S681 Financial Statistics
Homework 1
Feb 8 2012
Cheng-Han Yu

Textbook: Ruppert, D. (2011) *Statistics and Data Analysis for Financial Engineering*, Springer, New York.

1. **Chapter 2 Problem 2** (1) Compute the log returns for GM and (2) plot the returns versus the log returns. (3) How highly correlated are the two types of returns? (The R function *cor* computes correlations.)

   Solution:
   (1) The R code is

```
GMlogReturn = log(GM_AC[2 : n]/GM_AC[1 : (n - 1)])
```

   (2)

```
plot(GMReturn, GMlogReturn)
```

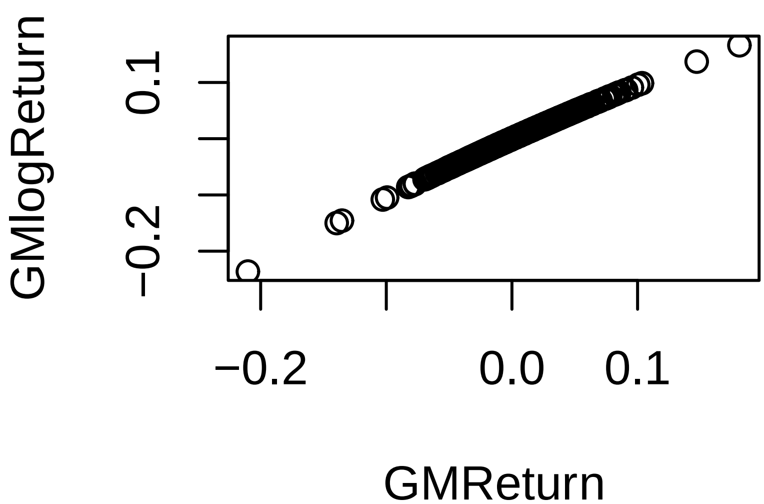


Figure 1: GM return versus GM log return

   (3)

```
cor(GMReturn, GMlogReturn)
0.9995408
```

Since the log function is an increasing function, we have a nearly perfect linear relationship between GMReturn and GMlogReturn.

2. **Chapter 2 Problem 4** What is the probability that the hedge fund will make a profit of at least 100,000?

   Solution:

   The R code is

```
niter = 1000
below = rep(0,niter)
above = rep(0,niter)
logPrice = rep(0,niter)
set.seed(2009)
for (i in 1:niter)
{
r = rnorm(100,mean=.05/253, sd=.23/sqrt(253))
for (j in 1:100)
{
logPrice[i] = log(1000000) + cumsum(r)[j]
if (logPrice[i] > log(1100000))
{
above[i]=1
below[i]=0
break
}else if (logPrice[i] < log(950000))
{
above[i]=0
below[i]=1
break
}
else
{
above[i]=0
below[i]=0
}
}
}
mean(above)
mean(below)
profit <- exp(mean(logPrice))-1000000
```

   After 100000 iterations by R, the simulated probability that the hedge fund will make a profit of at least $100000 is 0.3888. (Sorry I don't know how to indent the braces in R code to the right by Latex.)

3. **Chapter 2 Problem 6** What is the expected profit from this trading strategy?

   Solution:

   By the above code, the simulated expected profit is 5979.124.

4. **Chapter 2 Exercise 5** Let r_t be a log return. Suppose that $r_1, r_2, ...$ are $i.i.d.N(0.06, 0, 47)$.

   (a) What is the distribution of $r_t(4) = r_t + r_{t-1} + r_{t-2} + r_{t-3}$?
   Solution:

   Since all r_i are iid normal random variables with mean 0.06 and variance 0.47, by the additivity of normality, $r_t(4) = r_t + r_{t-1} + r_{t-2} + r_{t-3}$ also follows normal distribution and its mean equals to $E[r_t(4)] = E[r_t + r_{t-1} + r_{t-2} + r_{t-3}] = E[r_t + r_t + r_t + r_t] = 0.06 + 0.06 + 0.06 + 0.06 = 0.24$, and the variance is $Var[r_t(4)] = Var[r_t + r_{t-1} + r_{t-2} + r_{t-3}] = Var(r_t) + Var(r_{t-1}) + Var(r_{t-2}) + Var(r_{t-3}) = 0.47 + 0.47 + 0.47 + 0.47 = 1.88$. That is, $r_t(4) \sim N(0.24, 1.88)$.

