ENSAE - IP Paris

Time Series Project

ARIMA Modeling of a Time Series

Authors:
Michael AICHOUN
Jules CHAPON

Supervised by:
Christian FRANCQ

2nd Year ENSAE - Academic Year 2022/2023

1 Data Presentation

1.1 Series Description

The series we have chosen represents the manufacturing of travel articles, leather goods, and saddlery in France. The values are monthly, ranging from January 1990 to February 2023, and the values are expressed on a base 100 relative to 2015. It can be downloaded in CSV format here.

In order to enhance subsequent results, we decide to remove values between 1990 and 2000, as well as the last two values of the series, since we will aim to predict them later.

Subsequently, we reindex the data to arrange them chronologically and format them according to the date format recognized by R. As a result, our post-processed series now consists of 278 observations and is ready for use.

The series, graphically presented below, initially appears to be non-stationary, as clear trends are identifiable.

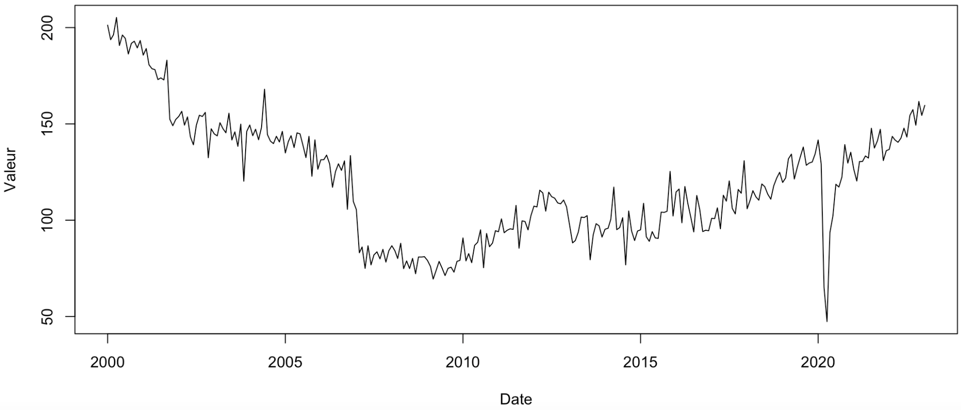


Figure 1: Graphical representation of the studied series

The autocorrelation graph of the series further confirms this suspicion of non-stationarity with a slow decay towards 0.

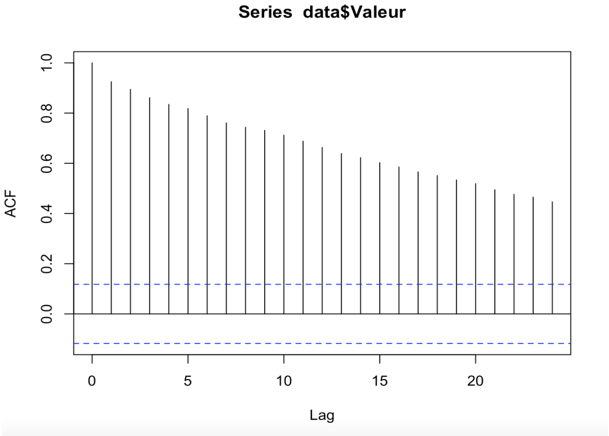


Figure 2: Autocorrelation graph of the series

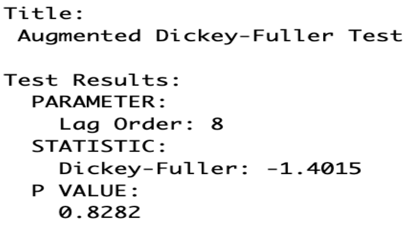
2 Unit Root Tests

To determine the nature of the series and perform the augmented Dickey-Fuller test, we wish to evaluate the presence of a constant, temporal dependence, and potential autocorrelations. To demonstrate this, we regress the series on all dates during which it is observed.

