4 Chapter 5: Verification diagnostics of ARMA models

The third step of the Box-Jenkins cycle is to confirm that the model in fact fits the data. There are two basic techniques here:

- Overfitting: i.e. add extra parameters to the model and use likelihood ratio or t-tests to check that these are not significant. (This method can be considered as another insight to model selection.)
- Residual analysis: calculate residuals from the fitted model and plot their residuals and their a.c.f., p.a.c.f., spectral density estimates, etc., to check that they are consistent with white noise.

Recall the definition of white noise: $\varepsilon_t \sim WN(0, \sigma^2)$ if $E(\varepsilon_t) = 0$ and $\mathrm{Cov}(\varepsilon_\mathrm{s}, \varepsilon_\mathrm{t}) = \delta_{\mathrm{s,t}}\sigma^2$, where δ_{ts} is the Kronecker symbol, i.e. covariance between ε_t and ε_s is σ^2, if $s = t$, and is 0 otherwise.

Thus, we need to check that residuals ε_t are:

- uncorrelated,
- homoscedastic,

  As typically we also would like to construct confidence and predictive intervals, i.e. CI and PI, which are based on the normal critical values, we also assume that the residuals are

- normally distributed.

Note that if the residuals ε_t are **i.i.d.** $N(0, \sigma^2)$ then all three above properties automatically satisfied. In this case, ε_t are not only uncorrelated, ε_t are independent. To be independent is much stronger property than to be uncorrelated.

While a given model may still have useful predictive value even when those three assumptions are violated, the confidence intervals, prediction intervals, and p-values associated with the t-statistics will generally be incorrect when those three assumptions do not hold.

In a residual analysis we attempt to assess the validity of the assumptions by examining the estimated residuals ε_t to see if they actually behave like i.i.d. $N(0, \sigma^2)$ random variables.

We perform our diagnostics analysis from a step-by-step verification of each assumption and illustrate it by application to residuals from ARMA(1,1) model for twice differenced **BJsales**

data. We shall begin again from the "by-eye" diagnostics and proceed further with the formal definitions and theoretical background.

4.1 Diagnostics "by eye"

A basic technique for investigating the aptness of a model is based on analyzing the residuals ε_t. If the model is apt, the observed residuals should reflect the three assumptions. A lot of useful diagnostic information may be obtained from a residual plot.

What may you notice from the residual plot?

- Change of variability with time indicates heterogeneity of variance of ε_t, i.e. heteroscedasticity.

- Systematic trends in the residuals can suggest correlations between the residuals or inadequateness of the proposed linear model.

- Obvious lack of symmetry (around 0) in the plot suggests possible lack of normality, or perhaps the presence of outliers.

Overall, there should be NO patterns in the residuals, ε_t should look as random as possible.

4.2 Assessing homoscedasticity

We plot the residuals ε_t vs. time. If the assumption of constant variance is satisfied, ε_t should fluctuate around the zero mean with more or less constant amplitude and this amplitude does not change with time. However, sometimes it is hard to assess homoscedasticity by eye. In this case, we can split residuals ε_t into a few groups and run the formal test for homogeneity of variances among the groups.

Consider k random samples $x_{i1}, x_{i2}, \ldots, x_{it_i}$ from the ith population with unknown mean μ_i, variance σ_i^2 and distribution function $F(\cdot)$, $i = 1, 2, \ldots, k$. Here $T = \sum_{i=1}^{k} t_i$ is the sample size and k is the number of groups which we divided our overall dataset to.

The null hypothesis is

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2,$$

vs. the alternative hypothesis

$$H_1 : \sigma_i^2 \neq \sigma_j^2, \text{ for at least one i} \neq \text{j}.$$

Available Tests for Homogeneity of Variances

- **Bartlett** (1938) proposed a test statistic $B = M/C$ for normal data, where

$$M = (T-k)\ln\left(\sum_{i=1}^{k}\frac{(t_i-1)s_i^2}{(T-k)}\right) - \sum_{i=1}^{k}(t_i-1)\ln(s_i^2), \tag{4.1}$$

  $T = \sum_{i=1}^{k} t_i$, $C = 1 + \left\{\left[\sum_{i=1}^{k}(t_i-1)^{-1}\right] - (T-1)^{-1}\right\}/\{3(k-1)\}$.

  The test statistic B is χ^2-distributed with k degrees of freedom. The statistic B is not robust to non-normality and more useful as a test for normality test rather than as a test for equality of variances (Box, 1953).

- **Box and Anderson** (1955) noticed that the effect of normality depends on the kurtosis $\kappa = \mu_4/\mu_2^2$, where μ_j is the j-th population moment. Let all data come from the same distribution F (not necessarily normal). Then **the Box-Anderson statistic** is

$$B_3 = \frac{2M}{(\kappa-1)}, \tag{4.2}$$

  where the sample estimate of κ is

$$\hat{\kappa} = \frac{T\sum_{i=1}^{k}\sum_{j=1}^{t_i}(x_{ij}-\bar{x_{i.}})^4}{\sum_{i=1}^{k}\sum_{j=1}^{t_i}(x_{ij}-\bar{x_{i.}})^2} \tag{4.3}$$

  and M is the same as in **Bartlett's** B.

$$B_3 \sim \chi_{k-1}^2.$$

  In small samples $\hat{\mu}_4$ is quite variable. Hence, a test without $\hat{\mu}_4$ is often preferred.

