# Project 1: Missing trading volume data

## Financial Time Series
## Course Code: TMS088/MSA410

**Siddhant Som & Devosmita Chatterjee**

# Theoretical Part

1. Let $\mu > 0$, $\sigma^2 > 0$ and let $Z \sim \text{WN}(\mu, \sigma^2)$. Let then $Y$ be the process defined by

$$Y_t = \sum_{j=0}^{q} \theta_j Z_{t-j}$$

for some coefficients $\theta_1, \cdots, \theta_q \in \mathbb{R}$, $q \in \mathbb{N}$, with $\theta_0 = 1$. $Y$ is called a moving average process of order $q$.
*Solution:* Given that $\mu > 0$, $\sigma^2 > 0$ and $Z \sim \text{WN}(\mu, \sigma^2)$, that is,

$$E(Y_t) = \mu, t \in \mathbb{N} \tag{1}$$

$$Var(Y_t) = \sigma^2, t \in \mathbb{N} \tag{2}$$

$$\gamma_Z(h) = \begin{cases} \sigma^2 & \text{if } h = 0, \\ 0 & \text{else} \end{cases} \tag{3}$$

and

$$Y_t = \sum_{j=0}^{q} \theta_j Z_{t-j} \tag{4}$$

for some coefficients $\theta_1, \cdots, \theta_q \in \mathbb{R}$, $q \in \mathbb{N}$, with $\theta_0 = 1$.
a) Show that for any $(t, h) \in \mathbb{Z}^2$, $E(Y_t) = E(Y_{t+h})$.
*Solution:*

$$E(Y_t) = E(\sum_{j=0}^{q} \theta_j Z_{t-j})$$

$$= \sum_{j=0}^{q} \theta_j E(Z_{t-j})$$

The above equation was written using the property of linearity of expectation. Since $Z \sim \text{WN}(\mu, \sigma^2)$, then $E(Z_t) = \mu$, $t \in \mathbb{Z}$ i.e the mean is independent of $t$. Substituting this in the above equation, we get:

$$E(Y_t) = \sum_{j=0}^{q} \theta_j \mu \ (\because Using\ (1))$$

$$= \mu \sum_{j=0}^{q} \theta_j$$

This equation is independent of t. Hence, calculating E($Y_{t+h}$) the same way, we get:

$$E(Y_{t+h}) = \mu \sum_{j=0}^{q} \theta_j$$

This concludes our proof.

b) Show that $(t,s,h) \in \mathbb{Z}^3$, $Cov(Y_t, Y_{t+h}) = Cov(Y_s, Y_{s+h})$
*Solution:*

$$Cov(Y_t, Y_{t+h}) = E(Y_t Y_{t+h}) - E(Y_t)E(Y_{t+h})$$

$$= E(\sum_{j=0}^{q} \theta_j Z_{t-j} \sum_{k=0}^{q} \theta_k Z_{t+h-k}) - E(\sum_{j=0}^{q} \theta_j Z_{t-j}) E(\sum_{k=0}^{q} \theta_k Z_{t+h-k})$$

$$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k E(Z_{t-j} Z_{t+h-k}) - \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k E(Z_{t-j}) E(Z_{t+h-k})$$

$$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k [E(Z_{t-j} Z_{t+h-k}) - E(Z_{t-j}) E(Z_{t+h-k})]$$

$$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k Cov(Z_{t-j}, Z_{t+h-k})$$

$$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \gamma_z(h + j - k)$$

This equation is independent of $t$ and we would get the same result when $t = s$. Hence $Cov(Y_t, Y_{t+h}) = Cov(Y_s, Y_{s+h})$.

