ARIMA and Dynamic Regression

Xiaodong Lin

ARIMA models

- So far, we have considered the **ARMA** family of models, which rely on the assumption of stationarity.
- We now consider a more general family that allows the modeling of nonstationary time series through the application of differencing.
- The simplest example is the random walk example we discussed previously. Recall that, we defined the random walk $\{X_t\}$ as

$$X_t = X_{t-1} + w_t, \quad \text{where } \ w_t \sim \mathrm{WN}(0, \sigma^2)$$

$\{X_t\}$ is a nonstationary AR(1) process. However, $\{\nabla X_t\}$ with

$$\nabla X_t = X_t - X_{t-1}$$

is a stationary process, being just the white noise w_t.

ARMIA models

- Consider the following nonseasonal model with trend

$$X_t = m_t + Y_t$$

  where m_t is a polynomial of order k and Y_t is a stationary process.
- $\{X_t\}$ is nonstationary since it has trend (polynomial).
- The operation $\nabla^k X_t$ removed the trend and yielded a stationary time series that can be analyzed with the ARMA.

Definition

If d is a nonnegative integer, then

$$\{X_t\} \text{ is an } \textbf{ARIMA}(p, d, q) \text{ process if}$$

$$Y_t = (1 - B)^d X_t \text{is an } \textbf{ARMA}(p, q) \text{ process.}$$

ARMIA models

- ARIMA(p,d,q) model
  AR: p=order of the autoregressive part.
  I: d=degree of first differencing involoved.
  MA: q=order of the moving average part.
- Examples:
  - White noise model: ARIMA(0,0,0)
  - Random walk: ARIMA(0,1,0) with no constant
  - Random walk with drift: ARIMA(0,1,0) with constant
  - AR(p): ARIMA(p,0,0); MA(q): ARIMA(0,0,q)
- Using backshift notation
  ARIMA(1,1,1) model with constant:

$$(1-\phi_1 B)(1-B)X_t = \phi_0 + (1+\theta_1 B)W_t.$$

equivalent to

$$X_t = \phi_0 + X_{t-1} + \phi_1 X_{t-1} - \phi_1 X_{t-2} + \theta_1 W_{t-1} + W_t.$$

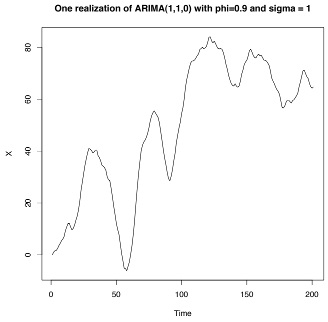
ARMIA models

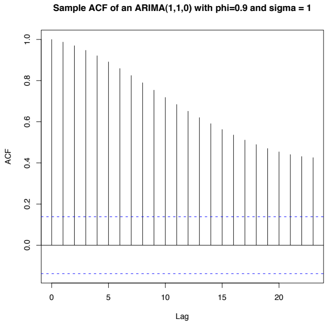
- ARIMA stands for Integrated ARMA. The model can be written as
$$\phi(B)(1-B)^d X_t = \phi_0 + \theta(B) w_t$$
where $\phi_0 = \mu(1-\phi_1 - \cdots - \phi_p)$
- The polynomial $\phi(z)(1-z)^d$ has a unit root with multiplicity d, the process is still nonstationary even if the roots of $\phi(z)$ are different from 1. However, $\{\nabla^d X_t\}$ is stationary.

Simulated example

Simulate an **ARIMA**$(1,1,0)$ with $\phi = 0.9$ and $\sigma^2 = 1$.

```
X <- arima.sim(list(order = c(1,1,0), ar = 0.9), n = 200)
```





The ARIMA model

- The need for **ARIMA** arises from the fact that the series $\{X_t\}$ is nonstationary.
- Apply **differencing** operator $\nabla = 1 - B$ until the transformed series $\{Y_t = \nabla^d X_t\}$ exhibits stationarity.
- Test non stationarity using unit-root test.

$$X_T = \phi_1 X_{t-1} + w_t.$$

Test $H_0 : \phi_1 = 1 \;\; v.s. H_1 : \phi_1 < 1$. Using the usual t-stat.

