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.

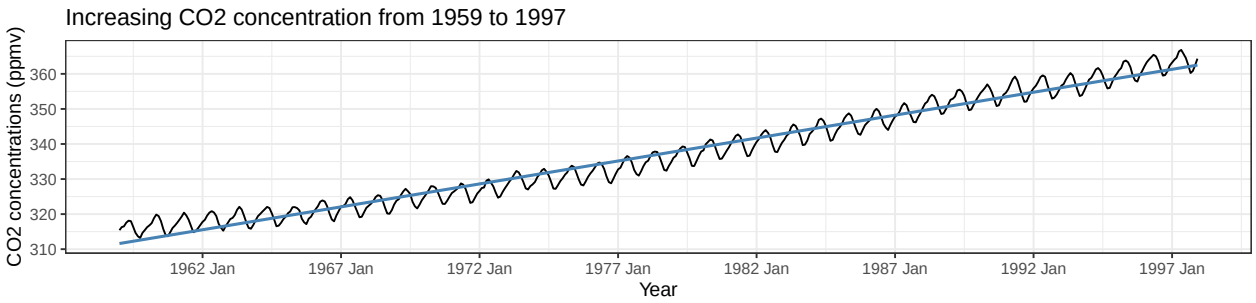
Geochemist Dr. Charles David Keeling's pioneering work in atmospheric carbon dioxide measurements fundamentally reshaped our understanding of the global carbon cycle and its impact on climate change. In 1958, Keeling initiated a long-term study at the Mauna Loa Observatory, producing the iconic "Keeling 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.

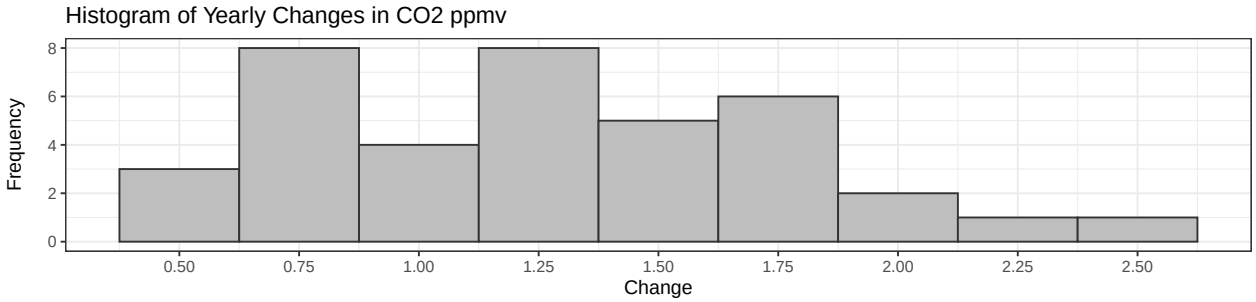
CO2 is classified as a "greenhouse gas," meaning that it traps heat in the atmosphere and lead to rising global temperatures when in high concentrations. It can be important to track Co2 levels as rising global temperatures can lead to imbalances in ecosystems and rising water levels that impact both 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

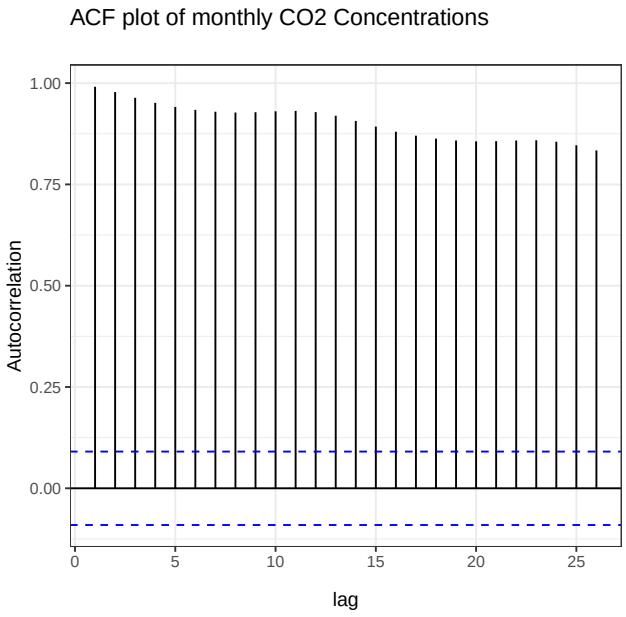
The current data is gathered from measurements made under Dr. Charles Keeling's study at the Mauna Loa Observatory in Hawaii (Cleaveland, 1993). Measurements were taken by a chemical gas analyzer sensor, with detections based on infrared absorption. This data measures monthly CO2 concentration levels 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).

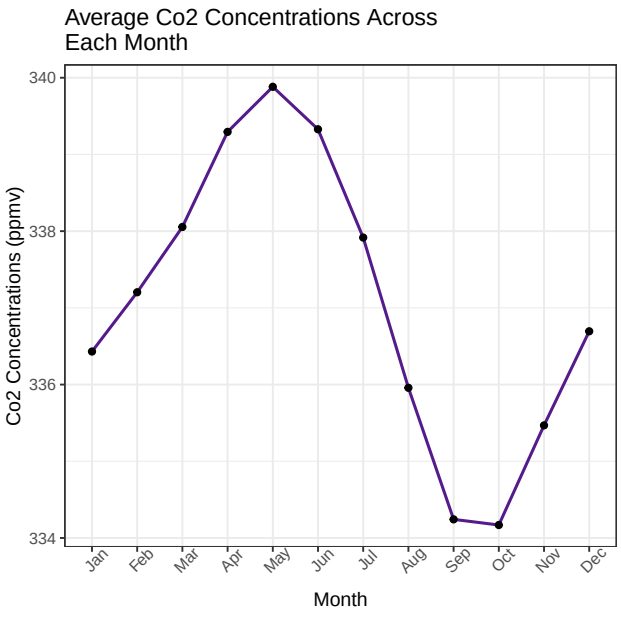
The time series shows a clear upward trend of global CO2 concentrations from 1959 to 1998, with an average increase in 1.26 CO2 ppmv and a standard deviation of .51 CO2 ppmv. Upon inspection of the yearly increases, the bulk of changing CO2 levels are between 0.5 and 2.0 CO2 ppmv.





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).

