Report from the Point of View of 1997

Introduction

In the 1950s, a geochemist, Charles David Keeling, began to collect air samples to measure the amount of carbon dioxide in the air, and these measurements would become his life's work. His initial samples indicated that the air contained more CO2 at night compared to the daytime. This finding suggested that plants absorbed CO2 during the day and released CO2 during the evening. Over time, his measurements showed a seasonal cycle, which was consistent with the seasonal patterns of terrestrial vegetation, and the CO2 data also would persistently increase over time. This rise in CO2 prompted questions about preconceived notions around the ocean's ability to absorb CO2 emitted by fossil fuels. If CO2 released by coal, natural gas, and petroleum remained in the air, this could account for the rise in CO2 over the years.

This finding is incredibly important, because the rising rate of CO2 has implications on the Earth's ability to release heat from the atmosphere. If heat isn't able to escape, then this would subsequently result in a warmer climate, which would become known as the "greenhouse effect". However, while the CO2 data demonstrated a clear upward trend over time, it was also marked by periods of decline which could not solely be attributed to plant growth seasonality. A deeper investigation revealed that El Nino weather patterns affected the amount of CO2 that was released in the air by vegetation and soils. Over time, the data would also show earlier seasonal starts, and when this data was layered with temperature data, it had become clear that the concerns of a warming planet years earlier were starting to manifest itself within the data. Since this connection has been established between rising CO2 and corresponding global temperatures, more attention has surrounded the topics of global warming and climate change. As a result, business and political decision makers are starting to incorporate this data into their longer-term strategies; however, their pace of action may be too slow to prevent some of dangers associated with climate change.

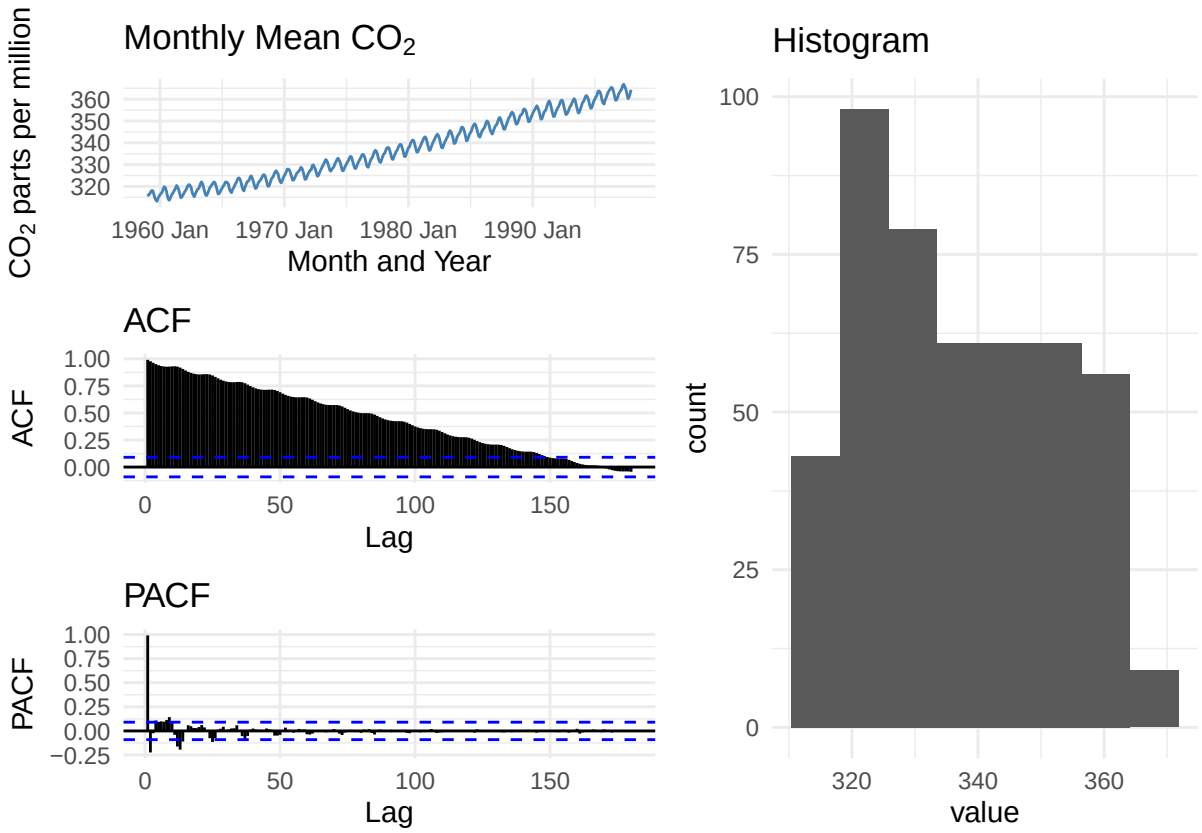
In the analysis below, we will confirm the trends identified above by Keeling himself and use this data to fit a model to make a forecast of CO2 levels in the future.

Exploratory Data Analysis

The data that we will analyze in this report is monthly data on the CO2 levels made at the Mauna Loa observatory from Jan 1959 to Dec 1997. According to the data documentation, the air at Mauna Loa is thought to be representative of much of the Northern Hemisphere and potentially the globe as well, as the observatory is at an altitude of 3400 meters and surrounded by bare lava, which allows for measurement of "background" air that is resistent to day-to-day fluctuations in CO2 levels.

The data is in units of "mole fraction", which according to the data source is "defined as the number of carbon dioxide molecules in a given number of molecules of air, after removal of water vapor. For example, 413 parts per million of CO2 (abbreviated as ppm) means that in every million molecules of (dry) air there are on average 413 CO2 molecules." The data that makes up our dataset was measured daily from Jan 1959 to Dec 1997, but in our dataset appears as an averaged mean per month in ppm units.

In raw form, the data appears as a matrix of doubles that represent the ppm measurements per month and year combination. Let's create some initial EDA plots that will allow us to better understand the data, starting by analyzing the time series, histogram, auto-correlation function (ACF), and partial auto-correlation function (PACF) plots.



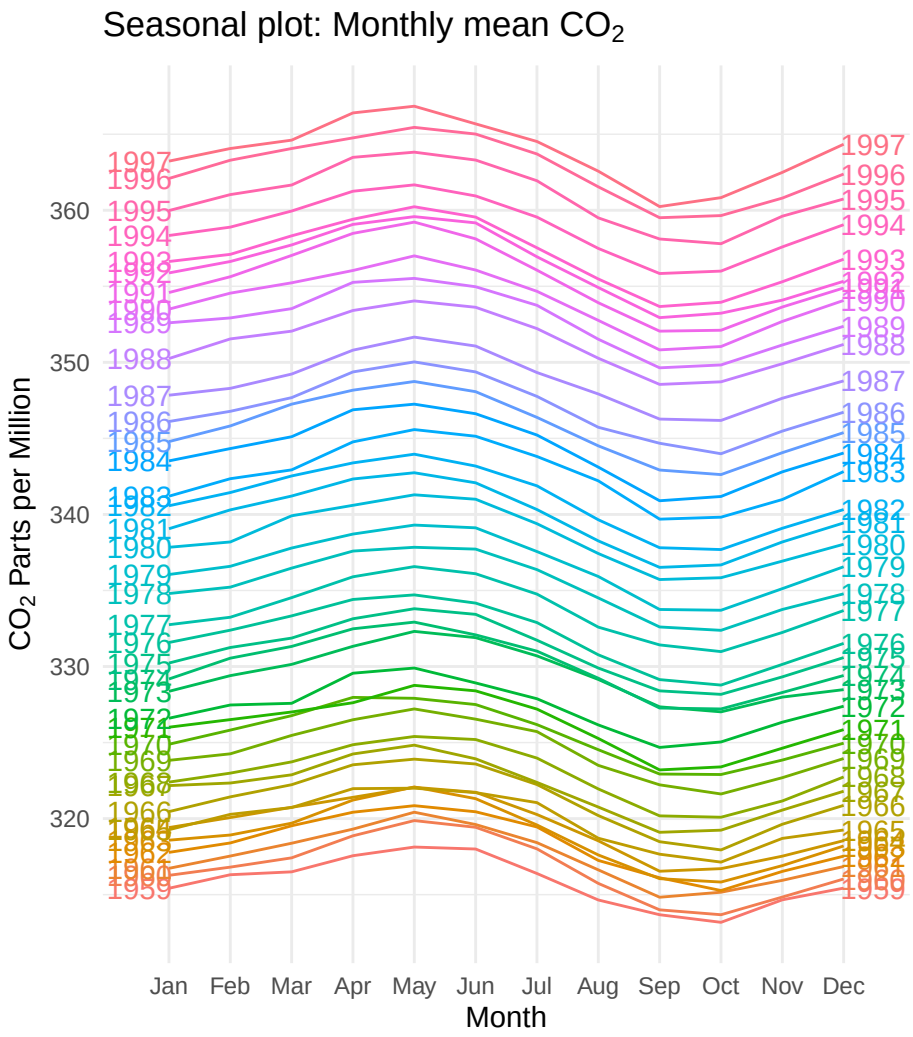
As we can see above, the data seems to follow a clear and increasing trend, with a distinct seasonal pattern that appears as "waves" that we would like to further analyze. The magnitude of the fluctuations do not appear to vary with the time series level, so in terms of decomposition, an additive model would likely fit this series best. The time series does not appear stationary from this plot, as the mean does not appear constant and in fact appears to increase over time, but the variance appears to be constant.

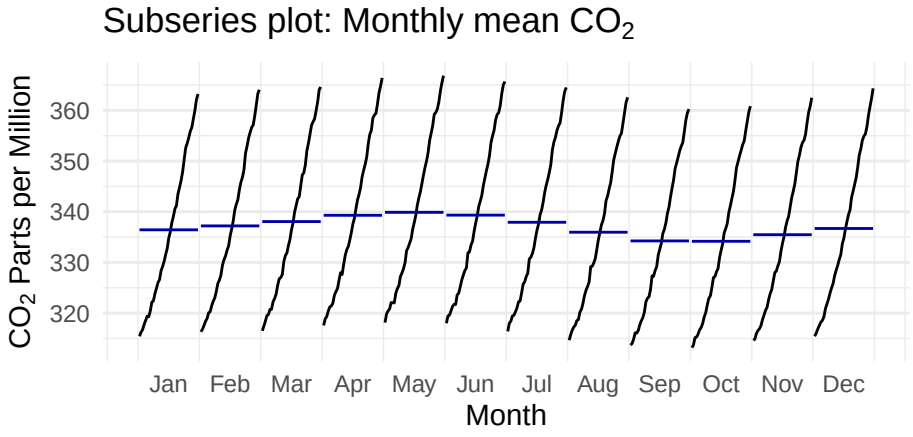
The ACF starts out close to 1 and declines slowly over time, losing significance but staying mostly positive and above the significance line until around lag 140. This slow decline in ACF is what we should see in a time series with a pronounced trend effect, and tracks with what we noticed in the previous time series plot. There appear to be "waves" in the ACF plot similar to the "waves" we also noticed in the time series plot, indicating that the there is a seasonal or cyclic component to our data. Thinking ahead to our modeling, the ACF seems to indicate an autoregressive component in our data generating process, as it declines slowly over time.

The PACF starts out with a single significant positive spike at lag 1, followed by a (relatively smaller) significant negative spike at lag 2, with oscillating clusters of positive and negative lags with much lower levels of significance as the lag number increases. Thinking ahead to our modeling, the PACF seems to indicate an moving average component in our data generating process, as it oscillates between positive and negative over time.

The histogram shows that the CO2 values seem to range between 300 and 380 ppm and do not appear normally distributed, with most values in between these two ranges.

Let's take a closer look at the seasonality in the data. We'll start by creating a season plot and a subseries plot of the data.

