US15: Linear Regression Analysis

João Amorim

June 9, 2024

1 Data and Results

1.1 Dataset

The dataset used for this analysis consists of two files: "water_consumption_updated.csv" and "Area.csv". The former contains information about water consumption in different parks, while the latter provides the area of each park.

1.2 Linear Regression

Linear regression is employed to model the relationship between the area of the park and the average monthly cost spent on water consumption. This analysis aims to predict the monthly cost associated with water consumption in each park based on its size.

1.2.1 Linear Regression Equation

The linear regression equation is given by:

$$\hat{y} = \beta_0 + \beta_1 x$$

where $\hat{y}$ is the predicted monthly water consumption cost, x is the area of the park, and β_0 and β_1 are the intercept and slope coefficients, respectively.

1.3 Best-Fitting Line

The linear regression model will yield the coefficients for the best-fitting line, allowing us to visualize the relationship between park area and monthly water consumption cost.

2 Analysis and Interpretation

The analysis will focus on interpreting the results obtained from the linear regression model.

2.1 Visualization

A scatter plot illustrating the best-fitting line along with the observed data points will be presented to visually assess the fit of the linear regression model.

2.2 Evaluation

The goodness of fit of the linear regression model will be evaluated using statistical metrics such as mean squared error (MSE).

2.2.1 Mean Squared Error (MSE)

The mean squared error is calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where n is the number of observations, y_i is the observed monthly water consumption cost, and $\hat{y}_i$ is the predicted monthly water consumption cost.

2.3 Conclusion

The conclusions drawn from the analysis will summarize the effectiveness of the linear regression model in predicting the monthly water consumption cost based on park area.