7 Chapter 7: Modelling seasonal ARMA processes. SARIMA

7.1 Example

If a time series contains a seasonal component S_t then the ordinary differencing methods of the previous section will not work. We will illustrate this with an example.

Example. Monthly production of chocolate confectionery in Australia: tonnes. July 1957 - Aug 1995. (see the top plot of Fig. 7.1).

Clearly data show trend, periodic and have non-constant variance. Therefore, we start from the variance-stabilizing transformation, in particular square root (middle) and log (bottom) (see Fig. 7.1).

We notice that both sqrt and log transforms have removed some heteroscedasticity of the data. (We shall not choose the best variance tstabilizing transformation now and will try to test both sqrt and log for a while.) However, we still have very strong seasonality and trend to deal with.

```
> ts.plot(ts(choc, frequency=12, start=c(1957, 7)),
  main="choc")

> ts.plot(ts(sqrt(choc), frequency=12, start=c(1957, 7)),
  main="sqrt(choc)")

> ts.plot(ts(log(choc), frequency=12, start=c(1957, 7)),
  main="log(choc)")
```

We will use here **seasonal differencing**. If Y_t has seasonal period m, then the **first seasonal difference** is defined to be

$$D_{m,t} = Y_t - Y_{t-m}$$

for $t = m+1, m+2, \dots, n$.

In this way we compare seasons with seasons, e.g. we subtract July from July and December from December. Seasonal differencing will (at least approximately) remove the seasonal component (see Fig. 7.2)

```
> sdsqrtchoc<- diff(sqrt(choc), lag=12)
> ts.plot(ts(sdsqrtchoc, frequency=12, start=c(1957, 7)),
main="seasonally differenced sqrt choc",
ylab="sd(sqrt(chocolate production))")
```

```
> sdlogchoc<- diff(log(choc), lag=12)
> ts.plot(ts(sdlogchoc, frequency=12, start=c(1957, 7)),
 main="seasonally differenced log choc",
 ylab="sd(log(chocolate production))")
```

We notice that seasonally differenced sqrt chocolate production are much more heteroscedastic. In contrast, seasonally differenced log chocolate production is almost homoscedastic. Thus, we shall utilize the log transformation.

Fig. 7.3 presents the acf and pacf plots of the seasonally differenced log transformed chocolate data.

Notice that seasonal differencing at lag 12 nearly removes **seasonality** but still there exists some minor **trend** in the data which we can remove by ordinary (classical) differencing, though, of course, it is a subjective opinion. In particular, compare the bottom of Fig. 7.2 and top of Fig. 7.4.

So it might make sense to apply also ordinary (classical) differencing to our data

```
> diff.sd.log.choc<- diff(sdlogchoc)
```

The order in which you apply seasonal and ordinary differencing does not matter. In particular, you can seasonally difference first, then apply ordinary differencing, or you can ordinary difference first and then apply seasonal differencing.

Now we can try to fit various ARMA and ARIMA models to this preprocessed data.

```
> model1<-arima0(log(choc), order = c(5,1,1),
seasonal = list(order=c(4,1,1),
period=12))

> model1

Call: arima0(x = log(choc), order = c(5, 1, 1), seasonal =
list(order = c(4, 1, 1),
    period = 12))

Coefficients:
         ar1     ar2     ar3     ar4      ar5      ma1     sar1     sar2     sar3     sar4
      0.2273  0.1399  0.0475  -0.028  -0.0074  -0.9357  -0.1317  -0.2211  -0.2293  -0.2023
s.e.  0.0005     NaN     NaN     NaN      NaN   0.0031      NaN
NaN      NaN  0.0002
         sma1
      -0.4349
s.e.   0.0002

sigma^2 estimated as 0.01012:  log likelihood = 387.27,  aic =
-750.55 Warning message: NaNs produced in: sqrt(diag(x$var.coef))
```