   (b) What is $P\{r_1(4) < 2\}$?
   Solution:

$$
\begin{aligned}
P\left(r_1(4) < 2\right) &= P(r_t(4) < 2) \\
&= P\left(\frac{r_t(4) - 0.24}{\sqrt{1.88}} < \frac{2 - 0.24}{\sqrt{1.88}}\right) \\
&= P\left(Z < 1.283612\right) \\
&= 0.9003611
\end{aligned}
$$

   where $Z \sim N(0, 1)$.

   (c) What is the covariance between $r_1(2)$ and $r_2(2)$?
   Solution:

$$
\begin{aligned}
Cov(r_1(2), r_2(2)) &= Cov(r_1 + r_0, r_2 + r_1) \\
&= Cov(r_1, r_2) + Cov(r_1, r_1) + Cov(r_0, r_2) + Cov(r_0, r_1) \\
&= 0 + Var(r_1) + 0 + 0 \\
&= 0.47.
\end{aligned}
$$

   (d) What is the conditional distribution of $r_t(3)$ given $r_t(2) = 0.6$?
   Solution:

   Given $r_{t-2} = 0.6$, $r_t(3) = r_t + r_{t-1} + r_{t-2} = r_t + r_{t-1} + 0.6$. Hence, $r_t(3)$ is also a normal random variable with mean $E[r_t(3)] = E[r_t + r_{t-1} + 0.6] = 0.06 + 0.06 + 0.6 = 0.72.$, and variance $Var[r_t(3)] = Var[r_t + r_{t-1} + 0.6] = Var[r_t] + Var[r_{t-1}] = 0.47 + 0.47 = 0.94$. That is, $r_t(3)|_{r_{t-2}=0.6} \sim N(0.72, 0.94)$.

5. **Chapter 2 Exercise 6** Suppose that $X_1, X_2, ...$ is a lognormal geometric random walk with parameters (μ, σ^2). More specifically, suppose that $X_k = X_0 \exp(r1 + ... + r_k)$, where X_0 is a fixed constant and $r_1, r_2, ...$ are i.i.d. $N(\mu, \sigma^2)$.

   (a) Find $P(X_2 > 1.3X_0)$?
   Solution:

   $X_2 = X_0 \exp(r_1 + r_2)$. Since X_2 is a lognormal geometric random walk with parameters μ and σ^2, $\log \frac{X_2}{X_0} = r_1 + r_2$ is normal distribution with mean 2μ and variance $2\sigma^2$, i.e. $\log \frac{X_2}{X_0} \sim N(2\mu, 2\sigma^2)$. Then

$$\begin{aligned} P(X_2 > 1.3X_0) &= P\left(\frac{X_2}{X_0} > 1.3\right) \\ &= P\left(\log \frac{X_2}{X_0} > \log 1.3\right) \\ &= P\left(\frac{\log \frac{X_2}{X_0} - 2\mu}{\sqrt{2}\sigma} > \frac{\log 1.3 - 2\mu}{\sqrt{2}\sigma}\right) \\ &= P\left(Z > \frac{\log 1.3 - 2\mu}{\sqrt{2}\sigma}\right) \\ &= 1 - P\left(Z < \frac{\log 1.3 - 2\mu}{\sqrt{2}\sigma}\right) \\ &= 1 - \Phi\left(\frac{\log 1.3 - 2\mu}{\sqrt{2}\sigma}\right) \end{aligned}$$

   where $\Phi(z)$ is the cdf of a standard normal distribution.

