Solution to Tutorial 3

Section 1: Seasonal Indicator Regression

1. The R code to fit the regression model **(10)**:

```r
library(astsa)
trend = time(jj) - 1970  #"center time" to reduce collinearity

#construct the Q matrix of indicator functions
#method 1
Q1 <- rep(c(1,0,0,0), 21)
Q2 <- rep(c(0,1,0,0), 21)
Q3 <- rep(c(0,0,1,0), 21)
Q4 <- rep(c(0,0,0,1), 21)
Ind <- cbind(Q1, Q2, Q3, Q4)

#method 2
#turn Quarters 1,2,3,4 into factors, this is one-hot encoding in machine learning
Q = factor(cycle(jj))

reg = lm(log(jj) ~ 0 + trend + Q, na.action=NULL)
#no intercept
#na.action=NULL to preserve ts class in jj when doing linear regression

(W = model.matrix(reg))  #view the design matrix
kappa(crossprod(W)) #condition number of W^TW, lower is better. Compare the condition number by not subtracting 1970 from time(jj).

summary(reg)
```

Then the output from `summary(reg)` is

```
Call:
lm(formula = log(jj) ~ 0 + trend + Q, na.action = NULL)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29318 -0.09062 -0.01180  0.08460  0.27644

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
trend 0.167172   0.002259   74.00   <2e-16 ***
Q1    1.052793   0.027359   38.48   <2e-16 ***
Q2    1.080916   0.027365   39.50   <2e-16 ***
Q3    1.151024   0.027383   42.03   <2e-16 ***
Q4    0.882266   0.027412   32.19   <2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1254 on 79 degrees of freedom
Multiple R-squared:  0.9935,    Adjusted R-squared:  0.9931
F-statistic:  2407 on 5 and 79 DF,  p-value: < 2.2e-16
```

2. Taking expectation on both sides of the regression model:

$$\mu_t = \mathrm{E}(X_t) = \beta(t - 1970) + \sum_{i=1}^{4} \alpha_i Q_i(t).$$

Let us write quarters $1, 2, 3$ and 4 in year 1960 as $1960.00, 1960.25, 1960.5, 1960.75$ respectively. Then

$$\mu_{1961.00} - \mu_{1960.00} = \beta(1961.00 - 1960.00) + \alpha_1 - \alpha_1 = \beta.$$

Therefore the estimated average log annual increase is

$$\widehat{\mu}_{1961.00} - \widehat{\mu}_{1960.00} = \widehat{\beta} = 0.167172.\mathbf{(5)}$$

3. Let us use the year 1960, the answer is the same for all other years:

$$\begin{aligned}\mu_{1960.75} - \mu_{1960.5} &= \beta(1960.75 - 1960.5) + \alpha_4 - \alpha_3 \\ &= 0.25\beta + \alpha_4 - \alpha_3.\end{aligned}$$

Then the estimated difference between the fourth quarter and the third quarter is

$$\begin{aligned}\widehat{\mu}_{1960.75} - \widehat{\mu}_{1960.5} &= 0.25\widehat{\beta} + \widehat{\alpha}_4 - \widehat{\alpha}_3 \\ &= 0.167172(0.25) + 0.882266 - 1.151024 = -0.2269646.\end{aligned}$$

In R, this can be computed by

```
b <- reg$coefficients
b[1]*0.25+b[5]-b[4]
```

Therefore, the average logged earnings decrease by 0.2269646 from the third quarter to the fourth quarter.**(5)**

4. If we include an intercept term in the model, then the sum of the columns corresponding to the indicator functions $Q_1(t), Q_2(t), Q_3(t)$ and $Q_4(t)$ will equal to the column of ones for the intercept. In other words, the columns in the design matrix $\boldsymbol{W}$ are not linearly independent anymore. Hence $\boldsymbol{W}^T\boldsymbol{W}$ is not invertible and the least squares estimates for $\beta, \alpha_1, \alpha_2, \alpha_3$ and α_4 are not unique. To solve this problem, R will discard $Q_4(t)$ by setting the estimate for α_4 as NA.

```r
regco = lm(log(jj) ~ trend + Ind)
summary(regco)
```

