Report from the Point of View of 1997

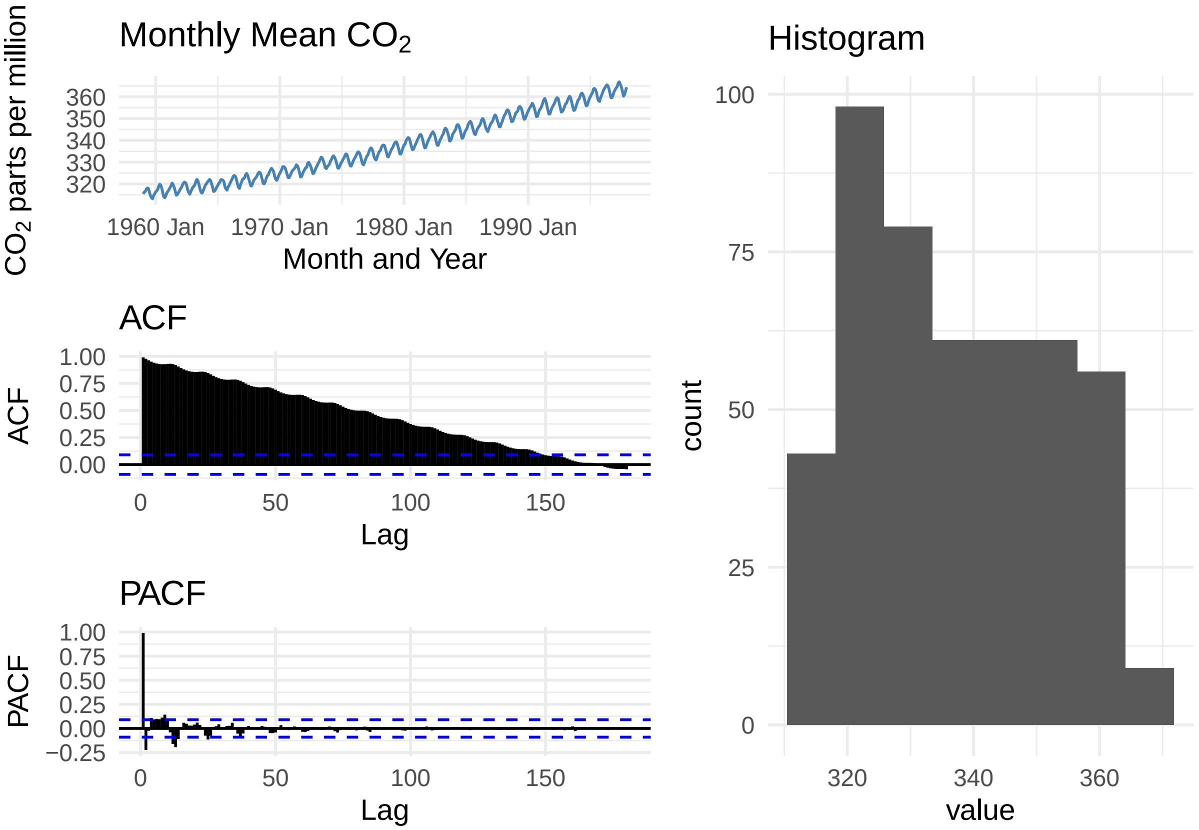
Introduction

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.

Let's take a look at the data. In raw form, it appears as a matrix of doubles that represent the ppm measurements per month and year combination, as appears below. Let's create some initial EDA plots that will allow us to better understand the data. Let's start by analyzing 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.

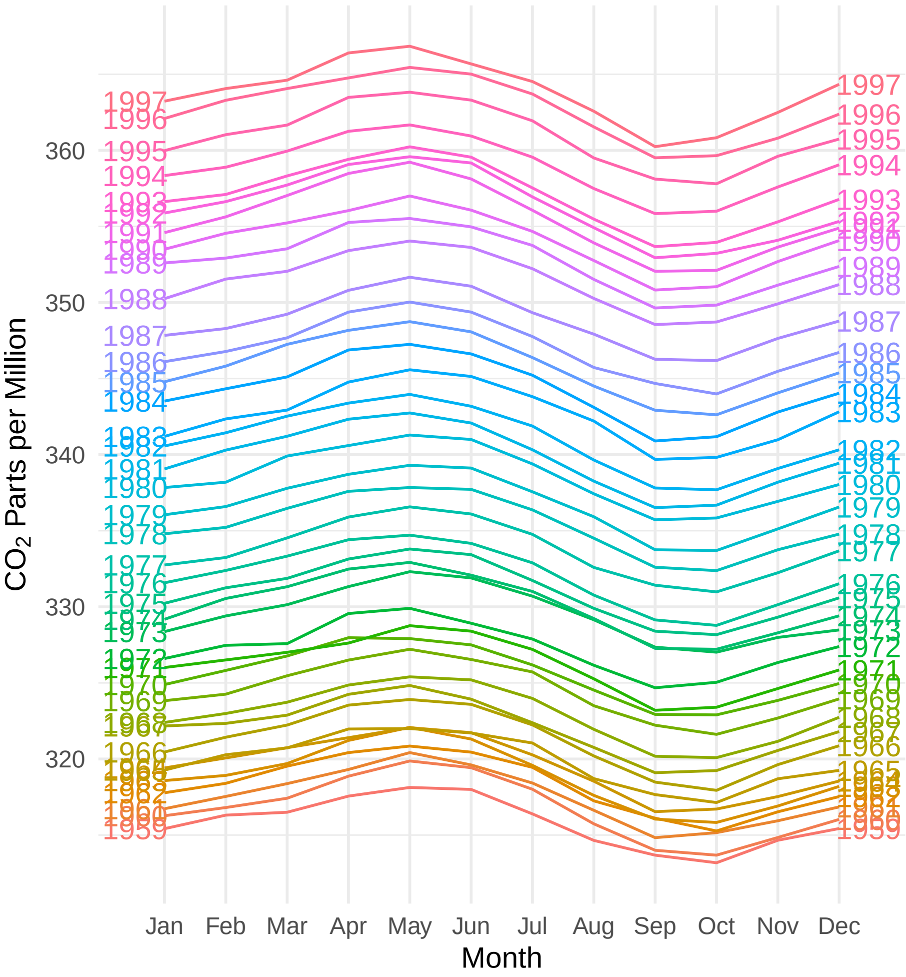
The ACF starts out close to 1 and declines slowly over time, losing significance but staying mostly positive and above the significance line until around lag 140. This slow decline in ACF is what we should see in a time series with a pronounced trend effect, and tracks with what we noticed in the previous 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.