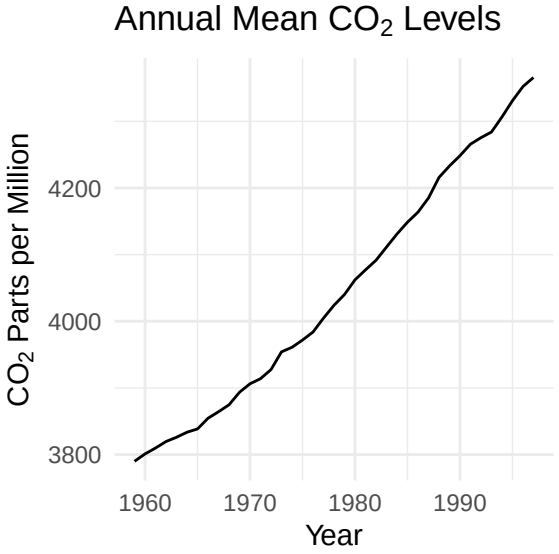




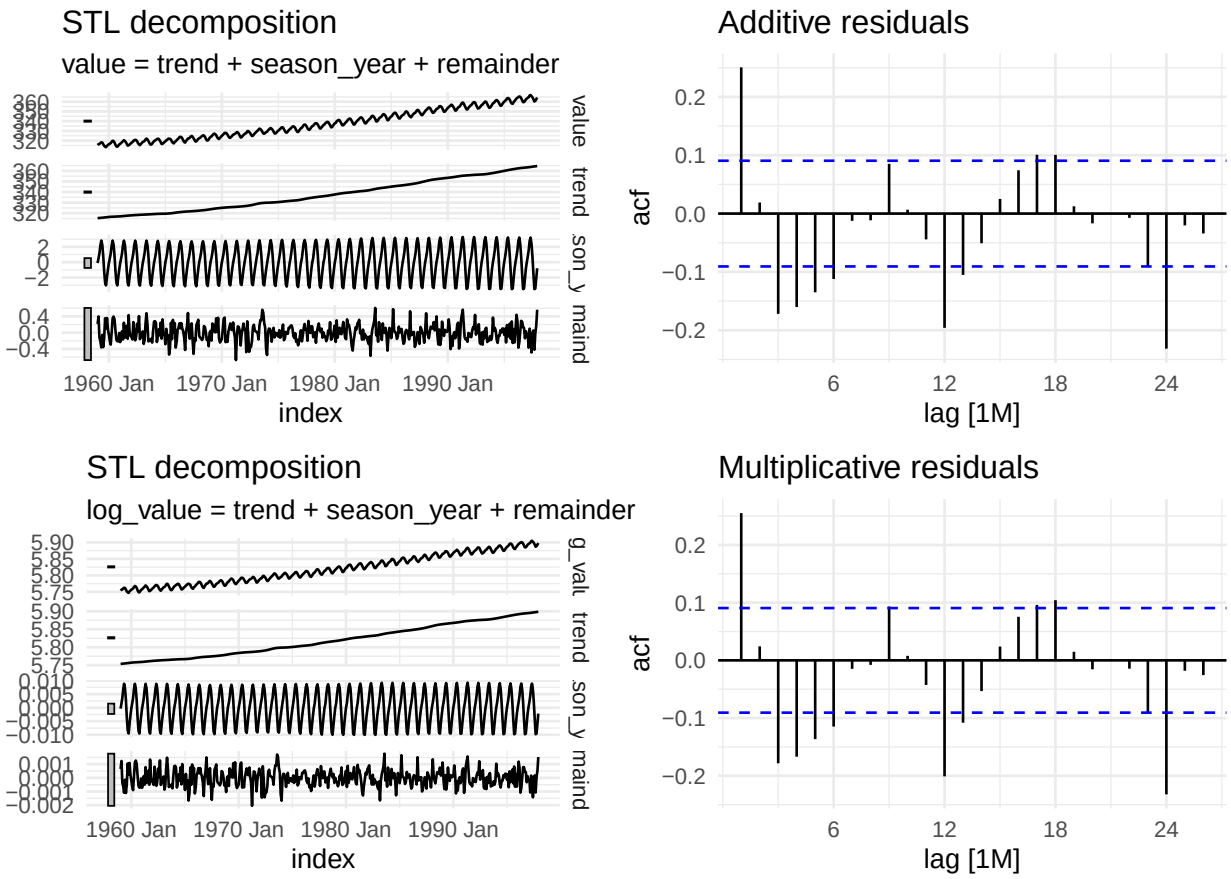
From these plots, we can see that the CO2 levels appear to increase from January through May, then decrease from June through October, hitting a low in October, and then increase again from November through end of the year. The source report does mention that plants and soil absorbing and emitting CO2 could influence these measurements. A possible explanation could be that, in the colder months, we can expect higher CO2 levels as plants die off, and in the warmer months, we can expect lower CO2 levels as plants thrive. Of course,

there is variability across the Northern Hemisphere in what counts as "colder" months and when plants thrive - for example, Hawaiian "winter" is temperate and plants can grow year round, so this could explain why the seasonal variability is slight across the year. We can again see the trend of CO2 levels increasing year by year, but the seasonality effect seems constant year after year without a noticeable increase in the magnitude of the fluctuations across years, again supporting an additive model.

We can confirm that our time series trend is increasing by removing our seasonality components and aggregating our data by year instead of month, as seen below.



We can also use both additive and multiplicative decomposition to remove both the trend and seasonal movements from our dataset and confirm that the variance is stationary. The results from these decompositions are plotted below.



From these plots, we can confirm again that the time series is trending upwards, as seen in the "trend" sections of the decomposition plots. However, upon closer look, the fluctuations within the seasonality of the time series seem to grow slightly larger over time, which we can see in the "season_year" section of the top left plot. In the multiplicative plot on the bottom left, the "season_year" plot appears just slightly more stable, tentatively supporting the idea that this might actually be a multiplicative time series, contrary to our earlier findings. In the next section we should look at applying a log transform on our series prior to modeling.

Looking at the residual plots, the residuals on both appear stationary, meaning that the decomposition methods we are using was able to eliminate deterministic components from the time series. However, they do not appear to be white noise, meaning that there is still correlation in the data.

Let's complete our EDA by running statistical tests to determine whether our model is stationary or non-stationary. We will run both the Augmented Dickey-Fuller (ADF) test and the Phillips Perron (PP) test to do this, under the following hypotheses:

H0: Time series is non-stationary

H1: Time series is stationary

```
##
##  Augmented Dickey-Fuller Test
##
## data:  co2
## Dickey-Fuller = -6.842, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary

##
```

```
##  Phillips-Perron Unit Root Test
##
## data:  co2
## Dickey-Fuller Z(alpha) = -92.68, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

Based on the ADF and PP tests, we can reject the null hypothesis that the time series is non-stationary. This is surprising as based on our visual analysis of the time series plots, the time series does not appear to be stationary with a mean that trends upwards. Because we know that both the ADF and PP tests have low power, we will move forward with the assumption that this time series is non-stationary based on our visual EDA.

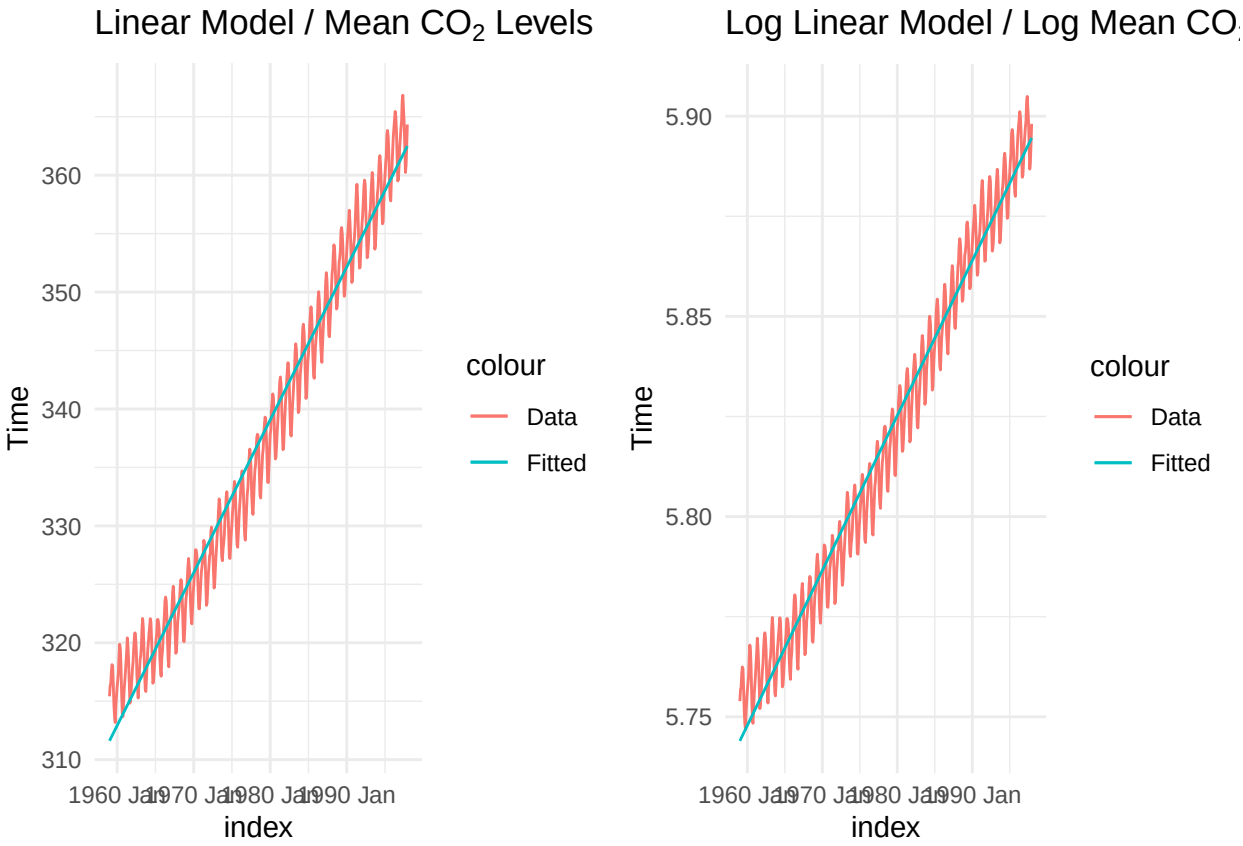
Modeling

Let's start by creating two linear models, one fit on our CO2 data and one fit on the log transform of our CO2 data. Since in our EDA we did notice that the fluctuations within the seasonality of the time series seem to grow slightly larger over time, potentially indicating a multiplicative series, we want to try out this log transform to see if it reduces variance in our model and leads to smaller residuals.

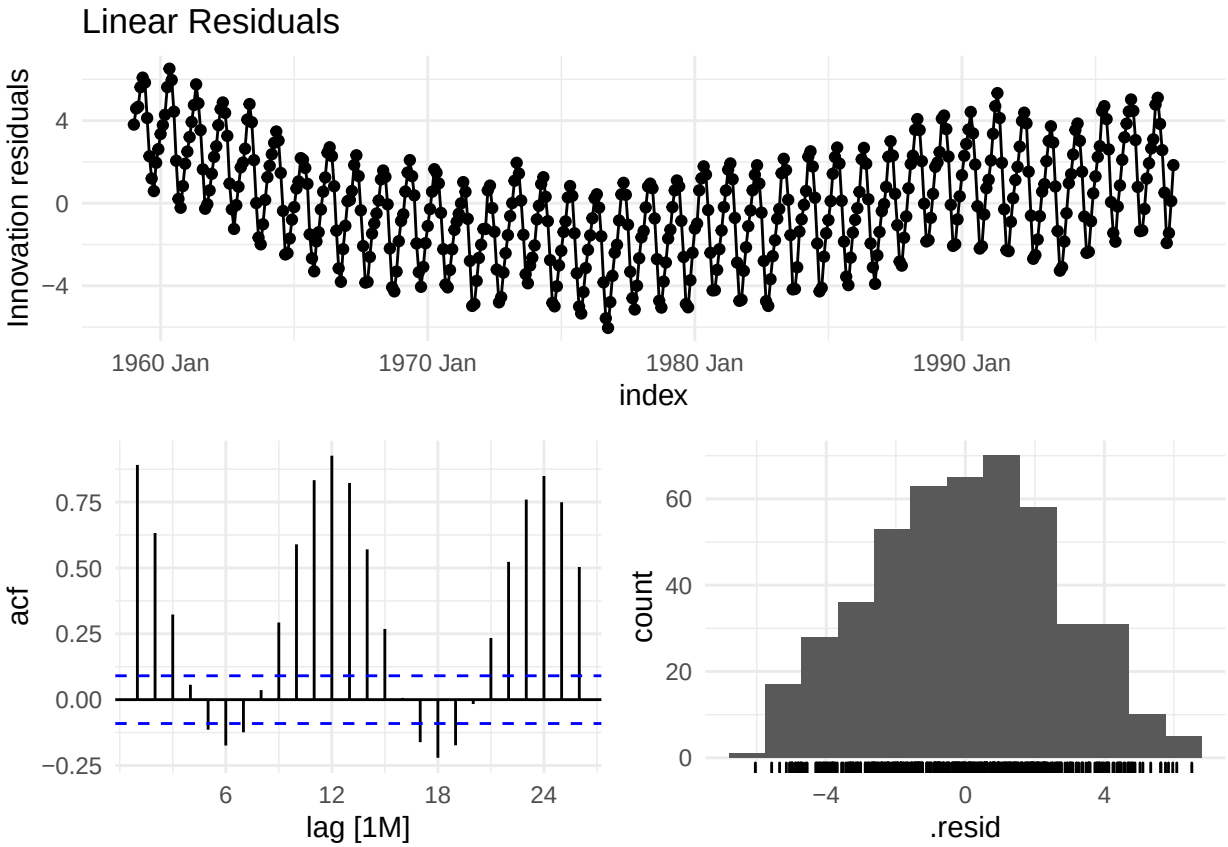
```
## Series: value
## Model: TSLM
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -6.039885 -1.947575 -0.001671  1.911271  6.514852
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.115e+02  2.424e-01  1284.9   <2e-16 ***
## trend()     1.090e-01  8.958e-04   121.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.618 on 466 degrees of freedom
## Multiple R-squared: 0.9695,  Adjusted R-squared: 0.9694
## F-statistic: 1.479e+04 on 1 and 466 DF, p-value: < 2.22e-16
```

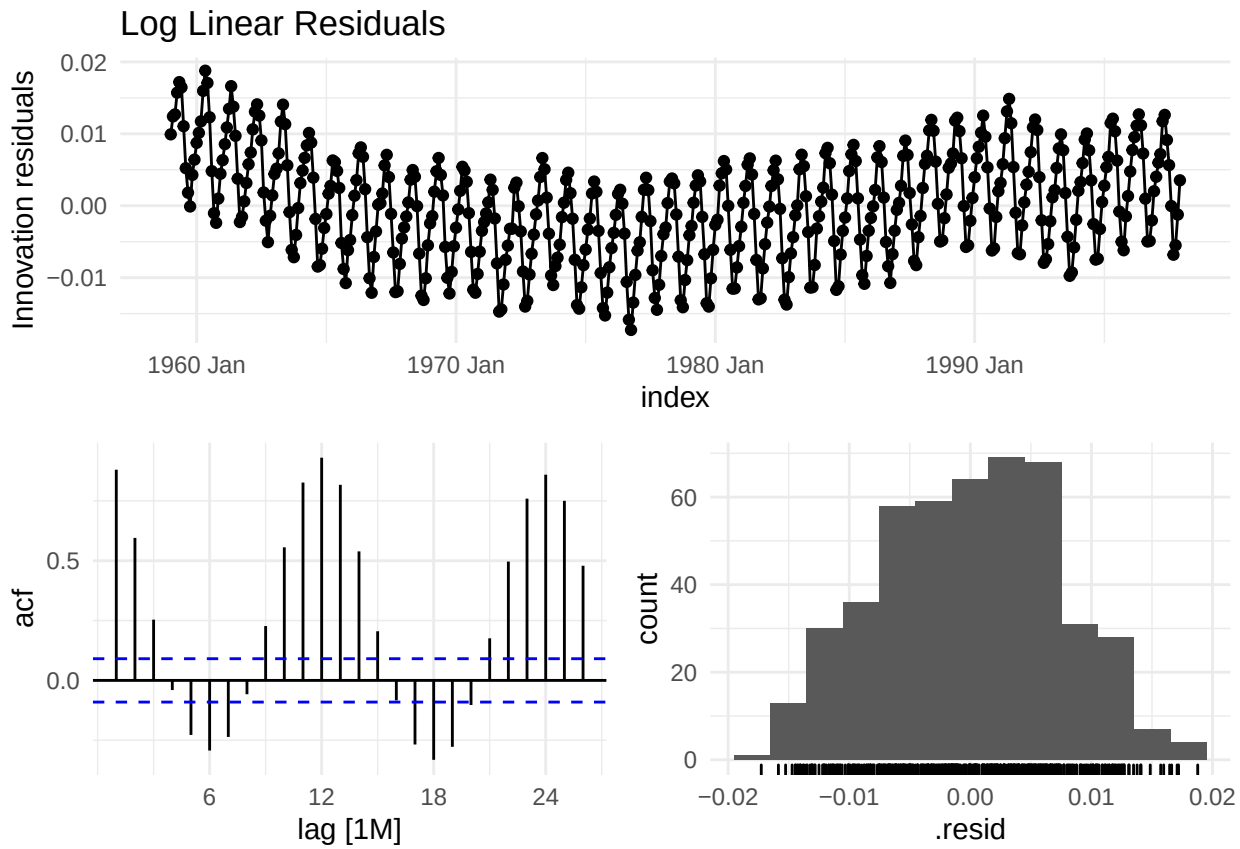
```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0172650 -0.0056145  0.0002764  0.0053760  0.0187770
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.744e+00  6.829e-04  8410.5   <2e-16 ***
## trend()     3.224e-04  2.523e-06   127.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007375 on 466 degrees of freedom
## Multiple R-squared: 0.9722,  Adjusted R-squared: 0.9722
## F-statistic: 1.633e+04 on 1 and 466 DF, p-value: < 2.22e-16
```

We can see from the output here that the R-squared values on both models are high and the input coefficients are highly significant on both. We will also plot the models on top of our data.



From these plots, both models seem to do a similar job fitting the data but fail to account for any of the seasonal "waves". Next, let's examine the model residuals.


Linear Residuals
Innovation residuals
4
0
−4
1960 Jan
1970 Jan
1980 Jan
1990 Jan
index
acf
0.75
0.50
0.25
0.00
−0.25
6
12
18
24
lag [1M]
count
60
40
20
0
−4
0
4
.resid



The residuals don't look much different here between the linear and log linear models, signaling that a log transformation is unnecessary for the linear model. In both models, the ACFs show a periodic oscillating pattern which signals that something is missing from our model. Based on our EDA, this is most likely the seasonality that we discussed but did not account for in this model. Additionally, the residuals appear somewhat normally distributed for both models. Let's run a Ljung Box test to understand whether we have white noise residuals, under the following hypotheses:

H0: Data are independently distributed.

H1: Data are not independently distributed.

