

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

Investigating the Keeling Curve and forecasting CO2 levels in Earth's atmosphere

Denny Lehman, Mingxi Liu, Aruna Bisht, Deepika Maddali

Contents

1	The Keeling Curve	1
2	Data	1
3	Linear model	3
3.1	ARIMA times series model	5
3.2	Atmospheric CO2 growth Forecast	8
A	Appendix: Model Robustness	11

1 The Keeling Curve

In the 1950s, the geochemist Charles David Keeling observed a seasonal pattern in the amount of carbon dioxide present in air samples collected over the course of several years. He was able to attribute this pattern to the variation in global rates of photosynthesis throughout the year, caused by the difference in land area and vegetation cover between the Earth's northern and southern hemispheres.

In 1958 Keeling began continuous monitoring of atmospheric carbon dioxide concentrations from the Mauna Loa Observatory in Hawaii and soon observed a trend increase carbon dioxide levels in addition to the seasonal cycle. He was able to attribute this trend increase to growth in global rates of fossil fuel combustion. This trend has continued to the present, and is known as the "Keeling Curve."

2 Data

The data measures the monthly average atmospheric CO2 concentration from 1959 to 1997, expressed in parts per million (ppm). It was initially collected by an infrared gas analyzer installed at Mauna Loa in Hawaii, which was one of the four analyzers installed by Keeling to evaluate whether there was a persistent increase in CO2 concentration.

Fig.1 shows a clear long-term upward trend, which is confirmed by Fig.2 where the growth rate for each year is above zero. Fig.2 also suggests the average growth rate after 1970 is higher than that before 1970, although there's no evidence of accelerating growth. The ACF plots in Fig.3 and Fig.4

suggest the original data is non-stationary but its first difference is stationary. More formally, the KPSS tests below confirm the observations above.

Table 1: KPSS test of original and 1st difference

	kpss_stat	kpss_pvalue
original	7.8173	0.01
1st_difference	0.0124	0.10

Another feature of the data is its robust seasonal pattern, with the peak in May and the bottom in October almost every year (see Fig.5). This seasonality can also be seen in Fig.4. Keeling believes it was the result of plant photosynthesis.

Fig.4 is the histogram of the remaining or irregular components after removing the trend and the seasonal components from the data with STL¹. It looks like a normal distribution without obvious outliers.

¹Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3–33.

Fig.1 Atmospheric CO2 concentration monthly average, parts per million (ppm)

