Unit 7 Live Session (Solutions)

Time Series Analysis Lecture 2: Regression with Time Series, An Intro to Exploratory Time Series Data Analysis, and Time Series Smoothing

Figure 1: South Hall

Class Announcements

- HW 7 is this week
- Teams for Lab-2 have been created.

Roadmap

Rearview Mirror

- Notion of stationarity, ergodicity, and dependency
- Basic Time Series models

Today

- Time series decomposition
- Time series smoothing methods
- Time series decomposition and forecasting using OLS

Looking Ahead

- Autoregressive Moving Average Models (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)

Start-up Code

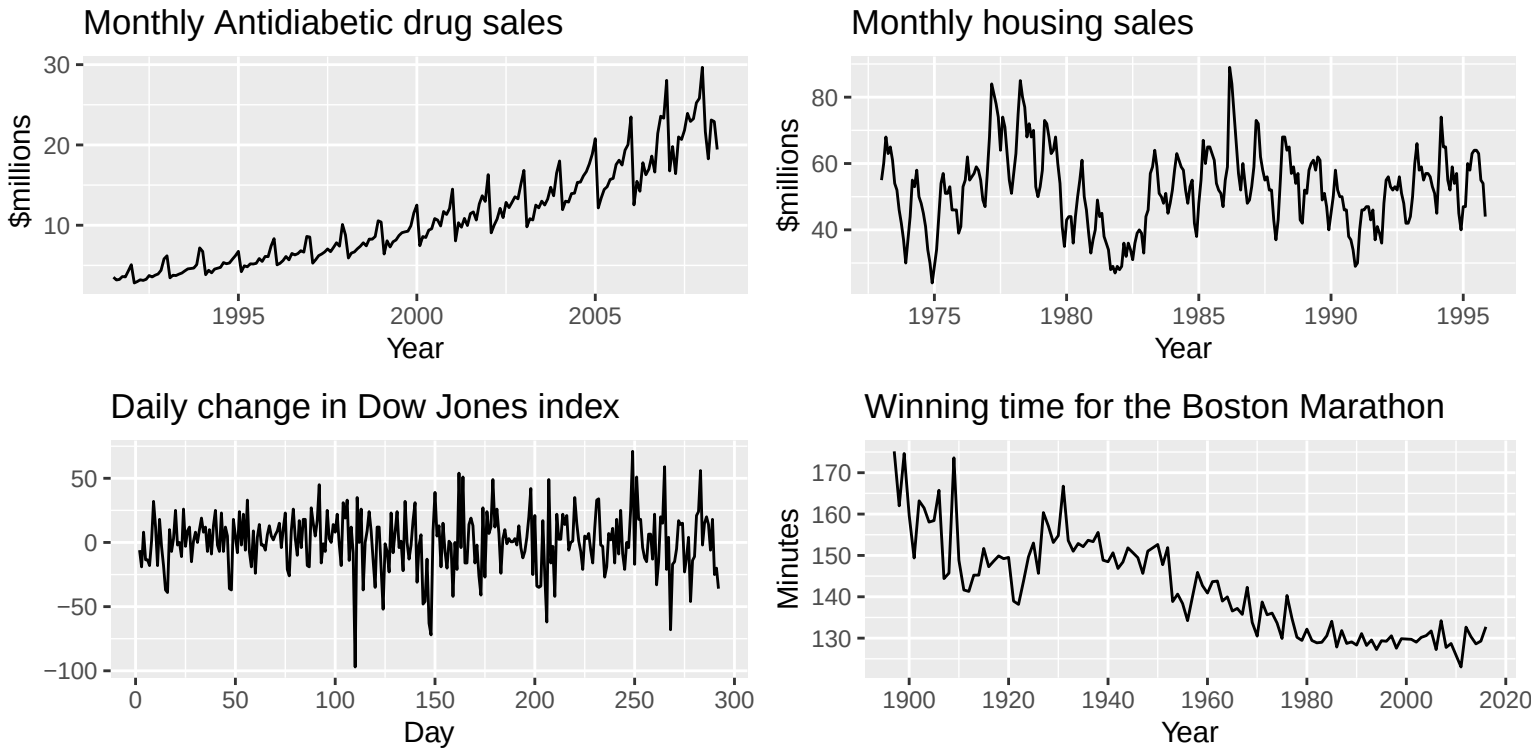
```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=80),tidy=TRUE)

# Load required libraries
## Load a set of packages inclusing: broom, cli, crayon, dbplyr , dplyr, dtplyr, forcats,
#googledrive, googlesheets4, ggplot2, haven, hms, httr, jsonlite, lubridate , magrittr,
#modelr, pillar, purrr, readr, readxl, reprex, rlang, rstudioapi, rvest, stringr, tibble,
#tidyr, xml2
library(tidyverse)
## To laod All data sets in the book "Forecasting: principles and practice"
#by Rob J Hyndman and George Athanasopoulos
library(fpp3)
library(fpp2)
## to create a infrastructure for tidy temporal data
library(tsibble)
### to work with date
library(lubridate)
## to use gg_season
library(feasts)
## Forecasting Models for Tidy Time Series
library(fable)
## To assemble multiple plots
library(gridExtra)
## for simulations
library(simts)
## To use TeX() to write expression in the title of plots
library(latex2exp)
```

Time series decomposition

Many time series include some or all of the following components.

- **Trend**: It is a long-term upward or downward movement of the data. The trend could be linear or nonlinear, and sometimes the trend might change over time.
- **Seasonal**: It is a regular movement in the data based on the season (e.g., every month/quarter/year). Seasonality is always of a fixed and known period.
- **Cyclical**: An oscillatory movement of a time series. Its length is not fixed, usually more than one year. In practice, the trend component is also assumed to include the cyclical component.
- **Irregular component** - a stationary process, such as white noise that is random.

a)- Given the above definitions, what are the possible components of the time series in the following graphs?



The monthly antidiabetic drug sales time series shows a clear and increasing trend. There is also a strong seasonal pattern that increases in size as the level of the series increases.

Monthly hosing sales time series shows strong seasonality each year and some solid cyclic behavior with a period of about 6–10 years.

Daily changes in dow joes index has no trend, seasonality, or cyclic behavior. There are random fluctuations that do not appear to be very predictable.

In the winning time of the Boston marathon, there is no seasonality but an apparent downward trend.

Time series decomposition: Mathematical representation

- To remove the deterministic components, we can decompose our time series into separate stationary and deterministic components.
- There are two common mathematical forms for decomposition:

1- Additive:

$$Y_t = T_t + S_t + E_t$$

2- Multiplicative:

$$Y_t = T_t \cdot S_t \cdot E_t$$

- **We call Y_t a trend stationary time series, if after removing the deterministic part, what remains E_t is a stationary series**

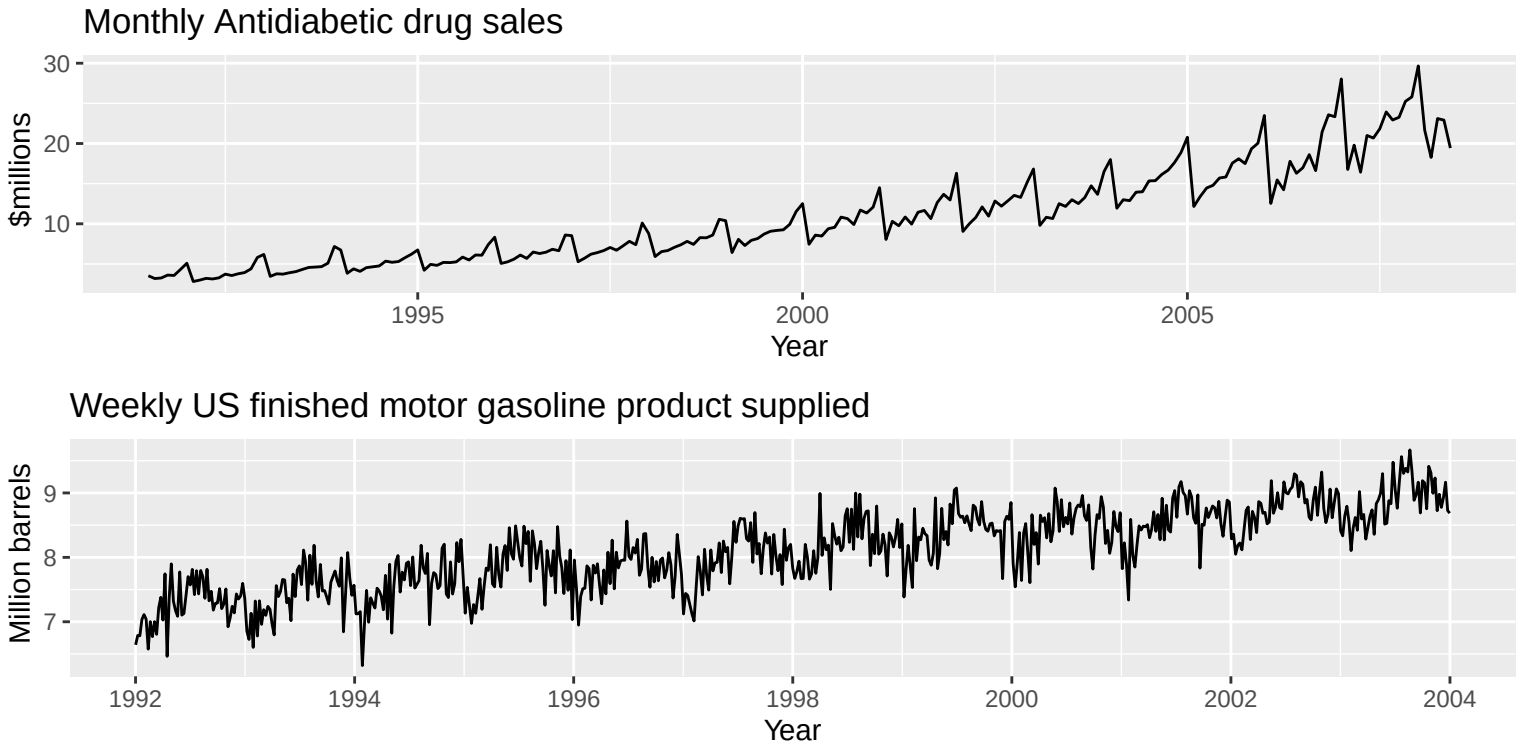
a)- Which type of decomposition might be appropriate for the following time series? Why?

In monthly antidiabetic drug sales, the seasonal component changes over time. A multiplicative model may be more appropriate if the seasonal effect increases as the trend increase.

For the gasoline product supplied, the magnitude of the seasonal fluctuations does not vary with the time series level, so An additive model is more appropriate.

b)- Is there any transformation that makes multiplicative decomposition simpler?