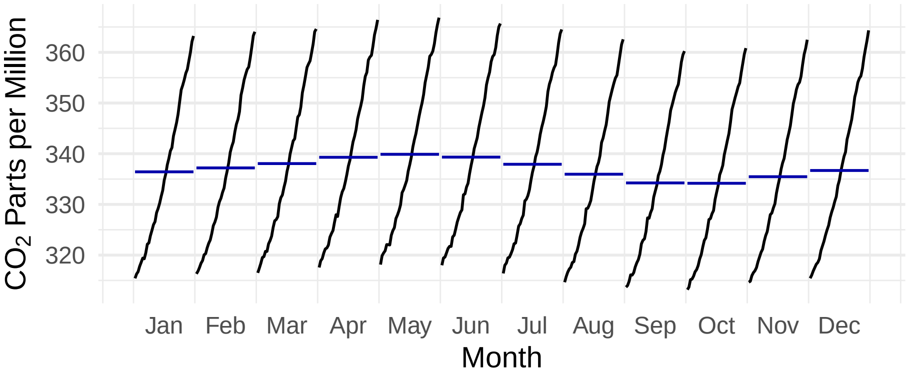
The histogram shows that the CO2 values seem to range between 300 and 380 ppm and do not appear normally distributed, with most values in between these two ranges.

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.

Seasonal plot: Monthly mean CO_2

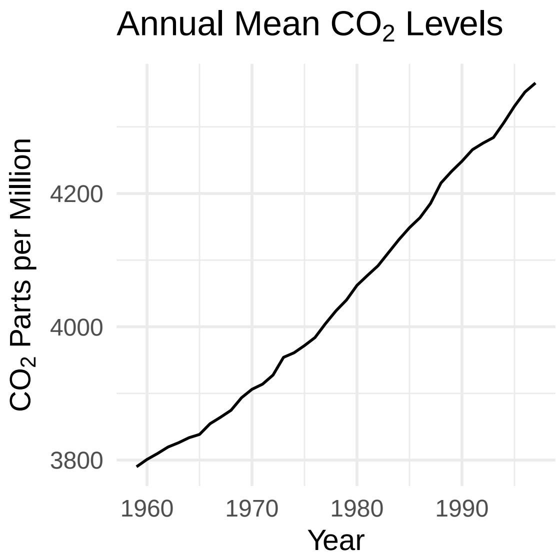


Subseries plot: Monthly mean CO_2

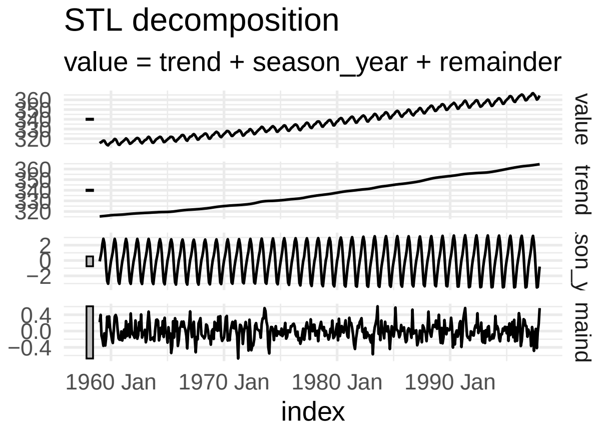


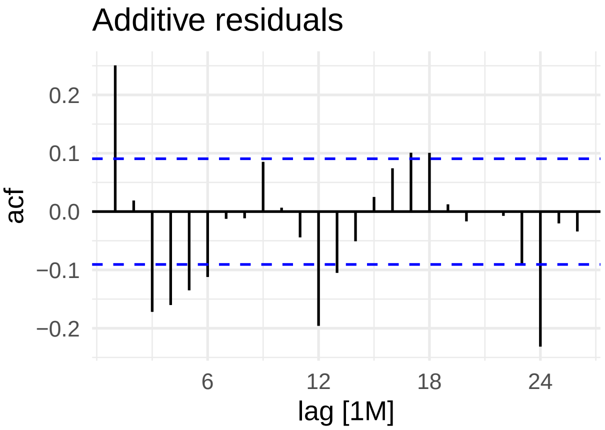
From these plots, we can see that the CO2 levels appear to increase from January through May, then decrease from June through October, hitting a low in October, and then increase again from November through end of the year. The source report does mention that plants and soil absorbing and emitting CO2 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.

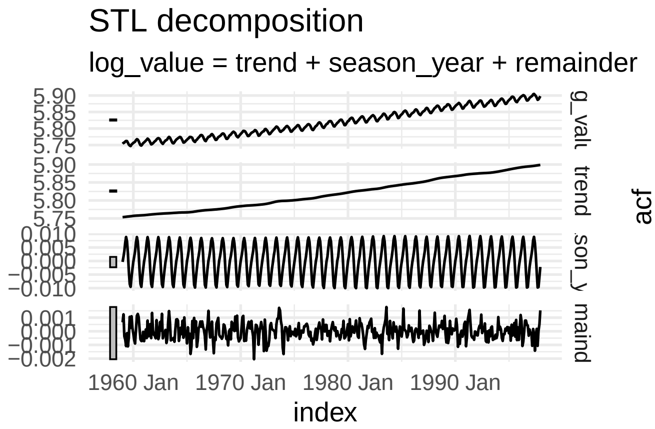
We can confirm that our time series trend is increasing by removing our seasonality components and aggregating our data by year instead of month, as seen below.

