SARIMA

Charlene P. Garridos & Ken Andrea Bahian

2023-12-14

Contents

1 R Libraries

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

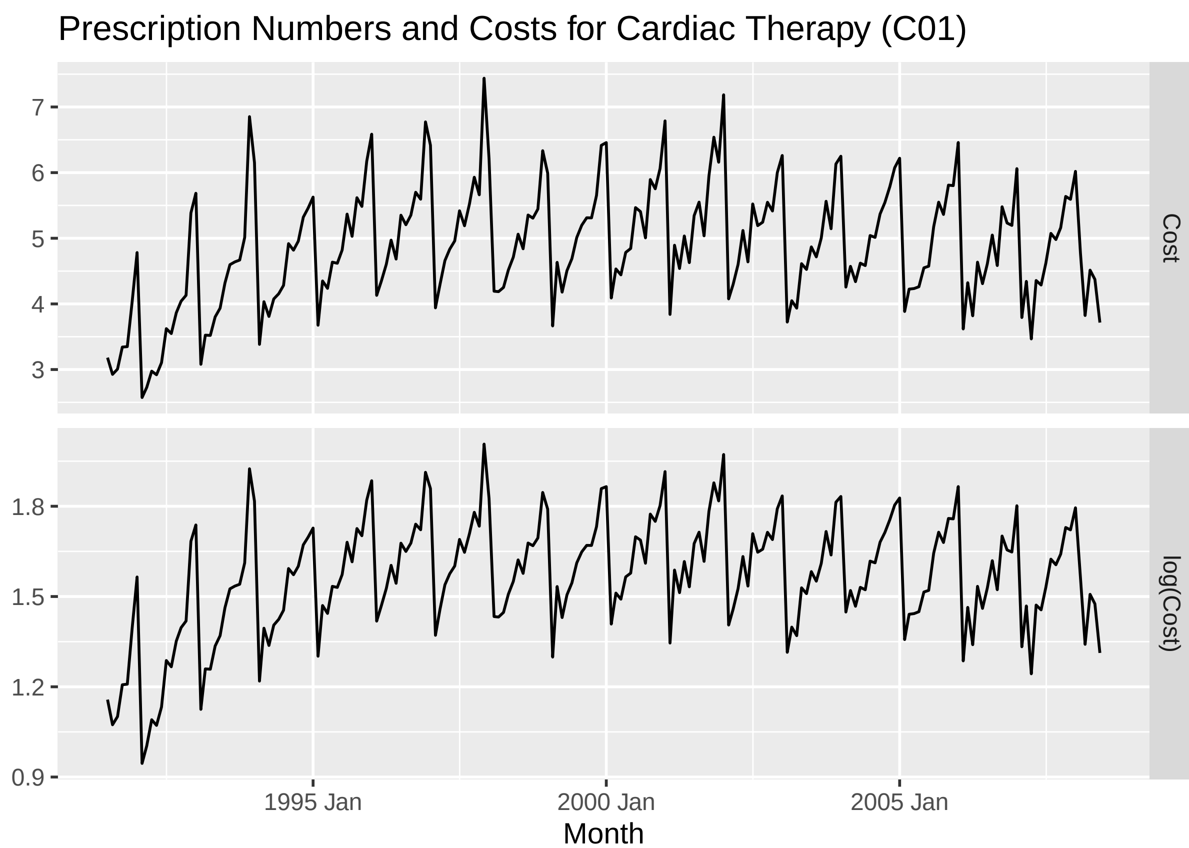
2 The Pharmaceutical Benefits Scheme (PBS) data Of 2021

```r
C01 <- PBS %>%
  filter(ATC2 == "C01")
print(C01)
```

```
## # A tsibble: 816 x 9 [1M]
## # Key:       Concession, Type, ATC1, ATC2 [4]
##       Month Concession   Type     ATC1  ATC1_desc ATC2  ATC2_desc Scripts   Cost
##       <mth> <chr>        <chr>    <chr> <chr>     <chr> <chr>       <dbl>  <dbl>
##  1 1991 Jul Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  205217 2.42e6
##  2 1991 Aug Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  170661 2.03e6
##  3 1991 Sep Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  159605 1.91e6
##  4 1991 Oct Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  158632 1.91e6
##  5 1991 Nov Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  135981 1.69e6
##  6 1991 Dec Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  139641 1.74e6
##  7 1992 Jan Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  122117 1.57e6
##  8 1992 Feb Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  161408 2.11e6
##  9 1992 Mar Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  188039 2.49e6
## 10 1992 Apr Concessional Co-paym~ C     Cardiova~ C01   CARDIAC ~  201915 2.70e6
## # i 806 more rows
```

