## Time Series Analysis in Health Research

Forecasting of Weekly mortality rate in Saskatchewan

Erfan Hoque
Dept. of Community Health and Epidemiology
University of Saskatchewan.

SEA 24th Annual Symposium Workshop, September 26, 2025

We will work in fpp3 package. Load the library fpp3. If fpp3 package is not installed, then we need to install it first and then load the library.

```
library(fpp3)
library(dplyr)
library(fable.prophet) # needed for Prophet model
```

This document is intended to serve as a rough template for what we covered in the last lectures on time series analysis in R.

## Weekly mortality Data

#### **Data Discription:**

This dataset (weekly mortality Data) provides weekly age-standardized mortality rates across Canadian provinces, disaggregated by sex, beginning in 2019. The rates are standardized using the 2011 Canadian population, allowing for consistent comparisons across regions and time periods regardless of demographic shifts.

Publisher: Statistics Canada Release https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310087901

Frequency: Weekly

Geography: Canada and Provinces (excluding Territories)

#### Set Directory

To check your current working directory in R, use: getwd()

This will return the path of the directory where R is currently reading and writing files like this

getwd()

#### Read Data

To read a different data you need to change the name of the data based on the data file name.

```
mortality <- read.csv("mortality_model_SK.csv")</pre>
```

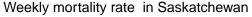
### tisbble objects

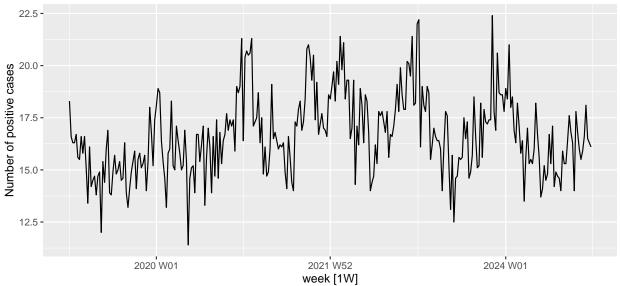
We need to create the data in *tisbble* object to use the functions from *fpp3* package.

```
## # A tsibble: 10 x 5 [1W]
##
     Date
                   week Rate pre_CA tem_CA
##
     <date>
                  <week> <dbl> <dbl>
   1 2019-01-05 2019 W01 18.3 0.0285
                                     -15.0
## 2 2019-01-12 2019 W02 16.6 0.0443 -16.6
## 3 2019-01-19 2019 W03 16.3 0.0259
                                     -25.2
## 4 2019-01-26 2019 W04 16.3 0.0366
                                     -20.0
## 5 2019-02-02 2019 W05
                         16.7 0.0452
                                     -25.6
## 6 2019-02-09 2019 W06 15.6 0.0161 -30.4
## 7 2019-02-16 2019 W07
                         15.5 0.0268 -22.4
## 8 2019-02-23 2019 W08 16.6 0.0142 -19.9
## 9 2019-03-02 2019 W09 15.8 0.0165 -21.4
## 10 2019-03-09 2019 W10 16.6 0.0103 -15.4
```

## Time series plot, patterns and decomposition

### Time plot



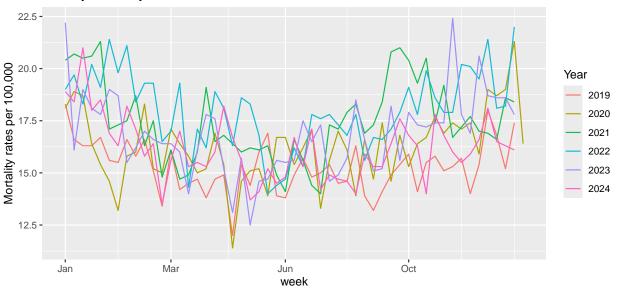


We can see some short trend and seasonality in the mortality rate.

### Seasonal plots

A seasonal plot is similar to a time plot except that the data are plotted against the individual "seasons" in which the data were observed. This enables the underlying seasonal pattern to be seen more clearly, and also allows any substantial departures from the seasonal pattern to be easily identified.

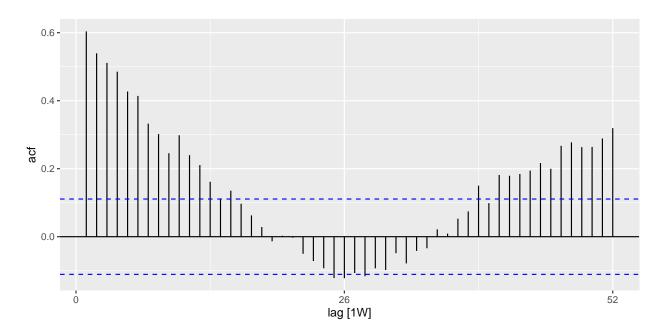
#### Weekly mortality rate in Saskatchewan



#### Autocorrelation

When data have a trend, the autocorrelations for small lags tend to be large and positive. When data are seasonal, the autocorrelations will be larger at the seasonal lags (i.e., at multiples of the seasonal frequency). When data are trended and seasonal, we see a combination of these effects.

```
mortality_ts %>%
  ACF(Rate, lag_max = 52) %>% autoplot()
```



## Time series decomposition

STL (Seasonal and Trend decomposition using Loess) is a versatile and robust method for decomposing time series.

```
dcmp <- mortality_ts %>%
  model(stl = STL(Rate))
components(dcmp)
```