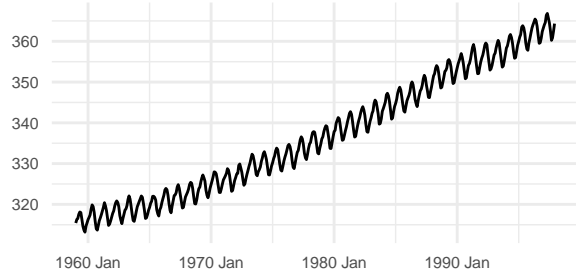


Fig.2 Annual growth rate of concentration, %

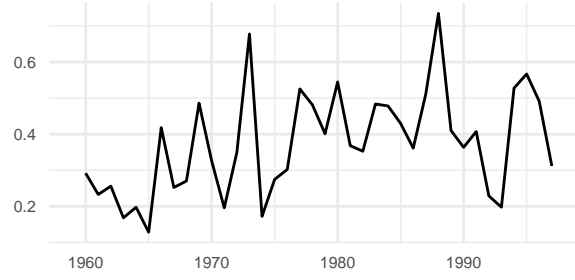


Fig.3 ACF of CO2 concentration

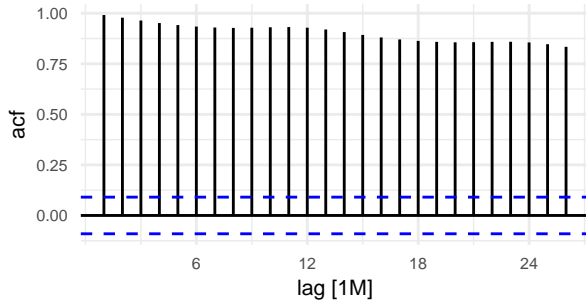


Fig.4 ACF of differenced CO2 concentration

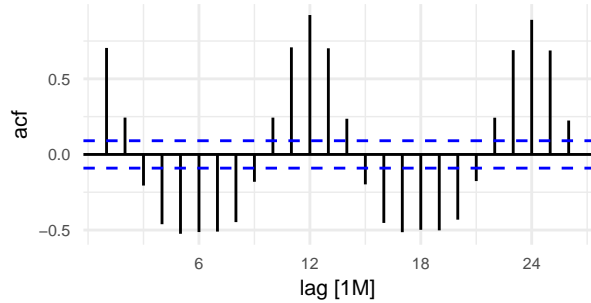


Fig.5 Seasonal plot of CO2 concentration

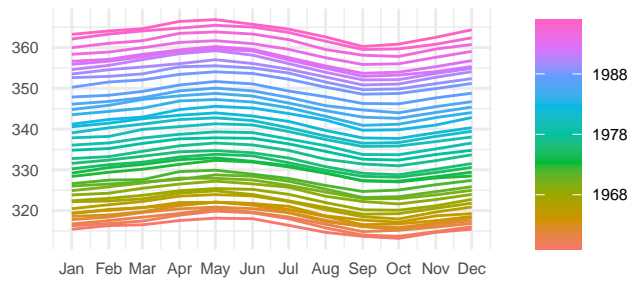
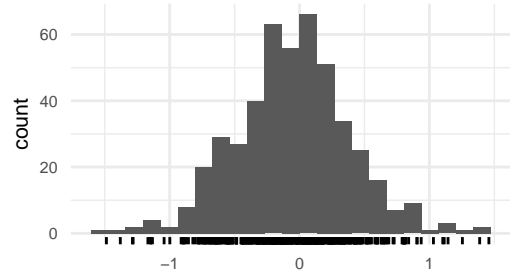


Fig.6 Histogram of irregular component by STL



3 Linear model

Before building the model, we need to consider whether the data need a log transformation. Normally a log transformation is needed when the data shows exponential growth or the variance expands or shrinks over time. From Fig.1 and Fig.2 we can see the slope or the growth rate of the data is stable, which suggests the growth is more close to linear instead of exponential. Also, Fig.5 shows the difference between the annual high and the annual low almost remained the same over the years, suggesting the variance is nearly constant. Therefore, the log transformation is not necessary. We can first fit the original data with a linear time trend model as:

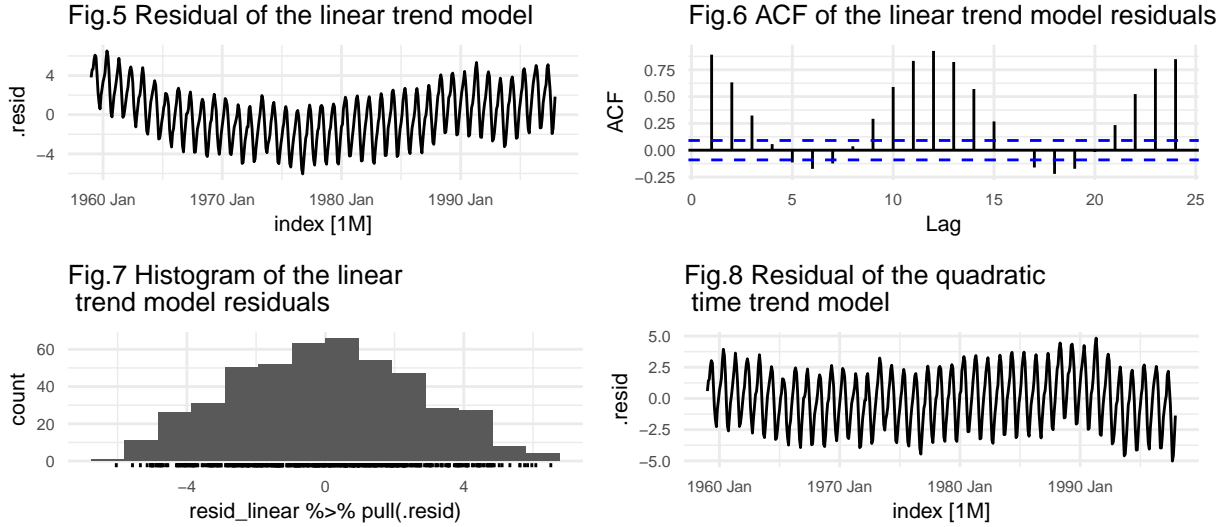
$$CO_2 = \beta_0 + \beta_1 t + \epsilon_t, \quad (1)$$

which gives the parameters as:

$$CO_2 = 311.5 + 0.11t + \epsilon_t \quad (2)$$

This linear trend model implies that the CO_2 concentration increased by 0.11/month on average from 1959 to 1997. However, the residual plots in Fig.5 to Fig.7 suggest this simple linear trend model is not adequate in the following two aspects.

First, the mean of the residual forms a “U” shape over time, suggesting a quadratic or higher-order polynomial time trend model may be more appropriate. For instance, the residual from a quadratic time trend model shows a more constant mean over time, as shown in Fig.8.