Since, in practice, it is much easier to deal with additive series $Yt = T_t + S_t + W_t$, we will transform the data by taking the logarithms



Time series decomposition: Trend and seasonal components detection

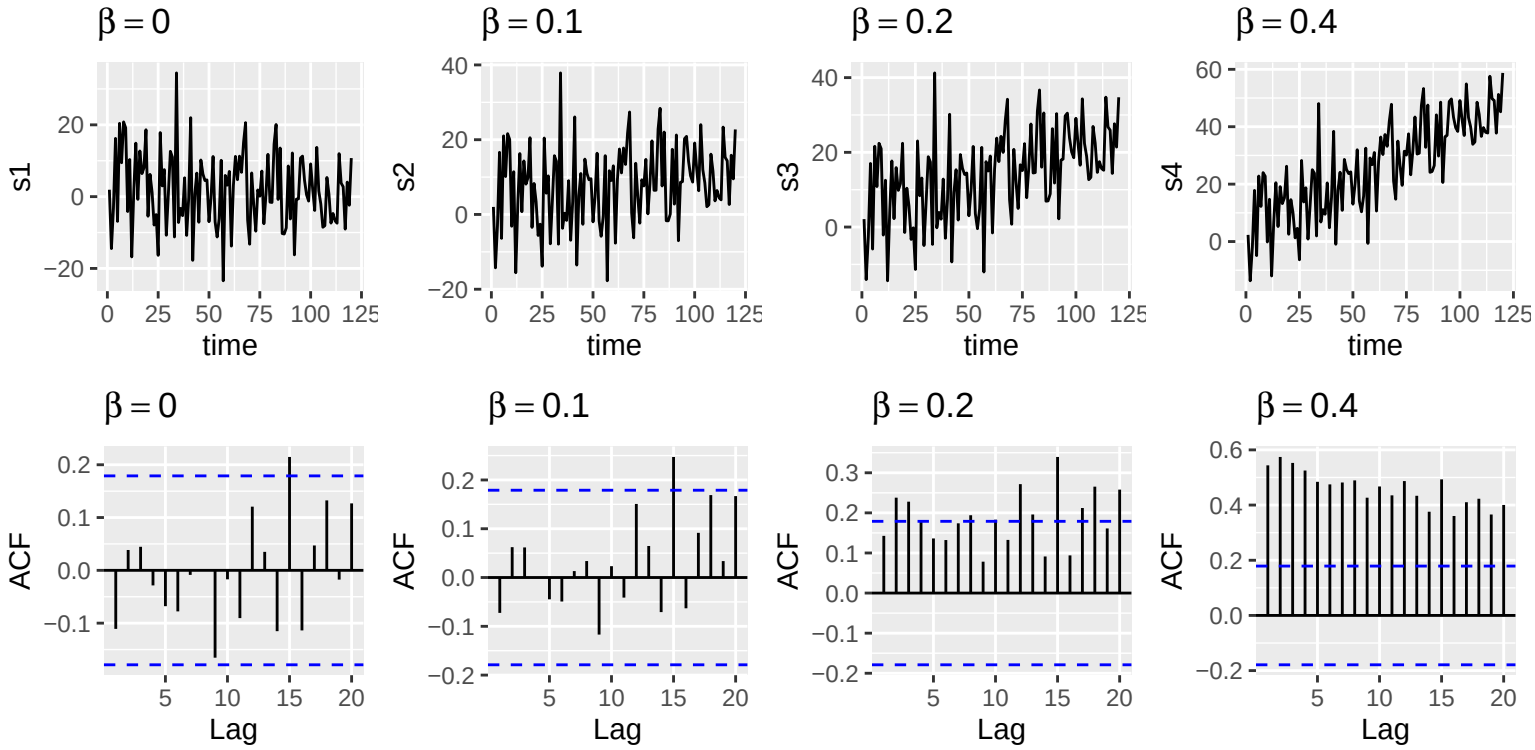
- The primary use of the ACF is to detect autocorrelation in the time series after removing the deterministic movement for the time series. ACF is also a helpful tool to determine if a time series has a trend or seasonal movement.

a)- The the following plots display the simulations of X_t for different values of β:

$$X_t = \beta t + W_t$$

- β is the slope over time and measures the change in Y_t over time.

- $\{W\}_{t=-\infty}^{\infty}$ is a white noise process with $E(W_t) = 0$ and $E(W_t^2) = \sigma^2$.
- How do the ACF and plot change when the trend becomes more pronounced?



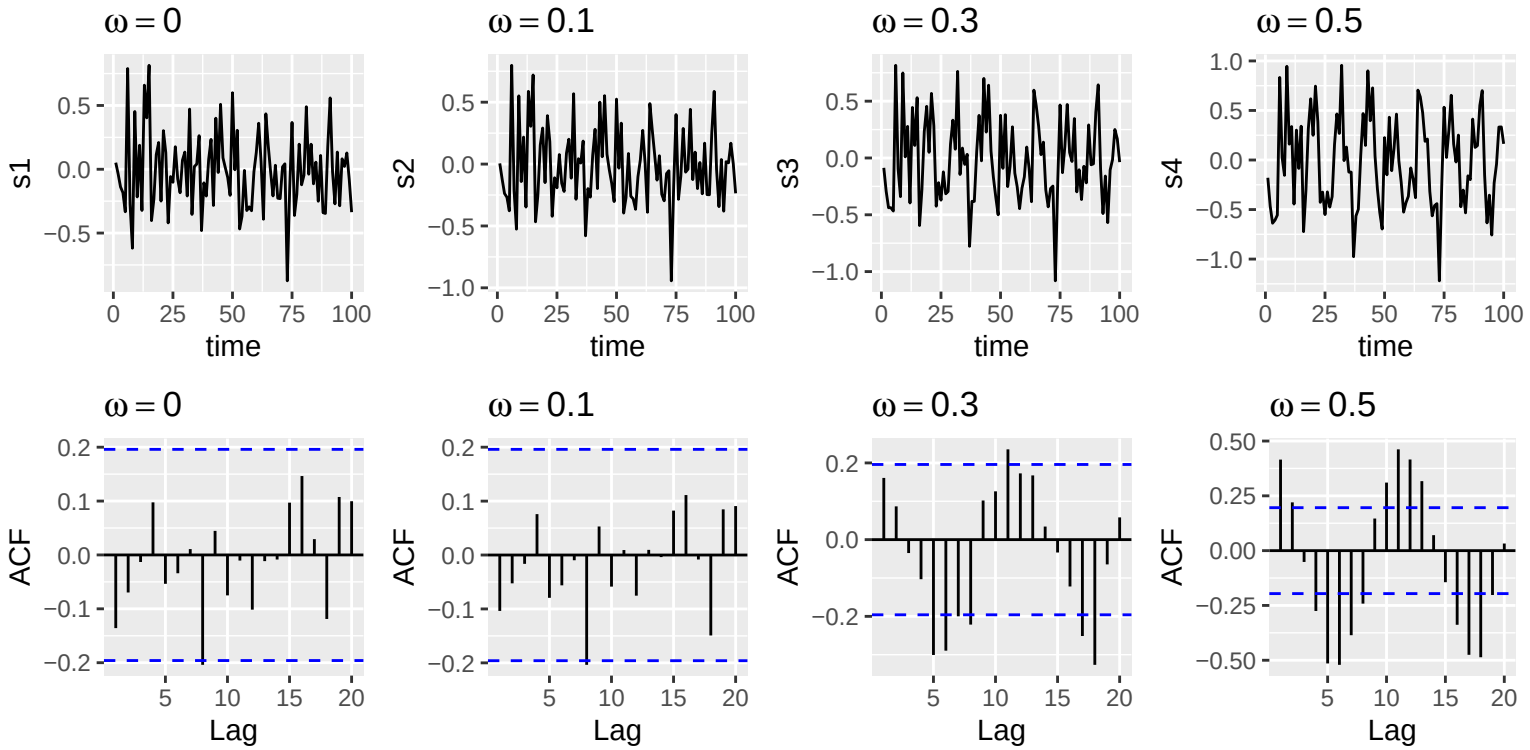
From the plot, we can see that X_t is a non-stationary series with a constant variance but a non-constant mean. And higher β or the more pronounced the trend leads to the slower declines in the ACF.

b)- The following time series process is from simulations of Y_t for different values of ω:

$$Y_t = \omega S_t + W_t$$

- ω measures the strength of the seasonal movement in Y_t over time.

- How do the ACF plots change with more pronounced seasonality?

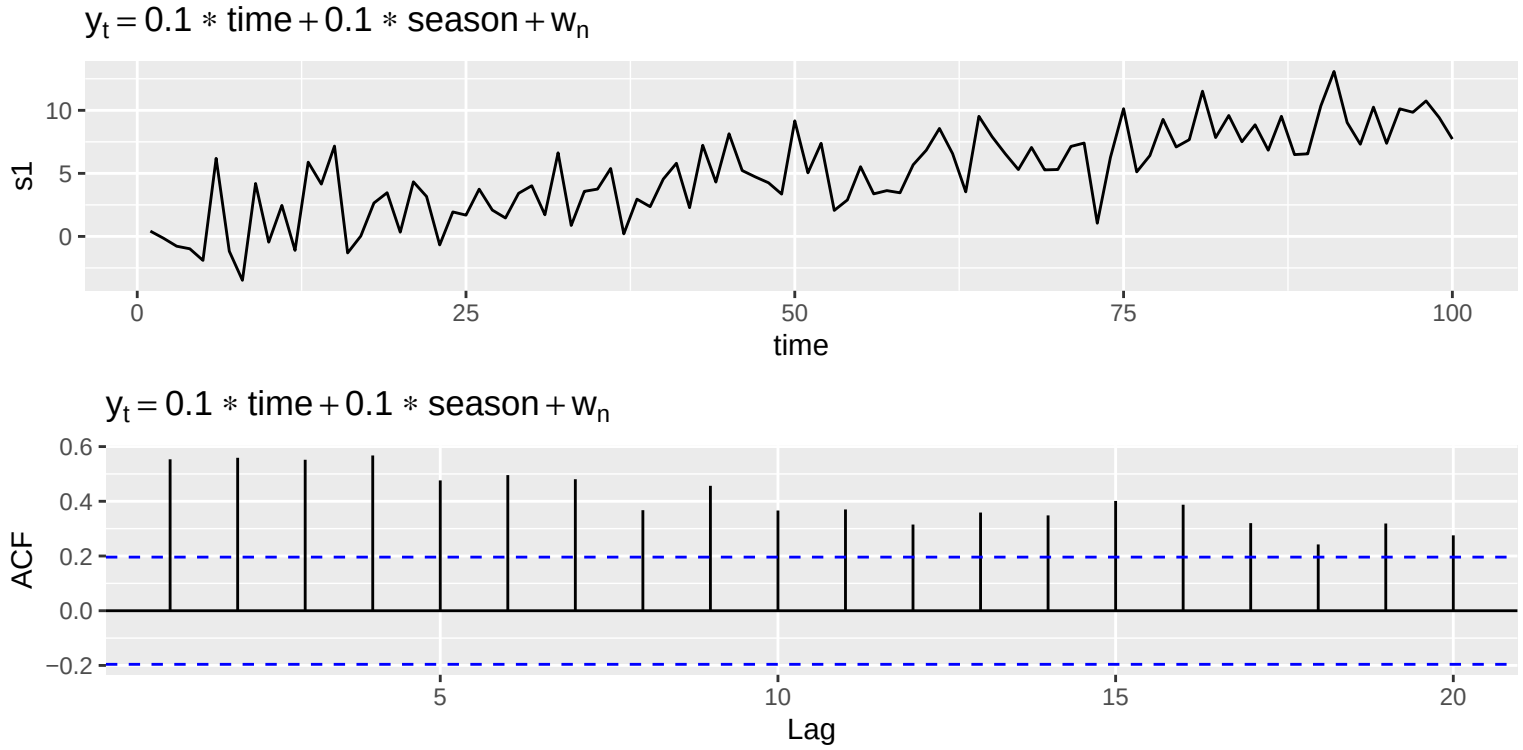


Based on these ACF plots, the more significant the amplitude of seasonal fluctuations, the more pronounced the oscillations are in the ACF.

c)- The following plot shows a simulation of Y_t with both trend and seasonal components:

$$Y_t = T_t + S_t + W_t$$

- What pattern does the ACF plot exhibit?



The ACF exhibits both a slow decline and oscillations movement.

Time series decomposition: Estimation

There are two approaches for estimating the deterministic movement in a time series:

1- Smoothing procedures

2- Linear regression

- What are the advantages and disadvantages of these two methods?

Time series decomposition using smoothing procedures

- Estimation of the deterministic component using a smoothing technique includes the following steps:

1- Estimate the trend using a smoothing procedure such as a moving average

2- De-trending the time series

- By subtracting the trend estimates from the time series for an additive decomposition
- By dividing the time series by the estimated trend values for a multiplicative decomposition

3- Estimating the seasonal factors from the de-trended series

- By calculating the mean (or median) values of the de-trended series for each specific period

4- Normalize the seasonal effects

- For an additive model, seasonal effects are adjusted so that the average of seasonal components is 0
- For a multiplicative model, the seasonal effects are adjusted so that they average 1

5- Calculate the irregular component

- For an additive model $\widehat{E_t} = Y_t - \widehat{T}_t - \widehat{S}_t$
- For a multiplicative model $\widehat{E_t} = \frac{Y_t}{\widehat{T}_t \cdot \widehat{S}_t}$

6- Analyze the residual component for stationarity: Decomposition aims to produce a stationary residual

- **The decomposition using smoothing techniques is usually quite successful at describing the time series in question. However, if they do not create an analytic expression for trend and seasonal parts like a moving average, they cannot easily be used for forecasting.**
- **One exception is Exponential smoothing which could be used either to produce smoothed data for presentation or to make forecasts.**

Time series decomposition using linear regression

- Linear regression can be used to estimate trend or seasonality terms in the following steps:

1- Estimate the trend or seasonal movement or both with the following specifications:

- A Linear trend: $Y_t = \beta_0 + beta_1 \cdot t + W_t$
- A quadratic trend: $Y_t = \beta_0 + beta_1 \cdot t + \beta_2 \cdot t^2 + W_t$
- A seasonal movement: $Y_t = \beta_0 + \sum_{i=1}^{s-1} \beta_i \cdot S_{it} + W_t$
- A linear trend with seasonal movement: $Y_t = \beta_0 + \beta_1 \cdot t + \sum_{i=2}^{s-1} \beta_i \cdot S_{it} + W_t$
- **We can include higher polynomial trends but it is not recommended for forecasting due to the risk of overfitting.**

2- Analyze the residual component:

- If it is white noise, we can use the estimated model for description and prediction.
- If it is not white noise but stationary, we can use a model to fit the stationary residuals, such as ARMA models.

