Topic: Regression Techniques for Ecological Modeling

Lesson Description:

This lesson introduces students to regression as a core machine learning approach for modeling and predicting continuous environmental variables. The lesson emphasizes both theoretical and applied perspectives, distinguishing between linear and non-linear regression techniques relevant to environmental and ecological studies. Students will explore the complete regression workflow, including data preparation, feature engineering, model training, evaluation using RMSE, and mitigation of overfitting and underfitting.
The selected use case focuses on predicting dissolved oxygen concentration from water temperature and related environmental variables in an estuarine shellfish farming area. This example illustrates how regression models can support real-world ecological monitoring and early warning systems for aquatic stress conditions.

Learning Outcomes:

After completing this unit, students will be able to:
1. Define regression problems and distinguish them from other machine learning tasks.
2. Explain the principles of simple and multiple linear regression and the logic of RMSE evaluation.
3. Develop and evaluate regression models using environmental datasets.
4. Identify and address common issues such as noise, multicollinearity, underfitting, and overfitting in regression problems with ecological data.
5. Interpret model outputs in ecological and environmental contexts.

Lecture Content

Introduction

Regression techniques are essential tools for modeling continuous environmental variables such as temperature, salinity, and pollutant concentrations. This section explains the difference between regression and classification tasks, introduces the concept of Root Mean Squared Error (RMSE), and discusses the importance of regression models for understanding and predicting ecological phenomena.

Context and motivation

Regression techniques are foundational in modeling many continuous-response environmental phenomena. In ecological and environmental sciences, one frequently seeks to predict a continuous variable (e.g., pollutant concentration, species biomass, river discharge, temperature anomaly) from a set of predictor variables (meteorology, landscape variables, catchment characteristics, land-use, etc.). Regression thus provides a bridge between observational/environmental data and explanatory, predictive modeling

— and sits at the crossroads of statistical ecology, hydrology, climatology, and machine learning.
Examples of regression problems in ecology and environmental sciences:

- Estimating the concentration of nitrogen dioxide (NO_2) or particulate matter ($PM_{2.5}$) in urban/industrial settings based on meteorological inputs, traffic/land-use variables, and atmospheric conditions.
- Modeling species abundance (or biomass) over time or space in response to environmental drivers such as temperature, precipitation, soil nutrients, habitat fragmentation, and human disturbance.
- Predicting hydrological responses like daily river discharge from rainfall, air temperature, humidity, antecedent soil moisture, and catchment attributes.
- Forecasting climate-related anomalies (e.g., temperature anomalies) or ecosystem responses (carbon fluxes, greenhouse-gas emissions) using continuous regression frameworks.

Regression vs. Classification Tasks

It is essential to distinguish regression from classification, as both fall under the family of supervised learning problems but serve different predictive objectives.
In classification, the target variable y is categorical — representing discrete labels or classes (e.g., “polluted” vs. “clean” water samples, or “presence” vs. “absence” of a given species). In contrast, in regression, y is continuous, meaning that its values can vary on a numerical scale without inherent class boundaries (e.g., predicting the exact concentration of nitrate in groundwater, or the biomass density in a forest plot).
In a classification setting, we are concerned with decision boundaries — where a new sample should be placed among known categories. In regression, we care about numerical precision — how closely our predicted value approximates the true continuous measurement.
By highlighting this distinction early in the lecture, students can better grasp why regression analysis is essential in environmental domains, where many targets are inherently continuous and require not only correct trend estimation but also an understanding of uncertainty and variance in predictions.

Why Regression Is a Supervised Learning Problem

Regression belongs to the family of supervised learning methods because model training requires labeled data — pairs of inputs and known target values ($\mathbf{X}_i, y_i$)
The model “learns” by adjusting its parameters to minimize the discrepancy between predicted values $\hat{y}_i = f(X_i)$ and actual observed outcomes, y_i.
The “supervision” arises from these known outcomes: the learning process is guided by the feedback from the error signal, i.e., how wrong the model’s predictions are compared to reality. The goal is to reduce the prediction error through successive approximations, guided by an objective function (often the mean squared error).

Usefulness of RMSE in Environmental and Ecological Regression

A crucial methodological aspect of regression modeling is the choice of an evaluation metric — a quantitative measure of how well the model's predictions align with observed data.
The Root Mean Squared Error (RMSE) is one of the most widely used metrics in environmental and ecological modeling, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

This evaluation metric is almost self-explanatory- in that sense, we will go from inside to perimeter. Squared errors indicate that an error is defined as the simple subtraction of true and predicted values, which is squared, avoiding the cancellation of positive and negative deviations. To obtain the total effect of all predicted values, these squared results are summarized. Ultimately, the interest lies in the average contribution of each data instance, which is calculated by dividing the total contribution by the number of instances. At last, the root operation restores the original units of measurement.

- RMSE provides an interpretable measure in the same units as the target variable, which is crucial for environmental scientists who must communicate results to stakeholders and policymakers.
- It reflects both bias (systematic error) and variance (dispersion of residuals), capturing the overall predictive reliability of the model.
- It's easy to plug the RMSE evaluation metric into optimization algorithms used in machine learning, thus facilitating automated parameter tuning.

Pointing out the main advantages of the RMSE:

- RMSE offers a clear and intuitive entry point for students: smaller values indicate better predictive accuracy.
- The quadratic penalty provides a natural illustration of why extreme prediction errors are disproportionately punished — an idea that aligns well with ecological risk reasoning.
- Comparing RMSE to alternative measures (e.g., Mean Absolute Error or Mean Bias Error) encourages critical reflection on metric selection and interpretation in environmental studies.

Technology Overview

Regression modeling in ecological and environmental sciences cannot be viewed solely through a purely mathematical procedure, but rather as a mathematically based approach

extended with a strong technological component that transforms heterogeneous environmental data into interpretable and verifiable models.

To make this process more accessible to students, different regression models (linear regression, Regression Trees, SVM for regression, and ANN) will be utilized on representative datasets from ecology and environmental sciences.
Through integration of open data sources, modern data processing libraries, and visualization techniques, the core principles of the technological and methodological aspects of regression modeling in ML for environmental science will be addressed.

Technological Integration and Datasets

First, a synthetic dataset is used to demonstrate the basic principles of simple and multiple linear regression in a controlled and easily interpretable setting. This allows students to focus on understanding model behavior without the complexity of real-world noise and missing data.

In the second part, a real marine dataset from the Copernicus Marine Environment Monitoring Service (CMEMS) is used, including sea surface salinity, chlorophyll concentration, and current velocity, to apply the same regression workflow to real environmental data and evaluate model performance and ecological relevance.

