SARIMA Procedures

Kyla Ayop | Jolia Keziah Balcita

December 14, 2023

Contents

I. INTRODUCTION

In this paper, we undertake a comprehensive exploration of SARIMA procedures which involves evaluating and identifying the most suitable sARIMA model for our dataset. The SARIMA model comprises Seasonal AutoRegressive (AR), Integrated (I), and Moving Average (MA) components, denoted as $(p, d, q) * (P, D, Q)_s$. Subsequently, we will also be deriving its equation that encapsulates the mathematical relationship between past observations, errors, and future predictions.

By achieving this with meticulous analysis, we enhance our understanding of the underlying patterns and nuances of our data, providing a concise and predictive framework for time series forecasting.

II. THE DATA

The data we will be using is from the Monthly Medicare Australia prescription data according to Medicare Australia throughout the year 1991-2008. Our main dataset is the Cost of the scripts of Vitamins in Australia.

```
library(fpp3)
library(urca)
```

```
PBS
```

```
## # A tsibble: 67,596 x 9 [1M]
## # Key:       Concession, Type, ATC1, ATC2 [336]
##       Month Concession   Type        ATC1  ATC1_desc ATC2  ATC2_~1 Scripts  Cost
##       <mth> <chr>        <chr>       <chr> <chr>     <chr> <chr>     <dbl> <dbl>
##  1 1991 Jul Concessional Co-payments A     Alimenta~ A01   STOMAT~   18228 67877
##  2 1991 Aug Concessional Co-payments A     Alimenta~ A01   STOMAT~   15327 57011
##  3 1991 Sep Concessional Co-payments A     Alimenta~ A01   STOMAT~   14775 55020
##  4 1991 Oct Concessional Co-payments A     Alimenta~ A01   STOMAT~   15380 57222
##  5 1991 Nov Concessional Co-payments A     Alimenta~ A01   STOMAT~   14371 52120
##  6 1991 Dec Concessional Co-payments A     Alimenta~ A01   STOMAT~   15028 54299
##  7 1992 Jan Concessional Co-payments A     Alimenta~ A01   STOMAT~   11040 39753
##  8 1992 Feb Concessional Co-payments A     Alimenta~ A01   STOMAT~   15165 54405
##  9 1992 Mar Concessional Co-payments A     Alimenta~ A01   STOMAT~   16898 61108
## 10 1992 Apr Concessional Co-payments A     Alimenta~ A01   STOMAT~   18141 65356
## # ... with 67,586 more rows, and abbreviated variable name 1: ATC2_desc
```

```
unique(PBS$ATC2)
```

```
##  [1] "A01" "A02" "A03" "A04" "A05" "A06" "A07" "A09" "A10" "A11" "A12" "A14"
## [13] "A15" "B01" "B02" "B03" "B05" "C01" "C02" "C03" "C04" "C05" "C07" "C08"
## [25] "C09" "C10" "D"   "D01" "D02" "D04" "D05" "D06" "D07" "D08" "D10" "D11"
## [37] "G01" "G02" "G03" "G04" "H01" "H02" "H03" "H04" "H05" "J01" "J02" "J04"
## [49] "J05" "J06" "J07" "L01" "L02" "L03" "L04" "M01" "M02" "M03" "M04" "M05"
## [61] "N02" "N03" "N04" "N05" "N06" "N07" "P01" "P02" "P03" "R"   "R01" "R03"
## [73] "R05" "R06" "S"   "S01" "S02" "S03" "V01" "V03" "V04" "V06" "V07" "Z"
```

