Regression Analysis Report

Introduction

This report presents an analysis of the relationship between ambient air temperature and energy output in a combined cycle power plant. By performing both linear and polynomial regression on a selected subset of data, we aim to model and understand how temperature variations influence energy production. The study involves data transformation, model fitting using closed-form solutions and iterative methods, and comparative analysis of different regression approaches. The findings provide insights into the efficiency of different regression approaches and the underlying dynamics between temperature and energy output.

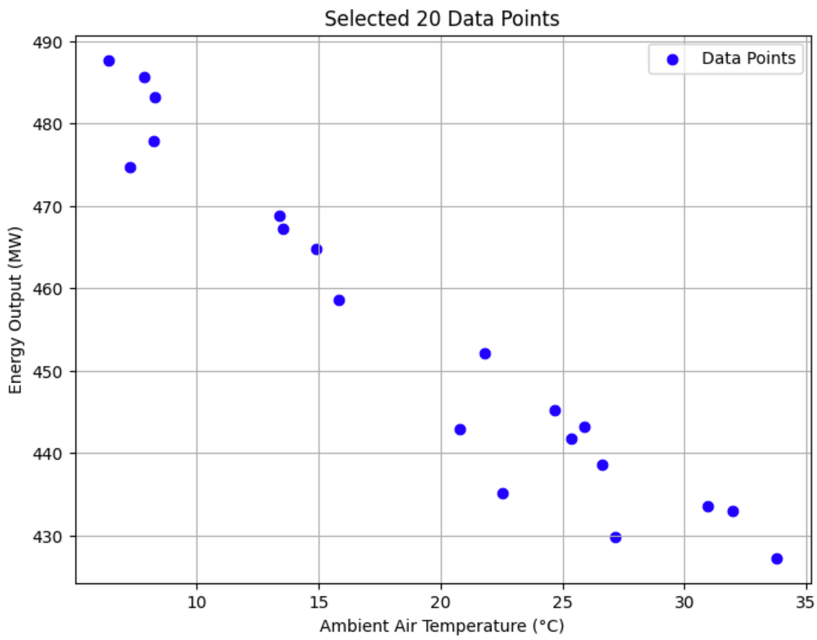
A.I. Linear Regression

Task 1: Data Selection and Visualisation

In the context of this assignment, we assigned a group number n = 1. The purpose of using the group number n is to divide the dataset into non-overlapping segments, allowing each group to analyse distinct portions simultaneously. This ensures computational efficiency and prevents data overlap among different groups. The dataset is partitioned such that each group works with a unique subset of 20 data points, starting at index 20n. Therefore, our subset for this iteration of the experiment comprises data points from index 20 to 39.

Visualisation:

A scatter plot of the selected data points was created to visualise the relationship between ambient air temperature (x) and energy output (y).



Observations:

Negative Correlation: There is a noticeable inverse relationship between ambient air temperature and energy output.
Trend: As the temperature increases from approximately 7°C to 34°C, the energy output decreases from around 488 MW to 427 MW.
Implication: Higher ambient temperatures are associated with reduced energy output, suggesting that the power plant's efficiency decreases with increasing temperature.

Task 2: Data Transformation

Z-Score Normalisation (Standardisation)

In our data transformation process, we applied Z-score normalisation (standardisation) to both predictor x and response y. Z-score normalisation is a technique that shifts data to have a mean of 0 and a standard deviation of 1, centring the dataset around the origin and scaling all values uniformly. This transformation is particularly useful for datasets with varying feature scales, as it provides a balanced framework that simplifies the modelling process.

Justification

We selected Z-score normalisation to enhance the performance of gradient descent algorithms, which are sensitive to feature scale differences. By standardising both x and, we allow the optimisation process to traverse the cost function’s landscape more efficiently and without bias towards larger-scaled features.

For example, without normalisation, a feature measured in large units, such as degrees in temperature, could dominate the model, skewing the optimisation and causing slower convergence. Standardisation eliminates this imbalance, enabling gradient descent to proceed smoothly across all features.

