Multiple Regression Analysis

Multiple regression analysis is an extension of simple linear regression that allows us to model the relationship between one dependent variable and two or more independent variables. This technique helps in understanding how multiple factors influence the dependent variable and predicting outcomes based on several predictors.

1. **Dependent Variable (Y)**: The outcome we are trying to predict or explain.
2. **Independent Variables ($X_1, X_2, ..., X_n$)**: The predictors or factors that we believe have an impact on the dependent variable.
3. **Regression Coefficients ($\beta_0, \beta_1, ..., \beta_n$)**: The parameters that represent the relationship between each independent variable and the dependent variable.
4. **Regression Equation**: The formula that represents the relationship between the dependent and independent variables.

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n + \varepsilon$$

   - β_0 is the intercept.
   - $\beta_1, \beta_2, ..., \beta_n$ are the coefficients for the independent variables.
   - ε is the error term.

Steps to Perform Multiple Regression Analysis

1. **Data Collection**: Gather data for the dependent variable and all independent variables.
2. **Data Entry**: Input the data into an Excel spreadsheet.
3. **Exploratory Data Analysis**: Visualize and summarize the data to understand its structure.
4. **Model Specification**: Define the regression model.
5. **Estimation of Parameters**: Use statistical software or Excel to estimate the coefficients.
6. **Model Diagnostics**: Check the validity of the model through residual analysis and other diagnostic tests.
7. **Interpretation**: Interpret the coefficients to understand the relationship between the variables.

Problem Statement

You are a data analyst working for a real estate company. Your task is to predict the house prices (Y) based on various factors such as square footage (X_1), number of bedrooms (X_2), and age of the house (X_3).

House Price (Y)	Square Footage (X_1)	Number of Bedrooms (X_2)	Age of House (X_3)
300000	200	1	10
320000	250	2	5
380000	300	2	20
420000	350	3	15
440000	400	4	10
390000	300	3	15
400000	350	3	10
450000	400	5	10

Steps to Perform Multiple Regression in Excel

1. **Enter Data into Excel**:
   - Input the given data into an Excel spreadsheet with each column representing one variable.
2. **Create the Regression Model**:
   - Go to the "Data" tab.
   - Click on "Data Analysis" and select "Regression."
   - Define the Input *Y* Range (dependent variable) and Input *X* Range (independent variables).
   - Specify the output range where you want the results to appear.
   - Check the "Labels" box if you have included column labels.
3. **Interpret the Output**:
   - The output will include the coefficients for the intercept and each independent variable.
   - Look at the *R*-squared value to assess the fit of the model.
   - Check the *p*-values to determine the significance of each coefficient.
   - **Regression Statistics**:
     - Multiple R: 0.996
     - R Square: 0.993
     - Adjusted R Square: 0.988
   - **Coefficients**:
     - Intercept: 155443.2
     - Square Footage: 520
     - Number of Bedrooms: 12821.5
     - Age of House: 2476.9

- **Interpretation**:
  - The R-squared value of 0.993 indicates that 99.3% of the variability in house prices is explained by the model.
  - For every additional square foot, the house price increases by $520.
  - Each additional bedroom increases the house price by $12,821.
  - Each additional year of the house's age increases the price by $2,477. (???)

Assignment 1

You are tasked with predicting the monthly electricity consumption kWh (*Y*) of households based on the following factors: household size (X_1), average monthly income (X_2), and number of electronic appliances (X_3).

Electricity Consumption (*Y*)	**Household Size (X_1)**	**Average Monthly Income (X_2)**	**Number of Appliances (X_3)**
500	3	4000	10
550	4	4200	11
650	5	4800	13
780	6	5200	16
850	7	5500	17
700	4	5000	14
800	5	5200	16
900	6	5600	18

Tasks

1. **Enter Data**: Input the provided data into an Excel spreadsheet.
2. **Perform Multiple Regression Analysis**:
   - Use the Data Analysis Toolpak in Excel to perform multiple regression analysis.
   - Define the dependent variable (Electricity Consumption) and independent variables (Household Size, Average Monthly Income, Number of Appliances).
3. **Interpret Results**:
   - Record the coefficients for the intercept and each independent variable.
   - Determine the *R*-squared value and interpret the fit of the model.
   - Check the *p*-values to determine the significance of each coefficient.
4. **Make Predictions**:
   - Predict the electricity consumption for a household with 5 members, an average monthly income of $5000, and 15 electronic appliances.

Assignment 2

You are an analyst working for a company that wants to predict the sales revenue (Y) based on various factors such as advertising expenditure (X_1), price of the product (X_2), and number of salespersons (X_3).

Data:

Sales Revenue (Y)	**Advertising Expenditure (X_1)**	**Price of Product (X_2)**	**Number of Salespersons (X_3)**
200000	50000	12	10
250000	55000	15	12
300000	65000	20	15
350000	75000	22	18
400000	80000	28	22
450000	95000	30	25
500000	110000	34	28
550000	120000	35	30
600000	130000	40	32
650000	140000	42	35

Tasks:

1. **Data Entry**:

o Enter the provided data into an Excel spreadsheet.

2. **Perform Multiple Regression Analysis**:

o Use the Data Analysis Toolpak in Excel to perform multiple regression analysis.

o Define the dependent variable (Sales Revenue) and independent variables (Advertising Expenditure, Price of Product, Number of Salespersons).

3. **Interpret Results**:

o Record the coefficients for the intercept and each independent variable.

o Determine the R-squared value and interpret the fit of the model.

o Check the p-values to determine the significance of each coefficient.

4. **Make Predictions**:

o Use the regression equation to predict the sales revenue for a scenario where the advertising expenditure is $150,000, the price of the product is $20, and the number of salespersons is 40.

The assumptions of a linear multivariate regression model:

1. Linearity of Relationships

- The relationship between the independent variables (predictors) and each dependent variable is linear. The model must be correctly specified, including all relevant predictors and excluding irrelevant ones.

2. Homoscedasticity (Constant Variance of Errors)

- The variance of the residuals (errors) is constant across all levels of the independent variables. In other words, errors do not systematically increase or decrease across the range of the data.

3. Independence of Error Terms

- The error terms (residuals) for different dependent variables should not be correlated. This means the covariance of error terms across equations is zero.

4. Normality of Error Terms

- The error terms for each dependent variable are normally distributed. This assumption is particularly important for hypothesis testing and confidence intervals.

5. Independence of Observations

- Observations must be independent of each other. This assumption is especially critical in cross-sectional data. For time-series data, residual autocorrelation should be tested (e.g., using the Durbin-Watson test).

6. No Multicollinearity

- The independent variables should not be highly correlated with one another. High multicollinearity can lead to unstable parameter estimates. This issue can be diagnosed using Variance Inflation Factors (VIF).

7. Full Rank of the Design Matrix

- The matrix of independent variables must have full rank, meaning no independent variable is a linear combination of others. If this assumption is violated, the model parameters cannot be uniquely estimated.

8. Sufficient Sample Size

- The sample size should be large enough relative to the number of parameters in the model. Generally, more predictors require a larger sample size to ensure reliable estimates.

9. Correct Model Specification

- The model should be properly specified, including all relevant variables and excluding irrelevant ones. Omitting key predictors or including unnecessary ones can lead to biased or inefficient estimates.

10. Non-Zero Variance of Independent Variables

- The independent variables must vary across observations; a variable with no variance cannot contribute to explaining the dependent variable.

Violations of these assumptions can lead to biased, inefficient, or inconsistent estimates. Diagnostic tests and graphical methods are often used to detect violations, and remedies such as data transformations, robust estimation methods, or alternative modelling approaches can be applied as needed.