MATH 4334: Mathemaical Modeling (#HW1b)

Reece Iriye

September 18, 2023

Problem 1: Suppose we are looking at data sets $\{x_i, y_i\}$ that seem to display proportionality – i.e. they seem to be well-described by a model of the form:

$$y = Ax$$

(a) write down an expression for the residual components $\{r_i\}$ associated with this model

$$\begin{aligned}\{r_i\} &= \{y_i - f(x_i)\}\\ &= \{y_i - Ax_i\}\end{aligned}$$

(b) by minimizing the sum of squares $E_2^2(A)$ of the residual, obtain a general formula for A.

$$E_2^2(A) = \sum_{i=1}^{n} (y_i - Ax_i)^2$$

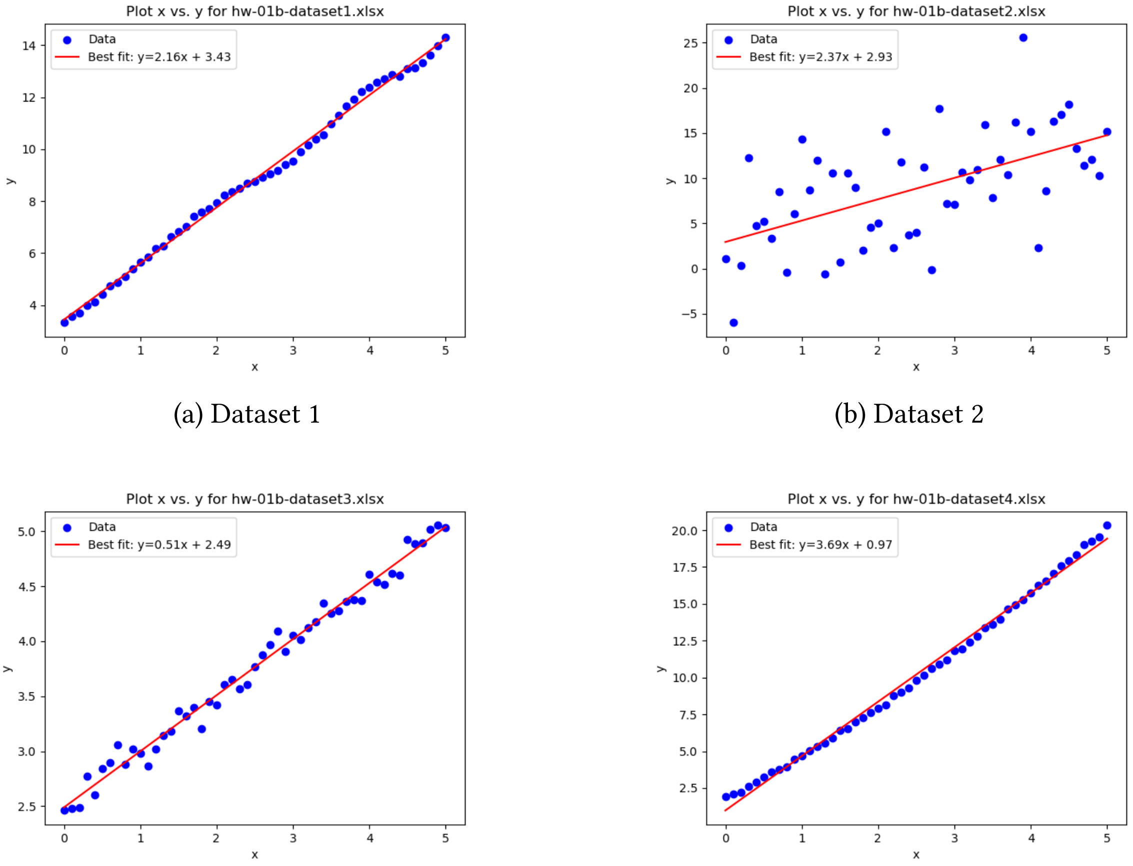
$$\begin{aligned}
&\implies \frac{dE_2^2}{dA} = \sum_{i=1}^{n} \frac{d}{dA}(y_i - Ax_i)^2 = 0\\
&\implies \sum_{i=1}^{n} 2(y_i - Ax_i)(-x_i) = 0\\
&\implies \sum_{i=1}^{n} Ax_i^2 - x_i y_i = 0\\
&\implies \sum_{i=1}^{n} Ax_i^2 - \sum_{i=1}^{n} x_i y_i = 0\\
&\implies \sum_{i=1}^{n} Ax_i^2 = \sum_{i=1}^{n} x_i y_i\\
&\implies A\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i\\
&\implies A = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

