. THINK! Think about the underlying DGP. (DO NOT USE STEPWISE REGRESSION METHODS!!!!! PLEASE!)
2. AIC and BIC model selection criteria (NOT adj. R-squared!!) – can give very different answers!
3. **Best method: Predict out of sample and see which model predict best! (RMSPE – root mean squared prediction error)**
4. t-tests and F-tests (p-values)
5. Omitted variables vs. inefficiency due to inclusion of irrelevant variables.
6. Test for serial correlation in the model error.

[Aside: k-fold and leave-one-out cross validation (LOO-CV) is a great way to choose a model!

You take your data and drop k observation (k < n), e.g. 5 out of 100, k= 1 is LOO-CV. Use the remaining to estimate and predict for the omitted observations, repeat with each subset of k obs until you have out of sample predictions for all n obs.]

[Aside: Flexible functional form

Transcendental logarithmic production function

Transcendental logarithmic consumption function

A quadratic model, with interaction terms, all the variables in logs]

**Seasonality**

Three ways to model seasonality:

1. Include seasonal dummy indicator variables in the model, one for each period of a year.

* E.g. If monthly data include 11 seasonal dummies + intercept (12 seasonal dummies, no intercept), , , etc.
* Test for seasonality - = no seasonality. Another joint F-test.

1. Include “seasonal lags” – the lag with a periodicity of a year, i.e. for monthly , for quarterly data .

E.g.

RMSE =

1. Seasonal differencing: take the difference of the variable over one year instead one period.

Seasonal difference for monthly data: (still a first difference!)

Using the package dynlm:

## FOR MODEL SPECIFICATIONS: take max lag in model and start at period = max lag + 1

For example, suppose we have quarterly data, if we estimate:

r0 <- dynlm(y~L(y,1),start=c(1,1), end=c(12,4)) () AR(1)

Actually start estimation at obs. 2 because we create one lag.

and

r1 <- dynlm(ly~L(y,1)+season(y),start=c(1,1), end=c(12,4))

()

This will be fine because we still start estimation at obs. 2 because we create one lag.

BUT … if we estimate:

r0 <- dynlm(y~L(y,1:2),start=c(1,1), end=c(12,4)) () AR(2)

This will use one less obs. than the other two models because of the extra lag – so actually it starts estimation at obs. 3

So when we try to use anova, etc., we will get an error message!

SOLUTION:

Start all the models at a late enough date that estimation starts at the same place, i.e. max lag in any model + 1. E.g. if one model has 2 lags and all the others have 2 or less, start at obs 3 for ALL models.

Seasonal difference, e.g.

Suppose is monthly data. The seasonal first difference is . Same for an explanatory variable, : .

Now we estimate a model:

Not the same as:

Not the same as:

Not the same as:

We can combine all of these (seasonal dummies, seasonal lags and seasonal difference!)

r0 <- dynlm(y~L(y,2),start=c(1,1), end=c(12,4)) () AR(2)

anova(r1,r0)

AIC(r1); BIC(r1)

**Serial correlation**

Suppose I estimate, instead of (1),

serial correlation = correlation from one time period to the next (in the same variable)

Ist order serial correlation = corr(, t = 1,2,3, …, n

2nd order serial correlation = corr(,

“Serial correlation” in the model error is a bad thing! It means there is systematic variation in the dependent that is not in the model (because it is in the error).

So include more lags, especially of the dependent variable, until (hopefully) the serial correlation is eliminated.

Use the Breusch-Godfrey LM test for serial correlation.

Null hypothesis is no serial correlation

So we want to fail to reject the null (at the 10% level).

If we reject the null, then we need add more lags to our model.

If we fail to reject, the model is fine as is (in terms of serial correlation).

**Granger causality test** (not granger causality, causality not casualty).

Test all the lags (plus the contemporaneous value) of one variable = Is there any effect over time of the variable on the dependent variable (any predictive power)?

E.g. Does US industrial production have any effect on the US-UK exchange rate over time?

Model:

Test

We found very weak to no evidence of a relationship (p-value = 0.1089)