W271 Lab 2: CO2 1997

Ken Trinh, Lisa Wu, Ray Cao, Sophie Yeh

Contents

0.1 (3 points) Task 0a: Introduction

If you are concerned about global warming (or question whether this is true), you may have heard about the "Keeling Curve" which is named after the scientist Charles David Kneeling. Kneeling started measuring and monitoring the accumulation of carbon dioxide (CO_2) in the atmosphere in 1958. Many scientists credit the Keeling curve with first bringing our attention to the current increase of CO_2 in the atmosphere. The one key question in people's minds is whether the CO2 increase trend observed in the past will continue and at what speed, over the next few decades. The answer to this question is critical to our policy makers and environmentalists, as the forecast CO_2 results will help them evaluate how concerned they should be and what actions to take to minimize the consequences. We will conduct the study of the CO_2 data set to answer this question. We plan to explore the data set and modeling alternatives to determine whether a reliable forecast model can be developed to forecast through the year of 2022.

0.2 (3 points) Task 1a: CO2 data

The CO2 data, tracking the atmospheric CO_2 level in part per million by volumne (ppmv), was measured continuously at the Mauna Loa Observatory in Hawaii from 1958 to the present day (the end of 1997). In the data set, there are a total of 468 monthly observations from January 1958 to December 1997 with no missing data. Prior to this effort, measurements of CO_2 concentrations had been taken on an ad hoc basis at a variety of locations. Keeling created a frequent and consistent measurement framework of CO_2. Keeling and his collaborators measured the incoming ocean breeze above the thermal inversion layer to minimize local contamination from volcanic vents. The data were normalized to remove any influence from local contamination. His work minimized the data noises due to measurement errors or differences.

Figure 1 shows that the CO_2 trend has steadily increased over time (close to a linear trend line), although the annual growth rate seems to be range-bound (mostly 0-0.75%) with no clear trend. CO_2 is a greenhouse gas, so the increasing trend has significant implications for global warming. From the histogram chart, we observed that the CO_2 levels are not normally distributed, ranging from 310 to 370 ppmv. There are no extreme outlines in this data set. Furthermore, from the CO_2 Decomposition graph, we observed trend, season and irregular components of the data set. Aside from the trend line discussed above, we observed strong seasonality, and the remaining irregular effect.

The boxplot in Figure 2 further validated the seasonal pattern in the Decomposition graph. The maximum level occurs in May and then decreases during the spring and summer as new plant growth takes CO_2 out of the atmosphere. After reaching a minimum in October, the level rises again in the late fall and winter as plants and leaves die off and decay, releasing CO2 back into the atmosphere. The difference between the highest and lowest monthly averages is 5.46 ppmv.

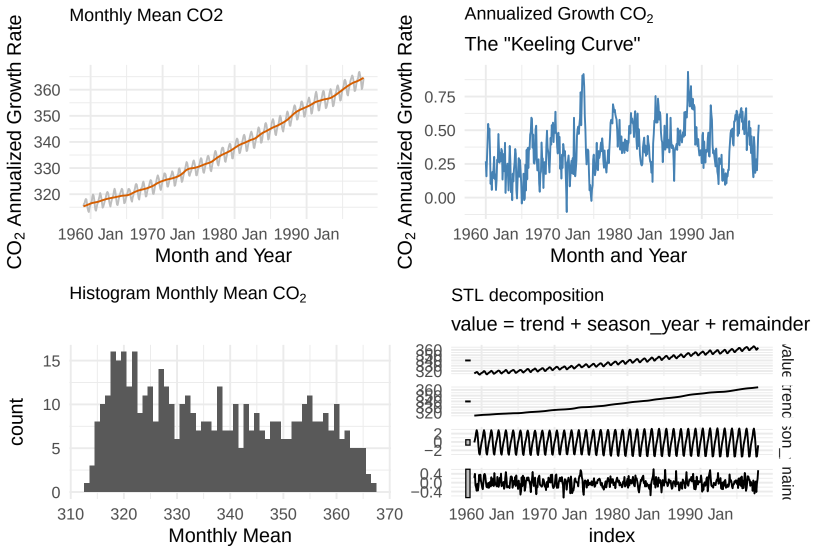
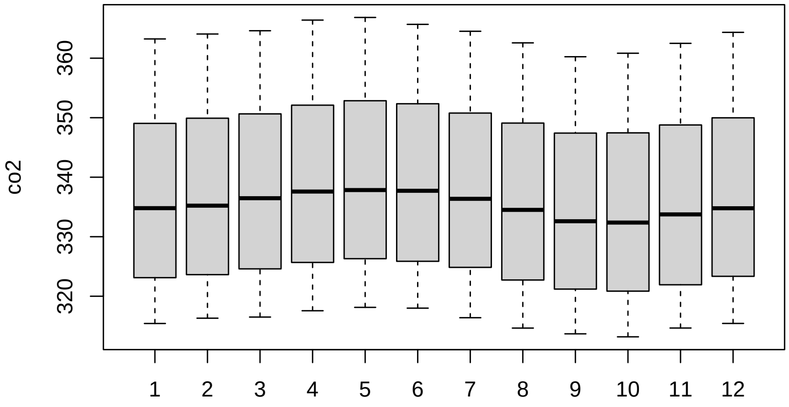