(c) apply your formula to the data set shown below, to determine A. You may leave A as a fraction.

$$\begin{aligned} A &= \frac{\sum_{i=1}^{6} x_i y_i}{\sum_{i=1}^{6} x_i^2} \\ &= \frac{(0)(0.1) + (1)(0.4) + (2)(0.6) + (3)(1.0) + (4)(1.1) + (5)(1.6)}{(0)^2 + (1)^2 + (2)^2 + (3)^2 + (4)^2 + (5)^2} \\ &= \frac{17}{55} \end{aligned}$$

Problem 2: For each of the four files of the form 'hw-01b-datasetX.xlsx':

(a) plot the data together with a best-fit linear model of the data.



(a) Dataset 1

(b) Dataset 2

(c) Dataset 3

(d) Dataset 4

Figure 1: *Plots of datasets with best-fit linear models*

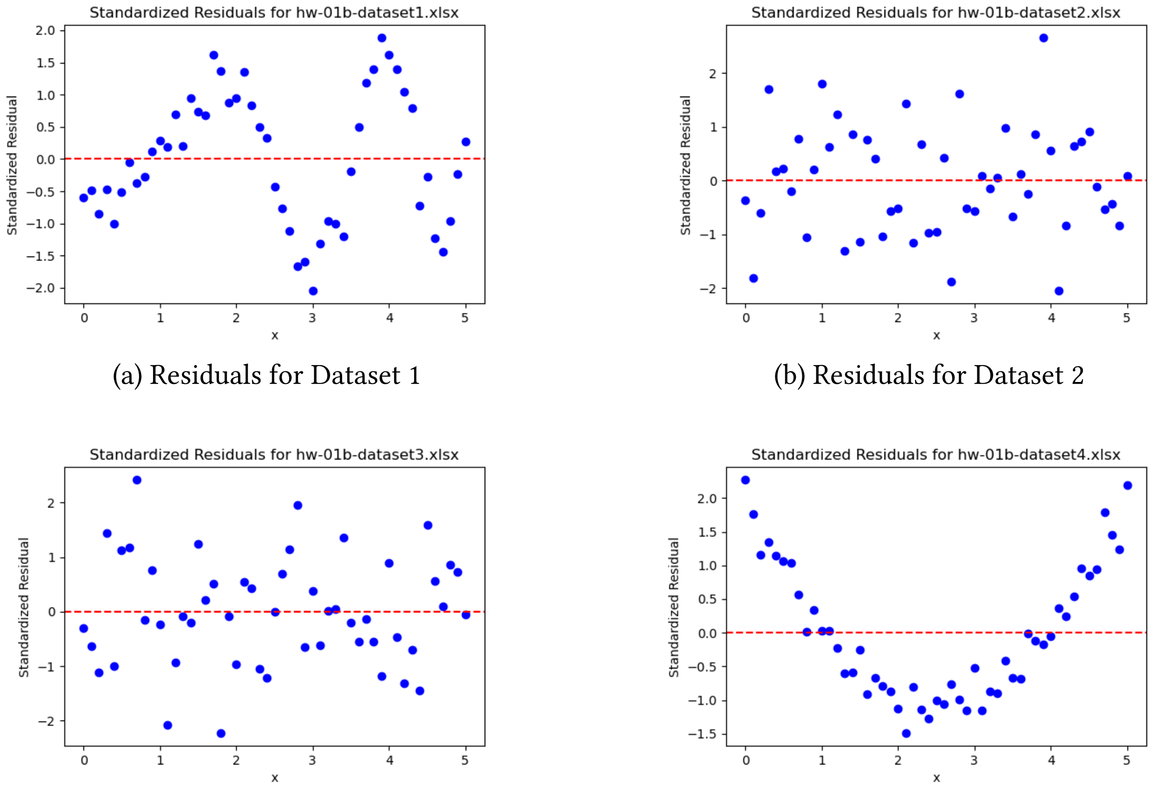
(b) calculate the UVR/EVR for the model, and plot the standardized residual.

For hw-01b-dataset1.xlsx
UVR: 0.0035
EVR: 0.9965

For hw-01b-dataset2.xlsx
UVR: 0.6725
EVR: 0.3275

For hw-01b-dataset3.xlsx
UVR: 0.0137
EVR: 0.9863

For hw-01b-dataset4.xlsx
UVR: 0.0055
EVR: 0.9945



(a) Residuals for Dataset 1

(b) Residuals for Dataset 2

(c) Residuals for Dataset 3

(d) Residuals for Dataset 4

Figure 2: *Standardized Residuals for the datasets*

(c) discuss whether any of the data points might be considered outliers

Dataset 1:

For dataset 1, I would not consider any of the points to be obvious outliers. In Figure 1(a), the points seem to follow a the same trend oscillating around the linear trend line with only some minor deviation from it where x is between around 0 and 2. In Figure 2(a), the trend I described about the points mostly oscillating around the trend line can be seen as well with the residual calculations, and there is not a clear singular point that breaks away from this trend.

Dataset 2:

For dataset 2, the data is more noisy, and I believe there is definitely at least 1 outlier that stands out, with a couple other points in question as well. The most notable point I am referring to is at $x = 3.9$, where $y \approx 25.57$ and the standardized residual is approximately 2.66. This point can be most notably seen in Figure 1(b), and it stands out in the standardized residual plot in Figure 2(b). Figure 2(b) also showcases the prominence of the 3 points with the points associated with three negative standardized residuals that have the largest order of magnitude, but I won't label them as outliers because there are some positive standardized residuals associated with points that have similar standardized residual magnitudes as well.

Dataset 3:

For dataset 3, the data is not as nearly noisy as in dataset 2 as we can see when comparing Figure 1(c) with Figure 1(b). However, the points are also not as close together as in Figure 1(a). When looking at Figure 2(c), I identify $x = 0.7$ and $y \approx 3.06$ as a significant outlier that can be visibly seen with respect to the entire dataset, despite the fact that the data as a whole lies much closer to the line of best fit when compared to Figure 1(b)'s line of best fit. The standardized residual is 2.42 for the point at $x = 0.7$ and $y \approx 3.06$. Given the plot of the residuals in Figure 2(c), I also believe that the points at the data points (1.8, 3.2), (1.1, 2.86), and (2.8, 4.1) are outliers, as their standardized residuals have a magnitude of 2.23, 2.08, and 1.96 respectively, with the next highest standard residual magnitude being 1.59, which is dramatically different.

Dataset 4:

For dataset 4, I do not believe there are any significant outliers in the data. My reasoning for this observation is because in Figure 2(d), we can see that the data follows a very specific pattern that is not linear. Regardless, the linear model still performs well on the data with an EVR of 0.9945, but because of this lack of randomness in the residuals, I do not believe that data points with higher standardized residual magnitudes are outliers. They fall directly into the observed trend.

(d) characterize the "magnitude / structure quadrant" of the residual

Dataset 1:

Figure 1(a) and figure 2(a) suggest that the data in dataset 1 falls under the high structure / small UVR quadrant. A pattern exists in the data that does not follow a linear relationship, but the data oscillates and hugs the line of best fit extremely closely to the point where the model's performance depends on the context of the problem.

Dataset 2:

Figure 1(b) and figure 2(b) suggest that the data in dataset 2 falls under the low structure / large UVR quadrant. The data does not follow any sort of clear pattern in the residual plot, which makes it have low structure, but the UVR is high.

Dataset 3:

Figure 1(c) and figure 2(c) suggest that the data in dataset 3 falls under the low structure / small UVR quadrant. The UVR calculation is small at 0.0137, and figure 2(c) showcases how the model does not have a distinct non-linear shape.

Dataset 4:

Figure 1(d) and figure 2(d) suggest that the data in dataset 4 falls under the high structure / small UVR quadrant. A pattern exists in the data that does not follow a linear relationship, but the data is bowed in and can be seen to likely follow a pattern of a degree two polynomial. Nevertheless, the performance of the model depends on the context of the problem.

(e) describe whether you would seek a better model: why or why not?

Dataset 1:

Seeking a better model in this scenario would depend on the context of the problem. If this model was used to predict medical diagnoses, I would likely be unhappy with the performance of the model. However, if the purpose was to predict trends in the market, I believe this would be a phenomenal model.

Dataset 2:

Seeking a better model in this scenario would depend on the context of the problem, as this model follows the trend of the data but does not capture variance well at all. In weather forecasting, this model would be great, but in the context of medical treatments or predicting loan defaults, a smaller UVR would be preferred.

Dataset 3:

I believe that the model performance in this context is great. The UVR is small, no clear pattern exists in the residual plot, and the data appears to closely hug the line of best fit in Figure 1(c) with no clear deviations.

Dataset 4:

Seeking a better model in this scenario would depend on the context of the problem. If this model was used for quality engineering and control in a context like aerospace engineering, I would likely be unhappy with the performance of the model. However, if the purpose was to predict trends in the market, I believe this would be a phenomenal model.

Problems 3: In this problem we will obtain and explore some data related to climate change.

(a) Find and download data on CO2 levels for the years 1850-present. Then find and download data on the global mean temperatures for the years 1850-present.

(b) In the field of climate science, the convention is to express temperature as the "Global Mean Temperature Anomaly" (GMTA) which is the difference in degrees Celsius between the global mean temperature now from its average value over some reference period. Different data sources use different reference periods, but we will follow the convention of the IPCC and use the years 1850-1900 for this purpose. Thus, if your data source uses a different range, just compute the average temperature over the years 1850-1900, and subtract that number from our temperature data. In addition, we know that pre-industrial CO2 levels were about 280 ppm, so we will subtract 280 from our CO2 values to obtain "excess CO2" (ECO2).

(c) Create a plot with the ECO2 levels vs time on the left, and GMTA levels vs time on the right. Note that the data on GMTA is much noisier over time than are the data on ECO2 levels, because GMTA is affected by several important factors besides ECO2 (such as El Nino cycles)

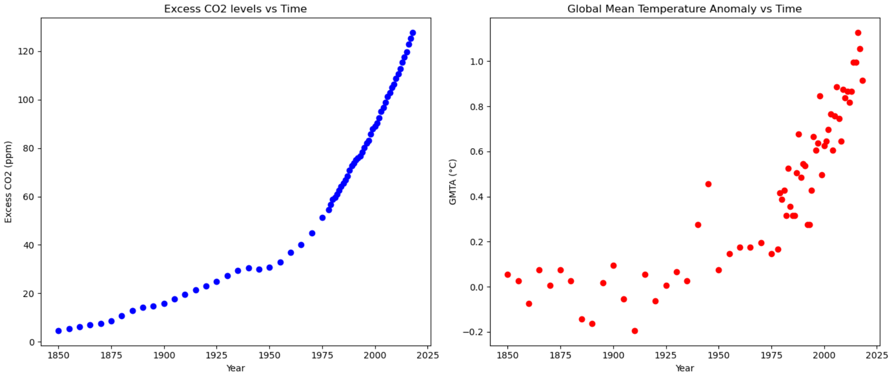


Figure 3: Time series of global CO2 concentrations and temperature anomalies.

(d) Next, eliminate the time variable by plotting the relationship between GMTA vs ECO2. The data is now in the same format as other exercises we have done in Unit 01. Does there appear to be a strong correlation, that could be fit by a simple, linear model?

