Time Series Analysis in Health Research

Lecture 3: Time Series Regression Models

Erfan Hoque

Dept. of Community Health & Epidemiology, University of Saskatchewan

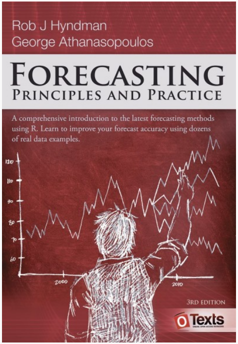
SEA 24th Annual Symposium Workshop, September 26, 2025

Schedule for this workshop

Week	Topic
Lecture 1	Introduction to forecasting, graphics
Lecture 2	Time Series toolbox and Forecasting Models
Lecture 3	Time Series Regression Models
Lecture 4	Hands-on session in R
Lecture 5	Advanced Forecasting Methods

Resources

Organization and presentation of material in these lectures will largely pull from



- Free and online (`https://otexts.com/fpp3/`)
- Data sets in associated R packages, R code for examples

Time series linear model

Time series linear model

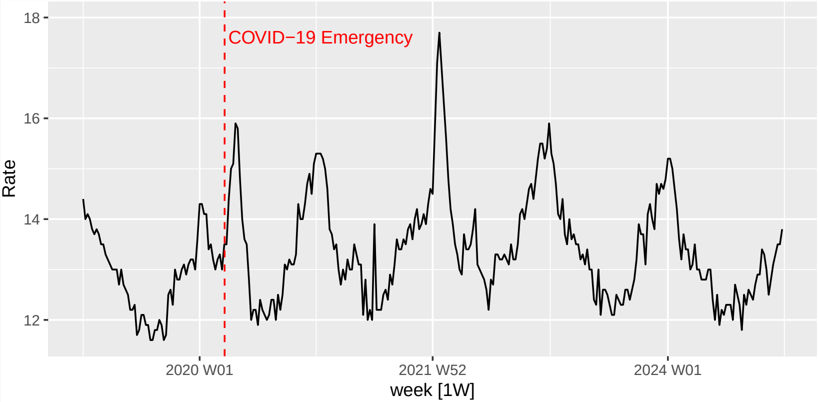
We will discuss regression models. The basic concept is that we forecast the time series of interest y assuming that it has a linear relationship with other time series x.

Multiple regression and forecasting

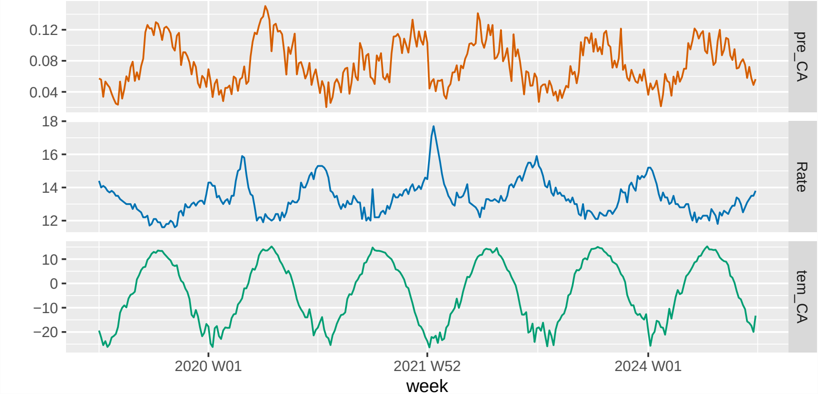
$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t.$$

- y_t is the variable we want to forecast: the "response" variable
- Each $x_{j,t}$ is numerical and is called a "predictor". They are usually assumed to be known for all past and future times.
- The coefficients $\beta_1, \ldots, \beta_k$ measure the effect of each predictor after taking account of the effect of all other predictors in the model. That is, the coefficients measure the **marginal effects** of predictor variables.
- ε_t is a white noise error term.