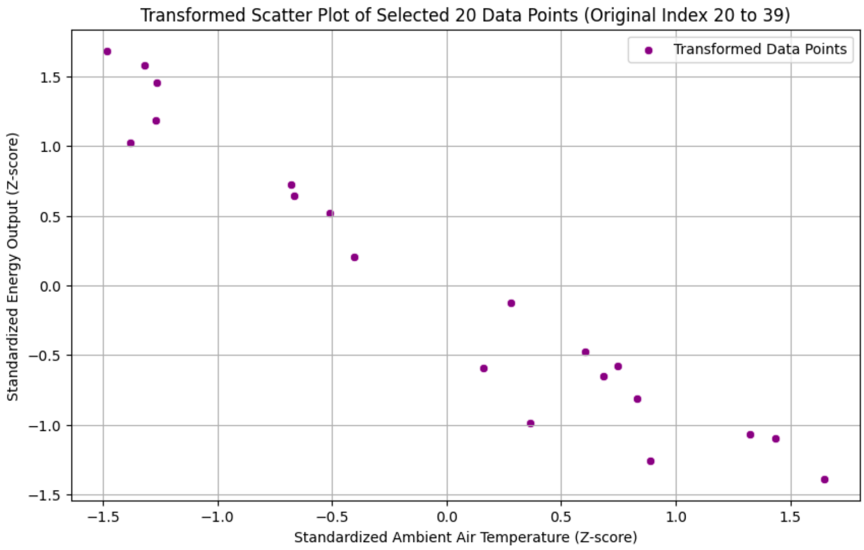
If we had not applied Z-score normalisation, the optimisation algorithm would have struggled with features of different magnitudes, potentially converging more slowly and less reliably due to the need for smaller steps to avoid overshooting the minimum.

We have therefore attained faster convergence as the algorithm avoids being biased toward larger-scaled features. Additionally, standardisation improves numerical stability, which is valuable when handling datasets with large ranges or extreme values. It minimises computational errors that can arise from very large or very small values, making calculations more reliable.

Another benefit is that it enables a balanced contribution from both x and y, preventing one variable from disproportionately influencing the model due to its scale. Finally, Z-score normalisation enhances interpretability, as the coefficients in the model now represent the expected change in y for a one standard deviation shift in x.

Anticipated Outcomes included expectations of faster convergence, more stable optimisation, and greater interpretability of model coefficients. With standardisation, we predicted that the gradient descent process would reach the optimal solution in fewer iterations. We also anticipated a reduction in oscillations, with a smoother path toward the minimum, and an improved interpretative value for model coefficients, making them more comparable across different features.

Actual Results and Observations confirmed our hypothesis. The actual results confirmed our hypothesis. After applying Z-score normalisation, both y achieved a mean close to 0 and a standard deviation close to 1, demonstrating successful transformation. This balanced scaling enabled a more stable and efficient optimisation process during regression. Importantly, the original high negative correlation of -0.95 between ambient air temperature and energy output was preserved, showing that Z-score normalisation retained the inherent inverse relationship in the data while simply rescaling the features.



Task 3: Compute Linear Regression Coefficients Using Closed-Form Solution

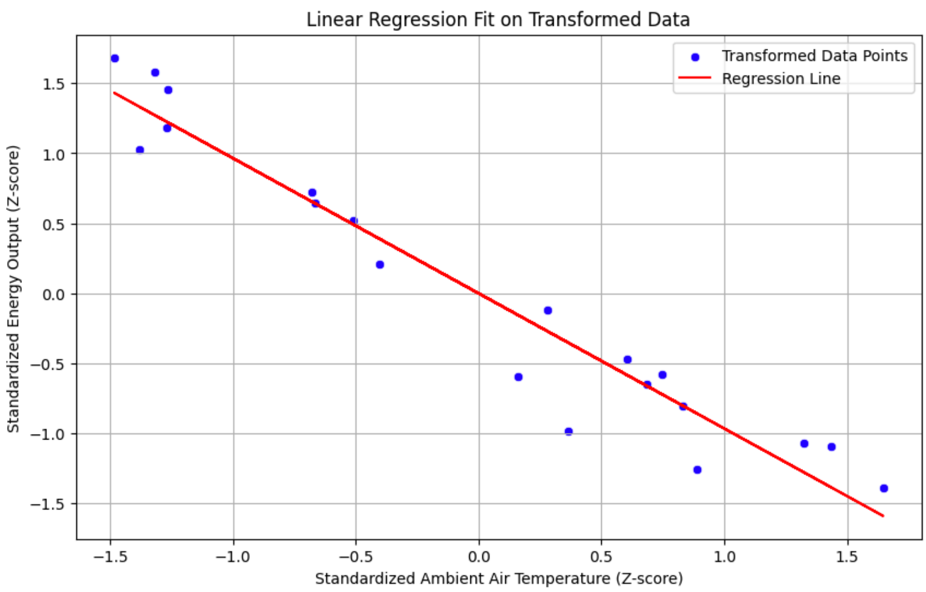
To compute the coefficients a (slope) and b (intercept) of the linear regression model $Y \approx aX + b$ using the closed-form solution (Normal Equation) and visualise the regression line against the standardised data.