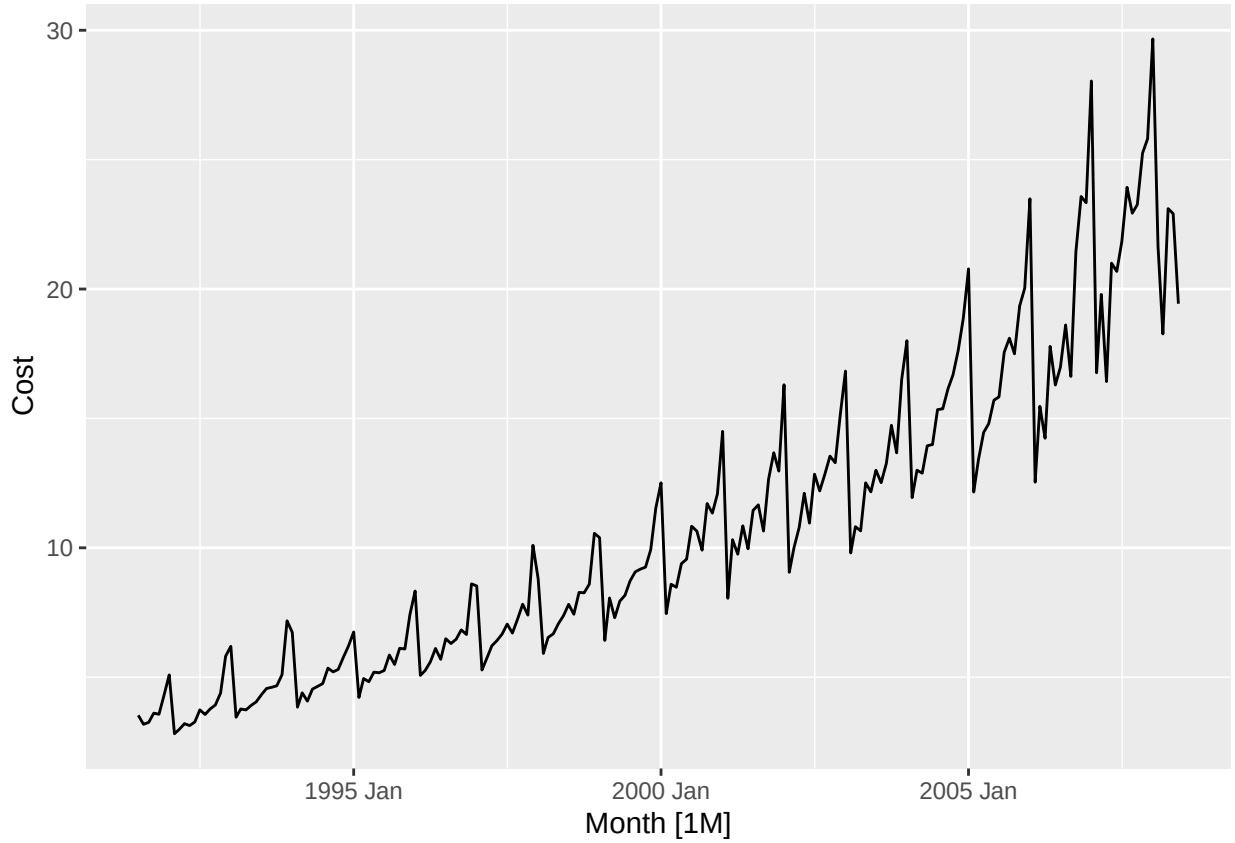
III. ARIMA MODELLING PROCEDURE

1 Historical Plot

Monthly vitamins sales in Australia. These are known as A10 drugs under the Anatomical Therapeutic Chemical classification scheme.

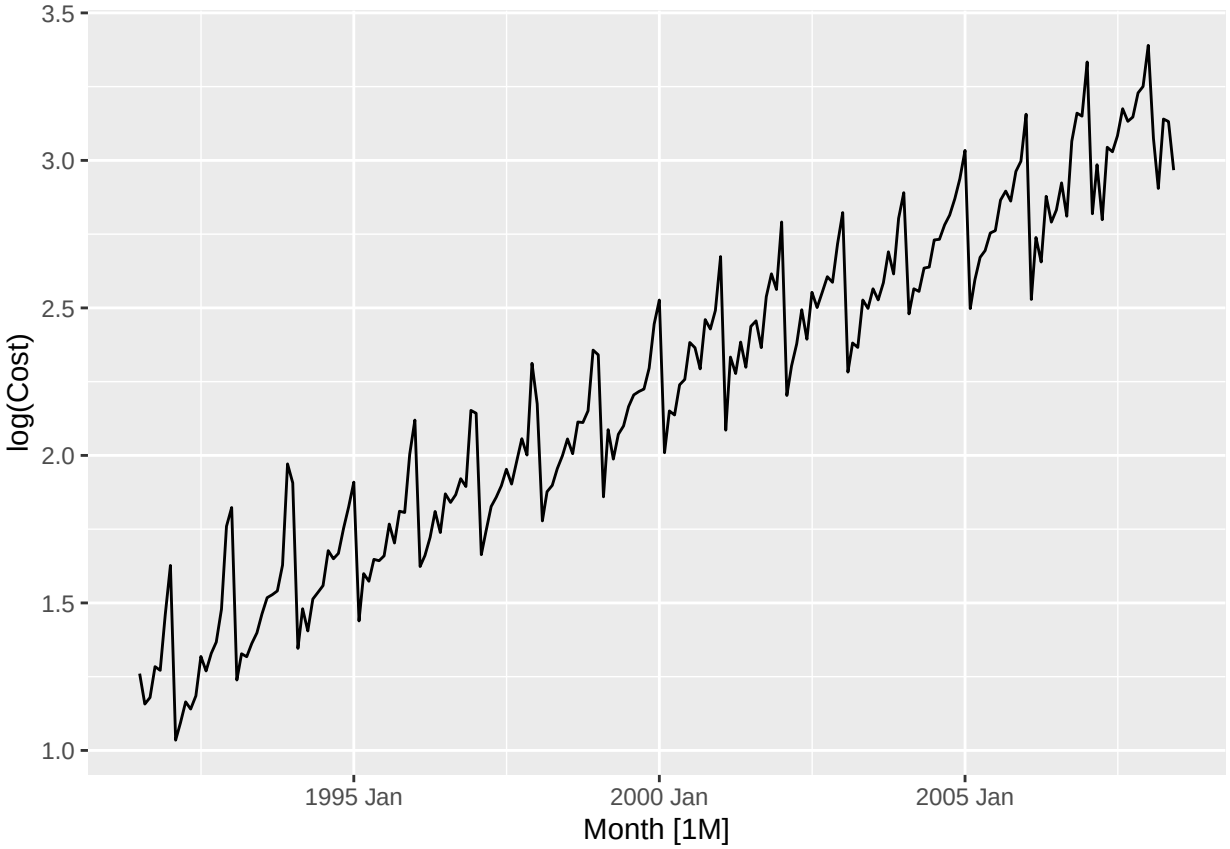
```
a10 <- PBS |>
  filter(ATC2 == "A10") |>
  summarise(Cost = sum(Cost)/1e6)
```

```
a10 |> autoplot(
  Cost
)
```



