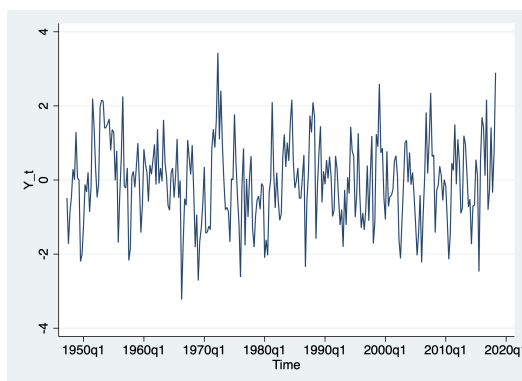


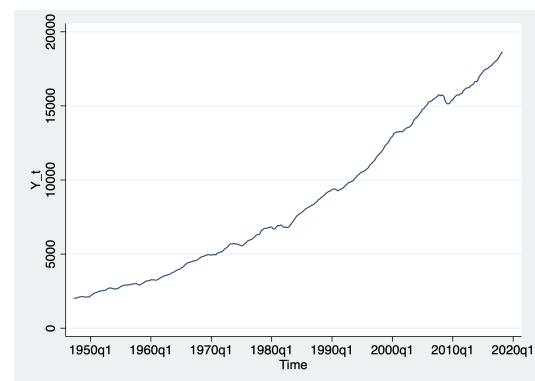
## Dis 12: Time Series (Cont'd)

### 1 Stationarity vs. Nonstationarity

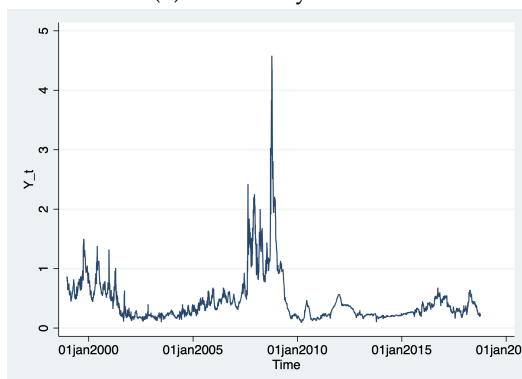
- Last week, we focused on forecasting variable  $Y_t$  that satisfies the stationarity assumption.
  - For stationary  $Y_t$ , two styles of models are proposed:  $AR(p)$  and  $ADL(p, q)$ .
  - And we saw how to use AIC and BIC to determine how many lags to include in an  $AR(p)$  or  $ADL(p, q)$  model.
- But stationarity is not always a given. There are a couple cases where stationarity would not hold:
  - Trend:** Data exhibits a persistent long-term movement over time.
  - Breaks:** Parameters in a regression model would change during the span of the sample due to changes in the economy (policy enactment, innovation, specific event, or else).
  - Seasonality:** Data exhibits similar movement for the same month / same quarter / same day of the week / same time of the day / etc.
- How to detect nonstationarity?
  - A very informal method: **Plot the time series**



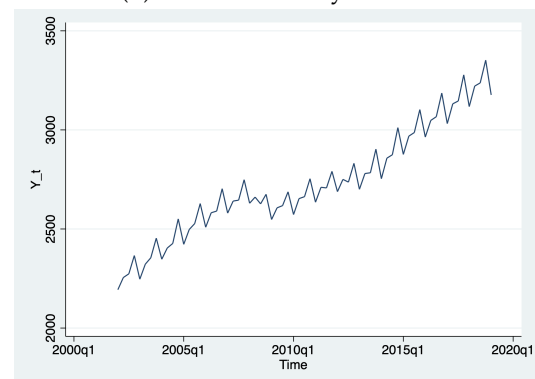
(a) Stationary series



(b) Nonstationarity: Trend



(c) Nonstationarity: Breaks



(d) Nonstationarity: Seasonality (and trend)