In addition, Fig.6, the ACF plot, indicates strong seasonal patterns exist in the residuals, suggesting we should consider seasonal factors in the model, and one solution is to include 11 dummy variables in the model to indicate the 12 months.

Based on the two points above, we compare the 2 candidates: a quadratic time trend model and a cubic one, as below.

$$\text{Quadratic time trend: } CO_2 = \alpha + \beta_0 t + \beta_1 t^2 + \sum_{i=1}^{11} \gamma_i Month_{it} + \epsilon_t \quad (3)$$

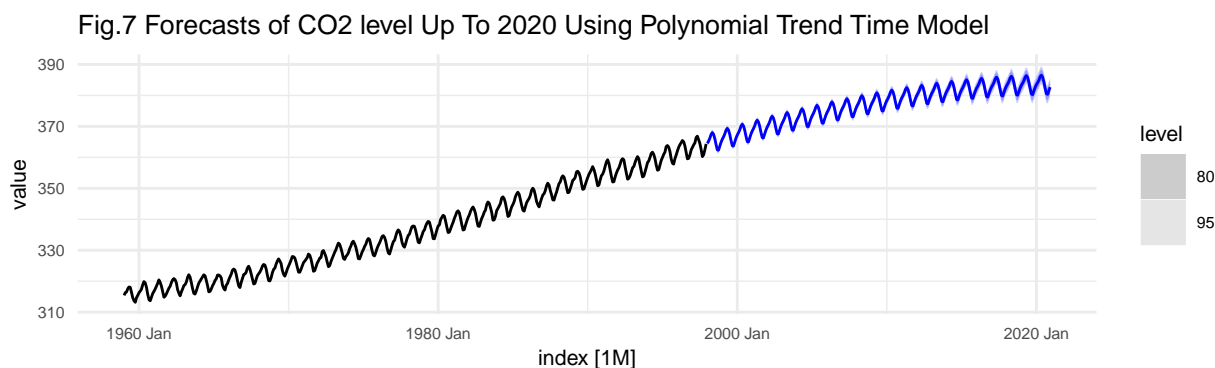
$$\text{Cubic time trend: } CO_2 = \alpha + \beta_0 t + \beta_1 t^2 + \beta_2 t^3 + \sum_{i=1}^{11} \gamma_i Month_{it} + \epsilon_t \quad (4)$$

We use the data before 1991 as the training set and the rest as the validation set (similar to an 80-20 split). Our final choice of the model depends on the combination of 2 guidelines: 1) the information criterion (AICc, BIC) from the model fitting process and 2) the root mean square error (RMSE) of predictions on the validation set, which are listed in Table.1. Both information criterion (AICc, BIC) and RMSE favor the cubic model. Therefore, the cubic time trend model becomes our final choice. Its details are in the Appendix. We plot the forecast of this model till 2020 in Fig.7. One thing to note is that because the coefficient of the cubic term is negative, the predicted values will eventually begin to decrease when predicting the far future. In fact, we can see from Fig.7 that the

predicted values have almost topped. If it doesn't make sense, we should confine our predicting interval to the near term.

Table 2: Information Criterion of model fitting and RMSE of validation

.model	AIC	AICc	BIC	RMSE
cubic	-659.6147	-658.1324	-596.4044	2.112194
quadratic	-639.1525	-637.8481	-579.8928	2.796572



3.1 ARIMA times series model

We will use the Box Jenkins process to find the best ARIMA model via the following steps:

- Determine the appropriate model from EDA
- Find the best parameters
- Examine the residuals using diagnostic plots and statistical tests

The EDA revealed that the time series of CO2 had both autoregressive and seasonal components. Considering the ACF plot's low slow decay of autocorrelation, we expect differencing to be a key part of any time series model. In addition, we predict that the model will require seasonal components to model the 12 month cycle of seasonal variations. Therefore, we expect a seasonal arima model (SARIMA) with differencing to be best.

In this section, we fit the best SARIMA model and analyze the results. We choose BIC as our information criteria for model selection. Simplicity is a desirable property in data science models to help explain the relationship between variables. We choose BIC as our information criteria because it penalizes complex models more than AIC or AICc and therefore selects more simple models with fewer parameters as the best ones. Lower BIC scores are better.

```
## Series: value
## Model: ARIMA(0,1,1)(1,1,2)[12]
##
## Coefficients:
##          ma1      sar1      sma1      sma2
##       -0.3482 -0.4986 -0.3155 -0.4641
## s.e.   0.0499  0.5282  0.5165  0.4367
```

```
##
## sigma^2 estimated as 0.08603: log likelihood=-85.59
## AIC=181.18 AICc=181.32 BIC=201.78
```

After searching over seasonal and non-seasonal P,D, and Q variables, the best model was an ARIMA(0,1,1)(1,1,2)[12] model with BIC score of 201.78. Next, we evaluate the model via diagnostic plots and statistical tests, concluding the Box Jenkins process.

