Dynamic Modeling and Forecasting in Big Data

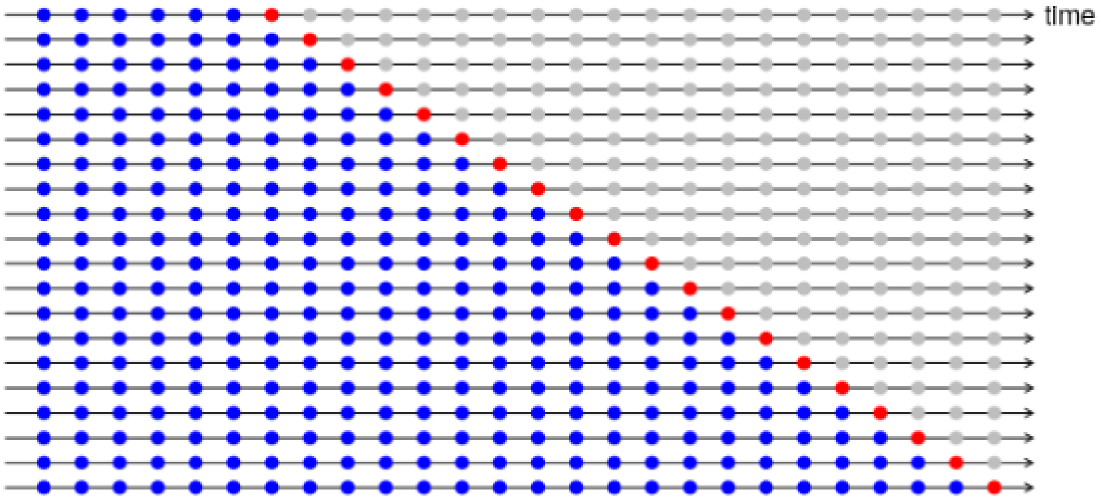
Transforming Data into Knowledge and Vision
Understanding the Power and Beauty of Data

Reduced form / time series models

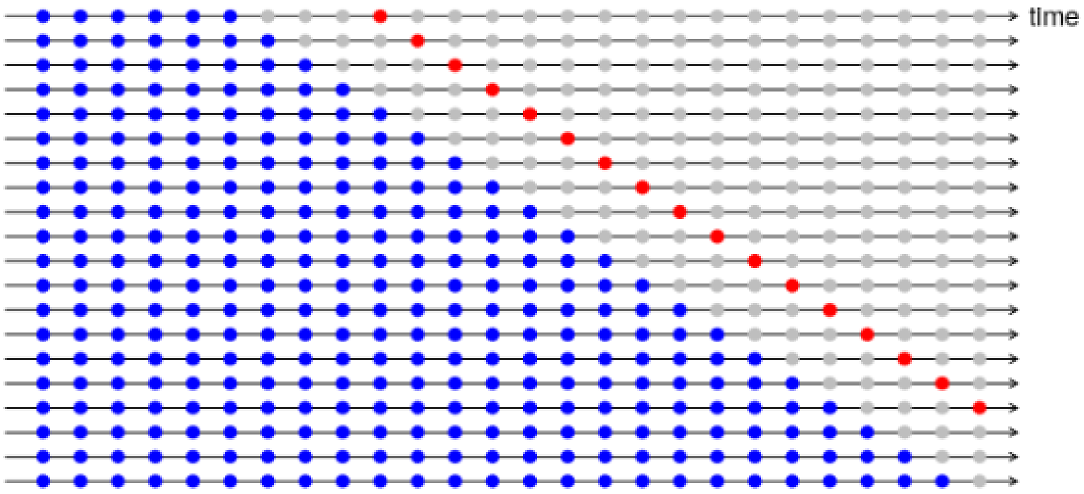
William Yu, PhD
Economist
UCLA Anderson Forecast

