CMU MSP 36618 Time Series Lecture 8: Unit Roots + Misc.

March 31, 2016

I. Unit roots and the random walk

a. AR1: $Y_t = \phi Y_{t-1} + \varepsilon_t$ or $Y_t = \Phi(\mathrm{B})^{-1}\,\varepsilon_t$ where $\Phi(\mathrm{B})^{-1} = (1-\phi B)^{-1}$ ε_t~N(0,σ^2)

b. If ϕ=1, then the process is no longer "mean reverting", and it is a non-stationary discrete-time continuous-space Gaussian random walk with variance $\propto$ time.

c. We can investigate random walks with the code in LNTS8.R.

II. About unit root testing for non-stationarity

a. **References:**

i. Testing for Unit Roots: What Should Students Be Taught? by John Elder and Peter E. Kennedy, *The Journal of Economic Education*, 32(2), pp. 137-146, 2001. (http://www.jstor.org/stable/1183489)

ii. Stationarity Issues in Time Series Models by David A. Dickey, SUGI 30 (192). (http://www2.sas.com/proceedings/sugi30/192-30.pdf)

b. **Four *realistic* scenarios**

i. Stationary around a fixed mean

ii. Stationary around a linear (or higher order) trend

iii. Non-stationary (random walk) without a trend

iv. Non-stationary (random walk) with a trend (preferred direction)

- Remember that finding a unit-root (random walk) suggests that we can and should analyze the (hopefully stationary) difference ∇Y_t instead of Y_t.

c. **Basic Dickey-Fuller equations (reformulated AR(1) with trend β)**

$Y_t = \rho Y_{t-1} + \alpha + \beta t + \varepsilon_t$

$Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + \alpha + \beta t + \varepsilon_t$

$\nabla Y_t = (\rho - 1) Y_{t-1} + \alpha + \beta t + \varepsilon_t \Leftrightarrow \nabla Y_t = \tau Y_{t-1} + \alpha + \beta t + \varepsilon_t$

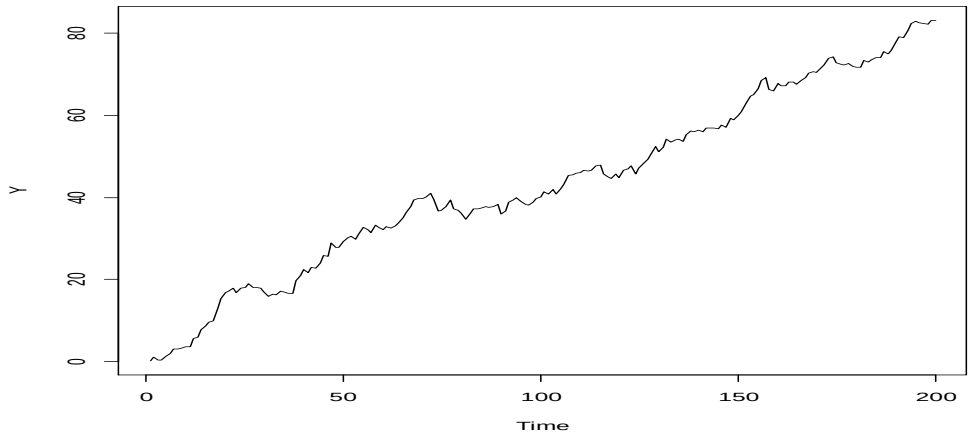
α is the "mean" and β is the "trend".

Testing if the coefficient for Y_{t-1} equals zero is the same as testing for ρ=1 for AR(1), i.e. we have H_0: $\rho - 1 = \tau = 0$ → H_0: ρ=1, so p>0.05 → unit root ***is*** present. (The statistic is labelled "τ" (tau), but the null sampling distribution is ***not*** "t".)

d. **Augmented Dickey-Fuller (ADF):** add more terms ("lags") for detecting unit roots in AR(1+lags) models

e. **Case 1: Y_t is growing** (or shrinking), e.g., DJIA or unemployment count
The scenarios are "$\rho=1$ and $\beta=0$" or "$\rho=1$ and $\beta\neq0$" or "$\rho<1$ and $\beta\neq0$". The first two are unit root (random walk) processes and the third is a stationary growth process around a trend line. First check "tau" $(1-\rho)/SE(1-\rho)$ for the "Trend" section of the ADF test: if $p<0.05$, reject a unit root and conclude that we have AR + trend (always reverting to the trend line). Otherwise we assume a unit root process and try a model for the first difference of Y. Of course, we have type-1 errors (conclude stationary growth when it is really non-stationary) at rate α, and type-2 errors at an unknown rate that is higher when the amount of data is small (due to low power).