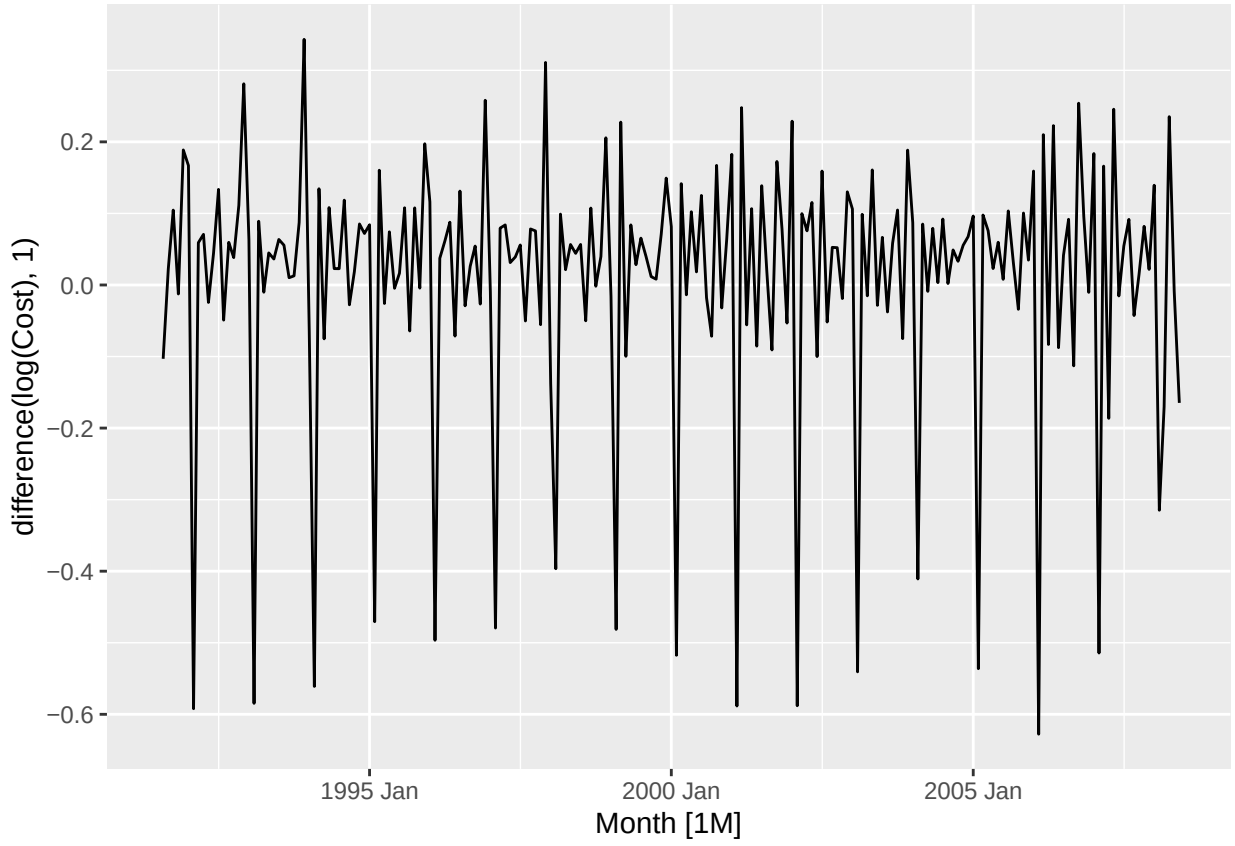
We take the log of the Cost to transform the volatility of the variance.

```
a10 |> autoplot(
  log(Cost)
)
```