We skip a few other simpler models because they provide convergence problems.

```
> model2<-arima0(log(choc), order = c(4,1,1),
seasonal = list(order=c(2,1,1), period=12))
> model2

Call: arima0(x = log(choc), order = c(4, 1, 1), seasonal =
list(order = c(2, 1, 1),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ar4      ma1    sar1     sar2     sma1
      0.2214  0.1367  0.0737  -0.0383  -0.9442  0.2201  -0.0476  -0.7767
s.e.  0.0484  0.0525  0.0518   0.0523   0.2603  0.0725   0.0596
0.0545

sigma^2 estimated as 0.01037:  log likelihood = 380.16,
aic = -742.32



> model3<-arima0(log(choc), order = c(3,1,1),
seasonal = list(order=c(2,1,1), period=12))
> model3

Call: arima0(x = log(choc), order = c(3, 1, 1), seasonal =
list(order = c(2, 1, 1),
    period = 12))

Coefficients:
         ar1     ar2     ar3      ma1    sar1     sar2     sma1
      0.2252  0.1381  0.0721  -0.9519  0.2244  -0.0483  -0.7727
s.e.  0.0481  0.0518  0.0514   0.2458  0.0725   0.0598   0.0551

sigma^2 estimated as 0.01038:  log likelihood = 379.89,
aic = -743.78



> model4<-arima0(log(choc), order = c(2,1,1),
seasonal = list(order=c(2,1,1), period=12))
> model4

Call: arima0(x = log(choc), order = c(2, 1, 1), seasonal =
list(order = c(2, 1, 1),
    period = 12))

Coefficients:
```

```
         ar1     ar2      ma1    sar1     sar2     sma1
      0.2213  0.1405  -0.9374  0.2192  -0.0564  -0.7738
s.e.  0.0476  0.0534   0.2158  0.0709   0.0588   0.0530

sigma^2 estimated as 0.01043:  log likelihood = 378.86,
aic = -743.72
```

We choose the model 3 as the best in terms of AIC.

Now we can predict 2 steps ahead for the log(choc) production

```
> predict(model3, n.ahead=2)
$pred Time Series: Start = 459 End = 460
Frequency = 1 [1] 9.242483 9.247498

$se Time Series: Start = 459 End = 460
Frequency = 1 [1] 0.1018926 0.1056290
```

Remark. Notice that now in order to predict the original chocolate production, you need to take inverse of log.

Finally, we perform our usual diagnostics (see Fig. 7.4):

```
> shapiro.test(model3$residual)

        Shapiro-Wilk normality test

data:  model3$residual W = 0.9904, p-value = 0.005337
```

Such a low p-value of the SW test is likely due to outliers, i.e. one in the left tail and two in the right tail. We can delete them and re-run the SW test again.

```
> k<-order(model3$residual)
> outliers<-c(k[1],
k[length(model3$residuals)-1], k[length(model3$residuals)])

> shapiro.test(model3$residual[-outliers])

        Shapiro-Wilk normality test

data:  model3$residual[-outliers] W = 0.9937, p-value = 0.06395
```

The residual diagnostics is quite appropriate. Hence, we conclude that the model is appropriate for forecasting.

7.2 Formal definition of SARIMA

Suppose we have r years of monthly data which we tabulate as follows:

Each column in this table may itself be viewed as a realization of a time series. Each row corresponds to a certain month and also may be viewed as a realization of a time series.

Suppose that each one of these 12 time series is generated by the same ARMA(P, Q) model. For example, the series corresponding to the j-th month, $X_{j+12t}, t = 0, \ldots, r-1$, satisfies a difference equation of the form

$$\begin{aligned} X_{j+12t} &= \Phi_1 X_{j+12(t-1)} + \ldots + \Phi_P X_{j+12(t-P)} \\ &+ U_{j+12t} + \Theta_1 U_{j+12(t-1)} + \ldots + \Theta_1 U_{j+12(t-Q)}, \end{aligned} \tag{7.1}$$

where

$$\{U_{j+12t}, t = \ldots, -1, 0, 1, \ldots\} \sim WN\left(0, \sigma_U^2\right).$$

For example, in the case of May this equation is rewritten for each year t as

$$\begin{aligned} X_{5+12t} &= \Phi_1 X_{5+12}(t-1) + \ldots + \Phi_P X_{5+12(t-P)} \\ &+ U_{5+12t} + \Theta_1 U_{5+12(t-1)} + \ldots + \Theta_1 U_{5+12(t-Q)}. \end{aligned} \tag{7.2}$$

However, since the same(!) ARMA(P, Q) model is assumed to apply to each month j, this equation may be rewritten for all years t as

$$\begin{aligned} X_t &= \Phi_1 X_{t-12} + \ldots + \Phi_P X_{t-12P} + U_t \\ &+ \Theta_1 U_{t-12} + \ldots + \Theta_Q U_{t-12Q}, \end{aligned} \tag{7.3}$$

which holds for each month $j = 1, \ldots, 12$.

We can rewrite this equation in a more compact form

$$\Phi(B^{12}) X_t = \Theta(B^{12}) U_t,$$

where $\Phi(z) = 1 - \Phi_1 z - \ldots - \Phi_p z^p$, $\Theta(z) = 1 + \Theta_1 z + \ldots + \Theta_Q z^Q$ and $U_t \sim WN\left(0, \sigma_U^2\right)$. We refer to this model as **the between-year model**.

Notice however that $E\left(U_t, U_{t+h}\right)$ is not necessarily 0 except when h is an integer multiple of 12. Indeed, it is unlikely that the 12 series corresponding to the different months are uncorrelated. To incorporate dependence between these series we assume that the $\{U_t\}$ sequence follows an ARMA(p, q) model

$$\phi(B) U_t = \theta(B) \epsilon_t, \qquad \{\epsilon_t\} \sim WN\left(0, \sigma^2\right).$$

This is **the between-months model**.

