Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Contents

1 Report from the Point of View of 1997

1.1 (3 points) Task 0a: Introduction

Climate change is an increasingly pertinent issue for scientists and policymakers alike, as global temperatures rise. It is crucial to understand the underlying reasons for this increase, and its relationship with carbon emissions. This report presents potential outcomes of this constant increase, and highlights the need to anticipate future impacts of carbon emission reduction efforts.

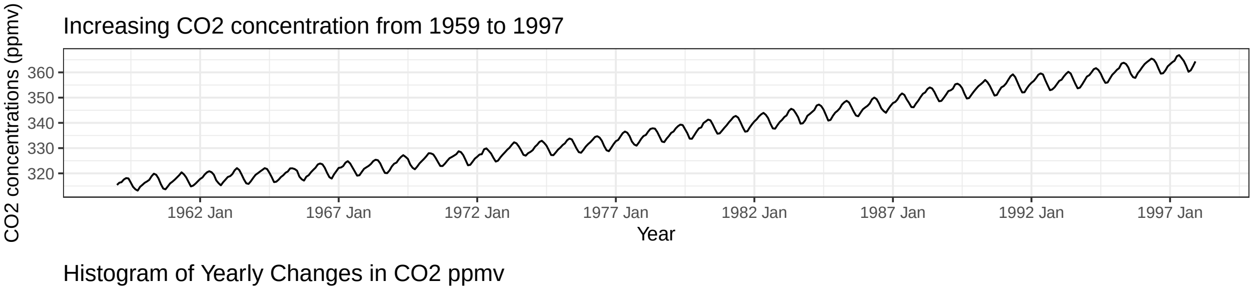
Geochemist Dr. Charles David Keeling's pioneering work in atmospheric carbon dioxide measurements fundamentally reshaped our understanding of the global carbon cycle and its impact on climate change. In 1958, Keeling initiated a long-term study at the Mauna Loa Observatory, producing the iconic "Keeling Curve," which revealed the steady rise of atmospheric CO2. His research confirmed that fossil fuel combustion was contributing to increasing CO2 levels, a discovery with profound social and political consequences. This work also paved the way for further investigations into other greenhouse gases and established benchmarks for testing climate models.

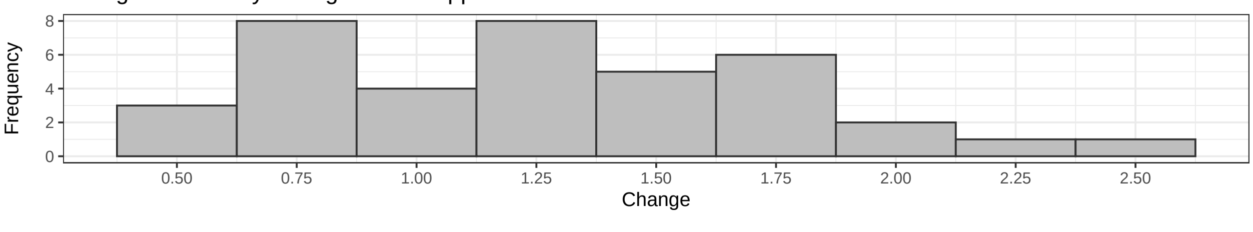
CO2 is classified as a "greenhouse gas," meaning that it traps heat in the atmosphere and lead to rising global temperatures when in high concentrations. It can be important to track Co2 levels as rising global temperatures can lead to imbalances in ecosystems and rising water levels that impact both animal and human life. Monitoring CO2 levels is critical because rising concentrations contribute to global warming, with severe consequences for ecosystems, sea levels, and both human and animal life. Understanding these trends is essential for assessing the long-term impact of human activities and guiding future policies.

1.2 (3 points) Task 1a: CO2 data

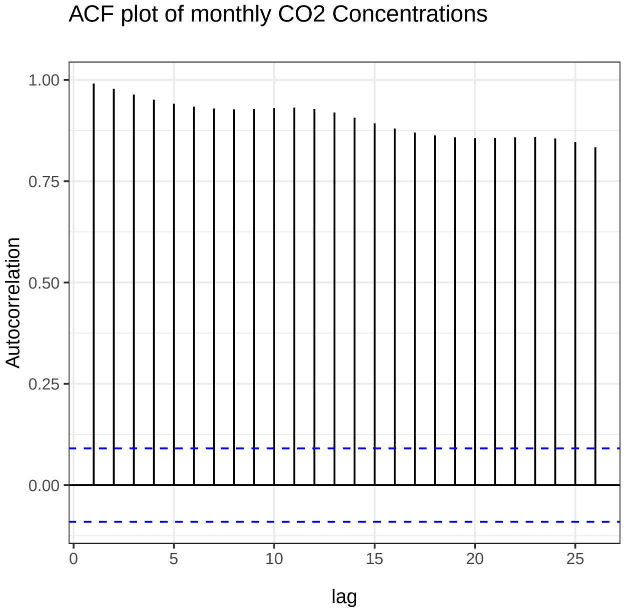
The current data is gathered from measurements made under Dr. Charles Keeling's study at the Mauna Loa Observatory in Hawaii (Cleaveland, 1993). Measurements were taken by a chemical gas analyzer sensor, with detections based on infrared absorption. This data measures monthly CO2 concentration levels from January 1959 to December 1997. Units are in parts per million of CO2 (abbreviated as ppmv) using the SIO manometric mole fraction scale. Dr. Keeling initially designed a device to detect Co2 concentrations to detect CO2 emitted from limestone near bodies of water. But his measurements revealed a pattern of increasing CO2 concentrations at the global scale, urging further need to continue tracking the gas (Keeling, 1998).