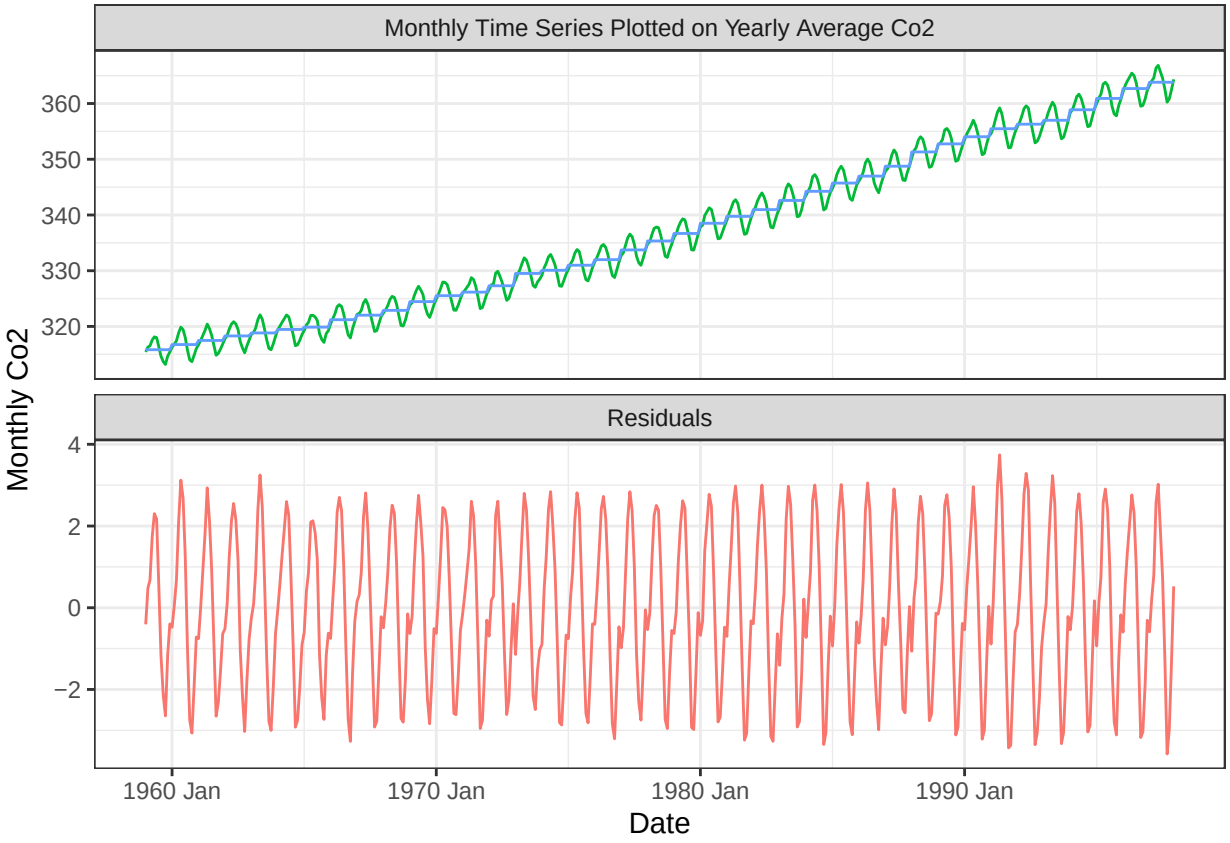
We now study the variance over time. While the average CO2 ppmv is consistently higher each year, We fit a yearly CO2 average on the monthly time series, and inspect the residuals from year to year. Although there are slight changes in the variance, they regress to a constant variance over time, according to a significant Augmented Dickey–Fuller Test, which suggests stationarity in variance. Thus, once we detrend the series by accounting for the yearly increases in CO2 ppmv, there is likely a constant variance over time, as well as a

Table 1: (#tab:seasonality irregularities)ADF Test Results

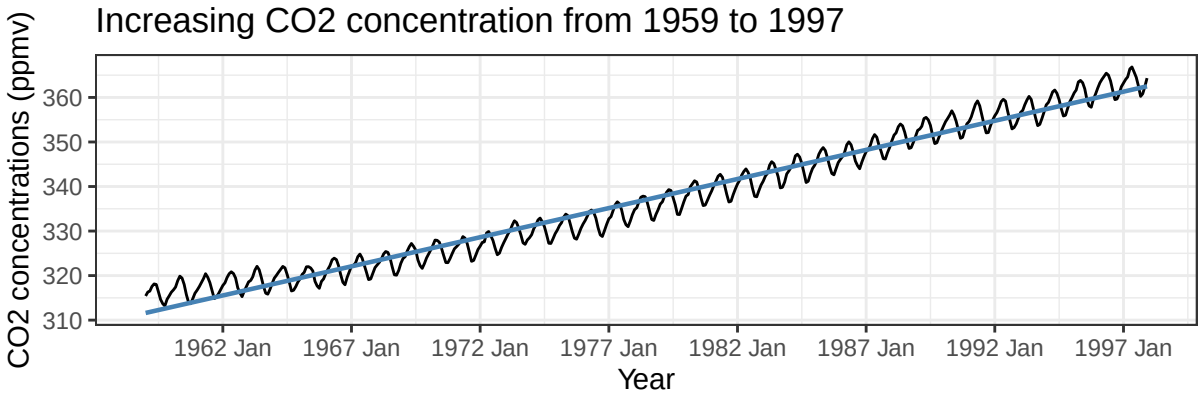
	Statistic	P_Value	Method	Alternative
Dickey-Fuller	-29.06892	0.01	Augmented Dickey-Fuller Test	stationary

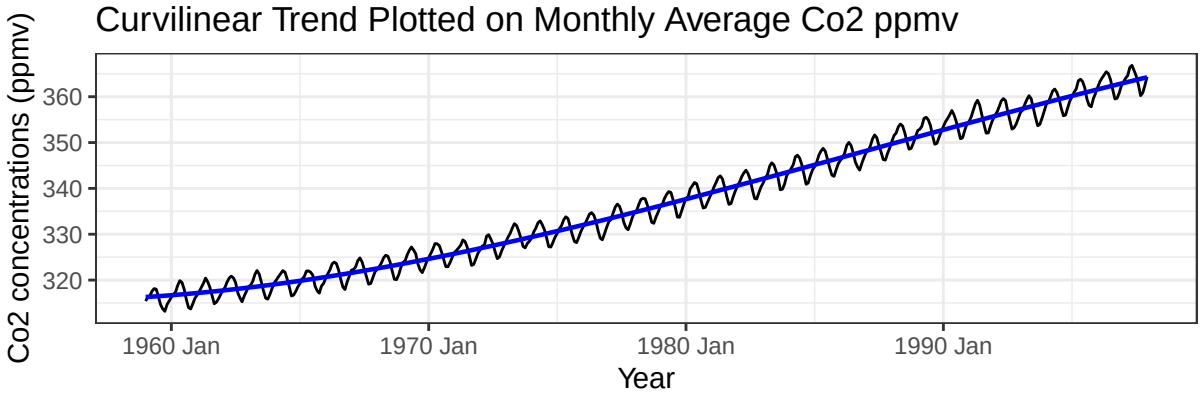
non-moving average.

```
## Warning in adf.test(.): p-value smaller than printed p-value
```



Variability in the yearly trend is also observed through plotting. While the average change in Co2 ppmv is 1.26, this increase varies per year, with some years experiencing heavier spikes in increasing Co2 concentrations than others. Additionally, upon further inspection of the monthly average trend plot above, the fitted line appears to be systematically overestimating values at certain points and underestimating values at other points. When fitting a curvilinear trend to the time series, we see a closer fit the central Co2 ppmv for each year