Fig.8 Residuals of the SARIMA model

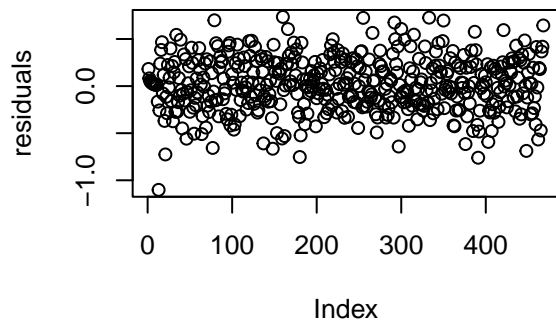


Fig.9 ACF plot of residuals

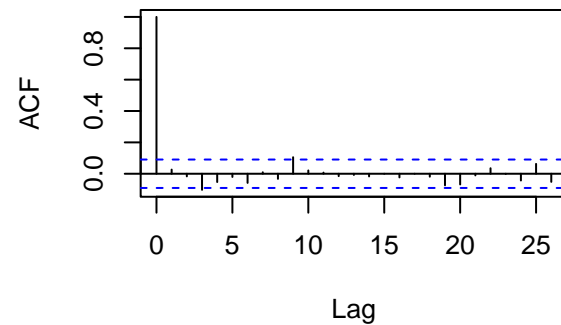


Fig.10 PACF plot of residuals

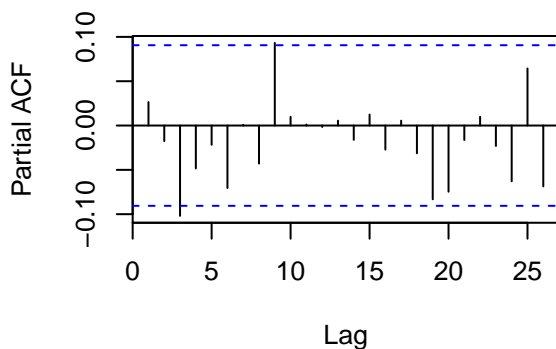
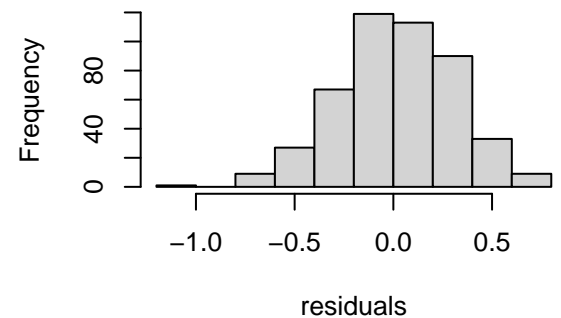


Fig.11 histogram of residuals



The residual plots (Fig 8-11) show that the SARIMA model was effective, with the residuals looking like stationary white noise (Fig 8). The time series has a mean of 0 with about constant variance, the ACF plot (Fig 9) shows no autocorrelation beyond the initial lag value. The PACF plot (Fig 10) appears to have a significant peak around the 3rd lag term, but this may be due to randomness, as it is barely passing the dashed blue line. The histogram (Fig 11) looks normally distributed at 0 with outliers creating a left tail.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

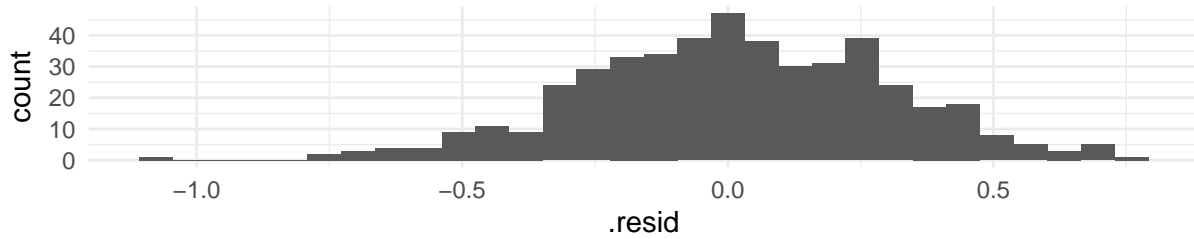
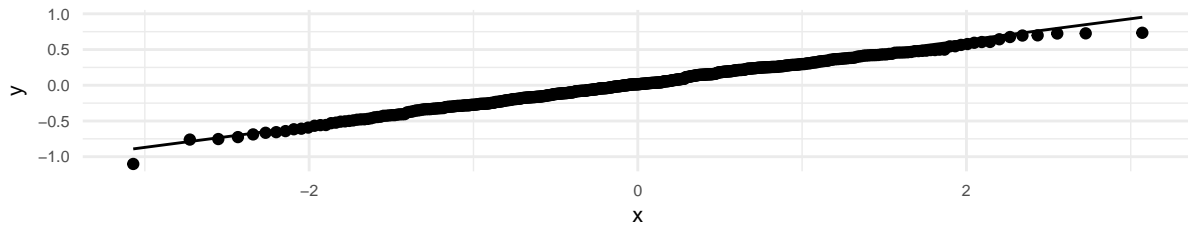