c) Show that $Y$ is stationary and give its autocovariance function.
*Solution:*

$$Var(Y_t) = Var(\sum_{j=0}^{q} \theta_j Z_{t-j})$$

$$= E((\sum_{j=0}^{q} \theta_j Z_{t-j})^2) - E((\sum_{j=0}^{q} \theta_j Z_{t-j}))^2$$

$$= \sum_{j=0}^{q} \theta_j^2 E(Z_{t-j}^2) - \sum_{j=0}^{q} \theta_j^2 (E(Z_{t-j}))^2 \ (\because Linearity)$$

$$= \sum_{j=0}^{q} \theta_j^2 [E(Z_{t-j}^2) - (E(Z_{t-j}))^2]$$

$$= \sum_{j=0}^{q} \theta_j^2 Var(Z_{t-j})$$

$$= \sum_{j=0}^{q} \theta_j^2 \sigma^2 \ (\because Using\ (2))$$

$$= \sigma^2 \sum_{j=0}^{q} \theta_j^2$$

$$< +\infty$$

From part a), $E(Y_t) = E(Y_{t+h})$ for any $(t,h) \in \mathbb{Z}^2$.
From part b), $Cov(Y_t, Y_{t+h}) = Cov(Y_s, Y_{s+h})$, $(t,s,h) \in \mathbb{Z}^3$.
Therefore, $Y$ is stationary.

The autocovariance function of $Y$ is given by

$$\gamma_Y(h) = Cov(Y_0, Y_h) = Cov(Y_t, Y_{t+h})$$

$$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \gamma_Z(h-k+j)$$

$$= \begin{cases} \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k \sigma^2 & \text{if } h-k+j=0,\ j=0,\cdots,q,\ k=0,\cdots,q \\ 0 & \text{else} \end{cases}$$

This condition is supposed to help you simplify the double sum, which you did not do.

↳ this notation does not make sense in this context.

$$= \begin{cases} \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} \sigma^2 & \text{if } |h| \le q \\ 0 & \text{if } |h| > q \end{cases}$$

The formula is correct but you did not provide any justification for it.

with $\theta_0 = 1$.

d) Prove that if $Z$ is a Gaussian process, then $Y_t$ is independent of $Y_{t+h}$ for any $t \in \mathbb{Z}$ and $|h| > q$.

*Solution:*

Given that $Z$=($Z_t$, $t \in \mathbb{Z}$) is white noise. Let $Z_t$ be Gaussian white noise (IID - independent and identically distributed).

→ the goal of the task was to prove independance, basically

First we have to show that $Y = \sum_{i=1}^{n} a_i Z_i, a_i \in \mathbb{R}$ is a Gaussian ~~process~~.
Let $M_Y(t)$ be the moment generating function of $Y$. Let $M_{Z_i}(t)$ be the moment generating function of $Z_i$. Since $Z_i$ are Gaussian variables each with mean $\mu$ and variance $\sigma^2$, we have that

What is the link with $Y_t$?

$$M_{Z_i}(t) = E[e^{tZ_i}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2},$$

$$M_{a_i Z_i}(t) = E[e^{t(a_i Z_i)}] = e^{a_i \mu t + \frac{1}{2}a_i^2 \sigma^2 t^2}.$$

Since $Z_i$ are independent with respect to each other and $Y$ is a linear combination of $Z_i$, we have that $M_Y(t)$ is a product of $M_{Z_i}(t)$, that is,

$$M_Y(t) = \prod_{i=1}^{n} M_{Z_i}(t) = \prod_{i=1}^{n} e^{a_i \mu t + \frac{1}{2}a_i^2 \sigma^2 t^2} = e^{\sum_{i=1}^{n} a_i \mu t + \frac{1}{2}\sum_{i=1}^{n} a_i^2 \sigma^2 t^2}$$

Thus, we have that $Y \sim \mathrm{N}(\sum_{i=1}^{n} a_i \mu, \sum_{i=1}^{n} a_i^2 \sigma^2)$.

Therefore, the linear combination of independent Gaussian random variables is also Gaussian. This implies that if $Z_t, \cdots, Z_{t-q}$ are $q+1$ independent Gaussian random variables ($Z \sim N(\mu,\sigma^2)$), then the random variable $Y$ is also Gaussian, where $Y$ is a linear combination of $Z_{t-j}$'s, $j = 0, 1, \cdots, q$ given by $Y_t = \sum_{j=0}^{q} \theta_j Z_{t-j}$, $\theta_j \in \mathbb{R}$. Rephrase

→ Not enough to prove that $Y$ is a Gaussian process.

