- Recall that any time series that is a stationary process has a constant mean function.
- So a process that has a mean function that *varies over time* must be nonstationary.
- For example, we have seen that $\{Y_t\}$ is nonstationary if

$$Y_t = \mu_t + X_t,$$

  where μ_t is a nonconstant mean function and X_t is a stationary time series with mean zero.
- Sometimes μ_t represents some deterministic trend.
- In other cases, time series data could exhibit nonstationarity, but there is no particular trend model that we believe holds (see R example for oil data).

AR with $|\phi| > 1$

- Consider a model of the form $Y_t = \phi Y_{t-1} + e_t$.
- When $|\phi| > 1$, we get an "explosive" (exponential growth) model in which the weights on past disturbance terms blow up (rather than dying out) as we go further into the past.
- See the R plot for a simulated example of such a series.
- In such series, $var(Y_t)$ tends to blow up as time increases, and for large t, $corr(Y_t, Y_{t-k}) \approx 1$.

Stationarity through Differencing

- If $\phi = 1$, then we get $Y_t = Y_{t-1} + e_t$, a nonstationary model which we can rewrite through *differencing* as $\nabla Y_t = e_t$, where $\nabla Y_t = Y_t - Y_{t-1}$.
- We have seen before that differencing (or the related approach of *detrending*) can convert nonstationary series into processes that can be modeled as stationary.
- In other situations, the second-difference model, in which we focus on $\nabla^2 Y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$, is stationary.
- This leads us to a general type of model in which the d-th difference is stationary.

The ARIMA Model

- A time series $\{Y_t\}$ is an autoregressive integrated moving average model if the d-th difference, denoted

$$W_t = \nabla^d Y_t$$

is a stationary ARMA model.
- Specifically, if $\{W_t\}$ is $ARMA(p, q)$, then $\{Y_t\}$ is $ARIMA(p, d, q)$.
- Often we consider $d = 1$ (first differences) or $d = 2$ (second differences).
- Consider the $ARIMA(p, 1, q)$ model, letting $W_t = Y_t - Y_{t-1}$:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

More on the $ARIMA(p, 1, q)$ Model

- The characteristic equation of this model can be shown to have one solution that is exactly 1 (hence the ARIMA model is nonstationary).
- The remaining solutions are the solutions of the characteristic equation of the stationary process ∇Y_t.
- For the $ARIMA(p, 1, q)$ model, we can write Y_t as

$$Y_t = \sum_{j=-m}^{t} W_j$$

  where $t = -m$ is some time earlier in the process than $t = 1$, when we first observed the time series.
- For the $ARIMA(p, 2, q)$ model, we can write Y_t as

$$Y_t = \sum_{j=0}^{t+m} (j + 1) W_{t-j}$$

Special Cases of ARIMA Models

- If the ARIMA process has no autoregressive terms, it becomes an *integrated moving average* process, denoted $IMA(d, q)$.
- If the ARIMA process has no moving average terms, it becomes an *autoregressive integrated* process, denoted $ARI(p, d)$.
- The simplest IMA process is the $IMA(1, 1)$ process:

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1}$$

- Since $W_t = Y_t - Y_{t-1} = e_t - \theta e_{t-1}$ here, we have, using the summation formula on the previous slide:

$$Y_t = e_t + (1-\theta)e_{t-1} + (1-\theta)e_{t-2} + \cdots + (1-\theta)e_{-m} - \theta e_{-m-1}$$

Properties of the $IMA(1,1)$ Process

- Here, the weights on the e_t's do not die out as we go back in time.
- Y_t is approximately a bunch of equally weighted white noise terms, plus a couple of white noise terms with different weights.
- The sizes of these weights depend on θ.
- It can be shown that $var(Y_t) = [1 + \theta^2 + (1-\theta)^2(t+m)]\sigma_e^2$.

$$corr(Y_t, Y_{t-k}) = \frac{1 - \theta - \theta^2 + (1-\theta)^2(t+m-k)}{[var(Y_t)var(Y_{t-k})]^{1/2}}$$

  which is near 1 for large m and small-to-moderate k.
- These imply that (1) as time goes on, $var(Y_t)$ gets larger and larger.
- And (2), the correlation between values of the process will be strongly positive for small lags ($k = 1, 2, \ldots$) and even moderately sized lags.