The presence of a significant trend in the series is confirmed by the coefficient associated with dates, which shows potential significance at 5%. However, it is important to note that this significance is not assured due to possible residual autocorrelation. Additionally, the regression constant is also significant at 5% (see detailed results in the appendix). Furthermore, a residual correlation test reveals significant correlation up to lag 8. Therefore, the series can be noted with $t \in T = [\![1, n]\!]$:

$$X_t = c + \alpha t + \beta X_{t-1} + \sum_{l=1}^{8} \phi_l X_{t-l} + \epsilon_t$$

To test the hypothesis of a unit root in the series, we use the augmented Dickey-Fuller test. The results of this test do not allow us to reject the hypothesis of a unit root presence, confirming that the series is non-stationary.



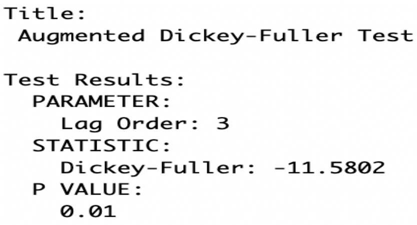
```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 8
  STATISTIC:
    Dickey-Fuller: -1.4015
  P VALUE:
    0.8282
```

Table 1: Stationarity test results

To analyze this time series using classical time series models, we need to stationarize it.

3 Stationarization

To stationarize the series, we apply a first difference to the series X. We denote $Y_t = \Delta X_t$, for all $t \in T$. Conducting a Dickey-Fuller unit root test, we reject the null hypothesis H0 at a significance level of 5%. This confirms that the differenced series Y_t is stationary. Therefore, there is no need for further differencing, and we choose to work with $Y_t = (1 - L)X_t$ as the working series.



```
Title:
 Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 3
  STATISTIC:
    Dickey-Fuller: -11.5802
  P VALUE:
    0.01
```

Table 2: Stationarity test results

4 Graphical Representation

Below are the graphical representations of the initial and final series. The raw series exhibits linear trends, while the differenced series is stationary.

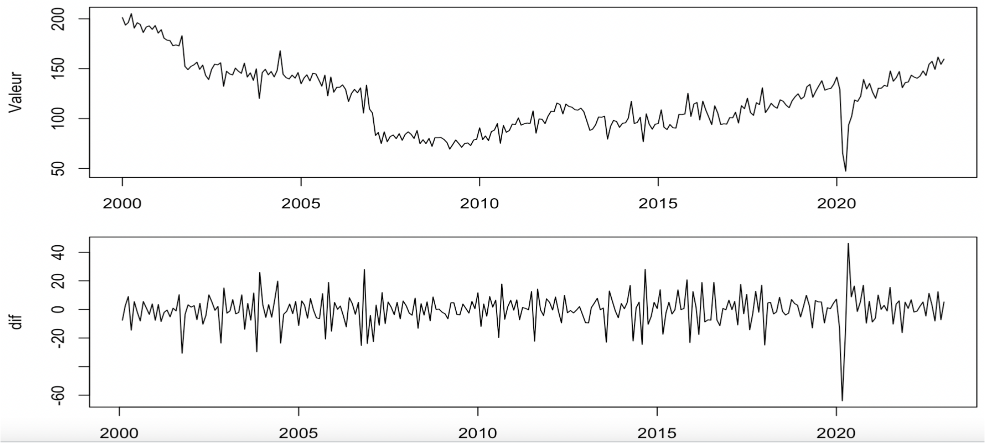


Figure 3: Representation of the initial series (first graph) and the stationarized initial series (second graph)

5 ARMA and ARIMA Modeling

5.1 ARMA Model

The methodology proposed by Box-Jenkins will be followed to identify the ARMA model. It includes the following steps:

- Step 1: Identification of maximum orders p_{max} and q_{max}, then verification of the validity of ARMA(p_{max},q_{max}) model
- Step 2: Selection of possible submodels
- Step 3: Test of fit for submodels
- Step 4: Test of submodel validity
- Step 5: Selection of the best submodel using information criteria

5.1.1 Identification

To determine the values of p_{max} and q_{max} for our ARMA model, we examine our autocorrelation functions, namely the ACF (autocorrelation function) and PACF (partial autocorrelation function).