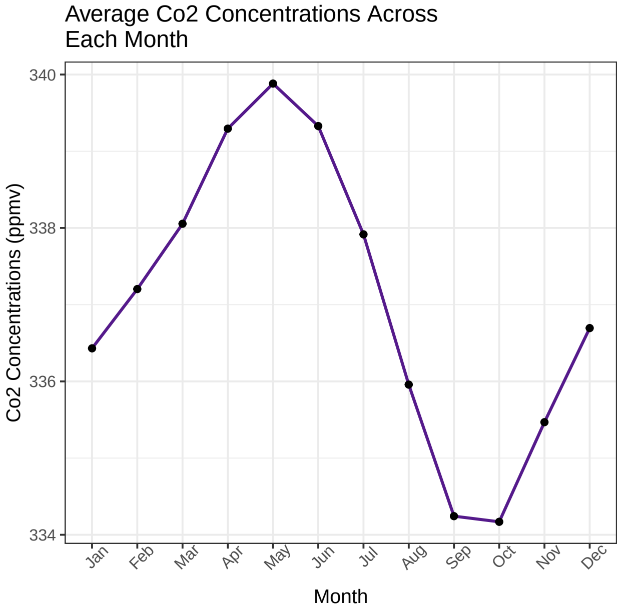
The time series shows a clear upward trend of global CO2 concentrations from 1959 to 1998, with an average increase in 1.26 CO2 ppmv and a standard deviation of .51 CO2 ppmv. Upon inspection of the yearly increases, the bulk of changing CO2 levels are between 0.5 and 2.0 CO2 ppmv.





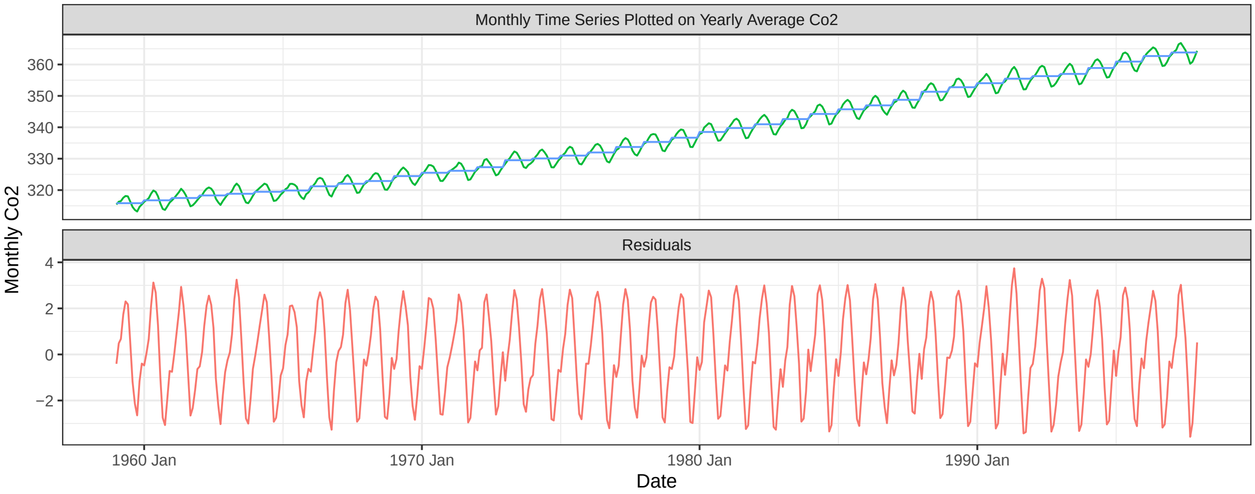
The time series also shows strong evidence of seasonality corresponding closely with the meteorological seasons of Autumn, Winter, Spring, and Summer. We now look at the ACF plot and average CO2 concentration for each month to gain further clarity on the seasonality.





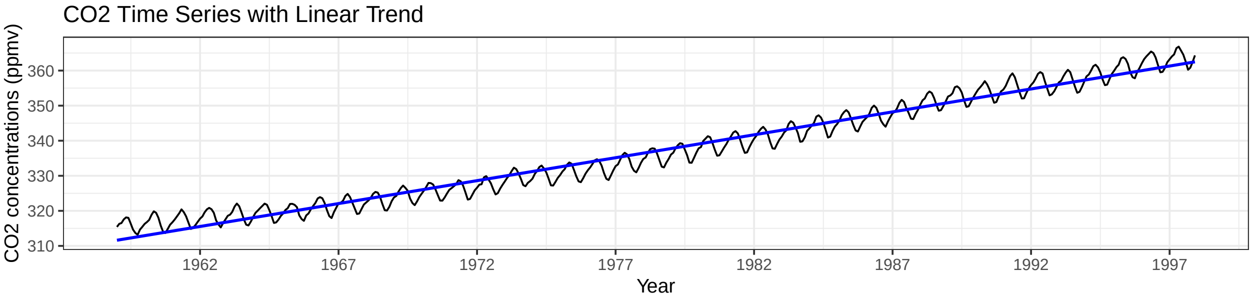
We also see a scallop/wave shaped pattern among correlations between the current value with growing lags. Clearer evidence of seasonality is shown when inspecting the monthly average the Co2 ppmv, when averaged across all years in the available data. CO2 contration peaks at the start of summer, and drops to a low in the fall, before rising again. This is likely due to the organic decomposition of plant life in these seasons (Keeling, 1960).