```
##
##  Box-Ljung test
##
## data:  linear_residuals
## X-squared = 373.94, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  linear_residuals
## X-squared = 850.26, df = 10, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  linear_log_residuals
## X-squared = 365.1, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  linear_log_residuals
## X-squared = 830.65, df = 10, p-value < 2.2e-16
```

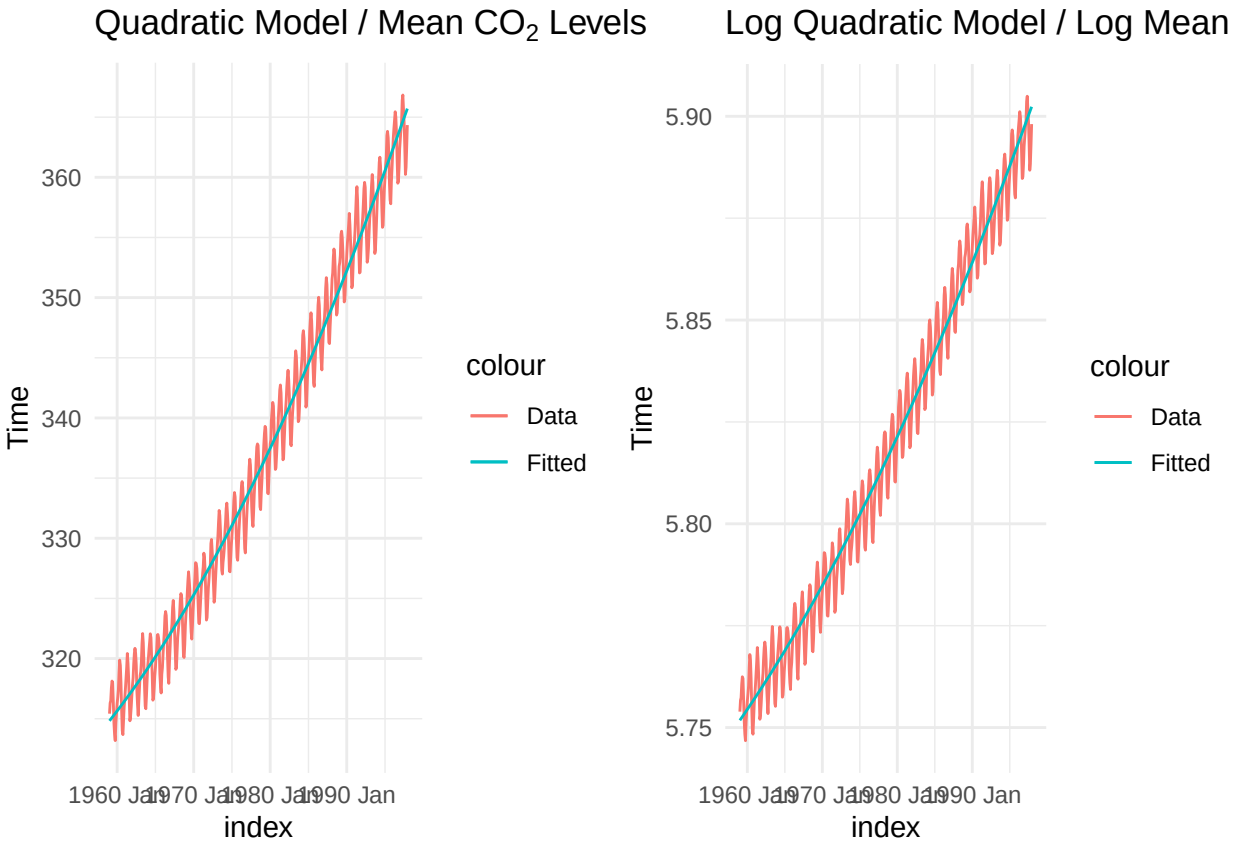
Based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data (likely the seasonality).

Let's repeat this process with a quadratic trend model to see if this model accounts for additional variance in our process.

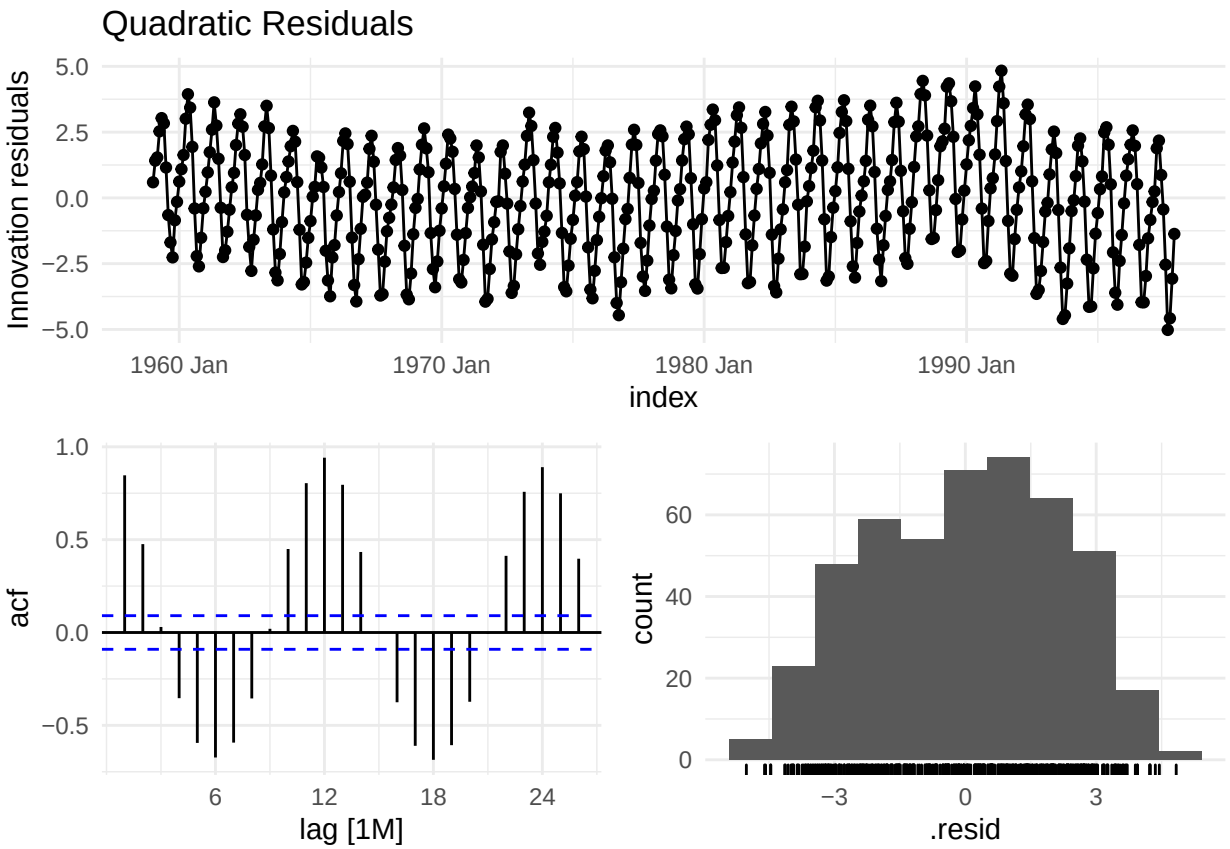
```
## Series: value
## Model: TSLM
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0195 -1.7120  0.2144  1.7957  4.8345
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.148e+02  3.039e-01 1035.65   <2e-16 ***
## trend()      6.739e-02  2.993e-03   22.52   <2e-16 ***
## I(trend()^2) 8.862e-05  6.179e-06   14.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.182 on 465 degrees of freedom
## Multiple R-squared: 0.9788,  Adjusted R-squared: 0.9787
## F-statistic: 1.075e+04 on 2 and 465 DF, p-value: < 2.22e-16
```

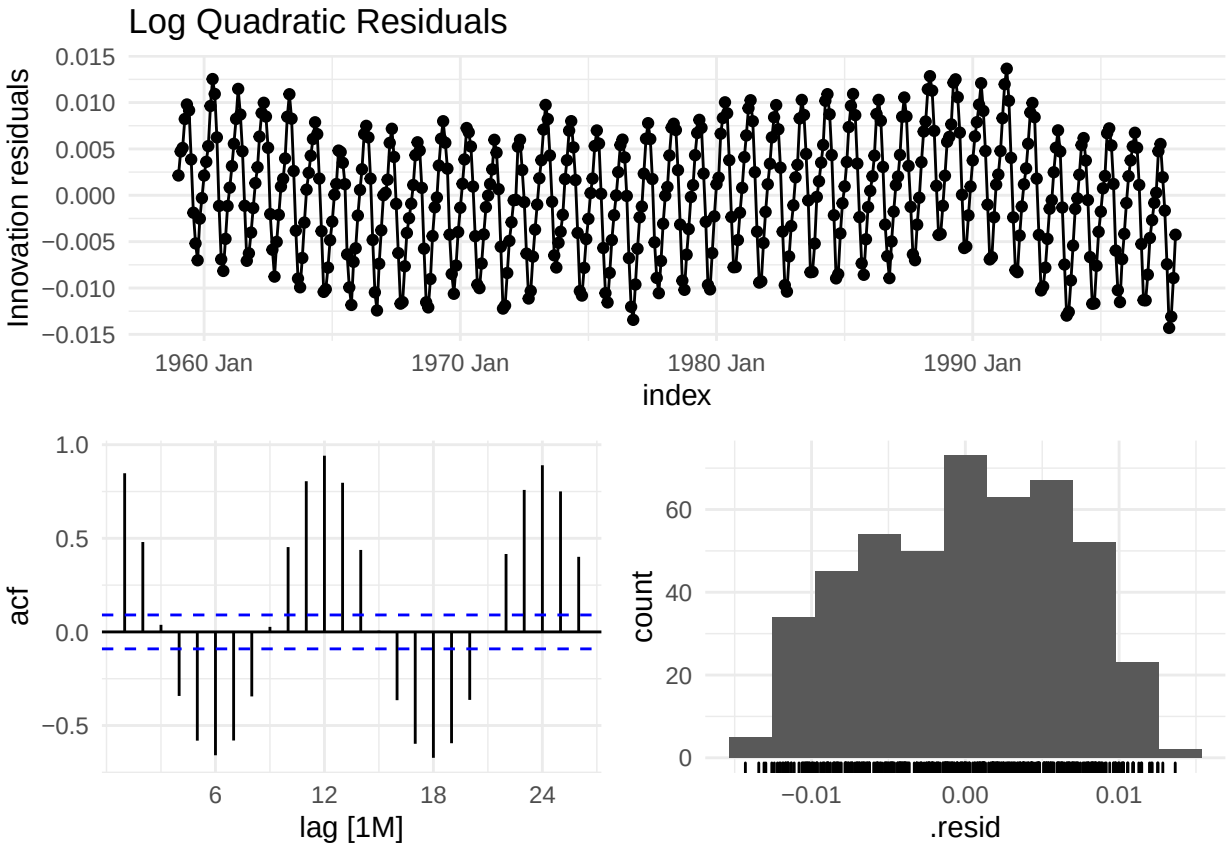
```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0143052 -0.0050832  0.0005277  0.0052757  0.0136508
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.752e+00  9.039e-04 6363.18   <2e-16 ***
## trend()      2.216e-04  8.900e-06   24.90   <2e-16 ***
## I(trend()^2) 2.149e-07  1.838e-08   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00649 on 465 degrees of freedom
## Multiple R-squared: 0.9786,  Adjusted R-squared: 0.9785
## F-statistic: 1.061e+04 on 2 and 465 DF, p-value: < 2.22e-16
```

We can see from the output here that the R-squared values on both models are high and the input coefficients are highly significant on both. We will also plot the models on top of our data.



From these plots, both models seem to do a similar job fitting the data. Next, let's examine the model residuals.


Quadratic Residuals
Innovation residuals
5.0
2.5
0.0
−2.5
−5.0
1960 Jan
1970 Jan
1980 Jan
1990 Jan
index
acf
1.0
0.5
0.0
−0.5
6
12
18
24
lag [1M]
count
60
40
20
0
−3
0
3
.resid



The residuals don't look much different here between both models, signaling that a log transformation is unnecessary for the quadratic model as well. In both models, the ACFs show the same periodic oscillating pattern which signals that the seasonality is likely missing from our model. The residuals also appear somewhat normally distributed for both models. Let's run a Ljung Box test again to understand whether we have white noise residuals under the same hypotheses as before.