- Suppose that group means μ_i are known. Then consider a smooth monotone function $G(|x_{ij}-\mu_i|)$, e.g. $(x_{ij}-\mu_i)^2$ and $|x_{ij}-\mu_i|$.

  Notice that $E(x_{ij}-\mu_i)^2 = \sigma_i^2$ while $E|x_{ij}-\mu_i|$ is also a measure of spread due to Pietra (Gastwirth, 1972) and related to a classical measure of income inequality.

  Thus, if we knew μ_i we could apply the standard ANOVA statistic to $(x_{ij}-\mu_i)^2$ or $|x_{ij}-\mu_i|$.

  Levene (1960) substituted μ_i by $\bar{x}_{i.} = \frac{1}{t_i}\sum_{j=1}^{t_i} x_{ij}$, treated $G(|x_{ij}-\bar{x}_{i.}|)$ as i.i.d. normal and then applied the classical ANOVA, or the F-test. Let $d_{ij} = |x_{ij}-\bar{x}_{i.}|$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, t_i$. Then **Levene's statistic** is

$$L = \frac{T-k}{k-1}\frac{\sum_{i=1}^{k}(\bar{d}_{i.}-\bar{d}_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{t_i}(d_{ij}-\bar{d}_{i.})^2}, \tag{4.4}$$

$$\bar{d}_{i\cdot} = \sum_{j=1}^{t_j} d_{ij}/t_j, \;\; \bar{d}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{t_j} d_{ij}/T.$$

Though d_{ij} are not $N(\mu_i, \sigma_i^2)$, the within group correlation is $O(1/n_i^2)$ while ANOVA is robust to non-normality.

Thus, L is F-distributed with $k-1$ and $T-k$ degrees of freedom.

The Levene's test is one of the most powerful and robust tests for homoscedasticity. Levene's test became so popular in many different applications that it gave birth to a whole set of modifications called a family of Levene's tests. One of such modifications is to substitute group means (classical measure of center) by robust group centers, i.e. group medians (Brown and Forsythe, 1974), or the trimmed mean (Lim and Loh, 1996).

```
###### time series plot with group splitting######
> ts.plot(arma11ML$residuals)
> abline(v=50, lwd=4)
> abline(v=110, lwd=4)
```

```
###### create a group factor, i.e. 1 indicates the 1st group########
###### 2 indicates the 2nd group, 3 indicated the 3d group  ########
> group<-c(rep(1,50), rep(2,60),
rep(3, (length(arma11ML$residuals)-110)))

######check that group and residuals are of the same length########
> length(arma11ML$residuals)
[1] 148
> length(group)
[1] 148

###### Bartlett test ##########
> bartlett.test(arma11ML$residuals,group)

        Bartlett test of homogeneity of variances

data:  arma11ML$residuals and group Bartlett's
```

```
K-squared = 8.1593,
df = 2, p-value = 0.01691

library(lawstat)

####### Levene's test with group centers being group means########
> levene.test(arma11ML$residuals,group)

        Classical Levene's test based on the absolute deviations
        from the mean

data:  arma11ML$residuals Test Statistic = 2.8831,
p-value = 0.05918


####### Levene's test with group centers being group medians########
> levene.test(arma11ML$residuals,group, option="median")

        Modified Robust Brown-Forsythe Levene-type test based on
        the absolute deviations from the median

data:  arma11ML$residuals Test Statistic = 2.8161,
p-value = 0.06312
```

Thus, the assumption of homoscedasticity is satisfied "so-so". This may be due to some outliers though. One can try to divide the sample data into 4 groups and re-run the analysis.

4.3 Checking that residuals are uncorrelated

We can construct acf and pacf plots of ε_t. There should be no significant spikes in both plots. (Of course, it is up to common sense since there might be some sampling errors especially for higher lags.)

Formal tests for Serial Correlation among the residuals:

1. **The Box-Pierce and Box-Ljung tests, or portmanteau tests**

We can test the joint or simultaneous hypothesis

$$H_{0,K} \quad : \quad \rho_1 = \rho_2 = \ldots = \rho_K = 0 \tag{4.5}$$

$$H_{1,K} \quad : \quad \rho_j \neq 0 \ \text{ for some } j \ \epsilon \ \{1, 2, \ldots, K.\} \tag{4.6}$$

There are two tests for the hypothesis (4.5). The first method is **the Box-Pierce test** (also called **the portmanteau test**) that is based on

$$\tilde{Q}_K = T \sum_{k=1}^{K} \rho_k^2,$$

where K is bigger than $p + q$ but much smaller than the sample size T, and ρ_k is the k-th sample autocorrelation of the residuals series.

An alternative method is **the Box-Ljung test** that replaces $\tilde{Q}_K$ by

$$Q_K = T(T+1) \sum_{k=1}^{K} \frac{\rho_k^2}{T-k}.$$

For large T under the null hypothesis $H_{0,K} : \rho_1 = \ldots = \rho_K = 0$, $\tilde{Q}_K$ and Q_K have the χ^2-distribution with $K - p - q$ degrees of freedom. Hence, we can construct tests based on $\tilde{Q}_K$ and Q_K.