   (b) Use (A.4) to find the density of X_1.
   Solution:

   (A.4) says $f_Y(y) = f_X(h(y))\,|h'(y)|$. We know that $X_1 = X_0 \exp r_1$. Then $r_1 = \log \frac{X_1}{X_0}$. Hence, by (A.4), we conclude that

$$\begin{aligned} f_{X_1}(x_1) &= f_{r_1}(h(x_1))\,|h'(x_1)| \quad r_1 = h(x_1) = \log \frac{x_1}{x_0} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left[\log \frac{x_1}{x_0} - \mu\right]^2}{2\sigma^2}\right) |1/x_1| \\ &= \frac{1}{\sqrt{2\pi}\sigma|x_1|} \exp\left(\frac{-\left[\log x_1 - (\log x_0 + \mu)\right]^2}{2\sigma^2}\right), \end{aligned}$$

   which is the probability density of X_1.

(c) Find a formula for the 0.9 quantile of X_k for all k.
Solution:

We know that $\log X_k = \log X_0 + r_1 + ... + r_k \sim N(\log X_0 + k\mu, k\sigma^2)$. To find the 0.9 quantile of X_k, $q_{0.9}$, we form the following probability:

$$
\begin{aligned}
P(X_k < q_{0.9}) &= 0.9 \\
\iff P(\log X_k < \log q_{0.9}) &= 0.9 \\
\iff P\left(Z < \frac{\log q_{0.9} - (\log X_0 + k\mu)}{\sqrt{k}\sigma}\right) &= 0.9
\end{aligned}
$$

Hence, $\frac{\log q_{0.9} - (\log X_0 + k\mu)}{\sqrt{k}\sigma} = 1.281552$. Thus, the formula for the 0.9 quantile of X_k for all k is

$$q_{0.9} = X_0 \exp(k\mu + 1.281552\sqrt{k}\sigma)$$

(d) What is the expected value of X_k^2 for any k? (Find a formula giving the expected value as a function of k.)
Solution:

We know that $\log X_k \sim N(\log X_0 + k\mu, k\sigma^2)$. The density of X_k is

$$\frac{1}{\sqrt{2\pi k}\sigma x_k} \exp\left(\frac{-\left[\log x_k - (\log x_0 + k\mu)\right]^2}{2k\sigma^2}\right)$$

The the second moment of X_k is

$$\int_0^\infty X_k^2 \frac{1}{\sqrt{2\pi k}\sigma x_k} \exp\left(\frac{-\left[\log x_k - (\log x_0 + k\mu)\right]^2}{2k\sigma^2}\right) dX_k$$

Substituting $X_k = X_0 \exp(r_k + \cdots + r_1) \equiv X_0 \exp(r_k(k))$, we have

$$E(X_k^2) = \int_{-\infty}^\infty [X_0 \exp(r_k(k))]^2 \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[\log x_0 + r_k(k) - (\log x_0 + k\mu)\right]^2}{2k\sigma^2}\right) dr_k(k)$$

$$= \int_{-\infty}^\infty X_0^2 \exp(r_k(k))^2 \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[r_k(k) - k\mu\right]^2}{2k\sigma^2}\right) dr_k(k)$$

$$= X_0^2 \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[r_k(k) - k\mu - 2k\sigma^2\right]^2}{2k\sigma^2}\right) \exp\left(\frac{2k\mu 2k\sigma^2 + 2^2k^2\sigma^4}{2k\sigma^2}\right) dr_k(k)$$

$$= X_0^2 \exp\left(2k\mu + \frac{2^2k\sigma^2}{2}\right) \underbrace{\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[r_k(k) - (k\mu + 2k\sigma^2)\right]^2}{2k\sigma^2}\right)}_{N(k\mu + 2k\sigma^2, k\sigma^2)} dr_k(k)$$

$$= X_0^2 \exp\left(2k\mu + \frac{2^2k\sigma^2}{2}\right). \quad \text{(The integral above is one.)}$$

(e) Find the variance of X_k for any k.