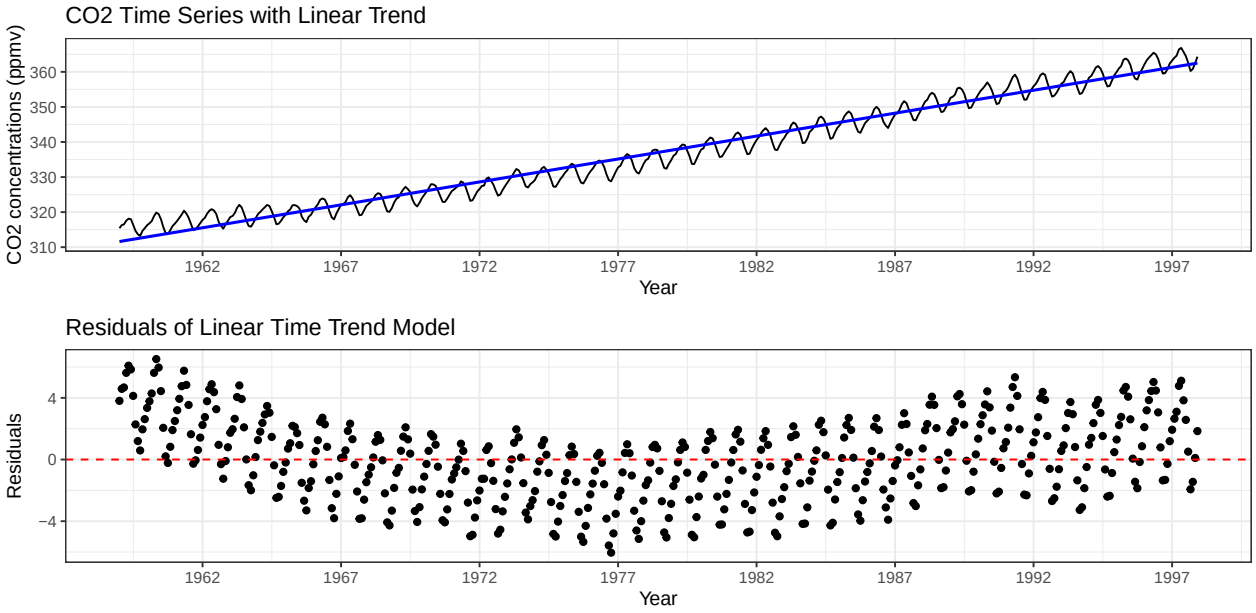




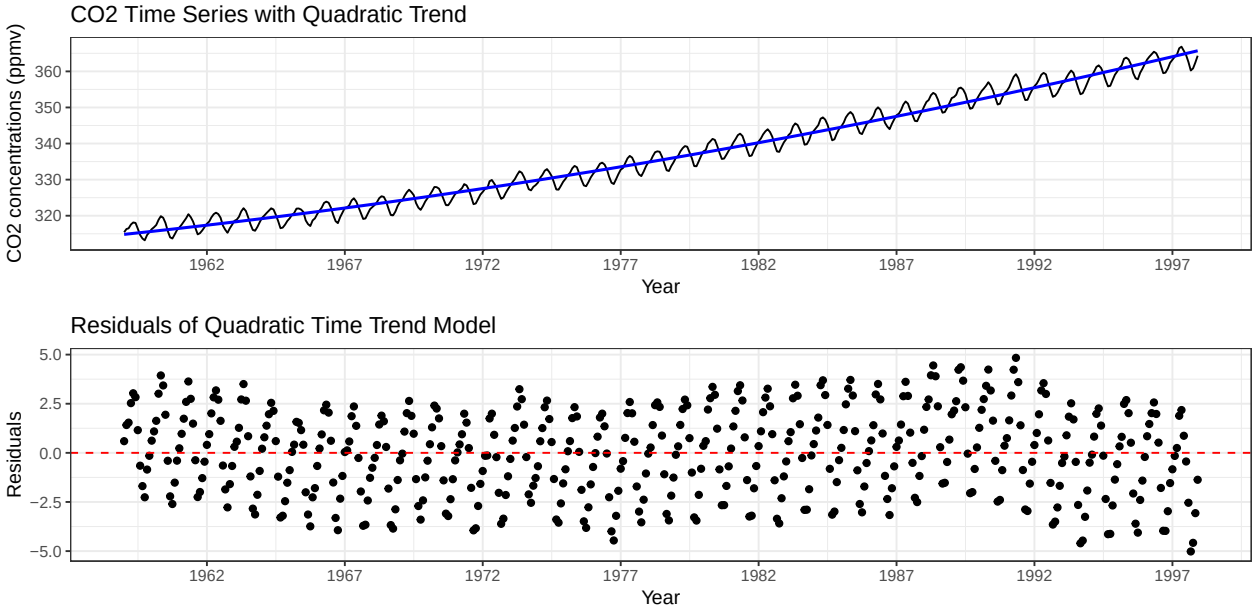
1.3 (3 points) Task 2a: Linear time trend model

Fit a linear time trend model to the `co2` series, and examine the characteristics of the residuals. Compare this to a quadratic time trend model. Discuss whether a logarithmic transformation of the data would be appropriate. Fit a polynomial time trend model that incorporates seasonal dummy variables, and use this model to generate forecasts to the year 2020.

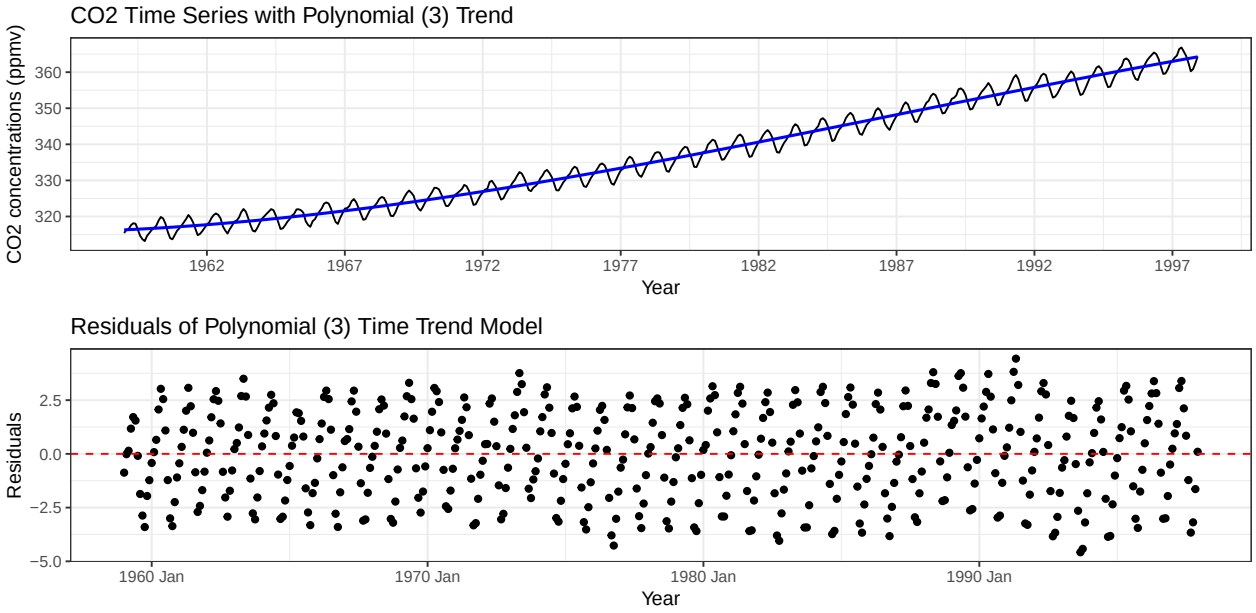
We now fit a linear time trend model the series, and examine the characteristics of the residuals.



