Statistical Modeling in Finance

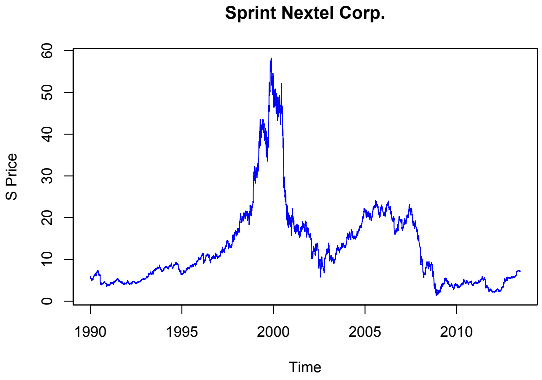
Final Report

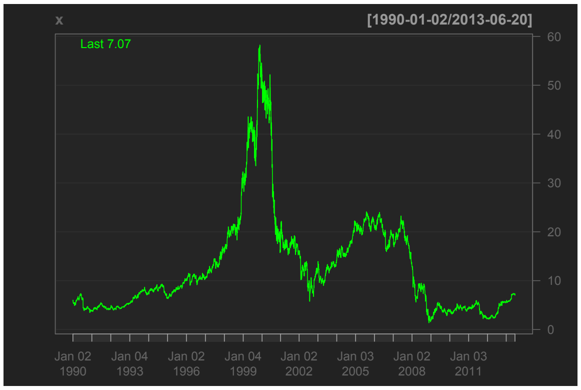
Paris 1 Summer School 2013

CHEN, Guo

ameliachennj@gmail.com

1.(a) We download the daily price of stock X, which is index ”S”. The plots are attached.



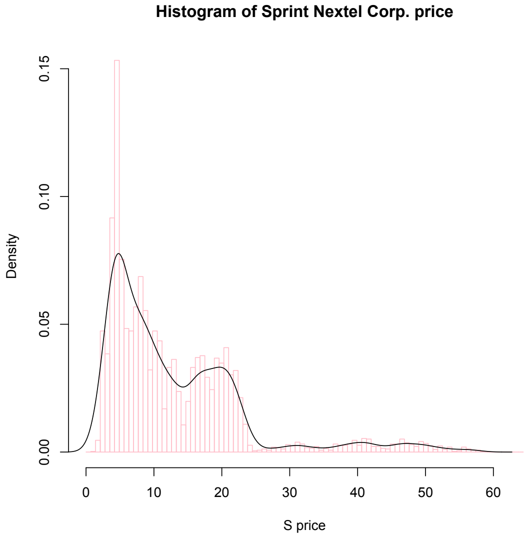


(b). Now we compute the basic statistics of X and have:

sample mean	standard deviation	skewness	ex kurtosis	min	max
12.8908	10.32845	1.818344	3.709197	1.37	58.24

The student t-test shows p value is less than 0.05. Hence we reject the null hypothesis that the mean is zero. So the conclusion is that the true mean of X is not zero at 5% significance level.

(c). The plot of a histogram with a density function for X is :



Using Kolmogorov-Smirnov test (K-S test) to compare pdf of X with a normal distribution(mean=0, sd=1), we find p-value is less than 0.05. So the pdf of X is not a standard normal distribution at 5% significance level, which is also obvious from the above plot.

Using K-S test, pdf of X is not a t-distribution with degree of freedom 4 at 5% significance level either. Indeed, from the plot, we see pdf of X has several peaks, so the guess is that it's a mixture of several distributions.