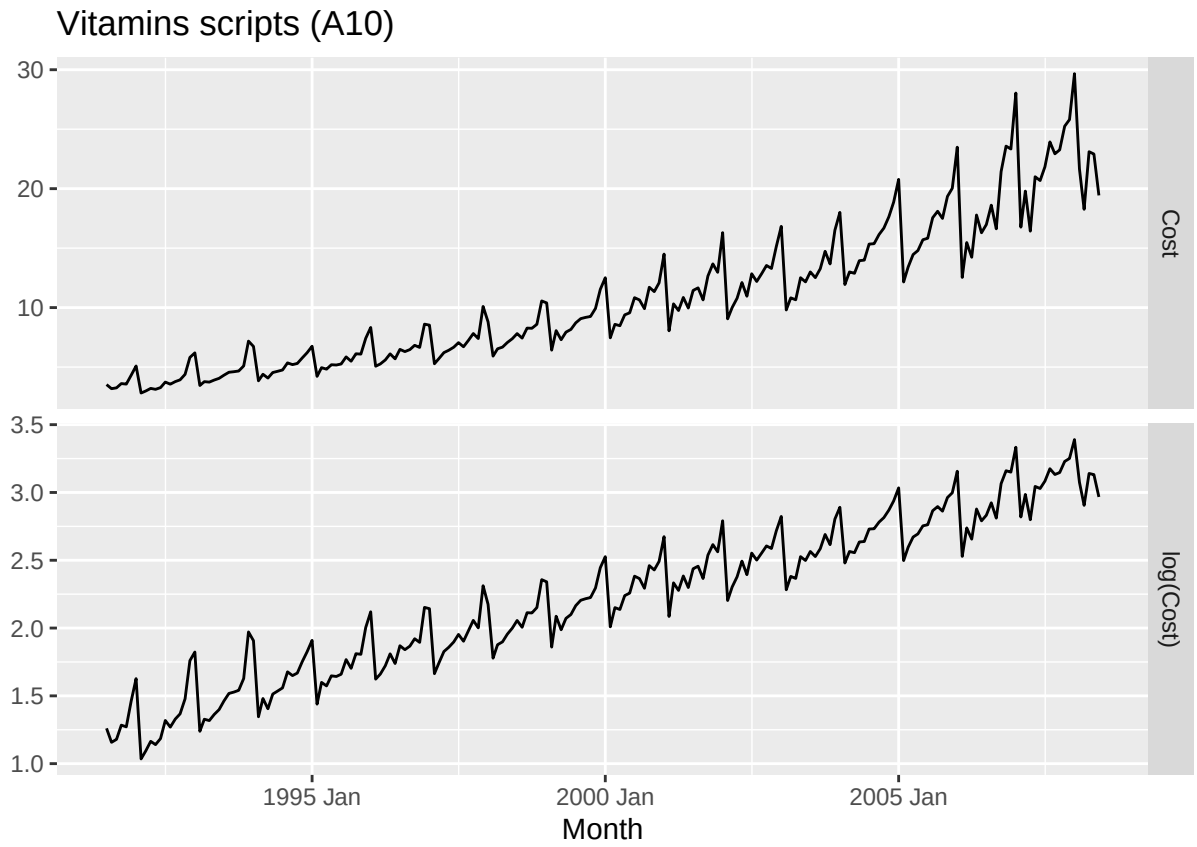


And then take the difference of order 1 to make the series stationary with constant mean and variance.

```
a10 |> autoplot(
  log(Cost) |> difference(1)
)
```




```
a10 |>
  mutate(log(Cost)) |>
  pivot_longer(-Month) |>
  ggplot(aes(x = Month, y = value)) +
  geom_line() +
  facet_grid(name ~ ., scales = "free_y") +
  labs(y="", title="Vitamins scripts (A10)")
```



Data from July 1991 to June 2008 are plotted in Figure. Hence, there is an increase in the variance with the level, so we take logarithm to stabilize or make the variance constant. The data are strongly seasonal and obviously non-stationary, so seasonal differencing is used after log transformation.