Solution:

$Var(X_k) = E(X_k^2) - [E(X_k)]^2$. We need to find $E(X_k)$. By the same logic of previous question, we have

$$\begin{aligned}
E(X_k) &= \int_{-\infty}^{\infty} X_0 \exp(r_k(k)) \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[\log x_0 + r_k(k) - (\log x_0 + k\mu)\right]^2}{2k\sigma^2} \right) dr_k(k) \\
&= \int_{-\infty}^{\infty} X_0 \exp(r_k(k)) \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[r_k(k) - k\mu\right]^2}{2k\sigma^2} \right) dr_k(k) \\
&= X_0 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[r_k(k) - k\mu - k\sigma^2)\right]^2}{2k\sigma^2} \right) \exp\left(\frac{2k\mu k\sigma^2 + k^2\sigma^4}{2k\sigma^2} \right) dr_k(k) \\
&= X_0 \exp\left(k\mu + \frac{k\sigma^2}{2}\right) \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi k}\sigma} \exp\left(\frac{-\left[r_k(k) - (k\mu + k\sigma^2)\right]^2}{2k\sigma^2} \right)}_{N(k\mu + k\sigma^2,\, k\sigma^2)} dr_k(k) \\
&= X_0 \exp\left(k\mu + \frac{k\sigma^2}{2}\right). \quad \text{(The integral above is one.)}
\end{aligned}$$

Hence,

$$\begin{aligned}
Var(X_k) &= E(X_k^2) - [E(X_k)]^2 \\
&= X_0^2 \exp\left(2k\mu + \frac{2^2 k\sigma^2}{2}\right) - \left(X_0 \exp\left(k\mu + \frac{k\sigma^2}{2}\right)\right)^2 \\
&= X_0^2 \exp\left(2k\mu + \frac{2^2 k\sigma^2}{2}\right) - X_0^2 \exp\left(2k\mu + k\sigma^2\right) \\
&= X_0^2 \exp\left(2k\mu + k\sigma^2\right)\left(\exp(k\sigma^2) - 1\right)
\end{aligned}$$

6. **Chapter 9 Problem 1** (a) Describe the signs of nonstationarity seen in the time series and ACF plots. (b) Use the augmented Dickey-Fuller tests to decide which of the series are nonstationary. Do the tests corroborate the conclusions of the time series and ACF plots?

Solution:

(a)

From the Figure 2, the time series of r, the 91-day treasure bill rate,is like a random walk, which is nonstationary because it does not appear to be mean-reverting, although it may not have momentum. At the middle of the Figure 2, the time series of y, the log of real GDP, has obvious momentum. it may not be stationary as well. We can not determine easily whether the time series of inflation rate, pi, is stationary or not, but it looks like a random walk too.

When checking the ACF of these three variables, all of them exist high autocorrelations even in short lags, which is a sign of nonstationarity. The ACF of *pi* decays faster than the other two, but it still does not decay exponentially fast. Hence, it may be a nonstationary or long term memory stationary process.