Example: Weekly mortality data

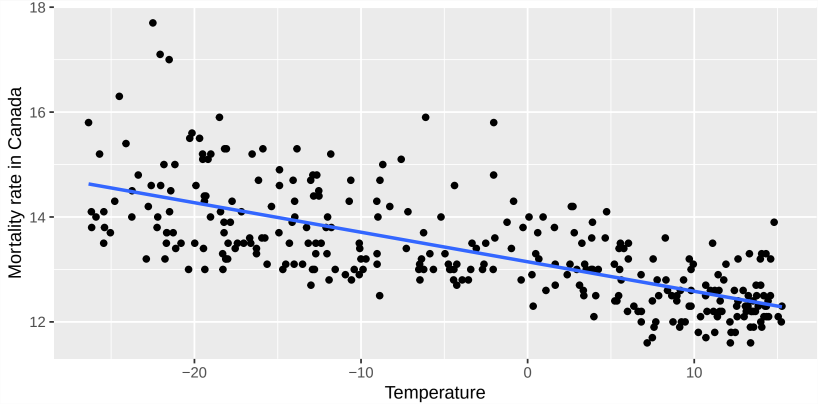


Example: Weekly mortality data and other predictors



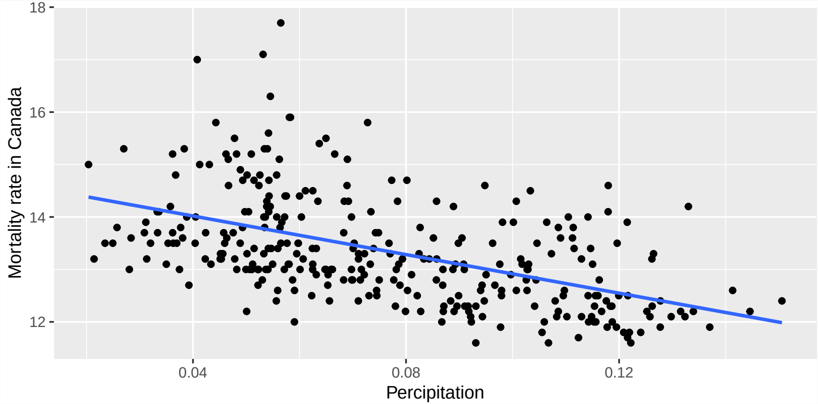
Example: Weekly mortality data

```r
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)
```