```
##
##  Box-Ljung test
##
## data:  quadratic_residuals
## X-squared = 337.42, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_residuals
## X-squared = 1213.2, df = 10, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_log_residuals
## X-squared = 338.34, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_log_residuals
```

```
## X-squared = 1186.5, df = 10, p-value < 2.2e-16
```

As before, based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data (likely again the seasonality).

Since the log transforms don't appear to reduce any of the variance in our models, I am going to compare the linear model directly with the quadratic model on non-transformed data. Our previous analysis shows that both models are behaving similarly but the quadratic model has smaller residuals compared to the linear model, making it our more favorable model. In addition to the residual analysis, I going to compare the two using the Bayesian Information Criteria (BIC) as my metric.

```
## # A tibble: 1 x 1
##     BIC
##   <dbl>
## 1  917.
```

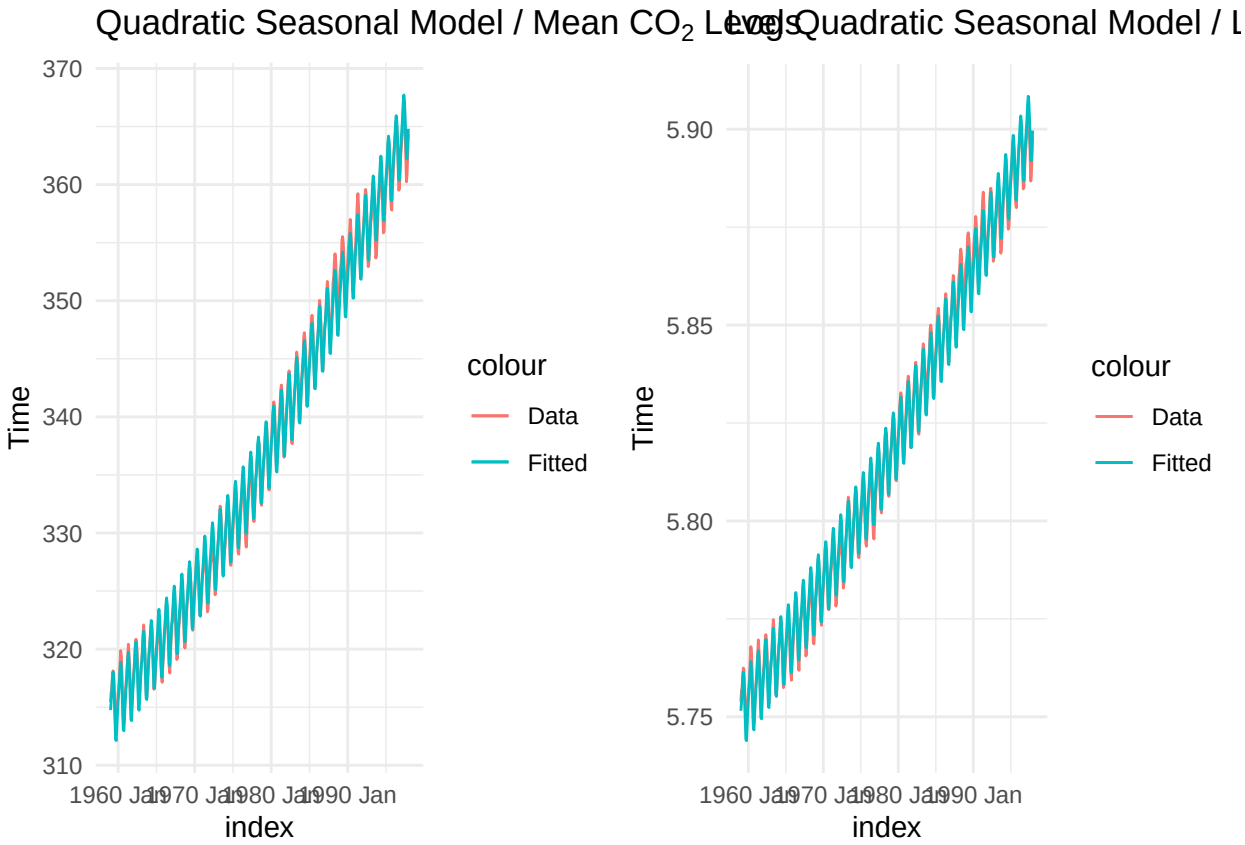
```
## # A tibble: 1 x 1
##     BIC
##   <dbl>
## 1  752.
```

The quadratic model has a lower BIC, so I will move forward with our modeling process using this quadratic model. Our next step is to fit a polynomial model that incorporates seasonal dummy variables, which will hopefully account for some of the variance that we failed to capture in our previous models. We'll create models using both the log transform and the normal CO2 values, although it has appeared from our past models that the log transform does not seem to account for much, if any, variance in our data.

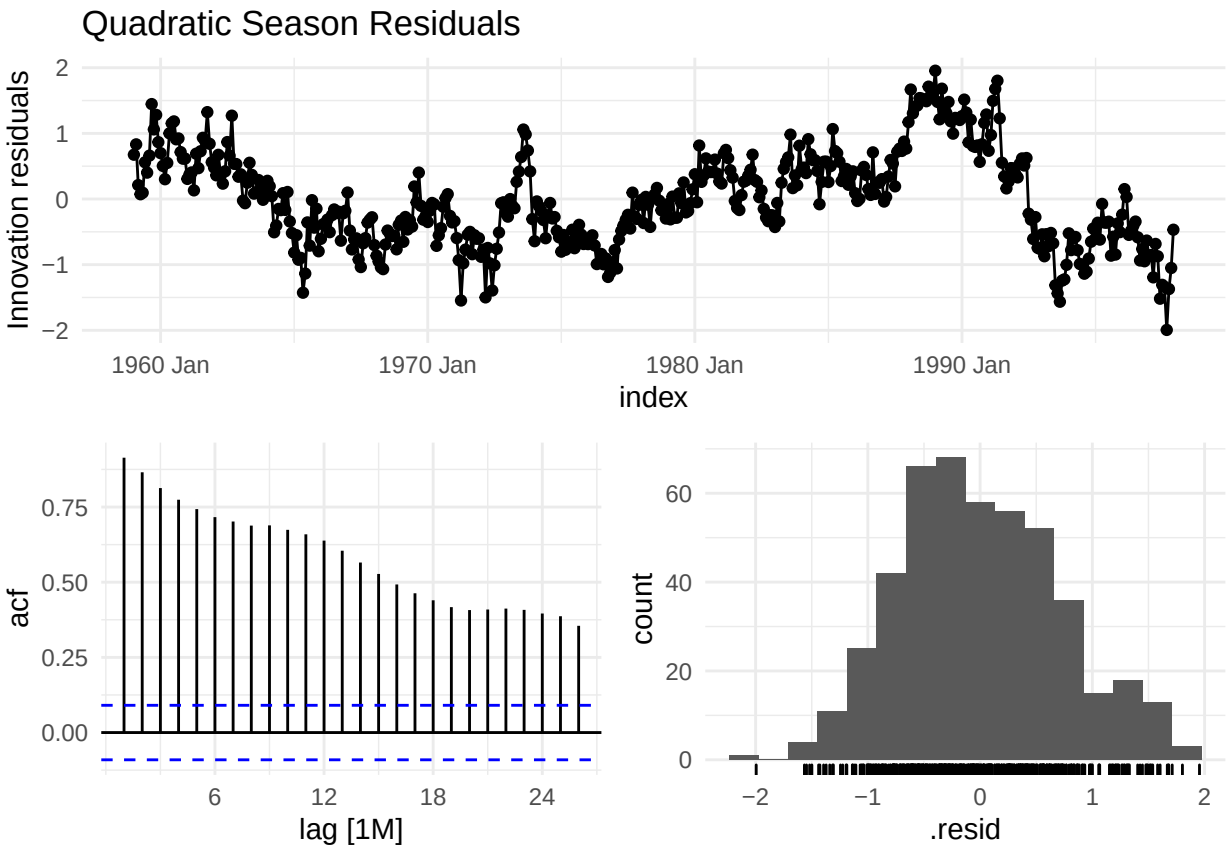
```
## Series: value
## Model: TSLM
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99478 -0.54468 -0.06017  0.47265  1.95480
##
## Coefficients:
##                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)     3.147e+02  1.494e-01 2105.894  < 2e-16 ***
## trend()         6.763e-02  9.929e-04   68.114  < 2e-16 ***
## I(trend()^2)    8.865e-05  2.050e-06   43.242  < 2e-16 ***
## season()year2   6.642e-01  1.640e-01    4.051 5.99e-05 ***
## season()year3   1.407e+00  1.640e-01    8.582  < 2e-16 ***
## season()year4   2.538e+00  1.640e-01   15.480  < 2e-16 ***
## season()year5   3.017e+00  1.640e-01   18.400  < 2e-16 ***
## season()year6   2.354e+00  1.640e-01   14.357  < 2e-16 ***
## season()year7   8.331e-01  1.640e-01    5.081 5.50e-07 ***
## season()year8  -1.235e+00  1.640e-01   -7.531 2.75e-13 ***
## season()year9  -3.059e+00  1.640e-01  -18.659  < 2e-16 ***
## season()year10 -3.243e+00  1.640e-01  -19.777  < 2e-16 ***
## season()year11 -2.054e+00  1.640e-01  -12.526  < 2e-16 ***
## season()year12 -9.374e-01  1.640e-01   -5.717 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.724 on 454 degrees of freedom
## Multiple R-squared: 0.9977,  Adjusted R-squared: 0.9977
## F-statistic: 1.531e+04 on 13 and 454 DF, p-value: < 2.22e-16
```

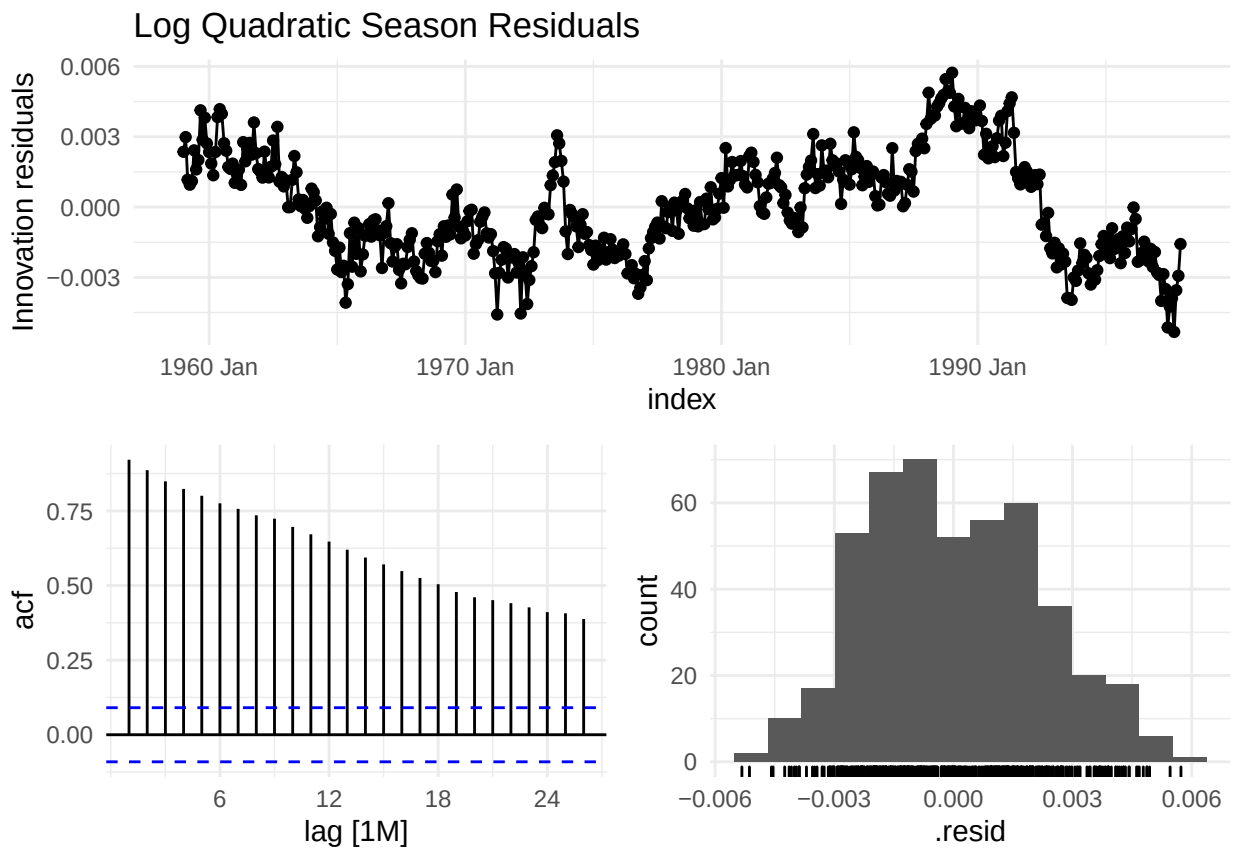
```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0053270 -0.0017362 -0.0001774  0.0015139  0.0057292
##
## Coefficients:
##                  Estimate Std. Error   t value Pr(>|t|)
## (Intercept)     5.751e+00  4.533e-04 12687.967  < 2e-16 ***
## trend()         2.223e-04  3.012e-06    73.817  < 2e-16 ***
## I(trend()^2)    2.150e-07  6.219e-09    34.566  < 2e-16 ***
## season()year2   1.969e-03  4.974e-04     3.959 8.73e-05 ***
## season()year3   4.163e-03  4.974e-04     8.371 7.16e-16 ***
## season()year4   7.498e-03  4.974e-04    15.075  < 2e-16 ***
## season()year5   8.911e-03  4.974e-04    17.916  < 2e-16 ***
## season()year6   6.965e-03  4.974e-04    14.004  < 2e-16 ***
## season()year7   2.480e-03  4.974e-04     4.986 8.78e-07 ***
## season()year8  -3.662e-03  4.974e-04    -7.362 8.61e-13 ***
## season()year9  -9.098e-03  4.974e-04   -18.290  < 2e-16 ***
## season()year10 -9.661e-03  4.974e-04   -19.423  < 2e-16 ***
## season()year11 -6.113e-03  4.974e-04   -12.290  < 2e-16 ***
## season()year12 -2.799e-03  4.974e-04    -5.627 3.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002196 on 454 degrees of freedom
## Multiple R-squared: 0.9976,  Adjusted R-squared: 0.9975
## F-statistic: 1.453e+04 on 13 and 454 DF, p-value: < 2.22e-16
```

Just from this summary, we can see that the R-squared values are slightly higher on both of these seasonal models than either the linear or quadratic models. We can also see that all the input coefficients are highly significant. Let's plot our models next.



From these plots, both models seem to do a similar job fitting the data and, unlike the previous models we've seen, actually account for the seasonal "waves" in our plot and track them pretty closely. Next, let's examine the model residuals.


Quadratic Season Residuals
Innovation residuals
2
1
0
−1
−2
1960 Jan
1970 Jan
1980 Jan
1990 Jan
index
acf
0.75
0.50
0.25
0.00
6
12
18
24
lag [1M]
count
60
40
20
0
−2
−1
0
1
2
.resid



The residuals again don't look much different here between both models, signaling that a log transformation is unnecessary for this seasonal quadratic model as well. In both models, the ACFs now shows highly significant and slowly declining lags, unlike the ACF plots on the previous models which showed an oscillating trend that was likely related to seasonality variance that our latest models are now capturing. The residuals appear somewhat normally distributed for both models. In comparison to both the non-seasonal linear and quadratic models, the residuals are small in magnitude. Let's run a Ljung Box test again to understand whether we have white noise residuals under the same hypotheses as before:

```
##
##  Box-Ljung test
##
## data:  quadratic_season_residuals
## X-squared = 393.48, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_season_residuals
## X-squared = 2758, df = 10, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  quadratic_log_season_residuals
## X-squared = 399.97, df = 1, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
```

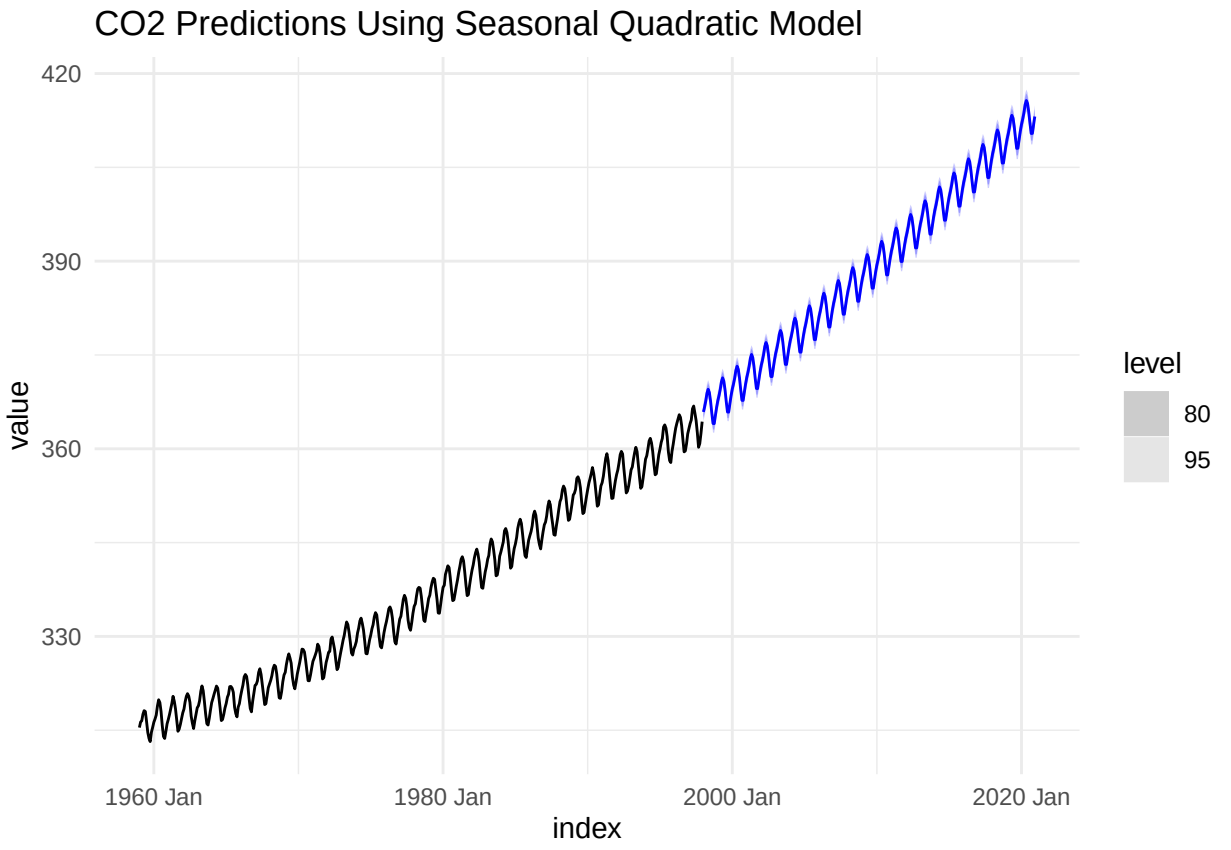
```
##
## data:  quadratic_log_season_residuals
## X-squared = 3040.2, df = 10, p-value < 2.2e-16
```

As before, based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data. This is somewhat surprising as our residuals are quite small, especially in comparison to our previous models. Let's compare the BIC of the quadratic seasonal model without the log transform to our previous BICs.

Let's compare the BIC of this model against the pure linear and quadratic models without seasonal components.

```
## # A tibble: 1 x 1
##     BIC
##   <dbl>
## 1 -224.
```

The BIC for our quadratic seasonal model is much smaller than the BICs for either our linear or quadratic model, and by this criteria is our best model that we have fit so far. Let's use this model to generate forecasts to the year 2020.



We can see that these predictions continue to follow the trend and seasonal fluctuations that we noticed in our data. Let's complete our linear modeling by examining the CLM assumptions for our best model, the seasonal quadratic model.

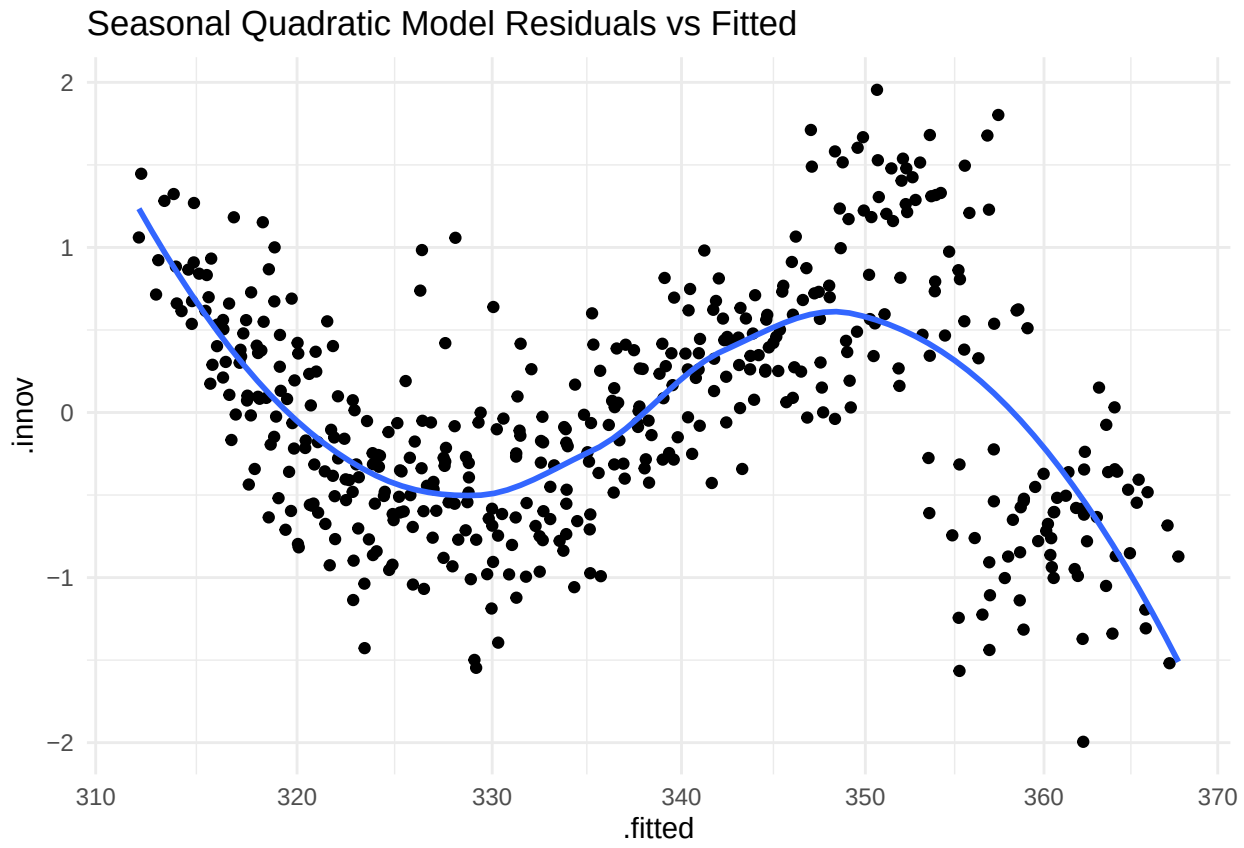
The first CLM assumption is that the underlying data generating process follows a linear model. We cannot check this assumption and will just assume it is true, and, if untrue, assume that we are just fitting the best linear model to the data.

The second assumption is ergodatic stationarity. Since this model is including trend and seasonal variables,

we are eliminating the deterministic parts in the process and are satisfying this assumption. From the ACF plot, the process does look stationary but the residuals do appear serially correlated.

The third assumption is no perfect multicollinearity. When we fit our model, R did not report any warning or missing coefficients, which means that this assumption is violated.

The fourth assumption is the zero conditional mean assumption. We can evaluate this by looking at these residuals versus the fitted value plot, plotted below. If this assumption is satisfied, there should be no relationship between the fitted values and the residuals, a.k.a. there should be a straight line in the plot.



The line is curved, so we cannot justify that the zero conditional mean assumption is satisfied, meaning that there are likely additional variables that should be included in our model.

The fifth CLM assumption is homoskedasticity. The residuals are quite spread out with a large range of variance, so it seems as though this assumption is not satisfied.

The last assumption is no serial correlation. From our ACF plot, the residuals do appear to be serially correlated. Additionally, the Ljung Box test returns a high statistic with a small p-value. We can reject the null hypothesis that there is no autocorrelation and conclude that our residuals here are autocorrelated, thereby not satisfying this assumption.

The residuals do look close to a normal distribution, which gives us more confidence in our generated prediction intervals.

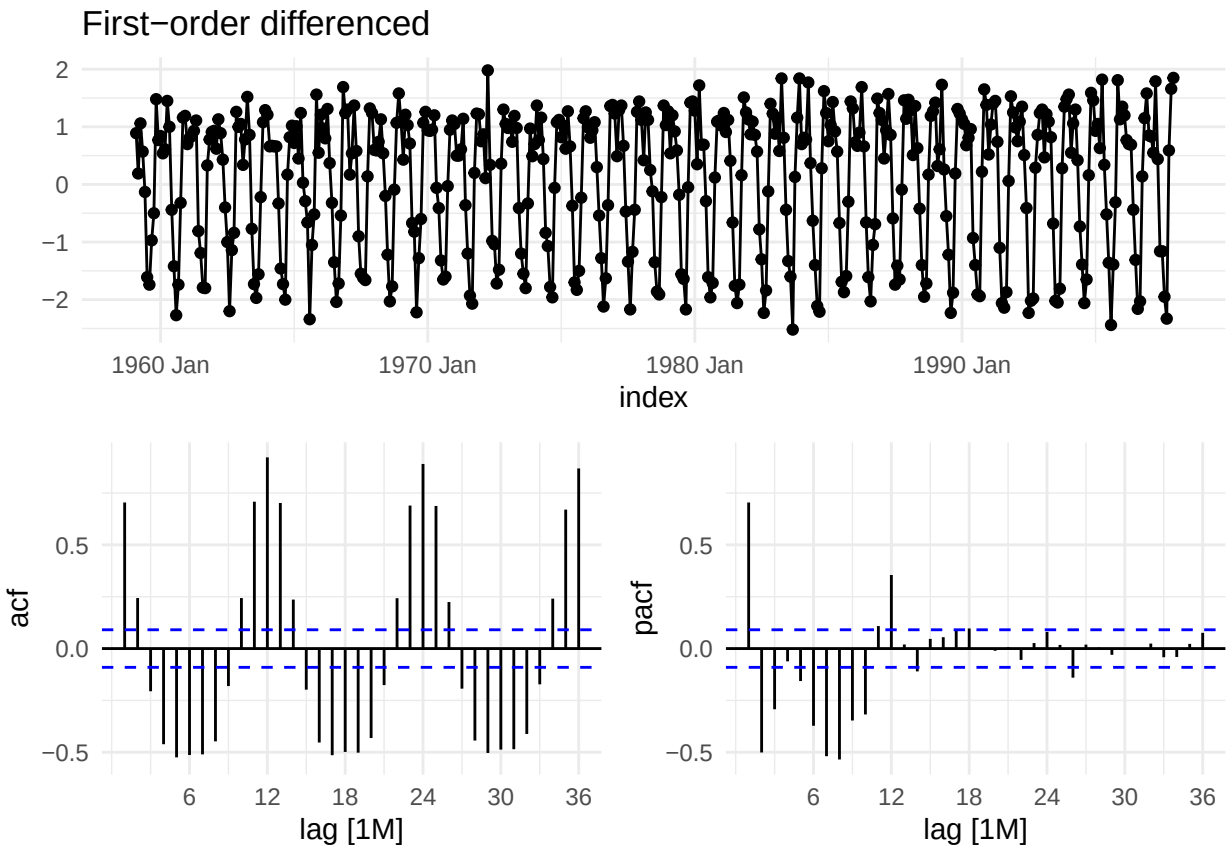
ARIMA Modeling

Now, let's choose an ARIMA model to fit to the CO2 data. Recall that our EDA indicated that the CO2 data had the following characteristics:

- non-stationary data series

- slowly decaying ACF plot, indicative of an AR process
- oscillating PCAF plot, indicative of an MA process
- seasonal data

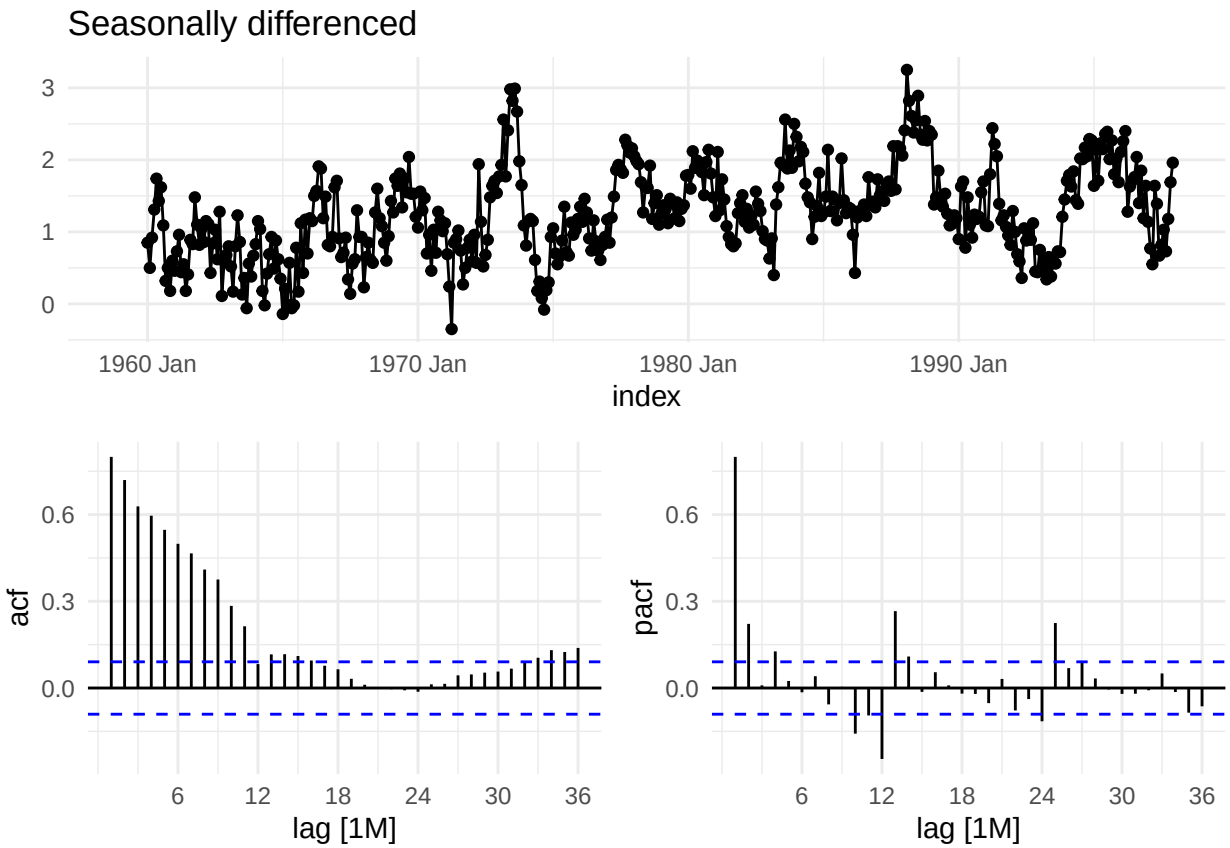
In order to transform the data into a stationary series, we can start by applying a first-order difference to the series and plotting the characteristics.



Based on the trend plot, the data seems to stabilize around a mean of -0.5, and it appears to have a constant variance, suggesting that the data only needs a first-order difference. The ACF plot continues to decay gradually but the oscillating behavior persists here. This indicates that the seasonal pattern is strong and stable after a first-order difference. As a result, we will want to use a seasonal difference in addition to this first difference. Since the seasonal behavior repeats every year, we will set m = 12, meaning the seasonal pattern repeats once every twelve months. The output of this seasonal model is below.