Fig.12 QQ plot of residuals



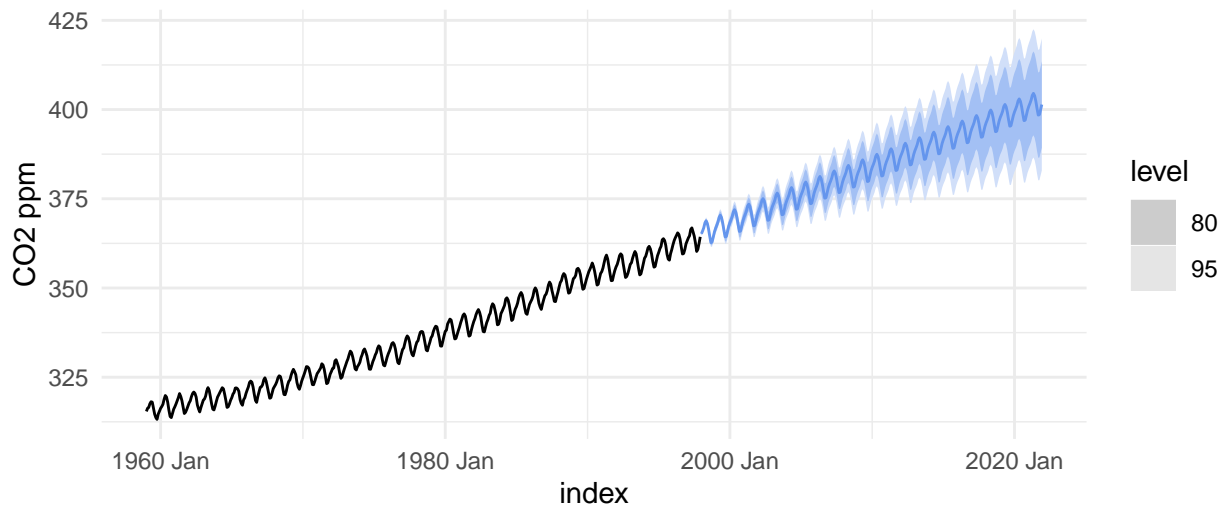
We test the residuals for stationarity with the augmented dickey fuller test. The augmented dickey fuller test has the null hypothesis that the data is non stationary. With a p-value of 0.01, we reject the null hypothesis because there is enough evidence to say that the residuals are stationary.

The Box-Ljung test has the null hypothesis that the data presented is independently distributed. When presented with the residuals of the ARIMA model, the test had p-values of 0.566 and 0.144 for lag =1 and lag = 10 respectively. For both of those lags, we fail to reject the null hypothesis and conclude that the data is independently distributed.

Finally, we visually inspect the histogram of the residuals and the qq plot to see if the residuals appear normally distributed. The histogram has the gaussian bell shaped curve with a few outliers. The qq plot shows that the data matches up with the normal distribution's quantiles. With these plots, we can confidently say that the residuals are visibly normally distributed.

To conclude, both diagnostic plots and statistical tests show that the residuals are stationary with mean 0, constant variance, and no autoregression or seasonality. We forecast our model to the year 2022 (Fig 13).