3 Stationarity and Volatility

With the use of KPSS Test to test for stationarity.

```
PBS |>
  filter(ATC2 == "A10") |>
  features(Cost, unitroot_kpss)
```

```
## # A tibble: 4 x 6
##   Concession   Type         ATC1  ATC2  kpss_stat kpss_pvalue
##   <chr>        <chr>        <chr> <chr>     <dbl>       <dbl>
## 1 Concessional Co-payments  A     A10        3.51        0.01
## 2 Concessional Safety net   A     A10        1.86        0.01
## 3 General      Co-payments  A     A10        3.39        0.01
## 4 General      Safety net   A     A10        1.33        0.01
```

```
PBS |>
  filter(ATC2 == "A10") |>
  features(difference(log(Cost)), unitroot_kpss)
```

```
## # A tibble: 4 x 6
##   Concession   Type         ATC1  ATC2  kpss_stat kpss_pvalue
##   <chr>        <chr>        <chr> <chr>     <dbl>       <dbl>
## 1 Concessional Co-payments  A     A10     0.00837         0.1
## 2 Concessional Safety net   A     A10     0.00847         0.1
## 3 General      Co-payments  A     A10     0.0832          0.1
## 4 General      Safety net   A     A10     0.00966         0.1
```

Thus, taking the first difference and the log resulted to a stationary data. Choose D = 1 and d = 0.

This is the seasonally differenced data.

```
a10 |> gg_tsdisplay(difference(log(Cost), 12),
                   plot_type='partial', lag_max = 24)
```

In the plots of the seasonally differenced data, there are spikes in the PACF at lags 12 and 24, and also seasonal lags in the ACF at 12 and 24. This may be suggestive of a seasonal AR(2) term. In the non-seasonal lags, there are three significant spikes in the PACF, suggesting a possible AR(3) term. The pattern in the ACF is not indicative of any simple model.

Spikes in PACF at lags 12 and 24 suggest seasonal AR(2) term. Spikes in PACF suggests possible non-seasonal AR(3) term. Initial candidate model:$ARIMA(3,0,0)(2,1,0)_{12}$

4 Examine ACF/PACF

From the PACF, suggests AR(4) model which results to ARIMA(4,1,0). From the ACF, it suggests MA(1) model which then be SARIMA(0,1,1). To explore more models we included the $ARIMA(3,0,4)(2,1,2)_{12}$model.