```
## # A tsibble: 468 x 4 [1M]
##       index value log_value sdiff_co2
##       <mth> <dbl>     <dbl>     <dbl>
##  1 1959 Jan  315.      5.75        NA
##  2 1959 Feb  316.      5.76        NA
##  3 1959 Mar  316.      5.76        NA
##  4 1959 Apr  318.      5.76        NA
##  5 1959 May  318.      5.76        NA
##  6 1959 Jun  318       5.76        NA
##  7 1959 Jul  316.      5.76        NA
##  8 1959 Aug  315.      5.75        NA
##  9 1959 Sep  314.      5.75        NA
## 10 1959 Oct  313.      5.75        NA
```

```
## # ... with 458 more rows
```



On its own, seasonal differencing does not appear to make the data stationary. However, based on the ACF plot, we can confirm that the seasonal differencing either removes or significantly reduces the oscillation as it does not seem apparent anymore. As a result, we can conclude that we need both a first difference and a seasonal difference on the series. We can confirm these observations using the following unit root tests. We will first use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test under the following hypothesis:

H0: The data are stationary

HA: The data are not stationary

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      7.82        0.01
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1    0.0124         0.1
```

```
## # A tibble: 1 x 2
##   kpss_stat kpss_pvalue
##       <dbl>       <dbl>
## 1      1.94        0.01
```

We can interpret the tests in the following manner:

- For the undifferenced data, the test statistic (7.81) is bigger than the 1% critical value, so the p-value is

less than 0.01, indicating that the null hypothesis is rejected. That is, the data are not stationary.

- For the first differenced data, the p-value is greater than 0.1, indicating the test fails to reject the null hypothesis. That is, the data are stationary.
- For the seasonally differenced data, the test statistic (1.94) is bigger than the 1% critical value, so the p-value is less than 0.01, indicating that the null hypothesis is rejected. That is, the data are not stationary.

Thus, we can conclude that we need both a first difference and a seasonal difference. But, how many differences do we need? To confirm that we are applying the appropriate amount of differences, we can use the following:

```
## # A tibble: 1 x 1
##   ndiffs
##    <int>
## 1      1
```

```
## # A tibble: 1 x 1
##   ndiffs
##    <int>
## 1      0
```

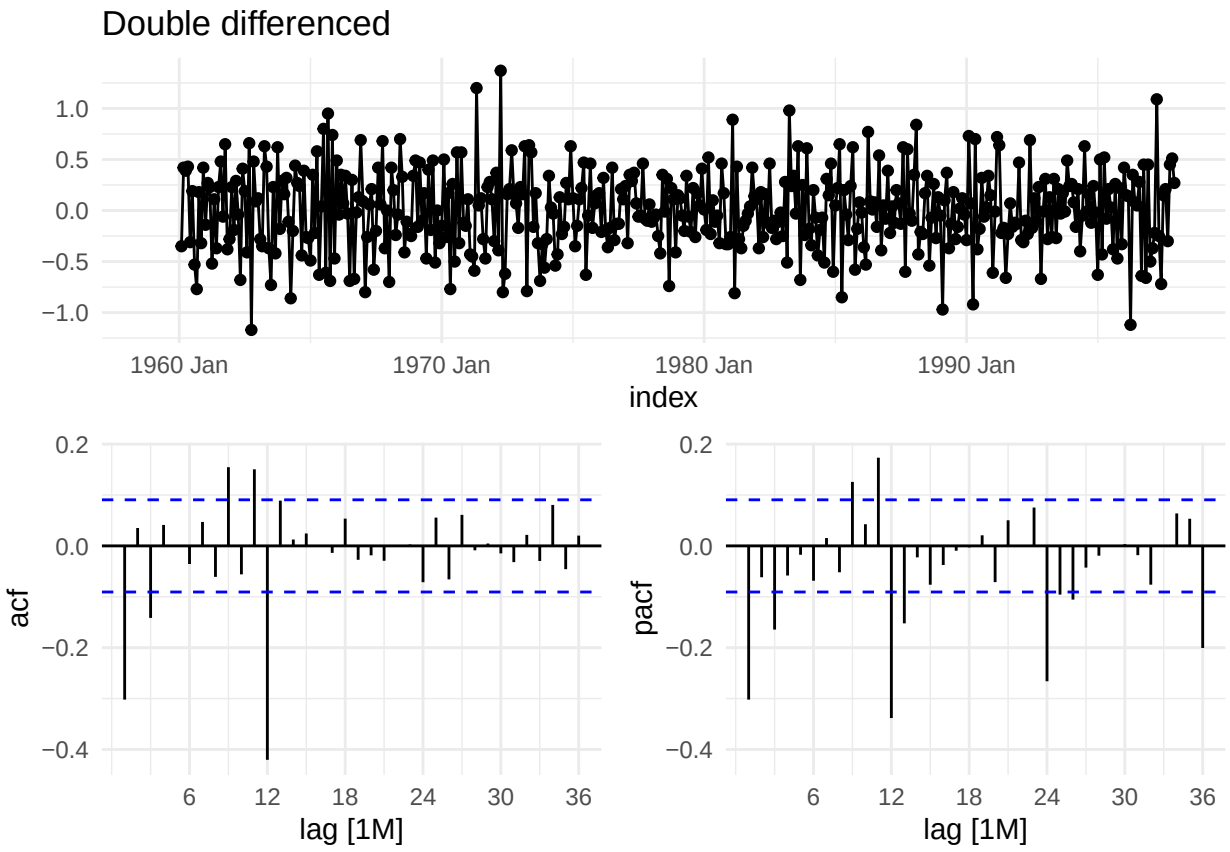
```
## # A tibble: 1 x 1
##   ndiffs
##    <int>
## 1      1
```

```
## # A tibble: 1 x 1
##   nsdiffs
##     <int>
## 1       1
```

```
## # A tibble: 1 x 1
##   nsdiffs
##     <int>
## 1       1
```

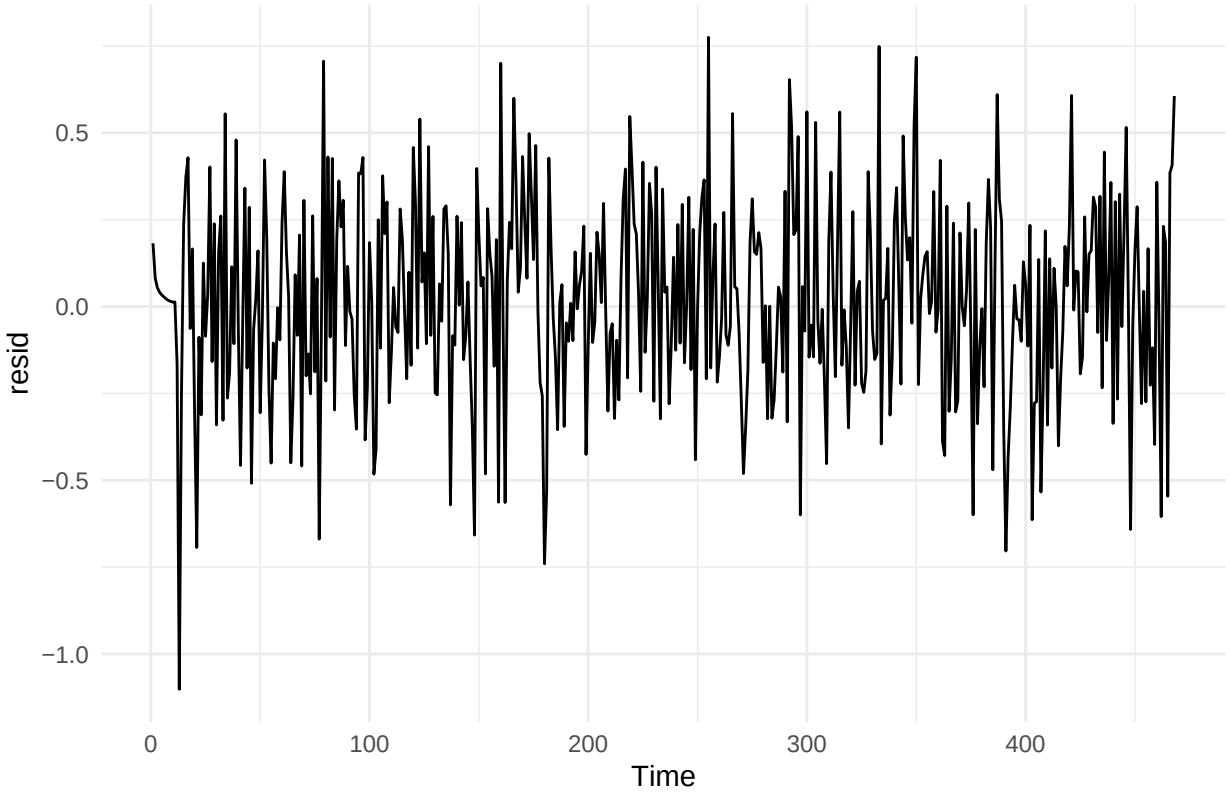
```
## # A tibble: 1 x 1
##   nsdiffs
##     <int>
## 1       0
```