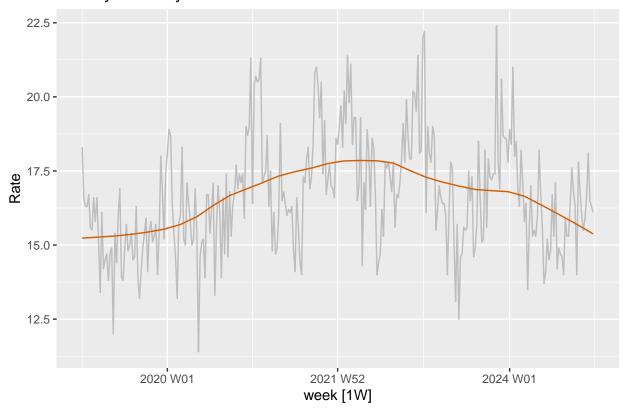
```
## # A dable: 313 x 7 [1W]
##
  # Key:
               .model [1]
##
  #:
              Rate = trend + season_year + remainder
##
                  week Rate trend season_year remainder season_adjust
      .model
##
      <chr>
                <week> <dbl> <dbl>
                                          <dbl>
                                                     <dbl>
                                                                    <dbl>
             2019 W01
                        18.3
                              15.2
                                           1.79
                                                    1.28
                                                                     16.5
##
    1 stl
##
    2 stl
             2019 W02
                        16.6
                              15.2
                                          2.63
                                                   -1.26
                                                                     14.0
##
    3 stl
             2019 W03
                        16.3
                              15.2
                                          1.94
                                                   -0.877
                                                                     14.4
             2019 W04
                        16.3
                              15.2
                                                   -0.577
                                                                     14.7
##
    4 stl
                                          1.63
##
    5 stl
             2019 W05
                        16.7
                               15.2
                                          1.57
                                                   -0.115
                                                                     15.1
##
    6 stl
             2019 W06
                        15.6
                              15.3
                                          1.09
                                                   -0.738
                                                                     14.5
    7 stl
             2019 W07
                        15.5
                              15.3
                                          0.334
                                                   -0.0877
                                                                     15.2
             2019 W08 16.6 15.3
                                          0.838
                                                    0.504
                                                                     15.8
##
    8 stl
```

```
## 9 stl 2019 W09 15.8 15.3 0.691 -0.153 15.1 ## 10 stl 2019 W10 16.6 15.3 1.17 0.168 15.4 ## # i 303 more rows
```

#### Trend-adjustment

```
mortality_ts %>%
  autoplot(Rate, color='gray') +
  autolayer(components(dcmp), trend, color='#D55E00') +
  labs(y = "Rate", title = "Weekly mortality rate in Saskatchewan")
```

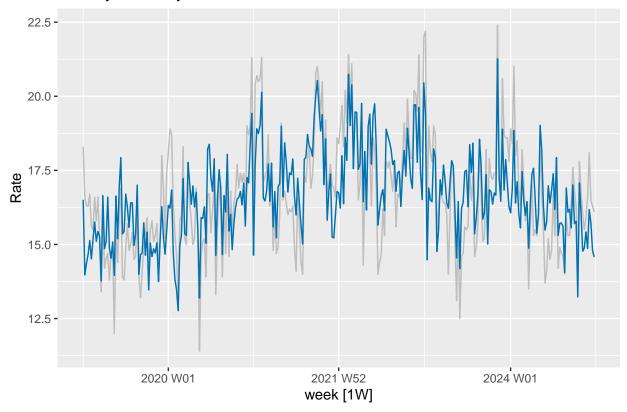
## Weekly mortality rate in Saskatchewan



#### Seasonal-adjustment

```
mortality_ts %>%
  autoplot(Rate, color='gray') +
  autolayer(components(dcmp), season_adjust, color='#0072B2') +
  labs(y = "Rate", title = "Weekly mortality rate in Saskatchewan")
```

## Weekly mortality rate in Saskatchewan

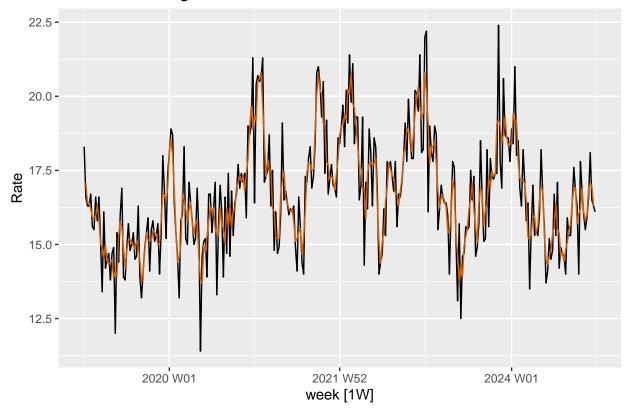


### Classical Decomposition

The traditional way to do time series decomposition is called Classical decomposition. The simplest estimate of the trend-cycle uses *moving averages* which is an average of nearby points.

## Warning: Removed 2 rows containing missing values or values outside the scale range
## ('geom\_line()').

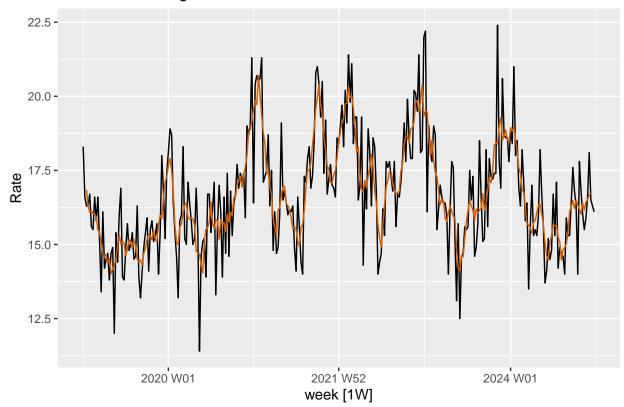
## 3-MA smoothing



```
# plot of 5-MA
mortality_ts_decom |>
  autoplot(Rate) +
  geom_line(aes(y = `5-MA`), colour = "#D55E00")+
  labs(title = "5-MA smoothing")
```

## Warning: Removed 4 rows containing missing values or values outside the scale range
## ('geom\_line()').





## Forecasting Models

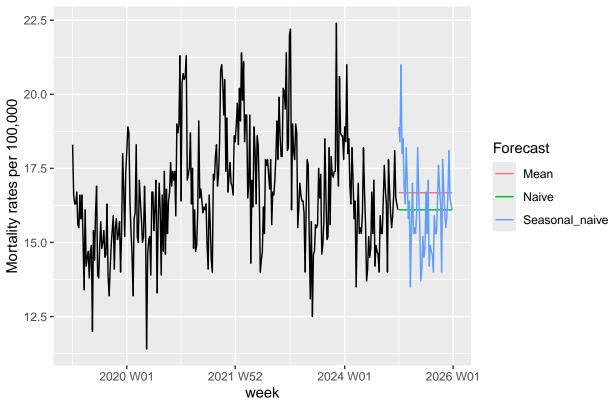
### Simple forecasting methods

The model() function trains models to data. We are going to forecast using Mean, Naive and Seasonal Naive method.

```
mortality_fit <- mortality_ts %>%
  model(
    Seasonal_naive = SNAIVE(Rate),
    Naive = NAIVE(Rate),
    Mean = MEAN(Rate)
)
```

We can now produce forecasts using the fitted models.