Fig.13 CO2 levels from 1959 to 2022

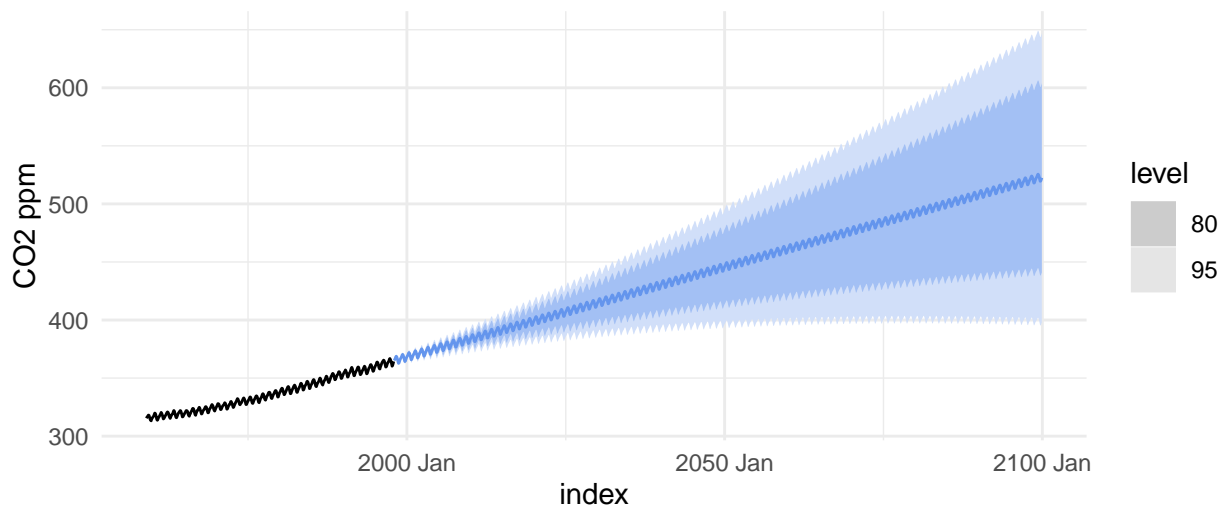


3.2 Atmospheric CO2 growth Forecast

Based on our model, the first time we could potentially see CO2 in 420 ppm is 2022-04-01 because that is when the upper 95% confidence interval (CI) of our model first reaches 420 ppm. The model's lower 95% CI hovers around 420, so there is no predicted final time. *Add why this level of CO2 is important*

The first time our model predicts the earth to reach 500 ppm CO2 on 2054-04-01, which is when the 95% CI reaches 500 ppm. The model's lower 95% CI never reaches 500, so there is no predicted final time. Below is the prediction of our model to the year 2100. Confidence intervals are shown fanning outward. The error of the predictions compounds overtime which expands the confidence intervals into a funnel shape. The farther out in time from the recorded data points, the less accurate the prediction.

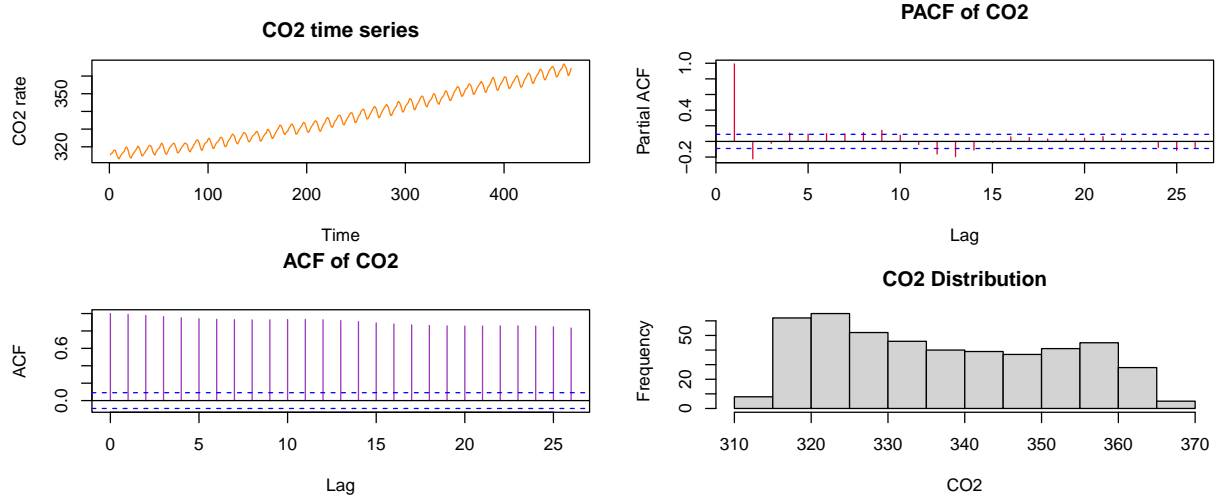
Fig.14 CO2 levels from 1959 to 2100



Introduce the question to your audience. Suppose that they *could* be interested in the question,