The joint probability function of ($Y_t$,$Y_{t+h}$) is given by

$Y_t$ and $Y_{t+h}$ are variables, not events

$$P(Y_t \cancel{\bigcap} Y_{t+h}) = P(\sum_{j=0}^{q} \theta_j Z_{t-j} \cancel{\bigcap} \sum_{k=0}^{q} \theta_k Z_{t+h-k}) \; (\because Using \; (4))$$

$$\cancel{= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j \theta_k P(Z_{t-j} \bigcap Z_{t+h-k})} \; (\because Linearity)$$

Absolutely not! The probability functions are not linear!

$$= \sum_{j=0}^{q}\sum_{k=0}^{q} \theta_j \theta_k P(Z_{t-j})P(Z_{t+h-k}) \; (\because Z_t \; are \; IID.)$$

$$= P(\sum_{j=0}^{q} \theta_j Z_{t-j})P(\sum_{k=0}^{q} \theta_k Z_{t+h-k}) \; (\because Linearity)$$

$$= P(Y_t)P(Y_{t+h})$$

Therefore, $Y_t$ is independent of $Y_{t+h}$ for any $t \in \mathbb{Z}$ and $|h| > q$.

# Practical Part

1. The file *intel.csv* contains daily trading volume data for the Intel Corporation stock at Nasdaq between March $19^{th}$, 2001 and November $30^{th}$, 2020 with $N = 4886$ data points. The variable *Volume* contains the original volume trading. The variable *VolumeMissing* is a copy of the variable *Variable* but has 100 data points missing, replaced by NaN entries. Your task is to work with the differenced log time series $Y = Y_t$, $(t = 1, \cdots, N-1)$, given by

$$Y_t := \log X_{t+1} - \log X_t$$

where $(X_t,\ t = 1, \cdots, N-1)$ are the data points in *VolumeMissing* and reconstruct the missing points using the theory of linear time series. Note that $Y$ will have more missing values than $X$ since, for a given t$\in\{1, \cdots, N-1\}$, $Y_t$ is missing if either $X_{t+1}$ or $X_t$ is missing.
2. Compute the time series $Y$ and plot the sample auto-correlation function (ACF) $\hat{\rho}_Y$(h) for $h = 0, \cdots, 20$. As we will see in the lectures, the ACF is consistent with $Y$ being a moving average process of order $q$, as defined in (1) (let us assume here and below that $\mu = 0$). Such a process has a feature that the sample ACF values $\hat{\rho}_Y(h)$ are approximately IID $\sim N(0, \mathrm{N}^{-1})$ for $h > q$. Based on this fact, choose a reasonable value of $q$ based on your computed ACF $\hat{\rho}_Y$, assuming that the data is a realization of (1).

*Solution:*
The time series $X$ where $X = (X_t, t = 1, \cdots, 4958)$ are the data points in *VolumeMissing* is computed in figure 1. For the given data *intel.csv*, $N = 4958$.