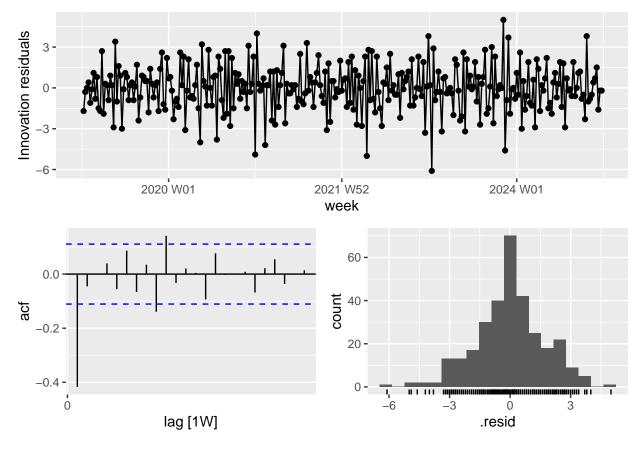




## Residual diagnostics

It is very important to do the residual diagnostic after fitting any model to check whether the residual assumptions have been satisfied or not.

```
fit.naive <- mortality_ts %>% model(NAIVE(Rate))
gg_tsresiduals(fit.naive)
```



Here, based on the plots, we can say the residual assumptions (uncorrelated, mean zero, constant variance) have not been satisfied for Naive model.

#### Tests for autocorrelation

Moreover, we can do Box-Pierce or Ljung-Box test to see whether the residuals are significantly different from a zero set.

H\_0: the series is white noise vs H\_1: the series is not white noise.

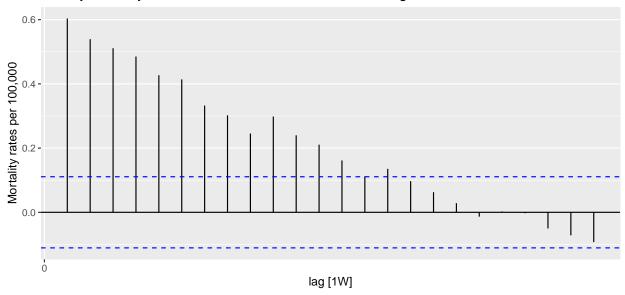
The results are significant (i.e., the p-values are relatively small < 0.05). Thus, we can conclude that the residuals are distinguishable from a white noise series.

### ARIMA / SARIMA models

Before fitting the ARIMA model, it is important to check the stationary of the data. If the data is non-stationary, then we need to make it stationary using the idea of differencing. Differencing helps to stabilize the mean (one way to make a non-stationary time series stationary).

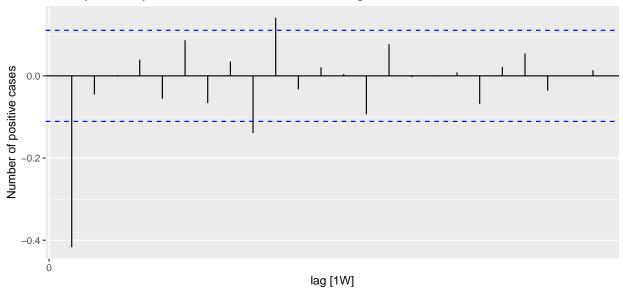
```
mortality_ts %>% ACF(Rate) %>% autoplot() +
    labs(title = "Weekly mortality rate in Saskatchewan without differencing",
    y = "Mortality rates per 100,000")
```

#### Weekly mortality rate in Saskatchewan without differencing



```
mortality_ts %>% ACF(difference(Rate)) %>% autoplot() +
    labs(title = "Weekly mortality in Saskatchewan with differencing",
    y = "Number of positive cases")
```

#### Weekly mortality in Saskatchewan with differencing



First order differencing seems make the mortality data stationary. But still there are some pattern left over.

Now let's fit the ARIMA/SARIMA model. The ARIMA frunction form fpp3 package does everything together (checking for stationarity of the data, doing any differencing if needed and then fitting the model using ARIMA or SARIMA based on data)

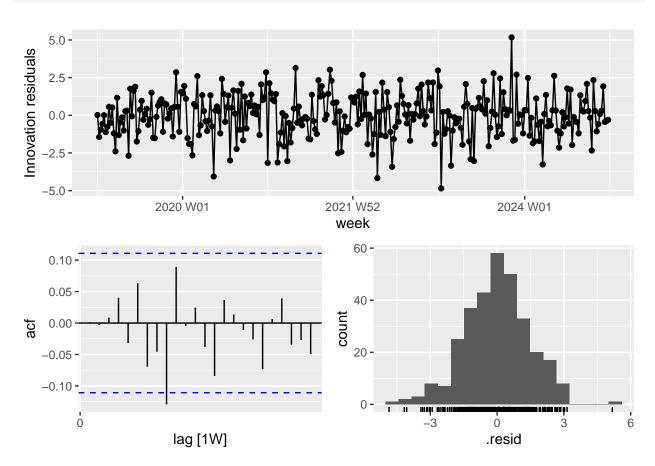
```
fit.arima <- mortality_ts |>
  model(ARIMA(Rate))
report(fit.arima)
```

```
## Series: Rate
## Model: ARIMA(1,1,1)(0,0,1)[52]
##
##
  Coefficients:
##
##
         0.0877
                 -0.6961
                          0.111
##
         0.0932
                  0.0711
                          0.064
##
## sigma^2 estimated as 2.129: log likelihood=-559.69
## AIC=1127.38
                 AICc=1127.51
                                 BIC=1142.35
```

We have SARIMA(1,0,3)(0,0,1)[52] model for this data.