The Box-Ljung test statistic Q_k is recommended on the grounds that the distribution of Q is closer to its χ^2_{K-p-q} limit than that of Q.

It is easy to see that small values of Q_K support $H_{0,K}$ while large values of Q_K would lead us to reject $H_{0,K}$. Hence the p-value p_K is computed as

$$p_K = P\left(Q > Q_K | \text{given } Q \sim \chi^2_{K-p-q}\right). \tag{4.7}$$

We can plot p-values p_k against the lag k using the R function **tsdiag**. The 0.05 rejection line is included to the plot. We reject $H_{0,K}$ if p_K falls beneath this line.

```
> arma11ML<-arima0(BJ2, order=c(1, 0, 1))
> tsdiag(arma11ML)
```

Note that for the residuals from the **BJsales** example all p-values are well above 0.05, indicating that there is no evidence against white noise.

2. **Test of Up and Down, or Wald-Wolfowitz runs test for randomness**

We can count the number of times a sequence of observations crossed the cut-off line, for example, the median line, and use this information to assess randomness of ε_t. Alternatively we count a succession of plus or minus signs, surrounded by the opposite signs. Each such a succession is called **a run**.

The formal test is the following:

When a sequence of T observations with n observations in positive runs and m observations in negative runs is a stationary random process with independent values generated from the continuous distribution, then the sampling distribution of the number of runs R has the mean and variance

$$E(R) = \frac{1+2nm}{n+m},\ \sigma^2 = \frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)}.$$

Since $n+m=T$, we can re-write it as

$$E(R) = \frac{1+2nm}{T},\ \sigma^2 = \frac{2nm(2nm-T)}{T^2(T-1)}.$$

Note that the only assumption for this test is that all sample observations come from continuous distribution.

The two-tail alternative is as follows

$$\begin{array}{lcl} H_0 & : & \text{Sequence generated by a random process} \\ H_1 & : & \text{Sequence generated by a process containing either} \\ & & \text{persistence or frequent changes in direction.} \end{array}$$

When positive autocorrelation (or persistence) is present, R will clearly be small. On the other hand, if the process involves frequent changes in direction (negative autocorrelation or antipersistence), R will be large.

When the number of observations is sufficiently large, i.e. $T > 30$, the runs test statistic R is based on the standardized normal test statistic

$$z^* = \frac{R - E(R)}{\sigma(R)}.$$

Here z^* follows approximately a standard normal distribution.

Runs test are very easy to interpret even to non-statistician. Runs test allows to assess only the first order serial correlation in the residuals, i.e. to test whether $\varepsilon \sim AR(1)$. Runs test is typically less powerful than other randomness tests, e.g. χ_2 tests or Bartels test.

```
> library(lawstat)
> runs.test(arma11ML$residuals)

        Runs Test - Two sided

data:  arma11ML$residuals Standardized Runs Statistic = 0.9898,
p-value = 0.3223
```

The p-value for the runs test for the ARMA(1,1) residuals of the twice differenced **BJsales** data is 0.32, which indicates that it is likely that residuals are NOT first order serially correlated.

3. **Bartels test**

Suppose that you have a set of observations

$$528, 348, 264, -20, -167, 575, 410, -4, 430, -122.$$

Then ranking R of this set gives

$$9, 6, 5, 3, 1, 10, 7, 4, 8, 2.$$

The Bartels test statistic (Bartels, 1982, JASA) that observations are random, or more precisely are not first order serially correlated, is based on Von Neumann's ratio of rankings R_i

$$\mathrm{RVN} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{T}-1}(\mathrm{R_i} - \mathrm{R_{i+1}})^2}{\sum_{\mathrm{i}=1}^{\mathrm{T}}(\mathrm{R_i} - \bar{\mathrm{R}})^2}. \tag{4.8}$$

RVN is asymptotically normally distributed. In particular,

$$0.5\sqrt{T}(\mathrm{RVN} - 2) = \mathrm{N}(0, 1). \tag{4.9}$$

Based on (4.9) we can construct the rejection region.

```
> library(lawstat)
> bartels.test(arma11ML$residuals)

        Bartels Test - Two sided

data:  arma11ML$residuals Standardized Bartels Statistic = 0.2583,
RVN Ratio = 2.042, p-value = 0.7962
```

The Bartels test provides the p-value of 0.80 for the ARMA(1,1) residuals of the twice differenced **BJsales**. Thus, there is no evidence that residuals are first order serially correlated.

4.4 Checking that residuals are normally distributed

Graphical methods visualize the distribution using graphs, such as histograms, stem-and-leaf plot, box plot etc. For instance, below (Fig. 4.5) you can see the histogram of the simulated normally distributed data and the ARMA(1,1) residuals with superimposed normal curve with the corresponding mean and standard deviation.

Another very popular graphical method of assessment normality is the quantile-quantile (Q-Q) plot. The Q-Q plot compares the ordered values of a variable with the corresponding ordered values of the normal distribution.