The residuals of the linear model 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

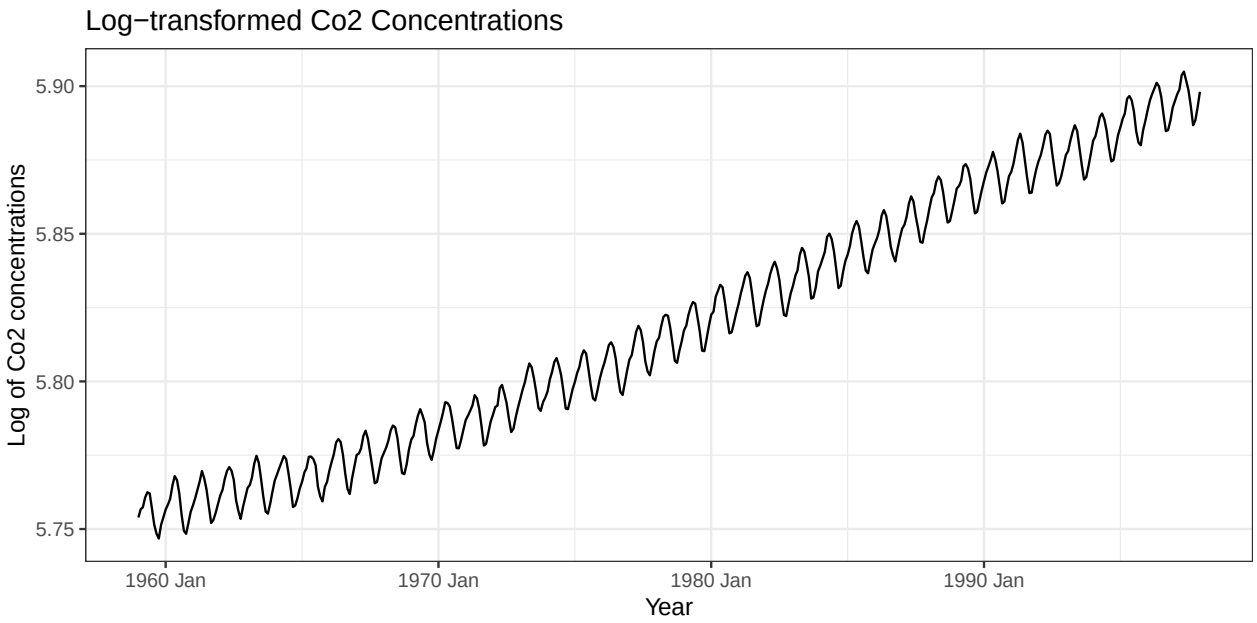


The quadratic model's residuals indicate a small reduction in variance, demonstrating a slightly improved fit. However, the cyclical behavior remains, indicating that seasonality is unaccounted for in the model still. There is also an overall non-random trend in the residuals, indicating that the model 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.



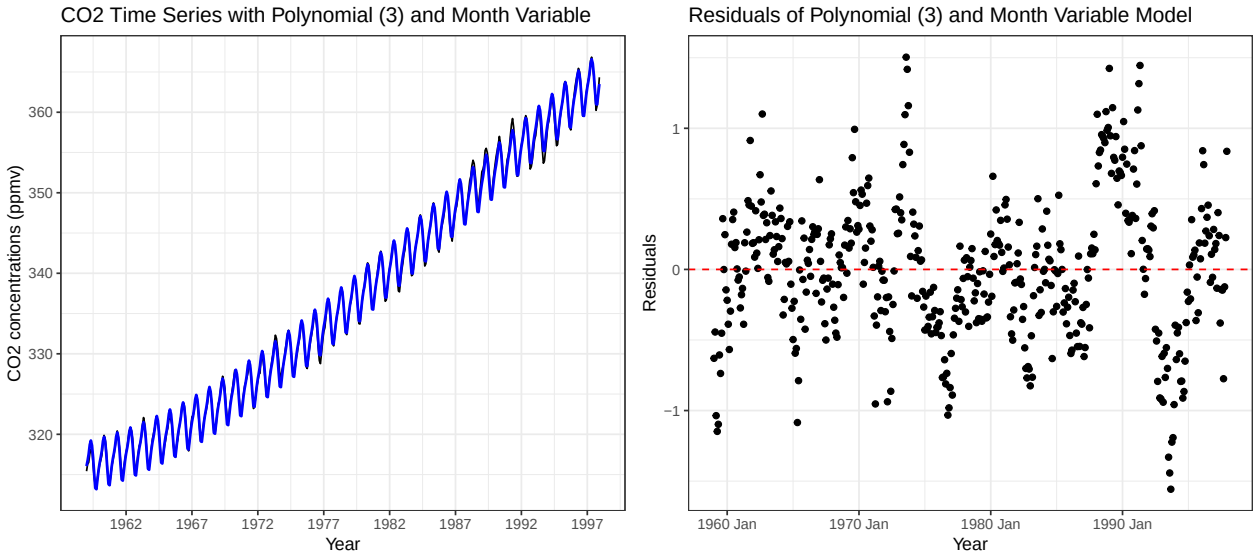
The third-order polynomial model demonstrates improved residual behavior compared to quadratic and linear models. We chose to stop at this order to prevent excessive overfitting, as higher-order polynomials showed diminishing returns in model performance.