Let's check the residual.

#### gg\_tsresiduals(fit.arima)



Here, based on the plots, we can say the residual assumptions (uncorrelated, mean zero, constant variance) have been satisfied for ARIMA model (still some pick is there!).

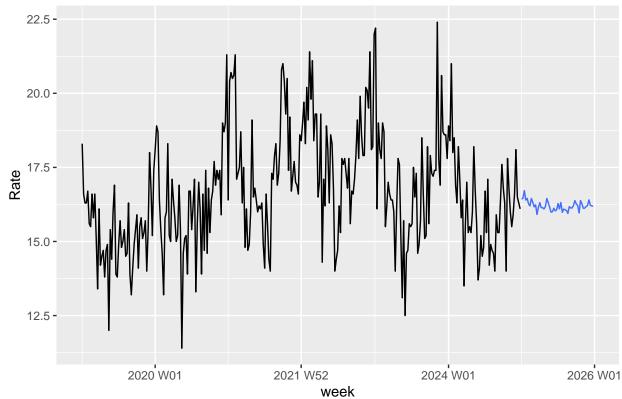
```
augment(fit.arima) %>%
features(.innov, ljung_box, lag = 36, dof=4)
```

The results are not significant (i.e., the p-values are relatively large). Thus, we can conclude that the residuals are not distinguishable from a white noise series.

Let's forecast mortality rate for next one year (52 weeks).

```
fit.arima %>% forecast(h = 52) %>% # h = "1 years"
autoplot(mortality_ts, level=NULL) +
labs(y = "Rate", title = "Weekly mortality rate in Saskatchewan using SARIMA")
```

## Weekly mortality rate in Saskatchewan using SARIMA



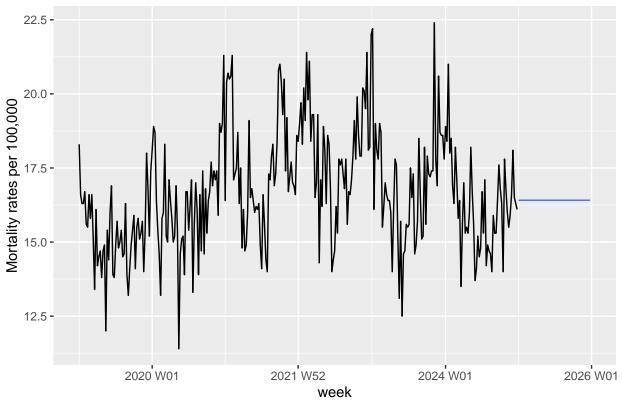
### Simple Exponential Smoothing

```
fit.ets <- mortality_ts %>% model(ETS(Rate))
report(fit.ets)
```

```
## Series: Rate
## Model: ETS(M,N,N)
##
     Smoothing parameters:
##
       alpha = 0.3615107
##
##
     Initial states:
##
        1[0]
    16.96287
##
##
##
     sigma^2:
                0.0076
##
##
        AIC
                 AICc
                            BIC
## 2032.487 2032.565 2043.725
```

```
fit.ets %>%
  forecast(h = "1 years") %>%
  autoplot(mortality_ts, level=NULL)+
  labs(title="Weekly mortality rate in Saskatchewan using ETS", y="Mortality rates per 100,000")
```

## Weekly mortality rate in Saskatchewan using ETS



### Neural Netwrok Model

Now, we will use Neural network autoregression model to forecast mortality rate using NNETAR() function that fits an NNAR $(p, P, k)_m$  model. If p and P are not specified, they are automatically selected.

```
fit.nn <- mortality_ts %>% model(NNETAR(Rate))
fit.nn

## # A mable: 1 x 1
## 'NNETAR(Rate)'
```

The result provides a NNAR(6,1,4)[52] model for mortality rate. Here, the last 6 observations are used as predictors, and there are 4 neurons in the hidden layer.

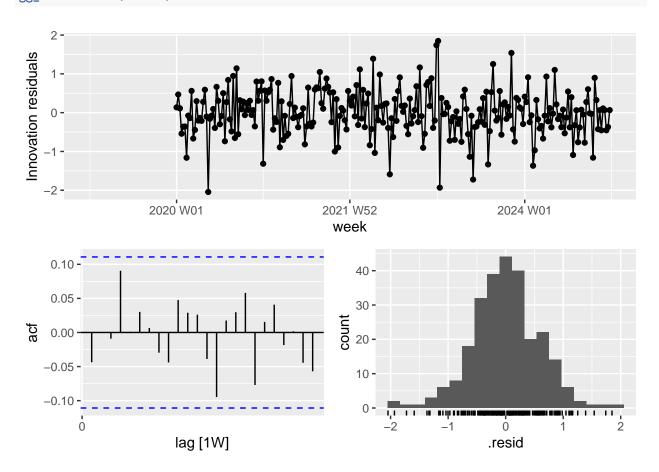
Let's check the residual.

##

#### gg\_tsresiduals(fit.nn)

## 1 <NNAR(10,1,6)[52]>

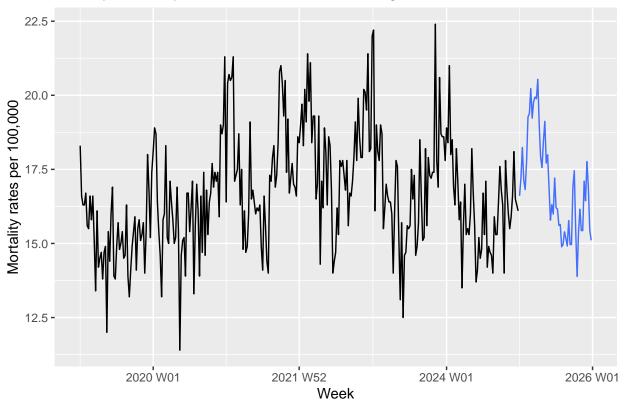
<model>



Let's forecast mortality rate for next one year (52 weeks).

```
fit.nn %>% forecast(h="1 years", times=1) %>%
  autoplot(mortality_ts, level=NULL) +
  labs(x = "Week", y = "Mortality rates per 100,000", title = "Weekly mortality rate in Saskatchewan us
```

