Pranav Kasela 846965

Arima Models

```r
library(ggplot2)
library(forecast)
library(tidyverse)
```

```r
df <- read.csv2("time_series_dataset.csv", dec = ".")
df$Data <- as.Date(df$Data)
head(df)
```

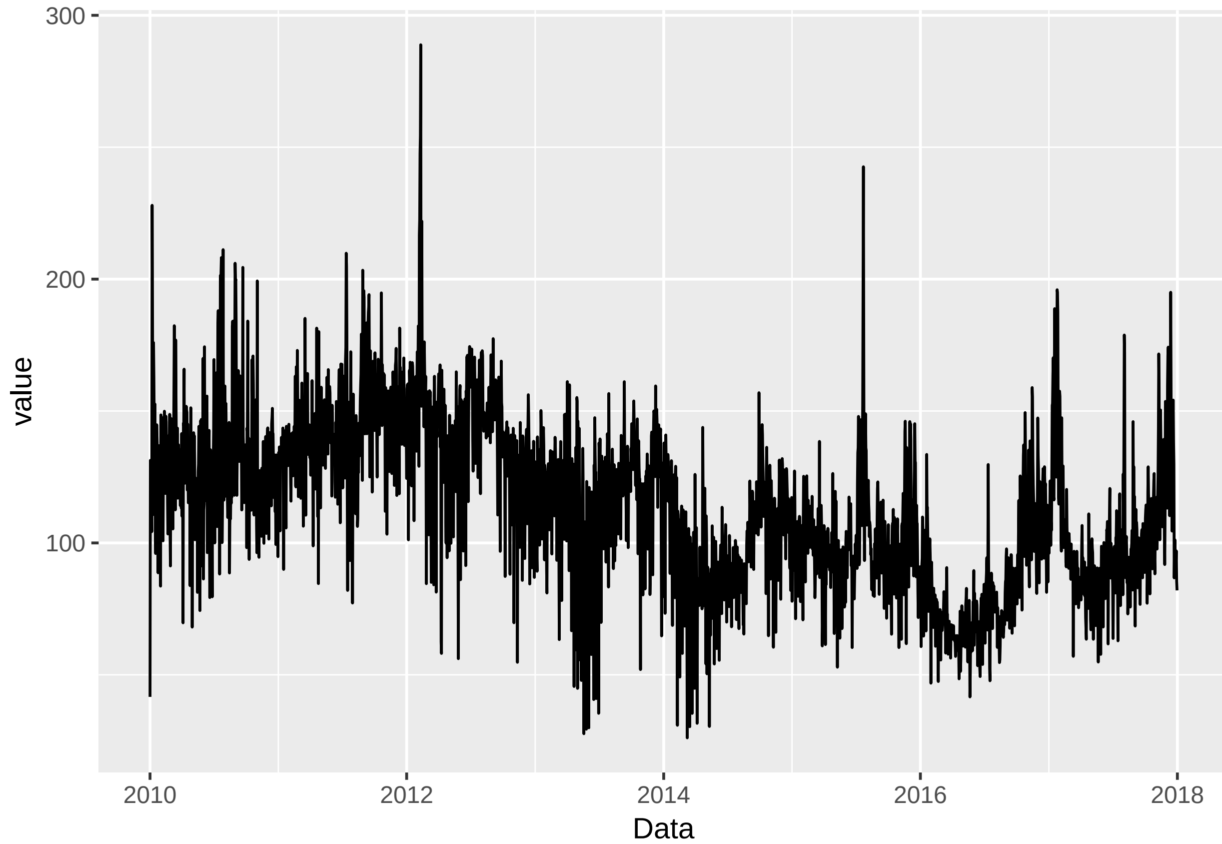
```
##         Data     value
## 1 2010-01-01  41.65104
## 2 2010-01-02 131.28660
## 3 2010-01-03 117.38812
## 4 2010-01-04 116.46128
## 5 2010-01-05 123.82376
## 6 2010-01-06 104.28556
```

```r
#Trying Augmeted Dickey-Fuller test to see if the series is stationary:
#$H_0$ is that the model is not stationary
tseries::adf.test(df$value)
```

```
## Warning in tseries::adf.test(df$value): p-value smaller than printed p-
## value
```

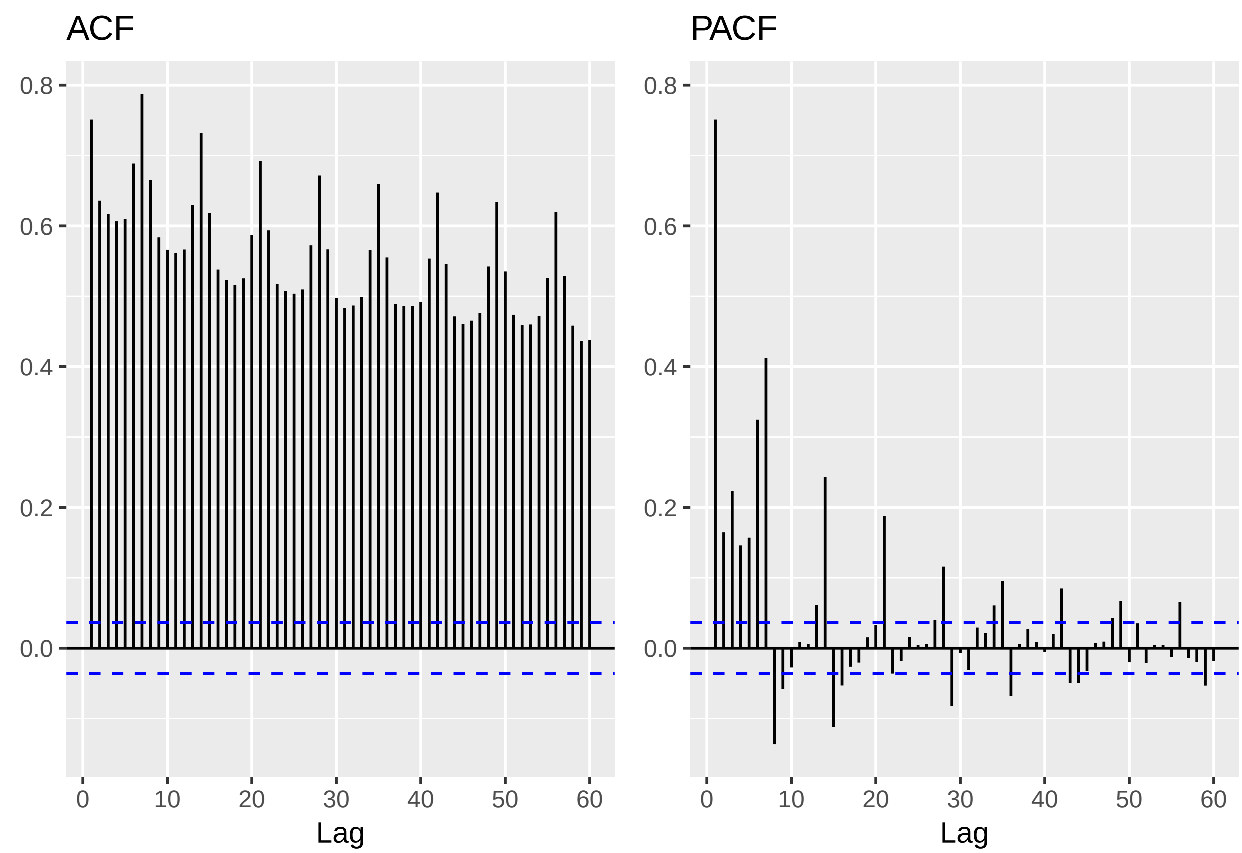
```
##
##  Augmented Dickey-Fuller Test
##
## data:  df$value
## Dickey-Fuller = -4.8605, Lag order = 14, p-value = 0.01
## alternative hypothesis: stationary
```

```r
train <- df[1:(nrow(df) - 365),] #last one year is for validation