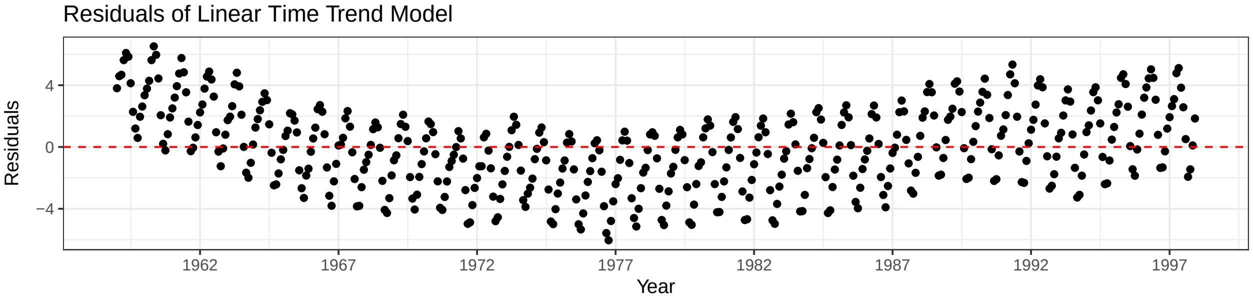
We now study the time series' stationarity. We first conduct the Augmented Dickey-Fuller Test to test the null hypothesis that the time series is not stationary. As seen in the time series plot for `co2`, we have a clear upward trend, suggesting non-stationarity. This is confirmed by a p-value of 0.2269 yielded by the test, which indicates insufficient evidence to reject the null hypothesis of non-stationarity. To look at stationarity in variance, we fit a yearly CO2 average on the monthly time series, and inspect the residuals from year to year. Although there are slight changes in the variance, they seem to regress to a constant variance over time. Thus, once we account for the yearly increases in CO2 ppmv, there is likely a constant variance over time.



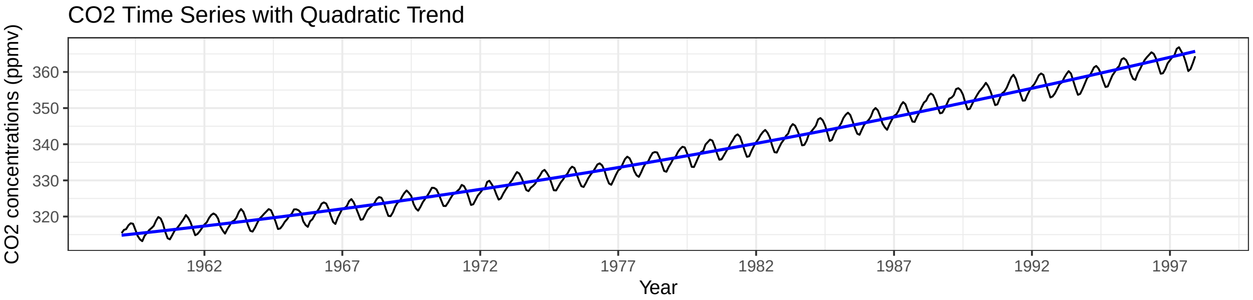
1.3 (3 points) Task 2a: Linear time trend model

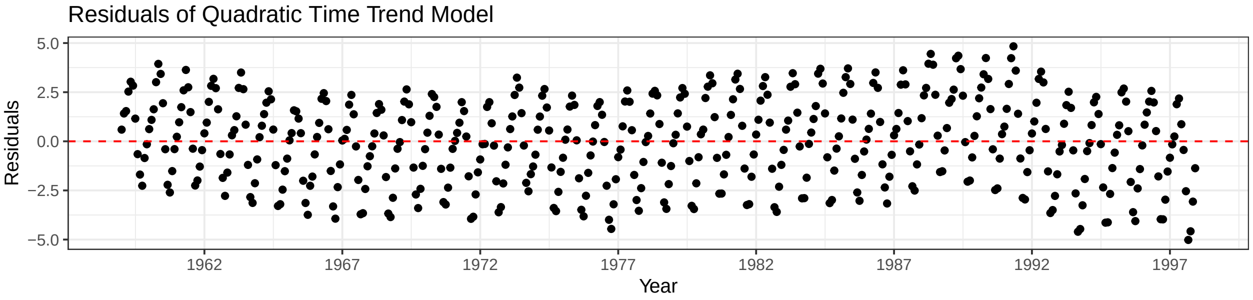
We now fit a linear time trend model the `co2` series, and examine the characteristics of the fit and residuals.


