Jacqueline Gomez

Simple Linear Regression - Overview

28 January 2024

PURPOSE

Simple linear regression is used to predict the value of a dependent variable (outcome) based on the value of an independent variable (predictor).

MODEL EQUATION

$$\boldsymbol{Y = b_0 + b_1 X + \epsilon}$$

- $\boldsymbol{Y}$ = dependent variable (the outcome you are trying to predict; continuous random variable)
- $\boldsymbol{X}$ = independent variable (the predictor variable; continuous or random variable)
- $\boldsymbol{b_0}$ = intercept (value of Y when X is 0)
- $\boldsymbol{b_1}$ = slope (change in Y for a one-unit change in X)
- $\boldsymbol{\epsilon}$ = 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

- **R-squared** $(\boldsymbol{R^2})$: Represents the proportion of variance in the dependent variable that can be explained by the independent variable. Ranges from 0 to 1; higher values indicate a better fit.
- **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.

INTERPRETATION OF RESULTS

- **Slope** $(\boldsymbol{b_1})$: Indicates how much the dependent variable is expected to increase (or decrease) when the independent variable increases by one unit.
- **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.
- **Model Complexity**: Only suitable for simple relationships; more complex relationships may require multiple regression or other techniques.