All analyses are performed in Google Colab using standard Python libraries:

- **pandas** and **NumPy** for data loading, cleaning, and numerical operations,
- **scikit-learn** for implementing regression models (Simple Linear Regression and Multiple Linear Regression) and computing evaluation metrics (RMSE and R^2),
- **Matplotlib** for visualization of data distributions, regression fits, and model diagnostics.

Simple Linear Regression is used as a baseline model to describe the relationship between a single environmental predictor (e.g., temperature) and a continuous response variable (dissolved oxygen). Its main advantage is interpretability and its ability to illustrate fundamental regression concepts.

Multiple Linear Regression extends this approach by incorporating additional predictors (e.g., depth and electrical conductivity) to better represent physical and ecological processes influencing dissolved oxygen variability. This model demonstrates how adding relevant environmental variables can improve explanatory power while maintaining model transparency.

Simple Linear Regression and Multiple Linear Regression: Concept, Method, and Demonstration

A simple linear regression relates one predictor to one response variable:

$$y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

- Where ε represents unexplained variability or measurement noise.

In multiple regression, we have more independent variables, $X_{i,}$ and one dependent variable.
The aim is to determine β_0 and β_1 (or β_k for multiple regression) such that the model minimizes the mean squared error between observed and predicted values.
This lesson demonstrates simple and multiple linear regression using environmental sensor measurements from an estuarine shellfish farming area. In the applied case study, dissolved oxygen concentration (mg/L) is modeled as a function of water temperature and additional predictors (e.g., depth and conductivity). This setup illustrates how linear models can provide interpretable baseline relationships and how multi-predictor models can improve explanatory power in ecological systems.

Steps:

Data preparation:
Load the environmental sensor dataset containing measurements of water temperature, dissolved oxygen concentration, and optional predictors such as depth and electrical conductivity. Convert all scaled values into physical units, inspect the dataset for missing values and outliers, and split the data into training and test sets.

Analysis:
Fit two regression models:
(i) a simple linear regression model using temperature as the sole predictor of dissolved oxygen (Temp → DO), and
(ii) a multiple linear regression model using temperature, depth, and conductivity as predictors of dissolved oxygen (Temp + Depth + Conductivity → DO).

Evaluation:
Evaluate both models using the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). Visualize predicted versus observed values and analyze residual plots to assess systematic errors and model assumptions.

Model refinement:
Examine multicollinearity among predictors and discuss its influence on model stability. As an extension, introduce regularization techniques such as Ridge and Lasso regression to reduce overfitting and improve model robustness.

Interpretation:
Interpret regression coefficients and residual patterns in ecological and environmental terms, considering physical and biological mechanisms such as oxygen solubility, photosynthetic activity, and water mixing processes.

Non-Linear Regression: Concept, Method, and Demonstration

Non-linear regression extends the idea of linear regression to relationships that a straight line cannot adequately capture. In many ecological and environmental processes, the response variable does not change linearly with its predictors — instead, it may exhibit thresholds, saturation, exponential growth, or cyclical patterns.

The formal definition of a non-linear regression can be expressed as:

$$y_i = f(X_i, \beta) + \varepsilon_i$$

- Where f(·) is a non-linear function of the predictors Xi and parameters β, and ε_i represents random error.

Common functional forms include polynomial, exponential, logarithmic, or power-law transformations, as well as kernel-based, tree-based, or ANN models when non-parametric flexibility is required.

The objective remains the same as in linear regression: to minimize the difference between observed (y_i) and predicted ($\hat{y}_i$) values, typically using the RMSE.
However, in non-linear regression, the fitted curve can take many possible shapes depending on ecological theory, data behavior, and feature transformations.

This demonstration focuses on modeling dissolved oxygen concentration (DO) as a function of water temperature using a synthetic dataset from an estuarine shellfish farming area. Polynomial regression can be used to illustrate how non-linear models capture curved response patterns when linear assumptions are insufficient.

Steps:

Data preparation:

Retrieve continuous measurements of water temperature and dissolved oxygen concentration from the estuarine shellfish farming monitoring system.

Supplement the real dataset with synthetic datasets designed to enhance non-linear patterns for didactic purposes. Handle missing values, normalize predictors, and split the dataset into training, validation, and test sets.

Analysis:

Apply Polynomial Regression to capture non-linear relationships between water temperature and dissolved oxygen concentration. Compare the polynomial model with Support Vector Regression (SVR), Random Forest Regression, and Artificial Neural Networks (ANN) to explore more complex non-linear structures.

Evaluation:

Calculate RMSE for all models and visualize predicted versus observed values using two-dimensional plots (temperature vs dissolved oxygen) and optional three-dimensional plots when additional predictors are included. Examine non-

linear response curves to detect ecological thresholds, such as temperature levels above which dissolved oxygen rapidly declines.

Model optimization:

Use 10-fold cross-validation to select the polynomial degree and tune kernel or ANN hyperparameters. Compare model performance on the test dataset by assessing both generalization error and model interpretability.

Interpretation:

Relate model findings to aquatic ecosystem processes and discuss implications for shellfish farming management, early warning of hypoxic conditions, and climate-driven stress in estuarine environments.

Evaluation metrics for regression problems

Evaluation metrics quantify the predictive performance of the model, while diagnostics assess how well the model structure aligns with the underlying ecological process. The goal is to understand not only *how accurate* the model is, but *why* it performs as it does — where it succeeds, where it fails, and what those errors imply.

The most widely used measures in ecological regression are:

Root Mean Squared Error (RMSE) – already mentioned:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

Sensitive to significant errors — ideal when extreme deviations have substantial ecological consequences (e.g., underestimating flood discharge or overpredicting pollutant levels).

Mean Absolute Error (MAE):

$$MAES = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

Less sensitive to outliers and provides insight into average deviation.

Coefficient of Determination (R^2):

This measure is not reliable for time-series data or spatial data. In other cases, it measures the proportion of variance in y that is explained by the model, which helps assess the overall explanatory strength.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- SSres → $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ → **Residual Sum of Squares** (unexplained variance)
- SS_{tot} → $\sum_{i=1}^{n} (y_i - \overline{y})^2$ → **Total Sum of Squares** (total variance in observed data)
- y_i = observed value
- $\hat{y}_i$ = model prediction
- $\acute{y}$ = mean of all observed values
- n = number of samples

Nash–Sutcliffe Efficiency (NSE):

Common in hydrology, expressing predictive skill relative to the mean of observations.

$$NSE = 1 - \frac{\sum (Q_{obs} - Q_{mod})^2}{\sum (Q_{obs} - \overline{Q})^2}$$

in which Q_{obs} refers to observations (in this case, of discharge Q), Q_{mod} refers to simulations of the same variable as the observations, and $\overline{Q}$ represents the average observation of the selected variable.