Computed Coefficients:

Intercept (b): -0.0000
Slope (a): -0.9660

Interpretation of the Results:

- **Intercept (b)**: In the context of standardised data, an intercept near zero indicates that when the ambient air temperature is at its mean, the energy output is also at its mean. This aligns with our understanding that the data has been centered.
- **Slope (a)**: The slope of -0.9660 indicates that for every one standard deviation increase in ambient air temperature, the energy output decreases by approximately 0.966 standard deviations. This near-unitary negative slope underscores the strong inverse relationship between the two variables.



Visualisation and Fit Quality:

The regression line was plotted alongside the standardised data points. The line closely follows the trend of the data points, validating the strength and direction of the relationship. The proximity of data points to the regression line indicates a good fit.

Task 4: Choose a Cost Function and Find Partial Derivatives

Hypothesis:

We hypothesised that using the Mean Squared Error (MSE) as the cost function would effectively quantify the discrepancy between the predicted energy outputs and the actual values. Given the strong negative correlation observed between ambient air temperature and energy output, MSE was expected to provide a robust measure for optimising our regression model.

Chosen Cost Function:

The Mean Squared Error (MSE) is defined as:

$$J(a, b) = \frac{1}{2n} \sum_{i=1}^{n} \left(y^{(i)} - (a x^{(i)} + b)\right)^2$$

Justification:

- **Penalises Larger Errors:** MSE emphasises larger errors due to the squaring of residuals, promoting a model that minimises significant discrepancies.
- **Differentiable:** Allows for the computation of gradients necessary for optimisation via gradient descent.
- **Widely Used:** MSE is a standard choice in regression problems, facilitating comparison with other studies.

Partial Derivatives:

With respect to a:

$$\frac{\partial J}{\partial a} = -\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(y^{(i)} - (a x^{(i)} + b)\right)$$

With respect to b:

$$\frac{\partial J}{\partial b} = -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - (a x^{(i)} + b)\right)$$

Evaluation of the Experiment:

The computed coefficients from the closed-form solution (intercept $-$ 0.0000 and slope -0.9660) were used to calculate the current cost, yielding an MSE of 0.0668. The partial derivatives at these parameters were close to zero ($\frac{\partial J}{\partial a} = -0.0000$, $\frac{\partial J}{\partial b} = -0.0000$), indicating that the parameters are at or near the minimum of the cost function.

Conclusion: The results strongly support the initial hypothesis. The MSE effectively captures the relationship, and the computed coefficients align with the observed high negative correlation. The low MSE and near-zero partial derivatives validate the reliability and accuracy of the regression model in capturing the underlying data trends.

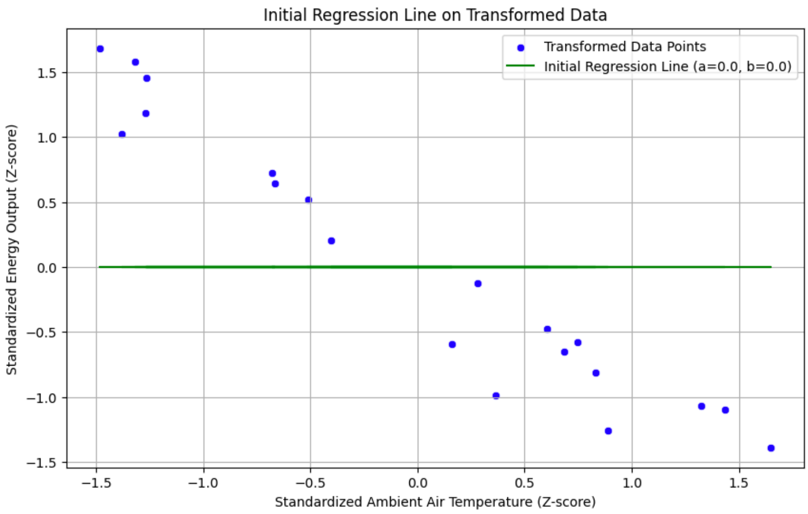
Task 5: Perform One Iteration of Gradient Descent

