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 wonder whether this is true or not), 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 CO_2 will continue to go up and at what speed, over the next few decades. The answer to this question is critical to our policy makers and environmentalists. The forecast CO_2 results will help them evaluate how concerned they should be and what actions to take to minimize the consequences. In order to answer this question, we will conduct the study of the CO_2 data set and develop a model(s) to forecast CO_2.

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 since 1958. This data has 468 monthly observations, from January 1958 to December 1997 (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.

Figure 1 shows that CO_2 level is trending up over time, with seasonal variability. The annual growth rate is mostly in the range of 0-0.75%, with a modest upward trend. The long run average of the annual growth rates is 0.37%. Since CO_2 is a greenhouse gas, 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 is no extreme outliers in this data set. Furthermore, the CO_2 Decomposition graph shows the upward trend, seasonal effect and irregular components of the data set.

In our analysis, we noted that the maximum level occurs in May and then decreases during the warm seasons as new plant growth takes CO_2 out of the air. After reaching a minimum in October, as plants die off in the cold weather, CO_2 is released back into the atmosphere. The difference between the peak and trough monthly averages is 5.46 ppmv.

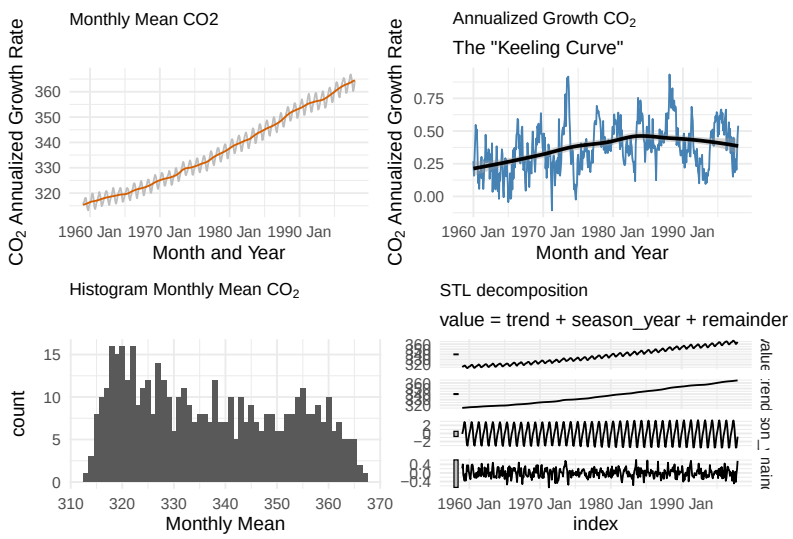

Figure 1: Atmospheric CO2 Level Time Series Overview

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

As shown in Figure 1 of the CO_2 Data section, the CO_2 time series data set follows closely to a linear trend line, with a slight curvature. The annual growth rates are range-bound, with a modest upward trend. We don't see a strong sign of exponential growth or increased variance over time. Hence we don't think it is necessary to perform a logarithmic transformation for this data set.

We will first develop a linear trend model, with time as the explanatory variable. Reported as Model (1) 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, shown in Figure 2 (left), are 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. See Model (2) results in Table 1. The right plot in Figure 2 also shows a curved residuals line. Variance still shows some level of heteroskedasticity. This model also does not adequately capture the data characteristics.

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

Finally we fitted a polynomial time trend model which incorporated seasonal dummy variables. We used the goodness-of-fit information criterion to select the polynomial degree that optimizes the model fit. The three goodness-of-fit metrics are AIC, AICc and BIC. Lower AIC, AICc and BIC score 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 and displayed the result in Figure 3. 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.

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

Figure 4 shows that the residuals have significant positive autocorrelation, which will underestimate the standard errors. Variance is not constant as well. Classical linear regression model assumptions are violated.

Table 1: Estimated Atmospheric CO2 Level

	Output Variable: CO2 Level in ppmv	
	(1)	(2)
linear time	1.307***	−49.191***
	p = 0.000	p = 0.000
quadratic time		0.013***
		p = 0.000
(Intercept)	−2,249.774***	47,702.940***
	p = 0.000	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