To determine p_{max}, we look at the largest p such that the autocorrelation function (ACF) is significant. This yields $p_{max} = 1$.

To determine q_{max}, we look at the largest q such that the partial autocorrelation function (PACF) is significant. This yields $q_{max} = 4$.

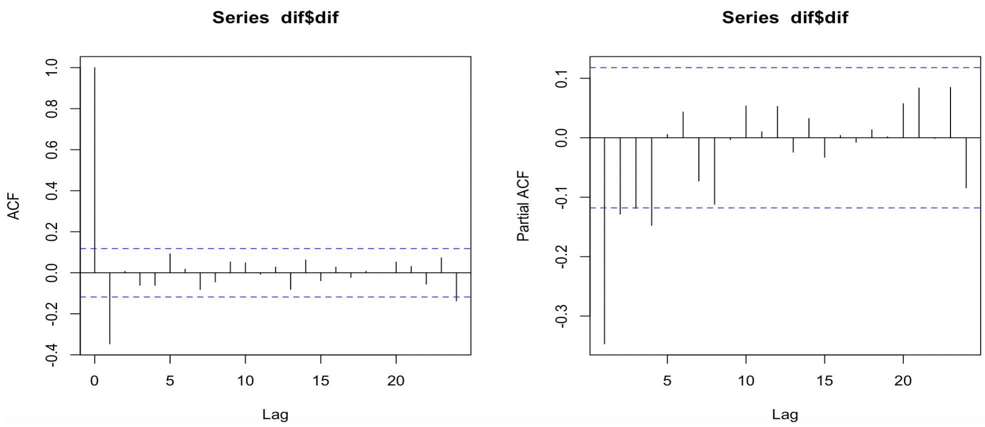
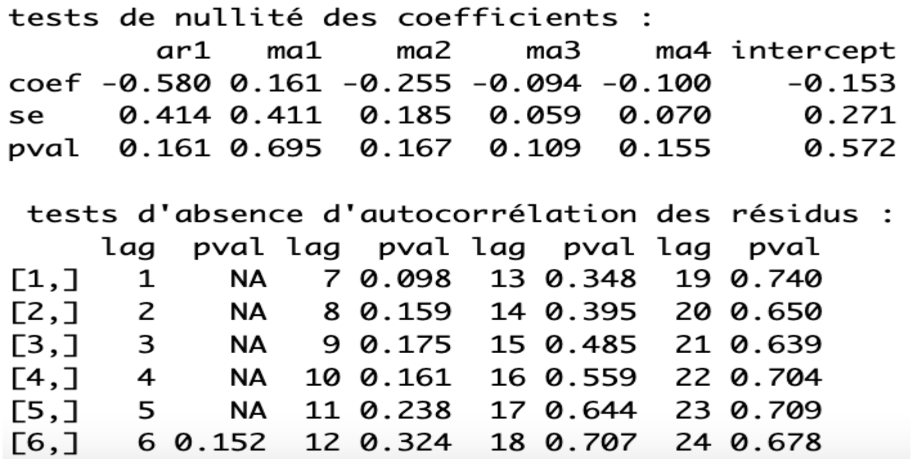


Figure 4: ACF and PACF of the differenced series

To validate the ARMA(1,4) model, we perform a Ljung-Box test to check if the residuals have significant autocorrelations. This test examines autocorrelations up to a specified order k and tests the null hypothesis of no autocorrelation in the residuals. We observe that the absence of autocorrelation is not rejected at 95% significance up to lag 24. Therefore, the model is valid.

