# Multiple Linear Regression (MLR)

**Step 1: Defining Target and Predictors**

**As shown in a heatmap, temperature clearly correlates with a lot of other parameters. Therefore, temperature will be the target of the MLR analysis (OLS Regression). The predictors are ‘p**H', 'DO', 'Conductivity', ‘Total\_N', 'Total\_P'. Year, Month and Season parameters are not included as they obviously correspond to temperature. Furthermore, the Year, Month and Season parameters are not in numerical format.

**Step 2: Analysing the OLS Regression Results**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

* **An R-squared value of 0.724, and an adjusted R-squared value of 0.723, tells us that 72.4% of the variance in temperature is explained by the model.**
* F-statistic score is 965.7, the model as a whole is statistically significant.
* All predictors except Total\_P are highly significant (p < 0.05)
* pH and DO are the most influential predictors
* Temperature tends to increase with pH and conductivity, but decrease with DO, Total\_N, and Total\_P.

There may be multicollinearity issues, which is expected as some predictors are correlated, which could affect stability of coefficients.

**Step 3: Residual Analysis**

**A diagram of blue dots

AI-generated content may be incorrect.**

* **No indication of heteroscedasticity as there is no funnel shape present**
* **No indication of non-linearity as there is no curved shape present**
* **There are only four distinct predicted fitted values: 5, 10, 15, 20**
* **There is a cluster of data points that do not appear randomly spread out, however this could be attributed to the data values being grouped into a small range of fitted values.**
* **This clustering is also likely due to the multicollinearity issues, as shown when VIF analysis is performed**

**Step 4: Variance Inflation Factor (VIF) Analysis**

**A screenshot of a graph

AI-generated content may be incorrect.**

* **The** VIF scores indicatesevere multicollinearity issues in the regression model
* pH, DO and Conductivity have extremely high VIF values
* **The predictors are highly correlated with each other, although this is expected with environmental parameters**
* **The** model has limited unique predictions, instead the predictors move together rather than independently
* **The model cannot reliably determine each variable's individual effect on the outcome**
* **Coefficient estimates will be unstable and unreliable**
* **Standard errors are inflated, making significance tests untrustworthy**
* **To address the extreme multicollinearity problem,** Principal Component Analysis (PCA) will be carried out.