Figure 1: Atmospheric CO2 Level Time Series Overview



cycle(co2, xlab = "Month", ylab = "CO2 (ppmv)", main = "Monthly Mean CO2")

Figure 2: Seasonality CO2 Level Monthly Distribution

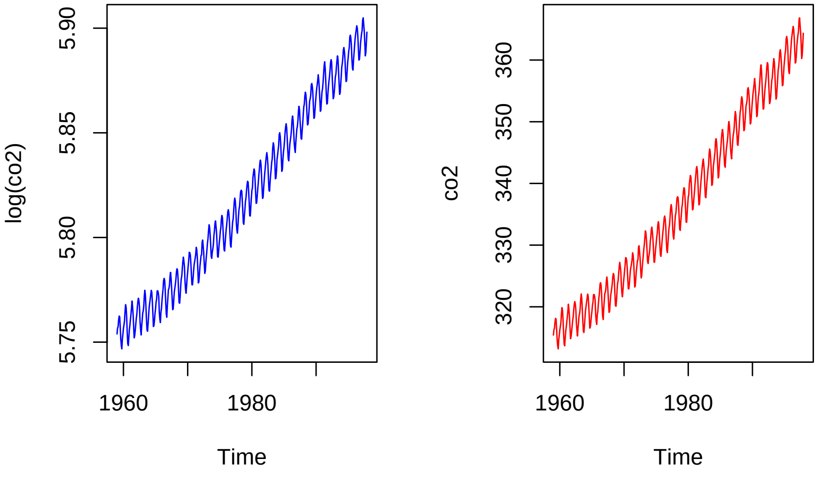


Figure 3: CO2 level regular versus logarithmic transformation

0.3 (3 points) Task 2a: Linear time trend model

As discussed in the CO_2 Data section above, the CO_2 time series data set follows closely to a linear trend line, so we will first explore a linear regression model for the data set. Given both the original data series and its annual growth rate are range-bound with no clear sign of exponential growth and increased variance with time, we don‘t think it is necessary to perform a logarithmic transformation. This is supported by both Figure 1 and Figure 3. In these plots, the series variance does not seem to change over time, and the trend does not have major curvature. Thus a logarithmic transformation is not needed for this data set.

In the linear trend model, we use time as the explanatory variable and CO_2 level as the response variable. Reported in Table 1, this model has an intercept term and a positive slope of 1.3. Both coefficients are statistically significant, with p-value less than 0.001. The model residuals, Figure 4 (lelf), is curved, which violates the assumption of independent and identically distributed residuals with zero mean expectation. Variance increases as the fitted values increase, which violates the homoskedasticity assumption of classical linear model. Clearly this simple model failed to sufficiently capture the data characteristics.

We then evaluated the quadratic model by adding the quadratic term of time. The coefficients are reported in Table 1 and are statistically significant. The right plot in Figure 4 shows that the residuals line is still curved. Variance still shows some small level of heteroskedasticity. This model also does not adequately capture the data characteristics.

