Multiple Linear Regression

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of Simple linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.
Multiple linear regression technique enables analysts to determine the variation of the model and the relative contribution of each independent variable in the total variance.

Multiple Linear Regression Line Equation is given by,

$Yi= \beta 0 + \beta 1xi1 + \beta 2xi2+.....+ \beta pxip + \epsilon$

Where:

- **yi** is the dependent or predicted variable
- **β0** is the y-intercept, i.e., the value of y when both xi and x2 are 0.
- **β1** and **β2** are the regression coefficients representing the change in y relative to a one-unit change in **xi1** and **xi2**, respectively.
- **βp** is the slope coefficient for each independent variable.
- **ϵ** is the model's random error (residual) term.

Understanding Multiple Linear Regression

Multiple regression is a type of regression where the dependent variable shows a **linear** relationship with two or more independent variables. It can also be **non-linear**, where the dependent and independent variables do not follow a straight line.
Both linear and non-linear regression track a particular response using two or more variables graphically. However, non-linear regression is usually difficult to execute since it is created from assumptions derived from trial and error.

Assumptions of Multiple Linear Regression:

1. **Linearity**: The relationships between the independent and dependent variables are linear. The best way to check the linear relationships is to create scatterplots and then visually inspect the scatterplots for linearity. If the relationship displayed in the scatterplot is not linear, then the analyst will need to run a non-linear regression or transform the data using statistical software.

2. **Independence**: Observations are independent of one another. To test for this assumption, we use the Durbin Watson statistic.
3. **Homoscedasticity:** The variance of the residuals remains consistent across all levels of the independent variables. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity.
4. **Normality**: Residuals adhere to a normal distribution. Multivariate normality occurs when residuals are normally distributed. To test this assumption, look at how the values of residuals are distributed. It can also be tested using two main methods, i.e., a histogram with a superimposed normal curve or the Normal Probability Plot method.
5. **No Multicollinearity**: Independent variables are not highly correlated with each other. When independent variables show multicollinearity, there will be problems figuring out the specific variable that contributes to the variance in the dependent variable. The best method to test for the assumption is the Variance Inflation Factor method.

Evaluation:

Assessing the performance of MLR models entails leveraging various metrics such as R-squared, adjusted R-squared, mean squared error (MSE), root mean squared error (RMSE), and F-statistic. These measures provide insights into the model's predictive power and overall significance.

Applications of Multiple Linear Regression in various domains:

A. **Economics and Finance**:
   1. **Economic Forecasting:** MLR models can predict economic indicators such as GDP growth, inflation rates, and unemployment rates based on factors like government spending, interest rates, and consumer sentiment.

2. **Financial Analysis:** MLR is used to analyze the relationship between financial variables such as stock prices, interest rates, and company performance, aiding in investment decision-making and risk management.

B. Marketing:

1. **Market Research:** MLR helps businesses understand consumer behavior by analyzing the relationship between marketing expenditures, demographics, and sales figures.
2. **Price Optimization:** MLR models can predict optimal pricing strategies by considering factors such as production costs, competitor prices, and consumer demand.

C. Healthcare:

1. **Clinical Research:** MLR is used to analyze clinical trial data to identify factors influencing treatment outcomes and patient responses.
2. **Healthcare Management:** MLR models predict healthcare utilization rates, hospital readmission rates, and patient satisfaction scores based on demographic information and healthcare services provided.

D. Environmental Science:

1. **Climate Modeling:** MLR is employed to analyze the relationship between environmental variables (e.g., temperature, precipitation) and climatic phenomena such as droughts, hurricanes, and heatwaves.
2. **Ecological Studies:** MLR helps ecologists understand the impact of environmental factors on biodiversity, species distribution, and ecosystem health.

E. Social Sciences:

1. **Education Research:** MLR models analyze educational data to identify factors influencing academic performance, graduation rates, and student retention.
2. **Sociological Studies:** MLR is used to study social phenomena such as crime rates, income inequality, and voting behavior by analyzing demographic, economic, and social data.

F. Operations Research:

1. **Supply Chain Management:** MLR models optimize inventory levels, production schedules, and distribution routes by analyzing factors like demand forecasts, lead times, and transportation costs.
2. **Quality Control:** MLR is employed to analyze manufacturing processes and identify factors contributing to product defects, allowing businesses to improve product quality and reduce waste.

G. Human Resources:

1. **Workforce Management:** MLR models predict employee turnover rates, absenteeism rates, and performance ratings based on factors like compensation, job satisfaction, and organizational culture.
2. **Recruitment and Selection:** MLR helps organizations identify predictors of job performance and make data-driven hiring decisions.

Let's Understand overall process using example of Predicting Housing Prices:

1. Dataset Acquisition:

- Data pertaining to housing prices typically include various features such as the size of the house, number of bedrooms and bathrooms, location (e.g., ZIP code), proximity to amenities (e.g., schools, parks), and historical sales data.
- Additionally, demographic information about the neighborhood, economic indicators, and housing market trends may also be incorporated.

2. Data Preprocessing:

- Missing values are handled through techniques like imputation or deletion.
- Categorical variables (e.g., location) are encoded using methods like one-hot encoding.
- Outliers may be identified and treated using techniques such as winsorization or transformation.

3. **Model Building:**
    - The dataset is divided into training and testing sets for model validation.
    - Multiple linear regression is applied, where the housing price (dependent variable) is regressed on various features such as house size, number of bedrooms, location attributes, and other relevant predictors.
    - The model equation takes the form:

Price = β0 + β1*Size + β2*Bedrooms + β3*Bathrooms + β4*Location + ε
Here, β0 represents the intercept,
β1-β4 denote the regression coefficients for each respective predictor, and
ε represents the error term.

4. **Model Evaluation:**
    - Performance metrics such as R-squared, adjusted R-squared, mean squared error (MSE), and root mean squared error (RMSE) are used to assess the model's goodness-of-fit and predictive accuracy.
    - Residual analysis is conducted to ensure the model assumptions hold, including checking for linearity, homoscedasticity, and normality of residuals.

5. **Interpretation and Inference:**
    - Interpretation of the regression coefficients allows for insights into the impact of each predictor on housing prices.
    - Positive coefficients indicate variables that have a positive effect on housing prices, while negative coefficients imply variables that exert a negative influence.
    - For example, a positive coefficient for house size suggests that larger houses command higher prices, while a negative coefficient for distance from city center indicates that houses farther from urban hubs may be priced lower.

6. **Prediction and Deployment:**
    - Once validated, the MLR model can be deployed to predict housing prices for new or unseen data.
    - Real estate agents, property developers, and prospective buyers can leverage these predictions to make informed decisions regarding pricing, investment, and purchasing strategies.

Below is the file link for code of Multiple linear Regression:
Multiple linear Regression