Example 1:

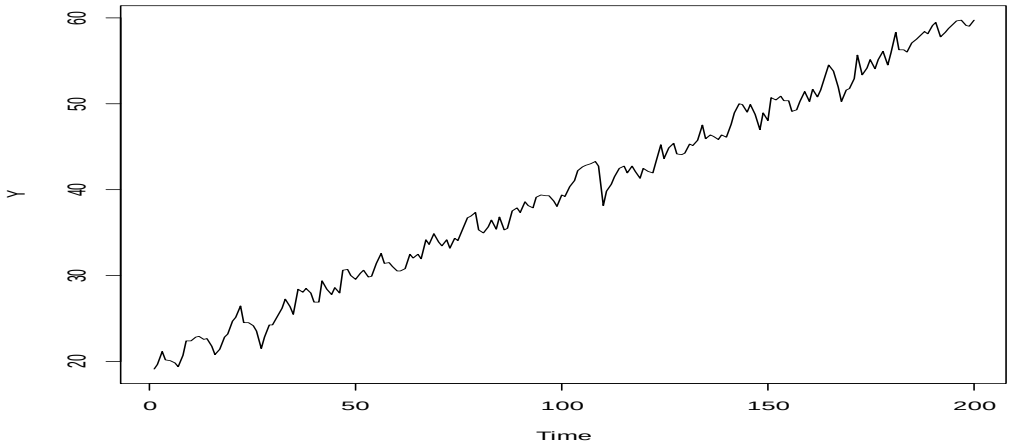



```
PROC ARIMA DATA=simulated;
  IDENTIFY VAR=y STATIONARITY=(ADF);
RUN; QUIT;
```

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	1.4251	0.9610	4.64	0.9999		
	1	1.4094	0.9597	4.00	0.9999		
	2	1.4005	0.9590	3.72	0.9999		
Single Mean	0	-0.5240	0.9248	-0.79	0.8203	16.67	0.0010
	1	-0.4933	0.9270	-0.71	0.8418	12.82	0.0010
	2	-0.6137	0.9181	-0.89	0.7908	12.22	0.0010
Trend	0	-10.2140	0.4123	-2.43	0.3619	3.04	0.5690
	1	-11.2237	0.3483	-2.47	0.3429	3.10	0.5571
	2	-12.2307	0.2921	-2.63	0.2662	3.59	0.4600

Example 2:



Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	0.8726	0.8894	2.32	0.9952		
	1	0.8828	0.8913	2.77	0.9987		
	2	0.8770	0.8902	3.14	0.9996		
Single Mean	0	-0.9533	0.8896	-0.71	0.8402	3.71	0.1266
	1	-0.6547	0.9149	-0.58	0.8701	4.89	0.0421
	2	-0.3484	0.9368	-0.36	0.9121	5.81	0.0169
Trend	0	-95.3645	0.0006	-7.87	<.0001	30.94	0.0010
	1	-108.876	0.0001	-7.31	<.0001	26.70	0.0010
	2	-133.330	0.0001	-6.98	<.0001	24.39	0.0010