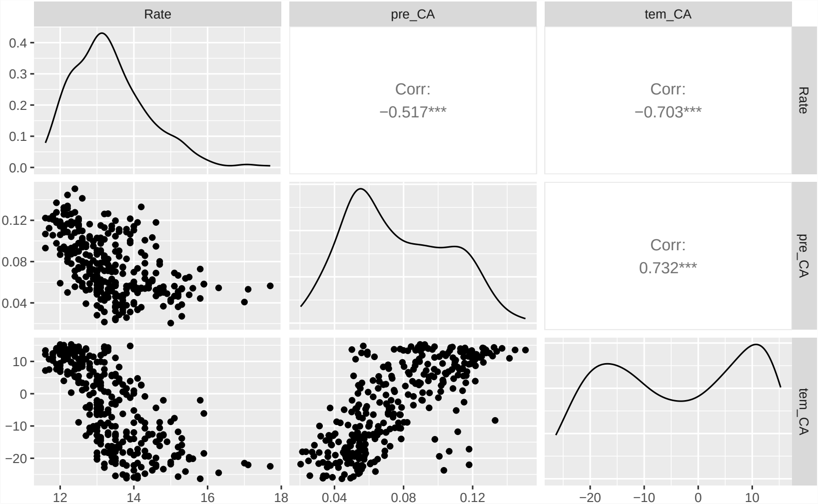


Example: Weekly mortality data

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



Example: Weekly mortality data

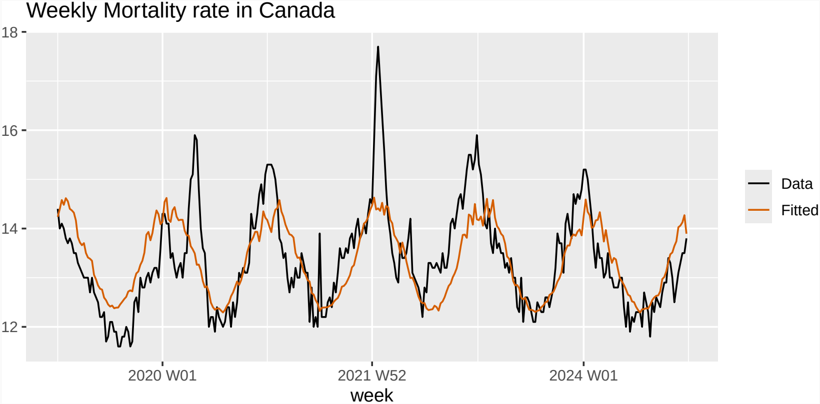


Example: Weekly mortality data

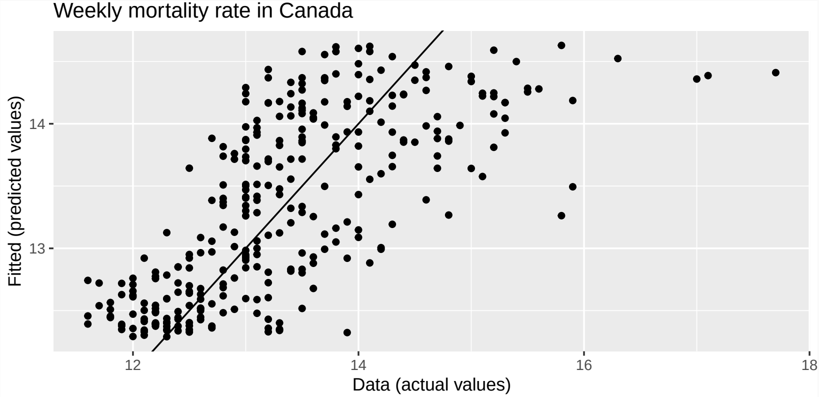
```r
fit_MR <- mortality_ts %>%
  model(lm = TSLM(Rate ~ tem_CA + pre_CA))
report(fit_MR)
```

```
## Series: Rate
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.29  -0.50  -0.14   0.40   3.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.16352    0.17872    73.7   <2e-16 ***
## tem_CA      -0.05596    0.00475   -11.8   <2e-16 ***
## pre_CA      -0.20707    2.10896    -0.1     0.92
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.738 on 310 degrees of freedom
## Multiple R-squared: 0.494,   Adjusted R-squared: 0.491
## F-statistic:  151 on 2 and 310 DF, p-value: <2e-16
```

Example: Weekly mortality data



Example: Weekly mortality data



Residual diagnostics

Multiple regression and forecasting

For forecasting purposes, we require the following assumptions:

- ε_t are uncorrelated and zero mean
- ε_t are uncorrelated with each $x_{j,t}$.

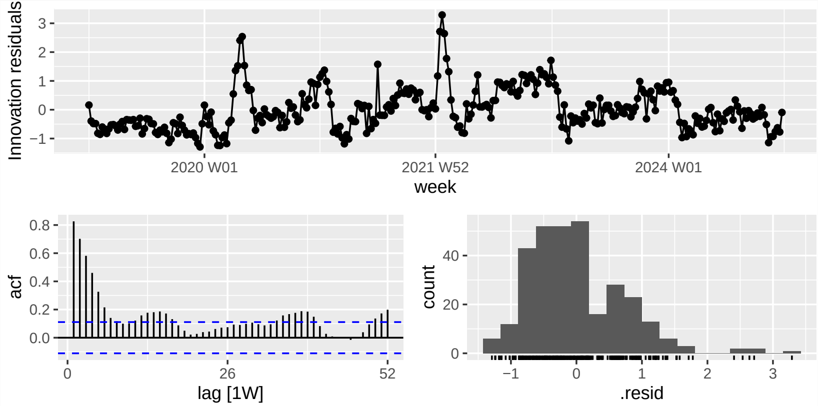
It is **useful** to also have $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$ when producing prediction intervals or doing statistical tests.