(5)

5. To plot the data and superimpose the fitted values (Figure 1), we use

```r
plot(log(jj), main="Johnson and Johnson Quarterly Earnings",
  ylab = "Log earnings", xlab="Year", lwd=2, col="black")
lines(fitted(reg), col="red", lwd=2)

legend(1960, 2, legend=c("J&J data", "Model"), col=c("black", "red"), lty=1, cex=1)
```

Let us plot the residuals (Figure 2):

```r
res <- log(jj) - fitted(reg)
plot(res, main="Residuals", ylab="")
```

(5)

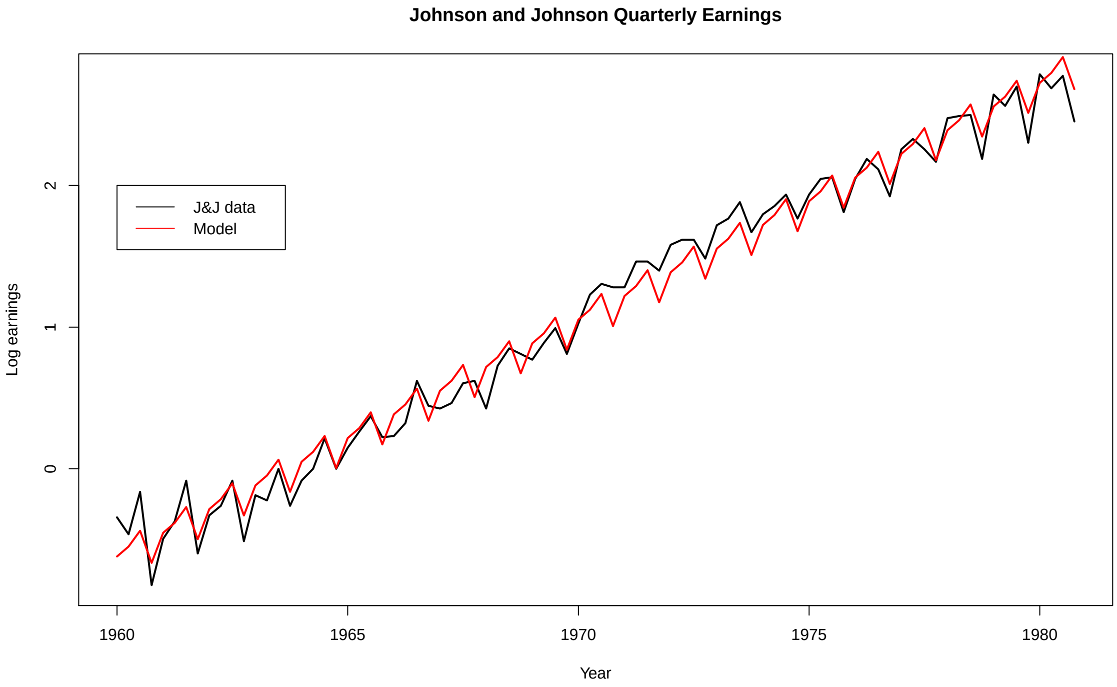

Figure 1: Johnson and Johnson log quarterly earnings (black) and the estimated log earnings in red.