Stata command to plot time series (after `tsset` time variable in data):

```
tsline VARNAME
```

– A slightly less informal method: **Plot the Autocorrelation Function (ACF) or Partial Autocorrelation Function (PACF)**

\* Recall from last discussion:

·  $j$ -th order autocorrelation =  $Corr(Y_t, Y_{t-j}) = \frac{Cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)Var(Y_{t-j})}}$

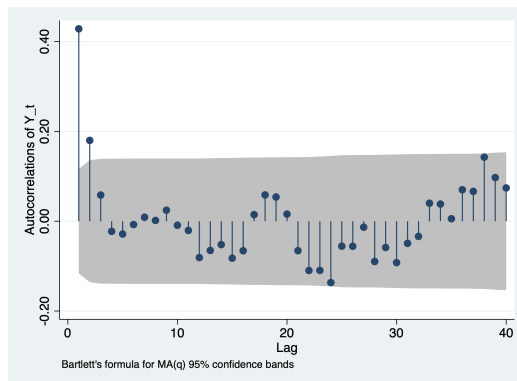
· Replacement of OLS independence assumption for stationary time-series data: When time gap becomes large,  $Y_t$  should become independent with its lag value.

\* These imply that, **if the data is stationary, then autocorrelation should quickly reduce to 0** (i.e. only after a few lags).

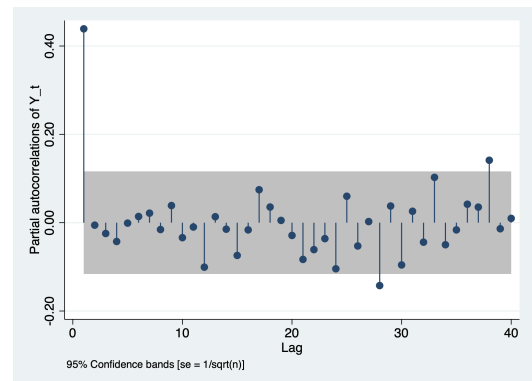
**If the data is nonstationarity, then autocorrelation will very slowly reduce to 0** (i.e. after many lags).

\* What about partial autocorrelation?

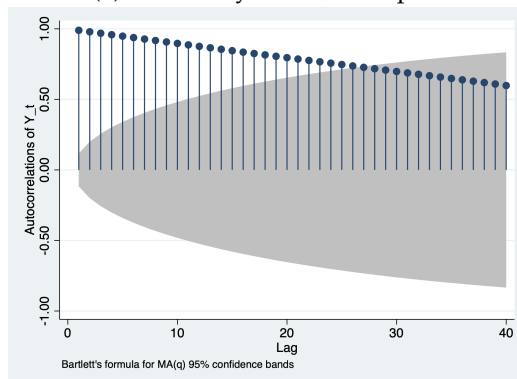
⇒ Similar to partial correlation: It controls for other  $X$  variables that could affect both the lag of  $Y$  and the current  $Y$  value.



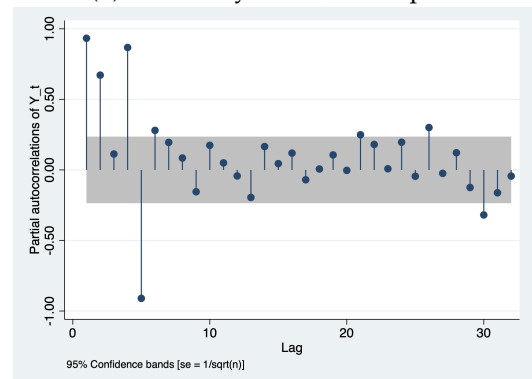
(a) Stationary series, ACF plot



(b) Stationary series, PACF plot



(c) Nonstationary series, ACF plot



(d) Nonstationary series, PACF plot

\* How to create ACF and PACF plots in Stata (after `tsset` time variable in data)?