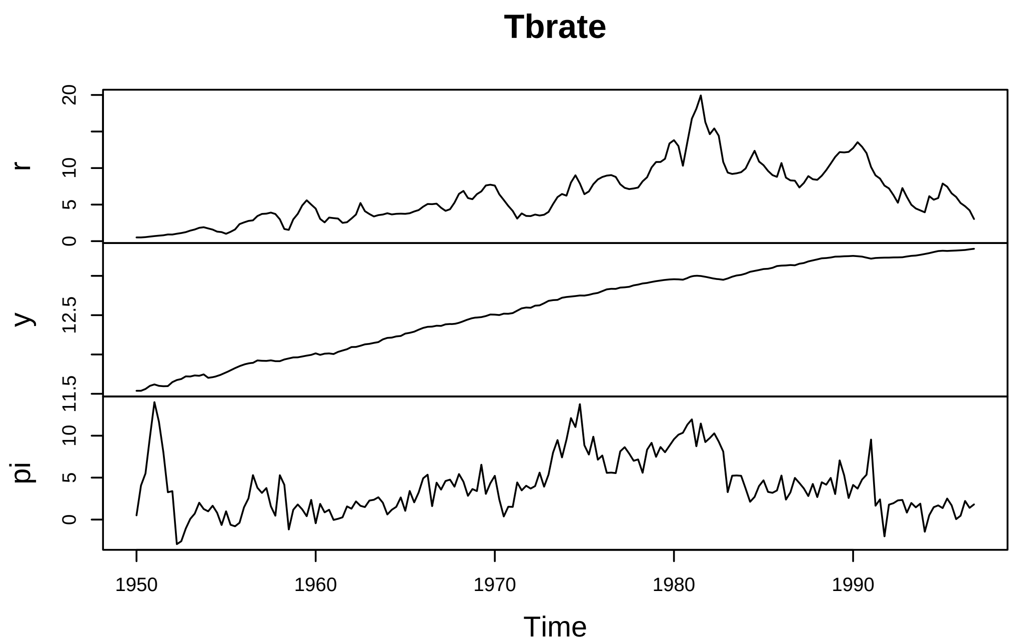


Figure 2: Time series of T-bill rate

(b)

The ADF test results for r, y and pi are the following:

```
Augmented Dickey-Fuller Test
data:  Tbrate[, 1]
Dickey-Fuller = -1.925, Lag order = 5, p-value = 0.6075
alternative hypothesis:  stationary

Augmented Dickey-Fuller Test
data:  Tbrate[, 2]
Dickey-Fuller = -0.3569, Lag order = 5, p-value = 0.9873
alternative hypothesis:  stationary

Augmented Dickey-Fuller Test
data:  Tbrate[, 3]
Dickey-Fuller = -2.9499, Lag order = 5, p-value = 0.1788
```

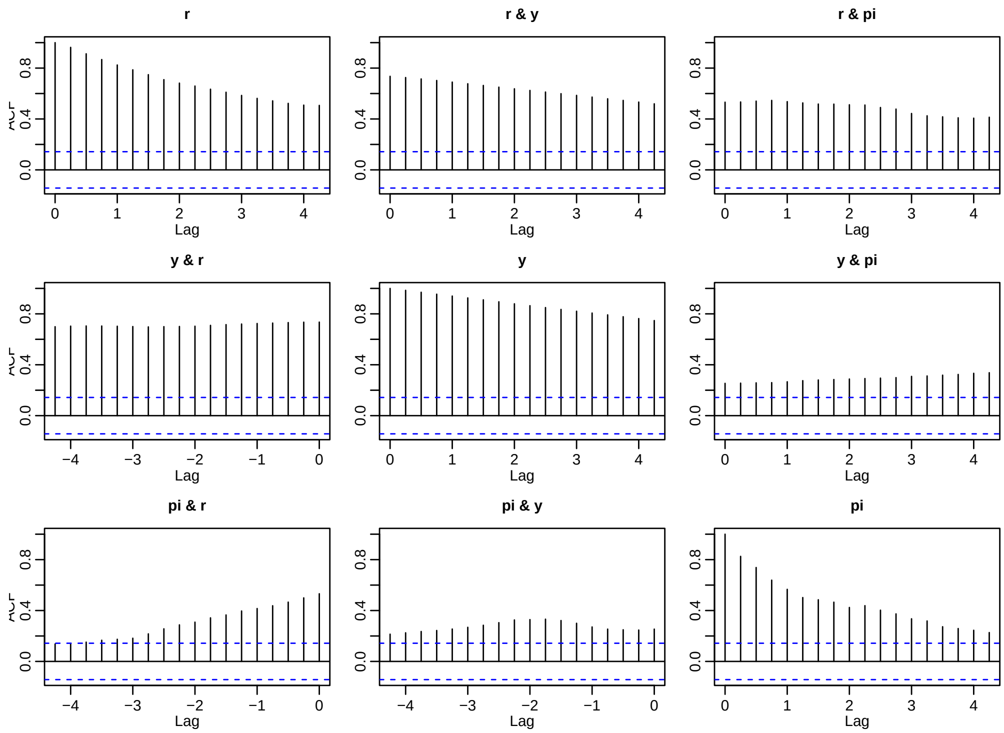


Figure 3: Sample ACF of T-bill rate

```
alternative hypothesis:  stationary
```

As we can see, all three tests do not reject their own null hypothesis, meaning that statistically these three processes are not stationary. Furthermore, the second test has the largest p-value. The momentum shows more obvious sign of nonstationarity. The third test has smaller p-value comparing with the other two. This corresponds to its stationary-like time series and ACF plot. However, it is still a nonstationary process. Hence, these three tests corroborate the conclusions of the time series and ACF plot we made in (a).