3 Historical Plot

```r
C01 <- PBS |>
  filter(ATC2 == "C01") |>
  summarise(Cost = sum(Cost)/1e6)
C01 |>
  mutate(log(Cost)) |>
  pivot_longer(-Month) |>
  ggplot(aes(x = Month, y = value)) +
  geom_line() +
  facet_grid(name ~ ., scales = "free_y") +
  labs(y="", title="Prescription Numbers and Costs for Cardiac Therapy (C01)")
```

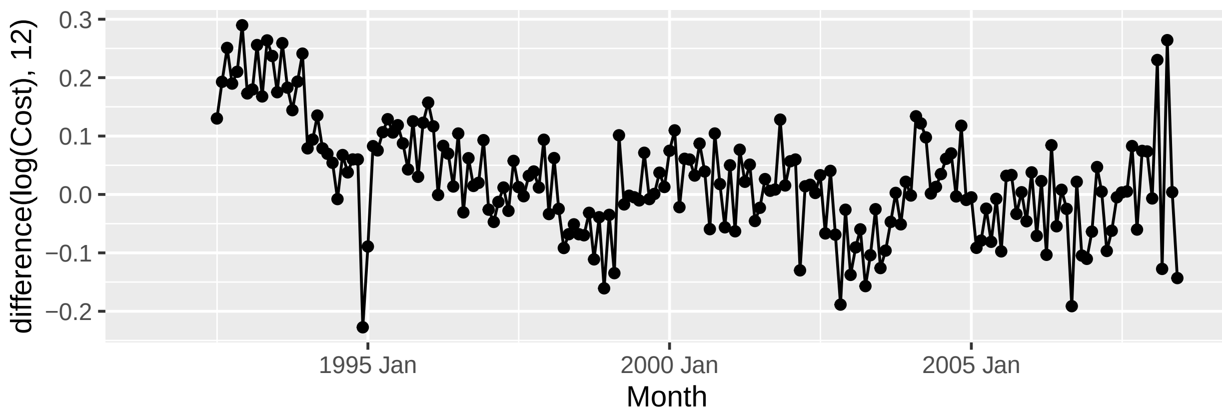


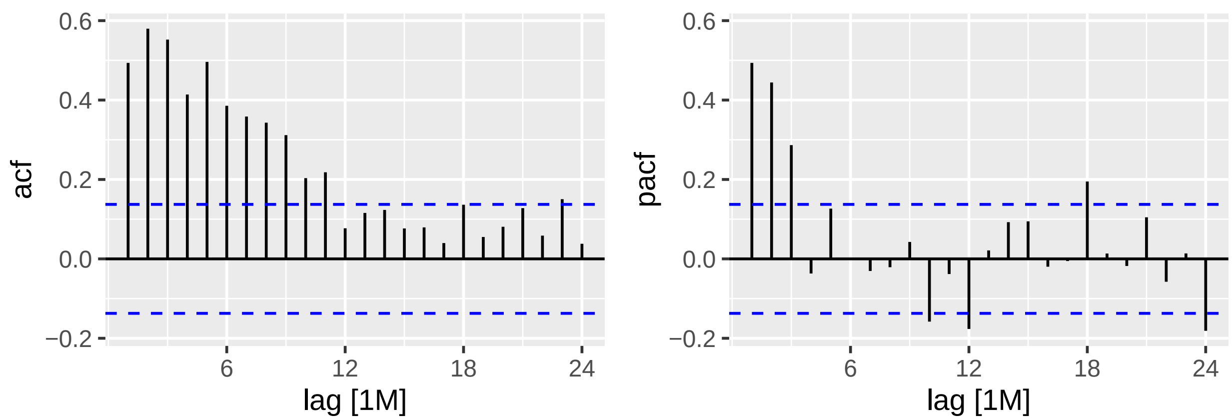
Interpretation

Data from July 1991 to June 2008 are plotted in the figure above. There is a small increase in the variance with the level, so we take logarithms to stabilise the variance.The data are strongly seasonal and obviously non-stationary, so seasonal differencing will be used.

4 The seasonally differenced data

```
C01 |>
  gg_tsdisplay(difference(log(Cost), 12),
                      plot_type='partial', lag_max = 24)
```





Interpretation

These are also clearly non-stationary. In the plots of the seasonally differenced data, there are spikes in the PACF at lags 12 and 24, but nothing at seasonal lags in the ACF. 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. Consequently, this initial analysis suggests that a possible model for these data is an ARIMA$(3, 0, 0)(2, 1, 0)_{12}$. We fit this model, along with some variations on it, and compute the AICc values shown in the figure below.

