STAT 223

Laboratory Activity #2 by Akbota Serikkyzy

MADfun

The function MADfun calculates the Mean Absolute Difference (MAD) of a given vector. The Python code for the function:

```
def MADfun(vec):
  if(len(vec)==0):
    return "Input should not be empty. Please try again!"
  elif (np.issubdtype(vec.dtype, np.number) == False):
    return "Input should be numbers. Please try again!"
  else:
    absdiff = np.sum(np.abs(vec[:, None] - vec))
    return absdiff / (len(vec) * (len(vec) - 1))
```

First, this function ensures that the input vector is not empty and numeric. After this, it calculates the absolute difference between every possible pair of elements in the vector, then sums all these differences. Finally, MAD is calculated by total sum of absolute differences is divided by the number of unique pairs.

The function worked well during testing, returning accurate results for vectors and catching errors for wrong inputs.

BAStd

BAStd (Standardized Batting Average) is a way to adjust a team's batting performance to compare it more fairly with other teams in the league. Positive values in BAStd show which teams are performing better than the average league performance. Negative values show which teams that are falling behind the league's overall performance. Zero indicates teams that are performing at the league's average.

In order to find the BAStd, we subtract the mean of BA from each team's BA, then divide that by the standard deviation of BA.

Statistics	Value
Mean	1.555×10^{-15}
Std	1.000
Min	-3.575
Q1	-0.736
Median	0.012
Q3	0.684
Max	2.552

Table 1: Statistics for BAStd

In the table 1, we can see that the mean and median are both close to zero, which suggests that the distribution is fairly symmetric. However, there are a few outliers. For example, the minimum value is -3.575, which means one team performed much worse than the league average. On the other hand, the maximum value is 2.552, showing that one team did significantly better than the others. The standard deviation is 1, which tells us that most teams' scores fall between -1 and 1. This means that most teams perform close to the average, with only a few standing out as much better or worse.

Relationship between HR and SLG

The scatterplot below shows different baseball statistics. HR (home runs) is the number of times a team hits the ball and scores by rounding all the bases in one play. SLG (slugging percentage) measures how powerful a team's hitters are by calculating the average number of bases they earn per at-bat. Division refers to the group a team belongs to within the league, such as East, Central, or West, and in the plot, each division is shown in a different color. WAR stands for wins above replacement, which estimates how many more wins a team has compared to a team made up of average players. In the scatterplot, WAR is shown by the size of each point, larger points meaning higher WAR.

From the figure 1, we can observe that SLG and HR are positively correlated, meaning that teams with higher home runs tend to also have higher slugging percentages. Additionally, most teams perform similarly regardless of their division, suggesting that division does not strongly affect performance in these areas. However, there are some divisions where teams perform well in SLG but poorly in HR. Despite this, in general, teams that perform better in HR also tend to perform well in SLG.

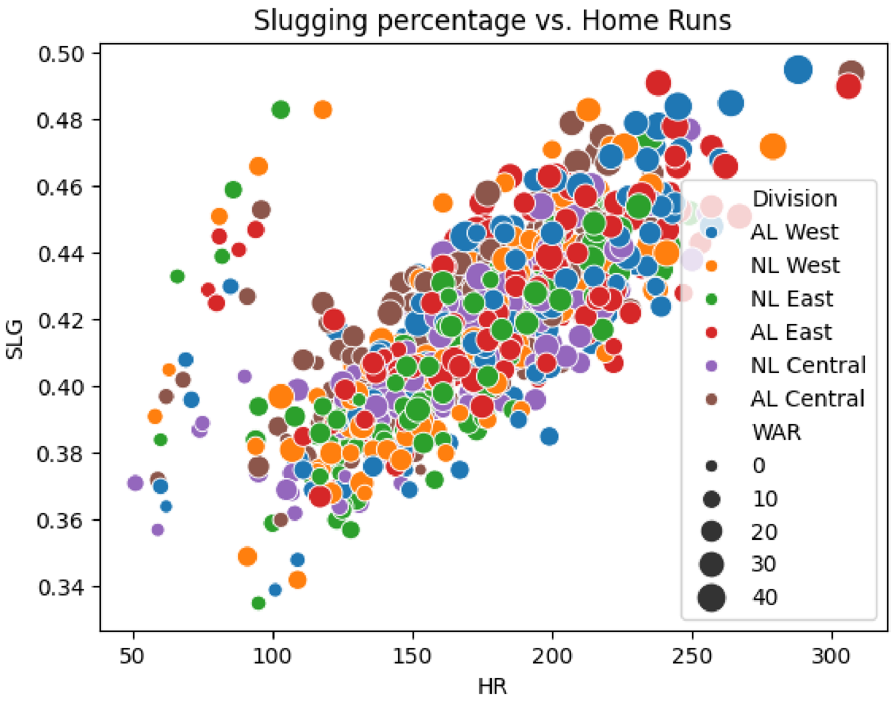

Figure 1: Relationship between HR and SLG

Relationship between wRC and WinPct

wRC (Weighted Runs Created) is an advanced baseball statistic that estimates the total number of runs a player contributes to their team. WinPct (winning percentage) shows how often a team wins. To explore the relationship between wRC and WinPct, we fit a linear regression model with wRC as the independent variable and WinPct as the dependent variable.

Variable	Coefficient	Standard Error	p-value
Intercept	-0.0345	0.021	0.102
wRC	0.0055	0.000	0.000

Table 2: Regression Results for WinPct and wRC

Statistic	Value
R-squared	0.447

Table 3: R-squared for the Regression Model (wRC and WinPct)

$$\text{The regression equation: WinPct} = -0.0345 + 0.0055 \times \text{wRC} \quad (1)$$

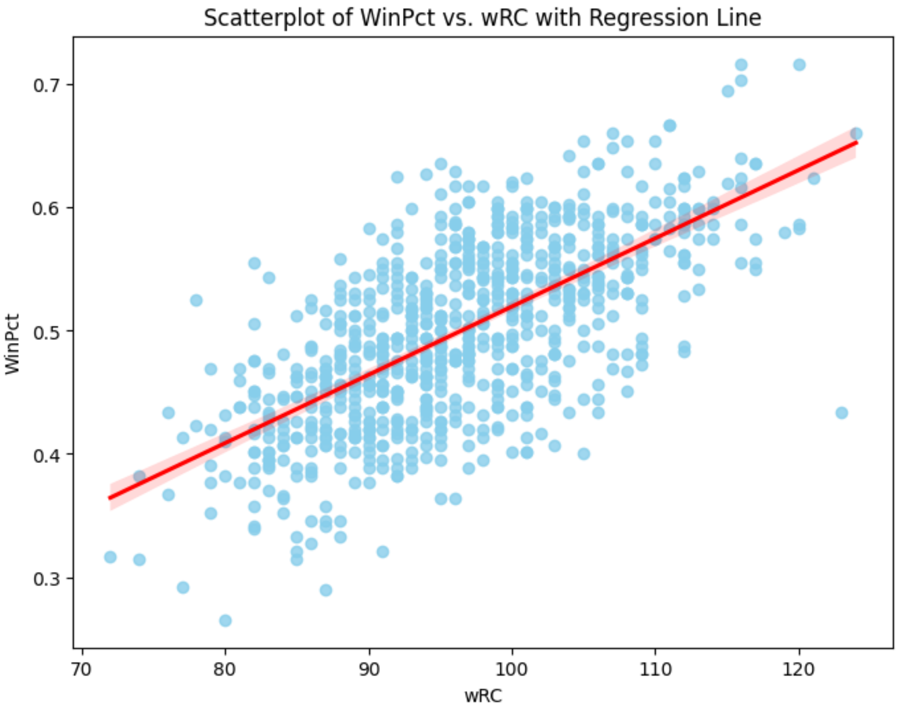

Figure 2: Relationship between wRC and WinPct

So, from the regression equation and figure 2, we can see that wRC and WinPct are positively correlated. This means that as a team's wRC increases, their winning percentage also tends to increase. The p-value for wRC is very close to zero, which tells us that this relationship is statistically significant. In other words, we have strong evidence that wRC is related to WinPct. On the other hand, the p-value for the intercept is 0.102 which means that the intercept is not statistically significant.

Looking at the regression results 2, the coefficient for wRC is 0.0055, which means that for each increase of one unit in wRC, we expect WinPct to increase by approximately 0.0055. The standard error for this estimate is very small, indicating that the coefficient has high precision.

The R-squared value is 0.447, which means that about 44.7% of the variation in WinPct can be explained by the wRC values. This is a strong relationship, especially since WinPct can also be influenced by many other factors not included in this model.

Overall, both the figure and the regression suggests that teams with higher wRC scores tend to have better win percentages.

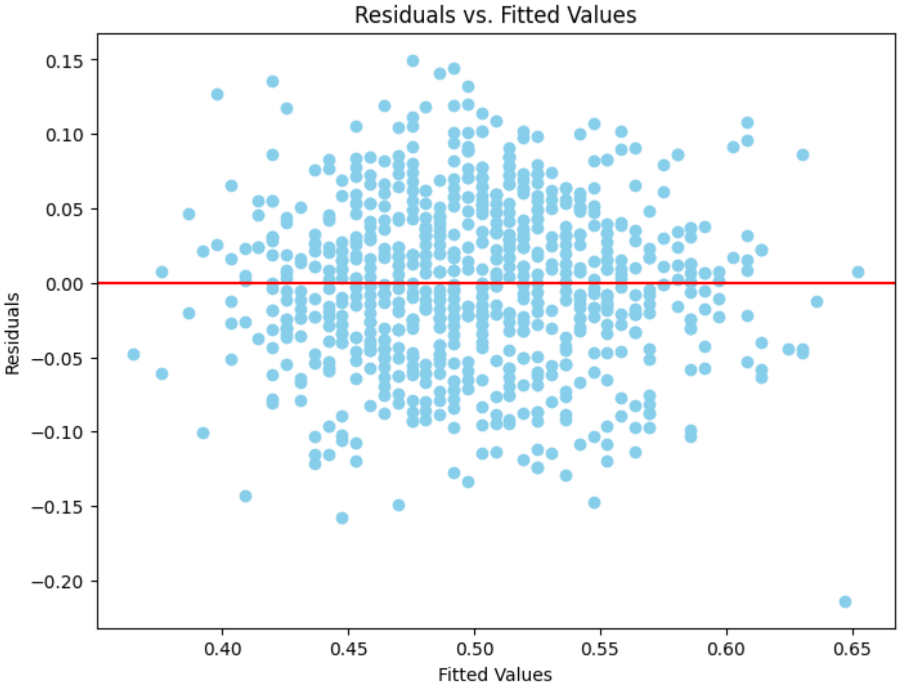

Figure 3: Residuals vs. Fitted Values

Figure 3 shows the residuals versus the fitted values. The points are spread out randomly around the red line, which means that the errors are evenly scattered and there is no strong pattern. This is a good sign since it suggests that the relationship between wRC and WinPct is likely linear and the model does not miss any major patterns.

Q-Q plot 4, the blue dots mostly follow the red line, which means that the residuals are pretty close to being normally distributed. There are a few small deviations at the ends, but overall it is normally distributed.

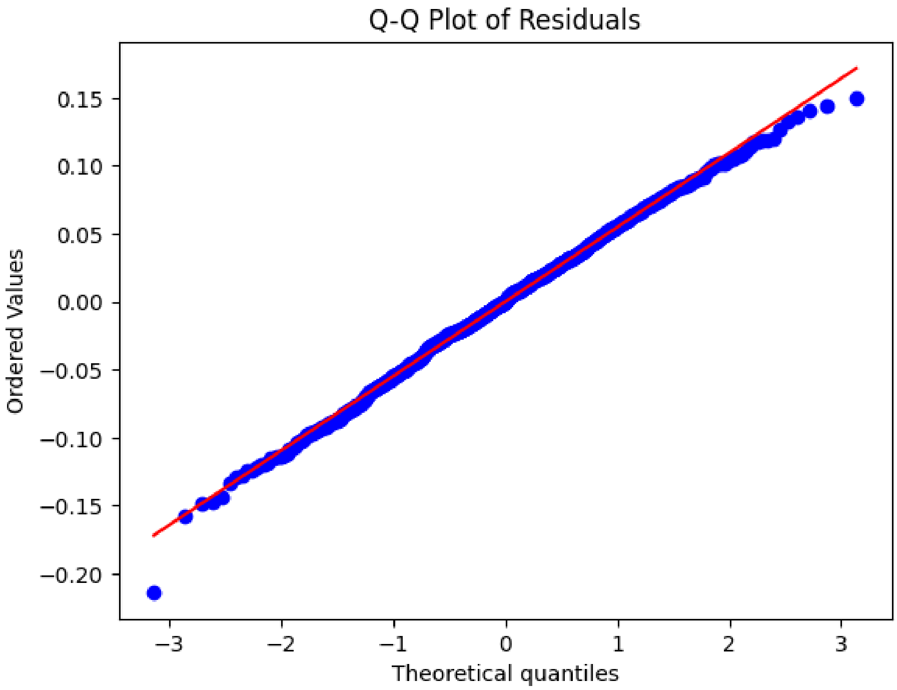

Figure 4: Q-Q plot

Relationship between HRA and SO

HRA, home runs allowed, and SO, strikeouts thrown, are both measures of pitching success. To explore the relationship between HRA and SO, we fit a linear regression model with HRA as the independent variable and SO as the dependent variable.

Variable	Coefficient	Standard Error	p-value
Intercept	636.5782	31.877	0.000
HRA	2.8717	0.184	0.000

Table 4: Regression Results for HRA and SO

Statistic	Value
R-squared	0.233

Table 5: R-squared for the Regression Model (HRA and SO)

$$\text{The regression equation: SO} = 636.5782 + 2.8717 \times \text{HRA} \tag{2}$$

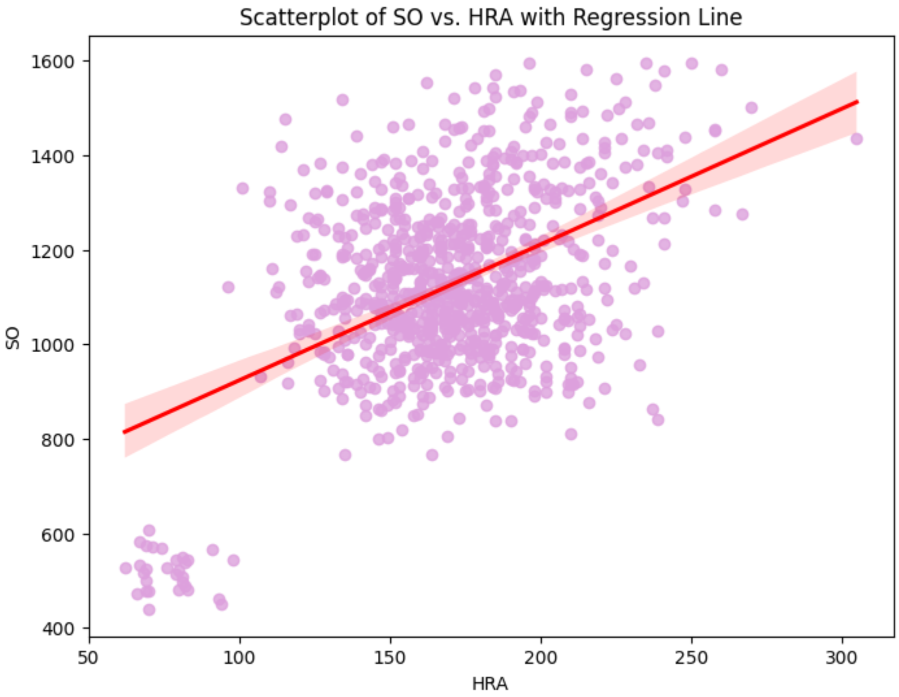

Figure 5: Relationship between HRA and SO

Looking at the regression equation and figure 5, we can see that HRA and SO are positively correlated. This means that as the number of home runs allowed increases, the number of strikeouts also tends to increase. The p-value is 0.000, which shows that this relationship is statistically significant. So, we have strong evidence that there is a real connection between HRA and SO.

According to the regression results in table 4, the coefficient for HRA is 2.8717. This means that for each additional home run allowed, we expect about 2.8717 more strikeouts. The standard error is small, which means this estimate is quite precise.

The R-squared value is 0.233, which means about 23.3% of the variation in strikeouts can be explained by home runs allowed. While this is not a very strong relationship, it still shows a meaningful connection.

Overall, both the scatterplot and the regression suggest that teams that allow more home runs also tend to record more strikeouts.

Looking at the figure 6, we can see that the values are fairly evenly spread, suggesting that the relationship between HRA and SO is likely to be linear. The histogram supports our earlier claim that HRA and SO have a positive relationship.

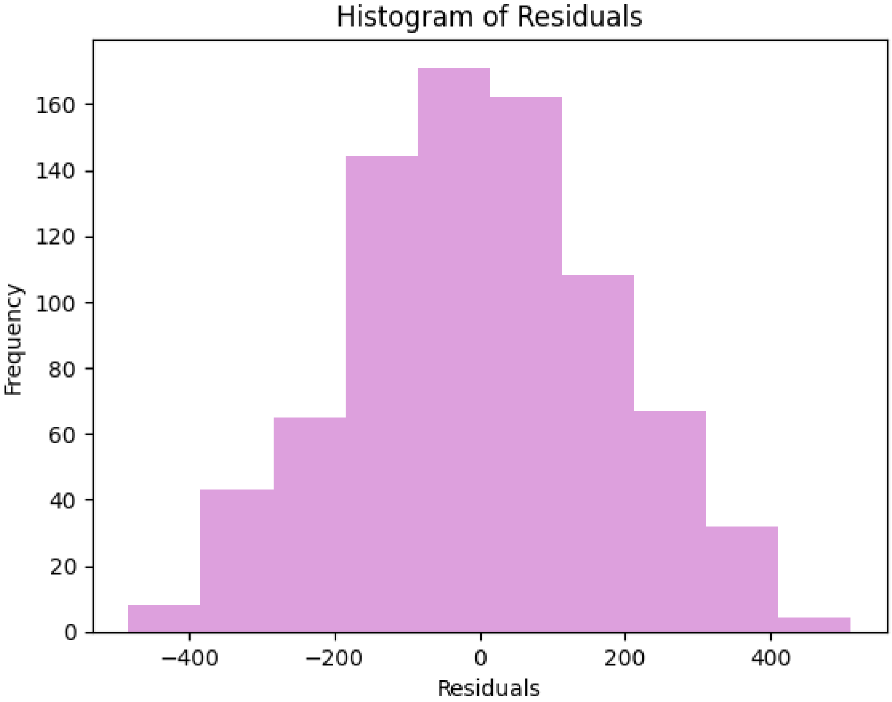

Figure 6: Histogram

The linear model shows a good fit for our data, and the residual analysis supports this linear relationship. However, the intercept value of 636.5782 creates a problem as it predicts SO levels around 637 when HRA is zero, which does not make much sense since all our actual SO measurements are much lower (between 100-300). This tells us that data points in the ranges of HRA (0-100) and SO (0-700) are way off from our best fit line. We can consider these as outliers that do not follow the overall trend. Even though our model works well for the main range of our data, we need to be careful with these extreme values.

Finally, the stated idea that ”the best pitchers tend to throw more strikeouts and give up fewer home runs” is not supported by our analysis. There could be a few reasons for this. First, our data focuses on team-level performance, not individual pitchers. That means the relationship might reflect team strategies or a mix of strong and weak pitchers, rather than individual pitcher skill. Second, even if two variables are positively related, it does not mean one causes the other since other factors might be influencing too.

Relationship between ERA and OPS

ERA, earned run average, is a measure of pitching success, with better pitchers having low ERAs. OPS is a measure of hitting success, with better hitters having high OPSs. To explore the relationship between ERA and OPS, we fit a linear regression model with OPS as the independent variable and ERA as the dependent variable.

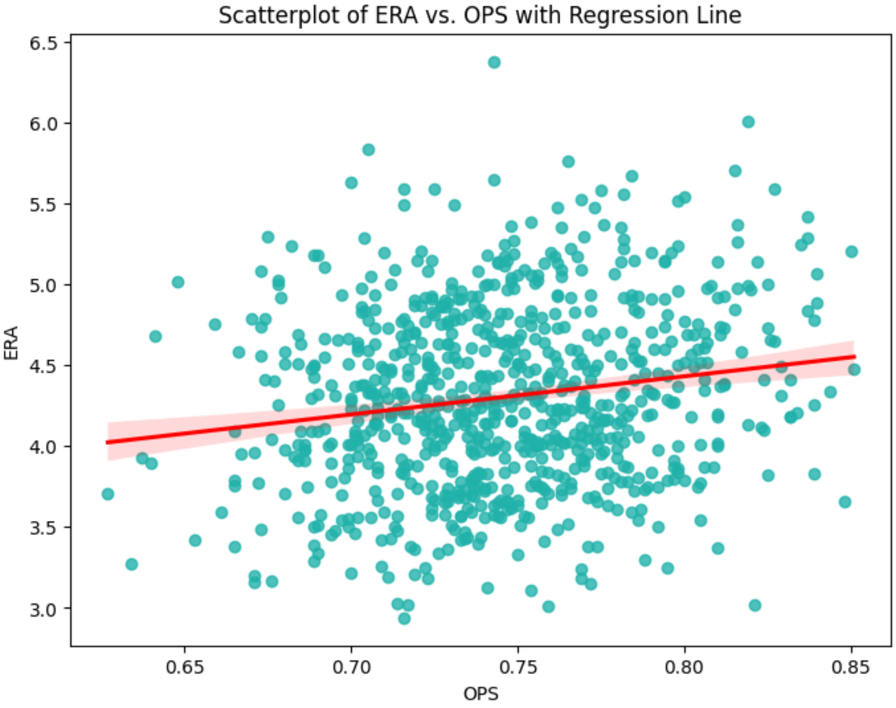

Figure 7: Relationship between ERA and OPS

$$\text{The regression equation: ERA} = 2.545 + 2.3588 \times \text{OPS} \tag{3}$$

Variable	Coefficient	Standard Error	p-value
Intercept	2.5450	0.364	0.000
OPS	2.3588	0.487	0.000

Table 6: Regression Results for ERA and OPS

Statistic	Value
R-squared	0.028

Table 7: R-squared for the Regression Model (HRA and SO)

Looking at the regression equation and figure 7, we can see that ERA and OPS are positively correlated, although the relationship is quite small. This means that as a team's OPS increases, its ERA also tends to increase slightly. The p-value is 0.000, which means the relationship is statistically significant.

In the regression results 6, the coefficient for OPS is 2.3588, which tells us that for every one-unit increase in OPS, ERA is expected to increase by about 2.3588. The standard error for the coefficient is 0.487, which shows how much the coefficient might vary.

The R-squared value is 0.028, meaning only 2.8% of the variation in ERA can be explained by OPS. This tells us that OPS does not strongly predict ERA.

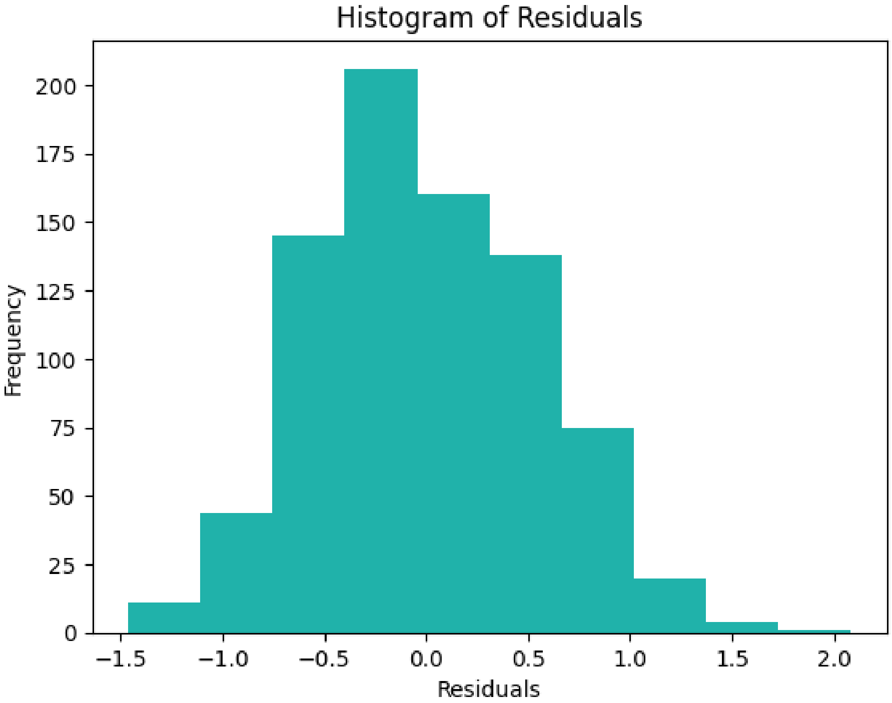

Figure 8: Histogram

Looking at the residual histogram in Figure 8, we can see that most residuals are centered around 0 and spread fairly evenly. This suggests that the residuals are approximately normally distributed, supporting the assumption of a linear relationship between ERA and OPS.