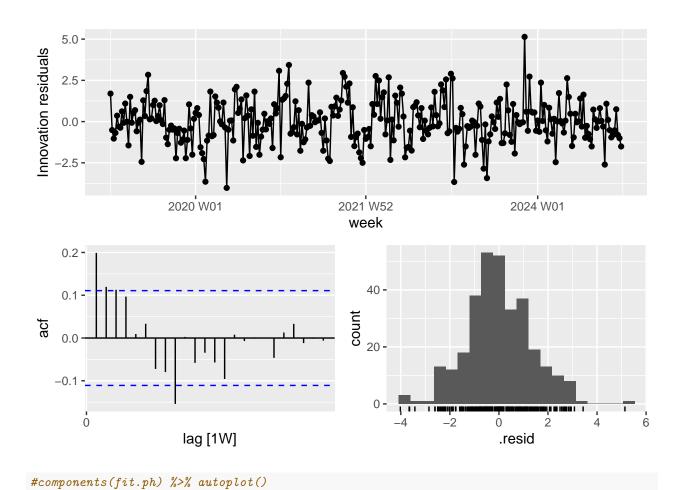




## Prophet Model

Prophet is an open-source forecasting tool developed by Facebook for time series forecasting. Prophet model is available via the fable.prophet package.

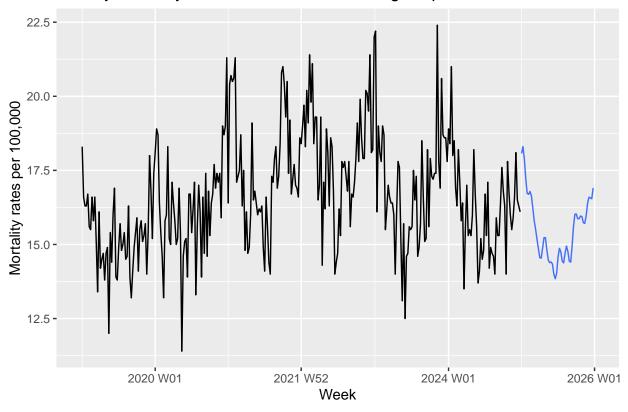
```
fit.ph <- mortality_ts %>%
  model(prophet(Rate ~ season(period = 52, order = 10)))
fit.ph |> gg_tsresiduals() # check residual
```



Let's forecast mortality rate for next one year (52 weeks).

```
fit.ph %>% forecast(h = "1 years") %>%
  autoplot(mortality_ts, level=NULL) +
  labs(x = "Week", y = "Mortality rates per 100,000", title = "Weekly mortality rate in Saskatchewan us
```

## Weekly mortality rate in Saskatchewan using Prophet



### Forecast accuracy: Comparing all models

We can fit all model together in one function and compare them based on some accuracy measures. Then choose the best one and forecast based on the best model.

```
mortality_fit <- mortality_ts %>%
model(
    Seasonal_naive = SNAIVE(Rate),
    Naive = NAIVE(Rate),
    Mean = MEAN(Rate),
    Arima = ARIMA(Rate),
    ETS = ETS(Rate),
    NNAR = NNETAR(Rate),
    #Prophet = prophet(FluPos)
    Prophet = prophet(Rate ~ season(period = 52, order = 10))
)
```

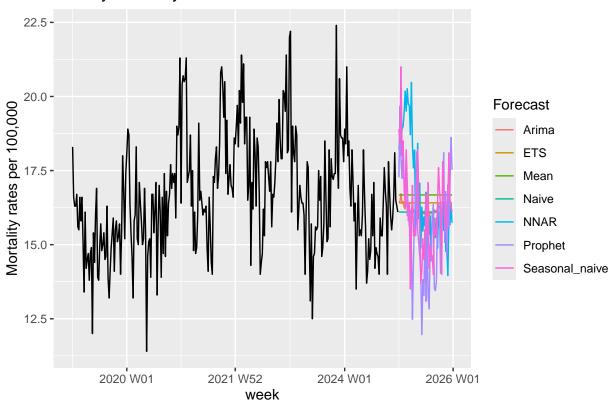
We can now produce forecasts for next 1 years (52 weeks) using the fitted models.

```
mortality_fc <- mortality_fit %>%
  forecast(h = "1 years", times=1)

mortality_fc %>%
  autoplot(mortality_ts, level = NULL) +
```

```
labs(title = "Weekly mortality reate in Saskatchewan",
    y = "Mortality rates per 100,000") +
guides(colour = guide_legend(title = "Forecast"))
```

### Weekly mortality reate in Saskatchewan



Let's compare all models.

### accuracy(mortality\_fit)

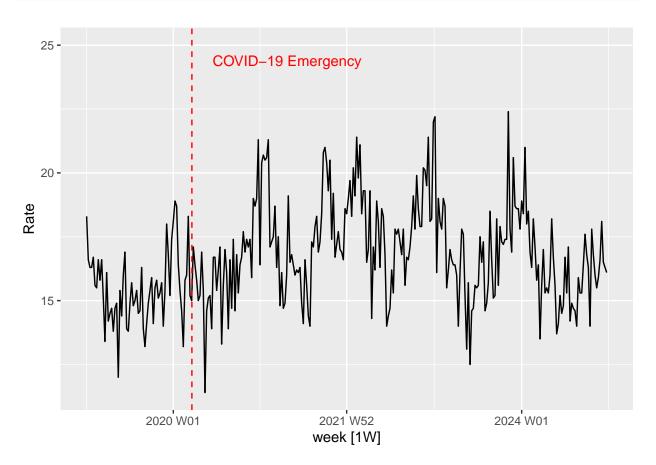
```
## # A tibble: 7 x 10
##
     .model
                                        RMSE
                                                                  MASE RMSSE
                                                                                  ACF1
                     .type
                                                MAE
                                                            MAPE
##
     <chr>
                     <chr>
                                 <dbl> <dbl> <dbl>
                                                           <dbl> <dbl> <dbl>
                                                                                 <dbl>
                                                     <dbl>
## 1 Seasonal_naive Traini~
                              1.77e- 1 2.20
                                              1.78
                                                     0.283 10.4
                                                                               0.401
## 2 Naive
                    Traini~ -7.05e- 3 1.71
                                              1.31
                                                    -0.574
                                                            7.96 0.735 0.776 -0.417
## 3 Mean
                    Traini~ -1.36e-15 1.92
                                                            9.11 0.849 0.871
                                             1.51
                                                    -1.30
## 4 Arima
                    Traini~ -1.36e- 2 1.45
                                             1.13
                                                    -0.718
                                                            6.85 0.634 0.659
                                                                               0.00128
## 5 ETS
                                             1.14
                    Traini~ -4.88e- 3 1.46
                                                    -0.666
                                                            6.91 0.639 0.664
                                                                               0.0337
## 6 NNAR
                    Traini~ -2.99e- 4 0.605 0.453 -0.283
                                                            2.71 0.255 0.275 -0.0463
## 7 Prophet
                    Traini~
                              6.87e- 5 1.33
                                             1.02
                                                    -0.654
                                                            6.17 0.574 0.605
```

