ECON-C4200 - Econometrics II

Lecture 9: Time series II

Otto Toivanen

Learning outcomes

- At the end of lecture 9, you know

1. how to formulate a time series model with several variables
2. what spurious correlation means in a time series context
3. how to model breaks in a time series
4. how to forecast
5. what the **M**ean **S**quared **F**orecast **E**rror is and how to estimate it
6. what the **Forecast interval** is

Time series with explanatory variables

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_2 X_{t-1} + u_t$$

- As with Y_t, can include multiple ("distributed") lags of X_t: $ADL(p, q)$ (for **A**utoregressive **D**istributed **L**ag model).

Stationarity for time series Y_t, X_t

- **Definition**: A time series Y_t, X_t is stationary if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{s+1}, X_{s+1}, Y_{s+2}, X_{s+2} \ldots, Y_{s+T}, X_{s+T})$ does not depend on s.
- Otherwise Y_t, X_t is **nonstationary**.
- Stationarity requries that in a **probabilistic** sense, the future is like the past.

What is the fuzz about stationarity?

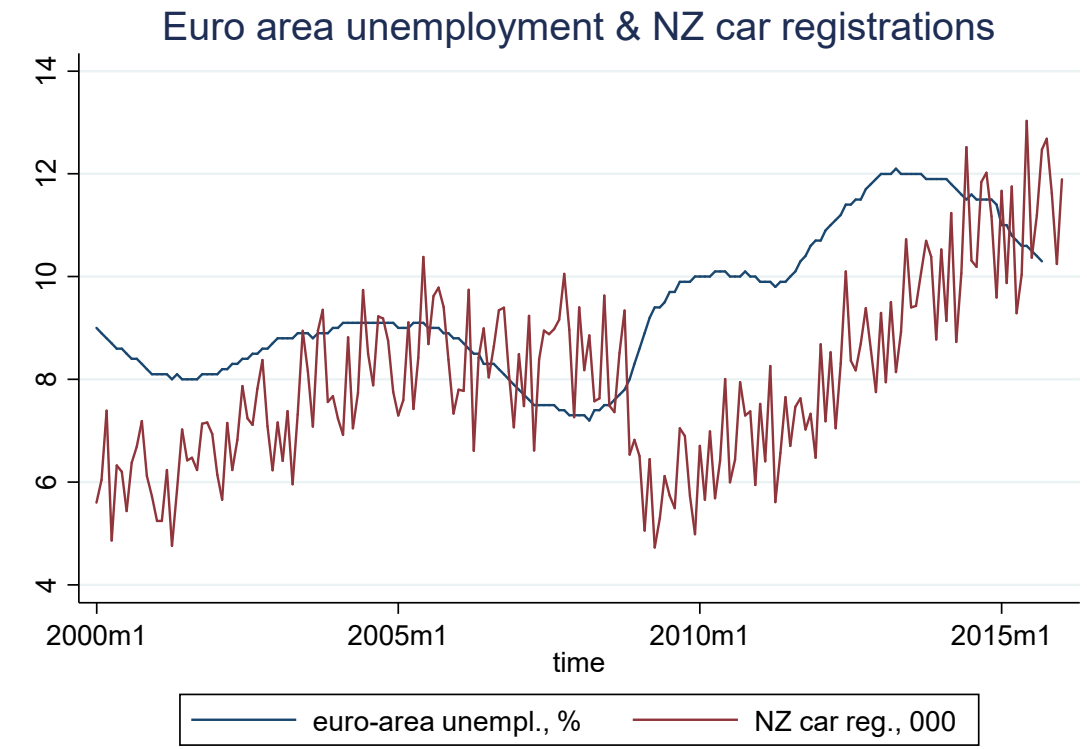
- If Y_t has a trend and X_t has a trend, the you may end up "explaining" a trend with a trend, with no real relation between Y_t and X_t. Such a relationship is called a **spurious** correlation.
- Demonstration: EU-unemployment and NZ car sales 2000m1 - 2016m1. **Note:** the model is badly specified on purpose.
- Spurious correlation yet another reason to test for a unit root.