3- If we have a few competing trend specifications, the best one can be chosen by AIC, BIC, RMSE (Root mean square error), or similar criteria.

4- Finally, forecasting can be achieved by forecasting the residuals and combining them with the forecasts of the trend and seasonal components.

Time series decomposition using both smoothing techniques and linear regression

- We can also combine smoothing techniques with linear regression in the following steps:

1- Estimate the trend, $\hat{T}_t^{\,1}$ via a smoothing method

2- Estimate and normalize the seasonal factors, $\hat{S}_t$, from the de-trended series

3- Deseasonalize the original data by removing the seasonal component $\hat{Y}_t = Y_t - \hat{S}_t$

4- Reestimate the trend, $\hat{T}_t^{\,2}$ from the deseasonalized data using a (polynomial) regression

5- Analyse the residuals $E_t = Y_t - \hat{S}_t - \hat{T}_t^{\,2}$ to verify if they are stationary and specify their model (if needed)

6- Forecast the series Y_{T+h}. Remember that $\hat{S}_t = \widehat{S_{t+d}}$ means that we can always forecast the seasonal component.

Differencing to de-trend and deseasonalize a time series

- Instead of using the smoothing technique or OLS regression to remove the deterministic movement in a time series, we can use differencing:

1- To remove a linear trend:

$$\bigtriangledown y_t = y_t - y_{t-1}$$

2- To remove a polynomial trend of degree k:

$$\bigtriangledown^k y_t = \bigtriangledown^{k-1}(y_t - y_{t-1}) = \bigtriangledown^{k-1} y_t - \bigtriangledown^{k-1} y_{t-1}$$

3- To remove a seasonal movement where $S_{t-d} = S_t = S_{t+d}$:

$$\bigtriangledown_d y_t = y_t - y_{t-d}$$

4- To remove both and trend seasonal movement, we need to apply both a non-seasonal first difference and a seasonal difference

- **Usually, the differences of order 1 or 2 are enough for removing a trend, and for seasonality, differences of order one are sufficient**
- **By differencing the data, our sample size is reduced.**
- **The interpretation also changes since we are now working with differences $\bigtriangledown y_t$, rather than levels of Y_t**

a)- Use the appropriate type of difference to remove the deterministic movement in Y_t completely.

$$Y_t = \beta_0 + \beta_1 \cdot t + S_t + Wt$$

- Where $S_t = S_{t+d}$

If we only apply a seasonal difference, we'll remove both seasonal and the trend movement, but instead of the trend, we'll have a constant $\beta_1 d$.

$$\begin{aligned}
\bigtriangledown_d y_t &= y_t - y_{t-d} \\
&= \beta_1 t - \beta_1 (t-d) + S_t - S_{t-d} + W_t - W_{t-d} \\
&= \beta_1 d + W_t - \bigtriangledown_d W_t
\end{aligned}$$

But if we want to remove the trend and seasonality altogether, we need to apply both a non-seasonal first difference and a seasonal difference

$$\begin{aligned}
\bigtriangledown_d y_t &= y_t - y_{t-d} \\
&= \beta_1 d + W_t - \bigtriangledown_d W_t \\
&= \beta_1 d + \bigtriangledown_d W_t - \beta_1 d - 1 - \bigtriangledown_d W_{t-1} \\
&= \beta_1 + \bigtriangledown_{d,1}^{1,1} W_t
\end{aligned}$$

Case Study: Airline passenger bookings in the U.S.

Introduction

An airline company usually needs to predict future demand before ordering new aircraft. The data science team obtained the latest U.S. carrier, foreign carrier, and individual airport passenger and flight data from the Bureau of Transportation Statistics (BTS). Their goal is to:

- **Describe how to model demand over time and predict future demand before ordering new aircraft**

Data Description and wrangling

The data set "Passengers_2019" contains a monthly time series of the number of domestic and international passenger bookings for 2010–2019.

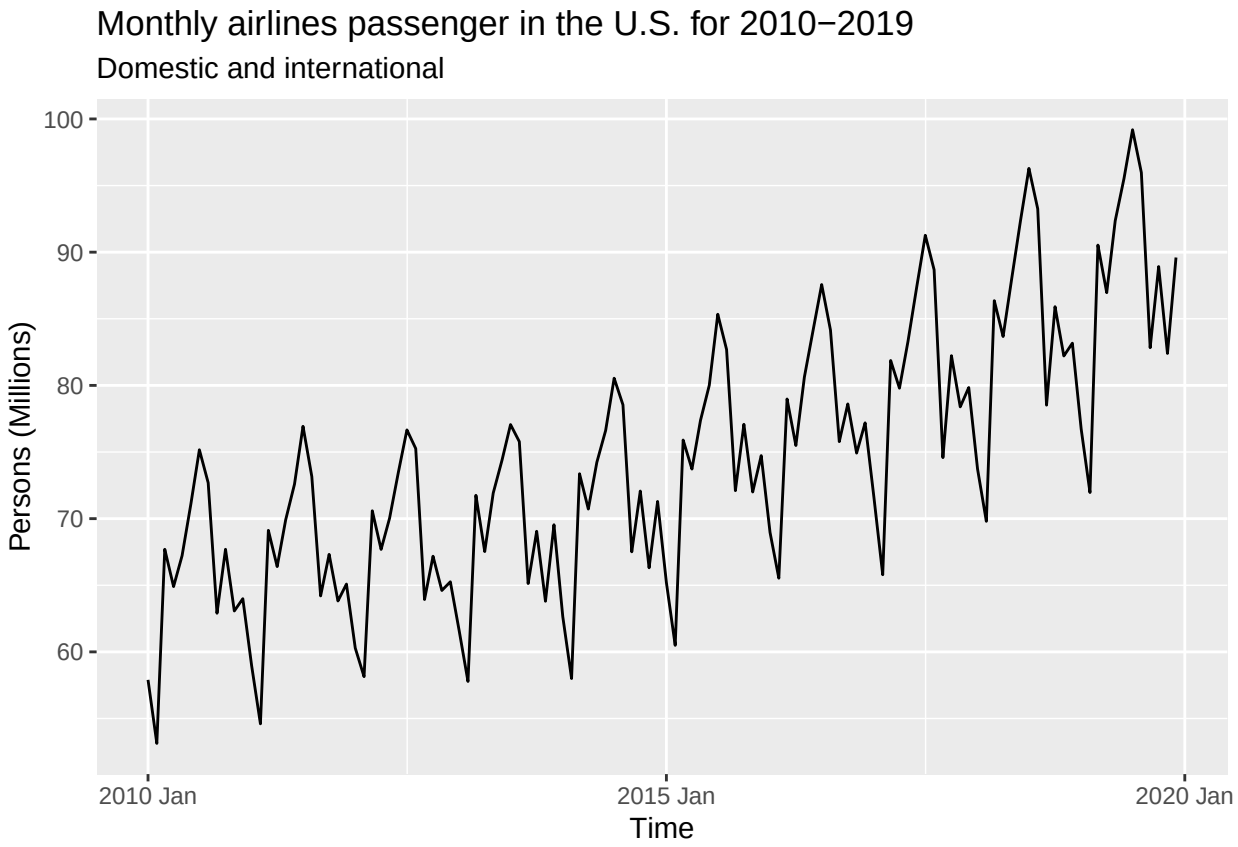
```
passenger <- read_csv("./data/Passengers_2019.csv")
passenger
```

```
## # A tibble: 120 x 3
##     Year Month    TOTAL
##    <dbl> <dbl>    <dbl>
##  1  2010     1 57895059
##  2  2010     2 53134779
##  3  2010     3 67703397
##  4  2010     4 64896774
##  5  2010     5 67223086
##  6  2010     6 71096629
##  7  2010     7 75169345
##  8  2010     8 72688674
##  9  2010     9 62903857
## 10  2010    10 67700757
## # ... with 110 more rows
```

```
df.ts <- ts(passenger[,3], start = c(2010,1), end=c(2019,12), frequency = 12)


## convert dataframe to  tsibble abject
df <- df.ts %>%
  as_tsibble(df.ts) %>%
  mutate(Total = round(value/1000000,2),
         log_total = log(Total)) %>%
  dplyr::select(index, Total, log_total)
```

Descriptive Statistics

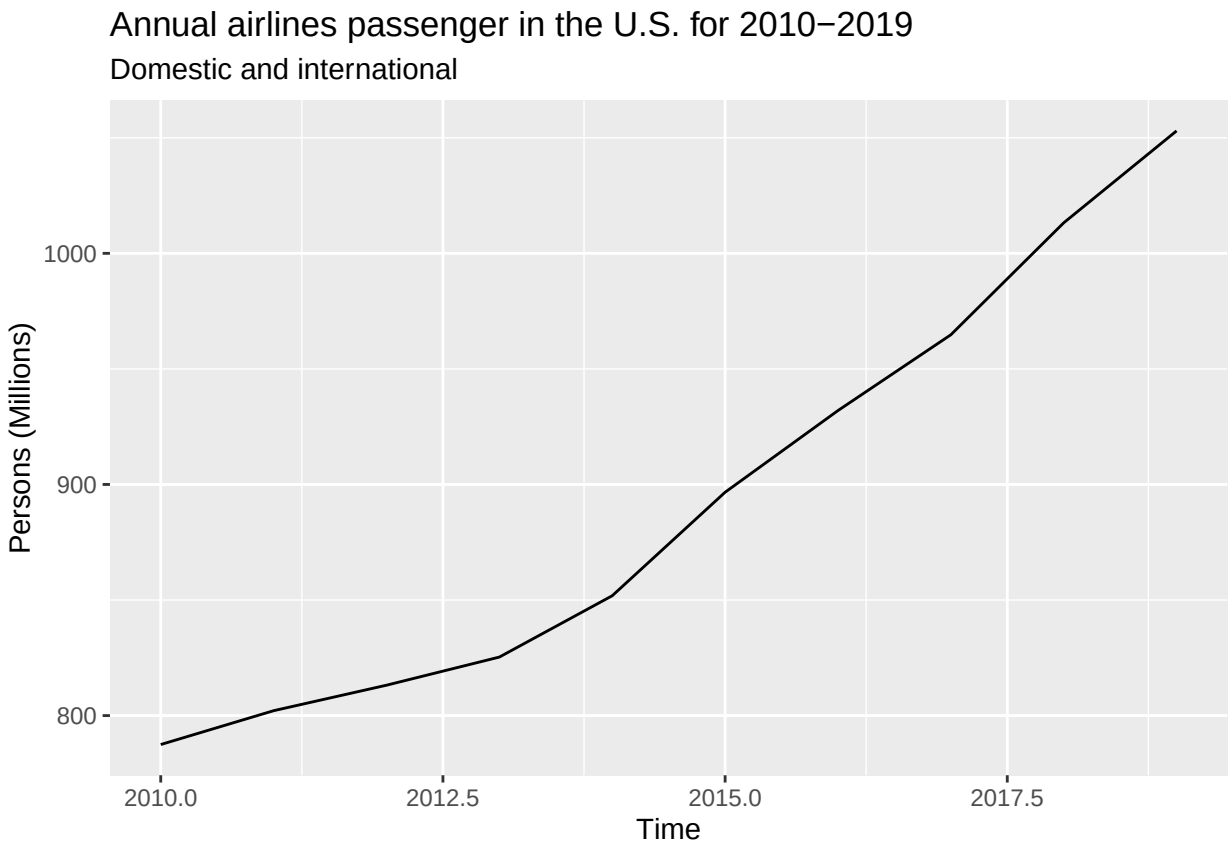
- Produce a time plot of the data and describe the patterns in the graph. Identify any unusual or unexpected fluctuations in the time series.
- Is there any indication of non-stationarity?