Finally we fit a polynomial time trend model and incorporate seasonal dummy variables. We will use the goodness-of-fit information criterion scores measurement to select the polynomial degree that optimizes the model fit. The three goodness-of-fit assessment measurements are AIC, AICc and BIC. Lower AIC, AICc and BIC indicates better model performance. Generally, BIC has a larger penalty for models with more parameters and therefore selects sparser models with fewer parameters compared to AIC and AICc. We ran both AIC and BIC assessments and displayed the result in Figure 5.

We use a range of 1 to 5 polynomial degrees for the trend variable and don’t recommend trying higher polynomial degrees to avoid over-fitting. This result shows that 3 is the optimal degree with the lowest AIC and BIC score.

```r
mod.lm1 <- lm(co2 ~ time(co2))
mod.lm2 <- lm(co2 ~ time(co2) + I(time(co2)^2))
```

Table 1: Estimated Atmospheric CO2 Level

	Output Variable: CO2 Level in ppmv	
	(1)	(2)
linear time	1.307*** p = 0.000	−49.191*** p = 0.000
quadratic time		0.013*** p = 0.000
(Intercept)	−2,249.774*** p = 0.000	47,702.940*** p = 0.000
Observations	468	468
R^2	0.969	0.979
Adjusted R^2	0.969	0.979
Note:	*p<0.05; **p<0.01; ***p<0.001	

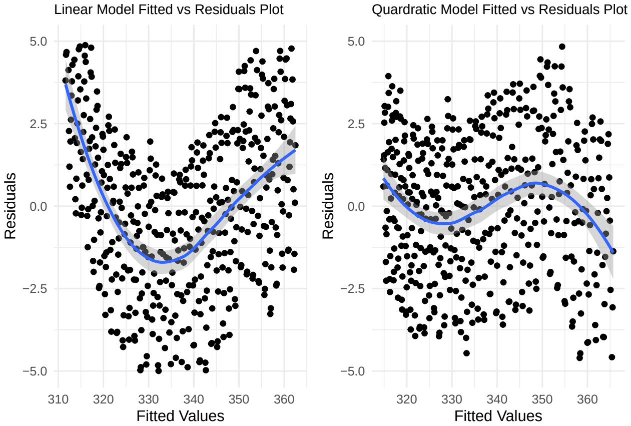


Figure 4: Residuals vs Fitted Plots: Linear Model (Left) Quadratic Model (Right)

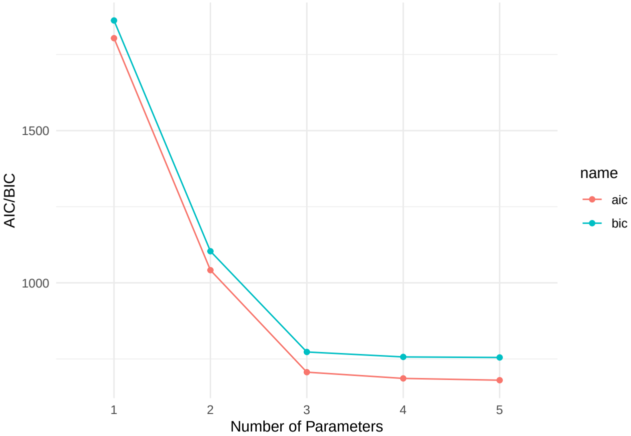


Figure 5: Polynomial trends selection using information criterions

```
mod.poly <- tslm(co2 ~ poly(trend, 3) + season)
```

In Figure 6, we plot the residuals and noted that residuals have significant positive autocorrelation (not randomly distributed) which will underestimate the standard errors. Variance is not constant as well. Classical linear regression model assumptions are violated. While further model improvement is needed, we will use the fitted polynomial model to forecast the CO_2 level through 2020 (Figure 7). This polynomial time trend with a seasonal dummy variable was able to capture the linear and seasonal trend when forecasted to 2020. There's a slight curve at the 2020 tail of the trend due to the 3rd degree polynomial.