- `adfTest(X, lag=5)`

```
STATISTIC:Dickey-Fuller: 0.4971
P VALUE:0.7737
```

Parameter Estimation

- Based on the original series X_t, we use $p = 1$, $d = 1$ and $q = 0$,

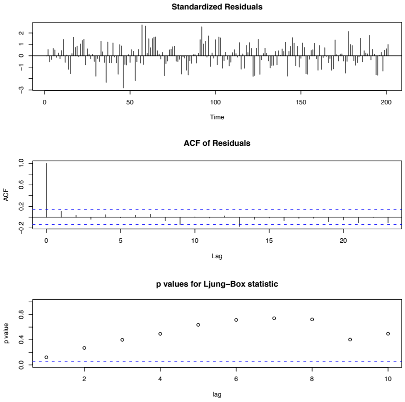
```
M1<-arima(X, order=c(1,1,0))
```

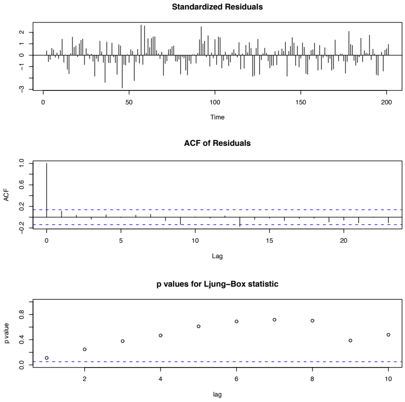
  We find $\hat{\phi} = 0.89$ with $\mathrm{se}(\hat{\phi}) = 0.0318$ and $\hat{\sigma}^2 = 0.8924$.
- Based on the differenced $Y_t = \nabla X_t$, we use $p = 1$, $d = 0$ and $q = 0$,

```
M2<-arima(Y, order=c(1,0,0))
```

  We find $\hat{\phi} = 0.88$ with $\mathrm{se}(\hat{\phi}) = 0.0322$ and $\hat{\sigma}^2 = 0.8906$.
- The diagnostics checks in both cases support the plausibility of the chosen model.

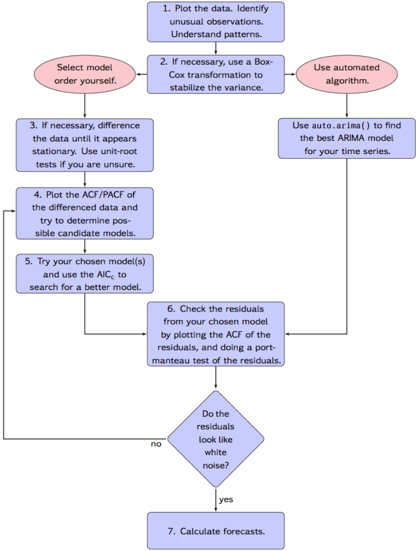
Diagnosis plots





General procedure: Applying differencing until the resulting series has a sample ACF that decays rapidly, and the differenced data can be fitted by a low-order ARMA process.

ARIMA modeling process



Regression with time series errors

- Regression between two (or more) time series
- Residual series of a standard regression has serial correlations.
- Problematic parameter inferences with the serial correlations are ignored. For instance, it may introduce biases in estimates and standard errors.
- Many applications: excess return of an individual stock to market index return; term structure of interest rate.
- Different model assumption to that of usual regression model.

Models

Basic model:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + a_t$$

$$\phi(B) a_t = \theta(B) w_t, \quad w_t \sim N(0, \sigma^2)$$

Alternative form

$$\phi(B)[y_t - (\beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t})] = \theta(B) w_t$$

Model building procedures

1. Run standard regression and y and x, obtain residuals
2. Build a time series model for the estimated residuals. This is to determine the order ect. on the time series component. We are not using the estimated coefficients here.
3. Joint estimate of both the regression component and the error time series component
4. Model checking and diagnostic. Refine and re-estimate.

Prediction

1. Joint estimate of both the regression component and the noise time series component.
2. obtain estimated a_t

$$\hat{a}_t = y_t - \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \cdots + \hat{\beta}_k x_{k,t}$$

3. Use $\hat{a}_t$ and the ARMA model to predict $\hat{a}_t(h)$. Then obtain

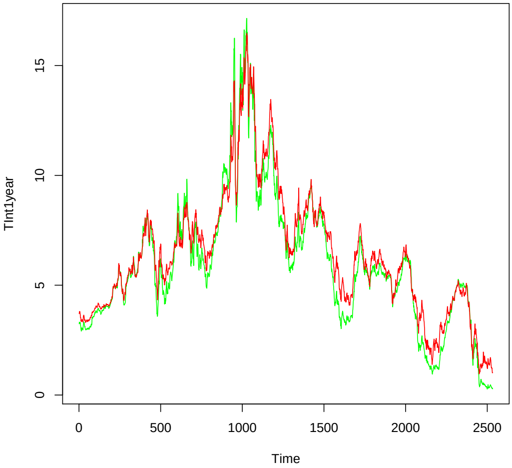
$$\hat{y}_t(h) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t+h} + \cdots + \hat{\beta}_k x_{k,t+h} + \hat{a}_t(h)$$

Example

U.S. weekly interest rate data: 1-year and 3-year constant maturity rates.

```
## Regression with time series error term.
TInt1year=read.csv("1yearTreasury1962to2010.csv",header=T)
### These are weekly data given in percents
TInt1year=
read.csv("1yearTreasury1962to2010.csv",header=T,skip=7)[,2]
TInt3year=
read.csv("3yearTreasury1962to2010.csv",header=T,skip=7)[,2]
plot.ts(TInt1year,col="green")
lines(TInt3year,col="red")
plot(TInt1year,TInt3year,lty=15,pch=20)
## a linear relation seems very appropriate
plot(diff(TInt1year),diff(TInt3year),lty=15,pch=20)
## changes in the rates
model1=lm(TInt3year~TInt1year)
summary(model1)
```