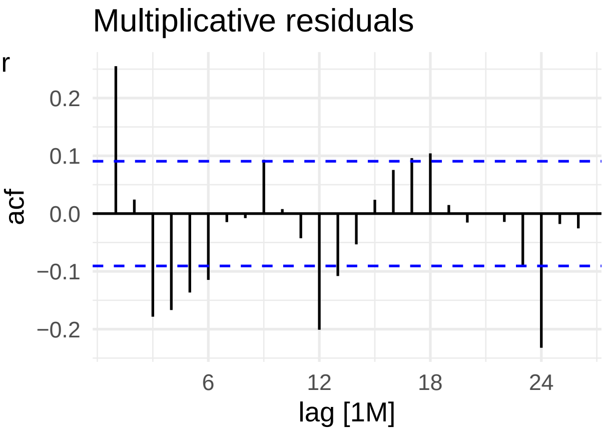


We can also use both additive and multiplicative decomposition to remove both the trend and seasonal movements from our dataset and confirm that the variance is stationary. The results from these decompositions are plotted below.









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 a bit more stable, supporting the idea that this might actually be a multiplicative time series. In the next section we should take a 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 from our visual analysis of the time series plots, the time series does not appear to be stationary as the mean 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.

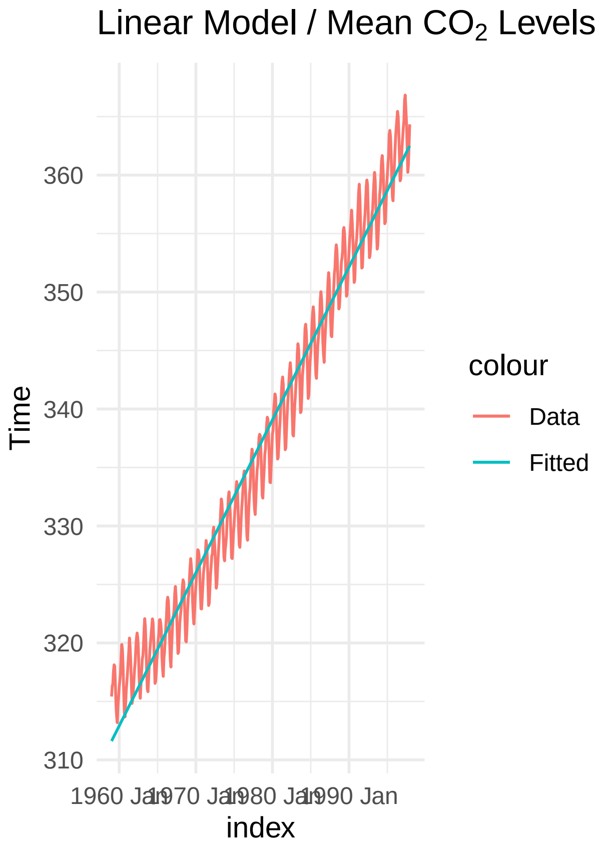
Modeling

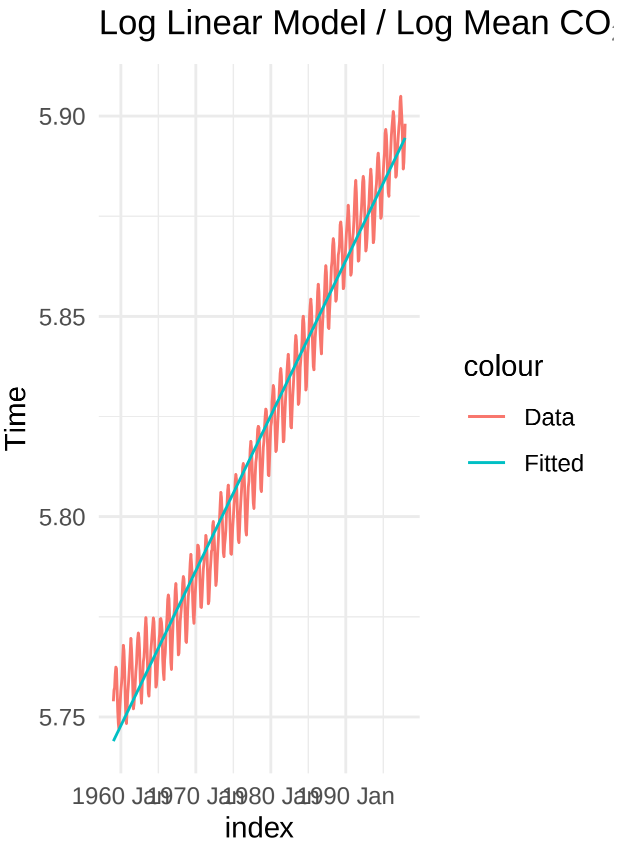