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Weekly mortality data: Residual diagnostic

```r
fit_MR |> gg_tsresiduals(lag=52)
```



Some useful predictors

Trend

Linear trend

$$x_t = t$$

- $t = 1, 2, \ldots, T$
- Strong assumption that trend will continue.

Dummy variables

- If a categorical variable takes only two values (e.g., `Yes'` or`No'), then an equivalent numerical variable can be constructed taking value 1 if yes and 0 if no. This is called a **dummy variable**.

	A	B
1	Yes	1
2	Yes	1
3	No	0
4	Yes	1
5	No	0
6	No	0
7	Yes	1
8	Yes	1
9	No	0
10	No	0
11	No	0
12	No	0
13	Yes	1

Dummy variables

- If there are more than two categories, then the variable can be coded using several dummy variables (one fewer than the total number of categories).

	A	B	C	D	E
1	Monday	1	0	0	0
2	Tuesday	0	1	0	0
3	Wednesday	0	0	1	0
4	Thursday	0	0	0	1
5	Friday	0	0	0	0
6	Monday	1	0	0	0
7	Tuesday	0	1	0	0
8	Wednesday	0	0	1	0
9	Thursday	0	0	0	1
10	Friday	0	0	0	0
11	Monday	1	0	0	0
12	Tuesday	0	1	0	0
13	Wednesday	0	0	1	0
14	Thursday	0	0	0	1
15	Friday	0	0	0	0

Beware of the dummy variable trap!

- Using one dummy for each category gives too many dummy variables!
- The regression will then be singular and inestimable.
- Either omit the constant, or omit the dummy for one category.
- The coefficients of the dummies are relative to the omitted category.

Uses of dummy variables

Seasonal dummies

- For quarterly data: use 3 dummies
- For monthly data: use 11 dummies
- For daily data: use 6 dummies
- What to do with weekly data?

Outliers

- If there is an outlier, you can use a dummy variable to remove its effect.

Public holidays

- For daily data: if it is a public holiday, dummy=1, otherwise dummy=0.

Intervention variables

Spikes

- Equivalent to a dummy variable for handling an outlier.

Steps

- Variable takes value 0 before the intervention and 1 afterwards.

Change of slope

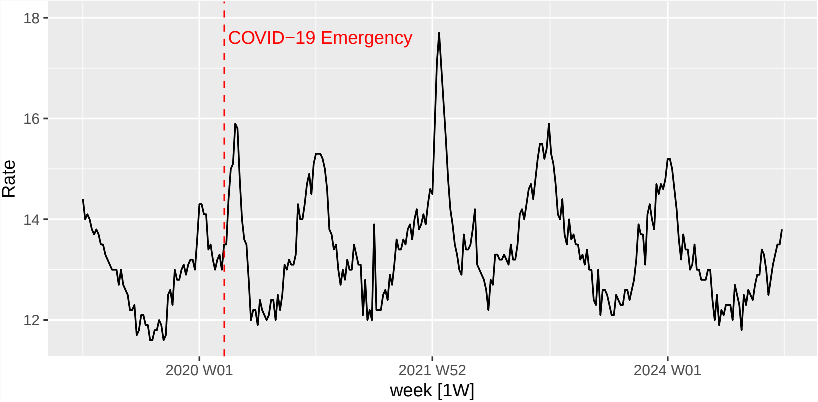
- Variables take values 0 before the intervention and values $\{1, 2, 3, \dots\}$ afterwards.

Holidays

For monthly data

- Christmas: always in December so part of monthly seasonal effect
- Easter: use a dummy variable $v_t = 1$ if any part of Easter is in that month, $v_t = 0$ otherwise.
- Ramadan and Chinese new year similar.