This assumption not only implies possible non-zero correlation between consecutive values of U_t , but also within the 12 sequences $\{U_{j=12t}, t = \ldots, -1, 0, 1, \ldots\}$, each of which was previously assumed to be uncorrelated. In this case the previous assumption on WN for U_t is no longer valid. However, it is usually the case that $E\left(U_t, U_{t+12j}\right)$ is small for $j = \pm 1, \pm 2, \ldots$.

Now combining the between-year and between-seasons equations and allowing for differencing leads us to the definition of the general seasonal multiplicative ARIMA process.

Definition The **SARIMA(p, d, q) × $(P, D, Q)_s$ process,
or SARIMA(p, P, d, D, q, Q).**

If $d, D \geq 0$, then $\{X_t\}$ is said to be a seasonal ARIMA$(p, d, q) \times (P, D, Q)_s$ process with period s if the differenced process $Y_t = (1-B)^d(1-B^s)^D X_t$ is a causal ARMA process

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\epsilon_t, \;\; \epsilon_t \sim WN\left(0, \sigma^2\right),$$

where $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p, \Phi(z) = 1 - \Phi_1 z - \ldots - \Phi_p z^P, \theta(z) = 1 + \theta_1 z + \ldots + \theta_Q z^q, \Theta(z) = 1 + \Theta_1 z + \ldots + \Theta_Q z^Q$.

Example. In our chocolate production example, we get the SARIMA$(3,1,1) \times (2,1,1)_{12}$ model

$$\phi(B)\Phi(B^{12})\big(1-B\big)\left(1-B^{12}\right)\log(\text{choc}) \quad = \quad \theta(B)\Theta(B^{12})\epsilon_t, \tag{7.4}$$

where $\epsilon_t \sim WN\left(0, \sigma^2\right)$.

The process $\{Y_t\}$ is causal if $\phi(z) \neq 0$ and $\Phi(z) \neq 0$ for $|z| \leq 1$. In applications, D is rarely more than 1 and P and Q are typically less than 3. Note that because of the basic algebra of the operators, it does not matter if the order of the various operators is interchanged.

The main steps of identification, estimation and verification in seasonal models are the same as in non-seasonal models. The main difference is that, in examining autocorrelations for both the initial identification and the final verification, particular attention must be paid to the values at or near multiples of the period s. For example, if the estimated autocorrelation $\hat{\gamma}_s$ is large but $\hat{\gamma}_{ks}$ is small for $k > 1$, this might be taken as an indication that $Q = 1$. The sample pacf coefficients at multiples of s are used in a similar way for the initial identification of P.

7.3 Periodically correlated processes

SARIMA models are by no means the last word on seasonal data. It should be noted that any seasonal ARMA model is simply a special case of a nonseasonal ARMA model, since both the autoregressive and moving average operators may be expanded out as ordinary (nonseasonal) operators. These models do not allow for seasonal variability in the covariances of the process, but in many practical applications, such variability may be observed from simple plots of the variances and low-order autocorrelations as a function of the time within the cycle (e.g. month of the year in the case of monthly data).

A simple example of a periodically correlated process is the PAR(1) model

$$X_{kM+m} = \phi^{(m)} X_{kM+m-1} + \sigma^{(m)} Z_{kM+m}, \quad 1 \leq m \leq M, \;\; k \geq 0, \tag{7.5}$$

in which there are $2M$ parameters $\phi^{(1)}, \ldots, \phi^{(M)}$, $\sigma^{(1)}, \ldots, \sigma^{(M)}$ and $\{Z_t\}$ is a white noise process with variance 1. The stationarity condition for this model is

$$\prod_{m=1}^{M} \left|\phi^{(m)}\right| < 1. \tag{7.6}$$

A simple extension of this is to allow the $\{Z_t\}$ process in (7.5) to be itself a stationary ARMA process, instead of just white noise. In that case the process is called PARMA.

One can of course think about more general extensions than this, e.g. allowing higher-order PAR terms and also introducing periodic MA terms, but the PARMA model just outlined is already quite complicated and probably good enough for most practical purposes.

Periodically correlated models can be considered as an alternative to the seasonal Box-Jenkins approach. In particular, if just the variances are seasonally dependent (i.e. we write $X_{kM+m} = \sigma_m Z_{kM+m}$ with $\{Z_t\}$ stationary), this may be detected by calculating sample standard deviations for each period m, and if these do appear to be non-constant, dividing through by the estimated σ_m values before fitting a stationary model to $\{Z_t\}$.