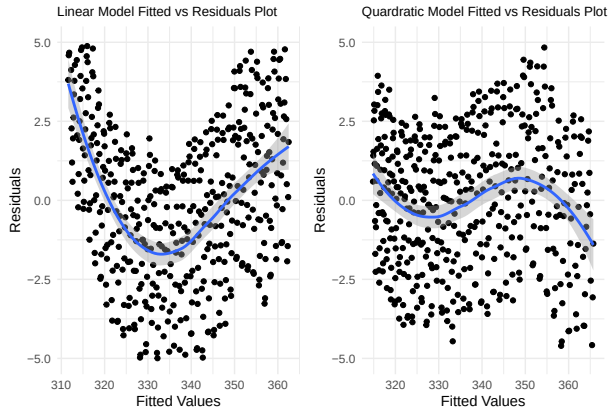

Figure 2: Residuals vs Fitted Plots: Linear Model (Left) Quadratic Model (Right)

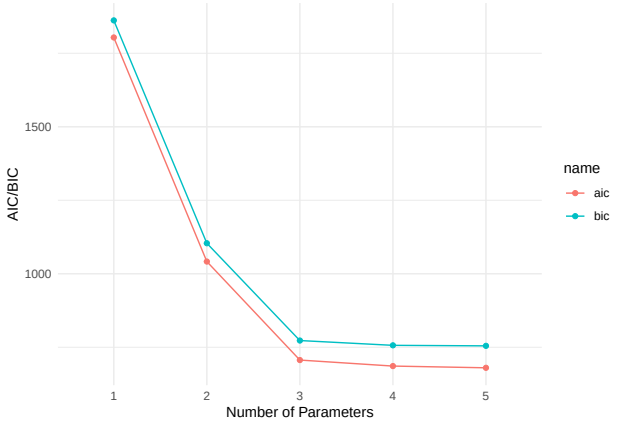

Figure 3: Polynomial trends selection using information criterions

While further improvements are necessary, we will use the fitted polynomial model to forecast the CO_2 level through 2020 (Figure 5). The forecast results capture the trend and seasonal effect well through 2020. After 2020, the forecast CO_2 level trends down due to the 3rd degree polynomial effect.

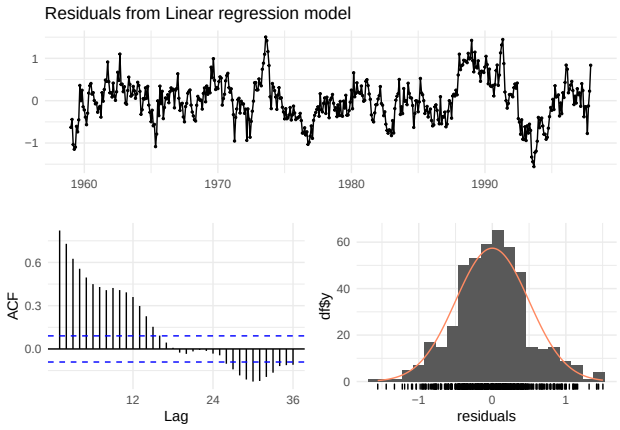

Figure 4: Polynomial Residuals Diagnostic Plots

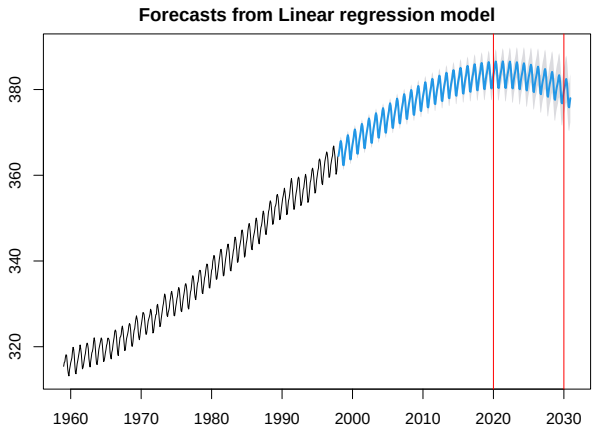

Figure 5: Linear model with polynomial trend and dummy seasonality forecast

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

We will fit a ARIMA model for this data set. This model has three parameters (p, d, q). p stands for the number of lag terms, 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 the MA terms needed to transform the data to a stationary time series.

As discussed earlier, this time series has a strong positive trend and seasonal effect and is non-stationary. Figure 6 shows that, after taking a first difference of the data, the resulting time series appears to oscillate around 0. To ensure that 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 believe that the time series after the first differencing is stationary. The ACF plot show significant cyclic lags due to seasonality, with no signs of dampening. The PACF plot has significant lags in the first 12 months which dampen after.