Based on the accuracy measures (say, RMSE, MAE, MAPE, MASE), it is found that Neural Net and then ARIMA models provide better forecasting performance compare to other models.

## Time Series Regression

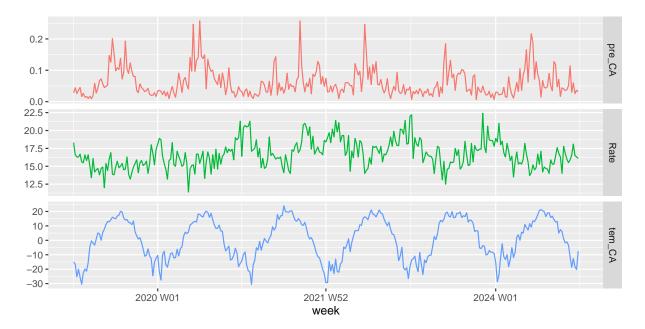
Now, we are going to do the time series regression models. We will forecast the weekly mortality rate based on some other predictor, say mean precipitation and temperature of week, trend of the data, any intervention effect (e.g. Covid-19 effect) and so on. We need to assume that morality rate has a linear relationship with these predictors.

First let's look at the time plot again.



Now, let's look at the relationship among variables.

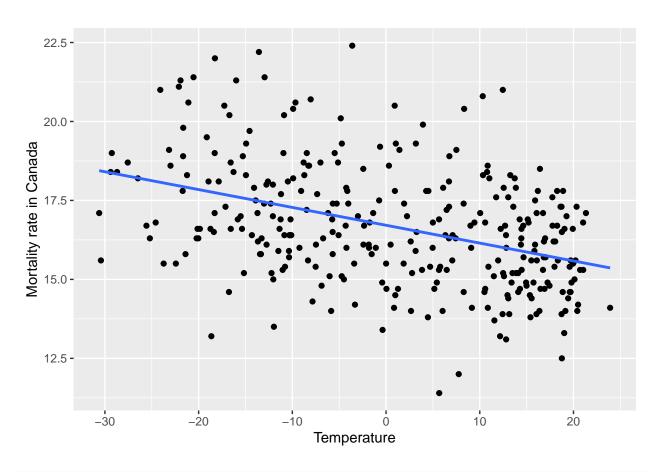
```
mortality_ts %>%
select(Rate, tem_CA, pre_CA, week) %>% pivot_longer(-week) |>
ggplot(aes(week, value, colour = name)) +
geom_line() + facet_grid(name ~ ., scales = "free_y") +
guides(colour = "none") + labs(y=" ")
```



Let's see the relation between weekly mortality rate and weekly mean temperature and mean precipitation.

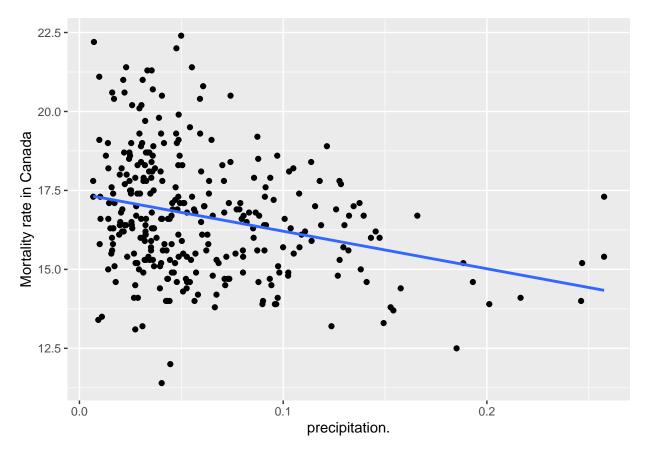
```
mortality_ts %>%
ggplot(aes(y = Rate, x = tem_CA)) +
labs(x = "Temperature", y = "Mortality rate in Canada") +
geom_point() + geom_smooth(method = "lm", se = FALSE)
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```



```
mortality_ts %>%
ggplot(aes(y = Rate, x = pre_CA)) +
labs(x = "precipitation.", y = "Mortality rate in Canada") +
geom_point() + geom_smooth(method = "lm", se = FALSE)
```

## 'geom\_smooth()' using formula = 'y ~ x'



Let's fit the time series regression model using covariates/predictors, say COVID-19 emergency intervention effect, trend variable (already built in fpp3 package as trend())

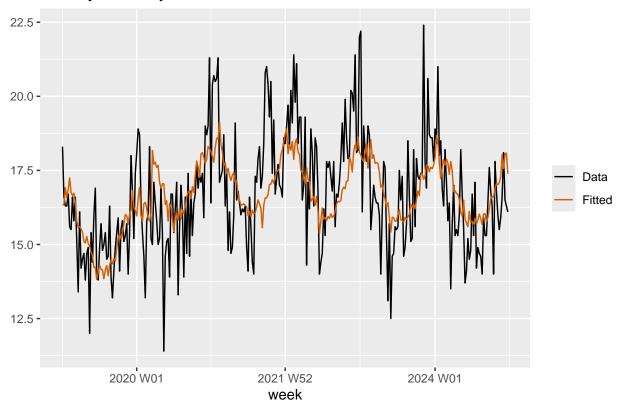
```
fit_MR <- mortality_ts %>%
model(lm = TSLM(Rate ~ tem_CA + pre_CA))
report(fit_MR)
```