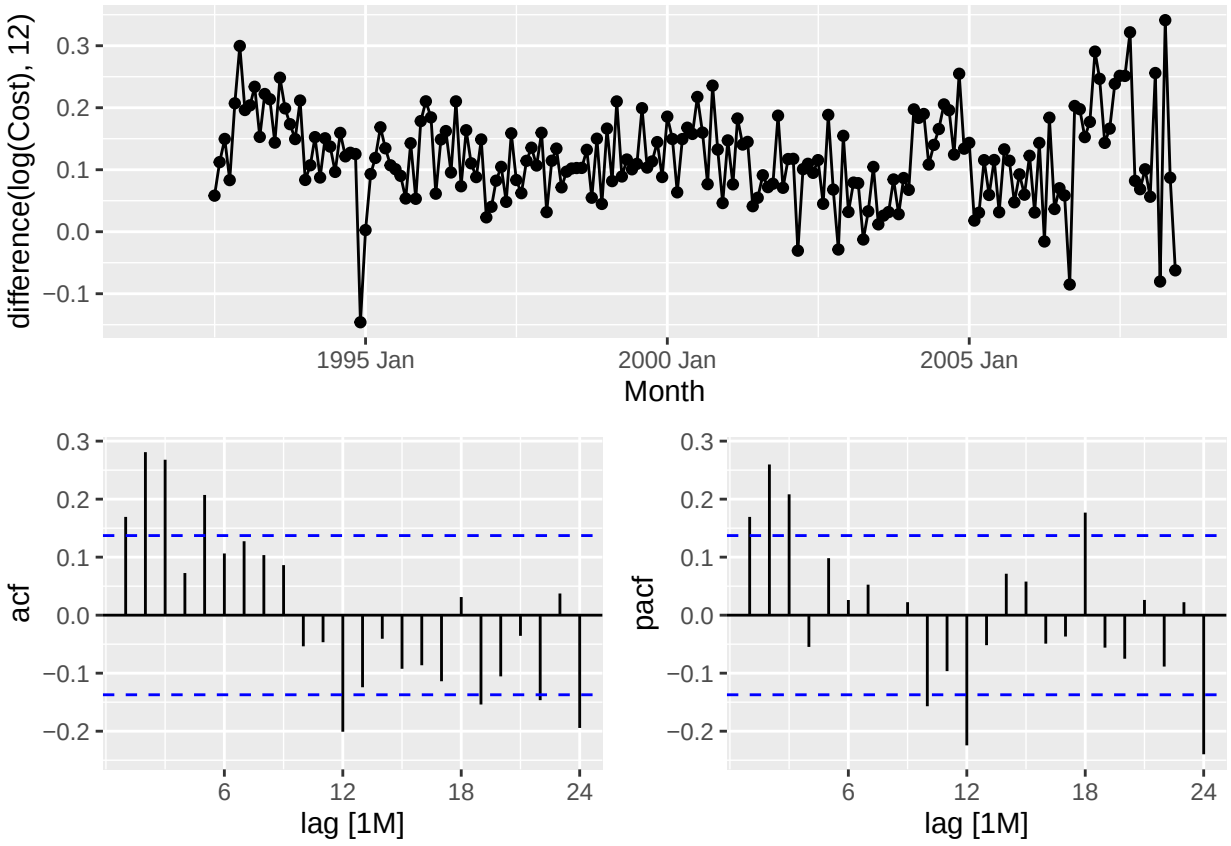


Figure 1: Seasonally differenced vitamins sales in Australia (in millions of scripts per month).

```
fit <- a10 |>
  model(arima300210 = ARIMA(log(Cost) ~ 0 + pdq(3,0,0) + PDQ(2,1,0)),
        arima004012 = ARIMA(log(Cost) ~ 0 + pdq(0,0,4) + PDQ(0,1,2)),
        arima201110 = ARIMA(log(Cost) ~ 0 + pdq(2,0,1) + PDQ(1,1,0)),
        arima301012 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(0,1,2)),
        arima301011 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(0,1,1)),
        arima300112 = ARIMA(log(Cost) ~ 0 + pdq(3,0,0) + PDQ(1,1,2)),
        auto = ARIMA(log(Cost), stepwise = FALSE, approx = FALSE))

fit |> pivot_longer(everything(), names_to = "Model name",
                          values_to = "Orders")
```

```
## # A mable: 7 x 2
## # Key:     Model name [7]
##   `Model name`                                Orders
##   <chr>                                      <model>
## 1 arima300210           <ARIMA(3,0,0)(2,1,0)[12]>
## 2 arima004012           <ARIMA(0,0,4)(0,1,2)[12]>
## 3 arima201110           <ARIMA(2,0,1)(1,1,0)[12]>
## 4 arima301012           <ARIMA(3,0,1)(0,1,2)[12]>
## 5 arima301011           <ARIMA(3,0,1)(0,1,1)[12]>
## 6 arima300112           <ARIMA(3,0,0)(1,1,2)[12]>
## 7 auto         <ARIMA(3,0,0)(2,1,1)[12] w/ drift>
```

5 Model Testing

Now we fit the models.

```
glance(fit) |> arrange(AICc) |> select(.model:BIC)
```

```
## # A tibble: 7 x 6
##   .model       sigma2 log_lik   AIC  AICc   BIC
##   <chr>         <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 auto        0.00338    272. -528. -527. -501.
## 2 arima301012 0.00371    261. -508. -507. -485.
## 3 arima300112 0.00374    261. -507. -507. -485.
## 4 arima301011 0.00386    258. -503. -503. -484.
## 5 arima300210 0.00400    255. -499. -499. -479.
## 6 arima201110 0.00468    243. -475. -475. -459.
## 7 arima004012 0.00779    196. -378. -378. -355.
```

Of these models, the best is the $SARIMA(3,0,0)(2,1,1)_{12}$ model since it has the smallest AICc value.

```
fit <- a10 |>
  model(auto = ARIMA(log(Cost), stepwise = FALSE, approx = FALSE))
report(fit)
```

```
## Series: Cost
## Model: ARIMA(3,0,0)(2,1,1)[12] w/ drift
## Transformation: log(Cost)
##
```

```
## Coefficients:
##          ar1     ar2     ar3    sar1     sar2     sma1  constant
##       0.0420  0.2596  0.3068  0.0868  -0.2640  -0.6873    0.0521
## s.e.  0.0731  0.0685  0.0754  0.1245   0.0993   0.1112    0.0015
##
## sigma^2 estimated as 0.003376:  log likelihood=271.77
## AIC=-527.54   AICc=-526.75   BIC=-501.48
```

The SARIMA(3,0,0)(2,1,1)_{12} with drift model has the equation:

The $SARIMA(3,0,0)(2,1,1)_{12}$ model with a drift and logarithmic transformation for the given series "Cost" can be expressed as follows:

$y_t = 0.0521 + 0.0420y_{t-1} + 0.2596y_{t-2} + 0.3068y_{t-3} + 0.086y'_{t-12} - 0.2640y'_{t-24} - 0.6873\epsilon_{t-1} + \epsilon_t$