CO2 Time Series with Linear Trend

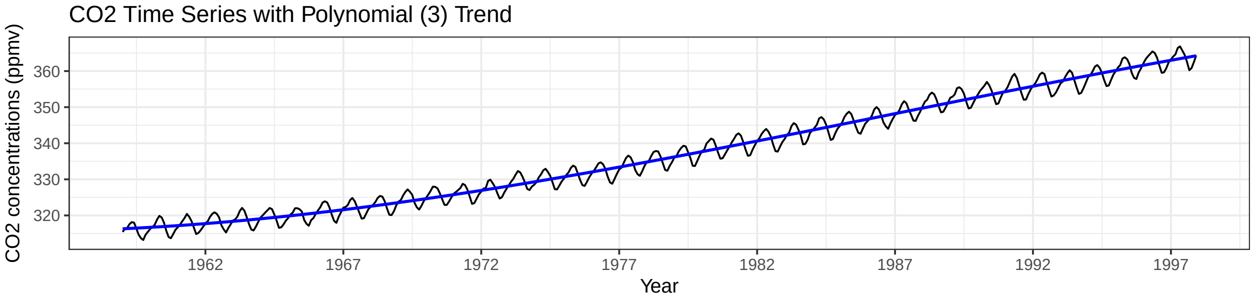

Residuals of Linear Time Trend Model

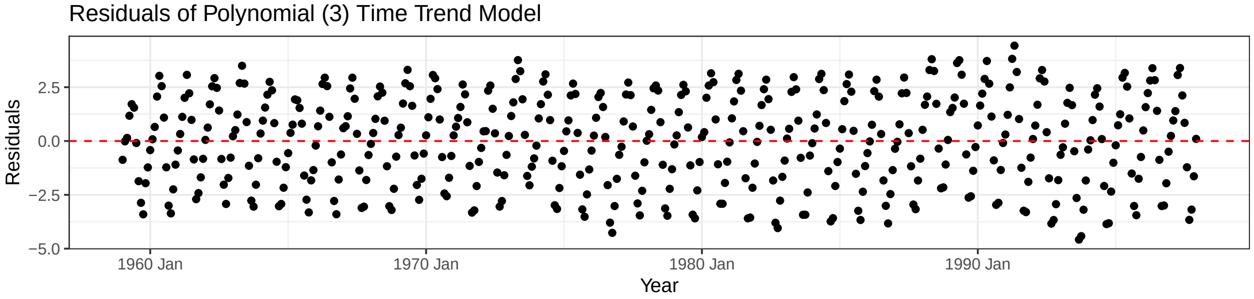
Upon inspection of linear fit, the fitted line appears to be systematically overestimating values at certain points and underestimating values at other points. This indicates that perhaps a higher order polynomial might produce a better fit of the overall trend. The residuals of the linear model also exhibit a cyclical, non-linear pattern, indicating that the model does not capture the seasonality in the data. The overall curve also suggests that the linear model insufficiently captures the overall trend. We now try a quadratic model, which may better capture the underlying trend.


CO2 Time Series with Quadratic Trend


Residuals of Quadratic Time Trend Model

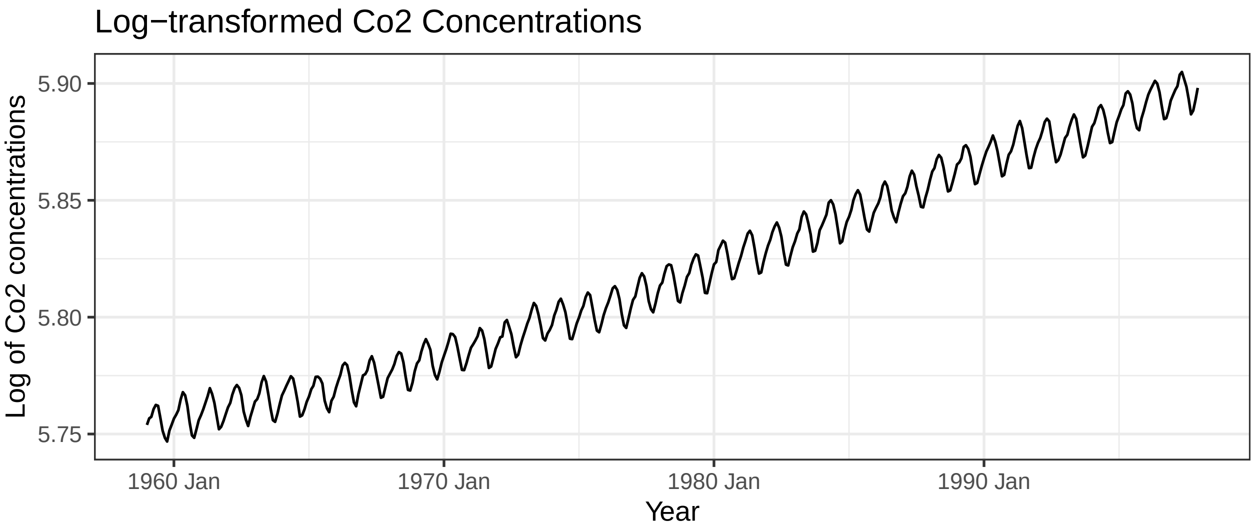
The quadratic model's residuals indicate a small reduction in variance, demonstrating a slightly improved fit. However, the cyclical behavior remains, indicating that seasonality is unaccounted for in the model still. There is also an overall non-random trend in the residuals, indicating that the model still may not capture all the structural details. We now fit a polynomial model to the data to see if there is an improved fit.