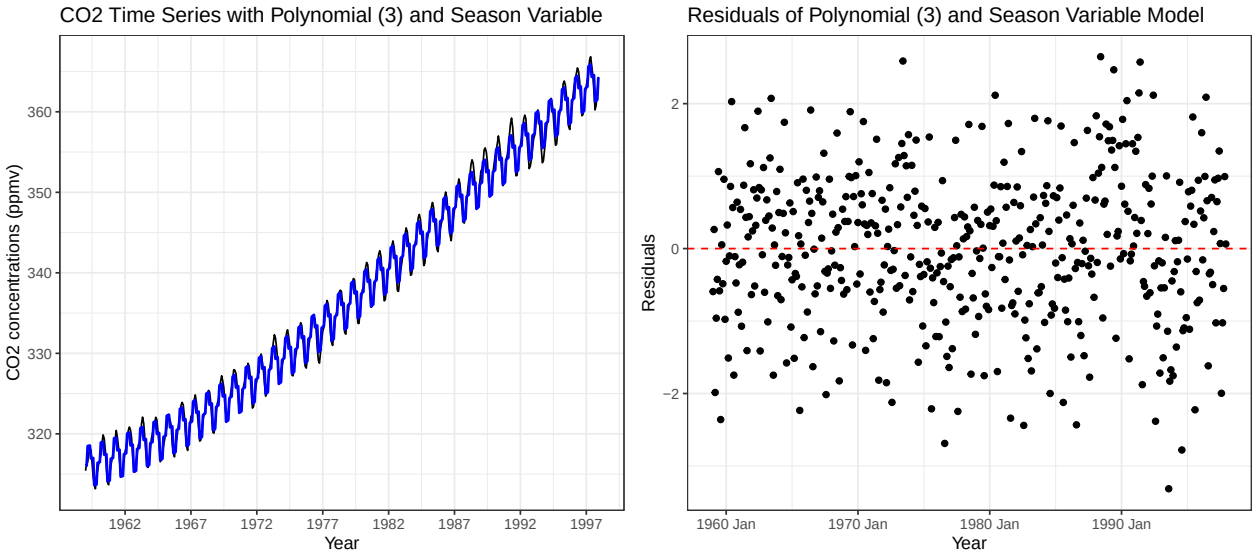
Apart from transforming the orders of the model, we were interested in data transformations - specifically logarithmic. As such we expermiented with a logarithmic dataset to observe the pattern of the data values.



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.

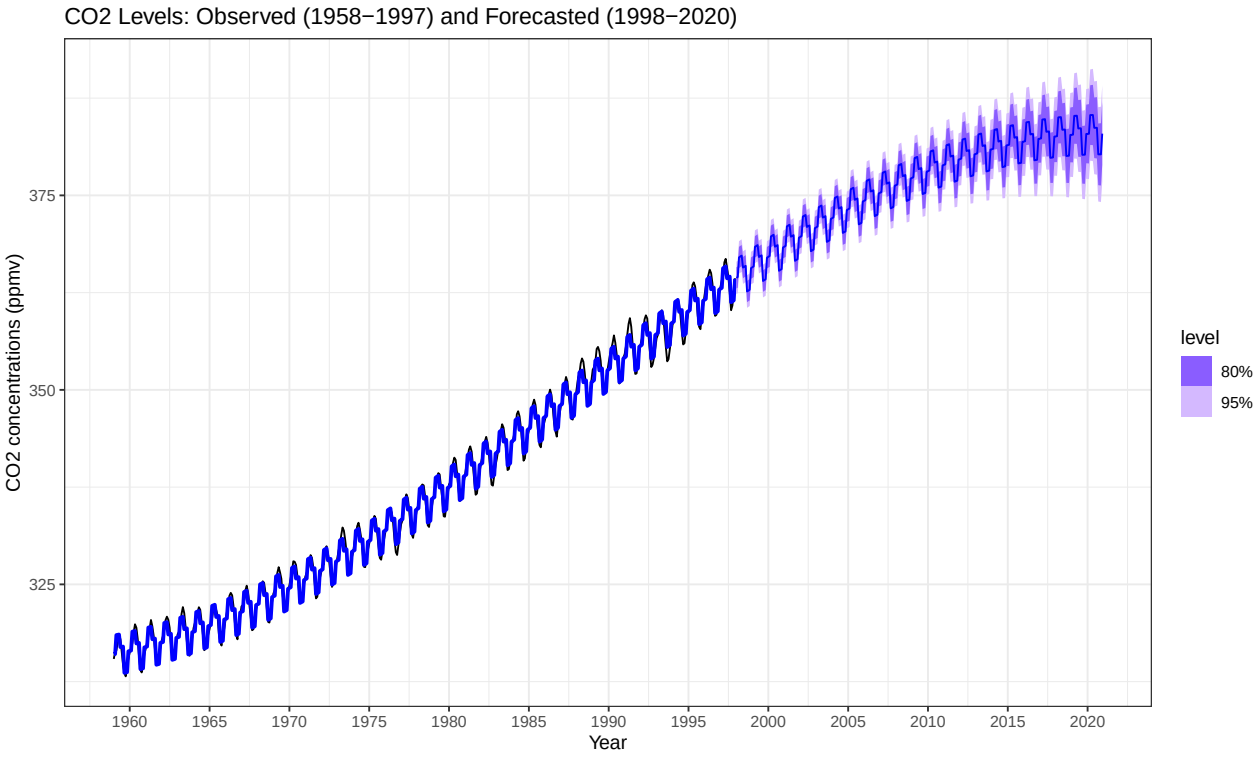
To address the cyclical behavior, we developed another polynomial model that includes each month as a variable. The average monthly CO2 emissions indicate significant cyclic patterns at the monthly level. By incorporating this variable, we anticipate an improvement in the fit of our time series model.





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, This adjustment centered the residuals around zero with a random distribution, though fluctuations remained between 2 and -2. We proceeded with this model, considering it the most robust.

Using the polynomial model with the `season` dummy variable, we then developed a forecast for CO2 emissions through 2020.



The forecast model using the `season` variable shows very poor performance, with future predictions deviating significantly from expected values, likely due to a misfitting model or improper scaling of the forecasted data. To resolve this, we turn to an ARIMA model, which may better capture the time series' underlying patterns

and improve forecast accuracy.

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

Following all appropriate steps, choose an ARIMA model to fit to the series. Discuss the characteristics of your model and how you selected between alternative ARIMA specifications. Use your model (or models) to generate forecasts to the year 2022.