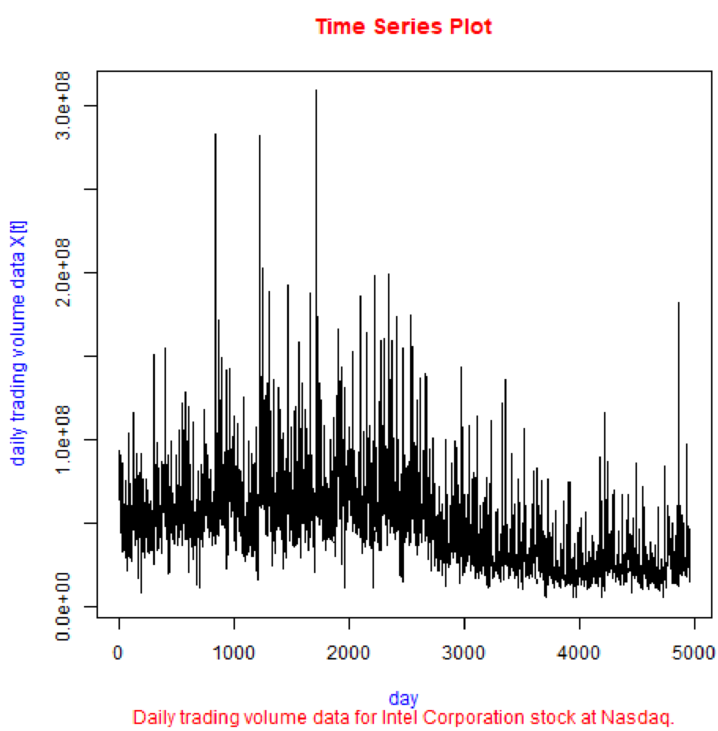


**Figure. 1** Time series $X$ where $X = (X_t,\ t = 1, \cdots, 4958)$ are the data points in *VolumeMissing*.

The time series Y where $Y = (Y_t,\ t = 1, \cdots, N-1)$ is computed in figure 2. Here $Y_t = \log X_{t+1} - \log X_t$ where $(X_t,\ t = 1, \cdots, N)$ are the data points in *VolumeMissing*. For the given data *intel.csv*, $N = 4958$.

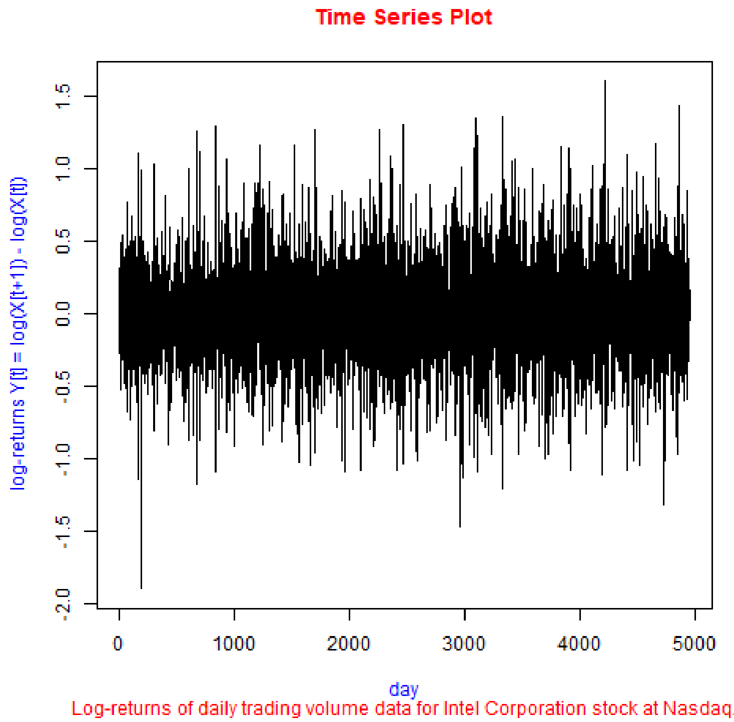


**Figure. 2** Time series $Y$ where $Y = (Y_t,\ t = 1, \cdots, 4957)$. Here $Y_t = \log X_{t+1} - \log X_t$ where $(X_t,\ t = 1, \cdots, 4958)$ are the data points in *VolumeMissing*.

ACF plot represents a bar chart of coefficients of correlation between a time series and it lagged values. In the ACF plot, the blue dashed lines represent an approximate 95% confidence interval which is given by C.I. $= \bar{x} \pm 1.96 s_x$ where $\bar{x}$ is the mean and $s_x$ is the standard deviation. The blue dashed lines give the values beyond which the autocorrelations are statistically significantly different from zero.

mean and standard deviation of what?

The sample auto-correlation function (ACF) for the given data *intel.csv* is plotted in figure 3. Since the sample ACF values $\hat{\rho}_Y(h)$ are approximately IID $\sim N(0, N^{-1})$, the mean is $\bar{x} = 0$ and the standard deviation is $s_x = \frac{1}{\sqrt{N}}$ where $N = 4958$. In figure 3, the blue dashed lines represent an approximate 95% confidence interval, given by, C.I. $= 0 \pm \frac{1.96}{\sqrt{N}} = \pm\frac{1.96}{\sqrt{4958}} = \pm 0.02783574$. From figure 3 and table 1, we see that the autocorrelations at lags 1, 2, 3 and 4 are out of these bounds while all the other autocorrelations at lags 5 - 20 are within these bounds. The only nonzero values in the theoretical ACF are for lags 1, 2, 3 and 4 while autocorrelations for higher lags 5 - 20 are almost zero. Autocorrelations are statistically significant at lags 1, 2, 3 and 4, but autocorrelations are non-significant for higher lags 5 - 20.