We note a couple of visuall indications of non-stationarity:

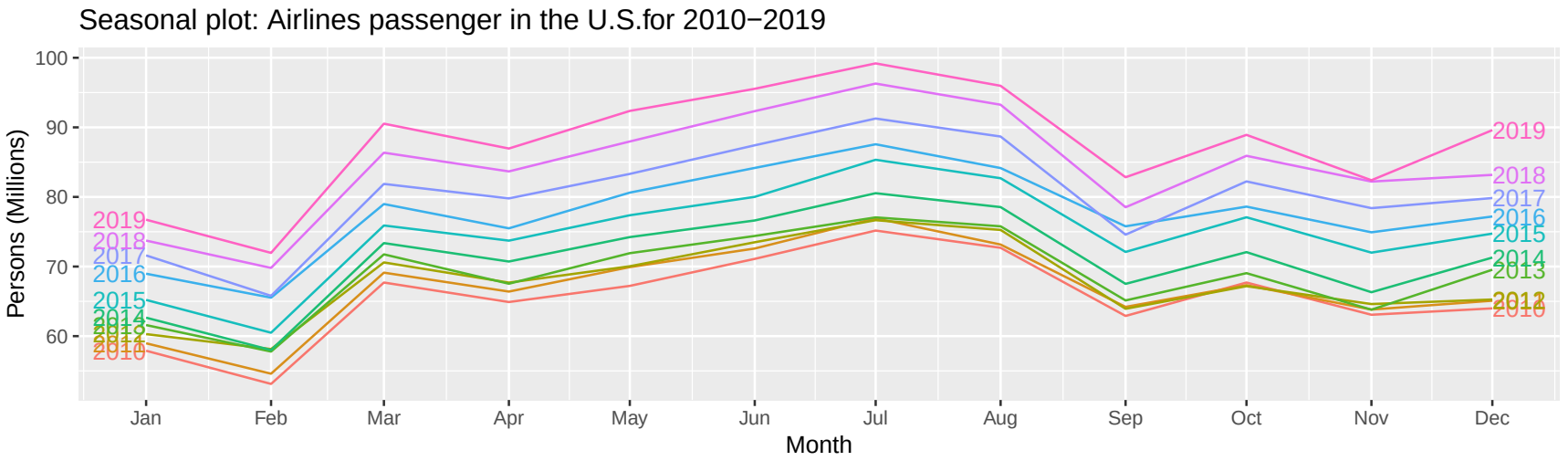
- **The mean of the process appears to be increasing as time increases**
- **The variance of the process (i.e., the fluctuations) appears to be smaller at the beginning and much larger at the end of the time series**
- **The monthly data appears to exhibit a seasonality pattern**

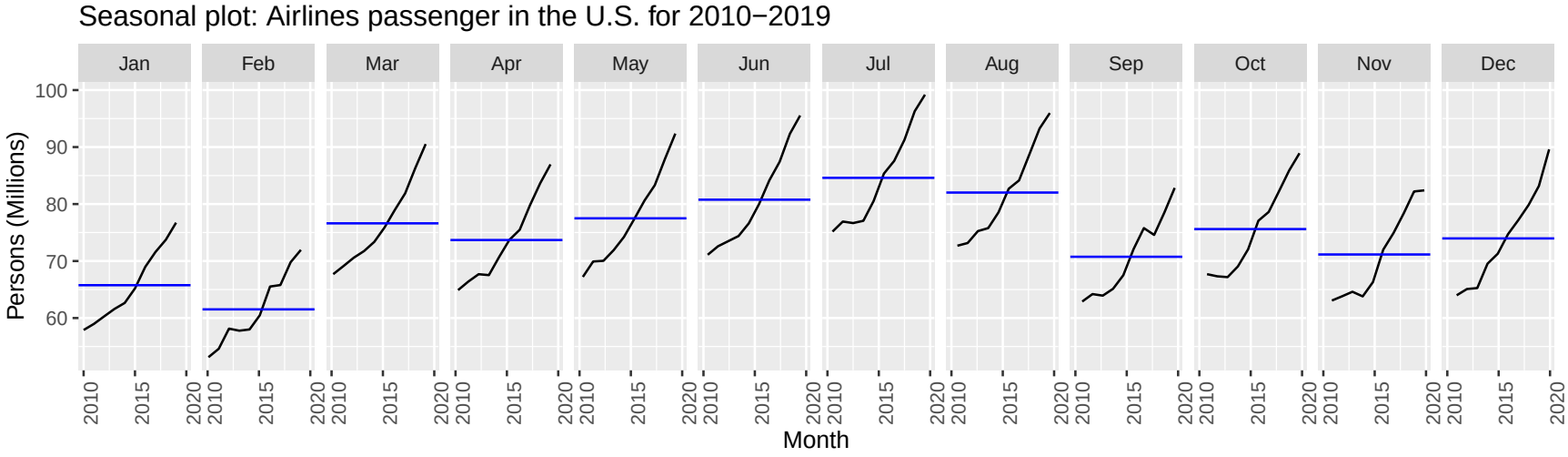
- To have a clearer view of the trend, we can remove the seasonal movement by aggregating the data to the annual level.



The number of passengers is increasing as time increases

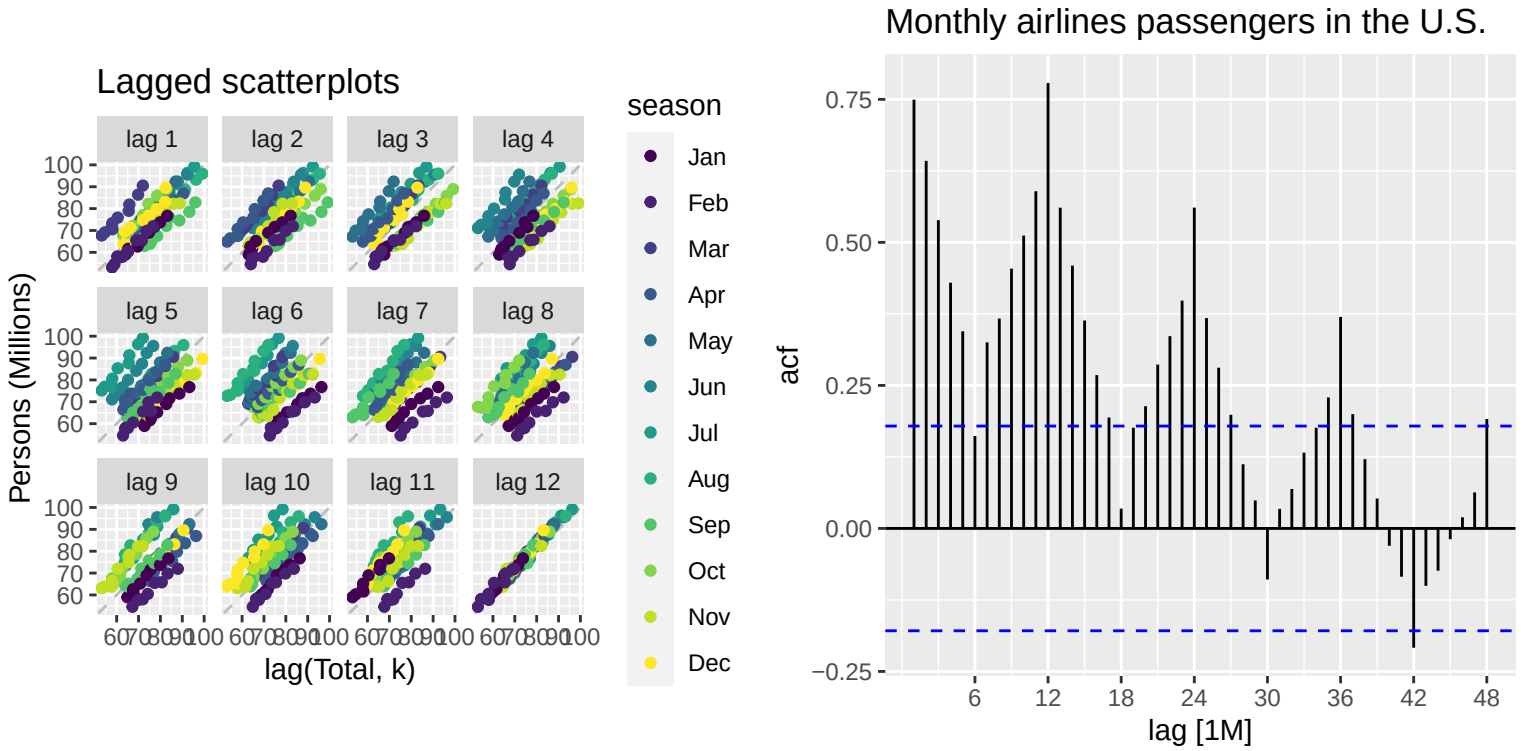
- We can view seasonality by using different types of plots.





From these plots, We see that there appears to be an increase in passengers between April and July each year, with the least amount of airline passengers in January and February of each year.

- One way to check for autocorrelation is to examine the ACF and lag plots. They are also helpful in determining if a time series has a trend and seasonality.
- What do you see in these two plots?



The lag plot shows a strong positive correlation, especially lag 12, reflecting the strong seasonality in the data.

The ACF decays very slowly; this indicates a trend effect. It also exhibits a larger correlation at around every 12th; this is an indication that there is a seasonal (or cyclic) component in our data.

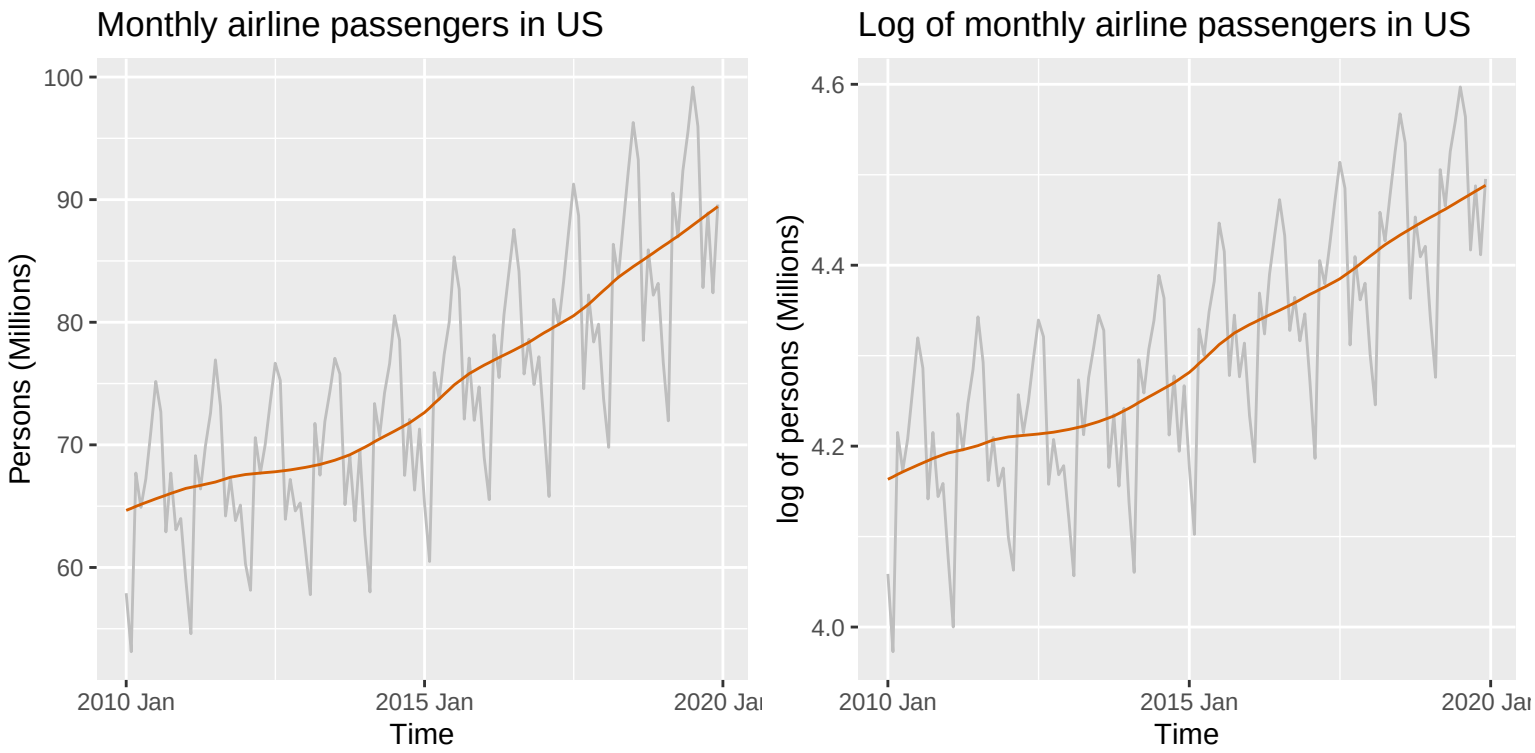
Model Development

Time series decomposition

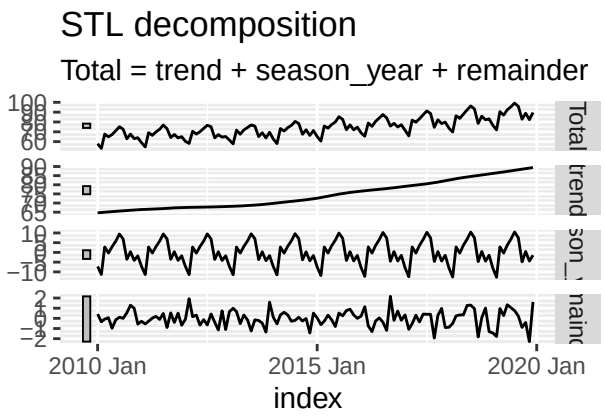
a)- Use both additive and multiplicative decomposition to remove trend and seasonal movement from the air passenger data set.