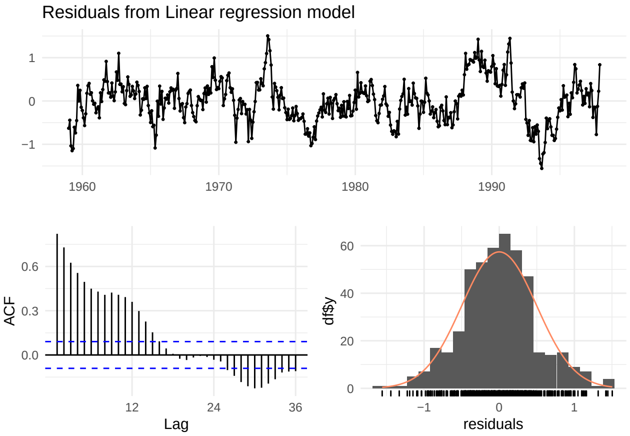


Figure 6: Polynomial Residuals Diagnostic Plots

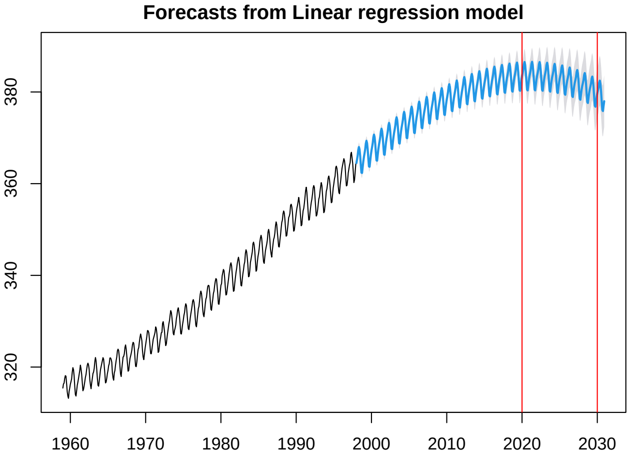


Figure 7: Linear model with polynomial trend and dummy seasonality forecast

0.4 (3 points) Task 3a: ARIMA times series model

ARIMA model has three parameters (p, d, q). p stands for the number of lag terms in the model, d stands for the number of times the raw observations are differenced, and q stands for the size of the moving average (MA) window. Typically for ARIMA, PACF plot indicates the lag order while ACF indicates how many MA terms are required to remove autocorrelation in the stationary series.

```
adf.test(diff(co2))
```

```
##
##  Augmented Dickey-Fuller Test
##
```