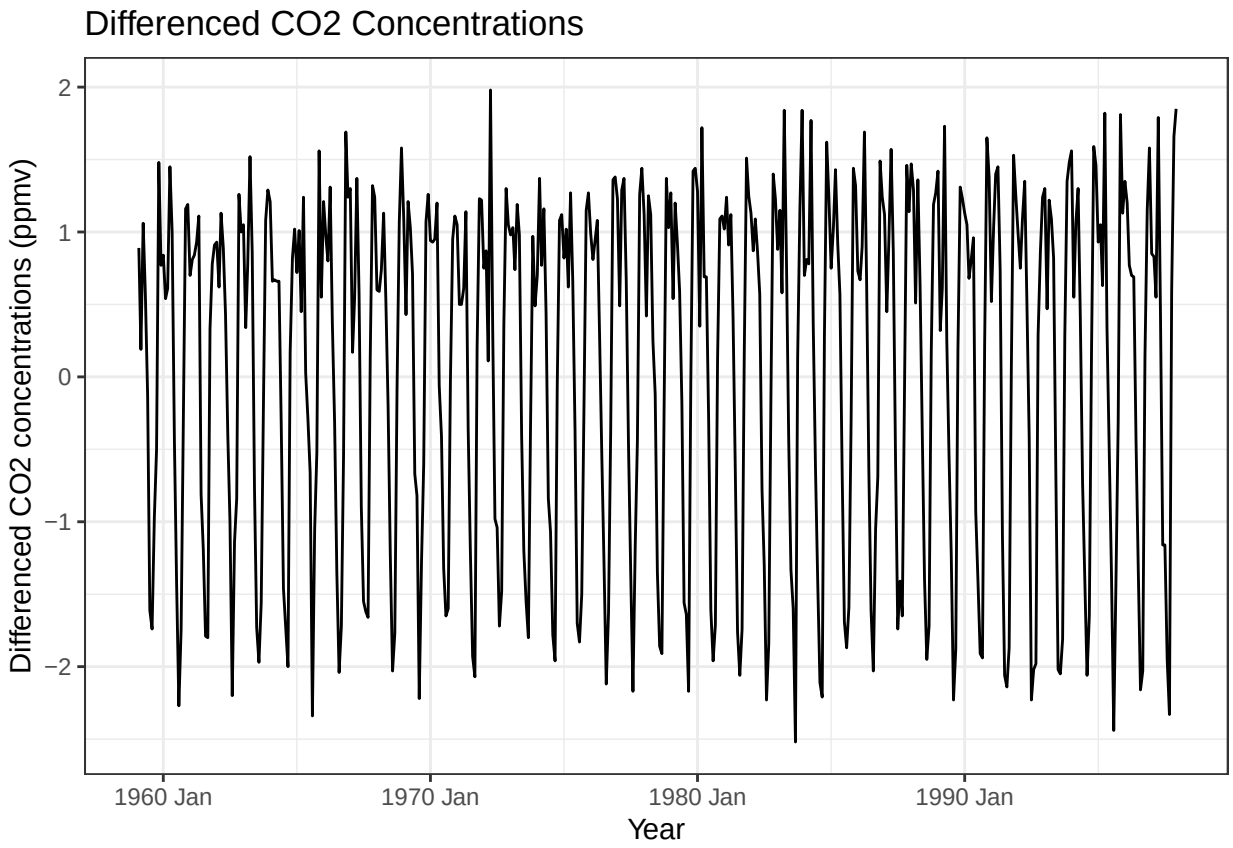
```
# checking for stationarity
adf.test(co2_tsib$value)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  co2_tsib$value
## Dickey-Fuller = -2.8299, Lag order = 7, p-value = 0.2269
## alternative hypothesis: stationary
```

With a p-value of 0.2269, the time series data fails to reject the null hypothesis of non-stationarity. As a result, we will proceed with differencing the data to make it stationary, which is a crucial step before fitting the ARIMA model effectively.

```
# First differencing the series to remove trend
# Added NA to ensure the data has equivalent numbers of rows.
co2_tsib$diff_value <- c(NA,diff(co2_tsib$value, differences = 1))

# Plot the differenced series to check if it looks stationary
ggplot(co2_tsib, aes(x = index, y = diff_value)) +
  geom_line() +
  labs(title = "Differenced CO2 Concentrations", x = "Year", y = "Differenced CO2 concentrations (ppmv)
```




```
# post-check for stationarity
# remove NA for adf test
adf.test(na.omit(co2_tsib$diff_value))
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  na.omit(co2_tsib$diff_value)
## Dickey-Fuller = -30.38, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

With first differencing successfully making the data stationary, we can now proceed with constructing the ARIMA (p,d,q) model using a d value of 1. The next step will involve fine-tuning the p and q parameters to further optimize the model for accurate forecasting.

```
# Fit ARIMA model by testing different lags using the AIC criterion
model.aic <- co2_tsib %>%
  model(ARIMA(value ~ 1 + pdq(0:10, 1, 0:10) + PDQ(0:2, 0, 0:2),
              ic = "aic", stepwise = FALSE))

# Report the best model
model.aic %>%
  report()
```

```
## Series: value
## Model: ARIMA(3,1,1)(0,0,2)[12] w/ drift
##
## Coefficients:
##          ar1      ar2      ar3      ma1    sma1    sma2  constant
```

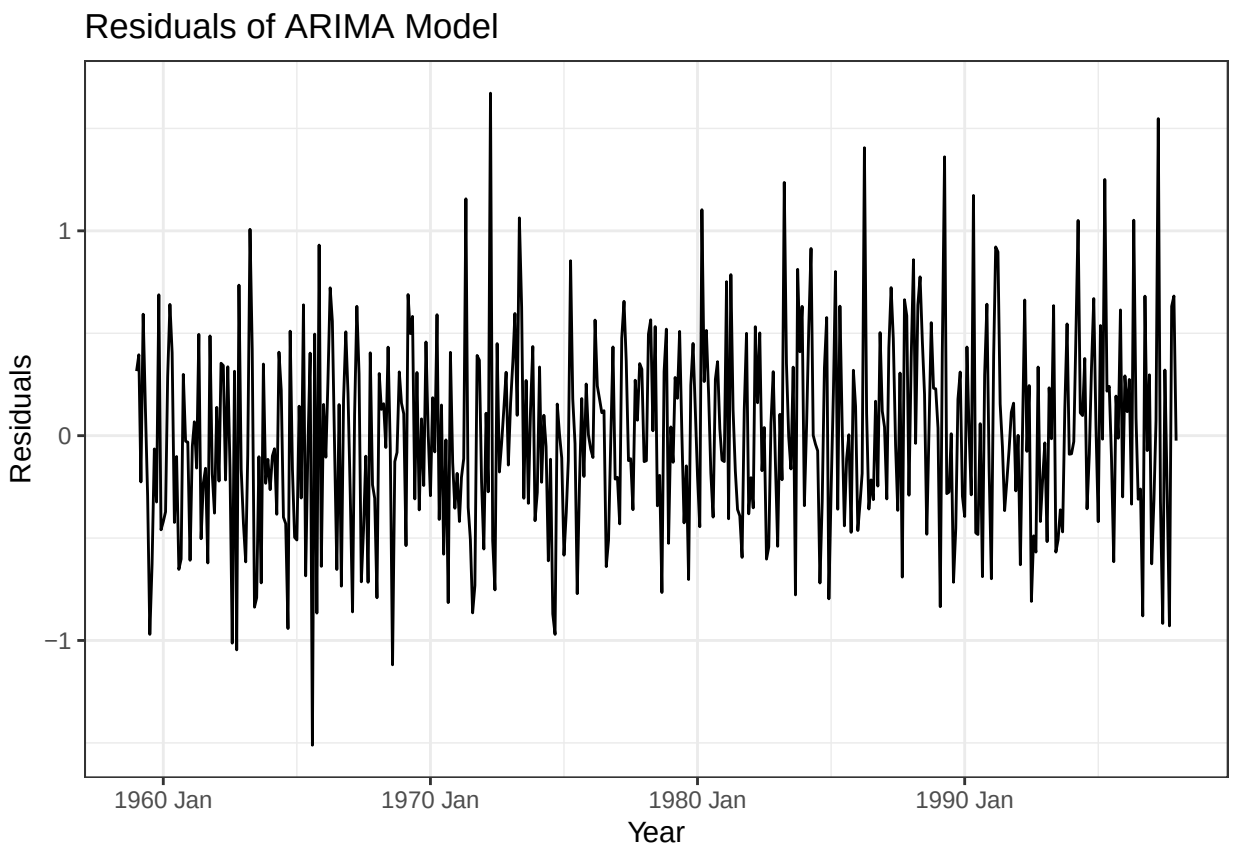
```
##       1.1159  -0.1776  -0.3450  -0.9252  0.6573  0.3792    0.0427
## s.e.  0.0497   0.0738   0.0456   0.0167  0.0506  0.0417    0.0034
##
## sigma^2 estimated as 0.2311:  log likelihood=-321.64
## AIC=659.29   AICc=659.6   BIC=692.46
```