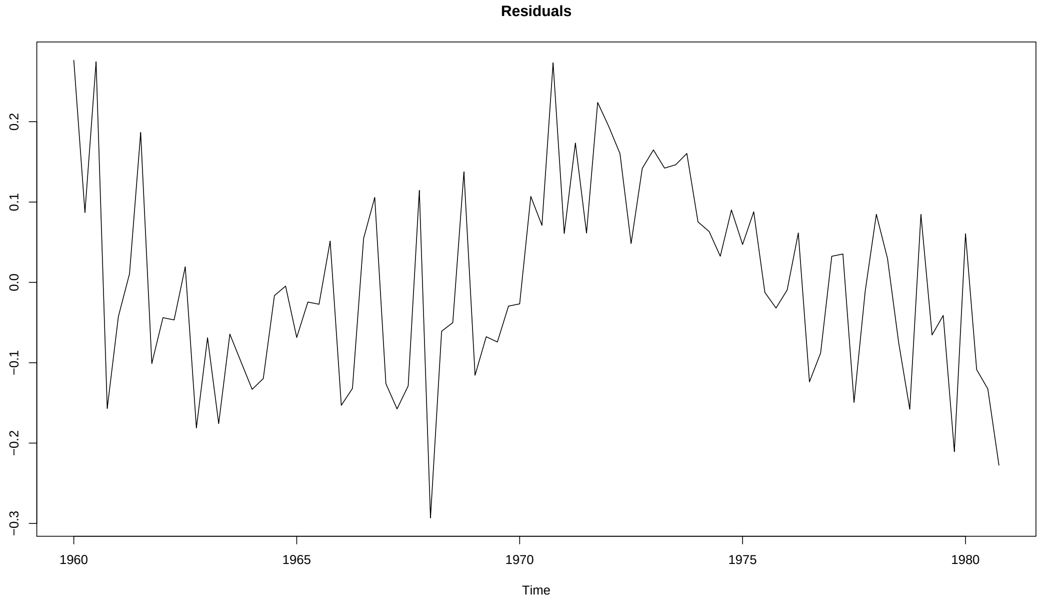

Figure 2: Residuals from fitting linear trend and seasonal indicator regression to the Johnson and Johnson data.

By looking closely at Figure 2, we see that the residuals still show periodic behavior. It seems that this period is approximately 10 years. Hence our model did not capture this periodic oscillation, and the residuals do not look like white noise.

Fourier analysis

Therefore, in addition to yearly cycles with period $s = 4$, it seems that there are other cycles present in the Johnson and Johnson data. In order to better identify the frequencies and periodicities of these hidden cycles, we will use **frequency domain** methods such as **fourier analysis**. The idea behind these methods is that to better find cycles in our data, it is more convenient to analyze time series in terms of frequencies rather than time as we have done so far. To do this, we transform the ACVF $\gamma(h)$ to its **spectral density**:

$$f(\omega) = \sum_{h=-\infty}^{\infty} e^{-2\pi i \omega h} \gamma(h)$$

where $i = \sqrt{-1}$ is the imaginary number and ω is **frequency** in Hertz (cycles per unit time).

Frequency and period are inversely related:

$$s = \frac{1}{\omega}.$$

Therefore analyzing the spectral density of our data will give us an idea about its cycles and periods.

However $f(\omega)$ is unknown in practice, so we estimate it using the **periodogram** $I(\omega)$. Technically, $I(\omega)$ is the squared modulus of the **discrete Fourier transform** of our time series data. Typically, we will smooth $I(\omega)$ using moving average filters called modified Daniell kernels. We can plot $I(\omega)$ versus frequency ω by

```
cycle = mvspec(res, log="no") #periodogram
```

In Figure 3, frequencies with spikes in the periodogram correspond to prominent cycles in our time series at these frequencies. Hence we see that the residuals contain cycles with frequencies approx $0.0\dot{8}$Hz, 0.98Hz and 1.96Hz respectively. The small peak attached to the larger peak near 0Hz is called **leakage** and it is an artifact from the estimating procedure. Applying the inverse relationship between frequency ω and period s, let us find the period corresponding to the frequency with maximum spike $0.0\dot{8}$Hz:

```
index = which(cycle$spec==max(cycle$spec))
maxfreq = cycle$freq[index] #maximum frequency
period = 1/maxfreq  #the cycle period with largest intensity
```

We see that the period corresponding to the largest cycle in Figure 2 is 11.25 years or 45 quarters.