Let X be a random variable having the property that the equation

$$P\left(X \leq x\right) = \alpha \tag{4.10}$$

has a unique solution $x = x_{(\alpha)}$ for each $0 < \alpha < 1$. That is, there exists $x_{(\alpha)}$ such that

$$P\left(X \leq x(\alpha)\right) = \alpha \tag{4.11}$$

and no other value of x satisfies 4.11. Then we will call $x(\alpha)$ the α^{th} *(population) quantile* of X. Note that any normal distribution has this uniqueness property. If we consider a standard normal $Z \sim N(0,1)$, then some well known quantile we get:

- $z_{(.5)} = 0$ (the median)

- $z_{(.05)} = -1.645$ and $z_{(.95)} = 1.645$

- $z_{(.025)} = -1.96$ and $z_{(.975)} = 1.96$

We call 0.25th, 0.5th, 0.75th quantiles the first, the second and the third quartiles correspondingly. The quartiles equally divide our data into 4 parts.

Now suppose $X \sim N(\mu, \sigma^2)$. By standardizing to $Z \sim N(0,1)$ we obtain

$$\alpha = P\left(X \leq x_{(\alpha)}\right) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x_{(\alpha)}-\mu}{\sigma}\right) = P\left(Z \leq \frac{x_{(\alpha)}-\mu}{\sigma}\right). \tag{4.12}$$

But we also have $\alpha = P(Z \leq z_{(\alpha)})$ by definition. Thus from the uniqueness property of the $N(0,1)$ it follows that

$$z_{(\alpha)} = \frac{x_{(\alpha)}-\mu}{\sigma} \quad \text{and hence} \quad x_{(\alpha)} = \sigma z_{(\alpha)} + \mu \tag{4.13}$$

Thus, if X is truly normal, a plot of the quantile of X vs. the quantiles of the standard normal should yield a straight line. A plot of the quantiles of X vs. the quantiles of Z is called a Q-Q plot.

Estimating quantiles from data. Let $X_1, \ldots, X_n$ be a sequence of observations. Ideally $X_1, \ldots, X_n$ should represent i.i.d. observations but we will be happy if preliminary tests indicate that they are homoscedastic and uncorrelated (see previous section). We order them from smallest to largest and indicate this by the notation

$$X_{(1/n)} < X_{(2/n)} < X_{(3/n)} < \ldots < X_{((n-1)/n)} < X_{(n/n)} \tag{4.14}$$

The above ordering has assumed no ties, but ties can be quite common in data, even continuous data, because of rounding. As long as the proportion of ties is small, this method can be used.

Note that the proportion of observations less than or equal to $X_{(k/n)}$ is exactly k/n. Hence $X_{(k/n)}$, called the k^{th} *sample quantile*, is an estimate of the population quantile $x_{(k/n)}$. The sample QQ-plot is obtained by plotting the sample quantiles vs. the quantiles of the standard normal. The *R* function 'qqnorm' produces a normal QQ plot of data and the function 'qqline' adds a line to a normal quantile-quantile plot which passes through the first and third quartiles.

```
> qqnorm(arma11ML$residuals)
> qqline(arma11ML$residuals)
```

The Figures 4.6 shows the Q-Q plot of the NYSE volume data and the simulated normal data with the same mean and standard deviation.

Note that one can standardize X robustly, i.e. subtract median rather mean and divide by robust standard deviation rather than classical standard deviation, and then compare with quantiles of $N(0,1)$. Such plots are called Robust QQ (RQQ) plots and might indicate more clearly the sources of non-normality, e.g. outliers or heavy tails, than the usual QQ plots. Such RQQ plots are implemented in **lawstat** as a function **rqq**.

Numerical test of assessment normality. Although visually appealing, these graphical methods do not provide objective criteria to determine the normality of variables.

One of the most popular numerical methods for assessment normality is the Shapiro-Wilk (SW) test (1965). The SW test is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance. It has been originally constructed by considering the regression of ordered sample values on corresponding expected normal order statistics. The SW statistics lies between 0 and 1. If the SW statistic is close to 1, this indicates normality of the data. The SW statistic requires the sample size T to be between 3 and 5000.

The SW statistic, W, is given by

$$\text{SW} = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum \left(x_i - \bar{x}\right)^2},$$

where $x_{(i)}$ are the ordered sample values ($x_{(1)}$ is the smallest) and the a_i are constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution.

```
> shapiro.test(arma11ML$residuals)

        Shapiro-Wilk normality test

data:  arma11ML$residuals W = 0.9935, p-value = 0.7503
```

The fit at the QQ plot looks satisfactory and the SW test provides the p-value of 0.75. Thus, we conclude that ε are normally distributed.