Let's start by creating two linear models, one fit on our CO2 data and one fit on the log transform of our CO2 data. Since in our EDA we did notice that the fluctuations within the seasonality of the time series seem to grow slightly larger over time, indicating a potentially multiplicative series, we 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
```

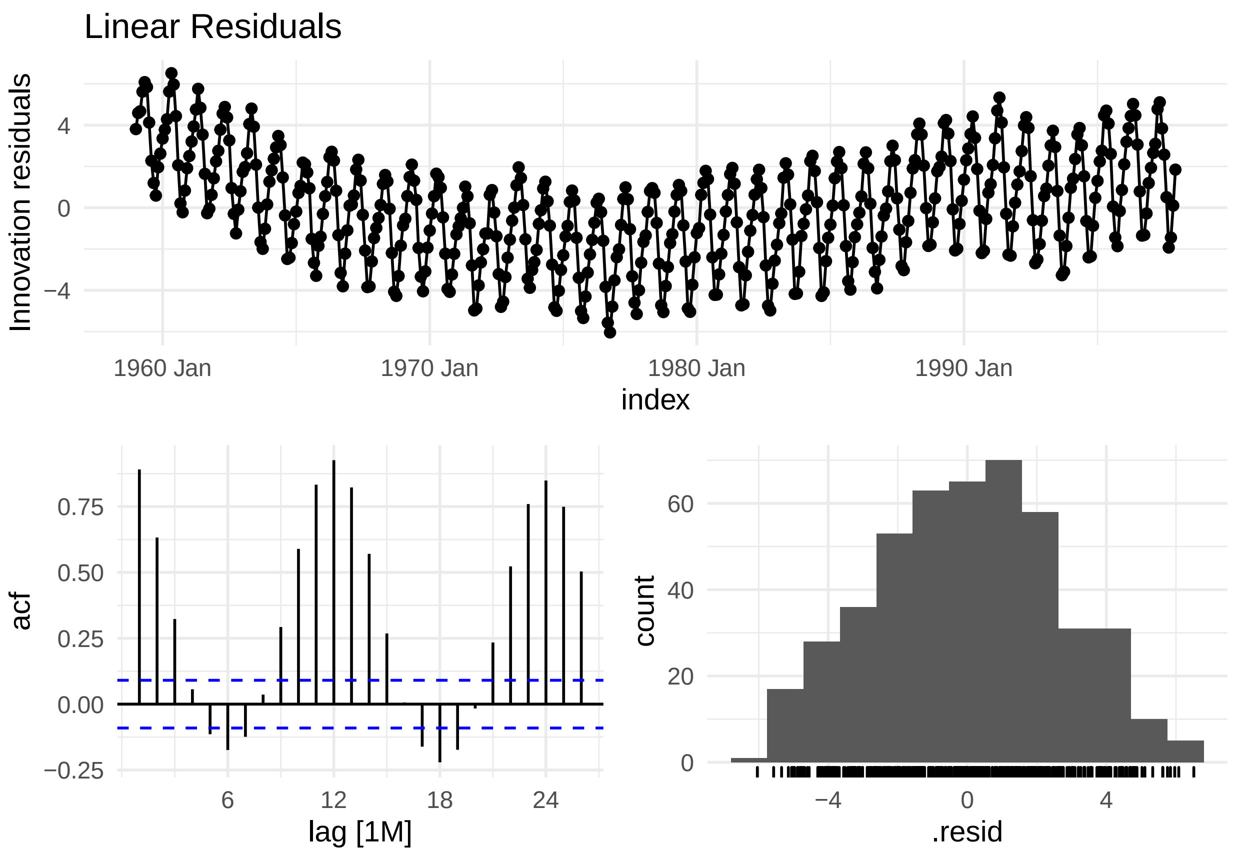
```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0172650 -0.0056145  0.0002764  0.0053760  0.0187770
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.744e+00  6.829e-04  8410.5   <2e-16 ***
## 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
```

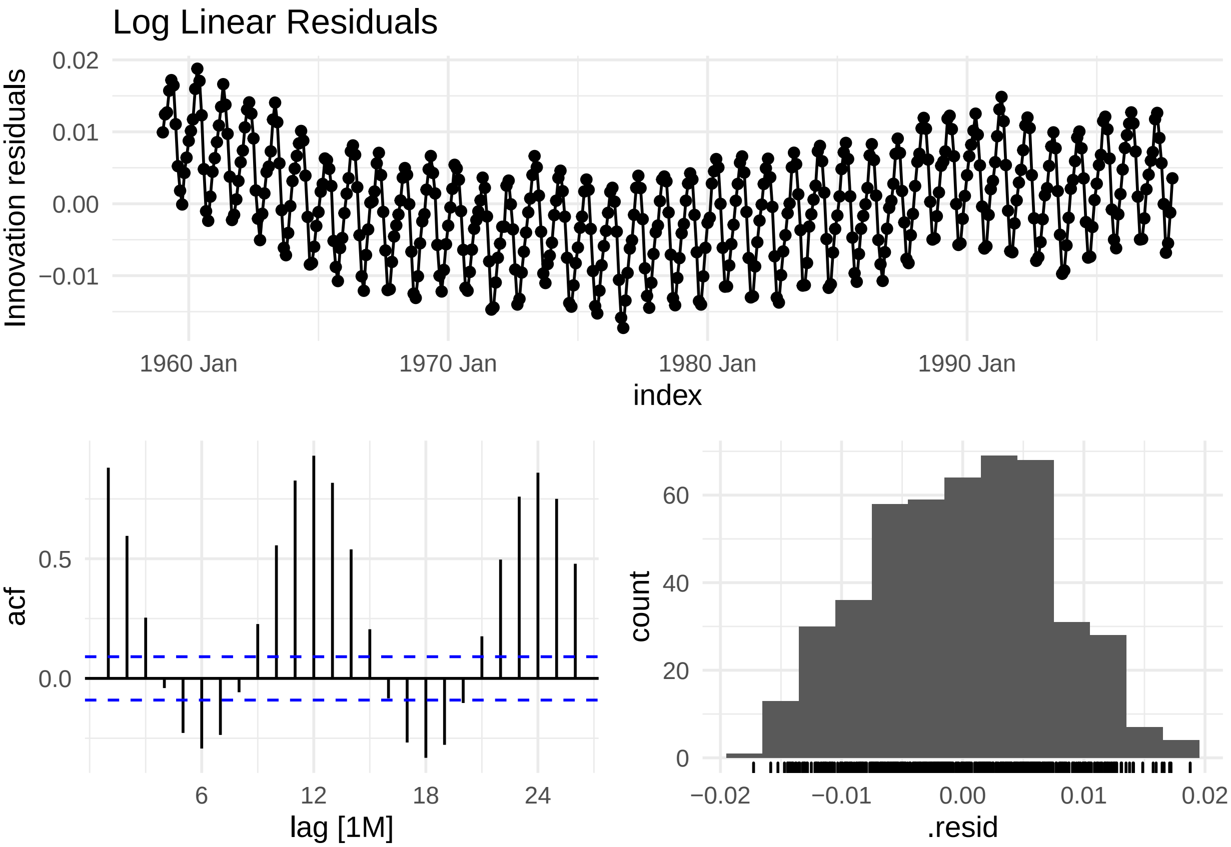
We can see from the output here that the R-squared values on both models are high and the input coefficients are highly significant on both. We will also plot the models on top of our data.





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
```