f. **Case 2: Y_t is not growing** (or shrinking), e.g., unemployment rate

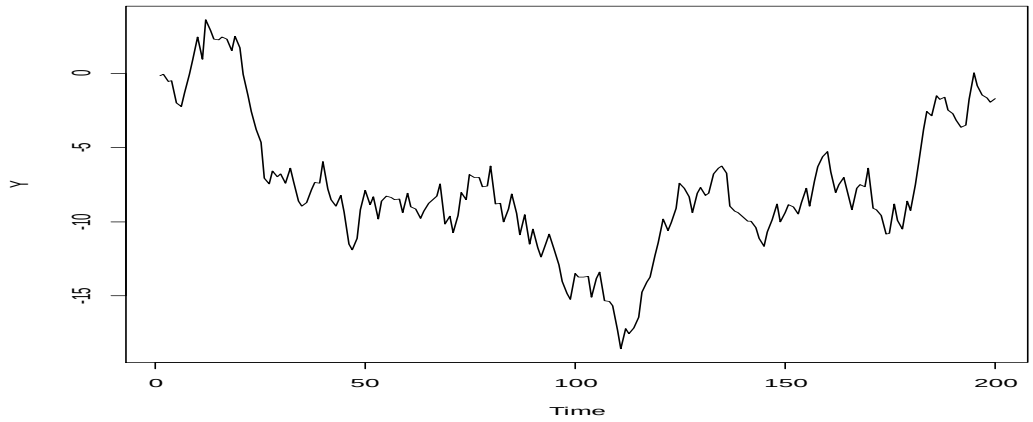
Here we want to fit

$\nabla Y_t = (\rho - 1)Y_{t-1} + \alpha + \varepsilon_t$

One possibility is $\rho=1$, in which case the mean varies unpredictably without a trend away from the current value. The other reasonable possibility is that $\rho<1$, in which case the time series is stationary around the mean $\alpha/(1-\rho)$. We should test for $\rho=1$ with $\alpha=0$ ("zero mean") ***if*** we are sure that Y_t is mean zero (e.g., we centered it). Otherwise include an intercept in the model ("single mean").

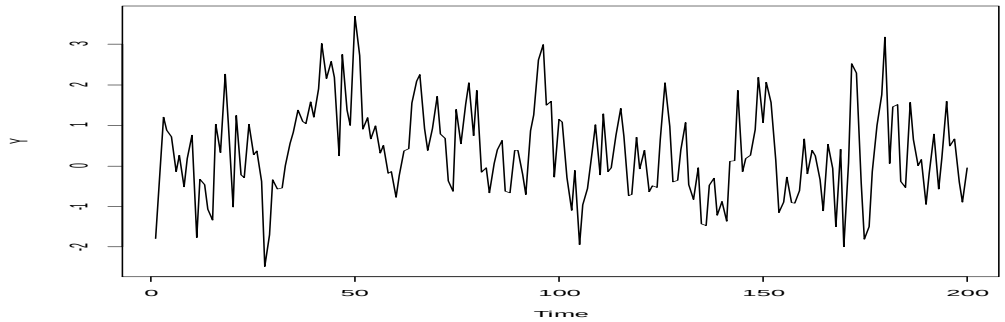
Similar to case 1, if the p-value (for H_0: $\rho=1$) is small, we conclude $\rho\neq1$ (stationary) but recognize the chance of a type-1 error (that the truth is non-stationarity). And if the p-value is large, then we cannot reject $\rho=1$ (conclude non-stationary) but recognize the chance for a type-2 error (especially with a short series) which would mean that the true situation is actually stationary. Here the type-1 error is more dangerous, because it will cause us to grossly underestimate the CI of anything beyond very short-term predictions.

Example 3:



Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-1.3664	0.4136	-0.82	0.3604		
	1	-1.3446	0.4169	-0.81	0.3649		
	2	-1.4702	0.3986	-0.84	0.3499		
Single Mean	0	-6.0759	0.3371	-1.83	0.3661	1.68	0.6439
	1	-6.2490	0.3237	-1.85	0.3570	1.71	0.6348
	2	-6.8051	0.2837	-1.87	0.3453	1.75	0.6241
Trend	0	-5.1797	0.8046	-1.52	0.8182	2.43	0.6911
	1	-5.2306	0.8007	-1.53	0.8160	2.52	0.6743
	2	-5.7241	0.7624	-1.58	0.7995	2.41	0.6959

Example 4:



Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-89.9599	<.0001	-7.67	<.0001		
	1	-80.4047	<.0001	-6.33	<.0001		
	2	-62.1628	<.0001	-5.15	<.0001		
Single Mean	0	-99.3626	0.0014	-8.19	<.0001	33.58	0.0010
	1	-94.5405	0.0014	-6.87	<.0001	23.58	0.0010
	2	-76.8595	0.0014	-5.60	<.0001	15.71	0.0010
Trend	0	-100.984	0.0001	-8.29	<.0001	34.41	0.0010
	1	-97.0699	0.0006	-6.97	<.0001	24.35	0.0010
	2	-79.4703	0.0006	-5.69	<.0001	16.18	0.0010

g. **Case 3: Trend unknown**

Main idea: If there is a trend, but a model is fit without a trend term, the only way the model can "fit the trend" is to erroneously claim a unit root. The lesser problem is that including a trend when it is not needed lowers the power of the tests.

Recommendation:

i. First examine the "Trend" section

ii. If p≤0.05, conclude no unit root (only 5% type 1 error)

   a. A simple transfer function model can test for trend vs. not.

iii. If p>0.05, conclude unit root (or no unit root, but erroneous due to low power)

   a. Now analyze ∇Y_t. A simple transfer function model can test for trend vs. not.

h. Example 1: Export data from previous class

This was a short time series, but based on an impression that it was not growing, it makes sense to read the "Single Mean" section:

Augmented Dickey-Fuller Unit Root Tests

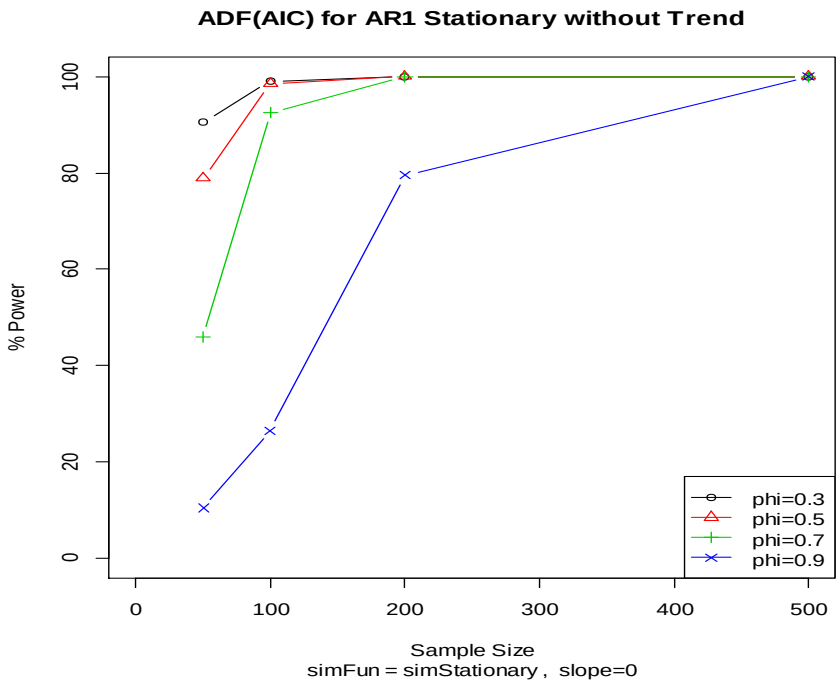
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Single Mean	0	-22.6286	0.0022	-3.82	0.0052	7.31	0.0010
	1	-30.4238	0.0003	-3.84	0.0050	7.40	0.0010
	2	-24.5872	0.0010	-3.01	0.0421	4.55	0.0667
	3	-15.1798	0.0243	-2.34	0.1657	2.79	0.3803

Now we conclude that there is probably no unit root. (If it really is non-stationary, then we would have a 5% chance of a type-1 error, i.e., the conclusion of a stationary series would be incorrect.) Another possibility is a unit root with a high order AR(4) process, but this seems unlikely both because power does diminish as the number of lags increases and because it is overly complex.