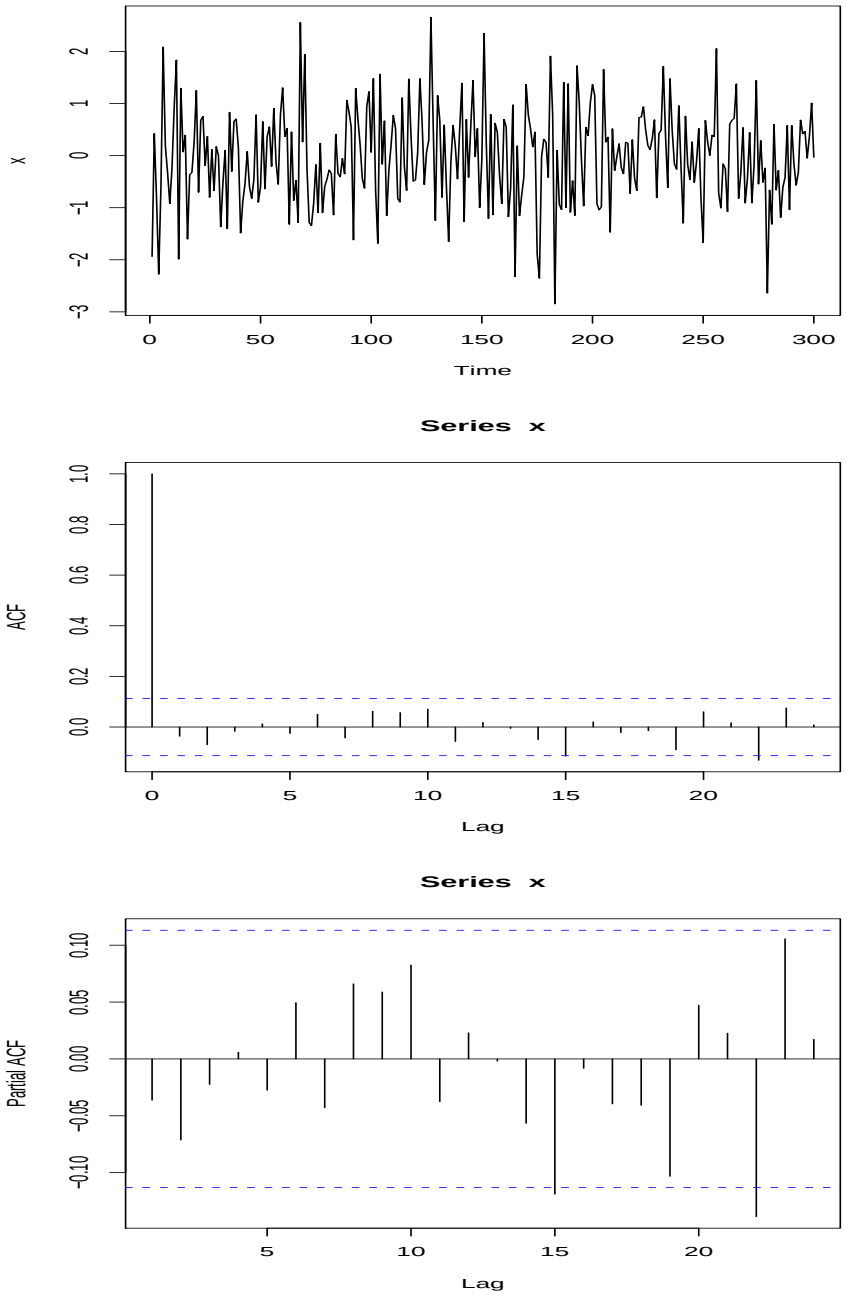

Figure 4.1: The time series, acf and pacf plots of "ideal" residuals. (Simulated 300 observations from $N(0,1)$)