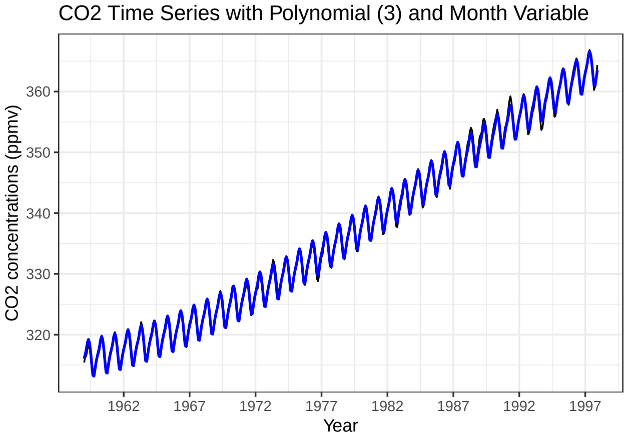
The third-order polynomial model demonstrates improved residual behavior compared to quadratic and linear models. We chose to stop at this order to prevent excessive overfitting, as higher-order polynomials showed diminishing returns in model performance.

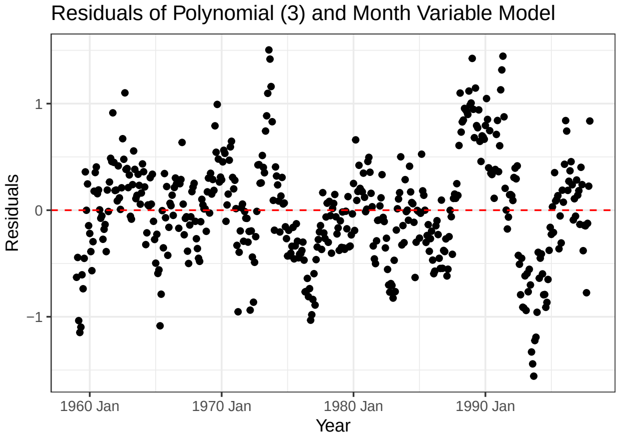
Apart from transforming the orders of the model, we were interested in data transformations - specifically logarithmic. As such we expermiented with a logarithmic dataset to observe the pattern of the data values.



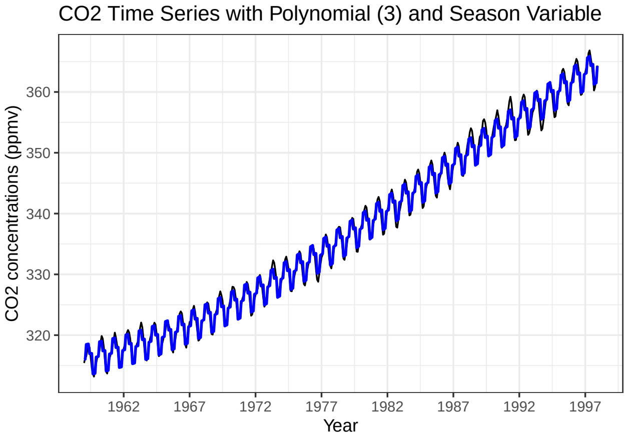
The logarithmic transformation reduces variance but offers minimal improvement compared to traditional plotting. This limited impact is likely due to the cyclical nature of the time series, which the transformation does not adequately address.

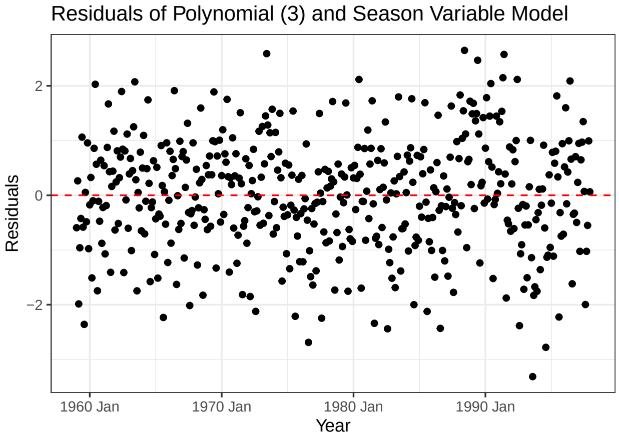
To address the cyclical behavior, we developed another polynomial model that includes each month as a variable. The average monthly CO2 emissions indicate significant cyclic patterns at the monthly level. By incorporating this variable, we anticipate an improvement in the fit of our time series model.