Hypothesis:

We hypothesised that initialising the regression parameters with arbitrary values and performing one iteration of gradient descent would lead to a reduction in the MSE cost function. This would demonstrate the effectiveness of gradient descent in optimising the regression model by adjusting the parameters to minimise the error.

Process and Results:

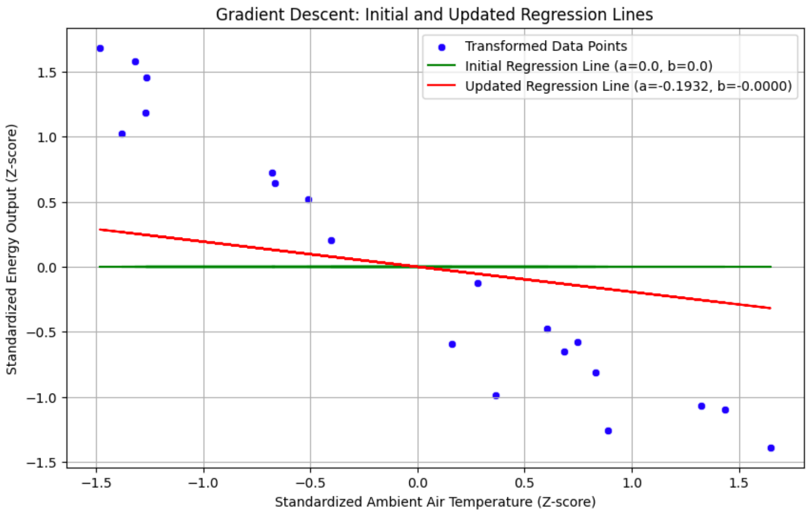
Initial Parameters: a = 0.0; b = 0.0
Initial Cost (MSE): 1.0000
Computed Gradients: $\frac{\partial J}{\partial a} = 1.9320 \quad \frac{\partial J}{\partial b} = 0.0000$



- **Updated Parameters After One Iteration:**

  a=−0.1932
  b = -0.0000

- **New Cost (MSE):** 0.6641



Interpretation: The cost decreased from 1.0000 to 0.6641, confirming that the gradient descent step effectively reduced the discrepancy between the model's predictions and the actual data. The slope a decreased from 0.0 to −0.1932, moving in the direction that reduces the cost and aligning with our expectation of a negative relationship between x and y.

Task 6: Iterate Gradient Descent Until Convergence

Hypothesis: We hypothesised that implementing an iterative gradient descent algorithm with a backtracking line search would efficiently find the optimal regression parameters that minimise the MSE cost function. The backtracking line search would dynamically adjust the learning rate to ensure stable and effective convergence.

Results:

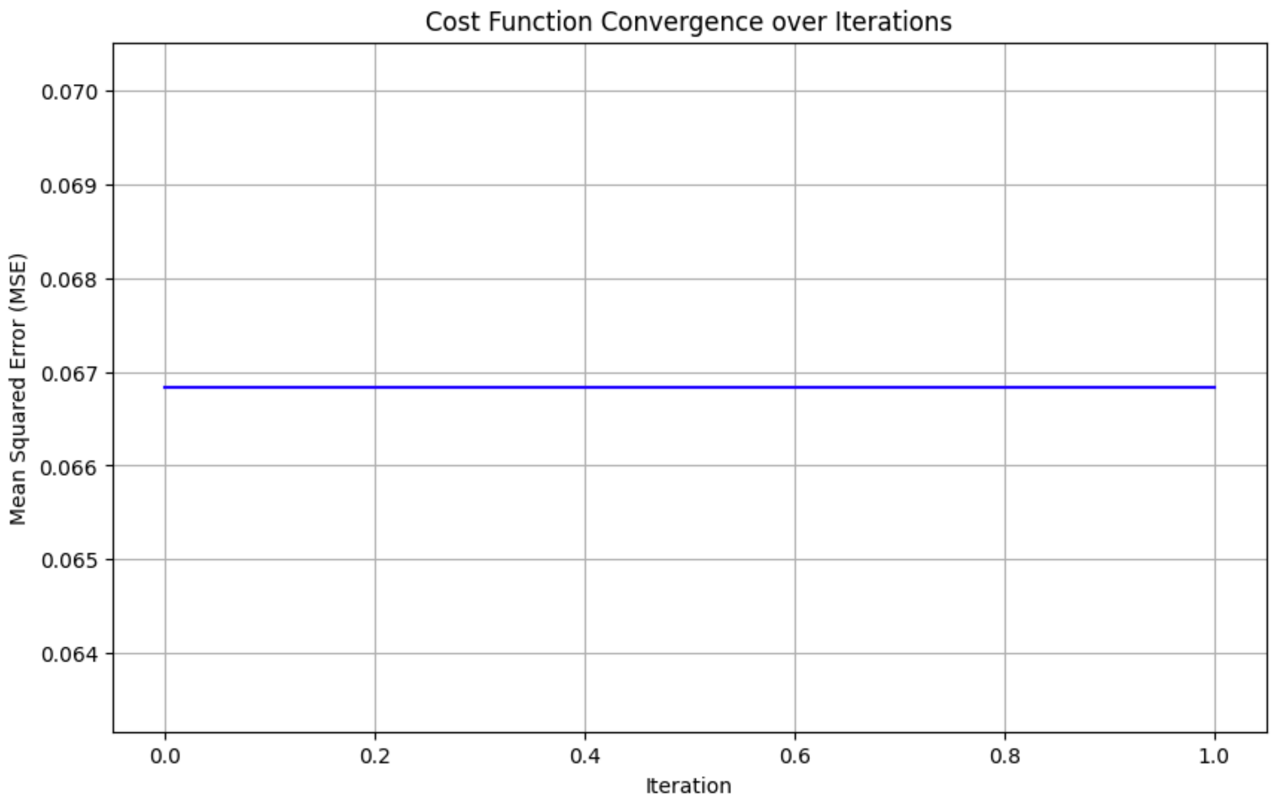
Convergence Achieved: At iteration 1
Final Parameters: a = -0.9660, b = -0.0000
Final Cost (MSE): 0.0668

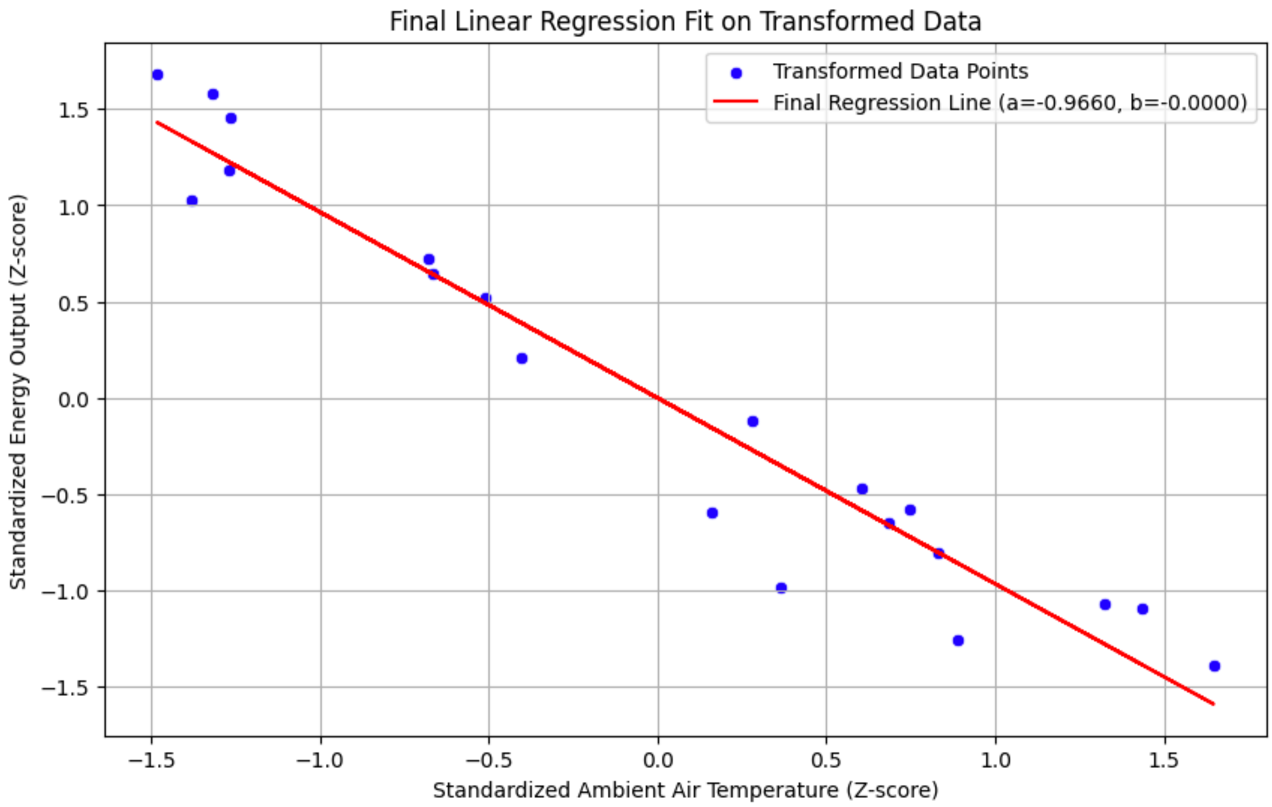
Interpretation:

The algorithm quickly found the optimal parameters due to the linear nature of the data and appropriate learning rate adjustment. The cost function decreased significantly, indicating that the gradient descent effectively optimised the regression parameters. The final regression line closely fits the data points, accurately reflecting the strong negative correlation between x and y.

Effect of Step Size:

If the learning rate is set too large, it can cause the optimisation process to overshoot the minimum, leading to divergence or oscillations. Conversely, a learning rate that is too small results in slow convergence, requiring many iterations to reach the minimum. The backtracking line search method addresses these issues by dynamically adjusting the learning rate, preventing overshooting and promoting stable convergence.





Task 7: "When solving a least squares problem, why might we prefer an iterative method over a "closed form" solution?"

While the closed-form solution provides an exact answer for linear regression, iterative methods offer several advantages. For large datasets, the closed-form solution can be computationally expensive due to the need to compute a matrix inverse, whereas iterative methods scale more efficiently with dataset size.

Additionally, iterative methods such as gradient descent provide greater numerical stability, especially when handling ill-conditioned matrices. They are also more flexible, as they can easily incorporate regularisation techniques and extend to non-linear models, making them a versatile choice for practical applications.

A.II. Polynomial Regression

Task 1: Define Cost Function and Update Rules for Quadratic Model

Quadratic Model:

We extended the linear model to a quadratic one to capture non-linear relationships:

$h_\theta(x) = \theta_2 x^2 + \theta_1 x + \theta_0$

Chosen Cost Function:

The Mean Quartic Error (MQE) was selected:

$$J(\theta) = \frac{1}{4n} \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right)^4$$

Partial Derivatives:

With respect to θ_0:

$$\frac{\partial J}{\partial \theta_0} = -\frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right)^3$$

With respect to θ_1

$$\frac{\partial J}{\partial \theta_1} = -\frac{1}{n} \sum_{i=1}^{n} x^{(i)} \left(y^{(i)} - h_\theta(x^{(i)})\right)^3$$

With respect to θ_2

$$\frac{\partial J}{\partial \theta_2} = -\frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)}\right)^2 \left(y^{(i)} - h_\theta(x^{(i)})\right)^3$$

Task 2: Optimise Quadratic Model with Gradient Descent

Optimisation Process:

Initial Parameters: $\theta_0 = 0.0, \theta_1 = 0.0, \theta_2 = 0.0$
Learning Rate (α): 0.01
Maximum Iterations: 1000
Convergence Criterion: Change in cost function less than 1×10^{-6}

Results:

Final Parameters:

$$\theta_2 = 0.1258$$
$$\theta_1 = -0.8752$$
$$\theta_0 = -0.1448$$

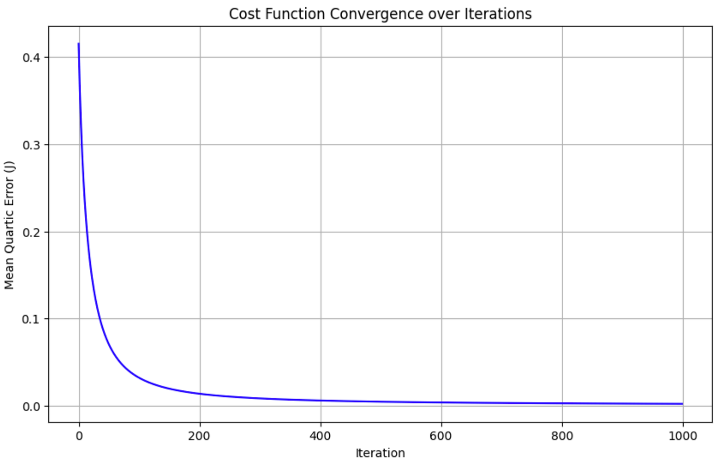
Final Cost (MQE): 0.002

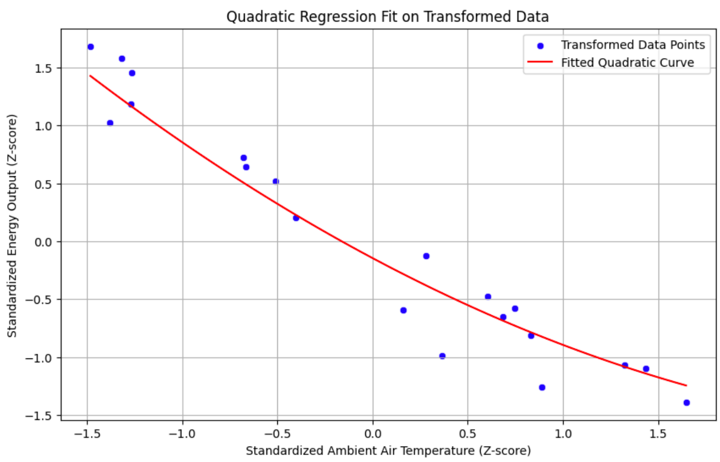
```
Shape of X: (20,)
Shape of Y: (20,)
Shape of X_poly: (20, 3)
Iteration 200: Cost = 0.013722, Theta = [ 0.08797277 -0.63820228 -0.03548725]
Iteration 400: Cost = 0.005947, Theta = [ 0.10458155 -0.75198954 -0.07488145]
Iteration 600: Cost = 0.003725, Theta = [ 0.11481421 -0.81048046 -0.10391522]
Iteration 800: Cost = 0.002723, Theta = [ 0.1214378  -0.84823352 -0.12664001]
Iteration 1000: Cost = 0.002183, Theta = [ 0.12581879 -0.87515697 -0.14481383]
Maximum iterations reached without full convergence.

Final Parameters after 1000 iterations:
Theta2 (quadratic term): 0.125819
Theta1 (linear term): -0.875157
Theta0 (intercept): -0.144814
```

Visualisation:

The fitted quadratic curve was plotted against the standardised data points. The curve closely follows the trend of the data points, demonstrating a good fit and capturing the curvature in the data.





Interpretation:

The model parameters reveals important features of the relationship between x and y. The positive quadratic term suggests that the relationship between x and y is not strictly linear but exhibits a curvature, capturing a non-linear trend in the data. Meanwhile, the negative linear term maintains the overall decreasing trend, indicating that as x increases, y generally decreases. Finally, the intercept is relatively small, which aligns with the fact that the data has been standardised and centred around zero.

Conclusion

This analysis confirms a strong inverse relationship between ambient air temperature and energy output in a combined cycle power plant. Linear regression effectively captured this relationship, with standardised data revealing that energy output decreases significantly as ambient temperature increases. The negative slope obtained from both the closed-form solution and iterative gradient descent validates this inverse correlation.

Extending the model to polynomial regression allowed us to capture non-linear patterns in the data, improving the fit and providing a more nuanced understanding of the relationship. The positive quadratic term in the polynomial model indicates a curvature that the linear model could not capture.

The use of gradient descent with appropriate cost functions proved effective in optimising both linear and polynomial models. Iterative methods demonstrated advantages over closed-form solutions, particularly in terms of computational efficiency and flexibility.