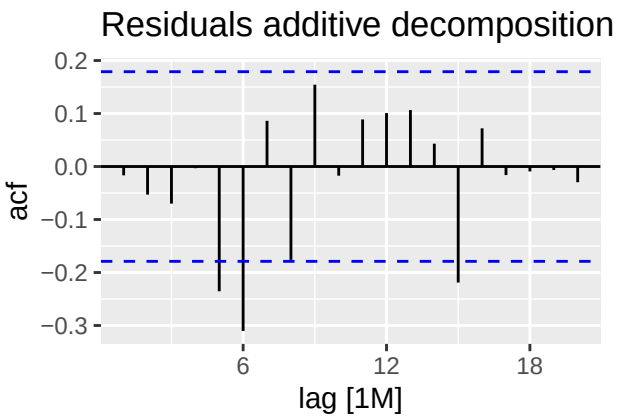
b)- Check the random component. Is it stationary?

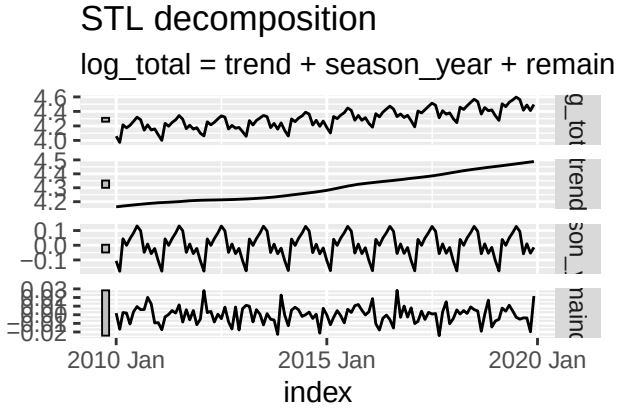
```
dcmp_add <- df %>%
  model(stl = STL(Total))

dcmp_multi <- df %>%
  model(stl = STL(log_total))

p31 <- components(dcmp_add) %>%
  as_tsibble() %>%
  autoplot(Total, colour="gray") +
  geom_line(aes(y=trend), colour = "#D55E00") +
  labs(y = "Persons (Millions)", x="Time",
    title = "Monthly airline passengers in US")

p32 <- components(dcmp_multi) %>%
  as_tsibble() %>%
  autoplot(log_total, colour="gray") +
  geom_line(aes(y=trend), colour = "#D55E00") +
  labs(y = "log of persons (Millions)", x="Time",
    title = "Log of monthly airline passengers in US")

grid.arrange(p31,p32, nrow = 1, ncol =2)
```

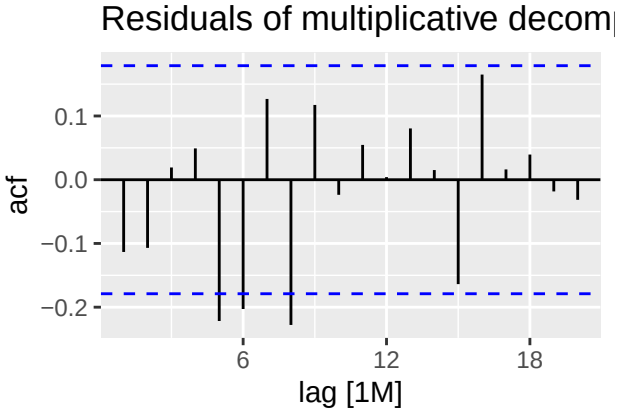



```
p33 <- components(dcmp_add) %>% autoplot()

p34<- components(dcmp_add)%>%
  ACF(remainder) %>%
  autoplot() + labs(title="Residuals additive decomposition")

p35 <- components(dcmp_multi) %>% autoplot()

p36<- components(dcmp_multi)%>%
  ACF(remainder) %>%
  autoplot() + labs(title="Residuals of multiplicative decomposition")

grid.arrange(p33,p34,p35 ,p36, nrow = 2, ncol = 2)
```









Since the variance or the fluctuations of this time series appear to be smaller at the beginning and much larger at the end of the time series, multiplicative is more appropriate

Looking at the residual plots from this decomposition, we note that although they are stationary, they do not appear to be white noise. This means that the decomposition method eliminates the deterministic components from this specific time series; still, some correlation remains in the data.

Modeling Deterministic Trend and Seasonality

- a)- Fit the following regression models to the logarithm of the number of airline passengers data with a linear trend, quadratic trend, and seasonal dummies variables.

$$log(AP_t) = \beta_0 + \beta_1 \cdot t + \epsilon_t$$

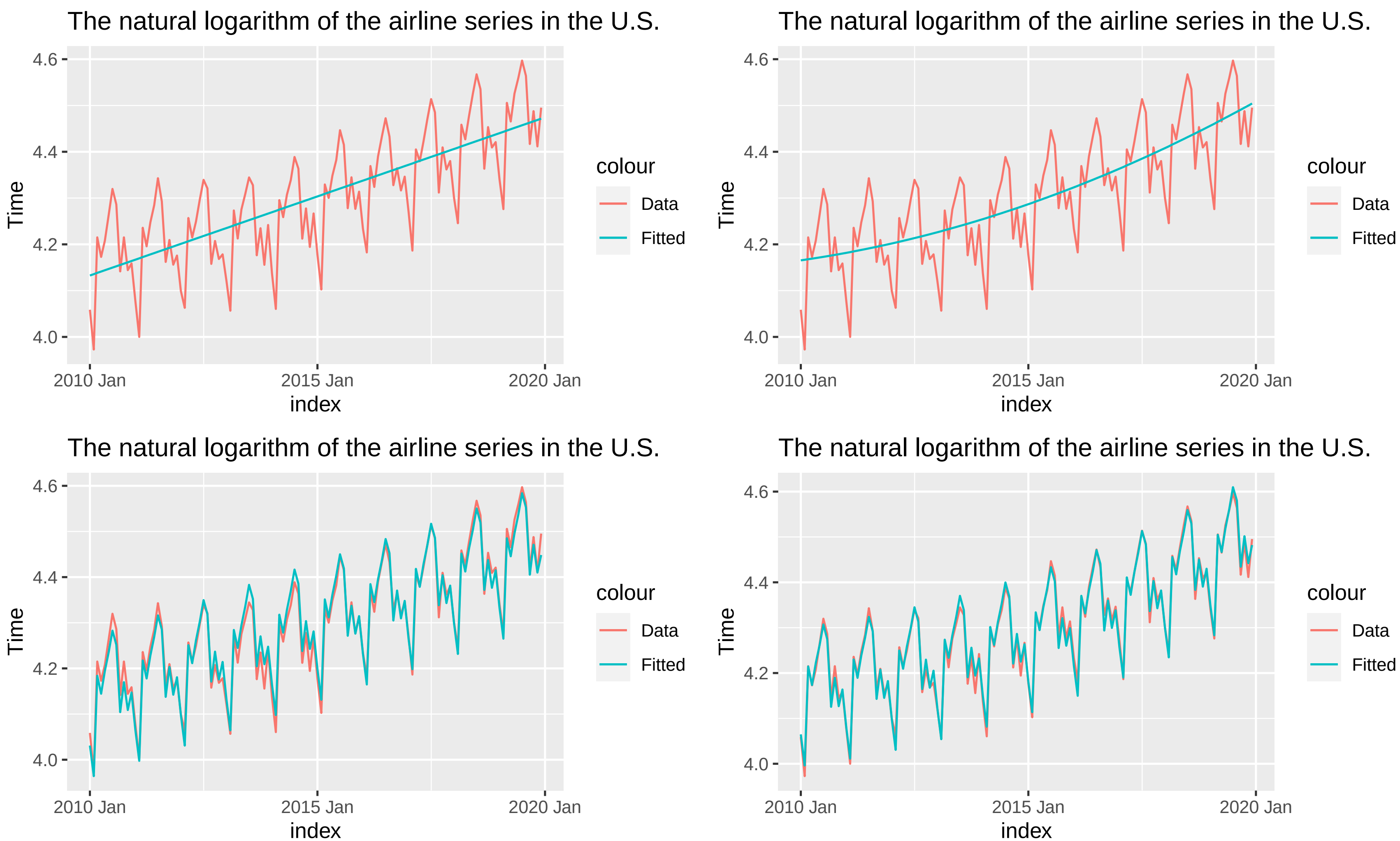
$$log(AP_t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \epsilon_t$$

$$log(AP_t) = \beta_0 + \beta_1 \cdot t + \sum_{i=2}^{12} \beta_i Month_i + \epsilon_t$$

$$log(AP_t) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \sum_{i=2}^{12} \beta_i Month_i + \epsilon_t$$

```
fit_linear <- df %>%
  model(trend_model = TSLM(log_total ~ trend()))


fit_quadratic <- df %>%
   model(trend_model = TSLM(log_total ~ trend()+I(trend()^2)))

fit_linear_season <- df %>%
   model(trend_model = TSLM(log_total ~ trend()+ season()))

fit_quadratic_season <- df %>%
  model(trend_model = TSLM(log_total ~ trend()+I(trend()^2)+ season()))
```

b)- Plot the fitted values against time and against the airline passenger time series? What is the best-fitting model?



From the fitted value plots, we can see that the first two models capture the trend quite well but do not capture the seasonal fluctuations.

The third model with linear trend and seasonal variables attempts to capture the seasonal effect; however, it underestimates the observed data.

The fourth model with quadratic trend and seasonal variables seems to do a good job capturing both trend and seasonal movement

in the data.

c)- How do we interpret the values of the coefficients?

```
fit_linear %>%  report()
```

```
## Series: log_total
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21177 -0.05431  0.01216  0.06556  0.16998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.129843   0.016103  256.47   <2e-16 ***
## trend()     0.002847   0.000231   12.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08765 on 118 degrees of freedom
## Multiple R-squared: 0.5629,  Adjusted R-squared: 0.5592
## F-statistic:   152 on 1 and 118 DF, p-value: < 2.22e-16
```

```
fit_quadratic_season  %>% report()
```

```
## Series: log_total
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0385737 -0.0081894  0.0005876  0.0080683  0.0341167
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.064e+00  5.878e-03 691.295  < 2e-16 ***
## trend()         1.028e-03  1.541e-04   6.669 1.20e-09 ***
## I(trend()^2)    1.455e-05  1.233e-06  11.795  < 2e-16 ***
## season()year2  -6.964e-02  6.484e-03 -10.740  < 2e-16 ***
## season()year3   1.475e-01  6.485e-03  22.752  < 2e-16 ***
## season()year4   1.055e-01  6.485e-03  16.263  < 2e-16 ***
```

```
## season()year5   1.527e-01  6.486e-03  23.548  < 2e-16 ***
## season()year6   1.914e-01  6.487e-03  29.503  < 2e-16 ***
## season()year7   2.354e-01  6.488e-03  36.277  < 2e-16 ***
## season()year8   2.016e-01  6.490e-03  31.068  < 2e-16 ***
## season()year9   5.124e-02  6.492e-03   7.894 2.85e-12 ***
## season()year10  1.140e-01  6.494e-03  17.557  < 2e-16 ***
## season()year11  5.011e-02  6.496e-03   7.715 7.00e-12 ***
## season()year12  8.553e-02  6.498e-03  13.163  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0145 on 106 degrees of freedom
## Multiple R-squared: 0.9893,  Adjusted R-squared: 0.9879
## F-statistic: 750.7 on 13 and 106 DF, p-value: < 2.22e-16
```

In models with seasonal dummy variables, all trend and dummy variables are statistically significantly different from zero, so we can conclude that the trend and seasonal effect are significant in our data.