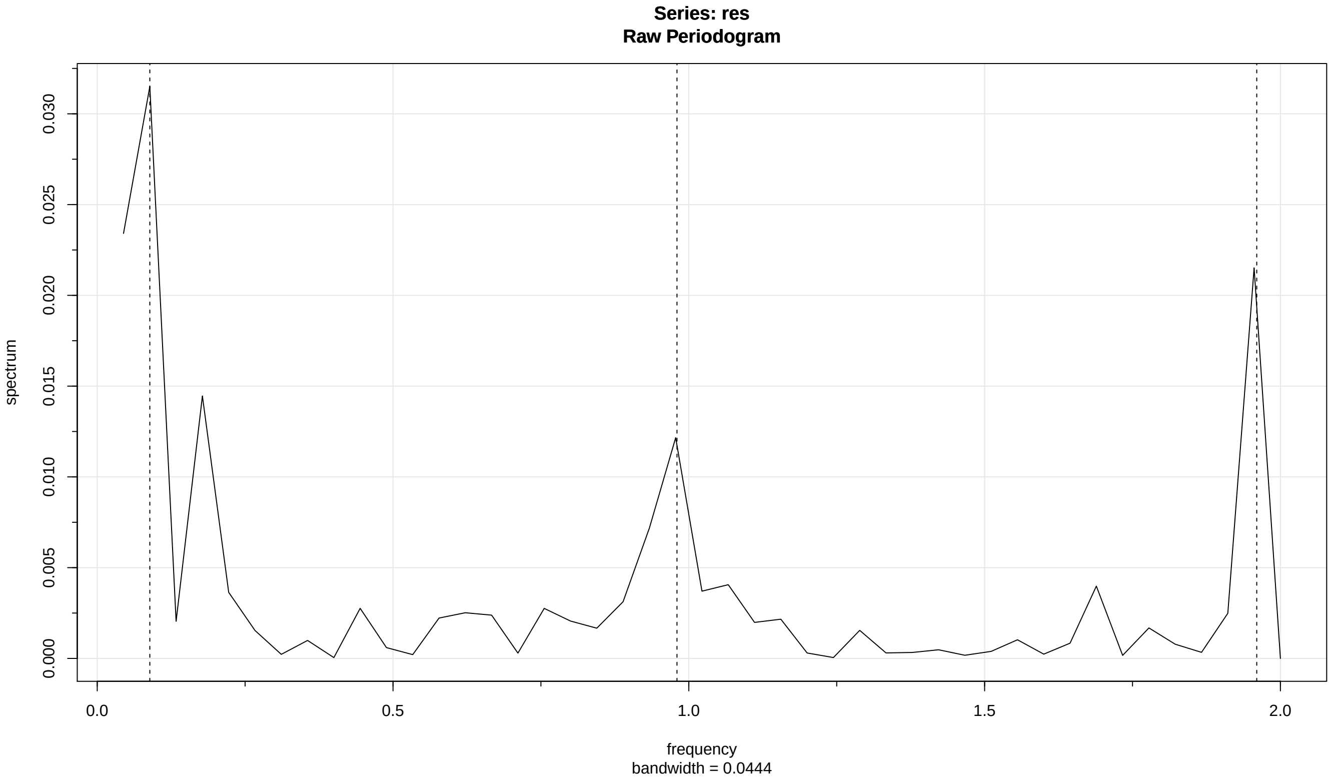

Figure 3: Periodogram of the residuals.

