6.3732 PSet 3 Part 1

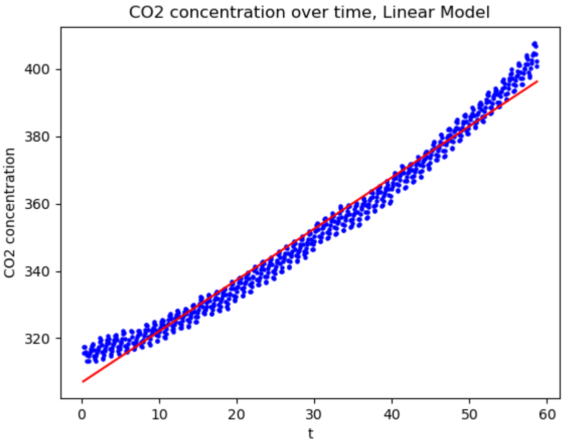
Stephen Andrews

March 15, 2025

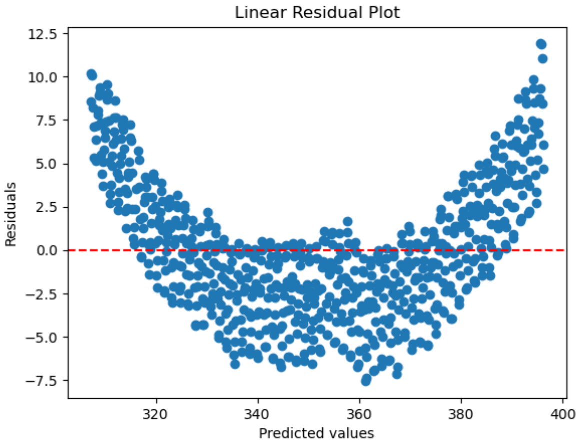
3.1:

(a)

The first design choice we make is how we deal with missing data. For this specific data, I decided to simply omit these data points. The reasoning is because later in the question, we look for seasonal trends and by doing a simple imputation we could be disturbing those trends. After running Linear Regression we get that $\hat{\alpha}_2 = 1.5239544$ and $\hat{\alpha}_1 = 306.83060094$. This gives us the following plot:



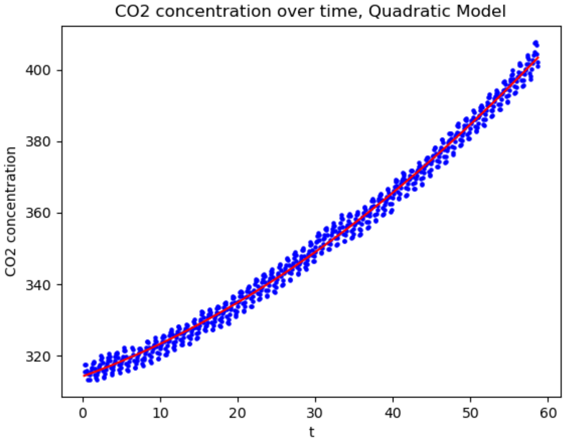
along with the following residual plot: that yields an R^2 of 0.976. While the fitted line does not look like a



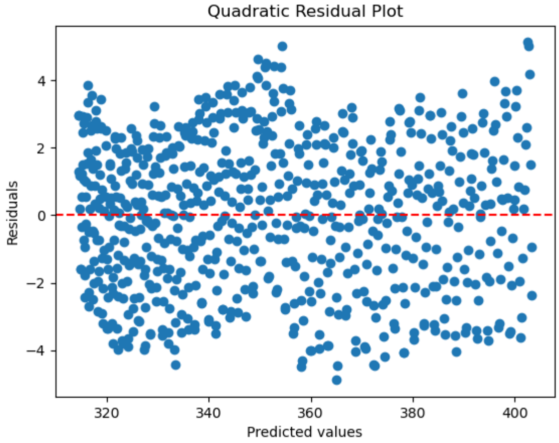
terrible approximation, in the residual plot we clearly there are some relationships we are not capturing with an affine function due to the parabolic shape of the residuals. This suggests we need to explore a more expressive hypothesis class.

(b)

Therefore, the natural next thing to try would include a polynomial transformation of our independent variable to capture the nonlinear relationship. Linear Regression on the transformed data gives us $\hat{\beta}_2 = 0.78335788$, $\hat{\beta}_3 = 0.012517$, and $\hat{\beta}_1 = 314.23913072$. This gives us the following plot:



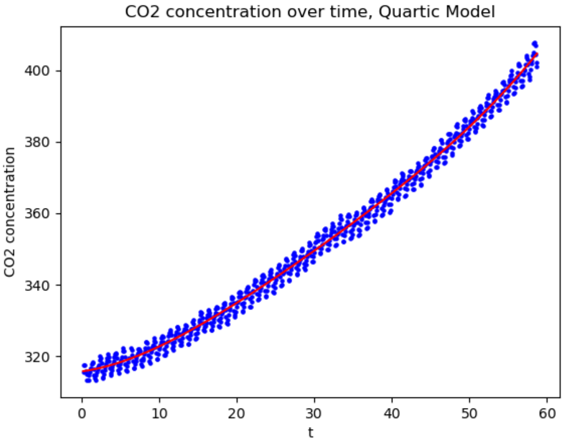
and the following residual plot:



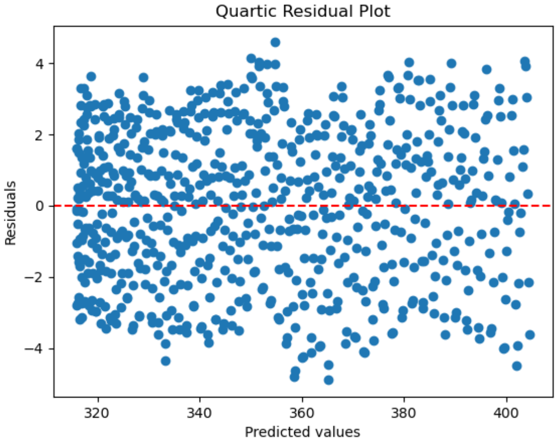
with an R^2 of 0.991. Based on the improved R^2, and the more normal distributed residuals this model seems to have captured the relationship between time and CO2 much better. However, it is still not perfect, so a natural follow up would be to increase the degree again.

(c)

We increase to degree 4 here and Linear Regression on this data gives us $\hat{\gamma}_2 = 0.2946$, $\hat{\gamma}_3 = 0.0489$, $\hat{\gamma}_4 = -0.0009484$, $\hat{\gamma}_5 = 0.00000798$, and $\hat{\gamma}_1 = 315.76663046$. This gives us the following plot:



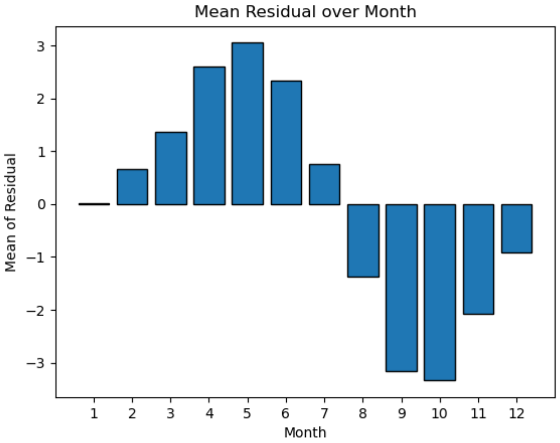
and the following residuals plot:



with a slightly improved R^2 of 0.993. While there are slight numerical improvements compared to the Quadratic Case, the benefit seems negligible. One possible way to select the order of your model could be similar to how we select the number of clusters in k-means with the elbow method. We could plot the R^2 of a sequence of a number of different orders and once the change in performance from one to the next no longer greatly improves, we can stop there.

(d)

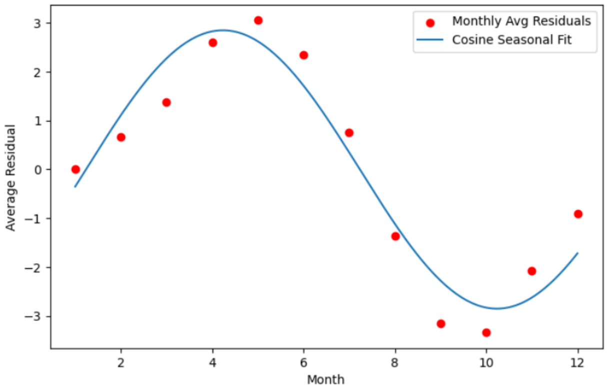
Averaging the residuals over all 12 months gets us the following bargraph:



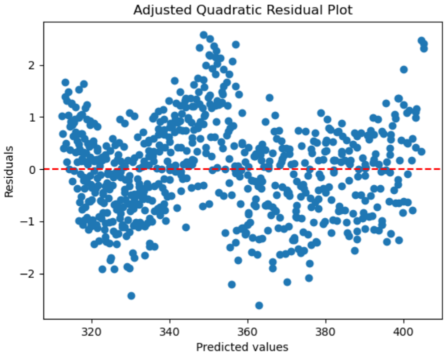
This suggests there is some cyclical relationship through each year that we are not capturing in our model. To address, we now fit a sinusoidal curve to these means be our P_i term. Python packages like scipy have this already implemented, so we get the following function:

$$P(x) = -0.003 + 2.273 \sin(\frac{2\pi x}{12}) - 1.716 \cos(\frac{2\pi x}{12})$$

which looks like this when plotted:



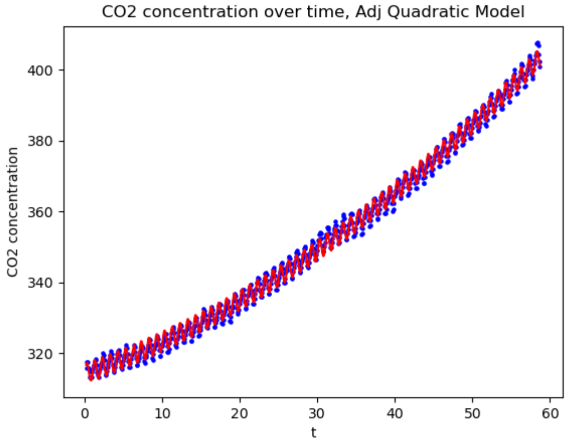
Combining the periodic function with out previous guesses give us the following residual plot:



While this looks a bit less normal, we get the highest R^2 so far of 0.998. In addition, in magnitude, these residuals are smaller than before. This provides strong evidence that this addition to our model helped.

(e)

Plotting the adjusted guesses gives us this plot:



A closer look at CO2 concentrations over time reveals that, alongside seasonal fluctuations, there is a pronounced long-term upward trend. Since 1958, CO2 levels have steadily risen, with an annual cycle superimposed on this overall increase.