Next, we need to test if the model is well-fitted. For this, we conduct the individual coefficient significance test. We need to examine the p-values for the highest-order coefficients (in this case, $p = 1$ and $q = 4$). If these p-values are greater than 0.05, then they do not reject their nullity at 95%, indicating that the model is not well-fitted. Conversely, if the p-values are less than 0.05, then they reject their nullity at 95%, indicating a well-fitted model.



```
tests de nullité des coefficients :
        ar1   ma1    ma2    ma3    ma4 intercept
coef -0.580 0.161 -0.255 -0.094 -0.100    -0.153
se    0.414 0.411  0.185  0.059  0.070     0.271
pval  0.161 0.695  0.167  0.109  0.155     0.572

 tests d'absence d'autocorrélation des résidus :
     lag  pval lag  pval lag  pval lag  pval
[1,]   1    NA   7 0.098  13 0.348  19 0.740
[2,]   2    NA   8 0.159  14 0.395  20 0.650
[3,]   3    NA   9 0.175  15 0.485  21 0.639
[4,]   4    NA  10 0.161  16 0.559  22 0.704
[5,]   5    NA  11 0.238  17 0.644  23 0.709
[6,]   6 0.152  12 0.324  18 0.707  24 0.678
```

Figure 5: Results for ARMA(1,4)

However, we notice that our model is not well-fitted, as the p-values of the highest-order coefficients are greater than 0.05. Therefore, we cannot select this model.

5.1.2 Estimation and Validation of Submodels

To choose a new model, we repeat this parameter validation method (absence of autocorrelation test and individual parameter significance test) for all possible models. These models are all ARMA(p,q) models where:

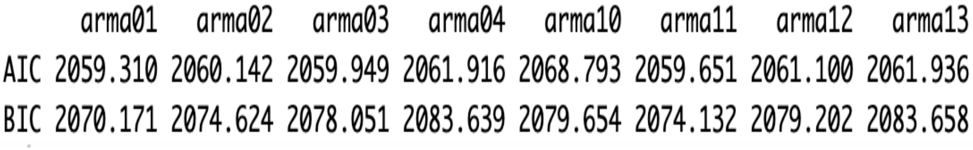
$$\begin{cases} 0 \leq p \leq p_{max} = 1 \\ 0 \leq q \leq q_{max} = 4 \end{cases}$$

We then keep only the models that are valid and well-fitted. However, only the ARMA(0,1) (or MA(1)) model is both valid and well-fitted. Hence, we keep it as the model.

5.1.3 Model Selection

The ARMA(0,1) model is the only model we kept since it was the only one that was both valid and well-fitted.

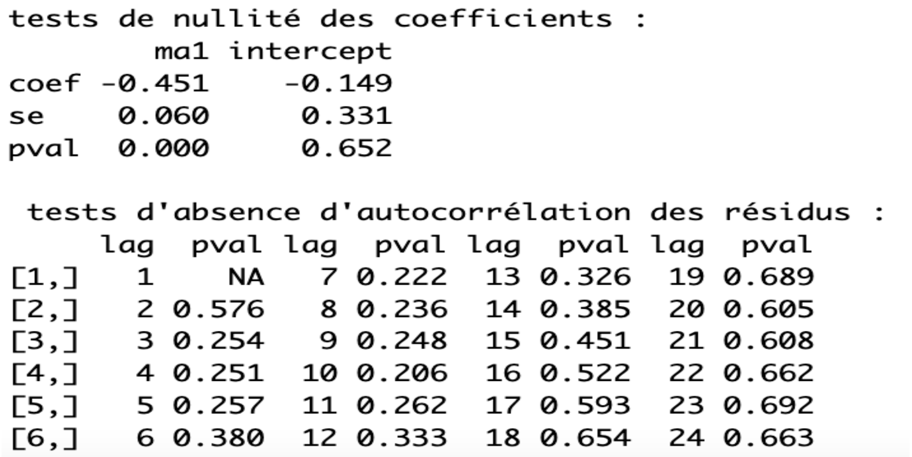
Nevertheless, we still compare the AIC and BIC criteria for all submodels. It can be observed that our ARMA(0,1) model minimizes both criteria. Thus, we choose it as the model.



```
      arma01   arma02   arma03   arma04   arma10   arma11   arma12   arma13
AIC 2059.310 2060.142 2059.949 2061.916 2068.793 2059.651 2061.100 2061.936
BIC 2070.171 2074.624 2078.051 2083.639 2079.654 2074.132 2079.202 2083.658
```