`ac VARNAME`

`pac VARNAME`

– A formal method to test stochastic trend: **Dickey-Fuller test**

\* Distinguish two types of trend:

1. **Deterministic trend:** Trend is a nonrandom function of time.
2. **Stochastic trend:** Trend is random and varies over time.

\* Although both types of trend imply nonstationarity ...

- ... deterministic trend can simply be dealt with in the model by adding time term(s)
- ... stochastic trend resembles a **random walk** process  $\Rightarrow$  causes all sorts of trouble

\* Random walk:

- $AR(1)$  random walk:

$$Y_t = Y_{t-1} + u_t$$

- $AR(1)$  random walk with a drift:

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

where  $u_t$  is i.i.d. random error.

- Both models imply that the best predictor of  $Y_t$  is its previous period value  $Y_{t-1}$  (with a constant  $\beta_0$  term added or not), given that  $u_t$  is random.
- Under random walk, coefficient on  $Y_{t-1}$  (call this coefficient  $\beta_1$ ) has that  $|\beta_1| = 1$ , which is why random walk / stochastic trend is also called a **unit root** process.  
(\*see Ch 15.7 for how the definition of unit root is adjusted for an  $AR(p)$  model.)
- \* Why is random walk / stochastic trend / unit root a bad thing?
  - **Makes forecasting useless:** The best forecaster of tomorrow's value is just today's value in an  $AR(1)$  random walk process.
  - **Causes problems in estimation:** OLS creates downward bias on estimator\*, and distribution of OLS estimator becomes nonnormal even in large sample time period.
  - **Leads to spurious regression:** Regress one random walk process onto another will imply a nonexistent relationship.
- \* Dickey-Fuller test can test for unit root process for an  $AR(p)$  model with and without deterministic trend component:
  - Without deterministic trend, for an  $AR(1)$  model without drift:

$$\Delta Y_t = \delta Y_{t-1} + u_t$$

- Without deterministic trend, for an  $AR(1)$  model with drift:

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

- With deterministic trend, for an  $AR(1)$  model:

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + u_t$$

- Hypothesis for all cases:  $H_0 : \delta = 0$  (stochastic trend),  $H_1 : \delta \neq 0$  (series is stationary).

---

\*Explanation on why OLS estimator is biased downward (math warning)

- In Stata (after `tsset` time variable in data),

```
dfuller VARNAME // without deterministic trend, without drift
dfuller VARNAME, drift // without deterministic trend, with drift
dfuller VARNAME, trend // with deterministic trend
```

```
dfuller VARNAME, lags(4) // AR(4), without deterministic trend or drift
```

- \* What to do when Dickey-Fuller test concludes that the series is not stationary?
  - Easy solution: The difference of the series usually then is stationary.

```
. dfuller Y_t
```

Dickey-Fuller test for unit root Number of obs = 284

	Test Statistic	Interpolated Dickey-Fuller		
		1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	<b>5.521</b>	<b>-3.457</b>	<b>-2.879</b>	<b>-2.570</b>

MacKinnon approximate p-value for Z(t) = **1.0000**

```
. gen Y_t_diff = Y_t - l.Y_t
(1 missing value generated)
```

```
. dfuller Y_t_diff
```

Dickey-Fuller test for unit root Number of obs = 283

	Test Statistic	Interpolated Dickey-Fuller		
		1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	<b>-10.792</b>	<b>-3.457</b>	<b>-2.879</b>	<b>-2.570</b>

MacKinnon approximate p-value for Z(t) = **0.0000**

- A formal method to test breaks: **Chow test and QLR test**

- \* If we **know** that data is breaking at certain date  $\tau$  (usually due to policy enactment), consider the following model:

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

where  $D_t(\tau)$  is a binary variable that equals to 1 after the breakdate  $\tau$ .

In this case, perform a joint F-test on  $\gamma_0 = \gamma_1 = 0$  tells us whether the forecast actually changes at breakdate  $\tau$ . This test is called **Chow test**.