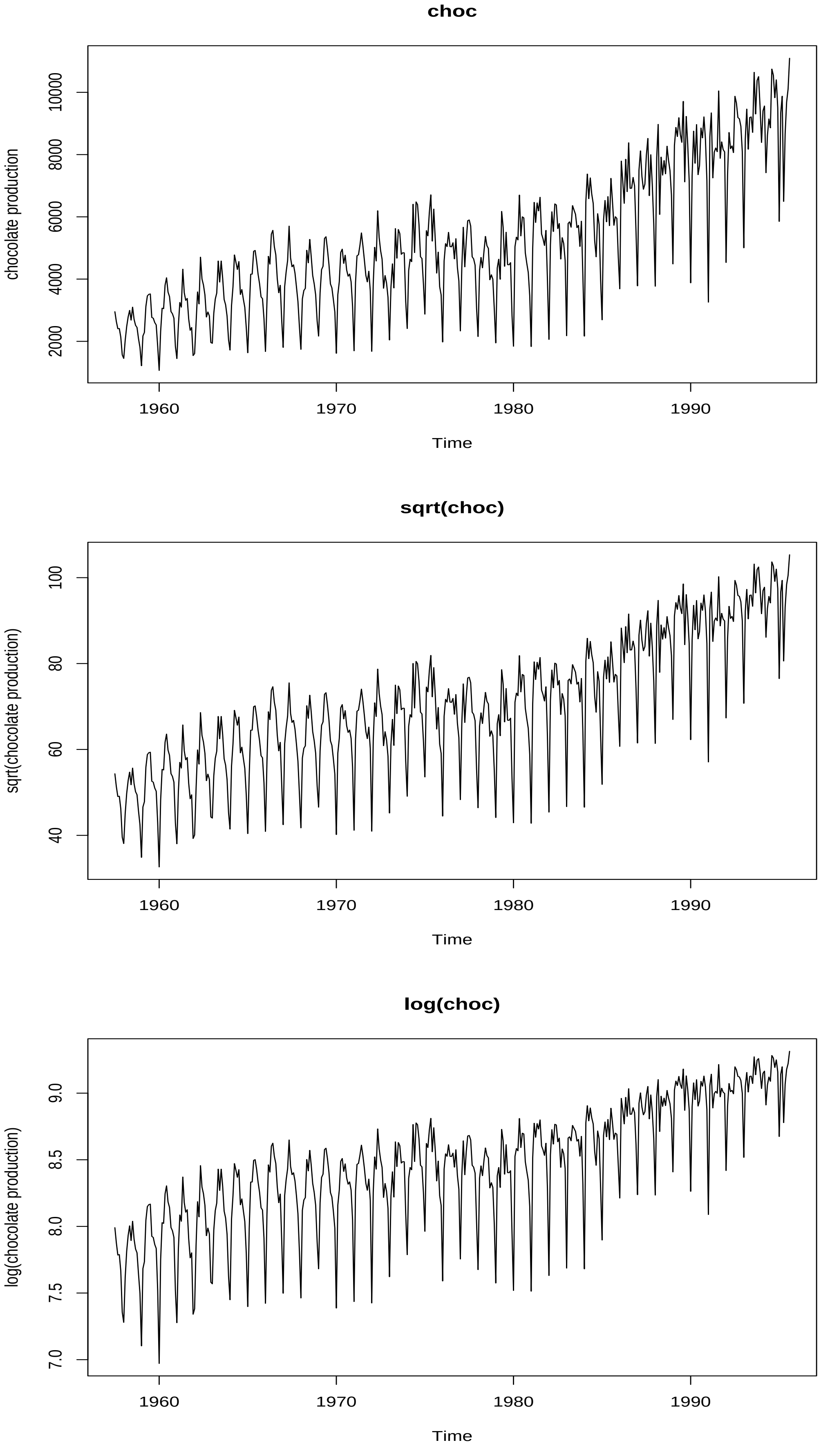
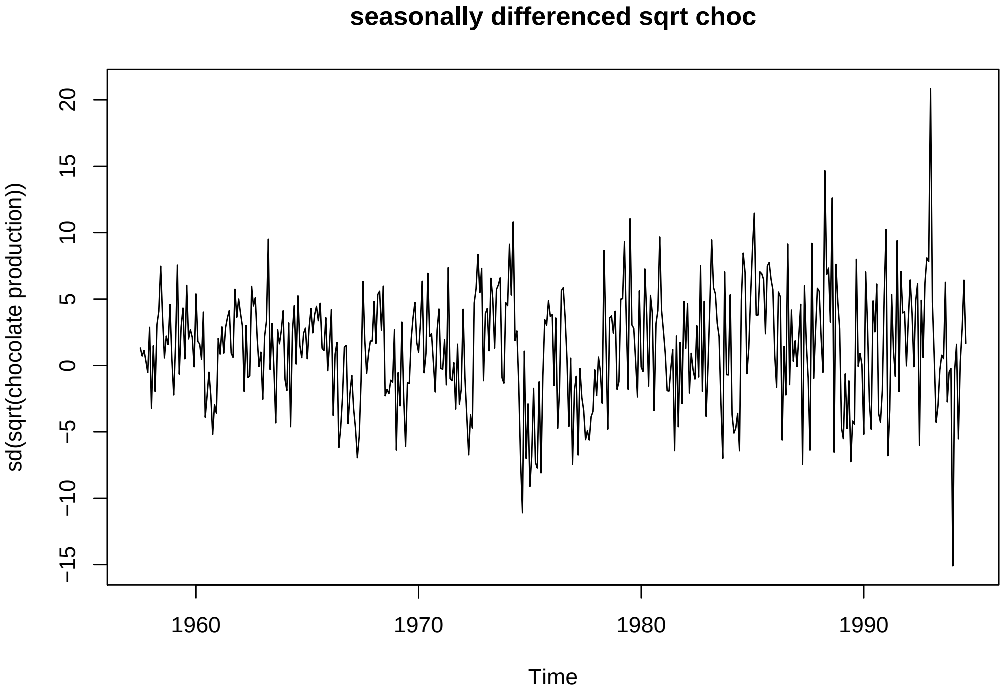