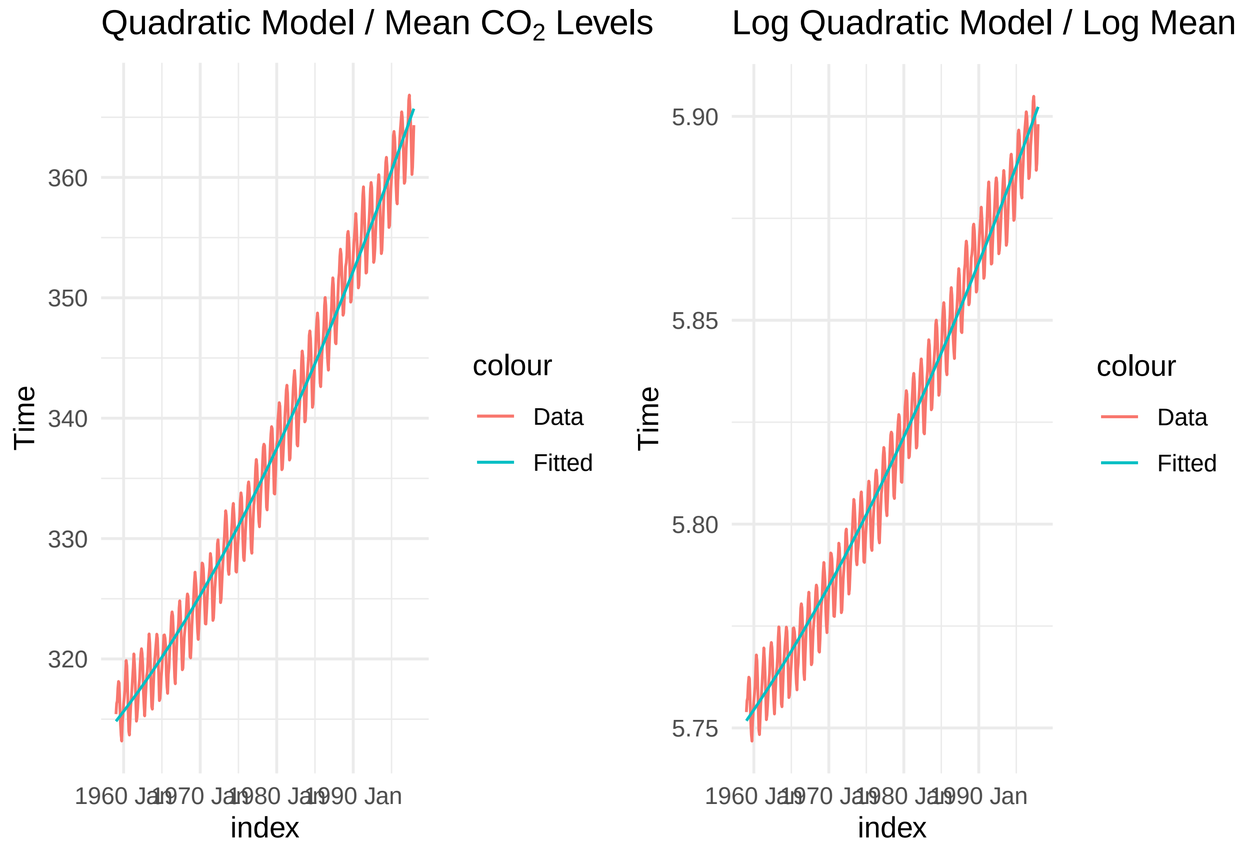
Based on the Ljung Box test, we can reject the null that the data is independently distributed up to 10 lags, which means that we likely do not have white noise residuals and both our models are failing to account for some variance in our data (likely the seasonality).

Let's repeat this process with a quadratic trend model.

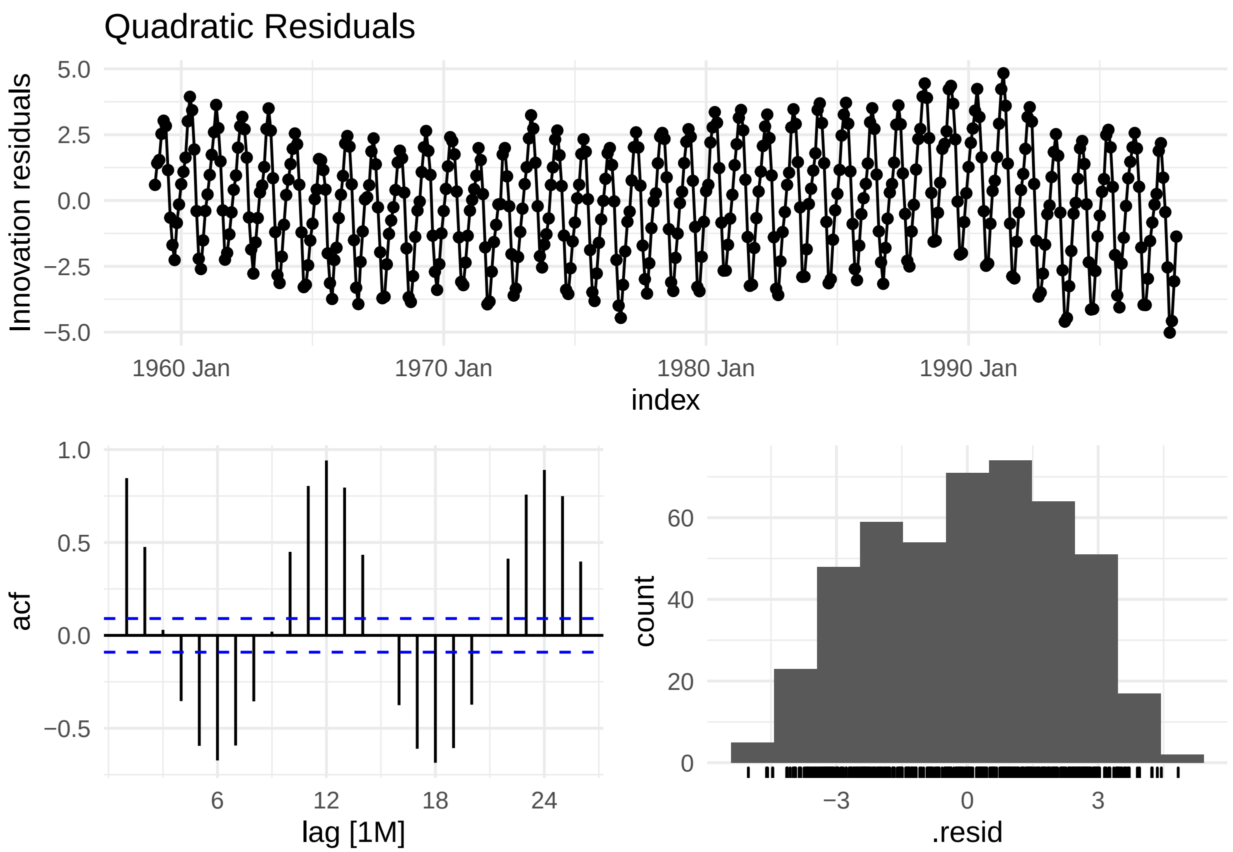
```
## Series: value
## Model: TSLM
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0195 -1.7120  0.2144  1.7957  4.8345
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.148e+02  3.039e-01 1035.65   <2e-16 ***
## trend()      6.739e-02  2.993e-03   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
```