Time series cross-validation
Out-of-sample/testset

One-step-ahead



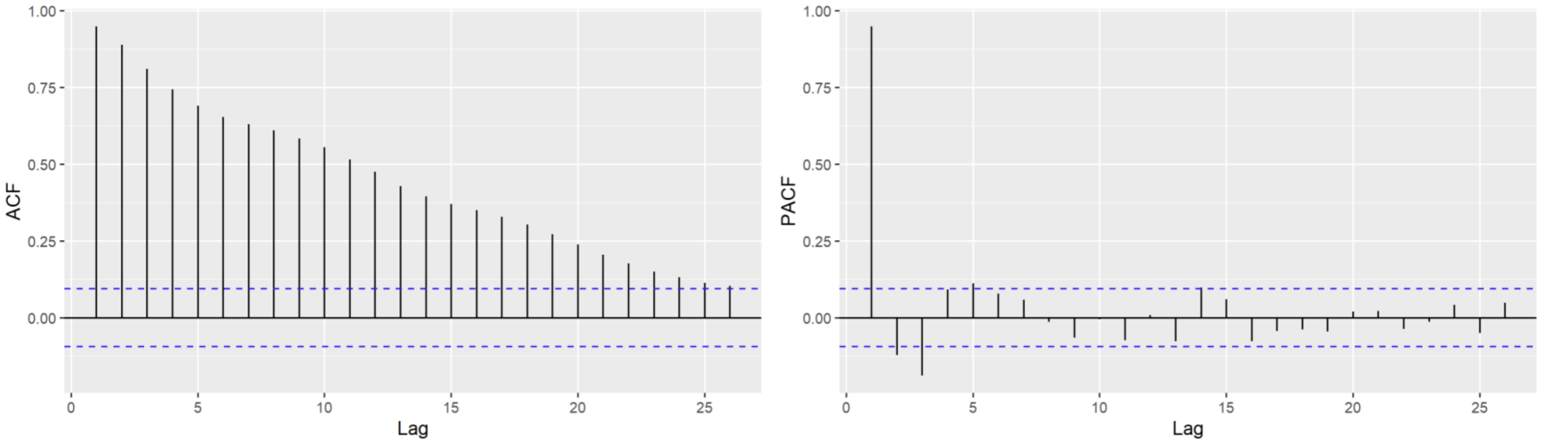
Four-step-ahead



The Box-Jenkins Model-ARMA models

- **Autoregressive model (AR)** use the variable's past values to predict the current/future variable
  - It is appealing because the predictor are observable
  - AR(1) – first order AR $y_t = \alpha + \beta y_{t-1} + \epsilon_t$ → $y_{t+1} = \alpha + \beta y_t + \epsilon_{t+1}$ → $y_{t+2} = \alpha + \beta y_{t+1} + \epsilon_{t+2}$ → …
  - AR(2) – second order AR $y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t$
  - AR(P)
- **Moving average model (MA)** use the past forecast errors (shocks) to predict the current/future variable
  - The shocks are not observable
  - MA(1) – first order MA $y_t = \alpha + \epsilon_t + \beta \epsilon_{t-1}$
  - MA(2) – second MA $y_t = \alpha + \epsilon_t + \beta_1 \epsilon_{t-1} + \beta_2 \epsilon_{t-2}$
  - MA(q)
- **ARMA** (Autoregressive and moving average)
  - ARMA(1,1) $y_t = \alpha + \beta_1 y_{t-1} + \epsilon_t + \beta_2 \epsilon_{t-1}$
  - ARMA(2,1) $y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \epsilon_t + \beta_3 \epsilon_{t-1}$

Preliminary check on ARMA orders



- Use ACF (autocorrelation) chart and partial autocorrelation (PACF) chart to decide orders of ARMA models.
- An AR process has a geometrically decaying ACF. Use PACF to decide AR order.
- An MA process has a geometrically decaying PACF. Use ACF to decide MA order.
- ACF vs PACF. The ACF are just “simple” or “regular” correlations b/w y_t and y_{t-p}. PACF measures the association b/w y_t and y_{t-p} after controlling other lags.

Autocorrelation function (ACF) vs PACF (Partial ACF)

- ACF shows how the time series correlates with itself at different lags.
- PACF indicates the correlation at lag k, removing the effect of any correlations due to terms at shorter lags. PACF could show how many lags of AR model.