test  <- df[(nrow(df) - 365 + 1):nrow(df),]

ggplot(data=train, aes(x=Data, y=value)) +
  geom_line()
```



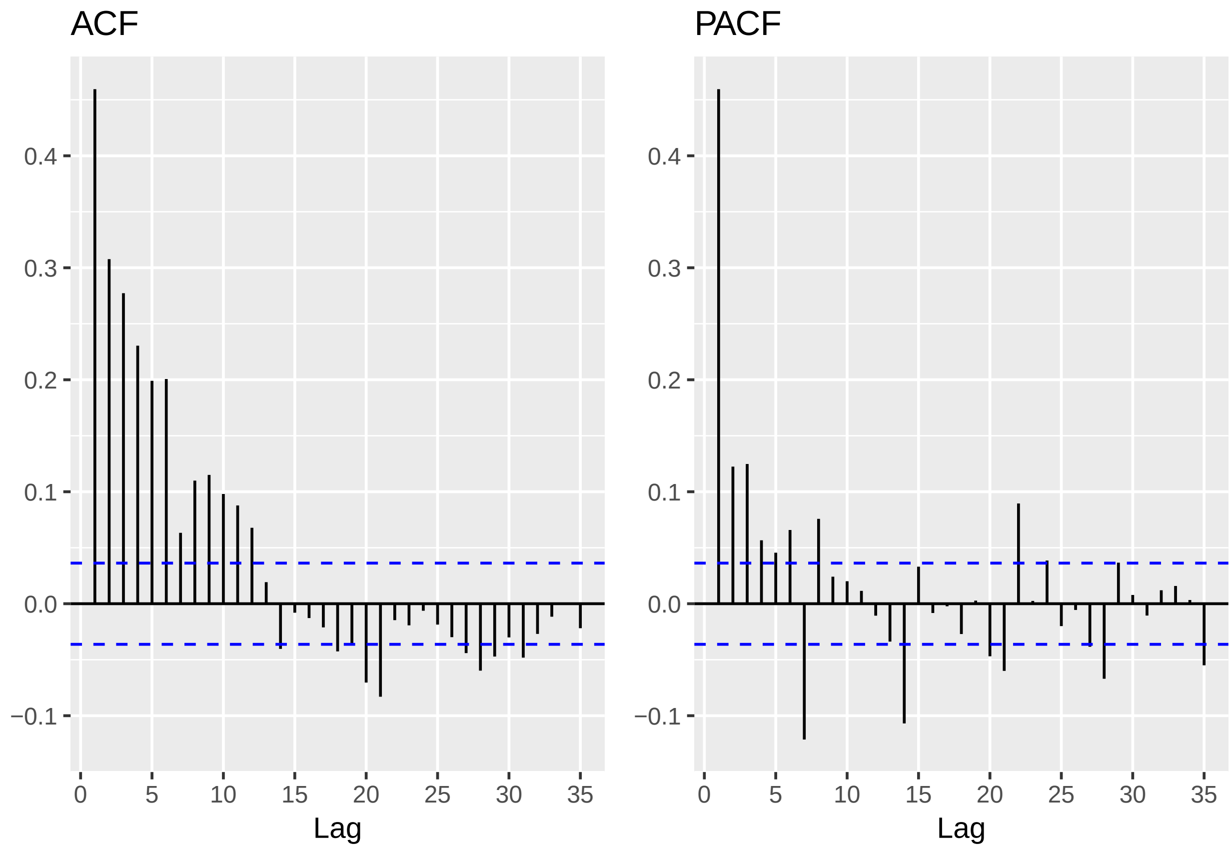

```
#This is the function ggtsdisplay of forecast package,
#but it has been modified so is doesn't plot the series,
#just the ACF and PACF plot, with the horizontal parameter
#the plot can be either horizontal or vertical
#The function has been simplified a lot, since we don't need
#all the complexity the original one has.
