# CP2403 - Assignment – Part 2 – Task 4: Multiple Regression

First Name:

Last Name:

|  |
| --- |
| **1: Data Selection** |
| * Data selected * Response variable - 'T\_degC' * Explanatory variable 1 - 'Salnty', * Explanatory variable 2 - 'O2ml\_L', * Explanatory variable 3 - 'PO4uM' |
| **2: Scatter plots between each explanatory variable and response variable** |
| R-value for Salnty vs T\_degC: -0.4886151763337268    R-value for O2ml\_L vs T\_degC: 0.828930796310497    R-value for PO4uM vs T\_degC: -0.9067362996016157    R-value for T\_degC vs T\_degC: 1.0 |
| **3: Summary of your pre-testing plan - List possible candidate combinations of individual regression models to compose a multiple regression model and your justification (e.g. why did you decide to apply such combination strategy?)** |

|  |
| --- |
| # Combination 1: Salinity and O2 concentration  model\_1 = sm.OLS(df[response\_variable], sm.add\_constant(df[['Salnty', 'O2ml\_L']])).fit()  **Justification:** Salinity and O2 concentration are commonly known factors influencing water temperature (T\_degC). Salinity affects water density, while O2 concentration can influence thermal properties. Combining these two variables captures their potential joint impact on temperature.  # Combination 2: O2 concentration and Phosphate concentration  model\_2 = sm.OLS(df[response\_variable], sm.add\_constant(df[['O2ml\_L', 'PO4uM']])).fit()  **Justification:**  O2 concentration and phosphate levels can both influence water quality and, consequently, temperature. The inclusion of phosphate concentration allows us to explore whether nutrient levels contribute to temperature variations.  # Combination 3: Salinity, O2 concentration, and Phosphate concentration  model\_3 = sm.OLS(df[response\_variable], sm.add\_constant(df[['Salnty', 'O2ml\_L', 'PO4uM']])).fit()  **Justification:**  This combination extends the analysis by including all three variables simultaneously. Salinity, O2 concentration, and phosphate levels may have interactive effects on temperature, and including all three provides a more comprehensive understanding of their combined impact. |
| **4: Pre-testing Regression analysis results (for each candidate (multiple) regression model)** |
| pre-testing Regression analysis results:  **Model 1:**  OLS Regression Results  ==============================================================================  Dep. Variable: T\_degC R-squared: 0.693  Model: OLS Adj. R-squared: 0.693  Method: Least Squares F-statistic: 4.717e+05  Date: Sat, 13 Jan 2024 Prob (F-statistic): 0.00  Time: 10:01:55 Log-Likelihood: -9.3303e+05  No. Observations: 418824 AIC: 1.866e+06  Df Residuals: 418821 BIC: 1.866e+06  Df Model: 2  Covariance Type: nonrobust  ==============================================================================  coef std err t P>|t| [0.025 0.975]  ------------------------------------------------------------------------------  const -18.9825 0.281 -67.656 0.000 -19.532 -18.433  Salnty 0.7012 0.008 85.986 0.000 0.685 0.717  O2ml\_L 1.7547 0.002 786.260 0.000 1.750 1.759  ==============================================================================  Omnibus: 736550.810 Durbin-Watson: 0.307  Prob(Omnibus): 0.000 Jarque-Bera (JB): 47099525504.054  Skew: 11.058 Prob(JB): 0.00  Kurtosis: 1645.702 Cond. No. 2.75e+03  ==============================================================================  Notes:  [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  [2] The condition number is large, 2.75e+03. This might indicate that there are  strong multicollinearity or other numerical problems.  **Model 2:**  OLS Regression Results  ==============================================================================  Dep. Variable: T\_degC R-squared: 0.850  Model: OLS Adj. R-squared: 0.850  Method: Least Squares F-statistic: 1.187e+06  Date: Sat, 13 Jan 2024 Prob (F-statistic): 0.00  Time: 10:01:55 Log-Likelihood: -7.8265e+05  No. Observations: 418824 AIC: 1.565e+06  Df Residuals: 418821 BIC: 1.565e+06  Df Model: 2  Covariance Type: nonrobust  ==============================================================================  coef std err t P>|t| [0.025 0.975]  ------------------------------------------------------------------------------  const 24.6181 0.029 840.507 0.000 24.561 24.675  O2ml\_L -1.2260 0.004 -279.169 0.000 -1.235 -1.217  PO4uM -5.9246 0.009 -674.667 0.000 -5.942 -5.907  ==============================================================================  Omnibus: 639653.930 Durbin-Watson: 0.291  Prob(Omnibus): 0.000 Jarque-Bera (JB): 44378481048.512  Skew: -7.936 Prob(JB): 0.00  Kurtosis: 1597.610 Cond. No. 55.6  ==============================================================================  Notes:  [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  **Model 3:**  OLS Regression Results  ==============================================================================  Dep. Variable: T\_degC R-squared: 0.858  Model: OLS Adj. R-squared: 0.858  Method: Least Squares F-statistic: 8.435e+05  Date: Sat, 13 Jan 2024 Prob (F-statistic): 0.00  Time: 10:01:55 Log-Likelihood: -7.7128e+05  No. Observations: 418824 AIC: 1.543e+06  Df Residuals: 418820 BIC: 1.543e+06  Df Model: 3  Covariance Type: nonrobust  ==============================================================================  coef std err t P>|t| [0.025 0.975]  ------------------------------------------------------------------------------  const -4.3704 0.192 -22.782 0.000 -4.746 -3.994  Salnty 0.8476 0.006 152.809 0.000 0.837 0.858  O2ml\_L -1.0975 0.004 -251.956 0.000 -1.106 -1.089  PO4uM -5.9739 0.009 -698.495 0.000 -5.991 -5.957  ==============================================================================  Omnibus: 559234.376 Durbin-Watson: 0.275  Prob(Omnibus): 0.000 Jarque-Bera (JB): 22850521557.938  Skew: -5.959 Prob(JB): 0.00  Kurtosis: 1147.233 Cond. No. 2.77e+03  ==============================================================================  Notes:  [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.  [2] The condition number is large, 2.77e+03. This might indicate that there are  strong multicollinearity or other numerical problems. |
| **5: Pre-testing Regression equation/line (for each candidate (multiple) regression model)** |
| **Pre-testing Regression equation/line:**  **Model 1:**  T\_degC = -18.98248767500521 + 0.7012372595170413 \* Salnty + 1.7547095749600117 \* O2ml\_L  **Model 2:**  T\_degC = 24.618076548733523 + -1.225999027198047 \* O2ml\_L + -5.9245607498703325 \* PO4uM  **Model 3:**  T\_degC = -4.370432329791711 + 0.8475680482039132 \* Salnty + -1.0975161402112414 \* O2ml\_L + -5.973922840472736 \* PO4uM |
| **6: Q-Q plot for each candidate (multiple) regression model** |
|  |
|  |
| **7: Conclusion from Q-Q plots** |
|  |
| **8: Residual Plot for each candidate model** |
| For each candidate model:   * Standardised Residual plot          * percentage of observations over 2 standardized deviation   Model 1 - Percentage over 2 std deviations: 5.186665520600538%  Model 2 - Percentage over 2 std deviations: 4.119152675109354%  Model 3 - Percentage over 2 std deviations: 4.428113002120222%   * percentage of observations over 2.5 standardized   Model 1 - Percentage over 2.5 std deviations: 2.622103795388994%  Model 2 - Percentage over 2.5 std deviations: 2.2381716425037723%  Model 3 - Percentage over 2.5 std deviations: 2.4979466315206387% |
| **9: Conclusion from Standardised Residual plots** |
|  |
| **10: Conclusion Overall** |
| * Can you select one best model among your candidate models? * Justify your selection.   **Model 1:**  - R-squared: 0.693  - Variables: Salinity, O2ml\_L  - Interpretation: This model explains approximately 69.3% of the variance in temperature. Both Salinity and O2ml\_L are significant predictors of temperature.  **Model 2:**  - R-squared: 0.850  - Variables: O2ml\_L, PO4uM  - Interpretation: This model shows an improved R-squared of 85%, indicating a better fit than Model 1. O2ml\_L and PO4uM are significant predictors of temperature.  **Model 3:**  - R-squared: 0.858  - Variables: Salinity, O2ml\_L, PO4uM  - Interpretation: Model 3 has the highest R-squared at 85.8%, suggesting a slightly better fit than Model 2. All three variables (Salinity, O2ml\_L, PO4uM) are significant predictors.  **Conclusion:**  Considering the higher R-squared values and the inclusion of all relevant variables, Model 3 appears to be the best among the candidate models. It provides the best balance between explanatory power and simplicity. Therefore, I would select Model 3 as the preferred multiple regression model for predicting temperature based on Salinity, O2 concentration, and Phosphate concentration. |