- \* If we **don't know** that data is breaking at certain date, we can test whether between time  $\tau_0$  and  $\tau_1$ , there exists a breakdate  $\tau$  (so that  $\tau_0 \leq \tau \leq \tau_1$ ).

Idea: Search for all  $\tau$ s over  $\tau_0$  and  $\tau_1$ , and find the largest F-statistic for a model similar to the one presented in Chow test. The date with the largest F-statistic can then be tested for statistical significance to determine whether it is a breakdate.

This test is called a **Quandt likelihood ratio (QLR) test**.

- \* How to deal with breaks?
  - If only some coefficients are breaking, include interaction terms between the predictors and the binary variable indicating breaking.
  - If all coefficients are breaking, run two separate regressions: one using the data before breakdate, one using the data after.

- Roadmap for dealing with time series data:

1. Test for stationarity of series by any (or a combination) of the three following methods:
  - (a) Plot time series
  - (b) Plot ACF or PACF
  - (c) Run Dickey-Fuller, Chow, and/or QLR test
2. If the data is stationary, use  $AR(p)$  or  $ADL(p, q)$  model. Number of lags can be determined by qualitative reasons, or AIC or BIC.
3. If the data is not stationary,
  - For stochastic trend, create a series of differences, and then regress on the differenced series.
  - For deterministic trend, include a time term  $t$  in your regression model.
  - For breaks, include interaction terms between the predictors and the binary variable indicating breaking, or estimating two separate regressions (using before and after breakdate data separately).

## 2 ARIMA Model

- An extension to  $AR(p)$  model, where  $ARIMA(p, d, q)$  stands for
  - $AR(p)$ : Autoregression of order  $p$
  - $I(d)$ : Integration (referring to taking  $d$  differences to the dependent variable)
  - $MA(q)$ : Moving average of order  $q$  (a way to model the error term in time-series)
- There's a Stata command, `arima`, that runs  $ARIMA$  model. The main addition to our current way of modeling is the  $MA$  component, but it usually doesn't differ greatly from a model with just the  $ARI$  components.

### 3 Problems

1. This week's exercise is to help you come up with the appropriate model for a time-series data.
  - (a) An initial hurdle when dealing with time-series data is to tell Stata which variable in my dataset records time. Let's practice using `tsset` on the following four sets of data:
    - Date information recorded in two columns: Cement data from Wooldridge  
<http://fmwww.bc.edu/ec-p/data/wooldridge/cement.dta>
    - Daily data: S&P 500 index  
<https://scaotravis.github.io/downloads/teaching/sp21-400/datasets/sp500.csv>
    - Monthly data: Personal savings rate  
<https://scaotravis.github.io/downloads/teaching/sp21-400/datasets/psr.csv>
    - Quarterly data: Private fixed investment  
<https://scaotravis.github.io/downloads/teaching/sp21-400/datasets/pfi.csv>
  - (b) Now that we have the quarterly data correctly `tsset`, let's figure out whether the `pfi` series is stationary before we propose any specific forecasting model. Plot the `pfi` series, its ACF and PACF. Can you guess whether this series is stationary or not?
  - (c) Perform two Dickey-Fuller tests
    - With drift
    - With trend

Determine whether the series has stochastic trend component at 5% size for each test.

(d) Based on your conclusion in part (b) and part (c), can we directly predict the `pfi` series using an  $AR(p)$  model? If not, what transformation on `pfi` needs to be done so that it is stationary?

(e) While we discussed last week that AIC and BIC can be used to determine number of lags to include in an  $AR(p)$  model, an alternative way is to use the ACF plot to decide how many lags to include.

Plot the ACF for the variable that you decide to perform forecast on, and write down the corresponding model you're going to estimate.

(f) Use the sample period before the first quarter of 2005 to estimate your proposed model, and use the rest of the data to construct pseudo out-of-sample estimate of MSFE (mean squared forecast error).