III. More on Augmented Dickey-Fuller Unit Root Testing

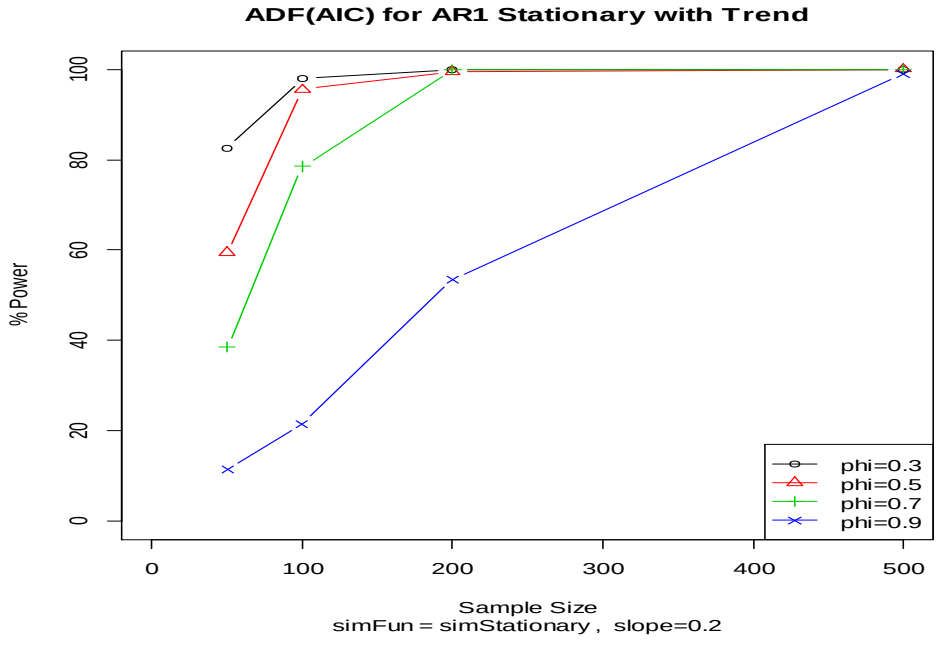
In R, package "urca" is one way to do unit root testing. The default for the (augmented) Dickey-Fuller test using `ur.df()` is to use a zero mean (`type="none"`) and 1 fixed lag which is most appropriate for finding a unit root in an AR(2) model with zero mean and no trend. Using `lags=0` gives the (non-augmented) Dickey-Fuller test which is most appropriate for detecting that the root of an AR(1) model is a unit root. The option `selectlags="AIC"` is nice because it chooses the number of lags using AIC. As in SAS, the use of a "zero mean" is *only* appropriate for when there is no evident trend *and* you have centered your data (or are otherwise assured of a zero mean). If you think there is no trend, but are unsure if the mean of the original data is zero, use a "Single Mean" (`type="drift"`). If you think there is a trend use "Trend" (`type="trend"`).

Power curve for simulated ***stationary*** AR(1) data ***without a trend*** analyzed as "Single Mean" with "AIC chosen lags":



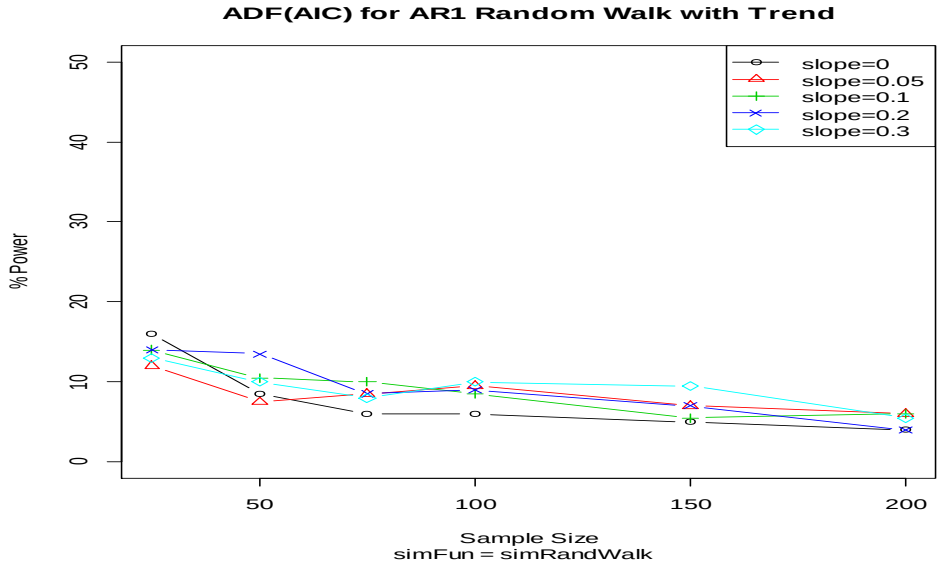
Conclude: ADF incorrectly retains "Unit root present" for short series, especially when the root is near one (e.g., phi=0.9). Power is even better with "Zero Mean" on centered data. When applied to only AR(1) data, setting `lags=0` gives more power, but you wouldn't *know* in advance that p=1 in most situations.