Table 3: Selection by AIC and BIC criteria

It would also have been possible to make predictions with different ARMA models and calculate the Root Mean Squared Error (RMSE) to assess the predictive performance of each model. This would have allowed us to choose the model with the best predictive performance. However, in our case, we have already identified an ARMA(0,1) model that is strongly favored and seems to be the most suitable for representing the data.



```
tests de nullité des coefficients :
        ma1 intercept
coef -0.451    -0.149
se    0.060     0.331
pval  0.000     0.652

 tests d'absence d'autocorrélation des résidus :
     lag  pval lag  pval lag  pval lag  pval
[1,]   1    NA   7 0.222  13 0.326  19 0.689
[2,]   2 0.576   8 0.236  14 0.385  20 0.605
[3,]   3 0.254   9 0.248  15 0.451  21 0.608
[4,]   4 0.251  10 0.206  16 0.522  22 0.662
[5,]   5 0.257  11 0.262  17 0.593  23 0.692
[6,]   6 0.380  12 0.333  18 0.654  24 0.663
```

Figure 5: Results for ARMA(0,1)

5.2 ARIMA Model

Let X_t be our series and Y_t be our first-differenced series. Based on the previous questions, this series follows an ARMA(0,1) model whose parameters we have estimated. Using these estimations, we can write the equation satisfied by our series:

$$Y_t = \epsilon_t - 0.45\epsilon_{t-1}$$

We denote Ψ as the polynomial such that:

$$\Psi(\epsilon_t) = \epsilon_t - 0.45\epsilon_{t-1}$$

Hence,

$$Y_t = \Psi(\epsilon_t)$$

The polynomial Ψ has a unique root $r = 2.22$, which is indeed outside the unit circle. Therefore, our model is invertible. Additionally, our model is an MA(1) and thus, by definition, causal (and there are no common roots).

Thus, our ARMA(0,1) model is a canonical ARMA.

Finally, since $Y_t = \Delta X_t$, we have:

$$(1 - B)X_t = \epsilon_t - 0.45\epsilon_{t-1}$$

i.e.

$$X_t = X_{t-1} + \epsilon_t - 0.45\epsilon_{t-1}$$

Our series X_t therefore follows an ARIMA(0,1,1).

6 Forecasting

6.1 Confidence Regions for (X_{T+1}, X_{T+2})

Let T be the length of the series. We assume that the residuals of the series X_t are Gaussian. We denote ${}_TX_{T+1}$ and ${}_TX_{T+2}$ as the best predictions for X_{T+1} and X_{T+2} given the values of $(X_t)_{t\leq T}$.

Recalling that we have:

$$\begin{cases} X_{T+1} = X_T + \epsilon_{T+1} - 0.45\epsilon_T \\ X_{T+2} = X_{T+1} + \epsilon_{T+2} - 0.45\epsilon_{T+1} \end{cases}$$

Thus,

$$\begin{cases} {}_TX_{T+1} =_T X_T +_T \epsilon_{T+1} - 0.45_T\epsilon_T \\ {}_TX_{T+2} =_T X_{T+1} +_T \epsilon_{T+2} - 0.45_T\epsilon_{T+1} \end{cases}$$

However,

$$\begin{cases} {}_T X_T = X_T \\ {}_T\epsilon_T = \epsilon_T \\ {}_T\epsilon_{T+1} =_T \epsilon_{T+2} = 0 \end{cases}$$