Careful with your use of the word significant (see below)

**tab. 1** The table represents the sample auto-correlation function (ACF) $\hat{\rho}_Y$(h) for $h = 0, \cdots, 20$

| Lag | Autocorrelation Function | Confidence Interval $= (-\frac{1.96}{\sqrt{4958}}, \frac{1.96}{\sqrt{4958}})$ = (-0.02783574,0.02783574) | ~~Statistical Significance~~ |
|---|---|---|---|
| 1 | -0.2964350058 | | Significant |
| 2 | -0.0955677625 | | Significant |
| 3 | -0.0468429283 | | Significant |
| 4 | -0.0338199004 | | Significant |
| 5 | 0.0183961382 | ✓ | Non-significant |
| 6 | 0.0012560940 | ✓ | Non-significant |
| 7 | -0.0161779941 | ✓ | Non-significant |
| 8 | -0.0152569478 | ✓ | Non-significant |
| 9 | -0.0118697314 | ✓ | Non-significant |
| 10 | 0.0183224359 | ✓ | Non-significant |
| 11 | 0.0094135502 | ✓ | Non-significant |
| 12 | -0.0023358759 | ✓ | Non-significant |
| 13 | -0.0121675147 | ✓ | Non-significant |
| 14 | 0.0001772998 | ✓ | Non-significant |
| 15 | -0.0060763378 | ✓ | Non-significant |
| 16 | -0.0119917437 | ✓ | Non-significant |
| 17 | -0.0070467143 | ✓ | Non-significant |
| 18 | 0.0078155343 | ✓ | Non-significant |
| 19 | -0.0038982318 | ✓ | Non-significant |
| 20 | 0.0076246353 | ✓ | Non-significant |

This is not statistical significance. You are just checking whether the values are within some bounds here...

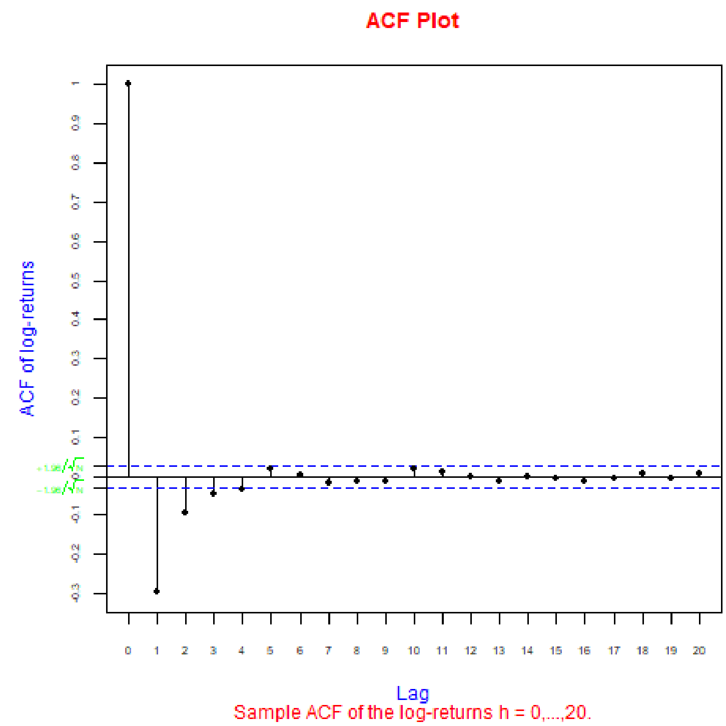


**Figure. 3** Sample auto-correlation function (ACF) $\rho_Y$(h) for $h = 0, \cdots, 20$.

The ACF is consistent with $Y$ which is a moving average process of order $q$, as defined in (1) (assuming that $\mu = 0$). Such a process has a feature that the sample ACF values $\hat{\rho}_Y(h)$ are approximately IID $\sim N(0, N^{-1})$ for $h > q$. Based on this fact, a reasonable value of $q$ is 4, based on our computed ACF $\hat{\rho}_Y$, assuming that the data is a realization of (1).

The full R code can be found in appendix A.

3. Let $\mathbb{M}$ be the set of indices of the missing values of $Y$. Use Corollary 2.4.6 in the lecture notes to write a program that, for each $t \in$ M, calculates the best linear predictor $b_t^l(Y^q)$ (but use our computed ACF $\hat{\rho}$). Here $Y^q := (Y_s : max(1, t\text{-}q) \leq s \leq min(N,\ t{+}q); s \notin \mathbb{M})$.