CO2 Time Series with Polynomial (3) and Month Variable

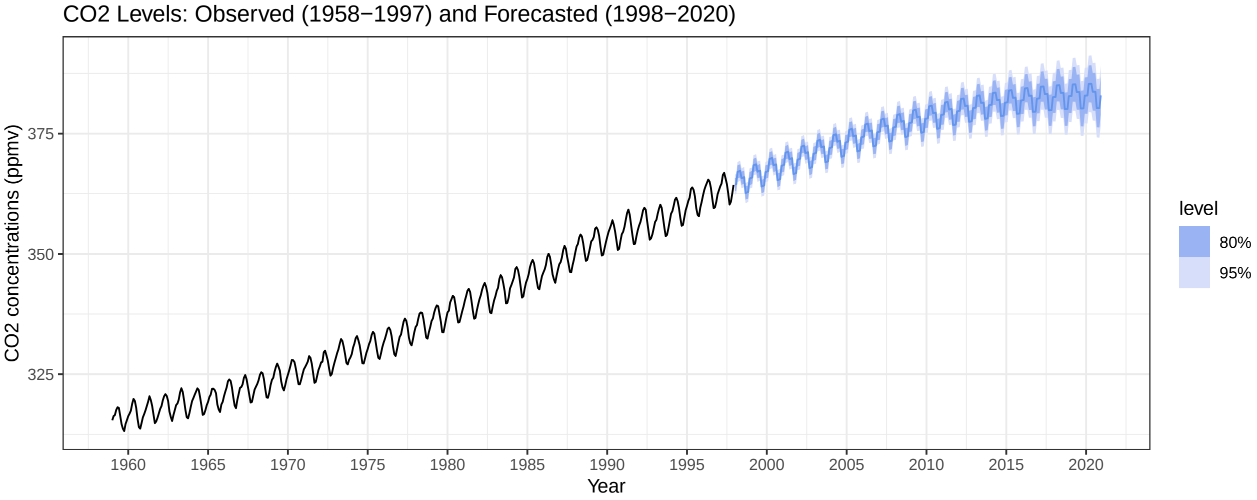

Residuals of Polynomial (3) and Month Variable Model

Incorporating the `month` dummy variable brought the residuals closer to zero, ranging between 1 and -1, but they still displayed a seasonal pattern. To refine the model, we grouped the months into quarters, to represent the seasons as a categorical variable, `season`.


CO2 Time Series with Polynomial (3) and Season Variable


Residuals of Polynomial (3) and Season Variable Model

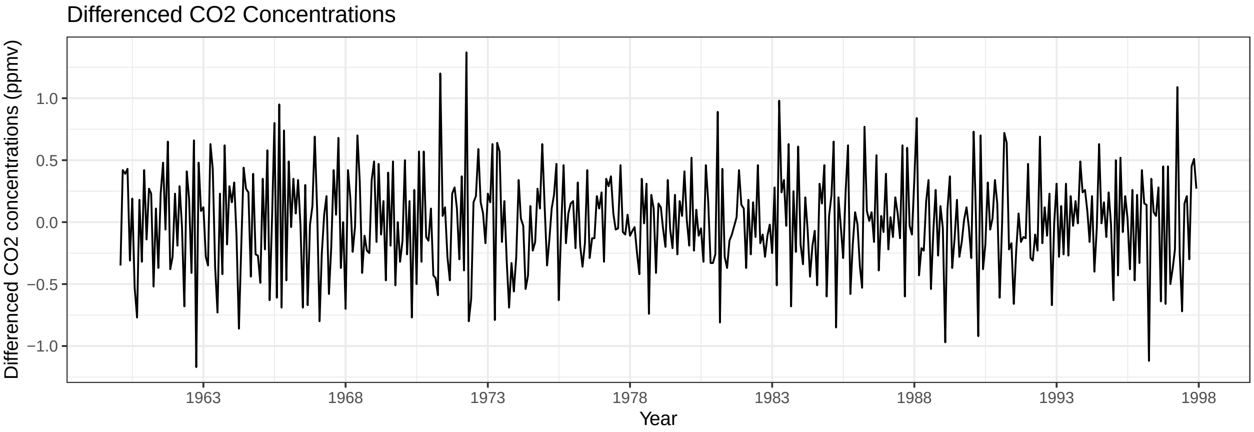
We see that using the `season` variable centered the residuals around zero with a random distribution, though fluctuations remained between 2 and -2. We now proceed with this model, the polynomial model with the `season` dummy variable, as it has the residuals plot that most look like a random distribution around the red line centered on zero, and developed a forecast for CO2 emissions through 2020.


CO2 Levels: Observed (1958−1997) and Forecasted (1998−2020)

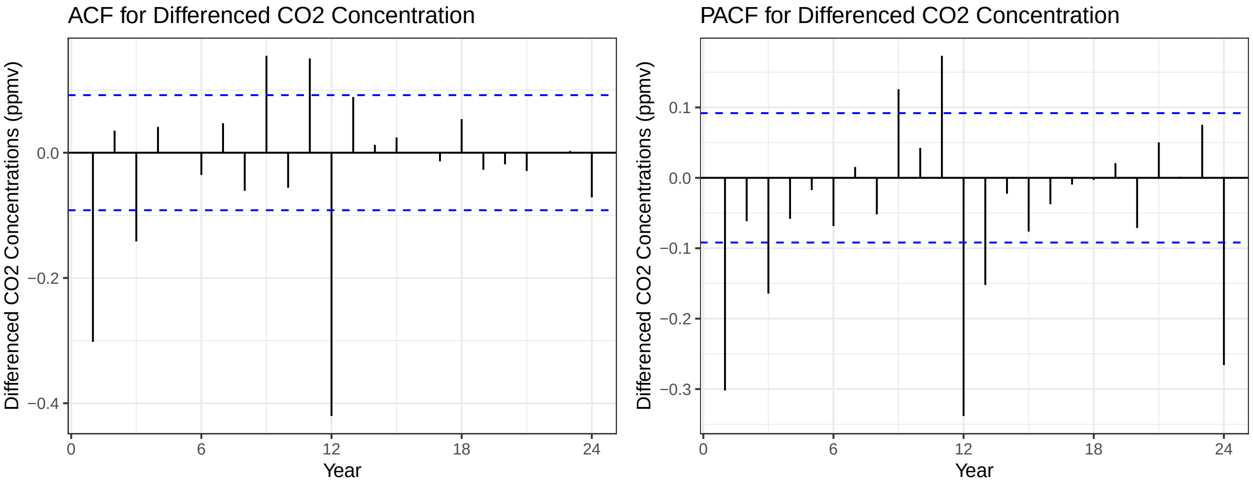
The forecast model using the `season` variable shows decent performance, and predicts for the upward trend to persist up to 2020, along with the annual seasonality. We will now explore an ARIMA model to see if it may better capture the time series' underlying patterns and improve forecast accuracy.