7. **Chapter 9 Problem 2** (a) Do the differenced series appear stationary according to the augmented Dickey-Fuller tests? (b) Do you see evidence of autocorrelations in the differenced series? If so, describe these correlations.
Solution:
(a)

The ADF test results for the differenced series are

```
Augmented Dickey-Fuller Test
data:  diff_rate[, 1]
Dickey-Fuller = -5.2979, Lag order = 5, p-value = 0.01
alternative hypothesis:  stationary
```

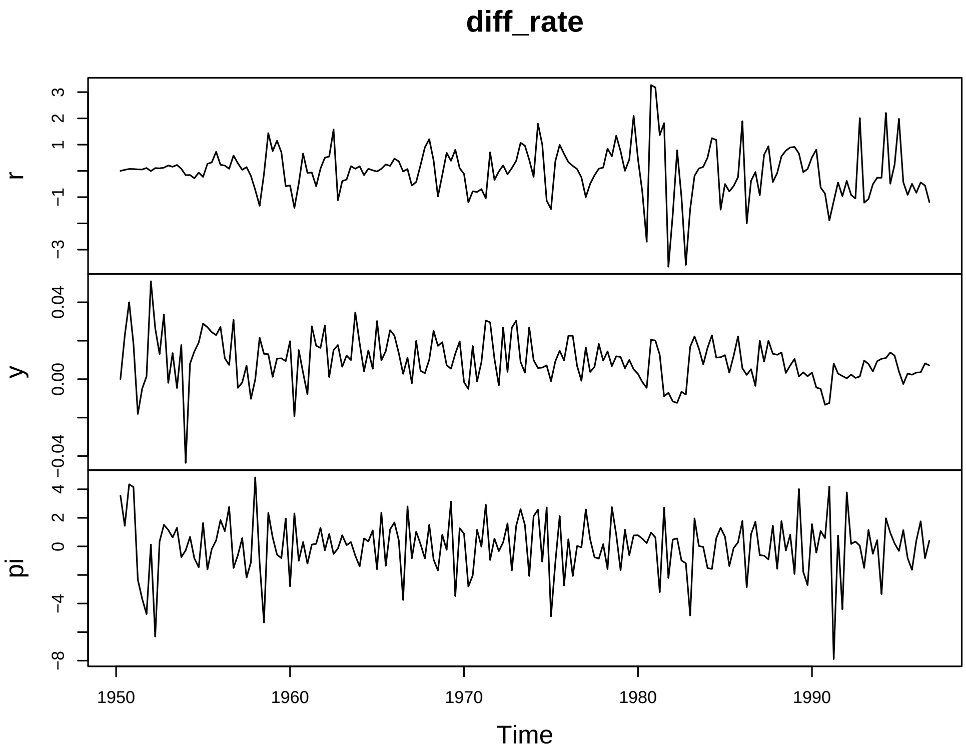


Figure 4: Sample ACF of differenced T-bill rate

```
Augmented Dickey-Fuller Test
data:  diff_rate[, 2]
Dickey-Fuller = -5.9168, Lag order = 5, p-value = 0.01
alternative hypothesis:  stationary

Augmented Dickey-Fuller Test
data:  diff_rate[, 3]
Dickey-Fuller = -7.6571, Lag order = 5, p-value = 0.01
alternative hypothesis:  stationary
```

Based on these results, we conclude that the differenced series appear to be stationary since all three p-values are less than 0.05, and we reject the null hypotheses.