Section 2: Theory Questions

1.(a) We are given that $\sum_{k=1}^{d} S_k = 0$ and $\mathrm{E}(Y_{jk}) = 0$. Therefore,

$$\begin{aligned}
\widehat{m}_j &= \frac{1}{d}\sum_{k=1}^{d} X_{jk} \\
&= \frac{1}{d}\sum_{k=1}^{d}(m_j + S_k + Y_{jk}) \\
&= m_j + \frac{1}{d}\sum_{k=1}^{d} S_k + \frac{1}{d}\sum_{k=1}^{d} Y_{jk}.
\end{aligned}$$

Then by taking expectations on both sides:

$$\begin{aligned}
\mathrm{E}(\widehat{m}_j) &= m_j + 0 + \frac{1}{d}\sum_{k=1}^{d}\mathrm{E}(Y_{jk}) \\
&= m_j,
\end{aligned}$$

since $\sum_{k=1}^{d} S_k = 0$ and $\mathrm{E}(Y_{jk}) = 0$. Hence $\widehat{m}_j$ is an unbiased estimator of m_j.

(b) To show that the seasonal component estimator satisfies the model assumptions, we need to show that $\sum_{k=1}^{d} \widehat{S}_k = 0$. To see why this is true, note that

$$\begin{aligned}
\sum_{k=1}^{d} \widehat{S}_k &= \frac{1}{b}\sum_{k=1}^{d}\sum_{j=1}^{b}(X_{jk} - \widehat{m}_j) \\
&= \frac{1}{b}\sum_{j=1}^{b}\left(\sum_{k=1}^{d} X_{jk} - d\widehat{m}_j\right) \\
&= \frac{1}{b}\sum_{j=1}^{b}(d\widehat{m}_j - d\widehat{m}_j) = 0.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\mathrm{E}(\widehat{S}_k) &= \frac{1}{b}\sum_{j=1}^{b}\left[\mathrm{E}(X_{jk}) - \mathrm{E}(\widehat{m}_j)\right] \\
&= \frac{1}{b}\sum_{j=1}^{b}(m_j + S_k - m_j) = S_k,
\end{aligned}$$

where the last inequality follows because $\mathrm{E}(X_{jk}) = m_j + S_k$ and $\mathrm{E}(\widehat{m}_j) = m_j$ as established above. Thus we conclude that $\widehat{S}_k$ is an unbiased estimator of S_k.

2(a) We are given that $S_t = S_{t-12}$ and $m_t = \beta_0 + \beta_1 t$. Then

$$\begin{aligned}
\nabla_{12} X_t &= X_t - X_{t-12} \\
&= m_t + S_t + Y_t - m_{t-12} - S_{t-12} - Y_{t-12} \\
&= \beta_0 + \beta_1 t - \beta_0 - \beta_1(t-12) + Y_t - Y_{t-12} \text{ because } S_t = S_{t-12} \\
&= 12\beta_1 + \nabla_{12} Y_t.
\end{aligned}$$

We then need to check that $\nabla_{12} X_t$ is weakly stationary. The mean function is

$$\begin{aligned}
\mathrm{E}(\nabla_{12} X_t) &= 12\beta_1 + \mathrm{E}(Y_t) - \mathrm{E}(Y_{t-12}) \\
&= 12\beta_1, \mathbf{(2)}
\end{aligned}$$

since $\mathrm{E}(Y_t) = 0$ for all t because $\{Y_t\}$ is a weakly stationary mean zero process. Then for any integer h,

$$\begin{aligned}
\mathrm{Cov}(\nabla_{12} X_{t+h}, \nabla_{12} X_t) &= \mathrm{Cov}(12\beta_1 + \nabla_{12} Y_{t+h}, 12\beta_1 + \nabla_{12} Y_t) \\
&= \mathrm{Cov}(\nabla_{12} Y_{t+h}, \nabla_{12} Y_t) \\
&= \mathrm{Cov}(Y_{t+h} - Y_{t+h-12}, Y_t - Y_{t-12}) \\
&= \mathrm{Cov}(Y_{t+h}, Y_t) - \mathrm{Cov}(Y_{t+h}, Y_{t-12}) - \mathrm{Cov}(Y_{t+h-12}, Y_t) \\
&\quad + \mathrm{Cov}(Y_{t+h-12}, Y_{t-12}) \\
&= 2\gamma(h) - \gamma(h+12) - \gamma(h-12). \mathbf{(2)}
\end{aligned}$$

Since the mean and autocovariance functions of $\nabla_{12} X_t$ do not depend on t, we conclude that $\nabla_{12} X_t$ is weakly stationary.