ggtsdisplay_2 <- function(x, lag.max, horizontal=TRUE, ...) {
    if (!is.ts(x)) {
      x <- ts(x)
    }
    if (missing(lag.max)) {
      lag.max <- round(min(max(10 * log10(length(x)), 3 * frequency(x)), length(x) / 3))
    }
    ######      END   CHECKING    ########

    # Set up grid for plots
    if (horizontal){
      gridlayout <- matrix(c(2, 3), nrow = 1)
    }
    else{
      gridlayout <- matrix(c(2, 3), nrow = 2)
    }
    grid::grid.newpage()
    grid::pushViewport(grid::viewport(layout = grid::grid.layout(nrow(gridlayout), ncol(gridlayout))))

    # Prepare Acf plot
```

```
    acfplot <- do.call(ggAcf, c(x = quote(x), lag.max = lag.max)) +
      ggplot2::ggtitle("ACF") + ggplot2::ylab(NULL)

    # Prepare last plot (variable)
    pacfplot <- ggPacf(x, lag.max = lag.max) + ggplot2::ggtitle("PACF") +
      ggplot2::ylab(NULL)
    # Match y-axis
    acfplotrange <- ggplot2::layer_scales(acfplot)$y$range$range
    pacfplotrange <- ggplot2::layer_scales(pacfplot)$y$range$range
    yrange <- range(c(acfplotrange, pacfplotrange))
    acfplot <- acfplot + ggplot2::ylim(yrange)
    pacfplot <- pacfplot + ggplot2::ylim(yrange)

    # Add ACF plot
    matchidx <- as.data.frame(which(gridlayout == 2, arr.ind = TRUE))
    print(
      acfplot,
      vp = grid::viewport(
        layout.pos.row = matchidx$row,
        layout.pos.col = matchidx$col
      )
    )

    # Add PACF plot
    matchidx <- as.data.frame(which(gridlayout == 3, arr.ind = TRUE))
    print(
      pacfplot,