The coefficients are:

Intercept (constant): 0.0521

Autoregressive coefficients: 0.0420, 0.2596, 0.3068

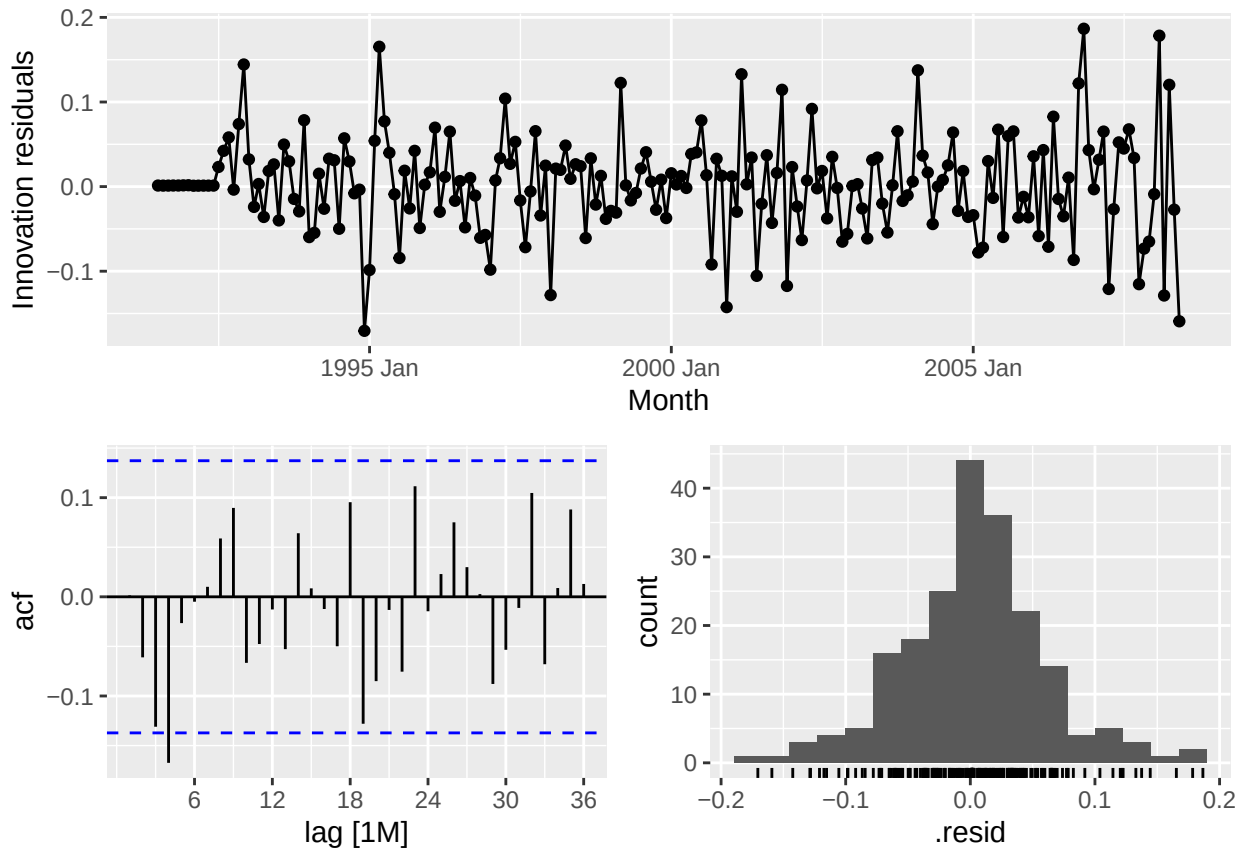
Seasonal Autoregressive coefficients: 0.0868, -0.2640

Seasonal Moving Average coefficient: -0.6873

The standard errors for each coefficient are provided as well.

6 Assessment of Residuals

```
fit <- a10 |>
  model(ARIMA(log(Cost) ~ 0 + pdq(3,0,0) + PDQ(2,1,1)))
fit |> gg_tsresiduals(lag_max=36)
```



The ACF plot of the residuals from the $SARIMA(3,0,0)(2,1,1)_{12}$ model shows One small but significant spike (at lag 4) out of 36 is still consistent with white noise. With the Histogram shows an almost bell shaped graph and the plot for the residuals are zero or closer to it.