Power curve for simulated ***stationary*** AR(1) data ***with a trend*** (analyzed as "trend"):



Conclude: Similar power to no-trend (but be sure to use "trend" portion of ADF output). Note that the "Single Mean" portion of the output almost always incorrectly says to retain the unit root null hypothesis.

Power curve for data with a ***unit root*** AR(1) with ***no trend or with trend***, analyzed as "trend"

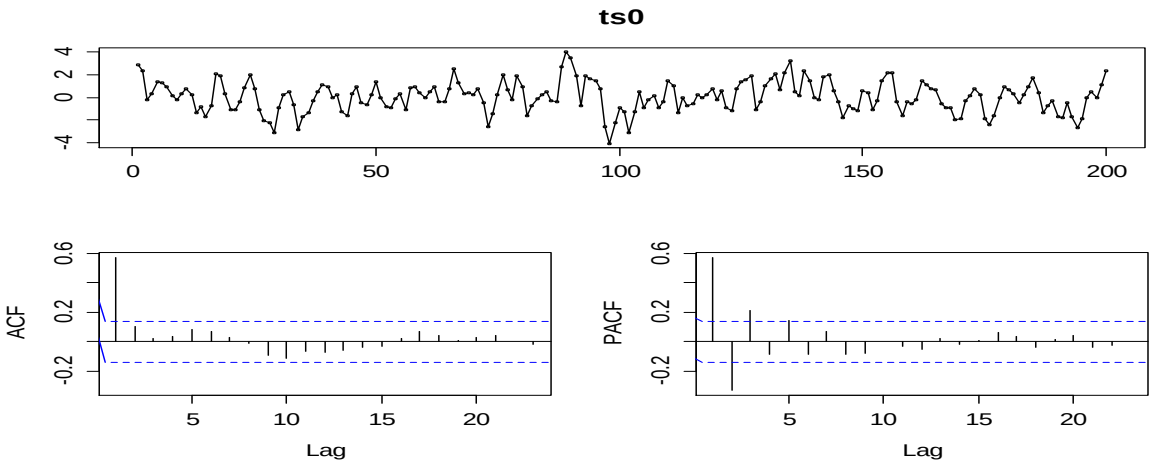


Note that slope=0 indicates no trend. We correctly retain H_0: "unit root present" all but about 5% of the time or a bit more when the series is short, even with no-trend data.

IV. Three examples

```
# Unit root tests
library(forecast) # for tsdisplay() and ndiffs()
library(tseries) # for adf.test()
library(urca) # for ur.df()

# no trend
ts0 = arima.sim(list(ar=c(0.3),ma=c(0.7)),200)
tsdisplay(ts0)
```