Figure 7.1: Chocolate production in Australia, 1957 - 1995.



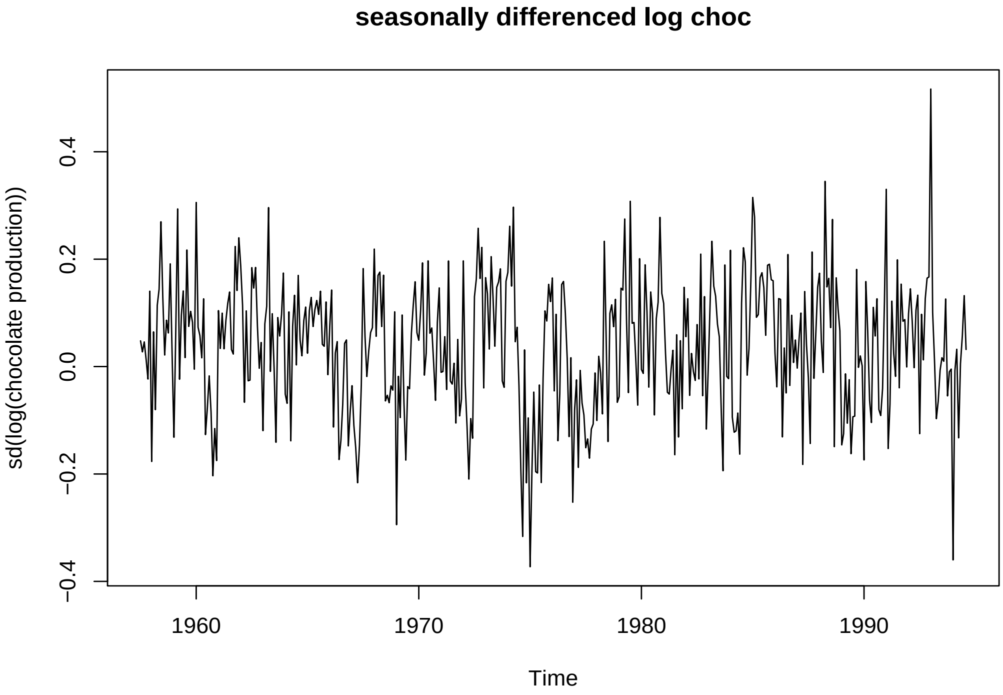


Figure 7.2: Seasonally differenced sqrt and log chocolate production in Australia, 1957 - 1995.