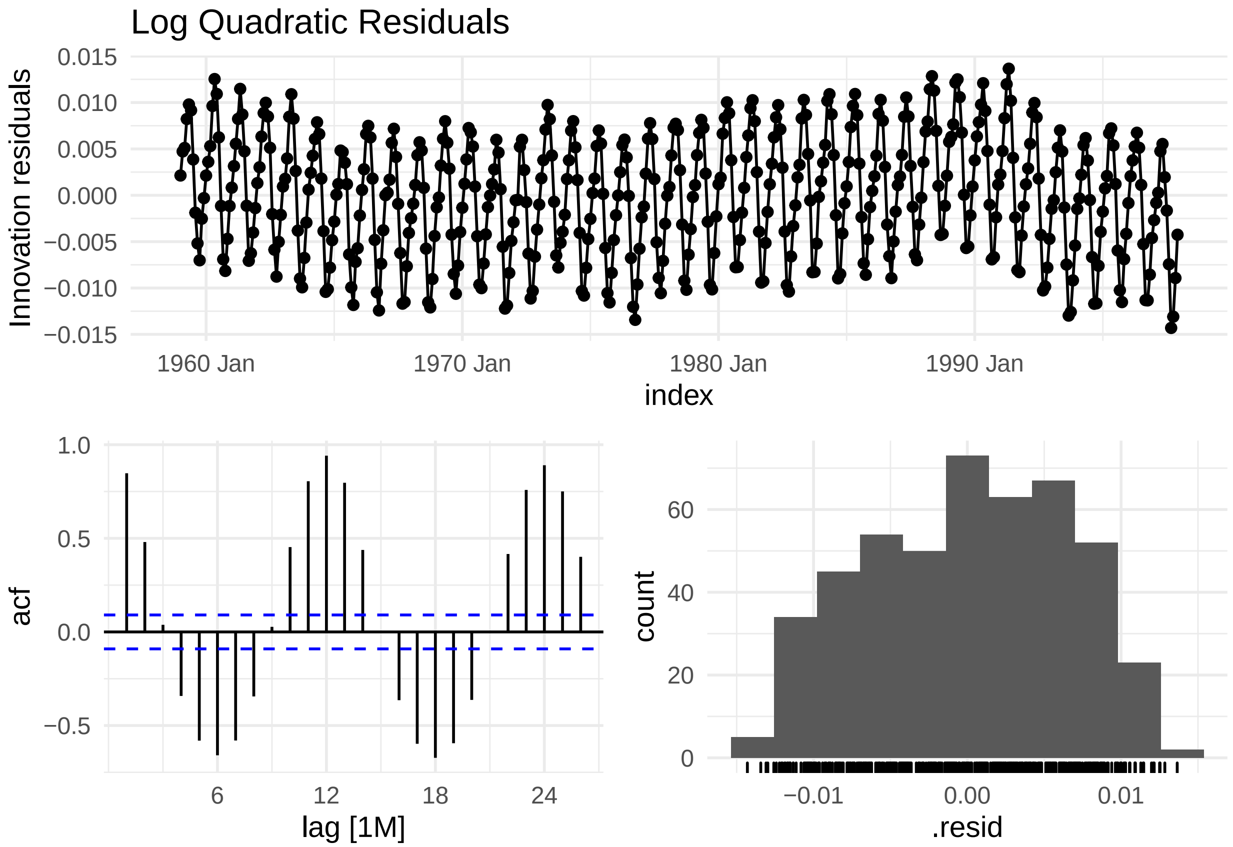
```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0143052 -0.0050832  0.0005277  0.0052757  0.0136508
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.752e+00  9.039e-04 6363.18   <2e-16 ***
## 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, so I am just 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 using the 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.

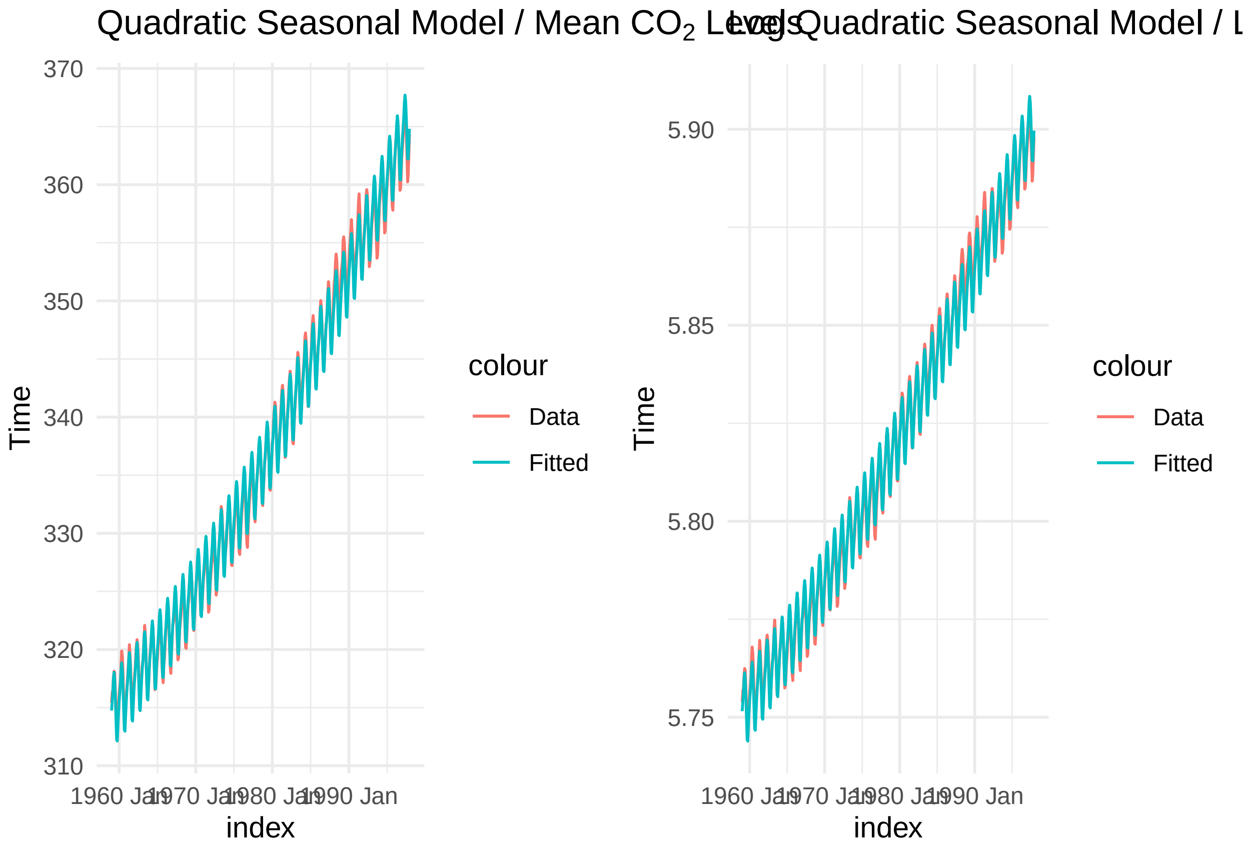
```
fit_quadratic_season <- co2_df %>%
  model(trend_model = TSLM(value ~ trend() + I(trend()^2) + season()))