##

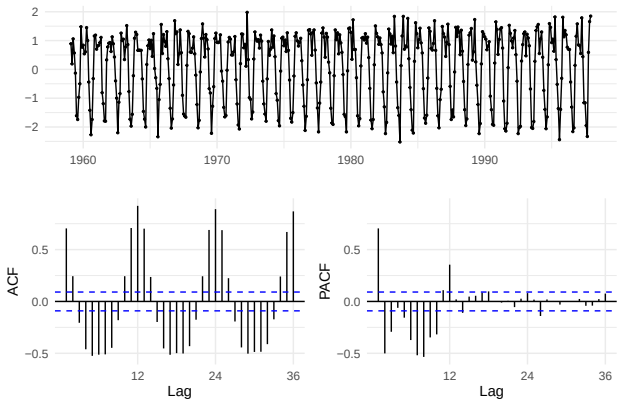

Figure 6: CO2 Level 1st order differencing ACF and PACF plots

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

Because the ACF has persistent significant lags while the PACF has dampening oscillations, the data leans towards a mixed of AR and MA process, with the first differencing order. Hence the ARIMA(p,1,q) model will be the most appropriate model form. We will use an iterative process to select the optimal AR and MA parameters based on the goodness-of-fit information criterion (AIC/AICc/BIC). We chose to use BIC for this process, as BIC has a larger penalty for models with more parameters and tends to select sparser models with fewer parameters compared to AIC and AICc.

```
# use auto arima and BIC to optimize AR and MA terms
mod.arima <- auto.arima(co2, d = 1, ic = "bic", trace = FALSE, seasonal = TRUE)
knitr::kable(mod.arima$coef,  col.names= "coefs", "latex")
```

	coefs
ma1	-0.3482279
sar1	-0.4985939
sma1	-0.3155139
sma2	-0.4641315

The final model is estimated to be ARIMA(0,1,1)(1,1,2)[12] with a BIC of 201.78. We analyzed the model residuals to evaluate model performance. Figure 7 shows that the residuals oscillates around 0. The ACF plot shows no significant autocorrelation, like a white noise process. The Ljung-Box test returns a large p-value of 0.3406, suggesting a very strong evidence that the residuals are stationary. Since the model residuals are stationary, we will use the model to forecast atmospheric CO2 level to 2022 (Figure 8). Noted that after 2010, the forecast starts to have a wider confidence interval.

```
##
##  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
```

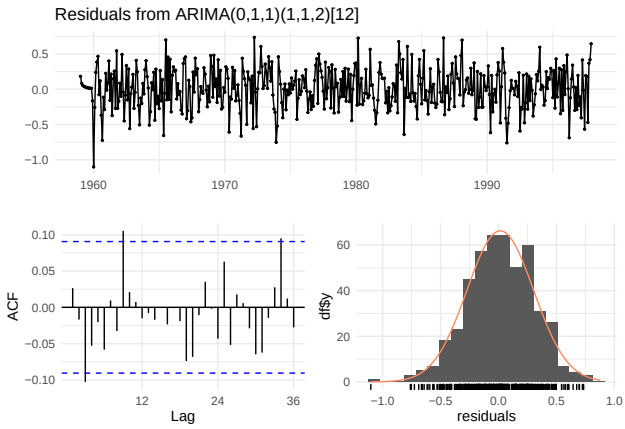

Figure 7: ARIMA(0,1,1)(1,1,2)[12] CO2 Level residuals

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

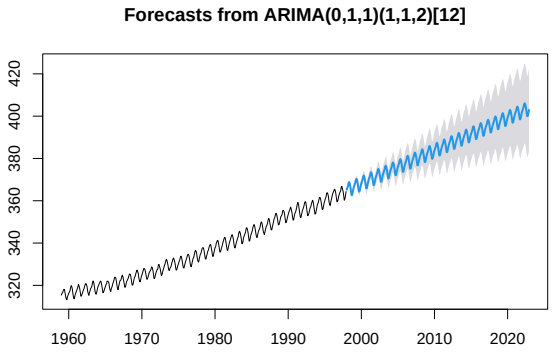

Figure 8: ARIMA(0,1,1)(1,1,2)[12] CO2 level 2022 forecast

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

We used the ARIMA model to forecast accumulated atmospheric CO_2 levels through 2100, to gauge when CO_2 levels will hit certain target. Our model forecasts that the atmospheric CO_2 level will reach 420 ppm by May 2031 and 500 ppm by Oct 2086. By Jan 2100, CO2 levels will reach 524 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 forecast has a wide confidence interval, the actual level accumulation could dramatically exceed the expected forecast level. Since CO_2 is a green house gas, any actions that we take now could prevent drastic damages in the future.

	Point.Forecast	Lo.95	Hi.95
May 2031	420	392	448
Oct 2086	499	397	602
Jan 2100	524	399	648