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3}$$

- Y$= \alpha + \beta_1 X_1$

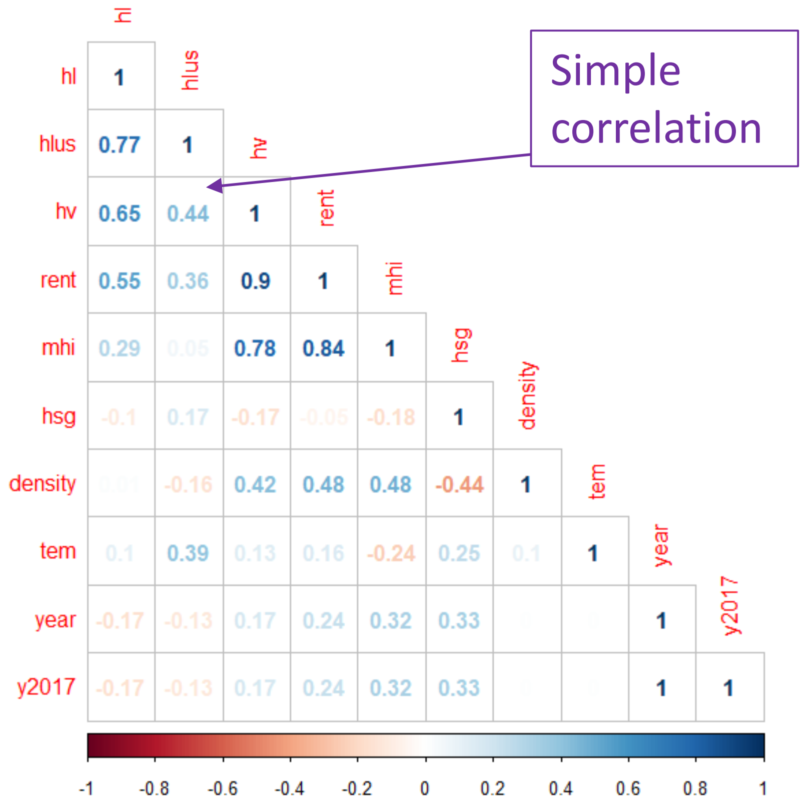
  hl hv

  homelessness home value

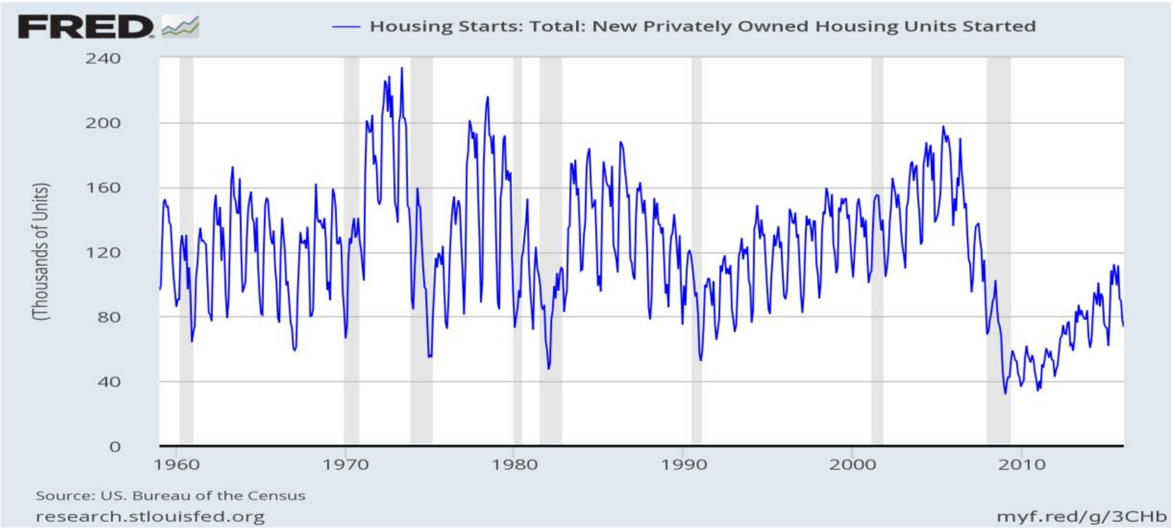
- Y$= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