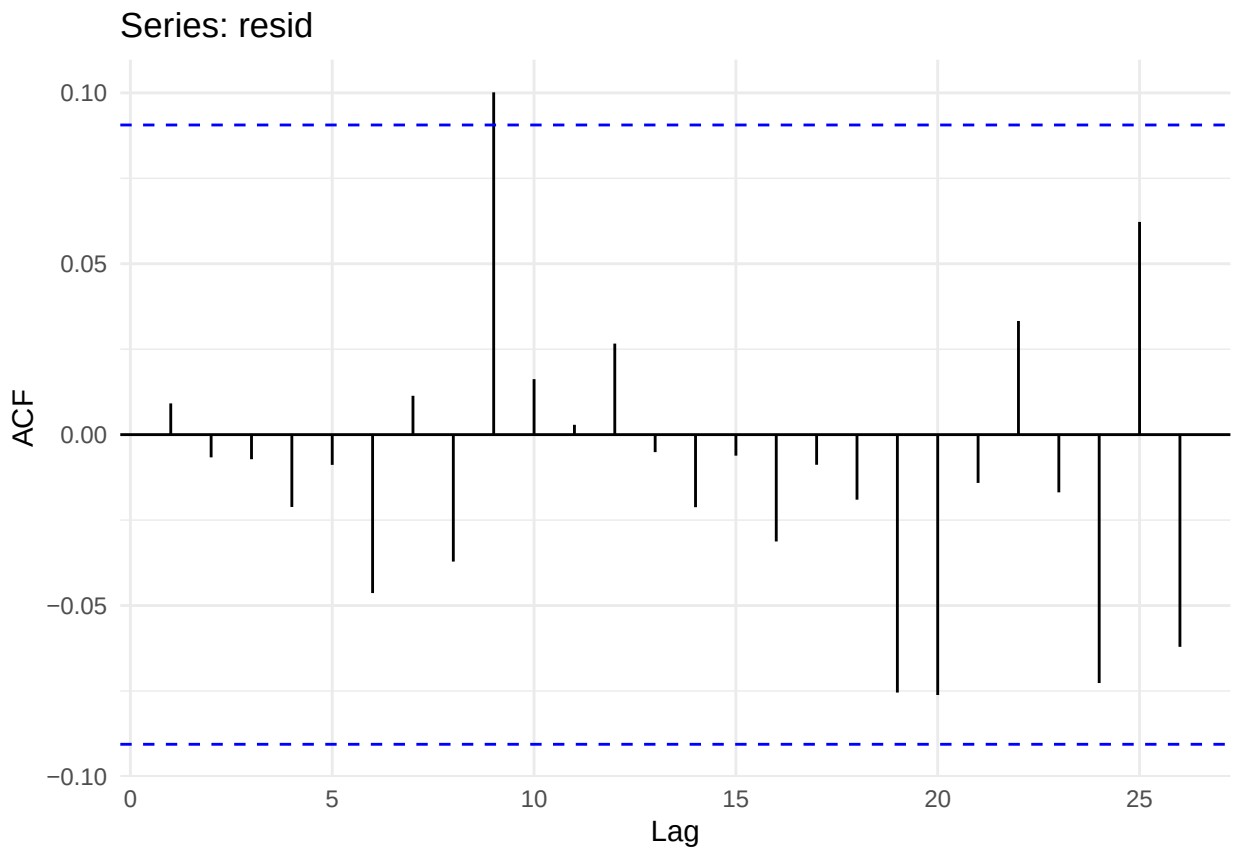
Here we can see that once we apply a first difference and a seasonal difference, the test returns 0, indicating no additional differences are required. It is important that we correctly identify the correct number of differences as we could potentially introduce false dynamics or autocorrelations into the time series.



We will now use the ACF and PACF plots of our double differenced data to determine an appropriate ARIMA model via brute force. The significant spike at lag 3 in the ACF suggests a non-seasonal MA(3) component. The significant spike at lag 12 in the ACF suggests a seasonal MA(1) component. So, we might begin with an ARIMA(0,1,3)(0,1,1)12 model, indicating a first difference, a seasonal difference, and non-seasonal MA(3) and seasonal MA(1) component. The output of fitting this model appears below.

```
##
## Call:
## arima(x = co2_df$value, order = c(0, 1, 3), seasonal = list(order = c(0, 1,
##     1), period = 12))
##
## Coefficients:
##           ma1      ma2      ma3     sma1
##       -0.3394  -0.0180  -0.0973  -0.8538
## s.e.   0.0475   0.0497   0.0467   0.0256
##
## sigma^2 estimated as 0.0816:  log likelihood = -83.43,  aic = 176.86
```


0.5
0.0
−0.5
−1.0
resid
0
100
200
300
400
Time




```
##
##  Box-Ljung test
##
## data:  resid
## X-squared = 7.0116, df = 10, p-value = 0.7243
```

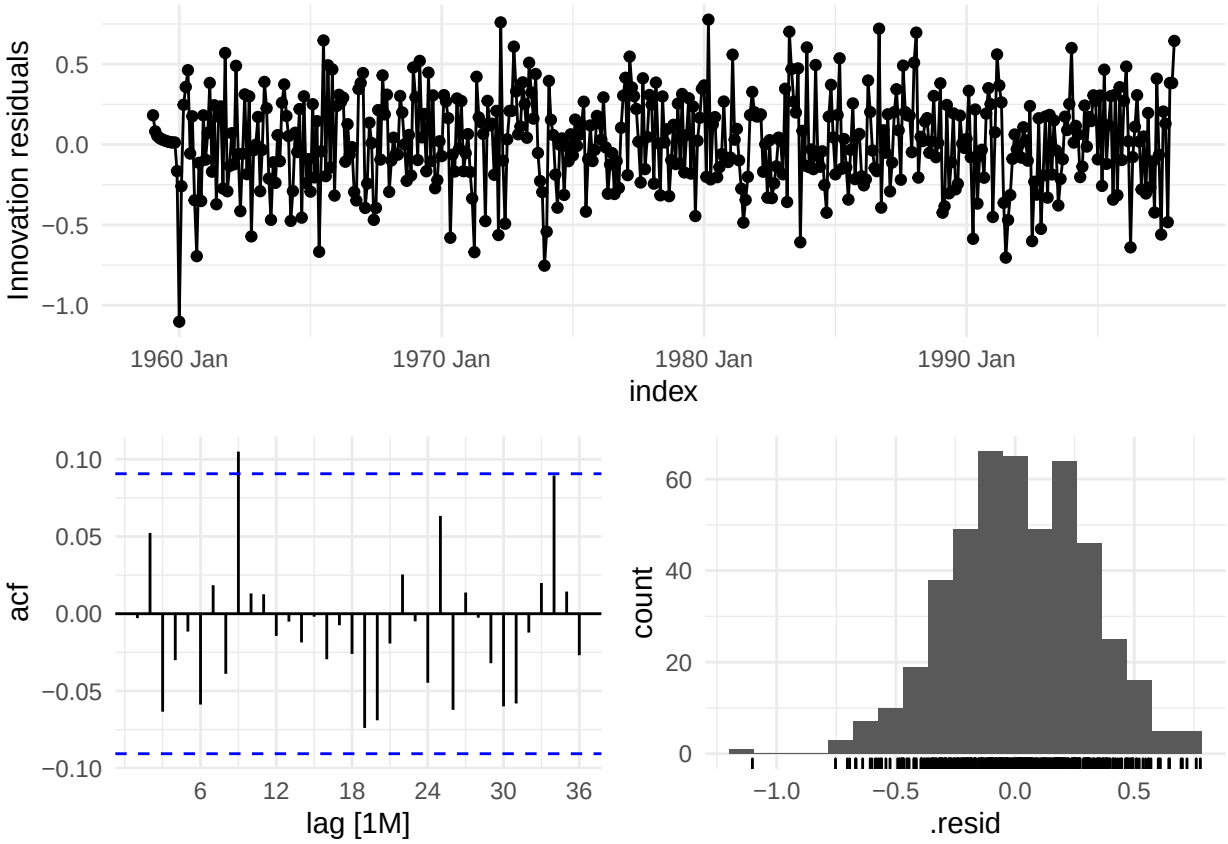
The ACF plot indicates that the residuals are following a white noise process as each autocorrelation is close to zero and 95% of spikes lie within significance thresholds. There is one small but significant spike at lag 8. Additionally, since the p-value from the Box-Ljung test is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no serial correlation in the data.

If we had evaluated the PACF as well, we may have chosen ARIMA(3,1,0)(1,1,1)12 model, indicating a first difference, a seasonal difference, non-seasonal AR(3) and MA(1) components, and seasonal AR(1) and AR(1) component. The significant spike at lag 3 in the PACF suggests a non-seasonal AR(3) component. The significant spike at lag 12 in the PACF suggests a seasonal AR(1) component. However, by comparing the AIC values of the first and second model, the first selection was a better fit based on the lower AIC value.

```
##
## Call:
## arima(x = co2_df$value, order = c(3, 1, 1), seasonal = list(order = c(1, 1,
##     1), period = 12))
##
## Coefficients:
##          ar1     ar2      ar3      ma1    sar1     sma1
##       0.2972  0.0842  -0.0635  -0.6335  0.0394  -0.8637
## s.e.  0.1822  0.0785   0.0578   0.1765  0.0545   0.0276
##
## sigma^2 estimated as 0.0815:  log likelihood = -83.15,  aic = 180.29
```

We can also try automatically fitting the model:

```
## Series: value
## Model: ARIMA(1,1,1)(1,1,2)[12]
##
## Coefficients:
##          ar1      ma1     sar1     sma1     sma2
##       0.2569  -0.5847  -0.5489  -0.2620  -0.5123
## s.e.  0.1406   0.1204   0.5881   0.5703   0.4820
##
## sigma^2 estimated as 0.08576:  log likelihood=-84.39
## AIC=180.78   AICc=180.97   BIC=205.5
```

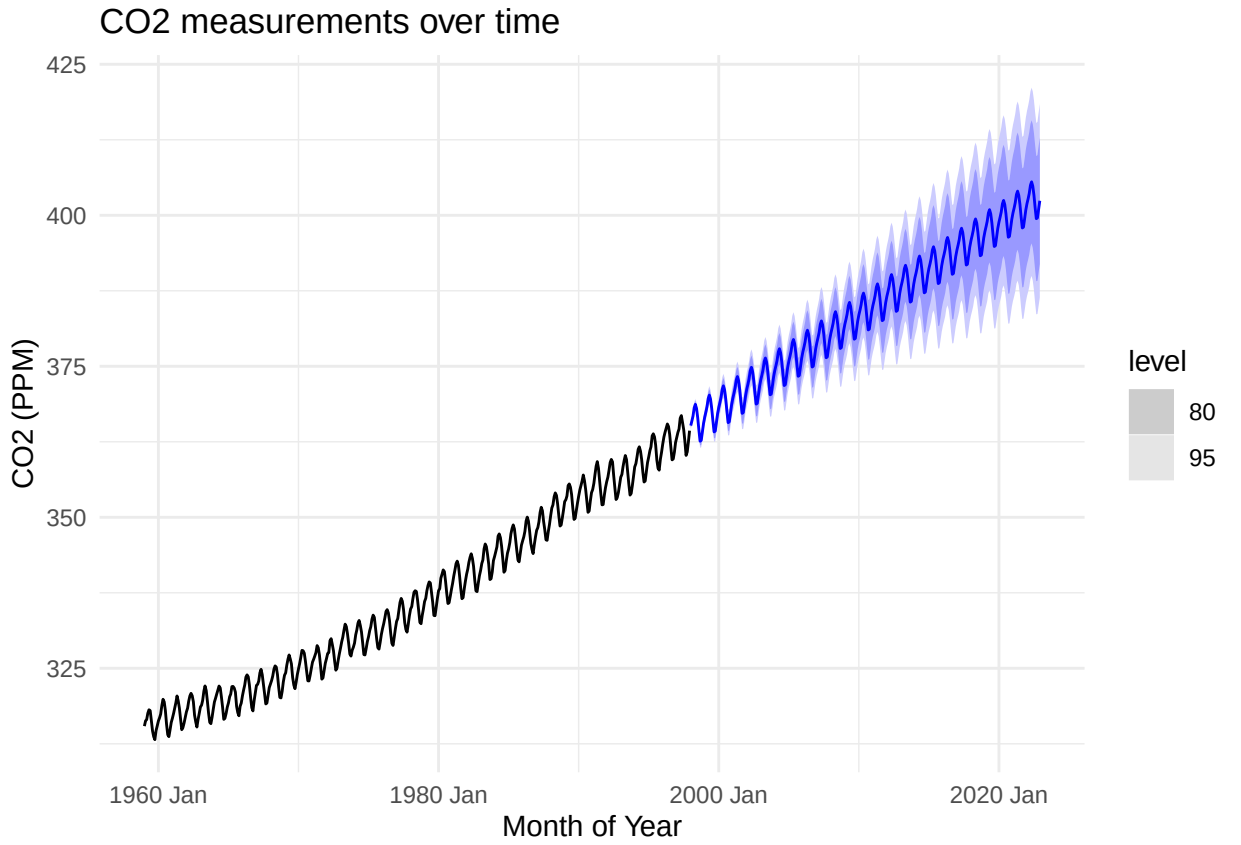



```
## # A tsibble: 468 x 3 [1M]
## # Key:       .model [1]
##    .model          index .resid
##    <chr>           <mth>  <dbl>
##  1 ARIMA(value) 1959 Jan 0.182
##  2 ARIMA(value) 1959 Feb 0.0821
##  3 ARIMA(value) 1959 Mar 0.0539
##  4 ARIMA(value) 1959 Apr 0.0411
##  5 ARIMA(value) 1959 May 0.0333
##  6 ARIMA(value) 1959 Jun 0.0276
##  7 ARIMA(value) 1959 Jul 0.0221
##  8 ARIMA(value) 1959 Aug 0.0177
##  9 ARIMA(value) 1959 Sep 0.0148
## 10 ARIMA(value) 1959 Oct 0.0129
```

```
## # ... with 458 more rows
```

```
##
##  Box-Ljung test
##
## data:  auto_resid[3]
## X-squared = 11.577, df = 10, p-value = 0.3143
```

The trend and ACF plots indicate that the residuals of the automatically fitted ARIMA model are following a white noise process. Autocorrelations are close to zero and 95% of spikes lie within significance thresholds. There is one small but significant spike at lag 9. Also, since the p-value from the Box-Ljung test is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no serial correlation in the data. However, the automatically fitted ARIMA model does not outperform our first guess as the AIC in the autofit model is higher. We will continue to proceed with the ARIMA(0,1,3)(0,1,1)12.



Again, we can see that these predictions continue to follow the trend and seasonal fluctuations that we noticed earlier in the data. However, our confidence intervals continue to widen with each year that we forecast.