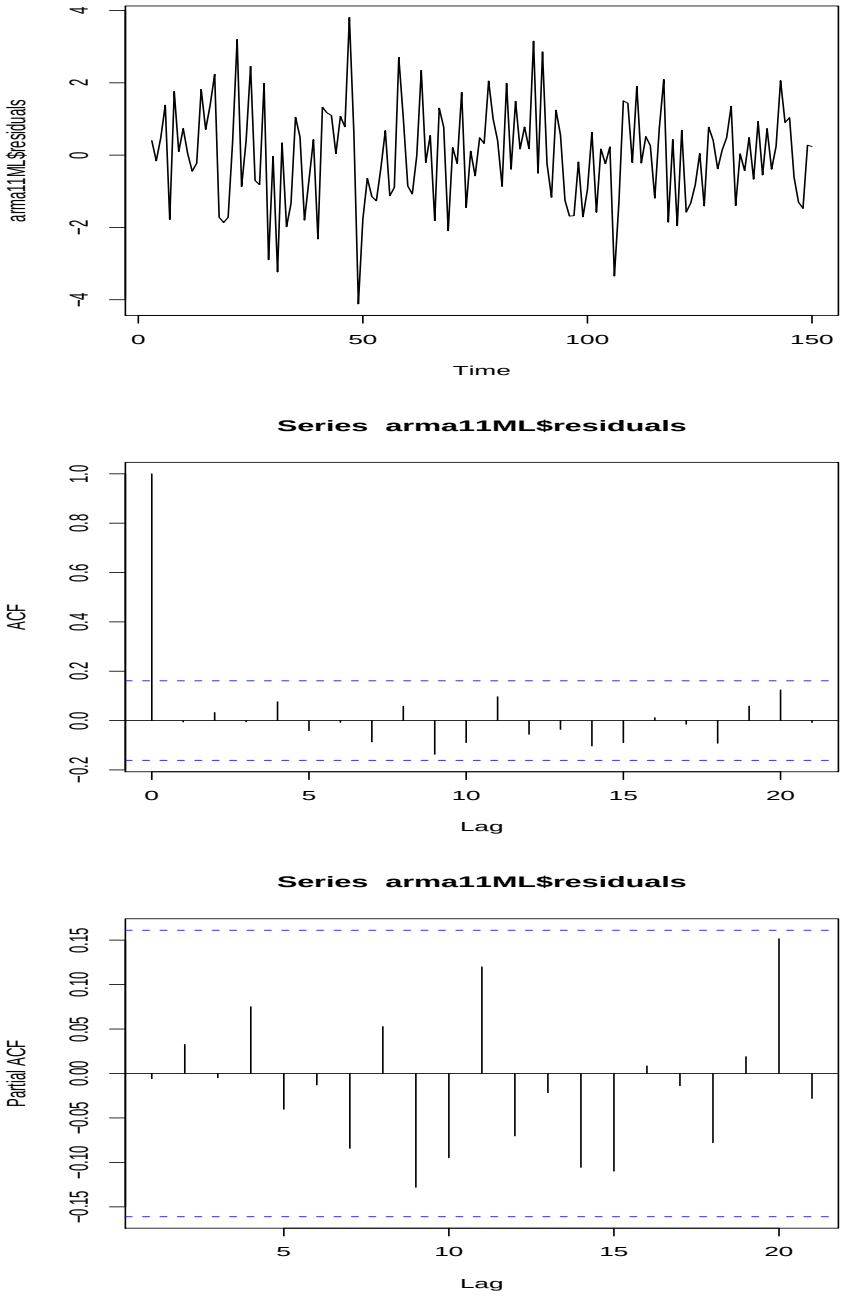

Figure 4.2: The time series, acf and pacf plots of residuals from ARMA(1,1) fot twice differenced **BJsales** data.

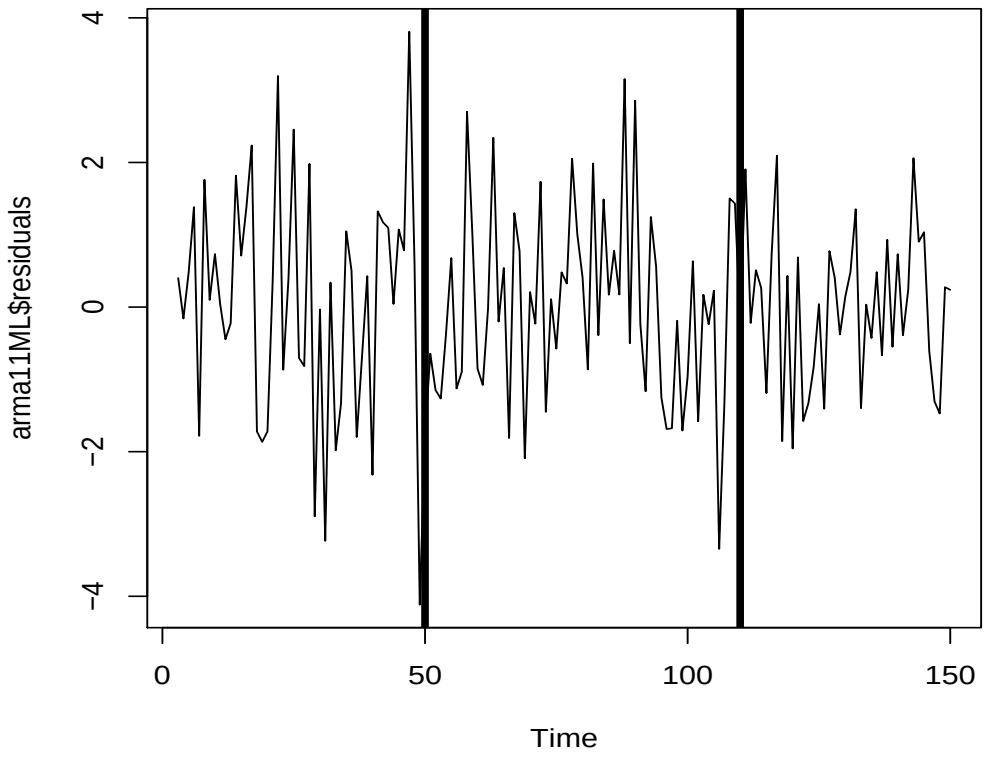

Figure 4.3: The time plot of residuals from ARMA(1,1) fot twice differenced **BJsales** data with splitting into groups.