```
## Series: Rate
## Model: TSLM
##
## Residuals:
##
       Min
                1Q Median
                                3Q
                                       Max
##
  -5.1107 -1.2302 -0.1048 1.1066 5.4669
##
## Coefficients:
##
                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.960834
                           0.180477
                                     93.978 < 2e-16 ***
                                     -6.090 3.35e-09 ***
## tem CA
               -0.049775
                           0.008174
               -4.153324
                           2.536057
                                     -1.638
                                               0.102
## pre_CA
##
                   0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
## Signif. codes:
##
## Residual standard error: 1.748 on 310 degrees of freedom
## Multiple R-squared: 0.1764, Adjusted R-squared: 0.171
## F-statistic: 33.19 on 2 and 310 DF, p-value: 8.7081e-14
```

Let's fit the model using other covariates/predictors, say Covid-19 intervention effect, trend variable (already built in fpp3 package as trend())

```
# create dummy for Covid-19 intervention effect
mortality_ts <- mortality_ts |>
mutate(covid19=ifelse(Date<as.Date("2020-03-17"),0,1))</pre>
fit_mortality <- mortality_ts |>
model(TSLM(Rate ~ tem_CA + pre_CA + covid19 + trend()))
report(fit_mortality)
## Series: Rate
## Model: TSLM
##
## Residuals:
##
      Min
               1Q Median
                              3Q
                                     Max
## -5.6463 -1.1411 -0.1157 1.0471 5.2008
##
## Coefficients:
##
               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.610218   0.255133   61.185   < 2e-16 ***
## tem CA
             ## pre CA
              -4.192882 2.340345 -1.792 0.0742 .
              2.055507
## covid19
                         0.315501
                                   6.515 2.96e-10 ***
              -0.001813
                         0.001402 -1.293
## trend()
                                            0.1972
## ---
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
## Residual standard error: 1.599 on 308 degrees of freedom
## Multiple R-squared: 0.3152, Adjusted R-squared: 0.3063
## F-statistic: 35.44 on 4 and 308 DF, p-value: < 2.22e-16
# Check the fitted values
augment(fit_mortality) |>
ggplot(aes(x = week)) +
geom_line(aes(y = Rate, colour = "Data")) +
geom_line(aes(y = .fitted, colour = "Fitted")) +
labs(y = NULL,
title = "Weekly Mortality rate in Canada"
) +
scale colour manual(values=c(Data="black",Fitted="#D55E00")) +
guides(colour = guide_legend(title = NULL))
```

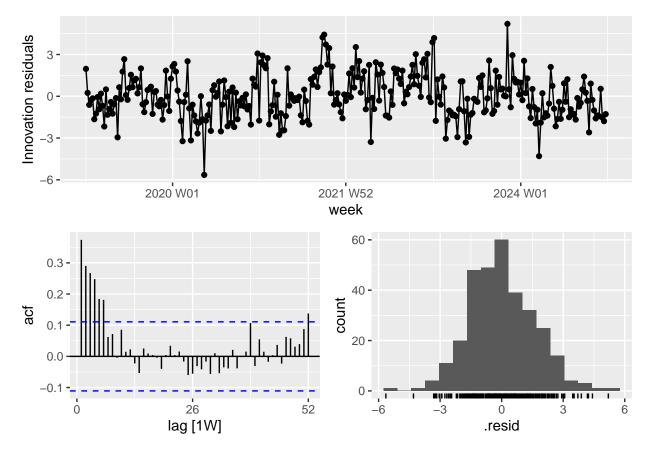
## Weekly Mortality rate in Canada



## Residual diagnostics

It is very important to do residual diagnostic after fitting the regression model so that we can verify the assumptions (residual's are uncorrelated, zero mean, constant variance, uncorrelated with other predictor).

fit\_mortality |> gg\_tsresiduals(lag=52)



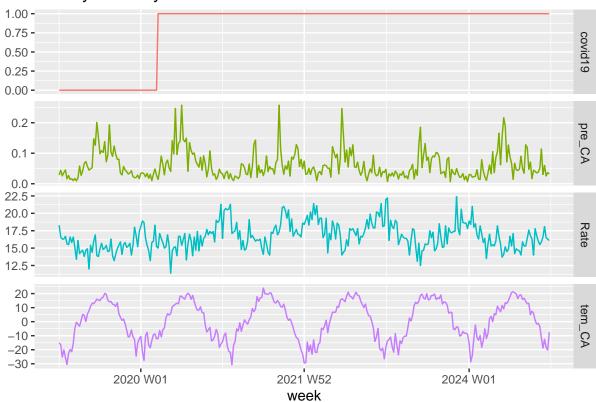
By looking at the auto correlation function of residuals, we can see still some correlation left in the residuals! What to do then??

Solution: we can use Dynamic regression model.

### Dynamic regression models

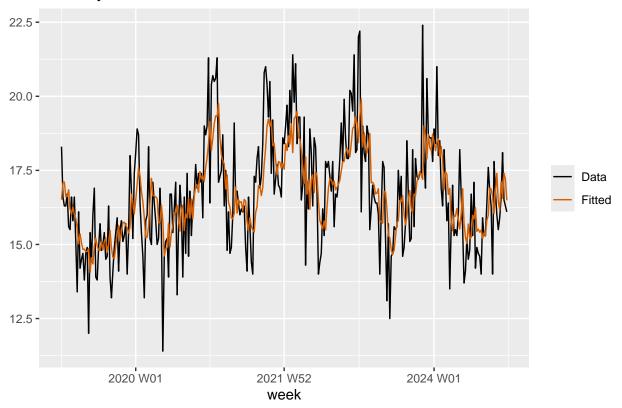
```
mortality_ts %>%
select(-Date) %>%
gather(key='variable', value='value') %>%
ggplot(aes(y=value, x=week, group=variable, colour=variable)) +
geom_line() + facet_grid(variable ~ ., scales='free_y') +
labs(y="",title ="Weekly mortality rate in Canada") +
guides(colour="none")
```

## Weekly mortality rate in Canada

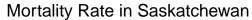


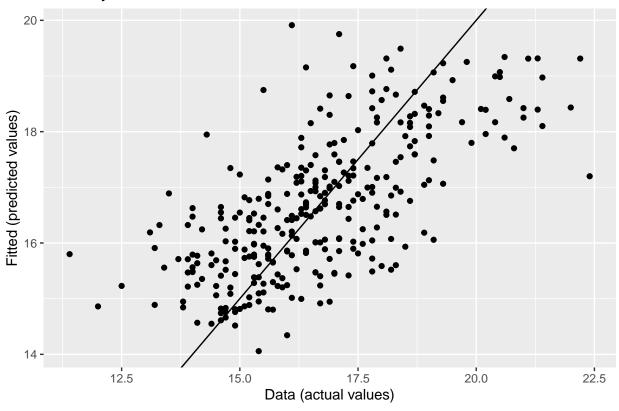
```
fit.dyn <- mortality_ts %>% model(ARIMA(Rate ~ tem_CA + pre_CA+ covid19+ trend()))
report(fit.dyn)
## Series: Rate
## Model: LM w/ ARIMA(1,0,1)(0,0,1)[52] errors
## Coefficients:
##
            ar1
                     ma1
                            sma1
                                   tem_CA
                                            pre_CA covid19
                                                             trend()
                                                                      intercept
##
         0.8591 -0.5749
                         0.1206
                                  -0.0289
                                           -3.0963
                                                     1.3523
                                                             -0.0007
                                                                         15.8954
## s.e. 0.0531
                  0.0789 0.0654
                                   0.0140
                                            2.2148
                                                     0.7391
                                                              0.0035
                                                                         0.5628
##
## sigma^2 estimated as 2.034: log likelihood=-551.78
                               BIC=1155.27
## AIC=1121.55
                AICc=1122.14
```

## Mortality rate in Saskatchewan