Weekly interest rates comparison



Usual linear regression

- Simple linear regression seem to be very good model with very high R squared.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.860075   0.022553   38.13   <2e-16 ***
TInt1year   0.925883   0.003375  274.31   <2e-16 ***
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.5178 on 2531 degrees of freedom
Multiple R-squared: 0.9675,     Adjusted R-squared: 0.9674
F-statistic: 7.525e+04 on 1 and 2531 DF,  p-value: < 2.2e-16
```

Strong linear relationship

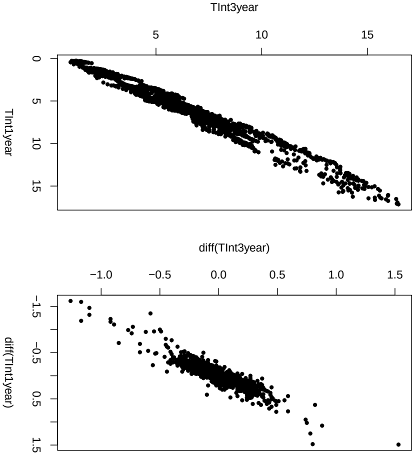


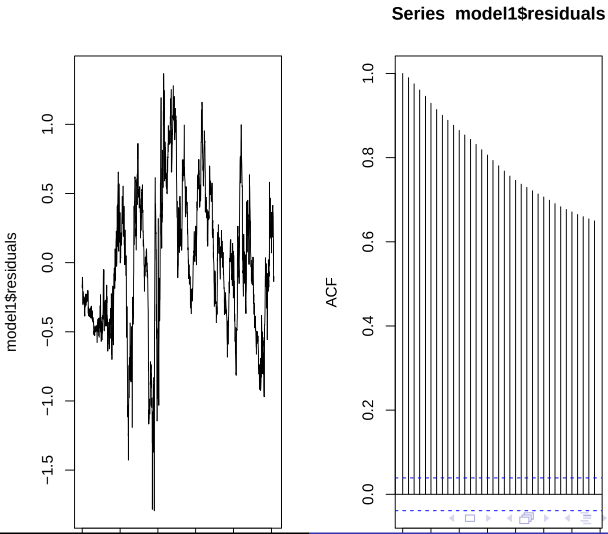
Figure: Linear relationship

Residual serial dependency

- So what's the problem? Look at the residual series.

```
plot(model1$residuals,type="l")
acf(model1$residuals)     ## clearly not white noise
Box.test(model1$residuals,lag=10,type="Ljung")
##null hypothesis= data is not correlated for the first x la
        Box-Ljung test
data:  model1$residuals
X-squared = 21758.7, df = 10, p-value < 2.2e-16
## unit root testing:
library(fUnitRoots)
ar(TInt1year)
adfTest(TInt1year,lag=30)  ## seems there is a unit root
Test Results:
  PARAMETER:
    Lag Order: 30
  STATISTIC:
    Dickey-Fuller: -0.834
  P VALUE:
```

Serial correlation



Stationarity

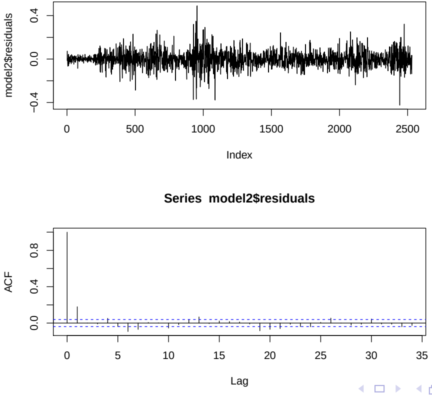
- Consider differences instead and run regression on the difference series.

```
c1=diff(TInt1year)
c3=diff(TInt3year)
model2= lm(c3~c1)
summary(model2)
lm(formula = c3 ~ c1)

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001317  0.0013768  -0.096    0.924
c1           0.7927730  0.0073689 107.583   <2e-16 ***
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.06928 on 2530 degrees of freedom
Multiple R-squared: 0.8206,     Adjusted R-squared: 0.8205
F-statistic: 1.157e+04 on 1 and 2530 DF,  p-value: < 2.2e-16
plot(model2$residuals,type="l")
acf(model2$residuals)
```

Serial correlation of the residuals of the differenced data

- The residual seems stationary now, but clearly not white noise. From the acf plot, we fit MA(1) model for simplicity.



Joint estimation

- ```
  m=arima(x=c3, order=c(0,0,1), xreg=c1, include.mean=F)
  m
  Call:
  arima(x = c3, order = c(0, 0, 1), xreg = c1, include.mean =
  Coefficients:
              ma1      c1
           0.1840  0.7943
  s.e.     0.0192  0.0076
  sigma^2 estimated as 0.004636:  log likelihood = 3210.51,  a
  ```
- The fitted model is $c_{3t} = 0.7943c_{1t} + a_t$, $a_t = w_t + 0.184w_{t-1}$, $\hat{\sigma} = 0.0678$. Thus

$$r_{3t} = r_{3,t-1} + 0.7943(r_{1t} - r_{1,t-1}) + w_t + 0.184w_{t-1}.$$

- Perform usual model diagnosis.
  ```