The $IMA(2,2)$ Process

- In the $IMA(2,2)$ process,

$$\nabla^2 Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

- Again, if we express Y_t as a linear combination of white noise terms, the weights on the e_t's do not die out as we go back in time.
- Again, $var(Y_t)$ gets larger as t increases.
- And again, the correlation between values of the process will be strongly positive for small lags ($k = 1, 2, \ldots$) and even moderately sized lags.
- See R plots for examples of graphs of simulated processes.

Constant Terms in ARIMA Models

- In the $ARIMA(p, d, q)$ process, $\nabla^d Y_t = W_t$ is a stationary $ARMA(p, q)$ process, which we assume to have mean zero.
- We can alter this, if necessary, to allow W_t to have a nonzero mean μ.
- One approach is to replace W_t everywhere with $W_t - \mu$:

$$W_t - \mu = \phi_1(W_{t-1} - \mu) + \phi_2(W_{t-2} - \mu) + \cdots + \phi_p(W_{t-p} - \mu) + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

More on Constant Terms in ARIMA Models

- Another approach is to add a constant term θ_0 into the model equation:

$$W_t = \theta_0 + \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

- If $E(W_t) = \mu$ for all t, then taking expected values of both sides of the above equation:

$$\mu = \theta_0 + (\phi_1 + \phi_2 + \cdots + \phi_p)\mu.$$

- Clearly, we can write μ in terms of θ_0, or θ_0 in terms of μ, so either approach is equivalent.

Still More on Constant Terms in ARIMA Models

- As an example, consider the $IMA(1,1)$ process with a constant term:
- We could express this as $Y_t = Y_{t-1} + \theta_0 + e_t - \theta e_{t-1}$, or as $W_t = \theta_0 + e_t - \theta e_{t-1}$.
- Then as a linear combination of white noise terms:

$$Y_t = e_t + (1-\theta)e_{t-1} + (1-\theta)e_{t-2} + \cdots + (1-\theta)e_{-m} - \theta e_{-m-1} + (t-m-1)\theta_0$$

- This has added a deterministic linear time trend (with slope θ_0) to the process.
- So over time the trend of the process would be expected to increase (or decrease) approximately linearly, depending on the sign of θ_0.
- To represent some general polynomial trend (not necessarily linear), we could consider $Y_t = Y_t' + \mu_t$, where μ_t is some polynomial in t and Y_t' is $ARIMA(p,d,q)$ with $E(Y_t') = 0$.

Other Transformations

- Differencing is not the only transformation that can be used to achieve stationarity.
- In many real time series, the variability of Y_t appears larger for later values of t.
- Suppose $Y_t > 0$ for all t, $E(Y_t) = \mu_t$, and $\sqrt{var(Y_t)} = \mu_t \sigma$.
- Then taking a Taylor series approximation of $\log(Y_t)$ and taking expected value and variance of that,

$$E[\log(Y_t)] \approx \log(\mu_t) \text{ and } var[\log(Y_t)] \approx \sigma^2.$$

- So if the standard deviation of the series is increasing proportionally with the mean of the series, then taking (natural) logarithms of the series values will yield a process with constant variance.
- Also, if Y_t is changing exponentially, then the logged series will change linearly.
- So the series of the first differences of the logged data should look stationary.

Percentage Changes and Logarithms

- This provides a natural form of transformation to use when the time series Y_t shows that the percentage change from one time period to the next is stable.
- In that case, taking the natural log and then taking first differences should produce a series $\nabla[\log(Y_t)]$ that is approximately stationary.
- See the examples of the electricity data (R plots of untransformed and transformed data), as well as the oil price data we examined previously.

Box-Cox Power Transformations

- A flexible family of transformations was given by Box and Cox:

$$g(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} \text{ for } \lambda \neq 0 \\ \log(x) \text{ for } \lambda = 0 \end{cases}$$

- A variety of different values of λ could be tried on a data set, and the "best" choice used.
- Note that $\lambda = 1/2$ corresponds to a square root transformation.
- $\lambda = -1$ corresponds to a reciprocal transformation.
- The Box-Cox transformation assumes the data values are all positive. If not, some constant could initially be added to all data values to make them all positive.
- A grid of λ values can easily be tried in R, and the λ that maximizes a normal log-likelihood criterion could be selected.
- See R example with the electricity data.