```r
# Extract residuals from the ARIMA model
residuals_arima <- residuals(model.aic)

# Plot the residuals over time
residual_plot <- autoplot(residuals_arima) +
  labs(title = "Residuals of ARIMA Model", x = "Year", y = "Residuals")
```

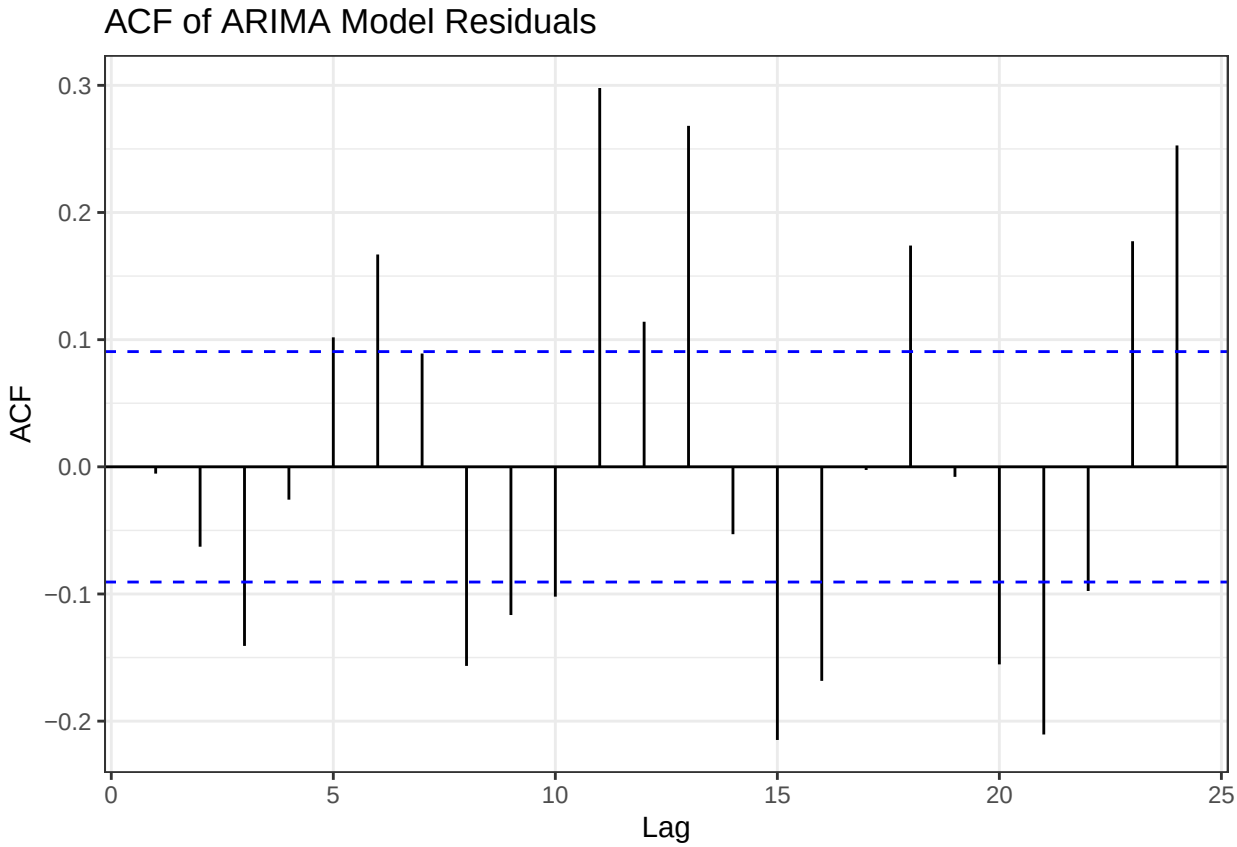
```
## Plot variable not specified, automatically selected `.vars = .resid`
```

```r
# Display the plot
residual_plot
```



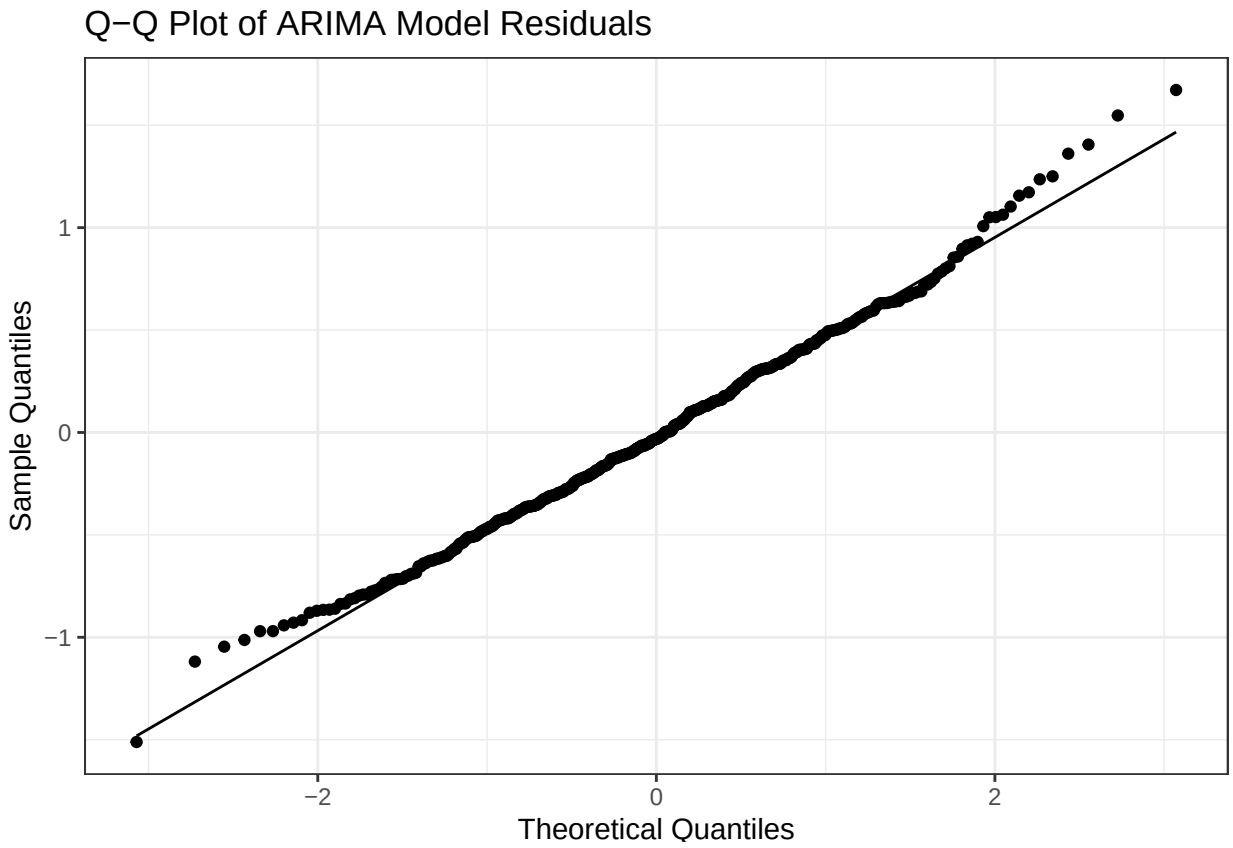

