Chapter 8 Linear Regression

A graph of a line of best fit

Description automatically generated

**What is Linear Regression?**

Linear regression is a simple statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (x) by fitting a straight line through the data.

**Explanation of the Diagram:**

* **Blue points:** Represent the data points, showing the observed relationship between the independent variable (x) and the dependent variable (y).
* **Red line:** The line of best fit, calculated to minimize the differences (errors) between the observed data points and the predictions.

This visually demonstrates how linear regression models the relationship by finding the optimal slope and intercept of the line

**Code Example**

The code (used in my hint videos for exercise) used **both statsmodels (OLS)** and **sklearn.linear\_model (LinearRegression)** for different purposes, as each library has distinct strengths:

**Reasons for Using Both:**

1. **statsmodels for Statistical Analysis**:
   * OLS from statsmodels provides a detailed statistical summary, including:
     + **p-values** for each predictor to assess significance.
     + **R-squared** and **adjusted R-squared** for model performance.
     + Diagnostic measures like confidence intervals, F-statistics, etc.
   * Statsmodels examine predictors and identify which variables (e.g., CoarseAggregate and FineAggregate) should be excluded based on their p-values.
2. **sklearn.linear\_model for Prediction and Scalability**:
   * LinearRegression from sklearn is more efficient for training and making predictions, especially for large datasets.
   * It integrates seamlessly with scikit-learn’s tools, such as pipelines, feature scaling, and model evaluation (e.g., cross-validation).
3. **Reason for Redundancy**:
   * The code may have started with statsmodels to explore and refine the model (e.g., removing insignificant predictors based on p-values).
   * Later, it switched to sklearn for ease of prediction, as LinearRegression is better suited for operational tasks, such as fitting on a cleaned dataset and predicting on new data.

**Analysis of the following result:**

A screenshot of a computer screen

Description automatically generated

To determine which variable to remove, focus on the **p-values** in the regression results. Variables with **p-values greater than 0.05** are not statistically significant at the 5% significance level, meaning they may not have a meaningful contribution to predicting the dependent variable.

**Variables with p-values greater than 0.05:**

* **CoarseAggregate**: p-value = 0.186
* **FineAggregate**: p-value = 0.272

These variables are candidates for removal because they are not statistically significant predictors of the dependent variable (CompressiveStrength) in the presence of the other variables.

**Recommendation:**

Remove both **CoarseAggregate** and **FineAggregate** because:

1. Their p-values are above the 0.05 threshold, indicating they do not significantly contribute to the model.
2. By removing them, you can simplify the model without sacrificing much predictive power (as indicated by the R-squared value).

**More on p-values**

A screenshot of a paper

Description automatically generated

A screenshot of a test

Description automatically generated

**Let’s look at R square value!**

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated