Contents

Chapter 1

Multivariate Linear Regression

1.1 Introduction

Multivariate linear regression is a statistical technique used to model the linear relationship between a dependent variable and multiple independent variables. Unlike simple linear regression, which involves only one independent variable, multivariate linear regression allows for the analysis of how changes in several predictors simultaneously affect the outcome. It's a cornerstone in statistical modeling and machine learning, providing a powerful tool for prediction, inference, and understanding complex relationships in data.

The core idea is to find a linear equation that best describes how the dependent variable can be predicted from a set of independent variables. This "best fit" is typically determined by minimizing the sum of squared differences between the observed and predicted values, a method known as Ordinary Least Squares (OLS).

1.1.1 Why it matters

- **Prediction:** It allows for forecasting the value of a dependent variable based on multiple influencing factors.

- **Inference:** It helps in understanding which independent variables have a significant impact on the dependent variable and the magnitude and direction of those effects.

- **Control for Confounding:** By including multiple variables, it can account for the effects of other factors, leading to a more accurate understanding of individual predictor relationships.

1.1.2 Scope

Multivariate linear regression is widely applied across various fields, including economics (predicting GDP based on multiple indicators), finance (stock price prediction), social sciences (factors influencing educational attainment), engineering (predicting material properties), and healthcare (predicting disease risk).

1.2 Deep Explanation

At its heart, multivariate linear regression seeks to establish a linear equation of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{1.1}$$

Where:

- Y: The dependent (response) variable.
- $X_1, X_2, \ldots, X_p$: The p independent (predictor) variables.
- β_0: The intercept, representing the expected value of Y when all $X_i = 0$.
- $\beta_1, \beta_2, \ldots, \beta_p$: The coefficients for each independent variable, representing the change in Y for a one-unit change in the corresponding X_i, holding all other variables constant.
- ϵ: The error term, representing the portion of Y that cannot be explained by the linear relationship with the predictors.

1.2.1 Matrix Notation

For computational efficiency, especially with large datasets, the equation is often expressed in matrix form:

$$Y = X\beta + \epsilon \tag{1.2}$$

Where:

- Y: A vector of observed dependent variable values ($n \times 1$).
- X: The design matrix ($n \times (p+1)$), containing a column of ones for the intercept and columns for each independent variable.
- β: A vector of coefficients ($(p+1) \times 1$), including $\beta_0, \beta_1, \ldots, \beta_p$.
- ϵ: A vector of error terms ($n \times 1$).

1.2.2 Estimating the Coefficients (Ordinary Least Squares - OLS)

The goal of OLS is to find $\hat{\beta}$ that minimizes the sum of squared residuals:

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \tag{1.3}$$

By setting the derivative to zero, we get the normal equation:

$$(X^T X)\hat{\beta} = X^T Y \tag{1.4}$$

Solving for $\hat{\beta}$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{1.5}$$

provided that $(X^T X)$ is invertible.

1.2.3 Assumptions of OLS Linear Regression

- **Linearity:** The relationship between Y and X is linear.
- **Independence of Errors:** The error terms are independent.
- **Homoscedasticity:** The variance of the errors is constant.
- **Normality of Errors:** Errors are normally distributed.
- **No Multicollinearity:** Independent variables are not highly correlated.
- **No Endogeneity:** Predictors are uncorrelated with the error term.

1.2.4 Interpretation of Coefficients

Each β_j represents the expected change in Y for a one-unit increase in X_j, holding all other variables constant.

1.2.5 Model Evaluation

- **R-squared (R^2):** Proportion of variance in Y explained by X:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} \tag{1.6}$$

- **Adjusted R-squared:** Adjusts for the number of predictors.
- **F-statistic:** Tests overall model significance.
- **t-tests:** Evaluate the significance of individual coefficients.
- **Residual Analysis:** Used to verify OLS assumptions.

1.3 Example: Python Implementation

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Generate synthetic data
np.random.seed(42)
n_samples = 100
hours_studied = np.random.rand(n_samples) * 10
attendance = np.random.rand(n_samples) * 5
previous_score = np.random.rand(n_samples) * 100

# True coefficients
beta_0 = 30
beta_hours = 5
beta_attendance = -3
```

```
beta_prev = 0.5

errors = np.random.randn(n_samples) * 5
scores = (beta_0 +
          beta_hours * hours_studied +
          beta_attendance * attendance +
          beta_prev * previous_score +
          errors)

data = pd.DataFrame({
    'Hours_Studied': hours_studied,
    'Attendance': attendance,
    'Previous_Score': previous_score,
    'Score': scores
})

X = data[['Hours_Studied', 'Attendance', 'Previous_Score']]
y = data['Score']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
   =0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

print("Model Coefficients:")
for feature, coef in zip(X.columns, model.coef_):
    print(f"{feature}: {coef:.2f}")
print(f"Intercept: {model.intercept_:.2f}")

print(f"\nMSE: {mean_squared_error(y_test, y_pred):.2f}")
print(f"R-squared: {r2_score(y_test, y_pred):.2f}")
```

Listing 1.1: Python Implementation for Multivariate Linear Regression

```
Model Coefficients:
Hours_Studied: 4.88
Attendance: -3.07
Previous_Score: 0.50
Intercept: 30.60

MSE: 26.68
R-squared: 0.88
```

Listing 1.2: Sample Output

1.4 Summary / Key Takeaways

- **Definition:** Models linear relationships with multiple predictors.
- **Equation:** $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
- **Matrix Form:** $Y = X\beta + \epsilon$
- **Estimator:** $\hat{\beta} = (X^T X)^{-1} X^T Y$
- **Evaluation:** Use R^2, Adjusted R^2, F-statistic, t-tests.