Applied Use Case: Regression Modeling of Dissolved Oxygen in an Estuarine Shellfish Farming Area

Background and Motivation

At a shellfish farming site located in an estuarine/riverine environment, water temperature and dissolved oxygen (DO, mg/L) are continuously measured using in situ sensors. Dissolved oxygen directly affects shellfish respiration and stress and is therefore a key indicator of ecosystem health and aquaculture sustainability.

The objective of this use case is to construct a **regression model** that estimates DO from water temperature. This model serves as a rapid indicator of potential unfavorable conditions: **when temperature increases, a decrease in oxygen concentration is expected.** Early detection of such conditions allows farmers to monitor and mitigate ecological stress risks more effectively.

Students compare model predictions with measured values and analyze situations in which the model fails. This provides an intuitive introduction to the idea that regression models can be improved by incorporating additional environmental predictors such as oxygen saturation (%), depth, conductivity, and turbidity.

Simple Linear Regression

Step 1 – Understanding of data

Dataset Description and Variables

The dataset used in this applied case study originates from continuous environmental monitoring at an estuarine shellfish farming site. Measurements were collected using in situ sensors and include water temperature and dissolved oxygen concentration, along with additional environmental variables used for model extension.
The dataset contains several thousand observations and does not include missing values in the primary variables used for modeling. All values were provided in scaled units and were converted into physical units (°C and mg/L) during preprocessing.

Dataset name:

In situ environmental monitoring system (synthetic dataset prepared for educational purposes)
The regression problem is defined as follows:

Target variable (y):
Dissolved oxygen concentration (DO, mg/L)
Primary predictor (X):
Water temperature (Temp, °C)

Additional predictors (for multiple regression):

- Depth (m)
- Electrical conductivity (µS/cm)
- (optional) oxygen saturation (%)

- (optional) turbidity

These variables were selected based on their physical and ecological relevance. Temperature directly affects oxygen solubility, while depth and conductivity reflect water column structure and mixing processes that influence oxygen availability.
Not all available dataset attributes were used in the regression models. Variables such as record identifiers and non-physical codes were excluded to avoid noise and overfitting and to maintain interpretability of the regression results.

This problem represents a **supervised regression task with a continuous target variable**.

Step 2 – Data inspection

The environmental sensor dataset was loaded using the Pandas library. All scaled measurements were converted into physical units (°C for temperature and mg/L for dissolved oxygen) to ensure physical interpretability of the regression models.

The dataset was imported from a semicolon-separated CSV file and preprocessed by converting scaled values into physical units (Figure 1). A preview of the first observations confirms correct data import and prepares the dataset for regression analysis (Figure 2).

```
# Load CSV (semicolon-separated)
df = pd.read_csv("/content/KrkaRiverDataset_synthetic_multi_R2_40pct.csv", sep=";")

# Convert scaled values to physical units
df["Temp_deg_C"] = df["Temp_deg_C"] / 100
df["HDO_mg/l"]   = df["HDO_mg/l"] / 100

# Quick inspection
display(df[["Temp_deg_C", "HDO_mg/l", "Depth_m", "SpCond_uS/cm"]].head())
```

Figure 1 Python code for loading the environmental sensor dataset from a semicolon-separated CSV file and converting scaled temperature and dissolved oxygen values into physical units (°C and mg/L).

	Temp_C	DO_mgL
0	14.81	6.32
1	15.26	4.96
2	16.43	7.09
3	15.68	7.14
4	16.21	5.73

Figure 2 Preview of the first observations after unit conversion, confirming correct data import and physical interpretability of temperature and dissolved oxygen values.

No missing values were detected in the primary variables used for regression modeling (Figure 3, Figure 4).

```
# Missing values check
print("Missing values:")
print(df[["Temp_deg_C", "HDO_mg/l"]].isnull().sum())
```

Figure 3 Python code used to check for missing values in temperature and dissolved oxygen variables.

```
Missing values:
Temp_deg_C    0
HDO_mg/l      0
```

Figure 4 Output showing that no missing values are present in the main variables used for regression modeling.

Descriptive statistics were used to inspect the range and variability of temperature and dissolved oxygen values and to identify potential outliers (Figure 5, Figure 6) .

```
# Descriptive stats
print("\nDescriptive statistics:")
display(df[["Temp_deg_C", "HDO_mg/l"]].describe())
```

Figure 5 Python code for computing descriptive statistics (count, mean, standard deviation, and range) for temperature and dissolved oxygen variables.

	Temp_deg_C	HDO_mg/l
count	5359.000000	5359.000000
mean	15.679737	6.104551
std	2.559164	1.119456
min	10.020000	4.000000
25%	13.285000	5.305000
50%	15.890000	6.070000
75%	17.990000	6.880000
max	20.790000	10.210000

Figure 6 Summary statistics of temperature and dissolved oxygen values used to assess data range and variability.

To explore the relationship between water temperature and dissolved oxygen concentration, a scatter plot was created (Figure 7, Figure 8).

```
# Scatter plot
plt.figure(figsize=(7,5))
plt.scatter(df["Temp_deg_C"], df["HDO_mg/l"],
            color="red", alpha=0.25, label="Observed data")
plt.xlabel("Temperature (°C)")
plt.ylabel("Dissolved Oxygen (mg/L)")
plt.title("Scatter plot: Temperature vs Dissolved Oxygen")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Figure 7 Python code for visualizing the relationship between temperature (°C) and dissolved oxygen (mg/L) using a scatter plot.

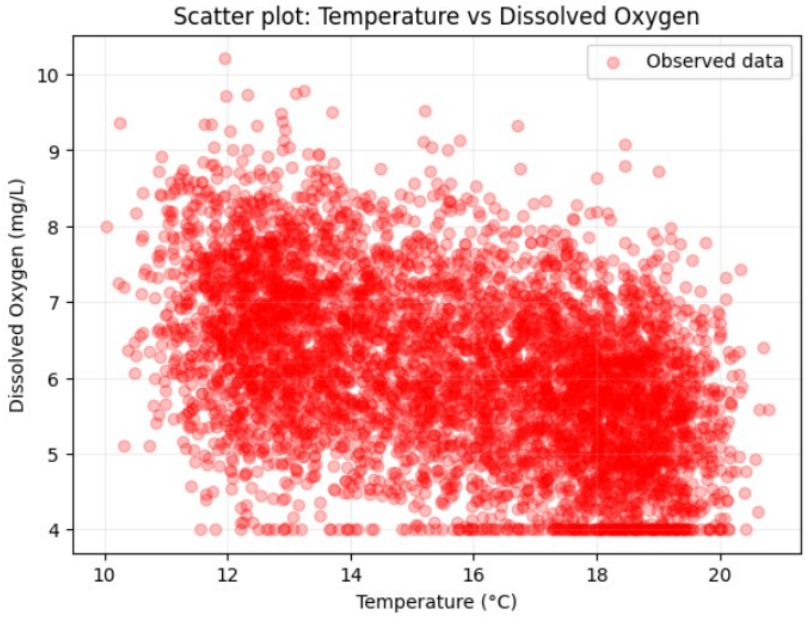

Figure 8 Scatter plot of dissolved oxygen (mg/L) versus water temperature (°C)

The scatter plot reveals a clear negative trend: dissolved oxygen decreases as temperature increases. This pattern is consistent with physical principles of gas solubility in water. The absence of strong curvature or threshold effects suggests that a linear regression model can serve as a suitable first approximation. The observed approximately linear relationship between temperature and dissolved oxygen supports the use of linear regression as a baseline model.

Step 3 – Dataset preparation for the Regression Model

For regression modeling, the relevant predictors and target variable were selected from the dataset.

The target variable was defined as dissolved oxygen concentration (DO), while predictor variables included water temperature for the simple regression model and water temperature, depth, and electrical conductivity for the multiple regression model.

The dataset was split into training and test subsets using an 80/20 ratio. The training set was used to fit the regression models, while the test set was reserved for independent performance evaluation.
This separation ensures an unbiased assessment of model generalization ability.

Step 4 – Baseline linear model (for comparison only)

Model Training

The dataset was split into training and test subsets (80/20 split). A simple linear regression model was fitted using temperature as the sole predictor (Figure 9).

```
# Define predictor and target
X_simple = df[["Temp_deg_C"]]
y = df["HDO_mg/l"]

# Train/test split
X_train_s, X_test_s, y_train, y_test = train_test_split(
    X_simple, y, test_size=0.2, random_state=0
)

# Fit model
lin_model = LinearRegression()
lin_model.fit(X_train_s, y_train)

# Predict
y_pred_simple = lin_model.predict(X_test_s)
```

Figure 9 Python code for training a simple linear regression model using water temperature as the predictor of dissolved oxygen concentration, including train–test split and prediction.

Model Equation and Performance

Model performance was evaluated using RMSE and R^2 (Figure 10,Figure 11).

```
# Evaluate
rmse_simple = np.sqrt(mean_squared_error(y_test, y_pred_simple))
r2_simple = r2_score(y_test, y_pred_simple)

print("SIMPLE LINEAR REGRESSION (Temp → DO)")
print(f"Equation: DO ≈ {lin_model.intercept_:.3f} + ({lin_model.coef_[0]:.3f}) * Temp")
print(f"RMSE = {rmse_simple:.4f} mg/L")
print(f"R²   = {r2_simple:.3f}")
```

Figure 10 Python code for evaluating the linear regression model and visualizing the regression line

```
LINEAR REGRESSION (Temp → DO)
Equation: DO ≈ 9.183 + (-0.197) * Temp
RMSE = 0.9644 mg/L
R²   = 0.204
```

Figure 11 Output showing the regression equation and performance metrics (RMSE and R²) for the simple linear regression model

To visually assess the fitted linear relationship, the regression line was plotted together with the observed data points (Figure 12, Figure 13).

```
# Plot Regression line
temp_grid = np.linspace(df["Temp_deg_C"].min(), df["Temp_deg_C"].max(), 300)
temp_grid_df = pd.DataFrame({"Temp_deg_C": temp_grid})
do_line = lin_model.predict(temp_grid_df)

plt.figure(figsize=(7,5))
plt.scatter(df["Temp_deg_C"], df["HDO_mg/l"],
            color="red", alpha=0.25, label="Observed data")
plt.plot(temp_grid, do_line,
         color="black", linewidth=3, label="Linear regression fit")
plt.xlabel("Temperature (°C)")
plt.ylabel("Dissolved Oxygen (mg/L)")
plt.title("Simple Linear Regression Fit (Temp → DO)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Figure 12 Python code for plotting the linear regression line and observed data

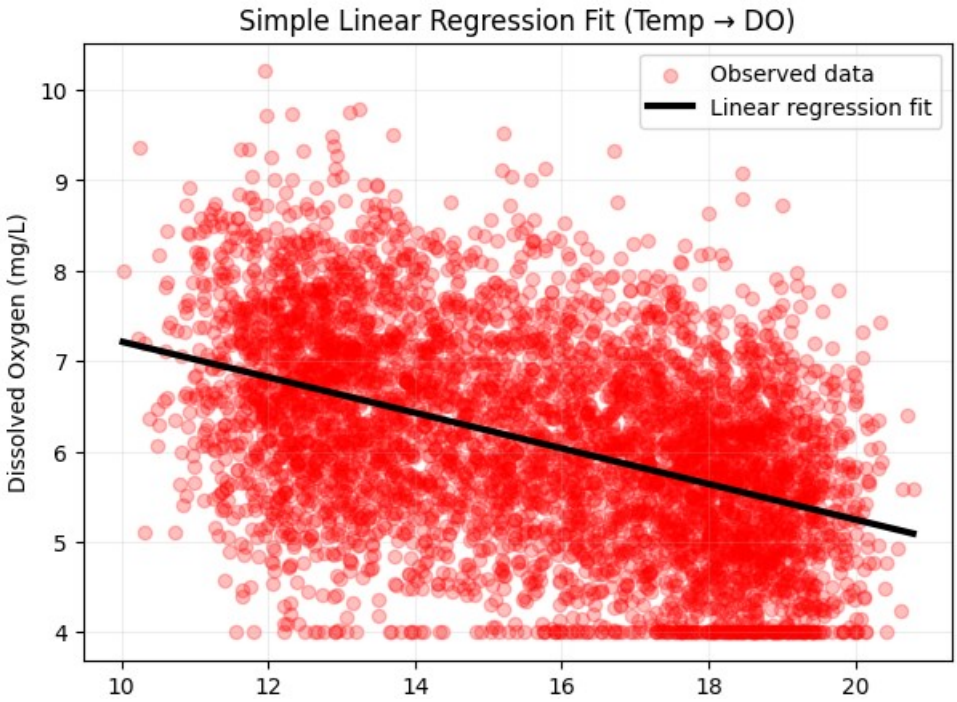

Figure 13 Linear regression fit between water temperature and dissolved oxygen concentration

The regression coefficient confirms a negative relationship between temperature and dissolved oxygen concentration, consistent with the physical principle that oxygen solubility decreases as water temperature increases.
The RMSE value of 0.964 mg/L indicates the average prediction error of the model, while the R^2 value of 0.204 suggests that temperature alone explains approximately 20% of the variance in dissolved oxygen concentration.

Model Diagnostics

Predicted vs Observed

A predicted versus observed plot was generated to examine how closely model predictions match measured dissolved oxygen values (Figure 14,Figure 15).

```
# Predicted vs Observed
plt.figure(figsize=(6,5))
plt.scatter(y_test, y_pred_simple,
            color="red", alpha=0.4, label="Predicted")
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()],
         color="black", linewidth=2, label="1:1 line")

plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Predicted DO (mg/L)")
plt.title("Predicted vs Observed (Simple Linear Model)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Figure 14 Python code for generating the predicted versus observed plot for the simple linear regression model

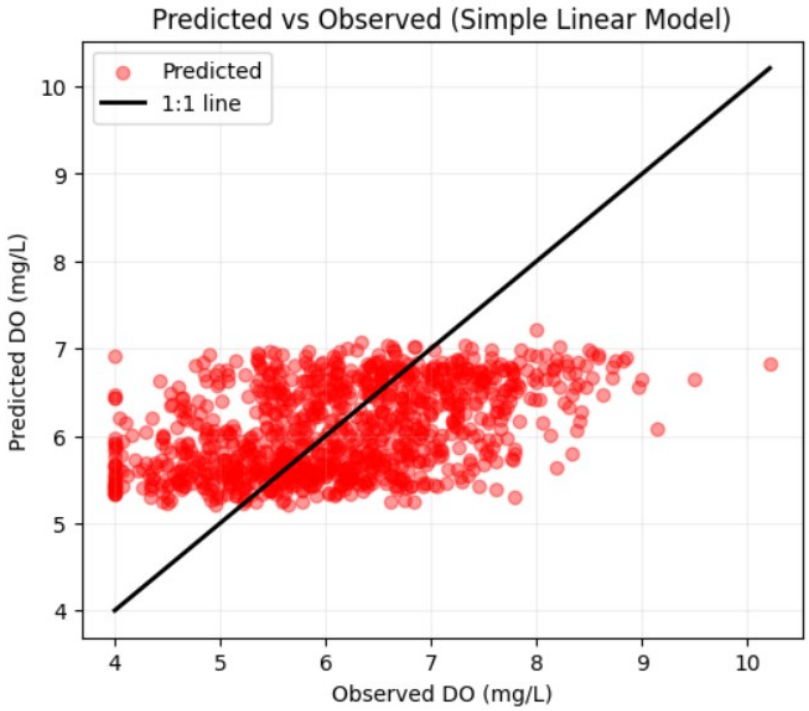

Figure 15 Predicted versus observed dissolved oxygen concentrations for the simple linear regression model

The points cluster around the diagonal line, indicating that the model captures the main decreasing trend of dissolved oxygen with increasing temperature. However, noticeable scatter remains, suggesting that temperature alone does not fully explain the variability in dissolved oxygen concentration.

Residual Analysis

Residuals were calculated as the difference between observed and predicted dissolved oxygen values (Figure 16, Figure 17).

```
# Residual plot
resid_simple = y_test - y_pred_simple

plt.figure(figsize=(6,5))
plt.scatter(y_test, resid_simple,
            color="purple", alpha=0.4, label="Residuals")
plt.axhline(0, color="black", linewidth=2)
plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Residual (Observed - Predicted)")
plt.title("Residual Plot (Simple Linear Model)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Figure 16 Python code for computing and plotting residuals for the simple linear regression model

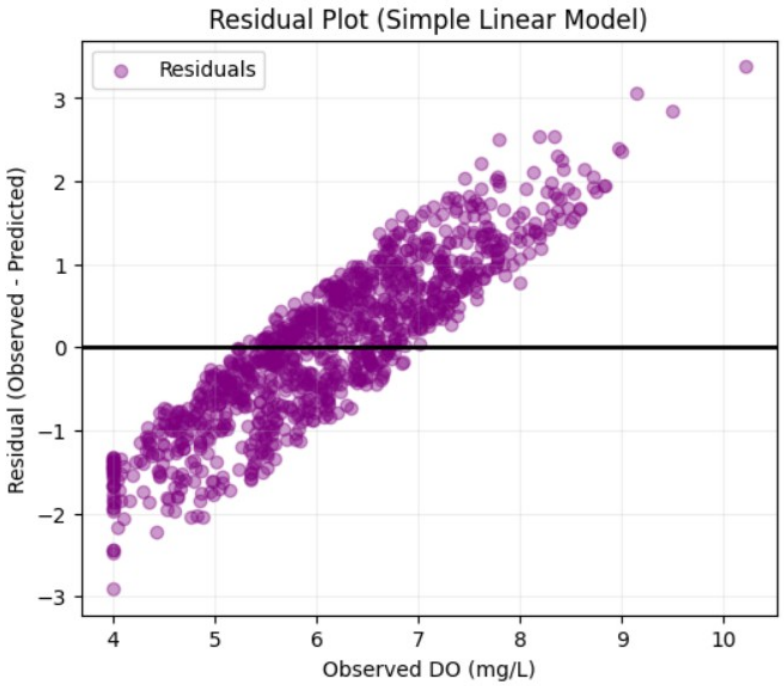

Figure 17 Residual plot for the simple linear regression model (observed minus predicted dissolved oxygen)

Residuals are distributed around zero without strong systematic structure, indicating that the linear model captures the dominant trend but leaves unexplained variability caused by additional ecological and physical processes such as water mixing and biological activity.

These results highlight the limitations of using temperature as the sole predictor and motivate the development of a multiple linear regression model incorporating additional environmental variables.

Multiple Linear Regression Model (Temp + Depth + Conductivity → DO)

Model Training

To improve predictive performance and account for additional physical drivers of dissolved oxygen variability, a **multiple linear regression model was developed using water temperature, depth, and electrical conductivity as predictor variables**.
These predictors were selected based on their ecological relevance. Water temperature directly controls oxygen solubility, while depth and electrical conductivity provide information on water column structure and mixing processes that influence oxygen distribution.
The dataset was split into training and test subsets using the same 80/20 ratio applied in the simple linear regression model to ensure a fair comparison of model performance. The model was trained on the training dataset and then used to predict dissolved oxygen concentrations for the test dataset (Figure 18).

```
# Define predictors and target
X_multi = df[["Temp_deg_C", "Depth_m", "SpCond_uS/cm"]]
y = df["HDO_mg/l"]

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(
    X_multi, y, test_size=0.2, random_state=0
)

# Fit multiple linear regression model
multi_model = LinearRegression()
multi_model.fit(X_train, y_train)

# Predictions
y_pred_multi = multi_model.predict(X_test)

# Evaluation metrics
rmse_multi = np.sqrt(mean_squared_error(y_test, y_pred_multi))
r2_multi = r2_score(y_test, y_pred_multi)

print("MULTIPLE LINEAR REGRESSION (Temp + Depth + Conductivity → DO)")
print("Intercept:", multi_model.intercept_)
print("Coefficients:", multi_model.coef_)
print(f"RMSE = {rmse_multi:.4f} mg/L")
print(f"R²   = {r2_multi:.3f}")
```

Figure 18 Python code for training and evaluating the multiple linear regression model using temperature, depth, and electrical conductivity as predictors

Model Equation and Performance

Model performance was evaluated using the Root Mean Squared Error (RMSE) and the coefficient of determination (R^2).