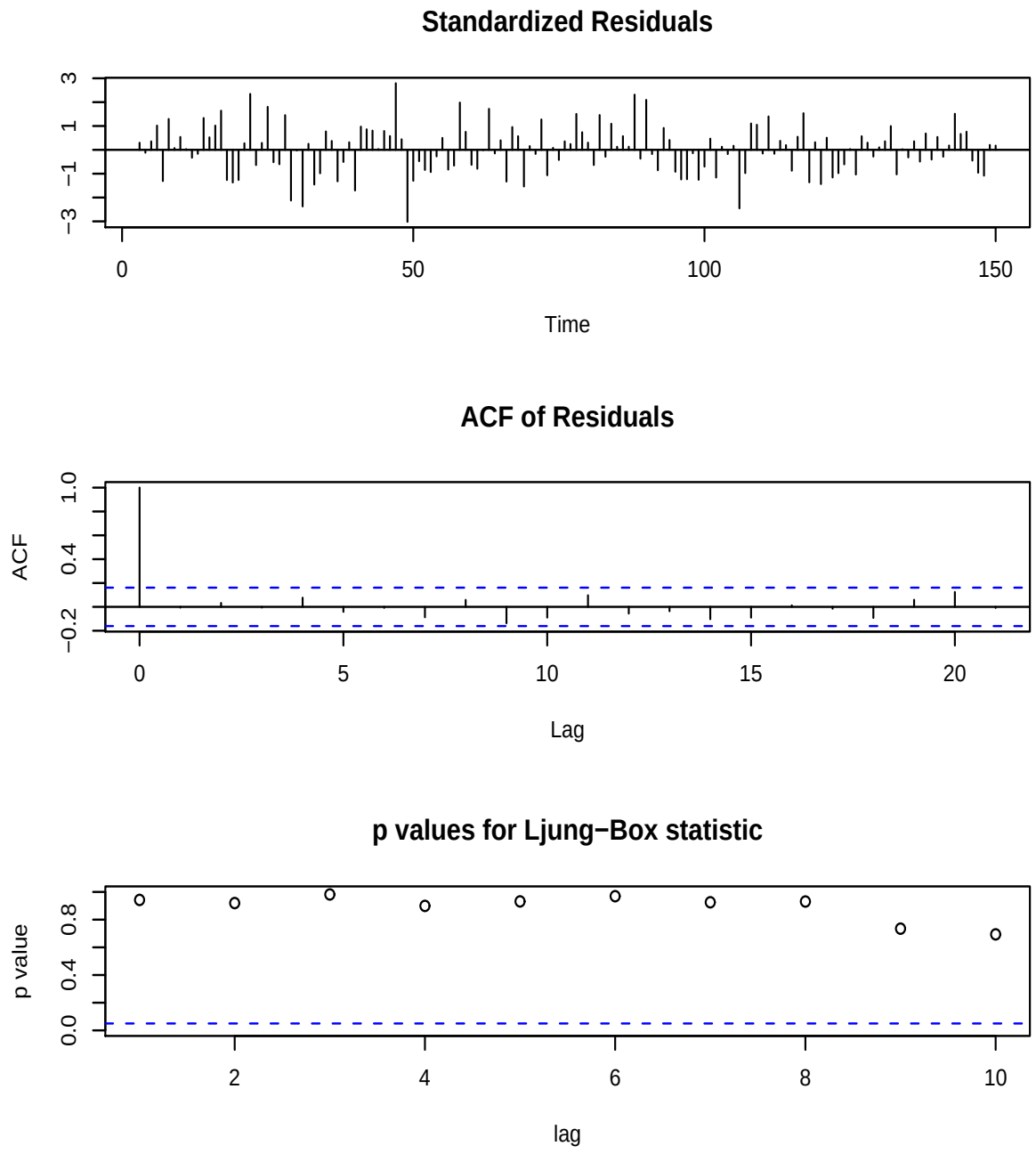

Figure 4.4: The **tsdiag** plot of residuals from ARMA(1,1) fot twice differenced **BJsales** data.

