W271 Group Lab 2 Team 2

Katt Painter, Bo He, Akanksha Chattopadhyay, Ian Vaimberg

POV: 1997

Context

Climate science has emerged as a leading field of interest in the 20th century. Efforts to bring awareness to the impact of human intervention on the environment, such as the burning of fossil fuels, have proved fruitful. The Intergovernmental Panel on Climate Change (IPCC) has reinforced these efforts. In 1990 they released the First Climate Assessment Report stating that "human activities are substantially increasing the atmospheric concentration of greenhouse gases" (IPCC, 1990). Greenhouse gases are a group of gases, such as carbon dioxide, methane, and nitrous oxide, that when present in higher concentrations in the atmosphere, raise the surface temperate of the Earth. Carbon dioxide is the most abundant greenhouse gas that is produced from human activity, namely energy production via fossil fuel combustion. Since the industrial revolution, energy consumption from petroleum and natural gas sources has risen dramatically. This report aims to investigate the following questions: How have the levels of atmospheric CO2 changed over time? And, is there an identifiable pattern that will persist into the future? Forecasting atmospheric carbon dioxide concentrations allows scientists to measure the corresponding impact to the global environment and justify the need for human intervention in the opposite direction.

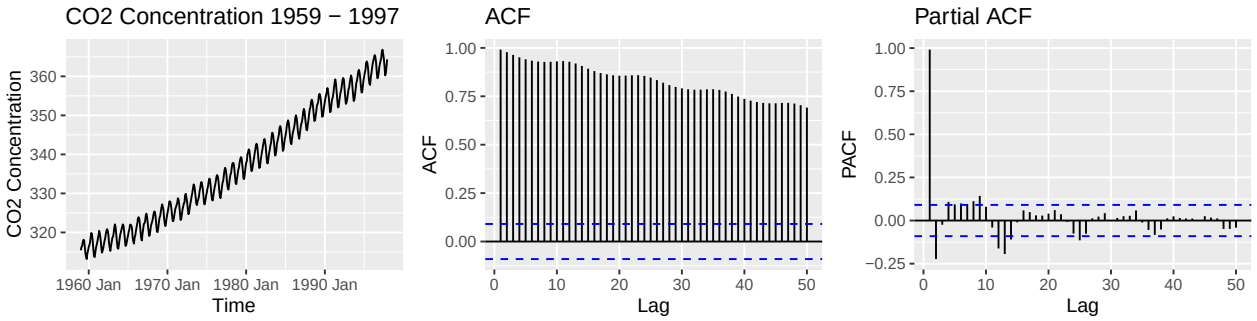
Data and Exploration

Charles Keeling was a research scientist who made it his life's work to survey the atmosphere in hopes of confirming Svante Arrhenius's theory that fossil fuel combustion is increasing the concentration of CO2 in the atmosphere. To this end, Keeling collected atmospheric CO2 concentration measurements at a number of sampling-stations including the Mauna Loa Observatory in Hawaii. These measurements were taken using a CO2 analyzer which detects the amount of infrared absorption present in a air sample and turns it into a mole fraction of CO2, defined as the total CO2 molecules divided by the total non-water vapor molecules in the air, measured in parts per million (ppm). This report uses the data collected at the Mauna Loa Observatory between January 1959 and December 1997. The dataset consists of 468 observations with each observation representing the monthly total atmospheric concentration of CO2 (ppm). The observations for February, March, and April 1964 were unavailable so the values in the dataset were generated via linear interpolation between the observations for January and May 1964.

In order to better understand the characteristics of this time series, we conducted a exploratory analysis prior to modeling. Figure 1 shows the time series plot for CO2 concentration, its autocorrelation plot, and its partial autocorrelation plot. The time series plot shows a clear positive trend as well as the presence of seasonality. The autocorrelation plot provides evidence to support the presence of both trend and seasonality as it decays with increasing lags and shows a spike at about every twelfth lag, indicating a seasonal cycle.

- Discuss seasonal/trend decomposition
- What about growth rates?
  - Some acceleration to the growth rates
  - Growth rate is mean reverting with given process

Figure 1. Time Series, ACF, and PACF



Models and Forecasts

While the exploratory analysis allows us to infer the underlying process of the time series, we cannot confirm this process without empirical modeling. In this section we will produce multiple models from the following classes, linear models and ARIMA models. Following the modeling and model evaluation, we will use the models to produce forecasts for what the CO2 concentration will be over the next 20 years.