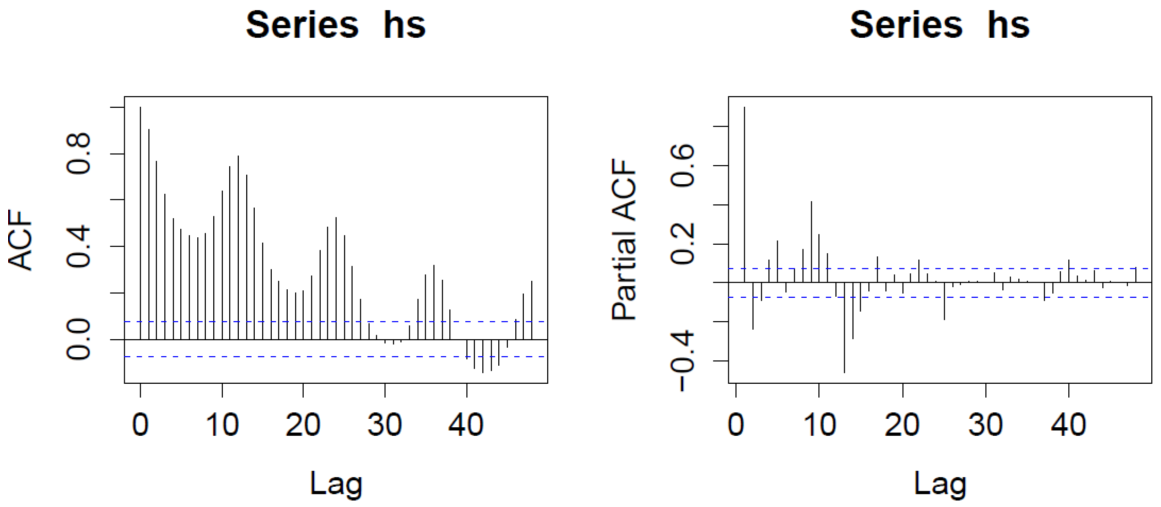
  hl hv mhi hsg

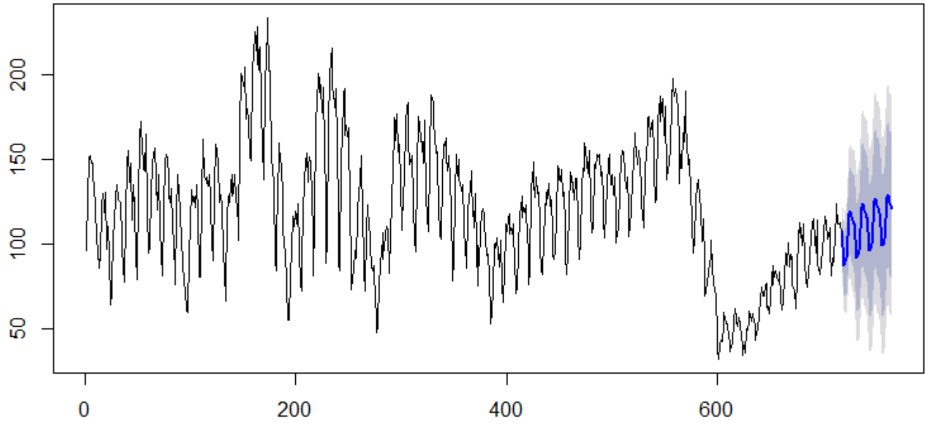
  homelessness home value median income home supply gorwth