Graph on EU unemployment and NZ car registrations

Stata code

```
1
2 twoway  tsline ue_euro || tsline registr, ///
3     title("Euro area unemployment & NZ car registrations") ///
4     xtitle("time") ///
5     legend(lab(1 "euro-area unempl., %") lab(2 "NZ car reg., 000")) ///
6     graphregion(color(white)) bgcolor(white)
```

Graph on EU unemployment and NZ car registrations



Corr(UE_euro, NZ_reg) = 0.40 (p-value 0.000).

1

Regression of EU unemployment on NZ car registrations

Stata code

```
regr ue_euro L.ue_euro  if year < 2016
estimates store L_ue_1
regr ue_euro L.ue_euro L.registr  if year < 2016
estimates store L_ue_2
estimates table L_ue_1 L_ue_2 , b(%7.4f) star(0.1 0.05 0.01) stat(N r2 r2_a aic bic)
estimates stat L_ue_1 L_ue_2
```

Regression of EU unemployment on NZ car registrations

```
. estimates table L_ue_1 L_ue_2 , b(%7.4f) star(0.1 0.05 0.01) stat(N r2 r2_a aic bic)
```

Variable	L_ue_1	L_ue_2
ue_euro		
L1.	0.9987***	1.0068***
registr		
L1.		-0.0174***
_cons	0.0196	0.0818
N	187	187
r2	0.9950	0.9954
r2_a	0.9950	0.9953
aic	-3.3e+02	-3.4e+02
bic	-3.3e+02	-3.3e+02

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

```
. estimates stat L_ue_1 L_ue_2
```

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
L_ue_1	187	-327.331	168.3473	2	-332.6947	-326.2325
L_ue_2	187	-327.331	175.2708	3	-344.5416	-334.8483

Note: N=Obs used in calculating BIC; see **[R] BIC note**.

Breaks in a time series

- A model may not stay constant, i.e., its coefficients may change during the sample period.
- If this happens, the series is said to have a **break**.
- How could you detect one?
- This is going to depend on whether you do or do not know when the break(s) occur.
- As an example, let's study an $ADL(1,1)$ model

Testing for a known break in the time series (Chow-test)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 [D_t(\tau) Y_{t-1}] + \gamma_2 [D_t(\tau) X_{t-1}] + u_t$$

- Notice the $D_t(\tau)$ variable: It takes value 1 iff $t \geq \tau$ and is 0 otherwise.
- $D_t(\tau)$ is interacted with (possibly all) the other variables, and it enters independently, too.
- γ_0 measures whether the constant changes = is different in the period $t = 1, ..., \tau - 1$ than in the period $t \geq \tau$.
- Similarly γ_1 and γ_2 measure whether the coefficients of Y_{t-1} and X_{t-1} are different in the two time periods.

Testing for a known break in the time series (Chow-test)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 [D_t(\tau) Y_{t-1}] + \gamma_2 [D_t(\tau) X_{t-1}] + u_t$$

- $H0$: no break. This is equivalent to

$$\gamma_0 = \gamma_1 = \gamma_2 = 0$$

- Can use the F-test.
- The Chow-test allows for
  1. multiple explanatory variables and lags
  2. the break affecting only a subset of explanatory variables
  3. for multiple breaks

Testing for an unknown break in the time series

- Now the break can be at any date t, $\tau_0 < t < \tau_1$.
- Approach: Try all possible break dates.
- How to choose among them?
- Answer: Always use the F-test (Chow) for a given break date, then compare the tests.

Testing for an unknown break in the time series

- This modified Chow-test is called the **Q**uandt **L**ikelihood **R**atio (QLR) - test.
- The QLR - statistic is the maximum of several F-tests and hence its distribution is not the same as that of an F-test. The distribution depends on
  1. the number of restrictions q
  2. the endpoints

Testing for an unknown break in the time series

- In practice, the endpoints are often chosen as $\tau_0 = 0.15\,T$, $\tau_1 = 0.85\,T$.
- Also the QLR-test can be amended to consider a subset of explanatory variables.
- The QLR - test can detect
  1. a single break
  2. multiple breaks and/or
  3. slow evolution of the regression function.

Forecasting

- A forecast is an **out-of-sample** prediction of the future.
- A forecast can be $h = 1, ...$ steps **ahead**.
- If $h > 1$, the forecast is **multistep**.
- $Y_{T+h|T}$ = forecast of Y_{T+h} using the **true** (but unknown) coefficients.
- $\hat{Y}_{T+h|T}$ = forecast of Y_{T+h} using the data $Y_1, ..., Y_T$ and the estimated coefficients.