5 Employ the AICc criterion

```
fit <- C01 |>
  model(
    arima301012 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(0,1,2)),
    arima301111 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(1,1,1)),
    arima301011 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(0,1,1)),
    arima301110 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(1,1,0)),
    arima301210 = ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(2,1,0)),
    arima300111 = ARIMA(log(Cost) ~ 0 + pdq(3,0,0) + PDQ(1,1,1)),
    arima300110 = ARIMA(log(Cost) ~ 0 + pdq(3,0,0) + PDQ(1,1,0)),
    arima302210 = ARIMA(log(Cost) ~ 0 + pdq(3,0,2) + PDQ(2,1,0)),
    arima302310 = ARIMA(log(Cost) ~ 0 + pdq(3,0,2) + PDQ(3,1,0)),
    auto = ARIMA(log(Cost), stepwise = FALSE, approx = FALSE)
```

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

```
## # A mable: 10 x 2
## # Key:     Model name [10]
##    `Model name`                     Orders
##    <chr>                           <model>
##  1 arima301012  <ARIMA(3,0,1)(0,1,2)[12]>
##  2 arima301111  <ARIMA(3,0,1)(1,1,1)[12]>
##  3 arima301011  <ARIMA(3,0,1)(0,1,1)[12]>
##  4 arima301110  <ARIMA(3,0,1)(1,1,0)[12]>
##  5 arima301210  <ARIMA(3,0,1)(2,1,0)[12]>
##  6 arima300111  <ARIMA(3,0,0)(1,1,1)[12]>
##  7 arima300110  <ARIMA(3,0,0)(1,1,0)[12]>
##  8 arima302210  <ARIMA(3,0,2)(2,1,0)[12]>
##  9 arima302310  <ARIMA(3,0,2)(3,1,0)[12]>
## 10 auto         <ARIMA(2,1,3)(0,1,1)[12]>
```

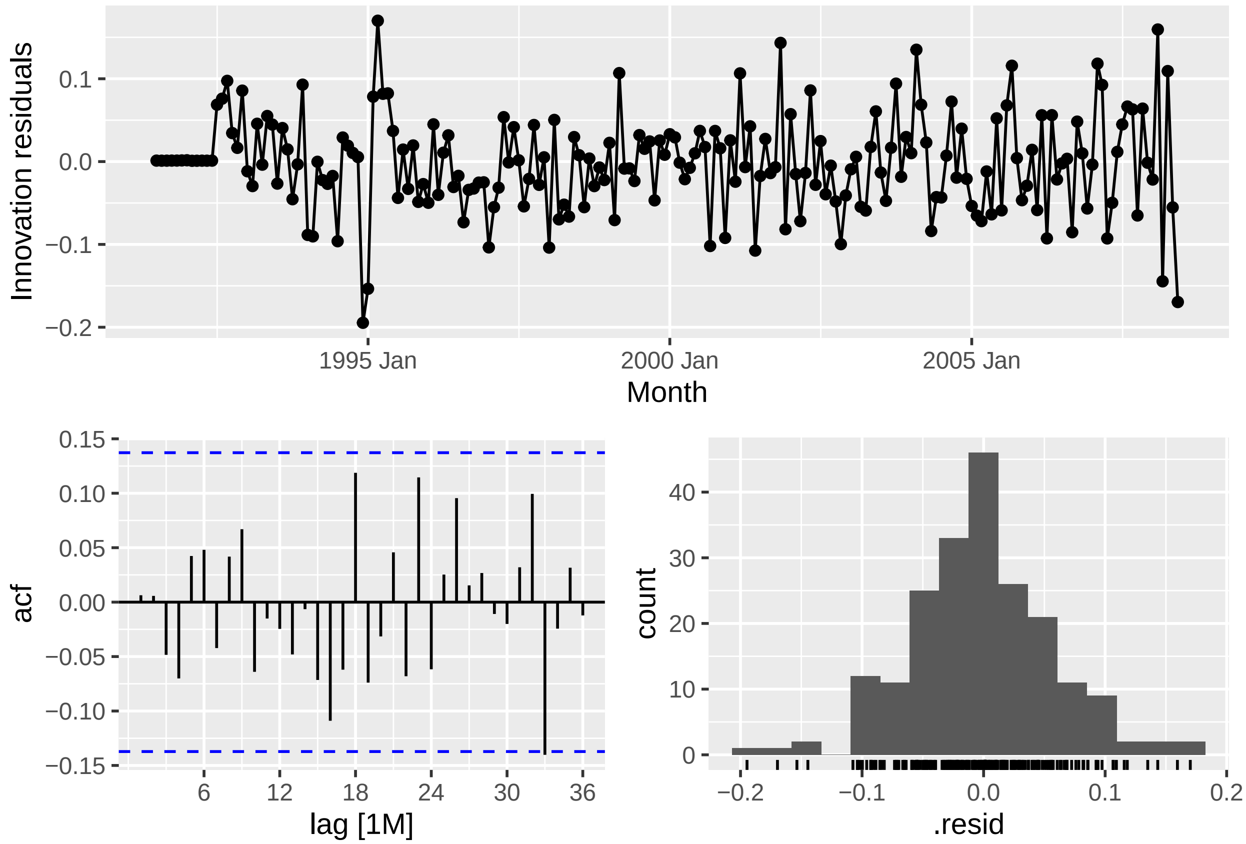
```r
glance(fit) |> arrange(AICc) |> select(.model:BIC)
```

```
## # A tibble: 10 x 6
##    .model       sigma2 log_lik   AIC  AICc   BIC
##    <chr>         <dbl>   <dbl> <dbl> <dbl> <dbl>
##  1 arima302310 0.00374    262. -507. -506. -477.
##  2 auto        0.00383    259. -504. -503. -481.
##  3 arima300111 0.00396    256. -501. -500. -481.
##  4 arima301012 0.00394    257. -501. -500. -478.
##  5 arima301111 0.00398    256. -499. -498. -476.
##  6 arima301011 0.00406    254. -497. -496. -477.
##  7 arima301210 0.00429    251. -489. -488. -466.
##  8 arima302210 0.00428    252. -489. -488. -463.
##  9 arima300110 0.00469    243. -476. -476. -460.
## 10 arima301110 0.00471    243. -475. -474. -455.
```