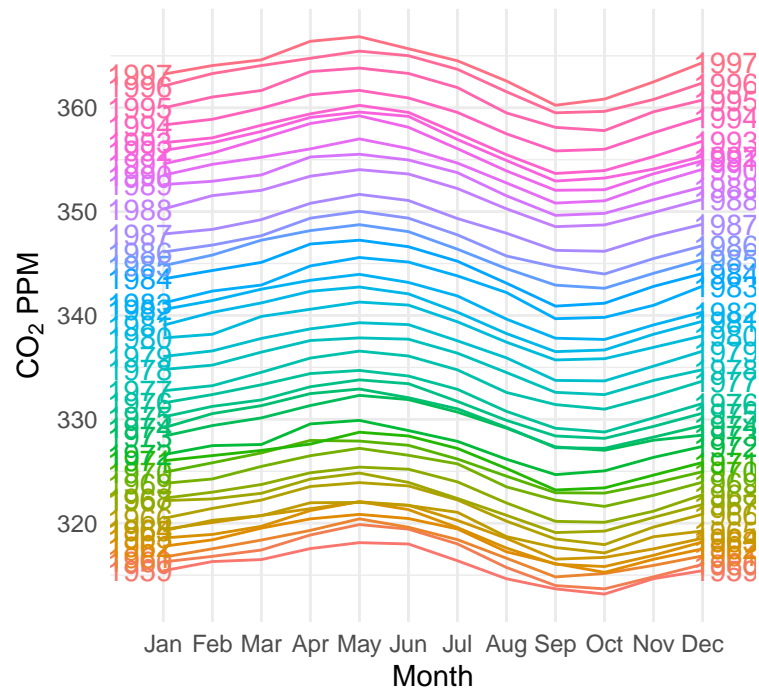
but they don't have a deep background in the area. What is the question that you are addressing, why is it worth addressing, and what are you going to find at the completion of your analysis. Here are a few resource that you might use to start this motivation. - Wikipedia - First Publication - Autobiography of Keeling An alarming pace of increase in global temperatures is being observed. Climate change is the term used to describe this occurrence. It has been proven that anthropogenic emissions, or greenhouse gas emissions brought on by human activities like deforestation and fossil fuel burning, are to blame. The phenomenon of climate change has been thoroughly studied by eminent scientists, such as Charles David Keeling, who is credited with developing the Keeling Curve, a visual representation of the rise in carbon dioxide content in the Earth's atmosphere from 1959 to the present. The following research topics will be addressed using the same data that supports the Keeling Curve: 1) How have carbon dioxide emissions grown over time? 2) How much carbon dioxide emissions should be anticipated? We will use data on atmospheric CO₂ concentrations gathered from Mauna Loa, Hawaii, in time-series analysis to respond to these study issues.

Our group will create a time-series model to demonstrate the growth in carbon emissions over the previous 38 years. What's more, we'll forecast how much carbon emissions will increase if nothing is done to lower anthropogenic emissions. These results are significant because rising carbon emissions have the potential to endanger our way of life by, among other things, causing higher temperatures, rising sea levels, and more frequent extreme weather events. We are hope that these results will be used to business strategies and initiatives that will lower our carbon footprint for the benefit of the environment. Additionally, we anticipate that these results will serve as motivation for employees at our organization to adopt actions that will lower their carbon emissions. ## (3 points) Task 1a: CO₂ data Conduct a comprehensive Exploratory Data Analysis on the `co2` series. This should include (without being limited to) a description of how, where and why the data is generated, a thorough investigation of the trend, seasonal and irregular elements. Trends both in levels and growth rates should be discussed (consider expressing longer-run growth rates as annualized averages). What you report in the deliverable should not be your own process of discovery, but rather a guided discussion that you have constructed so that your audience can come to an understanding as succinctly and successfully as possible. This means that figures should be thoughtfully constructed and what you learn from them should be discussed in text; to the extent that there is *any* raw output from your analysis, you should intend for people to read and interpret it, and you should write your own interpretation as well. ## Exploratory Data Analysis In this report, we are analyzing the monthly CO₂ levels data, captured at Mauna Loa observatory between Jan 1959 and Dec 1997, presented as monthly mean in PPM units. The unit of the data is in "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 (PPM) of CO₂ means that in every one million molecules of (dry) air there are on average 413 CO₂ molecules." According to the documentation, the air at Mauna Loa is considered to be representative of most of the northern hemisphere and potentially the globe as well, as the observatory is situated at an altitude of 3400 meters and surrounded by bare lava of the active volcano. Original data CO₂ represents CO₂ observational data for a particular year-month combination between Jan 1959 and Dec 1997. In the following section, we are performing initial EDA to better understand the data, starting by analyzing the time series, histogram, auto-correlation function (ACF), and partial auto-correlation function (PACF) plots.