Forecasting

- Consider a simple AR(1) model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

- The one-step ahead forecast is:

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \mathbb{E}[u_{T+1}] = \hat{\beta}_0 + \hat{\beta}_1 Y_T$$

- The two-step ahead forecast is:

$$\begin{aligned}\hat{Y}_{T+2|T} &= \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{T+1|T} + \mathbb{E}[u_{T+2}] \\ &= \hat{\beta}_0 + \hat{\beta}_1(\hat{\beta}_0 + \hat{\beta}_1 Y_T) \\ \hat{Y}_{T+2|T} &= \hat{\beta}_0(1 + \hat{\beta}_1) + \hat{\beta}_1{}^2 Y_T\end{aligned}$$

- Can you work out $\hat{Y}_{T+3|T}$, $\hat{Y}_{T+h|T}$ (think: geometric series) ?

The **M**ean **S**quared **F**orecast **E**rror, MSFE

- The MSFE measures the quality of the forecast.

$$MSFE = \mathbb{E}[(Y_{T+h} - \hat{Y}_{T+h})^2]$$

- The **R**oot MSFE (RMSFE) is the square root of MSFE.
- It has the same units as the dependent variable (its standard error).

The **M**ean **S**quared **F**orecast **E**rror, MSFE

- Just like in machine learning, the Oracle forecast is based on the true coefficients of the model (fixing the model) with the following MSFE:

$$MSFE = \mathbb{E}[u_{T+h}^2]$$

- Think of the ADL(1,1) model

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \delta_1 X_{t-1} + u_t$$

- For it, the MSFE for $h = 1$ is given by

$$MSFE = \mathbb{E}[u_{T+1}]^2 + \mathbb{E}[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta_1) X_T]^2$$

The **M**ean **S**quared **F**orecast **E**rror, MSFE

- If the sample is large and the number of explanatory variables small, then most of the MSFE comes from $\mathbb{E}[u_{T+h}^2]$.
- Under stationarity, the mean forecast error is zero and then $\mathbb{E}[u_{T+h}^2] = var(u_{T+h})$.
- Just like in machine learning, the second term in MSFE comes from the fact that the estimated coefficients are different from the true coefficients.

Estimating the MSFE

- Simplest estimate of MSFE, assuming stationarity and ignoring the error from coefficients, is obtained using the **s**tandard **e**rror of the **r**egession (SER):

$$MSFE_{SER} = s^2_{\hat{u}} = \frac{SSR}{T - p - 1}$$

- The **F**inal **P**rediction **E**rror (FPE) calculation takes the error from coefficients into account, but assumes homoskedasticity:

$$MSFE_{FPE} = \frac{T + p + 1}{T} s^2_{\hat{u}} = \frac{T + p + 1}{T - p - 1} \frac{SSR}{T}$$

Estimating the MSFE

- Finally, one can estimate MSFE by using **p**seudo **o**ut-**o**f-**s**ample forecast (POOS).
- POOS requires neither stationarity nor homoskedasticity.
- POOS is based on you "simulating" what the forecasting error would hav been, had you estimated the model in real time.

Estimating the MSFE

- $MSFE_{POOS}$ involves the following steps:
  1. choose a date $P <$ for the start of your POOS sample. For example, $P = 0.8T$.
  2. estimate your model using the data $s = P - 1$ and calculate $\hat{Y}_s = \hat{Y}_P$ and $\hat{u}_s = Y_s - \hat{Y}_s$
  3. repeat this for all $s = P - 1, ..., T - 1$.
  4. take the squares of $\hat{u}_s$ and sum them up.