Interpretation

Exploring various models allowed us to determine distinct values for AIC, AICc, and BIC for each model. Among these, the ARIMA model ARIMA$(3, 0, 2)(3, 1, 0)_{12}$ displayed the lowest values across AIC, AICc, and BIC, indicating its superior fit compared to the other models.

```r
fit <- C01 |>
  model(ARIMA(log(Cost) ~ 0 + pdq(3,0,2) + PDQ(3,1,0)))
fit |> gg_tsresiduals(lag_max=36)
```



Interpretation

There isn't a notable and substantial spike and the data pattern remains in line with a white noise pattern. To be sure, we use a Ljung-Box test, being careful to set the degrees of freedom to match the number of parameters in the model.

```
augment(fit) |>
  features(.innov, ljung_box, lag = 36, dof = 8)
```

```
## # A tibble: 1 x 3
##   .model                                                  lb_stat lb_pvalue
##   <chr>                                                     <dbl>     <dbl>
## 1 ARIMA(log(Cost) ~ 0 + pdq(3, 0, 2) + PDQ(3, 1, 0))        31.1     0.313
```

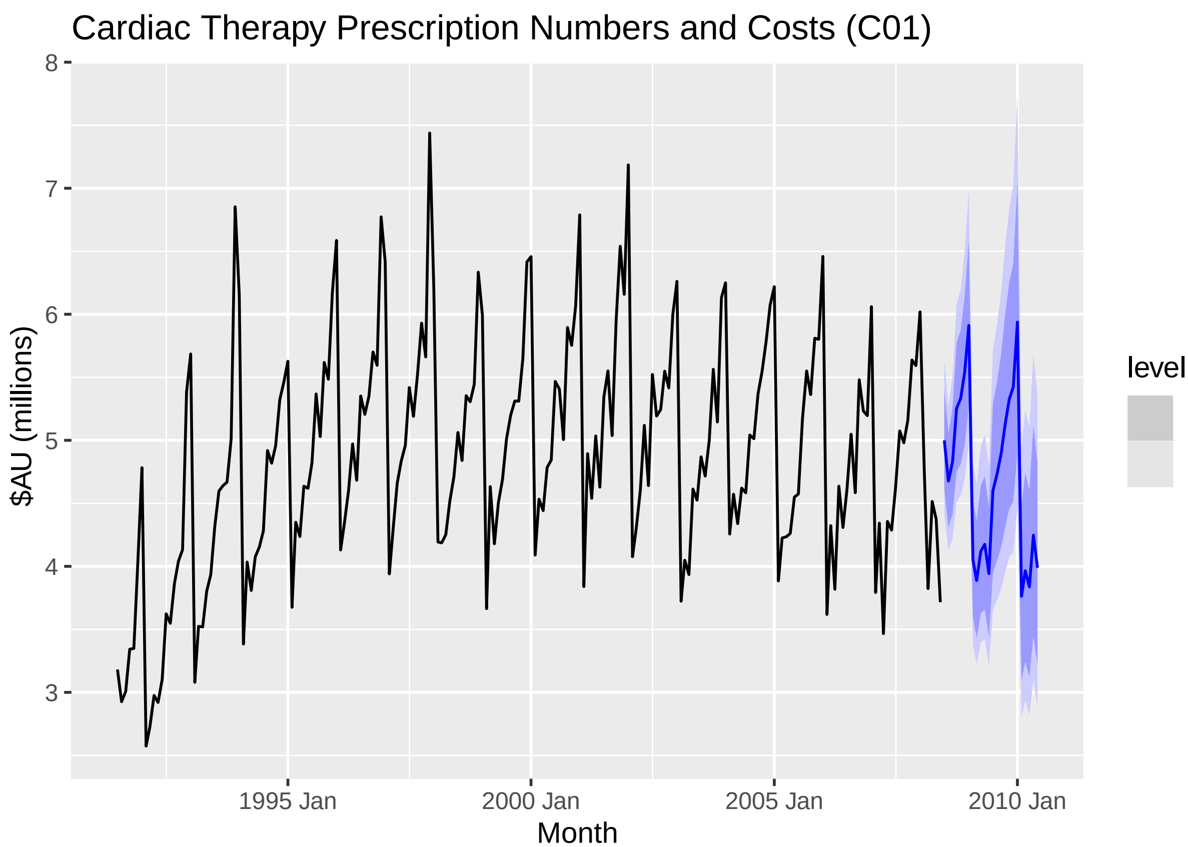
Interpretation

The high p-value provides confirmation that the residuals resemble white noise, meeting the criteria for our analysis.

Consequently, we've developed a seasonal ARIMA model that successfully clears the necessary assessments and is now primed for forecasting. The figure below depicts forecasts generated by the model for the upcoming years.

6 Test set evaluation

```
C01 |>
  model(ARIMA(log(Cost) ~ 0 + pdq(3,0,1) + PDQ(3,1,0))) |>
  forecast() |>
  autoplot(C01) +
  labs(y=" $AU (millions)",
       title="Cardiac Therapy Prescription Numbers and Costs (C01)")