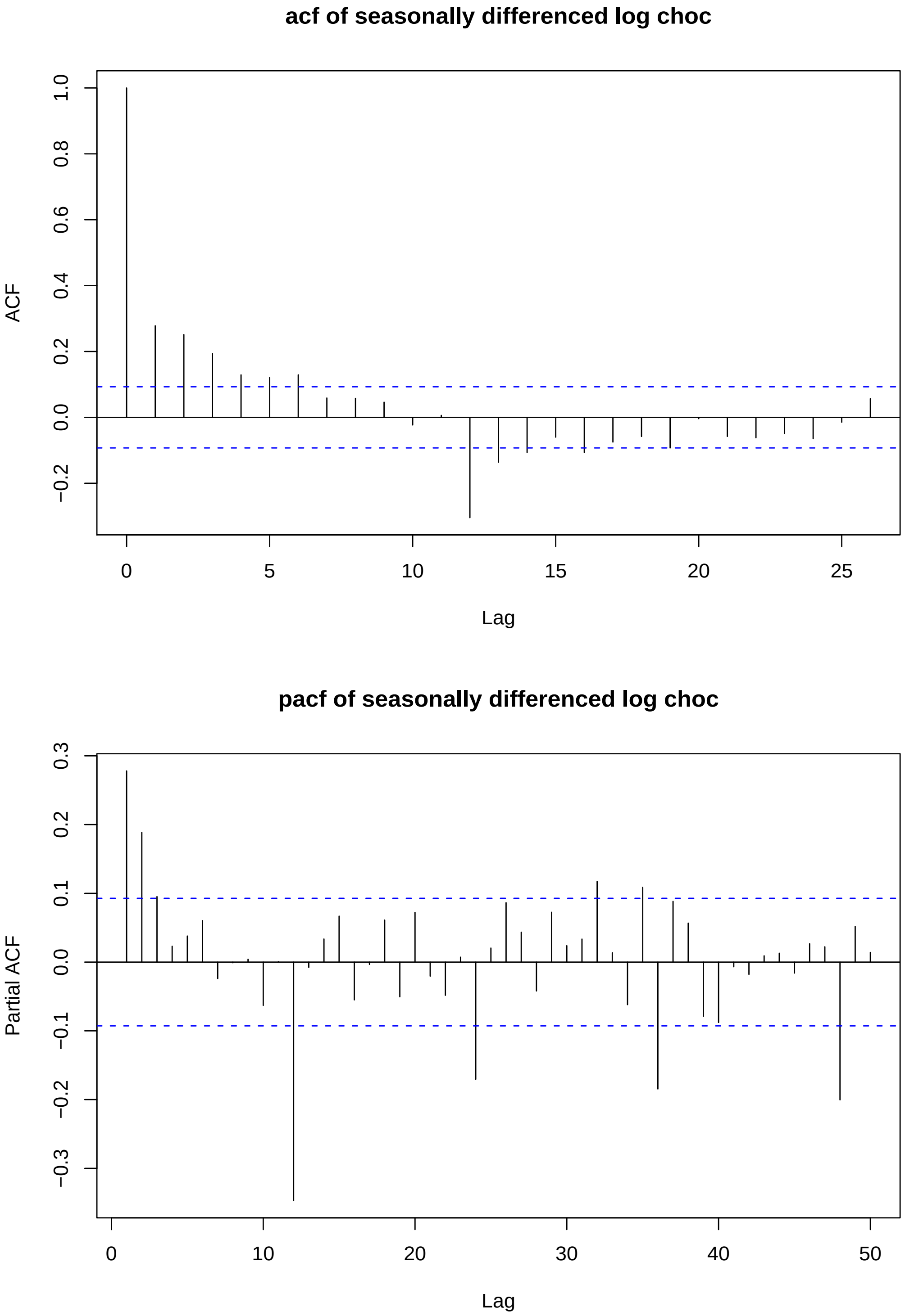


Figure 7.3: Acf and pacf plots of seasonally differenced log chocolate production in Australia, 1957 - 1995.