```r
# Plot the ACF of the residuals
acf_plot <- ggAcf(residuals_arima) +
  labs(title = "ACF of ARIMA Model Residuals", x = "Lag", y = "ACF")

# Display the ACF plot
print(acf_plot)
```




```r
# Q-Q plot to check normality of residuals
qq_plot <- ggplot(data = as.data.frame(residuals_arima), aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q Plot of ARIMA Model Residuals", x = "Theoretical Quantiles", y = "Sample Quantiles")

# Display the Q-Q plot
print(qq_plot)
```




```
# Ljung Box Test on residuals
# Calculate the number of observations
N <- nrow(co2_tsib)

# Determine the number of lags (using rule of thumb: sqrt(N))
lags <- floor(sqrt(N))

# get residuals
resid.ts<-model.aic %>%
  augment() %>%
  select(.resid) %>%
  as.ts()

Box.test(resid.ts, lag = lags, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  resid.ts
## X-squared = 226.59, df = 21, p-value < 2.2e-16
```

```
# # Fit ARIMA model by testing different lags using the AICc criterion
# model.aicc <- co2_tsib %>%
#   model(ARIMA(value ~ 1 + pdq(0:10, 1, 0:10) + PDQ(0:2, 0, 0:2),
#               ic = "aicc", stepwise = FALSE))
#
# # Report the best model
# model.aicc %>%
```

```
#   report()
#
# # Fit ARIMA model by testing different lags using the BIC criterion
# model.bic <- co2_tsib %>%
#   model(ARIMA(value ~ 1 + pdq(0:10, 1, 0:10) + PDQ(0:2, 0, 0:2),
#               ic = "bic", stepwise = FALSE))
#
# # Report the best model
# model.bic %>%
#   report()
```

The AIC analysis identified the optimal ARIMA model parameters as p=3, d=1, and q=1. The residuals are centered around zero with random fluctuations, indicating the model has captured most of the underlying structure. However, the ACF plot shows significant spikes at lags 10 and 24, indicating some remaining autocorrelation, and the Ljung-Box test confirms this with a p-value of $< 2.2e\text{-}16$, suggesting the need for additional terms to improve the model fit. Despite this, the QQ plot indicates that the residuals follow normality well. With these considerations, we can proceed with forecasting the time series data through 2022.

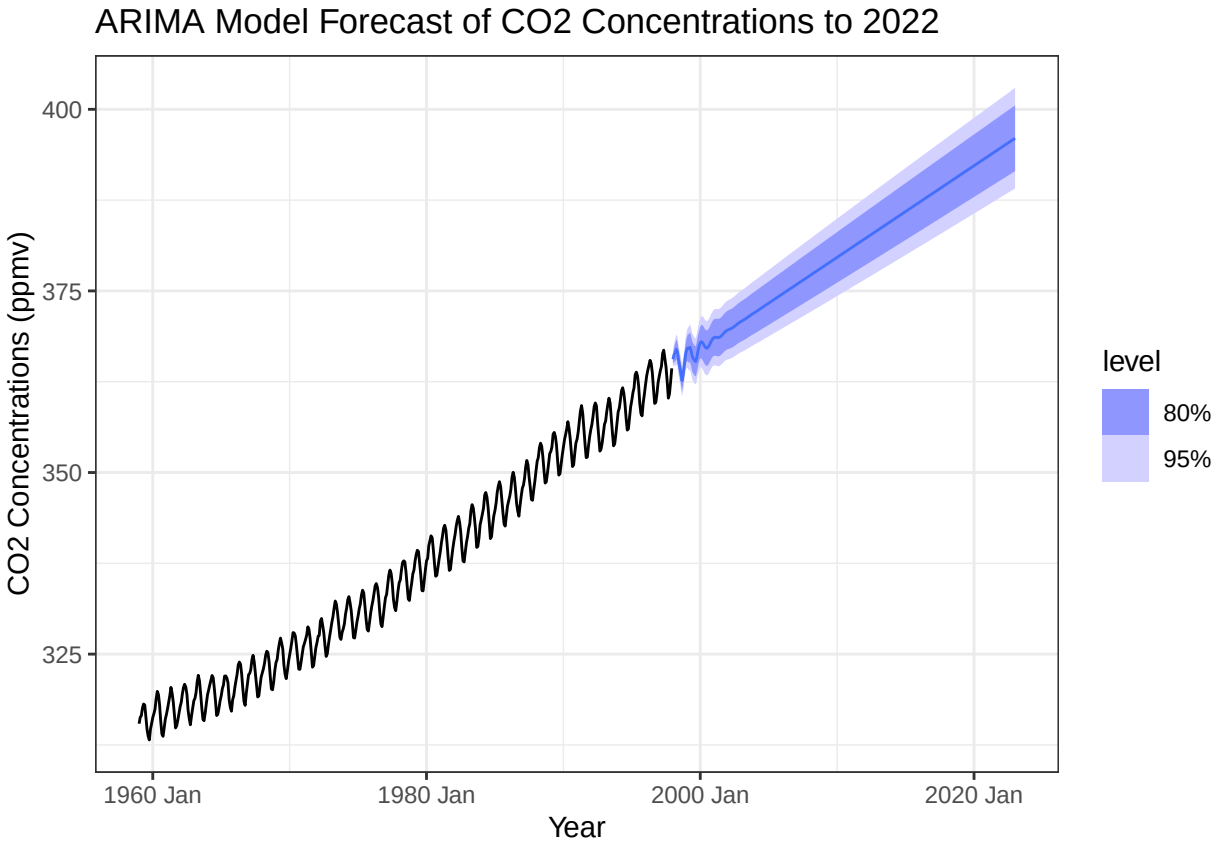
```
# Determine how many months to forecast (from the last observation to Dec 2022)
last_observation <- max(co2_tsib$index)
end_of_2022 <- as.Date("2022-12-31")

# Calculate the number of months to forecast
horizon <- as.numeric(
  difftime(end_of_2022, last_observation, units = "weeks")) / 4.34524  # Convert weeks to months

# Forecast using the ARIMA model until 2022
forecast_arima <- model.aic %>%
  forecast(h = round(horizon))  # Use calculated horizon

# Plot the forecast
forecast_plot <- forecast_arima %>%
  autoplot(co2_tsib) +
  labs(title = "ARIMA Model Forecast of CO2 Concentrations to 2022",
       x = "Year", y = "CO2 Concentrations (ppmv)")

# Show the plot
print(forecast_plot)
```



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

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 2100. How confident are you that these will be accurate predictions?

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?