Linear Models

Three linear models were considered, these models are describe by the following set of equations:

$$y_t = t + \epsilon_t \tag{1}$$

$$y_t = t + t^2 + \epsilon_t \tag{2}$$

$$y_t = t + t^2 + d_{2,t} + d_{3,t} + ... + d_{11,t} + d_{12,t} + \epsilon_t \tag{3}$$

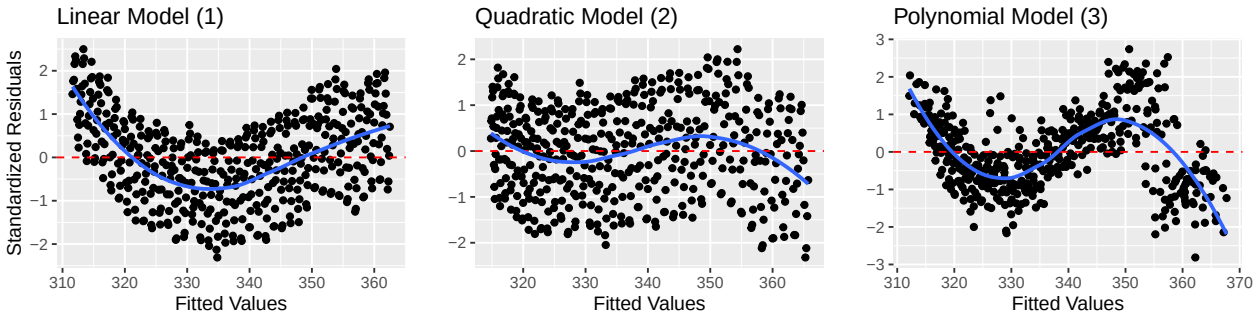
where t is the time index (i.e., year-month) represented as an integer and d is a dummy variable representing each month. For example d_2 would correspond to the month of February, d_3 would correspond to the month of March, and so on.

The following table reports the coefficients that produce the best fit for each of the previously defined models. It also reports several performance metrics such as R-squared, adjusted R-squared, F statistics, etc. The table suggests that each model explains a large amount of the variation in CO2 emissions with adjusted R-squared values ranging from 0.97 to 0.99. This indicates that these models appear to perform extremely well. The p-values associated with each predictor are also significant, indicating that each predictor increases the explanatory power of the model. However, in order to determine the appropriateness of using linear models to estimate CO2 concentrations we must examine the residuals of each model.

	Response: Atmospheric CO2 Concentration		
	(1)	(2)	(3)
Month Index	0.11^{***} (0.001)	0.07^{***} (0.003)	0.07^{***} (0.001)
Month Index^2		0.0001^{***} (0.0000)	0.0001^{***} (0.0000)
Feb			0.66^{***} (0.16)
March			1.41^{***} (0.16)
April			2.54^{***} (0.16)
May			3.02^{***} (0.16)
June			2.35^{***} (0.16)
July			0.83^{***} (0.16)
Aug			-1.23^{***} (0.16)
Sep			-3.06^{***} (0.16)
Oct			-3.24^{***} (0.16)
Nov			-2.05^{***} (0.16)
Dec			-0.94^{***} (0.16)
Constant	311.50^{***} (0.24)	314.76^{***} (0.30)	314.68^{***} (0.15)
Observations	468	468	468
R^2	0.97	0.98	1.00
Adjusted R^2	0.97	0.98	1.00
Residual Std. Error	2.62 (df = 466)	2.18 (df = 465)	0.72 (df = 454)
F Statistic	$14{,}794.94^{***}$ (df = 1; 466)	$10{,}749.90^{***}$ (df = 2; 465)	$15{,}314.77^{***}$ (df = 13; 454)
Note:			*p<0.05; **p<0.01; ***p<0.001

Figure 2 shows the standardized residuals against the fitted values for each linear model. There are clear patterns present in each of these plots. The first model shows a U-shaped curve that indicates the model is under predicting CO2 concentrations that are in the mid-range and over predicting CO2 concentrations closer to the edges. Whereas the quadratic model appears to oscillate in whether it is over or under predicting CO2 concentrations. The polynomial model appears to also have a U-shaped pattern. However, compared to the first model, its errors appear to be larger, especially at edge values. Nevertheless, the residuals appear to be homoskedastic for all three models which suggests that no transformation (i.e., logarthmic) is needed. Overall, these plots suggest that linear models are not appropriate for this particular data set. They do not appear to adequately capture the underlying time series process. So we must turn to ARIMA models.