Seasonal ARMA model

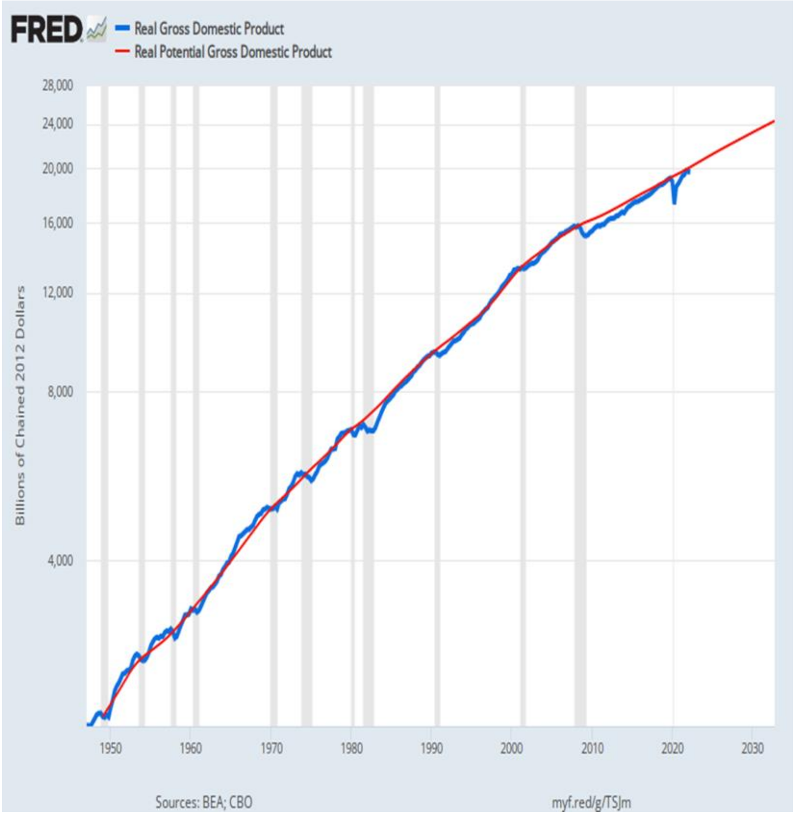






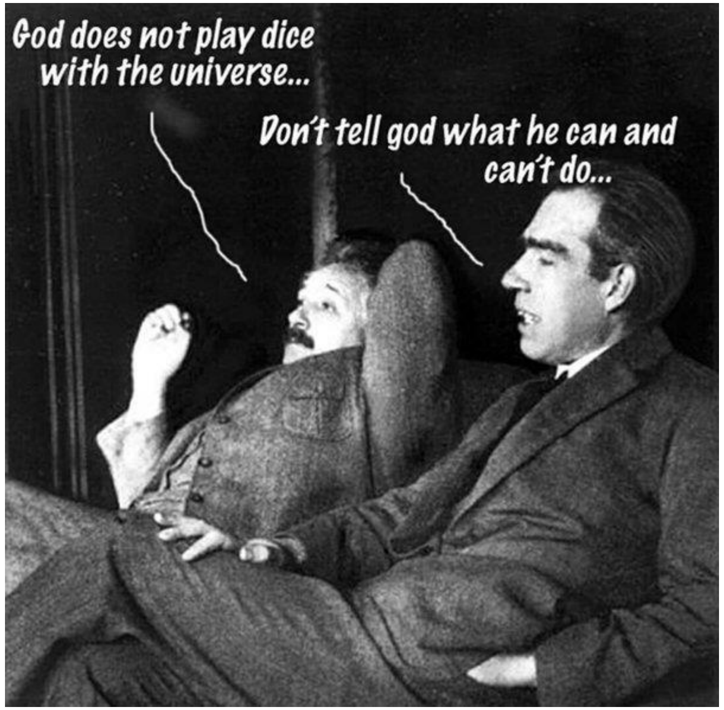
Stationarity vs Nonstationarity

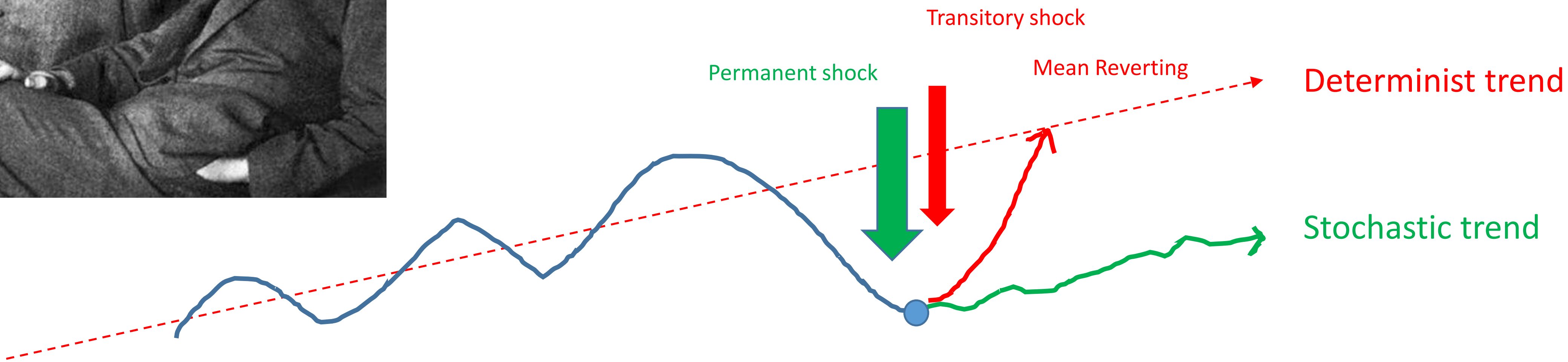
- Q: Stationarity is the necessary condition for forecasting. What should we do if the data is not stationary? Give up? NO WAY!
- Solutions: Transform the non-stationary data to stationary data.
  - *Logarithms transformation (or Cox-Box transformation)*
  - ***Detrend*** for data with a **deterministic trend**
  - ***Take first difference*** for data with a **stochastic trend:** Y_t - Y_{t-1}
  - **Growth rate:** $\ln(Y_t)$ - $\ln(Y_{t-1})$
- Remember Model 1, 2, 3, trend model, which is deterministic (secular).
  - After fitting the trend, we can go further to estimate the rest parts, such as seasonal and cycle components.
- Stochastic trend says the trend is not deterministic. It is also called ***random walk***.



Random walk

- Image a drunk man walking down the street.
  - Can he walk a straight line?
  - Can anyone predict his next move?
- Stationary series/deterministic trend:
  - The shock hitting the series is temporary.
  - Any deviation will come back to the trend. (Mean Reversion).
  - ACF die out exponentially.
  - I(0) process. $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$ where $\beta_1 < 1$, which is a typical AR(1) model.
- Stochastic trend:
  - The shock hitting the series is permanent.
  - Any deviation will NOT come back to the trend.
  - ACF die out hardly.
  - I(1) process. $Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$ where $\beta_1 = 1$. We call this process having a Unit Root.


God does not play dice with the universe...
Don't tell god what he can and can't do...