(b) For the mixed model $X_t = m_t S_t + Y_t$, we have

$$\begin{aligned}
\nabla_{12} X_t &= X_t - X_{t-12} \\
&= m_t S_t + Y_t - m_{t-12} S_{t-12} - Y_{t-12} \\
&= (m_t - m_{t-12}) S_t + \nabla_{12} Y_t \qquad \text{since } S_t = S_{t-12} \\
&= [\beta_0 + \beta_1 t - \beta_0 - \beta_1(t-12)] S_t + \nabla_{12} Y_t \\
&= 12\beta_1 S_t + \nabla_{12} Y_t.
\end{aligned}$$

Let us look at the mean function

$$\begin{aligned}
\mathrm{E}(\nabla_{12} X_t) &= 12\beta_1 S_t + \mathrm{E}(Y_t) - \mathrm{E}(Y_{t-12}) \\
&= 12\beta_1 S_t.
\end{aligned}$$

However since the mean function depends on t, we conclude that $\nabla_{12} X_t$ is not weakly stationary. Now using again $S_t = S_{t-12}$, it is possible to get rid of S_t by doing ∇_{12} another time to yield

$$\begin{aligned}\nabla_{12}^2 X_t = \nabla_{12}(\nabla_{12}X_t) &= \nabla_{12}(12\beta_1 S_t + \nabla_{12}Y_t)\\ &= 12\beta_1 S_t + Y_t - Y_{t-12} - 12\beta_1(S_{t-12}) - Y_{t-12} + Y_{t-24}\\ &= Y_t - 2Y_{t-12} + Y_{t-24}\\ &= \nabla_{12}^2 Y_t.\mathbf{(3)}\end{aligned}$$

In this case, we have $\mathrm{E}(\nabla_{12}^2 X_t) = \mathrm{E}(\nabla_{12}Y_t) = \mathrm{E}(Y_t) - 2\mathrm{E}(Y_{t-12}) + \mathrm{E}(Y_{t-24}) = 0$. For any integer h, the autocovariance function is

$$\begin{aligned}\mathrm{Cov}(\nabla_{12}^2 X_{t+h}, \nabla_{12}^2 X_t) &= \mathrm{Cov}(\nabla_{12}^2 Y_{t+h}, \nabla_{12}^2 Y_t)\\ &= \mathrm{Cov}(Y_{t+h} - 2Y_{t+h-12} + Y_{t+h-24}, Y_t - 2Y_{t-12} + Y_{t-24})\\ &= \mathrm{Cov}(Y_{t+h}, Y_t) - 2\mathrm{Cov}(Y_{t+h}, Y_{t-12}) + \mathrm{Cov}(Y_{t+h}, Y_{t-24})\\ &\quad - 2\mathrm{Cov}(Y_{t+h-12}, Y_t) + 4\mathrm{Cov}(Y_{t+h-12}, Y_{t-12}) - 2\mathrm{Cov}(Y_{t+h-12}, Y_{t-24})\\ &\quad + \mathrm{Cov}(Y_{t+h-24}, Y_t) - 2\mathrm{Cov}(Y_{t+h-24}, Y_{t-12}) + \mathrm{Cov}(Y_{t+h-24}, Y_{t-24})\\ &= 6\gamma(h) - 4\gamma(h+12) - 4\gamma(h-12) + \gamma(h+24) + \gamma(h-24).\mathbf{(3)}\end{aligned}$$

Since the mean and autocovariance functions of $\nabla_{12}^2 X_t$ do not depend on t, we conclude that $\nabla_{12}^2 X_t$ is a weakly stationary process.