Thus,

$$\begin{cases} {}_T X_{T+1} = X_T - 0.45\epsilon_T \\ {}_T X_{T+2} =_T X_{T+1} \end{cases}$$

Hence,

$$\begin{cases} {}_T X_{T+1} = X_T - 0.45\epsilon_T \\ {}_T X_{T+2} = X_T - 0.45\epsilon_T \end{cases}$$

This leads to

$$\tilde{X} = \begin{pmatrix} X_{T+1} -_T X_{T+1} \\ X_{T+2} -_T X_{T+2} \end{pmatrix} = \begin{pmatrix} \epsilon_{T+1} \\ \epsilon_{T+2} + 0.55\epsilon_{T+1} \end{pmatrix}$$

We set $\mathbb{V}(\epsilon_t) = \sigma^2 > 0$, and as the ϵ_t are uncorrelated (i.e. white noise), we have:

$$\begin{cases} \mathbb{V}(X_{T+1} -_T X_{T+1}) = \sigma^2 \\ \mathbb{V}(X_{T+2} -_T X_{T+2}) = (1 + 0.55^2)\sigma^2 \\ Cov(X_{T+1} -_T X_{T+1}, X_{T+2} -_T X_{T+2}) = 0.55\sigma^2 \end{cases}$$

The vector $\tilde{X}$ then follows a normal distribution $\mathbb{N}(0, \Sigma)$ where:

$$\Sigma = \begin{pmatrix} \sigma^2 & 0.55\sigma^2 \\ 0.55\sigma^2 & (1 + 0.55^2)\sigma^2 \end{pmatrix}$$

with $det(\Sigma) = \sigma^4 > 0$

Thus, Σ is invertible. We then have $\tilde{X}^T \Sigma^{-1} X \ \sim \ \chi^2$

As a result, we obtain two 95% confidence intervals for X_{T+1} and X_{T+2}:

$$X_{T+1} \in \left[{}_T X_{T+2} \pm 1.96\sigma^2\right]$$

and

$$X_{T+2} \in \left[{}_T X_{T+2} \pm 1.96(1 + 0.55^2)\sigma^2\right]$$

6.2 Assumptions for Obtaining Confidence Region

To obtain the results from the previous question, we made several assumptions:

- The innovations ϵ_t are i.i.d. and follow a centered reduced normal distribution.

- The variance of the innovations is known. If it were not the case, we would have needed to estimate it to obtain our confidence intervals.
- The estimator of our ARIMA(0,1,1) model is convergent since we used the estimated parameters as the true parameters.

7 Open Question

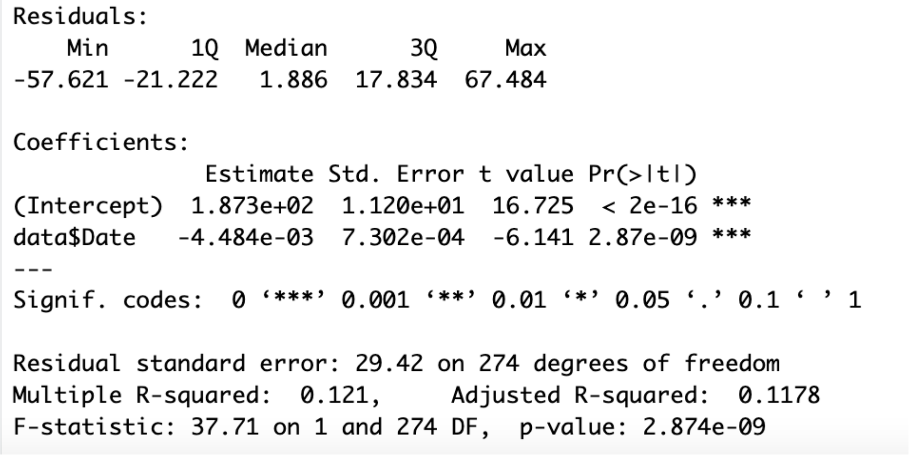
It is possible to improve the forecasting of the variable X_{t+1} by using information from the variable Y_{t+1} if the latter instantly causes the variable X_t in the Granger sense. This means that adding the variable Y_t in the prediction model will yield forecasts for X_t different from those based solely on the history of X_t. In other words, the variable Y_t will be useful for predicting X_t.