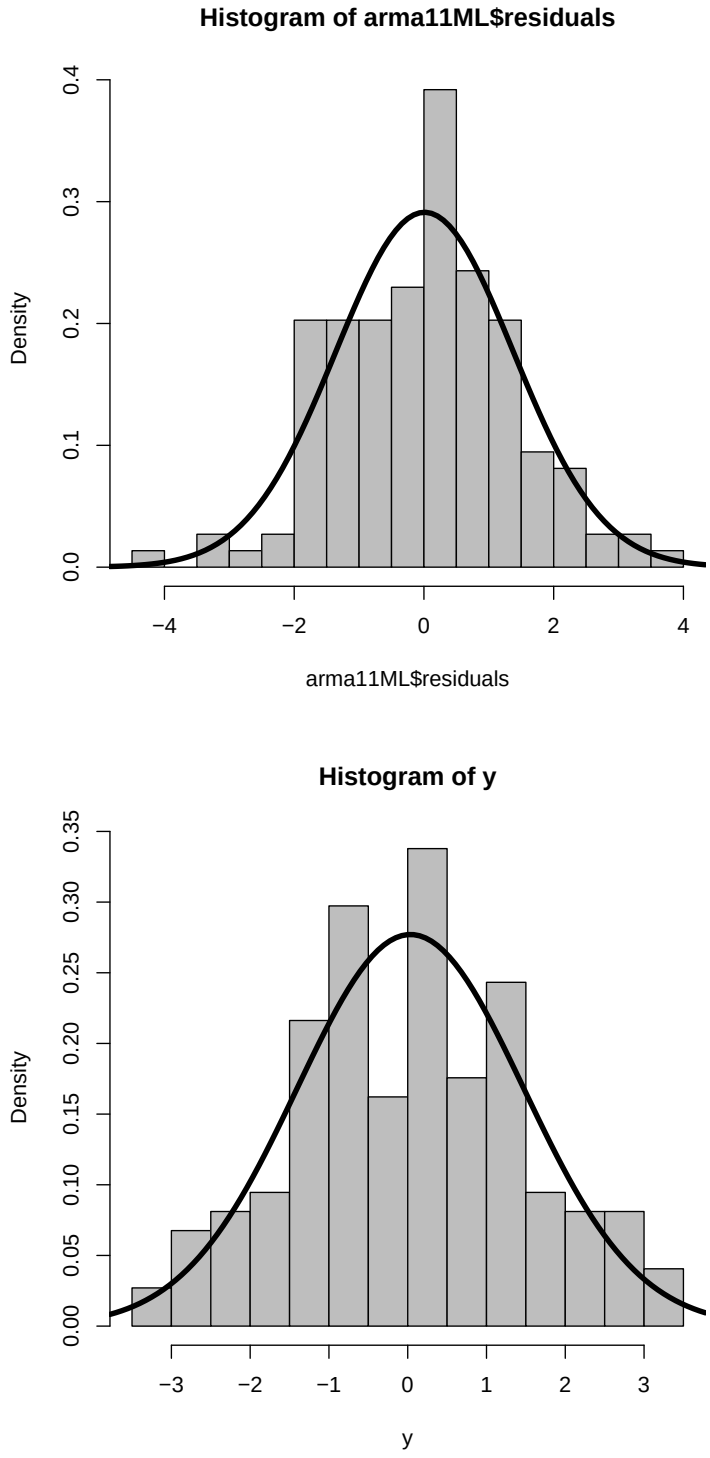

Figure 4.5: Histograms of residuals from ARMA(1,1) for twice differenced **BJsales** data and simulated normal data with the same mean and standard deviation with the superimposed theoretical normal curve.

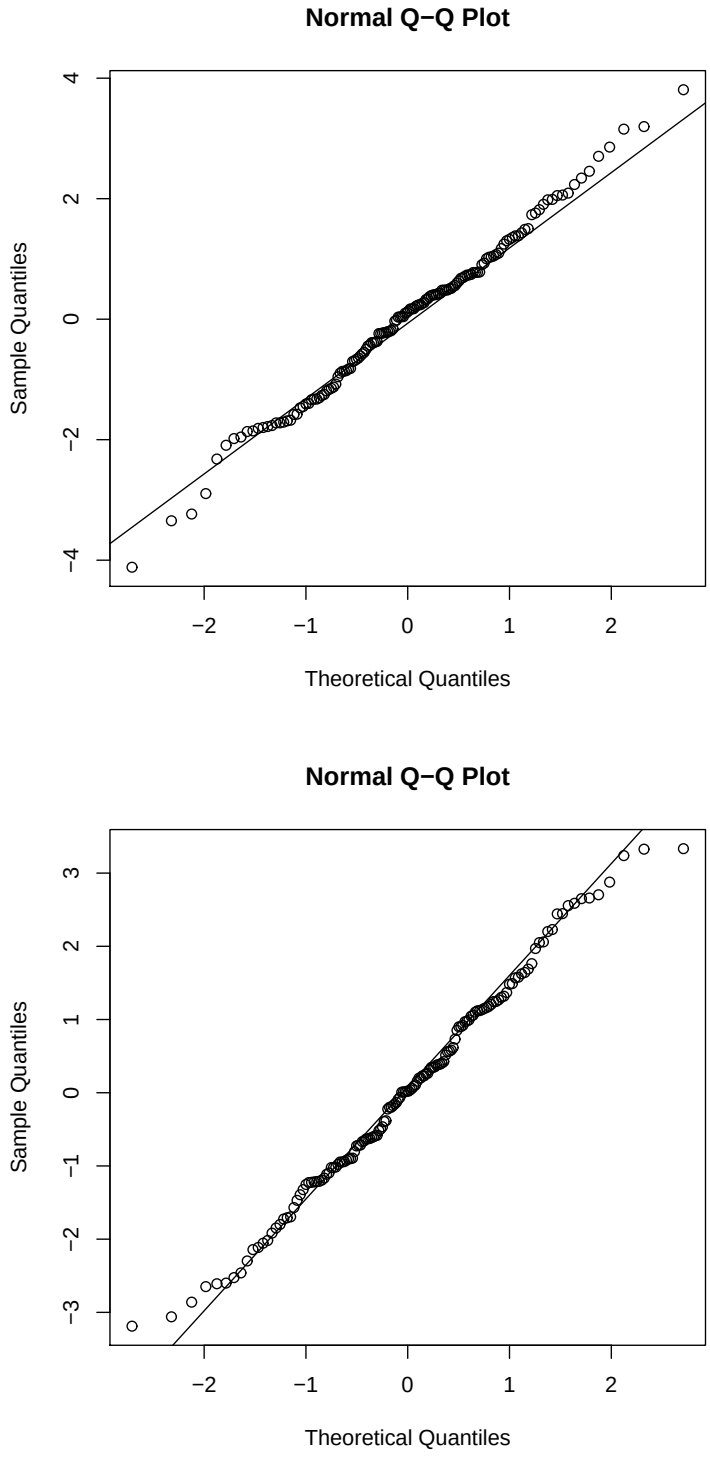

Figure 4.6: QQplots of residuals from ARMA(1,1) for twice differenced **BJsales** data and simulated normal data with the same mean and standard deviation.