1.4 (3 points) Task 3a: ARIMA times series model

As seen in our EDA, there is non-stationarity. Thus, we will proceed to difference the data to make it stationary, both at the 1st lag followed by the 12th lag to account for seasonality, which is a crucial step before fitting the ARIMA model effectively.



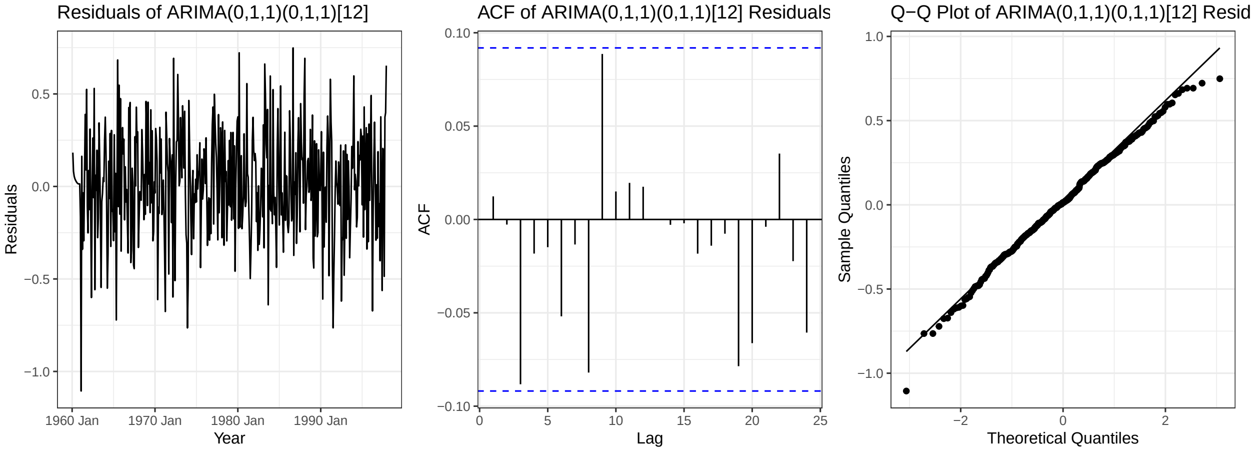
The plot of the differenced time series does look more stationary in the mean and variance; which is confirmed by a 0.01 p-value yielded from the Augmented Dickey-Fuller Test, indicating sufficient evidence to reject the null hypothesis of non-stationarity. We now look at the ACF and PACF plots of the differenced time series, to inform how we should construct our ARIMA model.



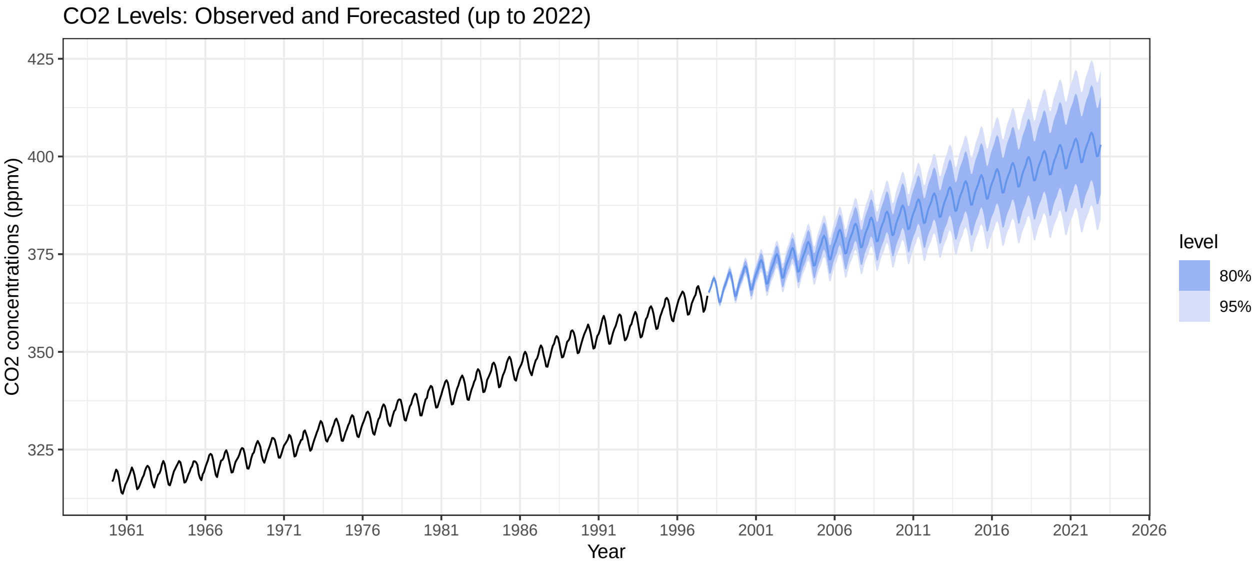
Both ACF and PACF plots show strong auto-correlation with lag 1, and the ACF cutting off strong after lag 1, and the PACF having a significant spike at lag 1, and tapering off a little more. This might indicate that our model has a MA(1) component. The spike at lag 12 in the ACF might also indicate a seasonal MA component.