fit_quadratic_log_season <- co2_df %>%
  model(trend_model = TSLM(log_value ~ trend() + I(trend()^2) + season()))

fit_quadratic_season %>% report()
```

```
## 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
```

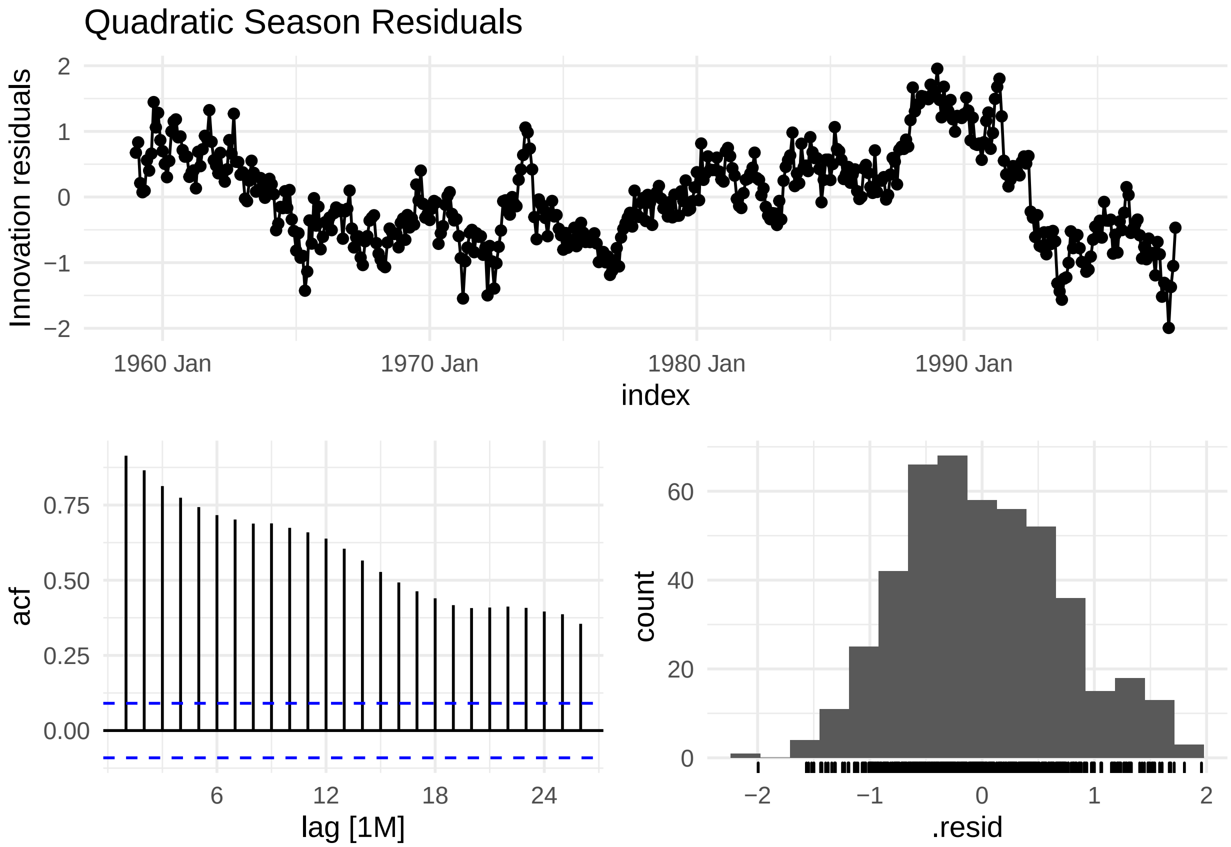
```
fit_quadratic_log_season %>% report()
```

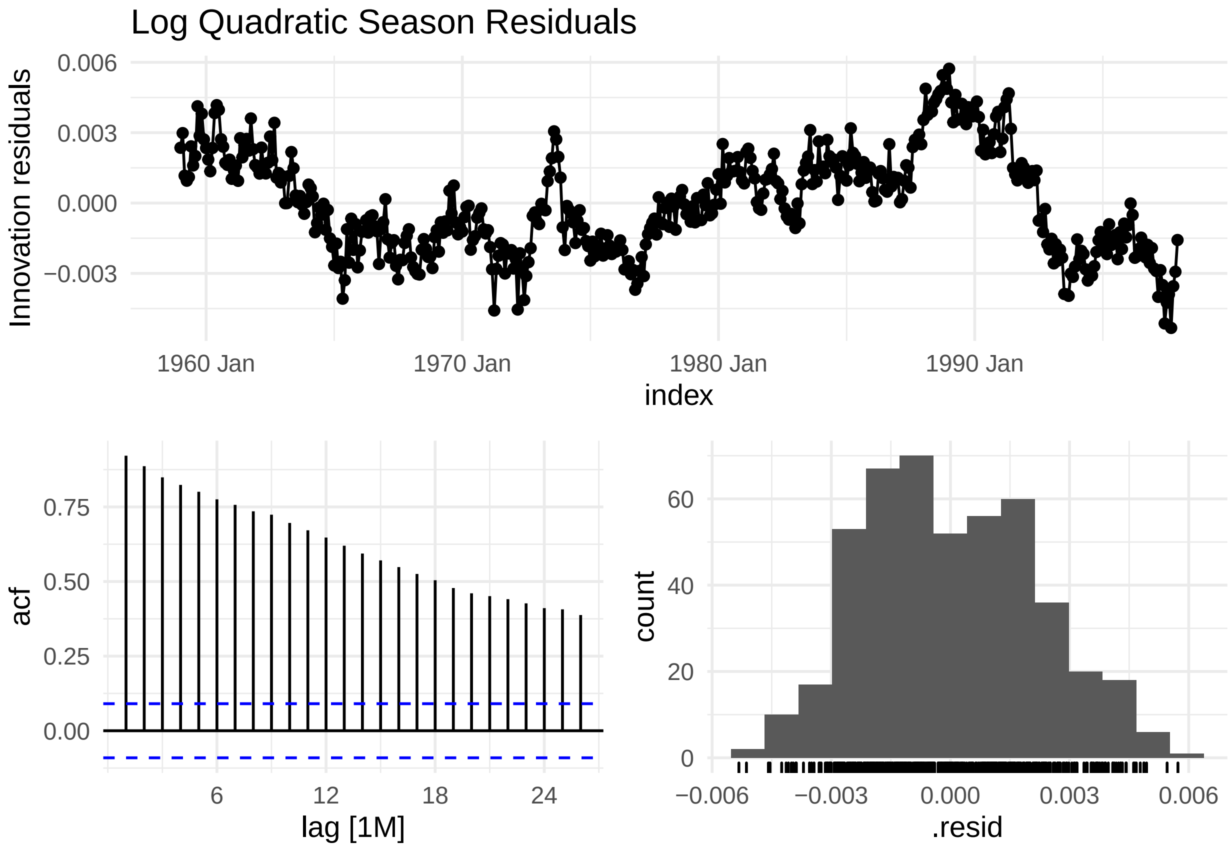
```
## Series: log_value
## Model: TSLM
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0053270 -0.0017362 -0.0001774  0.0015139  0.0057292
##
## Coefficients:
##                  Estimate Std. Error   t value Pr(>|t|)
## (Intercept)     5.751e+00  4.533e-04 12687.967  < 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. 