Revisit weekly mortality data



Regression model:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \text{Temp}_t + \beta_3 \text{Per}_t + \beta_4 \text{Covid-19}_t + \varepsilon_t$$

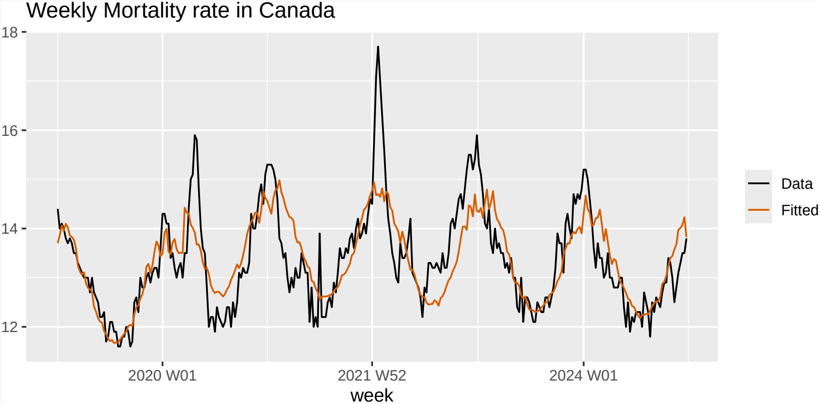
- $\text{Covid-19}_t = 1$ if t is after 2021-3-17 and 0 otherwise.

Revisit weekly mortality data

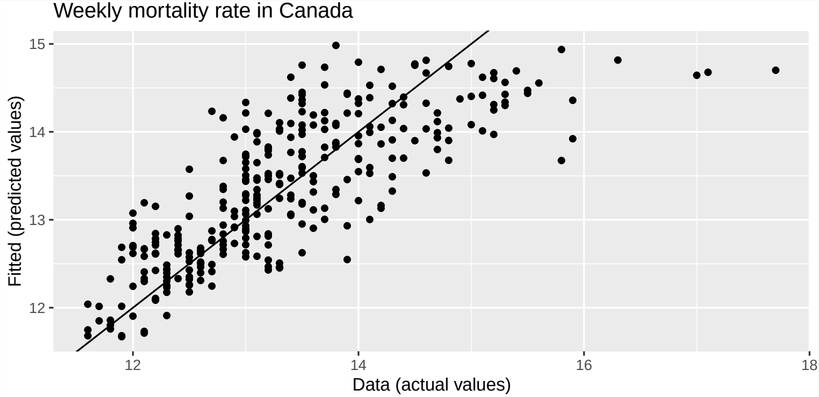
```r
mortality_ts <- mortality_ts |>
  mutate(covid19=ifelse(Date<as.Date("2020-03-17"),0,1)) # create dummy
fit_mortality <- mortality_ts |>
  model(TSLM(Rate ~ tem_CA + pre_CA + covid19 + trend()))
report(fit_mortality)
```

```
## Series: Rate
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.534 -0.424 -0.061  0.351  2.999
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.562091   0.179630   69.93   <2e-16 ***
## tem_CA      -0.059391   0.004217  -14.08   <2e-16 ***
## pre_CA      -0.191524   1.862001   -0.10   0.9181
## covid19      1.153946   0.127733    9.03   <2e-16 ***
## trend()     -0.002128   0.000566   -3.76   0.0002 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Revisit weekly mortality data

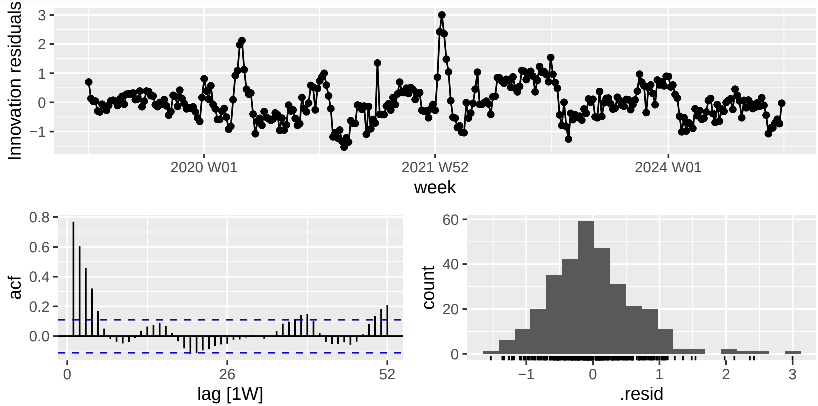


Revisit weekly mortality data



Revisit weekly mortality data