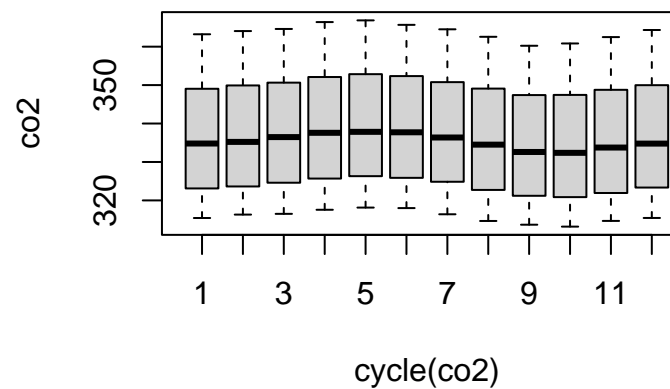
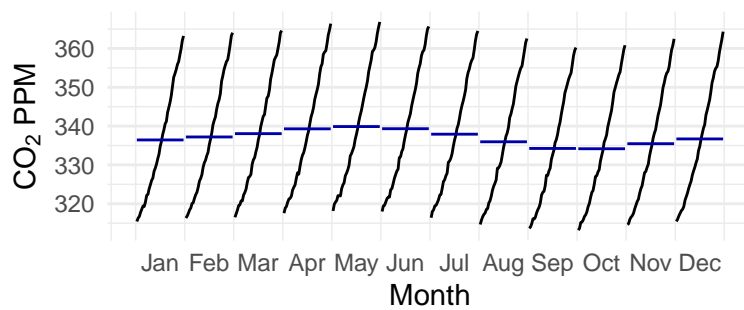
Observations The top-left graph shows the time series data from January 1959 to January 1997. The data shows a clear linear trend and an annual seasonal pattern. The linear trend means that the data is increasing over time, and the annual seasonal pattern means that the data is higher in some months of the year than others. The top-right graph shows the autocorrelation function (ACF) of the series. The ACF measures the correlation between a time series and its lagged values. The ACF shows spikes at lags of 12 months, which indicates that there is annual seasonality in the data. The ACF also shows a gradual decline, which indicates that there is a linear trend in the data. The bottom-left graph shows the partial autocorrelation function (PACF) of the series. The PACF is similar to the ACF, but it removes the effects of the intervening lags. The PACF also shows spikes at lags of 1, 2, 12, and 13. These spikes suggest that there is some seasonality in the series, but that the seasonality is not perfectly periodic. The PACF also shows a significant spike at lag 1, which indicates that there is a strong relationship between the current value of the series and its previous value. This is consistent with the linear trend that was observed in the ACF plot. The PACF's strong relevance at various lags may indicate that the time series is an ARMA process. An ARMA process is a type of stochastic process that is characterized by both autoregressive (AR) and moving average (MA) terms. The bottom-right graph shows the histogram of the data. The histogram shows that the data is not normally distributed. This is not ideal for asymptotic confidence interval estimation, which requires the residuals to be normally distributed. In summary, the time series data shows some seasonality and a linear trend. The PACF plot suggests that the time series may be an ARMA process. However, the data is not normally distributed, which is not ideal for asymptotic

Monthly Mean Seasonal Plot for CO₂ PPM



confidence interval estimation.

Monthly Mean Plot for CO₂ PPM Level



Observations The plots show that the CO2 level...

A Appendix: Model Robustness

```
fit_poly %>% dplyr::select(cubic) %>% report()
```

```
## Series: value
## Model: TSLM
##
## Residuals:
```

```

##           Min           1Q           Median           3Q           Max
## -1.172427 -0.262499 -0.003472  0.264990  1.476396
##
## Coefficients:
##           Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    3.157e+02  1.094e-01 2884.373 < 2e-16 ***
## trend()         4.471e-02  1.922e-03  23.266 < 2e-16 ***
## I(trend()^2)    1.856e-04  1.159e-05  16.016 < 2e-16 ***
## I(trend()^3)   -9.331e-08  1.979e-08  -4.715 3.44e-06 ***
## season()year2   6.565e-01  1.036e-01   6.334 6.94e-10 ***
## season()year3   1.383e+00  1.036e-01  13.342 < 2e-16 ***
## season()year4   2.495e+00  1.036e-01  24.076 < 2e-16 ***
## season()year5   2.986e+00  1.036e-01  28.809 < 2e-16 ***
## season()year6   2.366e+00  1.036e-01  22.829 < 2e-16 ***
## season()year7   9.035e-01  1.037e-01   8.717 < 2e-16 ***
## season()year8  -1.123e+00  1.037e-01 -10.832 < 2e-16 ***
## season()year9  -2.908e+00  1.037e-01 -28.050 < 2e-16 ***
## season()year10 -3.139e+00  1.037e-01 -30.273 < 2e-16 ***
## season()year11 -1.970e+00  1.037e-01 -19.001 < 2e-16 ***
## season()year12 -8.910e-01  1.037e-01  -8.592 2.44e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4145 on 369 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9988
## F-statistic: 2.194e+04 on 14 and 369 DF, p-value: < 2.22e-16

```