*Solution:*
Corollary 2.4.6 states that
For $X = (X_t,\ t \in \mathbb{Z})$ (the series being predicted) and $X^n = (X_1, X_2, \cdots, X_n)$ (set of random variables being used for prediction.), assuming that $X$ is stationary with mean $\mu$ and autocovariance function $\gamma$, the coefficients $(a^i,\ i = 0, \cdots, n)$ of the best linear predictor are determined by the following linear equations:

$$a_0 = \mu(1 - \sum_{i=1}^{n} a_i) \tag{5}$$

ok

$$\Gamma_n(a_1, a_2......a_n)' = (\gamma(t - t_n), ...., \gamma(t - t_1))' \tag{6}$$

with

$$\Gamma_n = (\gamma(t_{n+1-j} - t_{n+1-i}))_{i,j=1}^{n} \tag{7}$$

The full R code can be found in appendix B.

4. Now compute the differenced series using all available data, i.e., compute $y = (y_t, t = 1, \cdots, N-1)$, given by

$$y_t := \log(x_{t+1}) - \log(x_t)$$

where $(x_t,\ t = 1, \cdots, N)$ are the data points in *Volume*. Let $\hat{Y}$ denote the modified process consisting of Y where each missing data point $Y_t$ has been replaced by the best linear predictor $b_t^l\ (Y^q)$ computed as in the previous task. Let $\check{Y}$ denote the modified process consisting of $Y$ where each missing data point $Y_t$ has been replaced by a value calculated by simple linear interpolation. Calculate and report the root mean squared errors for the two series: ok

$$\sqrt{M^{-1} \sum_{t \in M} (y_t) - \hat{Y}_t)^2} \tag{8}$$

and

$$\sqrt{M^{-1} \sum_{t \in M} (y_t) - \check{Y}_t)^2} \tag{9}$$

where $M = |\mathbb{M}|$, the number of indices in M.

*Solution:*
The root mean square error value for the series described by equation (9) was found to be ~~0.5813~~ while the root mean square error for the series described by equation (10) was found to be 0.5013.

The full R code can be found in appendix C.

No interpretation of the results?

# Appendix A

**Read the given data intel.csv.**
*data = read.csv("intel.csv",header=TRUE)*
*data*
**Number of datapoints**
*length(data[,1])*
**Number of missing values in VolumeMissing**
*sum(is.na(data[,2]))*
**Compute and plot the time series X where X[t] are the data points in VolumeMissing.**
**png("TimeSeries.png")**
*X = data[,2]*
**VolumeMissing**
*plot(X, type = "l", main = "Time Series Plot", sub = expression("Daily trading volume data for Intel Corporation stock at Nasdaq."), xlab = "day", ylab = "daily trading volume data X[t]", col.main = "red", col.sub = "red", col.lab = "blue")*
**dev.off()**

**Compute and plot the differenced log time series Y[t] = X[t+1] - X[t] where X[t] are the data points in VolumeMissing.**
**png("logTimeSeries.png")**
*X = data[,2]* **VolumeMissing**
*Y = matrix(, nrow = length(X)-1, ncol = 1)*
*for (i in 1:length(X)-1)*
*Y[i] = log(X[i+1]) - log(X[i])* ✓

*plot(Y, type = "l", main = "Time Series Plot", sub = expression("Log-returns of daily trading volume data for Intel Corporation stock at Nasdaq."), xlab = "day", ylab = "log-returns Y[t] = log(X[t+1]) - log(X[t])", col.main = "red", col.sub = "red", col.lab = "blue")*
**Compute the sample autocorrelation function (ACF) of the log-returns.**
*L = acf(Y, lag = 20, type = "correlation", na.action = na.pass, plot = FALSE)* ✓
*ACF = Lacf*

**Compute whether autocorrelations are statistically significant.**
*for (i in 1:21)*
*if (ACF[i] > 1.96/sqrt(4958))*
*print(i)*
*else if (ACF[i] < -1.96/sqrt(4958))*
*print(i)*
*end*
*end*
**Plot the sample autocorrelation function (ACF) of the log-returns.**
**png("ACF.png")**
*plot(L, xaxt = 'no', main = "", sub = expression("Sample ACF of the log-returns h = 0,...,20."), ylab = "ACF of log-returns", xlab = "Lag", col.sub = "red", col.lab =*