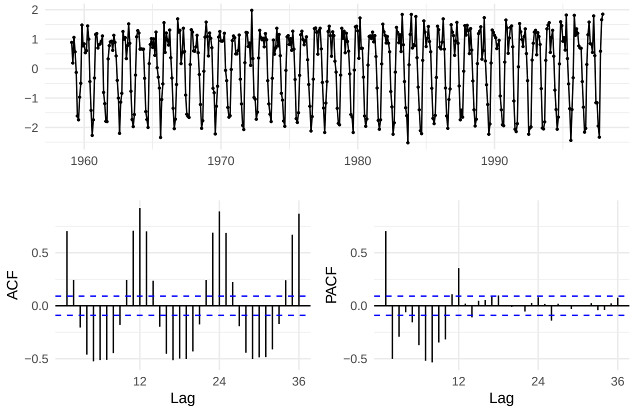


Figure 8: CO2 Level 1st order differencing ACF and PACF plots

```
## data:  diff(co2)
## Dickey-Fuller = -30.38, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

As noted earlier, the time series plot has a strong positive trend and seasonal effect and is non-stationary. Figure 8 shows that, after taking a first difference of the data, the time series plot appears to oscillate around 0. To ensure the it is stationary, we applied the Augmented Dickey-Fuller test, which returns a significant p-value less than 0.05. Thus, we have sufficient evidence to reject the null hypothesis and that the time series after the first differencing is stationary. The ACF has significant cyclic lags due to seasonality, but there are no signs of dampening. The PACF plot has several significant lags until the autocorrelation dampen.

Because the ACF has persistent significant lags while the PACF has dampening oscillations, the data leans towards an AR(3) process. With the differencing component, the ARIMA(p,1,q) model will be the most appropriate. The correct MA and AR parameters will be tested. To select the best model for CO2 level, the BIC information criterion is used because BIC has a larger penalty for models with more parameters and therefore selects sparser models with fewer parameters compared to AIC and AICc.

	coefs
ma1	-0.3482279
sar1	-0.4985947
sma1	-0.3155132
sma2	-0.4641321

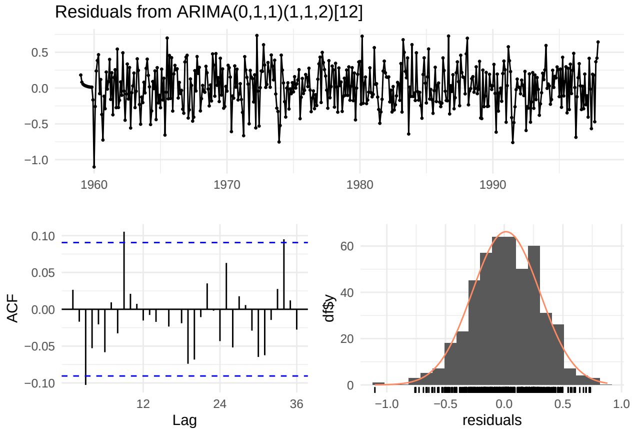

Residuals from ARIMA(0,1,1)(1,1,2)[12]

Figure 9: ARIMA(0,1,1)(1,1,2)[12] CO2 Level residuals

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(1,1,2)[12]
## Q* = 21.999, df = 20, p-value = 0.3406
##
## Model df: 4.   Total lags used: 24
```

The final model is estimated to be ARIMA(0,1,1)(1,1,2)[12] with a BIC of 201.78. Checking the residuals of the model in Figure 9, the residuals plot oscillates around 0. The ACF autocorrelations are all below or only slightly over the threshold value, and the distribution is Gaussian. The Ljung-Box test returns a large p-value of 0.3406, suggesting that there is sufficient evidence to reject the null hypothesis and the residuals are stationary. Since the model residuals are stationary, we decided to perform a forecast on atmospheric CO2 level to 2022 (Figure 10). Noted that in the figure after 2010, the forecast starts to curve with a wider confidence interval.

```
arima_pred <- forecast::forecast(mod.arima, level = c(95), h = 25 * 12)
plot(arima_pred)
```

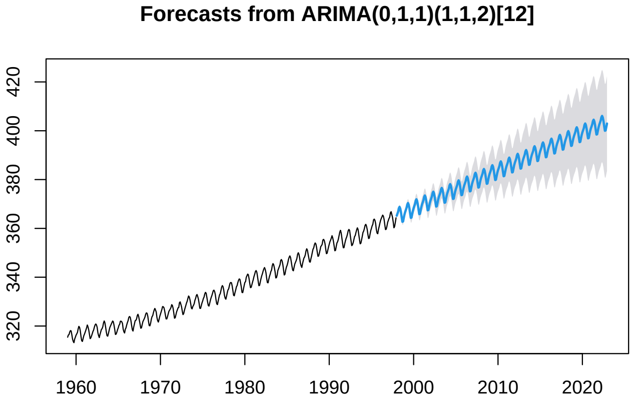


Figure 10: ARIMA(0,1,1)(1,1,2)[12] CO2 level 2022 forecast

0.5 (3 points) Task 4a: Forecast atmospheric CO2 growth

```
##          Point.Forecast    Lo.95    Hi.95
## May 2031       420.0598 392.3004 447.8192
## Oct 2086       499.4602 396.8727 602.0478
## Jan 2100       523.5351 398.8328 648.2374
```

To demonstrate future accumulated atmospheric CO_2 level, we ran a forecast to see when CO_2 level will hit certain target. Based on the model forecasts the atmospheric CO_2 level is expected to reach 420 ppm by May 2031 and 500 ppm by Oct 2086. By Jan 2100, CO2 levels will be at 523.5 ppm. We are not confident about these predictions because the lower bound of the confidence interval has plateaued at approximately 390 ppm while the upper bound continues to grow higher. While the model results has a wide confidence interval, the actual level accumulation could dramatically exceed the forecast level. Since CO_2 is a green house gas, any actions that we take now could prevent drastic damage in the future.