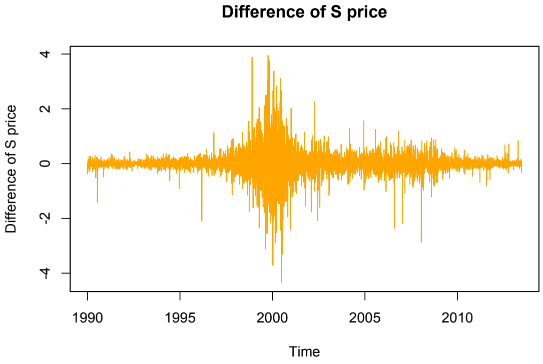
(d). At first I use adf.test to check whether X is stationary. The p-value is equal to 0.6956. The alternative hypothesis is ”stationary”. So the conclusion is that we cannot reject the null hypothesis, i.e. we conclude X is not stationary.

Then want to show X is not stationary by using moments:

Select 1st to 1000th data to form subgroup x1, select 2001st to 3000th data to form subgroup x2. By using t-test again, we find the means of these two subgroups are not equal by 5% significance level. We use F-test and find the variances of these two subgroups are not equal by 5% significance level either.

Hence we conclude X is not stationary, since mean is not constant everywhere, nor is the variance.

(e). Take the difference of the price at t and t-1, the transformed data Y is stationary. The plot is below. The adf.test shows p-value=0.01, which is less than 0.05. Hence we reject the null hypothesis that Y is not stationary. By using this test, the conclusion is Y is stationary.



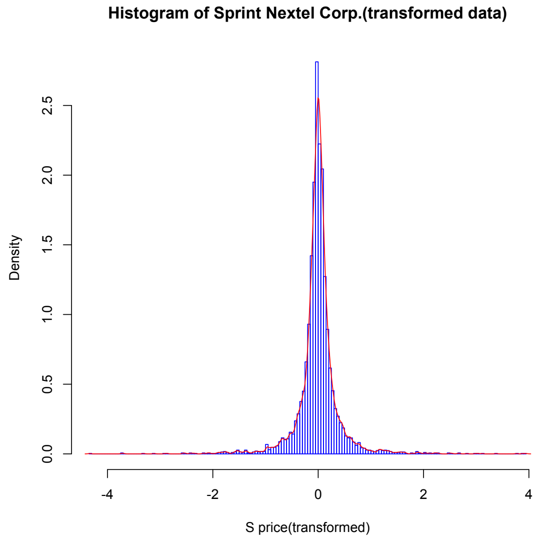
The kpss.test shows p-value = 0.1. Hence we cannot reject the null hypothesis that Y is level or trend stationary. By using this test, the conclusion is Y is stationary too.

(f). Now we compute the basic statistics of Y and have:

sample mean	standard deviation	skewness	ex kurtosis	min	max
0.0001844956	0.4174271	0.02346547	18.683004	-4.34	3.95

To see whether the mean of Y is zero, the student t-test shows p-value = 0.9729. Hence we cannot reject the null hypothesis that the mean is 0 at 5% significance level. We conclude Y has mean 0.

The plot of a histogram with a density function for Y is :

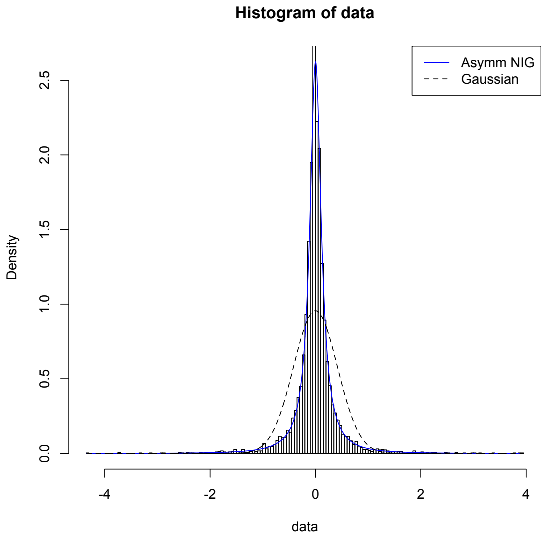


Now want to fit a NIG distribution on data Y. Using mle method we can find the parameters. The fitted NIG distribution has parameters $\bar{\alpha} = 0.094682168$, $\mu = 0.001906830$, $\Sigma = 0.427528775$, $\gamma = -0.001712437$. For details of the interpretation of these parameters, refer to:

`http://stat.ethz.ch/CRAN/web/packages/ghyp/vignettes/Generalized_Hyperbolic_Distribution.pdf`

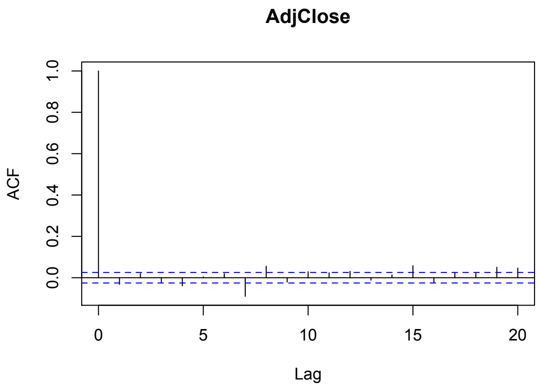
The authors designed the package that I used to find the fitted NIG distribution.

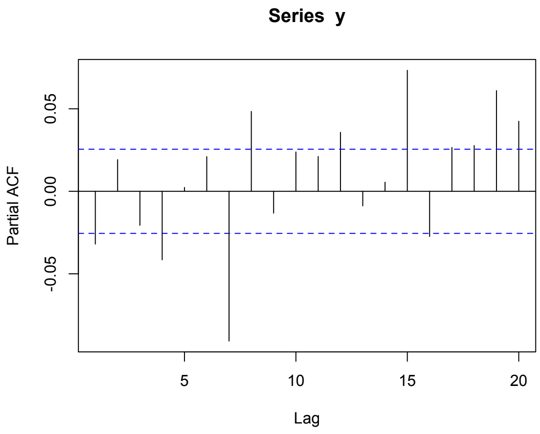
The fitted NIG distribution on the histogram of Y looks like this:



This fit looks good.

(g). The autocorrelation function(acf) and partial autocorrelation function(pacf) associated with Y are shown below:


AdjClose
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Lag


Series y
Partial ACF
-0.05
0.00
0.05
5
10
15
20
Lag

We notice PACF doesn't cut off after finite lags, which implies Y is not exactly an AR process. ACFs are not significantly small after finite lags, which implies Y is not exactly a MA process either. Since both ACFs and PACFs die down(if we choose lag=500 or bigger, we can see the pattern obviously), Y may be more likely a mixed ARMA process.

Y also doesn't satisfy normal assumption, as can be checked by `normalTest`.

(h). Using Ljung-Box test, we find there is serial correlation in Y.

(i). AR(1) model:

w(t)=-0.032 w(t-1)+e(t), where w(t)=y(t)-0.0002. ϕ has standard error 0.013.
Now test

$$H_0 : \phi = 0 \quad \text{vs} \quad H_1 : \phi \neq 0$$

Under normal assumption of e(t), 95% confidence interval for ϕ is from -0.032-1.96*0.013=-0.05748 to -0.032+1.96*0.013=-0.00652. Since 0 is not inside, we reject the null hypothesis. The coefficient ϕ is significant.

Also we find the residuals don't satisfy the normal assumption, have mean zero(by t-test) and the variance of the residuals is less than the variance of Y.

MA(1) model:

y(t)=0.0002+e(t)-0.0308e(t-1). θ has standard error 0.0128.
Now test:

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \neq 0$$

Under normal assumption of e(t), 95% confidence interval for θ is from -0.0308-1.96*0.0128=-0.055888 to -0.0308+1.96*0.0128= -0.005712. Since 0 is not inside, we reject the null hypothesis. The coefficient θ is significant.

Also we find the residuals don't satisfy the normal assumption, have mean zero(by t-test) and the variance of the residuasl is less than the variance of Y.