Portmanteau test of the residuals

```
augment(fit) |>
  features(.innov, ljung_box, lag = 36, dof = 6)
```

```
## # A tibble: 1 x 3
##   .model                                              lb_stat lb_pvalue
##   <chr>                                                 <dbl>     <dbl>
## 1 ARIMA(log(Cost) ~ 0 + pdq(3, 0, 0) + PDQ(2, 1, 1))    38.3     0.142
```

A portmanteau test (setting K=6) returns a large p-value enough to suggest that the residuals are white noise. With the null hypothesis that states that the series is white noise or independent and identically distributed. With an alpha of 0.05, the Ljung-Box test results provide strong evidence that we do not reject the null hypothesis that the fit model of $SARIMA(3,0,0)(2,1,1)_{12}$ is white noise. The p-value of approximately 0.1422364 indicates that the series is random or white noise or independent and identically distributed.

7 Forecasting

Now, that the residuals look like white noise. Forecasts from the $SARIMA(3,0,1)(0,1,2)_{12}$ model applied to the A10 monthly script sales data. With 80% and 95% prediction intervals shown.

```
a10 |>
  model(ARIMA(log(Cost) ~ 0 + pdq(3,0,0) + PDQ(2,1,1))) |>
  forecast() |>
  autoplot(a10) +
  labs(y=" $AU (millions)",
       title="Vitamins scripts (A10) sales data")
```

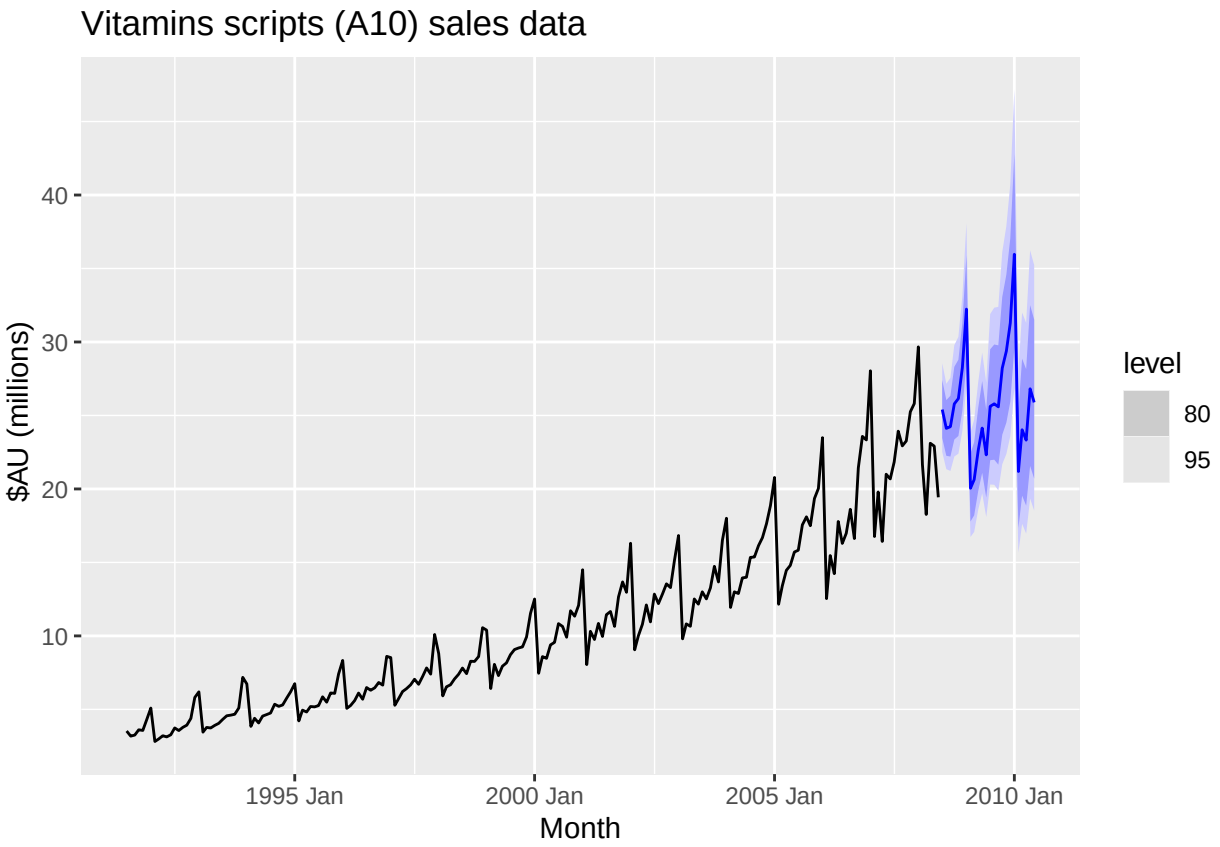


Figure 2: Forecast Cost of the scripts of Vitamins in Australia.