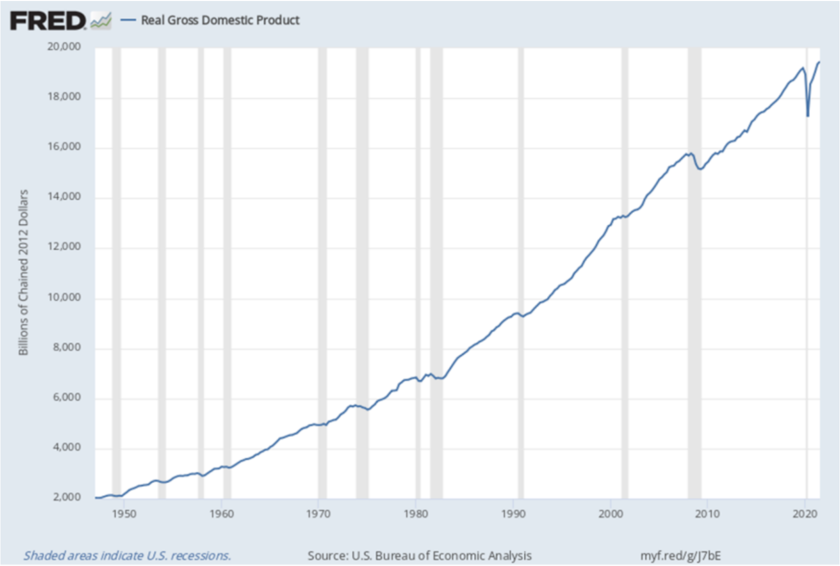

Transitory shock
Permanent shock
Mean Reverting
Determinist trend
Stochastic trend

ARIMA model

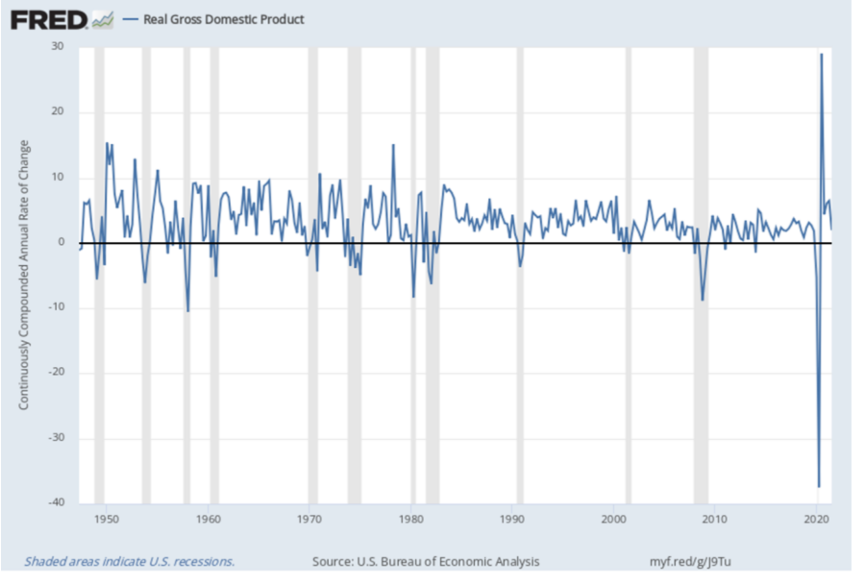
(Autoregressive Integrated Moving Average)

- Take difference to ensure stationarity!
- $\Delta Y_t = Y_t - Y_{t-1}$
- If the process could be converted to stationary process I(0) by taking difference once, called first difference, it is called I(1) process, d=1.
- If the process could be converted to stationary process I(0) by taking difference twice, called second difference, it is called I(2) process, d=2.
- Therefore, we call this ARIMA (p,d.q) model.
- For example, ARIMA (2,1,3) is the model that takes the first difference of the data and then the cyclical part could be modeled by ARMA(2,3).
- Before, when we model the GDP and housing price, we didn't model their level data directly. Rather we model their growth rate. That is the spirit of the first difference.