- You get

$$MSFE_{POOS} = \frac{1}{P} \sum_{s=T-P+1}^{T} \hat{u}_s^2$$

Constructing the **Forecast Intervals**

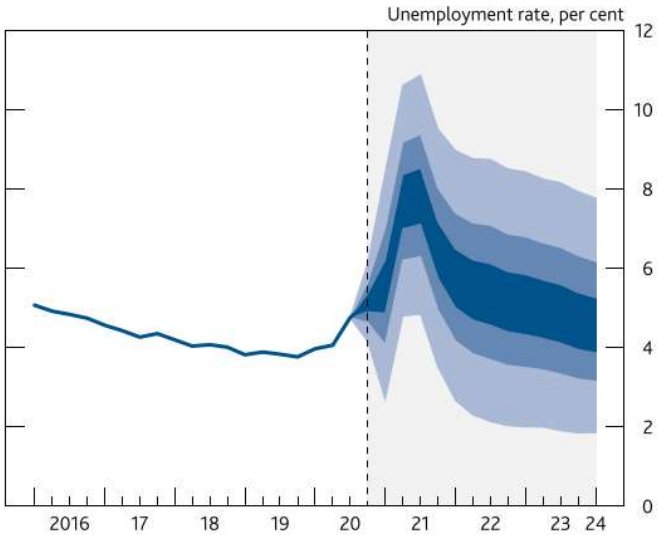
- If $\hat{u}_{T+1}$ is normally distributed, then the 95% forecast interval is given by

$$\hat{Y}_{T+1|T} \pm 1.96 \times RMSFE$$

- This is exactly correct only if $\hat{u}_{T+1}$ is normally distributed, but is often used as it is a good approximation.
- Frequently 67% forecast intervals used, i.e., $\hat{Y}_{T+1|T} \pm RMSFE$ are used.

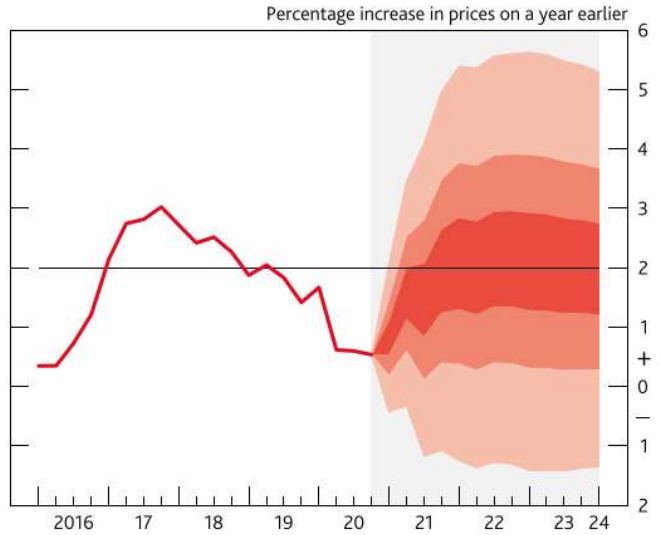
Example: BoE Monetary report forecasts for unemployment and inflation

Chart 1.3: Unemployment projection based on market interest rate expectations, other policy measures as announced



The fan chart depicts the probability of various outcomes for LFS unemployment. It has been conditioned on the assumptions in **Table 1.A** footnote (b). The coloured bands have the same interpretation as in **Charts 1.1** and **1.2**, and portray 90% of the probability distribution. The calibration of this fan chart takes account of the likely path dependency of the economy, where, for example, it is judged that shocks to unemployment in one quarter will continue to have some effect on unemployment in successive quarters. The fan begins in 2020 Q4, a quarter earlier than for CPI inflation. That is because Q4 is a staff projection for the unemployment rate, based in part on data for October and November. The unemployment rate was 5.0% in the three months to November, and is projected to be 5.1% in Q4 as a whole. A significant proportion of this distribution lies below Bank staff's current estimate of the long-term equilibrium unemployment

Chart 1.4: CPI inflation projection based on market interest rate expectations, other policy measures as announced



The fan chart depicts the probability of various outcomes for CPI inflation in the future. It has been conditioned on the assumptions in **Table 1.A** footnote (b). If economic circumstances identical to today's were to prevail on 100 occasions, the MPC's best collective judgement is that inflation in any particular quarter would lie within the darkest central band on only 30 of those occasions. The fan chart is constructed so that outturns of inflation are also expected to lie within each pair of the lighter red areas on 30 occasions. In any particular quarter of the forecast period, inflation is therefore expected to lie somewhere within the fans on 90 out of 100 occasions. And on the remaining 10 out of 100 occasions inflation can fall anywhere outside the red area of the fan chart. Over the forecast period, this has been depicted by the light grey background. See the box on pages 48–49 of the May 2002 *Inflation Report* for a fuller description of the fan chart and what it