The interpretation of the linear trend coefficient in the first model is that the number of passengers increases 0.28% per year on average. In the fourth model with the quadratic trend, since both coefficients of the linear trend and the quadratic trend are positive, the coefficient of the linear trend increases over time.

In the model with the seasonal variable, January is the base month, and the other coefficients are changes in the average number of passengers compared to the January holding trend variable constant. For example, the number of passengers is almost 7% lower in February compared to January, holding the trend constant

Regression Diagnostic Results

- The OLS procedure for time series data has good asymptotic properties under the following assumption (from Introductory Econometrics by Jeffrey Wooldridge):

1-**Linearity**: The stochastic process $\{(x_{1t}, x_{2t}....x_{kt}, y_t) : t = 1, 2, ..., n\}$ follows the linear model:

$$y_t = \beta_0 + \beta_1 \cdot x_{1t} + ... + \beta_k \cdot x_{tk} + u_t$$

- **The x_{tj} can be lagged dependent**

2-**Ergodic stationarity**: The stochastic process $\{(x_{1t}, x_{2t}....x_{kt}, y_t) : t = 1, 2, ..., n\}$ is stationary and ergodic.

- Random sample assumption for cross-sectional data is replaced with this assumption. Intuitively, This assumption allows observations to be dependent but:
  - Stationarity does require that the any correlation between $(x_{1t}, x_{2t}....x_{kt}, y_t)$ and $(x_{1t+h}, x_{2t+h}....x_{kt+h}, y_{t+h})$ is the same across all time periods t and only depends on h.
  - Ergoidicity does require that correlation between the $(x_{1t}, x_{2t}....x_{kt}, y_t)$ and $(x_{1t+h}, x_{2t+h}....x_{kt+h}, y_{t+h})$ disappears sufficiently quickly as observations get farther apart and $h \to \infty$

3- **No Perfect Collinearity**

4- **Zero conditional mean or predetermined regressors** All explanatory variables $(x_{t1}, ..., x_{tk})$ are uncorrelated with the contemporaneous error term u_t:

$$E(x_{tk} \cdot u_t) = 0$$

- For all t and k.
- This assumption limits how ϵ_t is related to the explanatory variables in other periods.

5- **Homoskedasticity**: The error u_t are contemporaneously homoskedastic

$$Var(u_t|x_{it}) = \sigma^2$$

6- **No serial correlation**: For all time period $t \neq s$ and i variables:

$$E(u_t u_s|x_{it}x_{is}) = 0$$

- **Under assumptions 1 to 4, the OLS estimators are consistent**

- **Under these six assumptions, the OLS estimators are asymptotically normally distributed, and the usual OLS standard errors, t statistics, F statistics, and LM statistics are asymptotically valid.**

Test for serial correlation Serial correlation means that error terms from different (usually adjacent) periods are correlated.

The serial correlation will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency.

With a positive serial correlation, the OLS estimates of the standard errors will be smaller than the actual standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are, inflating t statistics. There will be a tendency to reject null hypotheses when they should not be rejected.

- There are different types of serial correlation.

  1- First-order AR(1) serial correlation:

$$u_t = \rho u_{t-1} + e_t$$

- $\rho < 1$
- u_t is the measured error term
- e_t is an uncorrelated random variable with mean zero and variance σ_e^2

We can use the Durbin-Watson test for AR(1) serial correlation. The null hypothesis and the Durbin-Watson (DW) statistic are:

- $H_0 : \rho = 0$
- $H_a : \rho \neq 0$

$$DW = \frac{\sum_{t=2}^{T}(\hat{u_t} - \hat{u}_{t-1})^2}{\sum_{t=2}^{T} \hat{u_t}^2}$$

2- Higher-Order AR(p) Serial Correlation:

$$u_t = \rho u_{t-1} + \rho_2 u_{t-2} + ... + \rho_q u_{t-q} + e_t$$

- To test for higher-order serial correlation, we can use the Ljung-Box test. The null hypothesis and the Ljung-Box statistic are:
  - $H_0 : \rho_1 = \rho_2 = ... = \rho_q = 0$

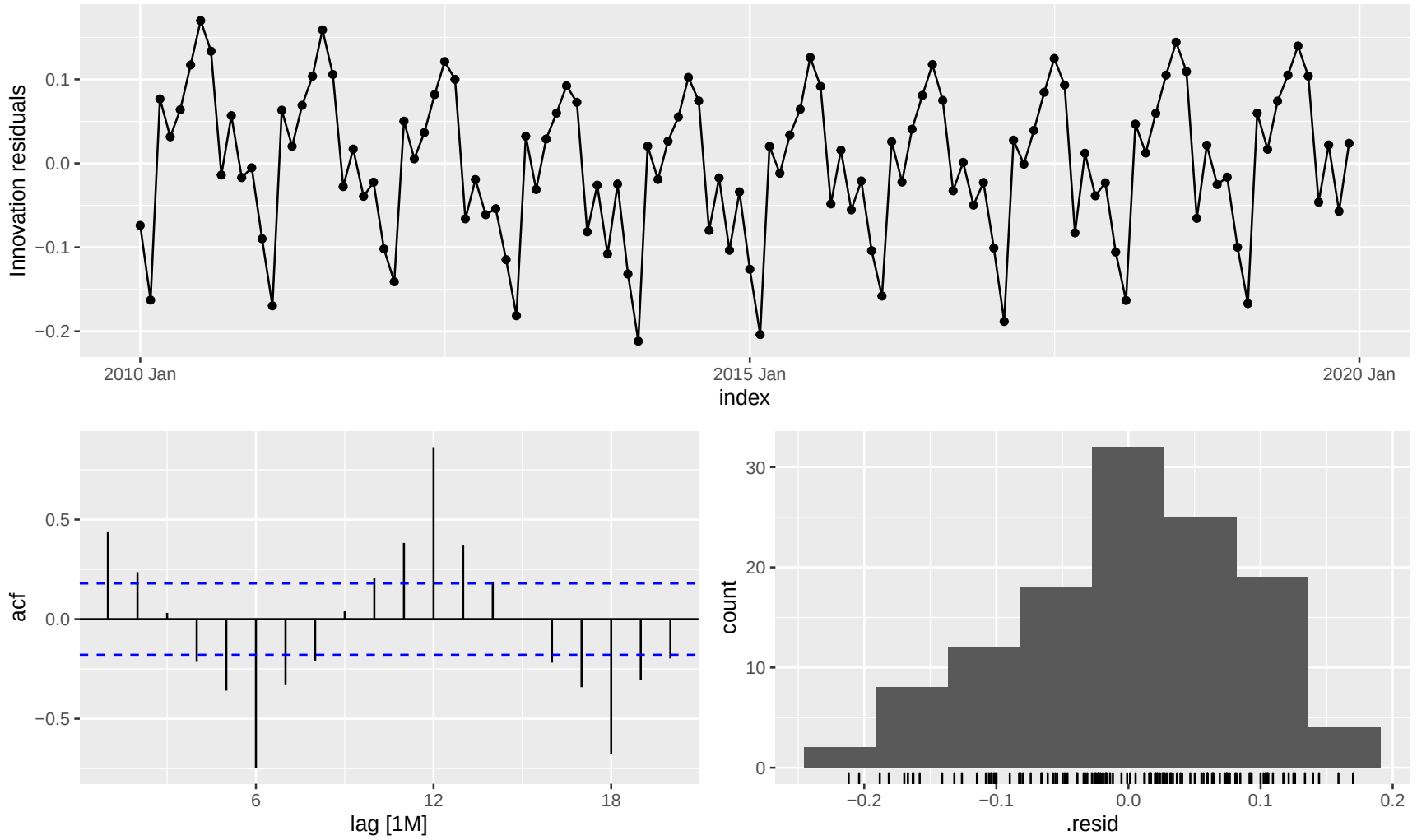
$$Q_h = T(T+2)\sum_{j=1}^{h} \frac{\hat{\rho}_j^2}{T-j}$$

- where h represents the maximum lag.

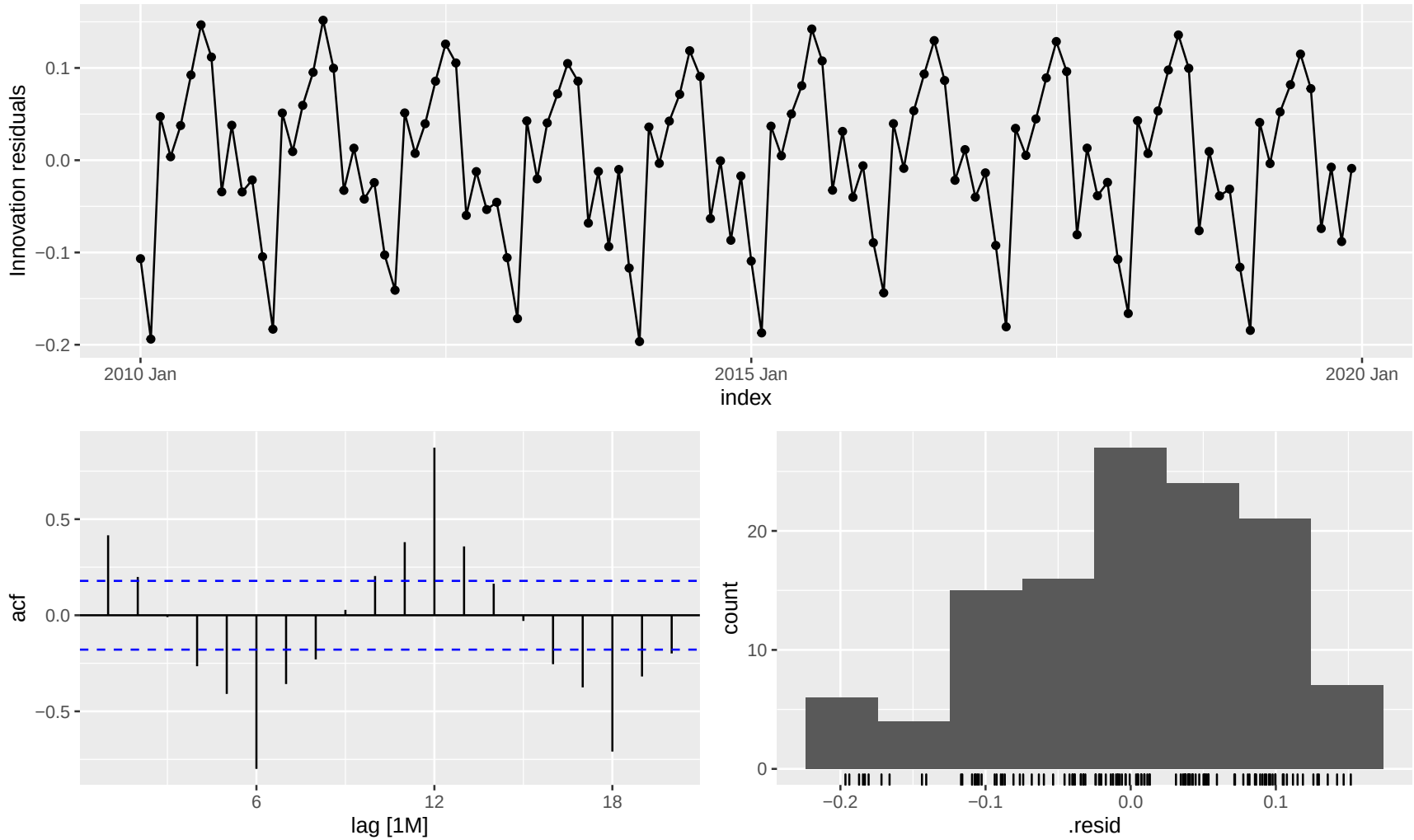
- Under the null hypothesis $Q_h \sim \chi_h^2$ since asymptotically $\hat{\rho}_j \sim N(0, \frac{1}{T})$ under the null and therefore, the statistic is proportional to the sum of squared standard normal random variables.

a)- Plot the residuals against time and the fitted values and try to check the 6 CLM assumptions. Do these plots reveal any problems with the model?