To verify this condition, we can use the equation of the causing variable and perform a global significance test of the coefficients in the matrix associated with past values. This will determine if the causing variable is statistically significant in predicting the caused variable.

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \sum_{l=1}^{d} \begin{pmatrix} A_l^{XX} & A_l^{XY} \\ A_l^{YX} & A_l^{YY} \end{pmatrix} \begin{pmatrix} Y_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

8 Appendices

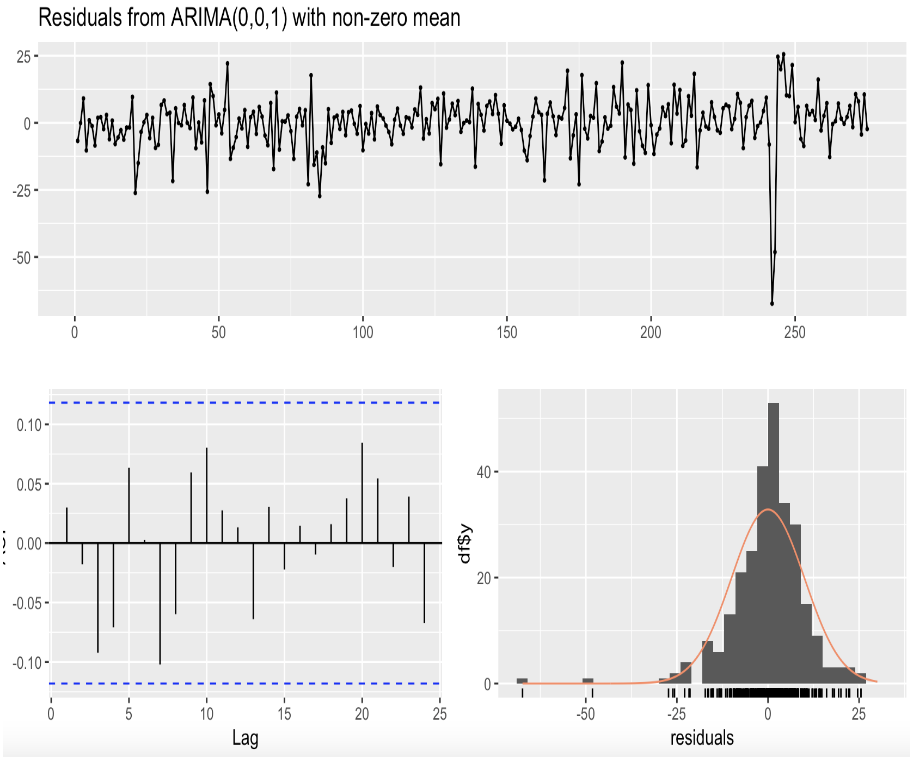
Regression Results on Dates



```
Residuals:
    Min      1Q  Median      3Q     Max
-57.621 -21.222   1.886  17.834  67.484

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.873e+02  1.120e+01  16.725  < 2e-16 ***
data$Date   -4.484e-03  7.302e-04  -6.141 2.87e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.42 on 274 degrees of freedom
Multiple R-squared:  0.121,     Adjusted R-squared:  0.1178
F-statistic: 37.71 on 1 and 274 DF,  p-value: 2.874e-09
```

Table 4: Regression results of the series on dates

Data on Residuals of the ARIMA(0,0,1) Model


Residuals from ARIMA(0,0,1) with non-zero mean
Lag
df$y
residuals