```
fit_mortality |> gg_tsresiduals(lag=52)
```



Still some correlation left in the residuals!

Revisit weekly mortality data

```r
augment(fit_mortality) |>
  features(.innov, ljung_box, lag = 52)
```

```
## # A tibble: 1 x 3
##   .model                                       lb_stat lb_pvalue
##   <chr>                                          <dbl>     <dbl>
## 1 TSLM(Rate ~ tem_CA + pre_CA + covid19 +~   518.          0
```

Forecasting with regression

Ex-ante versus ex-post forecasts

- *Ex ante forecasts* are made using only information available in advance.
  - require forecasts of predictors
- *Ex post forecasts* are made using later information on the predictors.
  - useful for studying behaviour of forecasting models.
- trend, seasonal and calendar variables are all known in advance, so these don't need to be forecast.

Dynamic Regression models

Dynamic regression models

- The regression models in previous part allow for the inclusion of a lot of relevant information from predictor variables, but do not allow for the subtle time series dynamics that can be handled with ARIMA models.
- Here, we consider how to extend ARIMA models in order to allow other information to be included in the models.
- Here, we will allow the errors from a regression to contain autocorrelation.

Regression with ARIMA errors

Regression with ARIMA errors

Regression models

$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$

- y_t modeled as function of k explanatory variables $x_{1,t}, \ldots, x_{k,t}$.
- In regression, we assume that ε_t was white noise series.
- Now we want to allow ε_t to be autocorrelated.

Example: ARIMA(1,1,1) errors

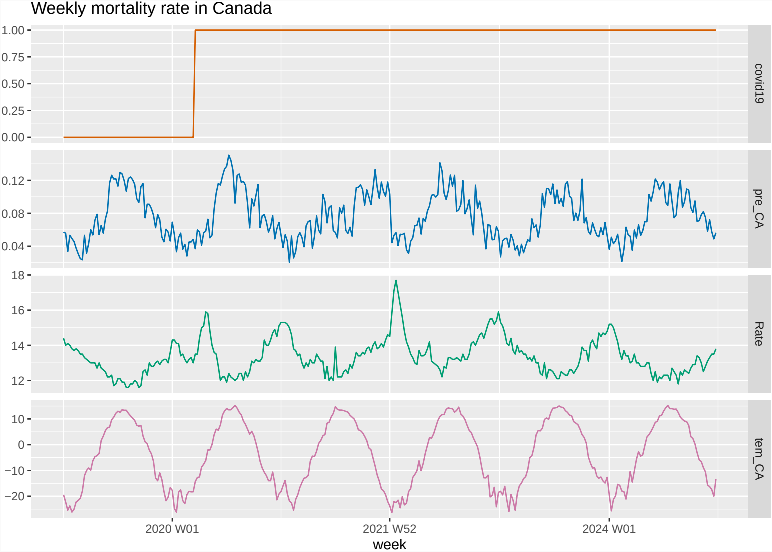
$$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \eta_t,$$
$$\eta_t \sim \text{ARIMA}(1,1,1),\ \varepsilon_t \sim IID(0, \sigma^2) \text{ is white noise.}$$

- Be careful in distinguishing η_t from ε_t.
- Only the errors ε_t are assumed to be white noise.
- In ordinary regression, η_t is assumed to be white noise and so $\eta_t = \varepsilon_t$.

Regression with ARIMA errors

- In R, we can specify an ARIMA(p, d, q) for the errors, and d levels of differencing will be applied to all variables $(y, x_{1,t}, \ldots, x_{k,t})$.
- Check that ε_t series looks like white noise.
- AICc can be calculated for final model.
- Repeat procedure for all subsets of predictors to be considered, and select model with lowest AICc value.

Example: weekly mortality rate in Canada



Example: weekly mortality rate in Canada