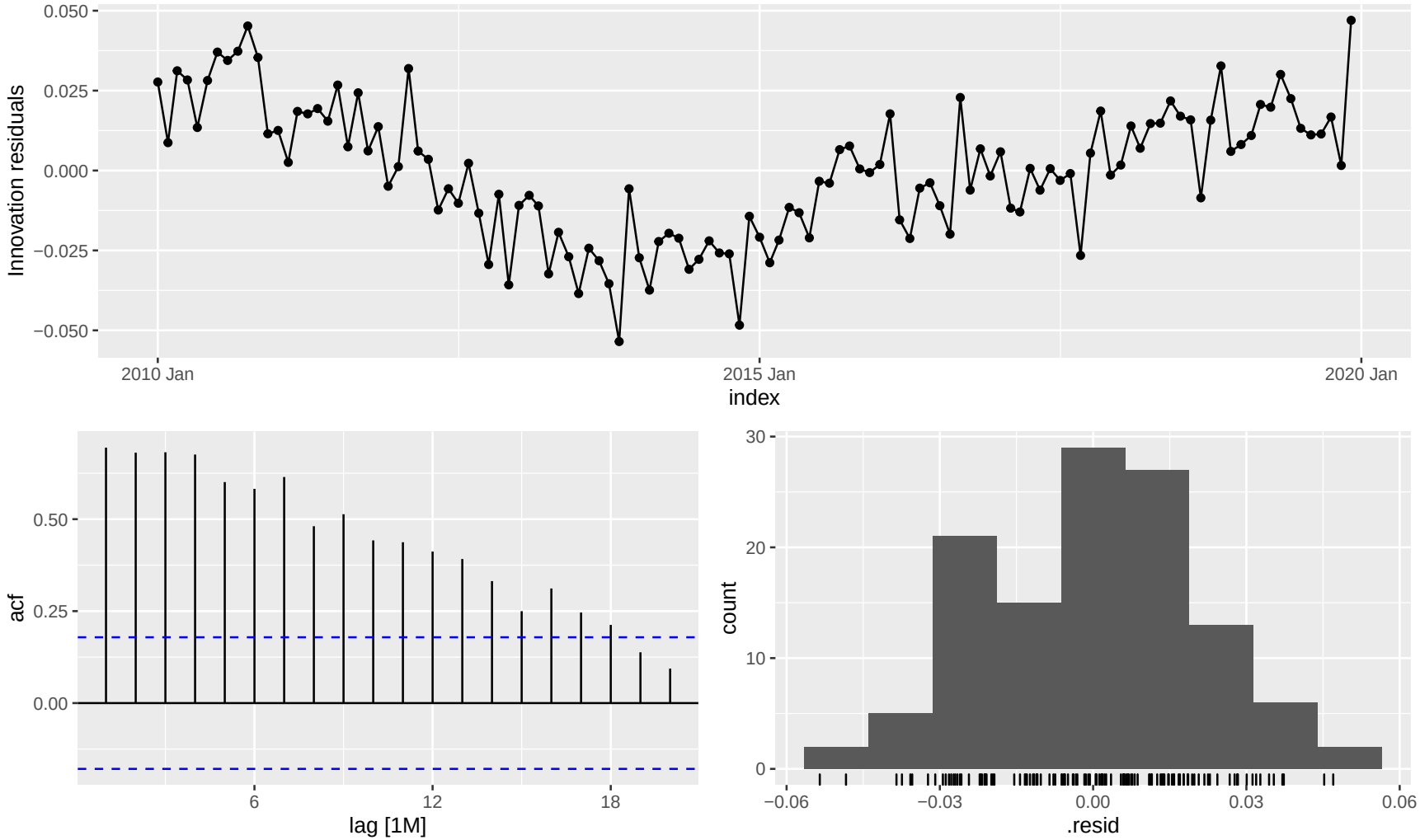
```
fit_linear %>% gg_tsresiduals()
```



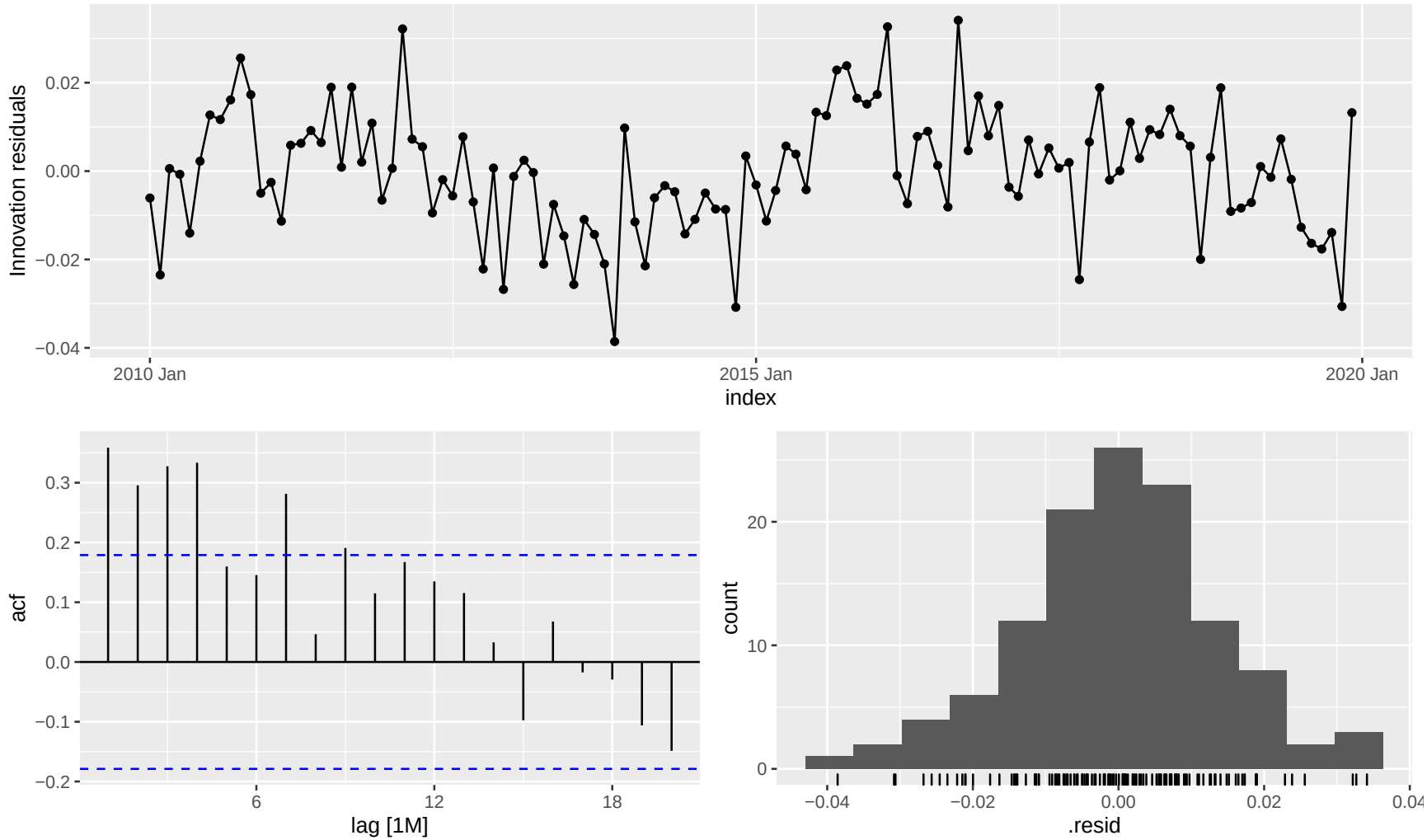

```
fit_quadratic %>% gg_tsresiduals()
```




```
fit_linear_season %>% gg_tsresiduals()
```



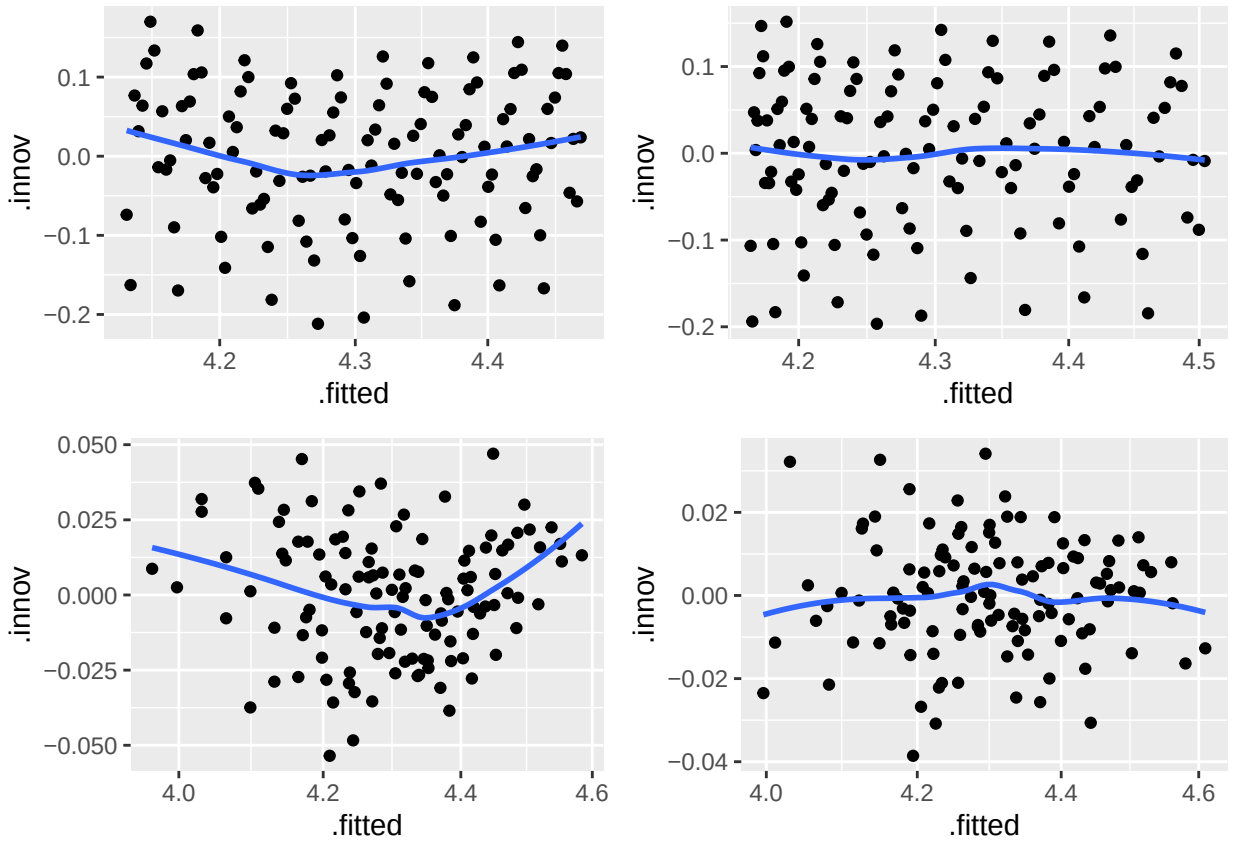

```
fit_quadratic_season %>% gg_tsresiduals()
```




```
p41<- augment(fit_linear) %>%
  ggplot(aes(x = .fitted, y = .innov)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  scale_x_log10()
```

