Jacqueline Gomez

Multiple Linear Regression - Overview

4 February 2024

PURPOSE

Multiple linear regression is used to predict the value of a **dependent variable** (outcome) based on the value of multiple **independent variables** (**predictors**).

MODEL EQUATION

$$\boldsymbol{Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k + \epsilon}$$

- $\boldsymbol{Y}$ = dependent variable (the outcome you are trying to predict; continuous random variable)
- $\boldsymbol{x_1, x_2, x_3, \ldots, x_k}$ = independent variable (the predictor variable; continuous or random variable)
- $\boldsymbol{b_0}$ = intercept (value of Y when X is 0)
- $\boldsymbol{b_1, b_2, b_3, \ldots, b_k}$ = coefficients for each independent variable (change in Y for a one-unit change in X)
- $\boldsymbol{k}$ = total count of independent variables (index of the last independent variable)
- $\boldsymbol{\epsilon}$ = error term (the difference between observed and predicted values)

ASSUMPTIONS

□ **Linearity**: The relationship between the independent and dependent variables is linear.

□ **Independence**: Observations are independent of each other.

□ **Homoscedasticity**: Constant variance of errors across all levels of X.

□ **Normality**: The residuals (errors) of the model should be approximately normally distributed.

MODEL FIT AND EVALUATION

- **F-statistics:** Used to determine if the overall regression model is statistically significant.
- **Adjusted R-squared**: Adjusted for the number of predictors in the model; useful when comparing models with different numbers of predictors.
- **p-value**: Tests the hypothesis that the slope (b_1) is significantly different from zero. A low p-value (< 0.05) indicates a significant relationship.
- **Variance Inflation Factor (VIF)**: Measures how multicollinearity is inflating the variance of coefficients. A VIF over 10 suggests a problematic correlation between variables.
- **Correlation Coefficient**: Represents the degree of the relationship between two independent variables.

Coefficient	Correlation	Multicollinearity
-1	Perfect negative	No multicollinearity
0	No correlation	Potential multicollinearity
1	Perfect positive	No multicollinearity

INTERPRETATION OF RESULTS

- **Coefficients** $(\boldsymbol{b_1, b_2, b_3, \ldots, b_k})$: Indicates how much the dependent variable is expected to increase (or decrease) when the independent variable increases by one unit, holding all other independent variables constant.
- **Intercept** $(\boldsymbol{b_0})$: The predicted value of Y when X is zero. Interpret with caution, especially if X cannot be zero in practical scenarios.

LIMITATIONS

- **Causation vs. Correlation**: Simple linear regression shows relationships but does not imply causation.
- **Outliers**: Influential outliers can skew results and lead to misleading conclusions.
- **Overfitting**: Including too many predictors can limit the model's performance on new data.
- **Model Complexity**: Only suitable for simple relationships; more complex relationships may require multiple regression or other techniques.