The ARIMA function behaved as expected, and returned an ARIMA(0,1,1)$(0,1,1)_{12}$ function, with BIC = 182.32. We will now look at the residuals for this model.

```
## Plot variable not specified, automatically selected `.vars = .resid`
```



The residuals look random, there are no significant autocorrelations in the ACF, and they closely follow a a normal distribution in the Q-Q plot. These all indicate that the model has captured most of the underlying structure of our time series. A Ljung-Box test also yielded a p-value of 0.6733, further confirming that there is insufficient statistical evidence to reject the null hypothesis that there is no autocorrelation. We can proceed with forecasting the time series data through 2022.



1.5 (3 points) Task 4a: Forecast atmospheric CO2 growth

We now forecast when atmospheric CO2 is expected to be at 420 ppm and 500 ppm for the first and final times, as seen in the table below. \begin{table}

\caption{(#tab:forecast 420 500)CO2 Levels and Forecasted Times with 80% Confidence Intervals}

CO2.Level	First.Month	First.Value	Last.Month	Last.Value
420 ppm	2031 May	420.1 (402.3, 438.0)	2035 Oct	420.4 (399.5, 441.3)
500 ppm	2083 Apr	500.4 (437.9,562.9)	2085 Dec	500.9 (435.7,566.2)

\end{table}

Our model also forecasts CO2 levels in the year 2100. Although these forecasts include a standard deviation, but these do not take into account the existing efforts to reduce global greenhouse gases, such as **insert efforts here**. Thus, since these are very human-activty dependent, it is unlikely to be super accurate.

Table 1: (#tab:forecast 2100)CO2 Forecasts in 2100

Date	Value	Standard_Deviation
2100 Jan	523.7	62.8
2100 Feb	524.6	62.9
2100 Mar	525.5	63.0
2100 Apr	526.8	63.0
2100 May	527.4	63.1
2100 Jun	526.7	63.2
2100 Jul	525.2	63.3
2100 Aug	523.1	63.3
2100 Sep	521.3	63.4
2100 Oct	521.4	63.5
2100 Nov	522.8	63.5
2100 Dec	524.2	63.6

2 Report from the Point of View of the Present

One of the very interesting features of Keeling and colleagues' research is that they were able to evaluate, and re-evaluate the data as new series of measurements were released. This permitted the evaluation of previous models' performance and a much more difficult question: If their models' predictions were "off" was this the result of a failure of the model, or a change in the system?

2.1 (1 point) Task 0b: Introduction

In this introduction, you can assume that your reader will have **just** read your 1997 report. In this introduction, **very** briefly pose the question that you are evaluating, and describe what (if anything) has changed in the data generating process between 1997 and the present.

2.2 (3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data.

The most current data is provided by the United States' National Oceanic and Atmospheric Administration, on a data page [here]. Gather the most recent weekly data from this page. (A group that is interested in even more data management might choose to work with the hourly data.)

Create a data pipeline that starts by reading from the appropriate URL, and ends by saving an object called `co2_present` that is a suitable time series object.

Conduct the same EDA on this data. Describe how the Keeling Curve evolved from 1997 to the present, noting where the series seems to be following similar trends to the series that you "evaluated in 1997" and where the series seems to be following different trends. This EDA can use the same, or very similar tools and views as you provided in your 1997 report.

2.3 (1 point) Task 2b: Compare linear model forecasts against realized CO2

Descriptively compare realized atmospheric CO2 levels to those predicted by your forecast from a linear time model in 1997 (i.e. "Task 2a"). (You do not need to run any formal tests for this task.)

2.4 (1 point) Task 3b: Compare ARIMA models forecasts against realized CO2

Descriptively compare realized atmospheric CO2 levels to those predicted by your forecast from the ARIMA model that you fitted in 1997 (i.e. "Task 3a"). Describe how the Keeling Curve evolved from 1997 to the present.

2.5 (3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models

In 1997 you made predictions about the first time that CO2 would cross 420 ppm. How close were your models to the truth?

After reflecting on your performance on this threshold-prediction task, continue to use the weekly data to generate a month-average series from 1997 to the present, and compare the overall forecasting performance of your models from Parts 2a and 3b over the entire period. (You should conduct formal tests for this task.)

2.6 (4 points) Task 5b: Train best models on present data

Seasonally adjust the weekly NOAA data, and split both seasonally-adjusted (SA) and non-seasonally-adjusted (NSA) series into training and test sets, using the last two years of observations as the test sets. For both SA and NSA series, fit ARIMA models using all appropriate steps. Measure and discuss how your models perform in-sample and (psuedo-) out-of-sample, comparing candidate models and explaining your choice. In addition, fit a polynomial time-trend model to the seasonally-adjusted series and compare its performance to that of your ARIMA model.

2.7 (3 points) Task Part 6b: How bad could it get?

With the non-seasonally adjusted data series, generate predictions for when atmospheric CO2 is expected to be at 420 ppm and 500 ppm levels for the first and final times (consider prediction intervals as well as point estimates in your answer). Generate a prediction for atmospheric CO2 levels in the year 2122. How confident are you that these will be accurate predictions?