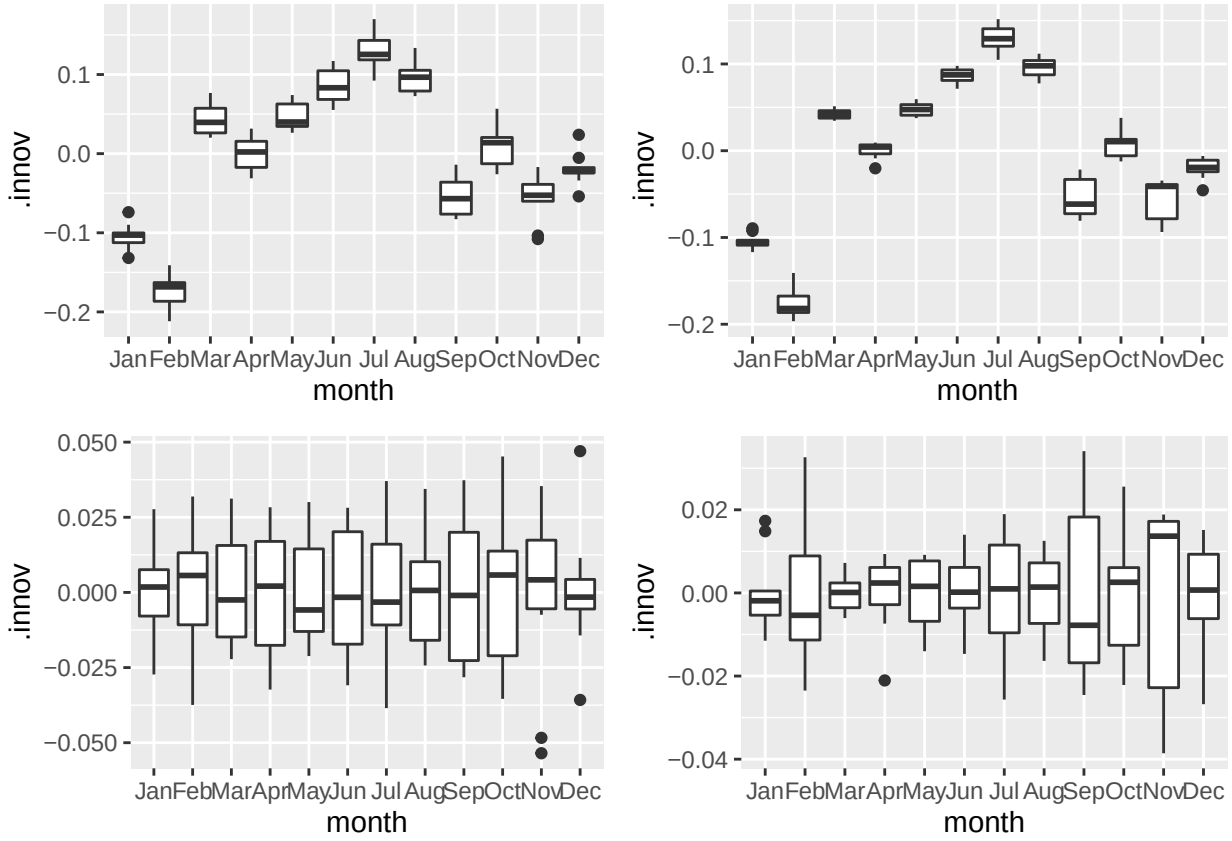
```r
p42<- augment(fit_quadratic) %>%
  ggplot(aes(x = .fitted, y = .innov)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  scale_x_log10()

p43<- augment(fit_linear_season) %>%
  ggplot(aes(x = .fitted, y = .innov)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  scale_x_log10()

p44<- augment(fit_quadratic_season) %>%
  ggplot(aes(x = .fitted, y = .innov)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  scale_x_log10()

grid.arrange(p41,p42,p43,p44, nrow = 2, ncol = 2)
```

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




```
p45<- augment(fit_linear) %>%
  mutate(month = month(index, label = TRUE)) %>%
  ggplot(aes(x = month, y = .innov)) +
  geom_boxplot()

p46<- augment(fit_quadratic) %>%
  mutate(month = month(index, label = TRUE)) %>%
  ggplot(aes(x = month, y = .innov)) +
  geom_boxplot()

p47<- augment(fit_linear_season) %>%
```

```
  mutate(month = month(index, label = TRUE)) %>%
  ggplot(aes(x = month, y = .innov)) +
  geom_boxplot()

p48<- augment(fit_quadratic_season) %>%
  mutate(month = month(index, label = TRUE)) %>%
  ggplot(aes(x = month, y = .innov)) +
  geom_boxplot()

grid.arrange(p45,p46,p47,p48, nrow = 2, ncol = 2)
```



b)- Test for autocorrelation of the residuals

```
augment(fit_linear) %>%
  features(.innov, ljung_box, dof = 2, lag = 24)
```

```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 trend_model    503.         0
```

```
augment(fit_quadratic) %>%
  features(.innov, ljung_box, dof = 3, lag = 24)
```

```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 trend_model    542.         0
```

```
augment(fit_linear_season) %>%
  features(.innov, ljung_box, dof = 13, lag = 24)
```

```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 trend_model    591.         0
```

```
augment(fit_quadratic_season) %>%
  features(.innov, ljung_box, dof = 14, lag = 24)
```

```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 trend_model    118.         0
```

Let's focus on the fourth model here since it has the best performance.

The first assumption is the linear population model. There's no way to check this assumption. We just assumed that it is true, and If it's not true, what we're doing is fitting the best linear model to the data.

The second assumption is *Ergodic stationarity*. By including trend and seasonal variables, we eliminate these deterministic parts, and this assumption should be satisfied, but we have to check the residual plots. From the ACF plot, we can see that it looks stationary, but residuals are serially correlated.

The third assumption is no perfect multicollinearity. Since R didn't report any warning or missing coefficients when we estimated

the model, this assumption is violated.

Assumption-4 is the zero conditional mean assumption. So one of the ways that we look at this assumption is to look at these residuals versus the fitted value plot. So if this assumption is satisfied, there's no relationship between the fitted values and the residuals, and we should see just kind of this Straight line in the residuals versus fitted value plot

But we can see this U-shaped going on in the middle of the plot. So in this particular case, we probably couldn't justify that the zero conditional mean assumption is satisfied here. Probably there are some other critical omitted variables like holidays that should be included in the model.

Assumpton-5 is homoskedasticity. If we look at the residuals v.s fitted values to plot, we can see roughly the same variance throughout the range residuals, and there are no severe deviations from this assumption.

The last and the most common problem in time series is the serial correlation. We can check this assumption by using ACF and also the ljung_box test.

From the ACF plot, it is clear that residuals are serially correlated. Also, Since the ljung_boxtest statistic is high ad the p-value is small, we reject the null hypothesis of no autocorrelation and conclude that the residuals are autocorrelated.

It is also helpful that the errors have normal distribution to produce prediction intervals easily. We can check those with the histogram of residuals. In this case, residuals almost look like a normal distribution.

Model Selection

- Recall that **Akaike Information Criterion (AIC)** is defined as:

$$AIC = -2 \times logL_k + 2 \times k$$

- where $logL_k$ is the maximized log-likelihood and k is the number of parameters in the model.

- One could normalize AIC by n, the number of observations used to estimate the model, and obtain

$$AIC = \frac{-2logL_k + 2k}{n} \approx ln(\hat{\sigma}^2) + \frac{2k}{n} + c$$

- where $\hat{\sigma}^2$ denotes the MLE of σ^2 and c is some constant.

- Also **Bayesian Information Criteria (BIC)** is given by:

$$BIC = ln(\hat{\sigma}^2) + k\frac{ln(n)}{n}$$

- Note that BIC imposes a greater penalty for the number of estimated model parameters than AIC. As such, BIC will always give a model whose number of parameters is no greater than that chosen under AIC.

- The information criterion model selection process should **NOT** be used as a substitute for careful examination of the characteristics of the estimated autocorrelation and partial autocorrelation functions; they can be used as supplementary guidelines.

- Critical examination of the residuals for model inadequacies should always be included as a major aspect of the overall model selection process.

a)- Now compare the three estimated models using AIC and BIC

```
glance(fit_linear) %>%
  dplyr::select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared      CV   AIC  AICc   BIC
##           <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1         0.559 0.00782 -580. -580. -572.
```

```
glance(fit_quadratic) %>%
  dplyr::select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared      CV   AIC  AICc   BIC
##           <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1         0.569 0.00773 -582. -582. -571.
```

```
glance(fit_linear_season) %>%
  dplyr::select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared       CV   AIC  AICc   BIC
##           <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1         0.972 0.000541 -902. -898. -863.
```

```
glance(fit_quadratic_season) %>%
  dplyr::select(adj_r_squared, CV, AIC, AICc, BIC)
```

```
## # A tibble: 1 x 5
##   adj_r_squared       CV    AIC  AICc   BIC
##           <dbl>    <dbl>  <dbl> <dbl> <dbl>
## 1         0.988 0.000238 -1001. -996. -959.
```

The model with the lowest AIC, corrected AIC, or BIC score is preferred. Based on all these three criteria, the fourth model with both quadratic trend and seasonal variables is the best model

Forecasting

- Recall that the predictions of y can be obtained using:

$$\hat{y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{t+1}$$

Assuming that the regression errors are normally distributed, an approximate 95% prediction interval associated with this forecast is given by:

$$\hat{y} \pm \cdot 1.96 \hat{\sigma}_u SE$$

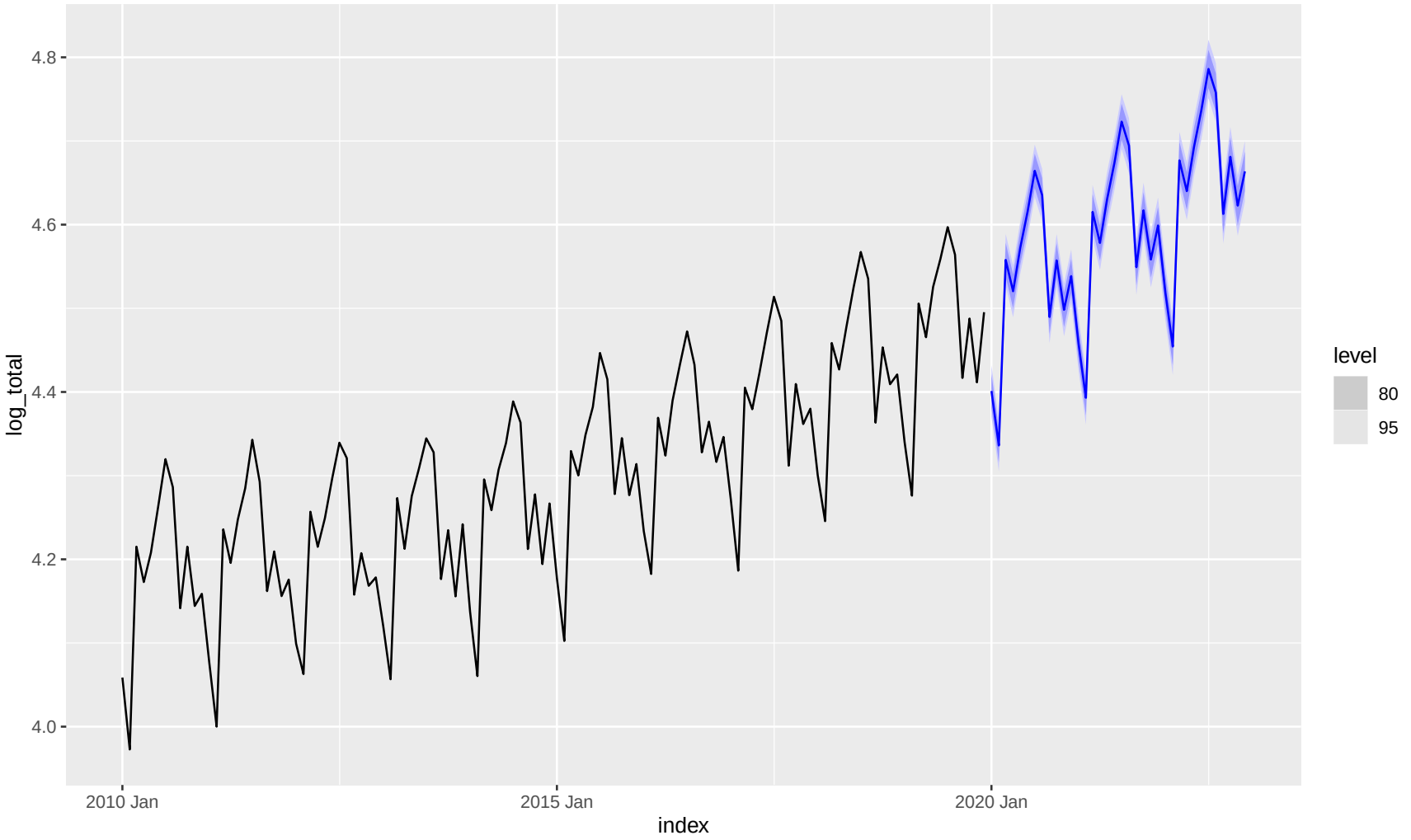
$$SE = \sqrt{1 + \frac{1}{T} + \frac{(x - \bar{x})^2}{(T-1)s_x^2}}$$

- Where T is the total number of observations, s_x is the standard deviation of the observed x values, and $\hat{\sigma}_u$ is the standard error of the regression.

- SE includes two sources of variance or uncertainty in our forecasts. First is σ^2, the variance of the errors in the population, and it does not change with the sample size. Second, the sampling error in $\hat{y}$ arises because we have estimated the β_0 and $beta_1$. Because β_0 and $beta_1$ have a variance proportional to $1/T$, the sampling error can be very small for large samples.

a)- Regardless of your answers to the above questions, use your regression model to predict the monthly airline passengers for 2020, 2021, and 2022. Produce prediction intervals for each of your forecasts.

```
future_df <- new_data(df, n = 36)

fit_quadratic_season %>%
  forecast(new_data = future_df) %>%
  autoplot(df)
```



We can see that the forecasts also have a trend and seasonal movement and fluctuations increase over time. But we all know these forecasts are incorrect due to an unpredicted shock in 2020, Covid-19!!