```
augment(fit.dyn) |>
  ggplot(aes(x = Rate, y = .fitted)) +
  geom_point() +
  labs(
    y = "Fitted (predicted values)",
    x = "Data (actual values)",
    title = "Mortality Rate in Saskatchewan"
) +
  geom_abline(intercept = 0, slope = 1)
```

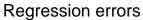


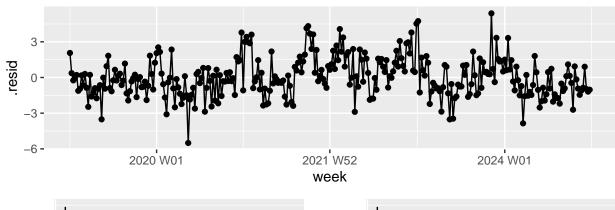


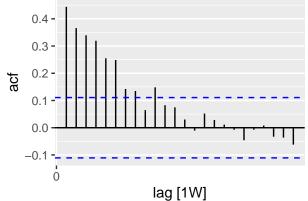
## Residual diagnostics

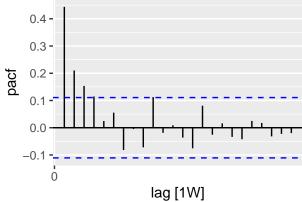
Recall that, in dynamic regression we have two errors: regression error ( $\eta_t$ , will show some correlation) and Arima error ( $\epsilon_t$ , will show no significant correlation).

```
# regression error
residuals(fit.dyn, type='regression') %>%
   gg_tsdisplay(.resid, plot_type = 'partial') +
   labs(title = "Regression errors")
```

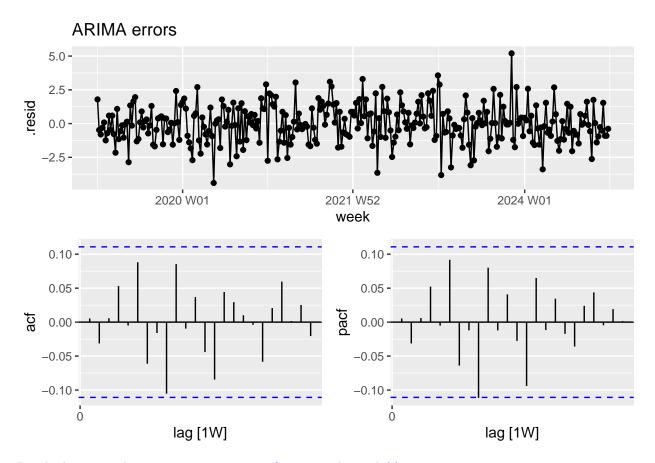








```
# Arima error
residuals(fit.dyn, type='innovation') %>%
gg_tsdisplay(.resid, plot_type = 'partial') +
labs(title = "ARIMA errors")
```



Residuals seems white noise, means no significant correlation left!

The results are not significant (i.e., the p-values are relatively large). Thus, we can conclude that the residuals are not distinguishable from a white noise series.

Comment: Overall, it shows that the dynamic regression perform better than the time series linear regression.

# Interactive graphics for Time Series data:

We are trying to developed interactive shiny graphics to address these issues for several of the most common time series data analyses. The interactive shiny graphics is used to generate an interactive visualization environment that contains several distinct graphics, many of which are updated in response to user input. These visualizations reduce the burden of exploratory analyses and can serve as a useful tool for the communication of results to non-statisticians.

## Link (still under working)

<a href="https://syedrizvi05.shinyapps.io/TimeSeries/">https://syedrizvi05.shinyapps.io/TimeSeries/</a>

# Acknowledgement:

This document was prepared with the help of a research student, *Syed Jafar Rizvi*, MSc Student, Dept. of Community Health and Epidemiology, University of Saskatchewan. Fell free to knock Rizbi at *jafar.rizvi@usask.ca* if you have any question.