(b)

As we can see in Figure 5, all three ACFs have some significant autocorrelation at lag 1, then decay very quickly within the blue dash line bounds,

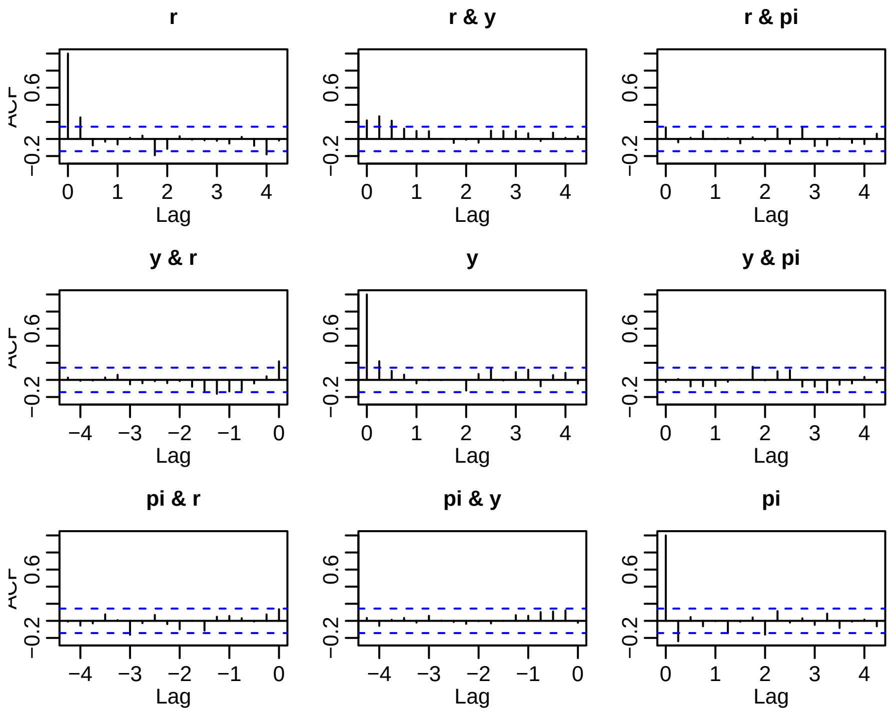


Figure 5: Sample ACF of differenced T-bill rate

showing that the differenced series become stationary. AR(1) or MA(1) of ΔTbrate could be considered.

8. **Chapter 9 Exercise 2** Next, fit AR(1) and AR(2) models to the CRSP returns:

```
arima(crsp,order=c(1,0,0))
arima(crsp,order=c(2,0,0))
```

(a) Would you prefer an AR(1) or an AR(2) model for this time series? Explain your answer.
Solution:

Both AR(1) and AR(2) fit the CRSP data pretty well. I would prefer AR(1) model. The reasons are following. After fitting AR(1) and AR(2), the results show:

```
fit1<-arima(crsp, order=c(1,0,0))
> fit1
Series:  crsp
ARIMA(1,0,0) with non-zero mean
Coefficients:
......... ar1 intercept
......... 0.0853 7e-04
```

```
s.e.  0.0198 2e-04
```
σ^2 `estimated as 5.973e-05:  log likelihood=8706.18`

```
AIC=-17406.37 AICc=-17406.36 BIC=-17388.86

fit2<-arima(crsp, order=c(2,0,0))
> fit2
Series:  crsp
ARIMA(2,0,0) with non-zero mean
Coefficients:
......... ar1 ar2 intercept
......... 0.0865 -0.0141 7e-04
s.e.  0.0199 0.0199 2e-04
```
σ^2 `estimated as 5.972e-05:  log likelihood=8706.43`

```
AIC=-17404.87 AICc=-17404.85 BIC=-17381.53
```

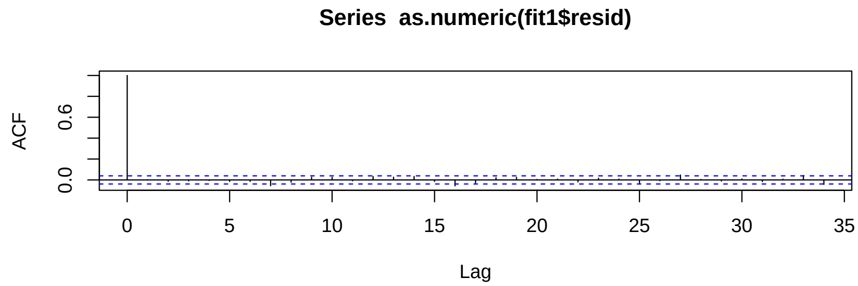
First, both AIC and BIC are smaller in AR(1) model than those in AR(2) model. Second, while both coefficients $\hat{\phi}_1$ are statistically significant since both are about 4.3 times their standard errors, the coefficient $\hat{\phi}_2$ seems not significant since the value of $\hat{\phi}_2$ is even smaller than its standard error.