Figure 2. Linear Model Evaluation



ARIMA Models

Due the strong presence of seasonality in the data, the next set of models are seasonal autoregressive integrated moving average models (SARIMA). The general notation (4) used to express these models is stated below as well as the short hand notation (5).

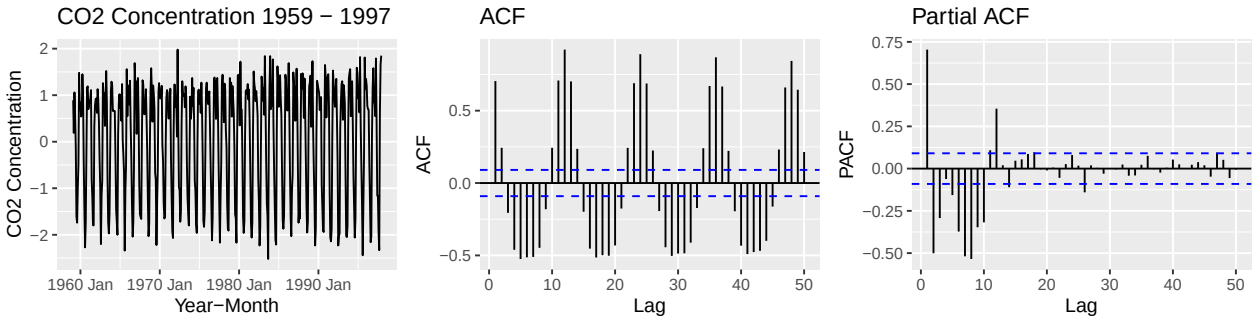
$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D x_t = \mu + \theta(B)\Theta(B^s)\epsilon_t \tag{4}$$

$$SARIMA(p,d,q)(P,D,Q)_s \tag{5}$$

These models include moving average (MA) terms, autoregressive (AR) terms, and differencing (D) terms at non-seasonal and seasonal lags. This allows the model to more accurately capture and account for the seasonality present in the data. To help determine the seasonal MA, AR, and D terms we took a first difference of the data and re-plotted Figure 1 with the differenced data. Figure 2 shows this plot. The plot indicates that after one difference the data is near stationary suggesting that the differencing term should be 1. We performed ADF and KPSS test to confirm stationarity. Moreover, the PACF shows a large spike followed by oscillations between positive and negative correlations which appear to weaken as the lag value increases. This indicates the presence of a higher order MA term. The plots do not appear to support the presence of a AR term.

```
## Warning: Removed 1 row containing missing values (`geom_line()`).
```

Figure 2. Time Series, ACF, and PACF After First Difference



Based on the assessment of Figure 2, we fit three SARIMA models of the form:

$$SARIMA(0,1,1)(0,1,1)_{12} \tag{6}$$

$$SARIMA(3,1,0)(3,1,0)_{12} \tag{7}$$

$$SARIMA(0,1,1)(1,1,2)_{12} \tag{8}$$

The final model was selected using the BIC scores with the ultimate winner being the $SARIMA(0,1,1)(0,1,1)_{12}$ model with a BIC value of 194.7. BIC was chosen due to its inherent penalty on adding additional terms leading to a well fitted and simpler model. Finding the ideal model is done through grid search of various parameters as we employed three simultaneous searches to see the most variety of different model specialties. All searches included non-zero difference terms based on the our EDA while one model focused on AR terms, another MA terms and one finally on both at the same time. To further assess the appropriateness of this model, we looked at a residual ACF plot which closely resembles that of a white noise process with lags rarely falling outside of the 95% confidence interval around 0 correlation. This supports the use of the SARIMA model to model the CO2 dataset.

Forecasts

- Forecasts
- Predictions for when CO2 is expected to be at 420 ppm and 500 ppm
- Interpretation/evaluation?

Conclusions

- Implications