      vp = grid::viewport(
        layout.pos.row = matchidx$row,
        layout.pos.col = matchidx$col
      )
    )

}

ggtsdisplay_2(train$value, horizontal = TRUE, lag.max = 60)
```



Possiamo vedere che nel PACF vi sono 7 ritardi e il ritrado stagionale che scende esponenzialemente al settimo ritardo, indicando la presenza di un SMA$(1)_7$ e vedendo ACF dai primi 2 ritardi stagionali ci convinciamo dell'esistenza di SMA$(1)_7$, inoltre vi è presente anche un SAR$(1)_7$. Stagionalità 7 indica un periodo settimanale in questa serie. Inoltre vista la discesa lenta e non geometrica della ACF potrebbe suggerire l'esistenza di una integrazione stagionale. Iniziamo ad aggiungere la parte stagionale prima e cerchiamo di capire dai residui come andrebbe aggiustato il modello.

```
mod1 <- Arima(train$value, c(0,0,0), list(order=c(1,0,1), period=7), lambda = "auto")
ggtsdisplay_2(mod1$residuals)
```

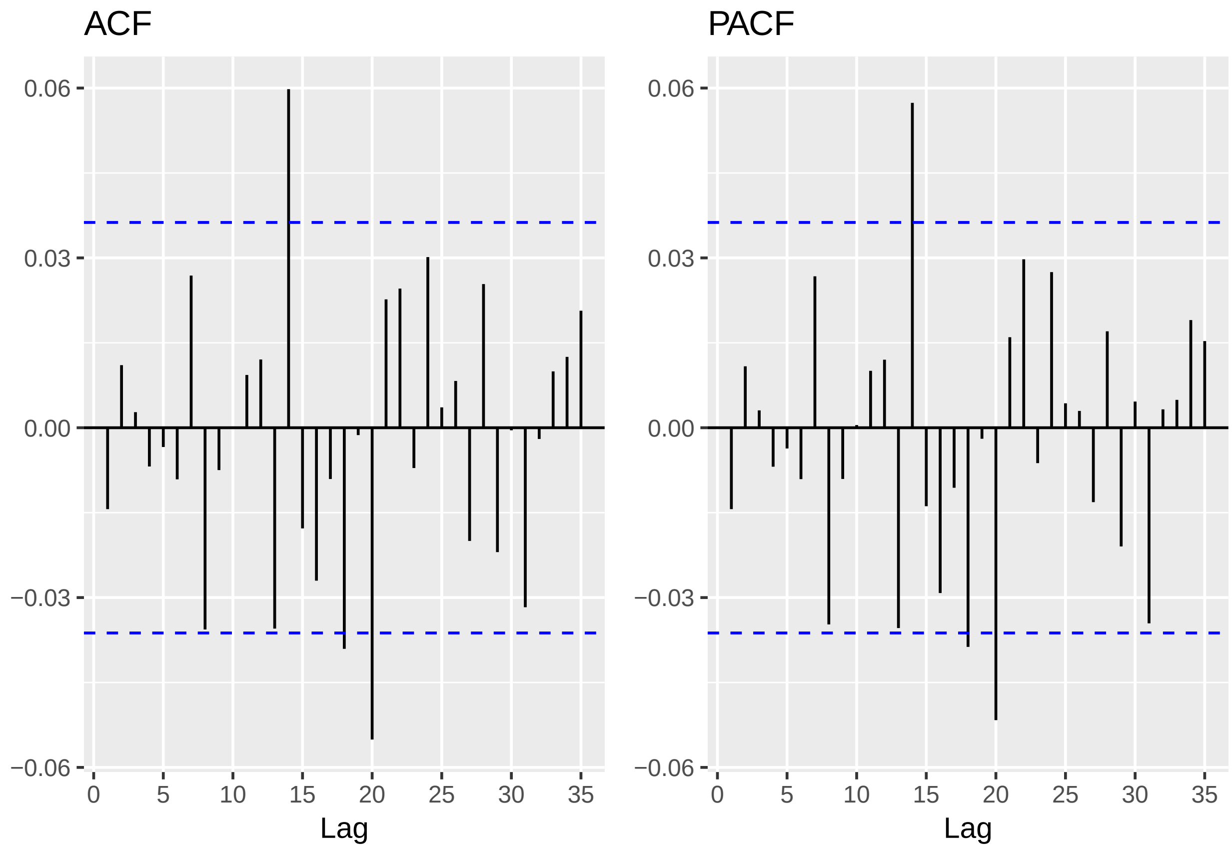



```
mod1
```

```
## Series: train$value
## ARIMA(0,0,0)(1,0,1)[7] with non-zero mean
## Box Cox transformation: lambda= 0.9437982
##
## Coefficients:
##         sar1     sma1     mean
##       0.9591  -0.5557  92.6251
## s.e.  0.0063   0.0229   2.5938
##
## sigma^2 estimated as 183.9:  log likelihood=-11767.49
## AIC=23542.97   AICc=23542.98   BIC=23566.89
```

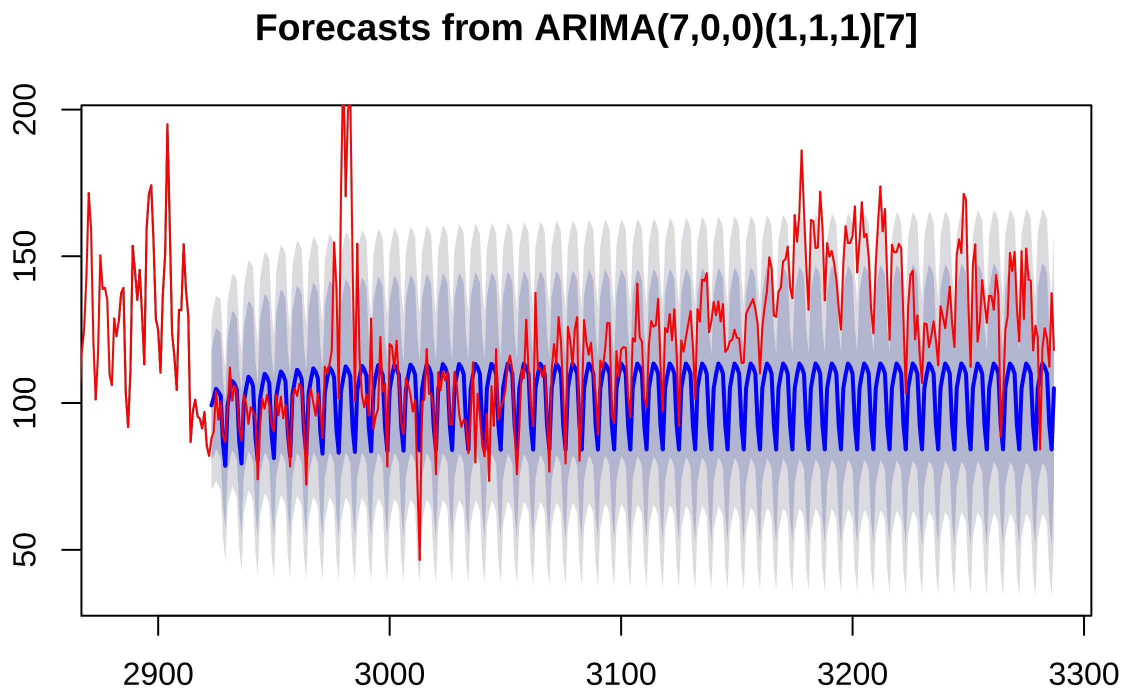
Vediamo anche che il coefficiente di SAR è molto vicino ad 1, quindi ha radice unitaria e ciò dice che esiste l'integrazione stagionale che sospettavamo prima.