ARMA(1) model:

w(t)= -0.8558 w(t-1) + 0.8307 e(t-1) +e(t), where w(t)=y(t)-0.0002. ϕ has standard error 0.0426, θ has standard error 0.0454.
Now test:

$$H_0 : \phi = 0 \quad \text{vs} \quad H_1 : \phi \neq 0$$

Under normal assumption of e(t), 95% confidence interval for ϕ is from -0.8558-1.96*0.0426= -0.939296 to -0.8558+1.96*0.0426=-0.772304. Since 0 is not inside, we reject the null hypothesis. The coefficient ϕ is significant.
Now test:

$$H_0 : \theta = 0 \quad \text{vs} \quad H_1 : \theta \neq 0$$

Under normal assumption of e(t), 95% confidence interval for θ is from -0.8307-1.96*0.0454=-0.919684 to -0.8307+1.96*0.0454= -0.741716. Since 0 is not inside, we reject the null hypothesis. The coefficient θ is significant.

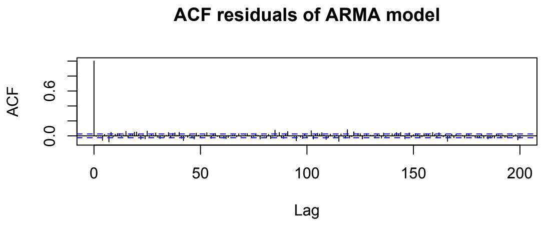
Also we find the residuals don't satisfy the normal assumption, have mean zero(by t-test) and the variance of the residuals is less than the variance of Y.

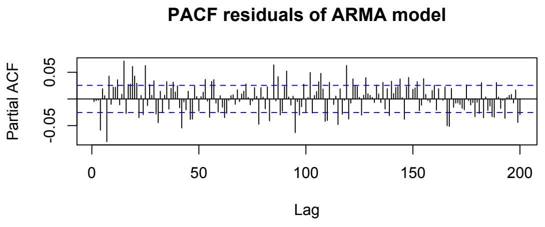
(j). The discussion in (g) implies ARMA model is a better model than AR model or MA model.

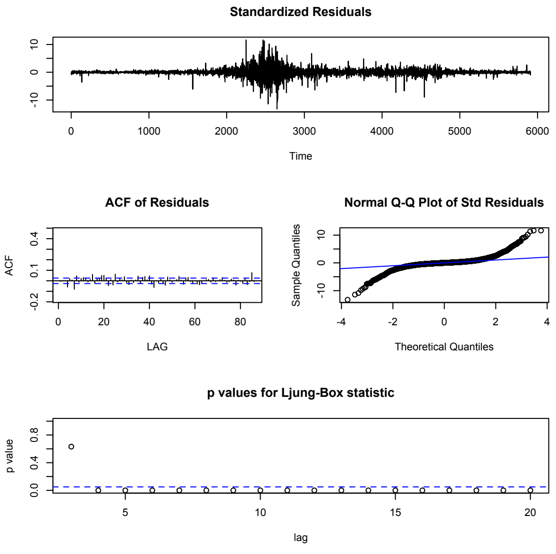
Also note ARMA(1,1) has the smallest AIC and AICc, the residuals from ARMA(1,1) have the smallest variance, hence this model is the most fitted of the three.

(k)&(l). From now on we focus on the ARMA(1,1) model.

By checking acf and pacf plots, we find the residuals are not exactly white noise, since ACFs(except at 0) are not all zeros (not inside the two dotted lines). Also the residuals don't satisfy the normal assumption.