```



Interpretation

The figure shows the chosen forecast model which is ARIMA$(3,0,2)(3,1,0)_{12}$

7 The equation for the ARIMA$(3,0,2)(3,1,0)_{12}$ model

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + b_{t-12} + \theta_{12,1} b_{t-12} + \theta_{12,2} b_{t-24} + \theta_{12,3} b_{t-36} \varepsilon_t$$

This equation represents the ARIMA(3,0,1)(1,1,1) model with a seasonality of 12, where

- Z_t is the observed time series.
- c is the constant term.
- ϕ_1, ϕ_2, ϕ_3 are the non-seasonal autoregressive coefficients.

- a_t is the white noise error term at time t.
- θ_1, θ_2 are the non-seasonal moving average coefficients.
- b_{t-12} is the seasonal lag-12 term.
- $\theta_{12,1}, \theta_{12,2}$ are the seasonal autoregressive coefficients.
- ε_t is the white noise error term.

Equation of the chosen model

```r
report(fit)
```

```
## Series: Cost
## Model: ARIMA(3,0,2)(3,1,0)[12]
## Transformation: log(Cost)
##
## Coefficients:
##           ar1     ar2     ar3     ma1     ma2     sar1     sar2     sar3
##       -0.1670  0.7017  0.4488  0.2577  -0.356  -0.5041  -0.4503  -0.3986
## s.e.   0.1405  0.1198  0.0944  0.1498   0.134   0.0773   0.0800   0.0816
##
## sigma^2 estimated as 0.003744:  log likelihood=262.39
## AIC=-506.78   AICc=-505.79   BIC=-477.46
```

By replacing the values in the given equation, we obtain the equation representing the chosen model.

$$Z_t = c - 0.1670Z_{t-1} + 0.7017Z_{t-2} + 0.4488Z_{t-3} + a_t + 0.2577a_{t-1} - 0.356a_{t-2}$$
$$-0.5041_{12,1}b_{t-12} - 0.4503_{12,2}b_{t-24} - 0.3986_{12,3}b_{t-36} + \varepsilon_t$$