Forecast atmospheric CO2 growth

If we generate predictions for when atmospheric CO2 is expected to be at 420 ppm and 500 ppm levels at the 95% confidence level, we can observe the wide confidence intervals noted in the previous graph. Predictions for 420ppm were as early as April 2022 and as late as December 2100. Surprisingly, when we make predictions on for 500ppm the confidence intervals narrow. According to this model's predictions, 500ppm CO2 levels will occur as early as March 2055 and as late as August 2086, as seen below.

```
## [1] "Apr 2022"
```

```
## [1] "Dec 2100"
```

```
## [1] 78.66667
```

```
## [1] "Mar 2055"
```

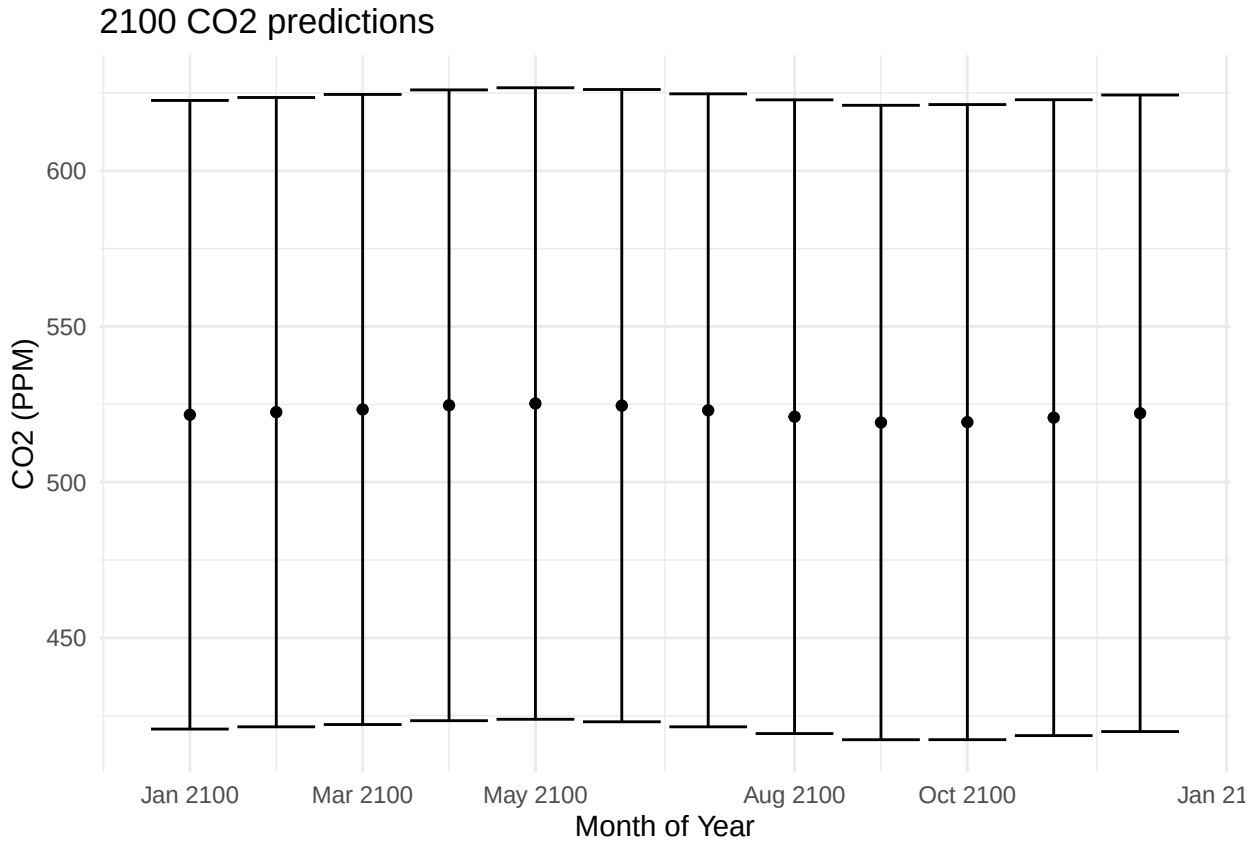
```
## [1] "Aug 2086"
```

```
## [1] 31.41667
```

When we look specifically at CO2 levels in 2100, we can expect the ppm values to be around 525 throughout the course of the year with fairly wide confidence intervals about 100ppms lower or higher.

```
##     time_index point_forecast          lo_95        high_95
##      "yearmon"      "numeric"      "numeric"      "numeric"
```

```
##    time_index point_forecast    lo_95  high_95
## 1    Jan 2100       521.6461 420.7014 622.5907
## 2    Feb 2100       522.4843 421.4223 623.5463
## 3    Mar 2100       523.3405 422.1614 624.5197
## 4    Apr 2100       524.6894 423.3936 625.9852
## 5    May 2100       525.2589 423.8467 626.6712
## 6    Jun 2100       524.5735 423.0449 626.1021
## 7    Jul 2100       523.0702 421.4254 624.7151
## 8    Aug 2100       521.0166 419.2556 622.7775
## 9    Sep 2100       519.1595 417.2825 621.0364
## 10   Oct 2100       519.2845 417.2918 621.2773
## 11   Nov 2100       520.7157 418.6073 622.8241
## 12   Dec 2100       522.1267 419.9027 624.3507
```



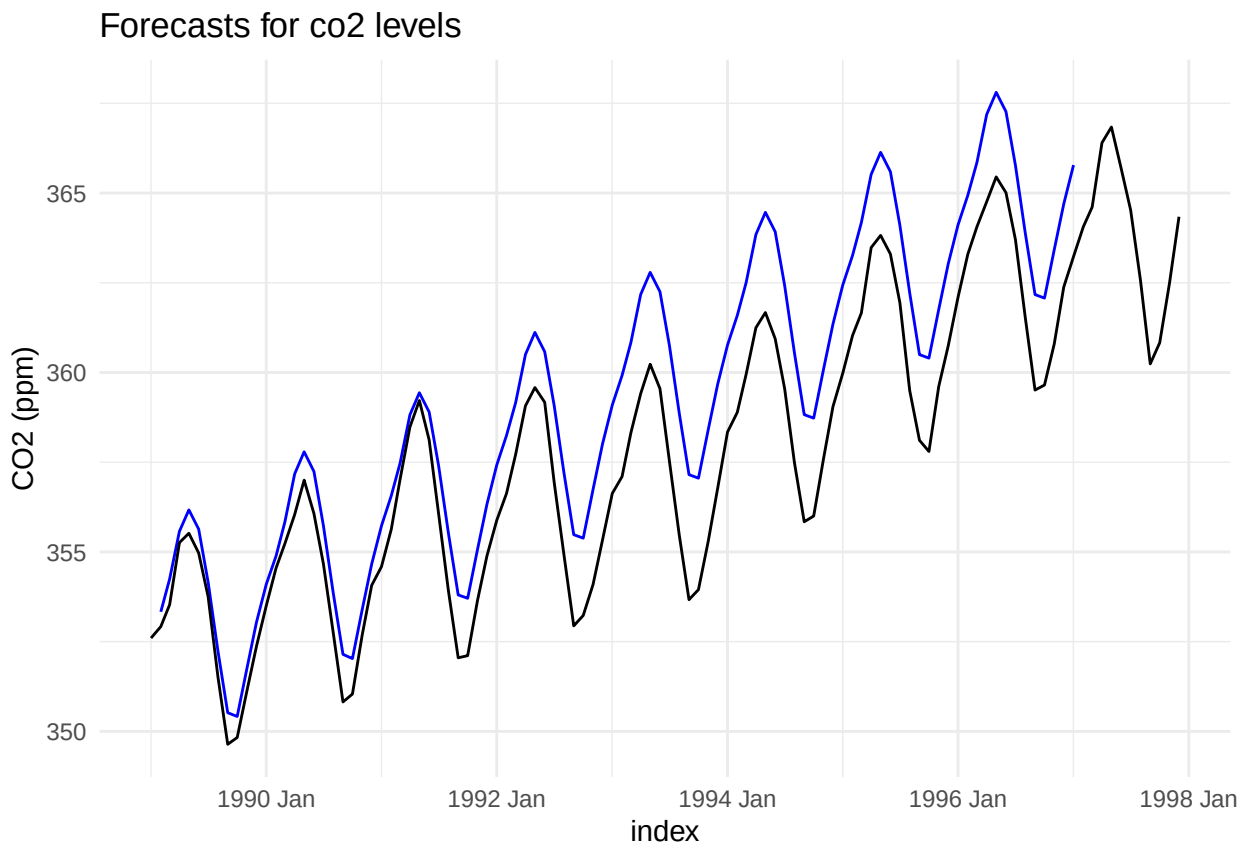
Based on the very wide confidence intervals, we're not very confident that these are accurate predictions. To assess the accuracy of our predictions more methodically, we can create a training set and a test set. From

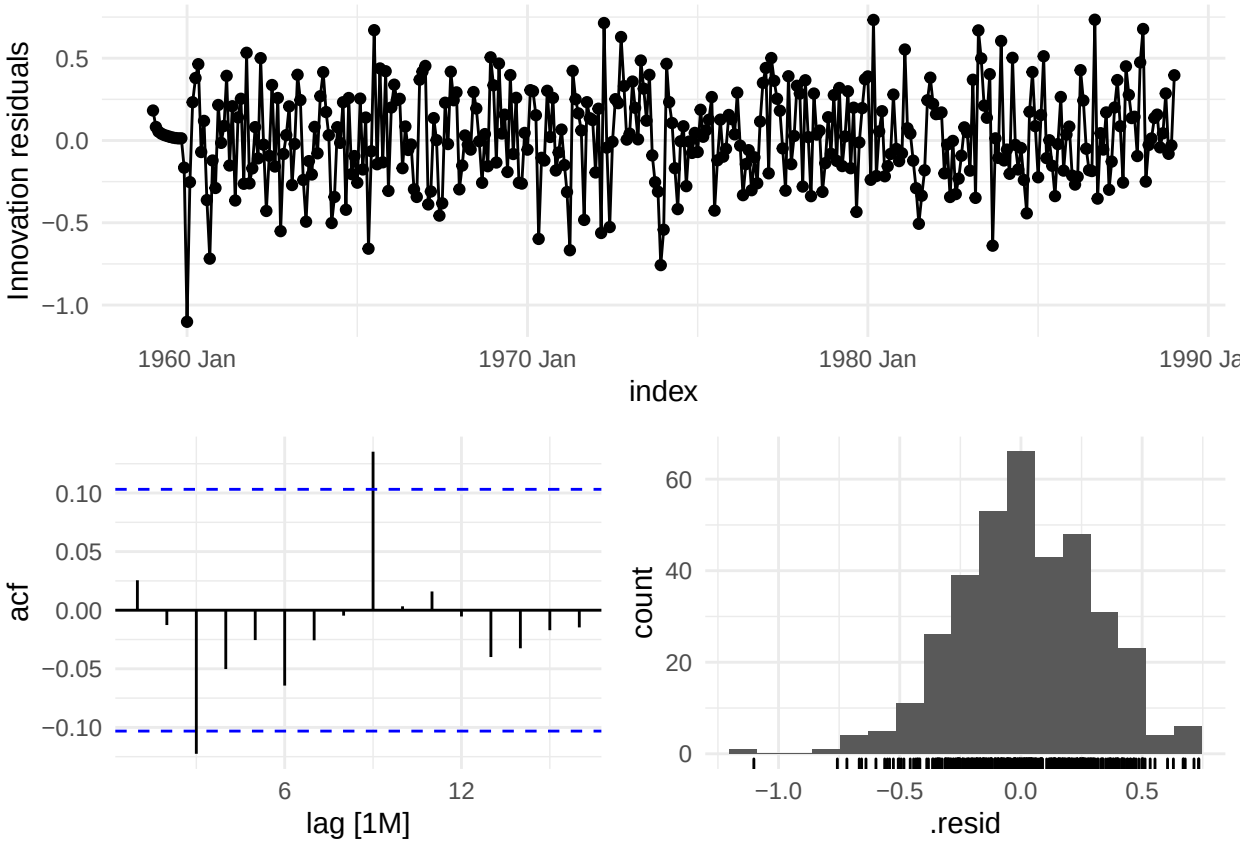
the plot, we can see that our predictions widen from the actuals with each consecutive year in just the last 8 years. While our predictions are probably a best guess given the information that we have, the ARIMA model assumes that seasonality is fixed over time, and this may be a strong assumption to make given that there are other factors that could likely impact CO2 over time.

When we observe the residuals, we also notice that there are two significant spikes in the ACF plot at lag 3 and lag 9. Our p-value from the Box-Ljung test is also greater than 0.05, meaning that we fail to reject the null hypothesis of no serial correlation. However, since both of these observations are close to the significance threshold, this may be cause for concern around the predictive power of our model, particularly for such a far projection in time, as seen below.

```
## # A tsibble: 361 x 3 [1M]
##       index value log_value
##       <mth> <dbl>     <dbl>
##  1 1959 Jan  315.      5.75
##  2 1959 Feb  316.      5.76
##  3 1959 Mar  316.      5.76
##  4 1959 Apr  318.      5.76
##  5 1959 May  318.      5.76
##  6 1959 Jun  318       5.76
##  7 1959 Jul  316.      5.76
##  8 1959 Aug  315.      5.75
##  9 1959 Sep  314.      5.75
## 10 1959 Oct  313.      5.75
## # ... with 351 more rows
```

```
## Series: value
## Model: ARIMA(0,1,1)(2,1,2)[12]
##
## Coefficients:
##           ma1     sar1     sar2     sma1     sma2
##       -0.3595  -0.1484  -0.0750  -0.6937  -0.1217
## s.e.   0.0575   0.5648   0.0658   0.5646   0.4868
##
## sigma^2 estimated as 0.08563:  log likelihood=-63.83
## AIC=139.66   AICc=139.9   BIC=162.77
```


Forecasts for co2 levels
CO2 (ppm)
365
360
355
350
1990 Jan
1992 Jan
1994 Jan
1996 Jan
1998 Jan
index




```
## # A tibble: 1 x 3
##   .model       lb_stat lb_pvalue
##   <chr>          <dbl>     <dbl>
## 1 ARIMA(value)    16.8    0.0778
```