ACF residuals of ARMA model
ACF
0.0
0.6
0
50
100
150
200
Lag


PACF residuals of ARMA model
Partial ACF
-0.05
0.05
0
50
100
150
200
Lag


Standardized Residuals
10
0
-10
0
1000
2000
3000
4000
5000
6000
Time
ACF of Residuals
ACF
0.4
0.1
-0.2
0
20
40
60
80
LAG
Normal Q-Q Plot of Std Residuals
Sample Quantiles
10
0
-10
-4
-2
0
2
4
Theoretical Quantiles
p values for Ljung-Box statistic
p value
0.0
0.4
0.8
5
10
15
20
lag

Now we want to detect whether there's an ARCH effect. If $\{e(t)\}_t \sim ARCH(1)$, then $\{e(t)^2\}_t \sim AR(1)$. Consider $e(t)^2$. Want to fit $e(t)^2$ an AR(1) model. The fitted model is

$$w(t) = 0.3237w(t-1) + u(t), \quad \text{where} \quad w(t) = e(t)^2 - 0.1738.$$

ϕ has standard error 0.0123. Want to test:

$$H_0 : \phi = 0 \quad \text{vs} \quad H_1 : \phi \neq 0$$

Under normal assumption of e(t), 95% confidence interval for ϕ is from 0.3237-1.96*0.0123=0.299592 to 0.3237+1.96*0.0123=0.347808. Since 0 is not inside, we reject the null hypothesis. The coefficient ϕ is significant.

Hence there is an ARCH effect since ϕ is nonzero.

Now examine the error term u(t) in this model. From acf and pacf plots, we find {u(t)} are not white noises either. {u(t)} don't satisfy the normal assumption, have mean zero (by t-test) and the variance is less than the variance of $\{e(t)^2\}$.

(m). The fitted model is

$$p(t+1) = 0.1442 \cdot p(t) + 0.8558 \cdot p(t-1) + 0.8307 \cdot e(t-1) + e(t) + 0.000171,$$

p(t) is the price, but the error term $e(t)$ is the difference between the price difference y(t) and the fitted value of y(t).

So in order to predict price p(t) more easily, we want to predict the price difference h(t) first. We have the model:

$$h(t) - 0.0002 = -0.8558(h(t-1) - 0.0002) + 0.8307e(t-1) + e(t).$$

We aim to predict the last 21 values of price. In the loop, we use the true value y(t-1) for h(t-1) in the above model and let e(t-1)=y(t-1)-h(t-1), then predict h(t). Then we can predict price from p(t+1)=h(t)+p(t). The last real 21 values of price are:

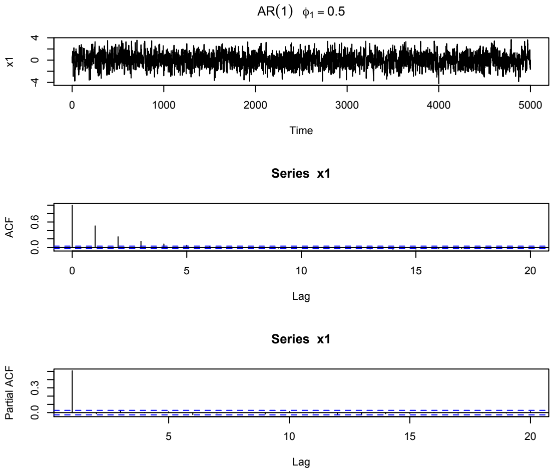
```
[1] 7.30 7.31 7.33 7.27 7.28 7.34 7.30 7.22 7.26 7.20 7.34
[12] 7.24 7.18 7.35 7.35 7.32 7.32 7.22 7.32 7.00 7.07
```

The predicted values are:

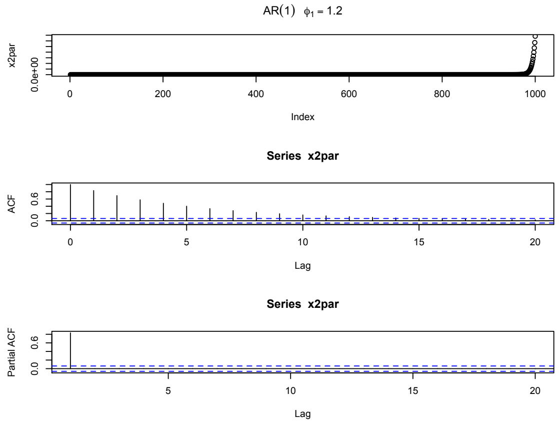
```
[1] 7.300000 7.310000 7.307496 7.309445 7.309704 7.309609
[7] 7.308553 7.310805 7.311313 7.310259 7.313012 7.307582
[13] 7.314974 7.310710 7.310356 7.311022 7.311593 7.311490
[19] 7.314457 7.309853 7.322080
```

The predictions are not too far away from reality.