```
mod2 <- Arima(train$value, c(7,0,0), list(order=c(1,1,1), period=7), lambda = "auto")
ggtsdisplay_2(mod2$residuals)
```



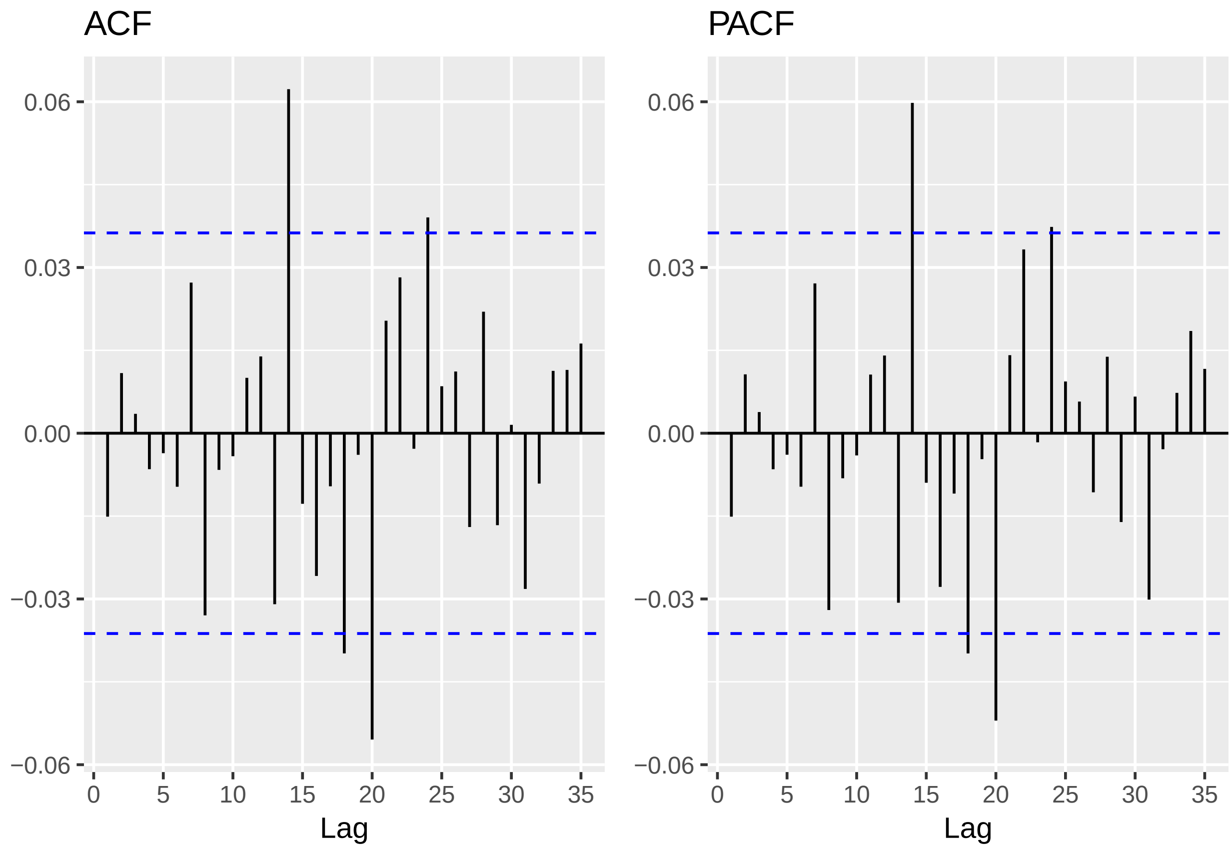
I residui sembrano essere rientrati nella banda tranne un residuo a 24 sia un ACF che PACF.