*"blue", axes = FALSE, xaxt = 'n', yaxt = 'n')*
*par(new = TRUE)*
*plot(Lacf, main = "ACF Plot", xlab = "", ylab = "", xaxt = 'n', yaxt = 'n', col.main = "red", pch = 20)*
*axis(1, at = seq(1,21,1), labels = seq(0,20,1), cex.axis = 0.6)*
*axis(2, at = seq(-1,1,0.1), labels = seq(-1,1,0.1), cex.axis = 0.6)*
*par(las = 2)*
*axis(2, at = 1.96/sqrt(4958), labels = expression(+1.96/sqrt(N)), cex.axis = 0.5, col.axis = "green")*
*axis(2, at = -1.96/sqrt(4958), labels = expression(-1.96/sqrt(N)), cex.axis = 0.5, col.axis = "green")*

# Appendix B

**Set of indices of the missing values of Y.**
M = which(is.na(Y))

**Initialize best linear predictors for all missing values.**
bestlinearpredictor = matrix(, length(M), 1)

**Loop starts**
for (j in 1:length(M))

t = M[j] - **Index of missing value**

q = 4 - **Lag**

N = 4958-1 - **Total number of data points**

max = max(1,t-q) - **Maximum**

min = min(N,t+q) - **Minimum**

int = seq.int(max,min) - **Sequence of integers**

**Sequence of integers which are missing and can not be used for prediction**
drop = intersect(int,M)

s = int[!int %in% drop] - **Set of indices without missing values**

n = length(Y[s]) - **Length of set of indices without missing values**

**Compute autocorrelation function for Y[s].**
L = acf(Y[s], lag = n-1, type = "correlation", plot = FALSE)

You're using the wrong ACF → Your ACF is based on a very small subset of the time series and therefore is not representative. You should have used the ACF computed in the previous task.

**Construct left hand side of linear system of equations from corollary 2.4.6.**
**Construct covariance matrix**
gamma = data.matrix(Lacf)
s1 = dim(gamma)
Gamma = matrix(1, s1, s1)
d = row(Gamma) - col(Gamma)
for (i in 1:(s1 - 1))
Gamma[d == i — d == (-i)] = gamma[i + 1]
end

? → This is not the formula in the lecture notes.

**Construct right hand side of linear system of equations from corollary 2.4.6.**
ACF = data.matrix(Lacf)
b = matrix(, length(s), 1)
for (k in length(s):-1:1)

*c = abs(t-s[k])*
*b[k] = ACF[c]*

**Solve the linear system of equations from corollary 2.4.6 to find the coefficients.**
*A = Gamma*
*a1 = solve(A, b)*

**Find mean**
*result.mean = mean(Y[-M])*

**Compute first coefficient**
*a0 = result.mean*(1-colSums(a1))*

**Compute best linear predictors for each missing value**
*bestlinearpredictor[j] = a0+t(Y[s])%*%a1*
*bestlinearpredictor*

Careful. Your first entry of a1 should be multiplied by the last entry of Y[s] (see formula).

# Appendix C

**Compute the differenced log time series Y2[t] = X2[t+1] - X2[t] where X2[t] are the data points in Volume.**

*X2 = data[,1]* **Volume**
*Y2 = matrix(, nrow = length(X2)-1, ncol = 1)*
*for (i in 1:length(X2)-1)*
*Y2[i] = log(X2[i+1]) - log(X2[i])*

**Compute modified Y where each missing value is replaced by best linear predictor**

*. Ycap = Y*
*Ycap[M] = bestlinearpredictor* ok
*RMSE = sqrt((1/length(M))*colSums(data.matrix((Y2[M]-Ycap[M])2)))*

**Compute modified Y where each missing value is replaced by a value using simple linear interpolation.**
**install.packages("baytrends")**
**install.packages("gapfill")**
*Ybar = fillMissing(Y)*
*RMSE2 = sqrt((1/length(M))*colSums(data.matrix((Y2[M]-Ybar[M])2)))*