2.(a) First we simulate an AR(1) process with $\phi = 0.5$, the acf and pacf functions look like below: In these plots, we see ACF decays exponentially as the lag increases. PACF cuts off at lag 1. So this is clearly an AR(1) progress.

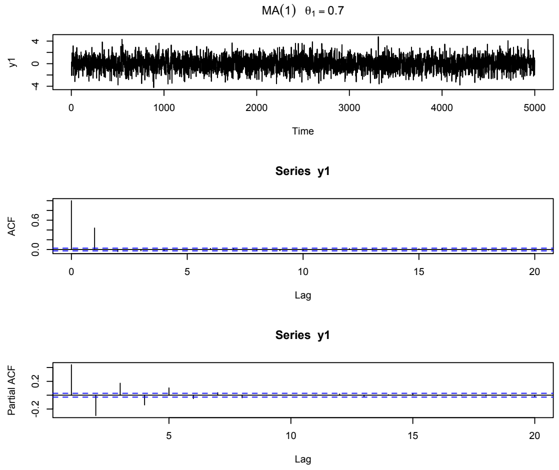


Second we simulate an AR(1) process with $\phi = 1.2$. I use a random variable from a normal distribution with mean 0, variance 0.01 as the error term. Because $\phi > 1$, the data explode (become too large). I choose the first 1000 data to plot acf and pacf functions, which look as below:

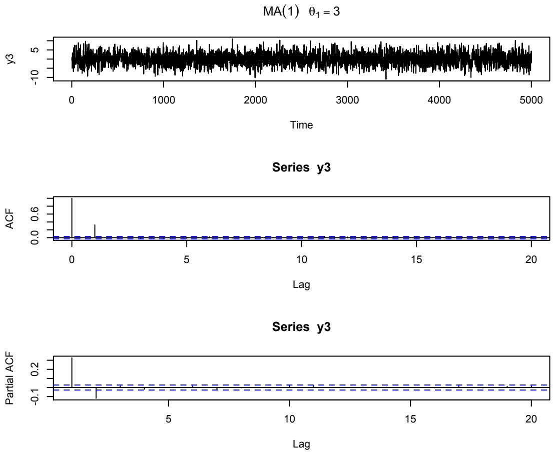


From the plots, for similar reasons, this is also observed as an AR(1) process. But the plot of the data is very different from the first one, since when $\phi > 1$ the data explode (becomes too big when n is large). Note the necessary and sufficient condition of stationarity is $|\phi| < 1$.

(b). First we simulate an MA(1) process with $\phi = 0.7$, the acf and pacf functions look like below: The plots show that ACF cuts off at lag 1 and PACF decays, so this is a MA(1) process. Second we simulate an MA(1) process with $\phi = 3$, the acf and pacf functions look like below:

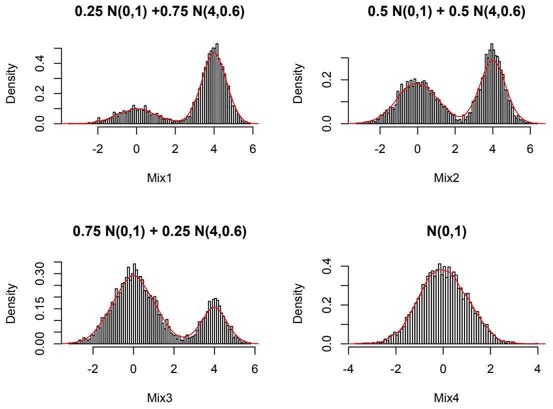


Same with the above observations, the plots show that ACF cuts off at lag 1 and PACF decays, so this is also a MA(1) process.