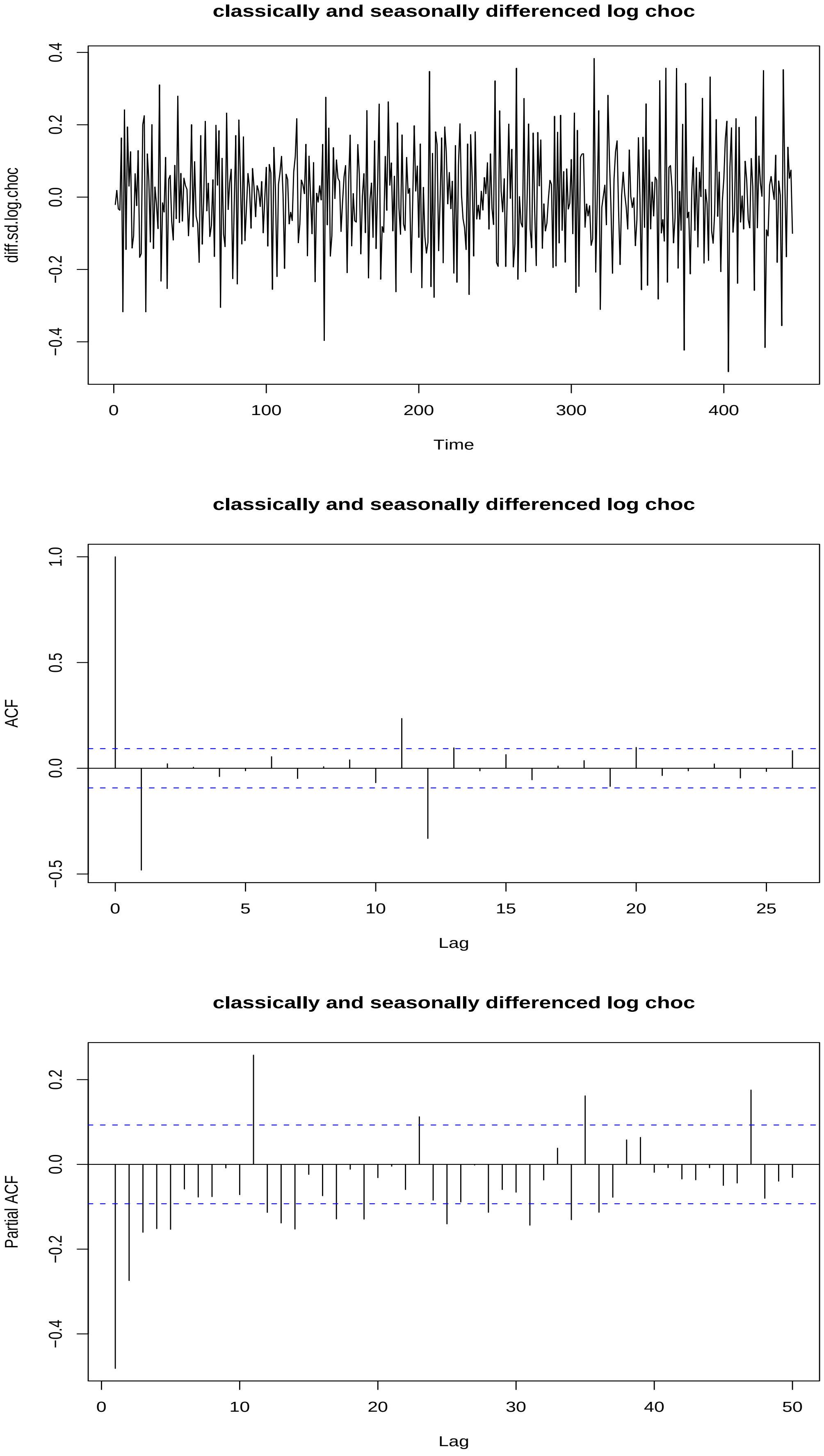


Figure 7.4: Time series, acf and pacf plots of classically and seasonally differenced log chocolate production in Australia, 1957 - 1995.