```r
plot(forecast(mod2, h=365), include = 40)
lines(df$value, col="red")
```

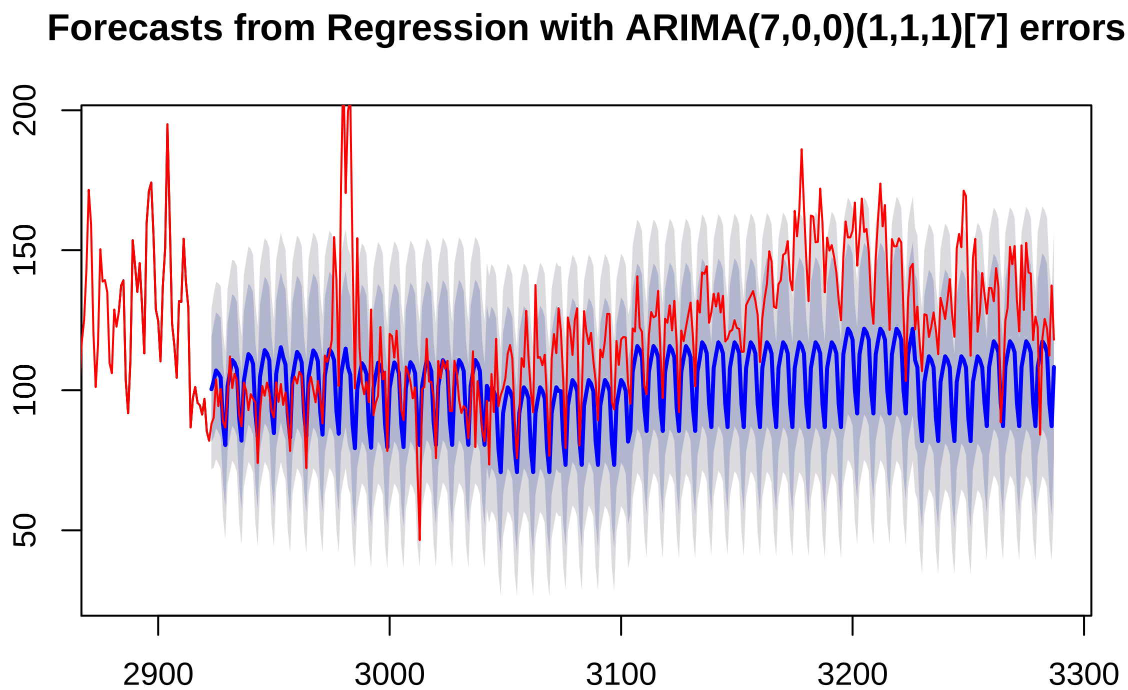


Mettiamo regressori sinusoidali e dummy settimanali.

```
fr <- 2*pi*outer(1:nrow(df), 1:6)/12
co <- cos(fr)
si <- sin(fr[,1:5])
colnames(co) <- paste0("cos",1:6)
colnames(si) <- paste0("sin",1:5)
xreg <- cbind(co, si)

#create dummy
df %>%
  mutate(month = months(df$Data), ind = 1) %>%
  spread(month, ind, fill = 0) -> more_reg
  #mutate(day = weekdays(df$Data), ind = 1) %>%
  #spread(day, ind, fill = 0)

#xreg <- as.matrix(cbind(xreg, more_reg[3:(ncol(more_reg)-1)]))
xreg <- as.matrix(more_reg[4:(ncol(more_reg)-1)])

mod1_reg <- Arima(train$value, c(7,0,0), list(order=c(1,1,1), period=7),
              xreg=xreg[1:(nrow(df)-365),])
ggtsdisplay_2(mod1_reg$residuals)
```




```
plot(forecast(mod1_reg, h=365,
              xreg=xreg[(nrow(df)-365+1):nrow(df),]),
     include = 40)
lines(df$value, col="red")
```


Forecasts from Regression with ARIMA(7,0,0)(1,1,1)[7] errors
200
150
100
50
2900
3000
3100
3200
3300