Compared to the simple linear regression model, the multiple linear regression model achieved a lower RMSE and a higher R^2 value, indicating improved predictive accuracy and an increased ability to explain variability in dissolved oxygen concentration (Figure 19).

The regression coefficients show that temperature remains the dominant predictor with a negative coefficient, confirming the inverse relationship between temperature and dissolved oxygen concentration. Depth and electrical conductivity contribute additional explanatory power, reflecting the influence of stratification and mixing processes on oxygen availability.

```
MULTIPLE LINEAR REGRESSION (Temp + Depth + Conductivity → DO)
Intercept: 21.181625209743284
Coefficients: [-0.1950301   0.57396502 -0.04874799]
RMSE = 0.8230 mg/L
R²   = 0.420
```

Figure 19 Output showing regression coefficients and performance metrics (RMSE and R^2) for the multiple linear regression model

These results indicate that the multiple linear regression model explains approximately 42% of the variability in dissolved oxygen concentration and provides a clear improvement over the simple temperature-only regression model.

Predicted vs Observed Values

A predicted versus observed plot was generated to visually assess the agreement between measured and predicted dissolved oxygen concentrations for the multiple regression model (Figure 20). The data points cluster more closely around the 1:1 diagonal line compared to the simple linear regression model, demonstrating improved predictive performance when additional environmental predictors are included.

Remaining deviations from the diagonal line indicate the influence of unmodeled biological and physical processes such as photosynthesis, respiration, and short-term hydrodynamic variability (Figure 21).

Overall, the predicted versus observed plot confirms that including depth and electrical conductivity as additional predictors enhances the model's ability to reproduce observed dissolved oxygen concentrations.

```
# Predicted vs Observed (Multiple Linear Regression)
plt.figure(figsize=(6,5))
plt.scatter(y_test, y_pred_multi,
            color="red", alpha=0.4, label="Predicted")
plt.plot([y_test.min(), y_test.max()],
         [y_test.min(), y_test.max()],
         color="black", linewidth=2, label="1:1 line")

plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Predicted DO (mg/L)")
plt.title("Predicted vs Observed (Multiple Linear Regression)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Figure 20 Python code for generating the predicted versus observed plot for the multiple linear regression model.

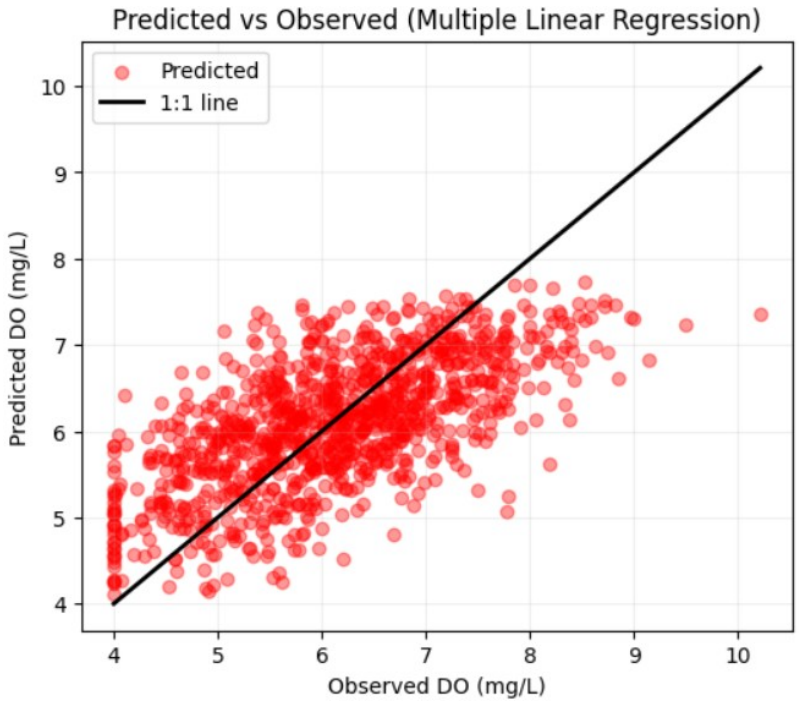

Figure 21 Predicted versus observed dissolved oxygen concentrations for the multiple linear regression model

Residual Analysis

Residuals were analyzed to identify systematic errors and assess the validity of linear model assumptions (Figure 22).

```python
# Residual plot (Multiple Linear Regression)
resid_multi = y_test - y_pred_multi

plt.figure(figsize=(6,5))
plt.scatter(y_test, resid_multi, color="purple", alpha=0.4, label="Residuals")
plt.axhline(0, color="black", linewidth=2)

plt.xlabel("Observed DO (mg/L)")
plt.ylabel("Residual (Observed - Predicted)")
plt.title("Residual Plot (Multiple Linear Regression)")
plt.legend()
plt.grid(alpha=0.2)
plt.show()
```

Figure 22 Python code for computing residuals and generating the residual plot for the multiple linear regression model

The residual plot shows that residuals are distributed around zero without a strong systematic pattern, indicating that the multiple linear regression model captures the main linear relationships between dissolved oxygen and the selected predictors (temperature, depth, and electrical conductivity)(Figure 23).

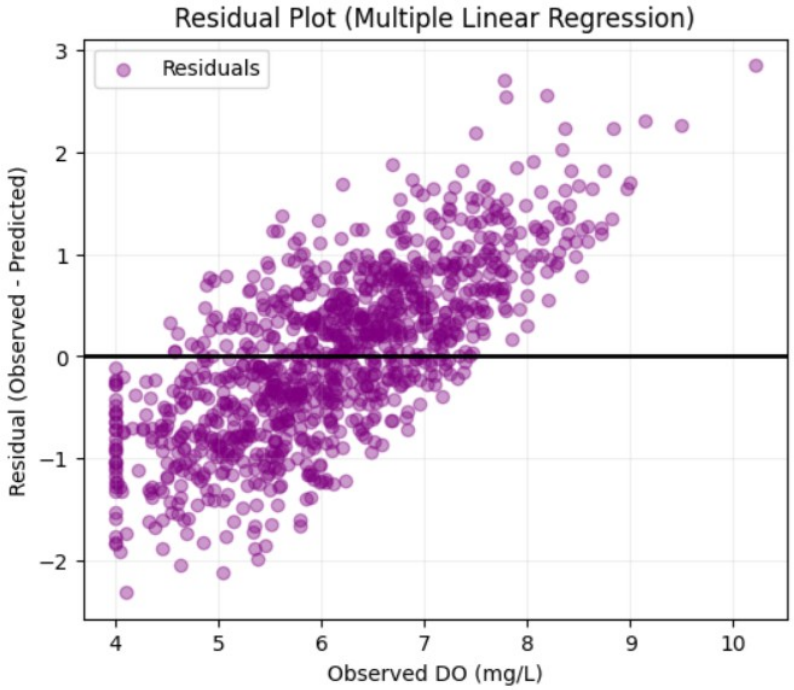

Figure 23 Residual plot for the multiple linear regression model showing the difference between observed and predicted dissolved oxygen concentrations

Compared to the simple linear regression model, the residuals exhibit a more homogeneous spread, suggesting reduced bias and improved model stability when additional environmental predictors are included.

However, some dispersion remains at higher temperature and lower oxygen ranges, which likely reflects the influence of nonlinear biological and physical processes such as photosynthesis, respiration, and short-term hydrodynamic mixing that cannot be fully represented by a linear model with a limited number of predictors.

Overall, the residual analysis confirms that the multiple linear regression model provides a reasonable approximation of dissolved oxygen dynamics while highlighting the need for more complex (non-linear) models to capture remaining variability.

Comparison with the Simple Linear Regression Model

The multiple linear regression model was compared with the simple linear regression model that used only water temperature as a predictor of dissolved oxygen concentration. The comparison shows that the multiple linear regression model achieved a lower Root Mean Squared Error (RMSE) and a higher coefficient of determination (R^2) than the simple linear regression model. This indicates improved predictive accuracy and an increased ability to explain variability in dissolved oxygen concentration when additional environmental predictors are included.

While the simple linear regression model captures the dominant inverse relationship between temperature and dissolved oxygen, it does not account for other important physical drivers such as water column structure and mixing conditions. **By incorporating depth and electrical conductivity, the multiple linear regression model better represents the complexity of estuarine oxygen dynamics.**

This comparison demonstrates the limitations of single-predictor regression models for ecological systems and highlights the importance of using multiple ecologically meaningful variables when modeling environmental processes.

Ecological Interpretation

The negative regression coefficient for temperature reflects the well-known physical principle that oxygen solubility decreases as water temperature increases. Warmer water holds less dissolved oxygen, which directly affects shellfish respiration and increases physiological stress in aquaculture environments.

The contributions of depth and electrical conductivity indicate the role of stratification, freshwater inflow, and mixing processes in shaping oxygen availability within the estuarine system. Variations in depth influence light penetration and biological activity, while conductivity serves as an indicator of salinity gradients and water mass exchange between riverine and marine sources.

Periods with larger prediction errors may correspond to episodes of enhanced photosynthetic oxygen production during daylight hours or sudden mixing events driven by tides, wind, or river discharge.

Overall, the multiple linear regression model provides both predictive capability and ecological insight by linking dissolved oxygen dynamics to key physical and biological mechanisms operating in estuarine shellfish farming areas. This demonstrates how regression modeling can support environmental monitoring and early warning systems for hypoxic stress conditions.

Practical Exercise: Regression in Environmental Sciences

Topic: Predicting chlorophyll-a concentration from physical oceanographic variables

Student work: Individual

Tools:
Python (Jupyter Notebook / Google Colab), pandas, numpy, scikit-learn, matplotlib, seaborn

Dataset:
Copernicus Marine Service (CMEMS) – sea surface salinity, chlorophyll, and current velocity.

Background and Motivation

Phytoplankton biomass, commonly represented by chlorophyll-a concentration (Chl-a), is a key indicator of marine ecosystem productivity and water quality. Variations in chlorophyll-a are influenced by physical drivers such as salinity and ocean currents, which affect nutrient availability, stratification, and mixing processes.
In this practical exercise, students apply regression techniques to model chlorophyll-a concentration using environmental predictors derived from Copernicus Marine Service (CMEMS) data. The task demonstrates how simple and multiple linear regression can be used to quantify relationships between biological and physical variables in marine systems and to support environmental monitoring and early warning applications.

Dataset Description

The dataset is derived from CMEMS marine products and contains time-series or spatial samples of surface ocean conditions. Each row corresponds to one observation point (time and/or location) and includes the following variables:

Target variable (y):

- Chlorophyll-a concentration (Chl-a, mg/m^3)

Predictor variables (X):

- Sea surface salinity (PSU)
- Surface current velocity (m/s)

Task Description

The student is expected to build and compare two regression models for predicting chlorophyll-a concentration:

1. **Simple Linear Regression:**
   Chl-a = f(Salinity)
2. **Multiple Linear Regression:**
   Chl-a = f(Salinity, Current Velocity)

Step-by-Step Guide

Step 1 – Setup

- Open Google Colab or Jupyter Notebook.
- Import required Python libraries (pandas, numpy, scikit-learn, matplotlib).

Step 2 – Load and Inspect the Dataset

- Load the CMEMS dataset from the provided CSV file.
- Display the first rows of the dataset.
- Report:
  - Number of observations (rows)
  - Number of variables (columns)
  - Target variable
  - Predictor variables
- Check for missing values.

Step 3 – Data Preparation

- Convert scaled values into physical units if necessary.
- Remove non-physical or identifier columns.
- Inspect descriptive statistics and potential outliers.
- Split the dataset into training and test sets (80/20).

Step 4 – Model 1: Simple Linear Regression (Baseline Model)

Train a simple linear regression model using salinity as the onlỳ sole predictor.
Model formulation:
Chl-a = f(Salinity)

- Fit the model on the training data.
- Predict chlorophyll-a on the test data.
- Compute:
  - RMSE
  - R^2

Visualizations:

- Scatter plot: Salinity vs Chlorophyll-a
- Regression line
- Predicted vs Observed plot
- Residual plot

Expected from student:

Briefly explain the physical meaning of the regression coefficient (how salinity relates to chlorophyll-a concentration).

Step 5 – Model 2: Multiple Linear Regression

Train a multiple linear regression model using both predictors:
Model formulation:
Chl-a = f(Salinity, Current Velocity)

- Fit the model on the training data.
- Predict chlorophyll-a on the test data.
- Compute:
  - RMSE
  - R^2

Visualizations:

- Predicted vs Observed plot
- Residual plot
-

Expected from student:

Interpret the coefficients of salinity and current velocity in ecological terms (mixing, transport, stratification).

Step 6 – Model Comparison

Compare the two models:

- Which model has lower RMSE?
- Which model has higher R^2?
- Does adding current velocity improve prediction performance?

Expected from student:

Provide a short explanation (1–2 sentences) of why the multiple regression model performs better or worse than the simple model.

Step 7 – Ecological Interpretation

Students should discuss:

- How salinity influences phytoplankton through freshwater input and nutrient supply.
- How current velocity affects chlorophyll distribution via mixing and transport processes.
- Why part of the variability remains unexplained (light availability, nutrients, biological processes).

What to Submit – Exercise Report

Format:
Jupyter Notebook or Google Colab with markdown and code cells. The report must contain:

a. Introduction
- What was modeled?
- Why is chlorophyll-a important for marine ecosystems?

b. Data Description
- Number of samples
- Predictor variables
- Target variable
- Handling of missing values

c. Methods
- Simple linear regression
- Multiple linear regression
- Train/test split ratio
- Evaluation metrics (RMSE, R^2)

d. Results
- RMSE and R^2 for both models
- Plots (scatter, predicted vs observed, residuals)
- Short comparison of models

e. Discussion
- Interpretation of coefficients
- Environmental meaning of errors
- Limitations of linear regression

f. Conclusion
- Summary of what was learned about regression modeling using marine environmental data

Assessment

PRACTICAL REGRESSION TASK

The Practical Regression Task provides students with a hands-on opportunity to apply regression models (simple and multiple linear regression) to a real environmental dataset obtained from the Copernicus Marine Service (CMEMS). It is intended to assess the students' ability to implement, validate, and interpret regression models for environmental decision support.

QUIZ

The quiz will be designed to assess conceptual understanding and applied reasoning gained through the lecture and practical exercise on regression techniques in environmental sciences. It will check whether students understand simple versus multiple linear regression, and the role of regression models in environmental monitoring and prediction.

SHORT WRITTEN REPORT

The Short Written Report serves as the final step in the Practical Regression Task, allowing students to synthesize, interpret, and communicate their findings from the regression modeling exercise.

Evaluation Criteria
1. QUIZ

Question 1
What type of machine learning problem is predicting dissolved oxygen concentration from environmental variables such as temperature and depth?
a) Classification
b) Regression
c) Clustering
d) Anomaly detection

Question 2
Which variable represents the target (dependent variable) in the applied use case of this lesson?
a) Water temperature
b) Electrical conductivity
c) Dissolved oxygen concentration
d) Depth

Question 3
Which of the following best defines a regression problem in machine learning?

a) Predicting a categorical label
b) Predicting a continuous numerical value
c) Grouping similar samples
d) Reducing the number of variables

Question 4
What is the main difference between simple linear regression and multiple linear regression?
a) Simple regression is unsupervised
b) Multiple regression uses more than one predictor variable
c) Simple regression is always more accurate
d) Multiple regression cannot be evaluated

Question 5
Why is regression considered a supervised learning method?
a) Because it does not require labeled data
b) Because the model learns without a target variable
c) Because the model is trained using known input–output pairs
d) Because it only works on small datasets

Question 6
What does the Root Mean Squared Error (RMSE) measure?
a) The proportion of explained variance
b) The average prediction error in the same units as the target variable
c) The number of correct predictions
d) The correlation between variables

Question 7
What is the main purpose of splitting a dataset into training and test sets?
a) To increase model complexity
b) To remove outliers
c) To evaluate model performance on unseen data
d) To reduce computation time

Question 8
Which situation suggests that a non-linear regression model may be more appropriate than a linear one?
a) The relationship between variables forms a straight line
b) The data show thresholds or curved response patterns
c) There is no noise in the data
d) Only one predictor is available

Question 9
What information does a predicted vs. observed plot provide?

a) It shows the distribution of predictors
b) It visualizes how closely predictions match measured values
c) It detects missing data
d) It selects the best predictors automatically

Question 10
Why are regression models useful in environmental sciences?
a) They replace physical measurements
b) They help predict and interpret continuous environmental variables
c) They eliminate uncertainty in observations
d) They only work for laboratory data

2. PRACTICAL REGRESSION TASK

Data Preparation and Preprocessing

- Correct loading and inspection of the dataset (shape, predictors, target variable).
- Appropriate handling of missing data and unit conversion where required.
- Transparent and reproducible preprocessing workflow (code is clean and steps are clearly explained).

Model Implementation

- Accurate setup and training of at least two regression models:
  - Simple Linear Regression
  - Multiple Linear Regression
- Appropriate use of train/test split.
- The code is organized and parameters are explicitly stated.

Model Evaluation and Diagnostics

- Correct calculation and reporting of regression metrics:
  - RMSE
  - R^2
- Logical comparison of models (e.g., “multiple regression improved R^2 compared to simple regression”).
- Use of diagnostic plots (predicted vs. observed, residual plots).
- Awareness of overfitting and underfitting.

Model Selection and Justification

- Clear justification of which model is preferred.
- Sensible explanation of the model’s generalization ability based on test data results.

Visualization and Presentation of Results

- Figures (scatter plots, regression fits, residual plots) are clearly labeled.
- Visuals support interpretation of model behavior.

Connection to Environmental Context

- The student explains how regression models can support environmental monitoring and early warning systems.

3. SHORT WRITTEN REPORT

Clarity and Structure of Report

- The report is logically organized (Introduction → Data → Methods → Results → Discussion → Conclusion).
- Steps of the workflow are visible in the notebook.
- Language is precise and concise.

Interpretation of Results and Discussion

- Regression results (RMSE, R^2, residuals) are accurately reported and meaningfully interpreted.
- The student reflects on why multiple regression performed better (or not) than simple regression.

Connection to Ecological and Environmental Context

- Findings are linked to real environmental meaning (e.g., predicting chlorophyll variability for marine ecosystem monitoring).
- The student demonstrates understanding of how data-driven regression can contribute to environmental decision-making.

Quality of Visual Presentation

- Figures are clearly labeled and referenced in the text.
- Visuals help explain model behavior, not just decorate the report.

Communication and Style

- Writing is clear, objective, and professional.
- Machine learning and environmental terminology are used correctly (predictor, target, RMSE, residuals, coefficient).
- The student acknowledges limitations (simplified model, linear assumptions, limited predictors).

Suggested Weighting

- Quiz: **20%**
- Practical Regression Task (notebook/code): **50%**
- Short Written Report: **30%**

References / Further Reading

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.
- Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.
- Copernicus Marine Service (CMEMS) Data Access Portal
- Open-access research articles (complete list will be provided on the Course e-learning system)