Besides, I did the Ljung-Box texts and checked the ACF of residuals for both models. By setting lag is 15, the testing results are

```
> test1<-Box.test(fit1$resid, lag=15, type="Ljung-Box", fitdf=1)
> test1
Box-Ljung test
data:  fit1$resid
X-squared = 22.2905, df = 14, p-value = 0.07284
> test2<-Box.test(fit2$resid, lag=15, type="Ljung-Box", fitdf=2)
> test2
Box-Ljung test
data:  fit2$resid
X-squared = 21.7227, df = 13, p-value = 0.05981
```

We shall accept the null hypothesis that the residuals are uncorrelated, at least small lags 15. The p-value in AR(1) model are slightly lager than that in AR(2) model. The Figure 6 shows that residuals seems to be uncorrelated as well.

In conclusion, AR(1) and AR(2) are both convincing to the some extent. I would sightly prefer AR(1) since it has smaller AIC and BIC, no non-significant coefficients and higher p-value in Ljung-Box test. Last, if we can use a simpler AR(1) model to fit the data well, as a fancy model does, why do we need to use a complicated one?



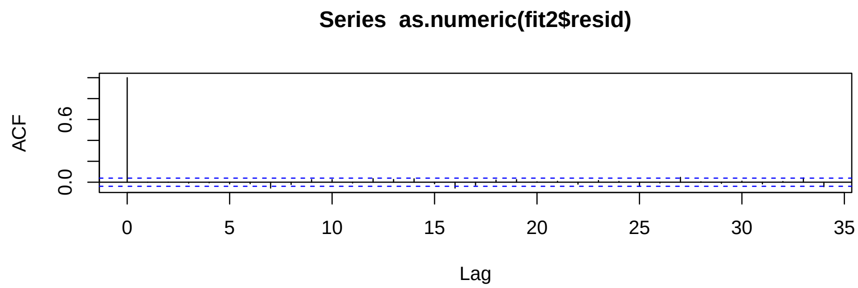


Figure 6: Sample ACF of residuals of AR(1) and AR(2)

(b) Find a 95% confidence interval for ϕ for the AR(1) model.
Solution:

The 95% confidence interval for ϕ for the AR(1) model is $\hat{\phi} \pm Z_{0.975} \cdot (s.e) = 0.0853 \pm 1.96 \cdot (0.0198) = (0.046492, 0.124108)$.

9. **Chapter 9 Exercise 3** Consider the AR(1) model $Y_t = 5 - 0.55Y_{t-1} + \epsilon_t$ and assume that $\sigma_\epsilon^2 = 1.2$.

(a) Is this process stationary? Why or why not?
Solution:

Since the $|\phi| = |-0.55| < 1$, the root of $\Phi(z) = 1 - \phi z = 0$ has a root greater than one. Hence, this process is (weakly) stationary.

(b) What is the mean of this process?
Solution:

$Y_t = (1-\phi)\mu + \phi Y_{t-1} + \epsilon_t = 5 - 0.55Y_{t-1} + \epsilon_t$. Therefore, $(1-\phi)\mu = (1+0.55)\mu = 5$. Hence the mean of this process is $\mu = 3.225806$ for all t.

(c) What is the variance of this process?
Solution:

The variance of this process is $Var(Y_t) = \dfrac{\sigma_\epsilon^2}{1-\phi^2} = \dfrac{1.2}{1-(-0.55)^2} = 1.72043$ for all t.

(d) What is the covariance function of this process?
Solution:

The covariance function of this process is $\gamma(h) = Cov(Y_t, Y_{t+h}) = \dfrac{\sigma_\epsilon^2 \phi^{|h|}}{1-\phi^2} = \dfrac{(1.2)(-0.55)^{|h|}}{1-(-0.55)^2} = 1.72043(-0.55)^{|h|}$ for all t and for all h.