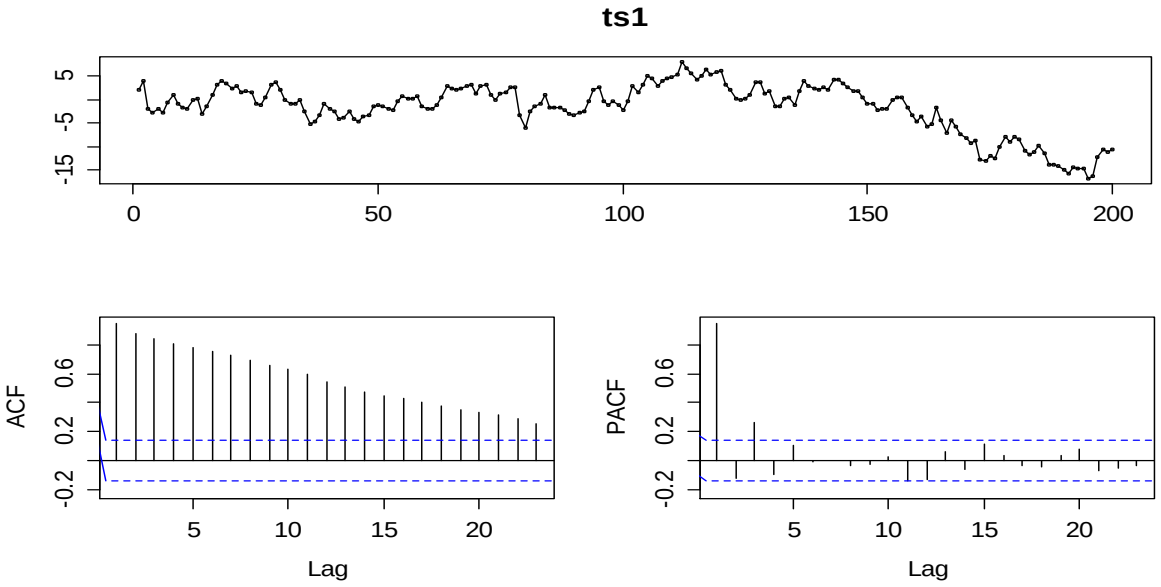


```
adf.test(ts0)  # in tseries
#  Augmented Dickey-Fuller Test
# data:  ts0
# Dickey-Fuller = -4.9846, Lag order = 5, p-value = 0.01
# alternative hypothesis: stationary
#
# Warning message: p-value smaller than printed p-value

summary(ur.df(ts0, type="drift", selectlags="AIC")) # in urca
# ###############################################
# # Augmented Dickey-Fuller Test Unit Root Test #
# ###############################################
# Test regression drift
# Value of test-statistic is: -9.1421 41.7893
#
# Critical values for test statistics:
#       1pct  5pct 10pct
# tau2 -3.46 -2.88 -2.57
# phi1  6.52  4.63  3.81

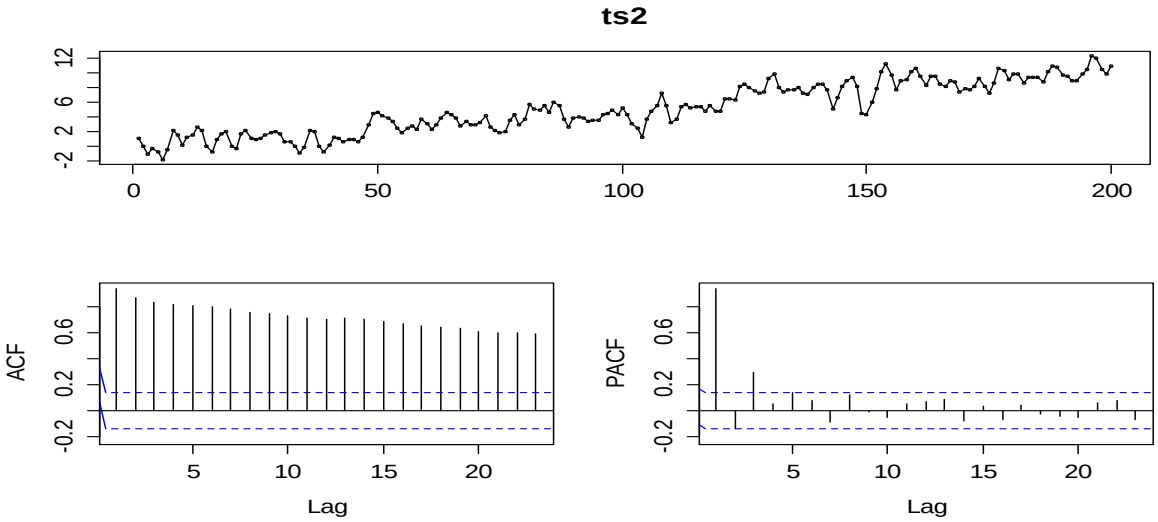
ndiffs(ts0)    # in forecast: choose "d"
# 0

# RW
ts1 = cumsum(rnorm(1:200)) +arima.sim(list(ar=c(0.3),ma=c(0.7)),200)
tsdisplay(ts1)
```




```
adf.test(ts1)  # p-value = 0.7545
summary(ur.df(ts1, type="trend", selectlags="AIC"))
# Value of test-statistic is: -2.6827 2.5345 3.6217
#
# Critical values for test statistics:
#       1pct  5pct 10pct
# tau3 -3.99 -3.43 -3.13
# phi2  6.22  4.75  4.07
# phi3  8.43  6.49  5.47
ndiffs(ts1) # 1
```

```
# ARIMA(1,0,1) with trend (tseries is misleading)
ts2 = (1:200)*10/200 + arima.sim(list(ar=c(0.3),ma=c(0.7)),200)
tsdisplay(ts2)
```




```
adf.test(ts2)  # p-value = 0.01217
summary(ur.df(ts2, type="trend", selectlags="AIC"))
# Value of test-statistic is: -8.8679 26.4181 39.3241
#
# Critical values for test statistics:
#   1pct  5pct 10pct
# tau3 -3.99 -3.43 -3.13
# phi2  6.22  4.75  4.07
# phi3  8.43  6.49  5.47
ndiffs(ts2) # 2
```