Note MA(1) process is always stationary. The condition of invertibility is $|\theta| < 1$.

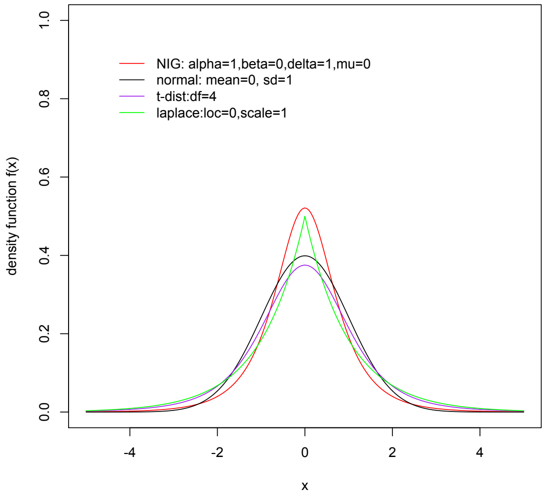
(c). We choose ω=0.25, 0.5, 0.75 and 1. The graphs are:



Each plot shows the pdf of a mixture of two random variables $\omega X_1 + (1-\omega)X_2$ with different ω. Here $X_1 \sim N(0,1)$ and $X_2 \sim N(4,0.6)$.

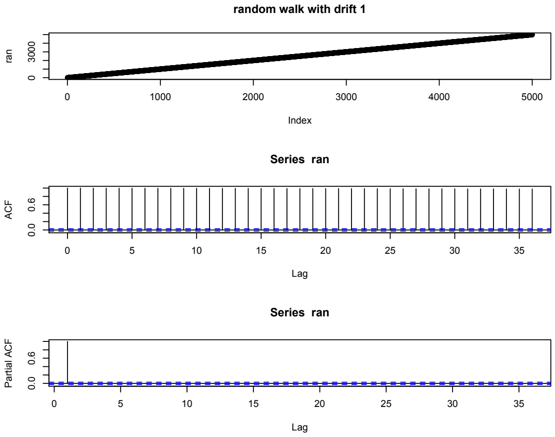
If $\omega \neq 0$, there are two peaks at 0 and 4, the centers of pdfs of X_1 and X_2 respectively. As ω becomes bigger, the part played by X_1 in the mixed distribution is more significant; if $\omega = 1$, the distribution is completely the pdf of X_1.

(d)&(e). First generate three distributions under the GH family: NIG distribution with parameters $\alpha = 1$, $\beta = 0$, $\delta = 1$ and $\mu = 0$; t-distribution with degree of freedoms 4 and Laplace distribution with loc=0 and scale=1. Then want to compare these distributions with standard normal distribution.



The observations are: Laplace distribution has a fatter tail than NIG distribution. NIG distribution has a fatter tail than t-distribution. And t-distribution has a fatter tail than standard normal distribution. This makes sense by looking at kurtosis: if we compute their respective kurtosis, in order of highest kurtosis we find: Laplace> NIG> t-dist>normal (distributions in in this example). Higher kurtosis implies fatter tail and that the distribution is more peaked, which is shown exactly in this plot.

(f). Define a vector ran, let the error term e be a random variable from a normal distribution with mean 0 and variance 0.01. Generate the random walk with drift 1 by iterating ran[i+1]=ran[i]+e[i+1]+1.
The plots look like below:



Note ran[n]=ran[0]+ n $+\sum_{i=1}^{n}$ e[i], i.e. ran is a trend with slope 1, which is also shown clearly in the plot of ran. The random walk without drift is an AR(1) process with $\phi = 1$, hence can be characterized by large nonvanishing spikes (close to the value 1) in acf plot. The random walk with drift is characterized by the same ACFs with that of the random walk without drift, so the sample ACF for ran approaches 1 for any finite lag.