```
fit <- mortality_ts %>% model(ARIMA(Rate ~ tem_CA + pre_CA+ covid19+ trend()))
report(fit)
```

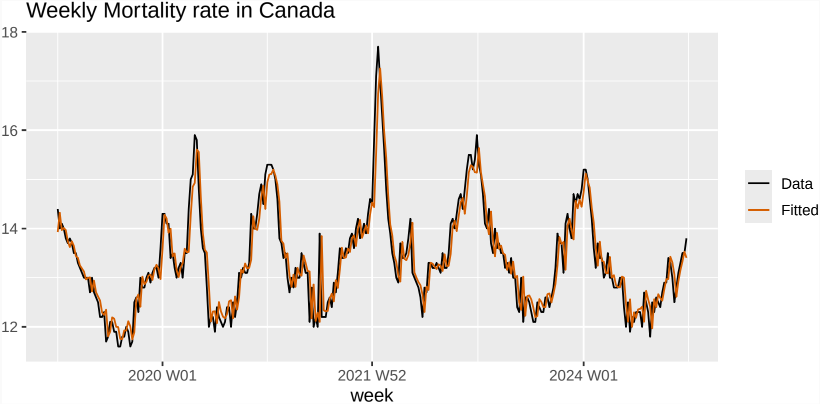
```
## Series: Rate
## Model: LM w/ ARIMA(1,0,0) errors
##
## Coefficients:
##          ar1   tem_CA  pre_CA  covid19  trend()  intercept
##       0.8927  -0.0144   0.434    0.628  -0.0015     13.043
## s.e.  0.0352   0.0109   1.356    0.370   0.0024      0.437
##
## sigma^2 estimated as 0.1591:  log likelihood=-154
## AIC=322   AICc=323   BIC=349
```

$$y_t = 13.043 + (-0.0144)\text{Temp}_t + 0.434\text{Per}_t + 0.628\text{Covid-19}_t + (-0.0015)t + \eta_t$$

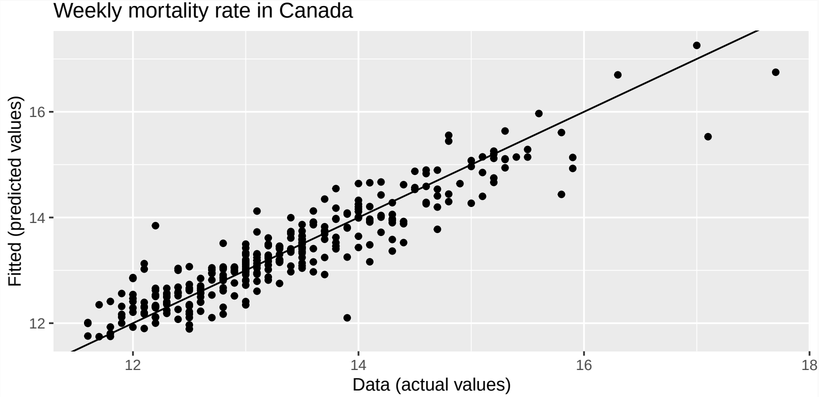
$$\eta_t = 0.8927\eta_{t-1} + \varepsilon_t,$$

$$\varepsilon_t \sim \text{N}(0, 0.159)$$

Example: weekly mortality rate in Canada

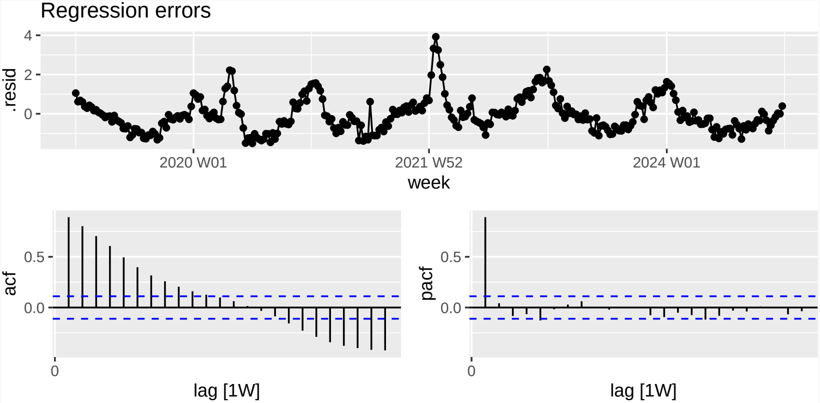


Example: weekly mortality rate in Canada



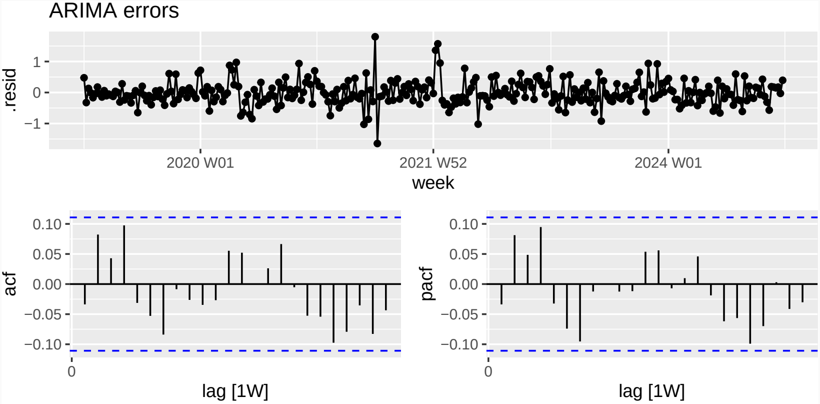
Weekly mortality data: Residual diagnostic

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



Weekly mortality data: Residual diagnostic