V. Package forecast's auto.arima()

auto.arima(ts0) # simulated as ARIMA(1,0,1)

```
# ARIMA(2,0,1) with zero mean
# Coefficients:       ar1     ar2     ma1
#                  0.1913  0.0105  0.7121
#            s.e.  0.1124  0.1006  0.0871
# sigma^2 estimated as 0.9807:  log likelihood=-282.34

auto.arima(ts1) # simulated as ARIMA(1,1,1)
# ARIMA(2,1,2)
# Coefficients:     ar1     ar2      ma1      ma2
#                0.1621  0.0691  -0.0311  -0.4690
#          s.e.  0.1635  0.1498   0.1444   0.1293
# sigma^2 estimated as 2.273:  log likelihood=-364.24

auto.arima(ts2) # simulated as ARMA(1,0,1) plus slope=0.05.
# ARIMA(2,1,2) with drift
# Coefficients:     ar1      ar2      ma1      ma2   drift
#                0.4947  -0.1929  -0.4611  -0.4972  0.0581
#          s.e.  0.1206   0.1004   0.1091   0.0964  0.0047
# sigma^2 estimated as 0.7505:  log likelihood=-252.53
# Warning message:
# Unable to fit final model using maximum likelihood. AIC value approximated
```

VI. Forecasting details (taken directly from Hyndman, chapter 8.8)

a. Point forecasts can be calculated using the following three steps.
   i. Expand the ARIMA equation so that y_t is on the left hand side and all other terms are on the right.
   ii. Rewrite the equation by replacing t by T+h.
   iii. On the right hand side of the equation, replace future observations by their forecasts, future errors by zero, and past errors by the corresponding residuals.

b. The calculation of ARIMA forecast intervals is much more difficult.

c. The first forecast interval is easily calculated. If $\hat{\sigma}$ is the standard deviation of the residuals, then a 95% forecast interval is given by $\hat{y}_{T+1|T} \pm 1.96\,\hat{\sigma}$. This result is true for all ARIMA models regardless of their parameters and orders.

d. Multi-step forecast intervals for ARIMA(0,0,q) models are relatively easy to calculate.

e. In Section 8.4, we showed that an AR(1) model can be written as an MA(∞) model. Using this equivalence, the above result for MA(q) models can also be used to obtain forecast intervals for AR(1) models.

f. More general results, and other special cases of multi-step forecast intervals for an ARIMA(p,d,q) model, are given in more advanced textbooks such as Brockwell and Davis.