3. By the definition of a white noise process, we have $\mathrm{E}(X_t) = 0$ for all t. Then if $h = 0$, we have $\mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Var}(X_t) = \sigma^2$. However if $h \neq 0$, we have $\mathrm{Cov}(X_{t+h}, X_t) = 0$ since white noise is a sequence of uncorrelated random variables. Therefore since the mean and autocovariance functions do not depend on t, we conclude that white noise is a weakly stationary process. For its autocorrelation function, note that $\rho(0) = \gamma(0)/\gamma(0) = 1$ by definition and $\rho(h) = \gamma(h)/\gamma(0) = 0$ for $h \neq 0$ because we showed above that $\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t) = 0$ for any $h \neq 0$. Hence,

$$\rho(h) = \begin{cases} 1, & \text{if } h = 0,\\ 0, & \text{if } h \neq 0.\end{cases}$$

4. By repeated substitutions:

$$\begin{aligned}X_t &= X_{t-1} + Z_t\\ &= X_{t-2} + Z_{t-1} + Z_t\\ &= X_{t-3} + Z_{t-2} + Z_{t-1} + Z_t\\ &= \vdots\\ &= X_0 + Z_1 + Z_2 + \cdots + Z_{t-2} + Z_{t-1} + Z_t\\ &= 5 + \sum_{i=1}^{t} Z_i.\end{aligned}$$

Then the mean function is

$$\mathrm{E}(X_t) = 5 + \sum_{i=1}^{t} \mathrm{E}(Z_i) = 5, \mathbf{(3)}$$

and the variance function is

$$\mathrm{Var}(X_t) = \sum_{i=1}^{t} \mathrm{Var}(Z_i) = t\sigma^2. \mathbf{(2)}$$

Therefore random walk processes are not weakly stationary. Let us compute the covariance function. For $a, b \in \mathbb{N}$,

$$\begin{aligned}
\mathrm{Cov}(X_a, X_b) &= \mathrm{Cov}\left(5 + \sum_{i=1}^{a} Z_i, \quad 5 + \sum_{i=1}^{b} Z_i\right) \\
&= \mathrm{Cov}\left(\sum_{i=1}^{a} Z_i, \quad \sum_{i=1}^{b} Z_i\right)
\end{aligned}$$

Suppose $a < b$, then

$$\begin{aligned}
\mathrm{Cov}(X_a, X_b) &= \mathrm{Cov}\left(\sum_{i=1}^{a} Z_i, \quad \sum_{i=1}^{a} Z_i + \sum_{i=a+1}^{b} Z_i\right) \\
&= \mathrm{Cov}\left(\sum_{i=1}^{a} Z_i, \sum_{i=1}^{a} Z_i\right) + \mathrm{Cov}\left(\sum_{i=1}^{a} Z_i, \sum_{i=a+1}^{b} Z_i\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^{a} Z_i\right) = \sum_{i=1}^{a} \mathrm{Var}(Z_i) = \sigma^2 a.
\end{aligned}$$

For the case $a > b$, we have

$$\begin{aligned}
\mathrm{Cov}(X_a, X_b) &= \mathrm{Cov}\left(\sum_{i=1}^{b} Z_i + \sum_{i=b+1}^{a} Z_i, \quad \sum_{i=1}^{b} Z_i\right) \\
&= \mathrm{Cov}\left(\sum_{i=1}^{b} Z_i, \sum_{i=1}^{b} Z_i\right) + \mathrm{Cov}\left(\sum_{i=b+1}^{a} Z_i, \sum_{i=1}^{b} Z_i\right) \\
&= \mathrm{Var}\left(\sum_{i=1}^{b} Z_i\right) = \sum_{i=1}^{b} \mathrm{Var}(Z_i) = \sigma^2 b.
\end{aligned}$$

Therefore,

$$\mathrm{Cov}(X_a, X_b) = \sigma^2 \min\{a, b\}. \mathbf{(5)}$$