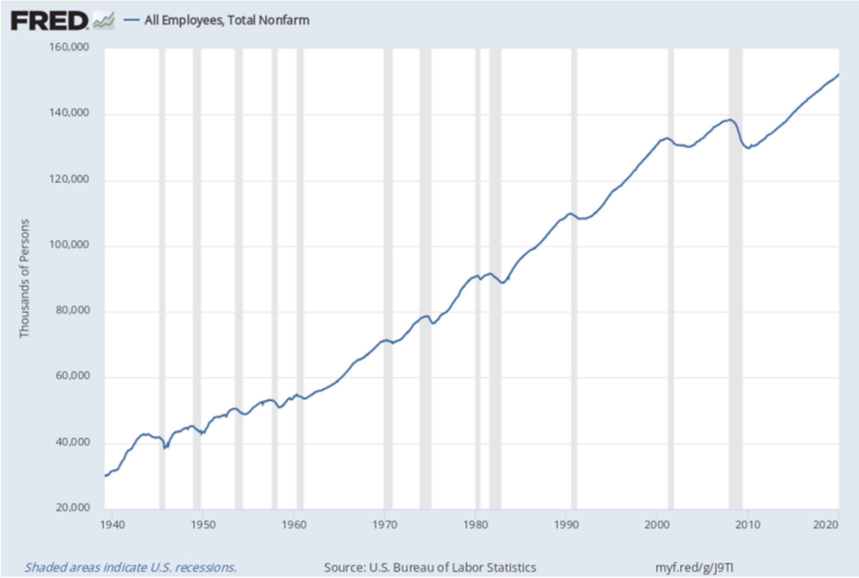
Real GDP level -- nonstationary



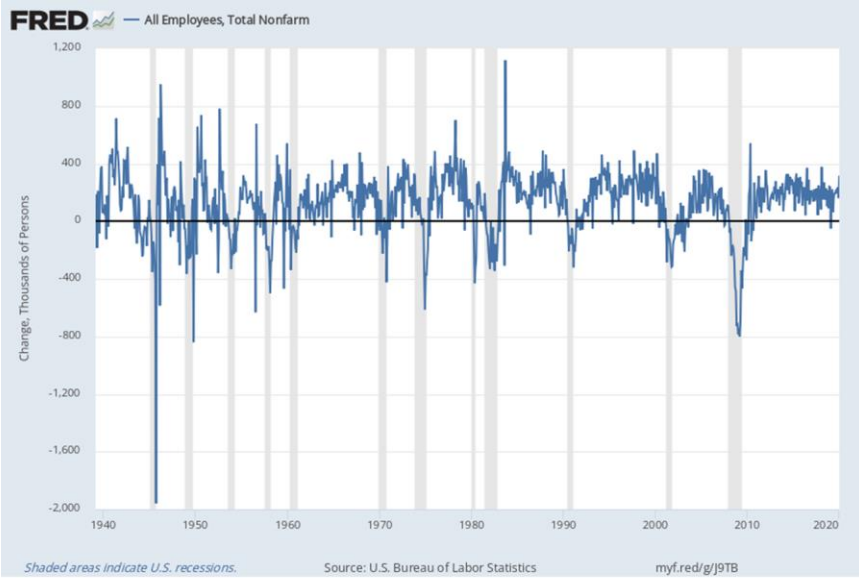
Real GDP growth rate -- stationary



Payroll job level -- nonstationary



Payroll job change -- stationary



- White noise: ARIMA(0,0,0)
- Autoregression: ARIMA(p,0,0)
- Moving average: ARIMA(0,0,q)
- Pure Random walk: ARIMA(0,1,0) (unpredictable)
- Random walk: ARIMA(p,1,q) (somewhat predictable)
- Random walk with drift: ARIMA (0,1,0) with a drift

SARIMA (Seasonal ARIMA) model

- SARIMA(p,d,q)(P,D,Q)$_m$
- M stands for frequency. If the data was recorded annually, m=1; quarterly, m=4, monthly, m=12.
- P is the order of the seasonal AR(P) process, D is the seasonal order of integration, A is th order of seasonal MA(Q) process.
- SARIMA(p,d,q)($0,0,0$)$_m$ = ARIMA(p,d,q)
- Example: m = 12. If P = 2, this means that we include two past values of the series at a lag that is multiple of m, which are y_{t-12} and y_{t-24}
- If D = 1, this means that a seasonal difference makes the series stationary $\Delta y = y_t - y_{t-12}$
- If Q = 2, we include past error terms ϵ_{t-12} and ϵ_{t-24}

Table 8.1 Appropriate frequency *m* depending on the data

Data collection	Frequency *m*
Annual	1
Quarterly	4
Monthly	12
Weekly	52

Table 8.2 Appropriate frequency *m* for daily and sub-daily data

Data collection	Frequency *m*				
	Minute	Hour	Day	Week	Year
Daily				7	365
Hourly			24	168	8766
Every minute		60	1440	10080	525960
Every second	60	3600	86400	604800	31557600

Source: Time Series in Python, Marco Peixeiro 2022