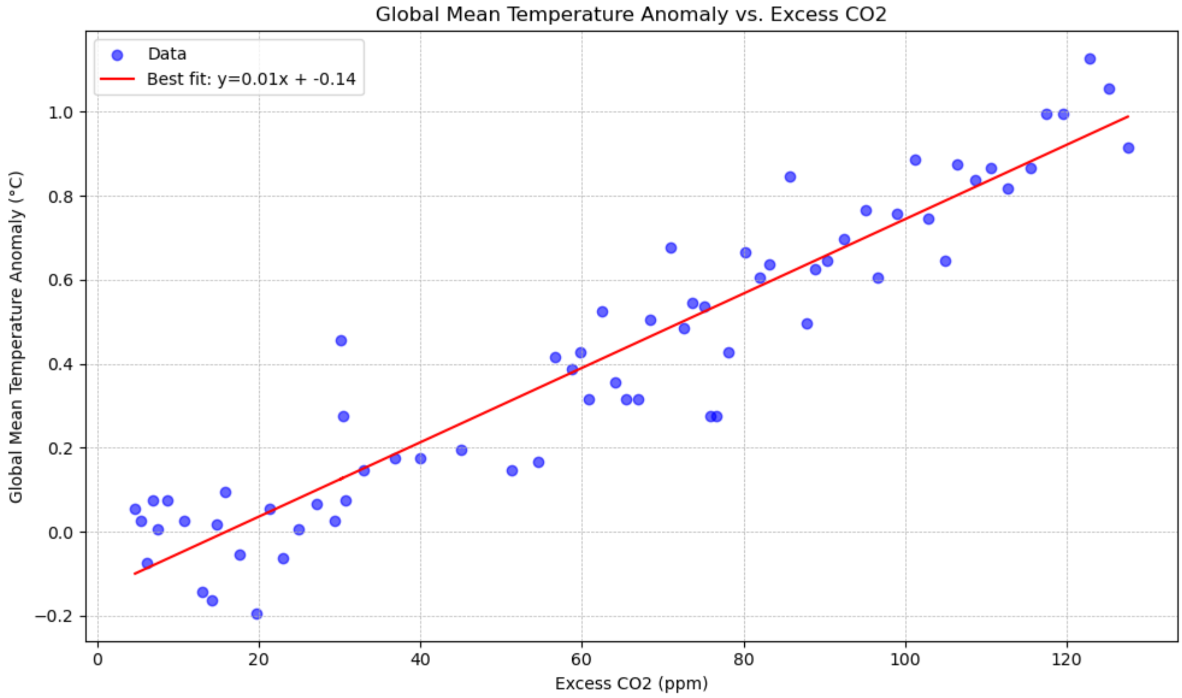


Figure 4: *Relationship between Global Mean Temperature Anomaly (GMTA) and Excess CO2 (ECO2)*

There appears to be a strong correlation that could be fit by a simple linear model in this case. There appears to be some noise in the data, but nevertheless, a linear trend can still be envisioned, and the correlation appears to be strong.

(e) For any years that look significantly different than their neighbors, determine whether unusual circumstances may have applied. If you find such combinations, you may remove those years from your data. Justify each removal.

The years that look off in the dataset are from 1945, 1993, and 1992, where GMTA equaled 0.456364, 0.276364, and 0.276364, while ECO2 equaled 30.10, 75.90, and 76.63. The deviated numbers in 1992 and 1993 can likely be explained by the Mt. Pinatubo explosion, which caused abnormal climate conditions to ensue, and also in 1945, the year where the atomic bombs were dropped in Hiroshima and Nagasaki. I will remove these pieces of data and refit a linear model to the data.

(f) Obtain a linear fit between ECO2 levels and the GMTA. State the fitted model and its coefficients. Discuss the value of R^2, and analyze the residual as we did in unit 1b.