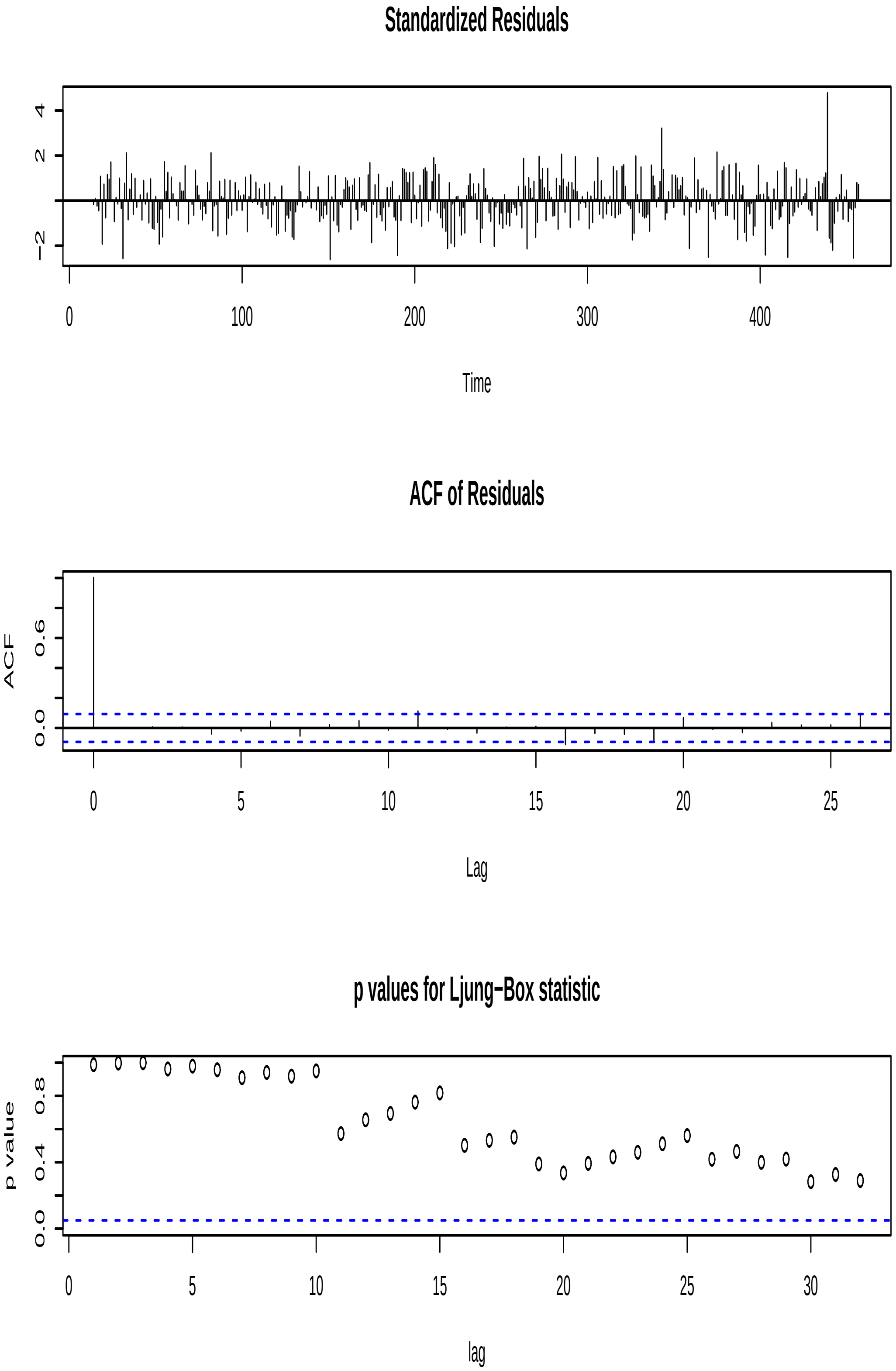


Figure 7.5: Residual diagnostics.

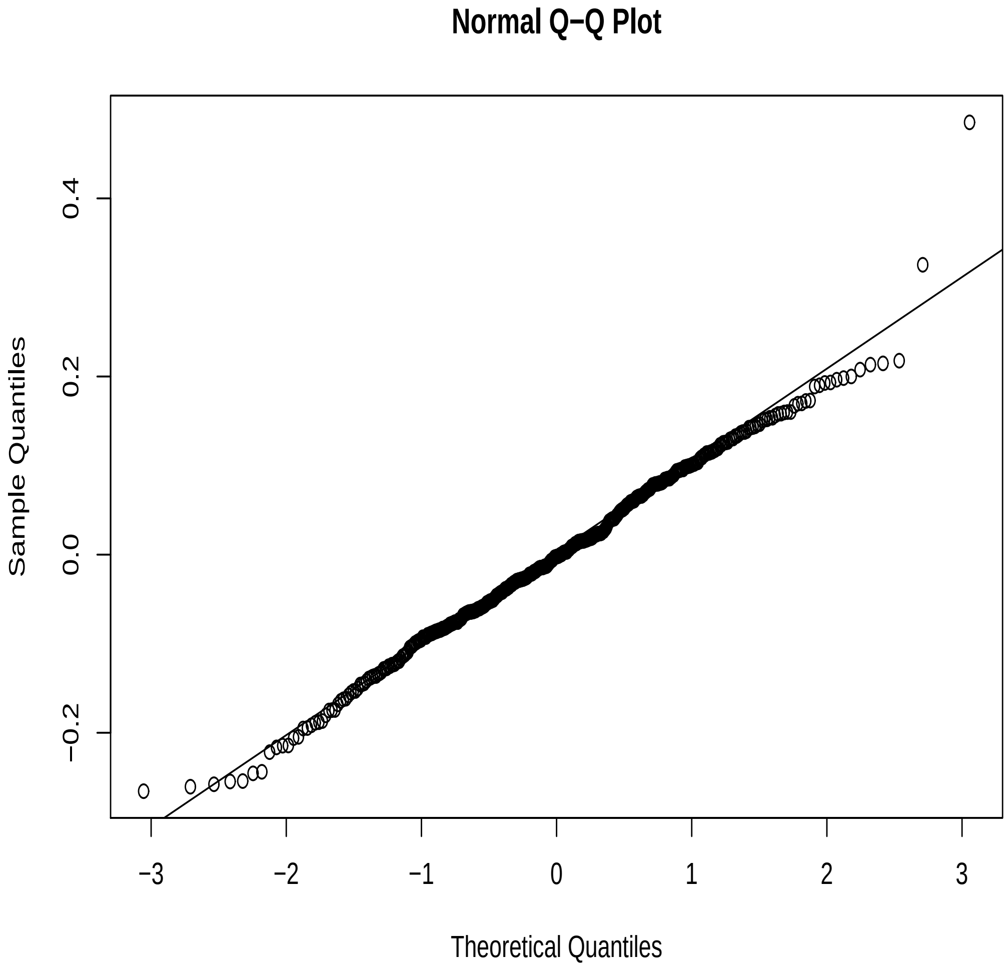


Figure 7.6: QQ plot of the residuals.