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



Residuals seems white noise

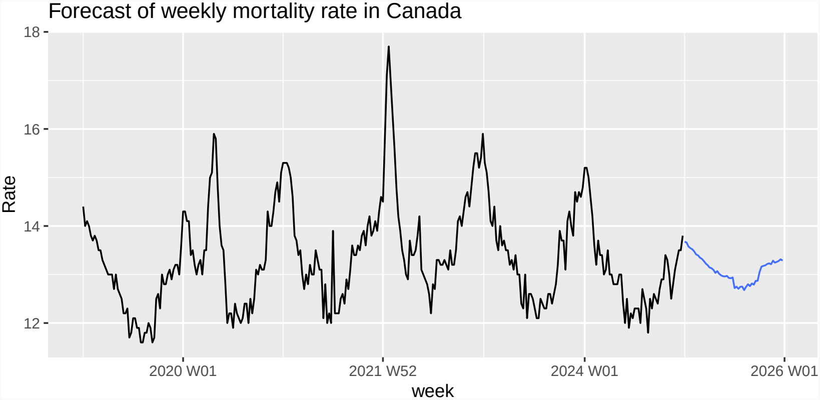
Weekly mortality data: Residual diagnostic

```r
augment(fit) |>
  features(.innov, ljung_box,lag = 52)
```

```
## # A tibble: 1 x 3
##   .model                                         lb_stat lb_pvalue
##   <chr>                                            <dbl>     <dbl>
## 1 ARIMA(Rate ~ tem_CA + pre_CA + covid19 ~    61.5     0.173
```

Forecasting of weekly mortality rate in Canada

```
forecast(fit, new_data = future_data) |>
  autoplot(mortality_ts, level=NULL) +
  labs(y = "Rate", title = "Forecast of weekly mortality rate in Canada")
```



Final thoughts!

This lecture provides the idea of time series regression models that can be used to forecast the time series of interest y assuming that it has a linear relationship with other time series x.

More details can be found in Chapters 7 and 10 of Forecasting: Principles and Practice (3rd ed, `https://otexts.com/fpp3/`), as well as in many other time series regression models resources.