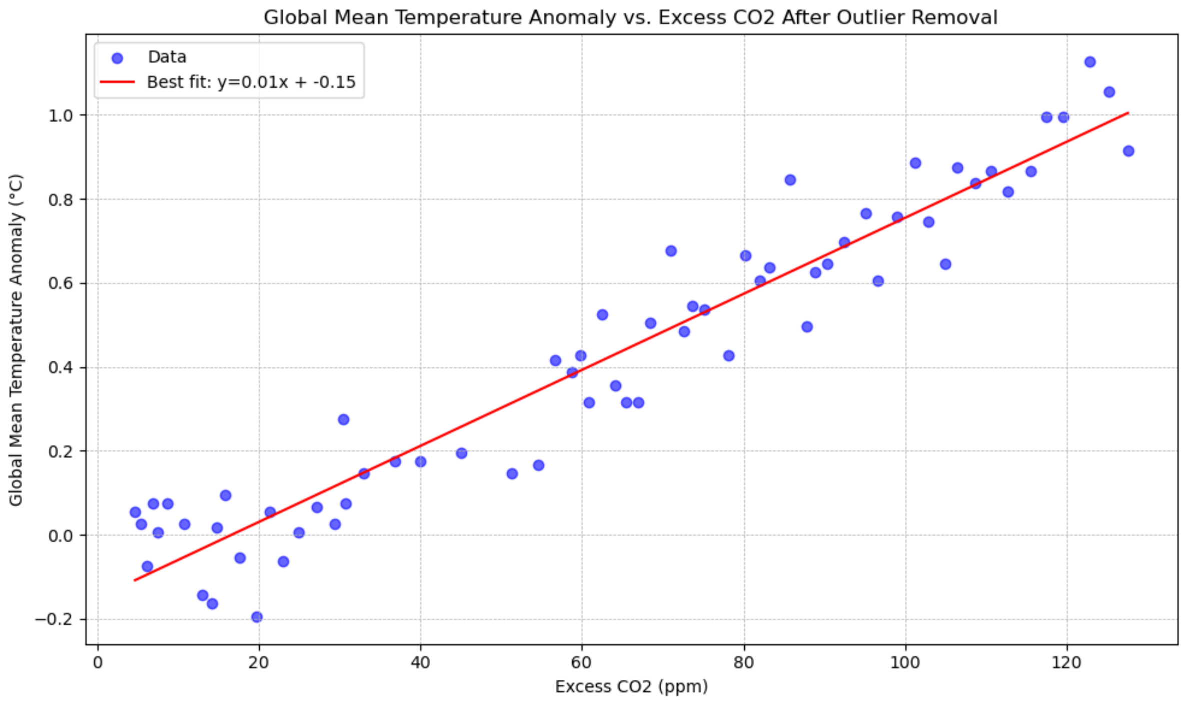


Figure 5: *Relationship between Global Mean Temperature Anomaly (GMTA) and Excess CO2 (ECO2) After Outlier Removal*

$$G\hat{M}TA = 0.00905 - 0.15067 \times ECO2$$

EVR (R^2): 0.9202
UVR: 0.0798

The R^2 value of 0.9202 suggests that a strong positive correlation between ECO2 and GMTA. While correlation does not necessarily imply causation, the strong linear relationship suggests that at least some sort of underlying commonality between the growth in ECO2 and the growth in GMTA should be explored and brought into question. The discussion is worth having based on Figure 5.

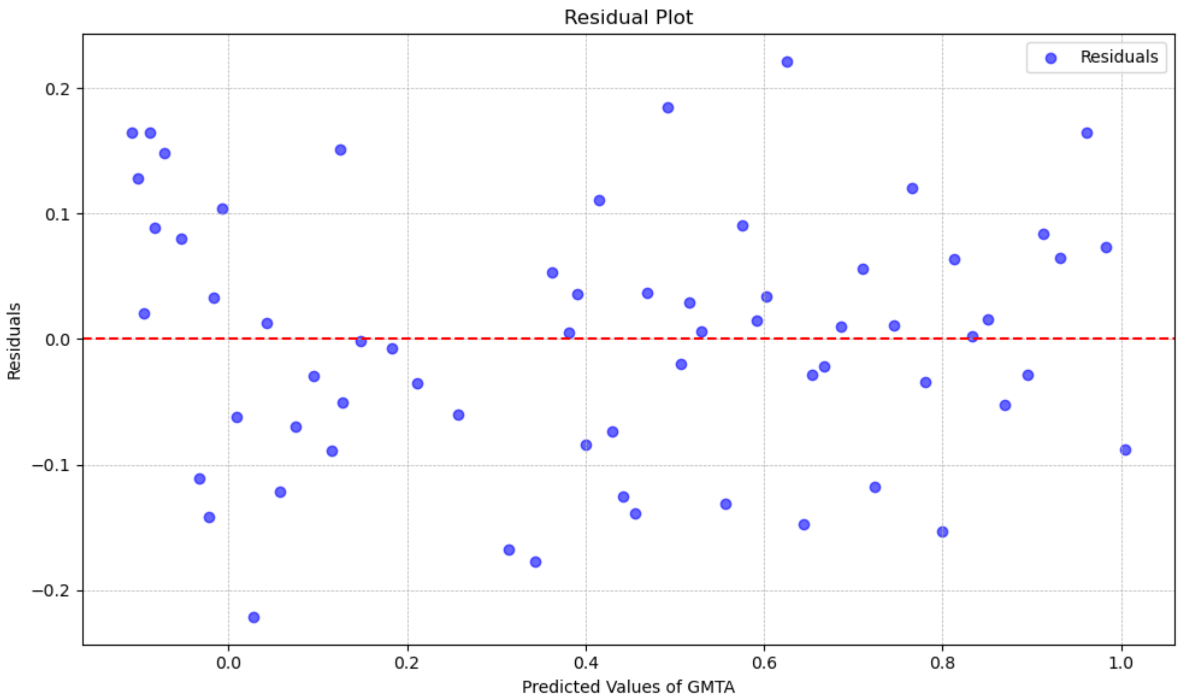


Figure 6: *Global Mean Temperature Anomaly (GMTA) Residual Plot Based on Linear Model in Figure 5*

The data in Figure 6 appears to be randomly scattered, meaning that the model likely follows a linear fit. The EVR is low, and there is no structure in the residuals, meaning that this linear model is strong in predicting GMTA based on CO2 levels.