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
##
```

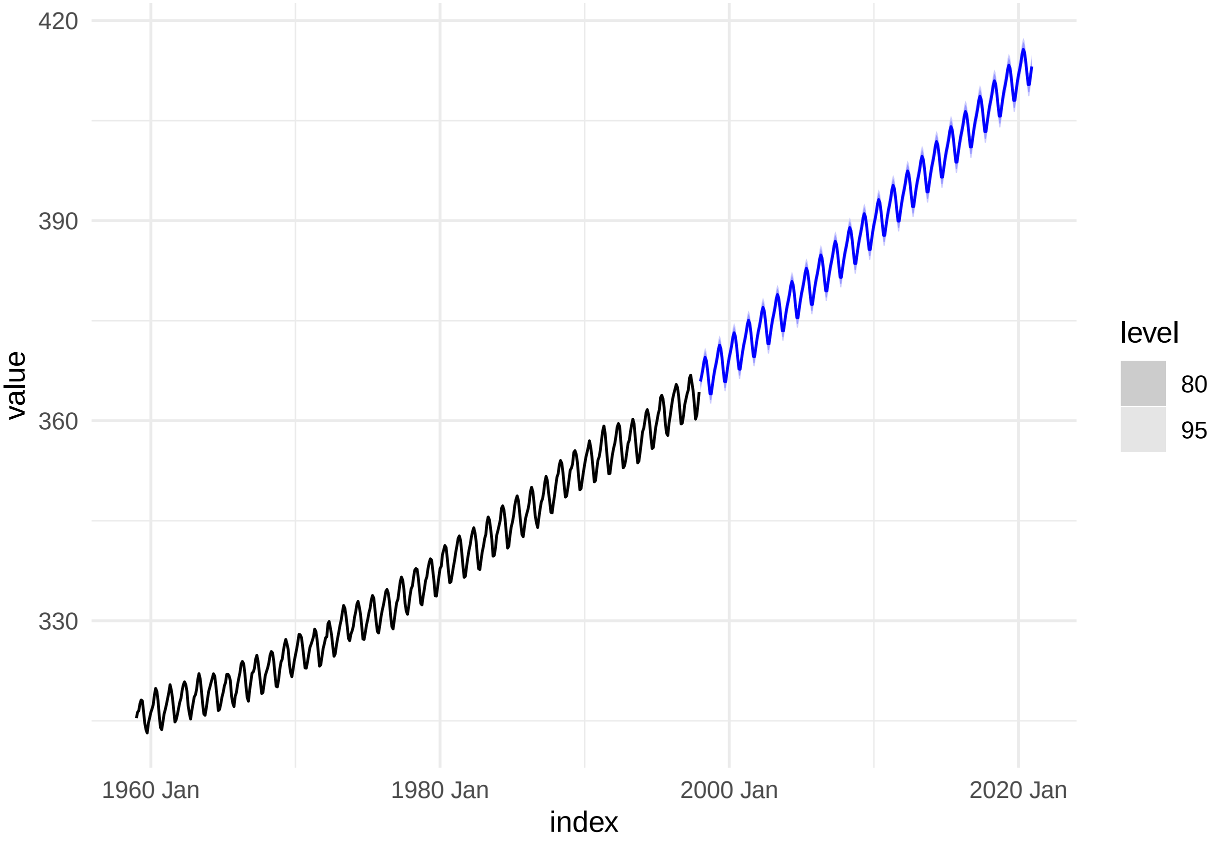
```
## data:  quadratic_log_season_residuals
## X-squared = 3040.2, df = 10, p-value < 2.2e-16
```

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 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 the trend and seasonal fluctuations that we noticed in our data.

TODO: Check CLM assumptions? https://github.com/mids-271/summer_22_central/blob/master/Live_session_and_solutions/LS_7_Solutions/LS-7-Solutions.pdf